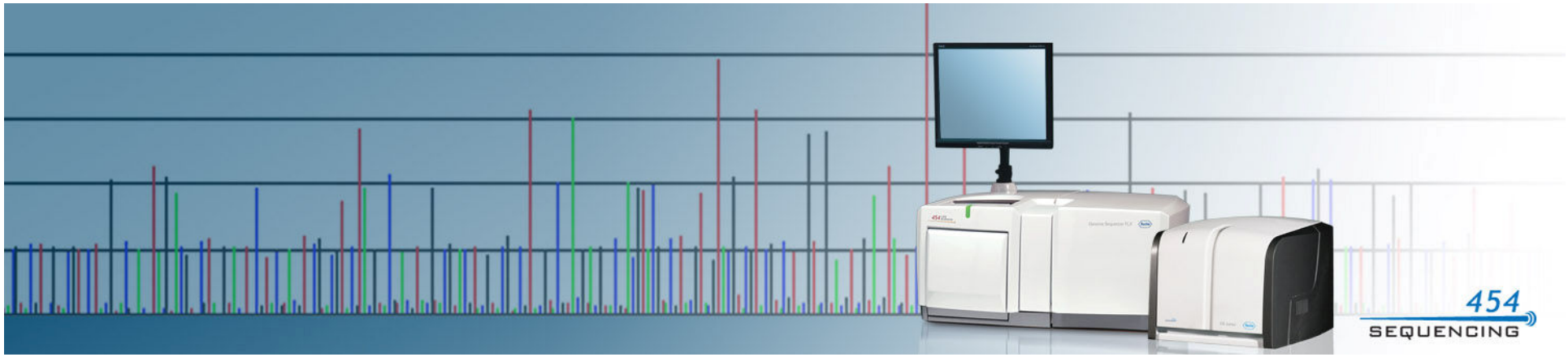
www.454.com

454 Sequencing Systems
GS Junior, GS FLX, and GS FLX+

Teri Rambo Mueller, Bioinformatics FAS

Sepideh Rogers, Molecular Biology FAS

Goli Shariat, Key Account Manager



IMPORTANT NOTICE

Intended Use

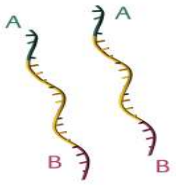
Unless explicitly stated otherwise, all Roche Applied Science and 454 Life Sciences products and services referenced in this presentation are intended:

For Life Science Research Only

Not for Use in Diagnostic Procedures

454 Sequencing Systems

Process Overview



1. Library Preparation

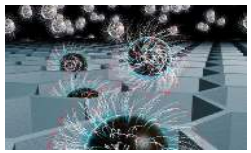
Incorporate 454 A & B sequences

ONE FRAGMENT

2. Emulsion PCR

Clonal amplification of library fragments

ONE BEAD



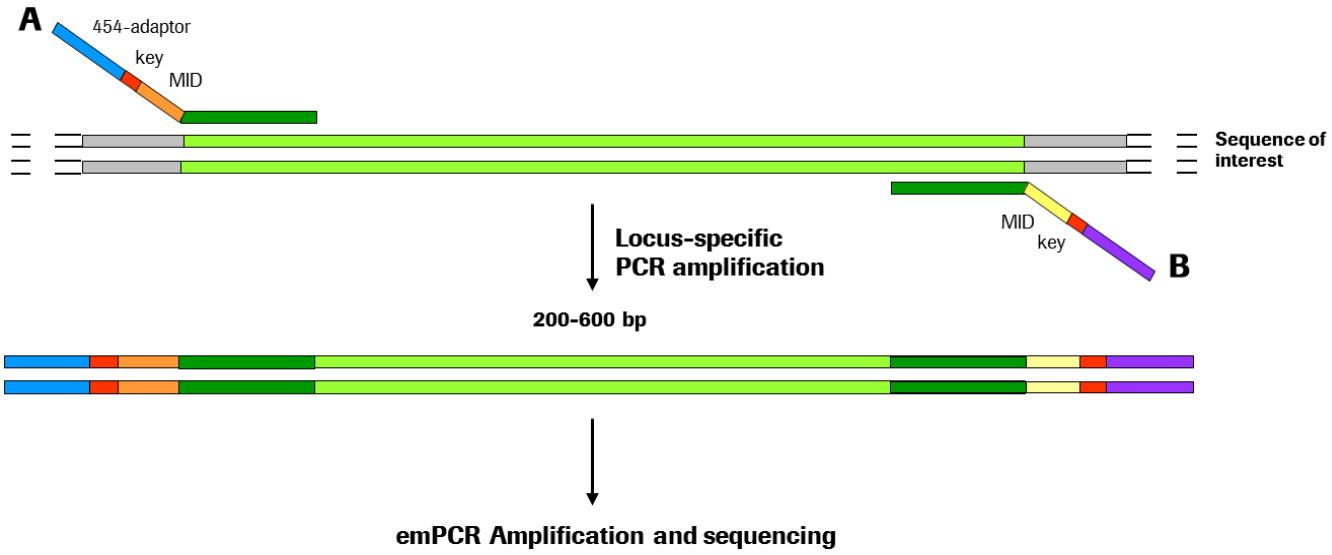
3. Sequencing

Load PicoTiterPlate device

ONE WELL = ONE READ

Library Preparation

Fusion Primer Approach

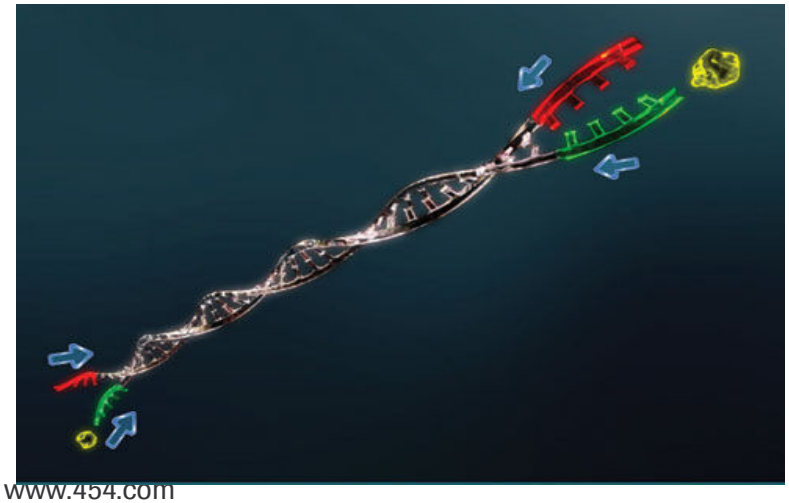


Consists of 454 A or B sequence + key (library identification) + MID (barcoded multiplexing) + target specific primer sequence

Experimental designs for unidirectional or bidirectional amplicon sequencing

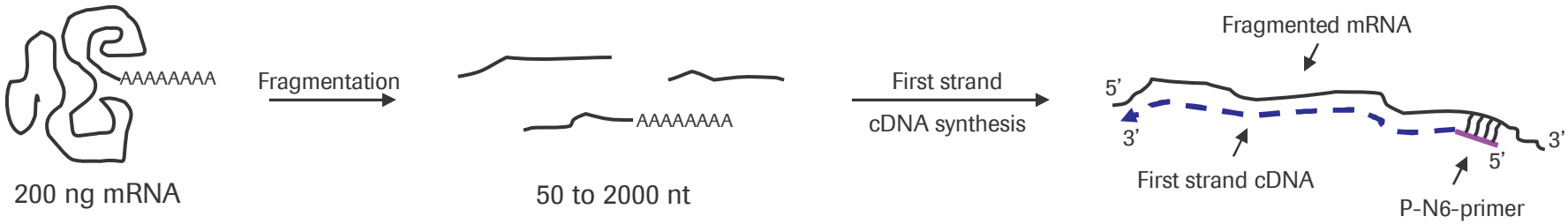
Library Preparation
Rapid Ligation Method

