AWARD NUMBER: W81XWH-20-1-0522

TITLE: Bridging Epigenomics and Patient Data to Detect Enhanceropathies in Cancer

PRINCIPAL INVESTIGATOR: Emily Hodges

CONTRACTING ORGANIZATION: Vanderbilt University, Nashville, TN

REPORT DATE: July 2022

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
July 2022	Annual	01Jul2021-30Jun2022

4. TITLE AND SUBTITLE

Bridging Epigenomics and Patient Data to Detect Enhanceropathies in Cancer

5a. CONTRACT NUMBER
W81XWH-20-1-0522

5b. GRANT NUMBER
W81XWH-20-1-0522

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Emily Hodges

E-Mail: emily.hodges@vanderbilt.edu

5d. PROJECT NUMBER
0011482279-0001

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Vanderbilt University, 110 21st Avenue
South, Suite 800, Nashville, TN, 37203-
2408

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The objectives of this proposal are to integrate gene regulation data with large-scale patient genotype and Electronic Medical Record (EMR) data to identify and characterize novel, non-coding genetic markers of lymphoma and blood cancers. In the first year of this award, we have made substantial progress towards the stated goals of the project, despite the delays brought on by COVID19. Specifically, we have identified new non-coding single nucleotide polymorphisms (SNPs) tagging gene enhancers associated with risk for lymphoid cancer and have further refined our protocols for SNP prioritization. We have developed new protocols for functionally analyzing non-coding SNPs using a new massively parallel reporter assay and have submitted a manuscript currently in press.

15. SUBJECT TERMS
Electronic health records, DNA methylation, Non-coding SNPs

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRDC
Unclassified	Unclassified	Unclassified	Unclassified	48	19b. TELEPHONE NUMBER *(include area code)*

TABLE OF CONTENTS

	Page
1. Introduction	4
2. Keywords	4
3. Accomplishments	4
4. Impact	9
5. Changes/Problems	9
6. Products	10
7. Participants & Other Collaborating Organizations	10
8. Appendices	11

INTRODUCTION: Genome-wide association studies have reported numerous non-coding single nucleotide polymorphisms (ncSNPs) that are associated with lymphoid leukemia. These ncSNPs are often found in and around enhancers of genes that control immune cell development and function. Enhanceropathies, diseases associated with germline variants within transcriptional enhancers, are driven by perturbation or co-option of gene regulatory circuitries. A serious knowledge gap in our ability to mechanistically link ncSNPs to associated cancer phenotypes creates a critical need to understand the basis of these "onco-enhanceropathies". The objective of this proposal is to use diverse functional genomic and patient data to determine cell-type specific gene regulatory patterns that can pinpoint ncSNP contributions to disease risk. Interpreting ncSNPs is challenging because genetically driven gene expression changes are context-dependent, with many only exhibiting disease-relevant effects in specific cell-types. Studying DNA methylation in normal blood cells, our lab has shown that cell-type specific hypomethylation, particularly within non-coding regions (ncHMRs), define cell-type specific enhancers. Preliminary work in our lab also indicates that immune cell-specific ncHMRs are enriched for reported SNPs linked to immunological traits, including cancer. Consequently, our central hypothesis posits that blood cell ncHMRs are likely to harbor previously unidentified ncSNPs associated with blood cell cancers. To test our hypothesis, we have proposed to identify "onco-enhanceropathies" in B cell specific enhancers using genotyping and patient data from the electronic health record. We further proposed to mechanistically interrogate genotype effects on enhancer function for ncSNPs associated with lymphoid cancers using large-scale, sequencing-based reporter assays and CRISPR-Cas9 genome editing.

KEYWORDS: DNA methylation, Non-coding Single Nucleotide Polymorphism, Enhanceropathies, Massively Parallel Reporter Assays, Lymphoid Cancers, Electronic Health Record, PheWAS, CRISPR-Cas9

ACCOMPLISHMENTS:

What were the major goals of the project?

The following tasks were included in Years 1-2 of the Statement of Work.

Major Task 1: Identify lymphoid "onco-enhanceropathies" in B cell enhancers using genotyping and patient data	Month	% Complete
Collect whole genome bisulfite sequencing (WGBS) datasets data	1	100
Determine enhancer hypomethylated regions (HMRs) from existing WGBS	1-3	100
Perform comparative analysis to isolate lymphoid enhancer HMRs	1-3	100
Apply for access to Vanderbilt EHR (BioVU) data	1	100
QC genotype data and impute LD (linkage disequilibrium) SNPs	1-3	100
Prioritize SNPs for PheWAS	3	100
Perform PheWAS for limited set of immune related ICD10 patient codes	3-6	100
Widen association tests to any PheWAS categories that are related to hematopoietic disorders and blood cancers (69 categories)	3-6	100
Perform test for enrichment among all phenotypes associated with HMR variants	6-9	100
Validate and Perform replication studies in UK Biobank cohort	6-9	50
Prioritize HMRs for functional studies with orthogonal data	9-12	75
Milestone Achieved: significant set of HMR variants associated with blood cancers identified and validated, methodology paper published	12	75
Major Task 2: Mechanistically interrogate genotype effects on enhancer function for ncSNPs associated with lymphoid cancers.		
Design oligos for cloning HMR regions into reporter constructs	9-12	100
Obtain DNA from BioVU or genotype matched HapMap samples	9-12	50
Clone HMR fragments and sequence verify clones	12-18	50
Perform reporter assays	12-18	75
Design and order CRISPR guide RNAs for select iHMRs	9-12	75
Assemble RNPs and nucleofect LCLs	12-18	25

Sequence validate insertions/deletions	12-18	0
Perform RNA-seq and analyze data	18-24	0
Milestone Achieved: Report findings from overall study	18-24	50

What was accomplished under these goals?

Major Activities:

During the second year of this award, we have focused on three major activities, including: 1. Revisions for the manuscript describing ATAC-STARR-seq, a large-scale, sequencing based reporter assay developed by my lab to characterize the regulatory activity of DNA methylation-defined enhancer regions, 2. Extensive functional and genetic characterization of lymphoid specific hypomethylated regions (HMRs) and preparation of a manuscript describing the results, and 3. Building tools to functionally characterize a B cell specific HMR locus harboring SNPs associated with lymphoid leukemia.

Specific Objectives and Key Outcomes:

1. Revision and publication of ATAC-STARR-seq to characterize transcriptional regulatory activity of DNA methylation-defined enhancers (See Appendix for more details):

Our objective was to validate and characterize the transcriptional activity of our cell-type restricted sets of B cell enhancer HMRs. We hypothesized that cell-type specific HMRs contain core sequence variants that drive differential enhancer activity. In our recently published manuscript, we presented important technological advances in massively parallel reporter sequencing. Massively parallel reporter assays (MPRAs) test the capacity of putative gene regulatory elements to drive transcription on a genome-wide scale. HMRs coincide with different classes of gene regulatory elements, the majority of which are derived from open chromatin. Since most gene regulatory activity occurs within accessible chromatin, recently described methods have combined assays that capture these regions—such as assay for transposase-accessible chromatin using sequencing (ATAC-seq)—with self-transcribing active regulatory region sequencing (STARR-seq) to selectively assay the regulatory potential of accessible DNA (ATAC-STARR-seq). To probe the functional activities of HMR-defined genomic sequences, we developed a powerful, multi-omic approach to isolate "driver" HMRs in specific cell contexts. Our ATAC-STARR-seq approach enables simultaneous profiling of chromatin accessibility, transcription factor occupancy, and regulatory activity all from a single DNA fragment library (**Figure 1**). ATAC-STARR-seq uses transposase-assisted cloning to pre-select accessible DNA from native chromatin. This is important for two reasons: 1) we narrow the scope from 100% of the genome to just ~2%, which allows for much deeper sequencing coverage of HMRs, and 2) we only assay HMRs capable of driving transcription endogenously, because they are derived from open chromatin. Together, our approach takes advantage of STARR-seq's unbiased cloning procedure and dynamic range while leveraging ATAC-seq to enrich for cell type specific *cis*-regulatory elements, allowing integrated functional assessment and prioritization of "driver" HMRs.

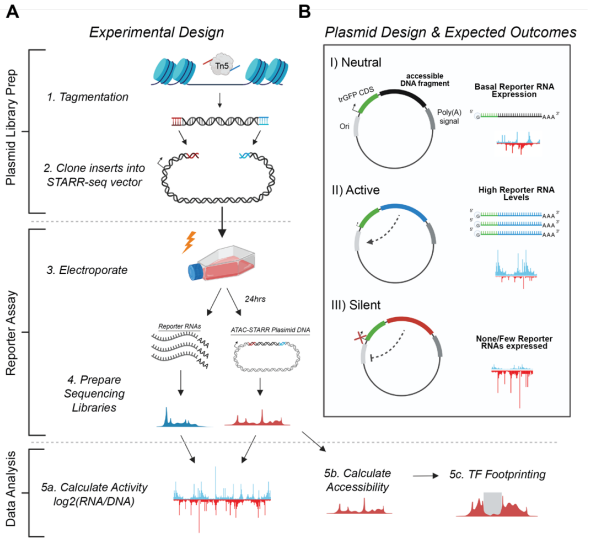


Figure 1. ATAC-STARR-seq methodology. *(A) The experimental design of ATAC-STARR-seq consists of three parts: plasmid library generation, reporter assay, and data analysis. Open chromatin is isolated from cells with Tn5 transposase. Fragments are cloned into a reporter plasmid and the resulting clones are electroporated into cells. Reporter RNAs (blue)—which are transcribed directly off the plasmid—and ATAC-STARR-seq plasmid DNA (red) are harvested, and Illumina-sequencing libraries are prepared and sequenced. The resulting data is analyzed to extract regulatory activity, chromatin accessibility, and TF footprints. (B) Reporter plasmid design and the expected outcomes for neutral, active, and silent regulatory elements. Each ATAC-STARR-seq plasmid within a library contains a truncated GFP (trGFP) coding sequence, a poly-adenylation signal sequence, an origin of replication (Ori) (which moonlights as a minimal core promoter), and the unique open chromatin fragment being assayed.*

Summary from the manuscript abstract: Our strategy, including important updates to the ATAC-STARR-seq assay and workflow, enabled high-resolution testing of ~50 million unique DNA fragments tiling ~101,000 accessible chromatin regions in human lymphoblastoid cells. We discovered that 30% of all accessible regions

contain an activator, a silencer or both. Although few MPRA studies have explored silencing activity, we demonstrate silencers occur at similar frequencies to activators, and they represent a distinct functional group enriched for unique transcription factor motifs and repressive histone modifications. We further show that Tn5 cut-site frequencies are retained in the ATAC-STARR plasmid library compared to standard ATAC-seq, enabling TF occupancy to be ascertained from ATAC-STARR data. With this approach, we found that activators and silencers cluster by distinct TF footprint combinations and these groups of activity represent different gene regulatory networks of immune cell function. Altogether, these data highlight the multi-layered capabilities of ATAC-STARR-seq to comprehensively investigate the regulatory landscape of the human genome all from a single DNA fragment source.

2. Functional and genetic characterization of lymphoid specific hypomethylated regions (HMRs) and preparation of a manuscript describing the results:

We are currently in the process of finalizing a manuscript describing the functional and genetic characterization of lymphoid specific HMRs. These results provide a substantial foundation for the work we have performed over the course of this funding award. The following is a summary of this work:

DNA methylation (DNAme) is essential for balanced multi-lineage cellular differentiation, but exactly how DNAme drives cellular phenotypes is unclear. While >80% of CpG sites are stably methylated, tens of thousands of discrete CpG loci form hypomethylated regions (HMRs). Because they lack DNAme, HMRs are considered transcriptionally permissive, but not all HMRs participate in genome regulation. Contrary to long-held views about the role of DNAme in controlling promoter function, most promoter HMRs are invariable across cell-types regardless of transcriptional status. In contrast, a subset of HMRs is cell-type specific and enriched for tissue specific enhancers. Recently, we showed that, while enhancer HMRs correlate with chromatin accessibility and other indicators of permissive chromatin, their temporal dynamics are distinct from chromatin—HMRs can persist long after accessibility is lost. These studies suggest that establishment of HMRs is an important step in enforcing cell identity, and that patterns of HMR establishment are more complex than previously appreciated.

To understand their functional significance, we systematically dissected HMR patterns across diverse cell types and developmental timepoints, including human embryonic stem cells, fetal spine and heart, adult hematopoietic lineage cells (hematopoietic stem cells, B cells, T cells, Macrophages), liver, and adrenal tissues. Unsupervised clustering of 102,390 distinct HMRs revealed that levels of HMR "sharedness" across cell-types reflects a developmental hierarchy supported by enrichment of stage-specific transcription factors and gene ontologies (**Figure 2**).

