Genome Biology

RESEARCH

Open Access

Benchmarking and building DNA binding affinity models using allele-specific and allele-agnostic transcription factor binding data

Xiaoting Li[1†], Lucas A. N. Melo[1†] and Harmen J. Bussemaker[1,2*]

[†]Xiaoting Li and Lucas A. N. Melo contributed equally to this work.

*Correspondence: hjb2004@columbia.edu

[1] Department of Biological Sciences, Columbia University, New York, NY 10027, USA
[2] Department of Systems Biology, Columbia University, New York, NY 10032, USA

Abstract

Background: Transcription factors (TFs) bind to DNA in a highly sequence-specific manner. This specificity manifests itself in vivo as differences in TF occupancy between the two alleles at heterozygous loci. Genome-scale assays such as ChIP-seq currently are limited in their power to detect allele-specific binding (ASB) both in terms of read coverage and representation of individual variants in the cell lines used. This makes prediction of allelic differences in TF binding from sequence alone desirable, provided that the reliability of such predictions can be quantitatively assessed.

Results: We here propose methods for benchmarking sequence-to-affinity models for TF binding in terms of their ability to predict allelic imbalances in ChIP-seq counts. We use a likelihood function based on an over-dispersed binomial distribution to aggregate evidence for allelic preference across the genome without requiring statistical significance for individual variants. This allows us to systematically compare predictive performance when multiple binding models for the same TF are available. To facilitate the de novo inference of high-quality models from paired-end in vivo binding data such as ChIP-seq, ChIP-exo, and CUT&Tag without read mapping or peak calling, we introduce an extensible reimplementation of our biophysically interpretable machine learning framework named PyProBound. Explicitly accounting for assay-specific bias in DNA fragmentation rate when training on ChIP-seq yields improved TF binding models. Moreover, we show how PyProBound can leverage our threshold-free ASB likelihood function to perform de novo motif discovery using allele-specific ChIP-seq counts.

Conclusion: Our work provides new strategies for predicting the functional impact of non-coding variants.

Keywords: Gene expression regulation, Non-coding variants, Transcription factors, Allele-specific binding, ChIP-seq, ChIP-exo, CUT&Tag, CTCF, EBF1, PU.1/SPI1, Motif discovery, Biophysically interpretable machine learning, Statistical modeling

Background

Transcription factors (TFs) bind to DNA and regulate transcription in a sequence-specific manner [1]. Genome-wide association studies (GWAS) have shown that the majority of significantly associated variants are located in non-coding regions and may therefore impact gene regulation [2]. One of the key drivers of phenotypic variation is variable TF-DNA binding [3], which is commonly caused by the disruption of TF binding sites by genetic variants.

The effect of genetic variation on TF binding can be investigated through an allele-specific approach, which can assess allelic differences in functional genomic readouts directly at heterozygous loci, thus better controlling for environmental differences between individuals and cell types. For instance, allele-specific binding (ASB) can be detected as a statistically significant imbalance between the number of mapped ChIP-seq reads containing the respective alleles of a single-nucleotide variant (SNV) [4]. AlleleDB [5] is a resource that provides ASB annotations based on the 1000 Genome Project.

Allelic preference at single-nucleotide variants (SNVs) may arise from the alteration of TF binding sites and therefore of TF occupancy [6]. In support of this, the AlleleDB [5] and ADASTRA [7] studies both reported concordance between ASB calls and motif disruption. However, an important fraction of ASB instances cannot be explained in terms of direct alteration of TF binding affinity [8, 9]. In these cases, the observed allelic imbalance may be due to variation in indirect binding via co-factors or in local chromatin accessibility. Prediction of SNV effects on TF binding is also limited by the quality of the TF binding models, especially for weaker sites [10].

Quantitative knowledge about a TF's ability to bind to specific DNA sequences is essential to understanding its function. TF binding models or "motifs" can be used to identify potential binding sites by scanning DNA sequences of cis-regulatory regions such as promoters and enhancers [1]. New techniques for probing TF-DNA interactions have greatly expanded our knowledge of TF binding specificity. High-throughput in vitro assays can readily characterize intrinsic TF binding preferences on a large scale. SELEX (systematic evolution of ligands by exponential enrichment) is one of the most widely used in vitro assays, assessing binding of a purified TF protein across a large pool of random DNA ligands through affinity-based selection over multiple rounds [11]. High-throughput SELEX data are available for hundreds of TFs [12, 13].

Various high-throughput in vivo assays can be used to probe TF binding landscapes under a specific biological condition in a particular cell type. Chromatin immunoprecipitation coupled with deep DNA sequencing (ChIP-seq) [14], in which proteins are crosslinked to their DNA binding sites before fragmentation and immunoprecipitation, is the most widely used assay. Thanks to collective efforts such as ENCODE, ChIP-seq data are available for hundreds of TFs and various cell types [15]. Other popular in vivo assays for probing TF binding include ChIP-exo [16] which adds an exonuclease digestion step and has several technical variants, and CUT&Tag [17] which does not require crosslinking.

Many computational studies have attempted to improve the prediction of genetic variant effects on TF binding. ProBound, a flexible machine learning framework recently developed in our lab [18], directly fits a biophysical model to multiple rounds of SELEX

data, while accounting for non-specific binding, dependencies between nucleotide positions, and multiple binding modes. ProBound can systematically and consistently analyze data from different types of SELEX experiments, and is capable of identifying low-affinity sites and capturing the impact of co-factors and DNA methylation. DNA binding models learned from high-throughput SELEX data using ProBound generally outperform binding motifs from other resources—including JASPAR [19], DeepBind [20], and HOCOMOCO [21]—when predicting in vivo DNA occupancy [18]. This suggested that ProBound may also have great potential for predicting the impact of genetic variation on TF-DNA binding, which is the topic of the present study.

