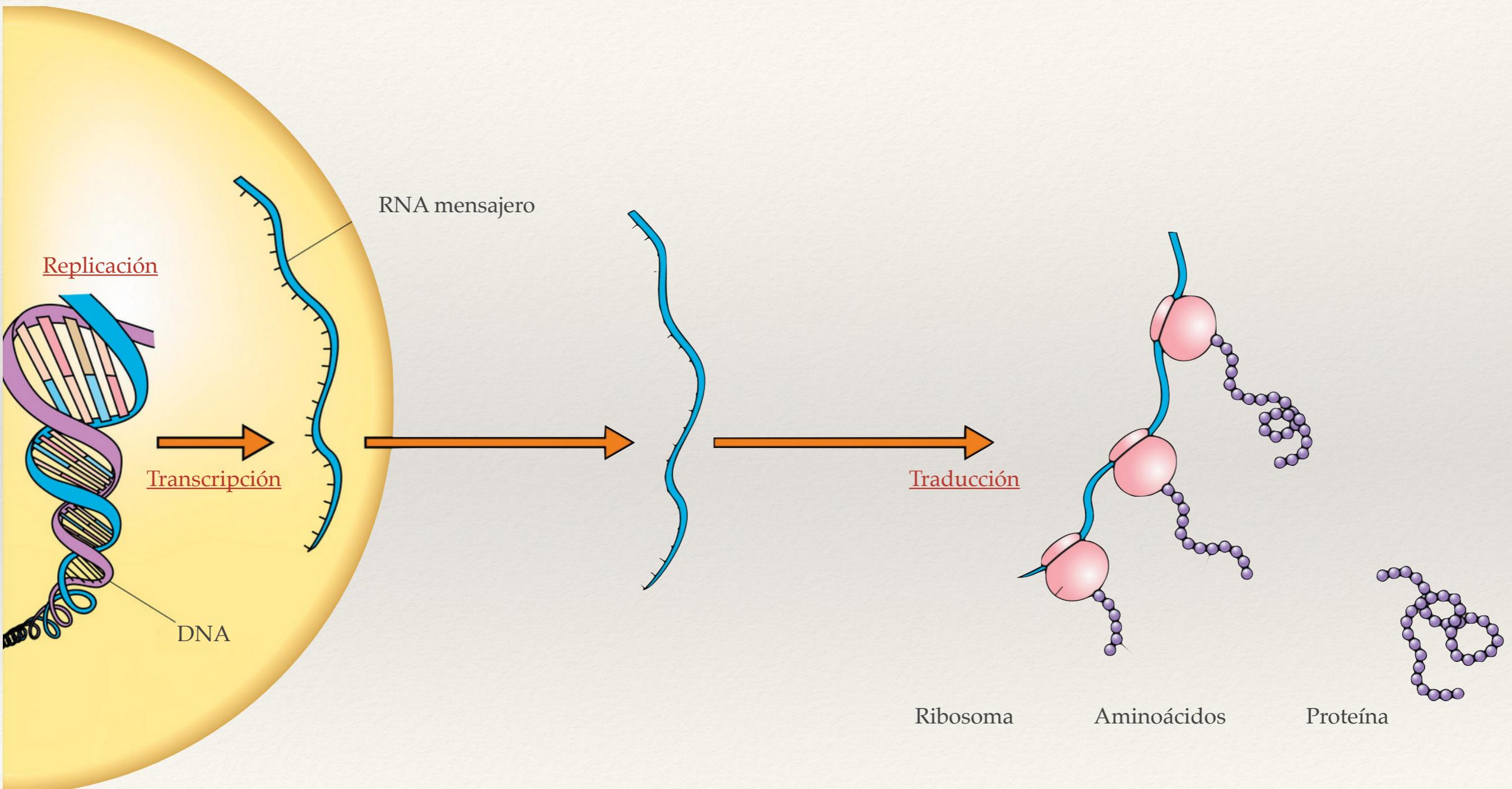


Introducción express a la Bioinformática

Tecnologías de secuenciación de nueva generación

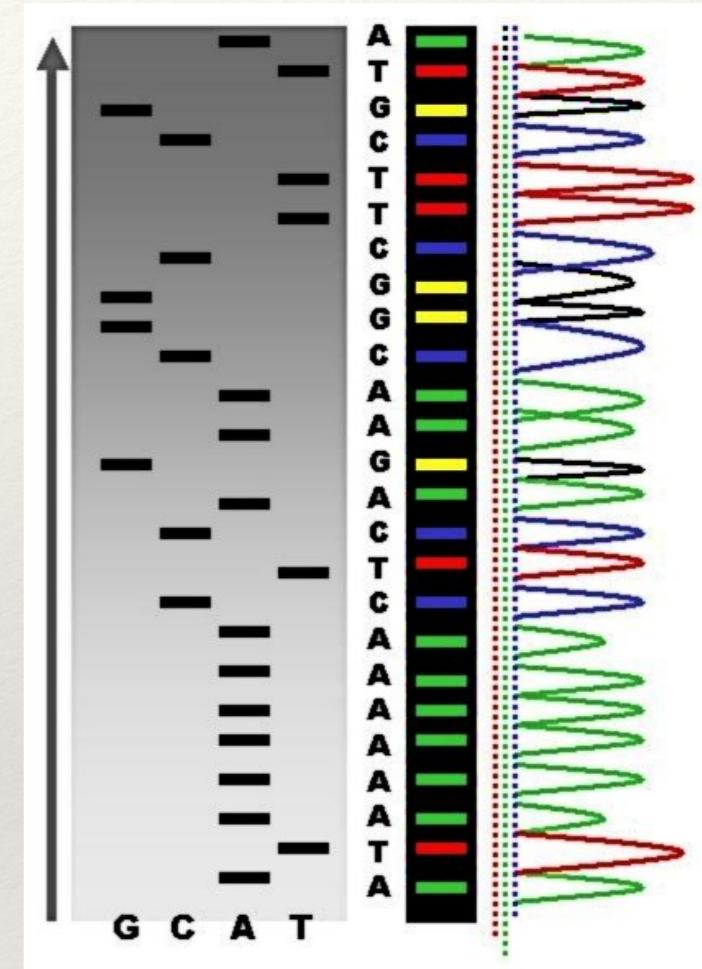
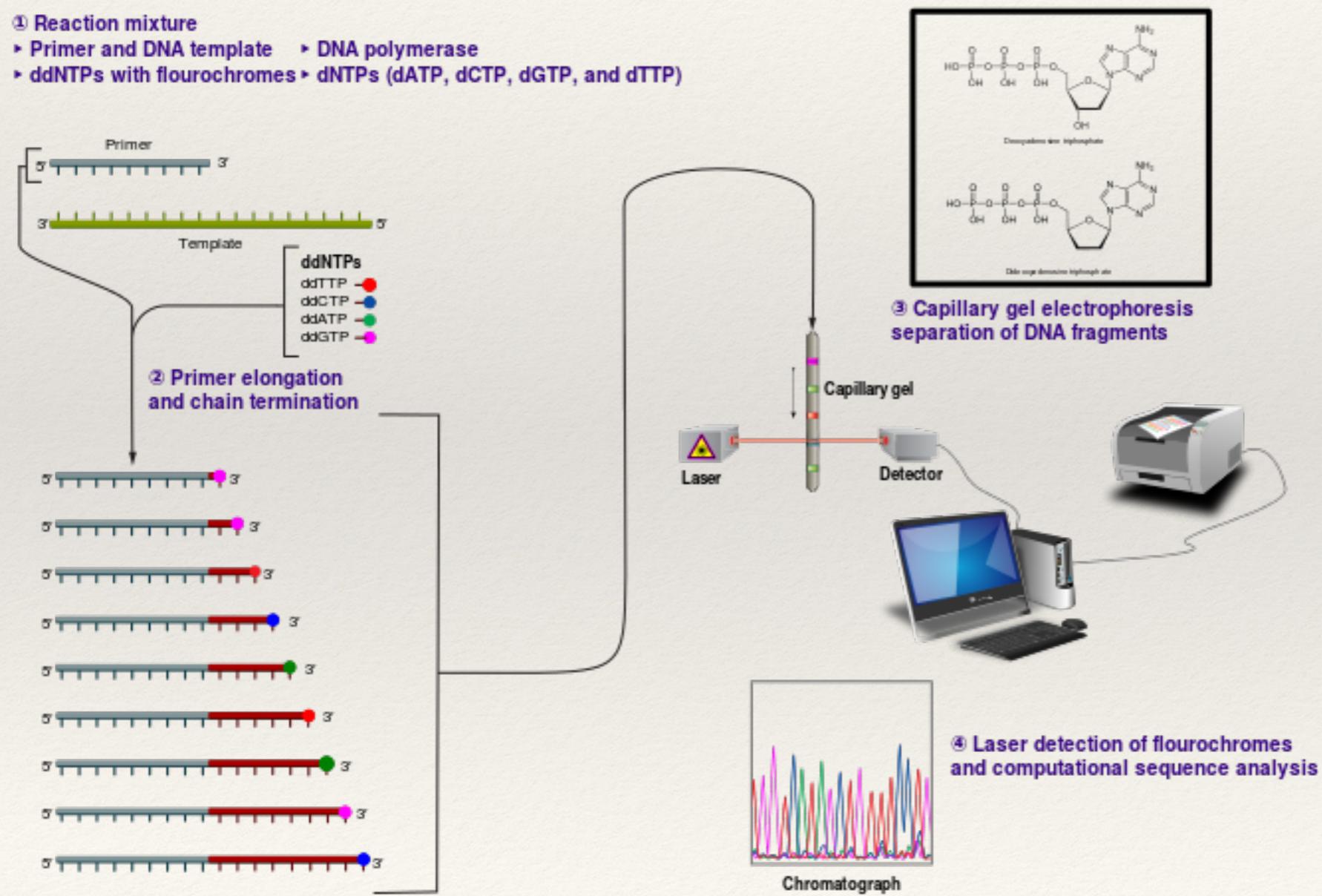
Hugo Tovar
Aarón Vázquez

