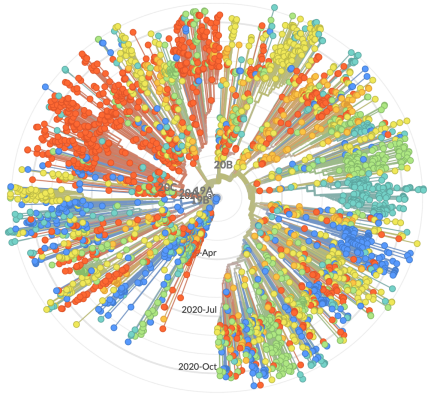


Session III:

Whole-Genome Analysis

Armando Reyes-Palomares

Department Biochemistry and Molecular Biology

Complutense University of Madrid

armandorp@ucm.es

Practical Sessions GAG

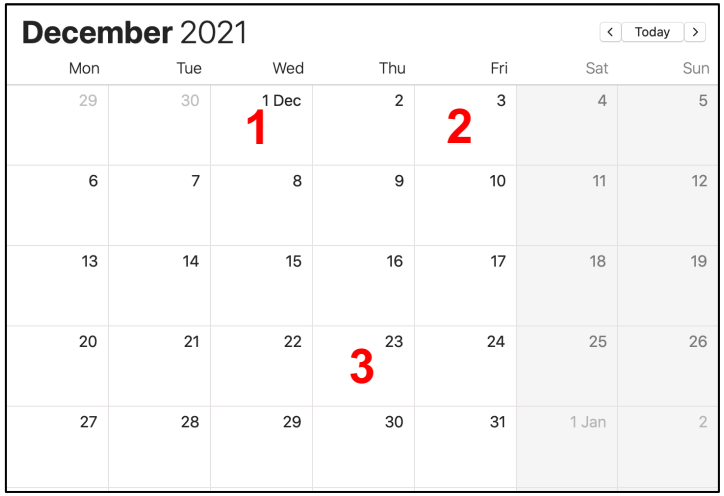
1. Introduction to R
2. Bioconductor: BiomaRt (Data mining)
3. Bioconductor: GenomicRanges (Coordinates)
4. Whole-Genome Analysis

armandorp@ucm.es

Proyecto Genómica Computacional

PROCESO:

1. Propuestas (1/12)
2. Visto Bueno a la propuesta (3/12)
3. **Entrega informe del proyecto (23/12)**
4. Presentación (14/1)



PROPUESTAS: 1 resumen 300 palabras

GRUPOS: 2 PERSONAS (máx.)

INFORMES: máximo 6 paginas (sin figuras, Referencias, anexos)

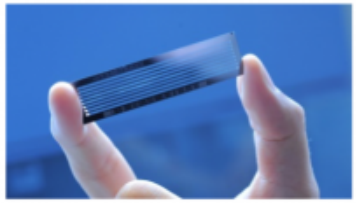
PRESENTACION: 15 min. + 5 de preguntas

armandorp@ucm.es

Table of contents

armandorp@ucm.es

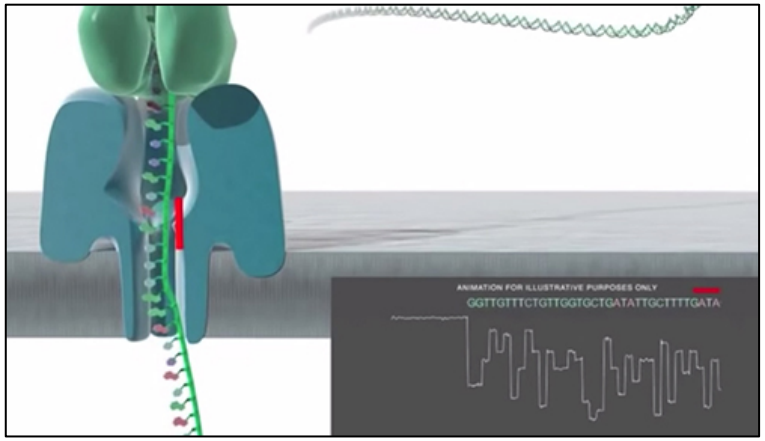
Illumina Platform



Chip, slide, flow cell…

HiSeq 2500

Oxford nanopore



armandorp@ucm.es

Illumina Platform

iSeq 100 System

MiniSeq System

MiSeq Series

NextSeq Series

Benchtop

Run Time	9–17.5 hours	4–24 hours	4–55 hours	12–30 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb
Maximum Reads Per Run	4 million	25 million	25 million †	400 million
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp

NextSeq Series

HiSeq Series

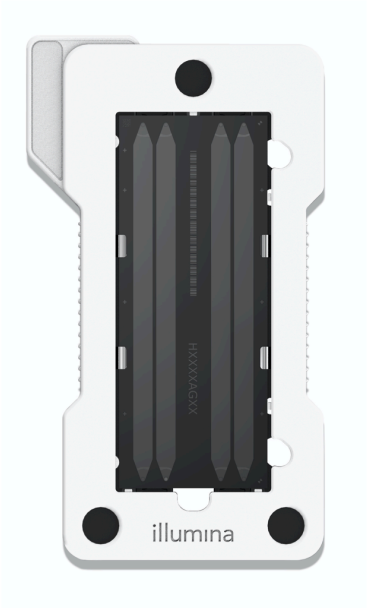
HiSeq X Series‡

NovaSeq 6000 System

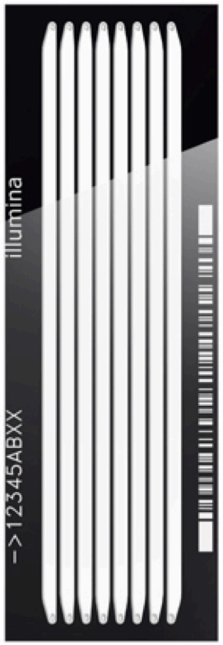
Production-Scale

Run Time	12–30 hours	< 1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)	< 3 days	16–36 hours (Dual S2 flow cells) 44 hours (Dual S2 flow cells)
Maximum Output	120 Gb	1500 Gb	1800 Gb	6000 Gb
Maximum Reads Per Run	400 million	5 billion	6 billion	20 billion
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp

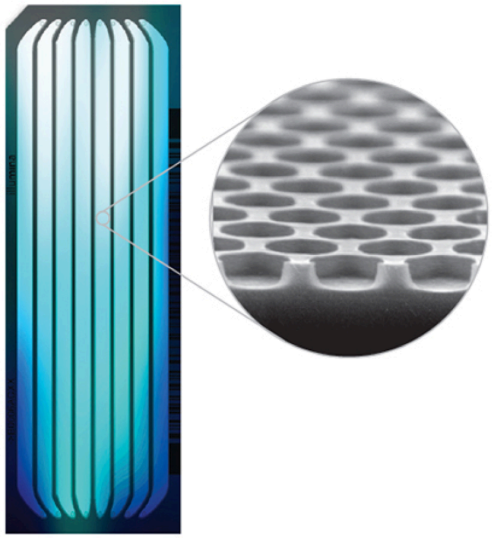
Illumina: flow cell



Nextseq500



HiSeq2500

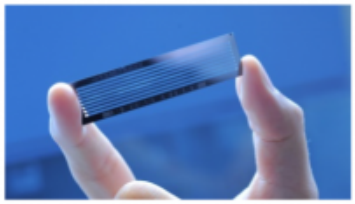


HiSeq3000/4000

Illumina: flow cell

armandorp@ucm.es

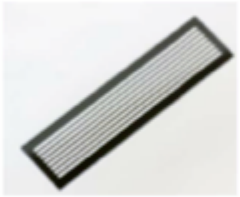
Illumina Platform



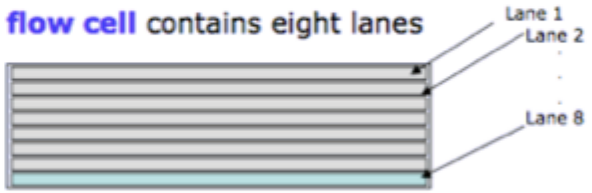
Chip, slide, flow cell.

