Detecting A-to-I RNA editing signatures in long RNA-Seq reads.

Lo Giudice C(1)† , Luigi Mansi, Pesole G(1,2), Picardi E(1,2)

(1) Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, Consiglio Nazionale delle Ricerche, Bari, Italy.
(2) Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari A. Moro, Bari, Italy.

† Email: ernesto.picardi@uniba.it

Motivation
RNA editing is an important co/post-transcriptional process able to increase transcriptome and proteome diversity. The most common type of RNA editing in humans is mediated by ADAR enzymes, which convert adenosine into inosine within double-stranded RNAs (dsRNAs). Alterations in RNA editing have been linked to various human disorders and this has greatly increased the interest toward high-throughput methods to detect A-to-I events at genomic scale. Compared to the second-generation sequencing, single-molecule real-time sequencing allows the production of long reads that are useful for a variety of genomic and transcriptomic applications. Long reads, however, are very noisy reaching an error rate up to 30% and the detection of A-to-I editing events in these reads appears quite challenging. In order to discover RNA editing changes in noisy long reads, we started developing a bioinformatics workflow to investigate the presence of suitable A-to-I signals in circular consensus PacBio reads from a dataset obtained by RNA-Seq analysis of U937 leukemia cell line.

Methods
Circular consensus PacBio reads from the U937 leukemia cell line were aligned to human genome (hg19 assembly) using two independent mappers, gmap [1] and minimap2 [2]. Variant calling on aligned reads was performed by mean of Reditools [3]. Repeat masked annotations were used to identify variants in Alu repeated elements. Known SNPs were removed using dbSNP annotations stored in UCSC. Custom python scripts were developed to calculate the distribution of observed base changes both globally and in Alu regions. Clusters of individual base changes were also taken into account to look at RNA editing signals.

Results
By comparing gmap and minimap2 local alignments to the genome we found, as expected for noisy reads such as PacBio, an overrepresentation of AC/TG mismatches more prominent in BAM files generated by minimap2. The high rate of AC/TG mismatches was probably due to different behaviors of the two aligners as well as sequencing errors. AG/TC mismatches, representing potential RNA editing candidates, appeared the second most abundant change. To look at RNA editing signals, we tried to mitigate the rate of AC/TG mismatches realigning long reads by Blat and filtering out reads mapping with an identity score lower that 80% and alignment

extent lower than 70%. After this procedure, we observed a significant increase in the number of AG/TC changes and a strong reduction of noisy AC/TG mismatches. The number of AG/TC changes was higher in Alu regions and the majority of such mismatches was organized in clusters, as expected in case of A-to-I RNA editing. A benchmark comparison with editing and hyper-editing events detected by using Illumina RNA-seq reads obtained from the same U937 sample has been carried out to assess prediction accuracy.