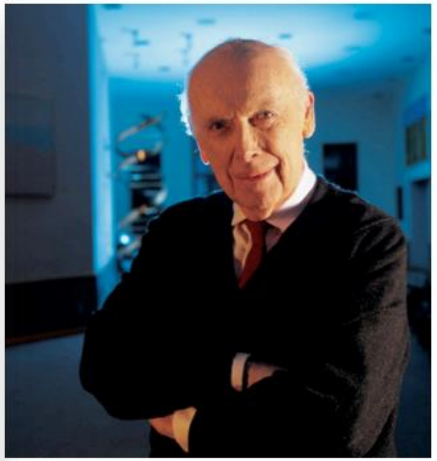
Genomas individuales: 454

nature Vol 452|17 April 2008|doi:10.1038/nature06884

LETTERS

The complete genome of an individual by massively parallel DNA sequencing

David A. Wheeler[1]*, Maithreyan Srinivasan[2]*, Michael Egholm[2]*, Yufeng Shen[1]*, Lei Chen[1], Amy McGuire[3], Wen He[2], Yi-Ju Chen[2], Vinod Makhijani[2], G. Thomas Roth[2], Xavier Gomes[2], Karrie Tartaro[2]†, Faheem Niazi[2], Cynthia L. Turcotte[2], Gerard P. Irzyk[2], James R. Lupski[4,5,6], Craig Chinault[4], Xing-zhi Song[1], Yue Liu[1], Ye Yuan[1], Lynne Nazareth[1], Xiang Qin[1], Donna M. Muzny[1], Marcel Margulies[2], George M. Weinstock[1,4], Richard A. Gibbs[1,4] & Jonathan M. Rothberg[2]†



Watson

- 6 meses luego de la publicacion del anterior
- Plataforma 454 FLX
- 24,5 billion bases en 93,2 milliones de lecturas
- 7.4X de covertura
- Menos de 1 million de dolares
- Completado en menos de 5 meses

Wheeler, D. A. et al. Nature (2008)

Genomas famosos!

Wednesday 06 April 2011

The Telegraph

HOME | NEWS | SPORT | FINANCE | COMMENT | CULTURE | TRAVEL | LIFESTYLE | FASHION | TECHNOLOGY

UK | World | Politics | Obituaries | Royal Wedding | Earth | Science | Health News | Education

Celebrity news

Glenn Close has genome sequenced to publicise mental illness

American actress Glenn Close has joined a handful of celebrities to have their genome sequenced in the name of science.

Close is a founder of the nonprofit group BringChange2Mind, which raises awareness about mental illness Photo: AP

Share:

Recommend

Tweet 0

Celebrity news

News » World News » North America » USA »

IN NEWS

Naomi Campbell opens pop up shop

Ozzy Osbourne genome sequenced

Genetic analysis of Black Sabbath star reveals he is more likely to experience hallucinations on marijuana and has increased risk of alcohol and cocaine addiction, researchers say

Tweet 0 Share 814

Comments (29)

Sean Michaels
guardian.co.uk, Friday 5 November 2010 11.30 GMT
Article history

A larger | smaller

Man of visions ... Ozzy Osbourne with wife Sharon at Jon Stewart's Rally to Restore Sanity and/or Fear on Saturday. Photograph: Scott Gries/EMPICS Entertainment

Music
Ozzy Osbourne · Black Sabbath · Metal · Pop and rock

Culture

Science
Genetics · Biology

More news

Related

8 May 2002
Rocker at odds with MTV over second series

6 Oct 2010
Ozzy Osbourne covers John Lennon song for charity

IGSR: The International Genome Sample Resource

Providing ongoing support for the 1000 Genomes Project data

IGSR and the 1000 Genomes Project

Populations: - African; - American; - East Asian; - European; - South Asian;

The International Genome Sample Resource (IGSR) was established to ensure the ongoing usability of data generated by the 1000 Genomes Project and to extend the data set. More information is available about the IGSR.

Links

Announcements

IGSR Sample Collection Principles

1000 Genomes Project Publications

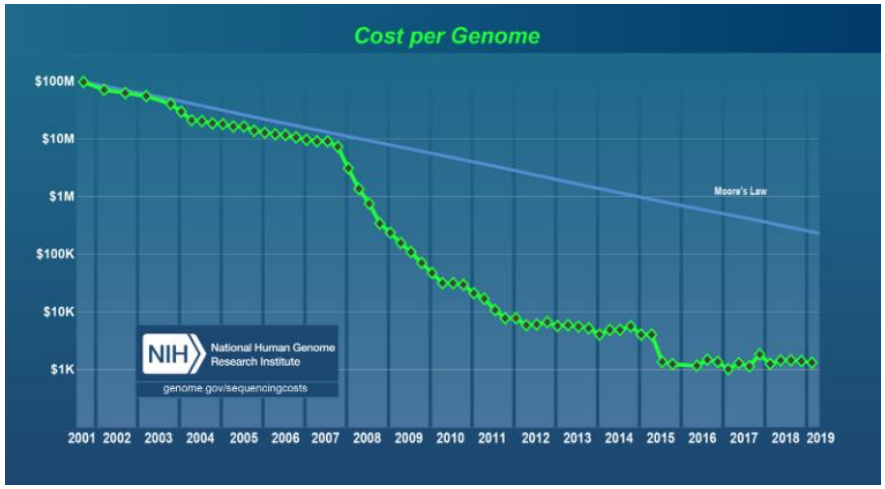
File formats

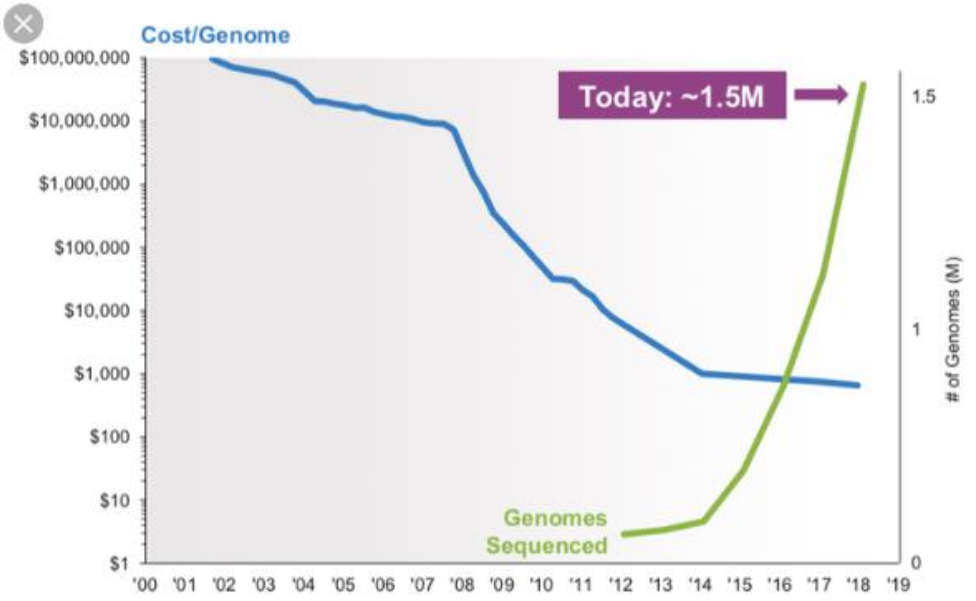
Software tools

Download data

Twitter

Costos y número de genomas completos





- Hoy en día hay millones de genomas secuenciados

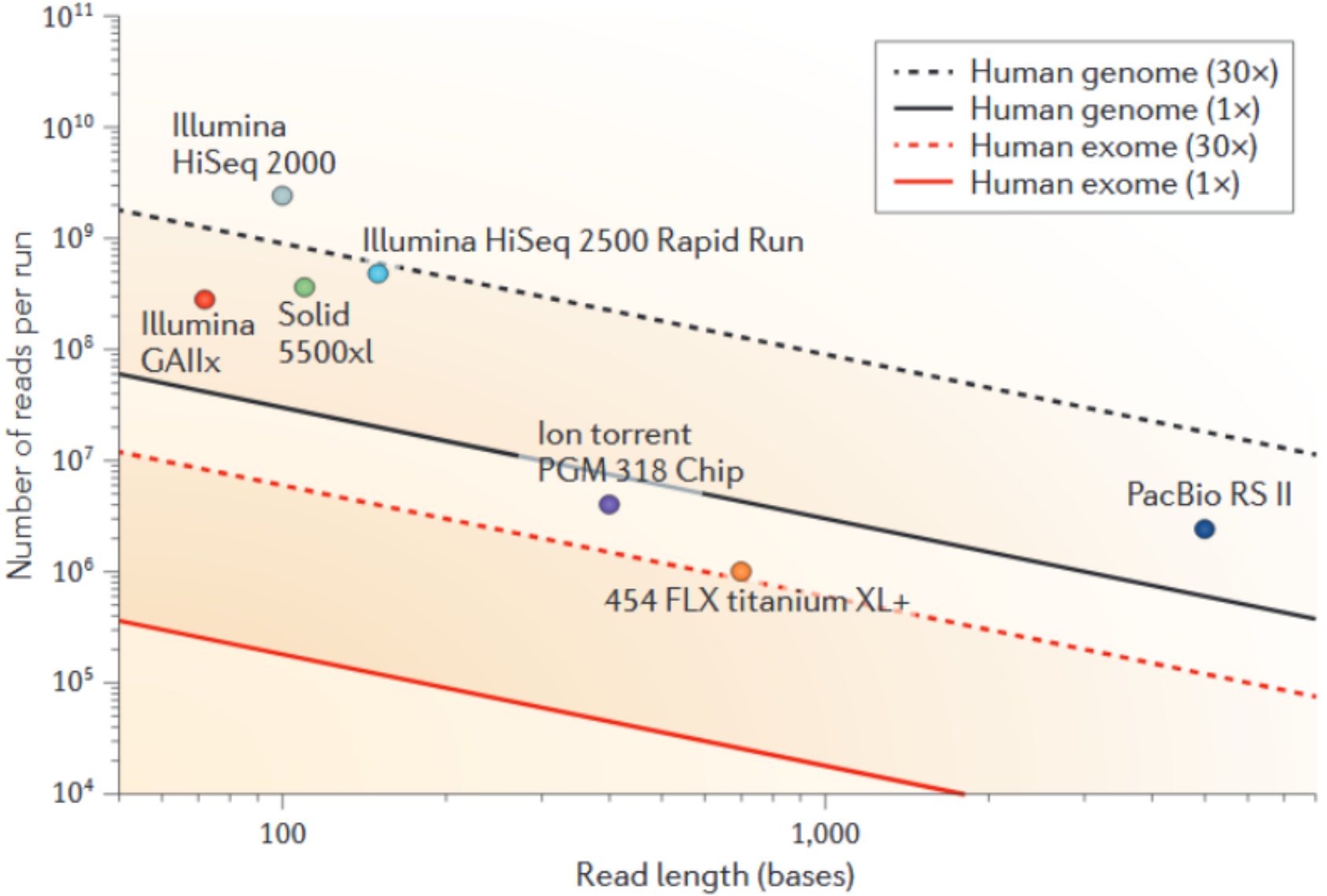
“Personal Genomics”

23andMe competitor Veritas Genetics slashes price of whole genome sequencing 40% to $600

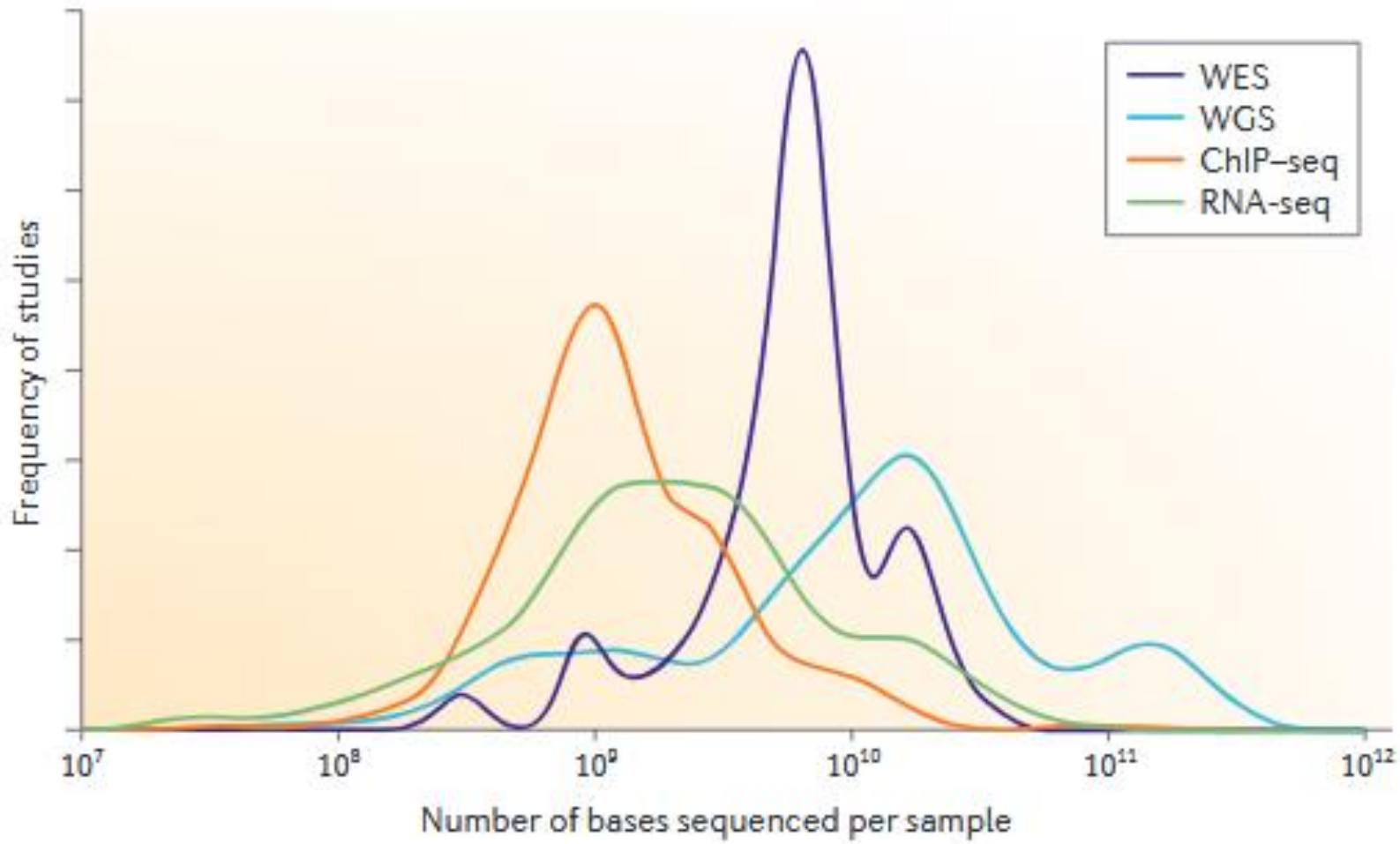
PUBLISHED MON, JUL 1 2019 • 9:30 AM EDT

“Personal Genomics”

