Ensamblaje de genomas y anotación

Genómica para bioinformática INB320
11 abril 2016
Eduardo Castro-Nallar, PhD
www.castrolab.org

"A man with a watch knows what time it is. A man with two watches is never sure." Segal's law

Resultado de la secuenciación

Resultado de la secuenciación

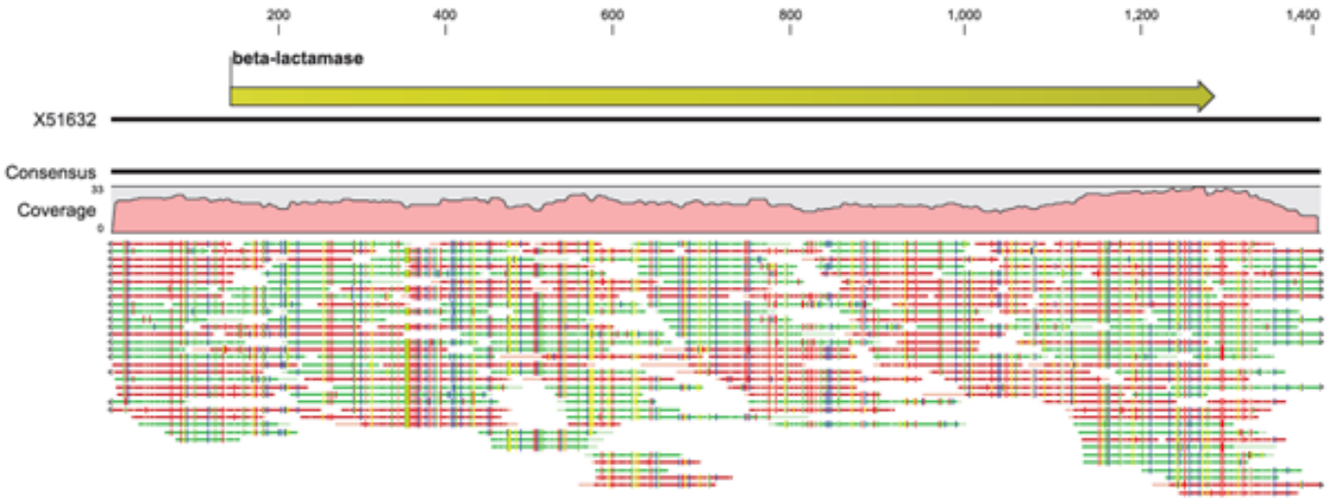
198D — -bash — 123×45

```
ACCFFFEGDGGGGGGGGGGGDF;FFCCECFCCCGFGGGGDFGGGCFFFF7F8FFFCFGG7CCFFFCEGGGCC=7,,7CEFGFG<FDE<FFFGF9FFFC,EFFC77CF7F7+6A=E8FGGFC<,
6AAFEFFFGGGG:FG+D@D,EBFAFDE,@>=;F9CF,@E?EC6E,7@997CE,C@E>CCCF>++=+;*;C8F4B*;*9>A)B<FFFABCFF*B>458@;35>8@DC7@(75)9)9<EEEEEFF
3?).9?3>>??1:1024).<:(2:1(23<<435)-:02:146AF96A
@M01447:49:000000000-A9789:1:1101:11006:1469 1:N:0:4
ACTTCACCCTGGTTGAAGAACGCCAGCACCAAACCTTCGGCGGCTTTCTCGATGAATTGCGGTTCGGCTTTCATGATGTCGGCGAAGAAGATGTTCGGCGACTTGCCGCCCAGTTCAACGGTG
GACGGAATGATGTTCTCGGCGGCCGCATGCATGATGTGCGAGCCCACTGGTGTGGAGCCAGTGAAGGCGATCTTGGC
+M01447:49:000000000-A9789:1:1101:11006:1469 1:N:0:4
ACCCGGGGGGGGGGGDGGGFFGGGGGDD<FE8<EFGGGGECECGGGGGGGGD8FAAAFF@E@EGGGDGGGGGG9FE9F<FGDCF:EEG7<<8BFFGGCCC7>FFE,EDCCFF+6@FF<DFF>F@
++FC77>D7>FCFGGGFC7E78FGGG>E*B+?;?CCF7CC=ECBF*AFCF??:5=?F=ECFF=C*0*>)7CGGG?=F
@M01447:49:000000000-A9789:1:1101:20842:1478 1:N:0:4
CCGTCGGGAATGGTCGCGCCGTAGATCGGTGACGGTTCGCTGGTGGCCGAACCACCCACGGCGAACCCGGTAGCCGTGGCCACCACGGCGGCGGTTGCAGCGACGAGTGCAGCAAGATGCAGG
CGCTTCATCACAGGTGATCCGCATCAATCACAGCCTTGGCAAAGTCTTGCGGGGCTTCCTGC
+M01447:49:000000000-A9789:1:1101:20842:1478 1:N:0:4
BCCGGGGG7FFGFGGFGEGGDFEGDFFFGF7DFD>FFGCEFFDG,@FFEC@FC8CFFCE77CCGGGGFE>F@CEF7F+>FE8FFGG7:FD+C+>FGGCA9C:*><*==<B,7<<7,:FBCCB+
4:FF<C@<<CF+<:F+?CGCCGE8C?+<E+A+<8@F@+<<99*<+<<<9<7*55CEFGGCCF
@M01447:49:000000000-A9789:1:1101:21118:1478 1:N:0:4
GCTGACTGTCGGGCCCCGAAGGCCCCCGAGAAAGTAATCGCCCAGTCCGCAGGTGGCGTGTAACAGCGGCCAGCAACACCGGAACCTTTCACTGATGCGCCCAGCTGTTACTTGGCCGGACGC
ATTGGGAAAACAACCAGCCCGCATGGAGACGTGCATGAACAAAGACGAAGCGTTTATCCGCCACTGTGCGGGCCGCAACCTAAGCAAATCCCATACAGC
+M01447:49:000000000-A9789:1:1101:21118:1478 1:N:0:4
@CCGFGGGGG<FCFFGGGGGG7FGGFGGDE>FCGGGFFFFGGGG?FFGGG7C7F?7FCGCGFGGGCFFGGGDEG8AFCGGDCCEGGDFGG9FE95C,C+BFC7AEFEFG9DFD,,>D@:EEC@
+CC<,:,,>CF7CGGGFGGGCFD>8DBCFFCG5B8F79;AC68C,CCEE7>5E=E8CFC8CDD<F4C+@<3DDD)9)/7F?9)6;*)96FF>FF0?24>
@M01447:49:000000000-A9789:1:1101:19737:1488 1:N:0:4
CCCCAGGTTAACGGTCTGGTCGACGGATTCAGGCTTAATGGAGCAAGCACCCGAGTGAACCGAACCAACGAAAGTCACAGTTCCATCTGCCGCATTTGCGACGGAGACTGAGCTCGCCAACAG
GCCGATAGACAACAACACGGTTTTCATTTTGCTATTCATCTACAGATTCCAACTCAAGTAAGTGTCATTCAATTACCCCAATACAGAAAATGCTCAGGGCGGCGACTTAACCAAGACACCCCT
TTGTTCGGTTGCTTGGTAG
+M01447:49:000000000-A9789:1:1101:19737:1488 1:N:0:4
@BCCCGGGGGGGGGGGGGFGGGGGGGEGGGGGFGGGFFGGFGGGGGDFGGGGCEFGFFGGGGEGGGGFFEFE>EGGCFFFGGGGCGGGDFFEFFFGGDG7@FF77:=AE8??AFFCFEECGFF
GDGD@C@F9FDFGDGGFCFGGGGGGFGGGG;DFAFGGGFGGFG7D7<EGGCCFECG66=B7>C9F9=CFGC7EFFFDFFC>F)?:A+;);>8A6C3@F(62(:909B:A49<>:223(8(1;?
BF261661<(7?9A<(84:
@M01447:49:000000000-A9789:1:1101:18372:1492 1:N:0:4
CTTGACCGAGTCGACCTTGTCGATGGAACGCTGGTTTTCCGGATCGAAGGTACGCAGGGTTTCGATTTCGTCGTCGAACAGGTCCATCCGGTAGGGCAGTTTGCTTCCCATCGGGAACAGGTC
CATCAGCGCGCCACGCACCGCGAACTCGCCGTGCTCGTACACCGTGTCGACGCAGCGATAGCCACTGGCCTCAAGCCGGGTGCGCATTTGCT
+M01447:49:000000000-A9789:1:1101:18372:1492 1:N:0:4
ACCCDFGGDGGGG7FGGGGGGC8FFCEFGGCFDGGFGGDE>@FCGDDGDCFFGGGG>FGGGGGGDFFGGGGG7FD77FC+C,FCFFGFB:7>@+BFCFCEED,5E,CB4C9A@+4+@DE=,CE
,8@<9=D7@=FE7C@E7FGDE8FGGCFC::*>1C>BB:*=*BCEF,=C*8C*:*2==*=5*;**2269DCCCFCFFF373>:C05>B)92@=
@M01447:49:000000000-A9789:1:1101:15204:1493 1:N:0:4
TCCGCCATCACCGAAGATCGACAGGTACGAGGTGTCGGTGCGCTGGTGATCCAGGCCGAGCAGCACGGTGTGGGTAATGTCGCCGGTGGCGAAGTCAGCCTGGAAATTATTGTCGACGGCGAA
CTGGCCGATGTTTTCGTCCACGTTGGTCGACCCGCGGCTGGTATTACCCTCGGCATCAATGGTCGAGTAGCCGTAGGAACCCACGGTCAGTT
+M01447:49:000000000-A9789:1:1101:15204:1493 1:N:0:4
CCCGGGGGGGGG7CEG>FGGDFEGGFFFG7;EGFFGE>F,CFGGDDFEDFEFFG,CFFE7CCFG8FCEG?C77CF9CGFGG8GGEDF+@F+6+:FFGFCGG8+,,,FEFEE?EC:+@FGF7@+
C>>:FGDEGGGGGGGCFEGCEEFF7,CEB5:BE*B:5CG5CF9F<>AFGFG=EE=G<<:F7CG09)>><FCGDF3C36*8>D<F@>7>:FB<
@M01447:49:000000000-A9789:1:1101:22503:1496 1:N:0:4
TGAAGTGTTCGATCAAGTGGCGCACCAGCACTTCGCCCACGCCAGGGCGGGTGCATTGCGGGTCCACCGCCAGGCACCAGAGGCTGCAGCCGTTCTCCGGGTCGTTGAAGGCTTTGTGGTGGT
TCAGGCCCATCACGCTGCCAATCACCGCGCCGCTGTCTTCATCCTCGGCCAGCCAATACACCGTGCCGCCTTGATGACGTGGCGTCCGCAGTGCCGGGTCCA
Eduardos-MBP-2:198D Ed$
```

