Variance-stabilized units for sequencing-based genomic signals

Faezeh Bayat[1] and Maxwell Libbrecht[2*]

Department of Computing Science, Simon Fraser University, Burnaby BC, Canada
{fbayat,maxwl}@sfu.ca

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 VSS, variance-stabilized signals for sequencing-based genomic signals. We generate VSS by learning the empirical relationship between the mean and variance of a given signal data set and producing transformed signals that normalize for this dependence. We demonstrate that these variance stabilized units have several desirable properties, including that differences in ChIP-seq signal across cell types indicate a difference in that gene's expression. VSS 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.

Keywords: variance stabilization · genomic signals · epigenomics · histone modification · chromatin accessibility · sequencing-based assays · ChIP-seq

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 extracting 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 [1, 2]. However, less work has been done to develop good units for genomic signals.

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 unstable; 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.

There has been a great deal of work on stabilizing the variance of RNA-seq data. One might expect that the location of each read were drawn independently from a multinomial distribution; this assumption implies that the read counts at each locus would follow a Poisson distribution. However, the variance of a Poisson distribution is always equal to its mean and this property does not hold true for real data sets, which often exhibit non-identical mean-variance relationships [3, 4]. Instead, many methods (including the widely-used DESeq method [5]) model read counts using a negative binomial distribution [6–10], which allows

these models to fit the mean-variance relationship of existing data. Most recently, a method called voom [11] learns an empirical, non-parametric mean-variance relationship to stabilize variance of RNA-seq signals in order to allow these signals to be analyzed with Gaussian distribution-based methods.

Despite the widespread use of variance stabilization in RNA-seq analysis, to our knowledge, no existing variance-stabilized units for genomic signals have been developed. 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 [12]. However, there are two limitations to this approach. First, each new application of genomic signals requires a substantial investment to implement and optimize such a statistical model. Second, variance unstable units impede hinder visualization.

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 [13, 14]. The most widely-used existing methods for SAGA either use a Gaussian distribution [15] or binarize the data to avoid having to find good units [16], although some work has been done to use a negative binomial distribution for SAGA as well [17].

Other than raw read counts, two units are currently used for genomic signals: fold enrichment and Poisson p-value [18, 19]. 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 signals (VSS) for sequencing-based genomic signals. Unlike alternative units, VSS are variance-stabilized: any pair of loci with VSS scores of x and $x + 1$ have the same difference in activity regardless of x. This method is inspired by the widely-used voom method for RNA-seq data [11]. We compare multiple replicates of the same assay to derive a mean-variance relationship, then use this relationship to derive a variance-stabilizing transformation.

2 Methods

2.1 ChIP-seq data

We acquired ChIP-seq data from the ENCODE consortium (`encodeproject.org`) for the histone modification H3K4me3 on six cell lines: GM12878, H1-hESC, HUVEC, K562, NHLF and GM06990. These data sets have ENCODE accession numbers ENCSR000AKA, ENCSR000AMG, ENCSR000AKN, ENCSR000AKU, ENCSR000DWZ and ENCSR000DQV, respectively. We used histone modifications H3K36me3, H3K4me1 and H3K27me3 on cell line H1-hESC with ENCODE accession numbers ENCSR925LJZ, ENCSR631RJR and ENCSR216OGD. In addition to human histone modification samples, we used mouse transcription factors ZC3H11A and MAFK on MEL cell line with ENCODE accession numbers ENCSR000ESF and ENCSR000ETK respectievly. These ChIP-seq data sets were processed with a uniform pipeline [20]. 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 [18]. P-value signal is defined as the log p-value of a Poisson model with a null distribution derived from the control [19].

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 [19]. These RNA-seq data sets were processed with a uniform pipeline that produces a TPM value for each gene [19]. 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 determining the mean-variance relationship for the input data set. We learn this by comparing multiple replicates of the same experiment. We designate two replicates as the *base* and *auxiliary* replicates, respectively. The following process is iterated for all possible choices of base and auxiliary (see below).