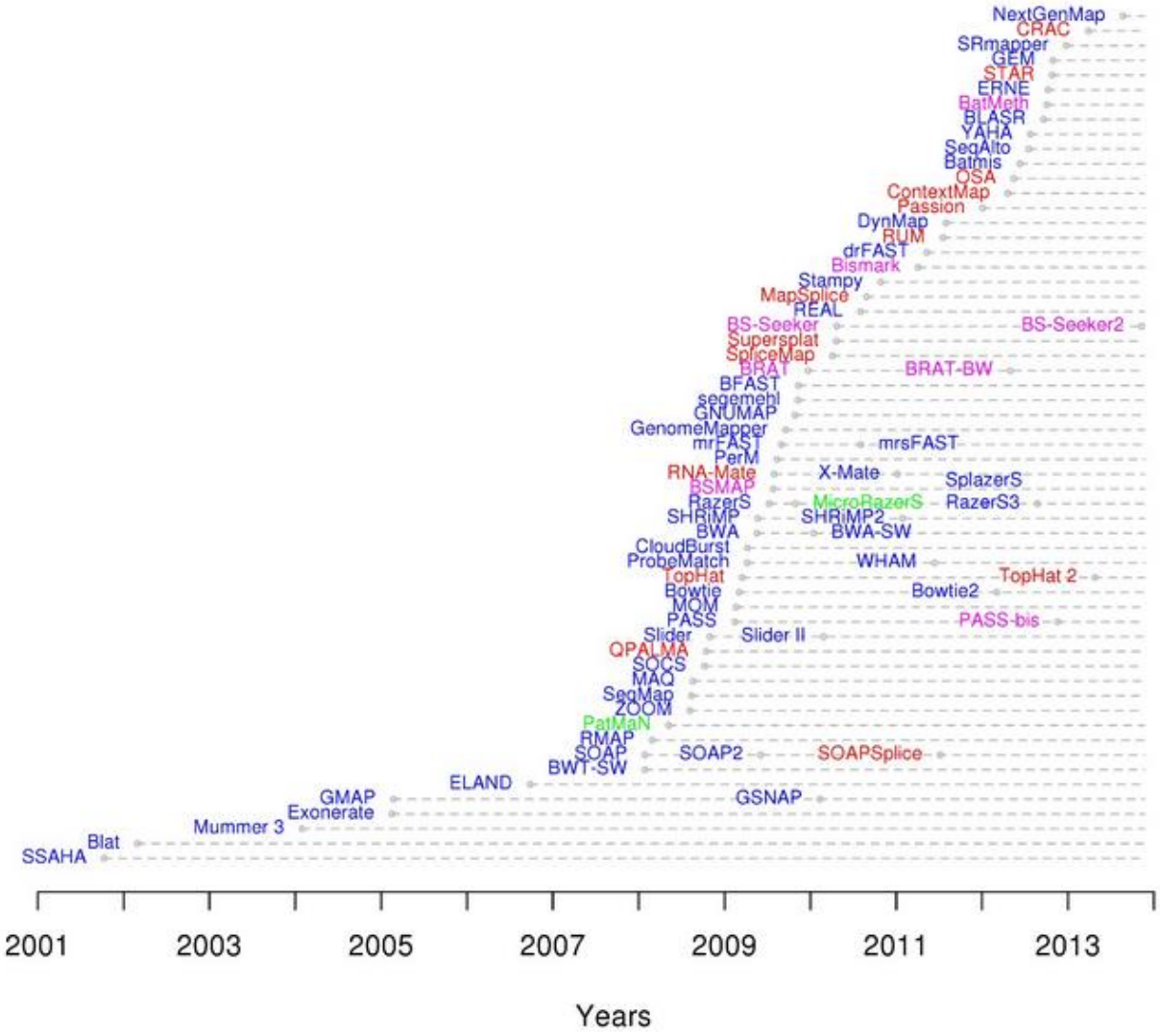
Genetic test company	blood sample	sent by post	USA[3]
ACGT, Inc	blood sample	sent by post	USA[4]
Admera Health LLC	blood sample	sent by post	USA[5]
Affiliated Genetics, Inc	blood sample	sent by post	USA[6]
Applied Biological Materials Inc	blood sample	sent by post	Canada[7]
CD Genomics	blood sample	sent by post	USA
CEN4GEN Institute for Genomics and Molecular Diagnostics	blood sample	sent by post	Canada[8]
ChunLab, Inc.	blood sample	sent by post	Korea, USA[9]
Core Life Sciences, Inc.	blood sample	sent by post	USA[10]
Complete Genomics	blood sample, frozen tissue, cell lines, or saliva[11]	send by post	USA
Clinical Microarray Core	blood sample	send by post	USA
DNA Link, Inc	blood sample	sent by post	Korea, USA[12]
Fulgent Diagnostics	blood sample	sent by post	USA[13]
Full Genomes Corporation	saliva sample	send by post	USA[14]
GENEWIZ, Inc.	blood sample	sent by post	USA[15]
Gene by Gene	blood sample (3 to 5cc)	send by post	USA
Genomix4Life	blood sample	sent by post	Italy[16]
Genomics Personalized Health	blood sample	sent by post	USA[17]
HIB Genomics Services Lab	blood sample	sent by post	USA[18]
Illumina	blood sample (4 to 8 ml vial)[19]	sent by post	USA[20]
Kinghorn Centre for Clinical Genomics	blood sample	sent by post	Australia[22]
Macrogen	blood sample	sent by post	Korea, Japan, USA, Netherlands[23]
Meports	saliva sample	sent by post	USA [24]
myGenomics, LLC	blood sample	sent by post	USA[25]
Novogene Corporation	blood sample	sent by post	USA, UK, Japan, China[26]
Quick Biology	blood sample	send by post	USA
Sequenom	blood sample	drawn on nearby sequenom location	USA
SBMRI Analytical Genomics Core Facility[28]	blood sample	sent by post	USA[29]
Sequentia Biotech SL	blood sample	sent by post	Spain[30]
SGI-DNA	blood sample	sent by post	USA[31]


Human genome (30×)
Human genome (1×)
Human exome (30×)
Human exome (1×)
Illumina HiSeq 2000
Illumina HiSeq 2500 Rapid Run
Illumina GAIIx
Solid 5500xl
Ion torrent PGM 318 Chip
PacBio RS II
454 FLX titanium XL+
Number of reads per run
Read length (bases)
10^{11}
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
10^{5}
10^{4}
100
1,000

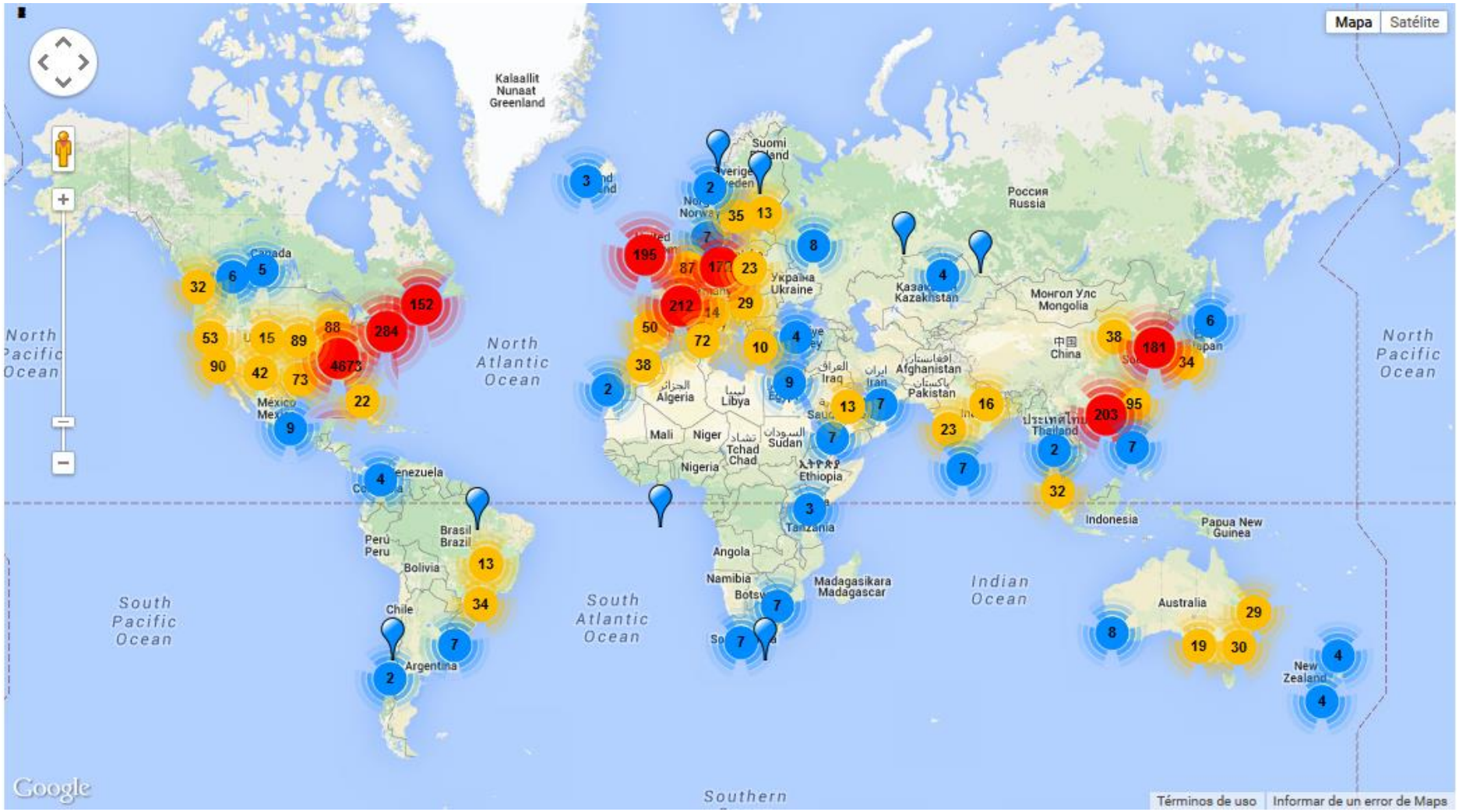
Número de bases secuenciadas



Software de análisis de secuencias cortas



Mapa mundial de secuenciadores de segunda generación

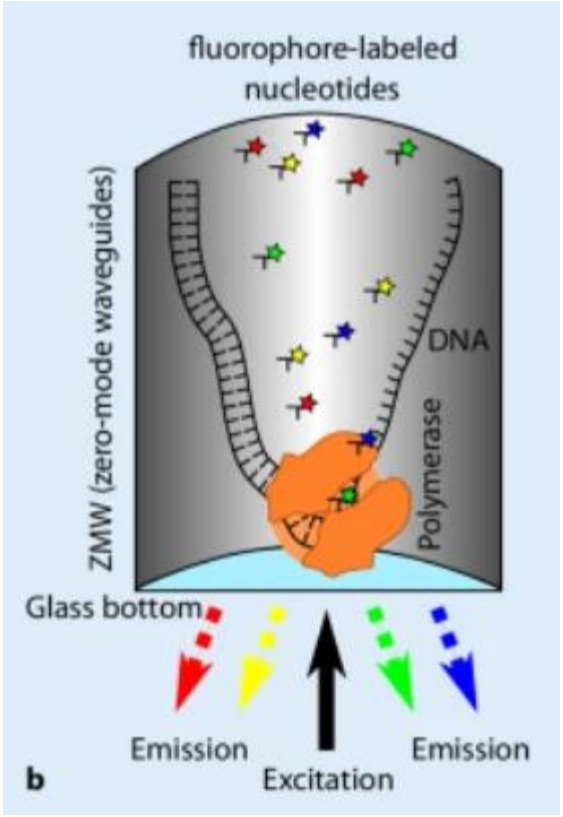


Secuenciadores de tercera generación

- Secuenciadores de molécula única
  - No hay amplificación de la muestra
  - Mayor tamaño de las lecturas
  - Actualmente mayor costo por base

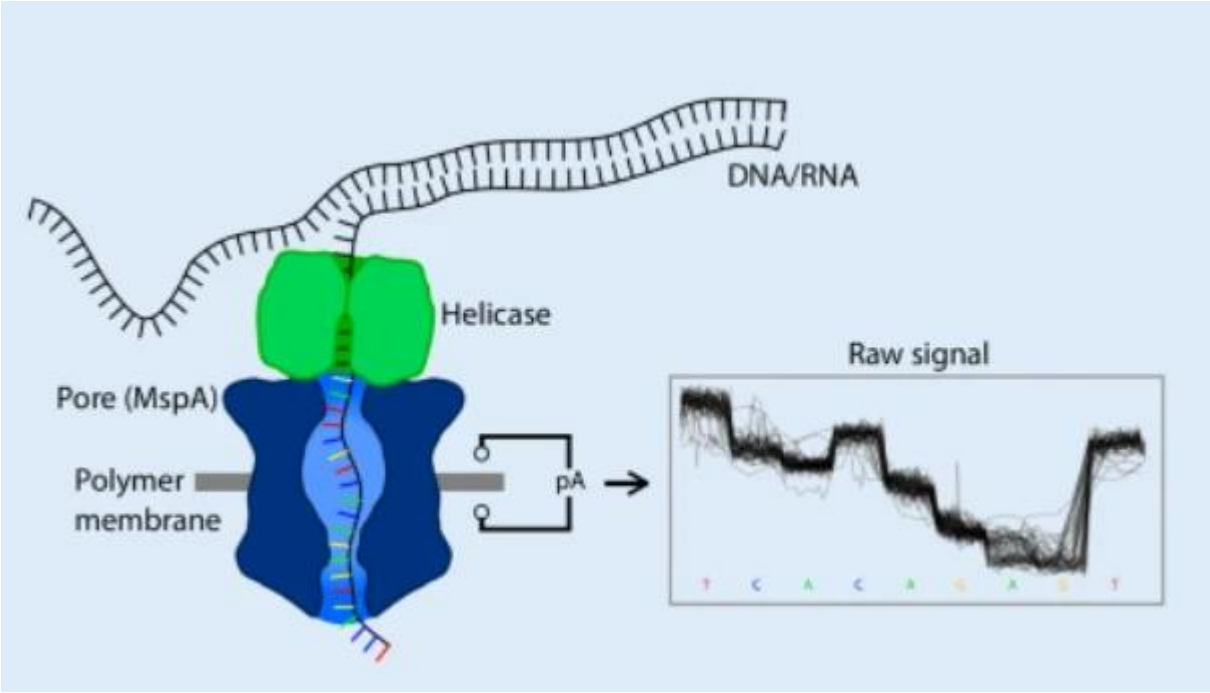
PacBio

- single-molecule real-time (SMRT) sequencing
- Polimerasa unida a un pocillo



Nanopore

- pasaje a través de un poro proteico modificado
- Se mide la variación de la corriente eléctrica a través del poro
- 450 bases por segundo



The MinION™ device: a miniaturised sensing system

Nanopore sensing technology can be miniaturised into a portable device for electronic single-molecule sensing.

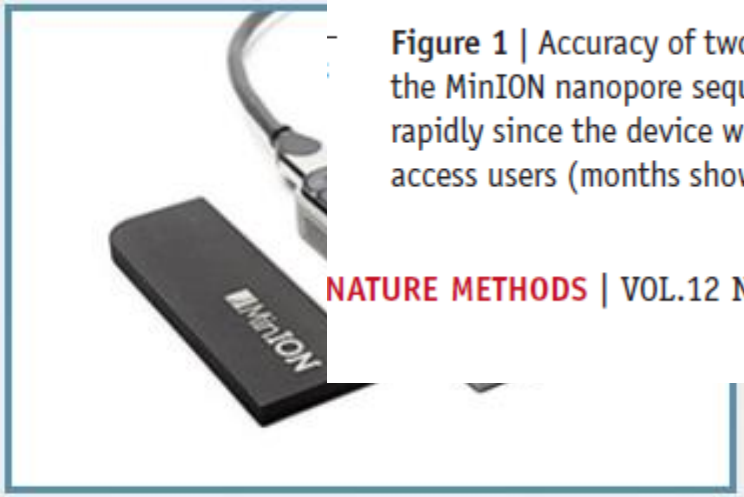
The MinION™ device is a small instrument that is compatible with consumable flow cells containing the proprietary sensor chip, Application-Specific Integrated Circuit (ASIC) and nanopores that are needed to perform a complete single-molecule sensing experiment. Plugging directly into a laptop or desktop computer through a USB port, it is a self-contained device to deliver real-time experimental data.

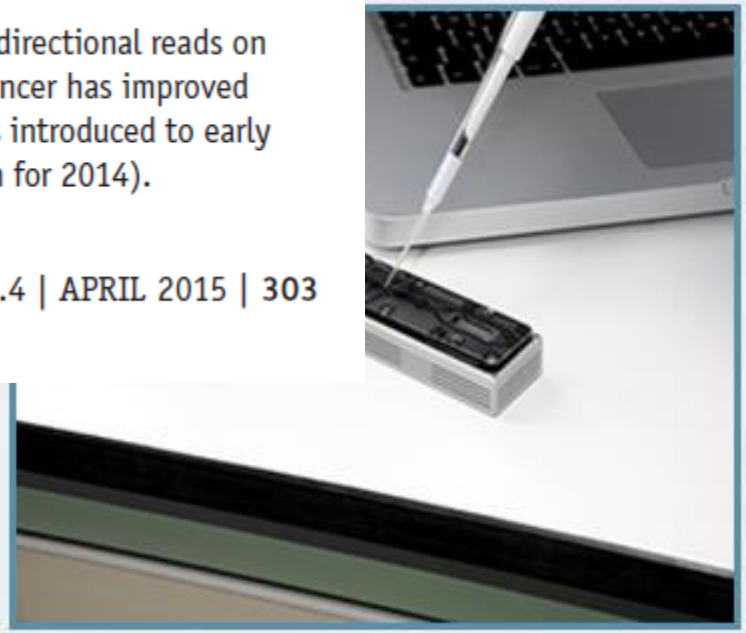
The MinION device is adapta ng techniques. Currently, several hundred participants in the Min xplore how its features - including long read lengths, real time di er a range of biological questions. The MAP initially focuses on D

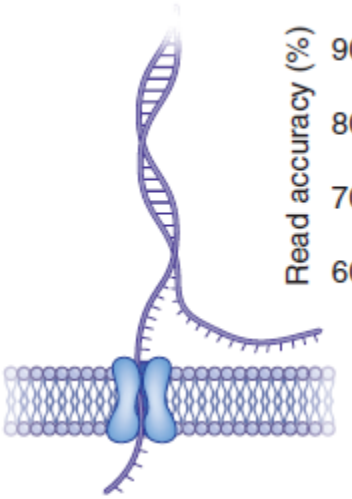
The MinION is operated using calling in real time in the cloud, with the option to perform these an

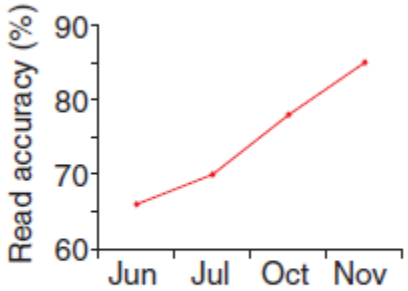
Oxford Nanopore is focused vs. The system is designed to be compatible with complex samp samples.

