Introducción al análisis de datos NGS

RICARDO A. VERDUGO, Ph.D.

Programa de Genética Humana, ICBM
Facultad de Medicina, U. de Chile

Abril 2019

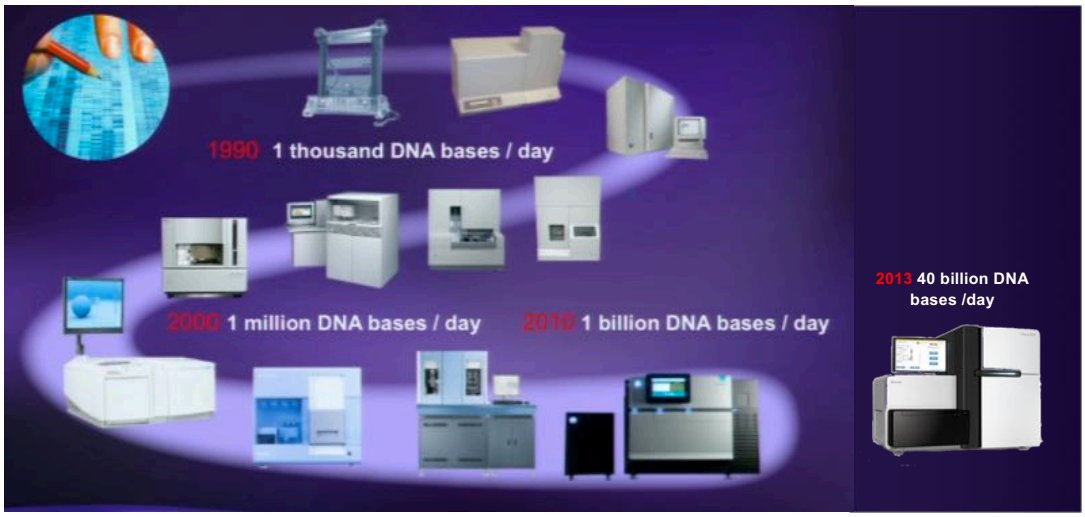
GENOMED-Lab
http://genomed.med.uchile.cl

Temas a cubrir

1. Secuenciación por síntesis
2. Datos de secuenciación masiva
3. Flujos de trabajo NGS
4. Control de calidad de los datos

Evolución de la técnica de secuenciación del ADN

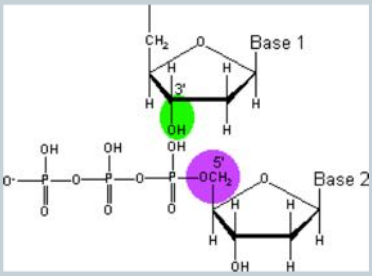
3



Polimerización de los nucleótidos

- El grupo 5 'de un nucleótidos trifosfato (dNTP) se acerca grupo hidroxilo 3' de una cadena de nucleótidos.

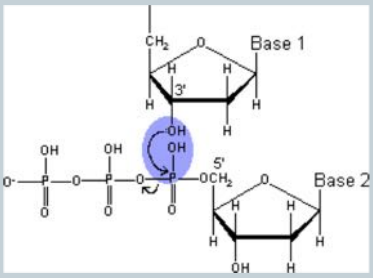
Cadena de ADN



dNTP

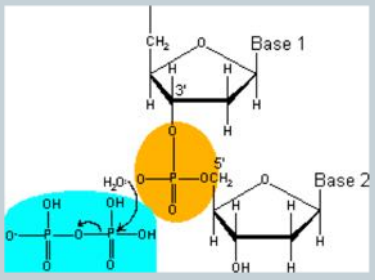
Polimerización de los nucleótidos

- El grupo hidroxilo 3’ forma un enlace con el átomo de fósforo más próxima al átomo de oxígeno 5’ del nucleótido libre (dNTP).
- El enlace entre el primer átomo de fósforo y el átomo de oxígeno que une a los próximos grupos fosfato se rompe.
- Se libera un protón (H^+) y un grupo OH^-)



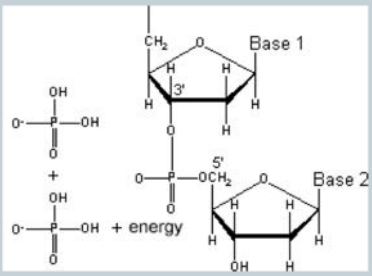
Polimerización de los nucleótidos

- Un nuevo enlace fosfodiester une los dos nucleótidos
- Se libera un grupo pirofosfato
- Se libera un protón (H^+)



Polimerización de los nucleótidos

- El grupo pirofosfato se hidroliza (se agregar agua)
- Se libera energía que se usa para la siguiente reacción



Más detealles en:
http://www.ncbi.nlm.nih.gov/books/NBK22513/
https://www.chem.wisc.edu/deptfiles/genchem/netorial/modules/biomolecules/modules/dna1/dna13.htm

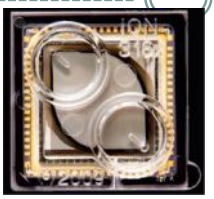
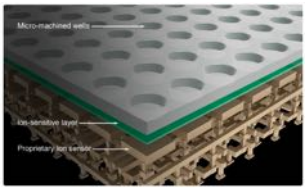
I. Secuenciación masiva

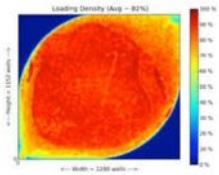
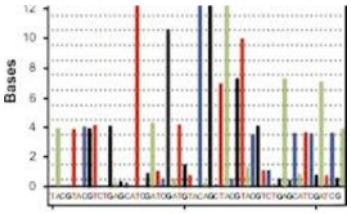
Secuenciación de última generación

