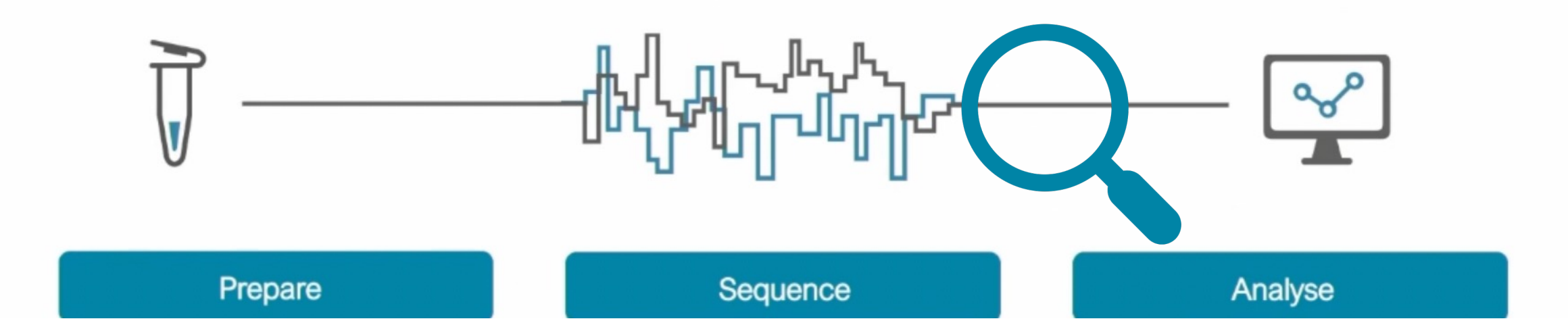
Introducción al llamado de bases de Nanopore

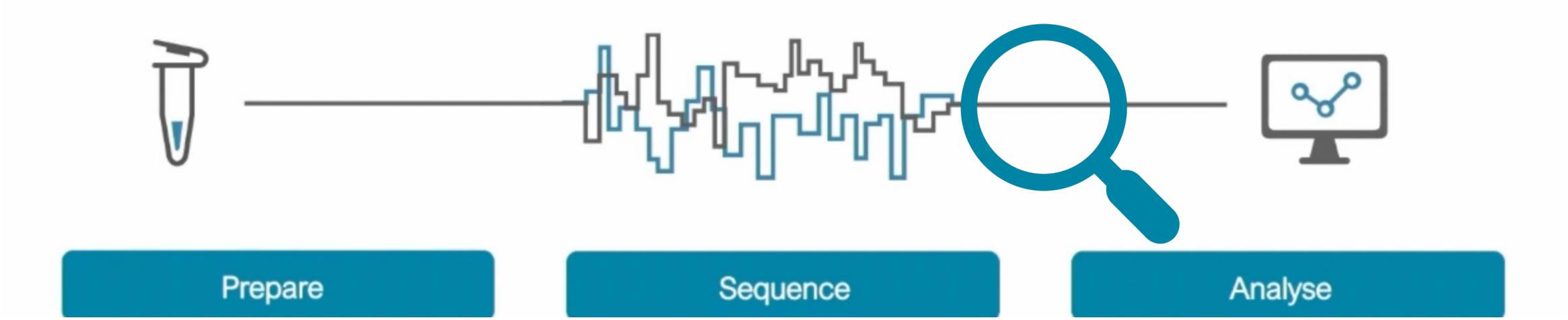
Daniel Bautista – Universidad del Magdalena – Santa Marta, Colombia
2021

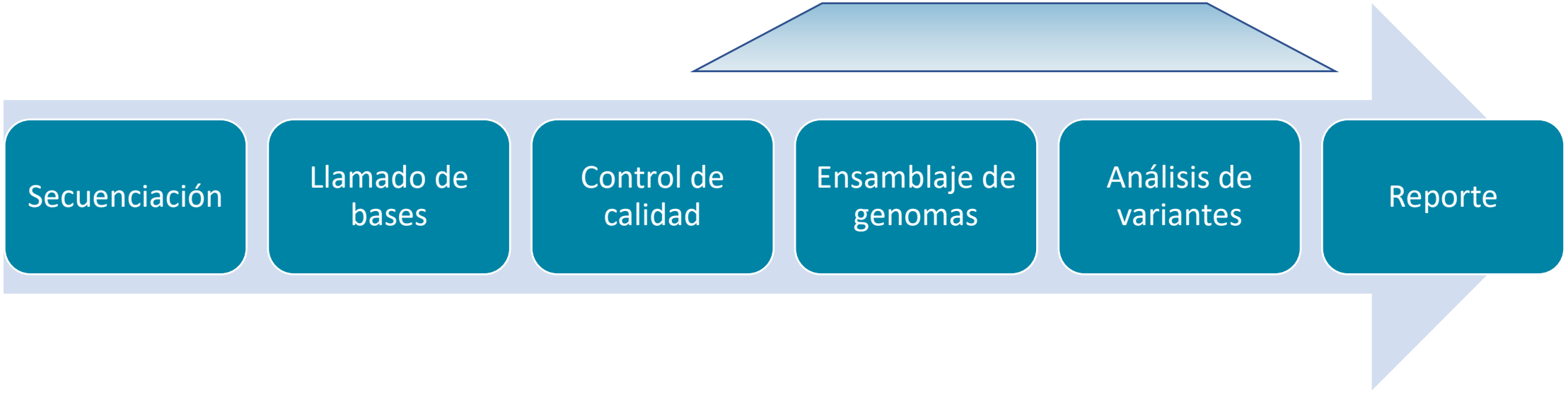
Pasos Procesamiento Datos Genómicos

Pasos Procesamiento Datos Genómicos



Pasos Procesamiento Datos Genómicos





¿Por qué usar tecnología Oxford Nanopore?

- No requiere una gran infraestructura, acomodación o equipos
- Tecnología moderna (escalable y portable)
- Diferentes aplicaciones (Tiempo Real, librerías)

