www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

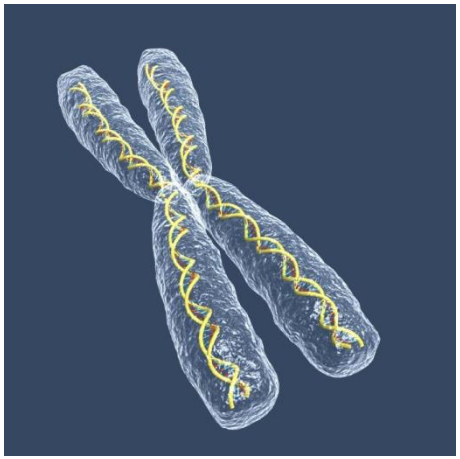
The HPC-Life Sciences scenario
Byte vs Flop?

Modesto Orozco

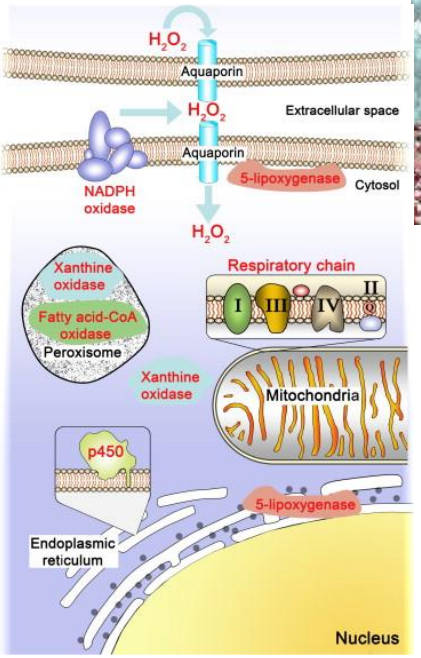
Henry Markram, Erik Lindahl, Paolo Carloni, Alfonso Valencia
Richard Lavery, Peter Coveney, Charles Laughton

