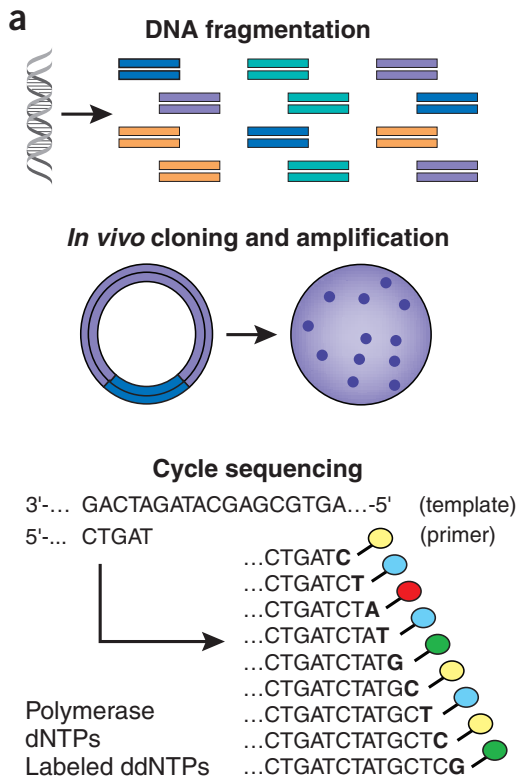
Next Generation Sequencing
Short introduction

David Gomez-Cabrero

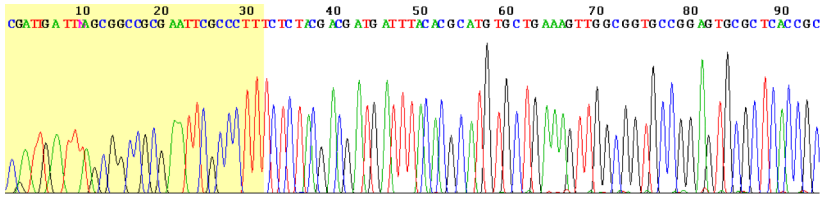
NGS: Short Introduction

History | **NGS** | **Bioinformatics**



Shendure and Ji, 2008, Nat. Biotech.