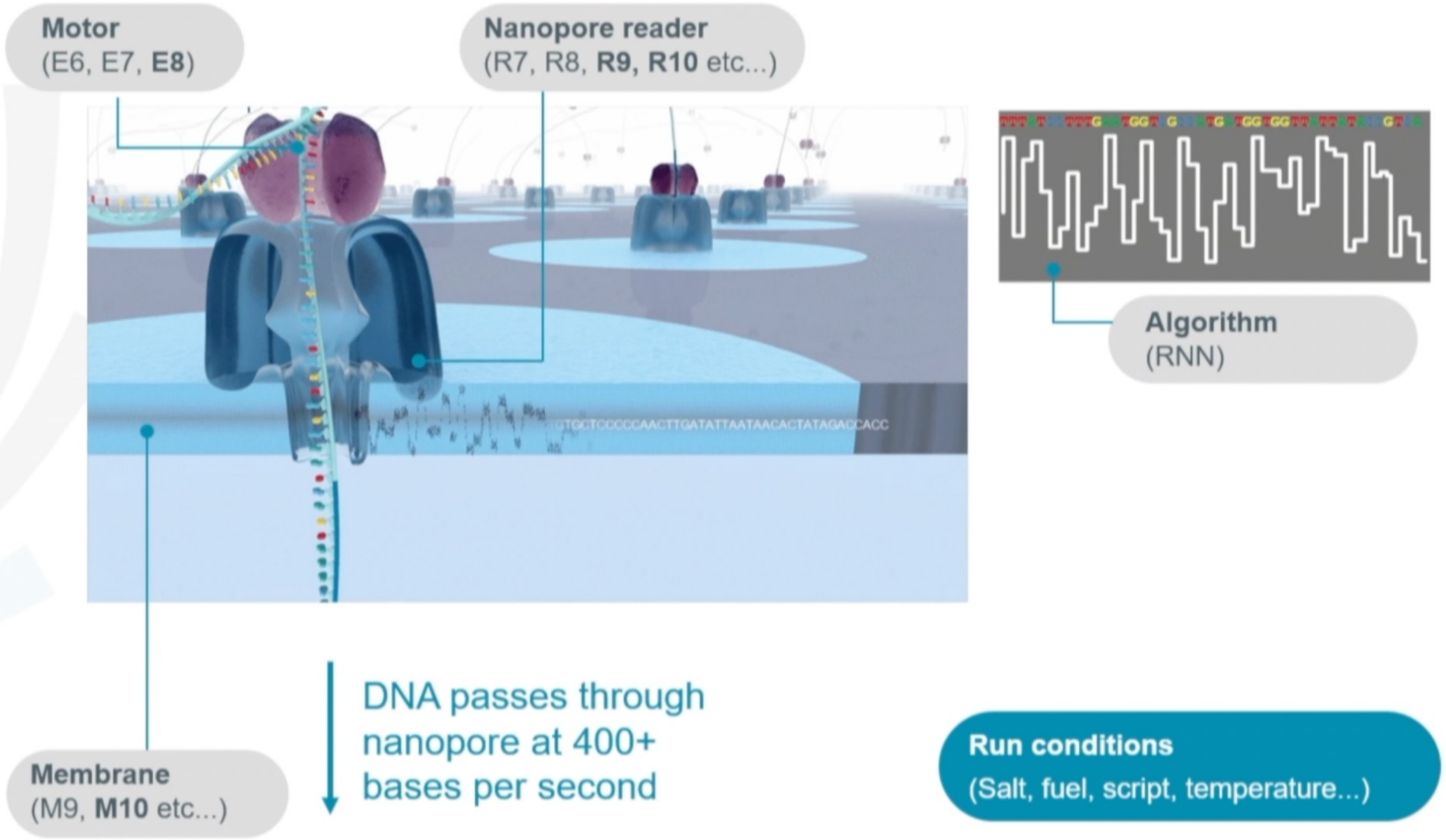
Desventajas

- No vamos a obtener # secuencias que Illumina -> menor consenso.
- % de error relativamente alto.
- Homopolimeros: cadena de secuencia repetidas de un mismo nucleótido.

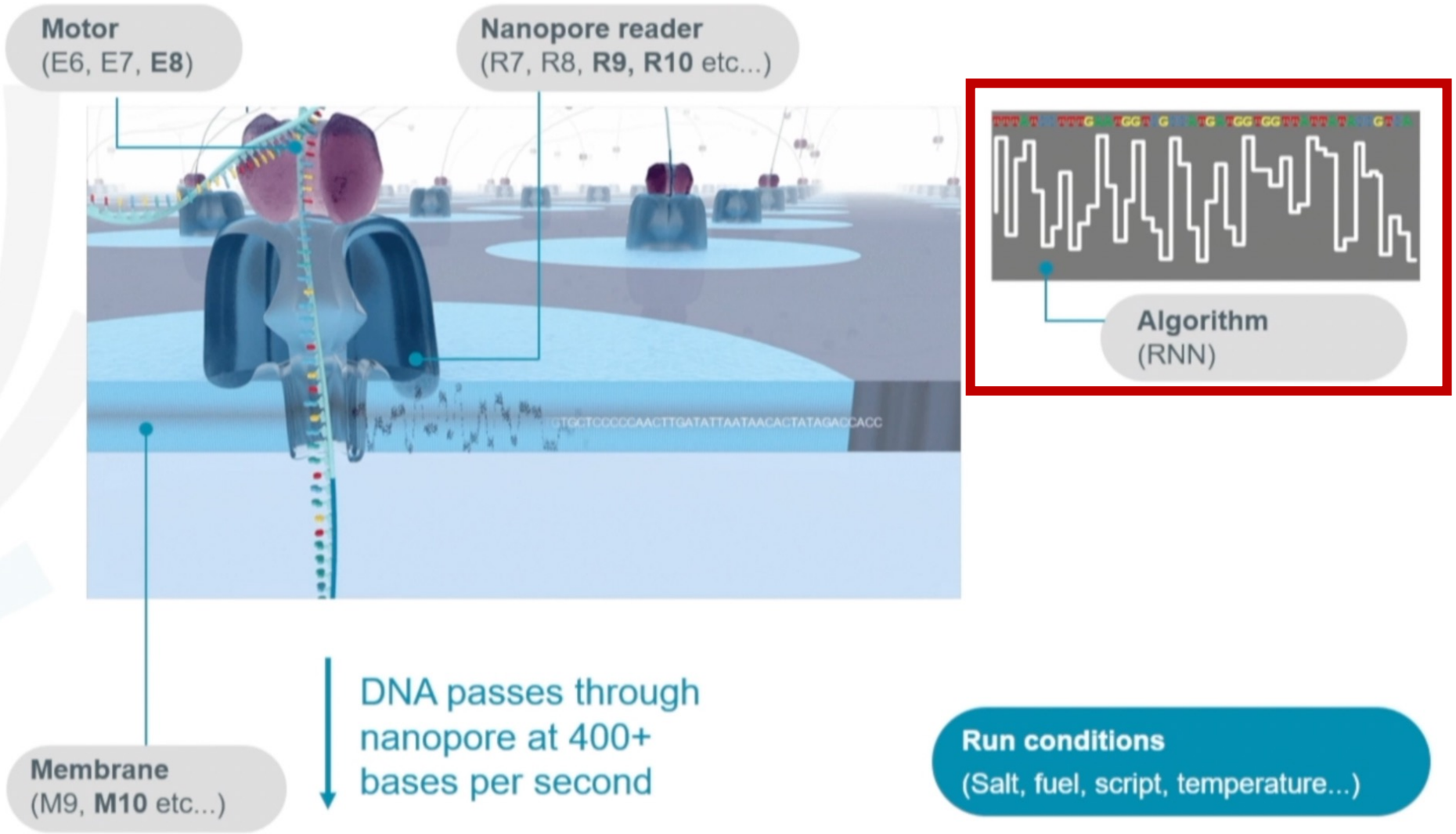
Chemistry	Raw read accuracy (modal)	Analytical tools	Sample
R9.4.1	98.3%	Production software MinKNOW 4.3 ("Super-accuracy" basecalling model), Guppy 5	Mixed genomes

