Computational tools and resources for prediction and analysis of gene regulatory regions in the chick genome

Mohsin A. F. Khan[1], Luz Mayela Soto-Jimenez[1,2], Timothy Howe[1], Andrea Streit[3], Alona Sosinsky[4] and Claudio D. Stern[1]

1. Department of Cell & Developmental Biology, University College London, Gower Street (Anatomy Building), London WC1E 6BT, U.K.
2. Programa de Ciencias Genomicas, Universidad Nacional Autonoma de Mexico, Campus Cuernavaca, Av. Universidad s/n Col. Chamilpa 62210, Cuernavaca, Morelos, Mexico.
3. Department of Craniofacial Development and Stem Cell Biology, King's College London, Guy's Campus, London SE1 9RT, U.K.
4. Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck College - University of London, Malet Street, London WC1E 7HX, U.K.

Abstract

The discovery of cis-regulatory elements is a challenging problem in bioinformatics, owing to distal locations and context-specific roles of these elements in controlling gene regulation. Here we review the current bioinformatics methodologies and resources available for systematic discovery of *cis*-acting regulatory elements and conserved transcription factor binding sites in the chick genome. In addition, we propose and make available, a novel workflow using computational tools that integrate CTCF analysis to predict putative insulator elements, enhancer prediction and TFBS analysis. To demonstrate the usefulness of this computational workflow, we then use it to analyze the locus of the gene Sox2 whose developmental expression is known to be controlled by a complex array of *cis*-acting regulatory elements. The workflow accurately predicts most of the experimentally verified elements along with some that have not yet been discovered. A web version of the CTCF tool, together with instructions for using the workflow can be accessed from

http://www.xxxx.com. For local installation of the tool, relevant Perl scripts and instructions are provided in the directory named “code” in the supplementary materials.

Introduction

The control of the precise spatial and temporal expression of genes is a fundamental aspect of development. In the developing embryo, the complex biological machinery that governs this precision has a remarkable capacity to process an enormous number of regulatory cues for various biological processes. This results in sets of time-dependent and tissue-specific regulatory outputs, critical in orchestrating different stages of embryonic development. Capturing these transcriptional activation states by the embryo at the right stage and time depends on several factors including the position of the gene in the genome, its local chromatin structure and the transcriptional regulatory elements associated with each gene (Maston *et al.*, 2006; Vogelmann *et al.*, 2011). Indeed the core promoters of genes, together with nearby proximal regulatory elements, are essential for proper initiation of transcription via recruitment of RNA polymerase II. However, their participation alone is not sufficient to regulate the process of transcription because distal *cis*-acting regulatory elements, such as enhancers, silencers, and insulators act in concert with promoters to streamline the process of transcription (Figure 1). This poses two challenges. First, regulatory elements such as enhancers are not necessarily located close to the genes they regulate, sometimes having the ability to act over considerably large distances in the genome. To understand how enhancers are constrained to act specifically within appropriate chromosomal domains is, therefore, a fundamentally important question. Second, the presence of multiple regulatory transcription factor binding sites (TFBS) within enhancers confers combinatorial control of regulation, making it difficult to decipher their role in the context of spatial and temporal gene expression.

Chromosome conformation capture (3C) studies have shown that long-range enhancer function can be mediated by chromatin loops, hence facilitating the juxtaposition of distant enhancer-bound transcription factors and their cognate promoters (Barrett *et al.*, 2012; Cullen *et al.*, 1993; Dekker *et al.*, 2002; Raab and Kamakaka, 2010). It has been proposed that the well characterised insulator element CCCTC-binding factor (CTCF) might be responsible for inducing these chromatin loops by binding to specific insulator sites, together with the DNA-binding protein Cohesin (Feeney and Verma-Gaur, 2012; Kim *et al.*, 2011). This CTCF-mediated looping mechanism may provide a physical basis for the segregation of functional domains by shielding biologically relevant enhancer-promoter interactions from inappropriate regulatory interactions outside of these functional domains (Cuddapah *et al.*, 2009; Dean, 2011; Kornblihtt, 2012). Therefore, CTCF binding sites can predict the position

of putative insulator regions, which can estimate the likely range of influence of genes and enhancers within a region.

Here we review the current bioinformatics methodologies and resources available for systematic identification of regulatory elements in the chick genome. We focus on computational methodologies available for the discovery of *cis*-acting regulatory elements and common approaches for Transcription Factor Binding Site (TFBS) analysis. In addition, we propose and make available, a novel workflow using computational tools that integrate tools for CTCF analysis to predict putative insulator elements, enhancer prediction and TFBS analysis. Finally, to demonstrate the usefulness of this computational workflow, we use it to analyze the gene Sox2, whose developmental expression is known to be controlled by a complex array of at least 25 *cis*-acting regulatory elements (Uchikawa *et al.*, 2003; Uchikawa *et al.*, 2004), comparing the results of our bioinformatics analyses with experimentally verified data from the literature.

