Atlas of Transcription Factor Binding Sites from ENCODE DNase Hypersensitivity Data Across 27 Tissue Types

Cory C. Funk[1], Segun Jung[2], Matthew A. Richards[1], Alex Rodriguez[2], Paul Shannon[1], Rory Donovan[3], Ben Heavner[5], Kyle Chard,[2] Yukai Xiao[2], Gustavo Glusman[1], Nilufer Erteskin-Taner[8], Todd E. Golde[9], Arthur Toga[6], Leroy Hood[1], John D. Van Horn[6], Carl Kesselman[7], Ian Foster[2], Seth Ament[3], Ravi Madduri[2]*, Nathan D. Price[1]*

1. Institute for Systems Biology, Seattle WA. 2. University of Chicago, Chicago IL. 3. Allen Institute for Cell Science, Seattle WA. 4. Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 5. University of Washington, Seattle WA. 6. USC Mark and Mary Stevens Neuroimaging and Informatics Institute, Los Angeles, CA. 7. USC Information Sciences Institute, Los Angeles, CA. 8. Mayo Clinic, Department of Neuroscience, Jacksonville, FL. 9. University of Florida, Department of Neuroscience, Gainesville, FL.

*Correspondence to madduri@mcs.anl.gov or nathan.price@systemsbiology.org

Abstract

There is intense interest in mapping the tissue-specific binding sites of transcription factors in the human genome to reconstruct gene regulatory networks and predict functions for non-coding genetic variation. DNase-seq footprinting provides a means to predict the genome-wide binding sites for hundreds of transcription factors (TFs) simultaneously. However, despite the public availability of DNase-seq data for hundreds of samples, there is neither a unified analytical workflow nor a publicly accessible database providing the locations of footprints across all available samples. Here, we describe the implementation of a workflow for uniform processing of footprints using two state-of-the-art footprinting algorithms: Wellington and HINT. Our workflow then scans footprints for 1,530 sequence motifs to predict binding sites for 1,515 human transcription factors. We tested our workflow using 21 DNase-seq experiments of lymphoblastoid cell lines, generated by the ENCODE project. We trained a machine learning model to predict TF binding sites, integrating footprints with additional biologically-related features. This model achieved a maximum MCC of 0.423 and an AUC of 0.943 compared to ENCODE ChIP-seq data for 62 TFs in the same cell type. We applied our workflow to detect footprints in 206 DNase-seq experiments from ENCODE, spanning 27 human tissues. These footprints describe an expansive landscape of TF occupancy in the human genome. Across all tissues, we detected high-quality footprints spanning 9.8% of all nucleotides in the human genome with scores found to enrich for true positives. The highest tissue-specific coverage was observed for samples in the brain (4.4%), followed by extra-embryonic structure (2.6%) and skin (2.4%). In addition, we report a more lenient footprinting call set, providing some evidence of TF occupancy in at least one tissue for 34% of all genomic positions. Our cloud-based workflow and a database with all footprints and TF binding site predictions are available at www.trena.org.

Keywords: Footprinting; Transcription Factors; Gene Regulation; Motifs; ENCODE;

INTRODUCTION

Regulation of gene expression by transcription factors (TFs) forms the basis for tissue and cell-type differentiation, arising from a complex interplay between TFs and the chromatin architecture in gene regulatory regions (Neph et al., 2012a). The importance of these regulatory regions is evidenced by the fact that >90% of haplotypes associated with common diseases contain no protein-coding variants and are strongly enriched in non-coding regulatory regions (Gusev et al., 2014). Identifying causal variants and target genes for risk haplotypes is a central challenge in human genetics. Understanding the impact of variants on TF occupancy and target gene expression is increasingly important as it is one way to provide mechanistic insight for a given locus. Several major efforts have aimed at expanding functional annotation of the genome for a variety of reasons. The ENCODE project is one such important effort, generating the largest amount of DNase I Hypersensitivity (DHS) data. DHS assays are predicated on accessibility of genomic DNA to DNase I. Regions of open chromatin are susceptible to cleavage by DNase I. Binding of transcription factors and/or other proteins results in altered accessibility and can result in a relative difference in the number of cleavage events in discrete regions along the genome, resulting in a footprint (Galas and Schmitz, 1978). Cleavage by DNaseI is known to have sequence preferences, possibly reflecting DNA topology (Sung et al., 2014). ENCODE has generated samples across 27 tissues, as well as many immortalized cell lines. Though ENCODE has produced several analysis workflows for other data types, there is presently no workflow that uses the DHS data for the purpose of producing footprints. Such a workflow is presented herein and made freely available to the scientific community.

Disrupting or altering the binding dynamics of a transcription factor to its cognate sequence is one way in which variants can influence gene expression. An example of this is found in the successful association between a non-coding regulatory region and disease-associated genes in the fat mass and obesity-associated (FTO) region, which is the strongest obesity-associated

region in the genome (Claussnitzer et al., 2015). A single-nucleotide variant was found to disrupt a conserved motif for the ARID5B repressor that resulted in an increase of IRX3 and IRX5 mRNA, causing a developmental shift in cell type from energy-dissipating beige (brite) adipocytes to energy-storing white adipocytes. More recently, rs9349379, a SNP found to be associated with 5 different vascular diseases was shown to fall within a distal enhancer for EDN1. Editing of the SNP by CRISPR/Cas9 modified expression of EDN1 (Gupta et al., 2017). Despite the clear change in EDN1 expression with CRISPR modification, the authors could not find a clear mechanism as they found little evidence of interaction between the enhancer and the EDN1 promoter in 4C experiments, but did find a point of common contact between the rs9349379 and another putative enhancer 300 kb away. Our own search for functional annotation of GWAS variants in bipolar disorder identified a variant in VRK2 which we found to be regulated by the transcription factor POU3F2, another gene previously implicated in bipolar disorder by GWAS (Pearl et al., 2017).

Variants identified through GWAS and/or other means are likely to exert their effects in ways that are context-specific. This fact was poignantly demonstrated for variants found in monocytes from 228 individuals where, under different time points and conditions (2h LPS, 24h LPS or IFN-gamma), a given variant could have the opposite effect on gene expression (Fairfax et al., 2014). As few large expression datasets are capable of characterizing multiple conditions to probe such context-specific relationships, the challenge of functional annotation becomes even more apparent.

Functional annotation of variants can help to uncover the relationships between transcription factors and target genes. Functional annotation of motifs can enable network analyses of expression data (RNA-seq, microarray) that identify which transcription factors are likely key drivers in various biological processes and/or disease (Ament et al., 2016). A challenge for

functional annotation is the variety of motif databases which contain motifs that range from highly similar to disparate and are often unique, depending on the database. This is compounded by the multiple mappings that can occur between a given motif and transcription factor, and *vice versa*.

