

Introducción

Biocomputación
Grado en Bioquímica

¿Qué es la biocomputación?

Pipette biologist. Microscopy biologist. Cell culture biologist. Have you ever heard any of those job titles? No, of course not. All are biologists, because it is the questions you address that matter, not the tools you use, and **computational biologists are just biologists using a different tool.**

Florian Markowetz

All biology is computational biology

<https://doi.org/10.1371/journal.pbio.2002050>

El objetivo de esta asignatura:

Enseñar el uso de herramientas
computacionales para resolver problemas
biológicos

¿Qué es la bioinformática?

Técnicas computacionales para resolver problemas biológicos

Técnicas (computacionales):

- Almacenar datos biológicos en bases de datos y hacerlo accesible
- Visualizar datos y resultados
- Comparar secuencias (alineamiento)
- Computación y algorítmica para analizar datos
- Análisis estadístico de datos
- Métodos de aprendizaje automatizado (predicción)
- ...

¿Bioinformática?

Problemas biológicos:

- Detectar genes en una secuencia anónima
- Predicción de función de secuencias
- Predicción de estructuras secundarias y/o terciarias
- Analizar datos experimentales (secuenciación masiva, expresión génica etc.)
- ...

¿Biología computacional?

Las técnicas y objetivos

- Secuencias de genes y proteínas
- Secuencias de genomas completos
- Datos de expresión génica
- Estructura de proteínas y RNA no-codificante
- Interacción entre proteínas y rutas metabólicas
- Información epi-genética
- Regulación de la expresión génica (TFBS, islas CpG)

Biología molecular
Genética
Genómica
Biomedicina
...



**Administración
de los datos y
de información**

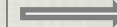
- Bases de datos
- Ontológicas



**Computación
Algorítmica**

- Programas de ordenador
- Servidores web

**Minería de
datos**



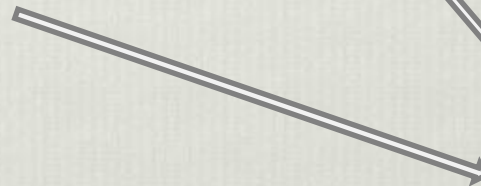
**Simulación
/modelos**



Datos



**Conocimiento
biológico**



Las cuestiones biológicas

Genotipo



Fenotipo

DNA
(genoma)



RNA



Proteína



Redes de
interacción



Células



Organismo /
Estado pato-
fisiológico

Genómica
comparativa
Epigenómica
Variación de
secuencia

Expresión génica,
expresión diferencial

Interacciones entre
proteínas
Redes de regulación
de la transcripción
Redes de
señalización
Metabolómica

Genética de
poblaciones
Genética humana
Estudios de
asociación

Detección de genes
Genes no-codificantes
Regiones reguladoras (sitios de unión a
factores de transcripción)

Estructura secundaria y/o terciaria
(plegamiento de la proteína)

Simulación del
funcionamiento
de una célula

alineamientos, bases de datos, búsqueda de patrones, aprendizaje automatizadas, métodos de redes

El ejemplo clásico: BLAST

La bioinformática aplica algoritmos computacionales a datos para responder a preguntas biológicas

Tenemos una secuencia de DNA o de proteína (frecuentemente anónima) como punto de partida:

¿A que otras secuencias se parece la secuencia de mi gen/proteína?

→ probablemente hay un gen con una secuencia “similar” bien caracterizado lo que permitiría obtener conocimiento acerca del gen estudiado

¿En que otras especies existe mi gen o la proteína?

→ responder a preguntas evolutivas

¿De que especie es?

→ en experimentos de metagenómica o interacciones entre especies – parasito/huésped

El ejemplo clásico:

Comparar la
secuencia de entrada
con **todas** las
secuencias de una
base de datos para
encontrar aquellas
secuencias (en la
base de datos) con
las que comparte un
**alto grado de
similitud de
secuencia**

Distribution of 27 Blast Hits on the Query Sequence

NG_001136 Homo sapiens gulonolactone (L-) oxidase, pseudogene (GU... S=98 E=3.8e-...

Color key for alignment scores

Query

1 450 900 1350 1800 2250

<40 40-50 50-60 60-80 >=200

FASTA (complete sequence)
FASTA (aligned sequences)
GenBank (complete sequence)
Hit Table (text)
Hit Table (CSV)
Text
XML
ASN.1

Available columns
Description
Max Score
Total Score
Coverage
E-value
Ident
Accession

Restore Defaults Ok Cancel

Descriptions

Sequences producing significant alignments

Select: All None Selected

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Max ident	Accession
☒ Rattus norvegicus GULO gene for L-gulonolactone oxidase, complete cds	850	2870	99%	0.0	96%	D12754.2
☒ Homo sapiens...						

Download GenBank Graphics Sort by: Query start position

Next Previous Descriptions

Rattus norvegicus GULO gene for L-gulonolactone oxidase, complete cds

Sequence ID: [dbj|D12754.2](#) Length: 6009 Number of Matches: 12

Range 1: 393 to 545 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
10 bits(366)	6e-36	135/162(83%)	9/162(5%)	Plus/Plus

Query 7 CTTGGTACCTGTGGCTAAACTTCCAGTCCCGTCTGCTGAGGTAAACCCAGAGCCGAGGT 66

Sbjct 393 CTTAGTATGGGTGGCTAAACCCAGTCTGCTGCTGAGGTAAACCCAGAGCCGAGGT 452

Query 67 TGCCTGACCACTGCATCTGCTGCTGCCAGGGCTTTGTTCAACTTCTGTGGGAAGCGTT 126

Sbjct 453 TGCCTGACCACTGC-----TACCCAGGGCTTTGTTCAACTTCTGTGGGAAGCGTT 503

Query 127 CAAGTCAAGTCTGCTGACCCCTGTTGATCGCTGGAATCATGTT 168

Sbjct 504 CAAGTCAAGTCTGCTGATCTCTGCTGATCTGGAATCATGTT 545

Range 2: 844 to 948 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
156 bits(357)	9e-35	98/105(93%)	0/105(0%)	Plus/Plus

Query 166 GGTCCATGGGTACAAAGGGGTCCAGTTCCTCAAACTGGGCGAAGACCTATGGCTGCAGTCC 225

Sbjct 844 GGTCCATGGGTACAAAGGGGTCCAGTTCCTCAAACTGGGCGAAGACCTATGGCTGCAGTCC 903

Query 226 AGAGATGTACTACCAGCCACCATCTGCGGGAGGTCAGAGAGGT 270

Sbjct 904 AGAGATGTACTACCAGCCACCATCTGCGGGAGGTCAGAGAGGT 948

Related Information

Gene - associated gene details

Map Viewer - aligned genomic context

GEO Profiles - microarray expression data

Bioinformática vs. Laboratorio húmedo

Nivel de usuario

Aplicar métodos desarrollados para responder a una pregunta biológica concreta

Bioinformática

¿Que función podría tener una secuencia anónima dada?

