Variance-stabilized units for sequencing-based genomic signals

Faezeh Bayat
Maxwell Libbrecht*
Department of Computing Science, Simon Fraser University, Burnaby BC, Canada

Abstract

Sequencing-based genomic signals such as ChIP-seq are widely used to measure many types of genomic biochemical activity, such transcription factor binding, chromatin accessibility and histone modification. The processing pipeline for these assays usually outputs a real-valued signal for every position in the genome that measures the strength of activity at that position. This signal is used in downstream applications such as visualization and chromatin state annotation. There are several representations of signal strength at a given that are currently used, including the raw read count, the fold enrichment over control, and log p-value of enrichment relative to control. However, these representations lack the property of variance stabilization. That is, a difference between 100 and 200 reads usually has a very different statistical importance from a difference between 1,100 and 1,200 reads. Here we propose SDPM, variance-stabilized units for sequencing-based genomic signals. We demonstrate that these variance stabilized units have several desirable properties. First, the difference between a pair of cell types of ChIP-seq signal at a gene's promoter is highly correlated with the difference in that gene's expression. Second, the ChIP-seq signal at a gene's promoter has a linear relationship with that gene's expression when the ChIP-seq is represented with variance-stabilized units; this correlation is stronger with variance-stabilized units than existing alternatives. SDPM units will eliminate the need for downstream methods to implement complex mean-variance relationship models, and will enable genomic signals to be easily understood by eye.

1 Introduction

Sequencing-based assays can measure many types of genomic biochemical activity, including transcription factor binding, histone modifications and chromatin accessibility. These assays work by selecting DNA fragments from a sample that exhibit the desired type of activity, sequencing the fragments to produce sequencing reads and mapping each read to the genome. Each of these assays produces a genomic signal—that is, a signal that has a value for each base pair in the genome. Examples include ChIP-seq measurements of transcription factor binding or histone modification and measurements of chromatin accessibility from DNase-seq, FAIRE-seq or ATAC-seq.

The usefulness of genomic signals requires they be expressed in units that are suitable for analysis. The natural unit of sequencing-based assays is the read count: the number of reads that mapped to a given position in the genome (after extending and shifting; see Methods). However, read count is a poor measure of the strength of a given signal, for reasons discussed below. For RNA-seq, which measures gene (or transcript) signals, several units have been proposed that perform better than read count, including RPKM, FPKM and TPM [13, 4]. However, less work has been done to develop good units for genomic signals.

*Corresponding Author. Email: maxwl@sfu.ca

In particular, we are interested in developing units for genomic signals with the desirable property of *variance stability*. That is, if we performed many replicates of the same assay, the variance of the signal should be constant from position to position. Existing units are variance in-stable; for example, a locus with 1,100 reads might get 1,200 reads in a subsequent experiment by chance, whereas a locus that went from 100 to 200 reads likely reflects a non-random change in activity.

Currently, the primary approach for analyzing genomic signals is to develop a complex statistical model that accounts for the behavior of read counts. In particular, most state-of-the-art genomic signal analysis methods for tasks such as peak calling use the negative binomial distribution because this distribution allows the model to account for a non-uniform relationship between the mean of the signal and its variance [14]. However, there are two limitations to this approach. First, a substantial investment must be made in each new application of genomic signals to implement and optimize such a statistical model. Second, without well-behaved units, visualization is difficult.

Because of the difficulty in developing and optimizing complex models, many existing methods use simple Gaussian models of genomic signals. Two prominent examples include imputation and semi-automated genome annotation (SAGA). Existing methods for imputation use mean squared error (MSE) as an evaluation metric, which is equivalent to log likelihood under a fixed-variance Gaussian model [8, 12]. The most widely-used existing methods for SAGA either use a Gaussian distribution [3] or binarize the data to avoid having to find good units [7], although some work has been done to use a negative binomial distribution for SAGA as well [11].

Other than raw read counts, two units are currently used for genomic signals: fold enrichment and Poisson p-value [9, 10]. Fold enrichment measures a genomic signal as the ratio of reads of the experiment to a control (such as ChIP Input). Poisson p-value measures a signal as the log p-value of a Poisson distribution test with a null hypothesis derived from a control distribution. Because these units are variance in-stable, existing methods attempt to stabilize the variance of these units by applying a transformation such as log or inverse hyperbolic sine (arcsinh). However, these transformations do not fully stabilize the variance (Results).