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 x_i, which has mean $\mu_i = \text{mean}(x_i)$. We further suppose that there is a relationship $\sigma(\mu)$ between the mean and variance of these distributions. That is, $\text{var}(x_i) = \sigma(\mu_i)^2$. 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 $\mu_j = 1/b \sum_{i \in I_j} x_i^{(2)}$ 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. (Figure 1). Based on the experiments, we chose to use $\beta = 10^6$ and $b = 10^5$ as the best combination of the parameters since it optimizes all evaluation metrics simultaneously (Results).

We used a smoothing spline to fit an estimated mean-variance curve. A smoothing spline estimator implements a regularized regression over the natural spline basis. We fit a function $\hat{\sigma}(\mu)$ using the estimated values of $\bar{\sigma}_j$. The spline coefficients w are selected to minimize,

$$(1-p) \sum_j w_j (\bar{\sigma}_j - \hat{\sigma}(\mu_j))^2 + p \int \left(\frac{d^2 \hat{\sigma}(\mu)}{dx^2} \right)^2 dx,$$

where μ and $\bar{\sigma}$ are a set of observations obtained from mean-variance data points. Variables $\hat{\sigma}(\mu)$, w, p represent smooth spline curve, weight coefficients and smoothing parameter. The variable p parameter varies between $(0, 1]$ such that $p = 0$ results in a cubic spline with no smoothing, and when p approaches zero the result is a linear function. To find the optimum value of *spar* parameter (p), first the smooth.spline function is called by activating the cross-validation in the smooth.spline (CV=TRUE). Following the cross-validation procedure, *spar* parameter is returned as the smoothing factor. We identified the optimal curve using the R function call `smooth.spline(means, sigmas, spar=p)`.

The process above assumed a fixed choice of base and auxiliary replicates. We repeat this process for each possible choice of base and auxiliary replicates (that is, for M replicates, the process above is repeated $M(M-1)$ times) and aggregate sufficient statistics across all iterations to produce an aggregated $\hat{\sigma}(\mu)$.

2.4 Calculating variance-stabilized signals

Having learned the mean-variance relationship, we compute VSS using the variance-stabilizing transformation [21]

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

where x is an untransformed signal and $\bar{\sigma}(u)$ is the learned standard deviation for a signal with mean u. This transformation is guaranteed to be variance-stabilizing; that is, $\mathrm{var}(t(x_i))$ is constant for all genomic positions i.

2.5 Alternative methods

We consider two alternative units for genomic signals that are typically used in existing analyses: fold enrichment [18] and Poisson p-value [19]. 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 [22]. Specifically, $\log(x)$ is variance-stabilizing when $\sigma(\mu) = s\mu$ for some constant s, and $\mathrm{arcsinh}(x)$ is variance-stabilizing when $\sigma(\mu) = s\sqrt{\mu^2+1}$ [23]. In the experiments below, we compare to both existing units under both existing transformations.

2.6 Likelihood analysis

We used a Gaussian log likelihood metric to measure the degree to which a transformation fits the mean-variance relationship. A transformation implicitly imposes a specific mean-variance relationship; the inferred variance for a given value equals the inverse of the derivative of the transformation (Methods). To measure the accuracy of this variance estimate, we used the Gaussian likelihood of the observed data. We chose a Gaussian distribution because this is the max-entropy distribution with a specific mean and variance. Specifically the log likelihood of a given data set is defined as

$$\sum_i \log N(x_i^{(2)} | \mu = x_i^{(1)}, \sigma = \hat{\sigma}(x_i^{(1)})). \tag{3}$$

This value is maximized when the inferred variance equals the variance of the data.

As noted above, a $\log(x+1)$ transformation implicitly assumes the mean-variance relationship $\hat{\sigma}(\mu) = \mu+1$ and the $\mathrm{arcsinh}(x)$ transformation assumes the mean-variance relationship $\hat{\sigma}(\mu) = s\sqrt{\mu^2+1}$.

