Bioinformatics Analysis Tools for NGS Data

Introduction

Philipp Rescheneder, Moritz Smolka

April 27, 2016

Center for Integrative Bioinformatics Vienna
Max F. Perutz Laboratories

CIBIV
Center for Integrative Bioinformatics Vienna

Bioinformatics

- Very Bad Things
  I've been doing bioinformatics for about 10 years now. I used to joke with a friend of mine that most of our work was converting between file formats. We don't joke about that anymore.

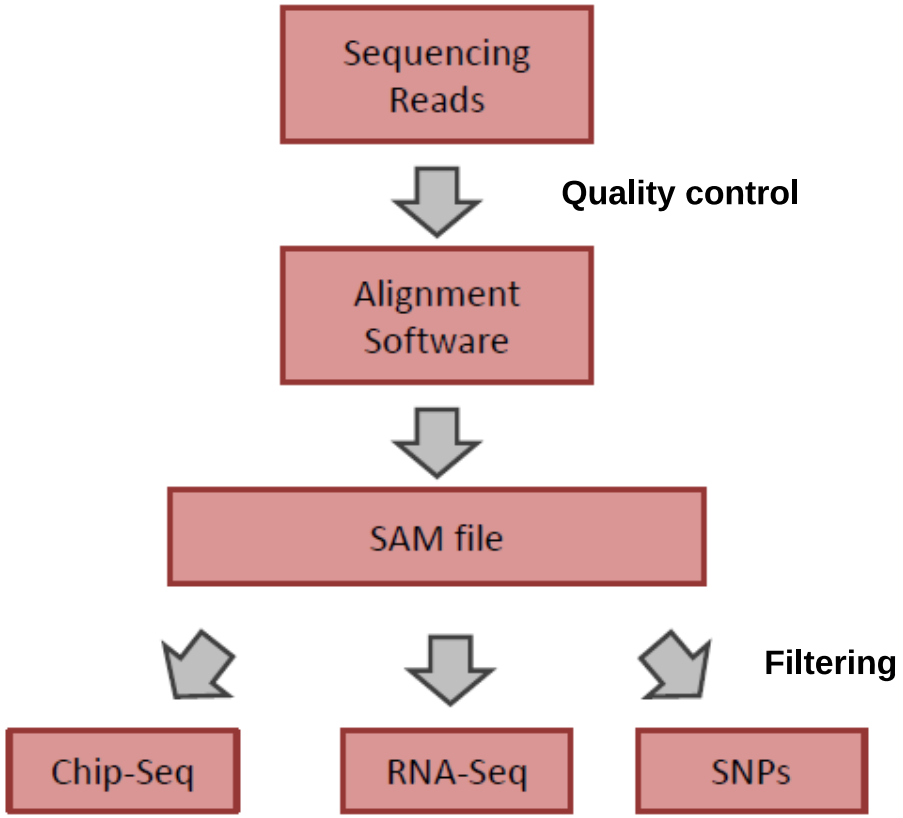
Bioinformatics

- **Very Bad Things**
  I've been doing bioinformatics for about 10 years now. I used to joke with a friend of mine that most of our work was converting between file formats. We don't joke about that anymore.
- **What Are The Most Common Stupid Mistakes In Bioinformatics?**
  Invent a new, weakly defined, internally redundant, ambiguous, bulky fruit salad of a data format. Again.

Bioinformatics

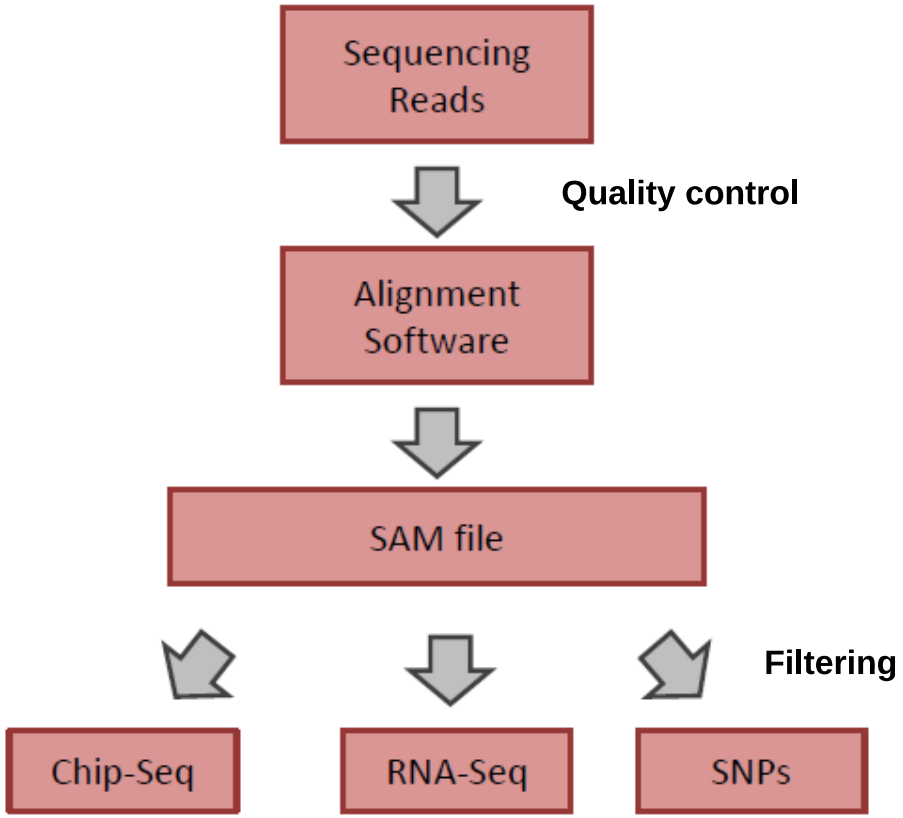
- **Very Bad Things**
  I've been doing bioinformatics for about 10 years now. I used to joke with a friend of mine that most of our work was converting between file formats. We don't joke about that anymore.
- **What Are The Most Common Stupid Mistakes In Bioinformatics?**
  Invent a new, weakly defined, internally redundant, ambiguous, bulky fruit salad of a data format. Again.
- **Why does each GO enrichment method give different results?**
  I'm new to GO terms. In the beginning it was fun, as long as I stuck to one algorithm. But then I found that there are many out there, each with its own advantages and caveats.

NGS analysis pipeline



- What we will talk about: most important file formats and tools for performing tasks that are required for all NGS experiments

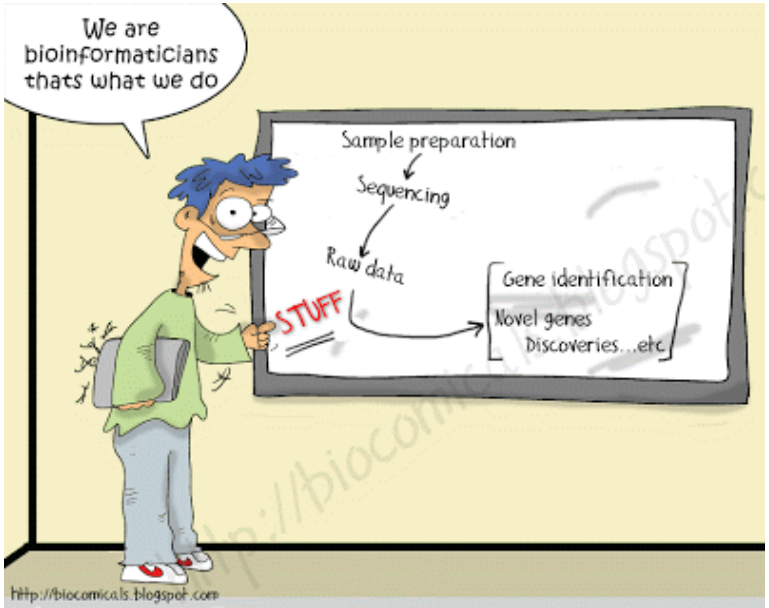
NGS analysis pipeline



- What we will talk about: most important file formats and tools for performing tasks that are required for all NGS experiments
- What we won't go into: full down-stream analysis of a specific protocol. So no R and no interpretation of specific results

