FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

Living with SOLiD, Helicos (and Solexa)

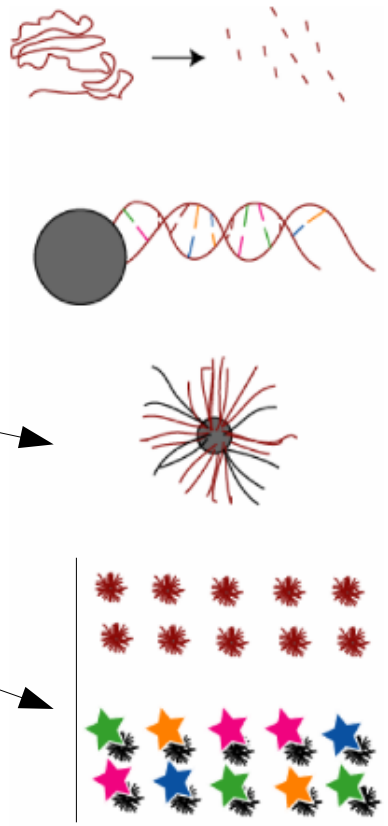
Workshop on next generation sequencing data analysis
CSC, 2.6.2010

Asta Laiho
The Finnish Microarray and Sequencing Centre
Turku Centre for Biotechnology
University of Turku

FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

Basics of the next generation sequencing technology

- Get and process DNA/RNA
- Attach it to something
- Extend and amplify signal with some color scheme
- Detect fluorochrome by microscopy
- Interpret series of spots as short strings of sequence (25 – 400 bp long)
- Multiple images are interpreted as 1-100 GB/run
- Assemble / align strings to reference sequence

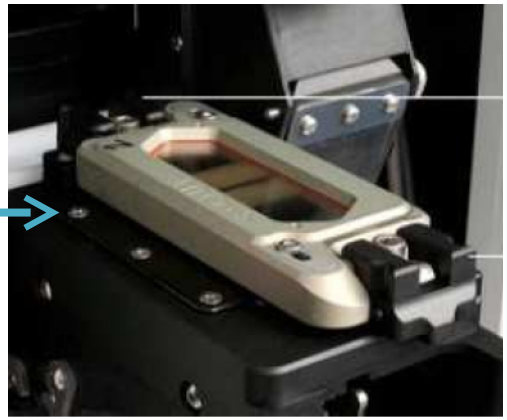


FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

applied biosystems
part of life technologies

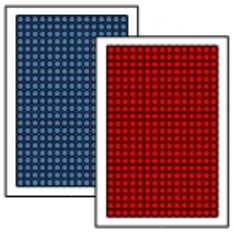
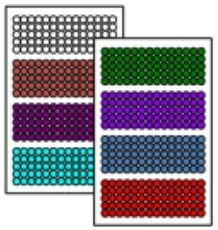
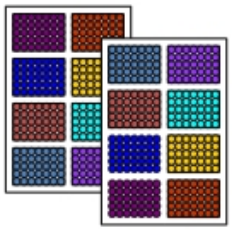
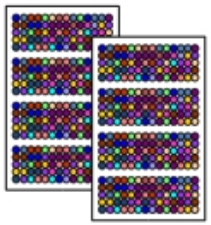
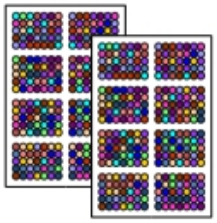
SOLiD4 system

(Sequencing by Oligonucleotide Ligation and Detection)

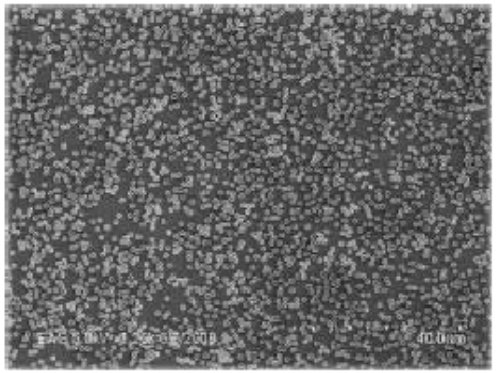
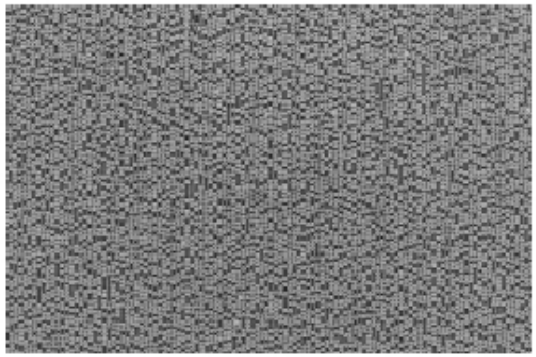
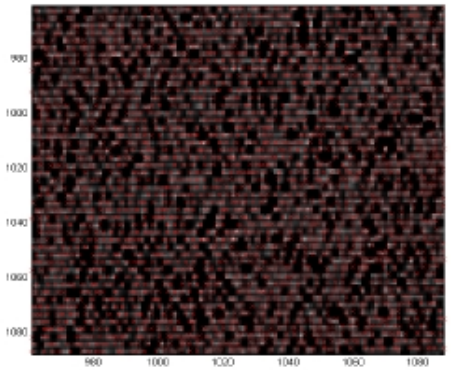


FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

Up to 1536 samples with multiplexing

	Unsegmented Full Slide	Quad	Octet
Individual Samples	 2 samples/run	 8 samples/run	 16 samples/run
Multiplexed Samples 96 barcodes	 ≤96 samples/slide x 2 slides/run ≤ 192 samples/run	 ≤20 samples/segment ≤4 segments/slide x 2 slides/run ≤ 768 samples/run	 ≤20 samples/segment ≤8 segments/slide x 2 slides/run ≤ 1,536 samples/run

FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

	SOLiD 3 Plus	SOLiD 4	SOLiD 4*hq*
			
GB/run	60 GB	100 GB	~300 GB
Beads per run	1 billion	1.4 billion	2.4 billion
Read length	2x50 bp	2x50 bp	2x75 bp
Bead size	1 micron	1 micron	0.75 micron

FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

How many reads are needed?

The number of reads needed is dictated by the complexity of application

Application	Complexity	Reads	Estimate mappable reads needed	Samples SOLID 4
Small RNA Discovery	Low to Mid	35 bp	~10M	up to 50 / slide
dGEx/SAGE	Low	35 bp	5M	100 / slide
Expression of annotated genes	Mid	50 bp	10-100M (human)	6 to 60 / slide
Whole Transcriptome Discovery	High (alternative transcripts & splicing)	50 bp	>100 million (human)	6 / slide
Allele Specific Expression	High (variants to be defined)	50 bp	>150 million (human)	4 / slide

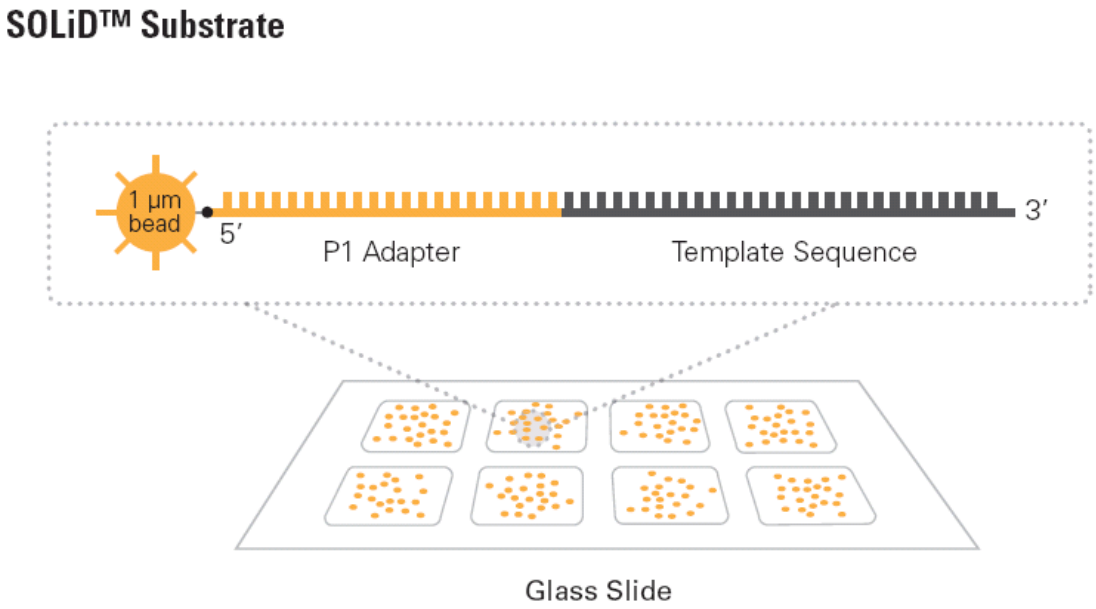
* Current best estimates from literature and internal research