Sanger Sequencing

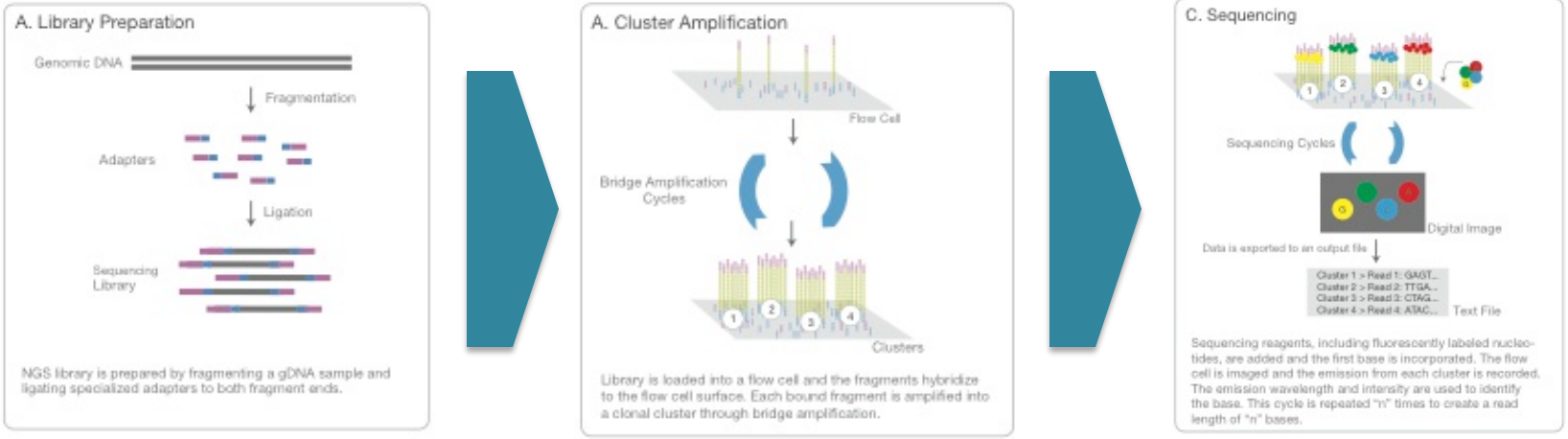


start of an example dye-terminator read

https://en.wikipedia.org/wiki/User:Loris

NGS: Short Introduction

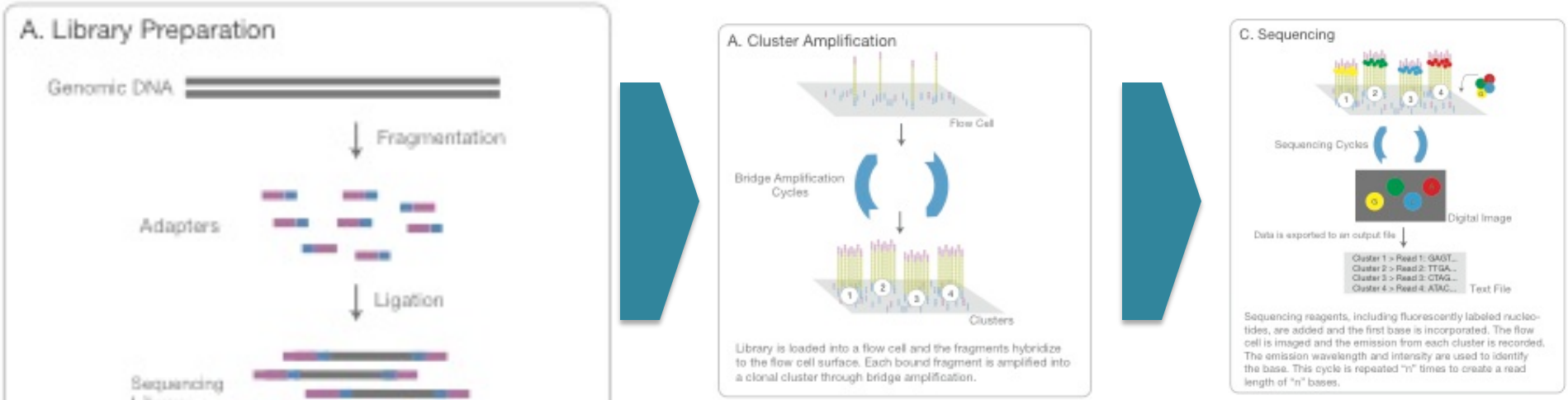
History | **NGS** | **Bioinformatics**



www.Illumina.com

NGS: Short Introduction

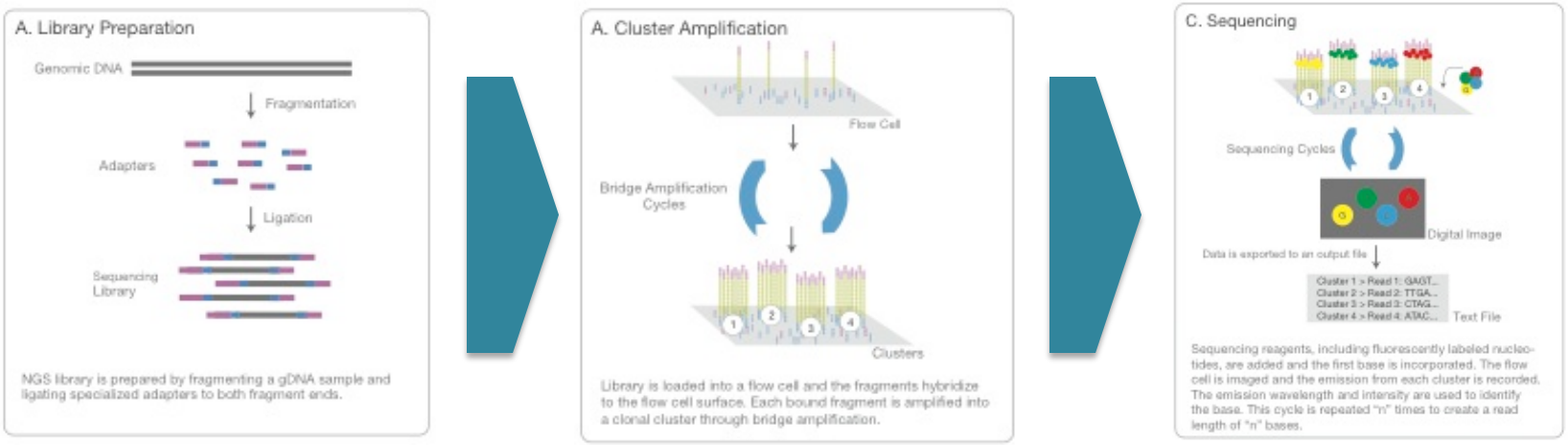
History | **NGS** | **Bioinformatics**

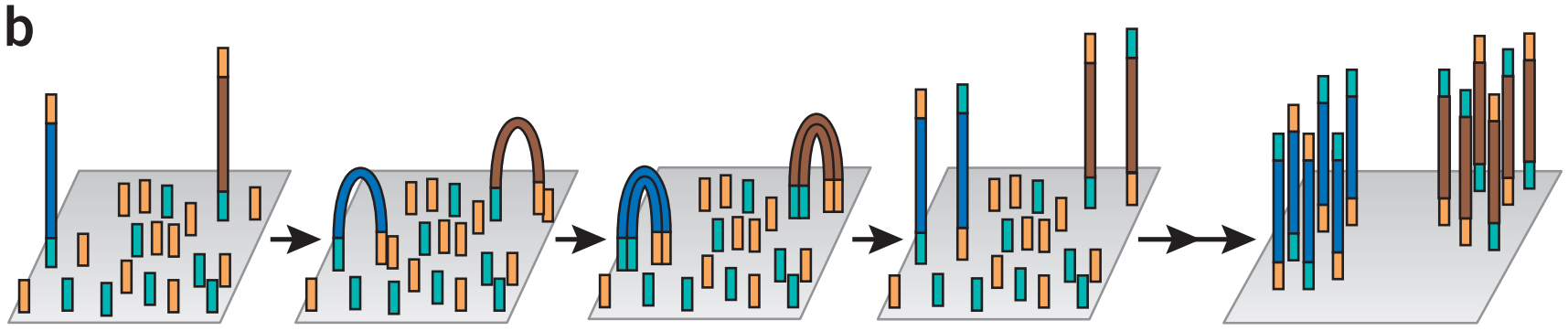


www.Illumina.com

NGS: Short Introduction

History | **NGS** | **Bioinformatics**

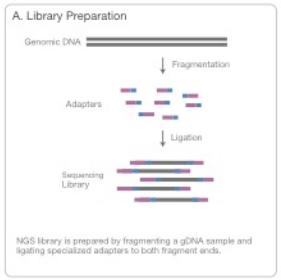


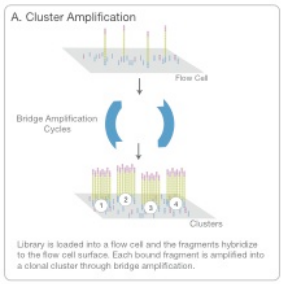


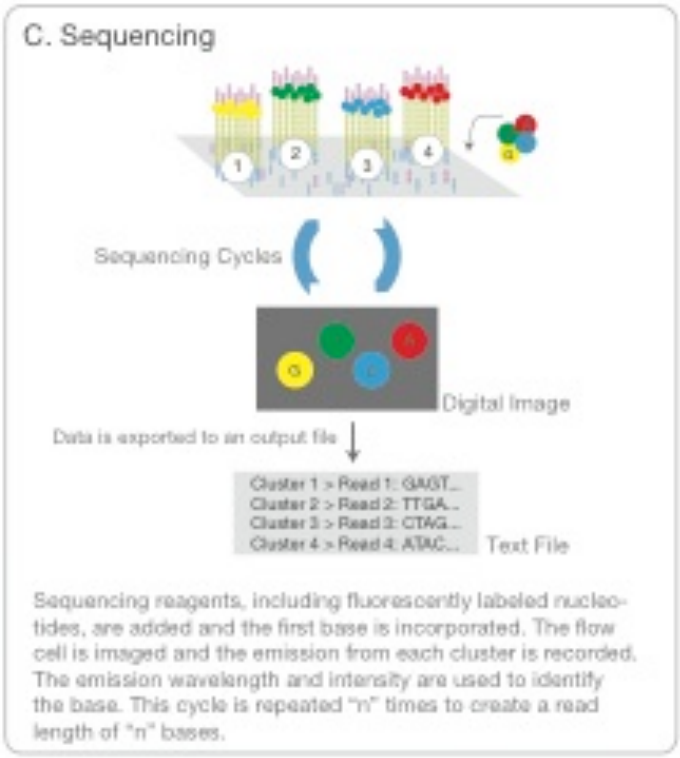
Shendure and Ji,
2008, Nat. Biotech.