Computational identification of footprints from high-throughput data is an area of active research. Several algorithms exist which utilize mapped DHS data to identify footprints. These algorithms typically use one of two different approaches: 1) calculate the relative number of DNase cleavage events along a sliding window of segments, agnostic about the absence or presence of a transcription factor binding motif (Boyle et al., 2011; Gusmao et al., 2014; Neph et al., 2012b; Piper et al., 2013; Sung et al., 2014); 2) begin with the known location for a transcription factor binding motif and model the DNAse cleavage patterns around it for all sites, genome-wide (Cuellar-Partida et al., 2012; Kahara and Lahdesmaki, 2015; Pique-Regi et al., 2011; Sherwood et al., 2014; Yardimci et al., 2014). Validation of these approaches typically has involved comparison of the footprints for individual TFs to binding sites found by ChIP-seq. We have created a workflow that utilizes two different footprinting algorithms, HINT and Wellington, and applied them to uniformly aligned DHS data across all available ENCODE samples. To generate as many putative footprints as possible (which can be filtered for quality and/or biological relevance in subsequent analyses), we have chosen liberal parameters in our mapping and footprinting, knowingly generating a high number of false positives (which can be filtered for quality and/or biological relevance in subsequent analyses). We do this to identify as many true positives as possible to start, and subsequently filter the footprints by training a model against ChIP-seq data. As our liberal approach for representation within the Atlas comes with a tradeoff of producing more false positives than otherwise while emphasizing being comprehensive, we show that identification of the footprints enables greater prediction accuracy of ChIP-seq peaks from 62 transcription factors beyond motif information alone. Independent of any downstream

machine learning models, all footprints generated have an associated score that can be used for downstream application and evaluation, and for filtering down to lower coverage at higher quality to suit the needs of any specific user.

METHODS

Footprinting workflows were created and executed using various tools and services built and operated as a part of the NIH Big Data to Knowledge (BD2K) Big Data for Discovery Science (BDDS) center (http://bd2k.ini.usc.edu). At a high level, these tools enabled authoring and orchestration of complex, multi-tool workflows, transparent and elastic scaling on cloud resources, reproducible analysis based on provenance captured using Minids and BDBags (detailed below). The scalable workflows were built using the cloud-based Globus Genomics service (Madduri et al., 2014). These workflows include data retrieval from ENCODE using our ENCODE2Bag service that creates a portable and identified data unit that encapsulates the entire results of an ENCODE query at a point in time. The resulting BDBag is used to run various analysis workflows in parallel to identify DNA footprints using cloud-based resources. The Globus Genomics platform, coupled with the BDDS tools, facilitates reproducibility of complex analysis for large cohorts through well-defined and publishable workflows.

BDBags, Minids

The input data from ENCODE consisted of all available DNAse Hypersensitivity (DHS) datasets from 27 tissue types. ENCODE provides metadata for each tissue type which was exported and included in the exported BDBag (Chard et al., 2016). BDBag is a format for defining a dataset and its contents by enumerating the data elements, regardless of their location (enumeration, fixity and distribution) and metadata. BDBags can be passed between services and materialized (by downloading data elements) only when needed. All datasets used in the workflow are identified using Minids—a method for uniquely identifying a dataset irrespective of its location

(identify, fixity). Minid and BDBag tooling provide mechanisms for exchanging datasets by name, without regard for location or size, and with assurance that the data has not been modified.

ENCODE2Bag Service and Globus Transfer

The ENCODE2Bag service provides a simple interface for exporting identified, verifiable collections of data from ENCODE. The service is given an ENCODE query and dynamically creates a BDBag, stored on Amazon S3, and identified with a Minid. The BDBag does not contain the large genomics files, but rather includes a manifest file which enumerates the files and includes a checksum for verifying integrity when accessed. The summary of the ENCODE query, represented as a Tab Separated Value file, is included in the BDBag as provenance metadata. Thus, given a BDBag a user may, at any point in the future, obtain the results of that ENCODE query executed at the original time—an important property for reproducibility. BDBag tooling abstracts the process by which the BDBag is "materialized". The analysis workflow requires only the Minid of the input dataset to execute the workflow. It transparently resolves the location of the BDBag, transfers it to the cloud-hosted analysis infrastructure, and uses BDBag tools to materialize the contents of that BDBag.

Globus provides reliable, secure, and high performance data transfer between Globus "endpoints" (Chard et al., 2014). Globus provides direct access to a variety of storage endpoints ranging from local POSIX file systems, through to cloud object stores (e.g., AmazonS3), high performance file systems, and even archival tape storage. Globus is able to orchestrate data transfer between any two systems by managing authentication with both endpoints, optimizing transfer configurations for transfer rate, recovering from errors, and notifying users of transfer status. We use Globus file transfer functionality to move large

amounts of data from repositories, institutional storage systems, and local computers to the high performance, cloud-hosted compute resources used by the workflow.

Globus Genomics

Globus Genomics is a cloud-hosted web service to enable rapid analysis of large genomics data. Globus Genomics has over 3000 computationally optimized tools and a collection of best practices analysis workflows along with data management tools built as part of the BDDS BD2K center and makes it easier for researchers to build high performance, reproducible bioinformatics workflows. Given a BDBag or Minid, Globus Genomics BDBag transfer tools are used to automatically and reliably download the raw files included in that BDBag. In this workflow, we used these tools to materialize the BDBag for each tissue. Each tissue type contained DHS data for multiple patients. In addition, each patient had a variable number of replicate sequence data. Footprints were generated for the same input data using two alignment seed-lengths of 16 and 20 units, respectively. The analysis of the data consisted of aligning each replicate sample using the SNAP-aligner (Zaharia et al., 2011). Once the alignment BAM files were produced for each replicate, they were merged using Samtools (Li et al., 2009). The merged BAM file was used to generate regions of open chromatin using F-Seq (Boyle et al., 2011) based on the recommended parameters by Koohy et al, with the minimum reported size reduced from 500 bases to 400 (Koohy et al., 2014). Wellington was run with the -fdrlimit set to -1, to be the most lenient in reporting. HINT was run using standard settings. Neither Wellington nor HINT were run using any cleavage bias correction (Gusmao et al., 2014; Piper et al., 2013). The footprints were then stored in a relational database for ease of query.

The size of the input data (2.5 TB) and variability in replicate quantity for all samples (1400 Fastq samples) made for a complex analysis (Figure 1). The Globus Genomics platform allowed us to automate this analysis through its support for transparent batch submission and parallelization

methods. We utilized Amazon EC2 r3.8xlarge instance type with 32 cpus and 244 gigabyte memory per node. The analysis of all tissues generated over 5 TB of data while using approximately 64,000 CPU hours (2000 node hours). The analysis of each tissue was executed in parallel. In addition, each patient and their replicates were executed in parallel, as well as each footprint algorithm

Alignment

For each tissue type, we started with the fastq files (851 files) available at https://www.encodeproject.org/. These files were encapsulated within a BDBag that captured, in an unambiguous manner, references to the raw data alongside complete metadata for processing (Chard et al., 2016). Some ENCODE experiments contain multiple biological samples, while others may contain only a single sample. An ENCODE experiment may contain single or paired-end reads, with varying depth of sequencing and varying read length in a single experiment.

