Introduction to the MiSeq System

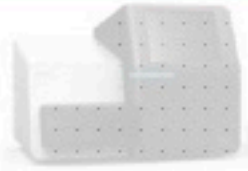
illumina

MiSeq Sequencing Workflow

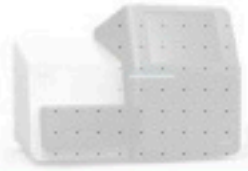
1. Library Preparation

2. Cluster Generation

3. Sequencing



4. Data Analysis



illumina

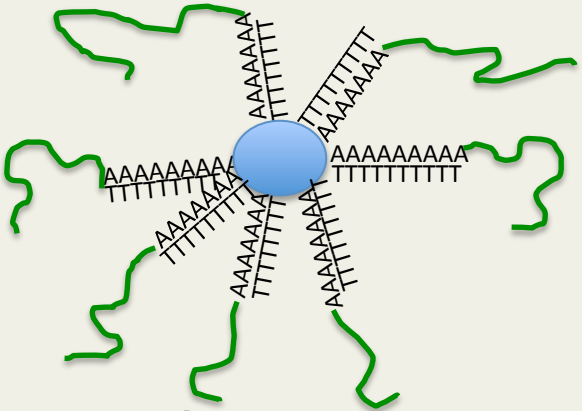
Sample Prep is Critical for Successful Sequencing

The aim of the sample prep step is to obtain nucleic acid fragments with adapters attached on both ends

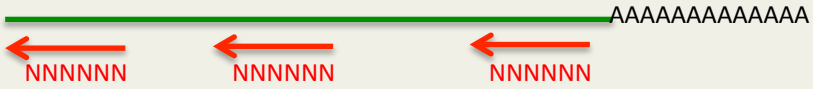
illumina

Total RNA isolation

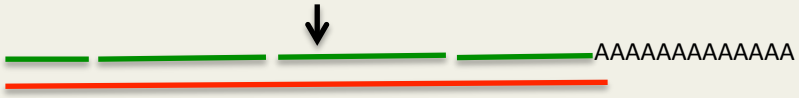
mRNA Isolation using Oligo(dT)
Magnetic Beads



First-Strand cDNA Synthesis with Random Primers



Second-Strand cDNA Synthesis

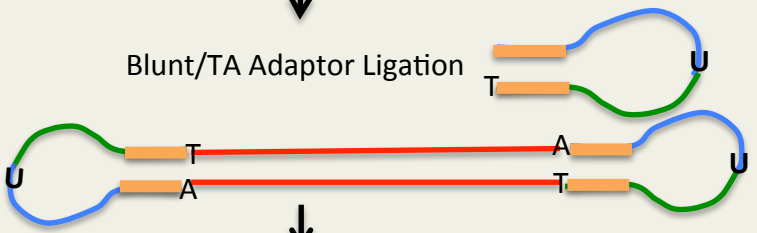


Purify the double stranded cDNA with AMPure magnetic Beads
(1.8X ratio Beads to cDNA volume)

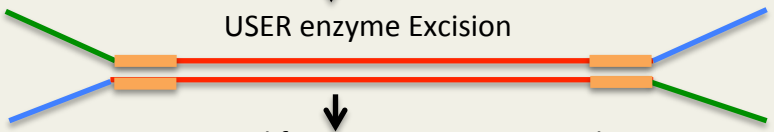
End Repair of double-stranded cDNA

Adenylation (A-Tailing)