NGS: Short Introduction

History | **NGS** | **Bioinformatics**

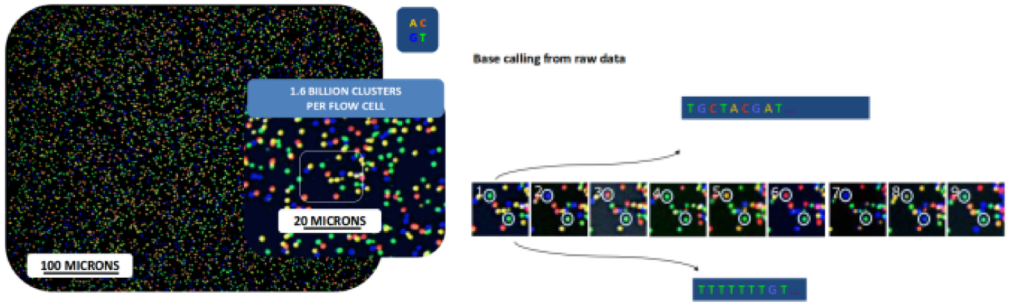






AT EACH CLUSTER

GLOBAL IMAGE



https://monsterbashseq.wordpress.com/2014/01/14/workshop-on-genomics-notes-day-2-sequencing-technologies/

NGS: Short Introduction

History | NGS | Bioinformatics

MULTIPLEXING

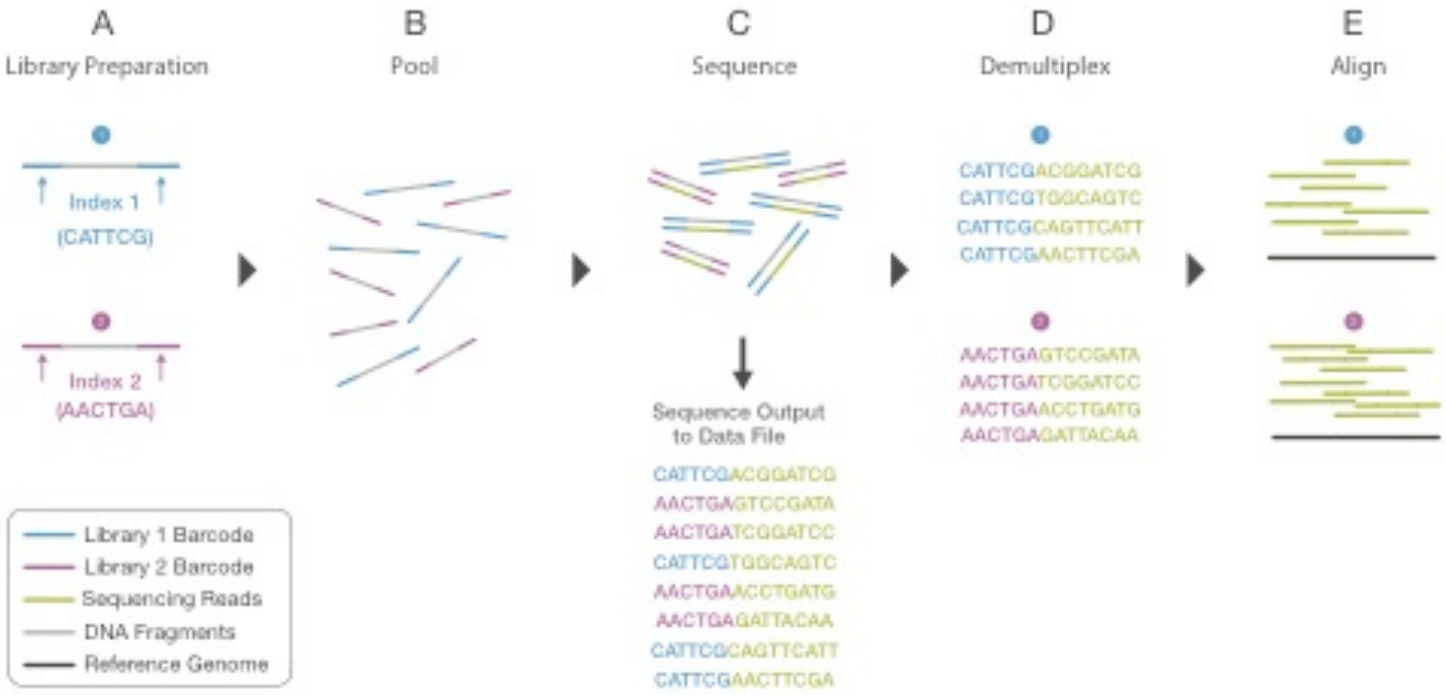


Figure 5: Library Multiplexing Overview.

a. Two distinct libraries are attached to unique index sequences. Index sequences are attached during library preparation.

b. Libraries are pooled together and loaded into the same flow cell lane.

c. Libraries are sequenced together during a single instrument run. All sequences are exported to a single output file.

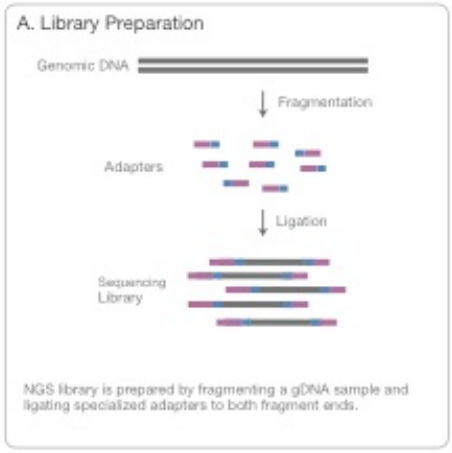
d. A demultiplexing algorithm sorts the reads into different files according to their indexes.

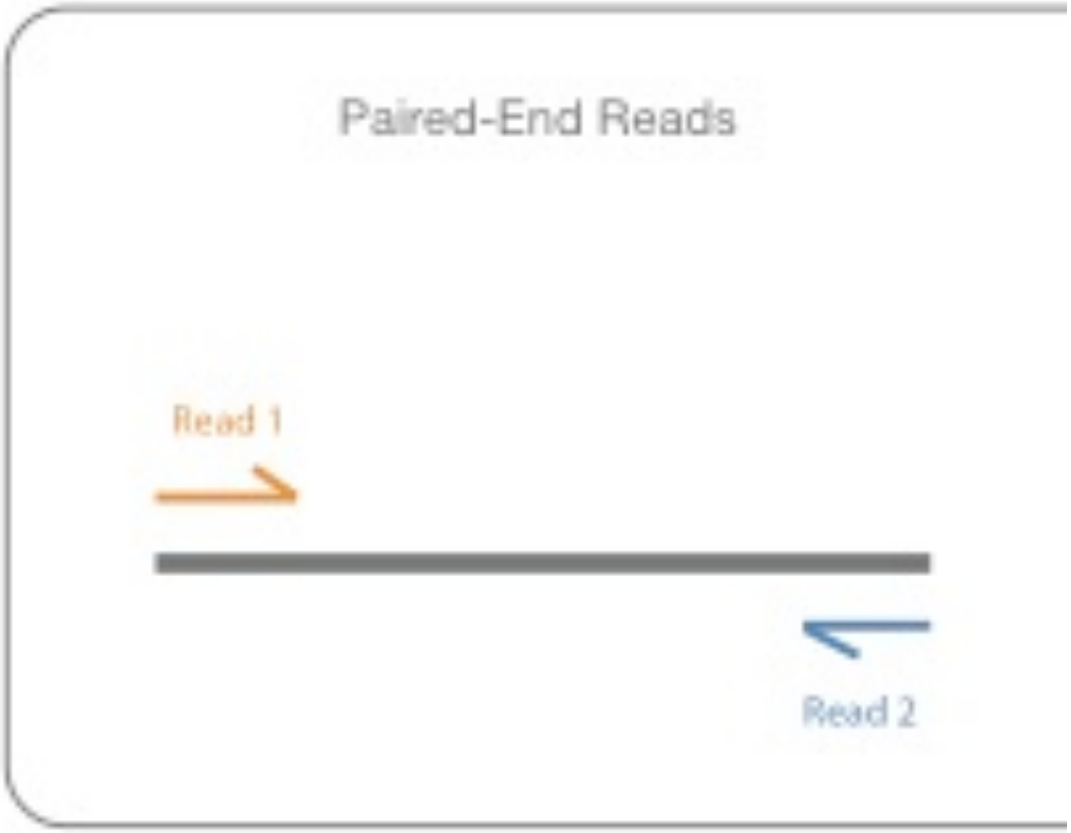
e. Each set of reads is aligned to the appropriate reference sequence.

