

## MIT Open Access Articles

### *Universal Count Correction for High-Throughput Sequencing*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

**Citation:** Hashimoto, Tatsunori B., Matthew D. Edwards, and David K. Gifford. "Universal Count Correction for High-Throughput Sequencing." Edited by Alice Carolyn McHardy. PLoS Comput Biol 10, no. 3 (March 6, 2014): e1003494.

**As Published:** <http://dx.doi.org/10.1371/journal.pcbi.1003494>

**Publisher:** Public Library of Science

**Persistent URL:** <http://hdl.handle.net/1721.1/86363>

**Version:** Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

**Terms of use:** Creative Commons Attribution



# Universal Count Correction for High-Throughput Sequencing

Tatsunori B. Hashimoto, Matthew D. Edwards, David K. Gifford\*

Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

## Abstract

We show that existing RNA-seq, DNase-seq, and ChIP-seq data exhibit overdispersed per-base read count distributions that are not matched to existing computational method assumptions. To compensate for this overdispersion we introduce a nonparametric and universal method for processing per-base sequencing read count data called FIXSEQ. We demonstrate that FIXSEQ substantially improves the performance of existing RNA-seq, DNase-seq, and ChIP-seq analysis tools when compared with existing alternatives.

**Citation:** Hashimoto TB, Edwards MD, Gifford DK (2014) Universal Count Correction for High-Throughput Sequencing. *PLoS Comput Biol* 10(3): e1003494. doi:10.1371/journal.pcbi.1003494

**Editor:** Alice Carolyn McHardy, Heinrich Heine University, Germany

**Received:** April 1, 2013; **Accepted:** January 8, 2014; **Published:** March 6, 2014

**Copyright:** © 2014 Hashimoto et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This study was funded by NIH grant no. 5-U01-HG007037, NSF grant no. 0645960, and the Qatar Computing Research Institute. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: gifford@mit.edu

This is a *PLoS Computational Biology* Methods article.

## Introduction

High-throughput sequencing is used in a variety of molecular counting assays [1] to study protein-DNA binding, transcription, and the dynamics of chromatin occupancy. ChIP-seq [2], used to study protein binding to the genome, captures short DNA fragments that are attached to a protein of interest after a chemical treatment that affixes proteins to nearby DNA molecules. High-throughput sequencing of these DNA fragments, followed by identifying their originating location in the genome, allows for the identification of read-enriched areas. These enriched regions correspond to locations where the protein of interest was bound, perhaps indirectly, to the DNA. RNA-seq [3,4], used to study gene expression, requires isolating the RNA content of a sample, converting it to DNA, and sequencing the resulting DNA library. Mapping or assembling the DNA reads and assigning them to exons or transcripts enables the genome-wide quantification of gene expression. DNase-seq [5,6] identifies regions of open chromatin and patterns of transcription factor binding by employing an enzyme that preferentially cuts DNA in accessible positions. Retrieving and sequencing the resulting library of DNA fragments, followed by identification of the originating locations, allows for a genome-wide characterization of chromatin occupancy. A unifying task in these analyses is comparing read count profiles, obtained from read mapping results, across varying biological samples or experimental conditions.

Although a myriad of specialized methods exist for analyzing read count data, it is frequently assumed (implicitly or explicitly) that read counts are generated according to a Poisson distribution with a local mean. The assumption is explicitly introduced by using the Poisson density directly as well as implicitly by relying on binned per-base counts in ranking and statistical testing (see Text

S1). When read count data exhibit overdispersed per-base read distributions, a Poisson model may produce erroneous or noisy results. This occurs because the data are not matched to the modeling assumption, resulting in incorrect assessments of statistical significance. While it is well known that the distribution of per-base counts within a single experiment is in fact typically overdispersed, there has not been a precise characterization of the degree of overdispersion and its effects on downstream analysis for general sequencing data.

We introduce a general and asymptotically correct preprocessing technique called FIXSEQ for correcting per-base and per-experiment read counts. FIXSEQ reduces noise and increases stability in subsequent inference procedures and complements existing literature on applications of heavy-tailed distributions [7–9]. Existing literature for preprocessing focuses on either de-duplication, which removes all but one read per base, or normalization techniques for RNA-seq data, which generally operate over exon-level counts. We have previously dealt with these problems when developing a ChIP-seq caller with adaptive count truncation [10] and found that this was effective in practice (see Text S1 and Table S1), but this work aims to construct a more general preprocessing scheme that works for any method and sequencing assay.

The normalization strategy of de-duplication is prevalent in multiple ChIP-seq peak callers [11], but less common in RNA-seq data analysis where highly-expressed transcripts may be expected to have many duplicate reads. However, a handful of RNA-seq processing algorithms remove duplicates as a conservative choice to avoid nonlinear PCR amplification errors [12,13].

Existing RNA-seq normalization techniques work at a higher conceptual level than FIXSEQ, using information from a local sequence context to correct exon- or transcript-level sums and reduce the impact of confounding noise covariates [7,14–18]. In our RNA-seq results we show that FIXSEQ can enhance the results of methods such as DEseq that already account for exon-level

## Author Summary

High-throughput DNA sequencing has been adapted to measure diverse biological state information including RNA expression, chromatin accessibility, and transcription factor binding to the genome. The accurate inference of biological mechanism from sequence counts requires a model of how sequence counts are distributed. We show that presently used sequence count distribution models are typically inaccurate and present a new method called FIXSEQ to process counts to more closely follow existing count models. On typical datasets FIXSEQ improves the performance of existing tools for RNA-seq, DNase-seq, and ChIP-seq, while yielding complementary additional gains in cases where domain-specific tools are available.

overdispersion as well as provide complementary information to methods that correct for mappability and GC content (see Figure S2). While these methods are valuable for RNA-seq and binned count statistics, they are less applicable to other sequencing data types and often have specific modeling assumptions that rely on the mechanisms of transcription, cDNA library preparation, and DNA sequencing. Covariate-based normalization techniques designed for other assays, like ChIP-seq, require identified binding sites as a prerequisite and are designed to correct only windowed read counts [19], rendering them unusable as preprocessing tools for algorithms which require per-base count data. Other assay-specific normalization tools (e.g. [20–23]) require extensive domain knowledge and application-specific modeling strategies and, while valuable, must be developed independently for each new assay.

In contrast to most of these existing strategies, FIXSEQ works at a lower and more general level, the per-base count, and attempts to decrease the false positive rate rather than recover lost signal caused by sequencing artifacts. This approach is applicable to many types of sequencing assays and downstream processing algorithms, without additional assumptions. This universal nature allows for FIXSEQ to be applied to any type of sequencing count data, without training phases or specialized model-building. However, in cases where applicable covariate-based or assay-specific normalization tools exist, they may be used in addition to FIXSEQ in order to leverage complementary gains (as in Figure S2). One additional consequence of our work is a generalization of the de-duplication heuristic into a broader and asymptotically correct preprocessing technique.

## Results

### Read counts in sequencing data are highly overdispersed

The distributions of per-base mapped read counts in all ChIP-seq and RNA-seq runs for the human embryonic stem cell type (H1-hESC or ES cells) in the ENCODE project and a set of K562 cell line DNase-seq experiments (see Text S1) show evidence of consistent and significant overdispersion (Figures 1, 2, and 3). This extra variation arises from a myriad of biological and technical sources, including true variation in factor binding signal or expression levels genome-wide, variation in molecular sequencing affinity, and variation in read mapping accuracy. The overdispersion we find is complex and cannot be directly categorized as the result of a well-known parametric distribution such as gamma-Poisson [24,25] or log-normal-Poisson [26] (Figures 1 and 2).