A review of current computational tools

The task of analysing gene regulation is complicated by the fact that it is context-dependent: genes are regulated in time and space, in different cell types, and also vary between different animals. The particular animal model being studied coupled with specific tissues or cell lines of interest and specific developmental or other physiological processes can affect how one goes about using the currently available computational resources. Analysis using Bioinformatics normally begins with the identification of promoters and enhancers. These methods tend to rely on nucleotide sequence conservation between orthologous genes as criteria to identify putative regulatory elements (Wasserman and Sandelin, 2004). It is generally accepted that the sequences close to a TSS (Transcription Start Site) may be functionally important. However, the identification of these regions is not straightforward and gains complexity with the addition of context-dependent alternative TSSs. The public resource 'Eukaryotic Promoter Database' (http://epd.vital-it.ch/) (Perier *et al.*, 2000) was among the first to make available a collection of non-redundant eukaryotic RNA polymerase II promoters, defined experimentally by a TSS. Although a useful resource, the approach relies on the identification of core promoter elements without taking into account that a single gene can have alternative TSSs. A number of programs have improved the success rate of TSS detection by using training sets containing known promoter regions and CpG islands (site of DNA methylation). Among the most popular ones are PromoterInspector from Genomatix (http://www.genomatix.de/) (Scherf *et al.*, 2000), FirstEF (http://rulai.cshl.org/tools/FirstEF/) (Davuluri, 2003), and Eponine (http://www.sanger.ac.uk/resources/software/eponine/) (Down and Hubbard, 2002). It is

however worth noting that these techniques suffer from some important limitations. First, not all of the TSSs reside proximally to a CpG island. Second, the correlation between CpG islands and promoter regions does not always have a syntenic relationship among different species. Alternative approaches using transcript data, are therefore necessary for further improvement in this area of research (Wasserman and Sandelin, 2004).

An analysis of gene regulation would be incomplete without identifying enhancers since they play a critical role in regulating tissue-specific gene expression (Jin *et al.*, 2011). The VISTA Enhancer Browser (http://enhancer.lbl.gov/) (Visel *et al.*, 2007) is a popular resource which facilitates comparative genome analysis for the purpose of discovering sets of highly conserved non-coding DNA segments in vertebrates, which can then be tested for enhancer activity. It provides a public database consisting of experimentally validated non-coding fragments found to be highly conserved across vertebrate species including chick, and showing enhancer activity in transgenic mice. As a part of the selection procedure prior to *in vivo* testing, conservation together with relevant experimentally-determined epigenetic enhancer marks (from ChIP-Seq experiments) are used as criteria to identify putative enhancer sequences. However, to date, only 1760 predicted elements from this database have been tested *in vivo*, of which just over half (893) were found to have enhancer activity (http://enhancer.lbl.gov/), indicating that in validated studies, ~ 51% of predicted enhancers (containing conserved TFBSs) have real biological function. Moreover, ChIP-Seq and other methods of active enhancer detection yield context-specific results – therefore data from established cell lines may not include information about the specific regulatory elements involved in the biological process of interest. Due to these limitations, alternative computational strategies for identifying other tissue-specific and time-dependent enhancers become important.

Over the years, phylogenetic footprinting has gained widespread popularity as the gold standard for computational prediction of *cis*-regulatory elements. This approach is based on the assumption that sequence comparison of orthologous genomic regions in closely related species can predict important biological functions (Woolfe *et al.*, 2005). Because mutations accumulate slowly within functional regions of genes, phylogenetic footprinting can identify enhancers as conserved segments of DNA containing similar sets of transcription factor binding sites retained through evolution. The availability of several genome assemblies has simplified the task of identifying and subsequently analyzing these conserved regions. Initially, this relied on constructing pairwise alignments between related species, but resources such as the UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgGateway) (Dreszer *et al.*, 2012), ENSEMBL (http://www.ensembl.org/index.html) (Flicek *et al.*, 2012), ECRbase (http://ecrbase.dcode.org/) (Loots and Ovcharenko, 2007), and Vista portal

(http://genome.lbl.gov/vista/index.shtml) (Brudno *et al.*, 2007) now use multiple species alignments to help carry out phylogenetic footprinting. The resulting inter-species conserved sequences can then be analysed for the presence of TFBSs.

Transcription factor binding site analysis can be performed using known motifs or by discovering enriched *de novo* motifs within the set of sequences. It is common practice to represent known motifs as either a consensus sequence or a Position Frequency Matrix (PFM), where the preference for each of the four bases A, T, G, and C is captured (Hannenhalli, 2008). The TRANSFAC (http://www.gene-regulation.com/pub/databases.html) (Matys *et al.*, 2006), and JASPAR (http://jaspar.cgb.ki.se/) (Sandelin *et al.*, 2004) databases are the two leading resources that compile these motifs from the literature but unlike TRANSFAC, JASPAR provides open data access to their matrices. In addition, the latter provides the JASPAR CORE database, containing a collection of manually curated, non-redundant profiles, which have been validated experimentally for multicellular eukaryotes. It is worth noting that DNA motifs recognized by transcription factors can be short and degenerate; therefore, computational approaches to identify TFBSs can suffer from high error rates. A reliable approach is to combine TFBS analysis with phylogenetic footprinting, as the occurrences of conserved binding sites across multiple closely related species suggests a greater likelihood of the sites being biologically functional.

The choice of a particular tool or resource should be determined by the type of biological question being investigated (Table 1). Several tools and resources are often used in parallel in the form of a workflow. The following section provides a proposed workflow for analysis of *cis*-regulatory elements useful for the study of gene regulation during chick development. This is most useful when performed in parallel with an experimental workflow, such as that described in the accompanying article (Streit *et al.*, 2013).

