Published in final edited form as:

Cell Rep. 2020 August 18; 32(7): 108029. doi:10.1016/j.celrep.2020.108029.

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

Cory C. Funk[1,11], **Alex M. Casella**[2,3,11], **Segun Jung**[4], **Matthew A. Richards**[1], **Alex Rodriguez**[4], **Paul Shannon**[1], **Rory Donovan-Maiye**[1], **Ben Heavner**[1], **Kyle Chard**[4], **Yukai Xiao**[4], **Gustavo Glusman**[1], **Nilufer Ertekin-Taner**[5], **Todd E. Golde**[5], **Arthur Toga**[6], **Leroy Hood**[1], **John D. Van Horn**[7], **Carl Kesselman**[8], **Ian Foster**[4,9], **Ravi Madduri**[4,9,*], **Nathan D. Price**[1,*], **Seth A. Ament**[2,10,12,*]

[1]Institute for Systems Biology, Seattle, WA 98109, USA

[2]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201, USA

[3]Medical Scientist Training Program, University of Maryland School of Medicine, Baltimore, MD 21201, USA

[4]Globus, University of Chicago, Chicago, IL 60637, USA

[5]Mayo Clinic, Department of Neuroscience, Jacksonville, FL 32224, USA

[6]Mark and Mary Stevens Neuroimaging and Informatics Institute, University of Southern California, Los Angeles, CA 90033, USA

[7]Department of Psychology, University of Southern California, Los Angeles, CA 90007, USA

[8]Information Sciences Institute, University of Southern California, Los Angeles, CA 90292, USA

[9]Data Science and Learning Division, Argonne National Laboratory, Argonne, IL 60439, USA

[10]Department of Psychiatry, University of Maryland School of Medicine, Baltimore, MD 21201, USA

[11]These authors contributed equally

[12]Lead Contact

SUMMARY

[*]Correspondence: madduri@anl.gov (R.M.), nathan.price@systemsbiology.org (N.D.P.), sament@som.umaryland.edu (S.A.A.).

AUTHOR CONTRIBUTIONS

Conceptualization, C.C.F., A.M.C., R.M., N.D.P., and S.A.A.; Methodology, A.M.C.; Software, C.C.F., A.M.C., S.J., M.A.R., A.R., P.S., R.D.-M., B.H., K.C., and Y.X.; Formal Analysis, C.C.F., A.M.C., M.A.R., R.D.-M., and S.A.A.; Data Curation, C.C.F., A.M.C., and S.A.A.; Writing – Original Draft, C.C.F., A.M.C., and S.A.A.; Writing – Review & Editing, all authors; Visualization, C.C.F., A.M.C., M.A.R., and S.A.A.; Supervision, R.M., N.D.P., and S.A.A.; Project Administration, R.M., N.D.P., and S.A.A.; Funding Acquisition, R.M., N.D.P., and S.A.A.

SUPPLEMENTAL INFORMATION

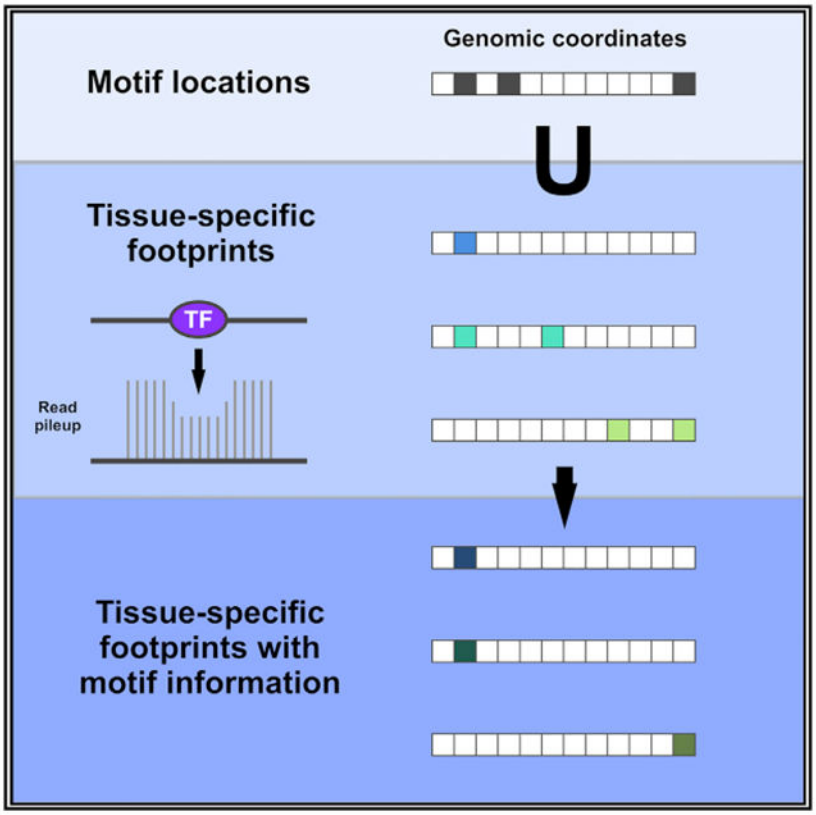
Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2020.108029.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Characterizing the tissue-specific binding sites of transcription factors (TFs) is essential to reconstruct gene regulatory networks and predict functions for non-coding genetic variation. DNase-seq footprinting enables the prediction of genome-wide binding sites for hundreds of TFs simultaneously. Despite the public availability of high-quality DNase-seq data from hundreds of samples, a comprehensive, up-to-date resource for the locations of genomic footprints is lacking. Here, we develop a scalable footprinting workflow using two state-of-the-art algorithms: Wellington and HINT. We apply our workflow to detect footprints in 192 ENCODE DNase-seq experiments and predict the genomic occupancy of 1,515 human TFs in 27 human tissues. We validate that these footprints overlap true-positive TF binding sites from ChIP-seq. We demonstrate that the locations, depth, and tissue specificity of footprints predict effects of genetic variants on gene expression and capture a substantial proportion of genetic risk for complex traits.

Graphical Abstract



In Brief

DNase-seq footprinting provides a means to predict genome-wide binding sites for hundreds of transcription factors (TFs) simultaneously. Funk et al. analyze data from the ENCODE consortium to create a resource of footprints in 27 human tissues, demonstrating associations of tissue-specific TF occupancy with gene regulation and disease risk.

INTRODUCTION

Regulation of gene expression by transcription factors (TFs) forms the basis for tissue- and cell-type differentiation arising from complex interplay between the TFs and the chromatin architecture in gene regulatory regions (Neph et al., 2012a; Tewhey et al., 2016). In humans, genetic perturbation of TF binding sites is thought to be an important mechanism by which single-nucleotide polymorphisms (SNPs) influence risk for human disease (ENCODE Project Consortium, 2012; Gusev et al., 2014; Maurano et al., 2012). Thus, characterizing the cell-type-specific occupancy of TFs at their genomic binding sites is a critical goal in genomics, providing insight into networks of TFs and their cell-type-specific target genes, as well as causal mechanisms underlying risk for human disease (Ament et al., 2018; Claussnitzer et al., 2015; Gupta et al., 2017; Moyerbrailean et al., 2016; Pearl et al., 2019).

Mapping human gene regulation requires comprehensive resources of tissue- and cell-type-specific TF binding sites. Major efforts over the past decade have produced vast quantities of public epigenomic data that have dramatically expanded the functional annotation of the human genome (Encode Project Consortium, 2004; Battle et al., 2017; Ward and Kellis, 2016), yet our understanding of cell-type-specific TF binding sites remains far from complete. Annotation of TF binding sites based solely on the locations of sequence motifs is imprecise because only ~1% of motif instances are occupied by a TF at any given time (Neph et al., 2012a). Similarly, information about the locations of promoters and enhancers lacks sufficient specificity because many genetic variants in these regions do not affect gene expression (Tewhey et al., 2016). TF occupancy can be ascertained with high sensitivity and specificity through chromatin immunoprecipitation followed by deep sequencing (ChIP-seq), in which an antibody specific to a TF is used to pull down genomic DNA fragments occupied by that TF in a given sample. However, high-quality ChIP-seq data have been generated for only a minority of all human TFs and often used standard cell lines rather than disease-relevant human tissues.

