NextGen Sequencing Technologies and Their Applications

Xuan Li

Email: LiXuan@sippe.ac.cn

Institute of Plant Physiology and Ecology, SIBS, CAS
(http://www.sippe.ac.cn)

NGS Platforms

Applied Biosystems
ABI 3730XL
1 Mb /day

Roche / 454
Genome Sequencer
FLX 500 Mb /run

HeliScope/
Single Molecule
Sequencer
1Gb / hour

Church/Dover Systems
Polonator
8000MB /run

Illumina / Solexa
Genetic Analyzer
40 Gb /run

Applied Biosystems/
SOLiD
60Gb /run

二代测序技术发展和应用
请勿复制

New Development

454 GS Junior System

Ilumina HiSeq 2000

ABI SOLiD4

Table 1. Manufacturer's specifications for instrument configuration and production of single end sequences from a single flow cell

Platform	Method	Template prep	Starting DNA (μg)	Instrument configuration	Throughput statistic	Data per run (Gbp)	Reagent cost per run ($)[a]	Run time
454 GS-FLX	Pyrosequencing	Emulsion PCR	3–5	Single picotiter plate, partitionable into 8 lanes	238-bp read[b]	0.1	8500	7.5h
Illumina 1G	Four-color SBS with reversible terminators[c]	Bridge PCR	0.1–1	Single flow cell, partitionable into 8 lanes	35-bp read	1.3	3000	3 d
ABI SOLiD	Oligonucleotide ligation with two-base, four-color encoding	Emulsion PCR	0.1–20	Independently controlled dual-flow cells, each partitionable into 8 lanes	35-bp reads, mapped to reference sequence allowing up to three mismatches	4	3400	7 d
Helicos HeliScope	Single-color SBS with virtual terminators	Not applicable	Not available	Single 25-lane flow cell	30-bp read	7.5	18,000	14 d

[a]Reagent costs are list prices.
[b]Average read length for a typical whole-genome library, using long read kit.
[c](SBS) Sequencing by synthesis.

Holt A. et al., Genome Res. 2008

454

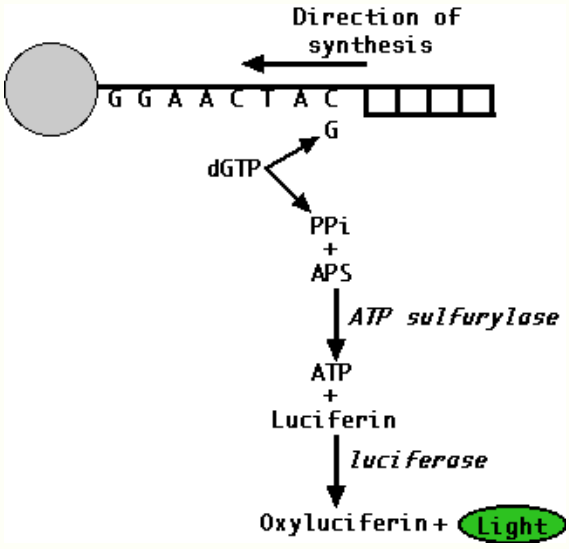
Solexa

SOLiD

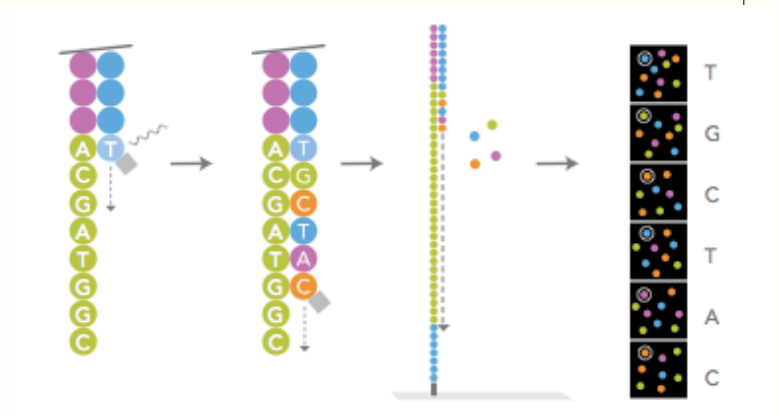
HeliScope

二代测序技术发展和应用

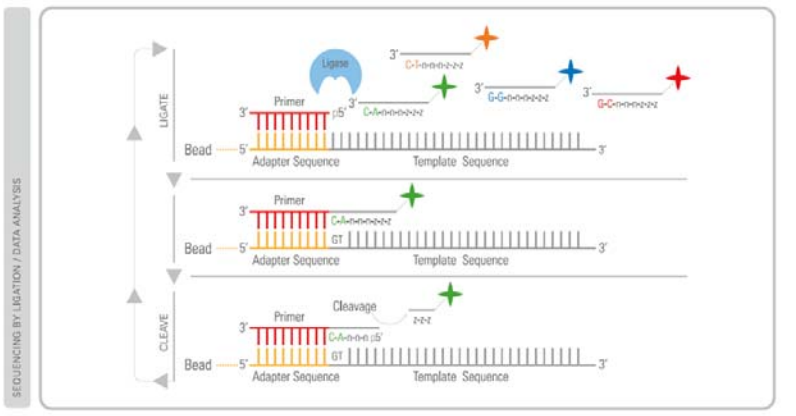
Pyrosequencing



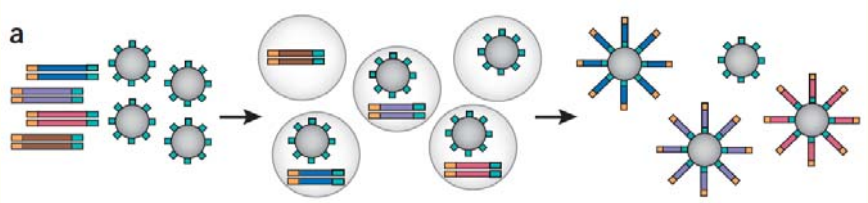
Four Color NT Substrate



Sequencing by Ligation

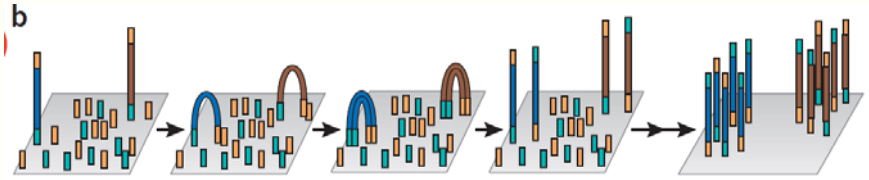


Emulsion PCR



- Fragments, with adaptors, are PCR amplified within a water drop in oil.
- One primer is attached to the surface of a bead.
- Used by 454, Polonator and SOLiD.

Bridge PCR



- DNA fragments are flanked with adaptors.
- A flat surface coated with two types of primers, corresponding to the adaptors.
- Amplification proceeds in cycles, with one end of each bridge tethered to the surface.
- Used by Solexa.