Secuenciación paralela

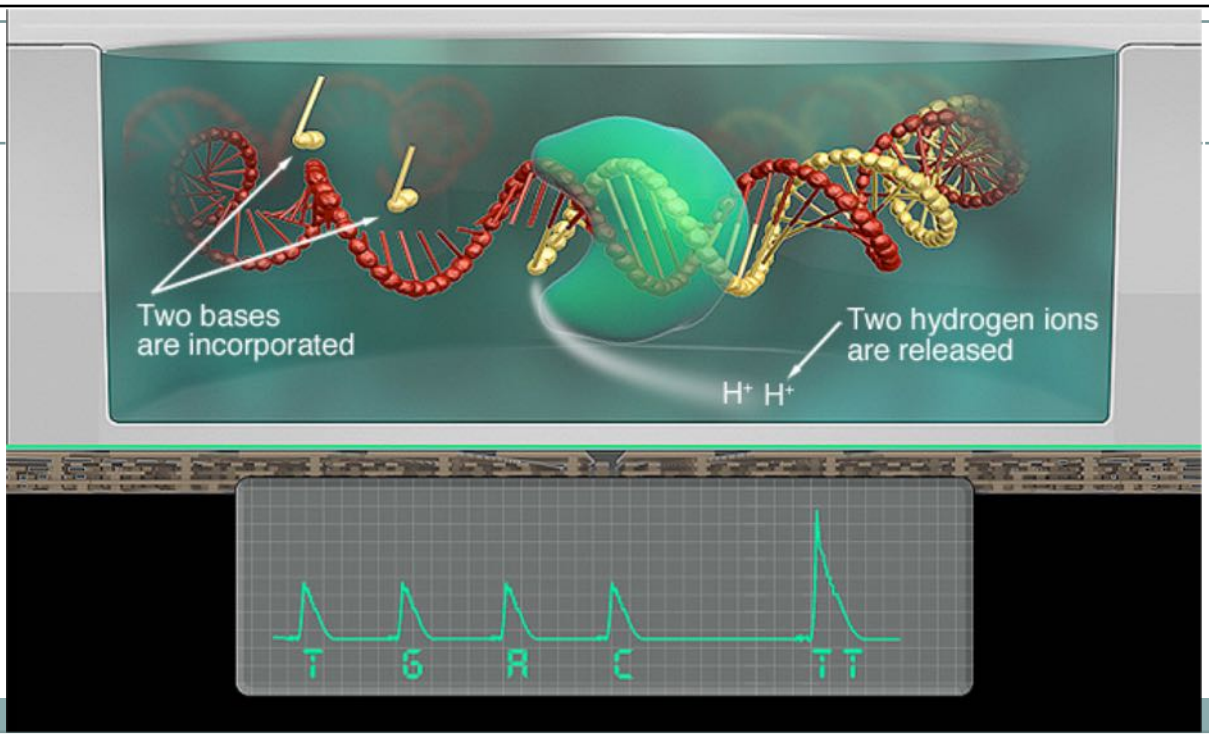
Next Generation Sequencing (NGS)

Ion Torrent – Semi conductor seq

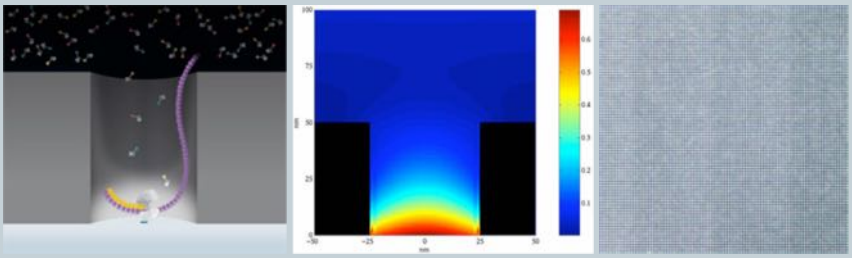




- Cada incorporación de un dNTP libera in ion H^+
- Detección por una placa detectora de iones
- Secuenciación por síntesis



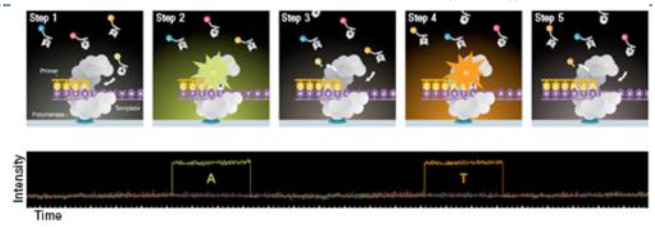
Pacific BioSciences

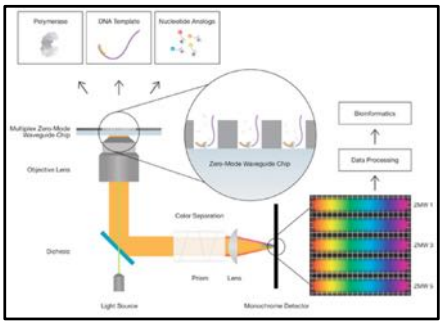


•Single Molecule, Real-Time (SMRT) Sequencing
•Average > 10,000 bp, some reads > 60,000 bp
•Single molecule sequencing suffers from poor signal/noise
•The Zero Mode Wave (ZMW) guide acts like a filter to restrict the fluorescence measurement to the well bottom

PACIFIC BIOSCIENCES

Pacific BioSciences



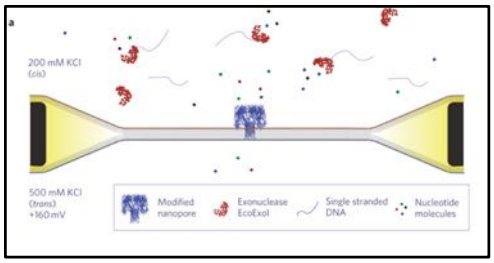


Real time sequencing thanks to engineered DNA polymerase and cleavable fluorophore.

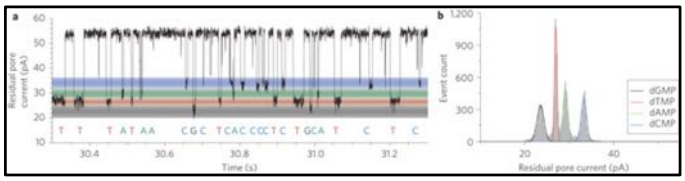
- 80,000 to 300,000 wells.
- 10,000 nt read length
- Strobe sequencing possible