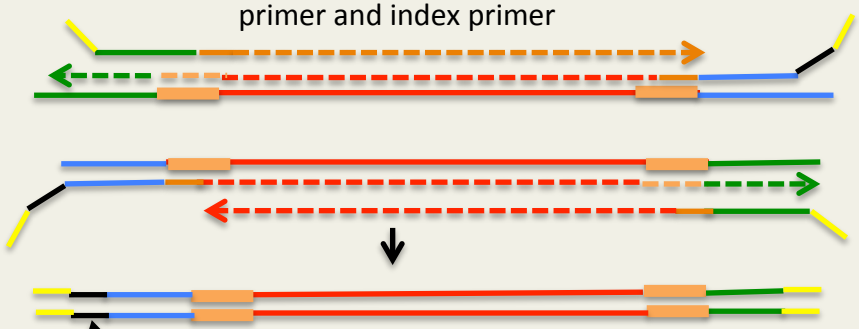
Blunt/TA Adaptor Ligation



USER enzyme Excision



PCR Amplification using a Universal
primer and index primer



Purify and size select cDNA Library using AMPure Beads

MiSeq Sequencing Workflow

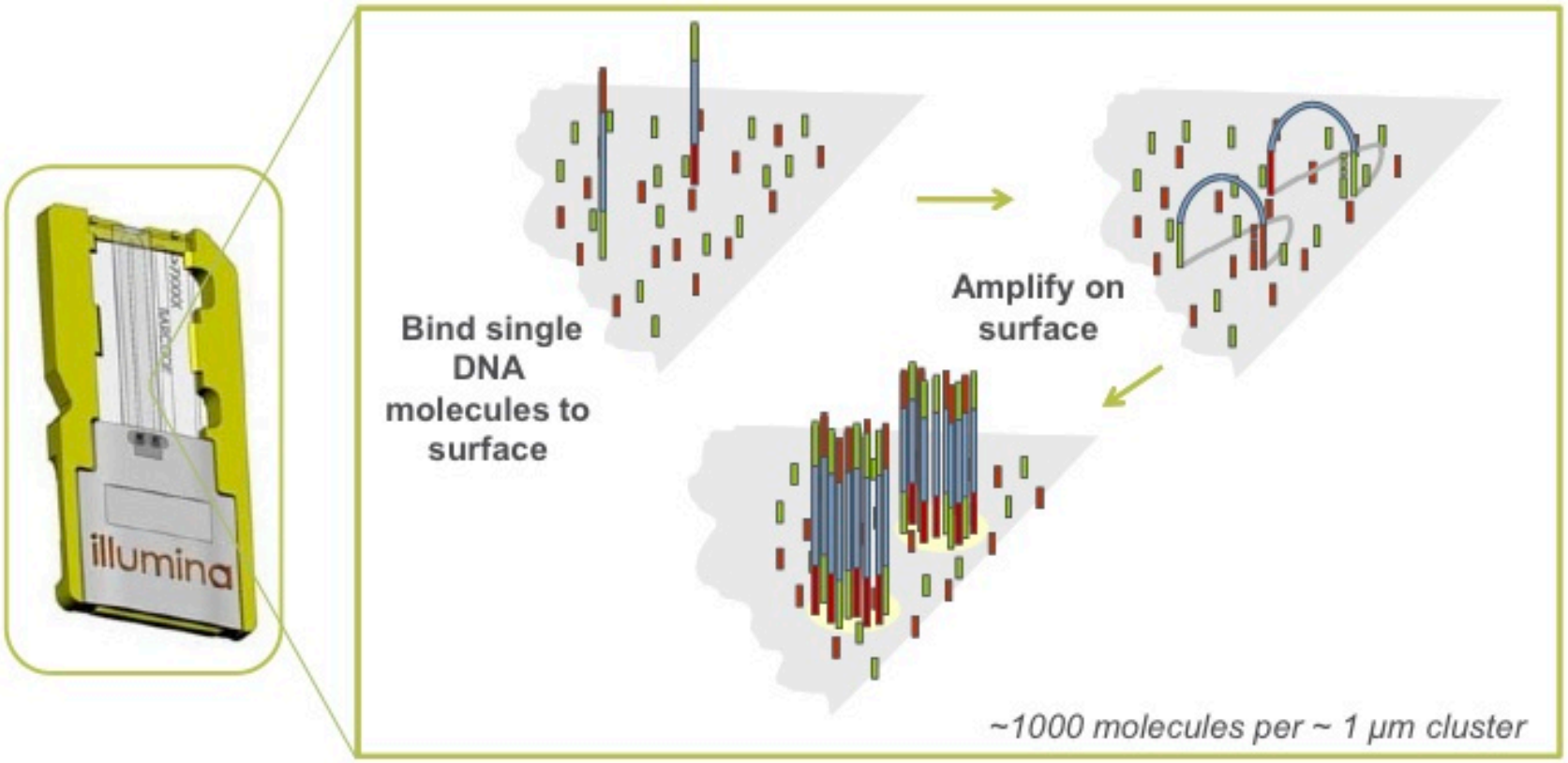
1 Library Preparation

2 Cluster Generation

3 Sequencing

4 Data Analysis

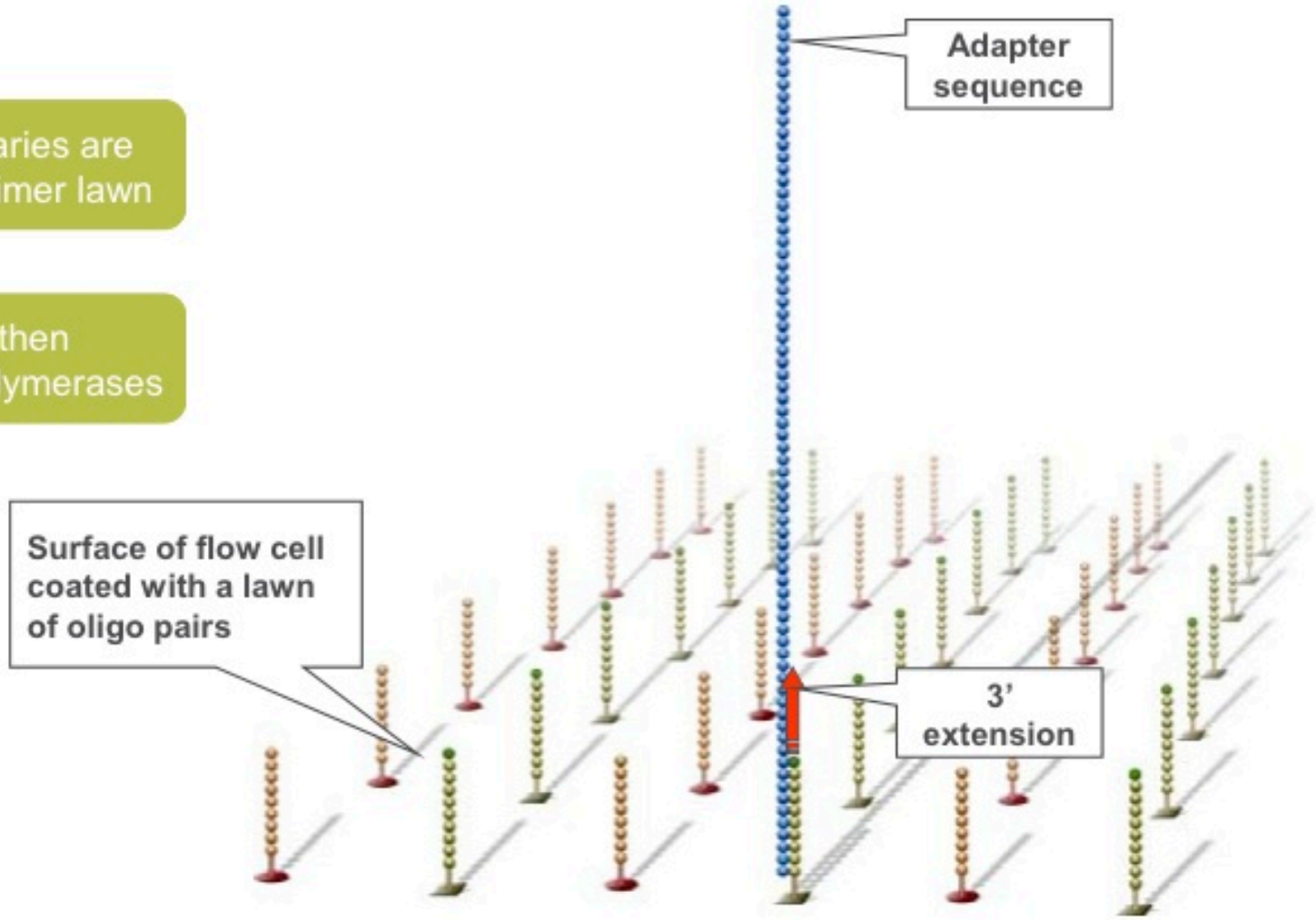
Cluster Generation



Hybridize Fragment & Extend

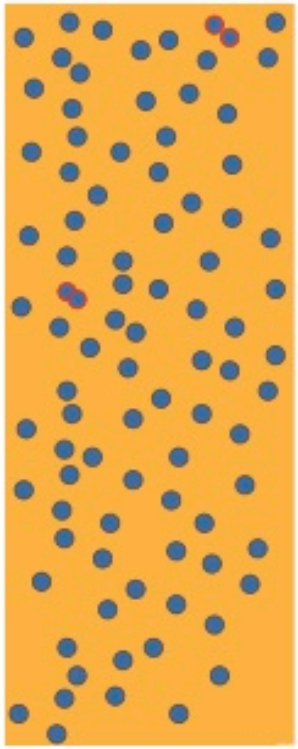
Single DNA libraries are hybridized to primer lawn

Bound libraries then extended by polymerases

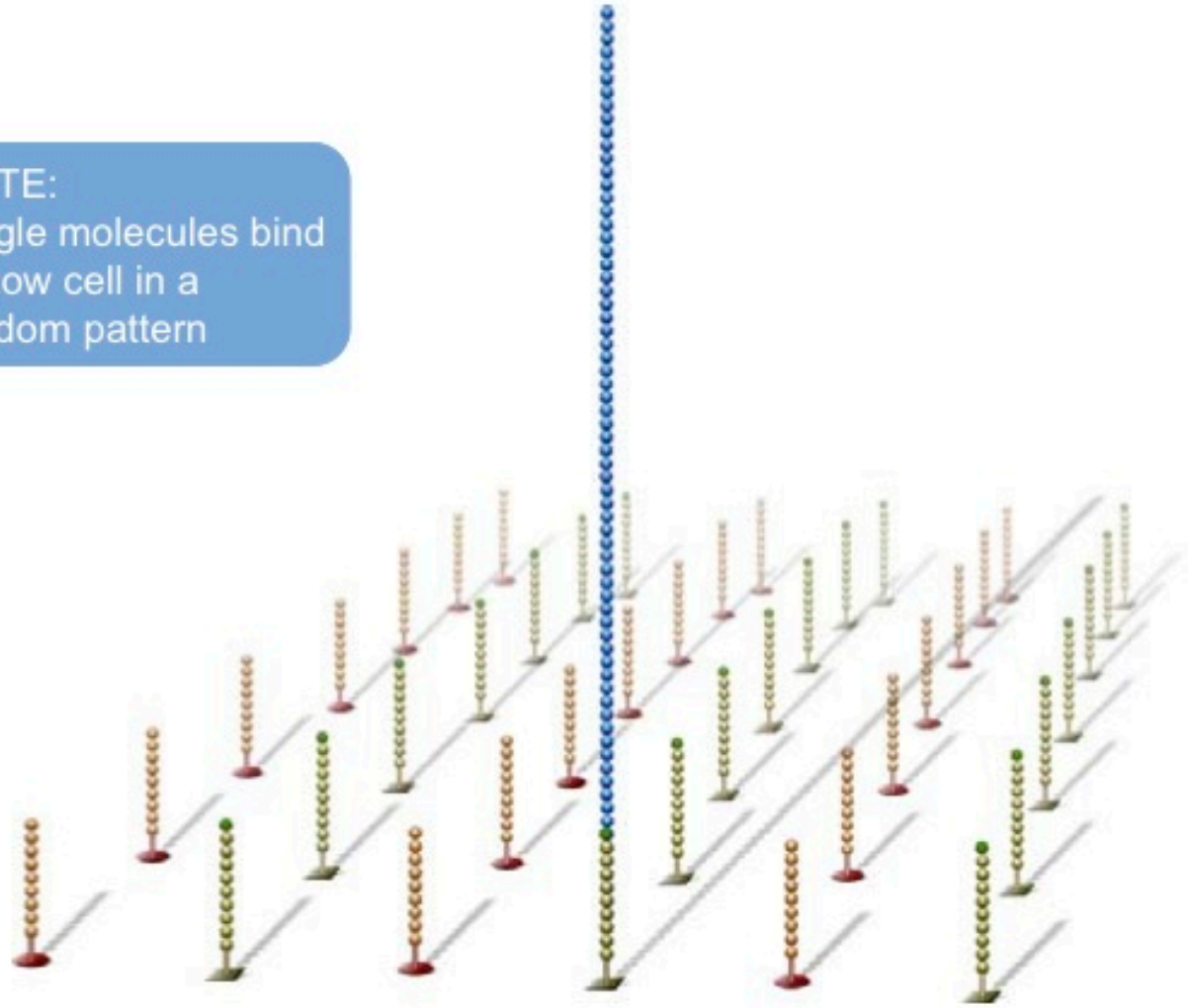


illumina

Hybridize Fragment & Extend



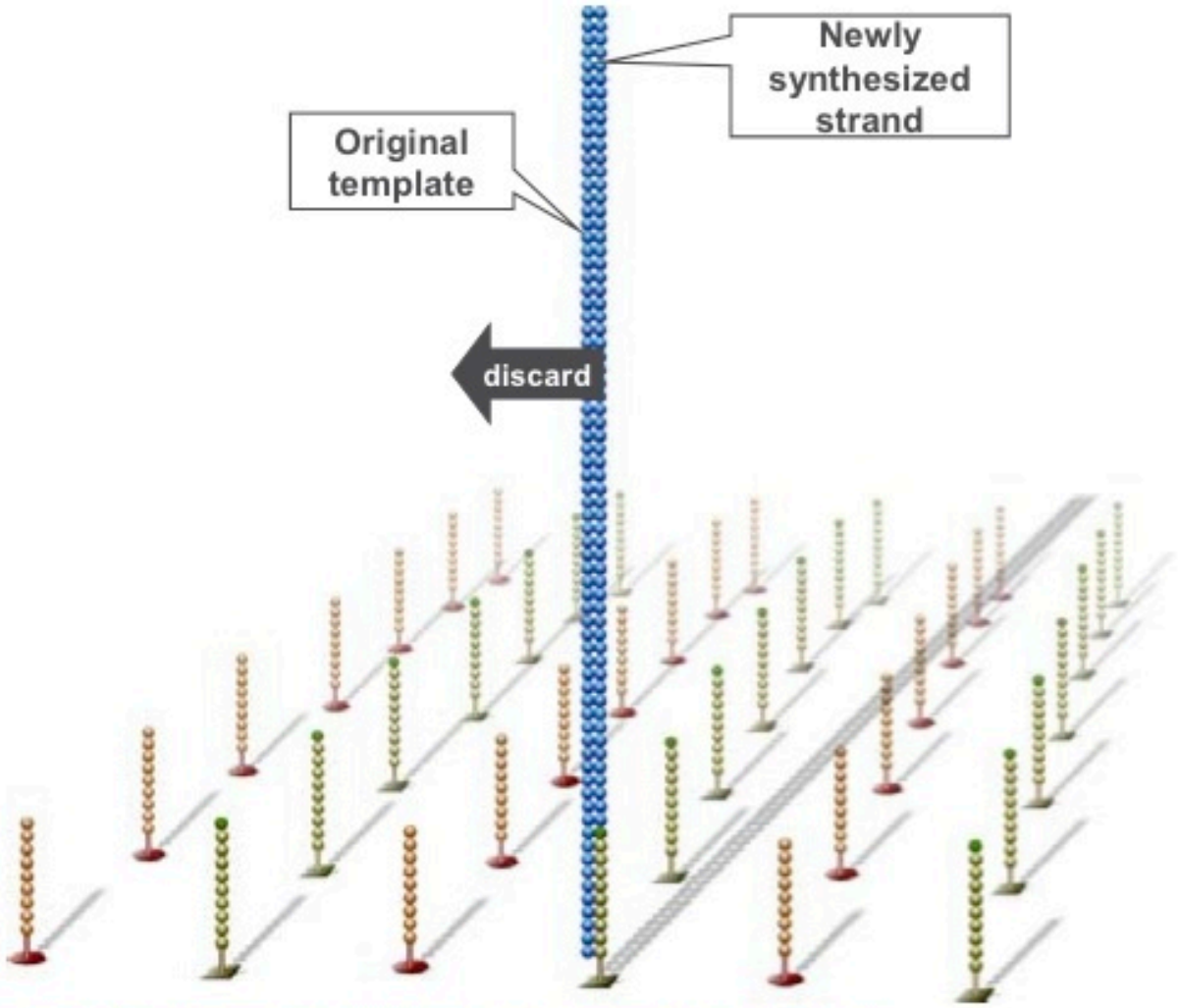
NOTE:
Single molecules bind to flow cell in a random pattern



illumina

Denature Double-Stranded DNA

- Double-stranded molecule is denatured
- Original template washed away
- Newly synthesized strand is covalently attached to flow cell surface