2.7 Variance instability evaluation

We developed a variance instability metric that evaluates the consistency of the transformed signals between different replicates of an experiment. Consider $t(x_i^{(1)})$ and $t(x_i^{(2)})$ as the transformed signals of $x_i^{(1)}$ and $x_i^{(2)}$ at ith genomic position. We define $d_i = t(x_i^{(1)}) - t(x_i^{(2)})$. Using the binning approach described above, we divided genomic positions to B bins of increasing value of $t(x_i^{(1)})$, where each bin is of size $b = 10000$. For bin j, let md_j be the mean of d_i for each value i in bin j. Let MD be the mean across all values of i. We define variance instability metric as

$$X = \frac{1}{b} \frac{\sum_{bin_j} (MD - md_j)^2}{MD^2}.$$

Signals with unstable variance will have large values of the variance instability metric.

2.8 Differential expression evaluation

In order to evaluate the quality of genomic signals relative to data sets not involved in training, we we developed a differential expression metric that compares signals to measurements of RNA-seq expression. This metric 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 Pearson correlation between dx and dg. We use H3K4me3 signal for this evaluation because this histone modification is expected to be indicative of gene expression. Signals with poor units will likely have a poor differential expression score because noisy high-variance signals (usually high-magnitude signals) will overwhelm the correlation.

2.9 Setting hyperparameters

In order to choose values for VSS's hyperparameters, we evaluated many possible combinations of values using the three evaluation measures including likelihood analysis, variance instability metric and differential expression analysis (Figure 1). The set of parameters is considered optimized if they minimize the very first two evaluation metrics and maximize the last one. Thus, based on these results, we chose $\beta = 10^6$ and $b = 10^5$ as the optimal set of the parameters as it satisfies all evaluation metrics simultaneously. We have also compared the optimization results with $\log(x+1)$ transformation which is shown by red dashed lines in the 1. The results indicate that our approach is outperforming the $\log(x+1)$ transformation in all investigated evaluation metrics.

