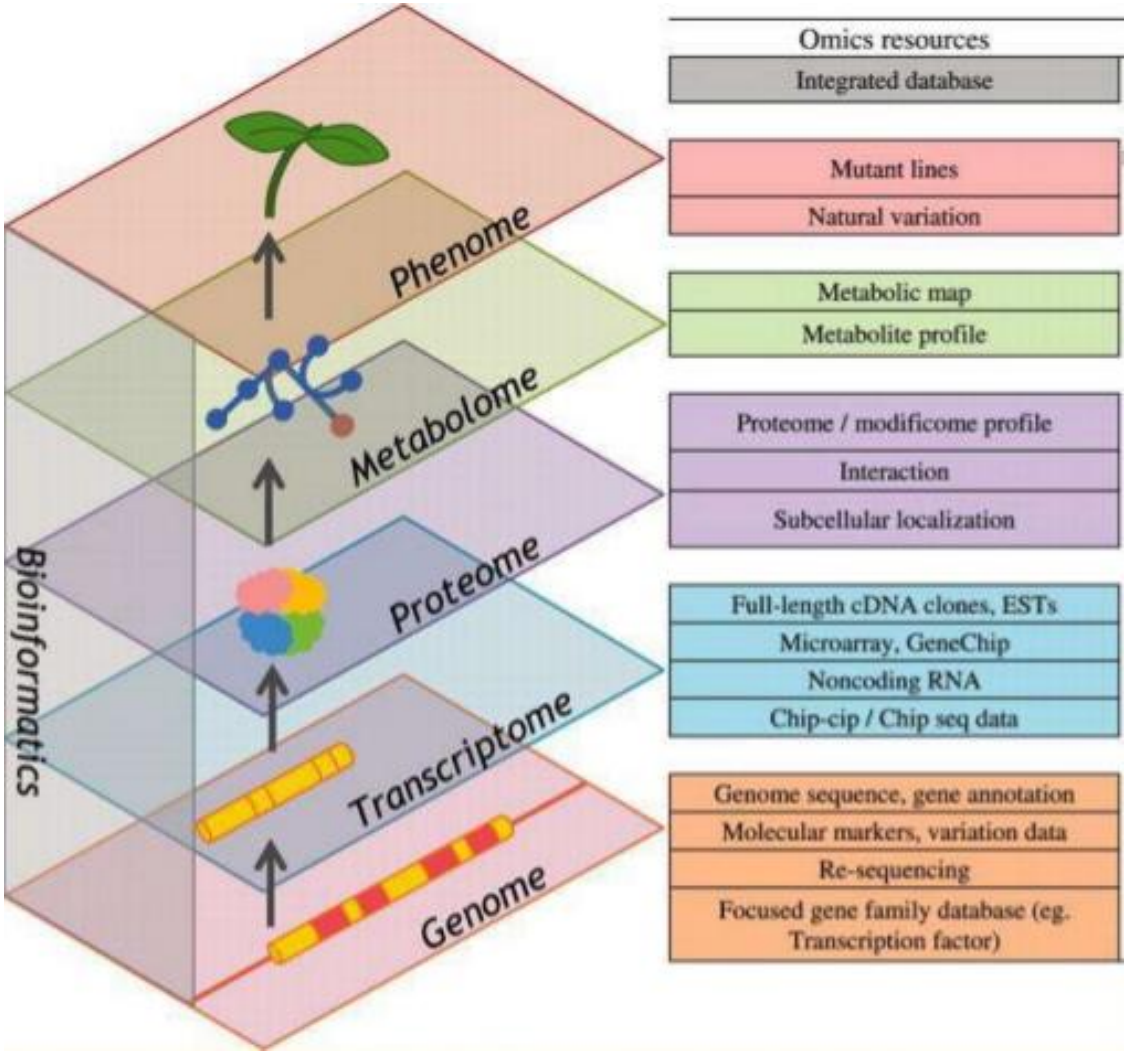
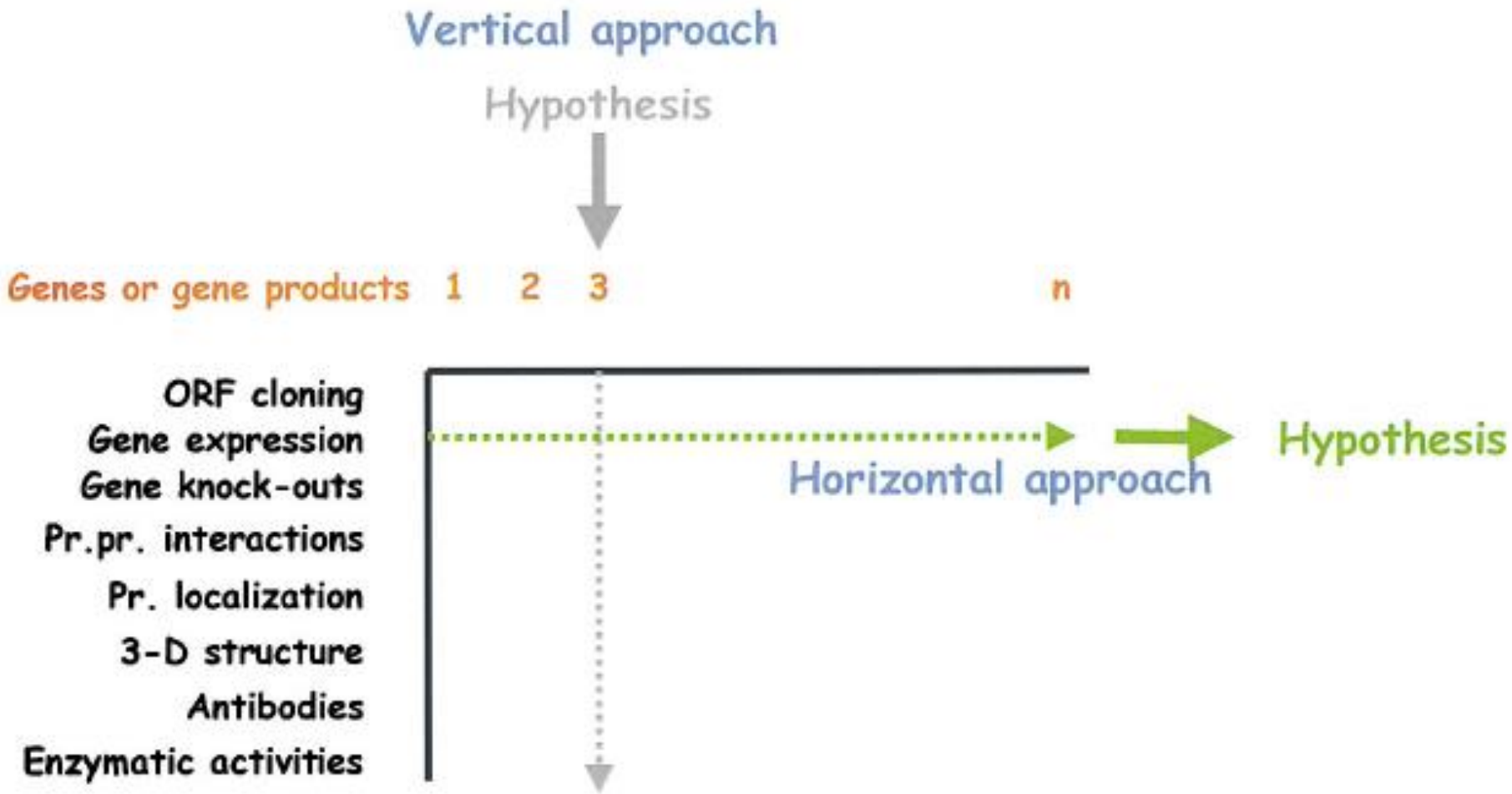
Ciencias “ómicas”

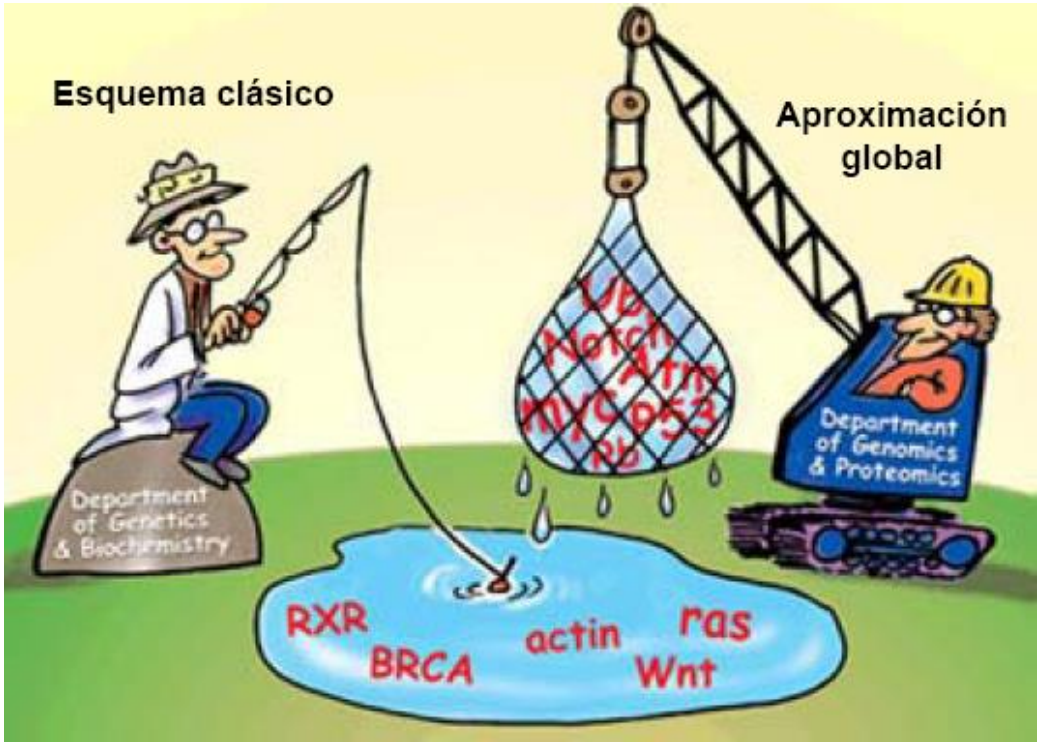


Ciencias “ómicas”

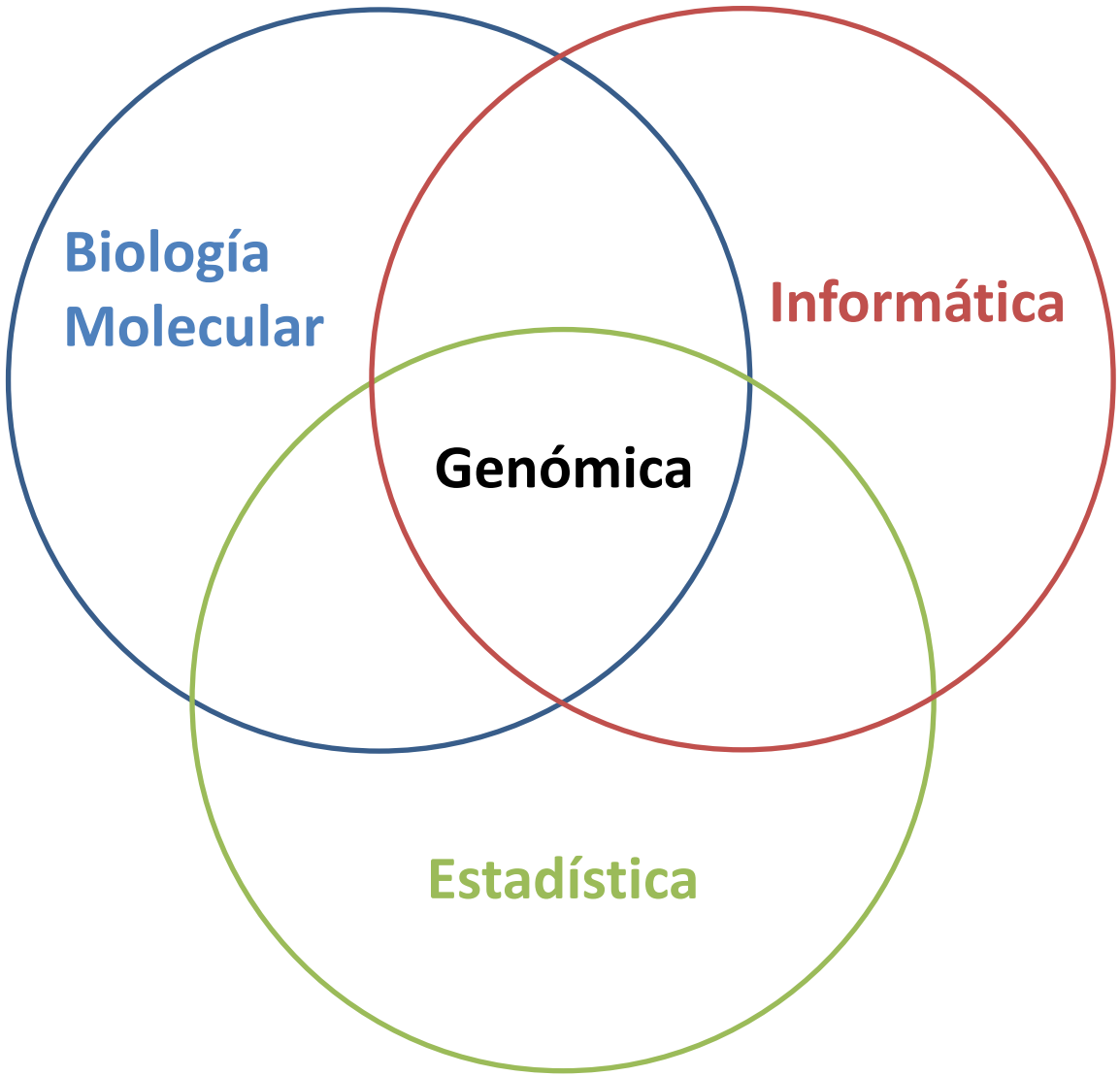


Genómica

- El estudio de la organización, función y evolución de los genomas
  - Genómica funcional
  - Genómica evolutiva
  - Genómica comparativa
  - Genómica estructural
  - …

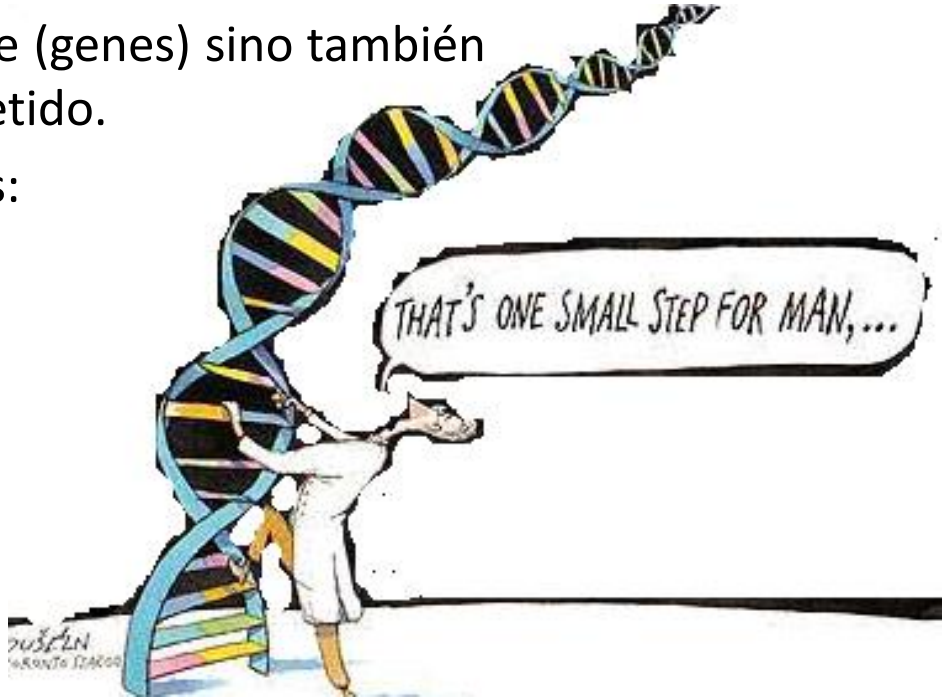


Genómica

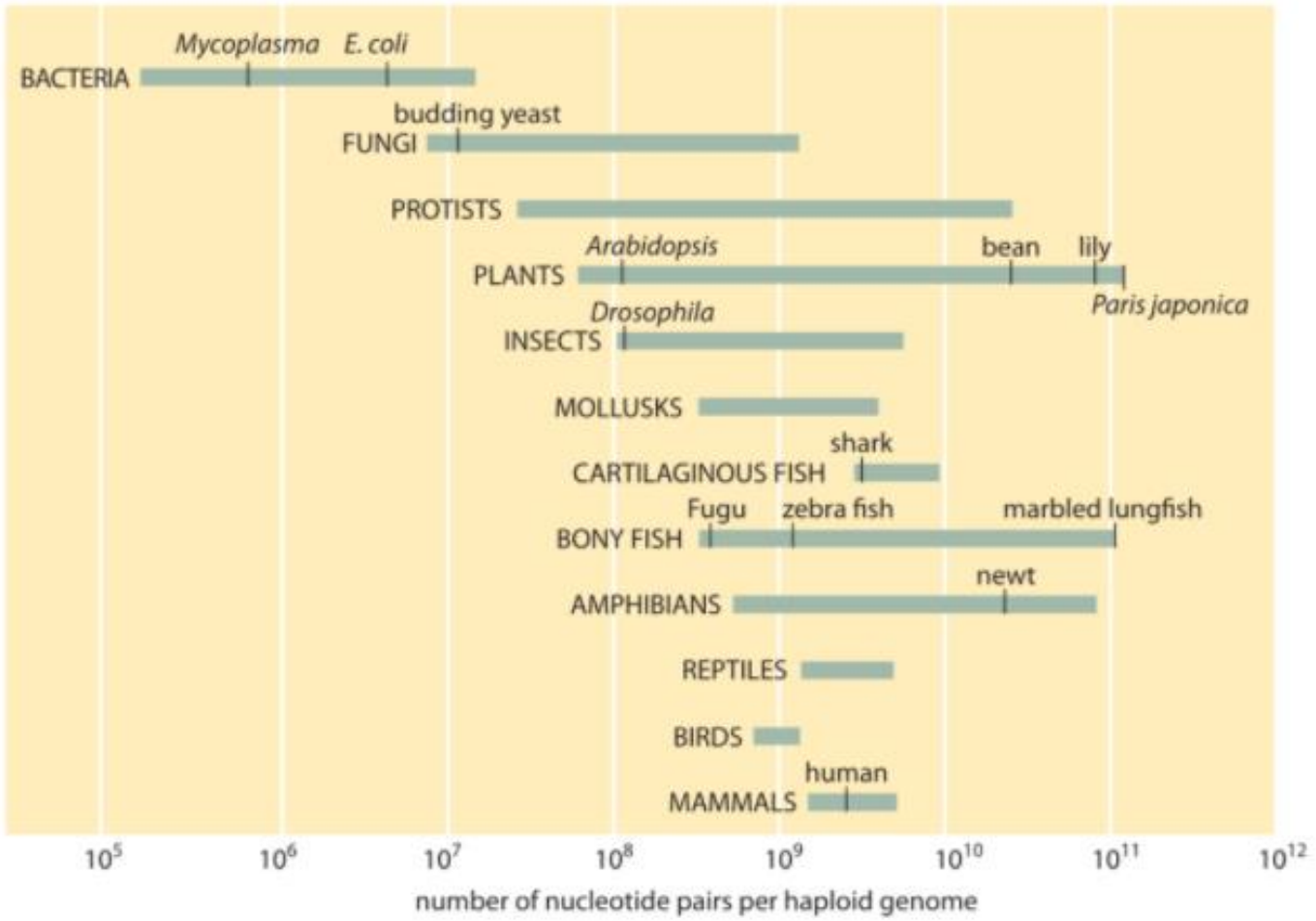


Genoma

- El genoma se puede definir como el contenido total de ADN de la célula.
- Incluye no sólo la porción codificante (genes) sino también las regiones intergénicas y ADN repetido.
- Los eucariotas tienen 2 o 3 genomas:
  - Genoma nuclear
  - Genoma mitocondrial
  - Genoma plastido
- Generalmente “genoma” se utiliza para referirse al genoma nuclear.