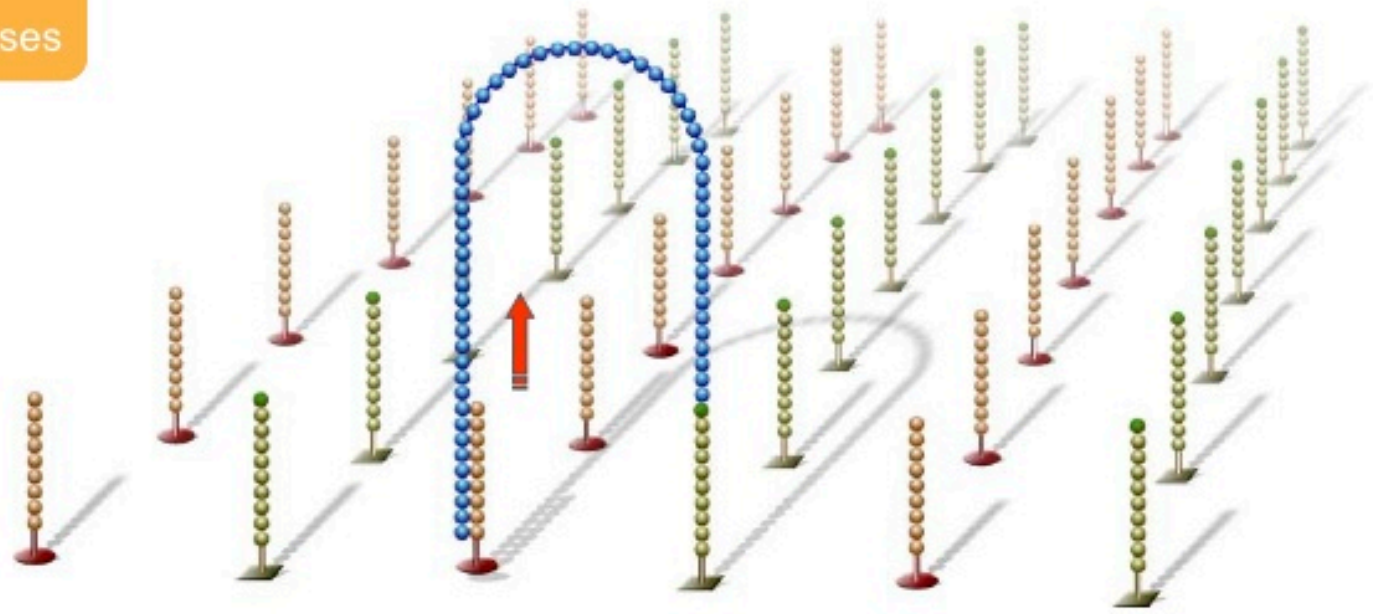


illumina

Bridge Amplification

Single-stranded molecule flips over and forms a bridge by hybridizing to adjacent, complementary primer

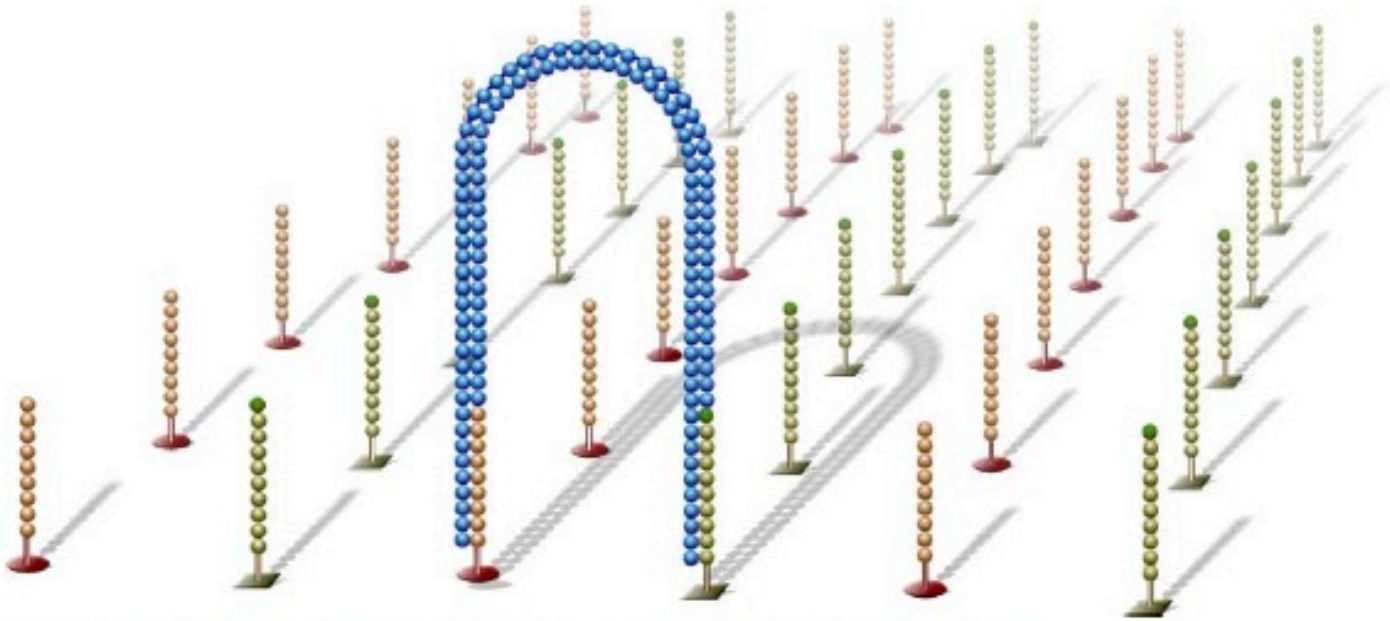
Hybridized primer is extended by polymerases