Fragmentation of long range PCR product, RNA, genomic DNA



Ligation of A/B adaptor sequences with fluorescent label

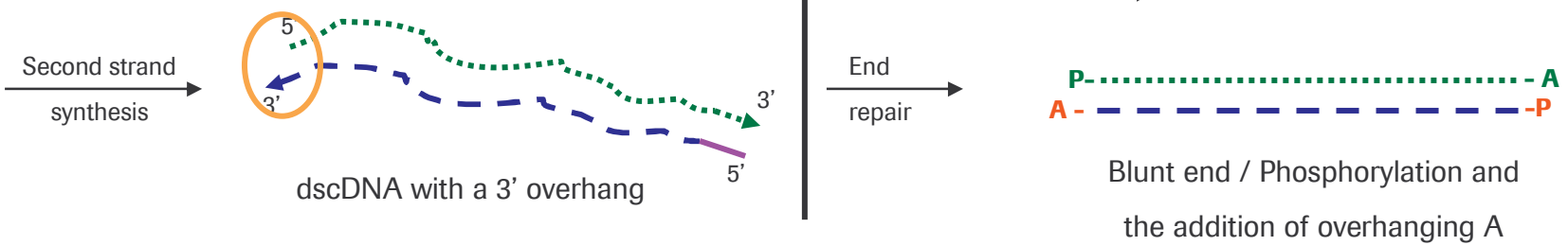
Library Preparation
cDNA Rapid Ligation



200 ng mRNA

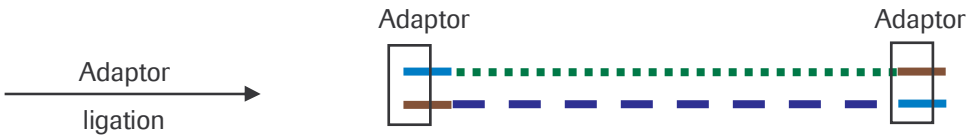
50 to 2000 nt

Enters the Rapid Library Prep Protocol



dscDNA with a 3' overhang

Blunt end / Phosphorylation and the addition of overhanging A



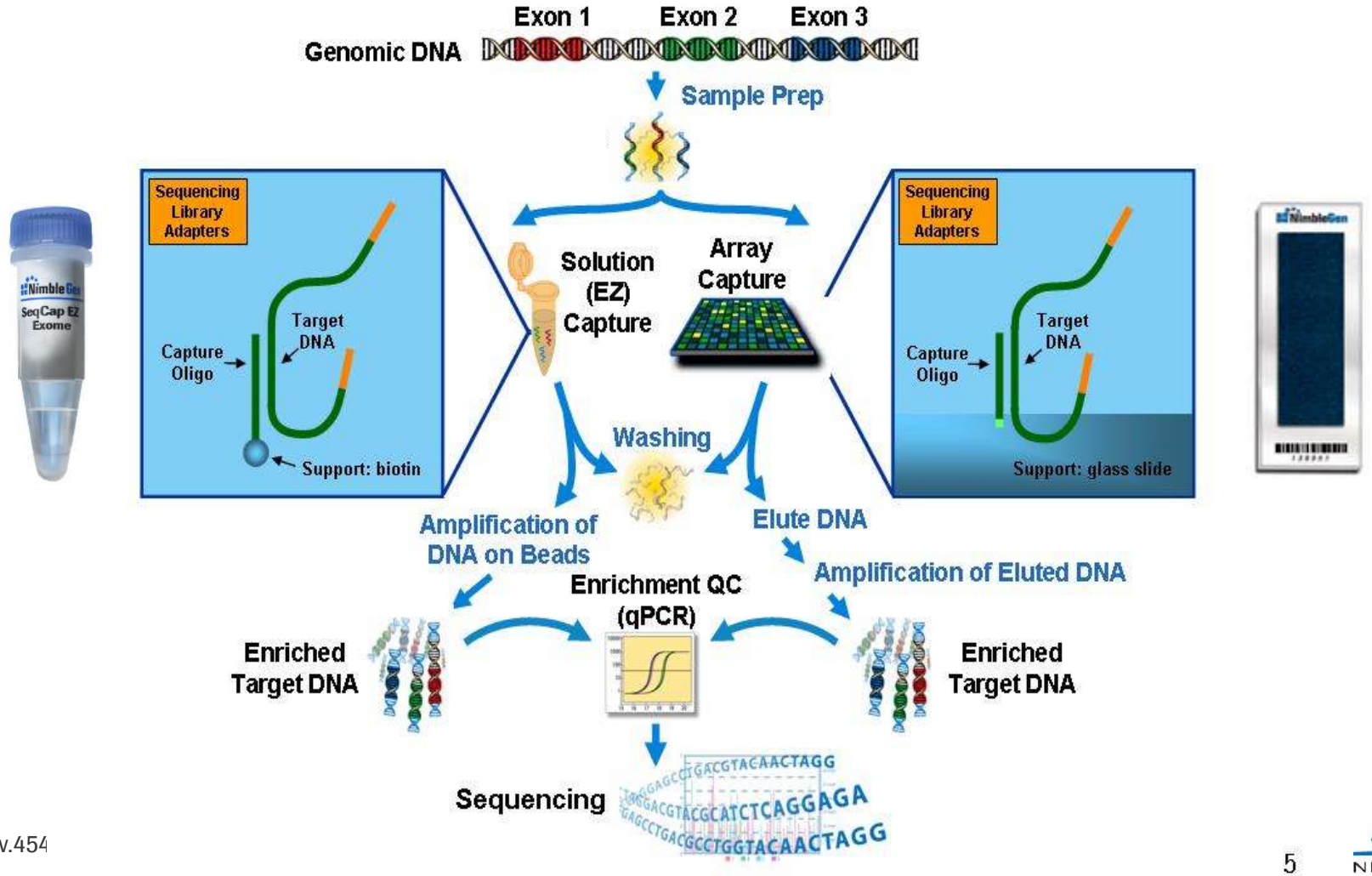
Final Library

GS FLX Titanium
emPCR amplification
&
Sequencing

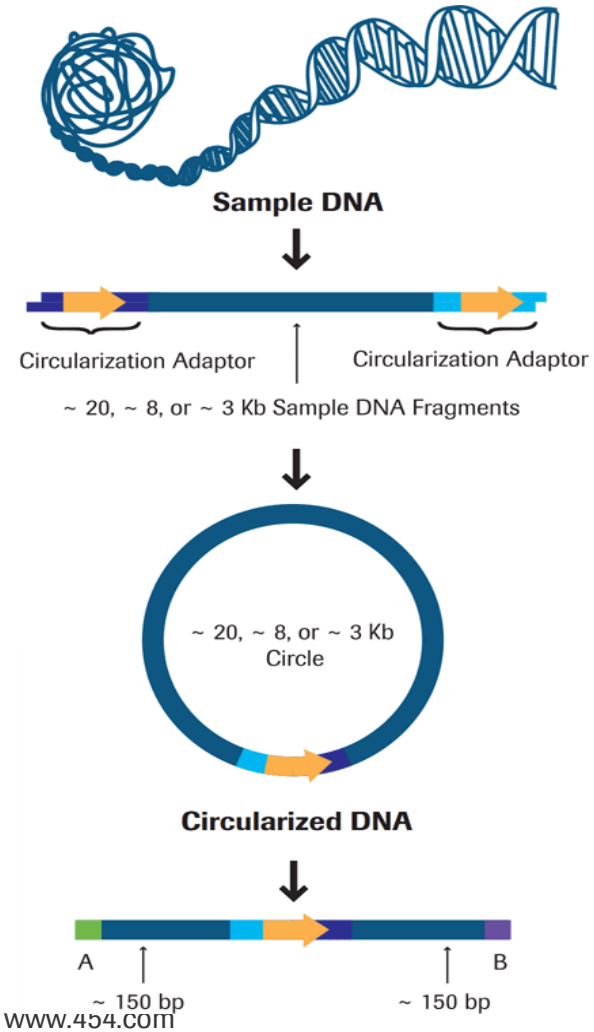
www.454.com

Library Preparation

Sequence Capture



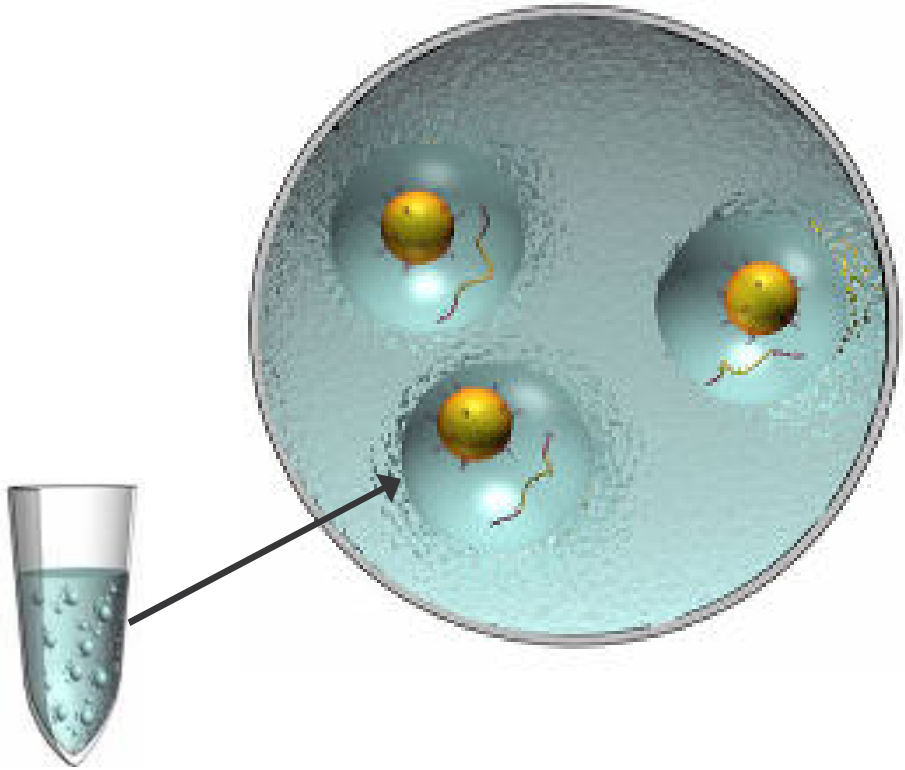
Library Preparation
Paired End Method



Paired End reads enable ordering & orientation of contigs

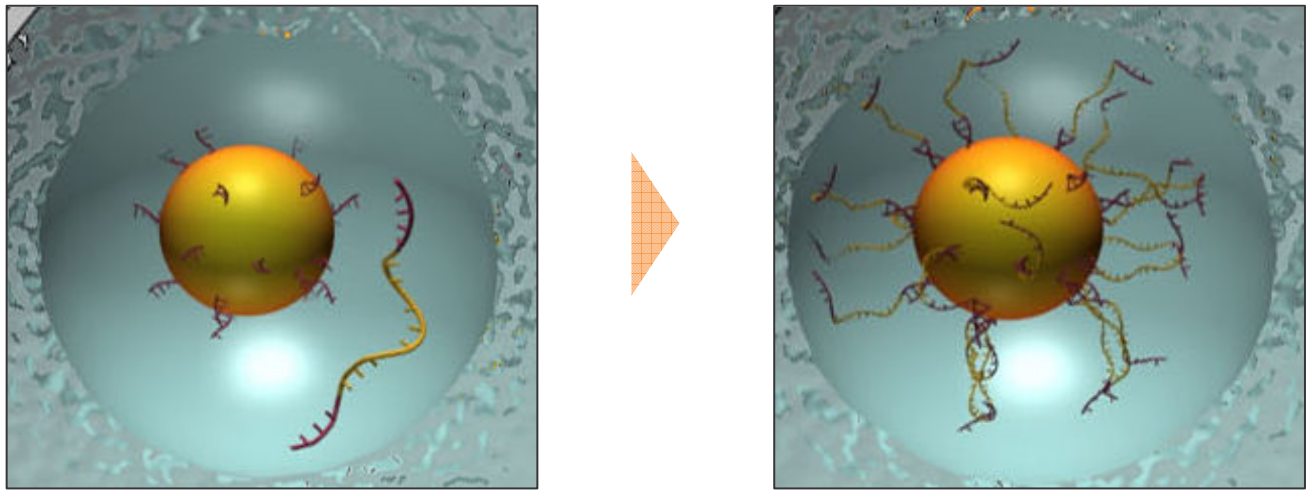
Emulsion PCR

Reagent Preparation



Emulsion oil and emPCR reagents are mixed to create water-in-oil microreactors

Emulsion PCR
Clonal Amplification



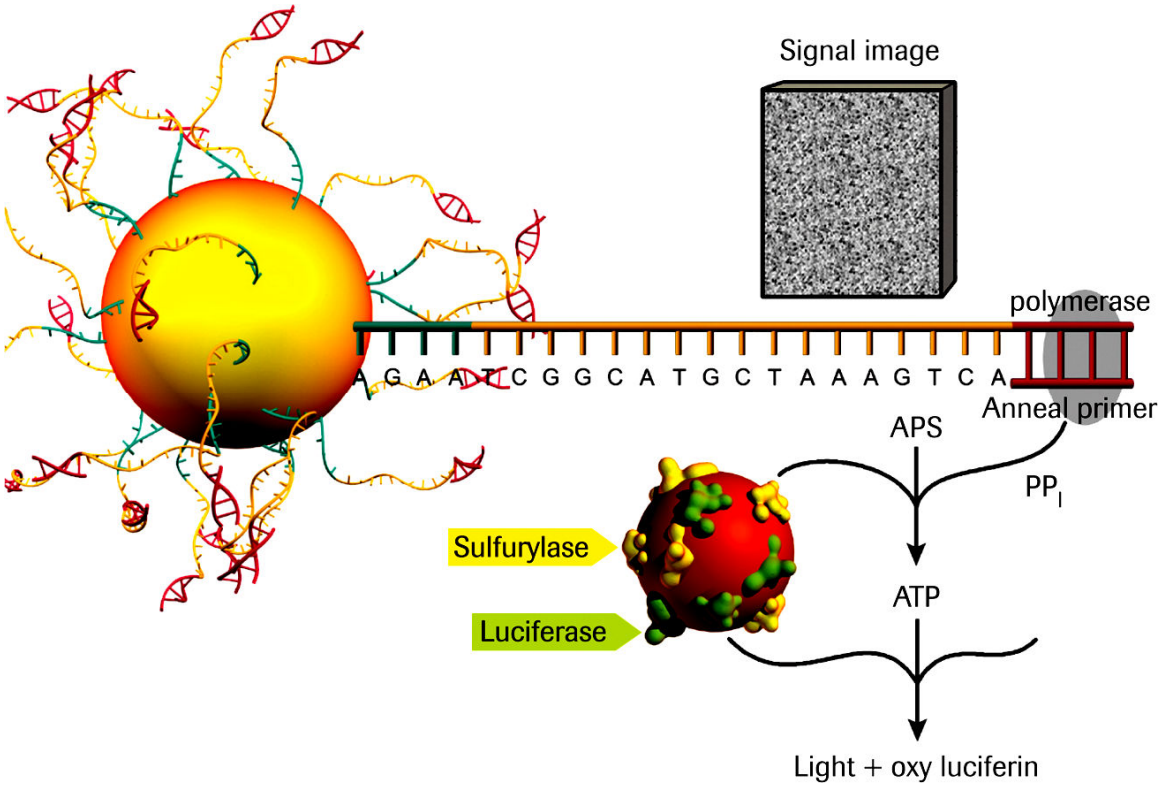
Emulsify beads & reagents to form water-in-oil microreactors

Clonal amplification of millions of beads in parallel

After PCR, each bead contains millions of copies of a single library fragment

454 Sequencing Systems

Key Concept



Polymerase adds nucleotide

Pyrophosphate is released (PPi)

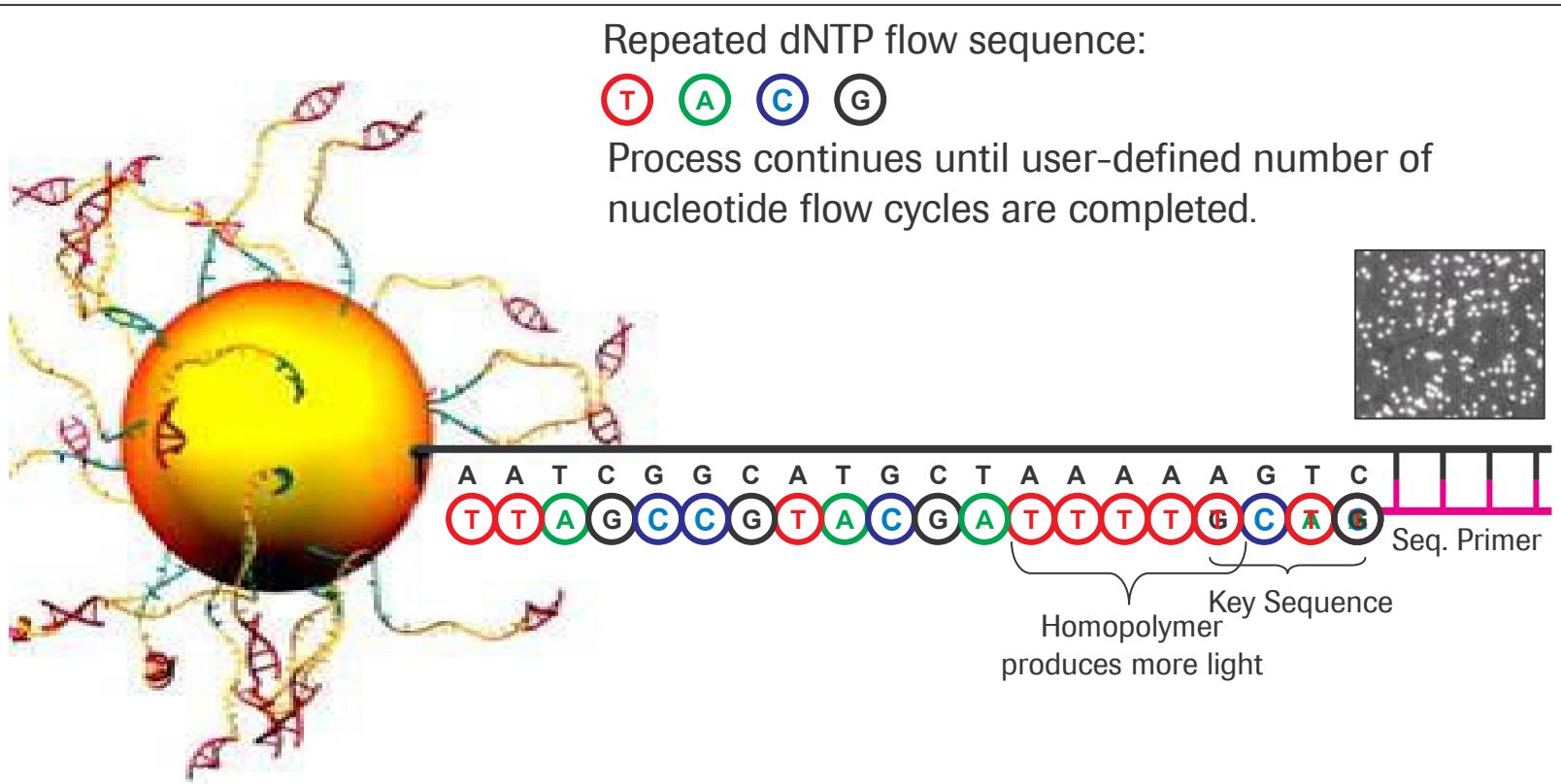
Sulfurylase converts PPi to ATP (in presence of APS)

