Pontificia Universidad Católica del Ecuador

Universidad de los Andes

ANÁLISIS DE DATOS DE SECUENCIACIÓN

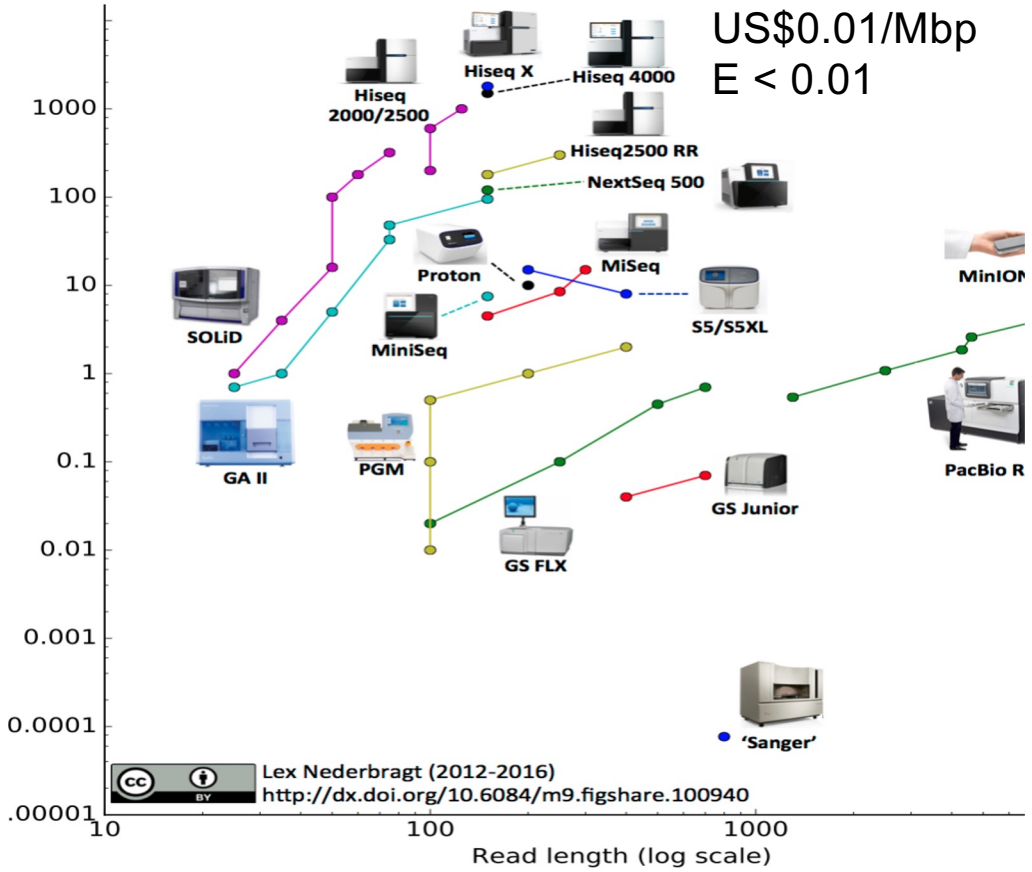
Jorge Duitama, PhD

Dpto Ingeniería de Sistemas
Universidad de los Andes

SGR
Sistema General de Regalías

TECNOLOGÍAS DE SECUENCIACIÓN

- Illumina puede producir hasta 1Tbp por corrida
- PacBio y Nanopore producen lecturas de más de 10Kbp