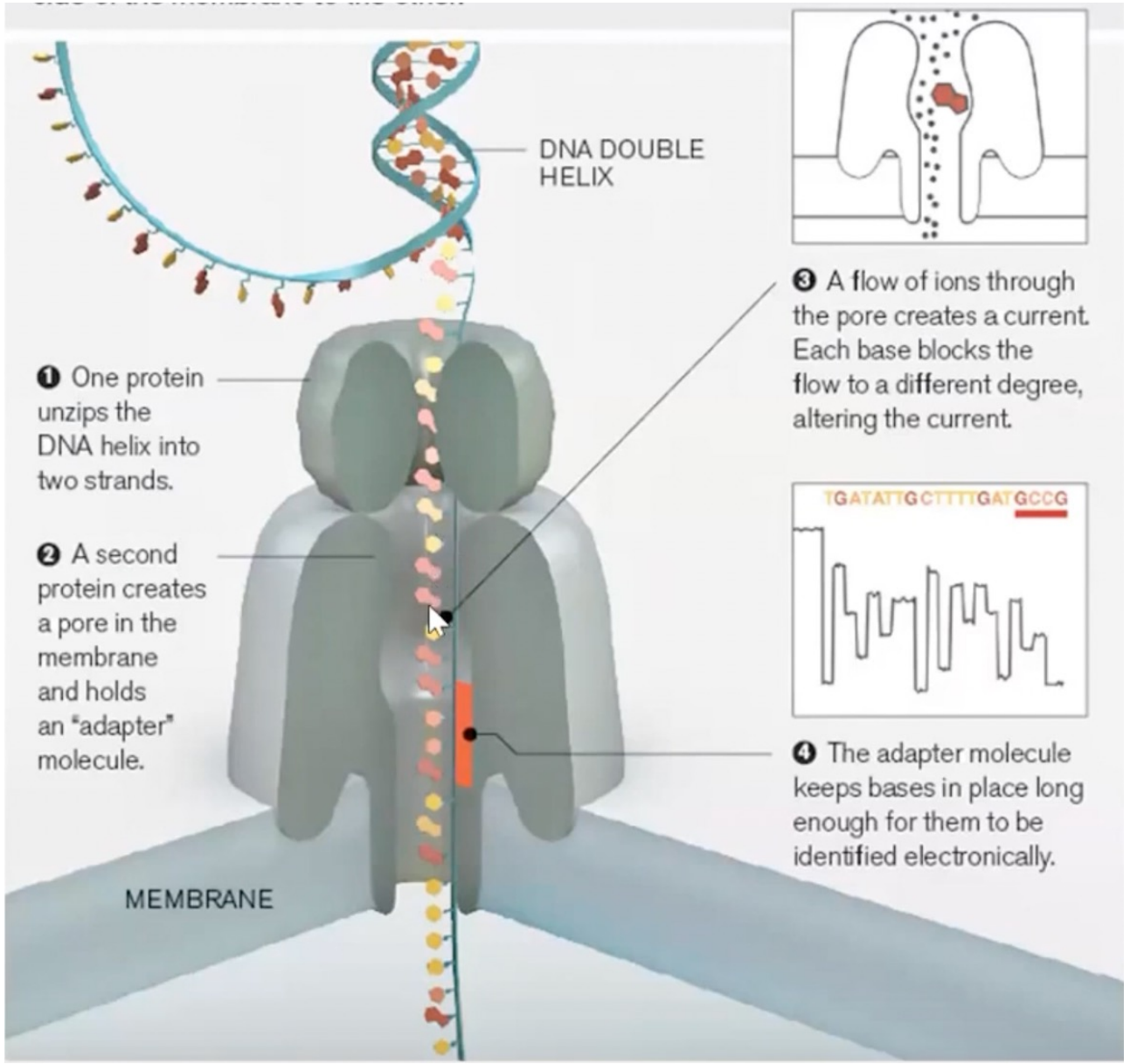
Chemistry	Consensus accuracy	Analytical tools	Sample
R9.4.1	Q50 at ~100X	Guppy basecall, Flye assembly, Medaka polish	Zymo mock community (bacterial)

