Bioinformática-introducción

Paula Alexandra Torres Quintero

Bioinformática

Semestre 2026-1

Biología

Universidad del Rosario

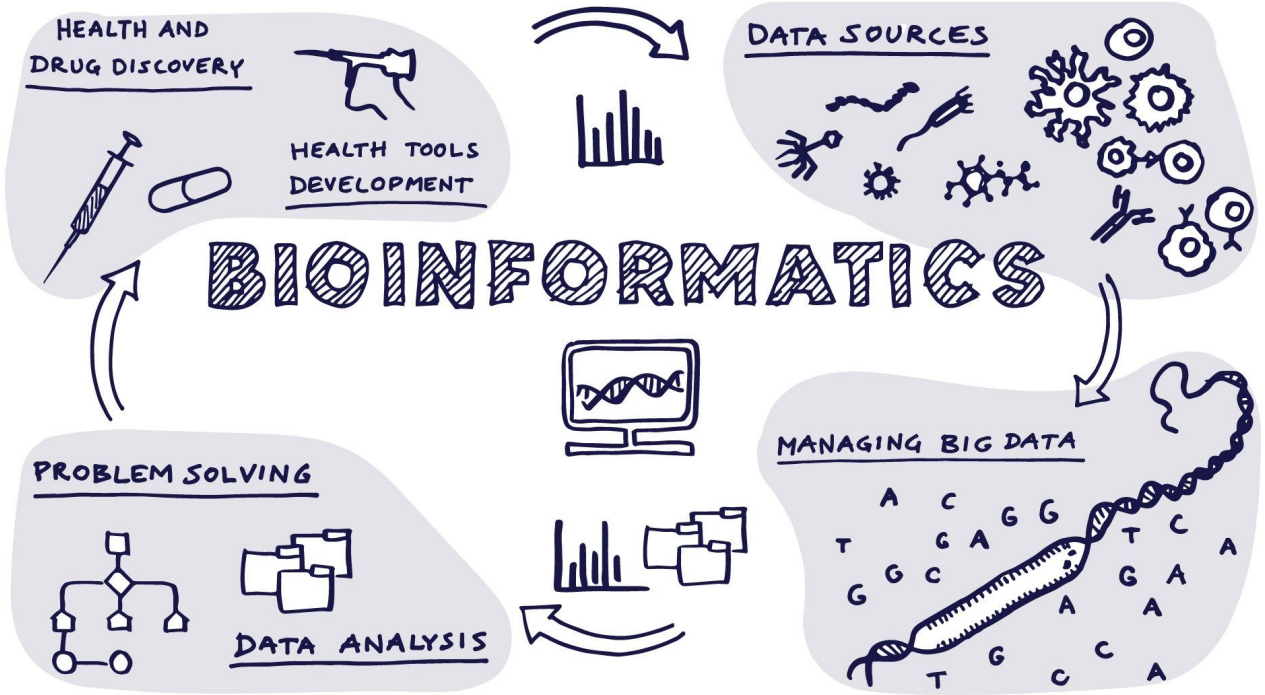
¿Qué es bioinformática?

Universidad del Rosario

Biología

¿Qué es bioinformática?

"Bioinformatics is a field that combines biology and computer science to study biological information. It involves using computers to analyze and interpret large amounts of data related to genetics, DNA, and proteins to understand how living organisms work."