Proposed workflow

Major challenges affecting the discovery of *cis*-acting regulatory elements include that they can be located very far from the gene they regulate, and that they can regulate several neighbouring genes, located up- or downstream, on either strand. Although comprehensive genome-wide studies of chromatin dynamics in multiple cell types suggest that majority of enhancer activity correlates with the expression of the most proximal genes, enhancers can have the ability to act as long-range regulators; sometimes occupying locations up to 1 megabases away from the gene they regulate (Chan and Kibler, 2005). Because of this, it is important to examine the regulatory architecture of the genome around the genes of interest before focussing on enhancer detection. CTCF is a well characterised insulator protein

known to facilitate shielding of genes within specific regulatory modules, thereby preventing them from being influenced by regulatory elements outside of these functional domains (Cuddapah *et al.*, 2009; Dean, 2011; Kornblihtt, 2012). Therefore, CTCF binding sites can predict the position of putative insulator regions, which can estimate the likely range of influence of genes and enhancers within a region. It is worth noting that in the event of several genes being present within an isolated putative insulator region, it is possible for them to either share the same, or have different regulatory elements. However, in both cases, the sphere of influence of such regulatory regions will be restricted to the length of the segregated insulated domain. The next step is to detect conserved non-coding segments of DNA that may act as enhancers within these insulator regions. Finally, candidate enhancer regions can be analyzed for the presence of transcription factor binding sites to predict regulatory mechanisms that can then be tested experimentally. Taking all of the above into account, we have developed a computational workflow (Figure 2).

CTCF insulator analysis

Genomic binding of CTCF at specific recognition sites induces chromosomal loops, providing a physical basis for domain segregation (Kim *et al.*, 2011). Therefore, our proposed workflow begins with the task of predicting these CTCF-specific binding sites in any chromosome and species of choice. The following is a description of a Perl script that we developed to automate this task (A web version of the program can be accessed from http://www.xxxx.com).

The JASPAR database contains a collection of 913 CTCF binding sites, represented as a Position Frequency Matrix (PFM). The PFM is defined as a $|\Sigma| \times m$ matrix, where m is the length of binding site and $\Sigma = \{A,T,G,C\}$ is the alphabet of permitted symbols, populated with $f(\sigma,j)$, the frequency of symbol σ at position j of the binding site. The result of this method of representation is that the preferences for each of four bases A, T, G, and C are captured at each position of the binding site (Figure 3a). The PFM for CTCF can then be used to scan entire chromosomes to predict CTCF binding sites. To perform this scanning, the CTCF PFM needs to be converted into a Position Weight Matrix (PWM) according to the following equation:

$$w(\sigma,j) = \log_2 \left(\left(f(\sigma,j) + \sqrt{N} \times b(\sigma) \right) / \left(N + \sqrt{N} \right) \right) / b(\sigma)$$

Where $w(\sigma,j)$ is the weight of nucleotide σ at position j, N is the total number of binding sites or the sum of all nucleotide occurrences in the column, and b is the prior background frequency of the nucleotide σ.

The sum of weights for corresponding nucleotides at each column of the matrix then estimates the likelihood of any sequence of length m to be an instance of a CTCF binding site and takes into account the GC content of the genomic region being scanned (Figure 3b).

The Perl script automates the above analysis; starting from the first nucleotide of a selected chromosome, calculates a weighted score in a one-nucleotide sliding window until both strands of the entire chromosome have been scanned for CTCF sites. This procedure is then repeated with randomly shuffled sequences from the same chromosome (to ensure that it maintains the same GC content as the original chromosome) and a probability distribution of weight scores is generated, comparing the number of occurrences of each given weight in the empirical distribution with that in the null distribution (Figure 3c). From this, the False Discovery Rate (FDR) is then computed as follows:

$$\mathbf{FDR} = V / V+S$$

Where V is the number of sites of a given weight in the control sample (random shuffled sequence) and S is the number of sites of a given weight in the test sample (actual chromosome) (Figure 3d). A P-value for each weight is also calculated as follows:

$$\mathbf{P} = A/B$$

Where A is the number of sites with weighted score equal to the cut-off and above in the control sample, and B is the total number of sites in the control sample.

The FDR together with the P-value for each calculated weight of the CTCF motif provides the user with statistical information from which a threshold of significance can be set. A weight score of ≥18.0 with an FDR and P-value of 0 for instance, might generate 1160 CTCF binding sites from the test sample none of which are false positives as indicated by its FDR. On the other hand, a weight score of ≥17.0 with an FDR of 8.5% and P-value of 7.5×10^{-7} might generate 1749 CTCF binding sites, 148 of which are expected to be false positives. After selection of a weight threshold by specifying a cut-off for the FDR, the program will display all CTCF sites with a weight equal to or above the user-defined threshold, together with their genomic coordinates in the input chromosome, weight score of each site, and the strand in which they appear.