Our previous study [18] included proof-of-concept that ProBound can be used to analyze single-end ChIP-seq data in a way that avoids peak calling, which ignores information from weakly bound regions. A DNA binding model inferred from ChIP-seq data for the glucocorticoid receptor was quantitatively consistent with a model derived from SELEX data for the same TF. The original implementation of ProBound however was not designed to able to handle DNA libraries with variable sequence lengths, which precluded optimal analysis of paired-end ChIP-seq data.

For this study, we created PyProBound, a machine learning framework based on PyTorch that reimplements the methodology of [18] in a more flexible and modular manner, and at the same time is fully backwards compatible with the original Java implementation of ProBound. We use PyProBound to learn TF binding models from paired-end ChIP-seq, ChIP-exo, and CUT&Tag data without the need for any peak calling. It is in fact not even necessary to map any reads to the genome if the DNA fragments defined by the read pairs are long enough. We do find that our peak-free approach requires explicit modeling of assay-specific technical biases, specifically the local sequence context dependence of the DNA fragmentation rate during sonication. However, as we will show, this is straightforward to implement within the PyProBound framework.

Using the widely studied human transcription factors CTCF, EBF1, and PU.1 (a.k.a. SPI1) as examples, we compare sequence-to-affinity models derived from various types of TF binding data in terms of their ability to predict the impact of genetic variation on TF binding. To this end, we construct a likelihood function that can quantify, on a genome-wide scale, to what extent allelic preference can be explained from DNA sequence alone. It does so without the need to make any calls of ASB at the level of individual variants. We show that the same genome-wide likelihood function can be leveraged to perform de novo motif discovery directly on allele-aware binding data using PyProBound. Taken together, our results underscore both the usefulness of resources such as AlleleDB and the extensibility of (Py)ProBound.

Results

Prediction of CTCF allele-specific binding events using sequence-to-affinity models

To assess to what extent sequence-to-affinity models can predict allele-specific binding effects, we used human ASB annotations from AlleleDB [5] and predicted the SNV's effect on TF binding affinity. We chose to focus on the insulator protein CTCF due to the abundance of variants (2231 SNVs) with statistically significant evidence of ASB for this factor [5]. The MotifCentral.org database contains hundreds of ProBound models trained on in vitro binding data from HT-SELEX [12] and SMiLE-seq [13] assays. We

used the MotifCentral model for CTCF to predict allele specific binding at heterozygous SNV loci that were previously found to have significant allelic bias in ChIP-seq coverage (Fig. 1A). For each SNV observed to have ASB, we predicted cumulative CTCF binding affinity from the DNA sequence around the SNV by summing over all offsets of the TF-DNA binding interface relative to the SNV, separately for the reference and the alternative allele (Fig. 1B; see "Methods" for details).

A key question is to what degree the direction of the in vivo allelic imbalance in TF binding is concordant with the difference in the binding affinity as predicted by the sequence-to-affinity model. We used the MotifCentral model for CTCF as derived from in vitro binding data (Fig. 2A) to predict the direction of allelic preference from DNA sequence alone for each ASB variant (Fig. 2B). The overall concordance without regard of the precise numerical value of the respective affinity for each allele, only the direction of the difference, was 62.4%. While this is significantly larger than the expected value of 50% in the random case (binomial test p-value $< 10^{-16}$), it is too far below 100% to be practically useful. However, two sequence-derived metrics can be used to stratify variants in terms of the likelihood that their allelic preference will be correctly predicted: First, we found that a large fold-difference in predicted affinity is associated with much higher concordance (Fig. 2C). Second, the predicted affinity of the strongest allele is informative in a complementary way, with substantial improvement in concordance seen over a remarkable thousand-fold range in affinity before it finally deteriorates (Fig. 2D); the latter may reflect both that the predictions of our model become less accurate at lower affinities and that the affinity-mediated effect of the variant on TF occupancy gets smaller compared to other contributions. When we imposed lower bounds on both of these metrics simultaneously, even higher levels of concordance can be achieved (Fig. 2E). Similar trends (Additional file 1: Figure S1) were seen for the only two other TFs represented in AlleleDB that have at least 100 variants with evidence for ASB (387 variants for PU.1, and 189 for EBF1). Taken together, these results point to the feasibility of sequence-based prediction of ASB via quantification of the difference in binding affinity between the two alleles, provided that it is possible to assess the reliability of such affinity predictions.

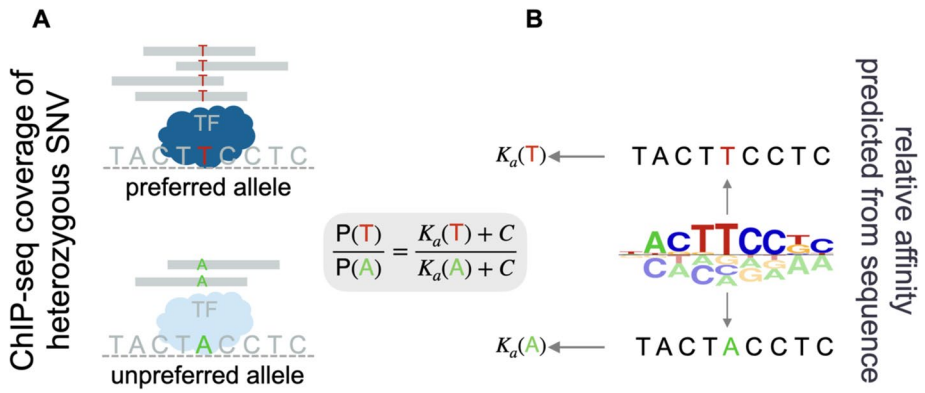

Fig. 1 Overview of the allele-specific binding data and binding affinity scoring. **A** For each variant, the preferred and unpreferred allele are defined in terms of ChIP-seq read coverage. **B** The binding affinity for each variant is computed as the sum of relative affinity scores over all possible offsets

