

Correction of transposase sequence bias in ATAC-seq data with rule ensemble modeling

Jacob B. Wolpe¹, André L. Martins^{2,3} and Michael J. Guertin^{1,2,3,*}

¹Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA, USA, ²Center for Cell Analysis and Modeling, University of Connecticut, Farmington, CT, USA and ³Department of Genetics and Genome Sciences, University of Connecticut, Farmington, CT, USA

Received November 30, 2022; Revised April 02, 2023; Editorial Decision May 15, 2023; Accepted May 19, 2023

ABSTRACT

Chromatin accessibility assays have revolutionized the field of transcription regulation by providing single-nucleotide resolution measurements of regulatory features such as promoters and transcription factor binding sites. ATAC-seq directly measures how well the Tn5 transposase accesses chromatinized DNA. Tn5 has a complex sequence bias that is not effectively scaled with traditional bias-correction methods. We model this complex bias using a rule ensemble machine learning approach that integrates information from many input k-mers proximal to the ATAC sequence reads. We effectively characterize and correct single-nucleotide sequence biases and regional sequence biases of the Tn5 enzyme. Correction of enzymatic sequence bias is an important step in interpreting chromatin accessibility assays that aim to infer transcription factor binding and regulatory activity of elements in the genome.

INTRODUCTION

Chromatin accessibility assays measure the relative frequency that exogenous enzymes access DNA. Chromatin accessibility is not a direct measure of molecular features of chromatin such as transcription factor occupancy or histone modification status. However, accessibility is considered a proxy measurement of regulatory element activity irrespective of the constellation of factors bound to the DNA (1,2). Accessible regions are enriched for transcription factor binding and histone modifications that are hallmarks of functional *cis*-regulatory elements (3–7). Deconvolution of genome-wide chromatin hypersensitivity data to infer individual transcription factor binding events remains a challenge (8).

ATAC-seq revolutionized the chromatin accessibility field by providing a straightforward method that requires fewer than 5000 cells (9). ATAC-seq leverages a hyperactive Tn5 transposase that directly inserts high throughput

sequencing adapters into accessible DNA to create sequencing libraries. The analysis of ATAC-seq data requires additional considerations beyond traditional hypersensitivity assays because the molecular biology of transposase function is distinct compared to DNase and other enzymes that are used to measure chromatin accessibility (10,11).

Although using Tn5 to measure chromatin accessibility is experimentally easier than using DNase, the dimeric Tn5 enzyme recognizes a wider region when interacting with DNA and this is reflected in the sequence bias of Tn5 (12). Characterizing and correcting enzymatic sequence bias is an essential step for accurate interpretation of chromatin accessibility data (13–15). Strategies to correct biases use a variety of sequence inputs to build models, including k-mer instances, position weight matrices, and long stretches of DNA. DNase bias can be directly scaled based on the 6-mer sequence centered on the DNase cleavage site (16–19). The advantage of direct k-mer scaling is the simplicity: reads are scaled by the expected/observed k-mer count ratio. If a k-mer is found less often than expected by chance, then the read is scaled to a higher value. However, direct k-mer scaling does not effectively correct Tn5 biases from ATAC-seq data (16,18,20), so other methods were developed to correct Tn5 bias. ATACCorrect scales individual ATAC-seq reads based upon a dinucleotide weight matrix representing Tn5 bias and introduces simulated reads to offset observed biases (21). Similar to weight matrix scaling, HINT-ATAC employs position dependency models, which account for interactions between weight matrix positions to model ATAC-seq bias (12). SELMA encodes k-mer data into a simplex vector, which is incorporated into a Hadamard matrix for all mono and dinucleotide combinations. These data are input into a linear regression model to capture and correct both DNase and Tn5 bias (22). More sophisticated methods that correct ATAC-seq data are often less interpretable than k-mer and weight matrix scaling. Seqbias trains a Bayesian network to encode nucleotide preferences using a 40 base window centered on the Tn5 recognition site (23). MsCentipede uses naked DNA cleavage to train a Bayesian multi-scale model of a Poisson distribution of reads to account for sequence bias (24). Another method

*To whom correspondence should be addressed. Tel: +1 860 679 1461; Email: guertin@uchc.edu

Research

Kinetic networks identify TWIST2 as a key regulatory node in adipogenesis

Arun B. Dutta,¹ Daniel S. Lank,² Roza K. Przanowska,³ Piotr Przanowski,³ Lixin Wang,³ Bao Nguyen,¹ Ninad M. Walavalkar,¹ Fabiana M. Duarte,⁴ and Michael J. Guertin^{5,6}

¹Department of Biochemistry and Molecular Genetics, ²Department of Pharmacology, ³Department of Biomedical Engineering, University of Virginia, Charlottesville, Virginia 22903, USA; ⁴Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA; ⁵Center for Cell Analysis and Modeling, University of Connecticut, Farmington, Connecticut 06030, USA; ⁶Department of Genetics and Genome Sciences, University of Connecticut, Farmington, Connecticut 06030, USA

Adipocytes contribute to metabolic disorders such as obesity, diabetes, and atherosclerosis. Prior characterizations of the transcriptional network driving adipogenesis have overlooked transiently acting transcription factors (TFs), genes, and regulatory elements that are essential for proper differentiation. Moreover, traditional gene regulatory networks provide neither mechanistic details about individual regulatory element–gene relationships nor temporal information needed to define a regulatory hierarchy that prioritizes key regulatory factors. To address these shortcomings, we integrate kinetic chromatin accessibility (ATAC-seq) and nascent transcription (PRO-seq) data to generate temporally resolved networks that describe TF binding events and resultant effects on target gene expression. Our data indicate which TF families cooperate with and antagonize each other to regulate adipogenesis. Compartment modeling of RNA polymerase density quantifies how individual TFs mechanistically contribute to distinct steps in transcription. The glucocorticoid receptor activates transcription by inducing RNA polymerase pause release, whereas SP and AP-1 factors affect RNA polymerase initiation. We identify *Twist2* as a previously unappreciated effector of adipocyte differentiation. We find that TWIST2 acts as a negative regulator of 3T3-L1 and primary preadipocyte differentiation. We confirm that *Twist2* knockout mice have compromised lipid storage within subcutaneous and brown adipose tissue. Previous phenotyping of *Twist2* knockout mice and Setleis syndrome *Twist2*^{−/−} patients noted deficiencies in subcutaneous adipose tissue. This network inference framework is a powerful and general approach for interpreting complex biological phenomena and can be applied to a wide range of cellular processes.

[Supplemental material is available for this article.]

Mature adipocytes contribute to a multitude of metabolic processes by regulating energy balance, producing hormones, and providing structural and mechanical support (Rosen and Spiegelman 2006). Adipocyte hyperplasia downstream from increased adipogenesis is associated with pathogenesis of obesity, type 2 diabetes, and cardiovascular disease (van Kruijsdijk et al. 2009; Unamuno et al. 2018). Adipogenic factors represent opportunities for intervention and possible mitigation of obesity-related sequelae (Ghaben and Scherer 2019; Ahmad et al. 2020). Adipocyte maturation is a tightly regulated process involving many chromatin and transcriptional changes downstream from transcription factor (TF) binding (Siersbæk et al. 2011; Tsankov et al. 2015; Thompson et al. 2016; Rauch et al. 2019; Madsen et al. 2020). Although prior studies have extensively characterized the TFs and gene expression changes required for adipogenesis (Rosen and Spiegelman 2006; Lefterova and Lazar 2009; Siersbæk et al. 2012), this work relied on measurements taken hours or days apart on cells undergoing adipogenesis. Molecular events, such as TF binding, chromatin remodeling, and redistribution of RNA polymerase, occur on a time scale of seconds to minutes (McNally et al. 2000; Chen et al. 2014; Duarte et al. 2016). Therefore, previous examinations of adipogenic signaling likely omitted multiple

waves of signaling and potential regulatory factors that may be critical to the process.

Molecular genomics assays can query transcriptional events with extremely high temporal resolution. Although each assay delivers a tremendous amount of information, each is limited in the biology that it measures. ChIP-seq directly quantifies chromatin occupancy of proteins, but the assay is dependent upon the availability of antibodies and is limited to a single factor at a time. ATAC-seq and DNase-seq assays quantify chromatin accessibility, which is an indirect measure of regulatory element (RE) activity (Boyle et al. 2008; Buenrostro et al. 2015). Combining accessibility data with TF motif analyses can accurately infer TF binding without the need for factor- and species-specific antibodies (Wu et al. 1979; Vierstra et al. 2020). Kinetic experiments can further increase the sensitivity of inferring dynamic TF binding, because changes in TF binding modulate local chromatin structure and accessibility (Guertin and Lis 2010, 2013; Siersbæk et al. 2011, 2014). However, these assays do not directly inform on changes in transcription and RNA polymerase dynamics. Although RNA-seq is a popular approach for measuring transcription, the assay relies on accumulation of mature RNA species over hours, making it inappropriate for rapid measurements. In addition, it is difficult to deconvolve mechanistic insights from RNA-seq data, which measure secondary and compensatory transcription as well as long-

Corresponding author: guertin@uchc.edu

Article published online before print. Article, supplemental material, and publication date are at <https://www.genome.org/cgi/doi/10.1101/gr.277559.122>. Freely available online through the *Genome Research* Open Access option.

