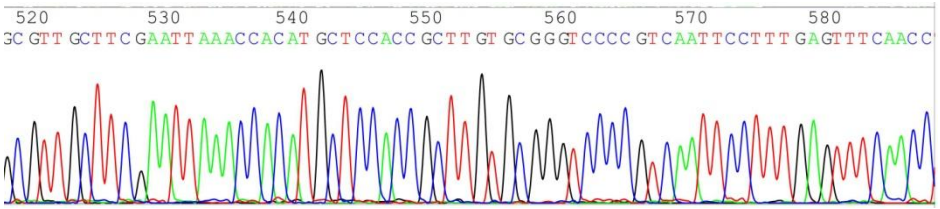
“SOLiD technology for high-througput sequencing”

Gennady V. Vasiliev

Institute of Cytology and Genetics SB RAS

First Generation of sequencers – Sanger based

96-capillary ABI 3730XL



Second Generation of sequencers – MPSS

High throughput models

	454-FLX Titanium	Illumina/ Solexa	SOLiD 4
Read length (bp)	240–400	2x100	2 x 50
Human genome resequencing cost ($)	1 000 000	20 000	5 000 - 10 000

Medium throughput models

454 Junior	Ion Torrent
240–400	100-200
N/A	N/A

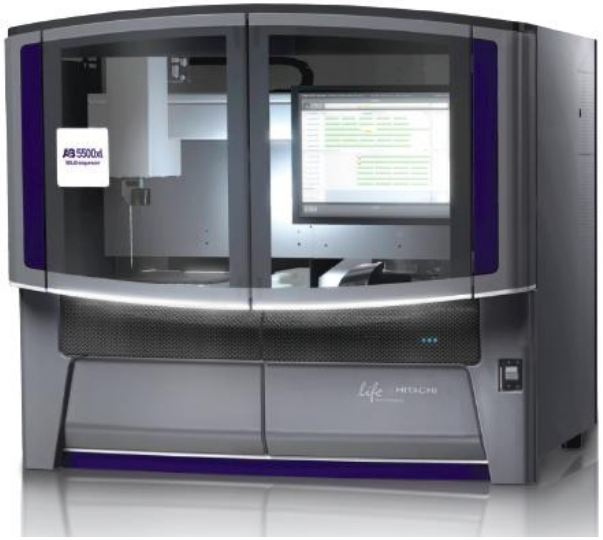
Third Generation of sequencers – Single molecular

	Helicos tSMS	PacBio SMRT	Complete Genomics
Read length (bp)	30	Up to 1 000	2 x 35
Human genome resequencing cost ($)	70 000	Less than 10 000	5000

SOLiD

SOLiD 4

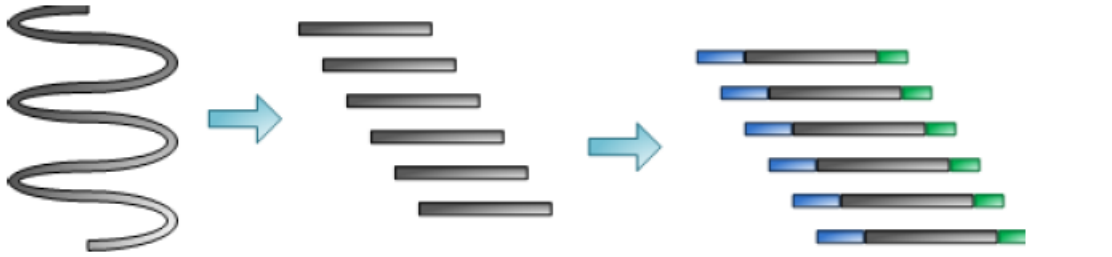
SOLiD 5500



SOLiD 4 and SOLiD 5

	SOLiD 4	SOLiD 4hg	SOLiD 5500	SOLiD 5500xl
Throughput per run	Up to 100 GB (1 hg, 30x)	Up to 300 GB (3 hg, 30x)	Up to 90 GB (1 hg, 30x)	Up to 180 GB (2 hg, 30x)
Samples number	Up to 8 per slide, 2 slide	Up to 4 per slide, 2 slide	1–6 (1 FlowChip)	1–12 (2 FlowChips)
Multiplexing	96 DNA and 48 RNA barcodes	-	96 barcodes for RNA and DNA	96 barcodes for RNA and DNA
Maximum read lengths	2 x 50 bp (Mate-Paired) 50 bp x 25 bp (Paired-End)	2 x 50 bp (Mate-Paired) 50 bp x 25 bp (Paired-End)	2 x 60 bp (Mate-Paired) 75 bp x 35 bp (Paired-End)	2 x 60 bp (Mate-Paired) 75 bp x 35 bp (Paired-End)

Step 1: Library preparation

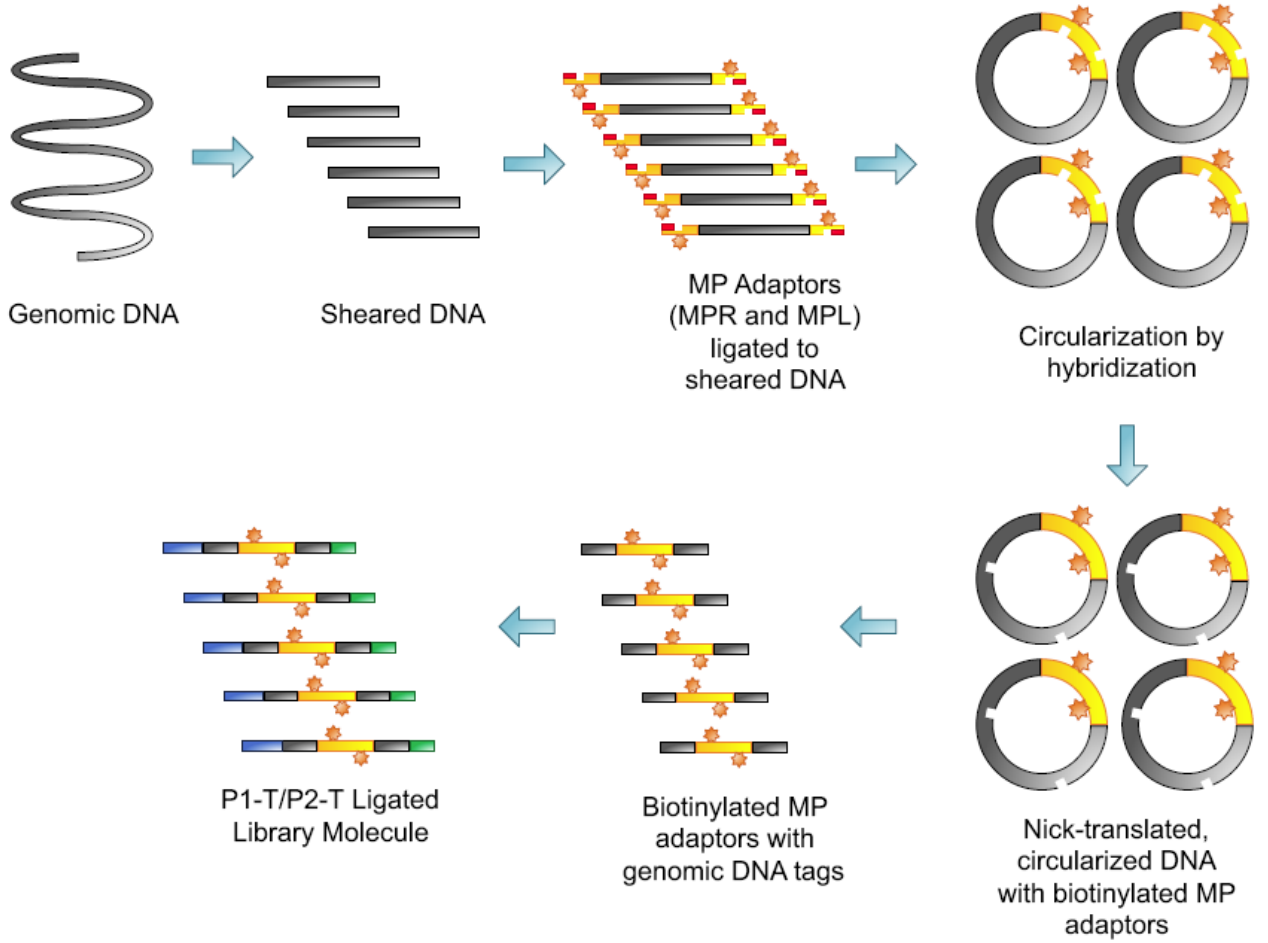


Genomic DNA | Sheared DNA | Library molecule ligated with P1-T and barcoded adaptors