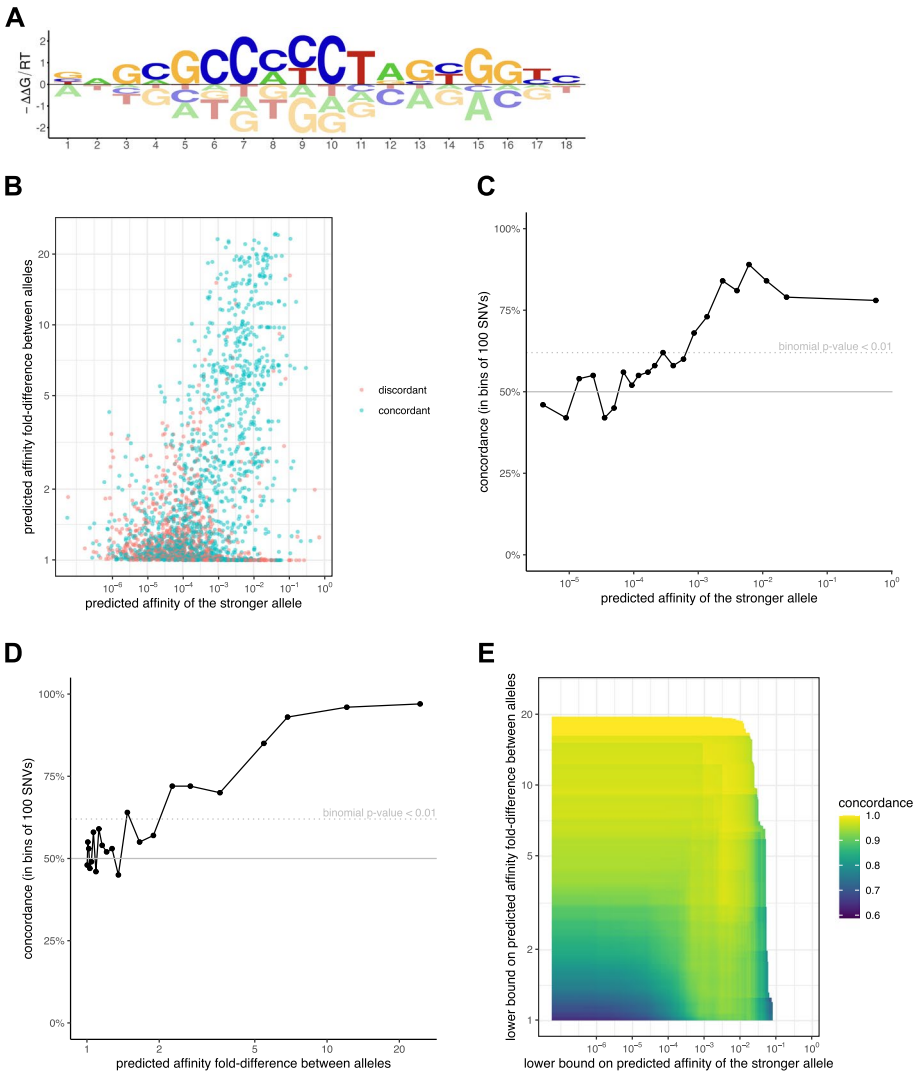

Fig. 2 Predicting allele-specific binding by CTCF using a sequence-to-affinity model. **A** Energy logo [22] representation of a CTCF binding model from MotifCentral capable of making accurate predictions of binding affinity. **B** Scatterplot of CTCF single-nucleotide variants (SNVs) at which significant ASB was detected in the AlleleDB study [5], colored by concordance. The *x*-value corresponds to the greater of the predicted affinities of the two alleles; the *y*-value corresponds to the ratio of predicted affinities of the two alleles. **C**, **D** Concordance of allelic preference, in bins of 100 SNVs as ranked by either of the axes in **B**. **E** Two-dimensional cumulative concordance of allelic preference as in **B** for all cutoffs with 20 or more SNVs

A metric for TF binding model quality based on allelic-specific ChIP-seq counts across the genome

Since for most TFs the fraction of variants for which allele-specific binding can be demonstrated on an individual basis is very small, we set out to find a way to aggregate below-threshold evidence for allelic preferences across the genome. We settled on a likelihood framework that uses the same beta-binomial distribution used by [5] to model allelic counts for AlleleDB; however, rather than using this model to reject the null hypothesis that both alleles of a given variant are equally probable, we used the likelihood function to quantify the performance of sequence-based predictors of allelic preference on a genome-wide scale (see "Methods" for details). Importantly, this was done

without regard to the statistical significance of allelic imbalance at the level of individual variants.

AlleleDB [5] used a beta-binomial test with a probability of 1/2 for both alleles to detect instances of ASB from allele-aware ChIP-seq data. By contrast, we here used allele-specific ChIP-seq counts aggregated across multiple individuals, and re-estimated the overdispersion parameter underlying the distribution by maximizing the likelihood function. We reasoned that using a beta-binomial model with variant-specific allelic probabilities based on predicted CTCF binding affinities should improve the likelihood, compared to the equal-preference control.

We found that this was indeed the case for the sequence-to-affinity model for CTCF in MotifCentral. To assess the statistical significance of the difference, we used bootstrapping to sample the log-likelihood distribution for each binding model separately (Fig. 3; see "Methods" for details). Compared to the equal-preference control, the MotifCentral model has a significantly higher mean log-likelihood (-3.2356 for control, -3.2089 for MotifCentral; Wilcoxon test $p < 10^{-8}$). Consistently, the maximum-likelihood estimate of the overdispersion parameter is lower (0.0916 for MotifCentral, 0.0990 for control). As expected, including all possible offsets between the binding model and the variant is important when predicting cumulative affinities from its flanking DNA sequence (Additional file 1: Figure S2). We saw modest further improvement in the likelihood when including contributions from offsets at which there was no overlap between the scored sequence and the variant, consistent with an interpretation that additional binding sites in the flanks would blunt the difference between the alleles when their effects on ChIP enrichment are additive.