Overview

- Sequencing data
- Quality control
- Read Mapping
- Working with mapped files
- Visualising mapped data
- Working with interval data
- Basics of RNA-Seq analysis
- Basics of SNP calling



Materials

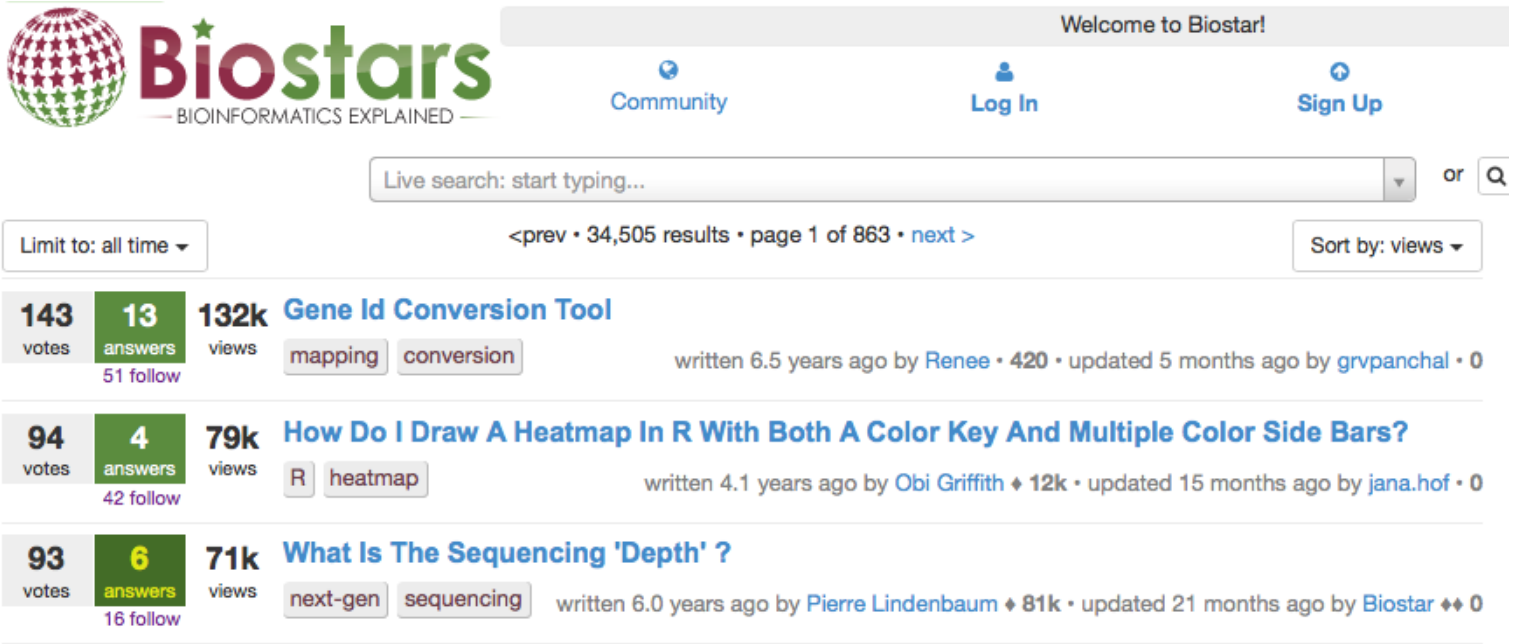
- Applied Bioinformatics course by Istvan Albert
  - `http://www.personal.psu.edu/iua1/2015_fall_852/main_2015_fall_852.html`

Materials

- Applied Bioinformatics course by Istvan Albert
  - `http://www.personal.psu.edu/iua1/2015_fall_852/main_2015_fall_852.html`
- SEQanswers
  - `http://seqanswers.com`

Materials

- Applied Bioinformatics course by Istvan Albert
  - http://www.personal.psu.edu/iua1/2015_fall_852/main_2015_fall_852.html
- SEQanswers
  - http://seqanswers.com
- Biostart
  - http://www.biostars.org

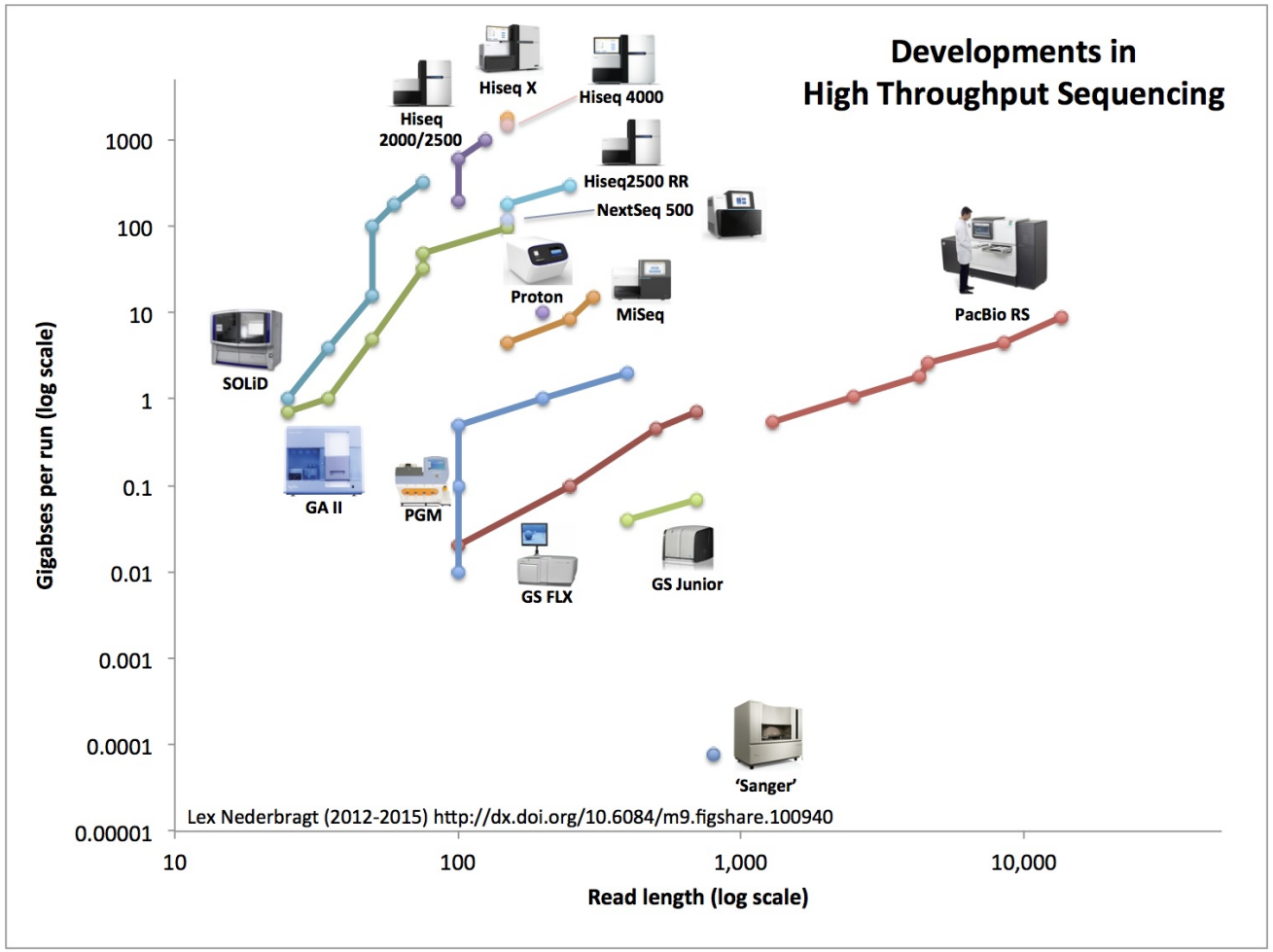


DNA Sequencing

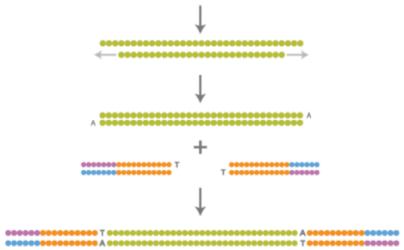
DNA sequencing is the process of **determining the precise order of nucleotides within a DNA molecule**. It includes any method or technology that is used to determine the order of the four bases (adenine, guanine, cytosine, and thymine) in a strand of DNA. The advent of rapid DNA sequencing methods has greatly accelerated biological and medical research and discovery.

