INTRODUCCIÓN A LA BIOINFORMÁTICA

JAZMÍN SÁNCHEZ PÉREZ
LICENCIATURA EN CIENCIAS GENÓMICAS

DRA. VALERIA SOUZA

INSTUTO DE ECOLOGÍA
UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

CLICK TO EDIT MASTER TITLE STYLE

- Unidad IV – Genómica
  - Introducción al estudio de genomas
  - Ensamble de genomas
  - Calidad de genomas
  - Anotación
  - Filogenias de genomas
- Unidad V – Metagenómica
  - Introducción al estudio de metagenómica
  - Metabarcording (Anotación taxonómica)
  - Shotgun metagenomics (Anotación taxónomica y funcional)
  - Índices de diversidad
  - R como herramienta

Click to edit the outline text format

Second Outline Level

Third Outline Level

Fourth Outline Level

Fifth Outline Level

Sixth Outline Level

Seventh Outline LevelClick to edit Master text styles

Second level

Third level

Fourth level

Fifth level

- *-OMICS* es en sufijo que describe algo grande, y se refiere al campo de estudio de las ciencias de la vida que se enfoca en información a grande escala.

- "The other twist to "omics" may be associated with the "Om" (pronounced "Aum"), an ancient Sanskrit intonation, which, like music, transcends the barriers of age, race, culture, and even species. "

Yadav, Satya P. "The Wholeness in Suffix -Omics, -Omes, and the Word Om." (2007)

¿QUÉ ES GENÓMICAS?

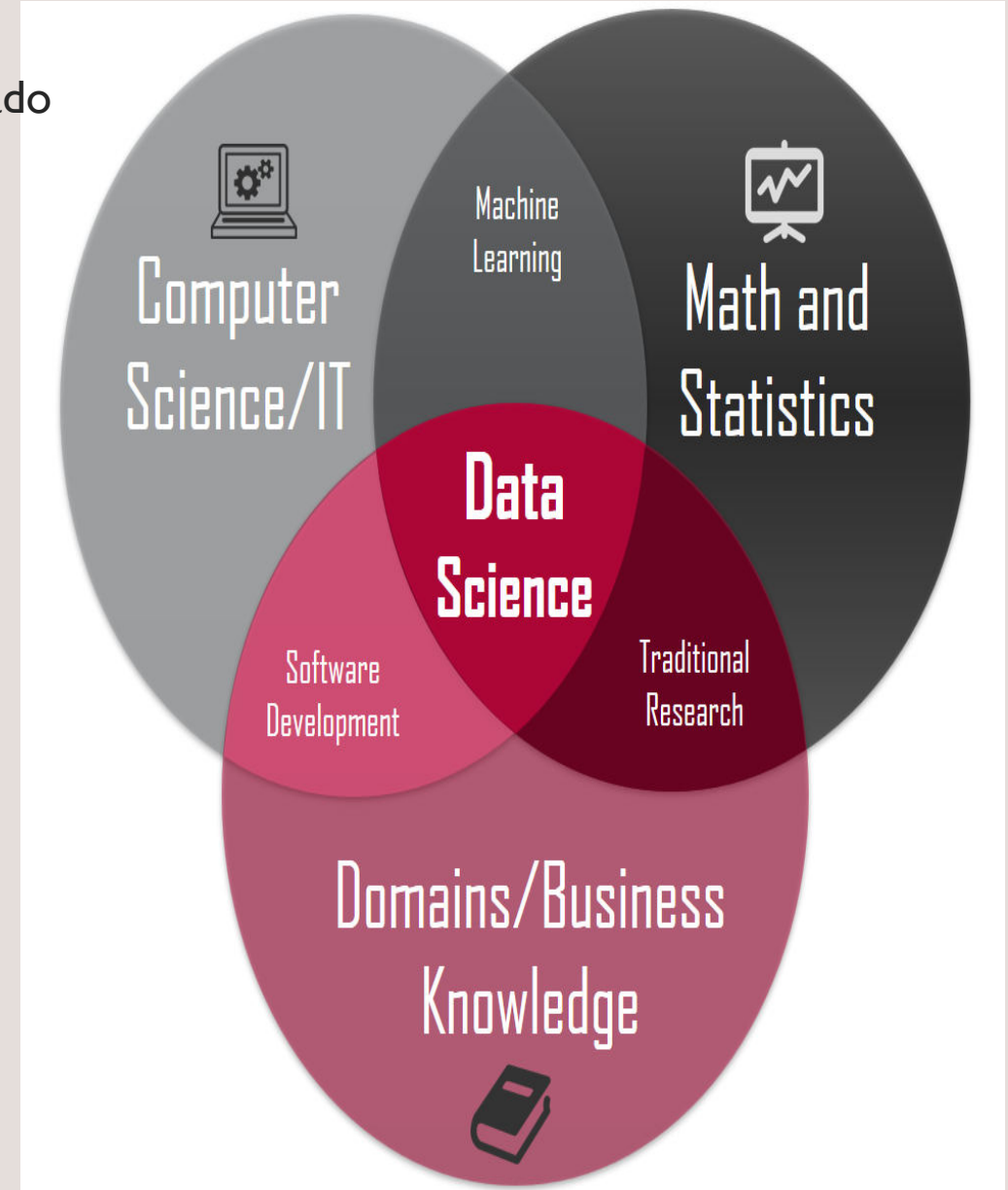
En un McDonald´s Raw Bar, con una cerveza en la mano y unos amigos…

Dr. Thomas H. Roderick, en 1986, acuña el termino *genomics*, como nombre para una revista que tratara de temas del genoma humano.

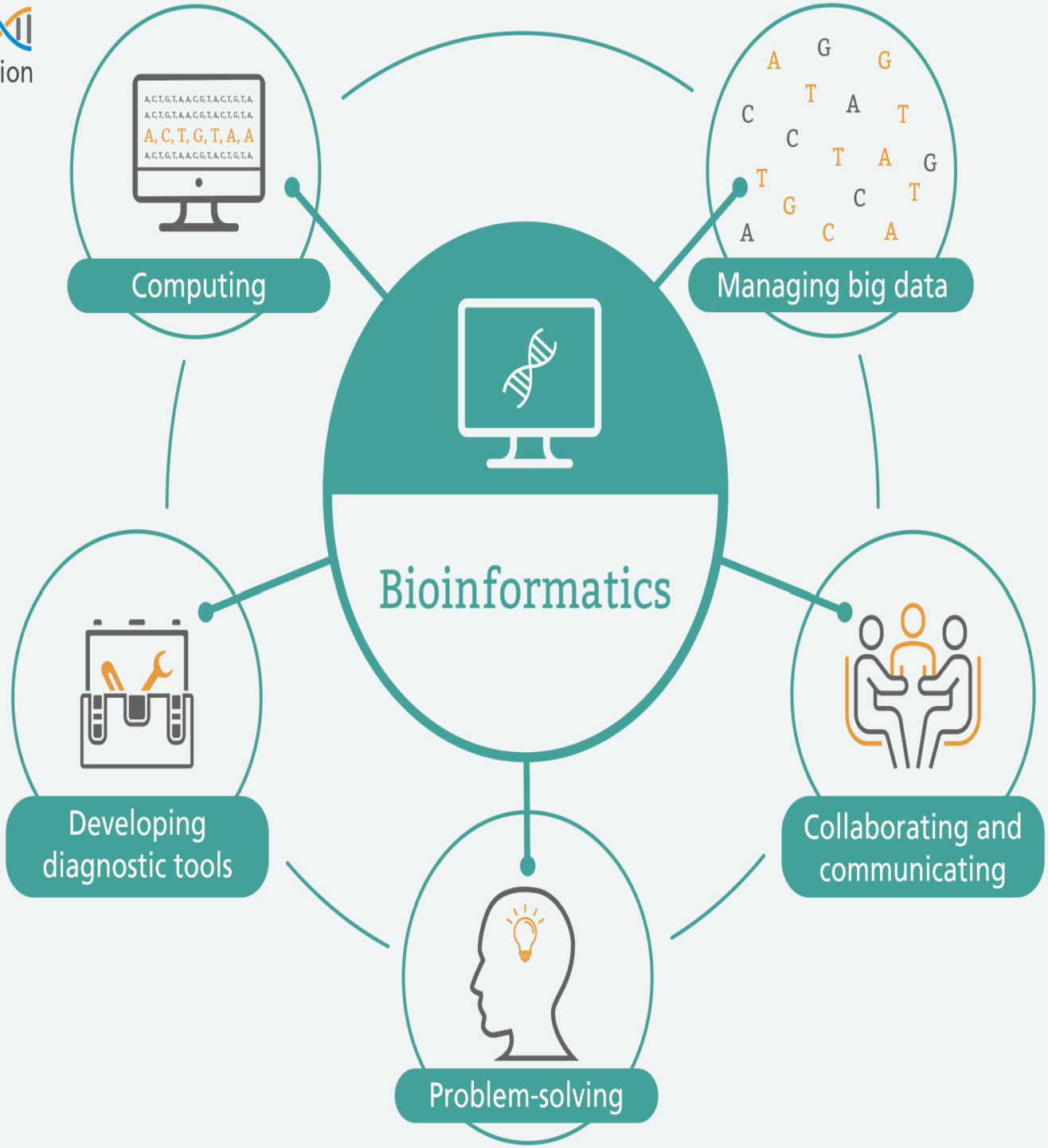
(Beer, Bethesda, and biology: how "genomics" came into being.)