Genomic footprinting is a higher-throughput approach that predicts TF genomic occupancy by combining information from open chromatin assays (such as DNase sequencing [DNase-seq]) with information about the locations of sequence motifs recognized by the DNA binding domains of TFs. DNase-seq assays are predicated on accessibility of genomic DNA to DNase I, where regions of open chromatin are susceptible to cleavage by DNase I. Binding of TFs and other DNA binding proteins can lead to a relative difference in the number of cleavage events in discrete regions along the genome, resulting in a footprint (Galas and Schmitz, 1978). Computational algorithms have been developed to identify footprints from high-throughput DNase hypersensitivity (DHS) data, typically using one of two strategies: (1) sliding window approaches in which the relative number of DNase cleavage events are counted along a sliding window of the genome, agnostic to the absence or presence of a TF binding motif (Boyle et al., 2011; Gusmao et al., 2014; Neph et al., 2012b; Piper et al., 2013; Sung et al., 2014) and (2) approaches that begin with the known location for a TF binding motif and model the DNase cleavage patterns around it for all sites in the genome (Cuellar-Partidaet al., 2012; Kähärä and Lähdesmäki, 2015; Pique-Regi et al., 2011; Sherwood et al., 2014; Yardımcı et al., 2014). Validation of these approaches typically has involved comparison of the footprints for individual TFs to binding sites found by ChIP-

seq. Notably, the computational identification of footprints from high-throughput data remains an area of active research, because existing algorithms detect genomic occupancy for only a subset of TFs. Moreover, because of redundancy in the sequence specificity of TFs, footprinting generally cannot distinguish which member of a TF family is occupying a footprint. Nonetheless, the accuracy and reproducibility of TF binding site predictions from footprinting analysis has begun to rival that of ChIP-seq, and DNase-seq footprinting has successfully been used to predict the binding sites for hundreds of TFs in a parallel approach.

One of the most important applications of comprehensive atlases of TF binding sites will be to functionally annotate genetic risk variants for human diseases. Many studies have shown that disease-associated SNPs are enriched in gene regulatory regions, including open chromatin regions identified through DNase-seq and ATAC-seq experiments (de la Torre-Ubieta et al., 2018; Finucane et al., 2015; Gusev et al., 2014; Maurano et al., 2012). However, genome-wide association study (GWAS) risk loci are defined by large sets of genetically correlated SNPs with similarly strong statistical associations to disease, of which only a subset are thought to be functional and causal for disease risk. It remains controversial how many of these causal SNPs disrupt gene regulation by altering the specific base pairs occupied by TFs versus other mechanisms. Several studies have identified risk loci for traits such as obesity and schizophrenia in which causal variants appear to functionally alter binding sites for key TFs (Claussnitzer et al., 2015; Gupta et al., 2017; Pearl et al., 2019). However, other studies question the generalizability of this insight and indicate that TF binding sites in existing databases do not fully predict causal variants (Moyerbrailean et al., 2016). One explanation for this discrepancy is that existing TF binding site databases do not include sufficient amounts of epigenomic data from disease-relevant tissues. Because the gene regulatory consequences of non-coding SNPs are likely to vary dramatically across tissues and cell types (Claussnitzer et al., 2015; Fairfax et al., 2014), these existing databases may miss context-specific effects of variants on TF occupancy. In addition, there is considerable variability in the sensitivity and specificity of footprinting algorithms, and it is unclear which approaches will be best suited for this task.

Here, we developed a comprehensive resource of genomic footprints across 27 human tissues, using data from 192 DNase-seq experiments from the Encyclopedia of DNA Elements (ENCODE). Before our work, there was no publicly available, scalable workflow using these data for the purpose of producing footprints. These analyses revealed an expansive landscape of tissue-specific genomic occupancy for 1,530 TFs. We validated our database based on ChIP-seq and expression quantitative trait loci (eQTLs), and we demonstrated that tissue-specific footprints are strongly and specifically enriched for disease-associated genetic variation. We have made our footprint database and the underlying cloud-based computational workflow available in a user-friendly and intuitive format (links available in STAR Methods) (Madduri et al., 2019).

RESULTS

A Comprehensive Atlas of Genomic Footprints across Human Tissues

ENCODE-generated DNase-seq FASTQ files from 192 experiments in 27 tissues were downloaded from the ENCODE data portal (https://www.encodeproject.org/). The tissue-specific genomic occupancy of 1,515 TFs was then predicted through genomic footprinting analyses using the workflow pictured in Figure 1A and detailed in STAR Methods. First, sequence reads were aligned to GRCh38 using SNAP (Zaharia et al., 2011). Because the DNase-seq data consist of short reads, we generated two alignments: one using the default 20 bp seed length (Seed20) and another using a 16 bp seed length (Seed16). We then identified regions of open chromatin in each of the 192 experiments using F-seq, followed by detection of footprints using both HINT and Wellington algorithms. Footprints detected in each of the 192 experiments were then grouped by tissue, producing 27 tissue-specific footprint maps, with separate maps for each seed size and footprinting algorithm. In general, seed size had only a modest impact. ~70% of the footprints had complete overlap between the two seed sizes (Figure S1A). In addition, we observed only a moderate relationship between the number of footprints found in a sample and the depth of sequencing (Figure S1B). Overall, HINT identified more footprints than Wellington.

Footprints from HINT and Wellington are identified without consideration of underlying motif sequence. Therefore, to predict which TFs occupy each footprint, we used Find Individual Motif Occurrences (FIMO) to create a catalog of all genome-wide instances of 1,530 sequence motifs recognized by 1,515 TFs (Grant et al., 2011). In addition to the motif-TF mappings provided by the aforementioned databases, we expanded the motif-TF mappings to incorporate families of TFs with similar DNA sequence specificity, using information from TFClass (Wingender et al., 2015) (Tables S1 and S2; Figure S2). This resulted in ~1.34 billion sequence-to-TF matches ($p < 10^{-4}$) before intersection with footprints, spanning almost 80% of the genome. These motif instances were then intersected with the footprints from Wellington and HINT to produce an atlas of predicted TF occupancy in each tissue.

When considering all samples from all tissues, the most liberal thresholds 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% (Figure 1B). Sample size and sequencing depth were the main determinants for the number of tissue-specific footprints identified in our atlas. However, intrinsic biological differences in tissue complexity also influence the number of distinct footprint locations. For example, we found strong overlap in footprint locations across the 46 experiments from skin (average pairwise Jaccard similarity index = 0.28), consistent with skin being a relatively homogeneous tissue. By contrast, the footprints detected in the 29 experiments from brain were less homogeneous (average pairwise Jaccard similarity score = 0.16), which likely reflects the highly specialized and disparate cell types and cell-type-specific gene regulation across brain regions. As a consequence, we identified more brain footprints than skin footprints, despite having 50% more skin samples.

An outstanding question is to what extent additional samples would add previously unseen footprints. To address this, using footprints derived from the HINT algorithm with seed length 20 (HINT20), we ordered the brain samples from most to fewest footprints and calculated the additional percentage of the genome covered by each sample (Figure 1C). The first sample contributed 3.25 million footprints spanning 1.75% of the genome, whereas the last sample added 235,000 novel footprints and 0.04% novel genome coverage. We repeated the same analysis using only high-quality footprints based on HINT and Wellington scores (see the next section). As expected, this analysis revealed even greater overlap across samples, because many footprints detected in only a single sample are of low quality (Figure 1C, bottom). These results suggest that at least for well-sampled tissues such as brain, our atlas captures most detectable footprints.