HiSeq 2500

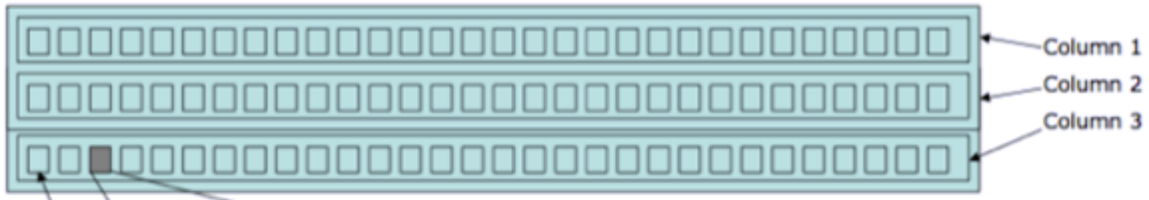
Flow cells



A **flow cell** contains eight lanes

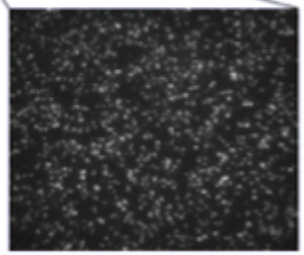


Each **lane/channel** contains **three columns** of tiles



Each **column** contains **100 tiles**

Tile



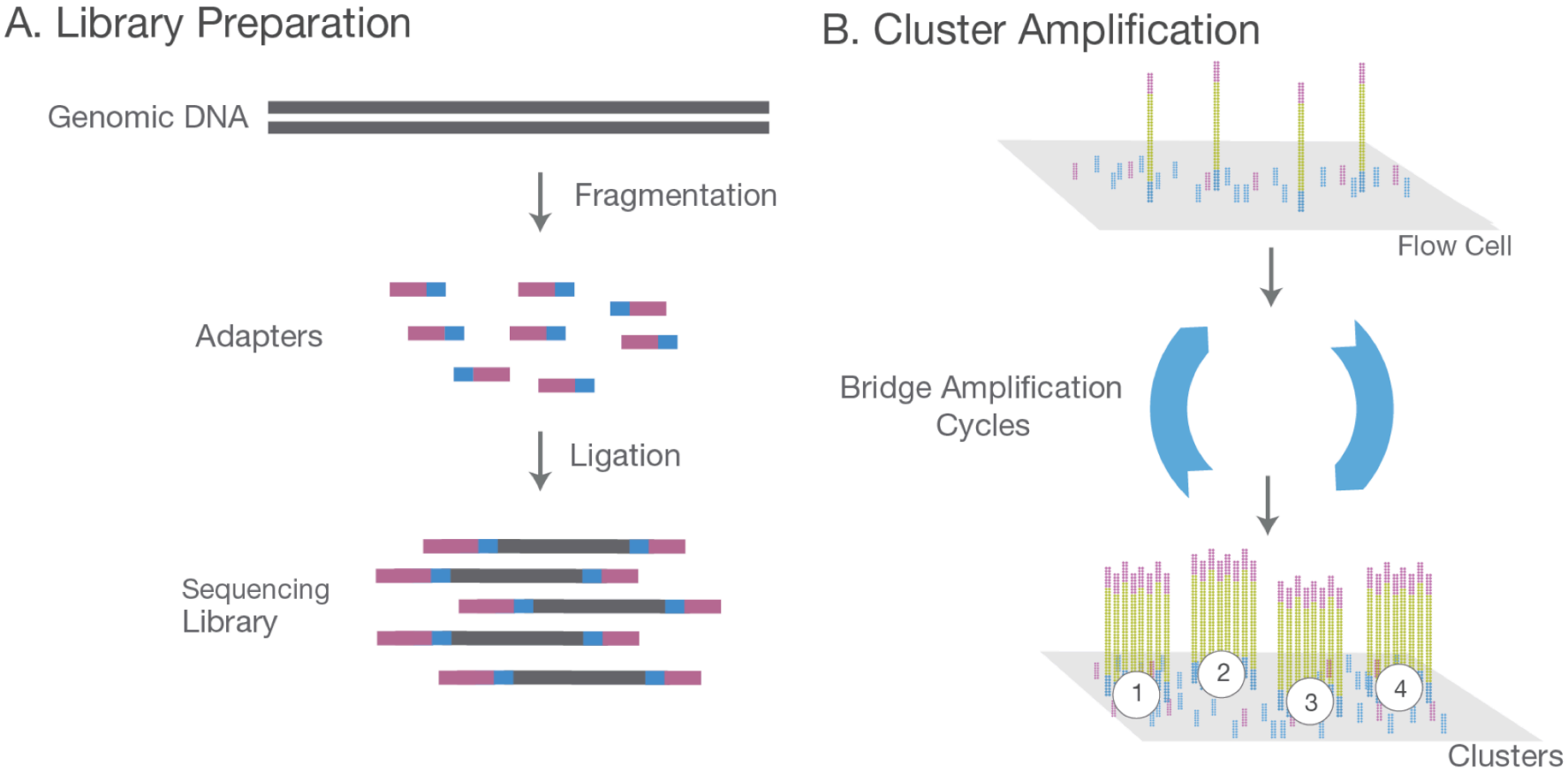
20K-30K Clusters

350 X 350 µm

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

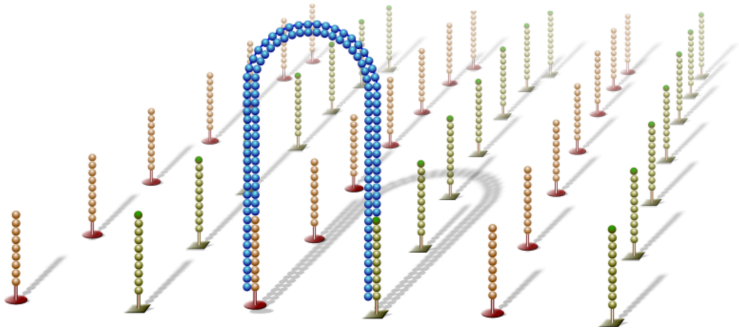
