454 Sequencing System Software Manual
Version 2.9

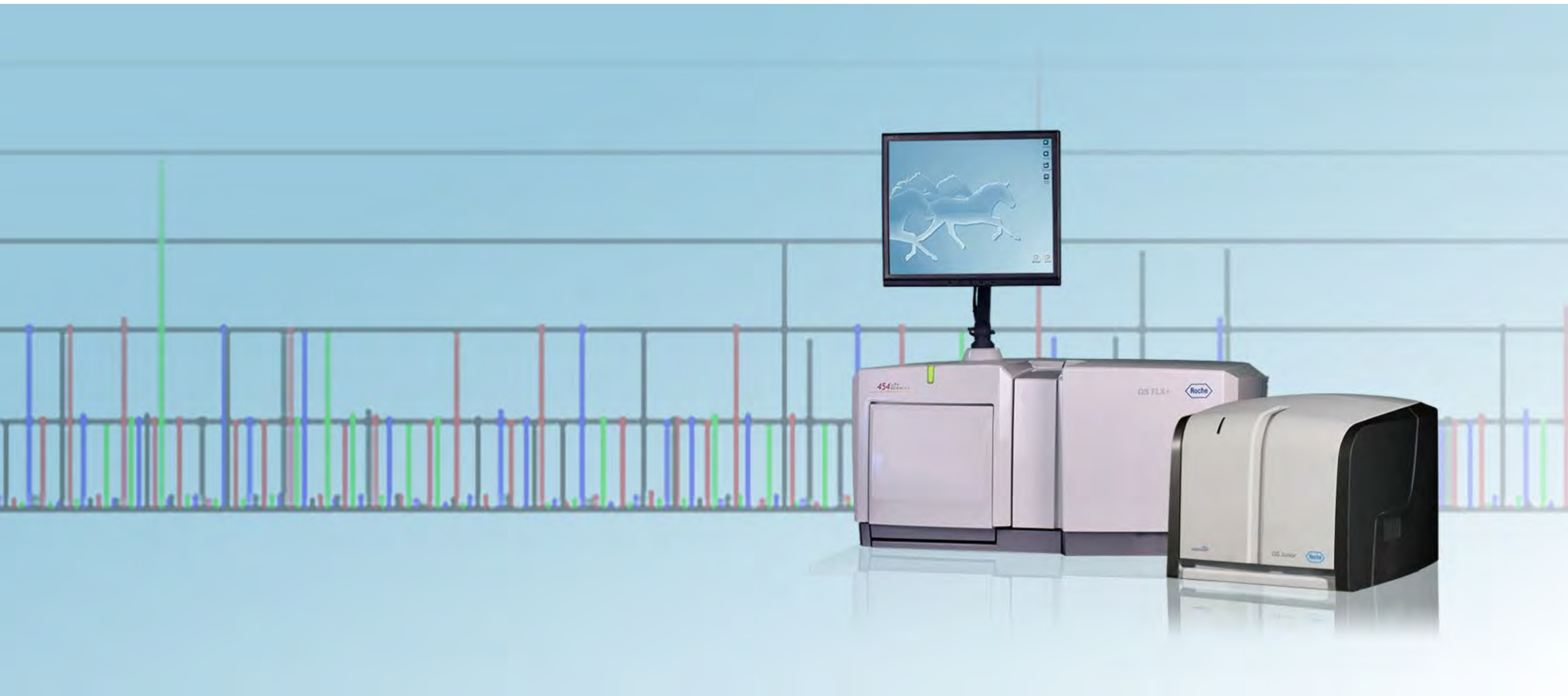
Part B:

GS Run Processor, GS Reporter, GS Run Browser, GS Support Tool

June 2013

	Instrument	Kit
✓	GS Junior	Junior
✓	GS FLX+	XL+
✓	GS FLX+	XLR70
✓	GS FLX	XLR70

For life science research only. Not for use in diagnostic procedures.



GS Run Processor, GS Reporter, GS Run Browser, GS Support Tool

1 GS Run Processor

The GS Run Processor application is identical for both the GS FLX and GS FLX+ Instruments. Unless otherwise indicated, **GS FLX+ Instrument** refers to both instruments.

The GS Run Processor application is identical for both the GS Junior and GS FLX+ Instruments. However, due to differences in Instrument hardware and software, some references in this manual will be specific to either the GS Junior or the GS FLX+ Instruments. There are four main differences: computing resources, camera resolution, the file format of stored images, and the size of the PicoTiterPlate or PTP device.

- The GS Junior Instrument does not have on-instrument computing capability, but instead includes an attendant PC with all of the 454 Sequencing software components installed. It is designed to perform image and signal processing during a run, as well as data analysis after a run. The 454 Sequencing software can optionally be installed on a datarig.
- Therefore, references to on-instrument software or computation are specific to the GS FLX+ Instrument, references to the GS Junior attendant PC are specific to the GS Junior Instrument, while references to a datarig can apply to either instrument.
- The GS Junior Instrument has a higher resolution camera compared to the GS FLX+ Instrument. Therefore, references to unbinned images are specific to the GS Junior Instrument.
- Image data from the GS FLX+ Instrument are stored as compressed .png image files, but the legacy .pif image format from older runs is still supported. Image data from the GS Junior Instrument are stored as .pif image files. Therefore, references to .png are specific to runs from the GS FLX+ Instrument, while references to .pif can apply to either runs from the GS Junior Instrument or older runs from the GS FLX+ Instrument.
- The PTP device on the GS FLX+ Instrument supports division into multiple (2, 4, 8 or 16) regions. The GS Junior Instrument PTP device only supports a single region. Therefore, any references to multiple regions are specific to the GS FLX+ Instrument.

The GS Run Processor application performs the data processing of a sequencing run to convert raw images into high quality sequence information. This is done in two phases: image processing (Section 1.2) and signal processing (Section 1.3), which together are called full processing. Signal processing itself is done in two phases: signal corrections (Section 1.3.1) and read quality filters (Section 1.3.2).

The various data processing paths that a user can follow are described in the Overview of this manual. The run processing types that are available during run setup in GS Sequencer or GS Junior Sequencer depend on the sequencing kit used (see Part A of this manual). The data processing options available in GS Run Processor depend on the data processing previously applied to the data in the selected R_ or D_ directory (see Section 3.4.3).

During a sequencing run, the CCD camera on the GS Junior or GS FLX+ Instrument takes an image of the PicoTiterPlate device for each nucleotide flow of the sequencing run protocol. This image capturing step generates the image files (.pif for GS Junior Instrument; .png for GS FLX+ Instrument) for each of the flows. The image analysis step finds raw wells (a well containing a DNA fragment that produced light due to base incorporations during the sequencing run) across the entire PTP device, and implements algorithms to normalize the background. This is followed by signal processing, which corrects the raw flow signals, trims or rejects reads using quality filters, and performs basecalling.

The data flow for the image and signal processing is illustrated in Figure 1.

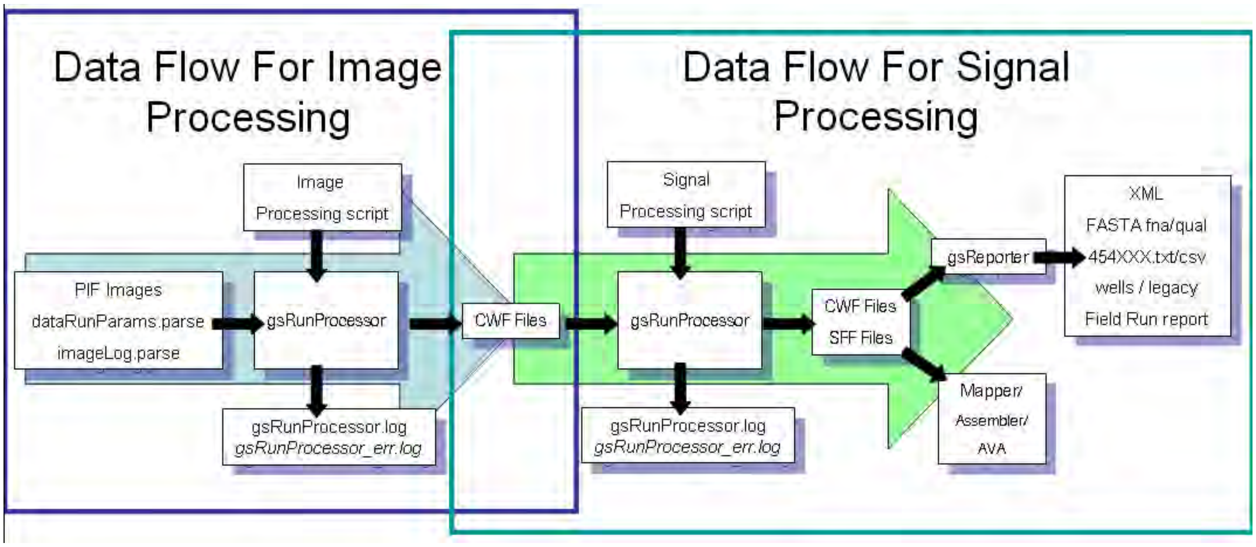


Figure 1: Data Flow for Image and Signal Processing.

The core of the GS Run Processor application is the gsRunProcessor executable, which contains algorithms for determining location of the loading regions of the PTP device, for image processing, signal correction, read filtering, basecalling and run metrics generation. The components of GS Run Processor are described in Table 1 below:

Component	Description
gsRunProcessor executable	Contains the core algorithms of the data processing application
startGsProcessor script	Calls the gsRunProcessor executable
Launching scripts (runImagePipe, runAnalysisPipe, runAnalysisFilter, *etc.*)	Provides a command line interface to the gsRunProcessor executable
Processing scripts (imageProcessing.xml, signalProcessing.xml, *etc.*)	A set of XML files passed to the gsRunProcessor executable containing default commands and parameters for various processing pipeline options
gsRunProcessorManager executable	Provides a simple job batching system for the gsRunProcessor executable in shared user environments
gsReporter executable	Extracts reports and other output data from the processed CWF files

Table 1: GS Run Processor components.

There are some environmental variables that control the gsRunProcessorManager application. One key variable is the GS_LAUNCH_MODE, which can operate in one of the modes listed in Table 2:

Value	Description
SINGLE	Starts a single copy of the gsRunProcessor (Default)
MULTI	Starts multiple copies of the gsRunProcessor, equal to the number of processors in the current workstation (Non-cluster)
MPI	Uses 'mpiexec' for launching jobs on a compute cluster. Refer to the *SysAdmin Guide* for details on configuring the gsRunProcessor suite for use on a non-Titanium cluster.
GSRPM	Starts the job using the gsRunProcessorManager. Will submit jobs to the same job queue as users who use gsRunBrowser to submit processing jobs. Recommended as a multi-user job queuing system.

Table 2: gsRunProcessorManager Environmental Variable GS_LAUNCH_MODE.

If GS_LAUNCH_MODE is set to GSRPM, the jobs are launched via the gsRunProcessorManager and the queued jobs can be monitored and aborted via the gsRunProcessorManagerCtrl command. This is the recommended configuration and is also the default.

The gsRunProcessorManager manages the gsRunProcessor application and launches and queues all processing jobs. The gsRunProcessorManager daemon is implemented as a system V script. If the Data Processing software was installed with system-level privileges, the gsRunProcessorManager will automatically start when the system is restarted. The following command (run as root) can be used to start the manager manually if necessary:

```
/etc/init.d/gsRunProcessorManager start
```

The following command can be used to view the job queue:

```
gsRunProcessorManagerCtrl queue
```

This command will return the job _ids of currently queued processing jobs. The following command can be used to abort a job that is in the job queue:

```
gsRunProcessorManagerCtrl abort job_id
```

1.1 Data Processing Pipelines

The data processing steps can be configured as a part of the sequencing run using the Instrument Procedure Wizard (see Part A of this manual), invoked post-run using the GS Run Processor Manager tool in the GS Run Browser application (see Section 3.4.3), or by using the command line interface (CLI) on an attendant PC or datarig (described below). Each data processing pipeline includes report generation by gsReporter (Section 2.1), plus the ability to run an optional user-provided script (postAnalysisScript.sh) for additional data processing or analysis (Section 5.2).

By default, the 454 Sequencing system software supports eleven data processing pipelines, although additional pipelines may be installed from the Add-Ons disc or customized using a template. See Section 1.3.5 for a comparison of pipeline stringencies.

- Image processing only
- Full processing for Shotgun or Paired End
- Full processing for Amplicons
- Full processing for Long Amplicons #1
- Full processing for Long Amplicons #2
- Full processing for Long Amplicons #3
- Signal processing only for Shotgun or Paired End
- Signal processing only for Amplicons
- Signal processing only for Long Amplicons #1
- Signal processing only for Long Amplicons #2
- Signal processing only for Long Amplicons #3

Data processing pipelines can be accessed through the CLI using data processing launch script commands, which launch a pipeline script that sends data processing commands to GS Run Processor. These scripts are XML-based text files located in *installDir*/apps/gsRunProcessor/etc/gsRunProcessor/, where installDir is the main software installation path (*e.g.* /usr/local/rig/ on the GS FLX+ Instrument or /opt/454/ on the GS Junior attendant PC).

The pipeline launch script commands runImagePipe, runAnalysisPipe runAnalysisPipeAmplicons, *etc.* are used to specify which of the default pipelines to run for a specified R_ or D_ directory (see Table 3). An additional launch script command, runAnalysisFilter, may be used to run customized signal processing scripts to reprocess sequencing data with user-modified filter settings. These signal processing (filter-only) pipelines skip the signal correction steps while still performing read rejection, read trimming, and basecalling. See Section 1.3.6 for a full description of the adjustment of filter parameters and the use of runAnalysisFilter.

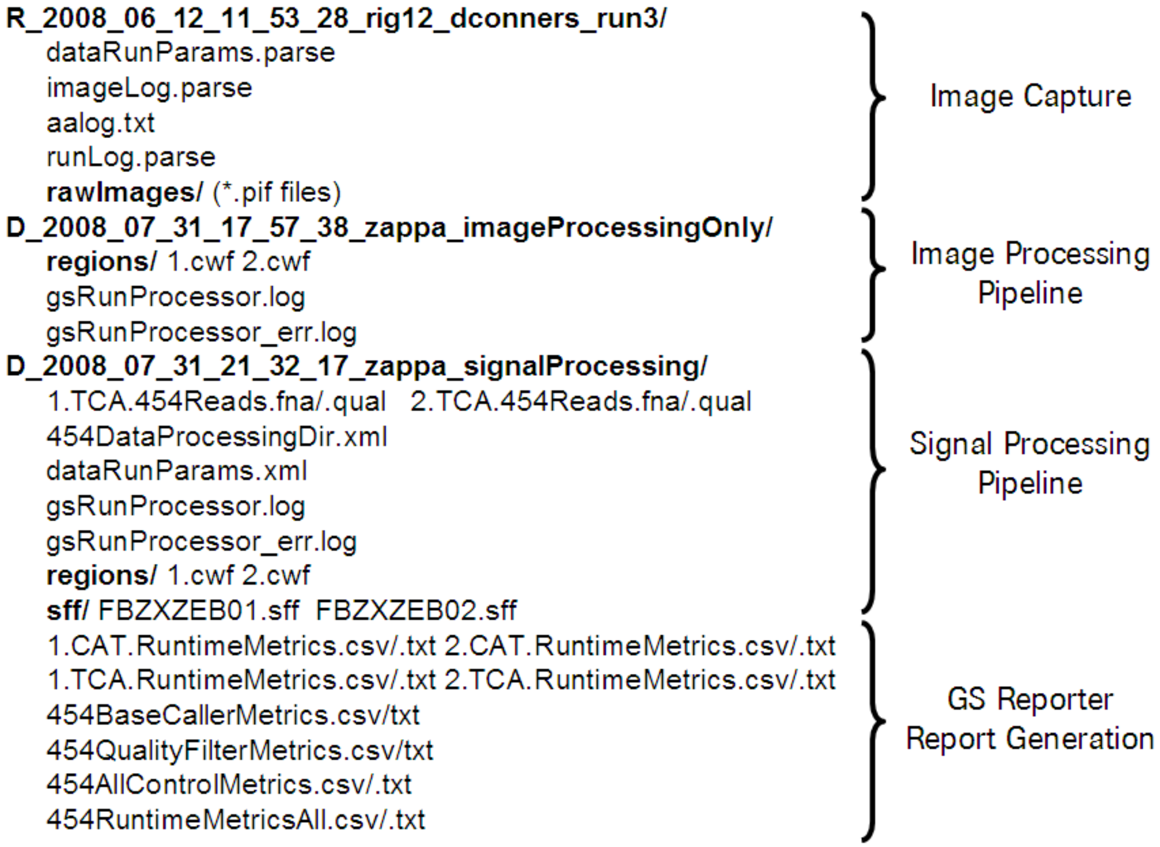
Data Processing Pipeline	Pipeline Launch Script Command R_ or D_ Directory	Pipeline Script
Image processing only	runImagePipe R_ Directory	imageProcessingOnly.xml
Full processing for Shotgun or Paired End	runAnalysisPipe R_ Directory	fullProcesssing.xml
Full processing for Amplicons	runAnalysisPipeAmplicons R_ Directory	fullProcesssingAmplicons.xml
Full processing for Long Amplicons #1	runAnalysisPipeLongAmplicons1 R_ Directory	fullProcesssingLongAmplicons1.xml
Full processing for Long Amplicons #2	runAnalysisPipeLongAmplicons2 R_ Directory	fullProcesssingLongAmplicons2.xml
Full processing for Long Amplicons #3	runAnalysisPipeLongAmplicons3 R_ Directory	fullProcesssingLongAmplicons3.xml
Signal processing for Shotgun or Paired End	runAnalysisPipe D_ Directory	signalProcessing.xml
Signal processing for Amplicons	runAnalysisPipeAmplicons D_ Directory	signalProcessingAmplicons.xml
Signal processing for Long Amplicons #1	runAnalysisPipeLongAmplicons1 D_ Directory	signalProcessingLongAmplicons1.xml
Signal processing for Long Amplicons #2	runAnalysisPipeLongAmplicons2 D_ Directory	signalProcessingLongAmplicons2.xml
Signal processing for Long Amplicons #3	runAnalysisPipeLongAmplicons3 D_ Directory	signalProcessingLongAmplicons3.xml
Signal processing (filter-only) for Shotgun or Paired End	runAnalysisFilter --pipe=custom.xml D_ Directory	custom.xml (see Section 1.3.6 for details) gsRunProcessor --template=filterOnly
Signal processing (filter-only) for Amplicons	runAnalysisFilter --pipe=custom.xml D_ Directory	custom.xml (see Section 1.3.6 for details) gsRunProcessor --template= filterOnlyAmplicons
Signal processing (filter-only) for Long Amplicons #1	runAnalysisFilter --pipe=custom.xml D_ Directory	custom.xml (see Section 1.3.6 for details) gsRunProcessor --template= filterOnlyLongAmplicons1
Signal processing (filter-only) for Long Amplicons #2	runAnalysisFilter --pipe=custom.xml D_ Directory	custom.xml (see Section 1.3.6 for details) gsRunProcessor --template= filterOnlyLongAmplicons2
Signal processing (filter-only) for Long Amplicons #3	runAnalysisFilter --pipe=custom.xml D_ Directory	custom.xml (see Section 1.3.6 for details) gsRunProcessor --template= filterOnlyLongAmplicons3

Table 3: Data Processing Pipelines, with associated Pipeline Launch Script Commands and Pipeline Scripts.

> The ‘paired end’ pipeline with associated launch script command ‘runAnalysisPipePairedEnd’ have been merged into the standard pipeline. Use ‘runAnalysisPipe’ to process paired end runs via the command line, or select ‘Processing for Shotgun or Paired End’ from the GS Run Browser or instrument GUIs.

An additional command, **startGsProcessor**, is used to prepare the directory structure for GS Run Processor jobs. It is a wrapper script that is called initially by all the ‘run’ launch script commands. Its use will be transparent to most users. It is described in the appendix Section 6.2 for reference.

The directory structure for a GS FLX+ system sequencing run on a datarig, after the data processing and report generation steps are carried out, is described below for a two-region sequencing run: (Note that for a GS Junior system sequencing run, the directory structure would be similar, but with a single region).



```
R_2008_06_12_11_53_28_rig12_dconners_run3/
    dataRunParams.parse
    imageLog.parse
    aalog.txt
    runLog.parse
    rawImages/ (*.pif files)
D_2008_07_31_17_57_38_zappa_imageProcessingOnly/
    regions/ 1.cwf 2.cwf
    gsRunProcessor.log
    gsRunProcessor_err.log
D_2008_07_31_21_32_17_zappa_signalProcessing/
    1.TCA.454Reads.fna/.qual   2.TCA.454Reads.fna/.qual
    454DataProcessingDir.xml
    dataRunParams.xml
    gsRunProcessor.log
    gsRunProcessor_err.log
    regions/ 1.cwf 2.cwf
    sff/ FBZXZEB01.sff  FBZXZEB02.sff
    1.CAT.RuntimeMetrics.csv/.txt 2.CAT.RuntimeMetrics.csv/.txt
    1.TCA.RuntimeMetrics.csv/.txt 2.TCA.RuntimeMetrics.csv/.txt
    454BaseCallerMetrics.csv/txt
    454QualityFilterMetrics.csv/txt
    454AllControlMetrics.csv/.txt
    454RuntimeMetricsAll.csv/.txt
```

Image Capture
Image Processing Pipeline
Signal Processing Pipeline
GS Reporter Report Generation

Figure 2: Sequencing run data file structure after full Data Processing and default report generation.

1.2 Image Processing

In the image processing step, the initial PPi or ATP flow is used to define the PTP regions across the plate. The specific background for each nucleotide type is also computed from the key flows and these are combined to subtract the background for each signal-producing location on the plate for the key flow images. By default, the last usable flow is defined based on preliminary well signal intensity levels, and only the included flows are used in the well finding process.

Prior to version 2.7, GS Run Processor on the GS Junior Instrument used images that had been 'binned' to a lower resolution to reduce noise. It is now possible to analyze 'unbinned' images for well finding, which leads to improvements in identification of well edges. The positions of local signal maxima in each included flow are accumulated into a consensus image. Local maxima in the consensus image are then merged if they are close enough to each other, and are used to define the center of each raw well. Flow signal information is extracted for each raw well and organized into composite well format (cwf) for further processing. Note that one consequence of this improved well finding process is that the software will not report the number of wells found until later in the run.

Image processing filter parameters are described in Table 4.

Image Processing Parameter	Description	Values	Shotgun/PE	Amplicon
minConsensusSignal	Minimum average intensity of the positive key flows to be considered as a potential well.	integer	GS Junior kit: 20* XLR70 kit: 20 XL+ kit: 20	GS Junior kit: 60 XLR70 kit: 20 XL+ kit: n/a
useBicubic	Use bicubic upsampling of images ('false' = bilinear).	true/false	All: true	All: true

Table 4: Image Processing Parameters. Stringency can be increased by using larger green values or lower red values.
***The imageProcessingOnly pipeline for the GS Junior system uses an alternative value of '60'.**

There is one CWF file generated for each sequencing run on a GS Junior Instrument, while for the GS FLX+ Instrument, there is one CWF file generated for each *region* of the PTP device. The total size of the raw images depends on the sequencing kit used, the number of regions, the flow pattern, and the length of the sequencing run in cycles or flows (Table 5).

Sequencing Kit	# Regions	Flow Pattern	Run Length	Raw Image Size
GS FLX Titanium Sequencing Kit XLR70	2	Cyclic	200 cycles	~18 GB
GS FLX Titanium Sequencing Kit XL+	2	Flow pattern A (Cyclic)	400 cycles	~35 GB
GS FLX Titanium Sequencing Kit XL+	2	Flow pattern B (Acyclic)	1779 flows	~38 GB
GS Junior Titanium Sequencing Kit	1	Cyclic	200 cycles	~10 GB

Table 5: Approximate raw image size of sequencing runs.

The image processing computation is fast enough to be run concurrently with a sequencing run using either the GS FLX Titanium Sequencing Kit XLR70 or GS Junior Titanium Sequencing Kit, but continues a few hours beyond the fluidics stage of a run using the GS FLX Titanium Sequencing Kit XL+.

The following files are generated by the sequencing run image capture and are used as input to the image processing step:

- dataRunParams.parse
- imageLog.parse
- ptpImage.png or ptpImage.pif
- runLog.parse
- aaLog.txt
- rawImages (subdirectory)- containing the image data in .png or .pif formatted files
    - 00000.png or 00000.pif (initial PPi or ATP flow image)
    - 00001.png, 00002.png, … or 00001.pif, 00002.pif, … (sequential nucleotide flow images)

Image data from the GS FLX+ Instrument are stored as compressed .png image files, but the legacy .pif image format from older runs is still supported. Image data from the GS Junior Instrument are stored as .pif image files.

For the 454 Sequencing system software applications version 2.3 and higher, all the image files are used by data processing applications, thus it is strictly required that these files not be altered.

1.2.1 Launching the Image Processing Pipeline

There are three ways to launch the image processing step of the GS Run Processor application:

- As part of the automated pipeline processing system of the GS Junior or GS FLX+ instruments (see GS Junior Sequencer or GS Sequencer application, in Part A of this manual)
- From the GS Run Processor Manager tool accessed via the GS Run Browser application (see section 3.4.3)
- From the command line interface (CLI)

The command line application can be launched by the runImagePipe command, which has the following command line structure:

```
runImagePipe [options] RUN_DIRECTORY
```

The RUN_DIRECTORY specified must contain a dataRunParams.parse file, the imageLog.parse file and the complete contents of the rawImages subdirectory. The options for the runImagePipe command are listed in Table 6.

Option	Description
--progress	Displays real-time progress of the job's processing.
--pipe=XMLFILE	Specifies a pipeline processing script file. (Specifying the .xml extension is optional.)
--reg=REGION_NUM	Specifies processing of data from a particular PTP Region. Regions can be specified as a range or as a comma-separated list ('1-4' or '1,2,5,9'). (Note: Multiple regions are only supported on the GS FLX+ Instrument.)
-verbose	Provides verbose log output useful for troubleshooting.

Table 6: runImagePipe command options.

1.2.2 Image Processing Output

The image processing step creates an output data directory of the format:

D_...imageProcessingOnly

The following files are generated by the image processing and are used as input to the signal processing step.

Output	Description
dataRunParams.xml	Contains the results of the region finding, a list of the Control DNA sequences detected in the data set, and meta-data about the sequencing run.
Region/regionnum.cwf	Composite Wells Format (CWF) files - Contains the raw flowgram information from the processed images, run metrics, run meta-data, all intermediate results and low-resolution image for each base flow.
gsRunProcessor.log	Contains logged messages from the image processing.
gsRunProcessor_err.log	Contains logged error messages from the image processing.

Table 7: Image processing output files.

1.3 Signal Processing

The signal processing step of the GS Run Processor application analyzes the signal data for each flow for all active wells of each loading region of the PicoTiterPlate device, using the flowgram data generated during the image processing step and stored in the .cwf files. The signal processing uses different internal parameters for signal processing of standard shotgun and paired end libraries compared to amplicon libraries, and thus the library type must be specified for signal processing jobs.

The choice of one of the amplicon signal processing pipelines is important for amplicon libraries, because they tend to generate much stronger signals than other library types. They consist of shorter fragments that amplify to a greater extent during emPCR Amplification. Very strong signals, coupled with the fact that many wells in runs with amplicon libraries contain similar sequences, can result in 'ghost wells', areas where light is detected but that do not truly correspond to DNA sequencing events. The special signal processing configuration for amplicons includes a number of amplicon-specific corrections, including an additional well density correction step (wellScreener) at the beginning of the signal correction phase.

Signal processing performs a series of normalization, correction, and quality filtering steps and then outputs the remaining (high quality) signals into flowgrams for each well (read). Signal processing also generates basecalls with associated quality scores for the individual reads, and outputs these data as Standard Flowgram Format (or SFF) files that contain all the sequence trace data for the reads. All the data analysis applications (GS *De Novo* Assembler, GS Reference Mapper or GS Amplicon Variant Analyzer) use these SFF files as input.

In a signal processing job, the GS Run Processor software performs the following operations:

1. Apply signal corrections (Section 1.3.1).
2. Apply read rejecting filters, based on sequence characteristics or signal quality (Key Pass, Primer-Dimer, Dot, Mixed and Ambiguous Read filters; Section 1.3.2.1, and Counting or Scoring Valley Filter; Section 1.3.2.2).
3. Apply read trimming/rejecting filters, which trim read ends for low quality or adaptor sequence and discard reads that have been trimmed too short (Signal Intensity, Signal Trimback, Adaptor Trim, Trimback Valley, and Basecall Quality Score Filters; Section 1.3.2.2).
4. Generate flowgrams and basecalled sequences with corresponding quality scores for all individual, high quality reads (*i.e.* those which passed all filters) and output to CWF and SFF files, one per PTP Region processed (Section 1.3.3). (Note: For the GS Junior Instrument, the entire PTP device is a single region).
5. Summarize the ten most common 50 base sequences observed during a run (Section 1.3.3.2).
6. Report 3'-adaptor matches (Section 1.3.3.3).
7. Generate metrics files with GS Reporter (Section 2.2).
8. Run a user-provided postAnalysisScript.sh script for additional data processing or analysis (Section 5.2).

1.3.1 Signal Corrections

Artifact and signal corrections are applied as the first step in signal processing, in the following order.

1. **Well density correction (wellScreener)** – filter out 'ghost' wells caused by high signal intensity (first pass, for amplicon pipelines only).
2. **Nucleotide normalization (nukeSignalStrengthBalancer)** – normalizes the signal strengths of different base incorporations.
3. **Inter-well crosstalk correction (blowByCorrector)** – corrects individual wells for the additional signal intensity conveyed by neighboring high intensity signal wells.
4. **Reagent flow event balancer (nucValveEventBalancer)** – corrects anomalous signal spikes due to reagent valve events (cyclic flow pattern, shotgun pipeline only).
5. **Signal global droop correction (globalDroopCorrector)** – correct for signal reduction over the course of the sequencing run (first pass, shotgun processing only). Not enabled for runs using XLR70 sequencing kit.
6. **CAFIE correction (cafieCorrector)** - CArry Forward & Incomplete Extension, corrects out-of-phase errors using a probability model in order to improve signal and decrease noise.
   - **Carry Forward** occurs when a trace amount of nucleotide remains in a well after the apyrase wash, perpetuating premature nucleotide incorporations for specific sequence combinations during the following flows. While this generally affects only a small percentage of DNA strands per bead, it causes those strands to continue to incorporate nucleotides out-of-phase with respect to the rest of the strands.
   - **Incomplete Extension** occurs when some DNA strands on a bead fail to incorporate during a nucleotide flow. This is more likely to occur with higher order homopolymers, and can be due to localized reagent concentration differences within the PTP device. Strands that fail to incorporate the appropriate nucleotide must wait for the next flow of that nucleotide to continue extending. If the following nucleotide in the DNA sequence flows before this happens (guaranteed with the cyclic TACG flow pattern), those strands will continue to incorporate out-of-phase.
7. **Signal global droop correction (globalDroopCorrector)** – correct for signal reduction over the course of the sequencing run (second pass, shotgun processing only). Not enabled for runs using XLR70 sequencing kit.
8. **Recursive CAFIE (recursiveCafieCorrector)** – uses an iterative approach to improve the CAFIE corrections after the primary CAFIE correction has been applied. Recursive CAFIE correction is applied by default for all runs except those using the XLR70 sequencing kit with a shotgun pipeline.
9. **Reagent flow event balancer (nucValveEventBalancer)** – corrects anomalous signal spikes due to reagent valve events (cyclic flow pattern, amplicon pipelines only).
10. **Nucleotide normalization (nukeSignalStrengthBalancer)** – normalizes the signal strengths of different base incorporations (second pass for amplicon pipelines only).
11. **Signal global droop correction (globalDroopCorrector)** – correct for signal reduction over the course of the sequencing run (third pass, shotgun processing only). First pass for runs using XLR70 sequencing kit.
12. **Individual Well scaler (individualWellScaler)** – well-by-well correction of well signal.
13. **Flow balancer (flowBalancer)** – flow-by-flow correction of 0-mer, 1-mer, and 2-mer signal heights (acyclic flow pattern only).
14. **Residual background subtraction and rescaling (mostLikelyErrorSubtractor)** – correct for residual background.
15. **Well density correction (wellScreener)** – filter out 'ghost' wells caused by high signal intensity (second pass for amplicon pipelines, first pass for shotgun pipeline).

The remaining corrected wells are considered to contain valid sequence information, albeit of varying quality. They are counted as Raw Wells in the totalRawWells output metric.

1.3.2 Read Quality Filters

1.3.2.1 Read Rejecting Filters

Read rejecting filters are applied after artifact filters and signal correction as a quick pass/fail test to discard no-information or low-information active wells. These stringent filtering algorithms ensure better results, even with a lower number of reads. Read rejecting filters are applied in the following order.

1. **Key Pass Filter** – passes reads that begin with a valid sequencing key. Sequencing keys are used to distinguish wells with library beads from those with control beads (see Table 8) and to discard false positive wells. The sequencing key is trimmed from the 5' end of the sequence in Basecaller. The passed read count is included in the *numKeyPass* metric.

Sequencing Key	Description	Chemistry	# Flows*
TCAG	Standard Library	All chemistries	8 flows
GACT	Rapid Library	GS Titanium and GS Junior titanium chemistry	9 flows
CATG	Type I Control (AvTF)	GS Titanium and GS Junior titanium chemistry	12 flows
ATGC	Type II Control (EcTF)	GS Titanium and GS Junior titanium chemistry	11 flows

Table 8: Sequencing keys. Controls (test fragments) are described in Section 7.1. *All sequencing keys are sequenced within the first twelve flows, which are identical between the acyclic flow pattern B and the cyclic flow pattern A.

DO NOT mix samples prepared with the Rapid Library protocol (GACT key) with samples prepared with the Standard Library protocol (TCAG key) in the same region of the PTP device.

2. **Primer-Dimer Filter (doShortSignalCheck)** – rejects reads with too few total flows. This can occur when adaptors combine with each other during emPCR to form short primer-dimer sequences. The Primer-Dimer Filter discards reads with the last positive flow occurring before the *shortSigLastPosFlowThresh*. A positive flow is defined as a normalized nucleotide incorporation signal (signal divided by signal per base) greater than 0.6. The count of discarded reads is included in the *numDotFailed* (not *numTrimmedTooShortPrimer*) metric, along with those discarded by the Dot Filter.
3. **Dot Filter (doDotCheck)** – rejects reads with too high a proportion of negative flows (flows with no nucleotide incorporation). The Dot Filter discards reads with more than the *dotFlowFractionCutoff* percentage of flows before the last positive flow occurring as dots. A 'dot' is a block of negative flows that is ended by a positive flow of one of the nucleotides in the block, or started and ended by positive flows of the same nucleotide. The count of discarded reads is included in the *numDotFailed* metric, along with those discarded by the Primer-Dimer Filter.
4. **Mixed Filter (doMixedCheck)** – rejects reads with too high a proportion of positive flows (flows with a nucleotide incorporation signal greater than 0.6). This can occur from a bead with two or more attached DNA fragments, a well containing more than one DNA bead, or signal contamination from a neighboring well. The Mixed Filter discards reads if more than the *mixedPositiveFraction* of flows occurring before the last positive flow have a positive signal. The count of discarded reads is included in the *numMixedFailed* metric, along with those discarded by the Ambiguous Read Filter.
5. **Ambiguous Read Filter (doClassifierCheck)** – rejects reads with too high a proportion of ambiguous flows at the beginning of the read. The Ambiguous Read Filter examines the first 168 flows, and discards reads with more than 20% ambiguous or fewer than 30% unambiguous positive or fewer than 30% unambiguous negative flows. A flow is considered ambiguous if the normalized flow signal value falls within a sliding range of between 0.35 & 0.75 for the first flow to between 0.475 & 0.625 for flow 168. The count of discarded reads is included in the *numMixedFailed* metric, along with those discarded by the Mixed Filter.

Read rejecting filter parameters are described in Table 9.

Read Rejecting Parameter	Description	Values	Shotgun/PE	Amplicon
doShortSignalCheck	Enable/disable the Primer-Dimer Filter.	true/false	All: true	All: true
shortSigLastPosFlowThresh	Minimum acceptable read length for the Primer-Dimer Filter.	# flows	All: 84	GS Junior: 84 XLR70: 84 XL+ cyclic: 84 LongAmp #1: 152 LongAmp #2: 152 LongAmp #3: 152
doDotCheck	Enable/disable the Dot Filter.	true/false	All: true	All: true
dotFlowFractionCutoff	Cutoff % for Dot Filter.	percentage	All: 5	GS Junior: 5 XLR70: 5 XL+ cyclic: 5 LongAmp #1: 2 LongAmp #2: 2 LongAmp #3: 2
doMixedCheck	Enable/disable Mixed Filter.	true/false	All: true	All: true
mixedPositiveFraction	Cutoff % for Mixed Filter.	percentage	All: 70	GS Junior: 70 XLR70: 70 XL+ cyclic: 70 LongAmp #1: 50 LongAmp #2: 50 LongAmp #3: 50
doClassiferCheck	Enable/disable the Ambiguous Signal Filter.	true/false	All: true	All: true

Table 9: Read Rejecting Filter Parameters. A 'false' parameter disables the filter, which will increase the number of reads, including low quality reads (not recommended). Filter stringency can be increased by using larger green values or lower red values. 'LongAmp' stands for 'long amplicon' processing with XL+ kit, flow pattern B.

1.3.2.2 Read Trimming/Rejecting Filters

Several chemical- and system-related effects can gradually degrade the signal over the course of a sequencing run, resulting in lower quality signal towards the 3' end. After read rejecting filters have been applied, low quality reads are trimmed from the 3' end until they either (a) exceed the filter quality threshold or (b) are discarded as too short. See Section 1.3.4 for a description of minimum read length parameters.