We quantified the degree of overdispersion with respect to a distribution by comparing per-base empirical log-likelihoods against the per-base maximum log-likelihood distributions for

the Poisson, negative binomial, and log-normal Poisson, where the per-base rates of the Poisson are assumed to be drawn from a log-normal distribution. For the negative binomial and log-normal Poisson, maximum likelihood distributions were found via numerical optimization with randomized restarts.

The deviation from Poisson is consistent across experiment and assay type, as shown in the left column of Figure 3. We would expect a completely linear histogram of log-counts but actually observe significant overdispersion, shown by large number of high-count bases.

Figure 1 shows that none of the parametric distributions we tested fit the observed counts well. The Poisson significantly underestimates the number of bases with more than one read, with the probability of having ten counts land on the same base estimated to be  $e^{40}$  times less than the observed number of counts. The negative binomial has previously been shown to be effective for modeling exon-level RNA-seq data, and we confirm that negative binomial fits per-base RNA-seq data well. However, we find that it fails to capture the wide variation in overdispersion for ChIP-seq, underestimating the high count bases by at most a factor of  $e^7$ . The log-normal Poisson fit shows that real-world sequencing data is not simply heavy-tailed; it has heavy tails whose shapes are dependent on the assay type. The log-normal Poisson is traditionally considered an extremely heavy-tailed distribution, and while it is relatively correct for ChIP-seq, it significantly overestimates the tail mass for DNase-seq experiments.

The wide variation in overdispersion level and type suggests that any single parametric approach is unlikely to be effective for all assay types. Instead of attempting to model each assay type with a separate parametric family, we will use nonparametric distributions that are flexible enough to fit all observed assay types well.

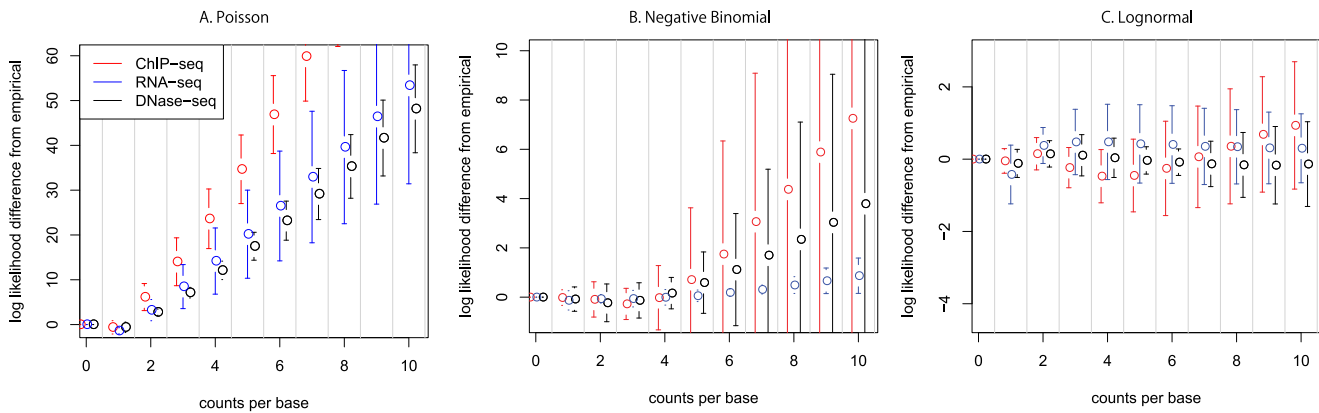
### Count correction via data transformation

While we have already seen that the Poisson assumption fails for most of the assay types we consider, it is not feasible or necessary to modify every analysis algorithm to use overdispersed distributions. Instead, for the class of inference algorithms which implicitly or explicitly assume Poisson counts and independent bases, such as most ChIP-seq callers, DNase-binding identifiers, and RNA-seq exon read counting methods, we can construct improved datasets with transformed counts that correct for overdispersion.

The two major nonparametric approaches to data transformation are quantile normalization, which matches input samples to a reference sample via their quantiles, and distribution matching, which fits a distribution to both the input and reference and constructs a mapping function between them.

Quantile normalization, which is a popular approach in the microarray literature, cannot easily be adapted to sequencing data, due to the large number of bases with equal counts. In order to rank normalize our observed counts to a Poisson, we would have to arbitrarily break ties between bases with equal reads, which could lead to spurious inference as well as force bases with non-zero counts to be discarded.

Instead of breaking ties, we employ a different approach to distribution mapping: given a distribution  $f$  over counts, we find the mapping which makes the density of the Poisson equal to that of the given distribution  $f$ . By fitting a distribution  $f$  to the observed counts, we avoid the problem of tied counts and allow for a continuous mapping. One advantage of viewing this approach under a distribution mapping framework is that it allows us to understand the theoretical basis of the de-duplication heuristic that is a popular preprocessing method for ChIP-seq data.



**Figure 1. Log-likelihood difference between maximum likelihood and observed distributions.** Differences in log-likelihood per base between the fitted model and the empirical distribution, also interpreted as the log-difference between observed and fitted counts. This error metric represents the error when calculating p-values or significance tests using a Poisson assumption. Three assay types are shown in each panel, analyzed by three models: (a) Poisson. (b) Negative binomial. (c) Log-normal Poisson. A model that fits the data would have points along the  $x=0$  line. Consistent deviation from zero by all distributions show that none of the distributions fit all assays well. doi:10.1371/journal.pcbi.1003494.g001

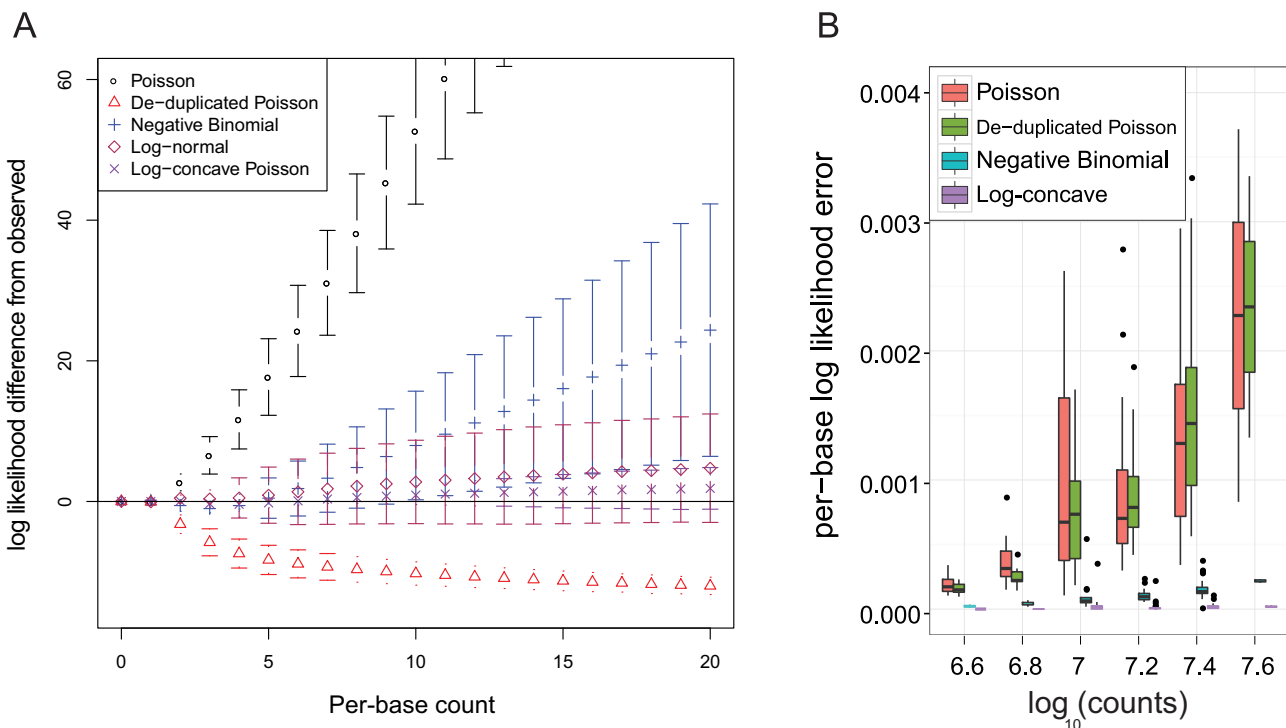
Our approach transforms the non-Poisson, curved count histogram on the left column of Figure 3 into the Poisson-like linear count histogram on the right column of Figure 3.