The ENCODE data was generated using short reads (<50 bases), resulting in a greater number of potential sequence matches than for longer reads. This ambiguity led us to produce alignments based on two different hash table seed lengths. Each fastq file (or paired-end files) was aligned to GRCh38 using the SNAP algorithm (Zaharia et al., 2011). SNAP uses a default seed length of 20. We additionally aligned to seed size 16, given the shorter sequence lengths. Using the experiment groupings from ENCODE, we produced 176 BAM files for each seed.

Identifying regions of open chromatin

Based on work from Koohy *et al.*, who compared four different approaches (F-seq, Hotspot, MACS and ZINBA) (Koohy et al., 2014) we used F-seq (Boyle et al., 2008) to identify regions of open chromatin from the aligned BAM files using the recommended parameters by Koohy *et al*. As stated in the F-Seq documentation, the results are non-deterministic because it uses a variable seed number in selecting a starting point for determining regions of open chromatin. The

seed sets the sliding frame at which regions are considered, leading to slightly different beginning and ending points of open-chromatin. The resulting regions (in BED format) vary slightly when repeated. The variable coverage on the edges becomes less of an issue with increased sample numbers.

RESULTS & DISCUSSION

In addition to generated the DHS data, ENCODE provides processed data for many but not all of these files. There is not a uniform workflow through which all the data has been processed. We set out to create such a workflow, as diagramed in Figure 1. As mapping of shorter reads has is less precise than longer reads, we aligned using two different seed lengths: 16 and 20 bases. Figure 2A compares the alignment overlap in lymphoblast. When accounting for the slight differences in location of the footprints (or small differences in footprint length), approximately 70% of the footprints had complete coverage between the two mappings. (Figure 2A). We also observed a weak relationship between the number of footprints found in a sample and the depth of sequencing (Figure 2C). HINT identifies more footprints relative to Wellington. The HINT samples also showed a wider range of total footprints compared to Wellington, likely due to HINT's higher sensitivity.

Collection and generation of the motif database

We utilized motifs and motif-transcription factor mappings from JASPAR, HOCOMOCO, UniPROBE, and SwissRegulon. Several redundancies occur between these databases, often containing position weight matrices that are similar or identical. A motif in one database can also be quite different from the motif in another database associated with the same transcription factor, resulting in different mappings. To reduce computation costs and avoid inclusion of redundant motifs, we updated and modified an existing R package, MotifDB (Shannon, 2017), to include the latest versions of all aforementioned databases. We evaluated the similarity of all

motifs using Tomtom (Gupta et al., 2007). Those that were significantly different (-log(p-value) ≥ 7.3) were retained. Each database contributed a unique set of motifs, in addition to providing additional mappings between motifs and transcription factors. The number of original motifs considered for each database and the number of motifs and transcription factor mappings considered after filtering with Tomtom is found in Table 1.

Mapping motifs to transcription factors

In addition to the mappings provided by each of the aforementioned databases, we integrated a large database of additional TF-motif mappings: TFClass (Wingender et al., 2015). The complete mapping can be accessed through MotifDB by calling the "associateTranscriptionFactors" method.

Collectively, our aggregation of motif databases and mappings contains 1,530 unique motifs recognized by 1,515 transcription factors. Many motifs were associated with a single transcription factor, while a few promiscuous motifs were associated with as many as 60 transcription factors (Figure 3). Reversing the association, many transcription factors were associated with one motif, while a few transcription factors were associated with > 100 motifs. The total number of motif-transcription factor mappings considered is 13,242.

Combining footprints with database of motifs

Footprints from HINT and Wellington are identified without consideration of the motif sequence. To efficiently identify all motif instances for a given footprint we create a catalog for all 1,530 motifs across the entire genome using FIMO (Grant et al., 2011). This resulted in ~1.34 billion matches (p-value < 10^{-4}) and covered almost 80% of the genome. To maximize coverage, and because of the potential imprecise nature of footprints, if any part of a known motif overlapped

with a single base of the footprint, an entry was created. Intersection was done by porting the motif instances and footprints into the GenomicRanges R package, using the “any” option. When considering all samples from all tissues, this liberal approach resulted in 34% coverage of the genome being represented in the Atlas for at least one tissue. The brain had the highest genome coverage at 14.9%, followed by skin (9.8%) and lymphoblast (8.9%). Urinary bladder had the lowest percentage of coverage at 1.1%. These genome coverages can represent both intrinsic biological differences across tissues as well as sample size differences in ENCODE. Based on our machine learning results (see below) we determined that filtering the HINT footprints for scores greater than 200 significantly enriches for true positives as defined by ChIP-seq hits. We also filtered the Wellington footprints, with a more negative score being better, for less than -27, also based on the machine learning results. These filters greatly reduced the percent coverage of the genome from 34% to 9.8% across all tissues (Figure 4A). In particular, the lymphoblast footprints were disproportionately reduced from 8.9% to 0.8%. The brain was also significantly reduced to 4.4%.

Footprints from the 29 brain samples (from multiple brain regions) accounted for 14.9% coverage of the entire genome. With the number of footprints identified related to the depth of sequencing, an outstanding question is to what extent additional samples will add previously unseen footprints. We ordered all the brain samples from greatest number of footprints to least number of footprints from the HINT20 samples and calculated the additional percentage of the genome covered by each sample (Figure 4B). Percent coverage enables us to account for footprints which may shift slightly across samples due to differences in mapped reads. Each of the deeper sequenced samples added additional coverage of the genome. As more samples are considered the ability of a new sample to increase coverage diminished. For example, the first sample added 1.75% genome coverage and 3.25 million hits while the last sample added 0.04% coverage which is ~235,000 hits. We performed the same analysis after filtering the footprints

based on the aforementioned HINT and Wellington scores. Filtering reduced the number of novel footprints added by each sample as seen by the slope of the line in the bottom panel of Figure 4B.

Machine learning on footprints

Having chosen liberal parameters to generate footprints with two different alignments, two different footprinting algorithms, and all motifs intersecting with any part of a footprint, we ingested this heterogeneous data with a flexible machine learning framework and performed a prediction task of identifying footprints with corresponding ChIP-seq hits. Machine learning was done using the R package XGBoost, with a maximum tree depth of 7 and 200 rounds of boosting and a logistic regression optimization criterion (Chen, 2016). As a benchmark for our footprints, we used the 77 ChIP-seq experiments generated by ENCODE in lymphoblasts. From our motif to transcription factor mappings, we were able to identify at least one motif for 72 of those transcription factors, for which 62 had information on their transcription factor class. It is this filtered set of 62 transcription factors, which mapped to a total of 264 motifs, that we used for our prediction task.

