|  |  |  |
| --- | --- | --- |
| **Title** | Halo\_sv2\_v1.0 Bioinformatics Validation | |
| **Summary Statement** | **The following documentation contains the protocol to run the Archer V 5.0.6 Analysis Software for Translocations at the Center of Personalized Diagnostics** | |
| **Prepared by** | **Ashkan Bigdeli** | Date: 10/19/2017 |
| **Approved by** |  | Date: |
| **Approved by** |  | Date: |
|  |  | Date: |
|  |  | Date: |
|  |  |  |
| **Retired by** |  | Date: |

Table of Contents

**Introduction**1

**Software and Reference Documentation**1

**Virtual Machine Configuration**2

**Analysis Protocol**3

**Introduction**

This document outlines the processes to convert raw Next-Generation Sequencing data to interpretation ready reports using the ArcherDX Version 5.0.6 analysis pipeline. Individual applications, algorithms, software, scripts, references and annotation mechanisms are clearly stated. All CPD developed code including the connection of discrete applications and utilized parameters are provided, and in the case of third party software and references, locations for retrieval of these resources are provided. All utilized bed files and scripts are provided with this document and expanded details can be found in the appendices. For validation documentation including acceptation and rejection criteria as well as the variant spectrum and laboratory processes please see (ask Robyn for name for doc).

**Software and Reference Documentation**

All utilized software, including third party and in-house developed, are provided below. Links for documentation and download, as well as the utilized version are provided. CPD developed code can be retrieved from our Github page (<https://github.com/faryabiLab/cpdUpenn>). For access to code and utilized servers contact the appropriate CPD personnel. For detailed information and additional downloads should one be needed contact an ArcherDX representative or refer to the software manual located at UPHSISILON\PathCPD\Informatics\Software\Archer.

|  |  |  |
| --- | --- | --- |
| **Third Party Programs** | **Version** | **Retrieval** |
| Archer Analysis Virtual Machine | 5.06.6 | CPD Isilon:/Software/ArcherDX |
| bcl2fastq | 1.8.2 | https://support.illumina.com/ |
| **CPD Developed Code** | **Version** | **Retrieval** |
| demultiplex\_archer.sh | 1.1 | https://github.com/faryabiLab/cpdUpenn/tree/archer |
| archer\_api\_parse.py | 1.1 | https://github.com/faryabiLab/cpdUpenn/tree/archer |
| **Associated Genomic Files** | **Type** | **Retrieval** |
| CKP0050-v2.1-Full\_Panel | bed | https://github.com/faryabiLab/cpdUpenn/tree/archer |
| CKP0050-v2.1-Full\_Panel | gtf | https://github.com/faryabiLab/cpdUpenn/tree/archer |
| **IT Domains** | **Purpose** | **Access** |
| UPHSISILON-PTP.uphs.pennhealth.prv | Storage | IT Provided |
| 170.212.141.107 | Demultiplexing | IT/CPD Provided |
| 170.212.137.168 | Archer VM Host | IT/CPD Provided |
| uphappndc539.uphs.upenn.edu | FileMaker Data | IT/CPD Provided |

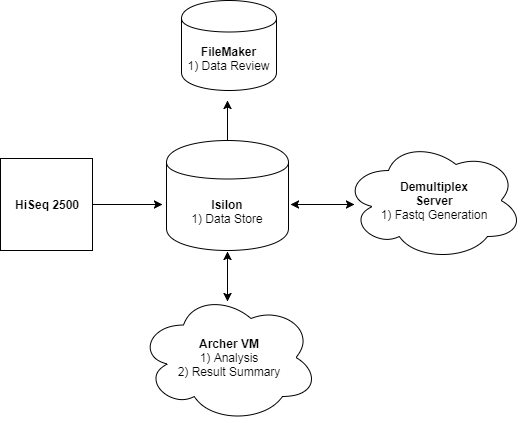
**Figure 1:** Software and dependencies required to run Archer Analysis

**Virtual Machine Configuration**

1. UPHS IS manages the IP 170.212.137.168, as well as the intended storage location UPHSISILON-PTP.uphs.pennhealth.prv, though the machine is capable of storing data locally.
2. Archer VM Analysis 5.0.6 is installed on the Cent OS 6.8 system.
   1. Installation has been modified to output analysis to UPHSISILON-PTP.uphs.pennhealth.prv/ifs/data/PathCPD/FromHPC/Archer. If attempting to remount, know that all permissions must be inherited for the web application to work correctly.
   2. Installation has been modified to process 5 samples in parallel by modifying /var/spool/torque/server\_priv/nodes~ and this process is outlined in the provided user manual.You must restart the web service to adjust the number of nodes.

