

**Dra Guadalupe del  
Rosario Quispe Saji**

Posdoc

Universidad Federal de Espiritu  
Santo

## **Ensamblaje de Genomas:**

**Tecnologías NGS  
Proceso de ensamblaje**

Tecnologías NGS

Tipos de archivos FASTQ, SAM/BAM

Proceso de Ensamblaje de genomas



# Secuenciamiento del DNA

Las tecnologías NGS (NextGeneration sequencing) vienen revolucionando la biología, la utilidad se extiende a los dominios de DNA, RNA y proteínas:

1. Permite determinar la secuencia de DNA de los genomas en todo el árbol de la vida
2. El resecuenciamiento de genomas.
3. Comparar 1 o más genomas a un genoma de referencia: genómica comparativa
4. Comparar diferencias genéticas dentro de un individuo en diferentes células
5. NGS aplicado al RNA (RNAseq) para medir niveles de transcripción en diferentes condiciones
6. Aplicar NGS a muestras ambientales (metagenómica)

# Secuenciamiento del DNA

El secuenciamiento permite la identificación del orden exacto de los pares de bases (A, T, G y C) en un segmento de DNA.

Por exemplo para el caso del Projeto Genoma Humano, foi determinado el orden de 3 billones de pares de bases (bp) que constituyen el DNA de los 24 cromossomos (~3,2 Gb).

Conocer la secuencia de las bases de un gene brinda importantes informaciones sobre sua estructura, función y relación evolutiva con otros genes (del mismo organismo o de organismos diferentes)



# Secuenciamiento de Sanger

Em 1970 Ray Wu desarrolla una estrategia de extension para secuencias de nucleotideos de DNA. Wu determinou dos terminaciones cohesas del DNA del fago lambda (1971)

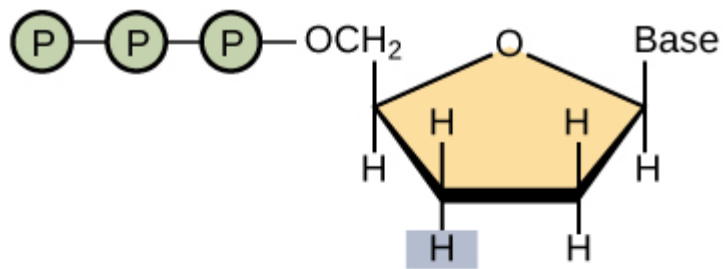
Esta tecnica vienen la ser la base usada en el secuenciamiento de sanger

En 1977 Sanger y colegas introdujeron la tecnica conocida como secuenciamiento de sanger o secuenciamiento de dideoxinucleotideos

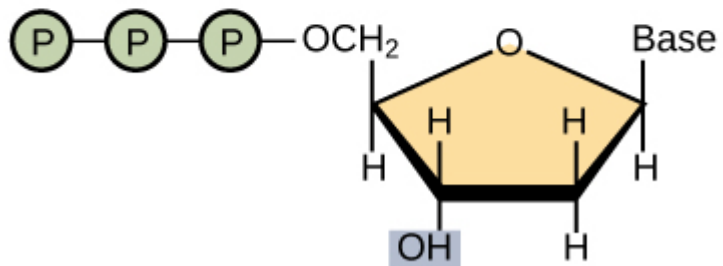
En el método Sanger el DNA blanco es copiado muchas veces, produciendo fragmentos de tamaños diferentes. Nucleotídeos “terminadores de cadeia” fluorescentes marcam los finais de los fragmentos y permitem que la secuencia seja determinada.

# Secuenciamiento de Sanger

Nucleótidos “terminadores de cadena” fluorescentes marcan el final de los fragmentos y permiten que la secuencia sea determinada: dideoxynucleótidos



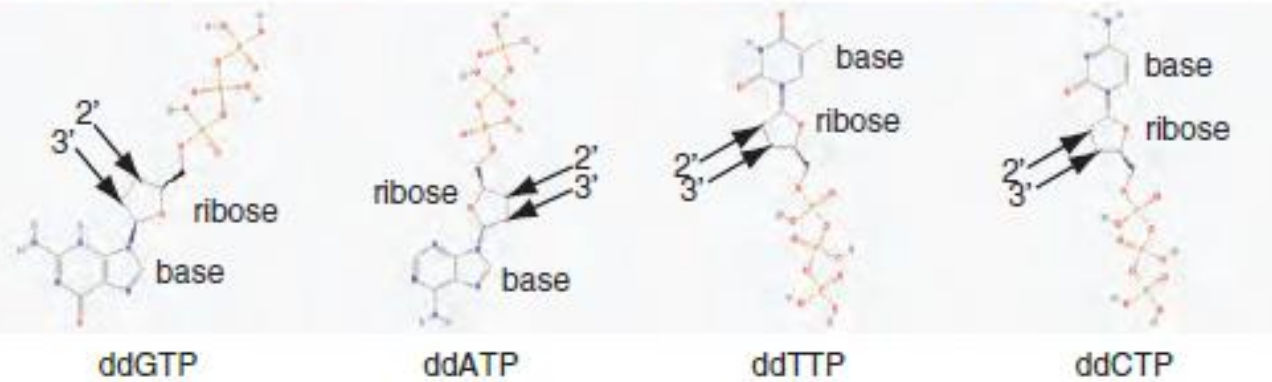
**Dideoxynucleotide (ddNTP)**



**Deoxynucleotide (dNTP)**

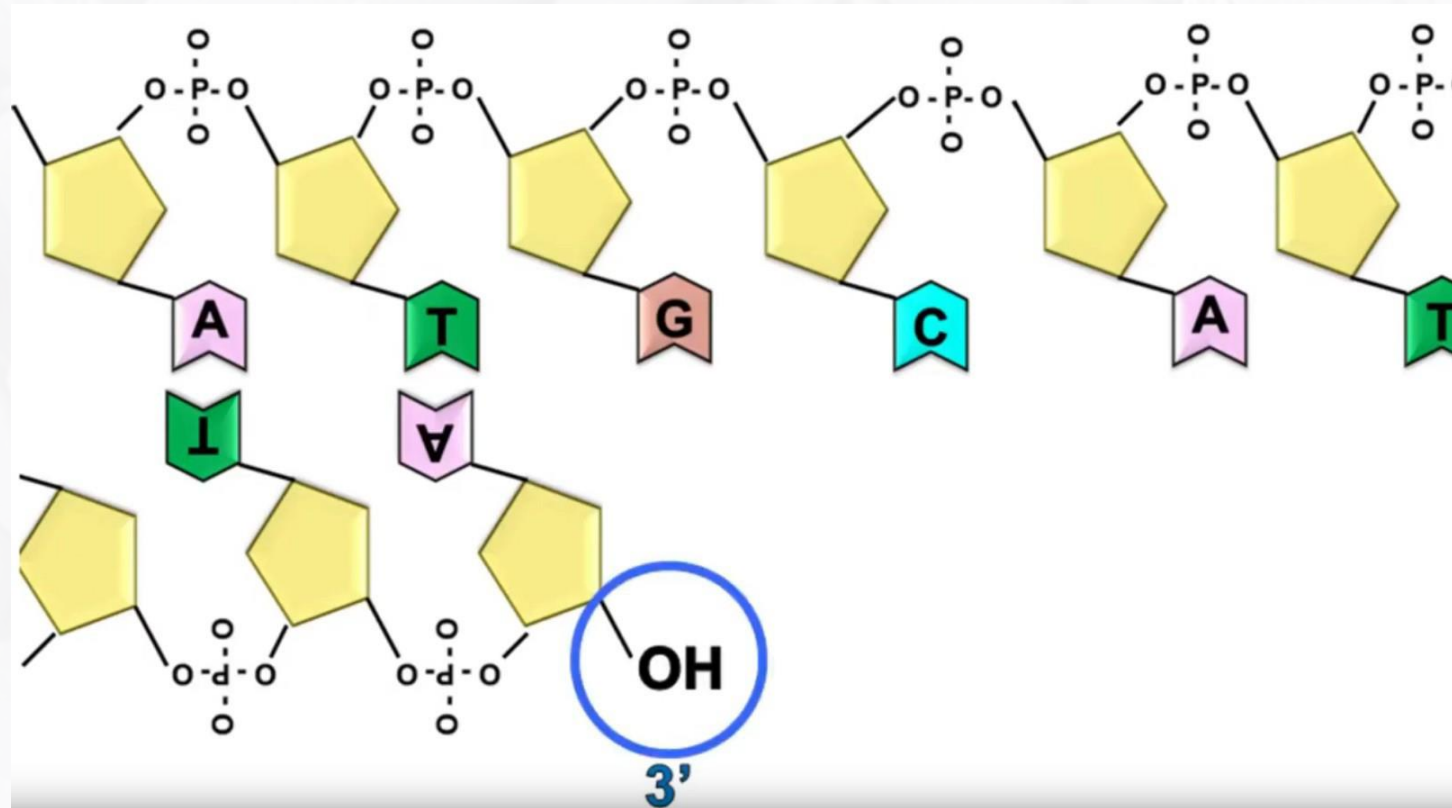
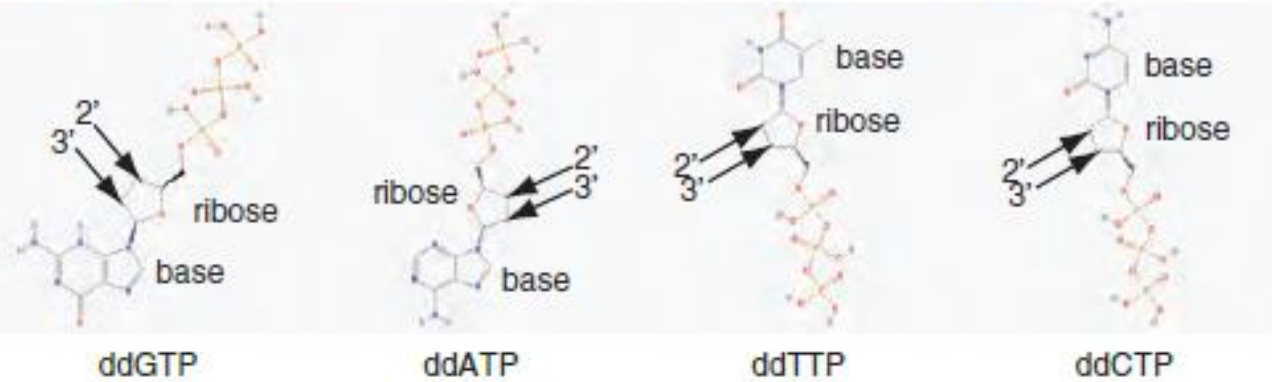


(a) Dideoxynucleotides (ddNTPs) ( -OH of dNTP is replaced by -H of ddNTP at the 2' ribose position)