BLAST

Buscar (mediante alineamiento) si existe una secuencia similar con función conocida en una base de datos

Laboratorio húmedo

¿El paciente esta infectado por un parásito dado?

PCR

Intentar amplificar una secuencia del parásito para detectar su presencia

Bioinformática vs. Laboratorio húmedo

Nivel de usuario

Aplicar métodos desarrollados para responder a una pregunta biológica concreta

Bioinformática

¿Que función podría tener una secuencia anónima dada?

BLAST

Buscar (mediante alineamiento) si existe una secuencia similar con función conocida en una base de datos



Laboratorio húmedo

¿El paciente esta infectado por un parásito dado?

PCR

Intentar amplificar una secuencia del parásito para detectar su presencia



Bioinformática vs. Laboratorio húmedo: Misma finalidad (responder a preguntas biológicas) pero diferentes métodos

Bioinformática vs. Laboratorio húmedo

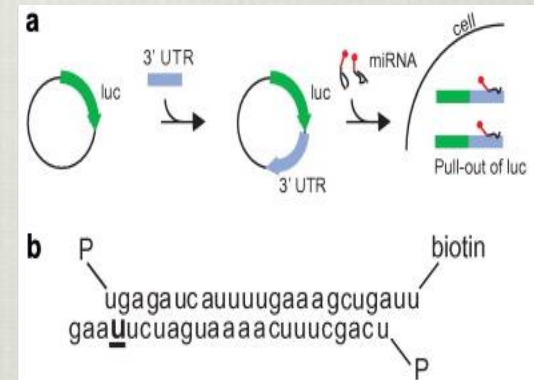
Nivel de desarrollo tecnológico: métodos nuevos para para responder a nuevas preguntas biológicas

¿A que genes regula un microRNA dado?

Predicción de dianas a partir de las secuencias basándose en modelos termodinámicos
Implementación del modelo en un algoritmo computacional

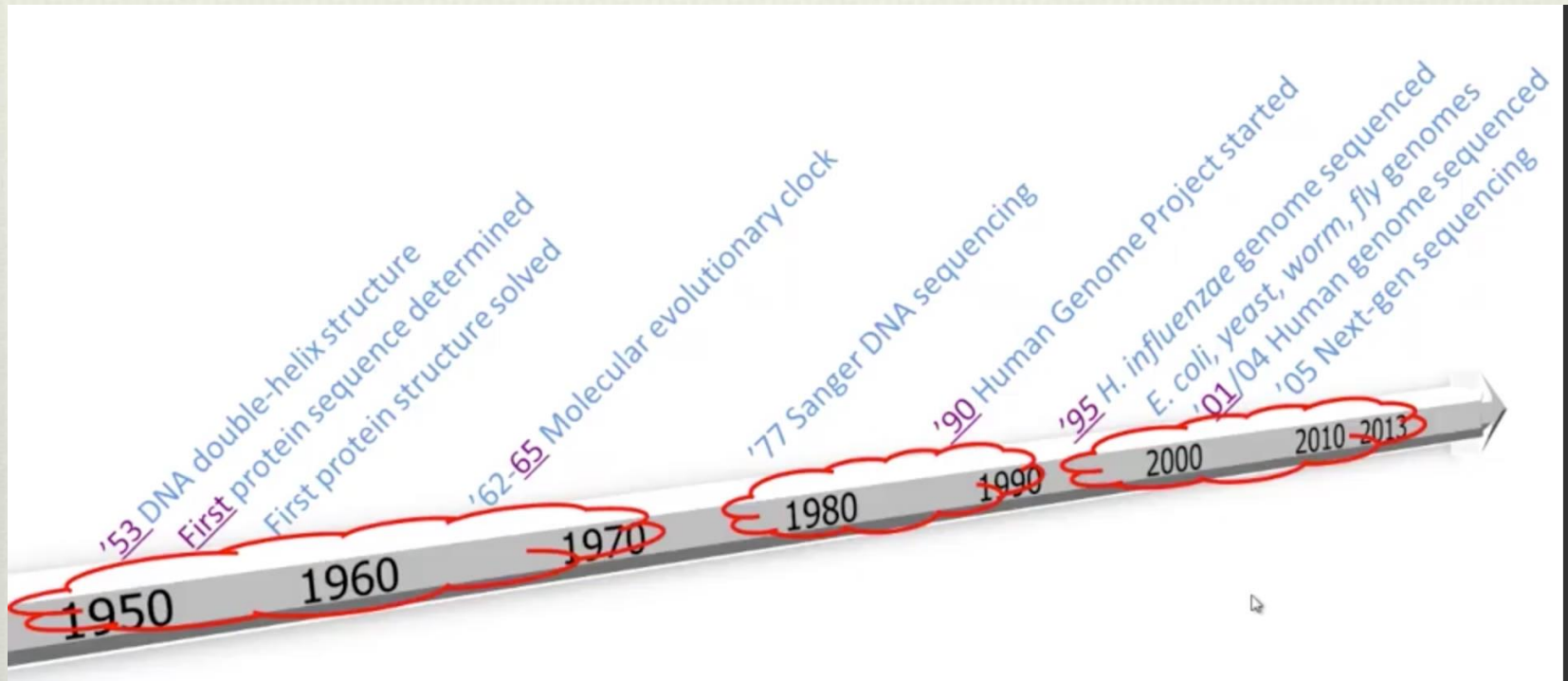
HTT 3' UTR	5' ...GAUUGC UUUUGUUUU--CCUGCUGG...
miR-214	3' UGACGGACAGACACGGACGACA

Biotinilar los microRNAs y capturar las uniones entre microRNAs biotinilados y los mRNAs
Desarrollar nuevos métodos bioquímicos



Bioinformática vs. Laboratorio húmedo: Misma finalidad (responder a preguntas biológicas) pero diferentes métodos

Orígenes y evolución de la bioinformática



La bioinformática tiene sus orígenes en la biología molecular de los años 60 del siglo pasado (aunque en aquel entonces nadie hablaba de bioinformática)

Dos eventos tienen especial relevancia para la evolución la bioinformática:

