



# Genome Sequencer System

Application Note No. 7 / April 2007



## Impact of Quality Filter Settings on Sequencing Accuracy



# Impact of Quality Filter Settings on Sequencing Accuracy

Corresponding author: Vinod Makhijani, 454 Life Sciences Corporation, Branford, CT, USA, Email: vmakhijani@454.com

## Introduction

The sequencing data (*i.e.*, individual reads) generated from a Genome Sequencer FLX System run are assessed for quality and accuracy with the help of various quality filters, after the image and signal processing algorithms have been applied, without *a priori* knowledge of the genome or template being sequenced. High-quality data is identified for subsequent assembly or mapping.

**The Valley Filter** is part of the signal intensity filter that filters out reads of questionable quality (Section 4.2.2.1 in the GS FLX Data Processing Software Manual). This filter can be adjusted to increase or decrease the stringency with which reads are filtered. For Genome Sequencer FLX (GS FLX) sequencing runs, the Valley Filter in the data analysis pipeline (Software Version 1.1.01 & 1.1.02) identifies ambiguous base calls (valley flows) within reads that have passed other filters (described in the GS FLX Data Processing Software Manual, Section 4.2), based on signal intensity, over the first 320 nucleotide flows (*i.e.*, 80 cycles of four successive nucleotide flows). The reads are sorted by their count of such valley flows. Good-quality reads have fewer valley flows. Reads with four or more valley flows are rejected by the filter.

Users have the ability to modify the stringency of the Valley Filter after the sequencing run data has been processed (instructions are provided in the

GS FLX Data Processing Software Manual, section 4.2.3). The stringency of the filtered reads can be controlled by modifying either the total number of flows over which valley flows are counted (using more flows improves the chance of reaching the rejection threshold, hence higher stringency), or the number of valley flows used as rejection threshold (a lower threshold is easier to meet, hence higher stringency). Increasing the stringency tends to improve the average accuracy of the filtered individual reads, though yields fewer reads; decreasing the stringency tends to increase the number of filtered high-quality reads at the expense of read accuracy. The overall impact of lower yield versus lower accuracy on consensus accuracy of mapped genomes for resequencing applications, and for *de novo* assembly of genomes, has not been studied yet.

In this application note, we evaluate the impact of the trade-off between yield and individual read accuracy, resulting from increased or decreased quality filter stringency, on the consensus accuracy of large contigs, assembled *de novo* or by mapping to a reference genome sequence at different levels of oversampling (or average coverage depth). Sequencing data for the *Escherichia coli* (*E. coli*) *K-12* genome from a standard GS FLX run was used as a test case in this study.

## Materials and Methods

### Materials

#### Equipment:

Genome Sequencer Instrument FLX  
GS FLX: Software Version 1.1.01

#### Reagents from Roche Applied Science:

##### Sample Preparation:

GS DNA Library Preparation Kit,  
GS emPCR Kit I (Shotgun)

##### Sequencing:

GS LR70 Sequencing Kit  
GS PicoTiterPlate Kit (70x75)\*

\* Bead Loading Gasket with 4 "medium" regions, each 14x43 mm  
Each "medium" region has approximately one-fifth the sequencing area of a large PicoTiterPlate.



**A detailed list of all required equipment and reagents is provided in the Genome Sequencer FLX User's Manuals and Guides.**

## Method of Approach

The sequencing method used here was a standard Genome Sequencer FLX run, performed with a GS PicoTiterPlate Kit (70x75) with four medium regions. *E. coli* library beads, obtained from the same batch of emulsion PCR, were used in all four regions. The sequencing data was processed with the standard data analysis pipeline using default values of the quality filter parameters (*i.e.*, reads with four or more valley flows over the first 320 nucleotide flows).

The same data set was then reprocessed by modifying the stringency of the filter settings. The most-stringent setting currently available in the software version 1.1.01 & 1.1.02, in which reads with two or more valley flows over the first 400 nucleotide flows are rejected, was used. The data was also processed using the least-stringent setting recommended in the GS FLX Data Processing Software Manual, in which only those reads with six or more valley flows over the first 168 nucleotide flows are filtered.

The high-quality reads obtained from all three filter settings were aligned against the *E. coli* K-12 reference sequence (approximately 4.64-megabase genome, obtained from ATCC®, Number 700926D™), using the GS Reference Mapper (Figure 1) in the Version 1.1.01 data processing software, to generate a consensus sequence of the *E. coli* K-12 library. For each level of filter stringency, the number of high-quality reads used for mapping was incrementally increased by using data from:

- (A) Region 1 only;
- (B) Region 1 and Region 2;
- (C) Regions 1, 2 and 3;
- (D) All 4 Regions.

Results were assessed for individual read accuracy, and for accuracy of the large (>500 base pairs) consensus base-called contigs through alignment against the reference sequence.

