Promoter-proximal CCCTC-factor binding is associated with an increase in the transcriptional pausing index

Sur Herrera Paredes¹,², Michael F. Melgar³ and Praveen Sethupathy²,⁴,⁵,*

¹Curriculum in Bioinformatics and Computational Biology, ²Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA, ³Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA, ⁴Department of Genetics and ⁵Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Received on May 9, 2012; revised on September 14, 2012; accepted on September 30, 2012

1 INTRODUCTION

Promoter-proximal pausing of RNA polymerase II (RNAPII) has emerged as a widespread and important step in the regulation of transcription (Core et al., 2008; Gilchrist and Adelman, 2012; Li and Gilmour, 2011; Seila et al., 2009). The role of pausing is not completely understood, but it is thought that pausing provides a means of regulating transcriptional elongation. Recent advances in high-throughput genomic technologies facilitated the discovery that RNAPII pausing at promoters is a widespread physiologically regulated phenomenon. The molecular underpinnings of pausing are incompletely understood.

Results: In this study, we analyzed CTCF binding sites and nascent transcriptomic data from three different cell types, and found that promoter-proximal CTCF binding is significantly associated with RNAPII pausing.

Contact: praveen_sethupathy@med.unc.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

2 METHODS

2.1 Datasets

Empirically determined CTCF-binding sites were downloaded from http://licr-renlab.ucsd.edu/download.html (CTCF chromatin immunoprecipitation followed by microarray analysis [ChiP-chip] in IMR90 cells), http://genome.ucsc.edu/ENCODE/downloads.html (CTCF chromatin immunoprecipitation followed by high-throughput sequencing [ChiP-seq] in MCF-7 cells) and Chen et al., 2008 (mES ChiP-seq). All coordinates were mapped to either the mm9 (mES) or the hg18 (IMR90, MCF-7) genome builds using the command line liftOver program with an identity threshold of 20–25%.

Analysis

Promoter-proximal pausing indices were calculated in the following manner: first, for each annotated RefSeq transcript longer than 3 kb, we set the transcription start site (TSS) as position 0, and defined 0–500 nt downstream as the promoter-proximal region and 500–gene end as the gene body. Second, we computed the density of RNAPII in both promoter-proximal regions and gene bodies by calculating Gro-seq reads/nt/mappability using the same mappability data as in Melgar et al., 2011. Third, for all genes that had at least 20–25% of CTCF-binding events are proximal to promoters (Jothi et al., 2008; Kim et al., 2007), we hypothesized that CTCF binding may be a cofactor in mediating promoter-proximal RNAPII pausing. To investigate this possibility, we analyzed genome-wide nascent RNA data (global nuclear run-on followed by sequencing; Gro-seq) and empirically determined CTCF-binding sites from the only three cell types for which both datasets are publicly available: human cell lines IMR90 (lung fibroblast) and MCF-7 (breast cancer) and mouse embryonic stem cells (mES).

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com
least one read mapping to the gene body (n = 24,988 for IMR90, n = 24,613 for MCF-7, n = 20,512 for mES), we computed pausing index as the ratio of RNAPII density in the promoter-proximal region to RNAPII density in the gene body, similar to previous studies (Core et al., 2008; Min et al., 2011).

Promoters were divided into two categories: those with and those without a CTCF-binding site overlapping X to Y nt downstream of the TSS, where X and Y were set in independent experiments to 0 and 100, 101 and 200, 201 and 300, 301 and 500, 501 and 1000 or 1001 and 2000. To compare pausing indices between categories, we computed the Jaccard index as follows: \( \frac{(x - y)}{(x + y)} \), where \( x \) is the median pausing index at CTCF-bound promoters and \( y \) is the median pausing index at promoters where CTCF is not bound. The significance of the difference in pausing indices between the two promoter categories was assessed by one-sided permutation and non-parametric Mann–Whitney–Wilcoxon (MWW) tests. This analysis was repeated for the same distances upstream of the TSS.

