Fundamentos y herramientas bioinformáticas para análisis genómicos

FACULTAD DE CIENCIAS
UDELAR | fcien.edu.uy

Institut Pasteur de Montevideo

Práctico ensamblaje y anotación genómica

Luisa Berná, PhD
lberna@pasteur.edu.uy

Unidad de Bioinformática - Laboratorio de Biología de Apicomplejos
Institut Pasteur de Montevideo

Laboratorio de Genómica Evoluitva - Facultad de Ciencias, UDELAR.

Objetivos

- Aprender a ensamblar genomas utilizando distintos tipos de secuencias y estrategias
- Afianzar conceptos sobre las distintas tecnologías de secuenciación
- Afianzar conceptos sobre algoritmos de ensamblaje
- Aprender a realizar anotación automática, y entender sus virtudes y limitaciones.

Cómo?

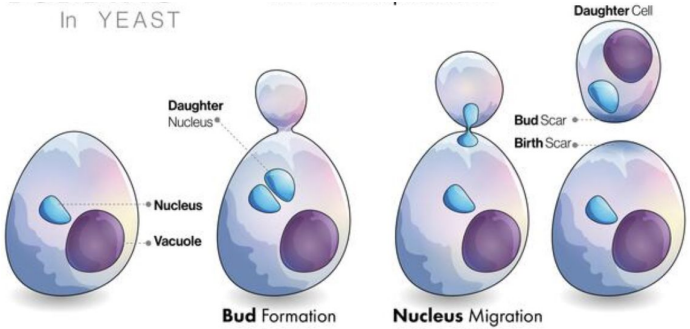
- Ensamblando el genoma de un organismo modelo (levadura) para el cual existen diferentes tipos de datos
  - Illumina (paired-end)
  - Nanopore
- Haciendo una predicción de genes

- A través de la práctica de distintas herramientas y la comparación y discusión de los resultados obtenidos

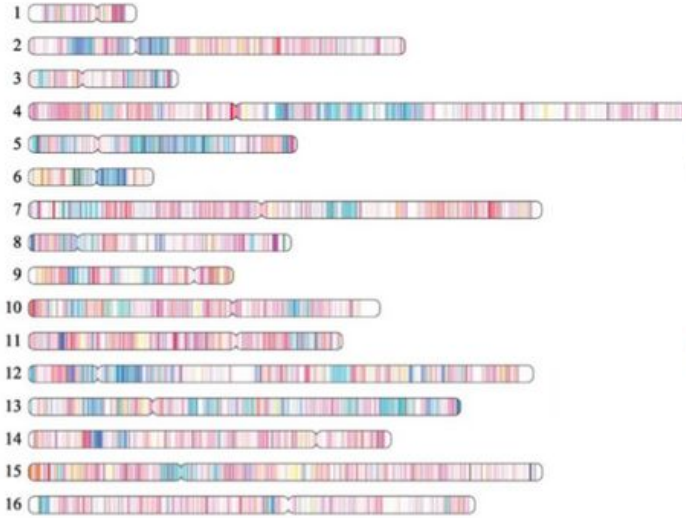
Modelo de estudio

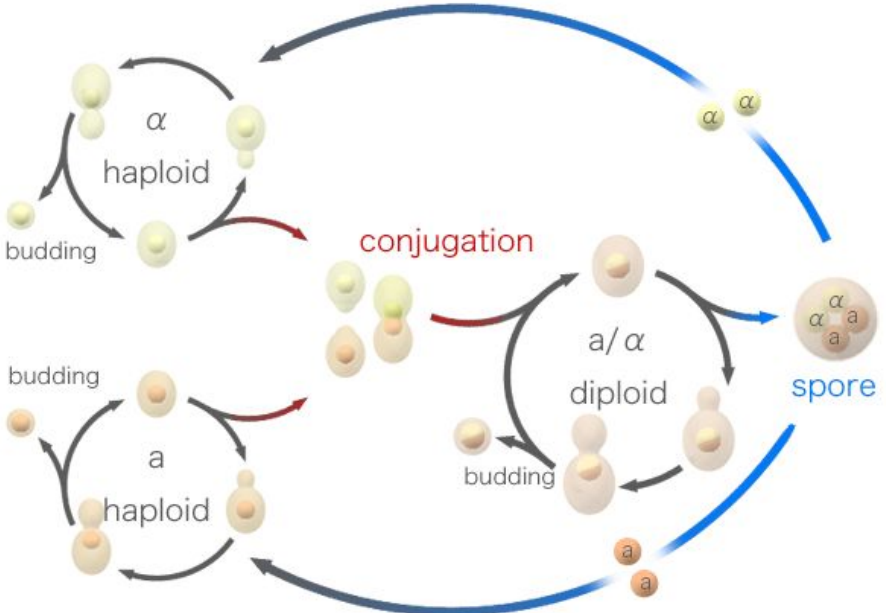
Para este práctico se utilizará la levadura (*Saccharomyces cerevisiae*). Este organismo tiene varias ventajas que lo hacen interesante para realizar un práctico de ensamblaje y anotación:

