|  |  |
| --- | --- |
| **{ Title }** | |
| **Responsible Department: CLS-NGS** | **Effective Date: { Publication Date }** |
| **Policy Basis for Procedure**  This SOP describes the standard processes for data management and archiving | |
| **Applicability**  This applies to the bioinformatics and lab personnel of the NGS Lab. | |
| **Description of Standard Procedure**  **PROCEDURE**  **Sequencing data generation and transfer to BioHPC**  The sequencing instruments split the flow cell into smaller segments called tiles in order to process the data efficiently, in parallel. The number of tiles varies depending on the flow cell type and instrument used. Sequencing data is generated from imaged clusters on a per tile process. Data from each tile is processed on RAM to reduce disk read/write without saving the actual image to the sequencer’s hard drive. The main output of this process is a basecall file (bcl for MiSeqDx, or bcl.bgzf for NextSeq or bcl.gz for HiSeq4000), and a file containing intensity and position data for all clusters (locs per MiSeqDx tile, for the NextSeq they contain data for all tiles per lane, for the HiSeq4000 this file contains data for all tiles in all lanes). Other output is generated to track pass filter status, quality scores, PhiX error rate, and other sequencing quality metrics.  Tile data is required for the conversion of the base called on the sequencer into fastq files. We use illumina’s bcl2fastq version 2.17.1.14, installed in BioHPC’s nucleus cluster. The output of the sequencers is saved in the run folders located at BioHPC (nucleus.biohpc.swmed.edu); the full path to the run folder is /project/PHG/PHG\_Illumina/BioCenter/<run folder>. The BioHPC folder is mounted on each sequencing instrument as a network folder accessible to the sequencing control software used to configure the run settings. Files are transferred automatically using illumina software installed on the sequencers. The software monitors the data transfer progress and has built-in functions to account for temporary interruptions of the network traffic. If the interruption is longer than 15 minutes on the MiSeqDx then the files are kept in a queue folder located on the sequencer (D:\Illumina\MiSeqTemp\<run folder>\Queued). Similarly if the interruption is longer than 60 minutes on the NextSeq or HiSeq4000 the files are kept in a queue folder: D:\Illumina\NextSeq Control Software Temp\<runfolder> or O:\Illumina\HiSeqTemp\<run folder>. From this queue folder all pending files need to be transferred to BioHPC by manually setting up a sftp session. There is no copy of the tile data retained in the sequencers if the file transfer is successful. The only files kept on the sequencer pertain to sequencing quality metrics and other files used for assessing run performance.  Retention period for clinical run folders located in BioHPC is 2 years unless there is a need to keep them for re-analysis or process development, or if the Medical Director approves an extension request. The fastq files generated by bcl2fastq use the data in the sequencing run folders, these fastq files are saved in /project/PHG/PHG\_Clinical/illumina/<run folder>. These fastq files are the input to all of our analysis pipelines.  **Analysis pipeline file locations and data backup generation**  Demultiplexing is processed in  /project/PHG/PHG\_Clinical/illumina/<run folder> and fastq files are symlinked to /project/PHG/PHG\_Clinical/cases/<project name>/<sample name>  Fastq files are archived to Information Resources(IR) where they will be backed up into Azure cloud storage. The command to archive fastq files is:  tar cf /project/PHG/IR\_Archival/YYMM/<run folder>.fastq.tar  /project/PHG/PHG\_Clinical/illumina /<run folder>  gzip /project/PHG/IR\_Archival/YYMM/<run folder>.fastq.tar  tar cf /project/PHG/IR\_Archival/YYMM/<run folder>runfolder.tar  /project/PHG/PHG\_Illumina/BioCenter /<run folder>  gzip /project/PHG/IR\_Archival/YYMM/<run folder>runfolder.tar  Bioinformatics analysis is initially processed in /project/PHG/PHG\_Clinical/processing/<run folder>  Upon successful completion of analysis all relevant files are rsynced to a folder of completed cases. This rsync is accomplished with the following command:  rsync –avz $case /project/PHG/PHG\_Clinical/cases  Where $case is the project name of the samples being analyzed. The folder “/project/PHG/PHG\_Clinical/cases” is a symlinked folder. The official folder where files are stored is “/archive/PHG/PHG\_Clinical/cases.” All rsync commands are generated automatically for each case with the script init\_workflow.pl and is written in a bash script named **run\_<run folder>.sh**.  