

Perspectivas en Bioinformática

Prof. Dr. José L. Oliver

Los orígenes...

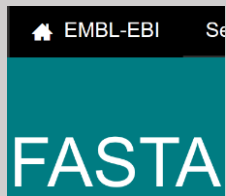
- La bioinformática que hoy conocemos ha surgido en torno a la genómica, jugando un papel clave en su desarrollo
- Pero sus fundamentos conceptuales vienen de mucho más atrás: Biología teórica, Biología computacional, Biocomputación...
- **La información es un concepto fundamental en Biología**
- Primero se descubrió su **transmisión**: leyes de Mendel
- **El contenido en información es lo que más distingue a un ser vivo de algo que no lo es** (Müller, 1916)
- **What is Life?** (Schrödinger, 1943): **Cristal aperiódico**, portador de información

Los orígenes...

- **Biología Molecular:** moléculas informacionales (ADN y proteínas) -> **la información biológica es digital**, no analógica
- **Un ser vivo no es más que la realización de un programa genético** (Waddington, Longuet-Higgins, Jacob, Luria...)
- Secuenciación de ADN y proteínas: el estudio del **contenido en información** de los seres vivos se convirtió en un tópico del mayor interés
- Pero el lanzamiento del **Proyecto Genoma** pronto produjo una acumulación explosiva de secuencias que reclamó la atención prioritaria de los bioinformáticos...

El desarrollo...

Secuenciación de proteínas y ADN, primeras bases de datos: EMBL-Bank (Heidelberg, 1980), GenBank (Los Alamos, 1982)...



El desarrollo...

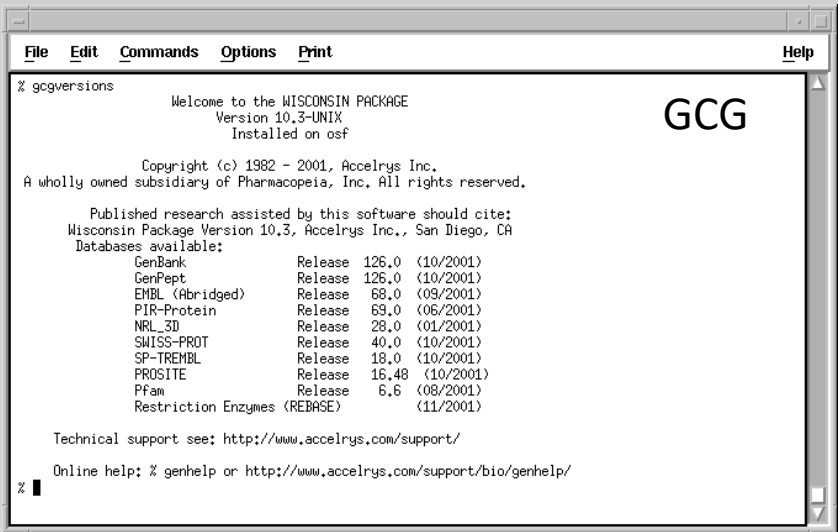
Swiss-Prot → UniProt:



3D structures: PDB, Cambridge (UK), Brookhaven (USA), 1971



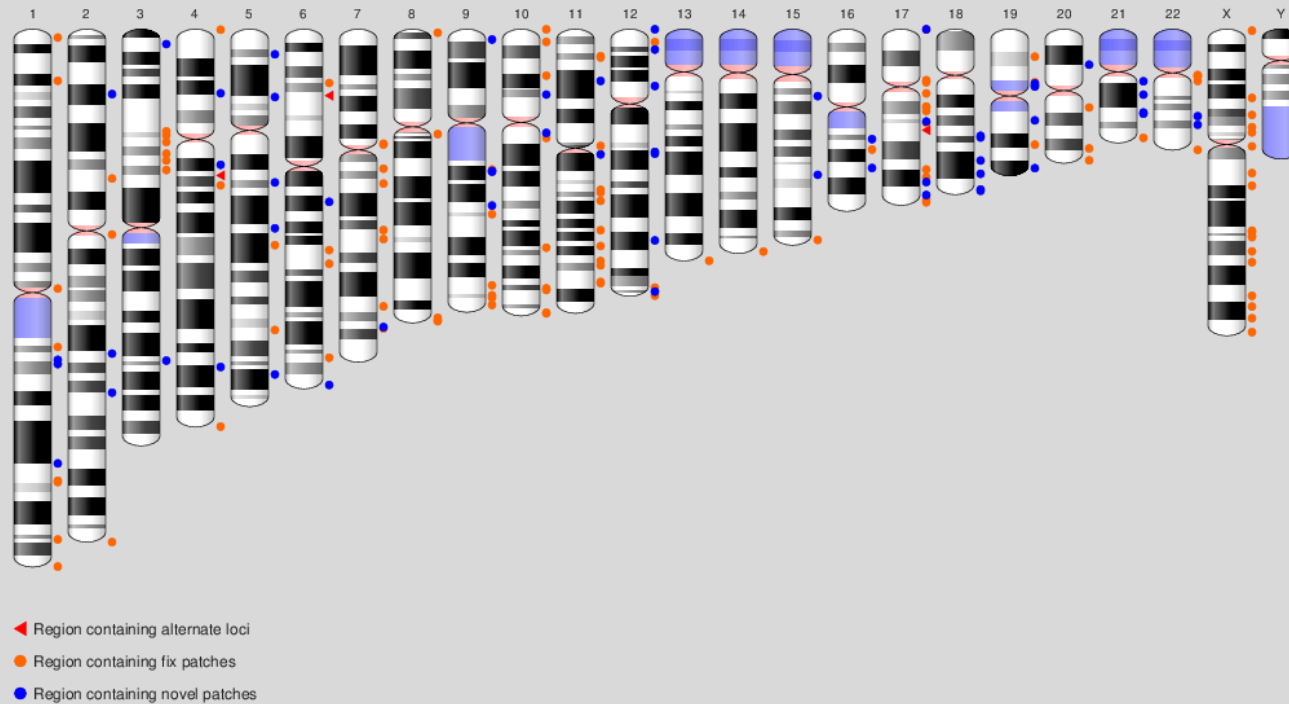
Software para análisis de secuencias: Genetic Computer Group (Wisconsin, 1982), EMBOSS (EMBL, 2000)