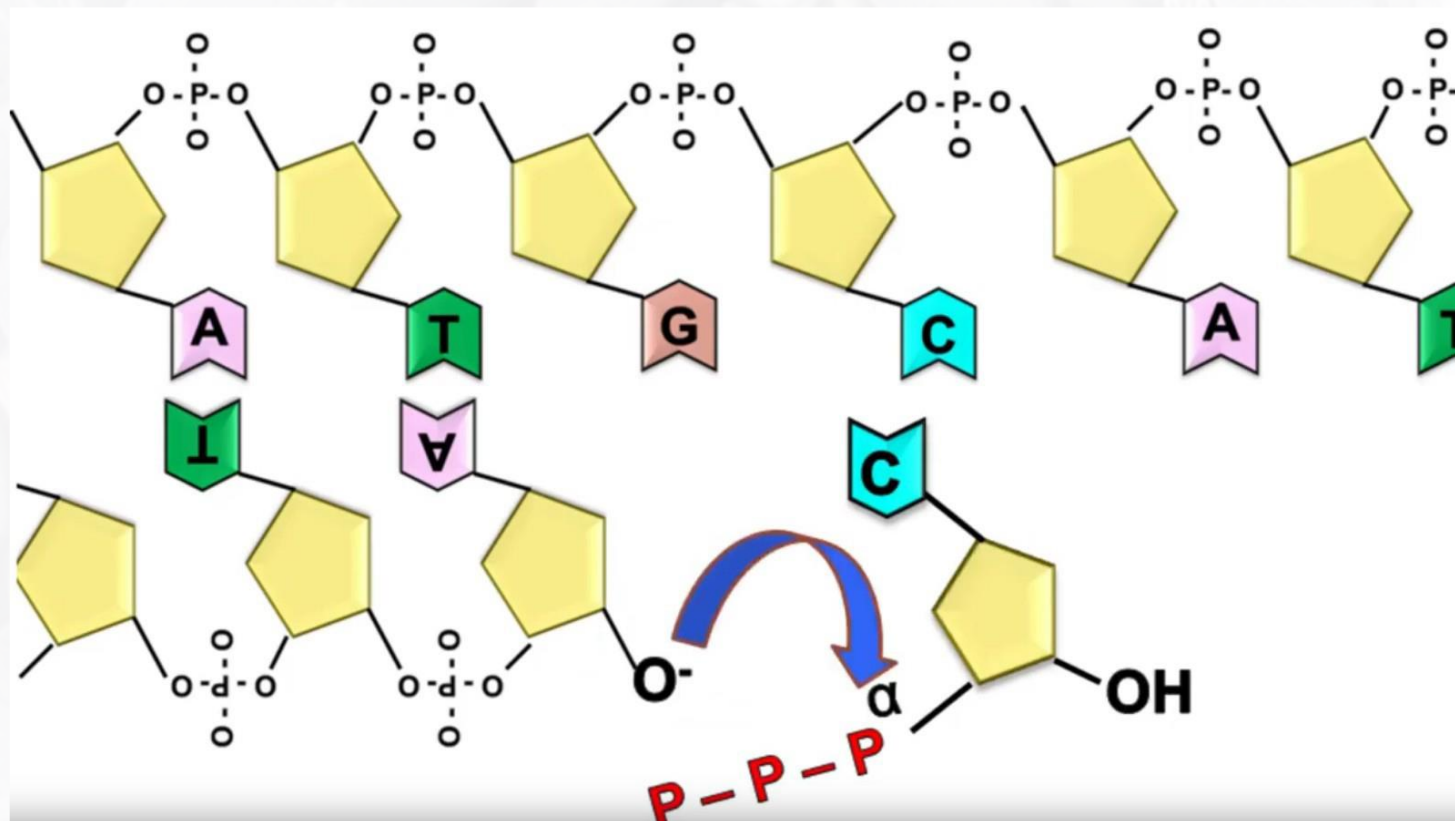
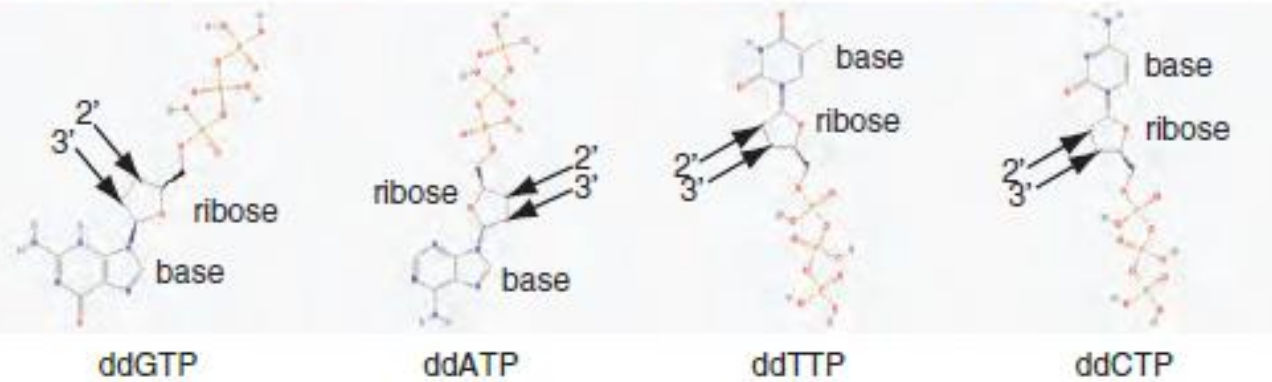
**secuenciamiento  
de Sanger**

(a) Dideoxynucleotides (ddNTPs) ( -OH of dNTP is replaced by -H of ddNTP at the 2' ribose position)





(a) Dideoxynucleotides (ddNTPs) ( -OH of dNTP is replaced by -H of ddNTP at the 2' ribose position)

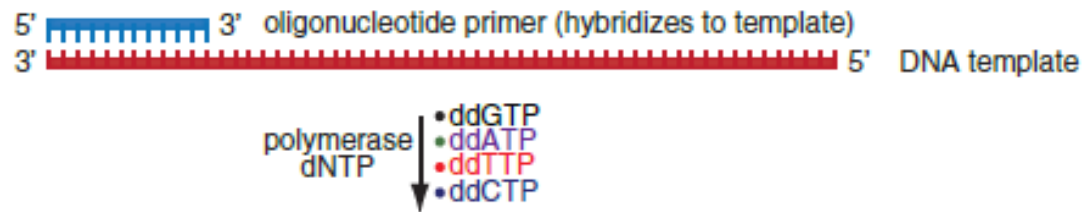


Sequencing  
Sanger

(a) Dideoxynucleotides (ddNTPs) ( -OH of dNTP is replaced by -H of ddNTP at the 2' ribose position)



(b) Primer elongation, chain termination upon incorporation of ddNTP, separation, detection



secuenciamiento  
de Sanger



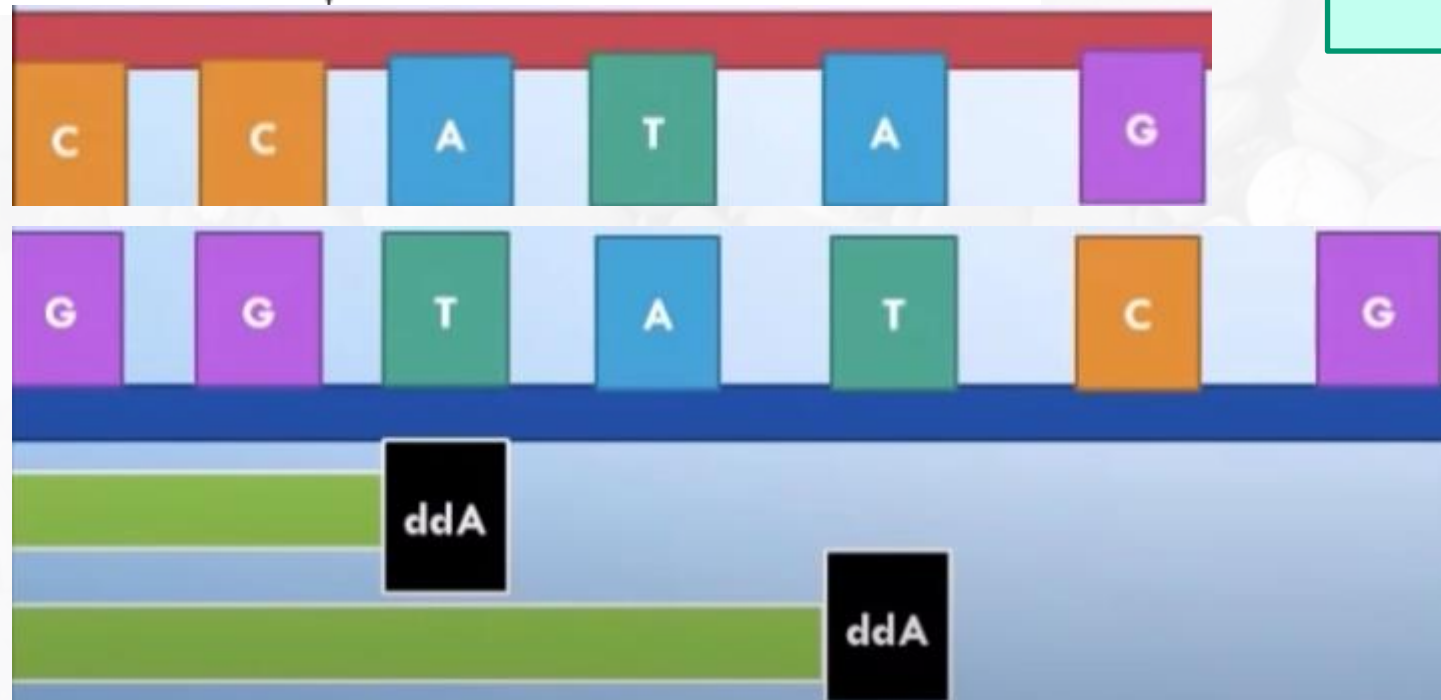
(a) Dideoxynucleotides (ddNTPs) ( -OH of dNTP is replaced by -H of ddNTP at the 2' ribose position)



(b) Primer elongation, chain termination upon incorporation of ddNTP, separation, detection

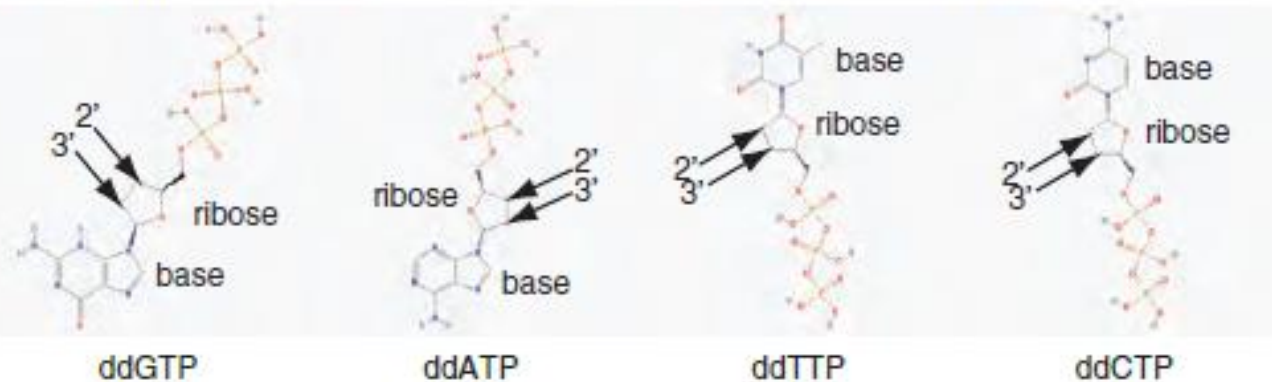
5' 3' oligonucleotide primer (hybridizes to template)  
3' 5' DNA template

polymerase  
dNTP  
• ddGTP  
• ddATP  
• ddTTP  
• ddCTP



secuenciamiento  
de Sanger

(a) Dideoxynucleotides (ddNTPs) ( -OH of dNTP is replaced by -H of ddNTP at the 2' ribose position)

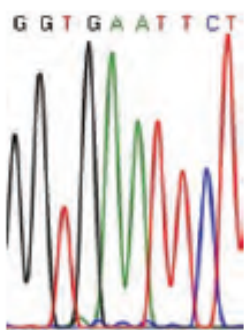


(b) Primer elongation, chain termination upon incorporation of ddNTP, separation, detection

5' 3' oligonucleotide primer (hybridizes to template)  
3' 5' DNA template

polymerase  
dNTP  
• ddGTP  
• ddATP  
• ddTTP  
• ddCTP

5' 3' Chain termination via incorporation of ddGTP  
5' 3' Chain termination via incorporation of ddGTP  
5' 3' Chain termination via incorporation of ddTTP  
5' 3' Chain termination via incorporation of ddGTP  
5' 3' Chain termination via incorporation of ddATP  
5' 3' Chain termination via incorporation of ddATP  
5' 3' Chain termination via incorporation of ddTTP  
5' 3' Chain termination via incorporation of ddTTP  
5' 3' Chain termination via incorporation of ddCTP  
5' 3' Chain termination via incorporation of ddTTP



Capillary gel electrophoresis to separate DNA

fragments by size  
Laser detection of labeled ddNTPs

Determination of DNA sequence  
inferred by pattern of chain termination

secuenciamiento  
de Sanger



# Secuenciamiento del DNA

## Etapas

- Los cromosomas, que varían en tamaño de 50 a 250 millones de bp, deben inicialmente ser fragmentados en fragmentos más cortos (Etapa de Subclonaje)