A **sequencing error** or mis-call occurs when a sequencing method calls **one or more bases incorrectly**, leading to an inaccurate **read**. Due to the vagaries of molecular biology, no laboratory-based DNA sequencing methods are perfectly precise; they are all known to mis-call bases occasionally in the machines.

Sequencing platforms



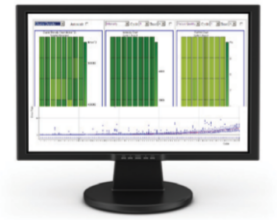
Illumina sequencing overview



Library Preparation
~2 h [15 min hands-on (Nextera)]
< 6 h [< 3 h hands-on (TruSeq)]

Cluster Generation
~5 h (<10 min hands-on)

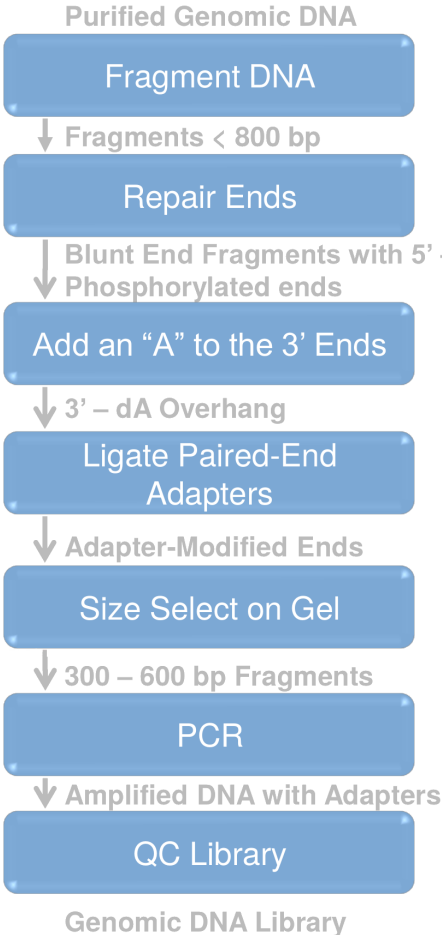
Sequencing by Synthesis
~1.5 to 11 days



CASAVA
2 days (30 min hands-on)

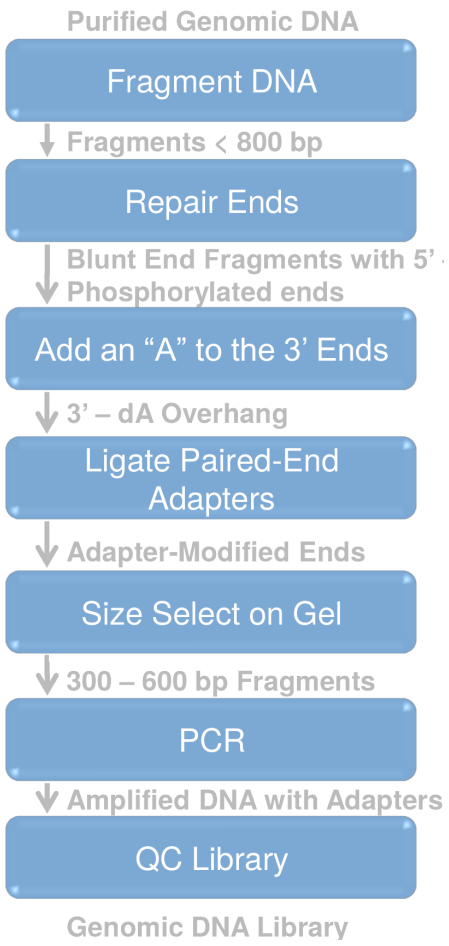
Library preparation

- Prepares sample nucleic acid for sequencing
  - Fragmenting
  - Generates double-stranded DNA (if necessary)
  - Flanks with Illumina adapters



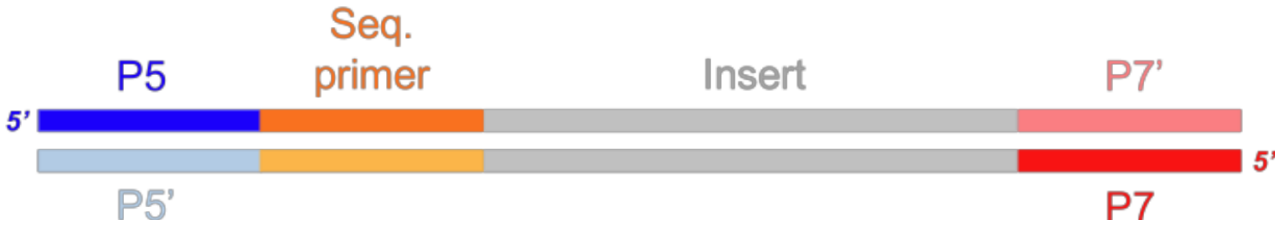
Library preparation

- Prepares sample nucleic acid for sequencing
  - Fragmenting
  - Generates double-stranded DNA (if necessary)
  - Flanks with Illumina adapters
- All preparation ends with the same general template structure
  - Double-stranded DNA flanked by adapters
  - Variables include: Insert Size, Adaptor type, Index
  - Fragmentation method might even influence down-stream analysis



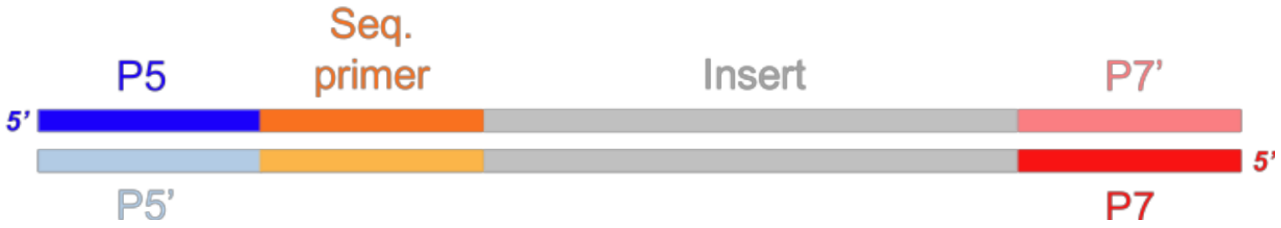
Adapter types (I)

- Single end sequencing:



Adapter types (I)

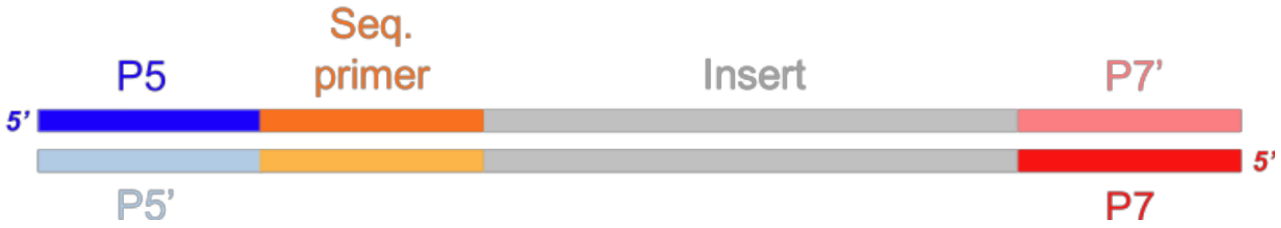
- Single end sequencing:



  - *P5 and P7:* bind to oligos on flow cell

Adapter types (I)