To unify the footprints generated in the multiple databases for lymphoblast, we joined all footprints based upon location in the genome to create one unified dataset. To account for the fact that the same footprints are often found in multiple samples from the same tissue, we retained the best score for each method and added as an additional metric the number of times a footprint was found at that location. As HINT is far more sensitive than Wellington, we scaled this count metric to one that captured the fraction of samples in which a given footprint was found. After the number of footprints for each location was summed, the highest number of occurrences was used as the denominator for all footprints in that method, resulting in a fractional representation for the occurrence metric. Additionally, we recognized that footprint-

motif intersections include overlap of any size, but regions with higher overlap might indicate higher-confidence cases. To capture this effect, we calculated the overlap distance between each motif and its footprints for both seed as a fraction of motif length.

JASPAR transcription factor class information was one-hot encoded in our feature matrix. GC content was calculated for each motif found within a footprint by using a window of 100 bases from the center on each side of the motif. Distance in base pairs (BP) to the nearest transcription start site (TSS) was calculated for each motif and transformed using the arcsinh (hyperbolic arcsine) function.

For purposes of training the model, we designated chromosomes 2 and 4 as a blind hold-out set for validation. Chromosomes 1, 3, and 5 were used to test the models as different parameters in architectures were explored. The remaining chromosomes were used to train the models. We trained 2 classes of models: 1) A basic logistic regression model; 2) A gradient boosted model, which aggregates an ensemble of decision trees to learn a nonlinear decision boundary. The boosted model was chosen based on its predictive power, as gradient boosted trees have been shown to offer state of the art performance for tasks of this nature (Olson et al., 2017). Regression models were constructed for their ease of interpretability, as well as for a baseline to which we compare the performance of the boosted models. We trained logistic regression models not only for all features in ensemble, but on each feature individually, in order to get an idea of which features were most predictive of ChIP-seq hits.

The number of footprints for a given motif (or set of motifs connected to a given transcription factor) is orders of magnitude larger than the number of ChIP-seq peaks. This imbalance Is challenging in our machine learning format, due to large memory requirements and poor signal to noise. To address this issue in our training set, we sampled 20 million hits for our 264 motifs,

combining those motif hits with our lymphoblast footprints, then filtering for a 10:1 ratio of negative-to-positives. We did not filter any of the ChIP-seq hits in our training set. This resulted in a more balanced training set in which the features associated with true positives could be better learned. We also used a statistical measure of performance, the Matthews Correlation Coefficient (MCC), that was designed to be robust to different sample sizes in the two classes being compared (Boughorbel et al., 2017).

To show how multiple motifs assigned to a single transcription factor can vary in performance for the predictive model, we trained the 26 motifs associated with the transcription factor ELK1. ChIP-seq hits were considered as true positives against footprints from HINT and Wellington. Often the same genomic location will contain matches for multiple motifs. We observed considerable variance for the different motifs (Figure 5). This result suggests that improved prediction could be achieved through selection of motifs.

To better understand the additional information contained in the quantitative scores from footprinting, we trained a model *devoid* of footprint values and related metrics. For this baseline model, we included all other information such as GC content and distance to the TSS, TF Class and motif score. We also only used motif locations where footprints were identified, as we wanted to test those relative to true positive ChIP-seq peaks, as that is a question of greater relevance to all motif instances. As expected, the boosted model achieved a relatively low MCC value of 0.322 and the linear model containing all regressors achieved a MCC of 0.274. Results can be seen in Figure 6A.

We then generated a model utilizing the full feature matrix with all footprinting data from Seed16 and Seed20. This full model achieved a MCC of 0.423 (Figure 6B). The linear model with all regressors performed comparably well with a maximum MCC of 0.403. Notably, a linear model

using only the best HINT20 scores performed slightly better with a max MCC value of 0.416. This linear model does not appear to be nearly as robust having a small threshold window relative to the boosted model and comprehensive linear model. The boosted model attained an AUC of 0.943 compared to the motif-only boosted model AUC of 0.811. In evaluating the models, the AUC was clearly driven by the high number of true negatives which comprise ___ of the ___ total observations. To illustrate this, we plotted a transformed score for Wellington and HINT along the x- and y axis, respectively, and colored each point with the prediction from the boosted model (Figure 7). The vast majority of true negatives are seen as having a lower HINT score, with the true-positives typically having a better HINT score. There appears to be little separation based on Wellington score. The majority of the false positives have high HINT scores. It should be kept in mind that false positive here is a “soft” assignment as a negative in a set of CHIP-seq experiment is not definitely for no binding under a different condition, and thus these high HINT scores could theoretically represent true positives in other cellular contexts.

While the gradient boosted decision tree models generated by XGBoost are somewhat opaque to human interpretation, some insightful metrics can be brought to bear. To identify those features most important for these models, we calculated the average gain in accuracy generated by each feature to the branches on which it occurs in the decision trees (Figure 8). In the motif-only model, where footprint scores are excluded, distance to the TSS was the most significant contributor to the prediction. In the boosted model, the dominant feature was the HINT20 score, followed by the HINT20 fraction, which represents the fraction of samples in which the footprint was observed. Several of the features we created contain informational redundancy. For example, a footprint found in multiple samples is more likely to have a good HINT or Wellington score. Similarly, a footprint with perfect overlap for a given motif is more likely to be found in multiple samples. Despite this redundancy, inclusion of these features contributed to the predictive power of our model.

CONCLUSION

We have created a uniform workflow for the ENCODE DNAse hypersensitivity data, generating two alignments and applying two different footprinting algorithms (HINT and Wellington). From multiple motif databases, we generated a non-redundant set of 1,530 unique motifs which corresponded to 1,515 transcription factors. Following intersection of the motifs and footprints from our liberal approach, we observed 34% of the genome to be covered. Filtering of the footprints based on the high-quality scores from HINT and Wellington resulted in covering 9.8% of the genome. We demonstrate the informational value of the footprints by training them on ChIP-seq data from 62 different transcription factors. Using baseline biologically important features (without footprint scores and metrics) we were able to predict ChIP-seq peaks at a relatively low MCC value of 0.322. With the footprinting score our predictive rate increased to 0.423 MCC. Importantly, this difference demonstrated that the footprinting information had a significant effect in improving predictive accuracy over motifs alone. Because of the inherent imbalance in the data, the primary driver of the AUC (and to a more muted extent the MCC) is the high number of true negatives. As the scores and metrics generated by machine learning are effectively tunable, we foresee those footprints which have good candidate scores but no corresponding ChIP-seq information are likely good transcription factor binding sites in other cellular contexts or conditions. The use of ChIP-seq as the gold standard has limitations and likely doesn't capture the full complement of TF binding sites. In this sense, our approach may offer a broader range of putative binding regions relevant to gene regulation. New approaches for identifying footprints are being developed by several groups. For example, last year SAGE Bionetworks had a DREAM Competition (https://www.synapse.org/#!Synapse:syn6131484) for new approaches to correctly identifying footprints, and this remains an active field of research. Our Globus Genomics workflow can be reproduced, extended with additional footprinting methods as new techniques become available, and is part of a family of interconnected tools

being built within our Big Data for Discovery Science (BDDS) center (http://bd2k.ini.usc.edu). We have made postgresql databases for all footprints in this analysis available at www.trena.org. Additionally, new DHS data will soon be released by ENCODE. Our approach and analysis can also be applied to future datasets, as they become available.