Exascale for Life Sciences applications

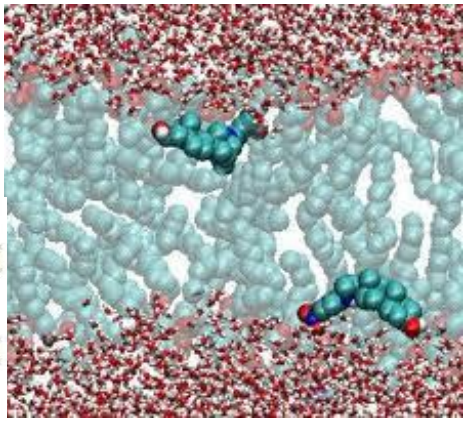
Omics



Systems Biology



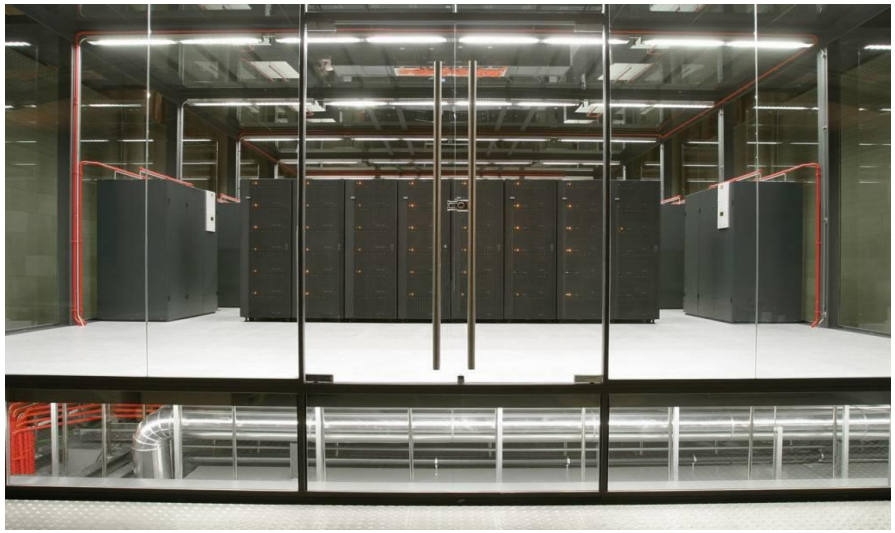
Molecular Simulation



Organ Simulation

The dual nature of computers

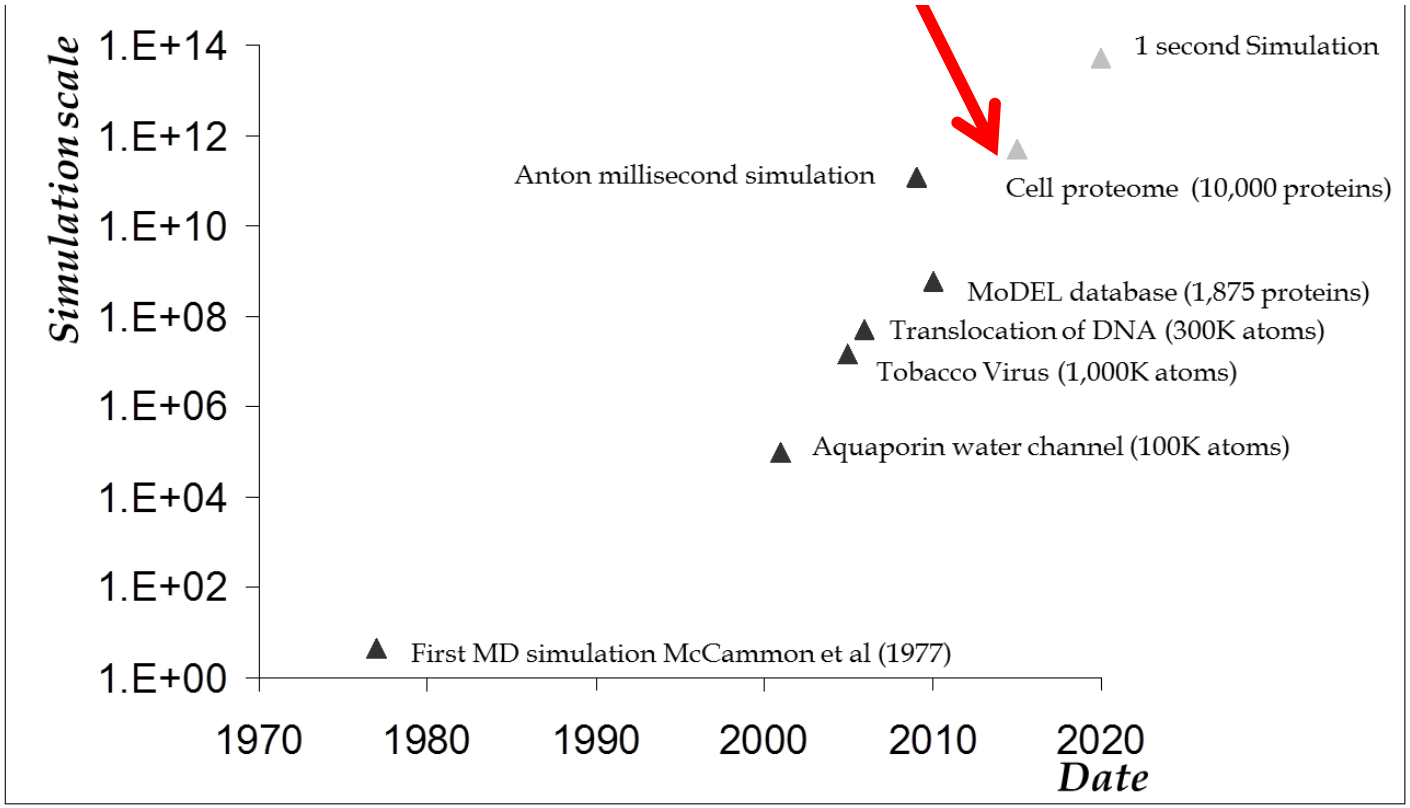
- Ordenador: a machine to manage data.
- Computador: a machine to do maths.



