Secuenciación

Rodrigo Santamaría

Secuenciación

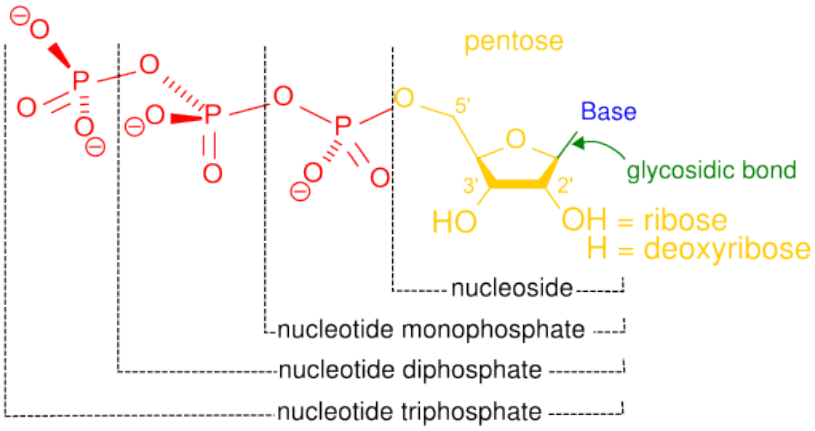
- Secuenciación de ADN
  - Secuenciación Sanger
  - Pirosecuenciación
  - Fabricantes
- Secuenciación de Genomas
- Secuenciación de Transcriptomas
- Perspectivas

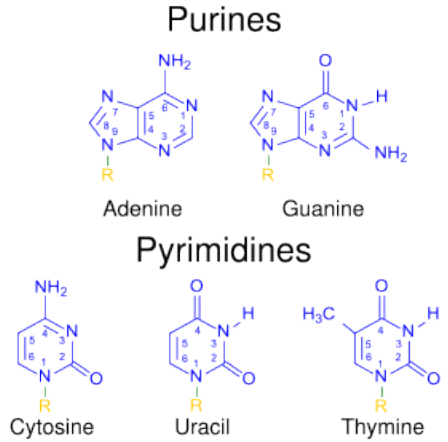
Secuenciación de ADN

- Secuenciación de ADN:
  - Determinar la secuencia de nucleótidos de una muestra de ADN
- Primeras técnicas de secuenciación
  - Método de Maxam-Gilbert (1976)
  - Método de terminación de cadena de Sanger (1977)
    - Se terminó imponiendo por ser más sencillo y preciso
    - Se basa en el uso de dideoxinucleótidos trifosfato (ddNTPs)
      - Nucleótido al que le faltan dos grupos –OH
      - De este modo, al entrar en contacto con la DNA-polimerasa ésta no puede enlazar otros nucleótidos al ddNTP

Secuenciación de Sanger

- **primer**: secuencia corta de ADN (unos 20 nucleótidos) que sirve de punto de inicio para sintetizar una secuencia
- **plantilla**: secuencia simple de ADN que queremos copiar
- **deoxinucleótido trifosfato** (dNTP): nucleótido unido a un trifosfato. Sanger usa en su lugar ddNTPs, sin los dos grupos -OH

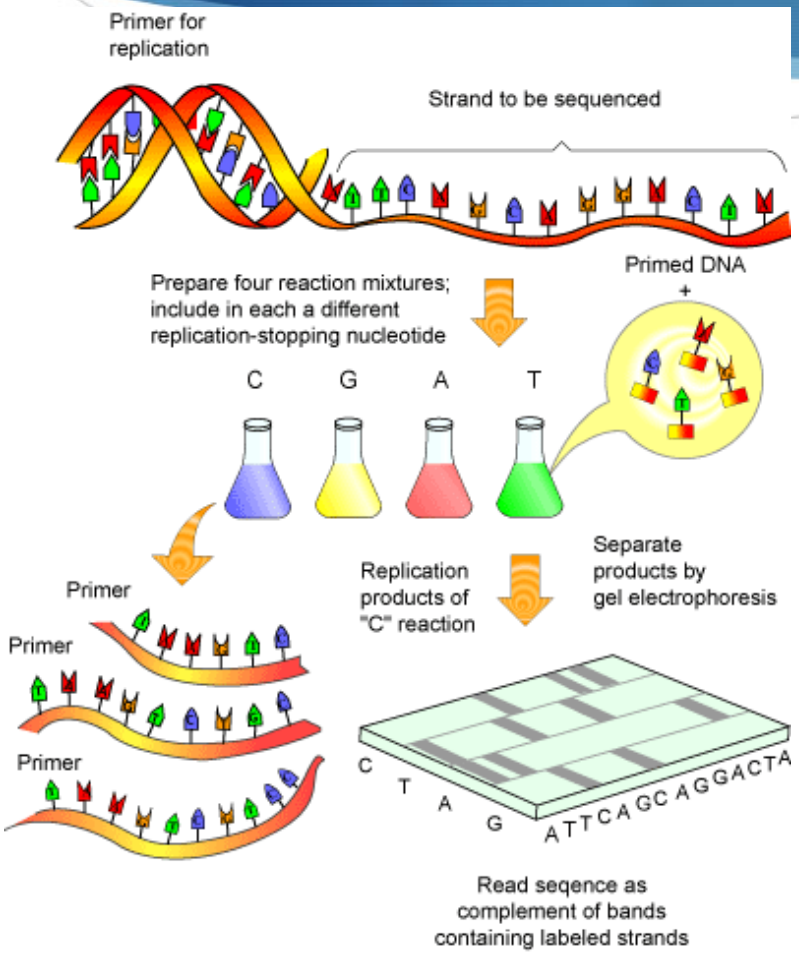


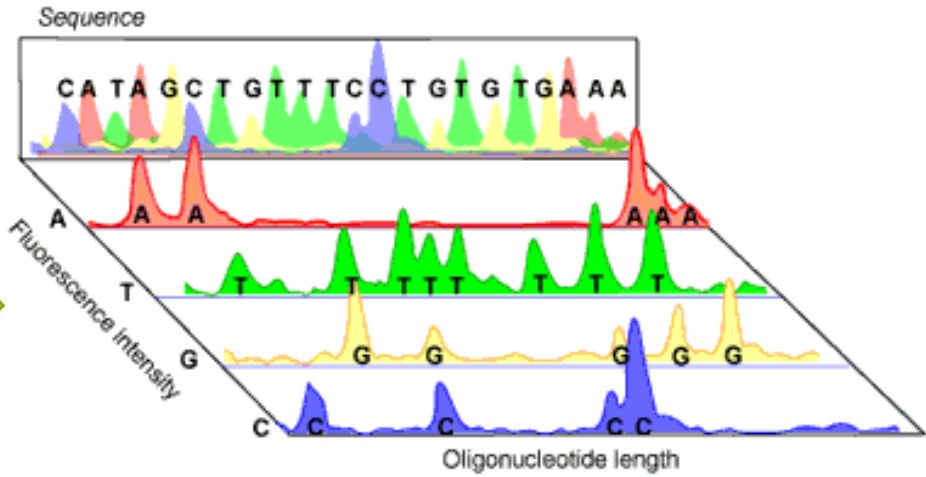


Secuenciación de Sanger

- Para la secuenciación necesitaremos
  - Una plantilla a copiar
  - Un primer para empezar la copia
  - DNA polimerasa para hacer la copia
  - ddNTPs que añadir al primer
    - Marcados con fluorescencia o radiactividad
    - Finalizan la copia tras añadirse
    - Tras varias iteraciones podemos tener copias de la longitud que queramos
- Se realizan cuatro reacciones separadas, cada una con un ddNTP distinto y marcado (ddATP, ddCTP, ddGTP, ddTTP)
  - El resultado es una secuencia complementaria a la plantilla, comenzando con el primer, marcada de distinta manera en cada nucleótido, y de la longitud que queramos según el número de reacciones que hagamos