REFERENCES

Ament, S.A., Pearl, J.R., Bragg, R., Skene, P.J., Coffey, S., Bergey, D.E., Plaisier, C., Wheeler, V., MacDonald, M., Baliga, N.S., *et al.* (2016). Genome-scale transcriptional regulatory network models for the mouse and human striatum predict roles for SMAD3 and other transcription factors in Huntington's disease. bioRxiv.

Boughorbel, S., Jarray, F., and El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS One *12*, e0177678.

Boyle, A.P., Guinney, J., Crawford, G.E., and Furey, T.S. (2008). F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics *24*, 2537-2538.

Boyle, A.P., Song, L., Lee, B.K., London, D., Keefe, D., Birney, E., Iyer, V.R., Crawford, G.E., and Furey, T.S. (2011). High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res *21*, 456-464.

Chard, K., Arcy, M.D., Heavner, B., Foster, I., Kesselman, C., Madduri, R., Rodriguez, A., Soiland-Reyes, S., Goble, C., Clark, K., *et al.* (2016). I'll take that to go: Big data bags and minimal identifiers for exchange of large, complex datasets. Paper presented at: 2016 IEEE International Conference on Big Data (Big Data).

Chard, K., Tuecke, S., and Foster, I. (2014). Efficient and Secure Transfer, Synchronization, and Sharing of Big Data. IEEE Cloud Computing *1*, 46-55.

Chen, T.G., Carlos (2016). XGBoost: Scalable Tree Boosting System (CoRR).

Claussnitzer, M., Dankel, S.N., Kim, K.H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., *et al.* (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N Engl J Med *373*, 895-907.

Cuellar-Partida, G., Buske, F.A., McLeay, R.C., Whitington, T., Noble, W.S., and Bailey, T.L. (2012). Epigenetic priors for identifying active transcription factor binding sites. Bioinformatics *28*, 56-62.

Fairfax, B.P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., *et al.* (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science *343*, 1246949.

Galas, D.J., and Schmitz, A. (1978). DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res *5*, 3157-3170.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017-1018.

Gupta, R.M., Hadaya, J., Trehan, A., Zekavat, S.M., Roselli, C., Klarin, D., Emdin, C.A., Hilvering, C.R.E., Bianchi, V., Mueller, C., *et al.* (2017). A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. Cell *170*, 522-533 e515.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol *8*, R24.

Gusev, A., Lee, S.H., Neale, B.M., Trynka, G., Vilhjalmsson, B.J., Finucane, H., Xu, H., Zang, C., Ripke, S., Stahl, E., *et al.* (2014). Regulatory variants explain much more heritability than coding variants across 11 common diseases. bioRxiv.

Gusmao, E.G., Dieterich, C., Zenke, M., and Costa, I.G. (2014). Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. Bioinformatics *30*, 3143-3151.

Kahara, J., and Lahdesmaki, H. (2015). BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data. Bioinformatics *31*, 2852-2859.

Koohy, H., Down, T.A., Spivakov, M., and Hubbard, T. (2014). A comparison of peak callers used for DNase-Seq data. PLoS One *9*, e96303.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Madduri, R.K., Sulakhe, D., Lacinski, L., Liu, B., Rodriguez, A., Chard, K., Dave, U.J., and Foster, I.T. (2014). Experiences building Globus Genomics: a next-generation sequencing analysis service using Galaxy, Globus, and Amazon Web Services. Concurrency and Computation: Practice and Experience *26*, 2266-2279.

Neph, S., Stergachis, A.B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J.A. (2012a). Circuitry and dynamics of human transcription factor regulatory networks. Cell *150*, 1274-1286.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., *et al.* (2012b). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83-90.

Olson, R., La Cava, W., Mustahsan, Z., Varik, A., and Moore, J. (2017). Data-driven Advice for Applying Machine Learning to Bioinformatics Problems.

Pearl, J.R., Bergey, D.E., Funk, C.C., Basu, B., Oshone, R., Shannon, P., Hood, L., Price, N.D., Colantuoni, C., and Ament, S.A. (2017). Genome-scale transcriptional regulatory network models of psychiatric and neurodegenerative disorders. bioRxiv.

Piper, J., Elze, M.C., Cauchy, P., Cockerill, P.N., Bonifer, C., and Ott, S. (2013). Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. Nucleic Acids Res *41*, e201.

Pique-Regi, R., Degner, J.F., Pai, A.A., Gaffney, D.J., Gilad, Y., and Pritchard, J.K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res *21*, 447-455.

Shannon, P.R., Matt (2017). MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs (Bioconductor).

Sherwood, R.I., Hashimoto, T., O'Donnell, C.W., Lewis, S., Barkal, A.A., van Hoff, J.P., Karun, V., Jaakkola, T., and Gifford, D.K. (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat Biotechnol *32*, 171-178.

Sung, M.H., Guertin, M.J., Baek, S., and Hager, G.L. (2014). DNase footprint signatures are dictated by factor dynamics and DNA sequence. Mol Cell *56*, 275-285.

Wingender, E., Schoeps, T., Haubrock, M., and Donitz, J. (2015). TFClass: a classification of human transcription factors and their rodent orthologs. Nucleic Acids Res *43*, D97-102.

Yardimci, G.G., Frank, C.L., Crawford, G.E., and Ohler, U. (2014). Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. Nucleic Acids Res *42*, 11865-11878.

Zaharia, M., J. Bolosky, W., Curtis, K., Fox, A., Patterson, D., Shenker, S., Stoica, I., Karp, R., and Sittler, T. (2011). Faster and More Accurate Sequence Alignment with SNAP, Vol 1111.

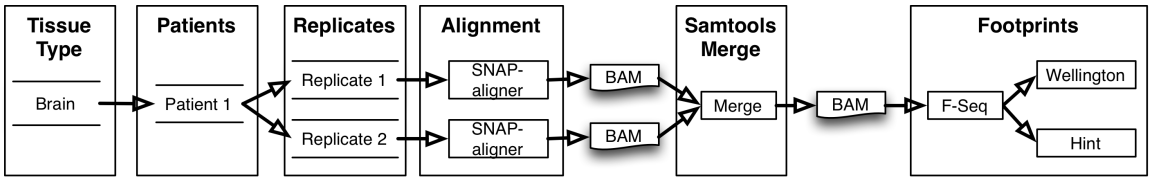

Figure 1. Footprints workflow overview - Each tissue type can have multiple quantity of patients and replicates. Each replicate is aligned using SNAP-aligner. All replicates for each patient are merged using Samtools. Finally, footprints for each BAM file are produced using Wellington and Hint and stored in a database.