Validation and Filtering of Footprints with ChIP-Seq and Machine Learning

Next, we sought to validate TF binding site predictions in our atlas and chose appropriate thresholds at which footprints reliably indicate TF occupancy. For this purpose, we compared footprints from 21 DNase-seq experiments in lymphocytes to predicted TF binding sites (peak regions) from ChIP-seq of 66 TFs in the same cell type. The genomic background for this analysis is the set of all genome-wide instances of the sequence motifs recognized by a given TF. On their own, these motif instances have an extremely high false-positive rate > 90%. We used the footprints from all 21 samples to define two scores at each genomic location: (1) the best footprint score, defined as the highest score at this location in any samples, and (2) the footprint fraction, defined as the proportion of independent samples with a non-zero footprint score. We then tested for a linear relationship between these footprint scores and the likelihood that a motif instance corresponded to a true-positive binding site from ChIP-seq, testing performance via the Matthews correlation coefficient (MCC), area under the receiver operator curve (AUROC), and area under the precision-recall curve (AUPR). The most accurate predictor was the best HINT20 score, which achieved a maximum MCC of 0.42, corresponding to AUROC > 0.9 (Figure 2). The high AUROC was driven by true negatives, which comprise 3,936,242 of the 4,110,504 total observations. Most true negatives had low HINT scores. True positives often had a high HINT score, but high HINT scores also had a significant false-positive rate (Figure S3). True and false positive here are soft assignments, because ChIP-seq experiments are imperfect predictors of TF occupancy.

We were curious whether performance could be improved by combining footprint scores from multiple algorithms with additional information about genomic context. We employed a supervised machine-learning approach, treating the ChIP-seq peaks as true positives. We employed two machine-learning algorithms: linear regression and gradient boosting trees implemented with XGBoost. We constructed and evaluated a comprehensive model that included as predictors the footprint scores from both HINT and Wellington using both the 16 bp and 20 bp seed sizes. Additional predictors included a score for the strength of the match to the sequence motif, TF class, guanine-cytosine (GC) content, and distance to a transcription start site (TSS). We compared this comprehensive model to predictions based on footprint scores alone, as well as to a baseline model that considered motif scores and genomic context but ignored footprinting data.

In the comprehensive model, gradient boosting and linear regression achieved maximum MCCs of 0.42 and 0.40, respectively (Figure 2), The predictor with the largest contribution to accuracy was the best HINT20 score, followed by the HINT20 footprint fraction (Figure 3). Prediction accuracy was lower in the baseline models but remained better than chance (gradient boosting, MCC = 0.32; linear regression, MCC = 0.27) (Figure 2). In these models, distance to the TSS was the most significant contributor to the prediction. Although the maximum MCCs of the HINT20 footprint-only versus comprehensive models were identical (0.42), the footprint-only model had a relatively small threshold window within which both true-positive and false-negative error rates were well controlled. Therefore, incorporating information about genomic context does not dramatically improve prediction accuracy but could potentially improve the robustness of these predictions.

We used machine-learning models to select appropriate cutoffs for high-quality footprints. We determined that a HINT score > 200 and a Wellington score < −27 were optimal filtering thresholds to control both false-positive and false-negative errors. Applying these filters reduced the percentage of coverage of the genome from 34% to 9.8% across all tissues (Figure 1B). This filtered estimate is in line with current estimates for the fraction of the genome that is actively involved in gene regulation. HINT20 footprints with scores > 200 were used in downstream analyses unless otherwise specified.

Footprints Predict Effects of Genetic Variants on Gene Expression

An important goal for footprinting is to predict the gene regulatory effects of non-coding SNPs. It has previously been shown that most haplotypes with *cis*-acting effects on gene expression (eQTLs) contain SNPs that are located within DNase I hypersensitive regions (Handel et al., 2017). However, DHS regions span a large fraction of the genome, and many SNPs within DHS regions have no evidence for influencing gene expression. It remains controversial whether footprints more precisely capture the causal variants on eQTL haplotypes: some recent studies found that only a small fraction of eQTL haplotypes overlap footprints (Handel et al., 2017; Moyerbrailean et al., 2016), whereas others have suggested stronger enrichment (Degner et al., 2012; Schwessinger et al., 2017). To address this question, we examined overlap between footprints in our database with eQTLs from the Genotype-Tissue Expression (GTEx) consortium.

We evaluated overlap between footprints (HINT20 score ≥ 200) from our database with eQTLs in 44 tissues from GTEx (v.V6p) (Battle et al., 2017). We focused on 1,561,655 genetic variants significantly associated with the expression of a nearby gene (<1 MB) and in the 95% credible set for that gene in at least one tissue, based on Bayesian fine mapping with CAVIAR (Hormozdiari et al., 2014); i.e., the set of variants with 95% likelihood to contain the causal eSNPs for the gene. Across all eQTL and footprint tissues, we found that 163,330 of these 1,561,655 variants intersected a TF binding site from our footprint database (TFBS-eQTLs). Counts of TFBS-eQTLs in individual tissues from our footprint database ranged from 743 (urinary bladder) to 71,692 (extra-embryonic structure) (Figure 4A). We tested whether this overlap was greater than expected by chance by mapping footprints to all 11,959,406 genotyped and imputed variants in the GTEx V6p dataset, followed by resampling permutations. We found significant enrichments ($p < 0.001$) for all 27 footprint

tissue × 44 eQTL tissue combinations. The overlap of footprints and eQTLs in mismatched tissues likely reflects that many of the strongest footprints and eQTLs are detected in multiple tissues (Battle et al., 2017). Sample size differs dramatically between tissues both in our footprint database and in GTEx, making it difficult to discern biologically relevant tissue-specific effects. Therefore, in subsequent analyses, we considered all eQTLs together, regardless of the tissue in which they were discovered.

We also determined whether eQTL SNPs with the highest likelihood of being causal variants from linkage-disequilibrium (LD)-based fine mapping with CAVIAR were also the most likely to overlap a footprint. eQTL variants that overlapped a footprint had higher posterior probabilities for being causal than eQTL variants that did not overlap footprints ($t = -61.4$, $p << 1e{-}308$). Indeed, we detected a strong positive association between a variant's posterior probability of being causal and the strength of enrichment for footprints that was consistent across footprints from all 27 tissues (Figure 4B). Focusing on the 3,193 eQTL variants with posterior probabilities > 0.8, we found that 29.2% (932) overlap a footprint. Resampling permutations indicated that this overlap for tissue-specific footprints is ~10- to 40-fold greater than expected by chance. These results suggest that a large fraction of eQTLs may be explained by causal variants that alter TF binding sites, with many of these effects captured by footprints in our database.

Tissue-Specific Footprints Are Enriched for Disease-Associated SNPs

Finally, we tested the hypothesis that high-scoring footprints are enriched for genetic variants associated with disease risk. To address this question, we studied genome-wide summary statistics from well-powered GWAS of eight immune-related traits and 27 psychiatric, behavioral, and cognitive traits (STAR Methods; Table S3). We hypothesized that heritability for immune traits would be specifically associated with footprints in lymphocytes, whereas heritability for neuropsychiatric traits would be specifically associated with footprints in the brain.

When considering all tissue-specific footprints from our database (HINT20 score > 0 in any sample), we found that footprints from brain tissue were strongly enriched for heritability for brain-related traits and footprints from lymphoblasts were strongly enriched for heritability for immune-related traits. However, because most base pairs that are open chromatin have a non-zero footprint score, this result is not distinguishable from previously reported enrichments of heritability in open chromatin. We therefore examined whether footprints with higher scores contributed more to heritability than footprints with lower scores. We used a partitioned heritability approach in which we divided footprints into deciles based on their maximum tissue-specific footprint scores. We found that footprints with the highest scores in brain contributed disproportionately to heritability to brain-related traits but were not strongly associated with immune traits (Figure 5A). Conversely, footprints with the highest scores in lymphoblasts contributed disproportionately and specifically to heritability in immune-related traits (Figure 5B). Interestingly, we also found that positions of open chromatin in the brain that had low footprint scores (bottom decile) contributed disproportionately to risk for brain-related traits. Motif enrichment analyses of the top versus bottom deciles indicated that these segments of open chromatin are enriched

for binding sites for distinct families of TFs. For instance, motifs recognized by several neurodevelopmental TFs (e.g., the LMX family) were disproportionately found in the bottom decile; these neurodevelopmental TFs are known to bind DNA more transiently than other TF classes, leaving a less distinct footprint signature (Baek et al., 2017) (Table S4). Altogether, our results support the hypothesis that the enrichment of disease risk in open chromatin disproportionately results from variants that affect TF binding and indicate that a footprint's score is positively associated with disease risk for many TFs.