NGS: Short Introduction

History | NGS | Bioinformatics

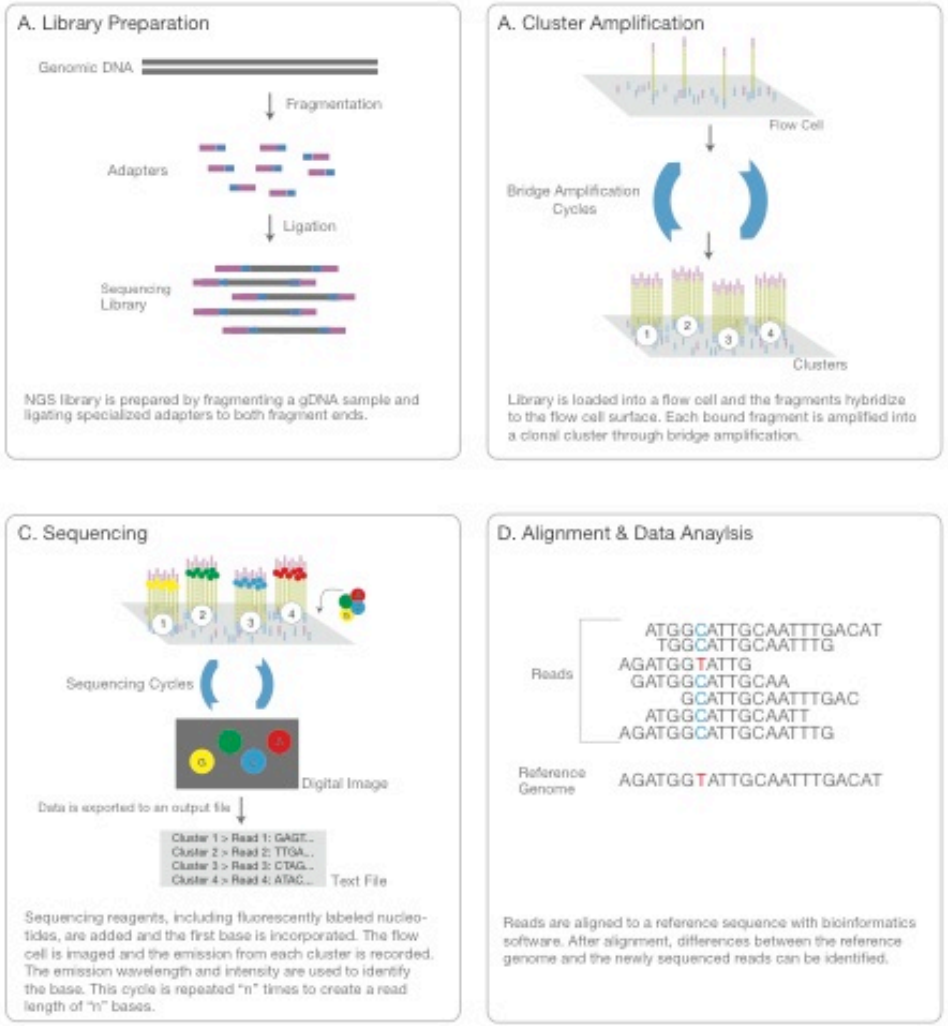
PAIR-END vs SINGLE-END





NGS: Short Introduction

History | **NGS** | **Bioinformatics**

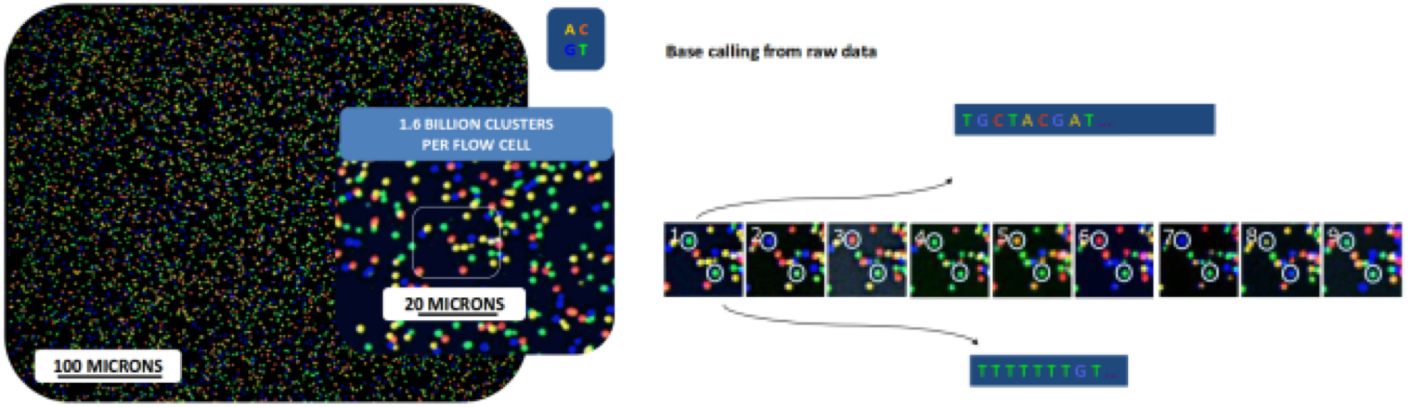


Where does bioinformatics starts?

