Methods xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Bioinformatics approaches to predict target genes from transcription factor binding data

Alexandra Essebier [*], Marnie Lamprecht, Michael Piper, Mikael Bodén [*]

The University of Queensland, Brisbane 4072, Australia

A R T I C L E I N F O

Article history:
Received 9 June 2017
Received in revised form 29 August 2017
Accepted 3 September 2017
Available online xxxx

Keywords:
Bioinformatics
Transcription factor
Enhancer
Promoter
Regulatory interactions
Cancer

A B S T R A C T

Transcription factors regulate gene expression and play an essential role in development by maintaining proliferative states, driving cellular differentiation and determining cell fate. Transcription factors are capable of regulating multiple genes over potentially long distances making target gene identification challenging.

Currently available experimental approaches to detect distal interactions have multiple weaknesses that have motivated the development of computational approaches. Although an improvement over experimental approaches, existing computational approaches are still limited in their application, with different weaknesses depending on the approach. Here, we review computational approaches with a focus on data dependency, cell type specificity and usability.

With the aim of identifying transcription factor target genes, we apply available approaches to typical transcription factor experimental datasets. We show that approaches are not always capable of annotating all transcription factor binding sites; binding sites should be treated disparately; and a combination of approaches can increase the biological relevance of the set of genes identified as targets.

Contents

[*] Corresponding authors.
E-mail addresses: a.essebier@uq.edu.au (A. Essebier), m.boden@uq.edu.au (M. Bodén).

http://dx.doi.org/10.1016/j.ymeth.2017.09.001
1046-2023/© 2017 Elsevier Inc. All rights reserved.

1. Introduction

Transcription factors are a set of DNA binding proteins that play a key role in regulation by providing delicate control over the expression of genes [1]. Transcription factors can also act as oncogenes or tumour suppressors, leading to uncontrolled growth and avoidance of apoptosis; they are therefore key targets for the diagnosis and treatment of cancer [2,3]. Understanding the role of a transcription factor in gene regulation requires knowledge of the transcription factor's binding sites and most importantly, its target genes.

Annotating a binding event to a target gene presents three key challenges:

- a single binding event can control multiple genes;
- a single gene can be coordinately controlled by multiple binding events; and
- binding sites can be involved in distal interactions facilitated by the formation of DNA loops.

Numerous approaches exist to annotate transcription factor binding sites to their target genes, with different approaches aiming to address neither, some or all of the key challenges. Here, we explore the factors that influence the identification of an accurate set of target genes, review annotation approaches that have previously been shown to reproduce linkages captured by chromatin conformation assays, and finally, present two case studies to demonstrate that treating binding events disparately and applying a variety of annotation approaches can increase the number of identified, biologically relevant target genes.

1.1. Binding context

Transcription factors do not act independently, instead relying on numerous genetic and epigenetic features to bind DNA and regulate gene expression. Therefore, when attempting to identify the genes targeted by a transcription factor, it is essential to use multiple data sources to determine not only where binding occurs but the binding context.

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is regularly used to identify binding sites of transcription factors [4–7]. ChIP-seq shows the hundreds to tens of thousands of specific locations at which a transcription factor binds including distal, proximal, intronic and intergenic regions. Binding preference across these regions depends on the transcription factor, cell type and condition, and genetic and epigenetic features [8].

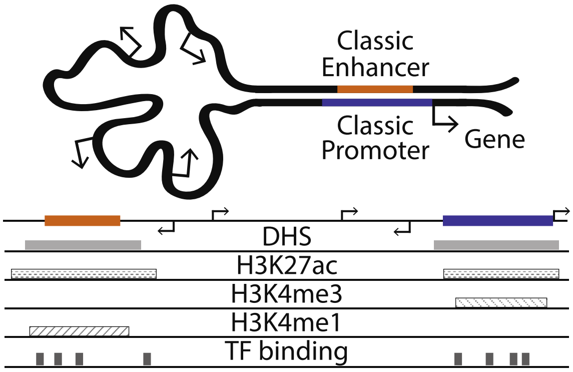

Fig. 1. Graphical and linear representations of a hypothetical long distance interaction. Through formation of a DNA loop, enhancer and promoter regions can be brought into contact over long distances. This interaction allows CRMs to coordinately regulate transcription. A DNA loop can contain genes whose expression are not altered by the paired CRMs. Active enhancers and promoters are historically characterised by H3K27ac and DNase I hypersensitivity (accessible chromatin) with enrichment of transcription factor binding. Promoters are generally marked with H3K4me3, while enhancers are marked with H3K4me1.

Key epigenetic features include histone modifications and chromatin accessibility. ChIP-seq is again used to detect different post-translational modifications to the histones around which the DNA is wrapped; a collection of ChIP-seq experiments can provide significant insights into regulation beyond a single transcription factor. Both DNase I hypersensitivity (DHS) and assay for transposase accessible chromatin (ATAC) sequencing identify regions of accessible chromatin: a feature generally required for transcription factor binding to occur. (So-called pioneer factors are a notable exception to this [9,10].)

Together, these assays have been used to identify and classify cis regulatory modules (CRMs): locations on the genome characterised by clustered transcription factor binding sites, distinct patterns of histone modifications and chromatin accessibility, specific sequence features, and evolutionary conservation [4]. *Enhancers* and *promoters* are both types of CRMs and have historically been viewed as distinct elements with unique roles in regulation [11,12]. Epigenetic patterns associated with enhancers and promoters are visualised in Fig. 1. They both drive target gene expression; promoters are recognised by general transcription factors and recruit RNA Pol II at sites proximal to known genes; enhancers are bound to by transcription factors at sites potentially (but not necessarily) distal to their target genes, where they can substantially modulate the transcriptional efficiency.

A transcription factor binding site outside a region with relevant binding context (see Fig. 1) is likely to be a false positive event, which are common in ChIP-seq analyses. Therefore, CRMs can be used to filter transcription factor binding sites. This presents two alternate ways to link a transcription factor binding site to a target gene: using the binding site alone or using the CRM that the transcription factor is bound to. A number of approaches exist for each of these concepts.

1.2. DNA loops

As shown in Fig. 1, DNA loops allow interaction between CRMs over distances of up to 1 Mb to facilitate the formation of regulatory clusters [13,14]. This secondary structure is dynamic, with a *many:many* relationship between genes and gene-distal CRMs. Enhancers and promoters can be brought into close proximity creating *regulatory clusters* that allow different CRMs to coordinately influence the transcriptional output of multiple genes [11,15]. Through this mechanism, a CRM is able to regulate multiple genes, and a gene can be regulated by multiple CRMs. These are two of the key challenges when annotating CRMs and transcription factors to target genes.

Linking CRMs to their targets over long distances is the third major challenge when annotating binding events. In the past, researchers have approached the problem by assigning a CRM to the nearest gene. Studies have shown that as few as 7% of DNA loops link a CRM to the nearest gene; it is common to observe a number of genes in the DNA loops whose expression is not influenced [16]. By proxy, transcription factors bound to CRMs involved in long distance interactions are often incorrectly annotated to target genes [17]. The interaction between two CRMs through DNA loops is referred to in the literature by a variety of names including enhancer-promoter pairs, enhancer-gene pairs, enhancer-target gene pairs and more. To move away from the enhancer/promoter dichotomy, we refer to such interactions as *regulatory partners*.

1.3. Experimental annotation approaches

Chromosome conformation capture techniques provide significant clues for identifying regulatory partners. There are a number