The MinION device is a mir igned to sense from complex system, designed for singl as blood/serum port of a laptop





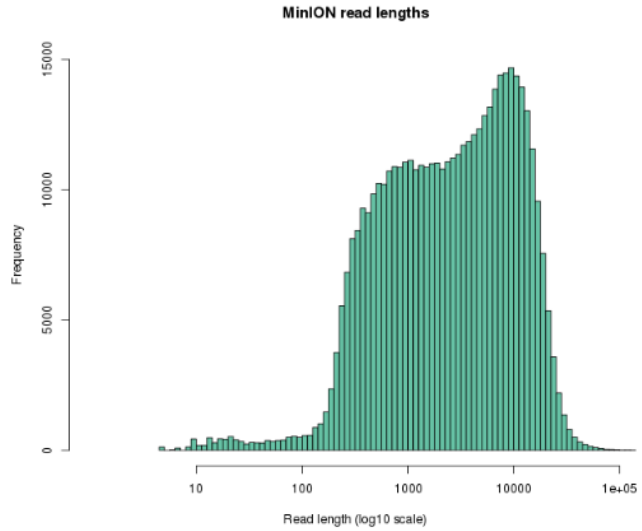


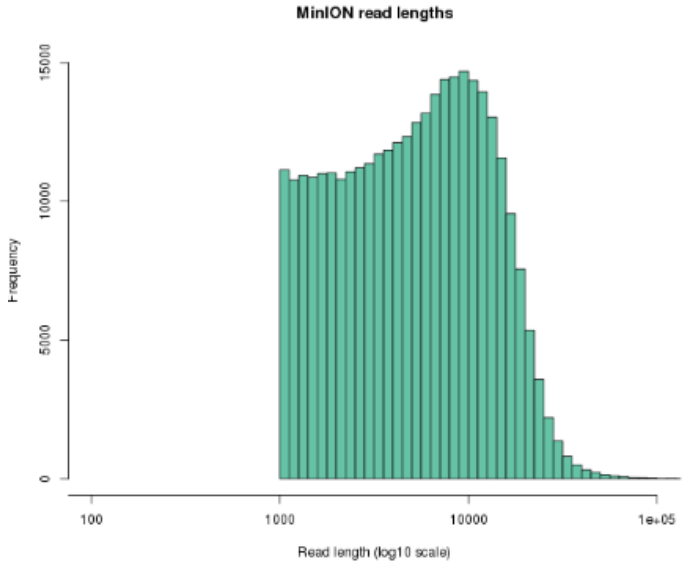


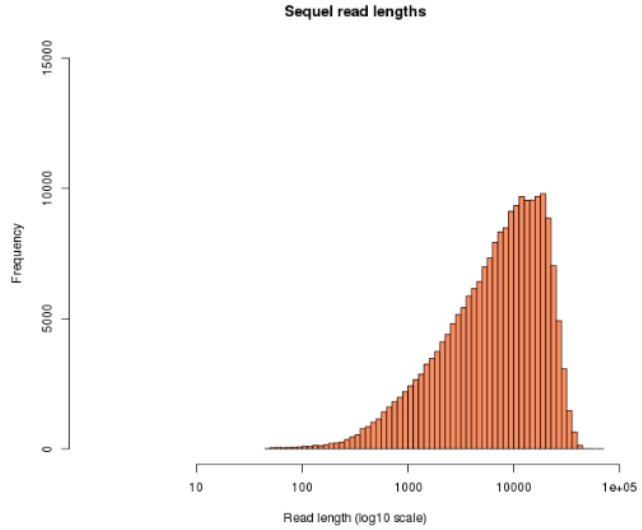
Katie Vicari/Nature Publishing Group

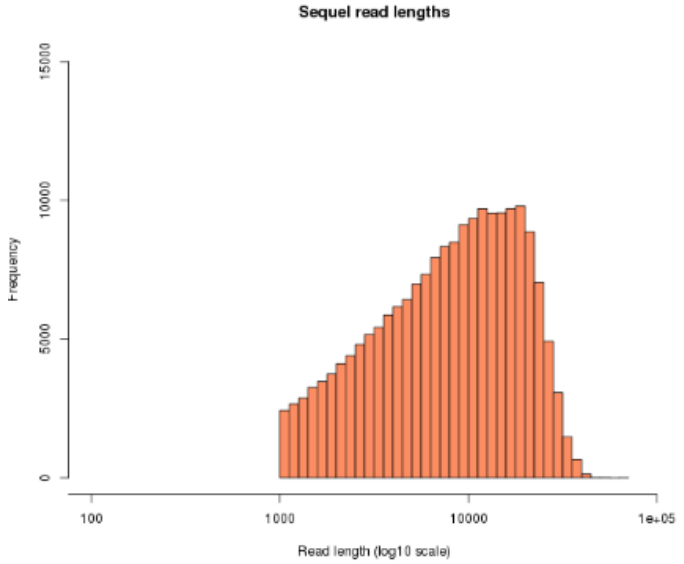
Figure 1 | Accuracy of two-directional reads on the MinION nanopore sequencer has improved rapidly since the device was introduced to early access users (months shown for 2014).

Comparación de las plataformas









Comparación de las plataformas

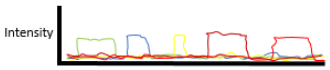
Table 1 Comparison of long-read sequencing methods

	Device	Device costs	Output max (avg) in Gb	Read length (avg/max)	Costs per Gb	Run time	Accuracy		Multiplexing capacity
							Long read	Consensus read	
ONT	Flongle	$	2 (1)	5–35 kb/>2 Mb	$$$	0.5–48 h	>Q10	>Q30	96
	MinION	$	30 (15)		$$				
	GridION	$$	150 (75)		$$				
	PromethION	$$$	15 Tb (4–6 Tb)		$	0.5–72 h			
PacBio	RS II	$$$$	2 (1)	5–15 kb/>60 kb	$$$$	0.5–6 h	>Q10	–	384
	Sequel	$$$	50[a]/20[b] (8–10)	5–30 kb/>200 kb	$$$	0.5–20 h		>Q30	
	Sequel II	?	300[a]/100[b] (?)		?	0.5–30 h			

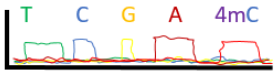
PacBio SMRT seq

DNA passes thru polymerase in an illuminated volume

Raw output is fluorescent signal of the nucleotide incorporation, specific to each nucleotide

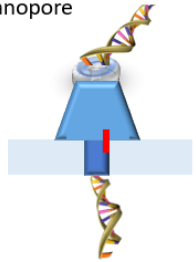


A,C,T,G have known pulse durations, which are used to infer methylated nucleotides

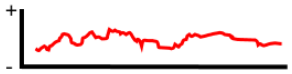


Oxford Nanopore

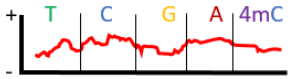
DNA passes thru nanopore



Raw output is electrical signal caused by nucleotide blocking ion flow in nanopore



Each nucleotide has a specific electric "signature"

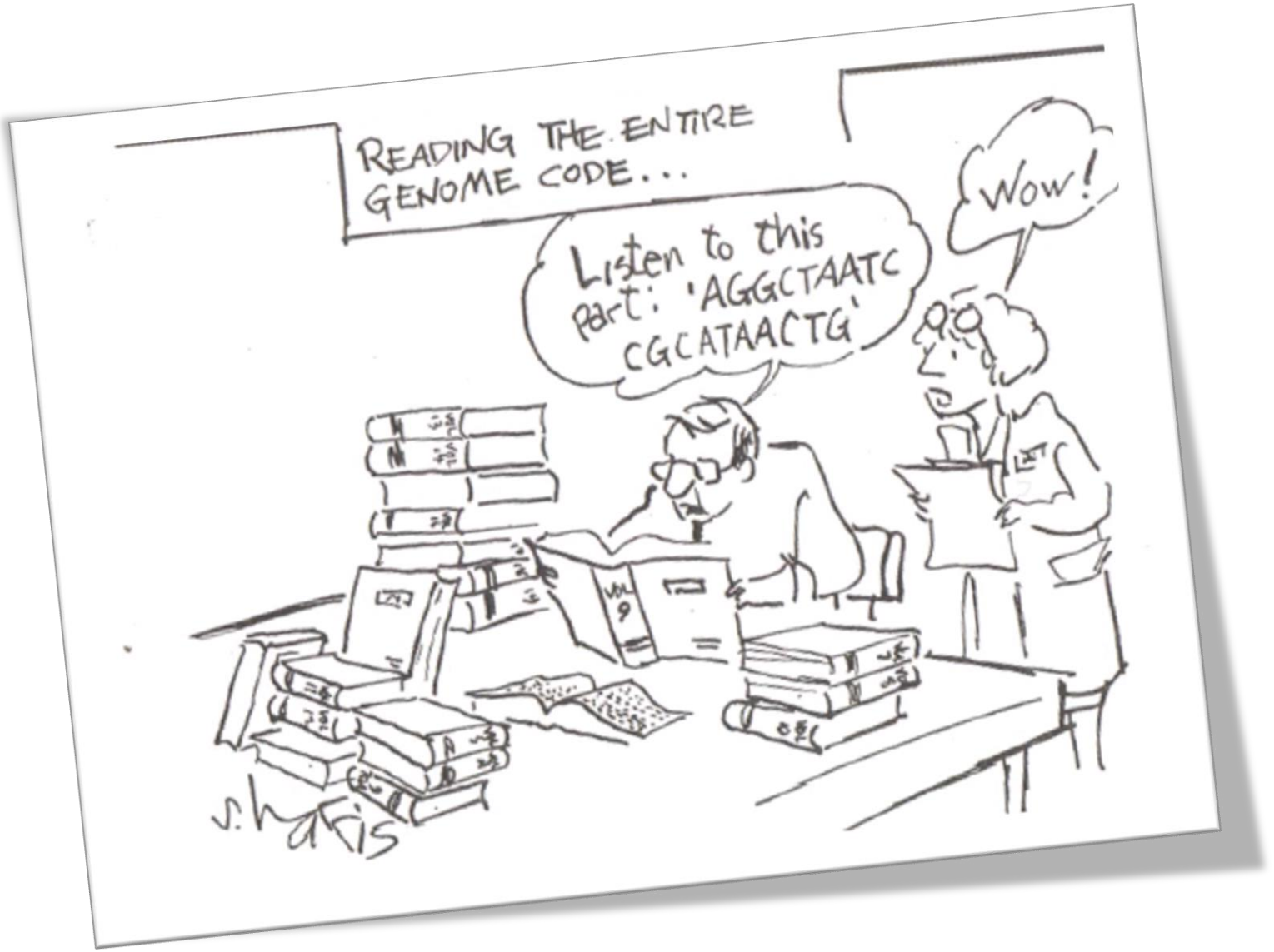


doi.org/10.1007/s11825-019-0249-z

Software de análisis de secuencias largas

Application	Tool		URL
Basecalling	Guppy	N	https://community.nanoporetech.com/downloads
	SMRT Analysis	P	https://www.pacb.com/support/software-downloads/
Alignment	BLASR	P	https://github.com/PacificBiosciences/blasr
	LAST	N/P	http://last.cbrc.jp/
	minimap2	N/P	https://github.com/lh3/minimap2
	NGMLR	N/P	https://github.com/philres/ngmlr
Assembly	Canu	N/P	https://github.com/marbl/canu
	FALCON	N/P	https://github.com/PacificBiosciences/FALCON
	Flye	N/P	https://github.com/fenderglass/Flye
	wtdbg2	N/P	https://github.com/ruanjue/wtdbg2
Error correction	HALC	P	https://github.com/lanl001/halc
	Medeka	N	https://github.com/nanoporetech/medaka
	Nanopolish	N	https://github.com/jts/nanopolish
SV calling	NpInv	N	https://github.com/haojingshao/npInv
	Pbsv	P	https://github.com/PacificBiosciences/pbsv
	Sniffles	N/P	https://github.com/fritzsedlazeck/Sniffles
	SVIM	N/P	https://github.com/eldariont/svim
SNP calling	DeepVariant	N/P	https://github.com/google/deepvariant
	GATK	N/P	https://software.broadinstitute.org/gatk/
	Medeka	N	https://github.com/nanoporetech/medaka
	Nanopolish	N	https://github.com/jts/nanopolish
DNA/RNA modifications	Nanopolish	N	https://github.com/jts/nanopolish
	SMRT Analysis	P	https://www.pacb.com/support/software-downloads/
	Tombo	N	https://github.com/nanoporetech/tombo
Tandem repeat analysis	NanoSatellite	N	https://github.com/arnederoeck/NanoSatellite
	nanoSTRique	N	–
	Pbsv	P	https://github.com/PacificBiosciences/pbsv
	STRetch	P	https://github.com/Oshlack/STRetch

Que hacemos con la secuencia??



Formato Fastq

- Formato de salida de los secuenciadores de nueva generación
  - Incluye información de secuencia
  - Incluye información de calidad para cada base (confiabilidad)

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

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

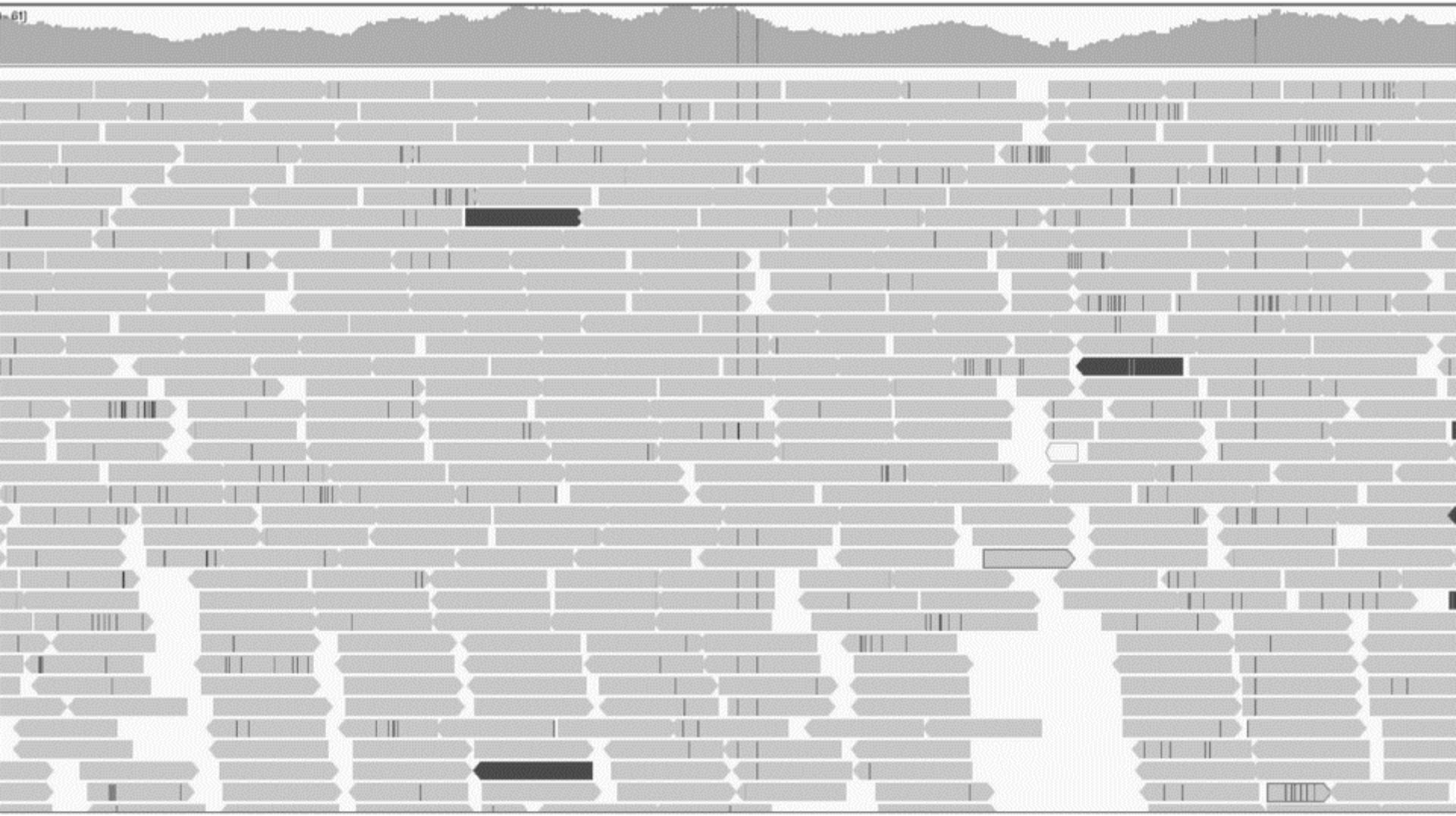
Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	Y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

Resecuenciado de Genomas

- Comparación a una referencia
  - Se obtienen las lecturas y se comparan a una referencia
    - Búsqueda de variantes, reordenamientos, etc

Mapeo de lecturas



Resecuenciado de Genomas

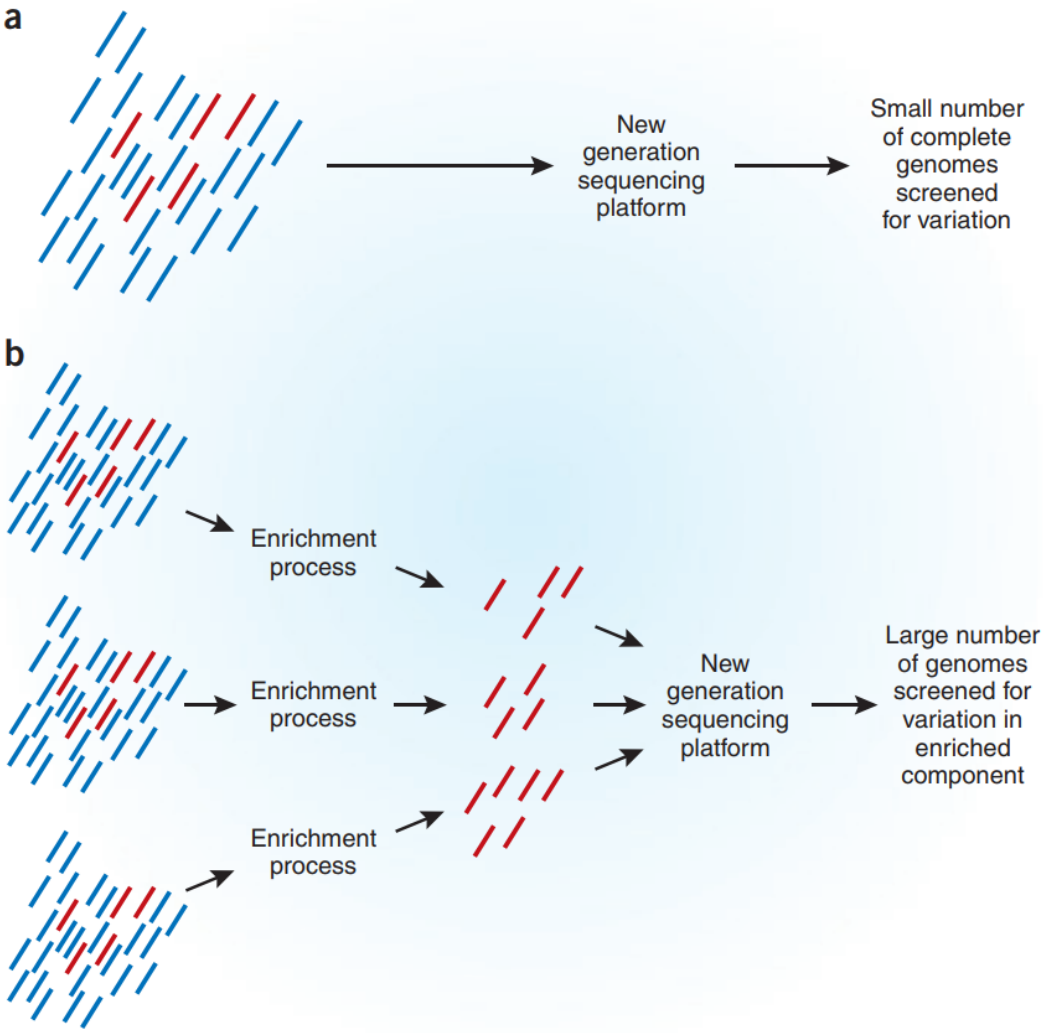


Figure 1 Enrichment procedures allow redistribution of sequencing throughput from all of a small number of genomes (**a**) to a small component of a large number of genomes (**b**).

