VSS: Variance-stabilized signals 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

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 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).

Read counts of genomic assays have a nonuniform mean-variance relationship, which poses a challenge to their analysis. For example, a locus with 1,000 reads in one experiment might get 1,100 reads in a replicate experiment by chance, whereas a locus with 100 reads might usually see no more than 110 reads by chance in a replicate. This property means that, for example, the difference in read count between biosamples is a poor measure of the difference in activity. To handle this issue, most statistical models of genomic signals—such as those used in peak calling—model the mean-variance relationship of read counts explicitly using, for example, a negative binomial distribution [21, 1, 18, 8, 22, 16, 10, 9, 7, 19, 23].

However, negative binomial models are challenging to implement and optimize, so many methods resort to Gaussian models. Two prominent examples include segmentation and genome annotation (SAGA) methods, such as Segway or IDEAS [11, 12, 2, 24, 25], and imputation methods such as ChromImpute, PREDICTD and Avocado [6, 20, 4]. In the former example, many SAGA methods use a Gaussian distribution to model the distribution of genomic signals given a certain annotation label (others binarize signal [5] or use a negative binomial read count model [17]). In the latter example, imputation methods optimize a mean squared error (MSE) objective function, which is equivalent to log likelihood in a Gaussian model.

Most Gaussian-based methods employ a variance-stabilizing transformation to handle the nonuniform mean-variance relationship. They most commonly use the log or inverse hyperbolic sine transformations (asinh), which have the formulae $\log(x + c)$ for a constant c (usually 1) and $\text{asinh}(x) = \log(x + \sqrt{x^2 + 1})$ respectively [13].

Variance-stabilizing transformations are also required for visualizing genomic signals. As previously noted, Euclidean distances in a 2D plot correspond to log likelihood in a Gaussian model, so a nonuniform mean-variance relationship complicates visualization. Therefore, most visualizations of genomic signals such as genome browsers [14] employ a variance-stabilizing transformation such as log.

Despite the widespread use of log and asinh transformations to stabilize variance, to our knowledge, no work has evaluated whether they in fact do so. The use of these transformations assumes that the signals have a specific mean-variance relationship (Methods). Here we show that, for many genomic signals, this assumptions is violated and thus existing transformations do not fully stabilize variance (Results). To solve this issue, we present VSS, a method that produces variance-stabilized genomic signals. VSS determines the empirical mean-variance relationship of a genomic signal by comparing replicates. It uses this empirical mean-variance relationship to produce a transformation function that precisely stabilizes variance.

The idea of VSS and preliminary results were presented as a poster at MLCB 2019. In this manuscript, we present new results from comprehensive evaluation of VSS, including the *variance instability* metric and an evaluation of how using VSS units as input to the SAGA algorithm Segway influences its results.

VSS source code is available at
`https://github.com/faezeh-bayat/Variance-stabilized-units-for-sequencing-based-genomic-signals`.

2 Methods

2.1 Identifying the mean-variance relationship

Our variance-stabilizing transformation depends on determining the mean-variance relationship for the input data set. We learn this relationship 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 the window size w such that it includes bins with weight at least 0.01; specifically, $w = -\beta \log(0.01)/b \log(2)$.

We used a smoothing spline to fit an estimated mean-variance curve $\hat{\sigma}(x)$.

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

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

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 respectively. 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)`. This process 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 $\hat{\sigma}(\mu)$.

Calculating variance-stabilized signals Having learned the mean-variance relationship, we compute VSS using the variance-stabilizing transformation [3]

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

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.

3 Results

3.1 Genomic signals 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 1a). 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. Moreover, the relationship does not match that expected by the currently-used $\log(x+1)$ and $\text{asinh}(x)$ 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. The observation that existing transformations are not variance-stabilizing was confirmed when we quantified this fit (Figure 1b). A transformation implicitly assumes that a data set has a specific mean-variance relationship; for example, a log transform assumes a linear mean-variance relationship. To measure the accuracy of a variance estimate, we used the log likelihood of a given mean-variance relationship estimate, which is maximized when the inferred variance equals the variance of the data. As expected, we found that a uniform variance model implied by using untransformed signals had a poor likelihood (average log density of -1.9), reflecting non-uniform variance (Figure 1b). We found that the variance estimates from the $\log(FE+1)$ and asinh(FE), where FE is the Fold enrichment signal, greatly improved the likelihood (average log density of -1.3 and -1.5 respectively). However, we found that mean-variance relationship learned by VSS had much better likelihood (average log density -1.2) than either transformation, indicating that the learned curve successfully models the mean-variance relationship of the data (Figure 1b).