Secuenciación Sanger

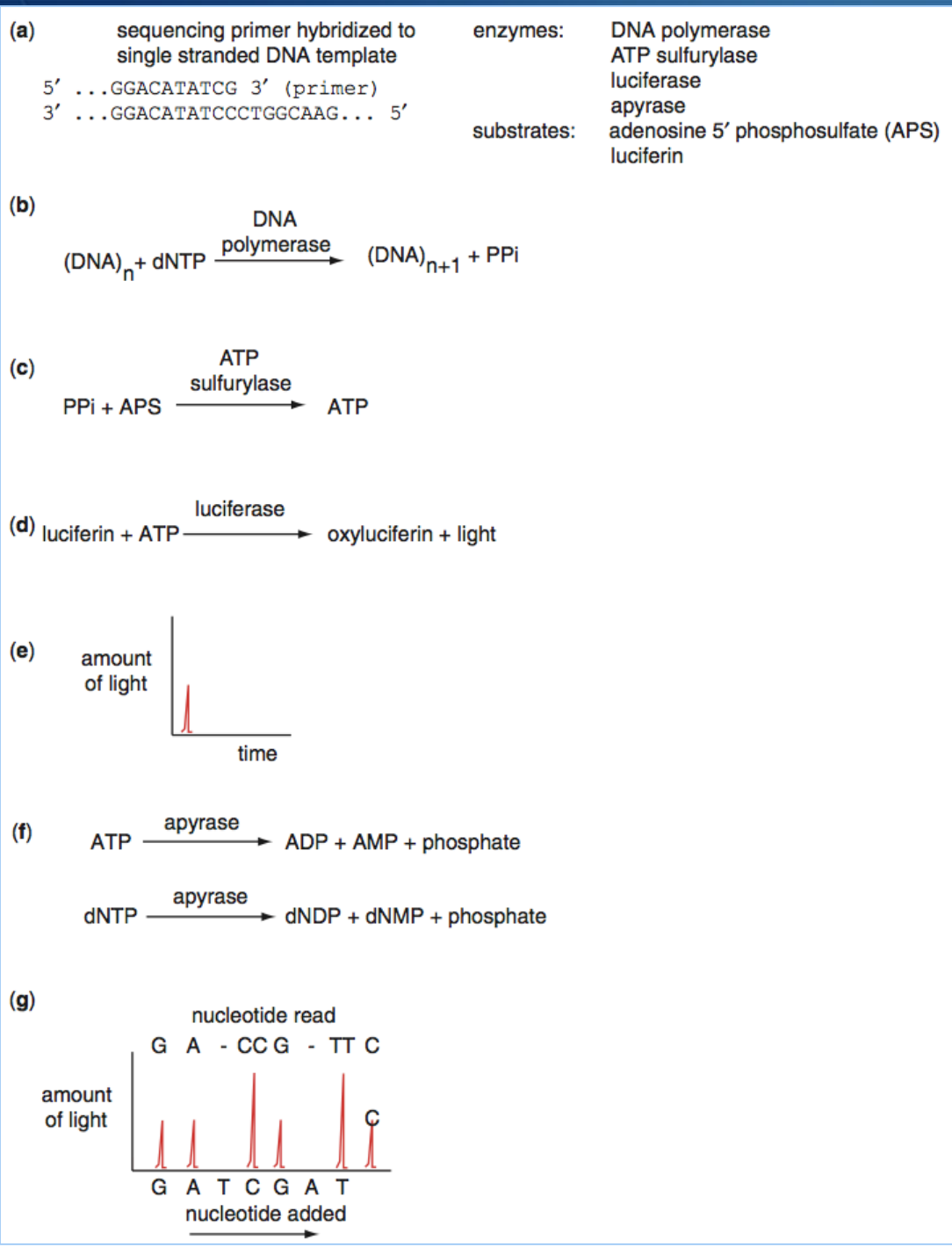




- http://www.scq.ubc.ca/genome-projects-uncovering-the-blueprints-of-biology/

Pirosecuenciación

- Una alternativa más moderna (1988) a la secuenciación Sanger, más rápida y barata, aunque menos precisa
  - Aprovecha el trifosfato que se libera cuando el dNTP se une a la cadena, en forma de pirofosfato (PPi)
  - Requiere de tres enzimas adicionales: ATP-sulfurilasa, luciferasa y apirasa
  - Requiere de dos sustratos adicionales: adenosin-fosfosulfato (APS) y luciferina
  - Con estas enzimas vamos a tener reacciones que de manera natural generan señales luminosas, ahorrándonos las marcas fluorescentes, y haciendo el proceso más rápido y barato



a) Partimos, como en Sanger, de una secuencia plantilla y un primer, pero con más enzimas y sustratos, y sin etiquetar los dNTPs

b) La polimerasa une un dNTP, liberando un PPi en el proceso

c) La ATP sulfurilasa convierte el PPi en ATP con ayuda del APS

d) La luciferasa convierte el ATP en luz, con ayuda de la luciferina

e) Medimos un pulso de luz

f) La apirasa se libra de los reactivos sobrantes para que el sistema esté limpio para el siguiente dNTP

g) El resultado final es, como en Sanger, picos de intensidad

Pirosecuenciación

- Ventajas
  - Muy rápido (hasta 700Mbps por ejecución)
  - Más barato que Sanger
  - Precisión bastante alta
- Desventajas
  - No puede generar secuencias muy largas (<1Kbs)
  - Problemas para medir homopolímeros (secuencias de varios nucleótidos iguales)

Fabricantes de secuenciadores

- Roche
  - Familia de secuenciadores 454
    - Versión paralelizada de la pirosecuenciación (enzima pirasa)
- Illumina
  - Familia de secuenciadores Solexa
    - Utilizan secuenciación de terminación reversible de ciclo
- Applied Biosystems
  - Familia de secuenciadores SOLiD
    - Secuenciación basada en ligación (enzima ligasa)

Secuenciación

Secuenciación de ADN

Secuenciación de Genomas

Fragmentación Shotgun

Ensamblado

Programas y formatos

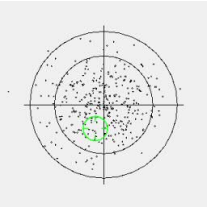
Secuenciación de Transcriptomas

Perspectivas

Secuenciación de genomas

- Las técnicas de secuenciación tradicionales sólo permiten secuenciar cadenas cortas de ADN
  - Necesitamos una estrategia para fragmentar un genoma y luego reensamblar el genoma secuenciado

- Las dos estrategias más usadas son:
  - Whole Genome Shotgun (WGS) sequencing
    - Para genomas pequeños, los fragmentos no se ordenan en jerarquías
  - Hierarchical shotgun sequencing
    - Para genomas más grandes, aunque en el fondo bastante similar
  - Ambas se basan en la técnica de **shotgun** para fragmentar las secuencias → fragmentación aleatoria



el disparo de una escopeta (shotgun) lanza fragmentos en una distribución aleatoria

Secuenciación de genomas

- Terminología
  - **BAC** (Bacterial Artificial Chromosome): constructo de ADN que utiliza las propiedades de los plásmidos bacterianos para replicarse.
  - **Fragmento**: secuencia contigua de ADN, sin huecos
  - **Contig**: conjunto se fragmentos solapantes de los que se puede obtener una secuencia más larga
    - Los contig de NCBI se refieren ya a la secuencia más larga en sí
  - **Scaffold** (andamio): serie de contigs ordenados, pero no necesariamente conectados en una secuencia sin huecos
  - **Run** (ejecución): un lanzamiento del proceso de secuenciación

Cobertura (coverage)

- Número medio de lecturas de cada nucleótido en la secuencia reconstruida
- Se estima una cobertura de 5x a 10x para que se pueda afirmar que el nucleótido está presente
- A partir de la cobertura se puede estimar la probabilidad de presencia del nucleótido
  - Y por tanto el porcentaje del genoma que se puede afirmar que está secuenciado correctamente

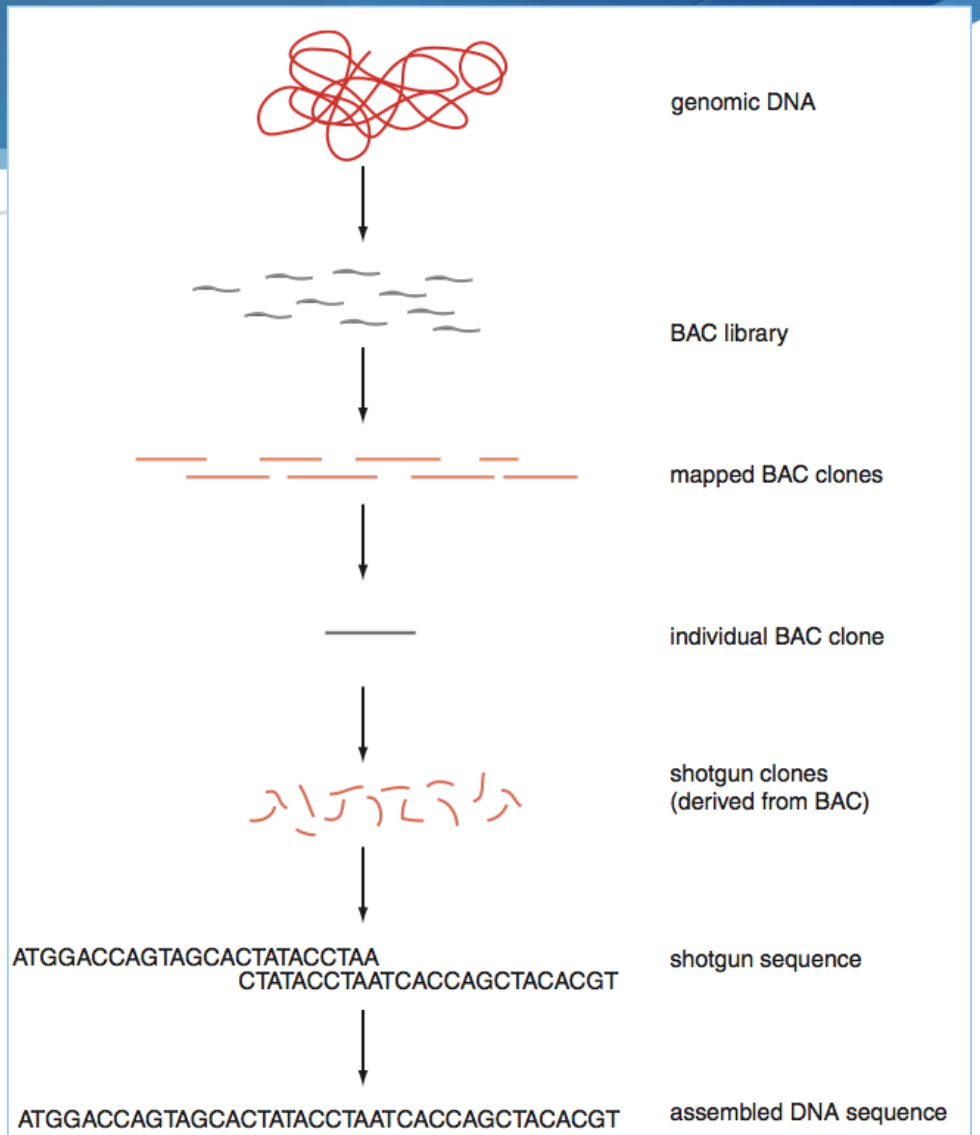
Fold Coverage	P_0	Percent Not Sequenced	Percent Sequenced
0.25	$e^{-0.25} = 0.78$	78	22
0.5	$e^{-0.5} = 0.61$	61	39
0.75	$e^{-0.75} = 0.47$	47	53
1	$e^{-1} = 0.37$	37	63
2	$e^{-2} = 0.135$	13.5	87.5
3	$e^{-3} = 0.05$	5	95
4	$e^{-4} = 0.018$	1.8	98.2
5	$e^{-5} = 0.0067$	0.6	99.4
6	$e^{-6} = 0.0025$	0.25	99.75
7	$e^{-7} = 0.0009$	0.09	99.91
8	$e^{-8} = 0.0003$	0.03	99.97
9	$e^{-9} = 0.0001$	0.01	99.99
10	$e^{-10} = 0.000045$	0.005	99.995