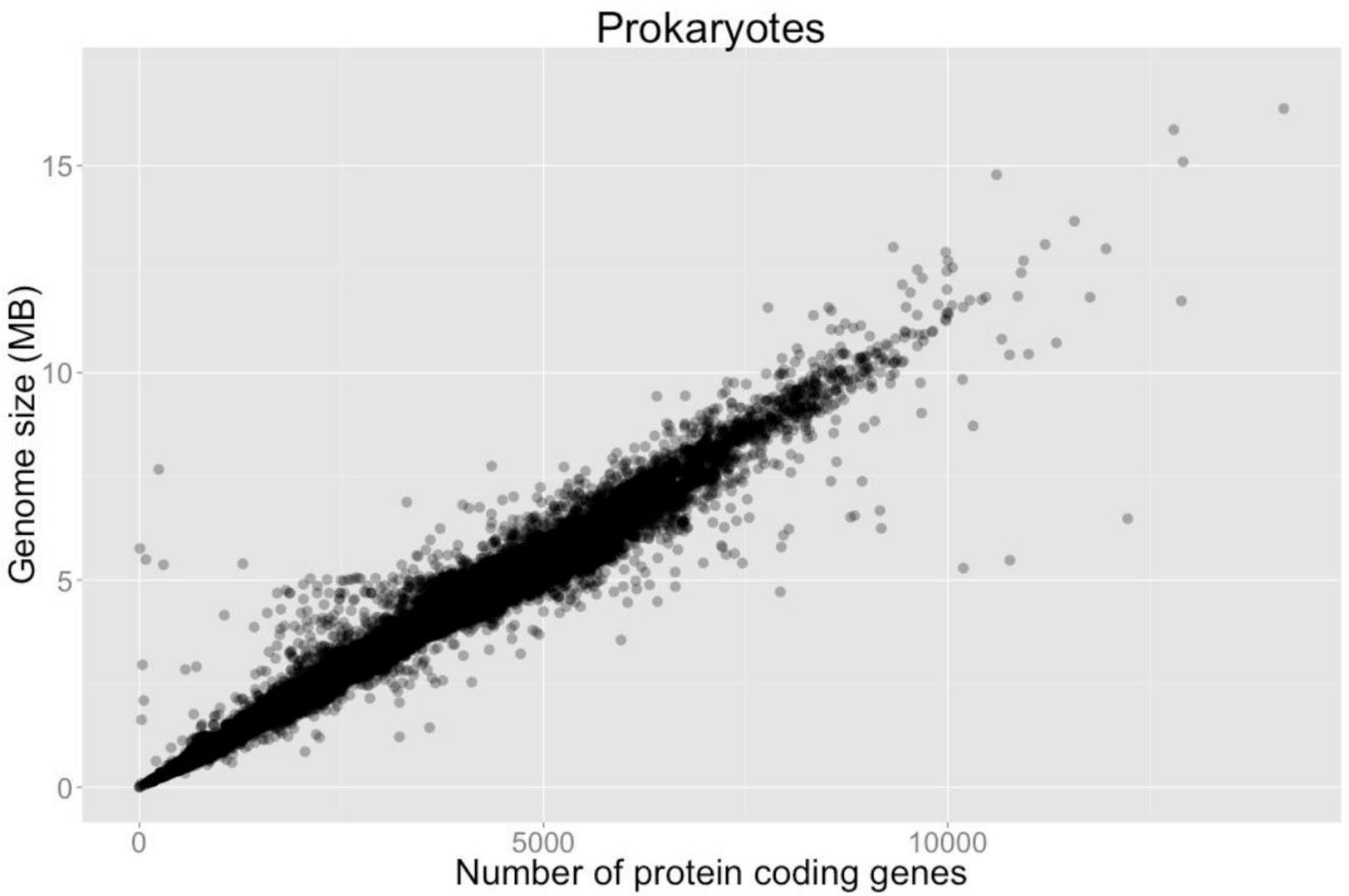
Tamaño de los genomas



- Paradoja del Valor C

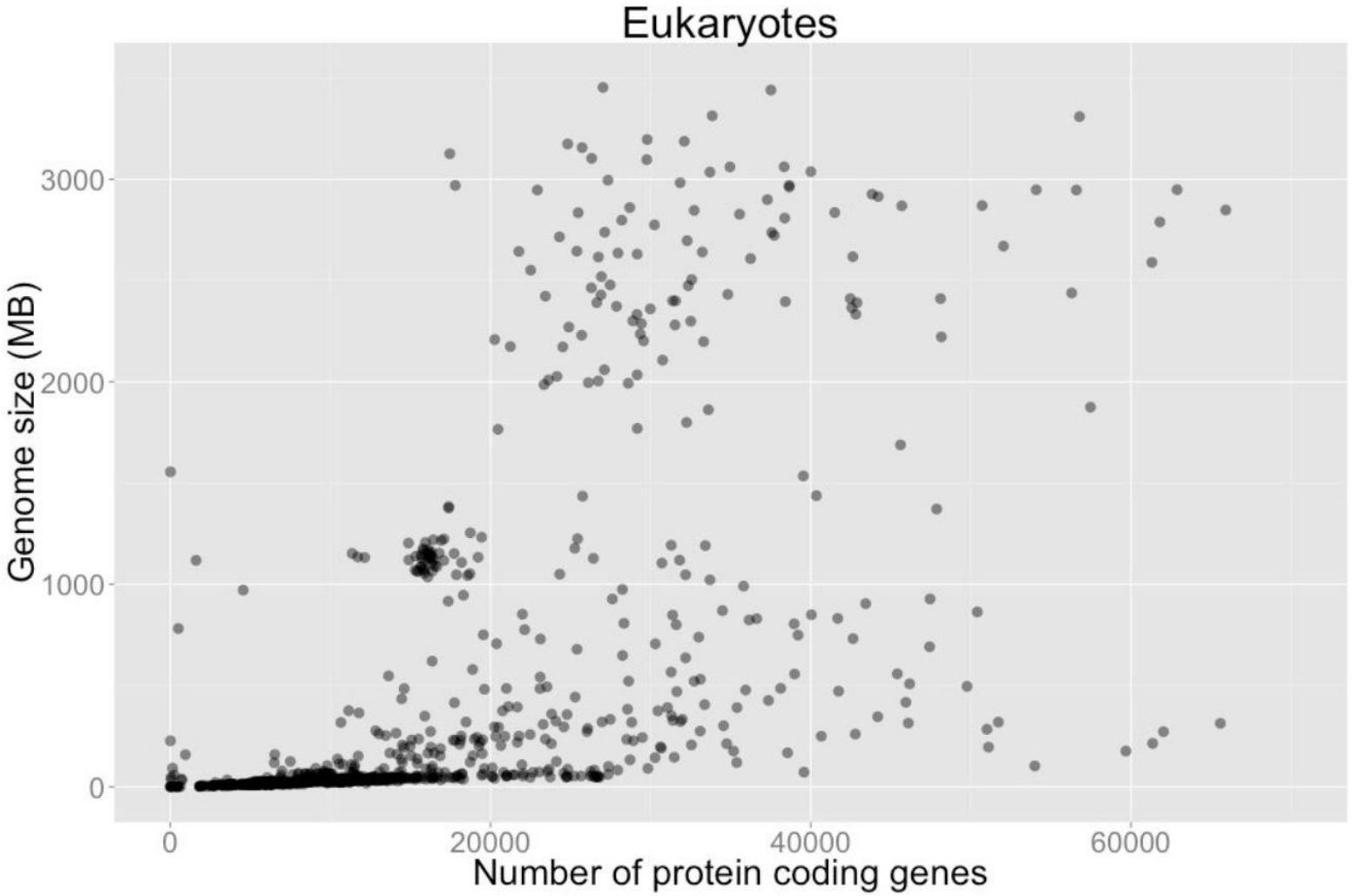
  - El tamaño del genoma no esta directamente relacionado con la complejidad del organismo

Tamaño de los genomas y número de genes



Tamaño de los genomas y número de genes

- No existe correlación entre el tamaño del genoma y la cantidad de genes

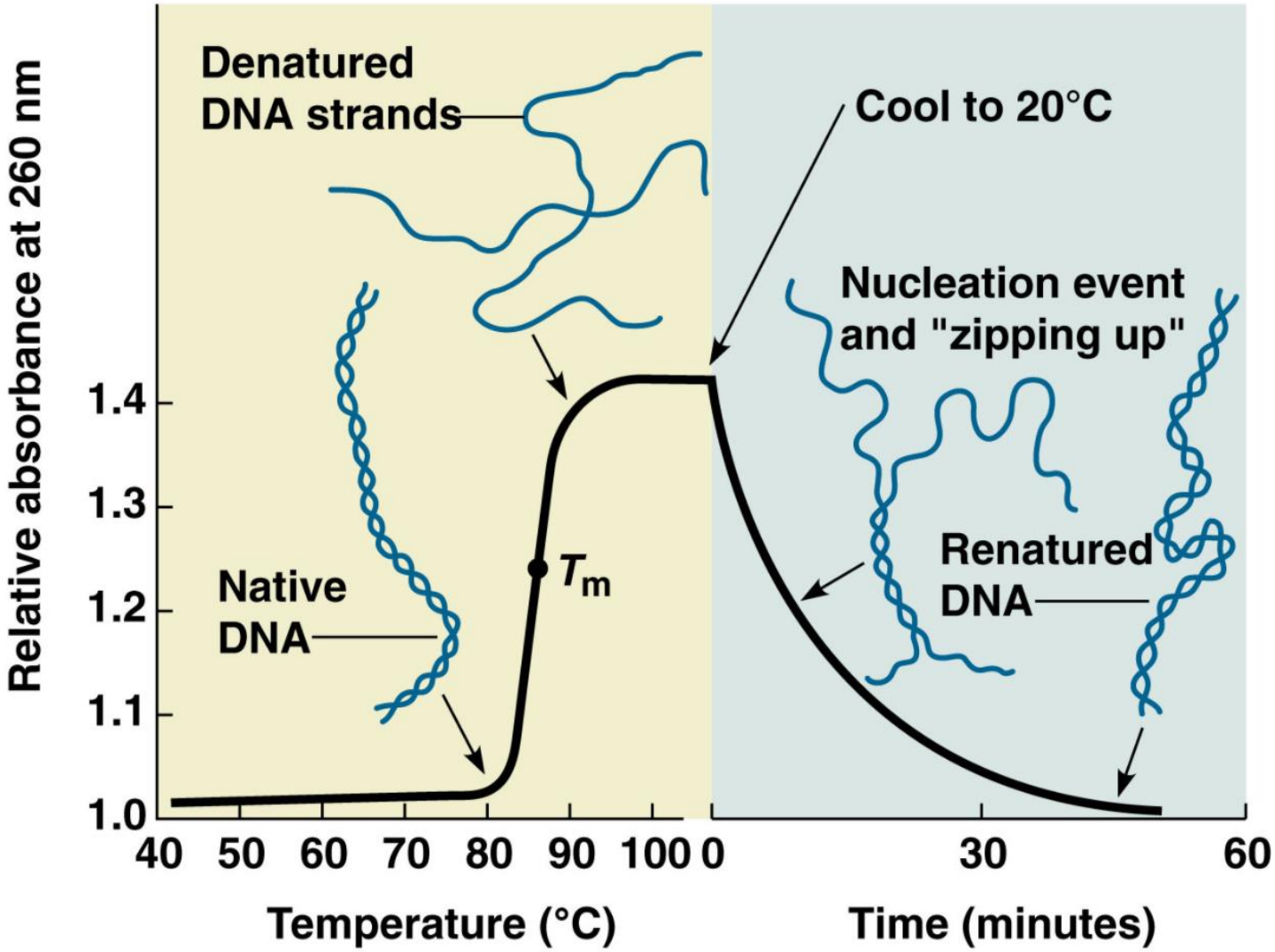


Experimentos de reasociación de ADN

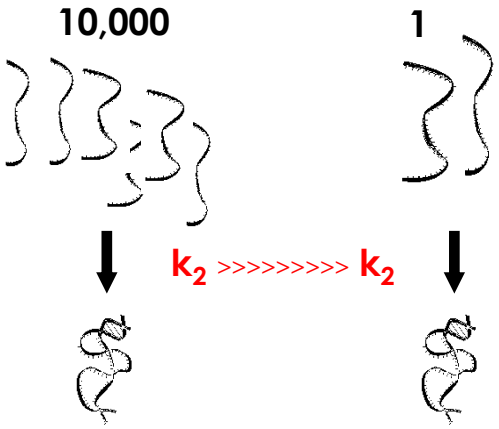
- Experimentos de Britten & Kohne:
  - Fragmentación de ADN
  - Desnaturalización
  - Medida del tiempo de reasociación

- Permiten determinar la complejidad del genoma

Experimentos de reasociación de ADN



Experimentos de reasociación de ADN



- Secuencias repetidas reasocian rápidamente.
- Secuencias únicas reasocian lentamente.

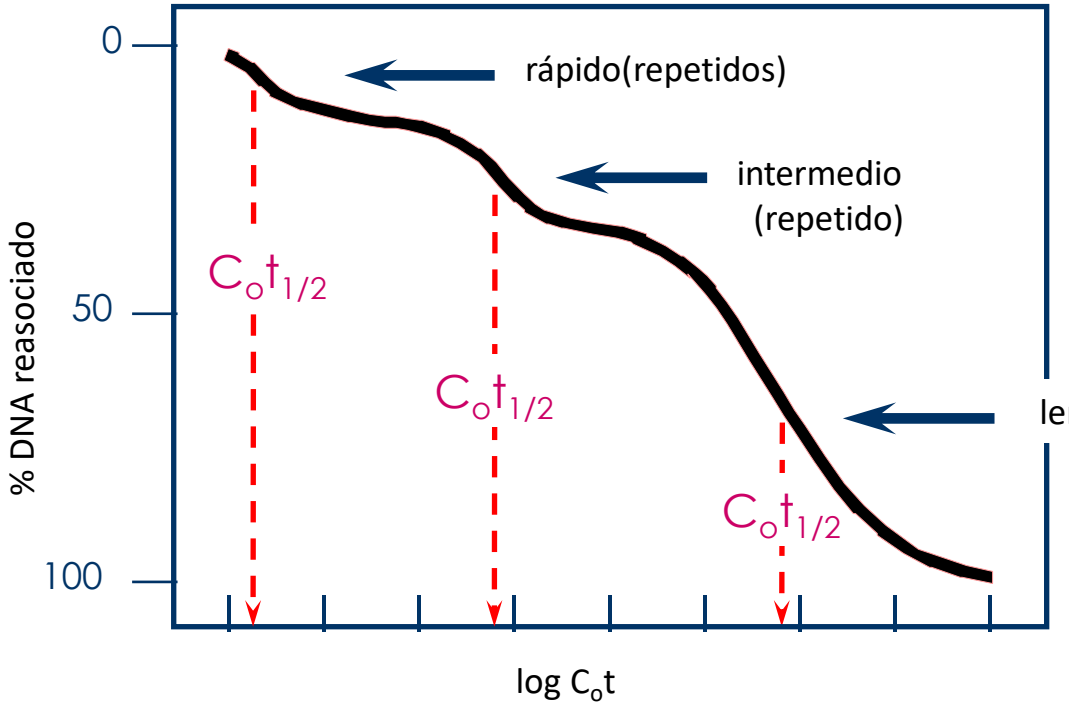
Cinética de reasociación del ADN genómico humano

$C_o t_{1/2} = 1 / k_2$

k_2 = constante de segundo orden

C_o = concentración de ADN

$t_{1/2}$ = tiempo medio de reacción



Fracciones obtenidas:

- reasociación rápida
- reasociación intermedia
- reasociación lenta

ADN repetido y copia única

Altamente Repetido (10-15%)

Medianamente Repetido (15-40%)

Copia Unica (50-75%)