El Proyecto Genoma Humano

- Fue una iniciativa internacional lanzada en 1990 'to identify and map **all of the genes** of the human genome from both a **physical** and a **functional** standpoint'
- La primera **secuencia de referencia** del genoma humano se publicó en 2003



Hoy ya vamos por el ensamblado de referencia hg38 (2013):

https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39

GRCh38.p13

Description: Genome Reference Consortium Human Build 38 patch release 13 (GRCh38.p13)

Organism name: Homo sapiens (human)

BioProject: PRJNA31257

Submitter: Genome Reference Consortium

Date: 2019/02/28

Assembly type: haploid-with-alt-loci

Release type: patch

Assembly level: Chromosome

Genome representation: full

RefSeq category: reference genome

There are already 227 assemblies for this organism!

Global statistics	
Number of regions with alternate loci or patches	358
Total sequence length	3,099,706,404
Total ungapped length	2,948,583,725
Gaps between scaffolds	349
Number of scaffolds	472
Scaffold N50	67,794,873
Scaffold L50	16
Number of contigs	998
Contig N50	57,879,411
Contig L50	18
Total number of chromosomes and plasmids	24
Number of component sequences (WGS or clone)	35,613

Contigs are continuous stretches of sequence containing only A, C, G, or T bases without gaps

Scaffolds are created by chaining contigs together using additional information about the relative position and orientation of the contigs in the genome

Alineadores gráficos (como HISAT2 and HISAT-genotype) -> representaciones gráficas del genoma, capaces de incorporar todas las variantes, tanto puntuales como estructurales, en un solo **pan-genoma**.

El desarrollo explosivo...

Next Generation Sequencing (NGS) -> Secuenciación de genomas a gran escala



454
Pyrosequencing (PS)



Illumina
Reversible Termination (RT)



SOLID
Sequencing by Ligation (SBL)