ANALYSIS

NGS: Short Introduction

History | **NGS** | **Bioinformatics**



ILLUMINA BASE CALL FILE: BCL

This part is hidden for the general user because the output provided by NGS facilities is directly sequences

bcl2fastq Conversion
User Guide

Version 1.8.4
FOR RESEARCH USE ONLY

CASAVA

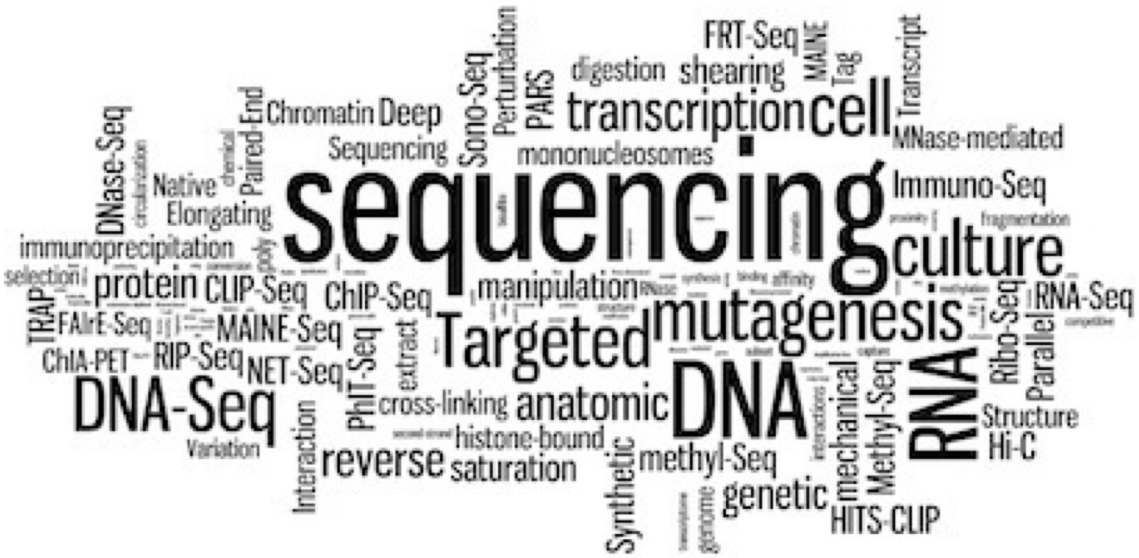
At Illumina, we are a dedicated to empowering your research through the continuous development of the most advanced genetic analysis tools. This rapid innovation occasionally results in product discontinuation, as we develop newer, more cutting-edge solutions. **We are discontinuing distribution of CASAVA software, and will continue its support until December 31, 2015. We remain committed to providing you with the highest-quality support and service during this transition.**

As a replacement for CASAVA, we have added functionality to the BaseSpace Isaac Enrichment and Isaac Whole Genome Sequencing apps available on BaseSpace and BaseSpace Onsite. In addition, the bcl2fastq Conversion Software is available to demultiplex and convert BCL files to FASTQ files on your local computer hardware. Illumina is committed to providing easy-to-operate, seamless solutions for the analysis of genomic data. BaseSpace and BaseSpace Onsite offer fully supported software solutions and are continuously optimized through ongoing development efforts.

NGS: Short Introduction

History | NGS | Bioinformatics

GENERAL PIPELINE?



DIFFERENT PIPELINES FOR EACH PROTOCOL

SHARED PARTS

RNA-seq

miRNA-seq

RRBS-seq

DNase-seq

single-cell RNA-seq

NGS: Short Introduction

History | NGS | Bioinformatics

All-seq

BCL to FASTQ

QUALITY CONTROL

How many clusters?

How many / Percentage fragments that have been identified?

…

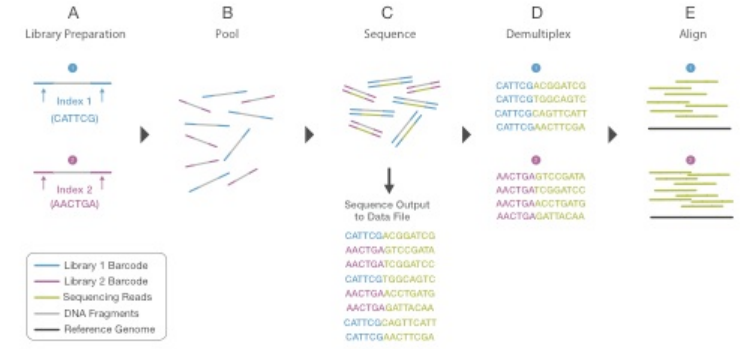


Figure 5: Library Multiplexing Overview.
a. Two distinct libraries are attached to unique index sequences. Index sequences are attached during library preparation.
b. Libraries are pooled together and loaded into the same flow cell lane.
c. Libraries are sequenced together during a single instrument run. All sequences are exported to a single output file.
d. A demultiplexing algorithm sorts the reads into different files according to their indexes.
e. Each set of reads is aligned to the appropriate reference sequence.

NGS: Short Introduction

History | NGS | Bioinformatics

All-seq

BCL to FASTQ

QUALITY CONTROL

FASTA FORMAT

```
> sequence: and information of it
ATCGATCGATCGATCGATCGATCG
```

Name, Details

Sequence

FASTQ FORMAT

```
@sequence: and information of it
GATTTGGGGTTCAAAG
+
!''*((((***+))%%
```

QUALITY VALUES FOR NUCLEOTIDES

NGS: Short Introduction

History | NGS | Bioinformatics

FASTQ FORMAT

All-seq

@sequence: and information of it
GATTTGGGGTTCAAAG
+
!''*((((***+))%%

QUALITY VALUES FOR NUCLEOTIDES

p = *probability that the corresponding base-call is incorrect*

From this is computed a Phred quality score:

$$Q_{sanger} = -10log_{10}(p)$$

This goes from 0 to 40-41, the larger the better

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
  ..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126
  0........................26...31.......40
                           -5....0........9.............................40
                                 0........9.............................40
                                    3.....9.............................40
  0.2......................26...31........41

 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
     with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
     (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

