Spring 5-12-2016

Evaluation of public genome references for RNA-Seq data analysis in Chinese Hamster ovary cells

Huong Le
Amgen

Chun Chen
Amgen

Chetan Goudar
Amgen

Evaluation of Public Genome References for RNA-Seq Data Analysis in Chinese Hamster Ovary Cells

Huong Le, Chun Chen, and Chetan T. Goudar

Drug Substance Technologies, Process Development
Amgen, Inc., One Amgen Center Drive, Thousand Oaks, CA 91320, USA

Recent advances in next-generation sequencing technologies have led to the emergence of RNA-Seq as the preferred transcriptomic tool in the biopharmaceutical industry. However, an important challenge with deploying RNA-Seq to characterize CHO cells is the absence of a common genomic reference for this species. In most published CHO cell transcriptomic studies, RNA-Seq reads are assembled into *de novo* genomic references which were subsequently used for mapping of the constituent reads. Such an approach makes it difficult to compare results across studies due to the incomplete and non-universal nature of those assemblies. To address this challenge, we evaluated two publicly available genomes and their derived transcriptomes at the NCBI Reference Sequence Database (RefSeq), including CHO-K1 genome (GCF_000223135.1) and Chinese hamster genome (GCF_000419365.1). When applied for a diverse set of 60 RNA-Seq samples, each with approximately 40 million reads, both genomes showed significantly better mapping rates (~75%) compared to their derived transcriptomes (53-63%). Despite similar annotation, gene content, and KEGG pathway coverage level in both genomes, only 69% of overlapping genes between these two genomes had consistent quantification (i.e., read count) across 60 RNA-Seq samples. Examining genes with quantification discrepancies in a genome browser provides an effective avenue to identify targets for potential genome improvement. Two metrics were proposed to assess the genome-specific difference (consistency) and the sample-specific difference (stringency). Genes with low stringency can introduce biases during the identification of differentially expressed genes and pathways. Given that both genomes for CHO cells are still incomplete, we propose utilization of both in RNA-Seq data analyses until a universal reference with refined genome assembly and gene model annotation is generated.