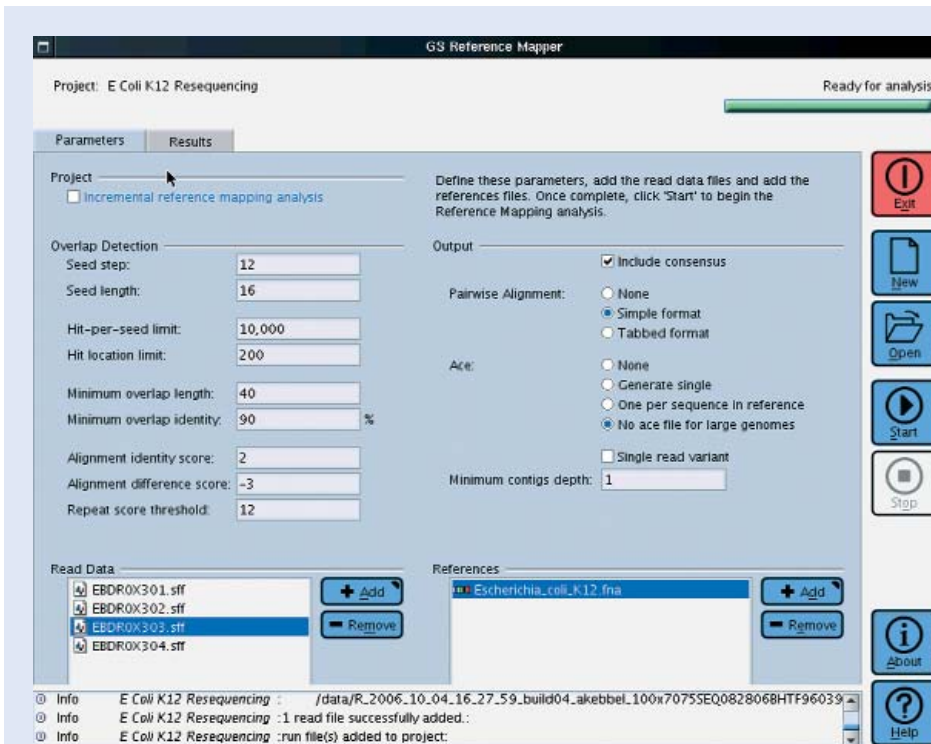
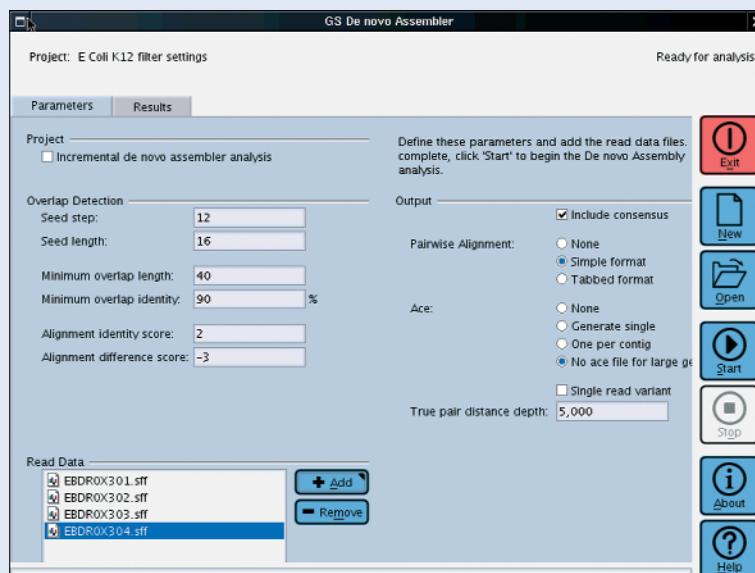


Figure 1: GS Reference Mapper

The GS *De Novo* Assembler (Figure 2) in the Version 1.1.01 data processing software was also used to assemble contig consensus sequences from the same data set. Once again, for each of the three levels of filter stringency (default, most, and least), the number of high-quality reads used for assembly was incrementally increased by using data from:

- (A) Region 1 only;
- (B) Region 1 and Region 2;
- (C) Regions 1, 2 and 3;
- (D) All 4 Regions.

The accuracy of the large contig consensus sequences was ascertained by mapping them against the *E. coli* K-12 reference sequence using MegaBLAST.



**Figure 2: GS *De Novo* Assembler**

## Results and Discussion

### Reference Sequence Mapping

Table 1 shows the results of the resequencing analysis for all three filter settings at different levels of oversampling, corresponding to the utilization of sequencing reads from one region, two regions, three regions, and four regions of the PicoTiterPlate device. Listed in the table are the average coverage depths for each case, the number of high-quality reads (and corresponding high-quality bases) and the actual number of bases that map against the reference *E. coli* sequence. The quality of individual reads is represented by the average map length and

the average cumulative read error over the first 150 and 238 bases. Higher-quality reads, in general, have longer mapped lengths and fewer read errors. The table also lists parameters that reflect the quality of the consensus base-called contigs (number of contigs, consensus genome coverage, and consensus accuracy).