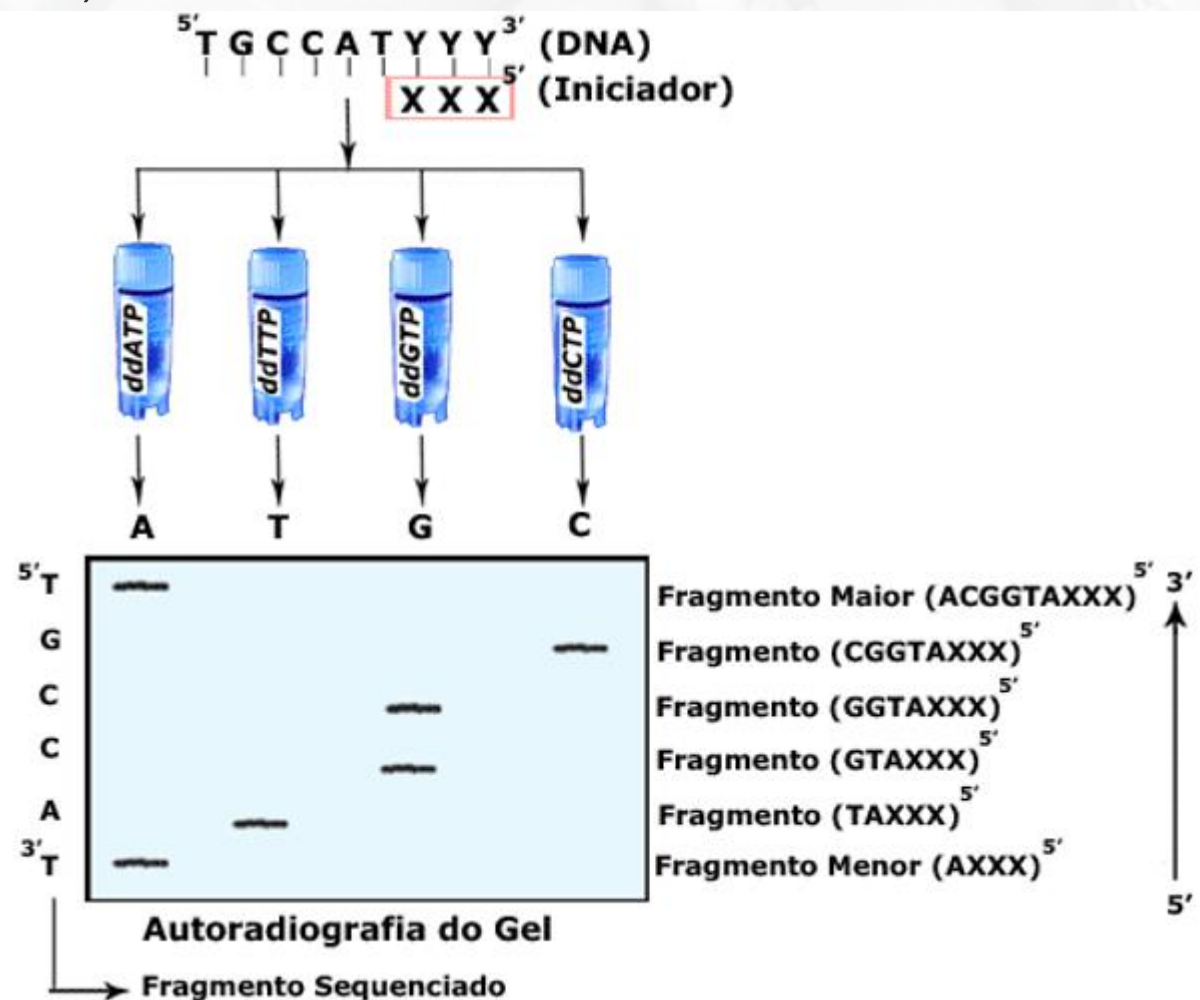


- Cada fragmento corto es usada como un molde para generar un conjunto de fragmentos
  - Cada fragmento difiere en el tamaño por una única base
  - Esta base será identificada en análisis posteriores

# Secuenciamento del DNA

## Etapas

- Los fragmentos son separados por eletroforesis en gel (Etapa de la Separación)



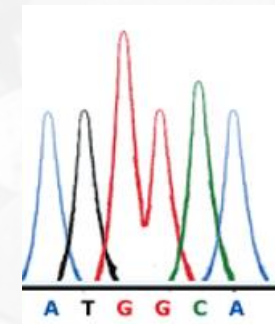
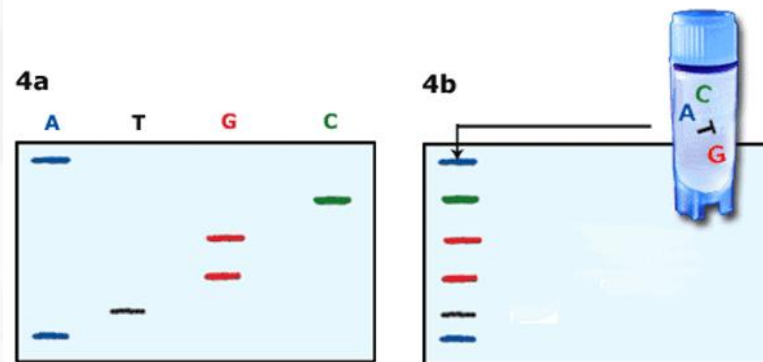


# Secuenciamiento del DNA

## Etapas

- A base final en la extremidad de cada fragmento es identificada (Etapa de base calling)
- Este proceso reconstruye la secuencia original de As, Ts, Cs, y Gs para cada fragmento generado en la primeira etapa
- Los secuenciadores automáticos analizan los eletroferogramas resultantes
- La salida es un cromatograma de cuatro colores que muestra los picos que representan cada una de las cuatro bases del DNA

Analisis de los  
produtos florescentes  
en canales separados  
(4ª) en canal unico  
(4b)do gel de  
poliacrilamida

























Eletroforegrama del gel

# Secuenciamiento del DNA

## Etapas

- Después que las bases “son leídas”, computadores se encargan de ensamblar las lecturas cortas (en bloques de aproximadamente 500 bases cada uno) en trechos largos continuos que serán analizados para filtrar errores, determinar regiones codificadoras de genes, y otras características

Gel:

	G	GCGAATGCGTCCACACGCTACAGGTG
	T	GCGAATGCGTCCACACGCTACAGGT
	G	GCGAATGCGTCCACACGCTACAGG
	G	GCGAATGCGTCCACACGCTACAG
	A	GCGAATGCGTCCACACGCTACA
	C	GCGAATGCGTCCACACGCTAC
	A	GCGAATGCGTCCACACGCTA
	T	GCGAATGCGTCCACACGCT
	C	GCGAATGCGTCCACACGC
	G	GCGAATGCGTCCACACG
	C	GCGAATGCGTCCACAC
	A	GCGAATGCGTCCACAA
	A	GCGAATGCGTCCACA
	C	GCGAATGCGTCCAC
	A	GCGAATGCGTCCA
	C	GCGAATGCGTCC
	C	GCGAATGCGTC
	T	GCGAATGCGT
	G	GCGAATGCG
	C	GCGAATGC
	G	GCGAATG
	T	GCGAAT



# Secuenciamiento del DNA

O secuenciamiento de sanger producía reads de alta calidad (tasa de error menor la 1% por base).  
Las nuevas tecnologías

**TABLE 9.1** Next-generation sequencing technologies compared to Sanger sequencing. Adapted from the companies' websites, [http://en.wikipedia.org/wiki/DNA\\_sequencer](http://en.wikipedia.org/wiki/DNA_sequencer), and literature cited for each technology.

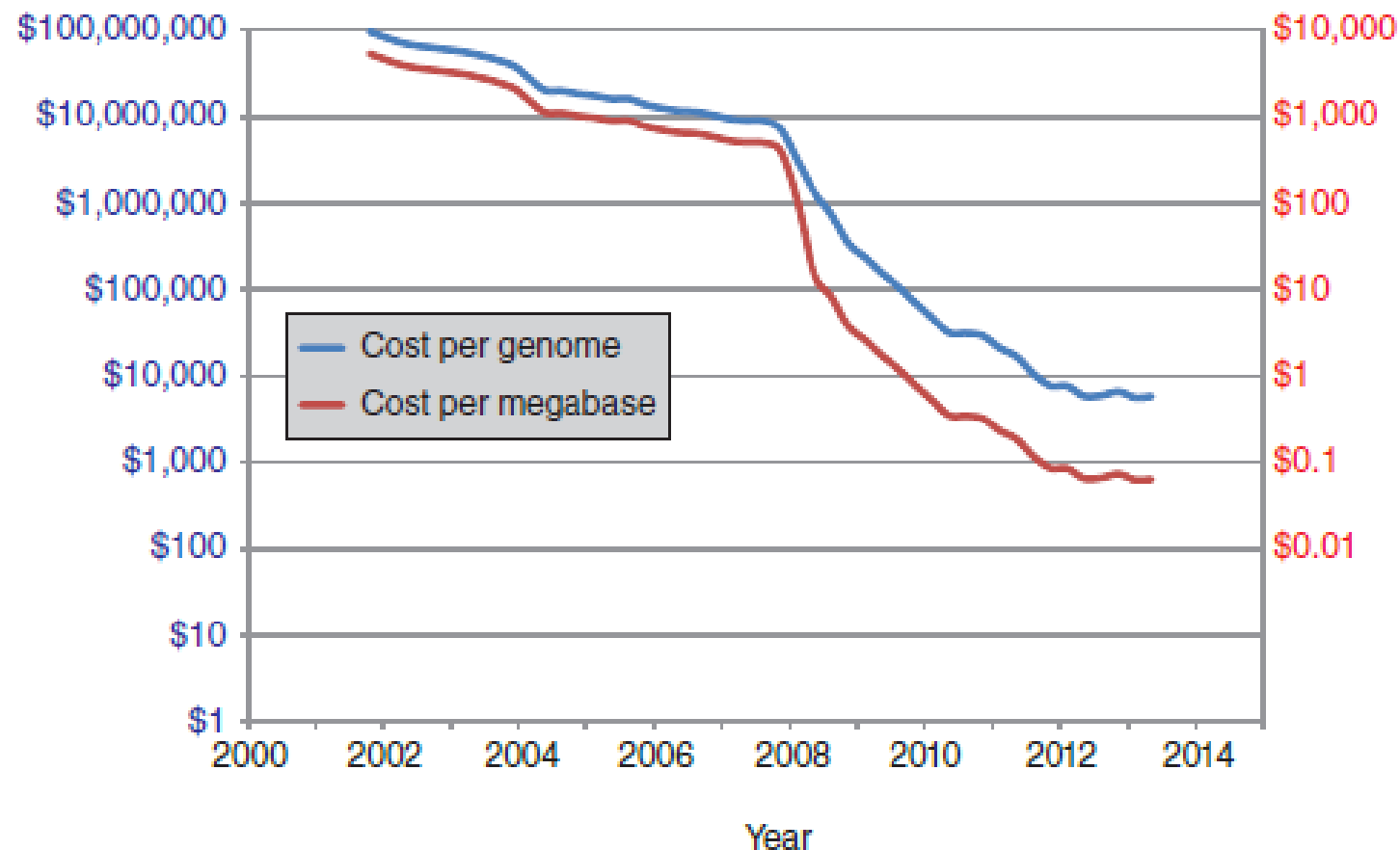
Technology	Read length (bp)	Reads per run	Time per run	Cost per megabase (US\$)	Accuracy (%)
Roche 454	700	1 million	1 day	10	99.90
Illumina	50–250	<3 billion	1–10 days	~0.10	98
SOLiD	50	~1.4 billion	7–14 days	0.13	99.90
Ion Torrent	200	<5 million	2 hours	1	98
Pacific Biosciences	2900	<75,000	<2 hours	2	99
Sanger	400–900	N/A	<3 hours	2400	99.90

# Secuenciamiento del DNA

Estimativa del costo del secuenciamiento

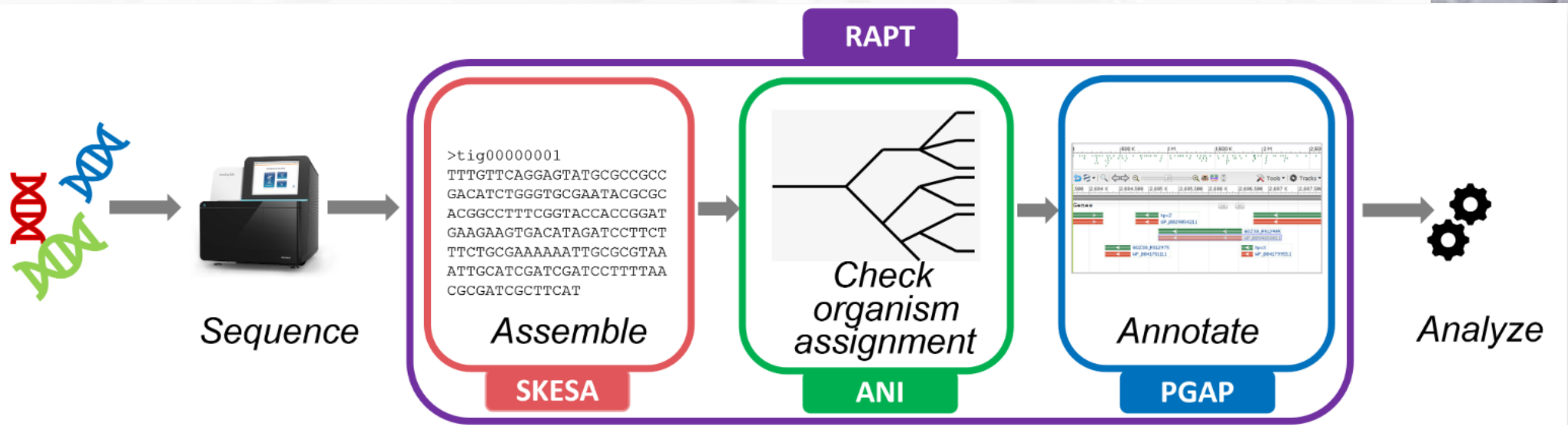
Projeto genoma humano : US\$ 13 billion 15 year