DISCUSSION

Here, we have described a uniform workflow for DNase genomic footprinting and generated a comprehensive atlas of TF binding sites in 27 human tissues. We validated these footprints using data from ChIP-seq and eQTL experiments. At optimal thresholds, footprints in our database span 9.8% of the human genome, describing an expansive landscape of tissue-specific TF occupancy. We found strong, tissue-specific enrichments of footprints for disease-associated SNPs from GWAS, demonstrating the utility of our database to characterize gene regulatory mechanisms underlying human disease.

Machine-learning approaches yielded several insights. First, footprinting information improved predictive accuracy compared with a baseline model. Because ChIP-seq is an imperfect gold standard, some footprints with no corresponding ChIP-seq may nonetheless be true binding sites for a TF. Footprinting may identify a broader range of putative binding regions relevant to gene regulation, particularly in light of the strong relationship found with eQTLs. As a future direction, integration of additional epigenomic data could provide additional predictive power to discern active versus inactive binding sites.

We also demonstrated strong enrichments of heritability for complex traits at the highest-scoring footprints, specifically in disease-relevant tissues. Given that most risk variants in GWAS fall within non-coding regions, this finding suggests that disruption of TF binding may be a common mechanism by which genetic risk is conferred. These results build on previous findings that heritability for complex traits is enriched in open chromatin regions. Annotating risk variants with footprint scores improves specificity and mechanistic insight compared with annotating these SNPs based only on chromatin state. This finding demonstrates the utility of our footprint atlas for fine mapping and other systems-level interrogations of complex genetic traits. We found that low-scoring footprints in the brain were highly associated with risk and that these footprints disproportionately contained motifs for developmental TFs. This indicates that caution should be taken when using hard footprint score cutoffs, especially in the brain.

This resource also represents a case study in the development of scalable cloud-based systems for large-scale data analysis (Madduri et al., 2019). The Globus Genomics workflow used to create this resource can readily be extended to new open chromatin datasets and footprinting algorithms as they become available, potentially including newly developed approaches for open chromatin profiling in thousands of single cells. This workflow 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 user-friendly fiat files for all

footprints in this analysis available at http://data.nemoarchive.org/other/grant/sament/sament/footprint_atlas.

STAR★METHODS

RESOURCE AVAILABILITY

Lead Contact—Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Seth Ament (SAment@som.umaryland.edu).

Materials Availability—This study did not generate new unique reagents.

Data and Code Availability—Footprint data files are freely available at http://data.nemoarchive.org/other/grant/sament/sament/footprint_atlas. Code and workflows available at https://github.com/globusgenomics/genomics-footprint.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

This study did not use experimental models.

METHOD DETAILS

Overview—We created and executed footprinting workflows 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 and multi-tool workflows, transparent and elastic scaling on cloud resources, reproducible analysis based on provenance captured using minids and Big Data Bags (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 data unit that encapsulates the entire results of an ENCODE query at a specific point in time. The resulting BDBag is passed as input to various analysis workflows that are executed 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 published workflows (Madduri et al., 2019).

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 a BDBag (Chard et al., 2016, IEEE Big Data, conference presentation). BDBag is a format for defining a dataset and its contents by enumerating the data elements, regardless of their location, and for associating 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 lightweight identifier for uniquely identifying a dataset. Minid and BDBag tools provide mechanisms for exchanging datasets by name, without regard for location or size, and with assurance that the data have not been modified.

ENCODE2Bag Service and Globus Genomics—The ENCODE2Bag service provides a simple web interface for exporting identified, verifiable collections of data from ENCODE. The service when given an ENCODE query, dynamically creates a BDBag that is 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 with their location(s) and checksum(s) for verifying integrity when accessed. The summary of the ENCODE query, represented as a Tab Separated Value file, is included in the BDBag as metadata to track and record provenance. 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 tools abstracts the process by which a BDBag is "materialized."

Globus Genomics is a cloud-hosted web service that enables rapid analysis of large genomics data. The service has over 3000 computationally optimized tools and a collection of best practices analysis workflows. Additionally, we added the data management tools built as part of the BDDS BD2K center to the service to make it easier for researchers to build high performance, reproducible bioinformatics workflows.

Globus provides reliable, secure, and high performance data transfer between Globus "endpoints" (Chard et al., 2014). Globus provides a common interface to a variety of storage systems 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 used 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.

The analysis workflows require only the minid of the input dataset to perform the analysis. The Globus Genomics service uses minid tools to transparently resolve the location of the BDBag, it then uses the BDBag tools to identify the contents of the dataset, and finally uses Globus to transfer the raw files to the cloud-hosted analysis infrastructure.

Scalable workflow for predicting Transcription Factor Binding Sites—In this workflow, we used the above-mentioned tools to materialize the BDBag for each tissue. Each tissue type contained DHS data for multiple samples. In addition, each sample 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. (2014), with the minimum reported size reduced from 500 bases to 400. 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 (Gusmaoet 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 (1591 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 68,771 CPU hours (2149.1 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 from the ENCODE portal (encodeproject.org). 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 each experiment.

The ENCODE data was generated using short reads (< 50 bases), resulting in a high number of potential sequence matches. This 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 386 BAM files for each seed.

Identifying regions of open chromatin—Based on work from Koohy et al. (2014), who compared four different approaches (F-Seq, Hotspot, MACS and ZINBA) we used F-seq (Boyle et al., 2008) to identify regions of open chromatin from the aligned BAM files using the same recommended parameters. 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.

Motif database curation—As footprints from HINT and Wellington are motif agnostic and do not include information on motif matches, we integrated the footprint locations with motifs and motif-transcription factor mappings from JASPAR, HOCOMOCO, UniPROBE, and SwissRegulon. There is considerable redundancy between these databases, which often contain 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 avoid inclusion of redundant motifs, we updated and modified an existing R package, MotifDB (Shannon and Richards, 2017), to include the latest versions of all four databases. We evaluated the similarity of all motifs using Tomtom (Gupta et al., 2007). Those motifs that were significantly different from the 2016 release of JASPAR (-log(p value) ≥ 7.3) were retained, yielding a total of 1,530 motifs. In addition to the mappings provided by each of the aforementioned databases, we also expanded the TF-motif mappings to incorporate families of TFs with very similar DNA sequence specificity, using information from TFClass (Wingender et al., 2015). The

complete mapping can be accessed through MotifDB by calling the "associateTranscriptionFactors" method. The number of original motifs considered for each database and the number of motifs and transcription factor mappings retained after filtering are found in Table S1.

Collectively, our aggregated collection 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. Two representative examples of these mappings are found in Figure S2. An entire map of all motifs and TFs can be found in the Table S2. 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 was 13,242.

Combining footprints with database of motifs—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.

ChIP-seq validation and machine learning models—We joined all footprints based upon location in the genome to create one unified dataset per tissue. 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 we summed the number of footprints for each location, we used the highest number of occurrences 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 validating the model, we designated chromosomes 2 and 4 as a hold-out set that was left untouched until the very end after all model parameter sets had been tested. 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 two classes of models: 1) a basic logistic regression model, and 2) a gradient boosted model, which aggregates an ensemble of decision trees to learn a nonlinear decision boundary. 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 the ensemble, but on each feature

individually, in order to get an idea of which features were most predictive of ChIP-seq hits. 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., 2018). We used the R package XGBoost to create this model using a maximum tree depth of 7, 200 rounds of boosting, and a logistic regression optimization criterion (Chen and Guestrin, 2016).

One challenge that we encountered in creating this model is that 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 problematic in the setting of this machine-learning format, as it increases memory requirements significantly and results in a poor signal-to-noise ratio. In order to address this issue in our training set, we sampled 20 million hits of 264 motifs, combined these motif hits with our lymphoblast footprints, then filtered 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 unbalanced sample sizes in the two classes being compared (Boughorbel et al., 2017).

eQTL Enrichment—Expression quantitative trait loci (eQTLs) from the Genotype Tissue Expression Consortium (GTEx; V6p 95% credible causal sets) (Battle et al., 2017) were downloaded from the UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database) on January 5, 2018. In addition, as a background set, we downloaded the table of all 11,959,406 genotyped and imputed variants from the GTEx V6p dataset ("GTEx_Analysis_2015-01-12_OMNI_2.5M_5M_450Indiv_chr1-22-X_genot_imput_info04_maf01_HWEp1E6_variant_id_lookup.txt.gz") from the GTEx web portal (https://www.gtexportal.org/home/; accessed March 16, 2018). GTEx variants were converted to hg38 coordinates using the UCSC Genome Browser's liftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver) with default parameters. We identified TF binding site-altering variants by intersecting the locations of GTEx variants with the locations of TF binding sites from DNase-seq footprinting, using the genomic coordinates of motifs that overlap a footprint with a HINT score > = 200. Statistical associations between footprints and eQTL posterior probabilities were calculated using the t.test() function in R. Statistical significance for overlap between variants that alter TF binding sites and variants that are eQTLs was calculated from 1,000 re-sampling permutations, drawing variants at random from the complete set of genetic variants in GTEx V6p.