- ★ Ampliamente conocido e interesante (pan/vino/cerveza)
- ★ Eucariota (complejidad intermedia)
- ★ Tamaño genómico reducido (12Mb)
- ★ 16 Cromosomas
- ★ Genes con Intrones (~10%)
- ★ EXISTEN datos disponibles de varias plataformas (illumina/PacBio/Nanopore)



I	230.218
II	813.184
III	316.620
IV	1.531.933
V	576.874
VI	270.161
VII	1.090.940
VIII	562.643
IX	439.888
X	745.751
XI	666.816
XII	1.078.177
XIII	924.431
XIV	784.333
XV	1.091.291
XVI	948.066





Published online 13 January 2018

Nucleic Acids Research, 2018, Vol. 46, No. 7 ***e38***
doi: 10.1093/nar/gky014

Complete genomic and transcriptional landscape analysis using third-generation sequencing: a case study of *Saccharomyces cerevisiae* CEN.PK113-7D

Piroon Jenjaroenpun[1,†], Thidathip Wongsurawat[1,†], Rui Pereira[2], Preecha Patumcharoenpol[1], David W. Ussery[1,3], Jens Nielsen[2,4] and Intawat Nookaew[1,2,3,*]

[1]Department of Biomedical Informatics, College of Medicine, University of Arkansas for Medical Sciences, Little Rock, AR 72205, USA, [2]Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg SE-412 96, Sweden, [3]Department of Physiology and Biophysics, College of Medicine, The University of Arkansas for Medical Sciences, Little Rock, AR 72205, USA and [4]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK2800 Lyngby, Denmark

Received September 11, 2017; Revised January 03, 2018; Editorial Decision January 04, 2018; Accepted January 05, 2018

1 - Pre-procesamiento de secuencias

- Descargar datos // Obtener datos a partir de secuenciaciones
- Analizar la calidad de las secuencias
- Filtrar si es necesario
- Re-analizar la calidad para determinar si los filtros utilizados fueron adecuados

2 - Ensamblaje DE NOVO

Vamos a ensamblar *de novo* un genoma de *Saccharomyces cerevisiae* usando reads obtenidos a partir de tecnologías diferentes:

- Illumina
- Nanopore ONT

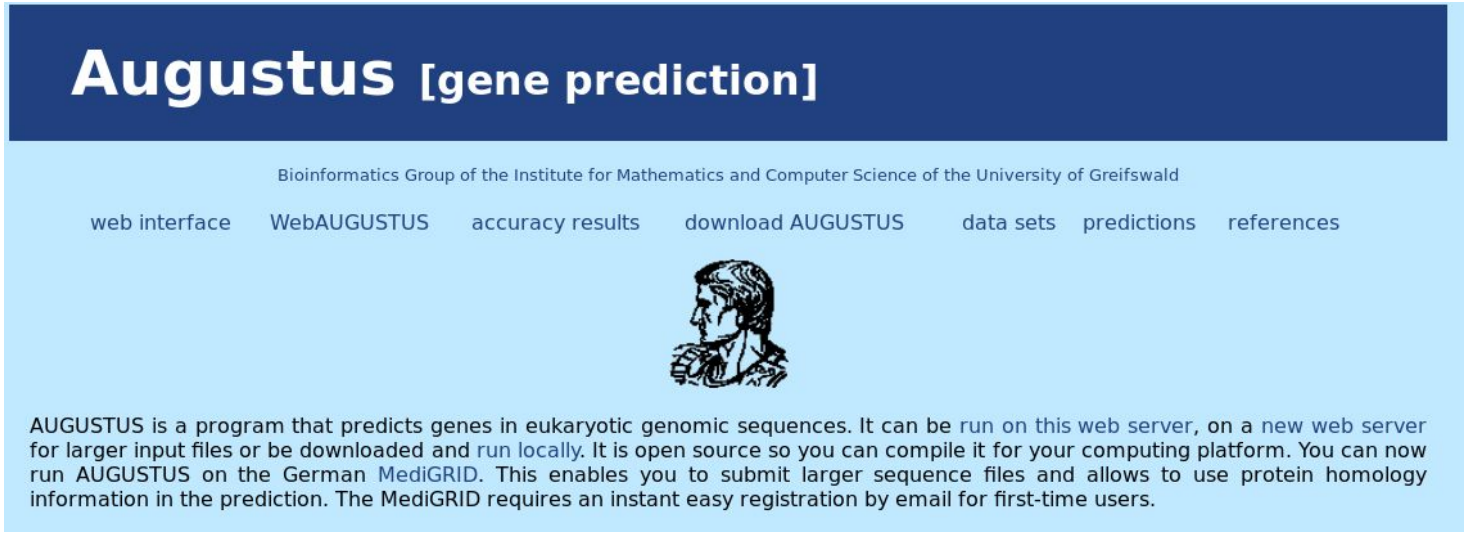
Vamos a utilizar distintos programas y comparar sus resultados