© 2023 Dutta et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

ANKLE1 cleaves mitochondrial DNA and contributes to cancer risk by promoting apoptosis resistance and metabolic dysregulation

Piotr Przanowski ^{1,2✉}, Róża K. Przanowska ^{1,2} & Michael J. Guertin ^{3,4✉}

Alleles within the chr19p13.1 locus are associated with increased risk of both ovarian and breast cancer and increased expression of the *ANKLE1* gene. *ANKLE1* is molecularly characterized as an endonuclease that efficiently cuts branched DNA and shuttles between the nucleus and cytoplasm. However, the role of *ANKLE1* in mammalian development and homeostasis remains unknown. In normal development *ANKLE1* expression is limited to the erythroblast lineage and we found that *ANKLE1*'s role is to cleave the mitochondrial genome during erythropoiesis. We show that ectopic expression of *ANKLE1* in breast epithelial-derived cells leads to genome instability and mitochondrial DNA (mtDNA) cleavage. mtDNA degradation then leads to mitophagy and causes a shift from oxidative phosphorylation to glycolysis (Warburg effect). Moreover, mtDNA degradation activates STAT1 and expression of epithelial-mesenchymal transition (EMT) genes. Reduction in mitochondrial content contributes to apoptosis resistance, which may allow precancerous cells to avoid apoptotic checkpoints and proliferate. These findings provide evidence that *ANKLE1* is the causal cancer susceptibility gene in the chr19p13.1 locus and describe mechanisms by which higher *ANKLE1* expression promotes cancer risk.

¹ Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, VA, USA. ² Department of Biomedical Engineering, University of Virginia School of Medicine, Charlottesville, VA, USA. ³ Center for Cell Analysis and Modeling, University of Connecticut, Farmington, CT, USA.

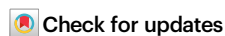
⁴ Department of Genetics and Genome Sciences, University of Connecticut, Farmington, CT, USA. ✉email: piotr.przanowski@gmail.com; guertin@uchc.edu

Intrinsic bias estimation for improved analysis of bulk and single-cell chromatin accessibility profiles using SELMA

Received: 1 September 2021

Accepted: 8 September 2022

Published online: 21 September 2022

Shengen Shawn Hu¹, Lin Liu², Qi Li³, Wenjing Ma^{1,4}, Michael J. Guertin⁵, Clifford A. Meyer^{6,7}, Ke Deng³, Tingting Zhang⁸ & Chongzhi Zang^{1,9,10} 

Genome-wide profiling of chromatin accessibility by DNase-seq or ATAC-seq has been widely used to identify regulatory DNA elements and transcription factor binding sites. However, enzymatic DNA cleavage exhibits intrinsic sequence biases that confound chromatin accessibility profiling data analysis. Existing computational tools are limited in their ability to account for such intrinsic biases and not designed for analyzing single-cell data. Here, we present Simplex Encoded Linear Model for Accessible Chromatin (SELMA), a computational method for systematic estimation of intrinsic cleavage biases from genomic chromatin accessibility profiling data. We demonstrate that SELMA yields accurate and robust bias estimation from both bulk and single-cell DNase-seq and ATAC-seq data. SELMA can utilize internal mitochondrial DNA data to improve bias estimation. We show that transcription factor binding inference from DNase footprints can be improved by incorporating estimated biases using SELMA. Furthermore, we show strong effects of intrinsic biases in single-cell ATAC-seq data, and develop the first single-cell ATAC-seq intrinsic bias correction model to improve cell clustering. SELMA can enhance the performance of existing bioinformatics tools and improve the analysis of both bulk and single-cell chromatin accessibility sequencing data.

Cis-regulatory elements in the genome play a critical role in transcription regulation by interaction with protein molecules such as transcription factors (TFs). These DNA elements are usually unwrapped from packed nucleosomes and are accessible in the chromatin structure^{1,2}. Genome-wide profiles of chromatin accessibility are a means to measure the global landscapes of active regulatory elements in different cell types. DNaseI hypersensitivity sequencing (DNase-seq) and the assay for transposase-accessible chromatin using sequencing

(ATAC-seq) have become widely used for the genomic profiling of chromatin structure and accessibility^{3,4}. Signal enrichments, or “peaks”, from DNase-seq or ATAC-seq data are considered to represent accessible chromatin regions and can be used for inferring regulatory elements or TF binding sites. In addition, DNase-seq and ATAC-seq data also exhibit footprint patterns, which are fine structures in the accessible chromatin where high-affinity protein-DNA interactions protect the DNA from DNaseI or Tn5-transposase cleavages^{4,5}.

¹Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA. ²Institute of Natural Sciences, MOE-LSC, School of Mathematical Sciences, CMA-Shanghai, and SJTU-Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University, Shanghai 200240, China. ³Center for Statistical Science and Department of Industrial Engineering, Tsinghua University, Beijing 100084, China. ⁴Department of Computer Science, Emory University, Atlanta, GA 30322, USA. ⁵Center for Cell Analysis and Modeling, Department of Genetics and Genome Sciences, University of Connecticut, Farmington, CT 06030, USA. ⁶Department of Data Science, Dana-Farber Cancer Institute, Boston, MA 02215, USA. ⁷Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA. ⁸Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, USA. ⁹Department of Public Health Sciences, University of Virginia, Charlottesville, VA 22908, USA. ¹⁰UVA Comprehensive Cancer Center, University of Virginia, Charlottesville, VA 22908, USA.

✉ e-mail: zang@virginia.edu

Article

Distinct *MUNC* lncRNA structural domains regulate transcription of different promyogenic factors

Roza K. Przanowska,^{1,2} Chase A. Weidmann,^{3,4} Shekhar Saha,¹ Magdalena A. Cichewicz,¹ Kate N. Jensen,¹ Piotr Przanowski,^{1,2} Patrick S. Irving,³ Kevin A. Janes,^{1,2} Michael J. Guertin,^{1,5} Kevin M. Weeks,³ and Anindya Dutta^{1,6,7,*}

¹Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, VA 22908, USA

²Department of Biomedical Engineering, University of Virginia School of Engineering, Charlottesville, VA 22908, USA

³Department of Chemistry, University of North Carolina, Chapel Hill, NC 27599, USA

⁴Department of Biological Chemistry and Center for RNA Biomedicine, University of Michigan Medical School, Ann Arbor, MI 48103, USA

⁵R. D. Berlin Center for Cell Analysis and Modeling, University of Connecticut, Farmington, CT 06030, USA

⁶Department of Genetics, University of Alabama, Birmingham, AL 35233, USA

⁷Lead contact

*Correspondence: duttaa@uab.edu

<https://doi.org/10.1016/j.celrep.2022.110361>

SUMMARY

Many lncRNAs have been discovered using transcriptomic data; however, it is unclear what fraction of lncRNAs is functional and what structural properties affect their phenotype. *MUNC* lncRNA (also known as ^{DRR}eRNA) acts as an enhancer RNA for the *Myod1* gene in *cis* and stimulates the expression of other promyogenic genes in *trans* by recruiting the cohesin complex. Here, experimental probing of the RNA structure revealed that *MUNC* contains multiple structural domains not detected by prediction algorithms in the absence of experimental information. We show that these specific and structurally distinct domains are required for induction of promyogenic genes, for binding genomic sites and gene expression regulation, and for binding the cohesin complex. *Myod1* induction and cohesin interaction comprise only a subset of *MUNC* phenotype. Our study reveals unexpectedly complex, structure-driven functions for the *MUNC* lncRNA and emphasizes the importance of experimentally determined structures for understanding structure-function relationships in lncRNAs.

INTRODUCTION

In 2004, the International Human Genome Sequencing Consortium found that roughly 70% of the nucleotides in the human genome are transcribed but not translated, thus comprising non-coding RNAs (ncRNAs) (International Human Genome Sequencing Consortium, 2004). Since then, many ncRNAs have been discovered and characterized, and a few functionally characterized, but even fewer have experimentally determined structural information. The majority of long non-coding RNAs (lncRNAs), defined as transcripts longer than 200 nucleotides, share signatures of coding RNAs: they are transcribed by RNA polymerase II, spliced, capped at the 5' end with 7-methylguanosine, and polyadenylated at the 3' end (Derrien et al., 2012). lncRNAs are not transient intermediaries to functional proteins, but functional macromolecules that drive cellular programs. It is thus important to understand how intrinsic higher-order lncRNA structures enable their function.