Read quality information can be visualized in a Signal Flowgram (see example in Figure 3). The x-axis corresponds to the individual nucleotide flows of the sequencing run and the y-axis is a measure of signal intensity. The processed signal flowgrams have the following characteristics;

- Each nucleotide type is plotted in a different color allowing for easy 'reading' of the sequence trace data from visual inspection of the flowgram.
- Nucleotide incorporations are normalized such that a signal intensity near 1.0 corresponds to a single nucleotide, a signal intensity near 2.0 corresponds to a dinucleotide homopolymer, *etc.*

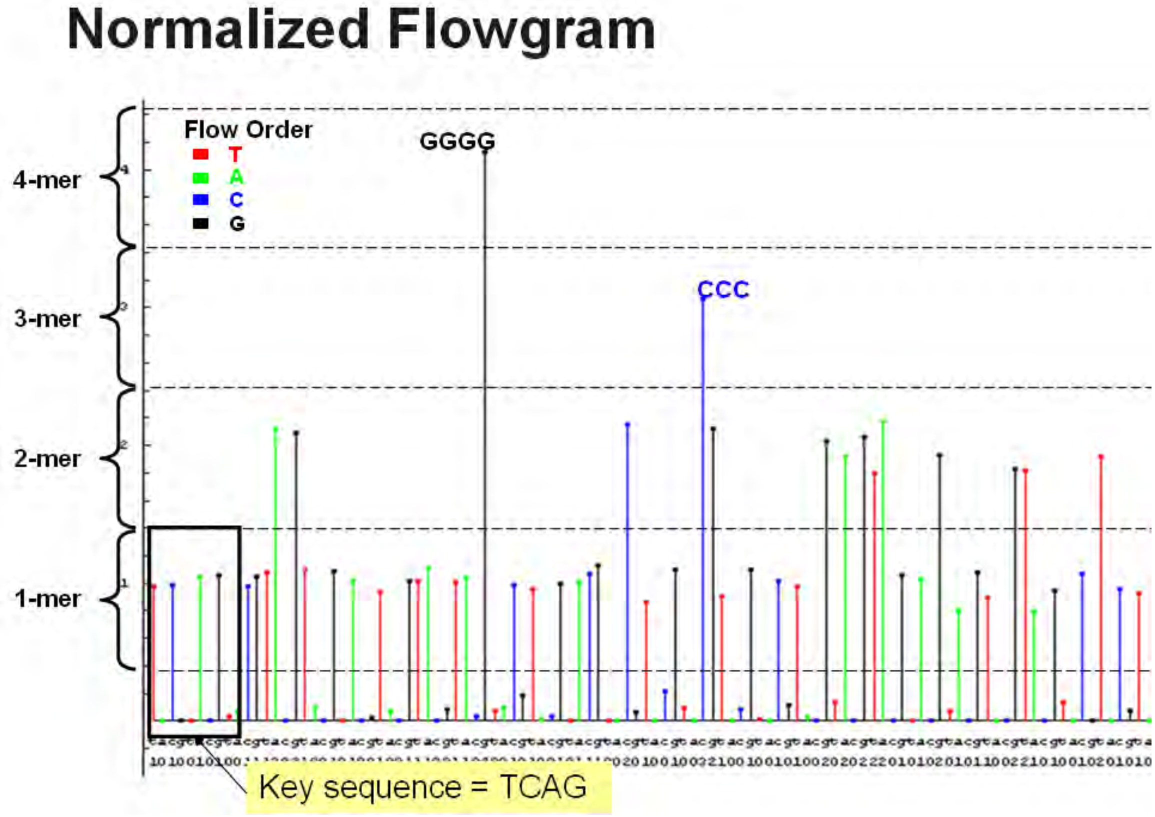


Figure 3: Normalized signal flowgram.

Read trimming/rejecting filters are applied in the following order.

1. **Signal Intensity Filter (filterToUse)** – scans forward from the first flow, and moves the trimpoint if the overall proportion of ambiguous flows remains below a ceiling value of *maxBadPercent* (3%). Flows are considered ambiguous if the normalized signal intensity falls in the range between 0.5 & 0.7. Reads with a final trimpoint less than *minLength* (83 flows) are discarded, with counts included in the *numTrimmedTooShortQuality* metric,
2. **Signal Trimback Filter (doTrimback)** – trims backwards from the last positive flow while ensuring that the tip of the 3' end of the read has a relatively low proportion of ambiguous or negative 'bad events', possibly caused by signal droop and/or CAFIE error accumulation. Flows are considered ambiguous if the normalized signal intensity falls in the range between 0.5 & 0.7, similar to the Signal Intensity Filter. Bad events are the sum of all ambiguous flows plus blocks of negative flows consisting of all four nucleotides with no intervening positive flows. Three increasingly stringent combinations of flow window and bad event threshold are evaluated, with the last combination (window = 2 flows, threshold = 0 bad events) ensuring that the trimpoint is set after an unambiguous positive flow. Reads trimmed shorter than *minLength* (83 flows) are discarded, with counts included in the *numTrimmedTooShortQuality* metric,

Signal Intensity and Signal Trimback Filter parameters are described in Table 10.

Signal Intensity/Trimback Parameter	Description	Values	Shotgun/PE	Amplicon
filterToUse	Enable/disable the Signal Intensity Filter.	leifilter/false	All: leifilter	All: leifilter
maxBadPercent	Ceiling value for the Signal Intensity Filter.	percent	All: 3%	All: 3%
doTrimback	Enable/disable Signal Trimback Filter.	true/false	All: true	All: true
minLength	Minimum acceptable read length for the Signal Intensity or Signal Trimback Filters.	# flows	All: 83	All: 83

Table 10: Signal Intensity/Trimback Filter Parameters. A 'false' parameter disables the filter, which may increase the length (and possibly the number) of reads, but will likely reduce overall read quality. Filter stringency can be increased by using larger green values or lower red values.

3. **Adaptor Trim Filter (doPrimerTrimming)** – trims 454 Sequencing system 3'-adaptor sequences (Table 11). Sequence matching occurs in flow space, and sequences with an average difference of less than 0.5 in the nucleotide incorporation signal compared to an idealized adaptor flowgram are trimmed. Compare with the adaptor match functionality in the Basecaller (Section 1.3.3.3). Reads trimmed shorter than *minLength* (83 flows) are discarded, with counts included in the *numTrimmedTooShortPrimer* metric,

3'-Adaptor Name	Adaptor Sequence, Capture Oligo Reverse Complement	Length
Lib-L, General	5′-CTGAGACTGCCAAGGCACACAGGGGATAGG-3′	30 bases
Lib-L, Rapid	5′-AGTCGTGGGAGGCAAGGCACACAGGGGATAGG-3′	32 bases
Lib-L, Rapid MID	5′-GGTCGGCGTCTCTCAAGGCACACAGGGGATAGG-3′	33 bases
Lib-A, Forward	5′-CTGATGGCGCGAGGGAGGCGATACG-3′	25 bases
Lib-A, Reverse	5′-CTGAGCGGGCTGGCAAGGCGCATAG-3′	25 bases

Table 11: Adaptor sequences trimmed by the Adaptor Trim filter. The underlined sequence in each adaptor is complementary to the capture oligo that is covalently attached to the sequencing bead, and should therefore lie at the far 3' end of the read.

Adaptor Trim Filter parameters are described in Table 12.

Adaptor Trim Filter Parameter	Description	Values	Shotgun/PE	Amplicon
doPrimerTrimming	Enable/disable Adaptor Trim Filter.	true/false	All: true	All: true
minLength	Minimum acceptable read length for Adaptor Trim Filter.	# flows	All: 83	All: 83
minPrimerMatchScore	Adaptor sequence match score, as a fraction of a normalized nucleotide incorporation.	fraction	All: 0.5	All: 0.5
useAmpliconPrimers	Use only sequencing adaptors appropriate for amplicons experiments, or specify each adaptor/primer explicitly inside a <primer></primer> element. *e.g.*: <primers> <primer>...</primer> <primer>...</primer> </primers>	true/false	All: false	All: true

Table 12: Adaptor Trim Filter Parameters. A 'false' parameter disables the filter, which leaves 3'-adaptor sequences attached to each read. Filter stringency can be increased by using larger green values or lower red values.

4. **Valley Filter (doValleyFilter)** – Filters or trims reads with many off-peak signal intensities. A valley flow is defined as an intermediate signal intensity, *i.e.*, a signal intensity occurring in the valley between the peaks for 0-mer and 1-mer incorporations, 1-mer and 2-mer incorporations or between 2-mer and 3-mer incorporations. The signal distribution of all reads of the run is used to define the peaks of the homopolymer incorporations relative to the valleys. One of the following three valley filters is applied when *doValleyFilter* = 'true'.

   - **Counting Valley Filter [read rejecting]** is applied when both *vfScanAllFlows* & *doValleyFilterTrimBack* are 'false', but has been superseded by the Scoring and Trimback Valley Filters. The Counting Valley Filter evaluates each flow up to a limit specified by *vfLastFlowToTest* (default = 320 flows). For each read, the number of borderline valley flows is compared to a threshold specified by *vfBadFlowThreshold* (default = 4 borderline flows). Reads that exceed the threshold are discarded, with counts included in the *numTrimmedTooShortQuality* metric.
   - **Scoring Valley Filter (vfScanAllFlows) [read rejecting]** is applied when *doValleyFilterTrimBack* is 'false' & *vfScanAllFlows* is 'true' or 'tiOnly', and is the default for amplicon processing. The Scoring Valley Filter calculates a valley score for each flow up to a limit specified by *vfScanLimit*. For each read, the average flow valley score is scaled by a factor (*vfTrimBackScaleFactor*) and compared to a threshold ratio calculated from *vfBadFlowThreshold* and *vfLastFlowToTest* (default = 4 bad flows per 320 flows). Reads with a scaled, average score that exceeds the threshold are discarded, with counts included in the *numTrimmedTooShortQuality* metric.
   - **Trimback Valley Filter (doValleyFilterTrimBack) [read trimming/rejecting]** is applied when *doValleyFilterTrimBack* is 'true', and is the default for shotgun processing and for two of the long amplicons pipelines. The Trimback Valley Filter utilizes the same mechanism as the Scoring Valley Filter, but with the additional feature that it scans reads backwards from the last flow of the read (specified by *vfScanLimit* = '4096') and trims flows until the scaled, average valley score for the read no longer exceeds the scoring valley filter threshold. This trimming is used to retain the higher quality portion of a read rather than discarding the entire read. Reads trimmed shorter than the larger of *minLength* or *qualityMinLength* (flow pattern B only) are discarded, with counts included in the *numTrimmedTooShortQuality* metric.

Counting Valley Filter [Default Value(s)]
Score = Count of borderline valley flows
Threshold = vfBadFlowThreshold [4 flows]
Last Flow Evaluated = vfLastFlowToTest [320 flows]

Scoring Valley Filter [Default Value(s)]
Score = Average (Flow Valley Scores) * vfTrimBackScaleFactor
Threshold = vfBadFlowThreshold / vfLastFlowToTest [4 / 320 = 0.0125]
Last Flow Evaluated = vfScanLimit [700 flows for amplicons, 1200 flows for long amplicons #1]

Trimback Valley Filter [Default Value(s)]
Score = Average (Flow Valley Scores) * vfTrimBackScaleFactor
Threshold = vfBadFlowThreshold / vfLastFlowToTest [4 / 320 = 0.0125]
Last Flow Evaluated = vfScanLimit [4096 (*i.e.* all flows) for shotgun, long amplicons #2 & #3]

Valley Filter parameters are described in Table 13.

Valley Filter Parameter	Description	Values	Shotgun/PE	Amplicon
doValleyFilter	Enable/disable all Valley Filters.	true/false	All: true	All: true
doValleyFilterTrimBack	Enable/disable the read trimming/rejecting Trimback Valley Filter (default shotgun processing).	true/false	All: true	GS Junior: false XLR70: false XL+ cyclic: false LongAmp #1: false LongAmp #2: true LongAmp #3: true
vfScanAllFlows	Enable/disable the read rejecting Scoring Valley Filter (default amplicon processing).	true/false tiOnly/flxOnly	All: false	GS Junior: tiOnly XLR70: tiOnly XL+ cyclic: tiOnly LongAmp #1: true LongAmp #2: true LongAmp #3: false
vfTrimBackScaleFactor	Scale factor (stringency) of the Scoring and Trimback Valley Filters.	arbitrary number	GS Junior: 1.6 XLR70: 0.7 XL+ cyclic: 2.2 XL+ acyclic: 2.2	GS Junior: 3.0 XLR70: 2.0 XL+ cyclic: 2.0 LongAmp #1: 3.0 LongAmp #2: 4.0 LongAmp #3: 2.0
vfBadFlowThreshold	The numerator of the threshold ratio used by the Trimback and Scoring Valley Filters.	# flows 2-6	All: 4	GS Junior: 4 XLR70: 4 XL+ cyclic: 4 LongAmp #1: 3 LongAmp #2: 3 LongAmp #3: 4
vfLastFlowToTest	The denominator of the threshold ratio used by the Trimback and Scoring Valley Filters.	flow # 168-400	All: 320	All: 320
vfScanLimit	The last flow of the scan window used by the Scoring and Trimback Valley Filters (4096 evaluates all flows).	flow #	All: 4096	GS Junior: 700 XLR70: 700 XL+ cyclic: 700 LongAmp #1: 1200 LongAmp #2: 4096 LongAmp #3: 4096
qualityMinLength	Minimum acceptable read length for doValleyFilterTrimBack with acyclic flow pattern (-1 disables).	# flows	GS Junior: -1 XLR70: -1 XL+ cyclic: -1 XL+ acyclic: 112	GS Junior: -1 XLR70: -1 XL+ cyclic: -1 LongAmp #1: -1 LongAmp #2: 700 LongAmp #3: 700
vfTrimBackMinimumLength	Minimum acceptable read length for doValleyFilterTrimBack with cyclic flow pattern.	# flows	All: 84	**All: 84**

Valley Filter Parameter	Description	Values	Shotgun/PE	Amplicon
vfMaxFailedPercent	The minimum percentage of ‘good’ flows required before a read is retained by the Trimback Valley Filter (100 disables this error check).	percentage 90-100	All: 100	All: 100
vfUseRollingWindows	Enable/disable a dynamic estimate of the valley thresholds between the 0, 1, 2, and 3-mers across the run.	true/false	All: true	All: true

Table 13: Valley Filter Parameters. A ‘false’ parameter disables the filter, which may increase the length (and possibly the number) of reads, but will likely reduce overall read quality. Filter stringency can be increased by using larger green values or lower red values. ‘LongAmp’ stands for ‘long amplicon’ processing with XL+ kit, flow pattern B.

5. **Basecall Quality Score Filter (doQScoreTrim)** – Technically, this filter is executed as part of basecalling (Section 1.3.3), but functionally it belongs with the other quality filters. The Basecall Quality Score Filter trims back from the 3' end of reads based on estimated (not final) quality scores derived from an internal calibrated signal histogram. The error rate in a sliding window is calculated from the quality scores, and multiplied by an empirical scaling factor, *QScoreTrimBackScaleFactor*. The window is slid leftwards until the estimated error rate in the window is less than *errorQScoreWIndowTrim*. Reads trimmed shorter than the larger of *QScoreTrimMinLength* (converted to flow space) or *qualityMinLength* (flow pattern B only) are discarded, but not counted in the *numTrimmedTooShortQuality* metric.

Basecall Quality Score Trimming Filter parameters are described in Table 14.

Quality Score Parameter	Description	Values	Shotgun/PE	Amplicon
doQScoreTrim	Enable/disable Basecall Quality Score Filter.	true/false	All: true	GS Junior: false XLR70: false XL+ cyclic: false LongAmp #1: false LongAmp #2: true LongAmp #3: true
QScoreTrimBackScaleFactor	Scale factor (stringency) of the Basecall Quality Score Filter.	arbitrary number	GS Junior: 1.2 XLR70: 0.9 XL+ cyclic: 1.5 XL+ acyclic: 1.5	GS Junior: 0.9 XLR70: 0.9 XL+ cyclic: 0.9 LongAmp #1: 0.9 LongAmp #2: 5.0 LongAmp #3: 1.5
QScoreTrimNukeWindowSize	Trimming window size.	# bp	All: 40	GS Junior: 40 XLR70: 40 XL+ cyclic: 40 LongAmp #1: 20 LongAmp #2: 30 LongAmp #3: 40
errorQscoreWindowTrim	Threshold of Basecall Quality Score Filter.	fraction	All: 0.010	All: 0.010
carryOverPeaks	Apply flowgram signal distribution statistics.	true/false	All: false	All: true
flowRadius	Compute smoothed signal distributions.	# flows	All: 16	All: 32
useCorrectionGlobalLimit	Enable/disable limit on flowgram correction.	true/false	All: false	All: true
qualityMinLength	Minimum acceptable read length for Basecall Quality Score Filter with acyclic flow pattern (-1 disables).	# flows	GS Junior: -1 XLR70: -1 XL+ cyclic: -1 XL+ acyclic: 112	GS Junior: -1 XLR70: -1 XL+ cyclic: -1 LongAmp #1: -1 LongAmp #2: 700 LongAmp #3: 700
QScoreTrimMinLength	Minimum acceptable read length beyond key.	# bp	All: 40	All: 40

Table 14: Basecall Quality Score Filter Parameters. A 'false' parameter disables the filter. Filter stringency can be increased by using larger green values or lower red values. 'LongAmp' stands for 'long amplicon' processing with XL+ kit, flow pattern B. Gray settings are ignored under default conditions.

1.3.3 Basecaller

The final processing stage of the signal processing pipeline is the basecaller, which generates final called sequences with corresponding quality scores, outputs results to CWF and SFF files, and generates a variety of metrics.

1.3.3.1 Cluster Builder

The cluster builder (amplicon processing only) groups reads that appear similar in flow space together, and calls bases on these groups as opposed to the entire read set (Table 15).

Cluster Builder Parameter	Description	Values	Shotgun/PE	Amplicon
nukeSpanStart	Define the start flow for the cluster builder algorithm.	# flows	All: 50	All: 50
nukeSpanEnd	Define the last flow for the cluster builder algorithm.	# flows	All: 600	GS Junior: 600 XLR70: 600 XL+ cyclic: 600 LongAmp #1: 1200 LongAmp #2: 1200 LongAmp #3: 600

Table 15: Cluster Builder Filter Parameters. LAmp refers to the three long amplicons pipelines, available only with the XL+ sequencing kit using flow pattern B. Gray settings are ignored under default conditions.

1.3.3.2 Sequence Composition Tool

The sequence composition tool summarizes the most common sequences observed during a sequencing run. It is useful for identifying both expected and contaminating sequences, particularly in amplicon or multiplex mapping projects. The ten most common 50 base sequences are output to the metrics.xml within the .cwf file, along with the length of the matching substring and the percent identity. Only matches across at least 50% of the sequence (25 bases) with an identity of 80% or greater will be reported.

1.3.3.3 Adaptor Match Tool

The adaptor match tool conducts a series of searches for sequences that may have originated from the 3'-adaptors used during library preparation, and reports them on the Adaptor Match tab of GS Run Browser (Section 3.10). Note that this tool is completely distinct from the adaptor trimming functionality in the read quality portion of the pipeline (Section 1.3.2.2). Although the adaptor match tool searches for the same adaptor sequences as the ones that are trimmed (Table 11), there are differences in the two searches that may result in a different set of reads and/or adaptor sequences being reported *vs*. trimmed.

The primary difference is that the adaptor match tool operates in nucleotide space (comparing base sequences), whereas the adaptor trim filter operates in flow space (comparing flowgrams). Also, the adaptor match tool uses a more relaxed match criterion toward the 3'-end of a read (after 700 bases, or in the last 100 bases of the read). This allows improved matching in the region where the 3'-adaptor is expected to be located and where read quality is lower.

> The adaptor match tool searches the entire read sequence, and will locate putative adaptor sequences regardless of whether or not they were trimmed by the adaptor trimming pipeline step, or by any other quality-based trimming.

1.3.4 Minimum Retained Read Length

Reads that are trimmed beyond a certain point are discarded. The minimum length parameters that control this minimum retained read length (Table 10) are generally defined in flowspace. A value of ~1.5 flows per base (cyclic flow pattern) or ~1.8 flows per base (acyclic flow pattern) may be used to roughly estimate the equivalent minimum retained read length in nucleotide space. For example, with a cyclic flow pattern run, *minLength* = 84 flows will result in a minimum read length of roughly 56 bases, including the four base sequencing key. With an acyclic flow pattern run, *qualityMinLength* = 700 flows will result in a minimum read length of roughly 390 bases.

> The average number of flows per base observed for a run will vary based on a number of conditions, including flow pattern, GC content, and homopolymer length. Relative to a genome with a balanced nucleotide composition, the number of flows per base for high or low GC content genomes might easily be 10% lower. Flows per base can also vary substantially from one amplicon sequence to another.
>
> When choosing an initial value for customizing one of the read length parameters or vfScanLimit, use the 1.5 (cyclic) or 1.8 (acyclic) flows per base as a starting point, but then adjust empirically. For high- or low-GC genomes, reduce the calculated number of flows by about 10%.
>
> **Example (high-GC content, acyclic flow pattern):** 700 bases * 1.8 flows per base * 0.9 = ~1134 flows

Filter Name	Minimum Length Parameter	Default Value	Discarded Read Metric
Primer-Dimer Filter	shortSigLastPosFlowThresh	84 - 152 flows	numDotFailed
Signal Intensity Filter	minLength	83 flows	numTrimmedTooShortQuality
Signal Trimback Filter	minLength	83 flows	numTrimmedTooShortQuality
Adaptor Trim Filter	minLength	83 flows	numTrimmedTooShortPrimer
Trimback Valley Filter	vfTrimBackMinimumLength qualityMinLength*	84 flows 112 - 700 flows	numTrimmedTooShortQuality
Basecall Quality Score Filter	QScoreMinLength qualityMinLength*	40 bases beyond key 112 - 700 flows	(n/a)

Table 16: Quality Filter Minimum Read Lengths. *Used with XL+ kit with flow pattern B only. When two minimum length parameters are active simultaneously, *e.g.* with flow pattern B runs, the final minimum read length will be controlled by the larger of the two values (in flow space), which will generally be qualityMinLength.

1.3.5 Pipeline Stringency

The standard pipeline choices provide a wide range of processing options that vary in stringency (Table 17).

Pipeline	Sequence Diversity	Sequencing Directionality	Valley Filter Type	Stringency	Example Amplicon Designs*
Shotgun/PE	high	unidirectional	trimming	low	(n/a)
Amplicons	low	bidirectional	rejecting	low	All
Long Amplicons #1	low	bidirectional	rejecting	high	Basic Universal Tailed
Long Amplicons #2	low	either	trimming	moderate	Long Range PCR
Long Amplicons #3	moderate	unidirectional	trimming	low	Ligated Adaptors One-Way Reads

Table 17: Pipeline choice, based on sample type and pipeline characteristics. *For more information regarding amplicon experimental design, refer to the research application guide entitled *454 Sequencing System Guidelines for Amplicon Experimental Design*, available at www.454.com/my454.

Each built-in pipeline was designed for a particular combination of sample characteristics (Table 17). Below are descriptions to help guide selection of an appropriate default pipeline, but only empirical reprocessing with other pipelines can identify the best pipeline for use with a particular sample.

The Shotgun/PE pipeline assumes a unidirectional sequencing application with randomly fragmented samples, which will yield a high signal variation within individual flows. Reads are trimmed, perhaps extensively, to minimize ambiguous (off-peak or valley) flows.

The standard Amplicons pipeline assumes a full-length, bidirectional sequencing application with low sequence diversity, which will yield a low signal variation within individual flows. Reads are discarded (rather than being trimmed) to eliminate reads with too many ambiguous (off-peak or valley) flows.

The Long Amplicons #1 pipeline is similar to the standard Amplicons pipeline, but with higher stringency. It will yield the lowest throughput, but will enforce the highest quality filtering.

The Long Amplicons #2 pipeline is similar to the Long Amplicons #1 pipeline, but is more suitable for (and will yield higher throughput for) particularly long amplicons or samples with content that is difficult to sequence. Although reads are trimmed, they are still more likely to be discarded to maintain quality, compared with the Shotgun/PE pipeline.

The Long Amplicons #3 pipeline has the greatest similarity to the Shotgun/PE pipeline of any of the amplicon pipelines. It is the most suitable for samples where bidirectional coverage is not required and reads can be trimmed with a relatively low stringency, for example for 16S metagenomics data. This pipeline will yield the greatest throughput of the amplicons pipelines, with the tradeoff that quality filtering will be somewhat less robust.

Note that all reads for amplicons shorter than the default qualityMinLength value of 700 flows (~390 bases) will be discarded when processed with valley filter trimming enabled (as in long amplicons #2 and long amplicons #3).

Table 18 provides a more detailed summary of the most significant valley and quality score functional distinctions across the five built-in pipelines. See Section 1.3.2.2 for a more complete description of the individual parameters.

Pipeline	Valley Filter[1]	Window Size[2]	Minimum Read Length[3]	Valley Stringency[4]	QScore Stringency[5]
Shotgun/PE	Trimback	4096 flows	GS Junior: 84 XLR70: 84 XL+ cyclic: 84 XL+ acyclic: 112	GS Junior: 1.6 XLR70: 0.7 XL+ cyclic: 2.2 XL+ acyclic: 2.2	GS Junior: 1.2 XLR70: 0.9 XL+ cyclic: 1.5 XL+ acyclic: 1.5
Amplicons	Scoring	700 flows	(n/a)	GS Junior: 3.0 XLR70: 2.0 XL+ cyclic: 2.0	GS Junior: 0.9 XLR70: 0.9 XL+ cyclic: 0.9
Long Amplicons #1	Scoring	1200 flows	(n/a)	XL+ acyclic: 4.0*	XL+ acyclic: 0.9
Long Amplicons #2	Trimback	4096 flows	XL+ acyclic: 700	XL+ acyclic: 5.3*	XL+ acyclic: 5.0
Long Amplicons #3	Trimback	4096 flows	XL+ acyclic: 700	XL+ acyclic: 2.0	XL+ acyclic: 1.5

Table 18: Significant Valley and Basecall Quality Score Filter differences across pipelines. Filter stringency can be increased by using larger green values or lower red values. Gray settings are ignored under default conditions.

[1] Enabled with doValleyFilterTrimBack = 'true' (Trimback Valley Filter) or with doValleyFilterTrimBack = 'false' (Scoring Valley Filter; vfScanAllFlows must also be 'true').

[2] Controlled by setting the value for vfScanLimit.

[3] Controlled by setting the value for qualityMinLength (for XL+ kit, flow pattern B, only) or vfTrimBackMinimumLength (shotgun pipeline, only). Applies to trimming pipelines, only.

[4] Controlled by setting the value of vfTrimBackScaleFactor . *These values in the table has been adjusted to allow comparison with other pipelines, which all use vfBadFlowThreshold = '4'.

[5] Controlled by setting the value for QScoreTrimBackScaleFactor (doQScoreTrim must also be 'true').

When customizing long amplicon processing pipelines, use doValleyFilterTrimBack and vfScanAllFlows to specify the type of valley filter to use, and concentrate on these four parameters. See Section 1.3.4 for additional details on 'flows per base' and Section 1.3.6 for creating custom 'filter-only' pipelines.

- [2] **vfScanLimit** (Scoring Valley filter, read rejecting) - To ensure high accuracy, set the window size to about the length of the smallest amplicon (using ~1.8 flows per base). Use a smaller window size if decreased accuracy at the 3' end is acceptable. This parameter should always be set to '4096' if trimming is allowed (*e.g.* with the Trimback Valley Filter).
- [3] **qualityMinLength** (Trimback Valley Filter & Basecall Quality Score Filter, read trimming) - Set the minimum trimmed read length to between 50% and 90% of the length of the smallest amplicon (using ~1.8 flows per base) in both the Valley Filter and the Basecall Quality Score Filter; *e.g.* between 675 and 1215 flows for a 750 base amplicon. Higher values will minimize trimming. This parameter should always be set to '-1' (disabled) if trimming is not allowed (*e.g.* with the Scoring Valley Filter).
- [4] **vfTrimBackScaleFactor** (Scoring or Trimback Valley Filter) - Increase for more stringent detection of ambiguous (off-peak or valley) flows, which will increase quality and decrease throughput (and *vice versa*).
- [5] **QScoreTrimBackScaleFactor** (Basecall Quality Score Filter) - Increase for increased trimming stringency in nucleotide space, based on quality scores. This parameter is ignored if trimming is not allowed (doQScoreTrim = 'false', *e.g.* with the Scoring Valley Filter).

1.3.6 Adjustable Pipeline Filter Parameters

Read quality filters may be turned off or adjusted to control the stringency of the output by editing an XML filter template file. A fragment of the default XML filter template file is shown in Figure 4.

```
<!-- Quality filter.-->
<qualityFilter>
  <!-- Enable or disable the trim back filter.  Normally, enabling
        this filter is recommended.
        default=true -->
  <doValleyFilterTrimBack>true</doValleyFilterTrimBack>

  <!-- This parameter controls the number of "bad" flows counted
        before the read is discarded.  Increase this number to
        increase the number of wells while reducing quality.
        default=4 -->
  <vfBadFlowThreshold>4</vfBadFlowThreshold>

  <!-- This is the last flow to test for the valley filter.  Set it
        to a larger number to make the filter more stringent.
        default=320 -->
  <vfLastFlowToTest>320</vfLastFlowToTest>

  <!-- Use this parameter to control the percentage of reads that
        can fail before the trimback filter discards the read.
        Setting it to 100 disables the trimback filter's error
        checking component (recommended).
        default=100 -->
  <vfMaxFailedPercent>100</vfMaxFailedPercent>

  <!-- Use this parameter to control the minimum number of good
        flows that must be present before the trim-back filter discards it.
        default=84 -->
  <vfTrimBackMinimumLength>84</vfTrimBackMinimumLength>

  <!-- This parameter controls the stringency of the
        valley filter.  A higher number rates ambiguous
        signals more strictly and more likely to throw
        out the read.

        default=0.7 for FLX, 1.6 for Junior, 2.2 for 400 cycle runs -->
  <vfTrimBackScaleFactor>0.7</vfTrimBackScaleFactor>
  <if test="starts-with(InstrumentModel,'GSJUNIOR_A')">
    <!-- This is for Junior runs only. -->
    <vfTrimBackScaleFactor>1.6</vfTrimBackScaleFactor>
  </if>
  <if test="Run.Flow.CycleCount &gt; 299">
    <vfTrimBackScaleFactor>2.2</vfTrimBackScaleFactor>
  </if>

  <!-- This parameter controls the last flow over
        which the valley filter scans.  For longer
        reads, lower quality signals at higher
        flows may penalize otherwise good data.  Set
        this limit to a number lower than the number of
        flows to enable this limit.

        default=4096 to disable -->
  <vfScanLimit>4096</vfScanLimit>

  <!-- This parameter controls the minimum
        flow length for passing reads that are trimmed
        for quality.  If the trimpoint is set by a quality
        filter, and is set shorter than this length, the
        whole read will be failed.
        default=84 -->
  <if test="Run.Flow.CycleCount &gt; 419">
    <qualityMinLength>112</qualityMinLength>
  </if>
</qualityFilter>
```

Figure 4: User-editable XML filter template file fragment, with default signal processing read quality filter settings used for a shotgun sequencing run (emphasis added).

This section provides instructions for customizing the XML filter-only template parameters for signal processing.

1. Open a command line terminal and change directory to the Data Processing folder ('D_') of the sequencing run whose reads are to be re-filtered.

   ```
   cd /Path/to/D_folder
   ```

2. Generate a template file in the D_ folder by typing one of the following two commands (the XML output template filename, to the right of the '>', can be any valid string ending with '.xml'):

   ```
   gsRunProcessor --template=filterOnly > myShotgunTemplate.xml
   gsRunProcessor --template=filterOnlyAmplicons > myAmpliconsTemplate.xml
   gsRunProcessor --template=filterOnlyLongAmplicons1 > myAmpliconsTemplate1.xml
   gsRunProcessor --template=filterOnlyLongAmplicons2 > myAmpliconsTemplate2.xml
   gsRunProcessor --template=filterOnlyLongAmplicons3 > myAmpliconsTemplate3.xml
   ```

3. Open the template file with a text editor, for example 'nedit' (on the GS FLX+ Instrument) or 'gedit' (on the GS Junior attendant PC). Type one of the following commands, specifying the template file you just created:

   ```
   nedit myShotgunTemplate.xml   or   gedit myShotgunTemplate.xml
   ```

4. Within the template, scroll down to the `<qualityFilter>` section (Figure 4) or `<baseCaller>` section (located just below `<qualityFilter>`). User-editable filter parameters that govern the selection of high quality reads are adjusted in these two sections. Edit filter parameters to either increase stringency (fewer, higher quality reads) or decrease stringency (more, lower quality reads). A short description and the default values for each filter in the template are shown in Table 13 (Valley Filter/Valley Trimback Filter Parameters) and in Table 14 (Basecall Quality Score Filter Parameters).
5. If desired, add other filter parameters and values. Use with care.
6. After the edits have been made, save and exit the text editor.
7. Change directory, to the parent run folder directory:

   ```
   cd ..
   ```

8. Launch the processing job, specifying the modified filter parameter script.

   ```
   runAnalysisFilter --pipe=/Path/to/myShotgunTemplate.xml /Path/to/D_folder
   ```

If the job is successfully launched, a new Data Processing directory with the name convention D_..myShotgunTemplate will be created in the run directory.

While it is possible to turn off the read rejecting filters that report as numDotFailed (doShortSignalCheck and doDotCheck) or numMixedFailed (doMixedCheck and doClassifierCheck), this will result in output of sub-standard quality data that is NOT recommended for use in subsequent data analysis. However, to turn off these filters for troubleshooting purposes, add the following lines under the <qualityFilter> section of the shotgun processing template.

```
<doShortSignalCheck>false</doShortSignalCheck>
<doDotCheck>false</doDotCheck>
<doMixedCheck>false</doMixedCheck>
<doClassifierCheck>false</doClassifierCheck>
```

1.3.7 Signal Processing with Acyclic Flow Pattern

1.3.7.1 Acyclic Flow Pattern and CAFIE Errors

Data acquisition using the GS FLX Titanium Sequencing Kit XL+ has the option of using one of two different flow patterns; the cyclic flow pattern A or the acyclic flow pattern B. One of the characteristics of the cyclic flow pattern is that once a particular nucleotide has flowed, each of the other three will flow before the first one flows again. A consequence of this behavior is that if a subset of DNA strands on a bead fails to fully incorporate, there is a 100% probability that the following base in the sequence will flow before the incompletely extended strands have a chance to 'catch up'. This subset of strands will continue to incorporate out-of-phase, which is known as an incomplete extension error (see Section 1.3.1, CAFIE errors).

With the acyclic flow pattern, any given nucleotide may flow repeatedly before all of the other nucleotides have had a chance to flow. If a second flow of the nucleotide occurs before a flow for the following base in the DNA sequence, the incomplete strands will complete their incorporation and be back in-phase ('catch up') with the remainder of the strands. This ability to catch up from an incomplete extension event depends on both the following flow(s) in the flow list and the following base in the DNA sequence; but the probability of catching up will always be as good as or better than with the cyclic flow pattern. Similarly, the other CAFIE error (carry forward error) can also be prevented or reversed by the non-uniform distribution of flows in an acyclic flow pattern.

Since CAFIE errors are one of the major limits on maximum read length, the acyclic flow patterns may yield longer overall trimmed read lengths, depending on how well the specific characteristics of the flow pattern match with the specific DNA sequences encountered.

1.3.7.2 Acyclic Flow Pattern and Raw (Untrimmed) Read Length

The positive effect of the reduction of CAFIE errors is counter-balanced by the possibility that a given read will get 'stuck' waiting for the next flow that matches the following base in the DNA sequence. With the four nucleotide cyclic flow pattern, a given nucleotide flow always occurs exactly four flows after the previous one, and a read never needs to wait for more than three flows to extend. With the acyclic flow pattern, the number of flows required to include at least one flow of each of the four nucleotides (defined as a flow set) will vary over the course of a run, and will average about 18% more than the exactly four flows required for the cyclic flow pattern.

Thus, the number of flows required to generate a given number of positive flows is higher for flow pattern B, again depending on how well the specific characteristics of the flow pattern match with the specific DNA sequences encountered. See Section 1.3.4 for a discussion of 'flows per base' and the relationship between read lengths in flow space *vs.* nucleotide space.

The balance between a shorter starting read length and longer reads resulting from reduced trimming is doubly dependent on the exact sequences encountered during sequencing. For most genomes, the net result is an increase in read lengths when using the acyclic flow pattern B compared to the cyclic flow pattern A. Results are expected to vary from genome to genome.

1.3.7.3 Acyclic Flow Pattern and Read Quality Filters

The predicted difference in flows per base between cyclic and acyclic flow patterns is also expected to lead to differences in the degree of trimming, and in the counts of reads discarded by various filters and reported as *numDotFailed*, *numMixedFailed*, and *numTrimmedTooShortQuality*. Two types of adjustments have been made to pipeline parameters; [a] altered cutoff values for the Dot and Mixed Filters (Section 1.3.2.1), and [b] the addition of a *qualityMinLength* parameter that controls the minimum length of acyclic flow pattern reads that have been trimmed by the Trimback Valley Filter or the Basecall Quality Score Filter (Sections 1.3.2.2 and 0). Even with these adjustments, somewhat different results are expected with flow pattern B because of the differences in the flow patterns themselves.

1.3.8 Launching the Signal Processing Pipeline

There are two commands that can be used to launch the signal processing step of the GS Run Processor application:

```
runAnalysisPipe [options] SOURCE_DIRECTORY

runAnalysisPipeAmplicons [options] SOURCE_DIRECTORY
```

The `Source_Directory` can be either an ‘R_’ directory to run both the image processing and signal processing steps on the data set, or a ‘D_’ directory to run only the signal processing step. If a run (‘R_’) directory is specified, the *.png (or legacy *.pif) files must be present in the rawImages subdirectory of the run directory. If a data (‘D_’) directory is specified, the directory must contain image processed results. If the data present in a ‘D_’ directory has also been signal processed, the runAnalysisPipe command will produce a warning message (note that this is only a warning; processing will still proceed).

Quality filter parameter settings can be customized by editing an XML filter template file (see Section 1.3.6).

The filter-only processing command is:

```
runAnalysisFilter --pipe=XMLFILE SOURCE_DIRECTORY
```

The signal processing and filtering command line arguments and output are described below:

Argument	Description
SOURCE_DIRECTORY	Path to an 'R_' (full processing) or a 'D_' directory (signal or filter processing).

Option	Description
--progress	Displays real-time progress of the job's processing.
--pipe=XMLFILE	Specifies a pipeline processing script file for signal processing using customized filter parameters (runAnalysisPipe) or to specify the input filter parameters for a filter-only signal processing job (runAnalysisFilter). (Specifying the .xml extension is optional.)
--reg=REGION_NUM	Only process data from a particular region of the PTP device. Regions can be specified as a range or as a comma-separated list ('1-4' or '1,2,5,9'). (Note: Multiple regions are only supported on the GS FLX+ Instrument.)
-verbose	Provides verbose log output useful for troubleshooting.

1.3.9 Signal Processing Output

The signal processing step creates an output data directory of the format:

D_...signalProcessing

The following files are generated by the signal processing and are used by the gsReporter to generate reports and output files for further data analysis;

Output	Description
region/regionnum.cwf	Composite Wells Format (CWF) files - Contains the corrected flowgram information, processing metrics from the data processing and low-resolution image for each base flow. One file per PTP region.
gsRunProcessor.log	Contains messages from the image processing.
gsRunProcessor_err.log	Contains error messages from the image processing.

With 454 Sequencing software applications version 2.01 or higher, while a signal processing job is in progress, the cwf files will have a .tmp extension. Users can verify if the pipeline has successfully finished by checking for .cwf files without .tmp suffixes in the regions sub-directory of the processing data (D_) directory.

2 GS REPORTER

The GS Reporter application is identical for both the GS Junior and GS FLX+ Instruments. However, due to differences in instrument hardware, some references in this manual will be specific to either the GS Junior or the GS FLX+ Instruments. The main difference is the PTP device.

The PTP device on the GS FLX+ Instrument supports division into multiple (2, 4, 8 or 16) regions. The GS Junior Instrument PTP device only supports a single region. Therefore, any references to multiple regions are specific to the GS FLX+ Instrument.

The GS Reporter is an application that can extract read trace information and run metrics from the CWF files produced by the GS Run Processor. The GS Reporter application is invoked by the signal processing or full processing jobs and is called automatically by the signal processing launch scripts to produce default files and reports.

2.1 gsReporter Executable

The gsReporter executable generates various human-readable files which can be used to examine the results of a sequencing run. These include files pertaining to read data, FASTA files (*.fna), associated base quality score files (*.qual), and legacy files (*.wells, *etc.*). The gsReporter also generates files containing the run metrics in text (*.txt) and comma-separated value (*.csv) files.

```
<gsReporter>
      <enable>true</enable>
      <options>
          --454RuntimeMetricsAll.txt --454RuntimeMetricsAll.csv
          --454AllControlMetrics.txt --454AllControlMetrics.csv
          --454BaseCallerMetrics.txt --454BaseCallerMetrics.csv
          --454QualityFilterMetrics.txt --454QualityFilterMetrics.csv
          --fna
          --qual
       </options>
</gsReporter>
```

Figure 5: gsReporter section of an XML processing script (options should appear on one line).