3- Evaluación de los ensamblajes

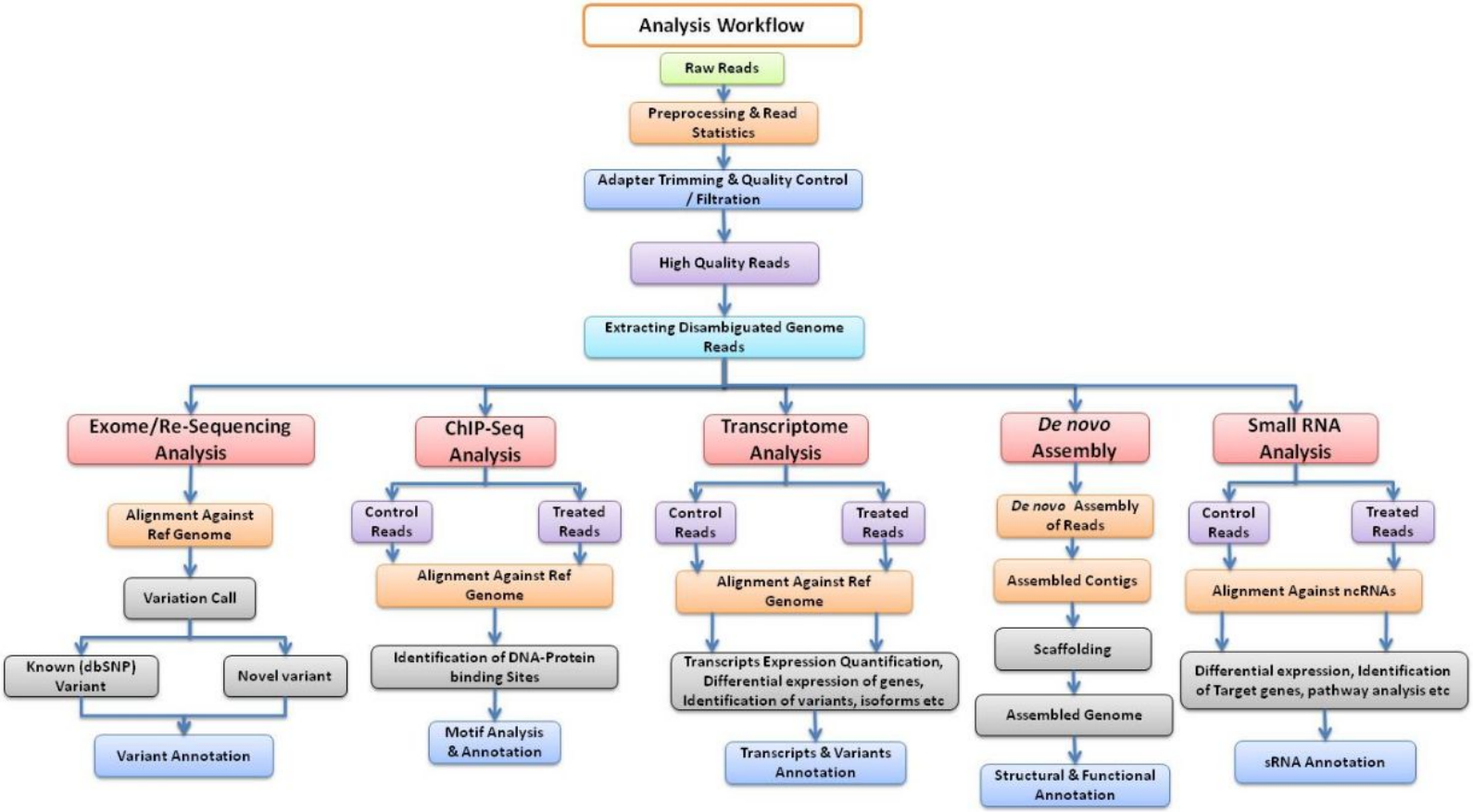
- **infoseq**
- **YASS**
- **BLAST**
- **Assemblytics**
- **BUSCO**

4- Búsqueda de genes

- AUGUSTUS



Formatos


Analysis Workflow
Raw Reads
Preprocessing & Read Statistics
Adapter Trimming & Quality Control / Filtration
High Quality Reads
Extracting Disambiguated Genome Reads
Exome/Re-Sequencing Analysis
Alignment Against Ref Genome
Variation Call
Known (dbSNP) Variant
Novel variant
Variant Annotation
ChIP-Seq Analysis
Control Reads
Treated Reads
Alignment Against Ref Genome
Identification of DNA-Protein binding Sites
Motif Analysis & Annotation
Transcriptome Analysis
Control Reads
Treated Reads
Alignment Against Ref Genome
Transcripts Expression Quantification, Differential expression of genes, Identification of variants, isoforms etc
Transcripts & Variants Annotation
De novo Assembly
De novo Assembly of Reads
Assembled Contigs
Scaffolding
Assembled Genome
Structural & Functional Annotation
Small RNA Analysis
Control Reads
Treated Reads
Alignment Against ncRNAs
Differential expression, Identification of Target genes, pathway analysis etc
sRNA Annotation

Algunas definiciones

- **Single read sequencing** = DNA fragments read from one end.