CTCF-bound sites can be classified into 1) constitutive sites, where CTCF will be bound at the same genomic location in different tissues and are therefore largely context-independent, and 2) labile sites, which may be involved in tissue-specific gene regulation. It is thought that the former are more likely to act as insulators (Martin *et al.*, 2011). For this reason, as well as because in most situations the most relevant cell line or tissue sample for the problem being studied will not have been analysed experimentally for CTCF binding, we decided to focus on identifying putative constitutive CTCF sites. Having computationally identified significant potential CTCF sites chromosome-wide in human, the next step at this stage of the workflow is to compare these sites to existing ChIP-Seq CTCF-enriched regions from several different tissue samples in human (downloaded from the UCSC genome browser) to see if they constitutively fall in the same genomic locations. For synteny analysis, the process of computationally predicting CTCF sites is then repeated in equivalent chromosomes in chick and mouse, followed by the use of existing ChIP-seq CTCF-enriched datasets (for both chick and mouse) generated from the laboratory of Gomez-Skarmeta (Martin *et al.*, 2011) to find constitutive sites in both species. Coincidence between these experimental results and the computational predictions should predict the most likely constitutive sites, and therefore syntenic putative insulators.

Enhancer discovery

Once candidate insulators encapsulating the gene of interest have been identified, the next stage is to discover enhancers likely to regulate the gene within the insulated region. DREiVe (Discovery of Regulatory Elements in Vertebrates) is a bioinformatics tool for identifying regulatory elements (such as enhancers) as evolutionarily conserved, order-independent clusters of short conserved DNA motifs in vertebrate species (http://dreive.cryst.bbk.ac.uk/) (Yeowell and Sosinsky, 2012). By integrating a traditional pattern discovery algorithm, SPLASH (Califano, 2000), with a novel local permutation clustering algorithm, it offers a platform which relies on the evolutionary conservation of transcription factor binding sites but without requiring prior knowledge of these transcription factors or their cognate binding sites.

DREiVe analysis begins with the task of identifying Short Conserved Motifs (SCMs), which occur at least once in each of a set of orthologous input sequences. This step is carried out by the SPLASH algorithm, which identifies SCMs as conserved motifs represented as regular expressions where rigid sites conserved across all species are denoted by their corresponding nucleic acid symbols (A, T, G, C) and variable positions as wildcards ('.'). There are two specific critical parameters used by SPLASH to determine the class of motifs identified. The first is the 'motif density', which is the minimum number of conserved

residues, *k* that occur over a window length, *w.* The second is the minimum number of matching residues, *l*, which defines the length of the motif. As an example, the constraints set by the parameters *k*=6, *w*=8, and *l*=9 would be satisfied by a motif such as 'AC.T.AGGTA..T'. This is because in a sliding window length of 8 residues, 6 of them are always conserved and the total number of conserved residues defining the length of the motif is equal to 9.

The next step is to discover Local Permutation Clusters (LPCs), which are subsets of conserved SCMs located within a user-defined maximum cluster length, *l* in each of the orthologous species. The discovery of these SCMs within a pre-defined cluster length is order-independent in the sense that the precise order of SCMs in each species-specific cluster is irrelevant to the discovery of LPCs. The PromoClust algorithm is used to detect maximal LPCs, followed by using a heuristic approach to assign a conservation score to each position of the input sequences equal to the length of the SCM. The LPCs that are assigned the highest conservation score are then reported as putative functional enhancers.

Transcription Factor Binding Site Analysis (TFBSA)

Following the identification of candidate enhancers using DREiVe, the next and final stage in our workflow is to scan conserved SCMs present in the DREiVe-predicted enhancers against a library of TRANSFAC and JASPAR PFMs. This enables us to detect sets of conserved transcription factor binding sites in each candidate enhancer sequence. For this, we use 'matrix-scan' from the Regulatory Sequence Analysis Tools (RSAT) workbench (http://rsat.ulb.ac.be/rsat/) (Thomas-Chollier *et al.*, 2011a). Matrix-scan accepts an unlimited number of sequences in FASTA format as the input and requires the user to provide a set of transcription factor matrices such as TRANSFAC and JASPAR PFMs. The program then scans the input sequences against each PFM and at each position of the input sequence, a sequence segment, *S* equal to the length of the PFM is assigned a weighted score (Ws). This is calculated as the log ratio between two probabilities as follows:

Ws = log[P(S|M)/P(S|B)]

Where

M = P(S|M) – the probability of the sequence segment, S, given the PFM model

and P(S|B) – the probability of the sequence segment, S, given the background model

Selecting an appropriate background model is a prerequisite for accurate pattern discovery because it is used to estimate the likelihood of sites occurring by chance alone. Matrix-scan

allows users to specify a particular Markov order as a background model, where an order of *n* suggests that the probability of each nucleotide base is reliant on *n* preceding nucleotide bases in the sequence. Likewise, a Markov order of 0 means that each residue does not depend at all on the preceding bases, and is therefore a Bernoulli model. Our complete workflow is illustrated in Figure 2.

Cis-regulatory analysis of chick *Sox2*

To test the usefulness of our computational workflow to identify biologically significant regions at each stage of *cis* regulatory analysis, we evaluated its potential by analysing the locus of the gene *Sox2. Sox2* is an important gene implicated in cell fate determination especially in embryonic stem cells and neural development (Collignon *et al.*, 1996; Papanayotou *et al.*, 2008; Pevny *et al.*, 1998; Streit *et al.*, 2000; Streit *et al.*, 1998; Uchikawa *et al.*, 1999; Wood and Episkopou, 1999). Several studies have revealed that it is uniformly expressed in the early neural tube, and is regarded as a pan-neural marker in early stages of embryonic development (Darnell *et al.*, 1999; Streit *et al.*, 1997). *Sox2* is expressed in multiple locations during early development, related to its involvement in the regulation of pluripotency in embryonic stem cells and early embryos (Kim *et al.*, 2008), early neural plate development (Rex *et al.*, 1997; Streit *et al.*, 2000; Streit *et al.*, 1997; Uwanogho *et al.*, 1995) and placodal development (Uchikawa *et al.*, 1999). Twenty-five separate enhancers have been identified experimentally by pioneering work from the laboratory of Hisato Kondoh (Uchikawa *et al.*, 2003). These enhancers are located within a region spanning 16.7 kb upstream and 32.5 kb downstream of the single exon *Sox2* gene in chick (Uchikawa *et al.*, 2003). Each enhancer has a specific activity, directing expression to one or a few specific sites of *Sox2* expression in the normal embryo.

