Temas actuales: Next Generation Sequencing (NGS)

Bioinformática

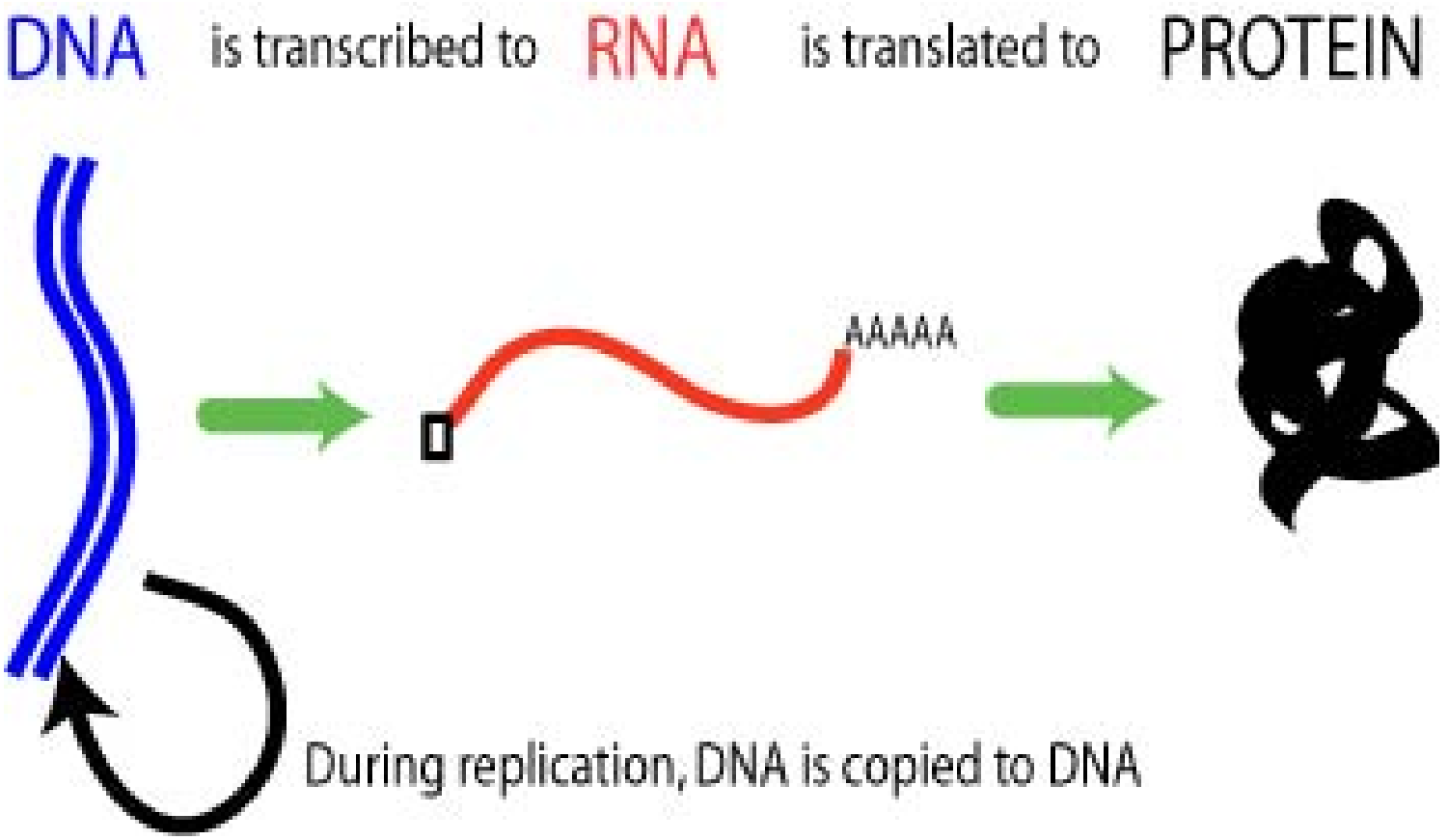
16-3-16

Elvira Mayordomo

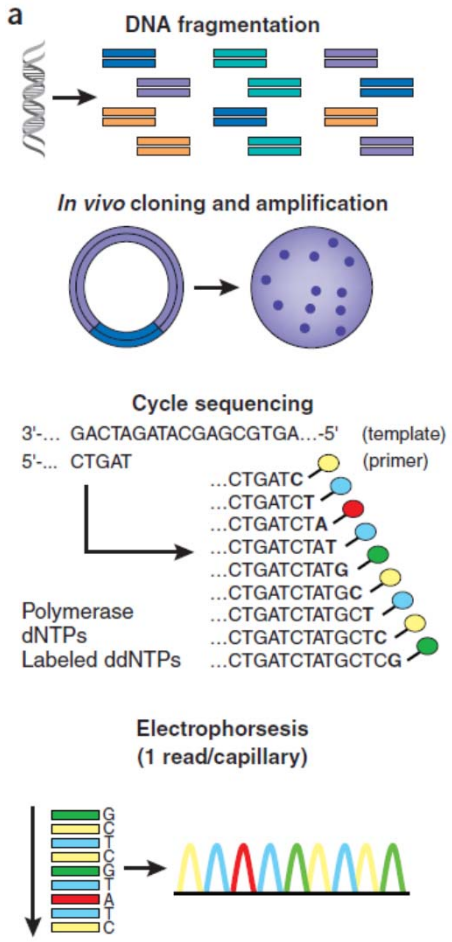
Veremos

- **Historia**
- Plataformas NGS
- Aplicaciones
- Retos bioinformáticos

Dogma central



Secuenciación de Sanger



- El DNA se fragmenta
- Clonación
- Reacción de secuenciación cíclica
- Separación por electroforesis
- Lectura con etiquetas fluorescentes

Sanger vs NGS

- La secuenciación de Sanger ha sido el **único método de secuenciación** de DNA durante **30 años**, pero ...
- ... había gran **necesidad** de una mayor eficiencia y de tecnologías de secuenciación más económicas ...
- NGS tiene la capacidad de **procesar en paralelo** millones de lecturas de secuencia en lugar de 96 a la vez (1/6 del coste)
- **Objeciones**: la fidelidad, la longitud de lectura, el coste de infraestructura, el manejo de grandes volúmenes de datos

NGS

- NGS = Secuenciación amplificada de una única secuencia
- Third generation sequencing = secuenciación de una única molécula