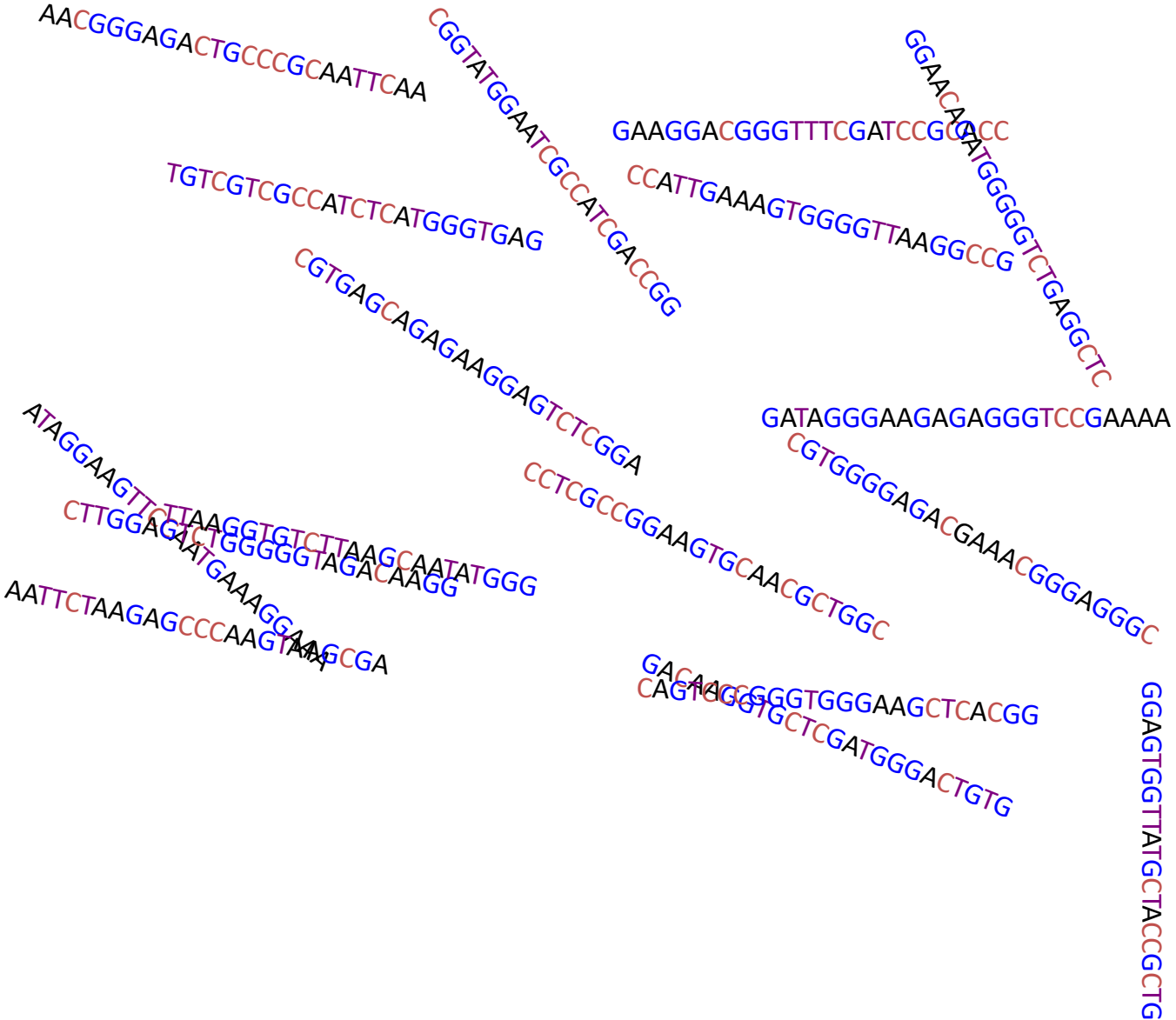


US$0.01/Mbp
E < 0.01

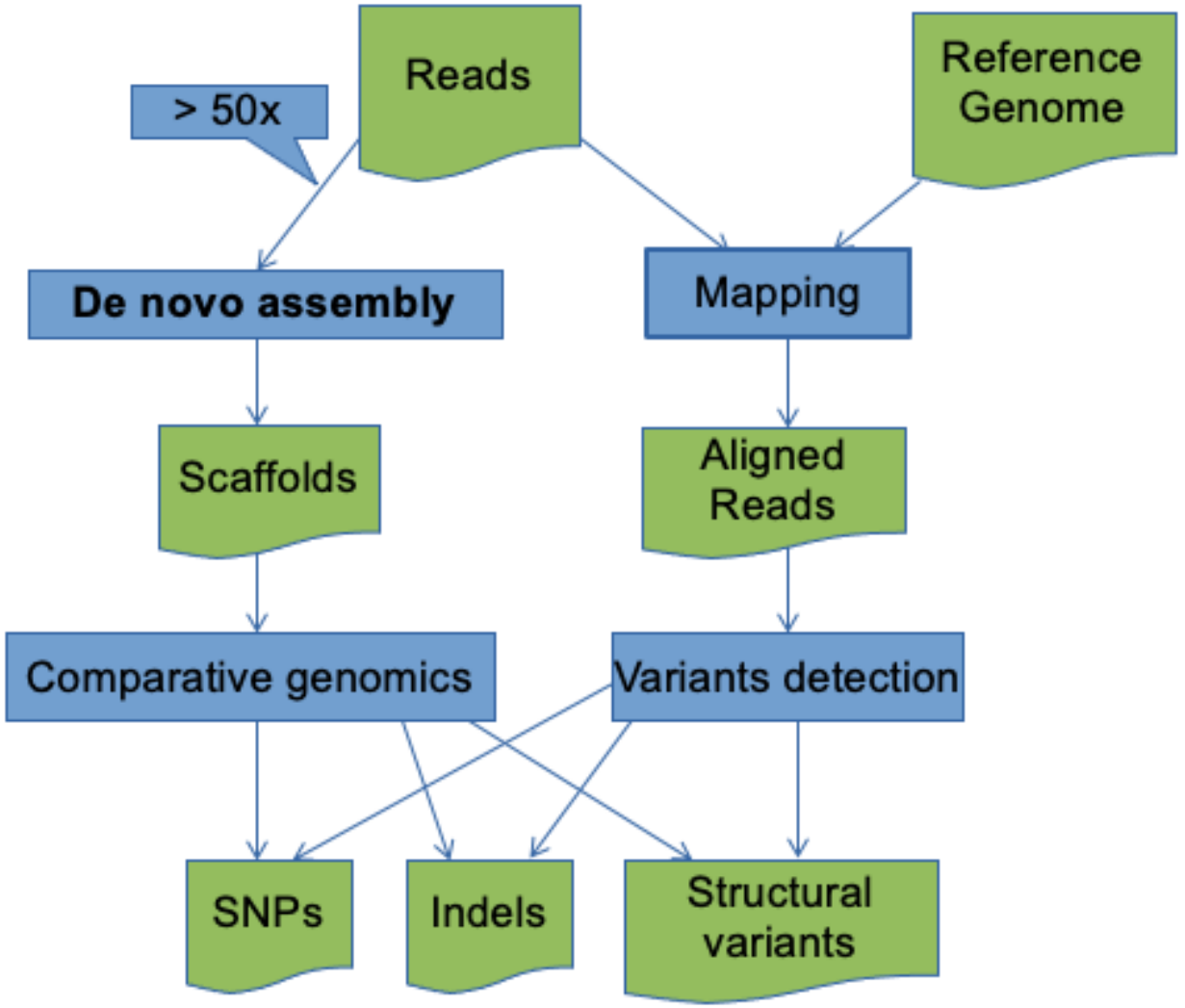
US$0.25/Mbp
E < 0.05

US$0.05/Mbp
E < 0.02

TECNOLOGÍAS DE SECUENCIACIÓN



BIOINFORMÁTICA PARA HTS

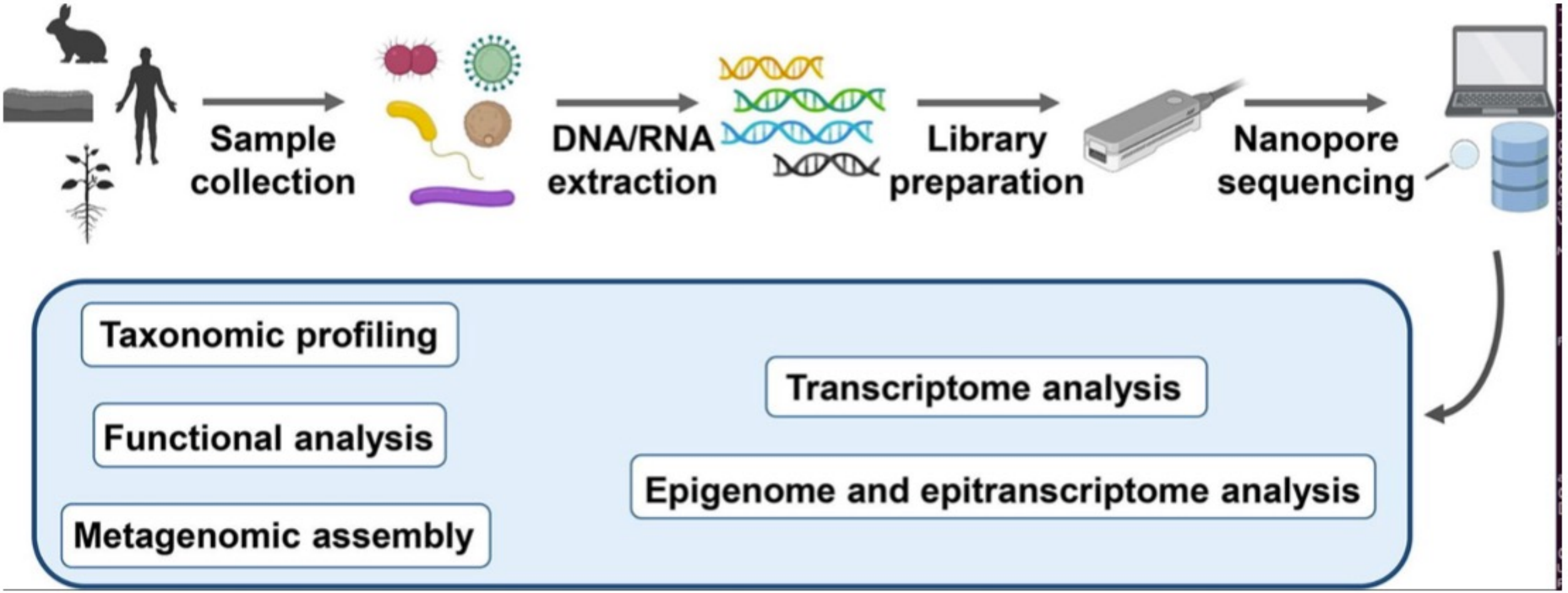


EJEMPLO CON TEXTO NATURAL

<nemo> <oy t> <mos > <nemo> <en b>
< cla> <orma> <enem> <emos> <tene>
<enem> <se d> <ritm> <atic> <orma>
<itmo> <ioin> < bio> <algo> <mos >
<info> <orit> <mos > < alg> <y te>
< de > <algo> <de a> <ioin> <n bi>
<n bi> <orma> < de > <e de> <lgor>
<en b> <enem> < alg> <clas> <oy t>
< alg> <mos > <atic> <mati> < alg>
<lgor> <rmat> <mos > <lase> < alg>

Tamaño estimado de secuencia: 50
Tamaño de lectura: 4
Número de lecturas: 50
Cubrimiento promedio: 4x
Texto: ?

What do you want to do with these long reads?

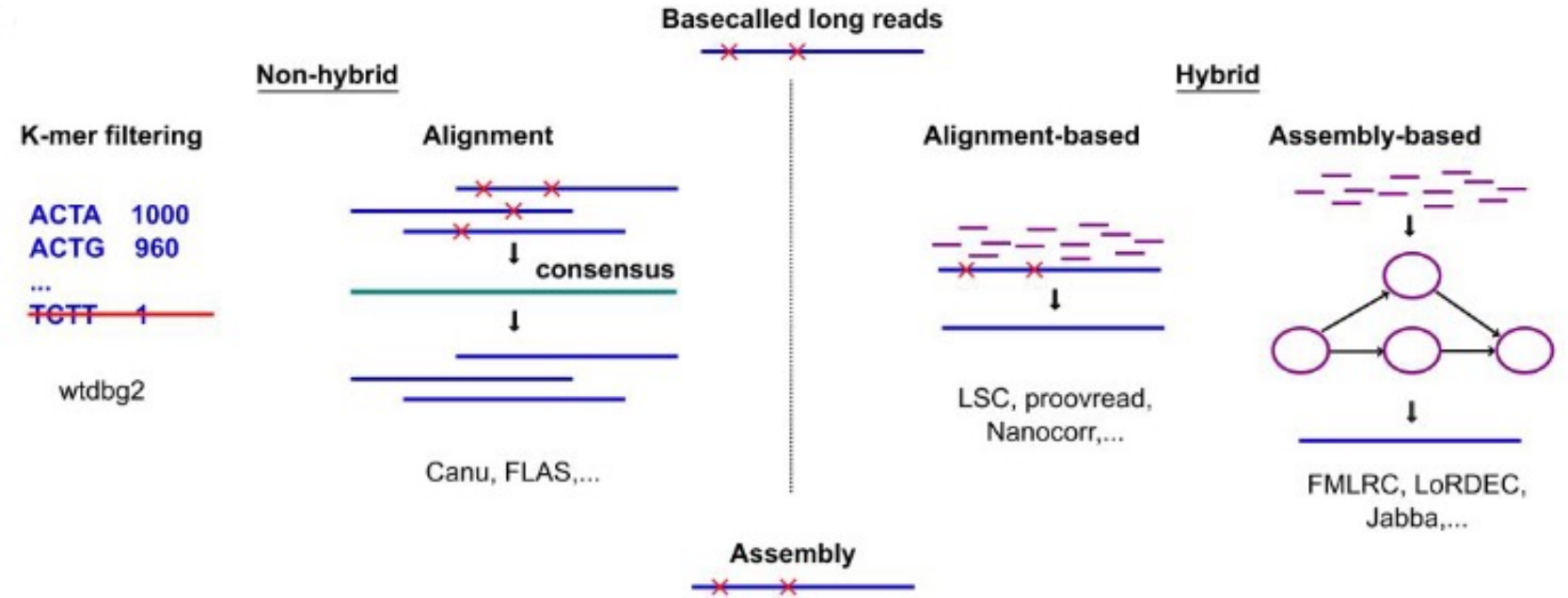


https://doi.org/10.1016/j.csbj.2021.02.020

Type	Reference	Application
Aligners/Alignment-based classifiers		
BLAST, MEGABLAST	[58,59]	Targeted; Shotgun
minimap2	[33]	Targeted; Shotgun
Alignment-free classifiers		
Kraken, Kraken2	[35,64]	Targeted; Shotgun