1976 – Viral RNA-genome bacteriofago MS2
1995 – El primer genoma eucarionte secuenciado *S.cerevisiae*
1996 – Ciencias de Datos
2016 – Secuencia del Genoma Humano

"In 2012, when Harvard Business Review called it "The Sexiest Job of the 21st Century", the term "data science" became a buzzword. "



Genomics Education Programme



GENOMICA

TOOLS & GENOME ASSEMBLE

CLICK TO EDIT MASTER TITLE STYLE

Full Report ▾ Send to: ▾

ASM190086v1

Organism name: Borreliella afzelii (spirochetes)
Infraspecific name: Strain: BO23
BioSample: SAMN05941949
BioProject: PRJNA350557
Submitter: NIH/NIAID
Date: 2016/12/09
Assembly level: Chromosome
Genome representation: full
GenBank assembly accession: GCA_001900865.1 (latest)
RefSeq assembly accession: GCF_001900865.1 (latest)
RefSeq assembly and GenBank assembly identical: yes
Assembly method: SPAdes v. 3.7.1
Expected final version: yes
Genome coverage: 2000.0x
Sequencing technology: Illumina NextSeq 500

IDs: 914131 [UID] 3786338 [GenBank] 3861958 [RefSeq]

See Genome Information for **Borreliella afzelii**

There are 11 assemblies for this organism

See more

History (Show revision history)

Comment

Annotation was added by the NCBI Prokaryotic Genome Annotation Pipeline (released 2013). Information about the Pipeline can be found here: https://www.ncbi.nlm.nih.gov/genome/annotation_prok/

Global statistics

Total sequence length	1,250,408

Access the data

Download the RefSeq assembly
Download the GenBank assembly
Download the full sequence report
Download the statistics report

Assembly Information

Assembly Help
Assembly Basics
NCBI Assembly Data Model

Related Information

BioProject
BioSample
Genome
Nucleotide INSDC
Nucleotide RefSeq
Taxonomy

Full Report ▾ Send to: ▾

Global statistics

Total sequence length	1,259,498
Total assembly gap length	0
Gaps between scaffolds	0
Number of scaffolds	32
Scaffold N50	905,394
Scaffold L50	1
Number of contigs	32
Contig N50	905,394
Contig L50	1
Total number of chromosomes and plasmids	6
Number of component sequences (WGS or clone)	32

be found here: https://www.ncbi.nlm.nih.gov/genome/annotation_prok/

Global statistics

Access the data
Download the RefSeq assembly
Download the GenBank assembly
Download the full sequence report
Download the statistics report

Assembly Information
Assembly Help
Assembly Basics
NCBI Assembly Data Model

Related Information
BioProject
BioSample
Genome
Nucleotide INSDC
Nucleotide RefSeq
Taxonomy

CLICK TO EDIT MASTER TITLE STYLE

chemotaxis protein CheR [Borreliella afzelii]

GenBank: APJ09116.1

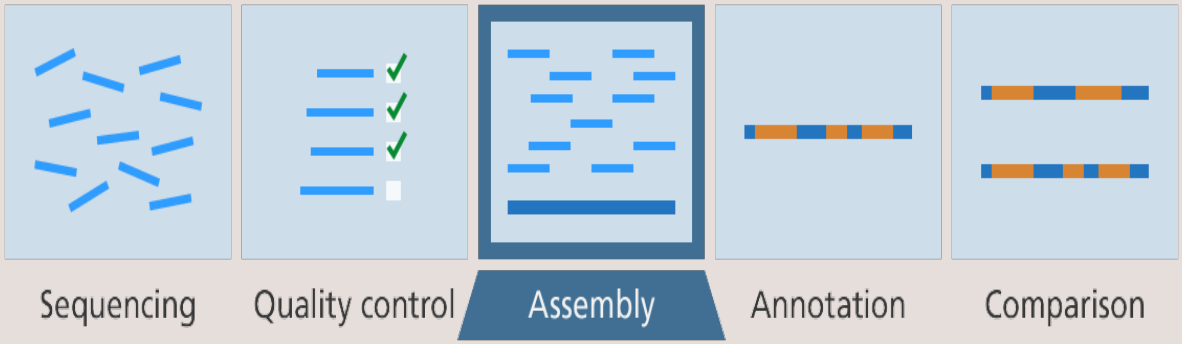
GenPept Identical Proteins Graphics

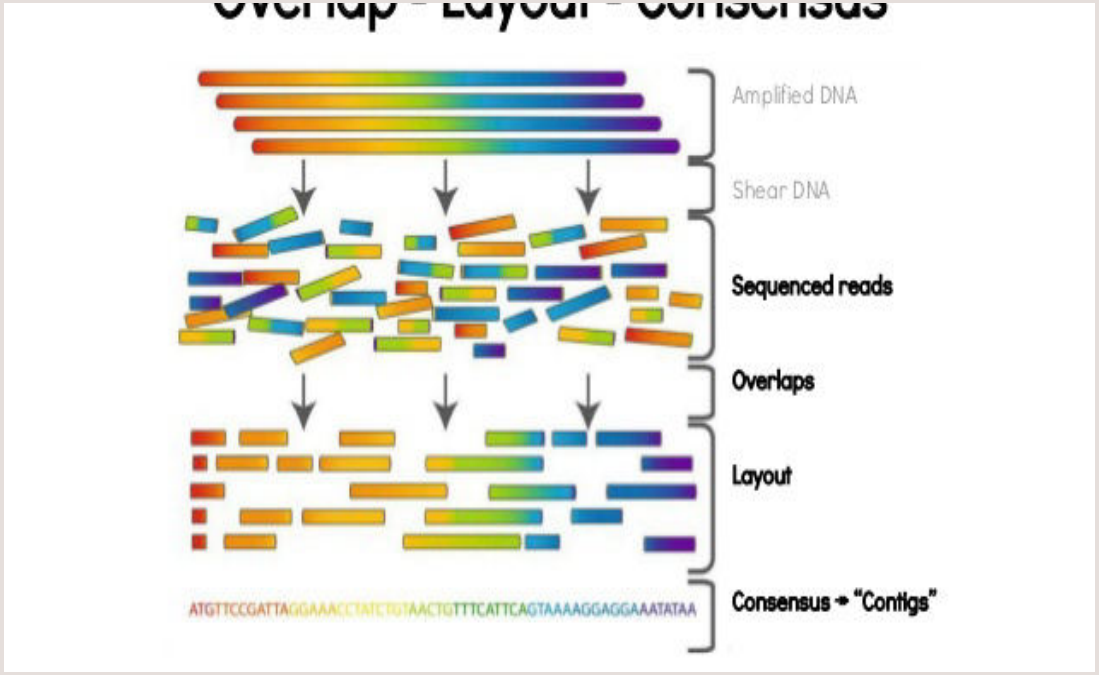
```
>APJ09116.1 chemotaxis protein CheR [Borreliella afzelii]
MNQNKFNLNINITKDELFRLIKIVYNNFGINLSEKKKMLIESRLSSLLKVKGFKNFTEYIDFLEKSAGNL
QLIELVDKISTNHTYFFRESKHFDFLNNKILPKLAEKILKSENSEIRIWSAGCSSGEEPYTIAMILKEYM
ENNKVNFKVKILATDISISVLHEAYEGIYPEDRTINLPKYLKTKYLNQLKDNKFQVKEILKKMVYFKKLN
LMDEKFPFSKKFDLIFCRNVMIYFDEKTRNDLANKFNYYLKNDSYLLIGHSETIRGNKNLKYIMPATYKK
N
```