applied biosystems
part of life technologies

FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

SOLiD sequencing chemistry

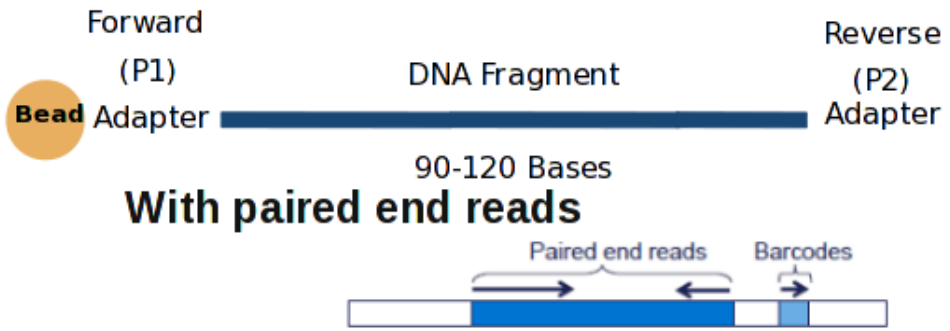
- The SOLiD chemistry is based on sequential ligation of fluorescently labeled oligonucleotide probes
- Available read lengts: **25bp, 35bp, 50bp**



FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

Applications and available libraries

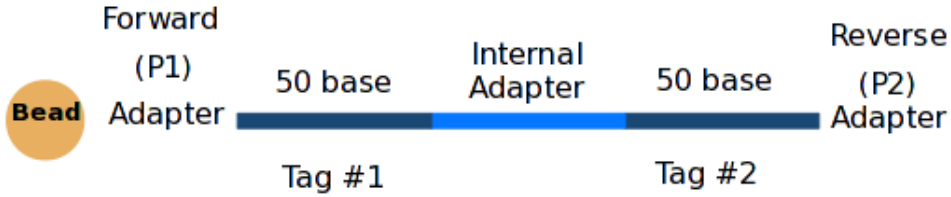
Fragment Library



starting material : 10ng-20ug

- Whole Transcriptome (RNA)
- Targeted Resequencing
- 3' SAGE or 5' SAGE
- ChIP-seq
- SNP Discovery
- *De novo* seq

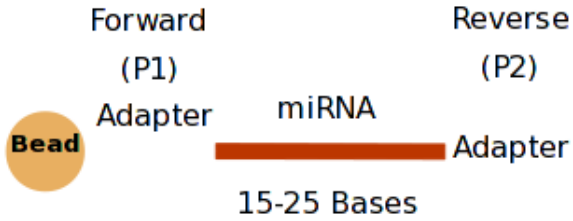
Mate Paired Library



starting material: 5ug-50ug

- Whole Genome Sequencing
- SNP Discovery
- Digital Karyotyping
- Methylation

Small RNA Library

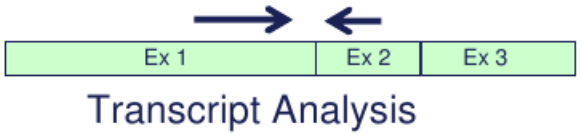


starting material:
total RNA: 0.5g-20ug RNA
purified small RNA fraction: 10-200ng

- miRNA Discovery and/or Profiling
- Gene Expression

FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

Paired end reads



Transcript Analysis

Fusion Transcript Detection

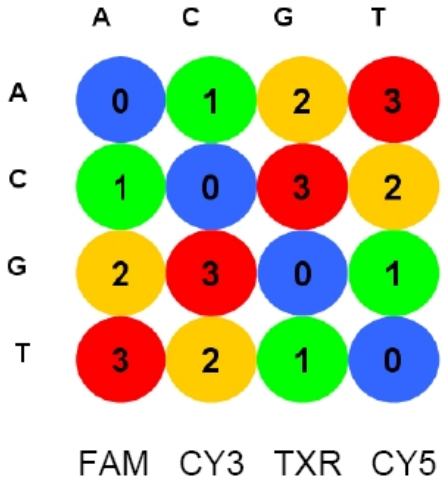
Structural Variation (SV) Analysis

Small & Large Indel Analysis

FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

Sequencing in color space

- Color space data



- Turned into 0,1,2,3 into digital data

```
>1279_26_18_F3
T13221021003120111002113333
>1279_26_41_F3
T12210131000231210012333321
>1279_26_71_F3
T01011013302301221130302223
>1279_26_192_F3
T32200031231222013000220000
>1279_26_254_F3
T02022103133320211120231200
>1279_26_332_F3
T31002301001011302322202300
>1279_26_373_F3
T01230121221133122232020310
>1279_26_380_F3
```

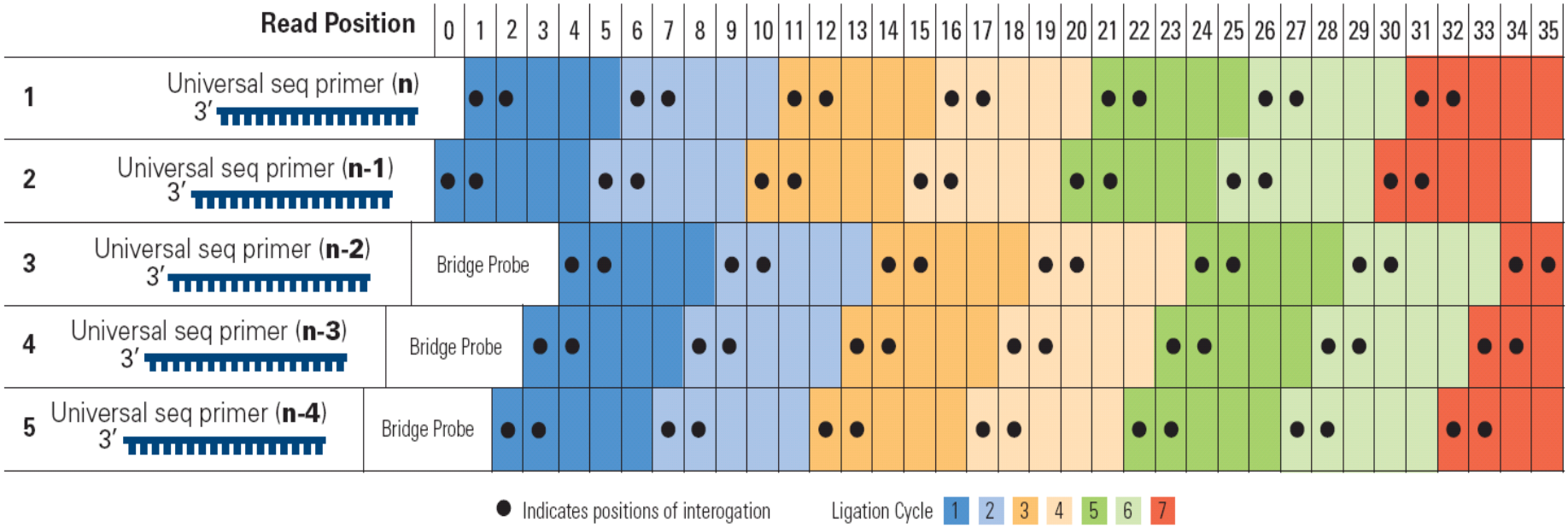
Each base is **interrogated twice,** by two **independent** reactions and probes

Error probability is the combined probability of two independent probes reading the same base wrong

This results in a **very high accuracy** (> 99.94 %)

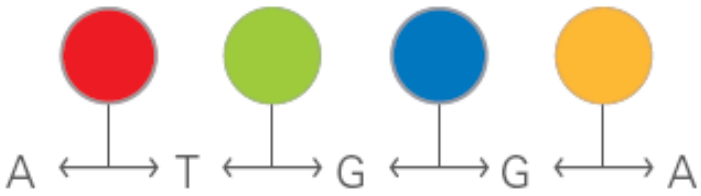
FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