Fragment Library

Mate-Paired Library

Basic 2 × 60 bp mate-paired library preparation workflow



T7 Exonuclease and S1 Nuclease digestion

Major types of SOLiD sequencing runs

Fragment sequencing

Paired-end sequencing

Mate-pair sequencing

Multiplex fragment sequencing

Multiplex paired-end sequencing

mRNA preparation for transcriptome experiment

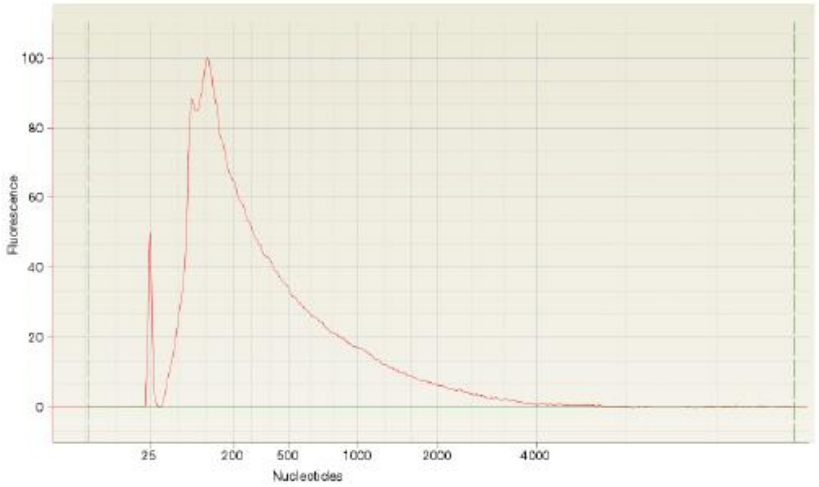


Figure 1 Size distribution of fragmented HeLa poly(A) RNA

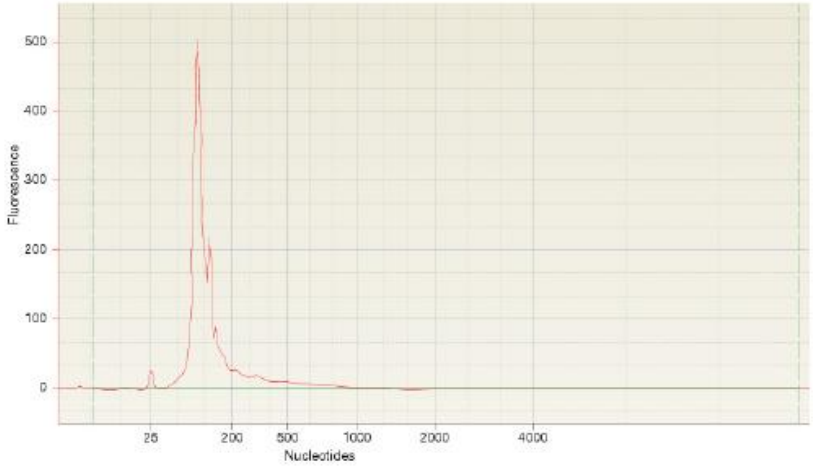


Figure 2 Size distribution of fragmented rRNA-depleted HeLa RNA

RiboMinus™ Concentration Module (Invitrogen)

PureLink™ RNA Micro Kit (Invitrogen)

Library preparation for transcriptome experiment

Fragmentation of whole transcriptome RNA

0.5–1 µg poly(A) RNA or 1 µg total RNA or rRNA-depleted total RNA

↓

Fragment the RNA

↓

Clean up the RNA

↓

Assess the yield and size distribution of the fragmented RNA

↓

Fragmented RNA

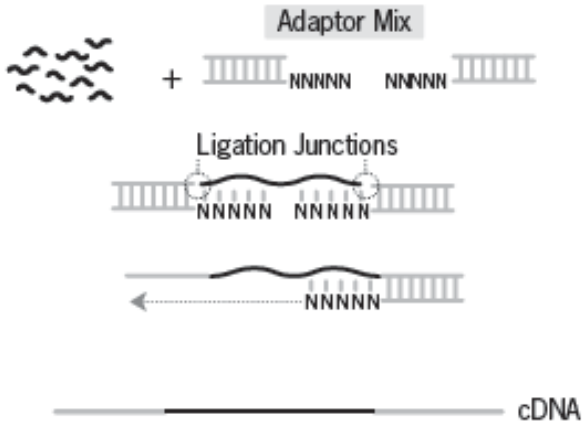
Library preparation for transcriptome experiment

Amplified library construction

Hybridize and ligate the RNA

Perform reverse transcription

Purify the cDNA



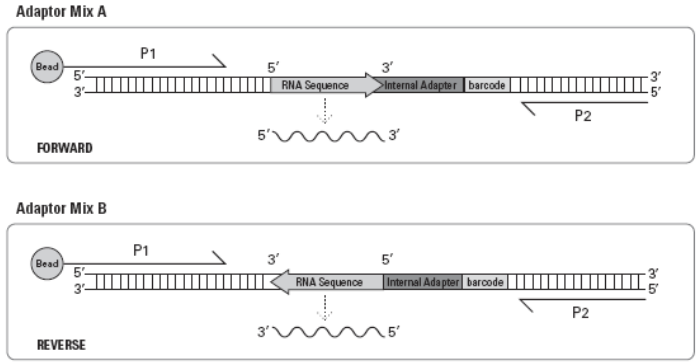


Figure 3 Adaptor Mix choice and RNA sequence

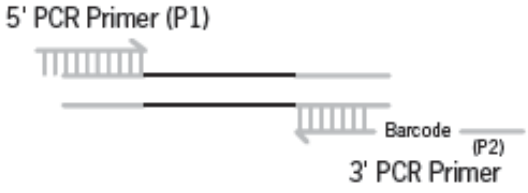
Library preparation for transcriptome experiment