PACIFIC BIOSCIENCES

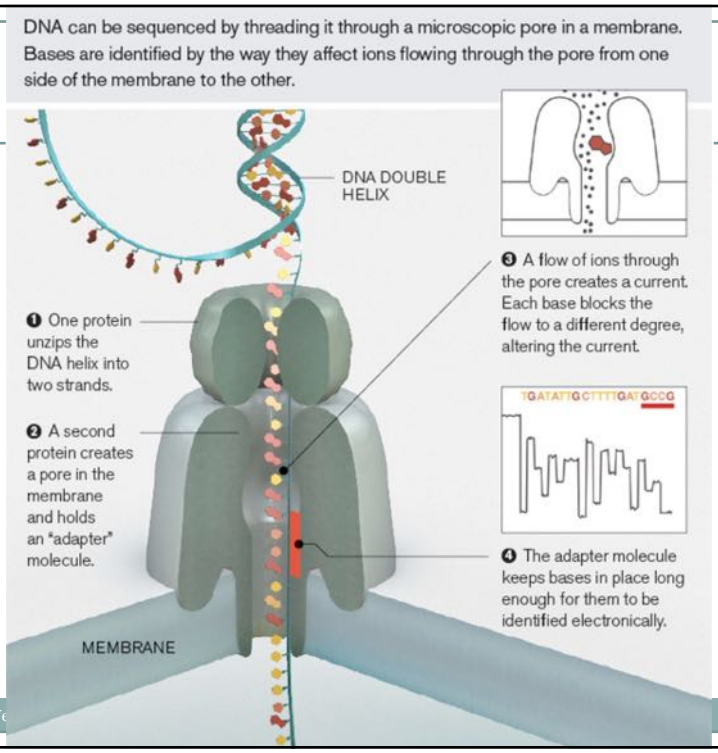
Oxford Nanopore



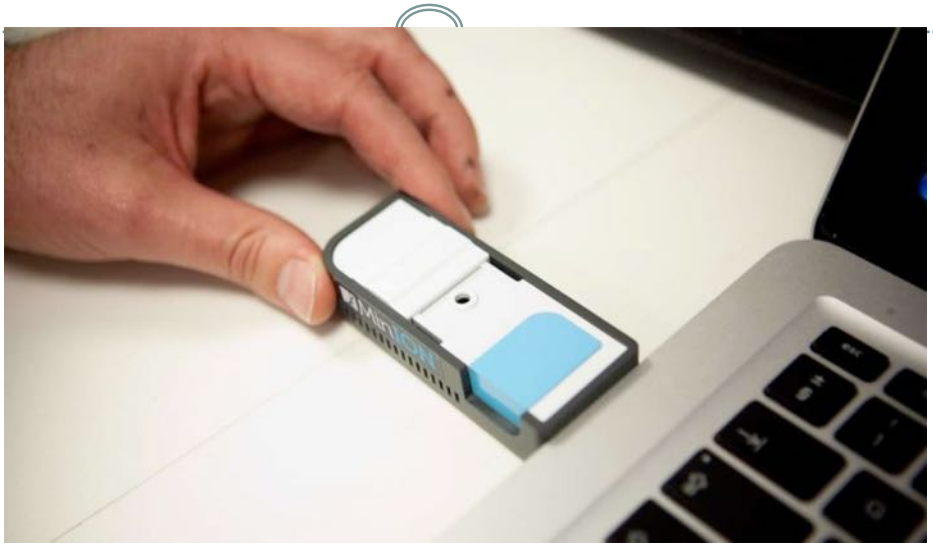
- Modifed αHemolysine with inserted Cyclodextrine
- Embedded in a lipid bilayer
- Flow of Nucleotide through the pore modifies the conductance
- Need to attach the Exonuclease to the pore



Oxford NANOPORE Technologies



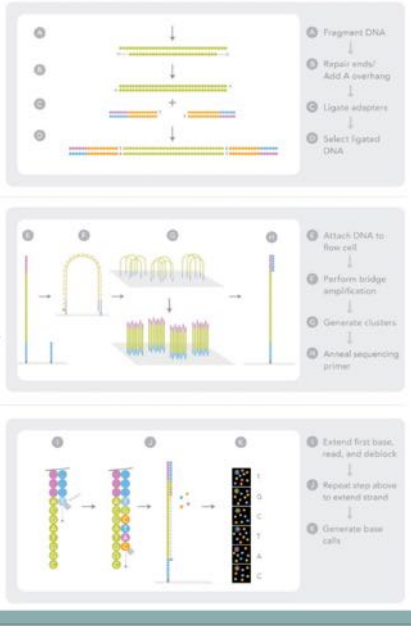
MinION Oxford Nanopore Sequencer



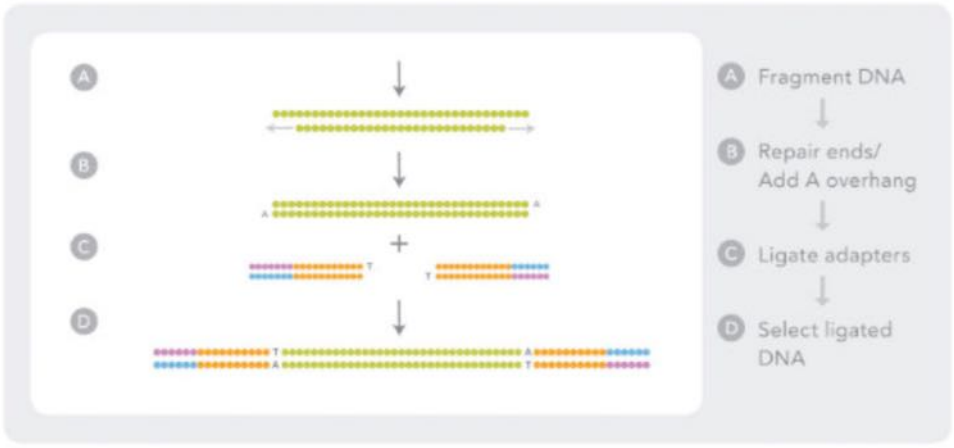
MinION : 512 pores (early access 2014) - disposable
GridION: 2000-8000 pores (TBD)
20 racks: human genome in 15 minutes

Illumina Technology

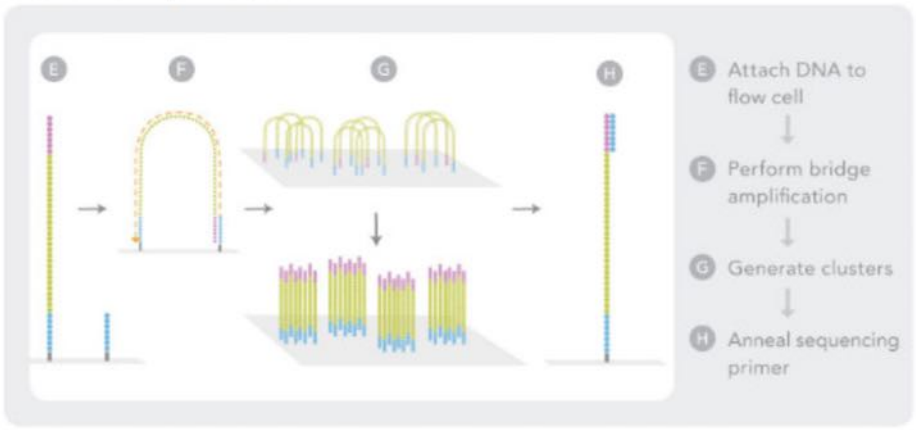
- Library Preparation
- Clonal amplification
- Sequencing by synthesis



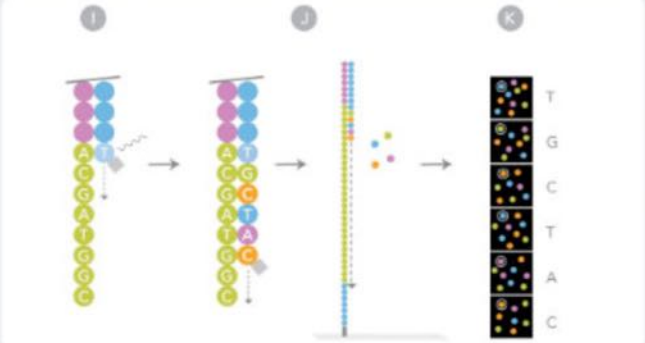
- Library Preparation (TruSeq)