Files retained include, but are not limited to:  sequence QC metrics files, original and deduped bam files, the vcf calls per tool and vcfs for the combined calls, copy number files for tumor dna, and translocation files for rnseq.  Further processing of files using the script unify\_case.sh is performed in the completed cases folder to produce final vcfs, translocations, tumor mutation burden and xml files. These files will also be stored together with the original processed files and fastq files.  A final archival of data will be archived to a folder accessible to IR.  Final archive to IR will be accomplished by the following command:  tar cf /project/PHG/IR\_Archival/YYMM/$case.tar  /project/PHG/PHG\_Clinical/cases/$case  gzip /project/PHG/IR\_Archival/YYMM/$case.tar  Logs of data archiving will be written with standard output data of sbatch script. If data is interrupted at any time the commands above can be ran manually. Exact commands can be found in **run\_<run folder>.sh**  **Information Resources data backup**   * Go to this source file to access the files:   + [\\lamella.biohpc.swmed.edu\phg\_backup\](file:///\\lamella.biohpc.swmed.edu\phg_backup\)        * Select the designated file for the month you wish to view   + Below is the screen shot of the files for August 2018- this shows the file name, date modified, file type, and the size of the file. You can order the files by any criteria shown by clicking on the top label.   C:\Users\Erika\Desktop\ir_pic1.png   * The files are manually transferred to **Microsoft Azure Storage Explorer** by dragging and dropping them from the source file under Local and Attached -> Storage Accounts -> swazrstrseq (External) -> Blob Containers -> archive201808   + Under “Blob Containers” there are separate containers for each designated month. This one is designated for August 2018.      * Once the files are uploaded, the storage file will look like this:   C:\Users\Erika\Desktop\ir_pic2.png   * The files should all match up with the original destination files from the source file. Depending on how large the file is, the times to transfer each file may vary.   For going back to previous months and updating the files, there are a lot of variations between the file sizes from what had been previously stored in the archive and the updated files. The previous file has to be deleted from Azure Storage Explorer and the larger copy from the file share has to be retransferred to Azure. For example, for “201711”,  C:\Users\Erika\Desktop\ir_pic3.png  The highlighted file was last modified in December 2017 and is only 64.0 GB.  C:\Users\Erika\Desktop\ir_pic4.png  The updated file is 225,245,540 KB. Therefore, the file on Azure has to be deleted in order to retransfer this new larger file. Along with this change, the new files added to “201711” are also dragged and dropped on to “archive201711” in Azure. | |
|  | |
| **Applicable Forms**   |  |  | | --- | --- | | Document ID# | Title of Document | | 1 | Data Management and Archiving | |  |  | |  |  |   **Related Documents**   |  |  | | --- | --- | | Document ID# | Title of Document | |  |  | |  |  | |  |  | | |
| **References** | |
| **Review, Revision, and Approval History**   |  |  |  |  | | --- | --- | --- | --- | | **History of Document** | | | | | Version | Date | Description of Change | Authored/Revised by | | 1 | 2018/10 | * New Document, Policy Tech upload, | Carlos Arana, Erika Villa, Brandi Cantarel, Andrew Quinn | |  |  |  |  | | |
|  | |
| **Contact for Further Information** | |

**Title:** { Title } **Version:** { Version Number } **Status:** { Status }

**Writer(s)**: { Writers > All Users > Name & Title }

**Approver(s)**: { Approvers > All Users > Name & Title } **Approved Date**: { Date Approved }

**Effective Date**: { Publication Date }

**Next Review Date**: { Next Periodic Review Date }