FLOPS and BYTES ARE EQUALLY IMPORTANT IN BIOLOGY?

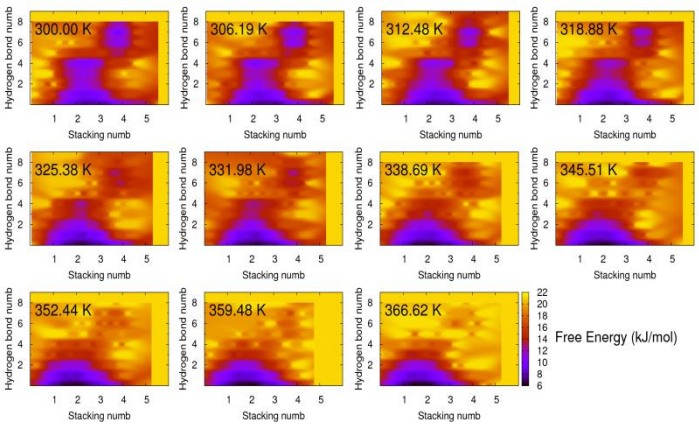
Evolution,…

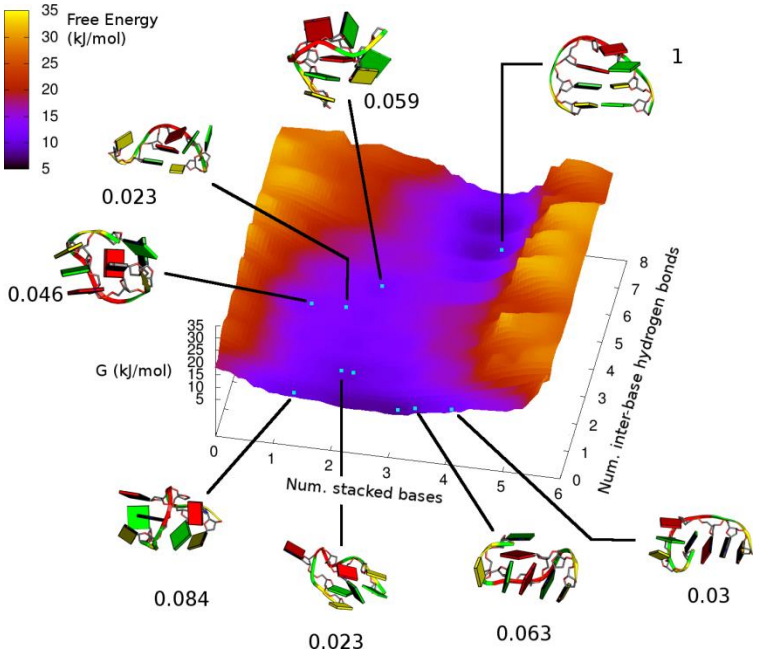
Mature HIV-1 capsid (64M atoms) NAMD