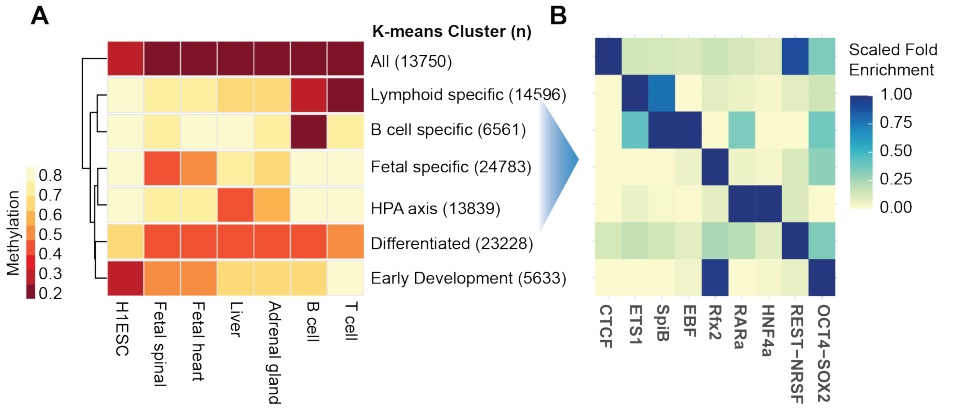


Figure 2. HMRs reconstruct developmental relationships. ***A.*** *Non-coding HMRs were k-means clustered based on their average CpG methylation values across 7 cell types and tissues. The k-means generated HMRs groups that are consistent with the biological relationships of their cell types.* ***B.*** *The TF motif enrichment of each k-means group reflects biological relationships captured in (**A**). Representative TFs were selected from top hits filtered by Benjamini-Hochberg adjusted p-value. Fold enrichment values are normalized from 0 to 1 across TFs. The background comparison file comprises HMRs across all represented cell types. HPA=hypothalamus-pituitary-adrenal axis.*

Using a pseudo-time course of embryonic to adult hematopoietic development, we further show that a majority of HMRs observed in differentiated cells (~70-75%) are established at early developmental stages and accumulate as development progresses. HMRs that arise during differentiation frequently (~35%) establish near existing HMRs (≤ 6kb away), leading to the formation of HMR clusters associated with stronger enhancer activity. Using SNP-based partitioned heritability from GWAS summary statistics of diverse traits and clinical lab values, we discovered that genetic contribution to trait heritability is enriched within HMRs. Moreover, heritability of cell-relevant traits increases with both increasing developmental specificity and HMR clustering, supporting the role for distinct HMR subsets in regulating normal cell function (**Figure 3**).

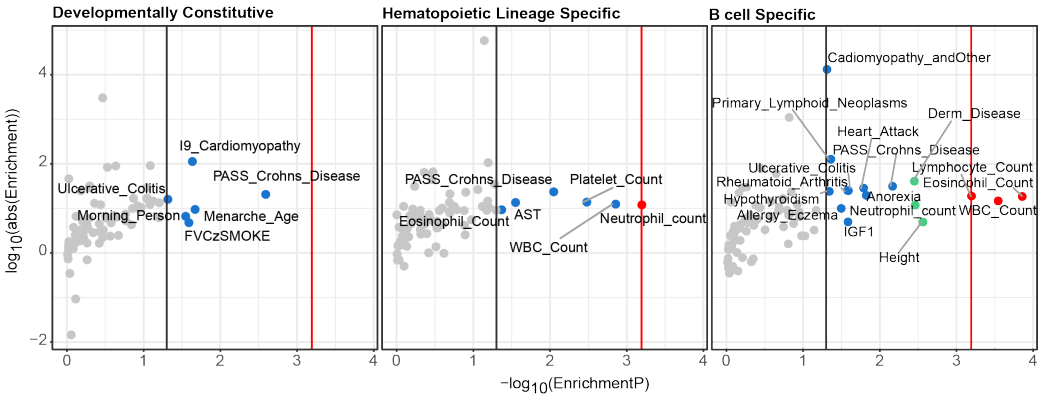


Figure 3. Partitioned heritability of different HMR categories. *Dotplots of partitioned heritability results across 78 traits are shown for three HMR sets identified in B cells. HMRs are grouped by the developmentally distinct cell type in which they were established. For example, "Developmentally Constitutive" HMRs are shared between embryonic stem cells and B cells, while "Hematopoietic Lineage specific" HMRs are shared between hematopoietic stem cells and B cells. Each HMR group was tested for enrichment of genetic heritability with a standard set of 98 base annotations against traits that include both clinical diseases as well as clinical lab values. The color of each point represents significance at nominal (blue), Benjamini-Hochberg FDR (green), and Bonferroni levels (red).*

Altogether, our findings reveal that HMRs can predict cellular phenotype by providing genetically distinct historical records of a cell's journey through development, and this attribute distinguishes DNAme from other epigenomic features. These studies lay the critical groundwork for understanding the relationship between B cell enhancer DNA methylation patterns, genetic variation, and lymphoid leukemia risk.

3. Building tools to functionally characterize a B cell specific HMR locus harboring SNPs associated with lymphoid leukemia:

We performed a phenome-wide association study (PheWAS) and a lab-wide association scan (LabWAS) for selected genotyped SNPs located within or proximal to B cell specific HMRs. A total of 94,474 patient DNAs from Vanderbilt's BioVU repository were genotyped on the Illumina MEGAEx array, and an additional 5,500 were genotyped on the Illumina OMNI-Quad array – covering more than 2 million individual SNP markers. Selected SNPs were tested for associations with 1,814 phenotype codes and 487 clinical laboratory measurements (e.g., cholesterol levels, white blood cells, glucose, etc.) using multivariable logistic or linear regression models, respectively. Our approach revealed a substantial number of previously uncharacterized, high-ranking HMR markers with suggested links to immune phenotypes. This analysis revealed a cluster of B cell specific HMRs that harbor SNPs (rs7976678) significantly linked to multiple immune cell traits including CD8/CD3/CD45 ratios (p=4.79x10^{-10}) and acute lymphoid leukemia (p=3.54x10^{-4}) (**Figure 4**). Despite CD27 being its nearest neighbor, rs7976678 is an eQTL for three other genes: TAPBPL (p=0.00043), ZNF384 (p=0.000017), and LTBR located ~40 kb upstream of the CD27 locus (**Figure 5**). ZNF384 itself is frequently mutated in acute lymphoblastic leukemia, and LTBR is involved in lymphotoxin signaling and lymphoid organ development.

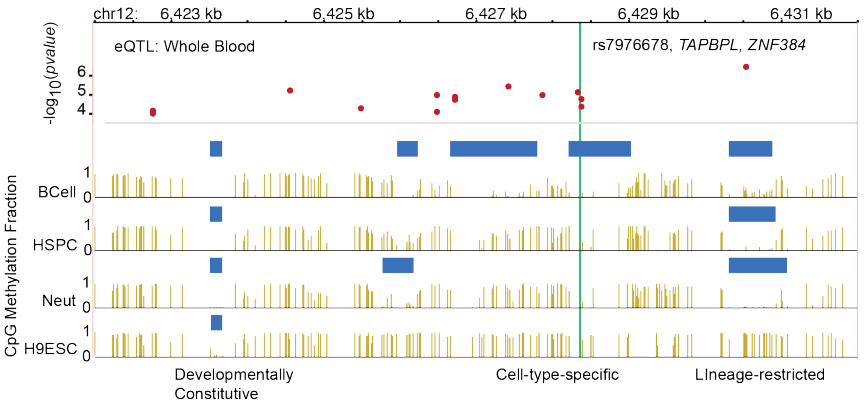


Figure 4. Genome browser tracks of whole genome bisulfite sequencing data and HMRs. *Methylation fractions for individual CpG sites are displayed (vertical yellow lines) for three hematopoietic cell types (B cells, Neutrophils, and hematopoietic stem and progenitor cells) and H9 embryonic stem cells. Genomic intervals representing HMRs are shown in blue boxes. HMRs falling into different "sharedness" categories are labeled below methylation tracks. Locations of eQTLs (red dots) identified in a single tissue (whole blood) are provided (GTex version 8). eQTLs cluster over cell-type specific HMRs. P-values indicate significance of association with transcript variance. Green line highlights location of rs7976678.*

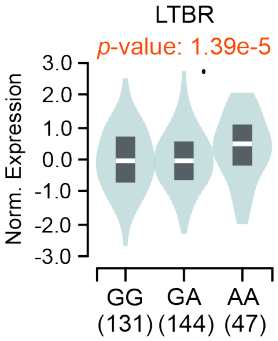


Figure 5. *Boxplot depicts rank normalized gene expression levels of individuals homozygous (G/G), heterozygous (A/G), or homozygous for the minor allele (A/A, MAF = 0.35). Data were obtained from GTEx.*

Together these data provide convergent evidence that this B cell HMR regulates ZNF384 and LTBR, and that it is significant for normal B cell function.

We have prioritized this locus for downstream functional dissections and are building tools for this analysis as described in Aim 2 of our original proposal. These tools include a) generation of luciferase reporter constructs to test the functional consequences of genetic variants on enhancer regulatory activity, and b) the selection of guide RNAs for CRISPR-Cas9-mediated deletions of individual HMRs within the locus to test the effect on transcription of associated genes (e.g., TAPBPL, ZNF384 and LTBR).

Discussion on stated goals not achieved:

Most of the goals stated in our Statement of Work have been achieved except for performing replications studies in the UK biobank cohort and the publication of a methods paper on combining DNA methylation with phenome association studies. This was due in large part to the fact that we needed to prioritize certain aspects of the project that included ATAC-STARR-seq paper revisions. It took nearly a year between first submission of the manuscript and final acceptance. We also prioritized writing of a manuscript that describes the use of comparative methylation profiling to identify cell-type specific hypomethylated regions in B cells, as well as other blood cell types, with detailed characterization of these regions including patterns of genetic heritability within these regions (described above).

What opportunities for training and professional development has the project provided?

The graduate students listed in the Statement of Work, Timothy Scott and Tyler Hansen, present their work related to this project and receive feedback on a monthly basis during Hodges lab meetings. Mr. Scott and Mr. Hansen have presented their research to the broader Vanderbilt community at the Genetics Interest Group, Biochemistry Student Association and Stem Cell and Progenitor Interest Group meetings. Mr. Hansen presented a poster describing ATAC-STARR-seq at the Cold Spring Harbor Systems Biology: Global Regulation of Gene Expression meeting which took place March 9 - March 12, 2022. Mr. Hansen and Mr. Scott have both submitted abstracts to be presented at the American Society for Human Genetics conference which will be held in Los Angeles, CA in October 2022. Both students have prepared manuscripts for publication (see Products) and have had opportunities to mentor rotation students throughout the year. Additionally, the project has provided opportunities to expand their training in genomic and genetic data analysis – both Mr. Scott and Mr. Hansen have sought online workshops and tutorials, including those available through our ongoing lab subscription to DataCamp.

How were the results disseminated to communities of interest?

During the past year, Dr. Hodges has presented data related to this project to the Vanderbilt community through virtual seminars held within the Department of Biochemistry and Vanderbilt Ingram Cancer Center. A manuscript describing the ATAC-STARR-seq method, a novel reporter assay developed as part of aim 2, was deposited in *BioRxiv (*bioRxiv. 2022.03.25.485870) and has been published as an Advanced Online Article in *Genome Research*. The accepted manuscript is attached to this progress report (see Appendices). A poster presentation describing the results of these studies was presented by graduate student Tyler Hansen at the Cold Spring Harbor Systems Biology: Global Regulation of Gene Expression meeting which took place March 9 - March 12, 2022.

What do you plan to do during the next reporting period to accomplish the goals?

During the final reporting period, we plan to complete and publish a paper describing the results accomplished within the scope of Aim 1. This manuscript describes use of comparative DNA methylation profiling to identify B cell enhancers combined with SNP-based heritability analysis of immune-related clinical phenotypes (including lymphoid cancers). This manuscript is currently in preparation for submission to *PLOS Genetics*. We have submitted two abstracts to be presented at the annual American Society for Human Genetics in October this year. Transitioning to Aim 2, we will perform ATAC-STARR-seq in HapMap derived lymphoblastoid cell lines to characterize the transcriptional output of B cell specific enhancers containing genetic risk variants identified in Aim 1. We have already selected candidate loci for CRISPR studies described in Aim 2 and plan to complete

those experiments within the next 6 months. CRISPR-Cas9 genome editing is used to create disease-associated enhancer variants in isogenic lymphoblastoid cells, which we will probe for target gene effects using RNA-seq.

IMPACT:

What was the impact on the development of the principal discipline(s) of the project?

These studies establish a fundamentally new way to understand how DNA methylation bridges the connection between genomes and phenomes, revealing important gene regulatory principles that are essential to understanding why epigenetic instability leads to specific disease outcomes.

What was the impact on other disciplines?

The new approaches developed to execute this project - including 1) ATAC-STARR-seq, and 2) combining DNA methylation and PheWAS to identify cancer risk loci - are broadly applicable to many different types of cancer. Additionally, these methods provide a novel route to identify enhancer-gene relationships, which will be of broad interest to the gene regulation field.

What was the impact on technology transfer?

Nothing to Report

What was the impact on society beyond science and technology?

The goals of this work are to establish methods for detecting and characterizing new, non-coding genetic markers associated with lymphoma and other blood cell cancers. By identifying new genetic risk mutations associated with lymphoma and blood cell cancers, this proposal addresses military relevant environmental risk factors associated with cancer (e.g., ionizing radiation, chemicals, infectious agents, environmental carcinogens, and stress). This approach will be broadly applicable to identify previously undetected gene regulatory contributions, not only to lymphoma and blood cell cancers, but other types of cancer as well. Ultimately, we expect our approach to accelerate the discovery and characterization of gene regulatory contributions to the molecular pathology of lymphoma and leukemia, uncovering new druggable gene targets, leading to a wider range of potential treatment options and better outcomes for patients.

CHANGES/PROBLEMS:

Changes in approach and reasons for change

Since the last reporting period, the only change to report is that we have extended our studies beyond the Vanderbilt Biobank cohort to perform genetic association analysis. The reason for this change is to increase the power of our analysis to identify phenotype associations. We will also perform individual GWAS for specific lymphoma and leukemia EHR codes within these larger cohorts. We originally described the use of such cohorts in our original proposal to replicate significant associations discovered in the Vanderbilt BioVU cohort. Two independent cohorts including the UK Biobank (N~340,000 participants of European ancestry) and the EMERGE network cohort comprising 95,000 individuals of diverse ancestries will be used for this purpose. Significant SNP-phenotype associations discovered in BioVU will be tested in both independent samples and will be considered “replicated” if the effect estimate is consistent with the original finding and the strength of the association exceeds $p = 0.05$ corrected for the number of independent tests performed. In addition to validation studies, these two cohorts provide an opportunity to run additional SNP-phenotype analysis for select lymphoid cancer phenotypes for which there are ample case counts (n>500).