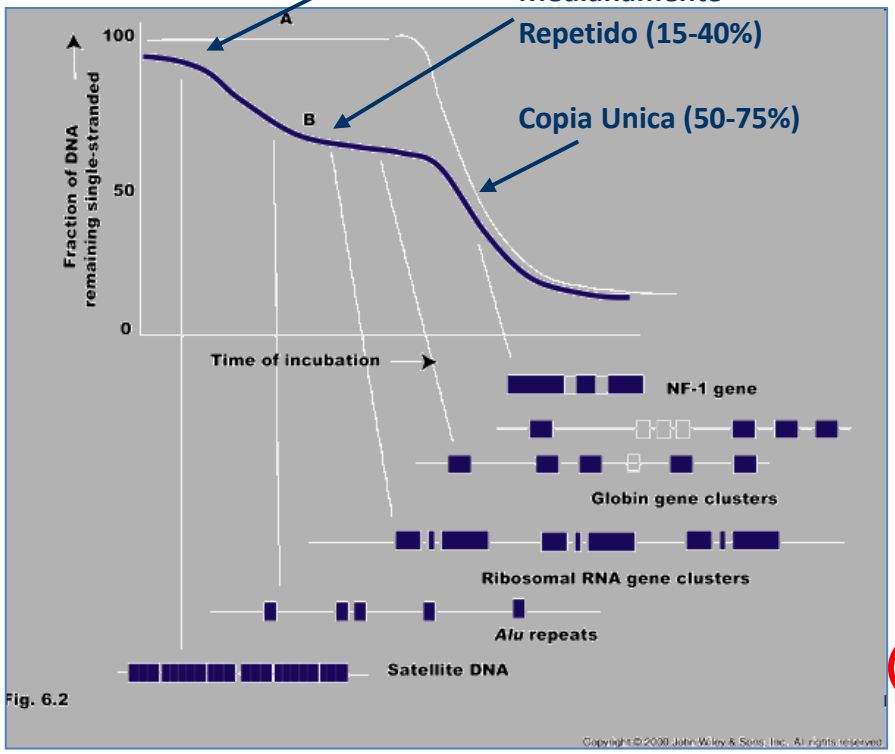
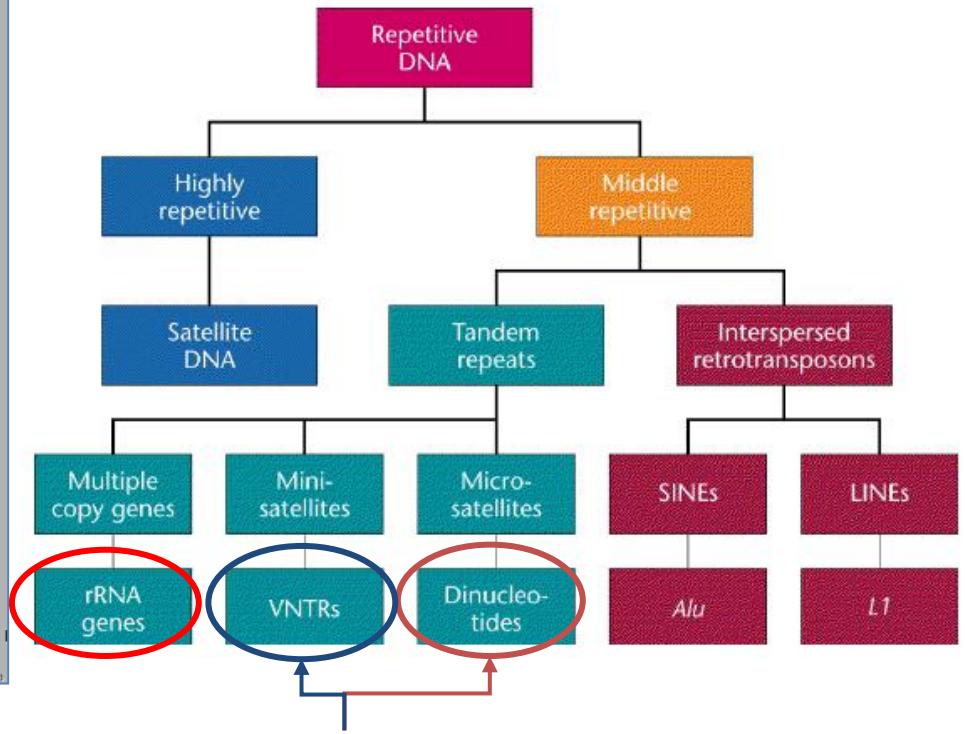


Fig. 6.2



Unico que contiene genes

Usados en identificación

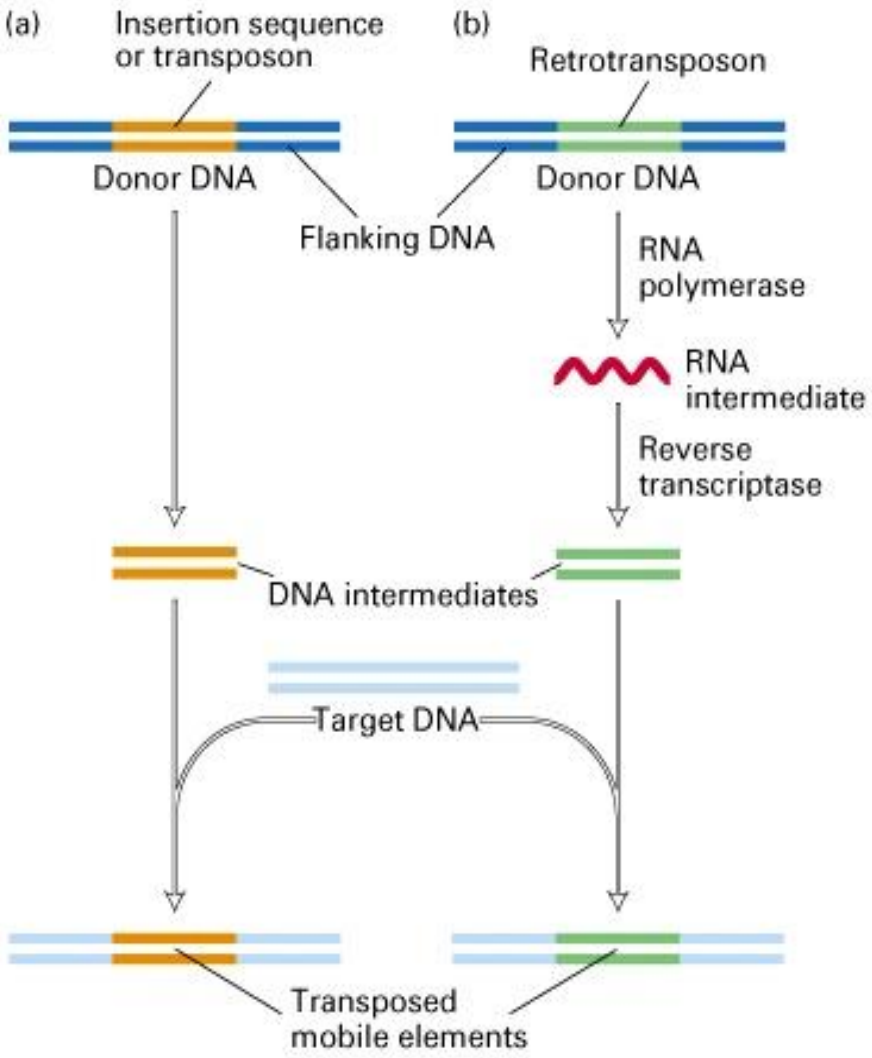
Hay cinclo clases pincipales de elementos repetidos

- Repetidos dispersos
- Pseudogenes
- Repetidos de secuencia simple
- ADN Satelite
- Duplicaciones segmentales

Repetidos dispersos

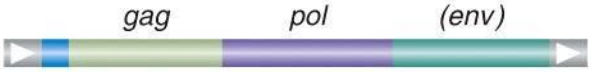
- Son secuencias derivadas de fenómenos de transposición
- En humanos constituyen alrededor de un 45% del genoma
- Se pueden clasificar de acuerdo al mecanismo de transposición
  - Intermediarios de ARN (retroelementos)
    - LTRs (Long-terminal repeat transposons) (RNA-mediated)
    - Long interspersed elements (LINEs); autónomos
    - Short interspersed elements (SINEs)(RNA-mediated); (incluyen la familia Alu)
  - Intermediarios de ADN (transposones de DNA)
    - 3% del genoma

Repetidos dispersos



Repetidos dispersos

Classes of interspersed repeat in the human genome

Element	Transposition	Structure	Length	Copy number	Fraction of genome
LINEs	Autonomous		1–5 kb	20,000–40,000	21%
SINEs	Nonautonomous		100–300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous		6–11 kb	450,000	8%
	Nonautonomous		1.5–3 kb		
DNA transposons	Autonomous		2–3 kb	300,000	3%
	Nonautonomous		80–3000 bp		

Pseudogenes

- Tienen un codón STOP prematuro o un cambio de marco que hacen que no produzcan una proteína no funcional. Se generan en por retrotrasposición o duplicación y perdida de función

- En humanos hay unos 19,000 pseudogenes descritos (un número solo un poco menor al número de genes). 11,000 no-procesados y 8,000 procesados (sin intrones)

Repetidos de secuencia simple

- Microsatélites

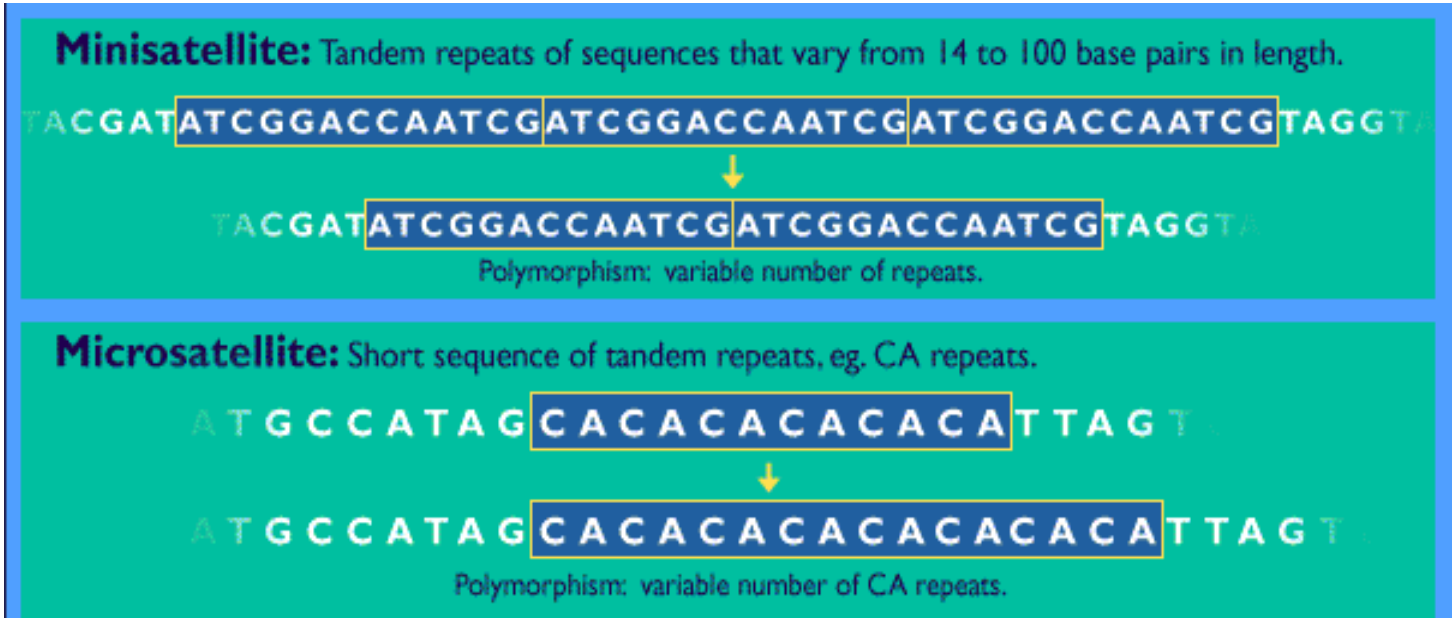
  Son repetidos en tándem

  La unidad repetitiva es de uno a 12-14pb

- Minisatélites: La unidad repetitiva es de mayor tamaño (hasta unas 500pb)

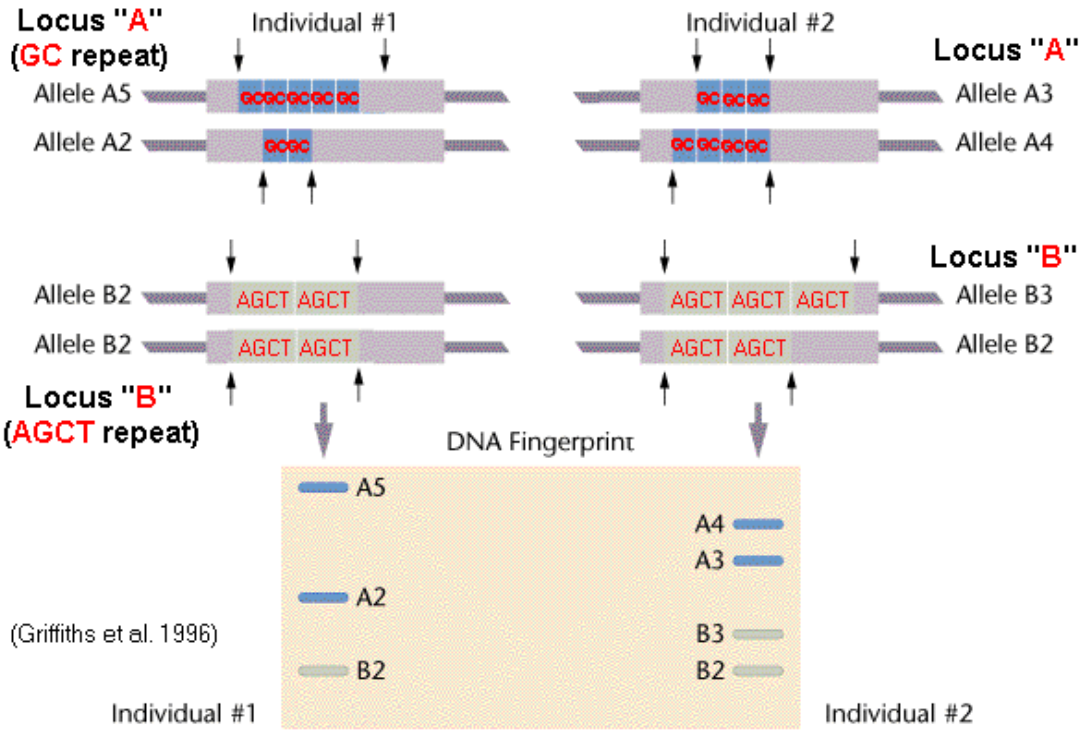
Repetidos de secuencia simple

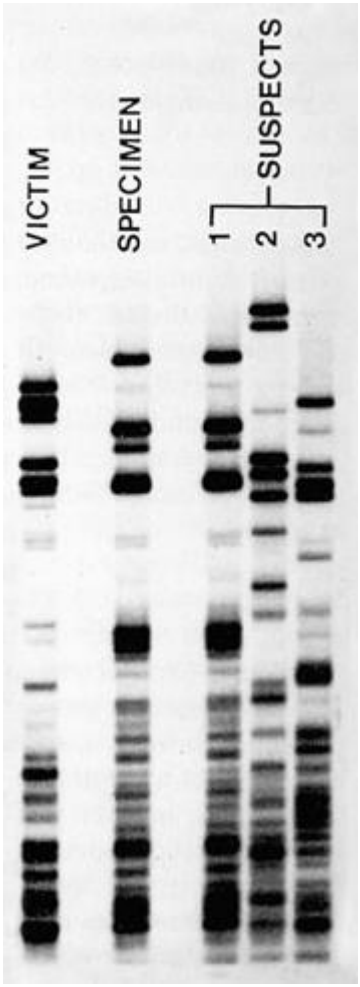
Fingerprinting



- Ambas son útiles para identificar individuos por DNA fingerprinting, puesto que la longitud de la región repetida es muy variable entre individuos.

VNTRs e identificación

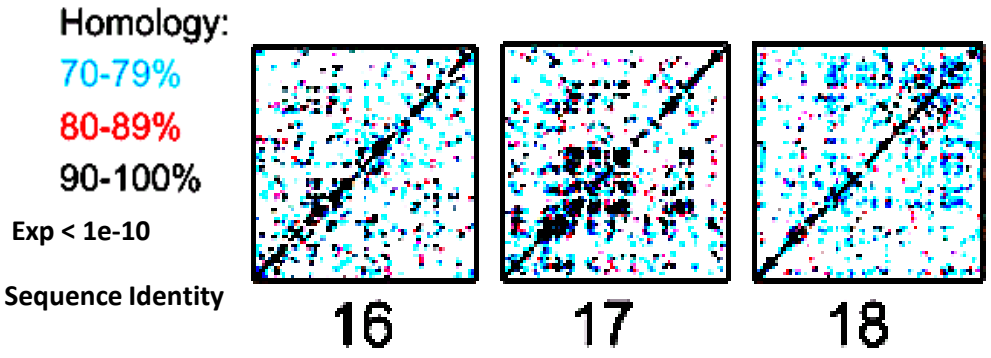




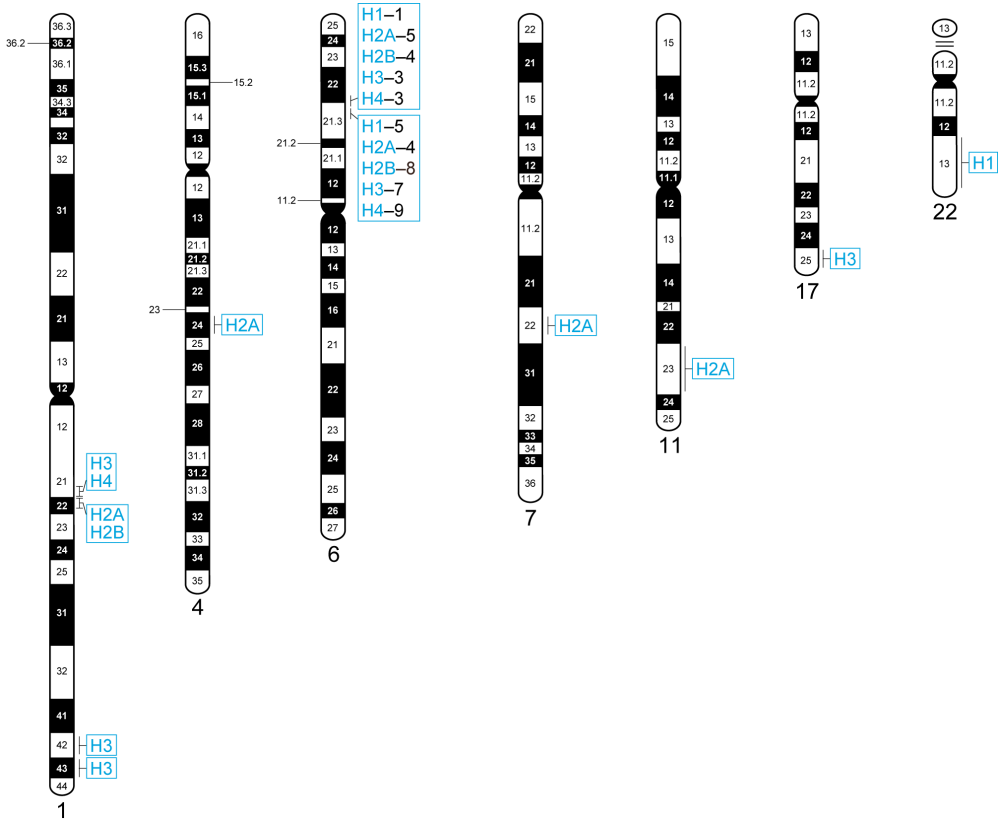
Duplicaciones segmentales

- Regiones grandes (10 a 300kb) que se encuentran dispersas en el genoma con baja copia (LCRs) y poseen un alto grado de identidad de secuencia.
- 5% del genoma humano

- El alto grado de homología de los LCRs los convierten en sustratos de recombinación (recombinación homóloga no alélica) generando reordenamientos cromosómicos (¨enfermedades genómicas¨)



Familias génicas

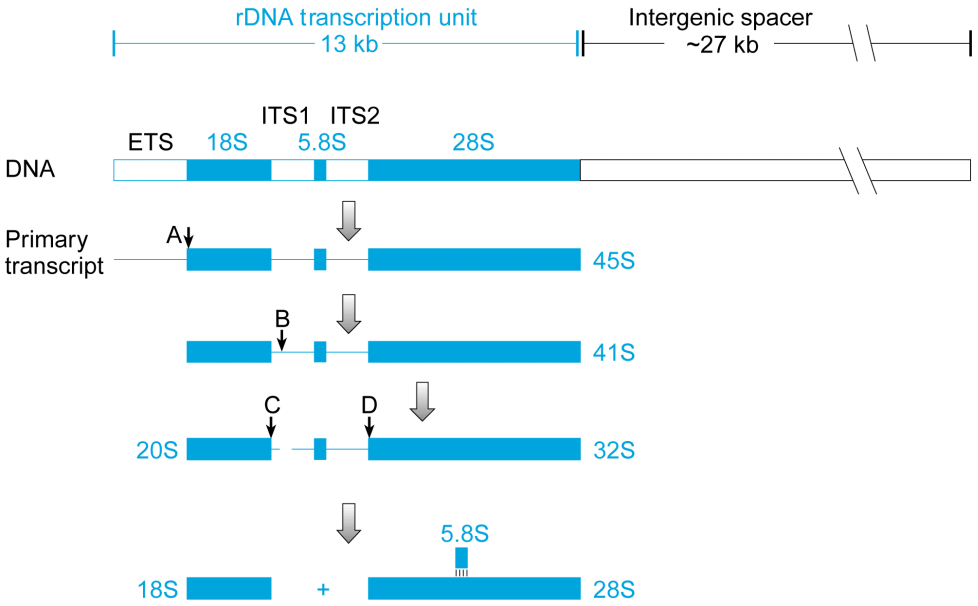


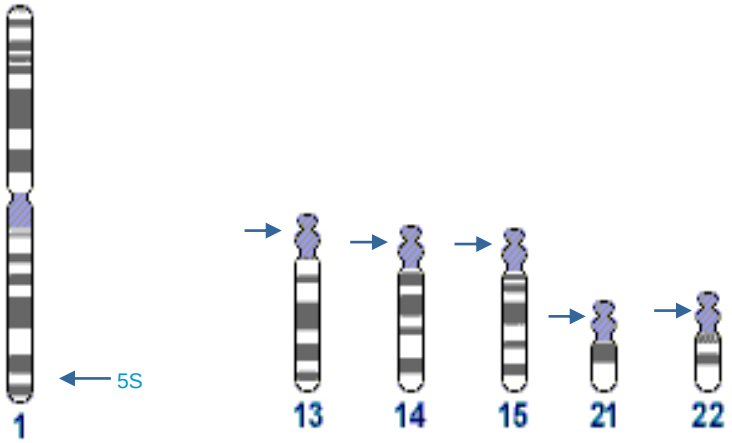
HISTONAS

- 11 clusters contienen los 60 genes de histonas.
- 2 clusters mayores en el cromosoma 6 y clusters menores incluyendo solo uno o dos subtipos.

rRNA

- 1 única unidad transcripcional repetida unas 50 veces en 5 clusters ubicados en los cromosomas 13, 14, 15, 21 y 22.
- El rRNA 5S codificado por varios cientos de genes en 3 clusters en el cromosoma 1.