probabilidad de un aminoácido

cualquiera no esté cubierto

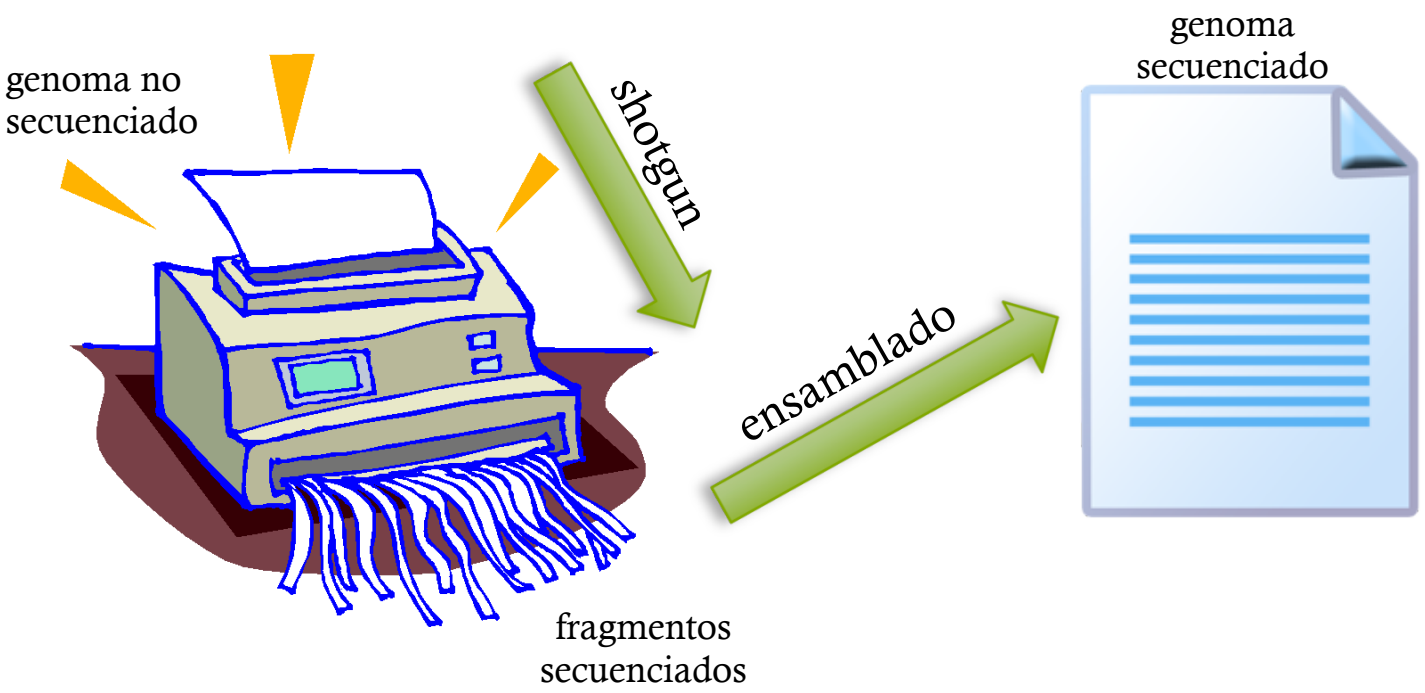
Shotgun sequencing

- El ADN genómico se fragmenta aleatoriamente (shotgun)
- Los fragmentos se clonan en forma de cromosomas artificiales de bacterias (BAC)
- Y se almacenan en bibliotecas según su tamaño (jerarquía)
  - fragmentos grandes (100-500kb)
  - cósmidos (~50kb)
  - plásmidos (~2kb)
- Los clones BAC se secuencian, obteniendo secuencias desordenadas
  - Se ordenarán con el ensamblado



Ensamblado

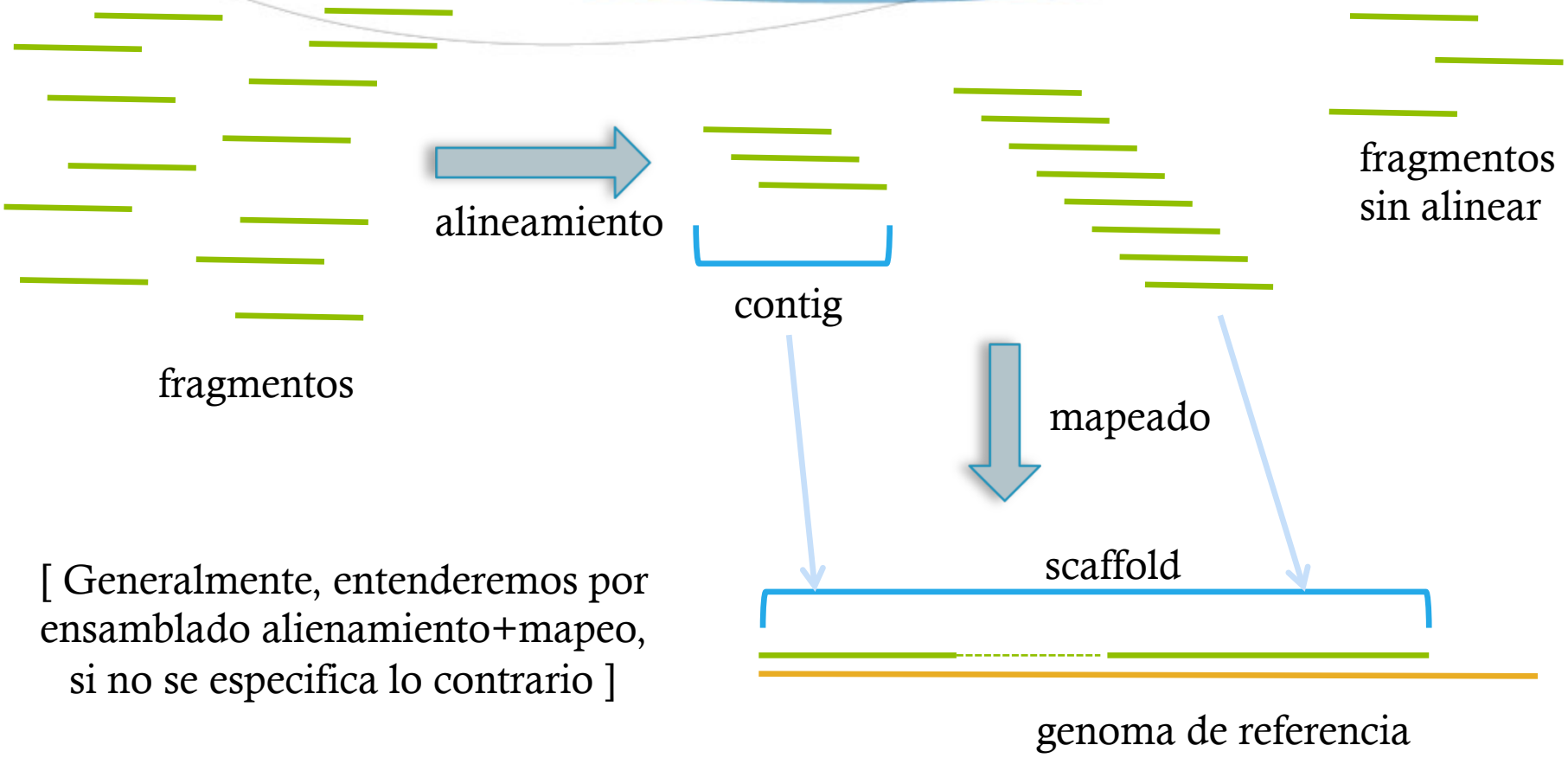
- Alineamiento y fusión de fragmentos de nucleótidos en una secuencia mayor



Tipos de ensamblado

- Según el fragmento
  - Secuenciación de fragmentos (genoma)
  - Secuenciación de ESTs (transcriptoma)
    - Más sencillo al ser más pequeño
- Según la secuencia de referencia
  - De novo: ensamblaje para formar una secuencia nueva
  - Mapeado: ensamblaje para formar una secuencia conocida
    - Nos centraremos en secuenciación y mapeo de genomas
    - Más adelante veremos el ensamblado de transcriptomas

Alineamiento y mapeo



[Generalmente, entenderemos por ensamblado alienamiento+mapeo, si no se especifica lo contrario]

Algoritmos de ensamblado

- Su fundamento son los algoritmos de alineamiento
- Algoritmo “avaricioso”
  1. Calcular alineamientos dos a dos entre todos los fragmentos
  2. Elegir los dos fragmentos con mayor solapamiento
  3. Fusionarlos en un único fragmento
  4. Repetir 2 y 3 hasta que no haya fragmentos solapados
- El proceso da un resultado subóptimo y es computacionalmente muy costoso, aparte de enfrentarse a problemas como las repeticiones.

Complejidad del ensamblado

- Complejidad computacional
  - Tenemos una cantidad enorme de fragmentos muy pequeños a alinear sobre un genoma completo
    - BLAST tardaría días en alinearlos todos
  - Sólo mantener en memoria los fragmentos y el genoma ocupa 12GB
- Un genoma generalmente contiene gran cantidad de secuencias de la longitud de los fragmentos repetidas (*repeats*)
  - Un 20% del genoma humano es repetitivo respecto a fragmentos de 32bp
  - Difícil determinar la posición de los repeats
- La tecnología de secuenciación puede tener errores y confundir al algoritmo de ensamblado

Transformación de Barrows-Wheeler (BWT)

- Técnica para optimizar memoria y detectar alineamientos
  1. Añade un carácter de terminación (p.ej. $ ó @) al final de la secuencia *S*
  2. Calcula todas las rotaciones de *S*, creando una matriz *M*
  3. Ordena alfabéticamente las filas de *M*
  4. Toma como salida la última columna, *S'*

- *M* tiene la propiedad de "mapeo principio-fin (LF)":
  - La *i*-ésima ocurrencia de un carácter en *S'* corresponde a la *i*-ésima ocurrencia de dicho carácter en *S*

Transformation			
Input	All Rotations	Sorted List of Rotations	Output Last Column
^BANANA@	^BANANA@ @^BANANA A@^BANAN NA@^BANA ANA@^BAN NANA@^BA ANANA@^B BANANA@^	ANANA@^B ANA@^BAN A@^BANAN BANANA@^ NANA@^BA NA@^BANA ^BANANA@ @^BANANA	BNN^AA@A

Imágenes de Wikipedia

BWT
Optimización de memoria

- La optimización de memoria viene dada porque *S'*, debido a la ordenación de *M*, va a contener más caracteres seguidos repetidos