Size select the cDNA

Amplify the cDNA

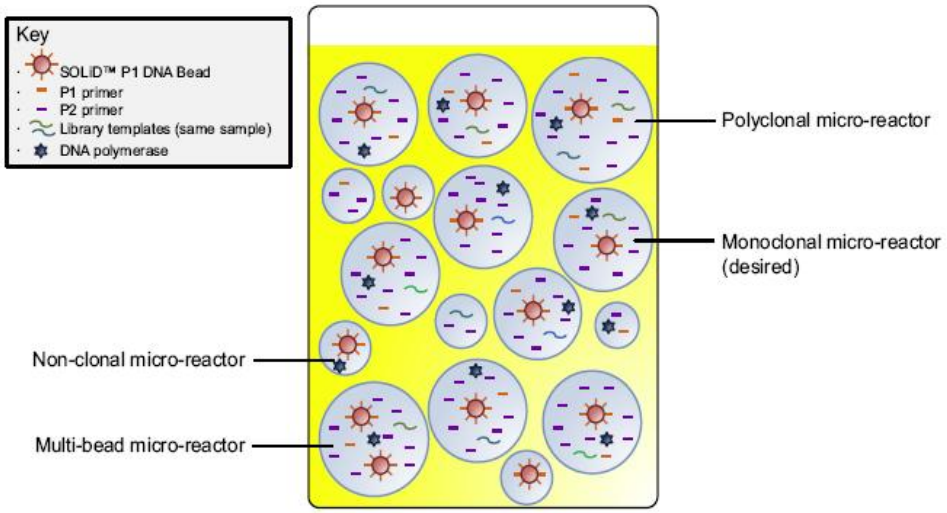
Purify the amplified DNA

Assess the yield and size distribution of the amplified DNA

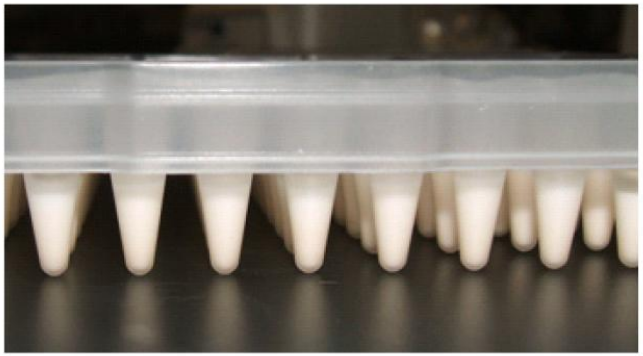


SOLiD™ System templated bead preparation and sequencing

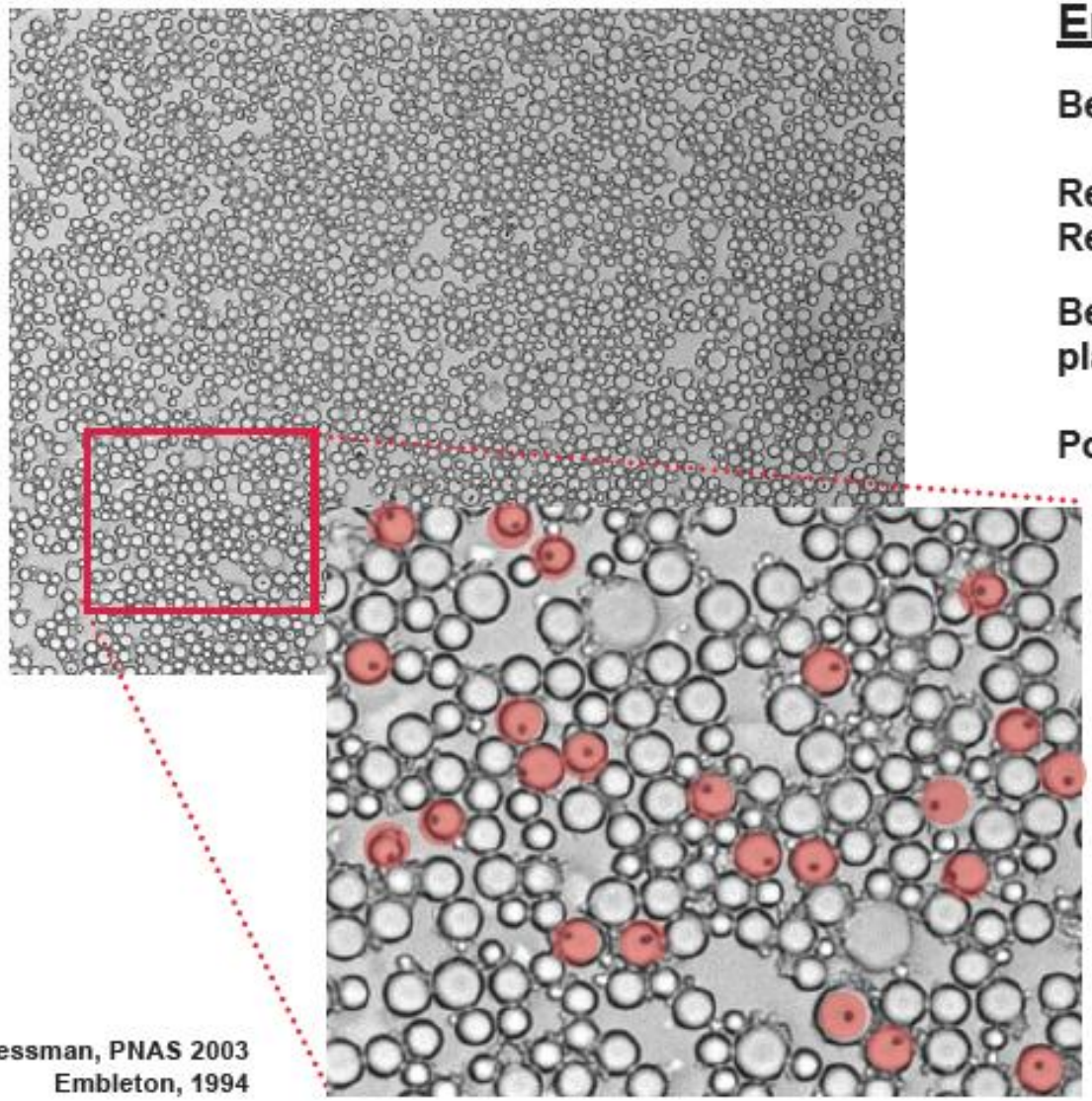
Step 2: E-PCR



Emulsion before amplification (ePCR).



SOLiD: результат ПЦР в эмульсии



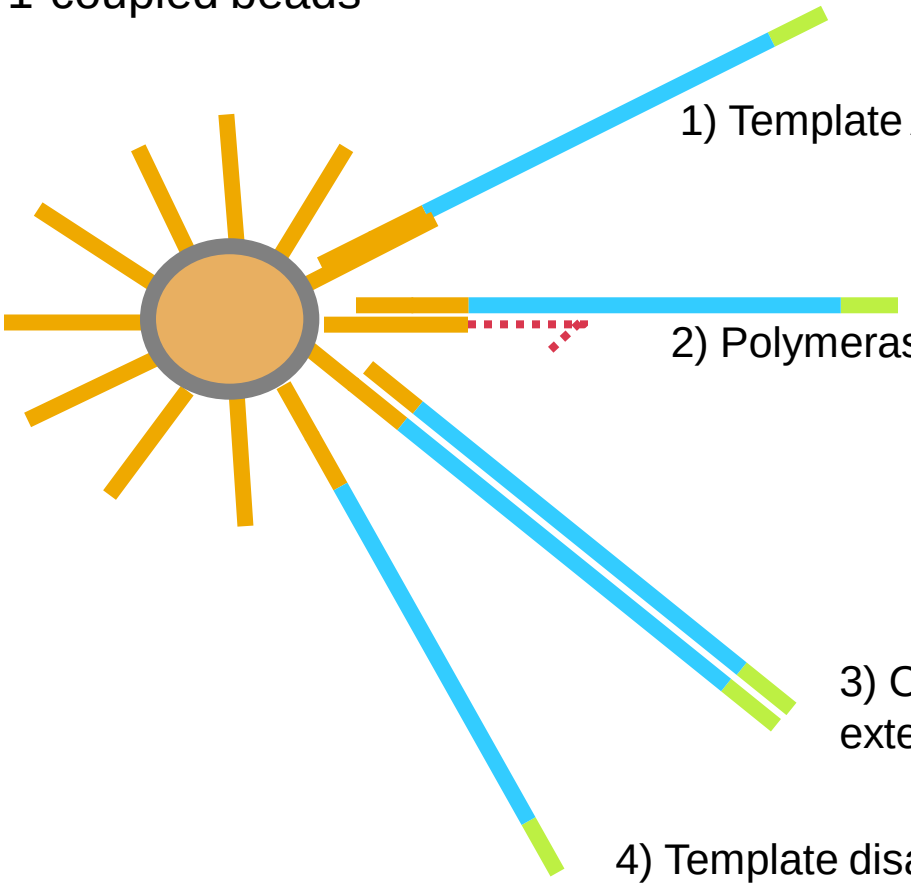
Dressman, PNAS 2003
Embleton, 1994

Emulsion Metrics

Bead size:	1 μm
Reactor size:	4 μm
Reactor volume:	34 fL
Beads / emulsion plate (96-well):	1.6×10^9
Post Enrichment:	$150 – 300 \times 10^9$

