

Curso de Métodos Estadísticos y Analíticos de Datos Genómicos

Leonardo Collado Torres

lcollado@ibt.unam.mx y lcollado@wintermexico.com

Lic. en Ciencias Genómicas

`www.lcg.unam.mx/~lcollado/`

Winter Genomics (WG) e Instituto de Biotecnología (IBT) de la UNAM

21 de Enero de 2010

Software para Datos de Secuenciación Masiva

- 1 Intro
- 2 Soluciones
- 3 La realidad
- 4 Tipos de análisis
- 5 Alineadores
- 6 Un caso de RNA-seq eucarionte

Por fin tenemos datos. . . ahora al análisis

Inicio

Intro

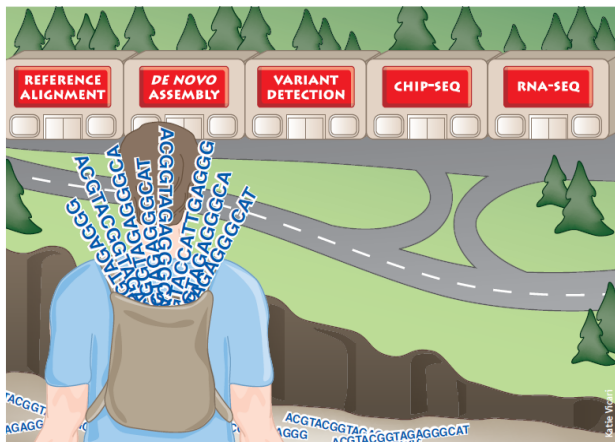
Soluciones

La realidad

Tipos de
análisis

Alineadores

Un caso de
RNA-seq
eucarionte



A gap exists between current sequence-generation and data-analysis capabilities.

¿Cuello de botella?

Next-generation gap

John D McPherson

There is a growing gap between the generation of massively parallel sequencing output and the ability to process and analyze the resulting data. New users are left to navigate a bewildering maze of base calling, alignment, assembly and analysis tools with often incomplete documentation and no idea how to compare and validate their outputs. Bridging this gap is essential, or the coveted \$1,000 genome will come with a \$20,000 analysis price tag.

El otro *problema*

- A Illumina, y me imagino que las otras, no les interesa desarrollar algoritmos que faciliten el análisis más allá de un nivel que consideran *suficiente*. Excepto en el análisis de imágenes.
- Lo dejan todo en manos de los investigadores :)

Además, todo cambia **MUY** rápido!!

Illumina's Cheap New Gene Machine

Matthew Herper, 01.12.10, 03:00 PM EST

A new DNA reader could turbocharge research into cancer and autism.



The biotech company **Illumina** is introducing a new machine that it says will decode a person's DNA in one week using \$10,000 worth of materials--five times cheaper than any other competing gadget on the market

The new machine, the HiSeq2000, will begin shipping next month with a cost of \$690,000 vs. \$500,000 for Illumina's current model. It is being unveiled today at J.P. Morgan's investment conference in San Francisco. The Beijing Genomics Institute will be the first customer, purchasing 128 of the new machines.

- $128 * 690000 = 88320000$ USD
- Broad, BGI, Sanger

¿México?

Slim invertirá 75 mdd en medicina genómica

En colaboración con el Instituto Nacional de Medicina Genómica de México y el *Broad Institute* secuenciarán el genoma del cáncer y la diabetes

Soluciones Integrales

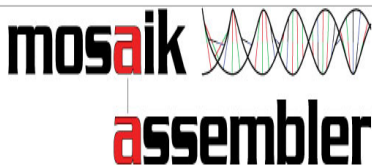
- Bioconductor
- CLCbio Genomics Workbench
- Mosaik
- Programas de las compañías de secuenciación: ELAND, Newbler, entre otros.
- SHORE...

CLCbio Genomics Workbench

CLC bio - the world's leading bioinformatics solution provider!

CLC bio is the world's leading bioinformatics solution provider. Next Generation Sequencing is a major focus area and CLC bio delivers the first and only comprehensive cross-platform analysis solution, which can analyze and visualize genomic, transcriptomic, and epigenomic data from all major platforms, like Illumina's Genome Analyzer, SOLiD by Applied Biosystems, 454 by Roche, and HeliScope by Helicos.

El problema principal es el precio, aunque está disponible en modo de prueba.

The logo for Mosaik assembler features the word "mosaik" in a bold, black, sans-serif font, with the "ai" in red. To the right of the text is a stylized graphic of a DNA double helix with red, green, and blue strands. Below "mosaik" is the word "assembler" in a larger, bold, black, sans-serif font, with the "a" in red. A thin vertical line connects the "i" in "mosaik" to the "a" in "assembler".