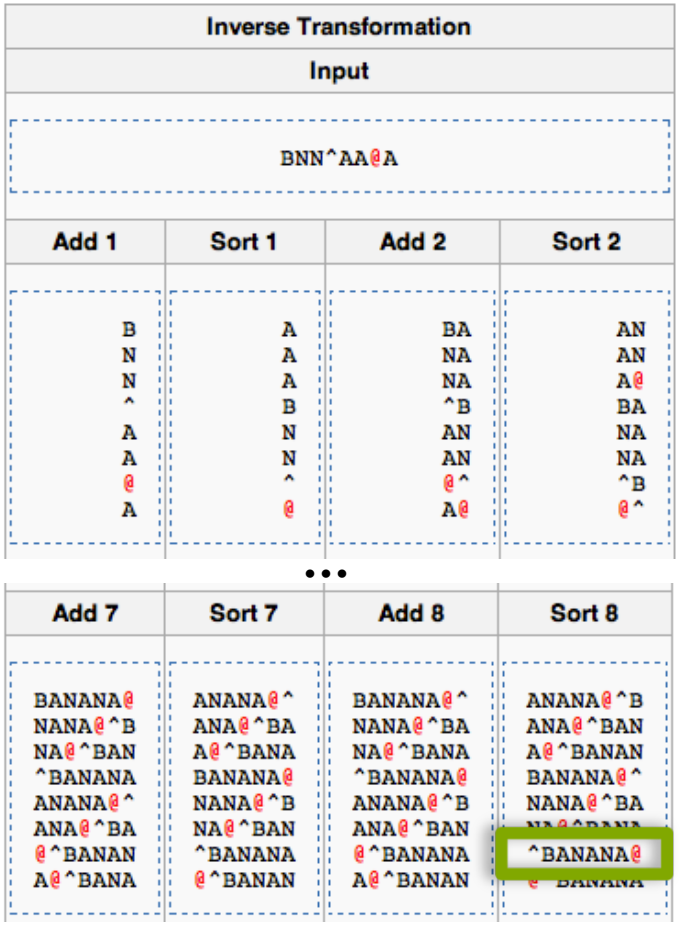
Input	SIX.MIXED.PIXIES.SIFT.SIXTY.PIXIE.DUST.BOXES
Output	TEXYDST.E.IXIXIXXSSMPPS.B..E.S.EUSFXDIIOIIIT

  - Entrada: 0 caracteres seguidos repetidos de 44
  - Salida: 13 caracteres seguidos repetidos de 44

- Tener muchos caracteres seguidos repetidos hace la compresión más sencilla y eficiente
  - Por ejemplo *AAAAAAA* se puede comprimir como *A7*

BWT
Recuperación

- Recuperación de S
  1. Hacemos una matriz M' con una sola columna S'
  2. Ordenamos la matriz alfabéticamente
  3. Añadimos a su final S' como columna
  4. Volvemos al paso 2 hasta que hayamos añadido S' tantas veces como su longitud
  5. La fila que tenga el carácter especial final en la última columna es S



BWT
Búsqueda de fragmentos

- Por la ordenación de *M*, sabemos que las filas que comienzan con la misma secuencia aparecen consecutivamente
- Esto permite determinar rápidamente si una secuencia contiene un fragmento
  - Lo podemos utilizar como técnica de alineamiento!

Programas de ensamblado

- Maq y SOAP:
  - No usan BWT si no hashing (más lento y ocupa más)
  - Permiten alineamientos con huecos
- Burrows-Wheeler Aligner (BWA)
  - Utiliza la transformación de BWT para optimizar recursos de memoria y alineamiento
- **Bowtie**
  - Usa BWT con una búsqueda mejorada para permitir mismatches
  - Muy rápido y bastante preciso
    - Alinea fragmentos cortos de ADN respecto al genoma humano a un ritmo de 25 millones de fragmentos de 35bps por hora
  - No permite alineamientos con huecos

Programas de ensamblado *de novo*

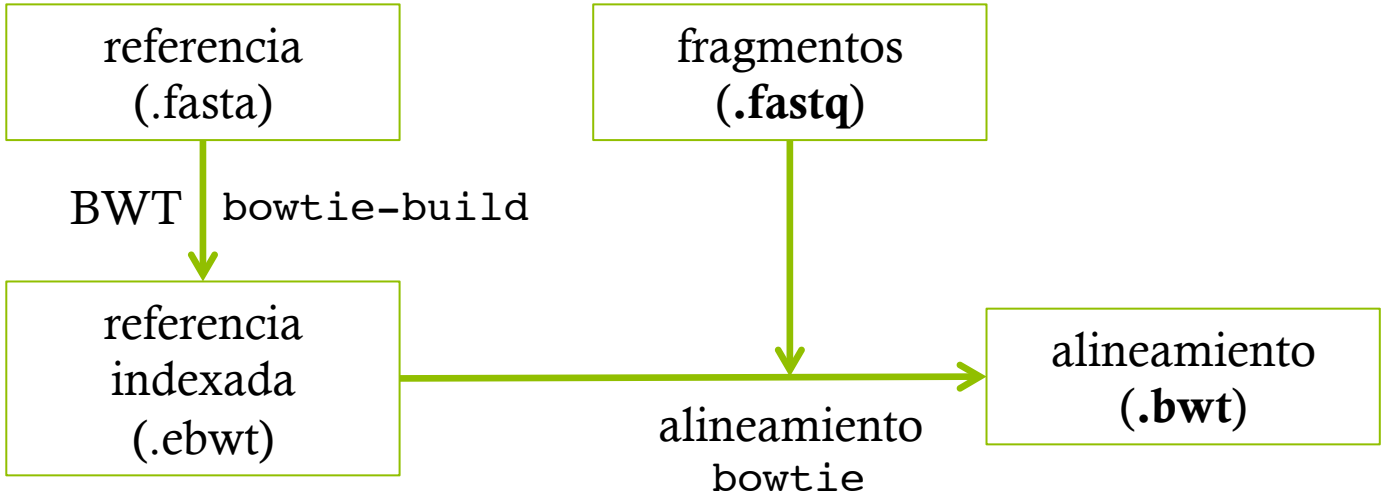
- Velvet
  - Zerbino et al.
- **ABySS**: ensamblador *de novo* en paralelo
  - Simpson et al.
  - Utiliza bowtie para el alineamiento
  - Utiliza grafos de De Brujin para la creación del scaffold
  - Herramienta para visualización del ensamblado: ABySS-Explorer
  - Versión para ensamblado de transcriptomas: Trans-ABySS

Bowtie

- Algoritmo de ensamblado basado en BWT
- Mejora la búsqueda de secuencias para permitir errores en la comparación, de modo que se puedan encontrar secuencias iguales salvo algunos nucleótidos distintos (mismatches)
- Divide por 10 los tiempos de ejecución de Maq, y por 100 los de SOAP, sin una degradación importante del ensamblado

workflow de ensamblado con Bowtie

- **Workflow** (o pipeline): secuencia de instrucciones/uso de programas para obtener un resultado



Fastq

- Fichero típico con las lecturas de salida del secuenciador
  - no ensambladas todavía
- Fastq = Fasta con calidades
  - 4 líneas por secuencia
    - @identificador
    - secuencia
    - +[identificador]
    - calidades (misma longitud que la secuencia)

Solexa

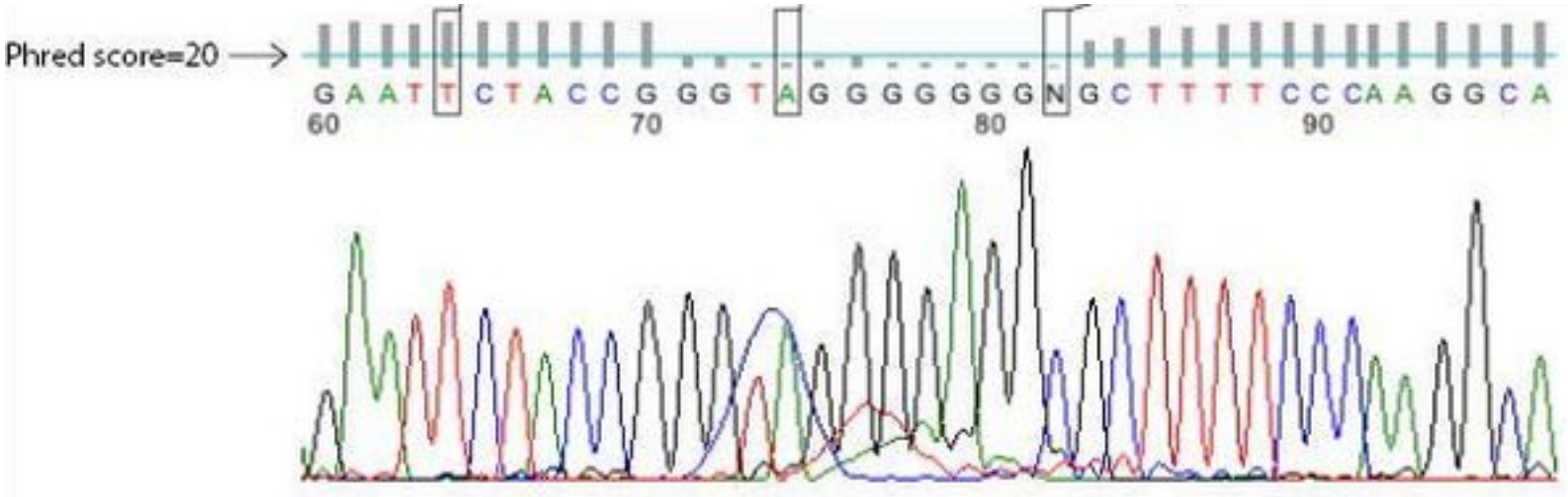
```
@HWI-EAS225:3:1:2:854#0/1
GGGGGGAAGTCGGCAAAATAGATCCGTAACTTCGGG
+HWI-EAS225:3:1:2:854#0/1
a`abbbbabaabbbabab^`[aaa`_N]b^ab^``a
```

NCBI

```
@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
```

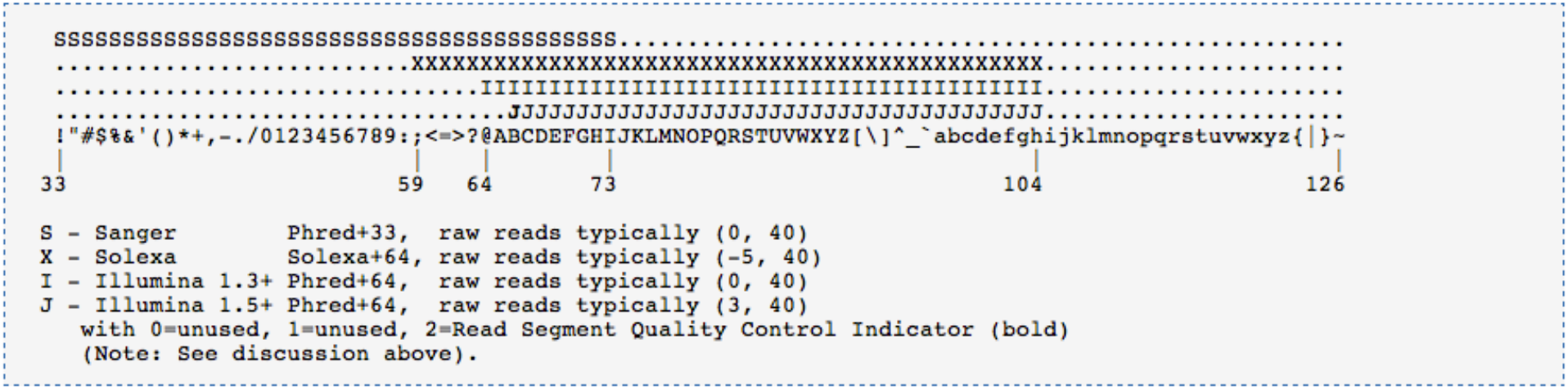
Fastq

- La calidad es una valor determinado por la probabilidad de que la base sea incorrecta (p)
  - Se calcula con el coeficiente de calidad de Phred: $Q = -10\log_{10}(p)$
    - Phred fue el primer programa de apoyo a la secuenciación, utilizado durante el Human Genome Project