Most published studies have predicted the secondary structures of lncRNAs computationally by leveraging thermodynamic properties of base pairing, base stacking, and other atomic interactions. Although modeling can be accurate for short RNA se-

quences, accuracy drops significantly as the length of the transcript increases (Miao et al., 2015; Deigan et al., 2009). Data from probing experiments can be used to restrain folding algorithms to yield more accurate structural models, especially for long RNAs (Li et al., 2020a; Deigan et al., 2009; Mustoe et al., 2019). SHAPE-MaP (2'-hydroxyl acylation analyzed by primer extension coupled with mutational profiling) chemical probing strategies have proven especially useful, as nearly every nucleotide is probed in a single experiment, and RNAs of nearly any length can be studied (Merino et al., 2005; Smola et al., 2015a, 2015b; Busan et al., 2019). Chemical probing studies have revealed that individual RNAs have distinct "structural personalities" (Weeks, 2021), a feature that likely applies to lncRNAs and to individual domains within large lncRNAs. High-resolution methods, especially of ribosome (Watson et al., 2020) and viral (Jaafar and Kieft, 2019) systems have revealed enormous complexity in RNA structure and that large RNAs tend to form smaller domains.

The lack of structure-function studies in the lncRNA field is a major limitation for determining the molecular mechanisms by which lncRNAs exert their functions. One of the best-characterized lncRNAs is *Xist* lncRNA, which is a master regulator of X



SOFTWARE

Open Access



PEPPRO: quality control and processing of nascent RNA profiling data

Jason P. Smith^{1,2}, Arun B. Dutta², Kizhakke Mattada Sathyan², Michael J. Guertin^{1,2*} and Nathan C. Sheffield^{1,2,3,4*} 

*Correspondence:

guertin@uchc.edu;

nsheffield@virginia.edu

¹Center for Public Health Genomics, University of Virginia, Charlottesville, USA

²Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, USA

Full list of author information is available at the end of the article

Abstract

Nascent RNA profiling is growing in popularity; however, there is no standard analysis pipeline to uniformly process the data and assess quality. Here, we introduce PEPPRO, a comprehensive, scalable workflow for GRO-seq, PRO-seq, and ChRO-seq data. PEPPRO produces uniformly processed output files for downstream analysis and assesses adapter abundance, RNA integrity, library complexity, nascent RNA purity, and run-on efficiency. PEPPRO is restartable and fault-tolerant, records copious logs, and provides a web-based project report. PEPPRO can be run locally or using a cluster, providing a portable first step for genomic nascent RNA analysis.

Background

Steady-state transcription levels are commonly measured by RNA-seq, but there are many advantages to quantifying *nascent* RNA transcripts: First, it measures the transcription process directly, whereas steady-state mRNA levels reflect the balance of mRNA accumulation and turnover. Second, nascent RNA profiling measures not only RNA polymerase occupancy, but also orientation by default, whereas traditional RNA-seq requires specific library preparation steps to capture orientation. Third, nascent RNA profiling measures unstable transcripts, which can be used to infer regulatory element activity and identify promoters and enhancers *de novo* by detecting bidirectional transcription and clustered transcription start sites (TSSs) [1, 2]. Fourth, nascent RNA profiling can be used to determine pausing and RNA polymerase accumulation within any genomic feature. These advantages have led to growing adoption of global run-on (GRO-seq), precision run-on (PRO-seq), and, most recently, chromatin run-on (ChRO-seq) experiments [3–5]. With increasing data production, we require analysis pipelines for these data types. While tools are available for downstream analysis, such as to identify novel transcriptional units and bidirectionally transcribed regulatory elements [1, 6–10], there is no comprehensive, unified approach to initial sample processing and quality control.



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

The following cover pages were presented in my senior appointments package as Associate Professor two years ago:



Comparative interactomes of HSF1 in stress and disease reveal a role for CTCF in HSF1-mediated gene regulation

Received for publication, August 3, 2020, and in revised form, November 10, 2020 Published, Papers in Press, November 18, 2020,
<https://doi.org/10.1074/jbc.RA120.015452>

Eileen T. Burchfiel^{1,2}, Anniina Vihervaara³, Michael J. Guertin⁴, Rocio Gomez-Pastor², and Dennis J. Thiele^{1,2,5,*}

From the Departments of ¹Biochemistry and ²Pharmacology and Cancer Biology, Duke University School of Medicine, Durham, North Carolina, USA; ³Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, USA; ⁴Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia, USA; and ⁵Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, North Carolina, USA

Edited by Ursula Jakob

Heat shock transcription factor 1 (HSF1) orchestrates cellular stress protection by activating or repressing gene transcription in response to protein misfolding, oncogenic cell proliferation, and other environmental stresses. HSF1 is tightly regulated via intramolecular repressive interactions, post-translational modifications, and protein-protein interactions. How these HSF1 regulatory protein interactions are altered in response to acute and chronic stress is largely unknown. To elucidate the profile of HSF1 protein interactions under normal growth and chronic and acutely stressful conditions, quantitative proteomics studies identified interacting proteins in the response to heat shock or in the presence of a poly-glutamine aggregation protein cell-based model of Huntington's disease. These studies identified distinct protein interaction partners of HSF1 as well as changes in the magnitude of shared interactions as a function of each stressful condition. Several novel HSF1-interacting proteins were identified that encompass a wide variety of cellular functions, including roles in DNA repair, mRNA processing, and regulation of RNA polymerase II. One HSF1 partner, CTCF, interacted with HSF1 in a stress-inducible manner and functions in repression of specific HSF1 target genes. Understanding how HSF1 regulates gene repression is a crucial question, given the dysregulation of HSF1 target genes in both cancer and neurodegeneration. These studies expand our understanding of HSF1-mediated gene repression and provide key insights into HSF1 regulation via protein-protein interactions.

Organisms are constantly challenged with adapting to stressful conditions such as protein misfolding, inflammation, environmental toxicants, increased temperature, and rapid cell

proliferation. A crucial player in the cellular stress response is heat shock transcription factor 1 (HSF1), a regulator of stress-protective gene transcription (1–4). In normal cells in the absence of acute stress, the majority of HSF1 is maintained in an inactive, monomeric state and resides predominantly in the cytoplasm (5, 6). This basal repression of HSF1 is achieved through protein interactions including a multichaperone complex (7–12), proteins that post-translationally modify HSF1, and repressive intramolecular interactions between leucine zipper regions (13–17). HSF1 is activated in response to stressful conditions through oligomerization and is retained in the nucleus where it binds heat shock elements (HSEs) adjacent to target genes or in distal regulatory elements to promote cell survival through several mechanisms (1, 2). Genes activated by HSF1 encode protein chaperones, the ubiquitin proteasome degradation machinery, cell cycle determinants, transcriptional regulators, and many other proteins involved in diverse processes (3). A more recently explored facet of HSF1 biology is its role in repressing gene transcription, including those involved in apoptosis, inflammation, and transcription (18–22).

Although HSF1 is primarily cytosolic and inactive in the absence of stress, a small fraction of HSF1 remains in the nucleus and binds to target genes in the absence of acute stress (23, 24). The characterization of HSF1 direct target genes in basal conditions and in response to acute stress, chronic protein misfolding diseases, or in cancer cells reveals cell-context-specific sets of HSF1-regulated genes, some of which contribute to pathogenesis (24–28). How HSF1 protein regulators are altered to influence different disease signatures is not well understood.

Owing to the central role HSF1 plays in coping with stresses that disrupt protein homeostasis, HSF1 is of great interest in neurodegenerative diseases that arise from chronic protein misfolding including Alzheimer's disease, Parkinson's disease, and Huntington's disease (HD) (5, 29–33). In HD, a CAG triplet nucleotide expansion in exon 1 of the Huntingtin gene results in a pathogenic polyglutamine (polyQ) expansion in the Htt protein (mutant Htt, mHtt) (34). mHtt protein aggregates in complexes with cellular components including cell signaling

This article contains [supporting information](#).

* For correspondence: Dennis J. Thiele, dthiele@sisupharma.com.

Present address for Anniina Vihervaara: Institute of Biotechnology, University of Helsinki, Helsinki, Finland.

Present address for Rocio Gomez-Pastor: Department of Neuroscience, University of Minnesota, Minneapolis, Minnesota, USA.

Present address for Dennis J. Thiele: Sisu Pharma, Inc, Durham, North Carolina, USA.