In this manuscript, we propose variance-stabilized units for sequencing-based genomic signals called SDPM ("standard deviations per million mapped reads"). Unlike alternative units, SDPM are variance-stabilized: any pair of loci with SDPM scores of x and $x + 1$ have the same difference in activity regardless of x. We do this by comparing multiple replicates of the same assay to derive a mean-variance relationship, then using this relationship to derive a variance-stabilizing transformation.

2 Methods

2.1 ChIP-seq data

We acquired ChIP-seq data from the ENCODE consortium for the histone modification H3K4me3 on four cell lines: GM12878, H1-hESC, HUVEC and K562 which have ENCODE accession numbers ENCSR000DRY, ENCSR019SQX, ENCSR000AKN and ENCSR000AKU, respectively. These ChIP-seq data sets were processed with a uniform pipeline [5]. Briefly, the ChIP-seq reads were mapped to the hg19 reference genome and reads were shifted and extended according to the estimated fragment length to produce a read count for each genomic position. As controls, ChIP-seq Input experiments were performed by the same labs. Two signals were produced: fold enrichment and log p-value. Fold enrichment signal is defined as the ratio of observed data over control [9]. P-value signal is defined as the log p-value of a Poisson model with a null distribution derived from the control [10].

2.2 RNA-seq data

For use in evaluation, we acquired RNA-seq data sets for each of the cell types above from the Roadmap Epigenomics consortium [10]. These RNA-seq data sets were processed with a uniform pipeline that produces a TPM value for each gene [10]. To stabilize the variance of these signals, we used an asinh transformation.

2.3 Identifying the mean-variance relationship

Our variance-stabilizing transformation depends on knowing the mean-variance relationship for input data. We learn this by comparing multiple replicates of the same experiment. Specifically, we assume that we have two replicates, which we term the *base* and *auxiliary* replicates respectively. Let the observed signal at position i be $x_i^{(1)}$ and $x_i^{(2)}$ for the base and auxiliary replicates respectively. Our model imagines that every position i has an unknown distribution of sequencing reads for the given assay $P_i(x)$, which has mean $\mu_i = \text{mean}(P_i(x))$. We further suppose that there is a relationship $\sigma(\mu)$ between the mean and variance of these distributions. That is, $\text{var}(P_i(x)) = \sigma(\mu_i)^2$. We are interested in learning $\sigma(\mu)$. Observe that x_i is an unbiased estimate of μ_i, and that $(x_i^{(1)} - x_i^{(2)})^2$ is an unbiased estimate of $\sigma(\mu_i)^2$. We use this observation to estimate the function $\sigma(\mu)$ as follows.

We first sort the N genomic positions $i \in \{1 \dots N\}$ by the value of $x_i^{(1)}$ and define bins with b genomic positions each. Let $I_j \subseteq \{1 \dots N\}$ be the set of positions in bin j. For each bin j, we compute $\mu_j = 1/b \sum_{i \in I_j} x_i^{(1)}$ and $\sigma_j^2 = 1/b \sum_{i \in I_j} (x_i^{(2)} - \mu_j)^2$. To increase the robustness of these estimates, we smooth across bins by defining

$$\bar{\sigma}_j^2 = \frac{\sum_{i=j-w}^{j+w} 2^{-b|j-w|/\beta} \sigma_i^2}{\sum_{i=j-w}^{j+w} 2^{-b|j-w|/\beta}}. \tag{1}$$

That is, we take the weighted average of $2w+1$ bins centered on j, where bin $j+k$ has weight $2^{-bk/\beta}$. β is a bandwidth parameter—a high value of β means that weight is spread over many bins, whereas a low value means that weight in concentrated on a small number of bins. We define w such that it ignores bins with weight less than 0.01; specifically, $w = -\beta \log(0.01)/b \log(2)$

The choice of b and β forms a bias-variance trade-off. Larger values of b and β lead to more observations contributing to each estimate $\sigma_j(\mu)$ and therefore result in a lower variance. In contrast, small values of b and β lead to a very homogeneous set of positions I_j and therefore less averaging across dissimilar positions. We compared multiple values of b and β to optimize this choice (Results).