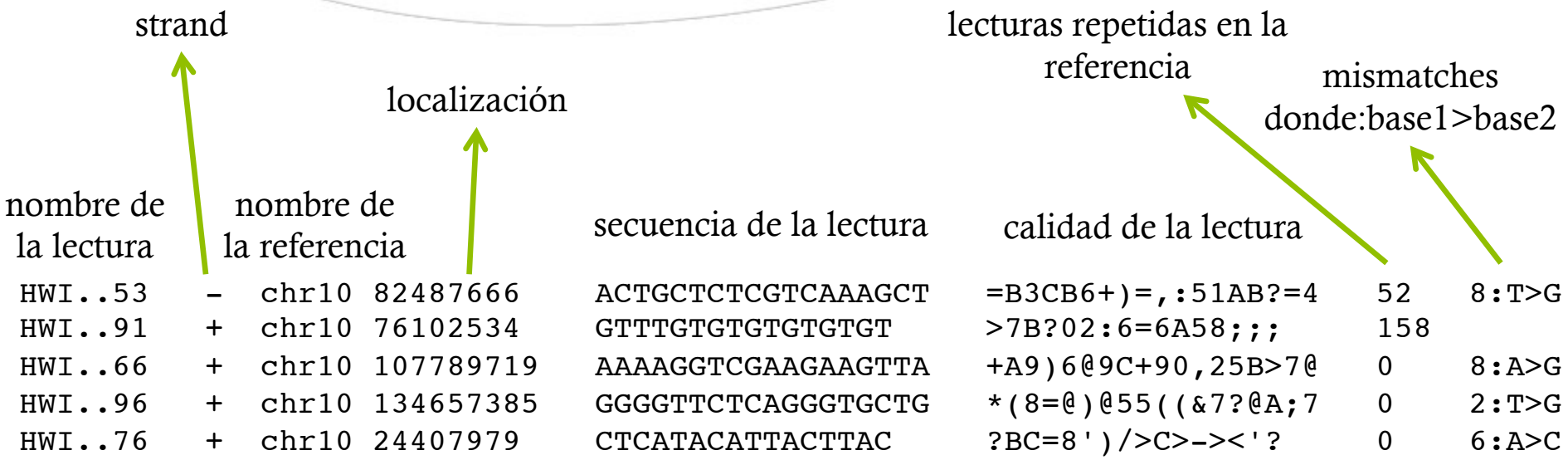


Fastq

- La calidad se escala a un cierto número de rangos
  - Se codifica con los caracteres ascii 33 a 73 (en Sanger)
  - En otras plataformas, se usan otras escalas y caracteres



Formato .bwt



- Simplificación del formato SAM (Sequence Alignment/Map)
  - Formato estándar para alineamientos
  - BAM: formato SAM en binario (comprimido)
  - Bowtie permite transformar .bwt a .sam/.bam

Consideraciones

- El ensamblado de secuencias es un campo muy nuevo, y de una carga computacional muy alta
- Esto limita las herramientas disponibles
  - Suelen ser herramientas de línea de comandos
    - No hay muchos programas gráficos o herramientas on-line
  - En muchos casos su uso en ordenadores de sobremesa es difícil, tanto en tiempo como en espacio
    - Ejecución en servidores o en paralelo (múltiples servidores)

Análisis post-alineamiento

- Una vez tenemos el alineamiento, para poder extraer conclusiones necesitamos métodos de manipulación e inspección
  - **SAMtools**: programa de línea de comandos para manipulación de alineamientos (ordenación, fusionado, filtrado, indexado…)
  - El paquete de R/BioConductor **shortReads** permite manipulaciones similares a SAMtools
    - El paquete **rtracklayer** permite conversiones a formatos aceptados por los principales navegadores de Genomas (UCSC, IGB)
- En el caso de secuenciación de transcriptomas, tendremos que estimar el nº de transcritos (algoritmos de conteo)

Secuenciación

Secuenciación de ADN

Secuenciación de Genomas

Secuenciación de Transcriptomas

RNA-Seq

Algoritmos

Perspectivas

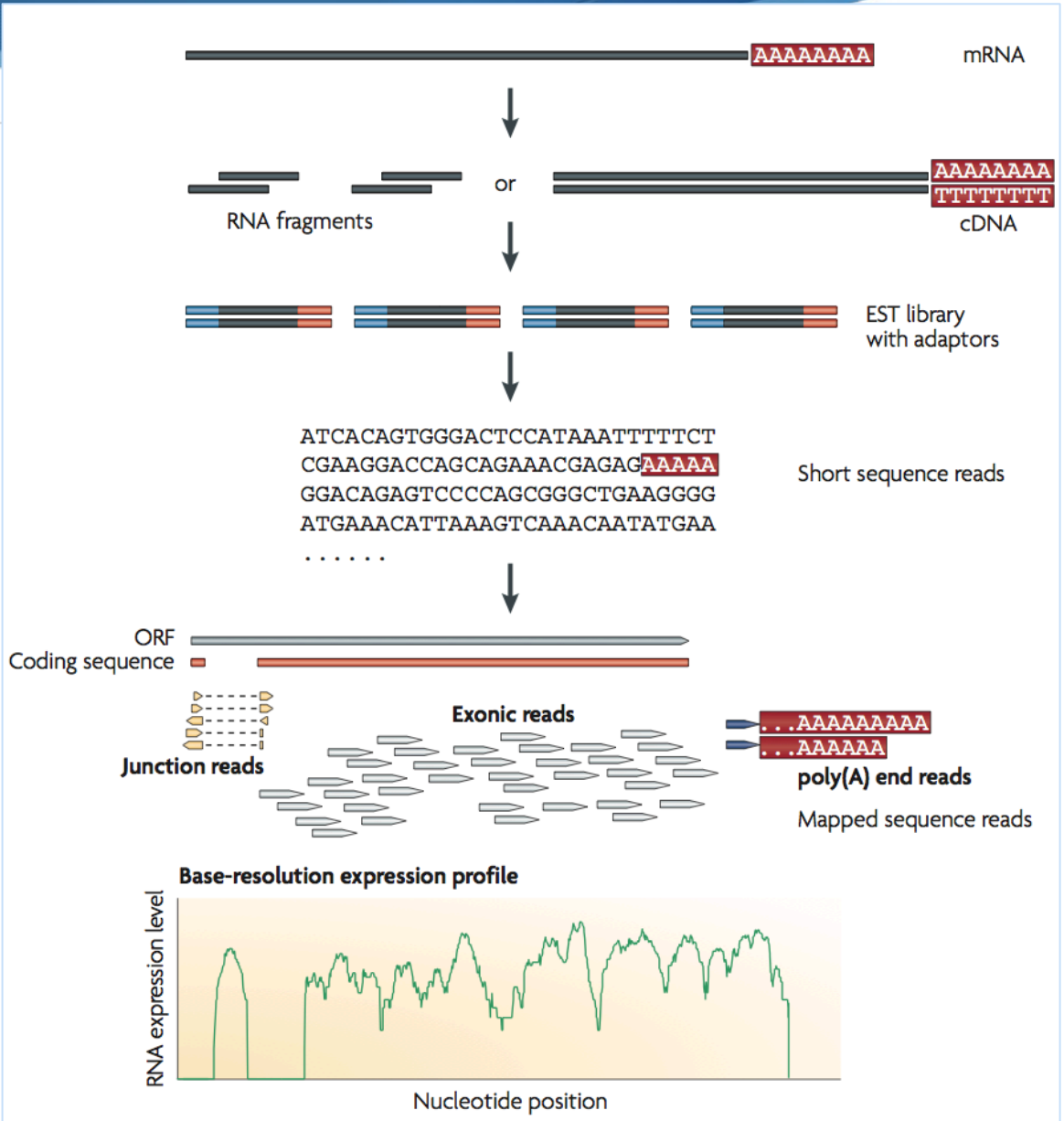
Secuenciación de transcriptomas

- Es la fusión natural de la filosofía de análisis de expresión con las técnicas de secuenciación de genomas
- Más sencilla que la secuenciación de genomas completos
  - El transcriptoma es más pequeño (< 2% del genoma)
- En vez de secuenciar una cadena de ADN secuenciamos el mRNA de una muestra
  - Si un gen se expresa mucho, habrá muchas copias de su mRNA

RNA-seq

- Toma el mRNA de una muestra y lo secuencia, obteniendo niveles de transcripción para cada secuencia corta de mRNA
  - No necesita hibridar con una plantilla como en microarray
    - No hay límite al número de copias
  - Permite obtener información de la secuencia durante la transcripción (detección de exones, SNPs, etc.)
  - No necesita conocer el genoma del organismo
  - Tiene un ruido de fondo muy bajo

- Partimos de un fragmento de ARN con una cola poli-A, o de su cDNA complementario (cola poli-T)
  - Suelen ser secuencias grandes, que se fragmentan en secuencias de 200-500 bps para secuenciarlas
- Los fragmentos se almacenan en una biblioteca de Expressed Sequence Tags (ESTs), con adaptadores (para activar la secuenciación)
- Se realiza la secuenciación y las lecturas se alinean con los ORFs conocidos:
  - Lecturas exónicas
  - Lecturas de unión ~interexónicas
  - Lecturas poli-A
- Se cuentan las lecturas para cada base → nivel de expresión



Z. Wang et al. *RNA-Seq: a revolutionary tool for transcriptomics*. Nature Perspectives, 2009

Análisis

- RNA-seq tiene un par de características especiales que requieren de análisis adicionales al de microarray
  - **Lecturas divididas** (splice o junction reads): lecturas que capturan RNA correspondiente a dos exones
    - Difíciles de alinear al genoma completo, sobre todo si usamos un alineamiento que no permite huecos
  - **Conteo:** necesidad de contar la cantidad de bases en cada posición (~cobertura) para determinar los niveles de expresión
    - Una vez obtenidos, tenemos una matriz de expresión analizable por los métodos tradicionales vistos para microarrays