KrakenUniq	[65]	Shotgun
Bracken	[66]	Targeted; Shotgun
Metamaps	[69]	Shotgun
Centrifuge	[34]	Targeted; Shotgun
Mash	[72]	Targeted; Shotgun
Long-read assemblers		
Canu	[90]	Shotgun
miniasm	[73]	Shotgun
wtdbg2	[91]	Shotgun
OPERA-MS	[95]	Shotgun
MetaFlye	[96]	Shotgun
MetaSPAdes	[74]	Shotgun
Sequence correction and polishing tools		
Nanopolish	https://github.com/jts/nanopolish	Targeted; Shotgun
Medaka	https://github.com/nanoporetech/medaka	Targeted; Shotgun
Metagenomic analysis pipelines		
MEGAN-LR	[60]	Shotgun
NanoCLUST	[25]	Targeted
Reticulatus	https://github.com/SamStudio8/reticulatus	Shotgun
MUFFIN	[70]	Shotgun
NanoSPC	[71]	Shotgun
BusyBee	https://ccb-microbe.cs.uni-saarland.de/busybee/	Shotgun

https://doi.org/10.1016/j.csbj.2021.02.020

Reads Correction or not?



Reads Correction process

Correction strategies (*hybrid*)

– External reads : Illumina

– Internal reads : Only long reads or long reads corrected by short ones

Correction pipeline (*non-hybrid*)

– Read alignment

– Consensus calling

Canu module,

Racon can also be used as a read error-correction tool.