Comparison of NextGen Technologies 2009

	Feature generation	Sequencing by synthesis
454	Emulsion PCR	Polymerase (pyrosequencing)
Solexa	Bridge PCR	Polymerase (reversible terminators)
SOLiD	Emulsion PCR	Ligase (octamers with two-base encoding)
Polonator	Emulsion PCR	Ligase (nonamers)
HeliScope	Single molecule	Polymerase (asynchronous extensions)

4

	Cost per megabase	Cost per instrument	Paired ends?	1° error modality	Read-length
454	~$60	$500,000	Yes	Indel	450 bp
Solexa	~$2	$430,000	Yes	Subst.	100 bp
SOLiD	~$2	$591,000	Yes	Subst.	100 bp
Polonator	~$1	$155,000	Yes	Subst.	13 bp
HeliScope	~$1	$1,350,000	Yes	Del	30 bp

454

Roche / 454 : GS FLX

The Genome Sequencer FLX System

Enable what has been impossible before

System features (Titanium Kits)

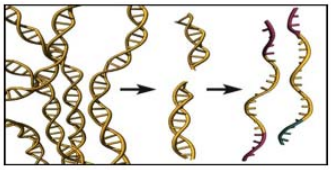
- **>1.000.000 sequence reads per run**
- **400 - 500 nt read length**
- **>500 MB per run (after Q filtering)**
- **>5.0 GB in 5 days**
- **>99% single read accuracy**
- **Low price per result (ca. 80% lower vs. current kits)**
- **All on todays´FLX**

- **Current (GS FLX Kits)**
- 400,000 reads per run
- 250 nt read length
- 100 MB per run
- 2 runs per day
- 1.0 GB in 5 days

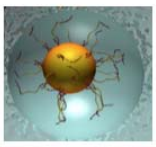
www.roche-applied-science.com

454 LIFE SCIENCES

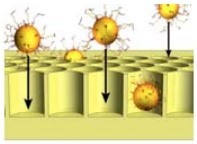
454 LifeSciences Sequencer - Process



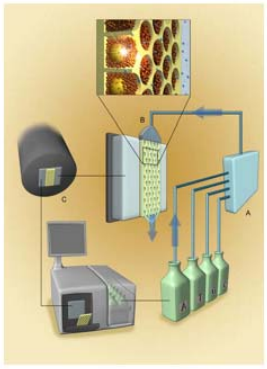
1) Prepare Adapter Ligated ssDNA Library



2) Clonal Amplification on 28 µ beads



3) Load beads and enzymes in PicoTiter Plate™



4) Perform Sequencing by synthesis on the 454 Instrument

454 Data Processing

GS Software Applications

Raw data production
Raw data processing

Instrument Software

Processed Data

Data browsing

GS Data Browser

Assembly

GS de novo Assembler

Mapping

GS Reference Mapper

www.roche-applied-science.com

454 LIFE SCIENCES

Image and Signal Processing Summary

GS FLX Output

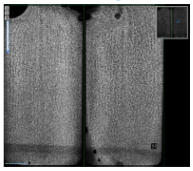


Image Processing *

Image Processing Output

- regions (folder)
  - dataRunParams.xml
  - 1.cwf
  - 2.cwf
- gsRunProcessor.log
- sRunProcessor_err.log

Signal Processing **

Signal Processing Output

- regions (folder)
  - 1.cwf
  - 2.cwf
- sff (folder)
  - C3U5GNB01.sff
  - C3U5GNB02.sff
- gsRunProcessor.log
- gsRunProcessor_err.log

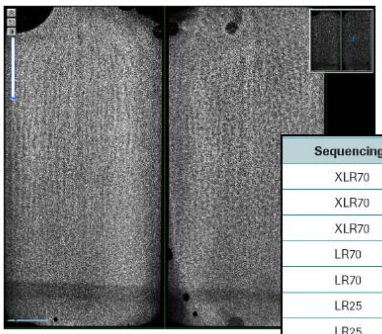
* ~20 min for a Titanium run on the instrument computer
** ~ 7 h for a FLX run on the instrument computer or
~ 8 h for a Titanium run on a cluster

www.roche-applied-science.com

454 LIFE SCIENCES

Image Capturing

Instrument Software



e.g. capturing 834 images for a XLR70 run with 200 cycles (Titanium)

Sequencing Kit	Number of Cycles	Raw Images
XLR70	200	28 Gb
XLR70	150	21 Gb
XLR70	100	14 Gb
LR70	100	14 Gb
LR70	42	5 Gb
LR25	100	6 Gb
LR25	42	3 Gb
SR70	42	5 Gb

www.roche-applied-science.com

454 LIFE SCIENCES

Raw Data Processing

GS Run Processor

- Image Processing
- Signal Processing
- *command line* based

for *Signal Processing* of
Titanium Runs

www.roche-applied-science.com

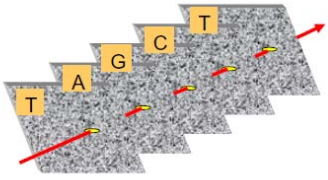
454 LIFE SCIENCES

Image Processing

GS Run Processor

```
runImagePipe RUN_DIRECTORY [options]
```

- Subtract background and normalize the images (at the pixel level)
- Find the centers of the active wells in the PicoTiterPlate device (*i.e.* the locations where sequencing reactions are taking place)
- For each active well, extract the raw signals from the images corresponding to all nucleotide or PPi flows
- Write the resulting flow signals to disk for further processing.



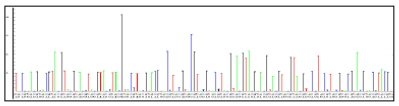
www.roche-applied-science.com

454 LIFE SCIENCES

Signal Processing
GS Run Processor

```
runAnalysisPipe [options] SOURCE_DIRECTORY
```

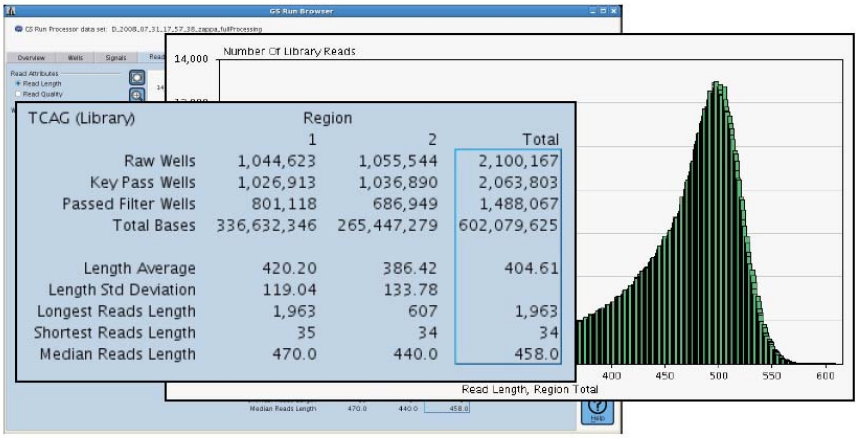
- Correct for interwell cross-talk between neighboring wells
- Correct for known "out-of phase" errors
- Correct for signal droop and subtract residual background signal
- Filter (pass or fail) the processed reads based on signal quality
- Trim read ends for low quality and primer sequence
- Generate "flowgrams" and basecalled sequences with corresponding quality scores for all individual, high quality reads (*i.e.* those which passed all filters). This information is stored as Standard Flowgram Format (SFF) files, to be used as input to the data analysis applications.