Flujo de la información genética



Secuenciación Sanger

F Sanger, A R Coulson. 1975. A Rapid Method for Determining Sequences in DNA by Primed Synthesis With DNA Polymerase. *J Mol Biol*, 94 (3), 441-448



> 500 nucleótidos

El formato fasta

- ❖ El formato FASTA es un formato basado en texto que almacena secuencias biológicas ya sea de aminoácidos o nucleótidos.
- ❖ Su nombre viene de un programa de alineación de secuencias de proteínas que se llamaba así en los años 80.
- ❖ El formato FASTA se compone de dos secciones por secuencia:

```
> Encabezado
GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTT
> hemoglobulina
VLSPADKTNV
KAAWGKVGAH
AGEYGAEALR
ERMFLSFPTT
KTYFPHFDLS
HGSAQVKGHG
KKVADALTNA
```

Secuenciación masiva

Sample Prep



Secuenciación masiva

Sample Prep

Tagmentation



Secuenciación masiva

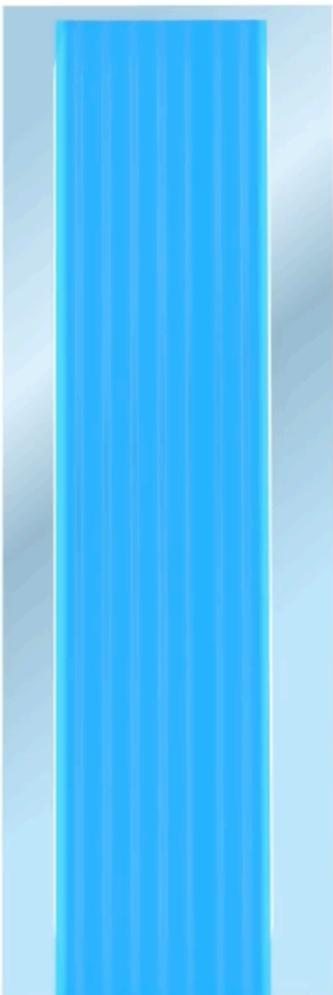
Sample Prep



Secuenciación masiva

Cluster Generation

Clustering is a process where each fragment is isothermally amplified

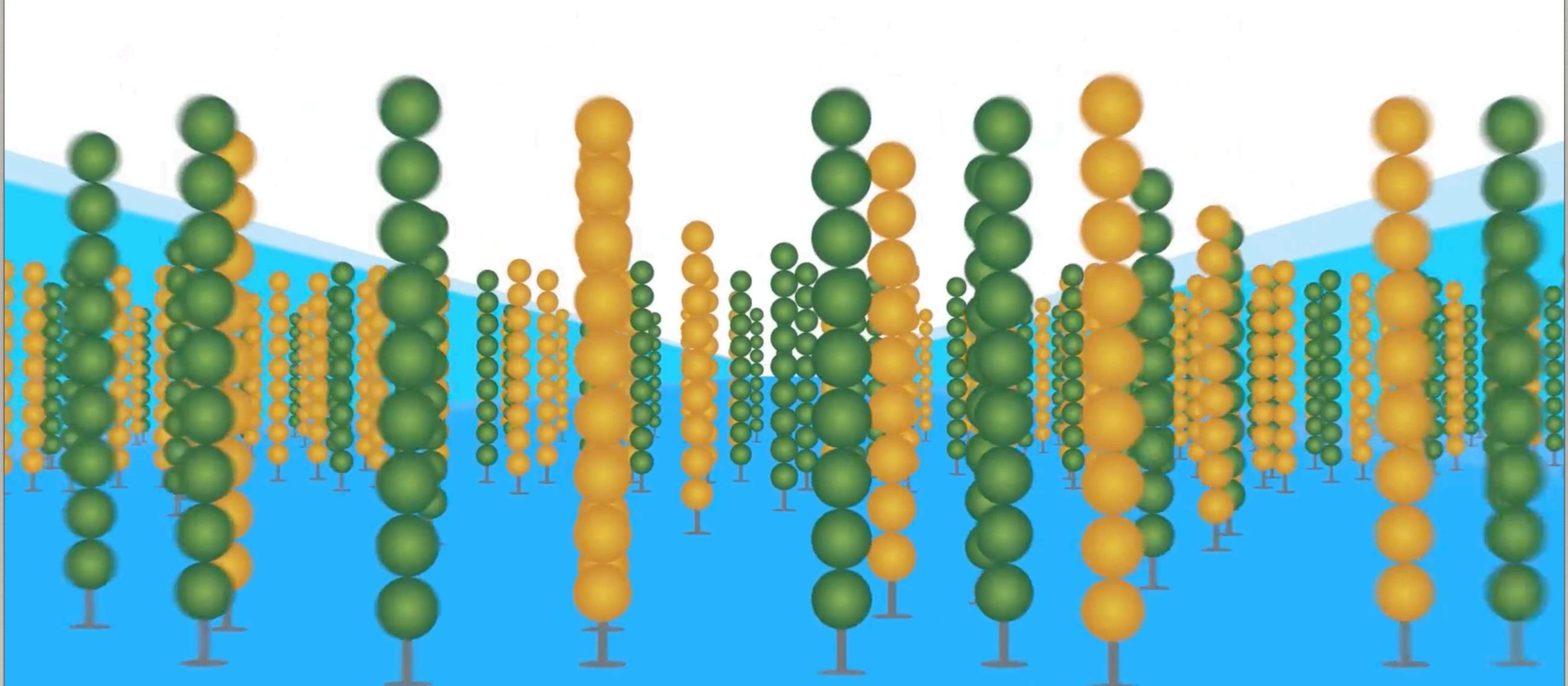


Secuenciación masiva

Cluster Generation

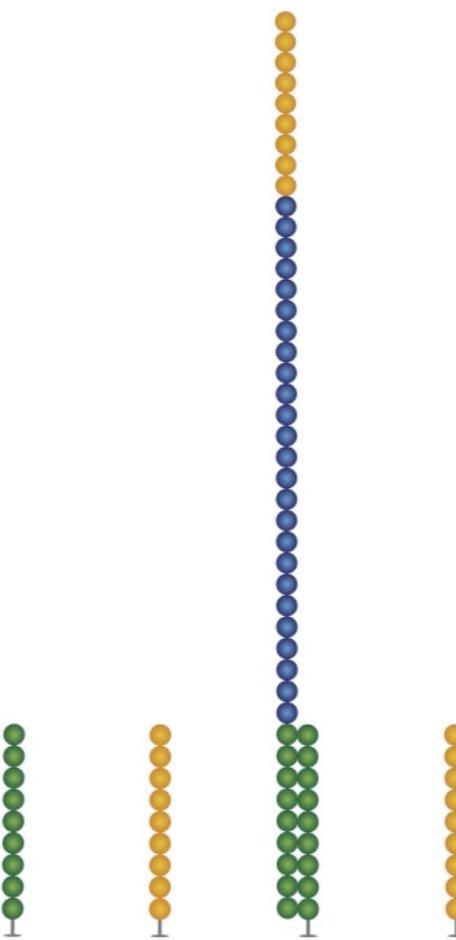
Secuenciación masiva

Cluster Generation



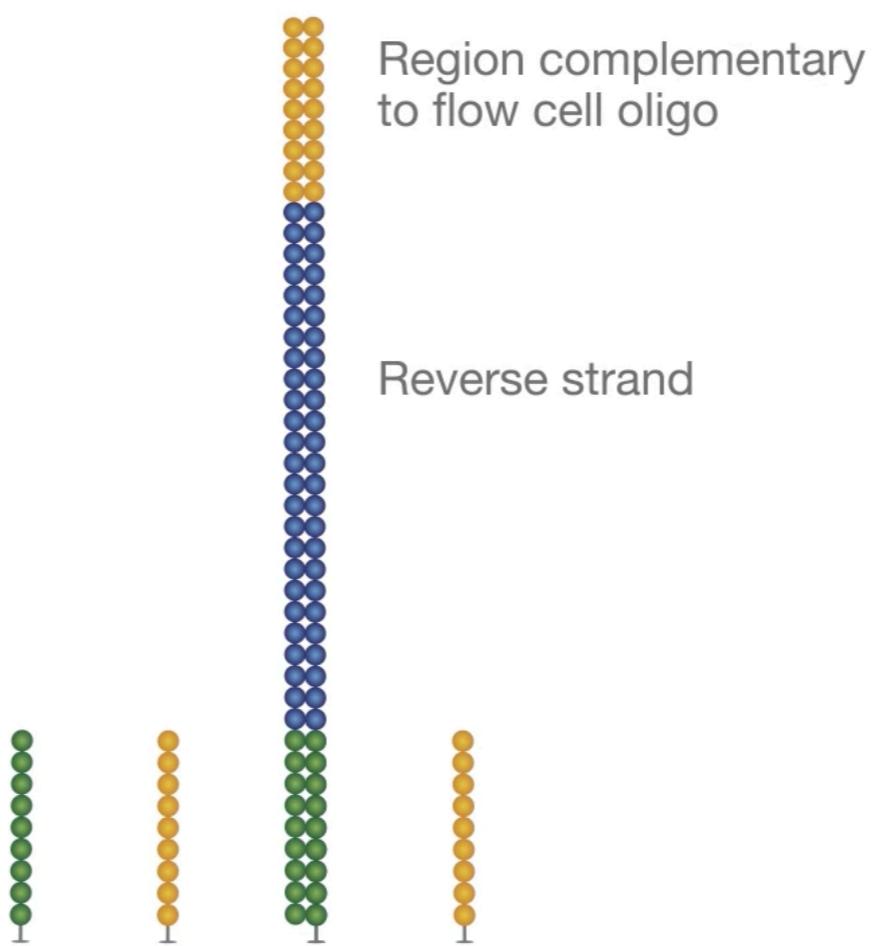
Secuenciación masiva

Cluster Generation



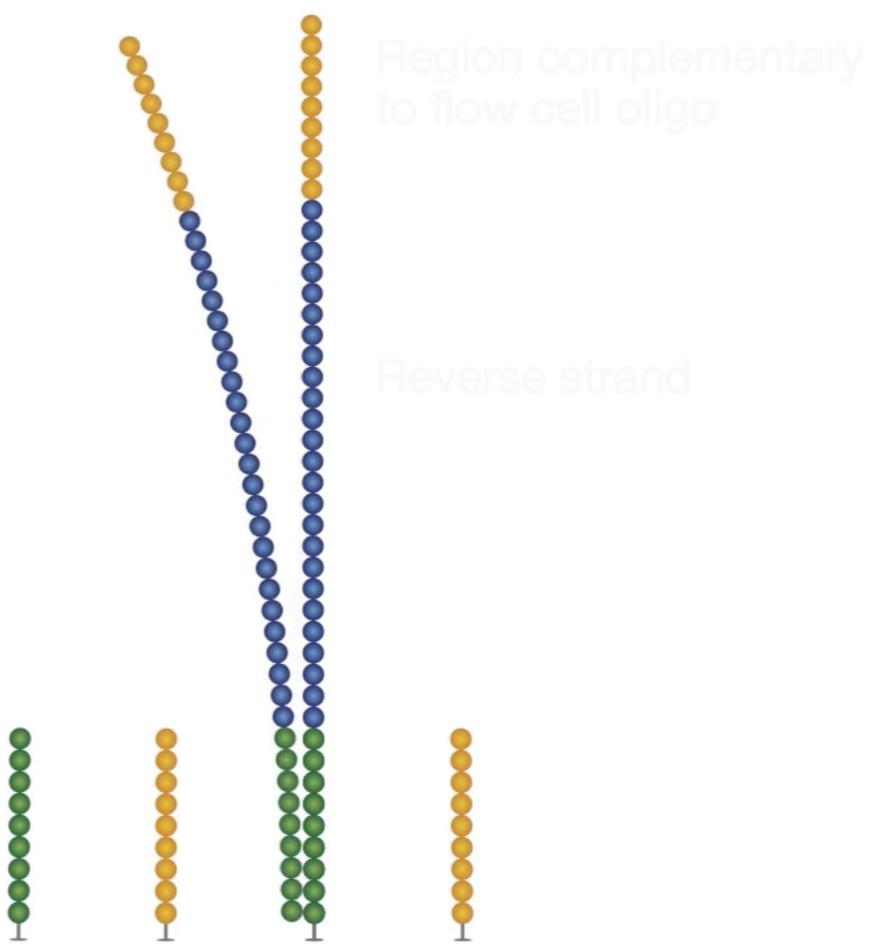
Secuenciación masiva

Cluster Generation



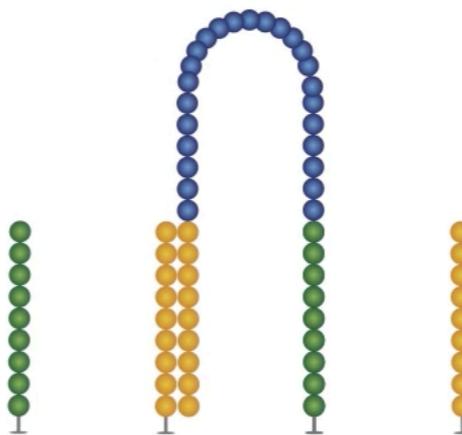
Secuenciación masiva

Cluster Generation



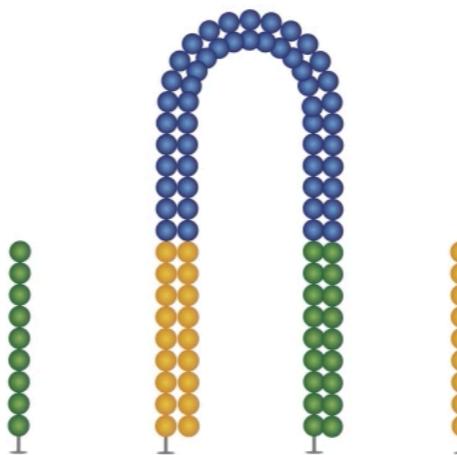
Secuenciación masiva

Cluster Generation