Luciferase hydrolyses ATP and uses luciferin to make light

Simultaneous sequencing of clonally amplified beads in picoliter-size wells

454 Sequencing Systems

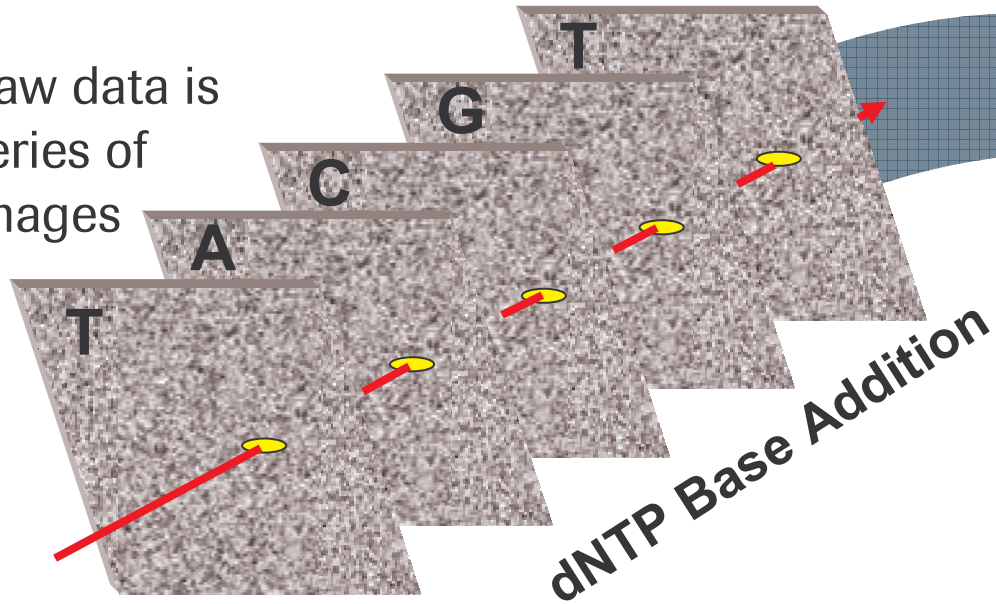
Sequencing-by-synthesis



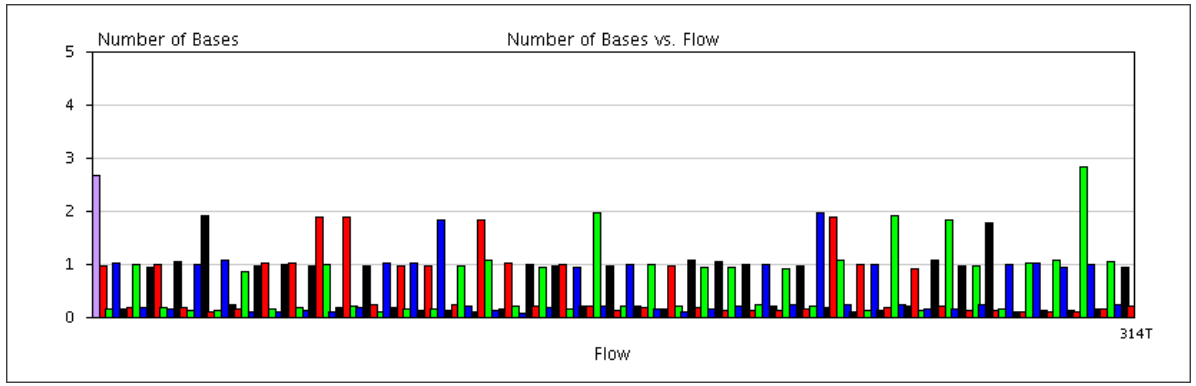
454 Sequencing Systems

Image Processing Overview

1. Raw data is series of images



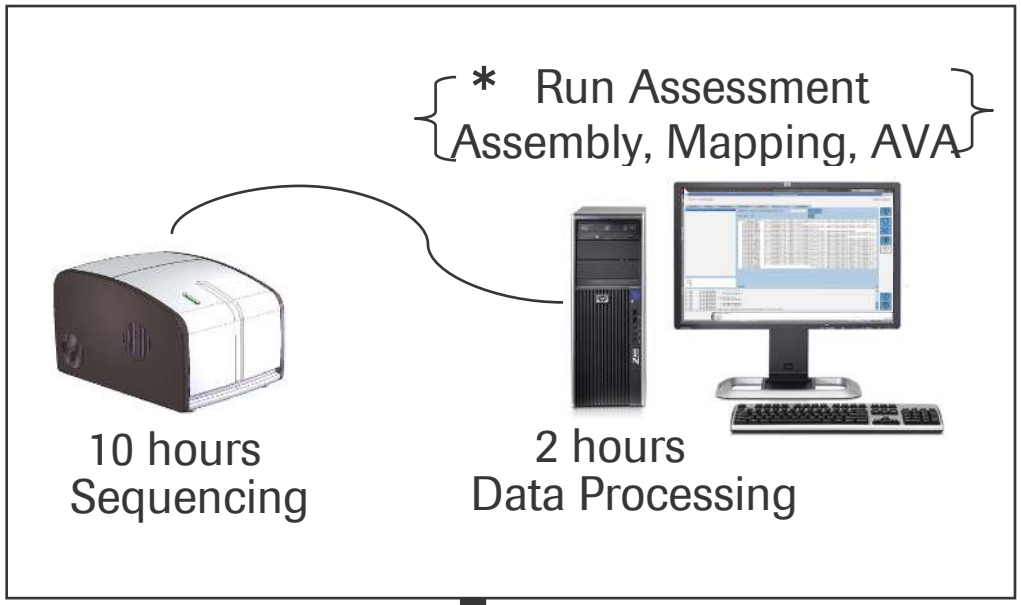
2. Each well's data extracted, quantified and normalized



3. Read data converted into "flowgrams"

Bioinformatic Workflow Options

GS Junior



* **Optional:**
Separate Datarig(s)

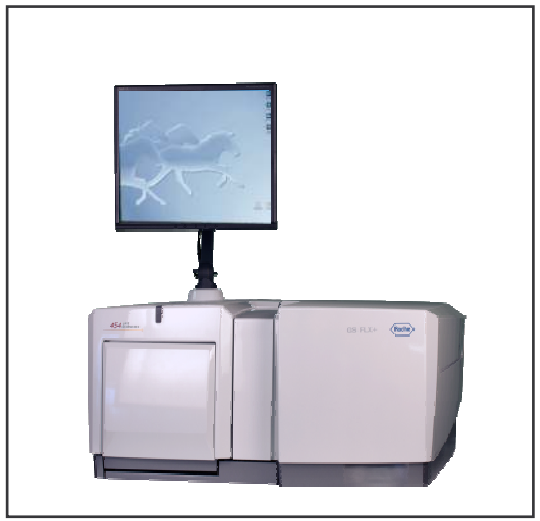
Run Assessment
Assembly, Mapping, AVA

Local Customer Network

Data Backup
Archival Storage System
Images, Analyzed data, Other data

Bioinformatic Workflow Options

GS FLX+ and GS FLX



10 hours Sequencing
and Image Processing

Signal Processing and
Assembly, Mapping, AVA

Local Customer Network

Data Backup
Archival Storage System
Images, Analyzed data, Other data

Run Metrics

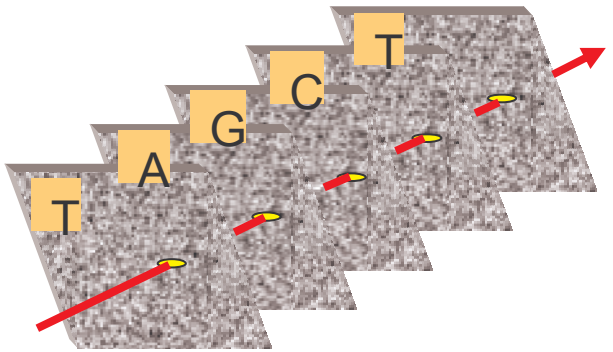
GS FLX+/GS FLX/GS Junior

	XL+	XLR70	GS Junior
Average Read Length	>420bp	>300bp	>300bp
Total Passed Filter Bases	>500 million	>360 million	19-27 million
Total Reads or totalPassFiltering	~1 Million	~1 Million	50,000-88,000
Total Run Size - including raw images	60 Gb	37 Gb	4 Gb
Raw Images Only	35 Gb	28 Gb	2.5 Gb

Data Processing
GS RunProcessor

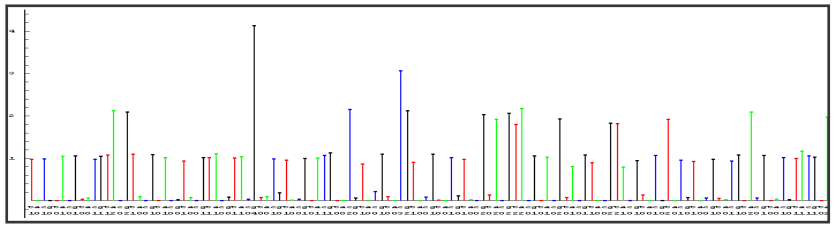
Image Processing

- image normalization, background subtraction
- locating „active“ wells on the PTP device



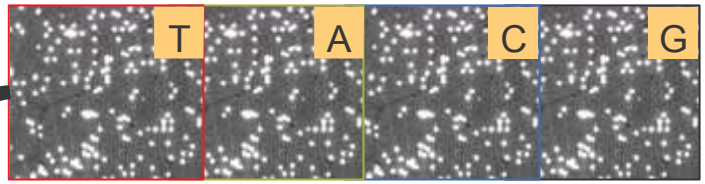
Signal Processing

- correction steps, e.g. crosstalk between neighboring wells
- quality filtering
- read trimming for good quality
- flowgram generation
- basecalling



Quality Filtering Fundamentals

GS RunProcessor



Read Rejecting Filters

Read Trimming Filters

Step	Filter Name
1	Keypass
2	Dot
3	Mixed
4	Signal Intensity
5	Primer
6	Valley or TrimBack

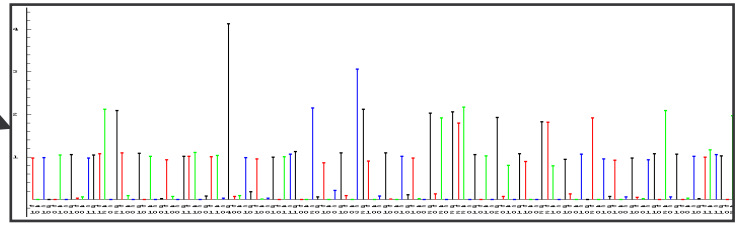


Image and Signal Processing Summary

GS RunProcessor

GS Junior Output

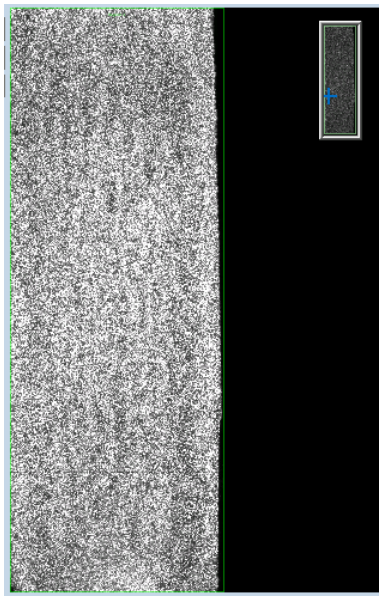


Image Processing Output

Image Processing

D_..imageProcesssing

- regions/1.cwf
- gsRunProcessor.log

Signal Processing Output

Signal Processing

D_..fullProcessing

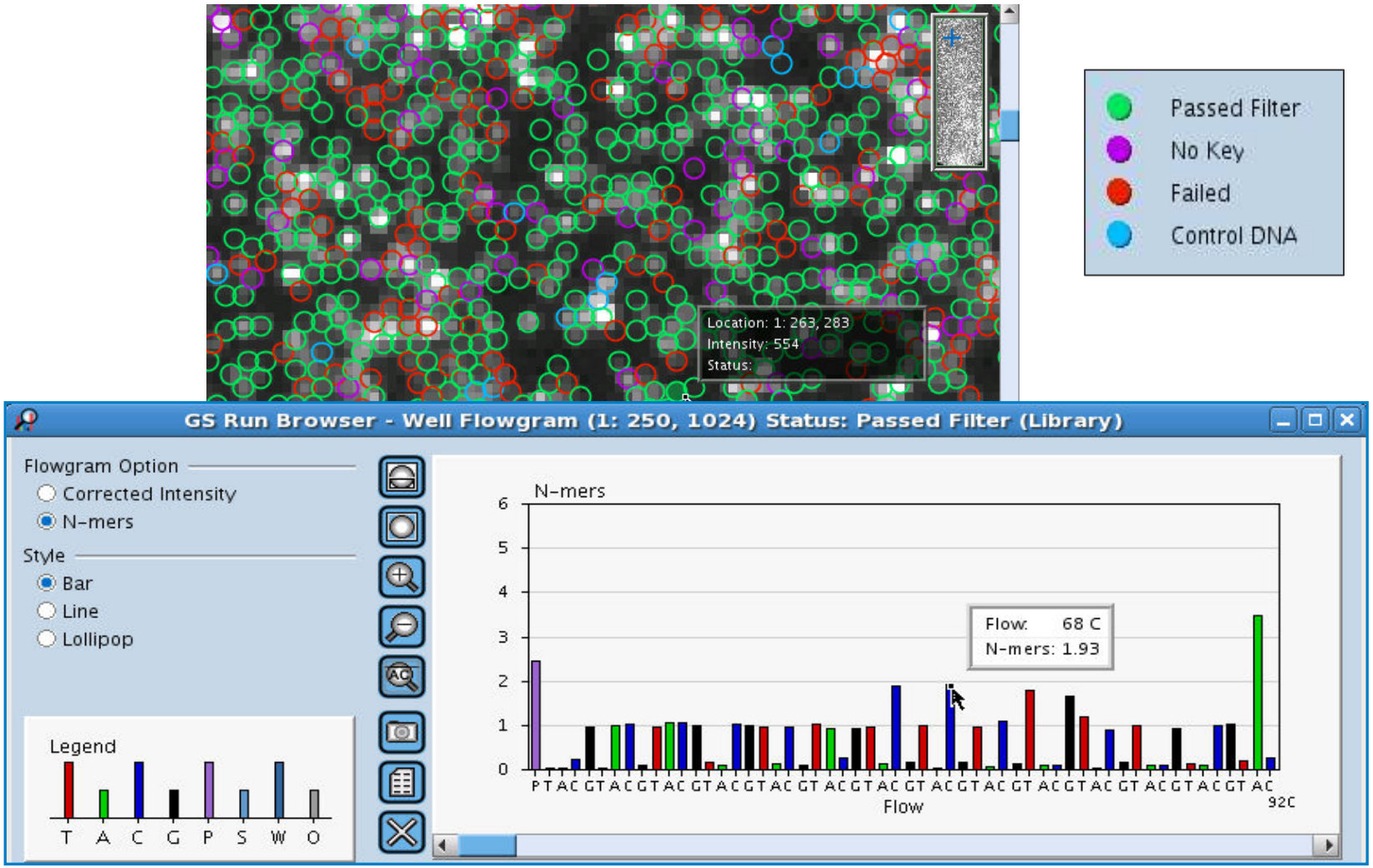
- regions/1.cwf
- sff/*.sff
- gsRunProcessor.log
- Metrics files- *.csv, *.txt,*.fna,*.qual