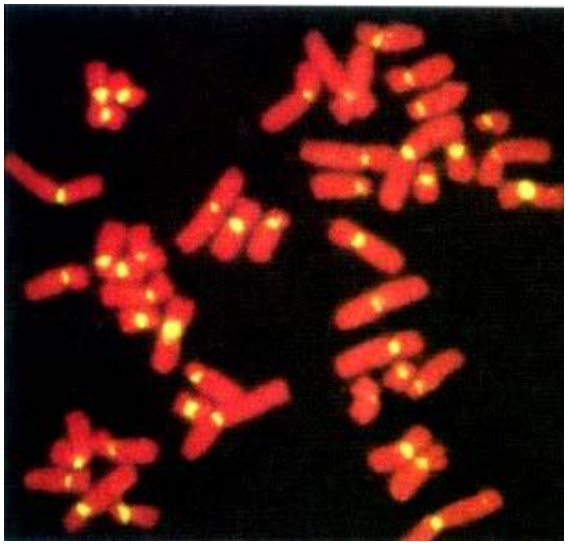


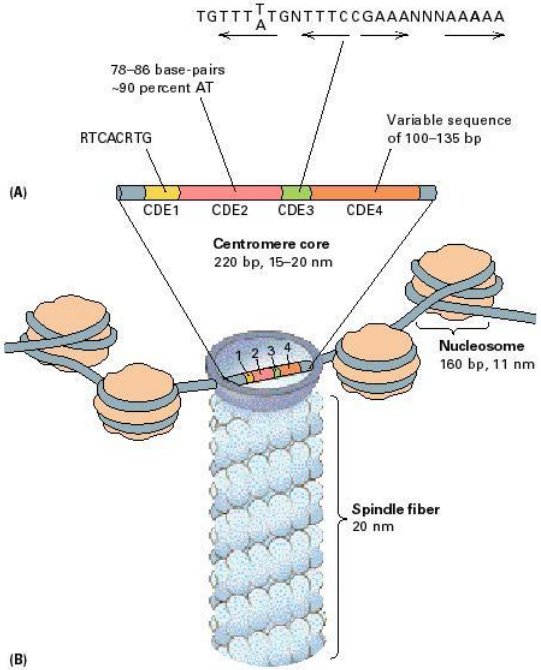


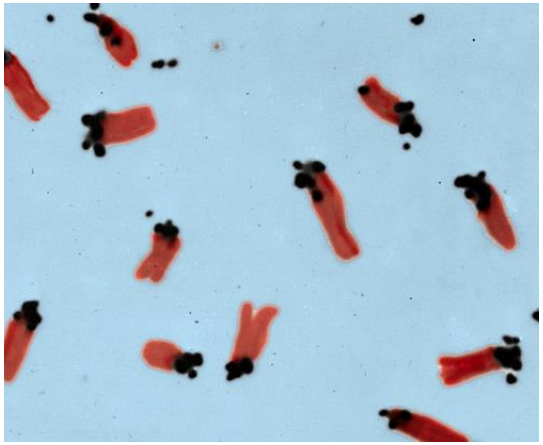
Bloques de repetidos en tandem (ADN satellite)

Repetidos teloméricos
Repetidos centroméricos

Son secuencias repetidas cortas del tipo mini o micro satélite
Pueden abarcar millones de pares de bases







Page 660

Identificación de secuencias repetidas

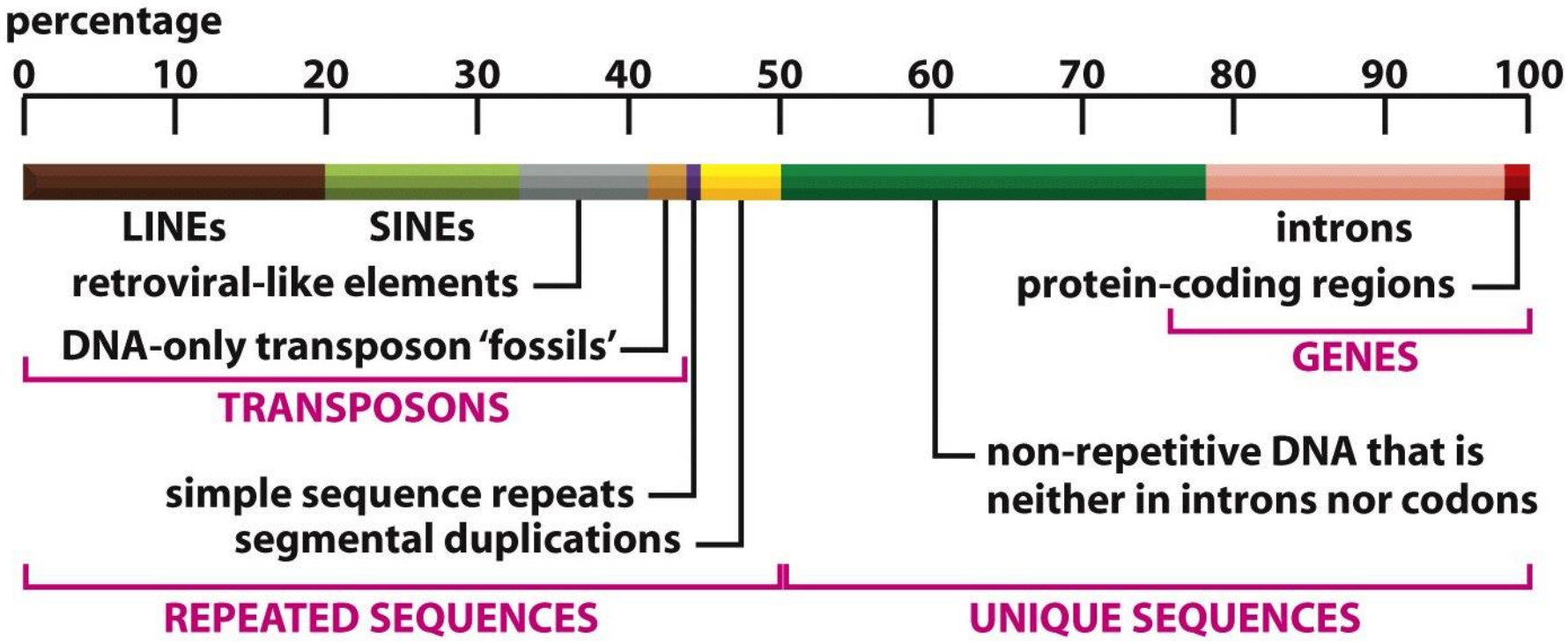
- Predicción *de novo*
  - Repeatexplorer
  - Repeatmodeler
  - REPET
- Homología con elementos previos
  - Repeatmasker
- Características estructurales conservadas
  - LTRharvest

Identificación de secuencias repetidas

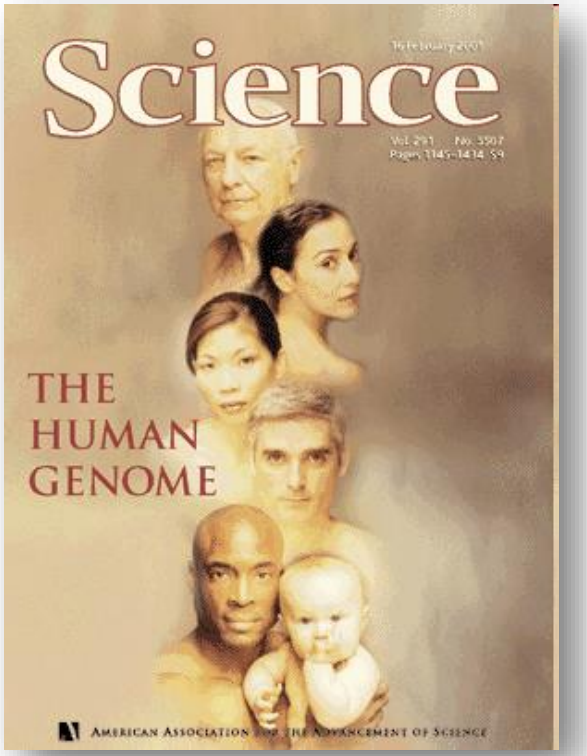
Summary:

```
==================================================
file name: RM2sequpload_1233250641
sequences:             1
total length:       2000 bp  (2000 bp excl N/X-runs)
GC level:          40.95 %
bases masked:        896 bp ( 44.80 %)
==================================================
                number of      length   percentage
                elements*    occupied  of sequence
--------------------------------------------------
SINEs:                  1         311 bp    15.55 %
      ALUs              1         311 bp    15.55 %
      MIRs              0           0 bp     0.00 %

LINEs:                  1         247 bp    12.35 %
      LINE1             1         247 bp    12.35 %
      LINE2             0           0 bp     0.00 %
      L3/CR1            0           0 bp     0.00 %

LTR elements:           1         338 bp    16.90 %
      ERVL              0           0 bp     0.00 %
      ERVL-MaLRs        1         338 bp    16.90 %
      ERV_classI        0           0 bp     0.00 %
      ERV_classII       0           0 bp     0.00 %

DNA elements:           0           0 bp     0.00 %
     hAT-Charlie        0           0 bp     0.00 %
     TcMar-Tigger       0           0 bp     0.00 %

Unclassified:           0           0 bp     0.00 %

Total interspersed repeats:       896 bp    44.80 %


Small RNA:              0           0 bp     0.00 %

Satellites:             0           0 bp     0.00 %
Simple repeats:         0           0 bp     0.00 %
Low complexity:         0           0 bp     0.00 %
--------------------------------------------------
```

Panorama General del Genoma Humano



Como se secuencia un genoma?



Proyecto público

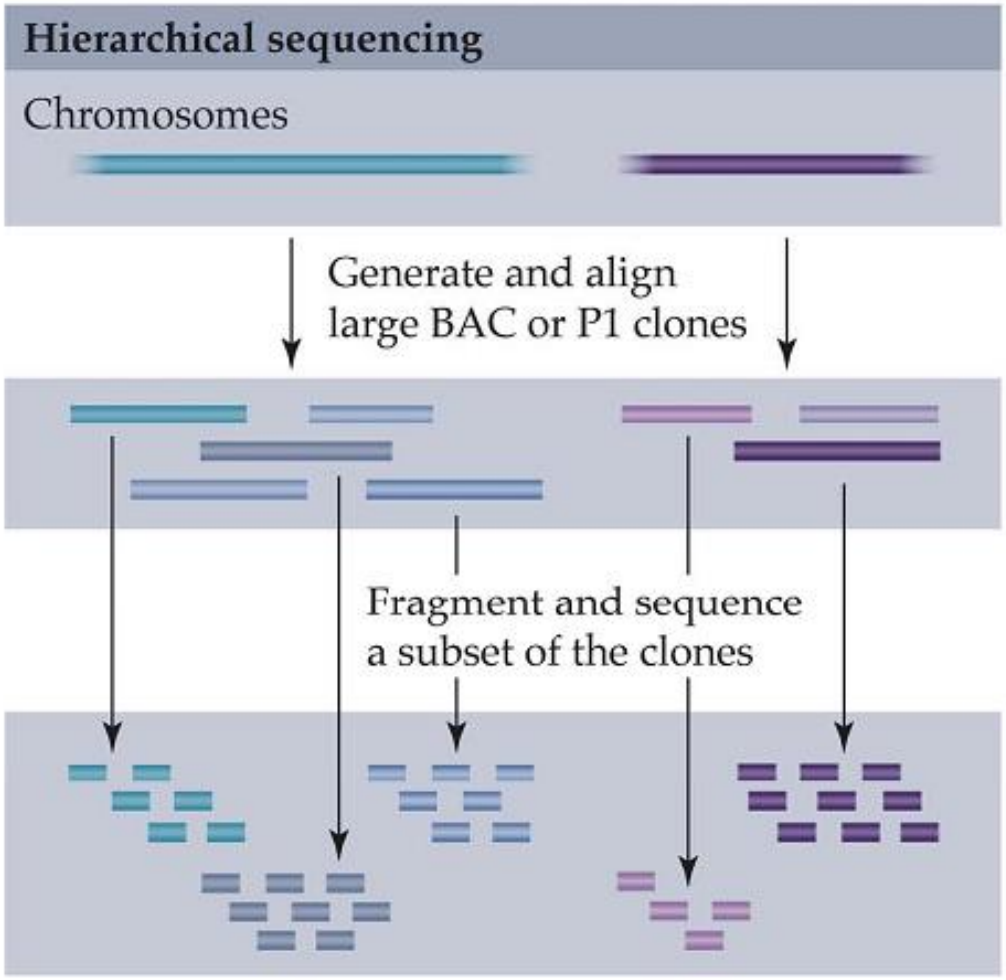
- Consorcio internacional
- NIH y Wellcome Trust
- Comienza en 1989
- Costo estimado $3 billiones en 15 años
- Encabezado por Watson y luego por Francis Collins
- Aprox. 15 individuos anonimos diferentes origenes

Diferentes estrategias

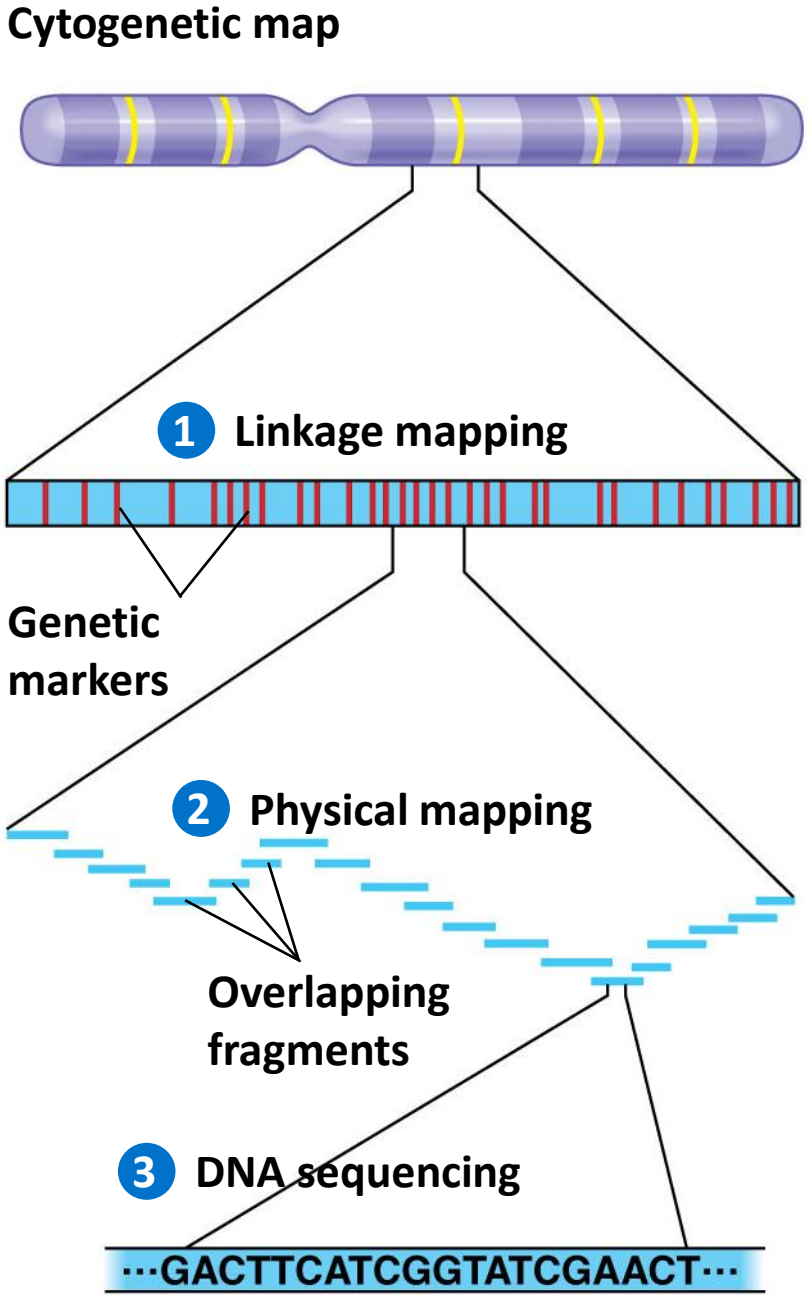
- Genero lecturas solapadas
  - "*primer walking*"
  - caro y lento
- Al azar
  - "*Shotgun*"
  - Muchas secuencias solapadas redundantes
  - Rápido y barato
- Estrategia intermedia
  - "*Clone by clone*"
  - Armo un set de clones ordenados
  - Los secuencio con una estrategia tipo "*shotgun*"