Clonal amplification



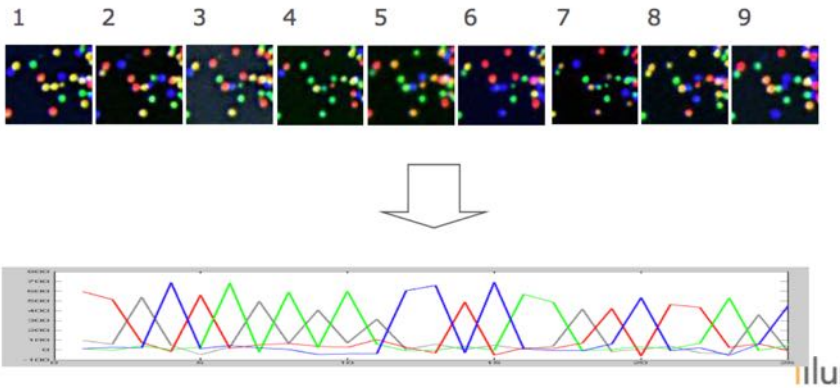
Sequencing by synthesis



Más detalles en http://bioinformaticsalgorithms.com/faqs/assembly.html

Image analysis identifies clusters (polonies, beads) and extracts intensity traces

- Detection: Find all clusters on the image
- Registration: Track clusters over multiple sequencing cycles
- Extraction: Provide intensity estimates for clusters in a given image



Illumina

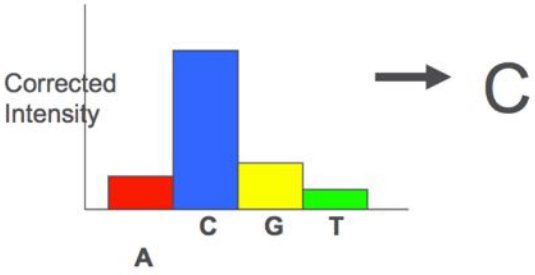
Visualizing
high-throughput Illumina
four-color sequencing-by-synthesis

W. Trimble
Argonne National Laboratory

http://bioinformaticsalgorithms.com/faqs/assembly.html

Base-calling has two aspects: Identifying the base-call and assigning a confidence estimate to the call

- Making a base-call is usually based on the intensity estimates
  - Signal-processing needs to correct for confounding factors:
    - Frequency cross-talk (optical detection mechanism)
    - Phasing effects (imperfect chemistry)
    - Signal decay
- Assignment of a confidence estimate or quality score is vital for downstream analysis
  - phred method can be extended to Next_gen technologies



Quality scores quantify the probability that a base-call is correct (or wrong)

Termninology:

- Base quality scores
  - Individual bases have quality scores which reflect the likelihood of the base being correct/incorrect
- Alignment scores
  - Probability than an alignment to a given position in the reference genome is correct
- Allele scores, SNP scores,
  - Probability that a given allele, SNP was observed (often conditional on the alignment being correct)
- Base and alignment scores are single read scores; SNP scores are consensus scores
  - Consensus calls use information from multiple reads

Phred scores

1. A base quality score assigned by the phred software (or a program based on the phred)
2. A quality score expressed on a logarithmic scale:
   - Q = - 10 log10(probability of an error)
   - Example: Q20 = 1% error probability

Capacidad de secuenciación de equipos Illumina disponibles en el mercado


MiSeq


NextSeq 500


HiSeq 2500

Run mode	N/A	Mid Ouput	High Output	Rapid Run	High Output
Flow cells processed per run	1	1	1	1 or 2	1 or 2
Output range	0.5 - 15 Gb	20 - 39 Gb	30 - 120 Gb	10 - 180 Gb	50 - 1000 Gb
Run time	5 - 65 hours	15 - 26 hours	12 - 30 hours	7 - 40 hours	< 1 day - 6 days
Single-end reads per flow cell	25 Million	130 Million	400 Million	300 Million	2 Billion
Max read length	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 125 bp

Número de muestras que pueden ser estudiadas en una corrida

Key Area	MiSeq	NextSeq 500	HiSeq 2500
DNA Sequencing			
Whole Genome - Large (e.g. Human)	-	1	1-10
Whole Genome - Small (e.g. E. coli)	1 - 96	120 -792	96 - 6660
Targeted - Exome or large panels	1	1 - 12	12 - 160
Targeted - Small gene panels	3	15 - 48	36 - 72
RNA Sequencing			
RNA Profiling	1 - 2	12 - 36	24 - 396
Transcriptome Analysis	-	3 - 10	8 - 96
Small RNA Analysis	1 - 5	25 - 80	60 - 792
Targeted RNA	384	384	6144
Regulation Applications			
ChIP-Seq	1	8 - 24	20 - 264
Methylation Analysis	-	1	

II. Basic NGS Workflow

Sample Prep

Isolation of material
PCR amplification
End repair, size selection

Sequencing

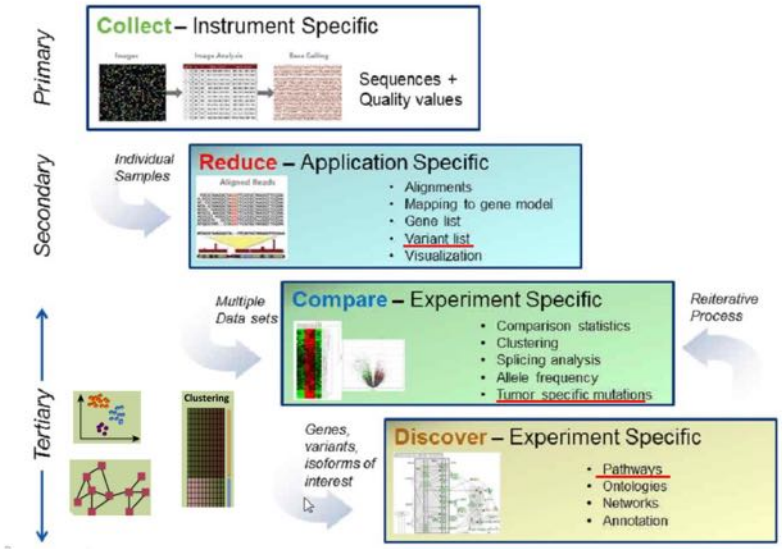
Library QC
Cluster generation
Instrument operation

