Next Generation Targeted Amplicon Resequencing with Long Sanger-Like Read Lengths on the GS FLX+ System

Dawn Gratalo, Mohammed Mohiuddin, Roger Winer, Mark Driscoll, Birgitte Simen, James Nealis

454 Life Sciences Corporation, A Roche Company, 1 Commercial Street, Branford CT, 06405, USA

1. Abstract

The Roche 454 GS FLX+ System features the unique combination of long reads, exceptional accuracy and high-throughput, making the system well suited for genomic projects that require large quantities of long reads such as *de novo* whole genome sequencing and assembly of large, complex organisms. With the launch of v2.9 system software, the long read and high accuracy sequencing capabilities will be available for amplicon applications. Here we describe our current development program leveraging an acyclic nucleotide flow pattern, Flow Pattern B, for highly accurate targeted sequencing of amplicons 700-800 base pair and longer. The longer reads allow more direct transition of existing Sanger amplicon designs to massively parallel sequencing with minimal re-design, direct haplotype phasing across longer spans, and improved identification of complex genetic variations including large insertions, deletions and block substitutions. The technique is also promising for metagenomic studies targeting 16S and 18S rRNA subunits by allowing full coverage of 6-7 variable regions in a single uni-directional amplicon read which enables the investigator to generate accurate diversity and abundance profiles. The advancements in long read sequencing will be applied to the GS Junior System as well. Please see poster #283 entitled *Long Sanger-Like Reads on a Benchtop Next Generation Sequencing Platform – GS Junior System*

2. Flow Pattern B: Reads lengths up to 1,000 bp and beyond

Available now with software version 2.8, a new acyclic flow list is supported for use with the GS FLX Titanium Sequencing Kit XL+.

- **Flow pattern A** – a cyclic 'TACG' flow pattern with 1600 flows that generates results similar to those obtained with a 400 cycle run in version 2.6 software.
- **Flow pattern B** - an acyclic flow pattern with 1779 flows that is expected to increase read length after all signal processing filters have been applied.

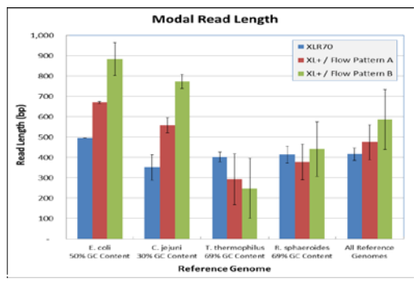


Figure 1) Comparison of Flow Patterns

Currently in development - Flow Pattern B for long amplicon sequencing to be released with v2.9 GS FLX+ System Software

Design Targets

- Robust sequencing of amplicons of 700-800 bp and beyond
- ~700,000 passed filter reads

3. Sequencing Performance

Shotgun Library : *E. coli* library run using GS FLX+ System with v2.8 software and Flow Pattern B

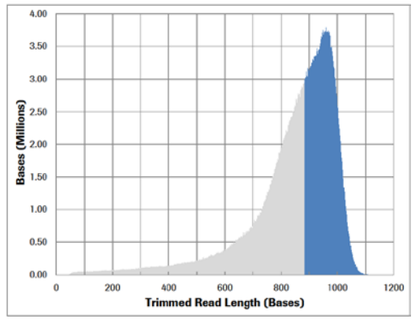


Figure 2) 50% of bases from 885 base and longer reads

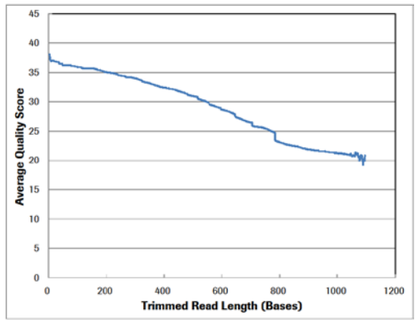
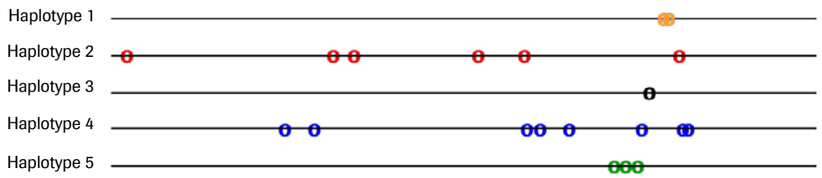


Figure 3) Average of assigned quality scores

Amplicon Library: Single 907 bp amplicon generated from 5 cell lines with different HLA types
Variants over full amplicon sequence were recovered at expected rates



Colored circles indicate positions where each haplotype has a unique variant. The 907 base pair amplicon was cloned from HLA-A and covers the intron 1 – exon2 – intron 2 – exon 3 – intron 3 regions of HLA-A.