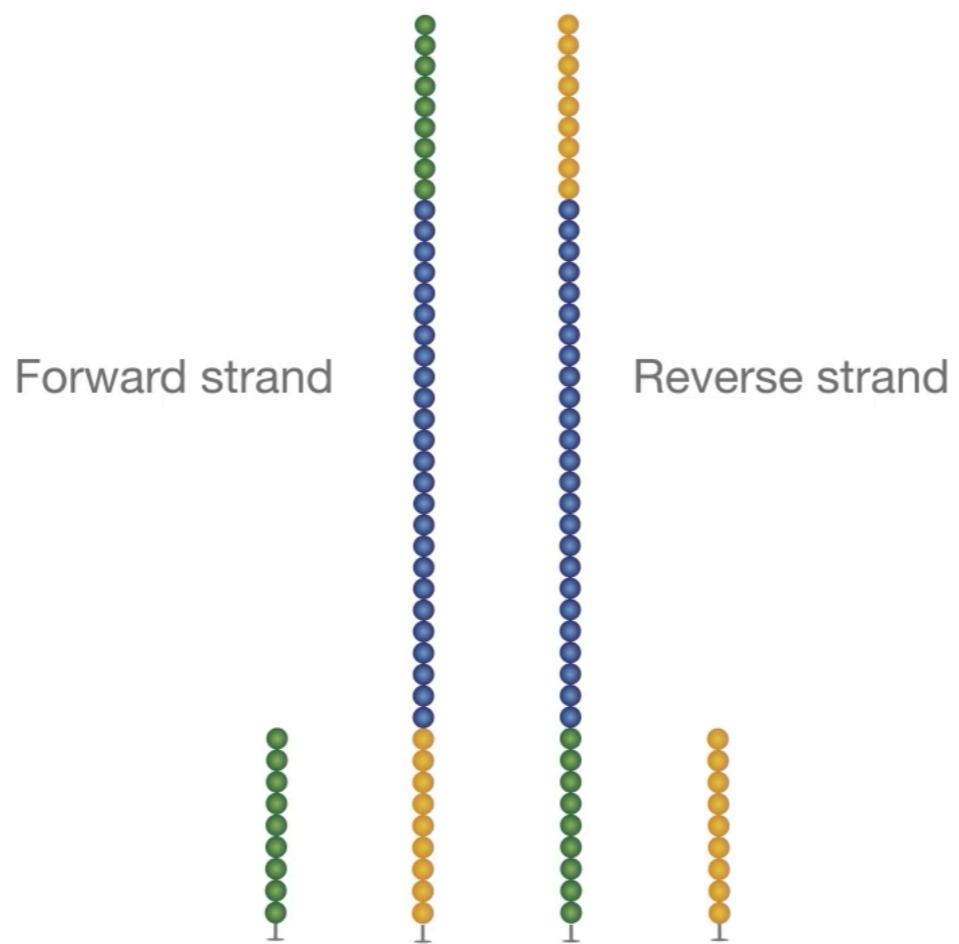
Secuenciación masiva

Cluster Generation



Secuenciación masiva

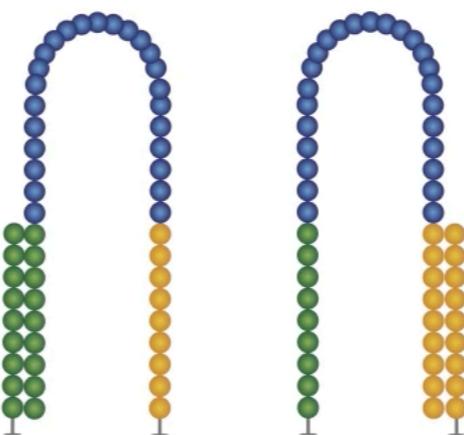
Cluster Generation



Secuenciación masiva

Cluster Generation

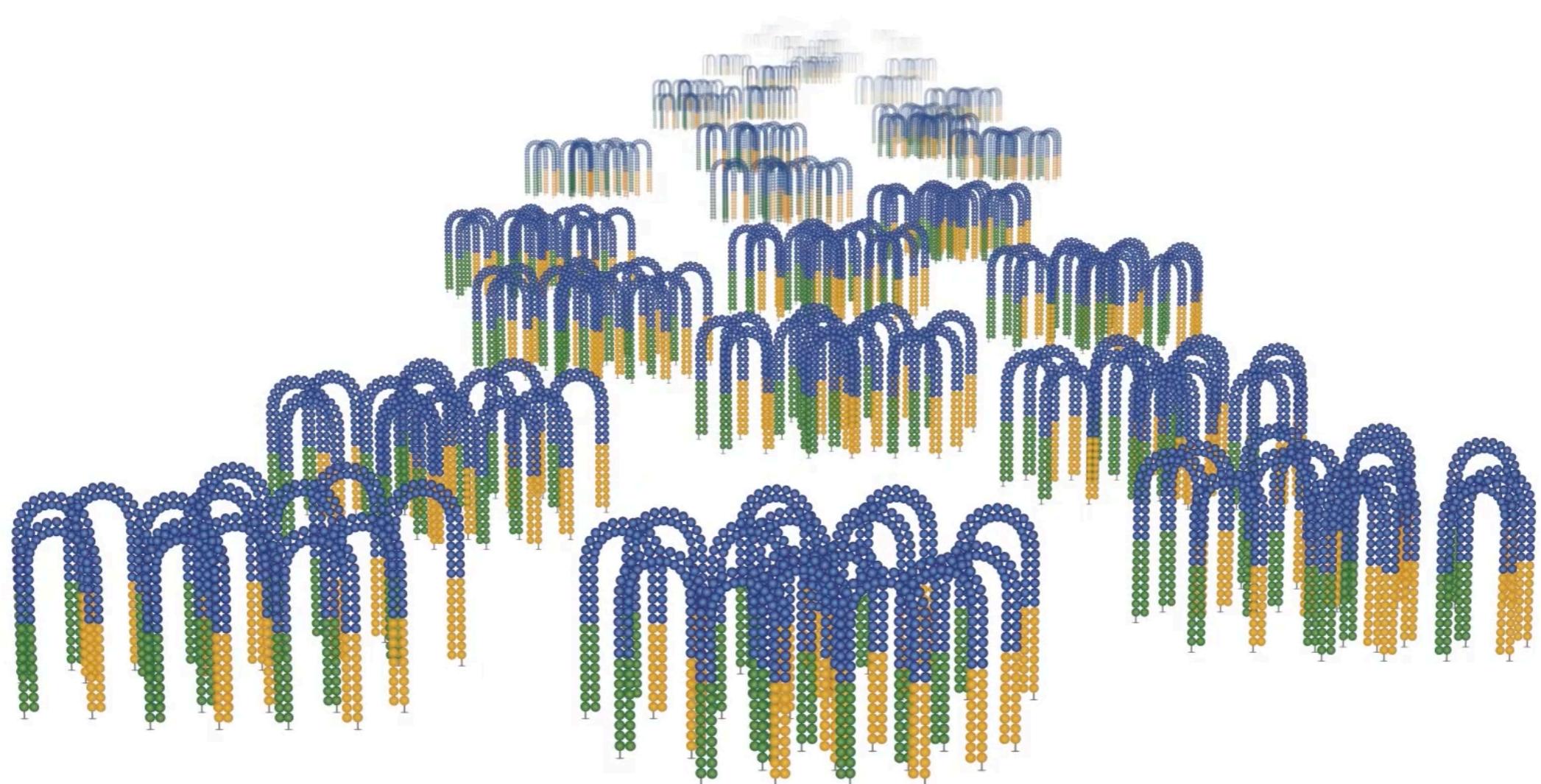
Bridge amplification



Secuenciación masiva

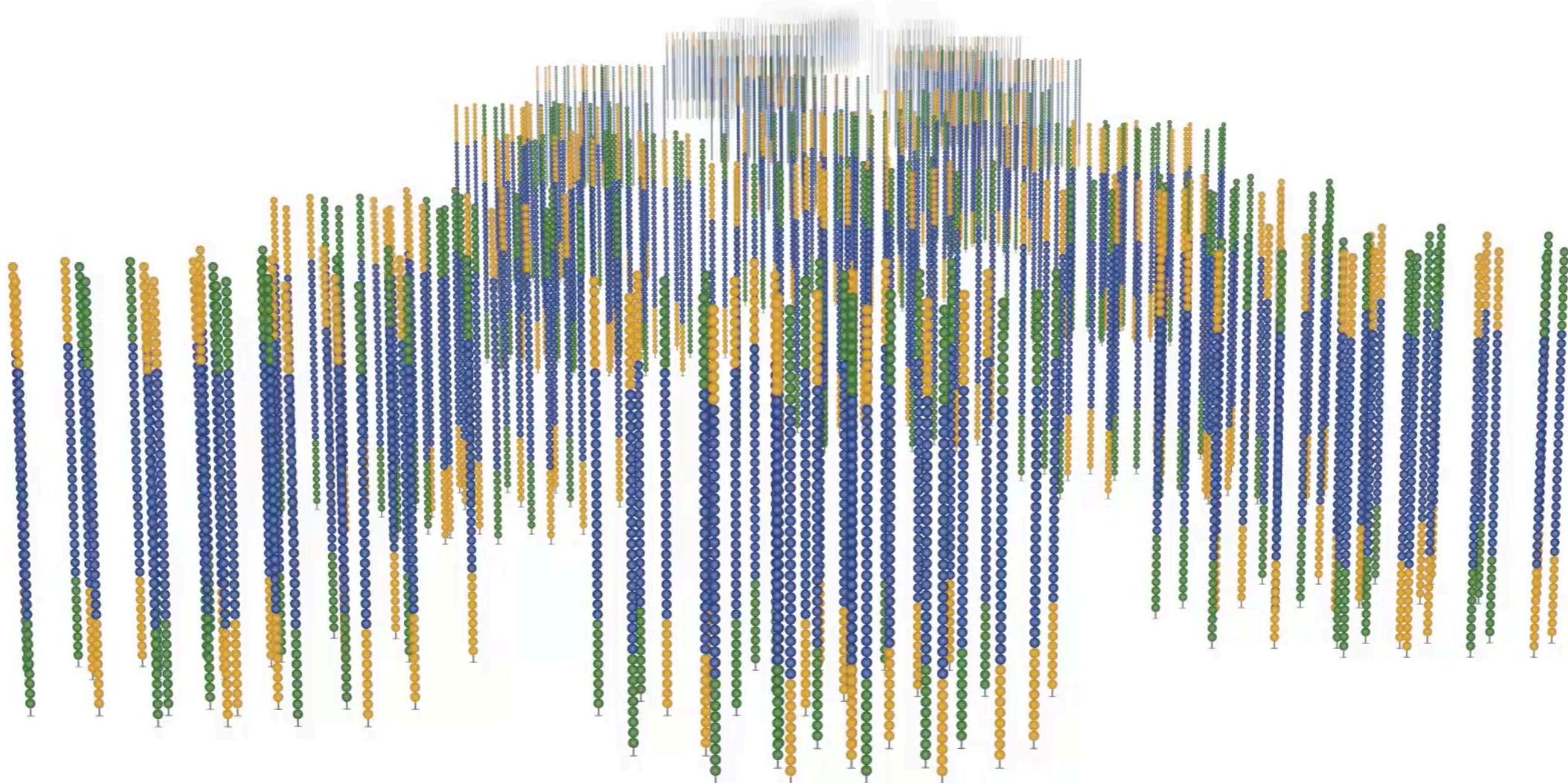
Cluster Generation

Bridge amplification



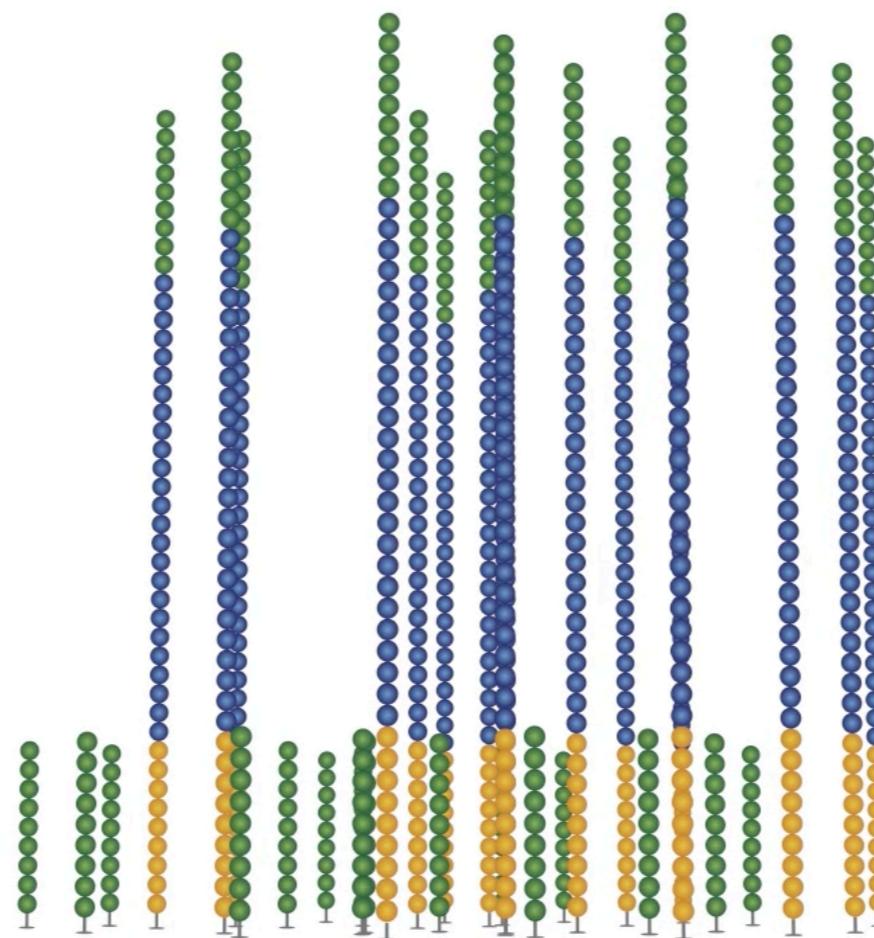
Secuenciación masiva

Cluster Generation



Secuenciación masiva

Cluster Generation



Secuenciación masiva

Sequencing



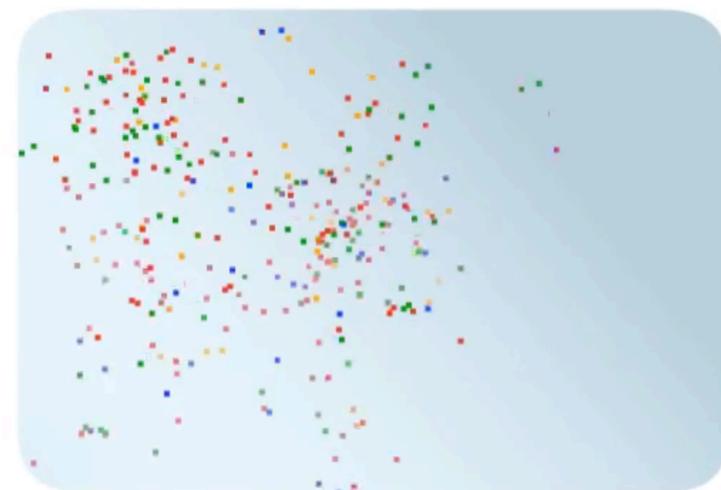
Secuenciación masiva

Sequencing



Secuenciación masiva

Sequencing

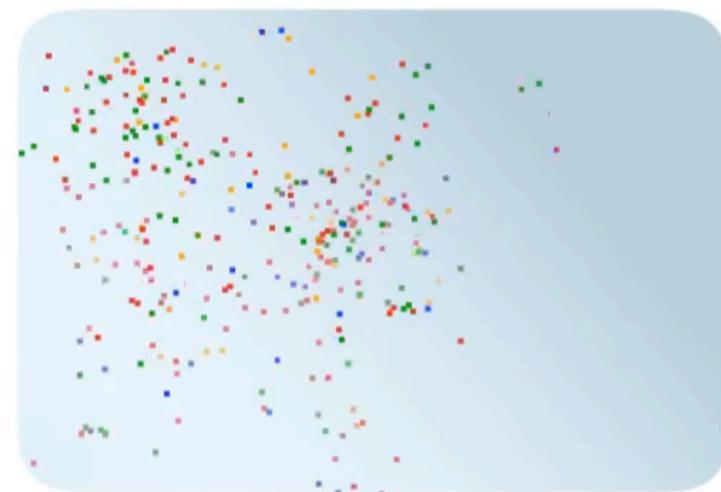


Flowcell

TACTACCTCAGCAGTAGTA
AGTAAGAAACAAAAGCAAT
TAAGGCTTACGCCGTACTA