The default output files of the gsReporter are specified in the gsReporter section of the signal processing XML pipeline script (Figure 5) and are indicated in the options table below by the shaded rows. The default location of the output files can be redirected to alternate locations, using the --out option of the gsReporter command, or to standard output, using the --console option, for piping of the gsReporter command in a script. The gsReporter will attempt to generate all output files specified even if it does not have the required information. For example, generating a FASTA file on an image-processed only CWF file will result in an empty .fna file and a warning message will be sent to the console.

The gsReporter command line has the following form:

```
gsReporter [OPTIONS] SOURCE_FILES
```

Command	Description
gsReporter	An application that extracts read trace information and run metrics from CWF files to generate various human-readable files, including FASTA files (*.fna), associated base quality score files (*.qual), legacy files (*.wells, *etc.*; files generated from pre-2.0 software), and run metrics files in text (*.txt) and comma-separated value (*.csv) formats.

Argument	Description
SOURCE_FILES	List of CWF data files to be processed.

Option*	Description
--console	Redirects output to standard output. Used when piping the gsReporter command in a script.
--out	Redirects output to an alternate (non-default) output location By default, the GS Reporter application attempts to write its output files to the locations they would have been placed by earlier versions of the software.
--dump	Dumps all XML in the corrected CWF file as a single document. (binary)
--info	Shows summary information about the CWF file.
--legacy	Simulate the output of a GS Junior or GS FLX+ sequencing run processed with pre-2.0 software.
-a, --all	Generates all output files.
--fna	Generates one FNA file per library key per region for a corrected CWF file.
--qual	Generates one qual file per library key per region for a corrected CWF file.
--meta	Extracts the meta information about a region's data as an XML file.
--metrics	Extracts the metrics information about a region's data as an XML file.
--history	Extracts the job history as an XML file.
--primerInfo	Extracts the 3'-adaptor name and trimpoint for each read, as set by the adaptor trim pipeline step as a tab-delimited file.
--wells	Generates a .wells file per region.
--cfValues	Generates one cfValues file per region for a corrected CWF file. This is a binary file that contains information about the CAFIE correction data. This file is used by the gsSupportTool for troubleshooting runs. (Binary)
--454RuntimeMetricsAll.txt	Generates a 454RuntimeMetricsAll.txt file for the run.
--454RuntimeMetricsAll.csv	Generates a 454RuntimeMetricsAll.csv file for the run.
--454BaseCallerMetrics.txt	Generates a 454BaseCallerMetrics.txt file for the run.
--454BaseCallerMetrics.csv	Generates a 454BaseCallerMetrics.csv file for the run.
--454RuntimeMetrics.txt	Generates one 454RuntimeMetrics.txt per key per region.
--454RuntimeMetrics.csv	Generates one 454RuntimeMetrics.csv per key per region.

Option*	Description
--454QualityFilterMetrics.txt	Generates a 454QualityFilterMetrics.txt file for the run.
--454QualityFilterMetrics.csv	Generates a 454QualityFilterMetrics.csv file for the run.
--cafieMetrics	Generates one cafieMetrics.csv file per region. (binary)
--droopEstimate	Generates one droopEstimate file per region. (binary)
--trimInfo	Generates one trimInfo file per region.(binary not used)
--xy	Generates one text file containing a list of well locations, per region.
--csv	Generates one file with the flows written out as comma-separated values per region.
--analysisParms.parse	Generates an approximation of a pre-2.0 analysisParms.parse file.
--revisedRegions.parse	Generates a single revisedRegions.parse file.

The binary files are an intermediate product of processing and do not contain metric information. They are used by the pipeline processing job.

2.2 GS Reporter Output

The default output files of the gsReporter application are the following:

- 454RuntimeMetricsAll.txt/.csv
- region.key.454RuntimeMetrics.txt/.csv
- 454QualityFilterMetrics.txt/.csv
- 454BaseCallerMetrics.txt/.csv
- 454AllControlMetrics.txt/.csv
- region.librarykey.454Reads.fna
- region.librarykey.454Reads.qual

where '.txt/.csv' denotes that both tab-delimited (*.txt) and comma-delimited (*.csv) files are output, and 'region.key.' denotes that one file per region-key combination is output.

The .fna and .qual files for the Control DNA reads are no longer automatically generated by the default gsReporter options but can generated by using the following constructs:

```
gsReporter --fna=control 1.cwf, or gsReporter --qual=TCAG 1.cwf
```

The 454AllControlMetrics.txt file contains the summary statistics for the Control DNA metrics and is useful in situations in which multiple control keys are used in a run (see Section 7.1 for more details on this). It is generated by default, but is not generated when using the --legacy option to gsReporter.

2.2.1 GS Reporter Metrics Files

The gsReporter application metrics file contents are shown, in part, below. The full contents are described in the appendix Section 6.5 along with the conventions for date, directory and other specifications.

Examples of the Metrics file sections are shown in txt format, in the Figures below.

```
/***************************************************************************
**
**     454 Life Sciences Corporation
**
**     Software Release: 2.9
**
**     Runtime Metrics Results
**
**     Run Name:
       R_2012_08_16_15_43_37_build11_bonvinl1_AFO_7075_2rsk93867961ECv28IntEval
**     Analysis Name: D_2013_05_29_15_36_01_britecm1_fullProcessing
**     Region Name:   All
**     Key Sequence:  All
**
**     File Created:  2012/08/16 15:43:37
**
***************************************************************************/

softwareVersion
{
       softwareVersionTag = "2.9";
}
```

Figure 6: 454RunTimeMetricsAll.txt comment header section.

```
runConditions
{
       runName      =
       "R_2012_08_16_15_43_37_build11_bonvinl1_AFO_7075_2rsk93867961ECv28IntEval";
       analysisName = "D_2013_05_29_15_36_01_britecm1_fullProcessing";

       PTPBarCode   = "779126";

       numberOfRegions = 2;
       numberOfCycles  = 445;
}
```

Figure 7: 454RuntimeMetricAll.txt run conditions group section.

```
        key
        {
                keySequence               = ATGC;
                numKeyPass                = 15664;
                numDotFailed              = 0;
                numMixedFailed            = 14;
                numTrimmedTooShortQuality = 306;
                numTrimmedTooShortPrimer  = 0;
                totalPassedFiltering      = 15342;
        }
        key
        {
                keySequence               = CATG;
                numKeyPass                = 10862;
                numDotFailed              = 1;
                numMixedFailed            = 3;
                numTrimmedTooShortQuality = 734;
                numTrimmedTooShortPrimer  = 0;
                totalPassedFiltering      = 10115;
        }
        key
        {
                keySequence               = GACT;
                numKeyPass                = 953751;
                numDotFailed              = 1355;
                numMixedFailed            = 4788;
                numTrimmedTooShortQuality = 186899;
                numTrimmedTooShortPrimer  = 1576;
                totalPassedFiltering      = 751421;
        }
```

Figure 8: 454QualityFilterMetrics.txt key group section.

```
basecallResults
{
        numReads  = 1570584;
        aveLength = 752.910;
        stdDev    = 221.448;
}
```

Figure 9: 454BaseCallerMetrics.txt basecall results section.

```
regionKey
{
        region = 1;
        key    = ATG;

        numberReads    = 15342;
        totalBases     = 11003802;
        averageLength  = 717.234,162.795;
        averageQuality = 35.165,6.580;

        lengthHistogram
        {
                lengthCount = 47,1;
                lengthCount = 48,1;
                lengthCount = 50,1;
                ...
                lengthCount = 960,1;
                lengthCount = 968,1;
                lengthCount = 1014,1;
        }
        qualityHistogram
        {
                qualityBinCount = 0,529;
                qualityBinCount = 10,9825;
                qualityBinCount = 11,7321;
                ...
                qualityBinCount = 38,187666;
                qualityBinCount = 39,756531;
                qualityBinCount = 40,4741073;
        }
```

Figure 10: 454BaseCallerMetrics.txt Region Key Group.

2.2.2 GS Reporter FNA and QUAL Files

The GS Reporter application also generates FASTA files (*.fna) and associated base quality files (*.qual) for the high quality reads generated. These are described below.

Output	Description
region.librarykey.454Reads.fna	The (trimmed) nucleotide sequences for the filtered reads of that region and key.
region.librarykey.454Reads.qual	This file contains the nucleotide quality scores (Phred-equivalent) for the high quality (filtered and trimmed) reads of that region and key. For a description of the process to compute individual base quality scores, see appendix Section 6.6.

2.2.3 Organization of a Data Processing Directory

Once the data processing has been completed and metrics and report files have been generated, these data structure can be quite complex. Table 19 below can be used to locate specific files within their expected directory if default locations were used.

Organization of a Data Processing Directory ('D_')

File name	A	B	C	D	E	Notes
`gsRunProcessor.log`	✓	✓				
`gsRunProcessor_err.log`	✓	✓				1
`454DataProcessingDir.xml`	✓	✓				2
`region.KEYL.454Reads.fna`		✓	✓	✓		3,7
`region.KEYL.454Reads.qual`		✓	✓	✓		3,7
`region.KEYC.454Reads.fna`			✓	✓		3,6,7
`region.KEYC.454Reads.qual`			✓	✓		3,6,7
`454BaseCallerMetrics.csv`		✓	✓	✓		3,4,7
`454BaseCallerMetrics.txt`		✓	✓	✓		3,7
`454QualityFilterMetrics.csv`		✓	✓	✓		3,4,7
`454QualityFilterMetrics.txt`		✓	✓	✓		3,7
`454RuntimeMetricsAll.csv`		✓	✓	✓		3,4,7
`454RuntimeMetricsAll.txt`		✓	✓	✓		3,7
`454AllControlMetrics.txt`	✓	✓	✓			3,7
`454AllControlMetrics.txt`	✓	✓	✓			3,7
`analysisParms.parse`			✓	✓		3,5
`revisedRegions.parse`			✓	✓		4
`454BaseCallerThresholds.txt`					✓	
`error.baseCaller2`					✓	
`error.bbcSelfTrain`					✓	
`error.cafieCorrection`					✓	
`regions/`	✓	✓				
`region0X.cwf`	✓	✓				
`region0X.metrics.xml`			✓			
`region0X.meta.xml`			✓			
`region0X.processingHistory.xml`			✓			
`region0X.wells`			✓	✓	✓	
`region0X.wells.KEY.454RuntimeMetrics.csv`			✓	✓	✓	3,4,7

File name	A	B	C	D	E	Notes
region0X.wells.KEY.454RuntimeMetrics.txt			✓	✓	✓	3,7
region0X.wells.cafieMetrics.csv			✓	✓	✓	3,4,7
region0X.wells.cfValues			✓	✓	✓	3,7
region0X.wells.droopEstimate			✓	✓	✓	3,7
region0X.wells.incValues			✓	✓	✓	3,7
region0X.wells.mleCorrectionInfo			✓	✓	✓	3,7
region0X.wells.trimInfo			✓	✓	✓	3,7
sff/	✓	✓				
ACCNOPREX.sff	✓	✓				

Table 19: Organization, in the 'D_' directory, of the files output by the GS Run Processor and the GS Reporter applications. The following codes are used in the file names: 'region' is the region number, 'ØX' is the zero-padded region number, 'KEYL' is the 3 letter library sequencing key, 'KEYC' is the 3 letter library sequencing key, 'ACCNOPRE' is the accession number prefix.

A: Generated by gsRunProcessor during processing
B: Generated by default
C: Generated directly by gsReporter from CWF files
D: Generated by gsReporter's '-- legacy' option
E: No longer generated

Notes:

1 - This file is only generated if there are warnings or errors generated during processing.
2 - This file is only created after the basecalling step is run and signifies to the data analysis software that the sff directory contains files suitable for further data analysis.
3- The default generation of these files can be controlled by adjusting the gsReporter options in the pipeline processing scripts.
4 - These files are generated for legacy purposes only. It is recommended that new applications use X.metrics.xml and X.meta.xml extracted from the CWF file *via* gsRunProcessor for report generating purposes.
5 - This file is deprecated and generated for legacy purposes only. X.processingHistory.xml and X.meta.xml are the canonical record of parameters used to process a run and a run's metadata, respectively.
6 – These .fna and .qual files were generated by default in previous versions of the software.
7 - These files may only be generated after all processing is complete. Specifically, the data to create these files are not available after the image processing only step of gsRunProcessor.

3 GS RUN BROWSER

The GS Run Browser application is identical for both the GS Junior and GS FLX+ Instruments. However, due to differences in instrument hardware and software, some references in this manual will be specific to either the GS Junior or the GS FLX+ Instruments. The main difference is the PTP device layout.

Image data from the GS FLX+ Instrument are stored as compressed .png image files, but the legacy .pif image format from older runs is still supported. Image data from the GS Junior Instrument are stored as .pif image files. Therefore, references to .png are specific to runs from the GS FLX+ Instrument, while references to .pif can apply to either runs from the GS Junior Instrument or older runs from the GS FLX+ Instrument.

The PTP device on the GS FLX+ Instrument supports division into multiple (2, 4, 8 or 16) regions. The GS Junior Instrument PTP device only supports a single region. Therefore, any references to multiple regions are specific to the GS FLX+ Instrument.

The GS Run Browser application allows the user to interactively view and analyze the results of sequencing runs performed on a GS Junior or GS FLX+ Instrument, to assess the general quality of a run or for troubleshooting when results are sub-optimal. It also allows one to launch data processing of the raw run data (*via* gsRunProcessorManager), and to prepare a data package that can be sent to Roche Customer Support for troubleshooting (*via* gsSupportTool, described in Section 3.10). It is available on the GS FLX+ Instrument, on the GS Junior attendant PC, or a datarig.

The application has the following tabs:

- Overview: run summary information
- Wells: Raw images of the PTP device captured during the run; Locations and status of all identified active wells, well density information for raw wells and key pass wells, and CAFIE correction data
- Signals: Raw or subtracted flowgrams for selected positions on the images; or fully processed flowgrams for all the data-generating wells detected during the run
- Reads: Read length and quality statistics for the sample library or the Control DNA reads
- Control DNA: Consensus flowgrams and accuracy metrics for the Control DNA sequence reads
- Filters: Raw signal statistics for the sample library or the Control DNA reads

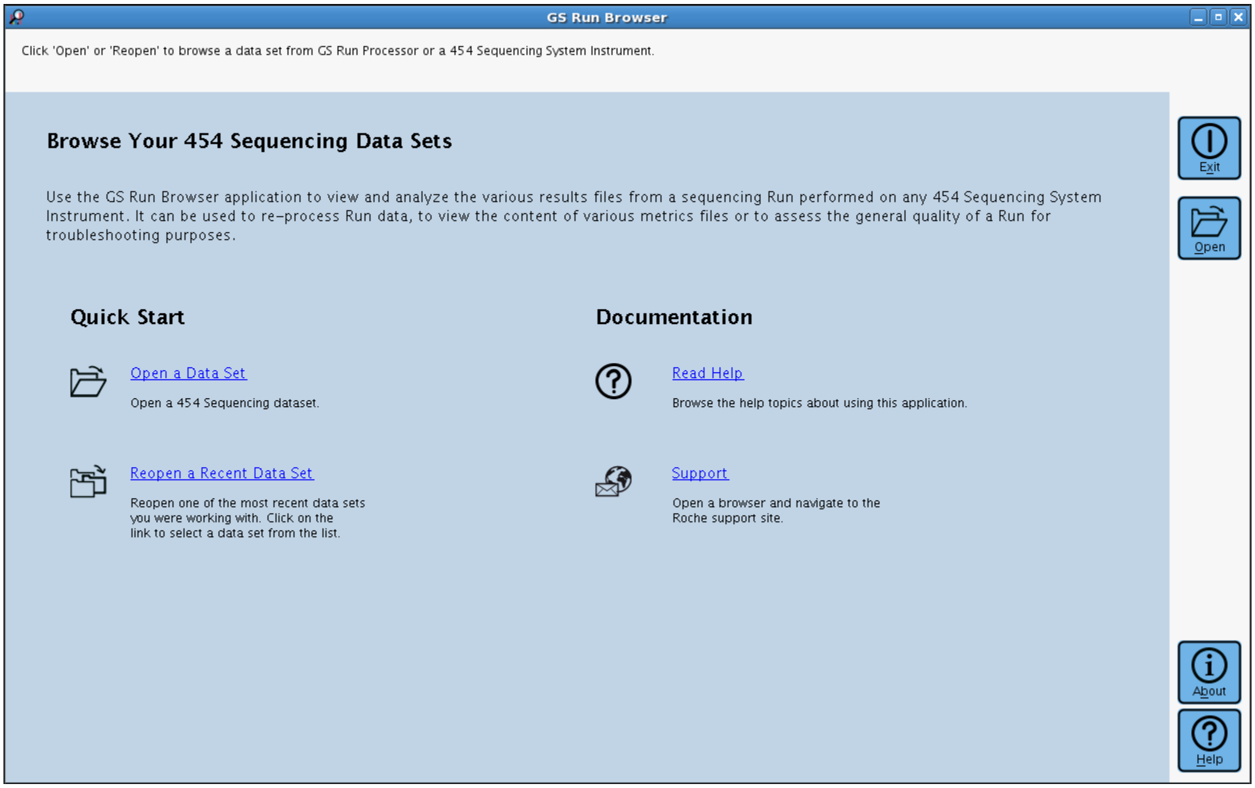
3.1 Launching the GS Run Browser

The GS Run Browser can be launched by double-clicking the desktop icon located on the GS Junior attendant PC desktop.

For all other hardware, including the GS FLX+ Instrument, open a terminal window and type the following command to launch the gsRunBrowser

```
gsRunBrowser
```

If the optional path to a data set is given on the command line, that data set will be loaded and shown in the application. Otherwise, an introduction window will be presented (Figure 11), ready to load a data set.



Click 'Open' or 'Reopen' to browse a data set from GS Run Processor or a 454 Sequencing System Instrument.

Browse Your 454 Sequencing Data Sets

Use the GS Run Browser application to view and analyze the various results files from a sequencing Run performed on any 454 Sequencing System Instrument. It can be used to re-process Run data, to view the content of various metrics files or to assess the general quality of a Run for troubleshooting purposes.

Quick Start

Open a Data Set
Open a 454 Sequencing dataset.

Reopen a Recent Data Set
Reopen one of the most recent data sets you were working with. Click on the link to select a data set from the list.

Documentation

Read Help
Browse the help topics about using this application.

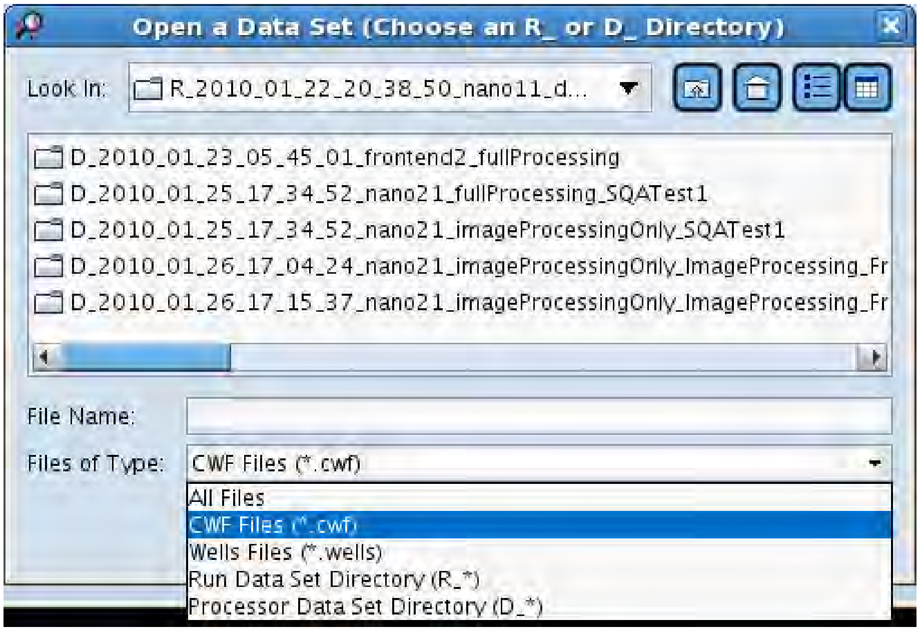
Support
Open a browser and navigate to the Roche support site.

Figure 11: The GS Run Browser main window, just after launching the application, when no data set directory is specified on the command line or launched from the desktop GS Run Browser icon.

3.2 Opening a Data Set

Clicking on the ‘Open’ button or on the ‘Open a Data Set’ text will open the ‘Open a Data Set’ dialog, prompting the user to select a data set to display (Figure 12). Either a run directory (R_...), a data processing directory (D_...), or individual cwf (.cwf) or wells (.wells) files may be selected:

- R_ directory– will load only the general run information and the captured Images, if available.
- D_ directory– will load the general run information, the captured images, and all the data processing results.
- .cwf files - Contain the uncorrected flowgrams from the image processing step or the corrected flowgrams from the signal processing step.
- .wells files – Legacy files generated by pre-version 2.0 software that contain well data.



Open a Data Set (Choose an R_ or D_ Directory)

Look In: R_2010_01_22_20_38_50_nano11_d...

D_2010_01_23_05_45_01_frontend2_fullProcessing
D_2010_01_25_17_34_52_nano21_fullProcessing_SQATest1
D_2010_01_25_17_34_52_nano21_imageProcessingOnly_SQATest1
D_2010_01_26_17_04_24_nano21_imageProcessingOnly_ImageProcessing_Fr
D_2010_01_26_17_15_37_nano21_imageProcessingOnly_ImageProcessing_Fr

File Name:

Files of Type: CWF Files (*.cwf)

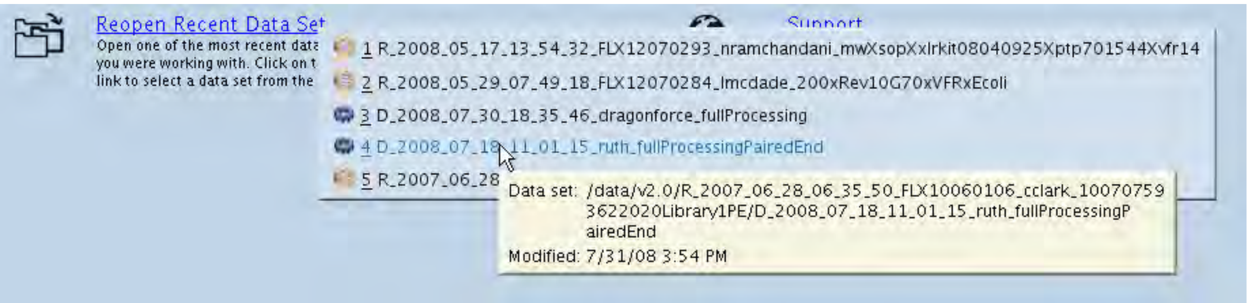
All Files
CWF Files (*.cwf)
Wells Files (*.wells)
Run Data Set Directory (R_*)
Processor Data Set Directory (D_*)

Figure 12: The ‘Open a Data Set’ dialog, used to select the data set.

The ‘Open’ button of the ‘Open a Data Set’ dialog is enabled only when a valid data set and the correct directory or file type has been selected. A progress bar will appear at the upper right corner of the GS Run Browser window while the data set is being loaded, showing the steps of the loading. If the data set fails to open, an error message window will appear (not shown). Alternatively, you can click-and-drag a proper data set from the OS file browser and drop it directly into the GS Run Browser window.

The application does not allow multiple .cwf files to be selected from different data processing directories. If you attempt to do so, the open button will be disabled and an error message will be displayed if you hover the mouse over the open button.

A list of recently opened data sets is always available. To access the list of previously opened data sets click the 'Reopen Recent Data Set' text from the introduction page, or click the 'Open' button while pressing the Shift key. The full path to a data set and the last time the data set was modified is displayed in a tooltip for reference (Figure 13).



Open one of the most recent data
you were working with. Click on t
link to select a data set from the

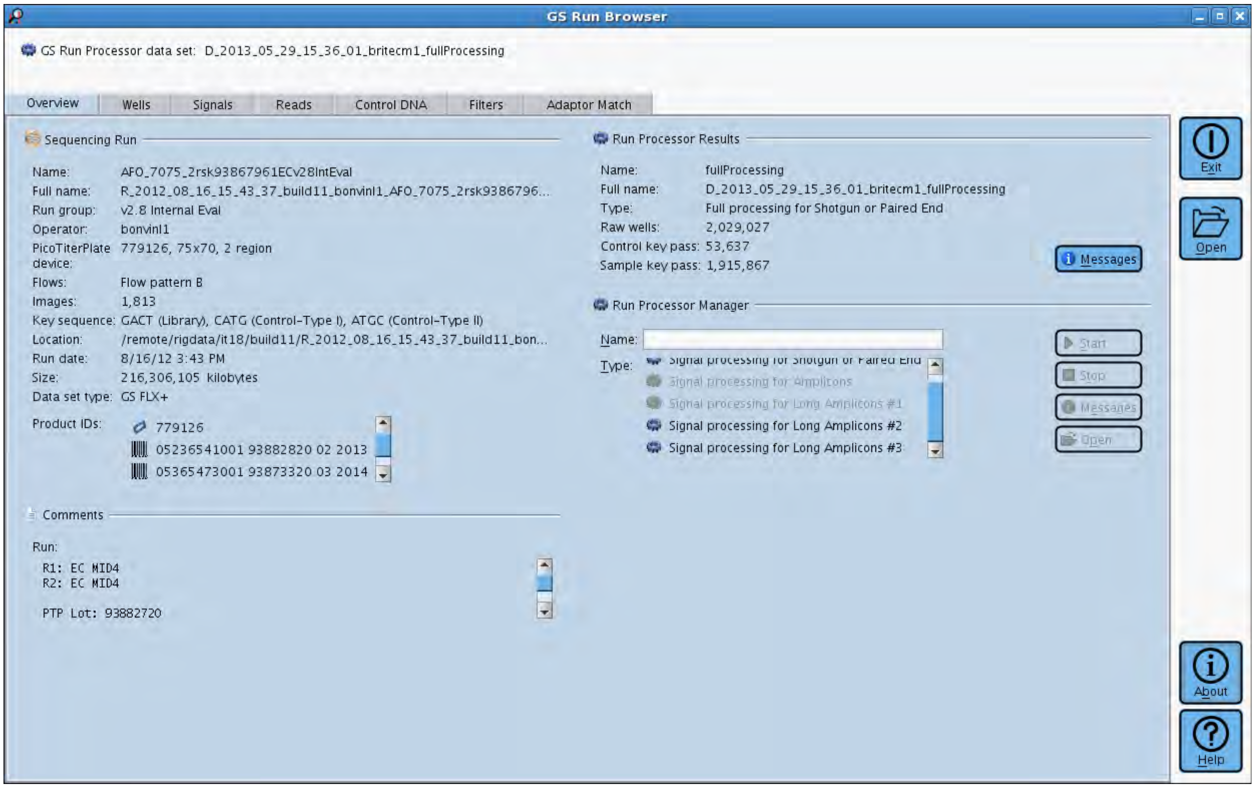
1 R_2008_05_17_13_54_32_FLX12070293_nramchandani_mwXsopXxlrkit08040925Xptp701544Xvfr14
2 R_2008_05_29_07_49_18_FLX12070284_lmcdade_200xRev10G70xVFRxEcoli
3 D_2008_07_30_18_35_46_dragonforce_fullProcessing
4 D_2008_07_18_11_01_15_ruth_fullProcessingPairedEnd
5 R_2007_06_28

Data set: /data/v2.0/R_2007_06_28_06_35_50_FLX10060106_cclark_10070759 3622020Library1PE/D_2008_07_18_11_01_15_ruth_fullProcessingP airedEnd
Modified: 7/31/08 3:54 PM

Figure 13: Pop up list of recent data sets, ready to reopen.

3.3 Overview of the GS Run Browser Interface

After a sequencing run data set has been loaded, it will be displayed in the GS Run Browser window (Figure 14). The full path of the data set is listed in the status area at the top of the page. Active data tabs are populated, inactive data tabs are grayed-out.



GS Run Processor data set: D_2013_05_29_15_36_01_britecm1_fullProcessing

Sequencing Run

Name: AFO_7075_2rsk93867961ECv28IntEval
Full name: R_2012_08_16_15_43_37_build11_bonvinl1_AFO_7075_2rsk9386796...
Run group: v2.8 Internal Eval
Operator: bonvinl1
PicoTiterPlate device: 779126, 75x70, 2 region
Flows: Flow pattern B
Images: 1,813
Key sequence: GACT (Library), CATG (Control-Type I), ATGC (Control-Type II)
Location: /remote/rigdata/it18/build11/R_2012_08_16_15_43_37_build11_bon...
Run date: 8/16/12 3:43 PM
Size: 216,306,105 kilobytes
Data set type: GS FLX+
Product IDs: 779126
05236541001 93882820 02 2013
05365473001 93873320 03 2014

Comments

Run:

```
R1: EC MID4
R2: EC MID4

PTP Lot: 93882720
```

Run Processor Results

Name: fullProcessing
Full name: D_2013_05_29_15_36_01_britecm1_fullProcessing
Type: Full processing for Shotgun or Paired End
Raw wells: 2,029,027
Control key pass: 53,637
Sample key pass: 1,915,867

Run Processor Manager

Name:
Type: Signal processing for Shotgun or Paired End
Signal processing for Amplicons
Signal processing for Long Amplicons #1
Signal processing for Long Amplicons #2
Signal processing for Long Amplicons #3

Figure 14: General view of the GS Run Browser window.

3.3.1 The Global Action Area Buttons

Four main buttons are always available in the right hand tool bar (see Figure 14):

- The Exit button closes the application.
- The Open button allows a user to browse the file system to find a data set to open. Clicking this button while pressing the Shift key evokes a pop up menu allowing the user to choose from a list of data sets recently opened with the GS Run Browser. Only one data set may be opened at a time, thus when opening a new data set, a message will appear asking the user to confirm closing of the current data.
- The About button opens a dialog window providing version information about the GS Run Browser application and a button to access the GS Support Tool used to help in the troubleshooting of potential sequencing run issues (see Section 3.10 for more details about the Support Tool).
- The Help button will provide instructions for where to find further information on the software, or will link you directly to a searchable version of the *454 Sequencing System Software Manual*.

3.3.2 The Tabs

The GS Run Browser application displays the various aspects of the sequencing run results in a series of 6 tabs, as listed below.

The **Overview** tab contains:

- a summary of the sequencing run on the left
- the GS Run Processor results, if available, on the top right
- the run Processor Manager interface, which allows the user to launch a new data processing job of the currently open sequencing run, on the bottom right.

The **Wells** tab displays three layers of information:

- the raw images captured during the sequencing run
- the regions found during image processing
- the wells found during image processing.

This tab also gives the user access to well flowgrams by right-clicking on areas of the PTP device image.

The **Signals** tab provides statistics on raw signal intensity and homopolymer length distributions across all the wells of the PicoTiterPlate device for any individual flow, for either sample library or Control DNA reads.

The **Reads** tab provides statistics on read length and quality score results, for either sample library or Control DNA reads.

The **Control DNA** tab displays the accuracy metrics for the Control DNA reads (% match to their reference sequences) and other Control DNA information.

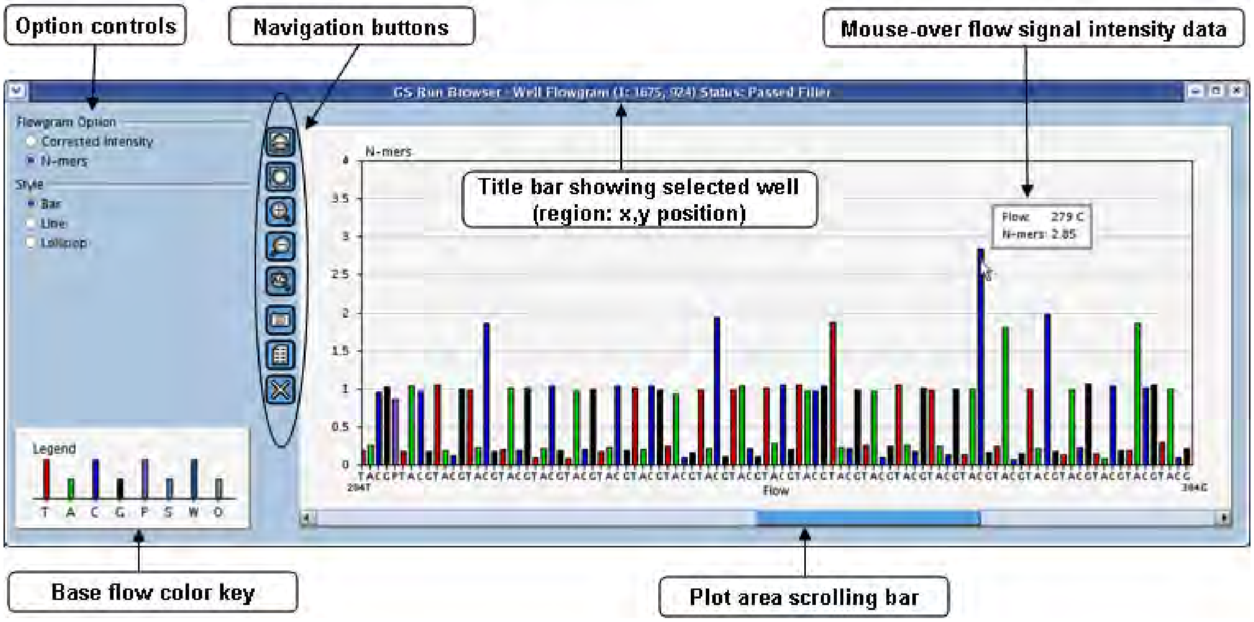
The **Filters** tab shows the quality filter information for the run, including the number of wells that failed and passed each filter, for either sample library or Control DNA reads.

The **Adaptor Match** tab provides statistics on the number of 3'-adaptor matches found by the adaptor match tool in the basecaller (Section 1.3.3.3), as well as a read length plots for adaptor match and no adaptor match reads.

3.3.3 The Buttons and Plots

The images, plots and data tables displayed in the various tabs of the GS Run Browser application are scrollable and/or zoomable graphical elements. They share certain common buttons and functions, *e.g.* to perform the scrolling and zooming. When they do appear, these graphic elements have some or all of the following features (see in Figure 15, an example of a well flowgram window, which has most of these elements):

- Scroll bars for horizontal and/or vertical scrolling (appearing below and to the right of the plot when necessary).
- A column of buttons along the upper left edge of the graphic elements, used for navigation (including various zooming functions) and/or to save snapshot images or text files of the displayed data (see Section 3.3.4 for details). For the image area of the Wells tab, this is replaced by a unique set of controls that are overlaid near the upper left corner of the image.
- Mouse functions (pointing, clicking or dragging the mouse, touchpad, pen, *etc.* over the graphical element) to view data values and adjust the zoom level.



Option controls

Navigation buttons

Mouse-over flow signal intensity data

Title bar showing selected well (region: x,y position)

Base flow color key

Plot area scrolling bar

GS Run Browser - Well Flowgram (1: 1675, 924) Status: Passed Filter

Flowgram Option
○ Corrected Intensity
● N-mers

Style
● Bar
○ Line
○ Lollipop

Flow: 279 C
N-mers: 2.85

Legend
T
A
C
G
P
S
W
O
N-mers
4
3.5
3
2.5
2
1.5
1
0.5
0
TACGPTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACG
294T
Flow
384C

Figure 15: An example Well Flowgram window showing many of the common graphical element functions.

3.3.4 The Navigation and Data Capture Buttons

The buttons appearing to the left of an element have common general functions. In some cases their specific meaning is adjusted to the context of the element:

<table>
<tr><th>Button</th><th>Name – Description</th></tr>
<tr><td>◎</td><td><b>Zoom to default size</b> – Some plots default to a zoom level different than ‘fit’. For those plots, this button resets the zoom level to the default initially shown.</td></tr>
<tr><td>⊜</td><td><b>Zoom to fit data</b> – Fit means ‘zoom all the way out.’ On plots, scale out to the limits of the data.</td></tr>
<tr><td>⊕</td><td><b>Zoom in</b> – Zoom in by a factor of 1.5. On image displays, this will zoom in by this factor, centered on the middle of the display. For plots, this button zooms only the y-axis scale (use the <b>Zoom to labels</b> or <b>mouse freehand zooming</b> functions described below to zoom the x-axis).</td></tr>
<tr><td>⊖</td><td><b>Zoom out</b> – Zoom out by a factor of 1.5. On image displays, this will zoom out by this factor, centered on the middle of the display. For plots, this will zoom only the y-axis scale and, unlike most zoom operations, this will zoom out past the data limits (to allow the user to get a better perspective of the data).</td></tr>
<tr><td>AC</td><td><b>Zoom to x-axis labels</b> – This button only appears in the flowgram viewers; it zooms the x-axis of the flowgram so that the nucleotide/flow characters can fit below the axis.</td></tr>
<tr><td>📷</td><td><b>Save chart as image file</b> – Save a snapshot image of the current view to disk. This will open a dialog asking for a location and file name, and then will save a PNG image file at the location specified. For Summary data tables, this saves the whole tabular area, while for all other displays the saved image contains only the visible region of the element.</td></tr>
<tr><td>▤</td><td><b>Save chart data to tab-separated values file</b> – Save an Excel-compatible, tab-delimited text file of the spreadsheet data for this graphic. This will open a dialog asking for the location and file name to save the file. It then saves the data, along with summary information describing where the data came from, as shown below (a file portion saved from a well flowgram):
<pre>
GS Run Browser - Well Flowgram (1: 1781, 2107) Status: Passed
Filter Corrected Intensity vs. Flow
GS Run Browser, Wed Aug 20 09:56:17 EDT 2009
Run
Directory=/data/flx/R_2009_02_08_17_02_43_build04_
mcorso_100x2575xecoli360k
Processor
Directory=/data/flx/R_2009_02_08_17_02_43_build04_mcorso_100x2575
xecoli360k/D_2008_07_31_11_11_55_enya_signalProcessing
Flow Corrected Intensity
5 PPI 3,696
8 T 511.5
9 A 112.69
10 C 491.5
11 G 150.12
...
</pre>
For plots, the spreadsheet file contains the data currently being plotted. For a tri-flowgram view (see Section 3.5.4), the data for all three plot subsections are saved to one file, with white space between subsections. For Summary data tables, the spreadsheet contains the table data. If a plot and a summary present the same data, there may be only one spreadsheet button, attached to whichever constitutes the superset of the data.</td></tr>
<tr><td>☒</td><td><b>Close</b> – Close a flowgram window. The ‘x’ icon in the upper right window corner will also close the window.</td></tr>
</table>

In addition to the buttons described above, it is possible to zoom in on a specific region of a plot by dragging a box (a 'marquee') around the region of interest. This is particularly useful for zooming in on the read length profiles on the Reads and Adaptor Match tabs. The zoom to default size button will return the plot to the original zoom scale.

The image area of the Wells tab has special navigation controls described in Section 3.5.2.4.