Actual or anticipated problems or delays and actions or plans to resolve them

Nothing to Report

Changes that had a significant impact on expenditures

Nothing to Report

Significant changes in use or care of human subjects

Nothing to Report

Significant changes in use or care of vertebrate animals

Nothing to Report

Significant changes in use of biohazards and/or select agents

Nothing to Report

PRODUCTS:

Journal publications

T. Hansen and E. Hodges; ATAC-STARR-seq reveals transcription factor bound activators and silencers across the chromatin accessible human genome; *Genome Research*; 2022; published; yes

Timothy Scott, Tyler Hansen, Evonne McArthur, Elizabeth Dorans, and Emily Hodges; Cross-tissue patterns of DNA hypomethylation reveal genetically distinct histories of cell development; *TBD*; 2022; in prep; yes

Website(s) or other Internet site(s)
Scripts for ATAC-STARR data analysis are posted on GitHub:
https://github.com/HodgesGenomicsLab/ATAC-STARR-seq.

An interactive version of the ATAC-STARR protocol is posted on protocols.io:
dx.doi.org/10.17504/protocols.io.bggijtue

Technologies or techniques
Protocols for ATAC-STARR-seq has been made publicly available on web portals mentioned above and have been submitted to *BioRxiv (*bioRxiv. 2022.03.25.485870). The ATAC-STARR paper has also been made "open access" as an advanced online publication in *Genome Research*.

Inventions, patent applications, and/or licenses
Nothing to Report

Other Products
Nothing to Report

PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

PDs/PIs:

Emily Hodges, PhD – No change
Scott Hiebert, PhD – No change
Tony Capra, PhD – No change

Other Personnel:

Tyler Hansen – No change
Timothy Scott – No change

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?

Nothing to Report

What other organizations were involved as partners?

Nothing to Report

APPENDICES:

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

ATAC-STARR-seq reveals transcription factor-bound activators and silencers across the chromatin accessible human genome

Tyler J. Hansen[1] and Emily Hodges[1,2,*]

[1] Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, TN, 37232, USA

[2] Vanderbilt Genetics Institute, Vanderbilt University School of Medicine, Nashville, TN, 37232, USA

* Correspondence: Tel: +1 615 875 9991; Email: emily.hodges@vanderbilt.edu

Running Title (50-character limit): Dissecting human gene regulation with ATAC-STARR

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

ABSTRACT

Massively parallel reporter assays (MPRAs) test the capacity of putative gene regulatory elements to drive transcription on a genome-wide scale. Most gene regulatory activity occurs within accessible chromatin, and recently described methods have combined assays that capture these regions—such as assay for transposase-accessible chromatin using sequencing (ATAC-seq)—with self-transcribing active regulatory region sequencing (STARR-seq) to selectively assay the regulatory potential of accessible DNA (ATAC-STARR-seq). Here, we report an integrated approach that quantifies activating and silencing regulatory activity, chromatin accessibility, and transcription factor (TF) occupancy with one assay using ATAC-STARR-seq. Our strategy, including important updates to the ATAC-STARR-seq assay and workflow, enabled high-resolution testing of ~50 million unique DNA fragments tiling ~101,000 accessible chromatin regions in human lymphoblastoid cells. We discovered that 30% of all accessible regions contain an activator, a silencer or both. Although few MPRA studies have explored silencing activity, we demonstrate silencers occur at similar frequencies to activators, and they represent a distinct functional group enriched for unique TF motifs and repressive histone modifications. We further show that Tn5 cut-site frequencies are retained in the ATAC-STARR plasmid library compared to standard ATAC-seq, enabling TF occupancy to be ascertained from ATAC-STARR data. With this approach, we found that activators and silencers cluster by distinct TF footprint combinations and these groups of activity represent different gene regulatory networks of immune cell function. Altogether, these data highlight the multi-layered capabilities of ATAC-STARR-seq to comprehensively investigate the regulatory landscape of the human genome all from a single DNA fragment source.

INTRODUCTION

Transcription is regulated by transcription factors (TFs) and the DNA sequences they bind, called *cis*-regulatory elements. Enhancers, which are a class of *cis*-regulatory elements, are

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

distally located from the genes they target and serve as key drivers of cell-type specific gene expression (Heinz et al. 2015). Because enhancers require TF binding, they are largely dependent on chromatin accessibility to elicit transcriptional activity. Therefore, chromatin accessibility is a vital regulator of enhancer function, and this is evidenced by the observation that ~94% of all ENCODE TF ChIP-seq peaks fall within accessible chromatin (Klemm et al. 2019). In any given cell type, only a small fraction (~2%) of the genome is accessible to TF binding (Thurman et al. 2012; Klemm et al. 2019). In this way, most enhancers are inaccessible and are less likely to drive transcription endogenously.

Enhancers are difficult to identify and validate because they lack uniform features and are less constrained by gene proximity than promoters (Gasperini et al. 2020). Massively parallel reporter assays (MPRAs) were developed to test the regulatory potential of thousands to millions of DNA sequences in parallel, providing high-throughput identification of putative enhancers. Overall, MPRAs test the regulatory potential of genomic regions by cloning them *en masse* into a reporter plasmid and leveraging high-throughput sequencing to quantify regulatory activity (Santiago-Algarra et al. 2017). Among the variety of different vector backbones and assay designs applied to MPRAs, Self-Transcribing Active Regulatory Region sequencing (STARR-seq) is uniquely designed to assay an entire genome for regulatory activity (Melnikov et al. 2012; Patwardhan et al. 2012; Arnold et al. 2013; Inoue et al. 2017; Maricque et al. 2017; Muerdter et al. 2018; Kircher et al. 2019). STARR-seq quantifies regulatory activity genome-wide by cloning randomly fragmented genomic DNA into the 3'UTR of the reporter plasmid. Thus, active enhancers drive transcription of themselves, and activity is quantified by the abundance of its own sequence in the transcript pool, removing the need for barcodes that some MPRAs employ. One major limitation of STARR-seq is that it is technically challenging to accommodate the massive size of the human genome; it requires large-scale cloning procedures and produces shallow sequencing coverage of human regulatory elements (Johnson et al. 2018). In addition, STARR-seq assays both accessible and inaccessible chromatin. Thus, many assayed

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

regions are derived from heterochromatin and are less likely to be transcriptionally active in the cell type in question.

To narrow the scope of the assay, recent methods have combined STARR-seq with techniques that capture accessible chromatin to specifically test the regulatory potential of accessible DNA (Buenrostro et al. 2013; Wang et al. 2018; Chaudhri et al. 2020; Glaser et al. 2021). As a result, these methods only sample a fraction of the human genome (~2%) while assaying nearly all regulatory elements capable of driving transcription endogenously, because they are derived from open chromatin. This approach remains comprehensive while enabling deeper sequencing coverage of biologically relevant genomic regions. Furthermore, integrated approaches have recently been described that combine measurements of chromatin accessibility with analysis of transcription and other epigenomic features from a single population of cells (Kelly et al. 2012; Clark et al. 2018; Barnett et al. 2020; Chen et al. 2022). Similarly, ATAC-STARR-seq has the potential to reveal multiple levels of gene regulatory information simultaneously, but this potential has not been explored. In addition, a complete understanding of gene regulatory activity is lacking with most MPRA approaches because silencing activity is largely overlooked, with a few recent exceptions (Doni Jayavelu et al. 2020; Pang and Snyder 2020; Kim et al. 2021); this is potentially due to technical caveats of distinguishing silencers from either that of missing data or interference from head-on transcriptional conflicts or post-transcriptional silencing mechanisms.

Here, we demonstrate a new workflow that substantially expands the capabilities of ATAC-STARR-seq to extract and measure gene regulatory information. Using this approach, we aimed to identify both activators and silencers, as well as to simultaneously profile chromatin accessibility, and perform TF footprinting. From a single ATAC-STARR-seq dataset, a multi-layered, integrated view of the human genome can be captured—a feature that has not been explored previously. We provide a protocol and code repository so that this new ATAC-STARR-seq workflow may be easily used and adopted by the field.

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

RESULTS

ATAC-STARR-seq Experimental Design

The ATAC-STARR-seq approach is divided into the three main parts: 1) ATAC-STARR-seq plasmid library generation, 2) reporter assay, and 3) data analysis (Figure 1A). To generate ATAC-STARR-seq plasmid libraries, nuclei are isolated from a cell type of interest and exposed to Tn5, the cut-and-paste transposase used in the ATAC-seq method (Buenrostro et al. 2013). Tn5 simultaneously cleaves DNA fragments within accessible chromatin and attaches customizable sequence adapters to their 5' ends. ATAC-STARR-seq adapters are designed to serve as homology arms for direct Gibson cloning into the STARR-seq reporter plasmid, which enables cloning of accessible DNA fragments *en masse*. The resulting ATAC-STARR-seq plasmid library consists of millions of unique plasmids each harbouring their own unique open chromatin-derived DNA fragment.

In our updated ATAC-STARR-seq workflow, we employ the STARR-seq Ori backbone, where the origin of replication (Ori) functions as the minimal promoter (Muerdter et al. 2018) (Supplemental Table S1). Each plasmid in the ATAC-STARR-seq plasmid library contains a truncated GFP (trGFP) coding sequence, a poly-adenylation signal sequence, the Ori, and the unique accessible DNA fragment being assayed (Figure 1B). Critically, the accessible region is cloned into the 3' UTR, so if the accessible region is active, it interacts with the Ori to drive self-transcription. Thus, an accessible region's level of activity is reflected by its own level of expression. Transcripts from ATAC-STARR-seq plasmids, termed "reporter RNAs", are expressed at basal levels from the activity of the Ori itself. This allows detection of silencing activity—the inhibition of the basal expression—in this assay.

Following its creation, the ATAC-STARR-seq plasmid library is transfected via electroporation into a given cell line. From the same flask of cells, both reporter RNAs and plasmid DNA are harvested 24 hours later, then prepared as Illumina sequencing libraries

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

and sequenced. Activity is calculated as the $\log_2$ ratio between normalized read counts from the reporter RNA and plasmid DNA datasets. The re-isolation of plasmid DNA recovers only the ATAC-STARR-seq plasmids that were successfully transfected, thus providing a more accurate representation of the “input” sample than sequencing without transfection. Supplemental Table S1 provides a comparison of experimental and analytical features as well as reported data metrics for the current ATAC-STARR design and previously reported approaches (Wang et al. 2018; Chaudhri et al. 2020).

ATAC-STARR-seq maintains library complexity and nucleosome profiles of Tn5 selected DNA fragments

Following the experimental design outlined above, we tagmented GM12878 cells and generated an ATAC-STARR-seq plasmid library that yielded about 50 million unique accessible DNA fragments (Supplemental Text, Supplemental Fig. S1A). For a total of three replicates, we then transfected the library into GM12878 cells and harvested both reporter RNAs and plasmid DNA from the same flask of cells 24 hours later. We chose 24 hours post-transfection to avoid significant effects from the plasmid-induced interferon gene response, and to ensure the data reflects steady-state regulatory properties of GM12878 accessible regions (Supplemental Text, Supplemental Fig. S1B) (Muerdter et al. 2018). Using the captured reporter RNAs and plasmid DNA, we prepared Illumina sequencing libraries for each replicate and submitted for sequencing.

The size distribution of the accessible DNA fragments remained consistent throughout the ATAC-STARR-seq procedure and displayed the characteristic nucleosome banding and DNA pitch typified by ATAC-seq fragment libraries (Supplemental Fig. S2A-B). Analysis of library complexity between replicates revealed an average maximum complexity of 90 million unique fragments for input DNA, and 10 million unique fragments for reporter RNAs (Supplemental Fig. S2C). The difference between RNA and DNA complexities is likely due to higher duplication rates in the RNA samples (Supplemental Table S2) driven by both the

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

expression of multiple transcripts per plasmid and more PCR cycles required for the RNA samples. In addition, for both RNA and DNA samples, replicates displayed high Pearson (r^2: 0.96-0.99) and Spearman's (ρ: 0.77-0.93) correlation coefficients indicating strong agreement among the three replicates assayed (Supplemental Fig. S3). Altogether the ATAC-STARR-seq sequence libraries demonstrated the necessary quality and complexity for downstream analysis.