Genoma de Craig Venter : US\$ 80 milhões

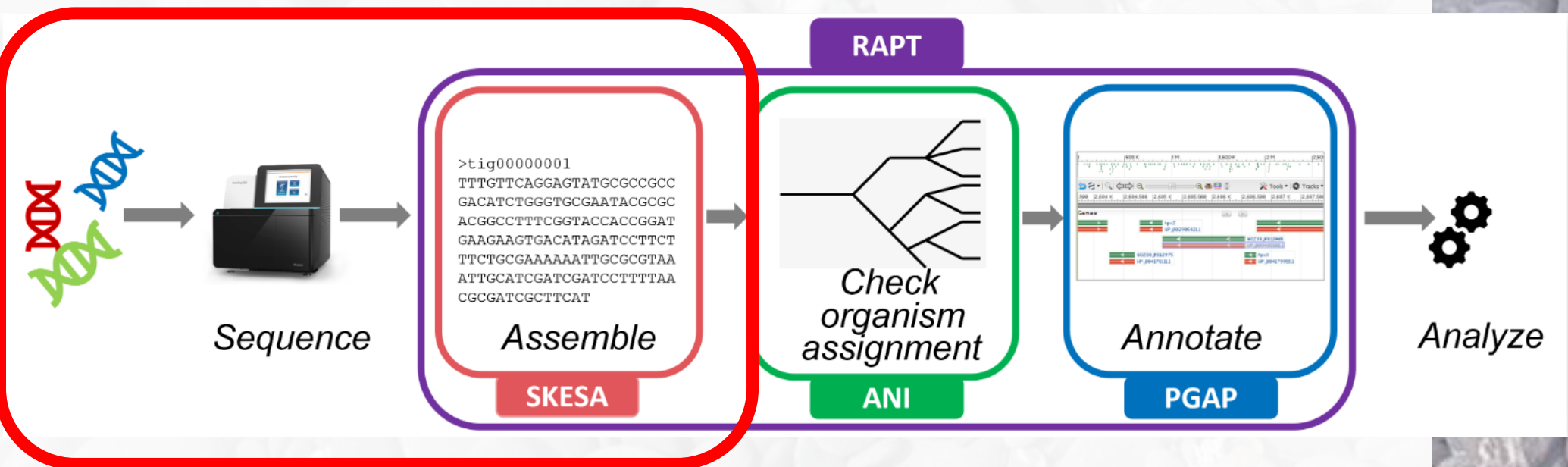




# Montaje y anotacion de reads

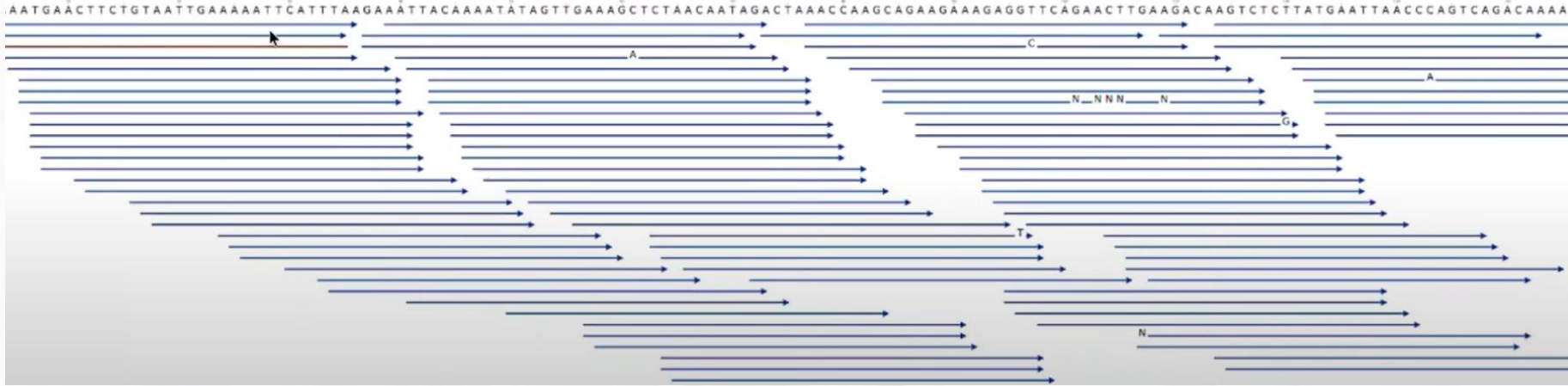


# Montaje y anotacion de reads





# Abordajes para el ensamblaje de genomas



## Ensamblaje por referencia

- Existe un genoma de referencia
- Permite generar contigs que se ensamblan en supercontigs

## Ensamblaje “de novo”

- Cuando no se tiene información previa del genoma
- Requiere mayores recursos computacionales

# Analise del NGS de DNA genomico

Stage	Examples/explanation	File formats
Laboratory work	Experimental design Library preparation Enrichment (capture)	Output: FASTQ-Sanger, FASTQ-Illumina
Next-generation sequencing	Platforms include Illumina, SOLID, Pacific Biosciences, other	
Quality assessment	Trimming, filtering Software: FastQC	FASTQ

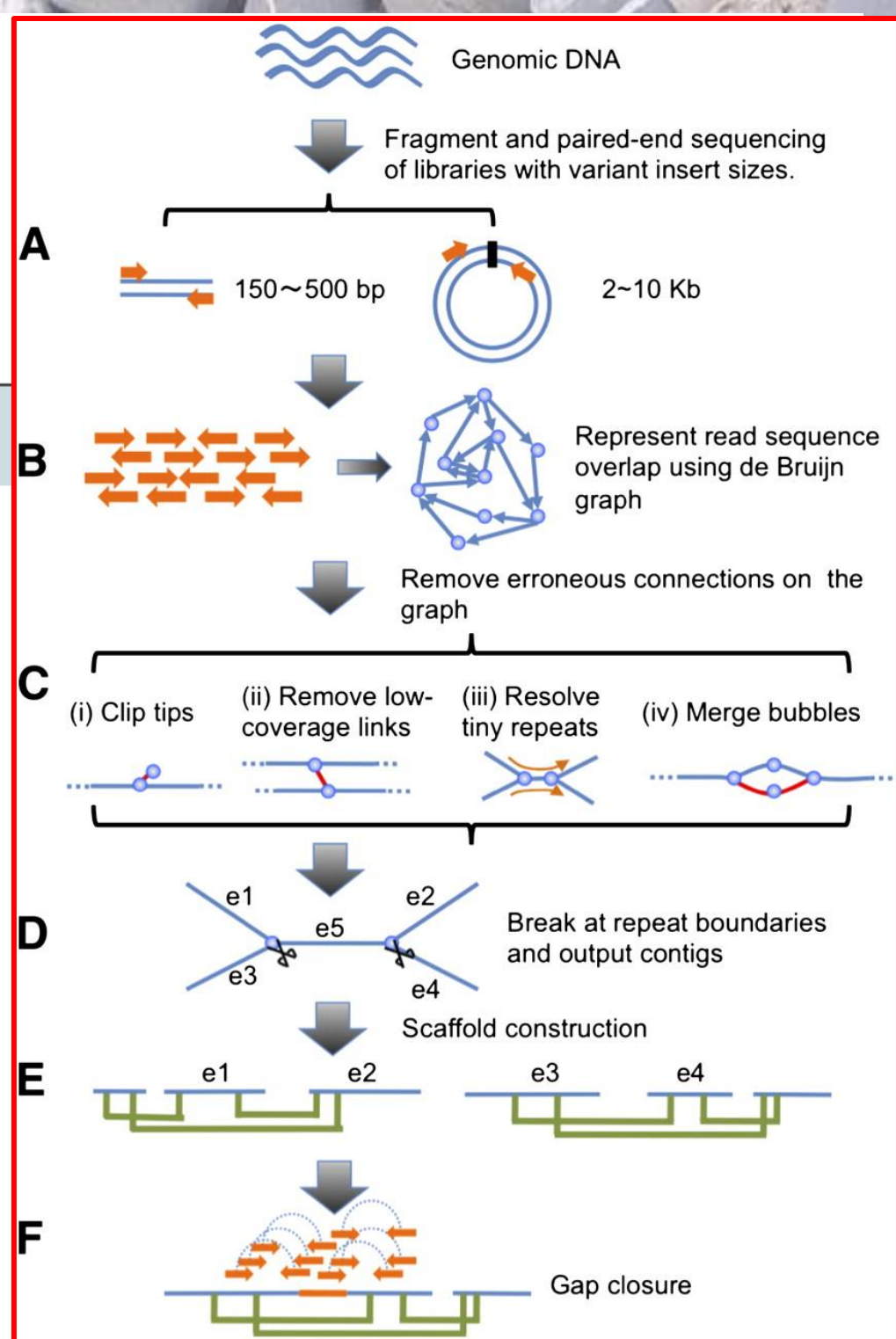
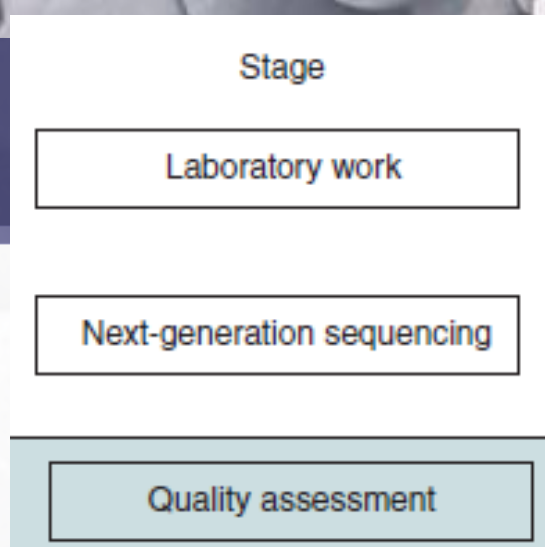
## 1. Ensamblaje de Novo