2.3.1 Calculating variance-stabilized signals

Having learned the mean-variance relationship, we compute SDPM using the variance-stabilizing transformation [6]

$$t(x) = C \int_0^x \frac{1}{\bar{\sigma}(u)} du, \tag{2}$$

where $\sigma(x)$ is the standard deviation of a variable with a mean of x which in our method, x is the weighted mean, and C can be any constant. This transformation is guaranteed to be variance-stabilizing; that is, $\text{var}(t(x_i))$ is constant for all genomic positions i.

2.4 Alternative methods

We consider two alternative units for genomic signals that are typically used in existing analyses: fold enrichment [9] and Poisson p-value [10]. Fold enrichment signal is defined as x'_i/c_i, where x'_i is the raw read count at genomic position i and c_i is the read count of a control experiment (e.g. ChIP-seq input sample). Poisson p-value signal is the log p-value of Poisson distribution test of whether the observed signal is greater than the control. To attempt to stabilize the variance, existing methods usually apply either a log or arcsinh transformation. These transformations are used because they are variance-stabilizing for certain mean-variance relationships [1].

Specifically, $\log(x)$ is variance-stabilizing when $\sigma(\mu)^2 = s\mu^2$ for some constant s, and $\text{arcsinh}(x)$ is variance-stabilizing when $\sigma(\mu)^2 = s\mu^2 + \lambda$ [2]. In the experiments below, we compare to both existing units under both existing transformations.

2.5 Differential expression evaluation

We developed two measures to evaluate the quality of units for genomic signals. The first is based on the objective that, when two samples are compared, a large difference in signal at a gene's promoter should indicate a large difference in that gene's expression. Specifically, for each gene k, let x_k^S and g_k^S be the signal and gene expression respectively. We define $dx_k = x_k^{S_1} - x_k^{S_2}$ and $dg_k = g_k^{S_1} - g_k^{S_2}$. We define the differential expression score as the correlation between dx and dg. We use Spearman correlation in order to remove the dependence of the score on the scale of the units. We use the histone modification H3K4me3 as the genomic signal for this evaluation because it is indicative of gene expression. Signals with poor units will likely have a poor differential expression score because high-variance signals (usually high-magnitude signals) will overwhelm the correlation

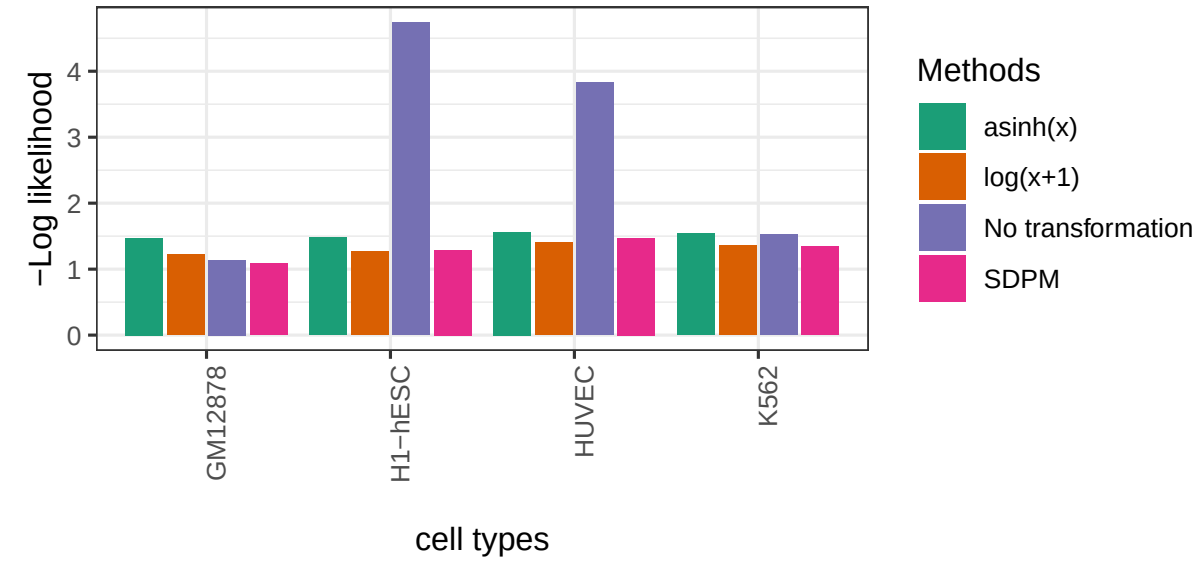
2.6 Linearity evaluation

Our second evaluation measure is based on the objective that genomic signals with good units should have a linear relationship with other types of data. Specifically, we define the linearity score as the Pearson correlation between g_k and x_k. We use the Pearson correlation because it specifically identifies the linear relationships between the two data sets.