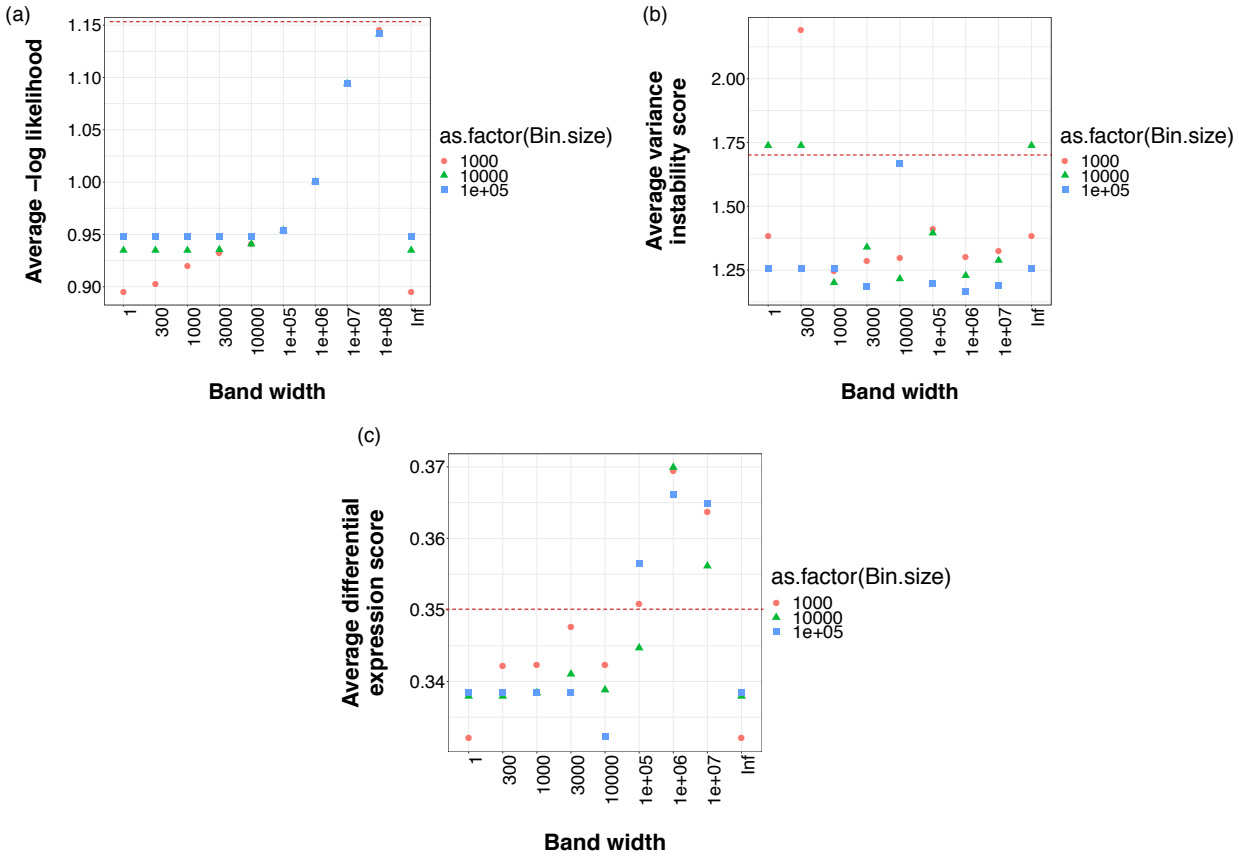

Fig. 1. Relationship between bin size and bandwidth in (a) likelihood analysis (b) variance instability analysis and (c) differential expression analysis. Among the band width values, the value Inf indicates that we applied no smoothing on the curve (Unweighted mean-variance curve). Red dashed line indicates the performance of $\log(x+1)$ transformation.

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 (Figure 2). 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 2b). Moreover, the relationship does not match that expected by the currently-used $\log(x+1)$ and $\operatorname{asinh}(x)$ transformations. A transformation implicitly imposes a specific mean-variance relationship; the inferred variance for a given value

equals the inverse of the derivative of the transformation (Methods). For example, the former transformation assumes a linear relationship (Methods). The observed mean-variance relationship does not precisely match the relationships assumed by either transformation, indicating that neither of these transformations is fully variance-stabilizing (Figure 2b).

The observation that existing transformations are not variance-stabilizing was confirmed when we quantified this fit (Figure 3). To measure the accuracy of a variance estimate, we used the Gaussian likelihood, which is maximized when the inferred variance equals the variance of the data (Methods). As expected, we found that a uniform variance model implied by using untransformed signals had a poor likelihood (average log density of -2.05), reflecting non-uniform variance (Figure 3b). We found that the variance estimates from the $\log(x+1)$ and asinh(x) greatly improved the likelihood (average log density of -1.32 and -1.52 respectively). However, we found that the learned mean-variance relationship had much better likelihood (average log density -0.97) than either transformation, indicating that the learned curve successfully models the mean-variance relationship of the data (Figure 4).

Moreover, we found that the mean-variance relationship differs greatly between experiments. For many histone modification ChIP-seq experiments like HeK4me3 H1-hESC and H3K4me3 HUVEC, a log transformation yields nearly-optimal fit, indicating that the data has a nearly linear mean-variance relationship (Figure 3a). However, other experiments, including most mouse transcription factor ChIP-seq and H3K4me3 in GM12878, have a very non-linear mean-variance relationship (Figure 3a). In fact, for some experiments, a log or asinh transformation has worse fit than no transformation, indicating that these transformations actually destabilize the variance (Figure 3a). Future work should investigate what properties of an experiment determine its mean-variance relationship. The mean-variance relationship learned by VSS 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.