Universidad del Rosario

Biología

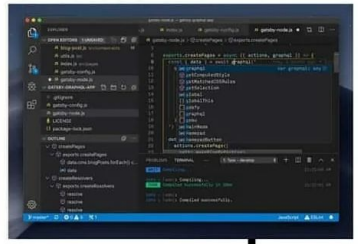
Biologists starter pack

De bata

De Bota

De BigData

macbook



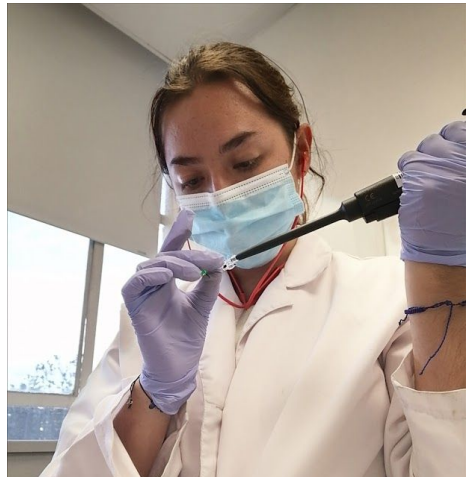
vs code

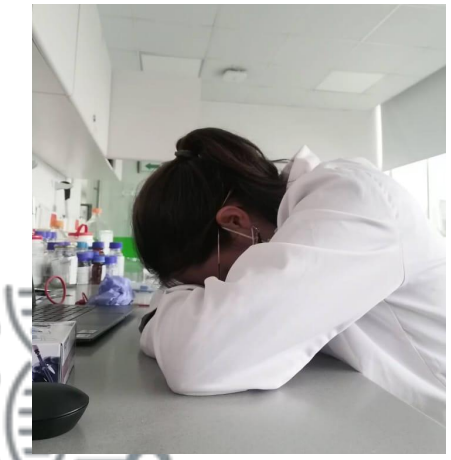
glasses must

False promises

Biologists starter pack

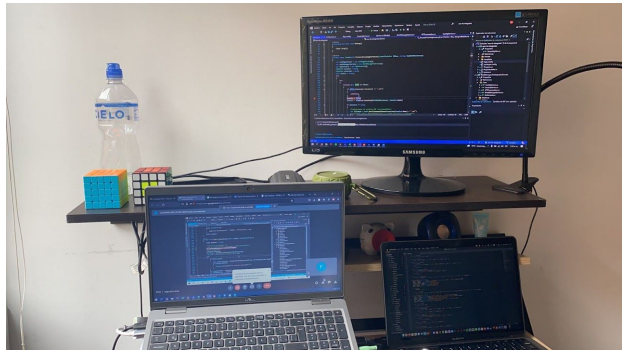