Estrategias para armar el rompecabezas

Terminología

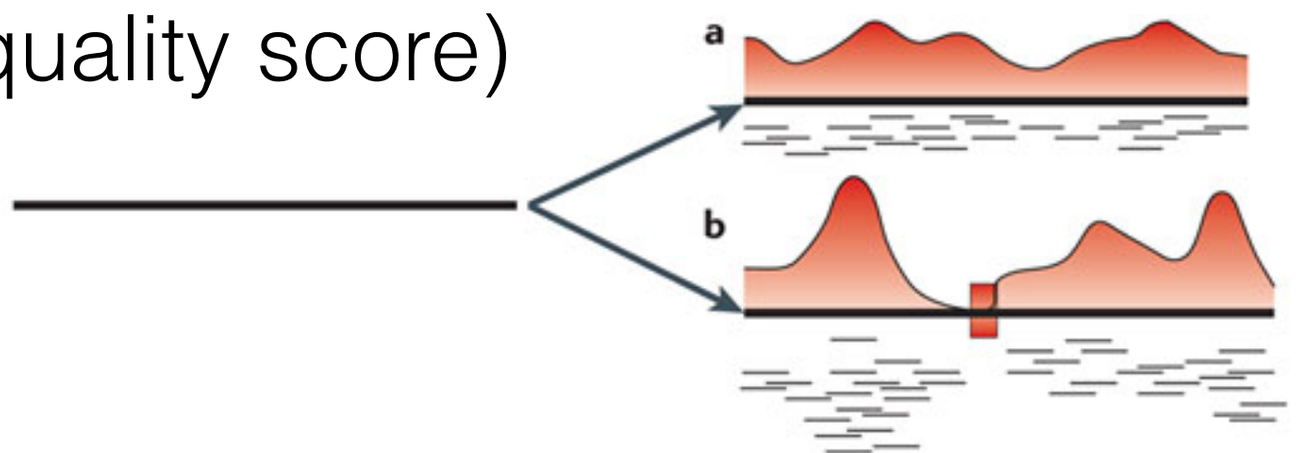
- Coverage



- Reads, contigs, scaffolds

- Profundidad (depth)

- Puntaje de calidad (quality score)



Phred Quality Score

The quality score of a base, also known as a Phred or Q score, is an integer value representing the estimated probability of an error, i.e. that the base is incorrect. If P is the error probability, then:

$P = 10^{-Q/10}$

$Q = -10 \log_{10}(P)$

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

http://drive5.com/usearch/manual/quality_score.html

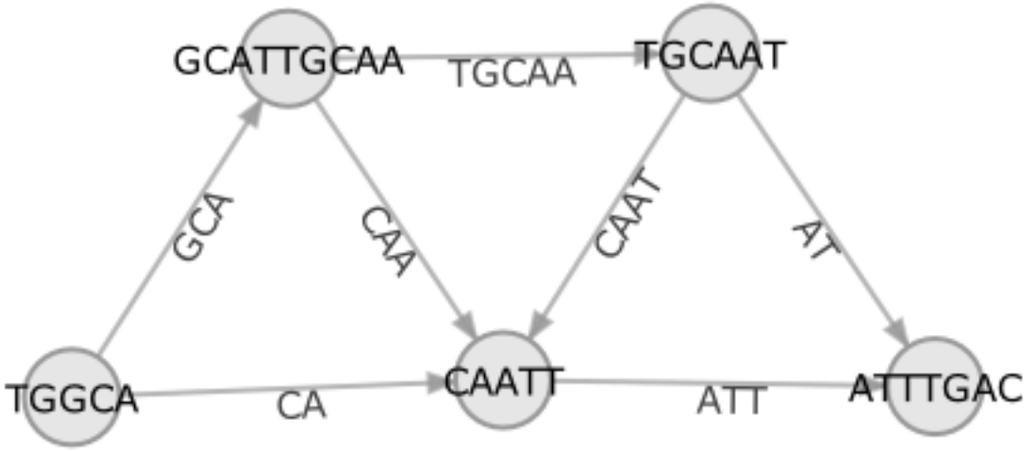
De novo

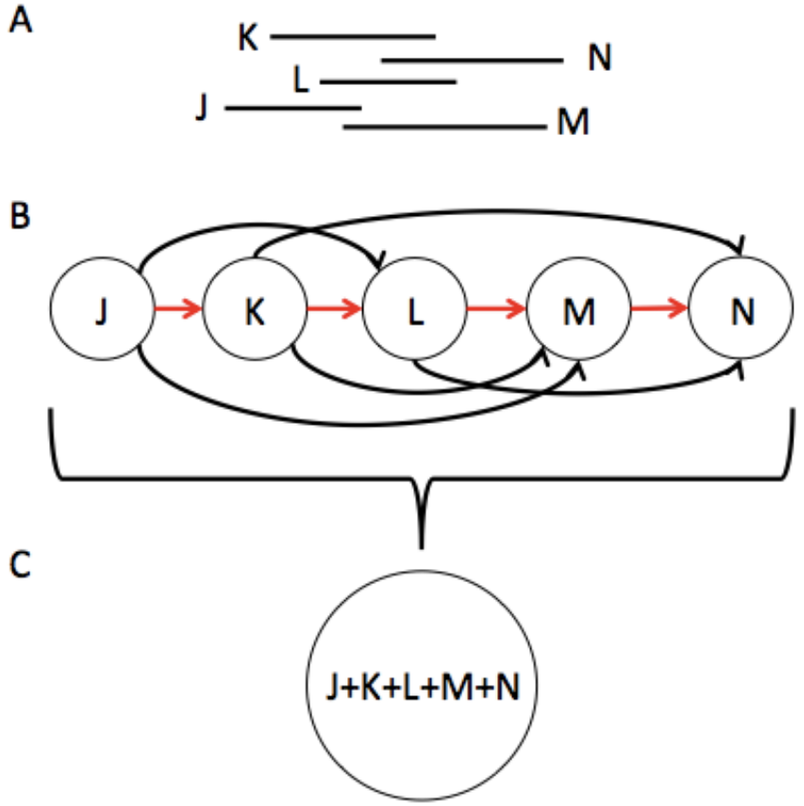
- Utilizar las reads por sí solas para reconstruir el genoma

- Dos estrategias: Overlay-layout-consensus y De Bruijn graphs

Overlay-layout-consensus

- Sobreponer reads por identidad de secuencia, unir reads sobrepuestas y encontrar un camino, formar un consenso

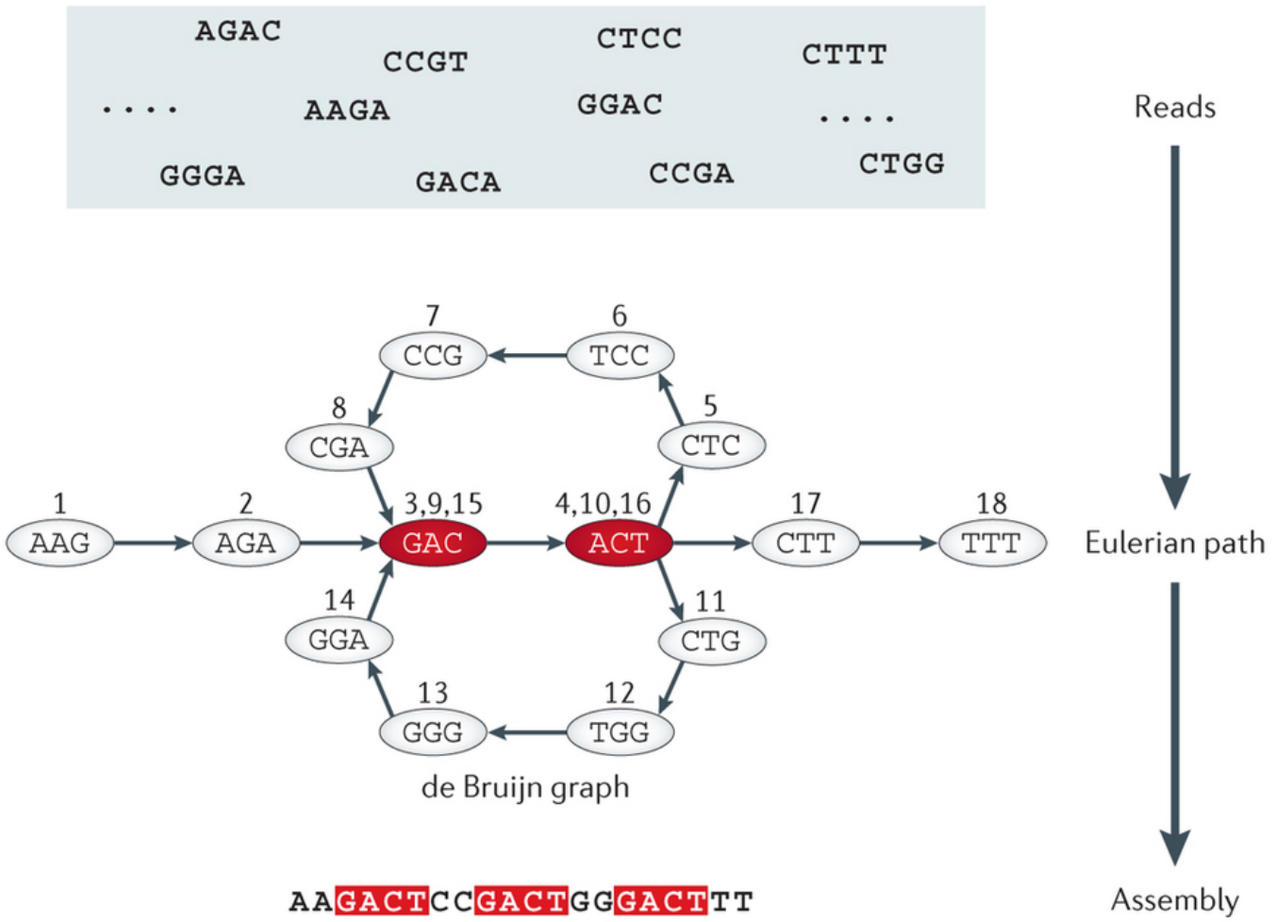




http://gcat.davidson.edu/phast/olc.html

De Bruijn graphs

- Fragmentar reads en pedazos de longitud K (llamados kmers), generar un gráfico sobrelapando kmers. Finalmente se forma una secuencia al trazar un camino donde cada kmer se visita una vez