Identifying putative insulators of *Sox2*

We started computationally by identifying statistically significant CTCF sites across human chromosome 3, which contains *Sox2*. We decided to use the human as a reference genome because the chick genome assembly is still incomplete and poorly annotated; syntenic relationships between human, mouse and chick are examined at a later stage in the analysis. CTCF analysis shows that using a weight cut-off of ≥18.0 (FDR=4.13%, P-value $=2.5 \times 10^{-7}$), 1160 statistically significant CTCF binding sites are found throughout chromosome 3 (Figure 4) (Supplementary table 1). Although our choice of cut-off was

somewhat arbitrary, we decided to use ≥18 because lower values increase the FDR whereas higher values decrease the number of detected CTCF binding sites significantly (Figure 4). We then compared the coordinates of this set of 1160 sites with those of ChIP-Seq peaks derived from 23 different human cell-lines for CTCF enrichment (from the UCSC genome browser) to identify the most likely constitutive CTCF sites. A total of 348 predicted CTCF sites were found to coincide in both sets of data (computationally predicted and all experimental ChIP-Seq sets), and we therefore define these as "constitutive" (Figure 4). The next step was then to supply these 348 sites to the UCSC genome browser as user tracks to identify the closest candidates encapsulating *Sox2.* From these 348 sites, we found a putative constitutive CTCF site (CTACCAGCAGGGGGCGCAC) (*hg19* coordinates chr3:181,427,485-181,427,504) ~2.2 kb upstream of *Sox2* and another (GTCTGCCCTCTAGAGGCCA) (*hg19* coordinates chr3:182,428,542-182,428,561) ~1 mb downstream of *Sox2* and ~100 kb upstream of the gene ATP11b (Figure 5a). Both sites are located on the sense strand, and have weight scores of 23.6 (FDR=0, P-value=0) and 19.06 (FDR=0, P-value=0) respectively. Furthermore, we repeated this CTCF analysis with equivalent regions in chick and mouse genomes, and found a syntenic region harbouring *Sox2* in both species (Figure 5b and 5c).

In chick, equivalent constitutive CTCF sites were found ~ 10 kb upstream and ~600 kb downstream of *Sox2*. An additional computationally-identified (but not constitutive) site was found ~ 300 kb downstream of *Sox2* and ~100 kb upstream of ATP11b. This, together with the constitutive site upstream of *Sox2,* forms a syntenic region equivalent to that in human (Figure 5b).

These findings collectively suggest the presence of a ~300 kb putative insulator region harbouring *Sox2* in chick, sharing synteny with an equivalent ~ 1 mb region in human and ~700 kb region in mouse (Figure 5).

Computational discovery of *Sox2* enhancers

Our next objective was to use DREiVe to discover putative enhancers within this ~ 1mb candidate insulator region containing *Sox2* in human*.* Performing DREiVe analysis using human as the reference genome (hg19 build) with the parameters k=6, w=8, and l=9 led to the discovery of 98 high scoring LPCs, conserved in human, mouse, chick, lizard, platypus, opossum, cow and elephant (conservation score > 2) (Supplementary Table 2). Within this 1 mb genomic window, a particularly dense ~70 kb region (7% of the 1 mb window) surrounding the *Sox2* gene contained 27 (28%) of the LPCs (Figure 5a). We discovered that

18 of 25 (72%) previously known enhancers of *Sox2* identified by the laboratory of Hisato Kondoh (Uchikawa *et al.*, 2003) overlap with 18 of 27 (67%) of these DREiVe-predicted LPCs (Figure 6). Among those identified by DREiVe were all of the neural *Sox2* enhancers, N1, N2, N3, N4, and N5, the nasal and otic placode enhancers, NOP-1 and NOP-2, and the spinal cord enhancers, SC1 and SC2, all conserved in chick. Among the 7 *Sox2* enhancers not identified by DREiVe included the late lens enhancer, L and the dorsal root ganglia enhancer, NC1 (Figure 6). Similarly, DREiVe predicted 9 conserved LPCs within this region that were not identified experimentally, which suggests some degree of complementarity between experimentally validated and computationally predicted *Sox2* enhancers. Moreover, it is worth considering that some of the 71 remaining highly conserved LPCs located within the *Sox2* insulators may contain novel enhancers for driving expression of *Sox2* in a context-dependent manner.

We also performed a separate but related analysis to identify *Sox2* enhancers using the bioinformatics software, EEL (Enhancer Element Locator). One of the key differences between EEL and DREiVe is that the former requires a set of transcription factor PFMs from JASPAR or TRANSFAC to locate enhancers sharing order-dependent binding sites between two orthologous species. DREiVe on the other hand, is a *de novo* method which does not rely at all on previous knowledge of binding sites, but rather locates enhancers sharing order-independent patterns across multiple species. Results from this analysis show that EEL only identified 6 of 25 (24%) of the *Sox2* enhancers, conserved in human and chick. Among those identified were the N2, N3, and N4 enhancers (Figure 6). This suggests that the order-independent nature of the methodology underlying DREiVe has greater sensitivity in identifying enhancers.