Secuenciación

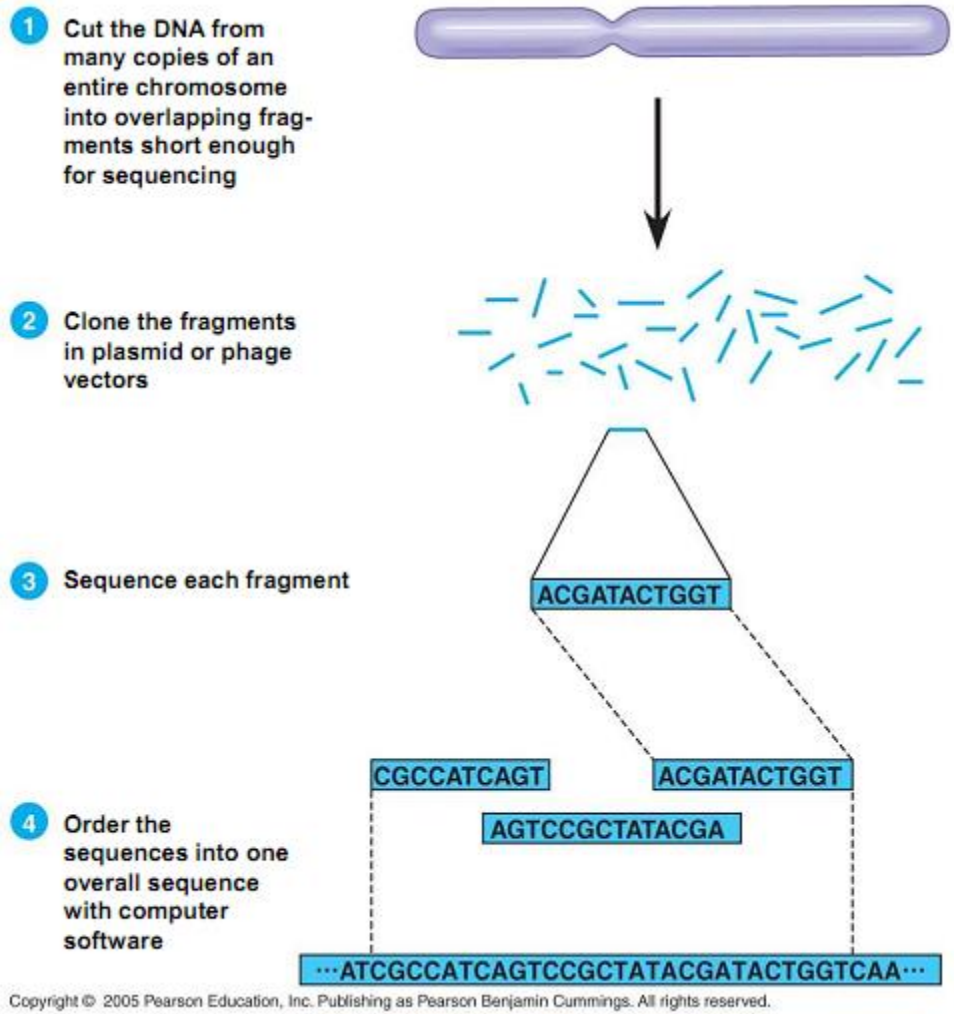
- Las secuencias de ADN son cortas. De unos 800pb con el método clásico de Sanger (Mucho mas cortas con los nuevos métodos)
- Como los genomas son grandes hay que construirlos a partir de secuencias mas pequeñas
- El secuenciado de genomas se basa en juntar secuencias obtenidas de fragmentos solapantes
- Para genomas pequeños la secuenciación “shotgun” funciona bien
- Para genomas grandes tiene algunos problemas…

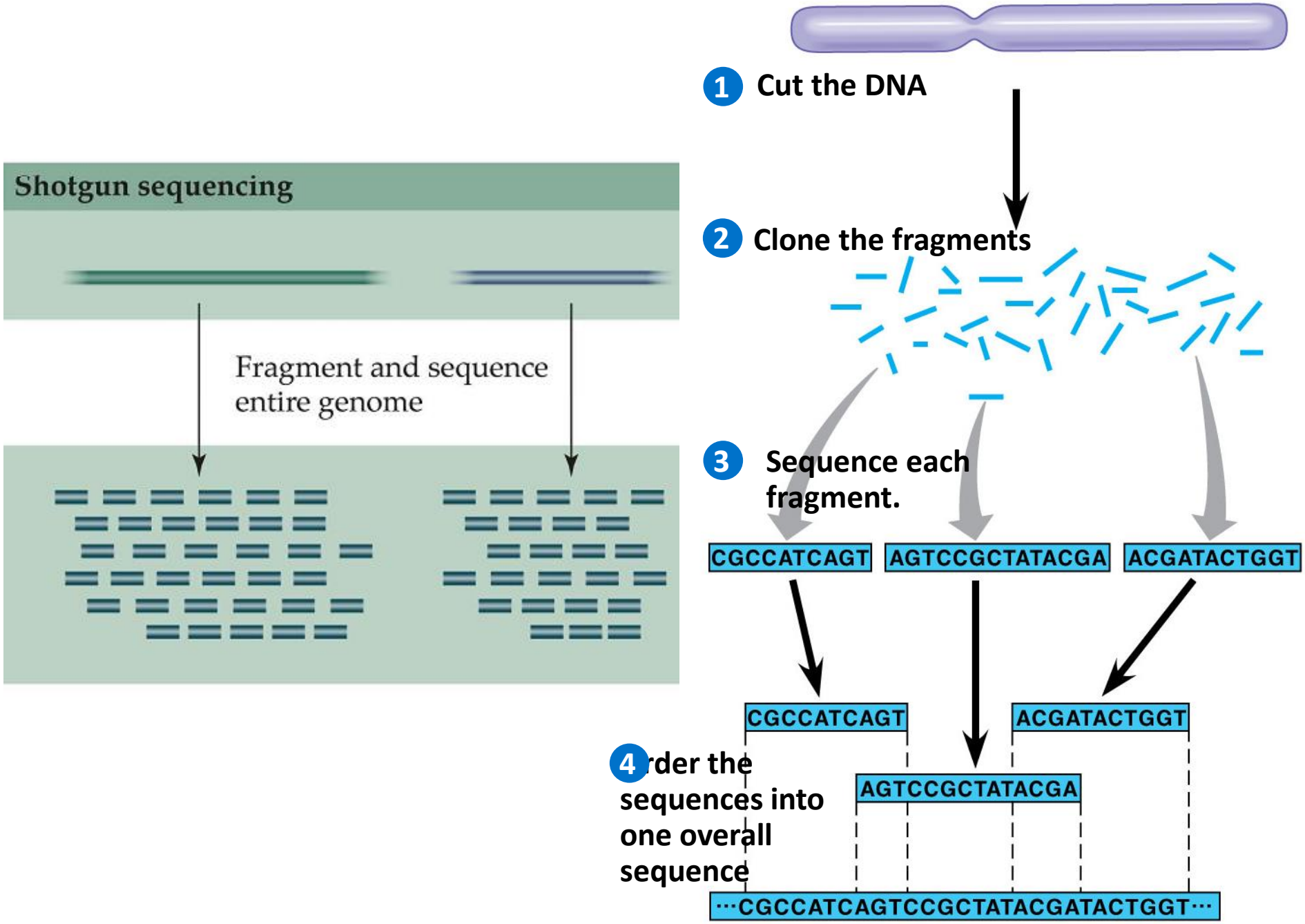


A PRIMER OF GENOME SCIENCE, Second Edition, Figure 2.7 (Part 1) ©

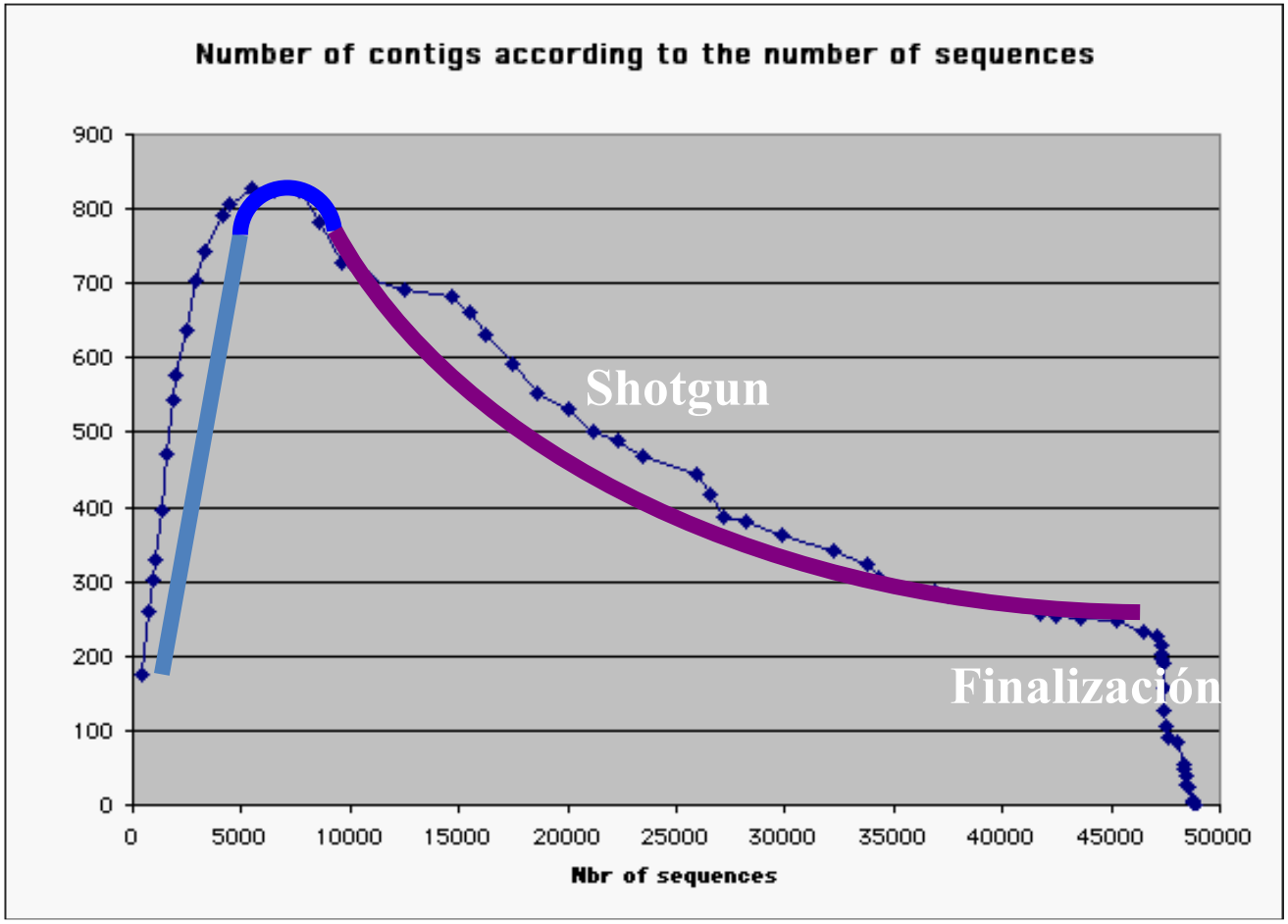


Secuenciación ¨Shotgun¨




Shotgun sequencing
Fragment and sequence entire genome
1 Cut the DNA
2 Clone the fragments
3 Sequence each fragment.
CGCCATCAGT
AGTCCGCTATACGA
ACGATACTGGT
4 Order the sequences into one overall sequence
CGCCATCAGT
AGTCCGCTATACGA
ACGATACTGGT
···CGCCATCAGTCCGCTATACGATACTGGT···

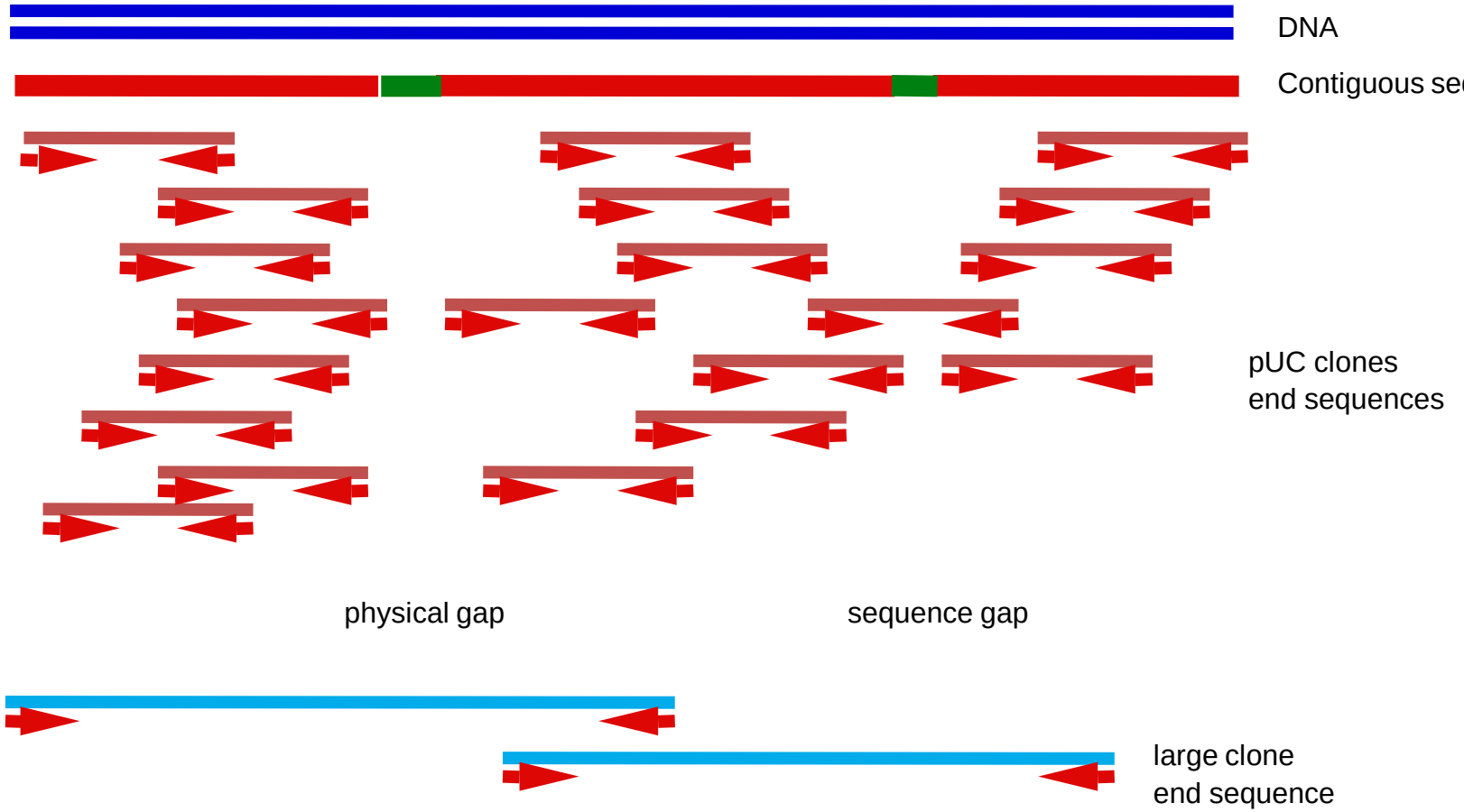
Secuencias vs Contigs obtenidos



Ensamblado de secuencias

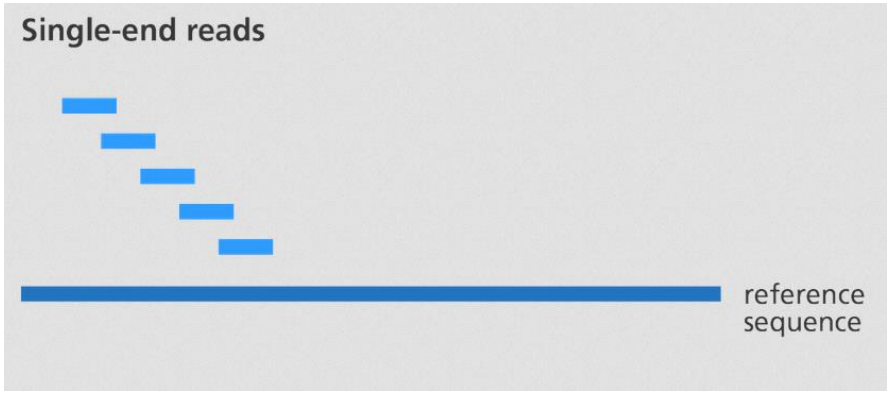
- Métodos de “fuerza bruta” (o “greedy”)
  - Tomo una lectura y busco entre las otras la que mejor solape (largo e identidad)
  - La agrego al contig
  - Continuo con esto hasta que no solapen mas lecturas
  - Si aun hay lecturas sobrantes, comienzo de nuevo para generar un nuevo contig
- Overlap-layout-consensus (OLC)
- Basados en grafos de de Bruijn