Partitioned Heritability Analysis—We utilized a partitioned heritability approach to characterize the relationship between footprint confidence scores and relevant phenotypes. First, we divided all the pooled footprints in a given tissue type into decile bins based on the score assigned to the best HINT20 score (1 = lowest scores, 10 = highest scores). We then used portioned LD Score Regression (LDSC) (Finucane et al., 2015) to assess each decile's contribution to heritability for several disease traits. The immune traits assessed were ulcerative colitis, type 1 diabetes, rheumatoid arthritis, primary biliary cirrhosis, multiple

sclerosis, lupus, Crohn's disease, and celiac disease. The neuropsychiatric traits included educational attainment, neuroticism, schizophrenia, and bipolar disorder, as well as 23 additional brain-related traits taken from the top 100 most heritable traits in the UK Biobank (Table S3) (Bentham et al., 2015; Bradfield et al., 2011; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Cordell et al., 2015; Dubois et al., 2010; Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011; Jostins et al., 2012; Okada et al., 2014; Okbay et al., 2016; Sawcer et al., 2011). The top and bottom brain deciles were compared using a chi-square test, and we used the residuals to determine over- and under-represented TFs in both deciles.

QUANTIFICATION AND STATISTICAL ANALYSIS

DNase-I genomic footprints were identified with HINT and Wellington. Thresholds for selecting high-quality footprints were evaluated via a gradient boosting model, comparing footprint locations to true-positive TF binding sites from ChIP-seq. Overlap of footprints with eQTLs was evaluated with bootstrap permutations and t tests. Enrichment of footprints for SNPs associated with risk for human traits was calculated with stratified LD score regression. Details are provided in the Method Details section, above.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported by the BDDS Center of the NIH Big Data to Knowledge program (U54 EB020406 to N.D.P. and L.H.), the National Institute of Aging (U01 AG046139 to N.D.P.; RF1 AG057443 to N.D.P.), the National Institute for General Medical Sciences Center for Systems Biology at the Institute for Systems Biology (P50 GM07654 to L.H. and N.D.P.), and the national Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative (R24MH114788, Owen White, PI; R24MH114815, Owen White and Ronna Hertzano, PIs; the National Human Genome Research Institute (5R01HG009018 to I.F. and R.M.,); and the National Institute of Mental Health (F30 MH120910 to A.M.C.).

REFERENCES

Ament SA, Pearl JR, Cantle JP, Bragg RM, Skene PJ, Coffey SR, Bergey DE, Wheeler VC, MacDonald ME, Baliga NS, et al. (2018).Transcriptional regulatory networks underlying gene expression changes in Huntington's disease. Mol. Syst. Biol 14, e7435. [PubMed: 29581148]

Baek S, Goldstein I, and Hager GL (2017). Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity. Cell Rep. 19, 1710–1722. [PubMed: 28538187]

Battle A, Brown CD, Engelhardt BE, and Montgomery SBGTEx Consortium; Laboratory, Data Analysis & Coordinating Center(LDACC)–Analysis Working Group; Statistical Methods groups–Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site–NDRI; Biospecimen Collection Source Site–RPCI; Biospecimen Core Resource–VARI; Brain Bank Repository–University of Miami Brain Endowment Bank; Leidos Biomedical–Project Management; ELSI Study; Genome Browser Data Integration & Visualization–EBI; Genome Browser Data Integration & Visualization–UCSC Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis & Coordinating Center (LDACC); NIH program management; Biospecimen collection; Pathology; eQTL manuscript working group (2017). Genetic effects on gene expression across human tissues. Nature 550, 204–213. [PubMed: 29022597]

Bentham J, Morris DL, Graham DSC, Pinder CL, Tombleson P, Behrens TW, Martin J, Fairfax BP, Knight JC, Chen L, et al. (2015). Genetic association analyses implicate aberrant regulation of

innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat. Genet 47, 1457–1464. [PubMed: 26502338]

Boughorbel S, Jarray F, and El-Anbari M (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS ONE 12, e0177678. [PubMed: 28574989]

Boyle AP, Guinney J, Crawford GE, and Furey TS (2008). F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics 24, 2537–2538. [PubMed: 18784119]

Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, Iyer VR, Crawford GE, and Furey TS (2011). High-resolution genome-wide *in vivo* footprinting of diverse transcription factors in human cells. Genome Res. 21, 456–464. [PubMed: 21106903]

Bradfield JP, Qu HQ, Wang K, Zhang H, Sleiman PM, Kim CE, Mentch FD, Qiu H, Glessner JT, Thomas KA, et al. (2011). A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. PLoS Genet. 7, e1002293. [PubMed: 21980299]

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, and Guestrin C (2016). XGBoost: A Scalable Tree Boosting System. arXiv, arXiv:1603.02754 http://arxiv.org/abs/1603.02754.

Claussnitzer M, Dankel SN, Kim KH, Quon G, Meuleman W, Haugen C, Glunk V, Sousa IS, Beaudry JL, Puviindran V, et al. (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N. Engl. J. Med 373, 895–907. [PubMed: 26287746]

Cordell HJ, Han Y, Mells GF, Li Y, Hirschfield GM, Greene CS, Xie G, Juran BD, Zhu D, Qian DC, et al.; Canadian-US PBC Consortium; Italian PBC Genetics Study Group; UK-PBC Consortium (2015). International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. Nat. Commun 6, 8019. [PubMed: 26394269]

Cuellar-Partida G, Buske FA, McLeay RC, Whitington T, Noble WS, and Bailey TL (2012). Epigenetic priors for identifying active transcription factor binding sites. Bioinformatics 28, 56–62. [PubMed: 22072382]

de la Torre-Ubieta L, Stein JL, Won H, Opland CK, Liang D, Lu D, and Geschwind DH (2018). The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. Cell 172, 289–304.e18. [PubMed: 29307494]

Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, et al. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482, 390–394. [PubMed: 22307276]

Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, Zhernakova A, Heap GA, Adány R, Aromaa A, et al. (2010). Multiple common variants for celiac disease influencing immune gene expression. Nat. Genet 42, 295–302. [PubMed: 20190752]

ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–640. [PubMed: 15499007]

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74. [PubMed: 22955616]

Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, Jostins L, Plant K, Andrews R, McGee C, and Knight JC (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science 343, 1246949. [PubMed: 24604202]

Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, Anttila V, Xu H, Zang C, Farh K, et al.; ReproGen Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; RACI Consortium (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet 47, 1228–1235. [PubMed: 26414678]

Galas DJ, and Schmitz A (1978). DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res. 5, 3157–3170. [PubMed: 212715]

Grant CE, Bailey TL, and Noble WS (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018. [PubMed: 21330290]

Gupta S, Stamatoyannopoulos JA, Bailey TL, and Noble WS (2007). Quantifying similarity between motifs. Genome Biol. 8, R24. [PubMed: 17324271]

Gupta RM, Hadaya J, Trehan A, Zekavat SM, Roselli C, Klarin D, Emdin CA, Hilvering CRE, 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. [PubMed: 28753427]

Gusev A, Lee SH,Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, Zang C, Ripke S, Bulik-Sullivan B, Stahl E, et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium; SWE-SCZ Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; SWE-SCZ Consortium (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet 95, 535–552. [PubMed: 25439723]

Gusmao EG, Dieterich C, Zenke M, and Costa IG (2014). Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. Bioinformatics 30, 3143–3151. [PubMed: 25086003]

Handel AE, Gallone G, Zameel Cader M, and Ponting CP (2017). Most brain disease-associated and eQTL haplotypes are not located within transcription factor DNase-seq footprints in brain. Hum. Mol. Genet 26, 79–89. [PubMed: 27798116]