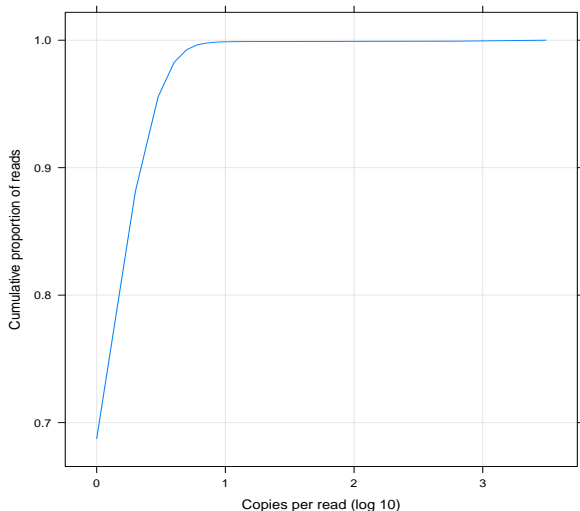
mosaik assembler

MOSAİK is a reference-guided assembler comprising of four main modular programs:

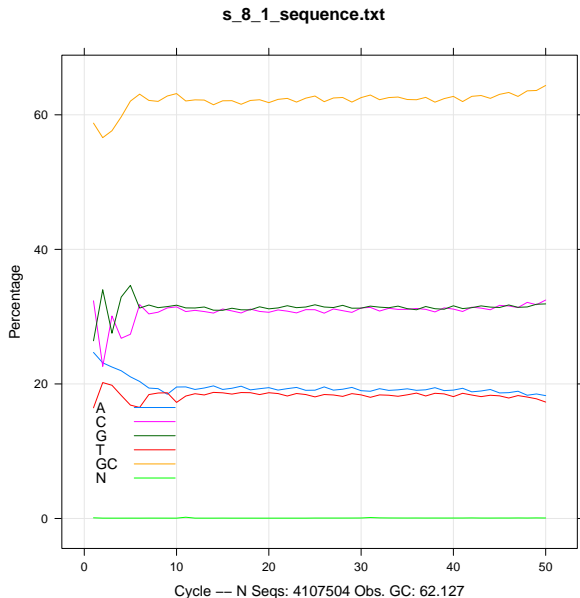
- 📦 MosaikBuild
- 📦 MosaikAligner
- 📦 MosaikSort
- 📦 MosaikAssembler.

- Es **muy** nuevo así que siguen arreglando errores, pero tiene potencial.
- ¿Pero se nos olvida algo?

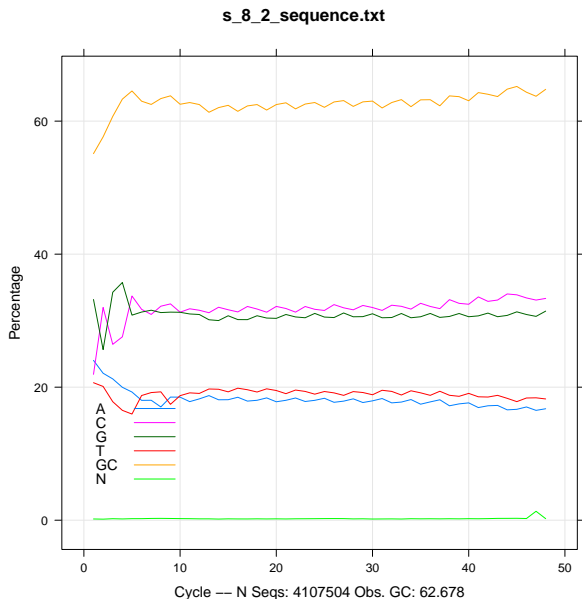
¿Tenemos muchas o pocas secuencias únicas?