We applied the same overall analysis to PU.1 and EBF1. Again, the MotifCentral model had statistically significant predictive performance when aggregating evidence across all variants (Additional file 1: Figure S3). Overall, our likelihood-based framework for

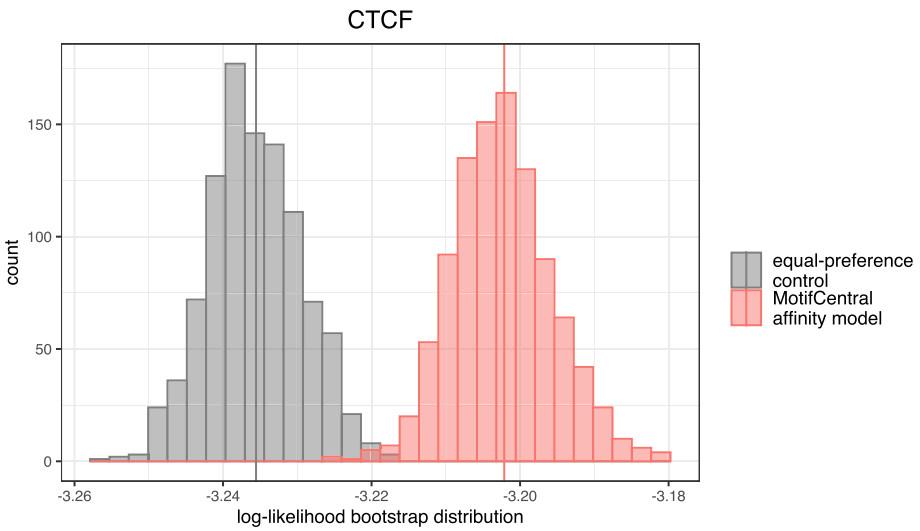

Fig. 3 Bootstrap distributions of log-likelihood for CTCF. The histograms show the bootstrap distributions of log-likelihood for 1000 resamples, based on the affinity-based likelihood model (allelic ratio predicted from genome sequence using binding model) or the control model (which assumes the alleles are equiprobable). The vertical line indicates the observed value of log-likelihood from each model

integrating evidence across the genome based on the beta-binomial distribution provides a way to quantitatively compare sequence-based predictors of allelic preference in an aggregate manner that is not limited by the number of variants that reaches statistical significance for calling ASB at the individual-locus level.

In vivo datasets can be affected by DNA fragmentation bias and overly short fragment lengths

So far, the TF binding models we used to score binding affinity were all derived from in vitro SELEX data. We wondered whether models derived in an allele-agnostic manner from in vivo assays would be better at explaining the allelic imbalances reflected in the allele-aware ChIP-seq data. To this end, we configured PyProBound to analyze paired-end in vivo data without peak calling (see "Methods" for details). Notably, when using the ChIP-seq data for CTCF from ENCODE, we found that the length of the merged read pairs was large enough for PyProBound to infer accurate binding models without any mapping to the genome.

Initially, when using PyProBound to fit a single binding mode intended to represent the DNA binding specificity of CTCF, the associated positional profile also learned by PyProBound indicated a strong bias at the ends of the ChIP-seq fragments (Additional file 1: Figure S4A). We interpreted this as confounding between the sequence preferences of CTCF binding and an unknown local sequence dependence of DNA fragmentation at the ends of the paired reads. To address this, we configured PyProBound to fit a more complex model that accounts for two multiplicative effects simultaneously: (i) the sequence dependence of the rate of DNA fragmentation during sonication at the two observed ends of the fragment; and (ii) the sequence dependence of CTCF binding, which determines the probability that the fragment is crosslinked with this TF during immunoprecipitation (see "Methods" for details).

This more sophisticated use of PyProBound when learning a sequence-to-affinity model from allele-agnostic ChIP-seq data for CTCF led to significantly improved performance when predicting ChIP-seq enrichment across the genome, and made the positional bias disappear (Additional file 1: Figure S4C). Note that a simpler approach in which we truncated each fragment to 200 bp around its center, which obscures the fragmentation bias due to the varying length of the fragments, performed less well on predicting enrichment during the immunoprecipitation step (Additional file 1: Figure S4B).

We found that PyProBound analysis of CUT&Tag and ChIP-exo data for CTCF did not require explicit fragmentation modeling, which allowed us to add additional flanking sequence to the genomic fragments defined by mapped read pairs to account for additional binding sites contributing to the enrichment achieved by the assay, improving the ProBound loss (from 1.5555 to 1.5308 for CUT&Tag after extending by 1300 bp, and from 1.3939 to 1.3917 for ChIP-exo after extending by 200 bp).

Taking DNA fragmentation bias into account when analyzing ChIP-seq data yields superior sequence-to-affinity models

To summarize the performance of the various alternative sequence-to-affinity models for CTCF thus generated, we again used our ASB likelihood metric (Fig. 4). The results were consistent with expectation in various regards: (Py)ProBound models trained on