8. Repeat Reset with , n-2, n-3, n-4 primers

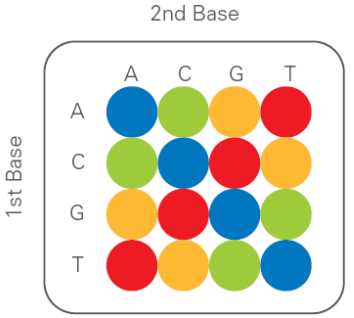


Double Interrogation

With 2 base encoding each base is defined twice



FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi



Advantages of Di-base Encoding

Double Base Interrogation

A C G G T C G T C G T G T G C G T

Color Space Sequence

A C G G T C G T C G T G T G C G T

Possible Base Change

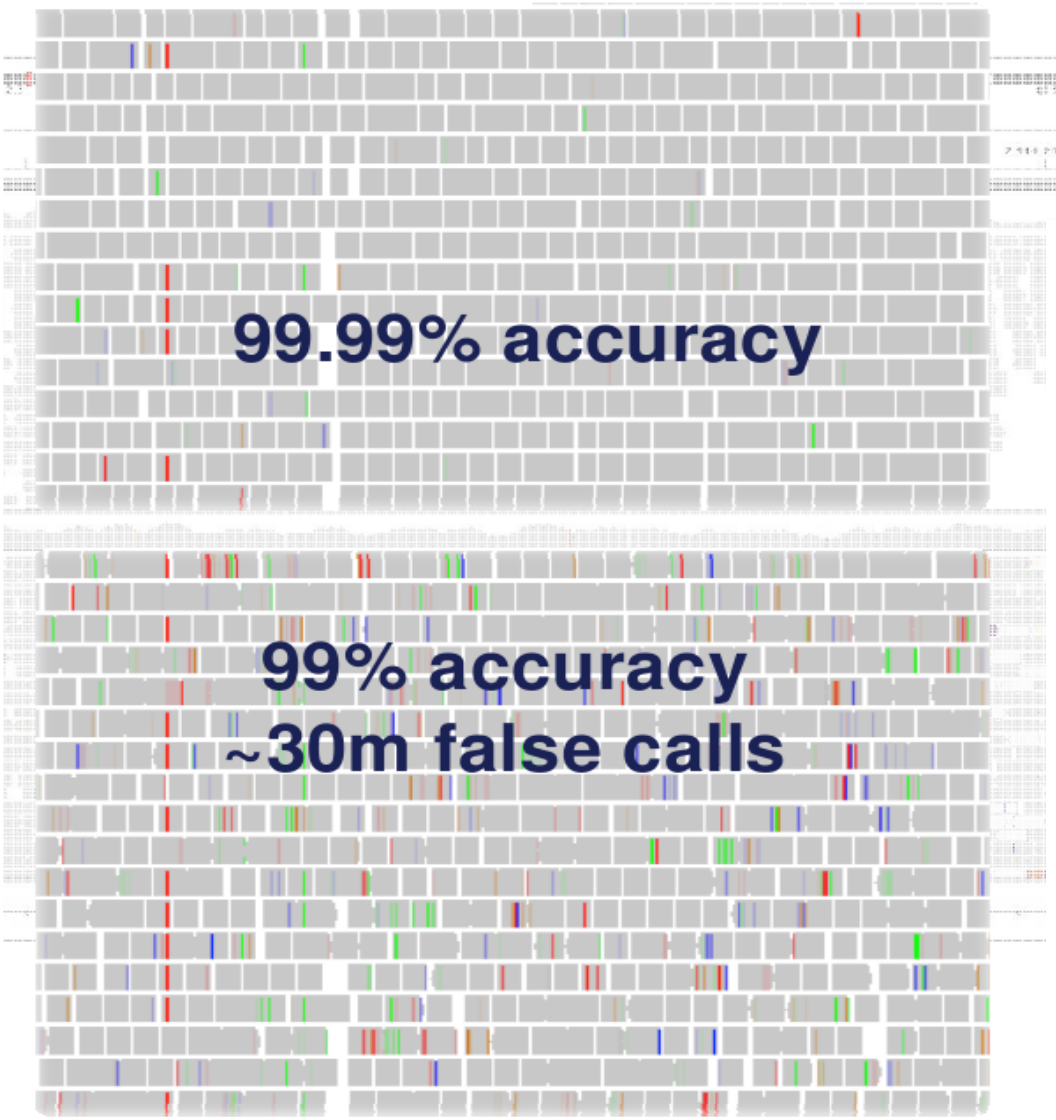
A C G G T C G C C G T G T G C G T

Measurement Error

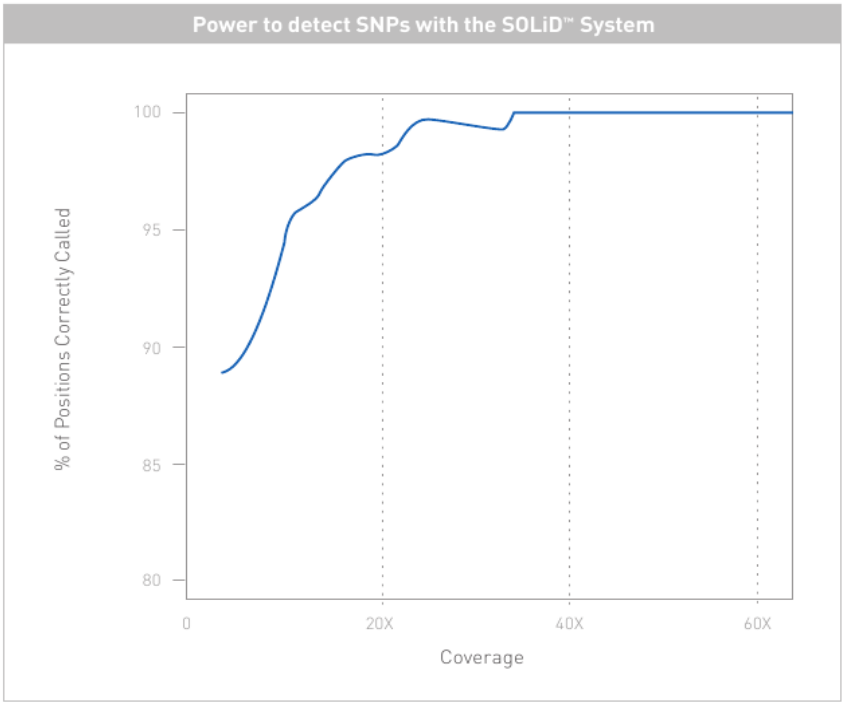
A C G G T C G T T A C A C A T A C

FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

About the importance of accuracy



Less sequencing is needed to identify rare variants

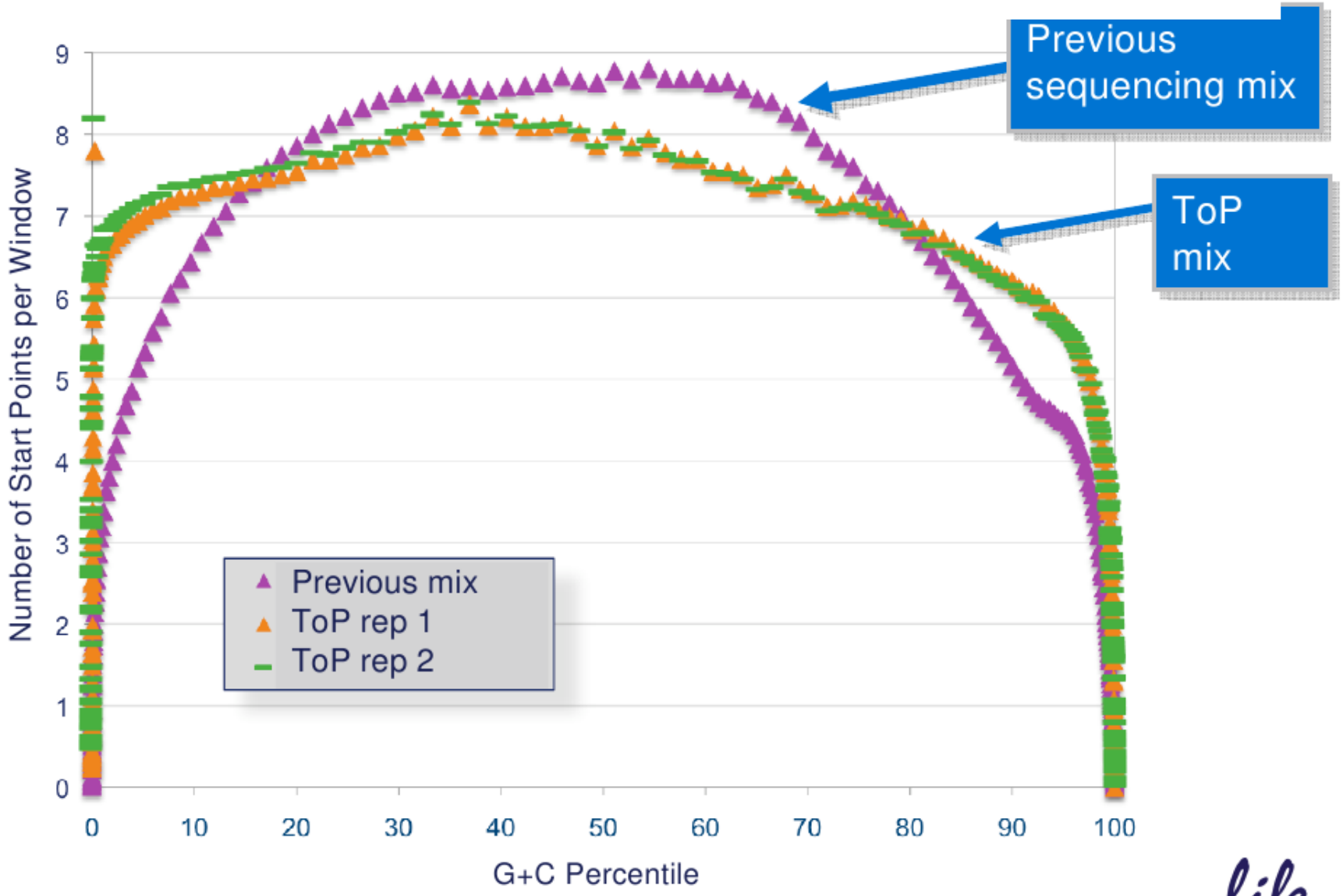


nstitute.org/igv
0 Genc

life technologies™

FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

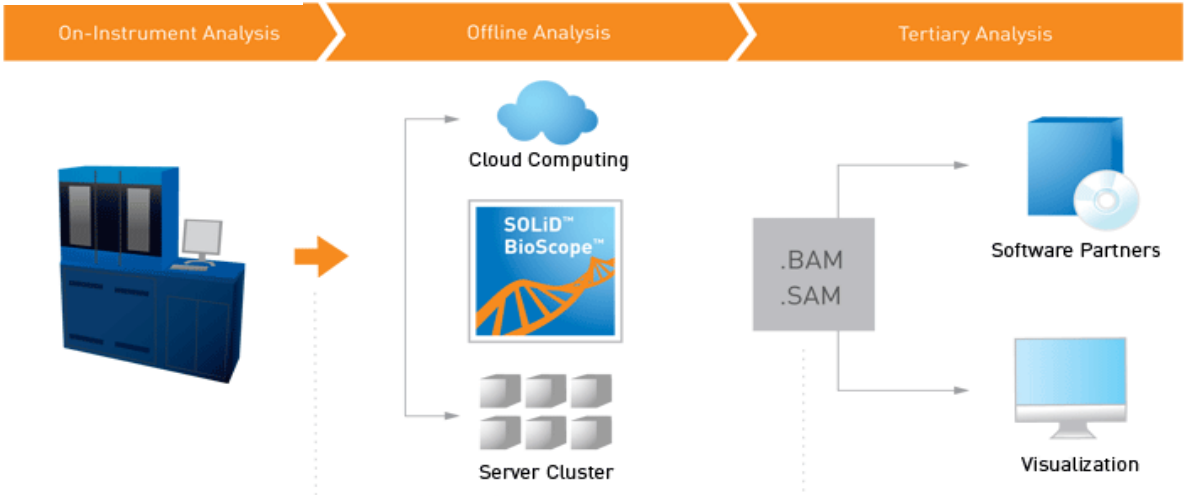
Coverage in AT rich regions



life technologies™

FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

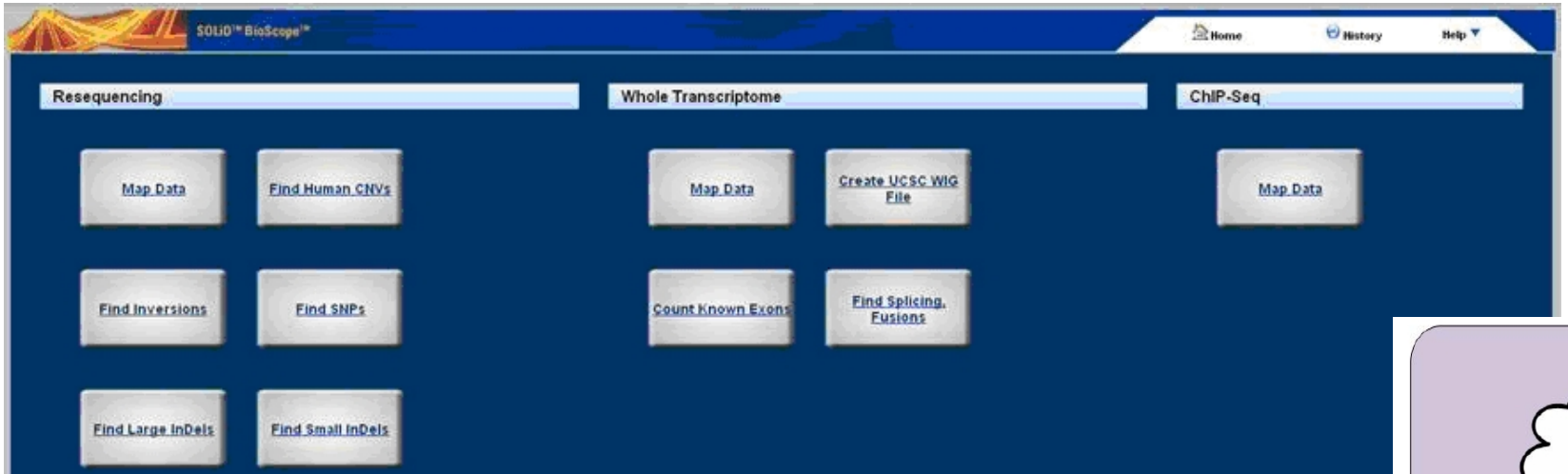
Data analysis workflow



	Primary Analysis	Secondary Analysis	Tertiary Analysis
Types of Analyses	• Image acquisition and bead processing • Quality metrics • Color calls	• Sequence Alignment • Sequence Stats • Consensus Calling • Create SAM/BAM file (base space) • Create QC file	• Visualization • Disease/gene annotations • Tag counting • Application-specific analysis ChIP Seq Whole transcriptome Resequencing De novo sequencing Methylation
Tools	• SOLiD™ Instrument Control Software (ICS) • SOLiD™ Experiment Tracking System (SETS)	• SOLiD™ BioScope™ Software • SOLiDBioScope.com™ • SOLiD™ Software Community Tools	

FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

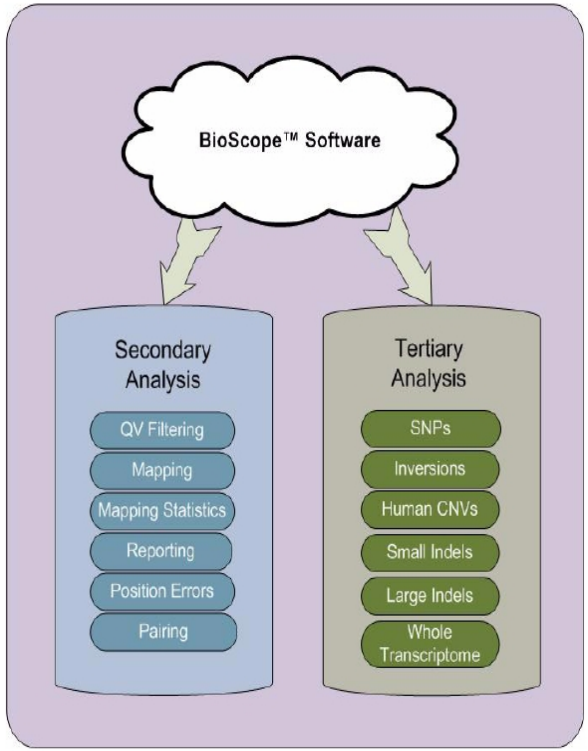
The SOLiD™ BioScope™ Software



• Command line and simple web interface for running application-specific sequence analysis tools.