E-PCR

P1-coupled beads



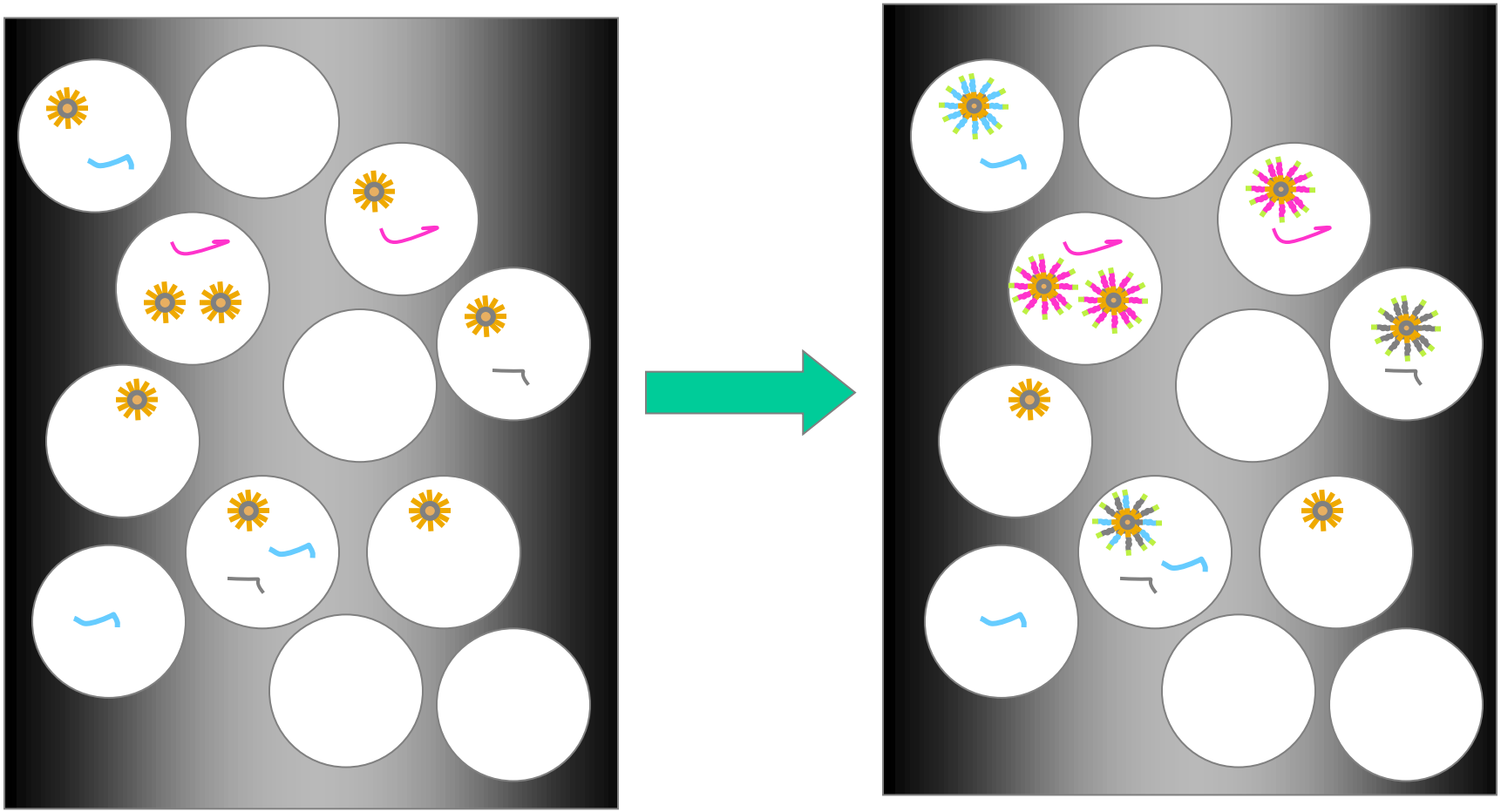
1) Template Anneals to P1

2) Polymerase extends from P1

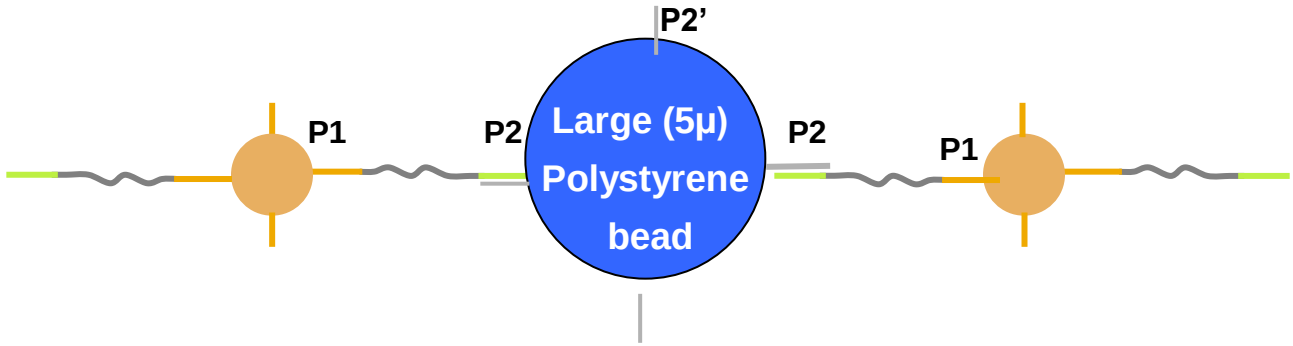
3) Complementary sequence is extended off bead surface

