Introduction to NGS technologies

Joaquín Panadero Romero
28 de septiembre de 2016

GDA
International Course on
Genomic Data Analysis

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

1

OUTLINE

1. Basics on the NGS technologies
2. Comparison across NGS platforms
3. Computing requirements
4. Tools for data analysis

2

Basics on NGS technologies

Millions of DNA molecules sequenced simultanously

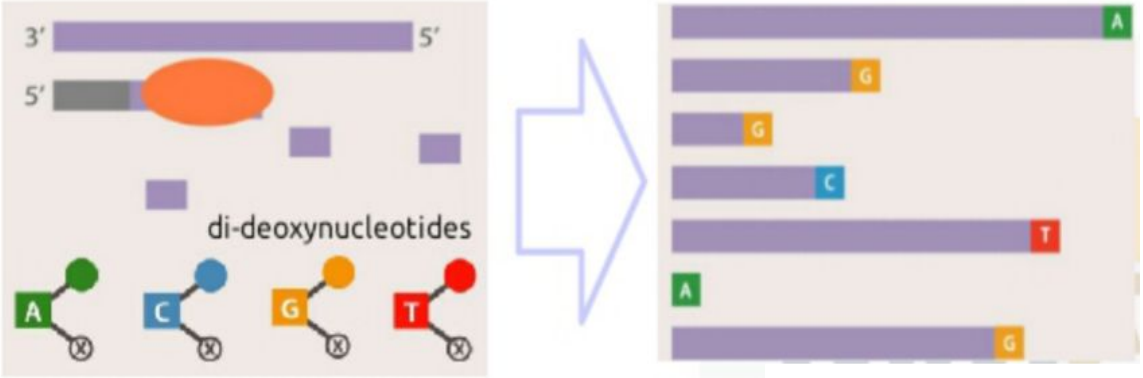
Types:

Sanger
Pyrosequencing
Sequencing by synthesis
Sequencing by ligation
Ion-Semiconductor sequencing

Sanger

Used nowadays in:

- Routine sequencing applications
- NGS data validation

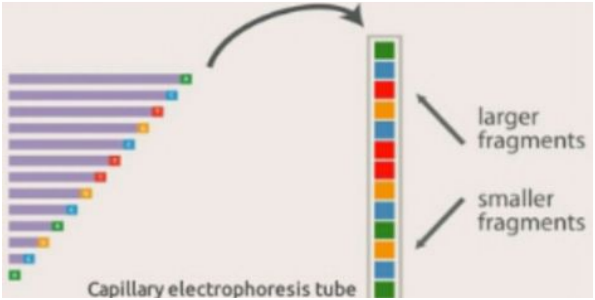


Multiple DNA fragments covering each base position

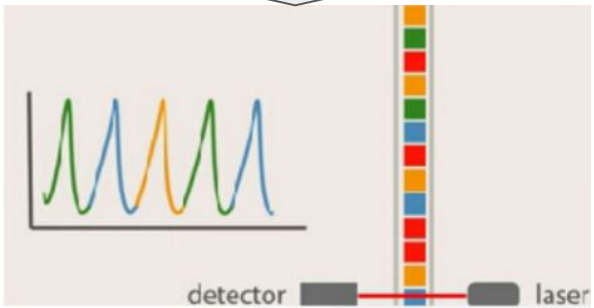
4

Sanger

DNA fragments move according their size



Light detected shows the base added at each position



GDA

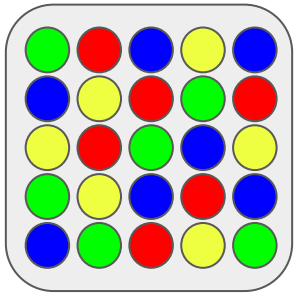
Commons among NGS technologies

Sample preparation

Sequencing machine

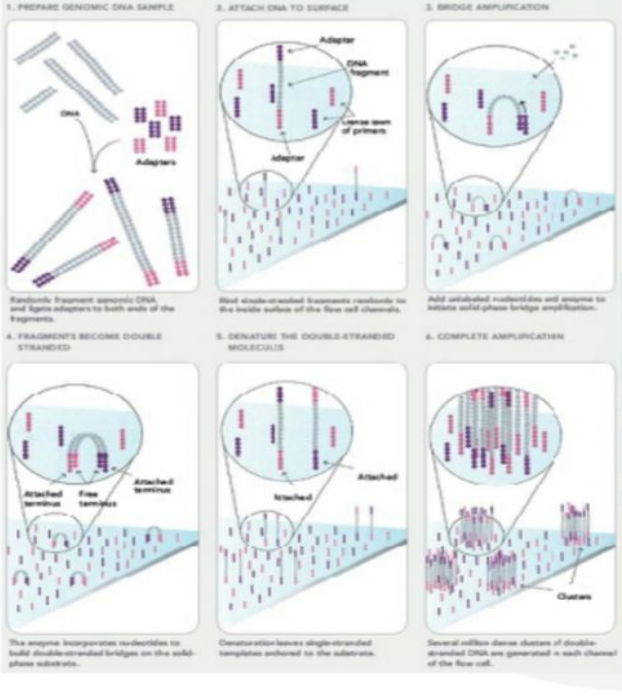
- bridge PCR
- emulsion PCR

Data output



Commons among NGS technologies