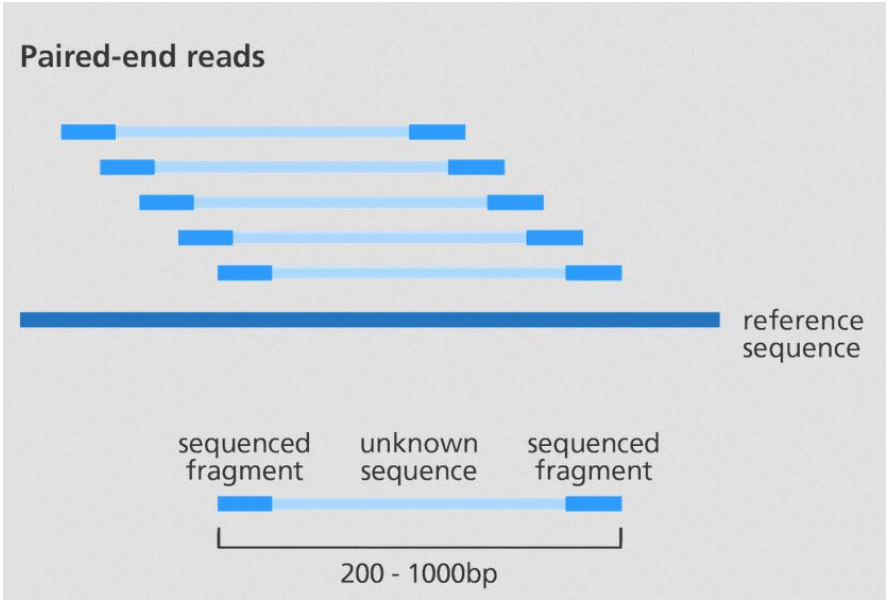
Ensamblado de secuencias



Ensamblado de secuencias

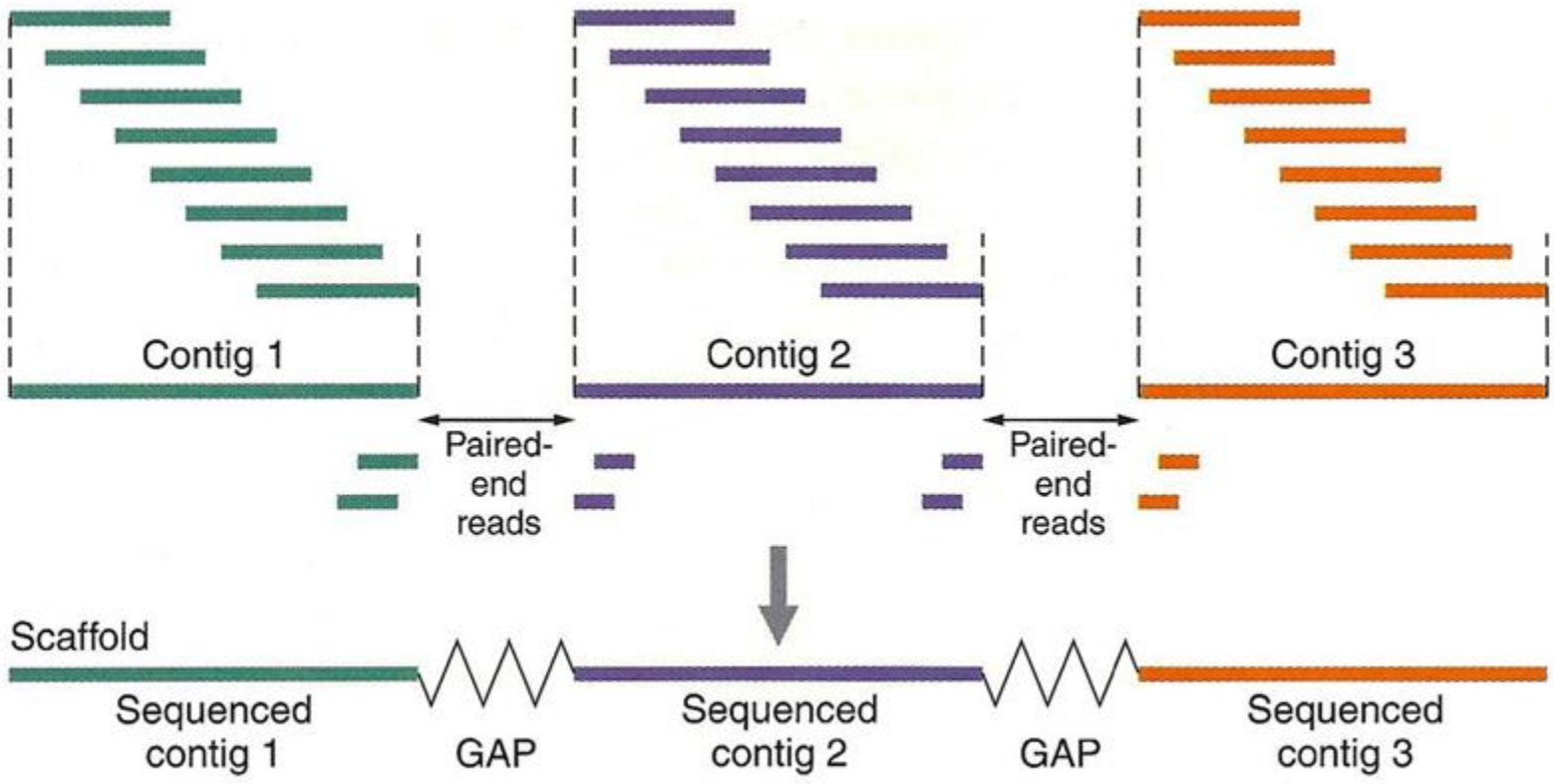
- Secuenciado de pares



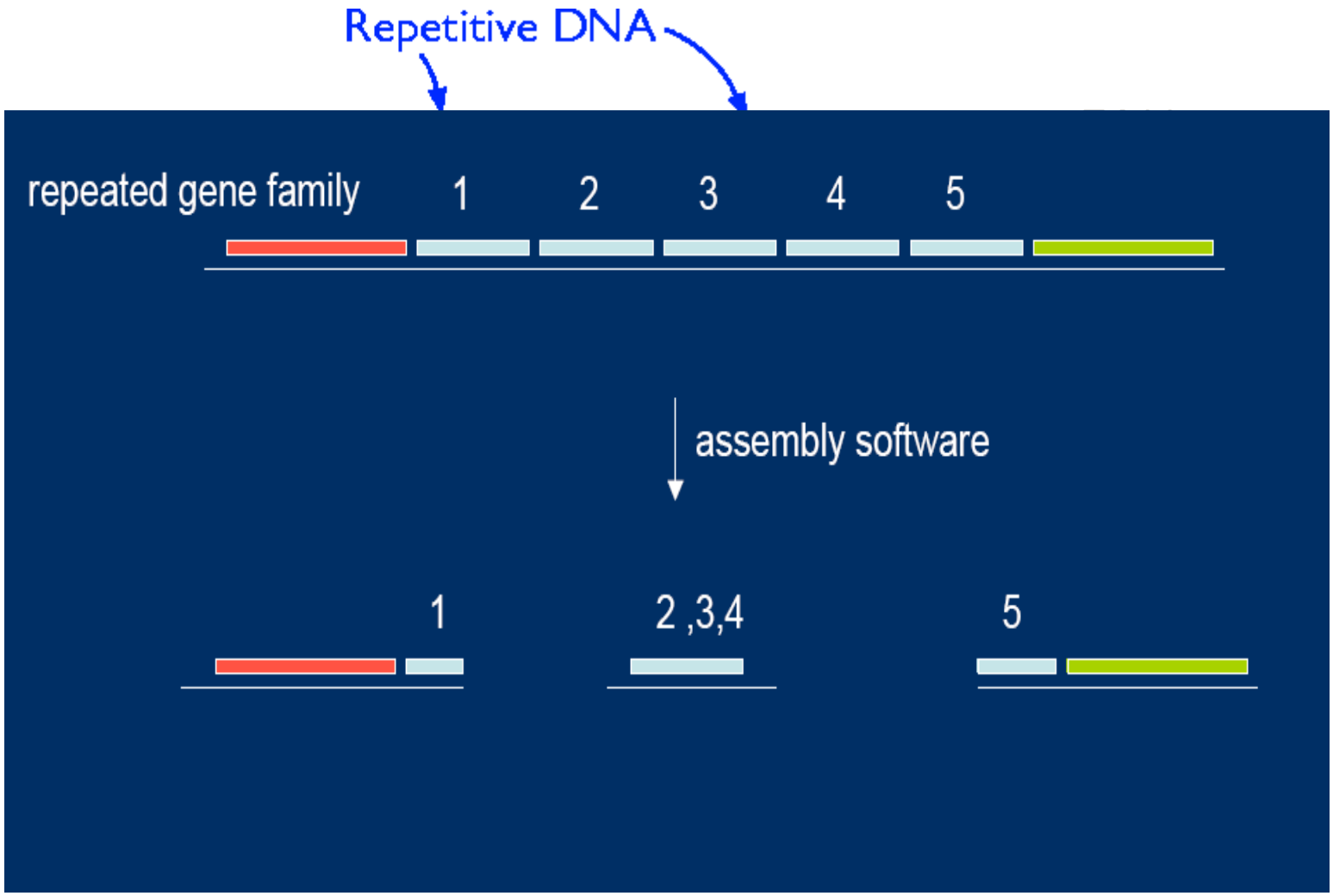


Ensamblado de secuencias

- Permiten unir contigs en scafolds

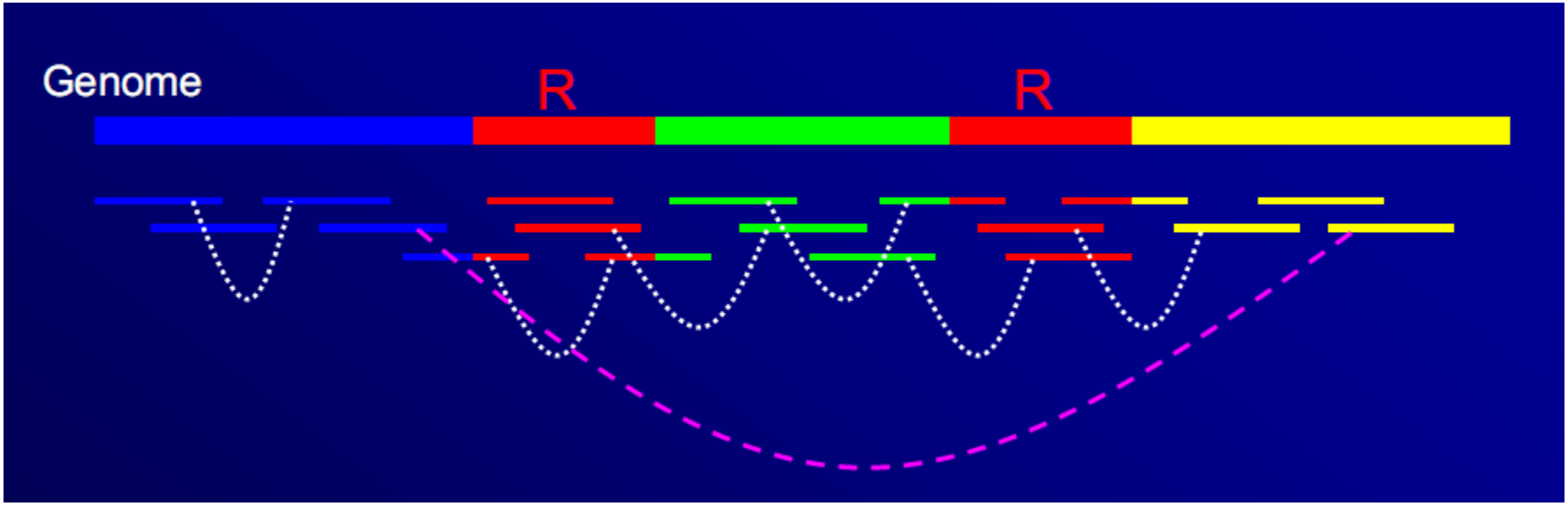


Problema…


Repetitive DNA
repeated gene family
1
2
3
4
5
assembly software
1
2 ,3,4
5

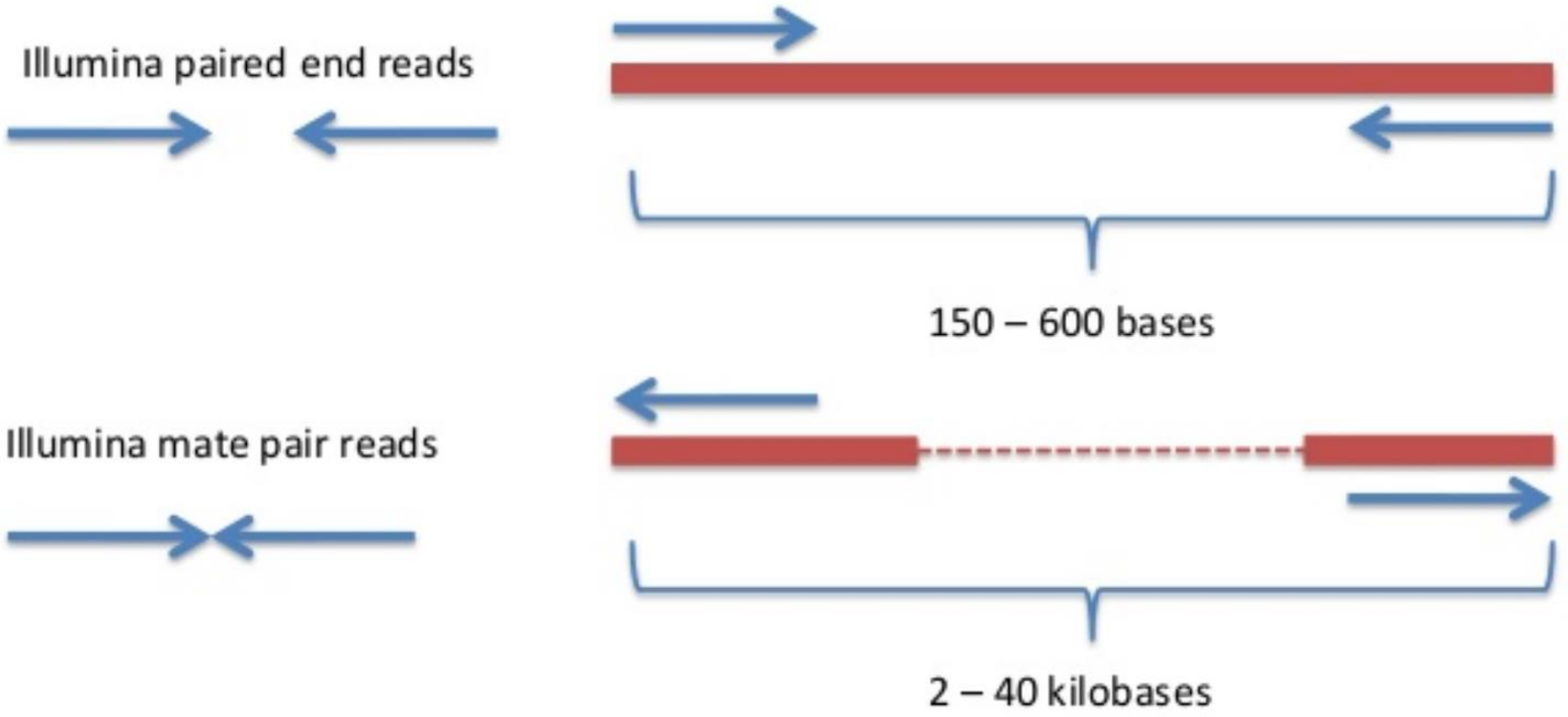
Solución…

- Hacer bibliotecas con tamaños de inserto diferentes
- Hacer lecturas pareadas y “mate pairs”

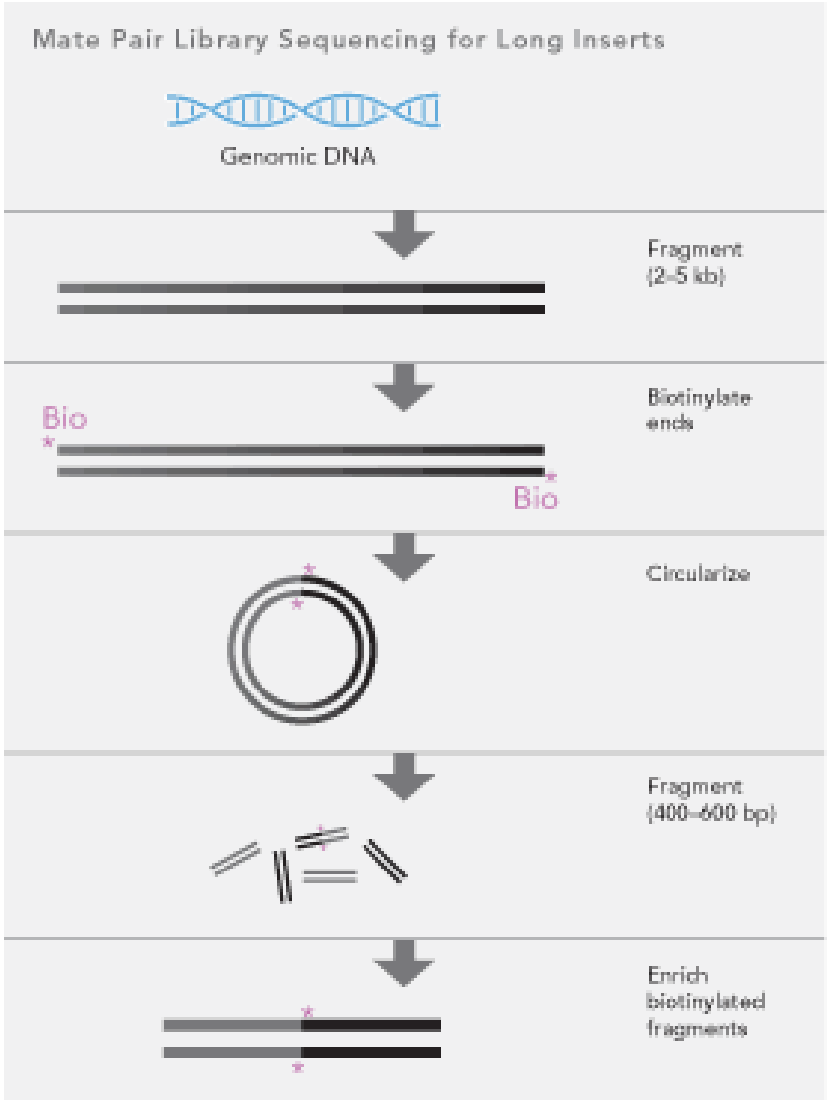


Solución…

- Hacer bibliotecas con tamaños de inserto diferentes
- Hacer lecturas pareadas y “mate pairs”



Solución…


Mate Pair Library Sequencing for Long Inserts
Genomic DNA
Fragment
(2–5 kb)
Bio
Bio
Biotinylate
ends
Circularize
Fragment
(400–600 bp)
Enrich
biotinylated
fragments

Cobertura

- En promedio cuantas veces leo cada base en el genoma ensamblado
- Se calcula como (*redundancia de cobertura*):

$$C = \frac{LN}{G}$$

- Donde L es el largo de las secuencias, N su número y G la cantidad de bases del genoma