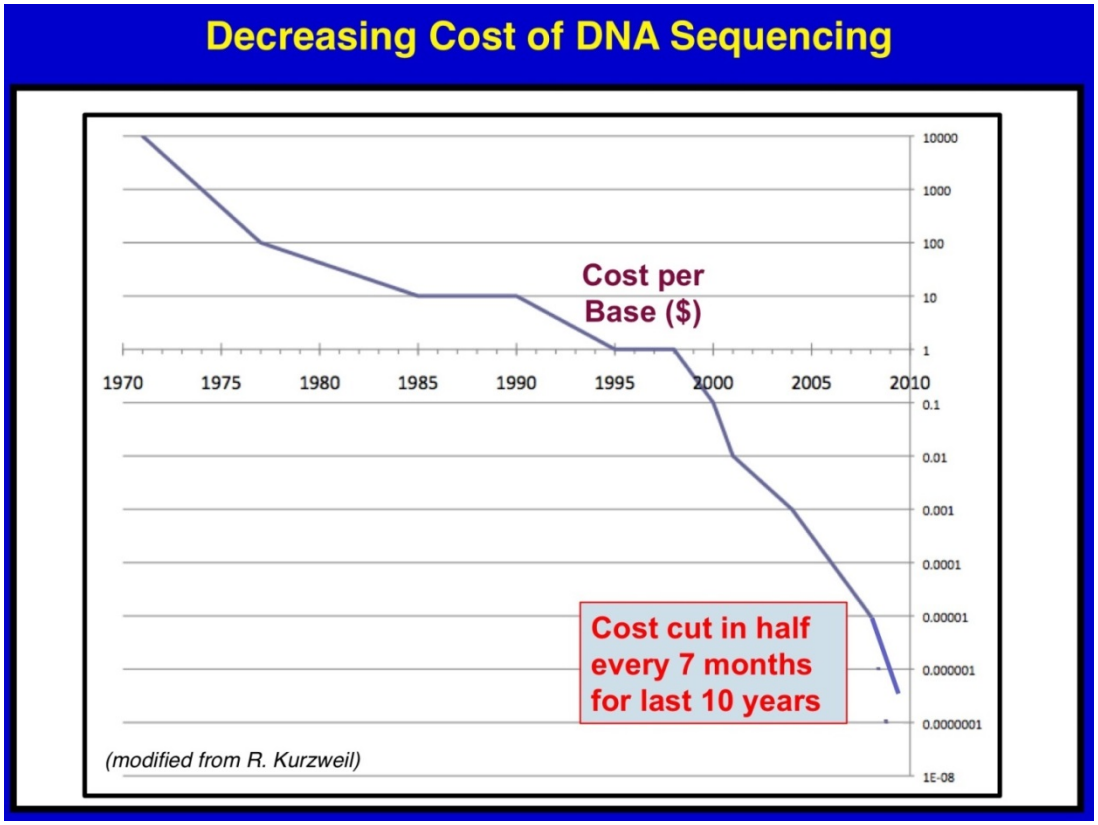
Veremos

- Historia
- **Plataformas NGS**
- Aplicaciones
- Retos bioinformáticos

Plataformas NGS

- Roche / 454 FLX: 2004
- Illumina Solexa Genoma Analyzer: 2006
- Applied Biosystems SOLId™ System: 2007
- Helicos Heliscope™ : 2008
- Pacific Biosciencies SMRT: 2010

Reducción de costes



Tres plataformas de secuenciación principales

- Roche 454
- **Illumina Solexa**
- Applied Biosystems SOLiD

Esquema general

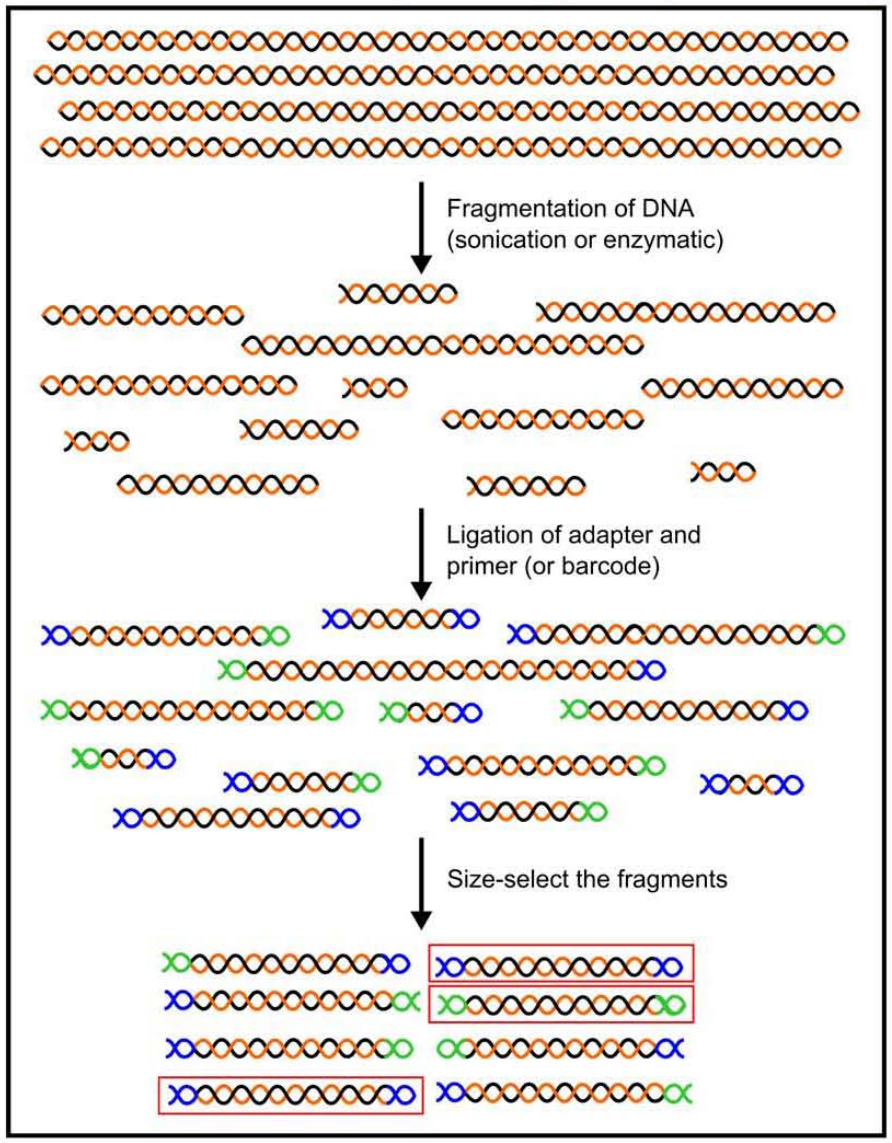
1. **Preparación biblioteca**
2. 2 opciones
    1. PCR por emulsión
    2. PCR polony en una transparencia
3. 3 opciones
    1. Amplificación de puentes (Illumina)
    2. Amplificación por temperatura (SOLiD)
    3. Pirosecuenciación (454)

1 Preparación de la biblioteca

- Se añaden a los fragmentos
  - Adaptadores
  - Primers (principios para replicación)
  - Códigos de barras

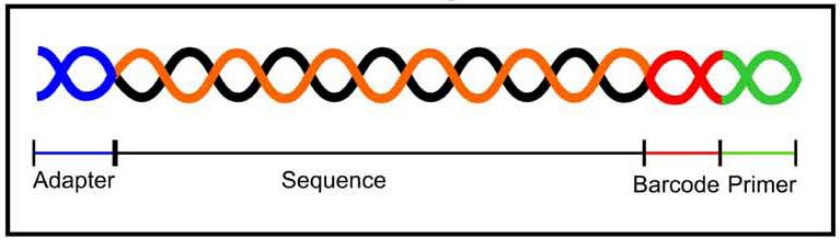
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing



Library preparation

Good fragments :



1- Preparación biblioteca

Esquema general

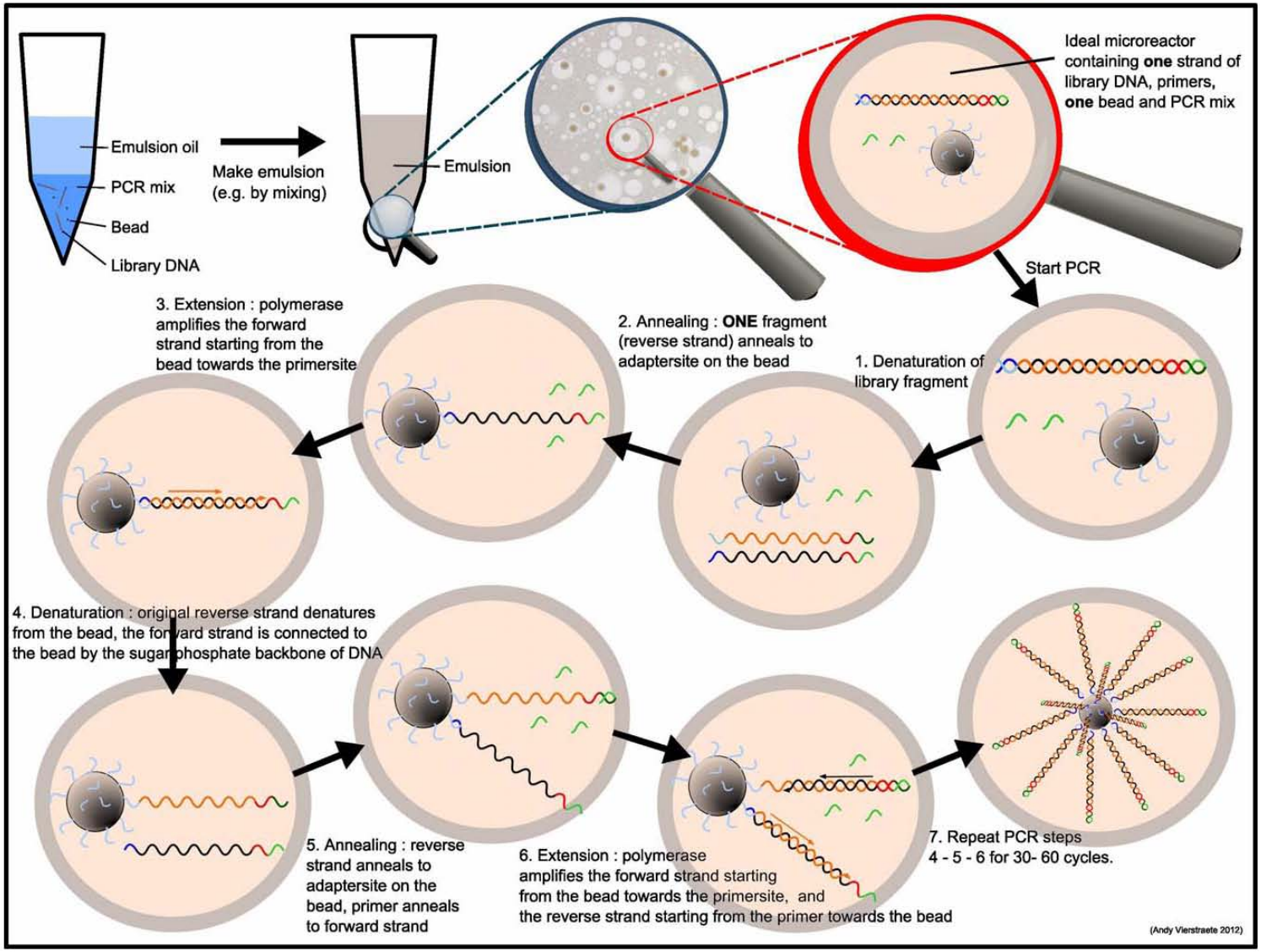
1. Preparación biblioteca
2. 2 opciones
    1. **PCR por emulsión**
    2. PCR polony en una transparencia
3. 3 opciones
    1. Amplificación de puentes (Illumina)
    2. Amplificación por temperatura (SOLiD)
    3. Pirosecuenciación (454)