Computational solutions for omics data

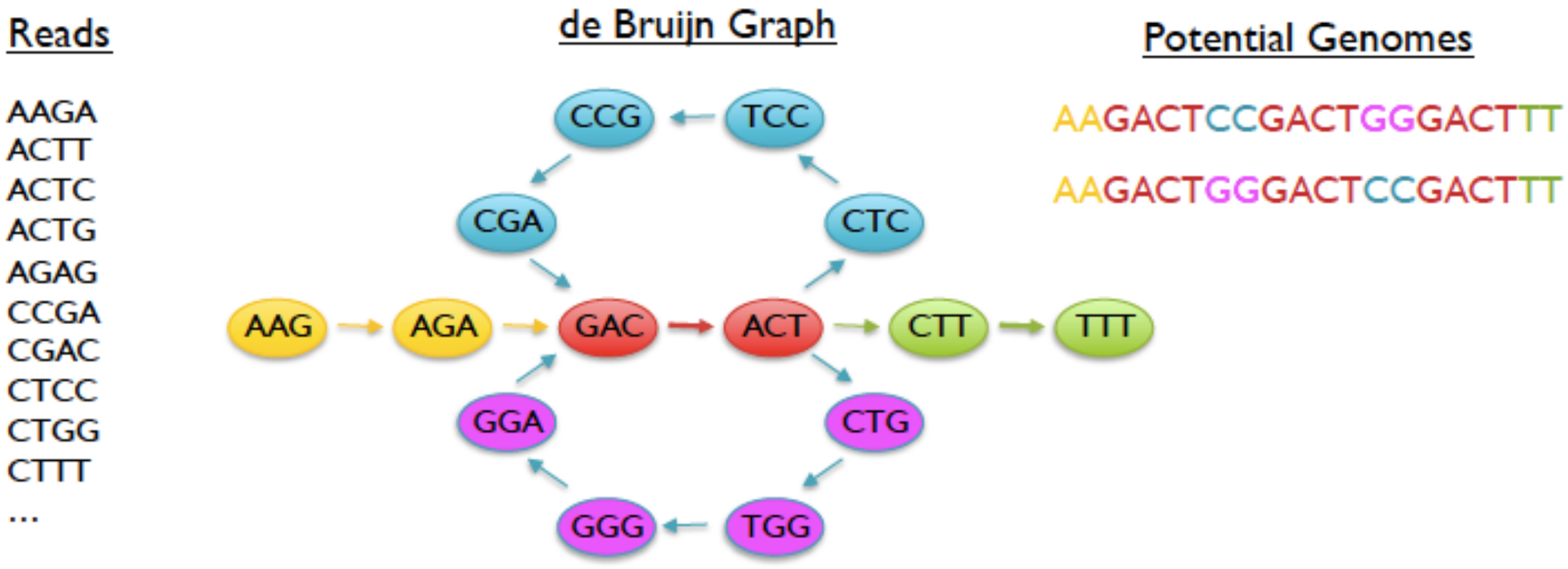
Bonnie Berger, Jian Peng & Mona Singh

Affiliations | Corresponding author

Nature Reviews Genetics **14**, 333–346 (2013) | doi:10.1038/nrg3433
Published online 18 April 2013

De Bruijn graphs

- Más de una solución para el mismo gráfico



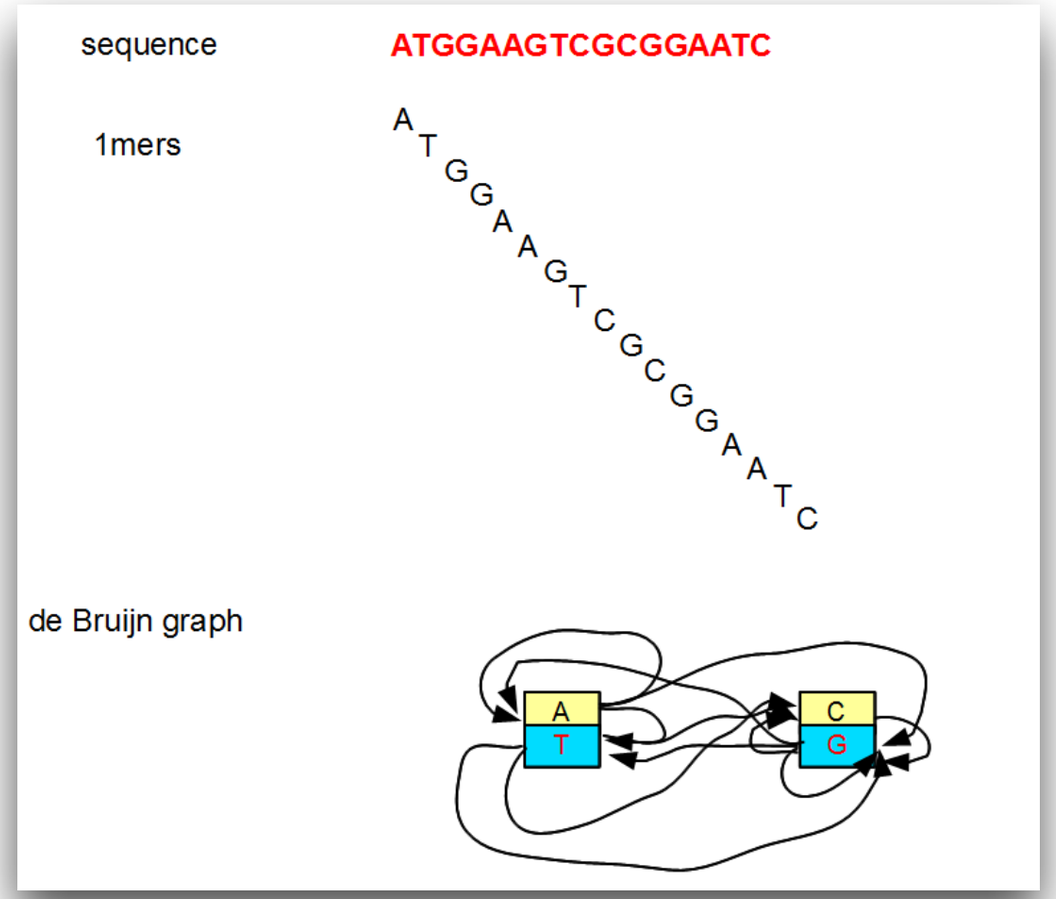
Michael Schatz, Cold Spring Harbor

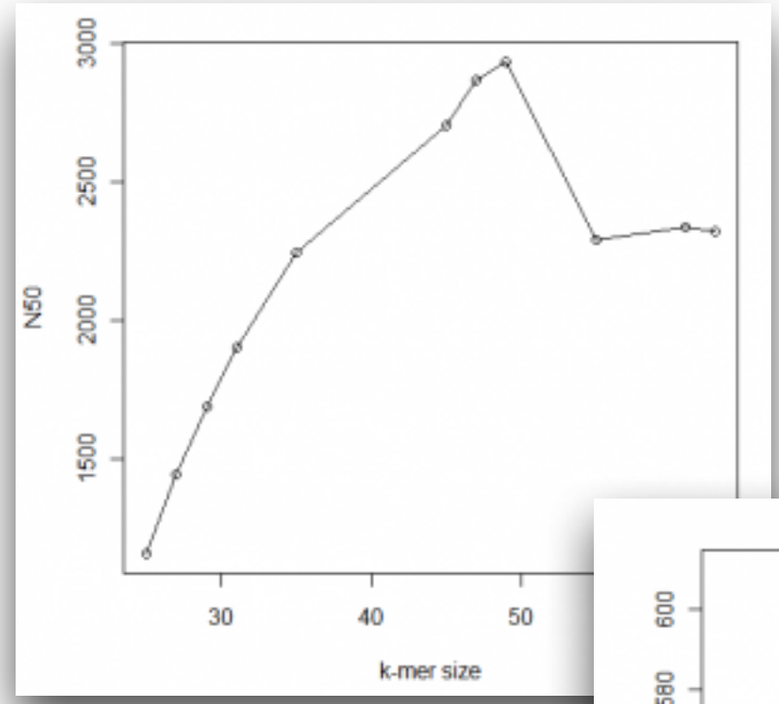
Consideraciones con ensamblaje de novo

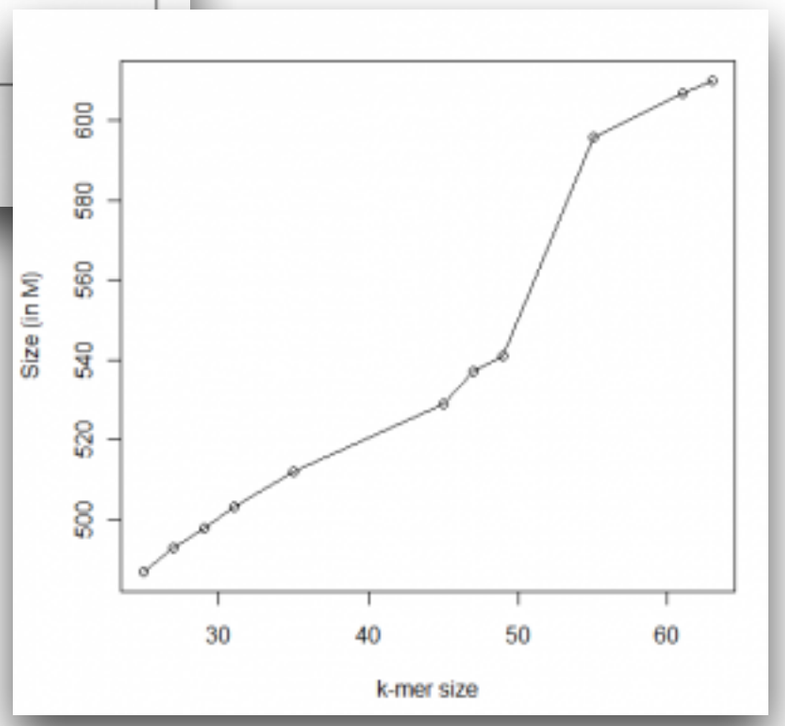
- DBG - valor de k, repeticiones más largas que k o que reads. Muchos errores con repeticiones
- DBG - mejor con reads cortas, aunque reads cada vez más largas
- DBG - requieren mucha memoria RAM, e.g., 140 GB a 2 TB
- OLC - lento, semanas en supercomputadora
- OLC - requiere calcular todas las combinaciones de reads
- OLC - errores cuando datos tienen mucha profundidad