ARF-AID: A Rapidly Inducible Protein Degradation System That Preserves Basal Endogenous Protein Levels

Kizhakke Mattada Sathyan,^{1,4} Thomas G. Scott,¹ and Michael J. Guertin^{1,2,3}

¹Biochemistry and Molecular Genetics Department, University of Virginia, Charlottesville, Virginia

²Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia

³Cancer Center, University of Virginia, Charlottesville, Virginia

⁴Corresponding author: sk8fz@virginia.edu

Inducible degron systems are widely used to specifically and rapidly deplete proteins of interest in cell lines and organisms. An advantage of inducible degradation is that the biological system under study remains intact and functional until perturbation, a feature that necessitates that the endogenous levels of the protein are maintained. However, endogenous tagging of genes with auxin-inducible degrons (AID) can result in chronic, auxin-independent proteasome-mediated degradation. The ARF-AID (auxin-response factor–auxin-inducible degron) system is a re-engineered auxin-inducible protein degradation system. The additional expression of the ARF-PB1 domain prevents chronic, auxin-independent degradation of AID-tagged proteins while preserving rapid auxin-induced degradation of tagged proteins. Here, we describe the protocol for engineering human cell lines to implement the ARF-AID system for specific and inducible protein degradation. These methods are adaptable and can be extended from cell lines to organisms. © 2020 The Authors.

Basic Protocol 1: Generation of ARF-P2A-TIR1 progenitor cells

Basic Protocol 2: Designing, cloning, and testing of a gene-specific sgRNA

Basic Protocol 3: Design and amplification of a homology-directed repair construct (C-terminal tagging)

Alternate Protocol 1: Design and amplification of a homology-directed repair construct (N-terminal tagging)

Basic Protocol 4: Tagging of a gene of interest with AID

Alternate Protocol 2: Establishment of an ARF-AID clamp system

Basic Protocol 5: Testing of auxin-mediated degradation of the AID-tagged protein

Keywords: ARF-AID • auxin • auxin-inducible degron • protein degradation

How to cite this article:

Sathyan, K. M., Scott, T. G., & Guertin, M. J. (2020). ARF-AID: A rapidly inducible protein degradation system that preserves basal endogenous protein levels. *Current Protocols in Molecular Biology*, 132, e124. doi: 10.1002/cpmb.124

INTRODUCTION

A diversity of molecular tools that disrupt genes are commonly used to gain mechanistic insight into protein function, and many of the methods available today disrupt gene

Gene expression

Defining data-driven primary transcript annotations with *primaryTranscriptAnnotation* in R

Warren D. Anderson ^{1,*}, Fabiana M. Duarte², Mete Civelek^{1,3,†} and Michael J. Guertin^{1,4,*}

¹Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA, ²Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA, ³Department of Biomedical Engineering, University of Virginia, Charlottesville, VA 22903, USA and ⁴Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22903, USA

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Co-Senior Authors.

Associate Editor: Inanc Birol

Received on September 26, 2019; revised on December 6, 2019; editorial decision on December 30, 2019; accepted on January 3, 2020

Abstract

Summary: Nascent transcript measurements derived from run-on sequencing experiments are critical for the investigation of transcriptional mechanisms and regulatory networks. However, conventional mRNA gene annotations significantly differ from the boundaries of primary transcripts. New primary transcript annotations are needed to accurately interpret run-on data. We developed the *primaryTranscriptAnnotation* R package to infer the transcriptional start and termination sites of primary transcripts from genomic run-on data. We then used these inferred coordinates to annotate transcriptional units identified *de novo*. This package provides the novel utility to integrate data-driven primary transcript annotations with transcriptional unit coordinates identified in an unbiased manner. Highlighting the importance of using accurate primary transcript coordinates, we demonstrate that this new methodology increases the detection of differentially expressed transcripts and provides more accurate quantification of RNA polymerase pause indices.

Availability and implementation: <https://github.com/WarrenDavidAnderson/genomicsRpackage/tree/master/primaryTranscriptAnnotation>.

Contact: warrena@virginia.edu or mjc7y@virginia.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Quantification of nascent transcription is critical for resolving temporal patterns of gene regulation and defining gene regulatory networks. Genome-wide nascent run-on sequencing methods are commonly used to quantify nascent transcription (Core *et al.*, 2008; Kwak *et al.*, 2013). Existing gene annotations are inadequate for both quantifying nascent transcripts and determining the RNA polymerase location relative to gene features. Analyses of run-on data indicate that annotated transcription start sites (TSSs) are often inaccurate. Moreover, transcription extends beyond the 3' polyadenylation region (Proudfoot, 2016), thereby rendering transcription termination sites (TTSs) distinct from annotated mRNA ends. Identification of more accurate TSSs and TTSs for primary transcripts is important for accurate transcript quantification from run-on data.

We present the R package *primaryTranscriptAnnotation* to directly infer TSSs and TTSs of annotated genes. This package also expands upon existing *de novo* transcription unit (TU) identification tools (Chae *et al.*, 2015; Heinz *et al.*, 2010) by facilitating the

assignment of gene identifiers to the identified TUs. Our improved annotations increase the detection of differential transcript expression and quantification of RNA polymerase pausing. This package improves precision in analyses of critical phenomena related to transcriptional regulation and can be easily incorporated into standard genomic run-on analysis workflows.

2 Description

The focus of this software package is to identify the coordinates of nascent transcripts corresponding to annotated genes. This tool provides a useful supplement to other tools used for nascent transcript analysis. For instance, dREG can be used to infer the coordinates of active transcription regulatory elements, such as enhancers, from nascent transcript data (Wang *et al.*, 2019). We distinguish two related tasks performed by our package: (i) integration of run-on data and existing gene annotations to refine estimates of TSSs and TTSs and (ii) combining the results of the first task with the results of an unsupervised TU identification method to annotate the TUs.

RESOURCE/METHODOLOGY

An improved auxin-inducible degron system preserves native protein levels and enables rapid and specific protein depletion

Kizhakke Mattada Sathyan,¹ Brian D. McKenna,¹ Warren D. Anderson,² Fabiana M. Duarte,³ Leighton Core,⁴ and Michael J. Guertin^{1,2,5}

¹Biochemistry and Molecular Genetics Department, University of Virginia, Charlottesville, Virginia 22908, USA; ²Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia 22908, USA; ³Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA; ⁴Department of Molecular and Cell Biology, University of Connecticut, Storrs, Connecticut 06269, USA; ⁵Cancer Center, University of Virginia, Charlottesville, Virginia 22908, USA

Rapid perturbation of protein function permits the ability to define primary molecular responses while avoiding downstream cumulative effects of protein dysregulation. The auxin-inducible degron (AID) system was developed as a tool to achieve rapid and inducible protein degradation in nonplant systems. However, tagging proteins at their endogenous loci results in chronic auxin-independent degradation by the proteasome. To correct this deficiency, we expressed the auxin response transcription factor (ARF) in an improved inducible degron system. ARF is absent from previously engineered AID systems but is a critical component of native auxin signaling. In plants, ARF directly interacts with AID in the absence of auxin, and we found that expression of the ARF PB1 (Phox and Bem1) domain suppresses constitutive degradation of AID-tagged proteins. Moreover, the rate of auxin-induced AID degradation is substantially faster in the ARF-AID system. To test the ARF-AID system in a quantitative and sensitive manner, we measured genome-wide changes in nascent transcription after rapidly depleting the ZNF143 transcription factor. Transcriptional profiling indicates that ZNF143 activates transcription in *cis* and regulates promoter-proximal paused RNA polymerase density. Rapidly inducible degradation systems that preserve the target protein's native expression levels and patterns will revolutionize the study of biological systems by enabling specific and temporally defined protein dysregulation.

[**Keywords:** auxin-inducible degron; auxin response factor; ZNF143; transcription factors; RNA polymerase pausing]

Supplemental material is available for this article.

Received May 21, 2019; revised version accepted July 18, 2019.

The function of proteins can be studied in cells using RNAi depletion, loss-of-function mutants, temperature-sensitive mutations, small molecule inhibitors, CRISPR interference, or nucleic acid aptamers. The two greatest limitations of these methods are (1) the chronic nature of permanently disrupting function and (2) the limited availability and specificity of inhibitor molecules. Inducible degradation methods, such as auxin-inducible degron (AID) (Nishimura et al. 2009) or degradation tag (dTAG) (Nabet et al. 2018) systems, can overcome these limitations. Rapidly regulated systems and inhibitors permit measurements of primary changes in molecular, cellular, and organismal phenotypes and subsequent tracking of

the cascade that accompanies immediate protein dysregulation. Gene-editing techniques now permit endogenous tagging of genes, which should preserve the target gene's expression levels and patterns. Endogenous fusion of target genes with rapidly inducible degradation tags have the potential to revolutionize the way we study biological systems by providing specific and temporally defined perturbation techniques for any protein of interest.

Auxin is a plant hormone that regulates various aspects of plant growth and development. In plants, auxin signaling triggers a rapid switch between transcriptional repression and transcriptional activation (Supplemental Fig. S1; Calderon-Villalobos et al. 2010; Lavy and Estelle 2016; Li et al. 2016). There are three key components in this signal

Corresponding author: guertin@virginia.edu

Article published online ahead of print. Article and publication date are online at <http://www.genesdev.org/cgi/doi/10.1101/gad.328237.119>. Freely available online through the *Genes & Development* Open Access option.