Data Browsing
GS Run Browser

- Browsing results of a sequencing run
- Visualization of the run results
- GUI for starting the data processing pipeline(s)

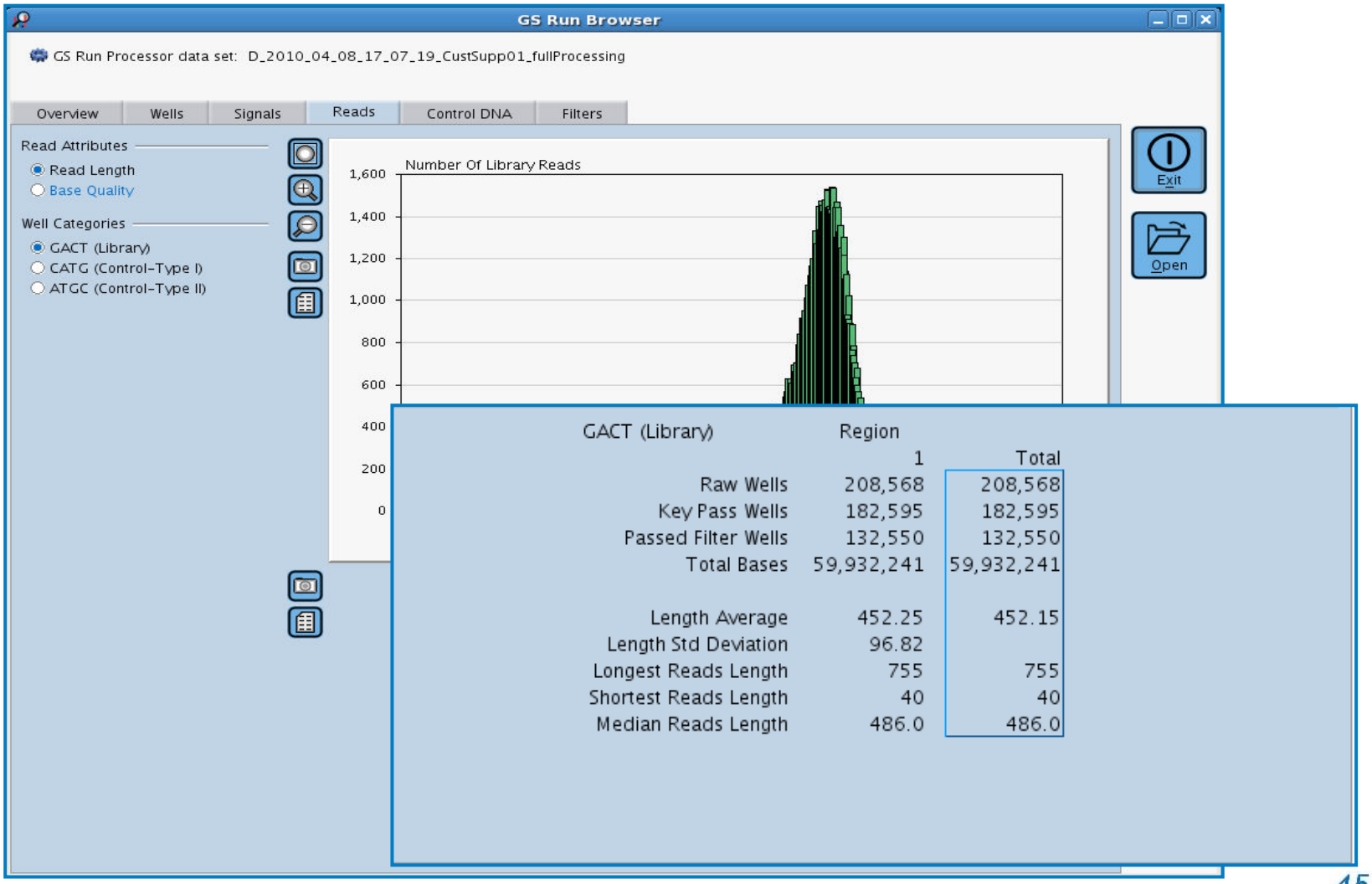
Individual Well – Flowgram

GS Run Browser



Read Length Distribution – Reads Tab

GS Run Browser



Data Analysis
Overview

GS de novo Assembler

Genome or Transcriptome de novo assembly; Large genomes; Hybrid assemblies

GS Reference Mapper

Targeted mapping using Sequence Capture; Genome or Transcriptome reference mapping; Variation detection including large rearrangements

GS Amplicon Variant Analyzer

Ultra-deep sequencing for variant analyses; Haplotyping

(SFF Tools)

SFF file manipulation

Performance on GS Junior Attended PC

GS de novo Assembler

C. jejuni de novo Assembly

- 118,000 HQ GS FLX Titanium reads containing 40 MB sequence data; 23X coverage
- 2 minutes

S. cerevisiae cDNA Assembly

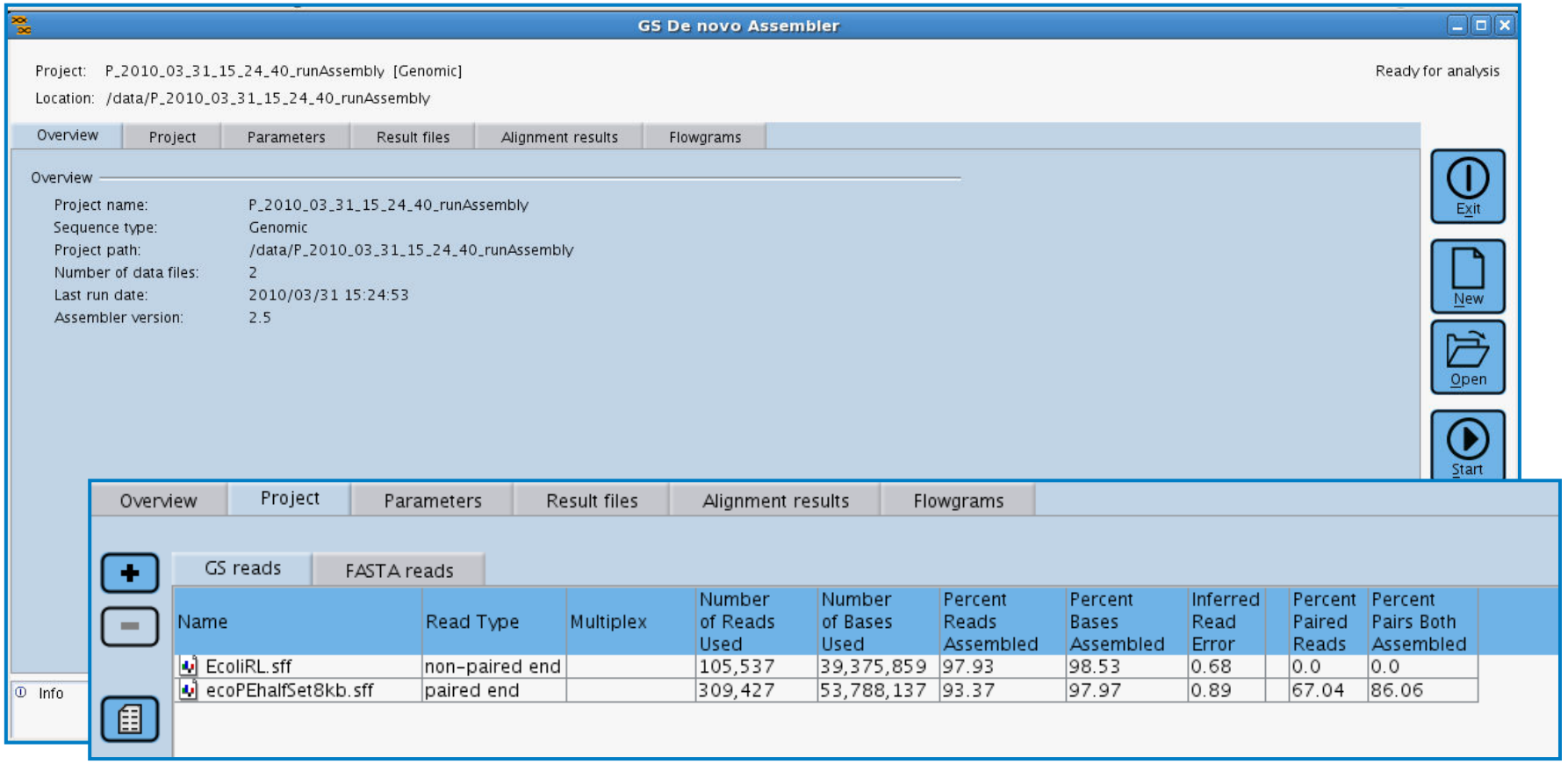
- 1,000,000 HQ GS FLX Titanium reads containing 4GB sequence data; 25X coverage
- 2 hours

D. melanogaster de novo Assembly

- 2,000,000 HQ GS FLX Titanium reads containing 9GB sequence data; 22X coverage
- 8 hours