De bata

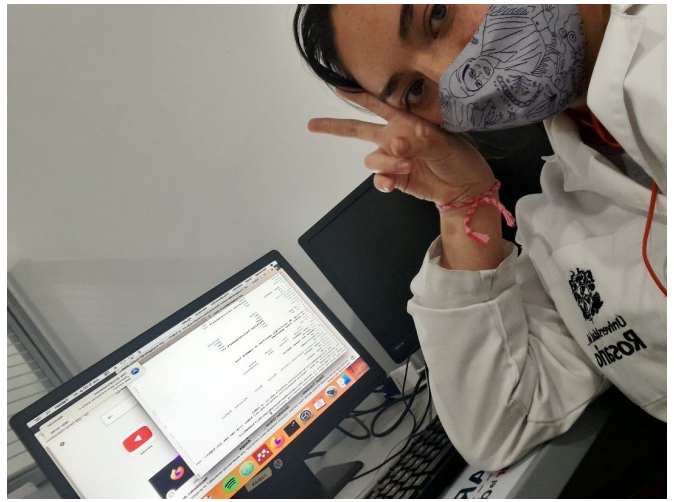




De Bota

De BigData

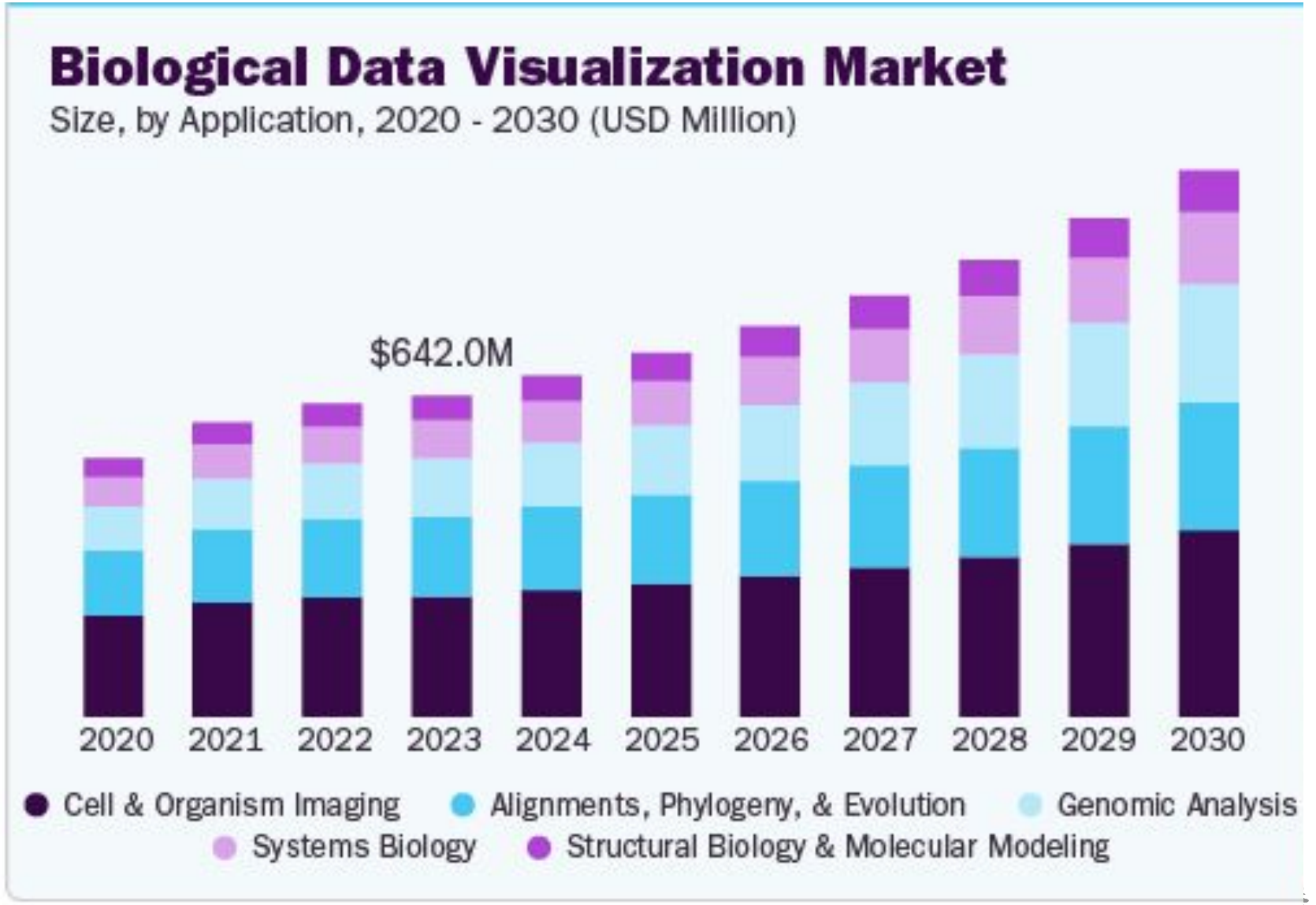




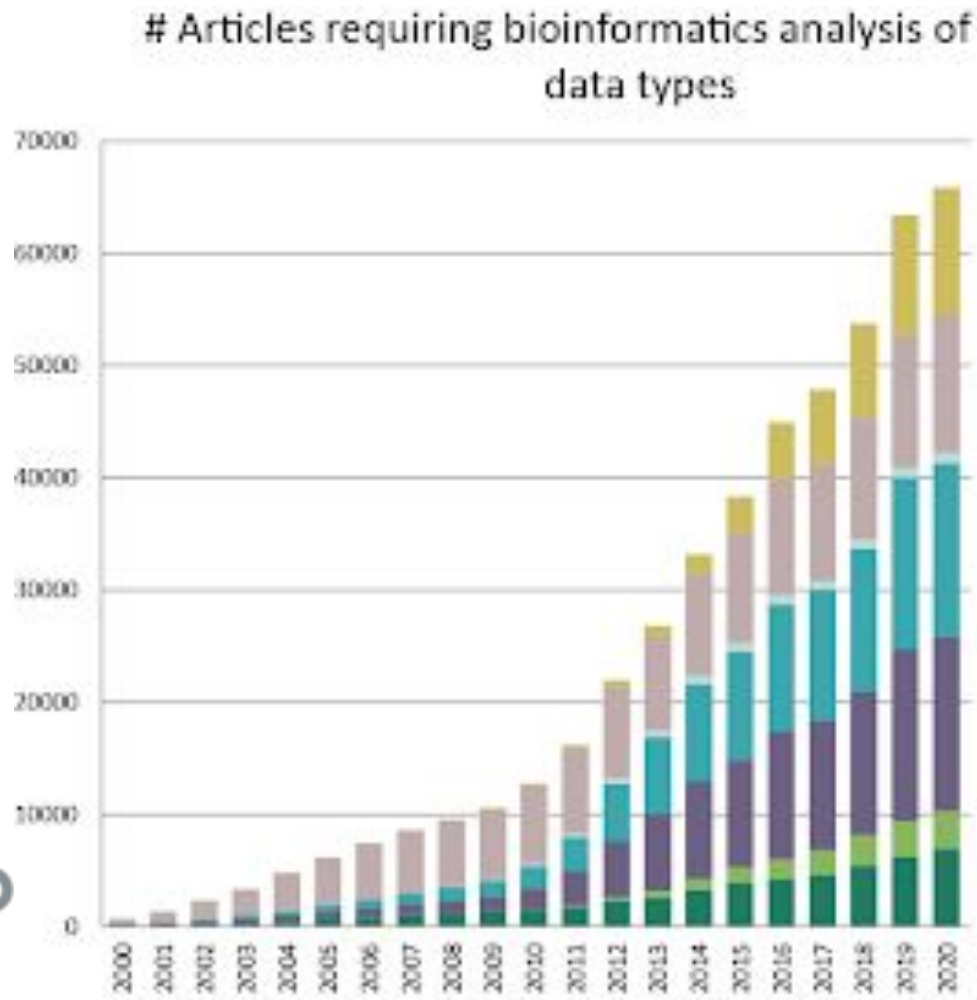
Universidad del Rosario

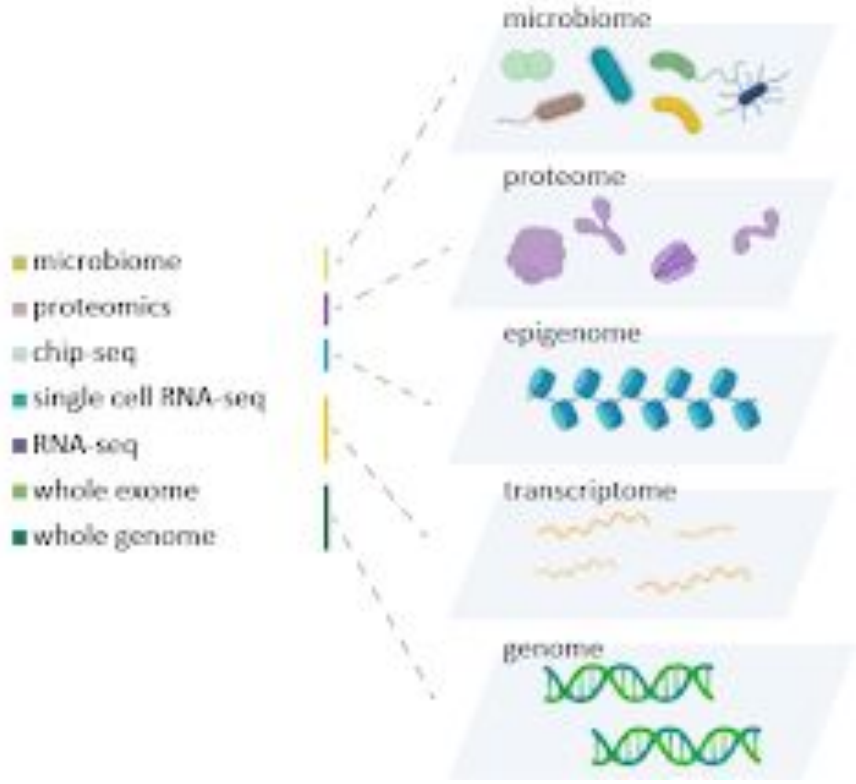
Biología

¿Por qué Bioinformática?



¿Por qué Bioinformática?