Oxford Nanopore



Oxford Nanopore

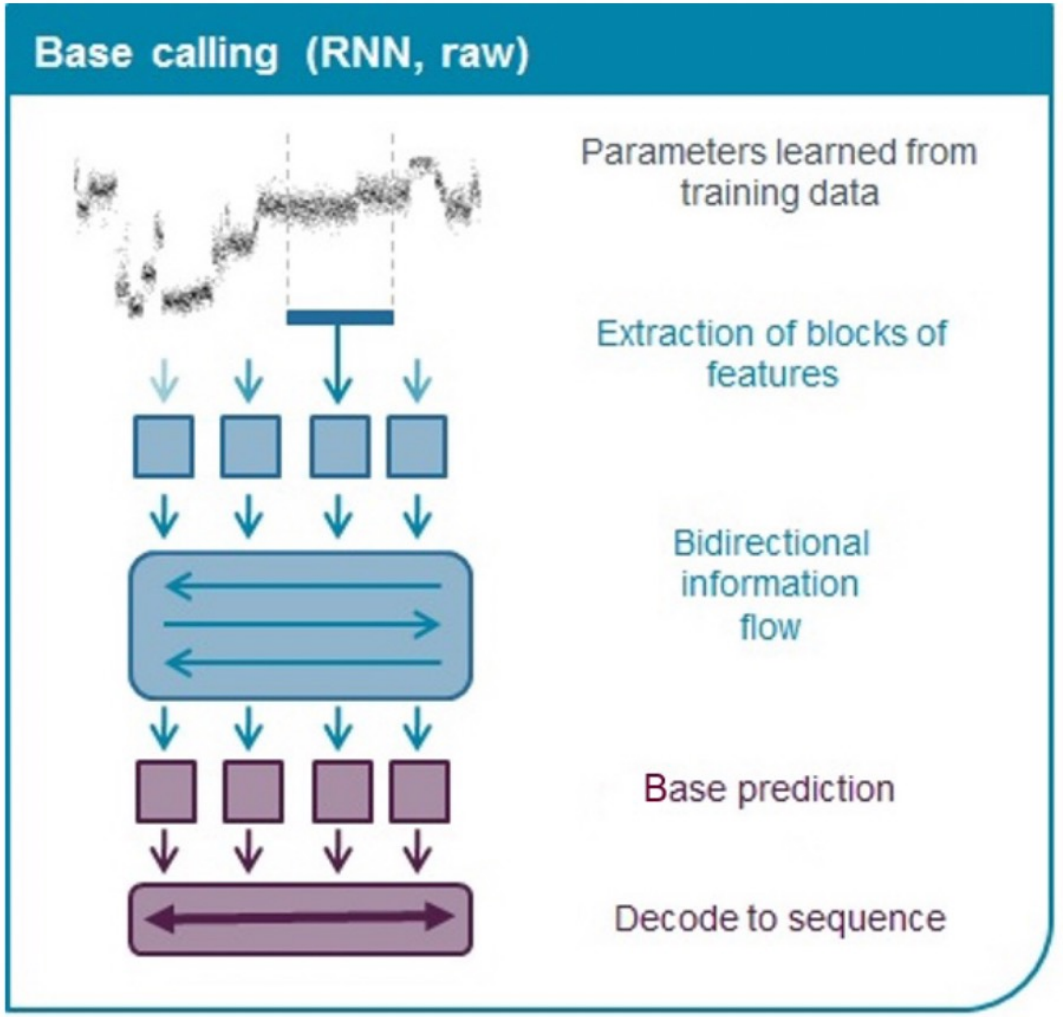




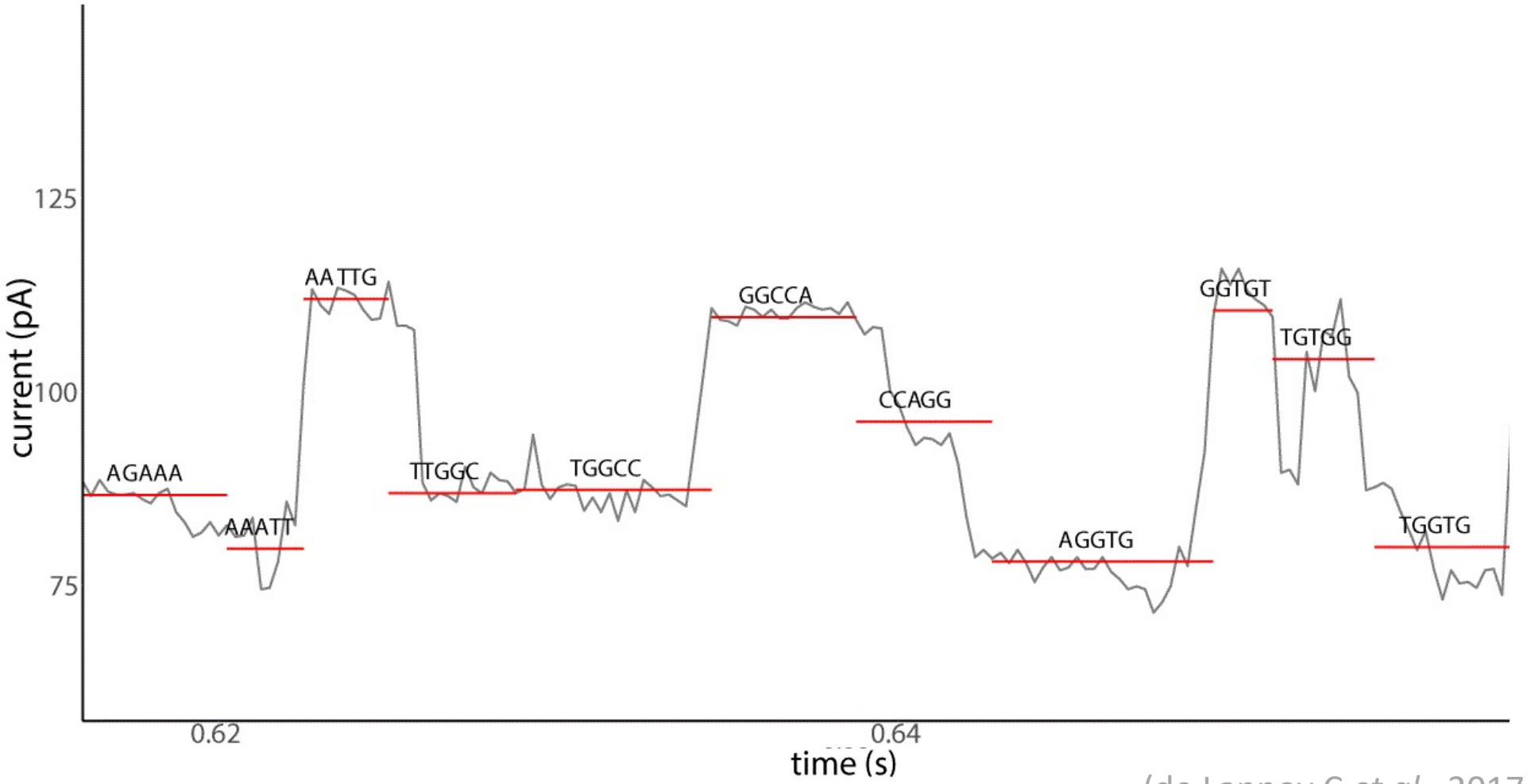
(de Lannoy C *et al.*, 2017)

¿Cómo determinar la secuencia nucleotídica?

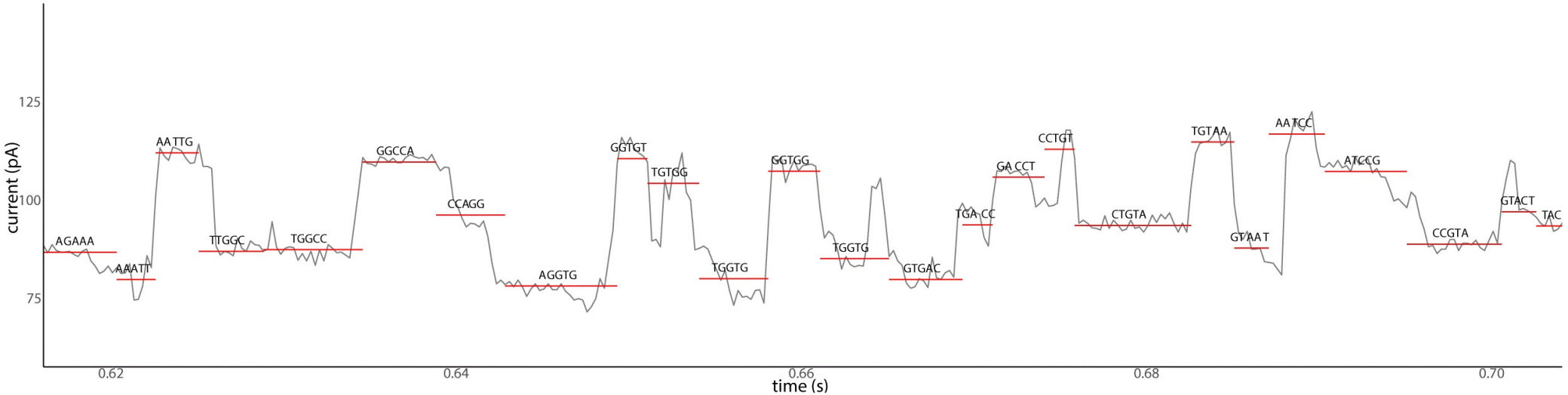
- Uso de algoritmos que puedan aprender con datos anteriores (Inteligencia Artificial).
- Es posible entrenar el algoritmo con datos propios.
- Solape de bases.



Llamado de bases



(de Lannoy C *et al.*, 2017)




```
@M01380:62:000000000-B547W:1:1102:20819:1013 1:N:0:MWI006 NGCCTCTT|1|NCTGCATA|1
NGTAGAGTTTGATTCTGGCTCAGGATGAACGCTGACAGAATGCTTAACACATGCAAGTCTACTTGATCCTTCGGGTGATGGTGGCGGACGGGTGAGTAACGCGTAAAGAACTTGCCCTGCAGTCTGGGACAACATTTGGAAACGAA
TGCTAATACCGGATATTATGCGAACTTCGCATGTAGCTCGTATGAAAGCTATATGCGCTGCAGGATAGCTTTGCGTCCTATTAGCTAGTTGGTGAGGTAACGGATCACCAAGGCCATGATCGGTAGCCGGGCTGAGTGTGTGAACG
GCCGCAAGG
+
#8BCCGGGGGGGGGGGGGFGGDFGFFGGFCFGGGDGFF8CEAFGFGGGGEDFGGGGGGFFGGGGGGGGGCFAF7C<+DDFGGGD8@EFFFGGGGFGGGGCCFGGDCGDD?,B?ECG?A<FGDFGGGGGGGFF8FGGFGGG9EFF7B
FFFFFFDGCFG7CEFAF@FG,3FGGGG,+FCECGG=CC9:CCFFGGF9>CFFCGGFGGGC*6<@@,9?FC@FG@EC88E?9F?F6>76+>AFC5C5EFAC6C**//02A=EGFEE437>:+1***122)/)/7*)9*:**)01
*)87)4),)-1:

@M01380:62:000000000-B547W:1:1102:16288:1015 1:N:0:MWI006 NGCCTCTT|1|NCTGCATA|1
NTACGTAGGGTTCGATCCTGGCTCAGGATGAACGCTAGCTACAGGCTTAACACATGCAAGTCGAGGGGCAGCATCATCAAAGATTGCTTTGATGGATGGCGACCGGCGCACGGGTGAGTAACACGTATCCAACCTGCCGACAACAC
TGGGATAGCCTTTCGAAAGAAAGATTAATACCGGATGGCATAATTATTACGCATGGGATAATTATTAAAGAATTTCGGTGGCCGATGGGGGTGCGTTACATTAGGCAGATGGCGGGGGAAAGGCCTACCAAAACAACGACGGATAG
GGTGTGTGG
+
#8@ACGG@BEFF87EFFFFF88CFGGFG,EECCF,CF:,,F<FECCFFDFGFGGFDCCEFFFGEGGG:@FCCDF8FFFGFGG8,9@,,?<C<CFGGEFF8FCCEEC7=7FFCG+8+AE<CBEGFEFF:BFFGFC8,,BF7@7CE8B
=FAB8,5,,7@FAE**><@,FCCFA@FFCC;,>11*5*>FGFG9,@C9,6=CEGG88+29+3?C+23+49<=9+?BFD8***3==/:=*;**/*1:C**+2+0:+3<C**+76==7*))*2979C**2)2)9)*)*.1>)87:.,9
*.,*4).4(

@M01380:62:000000000-B547W:1:1102:15376:1016 1:N:0:MWI005 NGCCTCTT|1|NTAAGGAG|1
```