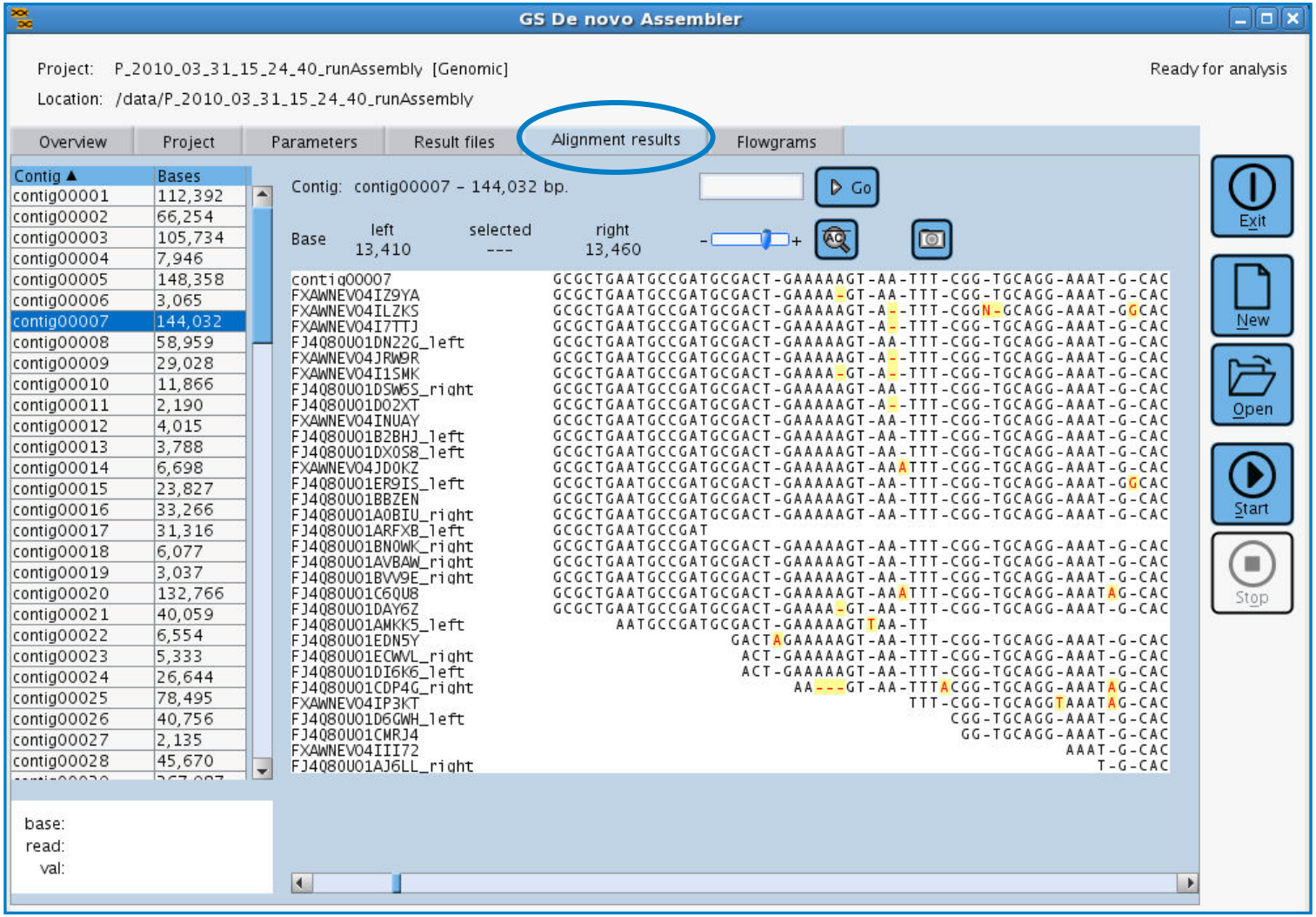
GUI Assembly – *Overview*

GS de novo Assembler



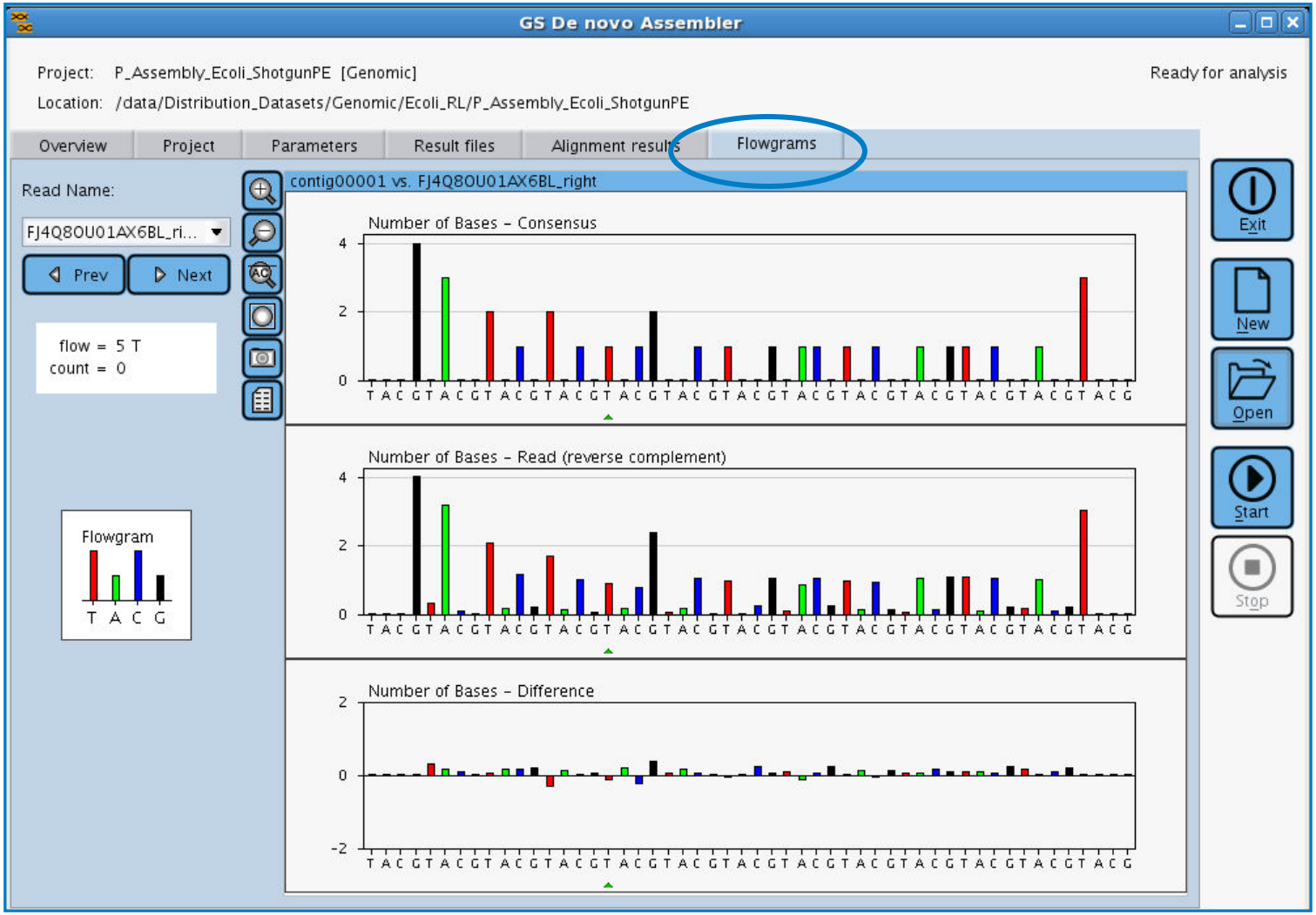
Alignment View

GS de novo Assembler



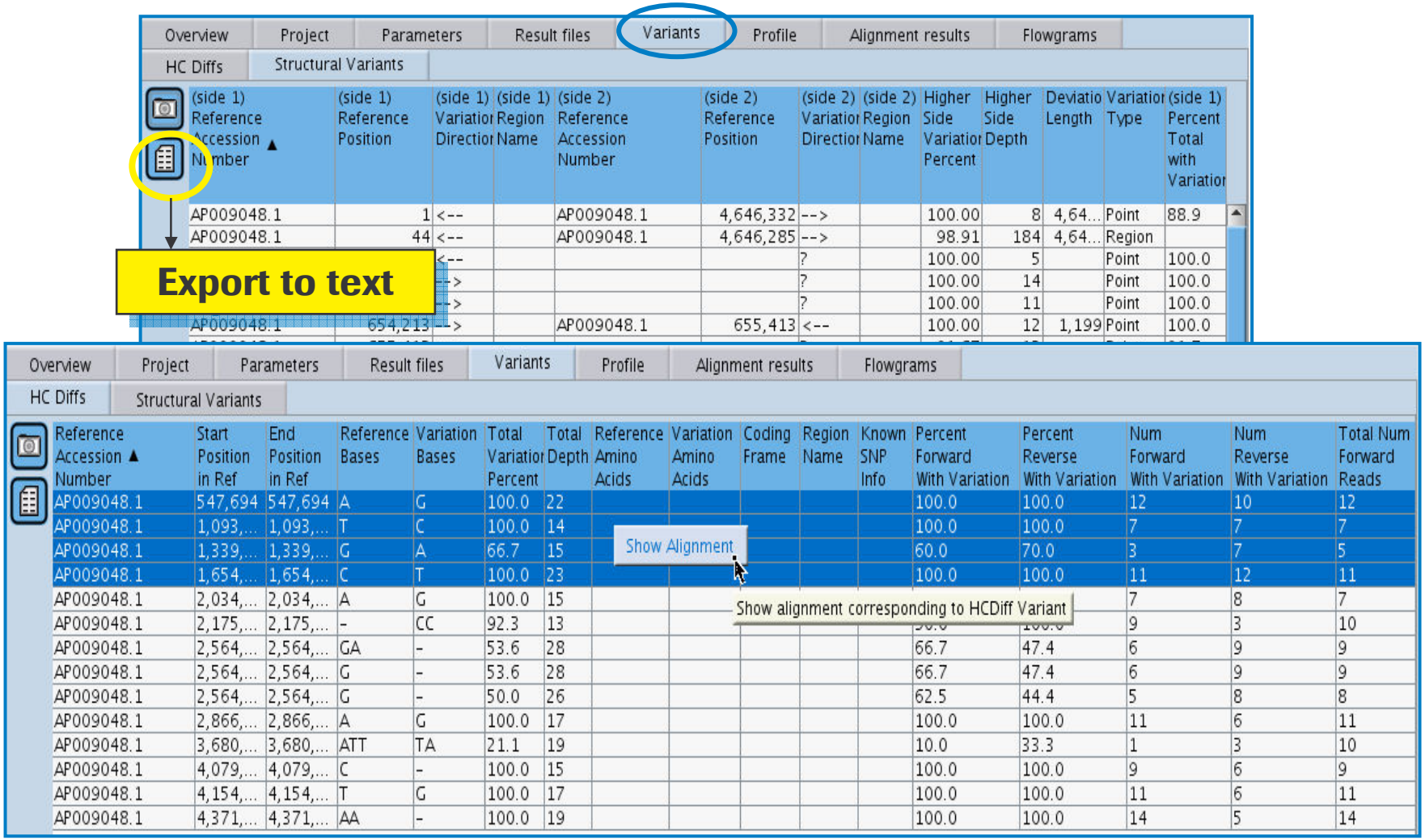
Flowgram View

GS de novo Assembler



Variations and Structural Rearrangements

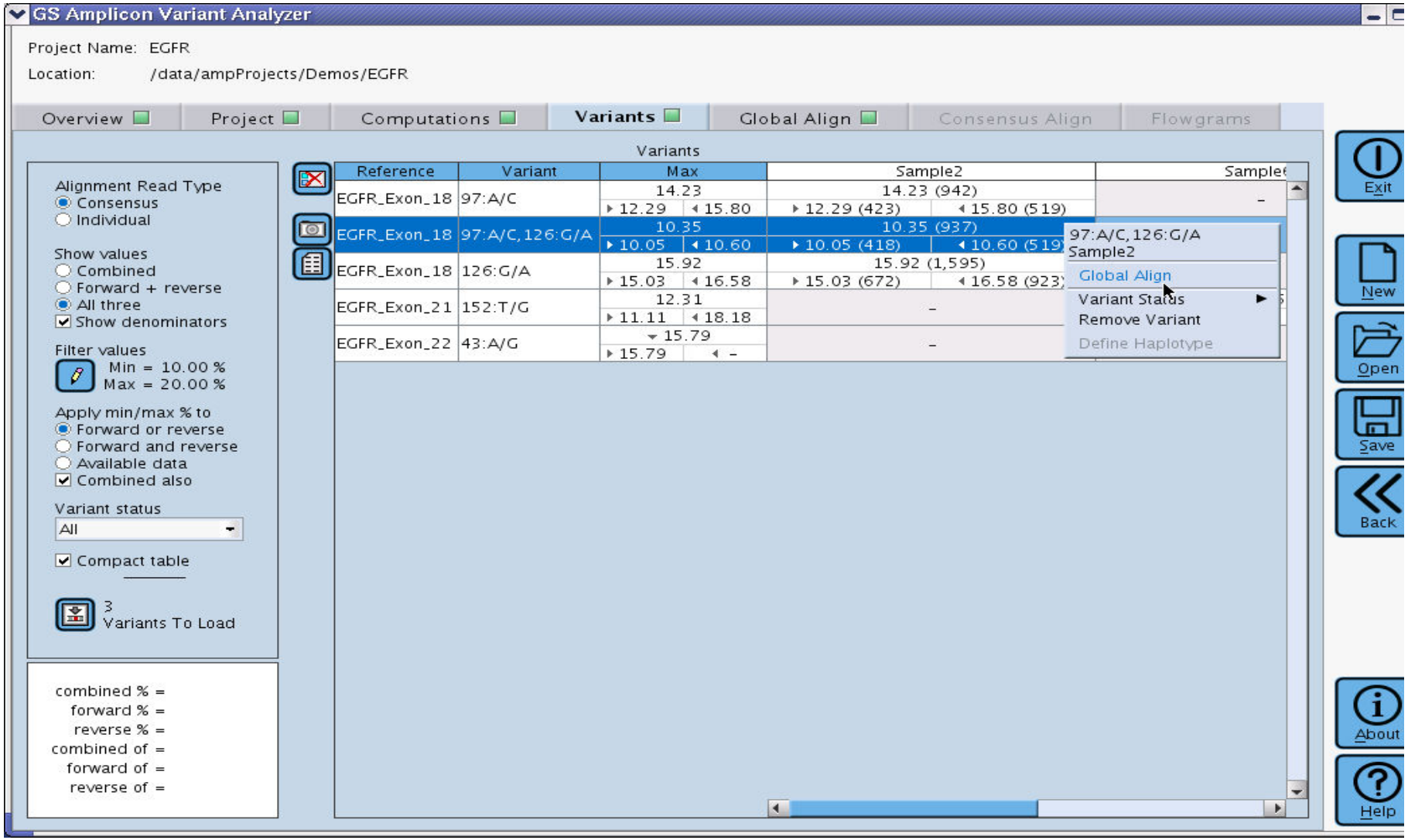
GS Reference Mapper



Reference Accession Number	Start Position in Ref	End Position in Ref	Reference Bases	Variation Bases	Total Variation Percent	Total Depth	Reference Amino Acids	Variation Amino Acids	Coding Frame	Region Name	Known SNP Info	Percent Forward With Variation	Percent Reverse With Variation	Num Forward With Variation	Num Reverse With Variation	Total Num Forward Reads
AP009048.1	547,694	547,694	A	G	100.0	22						100.0	100.0	12	10	12
AP009048.1	1,093,...	1,093,...	T	C	100.0	14						100.0	100.0	7	7	7
AP009048.1	1,339,...	1,339,...	G	A	66.7	15						60.0	70.0	3	7	5
AP009048.1	1,654,...	1,654,...	C	T	100.0	23						100.0	100.0	11	12	11
AP009048.1	2,034,...	2,034,...	A	G	100.0	15								7	8	7
AP009048.1	2,175,...	2,175,...	-	CC	92.3	13								9	3	10
AP009048.1	2,564,...	2,564,...	GA	-	53.6	28						66.7	47.4	6	9	9
AP009048.1	2,564,...	2,564,...	G	-	53.6	28						66.7	47.4	6	9	9
AP009048.1	2,564,...	2,564,...	G	-	50.0	26						62.5	44.4	5	8	8
AP009048.1	2,866,...	2,866,...	A	G	100.0	17						100.0	100.0	11	6	11
AP009048.1	3,680,...	3,680,...	ATT	TA	21.1	19						10.0	33.3	1	3	10
AP009048.1	4,079,...	4,079,...	C	-	100.0	15						100.0	100.0	9	6	9
AP009048.1	4,154,...	4,154,...	T	G	100.0	17						100.0	100.0	11	6	11
AP009048.1	4,371,...	4,371,...	AA	-	100.0	19						100.0	100.0	14	5	14