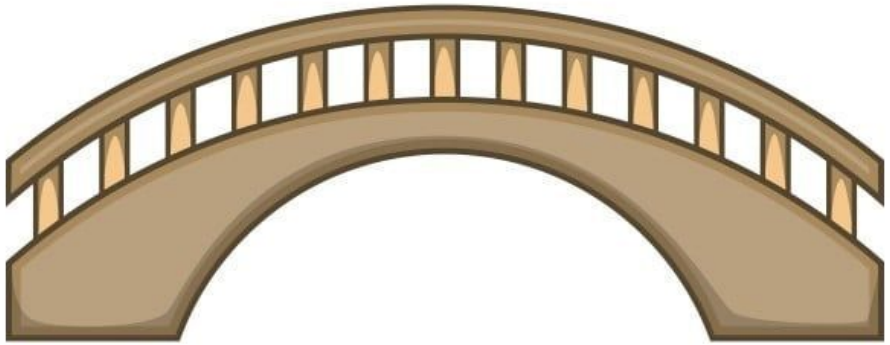




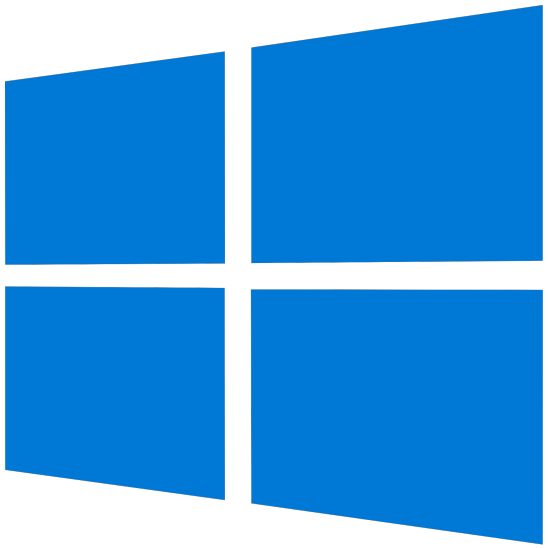
Universidad del Rosario

Biología

¿Qué es un sistema operativo?



¿Qué es un sistema operativo?



¿Por qué algunos biólogos no usan Windows?

Universidad del Rosario

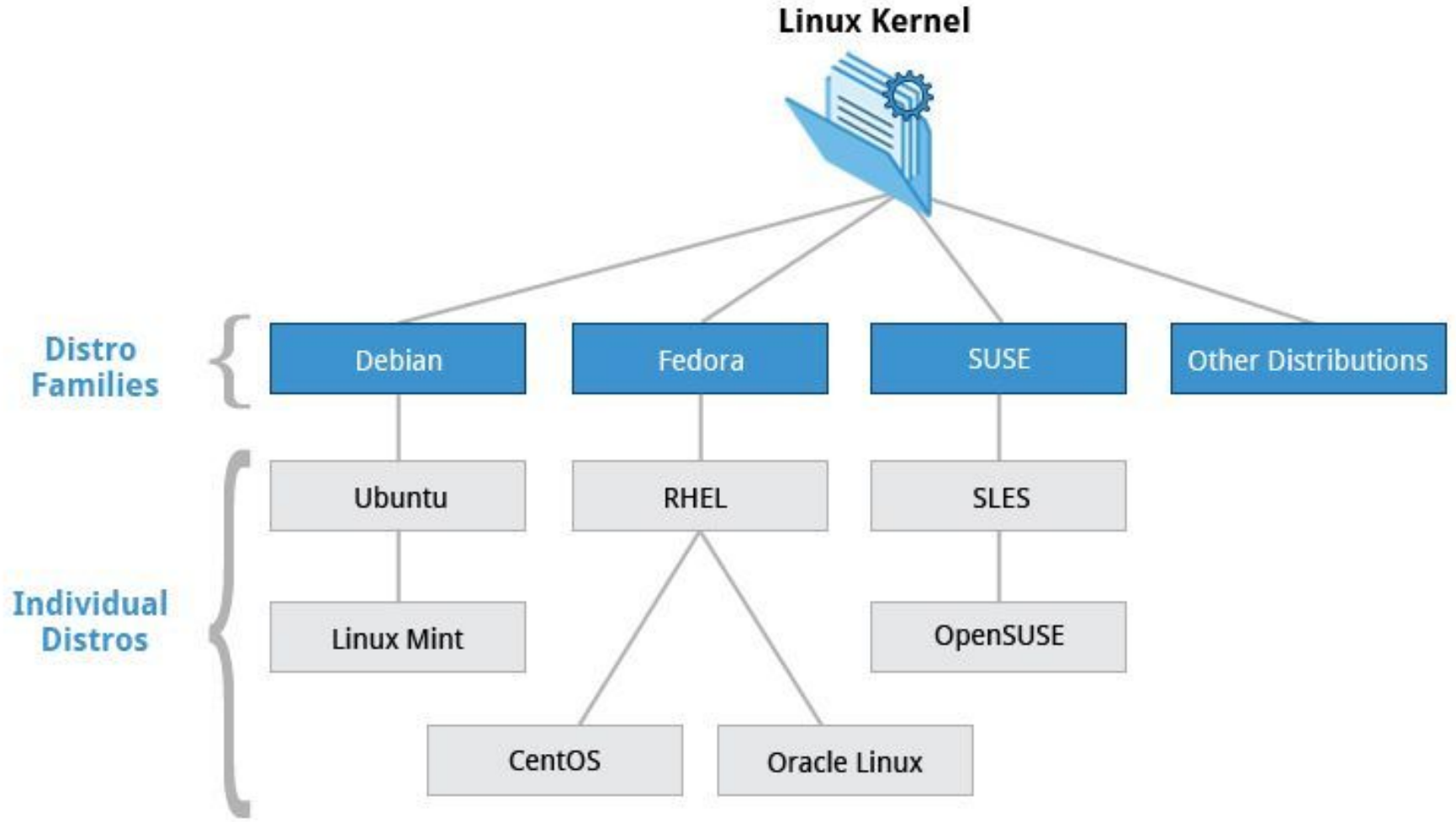
Biología

¿Por qué algunos biólogos no usan Windows?