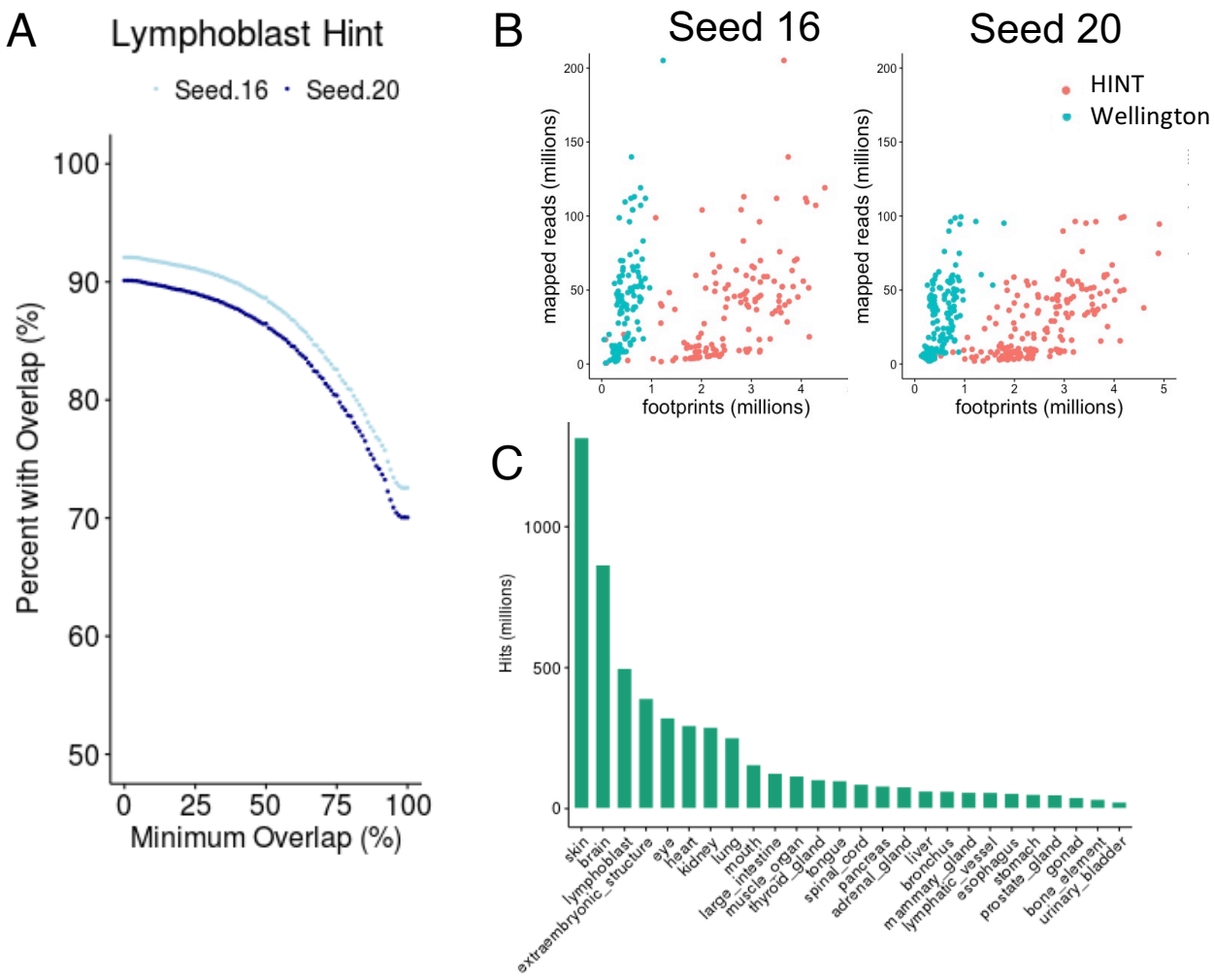

Figure 2. A) Footprints from the Seed16 alignment is compare to Seed20 alignment, considering the percent overlap. B) The number of footprints compared to the number of mapped reads for HINT and Wellington for all samples. C) The total number of hits (footprints and all intersecting motifs) per tissue type.

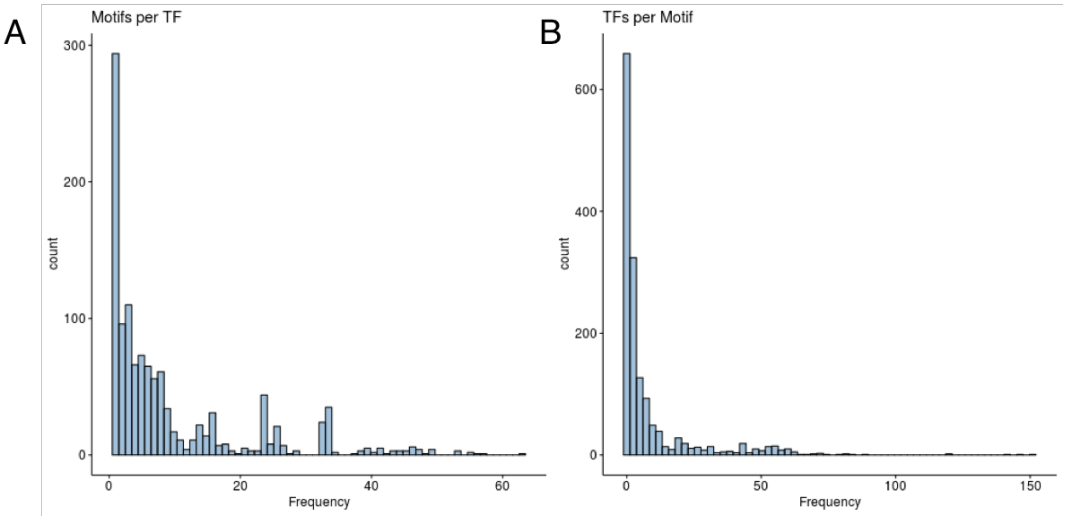

Figure 3. Motifs from JASPAR, HOCOMOCO, UniProt and SwissRegulon were combined into a non-redundant set of 1,530. The number of motifs per transcription factor is plotted on the left. The number of transcription factors per motif is on the right.

