Informatic Tools for Single-Nucleotide Analysis of RNA-Seq

Zhuangli Liang, Qing Zhang, Jaegyoon Ahn, Xinshu Grace Xiao
Department of Integrative Biology and Physiology, UCLA

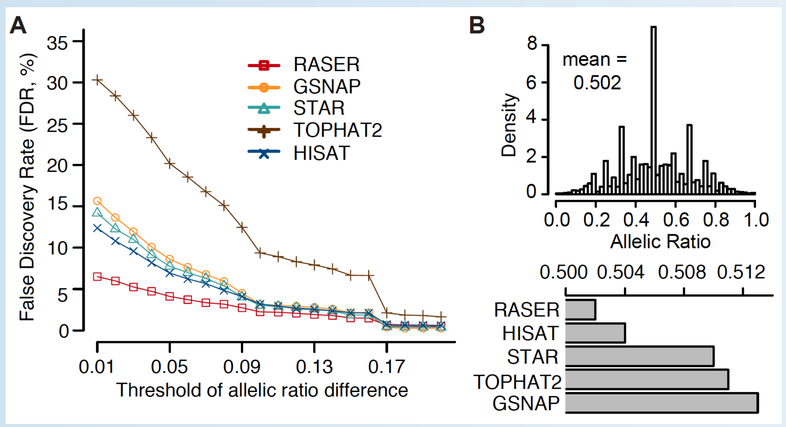
INTRODUCTION

This project aims to develop informatic tools for single-nucleotide analysis of cancer RNA sequencing (RNA-Seq) data. RNA-Seq is becoming an essential tool in both basic and clinical cancer research. As a result, numerous research groups are generating their own RNA-Seq data sets. In addition, large consortium efforts, such as the TCGA project, are producing an extraordinary amount of RNA-Seq data that are invaluable resources to the research community. The wide adoption of RNA-Seq calls for effective and user-friendly informatic tools that can extract information of important biological relevance. A major advantage of RNA-Seq is its capacity to provide information at the single-nucleotide level. Tools that harness this information are relatively scarce. As a result, single-nucleotide analysis is not yet a widely adopted procedure in RNA-Seq informatics. This type of analysis can potentially reveal important biological insights. However, a number of challenges exist in the identification, quantification and functional prediction of single nucleotide variants (SNVs) in RNA. We have developed a suite of methodologies to address these challenges. We will further improve these methods and build user-friendly informatic tools and web portals to identify and analyze SNVs in cancer RNA-Seq data. These tools will facilitate a broad spectrum of SNV analysis, ranging from raw read mapping to functional prediction of SNVs in affecting alternative splicing or RNA stability. With no additional experimental cost, information of SNVs is readily extractable in all RNA-Seq data sets. A full exploration of this information could provide novel insights and maximize the scientific value of the still costly RNA-Seq data.

METHODOLOGY

READ ALIGNMENT

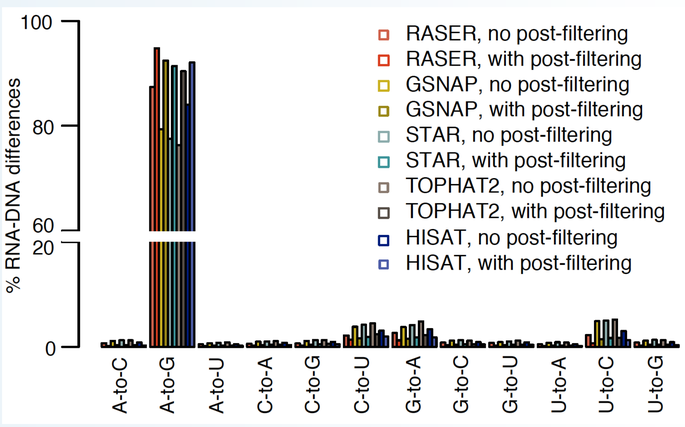
RASER: Read Aligner for SNPs and Editing sites of RNA



Aligner performance

Left: estimation of allelic ratios of SNPs.