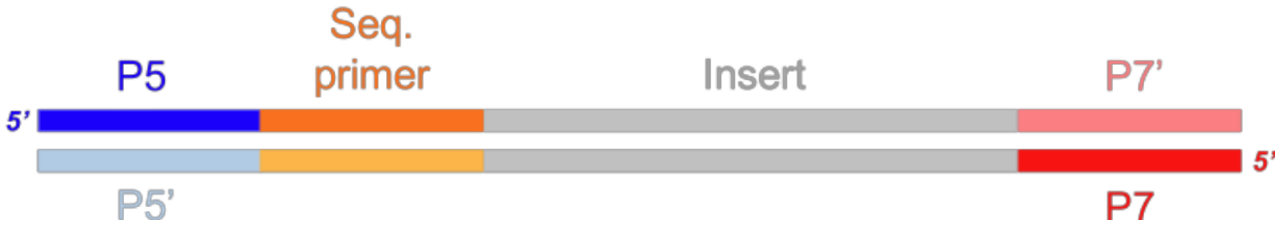
- Single end sequencing:



  - *P5 and P7:* bind to oligos on flow cell
  - *Sequencing primer:* starting point for sequencing

Adapter types (I)

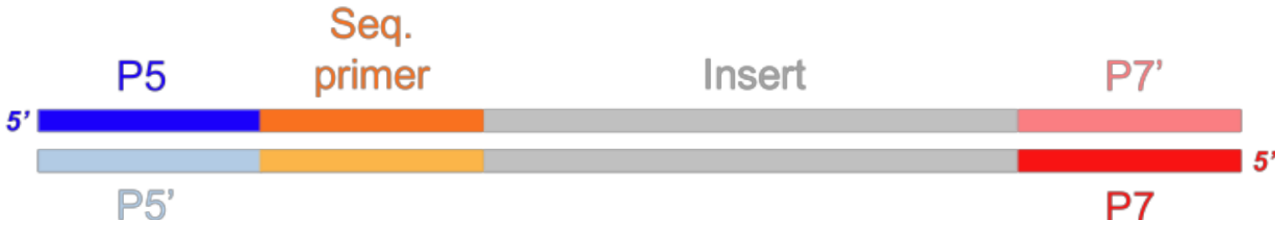
- Single end sequencing:



  - *P5 and P7:* bind to oligos on flow cell
  - *Sequencing primer:* starting point for sequencing
  - *Insert:* DNA that will be (partly) sequenced

Adapter types (I)

- Single end sequencing:



  - *P5 and P7:* bind to oligos on flow cell
  - *Sequencing primer:* starting point for sequencing
  - *Insert:* DNA that will be (partly) sequenced

- Applications: ChIP or low coverage resequencing projects

Adapter types (II)

- Paired end sequencing:

  - Applications: Most applications, #1 whole genome shotgun assembly

Adapter types (II)

- Paired end sequencing:

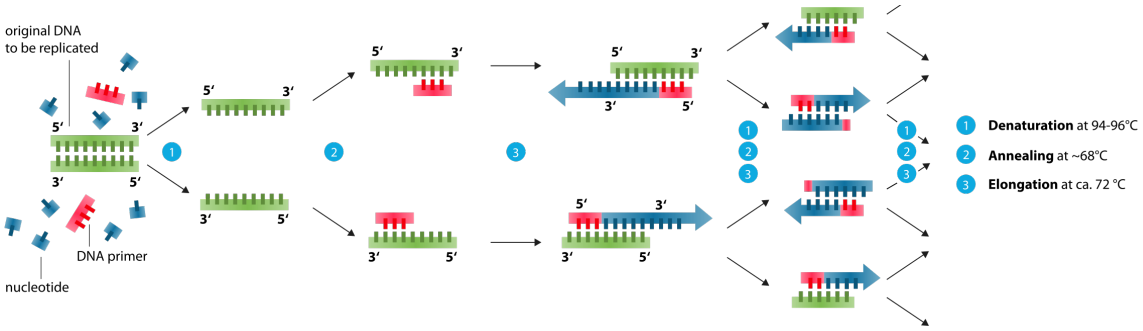
  - Applications: Most applications, #1 whole genome shotgun assembly

- Multiplex paired end sequencing:

  - Allows multiple libraries per lane (12 Index tags available x 8 lanes = 96 libraries per flowcell)

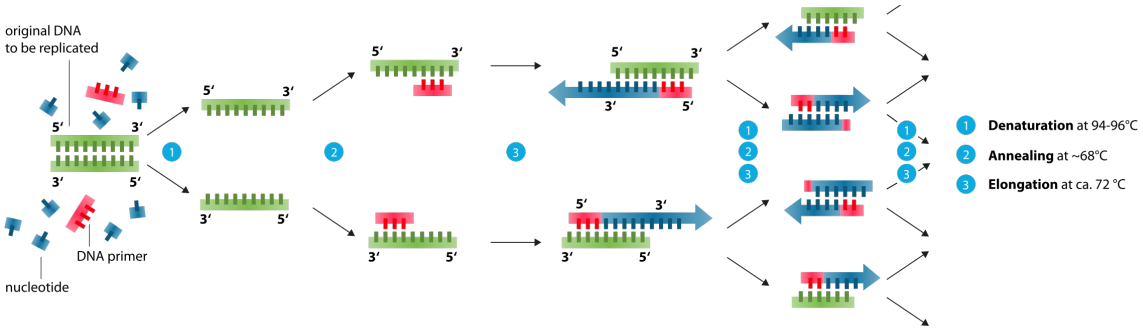
PCR amplification

Polymerase chain reaction: A technology used to amplify a single copy of a piece of DNA across several orders of magnitude, generating thousands to millions of copies.



PCR amplification

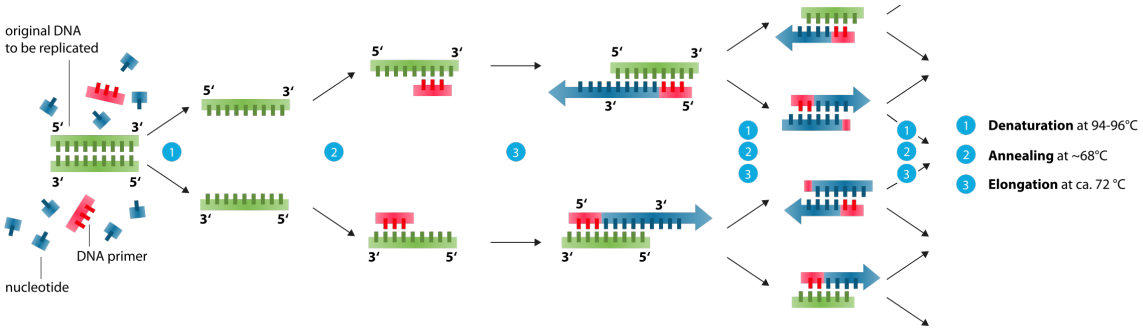
Polymerase chain reaction: A technology used to amplify a single copy of a piece of DNA across several orders of magnitude, generating thousands to millions of copies.



- Selectively enrich DNA fragments with adapter molecules on both ends

PCR amplification

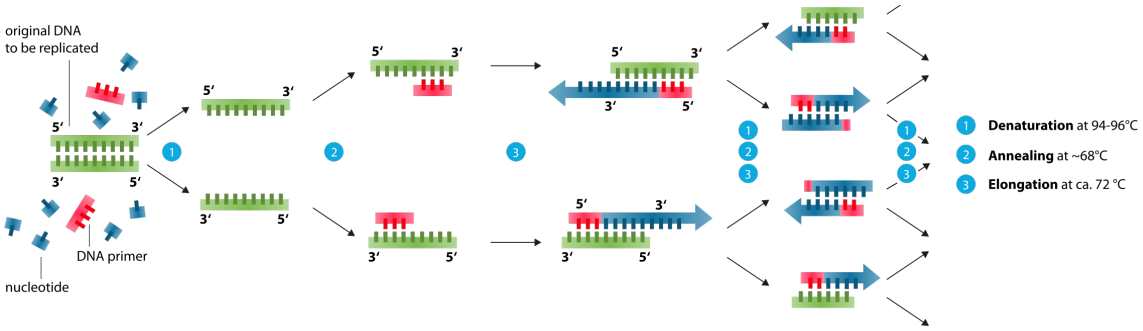
Polymerase chain reaction: A technology used to amplify a single copy of a piece of DNA across several orders of magnitude, generating thousands to millions of copies.