2.1 PCR por emulsión

- Se añaden microesferas recubiertas con cebadores (beads)
- El fragmento de DNA se desnaturaliza en 2 hebras
- Una hebra se une a la microesfera y empieza a replicarse
- Se consiguen de 30 a 60 copias

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing Emulsion PCR



Esquema general

1. Preparación biblioteca
2. 2 opciones
    1. PCR por emulsión
    2. **PCR polony en una transparencia**
3. 3 opciones
    1. Amplificación de puentes (Illumina)
    2. Amplificación por temperatura (SOLiD)
    3. Pirosecuenciación (454)

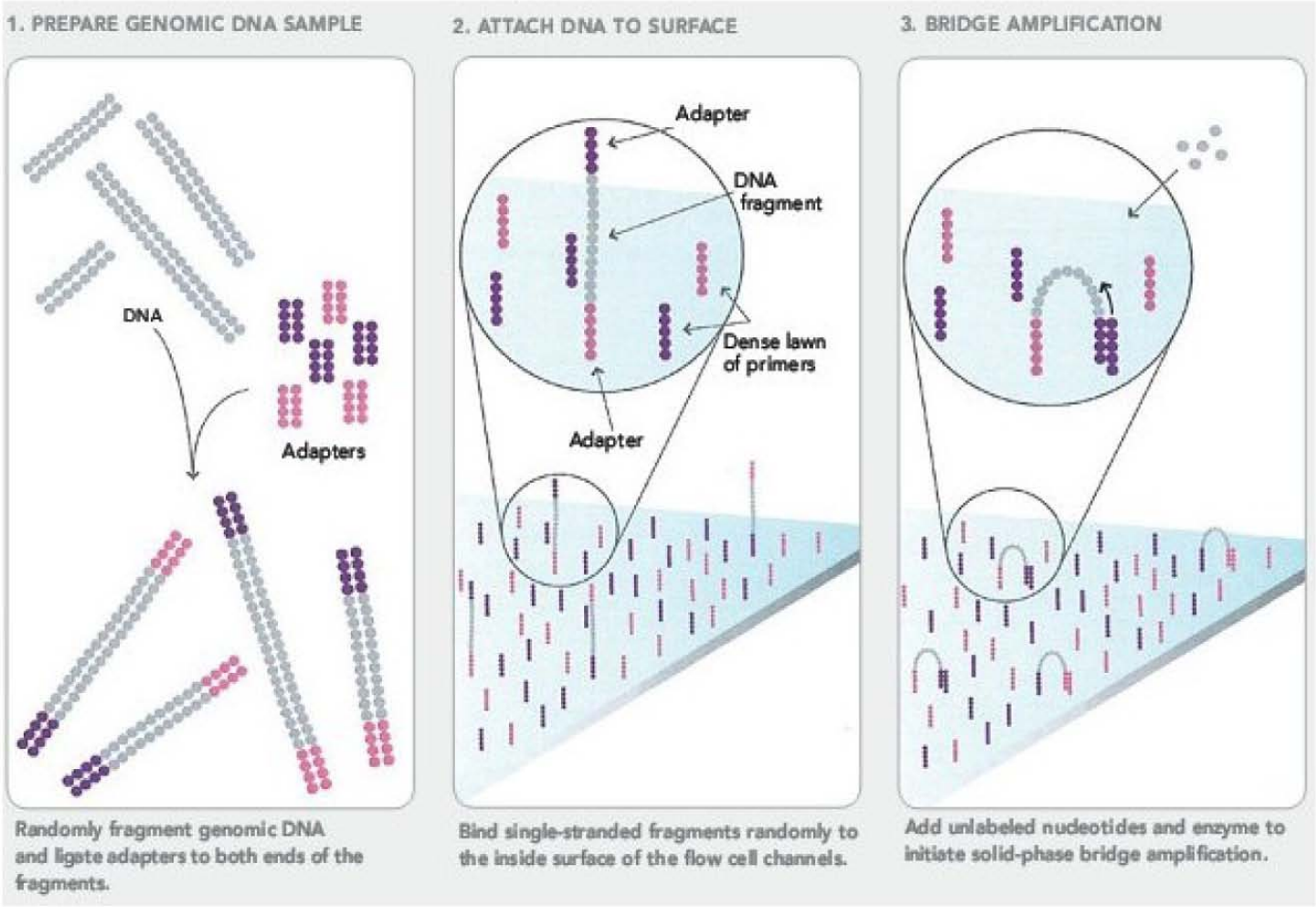
2.2 PCR polony en una transp.

- Pegar el DNA a una superficie con muchos primers
- Provocar la formación de puentes

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing "Polony" PCR

Bridge amplification : Illumina



Esquema general

1. Preparación biblioteca
2. 2 opciones
    1. PCR por emulsión
    2. PCR polony en una transparencia
3. 3 opciones
    1. **Amplificación de puentes (Illumina)**
    2. Amplificación por temperatura (SOLiD)
    3. Pirosecuenciación (454)