Assembly without reads correction

- Miniasm, Smartdenovo, Flye are members of this “new” family
- Improves speed
- Can work with less read depth.
- Can also assemble corrected reads

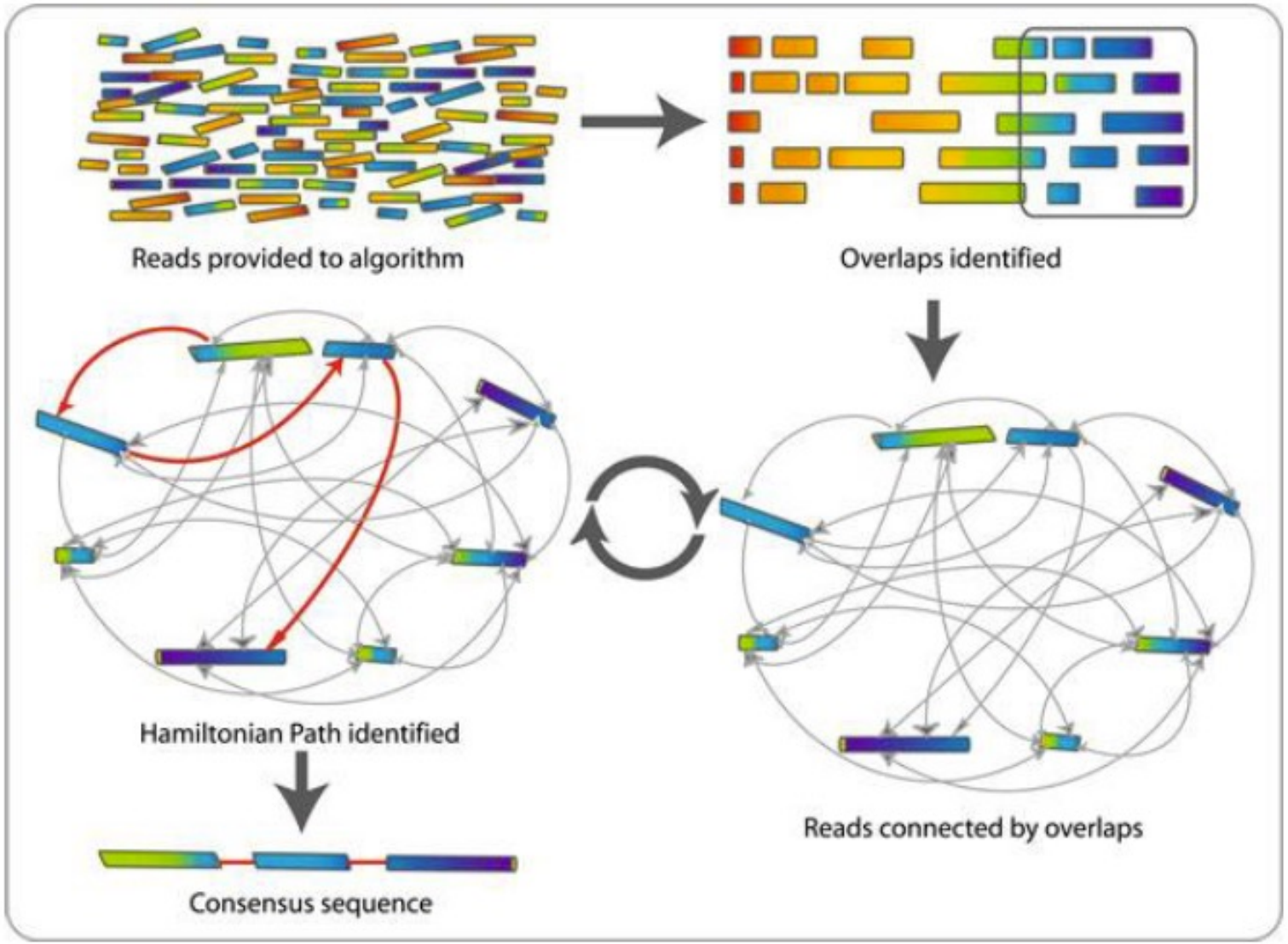
https://genomebiology.biomedcentral.com/track/pdf/10.1186/s13059-020-1935-5.pdf

What assembler to use over my favorite organism?

Long reads simplify genome assembly, with the ability *to* span repeat-rich sequences (characteristic of antimicrobial resistance genes) and structural variants. Nanopore sequencing also shows a lack of bias in GC-rich regions, in contrast to other sequencing platforms. To perform microbial genome assembly, we suggest using the third-party de novo assembly tool Flye. We also recommend one round of polishing with Medaka.
https://nanoporetech.com/sites/default/files/s3/literature/microbial-genome-assembly-workflow.pdf

For assembly, ONT recommend sequencing a human genome to a minimum depth of 30x of 25–35 kb reads. However, sequencing to a depth of 60x is advisable to obtain the best assembly metrics. We also recommend basecalling in high accuracy mode. Greatest contig N50 is usually obtained with Shasta and Flye. Polishing/Correction is also recommended (Racon and Medaka).
https://nanoporetech.com/sites/default/files/s3/literature/human-genome-assembly-workflow.pdf

Overlap–layout–consensus genome assembly algorithm (OLC)



Canu, Flye, Miniasm, Raven, Smartdenovo, Shasta

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3055744/

EJEMPLO CON TEXTO NATURAL

<rmat> <nfor> <gori> <mos > <y te>
<enem> <lase> <ritm> <se d> <n bi>
< en > <algo> <y te> < alg> <nfor>
<y te> <oy t> <bioi> <tene> <clas>
<atic> < ten> <s en> <hoy > <bioi>
<tene> <mos > <ase > <emos> <atic>
<lgor> <orma> <oinf> <nemo> < ten>
<itmo> <lase> <emos> <orma> <nfor>
<lgor> <n bi> <info> <itmo> < cla>
< cla> <itmo> <itmo> <os e> <en b>
<s en> <atic> <info> <ritm> <form>
< en > <mati> <y te> <nfor> <tene>
< en > <nemo> <itmo> <nemo> <mos >
<mos > <enem> <nfor> <e de> <mati>
< ten> <y te> <clas> < ten> < en >
<ase > <nfor> <lgor> <se d> < bio>
<oinf> <ritm> <nemo> <orit> <s cl>
<algo> <hoy > <s en> <os e> <e de>
<form> <n bi> <mati> <clas> <se d>
<rmat> <emos> <oinf> <mati> < alg>

Tamaño estimado de secuencia: 50
Tamaño de lectura: 4
Número de lecturas: 100
Cubrimiento promedio: 8x
Texto: ?