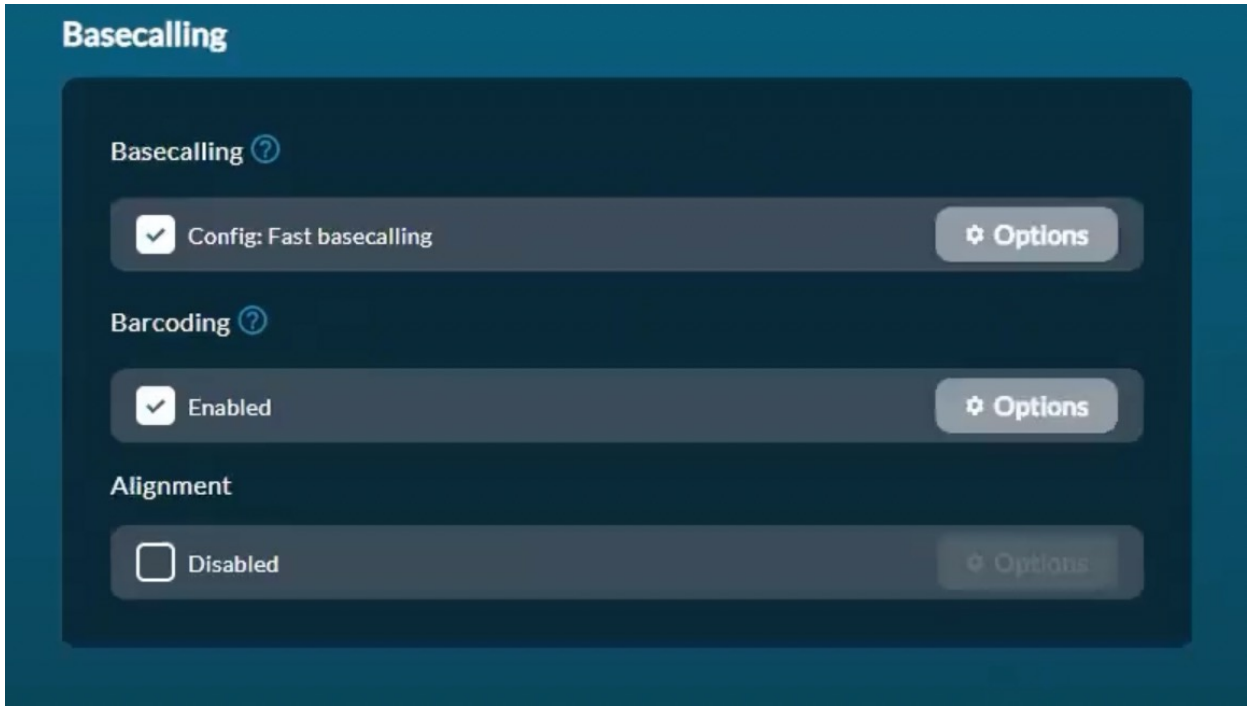
Ajustes del algoritmo

- Muchas maneras de optimizar el problema
- Mayor resolución en la determinación de las bases requiere de más tiempo.
- Dependiendo de nuestras necesidades.

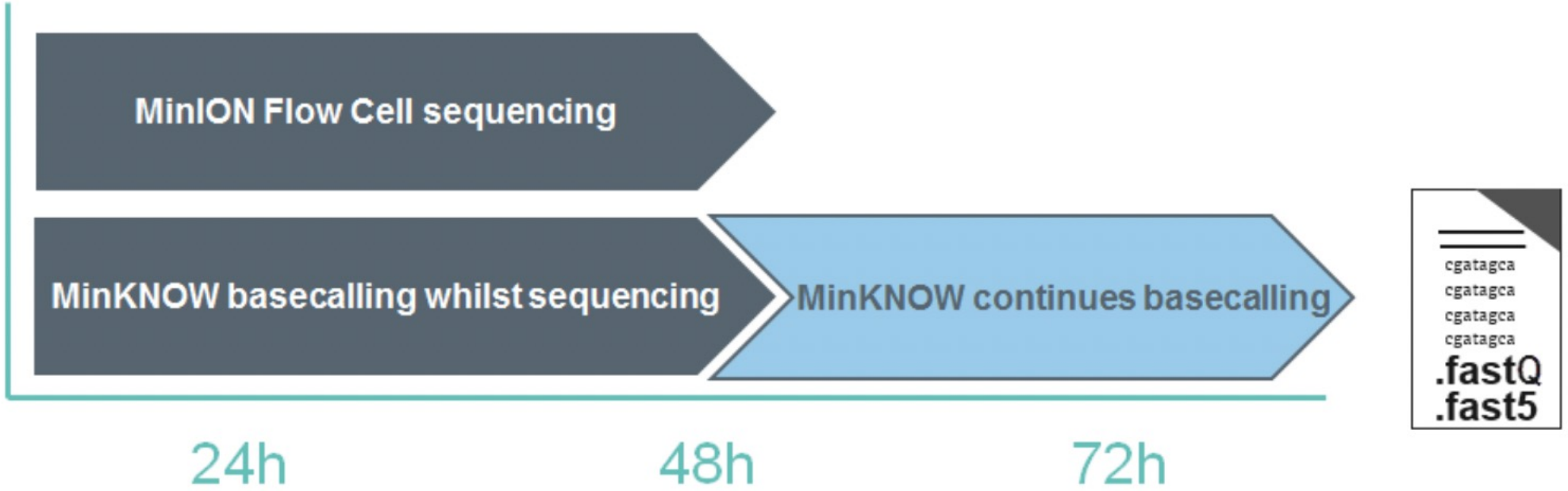
Modelos

- Fast está diseñado para seguir el ritmo de generación de datos en los dispositivos
- HAC proporciona una mayor precisión de lectura bruta que el modelo Fast. 5 a 8 veces más intensivo desde el punto de vista computacional.
- Super accurate tiene una precisión de lectura bruta aún mayor ~3 veces más intensivo que el modelo HAC.

Model	R9.4.1 modal accuracy	R10.3 modal accuracy
Fast	95.8	91.4
HAC	97.8	95.7
sup	98.3	97.5



Live basecalling



Tiempo de espera

Device	Basecalling speed in Gbases per hour			Keep-up, number of flow cells
	Fast model	HAC model	sup model	Fast model
PromethION P24	82	19	5.2	23
GridION	34	5	2	51
MinION Mk1C	0.8	0.07	-	1.25
MinION CPU, high-spec laptop	0.13	0.014	0.003	0.21
MinION GPU, RTX2070 laptop	14	1.6	0.06	22

De manera local

- Puede ser mucho más rápido.
- Primero en recibir actualizaciones de software.
- Permite usar opciones más avanzadas.