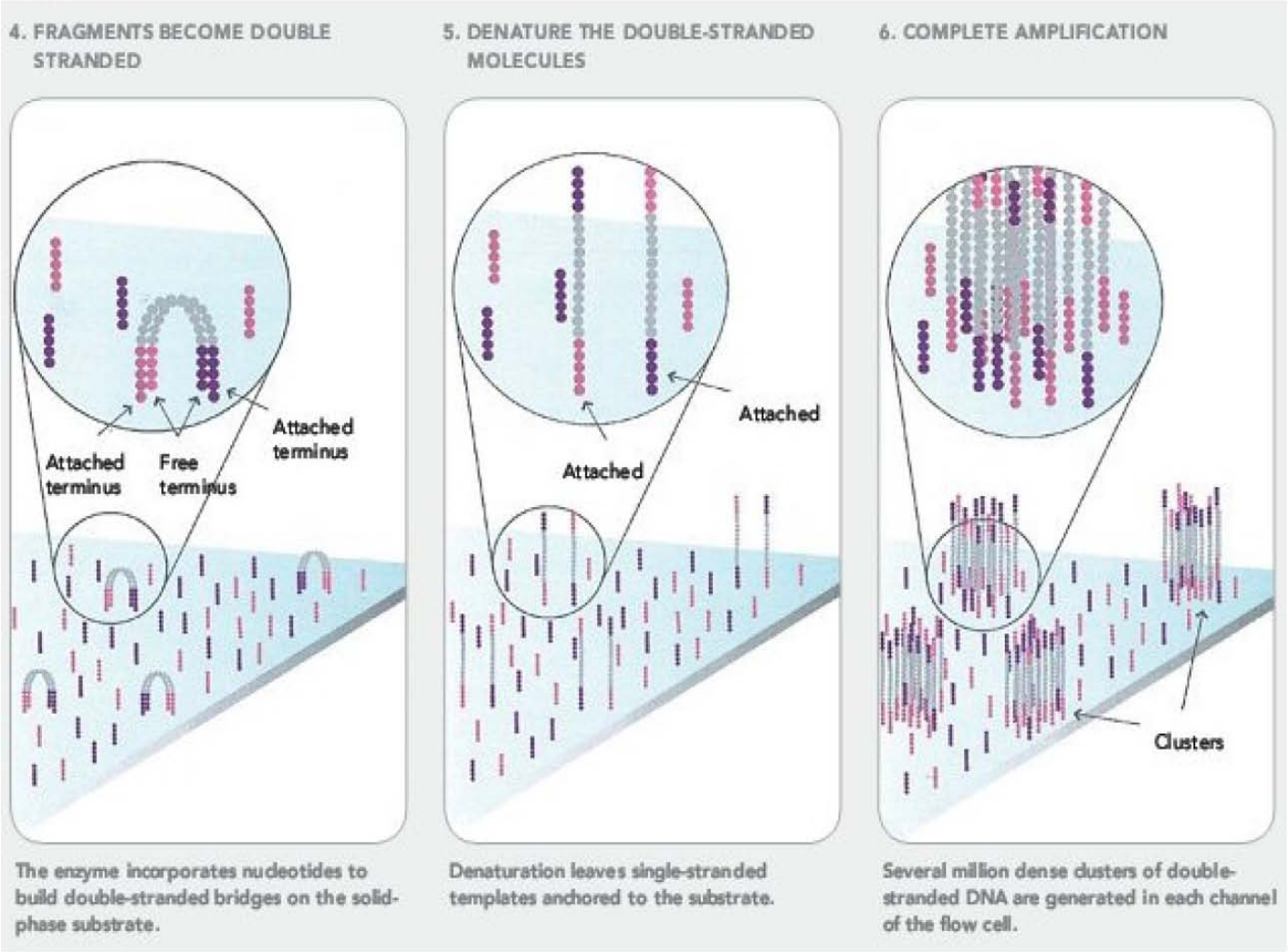
3.1 Amplificación de puentes

- Se construyen dobles puentes
- Se rompen los puentes (el resultado es duplicación)
- Repetir el proceso construcción puentes – duplicación - rotura

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing "Polony" PCR

Bridge amplification : Illumina



Esquema general

1. Preparación biblioteca
2. 2 opciones
    1. PCR por emulsión
    2. PCR polony en una transparencia
3. 3 opciones
    1. Amplificación de puentes (Illumina)
    2. **Amplificación por temperatura (SOLiD)**
    3. Pirosecuenciación (454)

3.2 Amplificación por temperatura

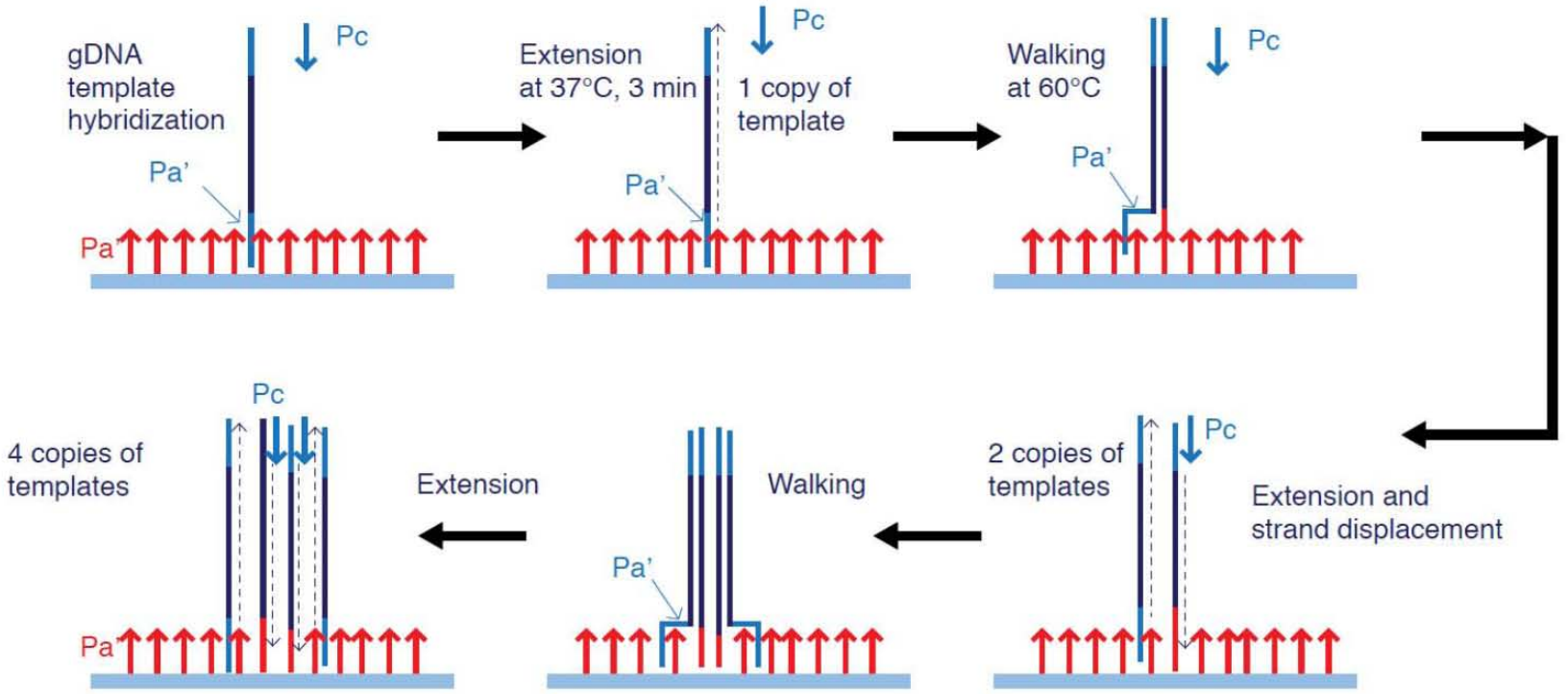
- Replicación y desplazamiento (andar) con distintos cambios de temperatura

- Todo a través de los enzimas adecuados

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing "Polony" PCR

Wildfire amplification : SOLiD



Wildfire chemistry schematic.

Esquema general

1. Preparación biblioteca
2. 2 opciones
    1. PCR por emulsión
    2. PCR polony en una transparencia
3. 3 opciones
    1. Amplificación de puentes (Illumina)
    2. Amplificación por temperatura (SOLiD)
    3. **Pirosecuenciación (454)**

