Table of Contents

[Document Owner 3](#_Toc20118845)

[Affected Parties 3](#_Toc20118846)

[Purpose 3](#_Toc20118847)

[Scope 3](#_Toc20118848)

[Overview of Target Enrichment Design 4](#_Toc20118849)

[Target Regions 5](#_Toc20118850)

[Padded Target Regions 6](#_Toc20118851)

[Probe Regions 7](#_Toc20118852)

[Non-overlapping Probe Regions 8](#_Toc20118853)

[Probe Coverage Gaps 9](#_Toc20118854)

[GatorSeq Workflow Overview 10](#_Toc20118855)

[GSBW: Transfer Raw Reads To Network Drive 11](#_Toc20118856)

[[2] Initiating GatorSeq Analysis 12](#_Toc20118857)

[Column descriptions 12](#_Toc20118858)

[GatorSeq NGS analysis initiation script 13](#_Toc20118859)

[GSBW: Auto Transfer Raw Reads To HiPerGator 15](#_Toc20118860)

[GSBW: Auto Run GatorSeq NGS Analysis 16](#_Toc20118861)

[GatorSeq NGS data analysis work flow 16](#_Toc20118862)

[MAP to Reference: BWA-MEM 16](#_Toc20118863)

[SAM to BAM: sambamba-view 16](#_Toc20118864)

[Sort BAM file: sambamba-sort 16](#_Toc20118865)

[Merge BAM files: sambamba-merge 17](#_Toc20118866)

[Mark Duplicates: sambamba-markdup 17](#_Toc20118867)

[Variant Calling: VarDictJava 17](#_Toc20118868)

[Tranlocation Calling 18](#_Toc20118869)

[GSBW: Auto Transfer Analysis Output to Network Drive 22](#_Toc20118870)

[Data Analysis Metrics Summary 23](#_Toc20118871)

[Enrichment Summary Table 23](#_Toc20118872)

[Read Level Enrichment Table 23](#_Toc20118873)

[Base Level Enrichment 24](#_Toc20118874)

[Coverage Summary for Target Regions 24](#_Toc20118875)

[Duplicate Summary 25](#_Toc20118876)

[Small Variants Summary 25](#_Toc20118877)

[GSBW: Variant Filtering & Reporting 26](#_Toc20118878)

[Near identical duplicate targets 26](#_Toc20118879)

[GSBW: LIS Entry & EMR 29](#_Toc20118880)

[Software 30](#_Toc20118881)

[Software List 30](#_Toc20118882)

[Software and Pipeline Update 30](#_Toc20118883)

[Reference Material 31](#_Toc20118884)

[Human Genome 31](#_Toc20118885)

[Download Source 31](#_Toc20118886)

[Removing Haplotypes 31](#_Toc20118887)

[Masking *CRLF2* on Y Chromosome 31](#_Toc20118888)

[Portable, Reliable, Reproducible, and Scalable Software Workflows 33](#_Toc20118889)

[Retrospecitve Analysis for Auditing 33](#_Toc20118890)

[GatorSeq Pipeline Code 34](#_Toc20118891)

[Version Control 34](#_Toc20118892)

[Version Traceability 34](#_Toc20118893)

[Code Location 34](#_Toc20118894)

[Computational Infrastructure 34](#_Toc20118895)

[Illumina NextSeq 500 34](#_Toc20118896)

[Computational Infrastructure 34](#_Toc20118897)

[Illumina NextSeq 500 34](#_Toc20118898)

[Pathology Network Drive 34](#_Toc20118899)

[ahc-path-data19 35](#_Toc20118900)

[HiPerGator 35](#_Toc20118901)

[Qiagen Clinical Insight (QCI) 35](#_Toc20118902)

[Data Backups 35](#_Toc20118903)

[Illumina NextSeq 500 35](#_Toc20118904)

[Pathology Network Drive 35](#_Toc20118905)

[ahc-path-data19 36](#_Toc20118906)

[HiPerGator 36](#_Toc20118907)

[QCI 36](#_Toc20118908)

[References 37](#_Toc20118909)

# Document Owner

UF Molecular Pathology

# Affected Parties

All UF Molecular Pathology employees handling GatorSeq workflow after NGS data had been generated.

# Purpose

This document describes various steps involved in GatorSeq workflow after NGS data is generated. There are two types of details described in this document, front-end technical details, which describe user interfaces that user interacts and second is backend computational technical details, which are hidden from user perspective and more useful for application developers. It is important for users not only to familiarize with front-end but also with back-end details as this will help them to debug/diagnose the problem in case of workflow breakdown.

# Scope

This SOP describes all details required by UF Molecular Pathology users for adeptly operating GatorSeq data processing workflow.

# Overview of Target Enrichment Design

GatorSeq assay is used to identify genetic sequence variants in following target regions: 179 human cancer genes exons, MSI, translocations, sample markers, ancestral informative markers, gender, and PGx loci. Targeted enrichment of these target regions was done using in-solution hybridization of complementary probes which were degined across the target regions. Below is the summary statistics of the targets of interest and probe regions, and their corresponding genomic coordinate files are listed.

## Target Regions

These are the target regions of following target regions: 179 human cancer genes exons, MSI, translocations, sample markers, ancestral informative markers, gender, and PGx loci. ("Gatorseq\_179Genes\_TLL\_MSI\_SAMPLEID\_AIMS\_GENDER\_PGX\_Targets\_2018\_11\_02.bed")

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| Number of targets (exons) | 3,482 |
| Total size of targets | 785,086 bp |
| Min size of targets | 3 bp |
| Max size of targets | 9,166 bp |
| Mean size of targets | 225.47 bp |
| q1 size of targets | 103 bp |
| Median size of targets | 132 bp |
| q3 size of targets | 183 bp |

## Padded Target Regions

These are the 250 bp padded target regions of following target regions: 179 human cancer genes exons, MSI, translocations, sample markers, ancestral informative markers, gender, and PGx loci. ("Gatorseq\_179Genes\_TLL\_MSI\_SAMPLEID\_AIMS\_GENDER\_PGX\_Targets\_250bp\_padded\_2018\_11\_02.bed")

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| Padded Size | 250 bp |
| Number of targets | 2554 |
| Total size of targets | 2271459 bp |
| Min size of targets | 507 bp |
| Max size of targets | 31,661 bp |
| Mean size of targets | 889.37 bp |
| q1 size of targets | 616 bp |
| Median size of targets | 654 bp |
| q3 size of targets | 854.25 bp |

## Probe Regions

These are the probe regions of following target regions: 179 human cancer genes exons, MSI, translocations, sample markers, ancestral informative markers, gender, and PGx loci. ("Gatorseq\_179Genes\_TLL\_MSI\_SAMPLEID\_AIMS\_GENDER\_PGX\_Probes\_2018\_11\_02.bed")

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| Number of probes | 8184 |
| Total size of probes | 981,925 bp |
| Min size of probes | 112 bp |
| Max size of probes | 120 bp |
| Mean size of probes | 119.97 bp |
| q1 size of probes | 120 bp |
| Median size of probes | 120 bp |
| q3 size of probes | 120 bp |

## Non-overlapping Probe Regions

These are the probe merged regions of following target regions: 179 human cancer genes exons, MSI, translocations, sample markers, ancestral informative markers, gender, and PGx loci. ("Gatorseq\_179Genes\_TLL\_MSI\_SAMPLEID\_AIMS\_GENDER\_PGX\_Probes\_2018\_11\_02.bed")

|  |  |
| --- | --- |
| **Statistic** | **Value** |
| Number of probe regions | 3511 |
| Total size of probe regions | 967,687 bp |
| Min size of probe regions | 117 bp |
| Max size of probe regions | 7,200 bp |
| Mean size of probe regions | 275.61 bp |
| q1 size of probe regions | 120 bp |
| Median size of probe regions | 240 bp |
| q3 size of probe regions | 240 bp |

