HiLive – Real-Time Mapping of Illumina Reads while Sequencing

Martin S. Lindner[1,#], Benjamin Strauch[1], Jakob Schulze[1], Piotr W. Dabrowski[1,2], Andreas Nitsche[2], Bernhard Y. Renard[1,*]

[1] - Research Group Bioinformatics (NG 4), Robert Koch Institute, Berlin, Germany

[2] - Centre for Biological Threats and Special Pathogens, Robert Koch Institute, Berlin, Germany

[#] - current affiliation: Karius Inc., Menlo Park, CA, United States of America.

Background

Next Generation Sequencing (NGS) is increasingly used in time critical setups, such as in clinical diagnostics or precision medicine. Today, the computational analysis of the massive amounts of data produced by modern devices is still a bottleneck on the way to the final interpretation of the experiment. Mapping reads to reference sequences is an essential step in many analysis pipelines. While read mapping algorithms have always been optimized for speed, they follow a sequential paradigm and only start after finishing of the sequencing run and conversion of files. The time while the sequencer is running is typically not used for data analysis.

We developed HiLive, the first general purpose read mapper that performs read mapping while the sequencer is still sequencing. HiLive makes use of the intermediate results generated by Illumina machines to perform read mapping and thereby drastically reduces crucial overall sample analysis time, e.g. in precision medicine.

Results

We present HiLive as a novel real time read mapper that is able to perform read mapping on the temporary, unfinished read data generated by Illumina sequencers. Such a strategy is facing mainly two problems: (i) Parallelism: > 1 billion reads are generated by the sequencer in parallel and need to be processed simultaneously to overcome the sequential paradigm of traditional read mappers. (ii) Incomplete information: Calculating the optimal alignment is not possible when the read is not completely sequenced. Therefore, many candidate alignments need to be stored for each read in the intermediate cycles. To address these problems, HiLive implements a k-mer based alignment strategy: the mapper continuously reads the intermediate BCL files created in each cycle of the instrument and extends initial k-mer matches by the increasingly produced data from the sequencer. We use exact and heuristic quality criteria to determine false alignments as early as possible without discarding true alignments. The overall memory footprint and required disk space is kept low by a slim implementation and data streaming.

We applied HiLive on real human transcriptome data to show that live mapping is technically possible and no compromise has to be made in comparison to traditional mappers. In our experiment, we mapped the 1.7 billion NGS reads generated in one Illumina HiSeq 1500 run to the human transcriptome. On a workstation size computer (32 cores), HiLive finished read mapping 9 min 53 s after the end of the sequencing run. Conversion of the BCL files to fastq files took already 48 min, and subsequent mapping with BWA took 12 h 31 min. Comparison to BLAST alignments shows that HiLive is on par with current read mappers, such as Bowtie 2, BWA, and Yara with slight advantages in sensitivity. These findings on the real data could be reproduced in an experiment based on simulated data.

Conclusions

We could show that live mapping of Illumina reads is technically and practically possible. Our tool HiLive allows a massive reduction in total sample analysis time by starting read mapping while the sequencer is still running. Although HiLive implements a completely different alignment strategy, the quality is comparable to other state of the art mappers.

HiLive is freely available from https://sourceforge.net/projects/hilive/ .