Data Analysis

QC and pipeline analysis
Data interpretation

Olson *et al.*

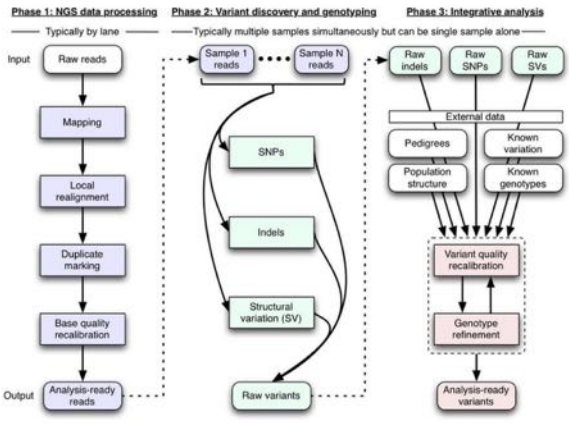
High Throughput Data Analysis Overview



Olson *et al.*

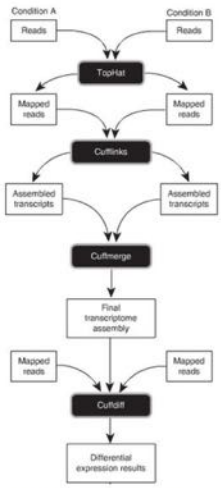
Typical Data Analysis Pipelines

Genotyping (GATK)



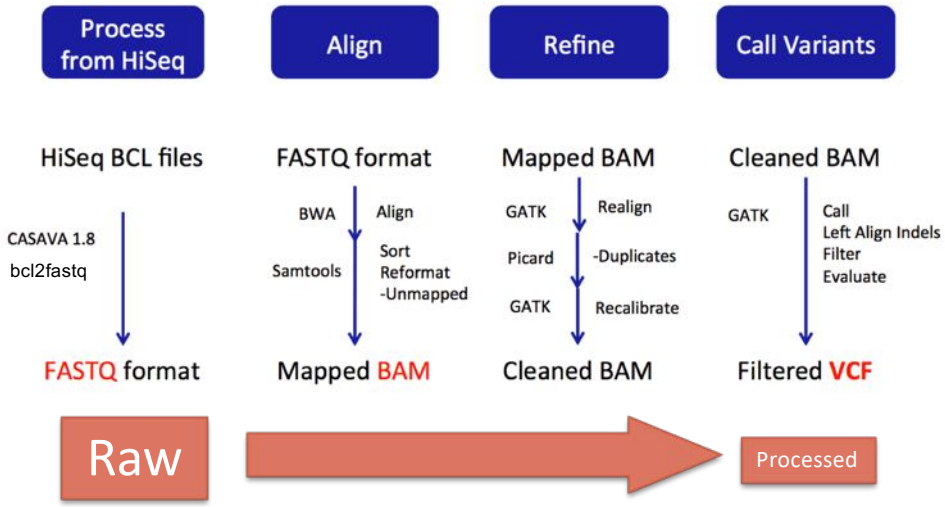
http://www.broadinstitute.org/gsa/wiki/images/7/7a/Overall_flow.jpg
http://www.broadinstitute.org/gatk/guide/topic?name=intro

RNA-seq (Tuxedo)



http://www.nature.com/nprot/journal/v7/n3/full/nprot.2012.016.html

NGS data files



File Sizes

	Depth	FASTQ.gz	BAM	VCF.gz
Genome	30x	100-200Gb	100-200 Gb	2Gb
Exome (50 Mb)	120x	5-10 Gb	10-30Gb	3-6Mb
Deep Targeted (150 kb)	1500x	300-500Mb	150Mb	3-6Mb

Assume 100 bp read length

Raw sequencing data: Fastq format

```
@IL31_4368:1:1:996:8507/2
TCCCTTACCCCCAAGCTCCATACCCTCCTAATGCCCACACCTCTTACCTTAGGA
+
FFCEFFFEEFFFFFFFFEFFEFFFFEFCFC<EEFEFFFCEFF<;EEFF=FEE?FCE
@IL31_4368:1:1:996:21421/2
CAAAAACTTTCACTTTACCTGCCGGGTTTCCCAGTTTACATTCCACTGTTTGAC
+
>DBDDB,B9BAA4AAB7BB?7BBB=91;+*@;5<87+*=/*@@?9=73=.7)7*
@IL31_4368:1:1:997:10572/2
GATCTTCTGTGACTGGAAGAAAATGTGTTACATATTACATTTCTGTCCCCATTG
+
E?=EECE<EEEE98EEEEAEEBD??BE@AEAB><EEABCEEDEC<<EBDA=DEE
@IL31_4368:1:1:997:15684/2
CAGCCTCAGATTCAGCATTCTCAAATTCAGCTGCGGCTGAAACAGCAGCAGGAC
+
EEEEDEEE9EAEEDEEEEEEEEEECEEAAEEDEE<CD=D=*BCAC?;CB,<D@,
@IL31_4368:1:1:997:15249/2
AATGTTCTGAAACCTCTGAGAAAGCAAATATTTATTTTAATGAAAAATCCTTAT
+
EDEEC;EEE;EEE?EECE;7AEEEEEE07EECEA;D6D>+EE4E7EEE4;E=EA
@IL31_4368:1:1:997:6273/2
ACATTTACCAAGACCAAAGGAAACTTACCTTGCAAGAATTAGACAGTTCATTTG
+
EEAAFFFEEFEFCFAFFAFCCFFEFEF>EFFFFB?ABA@ECEE=<F@DE@DDF;
@IL31_4368:1:1:997:1657/2
CCCACCTCTCTCAATGTTTTCCATATGGCAGGGACTCAGCACAGGTGGATTAAT
(...)
```

- Instrument serial #
- Lane
- Swath
- X coord
- Y coord
- Read direction

Fastq Read ID Variations

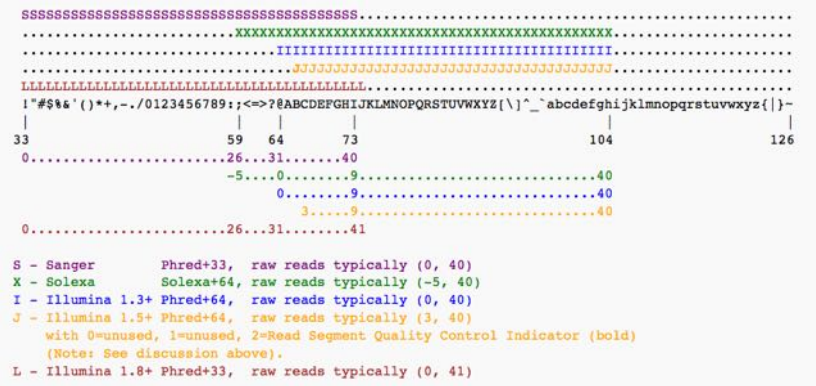
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