3.3.5 The Mouse Functions

When the mouse cursor is located over the flowgram, additional functions can be performed by moving, clicking or dragging the mouse:

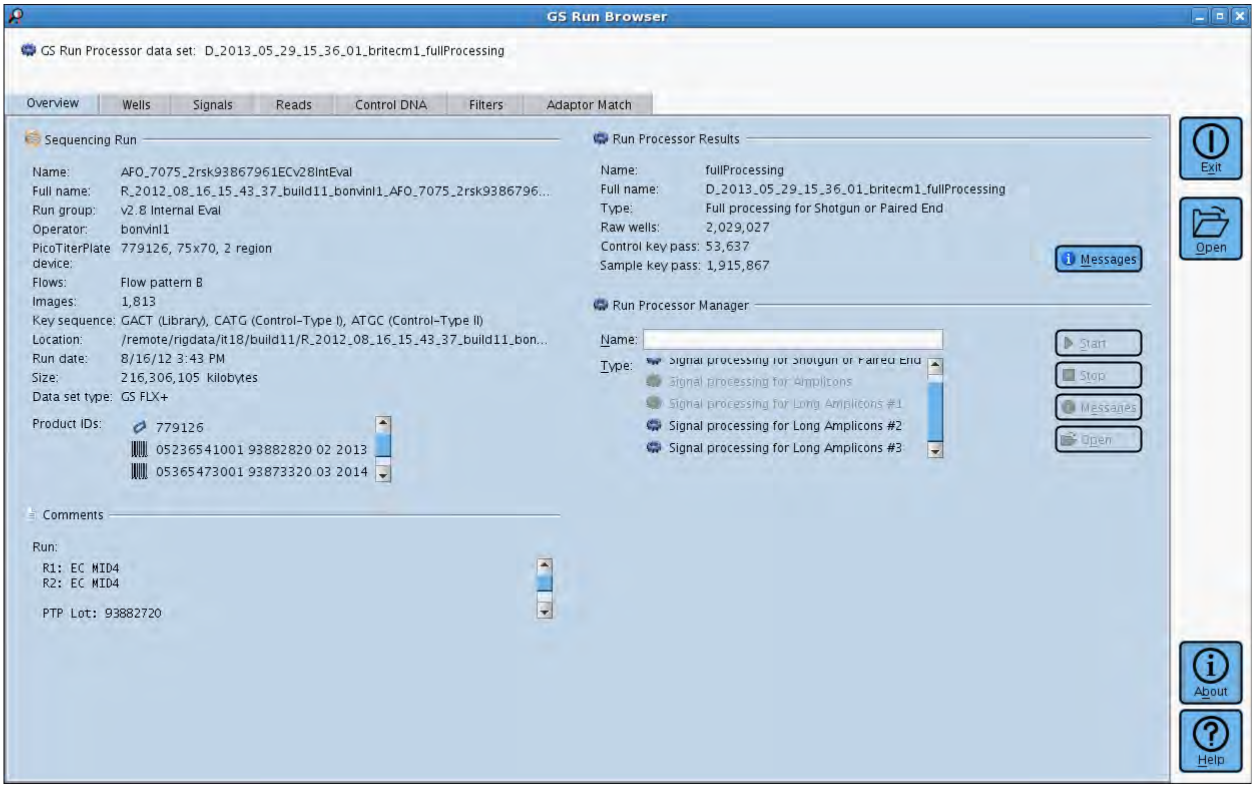
- **Mouse Tracker** – Whenever the cursor is located on data shown in an image or plot, the numeric data values for that data is shown in a related Mouse Tracker window that pops up near the location of the cursor. This allows the user to see the specific numeric value for any data point (see example in Figure 15).
- **Drag Image** – If the displayed image is larger than the viewing frame on screen, the mouse can be used to drag the image within the frame. Left-click on the image and drag the mouse in the direction needed.
- **Freehand Zoom In** – The mouse can be used to zoom in on specific regions of the image or of a plot. To zoom in on a plot, hold the left mouse button down and drag a box around the area of interest (not shown). Releasing the button zooms to that region. On the image area of the Wells tab, this function is accomplished with the right mouse button rather than the left, because of the 'left click and drag' function just described.
- **Freehand Zoom Out** – For plots only, right-clicking the plot will cause the plot to zoom out by a factor of 1.5 in both the X and Y directions, centered on the middle of the current view. This zoom will not zoom farther than the limits of the data.

3.4 The Overview Tab

The Overview tab is the default tab shown after a data set has been selected for viewing in the GS Run Browser. It is always populated when a run or a data processing data set is opened. A run data set is in a directory prefixed with an ‘R_’, whereas a Data Processing data set is prefixed with a ‘D_’.

The Overview tab contains three areas (Figure 16):

- the **Sequencing Run** area displays summary information about the sequencing run for data set currently open
- the **Run Processor Results** area shows a summary of the data processing results including the total number of raw wells, the number of key pass wells with the Control DNA key and with the sample library key
- the **Run Processor Manager** area includes controls that allow the user to process (or re-process) the data of the open data set.



GS Run Processor data set: D_2013_05_29_15_36_01_britecm1_fullProcessing

Sequencing Run

Name: AFO_7075_2rsk93867961ECv28IntEval
Full name: R_2012_08_16_15_43_37_build11_bonvinl1_AFO_7075_2rsk9386796...
Run group: v2.8 Internal Eval
Operator: bonvinl1
PicoTiterPlate device: 779126, 75x70, 2 region
Flows: Flow pattern B
Images: 1,813
Key sequence: GACT (Library), CATG (Control-Type I), ATGC (Control-Type II)
Location: /remote/rigdata/it18/build11/R_2012_08_16_15_43_37_build11_bon...
Run date: 8/16/12 3:43 PM
Size: 216,306,105 kilobytes
Data set type: GS FLX+
Product IDs: 779126
05236541001 93882820 02 2013
05365473001 93873320 03 2014

Comments

Run:

```
R1: EC MID4
R2: EC MID4

PTP Lot: 93882720
```

Run Processor Results

Name: fullProcessing
Full name: D_2013_05_29_15_36_01_britecm1_fullProcessing
Type: Full processing for Shotgun or Paired End
Raw wells: 2,029,027
Control key pass: 53,637
Sample key pass: 1,915,867

Run Processor Manager

Name:
Type: Signal processing for Shotgun or Paired End
Signal processing for Amplicons
Signal processing for Long Amplicons #1
Signal processing for Long Amplicons #2
Signal processing for Long Amplicons #3

Figure 16: GS Run Browser Overview tab.

3.4.1 Sequencing Run Area

This area shows basic information about the sequencing run that produced the data set, as shown in Figure 16. The source of information is as follows:

- For a run data set, the data shown comes from the dataRunParams.parse file.
- For a Data Processing data set, the source of the information displayed depends on the software version used to process the data:
    - For sequencing runs processed with the GS Run Processor version 2.0.00 or later, the information comes from the *region*.cwf files.
    - For sequencing runs processed with software version 1.1.03 or earlier, the information comes from the dataRunParams.parse and imageLog.parse files.

3.4.2 Run Processor Results Area

This area of the Overview tab shows a summary of the data processing results, if any. The source of information is as follows:

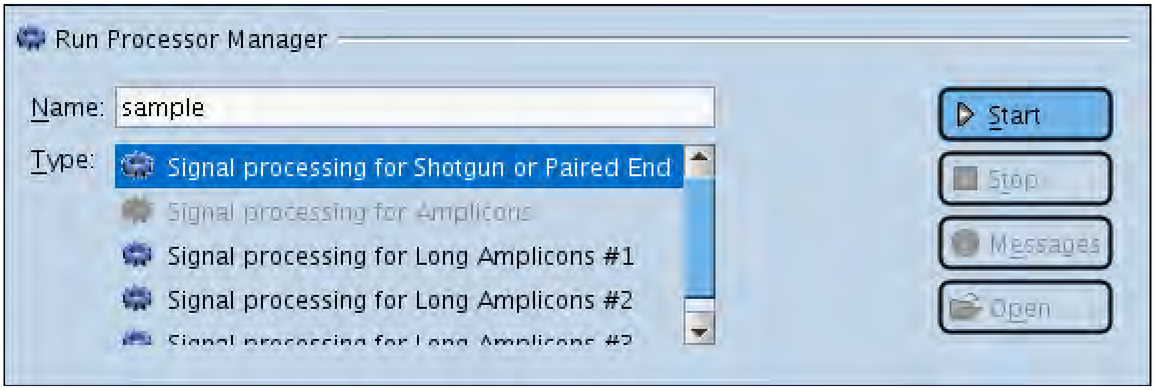
- For a run data set, there are no data processing results, so the run Processor Results area of the Overview tab contains no data and is grayed out. (Also, the Wells tab will display only the images, and all the other tabs will be unavailable.)
- For a Data Processing data set, the source of the information displayed depends on the software version used to process the data:
    - For sequencing runs processed with the GS Run Processor version 2.0.00 or later, the information comes from the *region*.cwf files.
    - For sequencing runs processed with software version 1.1.03 or earlier, the information comes from the files present in the D_ directory.

3.4.3 Run Processor Manager

The Run Processor Manager allows a user to process the data set currently displayed. If the GS Run Processor application is not currently available, the Run Processor Manager is disabled (grayed out), and the statement ‘No Run Processor Manager available’ appears. This may happen if the Run Processor Manager is not enabled on the same machine as the one where the GS Run Browser was started.

> Full processing and signal processing are not supported for acyclic flow pattern B data on GS FLX+ instruments or on 32 bit datarigs.

Data processing jobs not configured as a part of the sequencing run can be launched post-run from the GS Run Browser (Figure 17). This tool can also be used to reprocess data sets that were processed previously; each round of processing will give rise to its own D_ directory.



Name: sample

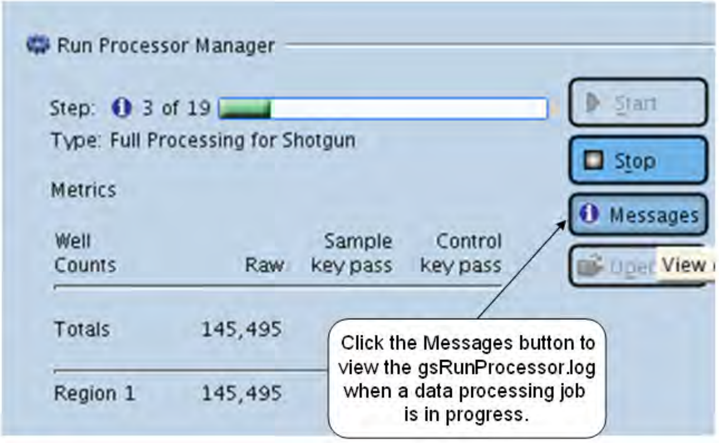
Type:
- Signal processing for Shotgun or Paired End
- Signal processing for Amplicons
- Signal processing for Long Amplicons #1
- Signal processing for Long Amplicons #2
- Signal processing for Long Amplicons #3

Figure 17: GS Run Processor Manager Tool in the GS Run Browser Overview tab.

To launch a data processing job:

- Type in a unique name for the processing job - This name will be appended to the data processing directory name (D_...) and the folder of processed results will be placed inside the sequencing run folder (R_...), parallel to any existing data processing folders.
- Select the processing type. The options below are the standard choices shipped with the software, but your site may have additional options.
  - If you opened an R_ directory, the options presented will be:
    - Image Processing Only
    - Full Processing for Shotgun or Paired End
    - Full Processing for Amplicons
    - Full Processing for Long Amplicons #1
    - Full Processing for Long Amplicons #2
    - Full Processing for Long Amplicons #3
  - If you opened a D_ directory, the options presented will be:
    - Signal Processing for Shotgun or Paired End
    - Signal Processing for Amplicons
    - Signal Processing for Long Amplicons #1
    - Signal Processing for Long Amplicons #2
    - Signal Processing for Long Amplicons #3
- Click the ‘Start’ button. There is a ‘Stop’ button if the processing job needs to be halted for any reason.

Once a job has been launched, the progress and statistics of the job is displayed, as shown on Figure 18.



Run Processor Manager

Step: 3 of 19

Type: Full Processing for Shotgun

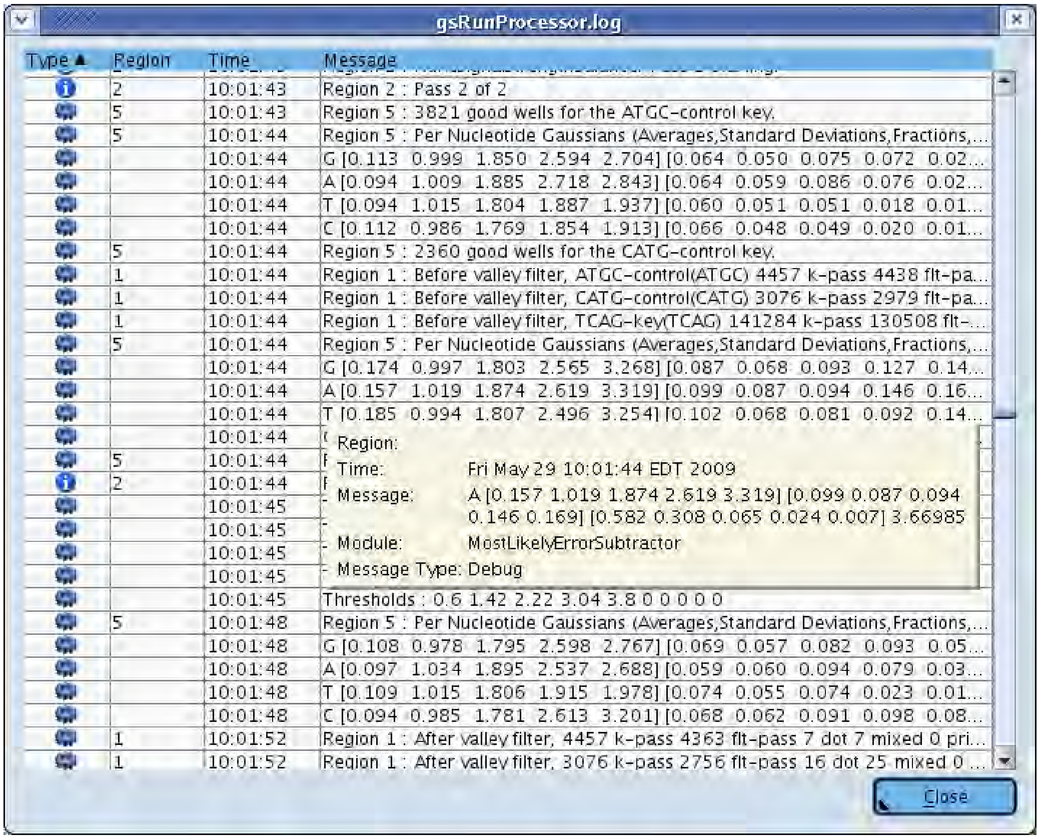
Metrics

Well Counts	Raw	Sample key pass	Control key pass
Totals	145,495		
Region 1	145,495		

Click the Messages button to view the gsRunProcessor.log when a data processing job is in progress.

Figure 18: GS Run Processor Manager Tool job-in-progress statistics.

In addition to the 'Start' and 'Stop' buttons for launching and halting a processing job, the gsRunProcessor.log file can be viewed in a separate pop-up window by clicking the 'Messages' button. The log file is presented in a spreadsheet view with log entries listed by timestamp (Figure 19). The tooltip shows additional information about the log entry under the pointer.



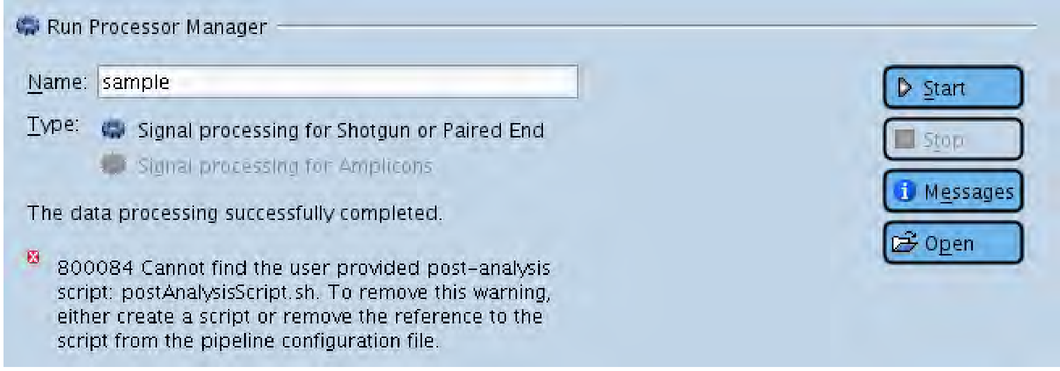
gsRunProcessor.log

Type ▲	Region	Time	Message
ℹ	2	10:01:43	Region 2 : Pass 2 of 2
⚙	5	10:01:43	Region 5 : 3821 good wells for the ATGC-control key.
⚙	5	10:01:44	Region 5 : Per Nucleotide Gaussians (Averages,Standard Deviations,Fractions,...
⚙		10:01:44	G [0.113 0.999 1.850 2.594 2.704] [0.064 0.050 0.075 0.072 0.02...
⚙		10:01:44	A [0.094 1.009 1.885 2.718 2.843] [0.064 0.059 0.086 0.076 0.02...
⚙		10:01:44	T [0.094 1.015 1.804 1.887 1.937] [0.060 0.051 0.051 0.018 0.01...
⚙		10:01:44	C [0.112 0.986 1.769 1.854 1.913] [0.066 0.048 0.049 0.020 0.01...
⚙	5	10:01:44	Region 5 : 2360 good wells for the CATG-control key.
⚙	1	10:01:44	Region 1 : Before valley filter, ATGC-control(ATGC) 4457 k-pass 4438 flt-pa...
⚙	1	10:01:44	Region 1 : Before valley filter, CATG-control(CATG) 3076 k-pass 2979 flt-pa...
⚙	1	10:01:44	Region 1 : Before valley filter, TCAG-key(TCAG) 141284 k-pass 130508 flt-...
⚙	5	10:01:44	Region 5 : Per Nucleotide Gaussians (Averages,Standard Deviations,Fractions,...
⚙		10:01:44	G [0.174 0.997 1.803 2.565 3.268] [0.087 0.068 0.093 0.127 0.14...
⚙		10:01:44	A [0.157 1.019 1.874 2.619 3.319] [0.099 0.087 0.094 0.146 0.16...
⚙		10:01:44	T [0.185 0.994 1.807 2.496 3.254] [0.102 0.068 0.081 0.092 0.14...
⚙		10:01:44	(
⚙	5	10:01:44	F
ℹ	2	10:01:44	F
⚙		10:01:45	-
⚙		10:01:45	-
⚙		10:01:45	-
⚙		10:01:45	-
⚙		10:01:45	Thresholds : 0.6 1.42 2.22 3.04 3.8 0 0 0 0 0
⚙	5	10:01:48	Region 5 : Per Nucleotide Gaussians (Averages,Standard Deviations,Fractions,...
⚙		10:01:48	G [0.108 0.978 1.795 2.598 2.767] [0.069 0.057 0.082 0.093 0.05...
⚙		10:01:48	A [0.097 1.034 1.895 2.537 2.688] [0.059 0.060 0.094 0.079 0.03...
⚙		10:01:48	T [0.109 1.015 1.806 1.915 1.978] [0.074 0.055 0.074 0.023 0.01...
⚙		10:01:48	C [0.094 0.985 1.781 2.613 3.201] [0.068 0.062 0.091 0.098 0.08...
⚙	1	10:01:52	Region 1 : After valley filter, 4457 k-pass 4363 flt-pass 7 dot 7 mixed 0 pri...
⚙	1	10:01:52	Region 1 : After valley filter, 3076 k-pass 2756 flt-pass 16 dot 25 mixed 0 ...

```
Region:
Time:          Fri May 29 10:01:44 EDT 2009
Message:       A [0.157 1.019 1.874 2.619 3.319] [0.099 0.087 0.094
               0.146 0.169] [0.582 0.308 0.065 0.024 0.007] 3.66985
Module:        MostLikelyErrorSubtractor
Message Type: Debug
```

Figure 19: gsRunProcessor.log file.

Error messages for processes that do not complete successfully are also displayed and described in detail in the area below the 'Type' field, listed by error message number (Figure 20). Each message has an associated icon for the type of log entry, as shown in Figure 21. The GS Run Browser supports older versions of the Run Processor Manager but will display a warning message when it detects a version older than 2.3.


Run Processor Manager

Name: sample

Type: Signal processing for Shotgun or Paired End
Signal processing for Amplicons

The data processing successfully completed.

800084 Cannot find the user provided post-analysis script: postAnalysisScript.sh. To remove this warning, either create a script or remove the reference to the script from the pipeline configuration file.

Figure 20: GS Run Processor Manager Tool job error messaging

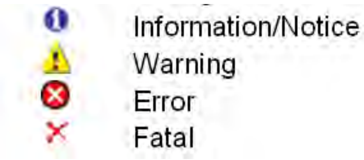


Figure 21: Message types and corresponding icons.

When a processing job has finished, the data set can be loaded into the GS Run Browser by clicking the 'Open' button (the fourth button in the Run Processor Manager section see Figure 20). To launch a processing job using the Run Processor Manager:

- Type in a unique name for the processing run data set. This name will be appended to the data processing directory name (D_...) and the folder of processed results will be placed inside the sequencing run folder (R_...), parallel to any existing data processing folders.
- Select the processing type – The options available depend on the data set being browsed.
- Click the 'Start' button. The 'Stop' button will become available, in case the processing job needs to be halted for any reason.

The data processing pipeline options are discussed in detail in Section 1.1.

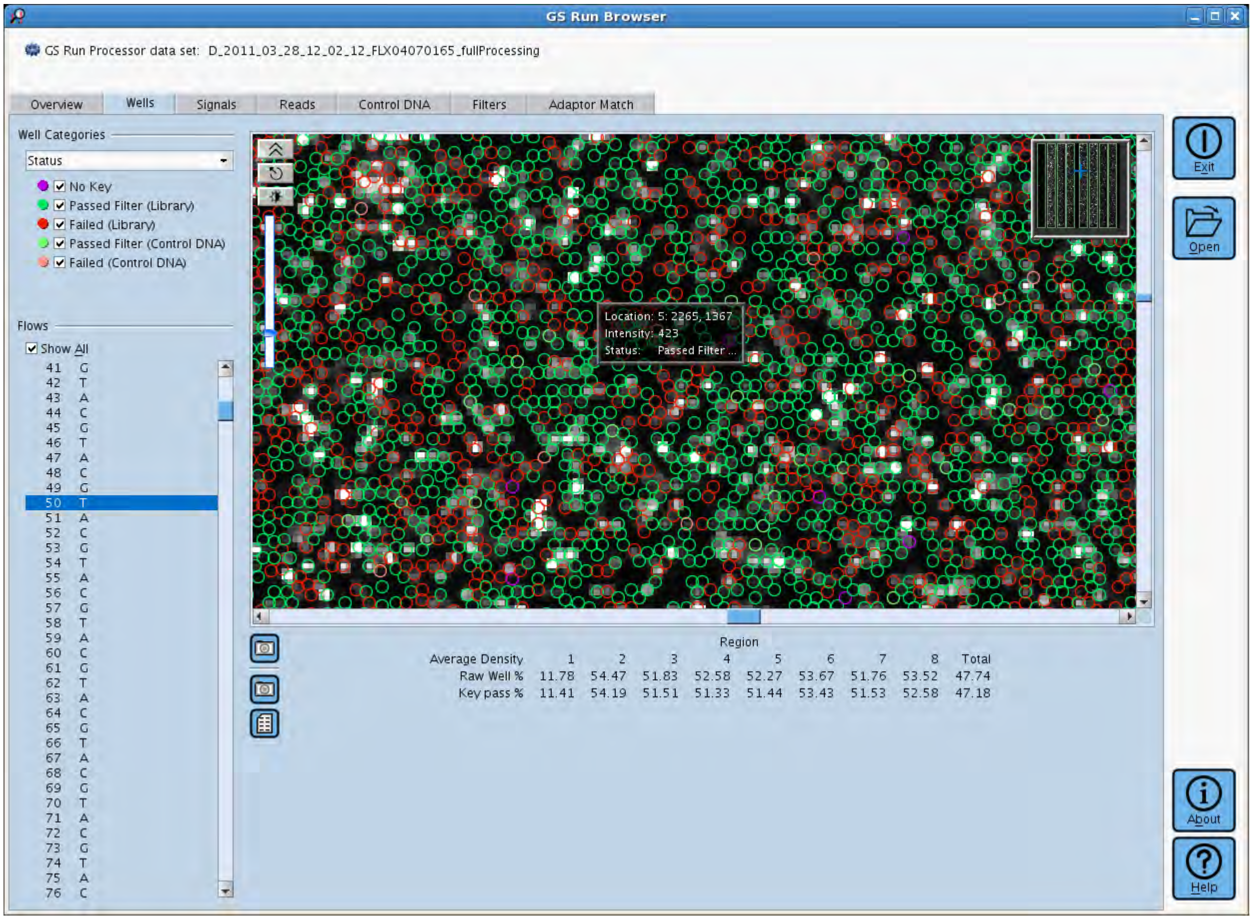
Once a data processing job completes, click on the 'Open' button in the Run Processor Manager section (see Figure 20) to view the results in the GS Run Browser. This will close any currently viewed data set and display the results of the newly processed data set.

3.5 The Wells Tab

The Wells tab (Figure 22) displays information about the wells identified during data processing; this information is overlaid on the raw images if they are available. The main display features include the following:

- Well indicators, shown as colored circles (where the color scheme can indicate a number of read type and quality attributes described below) overlaid on the image of the PTP device for each base flow
- Access to the fully processed well flowgrams generated by the data processing software

Additionally, well density statistics are displayed for each region of the PTP device and for the whole PTP device, below the image area.



GS Run Processor data set: D_2011_03_28_12_02_12_FLX04070165_fullProcessing

Well Categories

Status

- ● ☑ No Key
- ● ☑ Passed Filter (Library)
- ● ☑ Failed (Library)
- ● ☑ Passed Filter (Control DNA)
- ● ☑ Failed (Control DNA)

Flows

☑ Show All

41	G
42	T
43	A
44	C
45	G
46	T
47	A
48	C
49	G
50	T
51	A
52	C
53	G
54	T
55	A
56	C
57	G
58	T
59	A
60	C
61	G
62	T
63	A
64	C
65	G
66	T
67	A
68	C
69	G
70	T
71	A
72	C
73	G
74	T
75	A
76	C

Location: 5: 2265, 1367
Intensity: 423
Status: Passed Filter ...

	Region								
Average Density	1	2	3	4	5	6	7	8	Total
Raw Well %	11.78	54.47	51.83	52.58	52.27	53.67	51.76	53.52	47.74
Key pass %	11.41	54.19	51.51	51.33	51.44	53.43	51.53	52.58	47.18

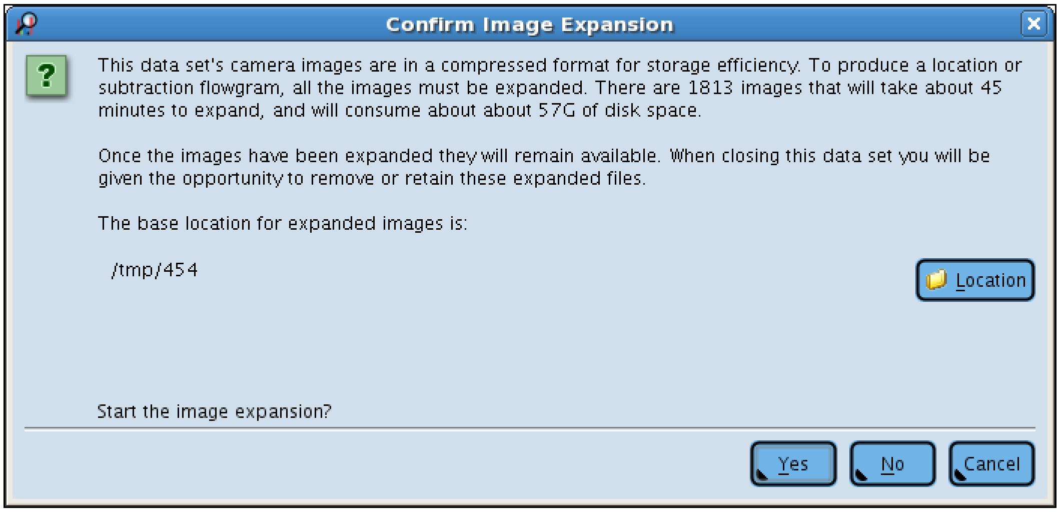
Figure 22: The GS Run Browser application's Wells tab.

For the Well data to be displayed, the *region*.cwf file(s) [or, for data sets processed with software v. 1.1.03 or earlier, the *region*.wells file(s)], located in the 'regions' sub-folder of the 'D_' data processing folder of the sequencing run, must exist. If no *region*.cwf (or *region*.wells) files exist for the run, the Wells tab will only display the images and the Flows selector, if available.

3.5.1 Expand Images (.png Image Files)

For the images to be displayed, there must be a rawImages directory (containing at least one .png or legacy .pif file) for the sequencing run being viewed. The information on which reagent was flowed at each step comes from the imageLog.parse file.

For runs from a GS FLX+ Instrument, images stored in the compressed .png format must be expanded to the .pif format before viewing. Right-click an image and choose Expand images from the contextual menu, which displays a dialog that allows you to control where the expanded images will be stored (see Figure 23). Click the Yes button to continue.


Confirm Image Expansion

This data set's camera images are in a compressed format for storage efficiency. To produce a location or subtraction flowgram, all the images must be expanded. There are 1813 images that will take about 45 minutes to expand, and will consume about about 57G of disk space.

Once the images have been expanded they will remain available. When closing this data set you will be given the opportunity to remove or retain these expanded files.

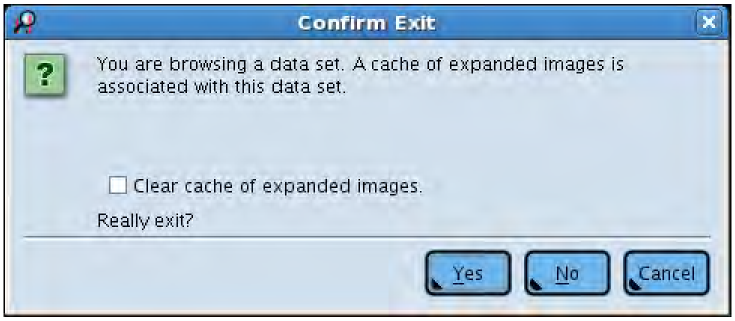
The base location for expanded images is:

/tmp/454

Start the image expansion?

Figure 23: Confirm Image Expansion dialog.

Expanding images can take several minutes, and the expanded files consume a substantial amount of disk space. When the run is closed in GS Run Browser, you will have the option of choosing to retain the expanded images for future performance benefit or to clear the cache of images for disk storage benefit (see Figure 24).


Confirm Exit

You are browsing a data set. A cache of expanded images is associated with this data set.

☐ Clear cache of expanded images.

Really exit?

Figure 24: Confirm Exit dialog with expanded images stored on disk.

Data sets from past sequencing runs may or may not contain the images, since images may be removed or archived to save storage space. All the tabs in the GS Run Browser will still be available even if the images are not present.

3.5.2 Wells Tab Features and Functionalities

The Wells tab contains many features with the functionalities described in the sub-sections below.

3.5.2.1 Well Categories

This is a drop down menu used to select the well display category (Figure 25). Each category has its own color chart to identify the specific attribute of each well, within the category (see Section 3.5.2.2). If the data from one or more regions of the PicoTiterPlate device is missing, '(Data unavailable)' will appear in the Well Categories drop down menu.

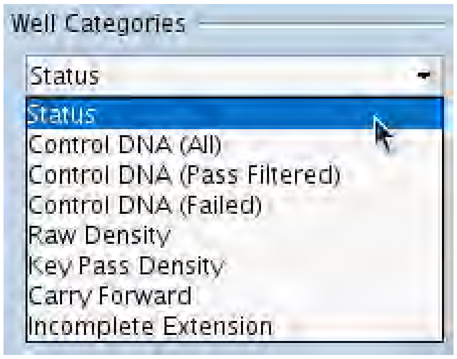


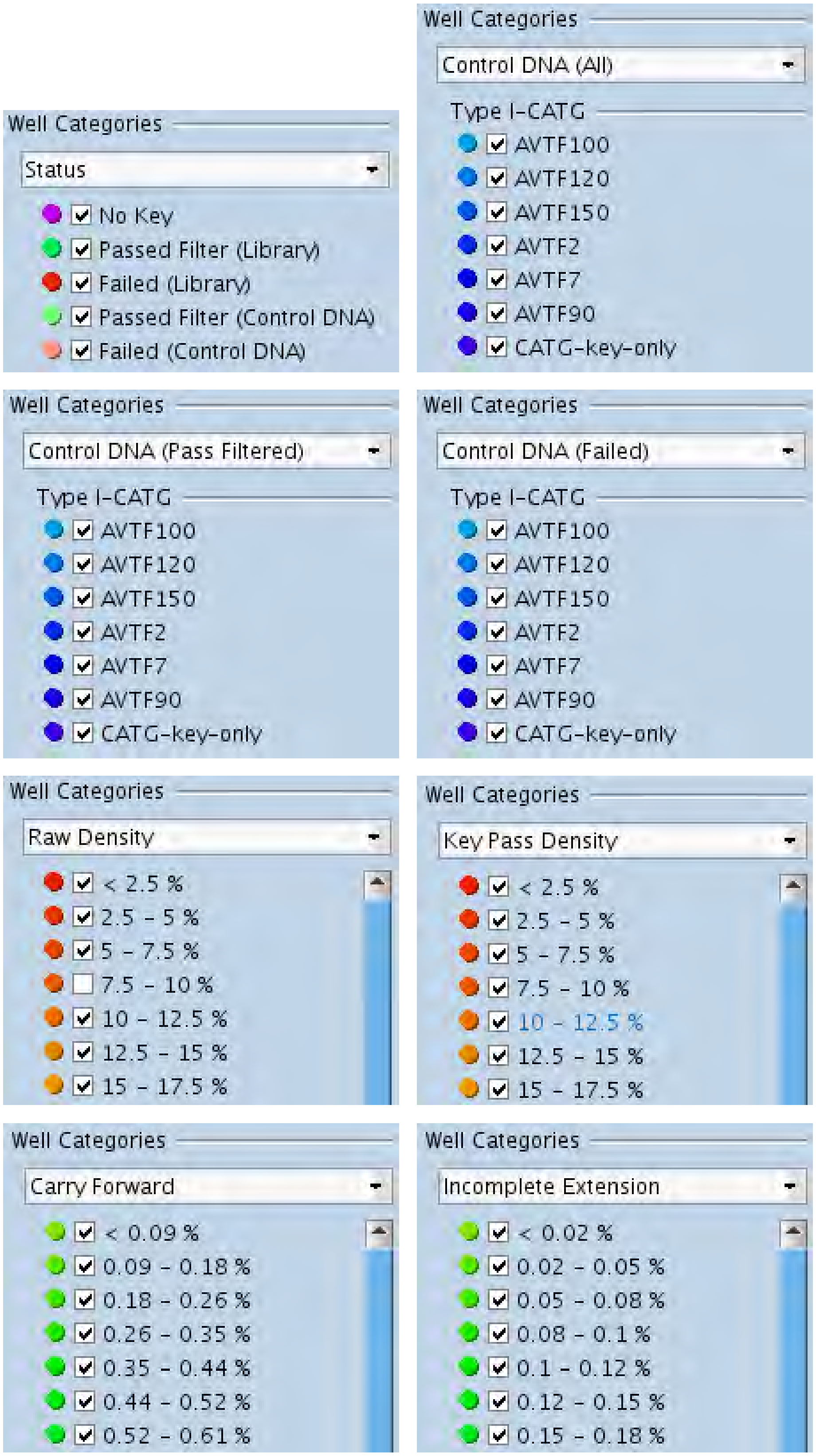
Figure 25: Wells Tab, Well Categories Menu.

- **Status** category
  - **No Key**: Identified as a well (generates signal), but not one with recognizable data
  - **Library Passed/Failed Filter**: Library read (key is TCAG or GACT) that either passed or failed quality filters. Note that the GS Run Browser does not make pass/fail decisions; it only reads results provided by the quality filtering algorithms of the GS Run Processor application.
  - **Control DNA Passed/Failed Filter**: Control DNA read (key is CATG or ATGC; see Section 7.1 for details) that either passed or failed quality filters.
- **Control DNA** category: This is similar to the Status category except that it uses separate colors for the reads matching each specific Control DNA sequence, or wells where an appropriate Control DNA key was detected but the sequence of the read does not match any of the known Control DNA fragments (see Section 7.1 for details on Control DNA keys). Wells where the library key or no key at all was detected are also displayed, using the same color scheme as for the Status category.
- **Raw Density** and **Key Pass Density**: Measures of local well crowding, calculated individually for all identified wells in which a signal is detected (Raw Density) and for all identified wells in which the initial signals match a known key, *i.e.* excluding the wells with a No Key status. For data sets processed with the software version 2.0.00 or later, this option will appear only if the signal processing step of Data Processing was carried out. For versions 1.1.03 or earlier, this will normally show (requires the region.wells files in the data set). Color gradient bins of +2.5% can be selected.
- **Carry Forward** and **Incomplete Extension**: Level of CAFIE correction applied to the first 400 flows of each well due to the detection of carry forward and incomplete extension events (described in Section 1.3.1). Although the level of correction applied is an indication of the relative amount of CAFIE error identified and corrected in a well, the CAFIE correction values are not comparable between reads sequenced with cyclic *vs.* acyclic flow patterns. This option will appear only if the signal processing step of Data Processing was carried out. Individual color gradient bins can be selected.

3.5.2.2 Color chart

Located just below the Well categories drop down menu (Figure 25), this area allows the user to select the range of well values to display in the main image area, for the category selected. These controls have the following general features and behaviors:

- Checking the box for an attribute (or numerical range) in the list causes wells matching that attribute (or range) to be displayed on the image. If an item in that list is not checked, the corresponding wells are not displayed. Shift-clicking deselects all attributes; or, if all are already deselected, it selects them all. Control-clicking deselects all attributes except the one clicked; or, if the user control-clicks on an attribute that is currently the only one selected, it selects all attributes except the one clicked.
- For the categories whose attributes are continuous variables, the color scheme uses the ‘color wheel’ order for the gradation of value ranges.
- For the attributes that denote ‘goodness’ or ‘badness’ of the wells, the ‘good’ values are indicated by green hues and the ‘bad’ values by red hues, when possible. For example, the most desirable values (wells marked green) for well density in the mid-range values (40-45%) but for CAFIE correction the wells with the least correction have the smallest values (closest to zero).
- The GS Run Browser ‘remembers’ attribute selection for each category when a user navigates to another tab and then returns to the Wells tab.