Data shows that when the filter stringency increases from default to most stringent, the number of high-quality filtered reads decreases with a concurrent increase in average cumulative accuracy of the individual reads. The converse is true when the filter

stringency decreases. The table clearly shows the consistency in quality of the individual reads across the four regions of the PicoTiterPlate device. There is very little variation in average cumulative read error rates when reads from more than one region are used for mapping, which implies that all four regions produce high-quality reads of comparable average accuracy.

With regard to construction of contigs and computing a consensus base-call sequence from the signals of the aligned reads, the table clearly shows the impact of yield (number of high-quality bases used for mapping) and individual read accuracy on consensus sequence accuracy. At low levels of oversampling when only one region is used ( $3\times$  -  $6\times$  average coverage depth), the default setting gives the highest consensus accuracy. The most-stringent setting results in the lowest consensus accuracy despite the extremely low individual read error rates (0.07%

cumulative error over the first 238 bases), primarily because of the very low coverage depth ( $3.3\times$ , as compared to  $4.8\times$  and  $5.9\times$ , for the default, and least-stringent setting, respectively). The least-stringent setting has lower consensus accuracy than the default setting despite higher coverage depth because the underlying reads used to build consensus sequences have more error. When the levels of oversampling are comparable, as is the case with the two-region/most-stringent data ( $6.4\times$ ) and the one-region/least-stringent data ( $5.9\times$ ), consensus error is much higher (almost  $3\times$ ) for the latter case due to the lower-accuracy reads. At higher coverage depths of  $13\times$  or more (*i.e.*, three or more regions for the default or least-stringent setting, and all four regions combined for the most-stringent setting), consensus accuracies are comparable ( $>99.998\%$ ) in all cases; the impact of individual read accuracy on consensus accuracy is much weaker than at lower coverage depths of  $6\times$  or less.

## Results and Discussion continued

**Table 1: Resequencing data for the *E. coli* K-12 genome at different levels of oversampling under different quality filter stringency settings.**

Number of Medium Regions	1	2	3	4
Default Filter Setting				
Average Coverage Depth	4.8	9.3	14.5	19.0
Number of High-Quality Reads	89,323	174,225	271,203	356,309
High-Quality Bases (MB)	22.49	43.84	68.30	89.75
Mapped Bases (MB)	22.06	43.03	67.06	88.13
Average Map Length (bp)	251	251	251	251
Cumulative Read Error @ 150 bp	0.11%	0.11%	0.11%	0.11%
Cumulative Read Error @ 238 bp	0.19%	0.19%	0.18%	0.18%
Number of Contigs	797	84	61	62
Consensus Coverage	97.15%	98.57%	98.65%	98.69%
Consensus Accuracy	99.9682%	99.9946%	99.9981%	99.9986%
Most-Stringent Setting				
Average Coverage Depth	3.3	6.4	10.2	13.4
Number of High-Quality Reads	61,824	119,468	191,050	251,118
High-Quality Bases (MB)	15.55	30.03	48.07	63.19
Mapped Bases (MB)	15.25	29.47	47.21	62.06
Average Map Length (bp)	251	251	251	251
Cumulative Read Error @ 150 bp	0.05%	0.05%	0.04%	0.04%
Cumulative Read Error @ 238 bp	0.07%	0.08%	0.07%	0.07%
Number of Contigs	1958	410	102	79
Consensus Coverage	90.55%	97.95%	98.54%	98.63%
Consensus Accuracy	99.9591%	99.9891%	99.9969%	99.9981%
Least-Stringent Setting				
Average Coverage Depth	5.9	11.6	17.8	23.4
Number of High-Quality Reads	111,446	219,302	336,666	443,286
High-Quality Bases (MB)	27.84	54.74	84.14	110.79
Mapped Bases (MB)	27.25	53.59	82.40	108.52
Average Map Length (bp)	249	249	250	250
Cumulative Read Error @ 150 bp	0.36%	0.37%	0.36%	0.36%
Cumulative Read Error @ 238 bp	0.50%	0.52%	0.50%	0.49%
Number of Contigs	368	62	60	62
Consensus Coverage	98.11%	98.61%	98.68%	98.71%
Consensus Accuracy	99.9678%	99.9961%	99.9983%	99.9984%

**De Novo Assembly**

Table 2 shows the results of the *de novo* assembly of the high-quality read data at different levels of oversampling (~ 3× – 24×) for all three filter settings. The trends, related to the trade-off between yield and individual read accuracy, are similar to that observed in the resequencing analysis. Once again, at low levels of oversampling (3× – 6×), when only one region is used, the default setting gives the highest consensus accuracy. Lower consensus accuracy for the higher, and lower, strin-

gency setting most likely results from lower oversampling, and lower read accuracy, respectively. Also, at comparable levels of oversampling (6.5× for the two-region/most-stringent data and 6× for the one-region/least-stringent data), consensus error is much higher (~ 2.5×) for the latter case due to the lower-accuracy reads. In the present case, at higher coverage depths of ~15× or more (*i.e.*, three or more regions for the default or least-stringent setting), consensus accuracies are quite high (>99.997%) in all cases.