- Proyecto genoma humano
- La aparición de métodos de secuenciación masiva

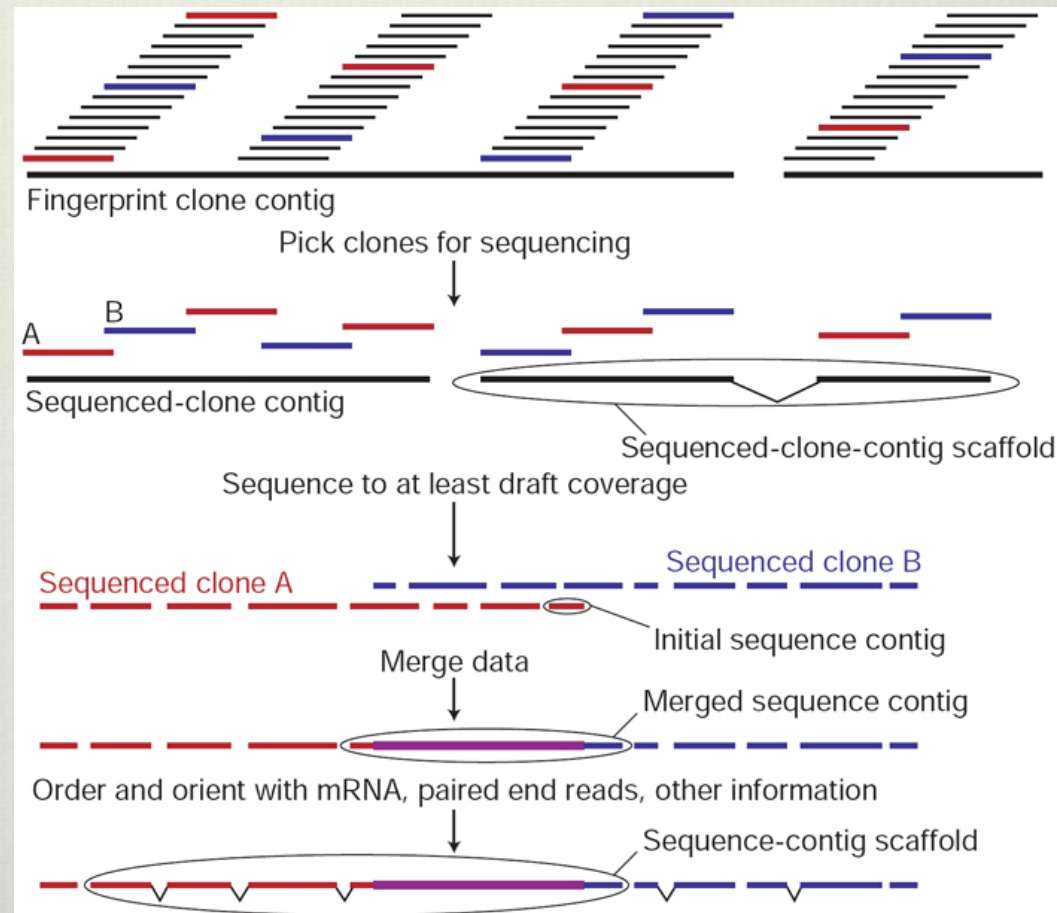
Proyecto genoma humano

El proyecto empezó en 1984 y terminó en 2003 con la publicación de la secuencia de la parte de eucromatina del genoma humano

- Determinación de la secuencia de bases nitrogenadas que forman el ADN humano
- Identificación de los genes en el genoma humano
- Hacer los resultados accesibles libremente

