DOI 10.18699/SBB-2022-664

Beta negative binomial mixture model facilitates identification of allele-specific gene regulation in high-throughput sequencing data

Meshcheryakov G.[1]*, Boytsov A.[2], Abramov S.[2], Kulakovskiy I.V.[1, 2]

[1] *Institute of Protein Research, RAS, Puschino, Russia*

[2] *Vavilov Institute of General Genetics, RAS, Moscow, Russia*

** georgy@vega.protres.ru*

Key words: allelic imbalance, beta negative binomial, mixture model

Motivation and Aim: High-throughput sequencing allows identifying allele-specific regulatory sites, where alternative alleles of homologous chromosomes have different chromatin state (e.g. allele-specific chromatin accessibility, ASC) or different transcription factor binding affinity (allele-specific binding, ASB). The ASC and ASB events are highlighted by single-nucleotide substitutions in read alignments and provide valuable information on functional consequences of single-nucleotide variants in non-coding regions. The key approach to identify AS events is to assess the allelic imbalance of read counts supporting alternative alleles in ChIP-Seq (for ASBs) or DNase-Seq/ATAC-Seq (ASC) read alignments.

There are two challenges in statistical scoring of the observed allelic imbalance at candidate sites. First, the relative allelic copy numbers (allelic dosage) of the particular genomic segments is generally unknown. This challenge is solved by BABACHI[1], that performs unsupervised estimation of the relative background allelic dosage (BAD) directly from variant calls. Knowing *BAD*, the expected frequency of two allelic variants is $f_1 = \frac{BAD}{BAD+1}$ and $f_2 = 1 - f_1 = \frac{1}{BAD+1}$.

The second challenge is to estimate whether the observed imbalance significantly deviates from the BAD estimates. In Abramov et al. [1], we used the negative binomial mixture model with f_1 and f_2. This approach works reasonably well for ASBs, but fails to model the observed distributions of allelic read counts for ASCs. ATAC-Seq and DNase-Seq data yield higher variance and the negative binomial model fails to adequately approximate the empirical distribution. Here, we propose an improvement by assumingfrequency to be a beta random variable.

Methods and Algorithms: Similarly to the previous work, we model the number of reads x_c supporting one of the alleles given a certain fixed number of reads c at the other allele. We assume x_c to be distributed as a mixture of two negative-binomial random variables with success rates f_1 and $f_2 = 1 - f_1$ respectively. In this study, we also let f_1 to be a beta random variable centered at $\mu = f_1 = \frac{BAD}{BAD+1}$. The latter can be conveniently achieved with reparametrization of $Beta(\alpha, \beta)$: $\alpha = \mu\kappa$, $\beta = (1-\mu)\kappa$ with mean $\mu$and concentration κ. Thus,

$x_c \sim \omega NB(r, p) + (1-\omega)NB(r, 1-p), p \sim Beta(\mu\kappa, (1-\mu)\kappa)$ (1)

where r, ω are active parameters and "+" is a mixing operator of two distributions, whereas $\omega, (1-\omega)$ are mixture weights (note that in the special case of $BAD = 1$ it

holds that $p = 1 - p$ and ω parameter is redundant). The proposed generative model at Equation 1 is simplified by integrating out the frequency parameter p via marginalization. The result is known as beta negative binomial distribution, and the final model attains the form of

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

Here r, κ, ω are obtained as maximum likelihood (ML) estimates. To fit the model parameters to the observed distribution of read counts x_c at various fixed c, we used JAX automatic differentiation library to obtain analytical gradients of ML objective function and scipy's SLSQP solver to obtain local minima.

Conclusion: The proposed method is integrated into the MixALime python software (https://github.com/autosome-ru/MixALime) and will facilitate reliable identification of allele-specific regulatory events in a wide repertoire of sequencing assays, including DNase-Seq and ATAC-Seq.

References

1. Abramov S. et al. Landscape of allele-specific transcription factor binding in the human genome. *Nat Commun.* 2021;12:2751.