EJEMPLO CON TEXTO NATURAL

<oritmos > < tenemos> <ritmos e>
<emos cla> <y tenemo> < clase d>
<nformati> <os clase> < en bioi>
<e algori> < clase d> <lgoritmo>
<itmos en> <os en bi> <informat>
<s en bio> <y tenemo> <os clase>
<ase de a> < algorit> <de algor>
< tenemos> <en bioin> <bioinfor>
<algoritm> <n bioinf> <nemos cl>
<clase de> <lase de > <lgoritmo>
<nformati> < de algo> <nformati>
<e de alg> <oritmos > < clase d>
<lase de > <s clase > < de algo>
<emos cla> <tmos en > <ioinform>
<nemos cl> <nformati> <oritmos >
<se de al> <e de alg> <oy tenem>
<mos en b> <algoritm>

Tamaño estimado de secuencia: 50
Tamaño de lectura: 8
Número de lecturas: 50
Cubrimiento promedio: 8x
Texto: ?

EJEMPLO CON ADN

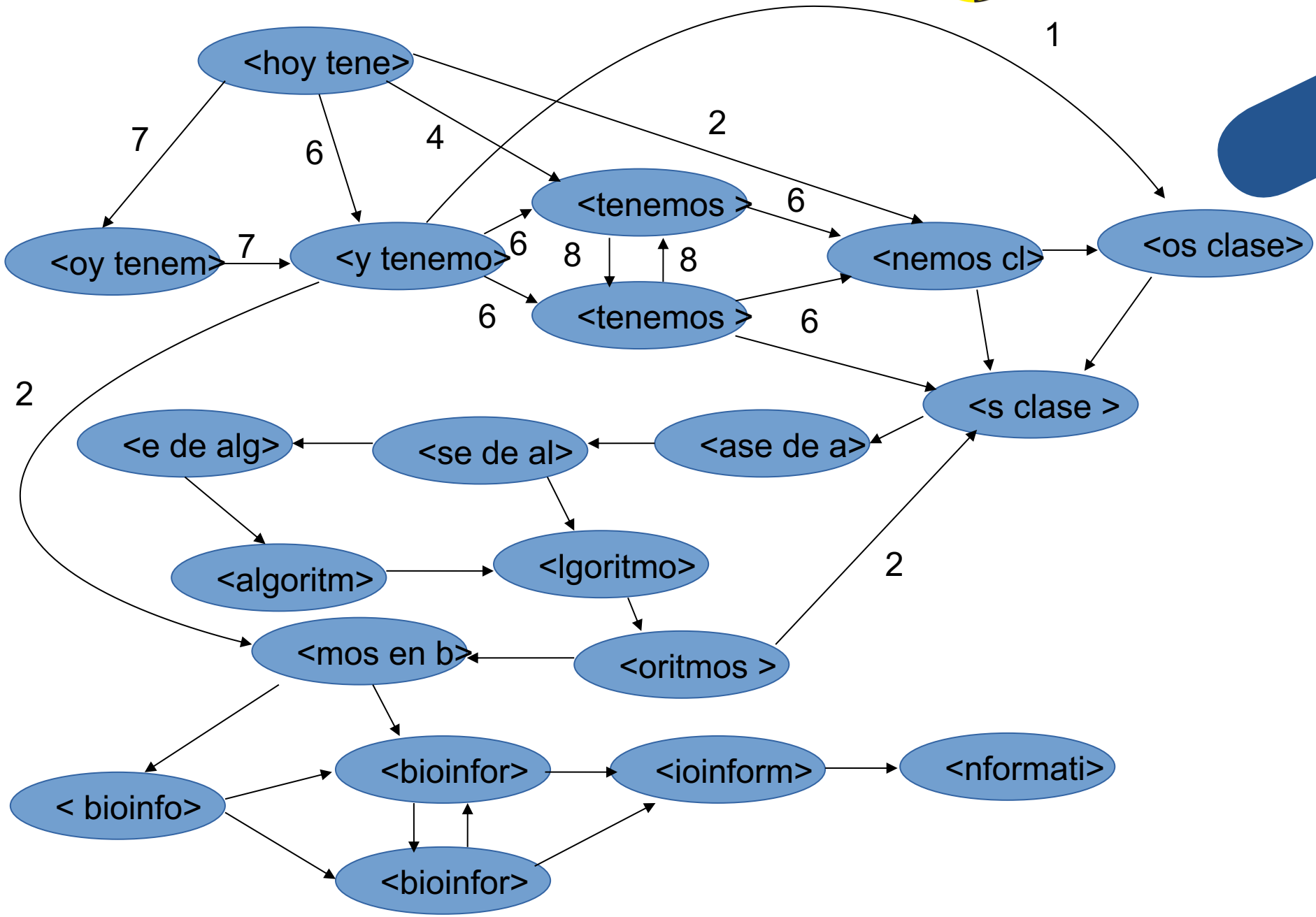
<TAGCTAAT> <GCTAGCTA> <AGCGTACT> <ACAGCGTA> <CAGCGTCG> <ACGTACGT> <TACTTGCG> <TACCGCTA>
<TACGTACC> <GTACGTAC> <GTACTTGC> <ACGTACCG> <CTAATAAC> <GCTAGCTA> <AACAGCGT> <AGCGTACT>
<TACGTACG> <GGCAGCGT> <GGCAGCGT> <ACGTACCG> <ACAGCGTA> <AGGCAGCG> <CGTACGTA> <ATAACAGC>
<TAATAACA> <TACTTGCG> <TACCGCTA> <TAACAGCG> <AACAGCGT> <CGTACTTG> <GCTAATAA> <TACGTACG>
<CGTCGTAC> <CGTACGTA> <ACGTACCG> <AGGCAGCG> <GTACGTAC> <GTCGTACG> <TCGTACGT> <GCTAGCTA>
<TACTTGCG> <AACAGCGT> <CAGCGTCG> <GTACTTGC> <GTACGTAC> <ACCGCTAG> <CAGCGTAC> <AGCTAATA>
<AGCTAATA> <ACGTACGT>