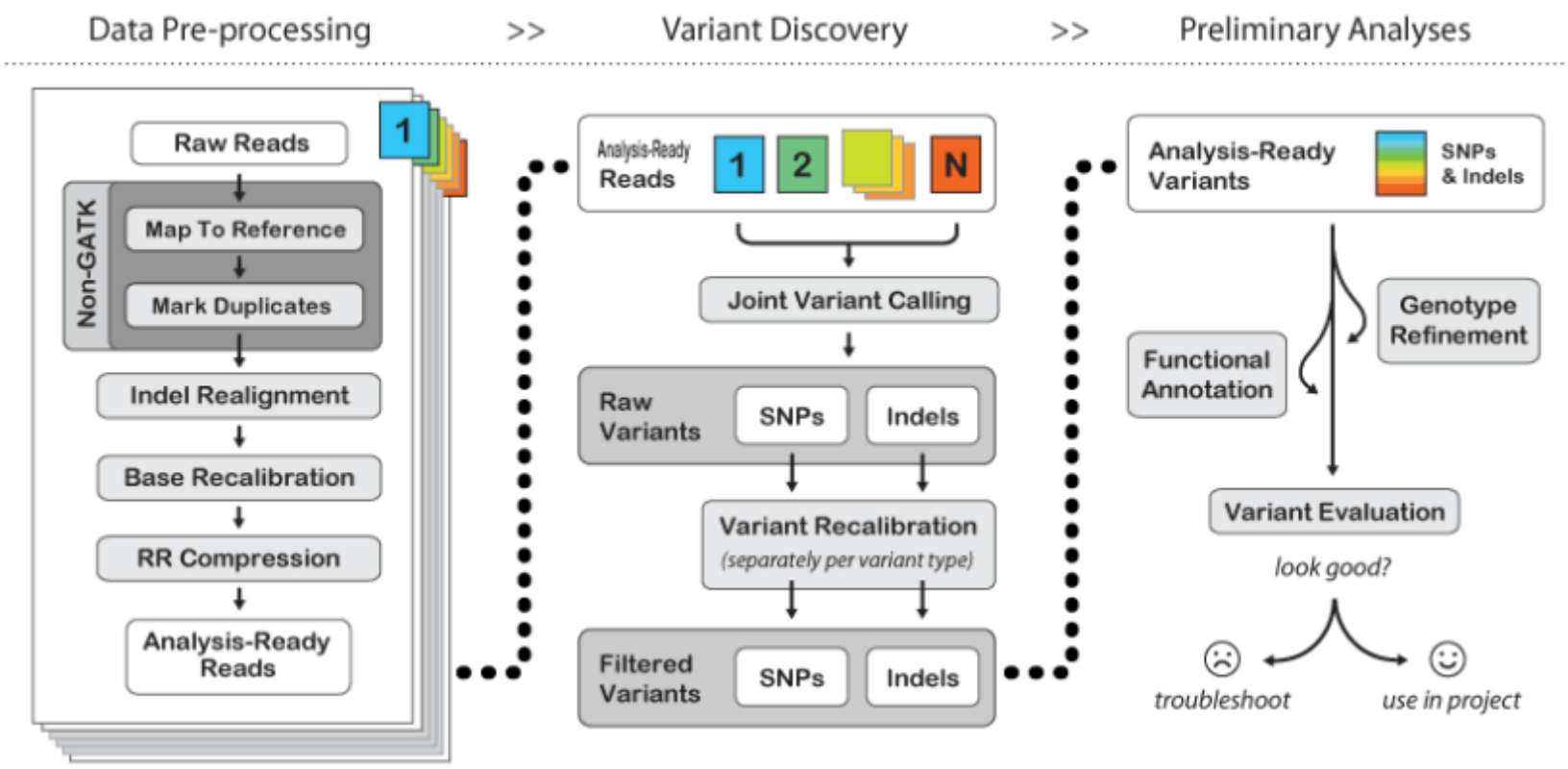
Stratton Nature (2008)