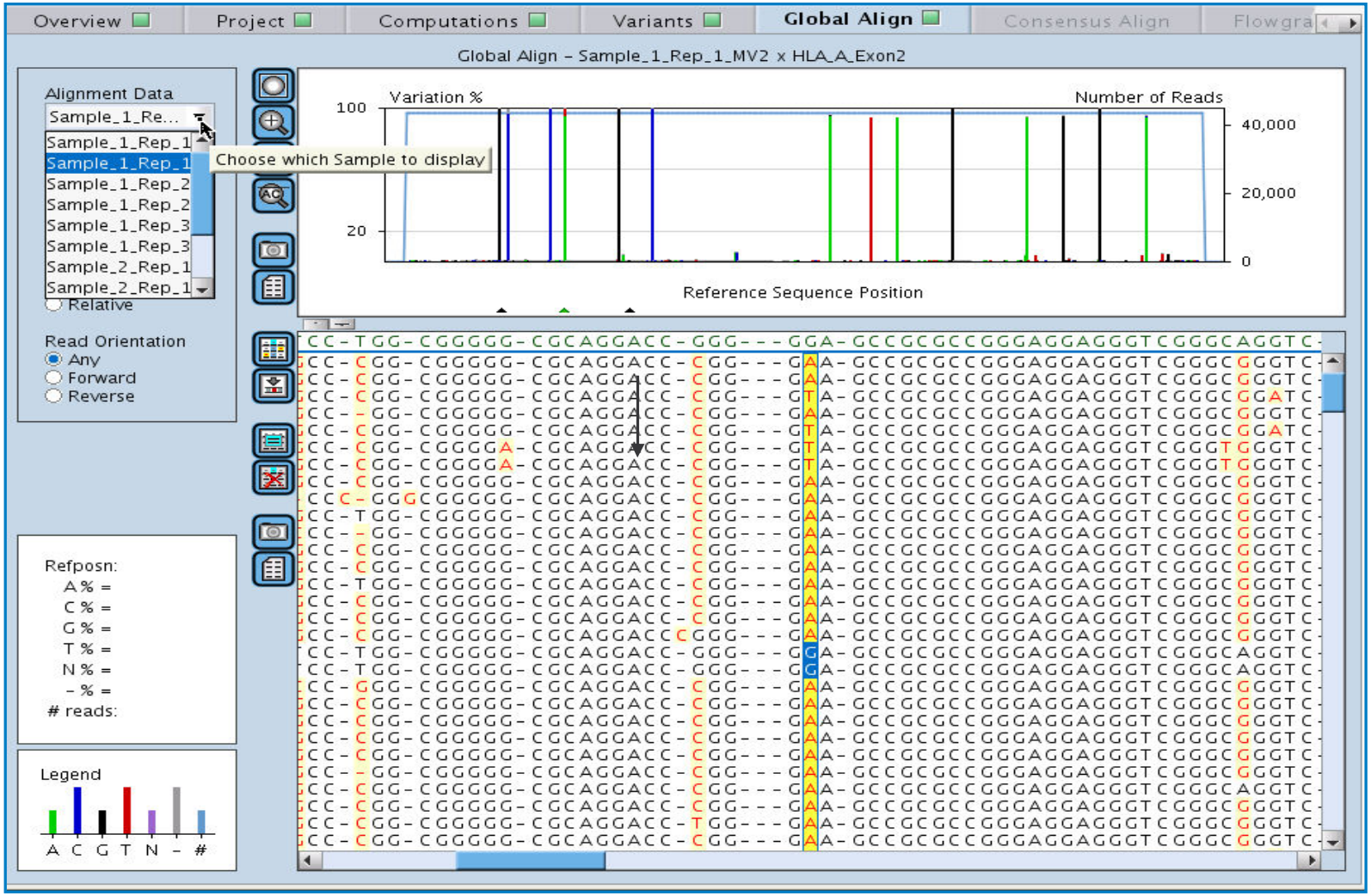
Variants Tab

GS Amplicon Variant Analyzer



Global Align Tab

GS Amplicon Variant Analyzer



Data Analysis
Overview

GS de novo Assembler

Genome or Transcriptome de novo assembly; Large genomes; Hybrid assemblies

GS Reference Mapper

Targeted mapping using Sequence Capture; Genome or Transcriptome reference mapping; Variation detection including large rearrangements

GS Amplicon Variant Analyzer

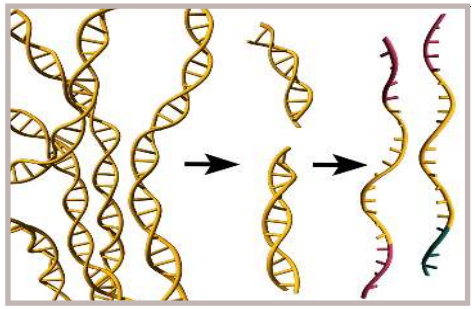
Ultra-deep sequencing for variant analyses; Haplotyping

(SFF Tools)

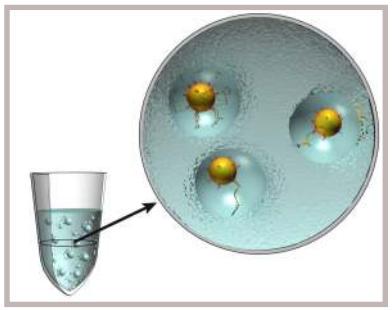
SFF file manipulation

454 Sequencing Systems

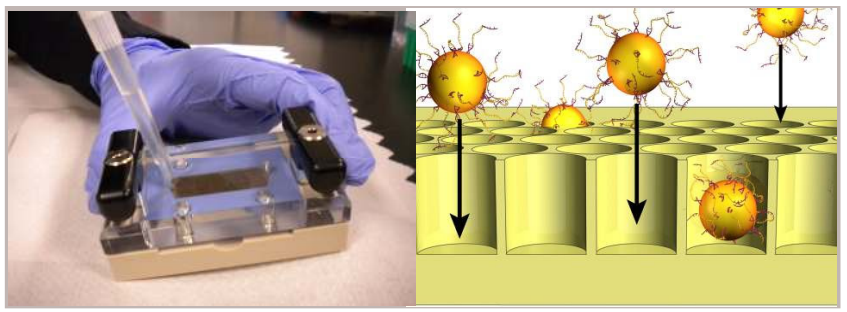
One Fragment, One Bead, One Read



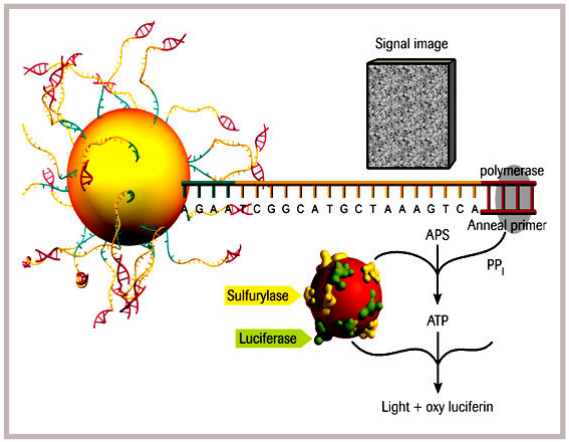
Shear DNA and add linkers



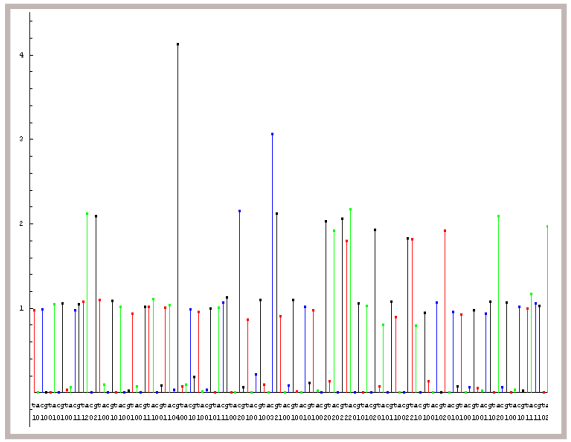
'Emulsion PCR' Clonal amplification



Deposition of beads into wells of PicoTiterPlate device



Sequencing-by-synthesis Detection of PPi release



Flowgram displays the sequence of reads based on light intensity

www.454.com

454 Sequencing Systems
First GS Junior Publication

Immunogenetics
DOI 10.1007/s00251-010-0481-9

BRIEF COMMUNICATION

Characterization of Mauritian cynomolgus macaque major histocompatibility complex class I haplotypes by high-resolution pyrosequencing

Melisa L. Budde · Roger W. Wiseman · Julie A. Karl · Bozena Hanczaruk · Birgitte B. Simen · David H. O'Connor

Received: 27 July 2010 / Accepted: 13 September 2010

Abstract Major histocompatibility complex (MHC) class I alleles of nonhuman primates have been associated with disease susceptibility, resistance, and resolution. Here, using high-resolution pyrosequencing, we characterized MHC class I transcripts expressed in Mauritian cynomolgus

Major histocompatibility complex (MHC) class I proteins are present on all nucleated cells and present peptides to circulating CD8+ T cells (Parham 1994). In humans, MHC class I loci are well defined with only three classical polymorphic genes per haplotype (HLA-A, HLA-B, and

454 Sequencing Systems
First GS Junior Publication

- Sequencing of MHC class I transcripts in macaques to discover all expressed transcripts from common class I haplotypes
- Sequenced 3 amplicons from ~440 to 620 bases
- Combination experiment
  - 7 individuals on GS FLX System, 3 using GS Junior System
  - Identified all sequences found previously
  - Discovered 2x more haplotypes than with previous Sanger-based approach
- 440-600 base amplicons allow resolution of haplotypes that are impossible with 190 base amplicons

CFTR Exon Resequencing

Utility of MIDs with Amplicon Pooling

11 Coriell samples with known CF variants

34 amplicons cover 27 exons total: 11 x 34 = 374 amplicons

Each sample was MID-labeled and pooled after multiplex PCR

Single GS Junior run with average coverage 182x

96% of the High Quality Reads mapped back to the CF gene region

---

AVA output – showing 5 of 11 samples vs. variants discovered

Reference	Variant	NA08338...	NA11277...	NA11281...	NA11283...	NA21551...
exon11_ref_0012	122:G/A	100.00	100.00	98.70	100.00	100.00
exon11_ref_0012	234-236:TCT/---	0.00	0.00	44.81	47.59	51.53
exon15_ref_0019	78:T/G	100.00	99.60	0.00	0.00	47.73
exon27_ref_0034	233:G/A	100.00	48.03	0.00	0.00	54.36

454 Sequencing Systems

Rare Variant Detection

AAC Accepts, published online ahead of print on 4 April 2011
Antimicrob. Agents Chemother. doi:10.1128/AAC.00091-11

1 **Concordance Between Two Phenotypic Assays and Ultra Deep**

2 **Pyrosequencing for Determining HIV-1 Tropism**

3

4 Adrien Saliou,[1] Pierre Delobel,[1,2] Martine Dubois,[1,3] Florence Nicot,[1,3] Stéphanie Raymond,[1,3]

5 Vincent Calvez,[4] Bernard Masquelier,[5] Jacques Izopet[1,3*] and the ANRS AC11 Resistance

6 Study Group

HIV “quasispecies” can use CCR5 and/or CXCR4 cell-surface receptors to enter cells

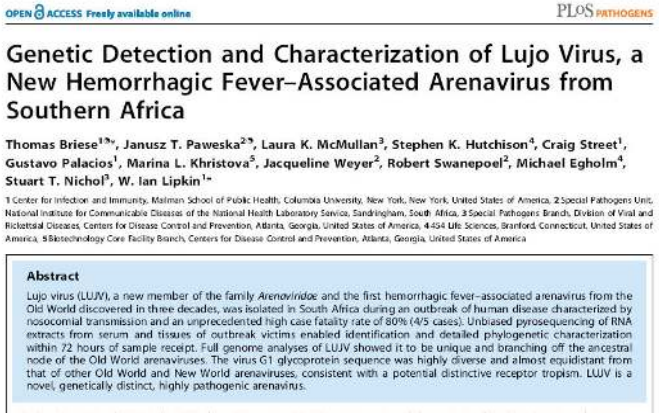
Drugs that block CCR5 receptors work only if CXCR4-binding variants are absent

454 Sequencing Systems

Rare Variant Detection

- 415 base cDNA amplicon covering V3 *env* region of HIV-1
- 23 individual samples pooled & loaded onto a single GS Junior run
- Obtained ~3,500 reads per sample
- GS AVA software used to align to reference
- Processed the reads using third party prediction software
- Detected quasispecies to 0.6% reliably
- Calculated mean error rate of .000853 from control plasmids

Pathogen Discovery
Metagenomics



OPEN ACCESS Freely available online PLoS PATHOGENS

Genetic Detection and Characterization of Lujo Virus, a New Hemorrhagic Fever–Associated Arenavirus from Southern Africa

Thomas Briese[1,9*], Janusz T. Paweska[2,9], Laura K. McMullan[3], Stephen K. Hutchison[4], Craig Street[1], Gustavo Palacios[1], Marina L. Khristova[5], Jacqueline Weyer[2], Robert Swanepoel[2], Michael Egholm[4], Stuart T. Nichol[3], W. Ian Lipkin[1*]

1 Center for Infection and Immunity, Mailman School of Public Health, Columbia University, New York, New York, United States of America, 2 Special Pathogens Unit, National Institute for Communicable Diseases of the National Health Laboratory Service, Sandringham, South Africa, 3 Special Pathogens Branch, Division of Viral and Rickettsial Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America, 4 454 Life Sciences, Branford, Connecticut, United States of America, 5 Biotechnology Core Facility Branch, Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America

Abstract

Lujo virus (LUJV), a new member of the family *Arenoviridae* and the first hemorrhagic fever–associated arenavirus from the Old World discovered in three decades, was isolated in South Africa during an outbreak of human disease characterized by nosocomial transmission and an unprecedented high case fatality rate of 80% (4/5 cases). Unbiased pyrosequencing of RNA extracts from serum and tissues of outbreak victims enabled identification and detailed phylogenetic characterization within 72 hours of sample receipt. Full genome analyses of LUJV showed it to be unique and branching off the ancestral node of the Old World arenaviruses. The virus G1 glycoprotein sequence was highly diverse and almost equidistant from that of other Old World and New World arenaviruses, consistent with a potential distinctive receptor tropism. LUJV is a novel, genetically distinct, highly pathogenic arenavirus.