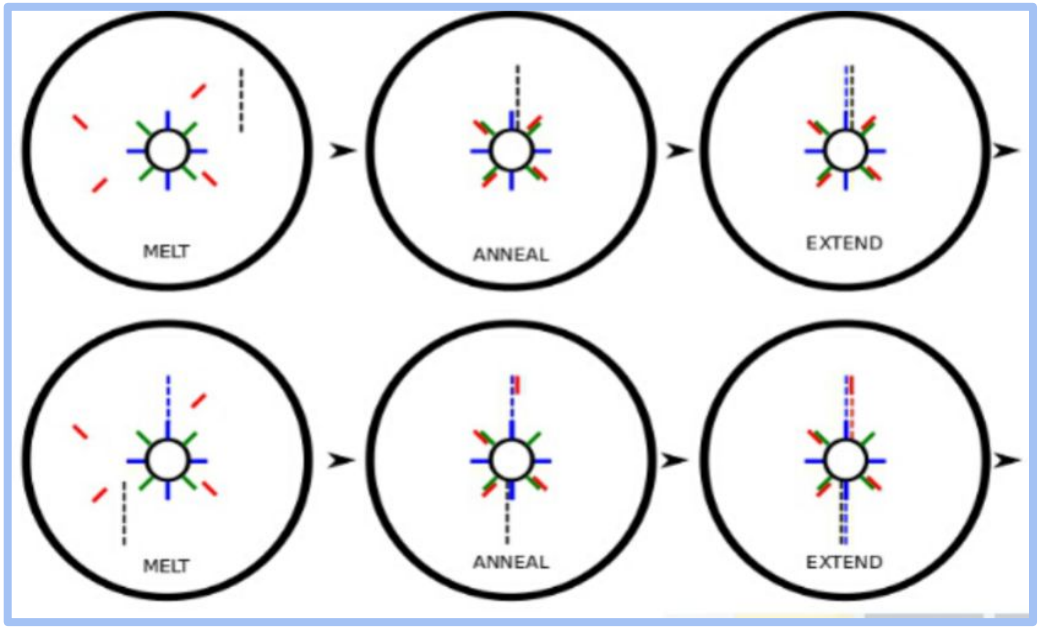
PCR bridge



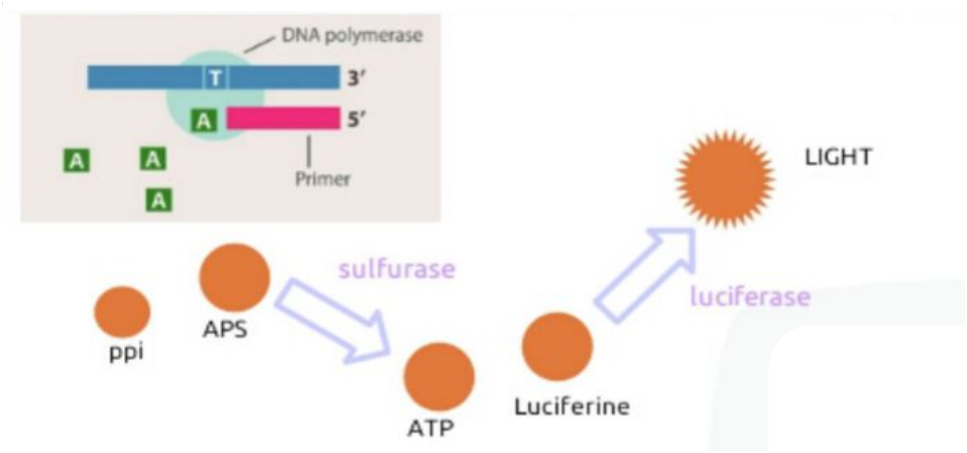
PCR bridge

Commons among NGS technologies

Emulsion PCR

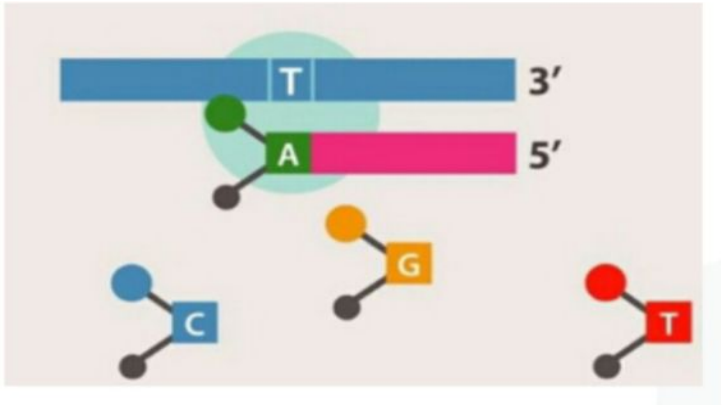


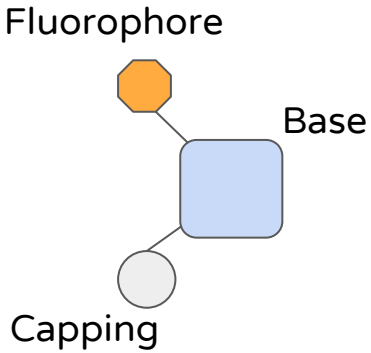
Pyrosequencing



- Large reads length generation
- High reagent cost
- High error rate over strings of 6+ homopolymers

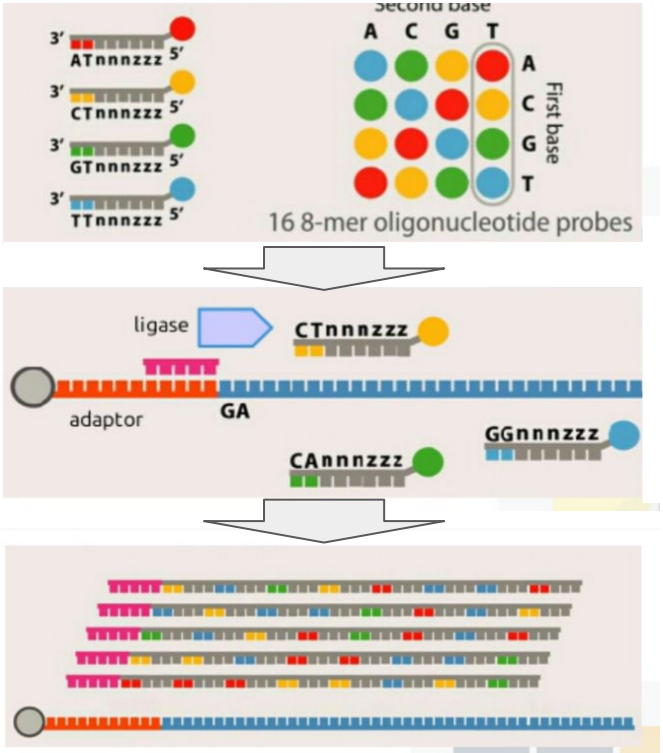
Sequencing by synthesis





- Overcomes homopolymer issue due to terminated nucleotides
- Increased error rate with read length