Opciones Guppy

- ~/ont-guppy/bin/guppy_basecaller \
- -i /fast5 \
- -s Analisis_S04112021/ \
- -c dna_r9.4.1_450bps_hac.cfg \
- --min_qscore 8 --barcode_kits "EXP-NBD196" \
- --require_barcodes_both_ends --compress_fastq --trim_barcodes \
- -x 'cuda:0' --chunk_size 2000 --chunks_per_runner 256 \
- --recursive

Opciones Guppy

- ~/ont-guppy/bin/guppy_basecaller \
- -i /fast5 \ Datos generados
- -s Analisis_S04112021/ \
- -c dna_r9.4.1_450bps_hac.cfg \
- --min_qscore 8 --barcode_kits "EXP-NBD196" \
- --require_barcodes_both_ends --compress_fastq --trim_barcodes \
- -x 'cuda:0' --chunk_size 2000 --chunks_per_runner 256 \
- --recursive

Opciones Guppy

- ~/ont-guppy/bin/guppy_basecaller \
- -i /fast5 \ Datos generados
- -s Analisis_S04112021/ \ Dónde guardar los datos
- -c dna_r9.4.1_450bps_hac.cfg \
- --min_qscore 8 --barcode_kits "EXP-NBD196" \
- --require_barcodes_both_ends --compress_fastq --trim_barcodes \
- -x 'cuda:0' --chunk_size 2000 --chunks_per_runner 256 \
- --recursive

Opciones Guppy

- ~/ont-guppy/bin/guppy_basecaller \
- -i /fast5 \ Datos generados
- -s Analisis_S04112021/ \ Dónde guardar los datos
- -c dna_r9.4.1_450bps_hac.cfg \ Modelo Redes Neuronales
- --min_qscore 8 --barcode_kits "EXP-NBD196" \
- --require_barcodes_both_ends --compress_fastq --trim_barcodes \
- -x 'cuda:0' --chunk_size 2000 --chunks_per_runner 256 \
- --recursive

Opciones Guppy

- ~/ont-guppy/bin/guppy_basecaller \
- -i /fast5 \ Datos generados
- -s Analisis_S04112021/ \ Dónde guardar los datos
- -c dna_r9.4.1_450bps_hac.cfg \ Modelo Redes Neuronales
- --min_qscore 8 --barcode_kits "EXP-NBD196" \
- --require_barcodes_both_ends --compress_fastq --trim_barcodes \
- -x 'cuda:0' --chunk_size 2000 --chunks_per_runner 256 \
- --recursive

Opciones de filtro y barconding

Opciones Guppy

- ~/ont-guppy/bin/guppy_basecaller \
- -i /fast5 \ Datos generados
- -s Analisis_S04112021/ \ Dónde guardar los datos
- -c dna_r9.4.1_450bps_hac.cfg \ Modelo Redes Neuronales
- --min_qscore 8 --barcode_kits "EXP-NBD196" \ Opciones de filtro y barconding
- --require_barcodes_both_ends --compress_fastq --trim_barcodes \
- -x 'cuda:0' --chunk_size 2000 --chunks_per_runner 256 \ Opciones de optimización
- --recursive