Algoritmos para RNA-seq

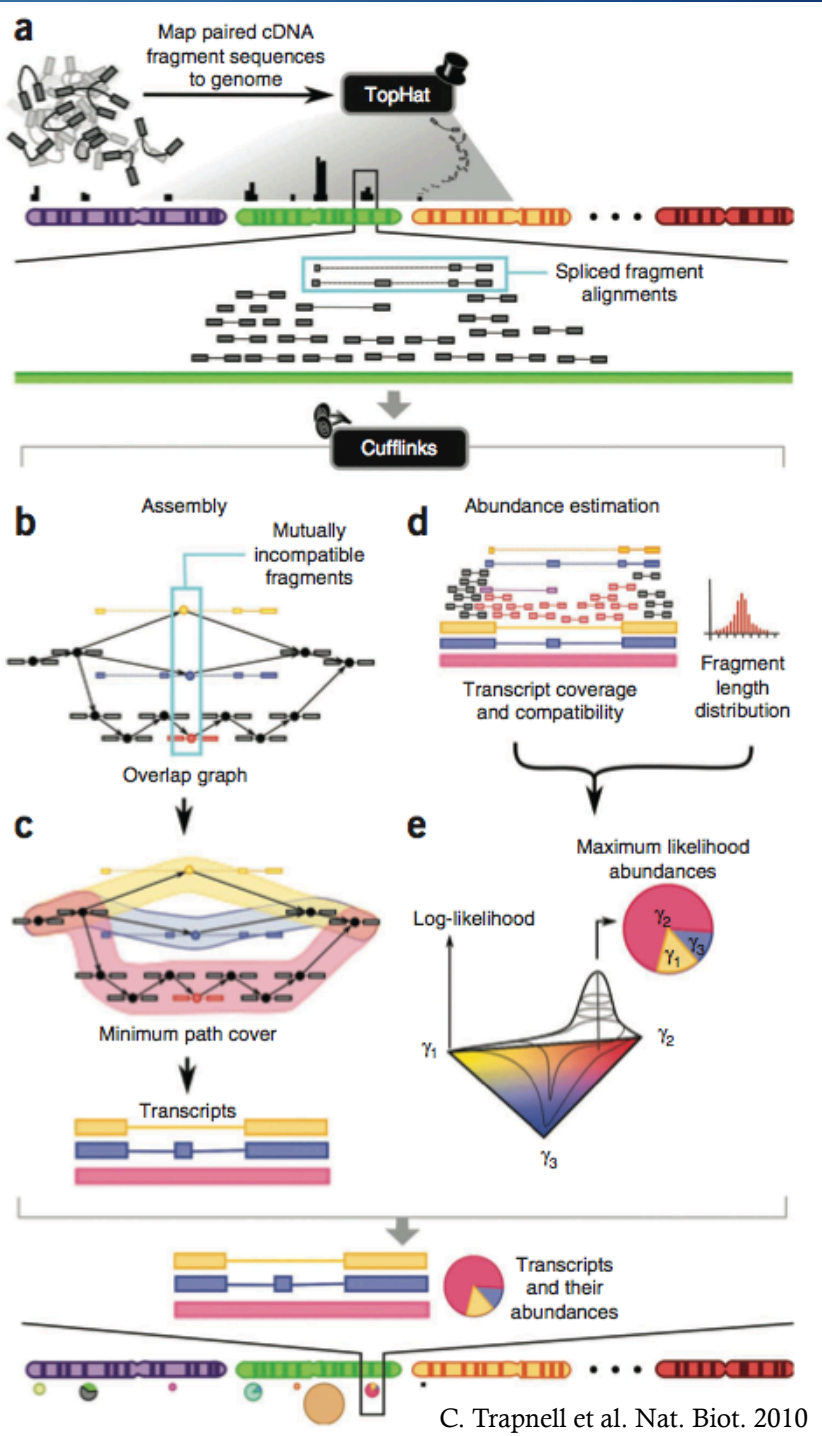
- Algoritmos de alineamiento: TopHat
- Algoritmos de conteo: Cufflinks y mmseq
- Análisis de expresión
  - Una vez tenemos los niveles de expresión, podemos utilizar cualquiera de los métodos vistos en el tema de microarrays
  - Pero tendremos más detalles
    - Secuencias reales de nuestra muestra, no sondas elegidas a priori por un fabricante
    - Splice variants: expresión de los distintos transcritos posibles

TopHat

- Programa de alineamiento que mapea transcritos a un genoma de referencia
- Primero realiza un alineamiento bowtie
  - Pero en RNA-seq algunas lecturas pueden caer en un intrón (juntion o splice read) → botwite no las alinea (gaps)
- A continuación intenta alinear las secuencias no alineadas por bowtie, que podrían ser splice reads, mediante un algoritmo más preciso que permita huecos (tipo maq)

Cufflinks

- Parte de TopHat, un alineamiento que contempla splice reads **(a)**
- Debe tratar con dos problemas: repeticiones en el genoma y splice reads
- **Splice reads**: utiliza teoría de grafos para detectar isoformas del gen **(b, c)**
  - Toma los splice reads y determina posibles transcritos según sus exones
- **Repeticiones**: estima qué fragmentos corresponden a cada transcrito y a partir de ahí su nivel de expresión **(d, e)**
  - Distribución uniforme
  - Corrección probabilística



C. Trapnell et al. Nat. Biot. 2010

ArrayExpressHTS

- El análisis de lecturas de transcriptomas es complicado
  - Múltiples herramientas a interconectar
  - Accesibles sólo desde línea de comandos
  - Muchas veces necesario aplicarlas desde servidores
- ArrayExpressHTS es un paquete para R/BioConductor que
  - Automatiza todo el workflow en una única herramienta
    - Permitiendo todas las opciones de las herramientas incluidas
    - Incluye la descarga de experimentos RNA-Seq de ArrayExpress
  - Gestiona el uso de servidores a través del servicio EBI Workcloud

ArrayExpressHTS

```
library(ArrayExpressHTS)
e <- ArrayExpressHTS("E-GEOD-16190")
```

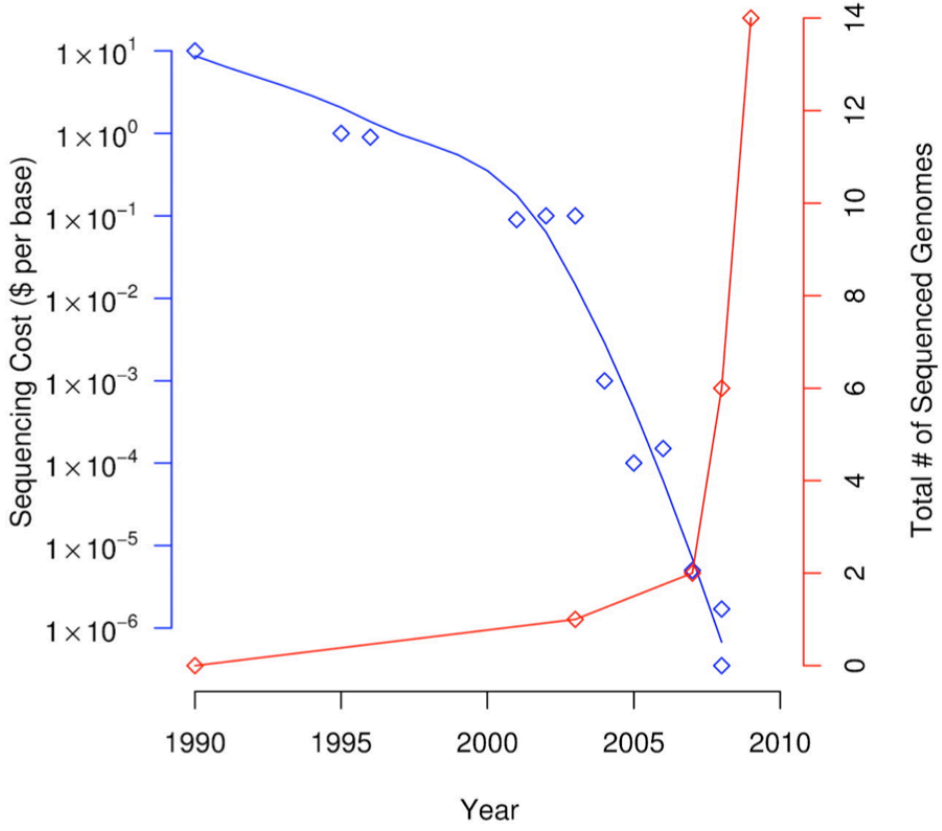
- Descarga los datos del experimento y ejecuta el análisis de secuencias realizando el alineamiento (TopHat) y conteo (Cufflinks)
- `e` contiene la matriz con los niveles de expresión
  - Se crea además una estructura de directorios con los datos crudos, informes de calidad, alineamientos y matriz de expresión
- Las opciones por defecto se pueden cambiar (parámetros y algoritmos)
  - http://www.ebi.ac.uk/Tools/rwiki/Wiki.jsp?page=ArrayExpressHTS%20Main%20Page

Secuenciación

Secuenciación de ADN
Secuenciación de Genomas
Secuenciación de Transcriptomas
Perspectivas

Perspectivas

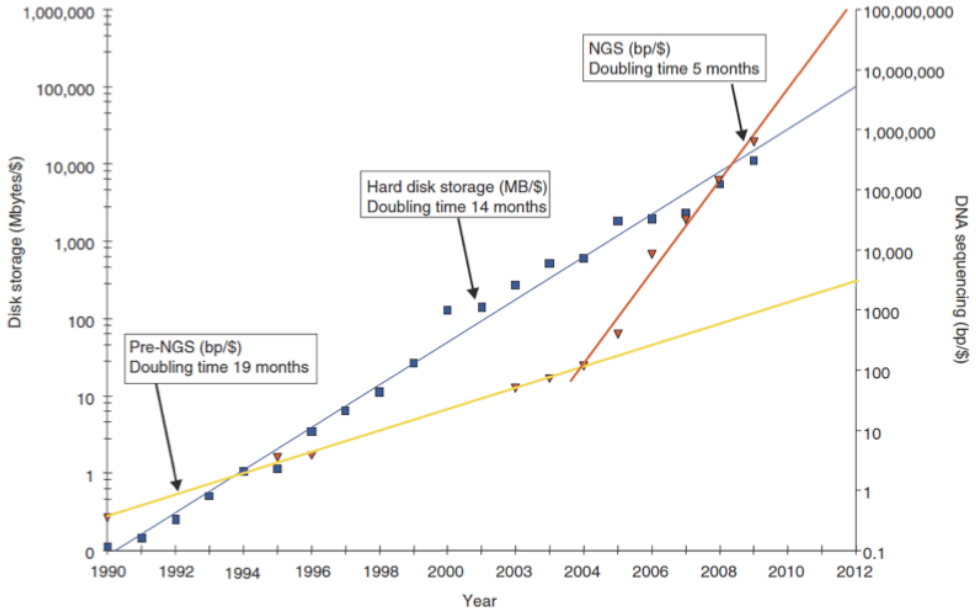
- Los costes de secuenciación bajan
  - Genoma humano ~$3 \cdot 10^{12}$ bp
    - Considerando un coste por base de 10^{-6} $ el coste total sería ~10^{6} €
  - Exoma humano ~1.7%
    - Coste ~20.000 €
- El número de genomas secuenciados aumenta exponencialmente