www.roche-applied-science.com

454 LIFE SCIENCES

Read Length Distribution – *Reads* Tab
GS Run Browser



TCAG (Library)	Region 1	2	Total
Raw Wells	1,044,623	1,055,544	2,100,167
Key Pass Wells	1,026,913	1,036,890	2,063,803
Passed Filter Wells	801,118	686,949	1,488,067
Total Bases	336,632,346	265,447,279	602,079,625
Length Average	420.20	386.42	404.61
Length Std Deviation	119.04	133.78	
Longest Reads Length	1,963	607	1,963
Shortest Reads Length	35	34	34
Median Reads Length	470.0	440.0	458.0

www.roche-applied-science.com

454 LIFE SCIENCES

GS Software Applications
Overview

Instrument Software
GS Run Processor
GS Reporter
GS Run Browser

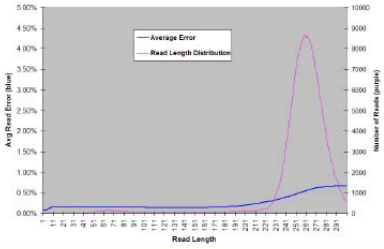
GS de novo Assembler
GS Reference Mapper
GS Amplicon Variant Analyzer

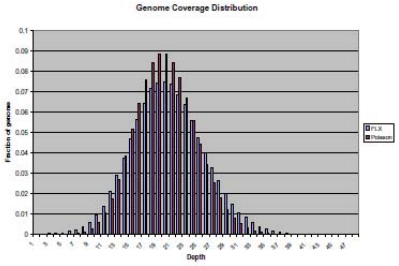
www.roche-applied-science.com

454 LIFE SCIENCES

Example: ***Streptococcus suis***

- 2 Mb bacterium (*Streptococcus suis*)
- De novo assembly yielded 43 contigs
- Mapping to reference shows total error of 6
- Accuracy of consensus sequence is 99.9997%.
- Genome coverage is even -> no bias!





请勿复制

Performance
GS de novo Assembler

E. coli (Bacterium, Genome Size: 4.6 MB)

447,994 GSFLX reads (~250 bp) with 112,8 MB in 7:01 min* (sw 1.1.03)

238,022 Titanium reads (~450 bp) with 100,0 MB in 18:43 min* (sw 2.0.00)

Arabidopsis (Plant, Genome Size: ~130 MB)

1.7 GB Titanium sequence data in ~30 h** (sw 2.0.00)

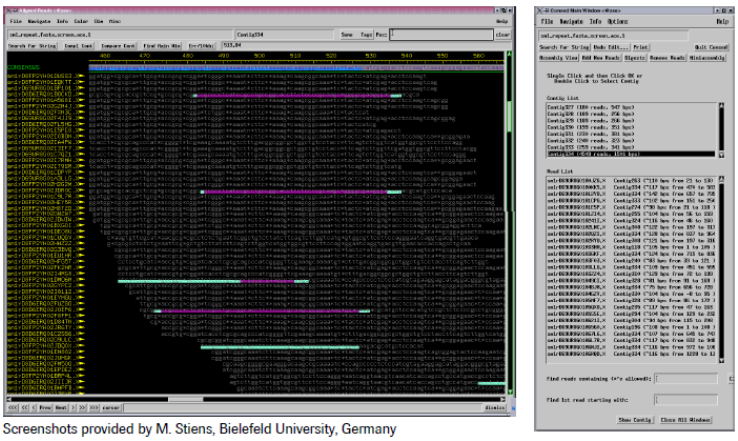
*Office PC with RHEL 5 (64-bit) 8 GB RAM. One CPU/core used!

** PC with RHEL (64-bit), 32 GB RAM. One CPU/core used!

www.roche-applied-science.com

454 LIFE SCIENCES

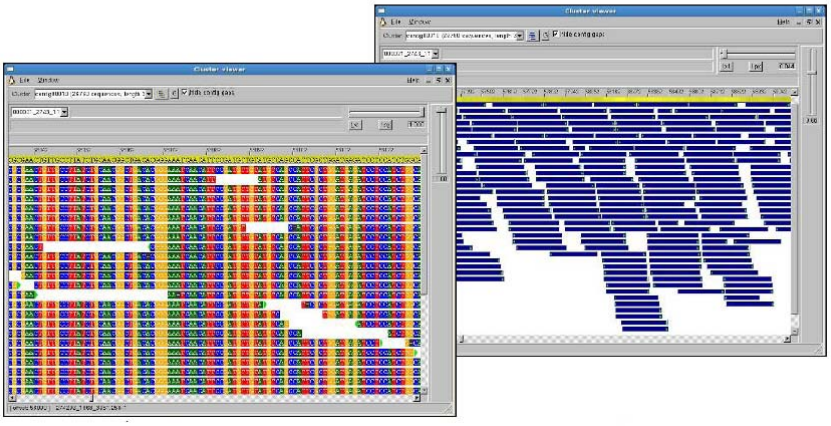
3rd Party Software Compatible Output
Consed from WashU



Screenshots provided by M. Stiens, Bielefeld University, Germany

3rd Party Software Compatible Output

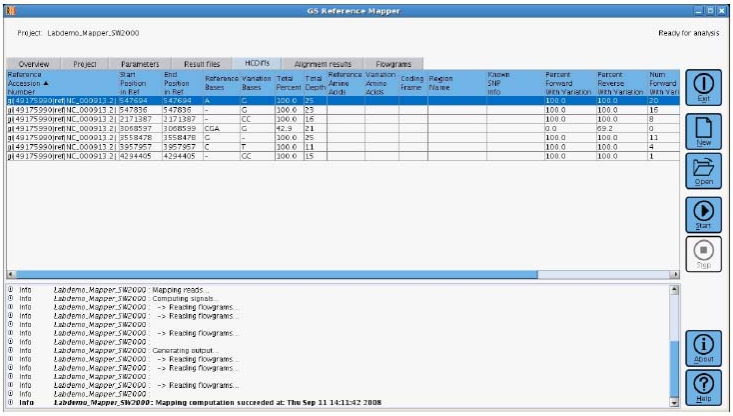
clview from TIGR



请勿复制 25

"High Confidence" Differences Table

GS Reference Mapper

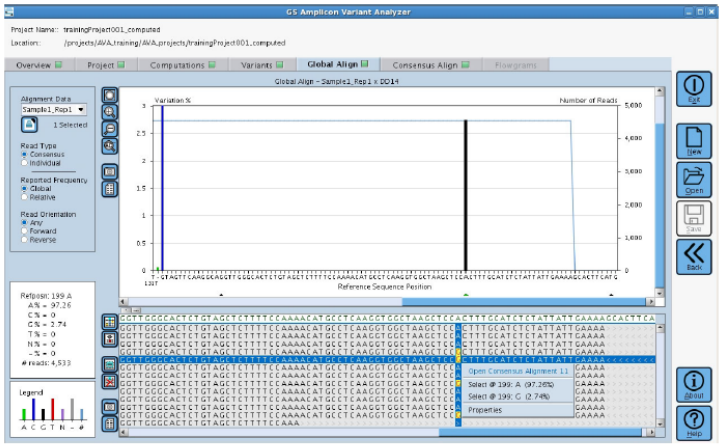


www.roche-applied-science.com

454 LIFE SCIENCES

请勿复制 26

Analyzing Amplicon Variants
GS Amplicon Variant Analyzer



www.roche-applied-science.com

454 LIFE SCIENCES

Helper *SFF Tool Commands*

sfffile [options] [MIDlist@] (sfffile | datadir) …

- SFF file(s) modification (e.g. merging of two or more SFF files; in- or exclude reads from a SFF file)

sffinfo [options] [- | sfffile] [accno …]

- information extraction from a SFF file (e.g. generating fasta and quality scores files from a SFF file)
- output: text file format

sff2scf locationstring [outputfile]