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al. (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 34, D590–D598. 10.1093/nar/gkj144. [PubMed: 16381938]

Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, and Eskin E (2014). Identifying causal variants at loci with multiple signals of association. Genetics 198, 497–508. [PubMed: 25104515]

Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA, et al.; International IBD Genetics Consortium (IIBDGC) (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491, 119–124. [PubMed: 23128233]

Kähärä J, and Lähdesmäki H (2015). BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data. Bioinformatics 31, 2852–2859. [PubMed: 25957350]

Koohy H, Down TA, Spivakov M, and Hubbard T (2014). A comparison of peak callers used for DNase-Seq data. PLoS ONE 9, e96303. [PubMed: 24810143]

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, and Carey VJ (2013). Software forcomputing and annotating genomic ranges. PLoS Comput. Biol 9, e1003118 10.1371/journal.pcbi.1003118. [PubMed: 23950696]

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, and Durbin R; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. [PubMed: 19505943]

Madduri RK, Sulakhe D, Lacinski L, Liu B, Rodriguez A, Chard K, Dave UJ, and Foster IT (2014). Experiences building Globus Genomics: a next-generation sequencing analysis service using Galaxy, Globus, and Amazon Web Services. Concurr. Comput 26, 2266–2279. [PubMed: 25342933]

Madduri R, Chard K, D'Arcy M, Jung SC, Rodriguez A, Sulakhe D, Deutsch E, Funk C, Heavner B, Richards M, et al. (2019). Reproducible big data science: A case study in continuous FAIRness. PLoS ONE 14, e0213013. [PubMed: 30973881]

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E,Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190–1195. [PubMed: 22955828]

Moyerbrailean GA, Kalita CA, Harvey CT, Wen X, Luca F, and Pique-Regi R (2016). Which Genetics Variants in DNase-Seq Footprints Are More Likely to Alter Binding? PLoS Genet. 12, e1005875. [PubMed: 26901046]

Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, and Stamatoyannopoulos JA (2012a). Circuitry and dynamics of human transcription factor regulatory networks. Cell 150, 1274–1286. [PubMed: 22959076]

Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, et al. (2012b). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489, 83–90. [PubMed: 22955618]

Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, Kochi Y, Ohmura K, Suzuki A, Yoshida S, et al.; RACI consortium; GARNET consortium (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506, 376–381. [PubMed: 24390342]

Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA, Turley P, Chen GB, Emilsson V, Meddens SF, et al.; LifeLines Cohort Study (2016). Genome-wide association study identifies 74 loci associated with educational attainment. Nature 533, 539–542. [PubMed: 27225129]

Olson R, La Cava W, Mustahsan Z, Varik A, and Moore JH (2018). Data-driven Advice for Applying Machine Learning to Bioinformatics Problems. Pac. Symp. Biocomput. 23, 192–542203. [PubMed: 29218881]

Pearl JR, Colantuoni C, Bergey DE, Funk CC, Shannon P, Basu B, Casella AM, Oshone RT, Hood L, Price ND, and Ament SA (2019). Genome-Scale Transcriptional Regulatory Network Models of Psychiatric and Neurodegenerative Disorders. Cell Syst. 8, 122–135.e7. [PubMed: 30772379]

Piper J, Elze MC, Cauchy P, Cockerill PN, 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. [PubMed: 24071585]

Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, and Pritchard JK (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. 21, 447–455. [PubMed: 21106904]

Psychiatric GWAS Consortium Bipolar DisorderWorking Group (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat. Genet. 43, 977–983. [PubMed: 21926972]

Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, Moutsianas L, Dilthey A, Su Z, Freeman C, Hunt SE, et al.; International Multiple Sclerosis Genetics Consortium; Wellcome Trust Case Control Consortium 2 (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476, 214–219. [PubMed: 21833088]

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature 511, 421–427. [PubMed: 25056061]

Schwessinger R, Suciu MC, McGowan SJ, Telenius J, Taylor S, Higgs DR, and Hughes JR (2017). Sasquatch: predicting the impact of regulatory SNPs on transcription factor binding from cell- and tissue-specific DNase footprints. Genome Res. 27, 1730–1742. [PubMed: 28904015]

Shannon P, and Richards M (2017). MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs (Version R package version 1.20). Bio-conductor. https://bioconductor.org/packages/release/bioc/html/MotifDb.html.

Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, and Gifford DK (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat. Biotechnol 32, 171–178. [PubMed: 24441470]

Sung MH, Guertin MJ, Baek S, and Hager GL (2014). DNase footprint signatures are dictated by factor dynamics and DNA sequence. Mol. Cell 56, 275–285. [PubMed: 25242143]

Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, Andersen KG, Mikkelsen TS, Lander ES, Schaffner SF, and Sabeti PC (2016). Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. Cell 165, 1519–1529. [PubMed: 27259153]

Ward LD, and Kellis M (2016). HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Res. 44 (D1), D877–D881. [PubMed: 26657631]

Wingender E, Schoeps T, Haubrock M, and Dönitz J (2015). TFClass: a classification of human transcription factors and their rodent orthologs. Nucleic Acids Res. 43, D97–D102. [PubMed: 25361979]

Yardımcı GG, Frank CL, Crawford GE, and Ohler U (2014). Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. Nucleic Acids Res. 42, 11865–11878. [PubMed: 25294828]

Zaharia M, Bolosky WJ, Curtis K, Fox A, Patterson D, Shenker S, Stoica I, Karp RM, and Sittler T (2011). Faster and More Accurate Sequence Alignment with SNAP. arXiv, arXiv:1111.5572 https://arxiv.org/abs/1111.5572.

Highlights

- Comprehensive map of TF occupancy in human tissues from DNase-seq footprints
- Footprints contain genetic variants associated with changes in gene expression
- Tissue-specific associations of footprints with genetic risk for complex traits

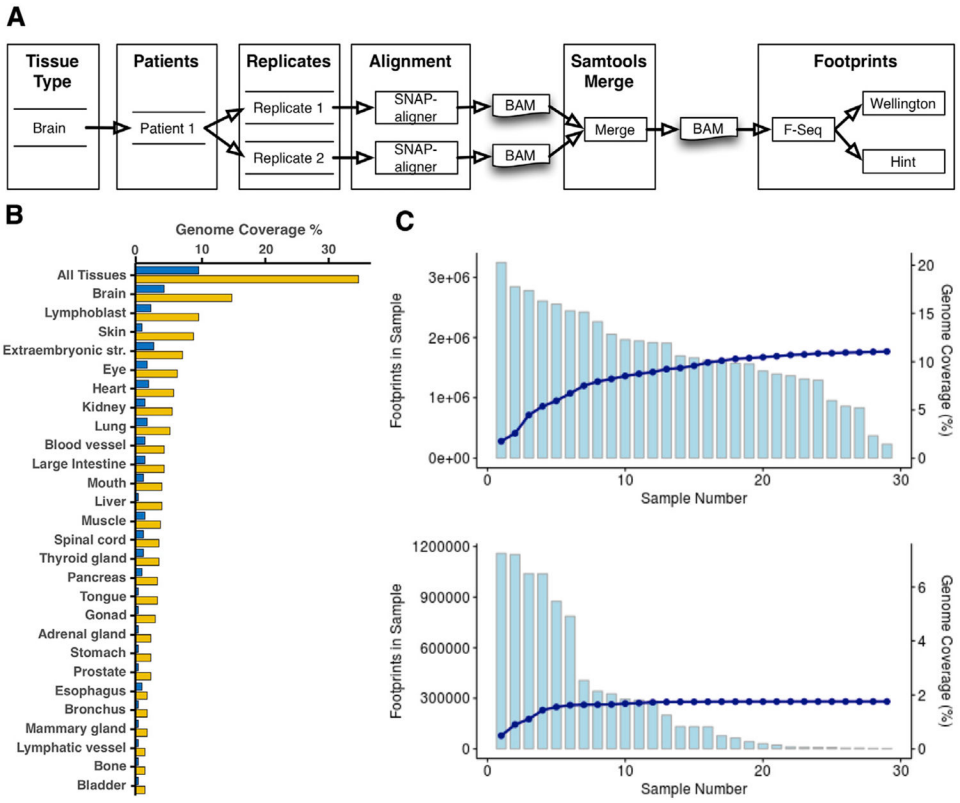


Figure 1. Footprint Atlas Workflow and Coverage Statistics

(A) Footprints workflow overview. Each tissue type can have multiple quantities 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.