# Analise del NGS de DNA genomico

## 1. Ensamblaje de Novo



# Analise del NGS de DNA genomico

1. Ensamblaje  
de Novo

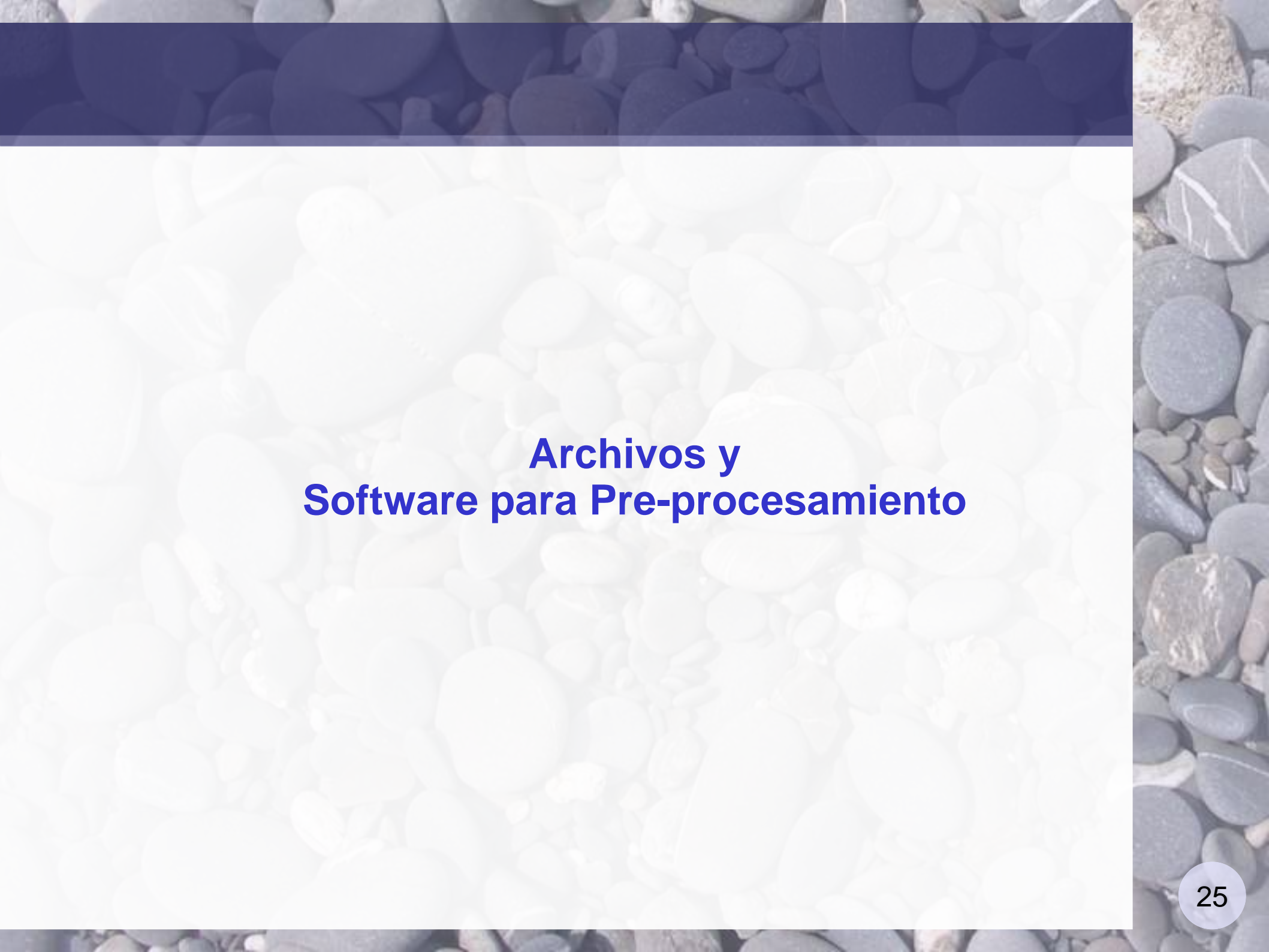


2. Mapeamento

Analysis pipeline

Stage	Examples/explanation	File formats
Laboratory work	Experimental design Library preparation Enrichment (capture)	
Next-generation sequencing	Platforms include Illumina, SOLID, Pacific Biosciences, other	Output: FASTQ-Sanger, FASTQ-Illumina
Quality assessment	Trimming, filtering Software: FastQC	FASTQ
Alignment to reference genome	Software: BWA, Bowtie2	Reference: FASTA Output: SAM/BAM
Variant identification	Single nucleotide variants (SNVs), structural variants (e.g. indels) Software: GATK, SAMTools Realignment, recalibration	Variant Call Format ( VCF/BCF)
Gene Identification	Comparison to public database (dbSNP, 1000 Genomes); functional consequence scores	
Annotation		
Visualization	Variant visualization; read depth; comparison to other samples Software: IGV, BEDTools, BigBED	
Prioritization	Discovery of relevant variants Software: PolyPhen-2, VEP, VAAST	VCF
Storage	Deposit data in ENA, SRA, dbGaP	BAM, VCF





## **Archivos y Software para Pre-procesamiento**

# Formatos de sequencias

## Formatos mais usados

- **FASTA** (.fasta, .fa, .fna)
- **FASTQ** (.fastq ou .fq)
- **SFF** (Sequence Flowgram Format: .sff)
- **CSFASTA** (Color-Space FASTA: .csfasta)
- **SRA** (Sequence Read Archive: .sra)
- **SRF** (Sequence Read Format: .srf)

## Outros formatos

<http://emboss.sourceforge.net/docs/themes/SequenceFormats.html>



## 1. Sequencia fastq extensões:fastq o fq

```
@SOLEXA01:1:1:27:1992#0/1
AGTACAAGAGACAGACATTCTTTTTTTTGACACAAG
+SOLEXA01:1:1:27:1992#0/1
\FFMXPYDDHJSUMVUJLPSNFRXZEDLNLHKHIT
```

### Identificadores Illumina

SOLEXA01	the unique instrument name
1	flowcell lane (8 lanes)
1	tile number within the flowcell lane
27	'x'-coordinate of the cluster within the tile
1992	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 ( <i>paired-end or mate-pair reads only</i> )

2. Codificación de calidad : codificada como un unico caracter de la tabela ASCII

$$F = 70 \text{ (ascii)} = 70 - 64 = 6 \text{ (Q}_{phred}\text{)} = 0,25 \text{ (P}_{error}\text{)}$$

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....  
.....XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX  
.....IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII  
.....JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ  
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMN  
|          |          |          |          |          |  
33         59        64         73         104        126
```

```
S - Sanger      Phred+33,  raw reads typically (0, 40)
X - Solexa     Solexa+64,  raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
```

$$Q_{phred} = -10 \log_{10} (P_{error})$$



2. Codificação de qualidade : codificada como um unico caracter de la tabela ASCII

$$F = 70 \text{ (ascii)} = 70 - 64 = 6 \text{ (} Q_{\text{phred}} \text{)} = 0,25 \text{ (} P_{\text{error}} \text{)}$$