Resecuenciado de Genomas

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001
NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:4
8:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:4
9:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:2
1:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:5
4:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTCT   G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:3
5:4       0/2:17:2       1/1:40:3
```

1. #CHROM
2. POS
3. ID
4. REF
5. ALT
6. QUAL
7. FILTER
8. INFO

Resecuenciado de Genomas



Resecuenciado de Genomas

Effect prediction details

Detailed desription of the effect predicted by SnpEff in the `Effect` and `Effect_Impact` sub-fields.

Notes:

- **Effect (Sequence Ontology)** Sequence ontology (SO) allows to standardize terminology used for assessing sequence changes and impact. This allows for a common language across all variant annotation programs and makes it easier to communicate using a uniform terminology. Starting from version 4.0 VCF output uses SO terms by default.
- **Effect (Classic)** This are the "classic" effect names usd by SnpEff, these can be accessed using the `-classic` command line option.
- **Effect impact** Effects are categorized by 'impact': {High, Moderate, Low, Modifier}. This are pre-defined categories to help users find more significant variants.

  ⚠ Impact categories must be used with care, they were created only to help and simplify the filtering process. Obviously, there is no way to predict whether a "high impact" or a "low impact" variant is the one producing a phenotype of interest.

Here is a list of effects and some brief explanations:

Effect Seq. Ontology	Effect Classic	Note & Example	Impact
coding_sequence_variant	CDS	The variant hits a CDS.	MODIFIER
chromosome	CHROMOSOME_LARGE DELETION	A large part (over 1% or 1,000,000 bases) of the chromosome was deleted.	HIGH
duplication	CHROMOSOME_LARGE_DUPLICATION	Duplication of a large chromoome segment (over 1% or 1,000,000 bases).	HIGH
inversion	CHROMOSOME_LARGE_INVERSION	Inversion of a large chromoome segment (over 1% or 1,000,000 bases).	HIGH
coding_sequence_variant	CODON_CHANGE	One or many codons are changed e.g.: An MNP of size multiple of 3	LOW
inframe_insertion	CODON_INSERTION	One or many codons are inserted	MODERATE

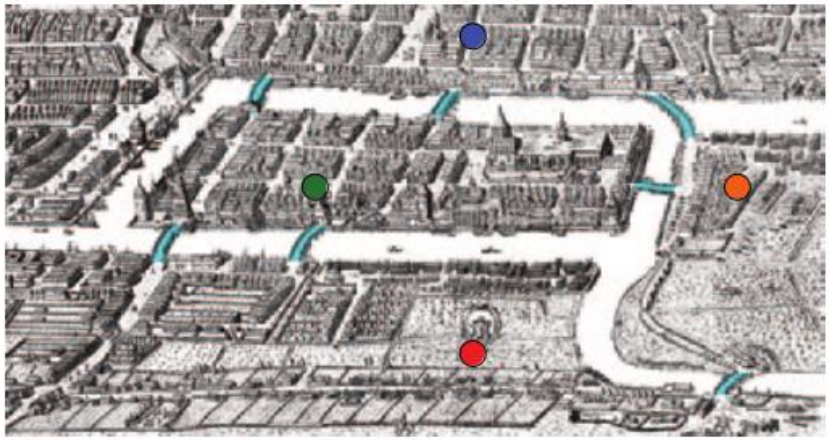
Ensamblado de secuencias cortas

- Existen ensambladores específicos
  - VELVET
  - Abyss
  - SOAPdenovo
  - …
- La mayoría se basan en encontrar el camino euleriano en un grafo de de Bruijn construido con k-mers solapantes obtenidos de las lecturas…

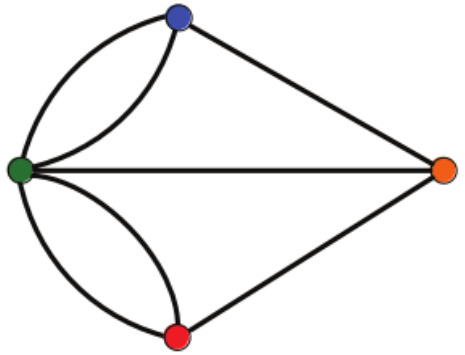