• Resequencing (mapping, SNP finding (DiBayes), Human copy number variations, inversions, small indels, large indels)
• ChIP-Seq
• Transcriptome analysis (mapping, fusion/splicing, counting, UCSC WIG Files creation)

• Results in GFF v3 and BAM formats



FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

BioScope tools and file formats

Software or bioinformatics tool	Input file type(s)	Output file type
Mapping tool	*.csfasta, *.fasta, *qv	*.ma(local)
Pairing tool	*.ma(local) [,*.fasta], *.qual, *.csfasta	*.bam
MatoBAM tool	—	Converts a *.ma file to a *.bam file.
Small Indel	*.bam	*.gff.3
Frag indel	*.ma, *.ma(local)	*.pas
diBayes	*.bam	*.gff.3, *.csfasta, consensus_calls
CNV - singleSample	*.bam	*.gff.3
Large Indel -singleSample	*.bam	*.gff.3
Large Indel -pairedSample	—	—
Inversion	*.bam	*.gff.3, *.txt
Position error	*.bam	position error
WT mapping	*.csfasta, *fasta, filter reference fasta, WT *.gtf reference	*.bam
Counttag	*.bam	*.gtf
SAM2wig	*.bam	.wig

Key

[ ] = optional
+ = 1 or more
*.ma = classic match file
*.ma(local) = match file with local alignment extensions
*.gff.3 = public, viewer-oriented *.gff v 3
*.gff 0.2 = SOLiD™ *.gff version 2

*.gff 3.5 = SOLiD™ *.gff version for 3.5 release

FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

Helicos
BioSciences Corporation

HeliScope Single Molecule Sequencer

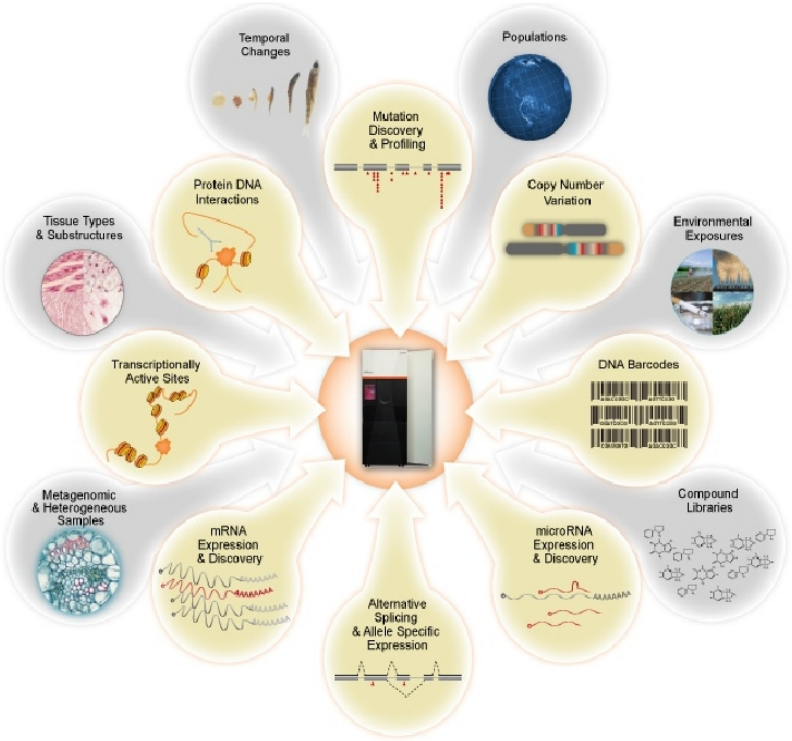
"the world's first DNA Microscope"

Allows direct measurement without amplification

FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

Applications

- **Whole genome resequencing**
- Targeted resequencing
- **Digital gene expression**
- **RNA-sequencing**
- Small RNA measurements
- **Copy number assessment**
- **Chromatin IP-sequencing**
- Methylation status



FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

Amplification-Based Sequencing Technologies

SOLiD

- Shear and Purify the DNA
- Repair DNA Ends and Purify Sample
- Ligate Adaptors to DNA
- Nick-Translation of the DNA
- Gel Size-Selection of the DNA
- Library Amplification
- Large-Scale Amplification /Purification of the Library
- Library Gel Purification
- Library QC and Quantitation

- *Emulsion PCR w/ ULTRA-TURRAX® Tube Drive from IKA®*
- Emulsion Break
- Templated Bead Wash
- Enrichment
- 3'-End Modification
- Bead Deposition

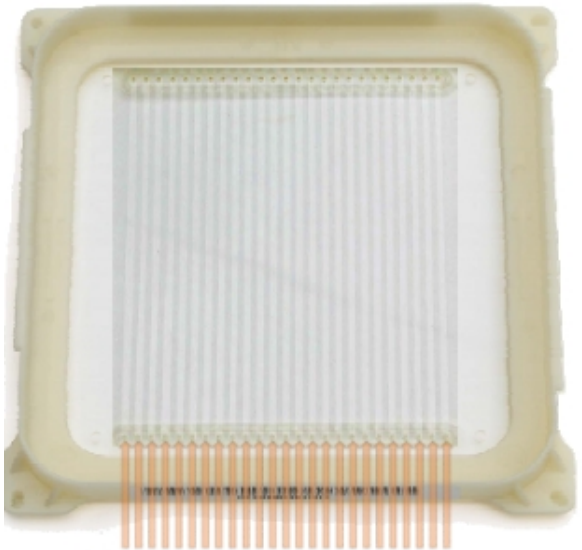
Illumina

- Fragment the Genomic DNA
- Perform End Repair
- Add 'A' Bases to the 3' End of Fragments
- Ligate Adapters to Fragments
- Gel Purify Ligation Products
- Enrich DNA Fragments by PCR
- Validate the Library
- Flow Cell Loading
- Cluster Amplification

Helicos BioSciences Corporation

- Fragment the Genomic DNA
- Size Selection of DNA and QC
- PolyA Tailing of the DNA and QC
- 3' Blocking
- Clean-up and Quantitation
- Flow Cell Loading

Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi



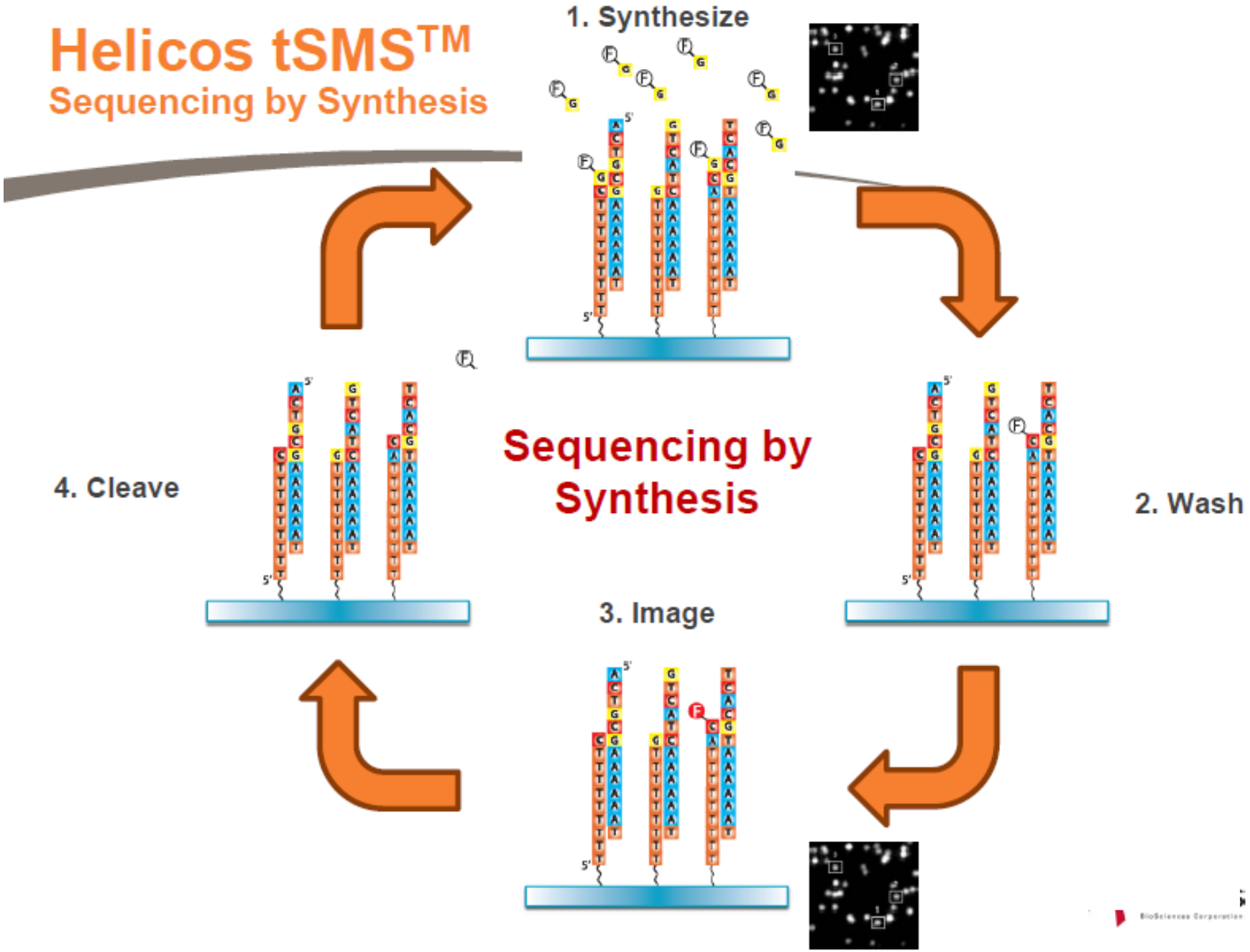
25 Channels – 2 flow cells per run

- 48 samples per run
  (one channel per flow cell for quality control)
- Universal for all applications

Strand output	12-20 million usable strands per channel 600mm – 1bn usable strands per run
Total output	21-33+ Gb/run (up to 10x Human Genome)
Read length	25 to 55 bases in length Median 33-36 bases
Accuracy	>99.995% concensus accuracy at >20x coverage

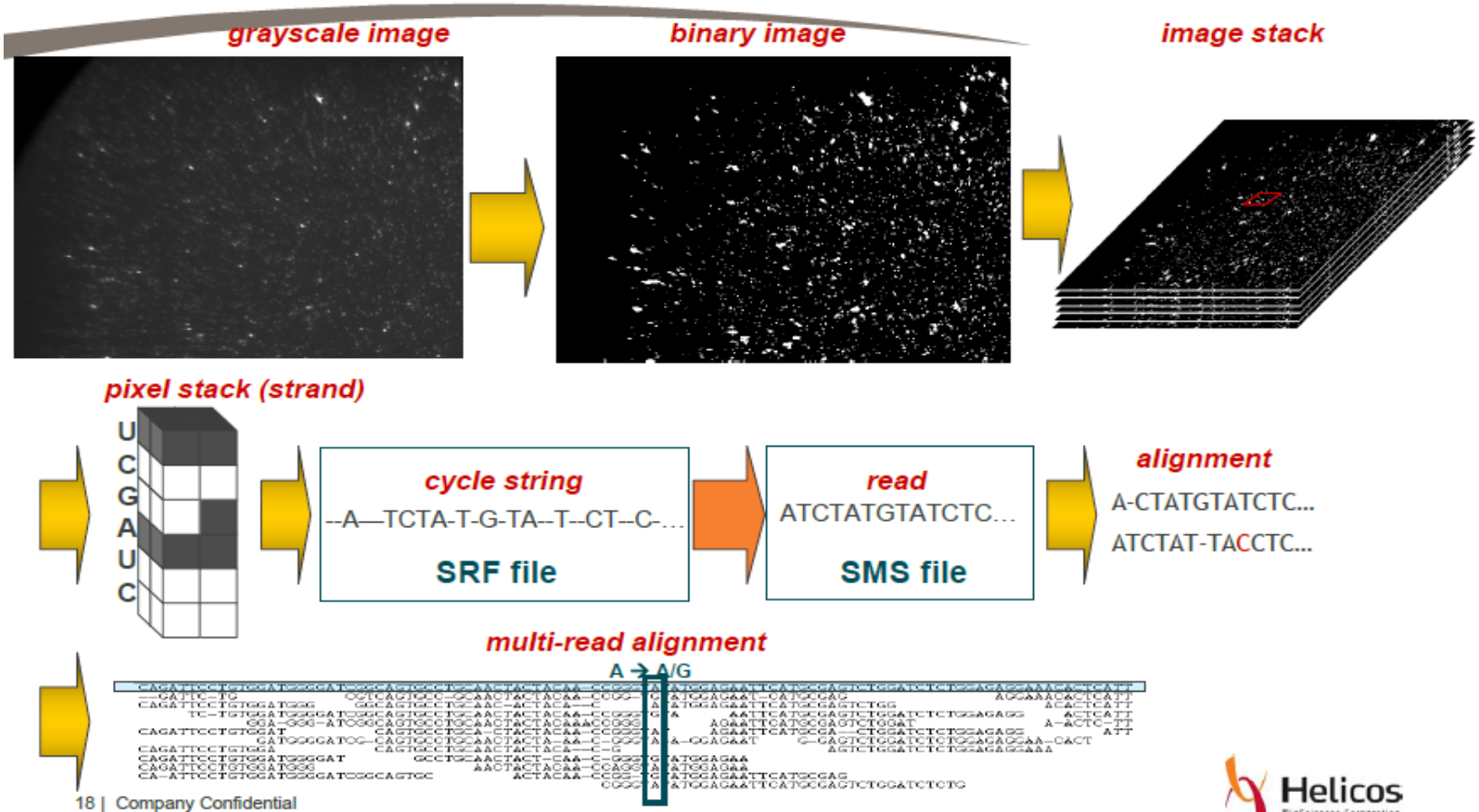
FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

Helicos tSMS™
Sequencing by Synthesis



FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

Data Structures



Helicos BioSciences Corporation

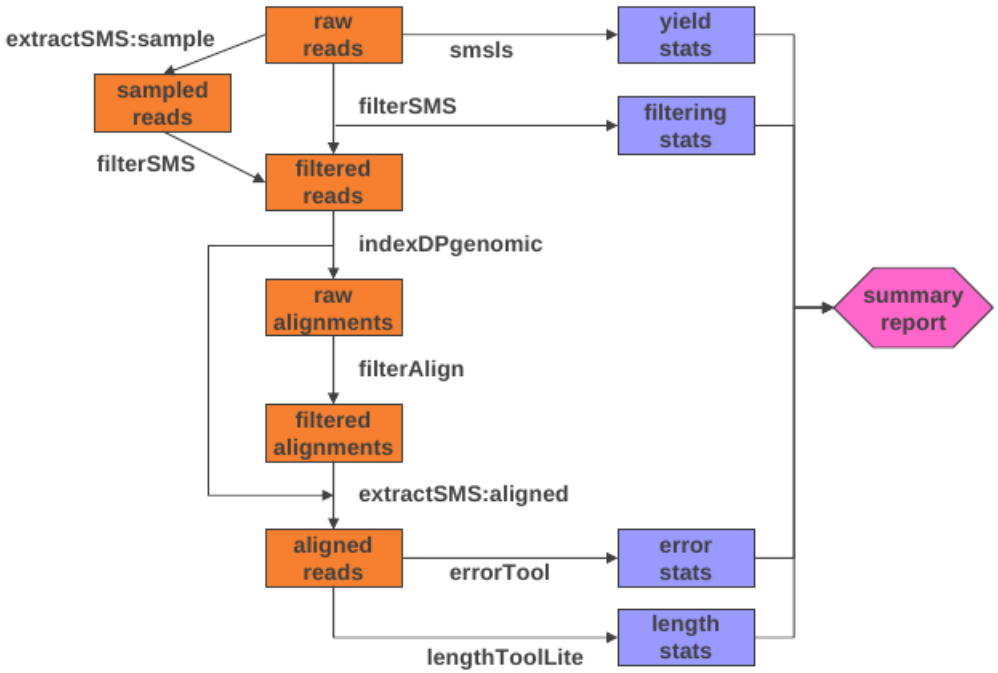
SRF File:
Sequences
Cycle strings
Strand Coordinates
Metadata:
Instrument configurations
Image registration coords
200 GB/run raw
80 GB/compressed