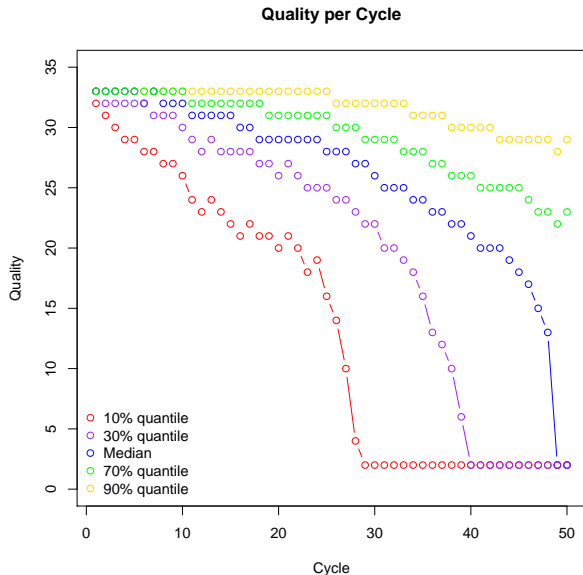
Frecuencia de NTs



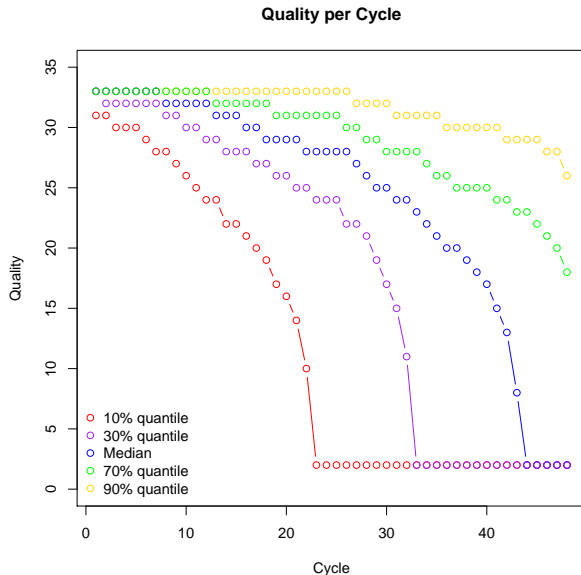
Frecuencia de NTs del otro par



¿Y la calidad?

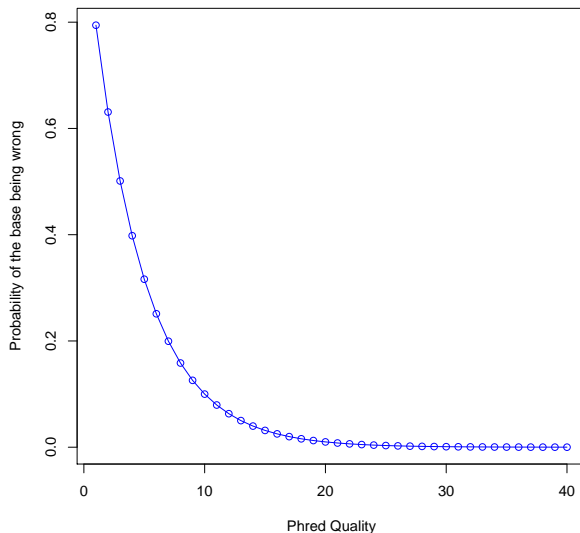


El par no pinta bien



Para entender mejor

Phred Quality to Probility

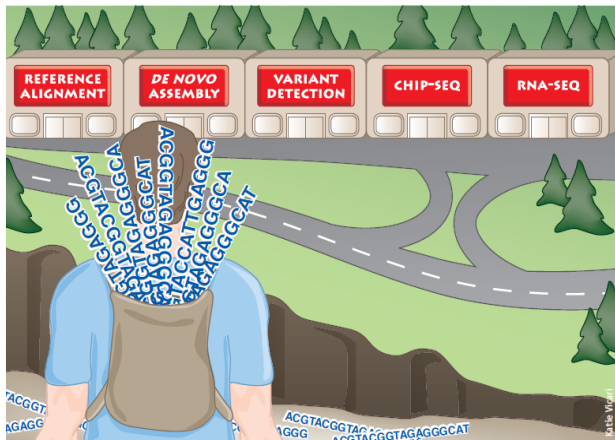


Así que...

Hay que filtrar!

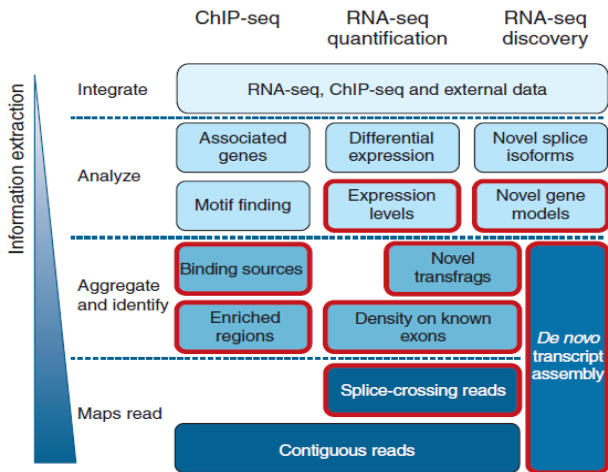
- Por calidad.
- Por un ciclo dado.
- Por la presencia de Ns.
- Las secuencias compuestas primordialmente de una sola base.
- Eliminar secuencia de adaptadores.
- Calidad del alineamiento.
- Lo que inventen :)

Recordando



A gap exists between current sequence-generation and data-analysis capabilities.