## Probe Coverage Gaps

Regions of the 179 GatorSeq target genes exons, which were not covered by probes. ("Gatorseq\_179Genes\_Targets\_Final\_not\_covered\_by\_probes.txt")

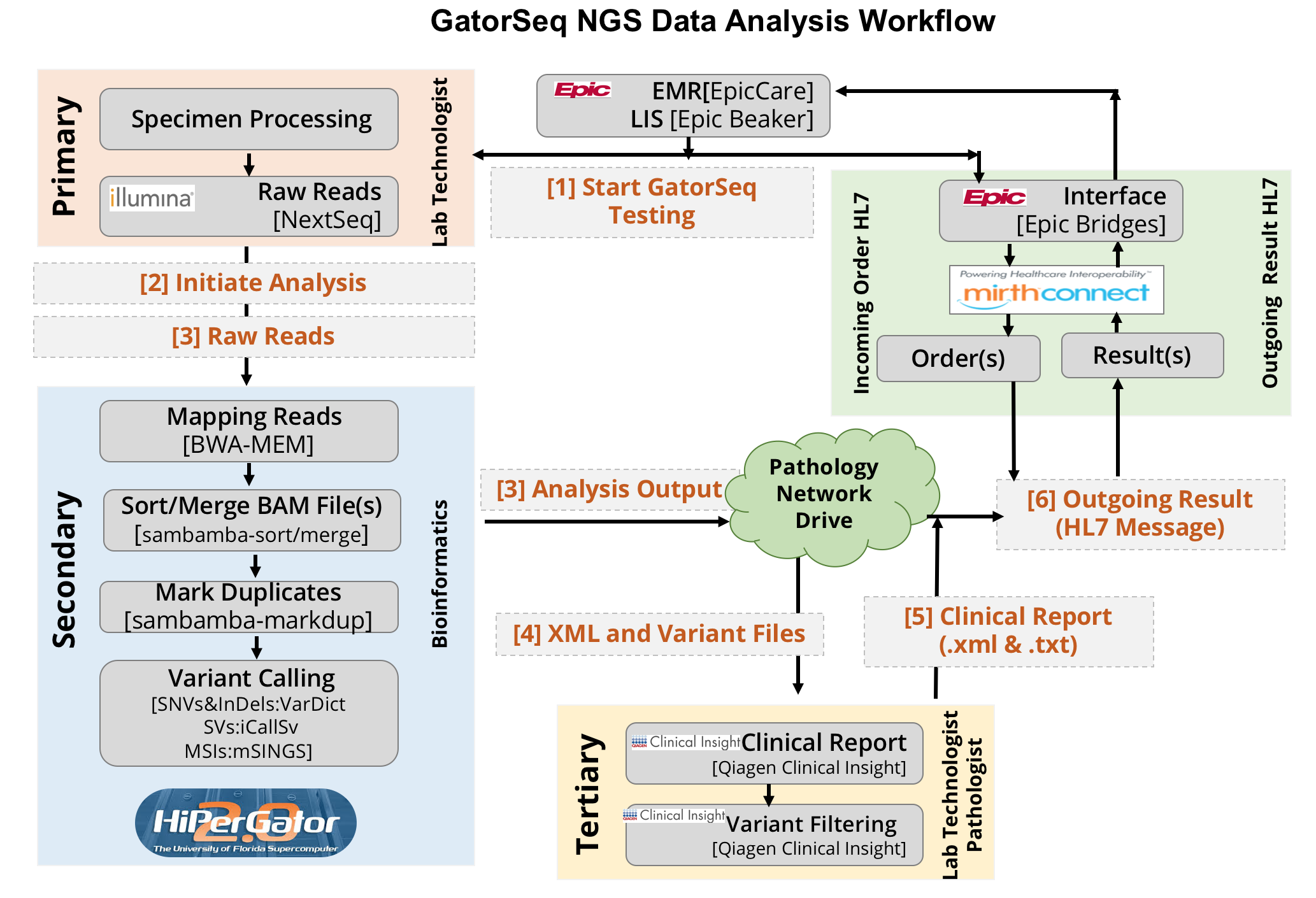
|  |  |
| --- | --- |
| **Statistic** | **Value** |
| Number of probe coverage gaps | 151 |
| Total size of probe coverage gaps | 10,079 bp |
| Min size of probe coverage gaps | 1 bp |
| Max size of probe coverage gaps | 2,302 bp |
| Mean size of probe coverage gaps | 66.75 bp |
| q1 size of probe coverage gaps | 12 bp |
| Median size of probe coverage gaps | 27 bp |
| q3 size of probe coverage gaps | 68.5 bp |
| Percent of target region of 652,264 bp | 1.54% |
| Number of probe coverage gaps of ≥500 bp \*\* | 1 (2,302 bp) |

\*\* BED coordinates of this probe coverage gap is "chr4 106159156 106161458 475\_176736\_54790(TET2)\_4a"

# GatorSeq Workflow Overview

GatorSeq data storage, analysis, and reporting is done in a very efficient and secure computational infrastructure, involving combination of commercial and in-house bioinformatics software tools and pipelines implemented on local and high-performance computing (HiPerGator) servers. Overview of this is depicted in Figure 1 and later sections of this document will further elloborate on each of the pipeline steps.

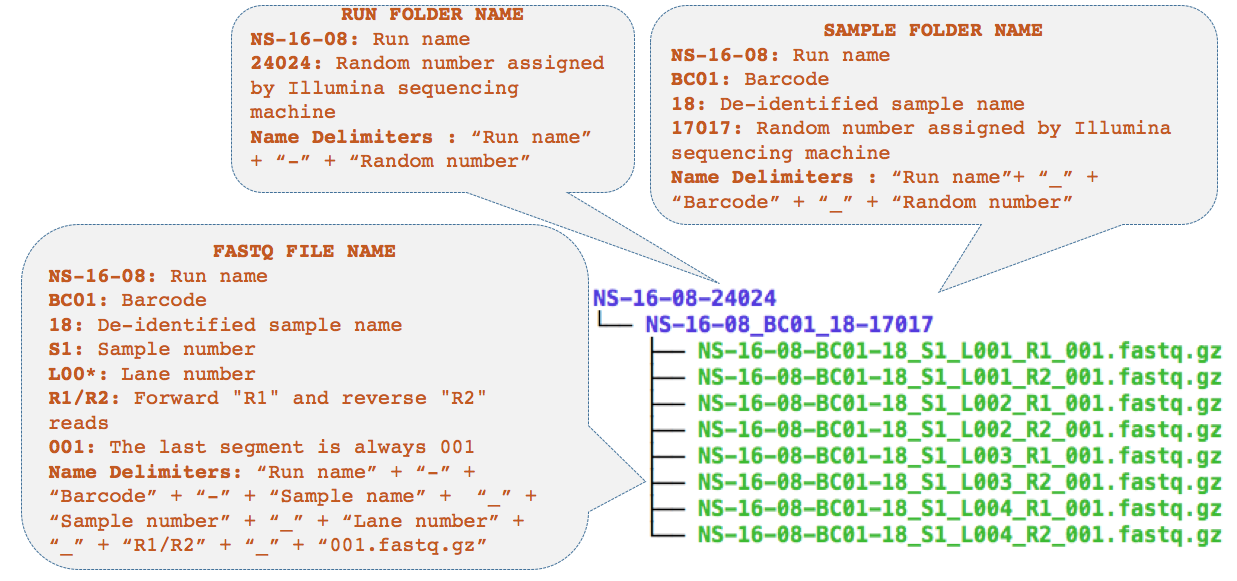
**Figure 1**: GatorSeq NGS data analysis workflow



# GSBW: Transfer Raw Reads To Network Drive

Raw sequencing data from Illumina sequencing run is demultiplexed and their corresponding fastq files are gzip compressed and are placed into their respective sample folders for each run on pathology network drive. In Figure 2, ‘NS-16-08’ Illumina run contain one sample with paired-end reads from four lanes.