- Selectively enrich DNA fragments with adapter molecules on both ends
- Amplifies the amount of DNA in the library

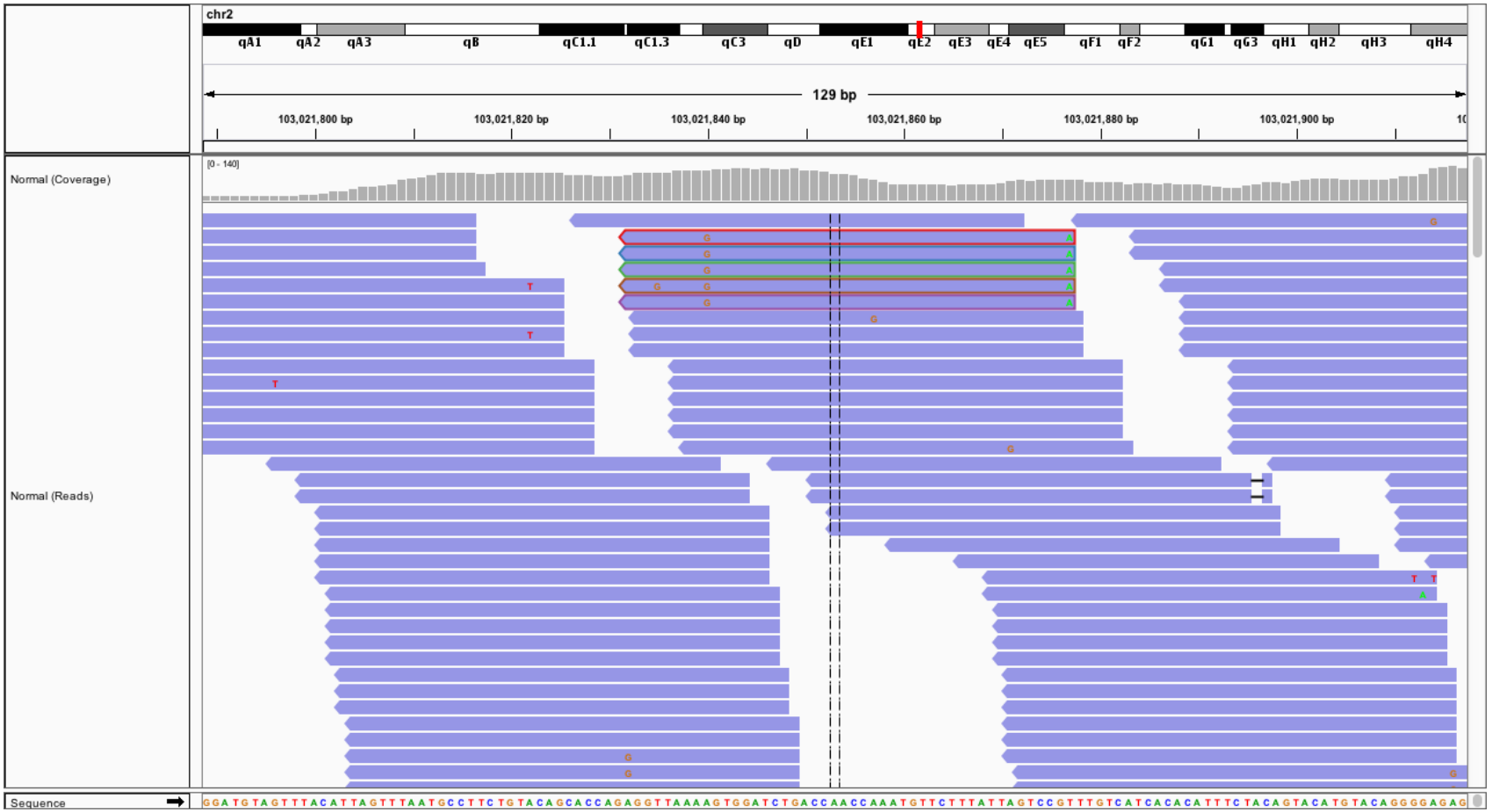
PCR amplification

Polymerase chain reaction: A technology used to amplify a single copy of a piece of DNA across several orders of magnitude, generating thousands to millions of copies.

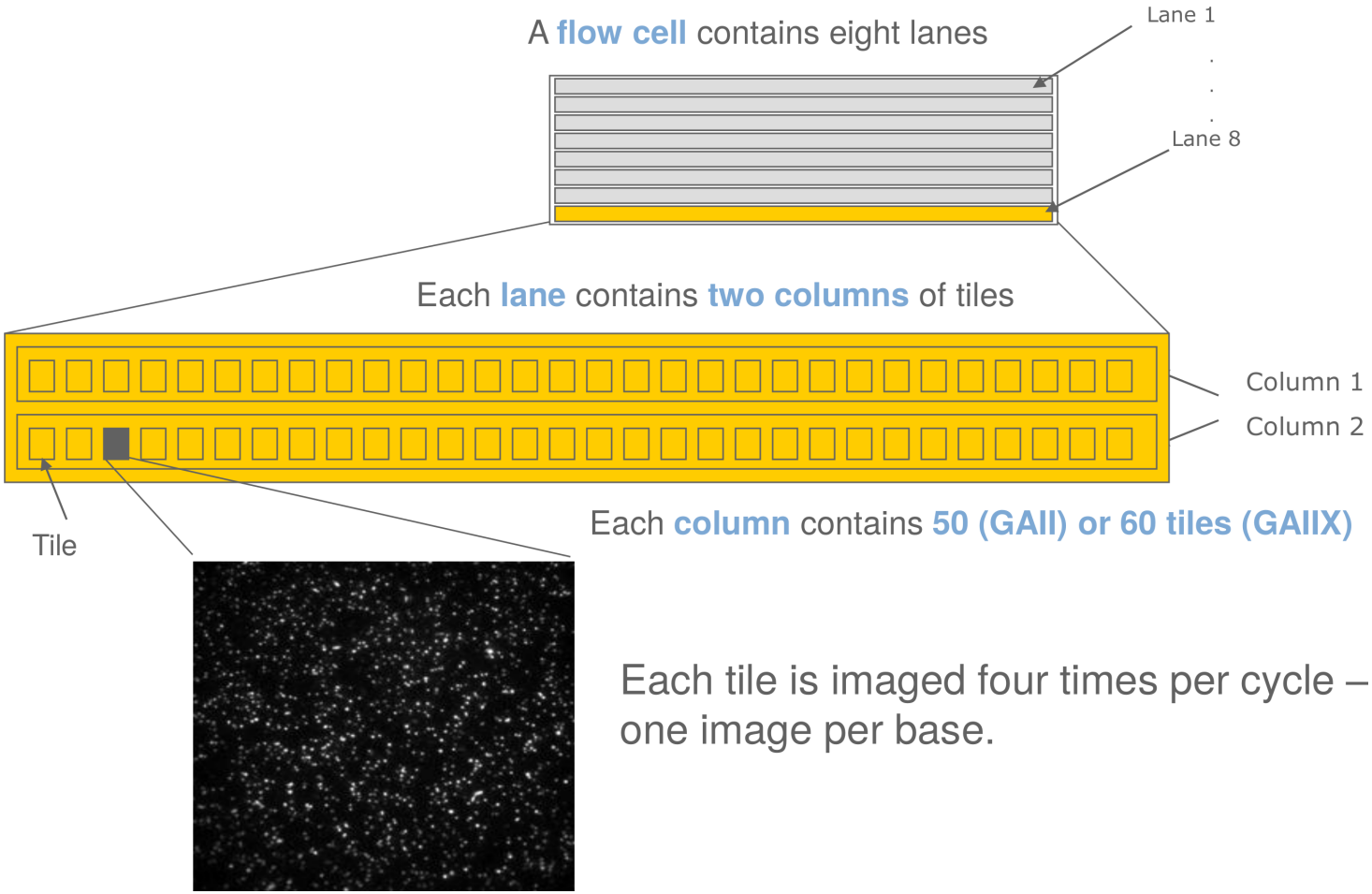


- Selectively enrich DNA fragments with adapter molecules on both ends
- Amplifies the amount of DNA in the library
- **Might introduce amplification bias or chimeric sequences**