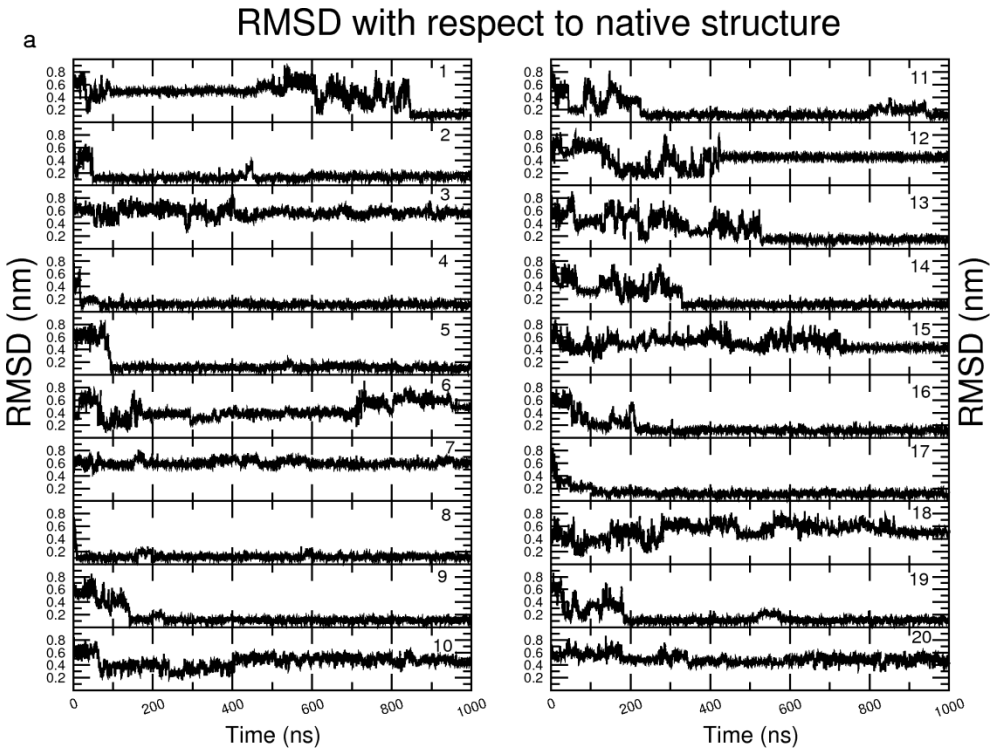


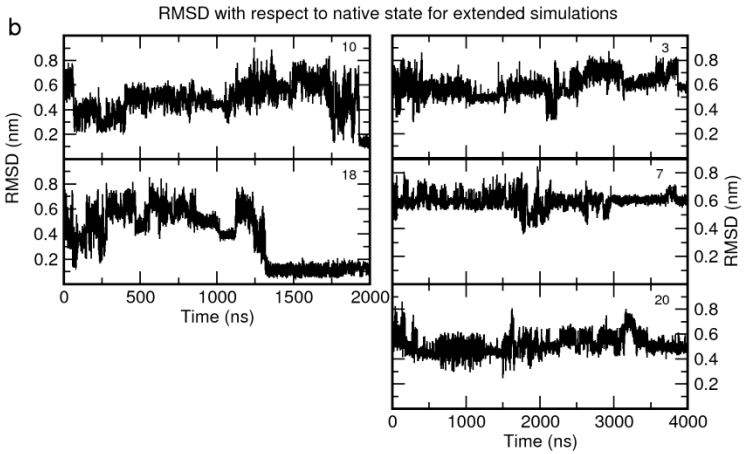
Simulation scale for long and large simulations is calculated multiplying simulation length (nanoseconds) with size (number of atoms).