Sequencing with prototype v2.9 software and Flow Pattern B recovered the variants at the expected frequencies across the entire 907 base pair sequence

Description	Expected Frequency	Observed Frequency of Unique Variants							
24:02:01:01 (cell line T1S1)	5%	1-A	1-B						
		5.2%	5.4%						
03:01 (cell line SAVC)	25%	2-A	2-B	2-C	2-D	2-E	2-F		
		23.1%	25.1%	24.7%	24.5%	25.0%	25.4%		
23:01 (cell line WT51)	57%	3-A							
		57.4%							
26:01 (cell line YAR)	1%	4-A	4-B	4-C	4-D	4-E	4-F	4-G	4-H
		0.7%	0.7%	1.1%	0.7%	0.8%	1.1%	0.8%	0.4%
68:01:01 (cell line LB)	12%	5-A	5-B	5-C					
		11.1%	11.1%	11.1%					

4. GS FLX+ Long Amplicon Applications

Environmental Sample: Majority of reads 1000 bp in length or greater
The 18S genes of 400 environmental samples were amplified using primers spanning approximately 1200 base pairs. The samples were processed using the published emPCR protocol for long amplicons and sequenced using the prototype v2.9 software with Flow Pattern B.

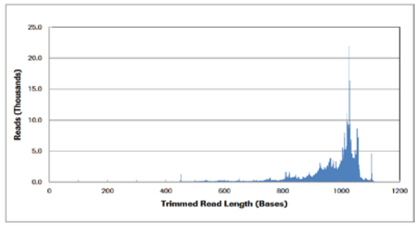


Figure 4) Amplicon read lengths of greater than 1000 bp

Cancer Gene Sequencing: High accuracy sequencing of complete 700 bp amplicon
A 700 base pair TET2-derived amplicon (containing exons 3-5 with introns), including 70 bp of adaptors, was sequenced to identify variants down to 1%. A variant was included at 1.3% to demonstrate the accurate detection of rare variants in both forward and reverse reads.

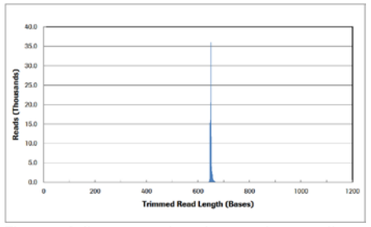


Figure 5) Sequenced entire 700 bp amplicon

Variant Frequency	1.3%
Variant Frequency Observed	1.2%
% forward reads	47%
% reverse reads	52%

Viral Deep Sequencing: Linked 837 bp HIV variants detected at 1.2%
Single HIV1 amplicon with linked variants L10R and L63P present at a low frequency were sequenced to test the detection capabilities. The link variant was detected bi-directionally .

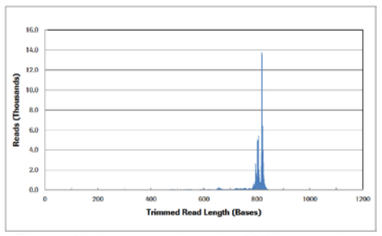


Figure 6) Sequenced entire 837 bp amplicon

Variant Frequency Observed	1.2%
% forward reads	54%
% reverse reads	46%

5. Summary

The acyclic nucleotide flow pattern on the GS FLX+ Sequencing System provided robust sequencing of long random fragment samples with the launch of v2.8 software. Now, with the launch of the v2.9 software in the second quarter of 2013, long read sequencing will be supported on amplicon samples as well. This enables the transition of Sanger amplicon designs onto the GS FLX+ Sequencing System with minimal redesign. There are no changes to any kits or instruments required and existing protocols for library preparation and emulsion PCR are used.