345,600 images for a 36-cycle run

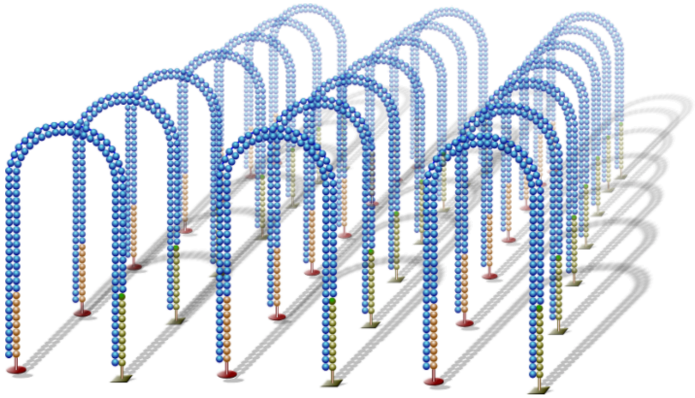
Illumina Platform

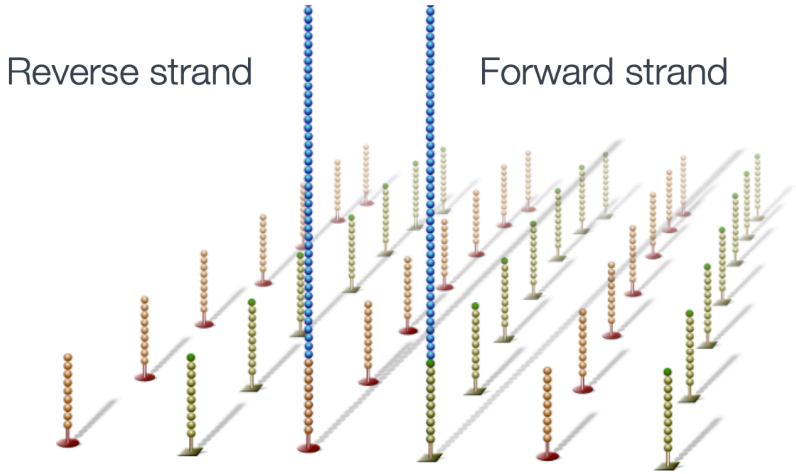


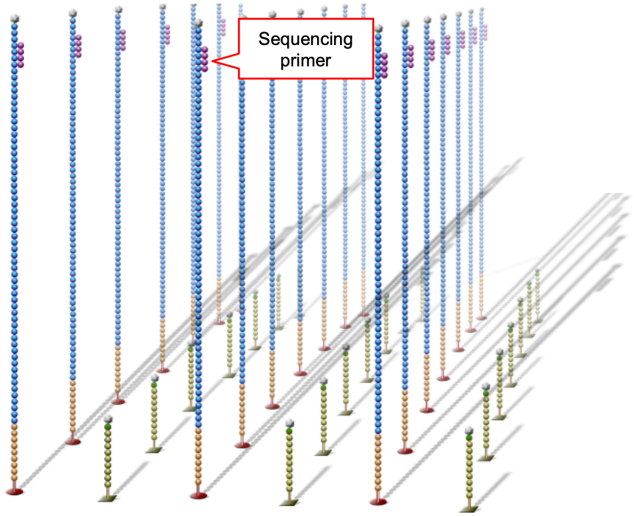
armandorp@ucm.es

Illumina: bridge amplification



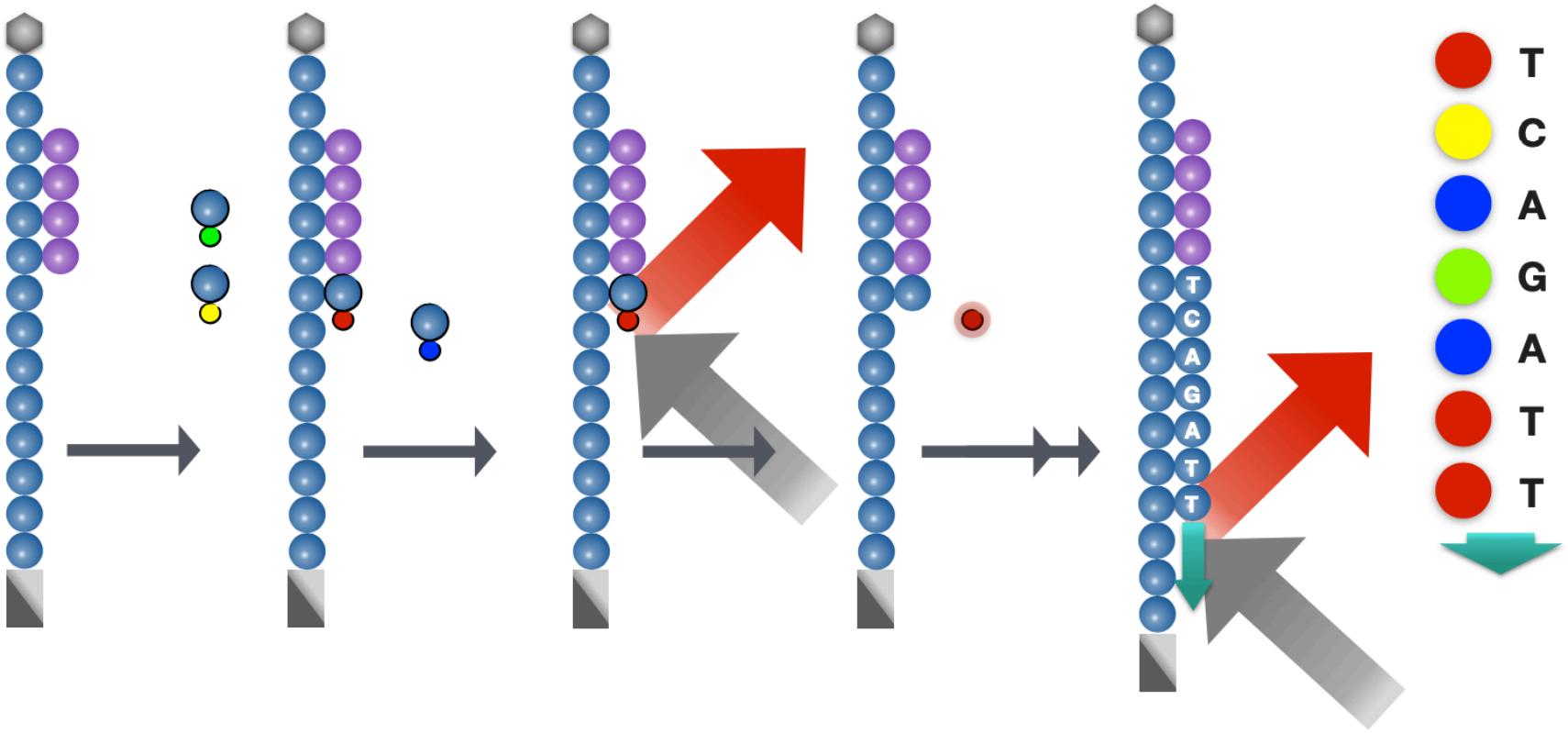






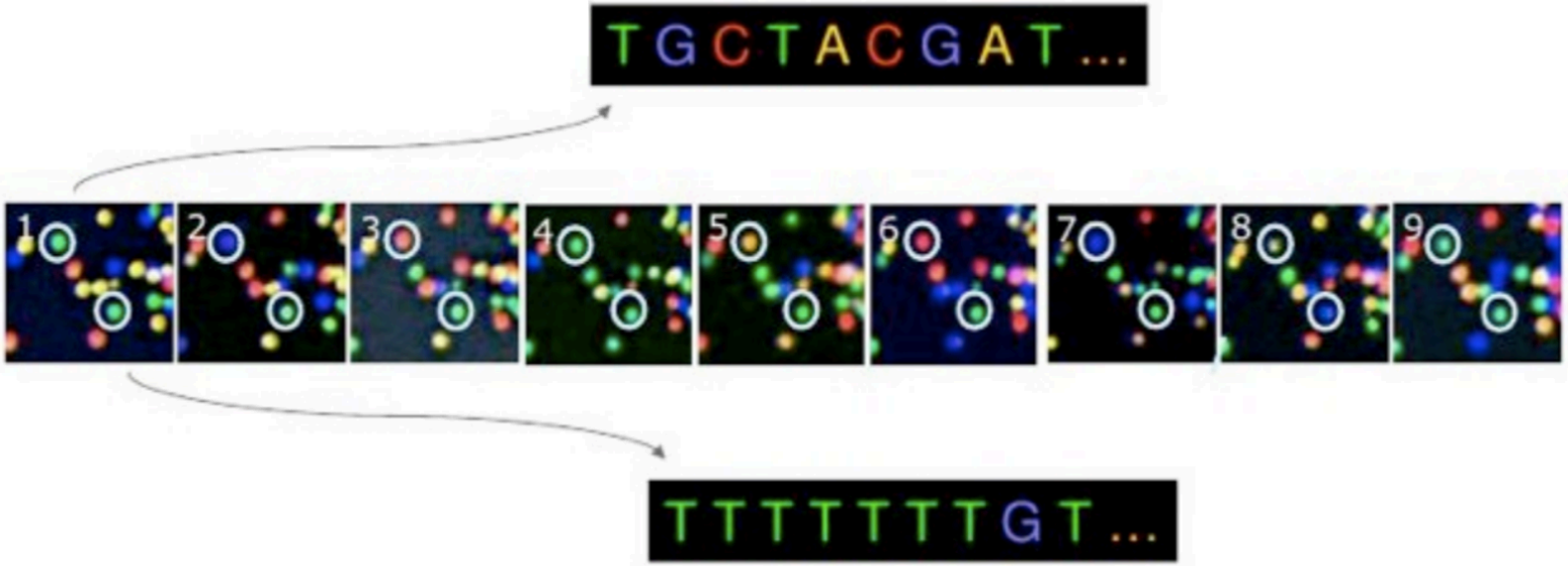
armandorp@ucm.es

Illumina: sequencing by synthesis



armandorp@ucm.es

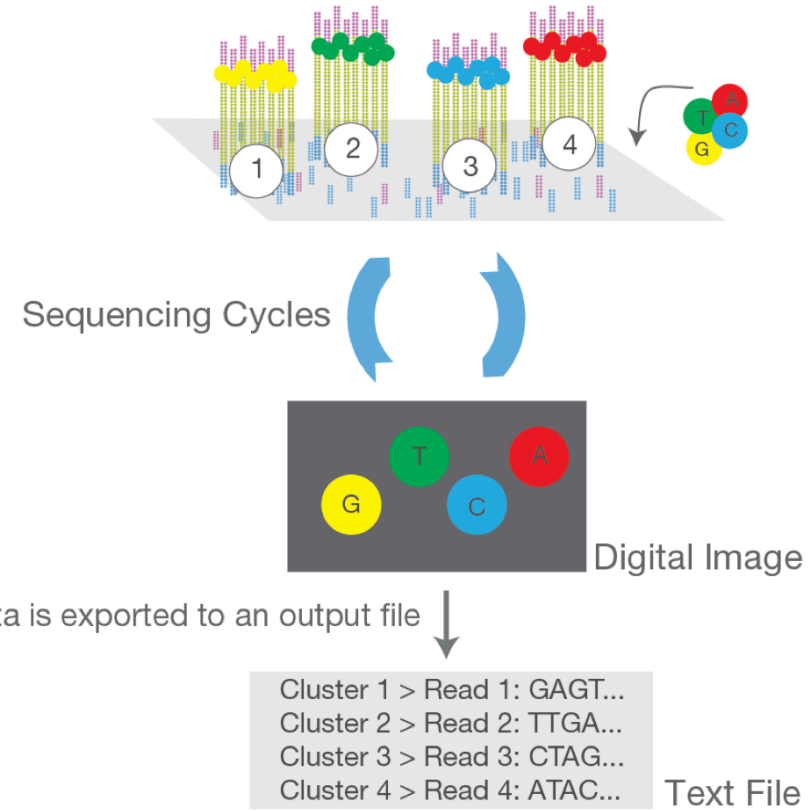
