Identification of RNA editing events in RNA-Seq data

We aligned the paired-end RNA-Seq reads to human (hg19) or mouse (mm10) genome using STAR aligner (Dobin et al., 2013). We then followed the GATK (Van der Auwera et al., 2013) workflow for calling variants in RNA-Seq (https://software.broadinstitute.org/gatk/documentation/article?id=3891) to identify all the mutations in each RNA-Seq library. We then restricted to the mutations within annotated mRNA transcripts, as well as restricting to A-to-G mutations in transcripts encoded by the forward strand and T-to-C mutations in transcripts encoded by the reverse strand. We also filtered mutations found in the dbSNP database since they are most likely DNA-level mutations. After that, we combined the filtered sets of RNA editing events from all RNA-Seq libraries of the same experiment, and counted the number of reads containing reference (A/T) and alternative (G/C) alleles from each library at each site.

Statistical test for difference in editing frequencies between conditions

We used beta-binomial distribution to model the RNA editing frequencies, which has also previously been applied to modeling allele frequencies in RNA-Seq reads (Parker et al., 2016; Yablonovitch et al., 2017). The beta-binomial distribution is the binomial distribution where the probability of success at each trial is not fixed, but instead drawn from the beta distribution. The probability functions of binomial distribution and beta distribution are:

$$P(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\pi(p|\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}$$

So the probability density function of the compound distribution, beta-binomial distribution, can be represented as:

$$f(k|n,\alpha,\beta) = \int_0^1 P(k|n,p)\,\pi(p|\alpha,\beta)\,dp$$

$$= \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\,dp$$

$$= \frac{\binom{n}{k}}{B(\alpha,\beta)} \int_0^1 p^{k+\alpha-1}(1-p)^{n+\beta-k-1} dp = \binom{n}{k} \frac{B(k+\alpha, n+\beta-k)}{B(\alpha,\beta)}$$

For convenience, it is common to reparametrize it as:

$$\mu = \frac{\alpha}{\alpha+\beta}$$

$$\rho = \frac{1}{\alpha+\beta+1}$$

so that the expectation and variance of the beta-binomial distribution are:

$$E(k|n,\mu,\rho) = n\mu$$

$$Var(k|n,\mu,\rho) = n\mu(1-\mu)[1+(n-1)\rho]$$

In this form, μ corresponds to the estimate of p, and ρ corresponds to the extent of over-dispersion. Both μ and ρ values are between 0 and 1.

When we use beta-binomial distribution to model the RNA editing events in RNA-Seq, n corresponds to the total number of reads overlapping with an RNA editing site and k corresponds to the number of reads with A-to-G mutations. In this scenario, beta-binomial distribution is a better model for read counts than binomial distribution since it takes the variability in mutation frequencies between biological samples into account. Under the null hypothesis, all samples have equal RNA editing level and the editing frequencies are drawn from the same beta distribution $\pi(\mu_0, \rho)$. Under the alternative hypothesis, the samples expressing dADAR-hMSI2 fusion protein have different RNA editing frequency than the control samples, and the frequencies come from two different beta distributions $\pi(\mu_1, \rho)$ and $\pi(\mu_2, \rho)$. For the read counts at each RNA editing site, we maximized the likelihood for both null and alternative hypotheses, and then computed the p-value using likelihood ratio test. The p-values from all sites were adjusted using Benjamin-Hochberg correction. The statistical computation was performed using R packages *VGAM* and *bbmle*.

References

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In Current Protocols in Bioinformatics, (Hoboken, NJ, USA: John Wiley & Sons, Inc.), p. 11.10.1-11.10.33.

Dobin, A., Davis, C. a, Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Parker, C.C., Gopalakrishnan, S., Carbonetto, P., Gonzales, N.M., Leung, E., Park, Y.J., Aryee, E., Davis, J., Blizard, D.A., Ackert-Bicknell, C.L., et al. (2016). Genome-wide association study of behavioral, physiological and gene expression traits in outbred CFW mice. Nat. Genet. *48*, 919–926.

Yablonovitch, A.L., Fu, J., Li, K., Mahato, S., Kang, L., Rashkovetsky, E., Korol, A.B., Tang, H., Michalak, P., Zelhof, A.C., et al. (2017). Regulation of gene expression and RNA editing in Drosophila adapting to divergent microclimates. Nat. Commun. *8*, 1570.