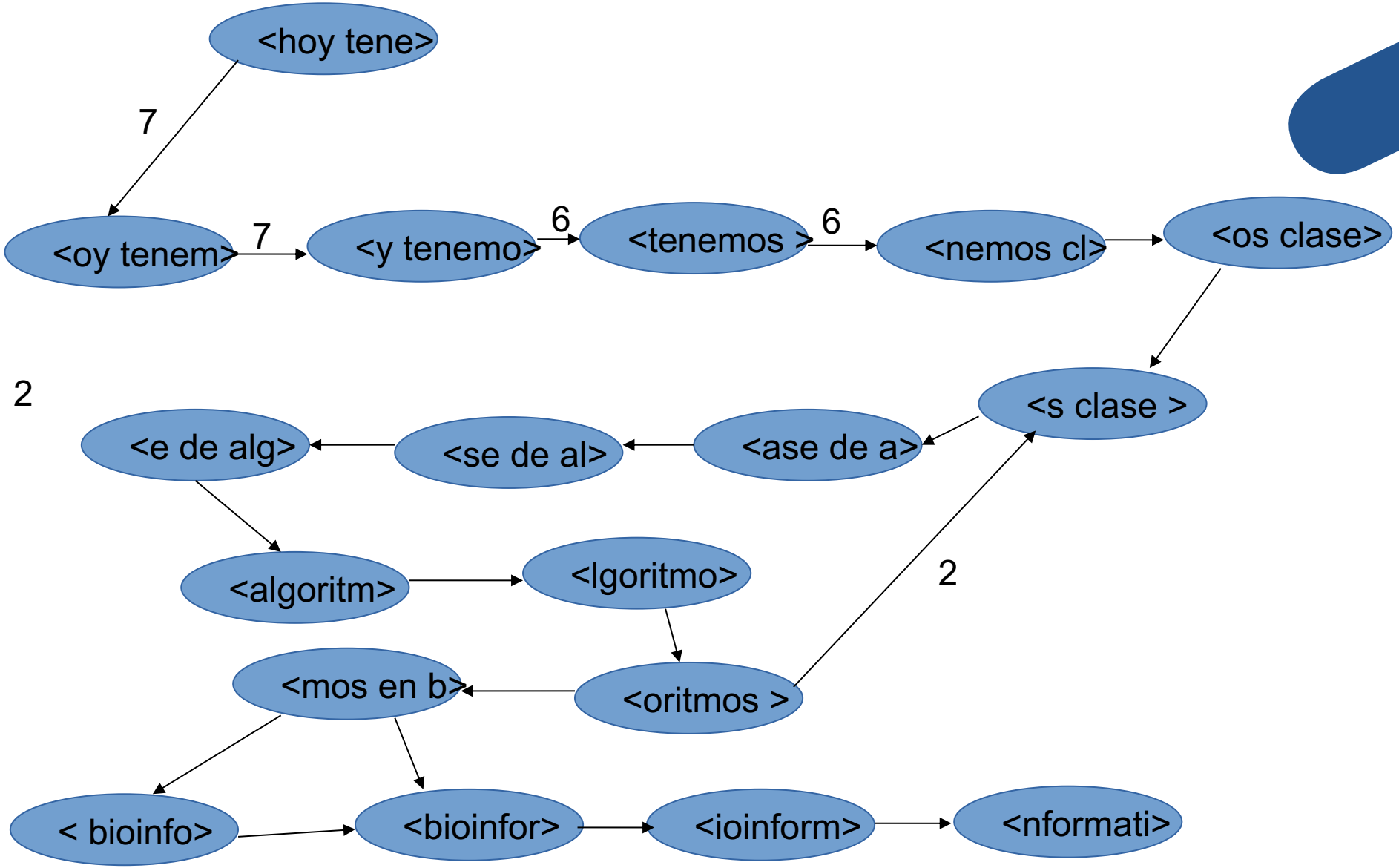
Tamaño estimado de secuencia: 50
Tamaño de lectura: 8
Número de lecturas: 50
Cubrimiento promedio: 8x

Secuencia: AGGCAGCGTCGTACGTACGTACCGCTAGCTAATAACAGCGTACTTGCGT

GRAFO DE ENSAMBLAJE

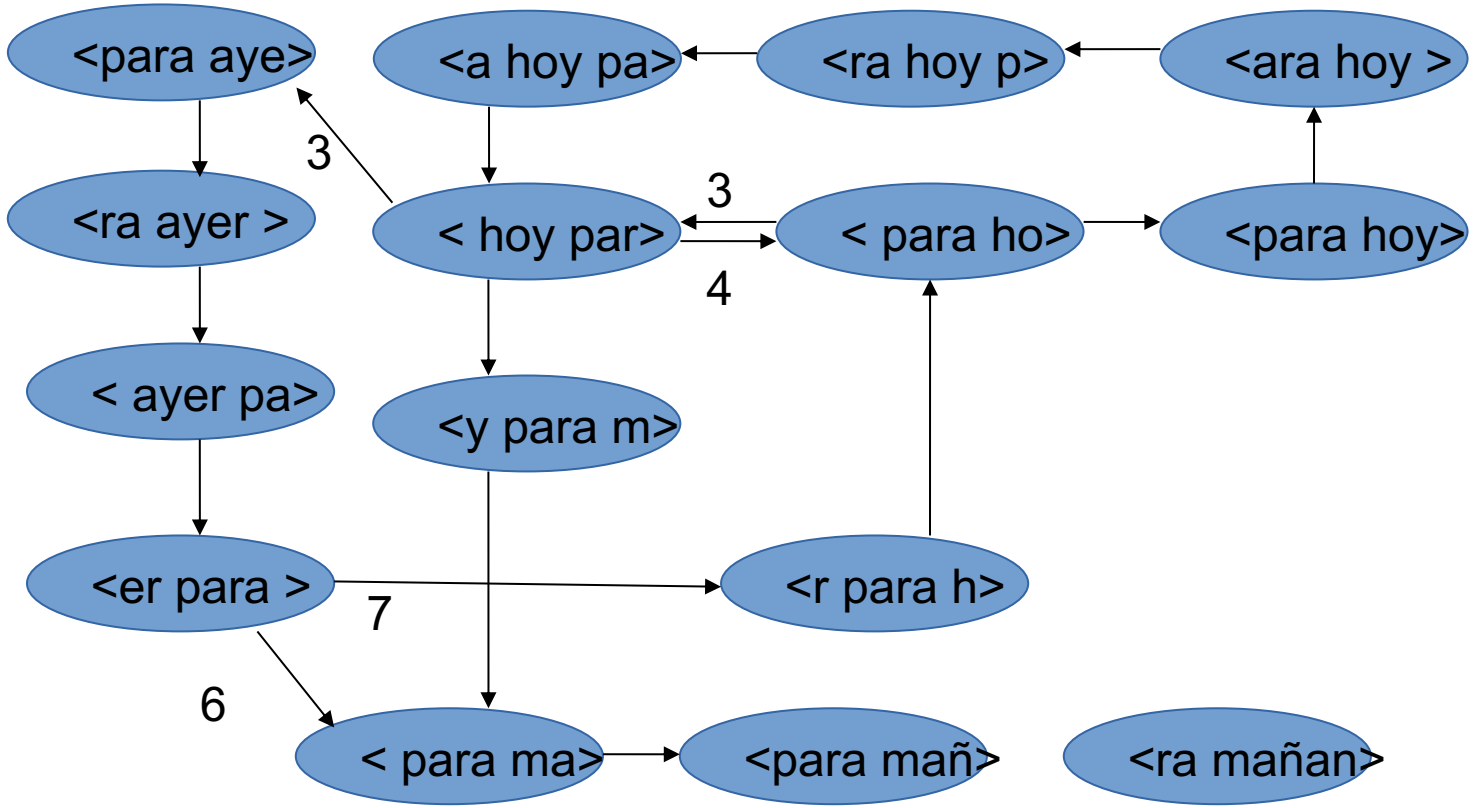
- Grafo dirigido
- Los vértices son las lecturas
- Hay un eje entre cada par de vértices si las lecturas se sobrelapan
- El número de bases en las que se sobrelapan es el peso del eje