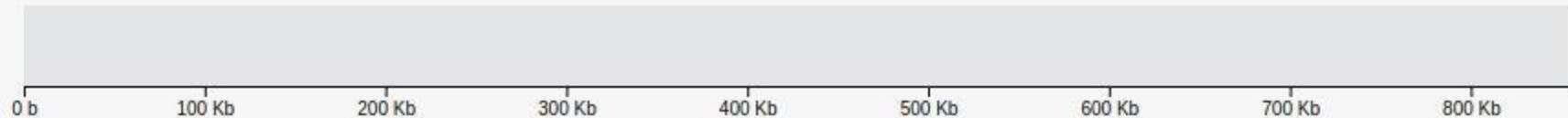
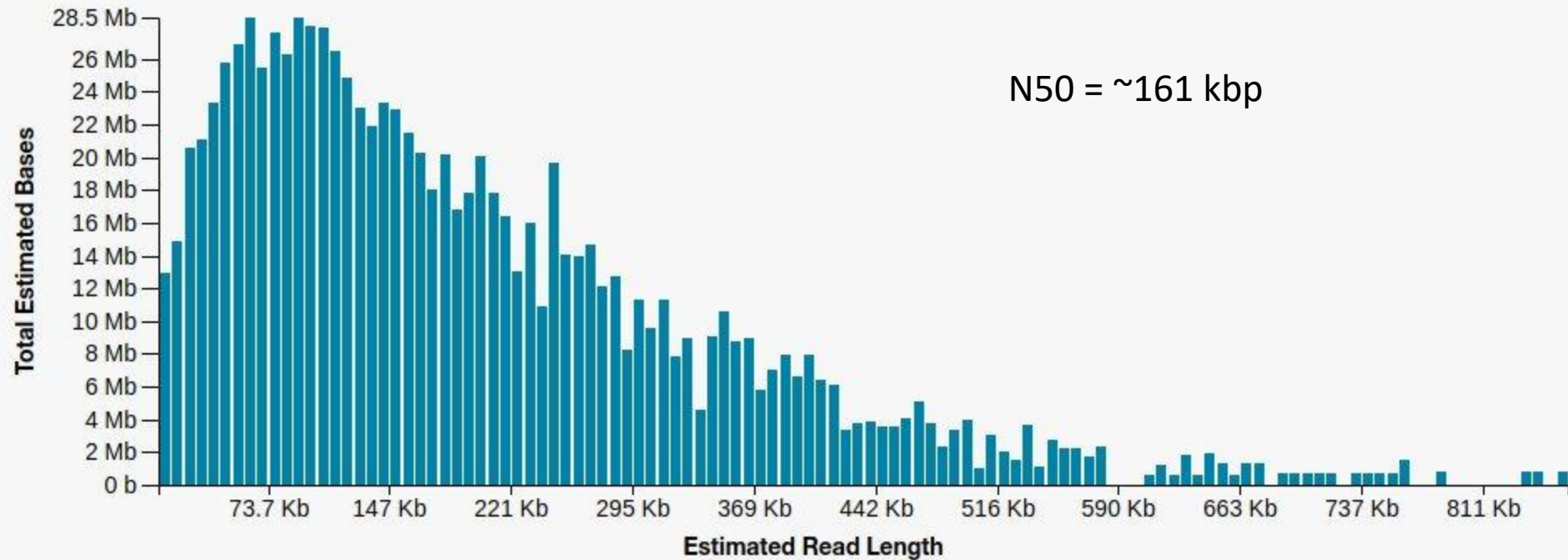


<http://www.ugr.es/~oliver>

Read Length Histogram

Summary read length distribution

Estimated N50: 160.96 Kb



Read Lengths

Read Counts

Reset

Telomere-to-telomere assembly of a complete human X chromosome

Date: 16th August 2019

Source: BioRxiv

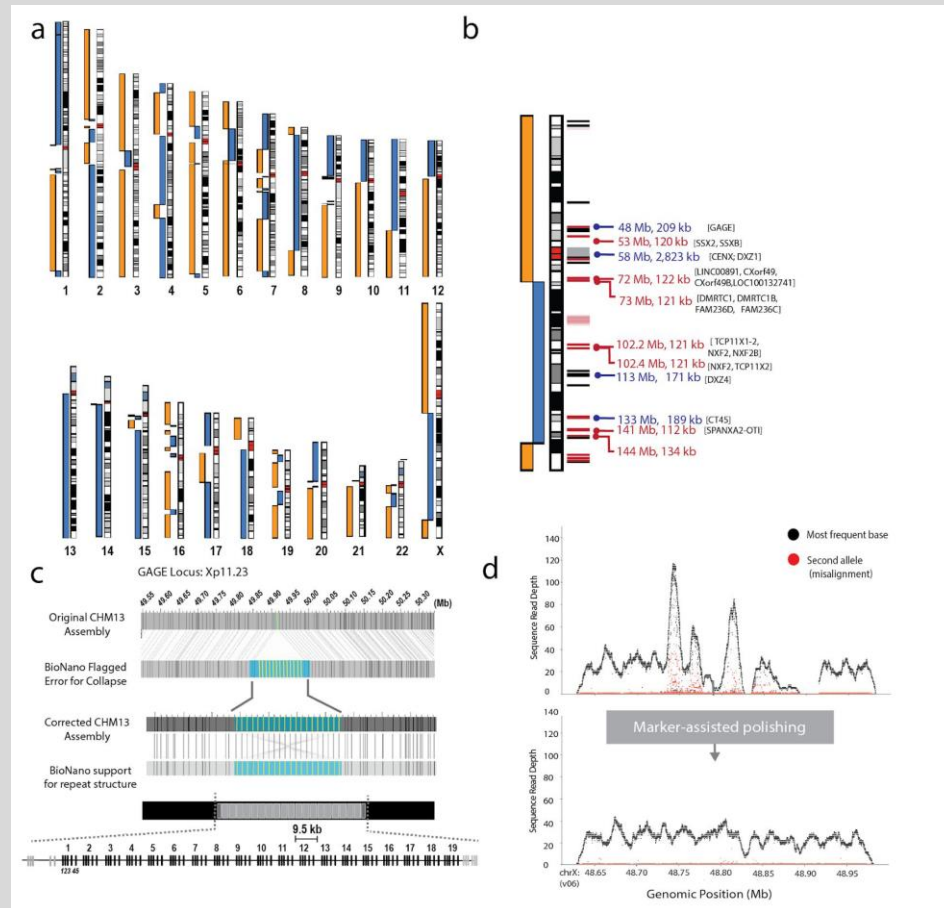
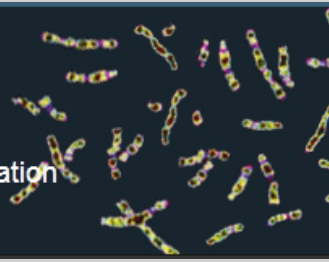