- **Paired-end sequencing** = DNA fragments read from both ends. Reads are typically up to a few hundred bp apart, but may also overlap.

- **Mate pair sequencing** – the ends of long DNA fragments (typically several kb apart) are fused and then excised for sequencing.

- **Long reads** – Much longer reads! Can thus improve de novo assembly, mapping certainty, transcript isoform identification, and detection of structural variants

Formatos

- FASTA
- FASTQ
  - Calidad PHRED

FASTA

Un formato de texto plano utilizado para representar secuencias de nucleótidos o proteínas. Los nucleótidos se representan con las letras A,C,G,T y los aminoácidos con el código de una letra:

- **Encabezado**
  descripción en una única línea que comienza con el símbolo ’>’. No hay espacios entre el ’>’ y el comienzo de la descripción. Puede contener un simple identificador o una descripción más compleja y comentarios
- **Cuerpo**
  la propia secuencia representada por la sucesión de letras. Pueden ir todas en una línea o en bloques (en general de 80 caracteres)

FASTA

```
>seq1
GATCAACGCAAAGGACTAAGCACTGCTGCCAAAAGCCACCAGCCCCAGAGACAACAGAGG
CTCCCAAATTTCTAGCCTCTGATCTCTGCCTCGGAACATTCTTGGGTCAAAATAAATGTG
CGATCGCTAGCTAAAACGTTTCGAT
```

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

multiFASTA

- Un archivo puede contener más de una secuencia. Cada secuencia comienza con su encabezado y tiene su cuerpo
- Un multifasta puede tener miles y miles de secuencias (por ejemplo uno con todas las proteínas de un genoma)

```
>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGKAAKKADRLAAEG
LVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRLKDPNKPEHK
IPQFASRKQLSDAILKEAEEKIKEELKAQGKPEKIWDNIIPGKMNSFIADNSQLDSKLTL
MGQFYVMDDKKTVEQVIAEKEKEFGGKIKIVEFICFEVGEGLEKKTEDFAAEVAAQL
>SEQUENCE_2
SATVSEINSETDFVAKNDQFIALTKDTTAHIQSNSLQSVEELHSSTINGVKFEEYLKSQI
ATIGENLVVRRFATLKAGANGVVNGYIHTNGRVGVVIAAACDSAEVASKSRDLLRQICMH
```

FASTQ Fasta Quality

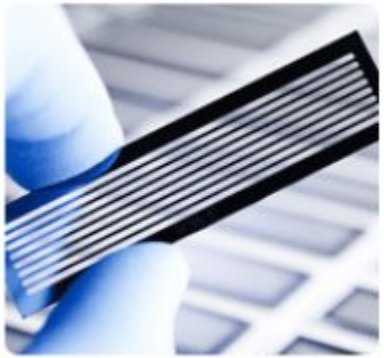
- Es el archivo que obtenemos de las distintas plataformas de secuenciación

- Posee información de secuencia (como el archivo FASTA:

```
@K00171:520:HLKJ5BBXX:1:1101:24535:1138 1:N:0:GTGTTCTA
AGTGGTCGGAGAAGTTTTCTGATGAGGGAGATTCGGACTCAGCCCTGGACTCGGTCTCTCCCTGCCCGTCCTCTCCAAAACAGATACACCTGGATGTGGAC
+
AAFFFAFFJJJFJJFAFJJJJFFFFJJJJJAAFJJJJJJJJJJAJJJJFFJJJJJ<FJJJJJJJJJJF<<A-AJJJJJJFJJFJJFJJJFFFJAFJFFJFJ
@K00171:520:HLKJ5BBXX:1:1101:26058:1138 1:N:0:GTGTTCTA
AGGAGGGGATTTCATGGGTAGACTGTGAAATTCCTGACCTCCCAAACACTCCATGTGATGTCAGCTCCACAGTCTTTTTCTATGTGTGTGTGTGTTTTG
+
AAFAFJJJFFJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJ<JFFJJJJJJJJJJJJJFJJJJJJJJJJJJFJJ<FFFJJJJJJJ
@K00171:520:HLKJ5BBXX:1:1101:26565:1138 1:N:0:GTGTTCTA
TCGGCCTGGGCCTGGGCCGGGGCCGCCTCAGGGAGCGGCAGGTAGGGCTGAGAGAAGCCGTCCACCAGGAGCACGTGCAGCGTGGCGGTGGCCGAGCGCGG
+
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFJJJJJJJFJJJJJJFJJFJJJJJJJJJJJFJFJJJJJJJJJ<FJJJJJJJJJJJJJJJJJJJJJ
@K00171:520:HLKJ5BBXX:1:1101:28777:1138 1:N:0:GTGTTCTA
GAGGCCTGTGGGTGCTGCAGGGCACCAGGGGGGTCACTAGGGAGATGCAGGGCGCGGGCCAGGAGGGGCGCCAGGGCAGCCGGGGGCCAGGCTGAGGCGAC
+
AAAAFJJJJJJJJJJJJJJJJFJFJJJJJJJJJ7FJFJJJJ<FJJJJJJJJJJFJJJJFAJJJ<JJFFFJJJFJJJFJJJJJJFFJJFJJJJJJJJJJJJ
```