```
>SEQUENCE_1  
1 9 7 15 20 21 16 26 31 37 38 ...  
31 13 23 29 31 33 35 30 29 34 ...
```

	Score	$P_{\text{erro}}$
	10	0.1
$Q_{\text{phred}} = -10 \log_{10} (P_{\text{error}})$	20	0.01
	30	0.001

Formato padrão para armazenamento para los datos de secuenciamiento NGS nos repositórios SRA, ENA y DRA.

### **SRA Toolkit**

Herramientas para importar en el formato SRA o exportar del SRA para outros formatos

**Alguns paquetes son:**

- sffload y sffdump
- fastqload y fastqdump

<http://www.ncbi.nlm.nih.gov/Traces/sra/?view=software>

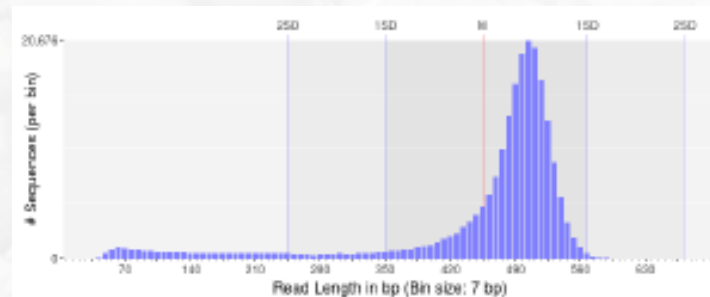


## Herramientas para chequear la calidad de los reads

**Assemblstats** : <http://community.g2.bx.psu.edu/tool/>

Métricas en archivos fasta

- Min read length
- Max read length
- Mean read length
- Standard deviation of read length
- Median read length
- N50 read length

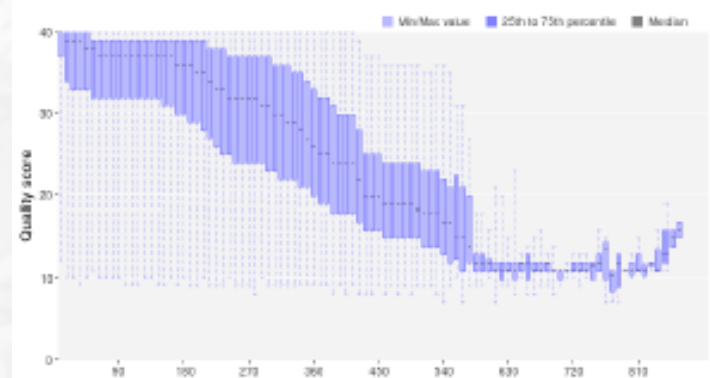


## PRINSEQ

<http://prinseq.sourceforge.net>

Métricas en archivos fasta, qual y fastq

- Filtros
- qualidade
- poly(A)
- Conteúdo de GC
- Duplicaciones



## FASTQC

<http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>

Necesita conversion para fastq

- sff2fastq <https://github.com/indraniel/sff2fastq>



## **Ensamblaje de novo**



## Ensamblaje “de Novo”

A Ensamblaje de genoma ofrece una representación consensual de un genoma, abarcando todos los Cromosomas (e elementos extracromosómicos, como genomas organelares y plasmídeos).

Consiste en la reconstrucción de la secuencia en su forma original, sin la consulta de secuencias previamente resueltas de genomas, transcritos y proteínas.

El ensamblaje es posible cuando el blanco es excesivamente muestreado con lecturas “*shotgun*” que se superponen.

# Ensamblaje “de Novo”

Se trata de una estructura jerarquica que mapea los datos de secuencias de fragmentos (reads) para una reconstrucción aproximada del genoma blanco (target) en la forma original;

$\text{lecturas}(\text{reads}) \Rightarrow \text{contigs} \Rightarrow \text{scaffolds}$

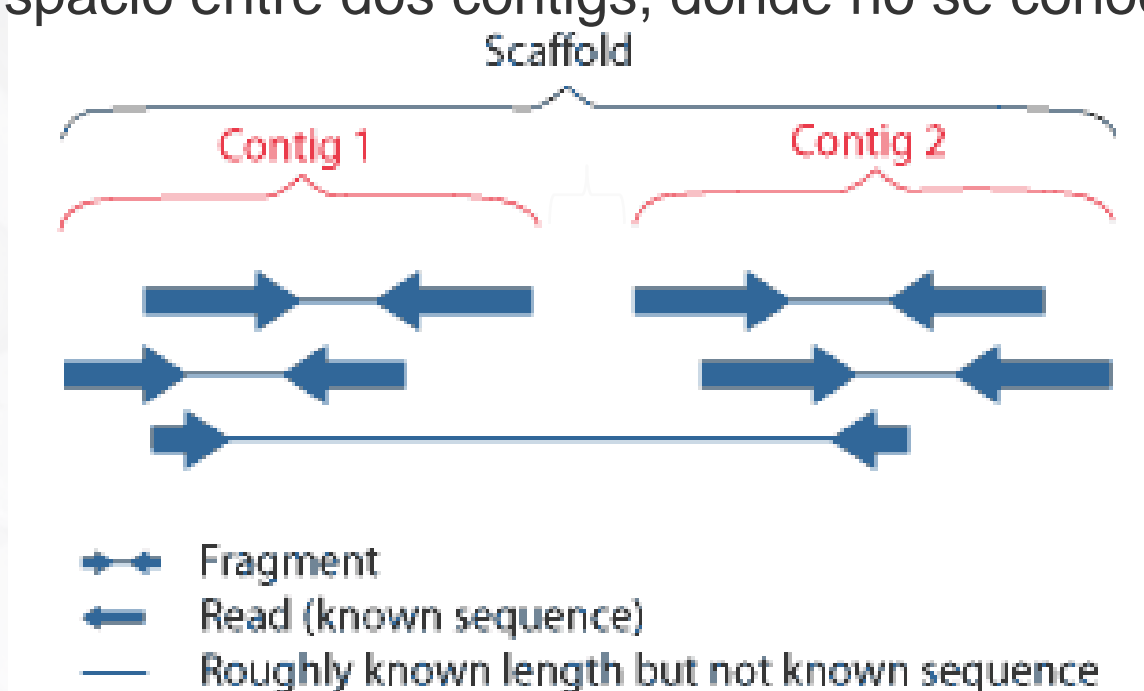
En este proceso se agrupan secuencias(reads) en contigs y *contigs* en *scaffolds* (*supercontigs*);

El ensamblaje sera posible cuando el blanco seja excessivamente secuenciado;



# Ensamblaje “de Novo”

- **Contig** – Alineamiento múltiplo de leituras de donde se extrae una secuencia consenso;
  - **unitig** – contig formado por la sobreposición de secuencias únicas de las leituras, o seja, sem ambiguidades;
- **Scaffold** – define la orden y orientación de los contigs y el tamaño de los gaps entre los contigs;
- **Singlets** – leituras no agrupadas en un contig;
- **gap** – Espacio entre dos contigs, donde no se conoce la secuencia;



# Ensamblaje “de Novo”

## Cobertura (*coverage*)

Seam :

N: Numero de reads

L : Tamanho del read

G : Tamanho del genoma

A cobertura pode ser calculada como el total de pares de bases secuenciadas  $[N \cdot L]$  dividido por el tamanho de la região de interes (genoma)  $[G]$ :

$$((N \cdot L) / G)$$

Exemplo: Genoma de 2Mbp (G)

10 milhões de reads (N) de 50bp (L)

$$\text{Cobertura} = (10.000.000 \cdot 50) / 2.000.000 = 25X$$

Este numero (25x) representa cuantas veces, en média, cada base del genoma foi secuenciada



## Profundidad (*depth of coverage*)

Requisitos para el secuenciamiento de genomas:

Exemplos:

Roche 454: J. Watson (3Gb ~7.4x)

[Wheeler *et al.*, 2008]

Illumina (52pb): Pan de la (*Ailuropo de la melanoleura*)  
(2.4Gb ~73x)

[Li *et al.*, 2010]

# Ensamblaje “de Novo”

Las estadísticas globales para ensamblajes incluyen:

1. El número total de scaffolds (incluyendo aquellos con o sin orientación conocida);
2. El scaffold N50 (o tamaño en pares de bases de modo que los scaffolds deste tamaño o mais incluam 50% das bases en la Ensamblaje);
3. El número total de contigs;
4. El contig N50 (tamaño del menor contig en el conjunto de los maiores contigs que combinados representam 50% de la Ensamblaje) – contiguity -uma medi de la de contiguidade, com valores maiores denotando mais montagens completas
5. Valores muito altos podem representar erros en la Ensamblaje y valores muito pequenos podem representar Ensamblaje incompleta;



# Ensamblaje “de Novo”



4. El contig N50 (tamaño del menor contig en el conjunto de los maiores contigs que combinados representan 50% de la Ensamblaje) – contiguity -uma medi de la de contiguidade, com valores maiores denotando mais montagens completas
5. Valores muito altos podem representar erros en la Ensamblaje y valores muito pequenos podem representar Ensamblaje incompleta;

## Modelo Lander-Waterman

Para estimar la cobertura, y estimar parâmetros :  
número esperado de *contigs* y tamaño de los *contigs*  
(Lander y Waterman, 1988)

L = tamanho das leituras

T = mínimo de sobreposição entre leituras

G = tamanho del genoma (pool de transcritos)

N = número de leituras

$c$  = cobertura ( $NL / G$ )

$\sigma = 1 - T/L$

$E(\#contigs) = Ne^{-c\sigma}$

$E(\text{tamanho del } contig) = L((e^{c\sigma}-1)/c+1-\sigma)$



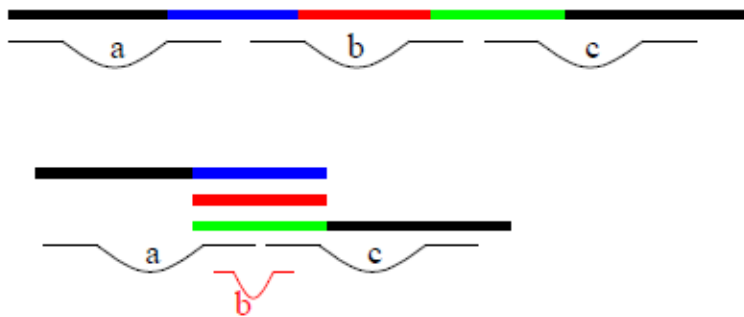
## Dificuldades en la Ensamblaje

- Contaminates nas amostras (e.g. Bacteria)
- Ribosomal RNA (pequenas y grandes sub-unidades)
- Artefatos gerados en la etapa de PCR (e.g. Quimeras y mutaciones)
- Repeticiones y genomas poliplóides (secuencias repetitivas torna la Ensamblaje mais difícil);
  - Utilización de leituras *paired-ends/mate-pairs* y suas propiedades de tamanho y orientación, estando un de los pares ancorado en una região única;
- Genes parálogos

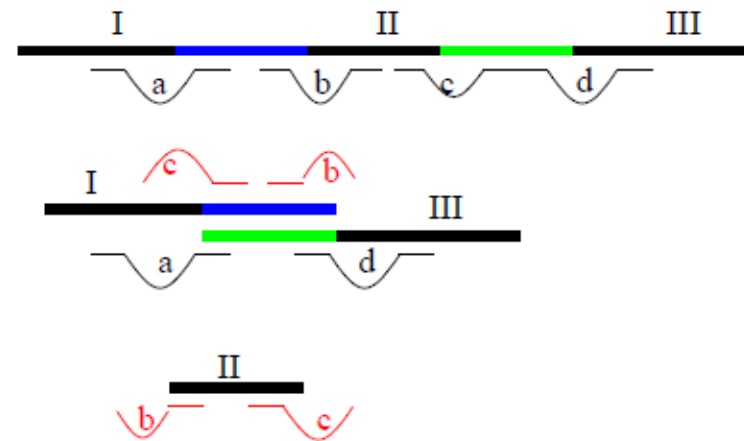
# Ensamblaje “de Novo”

## Problemas causados por repeticiones

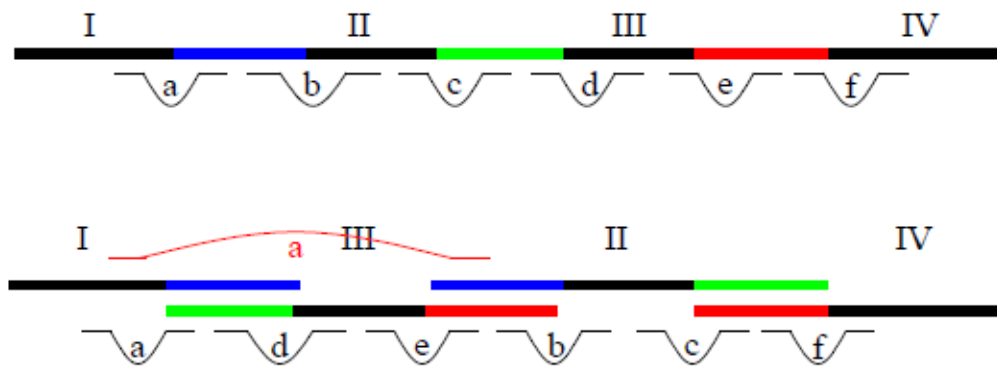
collapsed tandem



excision



rearrangement





# Ensamblaje “de Novo”

## Software para Ensamblaje de novo

TABLE 9.2. Software for genome assembly.

Assembler	Reference	URL
ABYSS	Simpson <i>et al.</i> (2009)	<a href="http://www.bcgsc.ca/platform/bioinfo/software">http://www.bcgsc.ca/platform/bioinfo/software</a>
ALLPATHS-LG	Gnerre <i>et al.</i> (2011)	<a href="http://www.broadinstitute.org/software/allpaths-lg/blog/">http://www.broadinstitute.org/software/allpaths-lg/blog/</a>
Bambus2	Koren <i>et al.</i> (2011)	<a href="http://www.cbcb.umd.edu/software">http://www.cbcb.umd.edu/software</a>
CABOG	Miller <i>et al.</i> (2008)	<a href="http://www.jcvi.org/cms/research/projects/cabog/overview/">http://www.jcvi.org/cms/research/projects/cabog/overview/</a>
SGA	Simpson and Durbin (2012)	<a href="https://github.com/jts/sga">https://github.com/jts/sga</a>
SOAPdenovo	Luo <i>et al.</i> (2012)	<a href="http://soap.genomics.org.cn/soapdenovo.html">http://soap.genomics.org.cn/soapdenovo.html</a>
Velvet	Zerbino and Birney (2008)	<a href="http://www.ebi.ac.uk/~zerbino/velvet/">http://www.ebi.ac.uk/~zerbino/velvet/</a>

Dois métodos principais son usados pelos montadores: a abordagem de sobreposição / layout / consenso e; os gráficos de Bruijn.

# Ensamblaje “de Novo

## “*k*-mers”

Subsecuencias de tamaño  $k$

En una secuencia de tamaño  $(L)$  há  $(L-k+1)$   $k$ -mers;

Ejemplo: secuencia de tamaño  $L=8$  tem 5  $k$ -mers com  $k=4$

**ACGTACGA**

ACGT

CGTA

GTAC

TACG

ACGA

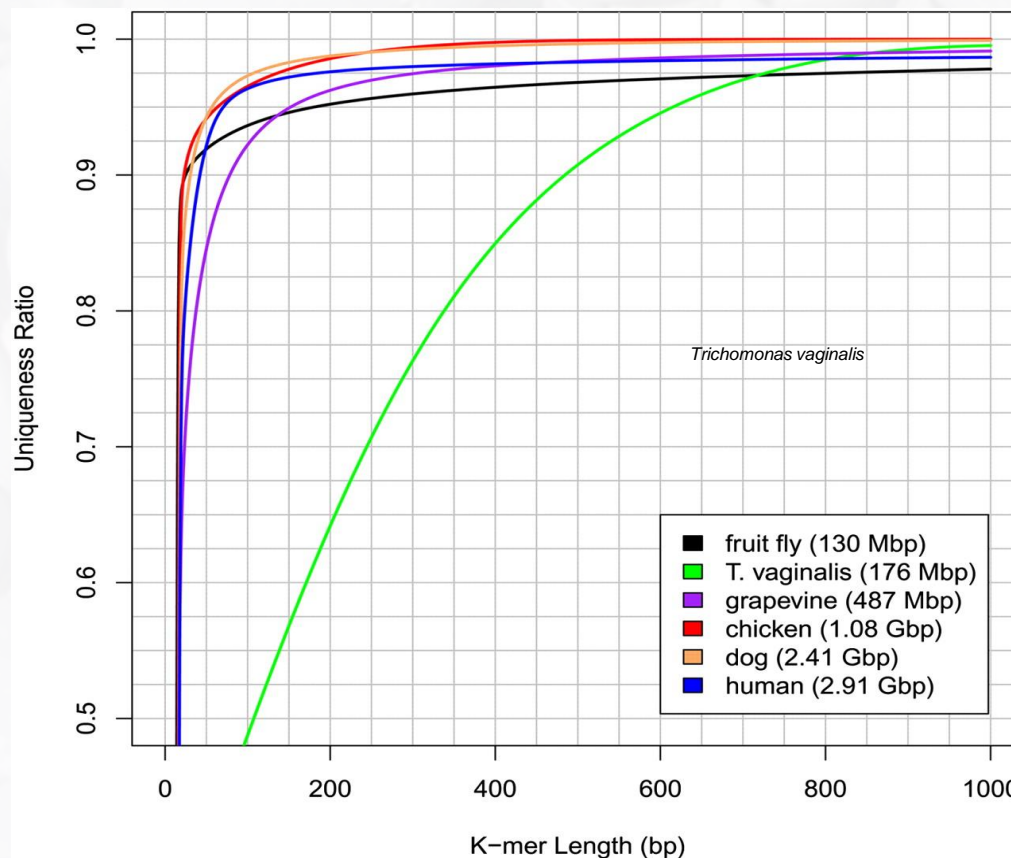


# Ensamblaje “de Novo”

***k*-mers** – secuencias de tamaño *k*

***k*-mers uniqueness ratio**

nro de *k*-mers distintas que ocurren una única vez en el genoma  
nro total de *k*-mers distintos que ocurren en el genoma

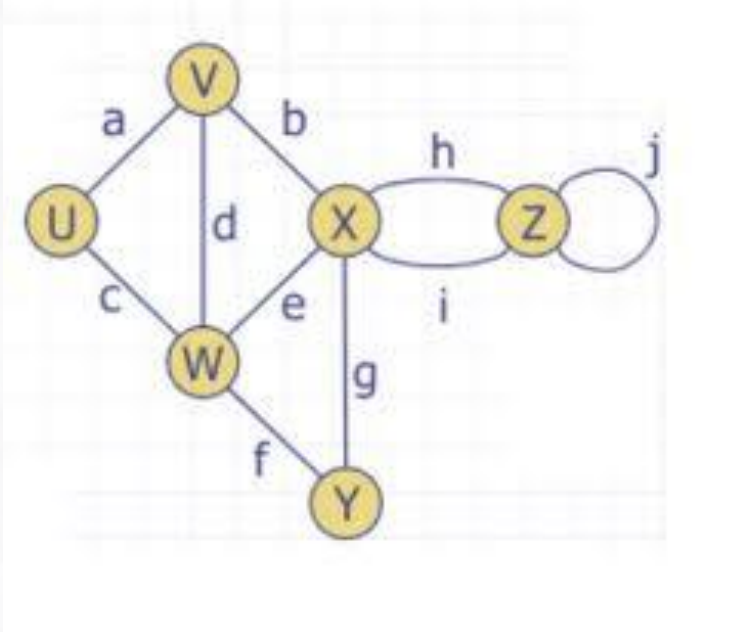


[Schatz et al., 2010]

# Ensamblaje “de Novo

**Grafo** es una estructura  $G(V, A)$  onde  $V$  es un conjunto não vazio de objetos denominados **nós** o **vértices** (*nodes/vertices*) e;

$A$  es un conjunto de pares não ordenados de  $V$ , chamado **arestas** o **arcos** (*edges/arcs*).



Nós (vértices):  $V = \{U, V, W, X, Y, Z\}$

Arestas (arcos):  $I_a = \{a, b, c, d, e, f, g, h, i, j\}$

Representación simplifiCada de un grafo qualquer

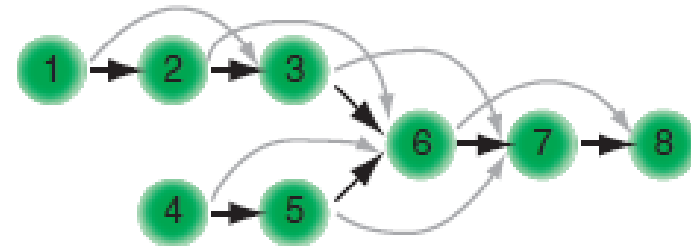


# Ensamblaje “de Novo”: Metodo de Bruijn

(a) Reads

1 ACCTGATC  
2 CTGATCAA  
3 TGATCAAT  
4 AGCGATCA  
5 CGATCAAT  
6 GATCAATG  
7 TCAATGTG  
8 CAATGTGA

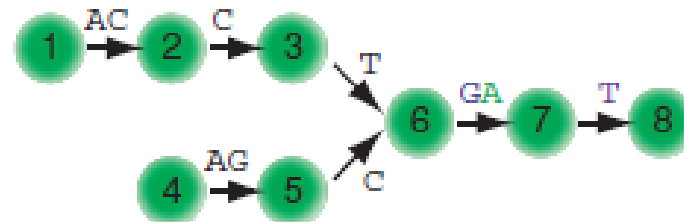
(b) Overlap graph



(c) de Bruijn graph

ACCTG ▶ CCTGA ▶ CTGAT ▶ TGATC  
AGCGA ▶ GCGAT ▶ CGATC  
GATCA ▶ ATCAA ▶ TCAAT ▶ CAATG ▶ AATGT ▶ ATGTG ▶ TGTGA

(d) String graph



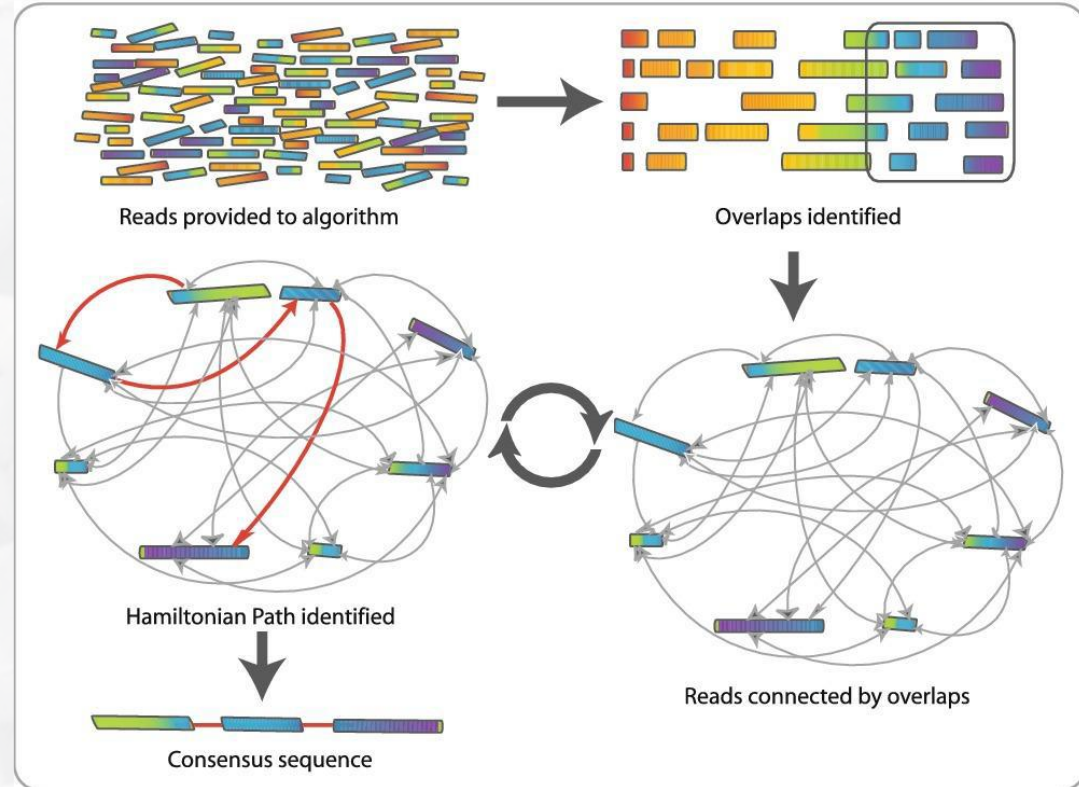
**FIGURE 9.9** Methods for genome assembly from short reads. (a) Example of 8 aligned reads (note that reads 4 and 5 only partially match reads 1–3). Colored nucleotides are identical for all aligned sequences. (b) Overlap graph represents a solution to the assembly. (c) de Bruijn graph breaks the reads into units of five nucleotide ( $k$ -mers with  $k = 5$  in this example). Colors of nucleotides match (a). Adapted from Henson *et al.* (2012) with permission from Future Medicine.

# Ensamblaje “de Novo” : *Overlap/Layout/Consensus* (OLC)

Etapas:

1º **Detecção de sobreposição;**  
Alineamiento pareado entre todas las lecturas – identificación de los pares com mejor *match* (alinhamento global + heurísticas [e.g. *seed & extend*]);

2º **Layout de los fragmentos**  
(Ensamblaje del *contig*);  
Construcción y manipulación del grafo de sobreposição  
(Analizar/Simplificar/Limpar);  
**Caminho Hamiltoniano;**



3º **Decisão de la secuencia** (Ensamblaje del consenso);  