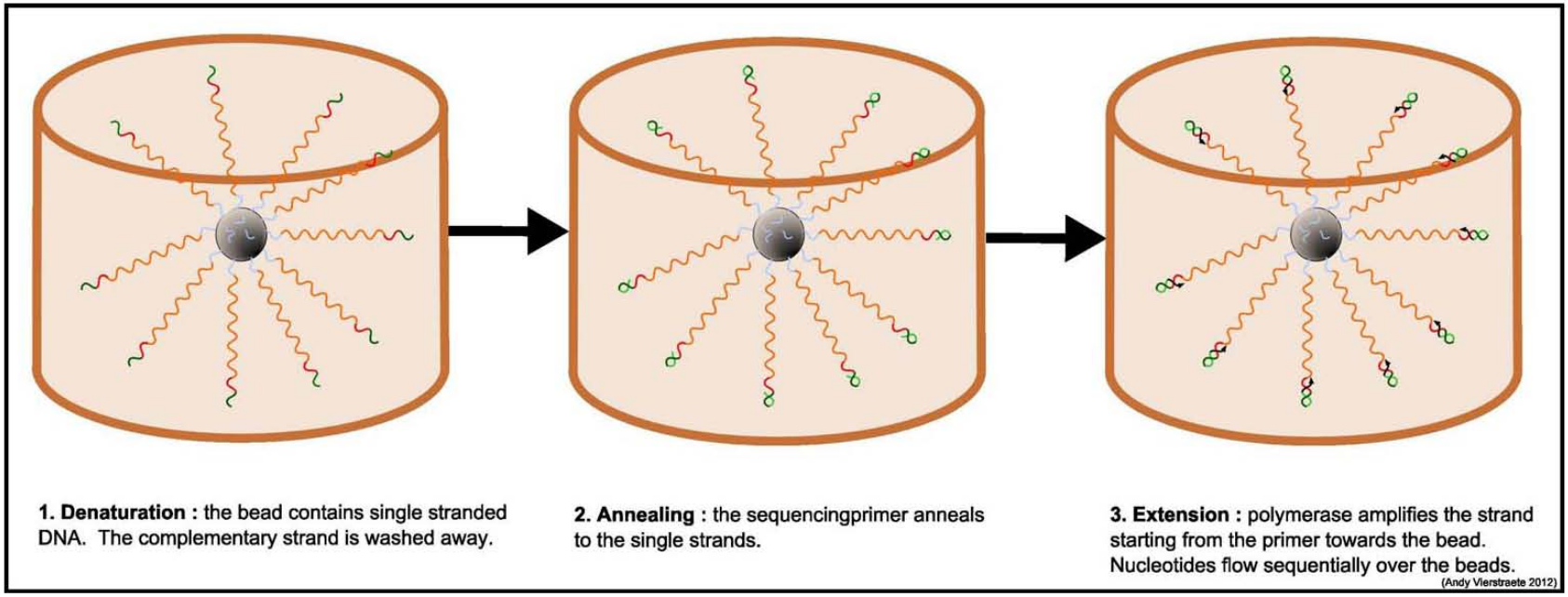
3.3 Pirosecuenciación

- Usa otra vez microesferas a las que se unen varios fragmentos
- Usando primers se produce la replicación

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Next Generation Sequencing : Amplified Single Molecule Sequencing

454 Sequencing / Roche Pyrosequencing



1. Denaturation : the bead contains single stranded DNA. The complementary strand is washed away.

2. Annealing : the sequencingprimer anneals to the single strands.

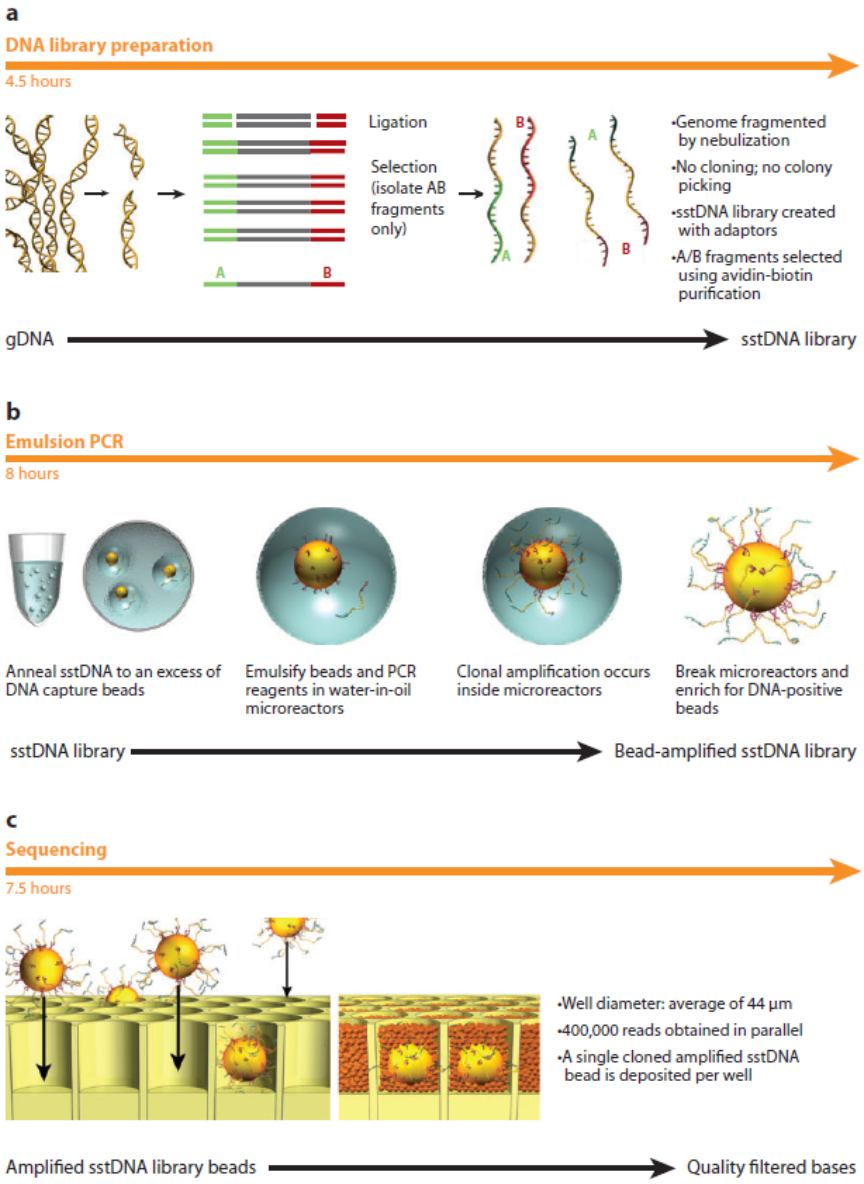
3. Extension : polymerase amplifies the strand starting from the primer towards the bead. Nucleotides flow sequentially over the beads.

(Andy Vierstraete 2012)

Falta el último paso

- Secuenciación: lectura de los trozos originales y replicados
- Bioquímica algo más complicada

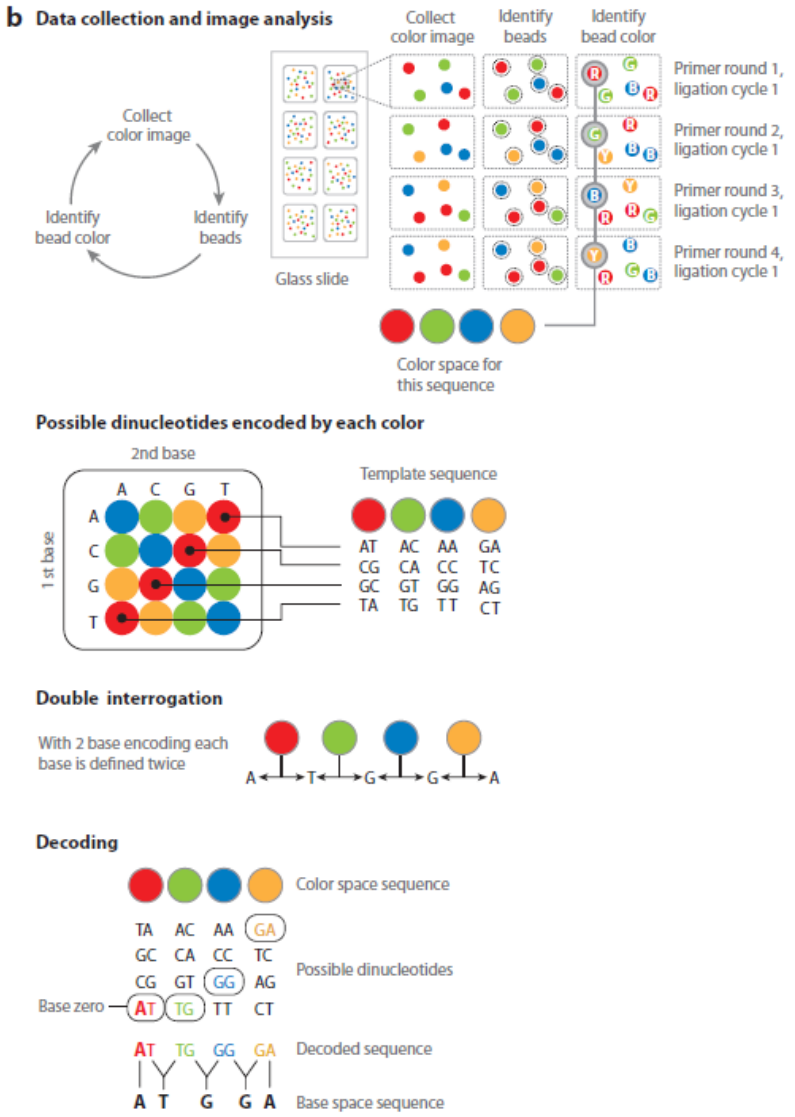
454



Maridis Annu. Rev. Genoma. Genet humano. 2008

SOLID

codificación por color



Maridis Annu. Rev. Genoma. Genet humano. 2008

Comparación de los métodos existentes

	Feature generation	Sequencing by synthesis
454	Emulsion PCR	Polymerase (pyrosequencing)
Solexa	Bridge PCR	Polymerase (reversible terminators)
SOLiD	Emulsion PCR	Ligase (octamers with two-base encoding)
Polonator	Emulsion PCR	Ligase (nonamers)
HeliScope	Single molecule	Polymerase (asynchronous extensions)