ChIP y RNA - seq



RNA-seq

a

De novo assembly of the transcriptome

Highly expressed gene



Lowly expressed gene



Read coverage must
be high enough to build
EST contigs (solid bar)

b

Map onto the genome



Read mapper must
support splitting reads
to record splices

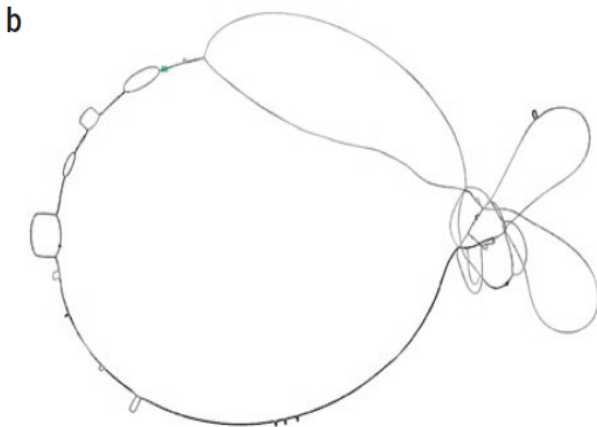
c

Map onto the genome and splice junctions

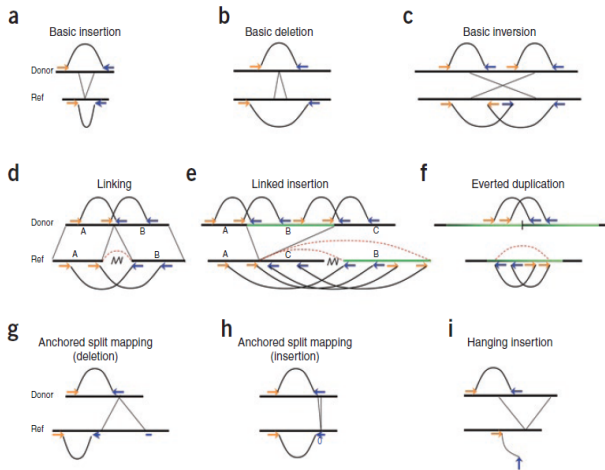


Splice junction
sequences from
either annotations
or inferred

Ensamblado de novo



Variación estructural



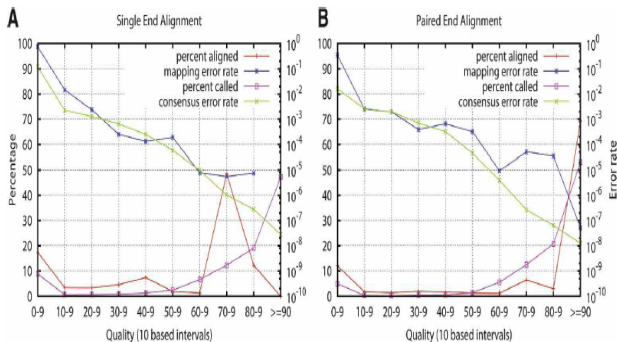
Lo básico

- Como pueden ver, hay que **filtrar** y **alinear** en todo tipo aplicación.

El famoso MAQ

- Todo gracias a **Heng Li** del Sanger.
- El primero en usar las calidades al momento de alinear.
- Bastante rápido.
- Trae un identificador de SNPs.
- Visualizador de alineamientos asociado: mapview.
- Muy bien documentado.

El famoso MAQ



El famoso MAQ

Table 2 | Sequencing statistics on personal genome projects

Personal Genome	Platform	Genomic template libraries	No. of reads (millions)	Read length (bases)	Base coverage (fold)	Assembly	Genome coverage (%) [*]	SNVs in millions (alignment tool)	No. of runs	Estimated cost (US\$)
J. Craig Venter	Automated Sanger	MP from BACs, fosmids & plasmids	31.9	800	7.5	De novo	N/A	3.21	>340,000	70,000,000
James D. Watson	Roche/454	Frag: 500 bp	93.2 [†]	250 [§]	7.4	Aligned [*]	95 [‡]	3.32 (BLAT)	234	1,000,000 [†]
Yoruban male (NA18507)	Illumina/Solexa	93% MP: 200 bp 7% MP: 1.8 kb	3,410 [†] 271	35 35	40.6	Aligned [*]	99.9	3.83 (MAQ) 4.14 (ELAND)	40	250,000 [†]
Han Chinese male	Illumina/Solexa	66% Frag: 150–250 bp 34% MP: 135 bp & 440 bp	1,921 [†] 1,029	35 35	36	Aligned [*]	99.9	3.07 (SOAP)	35	500,000 [†]
Korean male (AK1)	Illumina/Solexa	21% Frag: 130 bp & 440 bp 79% MP: 130 bp, 390 bp & 2.7 kb	303 [†] 1,156	36 36, 88, 106	27.8	Aligned [*]	99.8	3.45 (GSNAP)	30	200,000 [†]
Korean male (SJK)	Illumina/Solexa	MP: 100 bp, 200 bp & 300 bp	1,647 [†]	35, 74	29.0	Aligned [*]	99.9	3.44 (MAQ)	15	250,000 ^{†*}
Yoruban male (NA18507)	Life/APG	9% Frag: 100–500 bp 91% MP: 600–3,500 bp	211 [†] 2,075 [†]	50 25, 50	17.9	Aligned [*]	98.6	3.87 (Corona-lite)	9.5	60,000 ^{†**}
Stephen R. Quake	Helicos BioSciences	Frag: 100–500 bp	2,725 [†]	32 [§]	28	Aligned [*]	90	2.81 (IndexDP)	4	48,000 [†]
AML female	Illumina/Solexa	Frag: 150–200 bp ^{††} Frag: 150–200 bp ^{§§}	2,730 ^{††} 1,081 ^{§§}	32 35	32.7 13.9	Aligned [*]	91 83	3.81 ^{††} (MAQ) 2.92 ^{§§} (MAQ)	98 34	1,600,000
AML male	Illumina/Solexa	MP: 200–250 bp ^{††} MP: 200–250 bp ^{§§}	1,620 ^{††} 1,351 ^{§§}	35 50	23.3 21.3	Aligned [*]	98.5 97.4	3.46 ^{††} (MAQ) 3.45 ^{§§} (MAQ)	16.5 13.1	500,000
James R. Lupski CMT male	Life/APG	16% Frag: 100–500 bp 84% MP: 600–3,500 bp	238 [†] 1,211 [†]	35 25, 50	29.6	Aligned [*]	99.8	3.42 (Corona-lite)	3	75,000 ^{††}