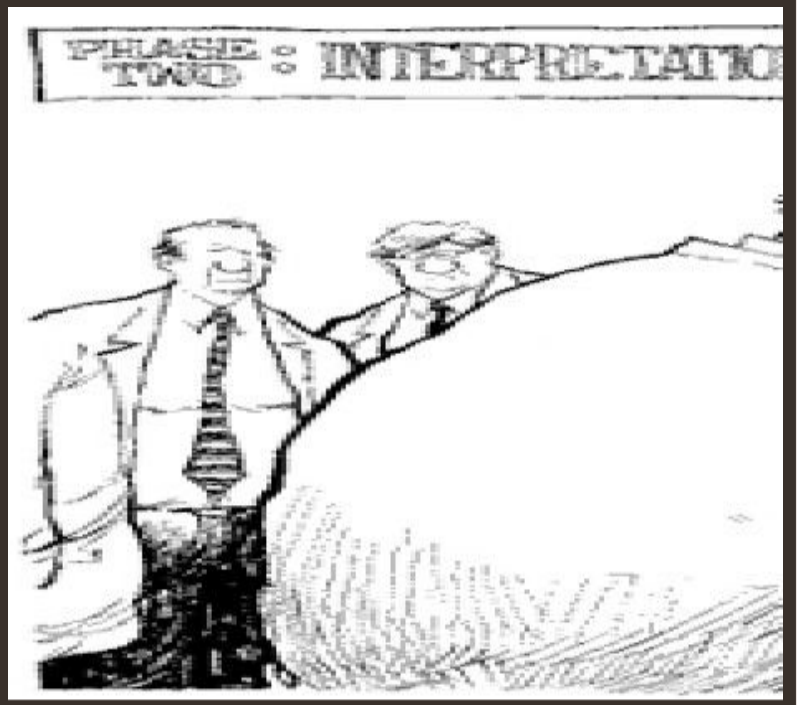
FASTQ FORMAT

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

ENSAMBLE DE GENOMAS







ENSAMBLE DE GENOMAS

Click to edit Master title style

- Ensamble *de novo* incluye dos partes:
  - El ensamble de la secuencia de nucleótidos
  - La anotación. La descripción de los elementos funcionales y estructurales dentro de la secuencia

Desventajas:

- Click to edit the outline text format
  - Second Outline Level
    - Third Outline Level
      - Fourth Outline Level
        - Fifth Outline Level
          - Sixth Outline Level
            - Seventh Outline Level Click to edit Master text styles
              - Second level
                - Third level
                  - Fourth level
                    - Fifth level
  - Se necesita mucho tiempo: para un genoma de eucarionte el ensamble puede tardar meses o incluso años. Un programa de ensamble o de anotación puede correr hasta por semanas.
  - Se necesita mucho dinero en la secuenciación
  - Se necesita mucho poder computacional

- La pregunta importante es: "¿Realmente necesito un genoma ensamblado y anotado para mi resolver pregunta de investigación?"

CHECKLIST PARA ANTES DE DISEÑAR TU PROYECTO DE ENSAMBLE DE GENOMA:

CLICK TO EDIT MASTER TITLE STYLE

- ❑ Para la extracción de ADN, seleccionar un individuo que sea un **buen representante** de la especie y que pueda **proveer suficiente ADN**
- ❑ **Extraer más ADN** del que creés necesitar.
- ❑ Recordar **extraer ARN y secuenciar ARN**, esto es muy funcional a la hora de ensamblar transcriptoma y para la anotación de funciones.
- ❑ **Decidir** de manera temprana **qué tecnología de secuenciación** se va a usar y que **herramientas de ensambles**. no quieres terminar con información que no puedes analizar con cualquier programa.

Click to edit the outline text format

- Second Outline Level
  - Third Outline Level
    - Fourth Outline Level
      - Fifth Outline Level
        - Sixth Outline Level
          - Seventh Outline Level

Click to edit Master text styles

- Second level
  - Third level
    - Fourth level
      - Fifth level

DISEÑO EXPERIMENTAL

NO.1 - PROPIEDADES DEL GENOMA

- Tamaño del Genoma
  - El número de lecturas *(reads)* depende del tamaño del genoma. El mínimo 60x *sequence depth*.
- Repeticiones
  - La cantidad y distribución de las repeticiones modifican la calidad del ensamble. Entre más repeticiones es mejor usar lecturas largas.
- Heterocigosidad
  - Los programas de ensamble colapsan alelos, si hay alta heterocigosidad las regiones homologas serán detectadas como regiones diferentes.
- Niveles de ploidia
  - Mejor usar tejidos haploides
- Contenido de GC
  - Niveles bajos o altos de GC en ciertas regiones causan problemas en ciertos secuenciadores como Ilumina.

Sequoia sempervirens

11 cromosomas
Hexaploide

Opuntia ficus-indica

11 cromosomas
Octoploide

DISEÑO EXPERIMENTAL

NO.1 - PROPIEDADES DEL GENOMA

- Tamaño del Genoma
  - El número de lecturas *(reads)* depende del tamaño del genoma. Ilumina 60x *sequence depth*.
- Repeticiones
  - La cantidad y distribución de las repeticiones modifican la calidad del ensamble. Entre más repeticiones es mejor usar lecturas largas.
- Heterocigosidad
  - Los programas de ensamble colapsan alelos, si hay alta heterocigosidad las regiones homologas serán detectadas como regiones diferentes.
- Niveles de ploidia
  - Mejor usar tejidos haploides
- Contenido de GC
  - Niveles bajos o altos de GC en ciertas regiones causan problemas en ciertos secuenciadores como Ilumina.

OTRAS COSAS A TOMAR EN CUENTA...

CLICK TO EDIT MASTER TITLE STYLE

- No intentar hacer *pool* de individuos.
  - Esto causa heterozigosidad en las secuencias.
- Whole Genome Amplification.
  - Se pueden crear secuencias quiméricas al fusionar secuencias que no están relacionadas.
- Presencia de otros organismos.
  - Puede haber contaminación de otros organismos que se introducen en el laboratorio, ADN humano, etc.
- ADN de organelos.
  - En ciertas células hay una mayor cantidad de ADN mitocondria o de cloroplastos que el ADN nuclear.

Click to edit the outline text format
Second Outline Level
Third Outline Level
Fourth Outline Level
Fifth Outline Level
Sixth Outline Level
Seventh Outline LevelClick to edit Master text styles
Second level
Third level
Fourth level
Fifth level

DISEÑO EXPERIMENTAL

NO.1 - PROPIEDADES DEL GENOMA

- Tamaño del Genoma
  - El número de lecturas *(reads)* depende del tamaño del genoma. Illumina 60x *sequence depth*.
- Repeticiones
  - La cantidad y distribución de las repeticiones modifican la calidad del ensamble. Entre más repeticiones es mejor usar lecturas largas.
- Heterocigosidad
  - Los programas de ensamble colapsan alelos, si hay alta heterocigosidad las regiones homologas serán detectadas como regiones diferentes.
- Niveles de ploidia
  - Mejor usar tejidos haploides
- Contenido de GC
  - Niveles bajos o altos de GC en ciertas regiones causan problemas en ciertos secuenciadores como Ilumina.