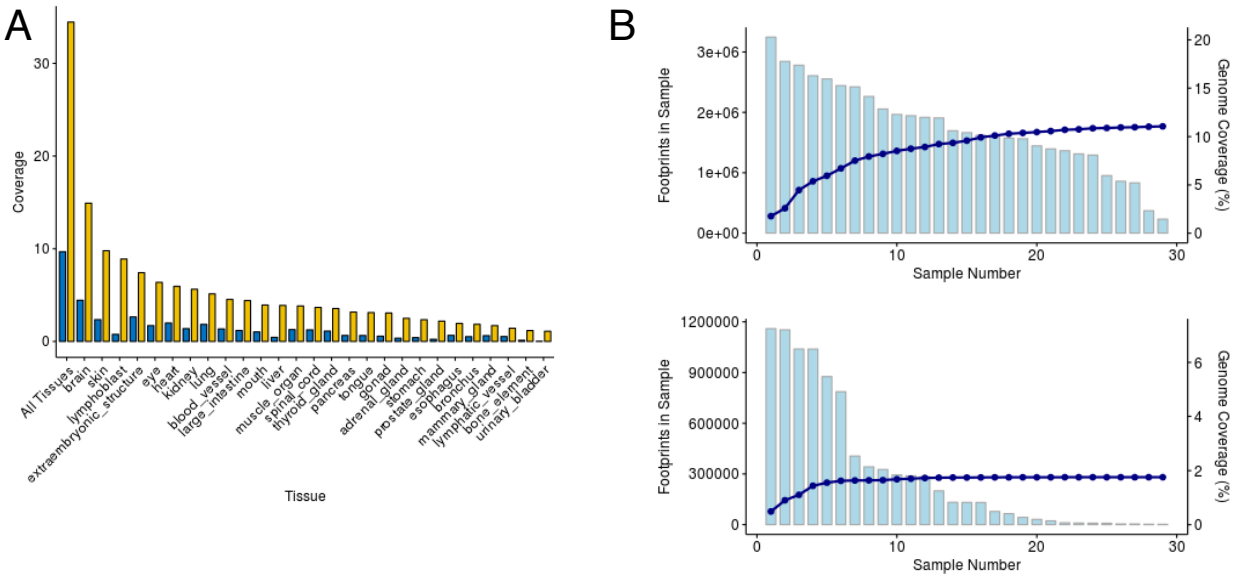

Figure 4. A) Percentage of the genome covered by the footprints for each tissue type and all tissues. Yellow is without filtering and dark blue is filtering HINT score > 200 and Wellington score < -27. (each method has its own scale and distribution) B) Footprints from the brain for HINT seed size 20 are ordered based on the number of footprints and summed. The light blue graphs represent the total number of footprints in each sample (top is without filtering on score; bottom is filtered as in panel A). The the dark blue line represents the cumulative percentage of the genome covered.

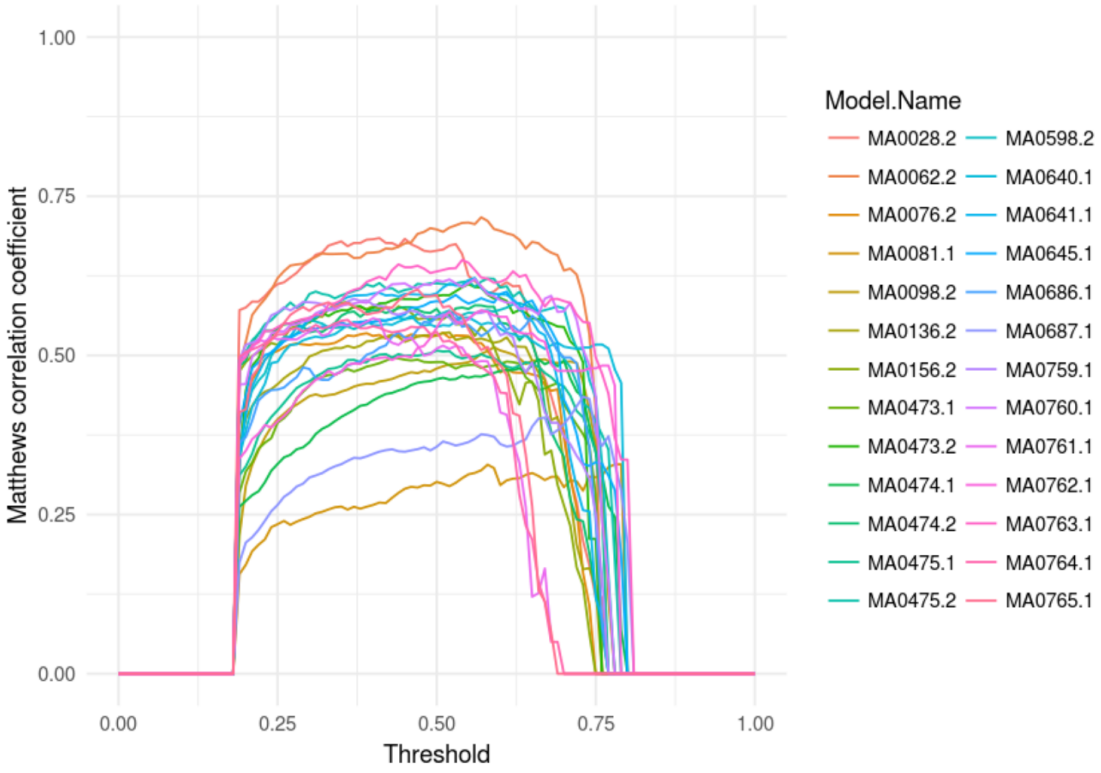

Figure 5. Mathew's correlation coefficient from XGBoost model for all motifs associated with the transcription factor ELK1.