Consideraciones con ensamblaje de novo

- DBG - valor de k, repeticiones más largas que k o que reads. Muchos errores con repeticiones







http://www.homolog.us/blogs/blog/2012/10/10/multi-kmer-de-bruijn-graphs/

Consideraciones con ensamblaje de novo

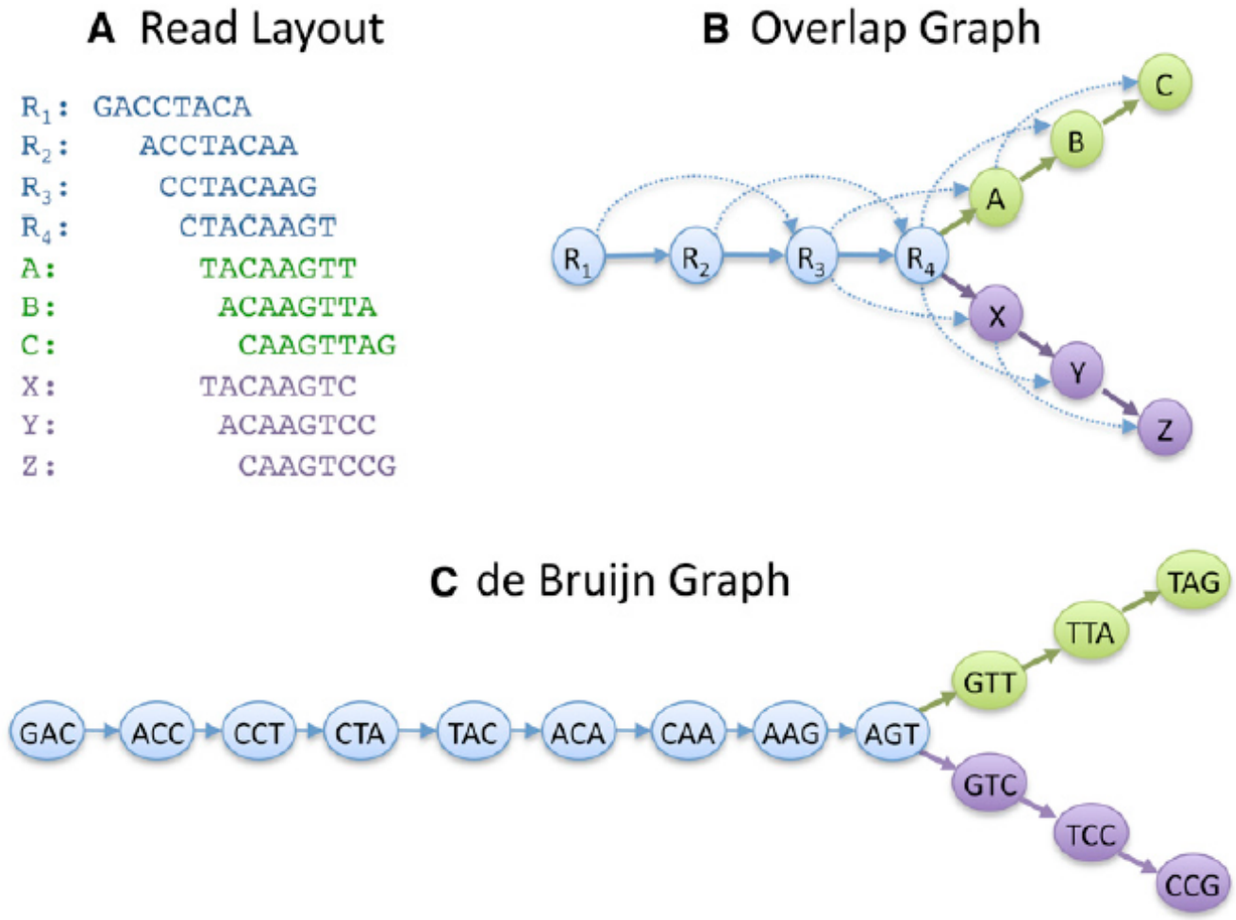


Figure 2. Differences between an overlap graph and a de Bruijn graph for assembly. Based on the set of 10 8-bp reads (*A*), we can build an overlap graph (*B*) in which each read is a node, and overlaps >5 bp are indicated by directed edges. Transitive overlaps, which are implied by other longer overlaps, are shown as dotted edges. In a de Bruin graph (*C*), a node is created for every *k*-mer in all the reads; here the *k*-mer size is 3. Edges are drawn between every pair of successive *k*-mers in a read, where the *k*-mers overlap by $k - 1$ bases. In both approaches, repeat sequences create a fork in the graph. Note here we have only considered the forward orientation of each sequence to simplify the figure.

Assembly of large genomes using second-generation sequencing

Michael C. Schatz, Arthur L. Delcher and Steven L. Salzberg

Genome Res. published online May 27, 2010
Access the most recent version at doi:10.1101/gr.101360.109

Consideraciones con ensamblaje *de novo*

- Mientras más largas las reads mejor es el ensamblaje, menos ambigüedad

- Se necesita mucho coverage

- El resultado está fragmentado

- Evaluar la calidad —> N50, mediana, media

- N50 = después de ordenar contigs, se divide la distribución de bases por la mitad, la longitud del contig donde esto ocurre es el N50

Por referencias

- Útil para estudios de resecuenciamiento, e.g., UK10K, GenomeTrakr

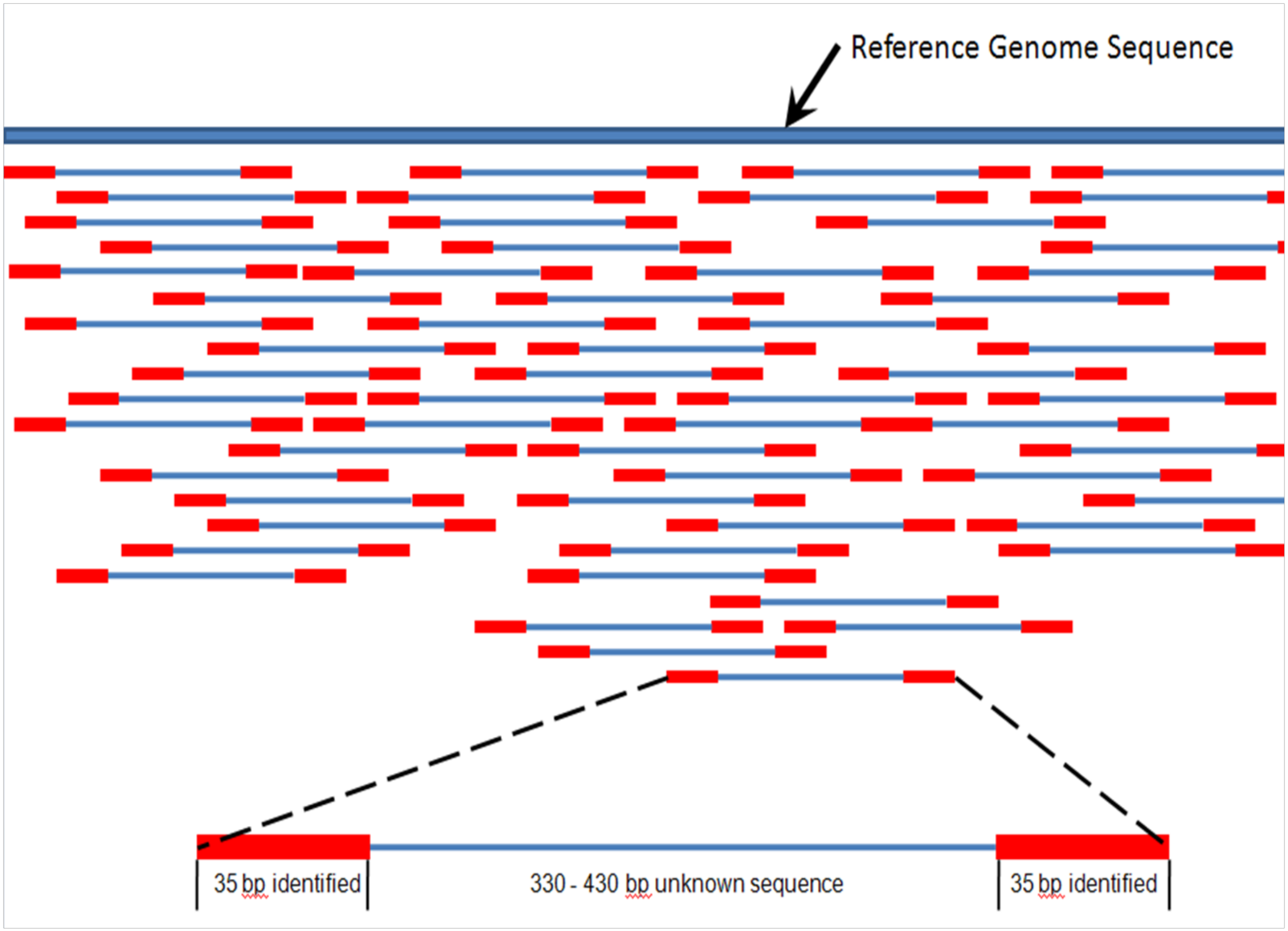
- Se usa un genoma ensamblado para “mapear” reads

- Computacionalmente más fácil que *de novo*

- Reads cortas pueden mapear en varias partes en la referencia

- Limita conocer la estructura de genomas nuevos, restringe reads a la referencia

Por referencias



¿Qué obtenemos al final del ensamblaje?

- Contigs o scaffolds
- Difícilmente se recupera el genoma completo, i.e., cromosomas lineales o circulares
- 100 contigs para bacterias es común
- “Finalizar” o “cerrar” es más caro y laborioso

¿Cuál implementación usar para mi genoma?

- Revisar Benchmarks

CSH PRESS GENOME RESEARCH

Assemblathon 1: A competitive assessment of de novo short read assembly methods

Dent Earl, Keith Bradnam, John St. John, et al.

Genome Res. 2011 21: 2224-2241 originally published online September 16, 2011
Access the most recent version at doi:10.1101/gr.126599.111

Bradnam *et al. GigaScience* 2013, **2**:10
http://www.gigasciencejournal.com/content/2/1/10

(GIGA)^n SCIENCE

RESEARCH Open Access

Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species

Keith R Bradnam[1*†], Joseph N Fass[1†], Anton Alexandrov[36], Paul Baranay[2], Michael Bechner[39], Inanç Birol[33],

GAGE: A critical evaluation of genome assemblies and assembly algorithms

Steven L. Salzberg,[1,7] Adam M. Phillippy,[2] Aleksey Zimin,[3] Daniela Puiu,[1] Tanja Magoc,[1] Sergey Koren,[2,4] Todd J. Treangen,[1] Michael C. Schatz,[5] Arthur L. Delcher,[6] Michael Roberts,[3] Guillaume Marçais,[3] Mihai Pop,[4] and James A. Yorke[3]

¿Cómo comparamos ensamblajes?