illumina

Bridge Amplification

Double-stranded bridge is formed

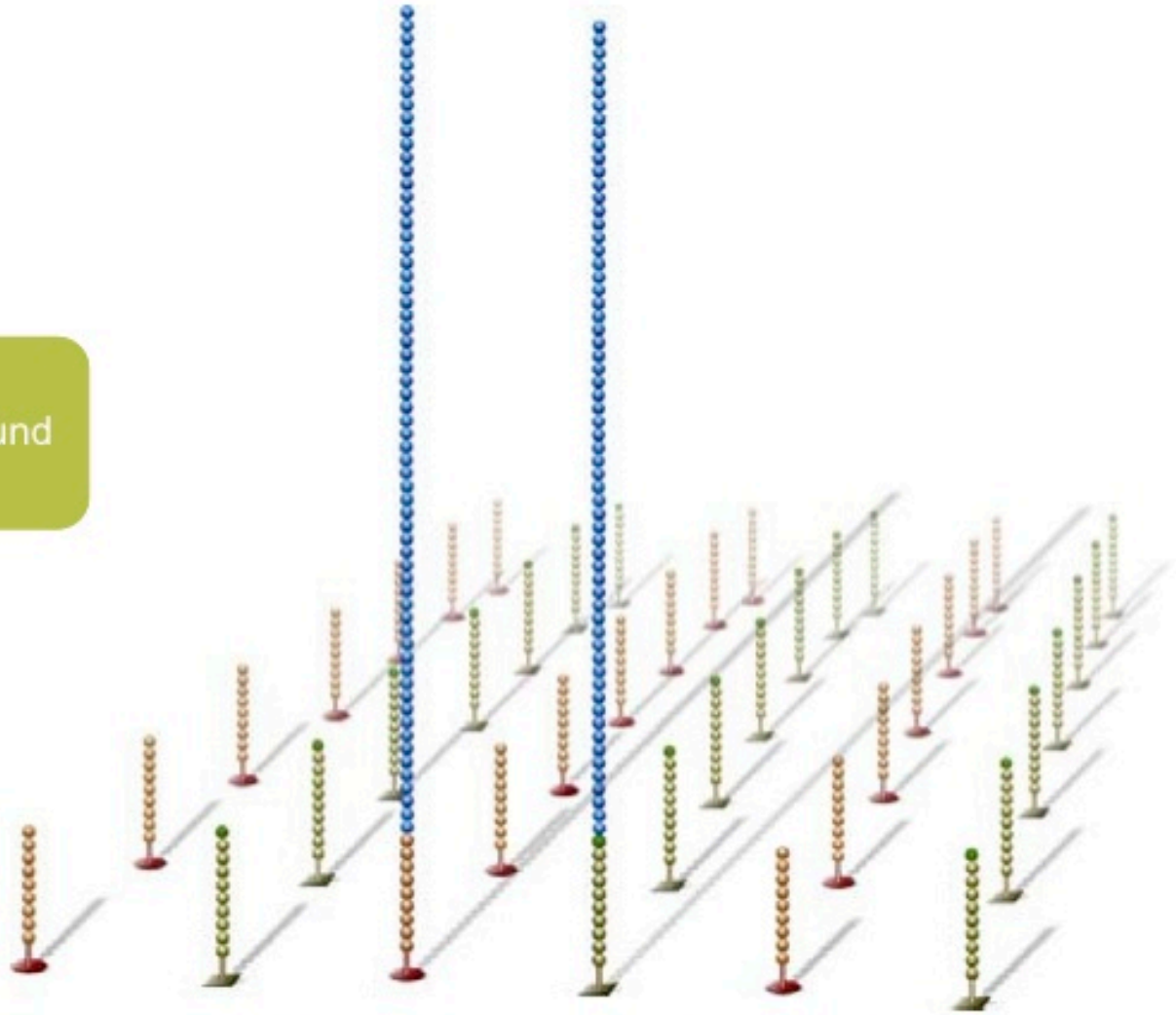


illumina®

Denature Double-Stranded Bridge

Double-stranded bridge is denatured

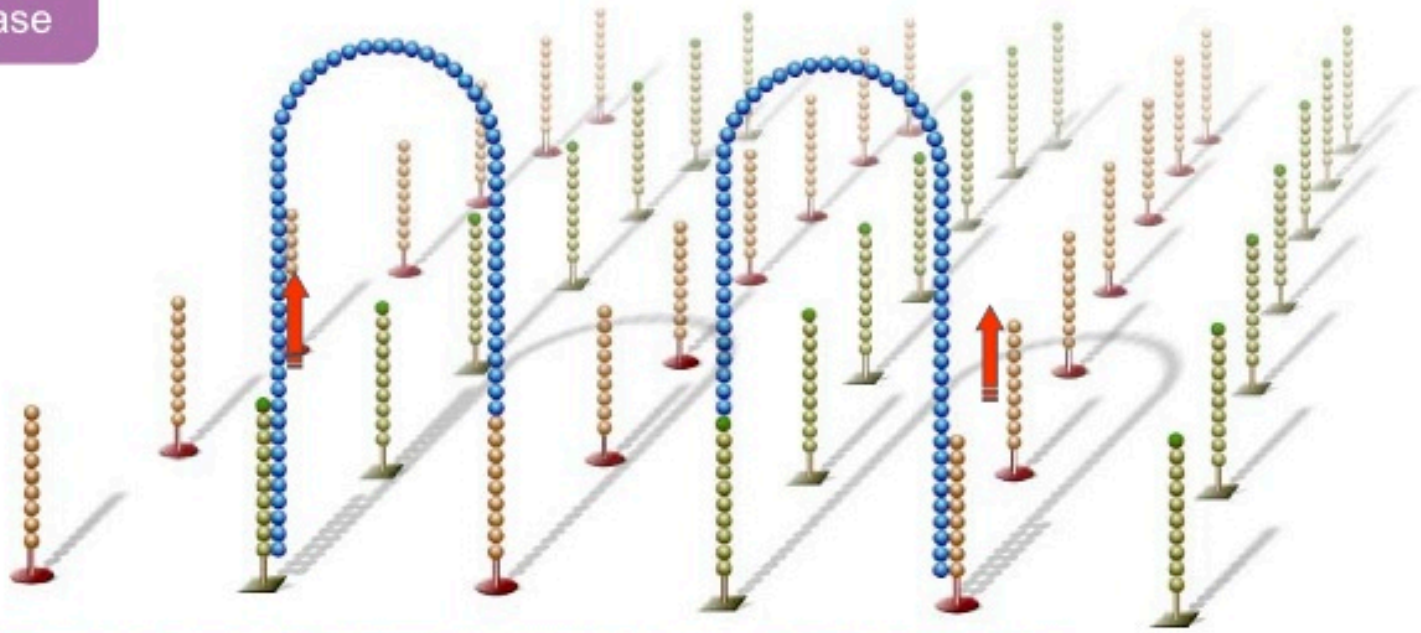
Result:
Two copies of covalently bound single-stranded templates



Bridge Amplification

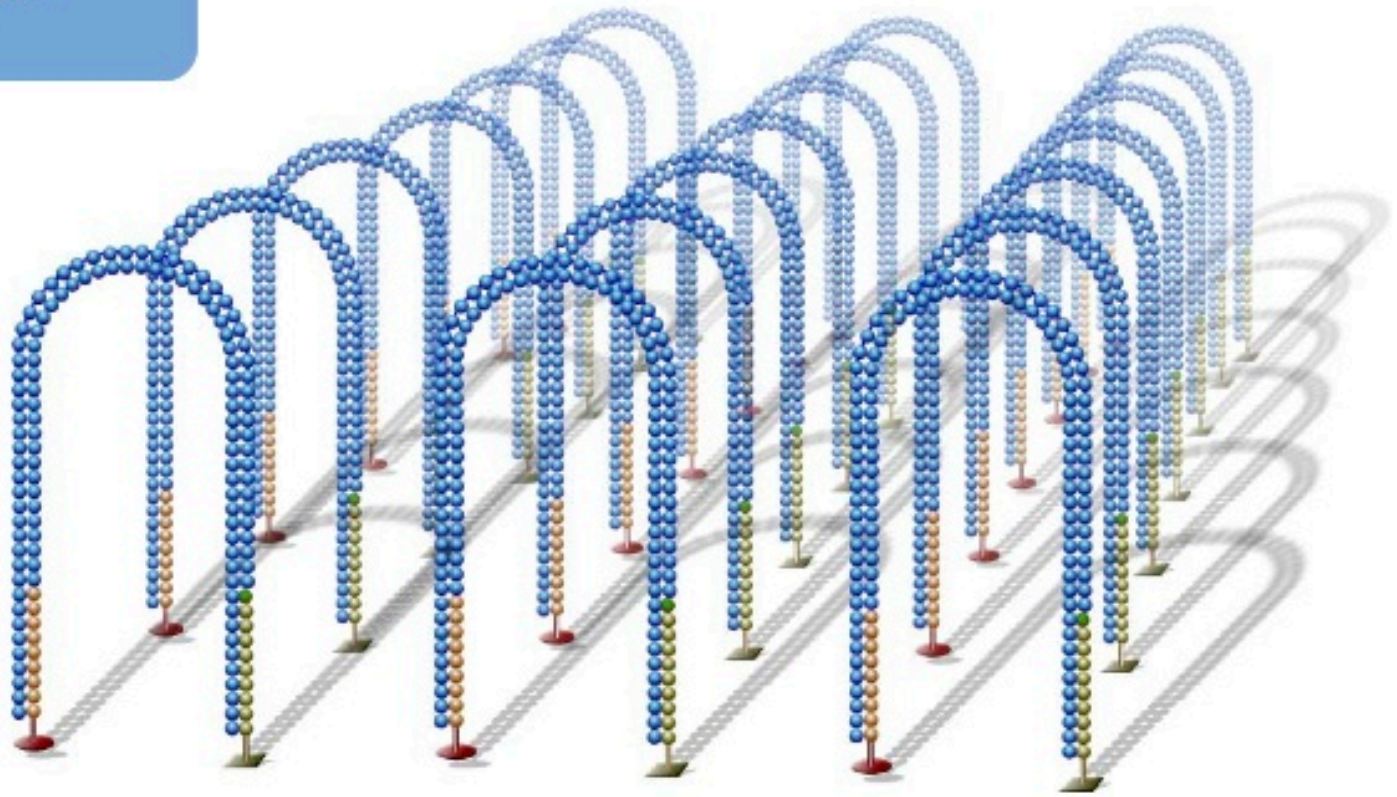
Single-stranded molecules flip over to hybridize to adjacent primers