**De-duplication acts as a degenerate data transform**

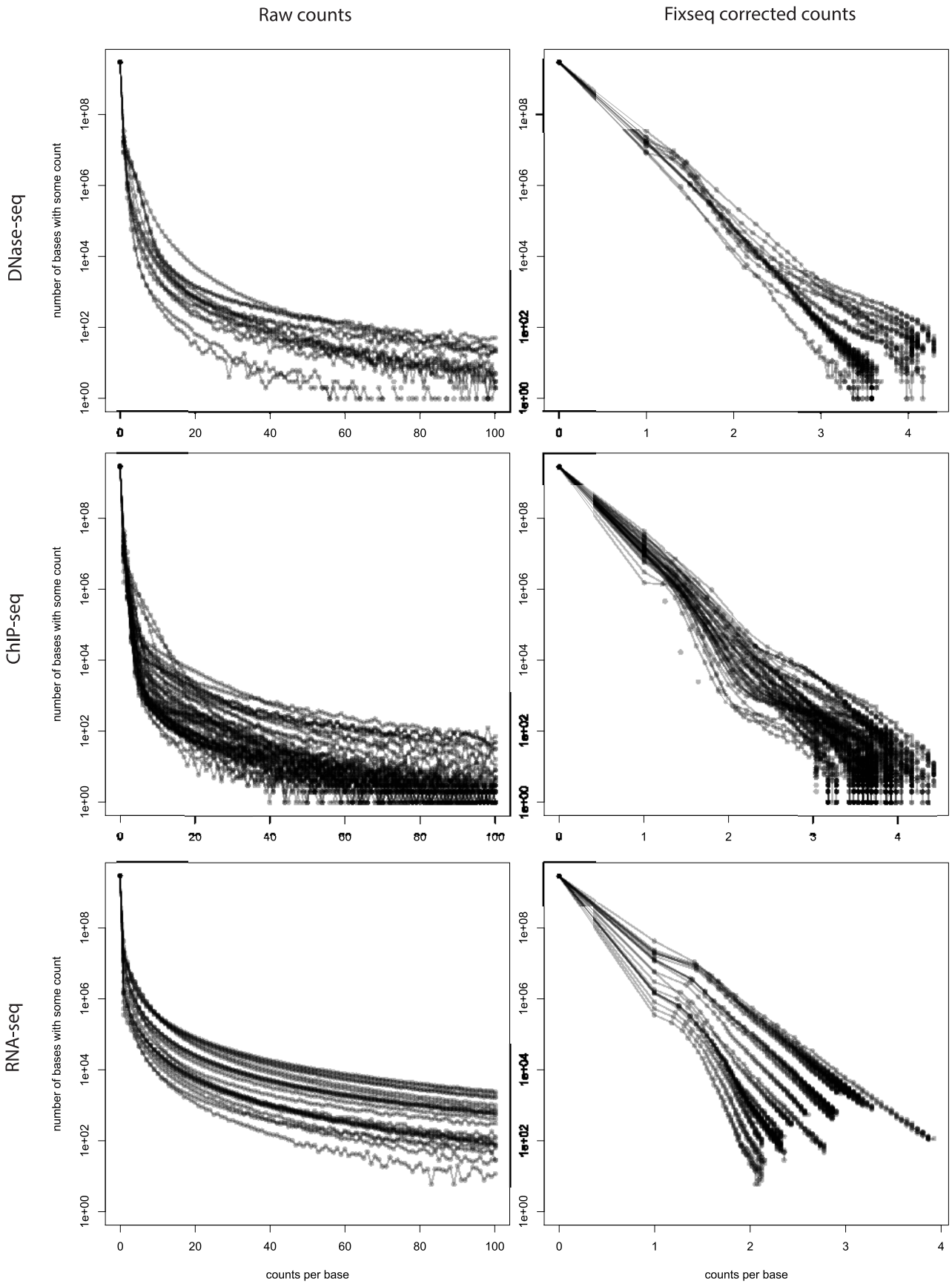
De-duplication, or removal of all but one read at each base position, has gained adoption in the ChIP-seq analysis literature as an effective way of reducing noise and improving replicate consistency [27,28]. ChIP-seq event callers such as MACS [29]

and SPP [30] either de-duplicate by default or strongly suggest enabling de-duplication.

The heuristic of de-duplication can be derived as a distribution mapping data transformation by assuming that the read counts arise from a degenerate count distribution, where the number of bases with non-zero reads is drawn from a binomial, and the number of reads at non-zero bases is drawn from a uniform noise component over  $[1, \infty)$ . In this case, the probability of all non-zero counts are equal, and they should be mapped to the same value



**Figure 2. Log-likelihood errors for ENCODE ChIP-seq data.** The distribution described as log-concave is the statistical model used in FIXSEQ. Subfigure (a) shows that de-duplication Poisson can control high per-base errors much like overdispersed models, but Subfigure (b) shows that de-duplication error rises rapidly as sequencing depth increases. doi:10.1371/journal.pcbi.1003494.g002



**Figure 3. Distribution of counts for 14 DNase-seq experiments, 23 RNA-seq experiments and 87 ChIP-seq experiments.** The left panel shows uncorrected counts, and the right shows counts after correction. Poisson distributed counts would follow a straight line; all experiments show significant deviation from linearity that is corrected by FIXSEQ. doi:10.1371/journal.pcbi.1003494.g003

after the transform. Conversely, any data transform preserving rank order will not fully de-duplicate but will instead monotonically re-weight counts.

De-duplication works well in practice by drastically reducing the error and additional variance from overdispersion, despite assuming that the data follow a degenerate distribution. Figure 2a shows the performance of various overdispersion correction methods using the per-base log-likelihood error for the Poisson, de-duplicated Poisson, negative binomial, and the log-concave Poisson distribution, which we use in FIXSEQ. These per-base errors reflect the expected error in statistical significance testing under window-based DNase-seq and ChIP-seq callers, as errors in the log-likelihood propagate directly into error in the Poisson test statistic.

Preprocessing by de-duplication does not continue to reduce per-base errors as sequencing depth increases. Figure 2b shows that the log-likelihood error per-base increases rapidly as a function of the total sequencing depth. The Poisson and de-duplicated Poisson both have average per-base error which increases similarly as a function of sequencing depth. This confirms the observation that as sequencing depth increases, more of the mappable genome will have at least one mapped read, leading to a loss of predictive power.

Therefore while de-duplication may be effective at lower sequencing depths, it relies upon a limited heuristic justification and will not remain effective as sequencing depths increase. On the other hand, FIXSEQ significantly outperforms de-duplication at modeling the observed data distribution as sequencing depth grows (Figure 2b) and is asymptotically consistent under relatively weak assumptions.

We compared three methods of count preprocessing: original (raw) counts, removal of all duplicates (de-duplication), and our novel preprocessing technique (FIXSEQ). These preprocessing schemes are compared across three assay types and in multiple experiments and in multiple contexts. We show that FIXSEQ consistently improves performance, with substantial improvements obtained in certain cases.