ATAC-STARR-seq faithfully captures chromatin accessibility with high signal-to-noise

The use of Tn5 on native chromatin to selectively clone chromatin accessible DNA fragments provides the opportunity to quantify not only reporter activity, but also chromatin accessibility simultaneously from the same plasmid library. This is because the same DNA fragments sequenced in a typical ATAC-seq workflow are contained in the ATAC-STARR-seq plasmids. Given the insert fragments from reisolated plasmids are sequenced, we asked if the resulting peak profiles recapitulate native ATAC-seq to measure chromatin accessibility. This is important because, in contrast to a typical ATAC-seq procedure, ATAC-STARR-seq involves several additional steps including cloning, transfection and reisolation, which could distort the content of the library such that it no longer represents its native profile in the genome. Specifically, mapped sequence reads derived from inserts of reisolated plasmids are counted at a given locus and this estimate infers the accessibility of the region at the time of tagmentation. This also reflects the number of plasmids that represent a given region within the reisolated ATAC-STARR-seq plasmid library. To test this, we processed the reisolated plasmid DNA as an Omni-ATAC-seq dataset and benchmarked against the GM12878 Omni-ATAC-seq dataset from Corces *et al*. 2017. Raw sequences obtained for both datasets were processed through identical workflows (see Methods). After collapsing read duplicates, we called peaks for each dataset using a variety of false-discovery rates (FDRs, Supplemental Table S3). To closely match the number of peaks previously reported by Corces *et al*. 2017 (~108,433), we chose two separate FDR thresholds—0.0001 for

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

ATAC-STARR-seq and 0.001 for the Corces data—yielding 101,904 and 89,829 accessible chromatin peaks respectively (Corces et al. 2017). The ATAC-STARR-seq and Corces *et al*. peak sets represent 2.22% and 2.11% of the genome, respectively, which agrees with previous reports (Figure 2A, (Thurman et al. 2012; Klemm et al. 2019)). Overall, 71% of ATAC-STARR-seq peaks are reproduced in the Corces *et al*. dataset, while 81% of Corces *et al*. peaks overlap the ATAC-STARR-seq dataset (Figure 2B; Jaccard index = 0.589), indicating strong agreement between these data despite substantial differences in ATAC-STARR DNA sample preparation. Furthermore, the fraction of reads in peaks score (FRiP), an ENCODE ATAC-seq standard measure of noise, is considerably higher for both ATAC-STARR-seq (0.74) and Corces *et al*. (0.526) than the ENCODE accepted standard (>0.2, Figure 2C), indicating minimal background in our dataset. The high signal-to-noise is also evident when looking at normalized read pileups at a representative locus (Figure 2D), where the signal mirrors the Corces *et al*. accessibility signal patterns. Based on these results, we conclude that ATAC-STARR-seq can accurately retain chromatin accessible peaks in the human genome with high signal-to-noise.

A sliding windows approach increases activity region calling sensitivity

ATAC-STARR-seq tests regulatory activity in DNA enriched for accessible chromatin. Unlike whole genome STARR-seq or other MPRAs, where the genomic DNA fragment distribution is relatively constant, read coverage varies substantially from peak-to-peak in ATAC-STARR-seq. In this way, ATAC-STARR-seq requires an analysis strategy that calls active and silent regulatory regions within accessibility peaks. To address this “peaks-within-peaks” problem, we developed an analytical approach using DESeq2 to normalize reporter RNA read counts to reisolated plasmid DNA read counts. DESeq2 additionally performs an independent filtering step which removes low count data confounders that can influence ratios and result in false positive peak calls (Love et al. 2014).

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

We tested two different approaches for regulatory activity analysis. The two approaches differ in how genomic regions are defined prior to differential analysis with DESeq2. Our "sliding window" method, defines regions by slicing accessible peaks into 50bp sliding bins with a 10bp step size (Figure 3A). Alternatively, the "fragment group" method, which is the approach used in Wang *et al*. 2018, synthesizes regions by grouping paired-end sequencing fragments by 75% or greater overlap (Supplemental Fig. S4A). Using a different set of genomic regions, both methods assign and count overlapping RNA and DNA reads to each genomic region and, using DESeq2, identify regions where the RNA count is statistically different from the DNA count at a Benjamini-Hochberg (BH) adjusted p-value < 0.1. The "sliding window" method yielded ~30,000 distinct active regions, while the "fragment groups" method yielded ~20,000 distinct active regions (Supplemental Fig. S4B). In addition, nearly all active regions defined using the fragment group method (95%) are also captured in the sliding window method regions (Supplemental Fig. S4C). Given this overlap and a 50% greater recovery with the sliding windows approach, we used the sliding windows method to call active ATAC-STARR-seq regulatory regions.

Because significance is the primary threshold in our region calling strategy, we examined the influence of replicate count on the number of active regions called (Supplemental Text, Supplemental Fig. S5). We found that, as expected, more replicates result in more active regions. However, we caution that these additional regions may represent a disproportionate number of false positives and may affect the outcomes of certain accuracy-sensitive applications like computational modelling. In this way, we believe three replicates are sufficient for most purposes. We also investigated the impact of read duplicates on replicate correlations and activity region calling, finding that duplicate removal substantially hindered region calling sensitivity despite yielding higher correlation coefficients between replicates (described in more detail in Supplemental Text, Supplemental Fig. S6).

Both short and long DNA fragments are required for comprehensive region calling

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

Because DNA fragment synthesis for MPRAs is limited to 200 bp including the adapters and barcode, a significant advantage of ATAC-STARR-seq and other capture-based MPRAs is the ability to measure activity of longer DNA sequences (Santiago-Algarra et al. 2017). To investigate the effect of fragment length on regulatory region calls, we divided mapped reads into short (>125bp) and long (<125bp) fragments and independently called active and silent regulatory regions; 125bp was chosen as it bisects the bimodal peak distribution displayed by RNA and DNA libraries (Supplemental Fig. S2B). Overall, read counts were similar for each sample after splitting into short and long groups (Supplemental Fig. S7A). Two to three times as many active and silent regions were called in the long fragment group compared to the short group (20,833 versus 10,789 for active and 16,872 versus 6,213 for silent). Nonetheless, a substantial number of regions are called within the short fragment group, although both fell short of the number of active and silent regions called when both long and short were used (Supplemental Fig. S7B). The regulatory regions called using long DNA fragments are larger than those called with short fragments, as expected (Supplemental Fig. S7C); however, they display little difference in TSS distance, indicating these groups are not comprised of different genomic annotations (Supplemental Fig. S7D). A critical observation is that only 23% of active regions called using short reads overlap active regions called using longer reads, revealing the two groups identify different regulatory regions in the genome (Supplemental Fig. S7E); this is also true for the silent regulatory regions, although to a lesser extent. Altogether this analysis reveals that short and long DNA fragments identify different regulatory region sets both in number and similarity. Therefore, to be as comprehensive as possible, STARR-seq assays should be designed to include both short and long DNA fragments rather than impose a size selection to remove smaller fragments.

ATAC-STARR-seq quantifies regulatory activity of open chromatin

In the sliding window approach, bins are classified as active or silent depending on whether RNA is enriched or depleted, respectively, and then like-bins are merged to collapse

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

overlaps (Figure 3A). Using this approach, we identified ~590,000 bins where RNA and DNA counts were significantly different (Figure 3B). More specifically, this analysis identified 251,895 (4.1%) active bins and 339,737 (5.5%) silent bins from the ~5.6 million total bins measured (Figure 3C). Overlapping bins were merged into 30,078 active and 21,125 silent regulatory regions (Figure 3D). It is important to note that more silent than active bins are called; however, because silent regions are generally larger (Figure 3E), merging overlapping bins results in fewer silent regions than active. Collectively, the active and silent bins represent ~9.5% of all bins measured, indicating that the majority of accessible DNA is transcriptionally neutral. Moreover, most accessible peaks do not have an active or silent region contained within them (69.5%), suggesting that most accessible regions are neutral regulatory regions according to our assay (Figure 3F). This suggests that the majority of accessible DNA has no regulatory potential in this cellular context or, alternatively, that ATAC-STARR-seq is not sensitive enough to measure weakly active or weakly silent regions. A recent study in mouse embryonic stem cells made the same observation using an orthogonal approach, suggesting this phenomenon is present in other mammalian species (Glaser et al. 2021). We note that a small percentage of accessible peaks (4.4%) contain both active and silent regions, demonstrating that there can be competing regulatory regions within the same accessible peak.

Active and silent ATAC-STARR-seq regions represent both proximal and distal *cis*-regulatory elements and lie within functional chromatin states

To gain insight into the regulatory features of active regions, we annotated both active and silent regions according to genomic location. Active regions are found in both promoter proximal and distal areas of the genome, with a majority occurring in intronic and intergenic sites (~55%), whereas silent regions coincide primarily with promoters (~75%) (Figure 4A). Functional classification of active and silent regions by the 18-state ChromHMM model (Roadmap Epigenomics Consortium et al. 2015) revealed that active regions consist of TSS

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

active, TSS flanking upstream, and Enhancer Active 1 chromatin states and are devoid of repressive states like Repressed Polycomb Weak and Quiescent (Figure 4B). By contrast, silent regions are slightly enriched for bivalent chromatin states (TSSBiv, EnhBiv), consistent with the observation that they are accessible but not active. Most silent regions also coincide with TSS Active and TSSFlnk ChromHMM states, which corroborates their promoter proximal locations; however, their designation as “active” by ChromHMM is somewhat puzzling considering these DNA fragments do not drive transcription in our assay. One explanation is that silent regulatory activity, as measured by episomal-based reporter assays, does not fully copy regulatory activity as predicted by ChromHMM. Alternatively, active promoters may confound the reporter assay by initiating transcription from the 3’UTR of the plasmid causing conflicts with active transcription from the Ori.

To further investigate if silent regions are a result of 3’UTR transcription initiation, we considered if an orientation bias existed in reporter RNAs levels. If 3’UTR transcription conflicts exist, we would expect many fewer reporter RNAs when transcription results in head-on conflicts rather than occurring in the same direction as the Ori. We therefore subset reads based on whether they arose from an insert cloned in a 3’ to 5’ direction or in a 5’ to 3’ direction (Supplemental Fig. S8A). We then assigned read counts to all bins analyzed (Supplemental Fig. S8B-C), the bins called active (Supplemental Fig. S8D-E), or the bins called silent (Supplemental Fig. S8F-G). Because this is expected to be a promoter-specific effect, we also split bins into proximal and distal based on location to the nearest transcription start site. In all cases, more than 95% of the bins do not display an orientation bias, which we defined as a normalized read count difference greater than five between orientations (Supplemental Methods, Supplemental Fig. S8H). Moreover, we observe high Pearson and Spearman’s correlation coefficients between orientations for all conditions (r^2: 0.80-0.91 and ρ: 0.73-0.90) and the minimal contribution of orientation bias to silent regions is in agreement with a previous report (Klein et al. 2020). For the <5% of regions that do display orientation bias, proximal bins are more affected than distal bins, as expected.

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

Altogether, ATAC-STARR-seq does not display a significant orientation bias and most of the 21,000 silent regions we observe result from legitimate silencing activity or another source.

Active and silent ATAC-STARR-seq regions are distinct functional classes and are enriched for specific histone modifications and TF motifs

To further investigate the chromatin landscape of the active and silent regions, we plotted ENCODE GM12878 ChIP-seq signal (The ENCODE Project Consortium et al. 2020) for EP300, CTCF, and histone modifications associated with active and repressed chromatin states (Figure 4C). As expected, active regions contain EP300 at their center with histone 3 lysine 27 acetylation (H3K27ac) more broadly distributed across the center; histone 3 lysine 4 mono-methylation (H3K4me1) is also present at distal regions, while histone 3 lysine 4 tri-methylation (H3K4me3) is at proximal regions. In addition, histone 3 lysine 27 tri-methylation (H3K27me3)—a bivalent repressive mark—is largely absent from active regions. Proximal silent regions, on the other hand, are enriched for H3K27me3 and H3K4me3. This suggests many of the proximal silent regions are accessible bivalent regulatory elements in lymphoblastoid cells. To support their designation as silent calls, we compared histone modification signal at accessible peaks that contain either a silent region, an active region, both a silent and active region, or neither, which we define as neutral accessible peaks (Supplemental Fig. S9A). Consistent with the observations above, silent accessible peaks contain more H3K27me3 signal and are devoid of H3K27ac signal relative to the other accessible peak types.

It is important to note that silent regions are distinct from neutral regions, which are defined as regions failing to reach significance in the RNA-DNA differential analysis. Overall, neutral regions exhibit baseline levels of histone modifications and distribution in genomic annotations like that of all accessible peaks (Supplemental Fig. S9B-C, Figure S4A-B). While neutral regions represent the majority of accessible peaks, it is possible that a subset are

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

weak enhancers as indicated by overlap with ChromHMM states, or regulatory elements that display activity in a different cellular context.

Our analysis of TF motifs within active and silent regions revealed prominent differences in motif enrichment. Distal silent regions are strongly enriched for CTCF and its counterpart BORIS, which is associated with diverse functions including gene repression and insulator activity (Figure 4D-E)(Kim et al. 2015). In addition, we found enrichment for the SP/KLF family several of which are known to be transcriptional repressors (Cao et al. 2010). By contrast, the most enriched TFs in active regions were the IRF family, the ETS family, subunits of the NF-kB complex, and general promoter TFs such as THAP11 and YY1. These data are consistent with our current understanding of immune gene regulation and regulatory element function, which together corroborates the quantification of regulatory activity with ATAC-STARR-seq.

ATAC-STARR-seq retains the ability to map *in vivo* TF binding

An inherent advantage of an ATAC-seq based approach is the ability to perform TF footprinting. Computational footprinting methods identify Tn5 cleavage events or “cut sites” from ATAC-seq data and, when combined with motif analysis, can identify TF binding sites with high accuracy (Bentsen et al. 2020; Yan et al. 2020). Since ATAC-STARR-seq produces similar high-quality chromatin accessibility peak profiles as standard ATAC-seq, we explored whether TF footprints were also preserved. We generated Tn5-bias corrected cut site signal files for both Corces *et al*. 2017 and ATAC-STARR-seq accessibility datasets and plotted cut site signal at all accessible CTCF motifs (Figure 5A) (Bentsen et al. 2020). As a control, we also plotted GM12878 CTCF ChIP-seq signal from ENCODE and ranked region order by highest mean ChIP-seq signal. We observed consistent depletion of Tn5 cut-sites for both Corces *et al*. 2017 and ATAC-STARR-seq accessibility at CTCF sites. Moreover, we only observe footprints at motifs with CTCF ChIP-signal, demonstrating the utility of TF footprinting to determine motifs that are bound or unbound by TFs. Given the

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

importance of TFs in driving enhancer function, this distinction is vital when dissecting transcriptional regulation in human cells.