DISEÑO EXPERIMENTAL

NO.3 – TECNOLOGÍAS DE SECUENCIACIÓN

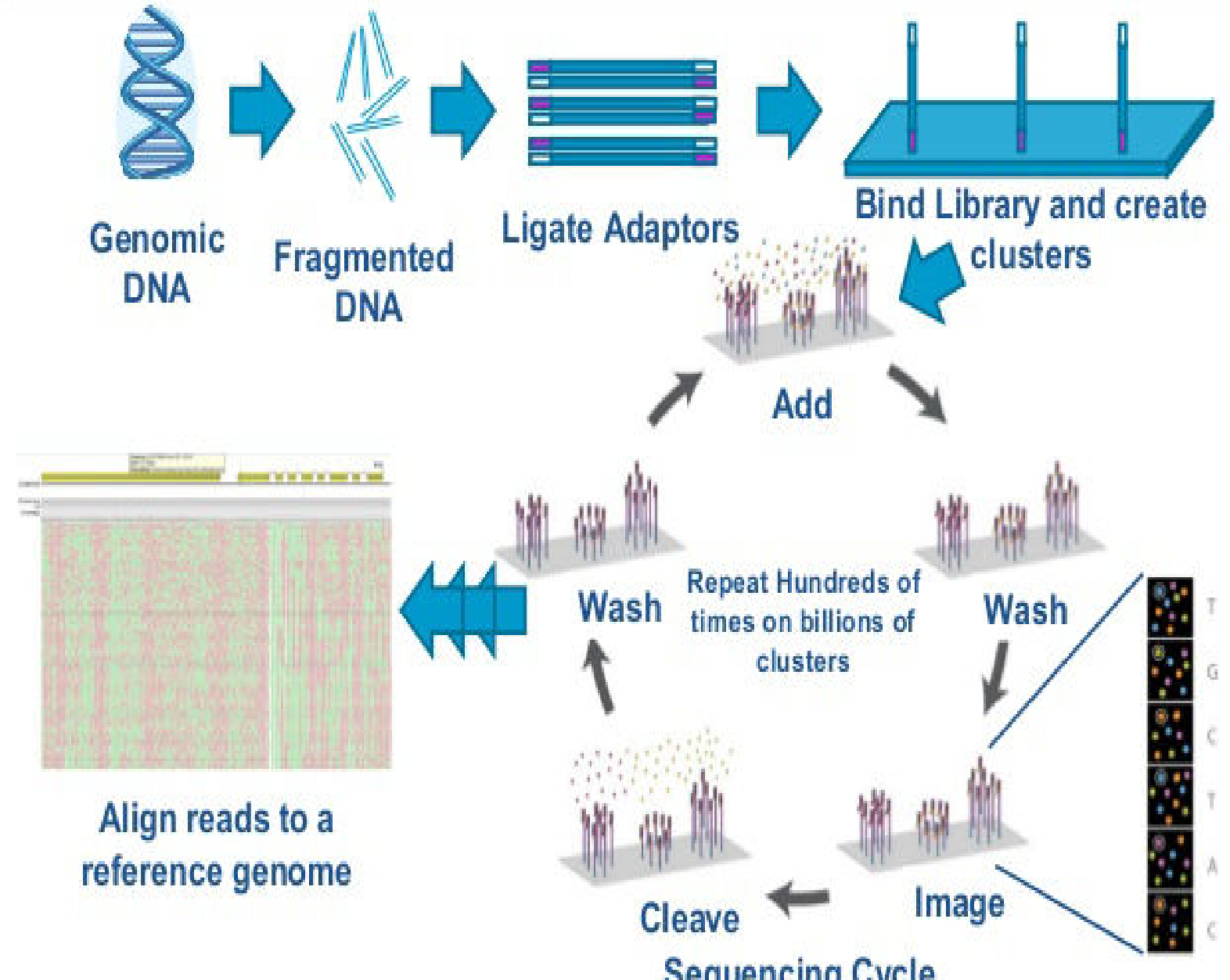
SEVENTH OUTLINE LEVELCLICK TO EDIT MASTER TEXT STYLE

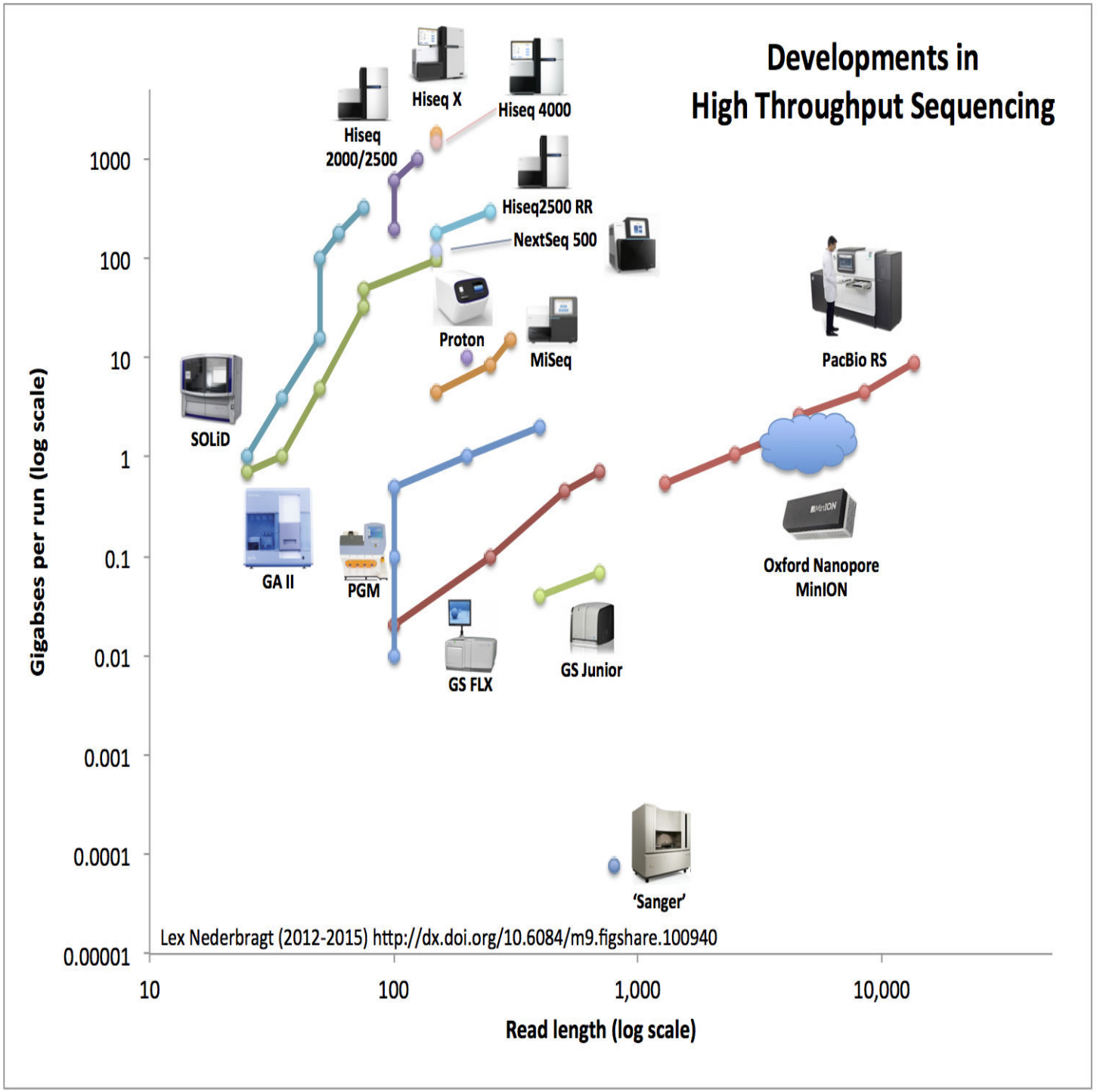
- Illumina y ion Torrent
  - Click to edit Master text styles
    - Second Outline Level
      - Third Outline Level
        - Fourth Outline Level
          - Fifth Outline Level
            - Sixth Outline Level
- Son más baratos
- Mayor cobertura
- Problemas con regiones de repetidas y con contenido de GC
- Lecturas de ~36 bp hasta ~500 bp
- Seventh Outline LevelClick to edit Master text styles
  - Second level
    - Third level
      - Fourth level
        - Fifth level
          - Sixth level
            - Seventh Outline LevelClick to edit Master text styles

SEVENTH OUTLINE LEVELCLICK TO EDIT MASTER TEXT STYLE

- PacBio y Nanopore
  - Click to edit Master text styles
    - Second Outline Level
      - Third Outline Level
        - Fourth Outline Level
          - Fifth Outline Level
            - Sixth Outline Level
- Son más caros
- Lecturas de mayor tamaño 10,00 bp hasta 100,000 bp
- Ensamblan mejor zonas repetidas (genomas de eucariontes)
- No son accesibles en todos los países
- Mayores requerimientos de la calidad del ADN
- Seventh Outline LevelClick to edit Master text styles
  - Second level
    - Third level
      - Fourth level
        - Fifth level
          - Sixth level
            - Seventh Outline LevelClick to edit Master text styles

2nd Generation Sequencing Overview




Developments in
High Throughput Sequencing
Hiseq X
Hiseq 4000
Hiseq
2000/2500
Hiseq2500 RR
NextSeq 500
Proton
MiSeq
PacBio RS
SOLiD
GA II
PGM
GS FLX
GS Junior
Oxford Nanopore
MinION
'Sanger'
Gigabses per run (log scale)
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1,000
10,000
Read length (log scale)
Lex Nederbragt (2012-2015) http://dx.doi.org/10.6084/m9.figshare.100940

DISEÑO EXPERIMENTAL

NO.1 - PROPIEDADES DEL GENOMA

- Tamaño del Genoma
  - El número de lecturas *(reads)* depende del tamaño del genoma. Illumina 60x *sequence depth*.
- Repeticiones
  - La cantidad y distribución de las repeticiones modifican la calidad del ensamble. Entre más repeticiones es mejor usar lecturas largas.
- Heterocigosidad
  - Los programas de ensamble colapsan alelos, si hay alta heterocigosidad las regiones homologas serán detectadas como regiones diferentes.
- Niveles de ploidia
  - Mejor usar tejidos haploides
- Contenido de GC
  - Niveles bajos o altos de GC en ciertas regiones causan problemas en ciertos secuenciadores como Ilumina.