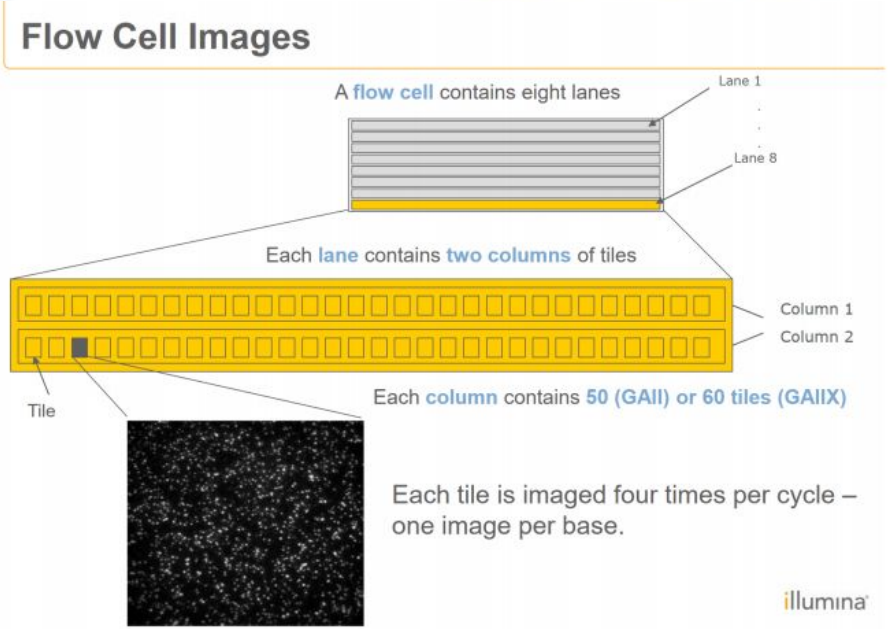
ezado ‘+’ y

FASTQ encabezado



```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*



FASTQ Calidad

PHRED SCORE es una medida de calidad **Q**

P es la probabilidad de que un nucleótido haya sido asignado incorrecta

$$Q = -10 \log_{10} P \qquad P = 10^{\frac{-Q}{10}}$$

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

FASTQ Calidad - código ASCII

Caracteres ASCII de control		
00	NULL	(carácter nulo)
01	SOH	(inicio encabezado)
02	STX	(inicio texto)
03	ETX	(fin de texto)
04	EOT	(fin transmisión)
05	ENQ	(consulta)
06	ACK	(reconocimiento)
07	BEL	(timbre)
08	BS	(retroceso)
09	HT	(tab horizontal)
10	LF	(nueva línea)
11	VT	(tab vertical)
12	FF	(nueva página)
13	CR	(retorno de carro)
14	SO	(desplaza afuera)
15	SI	(desplaza adentro)
16	DLE	(esc.vínculo datos)
17	DC1	(control disp. 1)
18	DC2	(control disp. 2)
19	DC3	(control disp. 3)
20	DC4	(control disp. 4)
21	NAK	(conf. negativa)
22	SYN	(inactividad sínc)
23	ETB	(fin bloque trans)
24	CAN	(cancelar)
25	EM	(fin del medio)
26	SUB	(sustitución)
27	ESC	(escape)
28	FS	(sep. archivos)
29	GS	(sep. grupos)
30	RS	(sep. registros)
31	US	(sep. unidades)
127	DEL	(suprimir)

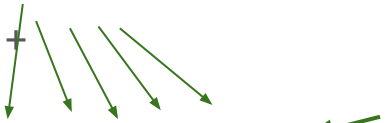
Caracteres ASCII imprimibles					
32	espacio	64	@	96	`
33	!	65	A	97	a
34	"	66	B	98	b
35	#	67	C	99	c
36	$	68	D	100	d
37	%	69	E	101	e
38	&	70	F	102	f
39	'	71	G	103	g
40	(	72	H	104	h
41	)	73	I	105	i
42	*	74	J	106	j
43	+	75	K	107	k
44	,	76	L	108	l
45	-	77	M	109	m
46	.	78	N	110	n
47	/	79	O	111	o
48	0	80	P	112	p
49	1	81	Q	113	q
50	2	82	R	114	r
51	3	83	S	115	s
52	4	84	T	116	t
53	5	85	U	117	u
54	6	86	V	118	v
55	7	87	W	119	w
56	8	88	X	120	x
57	9	89	Y	121	y
58	:	90	Z	122	z
59	;	91	[	123	{
60	<	92	\	124	|
61	=	93	]	125	}
62	>	94	^	126	~
63	?	95	_		

ASCII extendido							