### DNase-seq

We evaluate our model on the ability to identify transcription factor binding sites based upon DNase-seq counts on the ENCODE human K562 DNase-seq data using two different methods: an unsupervised task using the CENTIPEDE binding site caller [31] and a supervised task using a linear classifier. The binding site predictions are compared against all matching ChIP-seq calls for the same factor on a matched cell type, and we evaluate the algorithm on the fraction of ChIP-seq calls we are able to recover. The details of the comparison, such as the PWM matching and cutoffs, follow the techniques used by CENTIPEDE.

In the unsupervised task shown in Figure 4, FIXSEQ shows small but consistent improvements on nearly all runs and all methods, and on many factors we show improvements up to a 0.3 increase in area under the curve (AUC), a metric of accuracy. These large performance increases indicate that FIXSEQ rescued an otherwise failed run.

### ChIP-seq

We tested FIXSEQ on 87 ES cell ChIP-seq experiments from the ENCODE project [32], using the ChIP-seq callers MACS [29]

and PeakSeq [33] on original counts, de-duplicated counts, and FIXSEQ processed counts with rounding (see Text S1). Following prior work in evaluating ChIP-seq caller accuracy [28,34], we selected two evaluation criteria: replicate consistency of q-values and the number of overlapping ChIP-seq events across replicates.

We evaluate quantile-quantile correlation for replicate consistency, as this allows us to evaluate the distribution of q-values generated by each method without pairing binding sites explicitly. The quantile-quantile (QQ) correlations are an effective means of detecting not only whether we call similar numbers of binding sites across replicates, but also whether our ChIP-seq call confidence is consistent across replicates. The quantile-quantile correlations across all analyzed ENCODE ChIP-seq experiments shown in Figure 5a strongly suggest that FIXSEQ stabilizes the distribution of q-values from PeakSeq and MACS. FIXSEQ outperforms both raw counts and de-duplication for PeakSeq and improves on de-duplication significantly for MACS ( $p < 3.1 \cdot 10^{-5}$ ).

An alternative measure of ChIP-seq experiment quality is the number and size of overlapping sites across replicates. FIXSEQ increases the number of overlapping sites in both methods, showing that FIXSEQ improves consistency of localization of sites as well as the ranking of ChIP-sites (Figure 5b).

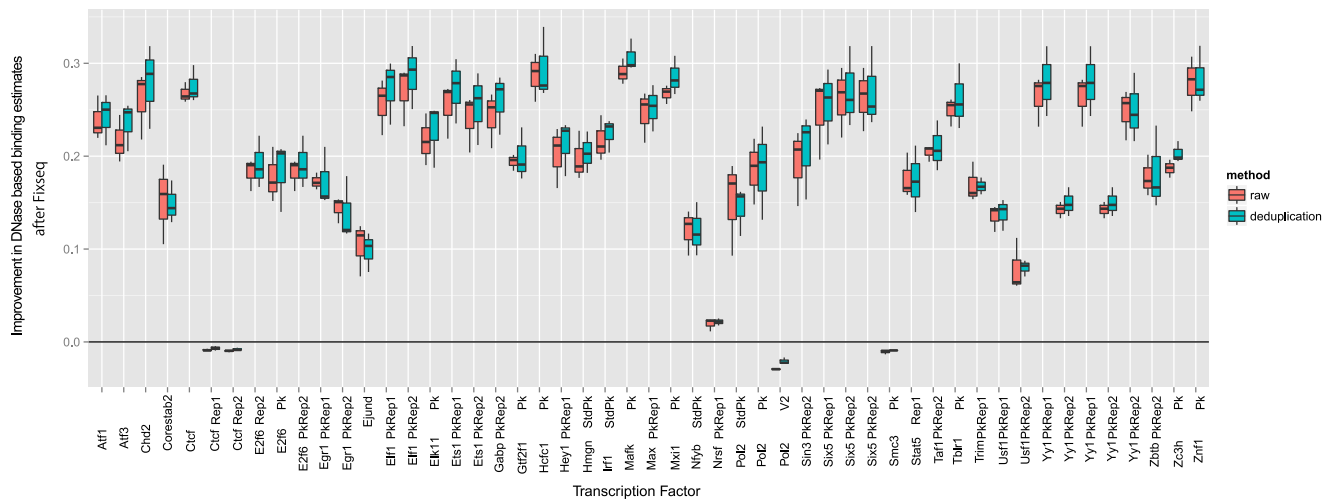
### RNA-seq

We ran FIXSEQ on all 23 ES cell RNA-seq datasets from ENCODE and evaluated the replicate consistency of the original read counts, de-duplication, and FIXSEQ. Using the ENCODE alignments, we followed the analysis technique suggested by DEseq [7] and mapped reads and adjusted counts to exons and generated exon-level counts.

Replicate consistency was measured in two ways: Spearman's rank correlation and the number of false positive differential expression events called by DEseq [7] across replicates. Spearman's rank correlation on exon counts was chosen to characterize the run-to-run variability between replicates, while DEseq was chosen to represent FIXSEQ's ability to enhance existing techniques that attempt to handle exon-level overdispersion.

The rank correlation between replicates shown in Figure 6 is significantly higher for FIXSEQ processed counts than for both raw counts and de-duplication ( $p < 1.1 \cdot 10^{-11}$ ). We make even greater improvements in the 75-bp single-end RNA-seq datasets, where it is possible that difficulty or ambiguity in read mapping causes single-base spikes that adversely affect replicate consistency. The RNA-seq correlations also support our earlier claims that de-duplication performance will begin to degrade as sequencing depth increases. In both the paired-end Caltech and Cold Spring Harbor Laboratory (CSHL) experiments we find that the original counts are on average better than the de-duplicated counts due to the higher coverage per base.

Our DEseq results in Table 1 are consistent with the correlation results, with FIXSEQ calling fewer differential exons across replicates despite being a less aggressive truncation scheme than de-duplication. Following the DEseq analysis pipeline, we used replicates to estimate exon-level overdispersion and identified exons differential across replicates at 0.05% FDR. The number of exons called differential across replicates in Table 1 are consistently the lowest for FIXSEQ out of all methods tested. Since



**Figure 4. AUC comparisons for baseline methods compared to FIXSEQ.** Boxplots depicting AUC improvement across multiple factors (boxes above zero represent improvement due to FIXSEQ). Variance was estimated using one thousand bootstrap replicates each. Transcription factors with no significant difference across methods ( $p=0.05$ ) are not shown. Of the 301 factors tested, 90 could be predicted nontrivially ( $AUC>0.55$ ) via CENTIPEDE. Of these, 51 show significant differences. doi:10.1371/journal.pcbi.1003494.g004

FIXSEQ is a shrinkage method, we would expect there to be fewer false positives between replicates under FIXSEQ than the original counts. For example, if we preprocess by deleting all counts, we would have a trivial zero false positive rate. However, this is likely not the method by which FIXSEQ decreases false positive rate since we outperform de-duplication, which is an even more aggressive shrinkage method. This suggests that the counts we retain are consistent between replicates.

## Discussion

We have shown that per-base count overdispersion is a widespread and consistent phenomenon in high-throughput sequencing experiments. While correcting for exon-level overdispersion has been studied in RNA-seq, per-base methods and corresponding tools for ChIP-seq and DNase-seq have largely been unexplored outside of aggressive count truncation methods particular to individual algorithms. One reason for the slow adoption of overdispersed models has been the empirical success of the de-duplication heuristic as a preprocessing scheme. However, we show that de-duplication assumes the data arise from a degenerate distribution, and that the performance of de-duplication will degrade as sequencing depth increases.

FIXSEQ corrects overdispersed sequence count data by assuming that the data arise from a flexible class of log-concave distributions. We designed a novel and fast inference algorithm for the class of Poisson log-concave distributions as well as effective rounding schemes. In a diverse array of validation tasks, including DNase-seq binding site identification, ChIP-seq peak calling, and RNA-seq self-consistency, FIXSEQ consistently increased performance compared to both original counts and de-duplication. In cases where domain-specific correction schemes exist, FIXSEQ can operate in conjunction with them to yield complementary gains. While not replacing other sophisticated methods that can model the intricate biological realities of a new sequencing assay, FIXSEQ aims to provide a useful solution for all count-based sequencing assays without modification for new protocols.