GCAGTAGTAAGAAACAAAAA
GGCTTACGCCGTACTACCT
TACTACCTCAGCAGTAGTA
AGTAAGAAACAAAAGCAAT
TAAGGCTTACGCCGTACTA
GCAGTAGTAAGAAACAAAAA
GGCTTACGCCGTACTACCT
TACTACCTCAGCAGTAGTA
AGTAAGAAACAAAAGCAAT
TAAGGCTTACGCCGTACTA
GCAGTAGTAAGAAACAAAAA
GGCTTACGCCGTACTACCT

Secuenciación masiva

Sequencing



Flowcell

TACTACCTCAGCAGTAGTA
AGTAAGAAACAAAAGCAAT
TAAGGCTTACGCCGTACTA
GCAGTAGTAAGAAACAAAAA
GGCTTACGCCGTACTACCT
TACTACCTCAGCAGTAGTA
AGTAAGAAACAAAAGCAAT
TAAGGCTTACGCCGTACTA
GCAGTAGTAAGAAACAAAAA
GGCTTACGCCGTACTACCT
TACTACCTCAGCAGTAGTA
AGTAAGAAACAAAAGCAAT
TAAGGCTTACGCCGTACTA
GCAGTAGTAAGAAACAAAAA
GGCTTACGCCGTACTACCT

Puntuación de calidad Phred

- ❖ El puntaje de calidad Phred es una medida de la calidad en la identificación de las bases generadas por un secuenciador automático de DNA