Figure 1. CHM13 whole-genome assembly and validation. (a) *De novo* assembly of the CHM13 genome using 39× of ultra-long Nanopore sequencing combined with 70× PacBio sequencing. Gapless contigs are illustrated as blue and orange bars next to the chromosome ideograms (two colors used only to better highlight contig breaks). Several chromosomes are broken only at centromeric regions (e.g. chr10, chr12, chr18, etc.). Large gaps between contigs (e.g. middle of chr1) indicate sites of large heterochromatic blocks or rDNA arrays where no GRCh38 reference sequence is available. (b) The X chromosome was selected for manual assembly, and was initially broken at three locations: the centromere (artificially collapsed in the assembly), a large segmental duplication (DMRTC1B, 120 kbp), and a second segmental duplication with a paralog on chromosome 2 (134 kbp). The relative placement of gaps in the GRCh38 reference are shown in the annotation track in black, known segmental duplications that are in red (with segmental duplications with paralogous sequence on the Y chromosome indicated in pink). Positions of repeats larger than 100 kb are indicated with the repeat sizing for use under a CC0 license.

T2T genomics!

1000 Genomes

A Deep Catalog of Human Genetic Variation



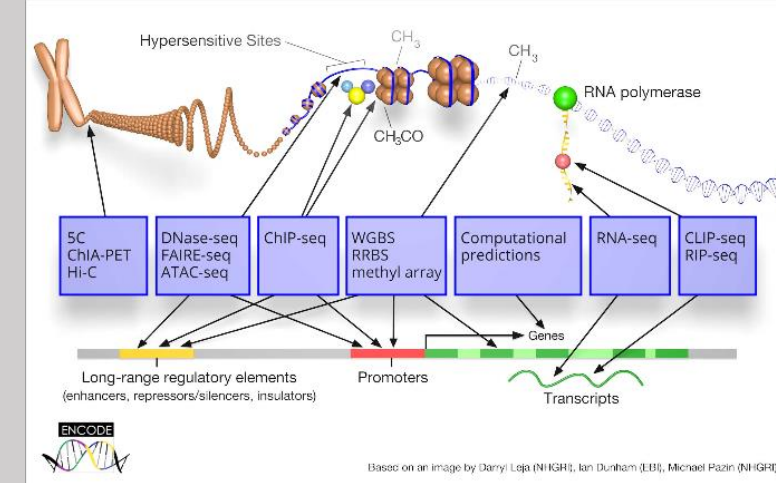
- Caracterización de la variación genética en el genoma humano
- La mayor parte de las variantes genéticas se localizan en las regiones no-codificadoras del genoma → elementos reguladores
- Cada ser humano lleva por término medio 200-300 variantes de pérdida de función (LOF)
- Una variante cada 20 bp!



ENCODE
Encyclopedia of
DNA Elements

El proyecto ENCODE

- La 'Encyclopedia of DNA Elements' (ENCODE) surge de una colaboración internacional iniciada en 2003.
- El objetivo de ENCODE es elaborar un catálogo exhaustivo de todos los elementos funcionales en el genoma humano, incluyendo tanto ARNs como proteínas, así como aquellos elementos reguladores que controlan el tipo celular y el momento del desarrollo en que un gen es activo.
- La cuestión es: la suma de los exones de los aprox. 21.000 genes humanos no llegan al 2% del genoma ¿para que sirve el 98% restante? ¿es ADN basura?



Principales hallazgos de ENCODE (Nature 515, 355–364, 2014)

La mayor parte del genoma (80.4%) se puede asociar con al menos una función en alguno de los 147 tipos celulares analizados. Puesto que puede haber hasta 2.000 tipos celulares, este porcentaje podría llegar a ser mucho más alto!

Muchas de los elementos funcionales encontrados se localizan en las regiones no-codificadoras (fuera de los genes).

Muchas enfermedades se asocian con un determinado factor de transcripción que varía entre tipos celulares.

v90 released!

September 16, 2019

New software version released!