Alinhamento Múltiplo de secuencias – normalmente baseado en la puntuación de los pares com sobreposição (sum-of-pairs o SP);  
Realiza ajustes en el layout se necessário;  
Normalmente la frequência de un nucleotídeo en determina de la posição determina la base consenso;

S1	A	T	C	T	C	G	A	-	-	G	A
S2	A	T	C	-	C	G	A	-	-	G	A
S3	A	T	G	T	C	G	A	C	-	G	A
S4	A	T	G	T	C	G	A	C	A	G	A
S5	A	T	-	T	C	A	A	C	-	G	A

match score= 5  
mismatch score= -1  
gap penalty= -3

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	
No. of match pairs	10	10	2	6	10	6	10	3	-	10	10	
No. of mismatch pairs	-	-	4	-	-	4	-	-	-	-	-	
No. of gap-base pairs	-	-	4	4	-	-	-	6	4	-	-	
Column score (Cs <sub>i</sub> )	50	50	-6	18	50	26	50	-3	-12	50	50	

Score for C3=  $2(5) + 4(-1) + 4(-3)$

SP score =  $\sum_{i=1}^{10} Cs_i = 323$



# Ensamblaje “de Novo” : *Overlap/Layout/Consensus* (OLC)

Utilizam el paradigma OLC:

Phrap (<http://www.phrap.org/>)

genomas

Sanger, 454

(Green, P., 1994 - unpublished)

CAP3 (<http://seq.cs.iastate.edu/>)

genomas, **cDNAs**

Sanger, 454

(Huang, X. *and* Madan, A., 1999)

MIRA (<http://sourceforge.net/projects/mira-assembler/>)

genomas, **cDNAs**

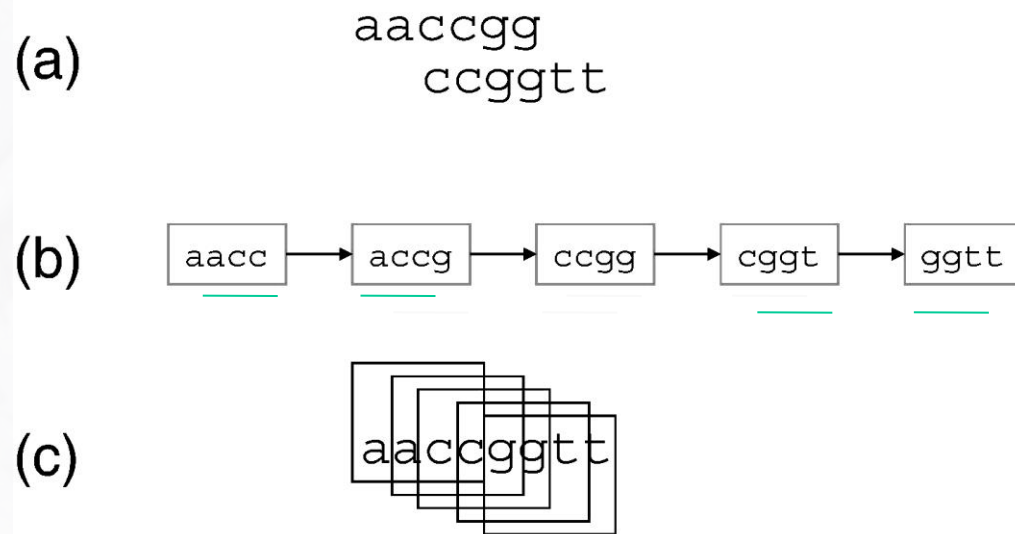
Sanger, 454, Solexa

(Chevreux, B. *et al.*, 1999) (Chevreux, B. *et al.*, 2004)

# Ensamblaje “de Novo”: Metodo de Bruijn

## Grafos k-mer

nodos – todas las subsecuencias de tamanho  $k$ ;  
aristas – todas las sobreposiciones ( $k-1$  bases) entre essas subsecuencias que son consecutivas en la secuencia original;  
Puede representar las múltiplas secuencias das leituras y implicitamente las sobreposiciones;



**aaccgg** (k-mer 4):  
aacc  
accg  
ccgg

**ccggtt** (k-mer 4):  
ccgg  
cggt  
gggt

### Grafo de de-Bruijn:

**nó** – subsecuencia ( $k$ -mer);  
**arestas** – sobreposiciones;

**Caminho Euleriano** – caminho que atravessa cada aresta una única vez  
(*contig*) – caminho simples;



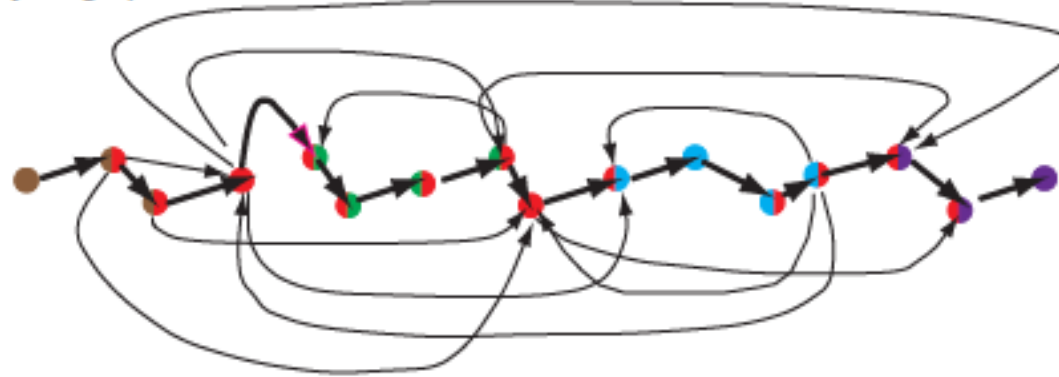
# Ensamblaje “de Novo”: Metodo de Bruijn

Ensamblaje  
eficiente de  
regiones  
repetitivas

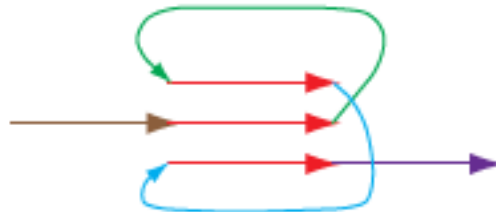
(a) DNA sequence with a triple repeat



(b) layout graph



(c) Construction of de Bruijn graph by gluing repeats



(d) de Bruijn graph



**FIGURE 9.10** Efficient assembly of repetitive DNA regions using a de Bruijn graph. (a) A genomic DNA segment is shown having four unique segments and three repeats. (b) The layout graph represents these repeats with a complex set of possible paths. (c) The de Bruijn graph is constructed by “gluing” repeats. (d) The de Bruijn graph represents repeat regions as edges rather than as a set of vertices in the layout graph.

# Ensamblaje “de Novo”: Metodo de Bruijn

## Vantajas

- Desarrollados para lidar com la alta complexidad y el grande volume de dados de los NGS;
- Rápida detecção de k-mers compartilhados - reduce costo computacional en relación a la busca de sobreposiciones en alinhamentos pareados;
- Não necessita comparaciones pareadas (todas x todas);

## Desvantagens

- Usam muita memória (tabla *hash k-mers*);
- Mais sensível a repeticiones y a errores de secuenciamiento;
- baixa sensibilidade (perde algumas sobreposiciones verdadeiras), dependendo do:
  - tamanho de k
  - tamanho de la sobreposição
  - taxa de erro nas leituras



# Ensamblaje “de Novo”: Metodo de Bruijn

Software basado en grafos de de-Bruijn:

VELVET /Oases (<http://www.ebi.ac.uk/~zerbino/velvet/>)

genomas, **cDNAs**

Solexa, SOLiD

(Zerbino, D.R. y Birney E., 2008)

ABYSS/Trans-ABYSS

(<http://www.bcgsc.ca/platform/bioinfo/software/abyss>)

genomas, **cDNAs**

Solexa, SOLiD

(Simpson, J.T, *et al.*, 2009) (Birol, I., *et. al.*, 2009)



## **Ensamblaje por referencia**



# Alinhadores o mapeadores de sequencias

## Bowtie

Usage:  
`bowtie [options]* <ebwt> {-1 <m1> -2 <m2> | --12 <r> | <s>} [<hit>]`

<m1> Comma-separated list of files containing upstream mates (or the sequences themselves, if -c is set) paired with mates in <m2>  
<m2> Comma-separated list of files containing downstream mates (or the sequences themselves if -c is set) paired with mates in <m1>  
<r> Comma-separated list of files containing Crossbow-style reads. Can be a mixture of paired and unpaired. Specify "-" for stdin.  
<s> Comma-separated list of files containing unpaired reads, or the sequences themselves, if -c is set. Specify "-" for stdin.  
<hit> File to write hits to (default: stdout)

### Input:

-q query input files are FASTQ .fq/.fastq (default)  
-f query input files are (multi-)FASTA .fa/.mfa  
-r query input files are raw one-sequence-per-line  
-c query sequences given on cmd line (as <mates>, <singles>)  
-C reads and index are in colorspace  
-Q/--quals <file> QV file(s) corresponding to CSFASTA inputs; use with -f -C  
--Q1/--Q2 <file> same as -Q, but for mate files 1 and 2 respectively  
-s/--skip <int> skip the first <int> reads/pairs in the input  
-u/--upto <int> stop after first <int> reads/pairs (excl. skipped reads)  
-5/--trim5 <int> trim <int> bases from 5' (left) end of reads  
-3/--trim3 <int> trim <int> bases from 3' (right) end of reads  
--phred33-quals input quals are Phred+33 (default)  
--phred64-quals input quals are Phred+64 (same as --solexa.3-quals)  
--solexa-quals input quals are from GA Pipeline ver. < 1.3  
--solexa1.3-quals input quals are from GA Pipeline ver. >= 1.3  
--integer-quals qualities are given as space-separated integers (not ASCII)

### Alignment:

-v <int> report end-to-end hits w/ <=v mismatches; ignore qualities  
or  
-n/--seedmms <int> max mismatches in seed (can be 0-3, default: -n 2)  
-e/--maqerr <int> max sum of mismatch quals across alignment for -n (def: 70)  
-l/--seedlen <int> seed length for -n (default: 28)  
--nomaground disable Maq-like quality rounding for -n (nearest 10 <= 30)  
-I/--minins <int> minimum insert size for paired-end alignment (default: 0)  
-X/--maxins <int> maximum insert size for paired-end alignment (default: 250)  
--fr/--rf/--ff -l, -2 mates align fw/rev, rev/fw, fw/fw (default: --fr)  
--nofw/--norc do not align to forward/reverse-complement reference strand  
--maxbts <int> max # backtracks for -n 2/3 (default: 125, 800 for --best)  
--pairtries <int> max # attempts to find mate for anchor hit (default: 100)  
-y/--tryhard try hard to find valid alignments, at the expense of speed  
--chunkmbs <int> max megabytes of RAM for best-first search frames (def: 64)

# Alineadores o mapeadores de secuencias

## Bowtie

### Reporting:

<code>-k &lt;int&gt;</code>	report up to <int> good alignments per read (default: 1)
<code>-a/--all</code>	report all alignments per read (much slower than low -k)
<code>-m &lt;int&gt;</code>	suppress all alignments if > <int> exist (def: no limit)
<code>-M &lt;int&gt;</code>	like -m, but reports 1 random hit (MAPQ=0); requires --best
<code>--best</code>	hits guaranteed best stratum; ties broken by quality
<code>--strata</code>	hits in sub-optimal strata aren't reported (requires --best)

### Output:

<code>-t/--time</code>	print wall-clock time taken by search phases
<code>-B/--offbase &lt;int&gt;</code>	leftmost ref offset = <int> in bowtie output (default: 0)
<code>--quiet</code>	print nothing but the alignments
<code>--refout</code>	write alignments to files refXXXXXX.map, 1 map per reference
<code>--refidx</code>	refer to ref. seqs by 0-based index rather than name
<code>--al &lt;fname&gt;</code>	write aligned reads/pairs to file(s) <fname>
<code>--un &lt;fname&gt;</code>	write unaligned reads/pairs to file(s) <fname>
<code>--max &lt;fname&gt;</code>	write reads/pairs over -m limit to file(s) <fname>
<code>--suppress &lt;cols&gt;</code>	suppresses given columns (comma-delim'd) in default output
<code>--fullref</code>	write entire ref name (default: only up to 1st space)

### Colorspace:

<code>--snpphred &lt;int&gt;</code>	Phred penalty for SNP when decoding colorspace (def: 30)
or	
<code>--snpfrac &lt;dec&gt;</code>	approx. fraction of SNP bases (e.g. 0.001); sets --snpphred
<code>--col-cseq</code>	print aligned colorspace seqs as colors, not decoded bases
<code>--col-cqual</code>	print  original colorspace quals, not decoded quals
<code>--col-keepends</code>	keep nucleotides at extreme ends of decoded alignment

### SAM:

<code>-S/--sam</code>	write hits in SAM format
<code>--mapq &lt;int&gt;</code>	default mapping quality (MAPQ) to print for SAM alignments
<code>--sam-nohead</code>	suppress header lines (starting with @) for SAM output
<code>--sam-nosq</code>	suppress @SQ header lines for SAM output
<code>--sam-RG &lt;text&gt;</code>	add <text> (usually "lab=value") to @RG line of SAM header

### Performance:

<code>-o/--offrate &lt;int&gt;</code>	override offrate of index; must be >= index's offrate
<code>-p/--threads &lt;int&gt;</code>	number of alignment threads to launch (default: 1)
<code>--mm</code>	use memory-mapped I/O for index; many 'bowtie's can share
<code>--shmem</code>	use shared mem for index; many 'bowtie's can share

### Other:

<code>--seed &lt;int&gt;</code>	seed for random number generator
<code>--verbose</code>	verbose output (for debugging)
<code>--version</code>	print version information and quit
<code>-h/--help</code>	print this usage message



## 1. Cabecera

Fuente de datos, seq de referencia, metodo de alineamiento, etc

Depende del alineador.

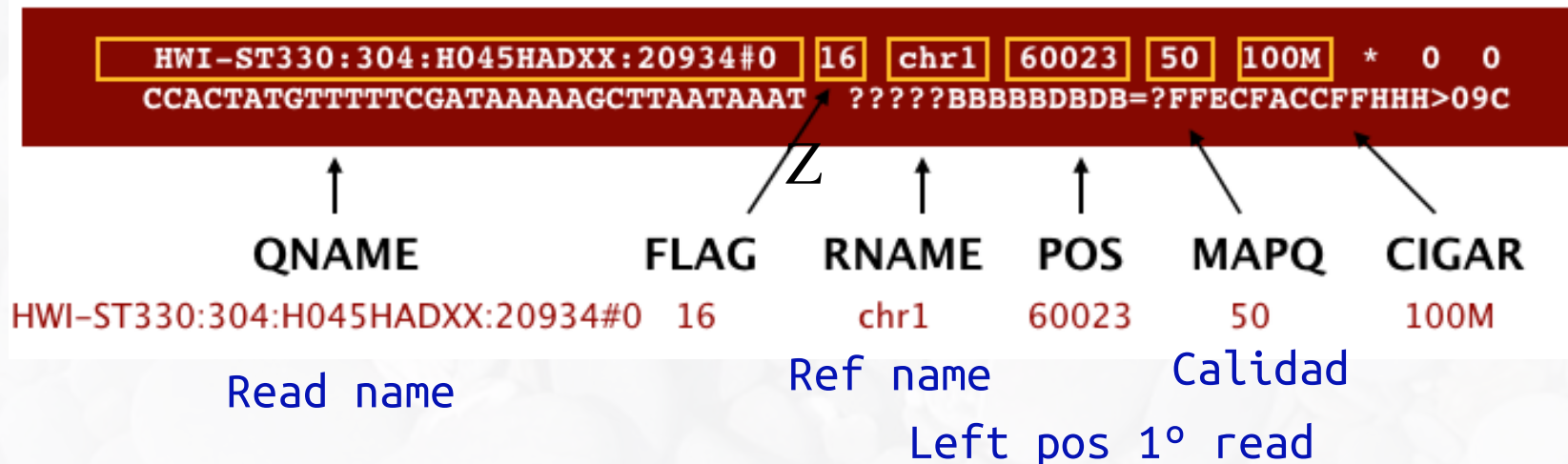
Cada Sección comienza con '@' seguido del codigo de 2 letras