TF motif enrichment analysis pointed to multiple ETS family members, including ETS1 which is an important immune cell regulator (Garrett-Sinha 2013) (Figure 4D). So, we asked whether ETS1 footprints are also present in our data. Unlike CTCF, ETS1 shares its motif with many other transcription factors, such as ETV4; therefore, footprinting cannot distinguish ETS1 and ETV4 binding sites. For this reason, we refer to TFs using their ENCODE-defined “archetypes”, which reflects the group of TFs that share the same motif (Vierstra et al. 2020). For each archetype, we performed footprinting against one of the TFs within an archetype to infer motifs bound by members of the group, such as ETS1 for the ETS/1 archetype. To assess the extent to which ETS1 footprints can be explained by ETS1 binding, we plotted GM12878 ETS1 ChIP-seq signal from ENCODE within both Corces *et al*. 2017 and ATAC-STARR-seq cut sites (Figure 5B). Indeed, ETS1 ChIP-seq signal explains the majority but not all the ETS/1 footprints present. We observe similar cut-site signal to Corces *et al*. 2017, further indicating that ATAC-STARR-seq can detect *in vivo* binding of transcription factors despite the additional cloning and transfection steps involved in producing ATAC-STARR-seq DNA libraries.

We performed footprinting for several more immune related TF archetypes to identify bound or unbound TF motifs (Figure 5C). For all TFs, bound motifs display substantially larger footprint depth than unbound motifs. Together, this indicates that ATAC-STARR-seq, when combined with footprinting, can identify regions of the genome where TFs are bound. This additional level of information can be leveraged in conjunction with accessibility and activity to understand the context of TF binding while circumventing the need to perform individual chromatin immunoprecipitations.

Collective profiling of accessibility, *in vivo* TF binding, and activity with ATAC-STARR-seq reveals distinct networks of gene regulation

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

Interrogating chromatin accessibility, TF binding, and regulatory activity together can be used to interpret locus-specific gene regulatory mechanisms. For example, active regulatory elements surrounding the B cell-specific expressed gene *ZBTB32* overlap IRF8 and NFKB1 footprints suggesting these regions are regulated by IRF8 and NFKB1 binding (Figure 6A). We also observe SP1 and KLF3 footprints overlapping a silent region at the *ETV2* locus, a gene lowly expressed in B cells, according to the Human Protein Atlas (Uhlen et al. 2015; Uhlen et al. 2019). Together this indicates that active and silent regions can, in part, be explained by the occupancy of these TFs.

To demonstrate the power of integrating TF footprints and regulatory regions on a global scale, we clustered active and silent regions based on the presence or absence of several TF footprints (Figure 6B-C). Footprints were selected based on top hits from the previous motif enrichment analysis (Figure 4D-E). Regulatory activity may be driven by one or multiple TF binding events that defines the cluster and is representative of a gene regulatory network in the genome. Indeed, we find that the putative target genes regulated by each unique group are enriched for distinct gene regulatory pathways and are often related to the TFs in the cluster (Figure 6D-E). For example, cluster C is primarily defined by the presence of IRF/1 and is enriched for interferon alpha/beta signalling. It is interesting that active clusters tend to be more associated with B cell function than silent clusters, which are more associated with general, non-B cell related pathways.

Altogether, these distinct gene regulatory networks provide an additional layer of insight into the mechanisms that control gene expression and showcase how integration of the multiple layers of gene regulatory information provided by ATAC-STARR-seq can narrow the focus of gene targets for active and silent regions. We envision such an analysis could be used to interpret the functional consequences of a dysregulated transcription factor or disease-associated genetic variants. We provide this level of detail from a single dataset, which further demonstrates the strong potential of our workflow to reveal distinct functional layers

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

of human gene regulation. The resolution we achieve here would not be possible without all three levels of regulatory information provided by ATAC-STARR-seq.

DISCUSSION

Genome-wide approaches that integrate measurements of multiple layers of gene regulation are needed to better understand enhancer function. By combining ATAC-seq with STARR-seq, ATAC-STARR-seq assays regulatory activity only within the context of accessible chromatin. This allows deeper coverage of regulatory elements by narrowing scope but remaining inclusive of nearly all active regulatory elements. In this report, we substantially expand the capabilities of ATAC-STARR-seq and present an improved workflow which uniquely permits simultaneous profiling of accessibility, TF occupancy, and regulatory activity from a single DNA fragment source. Specifically, we implement key experimental and analytical improvements and present data rationalizing the decisions we make. Experimentally, we adapt a modified tagmentation protocol (Omni-ATAC) to remove mitochondrial DNA from the DNA fragment pool. We also utilize the Ori as the minimal promoter on the STARR-seq backbone which improves reporter RNA expression, recovery, and dynamic range over the super core promoter (SCP1) backbone (Muerdter et al. 2018; Klein et al. 2020). Furthermore, we reisolate the transfected plasmid DNA to capture only the DNA that is available to cells, which is a more accurate measure of the input than sequencing prior to transfection. Reisolating plasmid DNA drives a greater degree of variance between samples and better reflects a true experimental replicate than sequencing the same DNA sample for each RNA replicate. Finally, we show that replicate number and inclusion of long and short fragment sizes are critical for comprehensive region calling.

Critically, we developed and tested a simple and sensitive region calling strategy that improves detection of regulatory regions including silencers. We also quantify chromatin accessibility and identify TF footprints, which is surprising given the added processing steps in ATAC-STARR-seq including cloning, transfection, and recapture of DNA libraries that can

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

dull or degrade footprint signal. This enabled us to stratify the active and silent regulatory regions into distinct gene regulatory networks defined by the presence of one or multiple TF footprints. Such an analysis typically requires multiple genomic sequencing assays, but we do this using a single dataset.

With this improved workflow, we identified 30,078 active regions and 21,125 silent regions in lymphoblastoid cells. Most active regions were distal to transcription start sites, enriched for functional active ChromHMM states, and were enriched for known B cell regulating-TF motifs such as IRF8 and NFkB. By contrast, the silencers are proximal to transcription start sites and enriched for CTCF and the SP/KLF TF family. Silent regions are also enriched for the bivalent marks H3K27me3 and H3K4me3 and may represent regulatory regions that are poised, particularly at promoters. Because our plasmid design places regulatory regions within the 3'UTR of the truncated reporter gene, it is possible that the lack of observed reporter RNAs at silent regions are a result of head-on transcriptional conflicts that arise from antisense transcription initiation from the 3'UTR. However, we show this minimally occurs in our system and the silent regions reflect true silencing activity or another source that has yet to be identified. While further studies may be needed to validate these silent regions, this work confirms that the silencers are a distinct class of regulatory element with distinct properties compared to active and neutral regions and warrant further investigation. Even with an increasing number of studies targeted at identifying silencers in the human genome, silencing regulatory regions remain an under-studied aspect of gene regulation and our approach provides a new strategy to investigate these elements on a global scale (Doni Jayavelu et al. 2020; Pang and Snyder 2020; Kim et al. 2021).

ATAC-STARR-seq data has several distinct attributes that require a tailored analysis strategy. Current MPRA bioinformatic tools and pipelines are not tractable for these data, because in ATAC-STARR-seq the input itself is enriched for accessible chromatin and the read pileup varies considerably within these loci. In this way, the analysis of our data

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

451 required calling essentially “peaks within peaks”. For this reason, it was critical to 1)
452 normalize RNA to DNA and 2) avoid regions of low count data, which is why we adapted
453 approaches using DESeq2. We also showed that including PCR duplicates was preferred
454 over collapsing duplicates. In the future it would be beneficial to introduce a unique
455 molecular identifier to the system—such as the strategy employed by UMI-STARR-seq
456 (Neumayr et al. 2019)—to collapse only the duplicates arising from PCR. While we show
457 comparisons of analysis strategies here, we believe that more information could be extracted
458 from this and future ATAC-STARR-seq datasets with improved analysis strategies. In recent
459 years we have seen the development of tailormade peak callers for whole genome STARR-
460 seq, such as CRADLE (Kim et al. 2021) and STARRPeaker (Lee et al. 2020); a similarly
461 tailored ATAC-STARR-seq peak caller could further improve the capabilities of the method.

462 While this study was limited to one condition, there are many potential applications of ATAC-
463 STARR-seq. With the ability to subset enhancers by TF occupancy, ATAC-STARR-seq
464 could be leveraged to investigate enhancer grammar by pairing measurable regulatory
465 activity with multiple TF footprints that drive enhancer function. This approach also has the
466 potential to identify dysfunctional gene regulatory networks in diseases like cancer where
467 neoplastic transformation can be driven by the dysfunction of a specific TF. Additionally, an
468 ATAC-STARR-seq plasmid library may be generated from one cell-type and tested in
469 another. This flexibility could be used as a tool to determine context dependent activity or
470 investigate enhancer and TF usage patterns during a differentiation time course.

471 In this study, we demonstrated that our improved ATAC-STARR-seq workflow is a powerful
472 approach enabling joint quantification of chromatin accessibility, transcription factor
473 occupancy, and regulatory activity. We further demonstrate how this single assay can
474 characterize the human genome at many functional levels from chromatin accessibility to
475 distinct gene regulatory networks. This method provides a state-of-the-art approach to
476 deeply investigate transcriptional regulation of the human genome. We provide a detailed

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

protocol and a well-documented code repository so that ATAC-STARR-seq may be easily used and adapted by the field.

MATERIAL AND METHODS

Cell Culture

GM12878 cells were obtained from Coriell and cultured with RPMI 1640 Media containing 15% fetal bovine serum, 2mM GlutaMAX, 100 units/mL penicillin and 100 µg/mL streptomycin. Cells were cultured at 37°C, 80% relative humidity, and 5% CO_2. Cell density was maintained between 0.2×10^6 and 1×10^6 cells/mL with a 50% media change every 2-4 days. All cell lines were regularly screened for mycoplasma contamination using the MycoAlert kit (Lonza).

Plasmids

The hSTARR-seq_ORI plasmid vector was a gift from Alexander Stark (Addgene plasmid #99296) and the pcDNA3-EGFP was a gift from Doug Golenbock (Addgene plasmid #13031). The bacterial stabs from Addgene were spread onto an LB agar plate containing 100µg/mL ampicillin and incubated at 37°C overnight. For each, a single colony was picked and grown in 50mL LB containing 100µg/mL ampicillin overnight at 37°C while shaking at 225rpm. Plasmid DNA was extracted using the ZymoPURE II Plasmid Midiprep kit (Zymo Research, #D4200).

The linear vector used in the ATAC-STARR-seq gibson cloning step was generated by a single 50µL PCR reaction using NEBNext® Ultra™ II Q5® Master Mix (NEB, #M0544S). While not necessary for this study, primers were designed to add the i5 barcode to the linearized vector; this allows for different ATAC-STARR-seq plasmid libraries to be pooled and tracked. Following this approach, a universal forward primer (Fwd_universal_STARR) and a reverse primer (Rev_N504_STARR) designed to add the N504 barcode were used

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

(primer sequences are provided in Supplemental Table S4). PCR Products were purified with the Zymo Research DNA Clean & Concentrator-5 kit. DNA yield was determined by Nanodrop, and purity was analysed by gel electrophoresis; the linearized vector was the only product observed on the gel.

Tagmentation

A total of eight tagmentation reactions were performed on 50,000 GM12878 cells each. We followed a slightly modified version of the Omni-ATAC approach used in Corces *et al.* 2017 (Corces et al. 2017). Specifically, twice as much Tn5 than described in the protocol was used. Standard Tn5 transposase was prepared in-house following the method described in Picelli *et al.* 2014 (Picelli et al. 2014). Standard Tn5 transposome was assembled as described in Barnett *et al.* 2020 (Barnett et al. 2020) with the following oligos: Tn5_1, Tn5_2_ME_comp, and TN5MEREV. Tagmented products were pooled together and purified with the Zymo Research DNA Clean & Concentrator-5 kit (#D4013). The entire elution was split and amplified via five-10µL PCR reactions. We used NEBNext® High-Fidelity 2× PCR Master Mix (#M0541S)—which is not a hot-start formulation—to first extend tagments before the initial denaturation step of PCR via the following cycling parameters: 72°C 5 min, 98°C 30s; 4 cycles of 98°C 10s, 62°C 30s, 72°C 60s; final extension 72°C 2 min; hold at 10°C. Forward and reverse primer sequences, Fwd_atac-starr_tag and Rev_atac-starr_tag, are provided in Supplemental Table S3. Amplified products were purified with the Zymo Research DNA Clean & Concentrator-5 kit and then analyzed for concentration and size distribution with a HSD5000 screentape (Agilent, #5067) on an Agilent 4150 TapeStation system. After amplification, we selected PCR products less than 500bp using SPRISelect beads (Beckman-Coulter, #B23317) at a 0.6× volume ratio of beads:sample. Selection was verified using a HSD5000 screentape.