Case from Sandton, South Africa

Infected paramedic during transfer, nurse at hospital, cleaning staff, and nurse of paramedic- 4/5 did not survive

Serum and tissue samples from victims were subjected to unbiased pyrosequencing, yielding within 72 hours of sample receipt, multiple discrete sequence fragments that represented approximately 50% of a prototypic arenavirus genome.

Recapitulated GS FLX System study in single GS Junior System run

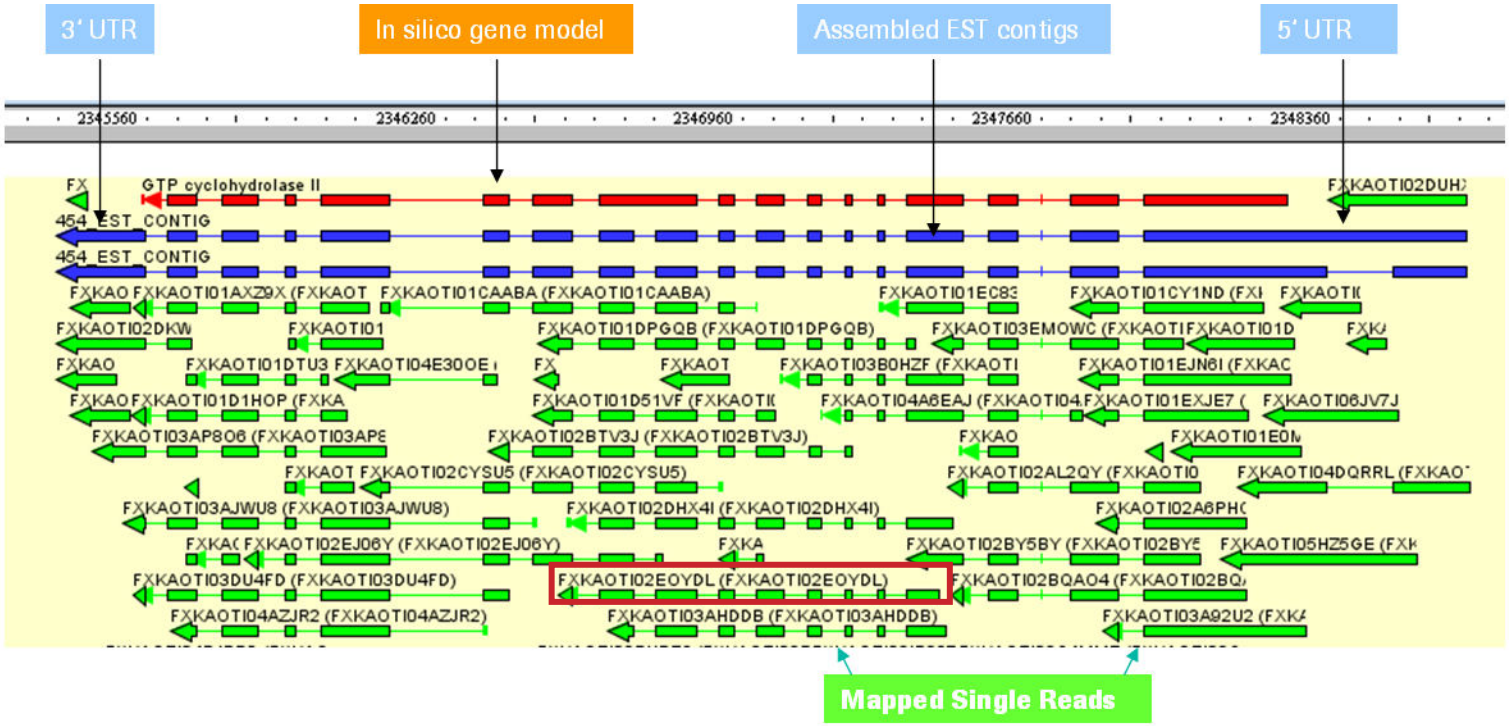
250 Hits to LuJo Virus covering 57% of the L-segment and 79% of the S-segment

www.454.com

454 SEQUENCING

Transcriptome Sequencing

One-step Discovery and Annotation



~75% of the EST contigs are full-length transcripts

Long 454 Sequencing single reads cover up to 9 exons

EST coverage is even across the transcript length

Long reads enable re-construction of full-length alternatively spliced variants

Bacterial *de novo* Assembly
8 Kb Paired End

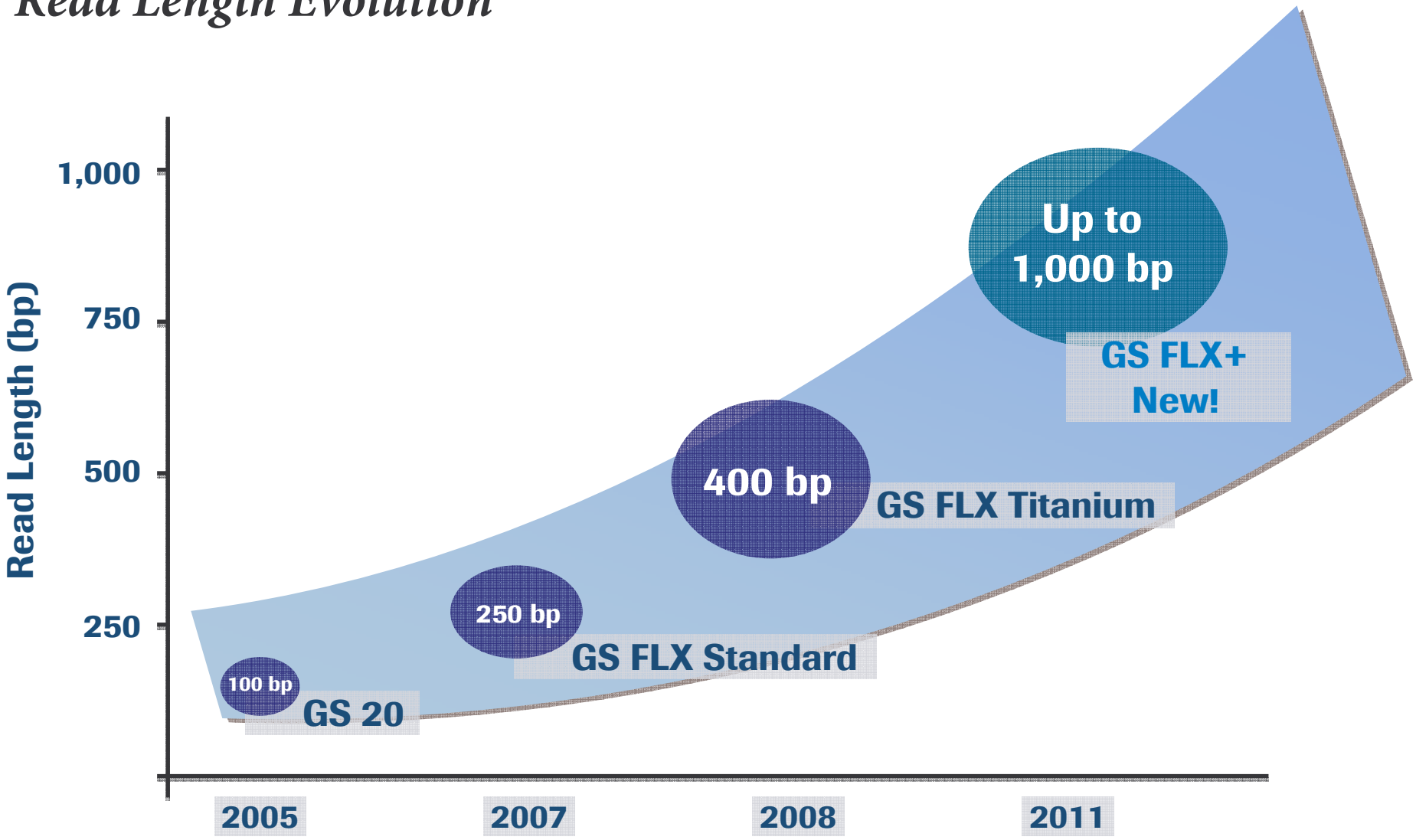
Microbial Genome Assembly				
Organism	***S. pneumoniae***	***E. coli***	***T. thermophilus***	***C. jejuni***
Number of Chromosomes	1	1	2	1
Large Scaffolds	1	1	2	1
Genome Size	2.2 Mb	4.6 Mb	2.1 Mb	1.6 Mb
N50 Scaffold Size	2.2 Mb	4.6 Mb	1.9 Mb	1.6 Mb
N50 Contig Size	26.1 kb	57.1 kb	10.5 kb	153.8 kb
Genome Coverage	99.6%	100%	100%	99.3%
Oversampling	25x	15x	33x	33x
Number of GS FLX Runs	1/4	1/4	1/4	1/4
Number of GS Junior Runs	2	2	2	2

GS FLX+ System

Delivering Sequencing Reads up to 1,000 bp

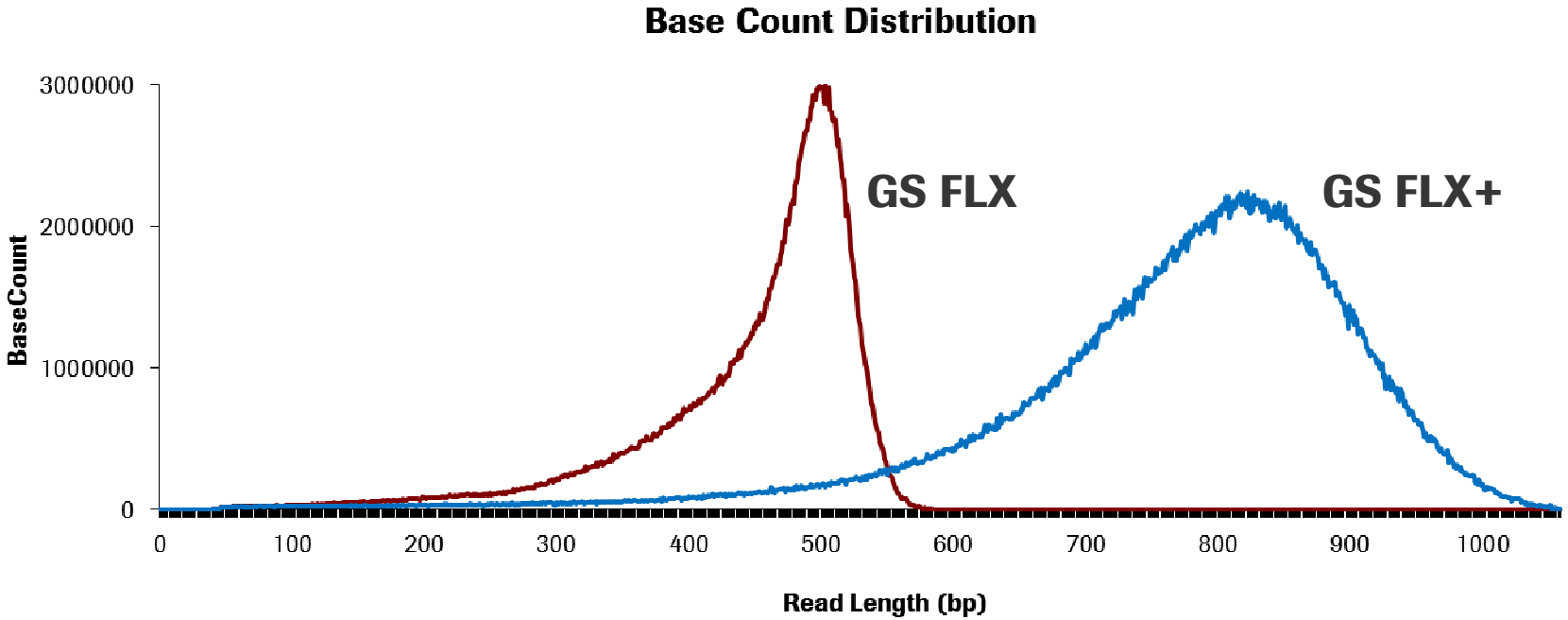
454 Sequencing Systems

Read Length Evolution



www.454.com

GS FLX+ Improvements
Read Length Comparison



New GS FLX+ generates significantly more bases > 500 bp

454 Sequencing Systems

XL+ Chemistry Comparison

	GS FLX+ System	
Sequencing Kit	**GS FLX Titanium XL+**	**GS FLX Titanium XLR70**
Read Lengths	Up to 1,000 bp	Up to 600 bp
Modal Read Length	700 bp	450 bp
Throughput Profile	- 85% of total bases from reads >500 bp	- 85% of total bases from reads > 300 bp
	- 45% of total bases from reads >700 bp	- 20% of total bases from reads > 500 bp
Typical Throughput	700 Mb	450 Mb
Reads per Run	~1,000,000 shotgun	~1,000,000 shotgun
Consensus Accuracy*	99.997%	99.995%
Run Time	23 hours	10 hours

Table 1: GS FLX+ System performance using either the GS FLX Titanium Sequencing Kit XL+ or XLR70. Actual results depend on specific sample and genomic characteristics. * Consensus accuracy at 15x coverage in *E. coli*.