The FIXSEQ method has the potential of broadly improving inference for high-throughput sequencing by bringing sophisticated overdispersion correction to a large number of existing analysis

pipelines, while being applicable to future assays without lengthy development and modeling cycles. Additionally, the modeling and inference results we presented can be used in new flexible analysis procedures for count data.

## Methods

Our count preprocessing method, FIXSEQ, consists of three components:

1. Parameter inference for a novel class of distributions called log-concave Poisson distributions.
2. A probability integral transform method to map counts generated under log-concave Poisson to a Poisson distribution.
3. Rounding techniques to adapt datasets to methods that utilize only integral counts.

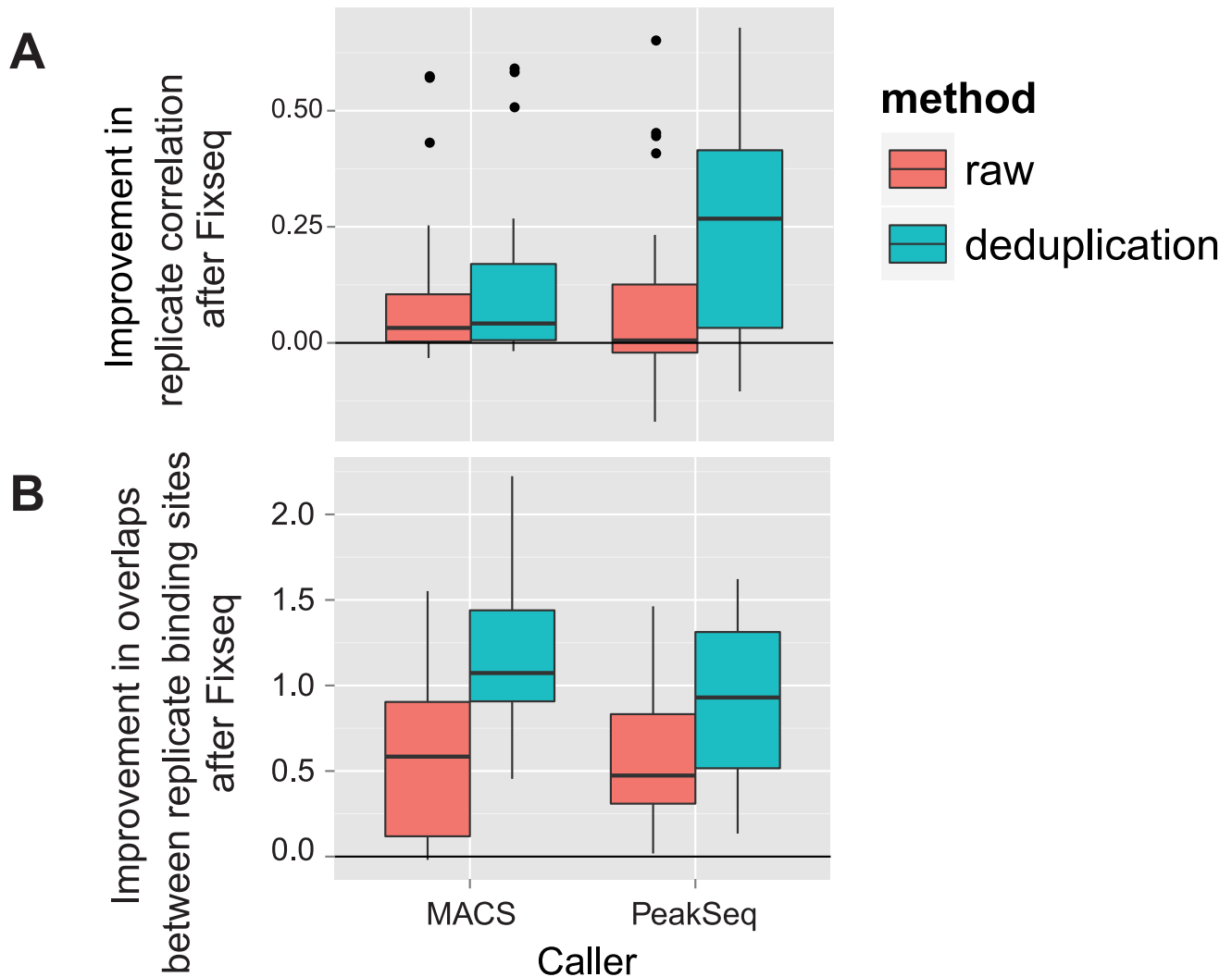
In the case that the algorithm downstream of our method is able to take weighted counts, FIXSEQ inherits all the favorable properties of the maximum likelihood estimator (MLE) and can guarantee unbiased and asymptotically consistent inference under the assumptions of per-base independence and log-concave count distributions.

### Poisson log-concave distributions

The challenge of constructing a universal preprocessor is finding a class of count distributions that is flexible enough to model a variety of assay types while remaining non-degenerate. We achieve this goal by letting the per-base rates of a Poisson distribution be drawn from a nonparametric class of distributions called log-concave. Log-concave distributions are a family of distributions  $f$  for which the log-density is a concave function. This allows us to write any log-concave function in terms of  $\phi$ , a concave function:

$$p(x|\phi) \propto \exp(\phi(x)).$$

The log-concave family includes a large family of unimodal distributions, such as most of the exponential family (including



**Figure 5. Event reproducibility for ChIP replicates.** Subfigure (a) shows that FIXSEQ increases the q-value correlation between replicates. Subfigure (b) shows the increase in the log number of bases overlapping between calls on replicates due to FIXSEQ. FIXSEQ calls have consistently higher peak overlaps between replicates for both MACS and PeakSeq. doi:10.1371/journal.pcbi.1003494.g005

common cases such as the normal, gamma with shape parameter greater than one, Dirichlet [35]. Important exceptions include all multi-modal distributions and distributions with super-exponential tails such as the  $t$ -distribution or the Cauchy.

In sequencing experiments log-concave Poisson families are capable of modeling zero-inflation as well as mixtures induced by copy number variation for low Poisson rates with  $\lambda < 1$ , where the overall distribution remains unimodal. If such distributions are needed, straightforward extensions for mixtures of log-concave distribution are well known [36].

Our algorithmic contribution is the use of compound log-concave distributions, where we use latent log-concave distributions which generate Poisson counts along the genome. Inference for latent log-concave distributions does not follow directly from recent results in log-concave density estimation because of the ambiguity of parameters in the latent space.