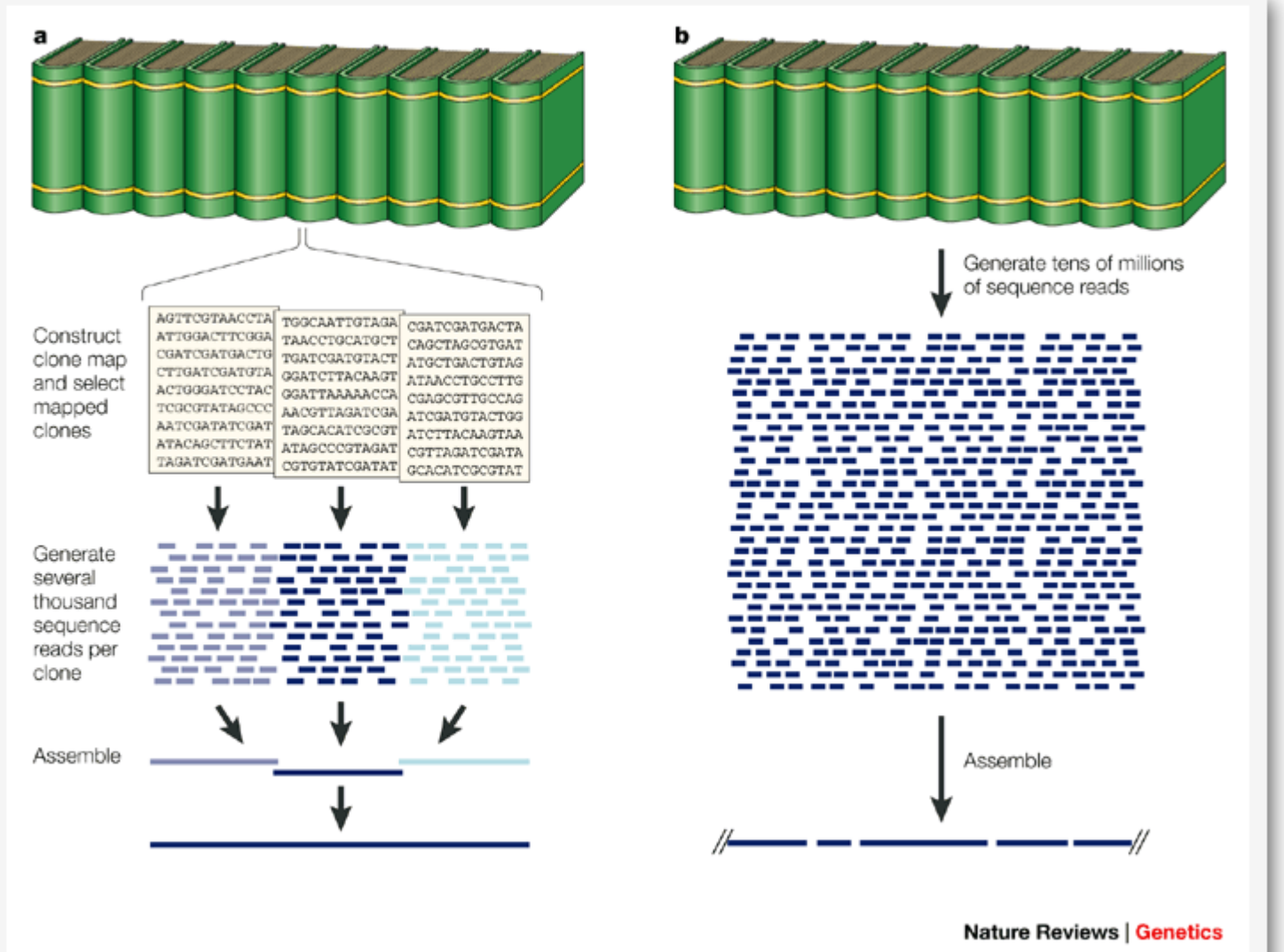
Importancia de la bioinformática

- Selección de los clones
- Ensamblar los contigs

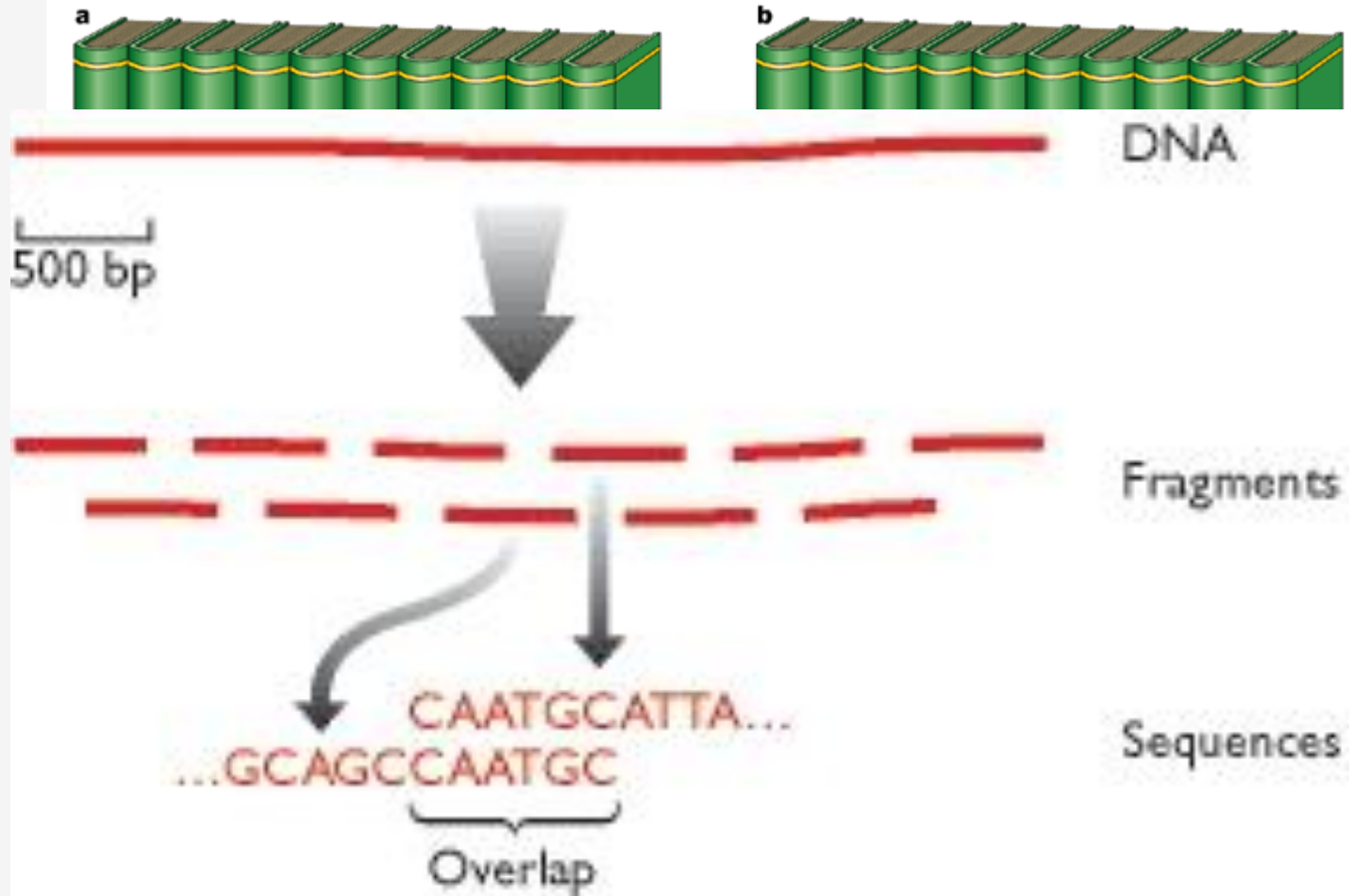


<https://www.nature.com/scitable/content/Initial-sequencing-and-analysis-of-the-human-16729>

Proyecto genoma humano



Proyecto genoma humano



Secuenciación masiva

illumina®

<https://www.illumina.com/>



Oxford
NANOPORE
Technologies

<https://nanoporetech.com/products>

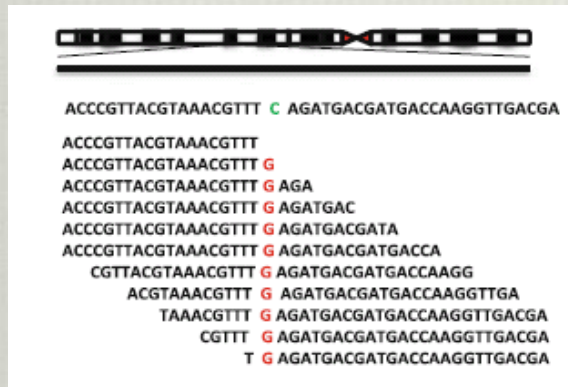
Secuenciación masiva

	SANGER	SECUENCIACIÓN MASIVA		
	Di-deoxy terminator	Roche 454 GS FLX (PS)	Illumina HiSeq 2000 (RT)	SOLID V4 (SBL)
Salida por proceso	1.6 Mb	600 Mb	200 GB	100 GB
Tiempo/Proceso	1h	10 h	9 d	11 d
Longitud media "reads"	800 pbs	400 pb	100 pb	75 pb
Salida por día	38.4 Mb	1.44 GB	22.2 GB	9 GB
Usos frecuentes	-	Secuenciación de novo Captura de exones	Resecuenciación Captura de exones Metagenómica	Resecuenciación Captura de exones Metagenómica

Secuenciación masiva

APLICACIONES

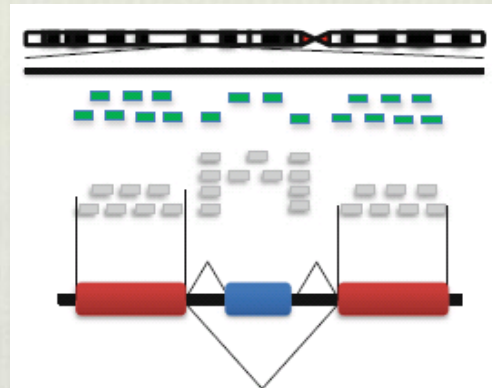
DNA sequencing



Detect sequence variants (SNV/SNPs, indels)