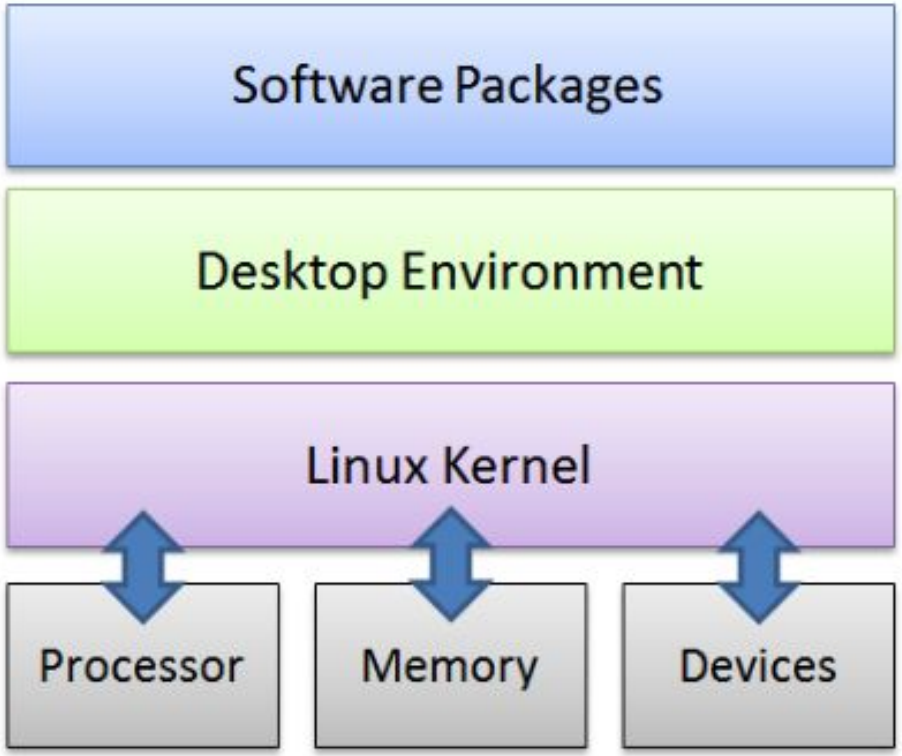
Universidad del Rosario

Biología

Distribuciones de Linux



Distribuciones de Linux



What the system can do

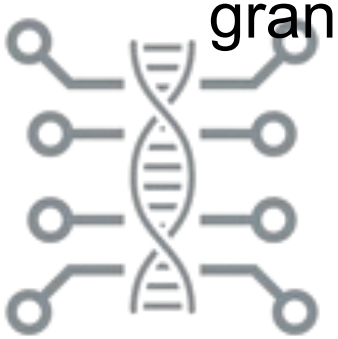
What the system shows

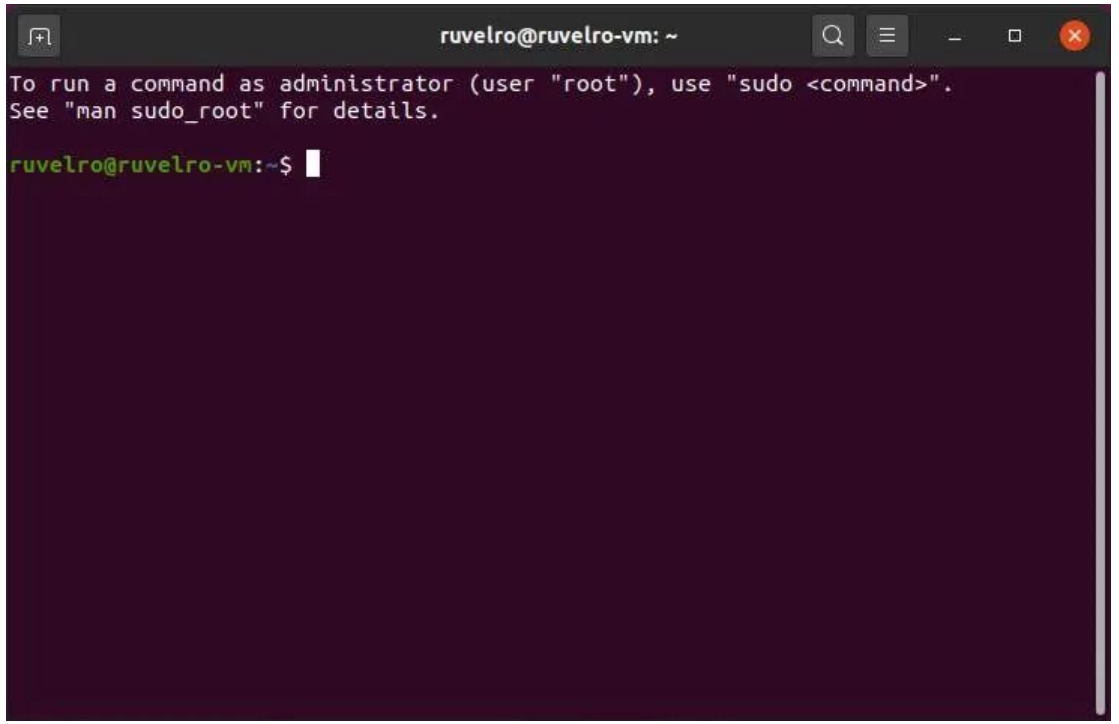
What the system controls

¿Por qué algunos biólogos no usan Windows?

La mejor parte!!

- ¡Es gratuito!
- Código abierto
- Hay mucho software en biología escrito
- Permite organizarse muy fácil
- Facilita el procesamiento de datos grandes.





Universidad del Rosario

Biología

Para esta clase necesitamos

- Un computador
- Un editor de texto
- Y un sistema operativo Unix

¿Cómo estamos de términos?