```
@HD The header line
VN: format version
SO: Sorting order of alignments
```

```
@SQ Reference sequence dictionary
SN: reference sequence name
LN: reference sequence length
SP: species
```

```
@PG Program
PN: program name
VN: program version
```

## 2. Alineamiento: cada linea posee 11 campos obligatorios



## FLAG

Para un determinado alinhamento  
Um flag pode estar  
ativado/desativado  
Indicando que la condição es  
verdadeira/falsa

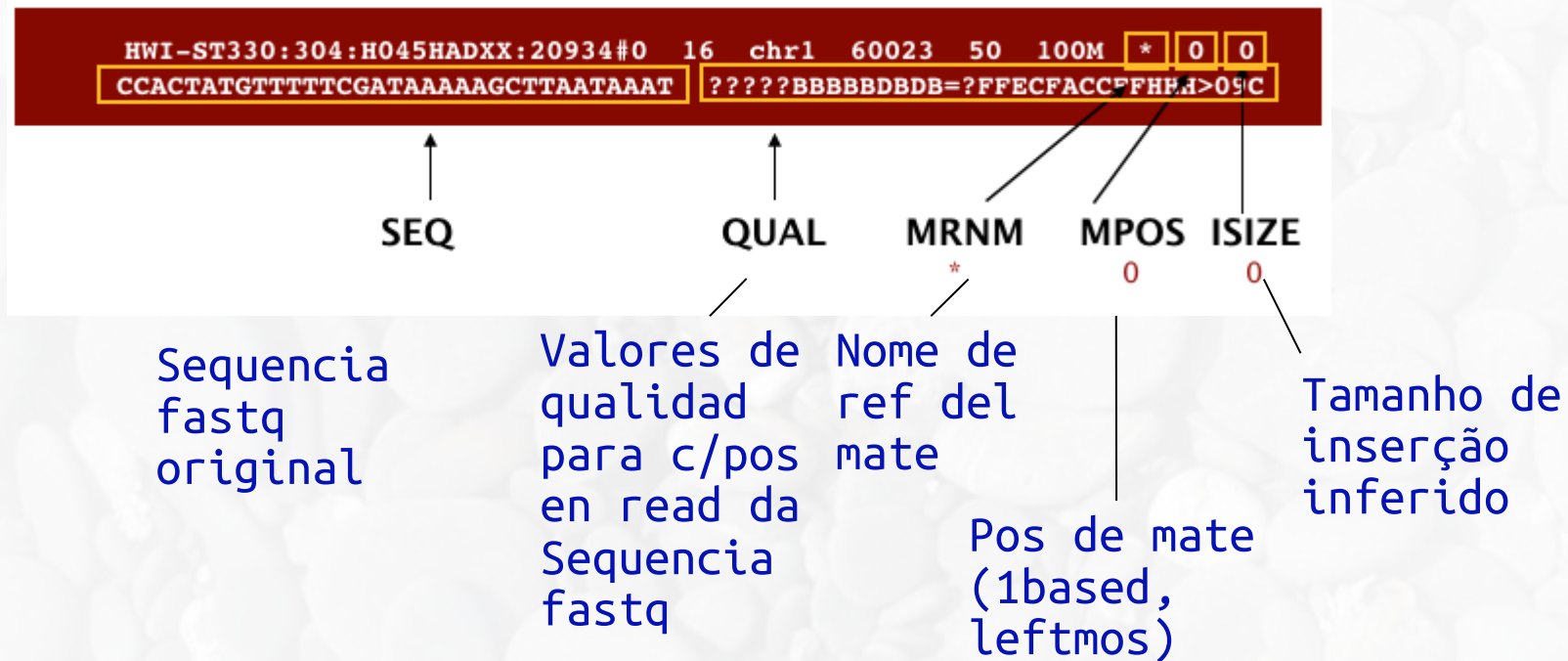
## CIGAR

Es una secuencia de letras y  
números que representam las  
ediciones o operaciones  
necessárias para corresponder la  
leitura à referência.

Flag	Description
1	read is mapped
2	read is mapped as part of a pair
4	read is unmapped
8	mate is unmapped
16	read reverse strand
32	mate reverse strand
64	first in pair
128	second in pair
256	not primary alignment
512	read fails platform/vendor quality checks
1024	read is PCR or optical duplicate



## 2. Alinhamento: cada linha possui 11 campos obrigatorios



# Alinhadores o mapeadores de sequencias

## Bowtie Exemplo de mapeamento

```
Bowtie -p 2 -l 20 -v 2 coli_k12.fna 1 sample1.fastq -2 sample2.fastq
```