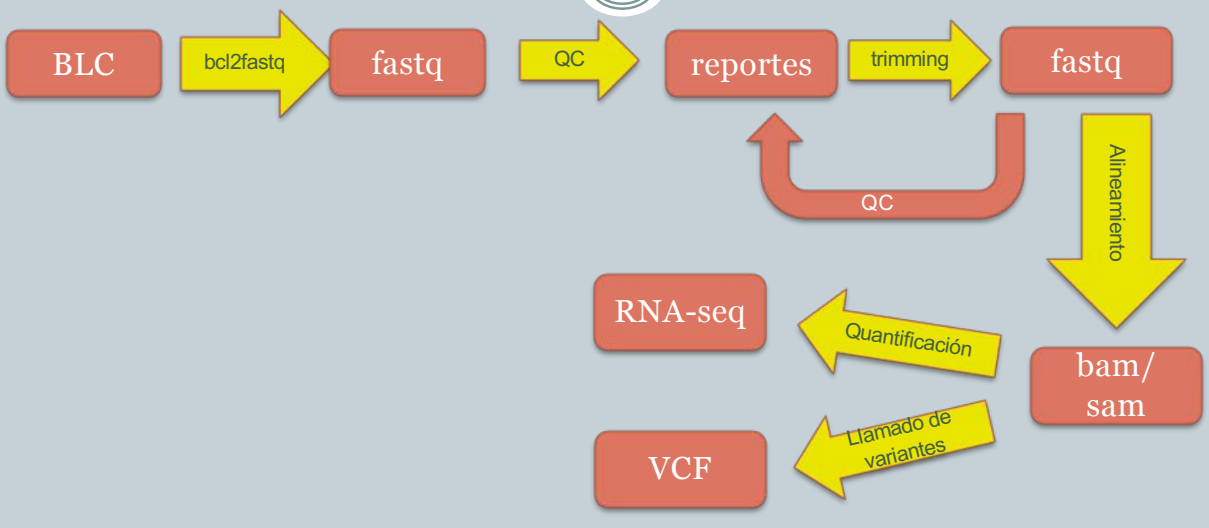
PHRED-like quality

Phred=-10log10(error)

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```



III. Standard NGS Analysis Pipeline



IV. Control de Calidad en NGS

37

1. Calidad de la muestra de ADN o ARN
   - Revisar degradación
   - Espectrofotometría (Nanodrop)
   - Fluorimetría (Pico- y Ribo-Green)
   - Electroforesis en gel o capilar (Bioanalyser, TapeStation, Fragment Analyzer). RNA Integrity Number (RIN) de Agilent para RNA (RIN>7)

Control de Calidad en NGS

38

1. Calidad de las secuencias
   - Sequence Analysis Viewer (SAV) en Basespace
   - FastQC[1]
   - MultiQC
   - FQC Dashboard (*Bioinformatics*:33(19):3137)

[1]https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/

Sequence Analysis Viewer (SAV)

39

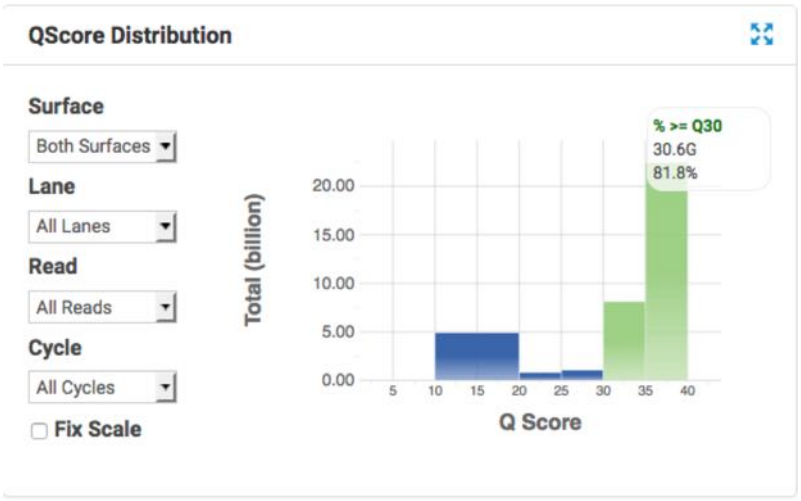
Per Read Metrics

	CYCLES	YIELD	PROJECTED YIELD	ALIGNED (%)	ERROR RATE (%)	INTENSITY CYCLE 1	%≥Q30
Read 1	76	16.72 Gbp	16.72 Gbp	1.55	1.34	3,531	83.60
Read 2 (I)	8	1.56 Gbp	1.56 Gbp	0.00	0.00	1,916	90.60
Read 3 (I)	8	1.56 Gbp	1.56 Gbp	0.00	0.00	1,670	89.58
Read 4	76	16.69 Gbp	16.69 Gbp	1.46	1.48	2,693	78.88
Non-Index Reads Total	152	33.41 Gbp	33.41 Gbp	1.50	1.41	3,112	81.24
Totals	**168**	**36.53 Gbp**	**36.53 Gbp**	**1.50**	**1.41**	**2,452**	**82.00**

Sequence Analysis Viewer (SAV)