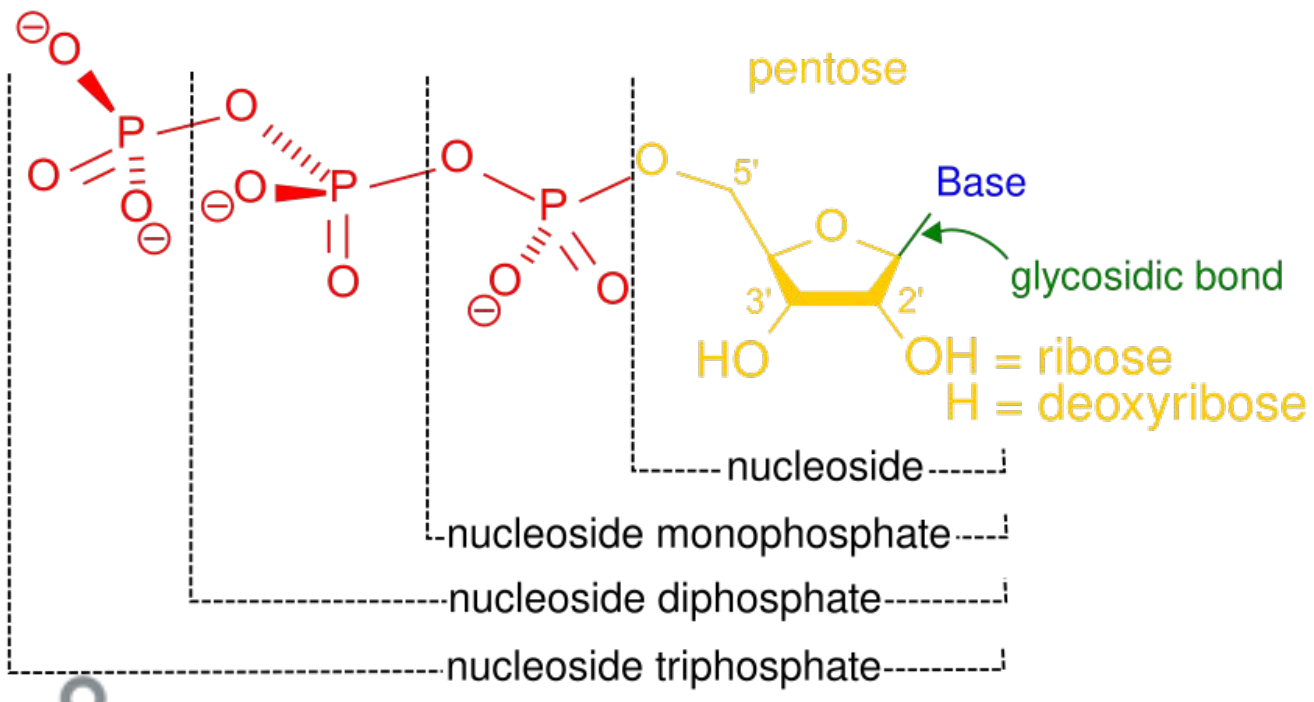
- Base/nucleotide
- Read
- Contig
- Scaffold
- Chromosome

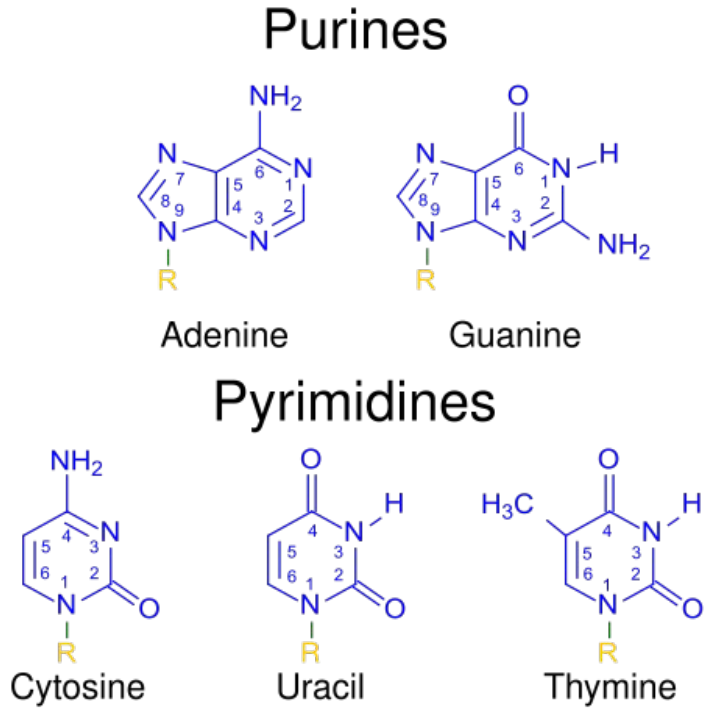
- FASTA
- FASTQ
- GFF
- GTF
- VCF
- SAM
- BAM

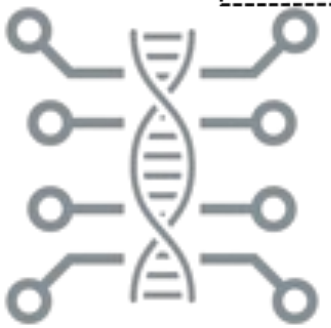
Universidad del Rosario

Biología

Base/nucleotide







Unidades estructurales y funcionales básicas de los ácidos nucleicos

Universidad del Rosario

Biología

Base/nucleotide

TTAACCTTGGTTTTGAACTTGAACACTTAGGGGATTGAAGATTCAACAACCCTAAAGCTTGGGGTAAAAC

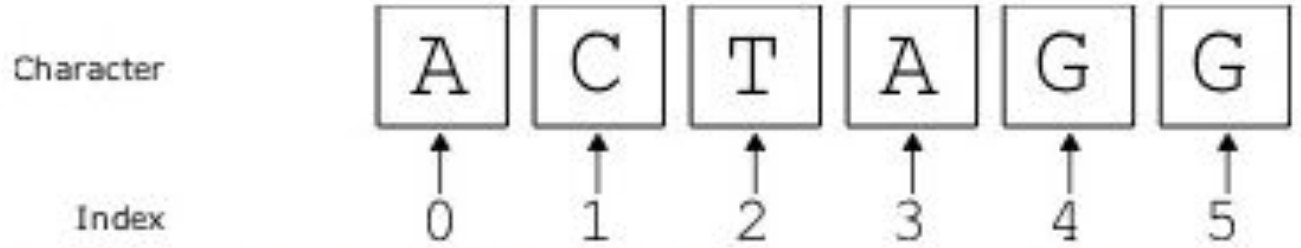


Figure 1.1: A simple representation of a character string. Each position of the string can be accessed by its index number, which goes from 0 to the length minus one.