© 2019 Sathyan et al. This article, published in *Genes & Development*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

Genetic and epigenetic determinants establish a continuum of Hsf1 occupancy and activity across the yeast genome

David Pincus^{a,†,*}, Jayamani Anandhakumar^{b,†}, Prathapan Thiru^a, Michael J. Guertin^c, Alexander M. Erkinen^{b,‡}, and David S. Gross^{b,*}

^aWhitehead Institute for Biomedical Research, Cambridge, MA 02142; ^bDepartment of Biochemistry and Molecular Biology, Louisiana State University Health Sciences Center, Shreveport, LA 71130; ^cDepartment of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908

ABSTRACT Heat shock factor 1 is the master transcriptional regulator of molecular chaperones and binds to the same *cis*-acting heat shock element (HSE) across the eukaryotic lineage. In budding yeast, Hsf1 drives the transcription of ~20 genes essential to maintain proteostasis under basal conditions, yet its specific targets and extent of inducible binding during heat shock remain unclear. Here we combine Hsf1 chromatin immunoprecipitation sequencing (seq), nascent RNA-seq, and Hsf1 nuclear depletion to quantify Hsf1 binding and transcription across the yeast genome. We find that Hsf1 binds 74 loci during acute heat shock, and these are linked to 46 genes with strong Hsf1-dependent expression. Notably, Hsf1's induced DNA binding leads to a disproportionate (~7.5-fold) increase in nascent transcription. Promoters with high basal Hsf1 occupancy have nucleosome-depleted regions due to the presence of "pioneer factors." These accessible sites are likely critical for Hsf1 occupancy as the activator is incapable of binding HSEs within a stably positioned, reconstituted nucleosome. In response to heat shock, however, Hsf1 accesses nucleosomal sites and promotes chromatin disassembly in concert with the Remodels Structure of Chromatin (RSC) complex. Our data suggest that the interplay between nucleosome positioning, HSE strength, and active Hsf1 levels allows cells to precisely tune expression of the proteostasis network.

Monitoring Editor
William P. Tansey
Vanderbilt University

Received: Jun 12, 2018
Revised: Oct 1, 2018
Accepted: Oct 11, 2018

This article was published online ahead of print in MBoc in Press (<http://www.molbiolcell.org/cgi/doi/10.1091/mbc.E18-06-0353>) on October 17, 2018.

[†]These authors contributed equally to this work.

[‡]Present address: Department of Pharmaceutical Sciences, College of Pharmacy and Health Sciences, Butler University, Indianapolis, IN 46208.

Author contributions: D.S.G., D.P., and J.A. designed the project. J.A., D.P., and A.M.E. conducted the experiments. D.P., P.T., and M.G. performed the bioinformatics analysis. D.P., J.A., A.M.E., and D.S.G. analyzed the data. D.P., J.A., and A.M.E. made the figures. D.S.G. and D.P. wrote the paper.

*Address correspondence to: David Pincus (pincus@wi.mit.edu) and David S. Gross (dgross@lsuhsc.edu).

Abbreviations used: Ac, acetylation; EMSA, electrophoretic mobility shift assay; GEO, Gene Expression Omnibus; GO, gene ontology; HS, heat shock; HSE, heat shock element; Hsf1, heat shock factor 1; NHS, non-heat shock; me, methylation; MEME, Multiple EM for Motif Elicitation; NFR, nucleosome-free region; rapa, rapamycin; RPM, reads per million mapped reads; RSC, Remodels Structure of Chromatin; RT, room temperature; SPMR, signal per million mapped reads; TE, Tris-EDTA; TSS, transcription start site; UAS, upstream activation sequence.

© 2018 Pincus, Anandhakumar, et al. This article is distributed by The American Society for Cell Biology under license from the author(s). Two months after publication it is available to the public under an Attribution-Noncommercial-Share Alike 3.0 Unported Creative Commons License (<http://creativecommons.org/licenses/by-nc-sa/3.0>).

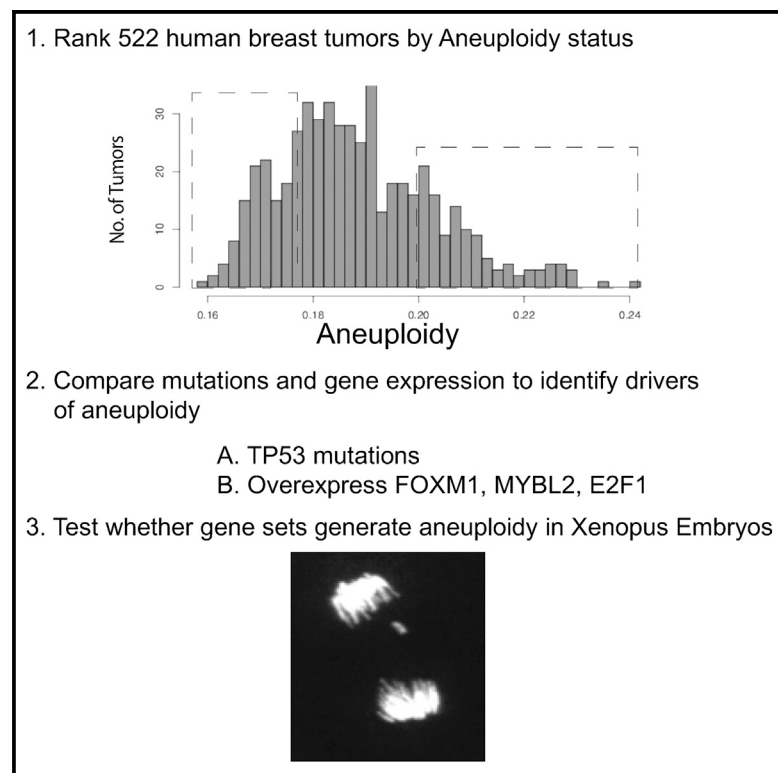
"ASCB®," "The American Society for Cell Biology®," and "Molecular Biology of the Cell®" are registered trademarks of The American Society for Cell Biology.

INTRODUCTION

The cellular response to thermal stress is directed by a transcriptional program whose basic components—sequence-specific activator, *cis*-response DNA element, and core target genes—have been conserved since the last common ancestor in the eukaryotic lineage (Verghese et al., 2012). Heat shock factor 1 (Hsf1), master activator of the heat shock response, is a winged helix-turn-helix transcription factor that recognizes DNA sequence motifs (heat shock elements) located upstream of genes encoding chaperones and other cytoprotective heat shock proteins (HSPs) (Wu, 1995; Gomez-Pastor et al., 2018). HSPs maintain cellular protein homeostasis (proteostasis) and are required in higher concentrations under stressful conditions to contend with unfolded cytosolic and nucleoplasmic proteins. This is principally achieved via Hsf1-mediated transcriptional up-regulation of the genes encoding these proteins. In addition to its evolutionarily conserved role, mammalian HSF1 contributes to oncogenesis by driving distinct transcriptional programs in both tumor cells and their supporting stroma (Dai et al., 2007; Mendiolo et al., 2012; Scherz-Shouval et al., 2014). HSF1 function also has been linked to normal development, neurodegenerative disease, and aging

Identification of Drivers of Aneuploidy in Breast Tumors

Graphical Abstract



Authors

Katherine Pfister, Justyna L. Pipka, Colby Chiang, ..., Michael J. Guertin, Ira M. Hall, P. Todd Stukenberg

Correspondence

pts7h@virginia.edu

In Brief

Pfister et al. analyzed TCGA sequence data to identify drivers of aneuploidy in breast tumors. TP53 is mutated in most aneuploid tumors, and a large number of genes that control mitosis are overexpressed. The oncogenes E2F1, MYBL2, and FOXM1 that regulate mitotic transcription drive the overexpression of mitotic proteins to lower the fidelity of chromosome segregation.

Highlights

- The authors analyze TCGA sequence data to identify highly aneuploid breast tumors
- Mitotic transcriptional regulators and target genes are overexpressed in aneuploid tumors
- Mutations in TP53 co-associate with the overexpression of mitotic transcription factors
- Overexpression of MYBL2, E2F1, and FOXM1 immediately reduces mitotic fidelity in tissues



Data and text mining

BART: a transcription factor prediction tool with query gene sets or epigenomic profiles

Zhenjia Wang¹, Mete Civelek^{1,2}, Clint L. Miller^{1,2,3,4},
Nathan C. Sheffield^{1,2,3,4}, Michael J. Guertin^{1,4} and
Chongzhi Zang^{1,2,3,4,5,*}

¹Center for Public Health Genomics, ²Department of Biomedical Engineering, ³Department of Public Health Sciences, ⁴Department of Biochemistry and Molecular Genetics and ⁵Cancer Center, University of Virginia, Charlottesville, VA 22908, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on December 18, 2017; revised on March 9, 2018; editorial decision on March 18, 2018; accepted on March 27, 2018

Abstract

Summary: Identification of functional transcription factors that regulate a given gene set is an important problem in gene regulation studies. Conventional approaches for identifying transcription factors, such as DNA sequence motif analysis, are unable to predict functional binding of specific factors and not sensitive enough to detect factors binding at distal enhancers. Here, we present binding analysis for regulation of transcription (BART), a novel computational method and software package for predicting functional transcription factors that regulate a query gene set or associate with a query genomic profile, based on more than 6000 existing ChIP-seq datasets for over 400 factors in human or mouse. This method demonstrates the advantage of utilizing publicly available data for functional genomics research.