The full model is as follows: per-base counts  $c_i$  are generated by per-base log-rates  $\eta_i$ , which are drawn from a log-concave distribution with density  $\exp(\phi(\eta_i))$ :

$$C_i \sim \text{Pois}(\exp(\eta_i))$$

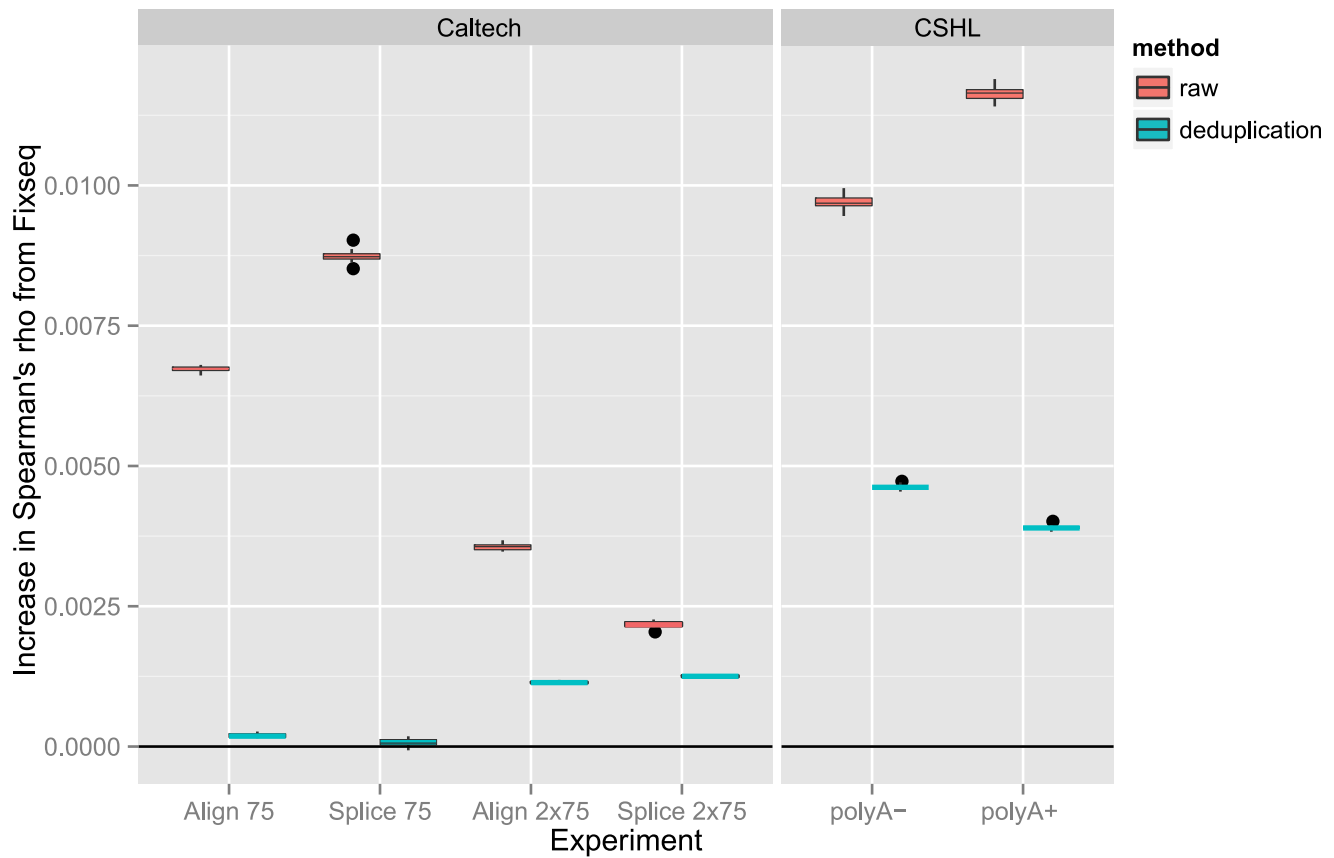
$$P(\eta_i) \propto \exp(\phi(\eta_i)).$$

Note that the two exponential operators above are intentional:  $\eta$  is a log-rate and therefore is exponentiated to become the Poisson rate, while  $\phi$  is a log-density and therefore is exponentiated to create an unnormalized density  $\exp(\phi(x))$ .

The form of this model naturally suggests an expectation-maximization strategy, which has been shown to be effective for clustering tasks [37]. However, while we can perform expectation maximization using numerical quadrature, we find in practice that the algorithm is unstable and converges extremely slowly.

Instead we propose an inference technique based upon accelerated proximal gradient descent. The marginal likelihood for counts can be written as:





**Figure 6. Bootstrapped differences in Spearman's rank correlation coefficient for ENCODE RNA-seq experiments.** Higher rank correlations across exon expression measurements between replicates indicate greater data quality and reproducibility. FIXSEQ increases the rank correlation coefficient for almost all experiments, though the improvements compared to de-duplication for the CSHL 75bp align and splice experiments are minimal.

doi:10.1371/journal.pcbi.1003494.g006

$$P(c) = \frac{\int_{-\infty}^{\infty} \text{Pois}(c; \exp(\eta)) \exp(\phi(\eta)) d\eta}{\sum_{k=0}^{\infty} \int_{-\infty}^{\infty} \text{Pois}(k; \exp(\eta)) \exp(\phi(\eta)) d\eta}.$$

The bottom term normalizes the log-concave distribution. Approximating the integral with a sum over  $N$  quadrature points  $\eta_i$  we obtain:

$$P(c) = \frac{\sum_{i=1}^N \text{Pois}(c; \exp(\eta_i)) \exp(\phi(\eta_i))}{\sum_{k=0}^{\infty} \sum_{i=1}^N \text{Pois}(k; \exp(\eta_i)) \exp(\phi(\eta_i))}.$$

Since  $\phi$  is always evaluated at the fixed points  $\eta_i$ , we can use the shorthand  $\phi_i = \phi(\eta_i)$  and let  $c_k$  be the number of bases with  $k$  counts. Then the maximum likelihood estimator for the observed data is given by:

$$\hat{\phi} = \operatorname{argmax}_{\phi \in \{\text{concave functions}\}} \sum_{k=1}^{\infty} c_k \log \left( \frac{\sum_{i=1}^N \text{Pois}(k; \exp(\eta_i)) \exp(\phi_i)}{\sum_{j=0}^{\infty} \sum_{i=1}^N \text{Pois}(j; \exp(\eta_i)) \exp(\phi_i)} \right).$$

Both the objective function and constraints are concave, and therefore we can use accelerated gradient descent to quickly find the global optimum [38]. In particular, we use a method called proximal gradient descent, which optimizes a objective function of

**Table 1.** Differential expression results across replicate experiments.

	Caltech Align 75	Caltech Splice 75	Caltech Align 75x2	Caltech Splice 75x2	CSHL polyA-	CSHL polyA+
Original	2903	1454	5719	2748	8403	6955
De-duplication	2559	1033	4951	2027	6640	5253
FIXSEQ	<b>2213</b>	<b>944</b>	<b>4230</b>	<b>1989</b>	<b>5319</b>	<b>4374</b>

Number of exons falsely called differentially expressed between biological replicates by DEseq at 5% FDR level; entries with the fewest false calls are bolded. There are 256324 total exons in the annotation set.

doi:10.1371/journal.pcbi.1003494.t001

the form  $\max_{x \in Q} llh(x)$  by repeatedly applying:

$$x_{t+1} = \text{Proj}_Q(x_t + \nabla llh(x)).$$

$\text{Proj}_Q(x)$  is defined as the projection of  $x$  onto  $Q$ .

Our gradient,  $\frac{\delta llh(\phi)}{\delta \phi_i}$ , is easily written in terms of the shorthand,  $g(k, \eta_i | \phi) = \text{Pois}(k; \exp(\eta_i)) \exp(\phi_i)$ , as:

$$\frac{\delta llh(\phi)}{\phi_i} = \left( \frac{\sum_{k=0}^{\infty} c_k \frac{g(k, \eta_i | \phi)}{\sum_{i=1}^N g(k, \eta_i | \phi)}}{\sum_{k=0}^{\infty} c_k} \right) - \frac{(\sum_{k=0}^{\infty} c_k) (\sum_{k=0}^{\infty} g(k, \eta_i | \phi))}{\sum_{k=0}^{\infty} \sum_{i=1}^N g(k, \eta_i | \phi)}.$$

This gradient has a straightforward interpretation: the first term is the distribution of  $\eta_i$  when observing the counts  $c_k$  and the second term is the distribution of  $\eta_i$  predicted from the prior  $\phi$  alone. The gradient works to minimize the difference between these observed and prior terms.

The projection operator  $\text{Proj}_Q(x)$  taking  $(\eta_i, \phi_i)$  and producing the closest concave  $\phi_i$  is the well-known concave regression algorithm [39].