NGS: Short Introduction

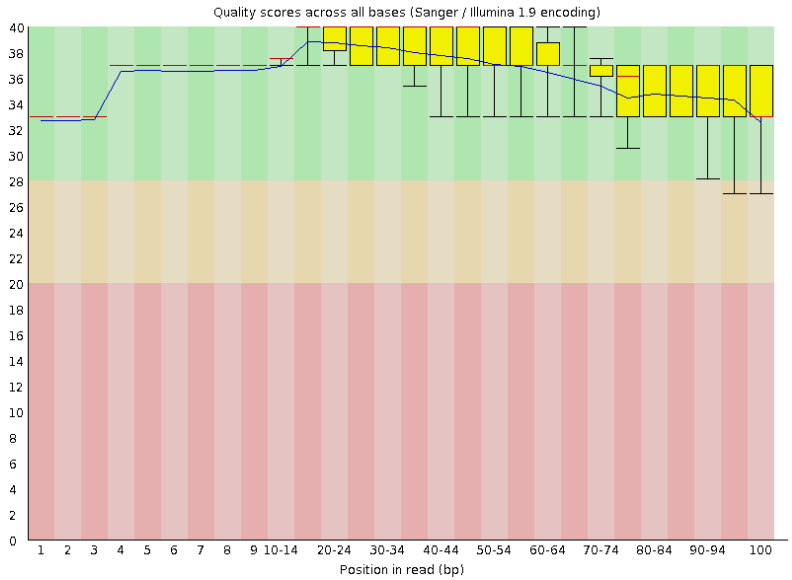
History | NGS | Bioinformatics

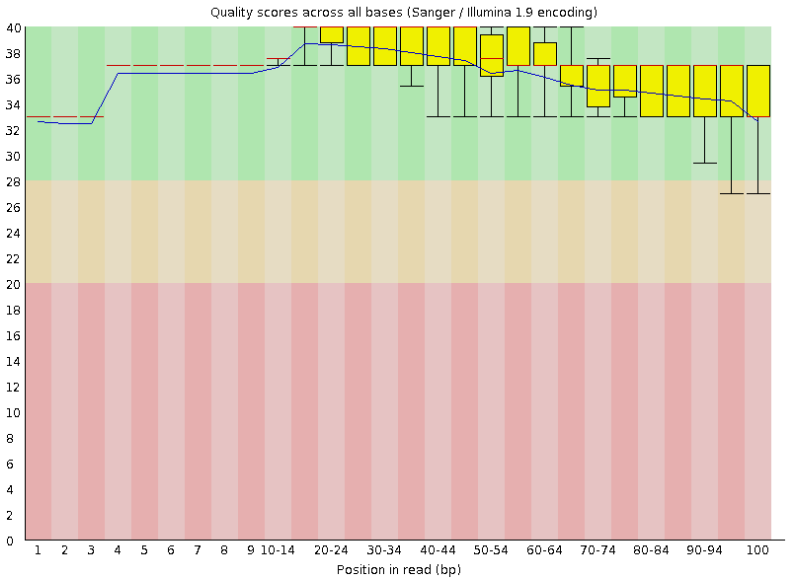
All-seq

FASTQ FORMAT

@sequence: and information of it
GATTTGGGGTTCAAAG
+
!''*((((***+))%%

$$Q_{sanger} = -10log_{10}(p)$$





FRAGMENT

FRAGMENT

READ 1

READ 2

NGS: Short Introduction

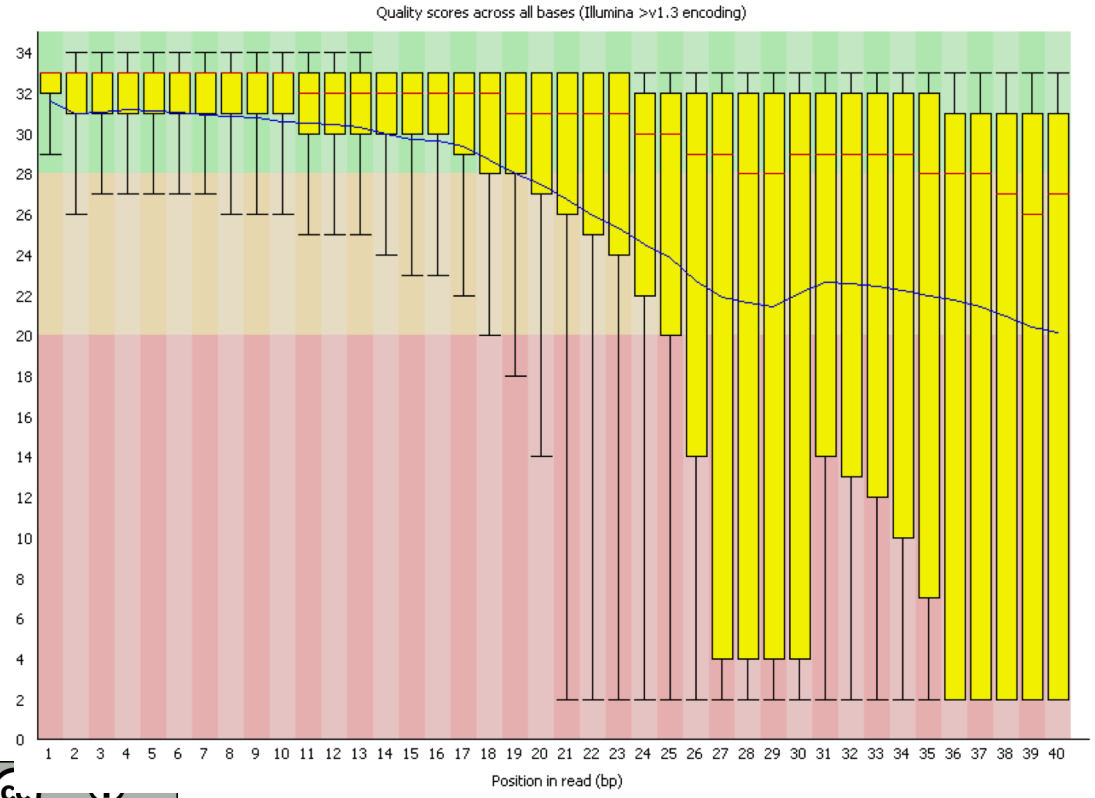
History | NGS | Bioinformatics

All-seq

FASTQ FORMAT

@sequence: and information of it
GATTTGGGGTTCAAAG
+
!''*((((***+))%%

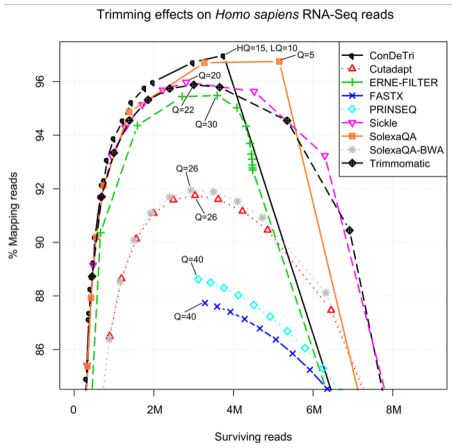
$$Q_{sanger} = -10log_{10}(p)$$



TRIMMING

Trim Galore!