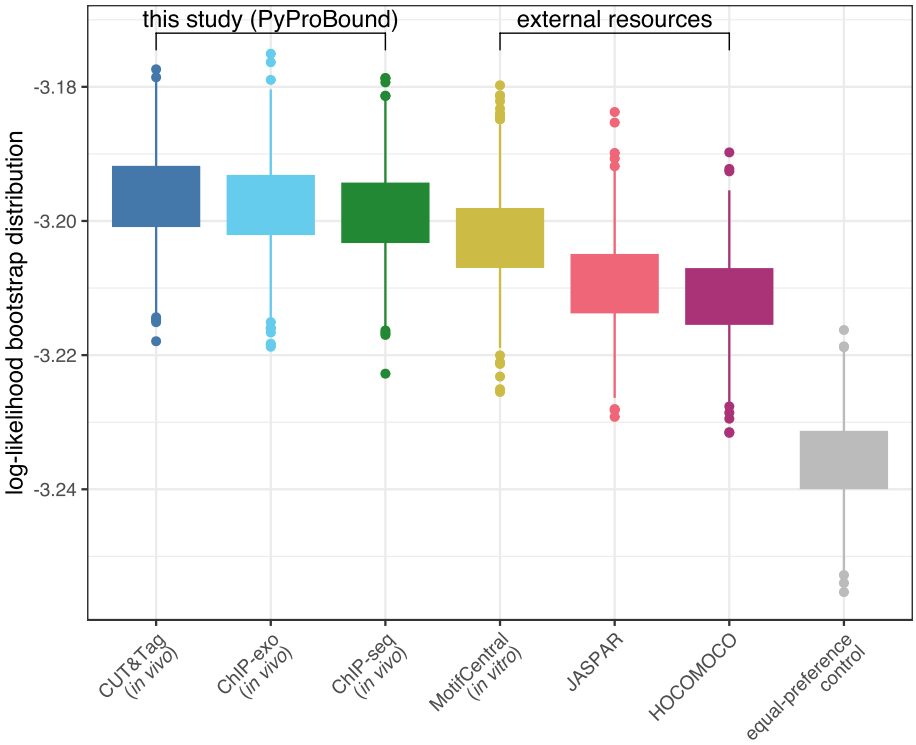

Fig. 4 Comparison between bootstrap distributions of log-likelihood across various models for CTCF binding. The boxplot shows the comparison in log-likelihood between the control model and affinity-based models based on different CTCF binding motifs, including a CUT&Tag-derived model introduced in this study and motifs from various resources

in vivo binding data in general outperform those trained on in vitro data. Additionally, ChIP-seq-derived PyProBound models trained directly on fragments outperform those trained on sequences extracted from genome alignments, as well as models trained on fragments that were all truncated to be the same length (log-likelihood −3.1989 for raw fragments; −3.2089 for length truncated; −3.2016 for genome alignment). CUT&Tag and ChIP-exo models improve in performance when trained with additional flanking sequences (log-likelihood improved from −3.1992 to −3.1963 for CUT&Tag after extending by 1300 bp, and from −3.1988 to −3.1979 for ChIP-exo after extending by 200 bp). Moreover, the PyProBound model inferred from CUT&Tag data showed better performance than either the PyProBound model inferred from ChIP-seq or ChIP-exo data or the motifs for CTCF available from JASPAR and HOCOMOCO (Fig. 4). Comparing the models on the less principled but perhaps more intuitive coefficient of determination (Pearson R^2) of binned allelic ratio yielded the same ranking in performance (Additional file 1: Figure S5).

Since models from external resources may not be well-equipped to predict relative binding affinities, we also wished to compare them on the simpler task of predicting only the direction of allelic preference, to investigate why they underperform on the ASB likelihood metric. Using the two metrics discussed above of predicted score fold-difference and predicted score of the strongest allele, we found that HOCOMOCO and JASPAR models of CTCF binding systematically underperform compared to MotifCentral in

terms of allelic preference concordance; the performance gap is strongest among the sequences that these models rank as the highest predicted binders (Additional file 1: Figure S6).

De novo motif discovery from allele-specific ChIP-seq counts

Our success in predicting allelic preference on a genome-wide scale using independently derived TF binding specificity models suggests that the direct modulation of TF binding by variants can explain allelic imbalance to a significant degree. We therefore wondered whether the same beta-binomial likelihood function could be leveraged to perform de novo motif discovery purely by trying to explain allelic imbalances in ChIP-seq counts from DNA sequence, while taking variation in combined ChIP-seq coverage among variants for granted. To our knowledge, such an approach, which controls for variation in chromatin context in a unique way by implicitly assuming a similar local molecular environment of a given variant on the respective homologs of the chromosome on which it resides, has not been explored before. We configured PyProBound to optimize the beta-binomial likelihood function not only with respect to the overdispersion parameter and non-specific binding coefficient as was done above, but also with respect to the position-specific free-energy parameters of the TF binding model (see "Methods"). Since we could not compare the resulting model on the benchmark data it was trained on, we instead compared its energetic parameters against previously published models (Fig. 5A). The ASB-derived model showed excellent agreement with the MotifCentral model (Pearson $R^2 = 0.890$; Fig. 5B), indicating that CTCF allelic imbalance is indeed driven by direct alteration of sequence-specific CTCF binding. We used the same motif discovery approach for EBF1, which resulted in a model that again had good agreement with the corresponding MotifCentral model (Pearson $R^2 = 0.817$; Additional file 1: Figure S7A). For PU.1, however, the ASB-derived model discovered additional specificity at one end of the binding mode compared to the MotifCentral model, leading to inferior agreement (Pearson $R^2 = 0.739$; Additional file 1: Figure S7B). A wider binding model capturing this additional specificity instead best compared to the HOCOMO model for PU.1 (Pearson $R^2 = 0.823$; Additional file 1: Figure S7C), indicating that a difference in in vitro and in vivo binding specificity for PU.1 may underlie these observations.

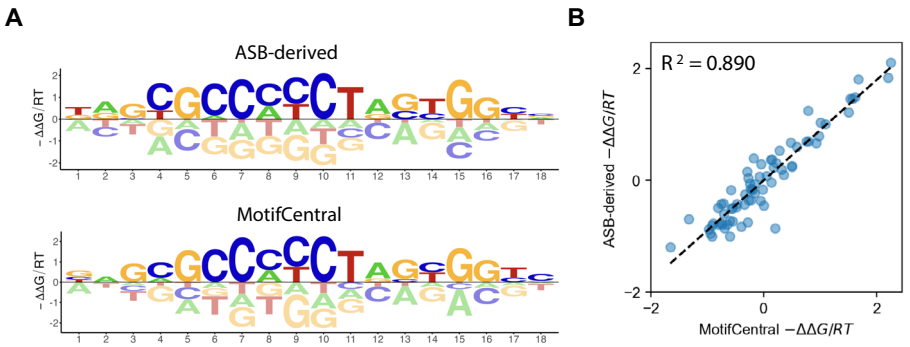

Fig. 5 Comparison of de novo ASB-derived and existing MotifCentral CTCF models. **A** Energy logos for binding model inferred from allele-aware CTCF ChIP-seq data from [5] using PyProBound, and from CTCF SELEX data using the original version of ProBound [18]. **B** Direct comparison of free energy parameters in the respective models, with each point corresponding a unique base/position combination in the logo

Discussion

In this study, we used allele-aware ChIP-seq data from AlleleDB [5] to assess the performance of DNA binding models when predicting allelic preference using sequence-based scoring of variation in binding affinity. The allele-specific approach is useful for assessing the prediction of genetic effects on in vivo TF binding. Leveraging the availability of many types of in vitro and in vivo binding data for the human transcription factor CTCF, we demonstrated that predictions of allelic preference made using sequence-based models trained on in vitro binding data, or trained on in vivo binding data in an allele-agnostic manner, can be highly concordant with empirical observations of allelic imbalance made using allele-aware ChIP-seq assays.