The inference algorithm is guaranteed to converge to a global optima of the quadrature approximation, which as the number of quadrature points increase will converge to the global optima. If there are sufficiently many quadrature points, Fixseq will converge to the log-concave distribution closest to the data-generating distribution in the KL-divergence sense [37]. For the results, we use one million quadrature points throughout.

When compared to the naive expectation maximization based method, our algorithm converges more quickly, with average runtime on our DNase datasets reducing from  $1.2 \pm 0.4$  hours per dataset for EM down to  $23 \pm 10$  minutes for the gradient based method on a standard laptop with Intel i7 2.5ghz, with a slight increase in goodness of fit for the gradient approach.

### Count adjustment via probability integral transform

Once we fit a log-concave distribution  $\phi$ , we need to be able to convert counts generated under the log-concave Poisson into those generated by the continuous extension of the Poisson. We will define the transformation from raw counts to processed counts via the probability integral transform.

Throughout this section, we will use the continuous extension of the Poisson PDF, CDF and the analogous densities for the log-concave compound distributions, defined below as:

$$P(c_i | \eta) = \frac{\eta^{c_i} \exp(-\eta)}{\Gamma(c_i + 1)}$$

$$F(c_i | \eta) = \int_0^{c_i} P(t | \eta) dt = \frac{\Gamma(c_i + 1, \eta)}{\Gamma(c_i + 1)}$$

$$Q(c_i | \phi) = \int_0^{\infty} P(c_i | \eta) \exp(\phi(\eta)) d\eta$$

$$H(c_i | \phi) = \int_0^{c_i} \int_0^{\infty} P(t | \eta) \exp(\phi(\eta)) d\eta dt.$$

Given the continuous extensions, we can apply the probability integral transform directly. Given  $x \sim Q(c_i | \phi)$ , we can generate a uniform random variable  $y \sim H(x | \phi)$ , from which we define the Poissonization transform,  $F^{-1}(H(x | \phi) | \eta) \sim P(c_i | \eta)$ . Since we are applying the probability integral transform to the continuous extensions of the Poisson, we are not guaranteed integral counts or consistency properties generally implied by the probability integral transform. However, for our purposes it is sufficient that the quantiles and densities are matched.

Our preprocessing function  $G(x, \eta) = F^{-1}(H(x | \phi) | \eta)$  takes any  $x$  distributed as Poisson-log concave with latent distribution  $\exp(\phi)$  and returns adjusted counts distributed as Poisson. This operation preserves all of the joint structure of  $x$  and acts as a black box which exchanges the Poisson assumption used in a method for a compound Poisson log-concave distribution assumption. Alternatively, one can consider using  $G(x, \eta)$  to be a re-weighting operation, which ‘fixes’ the underestimated tail density of the Poisson.

Examples of the  $G$  function for various ENCODE assays are shown as Figure S1.

Finally,  $G(x, \eta)$  contains a free parameter  $\eta$  which we can choose freely. While any  $\eta$  would be essentially equivalent, we choose to set  $\eta$  to be the median of the latent density throughout our results.

### Rounding schemes

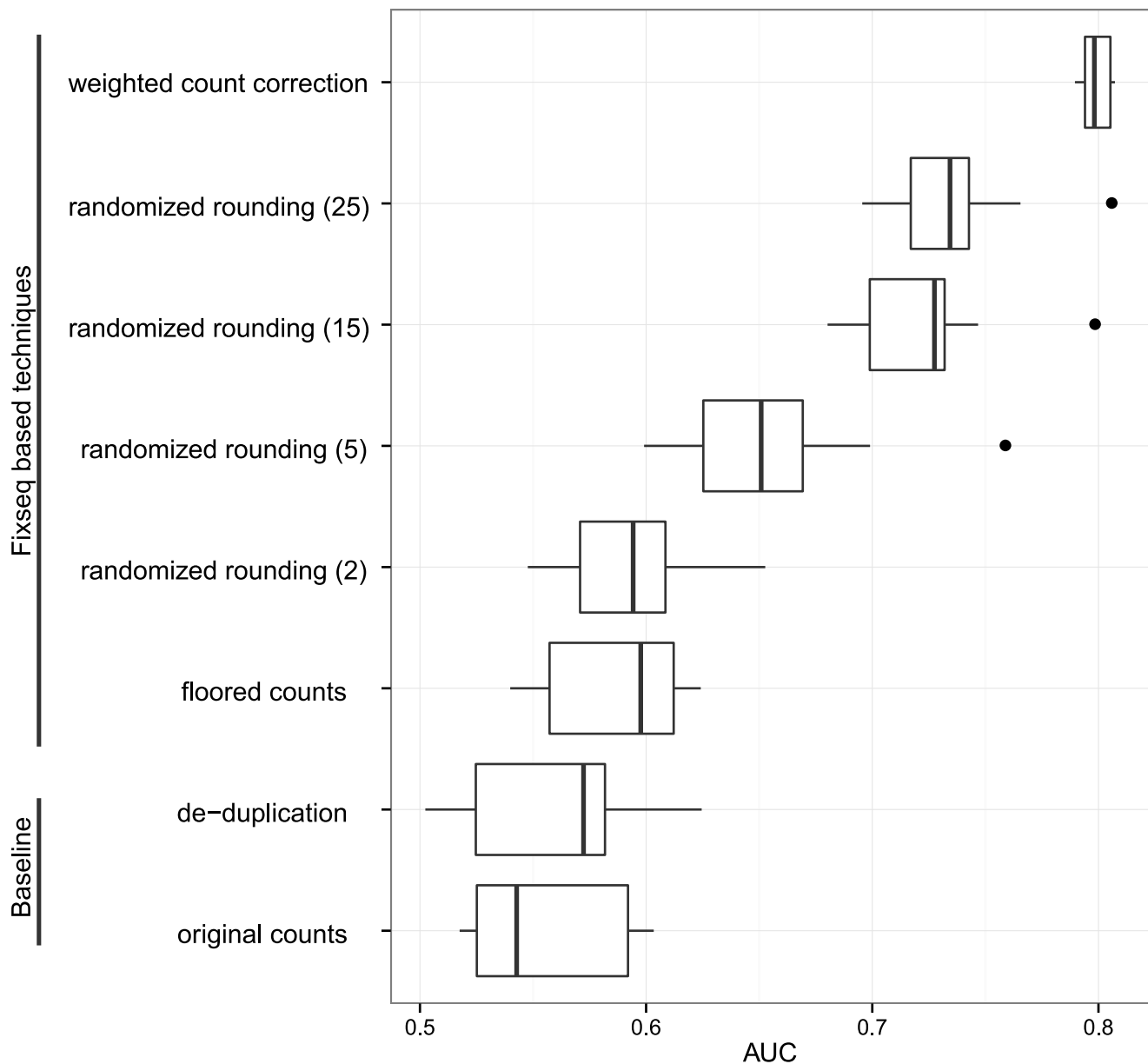
While some algorithms, such as CENTIPEDE [31] for DNase-seq binding, can take weighted (fractional) read counts, many existing algorithms will only accept integral counts. We therefore develop two rounding schemes that can improve performance while providing integral counts.

The straightforward count flooring schemes, where  $G(x, \hat{\eta}) \rightarrow \lfloor G(x, \hat{\eta}) \rfloor$  can be thought of as generalizations of de-duplication. In a typical DNase-seq experiment with 100 million reads, we find that flooring results in bases with 5 counts or less being de-duplicated, and those with 6 or more being reduced to two reads per base. While in the low-count cases, flooring is nearly identical to de-duplication, as sequencing depth increases, we expect our floored predecessor to begin strongly outperforming de-duplication.