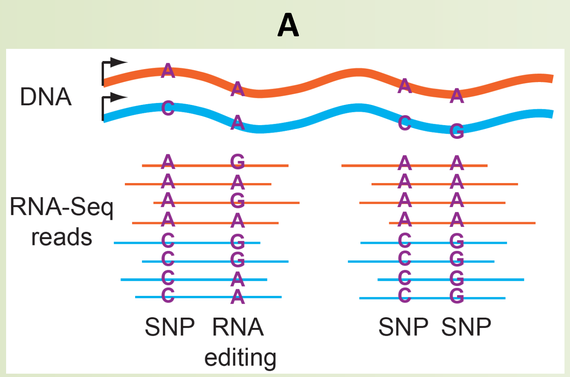
Right: identification of RNA editing sites.

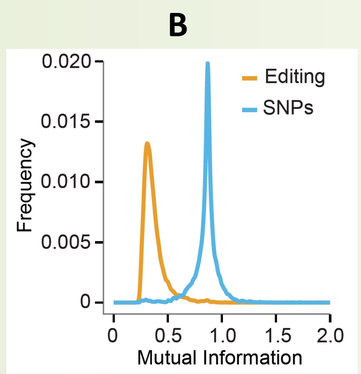


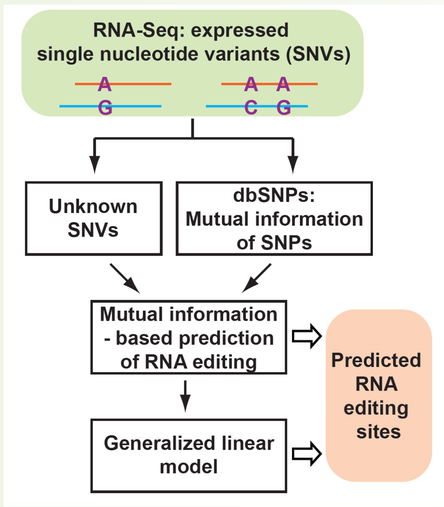
SNV CALLING, ARTIFACT FILTERING

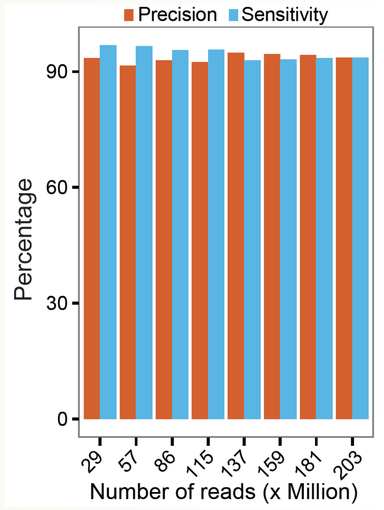
GIREMI: segregation of genetic variants and RNA editing sites

(A) Basic principle of GIREMI, utilizing linkage disequilibrium of SNVs in RNA-Seq reads. **(B)** Mutual information of pairs of editing sites or pairs of SNPs in RNA-Seq reads (GM12878 cells). **(C)** Flow chart of GIREMI. **(D)** Performance of GIREMI in identifying RNA editing sites given different read coverages.





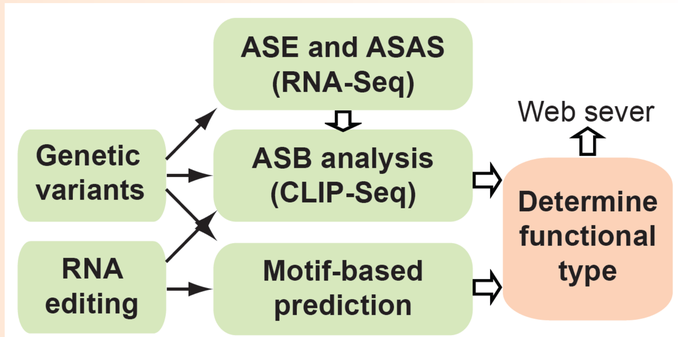




SEGRAGATION OF GENETIC VARIANTS & RNA EDITING SITES

FUNCTIONAL PREDICTION OF SNVS

ASE: allele-specific expression.
ASAS: allele-specific alternative splicing.
ASB: allele-specific binding.



References & Acknowledgements

RASER: Ahn & Xiao, *Bioinformatics*, 2015.
GIREMI: Zhang & Xiao, *Nature Methods*, 2015.
SNV calling and filtering: Bahn et al, *Gen Res*, 2012
Functional prediction: Li et al, *Nucleotide Acids Res*, 2012. Hsiao et al, *Gen Res*, 2016.

Funding: NCI, NHGRI