PCR DUPLICATES



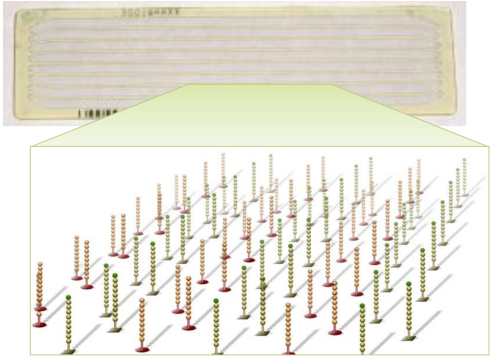
Flow cell



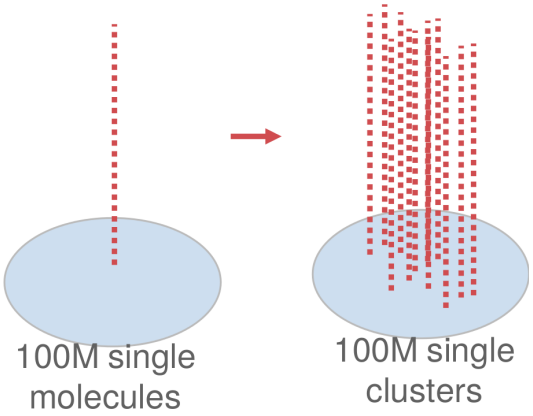
Each tile is imaged four times per cycle – one image per base.

Cluster generation

- Cluster Generation turns libraries into clonal clusters on a flow cell

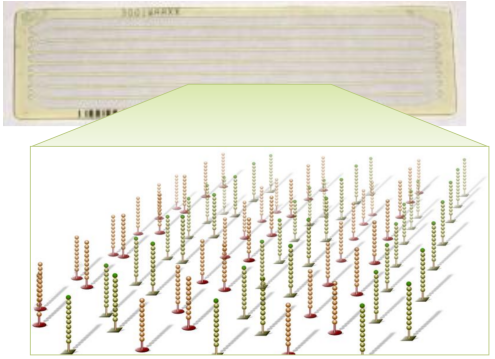


Surface of flow cell coated with a lawn of oligo pairs

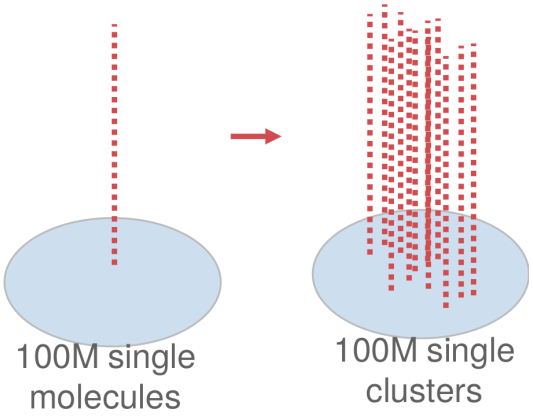


Cluster generation

- Cluster Generation turns libraries into clonal clusters on a flow cell
- This is done using a process called bridge amplification

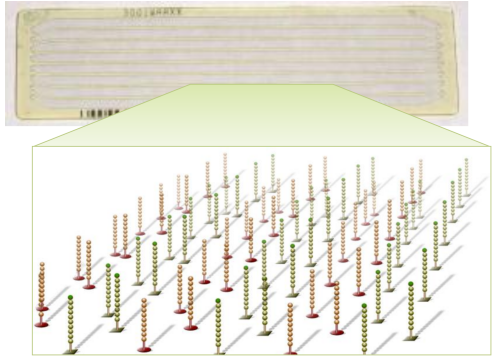


Surface of flow cell coated with a lawn of oligo pairs

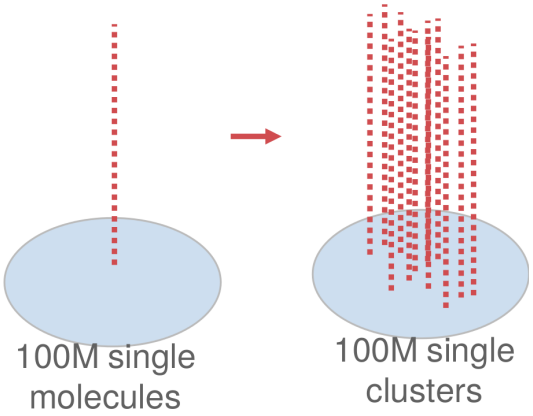


Cluster generation

- Cluster Generation turns libraries into clonal clusters on a flow cell
- This is done using a process called bridge amplification
- Massively parallel

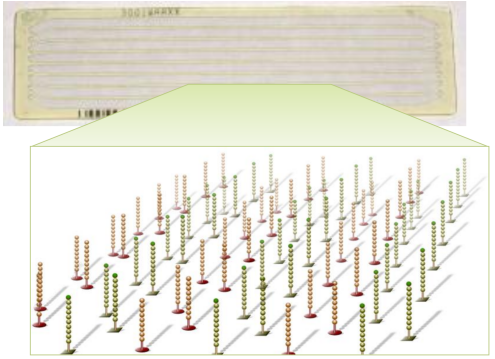


Surface of flow cell coated with a lawn of oligo pairs

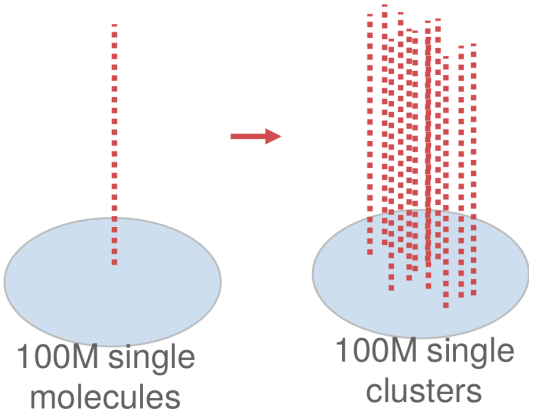


Cluster generation

- Cluster Generation turns libraries into clonal clusters on a flow cell
- This is done using a process called bridge amplification
- Massively parallel
- Cluster Station/cBot delivers fluidics and controls temperature