3.2 Differences between replicates are stabilized after transformation

To measure whether a given transformation stabilizes variance in a given signal data set, we defined the variance-instability metric. This metric is defined as the variance of mean squared between-replicate differences across bins defined by signal value. Signals with unstable variance will have a large value for this metric. We found that signals transformed using VSS have better variance stability than either untransformed signals or signals after alternative transformations (Figure 1c). Fold enrichment (FE) signals transformed by either $\log(x+1)$ and $\text{asinh}(x)$ on had an average of 1.8 variance instability, whereas VSS have instability of 1.5.

3.3 VSS signals improve segmentation and genome annotation (SAGA) algorithms

To evaluate the efficacy of transformed signals as input to Gaussian models, we use segmentation and genome annotation (SAGA) as an example. Segmentation and genome annotation algorithms are widely used to integrate genomic data sets and annotate genomic regulatory elements [11, 12, 2, 24, 25]. Following previous work [24, 15], we evaluated the quality of an annotation by the correlation between the label of a gene body and whether that gene is expressed as measured by RNA-seq. We evaluated this correlation across multiple cell types and model initializations. We believe that high quality input signals will lead to a high quality annotation. We used the SAGA algorithm Segway [11] annotation for this analysis.

We found that using variance-stabilized signals from VSS improves annotations produced by SAGA algorithms (Figure 1g-l). As had been previously observed [11], using non-stabilized fold enrichment (FE) signal results in poor performance (mean r^2=0.47, Figure 1j). To account for this, Segway recommends using an asinh transform; doing so substantially improves performance (mean r^2=0.57, Figure 1j). VSS produces similar results to asinh on FE data (mean r^2=0.57, $p = 0.28$). However, VSS outperforms asinh when using raw or log Poisson p-value (LPPV) as the base signals ($p = 0.028$ and $p = 0.007$ respectively, paired one-sided Wilcoxon rank-sum test). Likewise, VSS outperforms a log transformation for FE and LPPV signals ($p = 0.023$ and $p = 0.013$ respectively). This improvement likely results from the fact that VSS stabilizes variance in all cases, whereas asinh does so only when data sets happen to have a specific mean-variance relationship.