Ensamblando secuencias cortas

- Problema de los puentes de Königsberg
  - Leonhard Euler (1736)

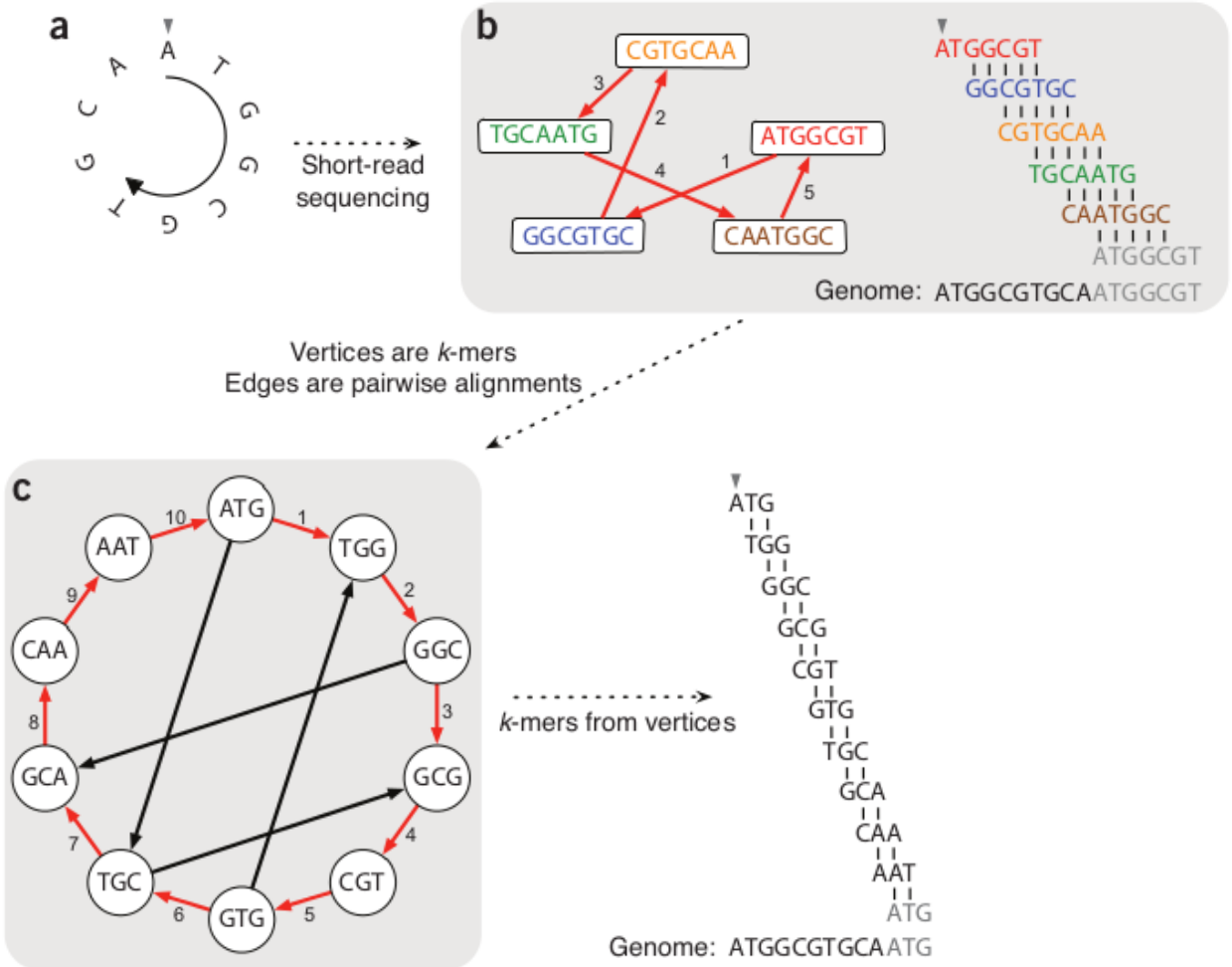
a



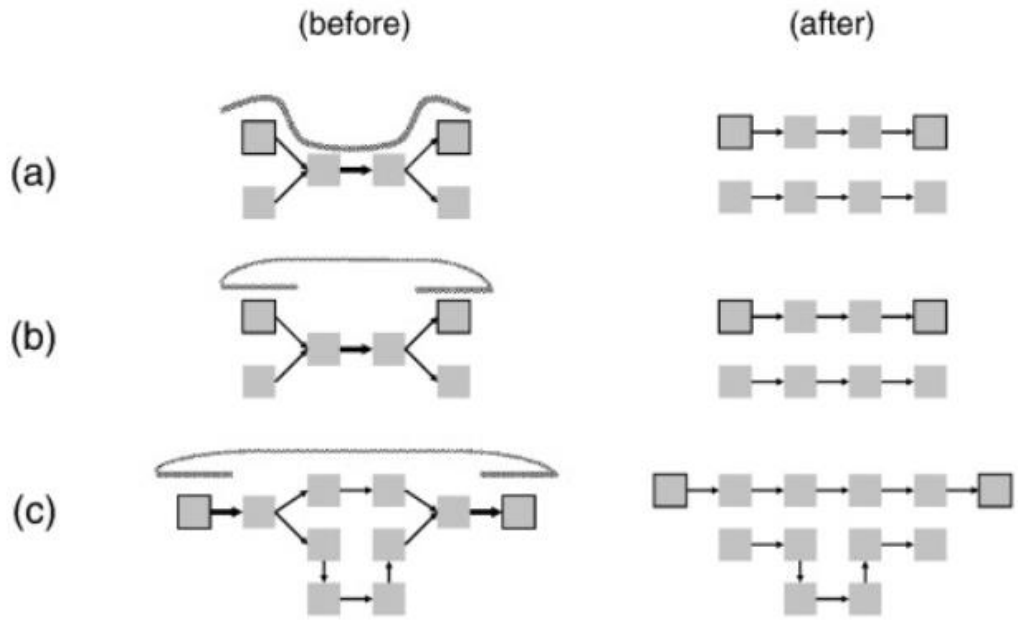
b



Ensamblando secuencias cortas



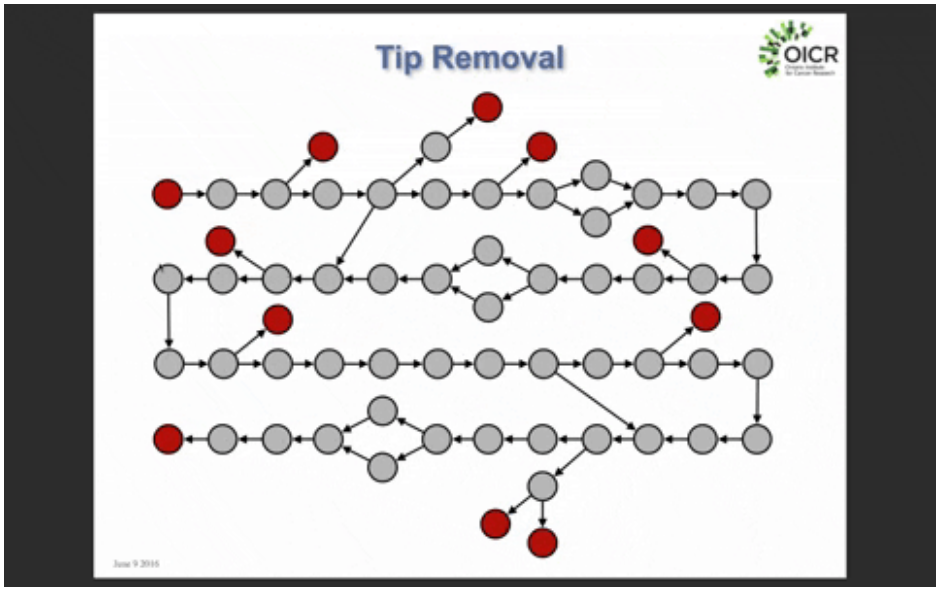
Resolución de nodos problematicos



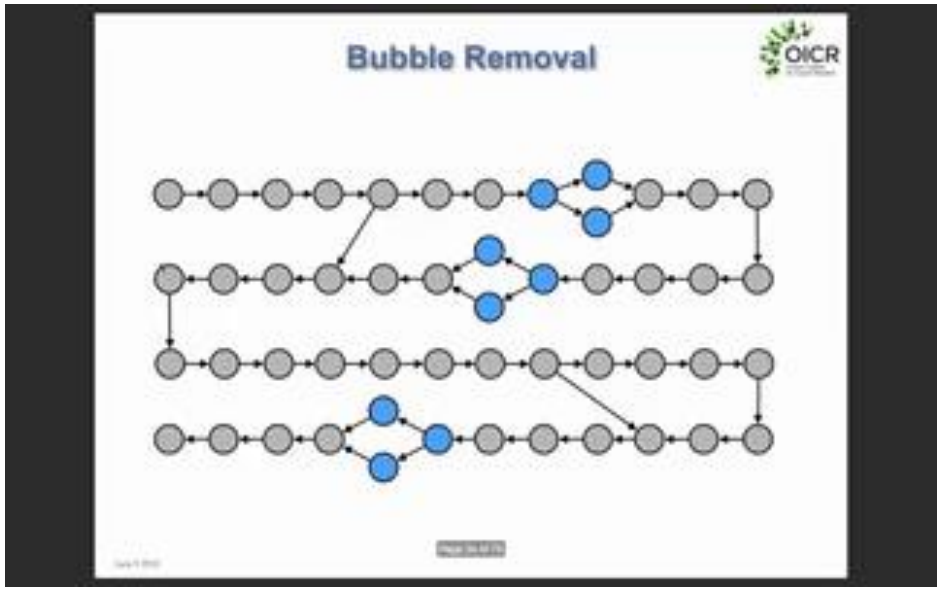
- Mapeo de las lecturas al grafo:
  - (a) Se mapea la lectura la cual atraviesa el nodo complejo
  - (b) Las lecturas pareadas se usan para resolver
  - (c) La distancia entre los pares se usan para resolver

Simplificación del grafo

Remoción de puntas



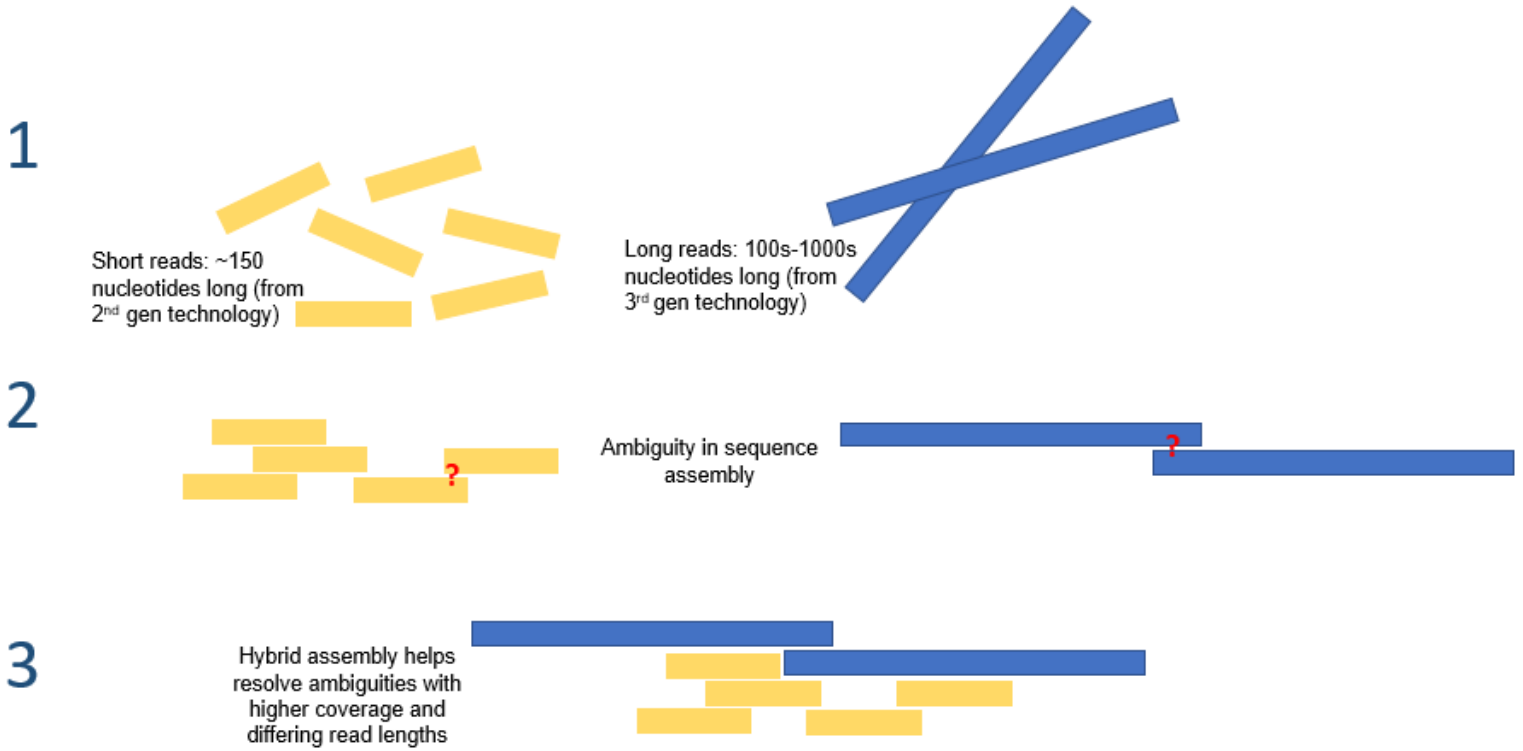
Remoción de burbujas



- Mayormente basado en la profundidad de lecturas de los nodos

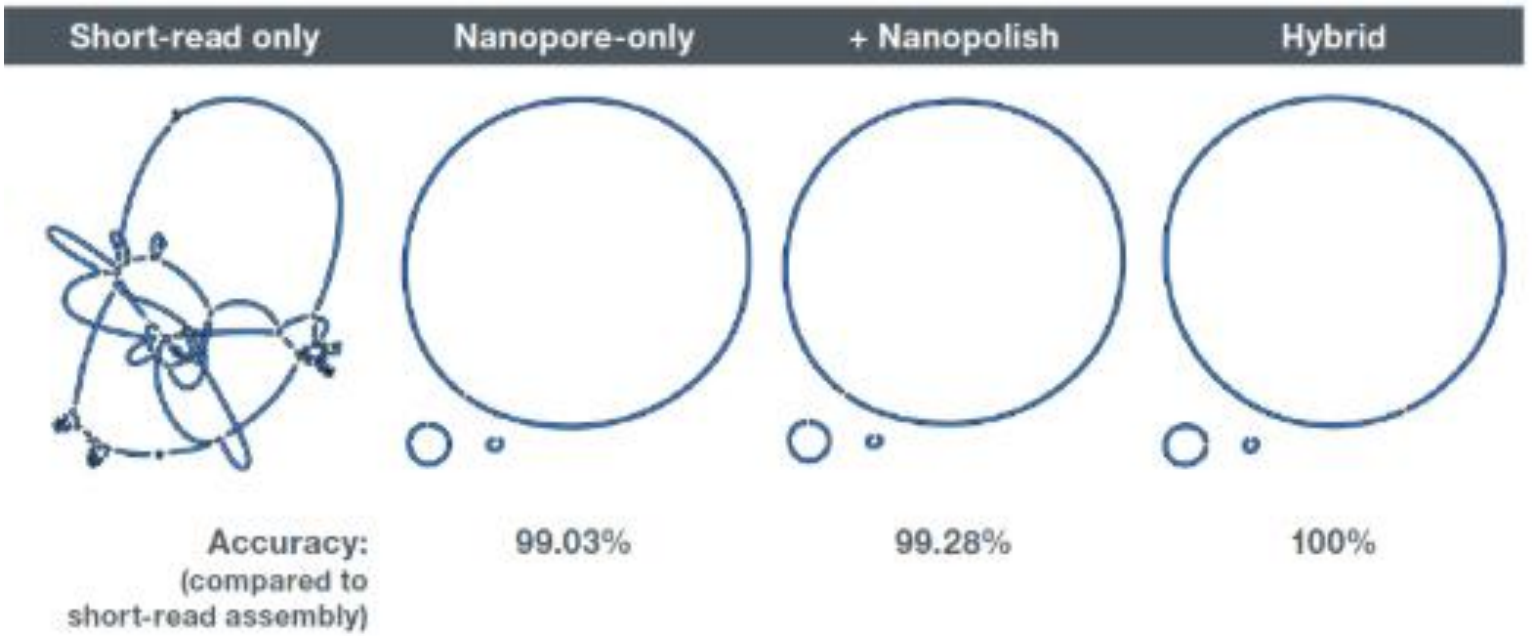
Ensamblado con lecturas largas

- Único
- Hibrido

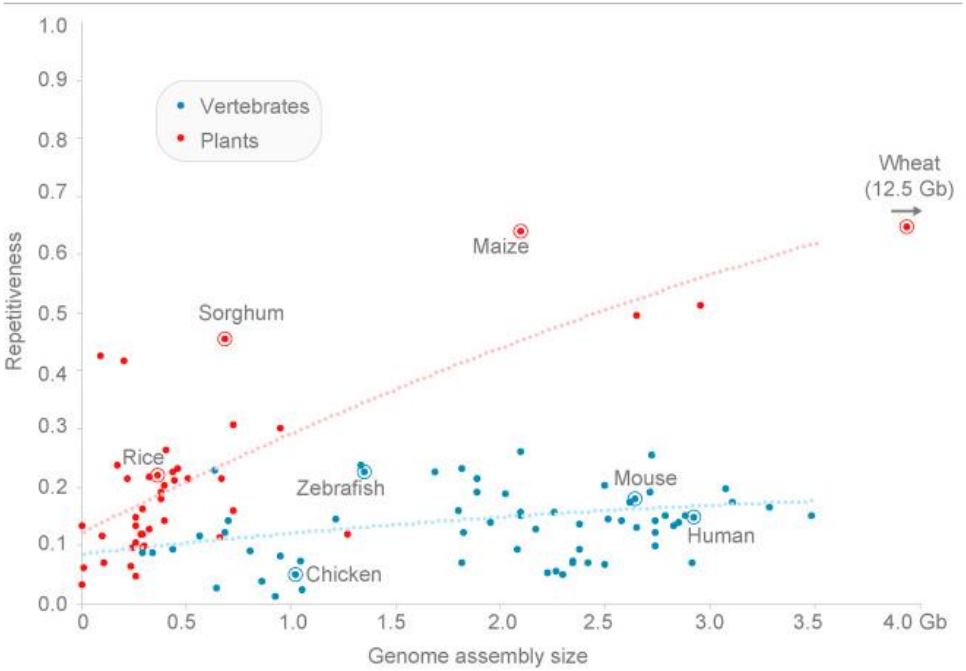


Ensamblado con lecturas largas

- Único
- Hibrido



Ensamblado con lecturas largas

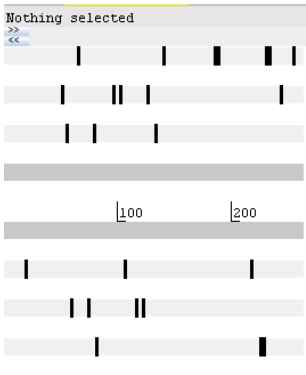


- Gran impacto en genomas muy repetitivos

El ensambaldo es el comienzo del proceso

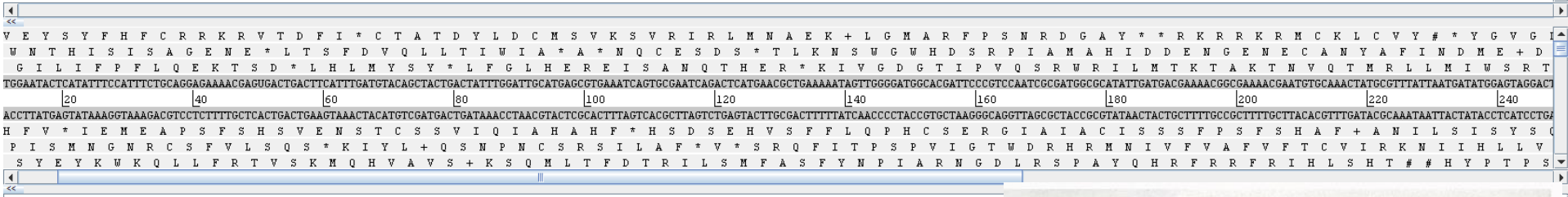
- Búsqueda de genes
  - Codificantes para proteínas
  - Pseudogenes
  - Genes para ARNs funcionales
    --tRNA
    --rRNA
    --snoRNA
    --snRNA
    --miRNA
  - Regiones reguladoras
  - Otras regiones funcionales

TCTGTGGATGTGGAATAC
TCATATTTCCATTTCTGC
AGGAGAAAACGAGTGACT
GACTTCATTTGATGTACA
GCTACTGACTATTTGGAT
TGCATGAGCGTGAAATCA
GTGCGAATCAGACTCATG
AACGCTGAAAAATAGTTG
GGGATGGCACGATTCCCG
TCCAATCGCGATGGCGCA
TATTGATGACGAAAACGG
CGAAAACGAATGTGCAAA
CTATGCGTTTATTAATGA
TATGGAGTAGGACTGGCG
TGAATCTGGTT.....

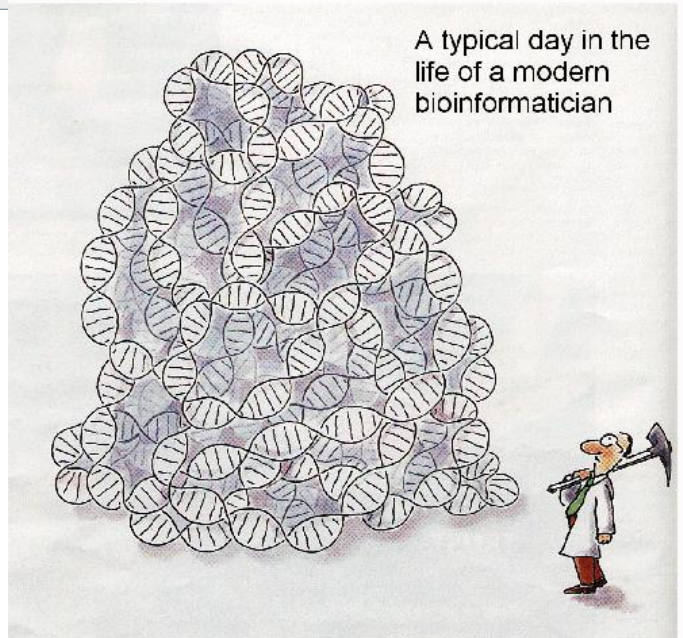


Que defino como gen?

Donde están?



Vamos a intentar encontrar patrones




Entry: GENETEST.SEQ GENETEST.TAB
Nothing selected
CDS
CDS
CDS
100
200
THISISAGENE*
IAMAHIDDENGENECANYAFINDME+
HERE IS ANOTHER*
20
40
60
80
100
120
140
160
180
200
220
240
CDS 17 52
CDS 93 134
CDS 167 244

Anotación del genoma secuenciado:

- Los genes para ARNs funcionales son difíciles de encontrar.
  - Pequeños
  - Sin ORFs
- Las regiones codificantes tienen características identificables
  - En procariotas se identifican de manera relativamente sencilla dada la alta densidad de genes
  - En eucariotas es mas complejo debido a que los exones son pequeños comparados con los intrones y regiones intergenicas

Lowe Lab

tRNAscan-SE Search Server

Search for tRNA genes in genomic sequence

---

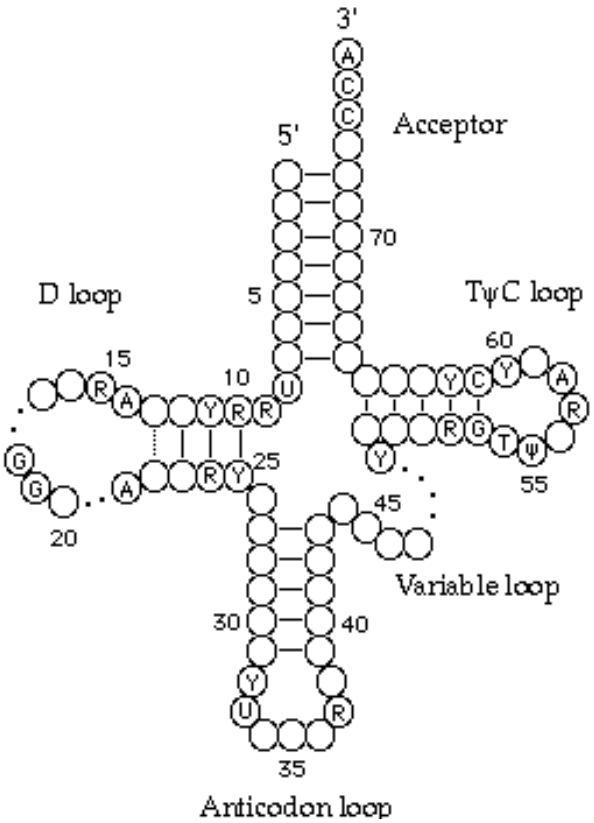
tRNAscan-SE 1.21

The principles underlying the tRNAscan-SE program are described in:

Lowe, T.M. and Eddy, S.R. (1997)
tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.