We also propose a more sophisticated randomized rounding scheme, where we take  $G(x, \hat{\eta})$  and let  $I(p)$  be a Bernoulli random variable with probability  $p$ , then the randomized round scheme generates simulated datasets whose counts round either up or down by the proximity of the adjusted count to its neighboring integers:

$$g' = \lfloor G(x, \hat{\eta}) \rfloor + I(G(x, \hat{\eta}) - \lfloor G(x, \hat{\eta}) \rfloor). \quad (\%)$$

We compared these schemes on DNase data, where the unsupervised classifier, CENTIPEDE, was capable of accepting weighted counts, allowing us to compare various rounding schemes to the direct weighting scheme using the same comparison method as our DNase-seq results. The results in Figure 7 show that floored counts provides a statistically significant, but similar, performance to de-duplication and randomized rounding strictly improves upon both schemes. Rounding is relatively dependent on the number of randomly-sampled replicates, with around thirty samples needed to achieve its peak performance.



**Figure 7. Performance comparison of rounding schemes on unsupervised DNase binding prediction.** All rounding schemes outperform baseline methods (bottom left) but only randomized rounding approaches performance of the weighted counts (top right). doi:10.1371/journal.pcbi.1003494.g007

The peak performance achieved by weighted counts is not achievable by any rounding scheme, but we find randomized rounding comes relatively close.

#### Availability

FIXSEQ is freely available for download at <http://cgs.csail.mit.edu/fixseq>.

#### Supporting Information

**Figure S1 Examples of latent  $\lambda$  distributions and mapping function.** Panel (a) shows the latent distribution of log-lambdas for various assays, while panel (b) plots the mapping function for various assays. (EPS)

**Figure S2 Comparison to a covariate-based correction method.** A comparison of rank correlation between replicate experiments is plotted for FIXSEQ, BEADS, and the two methods run in series. Measurements within each boxplot are computed via bootstrapping. (EPS)

**Table S1 Comparison to a specialized ChIP-seq event caller.** Correlation in q-value across replicates is shown for a set of hESC CTCF ChIP-seq experiments, with varying count preprocessing schemes. (PDF)

**Table S2 Analyzed ChIP-seq experiments.** Accession numbers and details for ChIP-seq experiments. (PDF)

**Table S3 Analyzed RNA-seq experiments.** Accession numbers and details for RNA-seq experiments. (PDF)

**Table S4 Analyzed DNase-seq experiments.** Accession numbers and details for DNase-seq experiments. (PDF)

**Text S1 Supplementary methods.** Supplementary results and a description of data sources and processing. (PDF)

## References

- Wold B, Myers R, et al. (2008) Sequence census methods for functional genomics. *Nature Methods* 5: 19–21.
- Park P (2009) Chip-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* 10: 669–680.
- Wang Z, Gerstein M, Snyder M (2009) Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10: 57–63.
- Ozsolak F, Milos PM (2010) Rna sequencing: advances, challenges and opportunities. *Nature Reviews Genetics* 12: 87–98.
- Boyle A, Davis S, Shulha H, Meltzer P, Margulies E, et al. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132: 311–322.
- Neph S, Vierstra J, Stergachis A, Reynolds A, Haugen E, et al. (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489: 83–90.
- Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biology* 11: R106.
- Spyrou C, Stark R, Lynch A, Tavare S (2009) Bayespeak: Bayesian analysis of chip-seq data. *BMC Bioinformatics* 10: 299.
- Ji H, Jiang H, Ma W, Johnson D, Myers R, et al. (2008) An integrated software system for analyzing chip-chip and chip-seq data. *Nature Biotechnology* 26: 1293–1300.
- Guo Y, Mahony S, Gifford DK (2012) High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS computational biology* 8: e1002638.
- Pepke S, Wold B, Mortazavi A (2009) Computation for chip-seq and rna-seq studies. *Nature methods* 6: S22–S32.
- Jones DC, Ruzzo WL, Peng X, Katze MG (2012) A new approach to bias correction in rna-seq. *Bioinformatics* 28: 921–928.
- Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM (2011) A powerful and exible statistical framework for testing hypotheses of allele-specific gene expression from rna-seq data. *Genome research* 21: 1728–1737.
- Srivastava S, Chen L (2010) A two-parameter generalized poisson model to improve the analysis of RNA-seq data. *Nucleic acids research* 38: e170.
- Li J, Jiang H, Wong WH (2010) Modeling non-uniformity in short-read rates in rna-seq data. *Genome Biology* 11: R50.
- Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L (2011) Improving rna-seq expression estimates by correcting for fragment bias. *Genome Biology* 12: R22.
- Hansen KD, Brenner SE, Dudoit S (2010) Biases in illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research* 38: e131–e131.
- Bullard J, Purdom E, Hansen K, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mma-seq experiments. *BMC bioinformatics* 11: 94.
- Cheung MS, Down TA, Latorre I, Ahringer J (2011) Systematic bias in high-throughput sequencing data and its correction by beads. *Nucleic acids research* 39: e103–e103.
- Yaffe E, Tanay A (2011) Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics* 43: 1059–1065.
- Li W, Jiang T (2012) Transcriptome assembly and isoform expression level estimation from biased rna-seq reads. *Bioinformatics* 28: 2914–2921.
- Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, et al. (2012) Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature Methods* 9: 999–1003.
- Benjamini Y, Speed TP (2012) Summarizing and correcting the gc content bias in high-throughput sequencing. *Nucleic acids research* 40: e72–e72.
- Thygesen HH, Zwinderman AH (2006) Modeling sage data with a truncated gamma-poisson model. *BMC bioinformatics* 7: 157.
- Bliss CI, Fisher RA (1953) Fitting the negative binomial distribution to biological data. *Biometrics* 9: 176–200.
- Bulmer M (1974) On fitting the poisson lognormal distribution to species-abundance data. *Biometrics* : 101–110.
- Chen Y, Negre N, Li Q, Mieczkowska JO, Slattery M, et al. (2012) Systematic evaluation of factors influencing ChIP-seq fidelity. *Nature Methods* 9: 609–614.
- Wilbanks EG, Facciotti MT (2010) Evaluation of algorithm performance in ChIP-Seq peak detection. *PLoS ONE* 5: e11471. doi: 10.1371/journal.pone.0011471.
- Zhang Y, Liu T, Meyer C, Eeckhoutte J, Johnson D, et al. (2008) Model-based analysis of chip-seq (macs). *Genome Biology* 9: R137.
- Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology* 26: 1351–1359.
- Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, et al. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research* 21: 447–455.
- Khatun J (2012) An integrated encyclopedia of dna elements in the human genome. *Nature* 489: 57–74. doi: 10.1038/nature11247.
- Rozowsky J, Euskirchen G, Auerbach R, Zhang Z, Gibson T, et al. (2009) Peakseq enables systematic scoring of chip-seq experiments relative to controls. *Nature Biotechnology* 27: 66–75.
- Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research* 22: 1813–1831.
- Cule M (2010) Theoretical properties of the log-concave maximum likelihood estimator of a mul-tidimensional density. *Electronic Journal of Statistics* 4: 254–270.
- Chang GT, Walther G (2007) Clustering with mixtures of log-concave distributions. *Computational Statistics & Data Analysis* 51: 6242–6251.
- Walther G (2009) Inference and modeling with log-concave distributions. *Statistical Science* 24: 319–327.
- Nesterov Y (2003) *Introductory Lectures on Convex Optimization: A Basic Course*. Springer.
- Hildreth C (1954) Point estimates of ordinates of concave functions. *Journal of the American Statistical Association* 49: 598.

**Dataset S1 Software.** Code, documentation, and test data implementing the FIXSEQ method. (ZIP)

## Author Contributions

Conceived and designed the experiments: TBH MDE DKG. Performed the experiments: TBH MDE. Analyzed the data: TBH MDE. Contributed reagents/materials/analysis tools: TBH MDE. Wrote the paper: TBH MDE DKG.