**Figure 2**: Raw Fastq file naming convension and structure.



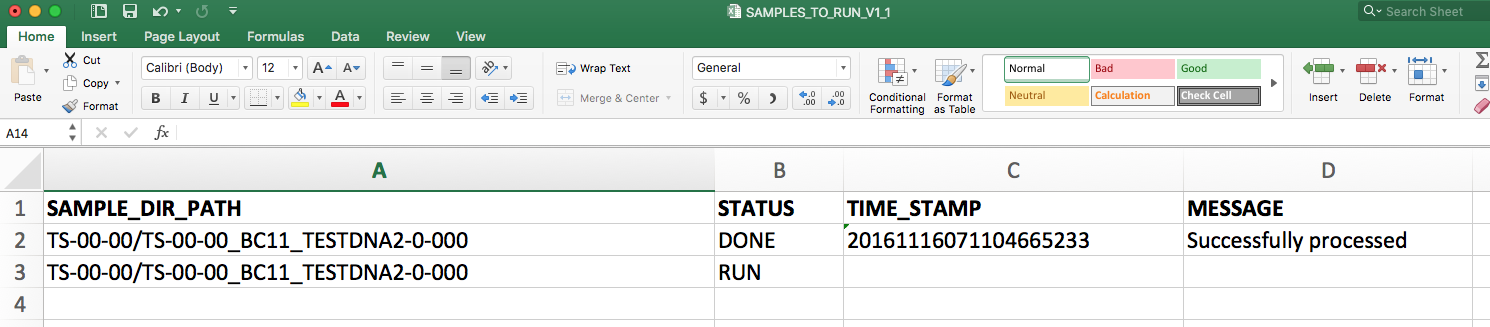
*File location on pathology network drive*:

PATH$/DRL/Molecular/NGS/NextSeq\_Fastq

# [2] Initiating GatorSeq Analysis

A lab user initiates GatorSeq NGS analysis by entering NGS sample details into an excel sheet named "GS\_Fastq\_to\_Analyze\_V1\_0.xlsx". This excel sheet contains four columns, first two columns are "SAMPLE\_DIR\_PATH" and "STATUS", for which user have to enter values, while values in the last two columns "TIME\_STAMP" and "MESSAGE" were generated by the system.

**Figure 3**: Excel sheet through which GatorSeq NGS analysis is initiated.



## Column descriptions

*Column 1 – "SAMPLE\_DIR\_PATH"*: This value contain NextSeq run name followed by forward slash followed by NextSeq sample name. Random number assigned by Illumina sequencing machine need to be excluded in run and sample names. For example, a valid value would be 'NS-16-08/NS-16-08\_BC01\_18' not 'NS-16-08-24024/NS-16-08\_BC01\_18-17017'. See above "Transfer Raw Reads To Network Drive" section for sample and run names descriptions.

*Column 2 – "STATUS"*: This value triggers the action of GatorSeq NGS pipeline. Valid values are 'RUN', 'SUBMITTED', 'FAILED', and 'DONE'. For initiatiating GatorSeq NGS pipeline, user will enter the value as 'RUN' and rest of three possible values are added by the GatorSeq NGS pipeline. See below for details on meaning and genearation of these values.

*Column 3 – "TIME\_STAMP"*: This is a twenty character value indicating the time at which GatorSeq pipeline for that particular sample is initiated. This has two primary functions, first it will allow to keep track of when a particular analysis was run. Second, it will be able to differentiate two GatorSeq outputs for a same sample ran at different time points. This scenario will arise, when user initiates a sample to run but accidentatly or on purpose reinitiates same sample again at later time point. Time stamp contains following information, 'year', 'month', 'day', 'hour', 'minute', 'second', and 'microsecond'. For example, a time stamp value of '20161104050225550505' can be interpreted as follows: 2016 (year), 11 (month), 04 (day), 05 (hour), 02 (minute), 25 (second) and (550505) microsecond.

TIP – As mentioned above time stamp can be used to differentiate two GatorSeq outputs for a same sample ran at different time points. However, it is strongly recommended to keep only one copy (i.e., delete unwanted copy) of GatorSeq analysis output for each sample to save file storage space and avoid confusion in the future. Excetptions can be made in cases where an initial report was singned off in to EPIC Beaker and later found out that the pipeline was not properly finished and contain only partial results; in this case GatorSeq analysis pipeline would be run twice, first analysis output being incorrect and second being correct. If both incorrect and correct reports are present in EPIC Beaker, then both GatorSeq analysis ouputs with different time points will be retained.

*Column 4 – "MESSAGE"*: This column contains various error, warning, and job status messages generated by GatorSeq analysis pipeline. See below for details on meaning and genearation of these values.

## GatorSeq NGS analysis initiation script

A python script ("gatorseq\_fastq\_to\_analyze.py") is scheduled to execute (in the background) every two minute on the ahc-path-data19 that will check "/path.ahc.ufl.edu/PATH$/DRL/Molecular/NGS/GatorSeq/ProdEnv/GS\_Fastq\_to\_Analyze\_V1\_0.xlsx" file on pathology network drive, which will have one of the following job statuses and their corresponding actions are listed below.

1. Status equal to 'RUN' : Python script will check whether sample fastq files are already uploaded to HiPerGator; see below "Auto Transfer Raw Reads To HiPerGator" section for details.
   1. If yes, then it will create a job trigger on HiPerGator and will mark job status to 'SUBMITTED' and appends following warning message "Currently job is running". This trigger will indicate HiPerGator to run GatorSeq NGS analysis; see below "Auto Run GatorSeq NGS Analysis" section for details.
   2. If no, then it will leave the job status as it is ('RUN') and appends following warning message "WARNING: Directory is not completely in sync with HPC folder".

DIGANOSTICS TIP – If the job status does not change from 'RUN' to 'SUBMITTED' within 10 minutes or so, then check whether pathology network drive is mounted on ahc-path-data19 under 'path-svc-mol' username. If not mounted, please mount. Wait still job status does not change please contact bioinformatics team for support.

1. Status equal to 'SUBMITTED' : Python script will check whether submitted GatorSeq NGS pipeline for the sample is successfully completed on HiPerGator and is downloaded to pathology network drive.
   1. If successfully completed and downloaded, then it will remove NGS output folder copy for the sample on HiPerGator and will mark job status to 'DONE' and appends following warning message "Successfully processed - check all files exist". Completed GatorSeq NGS pipeline’s ouput is located on HiPerGator and is downloaded to pathology network drive, which is mounted on ahc-path-data19 using a bash script ("rsync\_fastq\_V1\_1.sh") is scheduled to execute (in the background) every five minutes on the ahc-path-data19; see below "Auto Transfer Analysis Output to Network Drive" section for details.
   2. If job is still running on HiPerGator or successfully completed and not yet downloaded, then it will leave the job status as it is ('SUBMITTED') and appends following warning message "WAIT: Either job is still running on HiPerGator or the download is underway on to pathology network drive".
   3. If unsuccessfully completed and downloaded, then it will remove NGS output folder copy for the sample on HiPerGator and will mark job status to 'FAILED' and appends following warning message "ERROR: check log files for failure details".

DIGANOSTICS TIP – If the job status does not change from 'SUBMITTED ' to either 'DONE' or 'FAILED' status within a 4 hours or please contact bioinformatics team for support.

DIGANOSTICS TIP – If several jobs are submitted simultaneously and at least one or two jobs started changing statuses then it is a good sign indicating jobs are being processed on HiPerGator and in due course of time (at most within few hours) all jobs will be completed.

1. Status equal to 'FAILED' : Python script does not do anything and skips the sample.
2. Status equal to 'DONE' : Python script does not do anything and skips the sample.

**NOTE**: GatorSeq workflow pipeline only intiates when "GS\_Fastq\_to\_Analyze\_V1\_0.xlsx" is closed and not being used by anyother user. This will avoid the scenario where running a job prematurely when technician has not finished entering the required information.