Illumina: base calling



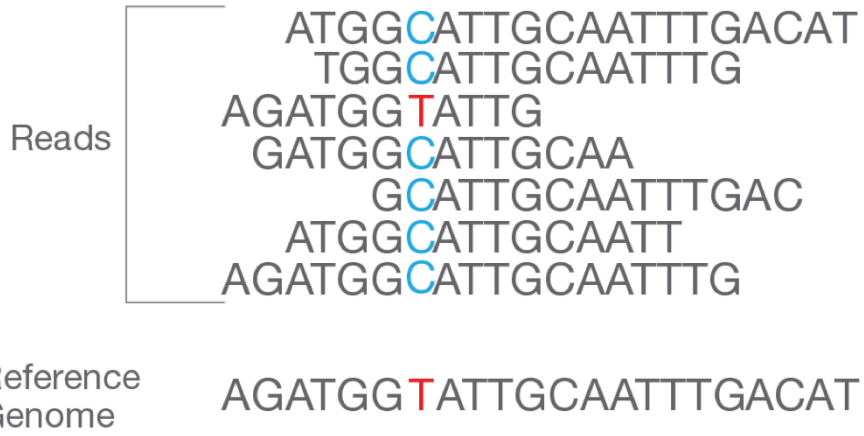
armandorp@ucm.es

Illumina Platform

C. Sequencing



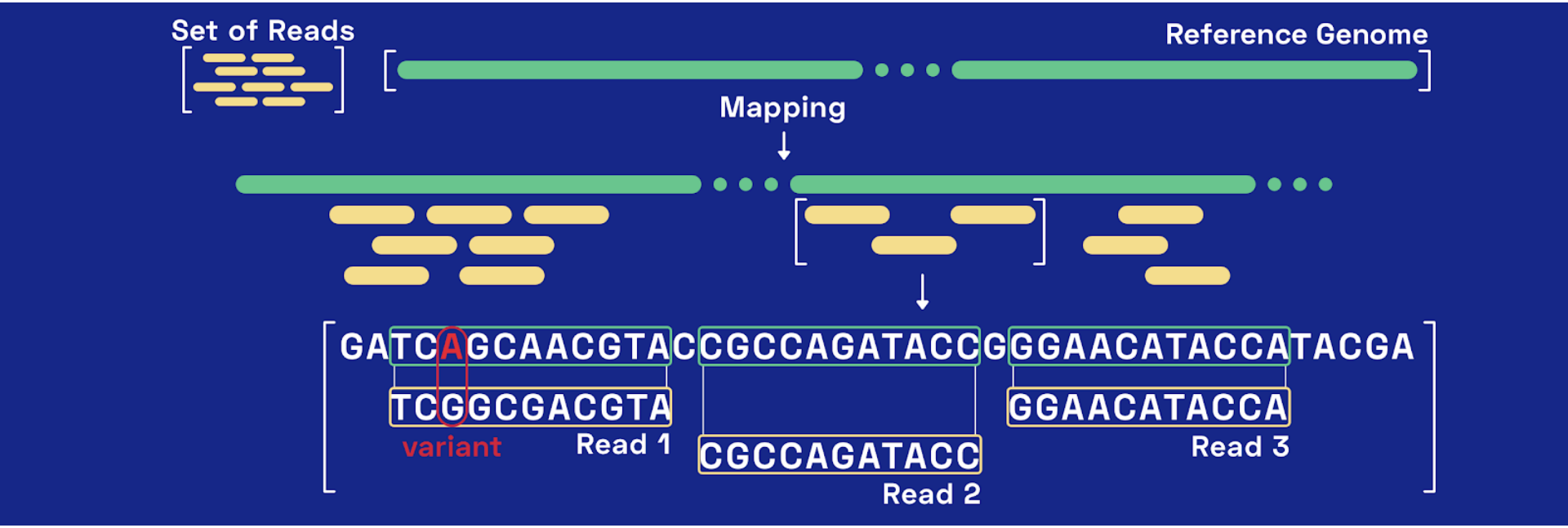
D. Alignment and Data Anaylsis



Nº clusters ~= Nº of reads
Nº sequencing cycles ~= Length of reads

armandorp@ucm.es

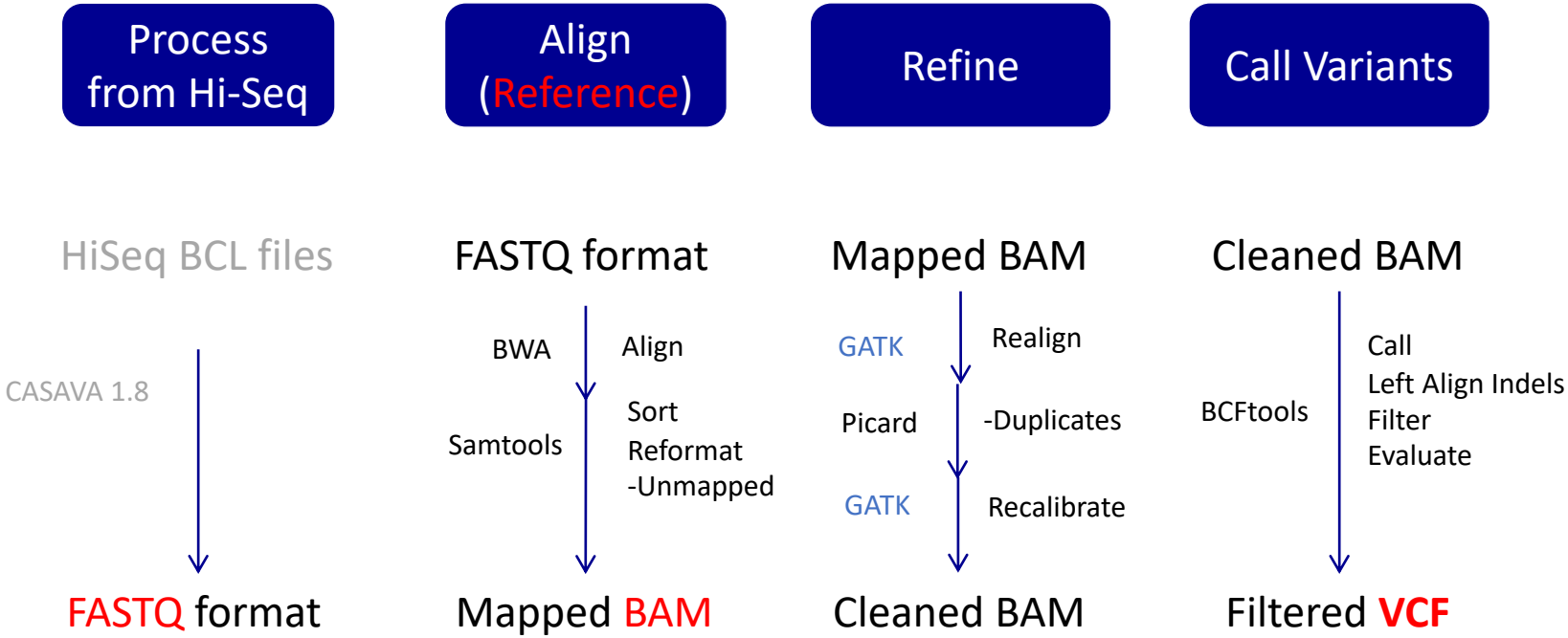
Aligning reads to Reference Genome



armandorp@ucm.es

Pipeline overview:

Alignment and Variant Calling



armandorp@ucm.es

Steps 1 and 2 in the tutorial

armandorp@ucm.es

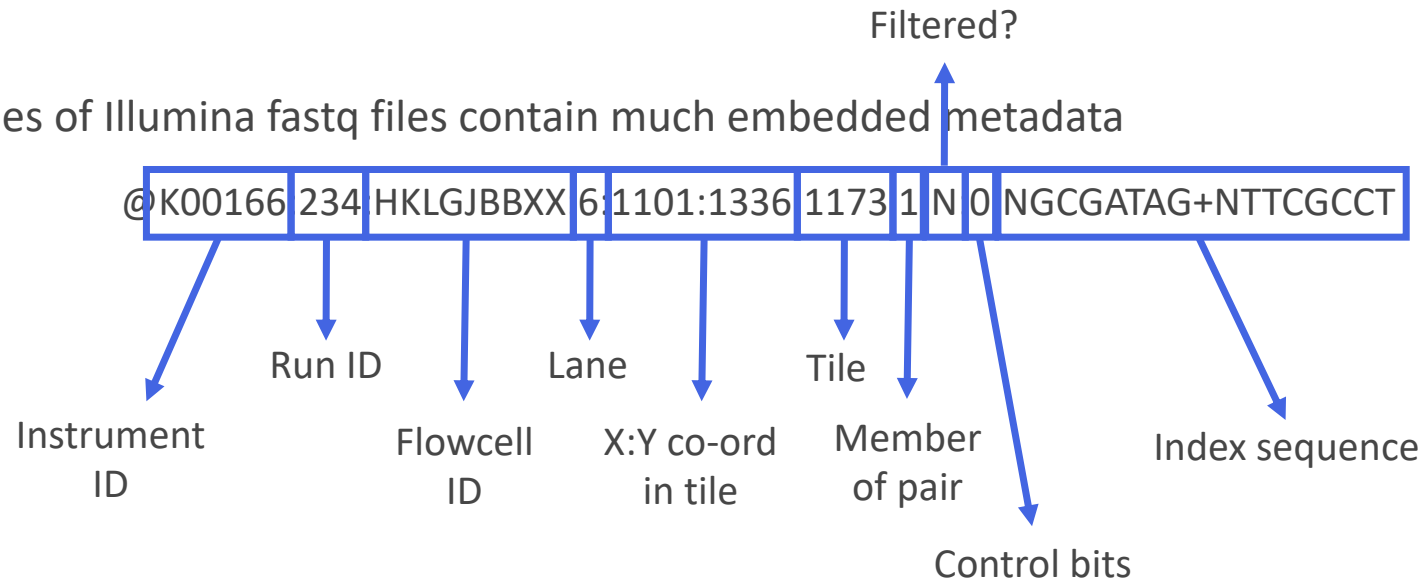
FASTQ format

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAAT
+
!''*((((***+))%%%++)(%%%%).1***
```

Four lines per sequence

1. '@' followed by ID, and optional description
2. Sequence
3. '+', optionally followed by ID
4. Phred Score Quality values (ASCII)

- Header lines of Illumina fastq files contain much embedded metadata



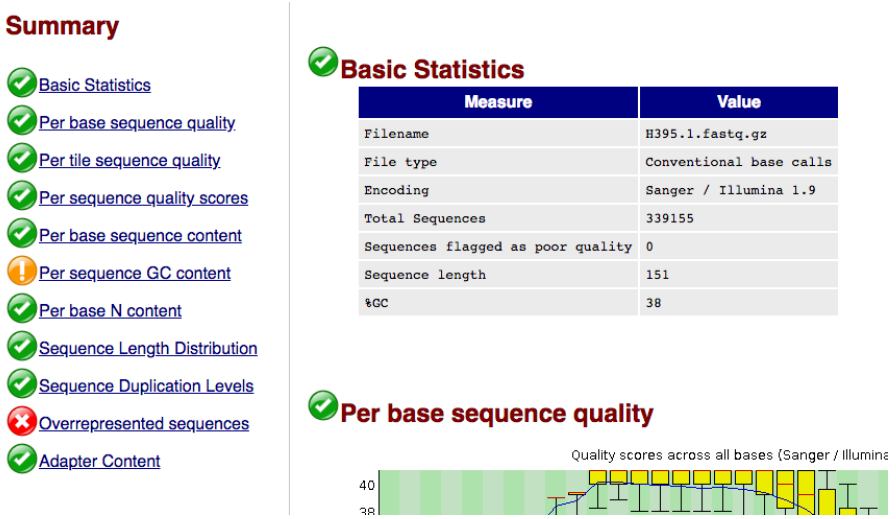
FASTQ format

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAAT
+
!''*((((***+))%%%++)(%%%%).1***
```

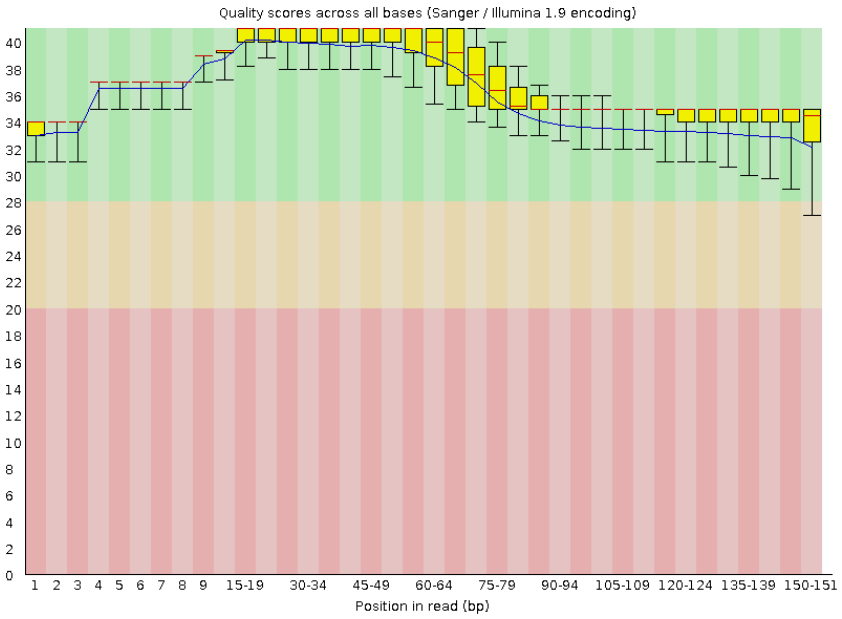
ASCII value	Character	Control character	ASCII value	Character	ASCII value	Character	ASCII value	Character
000	(null)	NUL	032	(space)	064	@	096	
001	☺	SOH	033	!	065	A	097	a
002	☻	STX	034	"	066	B	098	b
003	♥	ETX	035	#	067	C	099	c
004	♦	EOT	036	$	068	D	100	d
005	♣	ENQ	037	%	069	E	101	e
006	♠	ACK	038	&	070	F	102	f
007	(beep)	BEL	039	'	071	G	103	g
008	◘	BS	040	(	072	H	104	h
009	(tab)	HT	041	)	073	I	105	i
010	(line feed)	LF	042	*	074	J	106	j
011	(home)	VT	043	+	075	K	107	k
012	(form feed)	FF	044	'	076	L	108	l
013	(carriage return)	CR	045	-	077	M	109	m
014	♫	SO	046	.	078	N	110	n
015	☼	SI	047	/	079	O	111	o
016	►	DLE	048	0	080	P	112	p
017	◄	DC1	049	1	081	Q	113	q
018	↕	DC2	050	2	082	R	114	r
019	‼	DC3	051	3	083	S	115	s
020	¶	DC4	052	4	084	T	116	t
021	§	NAK	053	5	085	U	117	u
022	▬	SYN	054	6	086	V	118	v
023	↨	ETB	055	7	087	W	119	w
024	↑	CAN	056	8	088	X	120	x
025	↓	EM	057	9	089	Y	121	y
026	→	SUB	058	:	090	Z	122	z
027	←	ESC	059	;	091	[	123	{
028	(cursor right)	FS	060	<	092	\	124	|
029	(cursor left)	GS	061	=	093	]	125	}
030	(cursor up)	RS	062	>	094	^	126	~
031	(cursor down)	US	063	?	095	_	127	⌂

Interpreting FastQC Results

1. For each fastq file, output directory contains
   - fastqname_fastqc.zip
   - fastqname_fastqc.html
2. Zip file contains separate outputs
3. HTML file provides combined report in single file
4. View this using a web browser

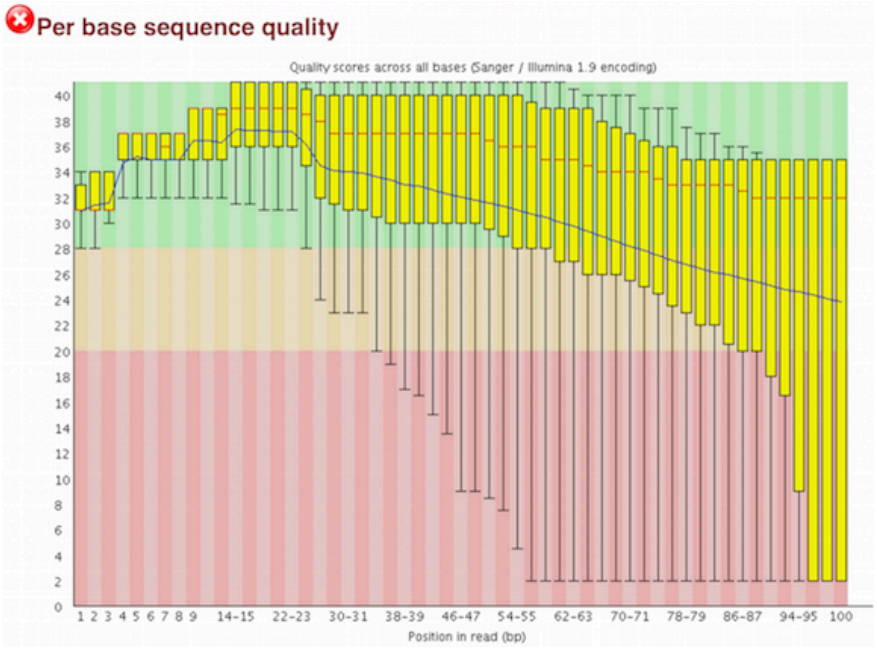


FastQC: Per-base sequence quality



- Box plot of quality scores along length or reads
  - red line: median
  - Blue line: mean
  - yellow box: inter-quartile range (25%-75%)
  - Whiskers: 10% & 90% range
- Quality score (y axis) is Phred-like value
- Typical cutoff 20
  - But depends on what you are doing…
- Typically see decrease in quality along read for Illumina sequence by problem data resolution: Read trimming

FastQC: Per-base sequence quality



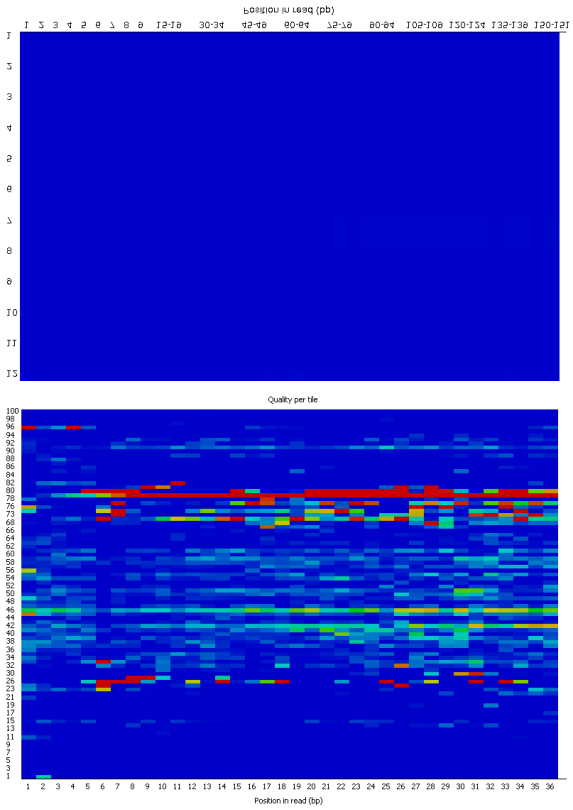
Signal decay: As sequencing proceeds, the fluorescent signal intensity decays with each cycle, yielding decreasing quality scores at the **3' end** of the read.