Availability and implementation: BART is implemented in Python and available at <http://faculty.virginia.edu/zanglab/bart>.

Contact: zang@virginia.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Transcriptional regulation of gene expression plays a critical role in many cellular processes, including cancer development and progression (Bradner *et al.*, 2017; Lambert *et al.*, 2018). Identification of functional transcription factors is essential for understanding gene regulatory mechanisms in such processes. In gene expression profiling studies, ontology and pathway analyses (Huang *et al.*, 2009; McLean *et al.*, 2010; Subramanian *et al.*, 2005) can identify functional annotations of differentially expressed genes; however, this approach is unable to predict transcription factors that regulate those gene sets. Most existing methods for cis-regulatory prediction rely upon detecting overrepresented DNA sequence motifs near the gene promoters to identify sequence-specific DNA-binding factors (Boeva, 2016). Such methods are limited by the context-specific nature of transcription factor activity and by multiple factors sharing

similar motifs (Jolma *et al.*, 2013). Moreover, most cis-regulatory events in mammalian genomes occur at distal enhancers, which cover much larger regions than promoters but without direct assignment to target genes; these regions are usually difficult to capture by motif scan alone (Shlyueva *et al.*, 2014).

Several methods have been developed to overcome these limitations of motif-based, promoter-biased approaches using comprehensive epigenomic information (Dozmorov, 2017), such as DNaseI hypersensitive sites (Sheffield *et al.*, 2013). Model-based analysis of regulation of gene expression (MARGE) is a method developed for modeling differential gene expression using a compendium of public H3K27ac ChIP-seq datasets (Wang *et al.*, 2016). By quantifying the regulatory potential of active enhancer histone mark H3K27ac on each gene in the genome from each ChIP-seq dataset, MARGE uses a semi-supervised learning approach to predict a genome-wide

Universal correction of enzymatic sequence bias reveals molecular signatures of protein/DNA interactions

André L. Martins¹, Ninad M. Walavalkar¹, Warren D. Anderson^{1,2}, Chongzhi Zang^{1,2} and Michael J. Guertin^{1,2,*}

¹Biochemistry and Molecular Genetics Department, University of Virginia, Charlottesville, Virginia, USA and ²Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA

Received February 10, 2017; Revised September 19, 2017; Editorial Decision October 16, 2017; Accepted October 18, 2017

ABSTRACT

Coupling molecular biology to high-throughput sequencing has revolutionized the study of biology. Molecular genomics techniques are continually refined to provide higher resolution mapping of nucleic acid interactions and structure. Sequence preferences of enzymes can interfere with the accurate interpretation of these data. We developed *seqOutBias* to characterize enzymatic sequence bias from experimental data and scale individual sequence reads to correct intrinsic enzymatic sequence biases. *seqOutBias* efficiently corrects DNase-seq, TACH-seq, ATAC-seq, MNase-seq and PRO-seq data. We show that *seqOutBias* correction facilitates identification of true molecular signatures resulting from transcription factors and RNA polymerase interacting with DNA.

INTRODUCTION

The field of molecular genomics emerged as classical molecular biology techniques were coupled to high-throughput sequencing technology to provide unprecedented genome-wide measurements of molecular features. Molecular genomics assays, such as DNase-seq (1,2), ChIP-exo (3) and PRO-seq (4,5), are converging on single-nucleotide resolution measurements. The enzymes that are routinely used in molecular biology and cloning have inherent and often uncharacterized sequence preferences. These preferences manifest more prominently as the resolution of genomic assays increases. Therefore, we developed *seqOutBias* (<https://github.com/guertinlab/seqOutBias>) to characterize and correct enzymatic biases that can obscure proper interpretation of molecular genomics data.

Enzymatic hypersensitivity assays, such as DNase-seq (1,2), TACH-seq (6) and ATAC-seq (7), have the potential to measure transcription factor (TF) binding sites genome-

wide in a single experiment. These assays strictly measure enzymatic (DNase, Tn5 transposase, Benzonase or Cyanase) accessibility to DNA and not a specific biological event, making data challenging to deconvolve. Standard algorithms scan for footprints, which are depletions of signal in larger regions of hypersensitivity (8–12). Many TFs, however, do not exhibit composite footprints if enzymatic cut frequency is averaged at all ChIP-seq validated binding sites with strong consensus motifs (10–13). Moreover, the inability to detect a footprint at any individual TF binding site results in high false negative rates for footprinting algorithms (14). Accurate footprinting is also confounded by the artifactual molecular signatures that result from enzymatic sequence preference (10–12). DNase footprinting algorithms can incorporate DNase cut preference data to abrogate this bias (12,14,15). However, no existing tools specialize in correcting intrinsic sequence bias for a diverse set of enzymes and experimental methodologies.

We find that correcting for enzymatic sequence bias highlights true molecular signatures that result from TF/DNA interactions. Despite the limitations of enzymatic hypersensitivity footprinting and sequence bias signatures, hypersensitive regions reveal a near-comprehensive set of functional regulatory regions in the genome (16). Therefore, we present *seqOutBias*, which calculates sequence bias from an aligned BAM file and corrects individual reads accordingly. While this software does not directly infer TF binding, correction of sequence bias provides a more accurate measurement of three key features of enzymatic hypersensitivity data: (i) raw peak height; (ii) footprint depth; and (iii) true molecular signatures. These measurements, taken together with DNA sequence, can be used to develop algorithms that infer TF binding genome-wide. Moreover, footprint depth and the presence of true molecular signatures are unique to each TF and these features should be characterized for each TF using corrected data in order to optimize TF-binding inference algorithms.

*To whom correspondence should be addressed. Tel: +1 434 924 0328; Email: guertin@virginia.edu

Parallel factor ChIP provides essential internal control for quantitative differential ChIP-seq

Michael J. Guertin¹, Amy E. Cullen², Florian Markowetz² and Andrew N. Holding^{2,*}

¹Department of Biochemistry and Molecular Genetics, School of Medicine, University of Virginia, Charlottesville, VA 22908, USA and ²Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, Cambridgeshire CB2 0RE, UK

Received September 11, 2017; Revised March 15, 2018; Editorial Decision March 21, 2018; Accepted March 28, 2018

ABSTRACT

A key challenge in quantitative ChIP combined with high-throughput sequencing (ChIP-seq) is the normalization of data in the presence of genome-wide changes in occupancy. Analysis-based normalization methods were developed for transcriptomic data and these are dependent on the underlying assumption that total transcription does not change between conditions. For genome-wide changes in transcription factor (TF) binding, these assumptions do not hold true. The challenges in normalization are confounded by experimental variability during sample preparation, processing and recovery. We present a novel normalization strategy utilizing an internal standard of unchanged peaks for reference. Our method can be readily applied to monitor genome-wide changes by ChIP-seq that are otherwise lost or misrepresented through analytical normalization. We compare our approach to normalization by total read depth and two alternative methods that utilize external experimental controls to study TF binding. We successfully resolve the key challenges in quantitative ChIP-seq analysis and demonstrate its application by monitoring the loss of Estrogen Receptor- α (ER) binding upon fulvestrant treatment, ER binding in response to estradiol, ER mediated change in H4K12 acetylation and profiling ER binding in patient-derived xenographs. This is supported by an adaptable pipeline to normalize and quantify differential TF binding genome-wide and generate metrics for differential binding at individual sites.

INTRODUCTION

ChIP combined with high-throughput sequencing (ChIP-seq) quantifies the relative binding intensity of protein/DNA interactions genome-wide for a single con-

dition (1–3). However, comparing relative intensities of binding between samples and between conditions is an ongoing challenge (4–8). Conventionally, correcting for sample-to-sample variability between conditions occurs at the analysis stage (9–12), but these methods assume that experimental variables remain constant between datasets and assume comparable genomic binding of the protein between conditions. In practice, different efficiencies in nuclear extraction, DNA shearing and immunoprecipitation present potential points within a typical ChIP-seq protocol (13) to introduce experimental variation and error (14). Analytical normalization methods exist to control for variability between samples of the same condition (14,15), but these methods cannot account for experimental variation between conditions (7). In order to approximate normalization between conditions the field has exploited a deficiency in ChIP-seq. In short, the total read depth is used as a normalization factor because the vast majority of ChIP-seq reads are outside of true transcription factor (TF) binding sites (8,9). Nonetheless, this approach does not control for any of the aforementioned causes of experimental variability and differences in DNA recovery can be interpreted as differential binding. Previous studies have aimed to resolve these challenges when analyzing genome-wide changes through the use of external spike-in controls (4,5). These methods rely on xenogeneic chromatin (i.e. from a second organism) and either a second species-specific antibody (5), or the cross-reactivity of a single antibody to the factor of interest (4) in both organisms.