**NOTE:** GatorSeq workflow pipeline will complain if lab technician enters wrong path name.

**NOTE:** When user starts the download job from Basespace, it will take sometime for it to download. However, if lab technician, even before fastq files are done downloading initiates the sample for running GatorSeq analysis, results in analysisng only partial data. To avoid this scenario, GatorSeq workflow pipeline only runs a sample when the download is fully completed.

# GSBW: Auto Transfer Raw Reads To HiPerGator

Raw reads are located on pathology network drive (as described in the previous section), which is mounted on ahc-path-data19. A bash script ("rsync\_fastq.sh") is scheduled to execute (in the background) every 15 minutes on the ahc-path-data19 that will automatically transfer raw reads from pathology network drive to HiPerGator.

# GSBW: Auto Run GatorSeq NGS Analysis

As mentioned in the above section, "gatorseq\_fastq\_to\_analyze.py" script on ahc-path-data19 will create a job triggers on HiPerGator for samples that need GatorSeq NGS analysis to be run. A bash script ("HiPerGator\_GatorSeq\_cronjob.sh") is scheduled to execute (in the background) every five minutes on the HiPerGator that will automatically checks for GatorSeq NGS analysis job triggers and executes the pipeline on HiPerGator.

## GatorSeq NGS data analysis work flow

Workflow for GatorSeq NGS data analysis is illustrated in the Figure 1. This workflow is performed separately for each of the patient samples. Data analysis steps that are listed under secondary section in the figure. First two steps (Mapping and Sorting) are performed separately on each sequencing lane corresponding to a patient sample and their outputs from the steps will be merged. This methodology helps parallelize data analysis execution and thus decreases the data analysis time.

Software suites and their versions used in the NGS workflow are as follows: are listed in "Prerequisite Software" section.

### MAP to Reference: BWA-MEM

Map the fastq reads for each illumina lane belonging to a sample using bwa.

Program Options

* -R Complete read group header line.
  + [ID:SAMPLE\_NAME + "\_" + lane\_num, LB:RUN\_NAME, SM:SAMPLE\_NAME, PL:ILLUMINA, PU:lane\_num]
* -t Number of threads.
  + [3]
* --M Mark shorter split hits as secondary
* hg19 reference genome

### SAM to BAM: sambamba-view

BWA-MEM alingments were converted to bam file format using sambamba-view.

Program Options

* --compression-level Set compression level for BAM output.
  + [0]
* --nthreads Number of threads to use.
  + [3]
* -- format Specify output format.
  + [bam]

### Sort BAM file: sambamba-sort

Sort BAM files using sambamba-sort.

Program Options

* --memory-limit Sets an upper bound for used memory.
  + [3G]
* --compression-level Set compression level for BAM output.
  + [0]
* --nthreads Number of threads to use.
  + [3]
* --tmpdir Use to output sorted chunks.

### Merge BAM files: sambamba-merge

Merge all bam files corresponding to each lane using sambamba-merge.

Program Options

* --compression-level Set compression level for BAM output.
  + [0]
* --nthreads Number of threads to use.
  + [3]

### Mark Duplicates: sambamba-markdup

Mark duplicates in the merged bam files of a sample using sambamba-markdup.

Program Options

* --compression-level Set compression level for BAM output.
  + [0]
* --nthreads Number of threads to use.
  + [3]
* --tmpdir Directory for temporary files.

### Variant Calling: VarDictJava

Variants are called on duplicate marked bam file on a sample using VarDictJava.

Program Options

* -G The reference fasta.
  + [hg19]
* -f The threshold for allele frequency.
  + [0.01]
* -N The sample name to be used directly.
  + [dynamic sample\_name]
* -b The indexed BAM file.
  + [dynamic bam\_file]
* -c The column for chromosome.
  + [1]
* -S The column for the region start.
  + [2]
* -E The column for the region end.
  + [3]
* -g The column for a gene name.
  + [4]
* -th Number of threads.
  + [3]
* -t Indicate to remove duplicated reads. Only one pair with identical start positions will be kept.
* -r The minimum # of variance reads.
  + [4]
* -B The minimum # of reads to determine strand bias.
  + [2]
* -m Minimum number of reads with mismatches more than this will be filtered and ignored. Gaps are not counted as mismatches.
  + [6]
* -P The read position filter. If the mean variants position is less that specified, it is considered false positive.
  + [5]
* -o The reads should have at least mean MapQ to be considered a valid variant.
  + [25]
* -q The phred score for a base to be considered a good call.
  + [25]

### Tranlocation Calling

GatorSeq analysis output alignment (\*.bam and \*.bam.bai) are fed into translocations detection software using iCallSV Software (version 0.0.6 - https://github.com/rhshah/iCallSV), which is the simillar pipeline used by MSKCC [1]. iCallSV requires a control file (--controlBam) for detecting translocations for which we ued HAP1 bam file was used (see below for the location). The final output contain only transversions and inversions, which are defined as translocations.

Control BAM File (Location on UF HPC):

'/ufrc/chamala/share/GatorSeq\_Share/GatorSeq\_Resources/DELLY\_CONTROLS/HAP1/NS-18-08\_BC706506\_p1\_20180414072002793846ProdEnv2.0.bam'

Tranlocation are called on duplicate marked bam file on a sample using iCallSV.

Program Options

* PYTHON
  + /apps/gcc/5.2.0/python/2.7.10/bin/python
* RHOME
  + /usr/local/anaconda/lib/R
* RLIB
  + /usr/local/anaconda/lib/R/library
* DELLY
  + /usr/local/anaconda/bin/delly
* DellyVersion
  + 0.7.6
* BCFTOOLS
  + /usr/local/anaconda/bin/bcftools
* REFFASTA - Path to hg19 Referece Fasta file
  + ucsc.hg19.fasta
* EXREGIONS - Path to file containing regions to exclude
  + human.hg19.excl.tsv
* HotspotFile - Path to file containing regions to where lenient threshold will be used
  + hotspotgenes.txt
* GenesToKeep - File containing genes to keep
  + genesToInclude.txt
* BlackListFile - Path to file containing regions to filter (Currently dummy file)
  + / blacklist.txt
* BlackListGenes - Path to file containing genes to filter (Currently dummy file)
  + blacklistgenes.txt
* SAMTOOLS
  + /usr/local/anaconda/bin/samtools
* ANNOSV
  + /usr/local/iannotatesv/1.0.9/bin/iAnnotateSV.py
* GENOMEBUILD
  + hg19
* DISTANCE
  + 3000
* CANONICALTRANSCRIPTFILE
  + /usr/local/iannotatesv/1.0.9/bin/data/canonicalInfo/canonical\_transcripts.txt
* UNIPROTFILE
  + /usr/local/iannotatesv/1.0.9/bin/data/UcscUniprotdomainInfo/hg19.uniprot.spAnnot.table.txt
* CosmicCensus
  + /usr/local/iannotatesv/1.0.9/bin/data/cosmic/cancer\_gene\_census.tsv
* CosmicFusionCounts
  + /usr/local/iannotatesv/1.0.9/bin/data/cosmic/cosmic\_fusion\_counts.tsv
* RepeatRegionAnnotation
  + /usr/local/iannotatesv/1.0.9/bin/data/repeat\_region/hg19\_repeatRegion.tsv
* DGvAnnotations
  + /usr/local/iannotatesv/1.0.9/bin/data/database\_of\_genomic\_variants/hg19\_DGv\_Annotation.tsv
* CalculateConfidenceScore
  + /usr/local/icallsv/0.0.6/iCallSV/R/Rscripts/calculateConfidenceScore.R
* GENOMEBUILD
  + hg19
* ReadLength
  + 150
* MAPQ
  + 20
* NumberOfProcessors
  + 4
* CaseAltFreqHotspot - Case Allele Fraction Hotspot
  + 0.05
* CaseCoverageHotspot - Total Case Coverage Hotspot
  + 5
* ControlAltFreqHotspot - Control Allele Fraction Hotspot
  + 0
* CaseAltFreq - Case Allele Fraction
  + 0.10
* CaseCoverage - Total Case Coverage
  + 10
* ControlAltFreq - Control Allele Fraction
  + 0
* OverallSupportingReads - Overall Supporting Read-pairs
  + 5
* OverallSupportingReadsHotspot - Overall Supporting Read-pairs Hotspot
  + 3
* OverallSupportingSplitReads - Overall Supporting splitreads
  + 0
* OverallSupportingSplitReadsHotspot - Overall Supporting splitreads Hotspot
  + 0
* CaseSupportingReads - Case Supporting Read-pairs
  + 2
* CaseSupportingSplitReads - Case Supporting splitreads
  + 0
* CaseSupportingReadsHotspot - Case Supporting Read-pairs Hotspot
  + 1
* CaseSupportingSplitReadsHotspot - Case Supporting splitreads Hotspot
  + 0
* ControlSupportingReads - Control Supporting Read-pairs
  + 3
* ControlSupportingReadsHotspot - Control Supporting Read-pairs Hotspot
  + 3
* ControlSupportingSplitReads - Control Supporting splitreads
  + 3
* ControlSupportingSplitReadsHotspot - Control Supporting splitreads Hotspot
  + 3
* LengthOfSV - Length of Structural Variant
  + 500
* OverallMapq - Overall Mapping Quality Threshold
  + 20
* OverallMapqHotspot - Overall Mapping Quality Threshold Hotspot
  + 5