- converts SFF files to scf files (for Consed)

fnafile [options…] (fastafile or PHDfile or SCFfile or directory) …

- pooling and convertion of FASTA, SCF, and PHD files to a single FASTA file (+ quality score file)

sffrescore [-f] (file | directory) …

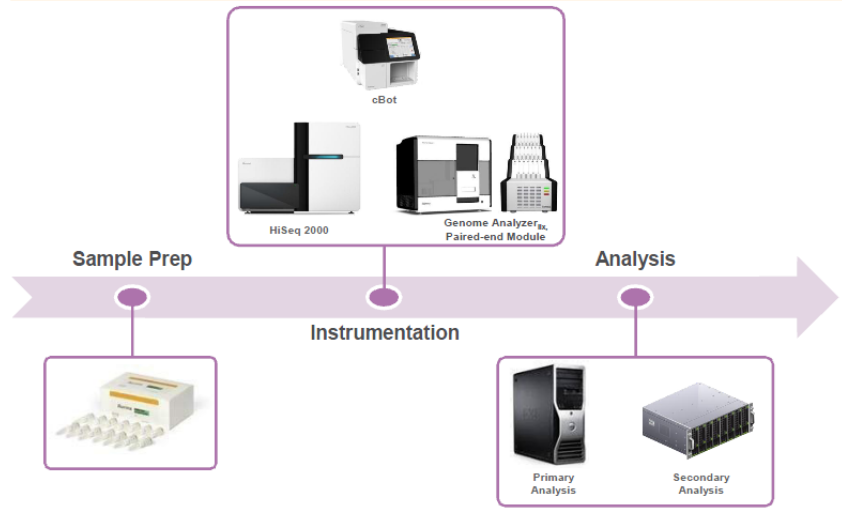
- rewrite existing SFF files with the new Phred-like quality scores

www.roche-applied-science.com

454 LIFE SCIENCES

Illumina

Illumina Sequencing Workflow



Paired end Sequencing process

1 *Library prep (~ 6 hrs)*

Fragment DNA
↓
Repair ends / Add A overhang
↓
Ligate adapters
↓
Select ligated DNA

2 *Automated Cluster Generation (~ 5 hrs)*

1-96 samples

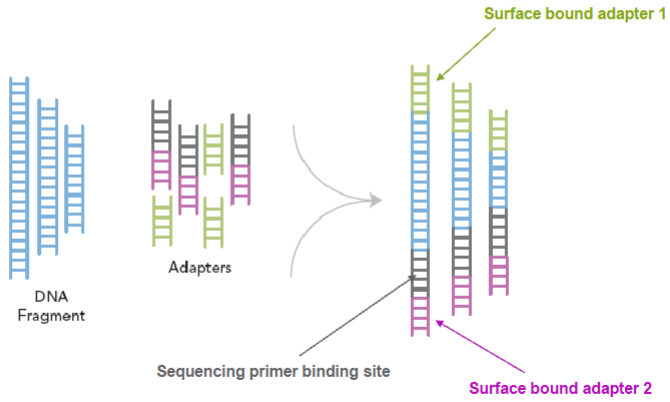
Hybridize to flow cell
↓
Extend hybridized oligos
↓
Perform bridge amplification