**Analysis Protocol**

To provide clarity to the operational protocol a flow chart is provided that describes the flow of data from a sequencer, through an intermediary server, processing environment, and deposit into the LIMS system.

****

**Figure 2:** Data flows from the sequencer to an intermediary server named “demux”. Here demultiplexing occurs and concatenated fastq files are shuttled back to the data store. The Archer VM consumes the fastq files and produces output which is then summarized and uploaded into the FileMaker LIMS System.

1. **Demultiplexing** 
   1. Upon run completion, the name of the run folder to be processed is required. Additionally, access to server 170.212.141 is required and can be obtained by authorized CPD personnel.
   2. Login to 170.212.141.107 using your credentials. The following command can be used to begin the process of demultiplexing and fastq generation (Appendix A).
      1. ./data1/HiSeqRun/ARCHER/demultiplex\_archer.sh \*Run Directory\* \*Out Directory\*
         1. The Run directory should be a completed HiSeqRun location that is mounted to the server (e.g, Z:\illumina\SeqOut\160909\_M01651\_0105\_AT073).
         2. The directory for output is typically /data1/HiSeqRun/ARCHER/ \*Run Directory\*, but has been provided as a parameter for flexibility.
      2. This program executes 3 operations.
         1. Demultiplexes the Run producing readable fastq files.
         2. Combines fastq files by read for, a requirement of for pipeline operation.
         3. Produces a file of reads each fastq file.
2. **Archer Analysis** 
   1. Visit <http://170.212.137.168/auth/login> and login. If you do not have an account, follow the steps for account creation. Additionally, you will require an administrator add you to the CPD\_LAB group so that you may view/process common samples.
   2. In the top menu bar navigate to “Perform Analysis” tab.
      1. Parameters for analysis
         1. Appropriate Name
         2. RNA Fusion
         3. Target Region = CKP0050-v2.1-Full\_Panel
      2. Advanced Parameters
         1. Error Correction
         2. Normalization value should be equal to 1.5 million, or the smallest number of reads obtained from demultiplexing out-put file suffix \_read\_counts.txt.
      3. Progress will be viewable in the “Home” tab, here you can see recently run samples as well as running jobs. For previous results, view “Past Results”.
3. **Data Summarization and Upload**
   1. To view run completion navigate to <http://170.212.137.168/past_results> where you will find the job status and, under the column listed as job and [ID] listed job name. If any errors occur, they will be viewable from here.
   2. Upon completion of an archer run a job hook summarizes read statistics as well as variants for upload to the CPD FileMaker Database and these files can be found using the Archer Job ID to navigate to the folder directly to the Isilon shared drive.
      1. /PathCPD/FromHPC/Archer/[JOB\_ID]/[JOB\_ID]\_RunStatsFinal.txt
      2. /PathCPD/FromHPC/Archer/[JOB\_ID]/[JOB\_ID]\_Run\_masterVarFinal.txt
      3. NOTE: Prior to 10/19/2017 samples where uploaded by logging into the Virtual machine and running the Following command “python archer\_parse.py /PathCPD/FromHPC/Archer/JOB\_ID Seq ID to generate the above mentioned files. This file can be found in the same github repository.
   3. Navigate to the appropriate SEQ\_RUN layout in FileMaker.
      1. Click “Import Run Data”
      2. When prompted for run stats, navigate to the appropriate archer run directory and select JOB\_ID\_RunStatFinal.txt.
      3. When prompted for the variant file, locate the JOB\_ID\_Run\_masterVarFinal.txt file.
      4. Quit the upload process.
   4. Upload run statistics per the laboratory SOP.