doi 10.18699/bgrs2024-12.3-20

Toward improved calling of allele-specific events in high-throughput sequencing data with a mixture of compound Dirichlet negative multinomial distributions

Meshcheryakov G.*, Kulakovskiy I.V.
Institute of Protein Research, RAS, Puschino, Russia
** georgy@vega.protres.ru*

Key words: allelic imbalance; multinomial distribution; compound models; mixture model

Motivation and Aim: Heterozygous sites in gene regulatory regions may have allele-dependent chromatin states or protein affinity. High-throughput sequencing efficiently reveals nucleotide variants at homologous chromosomes of the human genome. The allelic imbalance of read counts supporting alternating alleles at individual sites reveals functional consequences of nucleotide substitutions in regulatory regions. Thus, it is possible to study the allele-specific gene regulation using the data from assays focused on genome regulatory regions, such as ChIP-Seq for transcription factor binding, or DNase- and ATAC-Seq for chromatin accessibility. Quantitative evaluation of the allelic imbalance requires a proper statistical framework accounting for noise, technical variability, and biases in the underlying read count distributions. Here we present a novel approach utilizing joint bivariate Dirichlet negative multinomial compound distributions to improve accuracy and better reflect the underlying nature of high-throughput sequencing experiments.
Methods and Algorithms: Statistical scoring of the allelic imbalance faces three challenges. First, the relative allelic copy numbers, the background allelic dosage, usually remains unknown. This challenge is solved by BABACHI [1] which performs unsupervised estimation of the relative background allelic dosage (BAD) directly from variant calls. At known BAD, the expected frequency of allelic counts is $f_1 = \frac{BAD}{BAD+1}$ and $f_2 = 1 - f_1 = \frac{1}{BAD+1}$. Then, it is reasonable to model counts distribution as a mixture of 2 distributions with frequencies f_1 and f_2.
The second challenge is to estimate whether the observed imbalance at individual variants significantly deviates from the expected imbalance arising from BAD. The straightforward approach is to use the binomial or the negative binomial distribution and estimate the success probability and the location shift parameters. Yet, those approaches fail to model the observed distributions of allelic read counts for particular types of experimental data, such as mapping accessible chromatin with ATAC-Seq or profiling transcriptional activity of regulatory elements with CAGE. In those cases, the (negative-) binomial model fails to approximate the empirical distribution adequately. The problem can be alleviated by assuming that frequencies follow the beta distribution, resulting in the beta-binomial or the beta negative binomial models [2].
The third challenge is to ensure robustness under model misspecifications. In practice, model assumptions might deviate from the true data-generating distribution. For instance, the joint datasets could include multiple sequencing batches or different cell types, each with its unique distribution parameters. Even if it is possible to specify the

model-generating process and all sources of noise, the analytical formulae for such models become intractable.

In our previous work [2, 3], we did not specify the correct model generating distribution and resorted to using the pseudolikelihood framework. There, we assumed conditional distributions $x_{|y}$ and $y_{|x}$ to be known (where (x, y) is a pair of reference and alternative allele read counts) and to follow the negative binomial or beta negative binomial distribution. Yet, estimating the model parameters from conditional distributions introduces biases into resulting estimates. This pitfall affects other common approaches, including those based on the beta-binomial distribution.

To see this, let the distribution of x given coverage $n = x + y$ follow the beta-binomial law with the success rate p and the concentration parameter κ. The log-likelihood function for the joint distribution of (x, y) is:

$$l(x, n|p, \kappa) = l(x|n, p, \kappa) + l(n|p, \kappa).$$

Maximizing it with respect to p and κ is equivalent to maximizing the conditional log-likelihood $l(x|n, p, \kappa)$ alone only if the latter marginal term doesn't depend on any of the 2 parameters. In this work we show, that that marginal likelihood, in fact, depends on the parameter κ, making its estimates obtained via maximizing the conditional log-likelihood unreliable.