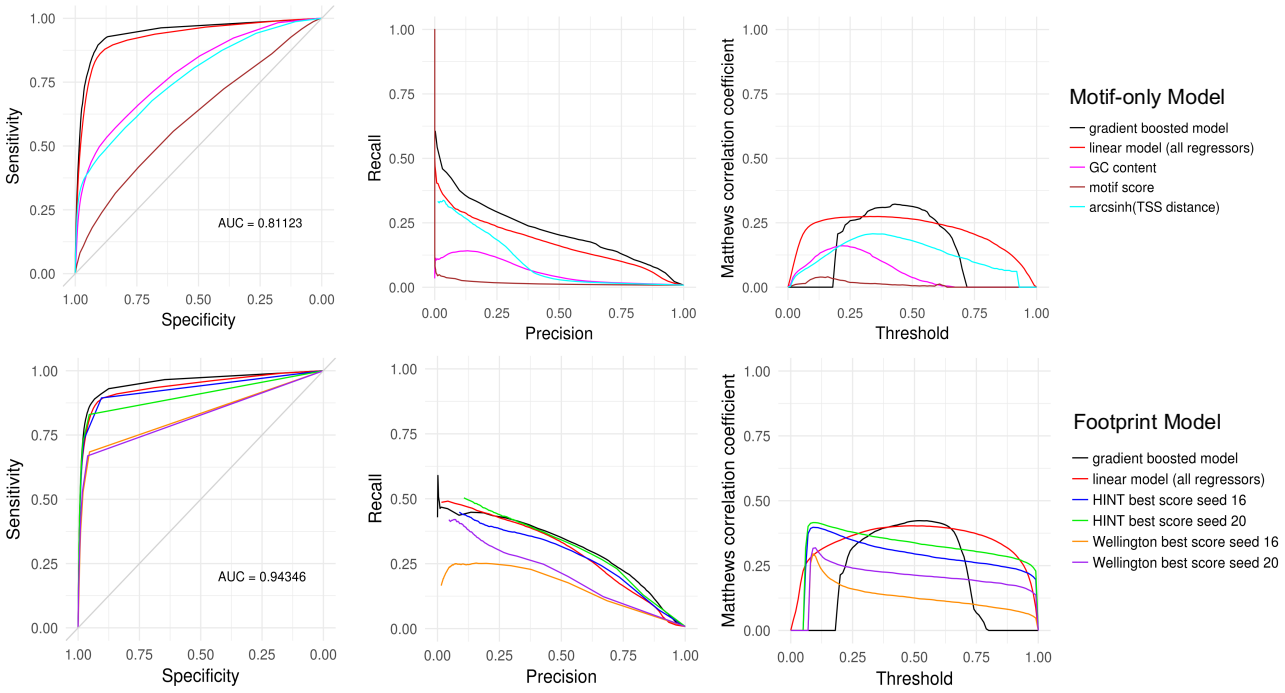

Figure 6. Machine learning results for XGBoost for the 62 transcription factors (264 motifs) training and testing the ENCODE-generated ChIP-seq samples using only motif information, TSS distance and GC content. A) Results using motifs devoid of footprint scores and metrics but including the following features: GC content, motif score, distance to TSS, and TF classes. B) Results for footprints generated from both Seed16 & Seed20 alignments using all aforementioned features, footprint scores and footprint metrics.

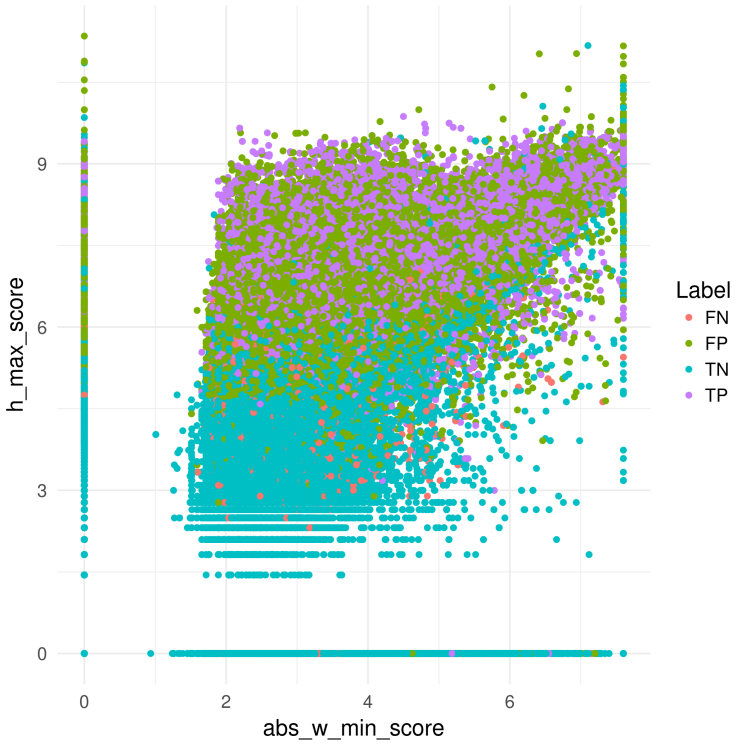

Figure 7. The relationship between the predictions made by the machine learning is plotted on top of the two footprinting scores. The absolute value of the hyperbolic arc sine of Wellington 20 and HINT 20 score are on the x- and y-axis, respectively. The colors represent all possible prediction outcomes (false negative, false positive, true negative, true positive) from the boosted model.

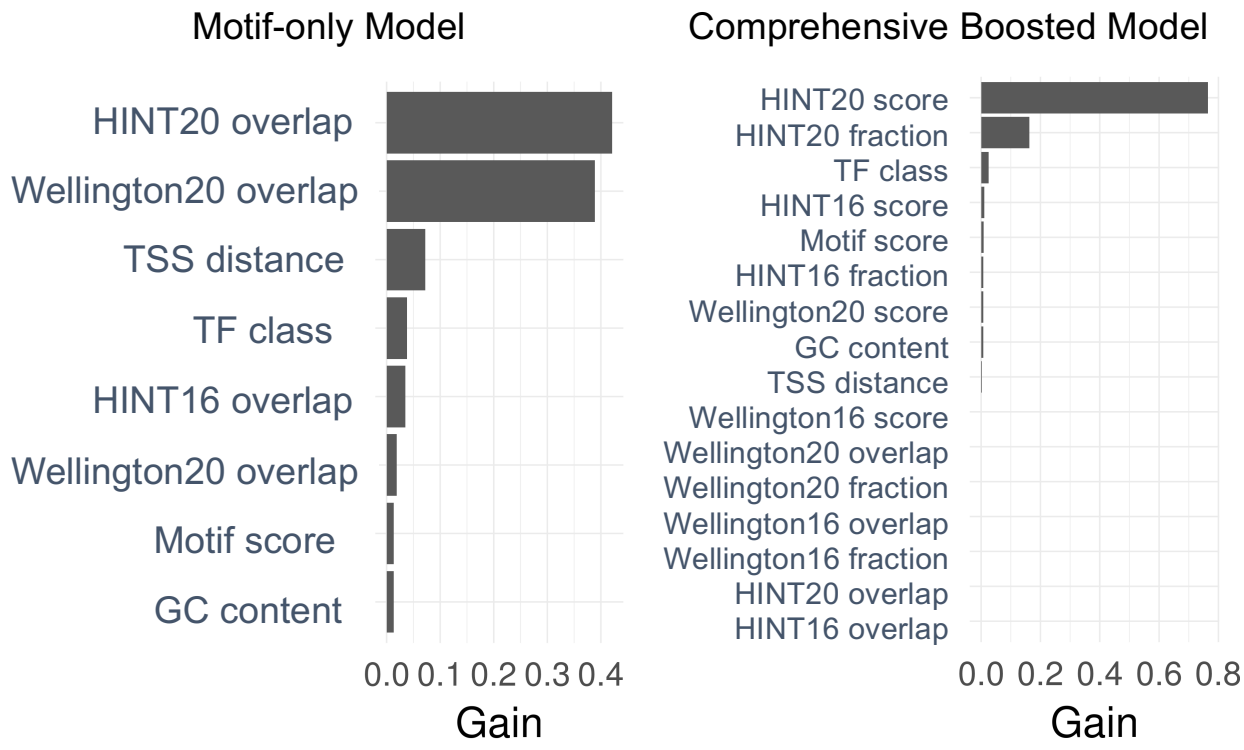

Figure 8. Importance matrix quantifying the contribution of each feature when trained and tested on the ENCODE ChIP-seq dataset for 62 transcription factors.

Origin	Motifs	Final Motifs	Total Mappings	Total TFs
jaspar2016	631	631	631	544
HOCOMOCOv10	1066	649	1679	604
UniPROBE	380	162	1104	363
SwissRegulon	684	88	3835	684
TFClass	NA	NA	8570	762

Table 1. Motif database sources for FIMO matching. Motifs represents the number of motifs in the original database. Final Motifs represent the number of motifs used after running Tomtom. Total Mappings are the motif-to-transcription factor mappings found in the associated metadata and Total TFs are the number of transcription factors for which a mapping is found in the corresponding database.