128	Ç	160	á	192	└	224	Ó
129	ü	161	í	193	┴	225	ß
130	é	162	ó	194	┬	226	Ô
131	â	163	ú	195	├	227	Ò
132	ä	164	ñ	196	─	228	õ
133	à	165	Ñ	197	┼	229	Õ
134	å	166	ª	198	ã	230	µ
135	ç	167	º	199	Ã	231	þ
136	ê	168	¿	200	╚	232	Þ
137	ë	169	®	201	╔	233	Ú
138	è	170	¬	202	╩	234	Û
139	ï	171	½	203	╦	235	Ù
140	î	172	¼	204	╠	236	ý
141	ì	173	¡	205	═	237	Ý
142	Ä	174	«	206	╬	238	¯
143	Å	175	»	207	¤	239	´
144	É	176	░	208	ð	240	≡
145	æ	177	▒	209	Ð	241	±
146	Æ	178	▓	210	Ê	242	‗
147	ô	179	│	211	Ë	243	¾
148	ö	180	┤	212	È	244	¶
149	ò	181	Á	213	ı	245	§
150	û	182	Â	214	Í	246	÷
151	ù	183	À	215	Î	247	¸
152	ÿ	184	©	216	Ï	248	°
153	Ö	185	╣	217	┘	249	¨
154	Ü	186	║	218	┌	250	·
155	ø	187	╗	219	█	251	¹
156	£	188	╝	220	▄	252	³
157	Ø	189	¢	221	¦	253	²
158	×	190	¥	222	Ì	254	■
159	ƒ	191	┐	223	▀	255	nbsp

FASTQ Calidad - código ASCII + 33

Caracteres ASCII imprimibles					
32	espacio	64	@	96	`
33	!	65	A	97	a
34	"	66	B	98	b
35	#	67	C	99	c
36	$	68	D	100	d
37	%	69	E	101	e
38	&	70	F	102	f
39	'	71	G	103	g
40	(	72	H	104	h
41	)	73	I	105	i
42	*	74	J	106	j
43	+	75	K	107	k
44	,	76	L	108	l
45	-	77	M	109	m
46	.	78	N	110	n
47	/	79	O	111	o
48	0	80	P	112	p
49	1	81	Q	113	q
50	2	82	R	114	r
51	3	83	S	115	s
52	4	84	T	116	t
53	5	85	U	117	u
54	6	86	V	118	v
55	7	87	W	119	w
56	8	88	X	120	x
57	9	89	Y	121	y
58	:	90	Z	122	z
59	;	91	[	123	{
60	<	92	\	124	|
61	=	93	]	125	}
62	>	94	^	126	~
63	?	95	_		

@SSR34543212 Saccharomyces cerevisiae
ATTCGCCAGGTCTAG
+



@SSR34543212 Saccharomyces cerevisiae CEN.PK113-7D
ATTCGCCAGGTCTAG
+
BC#4>RTWX ! ! [KU>

Base1: Phred value=33, sumamos 33 = 66 (ascii 66 = B)
Base2: Phred value=34, sumamos 33 = 67 (ascii 67 = C)
Base3: Phred value=2, sumamos 33 = 35 (ascii 35 = #)
Base4: Phred value=19, sumamos 33 = 52 (ascii 52 = 4)
Base15: Phred value= 29, sumamos 33 = 62 (ascii 62 = >)

Qué pasa con los reads largos?