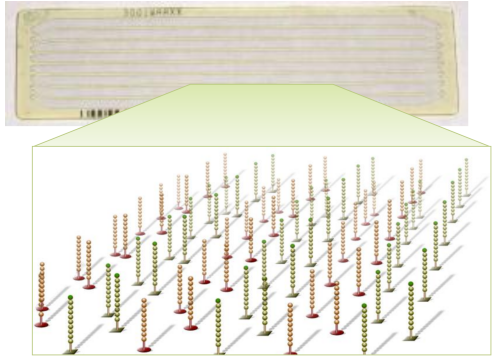


Surface of flow cell coated with a lawn of oligo pairs

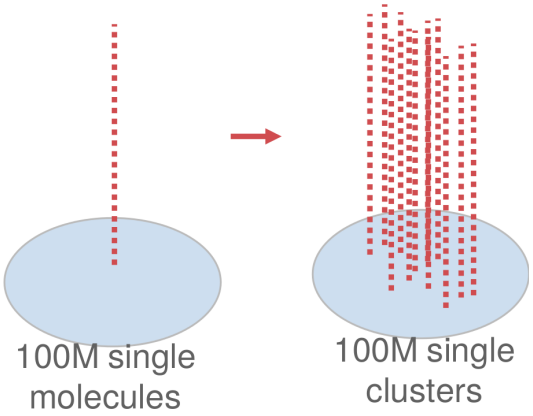


Cluster generation

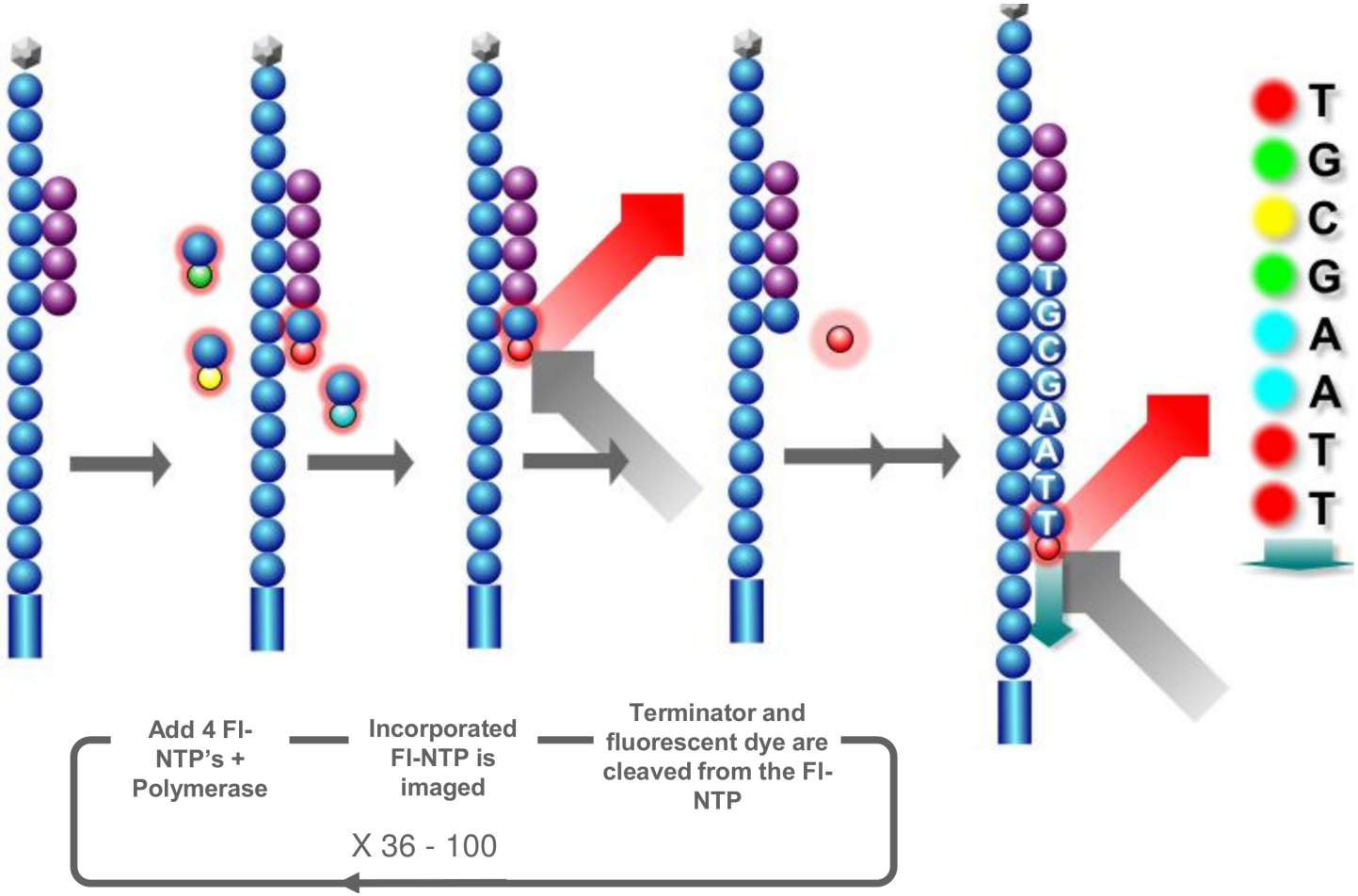
- Cluster Generation turns libraries into clonal clusters on a flow cell
- This is done using a process called bridge amplification
- Massively parallel
- Cluster Station/cBot delivers fluidics and controls temperature
- Sampling process, might introduce bias



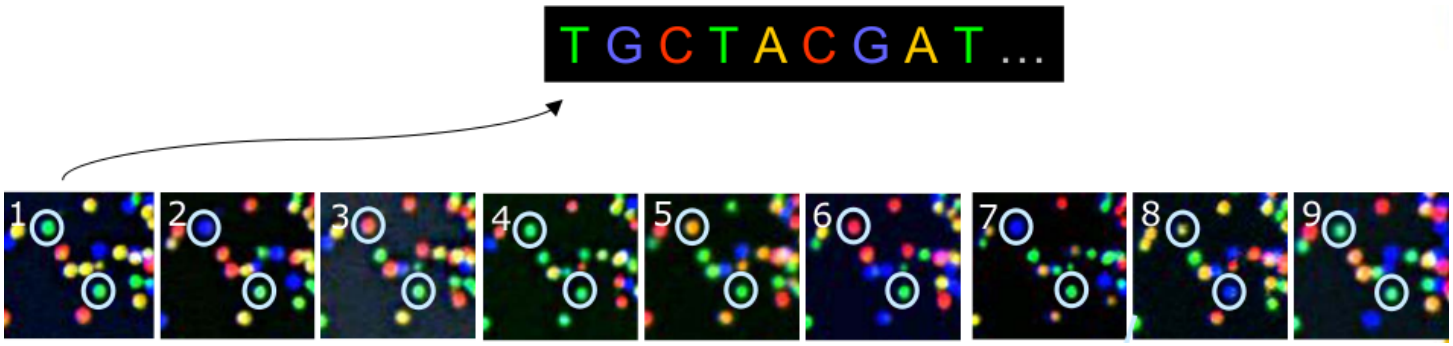
Surface of flow cell coated with a lawn of oligo pairs



Sequencing by synthesis

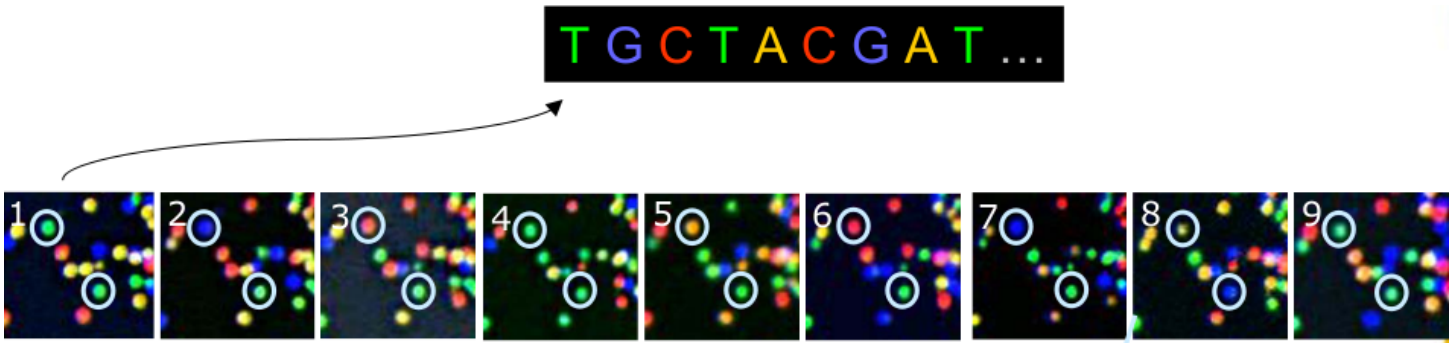


Base calling



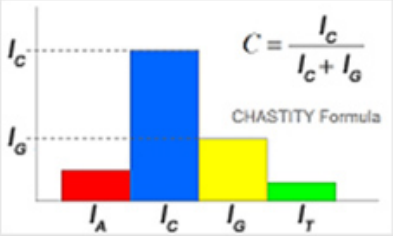
- Fluorescence signals are converted into sequence data

