Whole genome sequencing with increased insert length eliminates over 25% of errors in challenging medically relevant genes

Isaku Tanida, PhD. (isaku.tanida@elembio.com), Element Biosciences, **Semyon Kruglyak**, Element Biosciences, **Kelly Blease**, Element Biosciences, **Bryan R. Lajoie**, Element Biosciences, **Sophie Billings**, Element Biosciences, **Isaku Tanida**, Element Biosciences, **Junhua Zhao**, Element Biosciences, **Shawn Levy**, Element Biosciences

There are several ongoing and past efforts to identify and address regions in the human genome that are not well resolved with short read technologies or with standard reference genome builds. We performed experiments to address some of these challenges in an extensible manner where technology platforms and analysis pipelines were standardized to evaluate what gains can be made with simple upstream experimental changes.

We performed whole genome sequencing using a HG002 library prepared with a range of increasing insert lengths, up to mean aligned insert size of over 1KB. By extending the insert length, and benchmarking against the standard NIST v4.2.1 truth, the total number of errors across the genome was decreased by ~27.8% compared to shorter/standard insert length libraries (34,402 total errors reduced to 24,829). This corresponded to an increased F1 score of .9967 and .9968 for SNPs and indels, respectively. Notably, we observed a 32.1% reduction in errors (from 595 to 404 total errors) in the Medically Relevant Gene (MRG) regions.

In this work we show how long insert length libraries were created and how sequencing conditions were modified to generate high quality data. We demonstrate that the primary mechanism of improvement is the ability to accurately align to regions that were previously inaccessible with shorter inserts. Specifically, longer inserts may span repetitive regions and offer an anchor that is leveraged for accurate alignment. False negatives, which make up the largest error category in the challenging medically relevant genes, are reduced by nearly a third. The work demonstrates that some of the benefits typically associated with longer read lengths can be achieved by increasing the insert length. Unlike longer read lengths, longer inserts neither increase the cost nor the time required to complete the sequencing run.