Cobertura

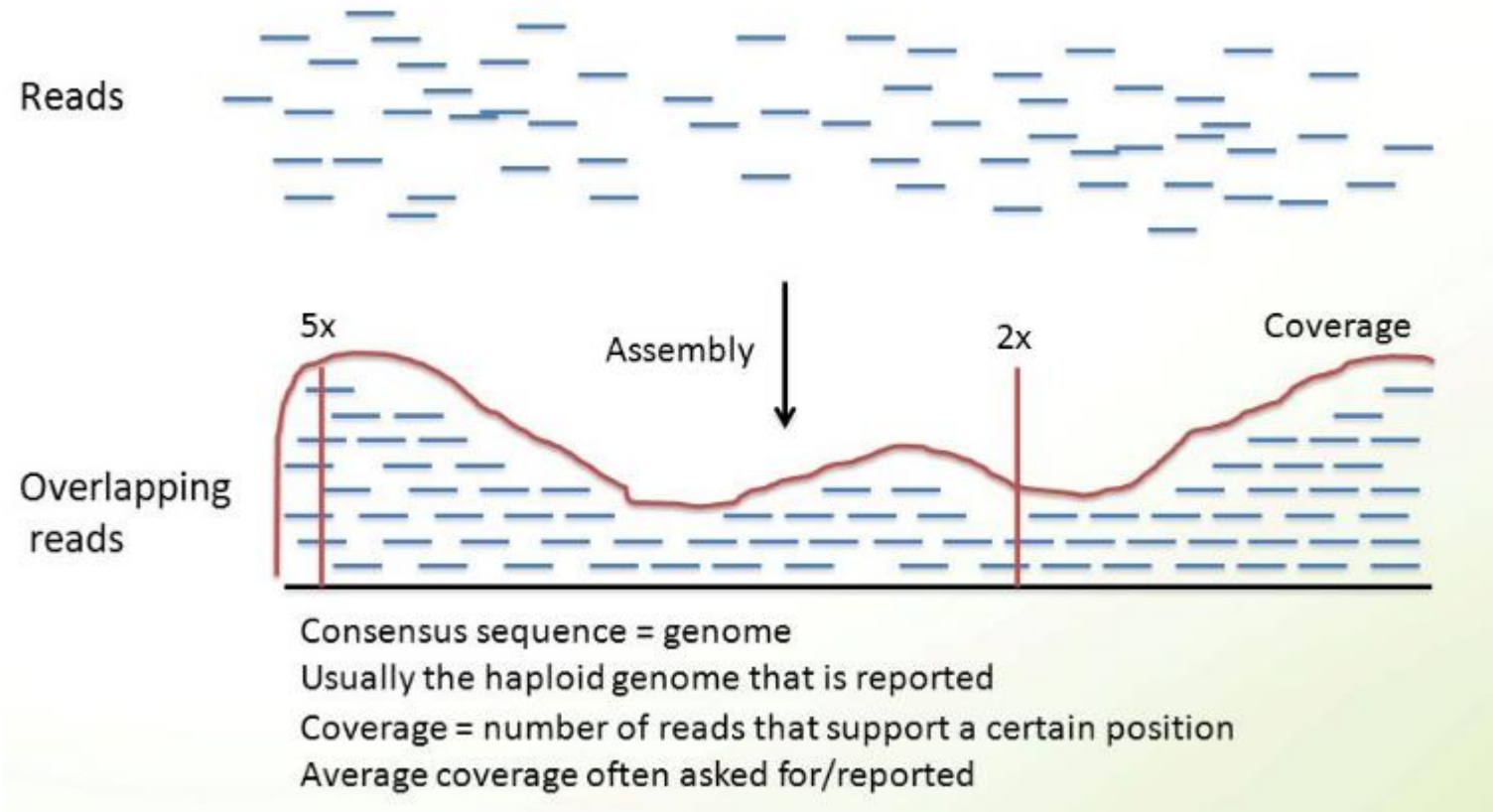
- La probabilidad de que una base no sea secuenciada depende de la cantidad de secuencias obtenidas y por lo tanto de la cobertura.

$$Po = e^{-c}$$

- 0,99 -> 4,6x

- $3x10^9$ -> 30.000.000 sin cobertura

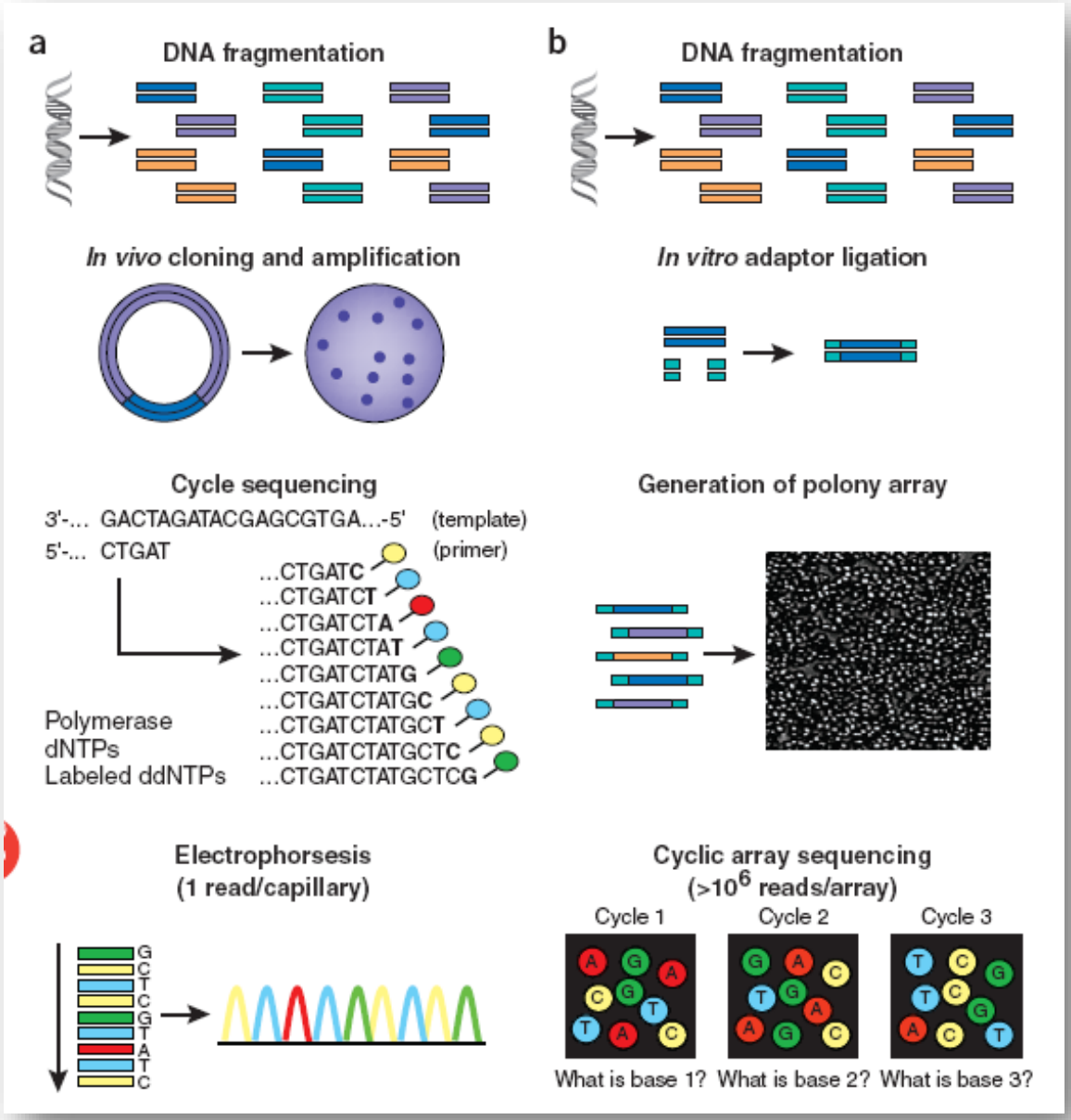
Cobertura


Reads
5x
Assembly
2x
Coverage
Overlapping
reads
Consensus sequence = genome
Usually the haploid genome that is reported
Coverage = number of reads that support a certain position
Average coverage often asked for/reported

Secuenciación de nueva generación

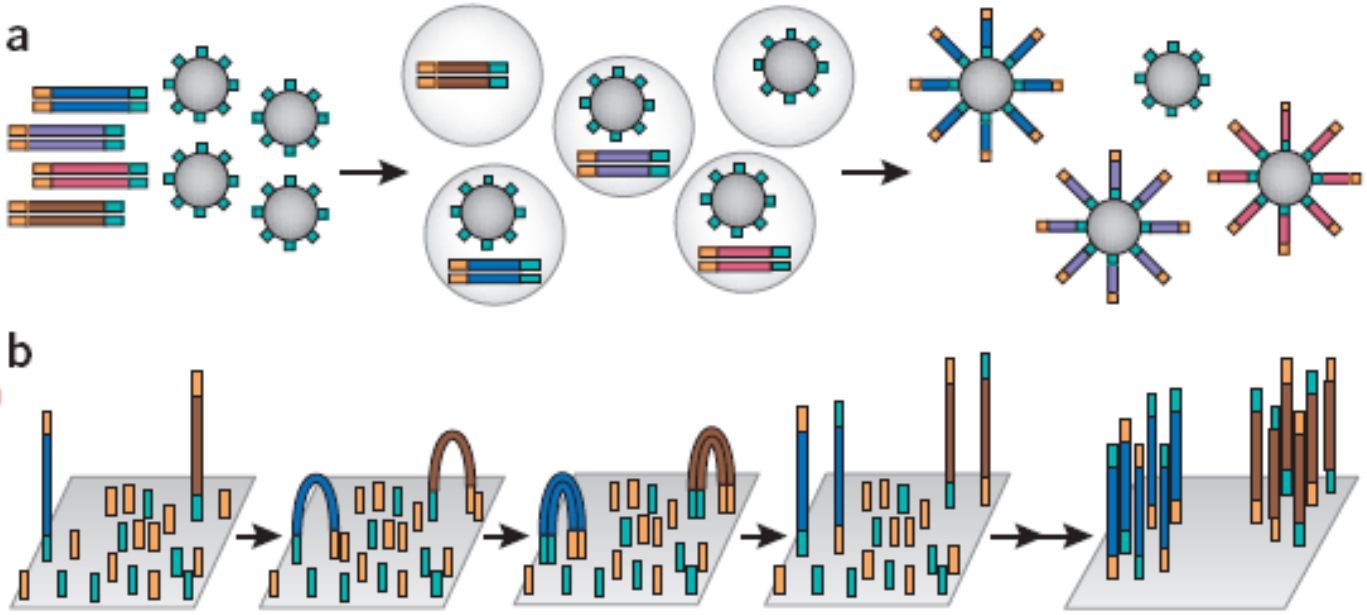
- Se basa en la paralelización de millones de reacciones de secuenciamiento
- Se generan cantidades enormes de datos

Sanger vs. nueva generación

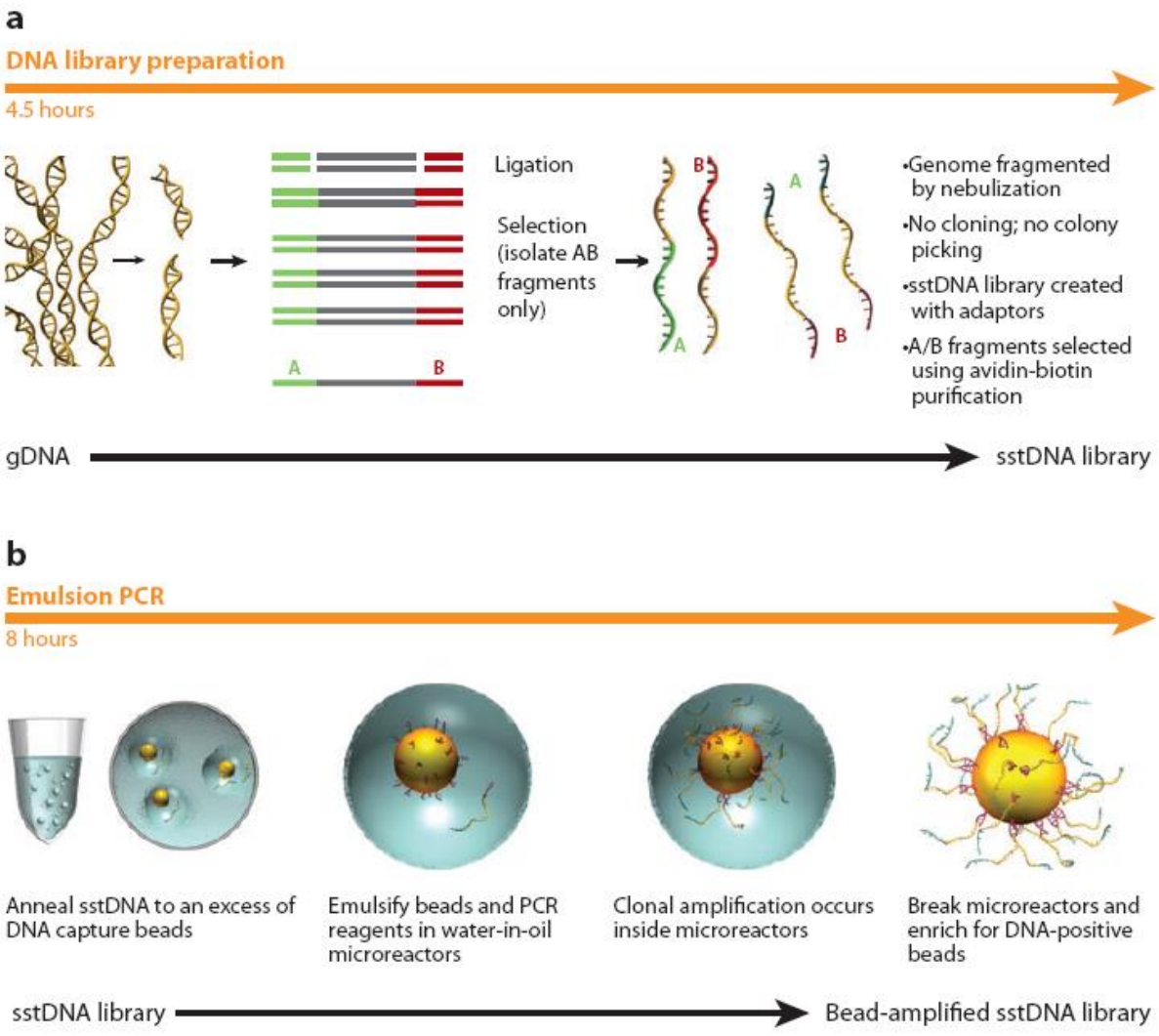


Generación de las polonys

- a- 454, Polonator, SOLiD, Ion Torrent
- b- Solexa
- PacBio, Oxford Nanopore- no amplifican (“single molecule”)



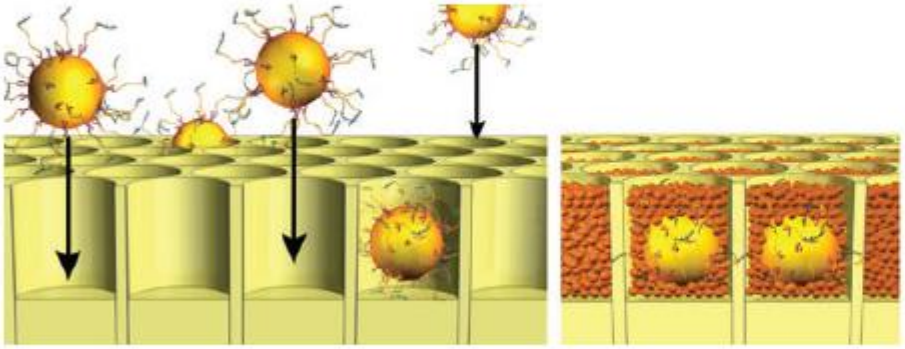
Pirosecuenciación (454 Roche)



c

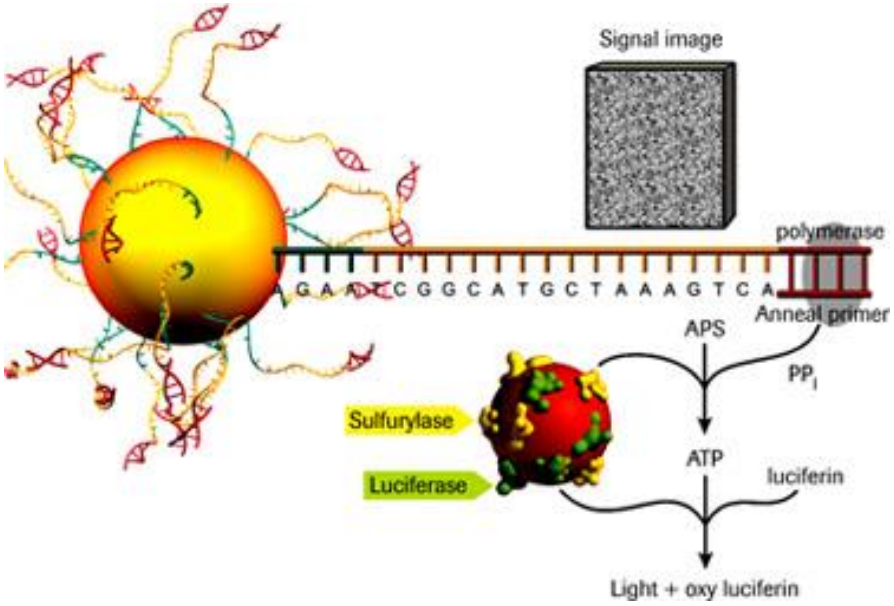
Sequencing

7.5 hours



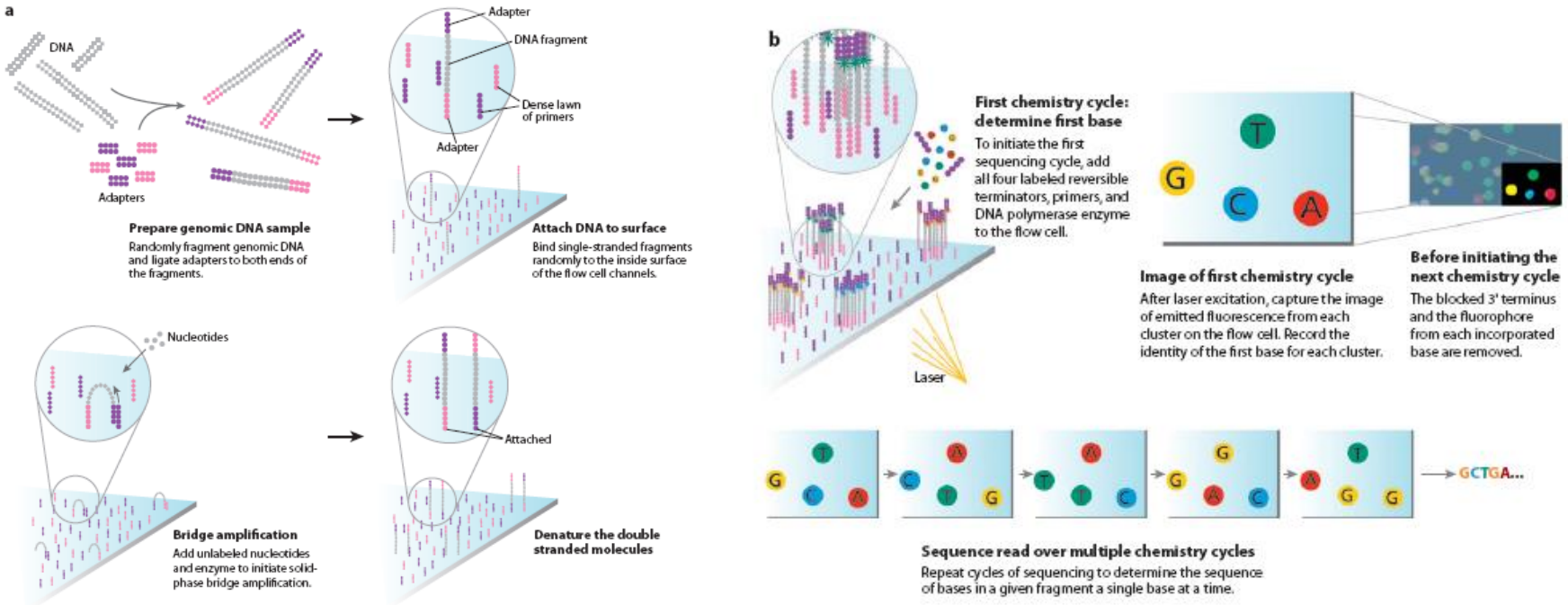
- Well diameter: average of 44 µm
- 400,000 reads obtained in parallel
- A single cloned amplified sstDNA bead is deposited per well