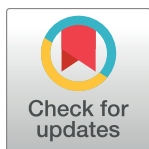
Here we present a method, termed parallel-factor ChIP, that utilizes a second antibody to provide an internal control. The process of utilizing a second antibody against the target chromatin avoids the need of a xenogeneic spike-in and controls for more experimental variables than previous methods. In contrast to spike-in methods, this approach controls for cell lysis conditions, immunoprecipitation efficiency and sonication fragment size. Moreover, parallel-factor ChIP is not dependent upon accurate quantification of spike-in chromatin. We present this method alongside the application of two xenogeneic methods for the analysis of the fold-change in TF binding between two conditions. Fur-

*To whom correspondence should be addressed. Tel: +44 1223 769 500; Fax: +44 1223 769 510; Email: andrew.holding@cruk.cam.ac.uk

EDITORIAL

Up For A Challenge (U4C): Stimulating innovation in breast cancer genetic epidemiology

Leah E. Mechanic¹*, Sara Lindström², Kenneth M. Daily³, Solveig K. Sieberts³, Christopher I. Amos⁴, Huann-Sheng Chen⁵, Nancy J. Cox⁶, Marina Dathe¹, Eric J. Feuer⁵, Michael J. Guertin⁷, Joshua Hoffman⁸, Yunxian Liu⁷, Jason H. Moore⁹, Chad L. Myers¹⁰, Marylyn D. Ritchie^{11,12}, Joellen Schildkraut¹³, Fredrick Schumacher¹⁴, John S. Witte⁸, Wen Wang¹⁰, Scott M. Williams¹⁴, U4C Challenge Participants[†], U4C Challenge Data Contributors[†], Elizabeth M. Gillanders¹



OPEN ACCESS

Citation: Mechanic LE, Lindström S, Daily KM, Sieberts SK, Amos CI, Chen H-S, et al. (2017) Up For A Challenge (U4C): Stimulating innovation in breast cancer genetic epidemiology. *PLoS Genet* 13(9): e1006945. <https://doi.org/10.1371/journal.pgen.1006945>

Published: September 28, 2017

Copyright: © 2017 Mechanic et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The U4C was funded by the National Institutes of Health (NIH) and the National Cancer Institute (NCI). This project was supported by contract HHSN261201200010I from the NCI Division of Cancer Control and Population Sciences to ICF Macro, which supported a task order 14DNB00071 to Sage Bionetworks and subcontract (123833) to SL for this project. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: SMW is a Section Editor of *PLOS Genetics*. LEM and EMG are employees of NCI. The authors have no other competing interests.

1 Epidemiology and Genomics Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America, **2** Department of Epidemiology, University of Washington School of Public Health, Seattle, Washington, United States of America, **3** Sage Bionetworks, Seattle, Washington, United States of America, **4** Department of Biomedical Data Sciences, Dartmouth College, Lebanon, New Hampshire, United States of America, **5** Surveillance Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America, **6** Division of Genetic Medicine, Vanderbilt University, Nashville, Tennessee, United States of America, **7** Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia, United States of America, **8** Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, United States of America, **9** Division of Informatics, Department of Biostatistics and Epidemiology, Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **10** Department of Computer Science and Engineering, University of Minnesota-Twin Cities, Minneapolis, Minnesota, United States of America, **11** Department of Biochemistry and Molecular Biology, Center for Systems Genomics, Pennsylvania State University, University Park, Pennsylvania, United States of America, **12** Biomedical and Translational Informatics, Geisinger Health System, Danville, Pennsylvania, United States of America, **13** Department of Public Health Sciences, University of Virginia School of Medicine, Charlottesville, Virginia, United States of America, **14** Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, United States of America

* These authors contributed equally to this work.

† Listing of the U4C Challenge Participants and U4C Challenge Data Contributors is provided in the Acknowledgments.

* mechanil@mail.nih.gov

Breast cancer remains a major public health burden, with an estimated 252,710 new cases and 40,610 deaths among women in the United States in 2017 [1]. To identify key genes and biological pathways potentially affecting disease risk, genome-wide association studies (GWAS) have been performed. At present, close to 100 common genetic variants have been associated with breast cancer [2–5]. However, these variants explain only a small proportion of the estimated genetic contribution to the risk of breast cancer [4]. GWAS analyses often report only results from single variant analyses, without exploring the impact of potential combinations or the interplay between variants. Therefore, in 2015, the National Cancer Institute (NCI) launched a challenge to inspire novel cross-disciplinary approaches to more fully decipher the genomic basis of breast cancer, called "Up For A Challenge (U4C)—Stimulating Innovation in Breast Cancer Genetic Epidemiology." The goal of U4C was to promote the development and/or implementation of innovative approaches to identify novel risk pathways—including new genes or combinations of genes, genetic variants, or sets of genomic features—involved in

RESEARCH ARTICLE

Identification of breast cancer associated variants that modulate transcription factor binding

Yunxian Liu¹, Ninad M. Walavalkar¹, Mikhail G. Dozmorov², Stephen S. Rich³, Mete Civelek^{3,4}, Michael J. Guertin^{1,3*}

1 Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia, United States of America, **2** Department of Biostatistics, Virginia Commonwealth University, Richmond, Virginia, United States of America, **3** Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, United States of America, **4** Department of Biomedical Engineering, University of Virginia, Charlottesville, Virginia, United States of America

* guertin@virginia.edu



OPEN ACCESS

Citation: Liu Y, Walavalkar NM, Dozmorov MG, Rich SS, Civelek M, Guertin MJ (2017) Identification of breast cancer associated variants that modulate transcription factor binding. PLoS Genet 13(9): e1006761. <https://doi.org/10.1371/journal.pgen.1006761>

Editor: Scott M. Williams, Case Western Reserve University School of Medicine, UNITED STATES

Received: November 16, 2016

Accepted: April 12, 2017

Published: September 28, 2017

Copyright: © 2017 Liu et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All data were deposited in the Gene Expression Omnibus (GEO) database (accession number GSE89013).

Funding: This research was supported by the Farrow Fellowship to YL, an NCI Cancer Center Support Grant P30 CA44579 subaward to MJG and YL, and an American Cancer Society #IRG 81-001-26 subaward to MJG. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Genome-wide association studies (GWAS) have discovered thousands loci associated with disease risk and quantitative traits, yet most of the variants responsible for risk remain uncharacterized. The majority of GWAS-identified loci are enriched for non-coding single-nucleotide polymorphisms (SNPs) and defining the molecular mechanism of risk is challenging. Many non-coding causal SNPs are hypothesized to alter transcription factor (TF) binding sites as the mechanism by which they affect organismal phenotypes. We employed an integrative genomics approach to identify candidate TF binding motifs that confer breast cancer-specific phenotypes identified by GWAS. We performed *de novo* motif analysis of regulatory elements, analyzed evolutionary conservation of identified motifs, and assayed TF footprinting data to identify sequence elements that recruit TFs and maintain chromatin landscape in breast cancer-relevant tissue and cell lines. We identified candidate causal SNPs that are predicted to alter TF binding within breast cancer-relevant regulatory regions that are in strong linkage disequilibrium with significantly associated GWAS SNPs. We confirm that the TFs bind with predicted allele-specific preferences using CTCF ChIP-seq data. We used The Cancer Genome Atlas breast cancer patient data to identify ANKLE1 and ZNF404 as the target genes of candidate TF binding site SNPs in the 19p13.11 and 19q13.31 GWAS-identified loci. These SNPs are associated with the expression of ZNF404 and ANKLE1 in breast tissue. This integrative analysis pipeline is a general framework to identify candidate causal variants within regulatory regions and TF binding sites that confer phenotypic variation and disease risk.

Author summary

The promise of effective personalized medicine is dependent upon the ability to identify genetic variants in the population that influence disease risk and then use this

ARTICLE

DOI: 10.1038/s41467-017-00151-0

OPEN

Transcriptional response to stress is pre-wired by promoter and enhancer architecture

Anniina Vihervaara^{1,2}, Dig Bijay Mahat², Michael J. Guertin³, Tinyi Chu^{4,5}, Charles G. Danko⁴, John T. Lis² & Lea Sistonen ¹

Programs of gene expression are executed by a battery of transcription factors that coordinate divergent transcription from a pair of tightly linked core initiation regions of promoters and enhancers. Here, to investigate how divergent transcription is reprogrammed upon stress, we measured nascent RNA synthesis at nucleotide-resolution, and profiled histone H4 acetylation in human cells. Our results globally show that the release of promoter-proximal paused RNA polymerase into elongation functions as a critical switch at which a gene's response to stress is determined. Highly transcribed and highly inducible genes display strong transcriptional directionality and selective assembly of general transcription factors on the core sense promoter. Heat-induced transcription at enhancers, instead, correlates with prior binding of cell-type, sequence-specific transcription factors. Activated Heat Shock Factor 1 (HSF1) binds to transcription-primed promoters and enhancers, and CTCF-occupied, non-transcribed chromatin. These results reveal chromatin architectural features that orient transcription at divergent regulatory elements and prime transcriptional responses genome-wide.