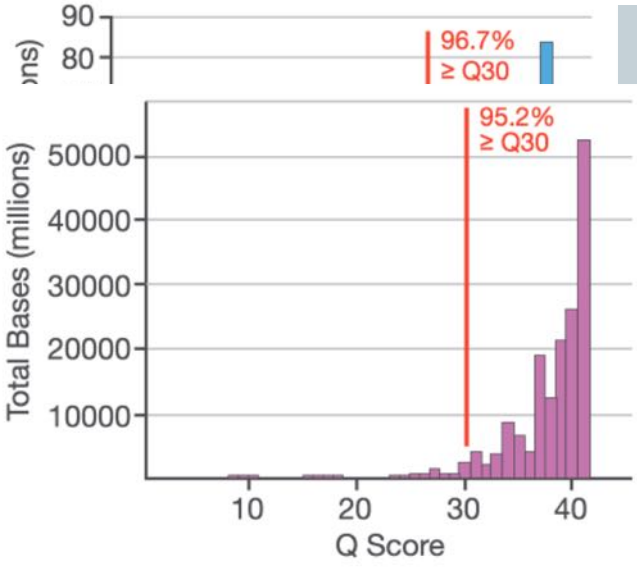
40

LANE	READ	CLUSTER PF (%)	%≥Q30	YIELD	ERROR RATE%	READS PF	DENSITY	TILES	LEGACY PHAS/PREPHAS (%)	INTENSITY	COMMENTS	STATUS
1	1	76.99 ±1.07	84.38	4.23 Gbp	1.16 ±0.16	56,450,488	339 ±9	72	0.211 / 0.166	3,865 ±334		QC Passed
	2(I)		91.34	395.00 Mbp	0.00 ±0.00				0.000 / 0.000	2,088 ±185		
	3(I)		90.14	394.72 Mbp	0.00 ±0.00				0.000 / 0.000	1,784 ±158		
	4		80.15	4.23 Gbp	1.29 ±0.18				0.223 / 0.189	2,919 ±303		
2	1	77.48 ±1.19	84.08	4.18 Gbp	1.23 ±0.24	55,713,284	332 ±6	72	0.228 / 0.174	3,586 ±177		QC Passed
	2(I)		90.65	389.79 Mbp	0.00 ±0.00				0.000 / 0.000	1,957 ±114		
	3(I)		90.08	389.66 Mbp	0.00 ±0.00				0.000 / 0.000	1,704 ±95		
	4		79.39	4.18 Gbp	1.36 ±0.21				0.223 / 0.194	2,765 ±167		
3	1	75.58 ±1.19	82.86	4.15 Gbp	1.50 ±0.26	55,376,164	339 ±4	72	0.221 / 0.158	3,252 ±384		QC Passed
	2(I)		90.38	387.60 Mbp	0.00 ±0.00				0.000 / 0.000	1,768 ±213		
	3(I)		89.01	387.35 Mbp	0.00 ±0.00				0.000 / 0.000	1,571 ±198		
	4		78.32	4.15 Gbp	1.64 ±0.23				0.217 / 0.191	2,511 ±312		
4	1	76.32 ±1.60	83.08	4.15 Gbp	1.46 ±0.34	55,336,016	335 ±9	72	0.230 / 0.161	3,419 ±419		QC Passed
	2(I)		90.01	387.31 Mbp	0.00 ±0.00				0.000 / 0.000	1,852 ±240		
	3(I)		89.08	386.99 Mbp	0.00 ±0.00				0.000 / 0.000	1,620 ±228		
	4		77.62	4.14 Gbp	1.64 ±0.34				0.229 / 0.199	2,575 ±369		



La calidad es dependiente de la tecnología, configuración, concentración de librería,

HiSeq
2×50 bp



Calidad Q30 en MiSeq

43

Quality Scores†

MiSeq Reagent Kit v2	MiSeq Reagent Kit v3
> 90% bases higher than Q30 at 1 × 36 bp	> 85% bases higher than Q30 at 2 × 75 bp
> 90% bases higher than Q30 at 2 × 25 bp	> 70% bases higher than Q30 at 2 × 300 bp
> 80% bases higher than Q30 at 2 × 150 bp	
> 75% bases higher than Q30 at 2 × 250 bp	

† A quality score (Q-score) is a prediction of the probability of an error in base calling. The percentage of bases > Q30 is averaged across the entire run.

Calidad Q30 en NextSeq

44

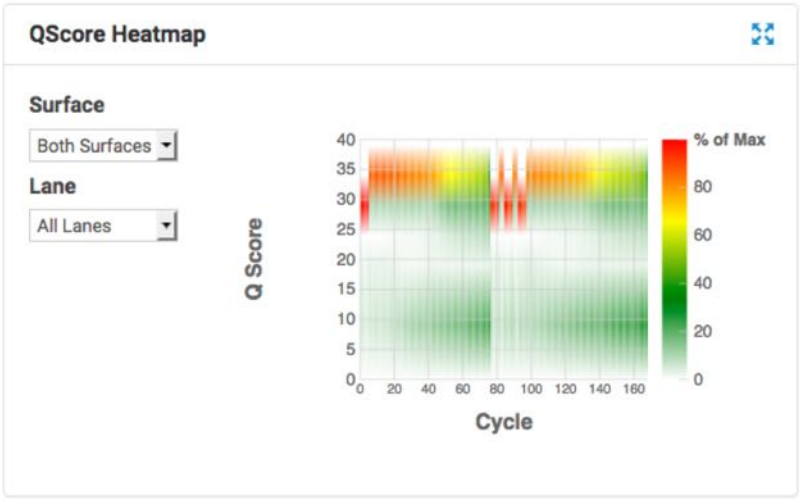
Quality Scores††

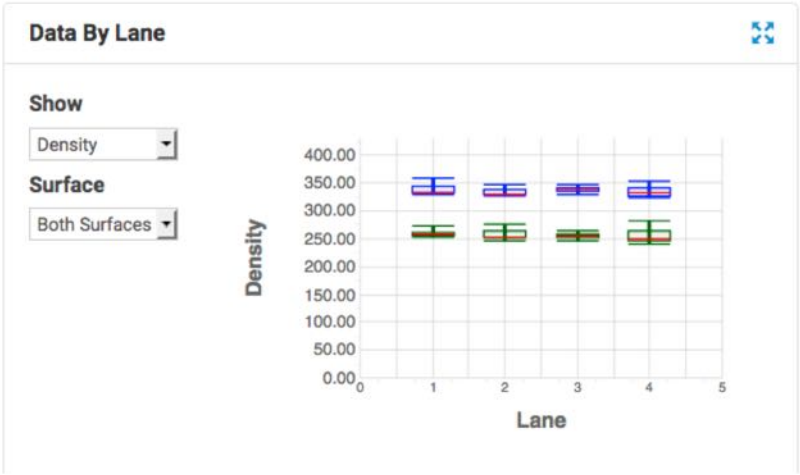
NextSeq 550 System High-Output Kit	NextSeq 550 System Mid-Output Kit
> 75% bases higher than Q30 at 2 × 150 bp	> 75% bases higher than Q30 at 2 × 150 bp
> 80% bases higher than Q30 at 2 × 75 bp	> 80% bases higher than Q30 at 2 × 75 bp
> 80% bases higher than Q30 at 1 × 75 bp	

††A quality score (Q-score) is a prediction of the probability of an error in base calling. The percentage of bases > Q30 is averaged across the entire run.

Valores históricos de calidad obtenidos en ChileGenomico

RUN NAME	INSTRUMENT	FLOWCELL ID	% PF	AVG %Q30	YIELD	OWNER	CREATED	SIZE	CYCLES	LANE & QC STATUS	RUN STATUS
WESSGPooling3	NB551...	HGLK5A...	76.60%	82.00%	36.53 Gbp	Ricardo ...	2017-11-2...	22 GB	76 | 76	QC Passed	Complete
Pool4MomiaNiñoelPlomo	NB551...	H3CY2B...	97.07%	96.91%	12.54 Gbp	Ricardo ...	2017-11-0...	7 GB	38 | 38	QC Passed	Failed
SEQMomia_DNAAntiguo	M0315...	0000000...	97.43%	90.52%	1.36 Gbp	Ricardo ...	2017-09-2...	2 GB	76 | 76	QC Passed	Complete
PoolExomeAgilentRun3	NB551...	HGL3NA...	87.40%	88.27%	26.64 Gbp	Ricardo ...	2017-08-0...	14 GB	76 | 76	QC Passed	Complete
WGSINIAII	NB551...	H5VC2B...	93.49%	83.12%	130.91 Gbp	Ricardo ...	2017-06-1...	64 GB	151 | 151	QC Passed	Complete
Poolprunus2	NB551...	HG2LYB...	94.35%	85.25%	122.51 Gbp	Ricardo ...	2017-06-0...	59 GB	151 | 151	QC Passed	Complete
AgilentExomasrun2	NB551...	HGLM3A...	83.80%	87.95%	35.23 Gbp	Ricardo ...	2017-05-2...	19 GB	76 | 76	QC Passed	Complete
AGILENTEXOME1	NB551...	HGL3YA...	86.13%	88.59%	29.42 Gbp	Ricardo ...	2017-05-2...	16 GB	76 | 76	QC Passed	Complete
EDP4-Run2	NB551...	HGL5HA...	77.39%	84.30%	32.48 Gbp	Ricardo ...	2017-05-1...	1 GB	76 | 76	QC Passed	Complete
Patagonia_WG_Run_1	NB551...	HWHYGB...	95.43%	88.16%	127.13 Gbp	Ricardo ...	2016-04-2...	59 GB	151 | 151	QC Passed	Complete