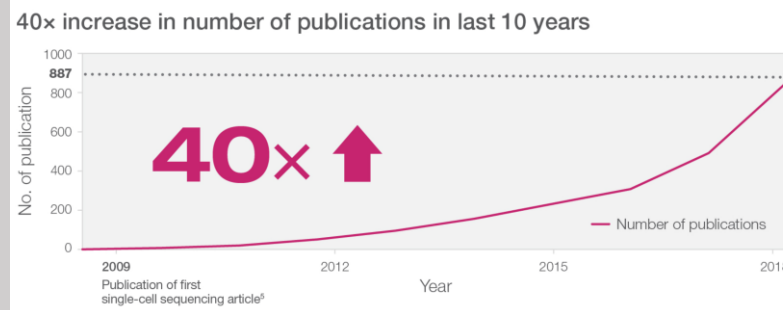
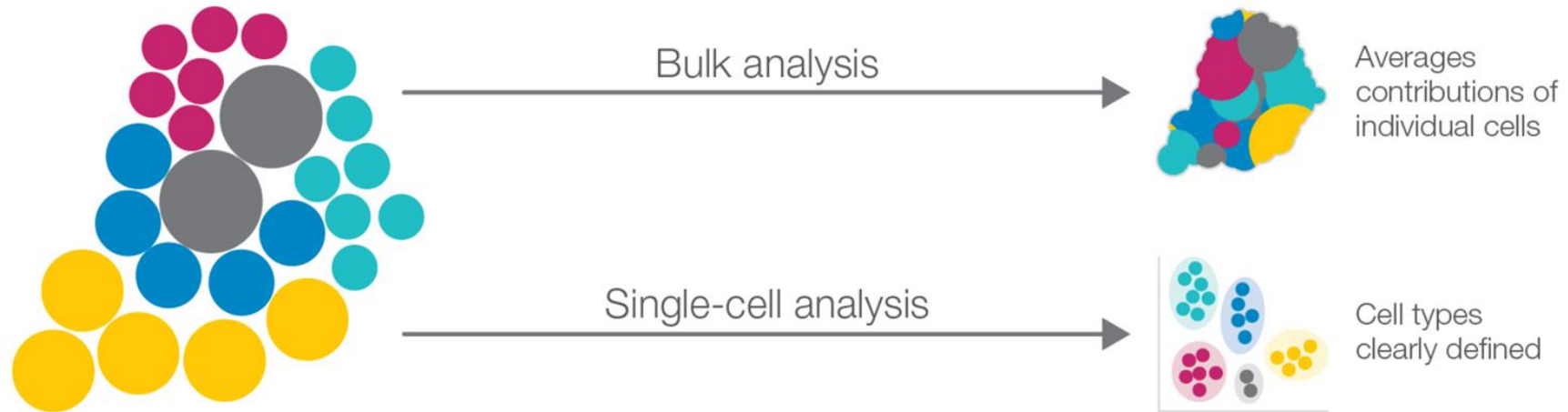
Secuenciación de células únicas