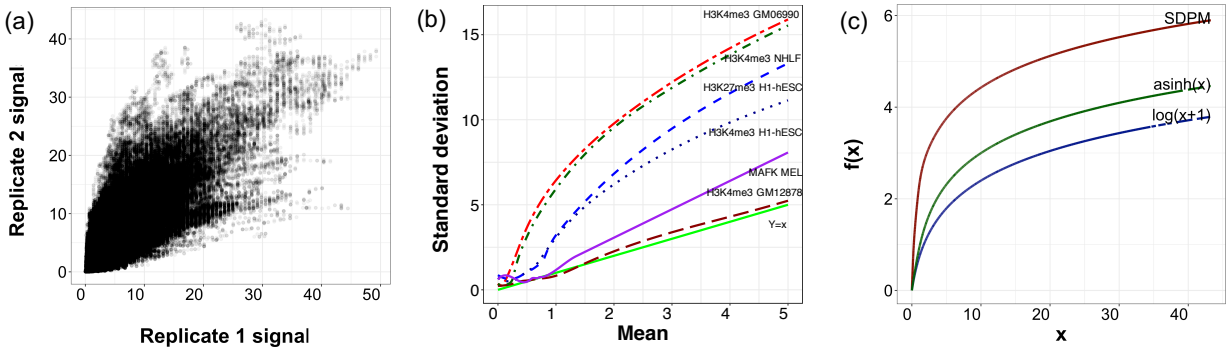

Fig. 2. Deriving the mean-variance relation. a) Replicate 1 vs Replicate 2 signals in H3K4me3 GM12878. b) Learned mean-variance relationships. c) Learned transformation function for H3K4me3 GM12878.

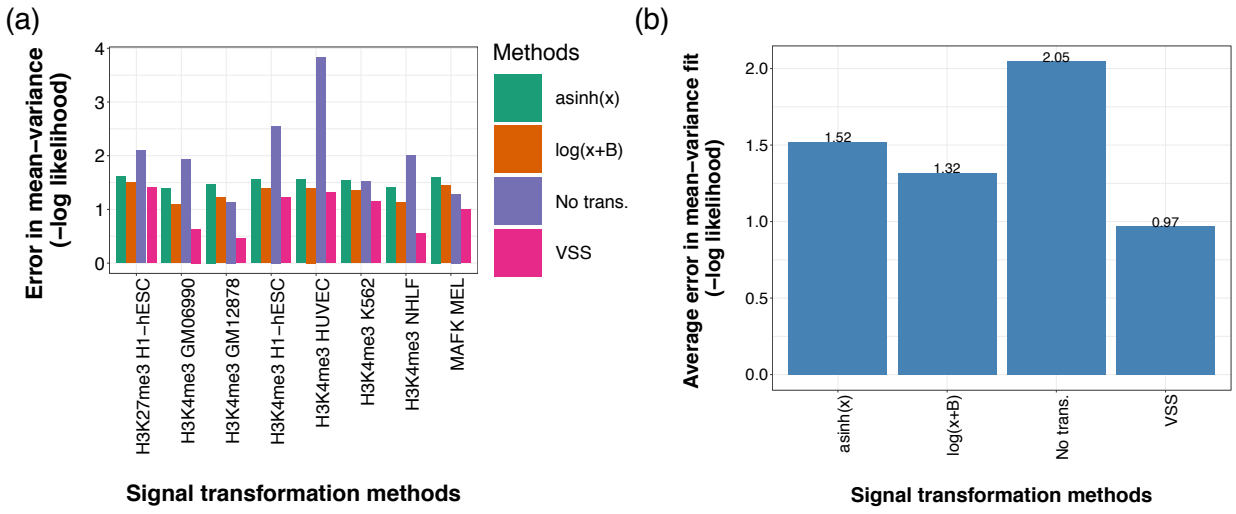

Fig. 3. (a) Goodness of fit to the mean-variance relationship, measured by Gaussian log likelihood (Methods). Lower values of negative log likelihood indicates better fit. Log likelihood was computed on chromosome 21; VSS's mean-variance relationship was trained on chromosome 22. (b) Same as (a), but averaged across data sets.