- GAEMR

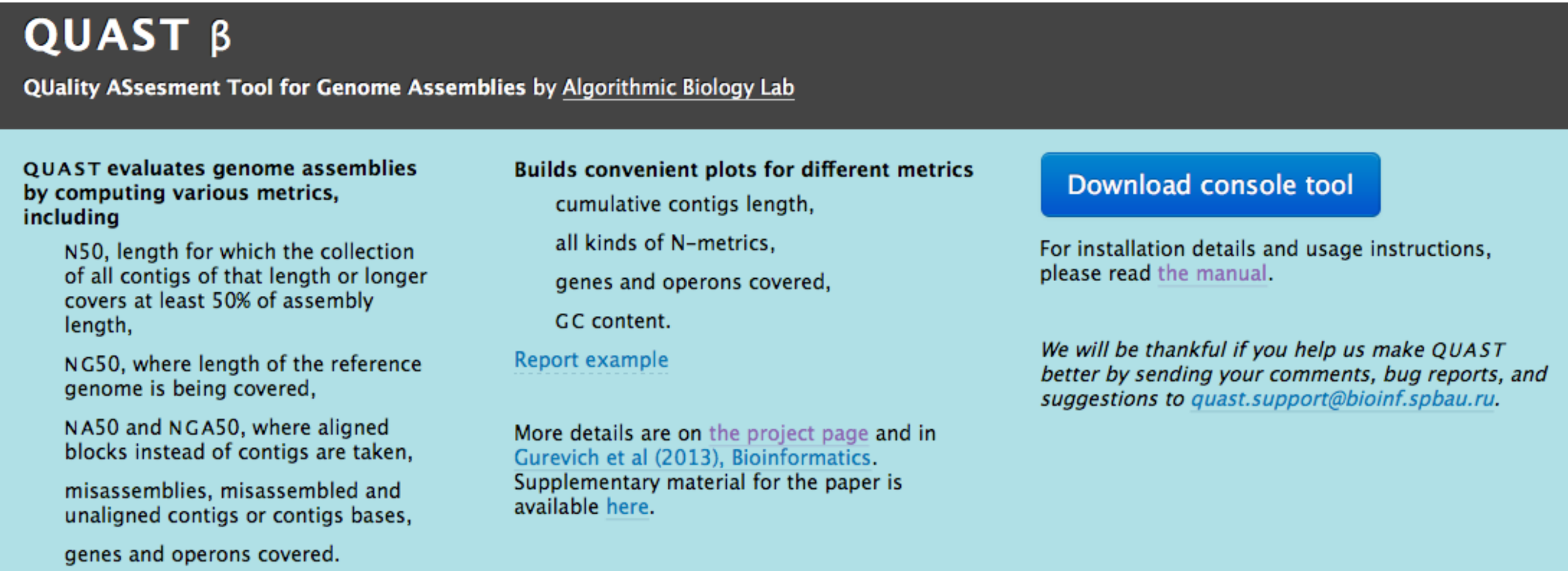
GAEMR

The **G**enome **A**ssembly **E**valuation **M**etrics and **R**eporting (GAEMR) package is an assembly analysis framework composed a number of integrated modules. These modules can be executed as a single program to generate a complete analysis report, or executed individually to generate specific charts and tables. GAEMR standardizes input by converting a variety of read types to Binary Alignment Map (BAM) format, allowing a single input format to be entered into GAEMR's analysis pipeline, hence enabling the generation of standard reports.

GAEMR's analysis philosophy is centered on contiguity, correctness, and completeness -- how many pieces in an assembly composed of, how well those pieces accurately represent the genome sequenced, and how much of that genome is represented by those pieces. By performing over twenty different analyses based on these principles, GAEMR gives a clear picture of the condition of a genome assembly. For a broadly-defined list of these analyses, see the Features section below.

¿Cómo comparamos ensamblajes?

- QUAST



Report Example

E. coli single-cell assemblies

4 639 675 bp, G+C content: 50.79%

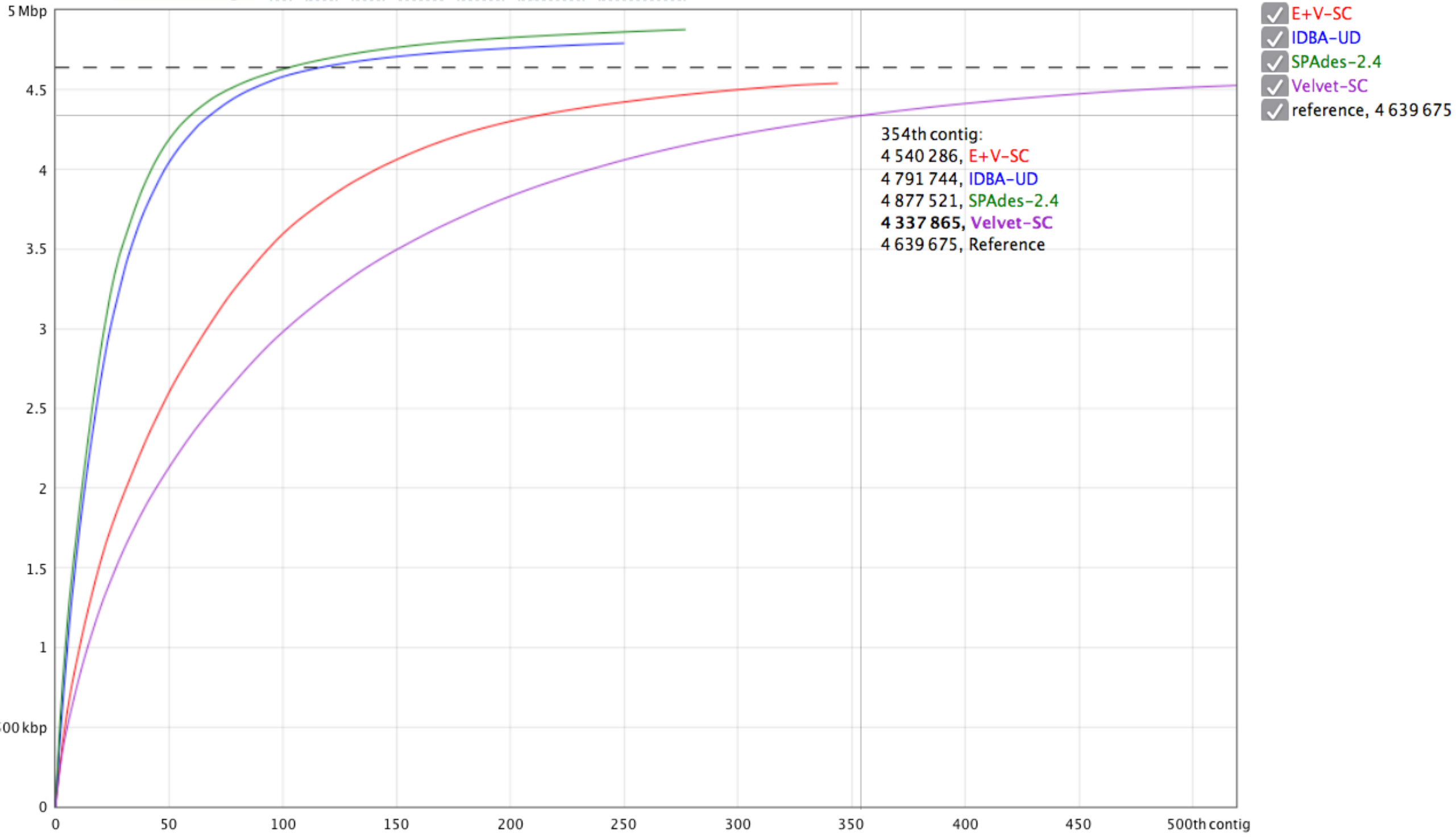
All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs.)

Worst Median Best

☑ Show heatmap

Statistics without reference	E+V-SC	IDBA-UD	SPAdes-2.4	Velvet-SC
# contigs	344	250	277	519
Largest contig	132 865	224 018	269 177	121 367
Total length	4 540 286	4 791 744	4 877 521	4 526 656
N50	33 616	96 947	106 927	20 445
Misassemblies				
# misassemblies	2	9	2	2
Misassembled contigs length	23 485	66 335	26 551	22 359
Mismatches				
# mismatches per 100 kbp	2.260	3.65	5.060	1.77
# indels per 100 kbp	0.7	0.2	0.7	0.92
# N's per 100 kbp	0	0	4.860	0
Genome statistics				
Genome fraction (%)	91.727	94.943	95.759	91.43
Duplication ratio	1.001	1.001	1.004	1.002
# genes	3767 + 160 part	4026 + 80 part	4046 + 102 part	3630 + 288 part
# operons	723 + 87 part	802 + 40 part	809 + 48 part	650 + 158 part
NGA50	32 051	96 947	110 539	19 791
Predicted genes				
# predicted genes (unique)	4258	4394	4417	4331
# predicted genes (>= 0 bp)	4258	4394	4490	4331
# predicted genes (>= 300 bp)	3643	3736	3784	3666
# predicted genes (>= 1500 bp)	524	559	559	515
# predicted genes (>= 3000 bp)	44	49	48	39

Plots: Cumulative length | Nx | NAx | NGx | NGAx | Genes | Operons | GC content



Contigs are ordered from largest (contig #1) to smallest.

¿Unir ensambles diferentes?

- Metassembler

Metassembler: merging and optimizing de novo genome assemblies

Alejandro Hernandez Wences and Michael C. Schatz

Genome Biology 2015 **16**:207 | **DOI:** 10.1186/s13059-015-0764-4 | © Wences and Schatz. 2015

Received: 28 July 2015 | **Accepted:** 1 September 2015 | **Published:** 24 September 2015

Abstract

Genome assembly projects typically run multiple algorithms in an attempt to find the single best assembly, although those assemblies often have complementary, if untapped, strengths and weaknesses. We present our metassembler algorithm that merges multiple assemblies of a genome into a single superior sequence. We apply it to the four genomes from the Assemblathon competitions and show it consistently and substantially improves the contiguity and quality of each assembly. We also develop guidelines for meta-assembly by systematically evaluating 120 permutations of merging the top 5 assemblies of the first Assemblathon competition. The software is open-source at http://metassembler.sourceforge.net.

Anotación genómica

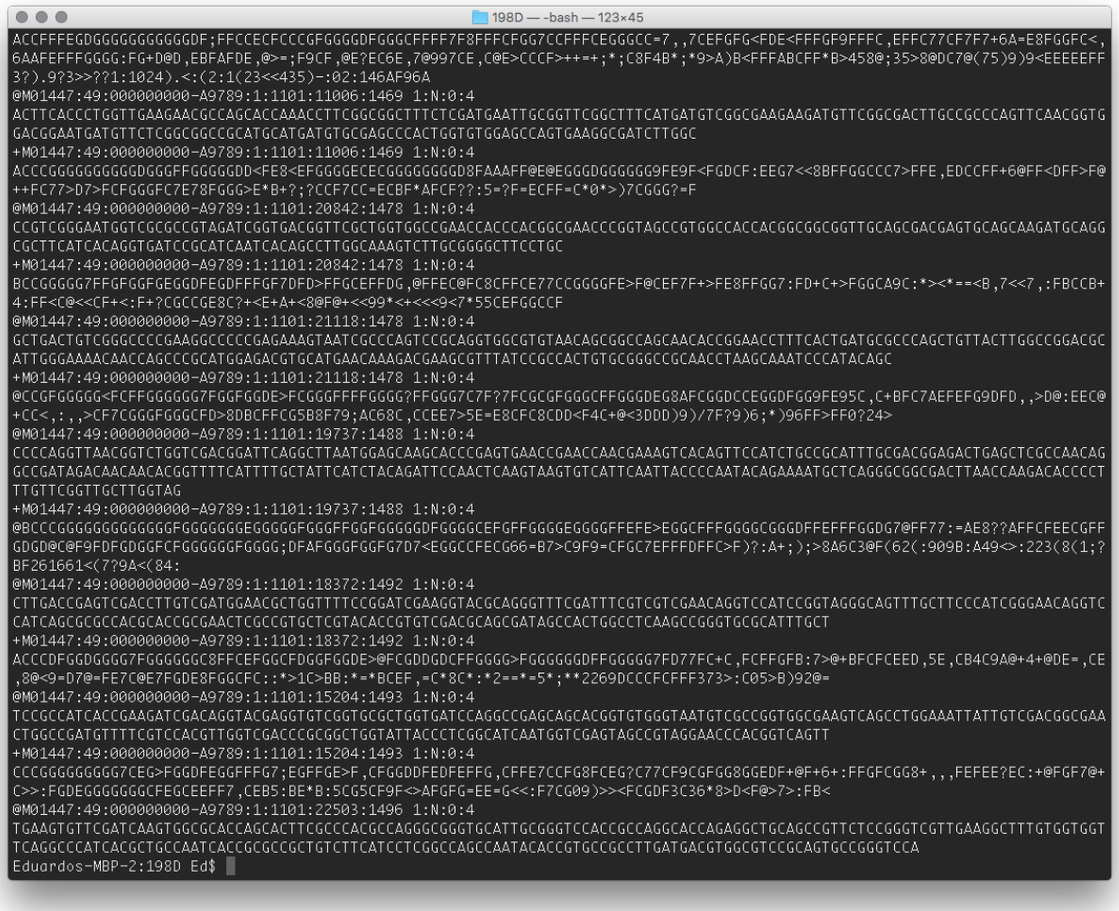
“a critical or explanatory note or body of notes added to a text”