Well Categories

Status

- ● ☑ No Key
- ● ☑ Passed Filter (Library)
- ● ☑ Failed (Library)
- ● ☑ Passed Filter (Control DNA)
- ● ☑ Failed (Control DNA)

Well Categories

Control DNA (All)

Type I-CATG

- ● ☑ AVTF100
- ● ☑ AVTF120
- ● ☑ AVTF150
- ● ☑ AVTF2
- ● ☑ AVTF7
- ● ☑ AVTF90
- ● ☑ CATG-key-only

Well Categories

Control DNA (Pass Filtered)

Type I-CATG

- ● ☑ AVTF100
- ● ☑ AVTF120
- ● ☑ AVTF150
- ● ☑ AVTF2
- ● ☑ AVTF7
- ● ☑ AVTF90
- ● ☑ CATG-key-only

Well Categories

Control DNA (Failed)

Type I-CATG

- ● ☑ AVTF100
- ● ☑ AVTF120
- ● ☑ AVTF150
- ● ☑ AVTF2
- ● ☑ AVTF7
- ● ☑ AVTF90
- ● ☑ CATG-key-only

Well Categories

Raw Density

- ● ☑ < 2.5 %
- ● ☑ 2.5 - 5 %
- ● ☑ 5 - 7.5 %
- ● □ 7.5 - 10 %
- ● ☑ 10 - 12.5 %
- ● ☑ 12.5 - 15 %
- ● ☑ 15 - 17.5 %

Well Categories

Key Pass Density

- ● ☑ < 2.5 %
- ● ☑ 2.5 - 5 %
- ● ☑ 5 - 7.5 %
- ● ☑ 7.5 - 10 %
- ● ☑ 10 - 12.5 %
- ● ☑ 12.5 - 15 %
- ● ☑ 15 - 17.5 %

Well Categories

Carry Forward

- ● ☑ < 0.09 %
- ● ☑ 0.09 - 0.18 %
- ● ☑ 0.18 - 0.26 %
- ● ☑ 0.26 - 0.35 %
- ● ☑ 0.35 - 0.44 %
- ● ☑ 0.44 - 0.52 %
- ● ☑ 0.52 - 0.61 %

Well Categories

Incomplete Extension

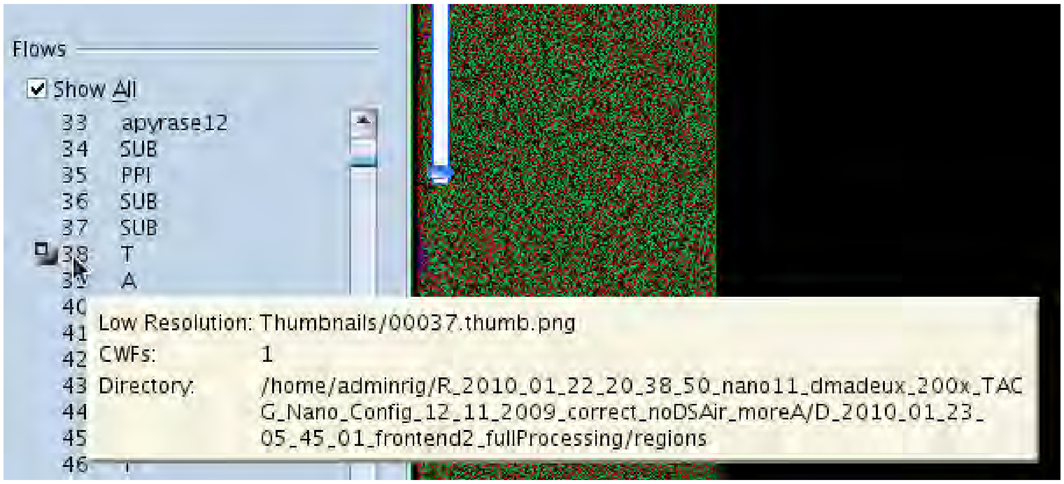
- ● ☑ < 0.02 %
- ● ☑ 0.02 - 0.05 %
- ● ☑ 0.05 - 0.08 %
- ● ☑ 0.08 - 0.1 %
- ● ☑ 0.1 - 0.12 %
- ● ☑ 0.12 - 0.15 %
- ● ☑ 0.15 - 0.18 %

Figure 26: Examples of color charts for well category display.

3.5.2.3 Flows

Located below the color chart, the Flow selector allows the user to choose an image to underlay the well circles, in the main image area. If images are available in the run data set, this selector is displayed; it lists the steps of the sequencing run where an image was taken by the instrument camera.

If any image is unavailable in the data set (*e.g.* it was removed), its appearance in the Flows list will depend on whether the current data set was opened from a raw instrument data set (R_ directory) or a data processing data set (D_ directory). If an R_ directory was opened, the image is grayed out in the Flows list. If a D_ directory was opened, the image will be enabled, but there will be a ⧉ icon next to it. This indicates that the raw image is unavailable, but a low-resolution image, stored in the .cwf file, is available. If you hover the mouse over the image, the tool tip will indicate Low Resolution, as shown in Figure 27.



Low Resolution: Thumbnails/00037.thumb.png
CWFs: 1
Directory: /home/adminrig/R_2010_01_22_20_38_50_nano11_dmadeux_200x_TACG_Nano_Config_12_11_2009_correct_noDSAir_moreA/D_2010_01_23_05_45_01_frontend2_fullProcessing/regions

Figure 27: Partial view of Flows, showing an image that is missing, but available in low-resolution format.

- The Flow tags have the following meanings:
  - SUB: Substrate buffer (background)
  - Apyrase##: flow used for Apyrase pulse calibration
  - PPI: Flow of PPi or ATP (illuminates all wells)
  - T, A, C, G: Nucleotide flow
- Clicking on a step displays the corresponding image in the main image area (and overview image area). After clicking in the Flows selector list, a user can use the keyboard arrow keys to navigate quickly up and down the list of images.
- When the Show All box is checked, all the images captured during the sequencing run are listed. If the box is not checked, only the images captured during nucleotide or PPi / ATP flows (*i.e.* excluding SUB and Apyrase flows) are listed.
- The wells themselves exist during all flows, so this selector has no effect on the display or coloring of the well circles. If the background image is a distraction, the user can disable the images by deselecting ‘Show image’ in the contextual menu that appears if a user right-clicks on the image.

3.5.2.4 Image area

This area provides a zoomable, scrollable view of the entire PicoTiterPlate device area, with well data overlaid on the camera image, if available. Overall, the image is comprised of four layers:

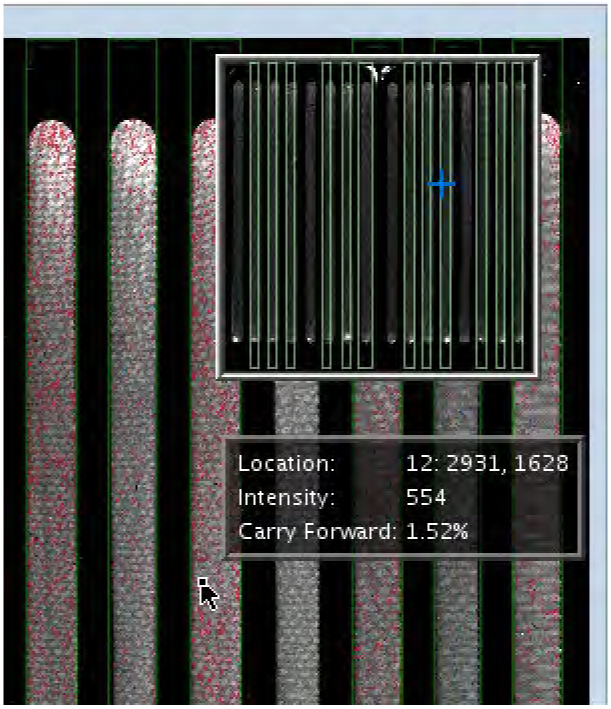
- The camera image selected; the displayed image is corrected to emphasize the well intensities and deemphasize the background intensities, for better visualization.
- Green rectangles delimit the bead loading regions of the sequencing run; the top of each region is labeled with the region name.
- Wells, per the options selected.
- Display control buttons and overview image

Settings chosen in the Options section are reflected in the image area, such that colored circles displayed at the x,y location of the wells identified during data processing reflect the Well Categories option, and the selected range of attribute values.

Wells remain in the same locations across all flows of a sequencing run, irrespective of what image is selected for display in the Flow selector (or of whether images are available at all in the data set).

Pausing or moving the mouse slowly over the main image area has the following effects (Figure 28):

- The well closest to the pointer is highlighted.
- The pointer position is marked by a blue '+' on the overview image area
- The Mouse Tracker box appears near the pointer, providing information about the image pixel under the pointer and the closest well. This includes the location under the cursor (region number and x,y coordinates), the corresponding pixel signal intensity value (for the flow selected, if images are available), and the attribute value for the closest (highlighted) well, in the category currently selected in the options. Note that setting the image brightness using the White Threshold slider does not change the intensity value of the pixel, only the brightness at which it is drawn on the screen.



Location: 12: 2931, 1628
Intensity: 554
Carry Forward: 1.52%

Figure 28: Wells Tab image area mouse tracker information and overview image.

Clicking on the image can have the following effects:

- A simple left-click selects the closest well for further action. A selected well is marked by an additional light blue circle around it.
- A double-click also applies to the closest well. This constructs a Well Flowgram (for wells that contain a library bead; TCAG or GACT key) or a Tri-Flowgram (for wells that contain a Control DNA bead, CATG or ATGC; see Section 7.1 for details), and opens the flowgram viewer to display it (see Sections 3.5.3 and 3.5.4 for a full description of well flowgrams and well tri-flowgrams, respectively).
- A right-click opens the contextual menu shown in Figure 29 and described below.
- A left-click-and-drag slides the image in the direction of the drag, revealing the part of the image that was just out of view.
- A right-click and drag zooms the image to the area circumscribed by the dragging action.

If a mouse wheel is available, rolling forward zooms in and rolling backward zooms out the image. The Magnification Slider control (see below) moves along with the mouse wheeling action.

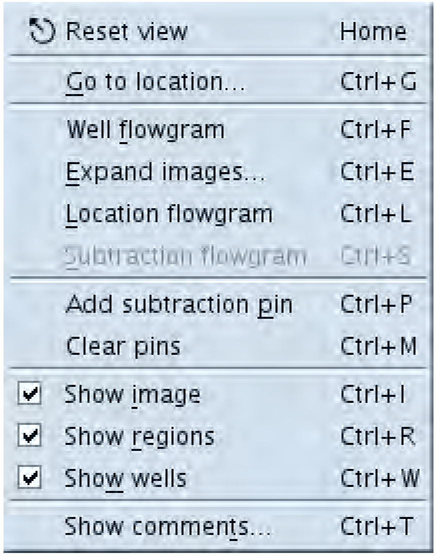
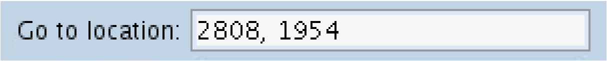


Figure 29: Contextual menu evoked by right-clicking the image area of the Wells tab.

The contextual menu of the image area offers the following actions:

- **Reset view**: resets the image to its default size and location, in the top-left corner position. The accelerator key is Home.
- **Go to location**: opens a data entry field just above the image area allowing a user to specify a location on the image (Figure 30), for immediate navigation when a user presses the ‘Enter’ key on the keyboard. Valid entries and their meanings are as follows:
    - a single positive integer scrolls the image to the top of the region number specified (*e.g.* 1-16).
    - two positive integers equal to or less than the number of the image pixels scroll the image to that x and y location. The two values can be separated with characters such as a space, comma, period or a slash.
    - a well ID or accession number scrolls the image to that location and selects the well

If the entry is not valid, an icon showing an ‘x’ in a red circle appears at the lower left corner of the field, and the image location does not change. The accelerator key is Ctrl-G.



Go to location: 2808, 1954

Figure 30: The 'Go to location' field, for the Image area of the Wells tab.

- **Well flowgram**: opens the ‘Well flowgram’ for the closest well. This will be a regular flowgram if the status of the well is ‘Passed filter’ (or ‘No key’, or ‘Failed’), and a tri-flowgram for ‘Control DNA’ wells. See sections 3.5.3 and 3.5.4 for a full description of Well flowgrams and Control DNA Tri-flowgrams, respectively. The accelerator key is Ctrl-F.
- **Expand images**: pre-expands .png camera images to the legacy .pif format to decrease image loading times (see Section 3.5.1). The accelerator key is Ctrl-E. This option is not available for runs from a GS Junior Instrument, which store .pif images.
- **Location flowgram**: opens the ‘Location flowgram’ for the image pixel on which the user clicked. See Section 3.5.5 for a full description of Location flowgrams. The accelerator key is Ctrl-L.
- **Subtraction flowgram**: this action requires that a pair of subtraction pins (see below) be first added on the image. Then, selecting this menu item opens the ‘Subtraction flowgram’ (Section 3.5.6) for the pair of image pixels selected by the pin, that is, for each flow of the sequencing run, the difference between the signals at the two ends of the pin. The accelerator key is Ctrl-S.
- **Add subtraction pin**: this action deposits a subtraction pin on the image (Figure 31). A subtraction pin is a visual indicator showing two image locations that can be used for calculating a subtraction flowgram. A small circle circumscribes the area that will be used for the calculation; a pentagon and a square around each of the location serve as ‘handles’ that can be dragged to position the circles at the locations of interest; a bar between the two squares shows which two locations are associated with other. The direction of the subtraction is: ‘circle in the pentagon’ (pin head) minus ‘circle in the square’ (pin tail). A user can add multiple subtraction pins on an image; the ‘active’ one, *i.e.* on which an eventual ‘Subtraction flowgram’ command would apply, is shown in blue; all the others are yellow. The accelerator key is Ctrl-P.
- **Clear pins**: removes all the subtraction pins currently in view. The accelerator key is Ctrl-M.
- **Show image**, **Show regions**, and **Show wells** check boxes: control whether the corresponding layers of the image area are displayed (checked) or not (unchecked).
- **Show comments** if enabled, will show all the run and region comments entered by the operator when the run was performed.

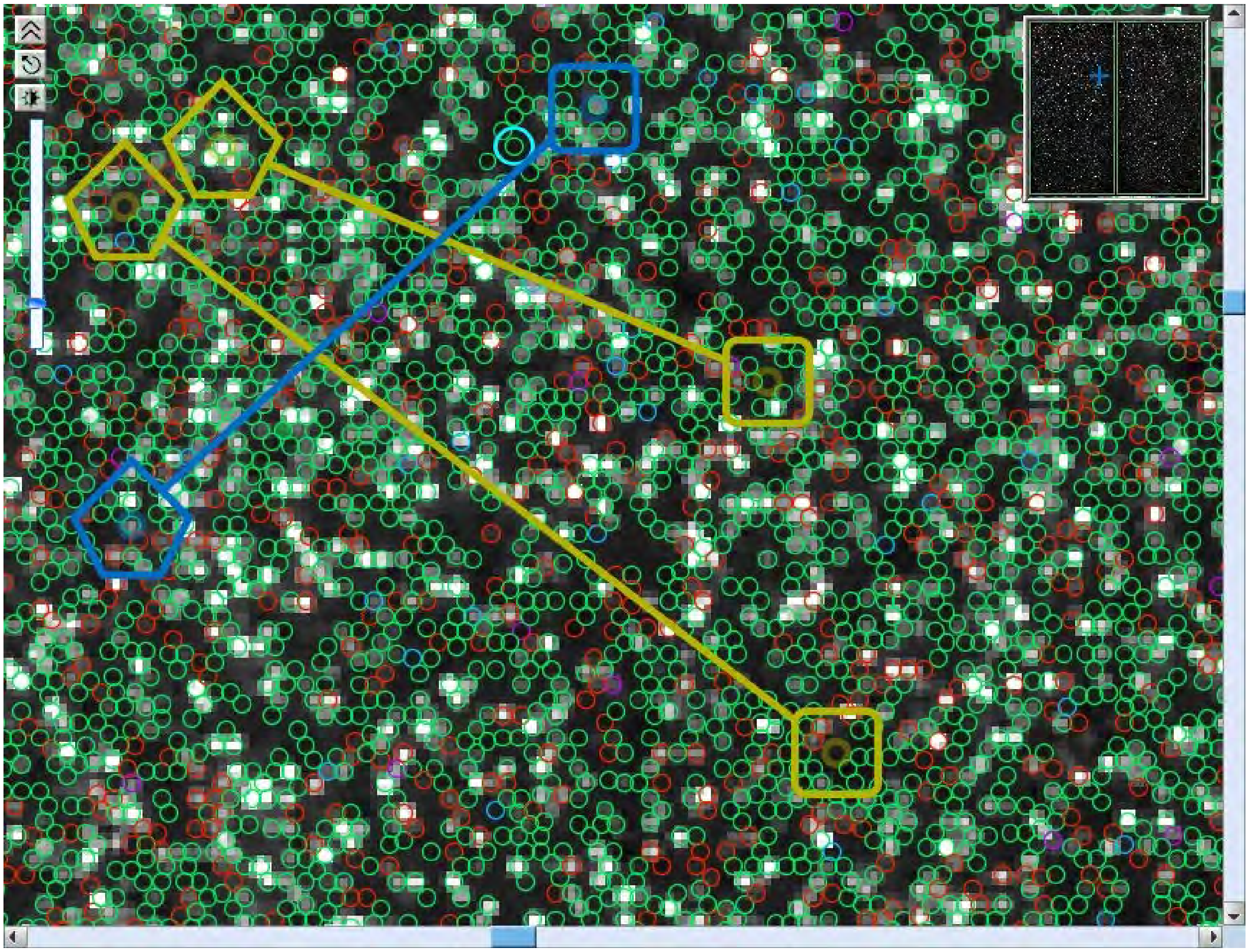
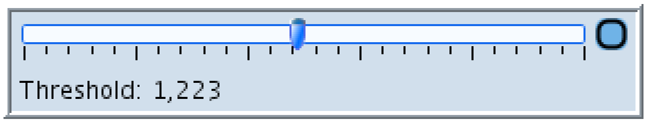


Figure 31: The image area of the Wells tab, with three subtraction pins. The one on which an eventual 'Subtraction flowgram' command would apply is in blue. The direction of the subtraction is: 'circle in the pentagon' (pin head) minus 'circle in the square' (pin tail).

The image area of the Wells tab has the following special navigation controls:

Button	Name – Description
︽	**Hide control**s – Hide the icons for the image area controls
︾	**Show controls** – Show the icons for the image area controls
↺	**Reset view** – Reset the view of the image to the default top left location and default magnification
✹	**Adjust white threshold –** Adjust the upper threshold for white values in the image. Clicking this button elicits a slider that can be used to set the white threshold (Figure 32). For the GS FLX+ Instrument, the range is from 0 to 2500 and the default setting is 1000. For the GS Junior Instrument, the range is from 0 to 8000, with a default setting of 2000. Clicking the blue button to the right of the slider resets the threshold to its default value. The white threshold value selected applies to all images.  Threshold: 1,223 **Figure 32: The Adjust White Threshold slider.** These default values can be changed by editing the GS Run Browser startup script, grRunBrowser.sh.
▯	**Magnification Slider –** Set the magnification of the image. Higher values reflect greater magnification (zoom in).
	Overview image: shows a small representation of the entire PicoTiterPlate device, with the image, if available, but without the wells, as an inset near the upper right corner of the image area. A blue '+' shows the current location of the mouse pointer, in the main image. Clicking (left or right) on the overview image scrolls the main image directly to the pixel on which was clicked.

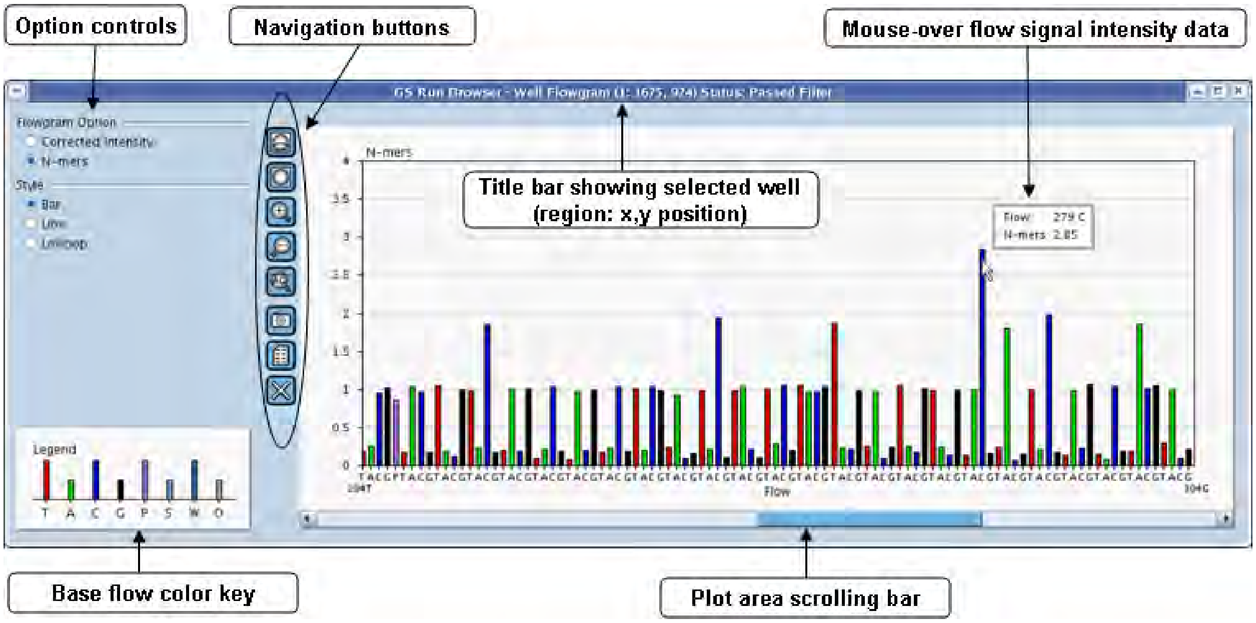
3.5.2.5 Average well density summary

This area provides the statistical averages for raw well density and key pass well density, calculated for each region and for the entire PicoTiterPlate device.

- For data sets processed with the GS Run Processor version 2.0.00 or later, these values are taken from the *[region]*.cwf files, if available. For older processed data sets, they are calculated by the GS Run Browser as the wells are loaded. For each well, the application counts the number of other wells in a local area around the well's center (~50 pixels); the raw well density calculation includes all wells, whereas the key pass density calculation omits wells that did not key pass (had a **No Key** status value). Note that the GS Run Processor and the GS Run Browser use different algorithms to perform these calculations, and will not yield identical well density values, but regions of fairly uniform density will have close results.
- An invisible divider is present between the image area and the well density summary table that can be dragged to adjust the part of the window allotted to these two areas.
- Image and Data Capture buttons are located to the left of the well density summary tables. The functions of these buttons are described in Section 3.3.4.

3.5.3 Well Flowgrams (for Wells Generating Library Reads)

Double-clicking the mouse in the main image area brings up the closest well's processed flowgram, which was generated by the GS Run Processor application and is used to determine the basecalled sequence for that well. In this flowgram view, all flows are shown rather than only the flows generating the trimmed sequence. The flowgram viewer for wells producing library reads (starting with the 'TCAG' or 'GACT' key), is shown below (Figure 33). The flowgram viewer for Control DNA wells is described in Section 3.5.4.



GS Run Browser - Well Flowgram (1: 1675, 974) Status: Passed Filter

Flowgram Option
○ Corrected Intensity
● N-mers

Style
● Bar
○ Line
○ Lollipop

Flow: 279 C
N-mers: 2.85

Figure 33: Well Signal Flowgram Plot.

Only one well flowgram window (or Control DNA well tri-flowgram window) can be open at a time; selecting the well flowgram action for another well on the main image area will replace any existing data in the well flowgram window. However, both a well flowgram window and a location or a subtraction flowgram window can be open at the same time (see Sections 3.5.3 through 3.5.6).

3.5.3.1 Well Flowgram Features and Functionalities

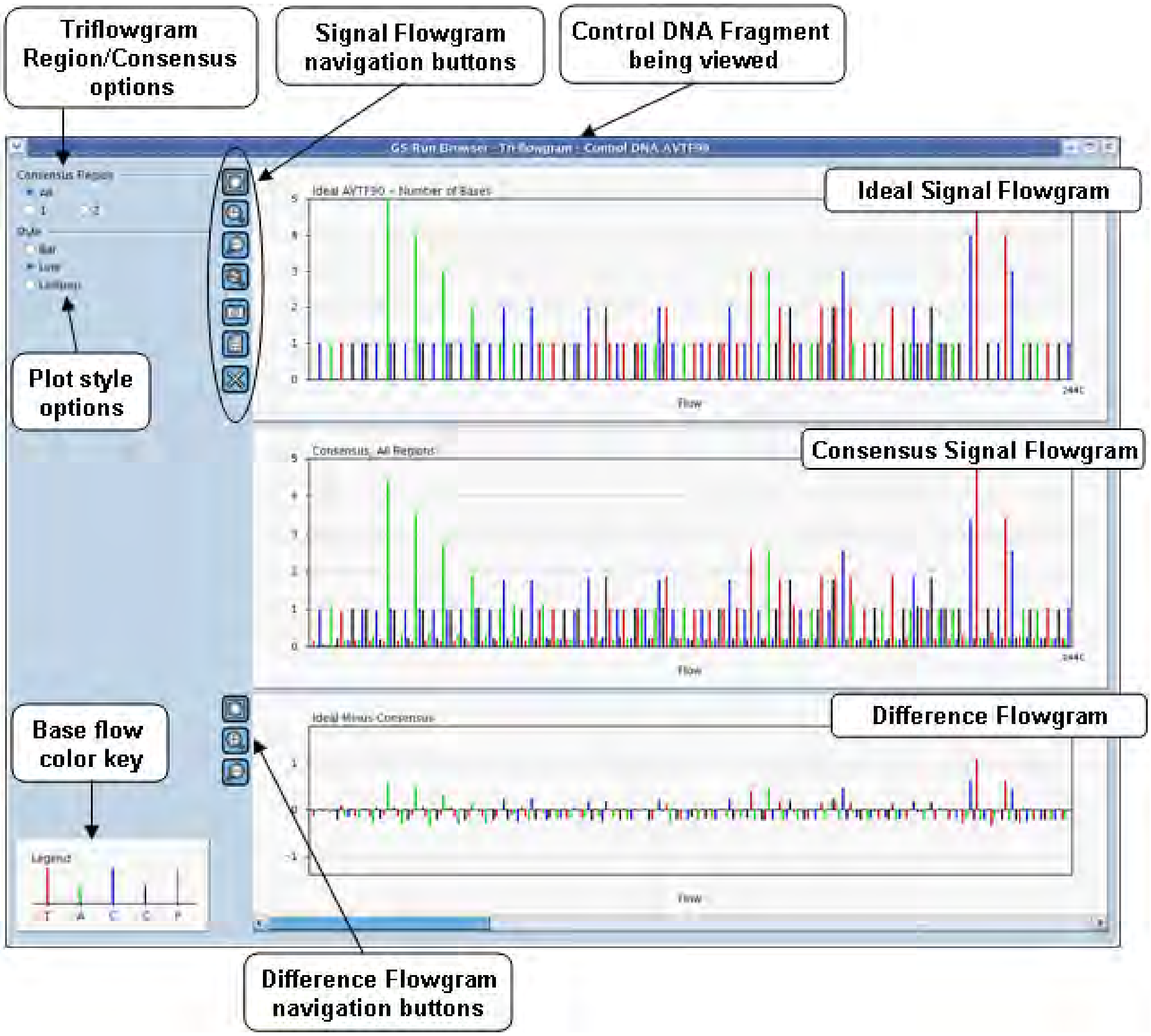
The well flowgram window has the following areas, with the functionalities described (see Figure 33):

- **Title bar**: The window title identifies the type of flowgram window (Well Flowgram), the PicoTiterPlate device region and the x,y location of the well, and the well's attribute value for the currently selected category.
- **Option controls**:
    - **Flowgram option**: Choice of the type of flowgram signal to display:
    - **Corrected Intensity** – Display the processed ('corrected') signal intensity as computed by the GS Run Processor application.
    - **N-mers** – Display each signal as an estimate of the number of bases extended at each flow (*i.e.*, the homopolymer length).
    - Choice of 3 plot **Styles**: Bars, Lines, or Lollipop. In all cases, the reagent flowed in each step is color-coded, per the legend. The lines and lollipop styles are narrower than the bar style, and will allow viewing more of a run without scrolling.
- **Navigation and data capture buttons**: All the buttons in this window have common functions: setting and adjusting the zoom level of the plot display, or allowing a user to save the data as a text file or snapshot image. See Section 3.3.4 for a description of the navigation and data capture button functions.
- **Plot display**:
    - The plot shares the common scrolling and zooming functions of the other GS Run Browser plots. See Section 3.3.3 for a description of the plot functions.
    - The default y-axis scale is set on the fly to values appropriate for visualizing the range of signals generated for the well flowgram displayed. The navigation buttons can be used to adjust this zoom level.
    - When the mouse pointer is over the plot, the mouse tracker shows the flow number, the reagent flowed, and the count/intensity for the flow under the pointer.
- **Signal legend**: This area shows the colors used to display the flowgram signals on the plot; each reagent is shown in a different color.

For data sets processed with the software version 1.1.03 or earlier, the nucleotide labels of the x-axis of the well flowgrams come from the imageLog.parse file. If this file is absent (*e.g.* if the D_ directory has been separated from its parent R_ directory), the software will attempt to infer what the flows were. If incorrect, this can be overridden by specifying the proper data set type using the drop down menu on the Overview tab (see Section 3.4.3).

3.5.4 Well Tri-Flowgrams (for Wells Generating Control DNA Reads)

When the selected well produces a Control DNA read (starting with the **'CATG'** or **ATGC**; see Section 7.1 for details), a 'tri-flowgram' plot appears (Figure 34). Tri-flowgrams allow the well's flowgram to be compared to an idealized flowgram of this Control DNA reference sequence, or to be compared to the consensus flowgram (see Section 3.8.2) generated from all the wells in the run that contain this Control DNA sequence.


Triflowgram Region/Consensus options
Signal Flowgram navigation buttons
Control DNA Fragment being viewed
GS Run Browser - Tri-flowgram - Control DNA AVTF90
Ideal Signal Flowgram
Ideal AVTF90 - Number of Bases
Consensus Signal Flowgram
Consensus, All Regions
Difference Flowgram
Ideal Minus Consensus
Flow
5
4
3
2
1
0
-1
244C
Plot style options
Base flow color key
Legend
T
A
C
G
F
Difference Flowgram navigation buttons

Consensus Region
● All
○ 1 ○ 2

Style
○ Bar
● Line
○ Lollipop

Figure 34: The tri-flowgram view of a Control DNA well (sequence AVTF90), at (region 2, position 2459, 877).

3.5.4.1 Well Tri-Flowgram Features and Functionalities

The tri-flowgram window contains two flowgram plots, plus a 'difference' flowgram showing the flow-by-flow difference between the top two flowgrams. This flowgram viewer has functionalities similar to the standard flowgram viewer, but with the following differences:

- Three flowgram plots are calculated per viewer window:
    - An 'ideal' flowgram constructed from theoretical values for the (known) sequence of the Control DNA fragment identified in the selected well. This is generated by taking the known nucleotide reference sequence and converting each homopolymer stretch into a corresponding signal (so, for example, 'AAA' becomes a signal of 3.0 in an A flow).
    - A 'consensus' flowgram calculated as a flow-by-flow average of all the wells that contained the same Control DNA sequence as the selected well.
    - The flowgram of the selected well.
- The **Options** area offers the following choices (specific to the tri-flowgram viewer):
    - Compare flowgrams:
        - **Ideal versus Well** – Display the idealized flowgram in the top plot, the selected well's flowgram in the middle plot, and the flow-by-flow difference of the two in the bottom plot.
        - **Consensus versus Well** – Display the consensus flowgram for the selected well's Control DNA sequence in the top plot, the selected well's flowgram in the middle plot, and the flow-by-flow difference of the two in the bottom plot.
        - **Ideal versus Consensus** – Display the ideal flowgram in the top plot, the consensus flowgram in the middle plot, and the flow-by-flow difference of the two in the bottom plot.
    - Consensus from:
        - **Well Region** – Average the signals of the reads (for this Control DNA sequence) only from the same region as the selected well.
        - **All Regions** – Average the signals of all the reads (for this Control DNA sequence) across the whole PicoTiterPlate device.

> Only the **N-mers** flowgram signals are shown, because the **Corrected Intensity** signals are not defined for ideal or consensus flowgrams or relevant in this type of comparison.

- The **Navigation and data capture buttons** and **flowgram display** have the following special functionalities (compared to the standard well flowgram view):
    - The scrolling and zooming of the three plots are correlated. All three plots scroll and zoom together along the x-axis, and the top two plots scroll and zoom together along the y-axis (the bottom 'difference' plot scrolls and zooms separately along the y-axis).
    - Since the bottom plot zooms separately along the y-axis, it has its own zoom buttons.
    - The horizontal bars between the plots allow for resizing the heights of the three plots.
    - The text file and snapshot image buttons will save a file containing the data or view for all three plots.

3.5.5 Location Flowgrams

Selecting the Location flowgram action from the right click menu from anywhere in the main image area brings up a 'raw' flowgram window for that location (Figure 35). The flowgram is constructed by computing, for each image in the sequencing run the average raw (non-corrected) signal intensity for the 9 pixels surrounding the selected location, and plotting these averages against the succession of reagent flows. (Note that this calculation does not give any consideration to the notion of 'wells'.)

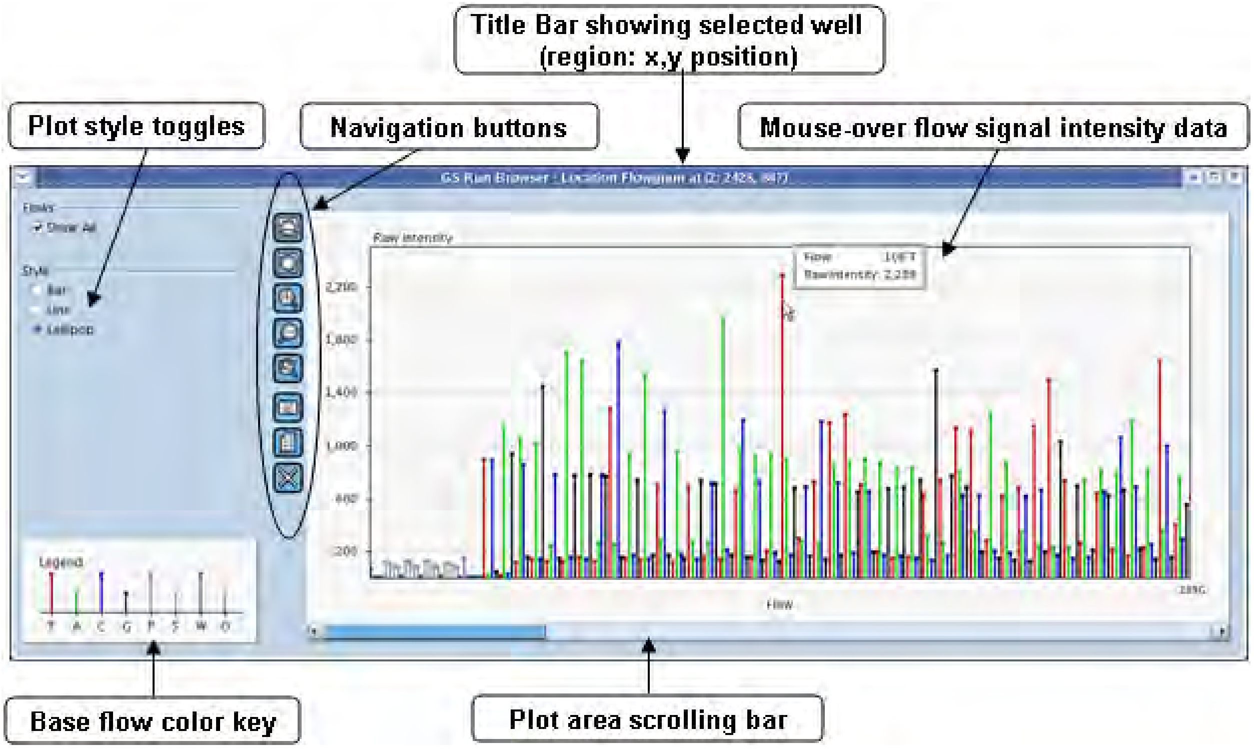


Figure 35: The Location flowgram of the Wells tab.

Only one Location flowgram window can be open at a time. Selecting the Location flowgram action for another location on the main image display will replace any existing data in a Location flowgram window. However, both a Location flowgram window and a Well flowgram or Tri-flowgram window can be open together.

3.5.5.1 Location Flowgram Features and Functionalities

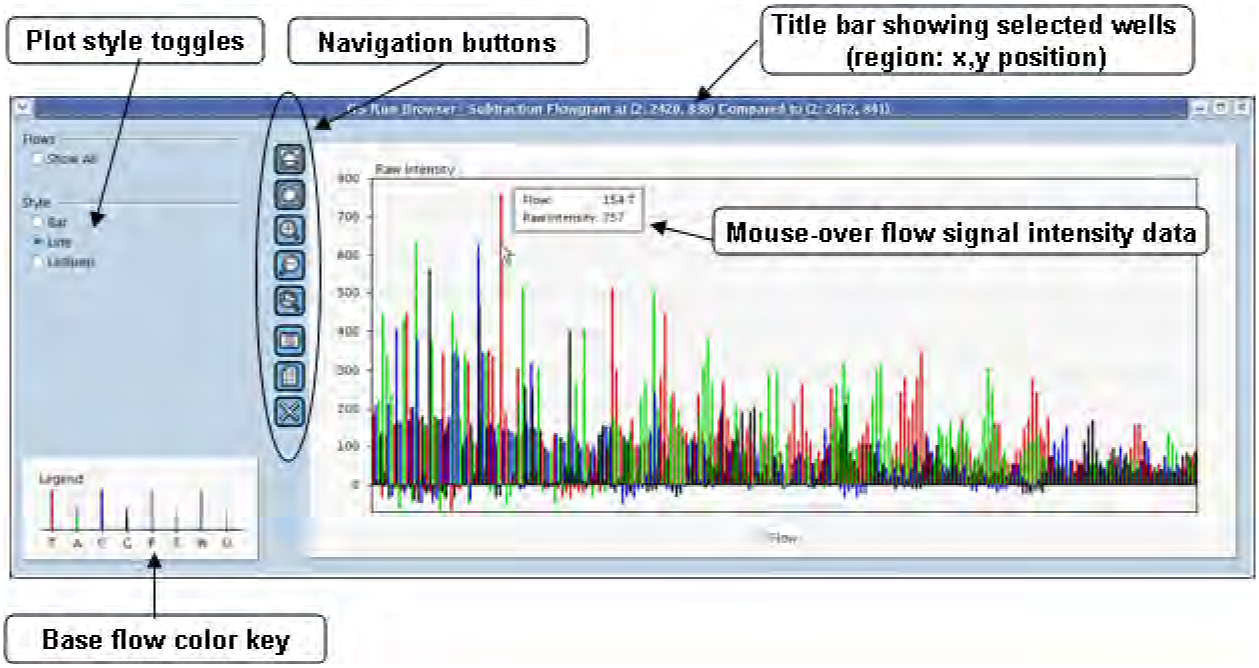
The Location flowgram window is very similar to the Well flowgram window (Section 3.5.3). The only four differences are:

- The window title bar identifies it as containing a Location flowgram.
- The flowgram plot displays raw signal intensities instead of corrected intensities.
- There isn't an option to display N-mers since these cannot be computed from non-normalized (corrected) data.
- A **Show all** check box allows a user to display (checked) or hide (unchecked) the flows that are neither nucleotides nor PPi/ATP flows.

3.5.6 Subtraction Raw Flowgrams

In addition to single location raw flowgrams, the GS Run Browser can generate a Subtraction (raw) flowgram from any two locations on the main image (Figure 36). Subtraction flowgrams are produced by right-clicking on a subtraction pin and selecting the Subtraction flowgram action, as described in Section 3.5.2.4. This is an advanced troubleshooting function that can be used in two distinct ways:

- to generate a raw flowgram that approximates a presumptive well's (in the pin head, pentagon) processed flowgram by subtracting the local background: when the signal intensity at the pin tail (square) location is lower than a set threshold (450 counts for 'A' flows and 150 counts for any other flow except PPi or ATP), the software will compute a simple local background subtraction.
- to compare the flowgrams from two well locations (or any two locations): when the signal intensity at the pin tail (square) location is above the set threshold, the subtraction flowgram simply compares the flowgrams from the two well locations (or any two locations). In this case, since different wells have different light-generating efficiencies, the calculation first normalizes the raw flowgrams at both locations by the average pixel intensity of their respective 'first ATP' signals, before doing the subtraction. This type of calculation is also always applied if the Subtraction flowgram is requested from an ATP image, irrespective of the pin tail signal.