SMS Files:
Sequences
Coords
Cycle strings
100 GB/run raw
60 GB compressed

FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

HeliSphere data analysis software

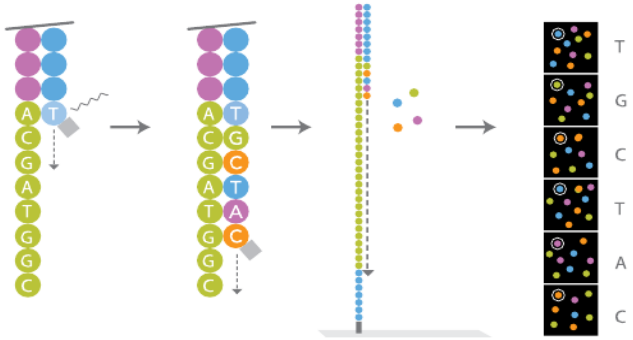
- An open-source LINUX software package
- Consists of a set of core tools and analysis pipelines.
- The core tools are command line tools that perform unit functions such as
  - alignment of reads to a reference
  - filtering of reads or alignments
  - report generation
  - file format conversions
  - etc.
- The pipelines organize sets of core tools to carry out a complex task, such as resequencing or digital gene expression analysis. The pipelines also provide managed parallel processing.



FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

Illumina Genome Analyzer IIx

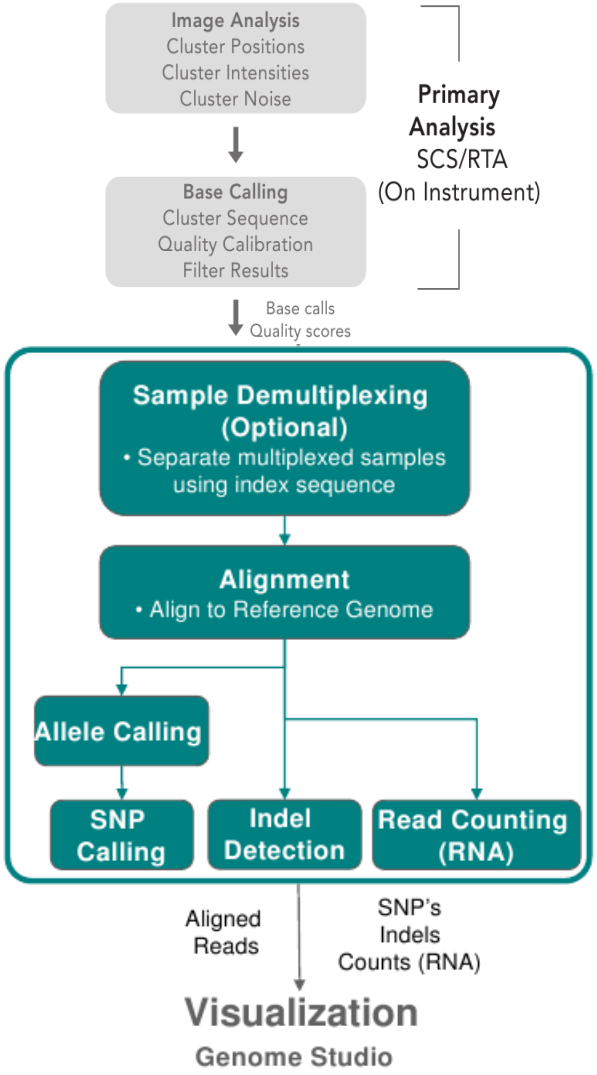
illumina®



- Illumina's sequencing by synthesis chemistry
- **8 lanes** per flow cell = 8 samples
- **Sample multiplexing** up to 12 samples per channel
- **Read length** 1x36bp - 2x100 bp on a paired-end flow cell
- **Read accuracy** greater than 98 %
  (per base raw accuracy averaged over 100 cycles per read)
- 30-40 million paired-read tags per channel
  (passing filtering with two or less mismatches)
- Up to **33 gigabases** of sequence data produced per paired end flow cell (> 70% of bases with Q>=30, 2*100bp read, allowing for filtered reads with two of fewer errors)

FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

Illumina data analysis



- **Primary analysis** software aligns reads to a reference sequence
- **CASAVA** software for SNP calls and RNA counts etc.
- **GenomeStudio** software for application specific data analysis:
  - DNA resequencing
  - ChIP sequencing
  - mRNA sequencing and transcriptome profiling
  - Genotyping
  - etc.

http://fmsc.btk.fi/

ngs@btk.fi

The Finnish Microarray and Sequencing Centre
Turku Centre for Biotechnology
Tykistökatu 6, Biocity
20520 Turku, Finland

FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

SOLiD files and file formats

Table 1. Average file sizes for various analyses.

	Image Data Size*	Primary Analysis Data†	Secondary Analysis Results in BAM Format‡
1 slide—tag (fragment 50 bp)	1.9 TB	0.8 TB	0.6 TB

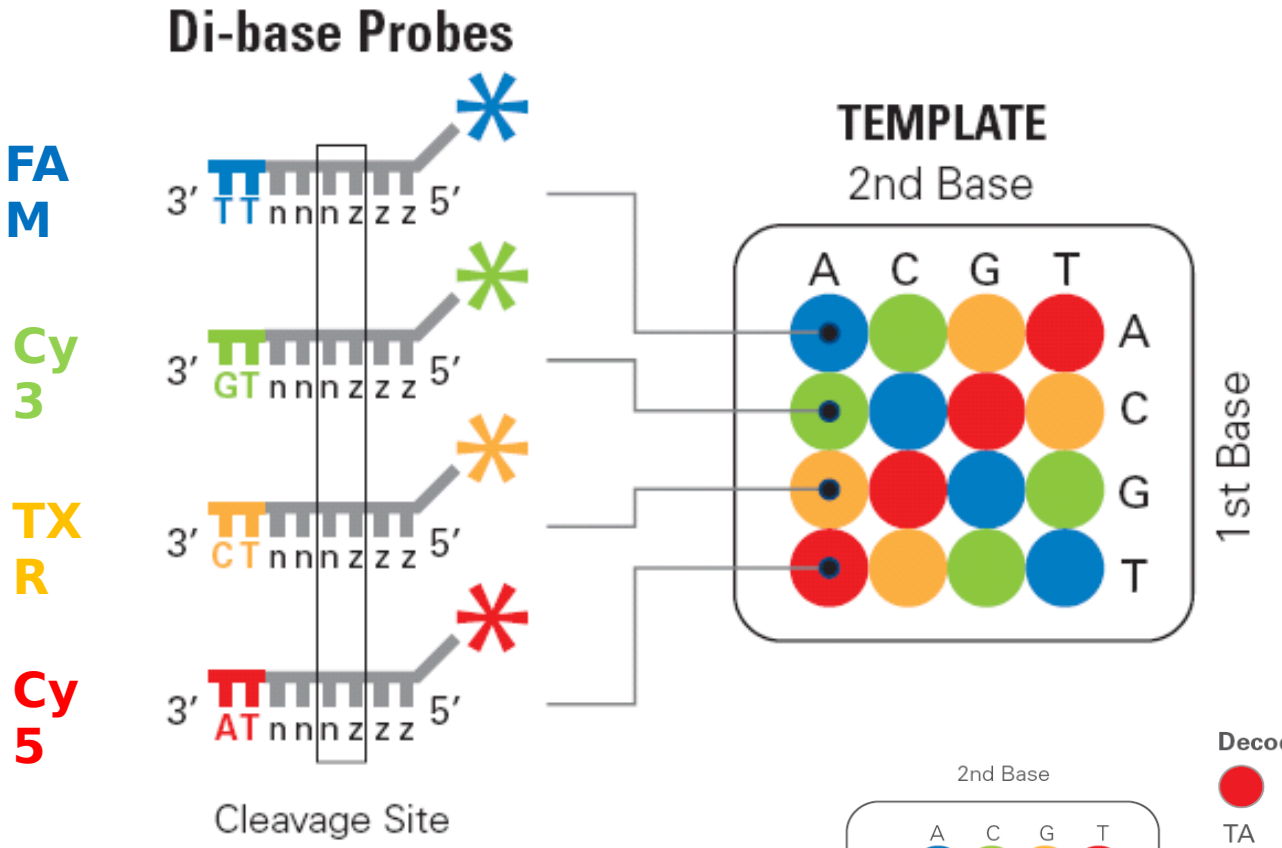
Average file sizes for various analyses under the following assumptions: 10 ligation cycles for each sequencing primer, 4 images per cycle, 1 for each dye.
A full slide contains more than 2,350 panels.
* Minimum space needed for images.
† Minimum space needed for primary analysis results (spch, csfasta, QV.qual).
‡ The size of the analysis result correlates directly with the throughput.