Mayor proporciones de errores

Errores aleatorios

Mejoras en los algoritmos que realizan la asignación de bases

Repositorios públicos - banco de datos

- **SRA (Sequence Reads Archive)**
  Un repositorio público con datos de secuenciaciones de distintas plataformas

- **GEO (Gene Expression Omnibus)**
  Un repositorio público con datos de expresión génica (Microarrays, NGS)

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

Announcement

NIH Request for Information (RFI) on SRA data format changes and plans.

Getting Started

- How to Submit
- How to search and download
- How to use SRA in the cloud
- Submit to SRA

Tools and Software

- Download SRA Toolkit
- SRA Toolkit Documentation
- SRA-BLAST
- SRA Run Browser
- SRA Run Selector

Related Resources

- Submission Portal
- Trace Archive
- dbGaP Home
- BioProject
- BioSample

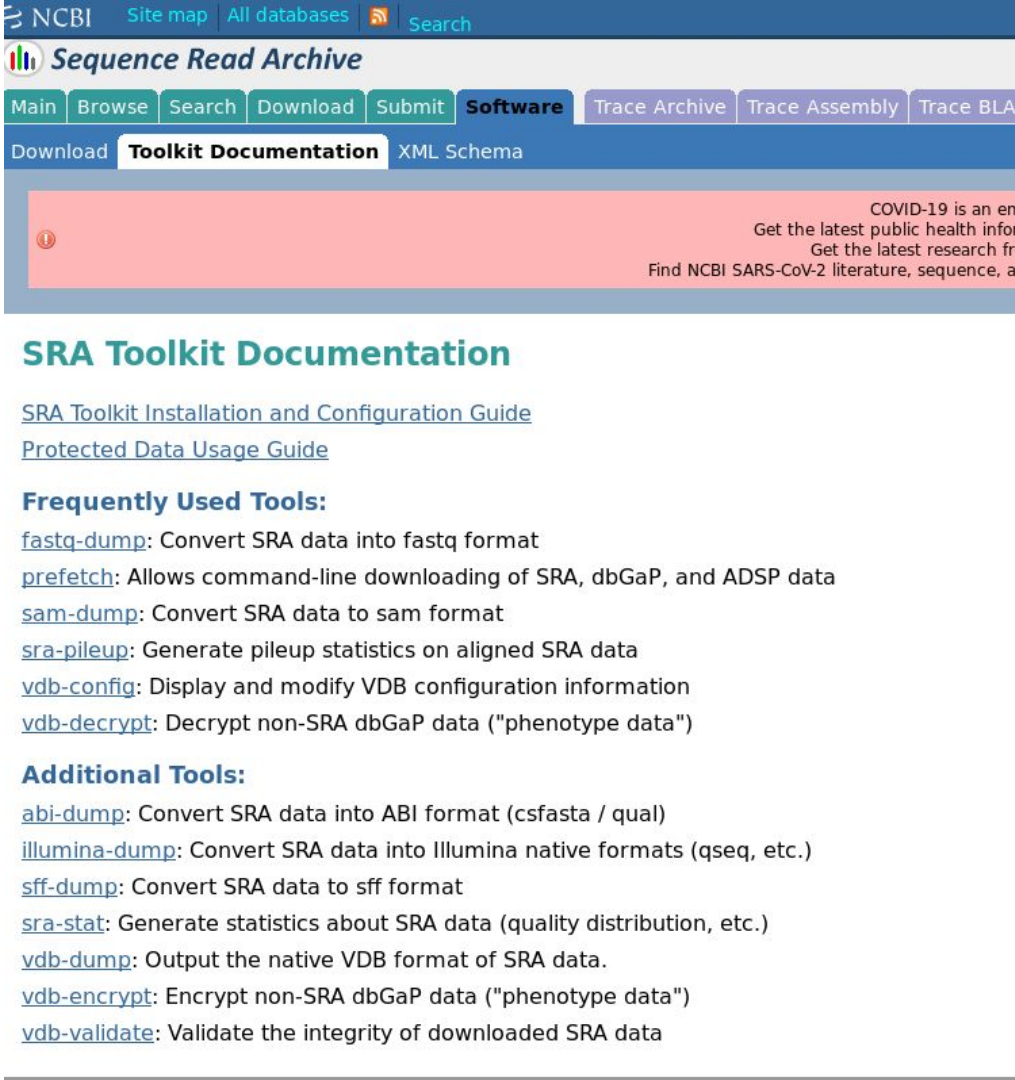
fastq-dump

Convierte datos SRA en formato FASTQ

fastq-dump [opciones] <IDENTIFICADOR>

Ej:

fastq-dump --splite-files SRR6074044.sra



fastq-dump

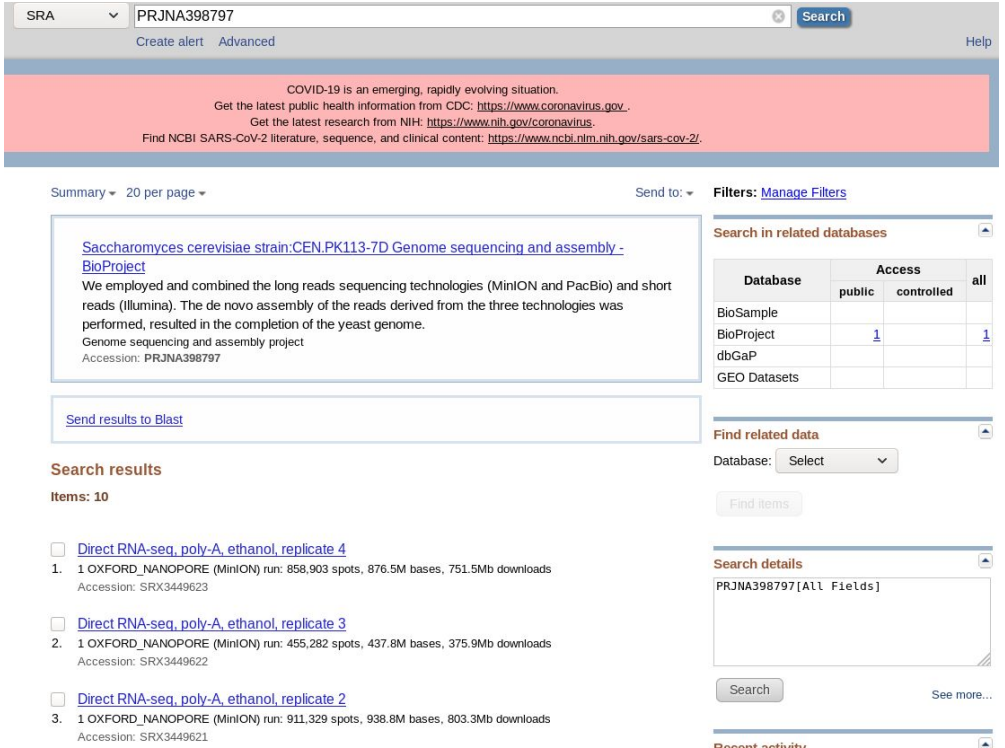
Convierte datos SRA en formato FASTQ

fastq-dump [opciones] <IDENTIFICADOR>

Ej:

fastq-dump --splite-files SRR6074044.sra

selected with the BluePippin system with a cut-off value of 9000 bp. We used one SMRTcell™ to sequence the DNA library on the PacBio Sequel instrument using the Sequel 2.0 polymerase and 600 min of movie time. The high quality PacBio reads are deposited in an SRA database under BioProject:PRJNA398797, SRP116559.



Obtención de datos de SRA

- Usando PREFETCH
  prefetch SRR8922830
  (genera un archivo SRR8922830.sra)
- Usando fastq-dump
  fastq-dump –Z –X 10 SRR8922830
  (imprime en pantalla (opción -Z) las primeras 10 secuencias (-X 10)
- Usando WGET