Salmonella Typhi con MAQ

Inicio

Intro

Soluciones

La realidad

Tipos de
análisis

Alineadores

Un caso de
RNA-seq
eucarionte

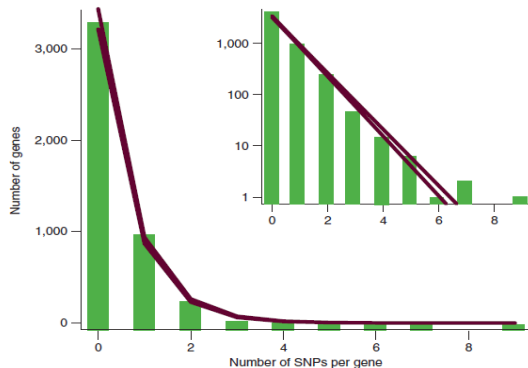
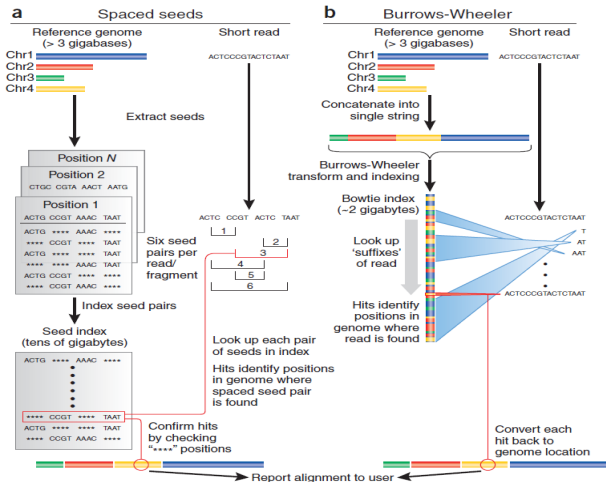
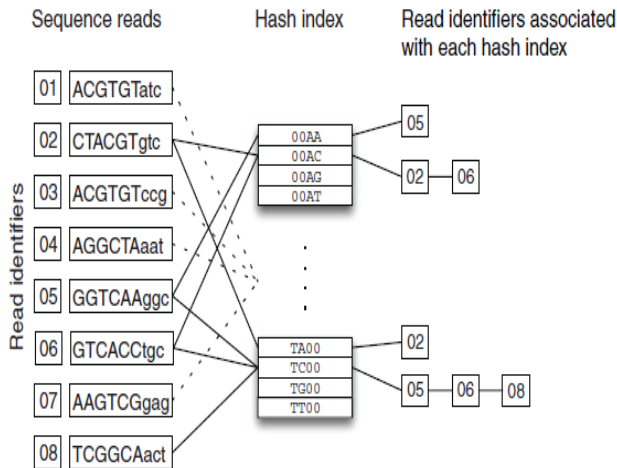


Figure 3 Distribution of number of SNPs per gene. Lines indicate 95% confidence interval of mean predicted values under a Poisson distribution fitted to the data shown in green. Inset shows gene count on a log scale to better show deviation from the Poisson model at high numbers of SNPs per gene.

Hay dos grandes categorías de alineadores



Hash Index - Spaced Seeds



- MAQ, SHRiMP, ELAND, SOAP, MOSAIK, ZOOM, BFAST, ...

Burrows-Wheeler Transform



^TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG\$

Genomic sequence

GGTTGGTCGGATTCCGAATCACGGAAAATT^AGATTCCSG

Transform

- Bowtie, BWA, SOAP2, ...
- Generalmente son **MUCHO** más **rápidos**.