We developed a likelihood framework based on the (over-dispersed) binomial distribution that can be used to quantify in an unbiased manner how well the predicted binding affinities can explain allelic preference on a genome-wide scale. This makes it possible to leverage the large number of variants for which allele-aware ChIP-data do not provide enough statistical power to make calls of ASB on an individual-variant basis. In fact, for most TFs, it is not possible to detect many instances of ASB using per-variant analysis of ChIP-seq counts. We propose that our method for aggregating evidence for allelic preference across the genome can be broadly useful for benchmarking sequence-based predictors of TF binding affinity across many different cellular contexts and many different TFs.

A key feature of our (Py)ProBound approach to building sequence-to-affinity models for TF binding is that each read in the training data contributes independently to the estimation of the binding free energy parameters in the model. By not relying on peak calling, we preserve the biophysical relationship between binding affinity and enrichment during the immunoprecipitation step in the assay. As expected, when comparing the performance of various binding models for CTCF when predicting allelic differences in TF binding as assayed in vivo using ChIP-seq, ChIP-exo, or CUT&Tag, models derived from in vivo data in an allele-agnostic manner generally perform better than those derived from in vitro binding data, even when the role of co-factors is not explicitly considered. In particular, ChIP-seq-derived models directly trained on genomic fragments defined by paired-end sequencing data outperform those trained on sequences extracted from genome alignments, perhaps due to genomic instability in the cell lines used. Additionally, CUT&Tag-derived models outperform ChIP-based approaches, perhaps by avoiding sequence-specific biases due to crosslinking. We also showed that our genome-wide likelihood function can be easily leveraged to perform de novo motif discovery from allele-aware ChIP-seq data, thanks to the flexible nature of our PyProBound software.

Taken together, our findings underscore the value of optimal methodology for estimating binding energy parameters, which indeed was our original motivation for developing ProBound and related tools [18, 23]. Note that ProBound already naturally accounts for any pre-existing structure in the ChIP input library by only aiming to explain how the immunoprecipitation step leads to further enrichment that can be explained in terms of the underlying DNA sequence. Therefore, this is distinct from the fragmentation bias that we explicitly account for with PyProBound using the technique described in this paper.

Conclusion

In the future, more sophisticated TF binding models may emerge that explicitly incorporate cooperative interactions among sets of interacting TFs, inferred using ProBound [18], the PyProBound platform presented here, or other methods. It is an appealing prospect that our ASB likelihood framework could again offer an objective and unbiased metric for evaluating the predictive performance of such more elaborate models in a relevant in vivo cellular context.

Methods

Allele-specific TF binding data

Allele-specific TF binding data were downloaded from the AlleleDB database [5], which contains allele-specific annotations for the 1000 Genomes variant catalog (https://doi.org/https://doi.org/10.1038/nature15393). The AlleleDB authors reprocessed hundreds of ChIP-seq assays on tissue samples from 14 human individuals, mapping reads to a personal genome, and identifying allelic imbalance using a beta-binomial test. The variants classified as "accessible" in AlleleDB comprises heterozygous loci that have at least the minimum number of reads needed to be statistically detectable, and consists of both variants with statistically significant ASB and non-ASB variants that can serve as controls. In the present study, we used the raw ChIP-seq counts for the reference and alternative allele as input to our modeling. We primarily examined CTCF in our analyses due to its relative abundance of ASB (2231 ASB variants, 44,422 other variants). We also analyzed data for EBF1 (189 ASB variants) and PU.1 (387 ASB variants).

Sequence-based prediction of protein binding affinity for ASB prediction

A TF can bind genomic DNA near a given SNV at various offsets on either strand, so we computed the relative association constant for each allele of the SNV as a "sliding-window" sum of relative binding affinities at each combination of offset and orientation x of the bound sequence S_x relative to the full sequence S that contains the variant:

$$K_a(S) = \sum\nolimits_x K_a(S_x) = \sum\nolimits_x \exp\left(-\frac{\triangle\triangle G(S_x)}{RT}\right)$$

Here $\triangle\triangle G(S_X) = \triangle G(S_X) - \triangle G(S_{\text{ref}})$ is the binding free energy penalty as predicted by the TF-DNA binding model, relative to an optimal reference sequence. We used the R language and the BSgenome.Hsapiens.UCSC.hg19 package from Bioconductor.org to construct reference and alternative DNA sequences with 29 bases of flanking sequence on each side of the variant, sufficient to accommodate binding models of various sizes.

TF-DNA binding models used in the analyses

We downloaded TF binding models derived from HT-SELEX data using ProBound [18] from MotifCentral.org. In addition, we collected TF binding motifs for CTCF, PU.1, and EBF1 from HOCOMOCO [21], and JASPAR [19]. These models were imported into PyProBound for the scoring of DNA sequences.