Hybridized primer is extended by polymerase



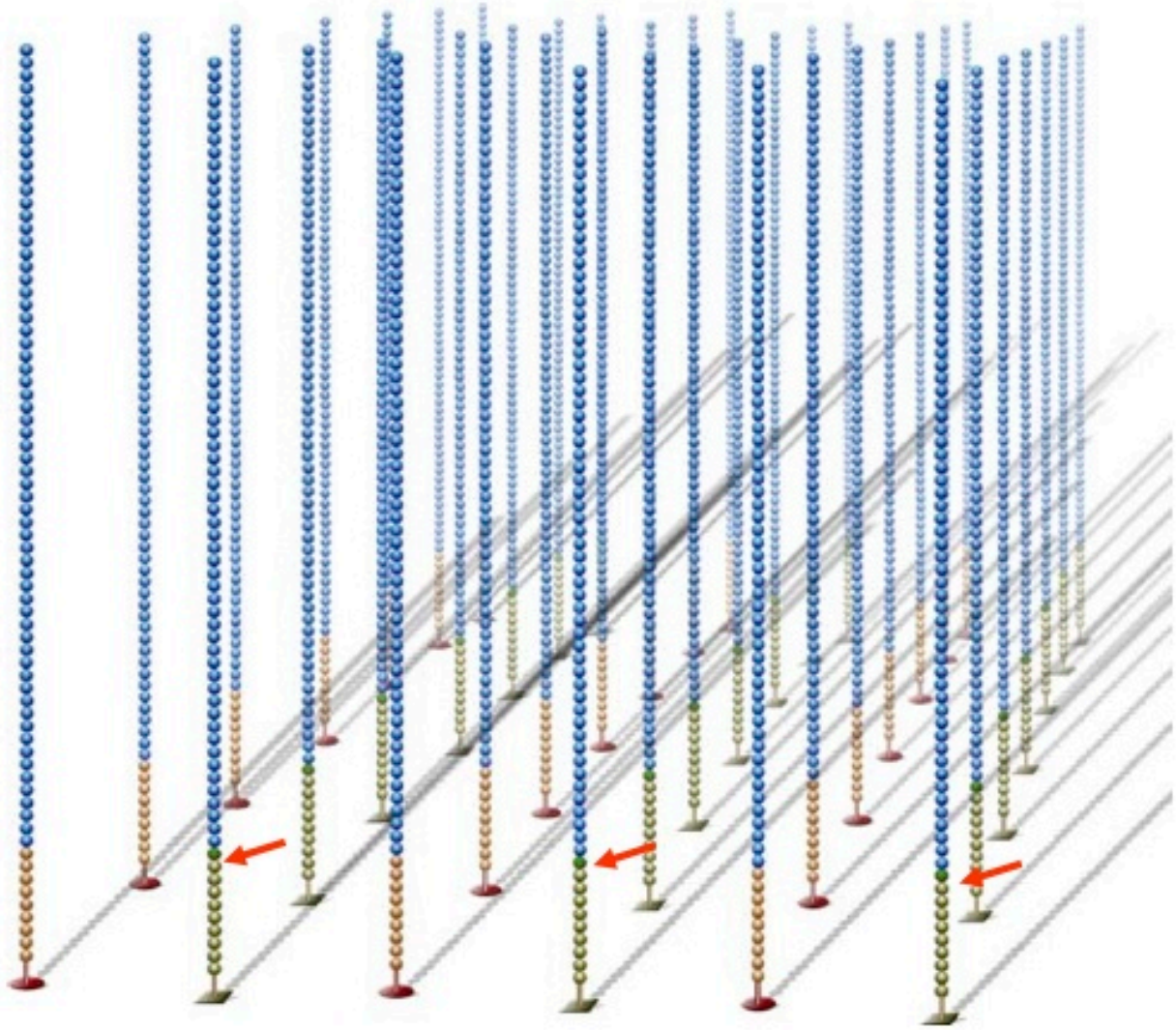
Bridge Amplification

Bridge amplification cycle repeated until multiple bridges are formed



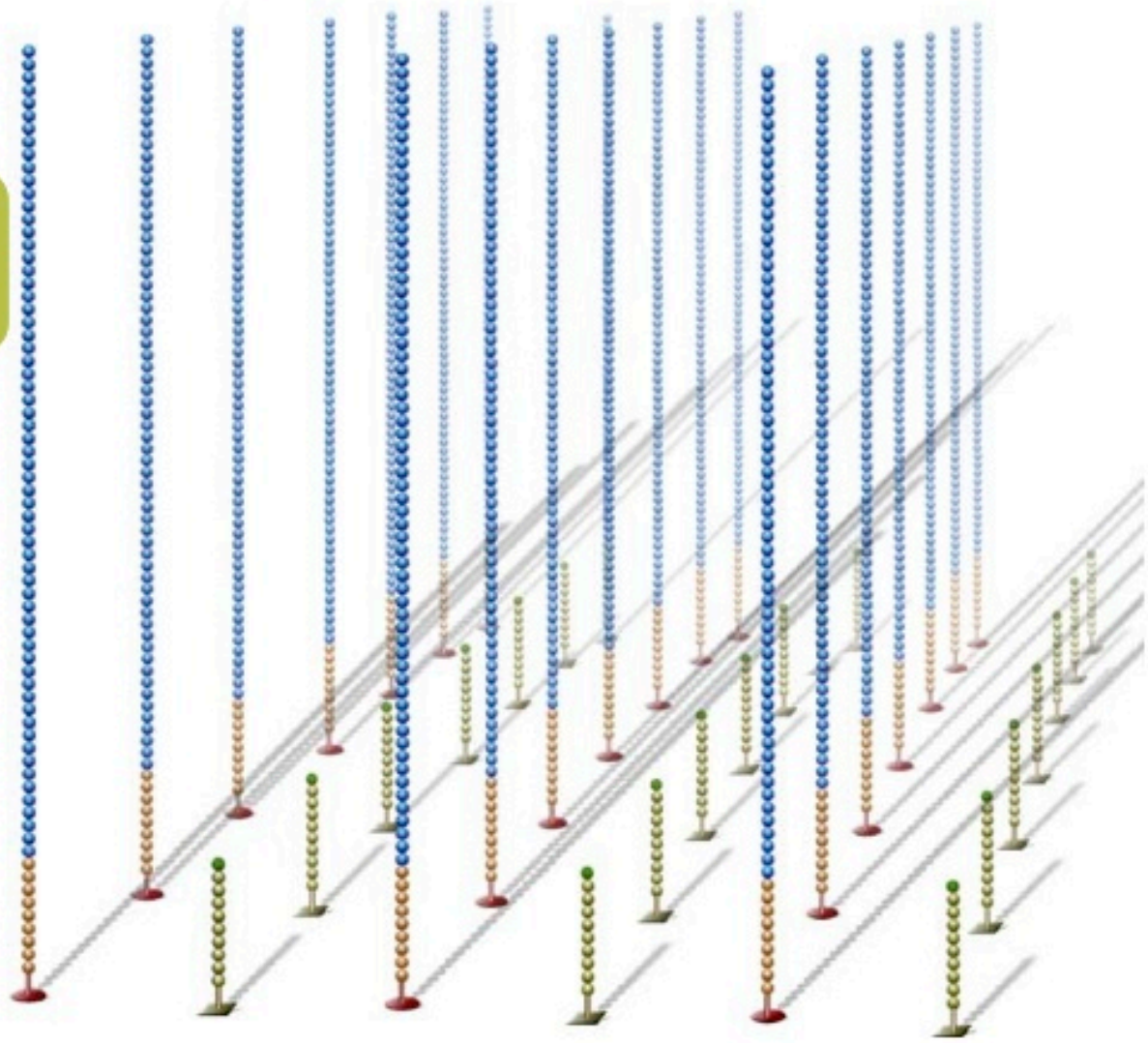
Linearization

dsDNA bridges are denatured



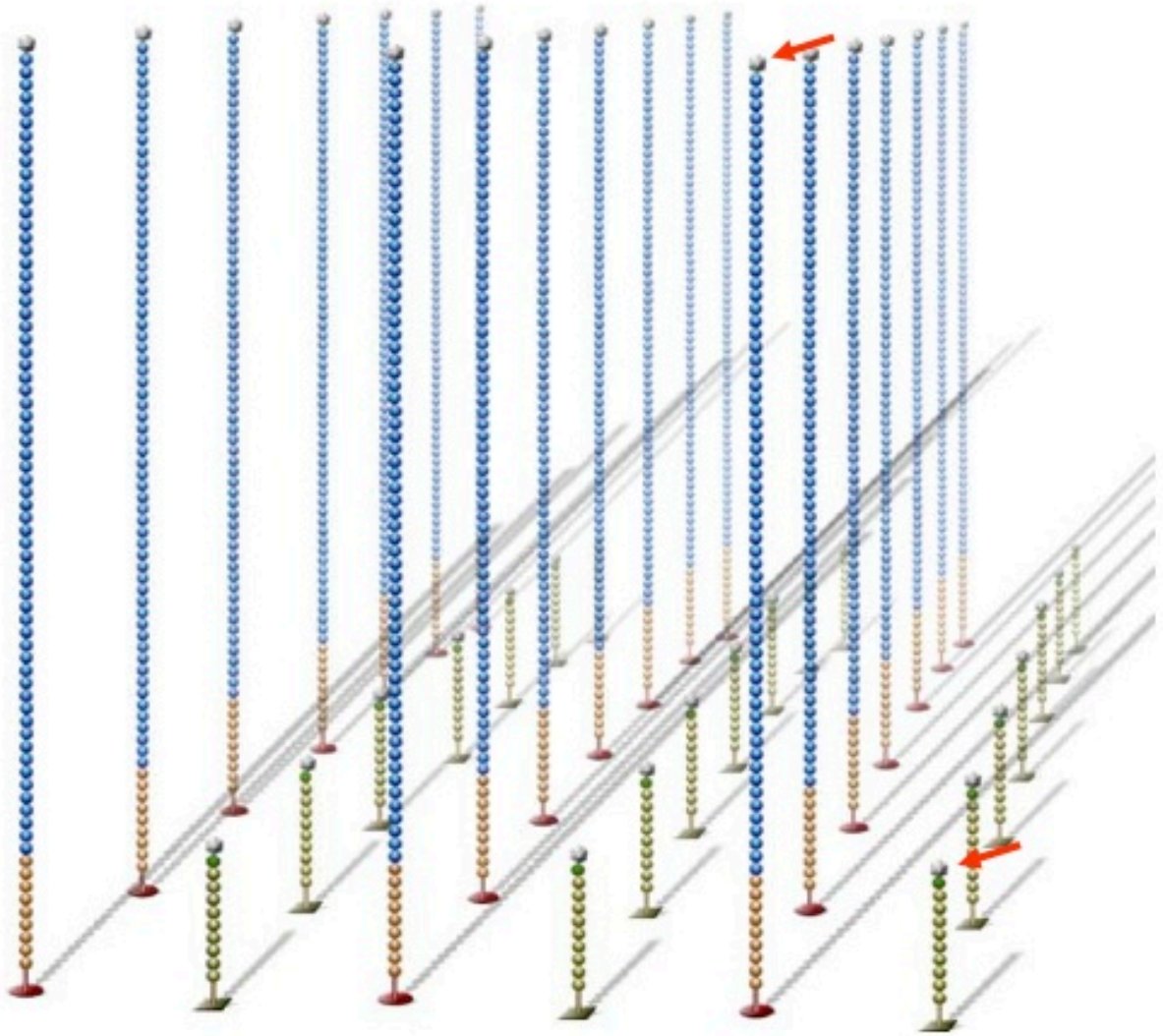
Reverse Strand Cleavage

Reverse strands cleaved and washed away, leaving a cluster with forward strands only



Blocking

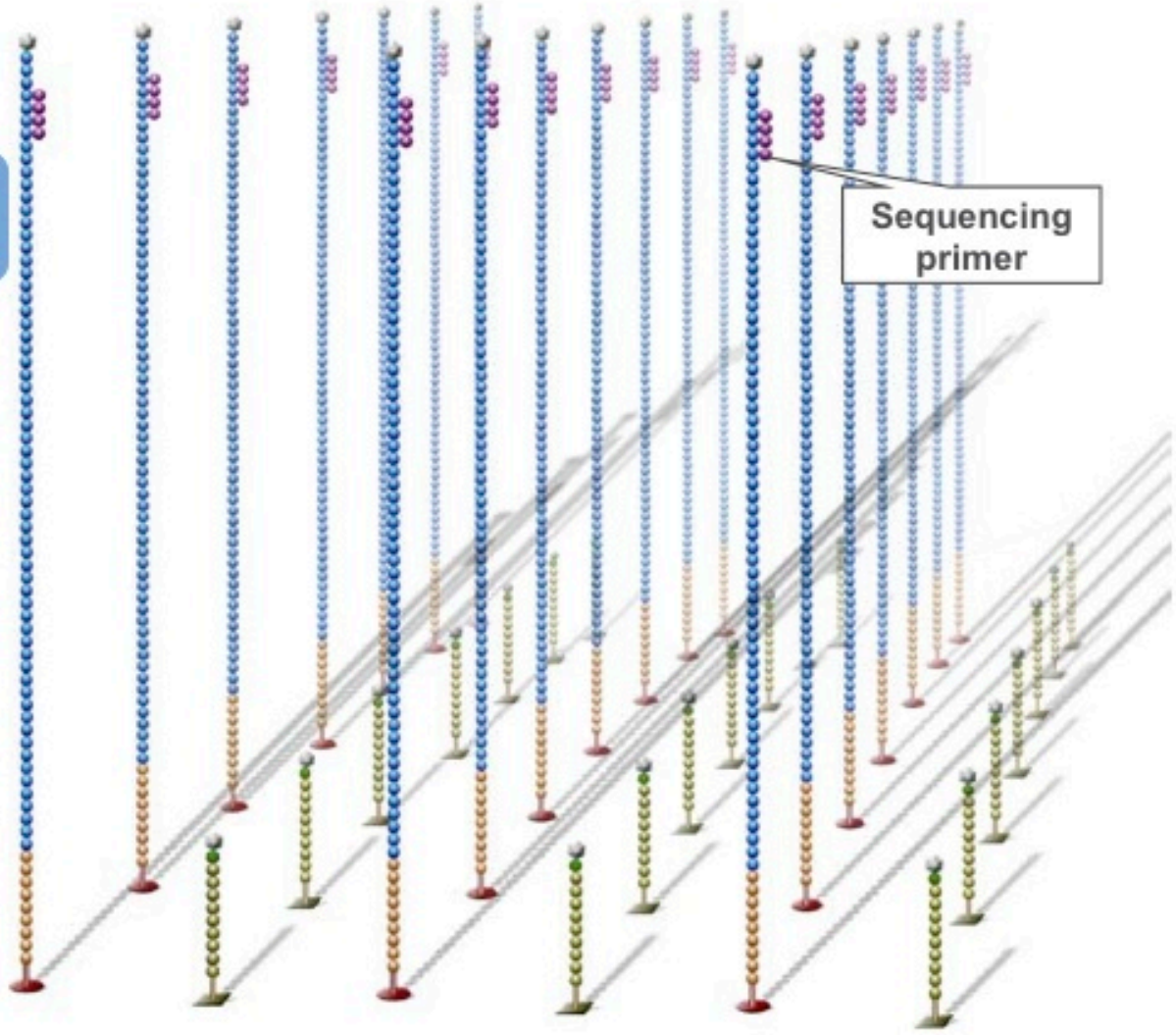
Free 3' ends are blocked to prevent unwanted DNA priming