<hoy tene>
<oy tenem>
<y tenemo>
<tenemos >
<tenemos >
<nemos cl>
<os clase>
<s clase >
<e de alg>
<se de al>
<ase de a>
<algoritm>
<lgoritmo>
<mos en b>
<oritmos >
< bioinfo>
<bioinfor>
<ioinform>
<nformati>
<bioinfor>
1
2
7
6
4
2
7
6
8
8
6
6
6
2
2


<hoy tene>
7
<oy tenem>
7
<y tenemo>
6
<tenemos >
6
<nemos cl>
<os clase>
2
<s clase >
<e de alg>
<se de al>
<ase de a>
<algoritm>
<lgoritmo>
2
<mos en b>
<oritmos >
< bioinfo>
<bioinfor>
<ioinform>
<nformati>

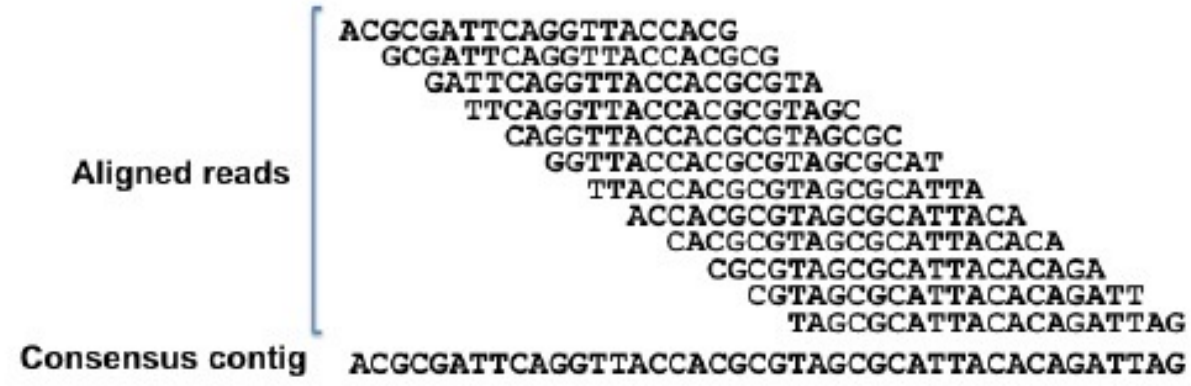
EJERCICIO CON TEXTO

< hoy par> < hoy par> < ayer pa> < ayer pa> <ara hoy >
<er para > <y para m> <para hoy> <er para > <para mañ>
<ra hoy p> <a hoy pa> <ara hoy > <para mañ> <ra ayer >
<r para h> <hoy para> <ra mañan> < hoy par> < para ho>
< para ma> <para mañ> <para aye> < hoy par> < ayer pa>



PROCESO DE ENSAMBLAJE

1. **Overlap**: Construir el grafo de sobrelapes (Overlap graph)
2. **Layout**: Encontrar el o los caminos en el grafo de sobrelapes que explican las lecturas
3. **Consensus**: Construir la secuencia de consenso a partir de las lecturas alineadas



http://image.slidesharecdn.com/110928clsioncodgenome-110929063927-phpapp01/95/how-to-sequence-a-large-eukaryotic-genome-15-728.jpg?cb=1357998263

Polishing / Correction

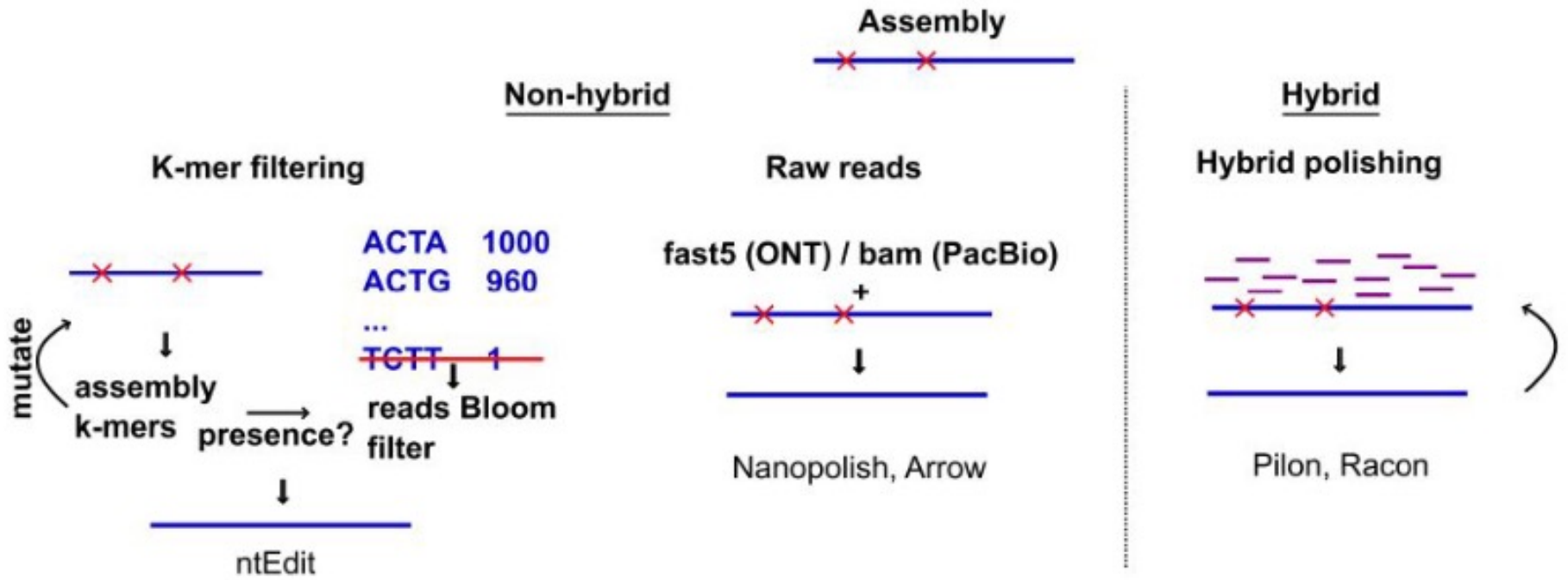
Racon correct raw contigs generated by rapid assembly methods which do not include a consensus step. It can polish with either Illumina data or data produced by third generation of sequencing. (recursive use)