Massively Parallel Cloning

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

Four 10µL gibson cloning reactions were performed with NEBuilder® HiFi DNA Assembly Master Mix at a vector:insert molar ratio of 1:2. As a negative control, we performed one cloning reaction substituting tagments with nuclease-free water. Gibson products were pooled and purified via ethanol precipitation as previously described in Sambrook & Russell (Sambrook and Russell 2006); we used glycoblue (150µg/mL) as a co-precipitant. Purified gibson products were electroporated into MegaX DH10B T1R Electrocomp™ Cells (Invitrogen, # C640003) using a Bio-Rad Gene Pulser. Three electroporations for the ATAC-STARR-seq sample (and 1 for the control) were performed with the following parameters: exponential decay pulse type, 2kV, 200Ω, 25µF, and 0.1cm gap distance. Pre-warmed SOC media (1mL) was added immediately following electroporation; the three reactions were pooled and incubated at 37°C for 1 hour. We confirmed cloning success by plating a dilution series—using a small aliquot from the ATAC-STARR-seq and negative control samples—onto pre-warmed LB agar plates containing 100µg/mL ampicillin and visualizing colonies 24 hours later. The remaining ATAC-STARR-seq transformation was added directly to a 1L LB liquid culture with 100µg/mL ampicillin and grown at 37°C while shaking at 225rpm overnight. The next day, plasmid DNA was harvested from the 1L culture using the ZymoPURE II Plasmid Gigaprep (Zymo Research, #D4204). Before prepping, we recorded a 1.633 optical density.

Electroporation

GM12878 cells were cultured so that cell density was between 400,000 and 800,000 cells/mL on day of transfection. Three replicates were performed on separate days. For each replicate, a total of 20 electroporation reactions was performed using the Neon™ Transfection System 100 µL Kit (Invitrogen, #MPK10025) and the associated Neon™ Transfection System (Invitrogen, #MPK5000). 121 million GM12878 cells were collected, washed with 45mL PBS, and resuspended in 2178µL Buffer R. For each reaction, 5 million cells (in 90µL Buffer R) were electroporated with 5µg of ATAC-STARR-seq plasmid DNA (in

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

552 10µL nuclease-free water) in a total volume of 100µL with the following parameters: 1100V,
553 30ms, and 2 pulses. Electroporated cells were dispensed immediately into a pre-warmed T-
554 75 flask containing 50mL of RPMI 1640 with 20% fetal bovine serum and 2mM GlutaMAX.

555 **Cell Harvest**

556 24 hours after transfection, the 50mL ATAC-STARR-seq flask was divided into two equal
557 volumes; plasmid DNA was extracted from one volume, while reporter RNAs were extracted
558 from the other. Plasmid DNA was isolated with the ZymoPURE II Plasmid Midiprep kit
559 (#D4200) and eluted in 50µL 10mM Tris-HCL pH 8.0. Prior to lysis, cells were washed with
560 25mL PBS to remove any extracellular plasmid DNA. Reporter RNAs were extracted in a
561 stepwise process. First, total RNA was isolated from the second volume of transfected cells
562 using the TRIzol™ Reagent and Phasemaker™ Tubes Complete System (Invitrogen™,
563 #A33251). Specifically, 5mL TRIzol was added to homogenize the washed and pelleted cells.
564 Next, polyadenylated RNA was isolated from total RNA using oligo(dT)25 Magnetic Beads
565 (NEB, #S1419S) at a 1µg Total RNA to 10µg beads ratio. We performed this step at 4°C and
566 eluted into 50µL10mM Tris-HCl pH 7.5. The extracted poly(A)$^+$ RNA was treated with DNase
567 I (NEB, #M0303S). This reaction was cleaned up using the Zymo Research RNA Clean &
568 Concentrator-25 kit (Zymo Research, #R1018).

569 **First-strand cDNA synthesis**

570 For each sample, ten 50µL reverse transcription reactions were carried out using
571 PrimeScript™ Reverse Transcriptase (Takara, #2680) and a gene specific primer
572 (STARR_GSP) as described by Muerdter *et al.* 2018 (Muerdter et al. 2018). Single-stranded
573 cDNA was treated with RNase A at a concentration of 20µg/mL in low salt concentrations
574 and cleaned up with a Zymo Research DNA Clean & Concentrator-5 kit.

575 **Illumina Sequencing Library Preparation**

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

For reisolated plasmid and reporter RNA samples, Illumina-compatible libraries were generated using NEBNext® Ultra™ II Q5® Master Mix and a unique combination of the following Nextera indexes: N504-N505 (i5) and N701-N702 (i7), see Supplemental Table S1 for primer sequences. DNA samples were amplified for 8 PCR cycles, while RNA was amplified for 12-13 cycles. In both cases, products were purified with the Zymo Research DNA Clean & Concentrator-5 kit and analyzed for concentration and size distribution using a HSD5000 screentape. Purified products were sequenced on an Illumina NovaSeq, PE150, at a requested read depth of 50 or 75 million reads, for DNA and RNA samples, respectively, on an Illumina NovaSeq 6000 machine through the Vanderbilt Technology for Advanced Genomics (VANTAGE) sequencing core. Reads were processed and analyzed as described in the supplemental methods. We provide guidelines for ATAC-STAR-seq quality control in the supplemental text.

DATA ACCESS

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE181317. Python scripts and additional code for ATAC-STARR-seq data analysis are available at GitHub (https://github.com/HodgesGenomicsLab/ATAC-STARR-seq) and as Supplemental Code. An interactive version of the protocol is posted on protocols.io (dx.doi.org/10.17504/protocols.io.b2nuqdew) and a pdf version of the protocol at publication date is included as a Supplemental file.

COMPETING INTEREST STATEMENT

None declared.

ACKNOWLEDGEMENTS

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

We thank Felix Muerdter, Sarah Fong, Tony Capra, and members of the Hodges Lab, especially Kelly Barnett, Tim Scott, Lindsey Guerin, Verda Agan, Elizabeth Dorans, and Ali Wilt for helpful feedback and discussions. We also thank Biorender.com for illustrations, Addgene for plasmids used in the study, the Dave Cortez Lab for use of their Bio-Rad Gene Pulser, and the Manny Ascano Lab for qPCR primers and helpful advice. We are grateful for support of the project and the time invested in producing this manuscript by the NIH awards [K22 CA184308-03 to E.H], Department of Defense Idea Award [W81XWH-20-1-0522 to E.H], and American Cancer Society (ACS) Institutional Research Grant (#IRG-15-169-56).

REFERENCES

Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**: 1074-1077.

Barnett KR, Decato BE, Scott TJ, Hansen TJ, Chen B, Attalla J, Smith AD, Hodges E. 2020. ATAC-Me Captures Prolonged DNA Methylation of Dynamic Chromatin Accessibility Loci during Cell Fate Transitions. *Mol Cell* **77**: 1350-1364 e1356.

Bentsen M, Goymann P, Schultheis H, Klee K, Petrova A, Wiegandt R, Fust A, Preussner J, Kuenne C, Braun T et al. 2020. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat Commun* **11**: 4267.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**: 1213-1218.

Cao Z, Sun X, Icli B, Wara AK, Feinberg MW. 2010. Role of Kruppel-like factors in leukocyte development, function, and disease. *Blood* **116**: 4404-4414.

Chaudhri VK, Dienger-Stambaugh K, Wu Z, Shrestha M, Singh H. 2020. Charting the cis-regulome of activated B cells by coupling structural and functional genomics. *Nat Immunol* **21**: 210-220.

Chen AF, Parks B, Kathiria AS, Ober-Reynolds B, Goronzy JJ, Greenleaf WJ. 2022. NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. *Nat Methods* **19**: 547-553.

Clark SJ, Argelaguet R, Kapourani CA, Stubbs TM, Lee HJ, Alda-Catalinas C, Krueger F, Sanguinetti G, Kelsey G, Marioni JC et al. 2018. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun* **9**: 781.

Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, Satpathy AT, Rubin AJ, Montine KS, Wu B et al. 2017. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods* **14**: 959-962.

Doni Jayavelu N, Jajodia A, Mishra A, Hawkins RD. 2020. Candidate silencer elements for the human and mouse genomes. *Nat Commun* **11**: 1061.

Garrett-Sinha LA. 2013. Review of Ets1 structure, function, and roles in immunity. *Cell Mol Life Sci* **70**: 3375-3390.

Gasperini M, Tome JM, Shendure J. 2020. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat Rev Genet* **21**: 292-310.

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

Glaser LV, Steiger M, Fuchs A, van Bommel A, Einfeldt E, Chung HR, Vingron M, Meijsing SH. 2021. Assessing genome-wide dynamic changes in enhancer activity during early mESC differentiation by FAIRE-STARR-seq. *Nucleic Acids Res* **49**: 12178-12195.

Heinz S, Romanoski CE, Benner C, Glass CK. 2015. The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol* **16**: 144-154.

Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, Ahituv N, Shendure J. 2017. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res* **27**: 38-52.

Johnson GD, Barrera A, McDowell IC, D'Ippolito AM, Majoros WH, Vockley CM, Wang X, Allen AS, Reddy TE. 2018. Human genome-wide measurement of drug-responsive regulatory activity. *Nat Commun* **9**: 5317.

Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones PA. 2012. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res* **22**: 2497-2506.

Kim S, Yu NK, Kaang BK. 2015. CTCF as a multifunctional protein in genome regulation and gene expression. *Exp Mol Med* **47**: e166.

Kim YS, Johnson GD, Seo J, Barrera A, Cowart TN, Majoros WH, Ochoa A, Allen AS, Reddy TE. 2021. Correcting signal biases and detecting regulatory elements in STARR-seq data. *Genome Res* doi:10.1101/gr.269209.120.

Kircher M, Xiong C, Martin B, Schubach M, Inoue F, Bell RJA, Costello JF, Shendure J, Ahituv N. 2019. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat Commun* **10**: 3583.

Klein JC, Agarwal V, Inoue F, Keith A, Martin B, Kircher M, Ahituv N, Shendure J. 2020. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat Methods* **17**: 1083-1091.

Klemm SL, Shipony Z, Greenleaf WJ. 2019. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* **20**: 207-220.

Lee D, Shi M, Moran J, Wall M, Zhang J, Liu J, Fitzgerald D, Kyono Y, Ma L, White KP et al. 2020. STARRPeaker: uniform processing and accurate identification of STARR-seq active regions. *Genome Biol* **21**: 298.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Maricque BB, Dougherty JD, Cohen BA. 2017. A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis-regulatory activity in neural cells. *Nucleic Acids Res* **45**: e16.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Jr., Kinney JB et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**: 271-277.

Muerdter F, Boryn LM, Woodfin AR, Neumayr C, Rath M, Zabidi MA, Pagani M, Haberle V, Kazmar T, Catarino RR et al. 2018. Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat Methods* **15**: 141-149.

Neumayr C, Pagani M, Stark A, Arnold CD. 2019. STARR-seq and UMI-STARR-seq: Assessing Enhancer Activities for Genome-Wide-, High-, and Low-Complexity Candidate Libraries. *Curr Protoc Mol Biol* **128**: e105.

Pang B, Snyder MP. 2020. Systematic identification of silencers in human cells. *Nat Genet* **52**: 254-263.

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM et al. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**: 265-270.

Picelli S, Bjorklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. 2014. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res* **24**: 2033-2040.

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317-330.

Sambrook J, Russell DW. 2006. Standard ethanol precipitation of DNA in microcentrifuge tubes. *CSH Protoc* **2006**.

Santiago-Algarra D, Dao LTM, Pradel L, Espana A, Spicuglia S. 2017. Recent advances in high-throughput approaches to dissect enhancer function. *F1000Res* **6**: 939.

The ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A et al. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**: 699-710.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75-82.

Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A et al. 2015. Tissue-based map of the human proteome. *Science* **347**: 1260419.

Uhlen M, Karlsson MJ, Zhong W, Tebani A, Pou C, Mikes J, Lakshmikanth T, Forsstrom B, Edfors F, Odeberg J et al. 2019. A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science* **366**.

Vierstra J, Lazar J, Sandstrom R, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F, Haugen E et al. 2020. Global reference mapping of human transcription factor footprints. *Nature* **583**: 729-736.

Wang X, He L, Goggin SM, Saadat A, Wang L, Sinnott-Armstrong N, Claussnitzer M, Kellis M. 2018. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat Commun* **9**: 5380.

Yan F, Powell DR, Curtis DJ, Wong NC. 2020. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol* **21**: 22.

FIGURE LEGENDS

Figure 1. Schematic of the ATAC-STARR-seq methodology. (A) The experimental design of ATAC-STARR-seq consists of three parts: plasmid library generation, reporter assay, and data analysis. Open chromatin is isolated from cells with the cut and paste transposase Tn5 and only large DNA fragments (>500bp) are removed. The open chromatin fragments are cloned into a reporter plasmid and the resulting clones—called an ATAC-STARR-seq plasmid library—are electroporated into cells. 24 hours later, both reporter RNAs (blue)—which are transcribed directly off the ATAC-STARR-seq plasmid—and ATAC-STARR-seq plasmid DNA (red) are harvested, and Illumina-sequencing libraries are prepared and sequenced. The resulting ATAC-STARR-seq sequence data is analyzed to extract regulatory activity, chromatin accessibility, and transcription factor footprints. (B) Reporter plasmid design and the expected outcomes for neutral, active, and silent regulatory

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

elements. Each ATAC-STARR-seq plasmid within a library contains a truncated GFP (trGFP) coding sequence, a poly-adenylation signal sequence, an origin of replication (Ori) (which moonlights as a minimal core promoter), and the unique open chromatin fragment being assayed. Since the accessible region is contained in the 3' UTR, the abundance of itself in the transcript pool reflects its activity. In this way, neutral elements do not affect the system and reporter RNAs are expressed at a basal expression level dictated by the minimal core promoter, the Ori. Accessible chromatin fragments that are active express reporter RNAs at a higher level than the basal expression level, while silent elements repress the Ori and reporter RNAs are expressed at a lower level than basal expression. Dashed boxes represent new components of the ATAC-STARR-seq assay design and workflow.