Sequencing by ligation

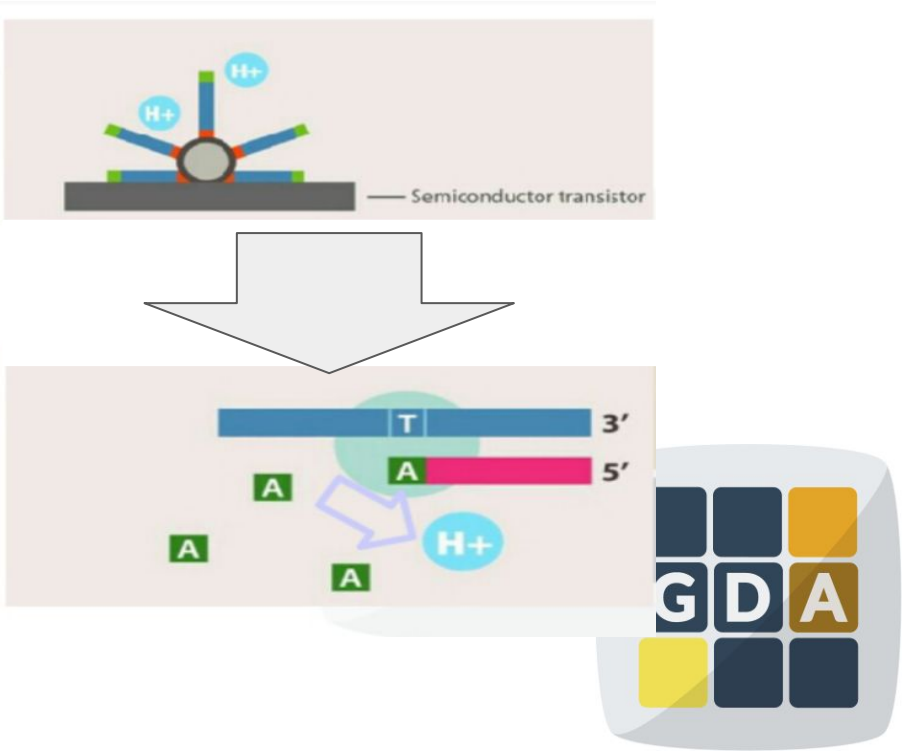


- Based on ligase instead of polimerase
- 5 x 7 ligation cycles
- Short sequences
- Overcome homopolymer problem

Ion-Semiconductor sequencing

- Beads are attached to semiconductor transistors

- Each time one nucleaotide is added, one H+ is released
- Semiconductor transistor detects changes on PH solution



Examples of NGS systems

General overview among NGS methodologies

Coverage of genome per run				
Pyrosequencing	0	0	5	151
Sequencing by synthesis	455	536	11k	323k
Sequencing by ligation	97	114	2k	69k
Ion semiconductor sequencing	3	4	74	2k

Applications

Whole genome sequencing
Variant Calling
RNA-seq
De novo sequencing and assembly
Chip-seq
Methyl-seq
Metagenomics

Computational infrastructure for NGS data analysis

Requirements:

Conditioned data center (server rooms)
Computing cluster (racks)
Many computer nodes (servers)
High performance and capacity storage
Fast networks
Skilled people in computing (sysadmins and developers)

Alternatives: cloud computing

Pros

Flexibility
You pay what you use
Don’t need to mantain a data center

Cons

Transfer datasets through the internet is slow
Lower performance
Privacy and security concerns
More expensive for big and long term projects

Storage

Which data do we want to keep?

- Raw data (fastq)
- Processed data (fastq, bam, sam…)
- Final results (vcf, excel, txt …)

How many storage resources are available?
How long?

BGI: Beijing Genomic Institute

Sequencing instruments
- Illumina HiSeq
- AB Solid system
- Ion Torrent

Informatics infrastructure
- 20576 cores cluster
- 17PB (petabytes)

CNAG: Centro Nacional de Análisis Genómico

Sequencing instruments
- 10 Illumina HiSeq 2000

Informatics infrastructure
- 850 cores cluster
- 7.5PB (petabytes)

How we do proceed?

Fatsq format

```
@K00211:91:H52J5BBXX:2:1101:9181:1297 1:N:0:NNNNNNNN
TCAAGCACCTTCTGCAGGCACCACTGGATGATGCTCAAGAAATTCTGCAAGCACGCTTCCCGATGCCACGTTACATCAACACGGAGCATG
GAGGCAGT
+
AFFJJJJJJJJJJJJJJJJJJJJFJFFJJJJJJJJJJFJJJJFF<77FJA-J--77<7AJ<FAJAJJJJJJJJFJJJJJJJJ<JFJJFJAJAJJJJJ
@K00211:91:H52J5BBXX:2:1101:9485:1297 1:N:0:NNNNNNNN
CAGGCCTTCCACGCTCTCACAGCTTCTGGAGGAAGGGAATCAATTTCGAGCAAATTACTTCTTACAACCAGAGCTCATGCCTTCACAACT
GGCCTTCA
+
AFAJJJJJJJJJJFJJJJFFJJJJJJJJJJJJJJJJJJJJJJJJJJAJJJFJJJJJJJFJAFFJJJJJFJJJJJJJJJJJJJJJJJJJFFFJFJJJ
@K00211:91:H52J5BBXX:2:1101:9567:1297 1:N:0:NNNNNNNN
GGGATTCATCATAGTAACTGTAACGGAACCTTTGACCCCAATGGCAGAAATCAGAAGAAACCACAAAGAGATTACTAGGATCCGCTAGAT
ATTTACTG
+
AFFJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJFJJJJFFJ<FJJAFJJJJJJJJJ<FFJJJJJJJJJJJAJJJJ<JJ<JFJJJJJJJJJJJJ
```

¿?