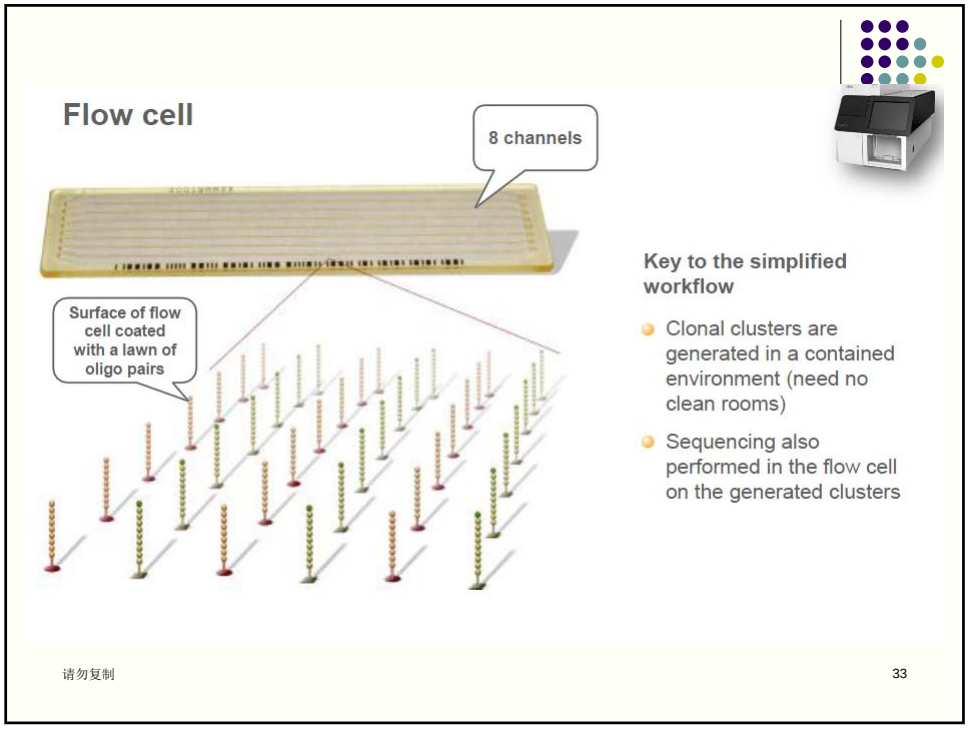
3 *Sequencing (~ 46 to 120 hrs)*

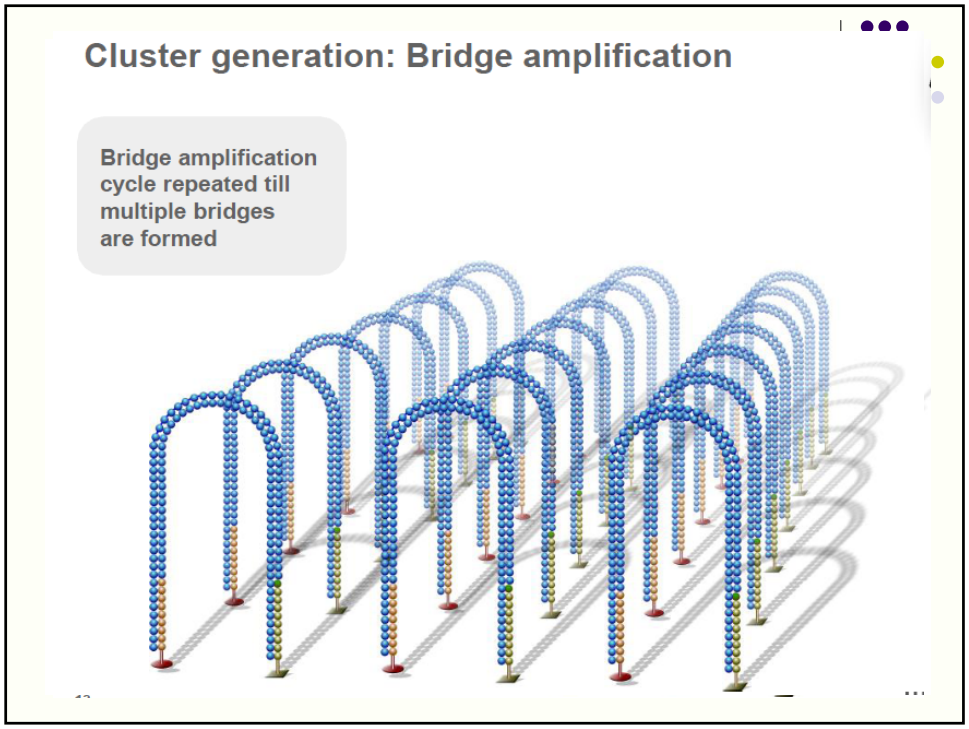
1-96 samples

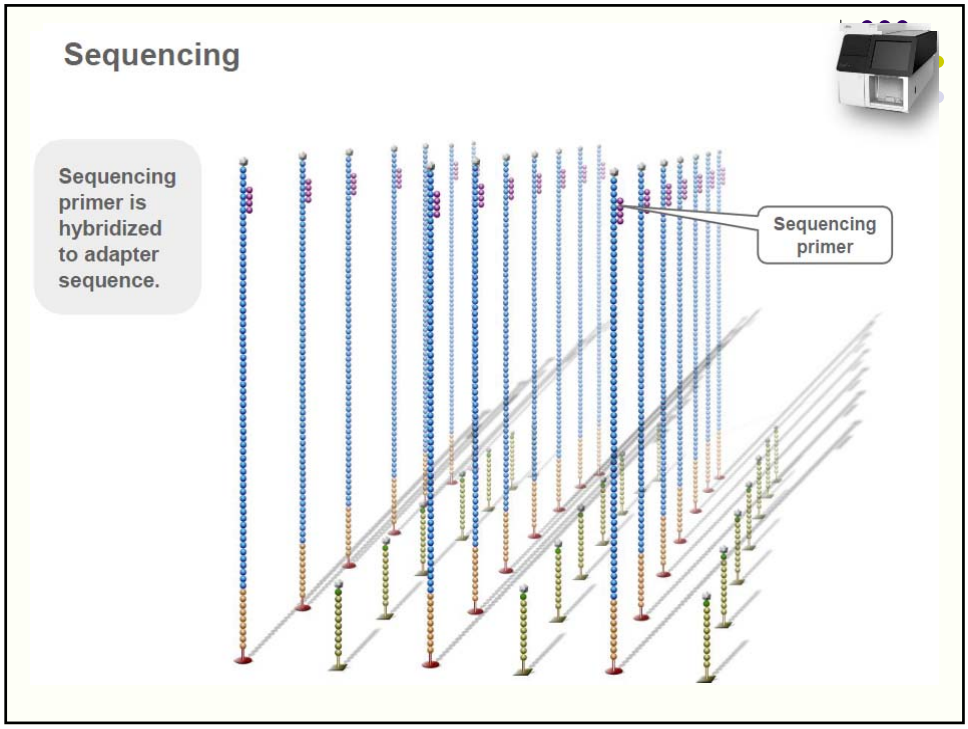
Perform sequencing on forward strand
↓
Re-generate reverse strand
↓
Perform sequencing on reverse strand

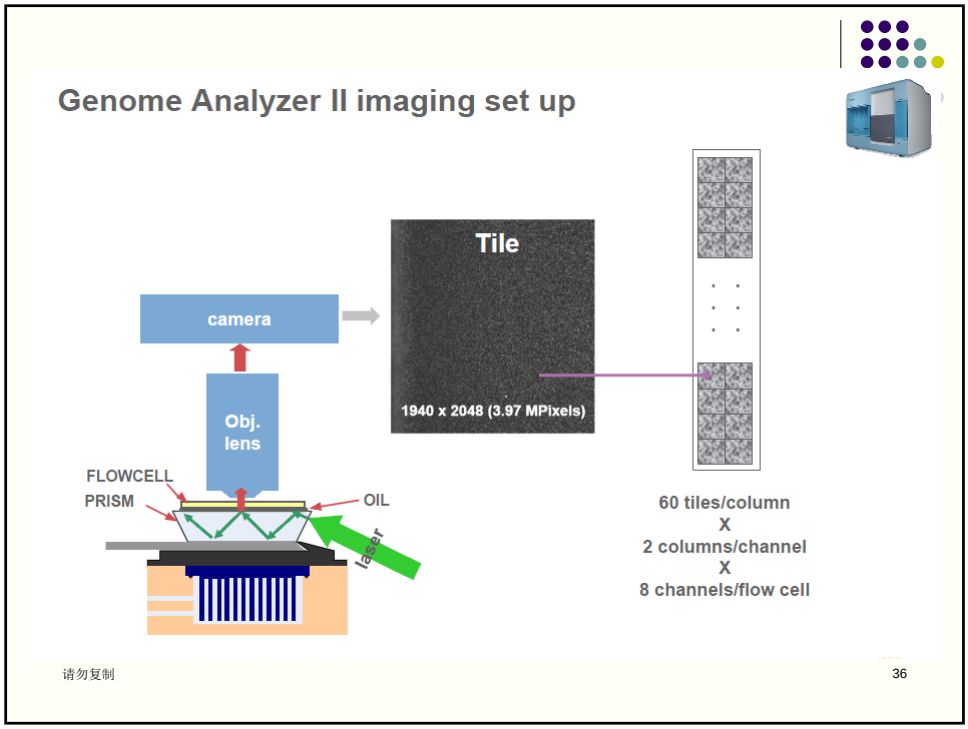
Sample Prep - Resequencing



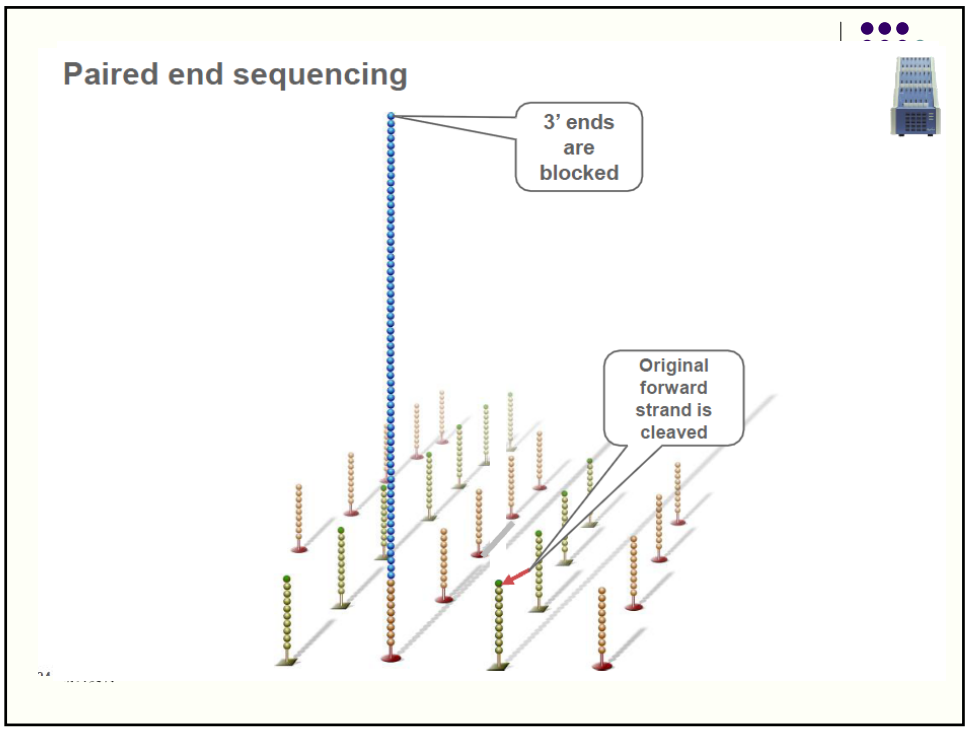

Flow cell
8 channels
Surface of flow cell coated with a lawn of oligo pairs
Key to the simplified workflow
Clonal clusters are generated in a contained environment (need no clean rooms)
Sequencing also performed in the flow cell on the generated clusters
请勿复制
33