(B) Percentage of the genome covered by the footprints for each tissue type and all tissues. Yellow is without filtering, and dark blue is filtering with HINT score > 200 and Wellington score < −27 (each method has its own scale and distribution).

(C) Footprints from the brain for HINT20 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 B). The dark blue line represents the cumulative percentage of the genome covered.

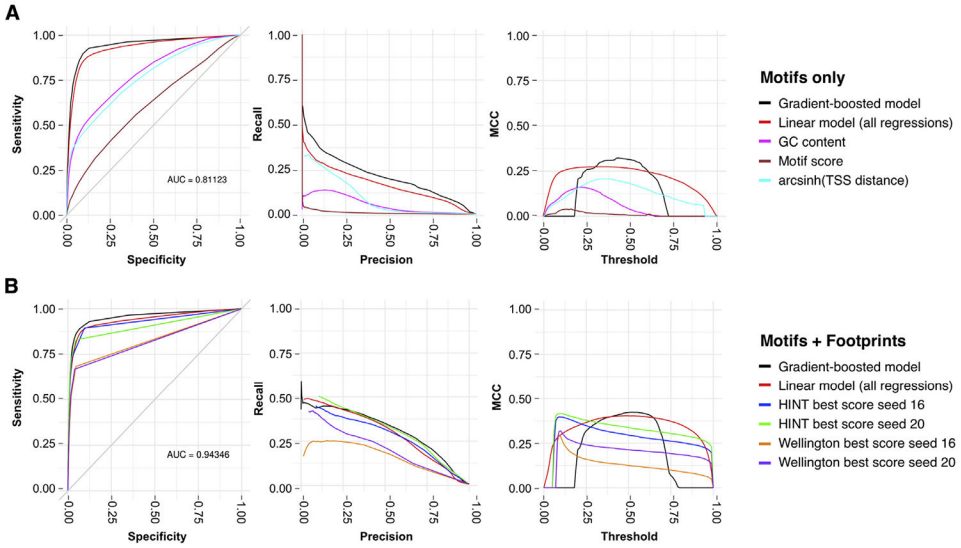


Figure 2. Predictive Performance on a Held-Out Test Set of a Gradient-Boosted Decision Tree (GBDT) Model of the 62 TFs (264 Motifs) in the ENCODE-Generated ChIP-Seq Samples

We compare with baseline models that use only motif information, TSS distance, and GC content and to a linear model that uses all of these.

(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 and Seed20 alignments using all aforementioned features, footprint scores, and footprint metrics. The GBDT model obtains the best performance by nearly all metrics, though the amount by which it outperforms the linear model on the footprint data is in some cases marginal enough that an interpretable linear model may be preferred for some applications.

The threshold in the third column refers to the decision boundary at which the continuous output of the models, which varies between zero and one, is thresholded and a classification decision is made. The aggregate models obtain good performance over a relatively wide range of thresholds compared with the models using individual methods.

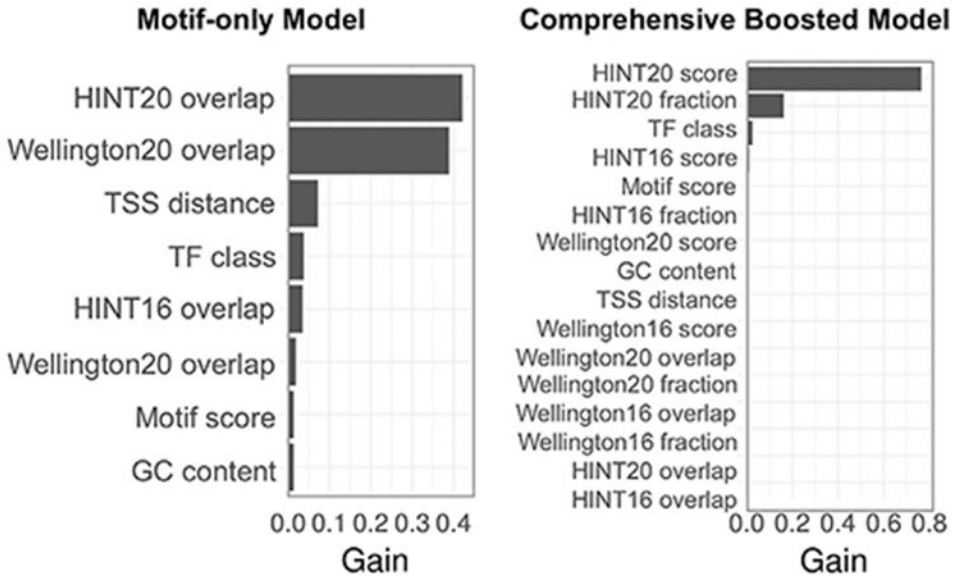


Figure 3. Importance Matrix Quantifying the Contribution of Each Feature when Trained and Tested on the ENCODE ChIP-Seq Dataset for 62 TFs

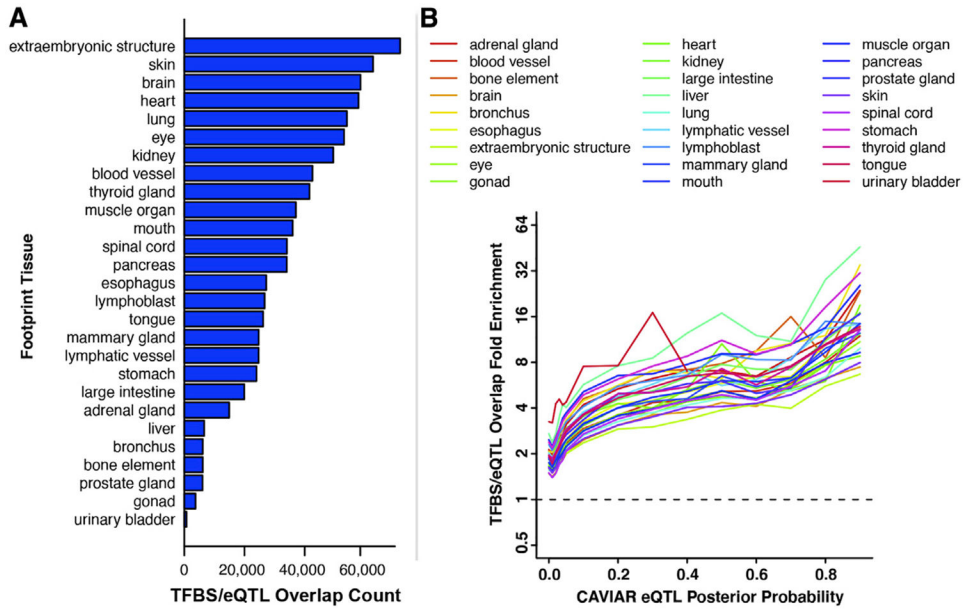


Figure 4. Footprints Overlap with Genetic Variants that Affect Gene Expression

(A) Counts of eSNPs overlapping predicted TF binding sites across all DHS tissues. Barplots indicate the total number of eSNPs overlapping footprints across all GTEx tissues.

(B) Comparison of CAVIAR eQTL scores with the fold enrichment for TFBS-eQTLs.