Antes de empezar práctica

- Vamos a descargar lo que necesitamos

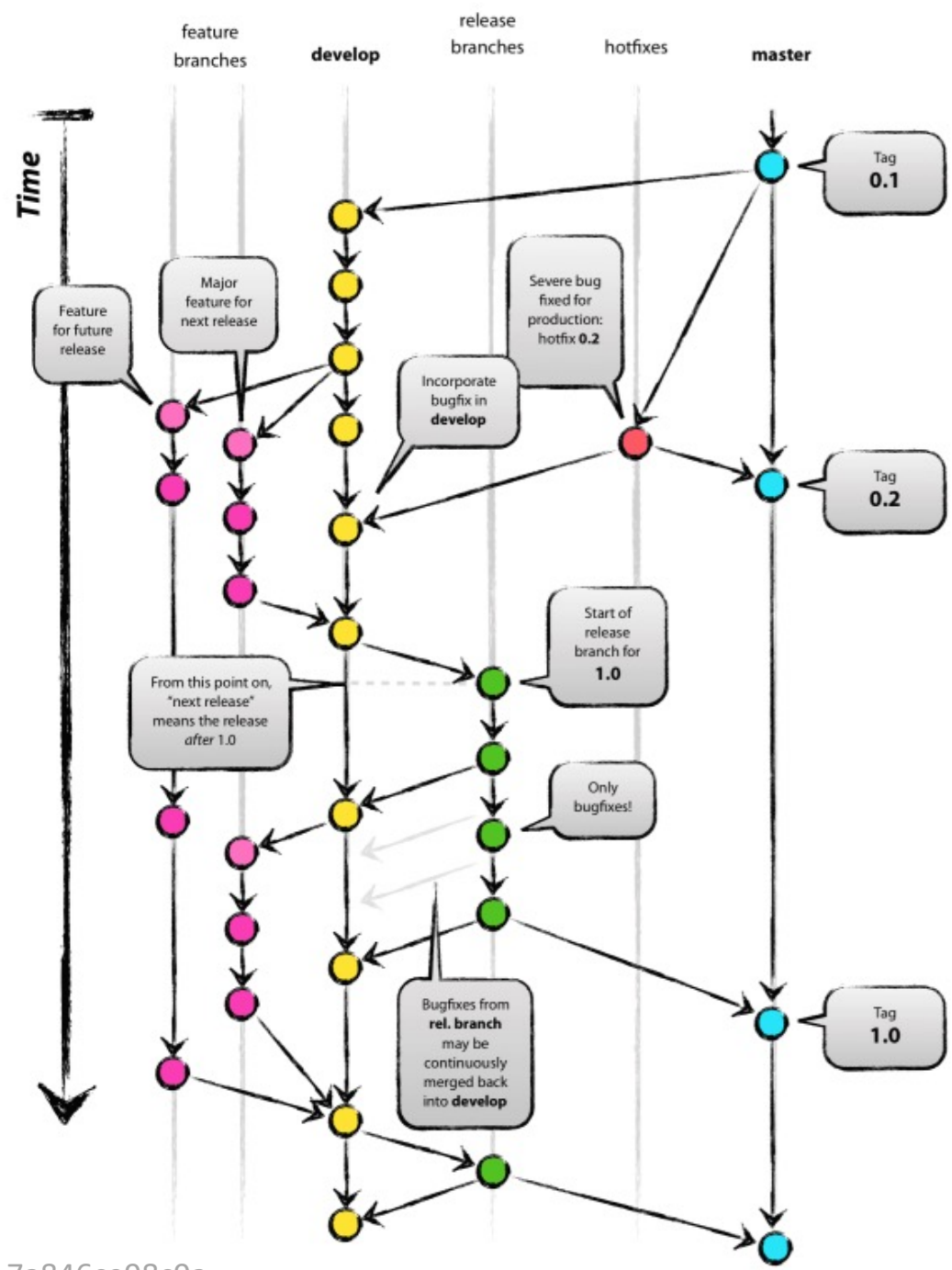
Antes de empezar práctica

- Vamos a descargar lo que necesitamos

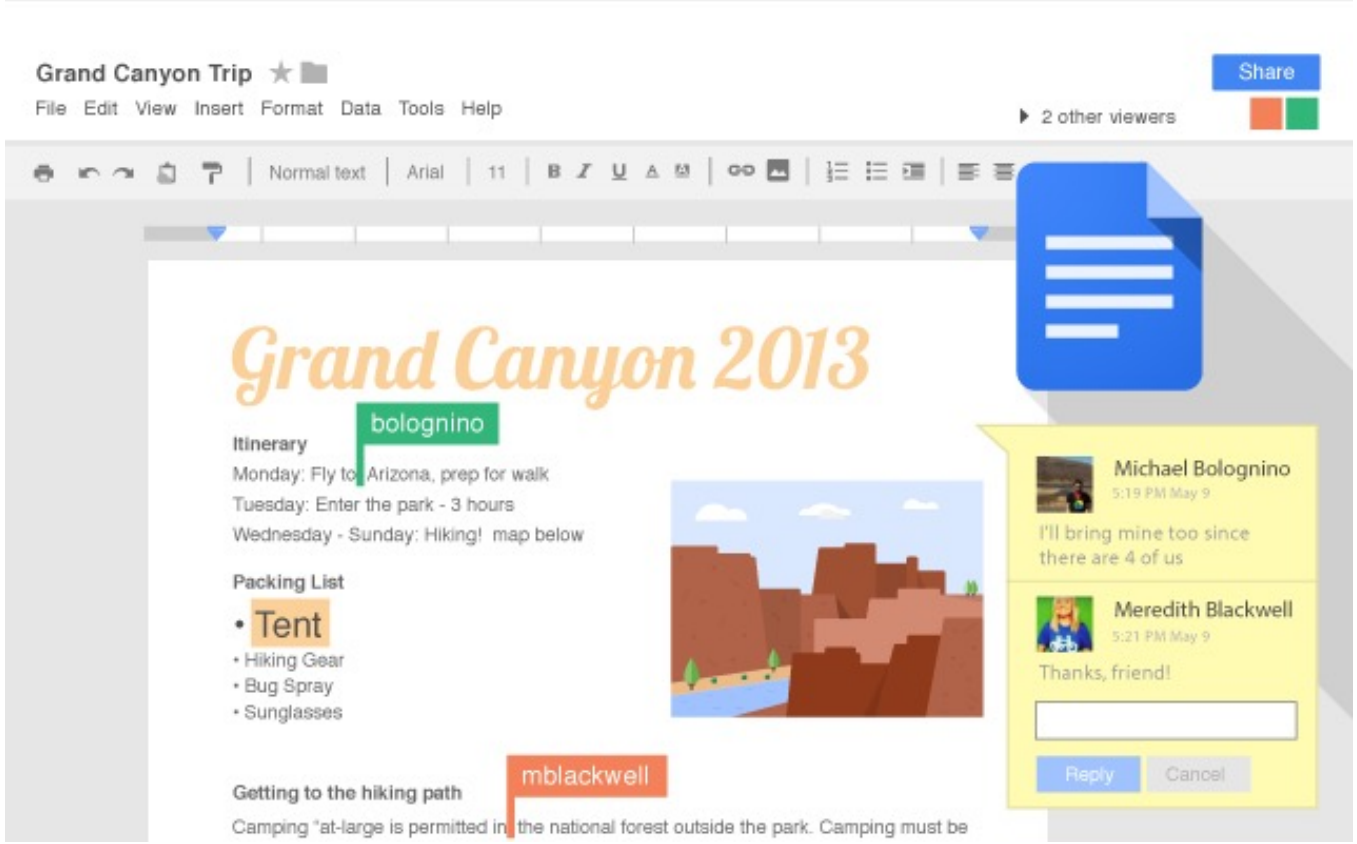
GitHub

El problema

- ¿Cómo tener una trazabilidad de todos los cambios que se le hacen a un programa?
- ¿Cómo ofrecer fácilmente las últimas versiones?
- Manera fácil de experimentar nuevos cambios en nuestro programa.



https://towardsdatascience.com/git-for-noobs-7a846ce98c9a



GitHub

https://blog.ida.cl/ida/funciones-herramientas-documentos-google-docs/

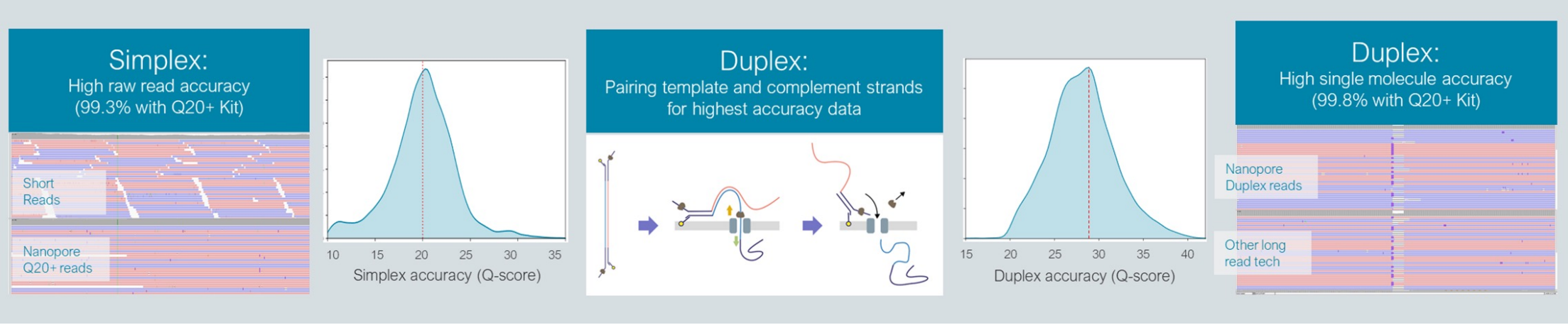
- Como desarrolladores permite:
- Sincronizar y compartir nuestro código (público o privado),
- Ver los cambios realizados.
- Interactuar con los usuarios.

- Como usuarios podemos:
- Ver el código de los programas y sus cambios.
- Utilizar el proyecto como base para nuestro uso.
- Descargar los archivos.

Práctica

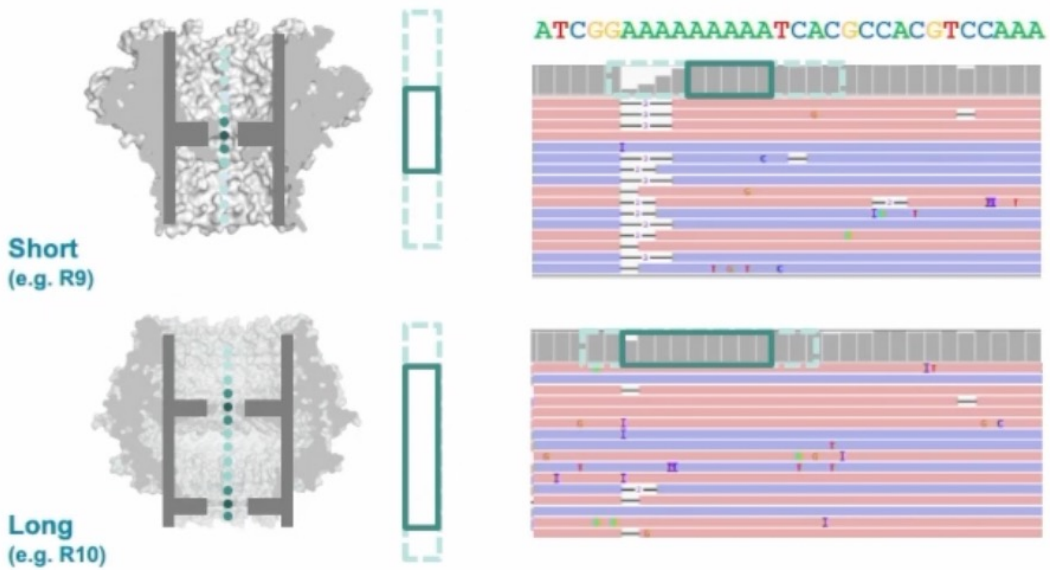
- Descargar el set de datos y realizar el llamado de bases con guppy.