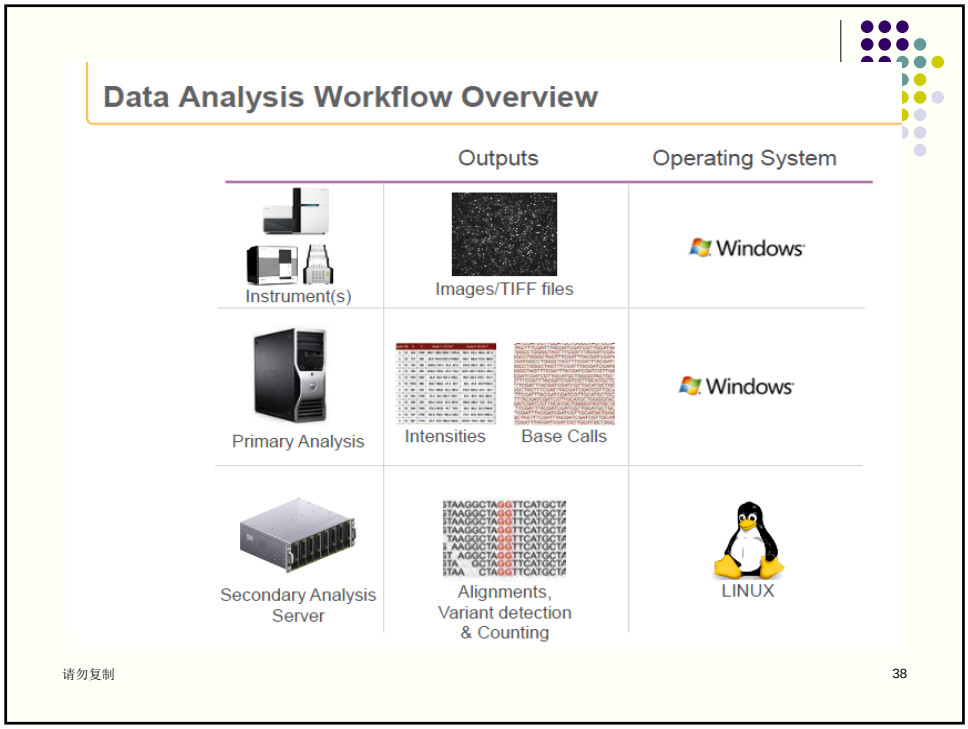

Cluster generation: Bridge amplification
Bridge amplification cycle repeated till multiple bridges are formed





二代测序技术发展和应用
请勿复制

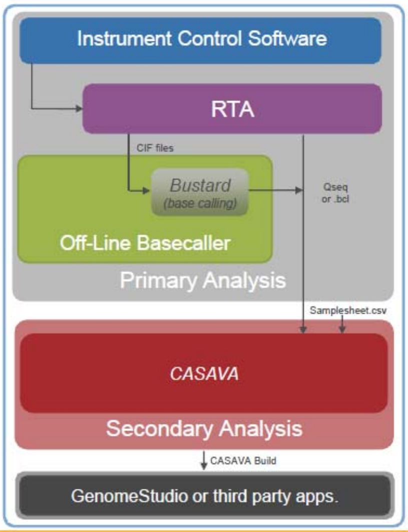




二代测序技术发展和应用
请勿复制

Primary Data Analysis Workflows

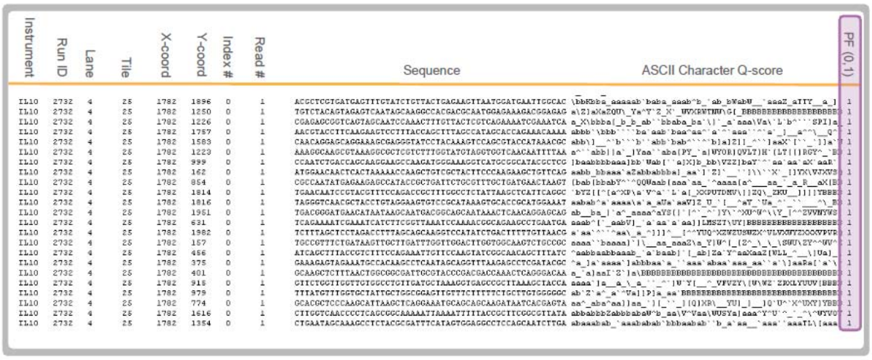
- Primary Analysis is performed by one of three workflows:
  1. The Instrument control software/RTA analyze images and produce base call files
  2. Data analysis from images
     - *(Genome Analyzer only. Not possible with the HiSeq 2000)*
  3. Offline basecalling from Intensity files



Note: Primary Analysis prior to v1.6 was called Pipeline

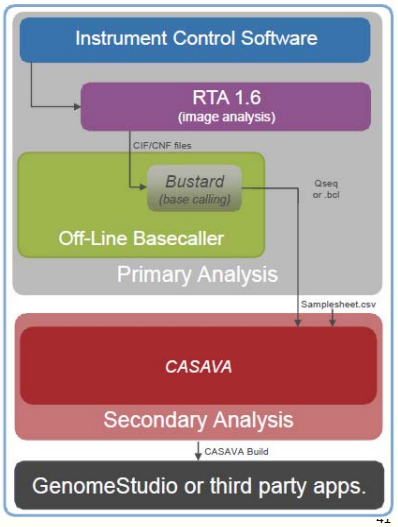
qseq.txt File

- Pass/Fail Filtering:
  - 1 = yes
  - 0 = no



OLB Workflows

- OLB provides the option to perform data analysis off-line on a UNIX machine
- OLB offers two different workflows:
  1. Data analysis from images
  2. Data analysis from intensities



Note: Primary Analysis prior to v1.6 was called Pipeline

Off-Line Basecalling

Test Command

Checks the syntax and required data is present

```
/path/to/Pipeline/bin/bustard.py --CIF
--GERALD=config.txt /path/to/RunFolder/Data/Intensities
```

Create MakeFiles

Builds the run folder structure and generates the MakeFiles

```
/path/to/Pipeline/bin/bustard.py --CIF
--GERALD=config.txt /path/to/RunFolder/Data/Intensities
--make
```

cd into the analysis directory

MakeFiles are created in the Bustard analysis directory

```
cd /path/to/RunFolder/Data/Intensities/Bustard1.4.0_DD-MM-
YYYY_user
```