EJEMPLO DE CLICK TO EDIT MASTER TITLE STYLE

MISIÓN: formar secuencias lo más largas posible con el menor número de gaps.

- Second Outline Level
  - Third Outline Level
    - Fourth Outline Level
      - Fifth Outline Level
        - Sixth Outline Level
- Seventh Outline LevelClick to edit Master text styles
  - Second level
    - Third level
      - Fourth level
        - Fifth level

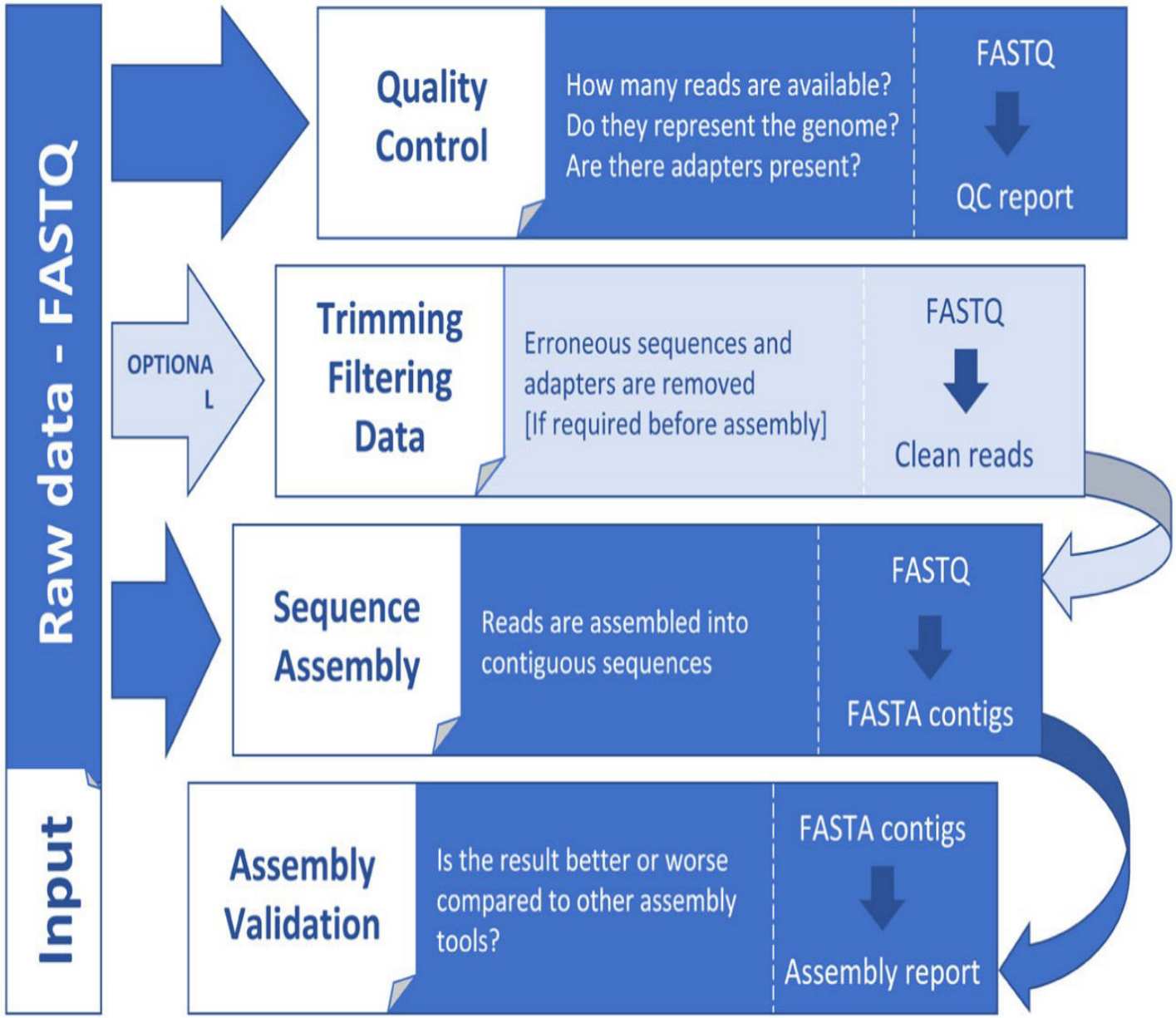


Figure 2. General steps in a genome assembly workflow. Input and output data are indicated for each step.

ENSAMBLE

1. CONTROLES DE CALIDAD

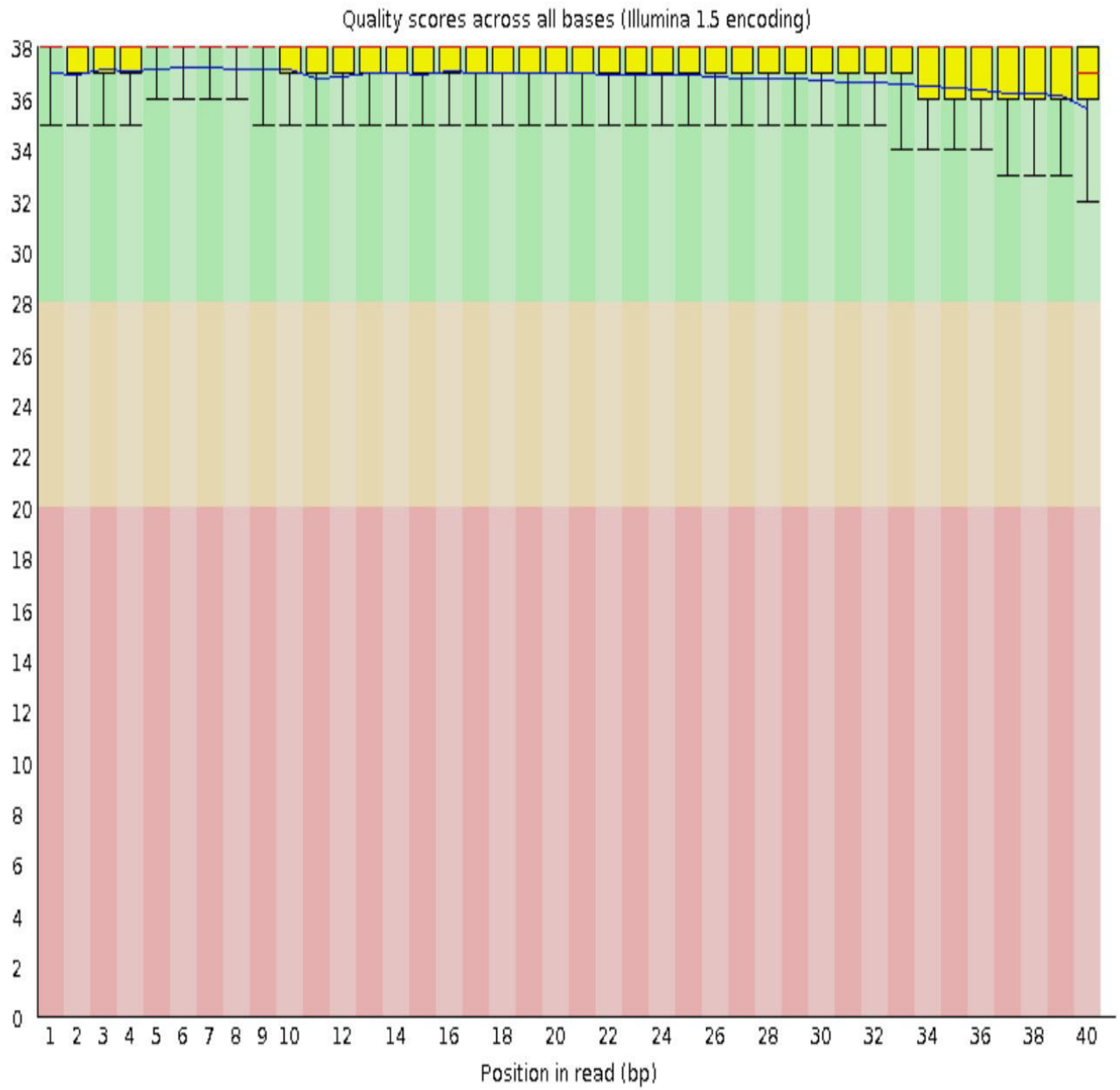
- Basic calling accuracy
- Trimming: eliminar adaptadores y eliminar zonas con poca calidad
- FASTQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FastQC Report