Nueva tecnología



Short and long “reader heads”

- Length of the main discrimination site (“read head”) affects accuracy
- Short read heads allow easier decoding of individual bases (R9 series)
- Longer read heads see more range and are more information rich (R10 series)

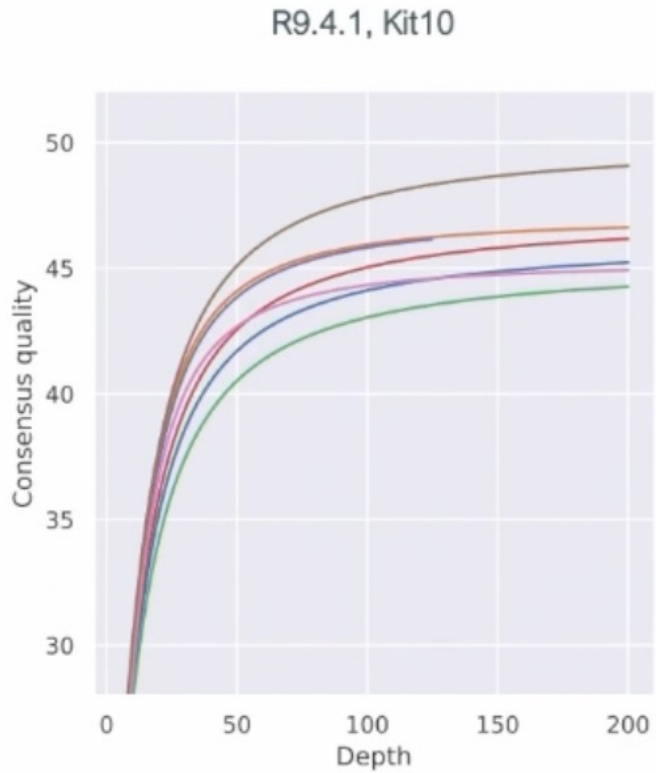


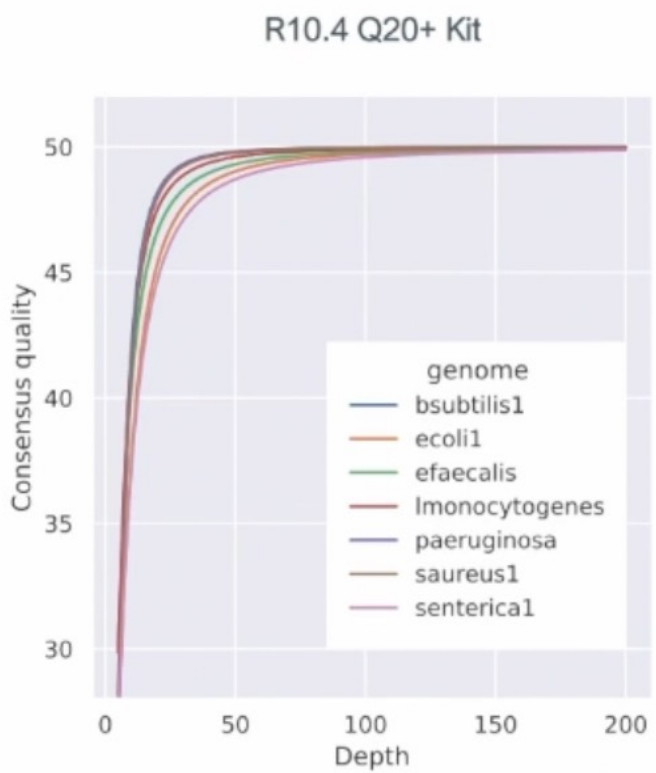
High accuracy results

Consensus accuracy

Results from our test set

- Latest results for R10.4, “Q20+” kit
  - “Sup” base calling
  - Zymo genomes
  - Flye assembly
  - Medaka consensus
- Comparison to R9.4.1, standard chemistry
- R10 series delivers higher consensus accuracy with far less data
- Team continue to train and improve base calling and consensus tools for both R9 and R10 series



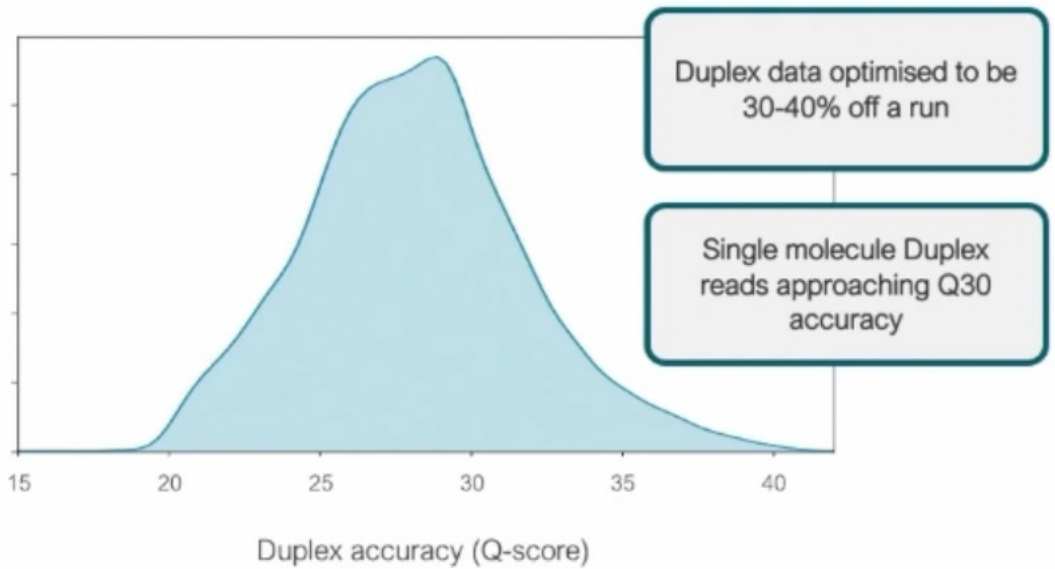


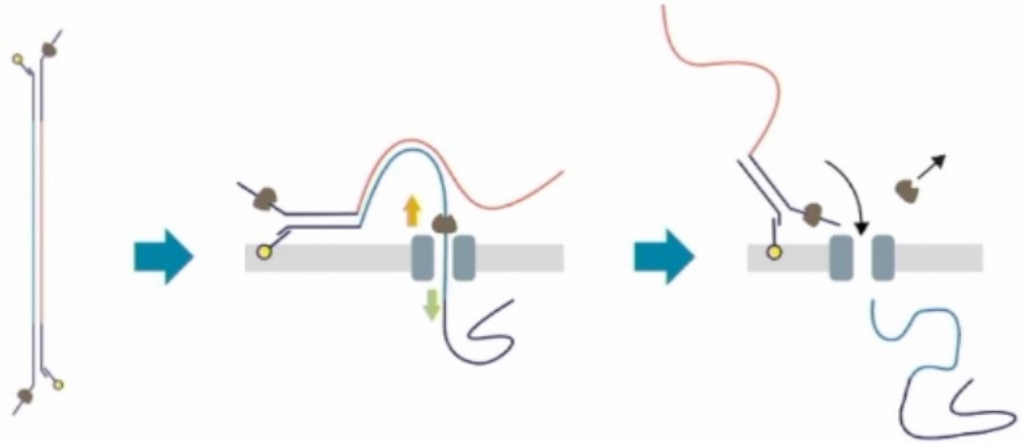
Nanopore accuracy

When we last spoke…

Duplex reads

- Possible when complement strand is sequenced immediately after template
- High duplex accuracy delivered by combining data template and complement
- New algorithms have been developed specifically for data combination
- Recent chemistries have optimised the amount of duplex data generated





Generating duplex data

- Chances of seeing the complement follow template increased with Q20+ chemistry
- Early protocols available in EA community
- Longest Duplex Q30 read to date: 156 kbase

Oxford NANOPORE Technologies