	Cost per megabase	Cost per instrument	Paired ends?	1° error modality	Read-length
454	~$60	$500,000	Yes	Indel	250 bp
Solexa	~$2	$430,000	Yes	Subst.	36 bp
SOLiD	~$2	$591,000	Yes	Subst.	35 bp
Polonator	~$1	$155,000	Yes	Subst.	13 bp
HeliScope	~$1	$1,350,000	Yes	Del	30 bp

Datos reales - espacio de nucleótidos

- Solexa (Illumina)

@ SRR002051.1: 8: 1: 325: 773 Longitud = 33
AAAGAACATTAAAGCTATATTATAAGCAAAGAT
+ SRR002051.1: 8: 1: 325: 773 Longitud = 33
IIIIIIIIIIIIIIIIIIIIIIIII'II @ I $) -
@ SRR002051.2: 8: 1: 409: 432 Longitud = 33
AAGTTATGAAATTGTAATTCCAATATCGTAAGC
+ SRR002051.2: 8: 1: 409: 432 Longitud = 33
IIIIIIIIIIIIIIIIIIIIIIIIIIIII07
@ SRR002051.3: 8: 1: 488: 490 Longitud = 33
AATTTCTTACCATATTAGACAAGGCACTATCTT
+ SRR002051.3: 8: 1: 488: 490 Longitud = 33
IIIIIIIIIIIIIIIIIIIIIIIIIIIII + i

Datos reales - espacio de color

- datos SOLID

> 1_24_47_F3
T1.1.23..0120230.32003330003003001002 2.00.0201.0201
> 1_24_52_F3
T2.3.21..2122321.213110332101132321002.11.0111.1222
> 1_24_836_F3
T0.2.22..2222222.010203032021102220200.01.2211.2211
> 1_24_1404_F3
T2.3.30..2013222.222103131323012313233.22.2220.0213
> 1_25_202_F3
T0.3213.111202312203021101111330201000313.121122211
> 1_25_296_F3
T0.1130.100123202213120023121112113212121.013301210

Espacio de color …

- AA, CC, GG, TT : 0
- AC, CA, GT, TG : 1
- AG, CT, GA, TC : 2
- AT, CG, GC, TA : 3

- T2.3.21..212232

Diferencia de salida de datos entre las tres plataformas

- espacio de nucleótidos vs. espacio de color
- Longitud del lecturas cortas

  454 (400 ~ 500 pb)> SOLID (70 pb) ~ Solexa (36 ~ 120 pb)

Veremos

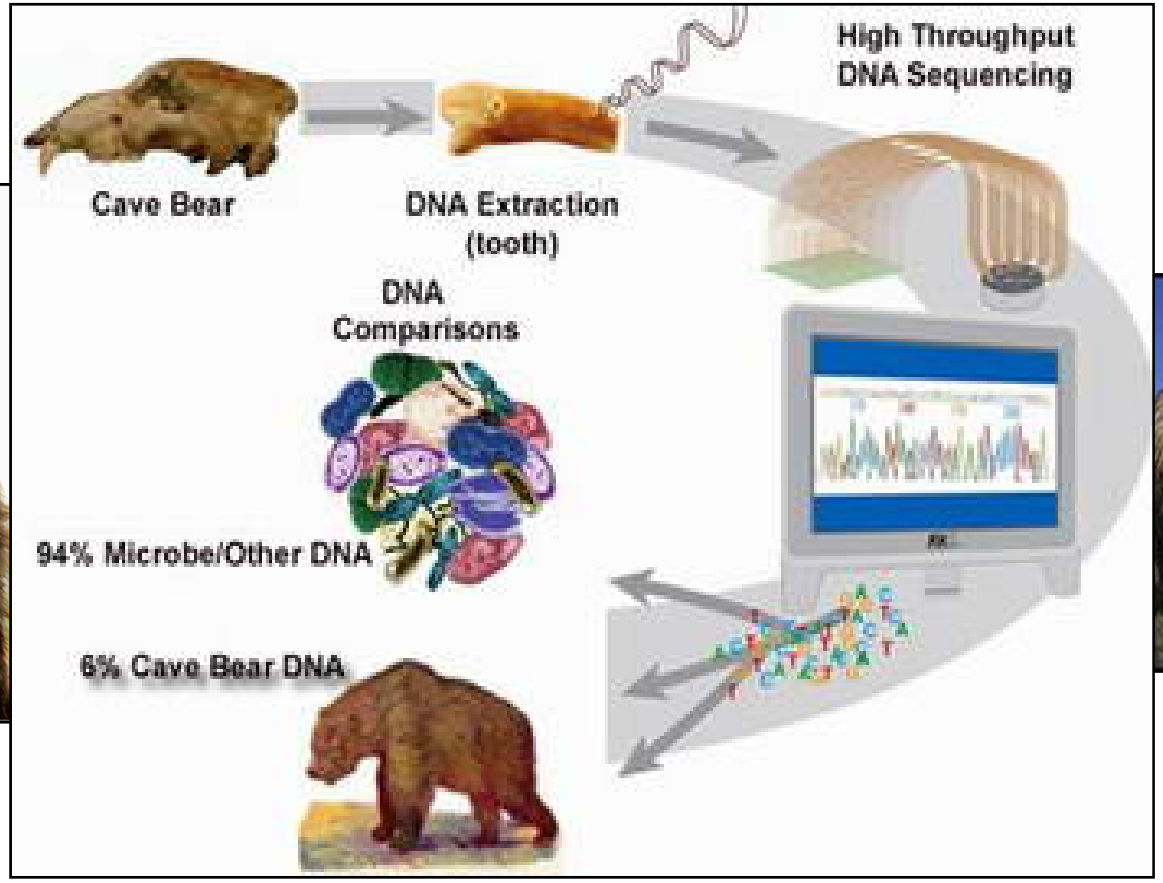
- Historia
- Plataformas NGS
- **Aplicaciones**
- Retos bioinformáticos

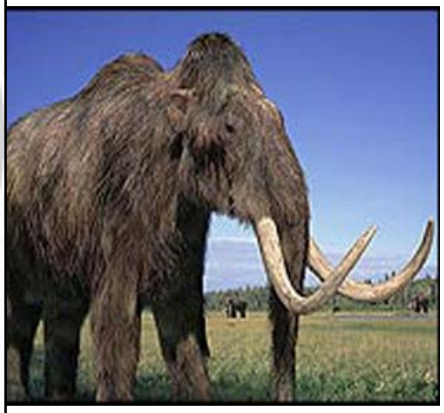
Aplicaciones

- Ensamblaje de novo de genoma
- Resecuenciación de genoma
- RNA-Seq (la expresión génica, estructura exón-intrón, los pequeños perfiles de RNA, y las mutaciones)
- CHIP-Sec (interacción proteína-DNA)
- perfiles epigenéticos