Universidad del Rosario

Biología

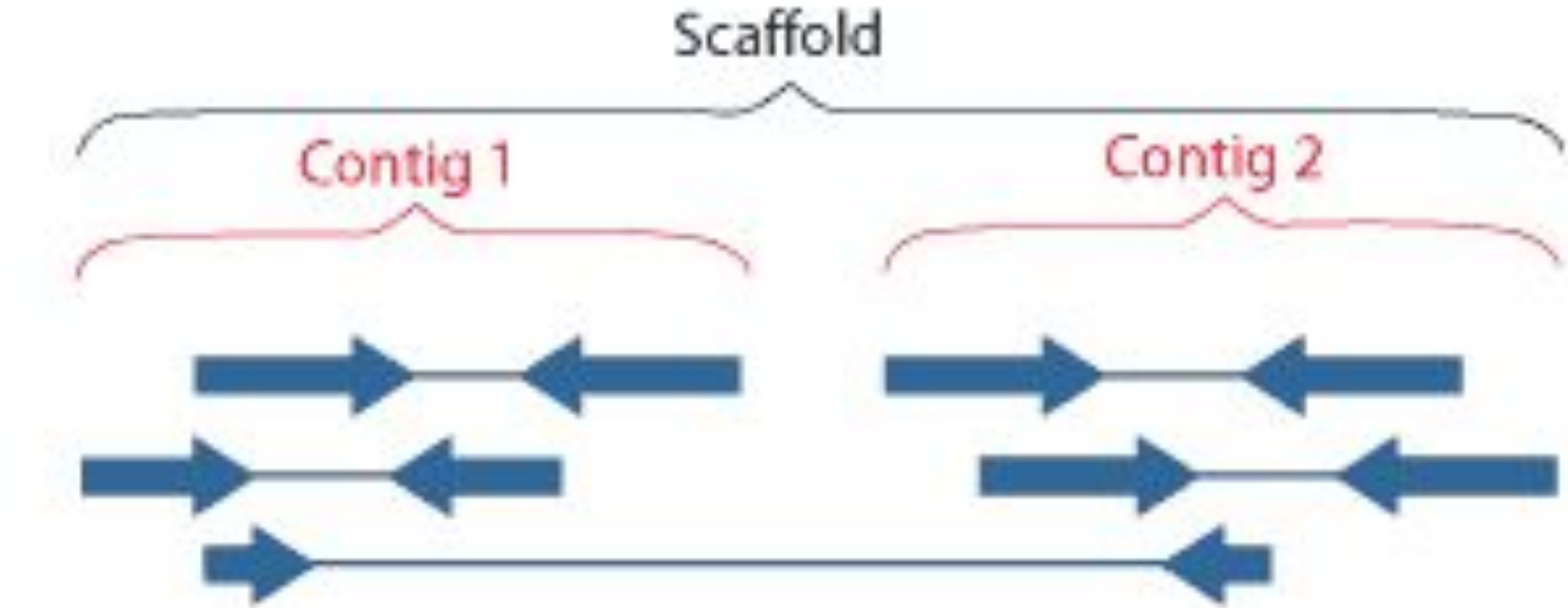
Contig

```
Read    TTAACCTTGGTTTTGAACTTGAACACTTAGGGGATTG
Read          TGGTTTTGAACTTGAACACTTAGGGGATTGAAGATTCAACAA
Read                AACTTGAACACTTAGGGGATTGAAGATTCAACAACCCTAAAGCT
Read                     CACTTAGGGGATTGAAGATTCAACAACCCTAAAGCTTGGG
Read                           ATTGAAGATTCAACAACCCTAAAGCTTGGGGT
Read                                 ACCCTAAAGCTTGGGGTA
Read                                       AGCTTGGGGTAAAAC
Contig TTAACCTTGGTTTTGAACTTGAACACTTAGGGGATTGAAGATTCAACAACCCTAAAGCTTGGGGTAAAAC
```

Un contig es la secuencia de consenso generada al alinear las lecturas entre sí.

Universidad del Rosario

Biología



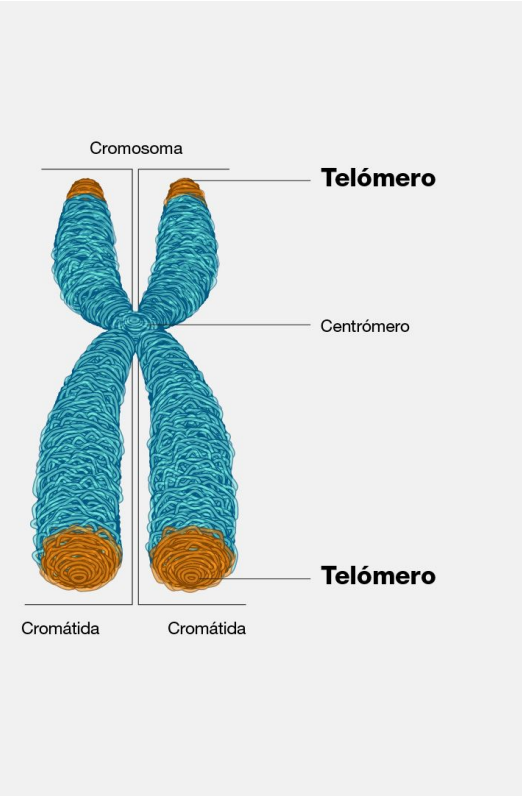
Fragment

Read (known sequence)

Roughly known length but not known sequence

Cromosoma

- Moléculas de ADN más grandes de una célula.
- Se pueden ordenar y orientar mediante un mapa genético o datos de Hi-C.
- El objetivo final de un proyecto de ensamblaje genómico es ensamblar lecturas en cromosomas en fase que representen a un individuo real.
- La mayoría de los ensamblajes cromosómicos producidos hoy en día no están en fase.