Beta-binomial model of genome-wide allelic effects on the binding affinity

To quantify how well the predicted binding affinity can explain the genome-wide ASB effects, we build a generalized linear model based on the beta-binomial distribution:

$$\boldsymbol{Y}_i \sim \text{BetaBinomial}\left(n_i, p_i, \rho\right)$$

Here $\boldsymbol{Y}_i = \left(Y_i^{\text{ref}}, Y_i^{\text{alt}}\right)$ denotes raw ChIP-seq counts Y for a variant with either reference or alternative alleles. In the beta-binomial distribution, $n_i = Y_i^{\text{ref}} + Y_i^{\text{alt}}$ plays the role of the sample size, and ρ is the over-dispersion parameter. The binomial success rate p_i was modeled in terms of the relative affinities for reference allele and alternative allele as follows:

$$p_i = \frac{K_a(S_i^{\text{alt}}) + C}{K_a\left(S_i^{\text{alt}}\right) + K_a\left(S_i^{\text{ref}}\right) + 2C}$$

Here S_i^{ref} and S_i^{alt} are the DNA sequences centered on the respective alleles of variant i; C represents background binding due to indirect effects or binding of other TFs. The parameters C and ρ were estimated by likelihood maximization using the R language. The likelihood was computed by the probability density function of the beta-binomial distribution within the R package, VGAM 1.1–5 (Vector Generalized Linear and Additive Models). The mean log-likelihood across variants was then computed with optimal parameters C and ρ. For the control model, a fixed $p_{\text{i}} = 1/2$ was used and the parameter ρ was estimated by likelihood maximization.

Bootstrapping of log-likelihood

To construct the sampling distribution of log-likelihood, the variants were resampled 1000 times with replacement and each time the parameters of the model were estimated using the same function. The distributions of log-likelihood from 1000 bootstraps were then constructed for the control model and affinity-based model separately.

In vivo binding data analysis with PyProBound

The algorithm follows the methodology published previously [18]. The quasi-Newton optimization method L-BFGS is used to optimize the Poisson loss function

$$\log \mathcal{L} = \frac{1}{\sum_{ir} k_{i,r}} \sum_{i,r} \left[k_{i,r} \log\left(\frac{\eta_r f_{i,r} \sum_{ir'} k_{i,r'}}{\sum_{ir'} \eta_{r'} f_{i,r'}} \right) - k_{i,r} - \log\left(k_{i,r}!\right) \right]$$

where $k_{i,r}$ is the observed count of fragment i in library r (either control or bound), $f_{i,r}$ is the predicted count of fragment i, η_r is a parameter that adjusts for the read depth, and $\eta_{\text{control}} f_{i,\text{control}} = 1$ by convention.

Unlike the original Java implementation of ProBound (http://github.com/RubeGroup/ProBound) [18], PyProBound can train on sequences of varying lengths. This allows for analysis of paired-end in vivo binding data directly without extending or truncating reads. The enrichment of reads in the bound libraries relative to the control libraries was modeled as the product of three factors: (i) $K_{a,\text{CTCF}}$ of the CTCF binding mode, summed

over all sliding windows; (ii) $K_{a,\text{left}}$ of the pair fragmentation mode scored only on the 10 base pairs in the DNA fragment; (iii) $K_{a,\text{right}}$ of the reverse-complemented fragmentation mode scored only on the last 10 base pairs. If the CTCF model was observed to exhibit strong bias near the ends, fragmentation modes were iteratively added until the fragmentation score of the highest-affinity sequence was lower than non-specific binding.

To each of these factors a trained non-specific binding parameter α^{NS} was added. Additionally, a multiplicative bias ω was trained for each consecutive set of five offsets, to detect for bias in binding at different positions within the fragment. The count $f_{i,\text{bound}}$ for a sequence S of length n is therefore

$$\left(\alpha^{\text{NS}}_{\text{CTCF}} + \sum_x \omega_{\lfloor x/5 \rfloor} e^{-\Delta\Delta G_{\text{CTCF}}(S_x)/RT}\right) \prod_i \left(\alpha^{\text{NS}}_{\text{left},i} + e^{-\Delta\Delta G_{\text{left},i}(S_{1:10})/RT}\right)\left(\alpha^{\text{NS}}_{\text{right},i} + e^{-\Delta\Delta G_{\text{right},i}(S_{n-9:n})/RT}\right)$$

PyProBound regularization

Two regularization terms were added to avoid overfitting, as published previously [18]. The first is an L2 regularization term with hyperparameter $\lambda = 10^{-6}$ for in vivo models or $\lambda = 10^{-4}$ for ASB-derived models (which required more severe regularization due to the low number of sequences trained on). The second regularization term is an exponential barrier that prevents numerical errors, and is defined as

$$\sum_i \left(e^{\theta_i - 40} + e^{-\theta_i - 40}\right)$$

where the sum is over all parameters of the model.

In vivo binding data used for de novo motif discovery using ProBound

Raw FASTQ files corresponding to the paired-end CTCF ChIP-seq were downloaded from ENCODE (encodeproject.org) using accession numbers ENCLB048DBS and ENCLB581JXH. Reads were pair-ended with BBMerge with parameters adapter = default k = 60 [24]. Raw FASTQ files corresponding to the CTCF ChIP-exo and CUT&Tag assays were downloaded from SRA (www.ncbi.nlm.nih.gov/sra) using accession numbers SRR6736394, SRR6736387, SRR8435051, and SRR8754587. Reads were mapped to GRCh38 with BBMap [25] using default parameters and then quality filtered using samtools view [26] with parameters –q 30 –F 1804 –f 2 and filtered for duplicate alignments before extracting the sequence.

Extension of in vivo binding fragments

For datasets that did not exhibit significant sequence bias at the ends of fragments, the reads were iteratively extended by 100 bp on either end. A control dataset was created by randomly permuting the newly appended sequences relative to the central fragments. Fragments were extended until the log-likelihood of the extended dataset was $< 5 \times 10^{-4}$ better than the control dataset.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03424-2.

Additional file 1. Supplemental figures S1-S7 with captions.

Additional file 2. Peer review history.

Acknowledgements
We thank Tuuli Lappalainen, Athena Tsu, Harshit Ghosh, H. Tomas Rube, and Chaitanya Rastogi for valuable discussions.

Review history
The review history is available as Additional file 2.

Peer review information
Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
HJB and XL developed the ASB likelihood methodology; LANM implemented PyProBound and was responsible for its application to various data sets; XL and LANM wrote all the software and performed all analyses under the supervision of HJB; XL, LANM, and HJB wrote the manuscript.