**Table 2: De novo assembly data for the *E. coli* K-12 genome at different levels of over-sampling under different quality filter stringency.**

Number of Medium Regions	1	2	3	4
Default Setting				
Average Coverage Depth	4.8	9.5	14.7	19.3
Assembly Contigs	1473	209	116	106
Assembly Coverage	91.81%	97.42%	97.42%	97.59%
Assembly Accuracy	99.9898%	99.9929%	99.9973%	99.9970%
Most-Stringent Setting				
Average Coverage Depth	3.4	6.5	10.4	13.6
Assembly Contigs	2422	832	221	139
Assembly Coverage	75.07%	95.14%	97.27%	97.41%
Assembly Accuracy	99.9864%	99.9944%	99.9956%	99.9964%
Least-Stringent Setting				
Average Coverage Depth	6.0	11.8	18.1	23.9
Assembly Contigs	886	137	107	110
Assembly Coverage	95.24%	97.39%	97.66%	97.63%
Assembly Accuracy	99.9858%	99.9957%	99.9971%	99.9981%

**Trade-Off Between Yield and Single-Read Accuracy at Different Filter Settings**

The GS Reference Mapper and GS *De Novo* Assembler both generate consensus base calls of the contigs by averaging the processed flow signals for each nucleotide flow (rather than the individual base calls) included in the alignment [1]. At higher coverage depth, there are more processed signals available for averaging at any given flow position. This effectively reduces the impact of individual flow signal errors on the consensus base calls (the probability of all reads having the same error at the same base position is generally low, unless there are sequence context-specific errors). When the coverage depth is low, the accuracy of each individual read used in the averaging becomes important.

A standard Genome Sequencer FLX run produces, on average, approximately 100 MB of high-quality data at the default filter setting. Increasing or decreasing the stringency of the filter setting will impact the number of high-quality bases available for mapping or for *de novo* assembly. For the pre-

sent case, we obtain 30% lower high-quality bases compared to default when the most-stringent filter setting is used. The corresponding number is 25% higher when the least-stringent filter setting is used. In other sequencing runs, depending upon the quality of the emPCR sample preparation, complexity of the genome, or other factors that may impact sequencing data quality, these numbers could be higher or lower. Therefore, depending upon the size of the genome being evaluated, different filter settings will result in different levels of oversampling and will also impact individual read accuracy differently. In studies where individual read accuracy is very important, more-stringent filter settings will be desirable because the accuracy of the filtered reads is higher. The default filter setting, however provides a fairly good trade-off between yield and read accuracy, and seems to work reasonably well at both low and high levels of oversampling as shown in the present case. It may, therefore, be a good starting point for analyzing data from resequencing or *de novo* sequencing runs, or for runs requiring very low individual read error.

## Conclusion

This study demonstrates how different quality filter settings within the Version 1.1.01 & 1.1.02 data analysis pipeline software affect the number of high-quality reads and the average accuracy of the individual reads. It shows how, depending upon the size of the genome being studied or the type of study being conducted, different filter settings may work well for different cases. For most applications, however, the default setting provided in the software may be a good place to start the analysis.

## References

1. Margulies, M., Egholm, M., *et al.*, "Genome sequencing in microfabricated high-density picolitre reactors" (2005) *Nature* 437:376-380.

**Notes:**

**Notes:**

**NOTICE TO PURCHASER**

RESTRICTION ON USE: Purchaser is only authorized to use the Genome Sequencer Instrument with PicoTiterPlate devices supplied by 454 Life Sciences Corporation and in conformity with the procedures contained in the Operator's Manual.

**Trademarks**

454, 454 LIFE SCIENCES, GENOME SEQUENCER, PICOTITERPLATE and emPCR are trademarks of 454 Life Sciences Corporation, Branford, CT, USA.

The ATCC trademark and trade name and any and all ATCC catalog numbers are trademarks of the American Type Culture Collection.

Other brands or product names are trademarks of their respective holders.

For more information, visit  
[www.genome-sequencing.com](http://www.genome-sequencing.com)



Diagnos**t**ics

Roche Diagnostics GmbH  
Roche Applied Science  
68298 Mannheim  
Germany  
[www.roche-applied-science.com](http://www.roche-applied-science.com)