1. Degrading fluorophores
2. A proportion of the strands in the cluster not being elongated

Phasing: As the number of cycles increases, the signal starts to blur as the cluster loses synchronicity.

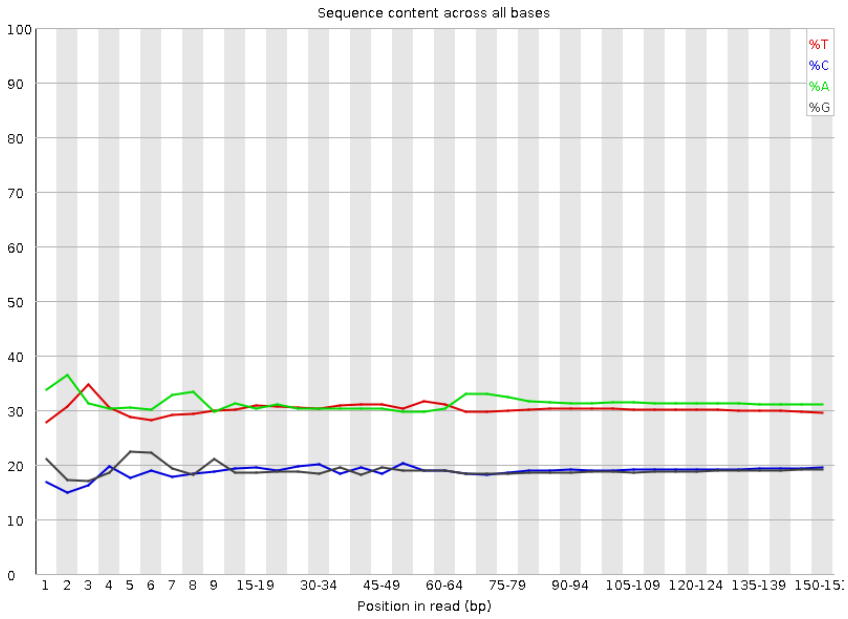
1. Incomplete removal of the 3' terminators and fluorophores
2. Incorporation of nucleotides without effective 3' terminators

FastQC – Per Tile Sequence Quality



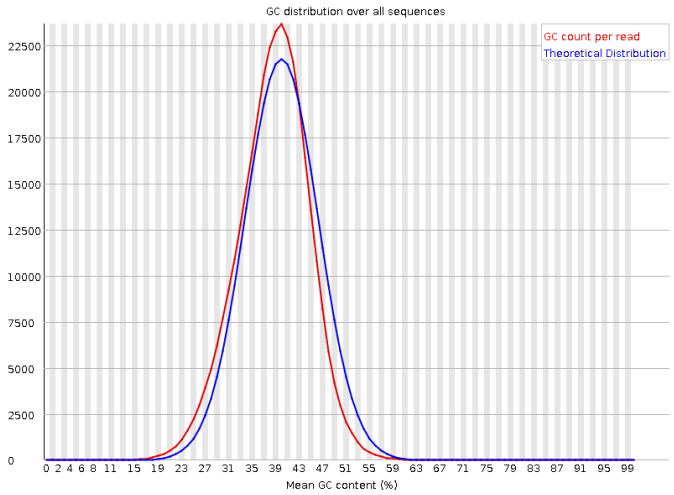
- Flowcell position of read determined from fastq ID
- Informs on localised problem on flowcell during run
- All blue – everything's good
- Otherwise…
  - Localised region: air bubble?
  - Broader region: dirt on flowcell?

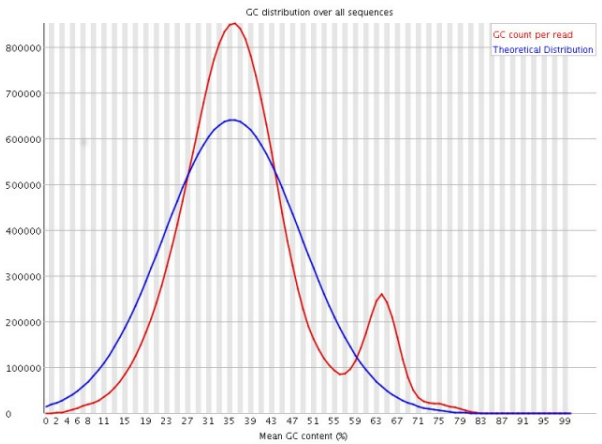
FastQC – Per-base sequence content



- Indicates proportion of A,C,T,G across length of read
- We seem to have a slightly AT-rich genome
- Common to see uneven distribution at beginning of read
  - Effect of library prep method
  - Technical bias but not a particular problem – affects base proportions but not specific sequences

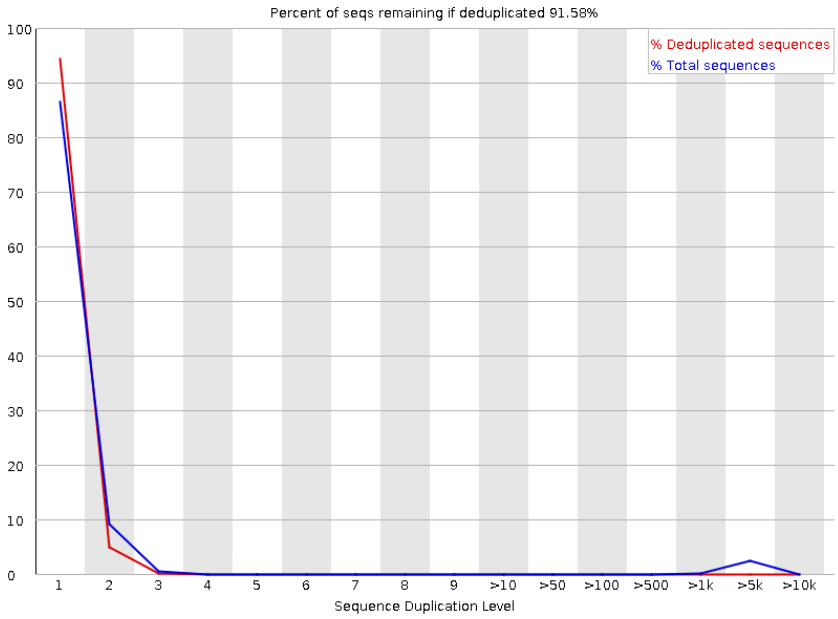
FastQC – Per-sequence GC content





- Distribution should be close to Gaussian (normal)
- Issues with library
  - Additional sharp peaks? Primer dimers?
  - Bimodal distribution? – contamination?

FastQC – Sequence Duplication Levels



- Indicates duplication level of sequences
  - Y-axis: % sequences
  - X-axis: Duplication level
- Duplicate sequences can occur through
  - Biological duplicates – real instances of replication of sequence
  - PCR duplicates – Overamplification of library
    - Insufficient diversity in library
  - Common in RNA-Seq – deep sequencing to find low abundance transcripts = number of copies of high abundance transcripts
  - Problematic for variant calling
  - Resolution: Removal of duplicate reads

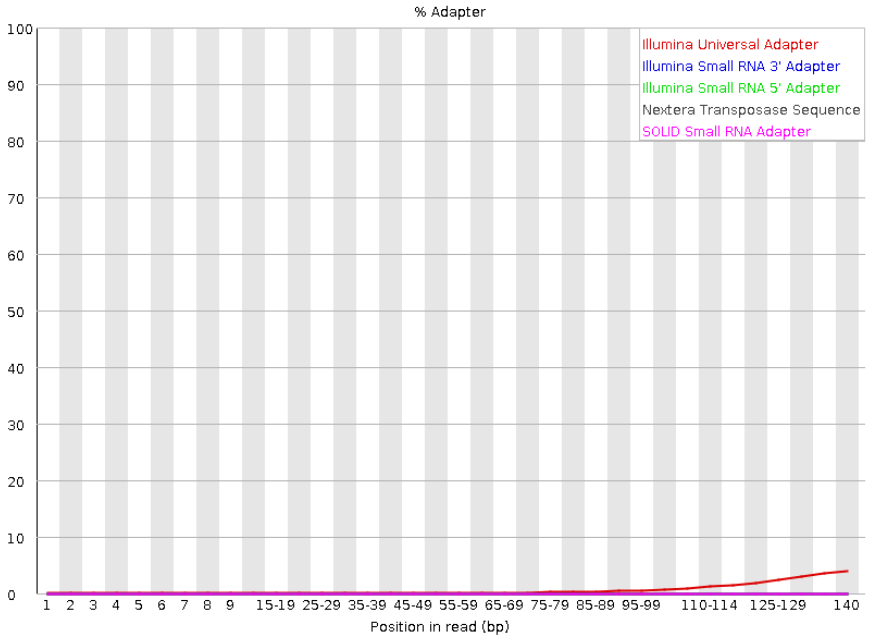
FastQC – Overrepresented sequences

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGC	9093	2.6810750246937243	TruSeq Adapter, Index 2 (100% over 50bp)
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATG	1034	0.30487535197770926	TruSeq Adapter, Index 2 (100% over 49bp)