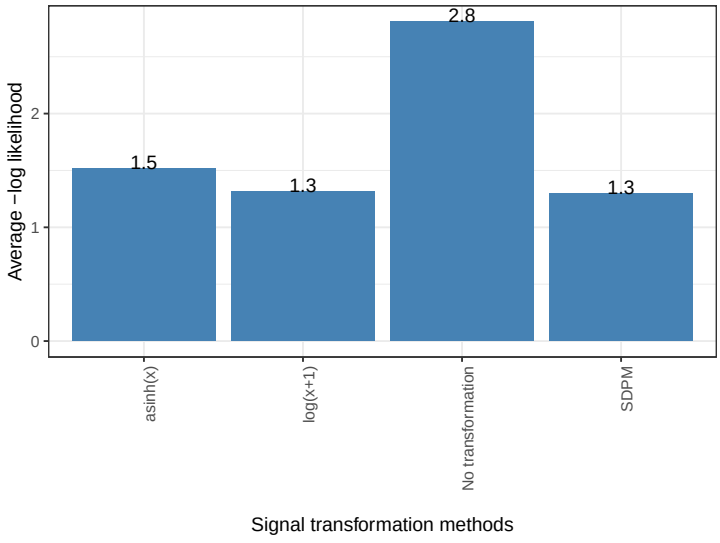
3 Results

3.1 Existing units are not variance-stabilized

To evaluate whether existing units for genomic signals have stable variance, we computed the mean-variance relationship for a number of existing data sets (Methods). As we expected, we found that the variance has a strong dependence on the mean; genomic positions with low signals experience little variance across replicates, whereas positions with high signals experience much larger variance (Figure 2c). Moreover, the relationship does not match that expected by the currently-used $\log(x+1)$ and $\text{asinh}(x)$ transformations: these transformations are variance-stabilizing when the mean-variance relationship is linear or linear plus constant respectively (Methods). In contrast, the observed mean-variance relationship does not precisely match either of these functions, indicating that neither of these transformations is fully variance-stabilizing.



(a)



(b)

Figure 1: Comparison of the SDPM method log likelihood with three other signal transformation methods; Fold enrichment, asinh(Fold enrichment) and log(x+1). Lower values for negative log likelihood are preferable. a)Log likelihood comparison. b)Average log likelihood comparison

3.2 We derived a variance-stabilizing transformation for genomic signals

Motivated by the observation that existing transformations are not variance-stabilizing, we used the learned mean-variance curve to derive a variance-stabilizing transformation (Methods). As expected, we found that this transformation successfully stabilized the variance of the signal. To measure the fit, we evaluated the log likelihood under a max-entropy model. Specifically, the max-entropy model given a specific mean and variance is a Gaussian distribution, so we evaluated the likelihood under a Gaussian model. As expected, we found that the learned mean-variance curve

