

PacBio Data Deep Dive: A Closer Look at HiFi Sequencing. PacBio May 24, 2021.

URL: <https://www.youtube.com/watch?v=1aqnLQ-nwYk>

CS0460375

Total Audio Length: 00:1:54

Transcribe: 00:09:50 - 00:11:45

Speaker: Since the launch of our Sequel system in 2015, this data flow has focused primarily on subreads, which takes us into our next topic for the day: continuous long reads. These are essentially the subreads that we extract from the polymerase reads. For the Sequel and Sequel II systems, our data model has been to practically preserve the polymerase reads for all ZMWs by partitioning them into scraps and subreads both of which are written to disk in unaligned BAM files that comply with the standard SAM specification but lack a cigar string. The subreads were the starting point for all secondary analysis pipelines and the scraps were really hardly ever used except in rare cases of troubleshooting. While this data model did preserve all the bases detected during primary analysis, these files were hundreds of gigabytes in size requiring huge amounts of data storage capacity, and computer resources for downstream analysis. But before we talk about our latest data model innovations, let's dig a little bit deeper into the subreads BAM file. Here's a single sub-read record from a BAM file produced by one of our Sequel II systems. In the SAM flag section of the record you can see there is no cigar string, indicating that the data are unaligned. You can also see that we use a number of custom tags to store information specific to our data type. Most important are the inter-pulse duration and the pulse-width tags shown in blue and orange respectively. These two tags contain the kinetic information that is critical to our CCS calling algorithm. There's also the tags for context flag and position start and end. [00:11:45]

END