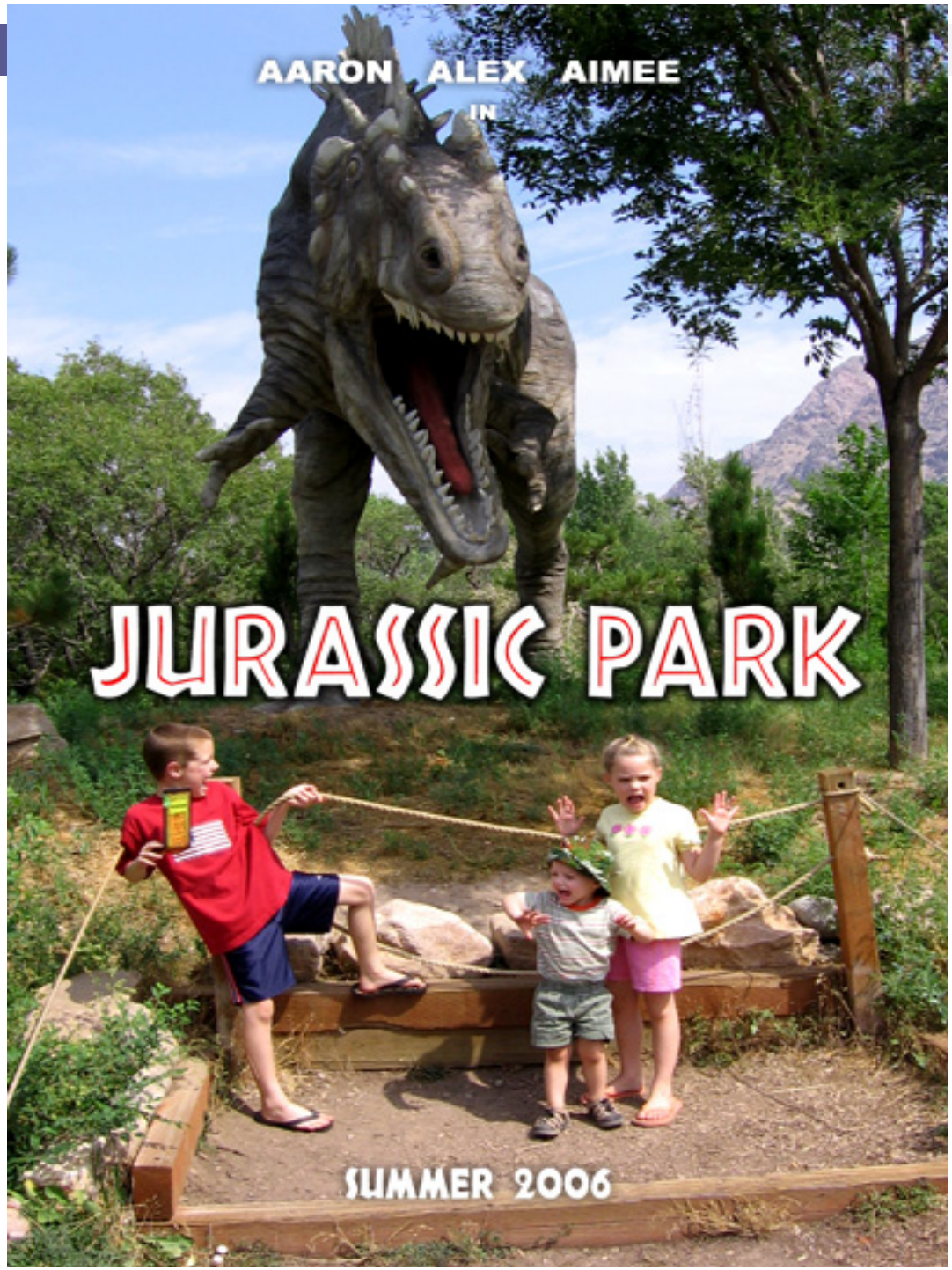
Los genomas antiguos resucitados

- **Estado de la muestra degradado →secuenciación mtDNA**
- **genomas nucleares de restos antiguos: oso de las cavernas, Mammuth, Neanderthal (10^6 bp)**



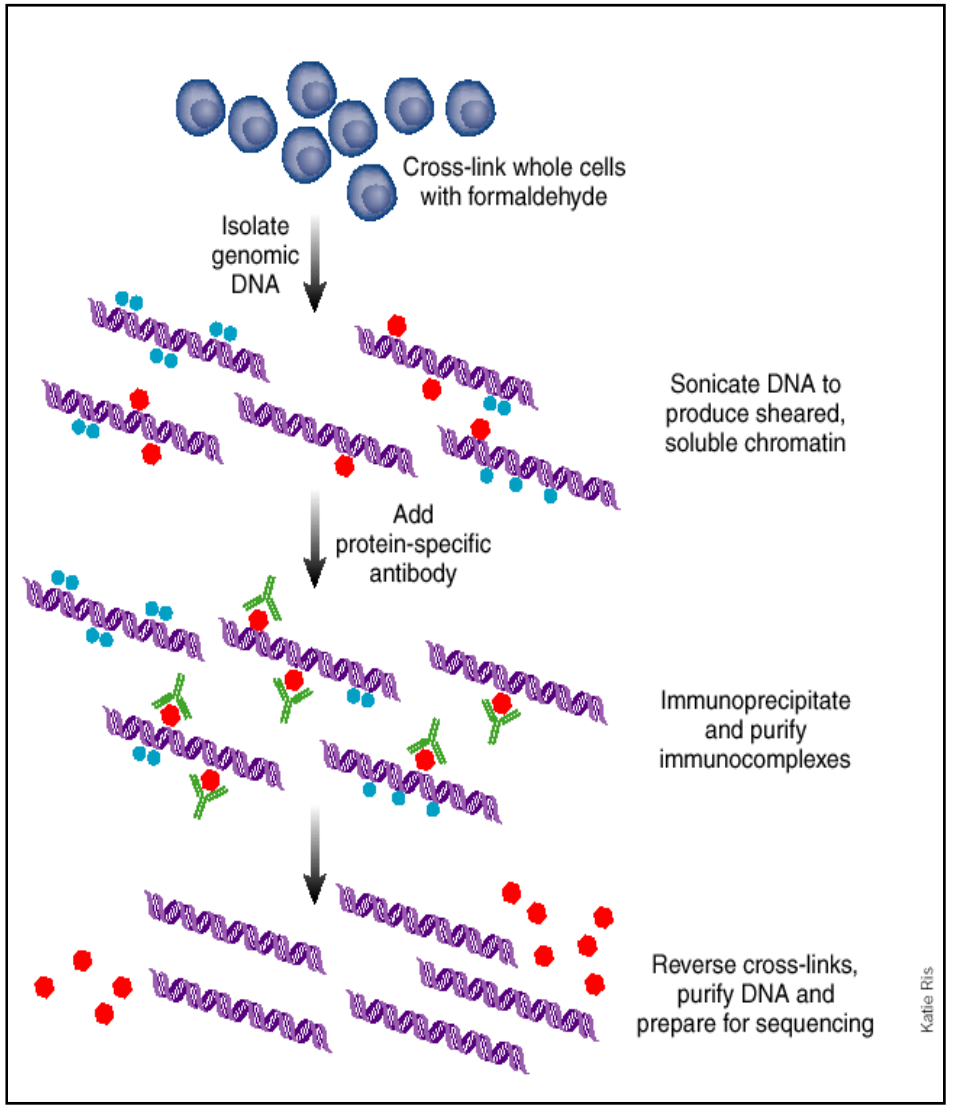


Los problemas de contaminación: los humanos modernos y coisolación del DNA bacteriano