Universidad del Rosario

Biología

FASTA

```
>gi|425153|gb|L26238.1|MUSHOME Mus domesticus (lbx) homeodomain mRNA, partial cds
CCATTTCAACAAGTACCTGACCAGGGCTCGGCGAGTGGAAGTTGCCGCTATTCTCGAGCTCAACGAAACT
CAAGTGAAAATT
```

```
Line 1: starts with “>” followed by ID
Line 2: Sequence data
```

FASTQ

```
Line 1: starts with “@” followed by ID
Line 2: Sequence data
Line 3: Starts with “+”       rest of the description is optional
Line 4:Quality score for each base in the sequence
```

```
@HISEQ:402:H147CADXX:1:1101:1250:2208 1:N:0:CGATGT
TGATGCTGCNAATTTTATTCAGTCAGCGGAGGGGGCTTACGTGTATTTTCTGCAACCTTT
+
CCCFFFFFH#4AFIJJJJJJJJIJJJJJJJJJJJJJJJJJJHHHHHHFFFFFFFEEEEED
```

Universidad del Rosario

Biología

GFF-General Feature Format

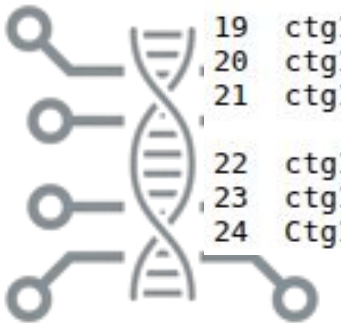
```
Column 1  seqID (e.g. chromosome/scaffold, genome id, etc..)
Column 2  Source (program used to generate or location of download)
Column 3  Feature type (gene, mRNA, CDS, exon, etc.)
Column 4  Start position of feature
Column 5  End position of feature
Column 6  Score (some program outputs will have a score of confidence for feature)
Column 7  Strand (+,-,.)
Column 8  Phase
Column 9  List of attributes in the format tag=value. Multiple attributes are separated by ";"
```

Universidad del Rosario

Biología

GFF-General Feature Format

```
0  ##gff-version   3
1  ##sequence-region   ctg123 1 1497228
2  ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN

3  ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001

4  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
5  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
6  ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3

7  ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
8  ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
9  ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
10 ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
11 ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003

12 ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
13 ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
14 ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
15 ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1

16 ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
17 ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
18 ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2

19 ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
20 ctg123 . CDS             5000  5500  .  +  2  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
21 ctg123 . CDS             7000  7600  .  +  2  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3

22 ctg123 . CDS             3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
23 ctg123 . CDS             5000  5500  .  +  2  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
24 Ctg123 . CDS             7000  7600  .  +  2  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```



Universidad del Rosario

Biología

GTF: Gene Transfer Formats

GTF es una ligera variación de GFF. Las primeras ocho columnas son iguales. La novena columna tiene una sintaxis diferente.

Example for GTF file

```
AB000381 Twinscan  CDS          380   401   .   +   0  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  CDS          501   650   .   +   2  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  CDS          700   707   .   +   2  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  start_codon  380   382   .   +   0  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  stop_codon   708   710   .   +   0  gene_id "001"; transcript_id "001.1";
```

GFF for comparison

```
1  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
2  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
3  ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
4
```

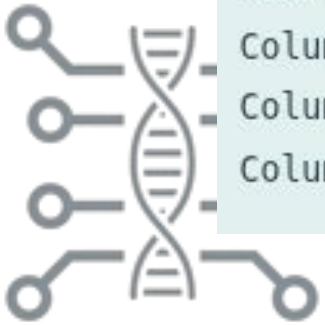
Universidad del Rosario

Biología

VCF: Variant Call Format

Archivo de texto para almacenar variantes de secuencia, SNP e InDels. A diferencia de otros formatos, VCF no almacena todos los datos genéticos redundantes que se comparten entre los genomas.

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  sample1 sample2 ...

Column 1: CHROM – chromosome name
Column 2: POS – position in the chromosome
Column 3:  ID – identifier
Column 4: REF – reference base(s) in the   reference genome
Column 5: ALT – alternate base(s) in the comparing sequence
Column 6: QUAL – quality score
Column 7: FILTER – filter status
Column 8: INFO – additional information
Column 9: FORMAT – genotype information
Column 10: sample–1
Column 11: sample–2 and so on …
```



Universidad del Rosario

Biología

VCF: Variant Call Format

```
##fileformat=VCFv4.2
##contig=<ID=2,length=51304566>
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#CHROM  POS ID  REF ALT QUAL FILTER  INFO FORMAT     SAMPLE1     SAMPLE2     SAMPLE3     SAMPLE4     SAMPLE5     SAMPLE6
2  81170  .  C  T   .  .   AC=9;AN=7424   GT:DP:GQ  0/0:4:12    0/0:3:9     0/1:1:3     0/1:9:24    1/0:4:12    0/0:5:15
2  81171  .  G  A   .  .   AC=6;AN=7446   GT:DP:GQ  0/1:4:12    0/0:3:9     0/0:1:3     0/0:9:24    0/1:4:12    0/1:5:15
2  81182  .  A  G   .  .   AC=5;AN=7506   GT:DP:GQ  0/0:5:15    0/0:4:12    0/0:5:15    0/0:9:24    0/0:4:12    0/0:4:12
2  81204  .  T  G   .  .   AC=2;AN=7542   GT:DP:GQ  1/0:5:15    0/0:9:27    0/0:10:30   0/0:15:39   0/0:9:27    1/0:13:39
```

Universidad del Rosario

Biología

SAM-Sequence Alignment/Map

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001    99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002     0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003     0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004     0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001   147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Header section

Alignment section

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

Universidad del Rosario

Biología

SAM-Sequence Alignment/Map

FLAGS

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

https://broadinstitute.github.io/picard/explain-flags.html

Universidad del Rosario

Biología