QScore Heatmap
Surface
Both Surfaces
Lane
All Lanes
Q Score
40
35
30
25
20
15
10
5
0
0 20 40 60 80 100 120 140 160
Cycle
% of Max
80
60
40
20
0


Data By Lane
Show
Density
Surface
Both Surfaces
Density
400.00
350.00
300.00
250.00
200.00
150.00
100.00
50.00
0.00
0 1 2 3 4 5
Lane

Densidad de Clusters

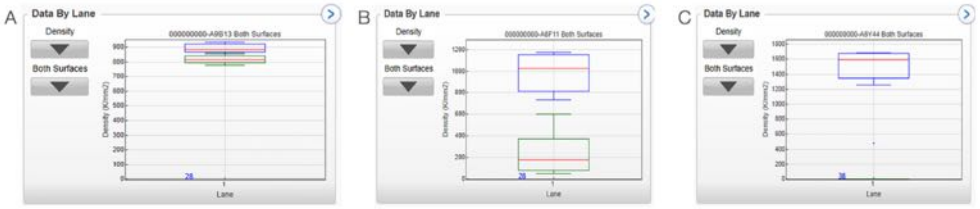
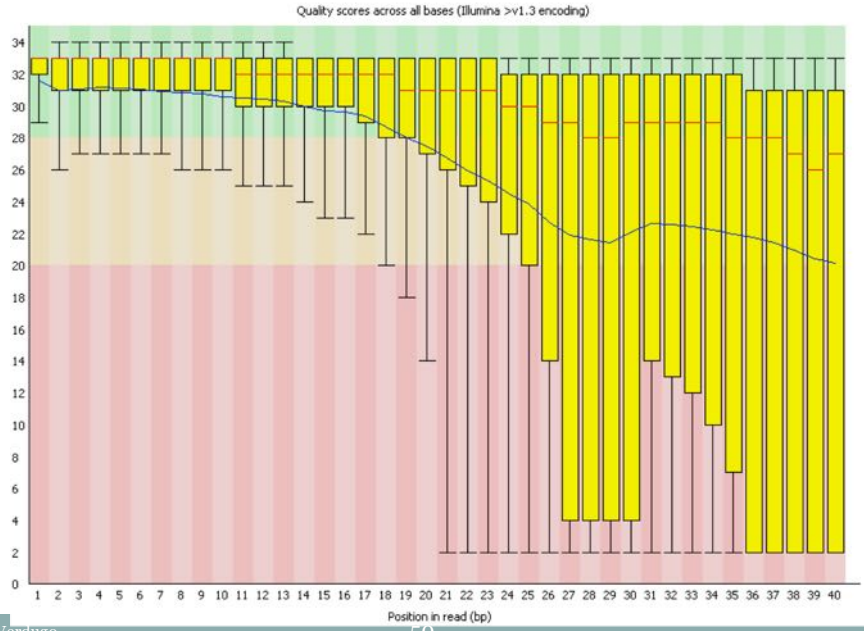


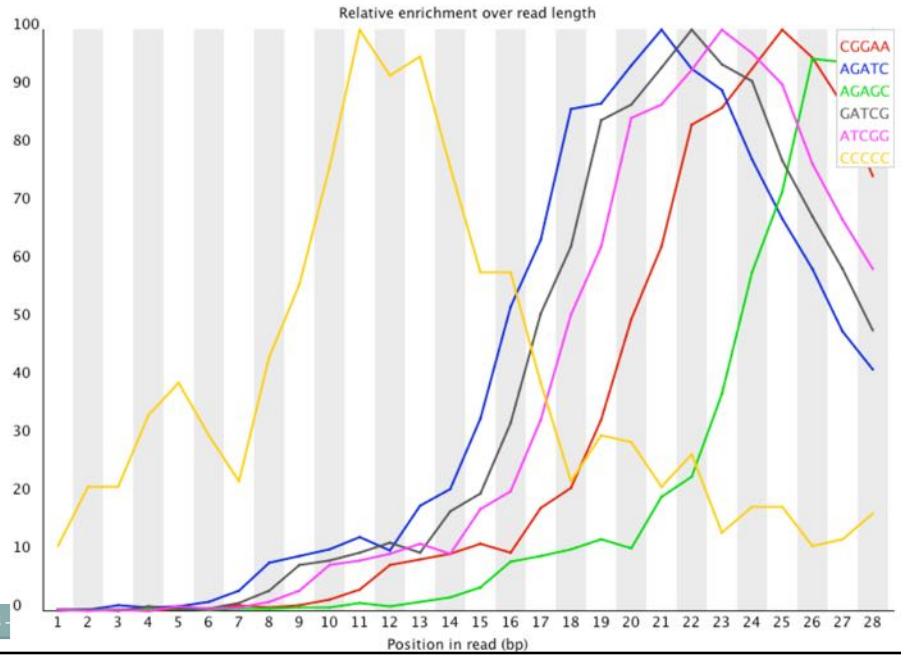
Figure 3: Data by Lane: Density. The blue boxes illustrate the raw cluster density range, the green boxes illustrate the %PF cluster density range, and the red lines indicate the median cluster density values. A) Optimal density. B) Overclustered. C) Severely overclustered.

Table 1: Optimal Raw Densities for Illumina sequencing systems

	MiniSeq™	MiSeq		NextSeq®	HiSeq 2500 Rapid Run (RR)	HiSeq 2500 High Output (HO)	
Versions	High and Mid Output	v2	v3	v2 High and Mid Output	v1 and v2	v3	v4
Raw Density (K/mm²)	170–220	1000–1200	1200–1400	170–220	850–1000	750–850	950–1050



KMERs


Relative enrichment over read length
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
Position in read (bp)
CGGAA
AGATC
AGAGC
GATCG
ATCGG
CCCCC