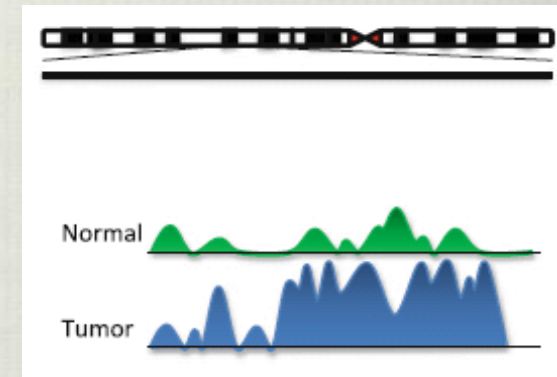
- Individual genomes
- Cancer genomes

Gene Expression



- Transcription
- small RNAs
- TF binding to DNA

Epigenomics

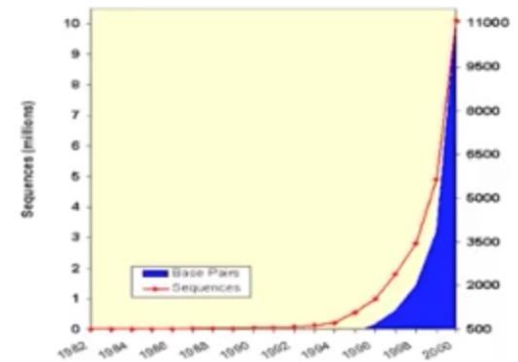


- DNA methylation
- Histone modifications

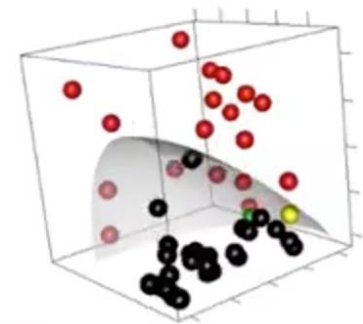
Los retos técnicos de la bioinformática

Opportunities and challenges hand-in-hand: the driving forces of bioinformatics

- High-throughput data
 - Huge amount
 - Explosive growth
 - Low signal-to-noise ratio
 - Multiple types
- Requirements for the methods
 - Data needs to be stored in efficient **ontology-based database** systems
 - The huge amount of data requires **efficient** algorithms
 - Exponential growth requires **scalable** methods
 - The low signal-to-noise ratio requires **accurate** methods
 - Multiple types of data require data **integrative** methods



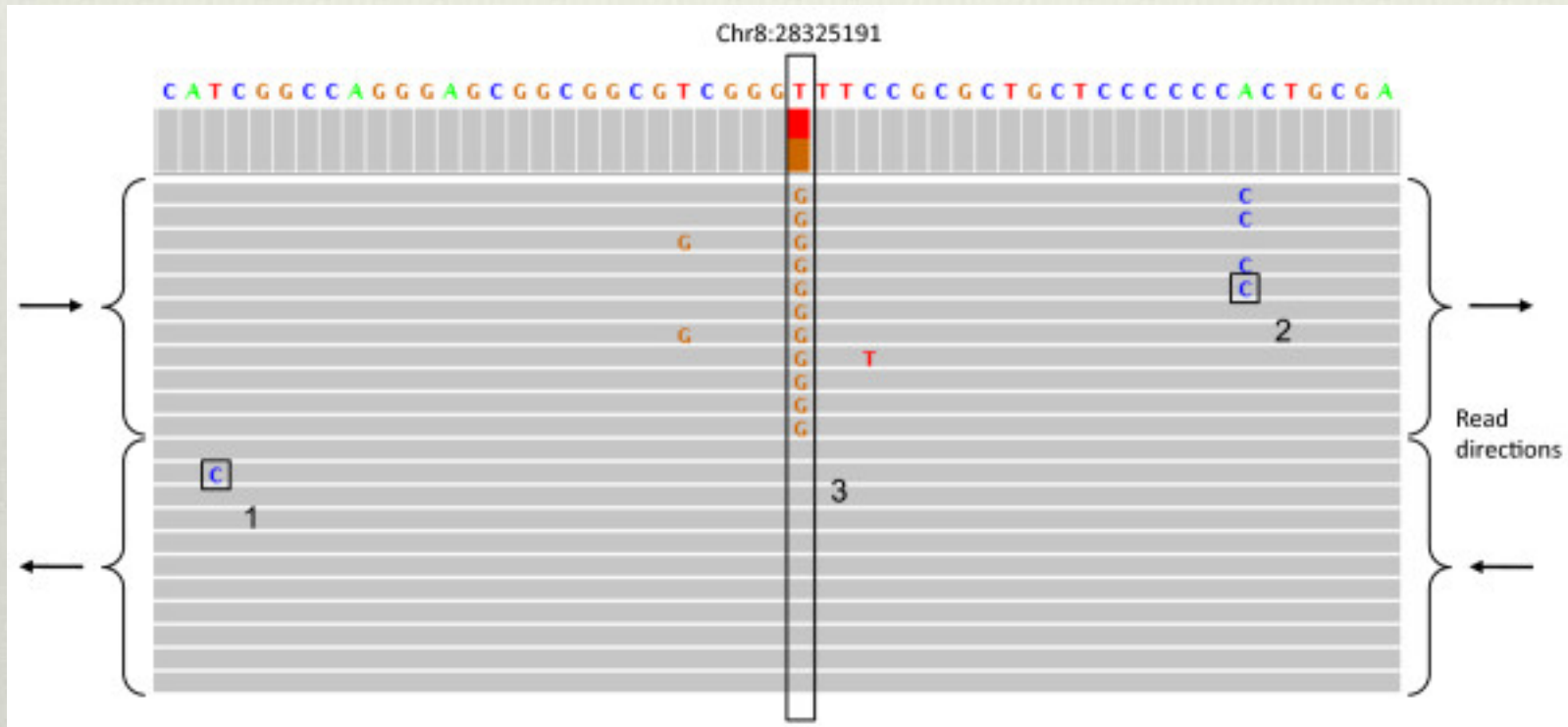
<ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt>