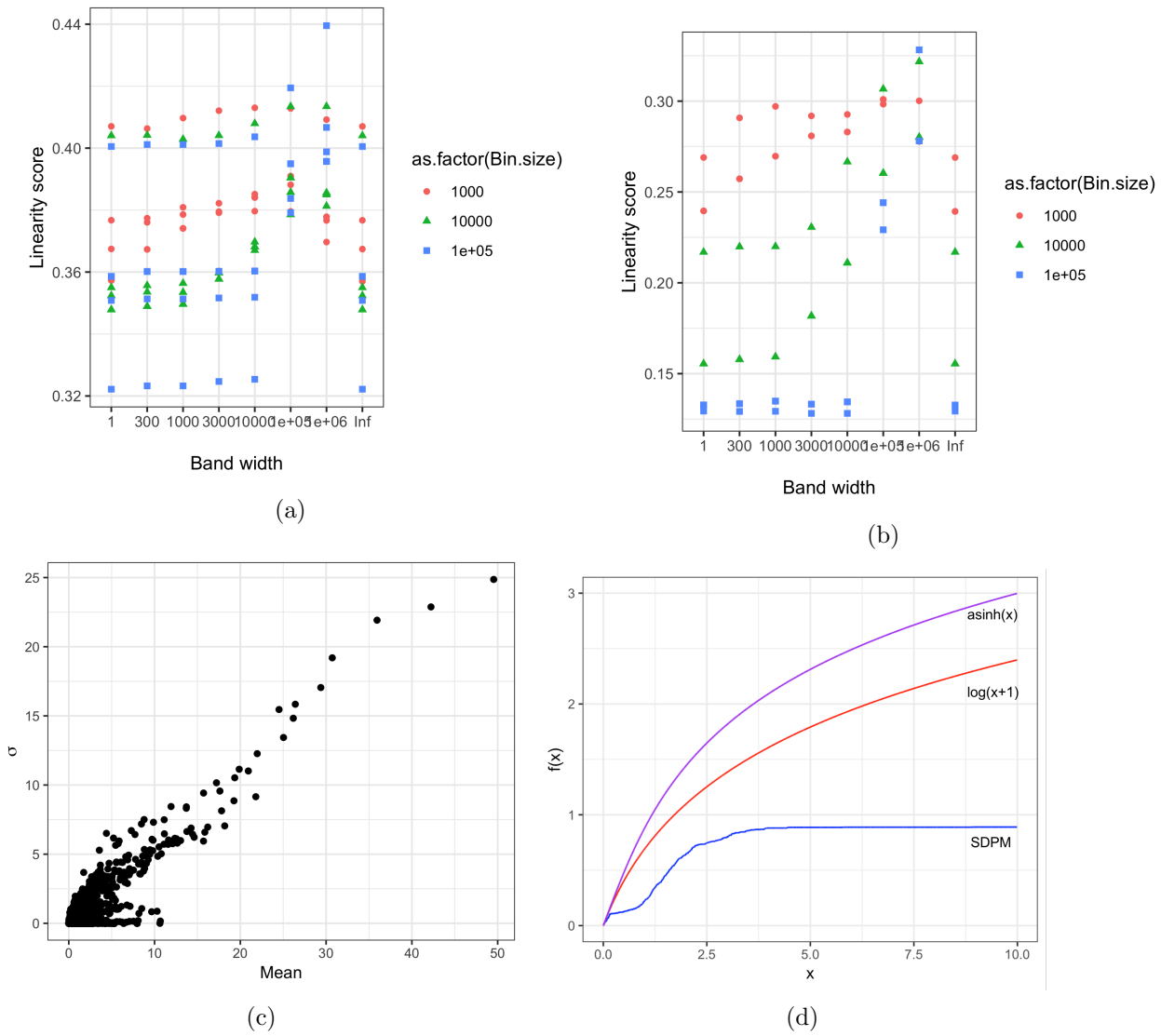


Figure 2: Deriving the mean-variance relation. a) Relation between bin size and bandwidth and Pearson correlation in gene expression analysis. b) Relation between bin size and bandwidth and Pearson correlation in differential expression analysis. c) Learned Mean-variance relationship derived from two replicates. d) Learned transformation function.

In (c), multiply each transformation by a constant so that the lines line up better.

had a poor likelihood (average log density of -2.8), reflecting non-uniform variance (Figure 1). We found that either linear or linear plus constant models—implied by the $\log(x+1)$ and asinh(x) transformations respectively—greatly improved the likelihood (average log density of -1.31 and -1.51 respectively). Our learned mean-variance relationship had the best likelihood (average log density -1.29), indicating that the learned curve successfully models the mean-variance relationship of the data.

Moreover, we found that the mean-variance relationship differs greatly from experiment to

experiment (Figure 1a). Some experiments (on GM12878 and K562) have near-uniform variance without transformation, whereas others (H1-hESC and HUVEC) have a near-linear mean-variance relationship. The mean-variance relationship learned by SDPM correctly captures these differences, as indicated by its good likelihood on all data sets. These differences indicates that it is necessary to learn a separate mean-variance relationship for each data set, rather than applying a single transformation (such as log or asinh) to every data set.