Summary

- 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

Per base sequence quality

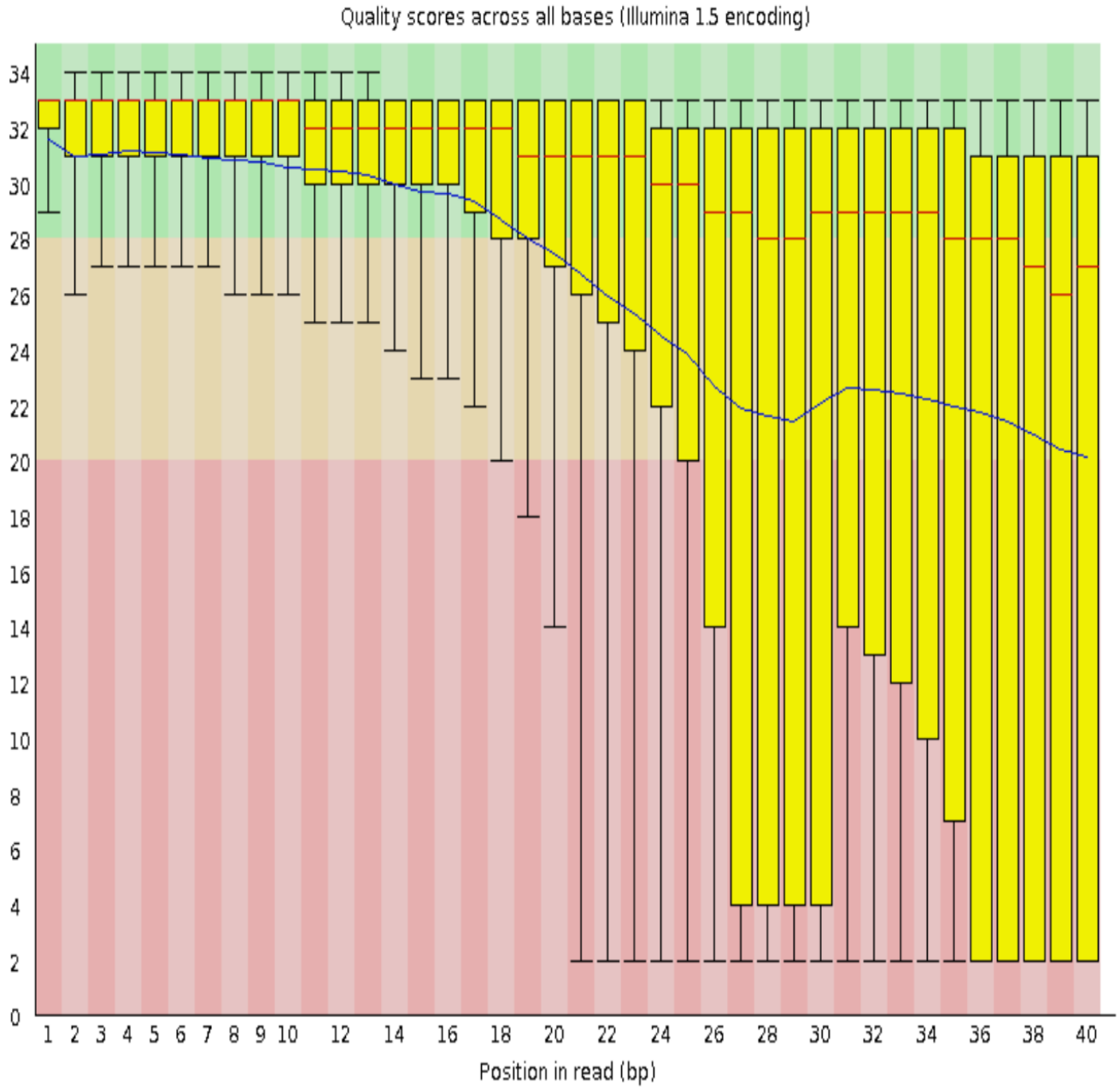


FastQC Report

Summary

- 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

Per base sequence quality

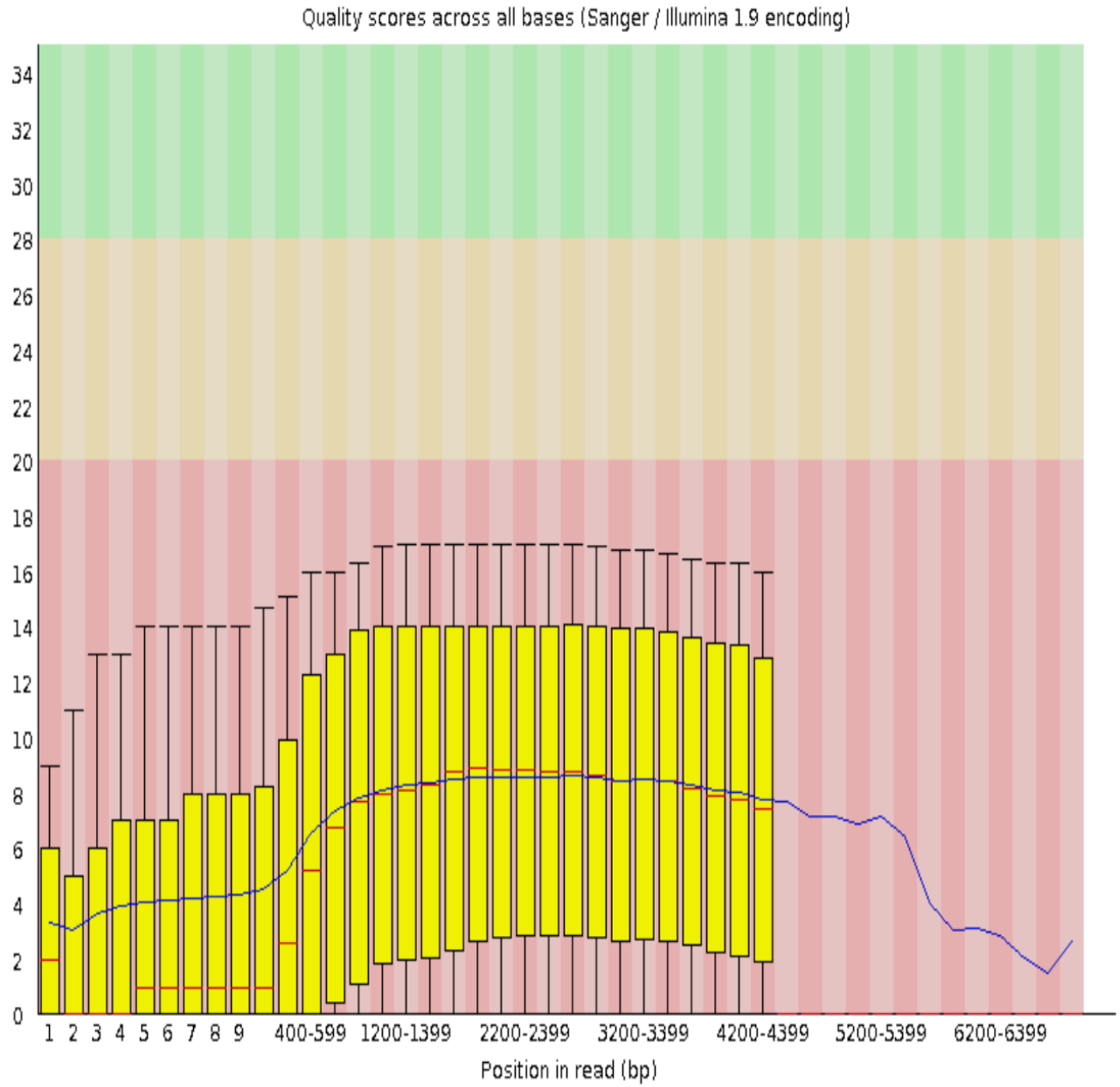


FastQC Report

Summary

- Basic Statistics
- Per base 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