Snyder M et al. Genes Dev. 2010;24:423-431

Perspectivas

- Genoma humano ~725MB
  - Los genomas varían menos de un 1%
    - Se pueden comprimir a 4MB si tenemos uno de referencia
- Estamos superando el límite en el que nos cuesta menos secuenciar que almacenar



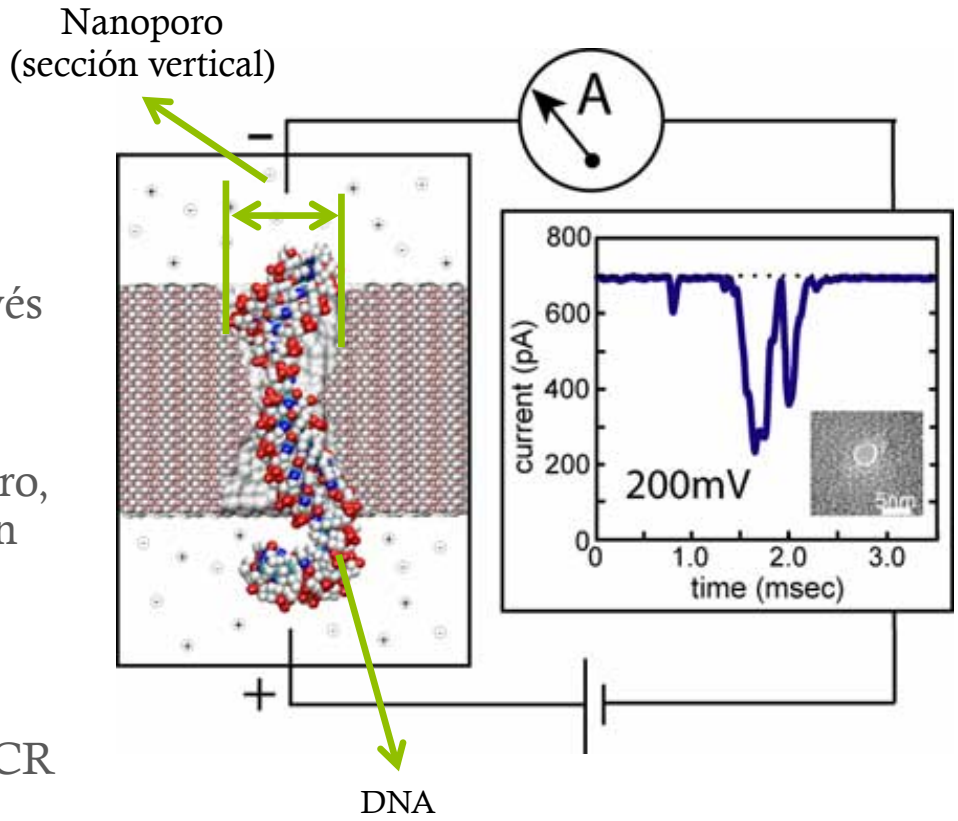
Stein, L. - Genome Biology (2010)
The case for cloud computing in genome informatics

La carrera por el genoma a 1000 $

- En 2003, la Fundación para la Ciencia Craig Venter ofreció un premio de 500.000 $ para el que consiguiera secuenciar un genoma por menos de 1000 $
- La Fundación X Prize ofrece 10 millones de dólares por secuenciar 100 genomas en 10 días, a un coste menor de 10.000$ por genoma
- El primer genoma (2003) costó $2.7 \cdot 10^9$ $
- A junio de 2011, su precio está entorno a 15.000 $

Técnicas de secuenciación de tercera generación

- **Nanoporos**
  - Agujeros con un diámetro de 10^{-9} m
  - Cuando se sumergen en un fluido conductor y se aplica un voltaje, se observa una pequeña corriente a través del nanoporo
    - sensible a su tamaño y forma!
  - Si se hace pasar ADN por el nanoporo, los cambios de corriente identificarán cada base
    - No hace falta un etiquetado químico de las bases
    - No hace falta un aumento por PCR



http://www.ks.uiuc.edu/Research/nanopore/

Técnicas de secuenciación de tercera generación

- Otras técnicas
  - Secuenciación con microfluidos
    - http://en.wikipedia.org/wiki/Microfluidic_Sanger_sequencing
  - Microscopía de electrones
    - http://en.wikipedia.org/wiki/Transmission_electron_microscopy_DNA_sequencing
  - Etc.
- En general tratan de reducir costes eliminando fases y reactivos en el análisis
- Todas estas técnicas están todavía en desarrollo

Table 1. *Individual genomes that have been sequenced and published*

Project	Technology	Paired end	SNPs; short Indel	SVs	New sequence	Fully phased genotyping	Reference
Reference	Sanger	No	NA	NA	NA	NA	Lander et al. 2001; Collins et al. 2003
European-Venter	Sanger	Yes	3 million; 0.3 million	0.2 million (>1000 bp)	1 M	Limited	Levy et al. 2007
European-Watson	454	No	3 million; 0.2 million	Limited	No	No	Wheeler et al. 2008
European-Quake	Helicos	No	3 million	Limited	No	No	Pushkarev et al. 2009
Asian	Illumina	Partially	3 million; 0.1 million	2700 (>100 bp)	No	No	Wang et al. 2008
HapMap sample; Yoruban 18507	Illumina	Yes	4 million; 10,000	100	No	No	Bentley et al. 2008
HapMap sample; Yoruban 18507	SOLiD	Partially	4 million; 0.2 million	5500 (unknown definition)	No	No	McKernan et al. 2009
Korean	Illumina	Yes	3 million	Limited	No	No	Ahn et al. 2009
Korean-AK1	Illumina	Yes	3.45 million; 0.17 million	~300 CNVs	No	No	Kim et al. 2009
Three human genomes	Complete genomics	Yes	3.2–4.5 million; 0.3–0.5 million	Limited (50,000–90,000 block substitutions)	No	Limited	Drmanac et al. 2009
AML genome and normal counterpart	Illumina	No	3.8 million; 700	Limited	No	No	Ley et al. 2008
AML genome	Illumina	Yes	64	Limited	No	No	Mardis et al. 2009
Melanoma genome	Illumina	Yes	32,000; 1000	51	No	No	Pleasance et al. 2010a
Lung cancer genome	SOLiD	Yes	23,000; 65	392	No	No	Pleasance et al. 2010b

Fifteen genomes have been sequenced from 13 individuals in addition to the original reference sequence. The HapMap cell line NA18507 has been sequenced independently three times. For the purposes of this tabulation, genomes deduced from both normal and disease are counted as one sequence. (NA) Not applicable.

Snyder M et al. Genes Dev. 2010;24:423-431

SCIENTIFIC AMERICAN

Features | Health

Ozzy Osbourne's Genome Reveals Some Neandertal Lineage

What genetic oddities does rock's Prince of Darkness and beheader of bats have entangled deep in his genetic code? Knome, the company that analyzed Ozzy's full genome, divulges some of the details in a Q&A

By Katherine Harmon | October 26, 2010 | 23

The one-time front man for heavy metal band Black Sabbath has joined the likes of DNA co-discoverer James Watson and Harvard University professor Henry Louis Gates on the short roster of people to have their full genome sequenced and analyzed.

HEAVY MENDEL: Scientists have thousands of interesting new mutations uncovered in Ozzy Osbourne's genome to puzzle over.
Image: WIKIMEDIA COMMONS/KAISERJNR

Ozzy Osbourne let a little blood to submit to the testing in July. Cofactor Genomics, a Saint Louis–based company, sequenced Osbourne's genome; Knome, Inc., which also helped raise money for the project, analyzed the data.

For his part, Osbourne was at first skeptical about the project, he explained in his October 24 *Sunday Times of London* column. But the platinum-record artist then began to wonder if he, in fact, might have something to offer science.

"Given the swimming pools of booze I've guzzled over the years—not to mention all of the cocaine, morphine, sleeping pills, cough syrup, LSD, Rohypnol…you name it—there's really no plausible medical reason why I should still be alive. Maybe my DNA could say why."

Ledford, H. Nature (2010)
Big science: The cancer genome challenge

http://www.1000genomes.org/

PROJECT DESIGN

Pilot Project

Three pilot studies provided data to inform the design of the full-scale project:

Pilot	Purpose	Coverage	Strategy	Status
1 - low coverage	Assess strategy of sharing data across samples	2-4X	Whole-genome sequencing of 180 samples	Sequencing completed October 2008
2 - trios	Assess coverage and platforms and centers	20-60X	Whole-genome sequencing of 2 mother-father-adult child trios	Sequencing completed October 2008
3 - gene regions	Assess methods for gene-region-capture	50X	1000 gene regions in 900 samples	Sequencing completed June 2009

Public releases of summary processed data and variant calls are available through the data tab of this website. The data from the pilot projects have been analyzed, especially to determine whether the strategy of 4X coverage is adequate to meet the goals of the Project. A paper describing the analyses of the pilot data and design of the full project was published in October in the journal *Nature* and is available here.