3 RESULTS

For each cell type, we divided promoters according to whether CTCF was bound upstream/downstream of the TSS (Section 2). The number of promoter-proximal CTCF-binding sites was highly variable across cell types (Supplementary Material 1).

GRO-seq data provide a density map of transcriptionally engaged RNAPII across the genome by purifying, sequencing and mapping nascent RNAs (Core et al., 2008). Using recently published GRO-seq data, we calculated RNAPII pausing indices at RefSeq-annotated promoters in three different cell types, and assessed whether promoter-proximal CTCF binding is associated with elevated RNAPII pausing indices (Section 2). We made two novel observations:

1. Genes that exhibit promoter-proximal CTCF binding (defined as within 100 nt downstream of the TSS) have a significantly greater RNAPII pausing index than genes that do not (Supplementary Material 1). For example, in IMR90 cells, the median pausing index at promoters with proximal CTCF-binding sites is \( \sim 2 \)-fold greater (permutation test \( P < 0.0001 \), Mann–Whitney–Wilcoxon test \( P = 1.8 \times 10^{-38} \) ) than at promoters without proximal CTCF-binding sites (Fig. 1). The highly overlapping range of pausing indices between the two categories of promoters indicates that while CTCF binding may be associated with pausing, it is by no means necessary or sufficient.

Regions of CTCF binding are enriched for motifs with elevated cytosine (C) content (Jothi et al., 2008). As such, the observed association between CTCF-binding events and increased pausing index could be explained by nucleotide composition (i.e. cytosine enriched regions may be responsible for the association with pausing, independent of CTCF binding). To test this, we repeated the analysis in each cell type with computationally predicted CTCF-binding sites (Section 2), based on the known C-rich motif (Kim et al., 2007). We did not observe any association with pausing (Fig. 1; Supplementary Material 1), which strongly suggests that sequence composition alone is not sufficient to explain the association signal, but rather, the physical binding of CTCF is required.

2. The association between CTCF binding and the promoter-proximal pausing index is position-dependent—that is, the further upstream/downstream from the TSS CTCF is bound, the less an association there is with RNAPII pausing (Fig. 2;
Supplementary Material 1). For example, in mES cells, CTCF binding within 101–200 nt downstream of TSSs is highly significantly associated with pausing (permutation test $P < 0.0001$); however, this association is dramatically weakened for CTCF binding within 1001–2000 nt upstream/downstream of TSSs (permutation test $P > 0.5$).

4 DISCUSSION

A recent study identified a novel role for CTCF in polymerase pausing at splice sites to facilitate exon inclusion (Shukla et al., 2011). Our findings provide the first indication that CTCF is associated with promoter-proximal pausing. It is important to note that the study does not demonstrate that CTCF binding is necessary or sufficient for promoter-proximal pausing. Substantial additional experimental work is required to uncover the molecular underpinnings of the observed association between CTCF binding and RNAPII dynamics.

The greatest degree of association with pausing is observed for CTCF-binding sites that are immediately downstream of the TSS. Therefore, an appealing mechanism for the observed association is steric hindrance—that is, stably bound CTCF may serve to hinder RNAPII processivity. If this is the case, it is likely that many other transcription factors with binding sites immediately downstream of TSSs also associate with RNAPII pausing. It will be possible to test this hypothesis as more molecular underpinnings of the observed association between CTCF binding and RNAPII dynamics.

ACKNOWLEDGEMENTS

The authors thank Dr Karen Adelman (National Institute of Environmental Health and Safety, National Institutes of Health) for valuable discussions, and Dr Michael L. Stitzel (National Human Genome Research Institute, National Institutes of Health) and members of the Sethupathy laboratory for helpful suggestions.

Funding: This work was supported in part by a National Institutes of Health R00 award (DK091318-02 to P.S).

Conflict of Interest: none declared.

REFERENCES