To predict regulation of *Sox2*, we subjected all 98 of the DREiVe-identified LPCs to transcription factor binding site analysis using ‘matrix-scan’ from the RSAT toolkit (Supplementary table 3). This identified several key conserved binding sites for important factors regulating early neural activity in the N1-N5 enhancers. In particular, putative sites for the HMG domain transcription factors Sox/LEF/TCF were found to be distributed among these early neural enhancers in both human and chick, together with sites for POU family proteins (Figure 7). These are again consistent with previous findings from the literature (Takemoto *et al.*, 2006; Uchikawa *et al.*, 2003), suggesting that the specific experimentally validated binding sites in the early neural enhancers of chick were accurately predicted by this computational approach.

Conclusions

Advances in computational biology and bioinformatics have made available a large number of public resources to facilitate *cis*- regulatory analysis suitable for the chick genome. This has in turn generated several different complementary techniques and methodologies for conducting computational analysis, each having its own set of strengths and weaknesses. In such a situation, the most effective approach is to select appropriate bioinformatics methodologies and to integrate them into a functional workflow to streamline the overall analysis. Here, we provide a new workflow integrating a novel tool for prediction of putative insulators (CTCF analysis), with a tool for enhancer prediction and TFBS analysis. We then test this approach by analysing the *Sox2* locus, which reveals a good correspondence between computationally predicted *cis*-regulatory sites and those that have been experimentally determined.

Acknowledgements

The development of these tools were made possible by an Advanced Investigator grant from the European Research Council (ERC) to CDS and project grants from BBSRC (to CDS and AS). We are grateful to Dr José-Luis Gómez-Skármeta (Seville, Spain) for useful advice on CTCF binding sites.

References

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS, editors. 2009. MEME SUITE: tools for motif discovery and searching. W202-208 p.

Barrett LW, Fletcher S, Wilton SD. 2012. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. Cell Mol Life Sci.

Brudno M, Poliakov A, Minovitsky S, Ratnere I, Dubchak I. 2007. Multiple whole genome alignments and novel biomedical applications at the VISTA portal. Nucleic Acids Res 35: W669-674.

Califano A. 2000. SPLASH: structural pattern localization analysis by sequential histograms. Bioinformatics 16: 341-357.

Carlson JM, Chakravarty A, DeZiel CE, Gross RH. 2007. SCOPE: a web server for practical de novo motif discovery. Nucleic Acids Res 35: W259-264.

Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T. 2005. MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics 21: 2933-2942.

Chan BY, Kibler D. 2005. Using hexamers to predict cis-regulatory motifs in Drosophila. Bmc Bioinformatics 6.

Collignon J, Sockanathan S, Hacker A, Cohen-Tannoudji M, Norris D, Rastan S, Stevanovic M, Goodfellow PN, Lovell-Badge R. 1996. A comparison of the properties of Sox-3 with Sry and two related genes, Sox-1 and Sox-2. Development 122: 509-520.

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.

Cullen KE, Kladde MP, Seyfred MA. 1993. Interaction between Transcription Regulatory Regions of Prolactin Chromatin. Science 261: 203-206.

Darnell DK, Stark MR, Schoenwolf GC. 1999. Timing and cell interactions underlying neural induction in the chick embryo. Development 126: 2505-2514.

Davuluri RV. 2003. Application of FirstEF to find promoters and first exons in the human genome. Curr Protoc Bioinformatics Chapter 4: Unit4 7.

Dean A. 2011. In the loop: long range chromatin interactions and gene regulation. Brief Funct Genomics 10: 3-10.

Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. Science 295: 1306-1311.

Down TA, Hubbard TJ. 2002. Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res 12: 458-461.

Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, Meyer LR, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Pohl A, Malladi VS, Li CH, Learned K, Kirkup V, Hsu F, Harte RA, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Diekhans M, Cline MS, Clawson H, Barber GP, Haussler D, James Kent W. 2012. The UCSC Genome Browser database: extensions and updates 2011. Nucleic Acids Res 40: D918-923.

Feeney AJ, Verma-Gaur J. 2012. CTCF-cohesin complex: architect of chromatin structure regulates V(D)J rearrangement. Cell Res 22: 280-282.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovcova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F,

Dunham I, Durbin R, Fernandez-Suarez XM, Harrow J, Herrero J, Hubbard TJ, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadissa A, Searle SM. 2012. Ensembl 2012. Nucleic Acids Res 40: D84-90.

Frith MC, Li MC, Weng Z. 2003. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. Nucleic Acids Res 31: 3666-3668.

Hannenhalli S. 2008. Eukaryotic transcription factor binding sites--modeling and integrative search methods. Bioinformatics 24: 1325-1331.

Hughes JD, Estep PW, Tavazoie S, Church GM. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol 296: 1205-1214.

Jin F, Li Y, Ren B, Natarajan R. 2011. Enhancers: multi-dimensional signal integrators. Transcription 2: 226-230.

Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. 2003. MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 31: 3576-3579.