- ❖ El puntaje Phred (Q) se define como la propiedad que está relacionada logarítmicamente con la probabilidad de error (P) en el llamado de la base:

$$Q = -10 \log_{10} P$$

Phred	Precisión del llamado de base	Probabilidad de llamado de base incorrecto
10	90%	1 llamado incorrecto en cada 10 pb
20	99%	1 en 100
30	99.9%	1 en 1,000
40	99.99%	1 en 10,000
50	99.999%	1 en 100,000
60	99.9999%	1 en 1,000,000

El formato fastq

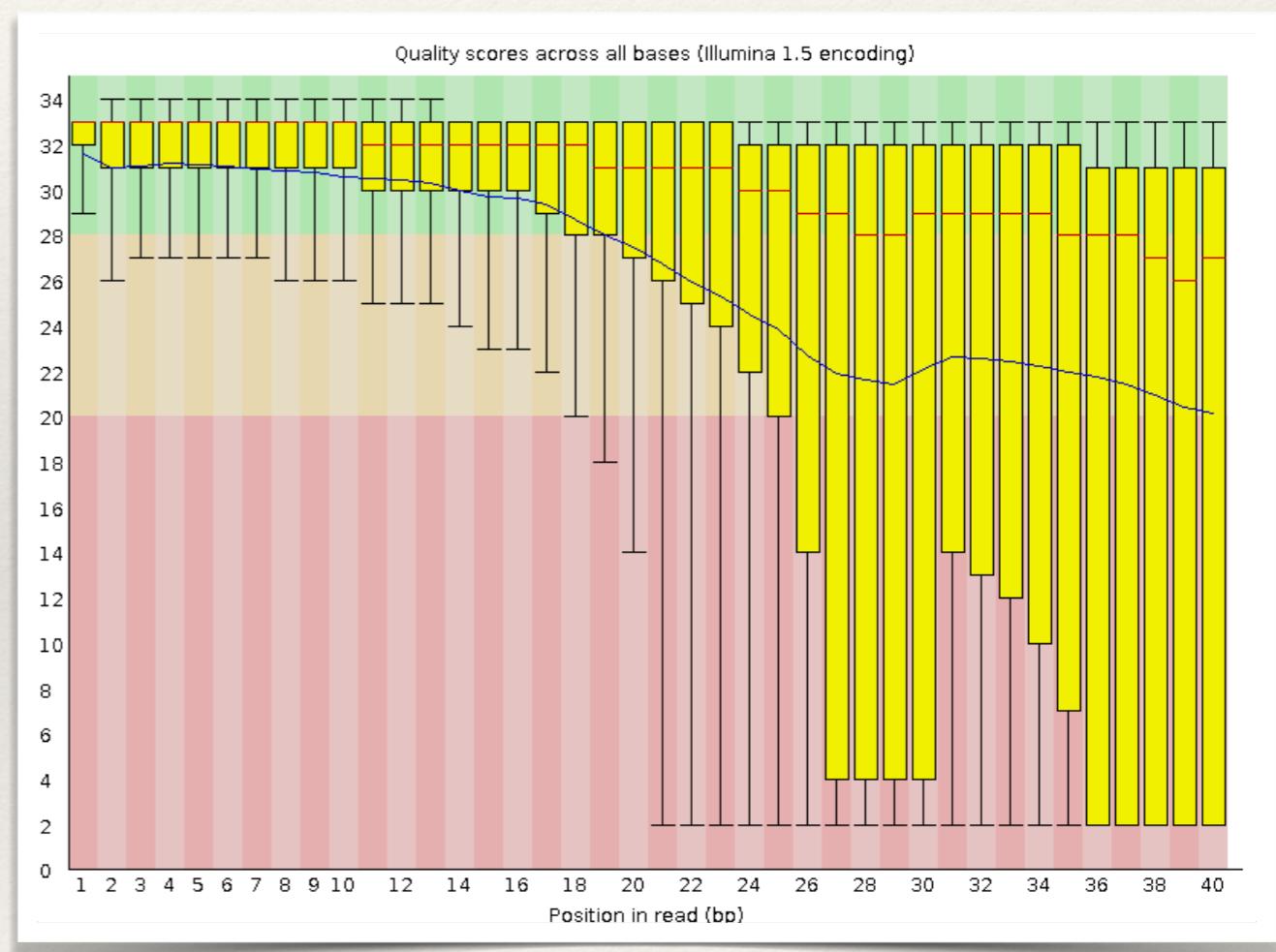
- ❖ El formato FASTQ es un formato basado en texto que almacena tanto la secuencia biológica y su correspondiente puntuación de calidad
- ❖ Tanto la secuencia como la puntuación de calidad están codificados con un solo carácter ASCII
- ❖ El formato FASTQ usa cuatro líneas por secuencia