illumina

Read 1 Primer Hybridization

Sequencing primer is hybridized to adapter sequence



illumina

MiSeq Sequencing Workflow

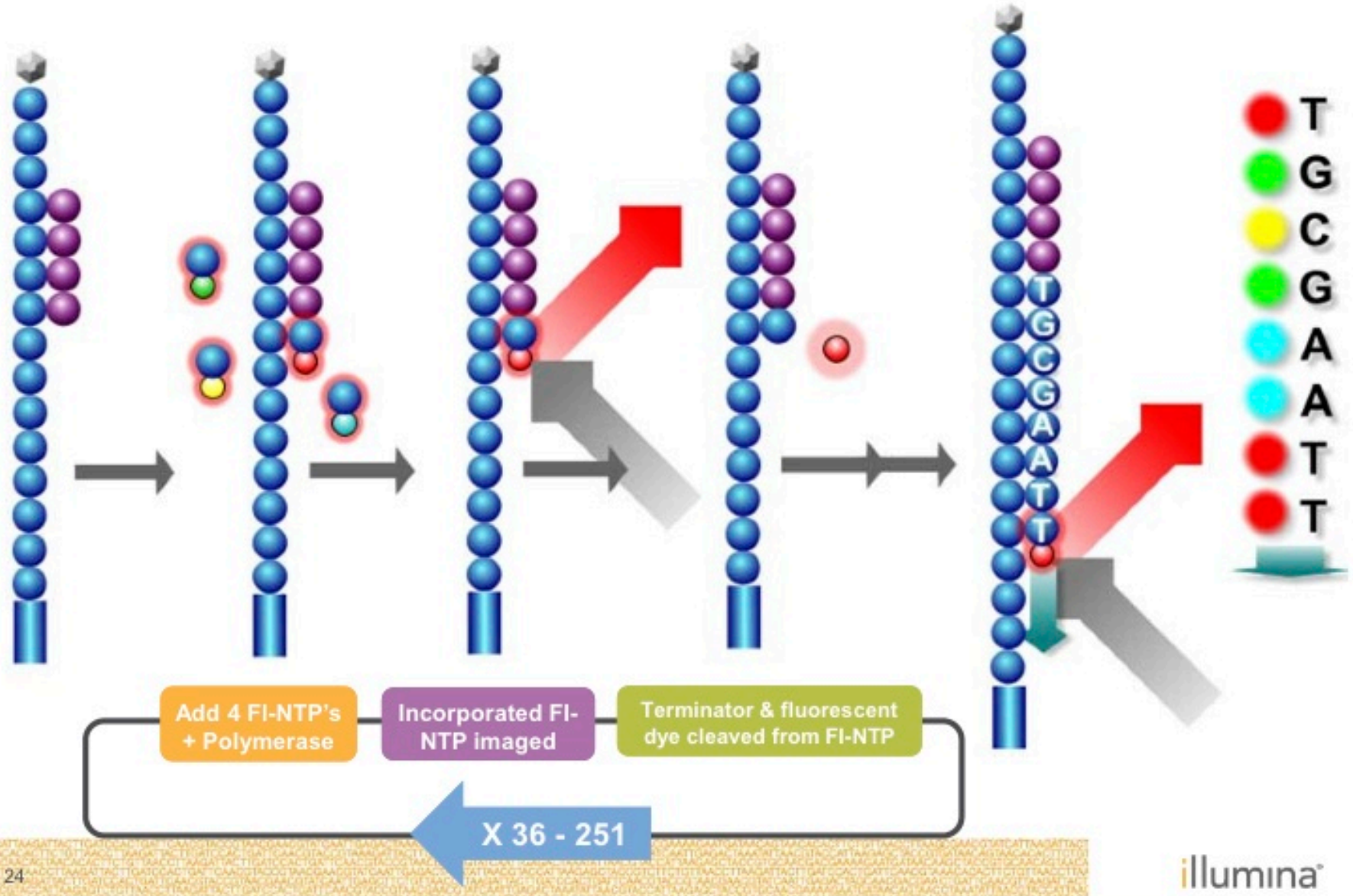
1 Library Preparation

2 Cluster Generation

3 Sequencing

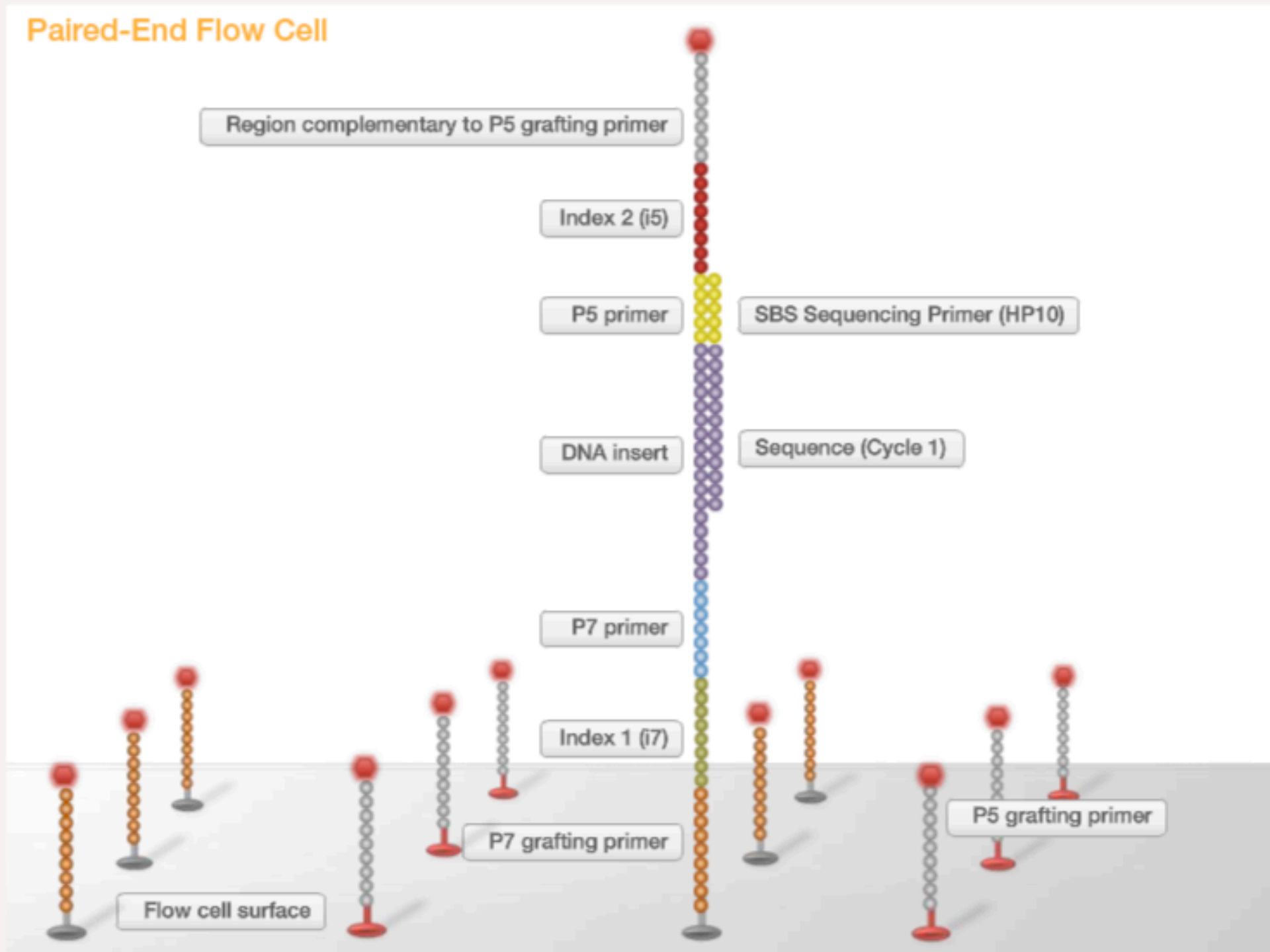
4 Data Analysis

Sequencing by Synthesis



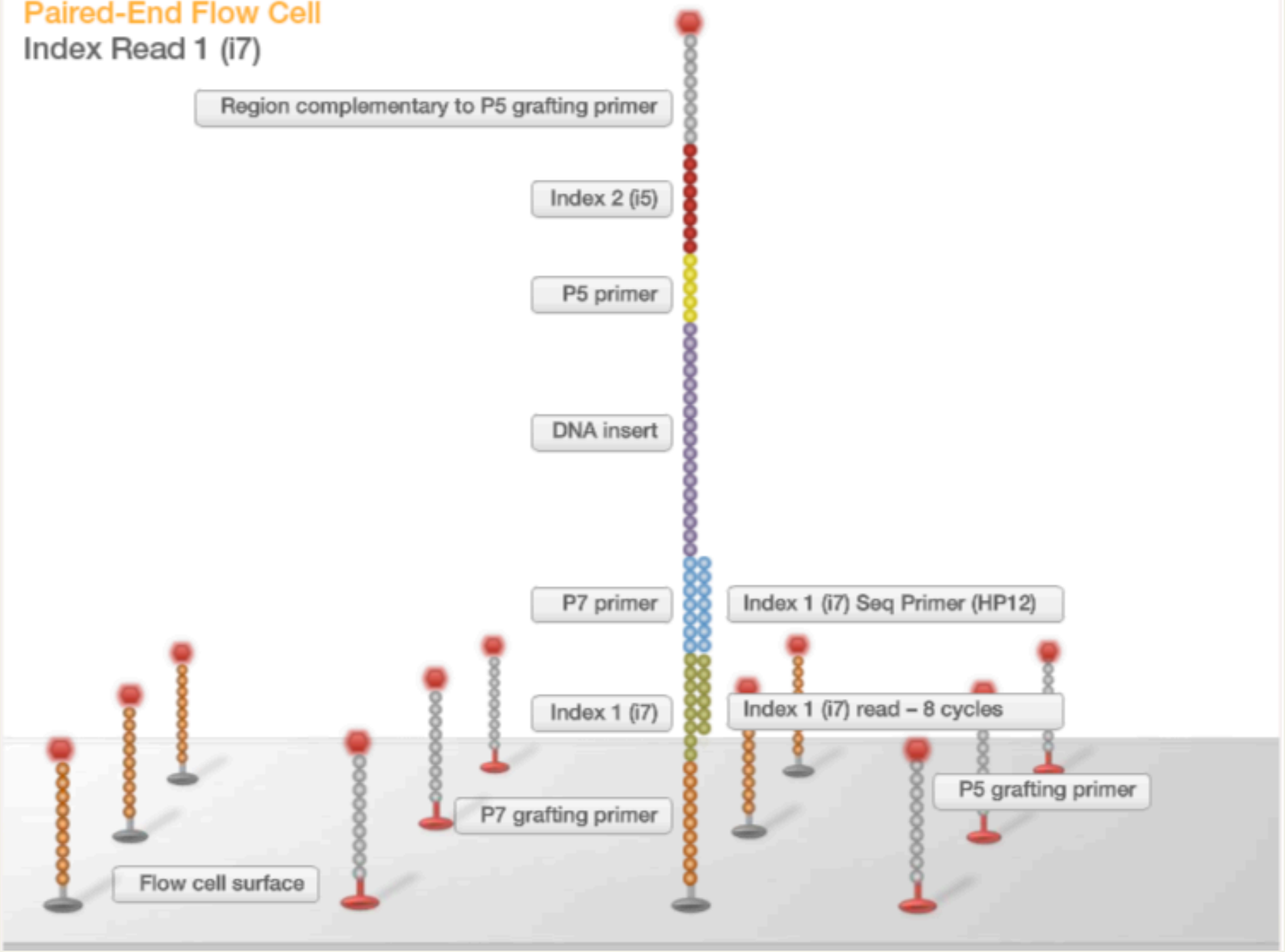
illumina

Paired-End Flow Cell



Paired-End Flow Cell

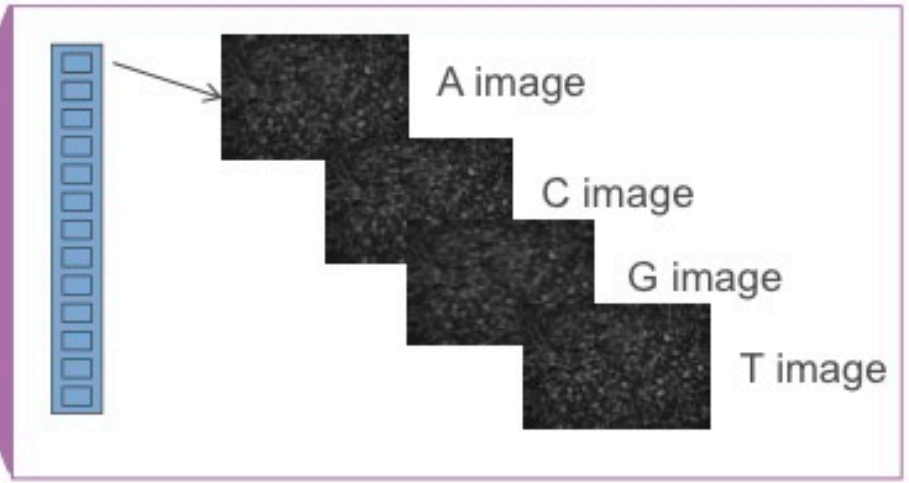
Index Read 1 (i7)



Sequencing

Clusters are images using LED and filter combinations specific for each fluorescently-labeled nucleotide

After imaging is complete for one section (tile), the flow cell is moved to the next tile and the process is repeated



Experimental Workflow

•Calculate coverage you want for your genome/transcriptome, and, using the coverage calculator, calculate how many samples will fit on your run.

