Massively Parallel Demultiplexing of Raw Illumina Sequencing Data

BR Myers[1], SML Griffith[1], KN Hetrick[1], BD Craig[1], D Mohr[1], K Roberts[1], MZ Mawhinney[1], AB Robinson[1], DR Leary[3], JD Newcomer[2], L Watkins[1], Jr., KF Doheny[1]

1 *Center for Inherited Disease Research (CIDR), Institute of Genetic Medicine, Johns University, Baltimore, MD*
2 *Johns Hopkins Applied Physics Lab (JHU/APL), 11100 Johns Hopkins Rd, Laurel, MD 20723*
3 *New York Genome Center, New York, NY.*

Background

The Center for Inherited Disease Research (CIDR) provides high quality genotyping and sequencing services and statistical genetics consultation to investigators working to discover genes that contribute to disease. To that end, raw base call files (BCL) generated by various Illumina technologies (HiSeq, MiSeq, etc.) must be converted to the FASTQ format for further data analysis. While Illumina's demultiplexing tools are typically used for this process, the CIDR software team has designed and implemented a distributed BCL demultiplexing tool.

Technology

The CIDR BCL conversion tool, written in Java, is integrated with CIDRSeqSuite, a multifaceted in-house software application designed to implement and run sequencing analysis pipelines in a distributed fashion across a computational cluster with relative ease for users. This allows us to highly parallelize each tile conversion and to integrate BCL demultiplexing in line with various analysis pipelines. Stand alone FASTQ generation is also an option for users.

Our previous method of generating FASTQ files was to demultiplex QSEQ files produced by Illumina's Offline Base Caller (OLB) which took many hours to complete whereas the current tool takes 5 to 20 minutes for a comparable dataset.

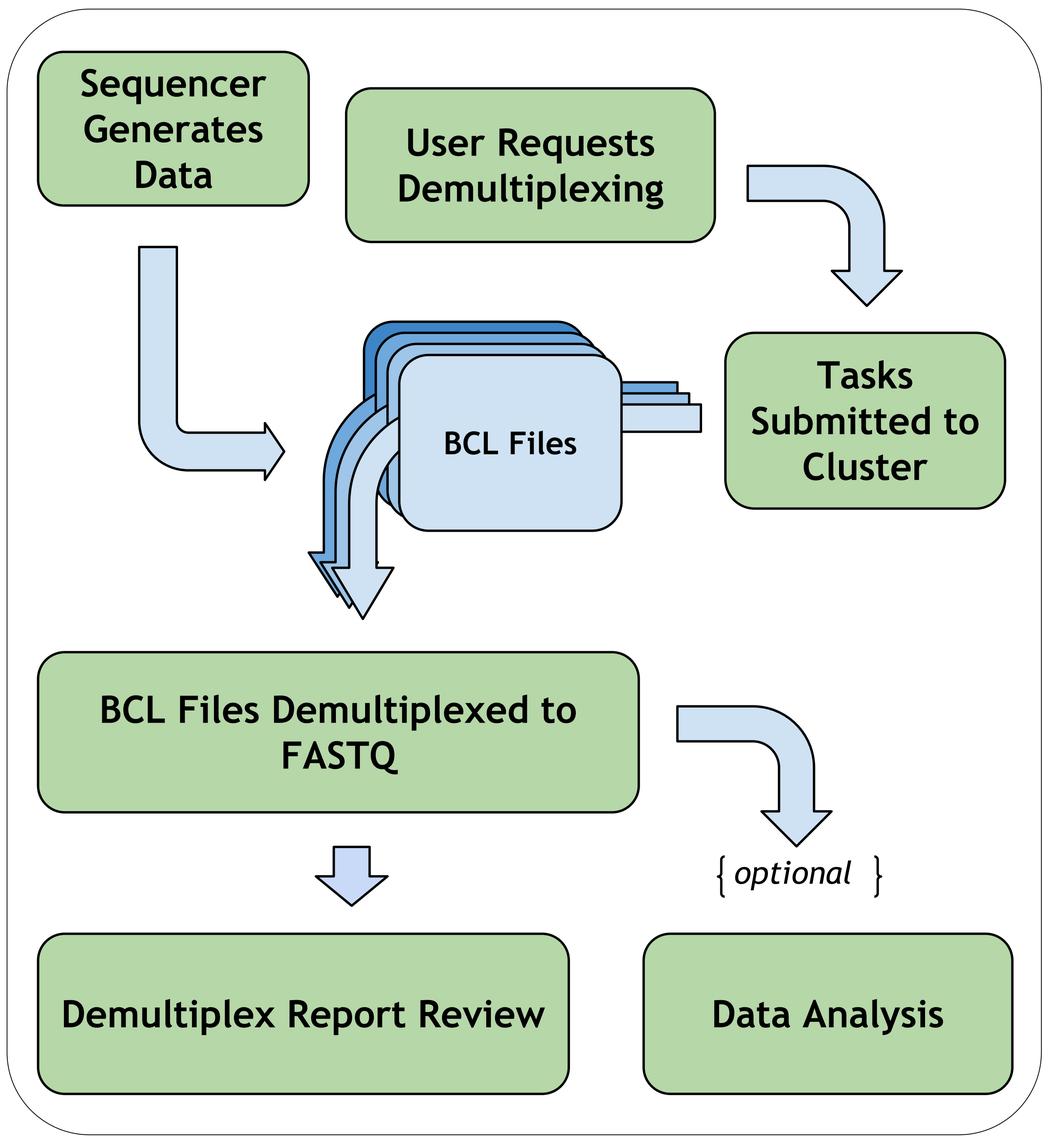
Additional Features

Autodetection: The conversion tool works for single- and paired- end runs, supports both currently used location formats (.clocs and .locs) and automatically detects which conversion to perform at a level abstracted away from the user.

Demultiplexing Reports: Detailed reports are written at the flowcell level on the sample index distribution discovered while demultiplexing. Users are notified by email when each report is written and can then quickly determine incorrect or underrepresented indices present in a given run. Email notification of a demux report denotes completion of demultiplexing for the corresponding flowcell.

Bad BCL Data Handling: If BCL files are discovered to be truncated or corrupt during demultiplexing, their associated tiles will be omitted from FASTQ generation. Users will be notified of this omission in real time via email.

BCL Conversion Workflow



Future

Work includes enabling dual-index demultiplexing and making the CIDR BCL conversion tool open source along with CIDRSeqSuite.