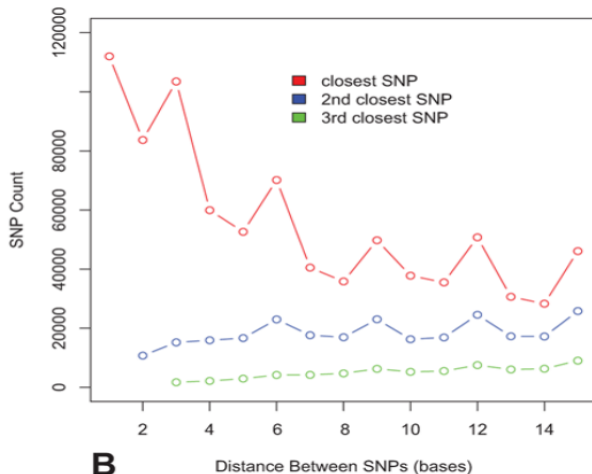
SHRiMP

- De los primeros en poder manejar datos de SOLiD.
- Implementa un alineamiento Smith-Waterman en el proceso. Aumenta la precisión.

Table 3. Color-space mapping accuracy of SHRiMP.

		Number of SNPs									
		0	1		2	3		4			
		Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
	0	85.7	83.2	84.8	81.3	83.5	76.6	80.6	65.2	75.6	46.8
Max	1	83.8	79.4	82.2	74.0	79.4	62.6	72.8	43.2	63.1	24.7
Indel	2	83.2	77.1	80.8	69.6	77.9	56.6	68.2	36.4	56.4	18.9
Length	3	80.7	71.0	79.6	64.2	73.6	48.3	66.5	31.5	57.1	16.6
	4	78.0	65.4	76.5	56.1	71.4	41.9	60.6	23.9	50.3	12.4
	5	75.9	58.9	73.0	48.1	69.7	36.6	57.0	21.3	46.0	12.7

¿Qué notan?



Son datos de *Ciona savignyi*

- Desarrollado por Ben Langmead
- Extremadamente rápido
- Similar a MAQ en la forma de uso
- Basado en el BWT
- Corre en paralelo

Bowtie vs otros

Table 3

Varying read length using Bowtie, Maq and SOAP

Length	Program	CPU time	Wall clock time	Peak virtual memory footprint (megabytes)	Bowtie speed-up	Reads aligned (%)
36 bp	Bowtie	6 m 15 s	6 m 21 s	1,305	-	62.2
	Maq	3 h 52 m 26 s	3 h 52 m 54 s	804	36.7×	65.0
	Bowtie -v 2	4 m 55 s	5 m 00 s	1,138	-	55.0
	SOAP	16 h 44 m 3 s	18 h 1 m 38 s	13,619	216×	55.1
50 bp	Bowtie	7 m 11 s	7 m 20 s	1,310	-	67.5
	Maq	2 h 39 m 56 s	2 h 40 m 9 s	804	21.8×	67.9
	Bowtie -v 2	5 m 32 s	5 m 46 s	1,138	-	56.2
	SOAP	48 h 42 m 4 s	66 h 26 m 53 s	13,619	691×	56.2
76 bp	Bowtie	18 m 58 s	19 m 6 s	1,323	-	44.5
	Maq 0.7.1	4 h 45 m 7 s	4 h 45 m 17 s	1,155	14.9×	44.9
	Bowtie -v 2	7 m 35 s	7 m 40 s	1,138	-	31.7

Bowtie: en paralelo

Table 4

Bowtie parallel alignment performance

	CPU time	Wall clock time	Reads mapped per hour (millions)	Peak virtual memory footprint (megabytes)	Speedup
Bowtie, one thread	18 m 19 s	18 m 46 s	28.3	1,353	-
Bowtie, two threads	20 m 34 s	10 m 35 s	50.1	1,363	1.77x
Bowtie, four threads	23 m 9 s	6 m 1 s	88.1	1,384	3.12x

Bowtie: creando índices

Table 5

Bowtie index building performance

Physical memory target (GB)	Actual peak memory footprint (GB)	Wall clock time
16	14.4	4 h 36 m
8	5.84	5 h 5 m
4	3.39	7 h 40 m
2	1.39	21 h 30 m

Resumiendo

Table 1 A selection of short-read analysis software

Program	Website	Open source?	Handles ABI color space?	Maximum read length
Bowtie	http://bowtie.cbc.b.umd.edu	Yes	No	None
BWA	http://maq.sourceforge.net/bwa-man.shtml	Yes	Yes	None
Maq	http://maq.sourceforge.net	Yes	Yes	127
Mosaik	http://bioinformatics.bc.edu/marthlab/Mosaik	No	Yes	None
Novoalign	http://www.novocraft.com	No	No	None
SOAP2	http://soap.genomics.org.cn	No	No	60
ZOOM	http://www.bioinform.com	No	Yes	240

Siempre queremos más :)