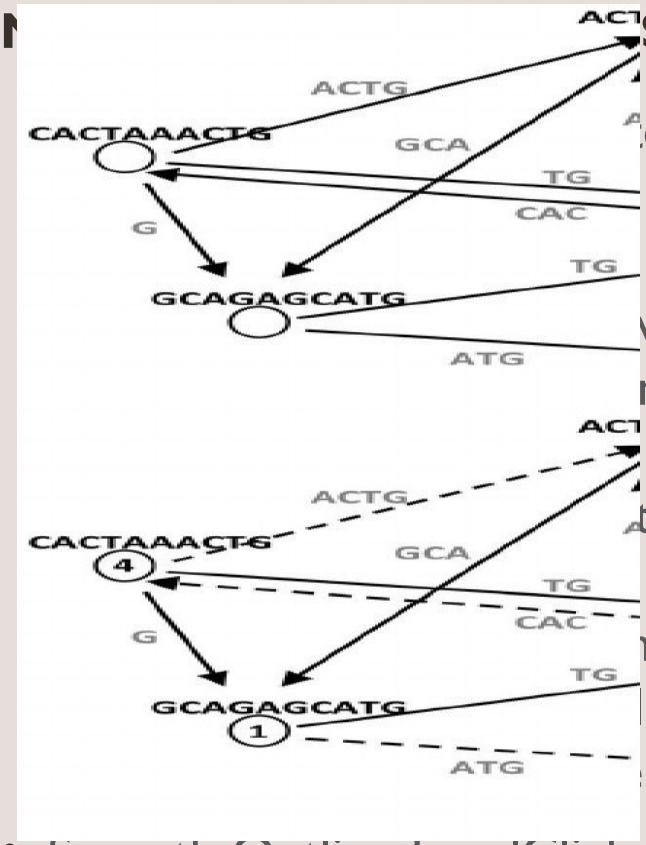
Per base sequence quality



CLICK TO EDIT MASTER TITLE STYLE

2. ENSAMBLE DE LA SECUENCIA

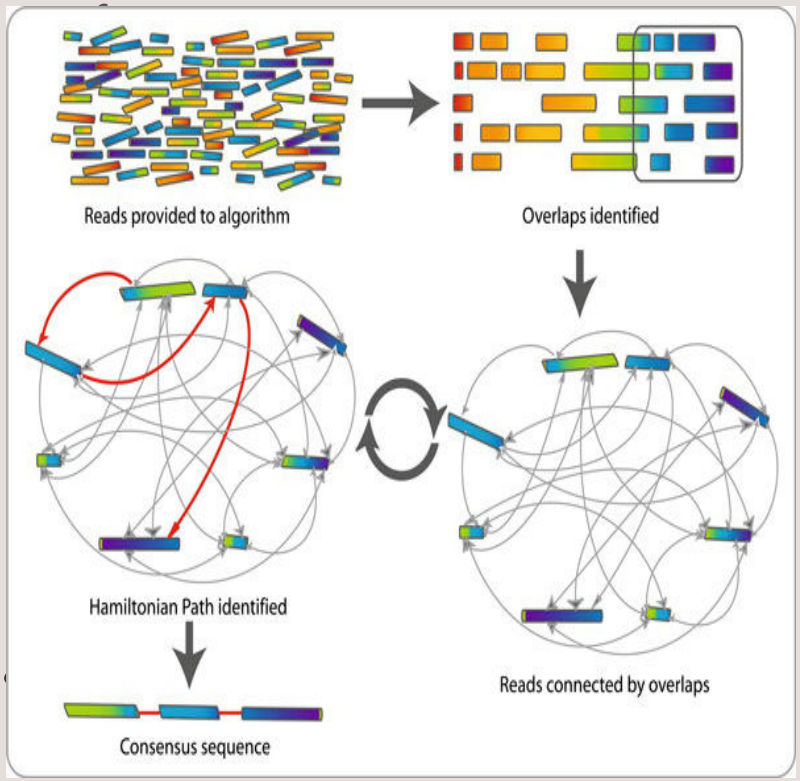
- Click to edit Master text styles
  - Second Outline Level
    - Third Outline Level
      - Fourth Outline Level
        - Fifth Outline Level
          - Sixth Outline Level
- Seventh Outline LevelClick to edit Master text styles



2. ENSAMBLE DE LA SECUENCIA

- Click to edit the outline text format
  - Second Outline Level
    - Third Outline Level
      - Fourth Outline Level
        - Fifth Outline Level
          - Sixth Outline Level

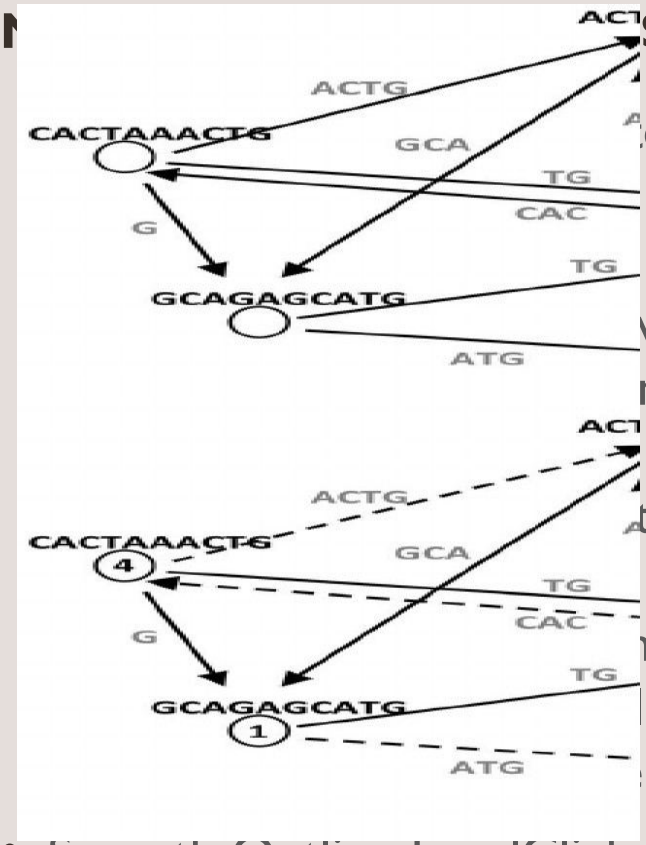
- Seventh Outline LevelClick to edit Master text styles