Execute MakeFiles

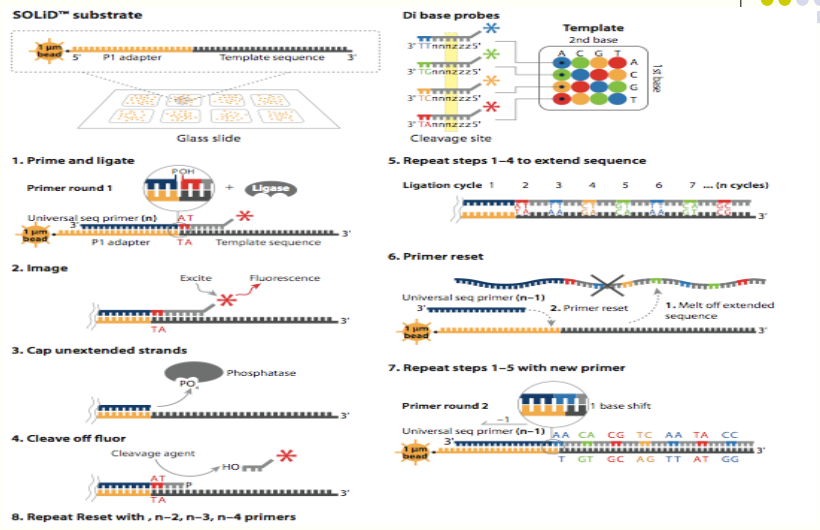
Starts the Pipeline run

```
nohup make -j <n> recursive
```

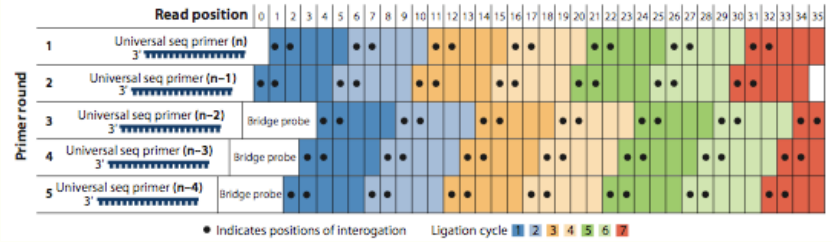
SOLiD

(Sequencing by Oligonucleotide Ligation and Detection)

ABI SOLiD



ABI SOLiD



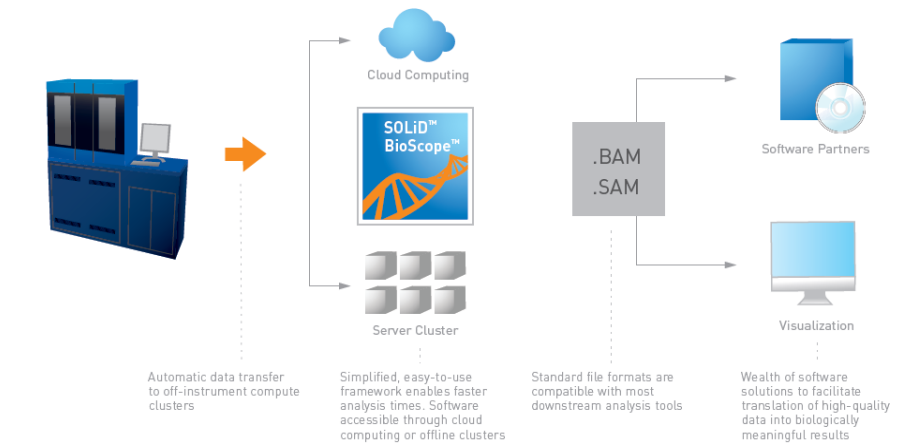
ABI SOLiD-3

- 60Gb per run/ 10 days / $14K per run
- Current read length = 35-50 bp
- Requires emPCR amplification, etc.
- Ligation-mediated sequencing with two-base encoding
- Paired end reads separation of 3kb available
- Limitation on two-base encoding requires reference sequence for alignment
- Platform not yet fully automated (walk-away)

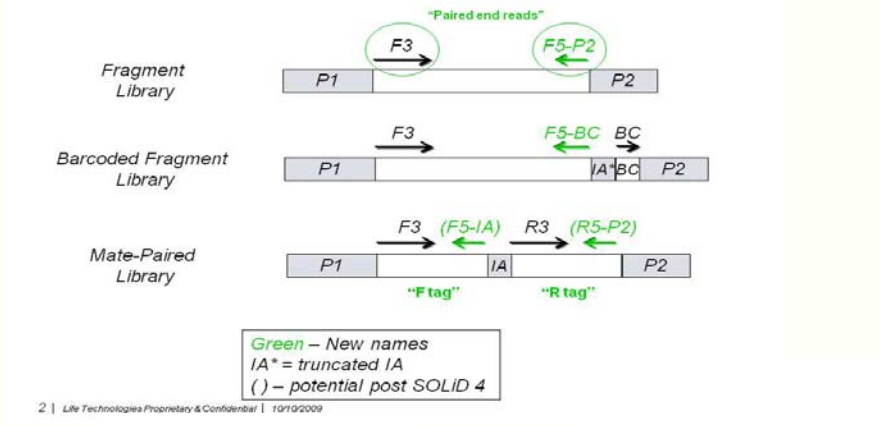
SOLiD Data Analysis

- On-Instrument Analysis: **ICS/SETS**
  - Paired End
  - Improved Multiplexing / Barcode support
  - Export
- Off-instrument Data Analysis: **Bioscope 1.2**
  - Data Format Standardization / SAM Tools
  - Paired End
  - Barcoding
  - RNA Seq update: New gene fusion and splice junctions detection
  - New SAET Tool
  - New ChipSeq Mapping / recommended analysis workflow
  - User Interface Improvements
  - Access to BioScope: cluster, cloud
- SOLiD Software Community
- Data Analysis / BioScope resources

SOLiD™ System Data Analysis Workflow



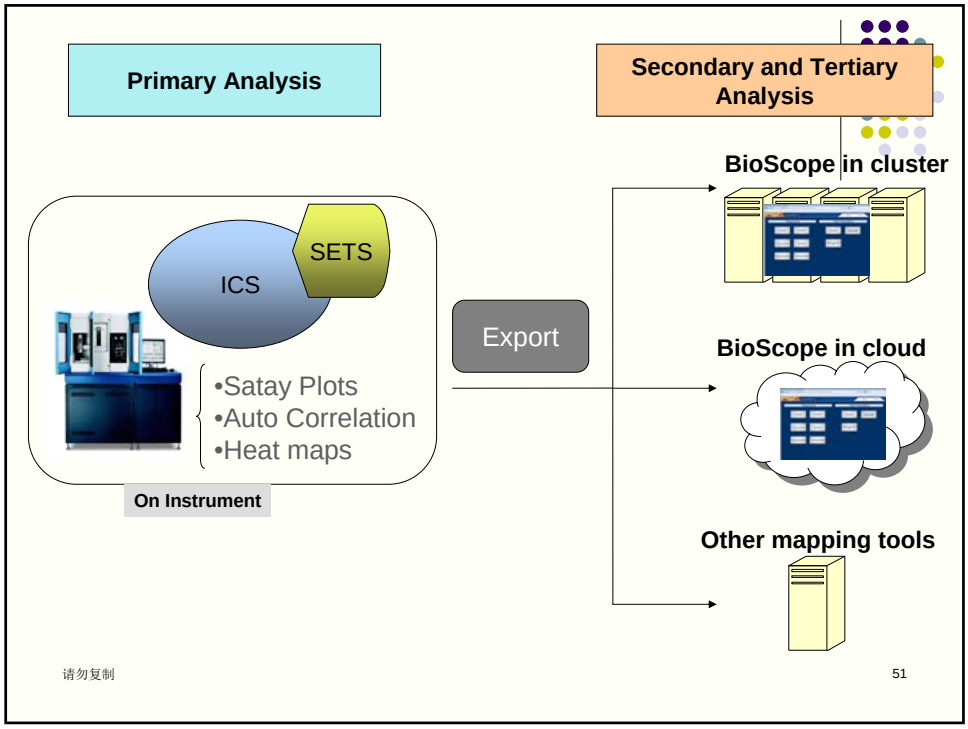
Paired-End:
Naming convention for sequence reads