Funding
This research was supported by NIH award R01MH106842 to H.J.B. and a PhRMA Foundation pre-doctoral fellowship in informatics to X.L.

Data availability
Allele-specific TF binding data were downloaded from the AlleleDB database (http://archive.gersteinlab.org/proj/alleledb/) [5] using the files ASB.auto.v2.1.aug16.txt.tgz and accB.auto.v2.1.aug16.txt.tgz.
The CTCF ChIP-seq data were downloaded from ENCODE with accession numbers ENCLB048DBS [27] and ENCLB581JXH [28]. The CTCF ChIP-exo and CUT&Tag data were downloaded from SRA with accession numbers SRR6736394 [29], SRR6736387 [29], SRR8435051 [30], and SRR8754587 [30].
All code (releases under the GPL-3.0 license) for processing data files, training models, and generating figures can be accessed at: https://github.com/BussemakerLab/AlleleSpecificBinding. Version v1.0 was permanently archived as https://doi.org/10.5281/zenodo.13941939.
The latest version of PyProBound (released under the MIT license) can be downloaded from https://github.com/BussemakerLab/PyProBound. Version v1.5.0 of PyProBound was permanently archived as https://doi.org/10.5281/zenodo.13937673. PyProBound documentation is available at https://pyprobound.readthedocs.io.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
H.J.B. is a co-founder and shareholder of Metric Biotechnologies, Inc. The other authors declare that they have no competing interests.

Received: 15 December 2023 Accepted: 17 October 2024
Published online: 31 October 2024

References

1. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. The human transcription factors. Cell. 2018;172:650–65.
2. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337:1190–5.
3. Deplancke B, Alpern D, Gardeux V. The genetics of transcription factor DNA binding variation. Cell. 2016;166:538–54.
4. McDaniell R, Lee BK, Song L, Liu Z, Boyle AP, Erdos MR, Scott LJ, Morken MA, Kucera KS, Battenhouse A, et al. Heritable individual-specific and allele-specific chromatin signatures in humans. Science. 2010;328:235–9.
5. Chen J, Rozowsky J, Galeev TR, Harmanci A, Kitchen R, Bedford J, Abyzov A, Kong Y, Regan L, Gerstein M. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. Nat Commun. 2016;7:11101.
6. Cavalli M, Pan G, Nord H, Wallerman O, Wallen Arzt E, Berggren O, Elvers I, Eloranta ML, Ronnblom L, Lindblad Toh K, Wadelius C. Allele-specific transcription factor binding to common and rare variants associated with disease and gene expression. Hum Genet. 2016;135:485–97.

7. Abramov S, Boytsov A, Bykova D, Penzar DD, Yevshin I, Kolmykov SK, Fridman MV, Favorov AV, Vorontsov IE, Baulin E, et al. Landscape of allele-specific transcription factor binding in the human genome. Nat Commun. 2021;12:2751.
8. Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, Migliavacca E, Wiederkehr M, Gutierrez-Arcelus M, Panousis NI, et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science. 2013;342:744–7.
9. Reddy TE, Gertz J, Pauli F, Kucera KS, Varley KE, Newberry KM, Marinov GK, Mortazavi A, Williams BA, Song L, et al. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. Genome Res. 2012;22:860–9.
10. Kribelbauer JF, Rastogi C, Bussemaker HJ, Mann RS. Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes. Annu Rev Cell Dev Biol. 2019;35:357–79.
11. Ogawa N, Biggin MD. High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro. Methods Mol Biol. 2012;786:51–63.
12. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al. DNA-binding specificities of human transcription factors. Cell. 2013;152:327–39.
13. Isakova A, Groux R, Imbeault M, Rainer P, Alpern D, Dainese R, Ambrosini G, Trono D, Bucher P, Deplancke B. SMiLE-seq identifies binding motifs of single and dimeric transcription factors. Nat Methods. 2017;14:316–22.
14. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007;316:1497–502.
15. Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, Myers Z, Sud P, Jou J, Lin K, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. 2020;48:D882–9.
16. Rossi MJ, Lai WKM, Pugh BF. Simplified ChIP-exo assays. Nat Commun. 2018;9:2842.
17. Kaya-Okur HS, Wu SJ, Codomo CA, Pledger ES, Bryson TD, Henikoff JG, Ahmad K, Henikoff S. CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat Commun. 1930;2019:10.
18. Rube HT, Rastogi C, Feng S, Kribelbauer JF, Li A, Becerra B, Melo LAN, Do BV, Li X, Adam HH, et al. Prediction of protein-ligand binding affinity from sequencing data with interpretable machine learning. Nat Biotechnol. 2022;40:1520–7.
19. Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranasic D, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2020;48:D87–92.
20. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33:831–8.
21. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 2018;46:D252–9.
22. Foat BC, Morozov AV, Bussemaker HJ. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. Bioinformatics. 2006;22:e141-149.
23. Rastogi C, Rube HT, Kribelbauer JF, Crocker J, Loker RE, Martini GD, Laptenko O, Freed-Pastor WA, Prives C, Stern DL, et al. Accurate and sensitive quantification of protein-DNA binding affinity. Proc Natl Acad Sci U S A. 2018;115:E3692–701.
24. Bushnell B, Rood J, Singer E. BBMerge - Accurate paired shotgun read merging via overlap. PLoS ONE. 2017;12:e0185056.
25. Bushnell B. BBMap: A Fast, Accurate, Splice-Aware Aligner. 2014.
26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
27. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. ENCSR617IFZ. ENCODE. https://www.encodeproject.org/experiments/ENCSR617IFZ/. 2016.
28. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. ENCSR007GUS. ENCODE. https://www.encodeproject.org/experiments/ENCSR007GUS/. 2016.
29. Rossi MJ, Lai WKM, Pugh BF. Simplified ChIP-exo assays. GSE110681. Gene Expression Omnibus (GEO). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE110681. 2018.
30. Kaya-Okur HS, Wu SJ, Codomo CA, Pledger ES et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. GSE124557. Gene Expression Omnibus (GEO). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1245572018. 2019.

Publisher's Note