•Isolate your RNA as usual, and get rid of rRNA if not needed

•Quantitate on the Bioanalyzer (for quality) and Qubit (for quantity) (may want to isolate a couple of extra and take the best ones).

•Prepare each library.

•Quantitate one more time with bioanalyzer and Qubit, then make calculation for dilutions to 4nM which is where the protocol starts on sequencing day.

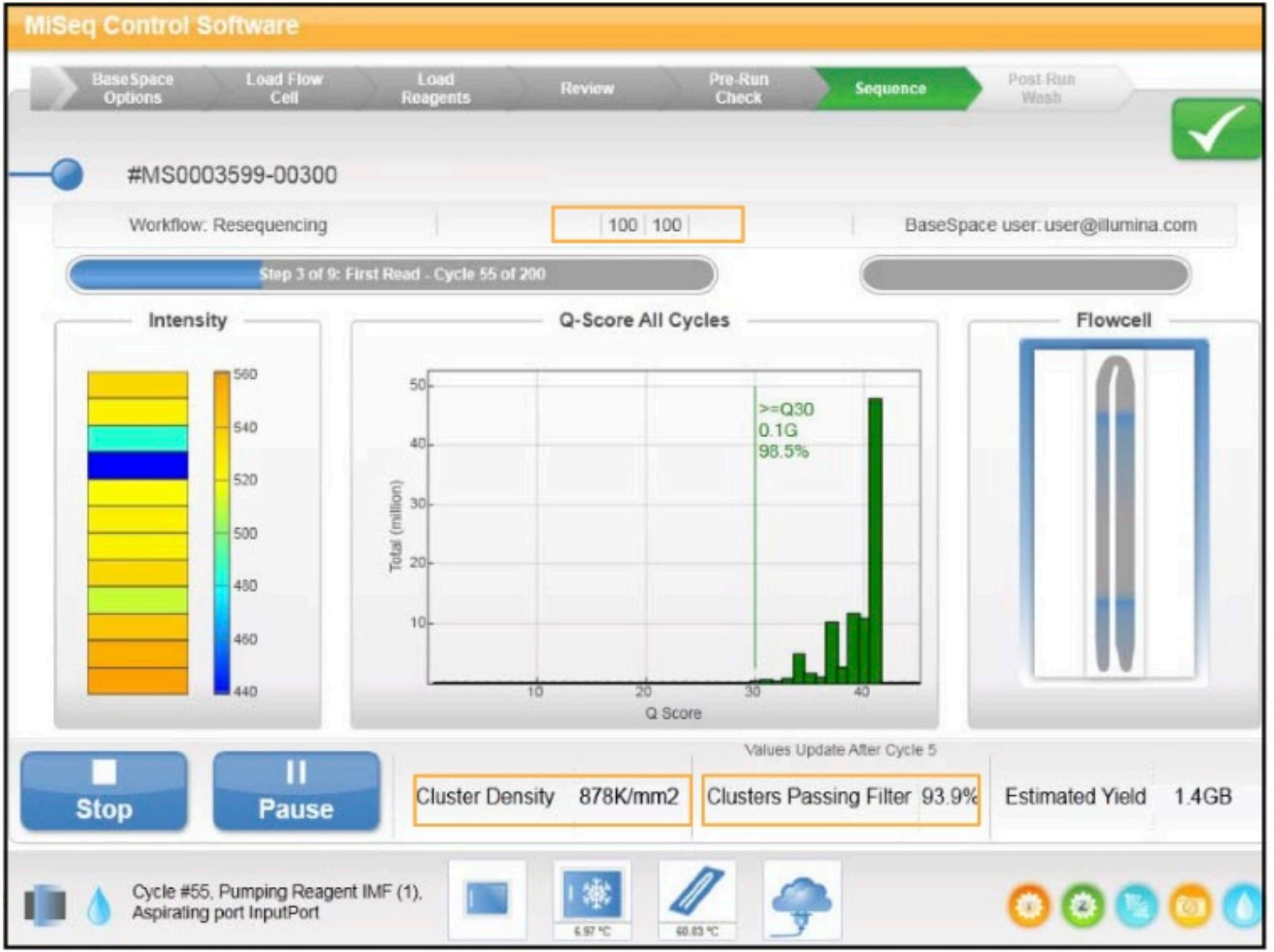
•On day of sequencing you will denature your 4nM libraries and dilute to desired molarity (generally 10-12pM for these libraries). You will do the same for the PhiX library purchased from Illumina, which you will put on every run as a control.

•Finally, pool all libraries and PhiX together in 650µl, with PhiX being 1-5% of pool and the rest of the libraries evenly represented. Load 600µl into cartridge and run.