  wget https://sra-downloadb.st-va.ncbi.nlm.nih.gov/sos1/sra-pub-run-12/SRR6352892/SRR6352892.1
- Usando FTP
  ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR607/SRR6074044/SRR6074044.sra

Usando WGET

Sequence Read Archive

Main | Browse | Search | Download | Submit | Software | Trace Archive | Trace Assembly | Trace BLAST

Studies | Samples | Analyses | Run Browser | Run Selector | Provisional SRA

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

Direct RNA-seq, poly-A, ethanol, replicate 4 (SRR6352892)

Metadata | Analysis | Reads | Data access

SRA archive data

SRA archive data is normalized by the SRA load process and used by the SRA Toolkit to read and produce formats like FASTQ, SAM, etc. The default toolkit configuration enables it to find and retrieve SRA runs by accession.

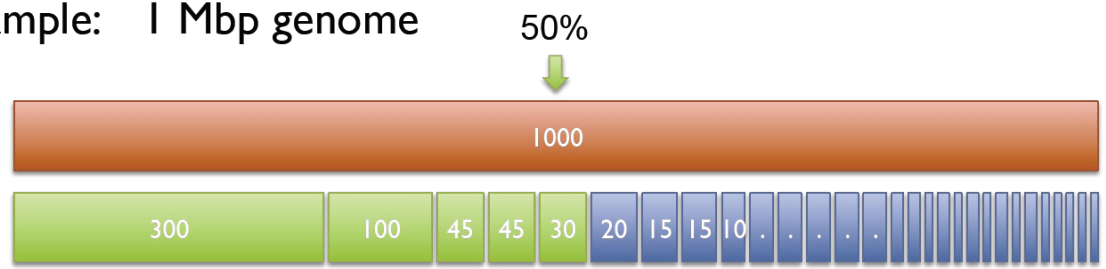
Public SRA files are now available from GCP and AWS cloud platforms as well as from NCBI. Access to most data in the cloud requires a user account with the cloud service provider. The user's account will incur costs for cloud compute or to copy data outside of the specified cloud service region.

Type	Size	Location	Name	Free Egress	Access Type
run	769,543 Kb	NCBI	https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos2/sra-pub-run-11/SRR6352892/SRR6352892.1	worldwide	anonymous
		NCBI	https://sra-downloadb.st-va.ncbi.nlm.nih.gov/sos1/sra-pub-run-12/SRR6352892/SRR6352892.1	worldwide	anonymous
		AWS	s3://sra-pub-run-5/SRR6352892/SRR6352892.1	s3.us-east-1	aws identity
		GCP	gs://sra-pub-run-5/SRR6352892/SRR6352892.1	gs.US	gcp identity

N50 size

Def: 50% of the genome is in contigs larger than N50

Example: 1 Mbp genome



N50 size = 30 kbp

(300k+100k+45k+45k+30k = 520k >= 500kbp)

Note:

N50 values are only meaningful to compare when base genome size is the same in all cases