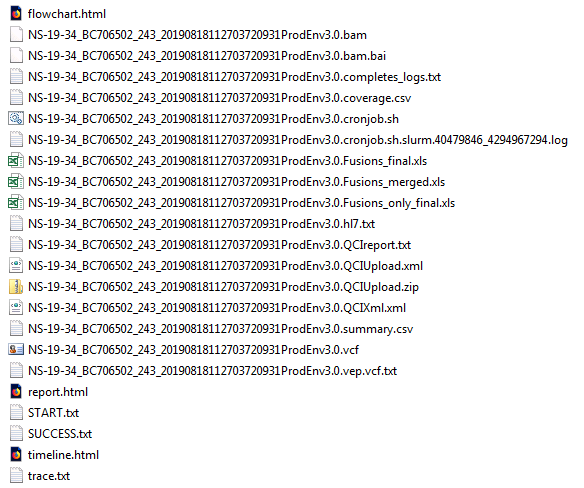
Note: "genesToInclude.txt " file contain following genes

ALK, CD74, ETV6, EWSR1, FGFR3, NTRK1, RET, ROS1, SDC4, SLC34A2

# GSBW: Auto Transfer Analysis Output to Network Drive

GatorSeq analysis output for samples are generated on HiPerGator, which need to be transferred over to pathology network drive that is mounted on ahc-path-data19. A bash script ("rsync\_output.sh") is scheduled to execute (in the background) every two minutes on the ahc-path-data19 that will automatically transfer GatorSeq analysis output from HiPerGator to pathology network drive. Figure 4 presents an example of files present in each output folder. All files in the Figure 4 are part of secondary analysis of GatorSeq bioinformatics pipeline except \*.QCI\* files which are part of the teritiary analysis depicted in Figure 1.

**Figure 4**: GatorSeq analysis output.



*Output location on Pathology Network Drive*:

/PATH$/DRL/Molecular/NGS/GenomOncology/NextSeq

## Data Analysis Metrics Summary

GatorSeq NGS analysis pipeline will produce a data analysis metrics file with ‘.metrics.txt’ file extension. Below tables describe each of the fields repored in the metrics file.

### Enrichment Summary Table

|  |  |
| --- | --- |
| **Statistic** | **Definition** |
| Total Length of Targeted Reference | The total length of the sequenced bases in the target region. |
| Padding Size | The length of sequence immediately upstream and downstream of the enrichment targets that is included for a padded target. |
| Total Length of Padded Targeted Reference | The total length of the sequenced bases in the target plus padded regions. |

### Read Level Enrichment Table

|  |  |
| --- | --- |
| **Statistic** | **Definition** |
| Total Reads [million] | The total number of reads passing filter present in the data set. |
| Total Aligned Reads [million] | The total number of reads passing filter present in the data set that aligned to the reference genome. |
| Percent Aligned Reads | The percentage of reads passing filter that aligned to the  reference genome. |
| Target Aligned Reads [million] | Number of reads that aligned to the target. |
| Padded Target Aligned Reads[million] | Number of reads that aligned to the padded target. |
| Read Enrichment | 100\*(Target aligned reads/Total aligned reads). |
| Padded Read Enrichment | 100\*(Padded target aligned reads/Total aligned reads). |
| Median Insert Size | The median insert sizes of aligned reads to the reference genome. Insert size is the length of the DNA that is "inserted" between the adapters (so adapters excluded) |
| Percent Proper Pair | The percentage of reads aligned in the correct orientation and within a defined insert size to the reference genome. |

### Base Level Enrichment

|  |  |
| --- | --- |
| **Statistic** | **Definition** |
| Total Bases[million] | The total number of bases present in the data set. |
| Total Aligned Bases[million] | The total number of bases present in the data set that aligned to the reference genome. |
| Target Aligned Bases[million] | Total aligned bases in the target region. |
| Padded Target Aligned Bases[million] | Total aligned bases in the padded target region. |
| Bases Enrichment | 100\*(Total Aligned Bases in Targeted Regions / Total Aligned Bases). |
| Padded Bases Enrichment | 100\*(Total Aligned Bases in Padded Targeted Regions / Total Aligned Bases). |

### Coverage Summary for Target Regions

|  |  |
| --- | --- |
| **Statistic** | **Definition** |
| Mean Coverage | Mean read depth in the targeted region. |
| Median Coverage | Median read depth in the targeted region. |
| Max Coverage | Max read depth in the targeted region. |
| Min Coverage | Min read depth in the targeted region. |
| q1 Coverage | First quantile of read depth in the targeted region. |
| q3 Coverage | Third quantile of read depth in the targeted region. |
| sstdev Coverage | Standard deviation of read depth in the targeted region. |
| Uniformity of Coverage [Pct > 0.2\*mean] | The percentage of targeted base positions in which the read  depth is greater than 0.2 times the mean region target coverage  depth. |
| Percent Target Coverage at 100X | Percentage targets with coverage greater than 100X. |
| Percent Target Coverage at 300X | Percentage targets with coverage greater than 300X. |
| Percent Target Coverage at 500X | Percentage targets with coverage greater than 500X. |

### Duplicate Summary

|  |  |
| --- | --- |
| **Statistic** | **Definition** |
| Percent Duplicate Reads | Percentage of paired reads that have duplicates. |

### Small Variants Summary

|  |  |
| --- | --- |
| **Statistic** | **Definition** |
| Total Passed Variants | The total number of variants present in the data set that passed the variant quality filters. |
| Total Failed Variants | The total number of variants present in the data set that failed the variant quality filters. |
| SNVs | The total number of SNVs present in the data set that passed the variant quality filters. |
| MNVs | The total number of MNVs present in the data set that passed the variant quality filters. |
| Insertions | The total number of Insertions present in the data set that passed the variant quality filters. |
| Deletions | The total number of Deletions present in the data set that passed the variant quality filters. |
| Indels | The total number of Indels present in the data set that passed the variant quality filters. |
| SNV Transitions/Transversions | Transition rate of SNVs that pass the quality filters divided by transversion rate of SNVs that pass the quality filters.  Transitions are interchanges of purines (A, G) or of pyrimidines (C, T). Transversions are interchanges between purine and pyrimidine bases (for example, A to T). |
| Total Het/Hom ratio | Number of heterozygous (Het) variants/Number of homozygous (Hom) variants. |
| SNV Het/Hom ratio | Number of Het SNVs/Number of SNVs Hom variants. |
| MNV Het/Hom ratio | Number of Het MNVs/Number of MNVs Hom variants. |
| Insertion Het/Hom ratio | Number of Het Insertions/Number of Insertions Hom variants. |
| Deletion Het/Hom ratio | Number of Het Deletions/Number of Deletions Hom variants. |
| Indel Het/Hom ratio | Number of Het Indels/Number of Indels Hom variants. |
| Insertion/Deletion ratio | Number of Insertion variants/Number of Deletion variants. |
| Indel/SNV+MNV ratio | Number of Insertions/Number of SNVs and MNVs variants. |

# GSBW: Variant Filtering & Reporting

GatorSeq analysis output variant (\*.vcf) file from previous step are automatically uploaded into QIAGEN Clinical Insight (QCI™) (Figure 1) from to pathology network drive using "QCI\_upload.py" script for doing further manual variant filtering/curation and generating molecular diagnostic report by Molecular Laboratory technologists/pathologists. One of the variant curation step involves flagging a variant if it falls in to near identical duplicate regions. Below section describes the near identical duplicate target regions. For more details on QCI fiterting and reporting, please refer to QCI SOP. Once QCI report is finalized it is automatically downloaded in to to pathology network drive as \*.xml file and converts to text format using "QCI\_download.py" script. QCI related files are listed as \*.QCI\* depicted under Figure 4.