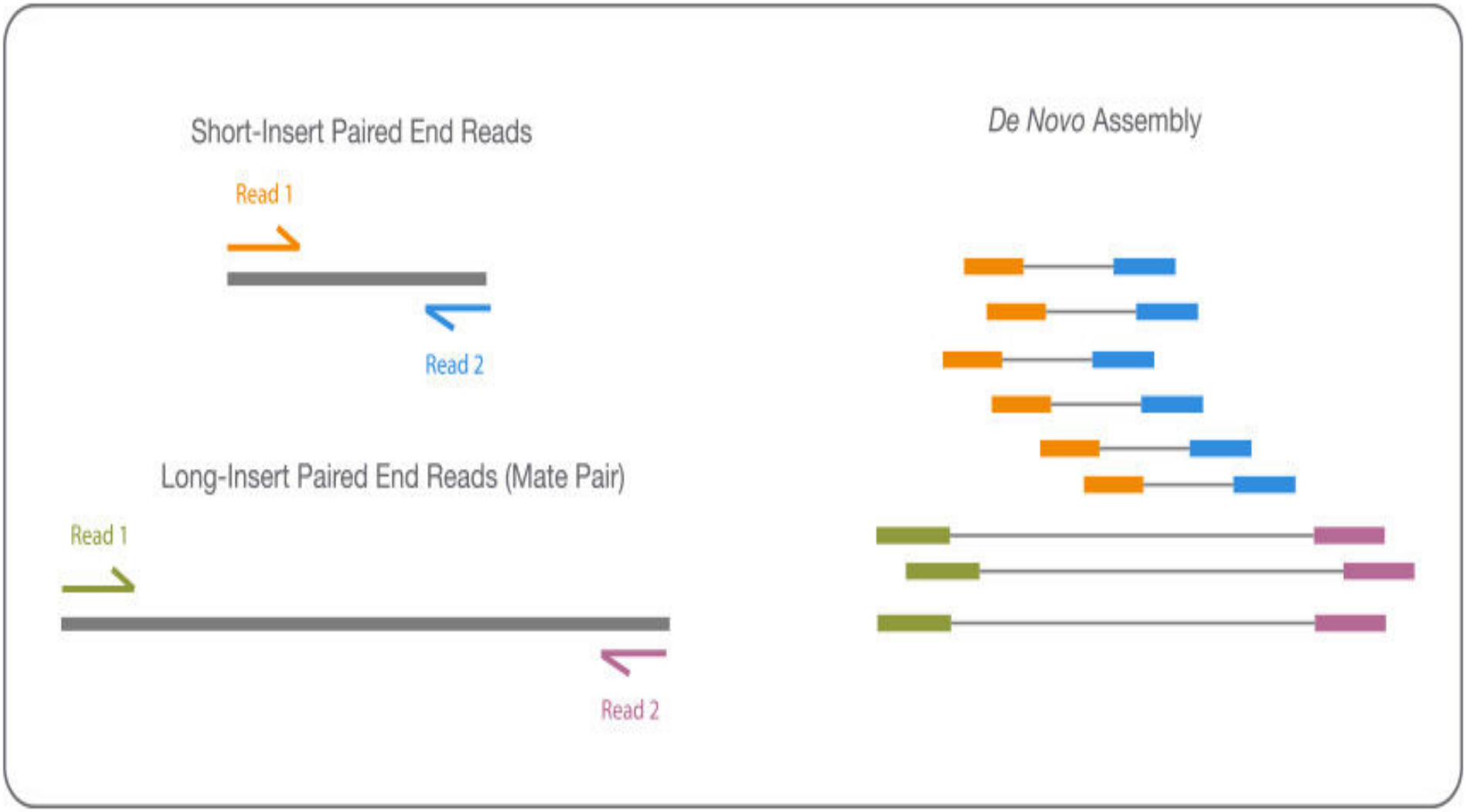


CLICK TO EDIT MASTER TITLE STYLE

2. ENSAMBLE DE LA SECUENCIA

- Click to edit Master text styles
  - Second Outline Level
    - Third Outline Level
      - Fourth Outline Level
        - Fifth Outline Level
          - Sixth Outline Level
- Seventh Outline LevelClick to edit Master text styles




Short-Insert Paired End Reads
Read 1
Read 2
Long-Insert Paired End Reads (Mate Pair)
Read 1
Read 2
De Novo Assembly

CLICK TO EDIT MASTER TITLE STYLE

3. DETERMINAR QUE ESTA LISTO - CALIDAD

- Click to edit the outline text format
  - Second Outline Level
    - Third Outline Level
      - Fourth Outline Level
        - Fifth Outline Level
          - Sixth Outline Level
            - Seventh Outline LevelClick to edit Master text styles
  - Métrica estándar. Entre más grande mejor.
  - Es el tamaño del scaffold o contig más pequeño que, sumano con los más largos que él, deben cubrir el 50% del genoma
  - NO IMPLICA QUE ESTÉ BIEN ENSAMBLADO
- Usar varios ensambladores y compararlos, correr el mismo ensamblador con distintos parámetros
  - Second level
  - Quast para comparar
    - Third level
      - Fourth level
        - Fifth Level
- Determinar cuantos genes de proteínas tiene el ensamble
  - herramienta BUSCO

1. ANOTACIÓN ESTRUCTURAL

- ¿Dónde están los genes y a qué se parecen? Sobre todo genes que codifican para proteínas

INTRÍNSECO

- Se centra en la información extraída de la secuencia
- Es intesivo ya que se necesitan crear modelos y los software tienen que entrenarse y optimizarse.
- Son específicos para cada genoma

EXTRÍNSECO

- Utiliza información adicional a la secuencia como lo son:
  - Zonas que codifican para proteínas
  - Sitios de splicing
  - Transcritos
- Es universal
- Se utilizan bases de datos, usualmente de polipéptidos
- No da información estructural del gen

1. ANOTACIÓN ESTRUCTURAL

- ¿Dónde están los genes y a qué se parecen? Sobre todo genes que codifican para proteínas

INTRÍNSECO

- Se centra en la información extraída de la secuencia
- Es intesivo ya que se necesitan crear modelos y los software tienen que entrenarse y optimizarse.
- Son específicos para cada genoma

EXTRÍNSECO

- Utiliza información adicional a la secuencia como lo son:
  - Zonas que codifican para proteínas
  - Sitios de splicing
  - Transcritos
- Es universal
- Se utilizan bases de datos, usualmente de polipéptidos
- No da información estructural del gen

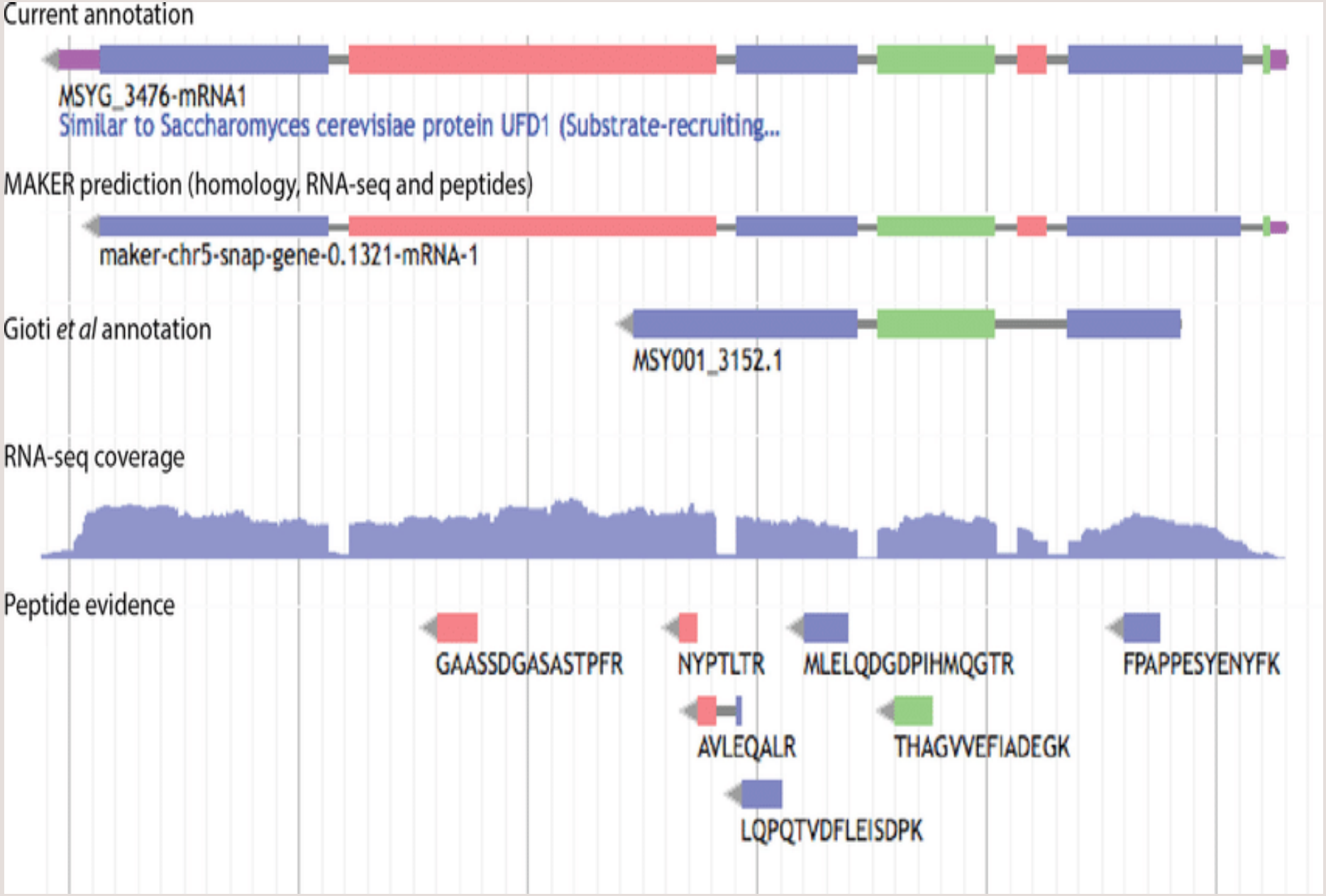
CLICK TO EDIT MASTER TITLE STYLE

1. ANOTACIÓN FUNCIONAL

- Asignar infromación biológica relevante como predicción de polipeptidos y los aspectos de los que se derivan
- Dos objetivos principales
  - Asignar elementos funcionales a genes
  - Hacer una revisión de calidad para los genes precedidos

Click to edit the outline text format

- Second Outline Level
  - Third Outline Level
    - Fourth Outline Level
      - Fifth Outline Level
        - Sixth Outline Level
          - Seventh Outline LevelClick to edit Master text styles
            - Second level
              - Third level
                - Fourth level
                  - Fifth level


Current annotation
MSYG_3476-mRNA1
Similar to Saccharomyces cerevisiae protein UFD1 (Substrate-recruiting...
MAKER prediction (homology, RNA-seq and peptides)
maker-chr5-snap-gene-0.1321-mRNA-1
Gioti et al annotation
MSY001_3152.1
RNA-seq coverage
Peptide evidence
GAASSDGASASTPFR
NYPTLTR
MLELQDGDPIHMQGTR
FPAPPESYENYFK
AVLEQALR
THAGVVEFIADEGK
LQPQTVDFLEISDPK