Kim J, Chu J, Shen X, Wang J, Orkin SH. 2008. An extended transcriptional network for pluripotency of embryonic stem cells. Cell 132: 1049-1061.

Kim YJ, Cecchini KR, Kim TH. 2011. Conserved, developmentally regulated mechanism couples chromosomal looping and heterochromatin barrier activity at the homeobox gene A locus. Proc Natl Acad Sci U S A 108: 7391-7396.

Kornblihtt AR. 2012. CTCF: from insulators to alternative splicing regulation. Cell Res 22: 450-452.

Loots G, Ovcharenko I. 2007. ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes. Bioinformatics 23: 122-124.

Martin D, Pantoja C, Minan AF, Valdes-Quezada C, Molto E, Matesanz F, Bogdanovic O, de la Calle-Mustienes E, Dominguez O, Taher L, Furlan-Magaril M, Alcina A, Canon S, Fedetz M, Blasco MA, Pereira PS, Ovcharenko I, Recillas-Targa F, Montoliu L, Manzanares M, Guigo R, Serrano M, Casares F, Gomez-Skarmeta JL. 2011. Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes (vol 18, pg 708, 2011). Nature Structural & Molecular Biology 18: 1084-1084.

Maston GA, Evans SK, Green MR. 2006. Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet 7: 29-59.

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE,

Wingender E. 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34: D108-110.

Papanayotou C, Mey A, Birot AM, Saka Y, Boast S, Smith JC, Samarut J, Stern CD. 2008. A mechanism regulating the onset of Sox2 expression in the embryonic neural plate. PLoS Biol 6: e2.

Perier RC, Praz V, Junier T, Bonnard C, Bucher P. 2000. The eukaryotic promoter database (EPD). Nucleic Acids Res 28: 302-303.

Pevny LH, Sockanathan S, Placzek M, Lovell-Badge R. 1998. A role for SOX1 in neural determination. Development 125: 1967-1978.

Raab JR, Kamakaka RT. 2010. Insulators and promoters: closer than we think. Nat Rev Genet 11: 439-446.

Rex M, Orme A, Uwanogho D, Tointon K, Wigmore PM, Sharpe PT, Scotting PJ. 1997. Dynamic expression of chicken Sox2 and Sox3 genes in ectoderm induced to form neural tissue. Dev Dyn 209: 323-332.

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: D91-94.

Scherf M, Klingenhoff A, Werner T. 2000. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. J Mol Biol 297: 599-606.

Sosinsky A, Honig B, Mann RS, Califano A. 2007. Discovering transcriptional regulatory regions in Drosophila by a nonalignment method for phylogenetic footprinting. Proc Natl Acad Sci U S A 104: 6305-6310.

Streit A, Berliner AJ, Papanayotou C, Sirulnik A, Stern CD. 2000. Initiation of neural induction by FGF signalling before gastrulation. Nature 406: 74-78.

Streit A, Lee KJ, Woo I, Roberts C, Jessell TM, Stern CD. 1998. Chordin regulates primitive streak development and the stability of induced neural cells, but is not sufficient for neural induction in the chick embryo. Development 125: 507-519.

Streit A, Sockanathan S, Perez L, Rex M, Scotting PJ, Sharpe PT, Lovell-Badge R, Stern CD. 1997. Preventing the loss of competence for neural induction: HGF/SF, L5 and Sox-2. Development 124: 1191-1202.

Streit A, Tambalo M, Chen J, Grocott T, Anwar M, Sosinsky A, Stern CD. 2013. Building gene regulatory networks: the chick as a perfect model system. Genesis (This issue).

Takemoto T, Uchikawa M, Kamachi Y, Kondoh H. 2006. Convergence of Wnt and FGF signals in the genesis of posterior neural plate through activation of the Sox2 enhancer N-1. Development 133: 297-306.

Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J. 2011a. RSAT 2011: regulatory sequence analysis tools. Nucleic Acids Research 39: W86-W91.

Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J. 2011b. RSAT 2011: regulatory sequence analysis tools. Nucleic Acids Res 39: W86-91.

Uchikawa M, Ishida Y, Takemoto T, Kamachi Y, Kondoh H. 2003. Functional analysis of chicken Sox2 enhancers highlights an array of diverse regulatory elements that are conserved in mammals. Developmental Cell 4: 509-519.

Uchikawa M, Kamachi Y, Kondoh H. 1999. Two distinct subgroups of Group B Sox genes for transcriptional activators and repressors: their expression during embryonic organogenesis of the chicken. Mech Dev 84: 103-120.

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. Mechanisms of Development 121: 1145-1158.

Uwanogho D, Rex M, Cartwright EJ, Pearl G, Healy C, Scotting PJ, Sharpe PT. 1995. Embryonic expression of the chicken Sox2, Sox3 and Sox11 genes suggests an interactive role in neuronal development. Mech Dev 49: 23-36.

Van Loo 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 Biology 9.

Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res 35: D88-92.

Vogelmann J, Valeri A, Guillou E, Cuvier O, Nollmann M. 2011. Roles of chromatin insulator proteins in higher-order chromatin organization and transcription regulation. Nucleus 2: 358-369.

Wasserman WW, Sandelin A. 2004. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5: 276-287.

Wood HB, Episkopou V. 1999. Comparative expression of the mouse Sox1, Sox2 and Sox3 genes from pre-gastrulation to early somite stages. Mech Dev 86: 197-201.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJ, Cooke JE, Elgar G. 2005. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3: e7.