Figure 2. ATAC-STARR-seq accurately quantifies chromatin accessibility. ATAC-seq data from Corces et al. 2017 is compared with ATAC-STARR-seq plasmid DNA data. (A) Fraction of the human genome represented by each peak set. (B) Venn diagram of peak overlap between the two datasets and the associated Jaccard Index. (C) Fraction of paired-end (PE) fragments in peaks—FRiP scores—for both samples. (D) Signal tracks comparing counts per million (CPM) normalized read count at a representative locus.

Figure 3. ATAC-STARR-seq quantifies regulatory activity within accessible chromatin. (A) Schematic of the sliding window peak calling method. Accessibility peaks are chopped into 50bp bins at a 10bp step size with the BEDTools makewindows function (options -w 50, -s 10). For each window, RNA and DNA reads are counted using Subread's featureCounts function. Differential analysis comparing RNA and DNA read count is performed with DESeq2. Significant bins are called at an Benjamini-Hochberg (BH) adjusted p-value < 0.1 and parsed into active or silent depending on $\log_2$ fold-change (FC) value (+/- zero). Finally, bins are collapsed into regions using the BEDTools merge function. Log_2FC scores are averaged across merged bins. (B) Volcano plot of $\log_2$FC scores against $-\log_{10}$-transformed BH adjusted p-value from DESeq2 for all bins analyzed. (C) The proportion of bins called as active or silent. (D) The number of regions defined as either active or silent. (E) Overlapping

Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

density plots of active and silent regulatory region size; dashed lines represent the medians in each case. (F) The proportion of accessible peaks that overlap an active or silent region, or both.

Figure 4. Regulatory regions defined by ATAC-STARR exhibit annotations, histone modifications, and TFs characteristic of their function. (A) Annotation of regulatory regions relative to the transcriptional start site (TSS). The promoter is defined as 2kb upstream and 1kb down-stream of the TSS. (B) Annotation of regulatory regions by the ChromHMM 18-state model for GM12878 cells. (C) Heatmaps of GM12878 ENCODE ChIP-seq signal and regulatory activity for proximal and distal ATAC-STARR-defined regulatory regions. Proximal regions were classified as within 2kb upstream and 1kb downstream of a TSS; all other regions were annotated as distal. Active and silent regions were ranked by mean activity signal for both proximal and distal regions. (D-E) Transcription factor motif enrichment analysis as quantified by HOMER. Fold-change values are relative to the default background calculated by HOMER.

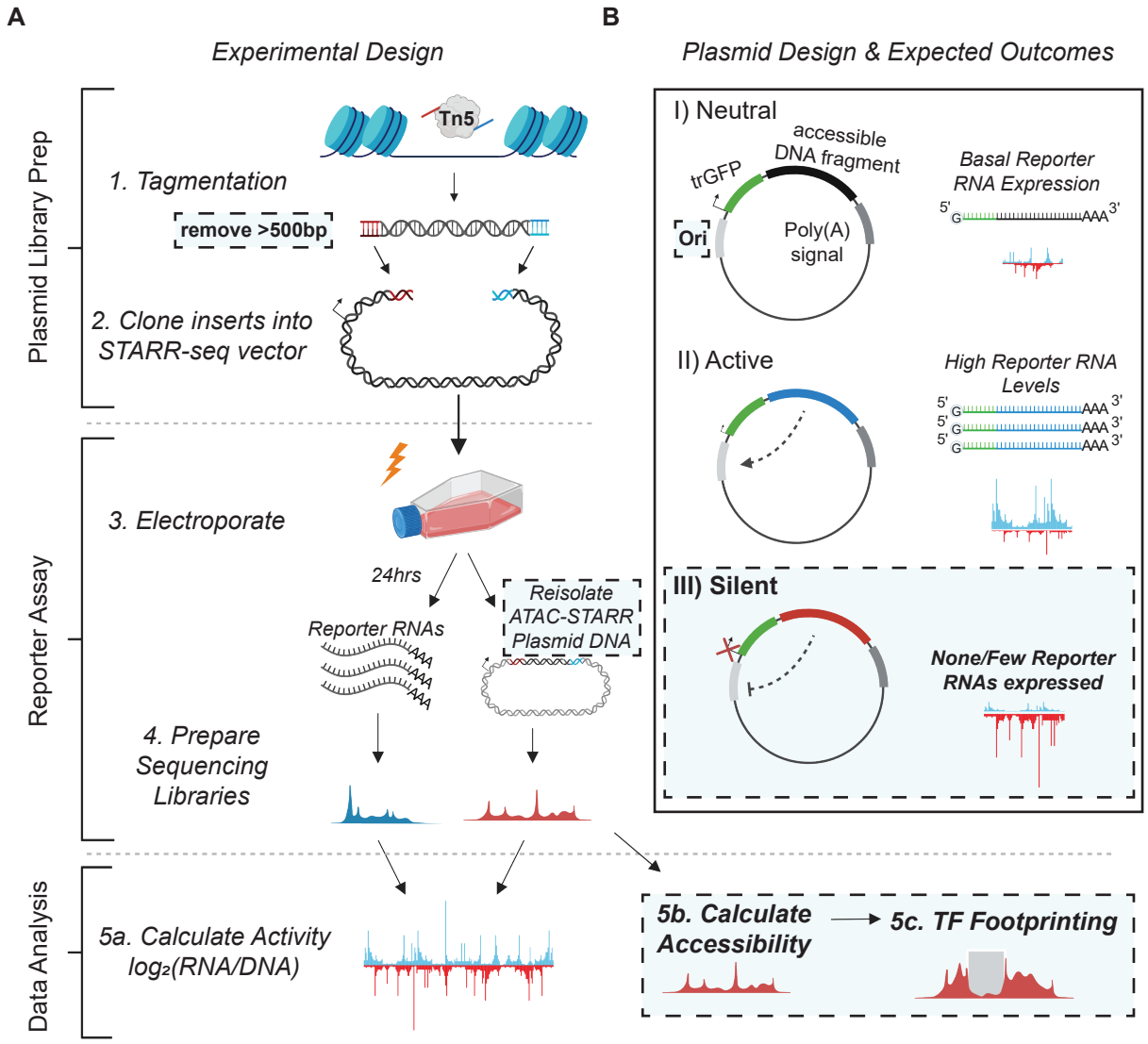
Figure 5. ATAC-STARR-seq identifies transcription factor footprints. (A) Comparison of ENCODE CTCF ChIP-seq signal to Corces et al. and ATAC-STARR-seq cut count signal for all accessible CTCF motifs. (B) Comparison of ENCODE ETS1 ChIP-seq signal to Corces et al. and ATAC-STARR-seq cut count signal for all accessible motifs with the ETS/1 motif archetype. For both, regions were ranked by largest mean ChIP-seq signal. (C) Aggregate plots representing mean signal for the TOBIAS-defined bound and unbound motif archetypes: CTCF, ETS/1, CREB/ATF/1, IRF/1, SPI, NFKB/2.

Figure 6. TF footprints stratify ATAC-STARR-defined regulatory regions into gene regulatory networks. (A) ATAC-STARR-defined chromatin accessibility, TF footprints, and regulatory regions at Chr19:35,611,232-35,798,446 (hg38). Signal tracks represent counts per million normalized read depth of chromatin accessibility. Zooms into *ETV2* and *ZBTB32* show that some regulatory regions are occupied by a SP1, KLF3, IRF8, or NFKB1 footprint.

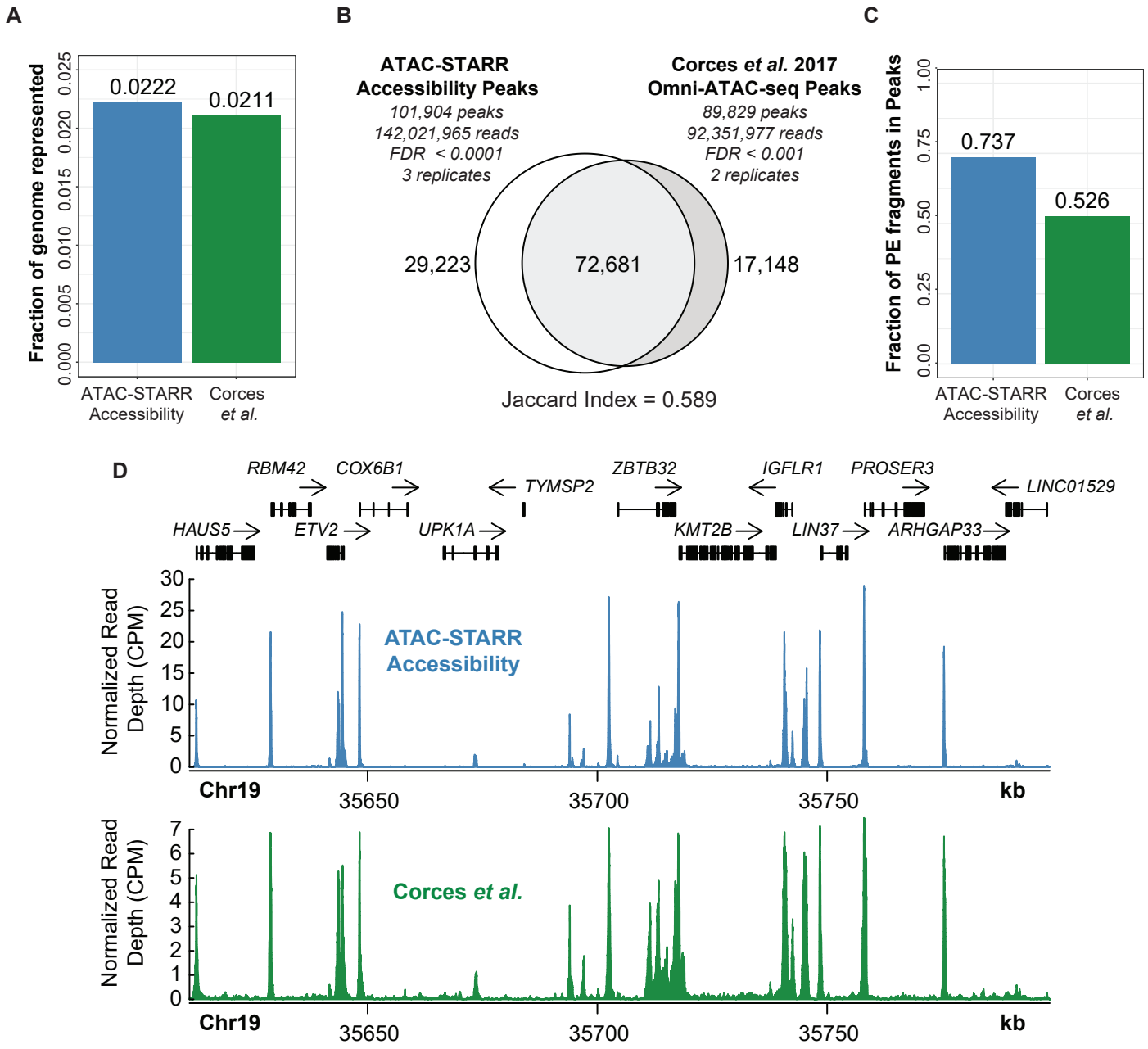
Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

(B-C) Heatmaps of clustered (B) active and (C) silent regions based on presence or absence of footprints for select TF motif archetypes. (D-E) Reactome pathway enrichment analysis for nearest-neighbor gene sets for each of the clusters. Genes counts for each cluster are displayed below their group identifier.

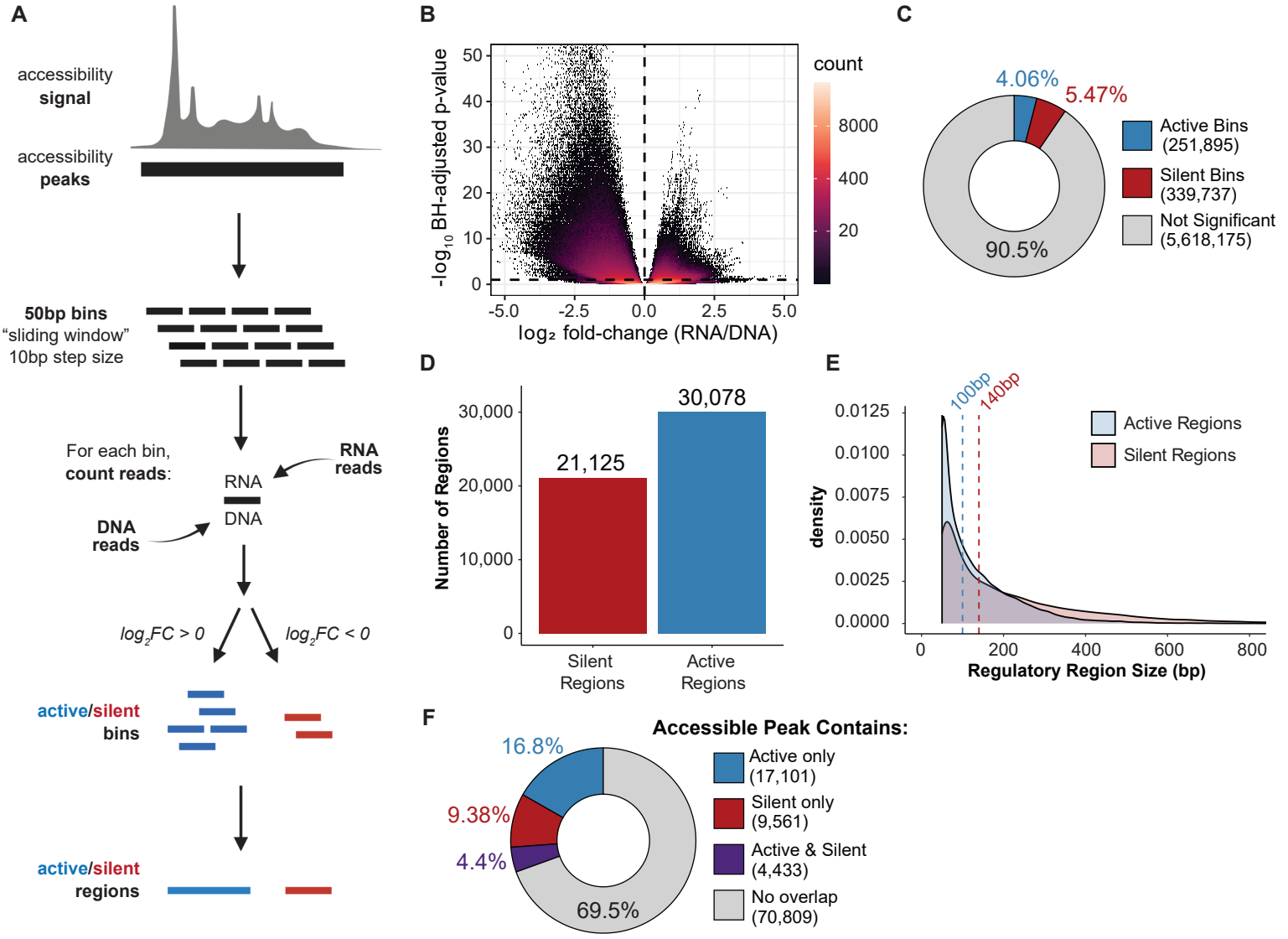
Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press