We generalize the existing models under the vein of the joint bivariate distribution, based on the Dirichlet negative multinomial distribution [3] (DNM):

$$(x, y) \sim \omega\, DNM(r_0, \alpha_0, \alpha_1, \alpha_2) + (1 - \omega) DNM(r_0, DNM(r_0, \alpha_0, \alpha_2, \alpha_1)), \quad (1)$$

where α_i are parameterized as $\alpha_0 = \frac{1-p_0}{p_0}\kappa, \alpha_1 = p\,\kappa, \alpha_2 = (1-p)\kappa,$ with p_0, r_0 controlling the coverage distribution, p being a parameter controlling the mapping bias ($p = f_1 = \frac{BAD}{BAD+1}$ if there is no bias), and κ being an overdispersion parameter. When dealing with data of $BAD = 1$, the mixture can be omitted by setting $w = 1$. The model can further accommodate a greater degree of overdispersion by assuming that $r_0 \sim ZTBinom(\mu, \sigma)$, where $ZTBinom$ is a zero-truncated Binomial distribution parameterized in terms of mean value μ and variance σ. It can be shown that it is possible to marginalize over r_0 and obtain an analytical form of the compound distribution [3, 4]. The parameters $p_0, p_1, p_2, r_0, \kappa$ and μ_0, σ (when dealing with the compound model) are estimated using the Local Maximum Likelihood (LML) method. Using LML instead of ML allows for capturing coverage-specific effects in the data. JAX automatic differentiation library to obtain analytical gradients of the likelihood objective function and scipy's SLSQP solver to obtain local minima.

After estimation of the model parameters, we compute p-values to identify ASEs using conditional distributions:

$$y_x \sim \omega BetaNB(r, p\kappa, (1-p)\kappa) + (1 - \omega) BetaNB(r, (1-p)\kappa, \mu\kappa), \quad (2)$$

where $r = r_0 + x$. Note that as $\kappa \to \infty$, $BetaNB$ converges to NB. Also note that alternatively, we could've conditioned on the coverage $n = x + y$ instead, obtaining a mixture of beta-binomial models instead.

If we use the compound model, then instead of $BetaNB$ we use $MCBNB$ (Marginalized Compound Beta Negative Binomial distribution) with the probability mass function

$$f(y|x, p, \kappa, r_0) \propto {}_3F_2(\kappa p + 1, y + 1, r - 1;\ 2, (1-p)\kappa - r;\ 1 - p), \quad (3)$$

where $r = r_0 + x$ and ${}_3F_2$ is a generalized hypergeometric function [5].

Conclusion: The proposed model is freely available as a part of the updated **MIXALIME** software at the Python PyPi package repository:

pip install mixalime

The detailed information on the software and documentation is available at github.com/autosome-ru/mixalime and at the website mixalime.georgy.top. The proposed method will facilitate improved reliability for identification of allele-specific regulatory events in a wide repertoire of sequencing assays, including CAGE-Seq and ATAC-Seq.

Funding: This study was supported by RSF 20-74-10075.

References

1. Abramov S. et al. Landscape of allele-specific transcription factor binding in the human genome. *Nat Commun*. 2021;12:2751. doi 10.1038/s41467-021-23007-0
2. Meshcheryakov G. et al. Beta negative binomial mixture model facilitates identification of allele-specific gene regulation in high-throughput sequencing data. In: Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2022). Novosibirsk, 2022;1107. doi 10.18699/SBB-2022-664
3. Meshcheryakov G. et al. MIXALIME: MIXture models for ALlelic IMbalance Estimation in high-throughput sequencing data. *arXiv*. 2023. doi 10.48550/arXiv.2306.08287
4. Meshcheryakov G. et al. Improved identification of allele-specific gene regulation in high-throughput sequencing data with the marginalized compound negative binomial distribution. Systems Biology and Bioinformatics (SBB-2023). Novosibirsk, 2023;44. doi 10.18699/SBB-2023-44
5. Abramowitz M., Stegun I. (Eds.). Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. NBS, 1964