“a note added to a text, book, drawing, etc., as a comment or explanation”

“the act of adding notes or comments to something : the act of annotating something”

diccionario Merriam-Webster

de información a mejor información




```
##gff-version   3
##sequence-region   chr8 1 100000
chr8    Gaze    gene    10503   11577   7.84    -       .       ID=GSVIVG01033678001;complete=1
chr8    Gaze    mRNA    10503   11577   7.84    -       .       ID=GSVIVT01033678001;Parent=GSVIVG01033678001;complete=1
chr8    Gaze    CDS     10503   10593   -0.4    -       1       ID=GSVIVT01033678001.cds1;Parent=GSVIVT01033678001;complete=1
chr8    Gaze    exon    10503   10593   .       -       1       ID=GSVIVT01033678001.exon1;Parent=GSVIVT01033678001
chr8    .       intron  10594   10900   .       -       .       Parent=GSVIVT01033678001
chr8    Gaze    CDS     10901   11396   5.32    -       2       ID=GSVIVT01033678001.cds1;Parent=GSVIVT01033678001;complete=1
chr8    Gaze    exon    10901   11396   .       -       1       ID=GSVIVT01033678001.exon2;Parent=GSVIVT01033678001
chr8    .       intron  11397   11465   .       -       .       Parent=GSVIVT01033678001
chr8    Gaze    CDS     11466   11577   5.46    -       0       ID=GSVIVT01033678001.cds1;Parent=GSVIVT01033678001;complete=1
chr8    Gaze    exon    11466   11577   .       -       1       ID=GSVIVT01033678001.exon3;Parent=GSVIVT01033678001
###
chr8    Gaze    gene    22057   23119   54.7    +       .       ID=GSVIVG01033677001;complete=1
chr8    Gaze    mRNA    22057   23119   54.7    +       .       ID=GSVIVT01033677001;Parent=GSVIVG01033677001;complete=1
chr8    Gaze    five_prime_UTR  22057   22166   1.22    +       .       ID=GSVIVT01033677001.utr2;Parent=GSVIVT01033677001
chr8    Gaze    exon    22057   22382   .       +       1       ID=GSVIVT01033677001.exon1;Parent=GSVIVT01033677001
chr8    Gaze    CDS     22167   22382   4.64    +       0       ID=GSVIVT01033677001.cds1;Parent=GSVIVT01033677001;complete=1
chr8    .       intron  22383   22496   .       +       .       Parent=GSVIVT01033677001
chr8    Gaze    CDS     22497   22550   12.1    +       0       ID=GSVIVT01033677001.cds1;Parent=GSVIVT01033677001;complete=1
chr8    Gaze    exon    22497   22550   .       +       1       ID=GSVIVT01033677001.exon2;Parent=GSVIVT01033677001
chr8    .       intron  22551   22650   .       +       .       Parent=GSVIVT01033677001
```

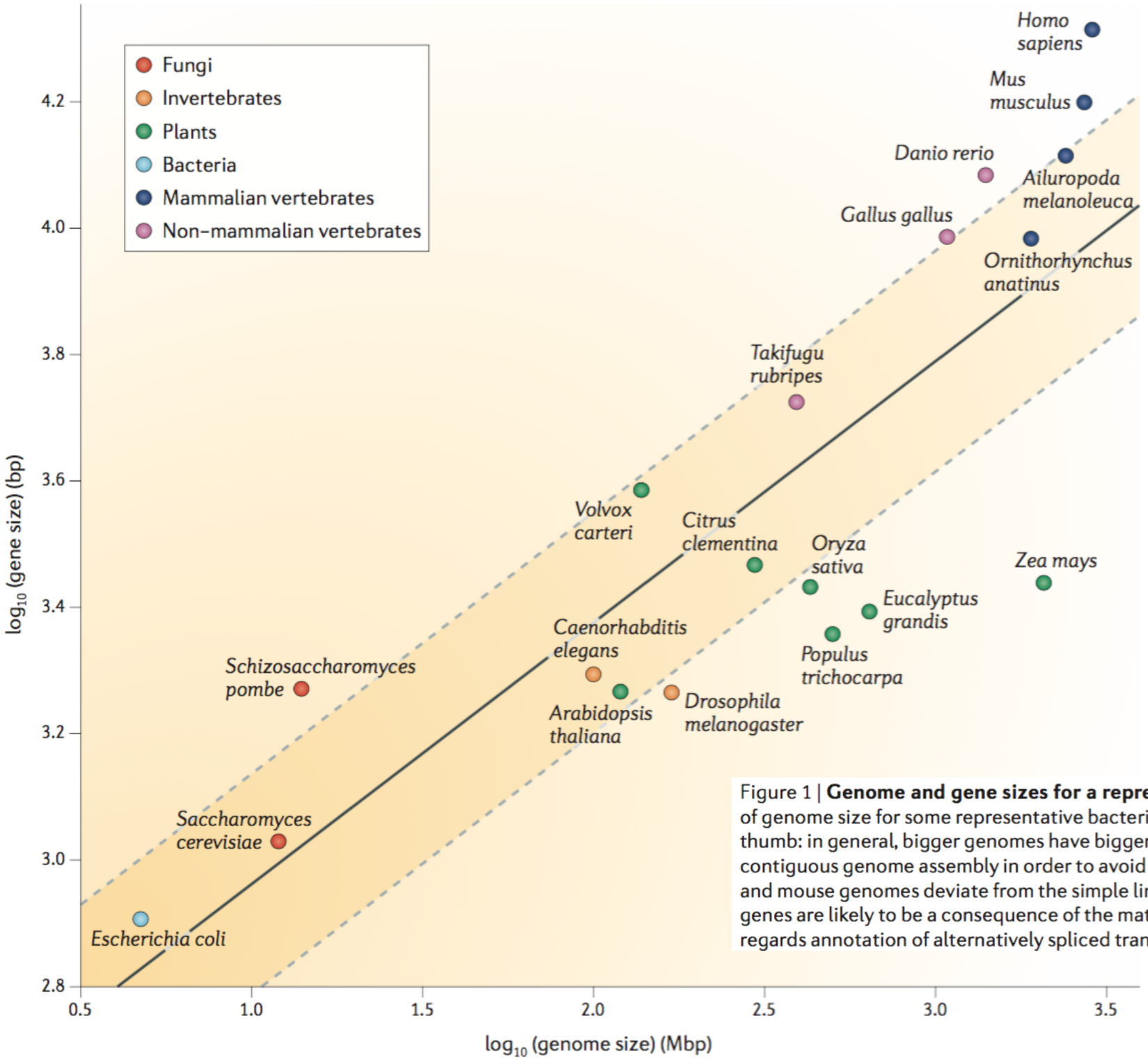


Figure 1 | **Genome and gene sizes for a representative set of genomes.** Gene size is plotted as a function of genome size for some representative bacteria, fungi, plants and animals. This figure illustrates a simple rule of thumb: in general, bigger genomes have bigger genes. Thus, accurate annotation of a larger genome requires a more contiguous genome assembly in order to avoid splitting genes across scaffolds. Note too that although the human and mouse genomes deviate from the simple linear model shown here, the trend still holds. Their unusually large genes are likely to be a consequence of the mature status of their annotations, which are much more complete as regards annotation of alternatively spliced transcripts and untranslated regions than those of most other genomes.

Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. May 2012:1-14. doi:10.1038/nrg3174.

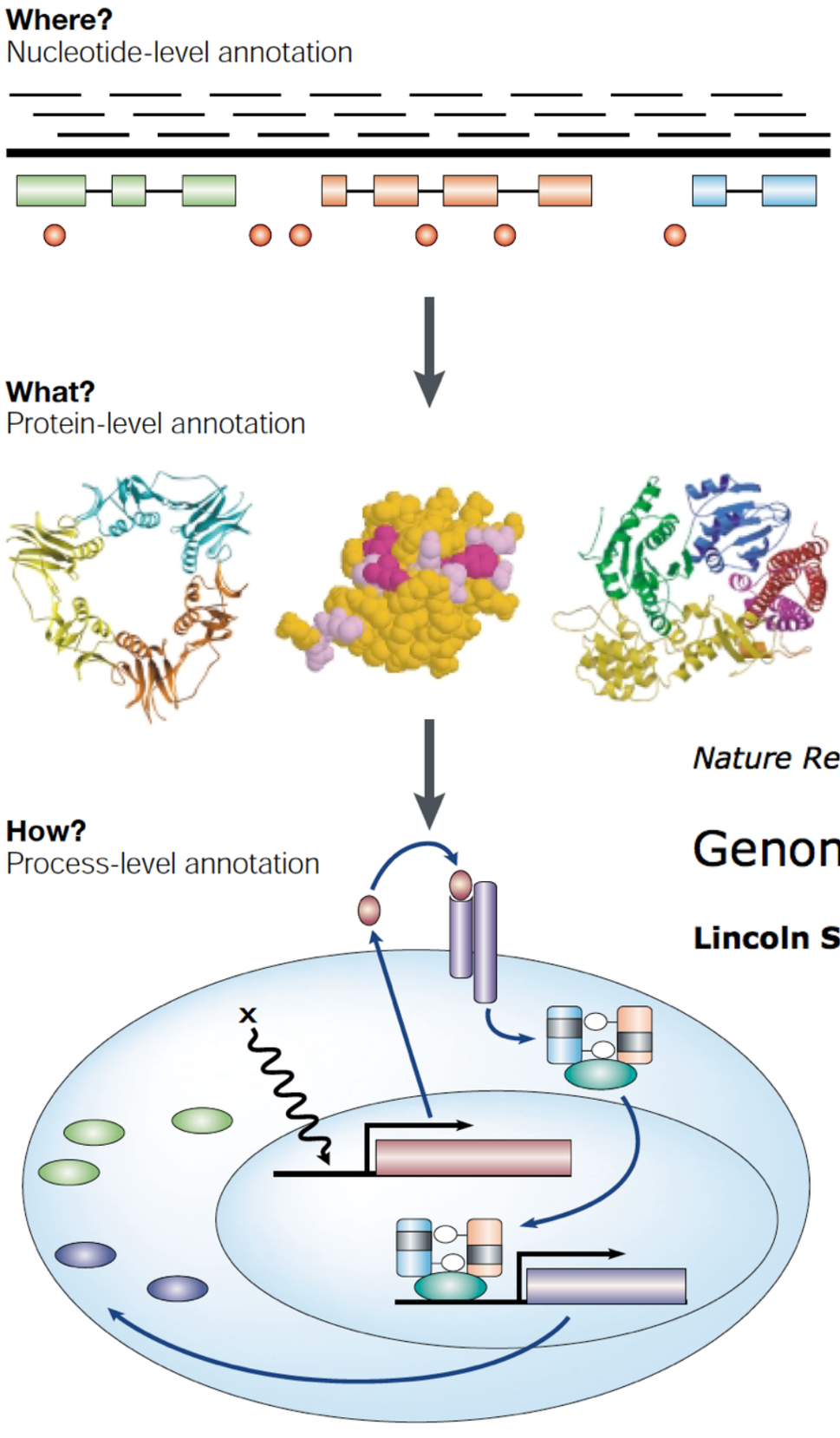
¿Por qué?