- Library should contain range of sequences
- Should not be a significant representation of particular sequence
- Overrepresented sequences may be
  - Contaminants
  - Biologically relevant
- Sequences of >0.1% of total listed, with attempted identification of source

FastQC – Adapter content

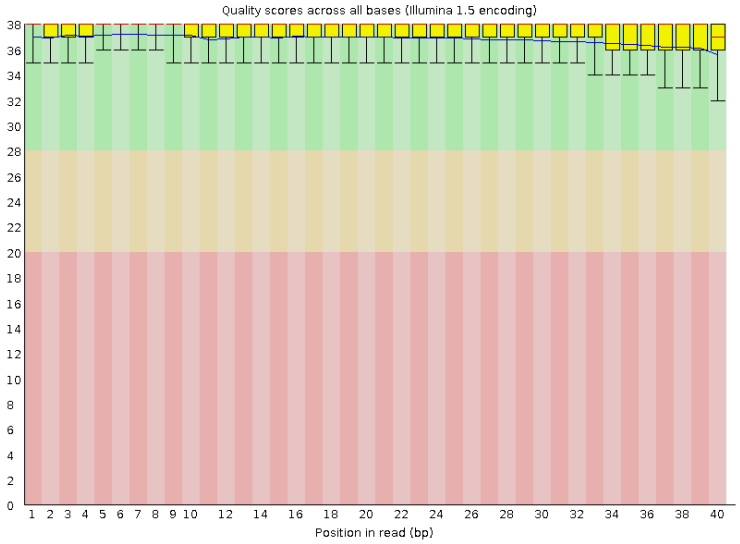


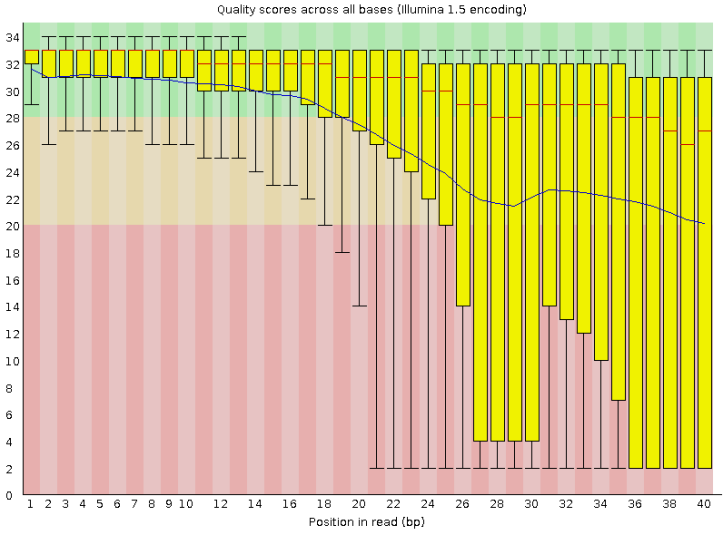
- Insert size shorter than read-length will result in read-through into adapter
- Plot of proportion of reads at each position which consist of adapter sequence
- May also be identified in ‘overrepresented sequences’ analysis
- Resolution: Carry out adapter trimming

FASTQC

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

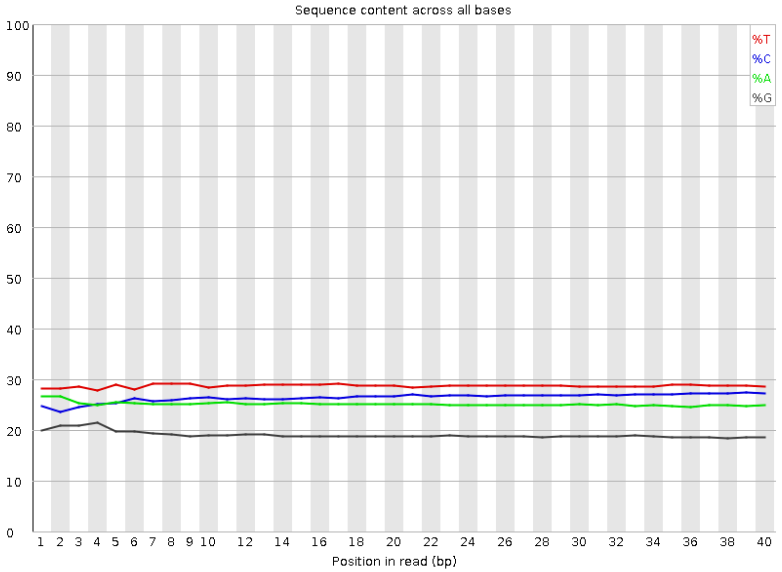
- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

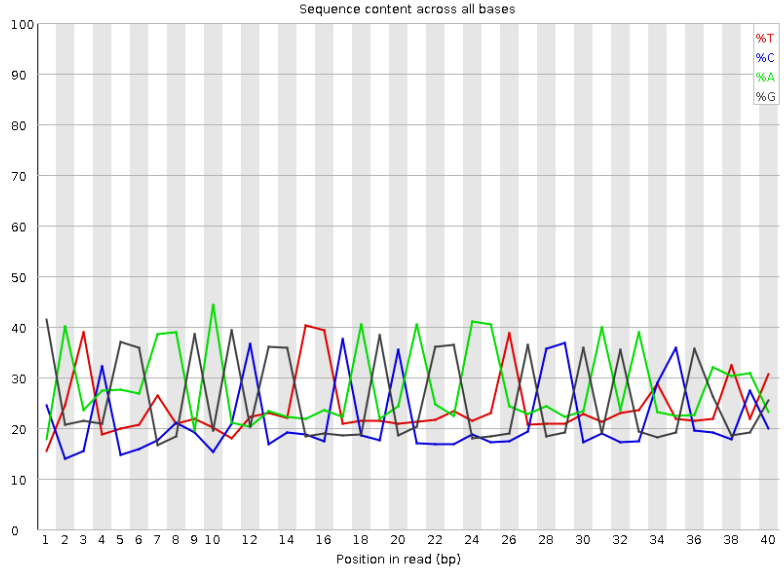


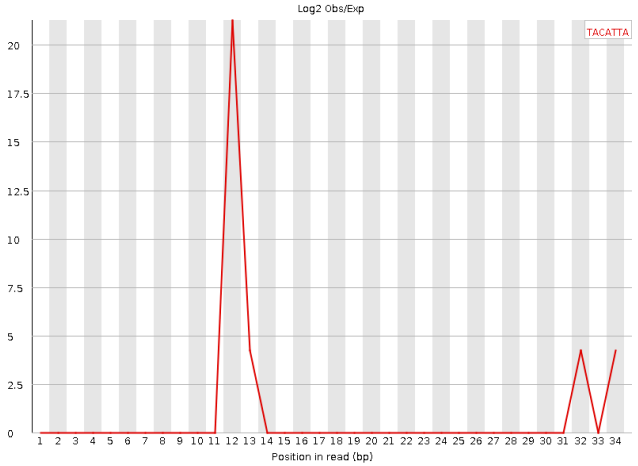


http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

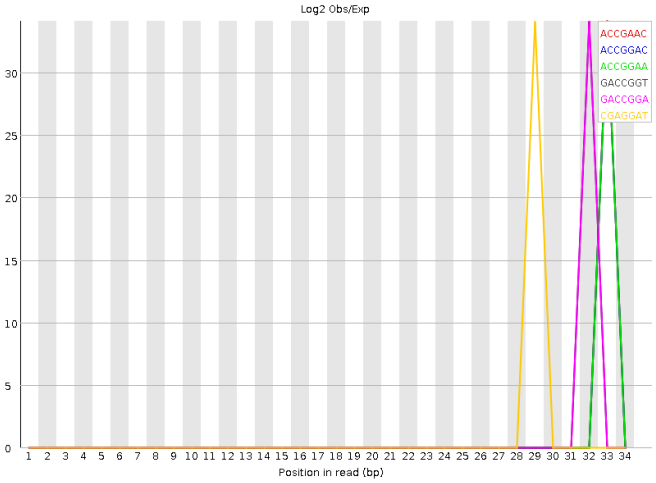
FASTQC Adapter Contamination







Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
TACATTA	40	0.003151852	21.2465	12



Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
ACCGAAC	35	1.0615131E-6	34.067673	33
ACCGGAC	30	1.4503141E-5	34.06767	33
ACCGGAA	55	3.092282E-11	34.06767	33
GACCGGT	20	0.0027499169	34.06767	32
GACCGGA	95	0.0	34.06767	32