http://www.bioinformatics.babraham.ac.uk/projec



Fabbro et al, 2013, PloS One

NGS: Short Introduction

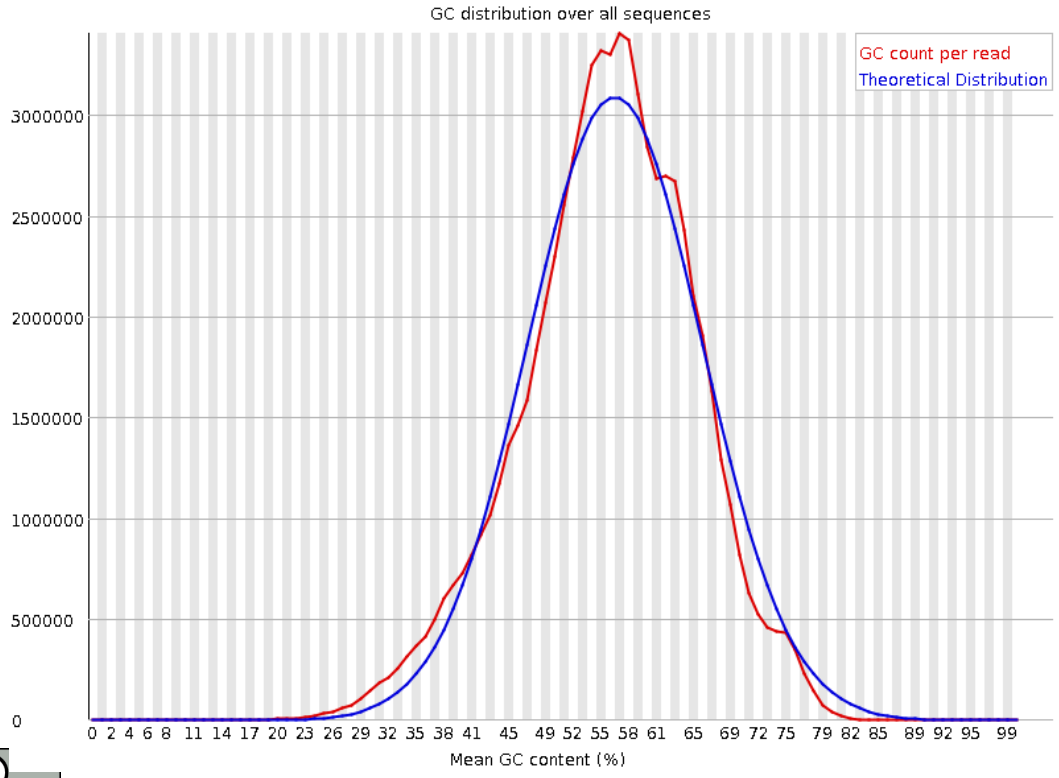
History | NGS | Bioinformatics

All-seq

FASTQ FORMAT

@sequence: and information of it
GATTTGGGGTTCAAAG
+
!''*((((***+))%%

Amount of CG content



Theoretical Distribution is different for RNA and DNA applications

Important to have similar proportions for all samples in a data-set (if expected)

NGS: Short Introduction

History | NGS | Bioinformatics

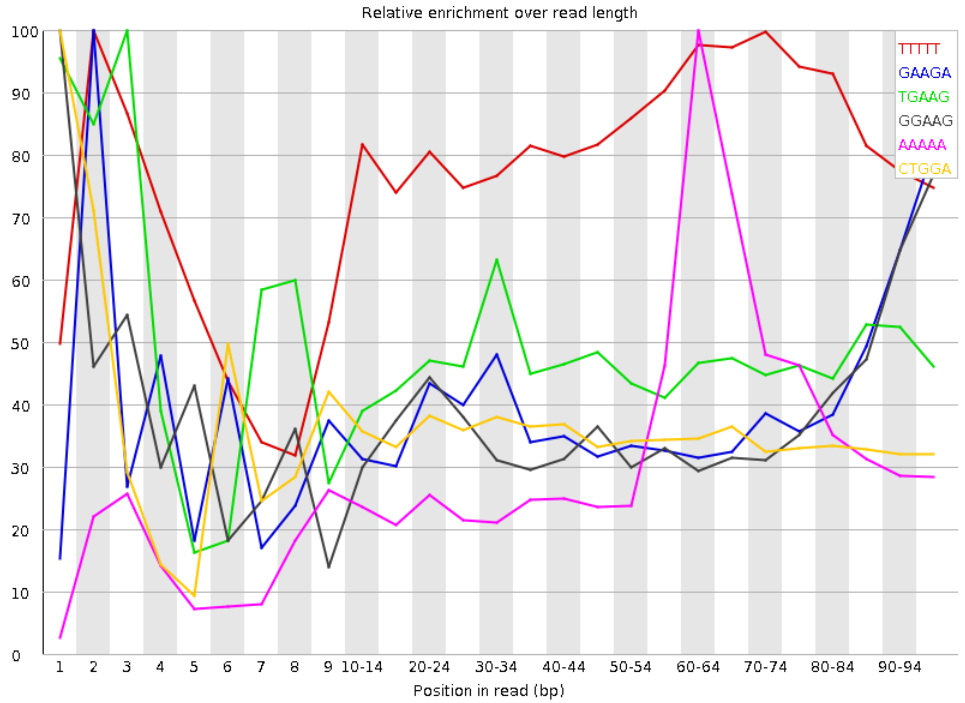
All-seq

FASTQ FORMAT

@sequence: and information of it
GATTTGGGGTTCAAAG
+
!''*((((***+))%%

K-mer over representation

- Over specific locations
- Over entire sequence



Adapters from Illumina?