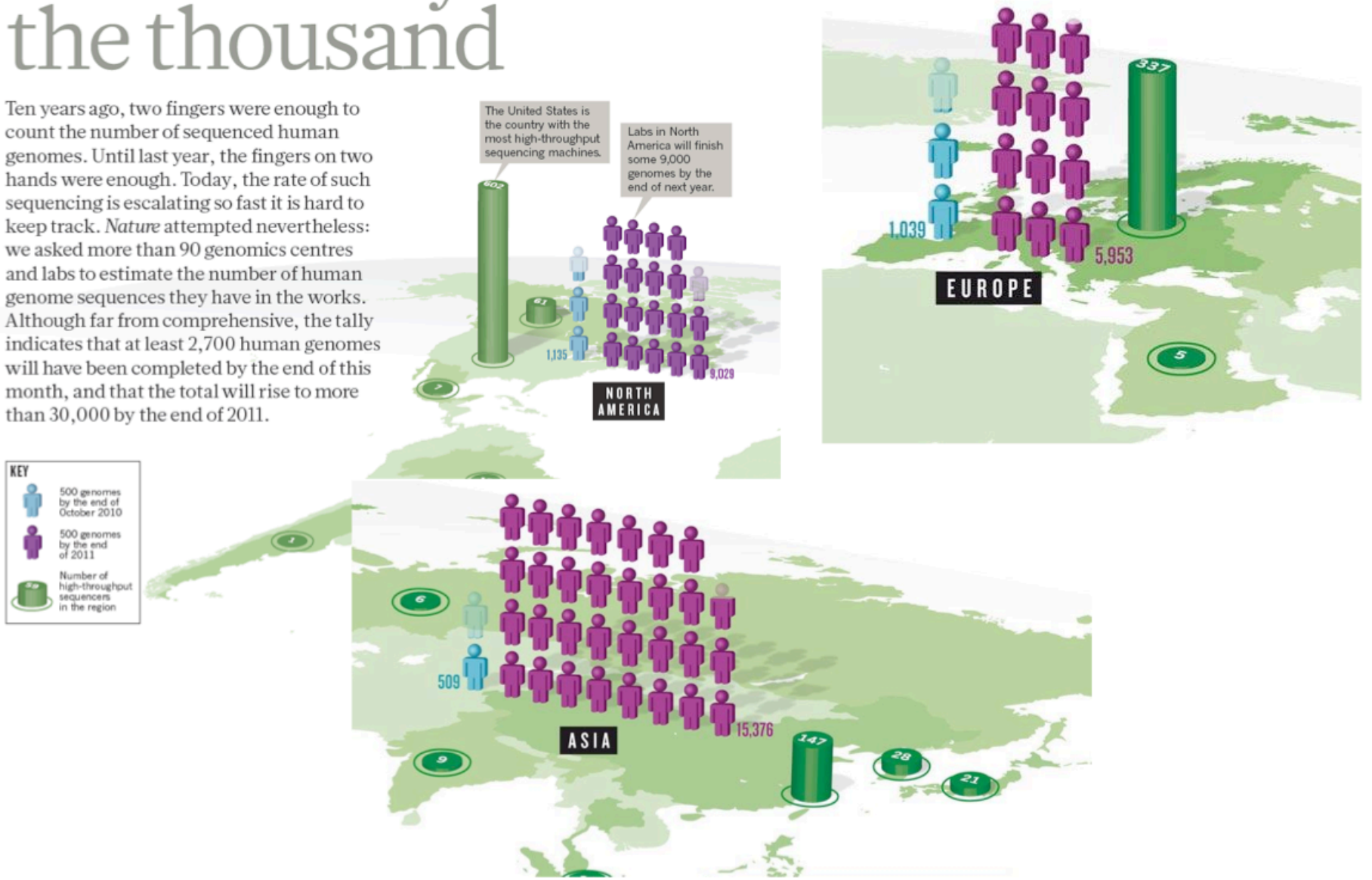
A summary of sequencing done for each of the three pilot projects is available here. And the list of samples and allocations is provided in a spreadsheet. Regions targeted in Pilot 3 are available on the project FTP site.

Main Project

The plan for the full project is to sequence about 2,500 samples at 4X coverage. The first set of samples for sequencing includes 1167 samples that already existed or could be collected quickly, from 13 populations, for sequencing in 2010 and early 2011. The second set includes 633 samples that are being collected, from 7 populations, for sequencing in early 2011. The third set, consisting of 700 samples, is expected to be available for sequencing in late 2011. Full details of the samples to be sequenced are below.

Genomes by the thousand

Ten years ago, two fingers were enough to count the number of sequenced human genomes. Until last year, the fingers on two hands were enough. Today, the rate of such sequencing is escalating so fast it is hard to keep track. *Nature* attempted nevertheless: we asked more than 90 genomics centres and labs to estimate the number of human genome sequences they have in the works. Although far from comprehensive, the tally indicates that at least 2,700 human genomes will have been completed by the end of this month, and that the total will rise to more than 30,000 by the end of 2011.



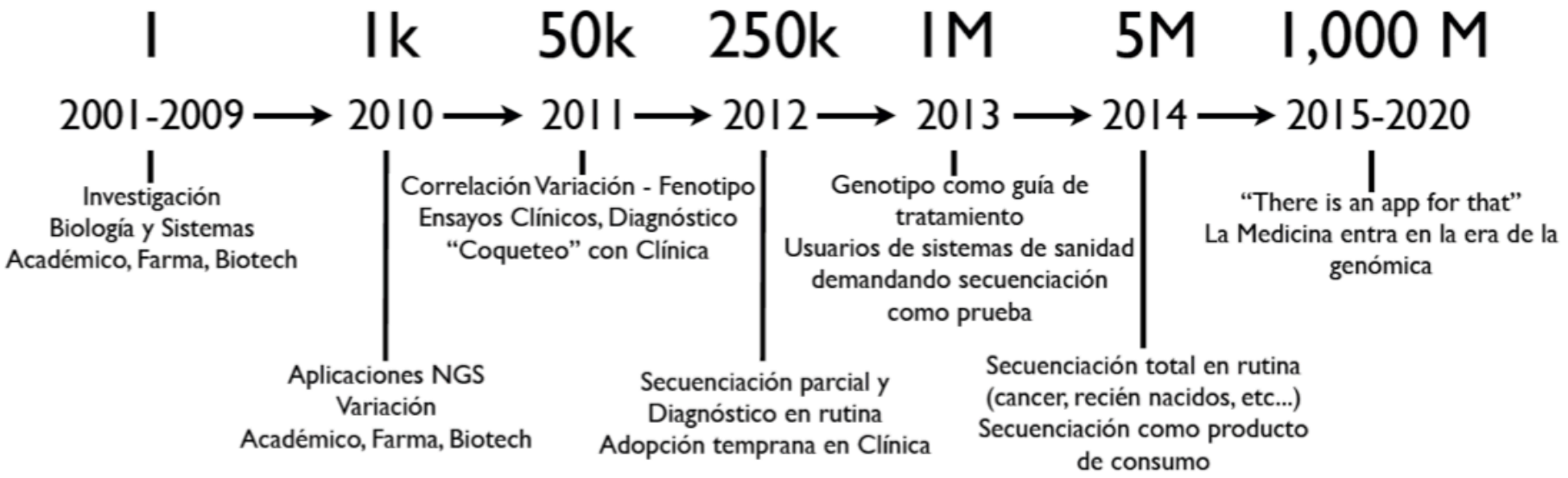
October 2010 | Nature 467, 1026-1027

GenomeQuest | Blog

Perspectives and Insights from GenomeQuest Executives

http://blog.genomequest.com/

Implications of exponential growth of global whole genome sequencing capacity



"sequencing capacity world-wide will continue to grow exponentially for at least the next 10 years"

Resumen

- La tecnología de microarrays mide el nivel de expresión (transcripción), resultando en una matriz de expresión de genes (filas) bajo distintas condiciones (columnas)
- El análisis de expresión génica tiene como objetivo determinar qué genes se encuentran diferencialmente expresados entre dos condiciones (estadística inferencial), y qué grupos de genes/condiciones tienen un patrón de expresión similar (estadística descriptiva)
- Existen una gran cantidad de métodos de análisis. Casi todos devuelven una salida. Lo difícil es confirmar que esa salida es válida desde un punto de vista biológico y estructural
- Muchos análisis caen en errores a la hora de identificar grupos, siendo los más corrientes 1) no hacer correcciones para contrastes de múltiples hipótesis, 2) no hacer una normalización adecuada, ni chequeos de la calidad de los arrays, 3) exceso de libertad paramétrica en el análisis de los datos y 4) exceso de limitaciones biológicas en el análisis de los datos
- Las correlaciones entre genes a nivel de expresión que queramos concluir como causales deben acompañarse de experimentos de laboratorio que aseguren que la relación a nivel transcriptómico se mantiene a niveles superiores (qRT-PCR, chIP-on-chip, etc.)

Cuestiones a debate

- Si tuvieras la oportunidad, ¿secuenciarías tu genoma? ¿qué implicaciones crees que tendría?
- Se estima que aproximadamente el 97% del genoma es "basura". O mejor dicho, que no se sabe cuál es su función ¿crees que es relevante secuenciar genomas completos? ¿cuáles son tus impresiones sobre las funciones que puede tener ese genoma "basura"?

Lecturas adicionales

- Pevsner, 2009: Ch 13 *Completed Genomes*
- Bowtie
  - Langmead et al. *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome*. Genome Biology 2009.
- Sobre tecnologías de secuenciación
  - Michael L. Metzker. *Sequencing technologies – the next generation.* Nature Reviews. 2010
- Sobre tecnología RNA-Seq
  - Wang et al. *RNA-Seq: a revolutionary tool for transcriptomics.* Nature Reviews. 2009

Ejemplo: Bowtie

- Partimos del genoma de referencia para el cromosoma 10 de humano
  - http://www.lcg.unam.mx/~compu2/cei/data/chr10.fa.gz
- Y una de estas lecturas para cuatro muestras
  - http://www.lcg.unam.mx/~compu2/cei/data/fastq_clean/
- Lo primero que hay que hacer es transformar el genoma de referencia para optimizar el mapeado
  - `bowtie-build [opciones] referencia refTransformada`
  - La referencia por defecto está en formato fasta
  - El nombre de refTransformada sin extensión

Ejemplo: Bowtie

- El resultado de la transformación son varios ficheros con extensión .ebwt
  - Algunos tienen también .rev., son los ficheros necesarios para hacer la transformación a la inversa y recuperar la secuencia original
- El segundo paso es realizar el alineamiento/mapeado
  - `bowtie [opciones] refTransformada {lectura1,…, lecturaN} [salida.bwt]`
  - Opciones
    - –q para lecturas .fastq, -f para lecturas .fasta, --sam para salida en .sam
    - -k n para devolver las n mejores coincidencias de cada lectura
    - -n m para permitir hasta m mismatches (-v m si lo hacemos ignorando calidad)
  - IMPORTANTE: controlar los parámetros es fundamental para evitar ejecuciones muy pesadas. Mantened bajo el nº de mismatches permitidas y el nº de coincidencias a devolver

Ejemplo: Bowtie