4) Template disassociates

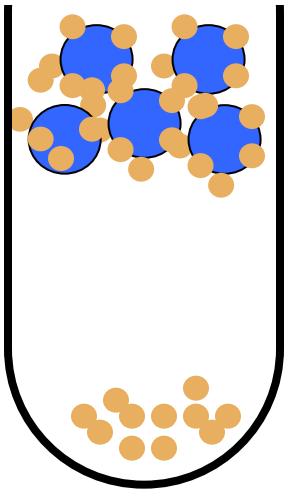
E-PCR



Step 3: Bead enrichment

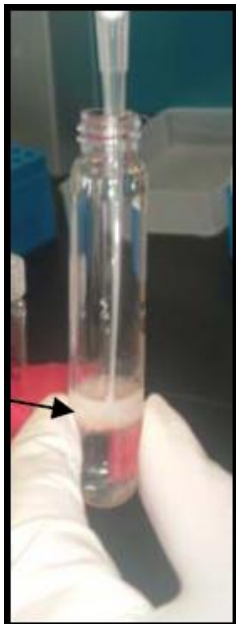


Centrifuge in glycerol gradient

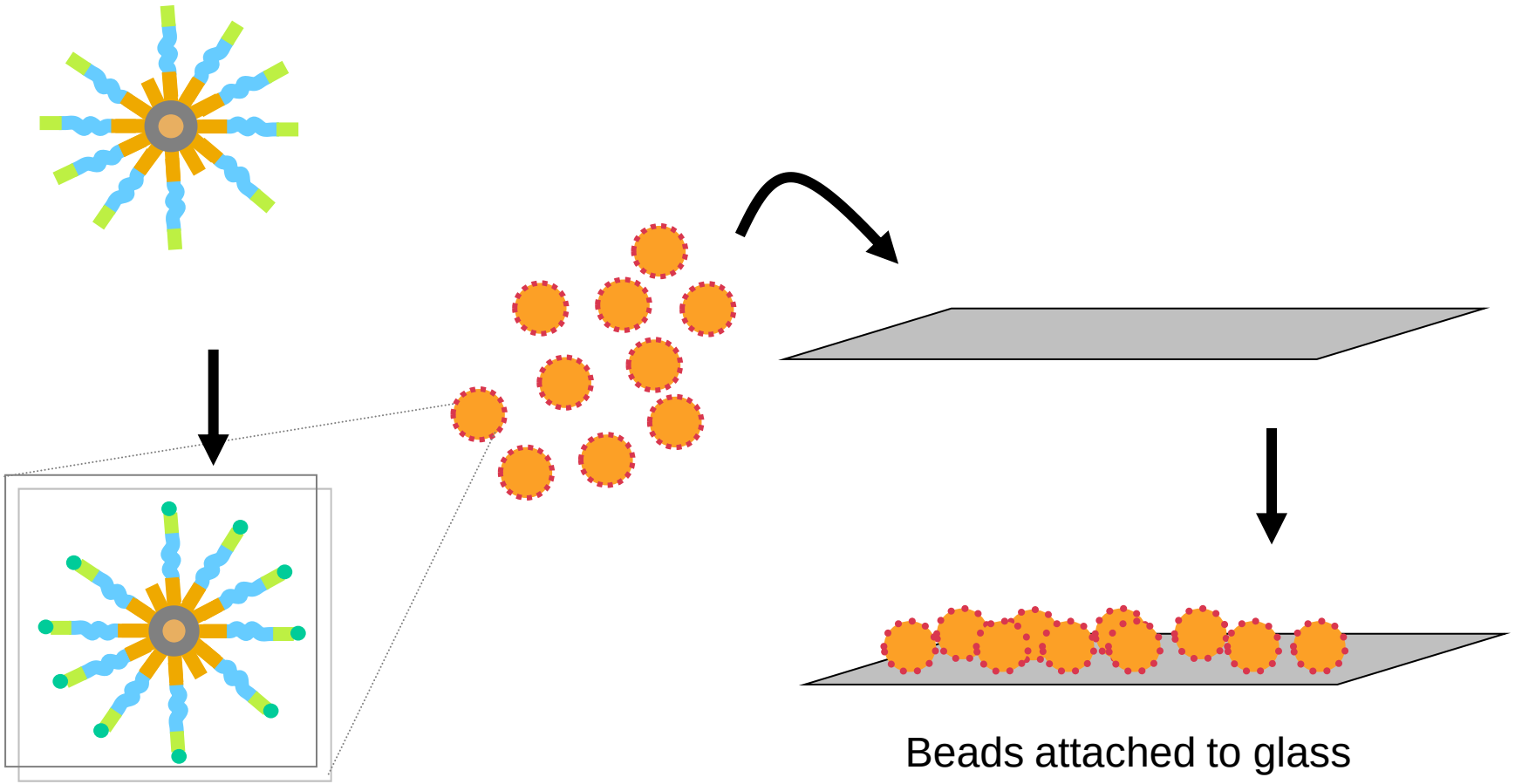


Supernatant
Captured beads with templates

Pellet
Beads with no template

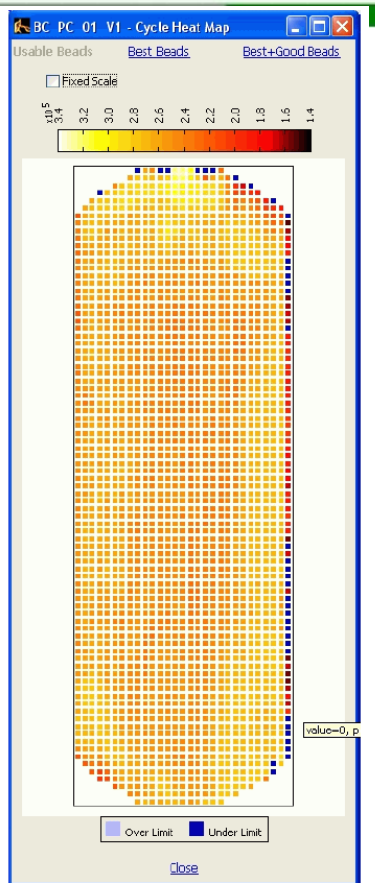


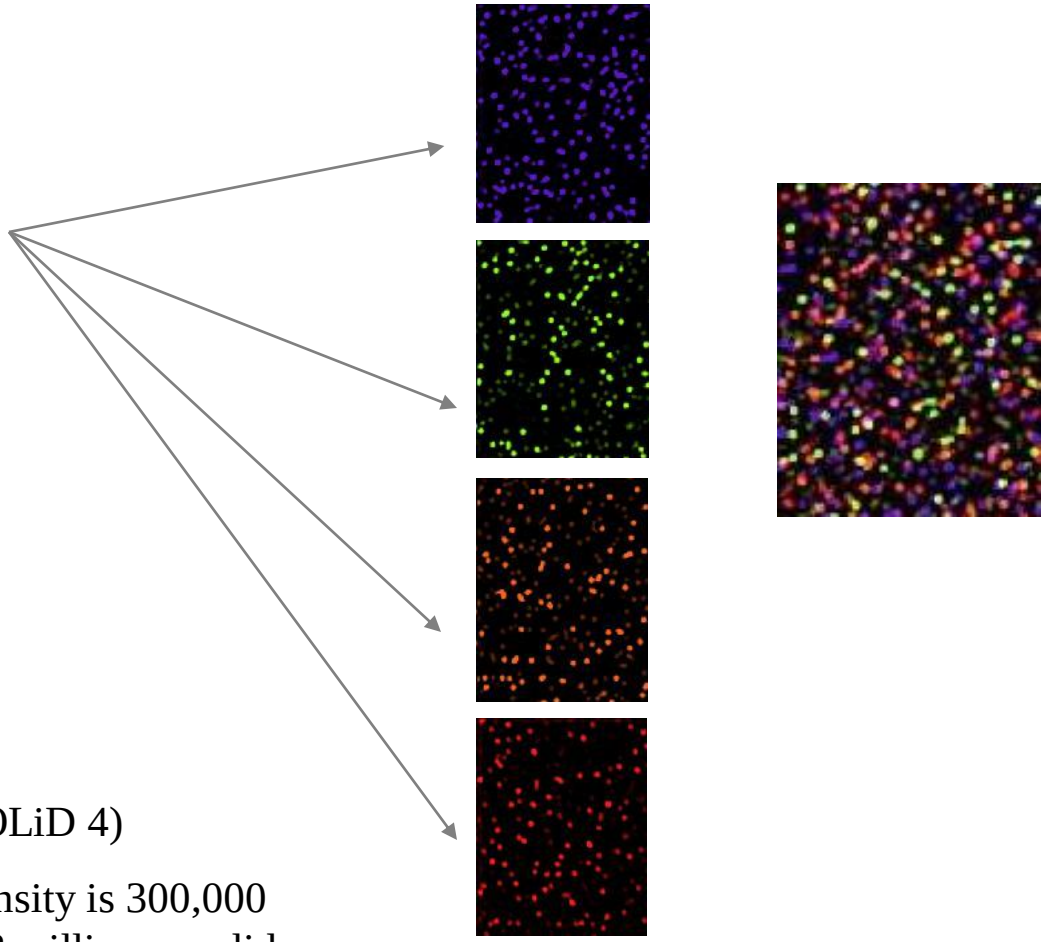
Step 4: 3′-end modification and beads deposition



Beads attached to glass surface in a **random** array

SOLiD





2357 panel per slide (SOLiD 4)

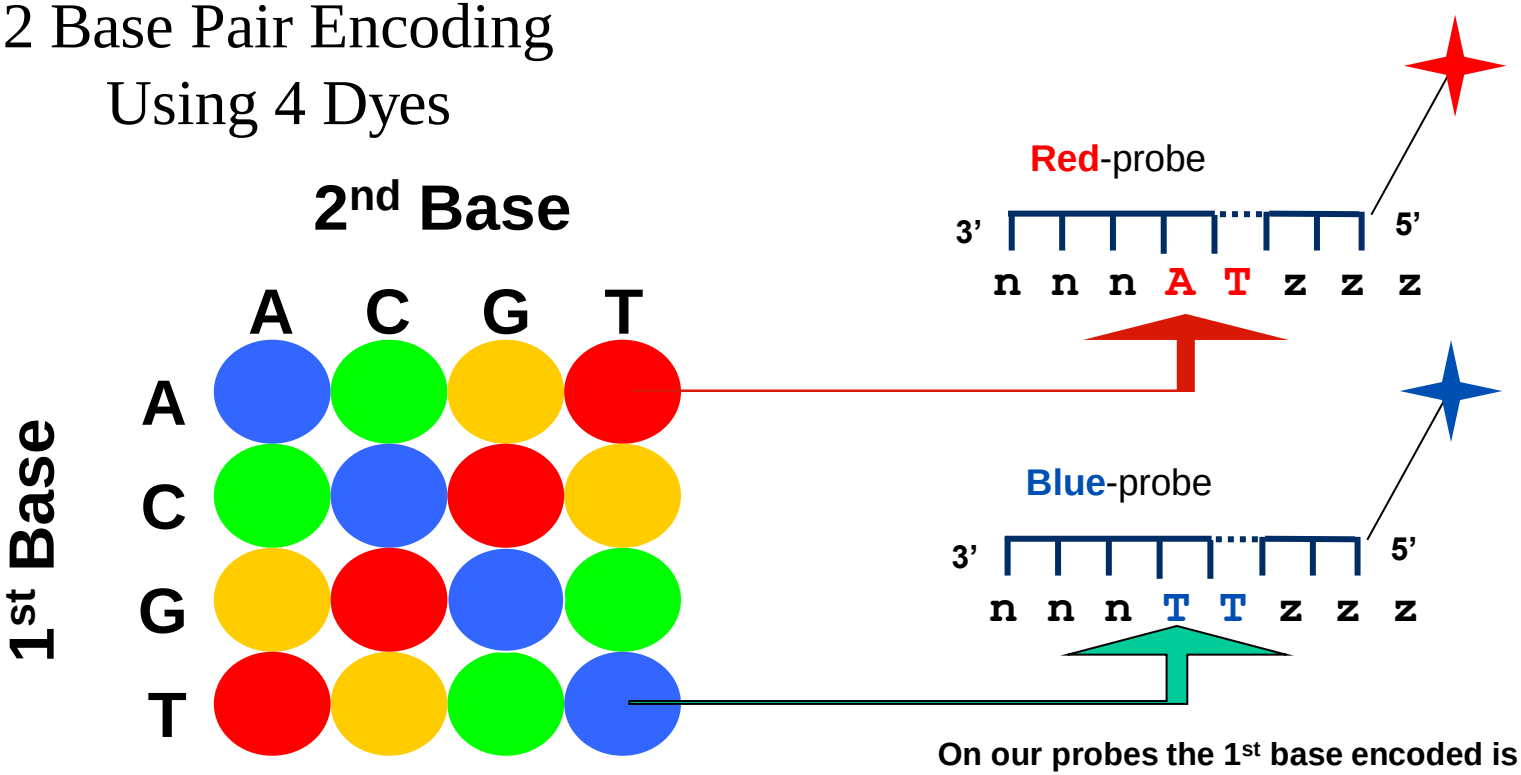
The targeted bead deposition density is 300,000 P2-positive beads per panel, 708 million per slide

SOLiD: Probe

1,024 Octamer Probes (4^5)

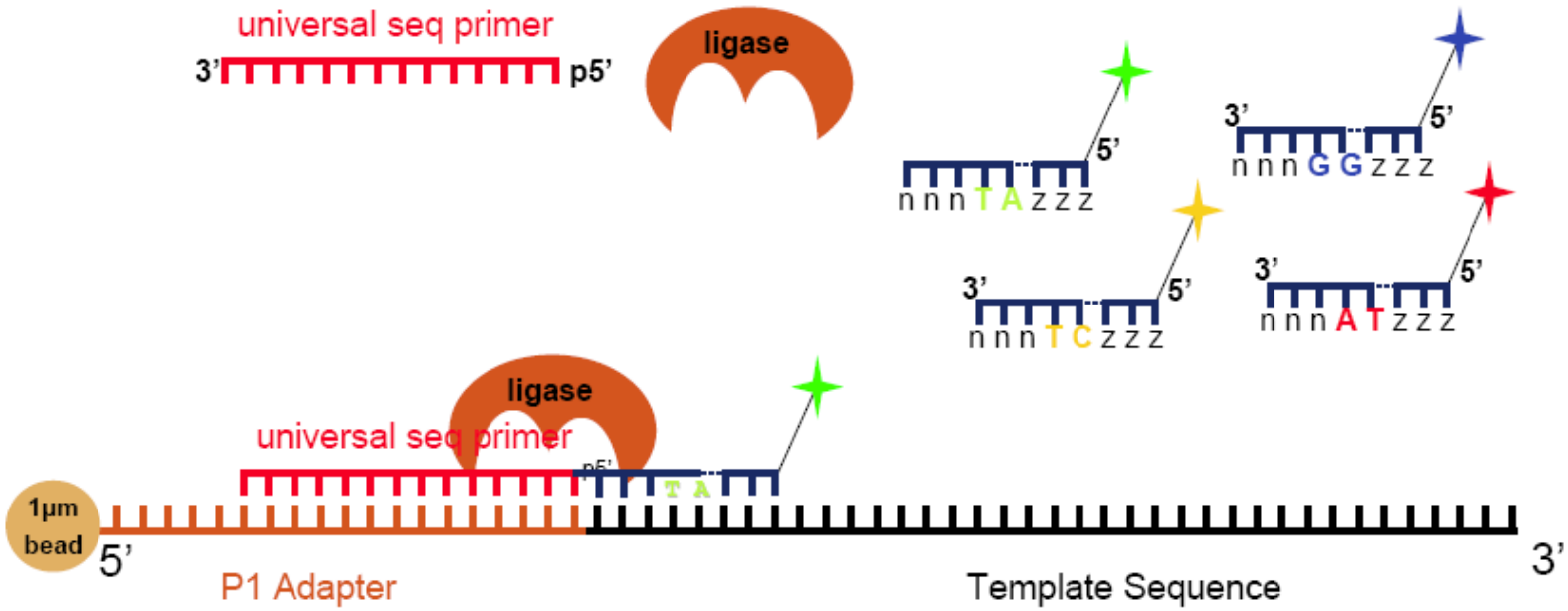
4 Dyes, 4 dinucleotides, 256 probes per dye