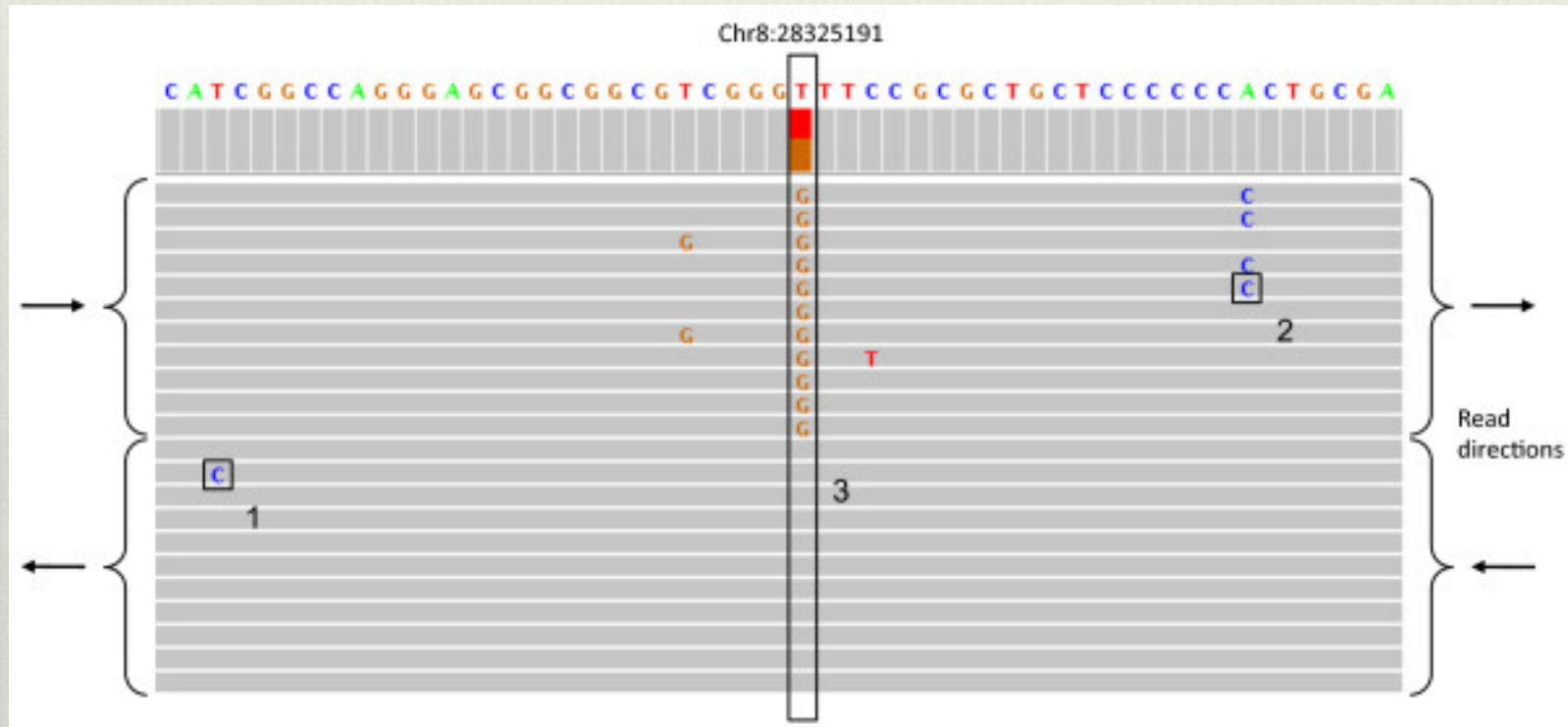
12:08

19:08

Los retos técnicos de la bioinformática



Los retos técnicos de la bioinformática

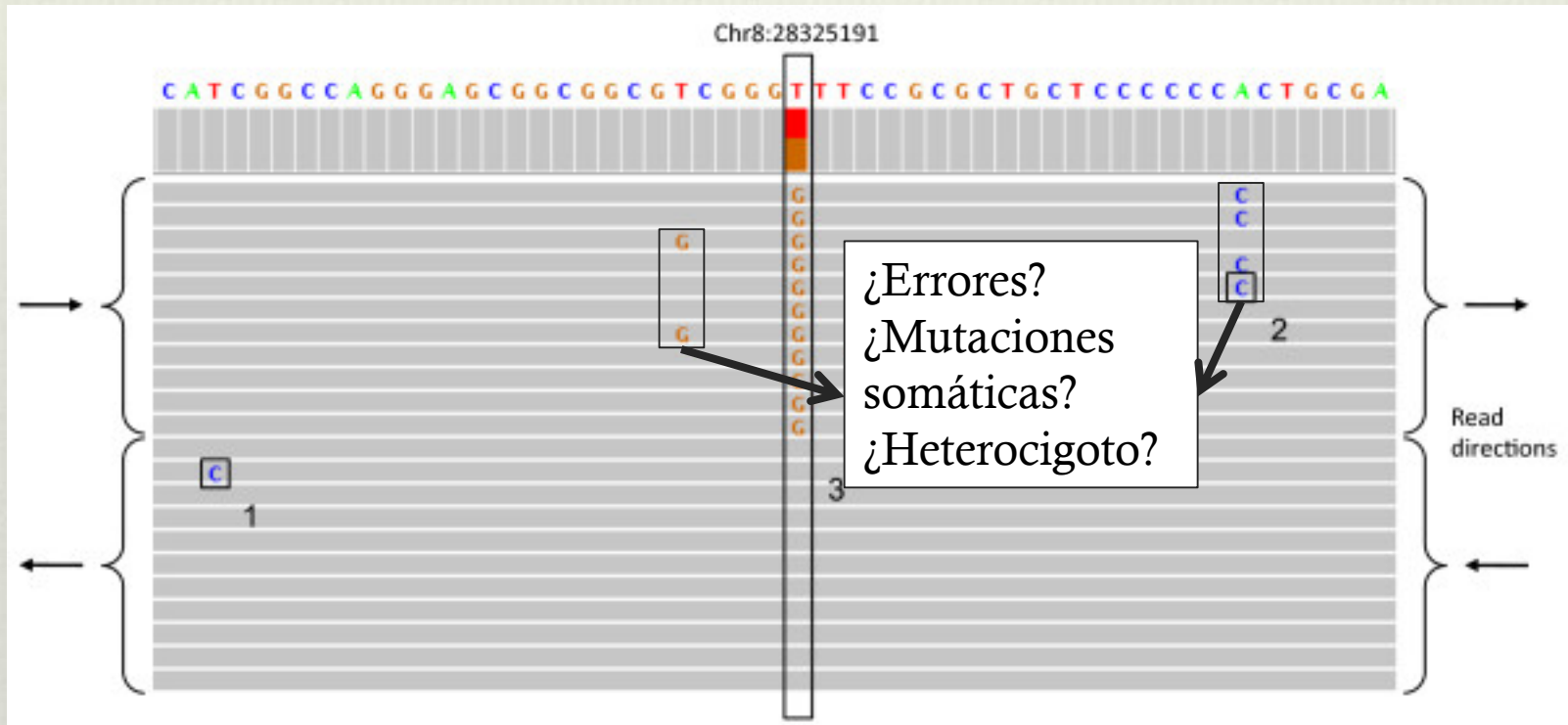


Heterozygote G/T

Detect germline variants
Inferential statistics

- Bayesian
- Frequentist

Los retos técnicos de la bioinformática



Se requiere métodos altamente específicos para no producir un elevado número de falsos positivos

Evolución de los retos técnicos

Años 90, principios 2000

Pocos datos experimentales →
La bioinformática intentaba
predecir o detectar elementos
genómicos o propiedades
epigenéticas dentro de
secuencias anónimas

- Predicción de TFBS
- Predicción de genes
- Predicción de estados de metilación
- Detección de islas CpG