AARON ALEX AIMEE
IN
JURASSIC PARK
SUMMER 2006

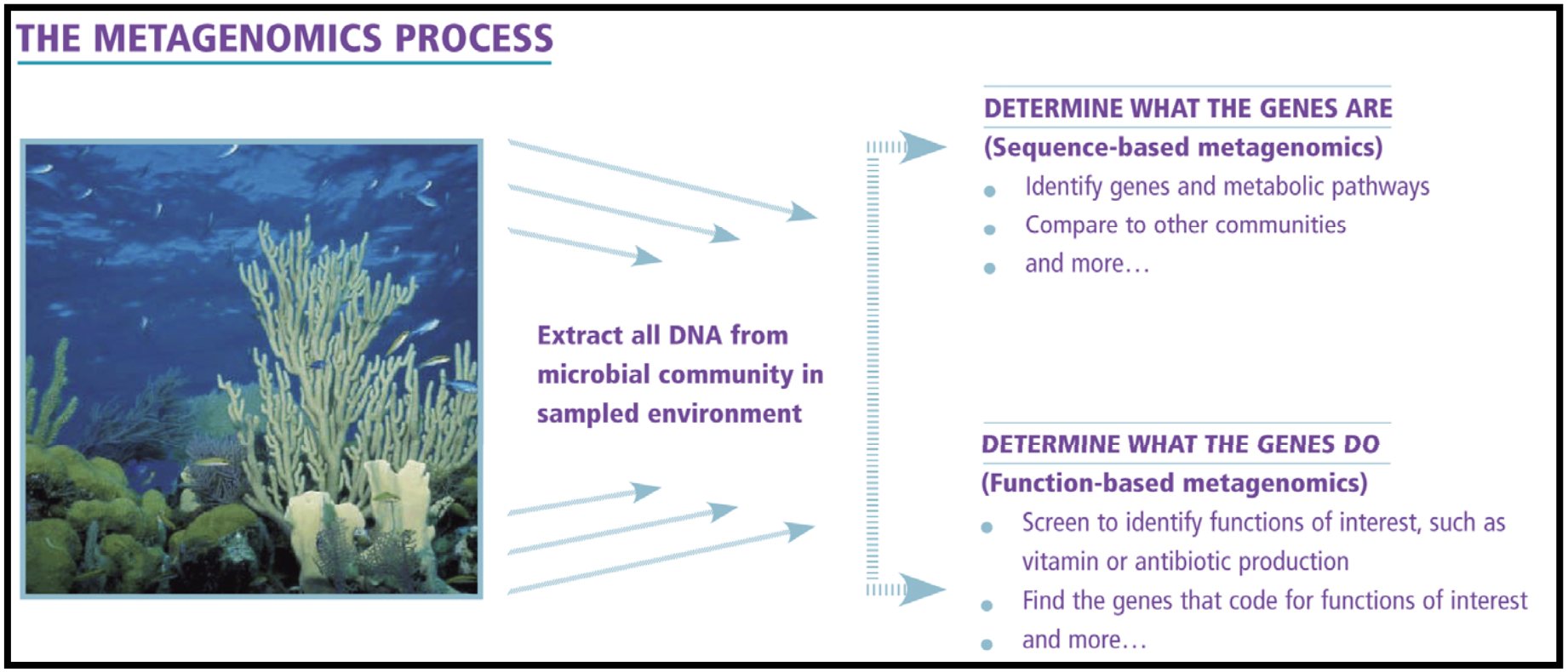
Epigenética: interacciones proteína-ADN a través de la secuenciación

- **pieza clave en la regulación de la expresión génica**
- **Recientemente, los estudios en todo el genoma de interacciones proteína-ADN**
- **Analizar factor de transcripción / estados de histona en el genoma humano**
- **Mejorar nuestra comprensión de la expresión génica en el contexto de los estímulos ambientales específicos**



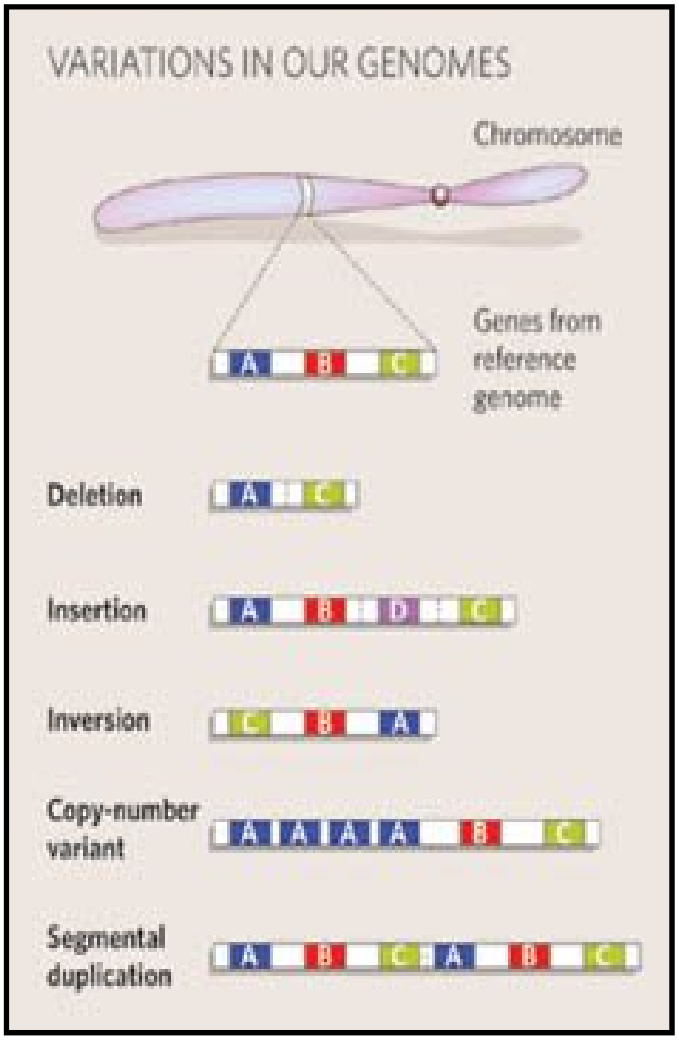
La metagenómica

- **Caracterización de la biodiversidad que se encuentra en la Tierra**
- **El creciente número de genomas secuenciados permite interpretar secuencias parciales obtenidas por muestreo directo de nichos ambientales.**
- **Ejemplos: lecho marino, el suelo, los arrecifes de coral, microbioma humano que puede variar según el estado de salud del individuo**



Definición de variabilidad en muchos genomas humanos

- variantes comunes aún no han explicado completamente la genética de enfermedades complejas →alelos raros también contribuyen
- También variantes estructurales, inserciones y deleciones grandes y pequeños
- Acelerar la investigación biomédica



Veremos

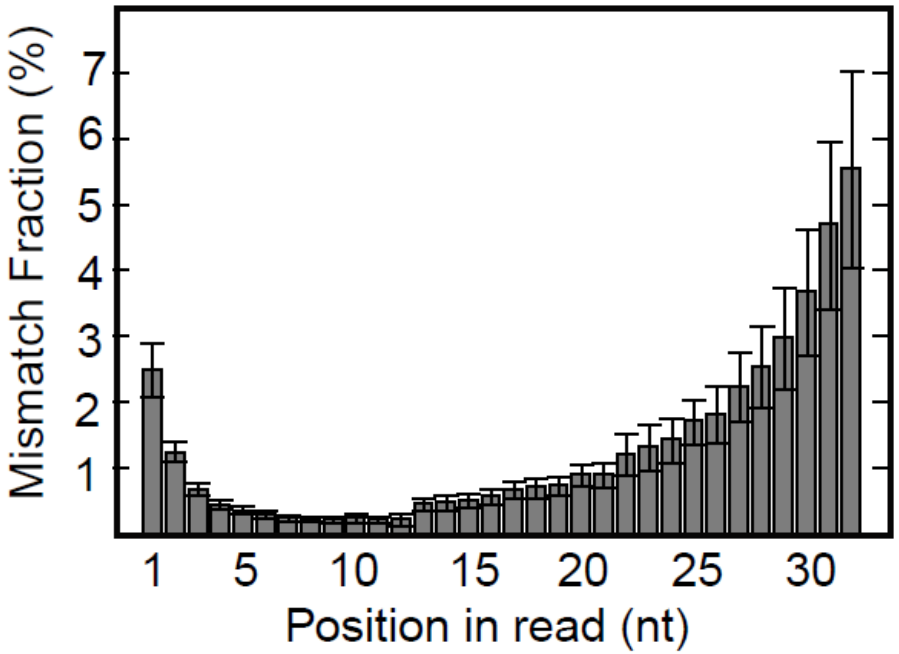
- Historia
- Plataformas NGS
- Aplicaciones
- **Retos bioinformáticos**

Cantidad enorme de datos

- Operación tipica con SOLiD:
  - ~ Archivo de imagen 2T
  - ~ archivo de texto 120G para extracción de conocimiento
  - ~ 75 M fragmentos cortos por muestra

Métodos eficientes para el almacenamiento y gestión de datos
Requiere cálculo paralelo de alto rendimiento

Tasas de error de secuenciación considerables



análisis de imágenes de alta calidad para identificación de bases

Retos bioinformáticos

- Métodos eficientes para almacenar, recuperar y procesar gran cantidad de datos
- Reducir los errores en el análisis de imágenes e identificación de bases (“base calling”)
- Métodos rápidos y precisos para alineación y montaje del genoma
- Nuevos algoritmos de extracción de conocimiento

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

Third Generation Sequencing : Single Molecule Sequencing

Helicos (BioSciences Corporation)

Helicos Genetic Analysis System

	Helicos
Read Length	35 bp
Throughput	35 Gb
Reads per run	600,000,000 - 1,000,000,000
Accuracy	97 %
Run Time	8 days

Workflow : Library preparation → Sequencing