GS Run Browser : Subtraction Flowgram at (2: 2420, 838) Compared to (2: 2452, 841)

Flows
□ Show All

Style
○ Bar
● Line
○ Lollipop

Flow: 154 T
Raw intensity: 757

Figure 36: The Subtraction flowgram of image pixels (region 2, position 2420, 838) minus (region 2, position 2452, 841). It is shown with a scale-to-fit y-axis, the lines style is chosen, and the wash steps are not displayed. The mouse pointer is over the data point of flow 154, a 'T' nucleotide flow that had substantially more signal in the pin head pixel than in the pin tail, as shown in the mouse tracker.

3.5.6.1 Subtraction Raw Flowgrams Features and Functionalities

A Subtraction flowgram window is identical to a Location flowgram window (Section 3.5.5), except that:

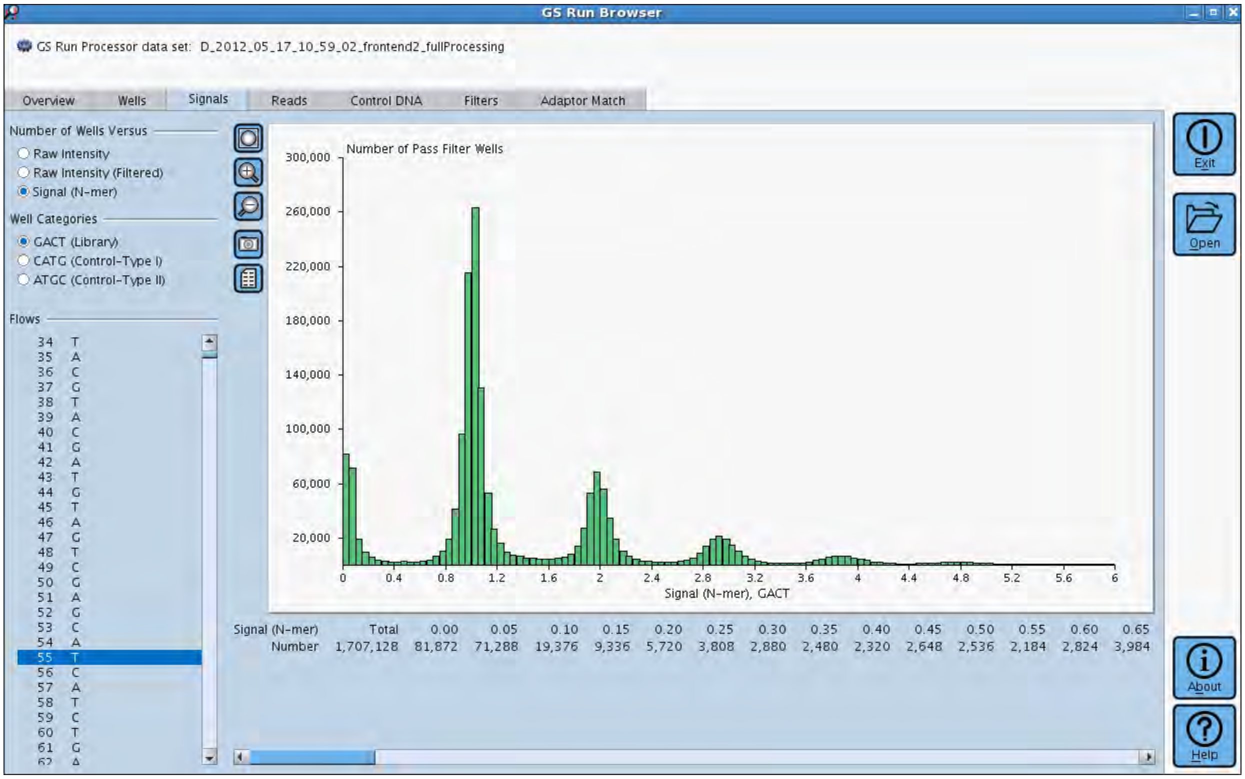
- The window title bar identifies it as containing a Subtraction flowgram, and references two pixel locations rather than one.
- the y-axis may contain negative values, since the subtraction may yield 'signals' below zero; this is akin to the 'difference plot' of a tri-flowgram window (Section 3.5.4).

3.6 The Signals Tab

The Signals tab (Figure 37) provides statistics on the distribution of the signals recorded during a run, and how they were interpreted by the data processing software in terms of the number of base incorporations at a given signal intensity. The tab displays either library well signals (sequencing key TCAG or GACT) or Control DNA well signals (sequencing key CATG or ATGC; see Section 7.1 for details), in one of the following three modes:

- **Raw Intensity**: the signal intensities as found in the *region*.cwf (or *region*.wells) files, using all the key passed wells for the library or Control DNA wells.
- **Raw Intensity (Filtered)**: the same signal intensity values, but using only the wells that both key passed and passed all quality filters.
- **Signals (N-mer)**: the basecalling estimates for the signal intensities, where the Raw Intensity (Filtered) signal values are converted into the estimated numbers of nucleotides that were incorporated in each well during the flow selected (for filter pass wells only).

The information displayed on this tab is not normally used for the evaluation of a sequencing run, but it can prove useful as an advanced tool to troubleshoot problems.



GS Run Processor data set: D_2012_05_17_10_59_02_frontend2_fullProcessing

Number of Wells Versus
○ Raw Intensity
○ Raw Intensity (Filtered)
● Signal (N-mer)

Well Categories
● GACT (Library)
○ CATG (Control-Type I)
○ ATGC (Control-Type II)

Flows
34 T
35 A
36 C
37 G
38 T
39 A
40 C
41 G
42 A
43 T
44 G
45 T
46 A
47 G
48 T
49 C
50 G
51 A
52 G
53 C
54 A
55 T
56 C
57 A
58 T
59 C
60 T
61 G
62 A

Signal (N-mer)	Total	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65
Number	1,707,128	81,872	71,288	19,376	9,336	5,720	3,808	2,880	2,480	2,320	2,648	2,536	2,184	2,824	3,984

Figure 37: The Signals tab shows a plot of the number of wells that produce final (passed filter) library reads and those reads interpreted by the data processing software as the extension of a given number of nucleotides.

For the Signals tab to be displayed, the *region*.cwf (or *region*.wells) file(s) located in the ‘regions’ sub-folder of the ‘D_’ data analysis folder of the sequencing run must exist. If no *region*.cwf (or *region*.wells) files exist for the run, the Signals tab will be unavailable.

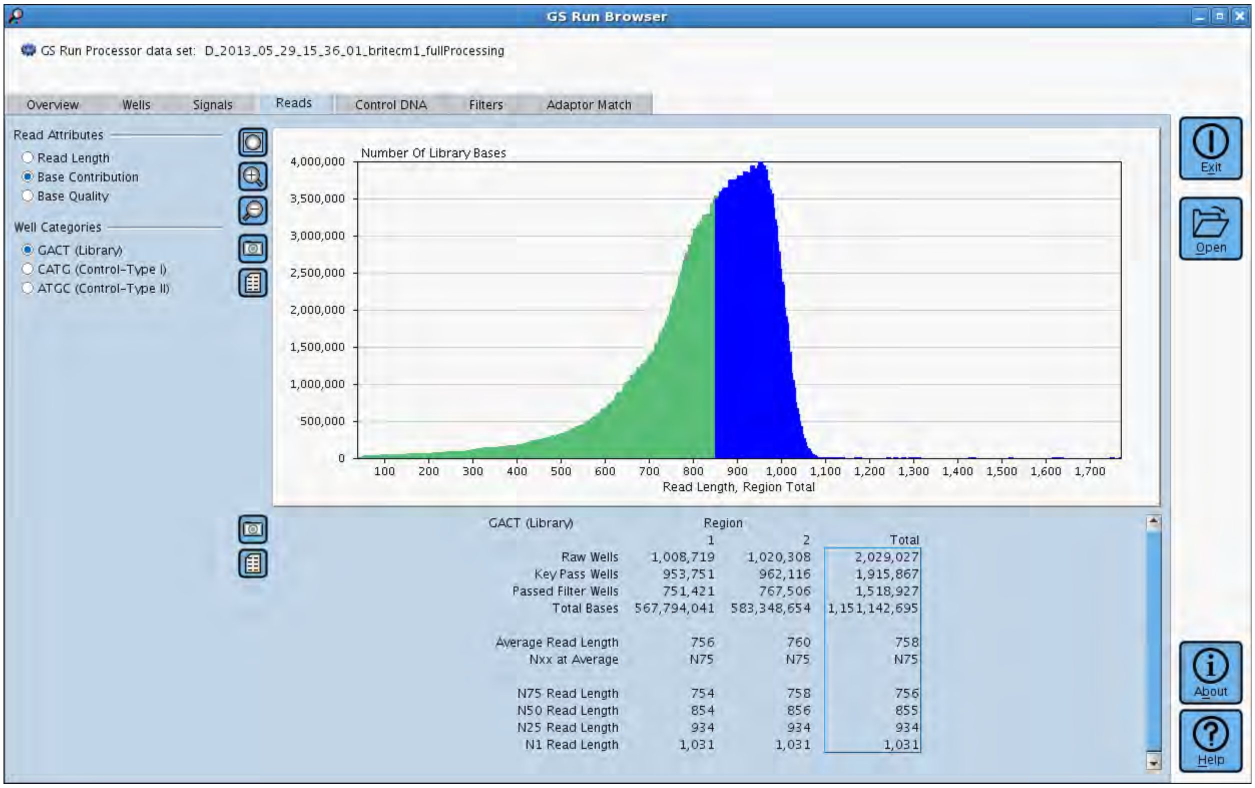
3.6.1 Signals Tab Features and Functionalities

The Signal tab contains many features with the functionalities described below (see Figure 37).

- The **Options** area (on the top left area of the tab) gives a choice of 6 plot types (each data set is available *for each flow*), broken down as follows:
  - **Number of Wells Versus:** For the well type selected (Library or Control DNA), and the flow selected, display the number of wells (y-axis) as a function of one of the following (x-axis):
    - **Raw Intensity** for all key pass wells
    - **Raw Intensity (Filtered)** for only filter pass (always also key pass)
    - **Signal (N-mer)** equivalent of the intensity values from the **Raw Intensity (Filtered)** signals, converted into number of nucleotides extended in the flow (*i.e.* 'homopolymer length', as estimated by the basecaller). This is the default view for library wells. Unless an anomaly is present in the Signal (N-mer) plot, it is unlikely that the Raw Intensity plots will need to be viewed.
  - **Well Categories**: For the signal type selected (Raw Intensity, Raw Intensity – filtered, or Signal N-Mer), display the data for the wells with the following keys:
    - **GACT (Rapid Library) TCAG (Standard library)** – Use the key passed wells identified as library wells.
    - **CATG** and/or **ATGC (control)** – Use the key passed wells identified as Control DNA wells. See Section 7.1 for details on Control DNA keys.
  - **Flows**: Located below the Well categories section, the Flows selector allows the user to select a flow from the sequencing run for viewing.
    - Only the nucleotide and PPi or ATP flows are listed (Since signal statistics for wash flows are irrelevant, there is no 'Show all' check box).
    - After clicking in the Flows selector, the keyboard arrow keys can be used to navigate quickly up and down this list.
- **Navigation and data capture buttons**: All the buttons on this tab have common functions: setting and adjusting the zoom level of the main image display or allowing the user to save the data as a text file or snapshot image. See Section 3.3.4 for a description of the navigation and data capture button functions.
- **Plot display**: The plot shares the common scrolling and zooming functions of the other GS Run Browser plots (see Section 3.3.3), with the following exceptions:
  - The default scales for the x and y axes have been set to fixed values that typically provide a useful view of the data. After first clicking the Flow selector, the keyboard arrow keys can be used to quickly scan from flow to flow; this constitutes a very useful troubleshooting tool for evaluating the signal distributions.
  - The bar widths (*i.e.* histogram 'bin sizes') are 100.0 for Raw Intensity plots, and 0.05 for Signal (N-mer) plots. These bar widths are fixed.
  - When the mouse pointer is over the plot, the mouse tracker shows the x-axis signal value and the number of wells having that signal value, in the selected flow.
- Summary:
  - This displays a text version of the data shown in the plot, listing the height of each bar in the plot, with zero height bars omitted.
  - An invisible divider is present between the plot area and the well density summary table that can be dragged to adjust the part of the window allotted to these two areas.

3.7 The Reads Tab

The Reads tab (Figure 38) provides statistics on the read length and read quality observed during the sequencing run for either the sample library reads (sequencing key TCAG or GACT) or the Control DNA reads (sequencing key CATG or ATGC; see Section 7.1 for details). It displays a plot of the read length or quality data either for a selected PTP Region or across the whole PTP device, and a table of region-by-region and whole PTP device statistics of related data.



GACT (Library)	Region 1	2	Total
Raw Wells	1,008,719	1,020,308	2,029,027
Key Pass Wells	953,751	962,116	1,915,867
Passed Filter Wells	751,421	767,506	1,518,927
Total Bases	567,794,041	583,348,654	1,151,142,695
Average Read Length	756	760	758
Nxx at Average	N75	N75	N75
N75 Read Length	754	758	756
N50 Read Length	854	856	855
N25 Read Length	934	934	934
N1 Read Length	1,031	1,031	1,031

Figure 38: The GS Run Browser application's Reads tab, showing the base contribution profile, which displays the number of bases contributed by reads of any given read length.

For the Reads tab data to be displayed, the *region*.cwf files must be present in the data processing directory of the selected run.

The base contribution profile displays the number of bases contributed by reads of each given length, which is the most appropriate way to view asymmetric read length distributions such as those obtained from shotgun sequencing runs. The base contribution percentile is the percentage of bases contributed by a specific read length or longer. The Nxx read length is the read length above which the indicated percentage of bases are contributed. For example 50% of bases are contributed by reads of the N50 read length or longer. The N50 read length provides a more consistent and reliable measure of read length than either average or modal read length.

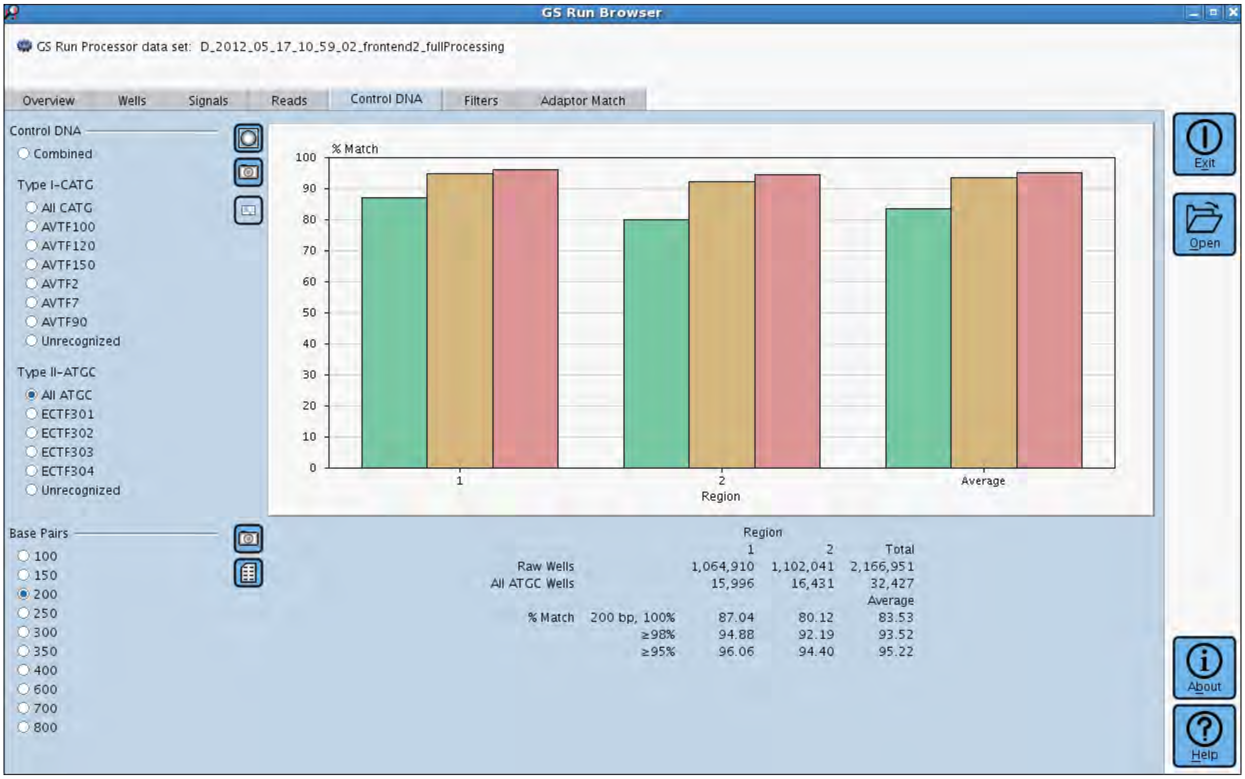
3.7.1 Reads Tab Features and Functionalities

The Reads tab contains many features with the functionalities described below (see Figure 38):

- The **Options** area (on the top left area of the tab):
    - **Read Attributes**: choice of data type to display:
        - Read Length
        - Base Contribution
        - Base Quality
    - **Well Categories**: choice of reads to display (only passed filter reads):
        - GACT (rapid library) or TCAG (standard library)
        - CATG and/or ATGC (control) - see Section 7.1 for details
- **Navigation and data capture buttons**: All the buttons on this tab have common functions: setting and adjusting the zoom level of the main image display or allowing a user to save the data as a text file or snapshot image. See Section 3.3.4 for a description of the navigation and data capture button functions.
- **Plot display**:
    - Graphic display of the data, as selected in the Options area for a single region or for the whole PTP device, as selected on the Summary.
    - For read length, the 'count' (y-axis) corresponds to the number of passed filter reads at a given length. For base contribution, the 'count' (y-axis) corresponds to the number of bases contributed by read at a given length. For read quality, the 'count' (y-axis) corresponds to the number of bases at a given quality score (PHRED equivalent; see Section 6.6).
    - The plot shares the common scrolling and zooming functions of the other GS Run Browser plots. See Section 3.3.3 for a description of the plot functions.
    - When the mouse pointer is over the Read Length plot, the mouse tracker shows the x-axis read length and the number of reads having that length, for the selected region or the entire PTP device.
    - When the mouse pointer is over the Base Contribution plot, the mouse tracker shows the x-axis read length, the number of bases contributed by reads of that length, and the percentile (Nxx) or percentage of bases contributed by reads of that length or longer.
- **Summary**:
    - This displays a text version of the data shown in the plot, for each of the regions and for the entire PTP device, plus additional intermediate data.
    - A blue rectangle identifies the data (single region or 'Total' column) currently displayed on the plot. When the mouse is moved over the data in another column, the highlight moves to this column; left-clicking the mouse selects the highlighted column (blue rectangle) for display in the plot.

3.8 The Control DNA Tab

The Control DNA tab (Figure 39) displays accuracy results for the Control DNA Beads that were spiked in the sequencing reaction, when available (see the *Sequencing Method Manual* for details about this procedure). Specifically, the Control DNA tab reports the percentage of Control DNA reads that match their reference sequence at 95%, 98%, and 100% accuracy, in each region of the PicoTiterPlate device. It also allows a user to view the 'consensus flowgram' for the reads from each Control DNA sequence (Section 3.8.2).



GS Run Processor data set: D_2012_05_17_10_59_02_frontend2_fullProcessing

Control DNA
○ Combined

Type I-CATG
○ All CATG
○ AVTF100
○ AVTF120
○ AVTF150
○ AVTF2
○ AVTF7
○ AVTF90
○ Unrecognized

Type II-ATGC
● All ATGC
○ ECTF301
○ ECTF302
○ ECTF303
○ ECTF304
○ Unrecognized

Base Pairs
○ 100
○ 150
● 200
○ 250
○ 300
○ 350
○ 400
○ 600
○ 700
○ 800

		Region		
		1	2	Total
Raw Wells		1,064,910	1,102,041	2,166,951
All ATGC Wells		15,996	16,431	32,427
				Average
% Match	200 bp, 100%	87.04	80.12	83.53
	≥98%	94.88	92.19	93.52
	≥95%	96.06	94.40	95.22

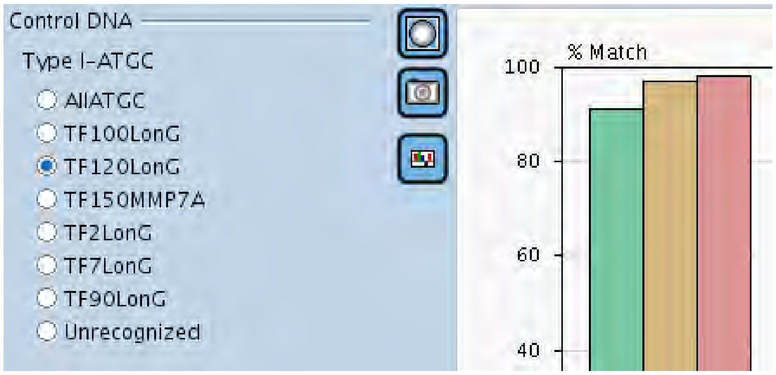
Figure 39: The GS Run Browser application's Control DNA tab.

For the Control DNA tab data to be displayed, the *region*.cwf files must be present in the Analysis sub-directory of the sequencing run.

3.8.1 Control DNA Tab Features and Functionalities

The Control DNA tab contains many features with the functionalities described below (see Figure 39):

- The **Options** area (on the top left area of the tab):
  - **Control DNA**: choice of data to display:
    - All Control DNA sequences included in the run
    - Any of the specific Control DNA sequences included in the run
    - Unrecognized reads, *i.e.* reads which begin with the Control DNA sequencing key (CATG or ATGC; see Section 7.1 for details) but do not match any of the corresponding Control DNA reference sequences
  - **Base Pairs**: choice of length over which to calculate match, *i.e.* show % match from the first base after the key up to this nucleotide in the reads. Only the options relevant to the read length of the data set are displayed.
  - A button to display the consensus flowgrams for each of the specific Control DNA sequences (as selected) in a **tri-flowgram viewer** (Figure 40). See Section 3.8.2 for a complete description of the Control DNA consensus flowgram view. This button is available only if the current selection is one of the Control DNA sequences; and grayed out if 'all' or 'unrecognized' are selected.



Control DNA

Type I-ATGC

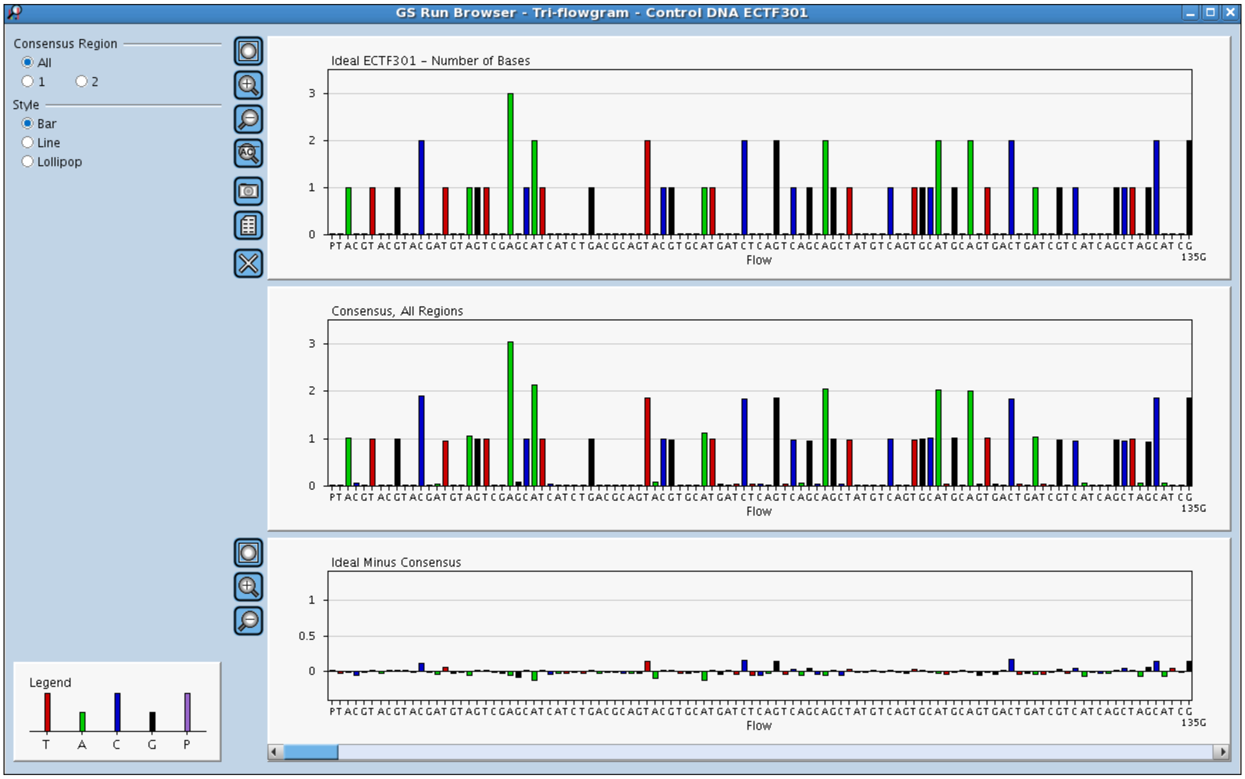
○ AllATGC
○ TF100LonG
● TF120LonG
○ TF150MMP7A
○ TF2LonG
○ TF7LonG
○ TF90LonG
○ Unrecognized

Figure 40: The 'Display Consensus Flowgram' button of the Control DNA tab, available when one of the Control DNA sequences is selected.

- **Navigation and data capture buttons**: Except for the 'Display Consensus Flowgram' button described above, all the buttons on this tab have common functions, allowing a user to save the data as a text file or snapshot image. See Section 3.3.4 for a description of the navigation and data capture button functions.
- **Plot display**:
    - Graphic display of the % match data for all key passed Control DNA reads, per the options selected for the individual regions of the PicoTiterPlate device and as an aggregate average.
    - This data is presented at 3 levels of accuracy:
        - 100%: The percent of key passed Control DNA reads that *exactly* matched the first N-bases of the corresponding known reference sequence.
        - e 98%: The percent of key passed Control DNA reads that had no more than 2% basecalling differences in the first N-bases compared to their known reference sequence.
        - e 95%: The percent of key passed Control DNA reads that had no more than 5% basecalling differences in the first N-bases compared to their known reference sequence.
    - The plot shares the common scrolling and zooming functions of the other GS Run Browser plots See Section 3.3.3 for a description of the plot functions.
    - When the mouse pointer is over the plot, the mouse tracker shows the x-axis value, and the % of reads among the Control DNA sequences selected that matched their respective reference sequences at 100, 98, and 95% accuracy over the selected length, in the region under the pointer or averaged over the entire PTP device.
- **Summary**: This area displays the number of Raw Wells and Control DNA reads for the Control DNA species specified; it also displays the percentage of matches between these reads and their reference sequence at the read length specified, for each region of the PicoTiterPlate device and as an aggregate average.

3.8.2 Control DNA Consensus Flowgrams

Clicking the ‘Open the flowgram’ button (in the Options area of the Control DNA tab), when one of the Control DNA sequences is selected, brings up the consensus flowgram for that sequence (Figure 41). A consensus flowgram is the flowgram constructed by averaging, for each nucleotide flow, the read flowgram signals of the reads identified as that reference sequence. This is presented as part of a tri-flowgram, along with the ‘ideal’ flowgram for that Control DNA sequence and the ‘difference’ flowgram, in a manner similar to the tri-flowgram of a Control DNA well, described before (Section 3.5.4).


GS Run Browser - Tri-flowgram - Control DNA ECTF301
Ideal ECTF301 – Number of Bases
3
2
1
0
PTACGTACGTACGATGTAGTCGAGCATCATCTGACGCAGTACGTGCATGATCTCAGTCAGCAGCTATGTCAGTGCATGCAGTGACTGATCGTCATCAGCTAGCATCG
135G
Flow
Consensus, All Regions
3
2
1
0
PTACGTACGTACGATGTAGTCGAGCATCATCTGACGCAGTACGTGCATGATCTCAGTCAGCAGCTATGTCAGTGCATGCAGTGACTGATCGTCATCAGCTAGCATCG
135G
Flow
Ideal Minus Consensus
1
0.5
0
PTACGTACGTACGATGTAGTCGAGCATCATCTGACGCAGTACGTGCATGATCTCAGTCAGCAGCTATGTCAGTGCATGCAGTGACTGATCGTCATCAGCTAGCATCG
135G
Flow
Legend
T
A
C
G
P

Consensus Region
● All
○ 1 ○ 2

Style
● Bar
○ Line
○ Lollipop

Figure 41: The tri-flowgram view of the AVTF90 Control DNA sequence, showing the consensus flowgram for both regions of the PicoTiterPlate device.

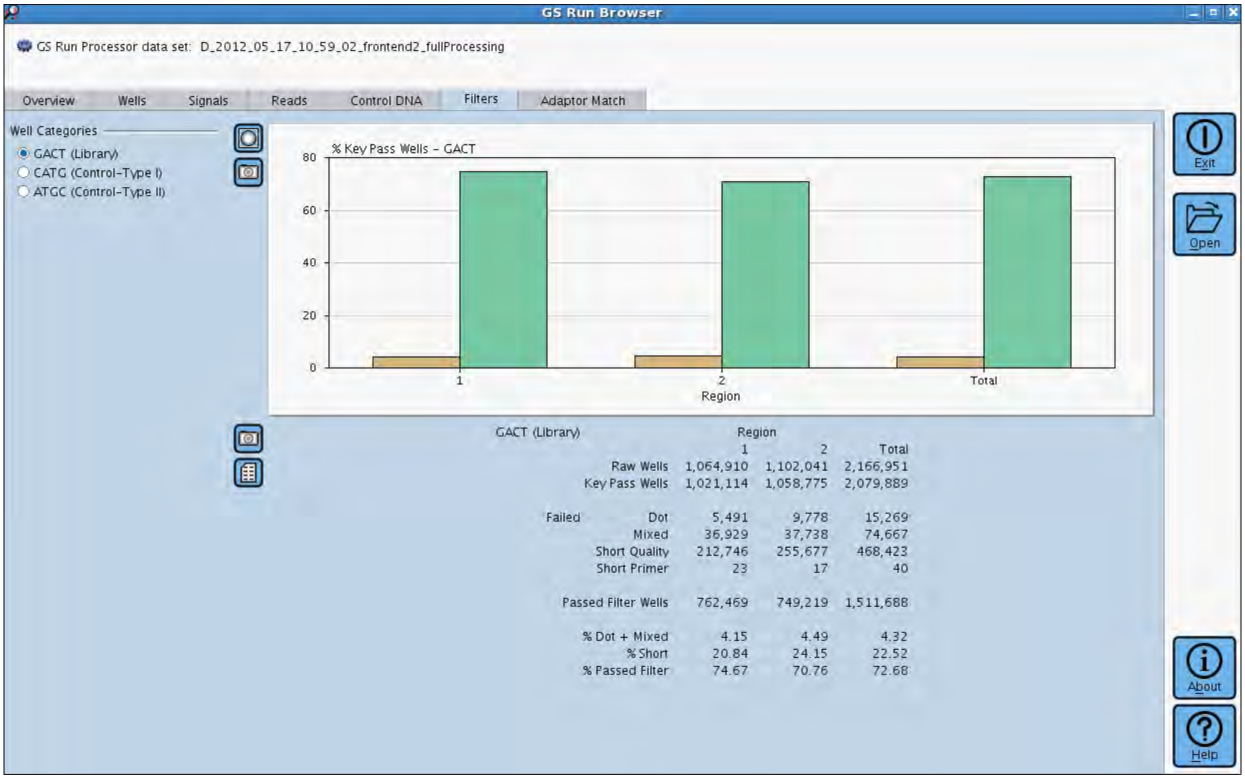
3.8.2.1 Control DNA Consensus Flowgram Features and Functionalities

The tri-flowgram view of a Control DNA consensus sequence is similar to that of an individual Control DNA read (see Section 3.5.4) and is described below.

- The **Options** area provides the following choices:
  - **Consensus Region**: choice of what PicoTiterPlate device **region** data to display:
    - **All** – Show the consensus flowgram formed by averaging all the read flowgrams across the entire PicoTiterPlate device, for the Control DNA sequence specified.
    - Individual region – Show the consensus flowgram of the reads from an individual region of the PicoTiterPlate device, for the Control DNA sequence specified.
  - **Style**: choice of Bars, Lines, or Lollipop plot styles. In all cases, the reagent flowed in each step is color-coded, per the legend. The lines and lollipop styles are narrower than the bar style, and will allow viewing of more of a run without scrolling.
- Navigation and data capture buttons:
  - All the buttons in this window have common functions: setting and adjusting the zoom level of the plot display or allowing a user to save the data as a text file or snapshot image. See Section 3.3.4 for a description of the navigation and data capture button functions.
  - The text file and snapshot image buttons will save a file containing the data or view for all three plots.
  - The bottom plot contains separate zooming buttons because the scrolling and zooming of the plots are tied together except for the y-axis of the bottom plot (see below).
- Plot display:
  - The top plot displays the idealized flowgram for the Control DNA sequence selected, generated by taking the known nucleotide reference sequence and converting each homopolymer stretch into a corresponding signal (*e.g.* ‘AAA’ becomes a signal of 3.0 in the next A flow). The middle plot displays the consensus flowgram calculated as a flow-by-flow average of all the reads matching this Control DNA sequence in the region(s) selected; and the bottom plot displays the flow-by-flow differences between the top two plots.
  - The plot shares the common scrolling and zooming functions of the GS Run Browser plots. See Section 3.3.3 for a description of the plot functions.
  - As with other tri-flowgrams, the scrolling and zooming of the three plots are tied together. All three plots scroll and zoom together along the x-axis, and the top two plots scroll and zoom together along the y-axis as well (the bottom ‘difference’ plot scrolls and zooms separately along the y-axis).
  - Since the bottom plot zooms separately, it has its own zoom buttons.
  - The horizontal bars between the plots allow for adjusting the heights of the three plots.
  - When the mouse pointer is over the plot, the mouse tracker shows the flow number, the reagent flowed and the N-mers count for the flow under the pointer.
- **Signal legend**: This area shows the colors used to display the flowgram signals on the plot; each reagent is shown in a different color.

3.9 The Filters Tab

The Filters tab (Figure 42) provides statistics on the read quality filters of the GS Run Processor application used to process the data of the sequencing run. The statistics are displayed, per region and in aggregate, for either the sample library reads (sequencing key **TCAG** or **GACT**) or the Control DNA reads (sequencing key **CATG** or **ATGC**; see Section 7.1 for details). The filtering statistics displayed are: **Key Pass**, **Dot, Mixed, Short Quality, and Short Primer**. The application of these filters is discussed in detail in Section 1.3.2.



GACT (Library)		Region		
		1	2	Total
	Raw Wells	1,064,910	1,102,041	2,166,951
	Key Pass Wells	1,021,114	1,058,775	2,079,889
Failed	Dot	5,491	9,778	15,269
	Mixed	36,929	37,738	74,667
	Short Quality	212,746	255,677	468,423
	Short Primer	23	17	40
	Passed Filter Wells	762,469	749,219	1,511,688
	% Dot + Mixed	4.15	4.49	4.32
	% Short	20.84	24.15	22.52
	% Passed Filter	74.67	70.76	72.68

Figure 42: The GS Run Browser application's Filters tab showing the results of the various quality filters for the sample library wells.

For the Filter tab data to be displayed, the *region*.cwf files must be present in the data processing directory of the selected run.

3.9.1 The Filters Tab Features and Functionalities

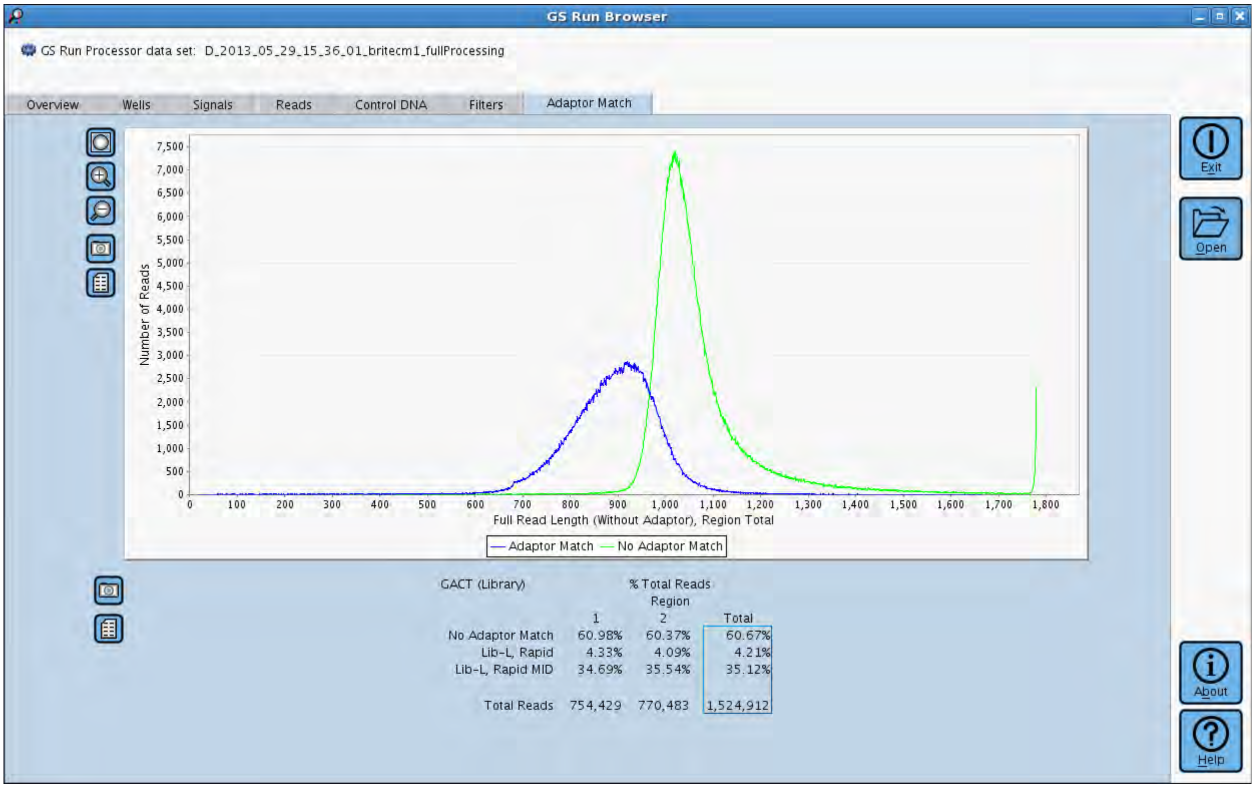
The Filter tab contains the features with the functionalities described below (see Figure 42):

- The **Options** area (on the top left area of the tab):
  - **Well Categories**: choice of reads to display:
    - GACT (rapid library) TCAG (standard library), or
    - CATG and/or ATGC (control) - see Section 7.1 for details on Control DNA keys
- **Navigation and data capture buttons**: All the buttons on this tab have common functions: setting and adjusting the zoom level of the main image display or allowing the user to save the data as a text file or snapshot image. See Section 3.3.4 for a description of the navigation and data capture button functions.
- **Plot display**:
  - The plot shows the number of reads that passed all filters and the number of reads that failed either the dot or the mixed filter (or both); these data are calculated as a percentage of key pass reads, for the sequencing key selected.
  - The plot shares the common scrolling and zooming functions of the GS Run Browser plots. See Section 3.3.3 for a description of the plot functions.
  - When the mouse pointer is over the plot, the mouse tracker shows the x-axis value, and the % of reads in the selected well category that passed all filters and that failed the dot and/or mixed filters, in the region under the pointer or averaged over the entire PTP device.
- Summary:
  - The Summary area shows all filter data, per PicoTiterPlate device region and in aggregate, including:
    - Number of **Raw Wells**
    - Then, for the key selected, the **number of reads** that:
      - passed key
      - failed each filter
      - passed all filters
    - and, for the key selected, the **percentage of key pass reads** that:
      - failed either or both the ‘dot’ or the ‘mixed’ filters (combined)
      - failed either or both of the ‘short’ filters (combined)
      - passed all filters

3.10 The Adaptor Match Tab

The Adaptor Match tab provides statistics on the number of 3'-adaptor matches found by the adaptor match tool in the basecaller (Section 1.3.3.3). It shows the length distribution for untrimmed library fragments up to (but not including) any identified 3'-adaptor sequence. Reads that sequenced all the way through to an identified adaptor sequence are displayed in blue (Adaptor Match), and reads where no adaptor was found are displayed in green (No Adaptor Match). The statistics are displayed, per region and in aggregate, for sample library reads (sequencing key **TCAG** or **GACT**). The Adaptor Match length histogram may be used to infer the approximate untrimmed library fragment length (shotgun or amplicon), as well as the fraction of library reads that were short enough to be fully sequenced.