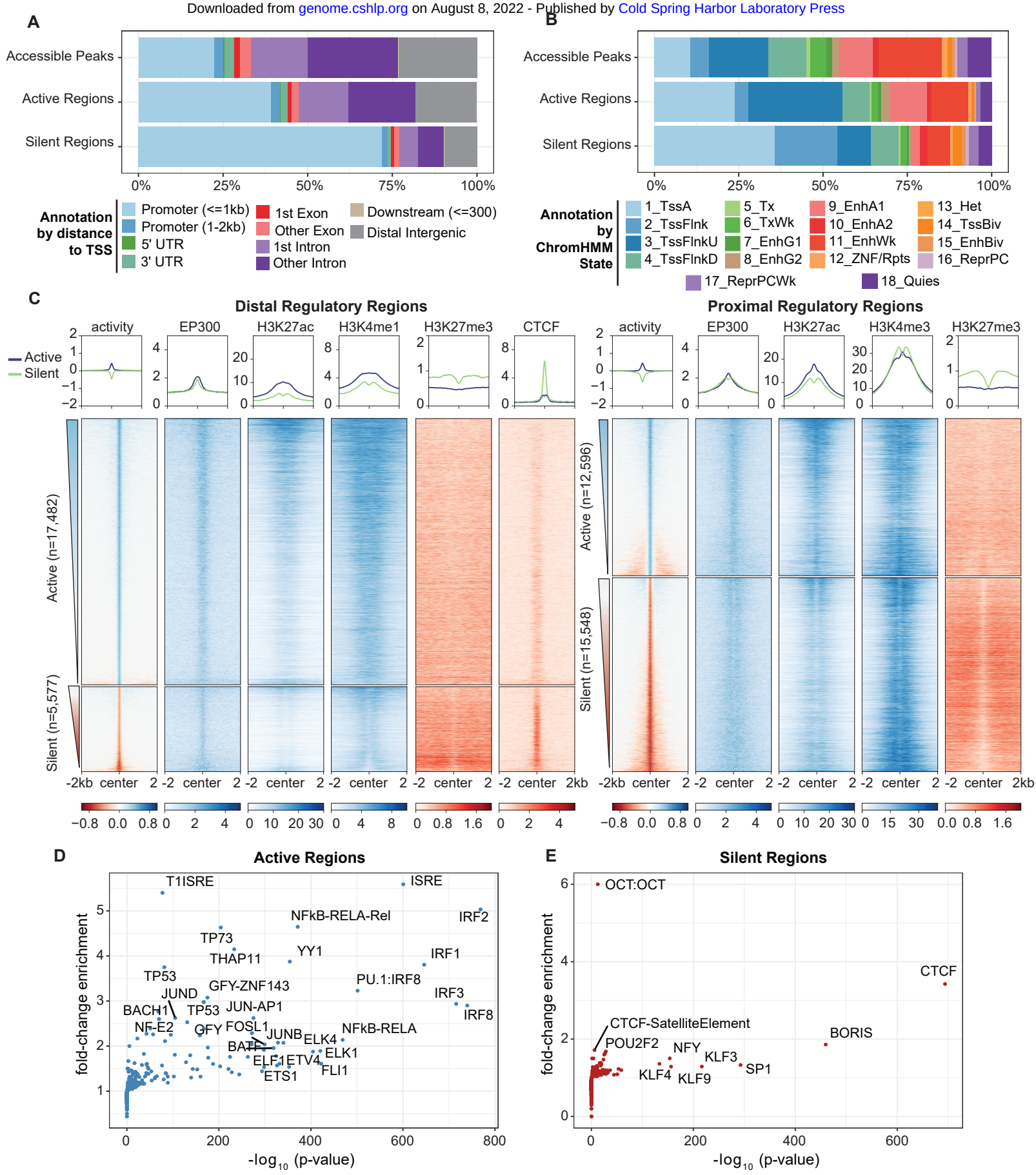
Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press



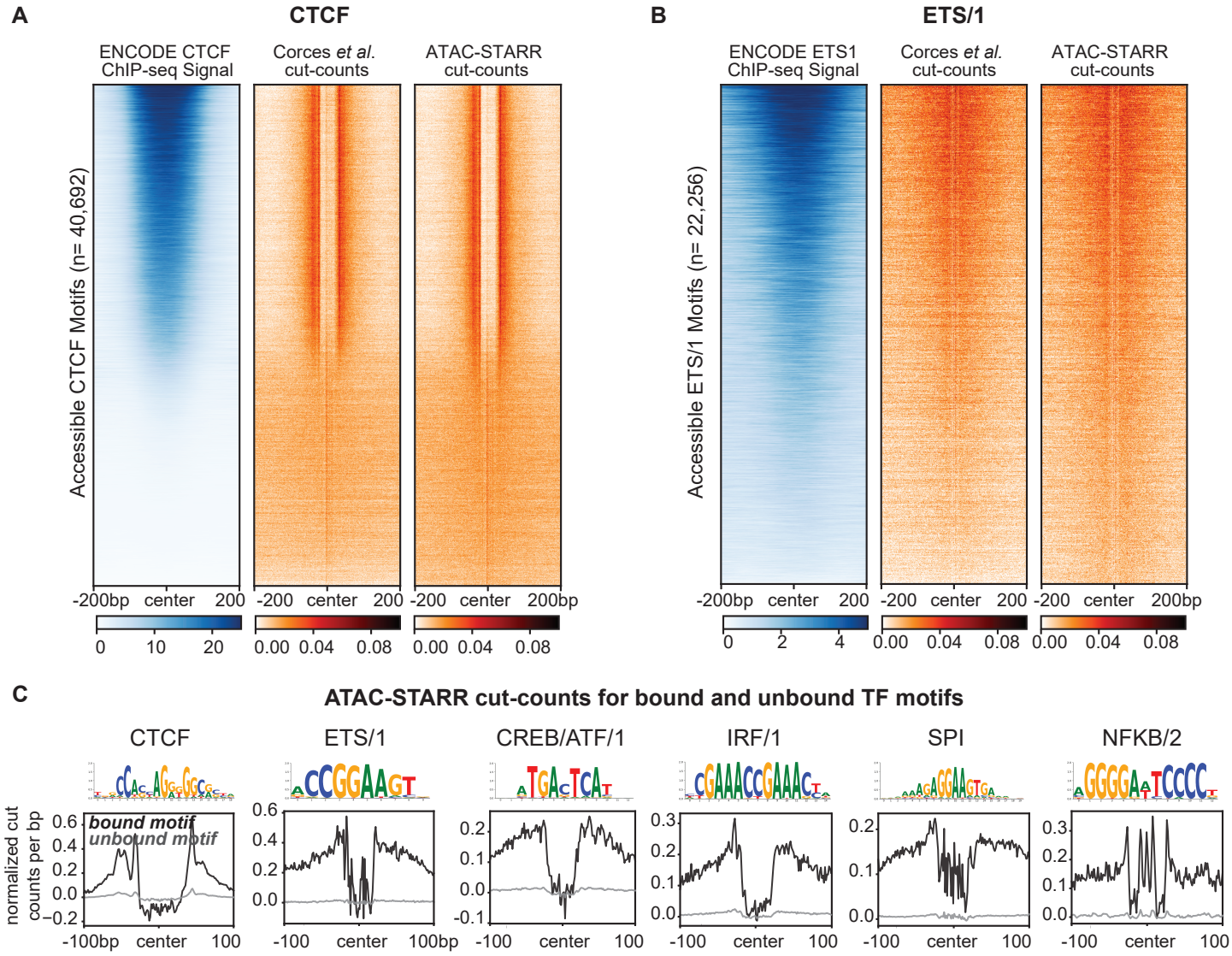
Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press



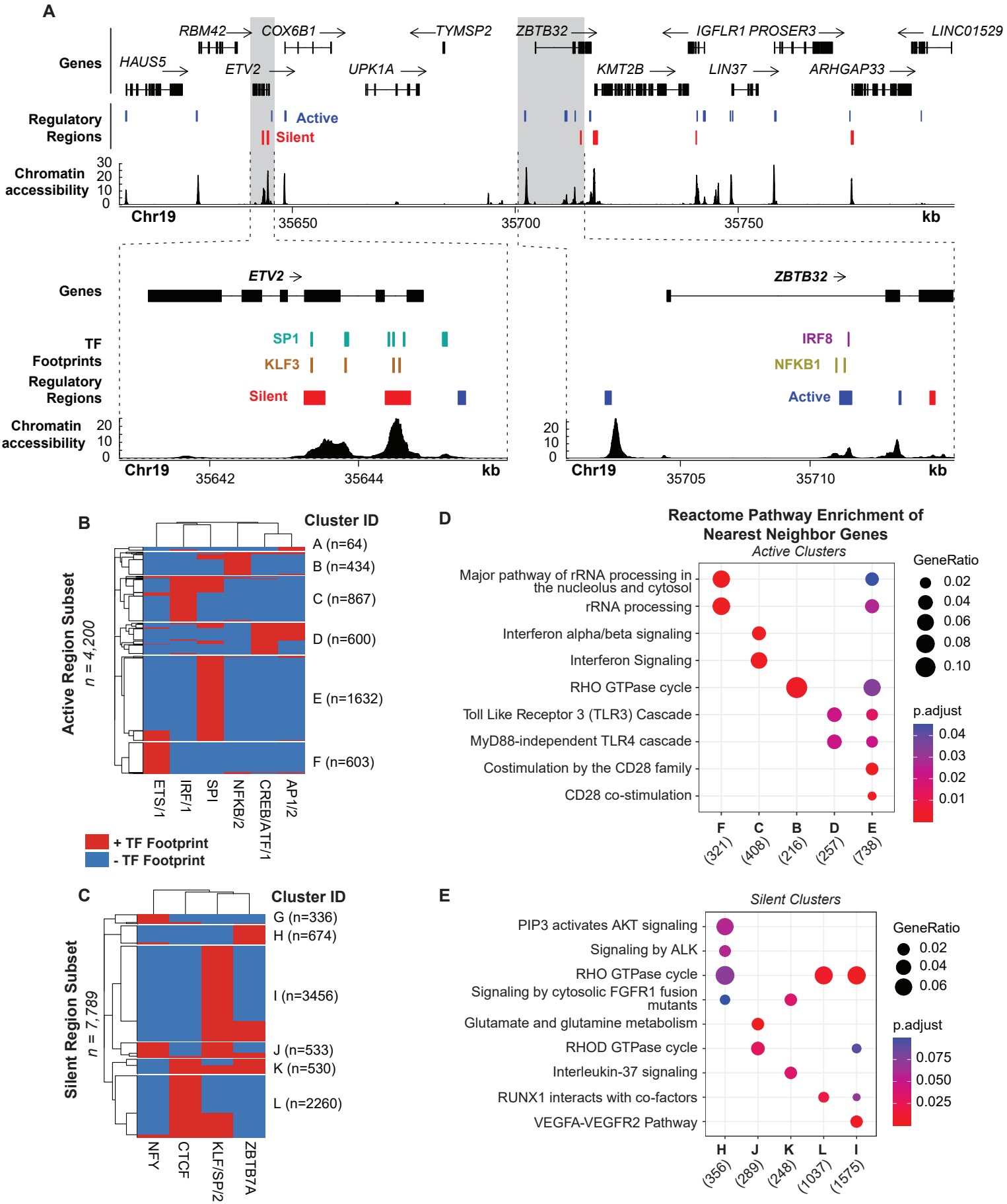
Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press



Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press



Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press



Downloaded from genome.cshlp.org on August 8, 2022 - Published by Cold Spring Harbor Laboratory Press

ATAC-STARR-seq reveals transcription factor-bound activators and silencers across the chromatin accessible human genome

Tyler J. Hansen and Emily Hodges

Genome Res. published online July 20, 2022
Access the most recent version at doi:10.1101/gr.276766.122

P<P Published online July 20, 2022 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This manuscript is Open Access.This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International license), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

Affordable, Accurate Sequencing.

gencove

To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press