Massive parallel simulations





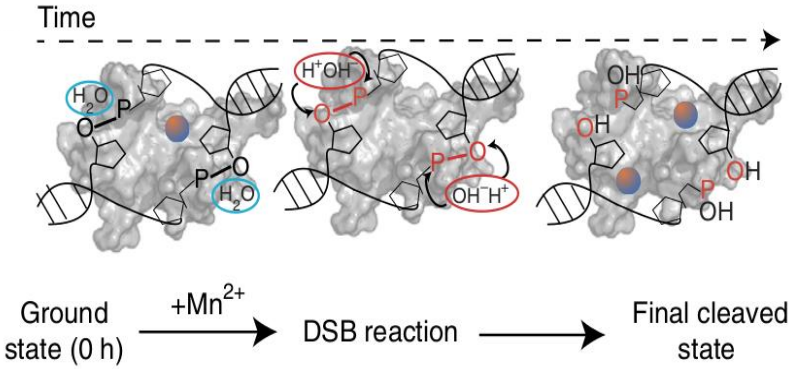




Portella & Orozco, Angew.Chem. 2010

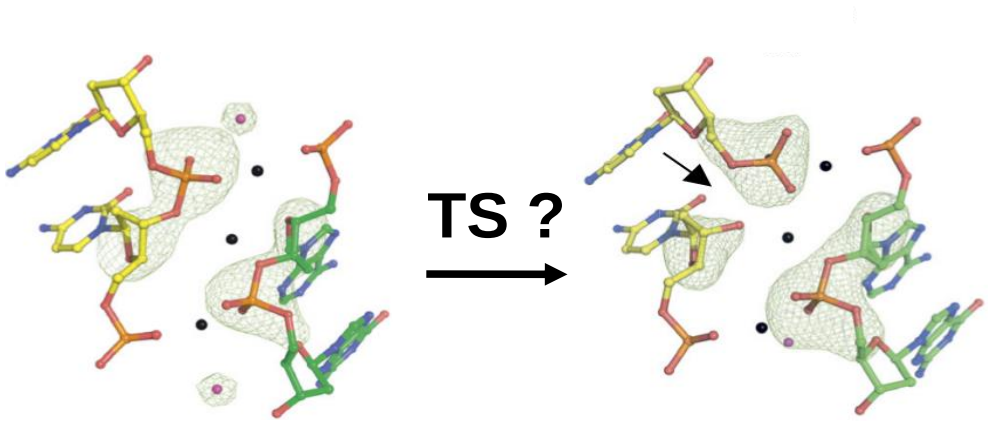
Integrate experiments

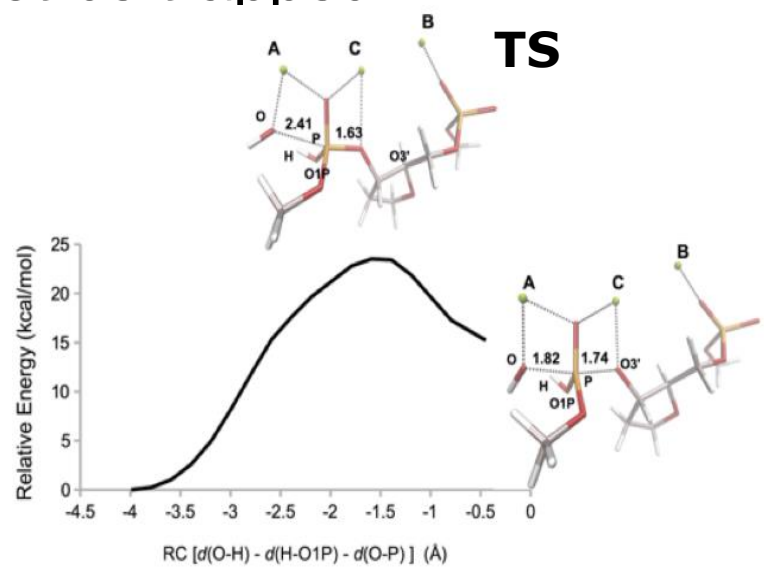
DNA double-strand break by the *I-Dmol* endonuclease



7 different catalytic intermediates were solved by X-ray, BUT reactive species cannot be trapped

We use a combination of **MD** and **QM** methods to fill the gaps and thus to complete the vision of the reaction, assign roles to the catalytic residues in the active site, estimate kinetic energy barriers ...