Análisis de metilación

Desde ~2005

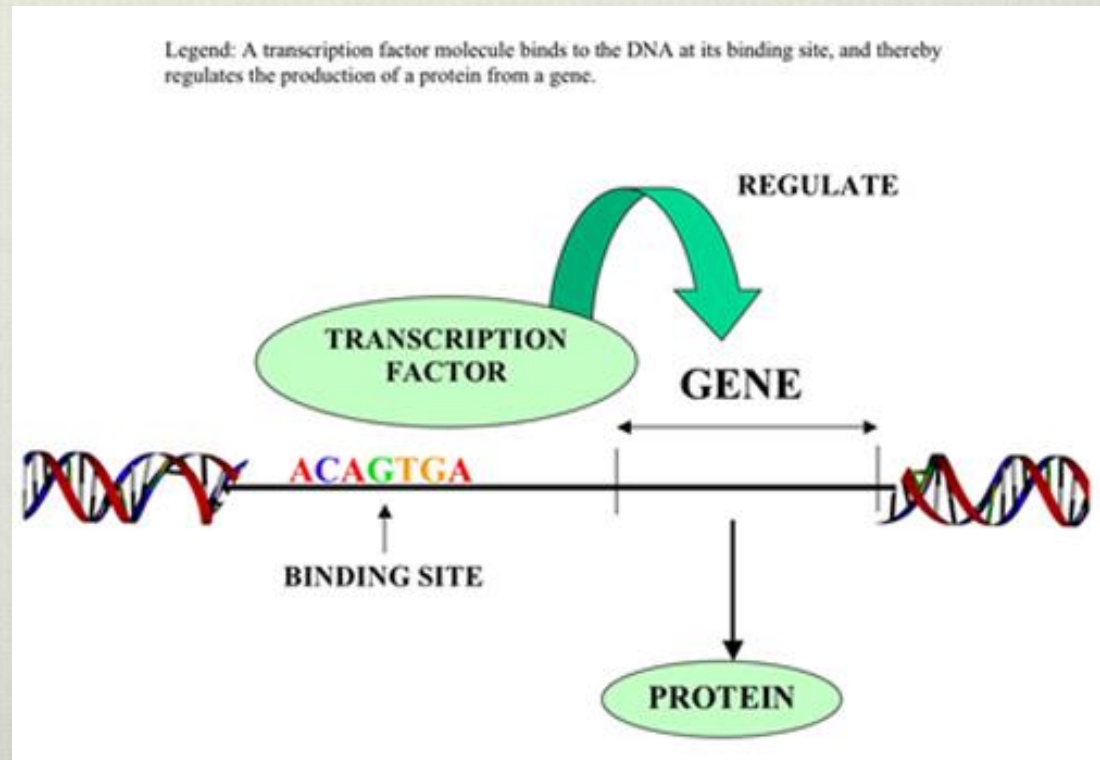
Aluvión de datos generados
mediante secuenciación masiva
(secuenciar genomas, expresión
génica, detectar modificaciones
químicas en el DNA y RNA,
etc...)

- Desarrollar algoritmos y métodos para el análisis de estos datos (exigentes debido a la enorme cantidad de datos)
- Bases de datos para almacenar esta cantidad de datos
- Analizar datos de secuenciación masiva para determinar los estados de metilación



Ejemplo 1: Sitios de unión a factores de transcripción

- Un factor de transcripción (TF) es una proteína que se une a una secuencia de DNA específica
- Controlan la expresión génica
- Pueden tanto promover (activadores - *activator*) o bloquear (represores-*repressor*) la unión de la polimerasa de RNA
- Estas proteínas tienen dominios específicos que les permite unirse al DNA
- Los TF se unen o al promotor o a un *enhancer* (“potenciador”)



Ejemplo 1: Sitios de unión a factores de transcripción

Antes de disponer de métodos de alto rendimiento

Generalizar las secuencias a las que se puede unir un TF (PFM)

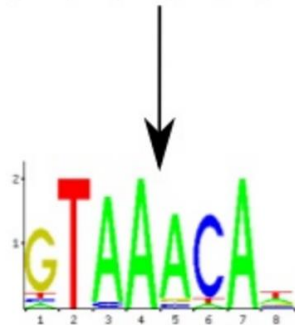
→ Buscar estas secuencias en una secuencia anónima

Known binding sites:

GTAACAAT
GTAAACAT
GTAAACAA
GTAAACAT
GTAAACAA
GTAAACAC
GTCAACAG
GTAAACAT
GTAAACAA
GTAAACAT
TTAAGTAA
ATAAACAA
CTAAACAG
GTAAACAT
GTAAACAA
GTAAACAT
GTAAACAC
GTAAACAT
GTAAACAT
GTAAACAG

Position Frequency Matrix:

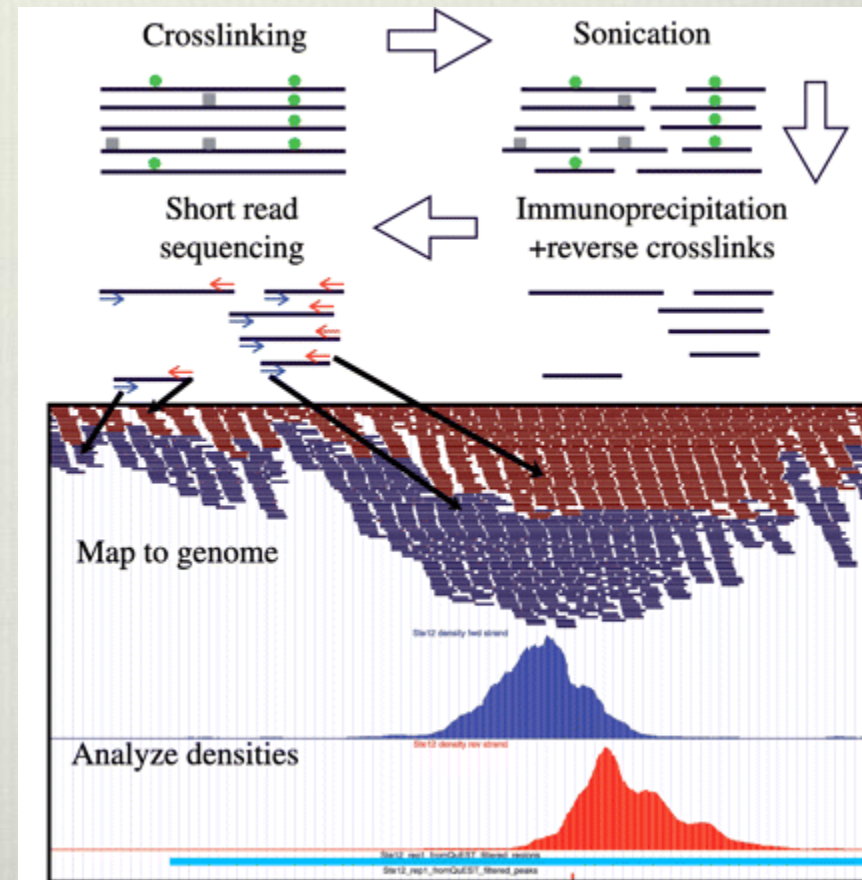
A [10 0 190 210 180 15 210 70]
C [10 0 20 0 15 180 0 25]
G [175 0 0 0 15 0 0 35]
T [15 210 0 0 0 15 0 80]