Medaka and **Nanopolish** create a consensus sequence of nanopore sequencing data. (mapping + consensus)

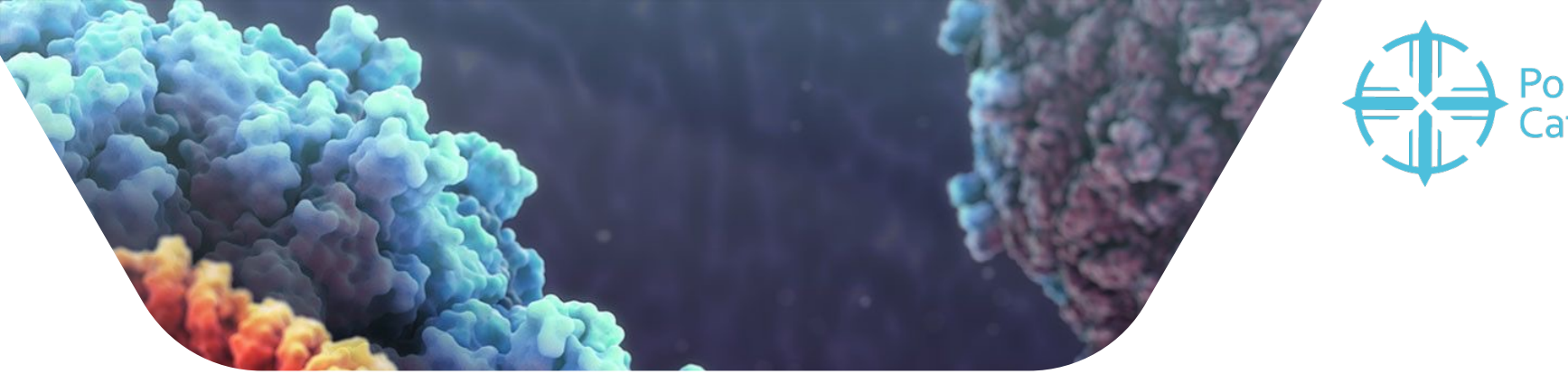
+ Medaka uses neural networks where Nanopolish uses HMMs.
+ Medaka uses basecalled reads, not the raw signal.
+ Medaka propose the ability to train one's own basecalling model

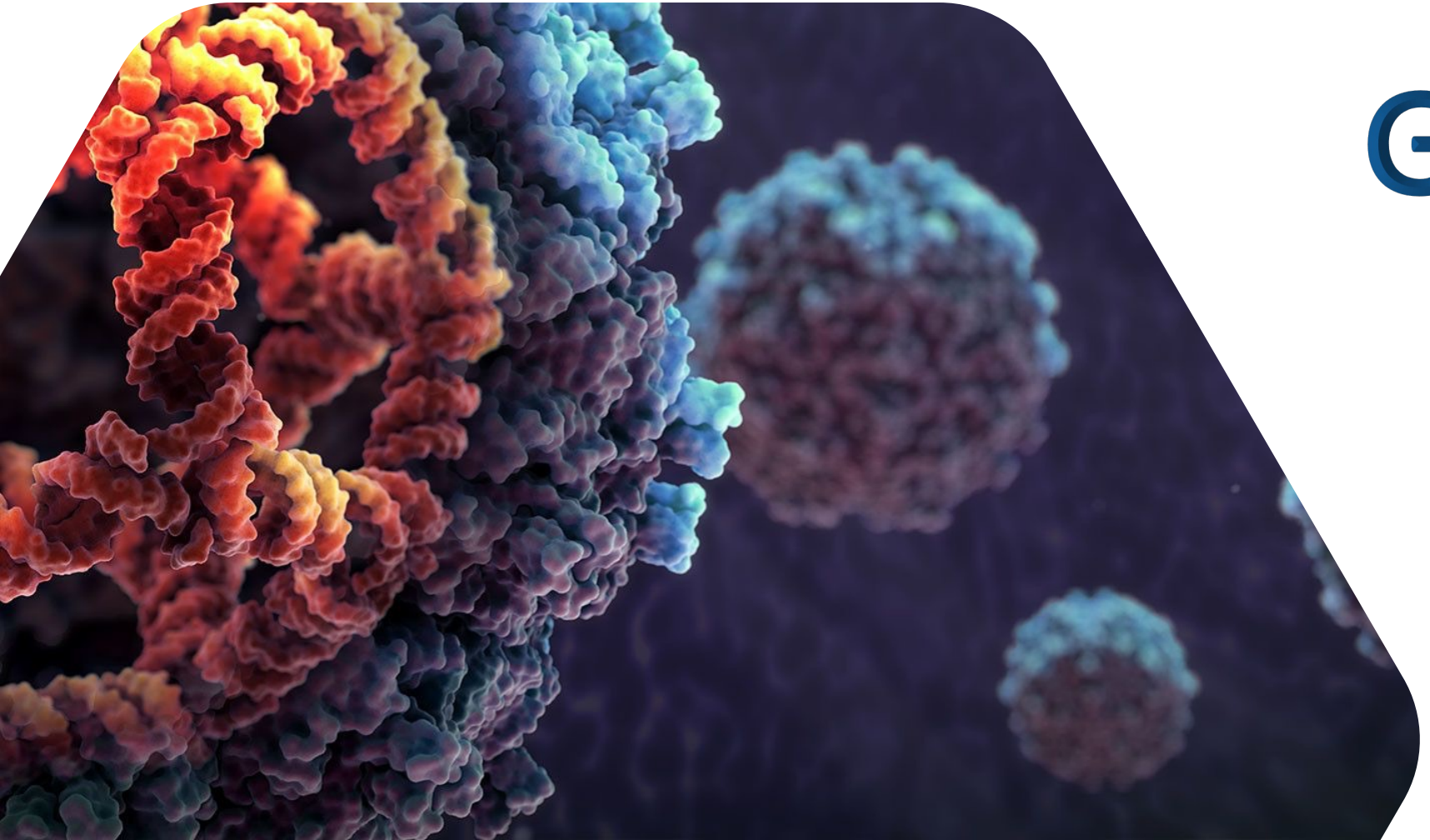
Pilon correct assemblies using illumina reads. (recursive use)

Autres : NeuralPolish , ntEdit



https://genomebiology.biomedcentral.com/track/pdf/10.1186/s13059-020-1935-5.pdf





Pontificia Universidad Católica del Ecuador

Universidad de los Andes

GRACIAS

SGR
Sistema General de Regalías