Figure 43 shows the Adaptor Match untrimmed length histogram for shotgun data, showing the distribution of library fragment sizes with and without an adaptor match.

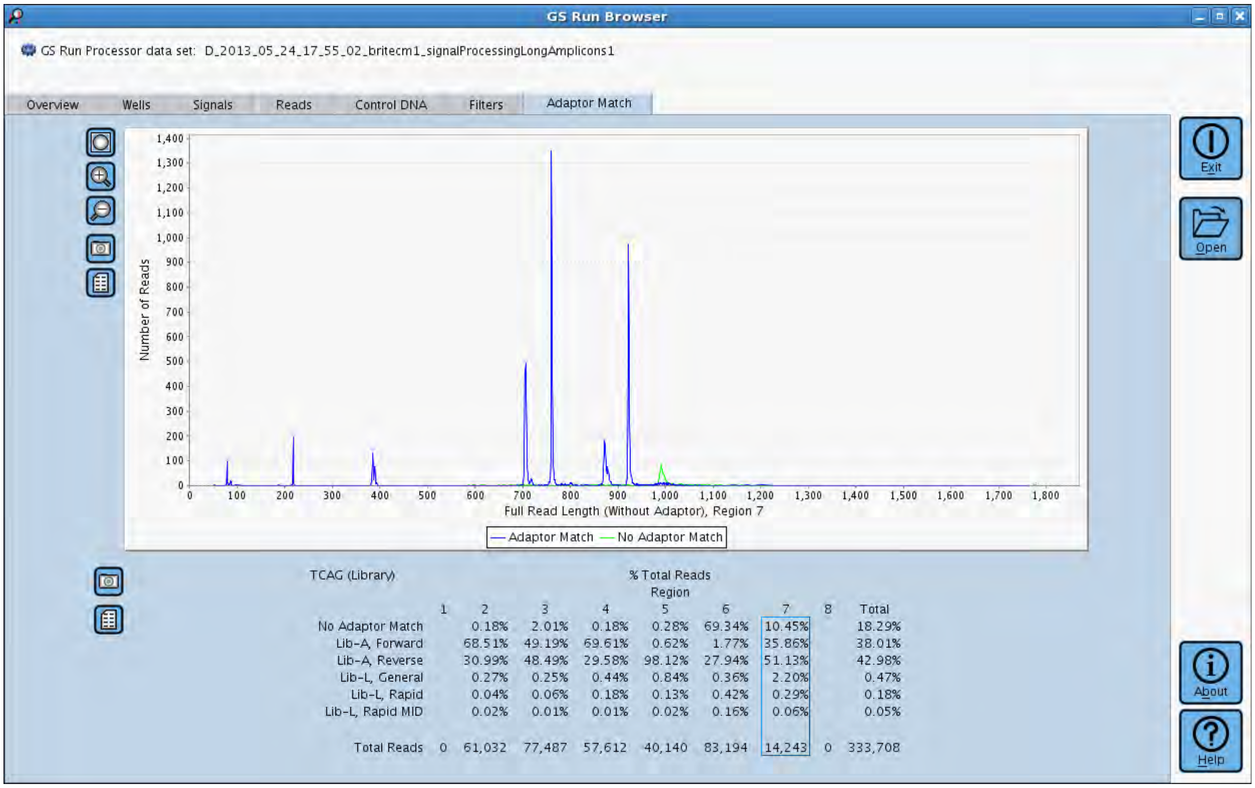


GACT (Library)	% Total Reads		
	Region		
	1	2	Total
No Adaptor Match	60.98%	60.37%	60.67%
Lib-L, Rapid	4.33%	4.09%	4.21%
Lib-L, Rapid MID	34.69%	35.54%	35.12%
Total Reads	754,429	770,483	1,524,912

Figure 43: Shotgun Adaptor Match. Full length reads that sequenced through to the 3'-adaptor, and can therefore be used to estimate library fragment length, are shown in blue.

Runs processed with amplicon pipelines will search for and match all five 3'-adaptor sequences associated with either Lib-A or Lib-L emPCR kits. On the other hand, runs processed with shotgun pipelines will only search for the 3'-adaptor sequences associated with the Lib-L emPCR kit, and only those associated with the specific library prep kit used (either Rapid or General).

Figure 44 shows the Adaptor Match untrimmed length histogram for amplicon data, showing the distribution of amplicon sizes with and without an adaptor match.



TCAG (Library)	% Total Reads								
	Region								
	1	2	3	4	5	6	7	8	Total
No Adaptor Match		0.18%	2.01%	0.18%	0.28%	69.34%	10.45%		18.29%
Lib-A, Forward		68.51%	49.19%	69.61%	0.62%	1.77%	35.86%		38.01%
Lib-A, Reverse		30.99%	48.49%	29.58%	98.12%	27.94%	51.13%		42.98%
Lib-L, General		0.27%	0.25%	0.44%	0.84%	0.36%	2.20%		0.47%
Lib-L, Rapid		0.04%	0.06%	0.18%	0.13%	0.42%	0.29%		0.18%
Lib-L, Rapid MID		0.02%	0.01%	0.01%	0.02%	0.16%	0.06%		0.05%
Total Reads	0	61,032	77,487	57,612	40,140	83,194	14,243	0	333,708

Figure 44: Amplicon Adaptor Match. Full length amplicons that sequenced through to the 3'-adaptor, and can therefore be used in a bidirectional analysis, are shown in blue.

The counts and types of adaptor sequence matches displayed here are similar to, but not identical to, the counts and types of adaptor sequences trimmed by the adaptor trim filter (Section 1.3.2.2). See Section 1.3.3.3 for an explanation of the differences in the way the adaptor match tool and the adaptor trim filter conduct their searches.

3.10.1 Adaptor Match Tab Features and Functionalities

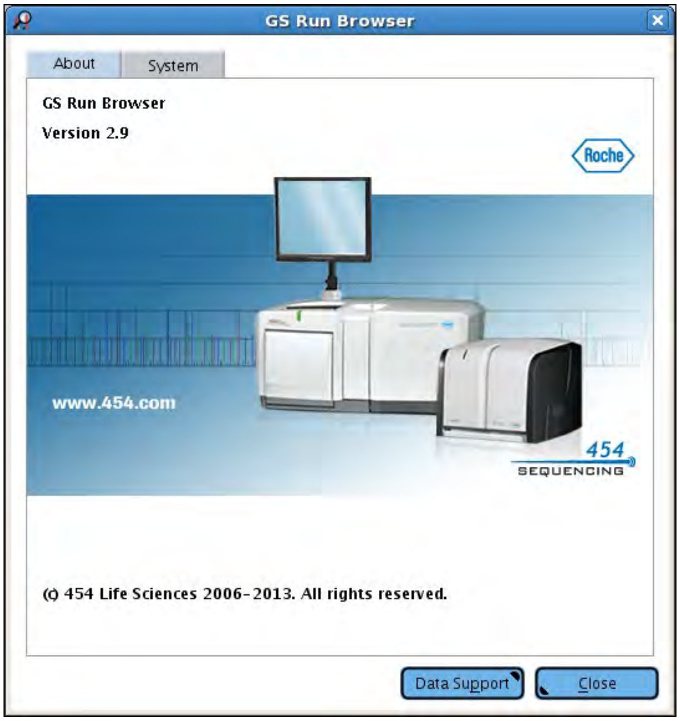
The Adaptor Match tab contains many features with the functionalities described below (see Figure 44):

- **Navigation and data capture buttons**: All the buttons on this tab have common functions: setting and adjusting the zoom level of the main image display or allowing a user to save the data as a text file or snapshot image. See Section 3.3.4 for a description of the navigation and data capture button functions.
- **Plot display**:
    - Graphic display of the data for a single region or for the whole PTP device, as selected on the Summary.
    - The 'count' (y-axis) corresponds to the number of passed filter reads at a given length, separated into those that have an adaptor match (blue) *vs*. those that have no adaptor match (green).
    - The plot shares the common scrolling and zooming functions of the other GS Run Browser plots. See Section 3.3.3 for a description of the plot functions.
    - When the mouse pointer is over the Adaptor Match plot, the mouse tracker shows the x-axis read length and the number of reads having that length, for the selected region or the entire PTP device.
- **Summary**:
    - Displays a text version of the data shown in the plot, for each region and for the entire PTP device. Adaptor match read counts are split into the counts for each of the 3' adaptor sequences found, and displayed as a percentage of the total passed filter reads.
    - A blue rectangle identifies the data (single region or 'Total' column) currently displayed on the plot. When the mouse is moved over the data in another column, the highlight moves to this column; left-clicking the mouse selects the highlighted column (blue rectangle) for display in the plot.

4 GS SUPPORT TOOL

The GS Support Tool is used to help in the troubleshooting of potential sequencing run issues or instrument problems for the GS Junior or GS FLX+ systems. The GS Support Tool will collect comprehensive information from the data set, which includes meta-data about the state of the GS Junior or GS FLX+ Instrument during the sequencing run, into a file that can be sent to Roche for a troubleshooting investigation.

The GS Support Tool can be accessed *via* the GS Run Browser GUI. Once a data set is loaded, clicking on the 'About' button in the global action area will bring up the following screen (Figure 45).

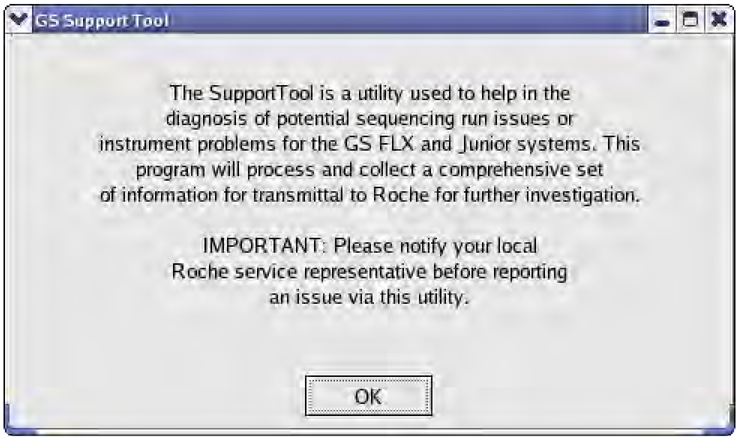


GS Run Browser
Version 2.9

(c) 454 Life Sciences 2006-2013. All rights reserved.

Figure 45: Access to the GS Support Tool *via* the 'About' button.

Clicking on the ‘Data Support’ button will bring up a confirmation dialog window (Figure 46) reminding the user to contact the Roche service representative to report the issue and open a support case.

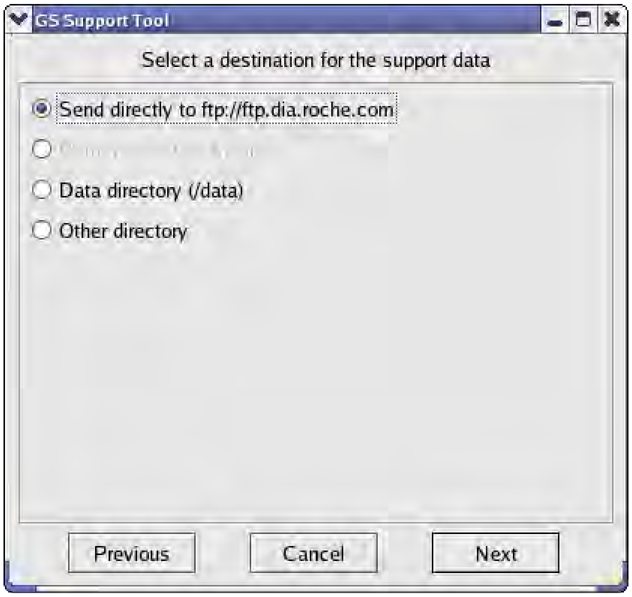


The SupportTool is a utility used to help in the diagnosis of potential sequencing run issues or instrument problems for the GS FLX and Junior systems. This program will process and collect a comprehensive set of information for transmittal to Roche for further investigation.

IMPORTANT: Please notify your local Roche service representative before reporting an issue via this utility.

Figure 46: GS Support Tool confirmation dialog.

The user is then asked to select a destination where the support data should be sent (Figure 47). The file can be sent directly to a Roche File Transfer Protocol (FTP) server (if the host machine has been configured for FTP transfers), it can be copied to a removable USB media device, copied to the /data directory or to another directory whose location would then need be specified.

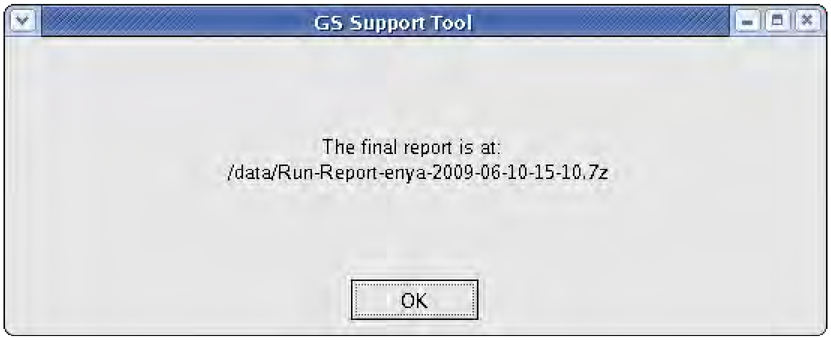


Select a destination for the support data

● Send directly to ftp://ftp.dia.roche.com
○ [illegible]
○ Data directory (/data)
○ Other directory

Figure 47: GS Support Tool report destination.

The name of the report generated start with 'Run-Report-' and includes a time stamp of the report generation, as shown on Figure 48.



The final report is at:
/data/Run-Report-enya-2009-06-10-15-10.7z

Figure 48: GS Support Tool report generation confirmation.

The GS Support Tool can also be invoked from a terminal window with the following command:

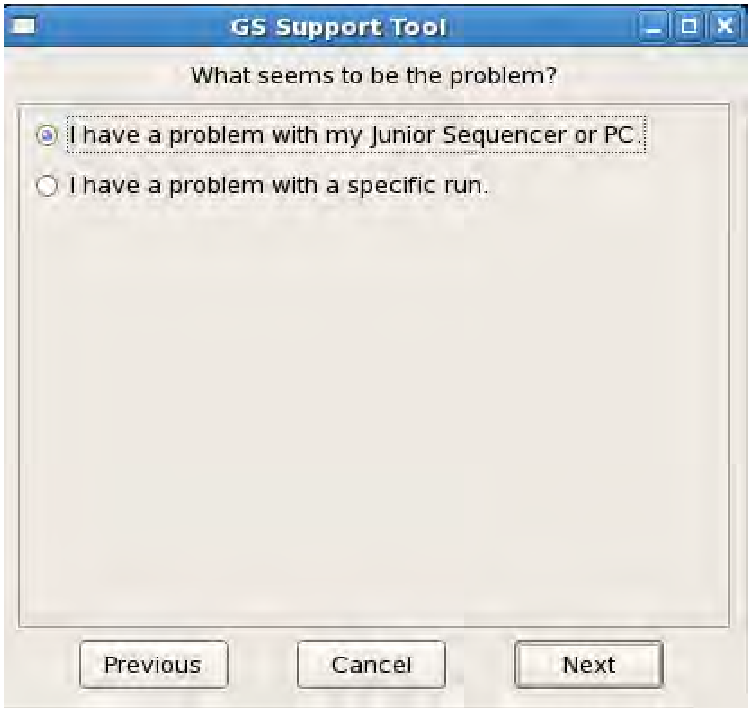
```
gsSupportTool [options] [SOURCE_DIRECTORY]
```

Argument	Description
SOURCE_DIRECTORY	Path to an 'R_' (full processing) or a 'D_' directory (signal or filter processing).

Option	Description
-h	Display the help message
-n	Do Not use the Network
-oDIR	Save output package to directory DIR
-p	Send file in plain text (*i.e.* never encrypt)
-f[SITE]	Sends file *via* ftp (default ftp.dia.roche.com).
-lUSER[:PASS]	Ftp login information (if needed)
-e[ADDRESS]	Emails the package (default service.sequencing@roche.com).
-mSERVER	Local e-mail server relay
-uUSER[:PASS]	Local e-mail server relay user login information (if needed)
-q	Quiet (no input or output).
-t	Forces text mode, even if X windows is available.
-V	Display gsSupportTool version

If no parameter is given to the FTP or e-mail destinations, defaults will be used.

When launched from the command line, an additional screen is initially displayed (Figure 49). This allows the user to specify if the problem is related to the GS Junior Instrument or attendant PC, or a specific run.


GS Support Tool

What seems to be the problem?

● I have a problem with my Junior Sequencer or PC.

○ I have a problem with a specific run.

Figure 49: Initial GS Support Tool screen when launched from the command line.

For this version of the gsSupportTool, the raw data is not required to be present in the same location as the processed directories. The cwf files in the processed folders contain compressed images, which are sufficient to create the support tool outputs.

5 Data Management Topics

Data management processes and scripts are used to archive sequencing run data (backupScript.sh) and to copy the sequencing run and image processed data to a datarig for signal processing and reporting (postAnalysisScript.sh).

5.1 backupScript.sh

The backupScript.sh is a user-modifiable script that is run automatically from the instrument and is responsible for copying the raw data, which cannot be recreated. For the GS FLX+ system, the script copies the data from the instrument to an archive system. For the GS Junior system, the script copies the data from the attendant PC to the archive system.

The following are the critical files in the R_ directory that are be backed up by default.

- dataRunParams.parse
- imageLog.parse
- runLog.parse
- aaLog.txt
- and all the PNG files under rawImages

The backup script is passed the full path of the R_ directory. It is run as the adminrig user and calls the ‘backupLog’ command to inform the instrument control of the status of the backup. It is important that any modifications or additions be made in the marked section. The lines containing ‘backupLog’ must be preserved. The RET_ERR variable passes the backup status to the GS Sequencer GUI Data tab and is set to 0 for success and 1 for error.

Below is an example backupScript.sh file content. The highlighted portion, between the sets of ‘========‘, is provided as an illustration only.

```
#!/bin/bash
###############################################################################
#
##
##    Filename:   backupScript.sh
##
##    Programmer: Bernard Puc
##
##    Description:
##
##          Runs at completion of the fluidics.
##
##    $Id: backupScript.sh,v 1.7 2005/01/20 15:59:23 bpuc Exp $
##
###############################################################################
#

if [ $# -ne 1 ]
then
      echo "Need to specify a directory on the command line."
      echo "Exiting."
      exit 1
fi

#
#     Initialize the backup status variable as failed
#
RET_ERR=1

#
#     Update the backup status file
#
backupLog $1 "running" "permanent"


# =======================================
# =======================================
#
#     Add custom backup code here...
#

# The first argument is the fully qualified pathname of the Run directory
RUN_SRC=$1

# The destination directory "/mnt/backupServer" must already be configured via
# nfs, samba, etc. Use on-demand mounting (autofs) rather than a persistent
# nfs mount.
DEST_SERVER=/mnt/backupServer
RUN_DEST=$DEST_SERVER/$RUN_SRC

# Copy all regular files (purposely allowing subdirectories to fail)
cp -p $RUN_SRC/* $RUN_DEST
if [ $? -ne 0 ]
then
      RET_ERR=1
else
      # Now explicitly copy subdirectory containing raw images
      cp -rp $RUN_SRC/rawImages $RUN_DEST
      if [ $? -ne 0 ]
      then
            RET_ERR=1
      fi
fi

# =======================================
# =======================================

#
#     Update the backup status file
#
if [ $RET_ERR -eq 0 ]
then
      backupLog $1 "complete" "permanent"
else
      backupLog $1 "failure" "permanent"
fi

#
#     End of script
#
exit 0
```

5.2 post-analysis script

Each processing pipeline chosen for a sequencing run on the GS Junior or the GS FLX+ Instrument can call a user-provided script, postAnalysisScript.sh, to copy the partially processed data (*e.g.* 'imageProcessingOnly') from the instrument to the datarig when the processing is complete. The script can optionally trigger an automated analysis or send notifications.

The following environmental variables are available to the post analysis script:

Environmental Variable	Description
$SOURCE_DIR	The directory from which the source data was read
$PIPELINE_NAME	The data analysis pipeline name
$SW_REVISION	The revision number of the data analysis pipeline
$GSREPORTER_BIN	The full path of the gsReporter executable

The full path of the directory containing the output files (the -o parameter to gsRunProcessor) is passed on the command line, then the name of each generated file relative to the output path is passed in the following order: CWF files, SFF files, then log/error files.

The order in which the postAnalysisScript.sh and the backupScript.sh are called is indeterminate. postAnalysisScript.sh is called when the analysis is complete, and backupScript.sh is called when the fluidics part of the run is complete. If both the partially processed data and the raw data are sent to the same destination, in creating directories, scripts should not assume one script was called before the other.

6 GS RUN PROCESSOR APPENDICES

The GS Run Processor application is identical for both the GS Junior and GS FLX+ Instruments. However, due to differences in instrument hardware, some references in this manual will be specific to either the GS Junior or the GS FLX+ Instruments. The main difference is the PTP device.

The PTP device on the GS FLX+ Instrument supports division into multiple (2, 4, 8 or 16) regions. The GS Junior Instrument PTP device only supports a single region. Therefore, any references to multiple regions are specific to the GS FLX+ Instrument.

6.1 gsRunProcessor executable

The gsRunProcessor command is not designed to be called directly by users but is primarily called from a launching script, by the gsRunProcessorManager or by a third-party job scheduling system such as SunGrid or PBS. The launching scripts handle all the job set-up, directory creation, and the launching of the gsRunProcessor executable itself. When the gsRunProcessor is called directly, its command line structure and arguments are as described below.

```
gsRunProcessor --prep [OPTIONS] sourceDirectory
gsRunProcessor --pipe=PIPELINE [OPTIONS] sourceDataFiles
```

Command	Description
gsRunProcessor	Runs image and/or signal processing algorithms, beginning with either the raw image files (*.png or legacy *.pif) or the output of the image processing step (the .cwf files) based on the processing launch script and pipeline processing script that invoked it.

Argument	Description
--prep	Either reads an existing dataRunParams.xml input XML file or creates one from the imageLog.parse and image files in the sourceDirectory and prepares the directory structure needed (D_ folder, regions and sff sub-folders) for the processing output.
sourceDirectory	Data directory containing files to be processed.
--pipe=XMLFILE	Specifies a pipeline processing script file. (Specifying the .xml extension is optional.)
sourceDataFiles	Data files to be processed.

Option	Description
--reg=REGION_NUM	Specifies which PTP Regions to analyze (*e.g.* '1-3,5')
-pXMLFILE, --pipe=XMLFILE	Specifies a pipeline processing script file (Specifying the .xml extension is optional.)
-rXMLFILE, --run=XMLFILE	Specifies a metadata file (*i.e.* dataRunParams.xml)
-iIMAGELOG, -imageLog=IMAGELOG	Specifies image information (*i.e.* imageLog.parse)
--analysisParms=FILE	Specifies the location of analysisParms.parse
-rDIR, --images=DIR	Specifies the raw image path
-lDEST, --log=DEST	Specifies the log file destination
-eDEST, --error=DEST	Specifies the error file destination
-oDIR, --out=DIR	Specifies the output directory
--intermediate[=type]	Outputs intermediate files(Type=wells or cwf)
-f, --force	Forces overwriting of output files
-u, --useDisk	Uses disk-backed storage to allow processing of larger data sets
-tTIME, --timeout=TIME	Specifies a timeout for waiting for new images
-nNAME, --name=NAME	Sets the processing job 'name'
-jJOBID, --job=JOBID	Overrides the automatic job ID creation
-T[XMLTEMPLATE], --template[=XMLTEMPLATE]	Creates a template XML script file of the type specified to standard output (*e.g.* '--template=filterOnly > filterTemplate.xml')
--remoteProgress [=HOST:PORT]	Sends progress information to an XML-RPC server like gsRunProcessorManager
--list[=ARG]	Lists choices for an option (*e.g.* '--list=template')
--mpi	Displays the MPI implementation (number of processors that will be used for the run)
-v, --verbose	Sends detailed messages to console & output files
-q, --quiet	Suppresses normal console messages
-P, --progress	Sends progress messages to the console
-V, --version	Displays the version of the software
-h, --help	Display available help information

There are a few of environmental variables that can be used to fine tune the gsRunProcessor's operation. Administrators and/or users do not normally need to set these variables after installation, however, when needed they can be used to locally define the environment variables for advanced users or special situations by inserting an environmental variable definition in a local or user's .bash_profile script. Site administrators may choose to adjust these values site-wide by editing /etc/profile or by adding a file in the /etc/profile.d directory.

Environmental Variable	Description
PIPELINE_PATH	Adds additional path for pipeline scripts. A colon (:) is used to separate paths when specifying more than one.
PHRED_TABLES_PATH	Overrides the default location of the .dat files used for PHRED based quality scoring.
LEGACY_DN_PATH	Adds additional path for legacy .dn files. A colon (:) is used to separate paths when specifying more than one.

6.2 startGsProcessor script

All the processing launch scripts with the 'run' prefix call the general purpose 'startGsProcessor' script which is responsible for calling the gsRunProcessor executable. Specifically, startGsProcessor is a wrapper script which first creates an output data directory with a regions subdirectory, and either reads an existing dataRunParams.xml input XML file or creates one from the imageLog.parse and image files, and finally calls the appropriate launching commands.

```
startGsProcessor --pipe=XMLFILE SOURCE_DIRECTORY
```

Command	Description
startGsProcessor	A wrapper script for the gsRunProcessor. It creates the output directories, prepares the run for processing and calls the appropriate launching commands.

Argument	Description
--pipe=XMLFILE	Specifies a pipeline processing script file. (Specifying the .xml extension is optional.)
SOURCE_DIRECTORY	Data directory containing files to be processed.

An important environmental variable that affects the operation of the `startGsProcessor` script is the '`GS_LAUNCH_MODE`' variable. This value can be set in each user's environment by adding an environmental variable definition line to the .bash_profile file in the user's home directory. Site administrators may have set these values site-wide in /etc/profile or by adding a file in /etc/profile.d.

```
GS_LAUNCH_MODE
```

Value	Description
SINGLE	Starts a single copy of the gsRunProcessor. (Default)
MULTI	Starts multiple copies of the gsRunProcessor, equal to the number of processors in the current workstation. (Non-cluster)
MPI	Uses 'mpiexec' for launching jobs on a compute cluster.
GSRPM	Starts the job using the gsRunProcessorManager. Will submit jobs to the same job queue as users who use gsRunBrowser to submit processing jobs. Recommended as a multi-user job queuing system.

6.3 GS Run Processor Manager

The gsRunProcessorManager is a server application that provides a simple queuing system for processing jobs that can launch, abort and monitor jobs, and report job progress. The gsRunProcessorManager is the mechanism used to launch processing jobs from the GS Run Browser application.

The command to start the gsRunProcessorManager daemon (run as root) is:

```
/etc/init.d/gsRunProcessorManager start
```

The gsRunProcessorManager command structure and options are:

```
gsRunProcessorManager OPTIONS
```

Command	Description
gsRunProcessorManager	Front end for management of a shared compute resource of gsRunProcessor

Option	Description
--daemon	Run application as a daemon
-CFILE, --conf=FILE	Loads an alternate config file
-pPORT, --port=PORT	Attaches to port number (default=4540)
--pid=FILE	Writes process ID (pid) to a file
--portFile=FILE	Writes port number to a file
-v, --verbose	Increases verbosity
-V, --version	Displays the version of the software
-h, --help	Display argument help information

The `gsRunProcessorManager` is controlled by a single XML configuration file, gsRunProcessorManager_conf.xml, found in /etc/gsRunProcessorManager/. Alternate configuration files may be specified on the command line by specifying a parameter to the -CFILE or --conf=FILE option. The parameters in the configuration file are described below.

Parameter	Values	Description
enableQueue	True, False	Enables or disables job queuing mechanism
log → file	gsRunProcessorManager.log path (/var/log)	Specifies the path to output the gsRunProcessorManager log file
log → autorotate	True, False	Enables or disables log autorotate such that the log file is not overwritten for each processing job.
outputDirectory	Path to sequencing run output data directory (/data)	Used to estimate if there is enough disk space available to start a new sequencing run
xmlrpc → port	Valid port number (4050)	Controls the port on which gsRunProcessor Manager listens for XML-RPC and HTTP requests
files → pid	gsRunProcessorManager.pid path (/var/run)	Specifies the path to output the gsRunProcessorManager pid file
files → port	gsRunProcessorManager.port path (/var/run)	Specifies the path to output the gsRunProcessorManager port file
mpi → enable	True, False	Enables or disables use of mpiexec for launching jobs. Equivalent to GS_LAUNCH_MODE=MPI.

The gsRunProcessorManager package also contains a small tool called gsRunProcessorManagerCtrl. It can be used to interact with the gsRunProcessorManager from command line scripts or other utilities.

```
gsRunProcessorManagerCtrl [OPTIONS] COMMAND [PARAMETERS]
```

Command	Description
gsRunProcessorManagerCtrl	Submits jobs to a gsRunProcessorManager. Not to be used directly, but to be called by shell scripts.

Commands	Description
Launch	Launches a processing job
Status	Gets the status of a processing job
Queue	Shows the status of the job queue
Abort	Interrupts processing of a job
queryVersion	Gets the version of the remote software

Option	Description
-HHOSTNAME, --host=HOSTNAME	Connects to host (default=localhost)
-m, --machine	Format output data suitable for scripting
-pPORT, --port=PORT	Uses port number (default=4540)
-v, --verbose	Increases verbosity
-V, --version	Displays the version of the software
-h, --help	Display argument help information

6.4 GS Run Processor Log Files – gsRunProcessor.log and gsRunProcessor_err.log

These files provide log information from the GS Run Processor job. Each line has a timestamp, and message. In the gsRunProcessor_err.log a timestamp, numeric error id, error severity, error message and source of the message is conveyed.

- 'Notice' or 'Information' messages are generated as the data processing progresses.
- 'Warning' messages are generated when an unexpected condition is encountered and should be investigated. Conditions that may result in warning messages but can be safely ignored include:
  - An unloaded / empty PTP Region (GS FLX+ Instrument only)
  - A region without library reads (*e.g.* a 'Control DNA Only' run) (GS FLX+ Instrument only)
  - A missing postAnalysisScript.sh script
- 'Error' messages are generated when the GS Run Processor encounters conditions from which it cannot automatically recover and will produce sub-standard or unusable results.
- 'Fatal' messages are generated when the GS Run Processor encounters errors that prevent the GS Run Processor from proceeding, such as missing input files or improper configurations. The GS Run Processor will exit after generating 'Fatal' messages and delete any partially generated output files.

All messages marked as 'Warning', 'Error' or 'Fatal' are placed in a 'gsRunProcessor_err.log' file in addition to being placed in the gsRunProcessor.log file. If no messages of this type are generated, the gsRunProcessor_err.log file will not be created.

6.5 GS Reporter Metric File Contents Descriptions

gsReporter Output Metrics File Contents:

454RuntimeMetricsAll / 454AllControlMetrics	454QualityFilterMetrics	454BaseCallerMetrics
Comment Header software version Run Name Analysis Name Region Name Key sequences file creation timestamp	Comment Header software version Run Name Analysis Name file creation timestamp	Comment Header software version file creation timestamp
Run Conditions Group Run Name Analysis Name PTP barcode Number of regions Number of cycles	Run Conditions Group Run Name Analysis Name PTP barcode Number of regions Number of cycles	Run Parameters Group Run Name Analysis Name PTP barcode Number of regions Number of cycles
Region Group Region Name Number of raw wells Number of Key Pass wells Key Group	Region Group Region Name Number of raw wells Number of Key Pass wells Key Group	BaseCaller Results Number of reads Average read length Standard deviation of read length
Key Group* keySequence keyPassWells keySignalPerBase ppi1, ppi2, ppi3 (avg signal, std dev.) *For the 454AllControlMetrics file, this group contains the combined metrics from all control reads.	Key Group keySequence numKeyPass numDotFailed numMixedFailed numTrimmedTooShortQuality numTrimmedTooShortPrimer totalPassFiltering	Region Key Group Region Name Key sequence Number of reads total number of Bases Average read length Average quality score Length histogram group (length bin counts) Quality Histogram Group (quality bin count)

Timestamps:

Convention	Description
YYYY/MM/DD HH:MM:SS or YYYY_MM_DD_hh_mm	YYYY - four digit year MM - two digit month (01 to 12) DD - two digit day (01 to 31 depending on month) hh - two digit hour (00 to 23) mm - two digit minutes (00 to 59)

Names and Strings:

Convention	Description
Rig Name	Network host name of the computer in the GS FLX+ Instrument or the GS Junior attendant PC.
User Name	Name of the user who executed the run (from the GS Sequencer sign in)
freeForm	Any alpha numeric string ('Unique run name') that was entered by the user for the sequencing run
AnalType	String identifying the data processing type that was run. 'imageProcessingOnly', for example, contains the image processed results.

Directory Names:

Type	Convention	Description
Run Name	R_YYYY_MM_DD_hh_mm_rigName_userName_freeForm	R_ stands for 'Run', the rest defined above
Analysis Name	D_YYYY_MM_DD_hh_mm_rigName_userName_AnalType or D_YYYY_MM_DD_hh_mm_rigName_userName_freeForm	D_ stands for 'Data', a the rest defined above

Comment Header Definitions:

Information	Description
Software Version	Version of the Data Analysis Software invoked
Run Name	R_YYYY_MM_DD_hh_mm_rigName_userName_freeForm
Analysis Name	D_YYYY_MM_DD_hh_mm_rigName_userName_AnalType or D_YYYY_MM_DD_hh_mm_rigName_userName_freeForm
Region Name	A number valid for the number of loading regions on the PicoTiterPlate device, *e.g.* for a two region device the region name can be '1' or '2'.
Key Sequences	Four base adaptor sequences for Library and Control beads
Timestamp	YYYY/MM/DD HH:MM:SS

Run Group Definitions:

Information	Description
Run Name	R_YYYY_MM_DD_hh_mm_rigName_userName_freeForm
Analysis Name	D_YYYY_MM_DD_hh_mm_rigName_userName_AnalType or D_YYYY_MM_DD_hh_mm_rigName_userName_freeForm
PTP barcode	The six digit PTP ID number.
Number of regions	The number of physically separate regions present on the PicoTiterPlate device (as defined by the Bead Deposition Gasket used during PTP device preparation). This value was set by the user during the sequencing run setup. Data from each region is processed separately, and each region has a complete set of metrics.
Number of cycles	The number of nucleotide flows divided by four.

Region Group Definitions:

Information	Description
Region name	A number valid for the number of loading regions on the PicoTiterPlate device, *e.g.* for a two region device the region name can be ‘1’ or ‘2’.
Number of Raw wells	The number of raw wells detected in this region of the PicoTiterPlate device. These are wells that are generating enough signal to be considered for further processing.
Number of Key Pass Wells	The total number of wells where the first four bases called match any defined key, detected in this region of the PicoTiterPlate device. A key pass well is assumed to contain a legitimate DNA read. The key pass wells are the wells that are further processed in the downstream data analysis pipeline.
Key Group	A sub group containing key information, described below.

The key subgroup is contained in the region group. The key sequence for sample library reads is GACT, **TCAG** and for Control DNA reads is **CATG** or **ATGC** (see Section 7.1 for details on Control DNA keys). This corresponds to the first four nucleotides in a read. This is dictated by the keys present (1) on Adaptors used to generate the DNA library and (2) on the Control DNA Beads provided in the sequencing kits. The keywords contained in the key group are:

Key Group Definitions:

Information	**Description**
Key sequence	First three bases of the actual key sequence; 'GAC' or 'TCA' for library reads and for Control DNA reads 'CAT' or 'ATG' (see Section 7.1 for details)
numKeyPass	The number of wells identified on the PicoTiterPlate device with a valid key sequence. A Key Pass well is assumed to contain a legitimate DNA read and is further processed in the downstream data analysis pipeline.
Key signal per base	A comma-separated list providing the average signal per base for the key signals and the corresponding standard deviation.
ppi1, ppi2, ppi3…	A set of comma-separated lists providing the average signal and the corresponding standard deviation, for each normalization flow during the sequencing run.
numDotFailed	The number of reads that failed the 'dot' filter; reads with too many negative flows. (Section 1.3.2)
numMixedFailed	The number of reads that failed the 'mixed' filter; reads with too many positive flows. (Section 1.3.2)
numTrimmedTooShortQuality	The number of reads that failed the length test because of quality trimming. (Section1.3.2.2)
numTrimmedTooShortPrimer	The number of reads that failed the length test because of Adaptor Trim Filter sequence trimming. (Section1.3.2.2)
totalPassFiltering	The number of wells that passed all five filters in the Signal Processing application. These are the wells that contain high quality DNA reads that are used (as trimmed) as input for data analysis post-run processing, or for basecalling of the reads.

BaseCaller Results Group Definitions:

Information	**Description**
Number of Reads	Number of reads from the wells that passed all five filters in the Signal Processing application.
Average Read Length	Average length across all reads that passed all five filters in the Signal Processing application.
Std Dev of Avg. Read Length	Standard deviation of the average length across all reads that passed all five filters in the Signal Processing application.

Region Key Group Definitions:

Information	**Description**
Region name	A number corresponding a loading region on the PicoTiterPlate device, *e.g.* for a two region PTP device the region name can be '1' or '2'.
Key sequence	First three bases of the actual key sequence; 'TCA' for library reads and for Control DNA reads 'CAT' or 'ATG' (see Section 7.1 for details)
Number of Reads	The total number of DNA reads that were retained (passed) after all quality filters were applied in this region with the specified key.
totalBases	Total number of DNA bases in the filtered reads (after trimming).
Average Read Length	A comma-separated list providing the average length of the DNA reads after filtering and trimming, and the corresponding standard deviation.
Average Quality Score	A comma-separated list providing the average quality score of all the DNA bases in these filtered reads (after trimming), and the corresponding standard deviation.
Length Histogram Group	A subgroup located within the regionKey group which provides a comma-separated list of the number of filtered reads at each read length observed in this region and with this sequencing key. Thus, on each line; a keyword, the read length (for the abscissa of the histogram), and the number of filtered reads with this trimmed read length.
Quality Histogram Group	A subgroup located within the regionKey group which provides a comma-separated list of the number of bases called at each quality score observed in this region and with this sequencing key. Thus, on each line; a keyword, the quality score (for the abscissa of the histogram), followed by the number of bases with this quality score.