2 Base Pair Encoding
Using 4 Dyes

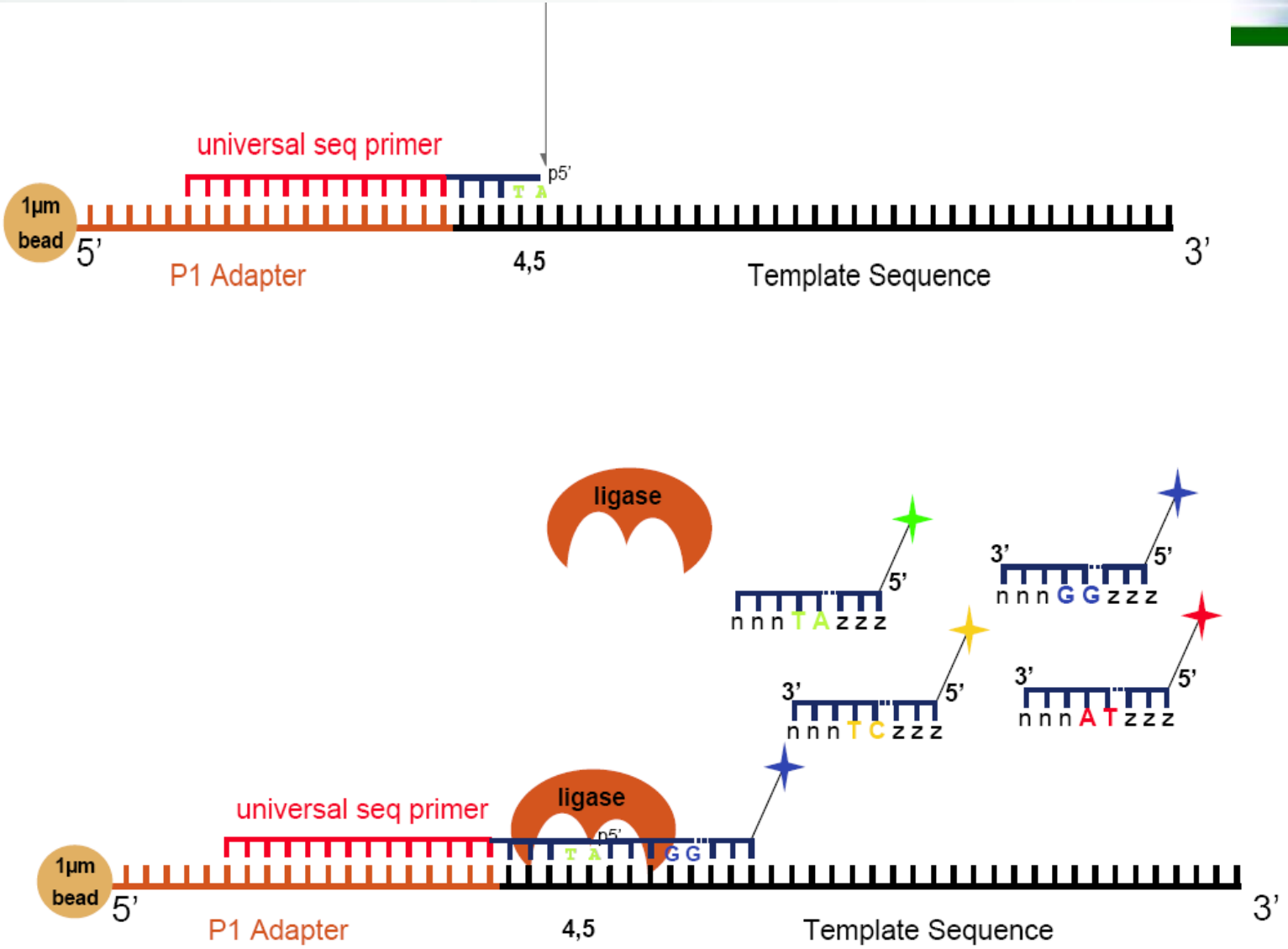


On our probes the 1st base encoded is position 4
the 2nd base encoded is position 5