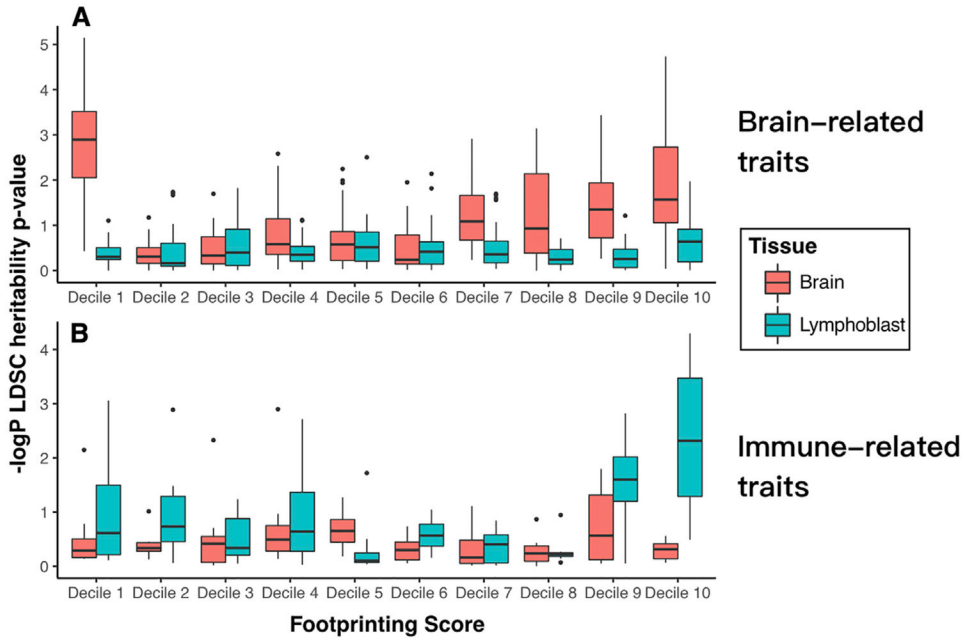


Figure 5. Partitioned Heritability of Tissue-Specific Footprints in Related GWAS by Footprint Confidence Score Decile

(A) Partitioned heritability of brain footprints by decile in 27 summarized brain-related traits. Box plots indicate the median and interquartile range of −log10 (p values) across the 27 traits.

(B) Heritability of lymphoblast footprint deciles in 8 summarized immune-related traits. Decile 1, lowest scores; decile 10, highest scores.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Footprint BED files	This paper	http://data.nemoarchive.org/other/grant/sament/sament/footprint_atlas/bed/
Footprint extended TSV files	This paper	http://data.nemoarchive.org/other/grant/sament/sament/footprint_atlas/extended/
ENCODE DNase-seq	https://www.encodeproject.org/	RRID: SCR_015482. For specific accessions/experiments, see http://data.nemoarchive.org/other/grant/sament/extended/
Expression quantitative trait loci (eQTLs)	Battle et al., 2017	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/
Background genotyped variants	Battle et al., 2017	https://storage.googleapis.com/gtex_analysis_v6/reference/GTEx_Analysis_2015-01-12_OMNI_2.5M_5M_450Indiv_chr1-22-X_genot_imput_info04_maf01_HWEp1E6_variant_id_lookup.txt.gz
Educational attainment GWAS	Okbay et al., 2016	https://data.broadinstitute.org/alkesgroup/sumstats_formatted/PASS_Years_of_Education2.sumstats
Schizophrenia GWAS	Schizophrenia Working Group Working Group of the Psychiatric Genomics Consortium, 2014	https://data.broadinstitute.org/alkesgroup/sumstats_formatted/PASS_Schizophrenia.sumstats
Neuroticism GWAS	Okbay et al., 2016	https://data.broadinstitute.org/alkesgroup/sumstats_formatted/PASS_Neuroticism.sumstats
Bipolar Disorder GWAS	Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011	https://data.broadinstitute.org/alkesgroup/sumstats_formatted/PASS_Bipolar_Disorder.sumstats
Alcohol intake frequency GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/bi7t4rekkhpa4ks/1558.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Alcohol usually taken with meals GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/ra3tw6s1kw31ywn/1618.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Age completed full time education GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/o1o31tevhou822f/845.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Job involves heavy manual or physical work GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/zswrzp8s19j28sz/816.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Job involves mainly walking or standing GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/de1u2yul4cffb1i/806.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Age at first live birth GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/bqjiqclxdb195d6/2754_irnt.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Fluid intelligence score GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/t3lrfj1id8133sx/20016_irnt.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Fedup feelings GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/kv6ltvrmkrugfy1/1960.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Frequency of tiredness lethargy in last 2 weeks GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/usoitcixaa39gtw/2080.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Mood swings GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/q4yv2y5u07z7qc6/1920.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Seen doctor GP for nerves anxiety tension or depression GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/ow0kr506vn2fiox/2090.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Number of incorrect matches in round GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/nb54tjjsoijf2x4/399_irnt.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Mean time to correctly identify matches GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/ysx8s20g8la9lm1/20023_irnt.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Daytime dozing sleeping narcolepsy GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/chtk02fbduvzv3r/1220.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Getting up in morning GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/0v2exws5j8z4yo7/1170.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Morningevening person chronotype GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/h6vnprrobkdia2d/1180.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Nap during day GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/q8fynq2rnkgttoi/1190.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Sleep duration GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/7tgjc9s68gp9d5a/1160.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Sleeplessness insomnia GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/jeolythrs18jk9p/1200.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Snoring GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/dvmbzveuc0htuj3/1210.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Current tobacco smoking GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/nwqshg5soaayh03/1239.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Ever smoked GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/2vxlmq7q7ozxgf9/20160.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Past tobacco smoking GWAS	http://www.nealelab.is/uk-biobank/	https://www.dropbox.com/s/29b2w1qc9erzlo4/1249.gwas.imputed_v3.both_sexes.tsv.bgz?dl=0
Celiac GWAS	Dubois et al., 2010	https://data.broadinstitute.org/alkesgroup/sumstats_formatted/PASS_Celiac.sumstats
Lupus GWAS	Bentham et al., 2015	https://data.broadinstitute.org/alkesgroup/sumstats_formatted/PASS_Lupus.sumstats
Primary biliary cirrhosis GWAS	Cordell et al., 2015	https://data.broadinstitute.org/alkesgroup/sumstats_formatted/PASS_Primary_biliary_cirrhosis.sumstats
Type 1 Diabetes GWAS	Bradfield et al., 2011	https://data.broadinstitute.org/alkesgroup/sumstats_formatted/PASS_Type_1_Diabetes.sumstats
Crohns Disease GWAS	Jostins et al., 2012	https://data.broadinstitute.org/alkesgroup/sumstats_formatted/PASS_Crohns_Disease.sumstats
Multiple sclerosis GWAS	Sawcer et al., 2011	https://data.broadinstitute.org/alkesgroup/sumstats_formatted/PASS_Multiple_sclerosis.sumstats
Rheumatoid Arthritis GWAS	Okada et al., 2014	https://data.broadinstitute.org/alkesgroup/sumstats_formatted/PASS_Rheumatoid_Arthritis.sumstats
Ulcerative Colitis GWAS	Jostins et al., 2012	https://data.broadinstitute.org/alkesgroup/sumstats_formatted/PASS_Ulcerative_Colitis.sumstats
Human reference genome NCBI build 38, GRCh37	Genome Reference Consortium	https://www.ncbi.nlm.nih.gov/grc/human
Software and Algorithms		
SNAP	Zaharia et al., 2011	http://snap.cs.berkeley.edu/
Samtools	Li et al., 2009	http://samtools.sourceforge.net/
F-Seq	Boyle et al., 2008	http://fureylab.web.unc.edu/software/fseq/
FIMO	Grant et al., 2011	http://meme-suite.org/doc/download.html
HINT	Gusmao et al., 2014	http://www.regulatory-genomics.org/hint/introduction/
Wellington	Piper et al., 2013	https://pythonhosted.org/pyDNase/
Tomtom	Gupta et al., 2007	http://meme-suite.org/doc/download.html

REAGENT or RESOURCE	SOURCE	IDENTIFIER
GenomicRanges	Lawrence et al., 2013	https://bioconductor.org/packages/release/bioc/html/GenomicRanges.html
XGBoost	Chen and Guestrin, 2016	https://xgboost.readthedocs.io/en/latest/
LiftOver	Hinrichs et al., 2006	https://genome-store.ucsc.edu
LDSC	Finucane et al., 2015	https://github.com/bulik/ldsc
Other		
JASPAR	MotifDB	http://jaspar.genereg.net/
HOCOMOCO	MotifDB	https://hocomoco11.autosome.ru/
UniPROBE	MotifDB	http://thebrain.bwh.harvard.edu/uniprobe/
SwissRegulon	MotifDB	http://www.swissregulon.unibas.ch/
BDBags	Madduri et al., 2019	https://github.com/fair-research/bdbag
ENCODE2Bag	Madduri et al., 2019	https://github.com/fair-research/encode2bag