mRNA-Seq whole-transcriptome analysis of a single cell

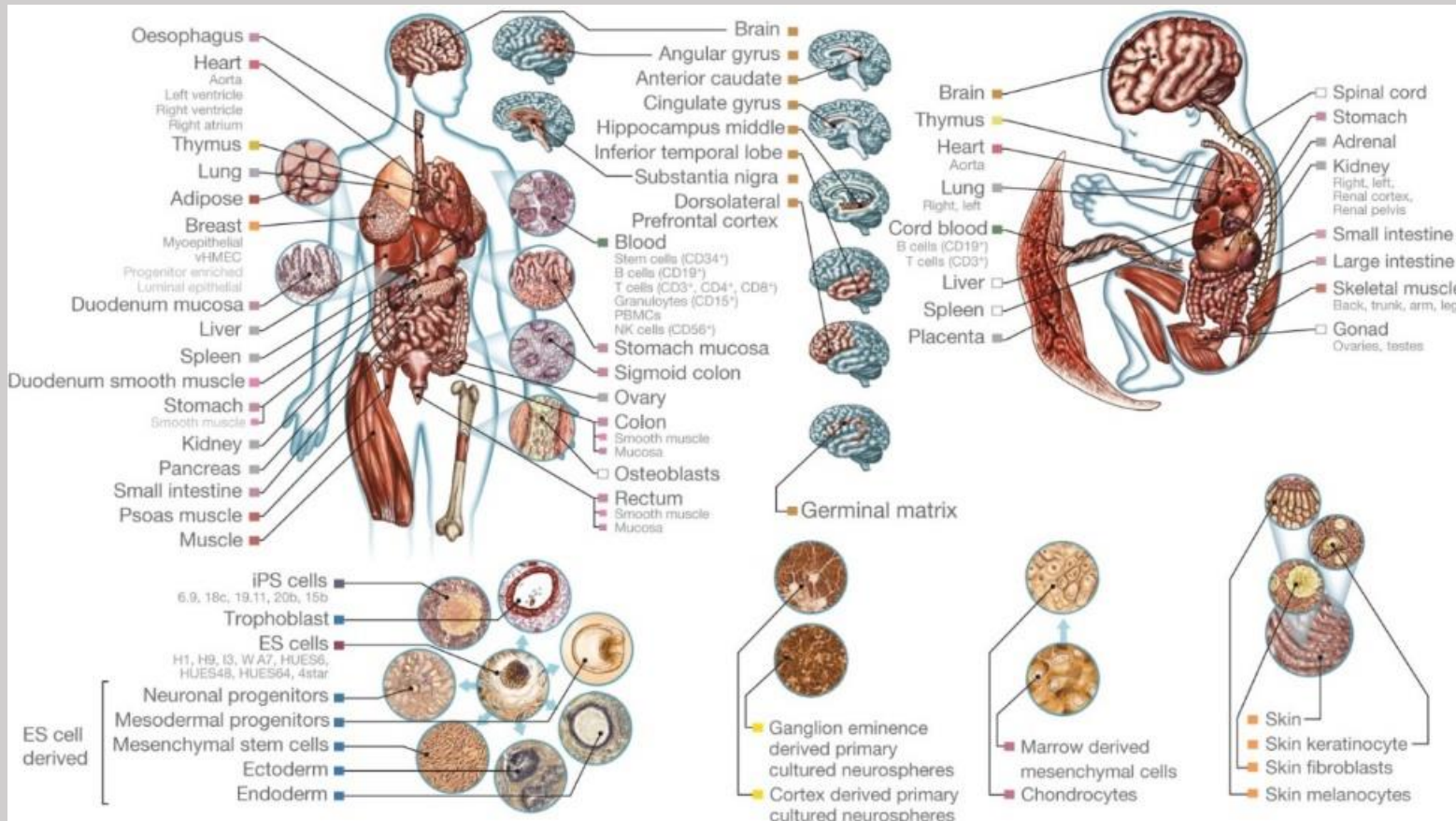
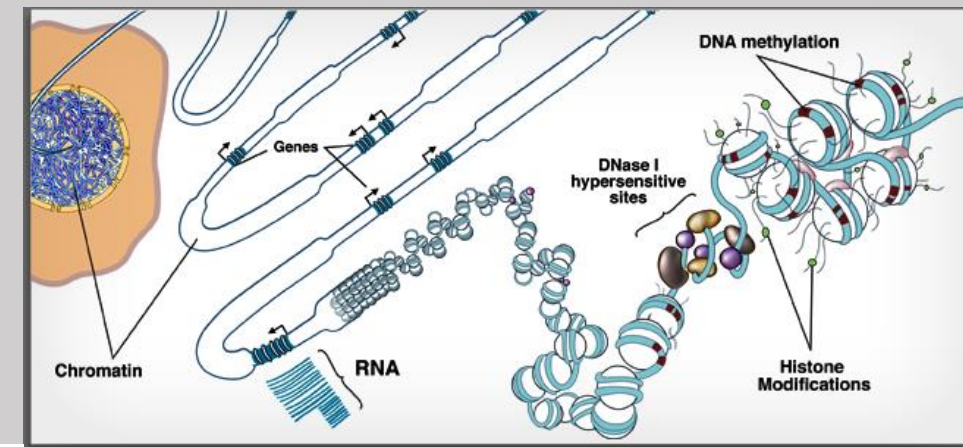
Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao ✉ & M Azim Surani ✉