3.3 Variance-stabilized units identify differences between cell types

We found that when genomic signals are represented in SDPM, differences in the signal between two cell types is predictive of a functional change between the cell types. To evaluate the quality of this predictiveness, for a given pair of cell types, we propose the differential expression score, which measures the correlation of the difference in H3K4me3 signal at a gene's promoter with the difference in that gene's expression (Methods). A high correlation indicates that We expect that units without stable variance will have low correlation because differences in signal will be overwhelmed by high-variance positions. We found that SDPM had a higher average differential expression score (0.33) than the other methods we tried (0.17-0.32).

3.4 SDPM have linear relationships with other measurements

We found that when genomic signals are represented in SDPM, they have linear relationships with other data sets (Figure 4). To evaluate this property, we calculated the Pearson correlation between the H3K4me3 SDPM signal at a gene's promoter and that gene's expression. Pearson correlation measures linear relationships, so the strength of Pearson correlation indicates the linearity of the relationship. We found that SDPM had a high Pearson correlation with gene expression—an average of 0.41 across the four cell types we tested. In contrast, other units for H3K4me3 signal had a less strong Pearson correlation (0.37-0.40, indicating a nonlinear relationship.

4 Discussion

In this manuscript we propose SDPM, units for sequencing-based genomic signals that have the desirable property of variance stability. This property is valuable for two reasons. First, SDPM signals to be used in downstream methods without the need for each method to implement a mean-variance relationship model. Second, SDPM signals can be easily analyzed by eye because the viewer does not need to take the mean-variance relationship into account when visually inspecting the data.

References

[1] Maurice S Bartlett. The use of transformations. *Biometrics*, 3(1):39–52, 1947.

[2] George EP Box. Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335, 1953.

[3] Rachel CW Chan, Maxwell W Libbrecht, Eric G Roberts, Jeffrey A Bilmes, William Stafford Noble, and Michael M Hoffman. Segway 2.0: Gaussian mixture models and minibatch training. *Bioinformatics*, 34(4):669–671, 2017.

[4] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong

Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):13, 2016.

[5] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57, 2012.

[6] Blythe P Durbin, Johanna S Hardin, Douglas M Hawkins, and David M Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18(suppl_1):S105–S110, 2002.

[7] Jason Ernst and Manolis Kellis. Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215, 2012.

[8] Jason Ernst and Manolis Kellis. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature biotechnology*, 33(4):364, 2015.

[9] Michael M Hoffman, Jason Ernst, Steven P Wilder, Anshul Kundaje, Robert S Harris, Max Libbrecht, Belinda Giardine, Paul M Ellenbogen, Jeffrey A Bilmes, Ewan Birney, et al. Integrative annotation of chromatin elements from encode data. *Nucleic acids research*, 41(2):827–841, 2012.

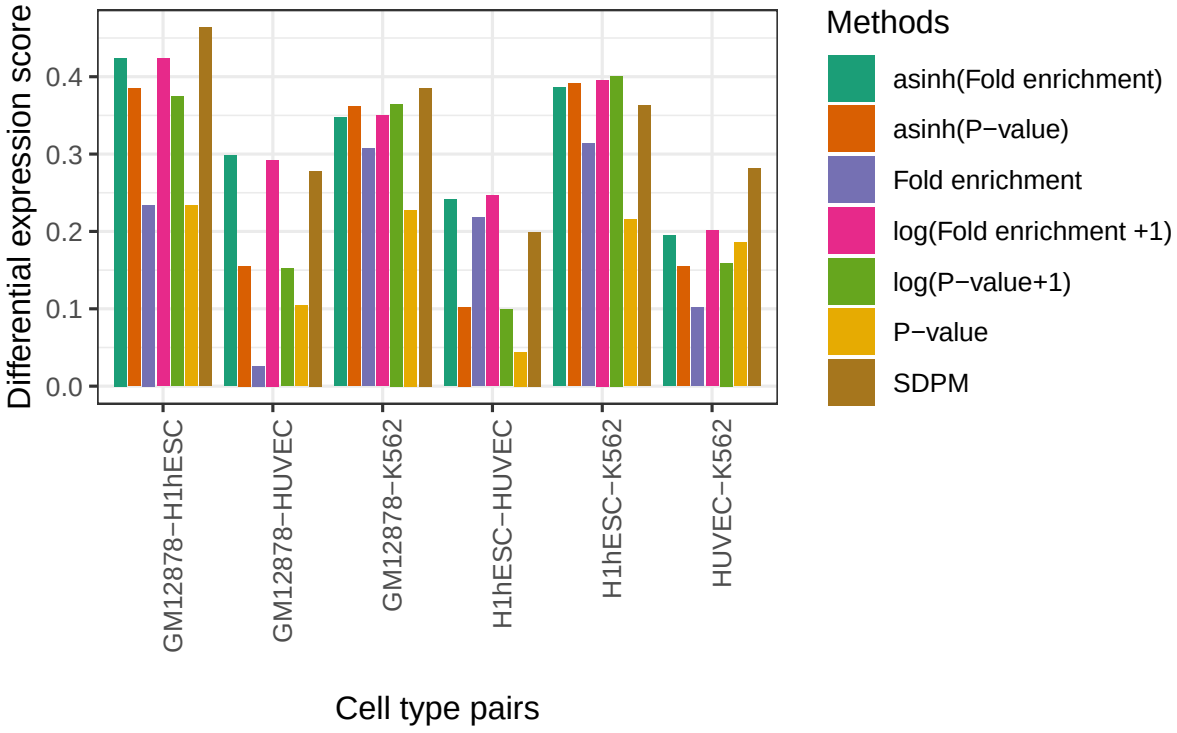
[10] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317, 2015.