Table 2. Data output files.

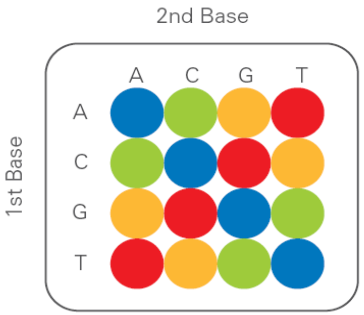
	File Type/Format	File Name Extension	File Content
Primary Analysis Files	Raw reads file	.csfasta	Color space reads
	QV quality value file	–QV.qual	Quality value for each color space sequenced
	Reads summary file	.stats	Statistics summarizing the number of reads collected in each panel on a slide
	Scaled Intensity Values File (optional)	–intensity.scaled [CY3|CY3|CY5|FTC|TXR].fasta	Color space reads
Secondary Analysis Files	Mapping file	.csfasta.ma	Sequence data mapped back to the reference sequence with quality values
	BAM	.bam	BAM (Binary Alignment Map) format is a generic format for storing large numbers of nucleotide sequence alignments

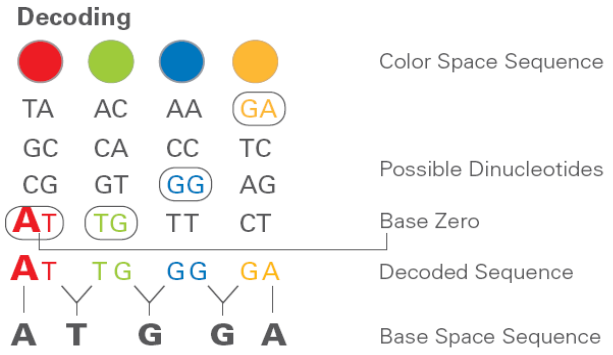
FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

SOLiD color space sequencing cont.



- Octamer probes
- 4 dyes, 4 dinucleotides

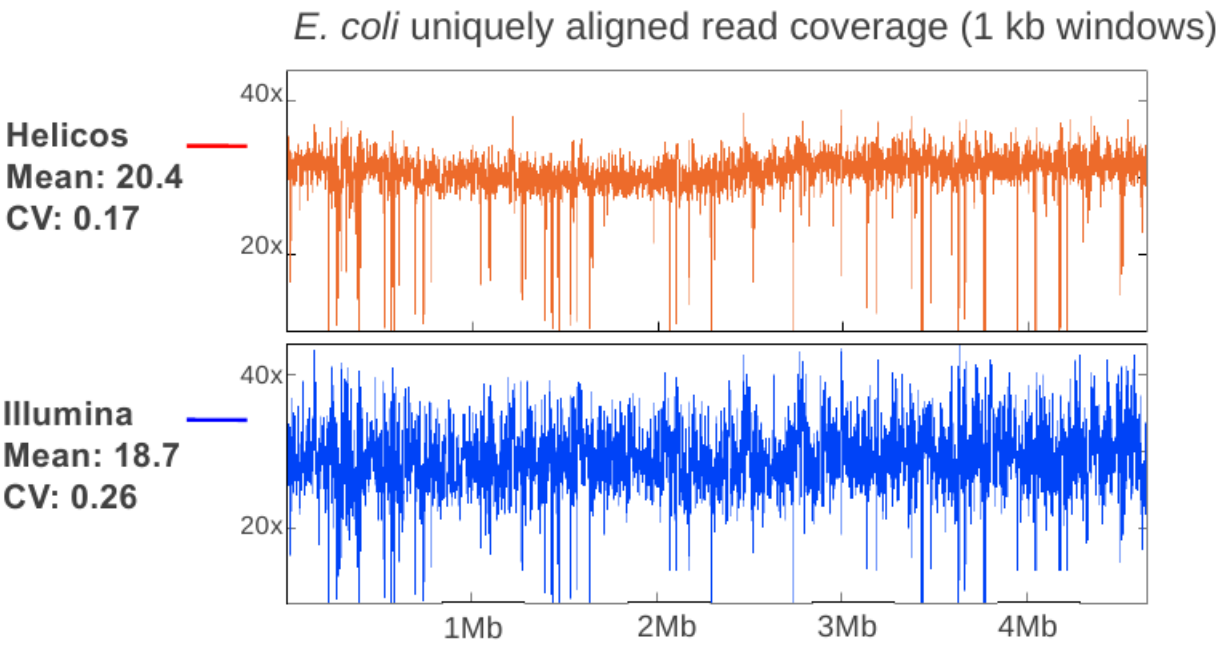




FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

Helicos data comparison

Even Coverage of the *E. coli* Genome



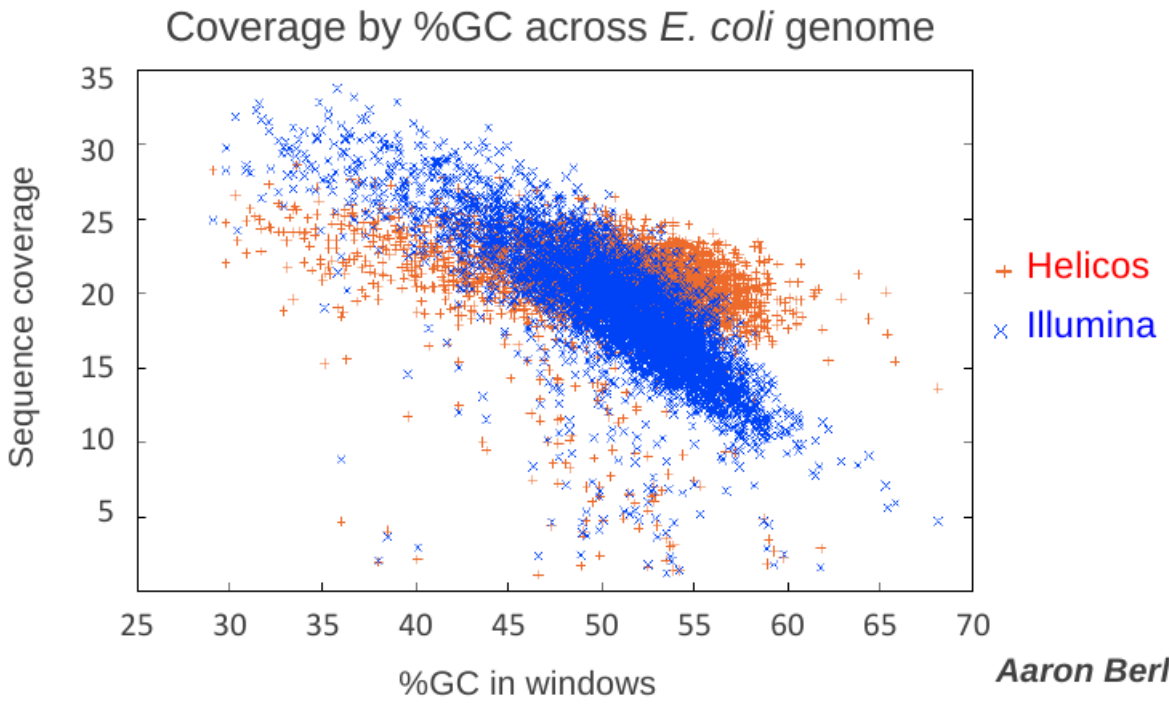
Identified 5 Variants from reference sequence – all five were true variants

Aaron Berlin

BROAD INSTITUTE

FMSC
The Finnish Microarray and Sequencing Centre
Biocity, 5th floor • Tykistökatu 6, 20520 Turku, FINLAND
Tel: +358 2 333 8634 • E-mail: fmsc@btk.fi • http://fmsc.btk.fi

Even representation by base composition



Aaron Berlin

BROAD INSTITUTE