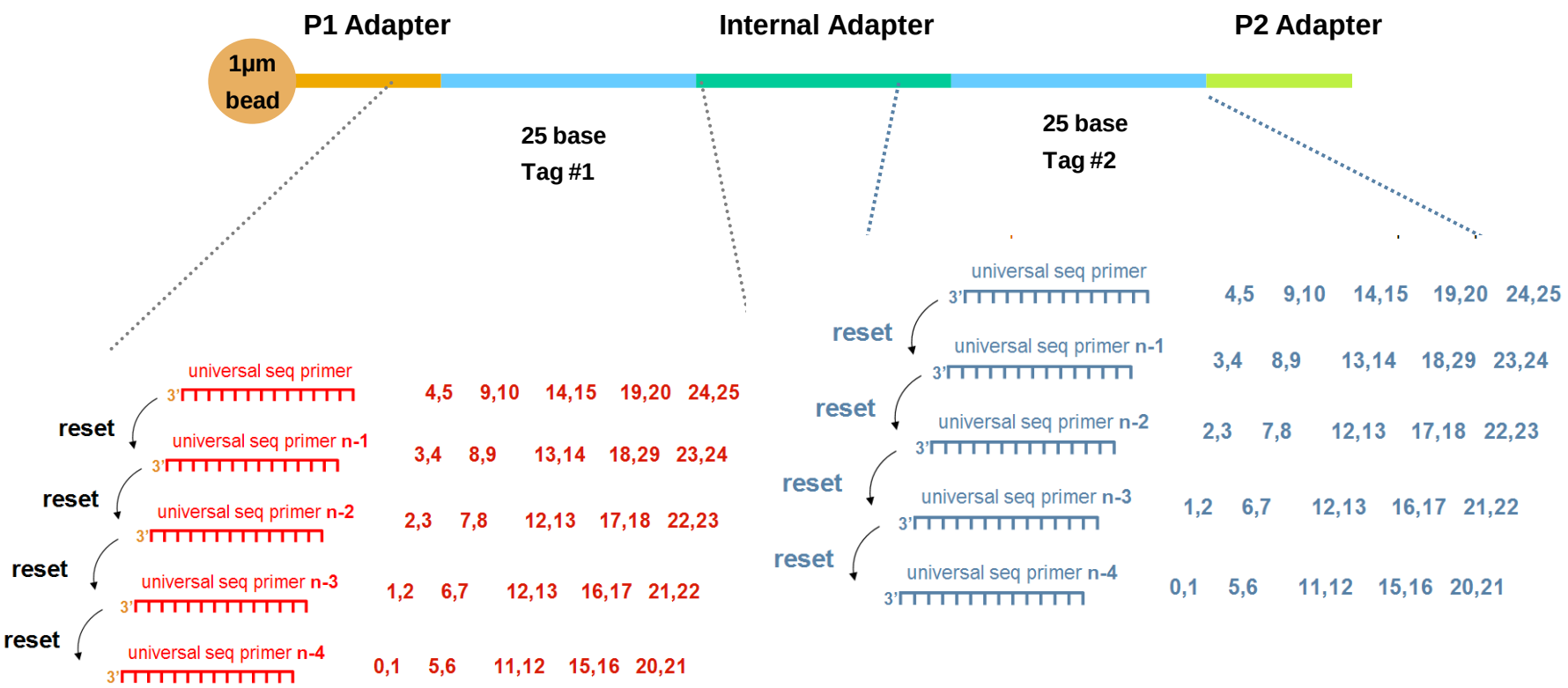
SOLiD



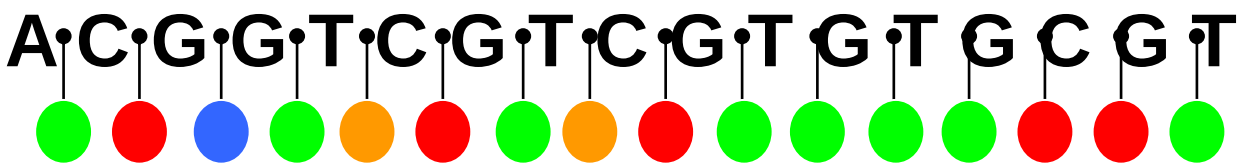
SOLiD

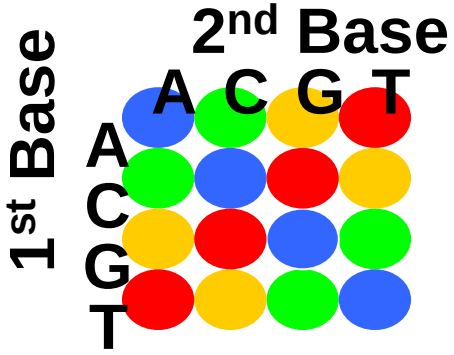


SOLiD

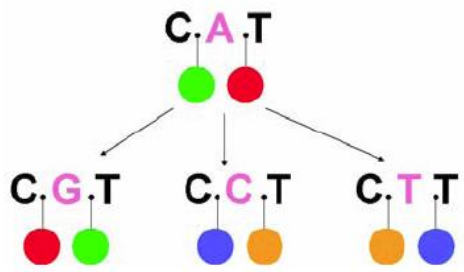


SOLiD read

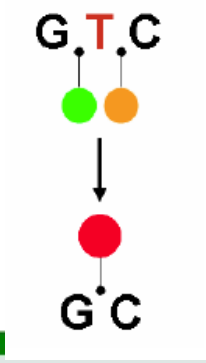




SNP

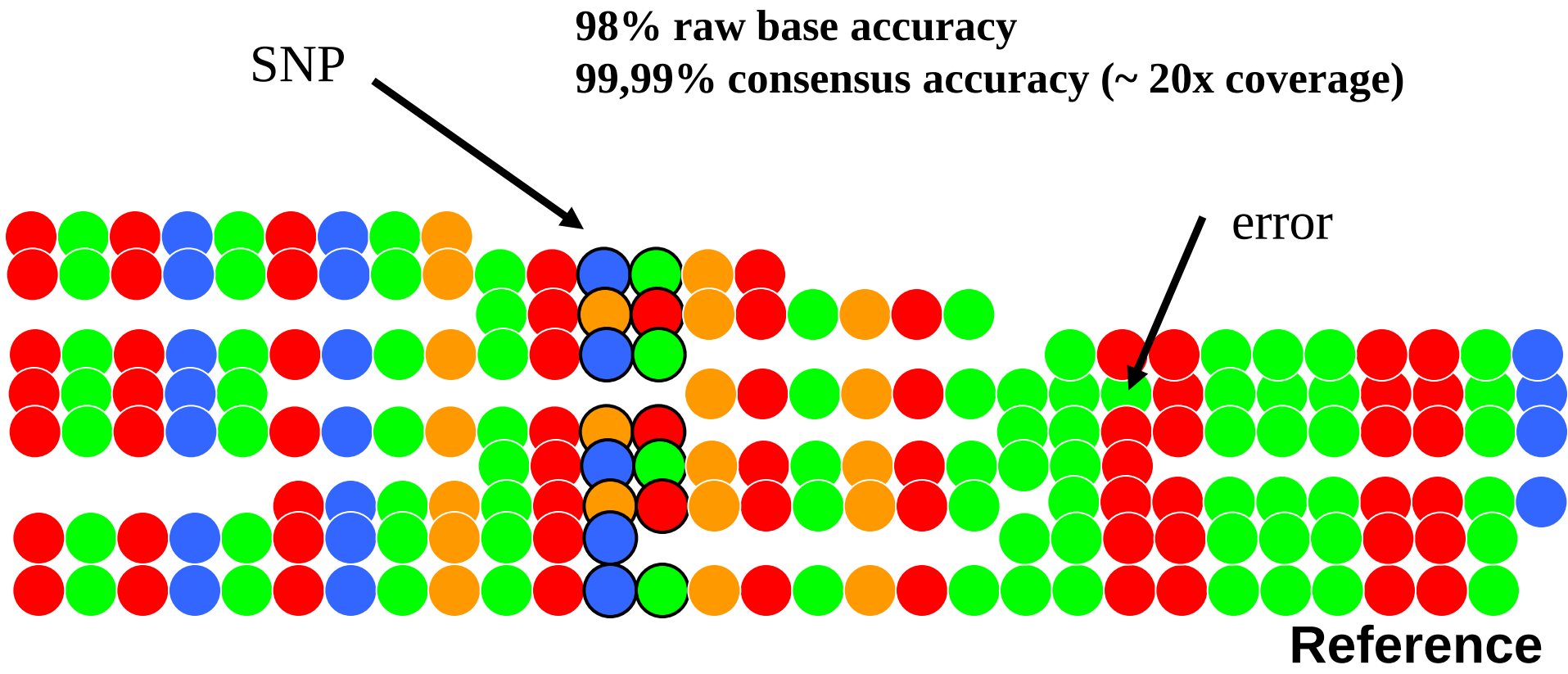


Insertion / Deletion



SOLiD: SNP vs sequencing error

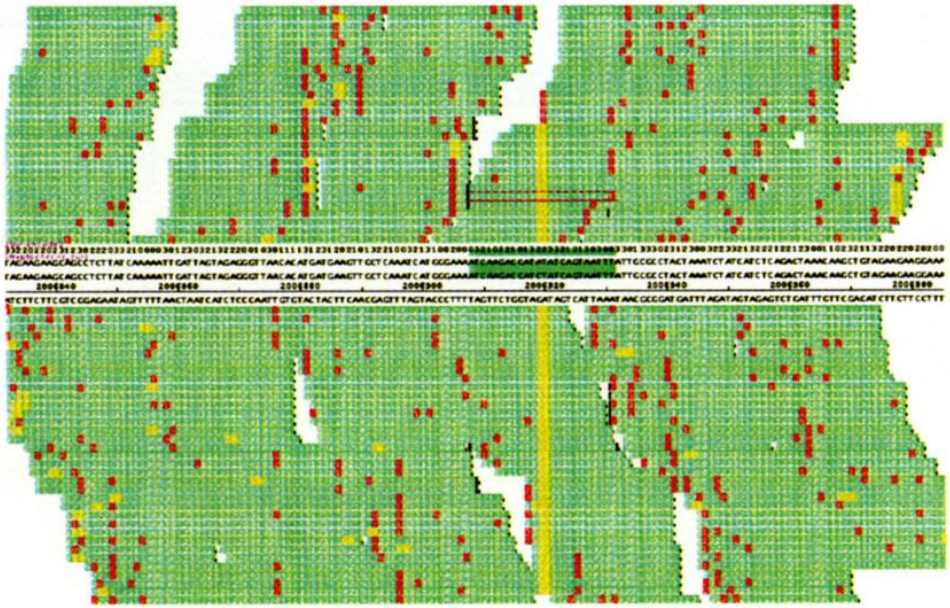
98% raw base accuracy
99,99% consensus accuracy (~ 20x coverage)



SNP 2 colors change

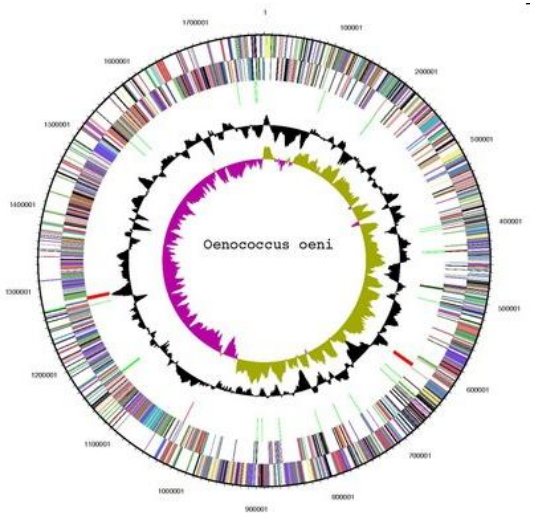
SOLiD: SNP and sequencing errors

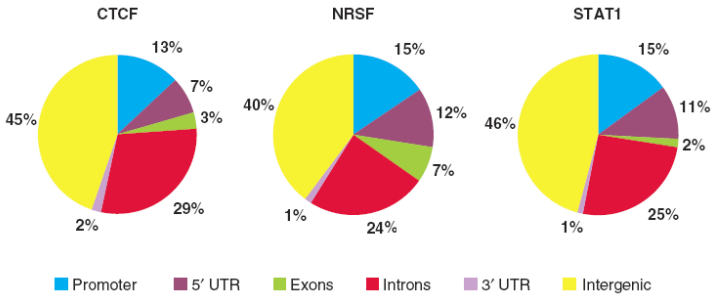
Secondary Analysis
Matching to Reference Sequences

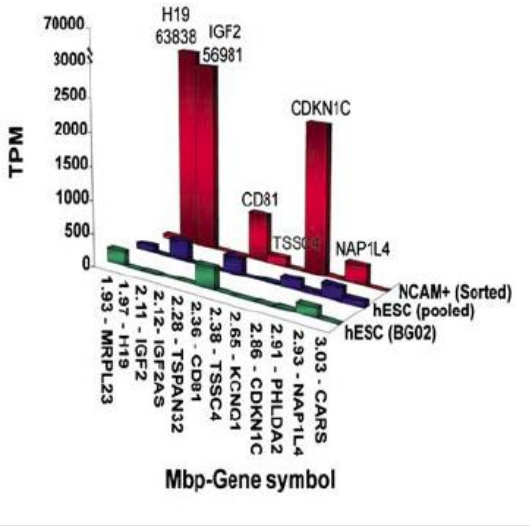


SOLiD usage

- Whole genome (eucariota or bacteria) resequencing
- Pathogen evolution study
- SNP discovery / associations studies
- Exome resequencing (using Agilent SureSelect tecnology)
- Whole transcriptome analysis
- Micro-RNA analysis
- Genome structural variation analysis / cancer genome study
- ChIP-Seq analysis
- Methylation analysis