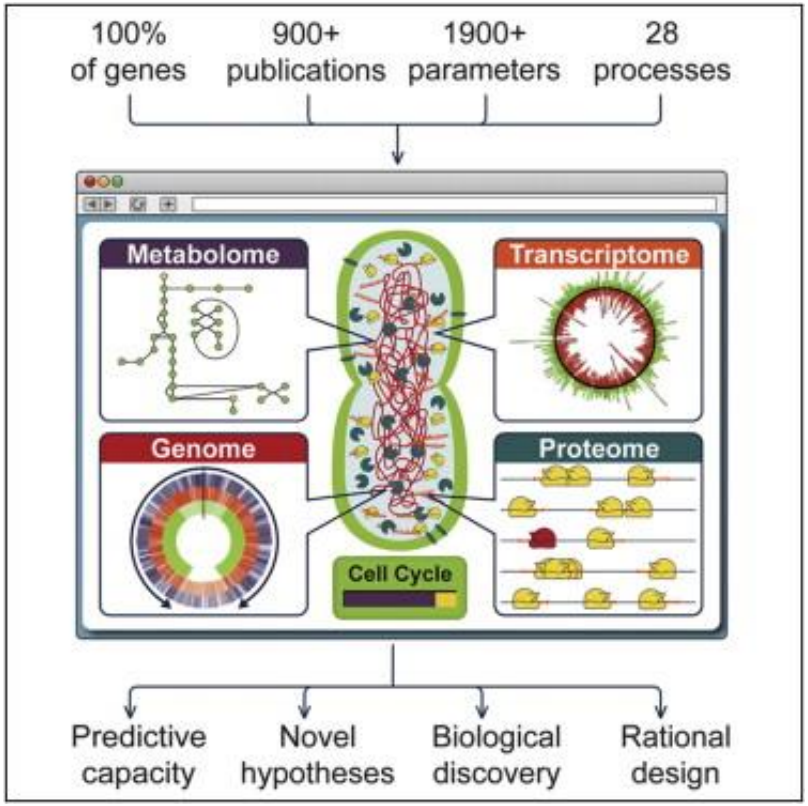


A reaction mechanism is proposed, involving the transient protonation of the PO4 group by the catalytic water

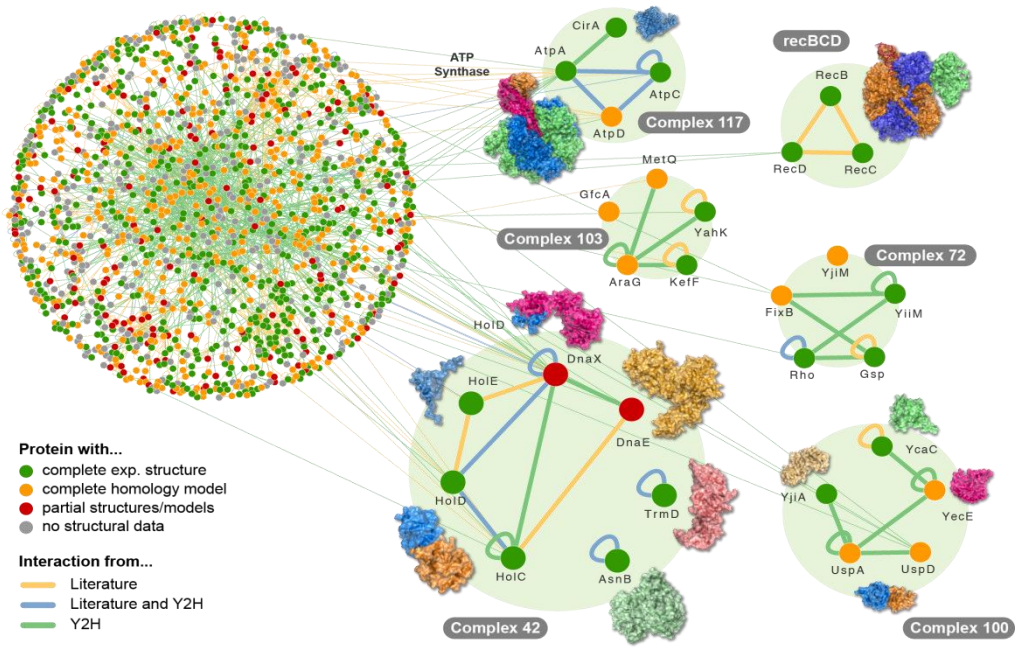
Molina, R.; et al.. *Nat. Struct. Mol. Biol.* **2014**. doi:10.1038/nsmb.2932

Whole-Cell simulation

- Whole-cell computational model of the life cycle of the human pathogen *Mycoplasma genitalium* from Genotype (Cell 2012)
- Entire organism (525 genes) modeled in terms of its molecular components
- Integration of multi-format data and very fragmented data
- experimental analysis directed by model predictions identified previously undetected kinetic parameters and biological functions
- 7 hours → 100 Gb of data



