Comparative Analysis of PacBio Libraries Reveals Non-Stochastic Biases in Sites of DNA Nicking

Ethan Alexander García Baker[1,2] *, Olivia Mendivil Ramos[1], Senem Mavruk Eskipehlivan[1], Sara Goodwin[1], Eric Antoniou[1] and W. Richard McCombie[1]

[1]Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724
[2]Department of Neuroscience, Dietrich School of Arts and Sciences, University of Pittsburgh, Pittsburgh, PA 15213
*Presenting Author

Long read sequencing (LRS) has become an invaluable tool for more complete genomic assembly and identification of genomic features that otherwise go undetected in short-read assemblies. Failed sequencing runs represent a significant expense to genomics facilities, resulting not only in wasted resources, but also irreversible loss of sample. This speaks to the need to identify the causes of failure, quantify and rectify them. Though there are several potential causes of poor performance of PacBio sequencing, we hypothesize that nicks disrupt the circular structure (SMRTbell) employed in PacBio sequencing, resulting in incomplete sequence capture and failure to generate a circular consensus sequence.

We quantified the contribution of DNA nicking in SMRTbell failure via fluorescent labeling of 3' ends. The results suggests poorly performing PacBio libraries do indeed have higher levels of nicking when compared to both undamaged genomic DNA and high-performing PacBio libraries (p-value=0.01). Therefore, DNA nicking plays an important role in failure of PacBio sequencing. Further experiments quantifying these nicks are underway in a parallel project from the CSHL group.

Furthermore, we generated *in silico* data via Monte-Carlo simulations which accommodate the inherent characteristics of PacBio sequencing. These simulations provide a framework to which actual PacBio data can be compared to identify nick preference for certain nucleotide permutations. Initial Monte-Carlo simulations were performed on sequencing data from the SKBR3 human breast cancer cell line, and suggested a non-stochastic pattern of nick occurrence over nucleotide pairs ($p<0.01$). Further *in silico* experiments were replicated in *E. coli*, *Z. mays*, and *S. cerevisiae*, revealing similar nick preference for specific nucleotide permutations across species. Our findings suggest a novel approach of predicting success of a library in long-read sequencing and identifying genomic features most susceptible to damage resulting bias during sequencing.