6.6 Phred-equivalent Base Quality Scores

Quality scores for individual called bases are determined by a method developed in collaboration with the Broad Institute(*Genome Research*,18(5): 763-70, 2008), whereby the methodology described by Ewing and Green (*Genome Research*, 8: 186-194, 1998) for the creation of quality scores as part of the Phred basecalling algorithm is applied to generating quality scores for 454 Sequencing reads. The quality scores computed for each called base are written to the CWF and SFF files (and optionally to a file paralleling the basecall FASTA file). Briefly, the method compares the properties of each base's flowgram signals against properties that have been found to correlate with accurate and/or error-prone signal information, using training sets of read data. A multivariate analysis of those properties determines the sets of property values that best describe 'bins' of basecalls, then assigns the training set accuracy rates of the basecalls in each bin as a quality score using the following scale:

$Q = -10 \log_{10} (\text{error rate})$

Then, when the basecalling of a new read is performed, the trace properties for each read are used to determine the bin in which each base falls, and the quality score associated with that bin is assigned to the basecall. For 454 Sequencing reads, the property metrics based on the following characteristics are used for the training and the quality score generation:

- Base position of the base in the read sequence.
- Flow signal intensity for the flow where the base is called.
- Flow signal intensity in the flows before and after the flow where the base is called, specifically the before and after flows where uncorrected CAFIE signal could remain in the flow of the basecall.
- Local variation of the flowgram signals in a window surrounding the flow where the base is called (*i.e.*, how far are those signals from the ideal 0.0, 1.0, 2.0, 3.0, … signals). This provides an estimate of the homopolymer accuracy.
- Overall ‘separation’ of the flowgram signals (*i.e.*, how greatly separated are the 0-mer signals from 1-mer signals).

To account for slightly different error tendencies, separate calibrations were performed using shotgun data obtained from runs using several subsets of GS FLX Titanium chemistry, including GS Junior runs, GS FLX runs, GS FLX+ runs using the cyclic flow pattern, and GS FLX+ runs using the acyclic flow pattern.

7 GS Run Browser Appendices

7.1 Control DNA (Test Fragment) Sequences

Control beads are sequencing beads with a set of defined control DNA (or test fragment) sequences attached. When included as part of a sequencing run, they can be used to help distinguish between problems with the sequencing run itself and issues with upstream portions of the protocol (library preparation and emPCR). Table 20 outlines the differences between two types of control DNA test fragments, including the sequencing key, names of the individual test fragments, and length. The 454 Sequencing system software automatically recognizes control reads and reports sequencing accuracy at various lengths on the Control DNA Tab in GS Run Browser (Section 3.8).

Type	Key	Name	Length
Type I	CATG	AVTF2 AVTF7 AVTF90 AVTF100 AVTF120 AVTF150	508 bp 524 bp 540 bp 560 bp 580 bp 600 bp
Type II	ATGC	ECTF301 ECTF302 ECTF303 ECTF304	952 bp 957 bp 949 bp 942 bp

Table 20: Control DNA (Test Fragment) Types. Type I and Type II controls differ primarily in sequencing difficulty (Type I controls are more difficult to sequence) and length (Type II controls are longer). Taken together as a set, the Type II controls were designed with the property of exhibiting a positive signal at every flow over a 200 cycle run (starting after the key). This can be useful for troubleshooting valve events, particularly with low-diversity amplicon sequencing runs. This special characteristic does not hold true for the later flows in a 400 cycle run, or for acyclic flow pattern runs.

The actual sequences for the Type I AVTF controls are displayed in Table 21, and those for the Type II ECTF controls in Table 22. These sequences are found in the sequences.xml stream of the .cwf file (see the Overview of this manual for additional details on the contents of the .cwf file). They are also located in the legacy analysisParms.parse file, which can be generated using the --analysisParms.parse option for gsReporter (Section 2.1).

Control	Sequence
AVTF2	5′-CATGCAAGCGTACGCACGTGGTTGTTAAAGCTTTTTTGAAAGTTAATCTCCTGGTTCACCGTCTGCTCGTATGCGGTTACCAGGTCGGCGGCCGCCACGTGTGCGCGCGCGGGACTAATCCCGGTTCGCGCGTCGGGCTCAAAGTCCTCCTCGCGCAGCAACCGCTCGCGATTCAGGCCATGCCGCAGCTCGCGCCCTGCGTGGAACTTTCGATCCCGCATCTCCTCGGGCTCCTCTCCCTCGCGGTCGCGAAACAGGTTCTGCCGCGGCACGTACGCCTCACGCGTATCACGCTTCAGCTGCACCCTTGGGTGCCGCTCAGGAGAGGGCGCTCCTAGCCGCGCCAGGCCCTCGCCCTCCTCCAAGTCCAGGTAGTGCCGGGCCCGGCGCCGCGGGGGTTCGTAATCACCATCTGCTGCCGCGTCAACCGCGGATGTCGCCCCTCCTGACGCGGTAGGAGGAGGGGAGGGTGCCCTGCATGTCTGCCGCTGCTCTTGCTCTTGCCATG-3′
AVTF7	5′-CATGTACCTCTCCGCGTAGGCGCTCGTTGGTCCAGCAGAGGCGGCCGCCCTTGCGCGAGCAGAATGGCGGTAGGGGGTCTAGCTGCGTCTCGTCCGGGGGGTCTGCGTCCACGGTAAAGACCCCGGGCAGCAGGCGCGCGTCGAAGTAGTCTATCTTGCATCCTTGCAAGTCTAGCGCCTGCTGCCATGCGCGGGCGGCAAGCGCGCGCTCGTATGGGTTGAGTGGGGGACCCCATGGCATGGGGTGGGTGAGCGCGGAGGCGTACATGCCGCAAATGTCGTAAACGTAGAGGGGCTCTCTGAGTATTCCAAGATATGTAGGGTAGCATCTTCCACCGCGGATGCTGGCGCGCACGTAATCGTATAGTTCGTGCGAGGGAGCGAGGAGGTCGGGACCGAGGTTGCTACGGGCGGGCTGCTCTGCTCGGAAGACTATCTGCCTGAAGATGGCATGTGAGTTGGATGATATGGTTGGACGCTGGAAGACGTTGAAGCTGGCGTCTGTGAGACCTACCGCGTCCATG-3′
AVTF90	5′-CATGCGCAAAAACGCAAAACGCAAACGCAACGCACCAGCCTATGCGCCTGGTCTGTACACCGTTCATCTGTCCTCTTTCAAAGTTGGTCAGTTCGGTTCCCTTATGATTGACCGTCGGACGAGCCCCTTTTTTGCTTTTCCCAGATGCATCCGGTGCTGCGGCAGATGCGCCCCCCTCCTCAGCAGCGGCAAGAGCAAGAGCAGCGGCAGACATGCAGGGCACCCTCCCCTCCTCCTACCGCGTCAGGAGGGGCGACATCCGCGGTTGACGCGGCAGCAGATGGTGATTACGAACCCCCGCGGCGCCGGGCCCGGCACTACCTGGACTTGGAGGAGGGCGAGGGCCTGGCGCGGCTAGGAGCGCCCTCTCCTGAGCGGCACCCAAGGGTGCAGCTGAAGCGTGATACGCGTGAGGCGTACGTGCCGCGGCAGAACCTGTTTCGCGACCGCGAGGGAGAGGAGCCCGAGGAGATGCGGGATCGAAAGTTCCACGCAGGGCGCGAGCTGCGGCATGGCCTGAATCGCGAGCGGTTGCTCATG-3′
AVTF100	5′-CATGATCCTATCCCTATCCCCTATCCCCCTATTCTCTTGAGGAGTCTCAGCCTCTTAATACTTTCATGTTTCAGAATAATAGGTTCCGAAATAGGCAGGGGGCATTAACTGTTTATACGGGCACTGTTACTCAAGGCACTGACCCCGTTAAAACTTTCATCAGTAACCCGTATCGTGAGCATCCTCTCTCGTTTCATCGGTATCATTACCCCCATGAACAGAAATCCCCCTTACACGGAGGCATCAGTGACCAAACAGGAAAAAACCGCCCTTAACATGGCCCGCTTTATCAGAAGCCAGACATTAACGCTTCTGGAGAAACTCAACGAGCTGGACGCGGATGAACAGGCAGACATCTGTGAATCGCTTCACGACCACGCTGATGAGCTTTACCGCAGCTGCCTCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGACGGTCACAGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCGCAGCCATGACCCAGTCATG-3′
AVTF120	5′-CATGAGGATAGGGATAGGGGATAGGGGGATAGGCTTAACTCAATTCTTGTGGGTTATCTCTCTGATATTAGCGCTCAATTACCCTCTGACTTTGTTCAGGGTGTTCAGTTAATTCTCCCGTCTAATGCGCTTCCCTGTTTTTATGTTATTCTCTCTGTAAAGGCTGCTATTTTCATCTTCTGCGCTAAGATTGTCAGTTTCCAAAAACGAGGAGGATTTGATATTCACCTGGCCCGCGGTGATGCCTTTGAGGGTGGCCGCATCCATCTGGTCAGAAAAGACAATCTTTTTGTTGTCAAGCTTGGTGGCAAACGACCCGTAGAGGGCGTTGGACAGCAACTTGGCGATGGAGCGCAGGGTTTGGTTTTTGTCGCGATCGGCGCGCTCCTTGGCCGCGATGTTTAGCTGCACGTATTCGCGCGCAACGCACCGCCATTCGGGAAAGACGGTGGTGCGCTCGTCGGGCACCAGGTGCACGCGCCAACCGCGGTTGTGCAGGGTGACAAGGTCAACGCTGGTGGCTACCTCTCCGCGTAGGCGCTCGTTGGTCCAGCAGAGGCGGCCGCCCTTGCGCGACATG-3′
AVTF150	5′-CATGGCGTTTTTGCGTTTTGCGTTTGCGTTGCGTTATAACCCAACCTAAGCCGGAGGTTAAAAAGGTAGTCTCTCAGACCTATGATTTTGATAAATTCACTATTGACTCTTCTCAGCGTCTTAATCTAAGCTATCGCTATGTTTTCAAGGATTCTAAGGGAAAATTAATTAATAGCGACGATTTACAGAAGCAAGGTCGCGGCGACACCTTTGCCACACGGGCTGAGGAGAAGCGCGCTGAGGCCGAAGCAGCGGCCGAAGCTGCCGCCCCCGCTGCGCAACCCGAGGTCGAGAAGCCTCAGAAGAAACCGGTGATCAAACCCCTGACAGAGGACAGCAAGAAACGCAGTTACAACCTAATAAGCAATGACAGCACCTTCACCCAGTACCGCAGCTGGTACCTTGCATACAACTACGGCGACCCTCAGACCGGAATCCGCTCATGGACCCTGCTTTGCACTCCTGACGTAACCTGCGGCTCGGAGCAGGTCTACTGGTCGTTGCCAGACATGATGCAAGACCCCGTGACCTTCCGCTCCACGCGCCAGATCAGCAACTTTCCGGTGGTGGGCGCCGAGCTGTTGCCCGTGCACTCCCATG-3′

Table 21: Type I (AVTF) Control DNA / Test Fragment Sequences.

Control	Sequence
ECTF301	5′-ATGCCTAGTAAACAATGTTCGATCCGGCGAAGTCTGCAAGAATCCAGCGCTGCCGGTTCGTCGGCGCTGTGCCGTGGAGCTGACCTGATCGACGACTCGTGCGATCGAGTATGGCTCGCCACCTAGAATCGCGACCTATCATCAACGAAGGCGACATCAATACGTCGCCGCGTTAAATCGTCACTTTCTGCGCACGTCAGCATTACCGCACAGTGCTGGAGTCACTTCCGTGCCGGTCTCTGCTGGTGACTTCGCAGACCGTCCTGACGTGCTAGGTGGCCGCTTCGCCTGGCAGCCGATGAACGGCGTAGATCGGTGCCCTCGCTGCTAGACTGGCGGTTACGAAATGGACGCGCGCATCTACTGTGCGAACGTGCTCGCTACCGGACCTGCCGGTATATGGTGAACACCAACACCTCGGCAGACCTCTCTGAGCCTAGCAGAGCTTCAACCTGGAAGTTCCGGTGACGATCACGAACGTATCGACGAAAGTTCAGGAATACGTTGCTACTACATCTAACGCTGACTGGATCGAATACTCTGACTGCACTTCTGTAGCGCAGCCGTCGTCTAGTCTCCGCCTGTCGTTCCGTTATCAGCTGACTGAACTTGACGCGCAGCGGGCACGTATCGTACTGTCCGGAAGGTGACGAACGCGTACCGTTAAAGCAGCCGTCTATCTGTGCTGAACGTAGGTATCGCAACTTGACGTACTGCTGGGTAATCCGGCAGAGATCACCGTGTTGCAGCGTCTCAGGGTGTAGAACTGGGTGCAGGGATTGAAATCGTTGATCCAGAAGTGGTTCGCGAAAGCTATGTTGGTCGTCTGGTCGAACTGCGTAAGAACAAAGGCATGACCGAAACCGTTGCCCGCGAACAGCTGGAAGACAACGTGGTGCTCGGTACGCGCATGACACGCAACAGGGGATAGGGACACGCACGCAACAGATGG-3′
ECTF302	5′-ATGCTCGACCGACGGTGTTGTAACGTGACGATTACACTACCTGTCTGCTATCGAACGAAGGCACTACGTTATCGCCCAGGCGAACTCCAACTTCGGATGAAGAAGGTCCACTTCGTAGAAGACCTGGTAACTTCGCCGTAGCAAAGGCGTAATCCTAGCTTGTTCAGCCGCGACCAGGTTGTACTACATGGACGTACTCCACAGACAGGTGGTATCCGTCGGTGACGTCTGATCCCGTCCTGGAACACGTATGACGACCAACCGTGCATTGATAGGGTGCGAACATGACAACGTCAGGCCGTTCCGACTCTGACGCGCTGATAACGCCGCTGGTTGGTACTGTATGGAACGTGCTGTTGACCGTTGACTCCGGTGTAACTAGCGGTAGCTAAACGTGGTGGTGTCGTTCGAGTACGTGGATGCTTCCCGTATCGTTATCGAGTTAACGAAGACGAGATGTATCCGGGTGAAGTCAGGTATCGACATCTACAACTGACCAAATACACGTTCTAACCAGAACACTGTATCAACCAGATGCCGTGTAGTGTCTCTGGGTGAACCGGTTGTAACGTGGCGACGTGCGTGGCAGACGGTCCGTACCACCGACCTCGGTGAACTAGGCGCTGGTCAGAACATGCGCGTACGCGTTCATGCCGTGAATGGTTACAACTTCGAAGACTCCATACCTCGTATCCGAGCGTGTTGTTCAGGTAAGACCGTTTCACCACCATCACATTCAGGAACTAGGCGTGTGTGTCGTGACACCAGCTGGGTCCGGAAGAGATCACCGCGTGACATCGAACGTGTGAAGCTGCGCTCTCCAAACTGGATGAATCCGGTATCGTTTACATTGGTGCGGAAGTGACCGGTGGCGACATTCTGGTTGGTAAGGTAACGCCGAAAGGTGAAACTCGCATGACACGCAACAGGGGATAGGGACACGCACGCAACAGATGG-3′
ECTF303	5′-ATGCGGTGTTATCCGACATTCACGGTCCTGGCCGATTGTACTGCAGGGTGATGACTCTCCCTGCCGTAACCGTTACGTCTATCGCACTGAACTCGAGCGCGACTACTGTAGGTGCGGTTGTTATAGGGTCCGTACGCTGACCGTTGCCGTAAGGCATGAGTTAAGTGTACTGGCCGTATCTGGAAGTTCCGGTTAGGCCGTGGCCGTGCTGGGCCGTGTGGTTAACACTACTGTGACACCAATCGACGTAGGTCCGCGTGGATCACGACGGCTTCTACTGCTGTAGAAGTCAATCGCTCCGCGTTATCGAACGTCTAGTCCGTAGATACAGCCGGTACAGACCGGTTATAAAGTCCGTTGACTCCATGACTCCCAATCGGTCGTGGTCAGACGTGAATTGATCATCGGTGACGTCAGACAGGTACCGCACTGGCTATCGATCGCCATCATCAACCAGCGCGATTCCGTATCAAATGTATCTATGTCGTCTATCGGCCAGAGCGTCCGACCATTTCTAACGTGGTACGTAAACTGAGAGCACGGCGCACTAGGCTAACACCATACGTTGTGGTAGCAACCGCGTCGTGAATCCGCGTGCACTGTCAATACCTGGCACCGTATAGCCGGTTGCGCAATGGGACGAATACTTCCGTGACCGCGGTGAAGATGACGCTGATCATTTACGACTGACCTGTCTACAGGCGTGTTGCTTACCGTCGAGATCTCCCTGCTGCTCCGTACGTCCGCCAGGACGTGAAGCATTACGCGACGTCTACCTCCACTCTCGTCTGCTGGAGCGTGCTGCACGTGTTAACGCCGAATACGTTGAAGCCTTCACCAAAGGTGAAGTGAAAGGGACCGGTTCTCTGACCGCACTGCCGATTATCGAAACTCAGGCGGGTGACGCATGACACGCAACAGGGGATAGGGACACGCACGCAACAGATGG-3′
ECTF304	5′-ATGCCGAAGCAATCCTAGCACGCGACGCGGCGCAGGTACCGGTAGGTGGTTGCAGTGACAAGATCGATACCAGAAGCTAGATCCGGATCGCGTAAGAACGAACTCTCCCACGTACGGCATCCTGCCGGAAGTAGTGCGGTAGAAAGCCGAGTTCGTACACGTATCTGCGAAAGCGTACCGTATCGATGTAACTGCTGGACGCTATACCTGCTGCACGGCGGAAGTTCGTGGAGCTCGAGCGGTACGTACGGTATGGCGAGCGGTGACGGTTATCGAATCCTTCCTCGATAGGTCGTGGTCCGTTGCTACCGTTCTAGGTACGTGAAGGTACTACTGCACAAGGGCGATATCGTTACTGTGTGGCGTTCGAATACGGTCGTGTTCGTGCGATAGCGTAACGAACTGTCAGGAAGTGCTGGAAGACGTCCGTCCATTCCGGTGGAAATCCTCGGCCTGTCCGCGTACCGGCTGCGGGTGATCGAAGTTACCGTTGTACGTGACGAGAAGAGCGCGTAGAAGTTGCACTCTATCGTACAGTAAATTCCGCGAAGTTACTAGGCGCGTCAGCGAGATCTACTCGTAGAACATGTTCGCCAACATAGACCGAAGGACGAAGTTCACGAAGTAGAATATCGTCCTGAAGGACAGACGTACAGGGTTCTGTCGAAGACGATCTCCGACTCCTTGCTGTACTGTCTACTGACGAAGTTAAACGTGAAGATCATCGGTTCTAGGCGTAGGTGGTATCACCGACGACGCCACTGGCTGACGGCGTCCAACGCCATCCTGGTTGGTCTAACGTACGTGCTGATGCCTCTGCACGTAAAGTGATTGAAGCGGAAAGCCTGGATCTGCGTTACTACTCCGTCATCTATAACCTGATTGACGAAGTGAAAGCGCATGACACGCAACAGGGGATAGGGACACGCACGCAACAGATGG-3′

Table 22: Type II (ECTF) Control DNA / Test Fragment Sequences.

GLOSSARY

A

Adaptor – a short strand of nucleic acid (an oligonucleotide) that serves to adapt a target DNA sequence to the 454 Sequencing system. The 5'-adaptor contains embedded amplification and sequencing primer sequences designed to function with the Lib-L or Lib-A emPCR kits, as well as a four base sequencing key that is used to identify key pass wells after a sequencing run. The 3' adaptor contains a sequence complementary to the capture oligo, which is covalently attached to the capture beads and used during emPCR to generate sequencing beads with millions of copies of a target DNA. The 5' sequencing key and 3'-adaptor are trimmed from a read during the signal processing phase of the processing pipeline, leaving the target sequence for subsequent analysis.

Accession number – a unique identifier given to a DNA or protein sequence record utilizing a universal accession naming convention created by the UniProt (SwissProt) Knowledgebase.

ATP (PPI) flow – in the pyrosequencing reaction, one molecule of Adenosine-5'-triphosphate (ATP) is synthesized for each unit nucleotide incorporation causing a flash of light (signal) whose intensity is proportional to the number of nucleotides incorporated. The initial ATP flow (GS Junior titanium and GS FLX Titanium chemistry) causes a simultaneous flash of light from all enzyme active wells and is used to define the PTP device loading regions and the shape of the background across the plate.

B

Basecalling – use of the relative signal intensity generated during an individual nucleotide 'flow' (incorporation) step to generate a quantitative representation of nucleotide incorporation (singlet or homopolymer stretch).

Background – the non-specific or non-resolvable signal intensity generated during the sequencing run which is corrected or subtracted to reduce the noise (signal to background ratio) and improve the information content of the signal flowgrams.

Base contribution profile – a plot displaying the number of bases contributed by reads at each observed value of read length.

Base contribution percentile – the percentage of bases contributed by reads of the reported read length or longer.

C

CAFIE (CArry Forward & Incomplete Extension) – corrects out-of-phase errors using a probability model in order to improve signal and decrease noise.

- **Carry Forward** occurs when a trace amount of nucleotide remains in a well after the apyrase wash, perpetuating premature nucleotide incorporations for specific sequence combinations during the following flows. While this generally affects only a small percentage of DNA strands per bead, it causes those strands to continue to incorporate nucleotides out-of-phase with respect to the rest of the strands.
- **Incomplete Extension** occurs when some DNA strands on a bead fail to incorporate during a nucleotide flow. This is more likely to occur with higher order homopolymers, and can be due to localized reagent concentration differences within the PTP device. Strands that fail to incorporate the appropriate nucleotide must wait for the next flow of that nucleotide to continue extending. If the following nucleotide in the DNA sequence flows before this happens (guaranteed with the cyclic TACG flow pattern), those strands will continue to incorporate out-of-phase.

Composite well format (*.cwf files) – contain the uncorrected flowgrams from the image processing step or the corrected flowgrams from the signal processing step.

Control DNA – fragments of DNA with a known sequence used in each sequencing run to determine the accuracy of the sequencing reaction signal intensity translation into basecalled sequence information.

Control DNA tab – GS Run Browser application tab, reports accuracy results for the Control DNA beads in terms of the % of reads that match their reference sequence at 95%, 98% and 100% accuracy, per PTP Region and total, *via* a histogram plot. The results can be viewed across all or for individual Control DNA sequences, for a given base pair length (Base Pair selection) over which the match is calculated.

D

Data processing – the processing of sequencing run raw images (*.png or *.pif files) to produce high quality basecalled reads in composite well format (*.cwf files) and standard flowgram format (*.sff files) for further data analysis. Full data processing is carried out in two steps; image processing and signal processing.

- **Image processing** – the process of converting image data into raw flow signals for each active well of the PTP device, where an active well contains a DNA fragment that produced light due to base incorporations during the sequencing run.
- **Signal processing** – the process of correcting the raw flow signals, trimming or rejecting reads using quality filters, and basecalling.

Data processing options – users can control data processing with respect to when to process (during or post sequencing run), where to process (on-instrument, on a GS Junior attendant PC, or on an external datarig), which steps to process (image processing, signal processing, full processing), and what to process (standard shotgun/paired end library or amplicon library).

Data processing pipeline – the series of data processing commands sent to GS Run Processor, as specified by one of the XML-based pipeline script files located in *installDir*/apps/gsRunProcessor/etc/gsRunProcessor/, where installDir is the main software installation path (*e.g.* /usr/local/rig/ on the GS FLX+ Instrument or /opt/454/ on the GS Junior attendant PC).

Dot – a block of negative flows (denoted as 'N' in a DNA sequence) that is ended by a positive flow of one of the nucleotides in the block, or started and ended by positive flows of the same nucleotide.

E

Environmental variable – a dynamically defined relationship, usually a [key, value] pair, used by a computer operating system to affect the way running processes will behave on a computer.

F

Filters tab – GS Run Browser application tab, reports statistics on the read quality filters used to process the signals into high quality (HQ) reads for library and control wells. Includes a histogram of the number of key pass wells, per PTP Region and total, for the % wells that passed all filters and the % wells that failed the Dot (null – not sufficient signal from DNA fragment on the bead) and/or Mixed (more than one DNA fragment on a bead) filters.

Flow – during a sequencing run, nucleotides are flowed sequentially across the PTP device, one at a time, as controlled by the run script. When the flowed nucleotide is complementary to the next homopolymer (including a single nucleotide) on the DNA template in any given well, the polymerase extends the nascent DNA strand in that well. Addition of one or more nucleotide(s) release(s) a corresponding number of pyrophosphate (PPi) molecules. One molecule of ATP is synthesized for each PPi released, causing a flash of light (signal) whose intensity is proportional to the number of nucleotides incorporated.

- **Key flows** – the first few nucleotide flows needed to sequence the library and control sequence keys. For the flow order 'TACG' and the key 'TCAG', the key flows would be **T**-A-**C**-G-T-**A**-C-**G** (incorporation in bold) and consist of eight flows.
- **Negative flow** – a well-specific attribute denoting a nucleotide flow where no signal is detected and thus no nucleotide incorporation is assumed.
- **Positive flow** – a well-specific attribute denoting a nucleotide flow where signal is detected and the intensity of the signal is related to the number of nucleotides incorporated.

Flow cycle – an invariant flow set consisting of exactly four nucleotide flows, repeating in a specific flow order.

Flow list – the series of nucleotide flows during a sequencing run, as specified by the run script.

Flow order – the repeated sequence of nucleotides flowed during each flow set of a cyclic flow pattern sequencing run, generally 'TACG'.

Flow pattern – the pattern of nucleotide flows during a sequencing run, as determined by the choice of run script.

- Cyclic flow pattern – a pattern of nucleotide flows characterized by a repeated cycle of four nucleotide flows, with each cycle (flow set) defined by a specific flow order.
- Acyclic flow pattern – a pattern of nucleotide flows characterized by a pattern that is not cyclic.

Flow set – the smallest group of nucleotide flows at any point in a flow list that includes at least one flow of each of the four nucleotides, with the simplest case being a four nucleotide flow cycle in a cyclic flow pattern.

Flowgram – data processing extracts information about the signal intensity in each well, over all flows. The signal intensity for each flow is plotted as a function of flow order, yielding a flowgram for the well. The signal intensity is proportional to the number of bases added (linear relationship); if no nucleotides is extended in that well during a flow, the signal will be very low (background); if one nucleotide is added, the signal will be similar in intensity to the key signal; if more than one nucleotide is added, the height of the signal will be correspondingly higher.

Flowgram gap – adjustments introduced to generate a common flowgram flow list that can be used to display the relative alignment between read and reference/consensus flowgrams; also known as a cycle shift when referring to cyclic flow pattern reads.

- Reference flowgram gap – a block of negative flows identified in the reference flowgram that is associated with a putative SNP or indel that is 'missing' positive flows in the reference relative to the read. The shaded flows would have been classified as a dot (too many negative flows in a row) if the flowgram had been derived from an actual sequencing run. Subtracting these shaded flows results in a theoretical flow list that could have been used to generate the reference flowgram.
- Read flowgram gap – a block of negative flows inserted into the read flowgram that is associated with a putative SNP or indel that requires 'extra' positive flows in the reference relative to the read. The shaded flows are duplicated from the adjacent flow list to aid in alignment with the reference. Subtracting these shaded flows results in the actual flow list used to generate the read flowgram.

H

Homopolymer – nucleotide sequence of varying length consisting of one uninterrupted nucleotide type, *e.g.* A (1-mer), A-A (2-mer), A-A-A (3-mer).

I

Instrument procedure wizard – a software application on the GS Junior and GS FLX+ instruments used to set up and launch a sequencing run.

Inter-well crosstalk correction – corrects individual wells for the additional signal intensity conveyed by neighboring high intensity signal wells.

K

Key – the sequencing key is a known sequence of four nucleotides located immediately downstream from the sequencing primer and, therefore, the first to be sequenced in each well.

- **library key** – the sequencing key used for the DNA library being sequenced. **'TCAG'** or **'GACT'**
- **control key** – the sequencing key used for the Control DNA used in the sequencing run. **'CATG'** or **'ATGC'**
  (see Section 7.1 for details)

Key sequence – the first bases of a sequencing key, used for matching the initial signal flow information from a well to a well categorization; Key Pass – matches a key, Fail – does not match a key, Library – matches the library DNA key, Control – matches the Control DNA key. Only the first three nucleotides are used because the fourth base to be incorporated may incorporate as a homopolymer instead of a single nucleotide, depending on the DNA fragment sequence, thus complicating the matching algorithm.

L

Legacy files, legacy formats – files and file formats generated by previous 454 Sequencing system software versions. Conversion of current formats to legacy formats and from legacy formats to current formats is enabled in some cases. See the GS Reporter and SFF Tools Sections of the 454 Sequencing System Software Manual for more details.

Library – a library is a collection of DNA fragments representative of the entire DNA sample to be sequenced. Each library is created from user-supplied purified DNA.

M

Modal read length – the read length in a Read Length Profile that occurs most frequently.

N

N50 contig size – the contig size for which 50% of bases in an assembly are contributed by contigs of this length or longer.

N50 read length – the shortest passed filter read length for which 50% of bases (rounded to the nearest whole percent) are contributed by reads of this length or longer. More generally, Nxx is the shortest read length for which xx% of bases are contributed by reads of this length or longer. For example, N1 is the read length for which only 1% of bases are contributed by reads of this length or longer (an effective measure of the longest read lengths obtained).

N50 scaffold size – the scaffold size for which 50% of scaffold bases (including bases inferred from gaps) are contributed by scaffolds of this length or longer.

Normalization flow – the initial ATP flow causes a simultaneous flash of light from all enzyme-active wells and is used to define the PTP device loading regions and the shape of the background across the plate.

Nucleotide normalization – a signal processing correction that normalizes the signal strengths of different base incorporations.

Nucleotide incorporation – polymerase extension of the nascent DNA strand in a well by a complementary nucleotide flowed across the PTP device.

O

Overview tab – GS Run Browser application tab, contains summary data of the sequencing run, summary data of the processing results, if carried out, and the GS Run Processor Manager used to launch data processing or reprocessing jobs for the currently selected run data set.

P

PHRED – a software program that reads DNA sequencer trace data, calls bases, assigns quality values to the bases, and writes the basecalls and quality values to output files.

PHRED score – a quality score logarithmically linked to the error probabilities for basecalling sequences derived from signal data for known sequences. Calculated based on several parameters related to peak shape and peak resolution at each base. Methodology described by Ewing and Green (*Genome Research*, 8: 186-194, 1998)

PicoTiterPlate device – a plate containing the DNA being sequenced. The PTP device is the interface between the fluidics and optics subsystems of the GS Junior and GS FLX+ systems. The side of the PTP device that is in contact with the fluidics subsystem, which delivers the reaction reagents, contains ~1.8 million microscopic (18.5 picoliter) wells in which the sequencing reactions take place. Each well is designed to contain a single, unique library bead carrying a clonally amplified DNA fragment. The bottom of each well is made of an optical fiber, which transmits light produced by the sequencing reaction across the thickness of the PTP device, to a camera, the optics subsystem which captures the raw images of the PTP device during each nucleotide flow. Each well wall is lined with a metalized finish to reduce well-to-well crosstalk and signal interference. The wells are organized into regions (or loading gaskets) of different configurations (2-16 regions), allowing flexibility in the depth and breadth of information captured in any single sequencing run.

Primer – a short strand of nucleic acid (an oligonucleotide) that serves as a starting point for DNA synthesis by DNA polymerase. In the 454 Sequencing system, primers are used during both emPCR (amplification primers) and during the sequencing run (sequencing primer). An amplicon fusion primer (used in pairs, with the Lib-A emPCR kit) is a 5'-adaptor fused to a template-specific primer at each end of a target sequence of interest.

Pyrosequencing – 'sequencing by synthesis'- sequencing of a single-strand DNA by synthesis of the complementary strand one base pair at a time. The added nucleotide pair is detected and coded.

Q

Quality score – a PHRED-like binned scored associated with a basecall, calculated for each nucleotide or homopolymer translation from signal space, based on the specific and local signal properties of the called base relative to a pre-calibrated 454 control DNA training set.

Quality filter – any of a series of filters used to assess and retain only high quality reads for further data analysis; including KeyPass, Primer-Dimer, Dot, Mixed, Ambiguous Read, Signal Intensity, Signal Trimback, Adaptor Trim, and the three Valley Filters (Counting, Scoring, and Trimback).

Q20 test – average base quality score > 20.

Q20 Read Length – read length at which the bases are 99% accurate or higher for all preceding bases.

R

Raw intensity – uncorrected signal intensity from a nucleotide flow for a read in a well of the PTP device.

Raw flow signals – uncorrected signal intensities from a sequencing run for a read in a well of the PTP device.

Raw image files (*.png) – images of the PTP device taken during each nucleotide flow of a sequencing run, capturing the light emitted from each active well due to nucleotide incorporation sequencing reactions.

Read – the sequence trace data derived from the flow signal of a DNA template.

- **Raw read** – the uncorrected sequence trace data derived from the flow signal of a DNA template.
- **High quality read** – the corrected sequence trace data derived from the flow signal. Corrections are applied for known chemical, biological and system artifacts and trimming is applied to retain the high quality signal intensity portion of the read.
- **Trimmed read** – a read that has had a portion of its 3' end trimmed to retain the relevant and/or high quality signal portion of the DNA template.
- **Unrecognized read** – reads which begin with the Control DNA sequencing key (CATG or ATGC; see Section 7.1 for details) but do not match any of the corresponding Control DNA reference sequences.

Read trace information – the linear sequence of a DNA template derived from signal data obtained during a sequencing experiment.

Read length – the length in nucleotides of a read.

Read length profile – a plot displaying the number of reads at each observed value of read length.

Read rejecting filters – applied as a pass/fail test and quickly discards no-information or low information active wells; KeyPass, Short Signal, Dot, Mixed, and Ambiguous Read .

Read trimming/rejecting filters – applied to retain the high information content portion of a read; Signal Intensity, Signal Trimback, Adaptor Trim, and the Valley Filter.

Reads tab – GS Run Browser application tab, contains statistics on the distribution of the read length and read quality for library and control wells and summary statistics per PTP Region and total, on the number of raw wells, key pass wells, passed filter wells, total bases, read length average, standard deviation, longest read length, shortest read length, and median read length.

Reagent flow event balancer – corrects anomalous signal spikes due to reagent valve events.

Run script – an instrument control (.icl) script file that specifies the type and duration of each flow during a sequencing run. The sequencing run script is automatically selected based on the instrument, PTP device, sequencing kit, flow pattern, and number of cycles specified in the Instrument Procedure Wizard of the GS Sequencer or GS Junior Sequencer GUIs. Other specialized run scripts can be selected using the Custom script service option. Run script files are located, by default, in /usr/local/rig/runScripts/ on the GS FLX+ instrument or /opt/454/apps/gsSequencer/runScripts/ on the GS Junior attendant PC.

S

Signal droop correction – correct for signal reduction during the eight-hour sequencing run exposure.

Signal flowgram – the linear sequence of a DNA template in a well of the PTP device during a sequencing run, inferred from the signal intensity for each nucleotide flow, plotted as a function of flow order.

Standard flowgram format file (***.sff**) – contains data on all the reads resulting from the data processing including the quality and adaptor trimming positions, flowgram information, called bases and their quality scores. This is a text file starting with a common header section, followed by a header and data section for each read.

Signal per base – the signal intensity calculated for a single nucleotide incorporation during a sequencing run, averaged over all homopolymer nucleotide incorporation signals in all Control DNA wells. It can be used to remove ghost wells, active wells with signal below 1-mer incorporations, and to estimate the number of copies of DNA template per bead in each well.

Signal processing – process of applying filters, corrections and trimming of the raw flow signals to produce high quality sequence information.

Signals tab – GS Run Browser application tab, contains statistics on the distribution of well intensity, filtered well intensity and N-mer (number or bases) signals recorded for each flow for control and library wells.

T

Titration – a process of determining or achieving the correct ratio of chemical reactants. For DNA sequencing runs, titration can be used to determine the optimal amount of DNA library to use in emPCR amplification.

V

Valley – off-peak signal intensity relative to the nearly quantized signal intensity per nucleotide homopolymer incorporation.

W

Well – conceptually, a well is a location on the raw image of the PTP device at which signal was observed during a nucleotide flow of a sequencing reaction. Physically, a well is an octagonal 18.5 picoliter compartment on the PTP device in which the sequencing reactions take place. Each well is designed to contain a single, unique library bead carrying a clonally amplified DNA fragment. The bottom of each well is made of an optical fiber, which transmits light produced by the sequencing reaction across the thickness of the PTP device, to the camera (optics subsystem) and each well wall is lined with a metalized finish to reduce well-to-well crosstalk and signal interference.

- **Raw wells** – wells identified as having measurable signal intensity during a sequencing run but that have not been filtered or corrected for sequence information content relevance.
- **Key Pass wells** – wells that have signal intensity matches of the initial nucleotide flows to known DNA Adaptor sequences used in the sequencing run.

Well density – a property estimating the closeness of active wells across the PTP device, based on the calculated signal per base.

Well density correction – calculates the signal per base to filter out ghost wells.

Well flowgram – a plot of the linear sequence of a DNA template inferred from the signal intensity observed for each nucleotide flow, plotted as a function of flow order. Accessible from the Wells tabs of the 454 Sequencing system software applications.

- **Consensus flowgram** – the flowgram of a Control DNA sequence constructed by averaging, for each nucleotide flow, the read flowgram signals of the reads identified for that reference sequence. Accessible from the Control DNA tabs of the 454 Sequencing system software applications.
- **Location (raw) flowgram** – a raw flowgram constructed by computing, for each image in the sequencing run, the average raw (non-corrected) signal intensity for the 9 pixels surrounding the selected location, and plotting these averages against the succession of reagent flows. (Note that this calculation does not give any consideration to the notion of 'wells'.) Accessible from the Wells tabs of the 454 Sequencing system software applications.
- **Subtraction (raw) flowgram** – a plot of the subtraction of any two location flowgrams, flow by flow. Created by the use of subtraction pins in location flowgram plots, accessible from the Wells tabs of the 454 Sequencing system software applications.
- **Tri-flowgram** – a plot of the subtraction of an idealized or consensus flowgram from an observed flowgram for a Control DNA reference sequence, flow by flow. Accessible from the Control DNA tabs of the 454 Sequencing system software applications.

Well status – a categorization of the signal intensity of a well based on a variety of criteria. Listed in several output files and viewable in several GUIs of the 454 Sequencing system software applications. Passed Filter – Library read that passed all quality filters; No Key – Identified as a well (generates signal), but not one with recognizable data; Failed – Library read that failed any of the quality filters; Control DNA – Control DNA read; Key Pass – matches a sequencing key, library or control.

Wells tab – GS Run Browser application tab, contains raw images of the PTP device which can be displayed for various well categories and the selected base flow. Also reports summary statistics of the average well density of the PTP device, per-region and total, for raw well, and key pass wells.

INDEX

Published by
454 Life Sciences Corp.
A Roche Company
Branford, CT 06405
USA

Notice to Purchaser
For patent license limitations for individual products please refer to: **www.technical-support.roche.com**.

For life science research only. Not for use in diagnostic procedures.

(9) 0613