## Near identical duplicate targets

Genomic targets for NGS sequencing in GatorSeq are enriched using oligonucleotide probe-based sequence enrichment methodology. The purpose of the oligonucleotide probe is to capture only genomic regions of interest. However, if the genomic region of interest have high sequence similarity with other parts of the genome like pseudogenes, parlogs genes, and repeats then sequence enrichment will capture off-target regions. These off-targent regions, if they are very high sequence similarity may result in false positive variant calling due to ambivalent mapping in the NGS data analysis and even mask the detection of genuine mutations residing in functional genes. *In silico* methodology described in Mandelker, Diana, et al. 2016 was implemented to find putative problematic target regions (from now on referred as near identical duplicate targets) and variants from these regions should be interpreted with caution. About 14,052 bp (0.22%) of the target region is classified as near identical duplicates.

Caution - From the 652,264 bp regions that we are targeting in the 179 gene panel, 14,052 bp (0.22%) have a near identical homolog regions in the genome which make them difficult or impossible to analyze with standard NGS approaches. These regions will be excluded from reporting or aberrations found will be reported with an appropriate disclaimer.

|  |  |
| --- | --- |
| **Probe** | **Location** |
| 472\_146043\_64783(RBM15)\_3 | chr1:110888929-110888976 |
| 469\_119176\_4853(NOTCH2)\_6a | chr1:120539619-120539955 |
| 469\_119175\_4853(NOTCH2)\_5a | chr1:120547951-120548211 |
| 469\_119174\_4853(NOTCH2)\_4a | chr1:120572528-120572610 |
| 469\_119173\_4853(NOTCH2)\_3a | chr1:120594097-120594748 |
| 469\_119172\_4853(NOTCH2)\_2a | chr1:120611947-120612317 |
| 473\_151296\_6391(SDHC)\_7 | chr1:161332118-161332330 |
| 462\_71755\_3020(H3F3A)\_4 | chr1:226259051-226259180 |
| 472\_139794\_5728(PTEN)\_9 | chr10:89725043-89725229 |
| 474\_161710\_9126(SMC3)\_29 | chr10:112363988-112364060 |
| 427\_14238\_472(ATM)\_32 | chr11:108159703-108159830 |
| 473\_151302\_6392(SDHD)\_6 | chr11:111965528-111965694 |
| 462\_77481\_440093(H3F3C)\_1 | chr12:31944692-31944830 |
| 465\_99860\_5604(MAP2K1)\_12a | chr15:66782839-66782953 |
| 469\_118582\_4763(NF1)\_9a | chr17:29527439-29527613 |
| 469\_118586\_4763(NF1)\_13a | chr17:29541468-29541603 |
| 469\_118588\_4763(NF1)\_15a | chr17:29548867-29549120 |
| 469\_118589\_4763(NF1)\_16a | chr17:29550461-29550585 |
| 469\_118591\_4763(NF1)\_18a | chr17:29553452-29553655 |
| 469\_118592\_4763(NF1)\_19a | chr17:29554235-29554309 |
| 469\_118594\_4763(NF1)\_21a | chr17:29556042-29556420 |
| 469\_118605\_4763(NF1)\_32a | chr17:29585373-29585520 |
| 427\_15264\_613(BCR)\_17 | chr22:23651610-23651670 |
| 427\_15265\_613(BCR)\_18 | chr22:23652510-23652620 |
| 427\_15266\_613(BCR)\_19 | chr22:23653883-23654105 |
| 427\_15267\_613(BCR)\_20 | chr22:23655073-23655208 |
| 427\_15269\_613(BCR)\_22 | chr22:23656738-23656901 |
| 470\_127812\_5290(PIK3CA)\_10 | chr3:178935997-178936122 |
| 470\_127813\_5290(PIK3CA)\_11 | chr3:178937018-178937065 |
| 470\_127814\_5290(PIK3CA)\_12 | chr3:178937358-178937523 |
| 470\_127815\_5290(PIK3CA)\_13 | chr3:178937736-178937840 |
| 470\_127816\_5290(PIK3CA)\_14 | chr3:178938773-178938945 |
| 475\_172896\_7097(TLR2)\_4 | chr4:154625935-154626289 |
| 473\_151209\_6389(SDHA)\_2 | chr5:223596-223683 |
| 473\_151210\_6389(SDHA)\_3 | chr5:224474-224636 |
| 473\_151211\_6389(SDHA)\_4 | chr5:225533-225677 |
| 473\_151212\_6389(SDHA)\_5 | chr5:225997-226162 |
| 473\_151213\_6389(SDHA)\_6 | chr5:228299-228448 |
| 473\_151215\_6389(SDHA)\_8 | chr5:233591-233760 |
| 473\_151216\_6389(SDHA)\_9 | chr5:235258-235454 |
| 473\_151217\_6389(SDHA)\_10 | chr5:236542-236714 |
| 473\_151277\_6389(SDHA)\_11 | chr5:240472-240591 |
| 473\_151278\_6389(SDHA)\_12 | chr5:251106-251218 |
| 473\_151279\_6389(SDHA)\_13 | chr5:251452-251901 |
| 473\_151280\_6389(SDHA)\_14 | chr5:254507-254621 |
| 473\_151281\_6389(SDHA)\_15 | chr5:256448-256535 |
| 469\_119424\_4869(NPM1)\_10 | chr5:170833402-170833658 |
| 469\_119426\_4869(NPM1)\_12 | chr5:170837530-170837569 |
| 470\_128930\_5395(PMS2)\_15 | chr7:6013029-6013173 |
| 470\_128929\_5395(PMS2)\_14 | chr7:6017218-6017388 |
| 470\_128928\_5395(PMS2)\_13 | chr7:6018226-6018327 |
| 470\_128927\_5395(PMS2)\_12 | chr7:6022454-6022622 |
| 470\_128926\_5395(PMS2)\_11 | chr7:6026389-6026664 |
| 470\_128926\_5395(PMS2)\_11 | chr7:6026682-6027251 |
| 470\_128924\_5395(PMS2)\_9 | chr7:6031603-6031688 |
| 470\_128920\_5395(PMS2)\_5 | chr7:6042083-6042267 |
| 470\_128919\_5395(PMS2)\_4 | chr7:6043320-6043423 |
| 470\_128918\_5395(PMS2)\_3 | chr7:6043602-6043689 |
| 470\_128917\_5395(PMS2)\_2 | chr7:6045522-6045662 |
| 466\_106280\_4233(MET)\_4a | chr7:116364153-116364218 |
| 464\_94998\_58508(KMT2C)\_35 | chr7:151882642-151882716 |
| 464\_94988\_58508(KMT2C)\_25 | chr7:151902190-151902310 |
| 464\_94987\_58508(KMT2C)\_24 | chr7:151904384-151904513 |
| 464\_94986\_58508(KMT2C)\_23 | chr7:151917607-151917820 |
| 464\_94985\_58508(KMT2C)\_22 | chr7:151919085-151919151 |
| 464\_94984\_58508(KMT2C)\_21 | chr7:151919657-151919767 |
| 464\_94983\_58508(KMT2C)\_20 | chr7:151921099-151921264 |
| 464\_94982\_58508(KMT2C)\_19 | chr7:151921519-151921701 |
| 464\_94981\_58508(KMT2C)\_18 | chr7:151927007-151927112 |
| 464\_94980\_58508(KMT2C)\_17 | chr7:151927304-151927406 |
| 464\_94979\_58508(KMT2C)\_16 | chr7:151932901-151933018 |
| 464\_94978\_58508(KMT2C)\_15 | chr7:151935791-151935911 |
| 464\_94977\_58508(KMT2C)\_14 | chr7:151944986-151945705 |
| 464\_94972\_58508(KMT2C)\_9 | chr7:151960100-151960215 |
| 464\_94971\_58508(KMT2C)\_8 | chr7:151962122-151962294 |
| 464\_94970\_58508(KMT2C)\_7 | chr7:151970789-151970952 |
| 464\_94964\_58508(KMT2C)\_1 | chr7:152132710-152132828 |
| 462\_70181\_2776(GNAQ)\_7 | chr9:80336238-80336429 |
| 462\_70175\_2776(GNAQ)\_1 | chr9:80646015-80646151 |
| 478\_202313\_8233(ZRSR2)\_11 | chrX:15840853-15841107 |
| 464\_90728\_7403(KDM6A)\_13 | chrX:44919266-44919401 |