[11] Alessandro Mammana and Ho-Ryun Chung. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome biology*, 16(1):151, 2015.

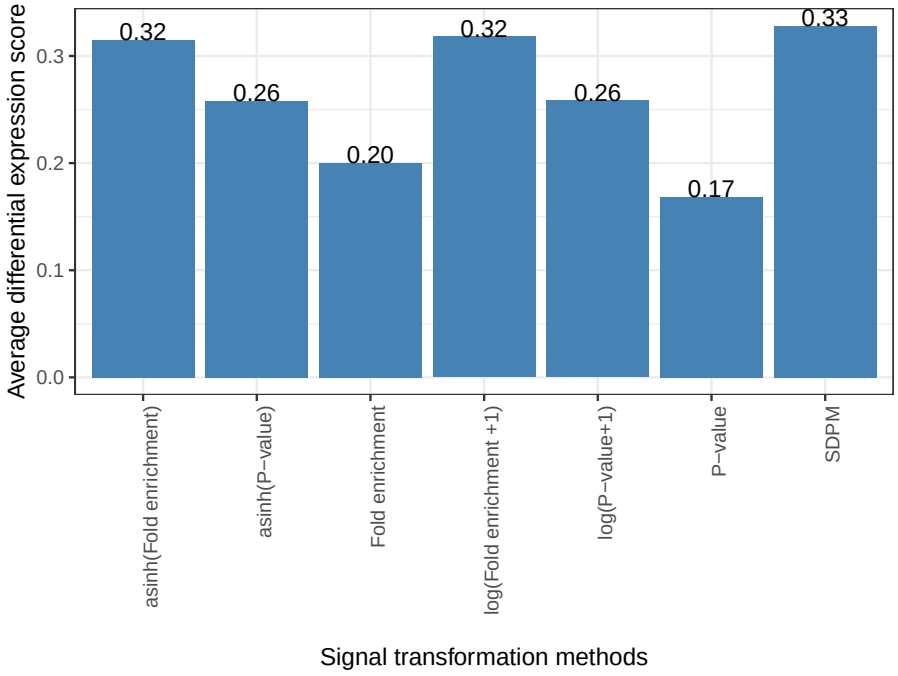
[12] Jacob Schreiber, Timothy J Durham, Jeffrey Bilmes, and William Stafford Noble. Multi-scale deep tensor factorization learns a latent representation of the human epigenome. *bioRxiv*, page 364976, 2018.

[13] Günter P Wagner, Koryu Kin, and Vincent J Lynch. Measurement of mrna abundance using rna-seq data: Rpkm measure is inconsistent among samples. *Theory in biosciences*, 131(4):281–285, 2012.

[14] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of chip-seq (macs). *Genome biology*, 9(9):R137, 2008.