Amplified sstDNA library beads → Quality filtered bases



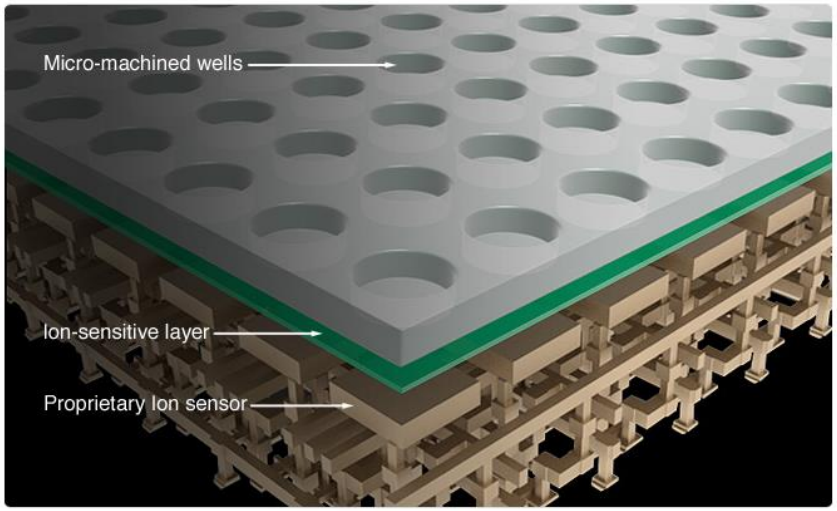
Illumina (Solexa)

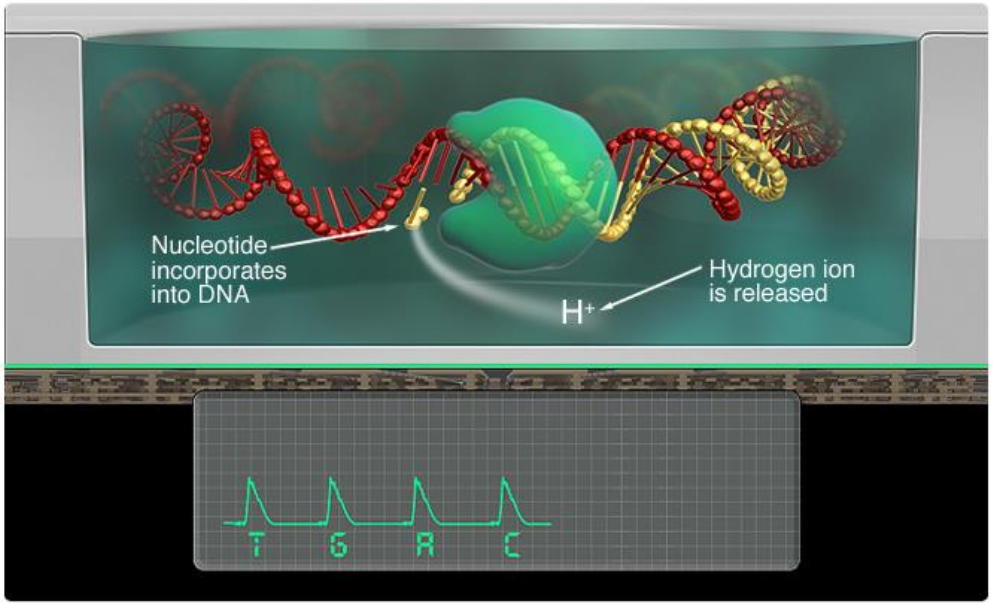
- Secuenciado por síntesis como el 454



Ion Torrent

- Medidor de Ph a nanoescala





Comparación entre las plataformas

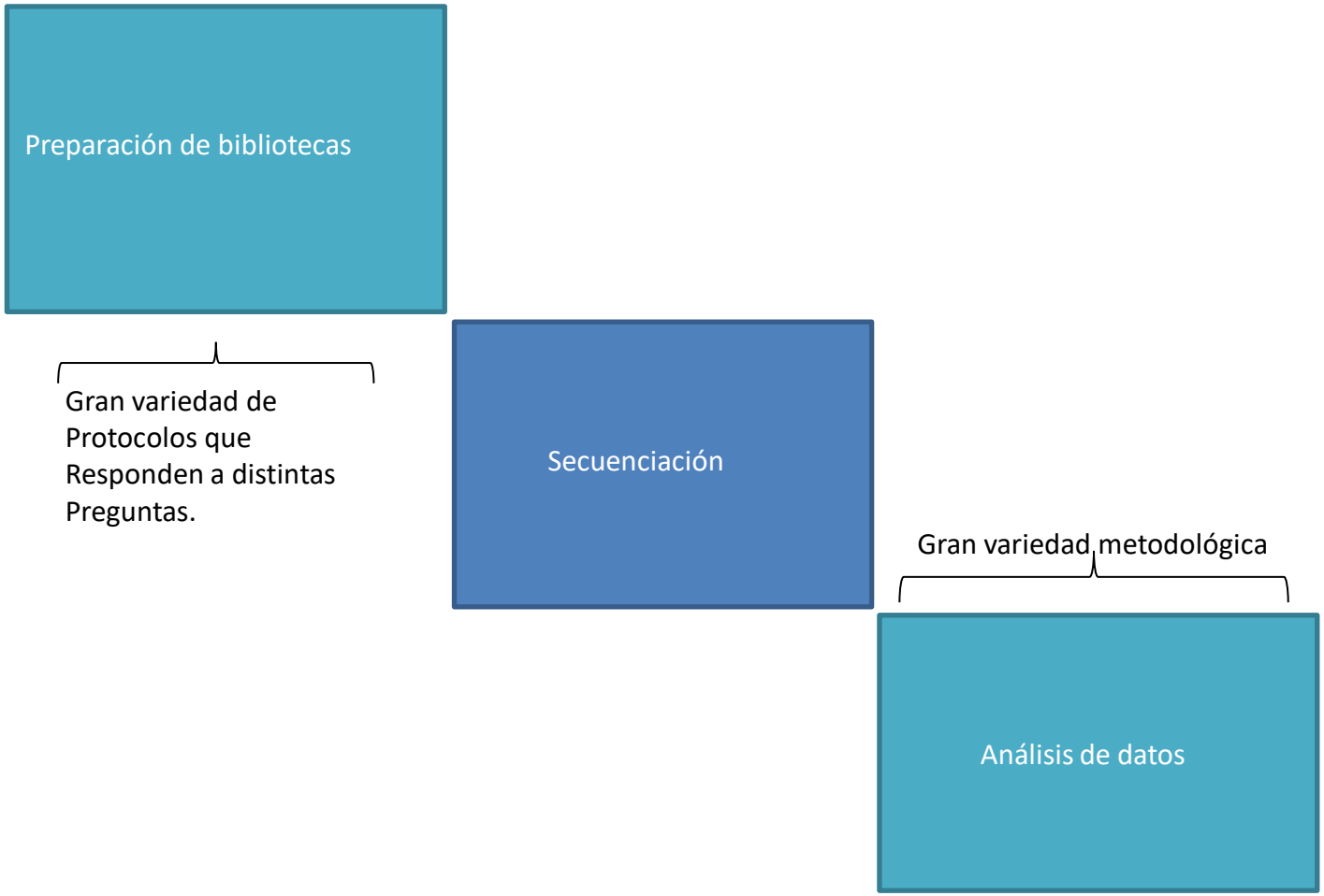
(a)

Sequencer	454 GS FLX	HiSeq 2000	SOLiDv4	Sanger 3730xl
Sequencing mechanism	Pyrosequencing	Sequencing by synthesis	Ligation and two-base coding	Dideoxy chain termination
Read length	700 bp	50SE, 50PE, 101PE	50 + 35 bp or 50 + 50 bp	400~900 bp
Accuracy	99.9%*	98%, (100PE)	99.94% *raw data	99.999%
Reads	1 M	3 G	1200~1400 M	—
Output data/run	0.7 Gb	600 Gb	120 Gb	1.9~84 Kb
Time/run	24 Hours	3~10 Days	7 Days for SE 14 Days for PE	20 Mins~3 Hours
Advantage	Read length, fast	High throughput	Accuracy	High quality, long read length
Disadvantage	Error rate with polybase more than 6, high cost, low throughput	Short read assembly	Short read assembly	High cost low throughput

(b)

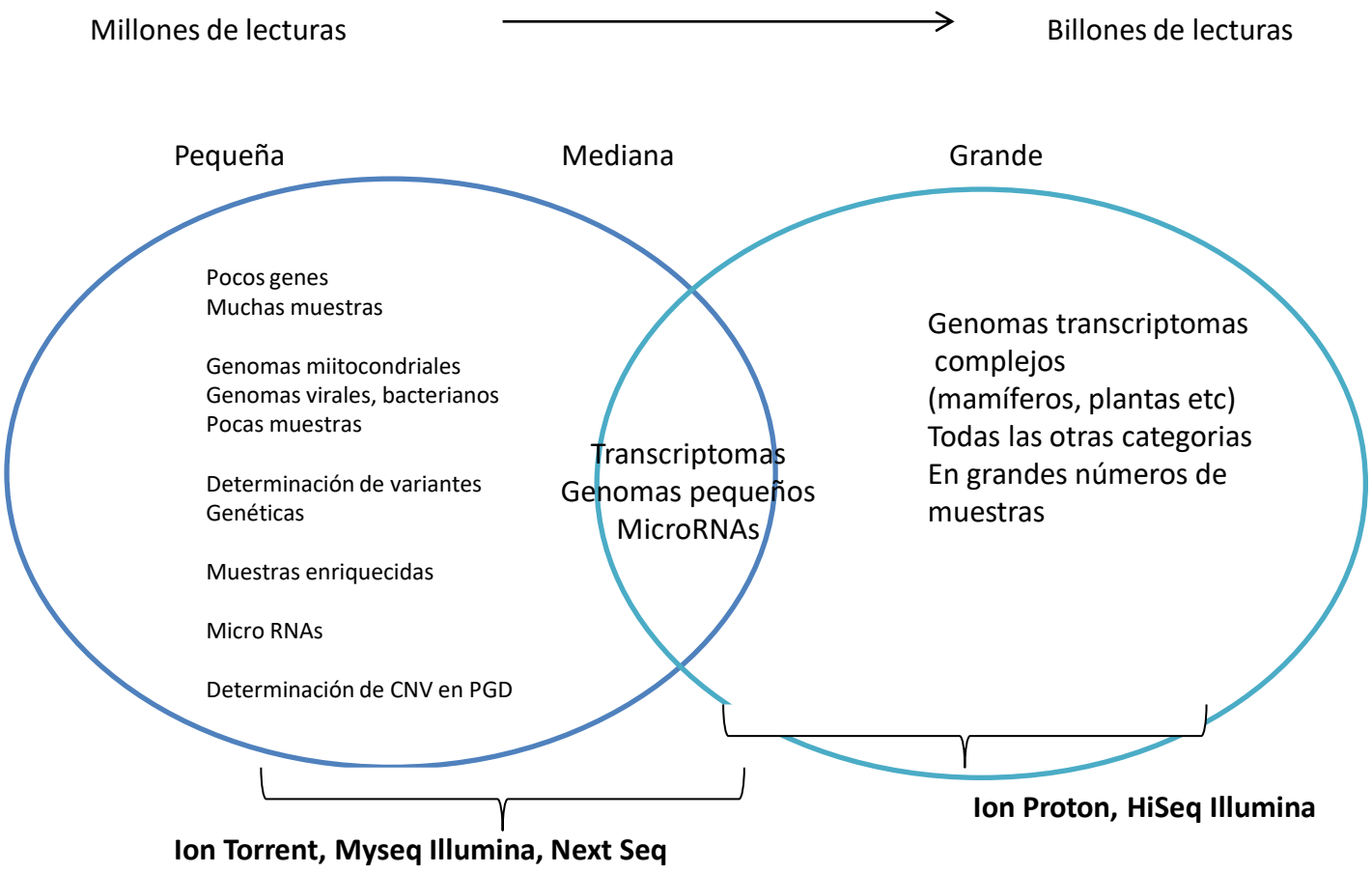
Sequencers	454 GS FLX	HiSeq 2000	SOLiDv4	3730xl
Instrument price	Instrument $500,000, $7000 per run	Instrument $690,000, $6000/(30x) human genome	Instrument $495,000, $15,000/100 Gb	Instrument $95,000, about $4 per 800 bp reaction
CPU	2* Intel Xeon X5675	2* Intel Xeon X5560	8* processor 2.0 GHz	Pentium IV 3.0 GHz
Memory	48 GB	48 GB	16 GB	1 GB
Hard disk	1.1 TB	3 TB	10 TB	280 GB
Automation in library preparation	Yes	Yes	Yes	No
Other required device	REM e system	cBot system	EZ beads system	No
Cost/million bases	$10	$0.07	$0.13	$2400

Comparación entre las plataformas

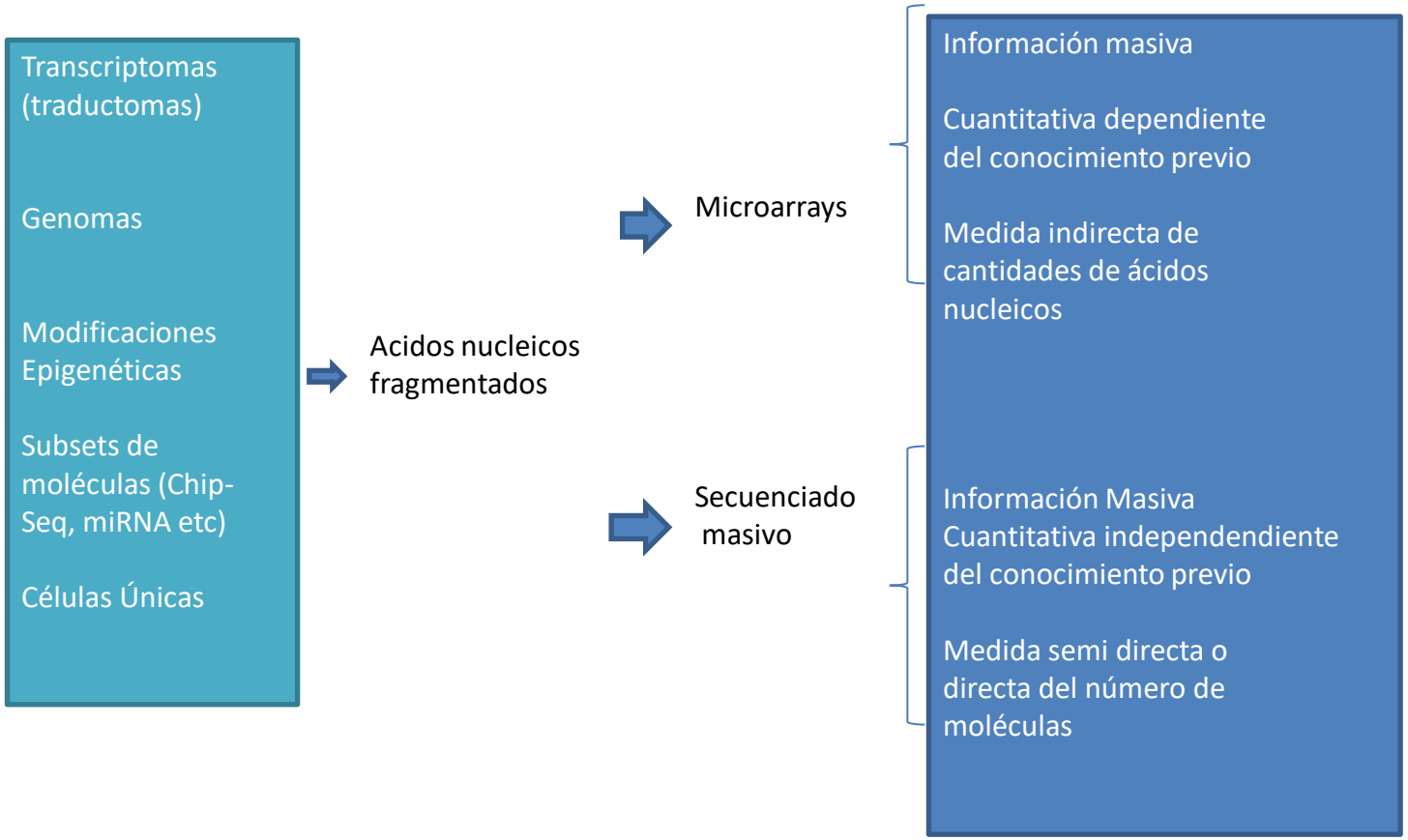


Comparación entre las plataformas

Escalas de trabajo y equipamiento.



Comparación entre las plataformas



Comparación entre las plataformas

(c)

Sequencers	454 GS FLX	HiSeq 2000	SOLiDv4	3730xl
Resequencing		Yes	Yes	
De novo	Yes	Yes		Yes
Cancer	Yes	Yes	Yes	
Array	Yes	Yes	Yes	Yes
High GC sample	Yes	Yes	Yes	
Bacterial	Yes	Yes	Yes	
Large genome	Yes	Yes		
Mutation detection	Yes	Yes	Yes	Yes

(1) All the data is taken from daily average performance runs in BGI. The average daily sequence data output is about 8 Tb in BGI when about 80% sequencers (mainly HiSeq 2000) are running.

(2) The reagent cost of 454 GS FLX Titanium is calculated based on the sequencing of 400 bp; the reagent cost of HiSeq 2000 is calculated based on the sequencing of 200 bp; the reagent cost of SOLiDv4 is calculated based on the sequencing of 85 bp.

(3) HiSeq 2000 is more flexible in sequencing types like 50SE, 50PE, or 101PE.

(4) SOLiD has high accuracy especially when coverage is more than 30x, so it is widely used in detecting variations in resequencing, targeted resequencing, and transcriptome sequencing. Lanes can be independently run to reduce cost.

Nuevas plataformas

illumina

Log in to get personalized account information. Quick Order View Cart

Contact Us MyIllumina Tools

APPLICATIONS SYSTEMS INFORMATICS CLINICAL SERVICES SCIENCE SUPPORT COMPANY Search

Subscribe | Follow us: | Select Language

Systems / **HiSeq X Ten**

Overview System Kits Support GET A QUOTE

Population power. Extreme throughput. $1,000 human genome.

The HiSeq X Ten is a set of ten ultra-high-throughput sequencers, purpose-built for large-scale human whole-genome sequencing.

Nuevas plataformas

- 3000 Gb por FC!
- 48 genomes per run using dual S4 flow cells
- 164 transcriptomas…
- Genoma por 100 dólares??