ABSTRACT

Chromatin Immunoprecipitation followed by massively-parallel sequencing (ChIP-Seq) is an indispensable tool in understanding the dynamics and evolution of regulatory circuitry of prokaryotes and eukaryotes. ChIP-Seq studies aim to decipher gene regulatory mechanisms by mapping genome-wide transcription factor binding sites (TFBSs). Aligning millions of short sequences (reads) to the reference genome is the first fundamental step in the analysis pipeline.

Whereas not all reads align to their reference genomes, the source of unaligned reads has not been systematically explored. We describe a computational approach to establish the source of unaligned reads from several major model organisms (*Homo sapiens, Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana*) and *Zea mays* in which ChIP-Seq efforts are underway as a first step in establishing the architecture of gene regulatory networks (GRN). The analysis of raw reads obtained from NCBI Short Read Archive (SRA) revealed a significant level of contamination in ChIP-Seq unaligned reads with sequences of bacterial and metazoan origin, irrespective of the source of chromatin used for the ChIP-Seq studies. In agreement with other sequencing studies, our results indicated that human sequences are the main source of contamination. Unexpectedly, however, was the observation that some of the selected unaligned reads data sets contained significant numbers of legitimate reads that have mappable properties, but were missed out by researchers in the alignment process. This highlights a need to improve the currently utilized alignment algorithms.