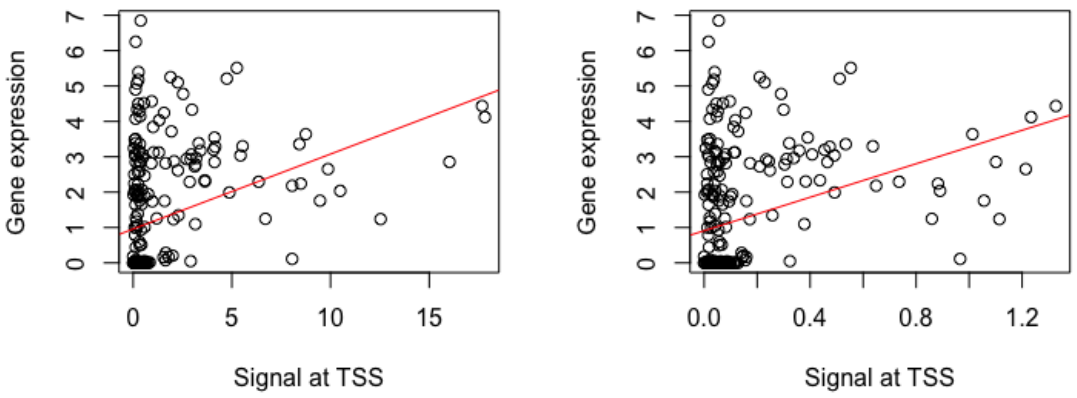


(a)

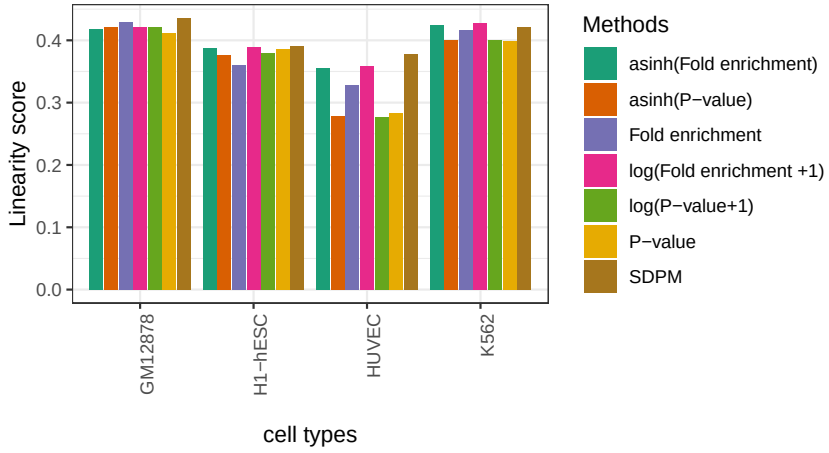


(b)

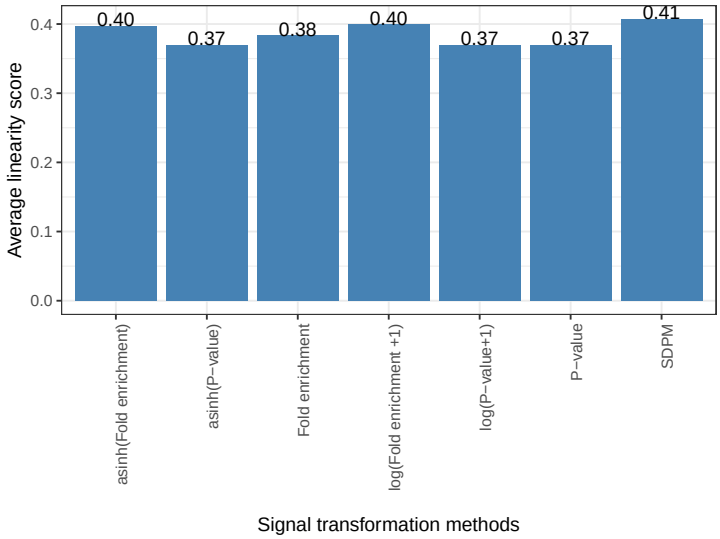
Figure 3: Differential expression evaluation. (a) Vertical axis indicates the differential expression score for a given pair of cell types, defined as the correlation between the difference in H3K4me3 value with difference in gene expression (Methods). (b) Same as (a), but averaged over all cell type pairs.



(a) (b)



(c)



(d)

Figure 4: (a,b) Relationship between H3K4me3 signal and gene expression for (a) fold enrichment units (b) SDPM units. (c) Linearity evaluation. Vertical axis indicates the linearity score, defined as the correlation between gene expression and H3K4me3 signal (Methods). (d) Same as (c), but averaged over cell types.