Nature Methods **6**, 377–382 (2009) | [Download Citation](#) ↓

Each cell is unique

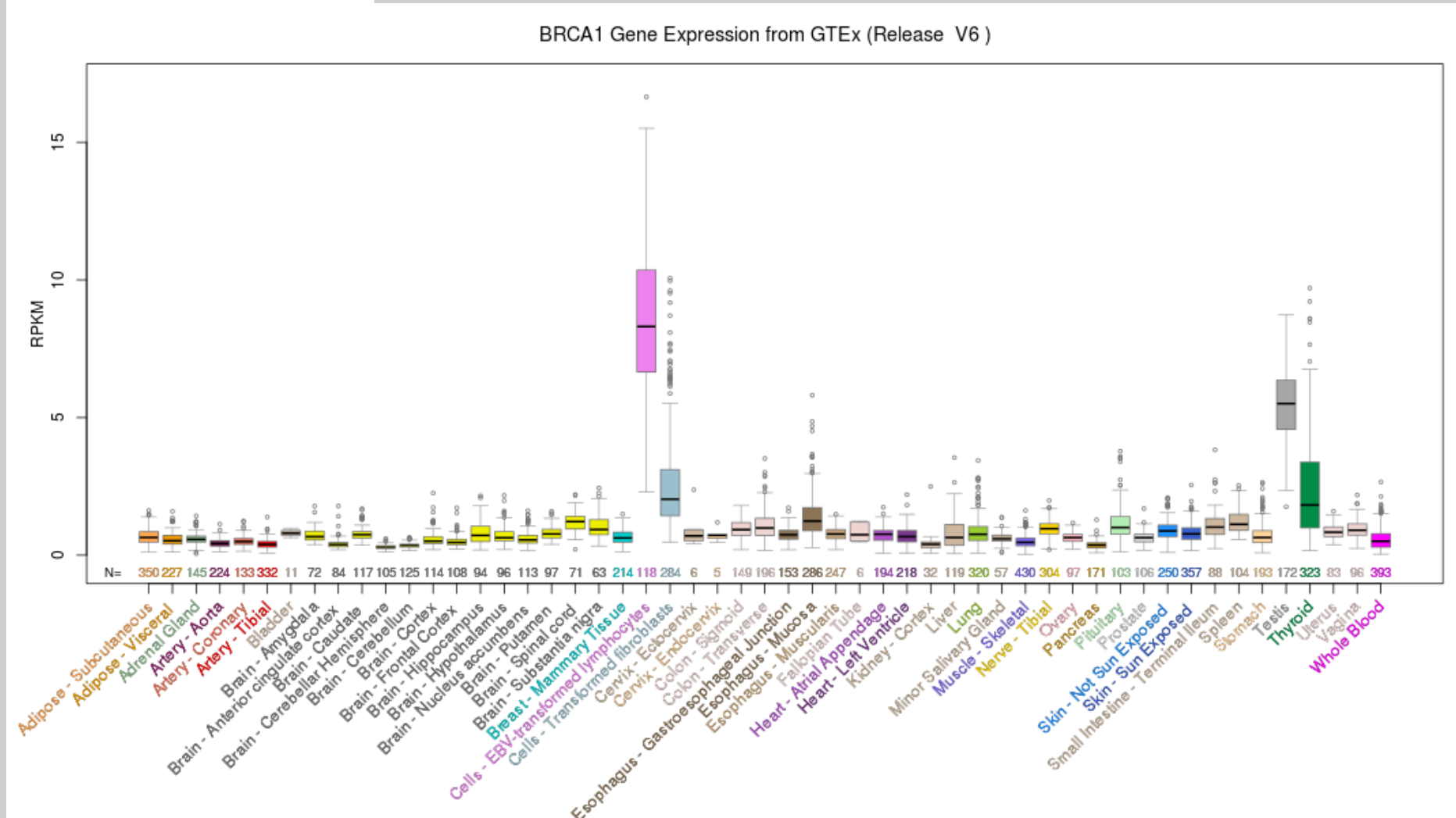


Epigenómica



Transcriptómica: RNA-seq -> Genotype-Tissue Expression (GTEx) Project

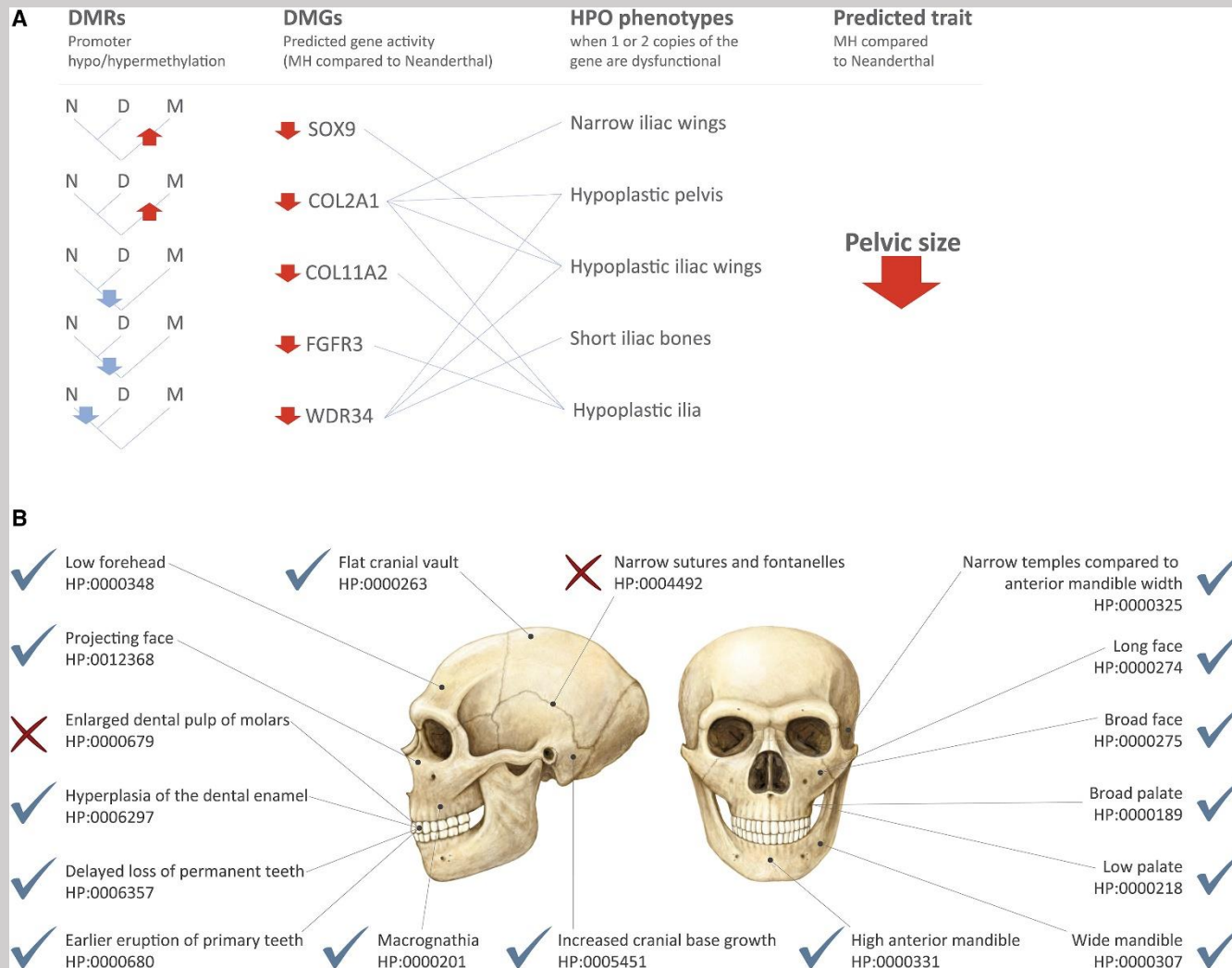
Ya contiene 17,382 muestras de 948 donantes y 54 tejidos