- Una secuencia por si sola no tiene mucho valor
- Es necesario asignar límites dentro de una secuencia para definir donde yacen elementos funcionales del genoma, e.g., genes, rRNAs, tRNAs, lncRNAs, promotores, sitios de unión de proteínas, etc.

Anotación genómica es:

- La información misma
- El proceso de anotar
- El manejo de datos —> formato, almacenamiento, distribución

Dónde, qué y cómo

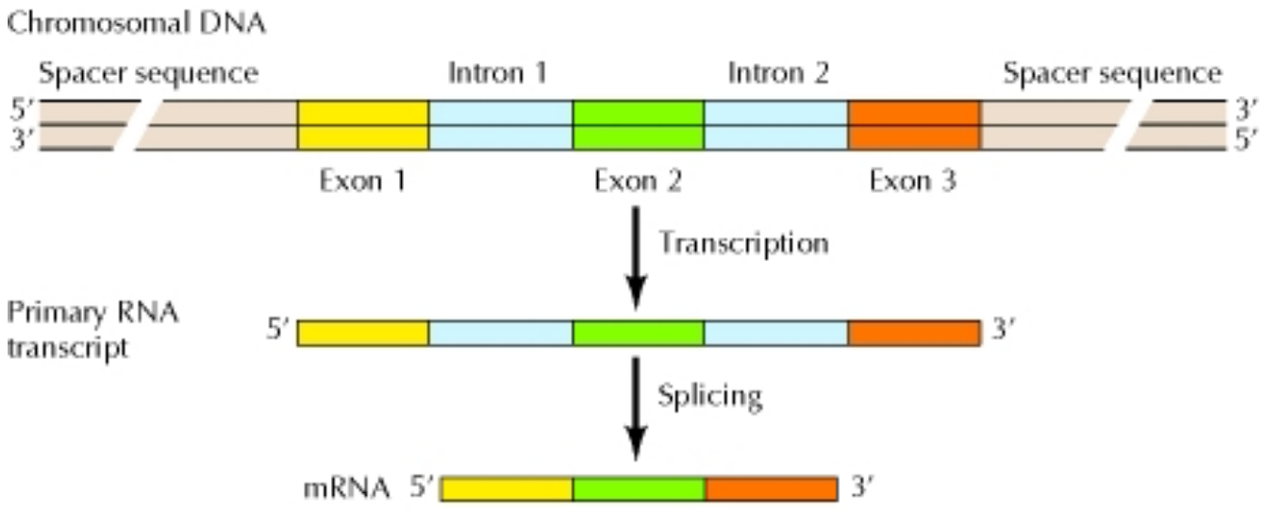


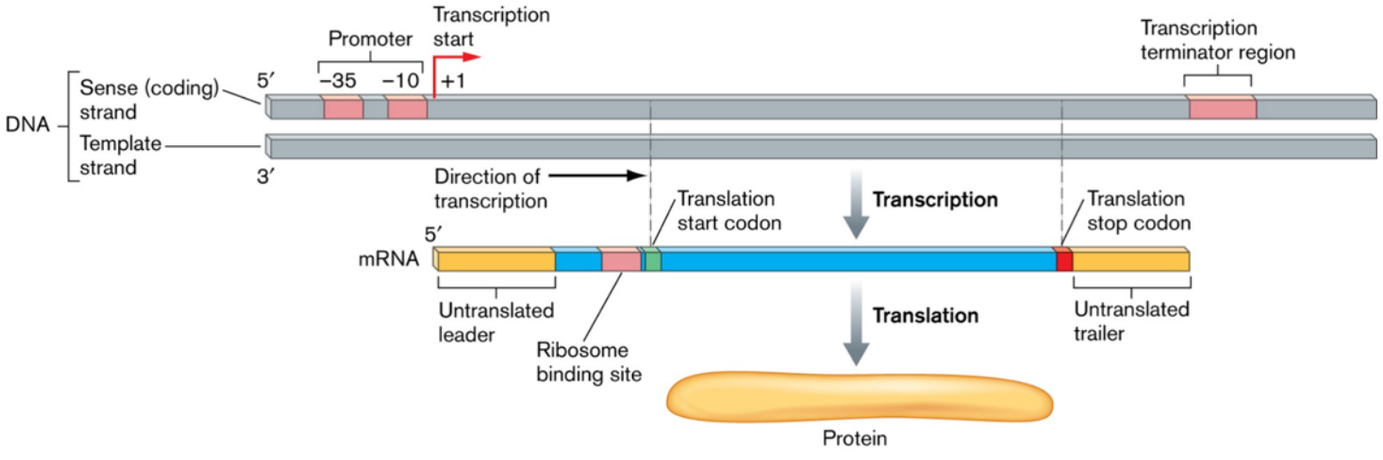
Nature Reviews Genetics **2**, 493-503 (July 2001) | doi:10.1038/35080529

Genome annotation: from sequence to biology

Lincoln Stein

Dónde: estructura de genes



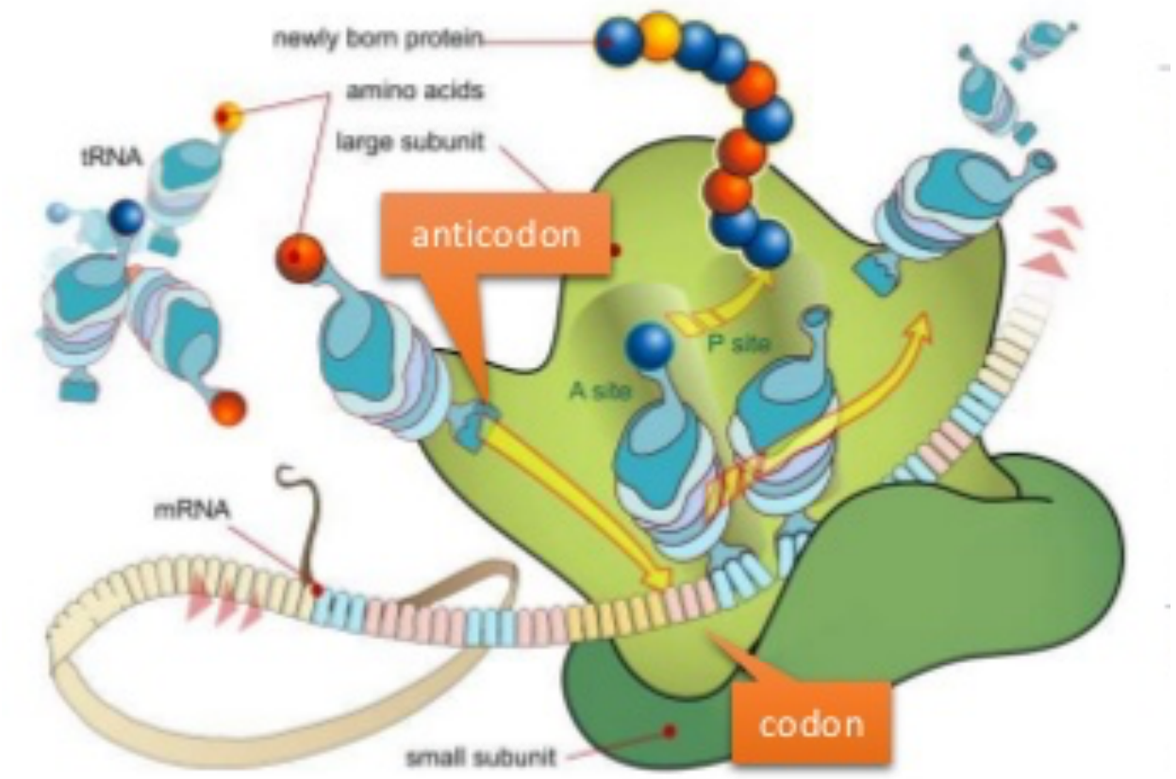


- ADN eucarionte envuelto en histonas, resulta en patrones repetitivos. Promotores están cerca de estos sitios

- Procaryotes no tienen intrones y regiones promotoras y codones de inicio están conservados

- Ambos difieren en uso de codones

Dónde: predicción de genes



- Gene finding/ gene prediction
- Uso de codones es especie específico
- Regiones funcionales como promotores, sitios de splicing, inicio de la traducción varian por especie

Dos metodologias clásicas

- ***Ab initio* o intrínsecos** —> solo a partir de la secuencia de DNA, busca señales inequívocas de la presencia de un gen o región de interés, e.g., codones de inicio/término, sitios de unión de factores de transcripción

- **Extrínsecos o por homología/evidencia** —> búsquedas en bases de datos curadas de proteínas, mRNAs o transcriptomas.

…En eucariontes

- **Extrínsecos o por homología/evidencia** —> búsquedas en bases de datos curadas de proteínas, mRNAs o transcriptomas.