Gatorseq\_179Genes\_Targets\_Homology\_Regions.bed

# GSBW: LIS Entry & EMR

QCI molecular report is automatically posted (Figure 1)to EPIC Beaker (LIS) automatically using "EPIC\_upload.py" script, which will be verified and once finalized it will be pused in to EpicCare (EMR).

# Software

## Software List

Below is the software used in GSBW.

|  |  |  |
| --- | --- | --- |
| **Software** | **Version** | **Reference** |
| BamUtil | 1.0.14 | http://genome.sph.umich.edu/wiki/BamUtil |
| BamTools | 2.5.1 | [2] |
| BedTools | 2.26.0 | [3, 4] |
| BWA-MEM | 0.7.17 | [5, 6] |
| Datamash | 1.4.0 | gnu.org/s/datamash |
| flock\*\* | 0.2.3 | https://github.com/discoteq/flock |
| iCallSV | 1.0.9 | https://github.com/rhshah/iCallSV |
| Miniconda3 | 4.6.14 | http://conda.pydata.org/miniconda.html |
| Perl | 0.1.16 | perl.org |
| Python\*\*\* | 3.5 | python.org |
| R | 3.2.2 | r-project.org |
| rtg-tools | 3.10.1 | realtimegenomics.com/products/rtg-tools |
| SAMtools | 1.3.1 | [7, 8] |
| Sambamba | 0.7.0 | [9] |
| Nextflow | 19.07.0 | [10] |
| SHYAML\*\*\* | 0.4.1 | https://github.com/0k/shyaml |
| VarDictJava | 1.6 | [11] |
| VCFtools | 0.1.16 | [12] |
| vt | 0.57721 | [13] |
| Singularity | 3.4.0 | [14] |

\*\* Only on ahc-path-data19; \*\*\* Both on ahc-path-data19 and HiperGator

## Software and Pipeline Update

GSBW uses both inhouse and open source software (listed in Prerequisite Software section) components. Below is the procedure for monitoring, recording, and implementing patch releases and upgrades to the GSBW.

**Monitoring**: Manual monitoring is performed every three months for any major patch releases and upgrades for software components used in GSBW. The patch releases and upgrades that have functionalilty impact on GSBW will be integrated in to GSBW.

**Recording & Implementation**: Manual check for any major patch releases and upgrades for software components used in GSBW. Major patch releases and upgrades for software components will be integrated in to GSBW and are documented in the GSBW version log “GSBW\_Pipeline\_Release\_Log.docx”. The extent of revalidation and/or confirmation depends on the upgrade’s impact on the functionality of GSBW and will be approved by the laboratory director.

# Reference Material

## Human Genome

### Download Source

Human genome version hg19 was downloaded on March 21st, 2014 from "http://ftp.broadinstitute.org/bundle/2.8/hg19". Currently this weblink is moved to "ftp://ftp.broadinstitute.org/bundle/hg19".

"ucsc.hg19.fasta": contains complete human genome

### Removing Haplotypes

"ucsc.hg19.fasta" file also contain nine haplotypes as listed below.

The seven haplotypes corresponding to the human chromosome 6 major histocompatibility complex (MHC) are listed in the below table along with their reference seqeunce.

|  |  |
| --- | --- |
| **Chromosome** | **Haplotype** |
| chr6 | Reference sequence (PGF) |
| chr6\_apd\_hap1 | 6-APD |
| chr6\_cox\_hap2 | 6-COX (previously CX) |
| chr6\_dbb\_hap3 | 6-DBB |
| chr6\_mann\_hap4 | 6-MANN |
| chr6\_mcf\_hap5 | 6-MCF |
| chr6\_qbl\_hap6 | 6-QBL |
| chr6\_ssto\_hap7 | 6-SSTO |

http://vega.sanger.ac.uk/info/data/MHC\_Homo\_sapiens.html

Horton, Roger, et al. "Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project." *Immunogenetics* 60.1 (2008): 1-18.

The two haplotypes corresponding to chr4 and chr17 are "chr4\_ctg9\_hap1" and "chr17\_ctg5\_hap1".

"ucsc\_hg19\_without\_halotypes.fa": contains complete human genome but without above nine haplotype sequences.

### Masking *CRLF2* on Y Chromosome

*CRLF2* belong to pseudoautosomal regions and thus have copies in X and Y chromosomes.

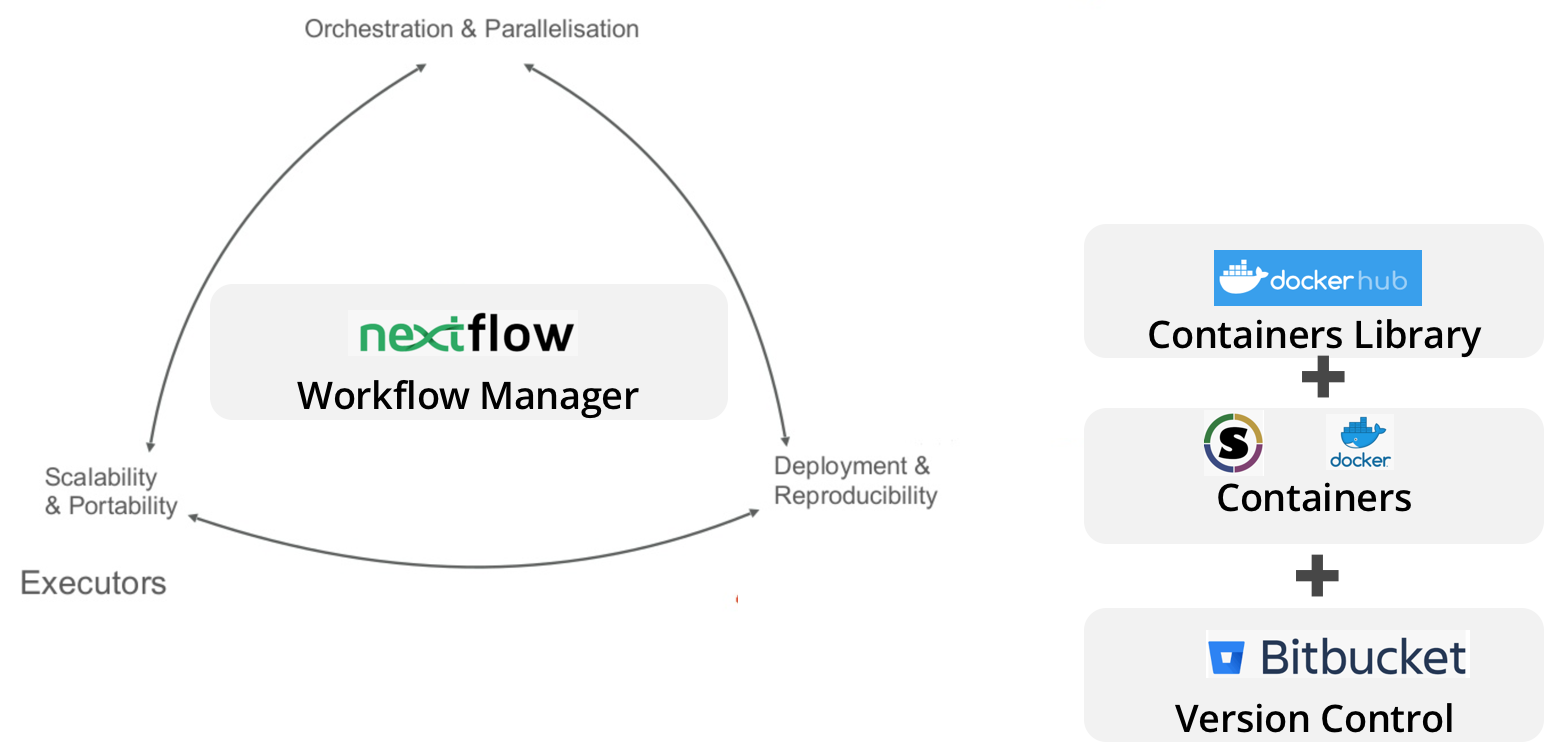
The homology of *CRLF2* targets (after flanking 100 bp each side) in GatorSeq were checked between X and Y chromosomes. They are 100% identical, in which case *CRLF2* target regions in the Y-chromosome of our reference genome are masked to avoid sequencing alignment and depth bias among male and female samples.