3.2 Differences between replicates are stabilized after transformation

We used the the variance instability metric to quantify the stability of transformed signals (Methods). This metric measures the degree to which differences between replicates vary for different magnitude of signal. Signals with unstable variance will have a large value for this metric.

We found that our VSS transformation is more consistent among different replicates in comparison with the $\log(x+1)$ and $\text{asinh}(x)$ transformations (Figure 4). Signals transformed by either $\log(x+1)$ and $\text{asinh}(x)$ had an average of 1.7 variance instability, whereas VSS have instability of 1.0. This indicates that VSS units are have more consistent signals among different replicates of an experiment (Figure 4).

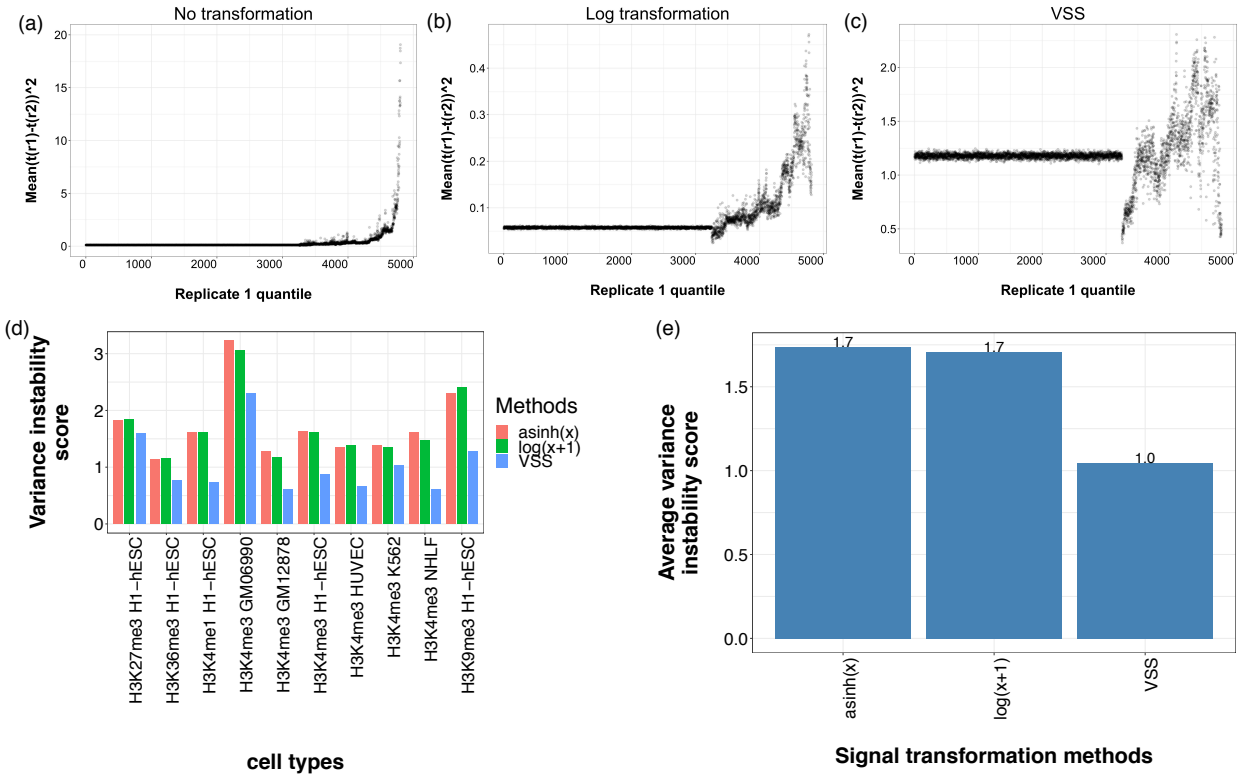

Fig. 4. Variance instability of transformed signals. (a,b,c) Each point corresponds to a bin, where binning is defined according to replicate 1 value (Methods). Horizontal axis indicates binning index. Vertical axis indicates squared difference of values between replicates in H3K4me3 NHLF. The flat line on the left half of each plot corresponds to positions where $x^{(1)} = 0$. Signals with stable variance show a flat (constant) trend on this plot; a trend (increasing or decreasing) indicates unstable variance. (a) Untransformed values. (b) $\log(x + 1)$ transform. (c) VSS transformation. (d) Variance instability score (Methods). Lower values