Nucleic Acids Res, 25, 955-964.

Instructions for using the tRNAscan-SE server and interpreting the output can be found in the tRNAscan-SE README file.

If you would like to run tRNAscan-SE locally, you can get the UNIX source code (gzip'd tar file).



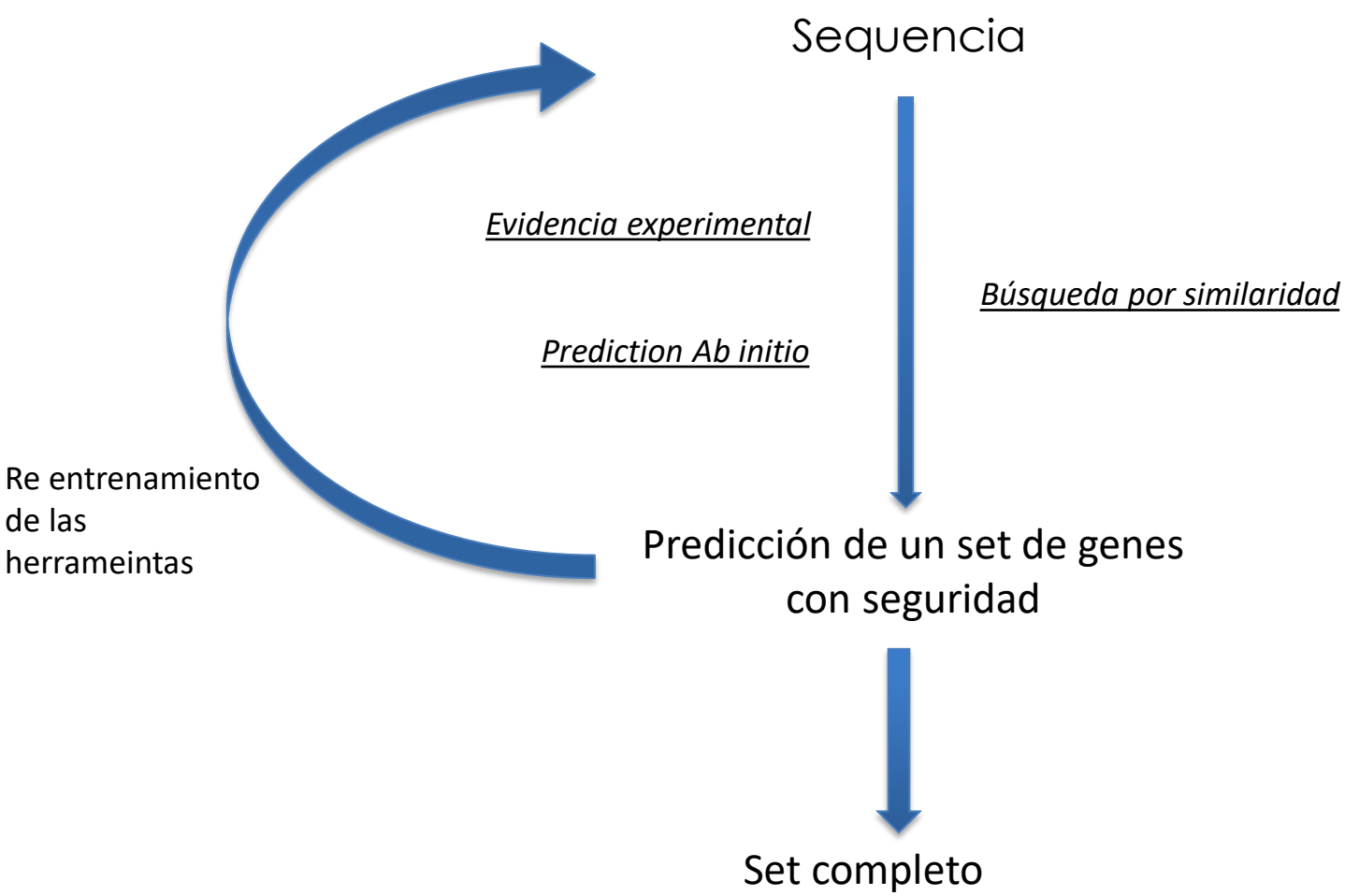
Anotación del genoma secuenciado:

- Búsqueda de ORFs
  - Marcos de lectura abiertos
    - Comienzan con un codon de inicio ATG
    - Finalizan en un codon de terminación
  - Son las **posibles** regiones codificantes

Diferentes aproximaciones

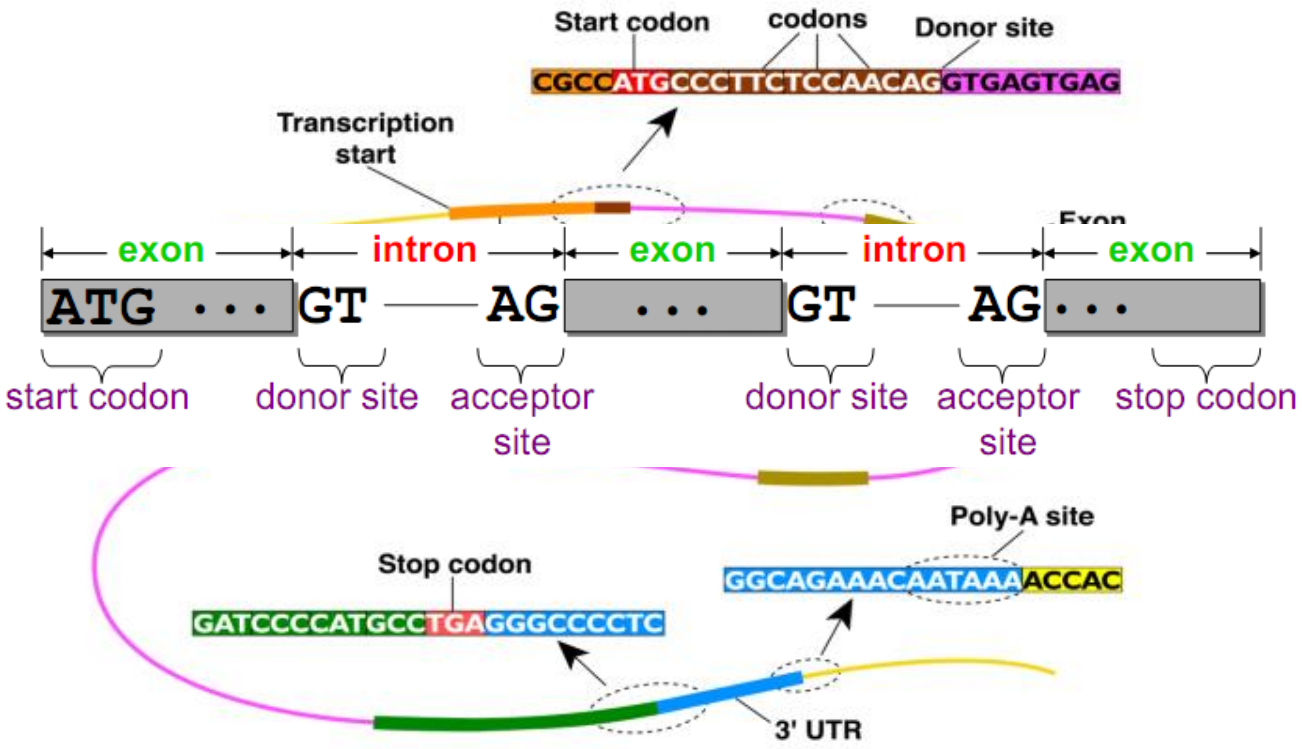
- Métodos predictivos
  - Buscan directamente en la secuencia de acuerdo a reglas definidas
- Métodos comparativos
  - Comparan con otras secuencias y extraen información (conservación)
- Métodos experimentales
  - Utilizan información obtenida de forma experimental

Procedimiento típico



Métodos predictivos

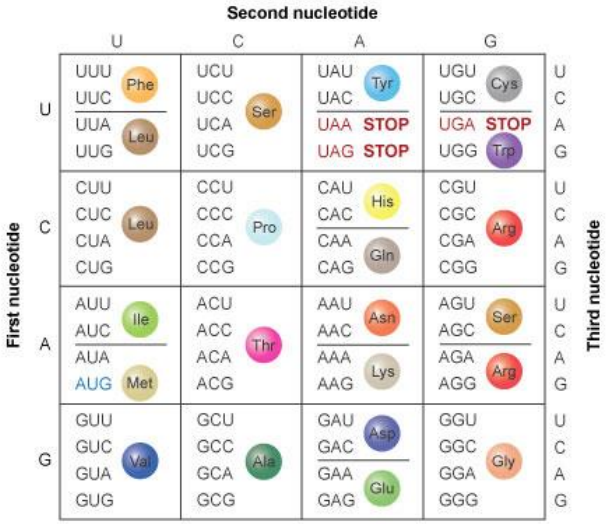
- Búsqueda de señales: motivos cortos delimitan las secuencias codificantes

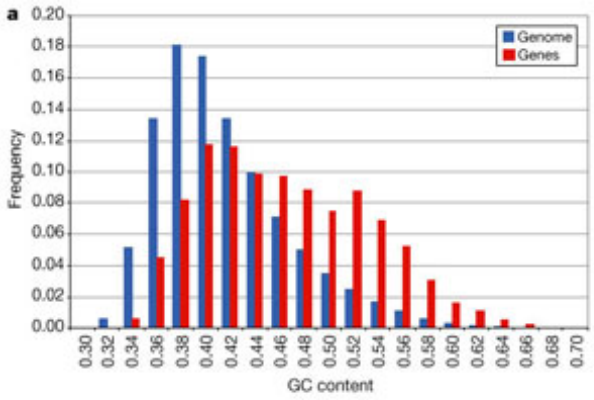


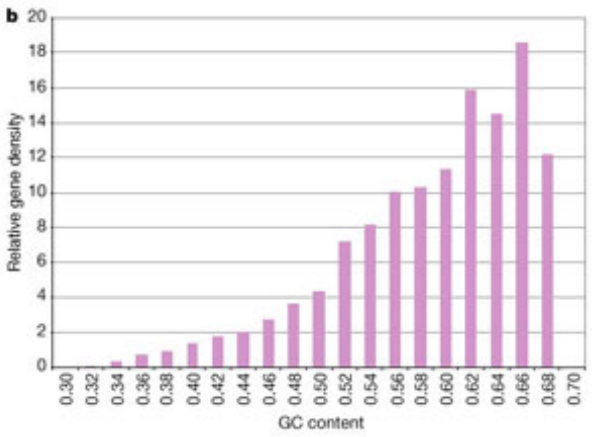
El contenido en GC varía entre los distintos genomas y dentro de un genoma determinado

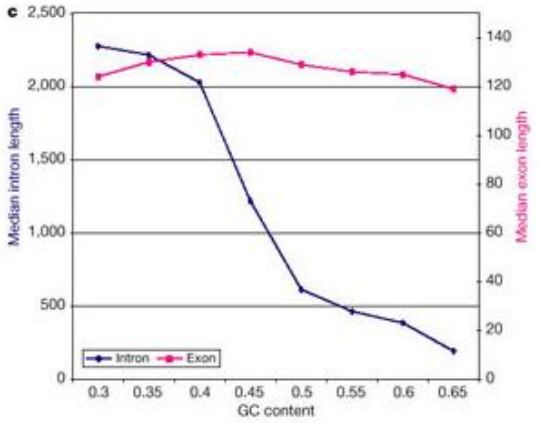
Escherichia coli	51%
Campylobacter jejuni	31%

- La variación esta mas restringida en las regiones codificantes
- Estas regiones tienen un contenido GC mas alto (fenómeno mas claro en genomas ricos en AT)

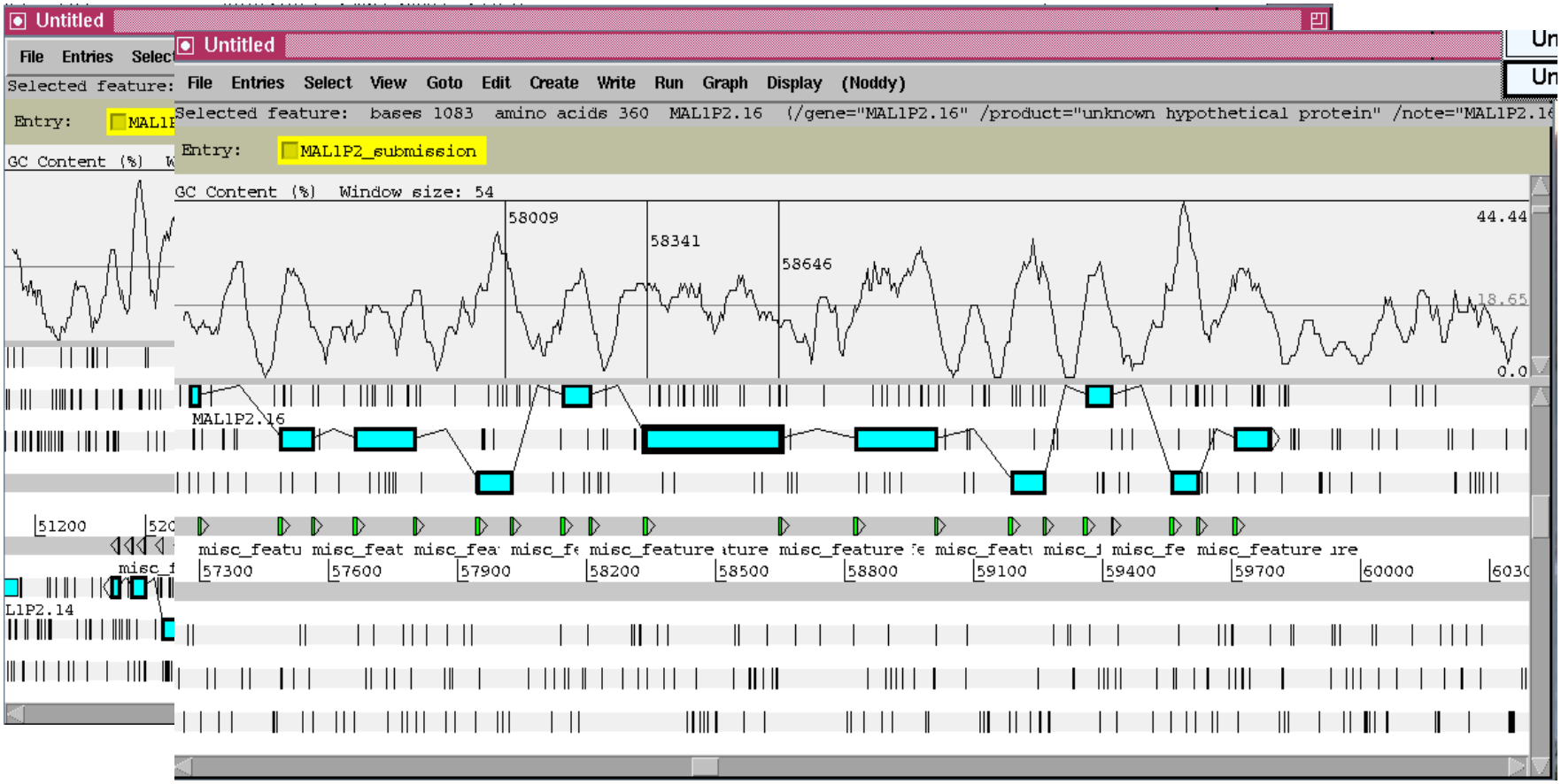








Visualización en Artemis



Métodos predictivos

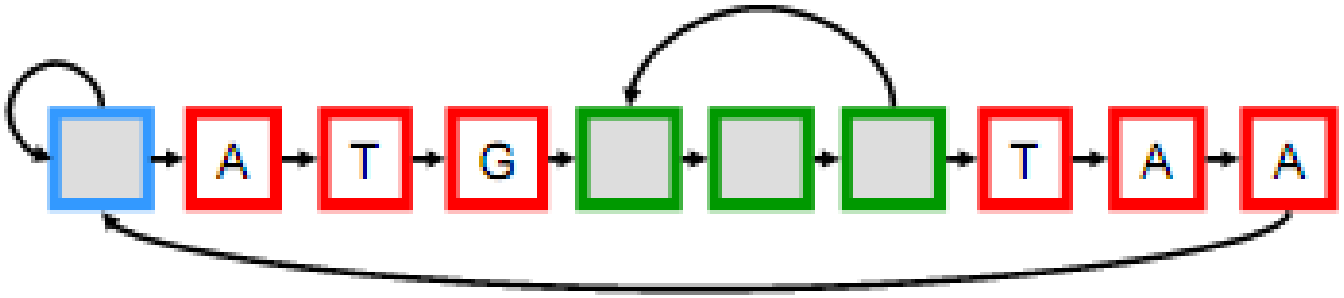
- Los primeros predictores hacían uso de reglas como las que comentamos antes.
  - Problemas:
    - Muchas reglas!
    - Muchas excepciones!
    - Los predictores se hacen demasiados complejos

- El análisis estadístico de las secuencias mejora las predicciones
  - Modelos ocultos de Markov

Modelos estadísticos (HMM)

- Los predictores mas recientes toman en cuenta modelos estadísticos para realizar la anotación
- Los observables son cada una de las bases
- Cada observable puede estar en diferentes estados
  - Exón
  - Intrón
  - Intergénica
- Cada uno de los estados tiene probabilidades de emisión particulares para cada una de las bases
- Los parámetros se ¨aprenden¨ de genes conocidos

Modelos estadísticos (HMM)



AAAGC ATG CAT TTA ACG AGA GCA CAA GGG CTC TAA TGCCG

The sequence of states is an annotation of the generated string – each nucleotide is generated in intergenic, start/stop, coding state

Modelos estadísticos (HMM)

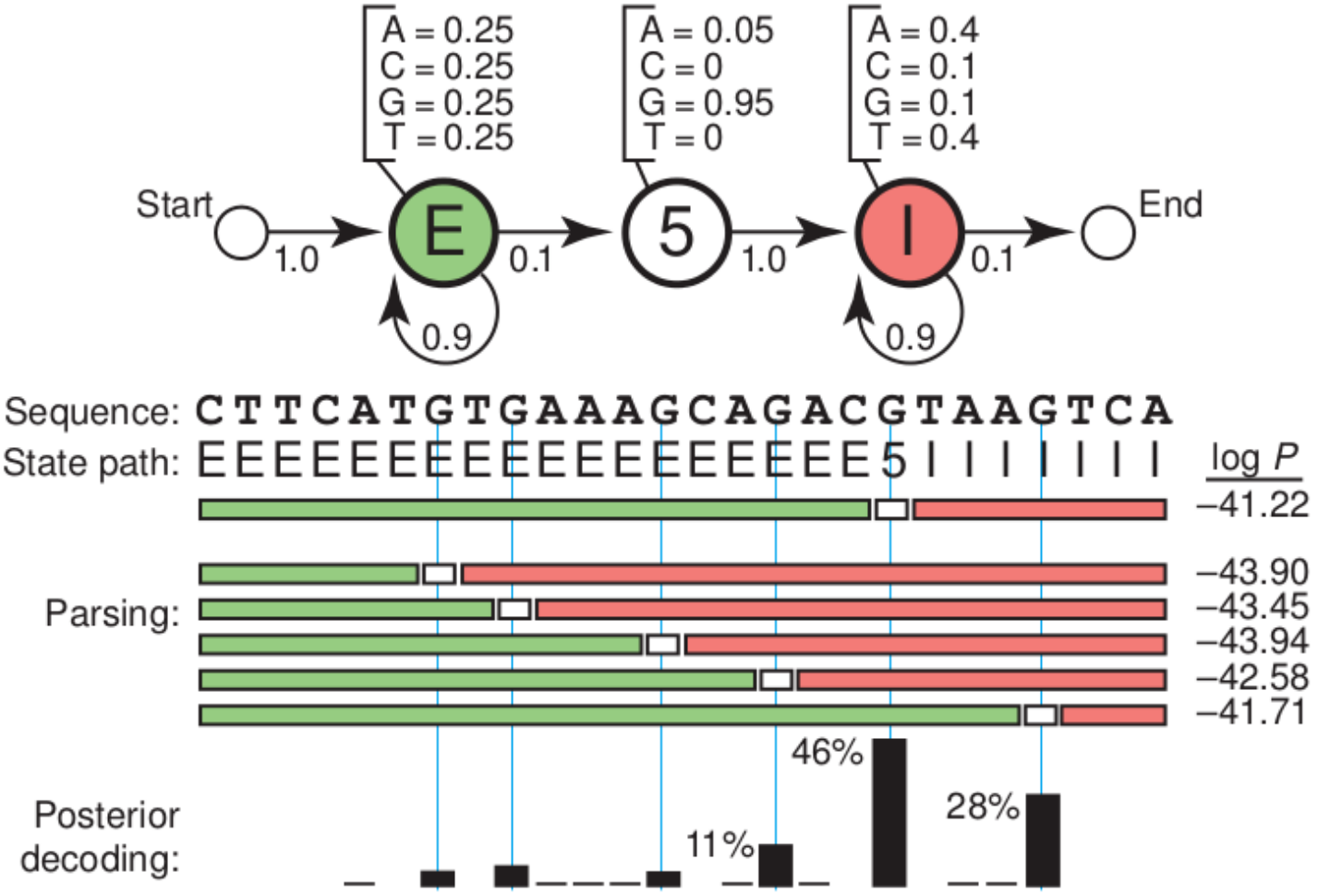


Figure 1 A toy HMM for 5′ splice site recognition. See text for explanation.

Modelos estadísticos (HMM)