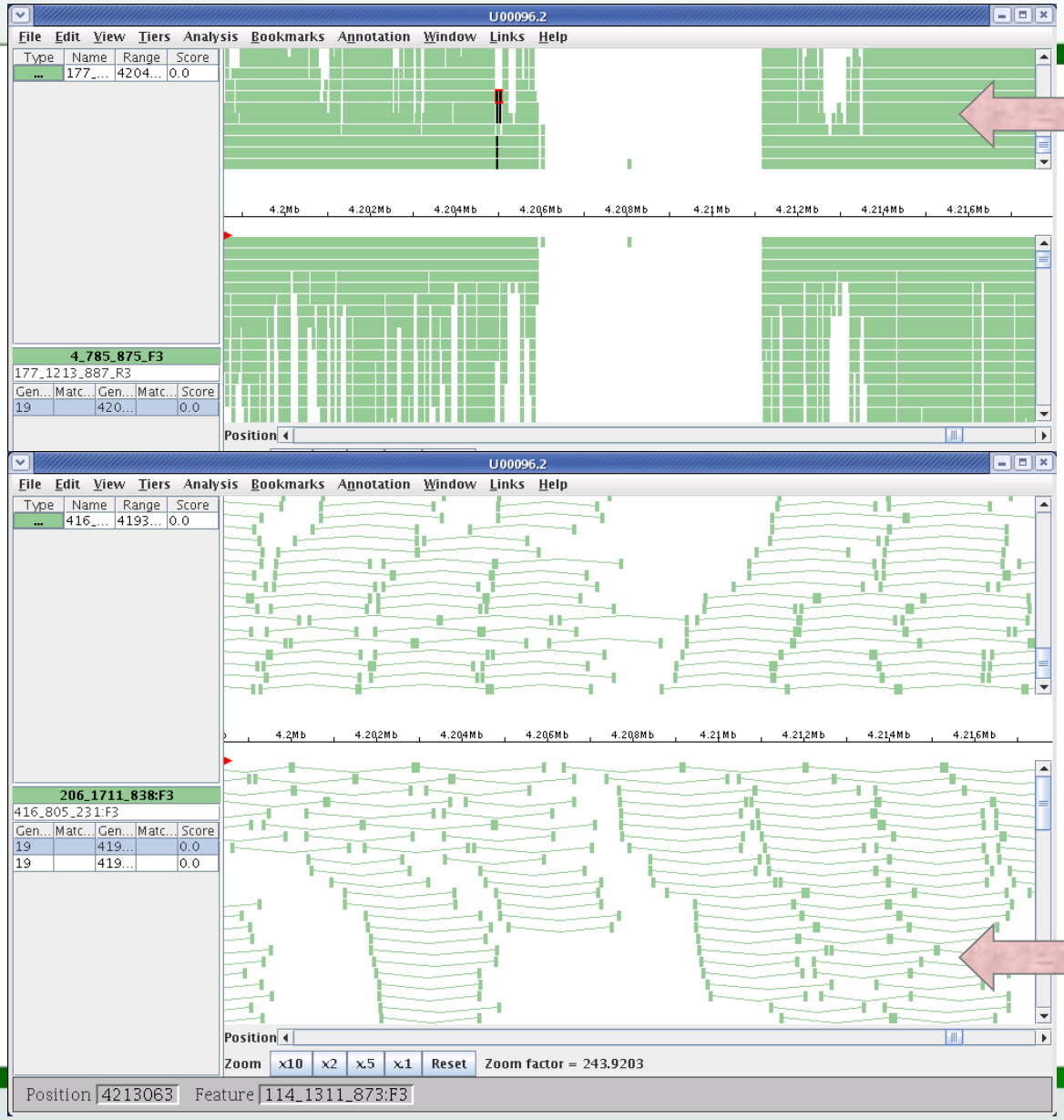


SOLiD shortcomings

Real life differs from advertising pictures

- Unusable for *de novo* sequencing
- Some reads are artificial – polyclonal beads result
- Two PCR reactions produce deviation from random reads distribution
- Multi-step sample preparation produce artifacts

Gaps in sequence up 5% of the reference

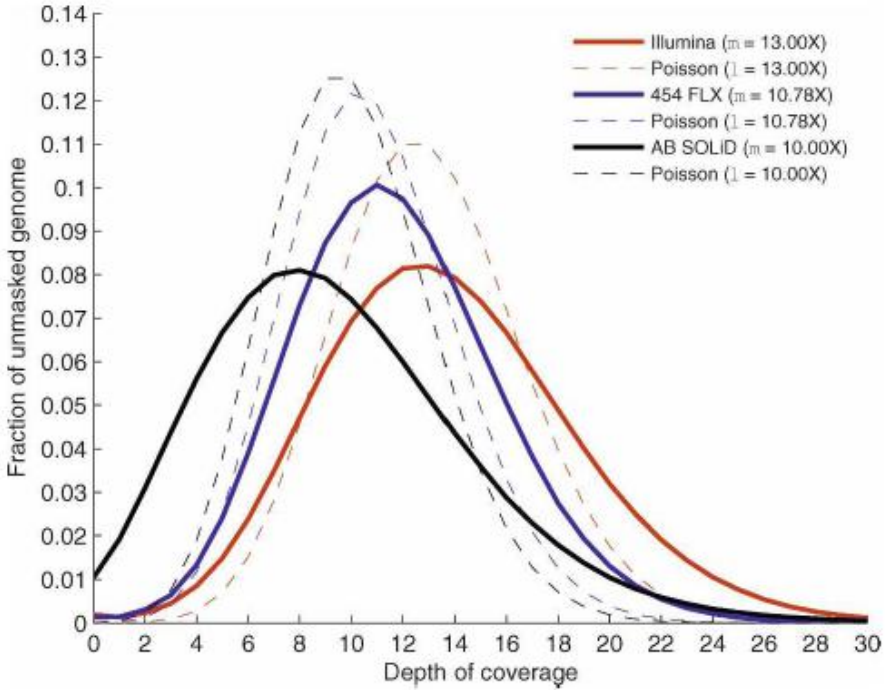


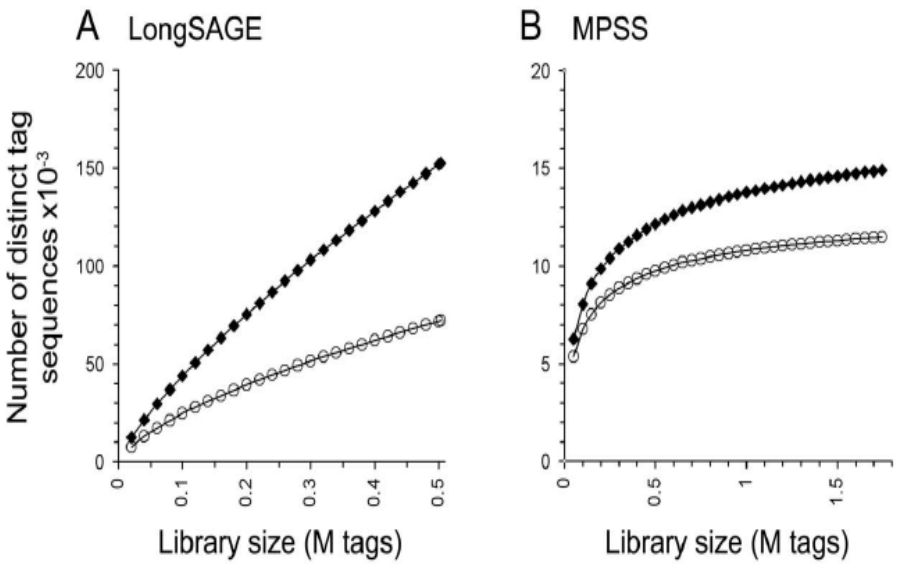
Coverage by uniquely placing single reads

- Duplicated rRNA genes: rrsE & rrIE
- No coverage by unique single reads
- Mapped with mate-paired reads

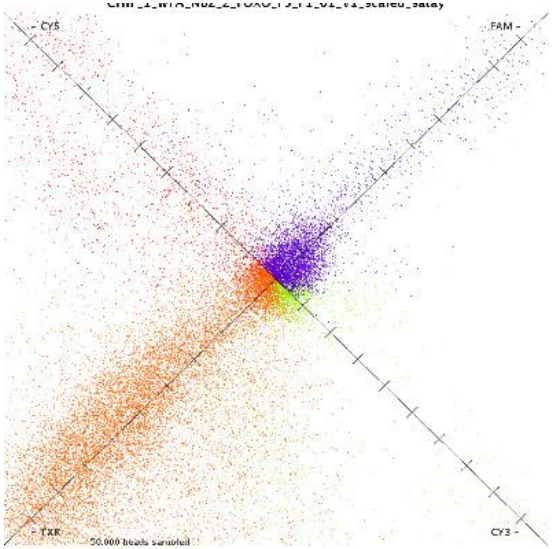
Coverage by mate-paired reads

Deviation from Poisson for MPSS reads

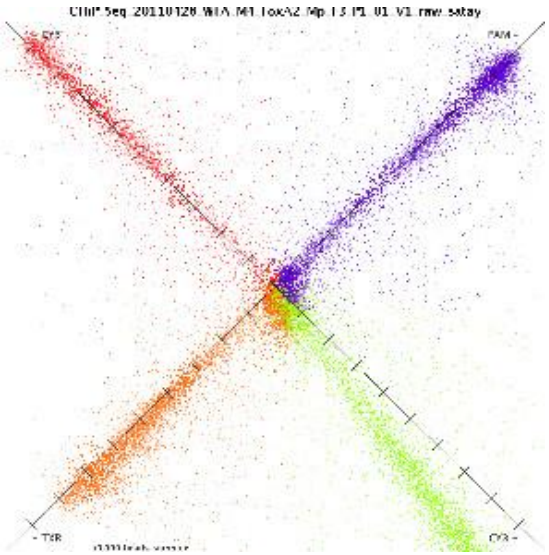




Multi-step sample preparation produce artifacts
Some reads are artificial – polyclonal beads result



Bad library



Usable library

Data quality drops with read lenth