MiSeq Output Calculations

	MiSeq with: - Upgraded hardware, or from September 2012 and later - MCS v2.3 or later - MiSeq Reagent Kit v3	MiSeq with: - Upgraded hardware, or from September 2012 and later - MCS v2.0 or later - MiSeq Reagent Kit v2
Reads/flow cell	25,000,000	16,000,000
Genome or region size (in bases)	**600,000**	**600,000**
Coverage	**100**	**100**
Total number of cycles (*e.g.* 300 for 2x150)	**150**	**50**
Total output required (in bases)	**60,000,000**	**60,000,000**
Output/flow cell (bases/flow cell)	**3,750,000,000**	**800,000,000**
Number of flow cells	**0.02**	**0.08**
Number of samples per flow cell	**62.50**	**13.33**

http://support.illumina.com/sequencing/downloads.ilmn


MiSeq Control Software
BaseSpace Options
Load Flow Cell
Load Reagents
Review
Pre-Run Check
Sequence
Post-Run Wash
#MS0003599-00300
Workflow: Resequencing
100 100
BaseSpace user: user@illumina.com
Step 3 of 9: First Read - Cycle 55 of 200
Intensity
560
540
520
500
480
460
440
Q-Score All Cycles
>=Q30
0.1G
98.5%
Total (million)
50
40
30
20
10
10
20
30
40
Q Score
Flowcell
Stop
Pause
Values Update After Cycle 5
Cluster Density 878K/mm2
Clusters Passing Filter 93.9%
Estimated Yield 1.4GB
Cycle #55, Pumping Reagent IMF (1), Aspirating port InputPort
6.97 °C
60.83 °C

BaseSpace | Dashboard | Runs | Projects | Apps | Public Data | Help | Search | Lou Bierwert | illumina

Runs » BPahangiAdultF101013

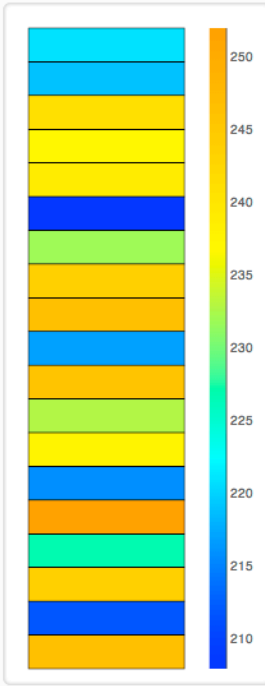
Charts

STATUS Extracted: 157 Called: 157 Scored: 157

Flowcell Chart

000000000-A5WBF Cycle 1 Bas...

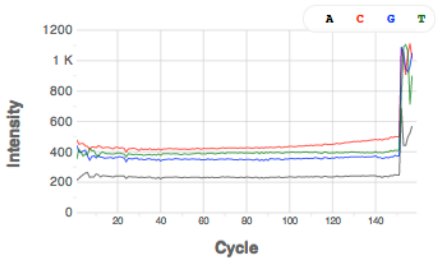
Show: Intensity
Surface: Top Surface
Cycle: Cycle 1
Base: Base A



Data By Cycle

000000000-A5WBF All Lanes All Bases

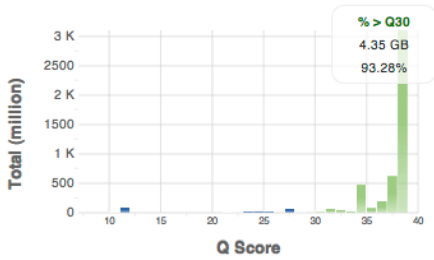
Show: Intensity
Surface: Both Surfaces
Lane: All Lanes
Base: All Bases



QScore Distribution

000000000-A5WBF All Lanes All Reads All Cycles

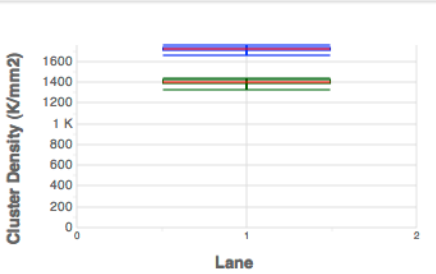
Surface: Both Surfaces
Lane: All Lanes
Read: All Reads
Cycle Up To: All Cycles



Data By Lane

000000000-A5WBF Read 1

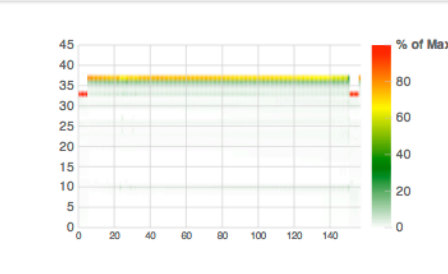
Show: Density



QScore Heatmap

000000000-A5WBF All Lanes

Surface: Both Surfaces
Lane: All Lanes



contact us

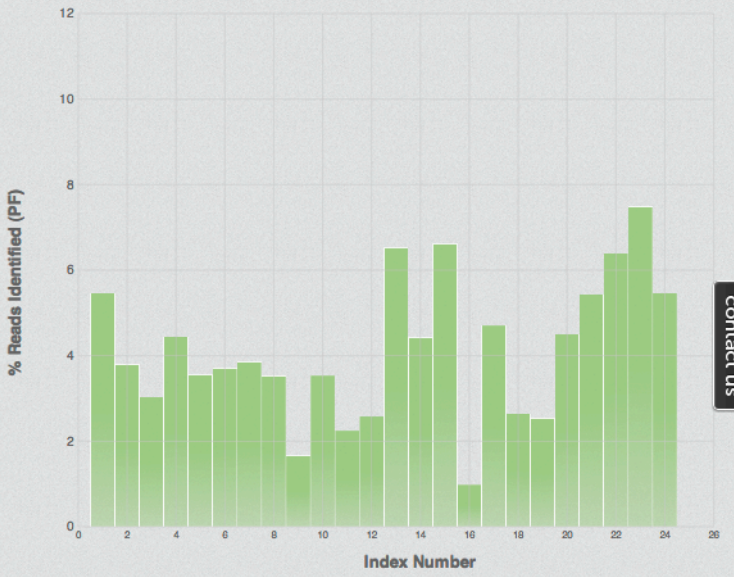
seqsystem_coverageca....zip

Show All

Reads Mapped to index ID

Totals Reads	PF Reads	% Reads Identified (PF)	CV	Min	Max
39100440	31866492	99.2045	0.4021	0.9886	7.4923

Index Number	Sample ID	Project	Index 1 (17)	Index 2 (15)	% Reads Identified (PF)
1	1	NA	ATCACG		5.4687
2	2	NA	CGATGT		3.7962
3	3	NA	TTAGGC		3.04
4	4	NA	TGACCA		4.4493
5	5	NA	ACAGTG		3.5587
6	6	NA	GCCAAT		3.7034
7	7	NA	CAGATC		3.8523
8	8	NA	ACTTGA		3.5304
9	9	NA	GATCAG		1.6697
10	10	NA	TAGCTT		3.5399
11	11	NA	GGCTAC		2.2518
12	12	NA	CTTGTA		2.5841
13	13	NA	AGTCAA		6.5291
14	14	NA	AGTTCC		4.4278
15	15	NA	ATGTCA		6.6197
16	16	NA	CCGTCC		0.9886
17	17	NA	GTCCGC		4.7116
18	18	NA	GTGAAA		2.6488



Bioinformatics on large datasets

Start by coming up with a workflow, then try to find a primary publication that did the same.

Sample workflow

QC Data -> Trim Data -> Assemble Data (if no reference de novo) -> Annotate Transcripts -> Map/Align reads to reference (either published or your new annotated de novo assembly) -> Count Reads that mapped to each transcript -> Normalize and Compare Read Counts

Bioinformatics on large datasets

Software Options For RNA-Seq

QC Data – FastQC – in Galaxy or Commandline (limits in Galaxy – data size, goes down)

Trim Data – FastQCTrimmer vs Quality Filter in Galaxy or Commandline

Assemble Data – Trinity in Galaxy or Commandline (limits in Galaxy – control – if on commandline, can normalize (eliminate repeated sequence (called digital normalisation) to give program less noise -> less memory needed, more speed for assembly)

SoapDeNovoTrans – make sure for RNA otherwise multiples reads of the same sequence will be viewed as repeats.

Annotate Transcripts – Trinotate, JAMg

Some websites: http://dnasubway.iplantcollaborative.org/
http://goblinx.soic.indiana.edu/src/yrGATE

Map/Align reads to reference – Tophat either in Galaxy or Commandline (only works if aligning to a reference genome), Bowtie, RSEM, BWA

Count Reads – Cufflinks either in Galaxy or Commandline, HTSeq, RSEM

Compare Read Counts-Use CuffMerge to bring together all data sets you want to compare and then run CuffDiff , DESeq2, eXpress, RSEM

Bioinformatics on large datasets

From here you will likely want to do:
Gene Ontology – DAVID and Cytoscape (networks), BioCyc or other functional genomic resources depending on organism (eg. EcoCyc is BioCyc for *E. coli*).

SNP detection, variant calling

There is also a very fun website called KeggAnime that gives you animation of your genes of interest in their respective pathways, http://biit.cs.ut.ee/kegganim/index.cgi

Bioinformatics on large datasets

A note on normalization:
RPKM switched to FPKM with the advent of paired end reads, making Reads per Kilobase per Million is no longer accurate as there were 2 reads per fragment. Thus, FPKM or Fragment per Kilobase per Million is now used.

HOWEVER, FPKM is a within sample normalization: it allows you to compare relative expression of genes or transcripts within a single sample to for instance plot expression of several genes in a single sample. FPKM is NOT comparable between samples because in addition to normalizing by transcript length and library size, it includes a sample-specific normalization constant. In addition, if you are comparing the expression of a gene across samples, that gene/transcript size is going to be the same across conditions (if comparing different species to each other, then FPKM may still be in play).

Thus, now it is recommended to use TPM (Transcripts per Million reads) as an equivalent metric that is comparable between samples. The math is explained beyond my understanding in a paper by Wagner:
http://lynchlab.uchicago.edu/publications/Wagner,%20Kin,%20and%20Lynch%20(2012).pdf

Cufflinks in Galaxy only lets you do FPKM – at the very least you should try one other method of analysis and compare your results.

Don't use sequencing to draw conclusions. Use sequencing to direct hypotheses that you then explore experimentally.

References and Places to Go

-Watch some training videos that pertain to you at Illumina.com http://support.illumina.com/training/sequencing_training.ilmn (Chemistry overview is wicked helpful).

-Register with Illumina.com to get your BaseSpace account and info on past and future webinars.

Integrated Genome Browsers	-UC Santa Cruz : hosts many genomes, including E.coli -IGV at Broad Institute: recommended by Illumina
Aligner	-Mauve
Mapping Viewer	-Tablet - Tview in Samtools
Gene Ontology	-DAVID -BioCyc and EcoCyc -Cytoscape -KEGG and KeggAnim

When all else fails, Google “How to” or “Manual” or “Tutorial” your problem or get an account at SeqAnswers and BioStar and ask the forums.

Some Illumina Tech Notes to Peruse

An Introduction to Illumina NGS Technology for Microbiologists:
http://res.illumina.com/documents/products/sequencing_introduction_microbiology.pdf

Estimating Sequencing coverage:
http://res.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf

Understanding Illumina Quality Scores:
http://res.illumina.com/documents/products/technotes/technote_understanding_quality_scores.pdf

And many more at:
http://support.illumina.com/sequencing/literature.ilmn