Elaine R Mardis
ChIP-seq: welcome to the new frontier
Nature Methods

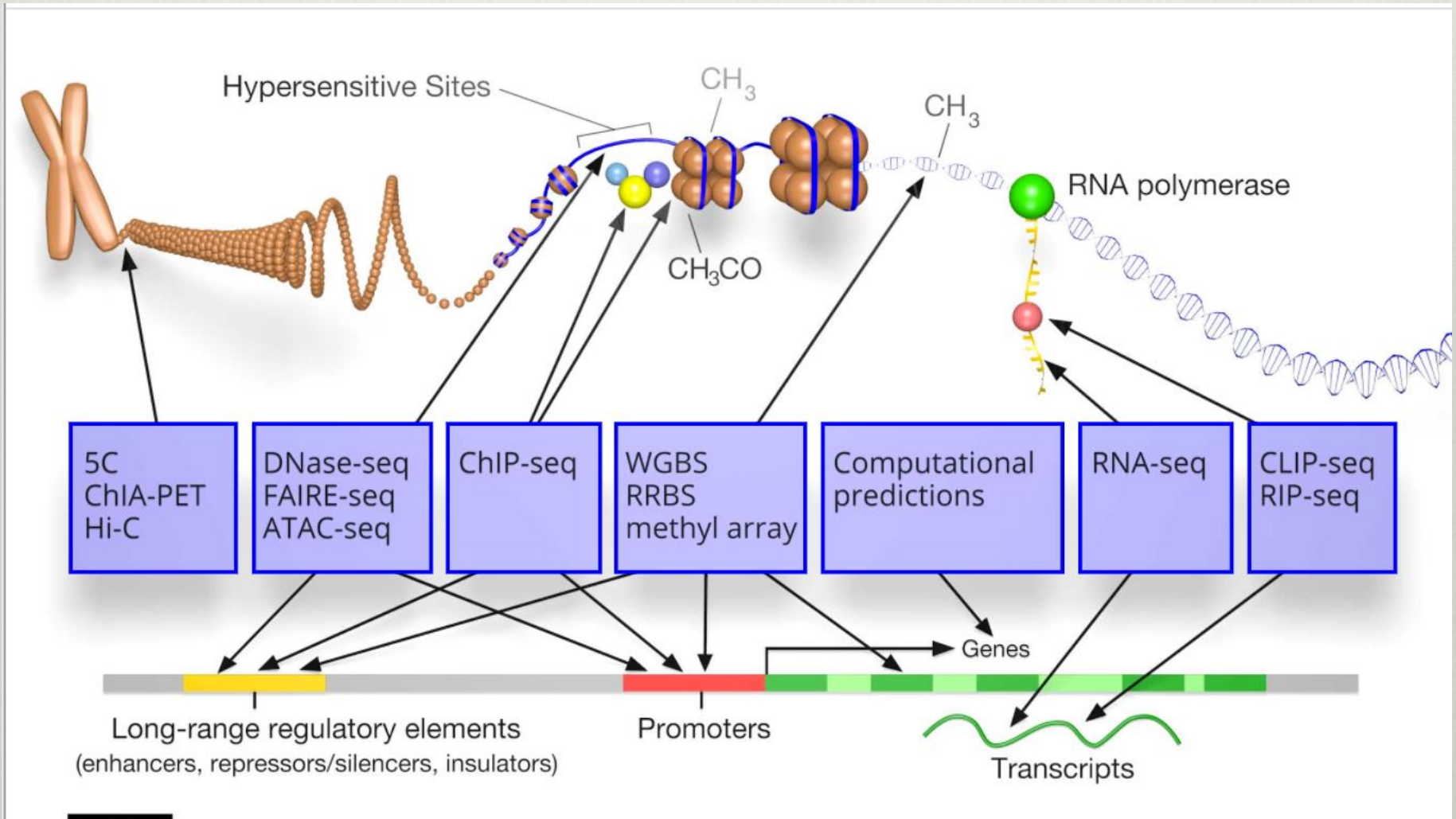
Detección experimental

1. chromatin immunoprecipitation (extraer /enriquecer secuencias de DNA a la que se une la proteína)
2. Detectar las secuencias a las que se une la proteína en el genoma



Los retos conceptuales: ENCODE

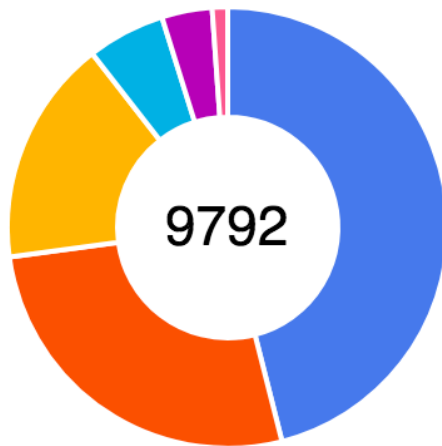
ENCODE: ¿que parte del genoma es funcional?



Los retos conceptuales: ENCODE

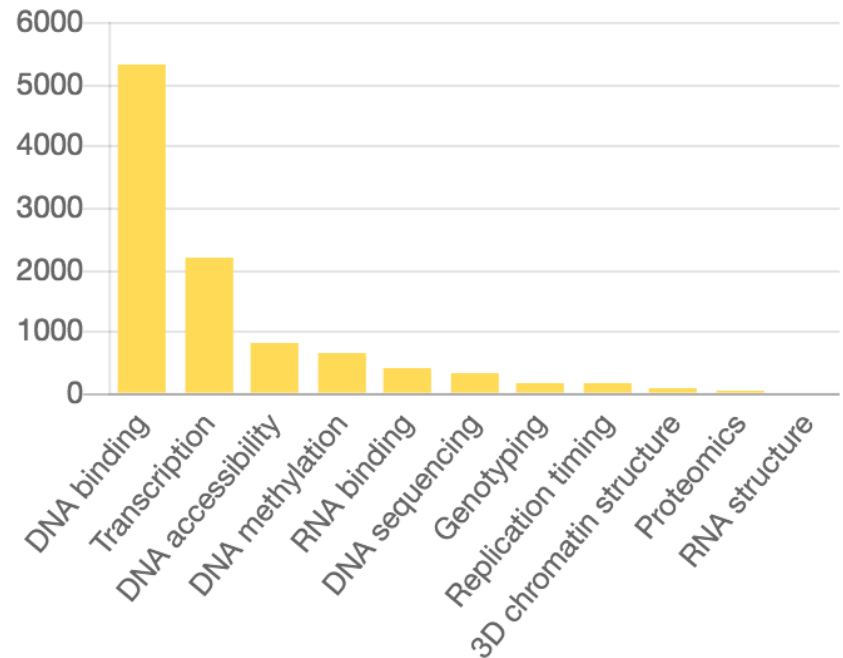
ENCODE: se analizaron casi 10,000 muestras mediante métodos de alto rendimiento

Biosample Type



- immortalized cell line
- tissue
- primary cell
- in vitro differentiated cells
- stem cell
- induced pluripotent stem cell line

Assay Categories



Los retos conceptuales: ENCODE

ENCODE: 13 publicaciones en 2012

<http://www.nature.com/encode/#/threads/transcription-factor-motifs/dna-elements>

ARTICLE

doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

Los retos conceptuales: 1000 genomas

Proyecto 1000 genomas: determinar toda la variación en las poblaciones humanas



<http://www.internationalgenome.org/home>

Los retos conceptuales: 1000 genomas

Proyecto 1000 genomas: determinar toda la variación en las poblaciones humanas

ARTICLE

OPEN

doi:10.1038/nature15393

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.