Yeowell K, Sosinsky A. 2012. Combination of alignment-based and alignment-free approaches for discovery of transcription regulatory regions in vertebrate genes demonstrates a high accuracy of prediction. submitted.

Software/ Tool	Description	Chick data analysis	Genome-wide	High throughput CTCF analysis	TFBS detection	Enhancer discovery
MEME (Bailey *et al.*, 2009)	Discovers conserved sequence motifs enriched in the users input sequences. Some of its programs include MEME (motif discovery), GLAM2 (motif discovery with gaps), and TOMTOM (motif-motif database searching).	YES, as long as chick sequences are provided as input.	NO- limited to number of input sequences.	NO	YES	NO
MATCHTM (Kel *et al.*, 2003)	Identifies TFBSs using an up-to-date library of TRANSFAC matrices. The algorithm uses a matrix similarity score (MSS) and a core similarity score (CSS) to assess the quality of a match between a TFBS and the users input sequence(s).	YES, as long as chick sequences are provided as input.	NO- limited to number of input sequences.	YES, as long as the CTCF matrix is provided.	YES	NO
MatInspector * (Carthariu s *et al.*, 2005)	Detects TFBSs using its own repository of TF matrices. This library of matrices consists of matrix families built with similar or functionally related TFBSs.	YES	NO	YES, as long as the CTCF matrix is provided.	YES	NO
ModuleMiner (Van Loo *et al.*, 2008)	Detects *cis*-regulatory motifs in co-expressed human genes. It uses a library of PFMs, and implements a whole-genome optimisation approach to look for specific signals in the input set that are not present in other genes.	NO	NO	NO	YES	YES
AlignACE (Hughes *et al.*, 2000)	Uses a Gibbs sampling technique to find patterns conserved in a set of DNA sequences.	YES	NO	NO	YES	NO
Cluster-Buster (Frith *et al.*, 2003)	Identifies *cis*-regulatory motifs by searching for regions of the sequence that resemble a statistical model of a motif cluster more than a background DNA model.	YES	NO	NO	YES	YES
SCOPE (Carlson *et al.*, 2007)	Conducts *de novo* identification of regulatory motifs in sets of co-regulated genes.	YES	NO	NO	YES	NO
matrix-scan (RSAT) (Thomas-Chollier *et al.*, 2011b)	Uses TF profiles from TRANSFAC or JASPAR to identify TFBSs for a set of given input sequences.	YES	NO	YES, as long as the CTCF matrix is provided.	YES	NO

DREiVe (Sosinsky *et al.*, 2012)	A method to identify putative regulatory regions by comparing orthologous genomic sequences. It integrates the well known SPLASH algorithm with a local permutation clustering (LPC) algorithm to discovery conserved motifs across multiple species.	YES	NO – limited to one gene at a time	YES	YES	YES

Table 1 – Complementary bioinformatics tools available for cis-regulatory analysis. Commercial products are highlighted with an asterisk.

Figure Legends

Figure 1. Regulation of Transcription – An overview of transcriptional regulatory elements, illustrating how distal regulatory elements can interact with the core promoter.

Figure 2. A proposed computational workflow for *cis*-regulatory analysis – The workflow can be divided into three principle stages; insulator analysis, enhancer prediction, and transcription factor binding site analysis.

Figure 3. Computational CTCF analysis – A) Representation of the CTCF matrix from the JASPAR database, B) the procedure used by our Perl script to calculate "weighted" scores of CTCF binding sites across the chromosome, C) probability distribution showing differences in frequencies of each weighted score between the empirical and null distributions, D) calculation of False Discovery Rates (FDR) and P values.

Figure 4. A comparison between computationally-identified and ChIP-Seq derived CTCF sites. All ChIP-seq datasets were downloaded from the UCSC genome browser (Dreszer et al., 2011).

Figure 5. A) An overview of the *Sox2* putative insulator region in Human. Red arrows highlight the constitutive CTCF sites found both up (*hg19* coordinates chr3:181,427,485-181,427,504) and downstream (*hg19* coordinates chr3:182,428,542-182,428,561) of *Sox2.* The green block shows DREiVe-identified LPCs which overlap with known *Sox2* functional enhancers in human. B) An overview of the equivalent syntenic region in Chick, with red arrows highlighting the CTCF sites and green blocks showing DREiVe-identified LPCs that overlap with known functional enhancers of *Sox2* in chick. Coordinates for CTCF site upstream of *Sox2*: *galGal3* chr9:18,000,253-18,000,272 and downstream of *Sox2*: *galGal3* chr9:17,684,565-17,684,584. C) Equivalent syntenic region in mouse. Coordinates for CTCF site upstream of *Sox2*: *mm9* chr3:34,546,807-34,546,826 and downstream of *Sox2*: *mm9* chr3:35,379,158-35,379,177. The UCSC genome browser was used to generate this representation.

Figure 6. Computational analysis of the *Sox2* locus. DREiVe-predicted enhancers are shown as red horizontal bars, EEL-predicted enhancers are shown as blue horizontal bars, and previously known enhancers of *Sox2* identified by Uchikawa et al., 2003 are shown as brown horizontal bars. Red rectangles display overlapping regions between computationally predicted and known enhancers.

Figure 7. Matrix Scan analysis of TFBSs found in the N1 and N2 enhancers of *Sox2.*