¹Faculty of Science and Engineering, Cell Biology, Åbo Akademi University, Turku 20520, Finland. ²Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA. ³Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia 22908, USA. ⁴Department of Biomedical Sciences, The Baker Institute for Animal Health, Cornell University, Ithaca, New York 14853, USA. ⁵Graduate Field of Computational Biology, Cornell University, Ithaca, New York 14853, USA. Correspondence and requests for materials should be addressed to J.T.L. (email: jtl10@cornell.edu) or to L.S. (email: lea.sistonen@abo.fi)

Genome-Wide Identification of Basic Helix–Loop–Helix and NF-1 Motifs Underlying GR Binding Sites in Male Rat Hippocampus

John R. Pooley,¹ Ben P. Flynn,¹ Lars Grøntved,² Songjoon Baek,³ Michael J. Guertin,⁴ Yvonne M. Kershaw,¹ Matthew T. Birnie,¹ Annie Pellatt,¹ Caroline A. Rivers,¹ R. Louis Schiltz,³ Gordon L. Hager,³ Stafford L. Lightman,¹ and Becky L. Conway-Campbell¹

¹Henry Wellcome Laboratories for Integrated Neuroscience and Endocrinology, University of Bristol, Bristol BS1 3NY, United Kingdom; ²Department of Biochemistry and Molecular Biology, University of Southern Denmark, 5230 Odense, Denmark; ³Laboratory for Receptor Biology and Gene Expression, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892; and ⁴University of Virginia School of Medicine, Charlottesville, Virginia 22908

Glucocorticoids regulate hippocampal function in part by modulating gene expression through the glucocorticoid receptor (GR). GR binding is highly cell type specific, directed to accessible chromatin regions established during tissue differentiation. Distinct classes of GR binding sites are dependent on the activity of additional signal-activated transcription factors that prime chromatin toward context-specific organization. We hypothesized a stress context dependency for GR binding in hippocampus as a consequence of rapidly induced stress mediators priming chromatin accessibility. Using chromatin immunoprecipitation sequencing to interrogate GR binding, we found no effect of restraint stress context on GR binding, although analysis of sequences underlying GR binding sites revealed mechanistic detail for hippocampal GR function. We note enrichment of GR binding sites proximal to genes linked to structural and organizational roles, an absence of major tethering partners for GRs, and little or no evidence for binding at negative glucocorticoid response elements. A basic helix–loop–helix motif closely resembling a NeuroD1 or Olig2 binding site was found underlying a subset of GR binding sites and is proposed as a candidate lineage-determining transcription factor directing hippocampal chromatin access for GRs. Of our GR binding sites, 54% additionally contained half-sites for nuclear factor (NF)-1 that we propose as a collaborative or general transcription factor involved in hippocampal GR function. Our findings imply a dose-dependent and context-independent action of GRs in the hippocampus. Alterations in the expression or activity of NF-1/basic helix–loop–helix factors may play an as yet undetermined role in glucocorticoid-related disease susceptibility and outcome by altering GR access to hippocampal binding sites. (*Endocrinology* 158: 1486–1501, 2017)

The hypothalamic–pituitary–adrenal (HPA) axis responds to stress with a delayed secretion of glucocorticoids promoting structural and neurochemical changes within the brain that subserve the adaptive response to stress, but can also contribute to maladaptation

resulting in disease states (1, 2). HPA axis abnormalities are well described as underlying features of a number of neuropsychiatric and neurodegenerative conditions, including major depression (3), posttraumatic stress disorder (4), social phobia (5), Huntington disease (6), and

ISSN Print 0013-7227 ISSN Online 1945-7170

Printed in USA

This article has been published under the terms of the Creative Commons Attribution License (CC BY; <https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. Copyright for this article is retained by the author(s).

Received 13 December 2016. Accepted 3 February 2017.

First Published Online 13 February 2017

Abbreviations: ANOVA, analysis of variance; ATAC-seq, assay for transposase-accessible chromatin sequencing; bHLH, basic helix–loop–helix; ChIP, chromatin immunoprecipitation; ChIP-seq, chromatin immunoprecipitation sequencing; DHS, DNase I hypersensitive site; DHS-seq, DNase I hypersensitive site sequencing; ER, estrogen receptor; GR, glucocorticoid receptor; GRE, glucocorticoid response element; HPA, hypothalamic–pituitary–adrenal; LDTF, lineage-determining transcription factor; NF, nuclear factor; nGRE, negative glucocorticoid response element; NSC, nonstressed control; PFC, prefrontal cortex; SDS, sodium dodecyl sulfate.

Transcription factors GAF and HSF act at distinct regulatory steps to modulate stress-induced gene activation

Fabiana M. Duarte,¹ Nicholas J. Fuda,^{1,3} Dig B. Mahat,¹ Leighton J. Core,^{1,4} Michael J. Guertin,² and John T. Lis¹

¹Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14855, USA; ²Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia 22908, USA

The coordinated regulation of gene expression at the transcriptional level is fundamental to development and homeostasis. Inducible systems are invaluable when studying transcription because the regulatory process can be triggered instantaneously, allowing the tracking of ordered mechanistic events. Here, we use precision run-on sequencing (PRO-seq) to examine the genome-wide heat shock (HS) response in *Drosophila* and the function of two key transcription factors on the immediate transcription activation or repression of all genes regulated by HS. We identify the primary HS response genes and the rate-limiting steps in the transcription cycle that GAGA-associated factor (GAF) and HS factor (HSF) regulate. We demonstrate that GAF acts upstream of promoter-proximally paused RNA polymerase II (Pol II) formation (likely at the step of chromatin opening) and that GAF-facilitated Pol II pausing is critical for HS activation. In contrast, HSF is dispensable for establishing or maintaining Pol II pausing but is critical for the release of paused Pol II into the gene body at a subset of highly activated genes. Additionally, HSF has no detectable role in the rapid HS repression of thousands of genes.

[Keywords: heat shock; transcription regulation; transcription factors; Pol II pausing]

Supplemental material is available for this article.

Received May 24, 2016; revised version accepted July 11, 2016.

The heat-shock (HS) response in *Drosophila melanogaster* has been an effective model system to discover and study mechanisms of transcription and its regulation (Guertin et al. 2010). This highly conserved protective mechanism (Lindquist and Craig 1988) is regulated at the transcriptional level by the HS transcription factor (HSF) (Wu 1995). When activated by stress, HSF potentially activates expression of HS genes, resulting in the accumulation of molecular chaperones, the HS proteins (HSPs), which helps the cell cope with stress-induced protein aggregation and misfolding (Lindquist and Craig 1988).

The transcriptional HS response has been studied largely using *Hsp70* as a model gene (Guertin et al. 2010). *Hsp70* maintains a promoter-proximally paused RNA polymerase II (Pol II) molecule 20–40 base pairs (bp) downstream from the transcription start site (TSS) that is released to transcribe the gene at a low level during normal nonstress conditions (Rougvié and Lis 1988; Rasmussen and Lis 1993). The transcription factor GAGA-

associated factor (GAF) is bound to the promoter of *Hsp70* prior to HS, and GAF is important for the establishment and stability of paused Pol II (Lee et al. 1992, 2008; Kwak et al. 2013). GAF has a key role in keeping the promoter region open and free of nucleosomes (Tsukiyama et al. 1994; Fuda et al. 2015), which allows the recruitment of general transcription factors and the initiation of transcription by Pol II. Upon HS induction, HSF trimerizes and is rapidly recruited to the promoter, where it binds to its cognate HS DNA elements (HSEs) (Xiao and Lis 1988). After binding, HSF directly and indirectly recruits coactivators and other factors (Lis et al. 2000; Saunders et al. 2003; Ardehali et al. 2009) that affect the chromatin structure and composition and promote the release of Pol II from the paused complex into productive elongation. This transition from the paused state into productive elongation depends critically on the positive elongation factor P-TEFb and has been shown to be a very general step that is essential for the regulation of virtually all genes across different species (Rahl et al. 2010; Jonkers et al. 2014). The net result of this molecular cascade is

Present addresses: ³Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, CA 94720, USA; ⁴Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269, USA. Corresponding authors: johnlis@cornell.edu, mjg7y@virginia.edu. Article published online ahead of print. Article and publication date are online at <http://www.genesdev.org/cgi/doi/10.1101/gad.284430.116>. Freely available online through the *Genes & Development* Open Access option.

© 2016 Duarte et al. This article, published in *Genes & Development*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.