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTT
+
! ' ' * ( ( ( ***+ ) ) % % % ++ ) ( % % % % ) . 1 *** - + * ' ' ) ) **55CCF>>>>CCCCCCCC65
```

BASE	G	A	T	T	G	G	G	T	T	C	A	A	A	G	C	A	G	T	A	T	C	G	A	T	C	A	A	A	T	A	G	T	A	A	A	T	C	C	A	T	T	T	G	T		
ASCII	!	'	'	*	(((*	*	*	*	+))	%	%	%	+	+)	(%	%	%	%)	.	1	*	*	*	-	+	*	'	'))	*	*	5	5	C	C	F	>
PHRED	33	39	39	42	40	40	40	40	42	42	42	43	41	41	37	37	37	43	43	41	40	37	37	37	37	41	46	49	42	42	45	43	42	39	39	41	41	42	53	53	67	67	70	62		

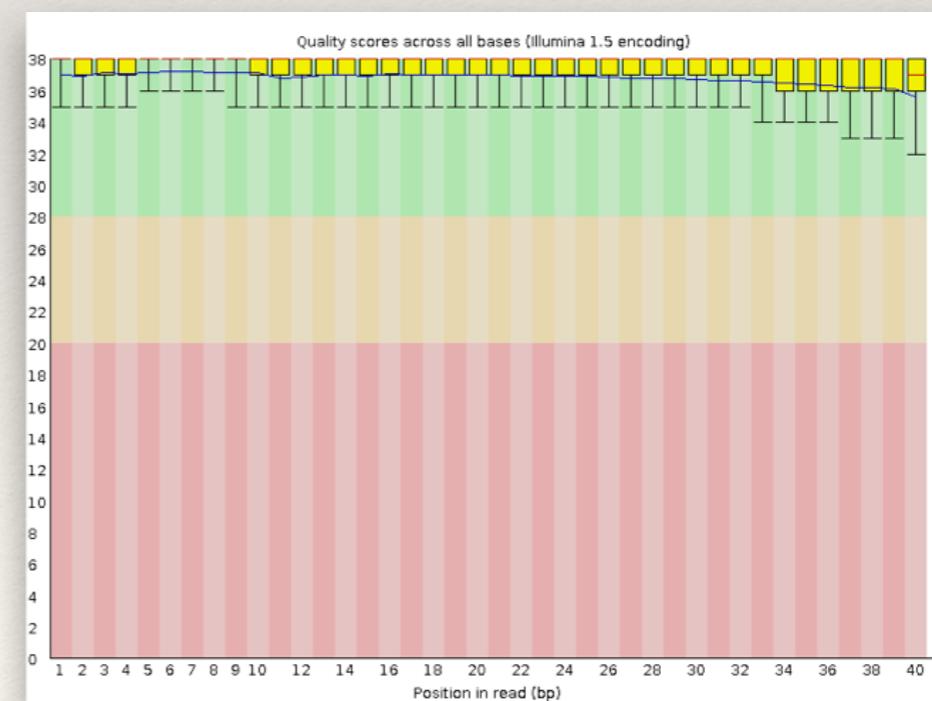
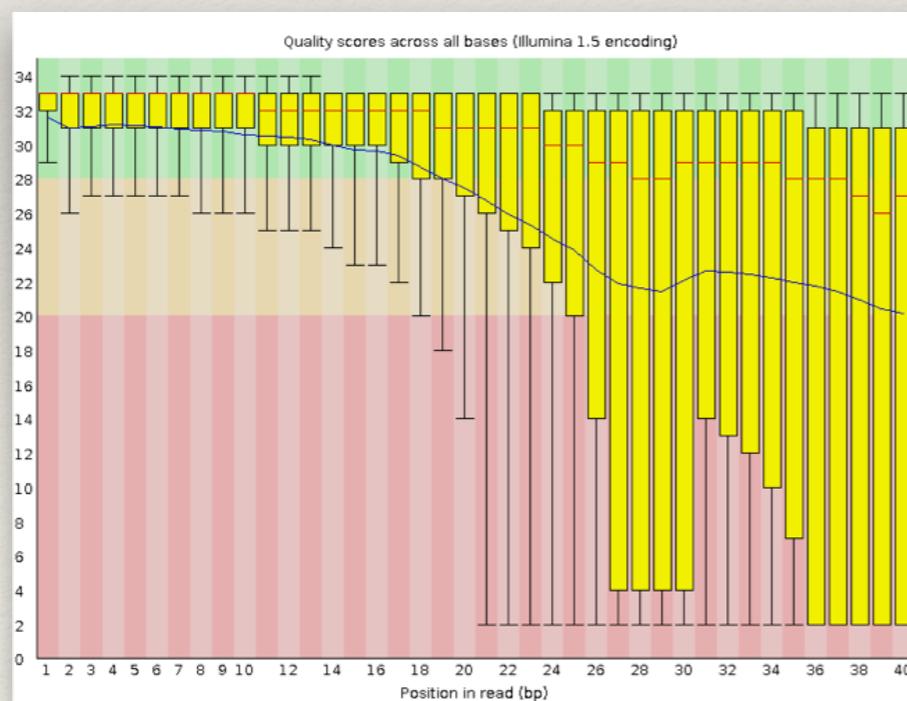
Control de calidad con FastQC

- ❖ FastQC trata de proveer una forma simple de hacer chequeo de control de calidad en las secuencias de los datos crudos provenientes de los secuenciadores de alto rendimiento



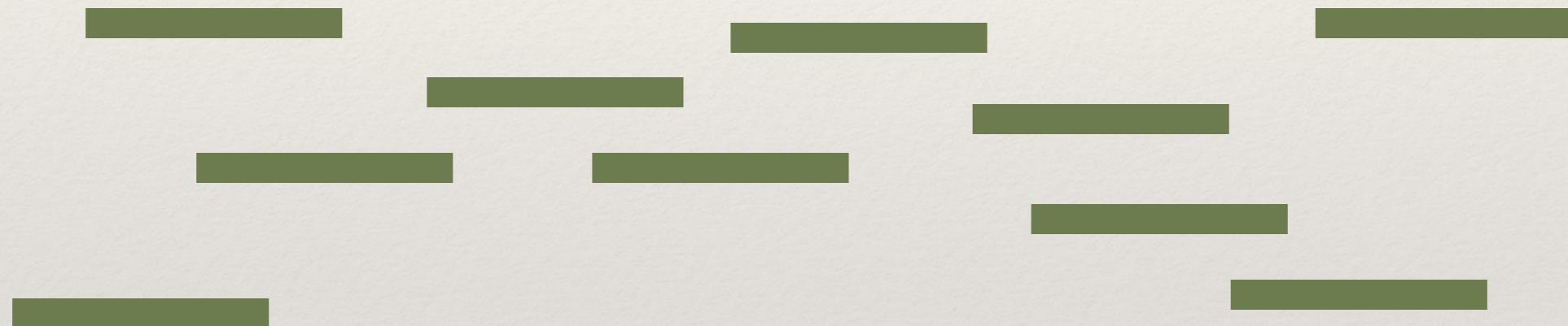
Filtrado y eliminación de adaptadores con Trimmomatic

- ❖ La finalidad es ayudar a filtrar las lecturas de mala calidad y los adaptadores propios de la secuenciación dejando los datos listos para la alineación



Alineamiento/mapeo a la referencia

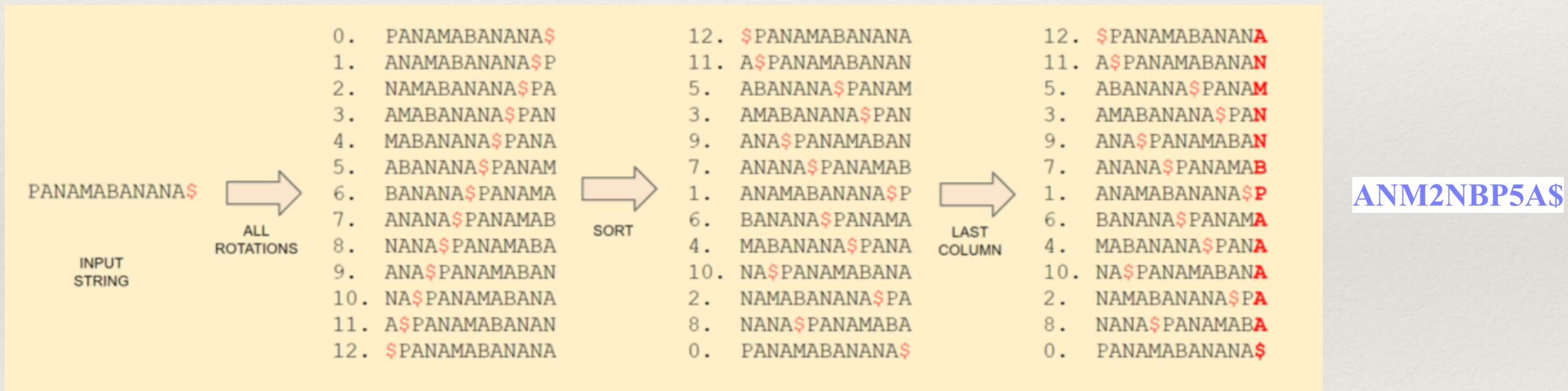
GENOMA DE REFERENCIA



LECTURAS DE SECUENCIACIÓN

Alineamiento/mapeo con BWA

- ❖ Derivado de la transformación de Burrows–Wheeler (BWT del inglés Burrows–Wheeler transform). Un algoritmo usado en técnicas de compresión. Inventado en 1994 por Michael Burrows y David Wheeler



El genoma se indexa utilizando esta transformación lo que hace la búsqueda de las millones de lecturas más rápida y eficiente

El formato SAM

- ❖ El formato Sequence Alignment Map (SAM) es un formato basado en texto para el almacenamiento de secuencias biológicas alineadas a una secuencia de referencia
- ❖ Los archivos BAM es una representación binaria de los archivos SAM

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	[0, 2 ¹⁶ – 1]	bitwise FLAG
3	RNAME	String	* [:rname:^*=] [:rname:] *	Reference sequence NAME ¹¹
4	POS	Int	[0, 2 ³¹ – 1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0, 2 ⁸ – 1]	MAPping Quality