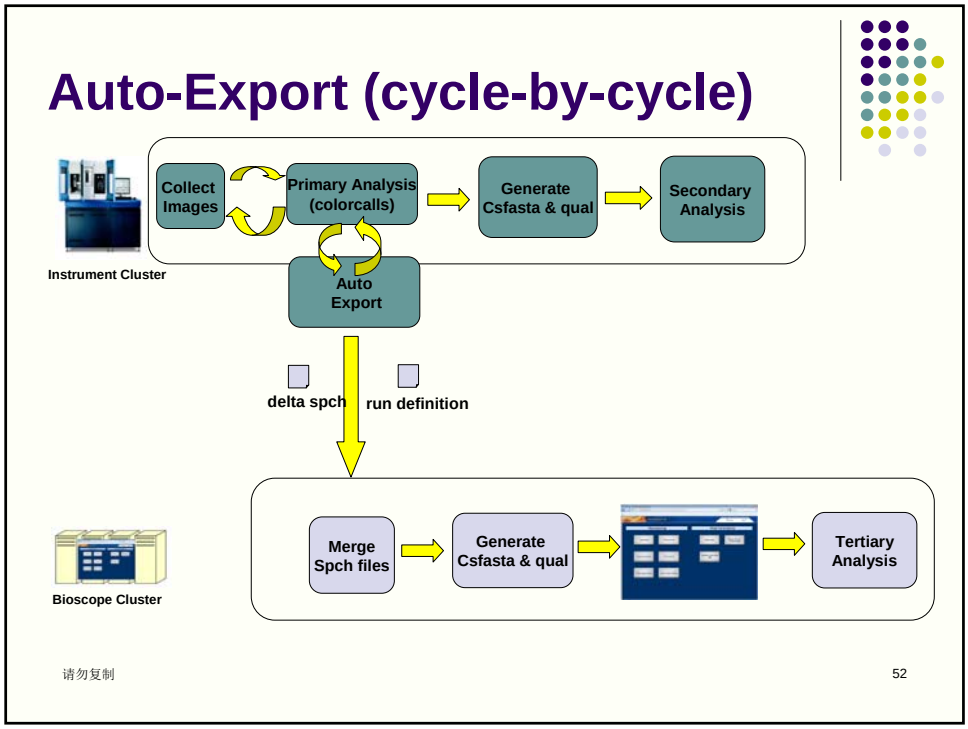


Paired-End: New Run Types in ICS

- Support paired-end run
  - Primer sets: F5-P2, F3
- Support paired-end multiplexing run
  - Primer sets: BC, F5-BC, F3
- New run protocols

Run Types
Paired-end *NEW*
Fragment
Mate Pair
Paired-end Multiplexing *NEW*
Fragment Multiplexing


Primary Analysis
Secondary and Tertiary Analysis
BioScope in cluster
ICS
SETS
Export
BioScope in cloud
•Satay Plots
•Auto Correlation
•Heat maps
On Instrument
Other mapping tools
请勿复制
51


Auto-Export (cycle-by-cycle)
Collect Images
Primary Analysis (colorcalls)
Generate Csfasta & qual
Secondary Analysis
Instrument Cluster
Auto Export
delta spch
run definition
Merge Spch files
Generate Csfasta & qual
Tertiary Analysis
Bioscope Cluster
请勿复制
52

Manual Export

Instrument Cluster

Collect Images

Primary Analysis (colorcalls)

Generate Csfasta & qual

Secondary Analysis

Option in SETS to export All run data or combination of Spch, csfasta & qual files

Bioscope Cluster

Tertiary Analysis

BioScope 1.2 Overview

Primary Image Analysis

Secondary Mapping Analysis

Tertiary Analysis

Visualization

Off-Instrument Compute Cluster

mapping

pairing

SNP

→WIG

mapping

counting

SAM/BAM

Fusion

mapping

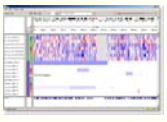
SNP

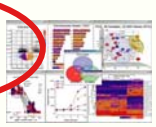
large indel

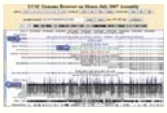
small indel

CNV

BioScope







SAM/BAM overview

- SAM is an alignment format developed by 1000 genome project that includes pairing information
- Similar to BED and GFF, but extended to include pairing details
- SAM refers to the generic format specification and the text file
- BAM is the compressible binary version of SAM
- Resources:
  - Main site http://samtools.sourceforge.net/
  - Format specification http://samtools.sourceforge.net/SAM1.pdf
  - Mailing lists http://sourceforge.net/mail/?group_id=246254

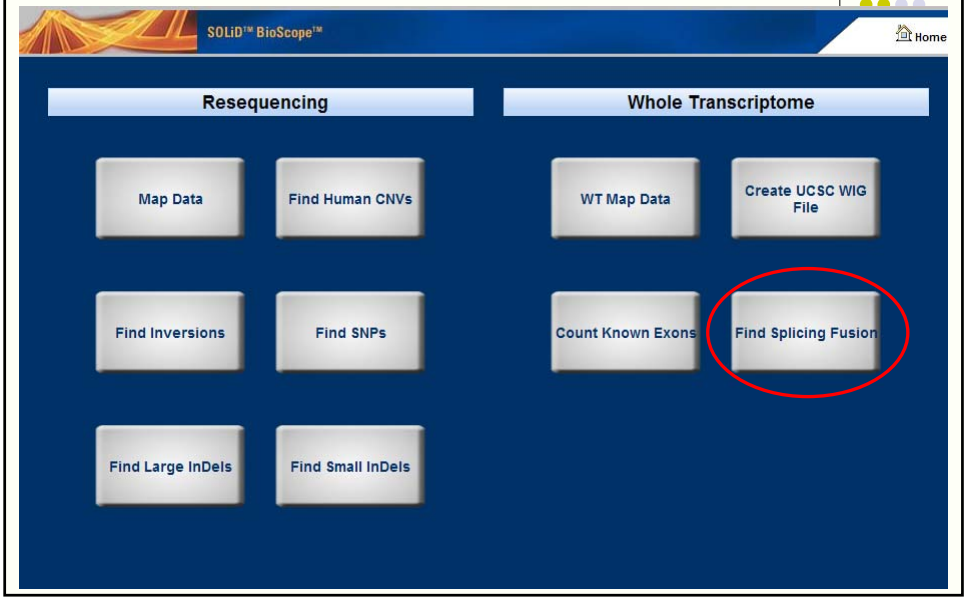
IGV to show multiple files

Supported library types in BioScope 1.2

	Fragment	Paired end	Mate pair
Resequencing			
Mapping	Yes	Yes	Yes
Human CNV	Yes	Yes	Yes
Inversion	No	No	Yes
SNP Finding	Yes	Yes	Yes
Large Indel	No	Yes	Yes
Small Indel	Yes*	Yes	Yes
Whole Transcriptome			
Coverage	Yes	Yes	n/a
Known Exons	Yes	Yes	n/a
Splice junctions	No	Yes	n/a
Gene Fusions	No	Yes	n/a

*ial plugin called Small Indel Frag

What's new in Bioscope 1.2 wt pipeline?




Thank you!