Base calling



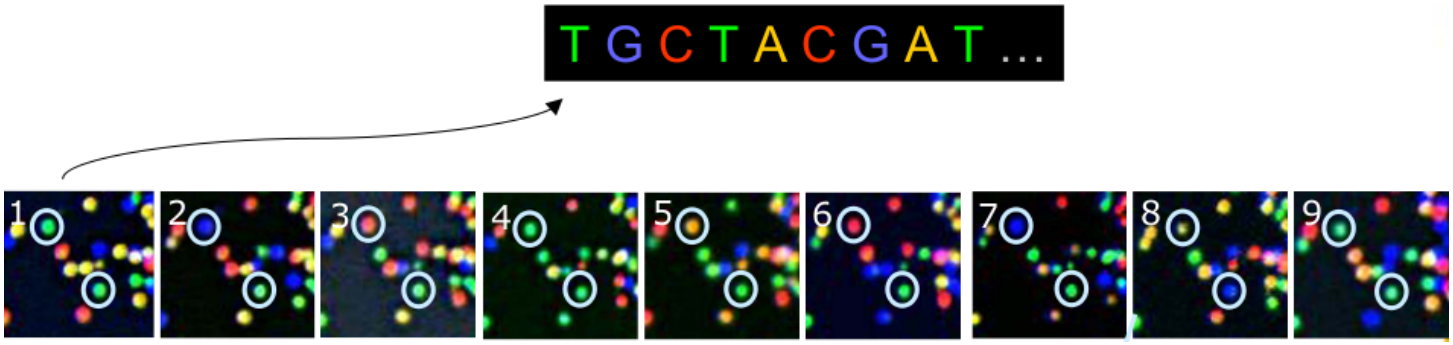
- Fluorescence signals are converted into sequence data
- Problems (phasing, fading, ...) result in **sequencing error (mostly mismatches)**

Solexa CHASTITY filtering: Individual bases generated from original image files have quality scores which reflect the probability that a base-call is correct (or wrong), this is quantified by CHASTITY Formula (as shown in the figure below).



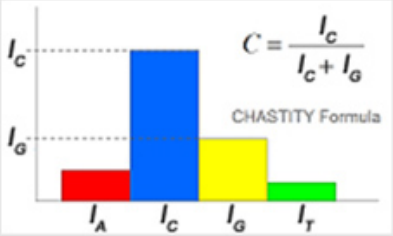
The chastity(C) of each base in the short reads is determined by the intensity of four colors (I_A, I_C, I_G, I_T here), the formula "the ratio of the highest (I_C here) of the four (base type) intensities to the sum of highest two(I_C and I_G here)." should be no less than 0.6 in the first 25 bases.

Base calling



- Fluorescence signals are converted into sequence data
- Problems (phasing, fading, ...) result in **sequencing error** (**mostly mismatches**)
- **Quality score** that reflects the probability that the base-call is correct

Solexa CHASTITY filtering: Individual bases generated from original image files have quality scores which reflect the probability that a base-call is correct (or wrong), this is quantified by CHASTITY Formula (as shown in the figure below).

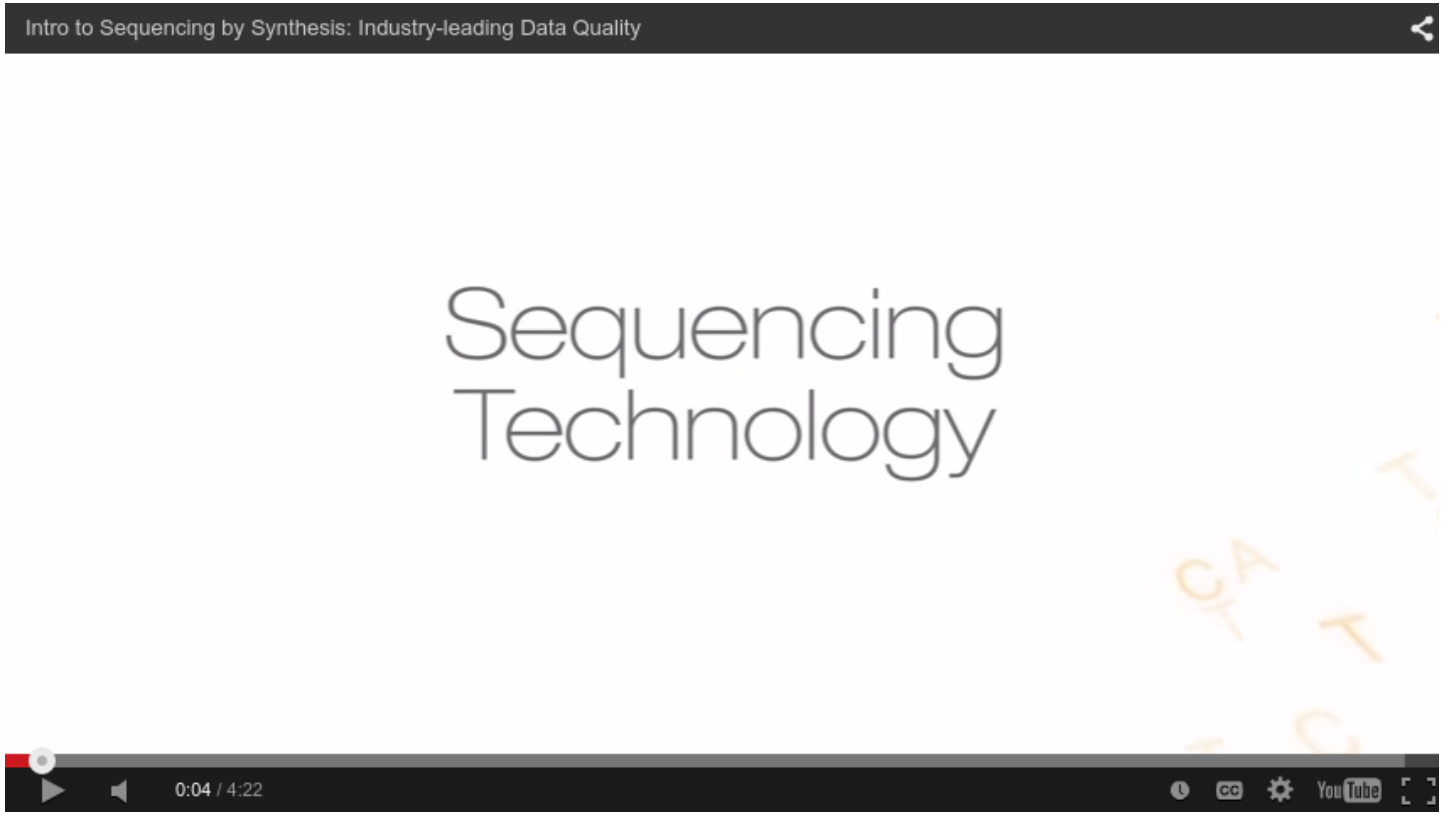


The chastity(C) of each base in the short reads is determined by the intensity of four colors (I_A, I_C, I_G, I_T here), the formula "the ratio of the highest (I_C here) of the four (base type) intensities to the sum of highest two(I_C and I_G here)." should be no less than 0.6 in the first 25 bases.

HiSeq performance metrics

	 MiniSeq System	 MiSeq Series	 NextSeq Series	 HiSeq Series	 HiSeq X Series*
Key Methods	Amplicon, targeted RNA, small RNA, and targeted gene panel sequencing.	Small genome, amplicon, and targeted gene panel sequencing.	Everyday exome, transcriptome, and targeted resequencing.	Production-scale genome, exome, transcriptome sequencing, and more.	Population- and production-scale whole-genome sequencing.
Maximum Output	7.5 Gb	15 Gb	120 Gb	1500 Gb	1800 Gb
Maximum Reads per Run	25 million	25 million†	400 million	5 billion	6 billion
Maximum Read Length	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp
Run Time	4–24 hours	4–55 hours	12–30 hours	<1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)	<3 days
Benchtop Sequencer	Yes	Yes	Yes	No	No
System Versions	• MiniSeq System for low-throughput targeted DNA and RNA sequencing	• MiSeq System for targeted and small genome sequencing • MiSeq FGx System for forensic genomics • MiSeqDx System for molecular diagnostics	• NextSeq 500 System for everyday genomics • NextSeq 550 System for both sequencing and cytogenomic arrays	• HiSeq 3000/HiSeq 4000 Systems for production-scale genomics • HiSeq 2500 Systems for large-scale genomics	• HiSeq X Five System for production-scale whole-genome sequencing • HiSeq X Ten System for population-scale whole-genome sequencing

Summary



- For more information see: GA Boot Camp

Future development: Nano Pore sequencing