indicate more stable variance. (e) Same as (d), but averaged across experiments. VSS's mean-variance relationship was trained on chromosome 22 and variance instability score was computed on chromosome 21. We omitted the variance instability value for untransformed signals because its value (mean of 324.3 across experiments) would distort the vertical axis.

3.3 Variance-stabilized units identify differences between cell types

We found that when genomic signals are represented in VSS, 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 cell type signal at a gene's promoter with the difference in that gene's expression (Methods). A high correlation indicates that differences in signal are informative at predicting other cellular phenomena. 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 VSS is slightly performing better than the asinh(x) and log(x+1) (0.37, 0.36 and 0.35, respectively) and outperforms the other methods that we tried (0.14-0.34) (Figure 5).

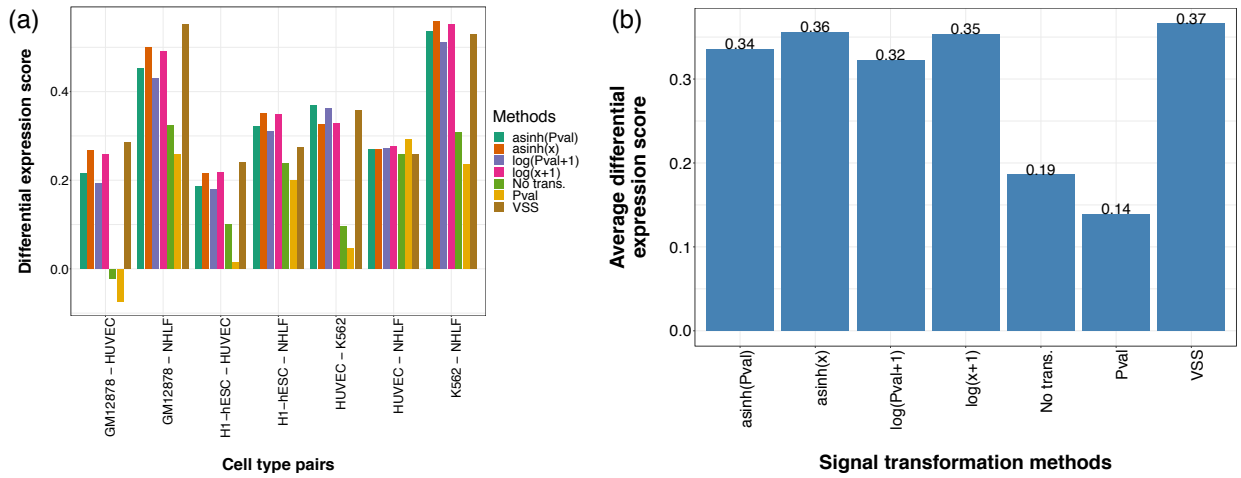

Fig. 5. (a) Differential expression evaluation. Vertical axis indicates the differential expression score for a given pair of cell types, defined as the correlation between the difference in H3K4me3 signal with difference in gene expression (Methods). Higher values indicate a better correlation between signal values and gene expression. Results are shown on chromosome 21 for an VSS model trained on chromosome 22. (b) Same as (a), but averaged over all cell type pairs.