RPKM stands for Reads Per Kilobase of transcript per Million mapped reads

Single-cytosine resolution
methylomes: DMRs, DMIs, DMCs





- DNA hypermethylation was used to infer gene downregulation
- Phenotype of downregulation was inferred from known loss-of-function phenotypes
- Neanderthal and chimp profiles reconstructed and tested against their known anatomy
- A reconstructed Denisovan anatomical profile was proposed

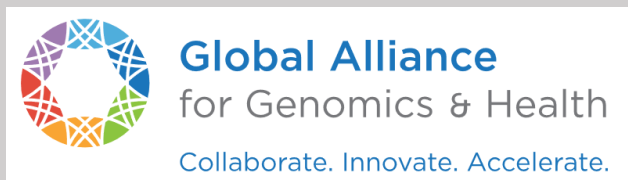


Reconstructing Denisovan Anatomy Using DNA Methylation Maps

David Gokhman, Nadav Mishol, Marc de Manuel, David de Juan, Jonathan Shuqrun, Eran Meshorer, Tomas Marques-Bonet, Yoel Rak, Liran Carmel
Cell, Volume 179, Issue 1, Pages 180-192.e10 (September 2019)

Áreas de investigación activa

- Biomarcadores
- Interactoma
- GWAS (phasing, tamaño de muestra)
- Genómica de caracteres complejos
- Análisis multi-ómicos
- Exomas
- Aprendizaje automático, clasificación
- Inteligencia artificial: algoritmos genéticos, redes neuronales
- Genómica personalizada
- Medicina de precisión
- Farmagenómica
- ...



Enabling responsible genomic data sharing
for the benefit of human health

The Personal Genome Project

<https://www.personalgenomes.org/>



Dedicated to creating public genome, health, and trait data. The PGP approach is to invite willing participants to publicly share their personal data for the greater good.

Bioinformática: ¿tecnología o ciencia?

- Los métodos son informáticos, estadísticos, IA... Pero los conceptos son biológicos
- Algunos opinan que está muriendo de éxito: una mayoría de biólogos ya la usan rutinariamente
- Otros (Birney, Valencia...) dicen que sólo está en sus comienzos
- La bioinformática hunde sus raíces en principios biológicos fundamentales y, además de sus múltiples aplicaciones, seguirá siendo imprescindible para resolver también problemas de ciencia básica

...vuelta al origen: What is Life?’

- La característica fundamental de la vida es el **flujo de información** a lo largo del tiempo: transcripción, traducción, código genético, sensores...
- Pero ¿qué es la información?
- ¿Puede tener efectos causales?
- ¿Se conserva, como la materia y la energía?
- ¿Aumenta o disminuye a lo largo del proceso evolutivo?

El origen de la vida o el de la consciencia, no podrán resolverse sin responder esas preguntas.

Y en ello la **bioinformática** está llamada a jugar un papel clave.

Perspectivas en Bioinformática

Prof. Dr. José L. Oliver