These technologies generate relatively short reads, typically from a few tens to a few hundred bases in length, with a general inverse relation between the total number of reads and the read length. In the context of whole human genome resequencing, on the order of a billion short reads are required to accurately resequence an individual genome, and this creates an unprecedented alignment problem of aligning this many reads to the reference human genome on a practical timescale of days. Using established dynamic programming algorithms [7] to align reads to the entire human genome is grossly impractical, since the computational cost is proportional to the target size. To reduce the cost resulting from

```
> (1e+09 * 50)/(2 * 3e+09)
```

```
[1] 8.333333
```

```
> 1e+09/(2e+07 * 8)
```

```
[1] 6.25
```


- Desarrollado por Nils Homer
- Rápido aunque no tanto como Bowtie
- Utiliza un alineamiento tipo Smith-Waterman. Es mucho más sensible!
- Acepta datos de SOLiD
- Mucho más robusto que cualquier otro **so far**
- Bastante nuevo.

BFAST: el plan

Step 1: Index the reference



Step 2: Find CALs using the index(es)



Step 3: Gapped local alignment

BFAST: creando un índice

Inicio

Intro

Soluciones

La realidad

Tipos de
análisis

Alineadores

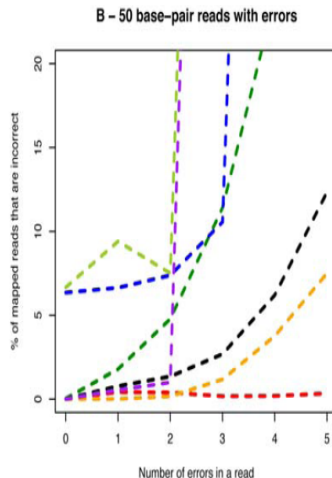
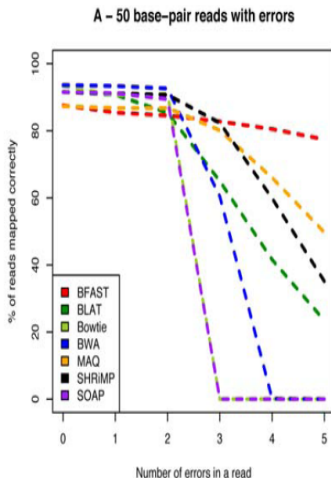
Un caso de
RNA-seq
eucarionte

A		C			
start		2-mer	order start	order end	
	AATCCCGATTACAGGGATT				
1	AATCCCGATTACAGGGATT	1. AA	-	-	
2	ATCCCGATTACAGGGATT	2. AC	1	1	
3	TCCCGATTACAGGGATT	3. AG	-	-	
4	CCCGATTACAGGGATT	4. AT	2	3	
5	CCGATTACAGGGATT	5. CA	4	4	
6	CGATTACAGGGATT	6. CC	5	5	
7	GATTACAGGGATT	7. CG	6	6	
8	ATTACAGGGATT	8. CT	-	-	
9	TTACAGGGATT	9. GA	-	-	
		10. GC	-	-	
		11. GG	-	-	
		12. GT	7	7	
		13. TA	8	8	
		14. TC	9	9	
		15. TG	-	-	
		16. TT	-	-	

B		D	
order	start	offset	match
	10100000011		
1.	2	ATCCCGATTACAGGGATT	
2.	8	ATTACAGGGATT	
3.	1	AATCCCGATTACAGGGATT	
4.	6	CGATTACAGGGATT	
5.	4	CCCGATTACAGGGATT	
6.	5	CCGATTACAGGGATT	
7.	7	GATTACAGGGATT	
8.	9	TTACAGGGATT	
9.	3	TCCCGATTACAGGGATT	

offset	ATCCCGATTATAG	match
0	A C : : : : : A T	X
1	T C : : : : : T A	X
2	C C : : : : : A G	✓

BFAST vs otros



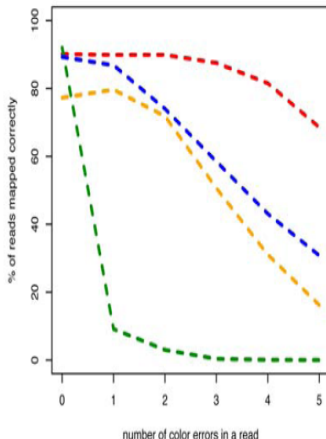
BFAST vs otros con datos reales

Table 2. Timing results of alignment algorithms on four different real-world datasets.