TRIMMING

NGS: Short Introduction

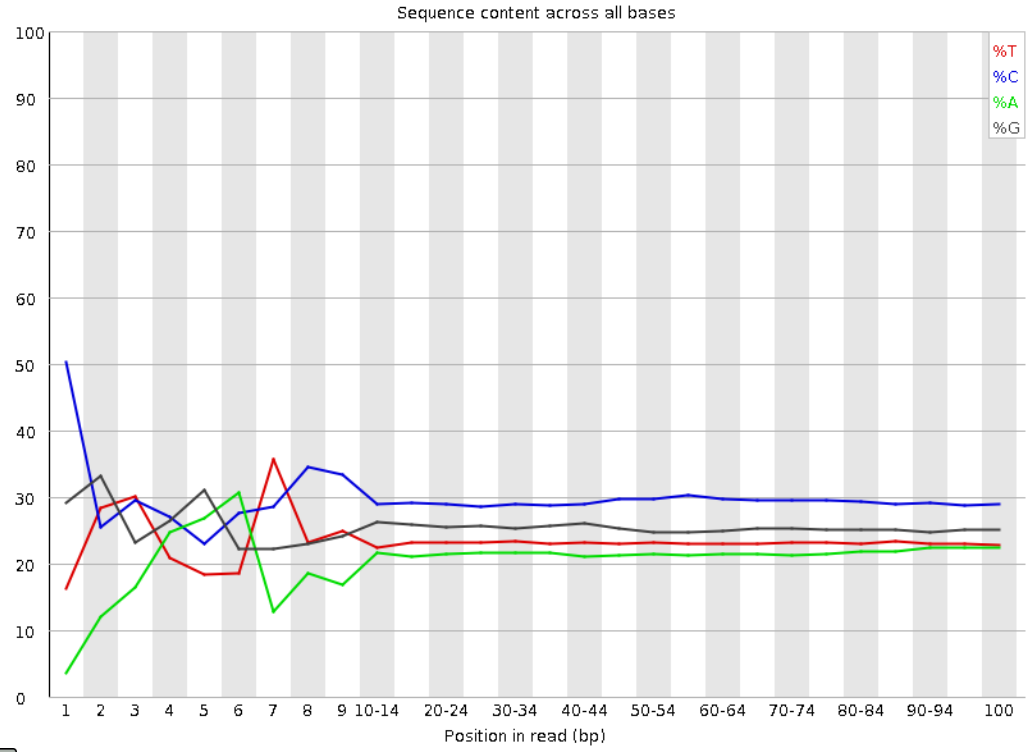
History | NGS | Bioinformatics

All-seq

FASTQ FORMAT

@sequence: and information of it
GATTTGGGGTTCAAAG
+
!''*((((***+))%%

Nucleotide percentage per-base



From FastQC

It's worth noting that some types of library will always produce biased sequence composition, normally at the start of the read. Libraries produced by priming using random hexamers (including nearly all RNA-Seq libraries) and those which were fragmented using transposases inherit an intrinsic bias in the positions at which reads start. This bias does not concern an absolute sequence, but instead provides enrichement of a number of different K-mers at the 5' end of the reads. Whilst this is a true technical bias, it isn't something which can be corrected by trimming and in most cases doesn't seem to adversely affect the downstream analysis.

NGS: Short Introduction

History | NGS | Bioinformatics

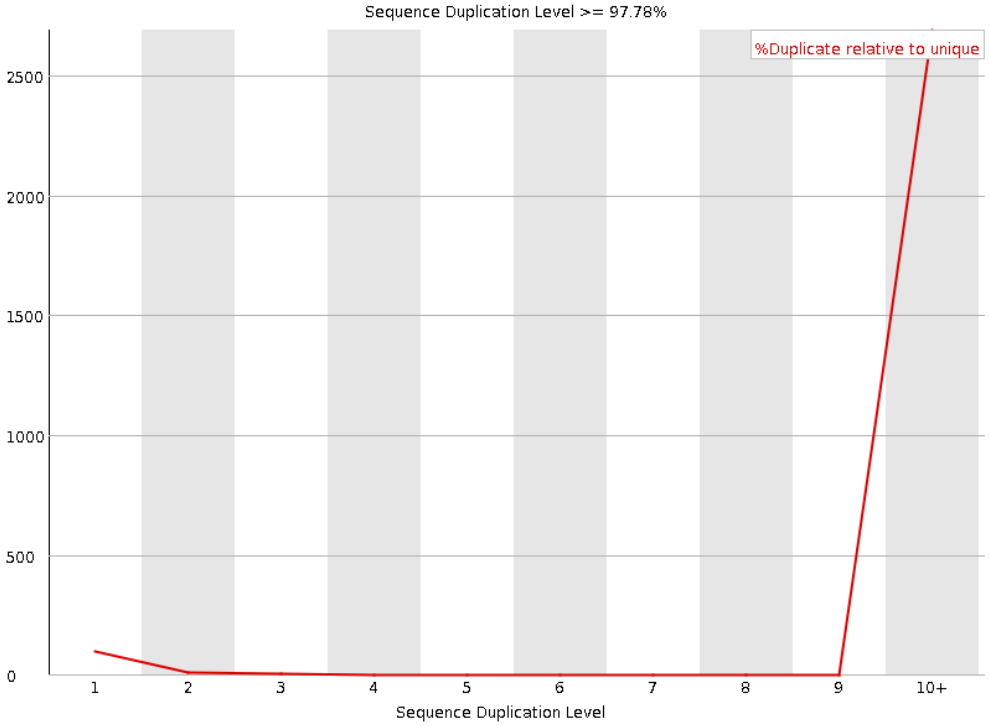
All-seq

FASTQ FORMAT

@sequence: and information of it
GATTTGGGGTTCAAAG
+
!''*((((***+))%%

Duplication level:

- PCR artifact
- Cluster scanned twice
- Biological duplication



DNA and RNA very different expectations

Important to have similar proportions for all samples in a data-set (if expected)

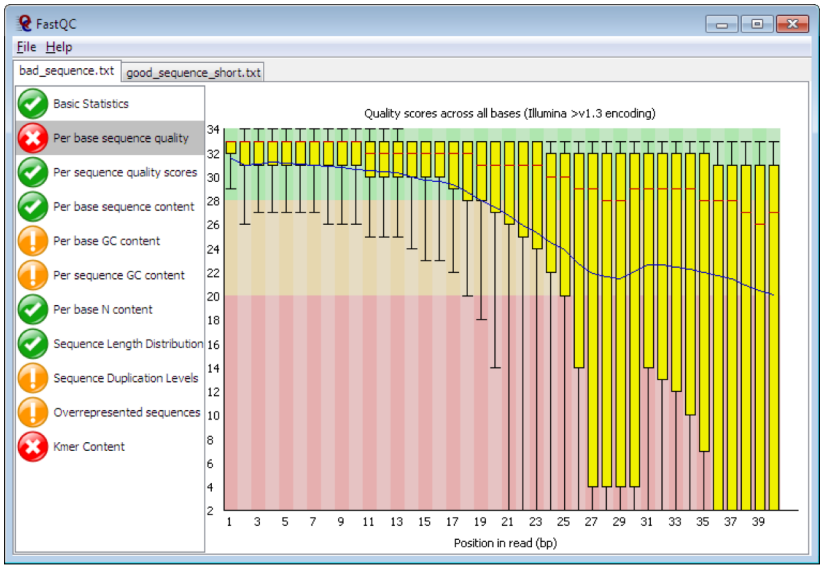
NGS: Short Introduction

History | NGS | Bioinformatics

All-seq

FASTQ FORMAT

@sequence: and information of it
GATTTGGGGTTCAAAG
+
!''*((((***+))%%



FastQC

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

http://www.bioconductor.org/packages/devel/bioc/vignettes/sbgr/inst/doc/sbgr.html

Picard
build passing
A set of tools (in Java) for working with next generation sequen
in the BAM (http://samtools.sourceforge.net) format.

http://broadinstitute.github.io/picard/

NGS: Short Introduction

History | NGS | Bioinformatics

All-seq

BCL to FASTQ

QUALITY CONTROL

FASTQ to SAM/BAM

RNA-seq

miRNA-seq

MAP TO A REFERENCE GENOME

Methodologies to map to the entire genome: BWA, Bowtie,…

Methodologies to map to selected parts of the genome: Tophat,…

RRBS-seq

More to be shown in the RNA-seq session

single-cell RNA-seq

DNase-seq

NGS: Short Introduction