6	CIGAR	String	* ([0-9]+[MIDNSHP=X])+	CIGAR string
7	RNEXT	String	* = [:rname:^*=] [:rname:] *	Reference name of the mate/next read
8	PNEXT	Int	[0, 2 ³¹ – 1]	Position of the mate/next read
9	TLEN	Int	[-2 ³¹ + 1, 2 ³¹ – 1]	observed Template LENgth
10	SEQ	String	* [A-Za-z.=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

NS500560:250:H2NGTBGX9:1:11101:15232:1104	83	chr16	20860356	60	75M	=	20860224	-207	CAAAAAACTCTTCTGCAACAGAAAGGACAATTAACTATTACGGCAATATA	EEEEEEEEEEEE/EEEAEeeeeeeeeeeeeeeeeeeeeeeeeeee
NS500560:250:H2NGTBGX9:1:11101:15232:1104	163	chr16	20860224	60	76M	=	20860356	207	AGTCAGGCCAACTGAAATGTAACCGGTAGAGATCCTGAATTATCTCTGTTGGCTCTGTTAAGAGGCT	AAAAAEeeeeeeeeeeeAEEAEEEEE/6EEAEeeeeeeeAEE
NS500560:250:H2NGTBGX9:1:11101:10668:1104	83	chr2	27382216	60	74M	=	27382108	-182	GCTCATCACATTCTGGGTCAAGAACACAGTCAGAATCTCCAGTCTCACTAAGGCAGCTAGAGGAGTATG	EEEEEEEEEAE6EEeeeeeeeeeeeeeeeeeeeeeeeeeee/EEEEE
NS500560:250:H2NGTBGX9:1:11101:10668:1104	163	chr2	27382108	60	76M	=	27382216	182	CTCCACCTGGTACTCACCTTCCACATGGATGACAATAACCGAAGCTCCCATTTCATCACGCTGGCTGATC	AAAAAEeeeeeeeeeeeAEEAEEeeeeeeeeeeeeeeeeeee

El formato SAM

- ❖ Las diferencias con la referencia son representadas en la columna CIGAR (Concise Idiosyncratic Gapped Alignment Report)

Por ejemplo:

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																		

Puede alienar así:

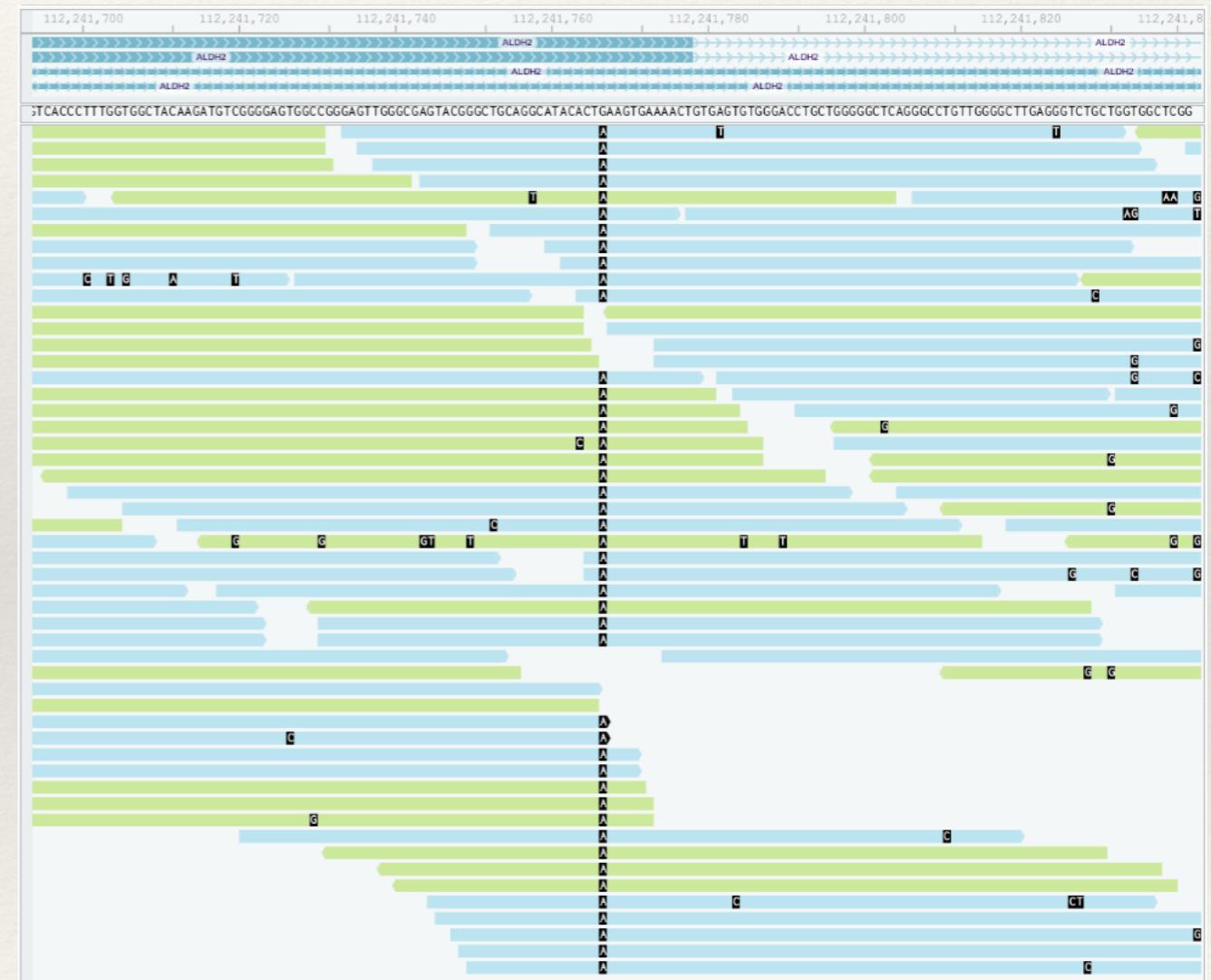
RefPos:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Reference:	C	C	A	T	A	C	T	G	A	A	C	T	G	A	C	T	A	A	C
Read:					A	C	T	A	G	A	A	T	G	G	C	T			

Con esta alienación el CIGAR quedaría así:

POS: 5
CIGAR: 3M1I3M1D5M

Variantes genéticas

- ❖ El genoma humano tiene 3,200,000,000 pares de bases.
- ❖ Los individuos diferimos en promedio entre el 0.5 y 1 % de nuestro genoma.
- ❖ El 1 % del genoma humano son 32,000,000 de pares de bases.



Variantes genéticas

Gen ABCC11: variaciones en este gen influyen en la cera de los oídos y en el olor.

Mal olor	5 horas	24 horas
TT	2.59	2.6
CT	3.26	3.4
CC	3.21	3.5
Usa desodorante	Si	No
TT	50%	50%
CT	86%	14%
CC	97%	3%