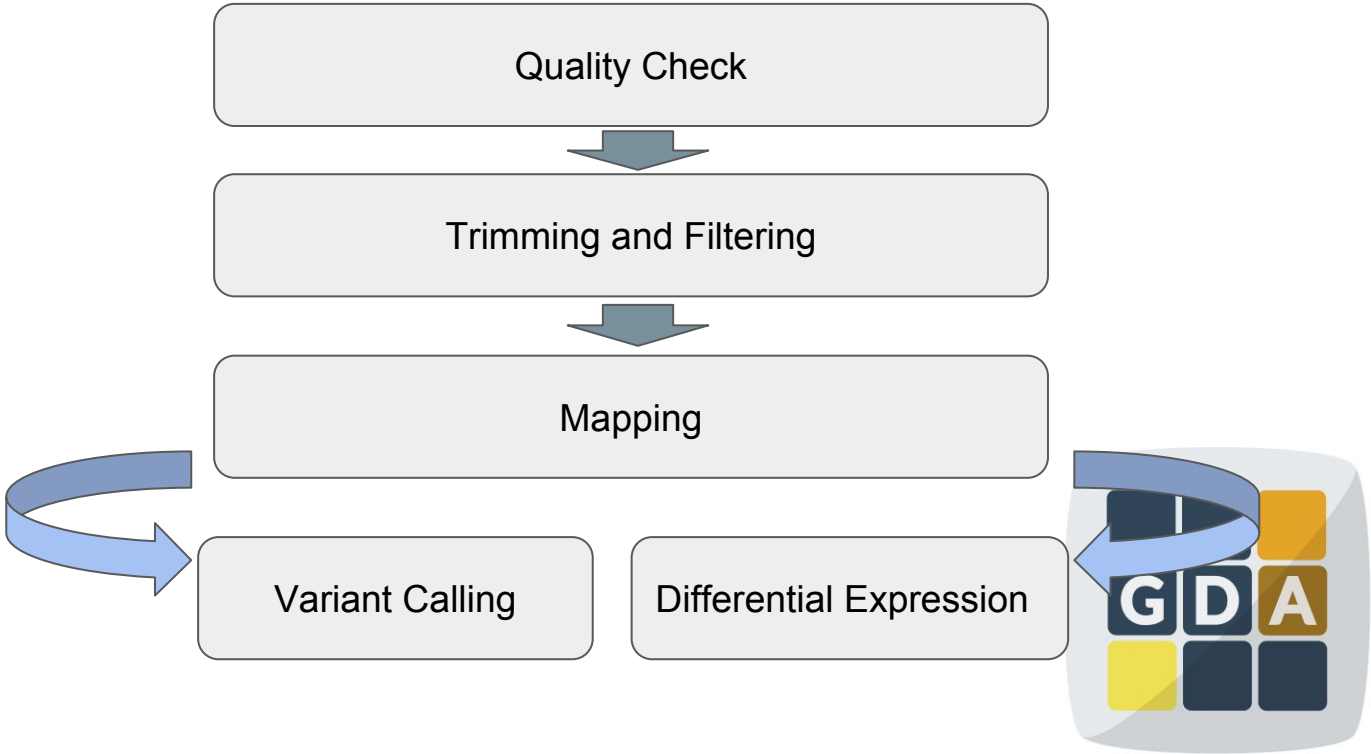
Tools on NGS data analysis

FastQC http://www.bioinformatics.babraham.ac.uk/projects/fastqc/	**Blast2GO** https://www.blast2go.com
cutadapt https://github.com/marcelm/cutadapt	**samtools** samtools.sourceforge.net
bowtie http://bowtie-bio.sourceforge.net/index.shtml	**vcftools** vcftools.sourceforge.net
bwa http://bio-bwa.sourceforge.net	**GATK** https://www.broadinstitute.org/gatk/
tophat https://ccb.jhu.edu/software/tophat/index.shtml	**qiime** http://qiime.org

Tools on NGS data analysis

cufflinks http://cole-trapnell-lab.github.io/cufflinks/	**mothur** www.mothur.org
abyss http://www.bcgsc.ca/platform/bioinfo/software/abyss	**bismark** http://www.bioinformatics.babraham.ac.uk/projects/bismark/
spades http://bioinf.spbau.ru/spades	**blast** https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download
glimmer https://ccb.jhu.edu/software/glimmer/	**augustus** http://augustus.gobics.de

Basic workflow



THANKS