4 Discussion

In this manuscript we proposed VSS, units for sequencing-based genomic signals that have the desirable property of variance stability. We found that the transformations that are currently used to attempt to stabilize variance—$\log(x+1)$ and $\operatorname{asinh}(x)$—do not fully do so. In fact, we found that the mean-variance relationship of genomic signals varies greatly between data sets, indicating that no single transformation can be applied to all data sets uniformly. Instead, one must use a method such as VSS to empirically determine the mean-variance relationship and use this to stabilize signals' variance.

Variance-stabilized signals will aid in all downstream applications of genomic signals. In particular, they are valuable for two reasons. First, VSS signals allow downstream methods to use squared error loss or Gaussian likelihood distributions, which are much easier to optimize than the existing practice of implementing a model that accounts for the mean-variance relationship. This will improve tasks such as chromatin state annotation and imputation that currently use Gaussian models.

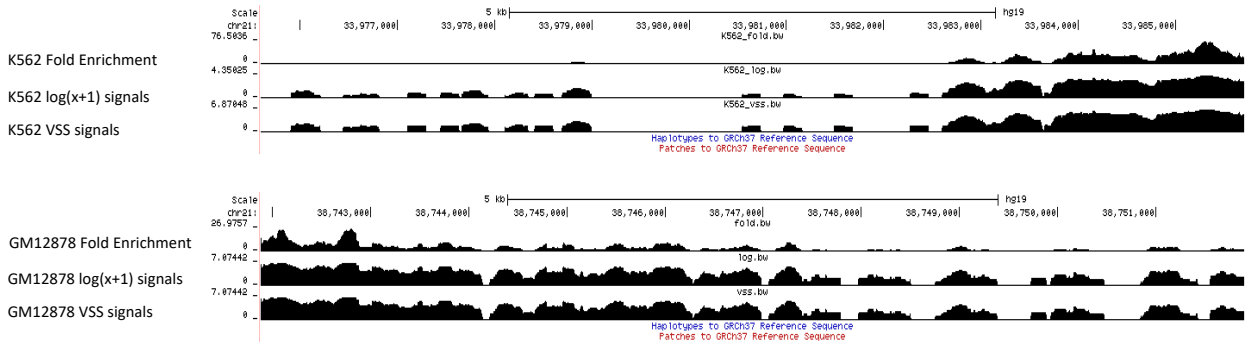

Fig. 6. Visualization of genomic signals (H3K4me3) in the UCSC genomic browser.

Second, VSS 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. For example, when viewing genomic

signals in a genome browser, variance-unstable signals often exhibit high peaks that swamp the vertical axis and flatten other variations in signal (Figure 6). Other methods for handling this problem—using a log transform or cutting off the vertical axis—can also be effective, but they lack the principled basis of VSS.

References

1. 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.
2. 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.
3. Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349, 2008.
4. Mark D Robinson and Gordon K Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.
5. Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010.
6. Lucy Whitaker. On the poisson law of small numbers. *Biometrika*, 10(1):36–71, 1914.
7. Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):1–15, 2019.
8. Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.
9. Marek Gierliński, Christian Cole, Pietà Schofield, Nicholas J Schurch, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, Karim Gharbi, Gordon Simpson, Tom Owen-Hughes, et al. Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics*, 31(22):3625–3630, 2015.
10. Xu Ren and Pei Fen Kuan. Negative binomial additive model for RNA-Seq data analysis. *bioRxiv*, page 599811, 2019.
11. Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29, 2014.
12. 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.
13. Jason Ernst and Manolis Kellis. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotechnology*, 33(4):364, 2015.
14. 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.
15. 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.
16. Jason Ernst and Manolis Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3):215, 2012.
17. 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.
18. 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.
19. 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.
20. ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57, 2012.
21. 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.
22. Maurice S Bartlett. The use of transformations. *Biometrics*, 3(1):39–52, 1947.
23. George EP Box. Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335, 1953.