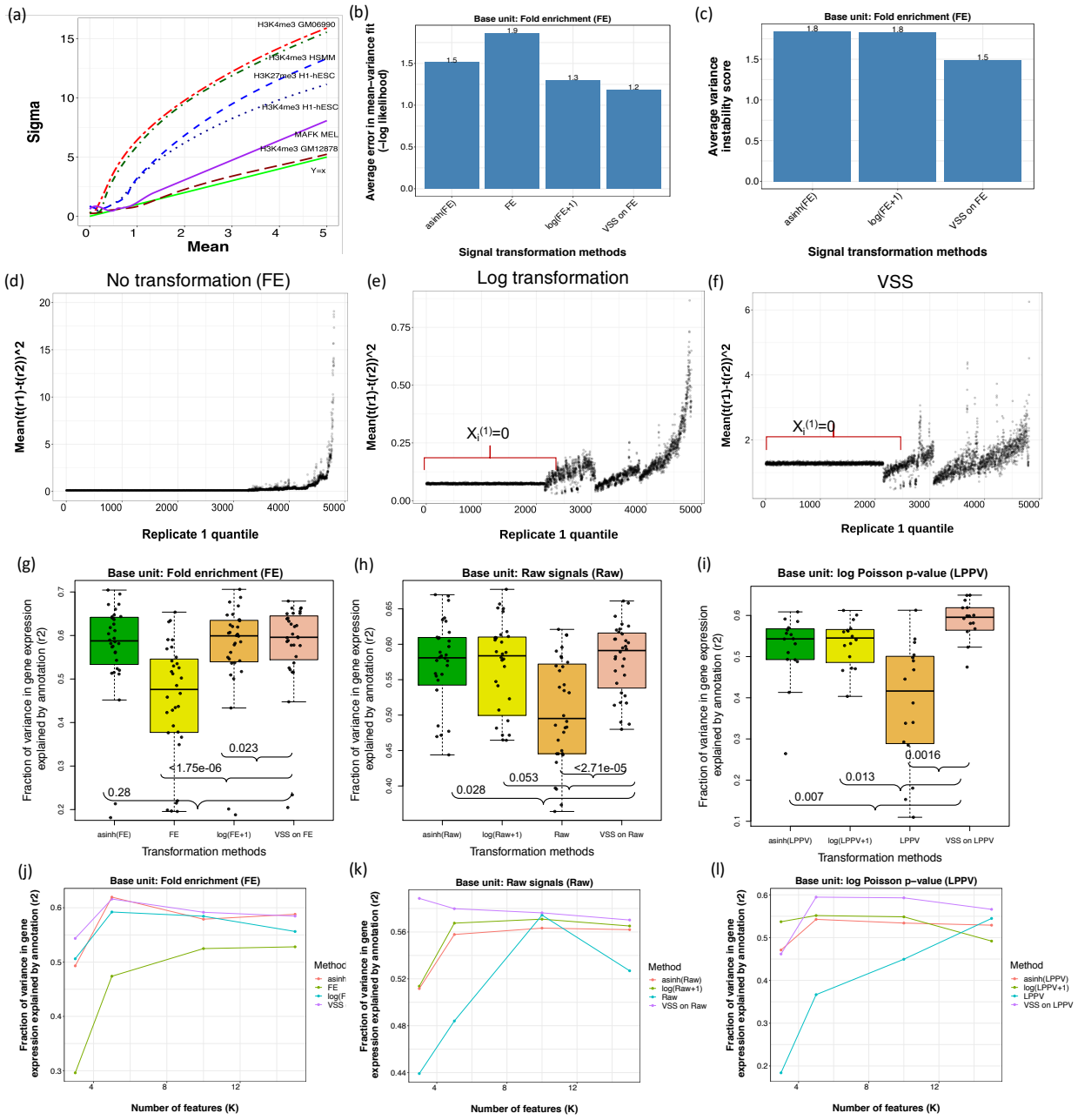

Fig. 1. General schematic of the VSS method. (a) Learned mean-variance relationships for several data sets. Horizontal and vertical axes denote mean and standard deviation respectively. (b) Goodness of fit to the mean-variance relationship averaged across experiments on Fold enrichment signals, measured by Gaussian log likelihood. Lower values of negative log likelihood indicates better fit. (c) Variance instability score averaged across experiments on Fold enrichment signals. Lower values indicate more stable variance. (d,e,f) Variance instability plots of transformed signals. Signals with stable variance show a flat (constant) trend on this plot; a trend (increasing or decreasing) indicates unstable variance. (g,h,i) and (j,k,l) Evaluation of annotations relative to gene expression on Fold enrichment, raw and P-value signals respectively. Vertical axis is the fraction of variance in gene expression explained (r2). (g,h,i) Horizontal axis are the transformation methods. (j,k,l) Horizontal axis is the number of features or states in a given model respectively (k).

References

1. Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010.
2. 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.
3. 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.
4. Timothy J Durham, Maxwell W Libbrecht, J Jeffry Howbert, Jeff Bilmes, and William Stafford Noble. Predictd parallel epigenomics data imputation with cloud-based tensor decomposition. *Nature communications*, 9(1):1–15, 2018.
5. Jason Ernst and Manolis Kellis. Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215, 2012.
6. Jason Ernst and Manolis Kellis. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature biotechnology*, 33(4):364, 2015.
7. 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.
8. Yuchun Guo, Shaun Mahony, and David K Gifford. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol*, 8(8):e1002638, 2012.
9. 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.
10. Arif Harmanci, Joel Rozowsky, and Mark Gerstein. Music: identification of enriched regions in chip-seq experiments using a mappability-corrected multiscale signal processing framework. *Genome biology*, 15(10):474, 2014.
11. Michael M Hoffman, Orion J Buske, Jie Wang, Zhiping Weng, Jeff A Bilmes, and William Stafford Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods*, 9(5):473, 2012.
12. 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.
13. Wolfgang Huber, Anja Von Heydebreck, Holger Sültmann, Annemarie Poustka, and Martin Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(suppl_1):S96–S104, 2002.
14. Donna Karolchik, Robert Baertsch, Mark Diekhans, Terrence S Furey, Angie Hinrichs, YT Lu, Krishna M Roskin, Matthias Schwartz, Charles W Sugnet, Daryl J Thomas, et al. The ucsc genome browser database. *Nucleic acids research*, 31(1):51–54, 2003.
15. Maxwell W Libbrecht, Oscar L Rodriguez, Zhiping Weng, Jeffrey A Bilmes, Michael M Hoffman, and William Stafford Noble. A unified encyclopedia of human functional dna elements through fully automated annotation of 164 human cell types. *Genome biology*, 20(1):180, 2019.
16. 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.
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. Naim U Rashid, Paul G Giresi, Joseph G Ibrahim, Wei Sun, and Jason D Lieb. Zinba integrates local covariates with dna-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome biology*, 12(7):R67, 2011.
19. Xu Ren and Pei Fen Kuan. Negative binomial additive model for RNA-Seq data analysis. *bioRxiv*, page 599811, 2019.
20. 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.
21. Lucy Whitaker. On the poisson law of small numbers. *Biometrika*, 10(1):36–71, 1914.
22. Haipeng Xing, Yifan Mo, Will Liao, and Michael Q Zhang. Genome-wide localization of protein-dna binding and histone modification by a bayesian change-point method with chip-seq data. *PLoS Comput Biol*, 8(7):e1002613, 2012.
23. 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.

24. Yu Zhang, Lin An, Feng Yue, and Ross C Hardison. Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic acids research*, 44(14):6721–6731, 2016.
25. Yu Zhang and Ross C Hardison. Accurate and reproducible functional maps in 127 human cell types via 2d genome segmentation. *Nucleic acids research*, 45(17):9823–9836, 2017.