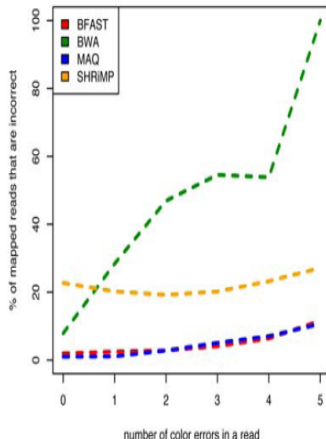
	Illumina 10.9 M 36 bp reads		Illumina 10.9 M 36 bp reads		Illumina 3.5 M 55 bp reads		Illumina 3.5 M 55 bp reads		ABI SOLiD 1 M 25 bp read	ABI SOLiD 1 M 25 bp read	ABI SOLiD 1 M 50 bp read	ABI SOLiD 1 M 50 bp read
	Time (s)	% mapped	Time (s)	% mapped	Time (s)	% mapped	Time (s)	% mapped	Time (s)	% mapped	Time (s)	% mapped
BFAST	43,775	32.1	47,474	69.6	9,590	66	42,856	72.5				
BLAT*	68,758	24.3	6,735,069	77.4	NA	NA	NA	NA				
Bowtie	2,270	13.1	857	55.7	NA	NA	NA	NA				
BWA	7,682	16	4,883	59.3	21,179	74.7	845	47.8				
MAQ	8,607	28.7	126,541	73.6	7,602	63.6	6,680	68.1				
SHRIMP*	186,764	14.9	324,380	83.3	2,977	2.4	32,644	70.4				
SOAP	11,938	13.3	131,248	62.4	NA	NA	NA	NA				

BFAST y color errors

A – 50bp reads with 1 SNP & 0–5 color errors



B – 50bp reads with 1 SNP & 0–5 color errors



Para todos gustos

Acuérdense de fijarse

- La sensibilidad.
 - La velocidad, si es crucial.
 - La memoria requerida.
 - Exploren los parámetros.
 - Chequen que estén bien los archivos de entrada. Cada programa puede usar uno diferente.
- Cuidado** con secuencias *paired-end* y *mate-pair*.

El tío SAM

- **Heng Li** y otros desarrollaron SAMtools que entre otras funciones, su objetivo es unificar formatos de salida de los alineadores.
- Está muy relacionado a su hermano BAMtools.
- <http://samtools.sourceforge.net/>

- Having a **large memory machine** - whatever route used here - is always useful. I would buy the largest machine which still has a reasonable linear trend of memory cost (at one goes up in memory, there is often a sharp increase in cost which is not linear. Buy just below that with lots of cores). This machine therefore, with the right number of cores, can be part of a *standard* farm without much cost penalty and can be used for these other tasks. This is often a **128GB** or **256GB** machine, but...you need to talk to vendors.

Programas a usar - Ben Langmead

Inicio

Intro

Soluciones

La realidad

Tipos de
análisis

Alineadores

Un caso de
RNA-seq
eucarionte

① TopHat

- ▶ Alinea las secs. para identificar uniones exón-exón
- ▶ Las secs. tienen que ser del mismo tamaño y no identifica indeles menores a cierto umbral.
- ▶ O todas son PE o todas son SE.
- ▶ No usa genoma de referencia.
- ▶ Usa Bowtie para identificar exones potenciales.
- ▶ Construye una db de posibles uniones y luego las confirma (3 tipos de evidencia).

② Cufflinks

- ▶ Ensambla secs. alineadas en transcritos y estima su abundancia.
- ▶ Mide la abundancia en RPKM: *reads per pk of exon model per million mapped reads*
- ▶ La version actual es beta.

Análisis

- Nos dan los datos: 100mil secuencias de 36pb de un experimento RNA-seq de *Drosophila melanogaster*.
- Leer en R, explorar las secuencias y filtrar: las que tienen Ns y no son del cromosoma. Nos quedamos con 55 %¹
- Calcular la cobertura con IRanges.
- Obtener la anotación del genoma usando biomaRt para conectarnos a ENSEMBL.
- Pasar la anotación a un objeto de IRanges.
- Calcular la cobertura por exón y por transcrito.
- Visualizar la cobertura en un *Genome Browser* como el de UCSC.
- Excluyendo solo las secs con Ns, corremos TopHat.
- Visualizar los archivos WIG y BED en un *Genome Browser*.
- Tan tan!

¹Podrían usar BioPython entre otras opciones

Referencias

- Next Generation Sequencing Analysis Focus de *Nature Methods*
- Trapnell y Salzberg
- Noticia Illumina
- Noticia Slim
- Artículo de MAQ
- Artículo Bowtie
- Artículo BFAST
- Artículo sobre Salmonella Typhi
- Software List de SEQanswers - **muy** útil!
- Metzker - Review
- Comunicación personal con Nicolas Delhomme

Información de mi sesión:

```
> sessionInfo()
```

```
R version 2.10.0 (2009-10-26)
```

```
i386-pc-mingw32
```

```
locale:
```

```
[1] LC_COLLATE=English_United States.1252
```

```
[2] LC_CTYPE=English_United States.1252
```

```
[3] LC_MONETARY=English_United States.1252
```

```
[4] LC_NUMERIC=C
```

```
[5] LC_TIME=English_United States.1252
```

```
attached base packages:
```

```
[1] stats      graphics  grDevices
```

```
[4] utils      datasets  methods
```

```
[7] base
```