Proceed to steps 3 and 4 of the tutorial

armandorp@ucm.es

SAM/BAM Format

```
[benpass align_genotype]$ samtools view allY.recalibrated.merge.bam
                                                    2     3     4   5     6          8     9
1 HW-ST605:127:B0568ABXX:2:1201:10933:3739        147   chr1  27675 60    101M  =     27588 -188
   10 TCATTTTATGGCCCCTTCTTCCTATATCTGGTAGCTTTTAAATGATGACCATGTAGATAATCTTTATTGTCCCTCTTTCAGCAGACGGTATTTTCTTATGC
   11 =7;:;<=??<=BCCEFFEJFCEGGEFFDF?BEA@DEDFEFFDE>EE@E@ADCACB>CCDCBACDCDDDAB@@BCADDCBC@BCBB8@ABCCCDCBDA@>:/
      RG:Z:86-191

HW-ST605:127:B0568ABXX:3:1104:21059:173553        83    chr1  27682 60    101M  =     27664 -119
      ATGGCCCCTTCTTCCTATATCTGGTAGCTTTTAAATGATGACCATGTAGATAATCTTTATTGTCCCTCTTTCAGCAGACGGTATTTTCTTATGCTACAGTA
      8;8.7::<?=BDHFHGFFDCGDAACCABHCCBDFBE</BA4//BB@BCAA@CBA@CB@ABA>A??@B@BBACA>?;A@8??CABBBA@AAAA?AA??@BB0
      RG:Z:SDH023
* Many fields after column 12 deleted (e.g., recalibrated base scores) have been deleted for improved readability
```

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{29}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next segment
8	PNEXT	Int	$[0,2^{29}-1]$	Position of the mate/next segment
9	TLEN	Int	$[-2^{29}+1,2^{29}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

SAM/BAM Format

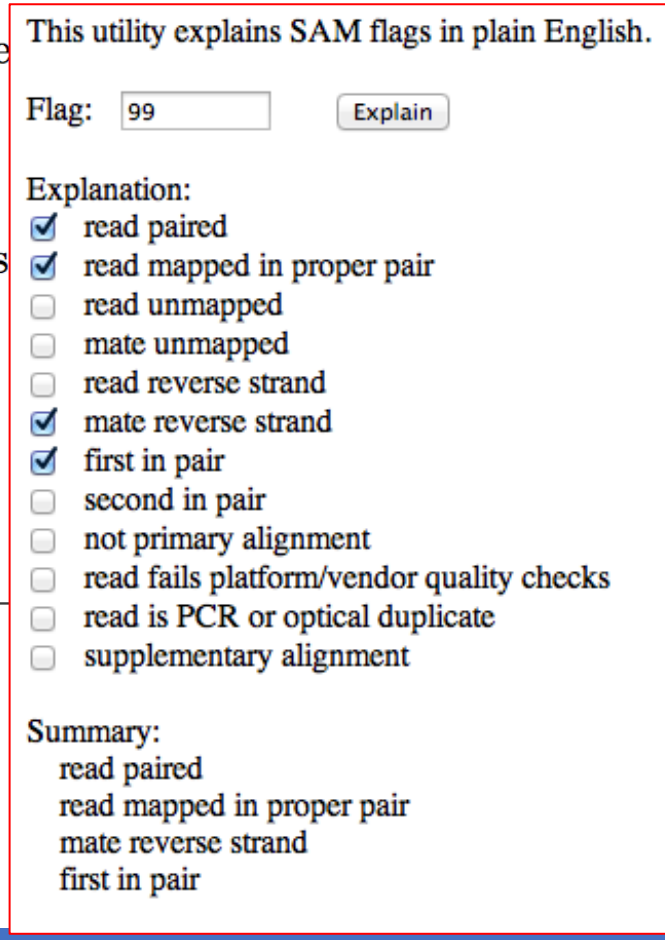
- **flexible** : compatible with multiple alignment programs
- **simple**: easy to generate and convert
- compact in file size
- works on a stream : low memory footprint
- Indexable for efficiency
- SAM is human readable, BAM is compressed

armandorp@ucm.es

SAM Flags

Bit	Description
0x1	template having multiple segments in sequencing
0x2	each segment properly aligned according to the aligne
0x4	segment unmapped
0x8	next segment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next segment in the template being revers
0x40	the first segment in the template
0x80	the last segment in the template
0x100	secondary alignment
0x200	not passing quality controls
0x400	PCR or optical duplicate
0x800	supplementary alignment

http://picard.sourceforge.net/explain-flags.html



CIGAR format

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

Before alignment

```
RefPos:      1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
Reference:   C  C  A  T  A  C  T  G  A  A  C  T  G  A  C  T  A  A  C
Read: ACTAGAATGGCT
```

After alignment

```
RefPos:      1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16 17 18 19
Reference:   C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T  A  A  C
Read:              A  C  T  A  G  A  A     T  G  G  C  T
```

```
POS: 5
CIGAR: 3M1I3M1D5M
```


armandorp@ucm.es

Proceed to steps 5 and 6 of the tutorial

armandorp@ucm.es

VCF format

HEADER

```
##fileformat=VCFv4.1
##fileDate=20090805
##tcgaversion=1.1
##vcfProcessLog=<InputVCF=<file1.vcf>,InputVCFSource=<caller1>,InputVCFVer=<1.0>,InputVCFParam=<a1,b>,InputVCFgeneAnno=<anno1.gaf>>
##reference=ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/special_requests/GRCh37-lite.fa
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
```

INFO meta-information

```
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
```

FILTER meta-information

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

FORMAT meta-information

```
##SAMPLE=<ID=NORMAL,Individual=TCGA-01-1000,File=TCGA-01-1000-1.bam,Platform=Illumina,Source=dbGAP,Accession=1234>
##SAMPLE=<ID=TUMOR,Individual=TCGA-01-1000,File=TCGA-01-1000-2.bam,Platform=Illumina,Source=dbGAP,Accession=4567>
##PEDIGREE=<Name_0=TUMOR,Name_1=NORMAL>
```

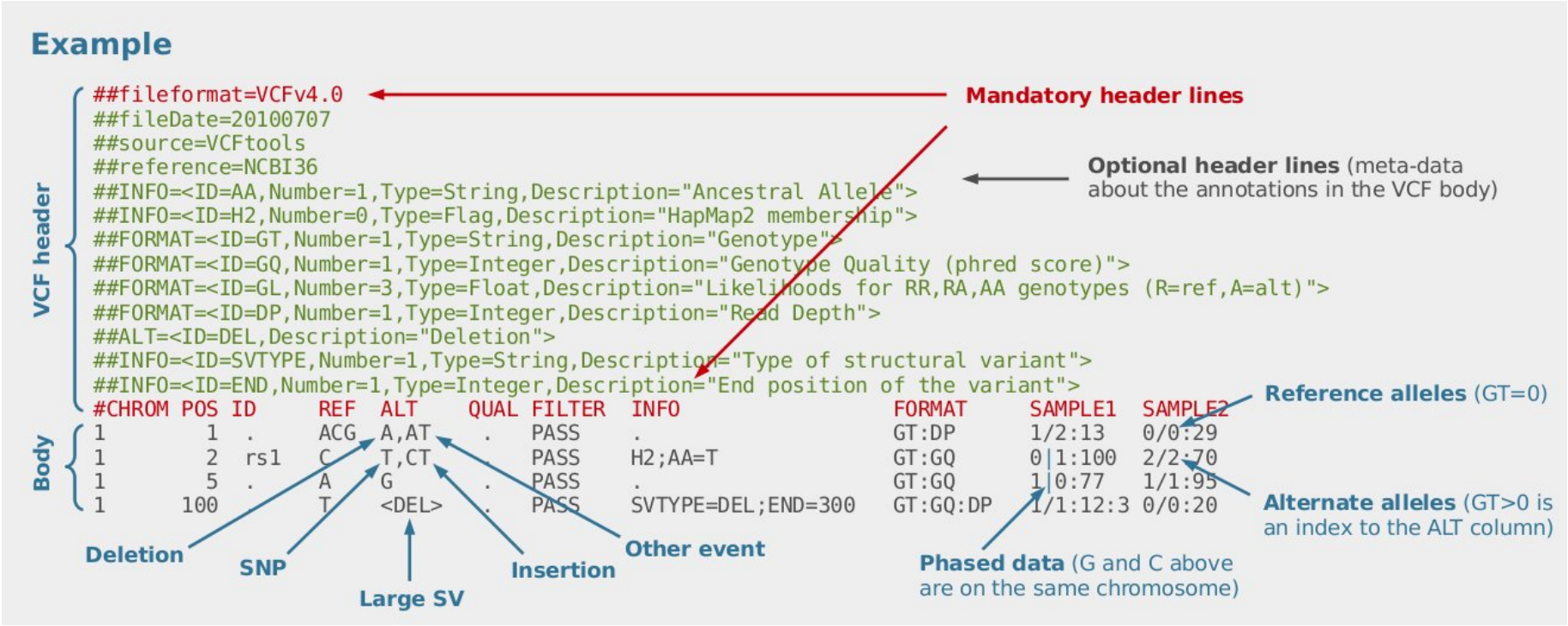
Fixed fields

Optional: FORMAT field specifying data type + Per-sample genotype data

BODY

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NORMAL	TUMOR
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51
20	1234567	microsat1	GTC	G,GTCTC	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2

VCF format


Example
VCF header
Body
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT . PASS . GT:DP 1/2:13 0/0:29
1 2 rs1 C T,CT . PASS H2;AA=T GT:GQ 0|1:100 2/2:70
1 5 . A G . PASS . GT:GQ 1|0:77 1/1:95
1 100 T <DEL> . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

Proceed to following steps in the tutorial

armandorp@ucm.es