```
CGATATTCGATTCTACGCGCGTATACTAGCTTATCTGATC
0111111122222221111112222111111222211110
```

transitions

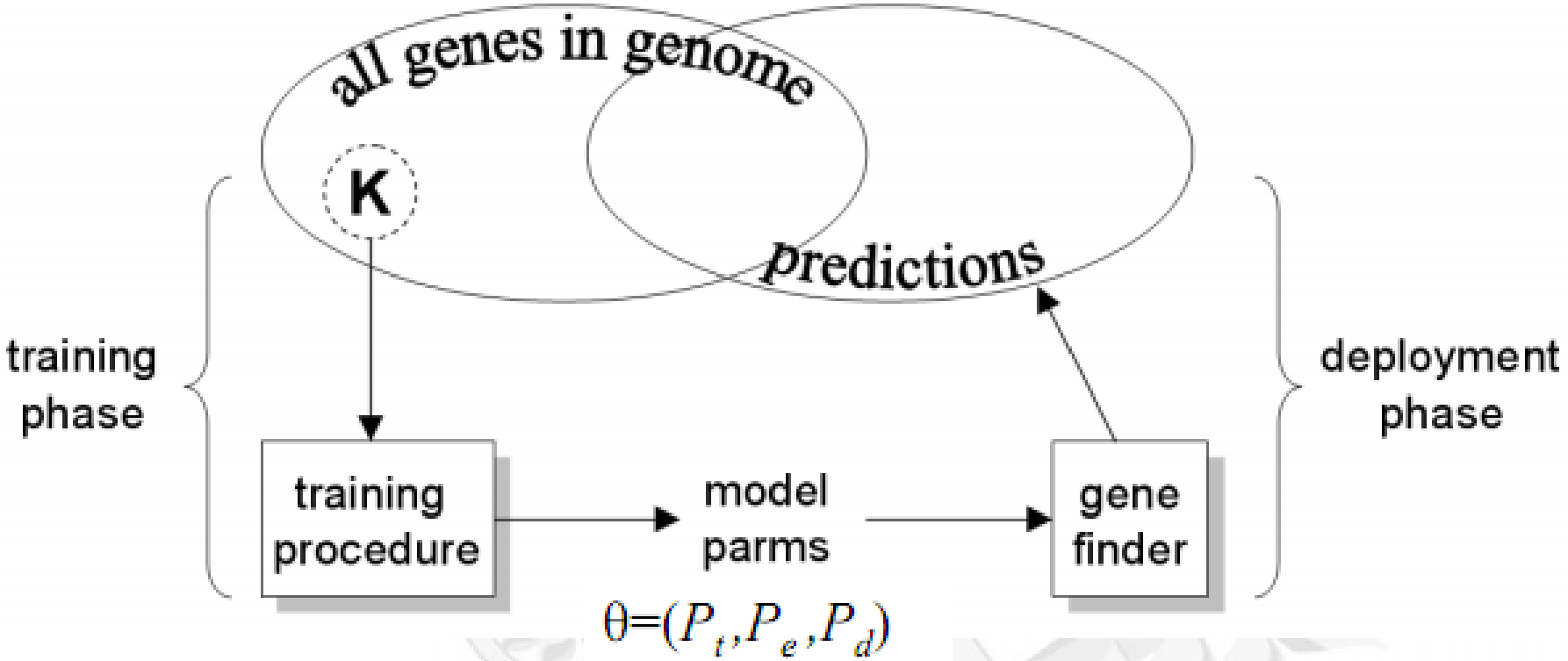
		to state		
		0	**1**	**2**
from state	**0**	0 (0%)	1 (100%)	0 (0%)
	1	1 (4%)	21 (84%)	3 (12%)
	2	0 (0%)	3 (20%)	12 (80%)

emissions

		symbol			
		A	**C**	**G**	**T**
in state	**1**	6 (24%)	7 (28%)	5 (20%)	7 (28%)
	2	3 (20%)	3 (20%)	2 (13%)	7 (47%)

- Los parámetros se ¨aprenden¨ de genes conocidos

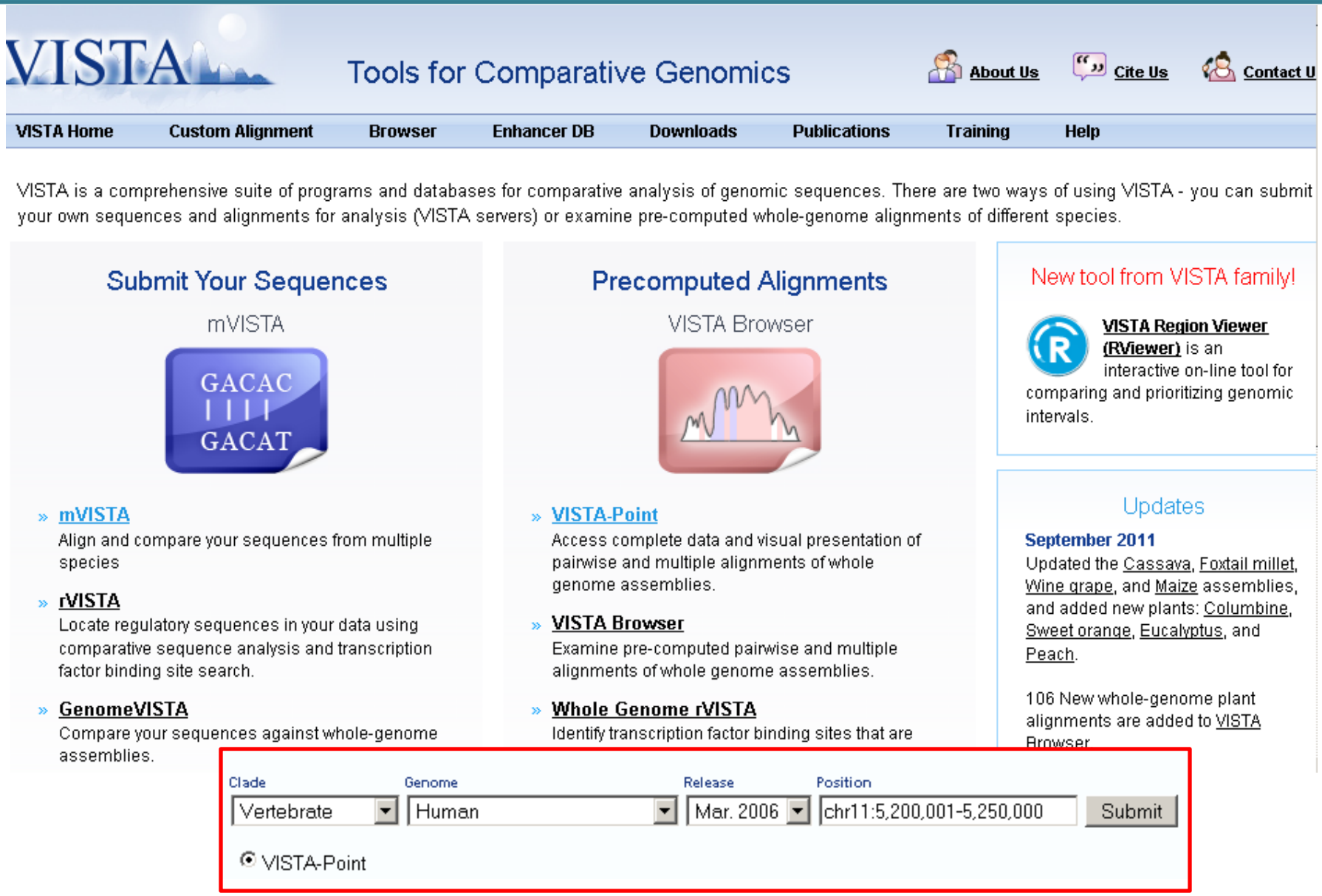
Modelos estadísticos (HMM)



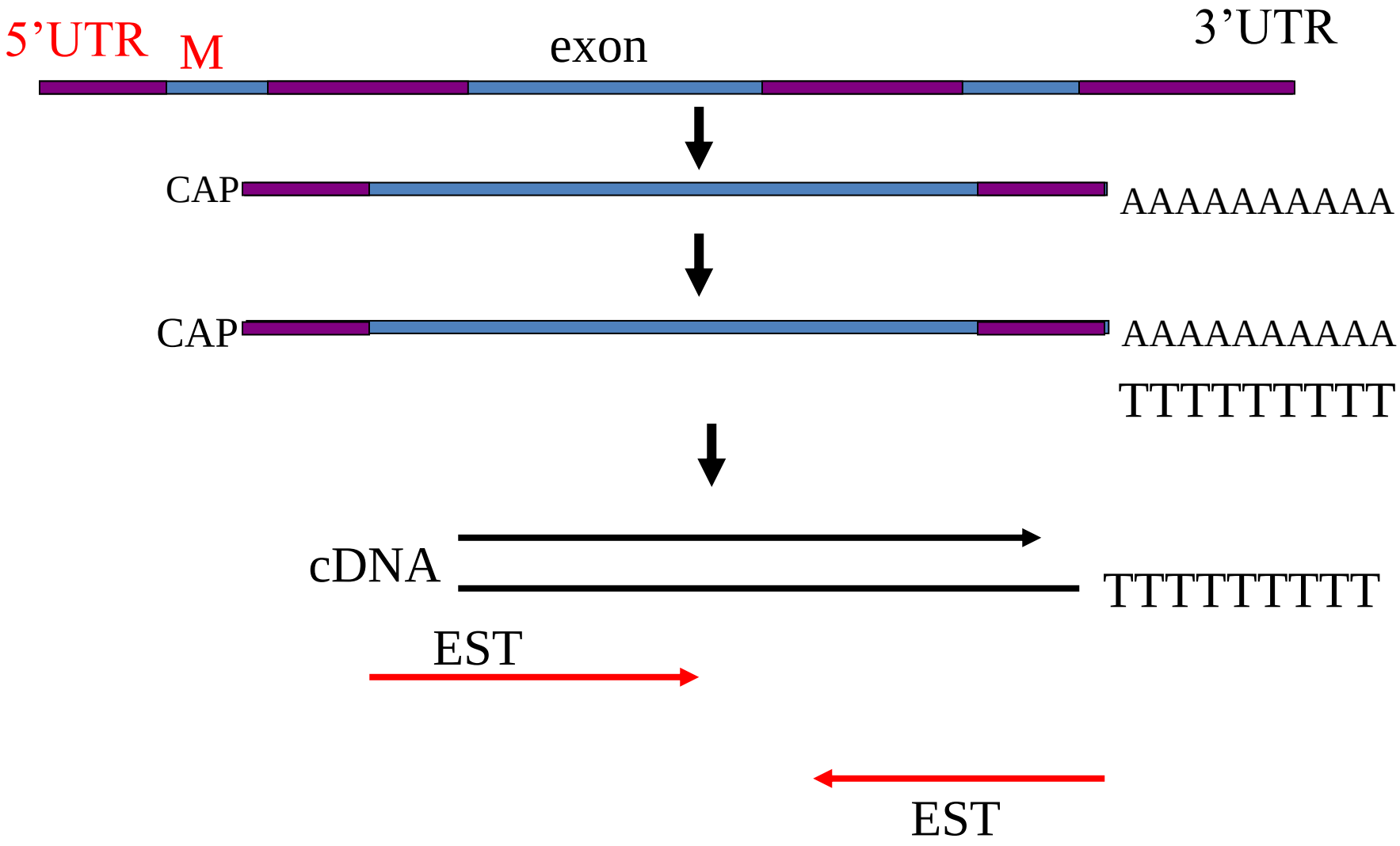
Predicción basada en homología

- Las secuencias codificantes son mas conservadas que las no codificantes debido a la presión selectiva
- Alineamiento de secuencias conocidas de genomas relacionados.
  - Podemos traducir nuestro genoma en los 6 marcos de lectura y alinear los péptidos obtenidos a proteínas conocidas (Blastx)
  - Nos puede dar idea de la función de la secuencia detectada

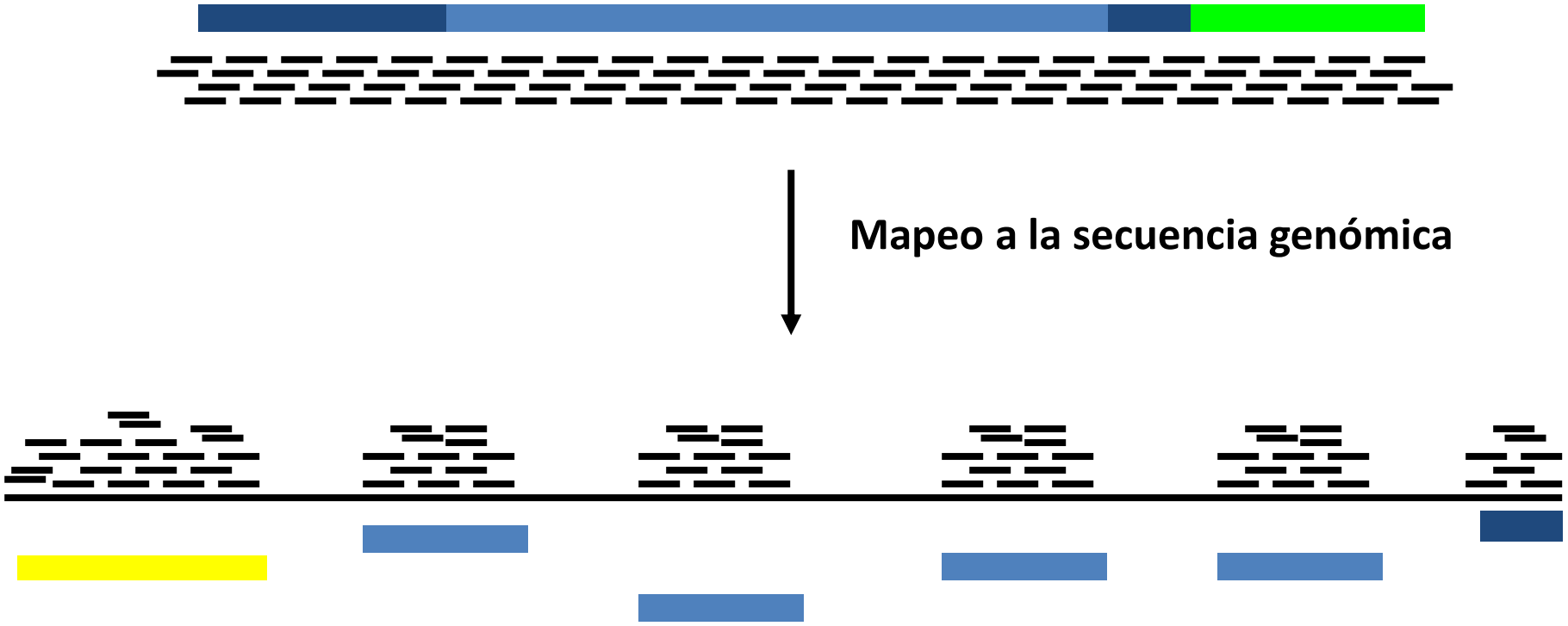
Predicción basada en homología



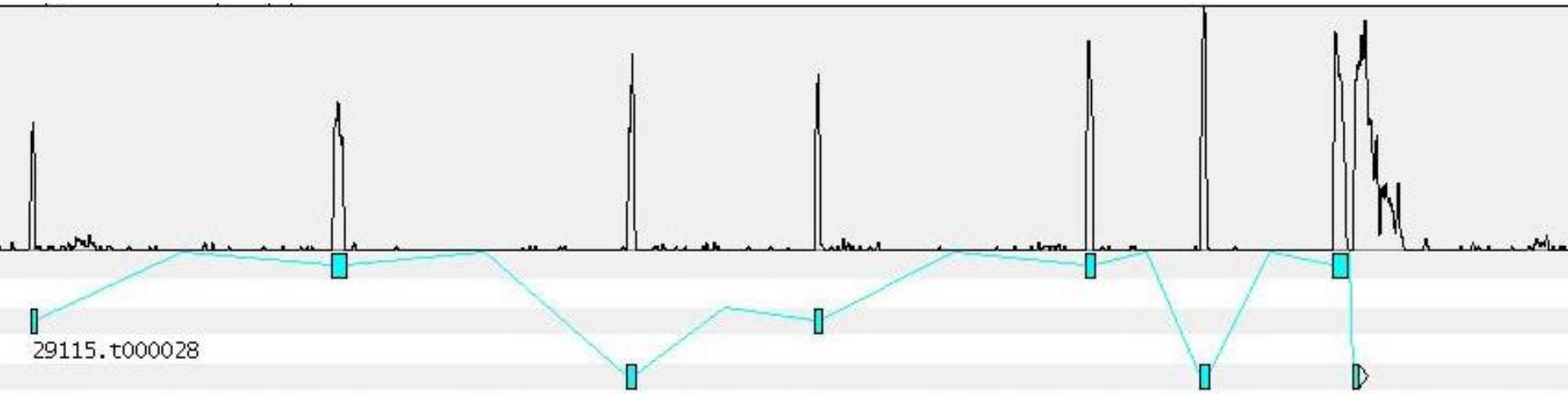
Evidencia experimental
ESTs



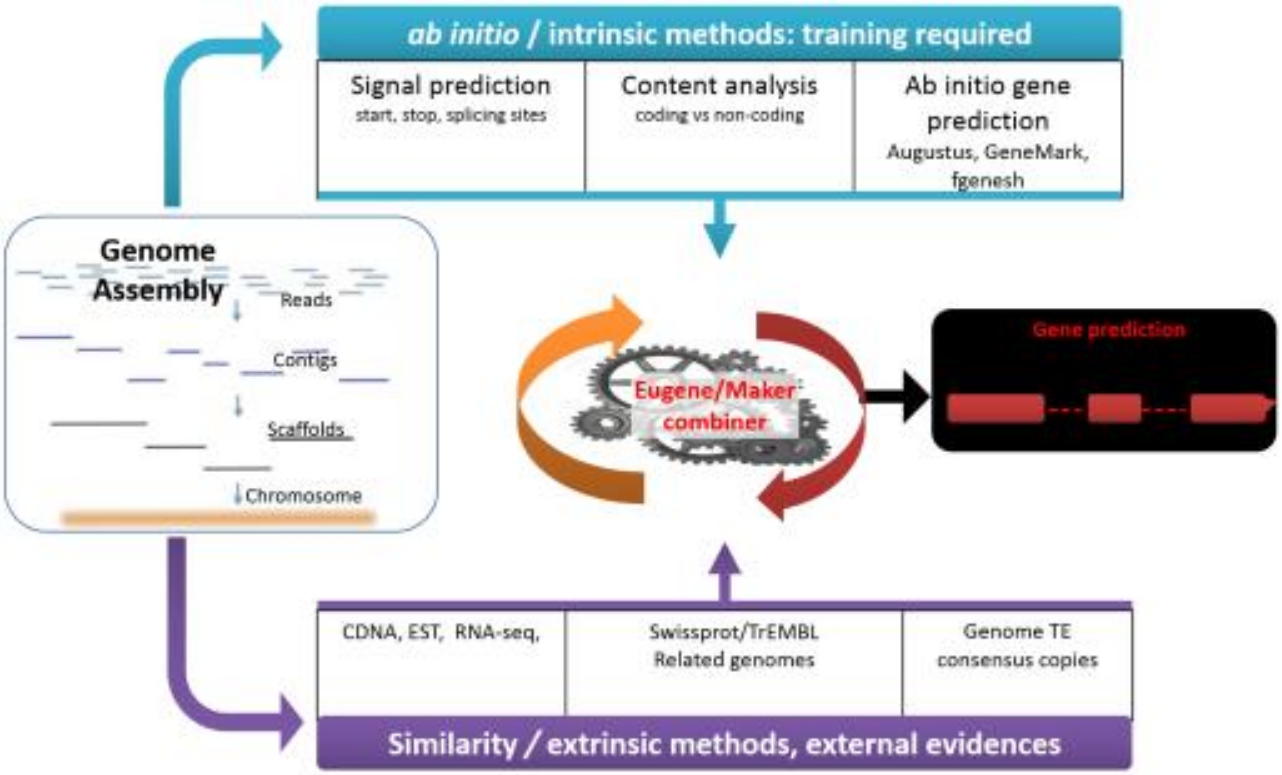
Evidencia experimental
RNAseq



Evidencia experimental
RNAseq

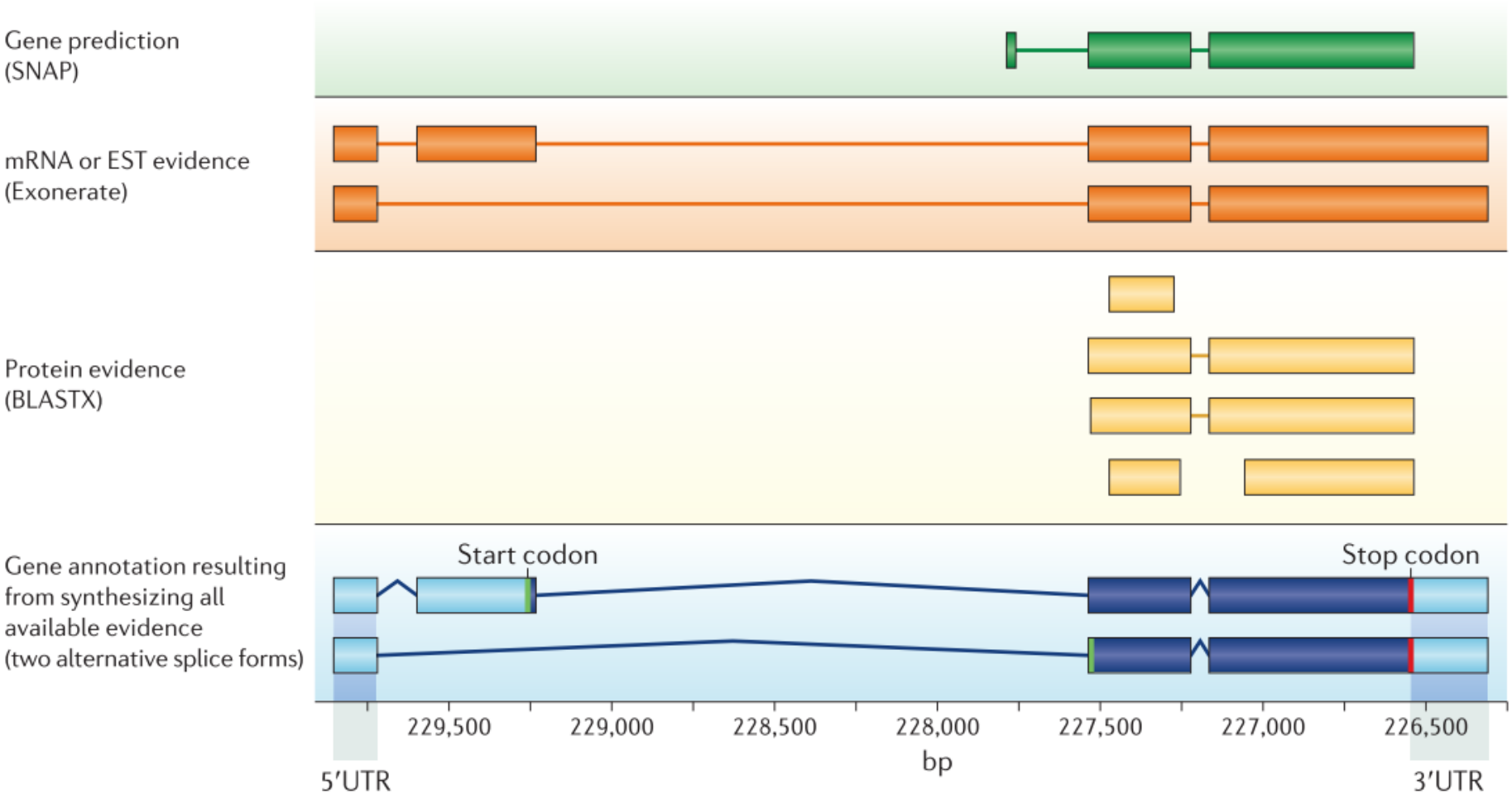


Anotación: Combinación de evidencia y curado

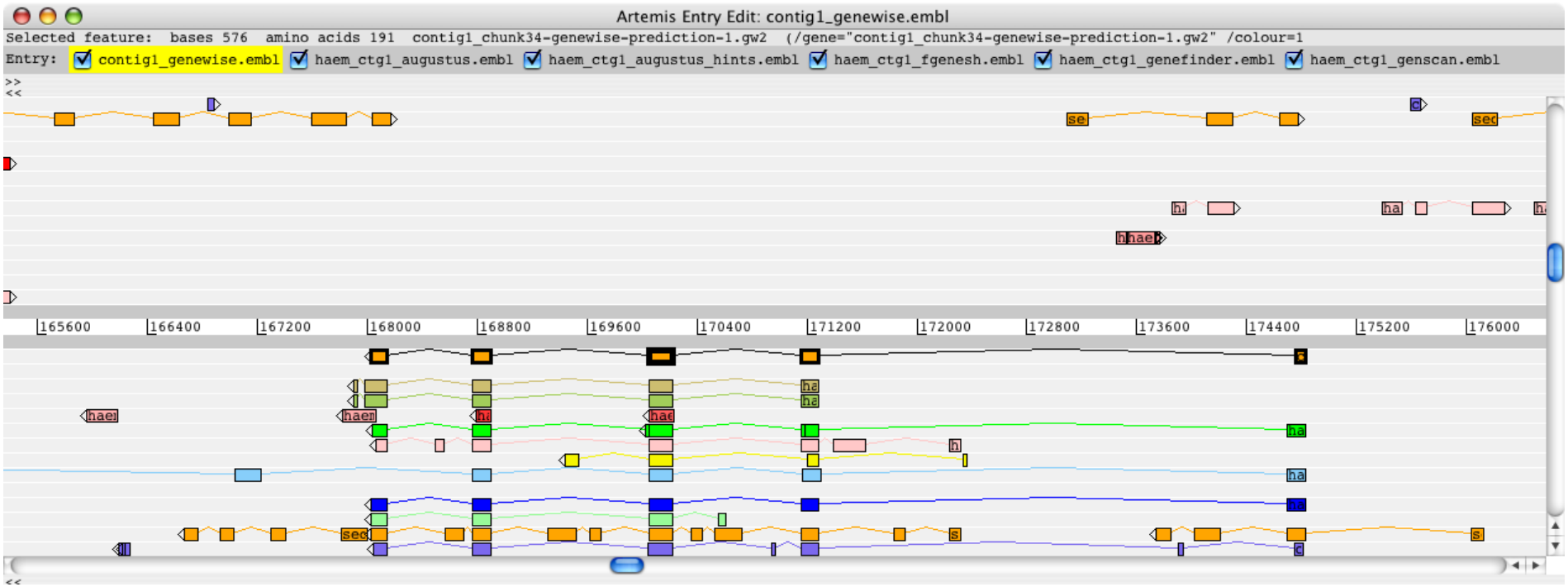


Dominguez Del Angel, F1000Research 2019

Anotación: Combinación de evidencia y curado



Anotación: Combinación de evidencia y curado



highlighted	manually reviewed gene structure
pale brown	hit to H. contortus EST cluster in Nembase found using PASA
brown-green	hit to H.contortus individual ESTs in NCBI database found using PASA
pink/red blocks	hits to Uniprot
bright green	twinscan prediction (**homology based**)
pale pink	snap prediction (ab initio)
yellow	hmmgene prediction (ab initio)
pale blue	genscan prediction (ab initio)
red	genefinder (ab initio)
dark blue	fgenesh prediction (ab initio)
jade green	augustus hints prediction (**homology based**)
orange	augustus prediction (ab initio)
purple	genewise prediction (**homology based**)

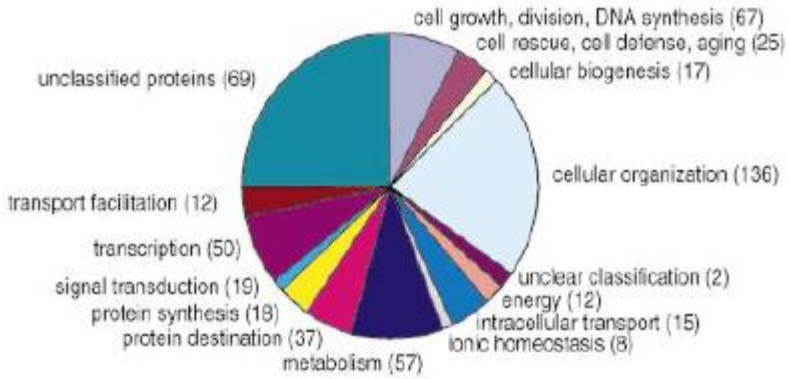
Que sigue???

Para que sirve?

Anotación funcional

Anotación funcional

- Búsqueda de similitud de secuencia
  - Inferir función por posible homología
    - A nivel global
      - Blast
    - Dominios y motivos
      - InterPro
        - » PROSITE, Pfam (HMM), PRINTS, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, GENE3D y PANTHER
- Señales particulares
    - Localización subcelular (SingnalP)
    - Dominios transmembrana
- Asignación de categorías en “gene ontologies” (GOs)



Anotación funcional

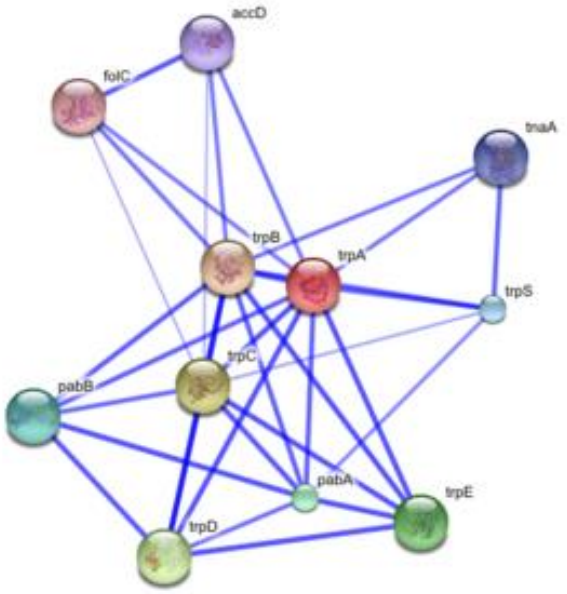
- Búsqueda de similitud de estructura
  - La estructura 3D de las proteínas es mas conservada que su secuencia
  - Se buscan las estructuras mas compatibles con la secuencia que se esta analizando
    - *"Protein threading"*

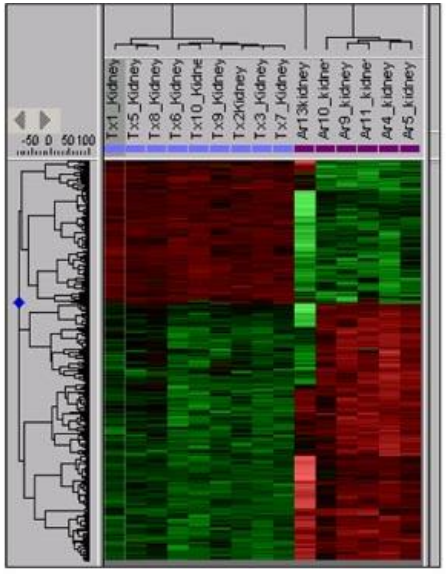
Anotación funcional

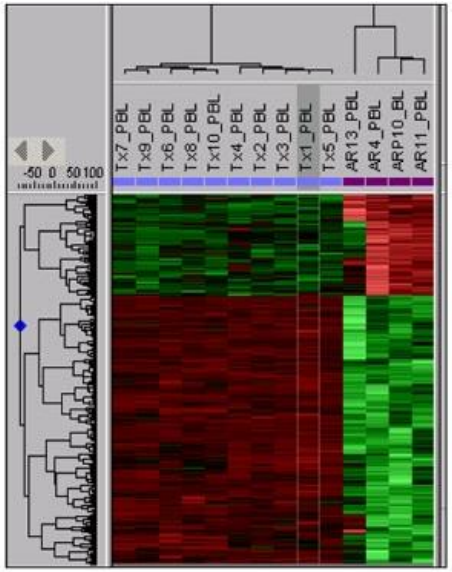
- Se asignan nombres de acuerdo a la evidencia:
  - “Putative”: alta similitud a genes conocidos o dominios funcionales
  - “Expressed”: con evidencia de EST (o RNA-seq) sin evidencia de función
  - “Hypothetical Protein Conserved”: Predicha por software de anotación y con secuencia similar en bases de datos
  - “ Hypothetical Protein”: Solo predicha por software de anotación

Anotación funcional

- *Guilt by association*:
  - Se anota de acuerdo a asociaciones entre genes
    - Interacciones proteína - proteína
    - Perfiles de expresión
  - Es una anotación general







Anotación funcional

- *Guilt by association*:

Anotación funcional