- **Eucariontes** —> *Spliced Alignment* = usar información experimental (transcritos) y de homología (proteínas de referencia)

…En eucariontes

- **Eucariontes** —> *Spliced Alignment* = usar información experimental (transcritos) y de homología (proteínas de referencia)
  - Matches = exons
  - Gaps = introns
  - GeneSeqer; Exonerate; GenomeThreader

Ab initio

- **Procariontes** —> más estudiados, se sabe qué buscar y genomas presentan cierta regularidad

    - ORF largos flanqueados por codones de inicio y término. Virtualmente no hay secuencias intergénicas

- **Eucariontes** —> sabemos menos, altamente variables. Sitios de unión para colas de poliA, islas CpG. Intrones y secuencias intergénicas + splicing alternativo lo hacen más complicado

    - Ventaja = intrones son más ricos en A/T que en exones

Predicción de genes

- Modelos génicos
- Coordenadas de inicio y término de elementos genéticos
- En eucariontes, no hay exones sobrelapantes, exones deben estar en el mismo marco de lectura, al juntar dos exones no se debe formar un codon de término

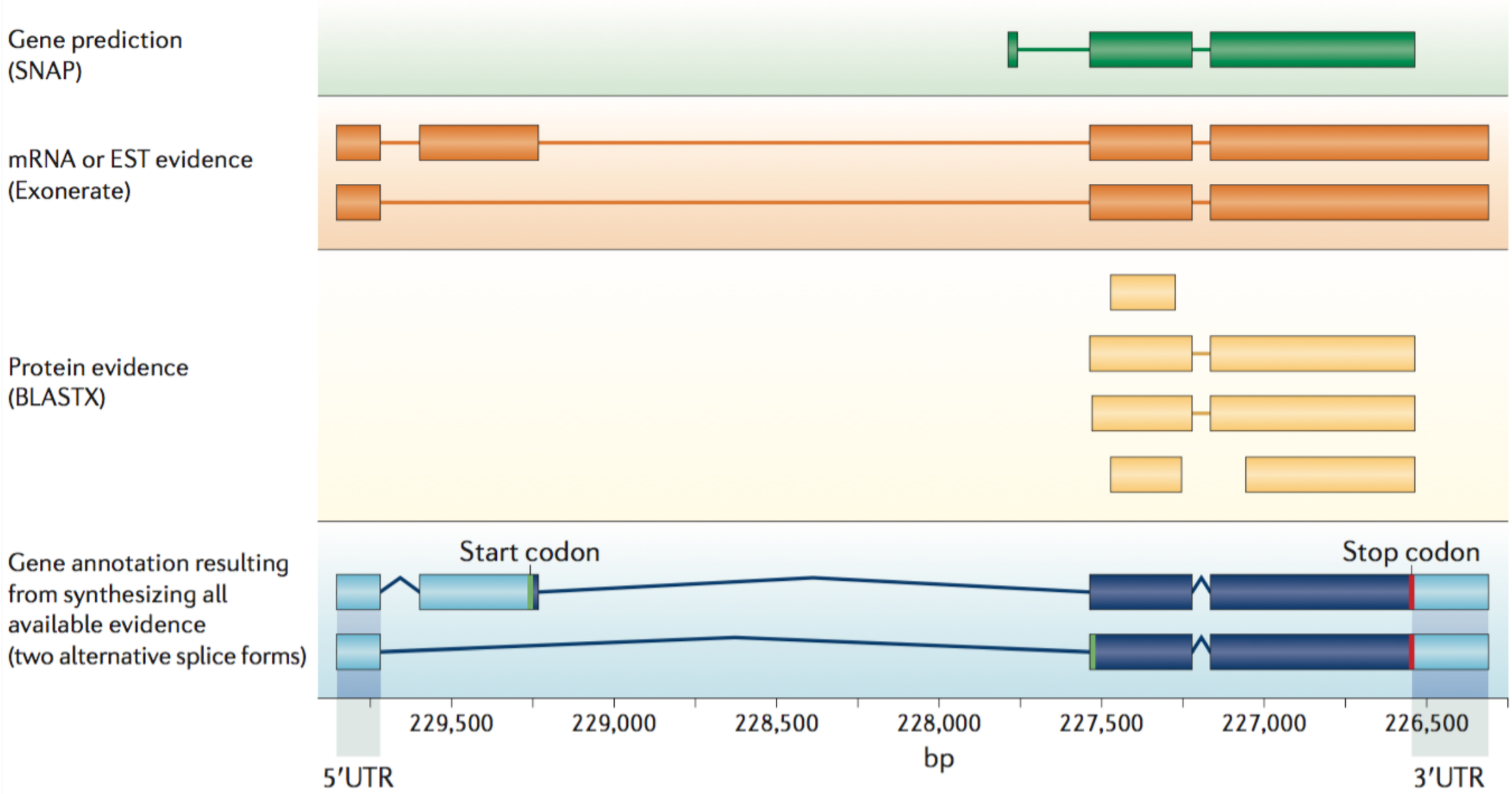
¿Qué tan bien funciona?

- **Procariontes** —> 50-70% por homología, resto *ab initio*. Difícil en genes que se sobrelapan

- **Eucariontes** —> 40% por homologia, resto *ab initio*. Refinación con RNASeq.

- Actualmente siempre se usa una combinación de distintos métodos y bases de datos para lograr mejores modelos génicos

Box 2 | **Gene prediction versus gene annotation**



Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. May 2012:1-14. doi:10.1038/nrg3174.

Implementaciones

- Augustus, GeneMark, SNAP en eucariontes
- Prokka (Glimmer, GeneMark, Extrínsecos) en procariontes
- tRNAScan/Aragorn
- Rfam y RNAmmer
- Hmmer

Estrategias por homología o extrinsecas

- Proteínas de referencia = Swissprot, Uniprot, RefSeq proteins. NR

- BLAST o BLAST-like programs

	Broad	**WUGSC**	**JCVI**	**BCM**
Blast Database	NR (bacteria)	NR (bacteria)	All Group NIAA-PANDA	NR (bacteria)
Min E value	10^{-10}	10^{-6} bit score=130	BlastP score cut off: 50 BlastP min E value: 0.1	10^{-5}
Min % identity	30%	30%	-	30%
Min query coverage	30%	30%	-	30%

http://hmpdacc.org/doc/CommonGeneAnnotation_SOP.pdf

Estrategias por homología o extrinsecas

- También se utilizan en “anotación funcional”
- función y metabolismo
- Gene Ontologies; KEGG pathways; OMA Browser; EggNOG database

¿Cuál estrategia es más adecuada?

- Usar siempre combinaciones de métodos intrínsecos y extrínsecos

- Sobre todo en eucariontes = inspeccionar visualmente las predicciones de genes, splicing junctions

- Implementaciones de software para evaluar ensamblajes/anotaciones = Apollo, yrGATE

- Incluso combinar anotaciones diferentes = MAKER; EvidenceModeler

"A man with a watch knows what time it is. A man with two watches is never sure." Segal's law

Option 1: predict

Option 2: predict and choose

Option 3: full-scale annotation pipelines

Run single *ab initio* gene predictor

Most likely CDS model for each gene

Run battery of *ab initio* gene predictors

Consensus-based chooser

Best consensus CDS model for each gene

Align ESTs, proteins and RNA-seq data to genome

Run battery gene predictors in evidence-driven mode

Consensus-based chooser

Best consensus CDS model for each gene

Post process gene predictions to add UTRs and alternatively spliced transcripts based on evidence

Consensus-based chooser

Best consensus mRNA model(s) for each gene

Evidence-based chooser

mRNA model(s) for each gene most consistent with evidence

Optional manual curation using genome browser

Manually curated gene models

Increasing time and effort

Increasing use of evidence

Increasing accuracy

Figure 2 | **Three basic approaches to genome annotation and some common variations.** Approaches are compared on the basis of relative time, effort and the degree to which they rely on external evidence, as opposed to *ab initio* gene models. The *y* axis shows increasing time and effort; the x axis shows increasing use of external evidence and, consequently, increasing accuracy and completeness of the resulting gene models. The type of final product produced by each kind of pipeline is shown in the dark blue boxes. Relative positions in the figure are for summary purposes only and are not based on precisely computed values. See TABLE 1 for a list of commonly used software components. CDS, coding sequence; EST, expressed sequence tag; RNA-seq, RNA sequencing; UTR, untranslated region.

¿Cuál estrategia es más adecuada?

First instantaneous, collaborative genomic annotation editor available on the Web.

Apollo is designed to support geographically dispersed researchers, and the work of a distributed community is coordinated through automatic synchronization: all edits in one client are instantly pushed to all other clients, allowing users to see annotation updates from collaborators in real-time during the editing process.

There are no installation requirements for Annotators!

Apollo is a plug-in for JBrowse, adding *User-created Annotations* and *DNA* tracks to the familiar main window.

Try Apollo at our Public Demo, built with the genome of several organism including the honey bee (*Apis mellifera*).

Software
Genome Biology
March 2006, 7:R58

First online: 19 July 2006

Open Access

yrGATE: a web-based gene-structure annotation tool for the identification and dissemination of eukaryotic genes

Matthew D Wilkerson, Shannon D Schlueter, Volker Brendel

¿Cuál estrategia es más adecuada?

BUSCO

QUEST FOR QUALITY

"BUSCO CALIDAD"

"BUSCO QUALIDADE"

Assessing genome assembly and annotation completeness with Benchmarking Universal Single-Copy Orthologs

Papers que deberían leer para interiorizarse

Review in Advance first posted online on April 22, 2015. (Changes may still occur before final publication online and in print.)

The Theory and Practice of Genome Sequence Assembly

Jared T. Simpson[1] and Mihai Pop[2]

[1]Ontario Institute for Cancer Research, Toronto, Ontario M5G ON3, Canada; email: jared.simpson@oicr.on.ca

[2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland 20742; email: mpop@umiacs.umd.edu

CSH PRESS GENOME RESEARCH

Assemblathon 1: A competitive assessment of de novo short read assembly methods

Dent Earl, Keith Bradnam, John St. John, et al.

Genome Res. 2011 21: 2224-2241 originally published online September 16, 2011
Access the most recent version at doi:10.1101/gr.126599.111

Bradnam *et al. GigaScience* 2013, **2**:10
http://www.gigasciencejournal.com/content/2/1/10

(GIGA)n SCIENCE

RESEARCH Open Access

Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species

Keith R Bradnam[1*†], Joseph N Fass[1†], Anton Alexandrov[36], Paul Baranay[2], Michael Bechner[39], Inanç Birol[33],

GAGE: A critical evaluation of genome assemblies and assembly algorithms

Steven L. Salzberg,[1,7] Adam M. Phillippy,[2] Aleksey Zimin,[3] Daniela Puiu,[1] Tanja Magoc,[1] Sergey Koren,[2,4] Todd J. Treangen,[1] Michael C. Schatz,[5] Arthur L. Delcher,[6] Michael Roberts,[3] Guillaume Marçais,[3] Mihai Pop,[4] and James A. Yorke[3]

STUDY DESIGNS

A beginner's guide to eukaryotic genome annotation

Mark Yandell and Daniel Ence

Abstract | The falling cost of genome sequencing is having a marked impact on the research community with respect to which genomes are sequenced and how and where they are annotated. Genome annotation projects have generally become small-scale affairs that are often carried out by an individual laboratory. Although annotating a eukaryotic genome assembly is now within the reach of non-experts, it remains a challenging task. Here we provide an overview of the genome annotation process and the available tools and describe some best-practice approaches.

D286–D293 *Nucleic Acids Research, 2016, Vol. 44, Database issue* *Published online 17 November 2015*
doi: 10.1093/nar/gkv1248

eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences

Jaime Huerta-Cepas[1], Damian Szklarczyk[2,3], Kristoffer Forslund[1], Helen Cook[4], Davide Heller[2,3], Mathias C. Walter[5], Thomas Rattei[6], Daniel R. Mende[7], Shinichi Sunagawa[1], Michael Kuhn[8], Lars Juhl Jensen[4], Christian von Mering[2,3,*] and Peer Bork[1,9,10,*]

“If I had more time, I would have written a shorter letter” Blaise Pascal