"ucsc\_hg19\_without\_halotypes\_CRLF2\_100bp\_padded\_chrY\_masked.fasta" : This file contains complete human genome but without above nine haplotype sequences and *CRLF2* targets masked on Y chromosome. This is the genome version that was used in our analysis.

# Portable, Reliable, Reproducible, and Scalable Software Workflows

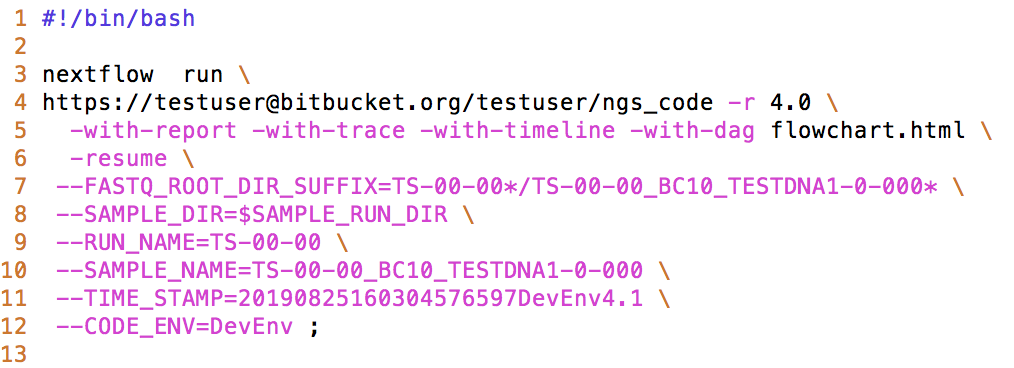
We are using industrial software development best practices to develop portable, reliable, reproducible, and scalable GatorSeq software workflows (see Figure 5).

**Figure 5**: Clinical grade software development model and tools.



## Retrospecitve Analysis for Auditing

For users to retrospectively re-run the analysis (e.g., auditing) all the users have to do is pass on “-r” option which will rerun the pipeline with exact OS environments and all bioinformatics software parameters.



# GatorSeq Pipeline Code

## Version Control

GatorSeq bioinformatics workflow code is tracked via version control system using UFRC Git server, hosted at "git.rc.ufl.edu". Various version changes of the pipeline is documented in the GSBW version log “GSBW\_Pipeline\_Release\_Log.docx”.

## Version Traceability

The specific version of the GSBW pipeline used to generate NGS data files are traceable for each patient from the following output log file “START.txt” and is listed in the line beginning with “GSBW\_VERSION”.

## Code Location

ahc-path-data19 :

"/home/path-svc-mol/GatorSeq/ProdEnv/gatorseq\_linux\_code"

GatorSeq (UF HPC):

"/home/path-svc-mol/GatorSeq\_Share/ProdEnv/gatorseq\_hpc\_code"

# Computational Infrastructure

## Illumina NextSeq 500

* On Medical device network
* Not accessed via internet

# Computational Infrastructure

## Illumina NextSeq 500

* On Medical device network
* Not accessed via internet

## Pathology Network Drive

* Holds NGS data analysis input and output
* Highly secure and is designed to hold PHI
* Redundant storage
* Automatic data backup
* Can only be accessed by designated personal of Molecular Path Labs
* Accessed via internal UF health network or using VPN

## ahc-path-data19

* Accessed via internal network / UF health VPN only
* Highly secure and is designed to hold PHI

## HiPerGator

* Communication between ahc-path-data19 and HiPerGator is via secure shell and is one-way i.e., ahc-path-data19 can access HiPerGator but not vice versa
* Can not store PHI
* De-identified NGS data is transferred to HiPerGator via secure shell
* Data analysis jobs are submitted on HiPerGator
* Data analysis output is transferred to ahc-path-data19 and removed from HiPerGator

## Qiagen Clinical Insight (QCI)

* Qiagen Cloud based application
* Accessed via internet
* Vender manages all data security compliances

## Data Backups

### Illumina NextSeq 500

Data is stored on Illumina’s BaseSpace server locally and does not have a direct backup of BaseSpace. However, after each sequencing run is done, the data is manually downloaded on to Pathology Network Drive, making it as a back up copy.

### Pathology Network Drive

Pathology network drive is part of the UF Academic Health Center (AHC) enterprise NAS and has the following data protection configured.

*Local backup system*: On the local system AHC perform hourly backups from 9am-5pm, daily backups for 45 days, and weekly backups for 12 weeks on rotation basis i.e., oldest copy is deleted to store the new copy. This means data can be recovered hourly during the day time from the hourly backup rotation, daily backup data can be recovered for up to 45 days from the daily backup rotation, weekly backup data can be recovered for up to 12 weeks from the weekly backup rotation.

*Remote backup system*: In addition to local backups, AHC replicate all the data daily to a offsite secondary NAS where additional 7 daily backups on rotation basis. If data storage suffers a catastrophic event the total data loss (RPO) could be up to 24 hours depending on the time the last replication completed. The secondary unit is bought back to online in less than an hour (RTO).

NOTE: RPO – Recovery Point Objective this is the maximum amount of data one could lose, i.e., 24 hours since that is the replication frequency.

NOTE: RTO – Recovery Time Objective this is the amount of time it would take to get the data back online and accessible if an event were to occur i.e., less than an hour in our case.

### ahc-path-data19

UF Health IT does the back up.

### HiPerGator

GatorSeq pipeline is revision control on GIT server, we should be able reinstate the pipeline within a day.

### QCI

Qiagen maintain all the backups of the database. Also, we back up the reported generated onto our Pathology Network Drive.

# References

1. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med. 2017.

2. Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 2011;27:1691–2.

3. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

4. Quinlan AR. BEDTools: the Swiss‐army tool for genome feature analysis. Curr Protoc Bioinforma. 2014;:11–2.

5. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26:589–95.

6. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60.

7. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.

8. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–93.

9. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31:2032–4.

10. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35:316.

11. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;:gkw227.

12. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.

13. Tan A, Abecasis GR, Kang HM. Unified representation of genetic variants. Bioinformatics. 2015;:btv112.

14. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS One. 2017;12:e0177459.