GS FLX+ Applications
De novo Assembly - Budgie Bird

- **Genome size** = 1.23 Gb
- **Question**: Do GS FLX+ extra-long sequencing reads improve assembly quality?
- **Experimental Design**
  - Shotgun sequence depth: 3.5X coverage each using XL+ and XLR70
  - Paired end sequence depth: 2.5X coverage – 3, 8 and 20 kb span using XLR70
- **Analysis**
  - Assembled data separately, compared results

Budgerigar Parrot

Data courtesy of the Erich Jarvis Lab, Duke University Medical Center

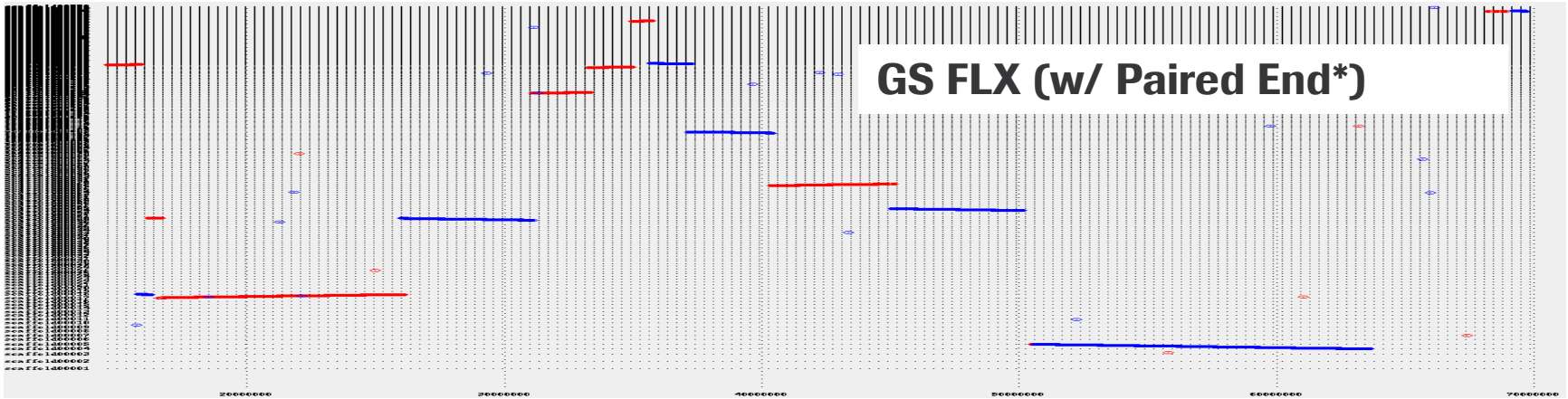
GS FLX+ Applications

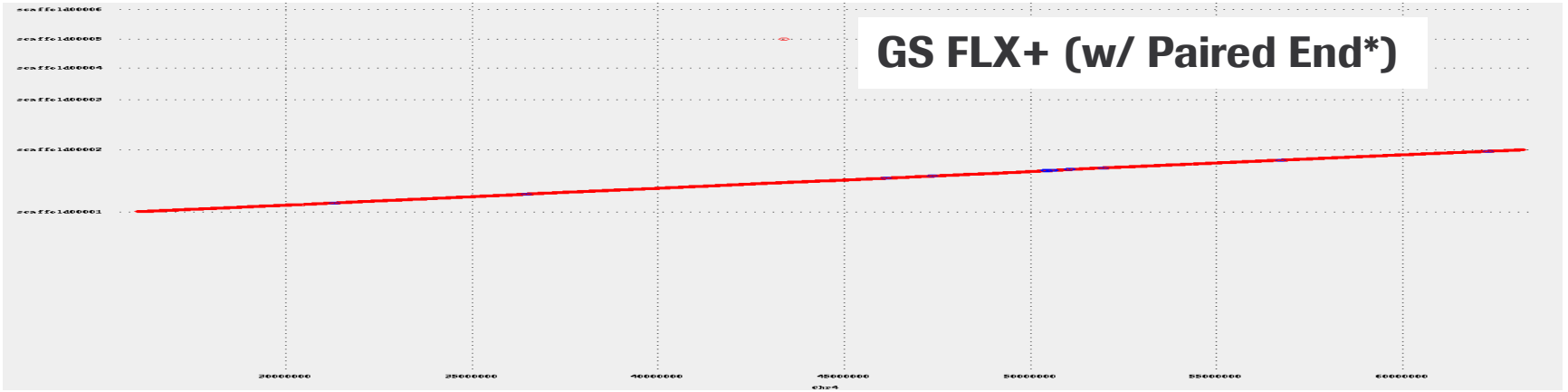
De novo Assembly - Budgie Bird

Assembly Metrics	GS FLX System (w/ 3, 8, 20 kb PE)	GS FLX+ System (w/ 3, 8, 20 kb PE)	Improvement
Sequence Depth	6X	6X	-
Number of Bases	8,018,686,780	8,019,891,335	-
Scaffold Metrics			
Number of Scaffolds	54 K	53 K	**-2% (fewer is better)**
Avg. Scaffold Size	22.5 Kb	23.1 Kb	**+3%**
N50 Scaffold Size	1.9 Mb	2.5 Mb	**+27%**
Largest Scaffold	14.0 Mb	15.6 Mb	**+11%**
Contig Metrics			
Number Of Contigs	418 K	302 K	**-28% (fewer is better)**
Avg. Contig Size	2.3 Kb	3.3 Kb	**+42 %**
N50 Contig Size	3.2 Kb	5.2 Kb	**+60 %**
Largest Contig	39 Kb	57 Kb	**+47 %**

Synteny Between Budgie and Reference

A Snapshot of Chr 4





Reference: Zebra Finch Chr 4 [25 MB-65 MB]

*Note: Same paired end reads (3, 8, 20 kb) used in both assemblies.

GS FLX+ Applications
Transcriptome Sequencing

- **Sample**: *Drosophila* larva; gene count 16,600; 21,784 transcripts
- **Question**: Do longer read lengths improve transcriptome assembly?
- **Experimental Design**:
  - Random primed *Drosophila* larva cDNA libraries (from mRNA)
  - GS FLX cDNA Rapid library protocol
  - Sequenced using XL+ and XLR70
- **Analysis**: Compared mapping and assembly metrics

Drosophila larva

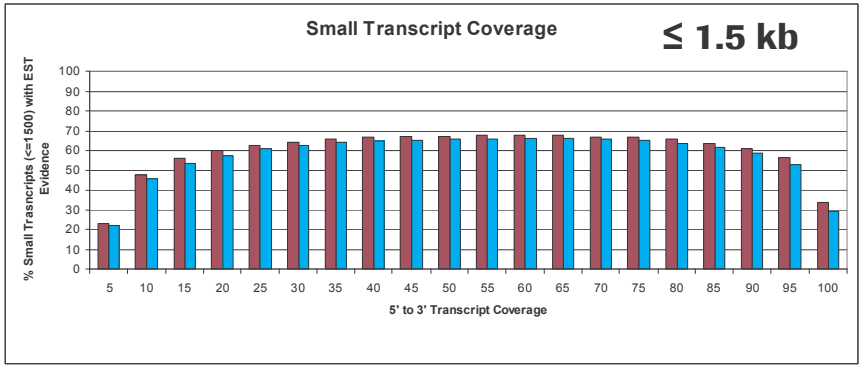
Transcriptome Sequencing
Transcript Coverage Analysis

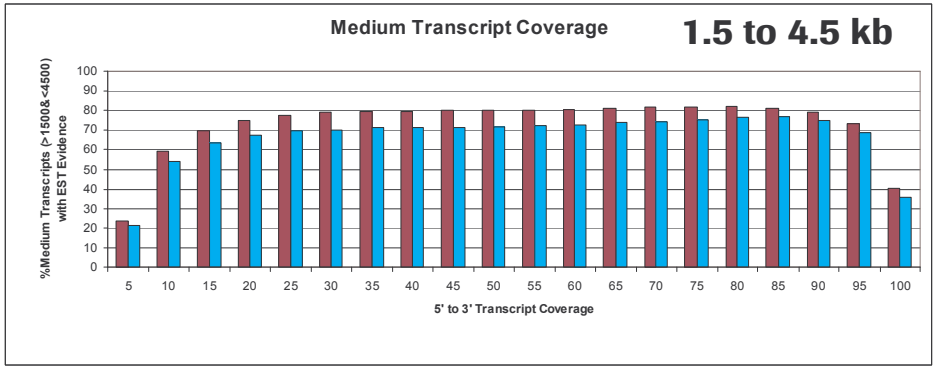
- Transcripts are grouped according to their lengths:
  - Small transcripts: ≤ 1.5 kb
  - Medium transcripts: 1.5 to 4.5 kb
  -Large transcripts: ≥ 4.5 kb

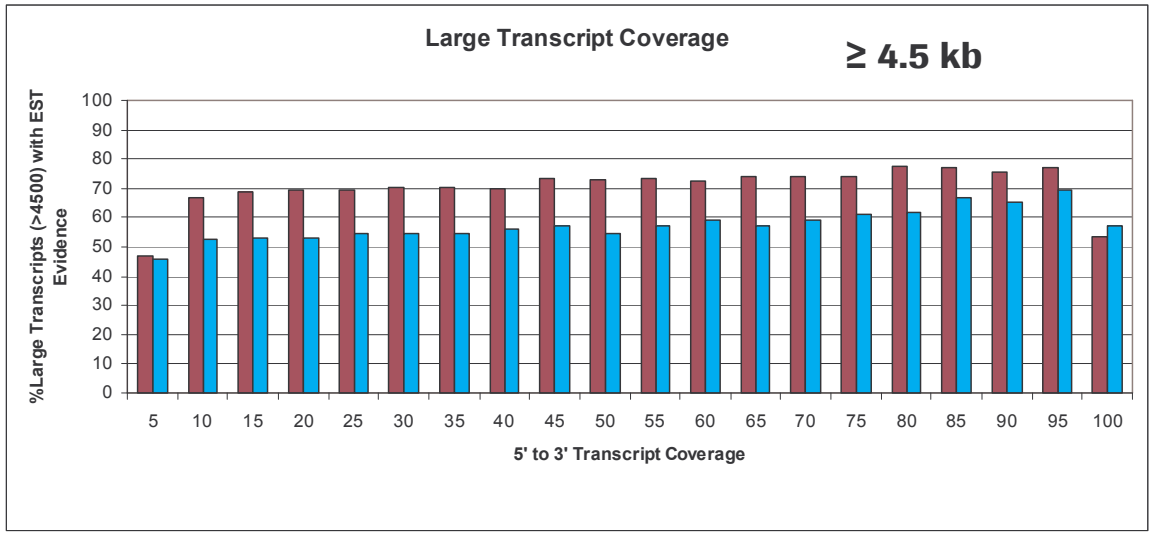
- Within each and every transcript, every 5% of its length would be mapped back to the reference and percentage of mappable hits are reported

- Same number of reads from XLR70 and XL+ are compared

Transcriptome Sequencing
Transcript Coverage







GS FLX+

GS FLX

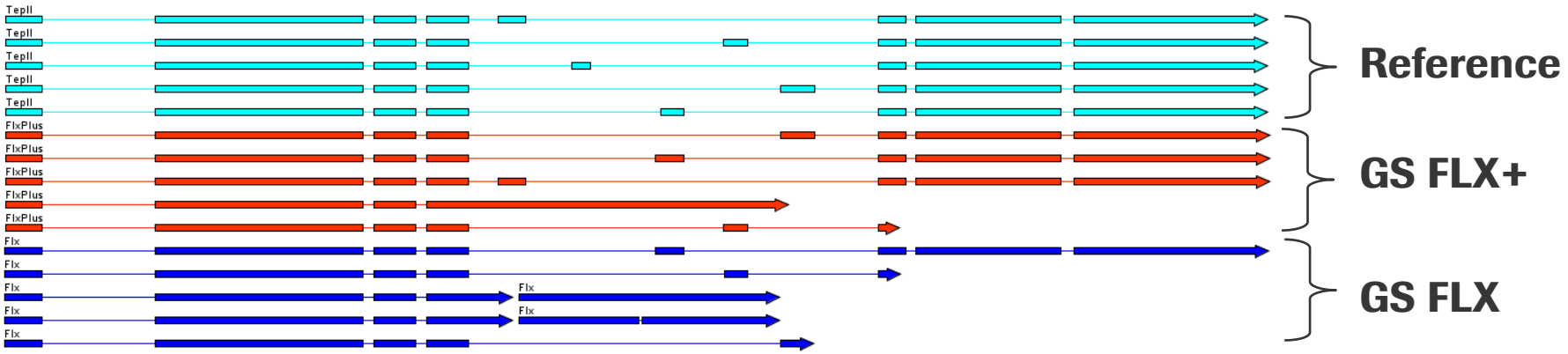
Extra long reads provide much better coverage of long transcripts.

Transcriptome Sequencing
Mapping Results

Mapping to the Chromosome	GS FLX	GS FLX+	Benefit
Number of Reads	1,557,790		
% Reads Mapped	94%	92%	
Number of Bases Mapped	547,105,393	732,429,546	**+34%**
Gene/Transcript/Exon Detection			
% Genes	86%	86%	**-**
% Transcripts	89%	89%	**-**
% Exons	83%	85%	**+2%**
% Avg. Transcript Length Covered	60%	74%	**+14%**
% Chromosome Coverage			
chr2L	24%	26%	**+2%**
chr2R	29%	31%	**+2%**
chr3L	25%	25%	**+2%**
chr3R	26%	29%	**+3%**
chr4	39%	40%	**+1%**
chrX	25%	28%	**+3%**

Extra long reads improve transcriptome mappability

Transcriptome Sequencing
Transcript Assembly Example



- Reference *TepII* transcript is **4,535** bp in length
- Assembled putative transcript from GS FLX+ EST data is an accurate reconstruction of the full length reference transcript, including the splice junctions

Reference Source: Flybase Version 5.18

Transcriptome *de novo* Assembly
GS FLX+ Advancements

Extra-long reads:

- Increase coverage of full-length transcripts, particularly transcripts with length ≥ 4.5 kb
- Span more exons and thus provide more information on splicing patterns and splice junctions
- Generate more accurate and complete assemblies by reducing ambiguity
- Increase the total bases or genes that can be mapped uniquely to the genome

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

www.454.com

454
SEQUENCING