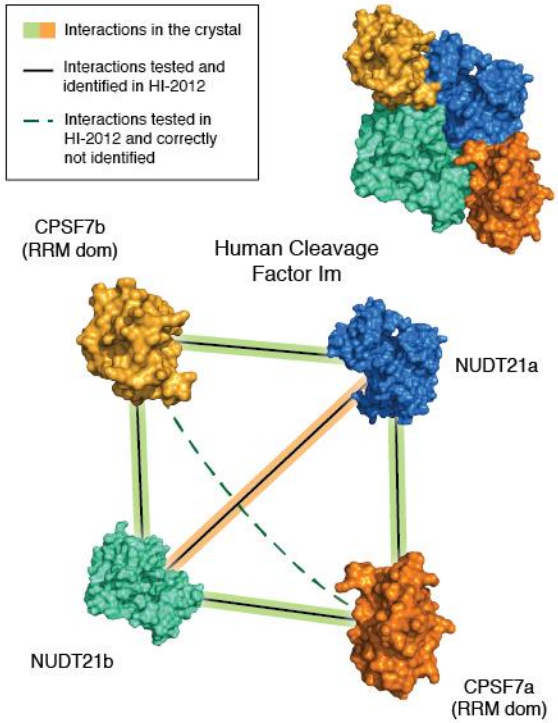
E. coli binary interactome network

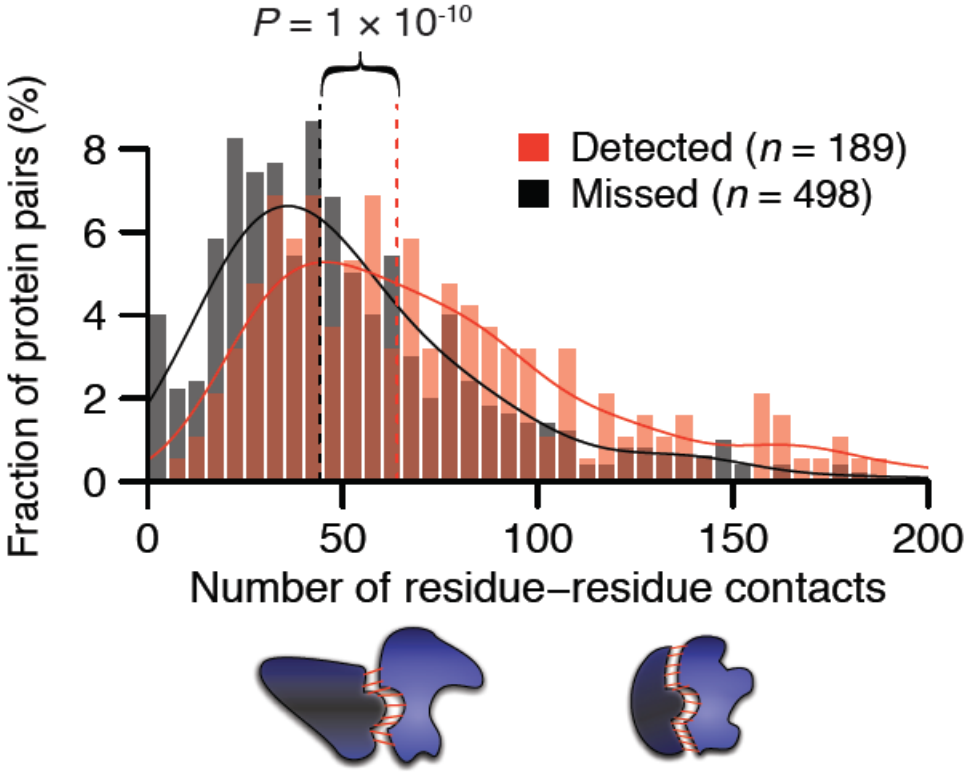


Rajagopala et al. *Nat Biotechnol* 2014

Human interactome network

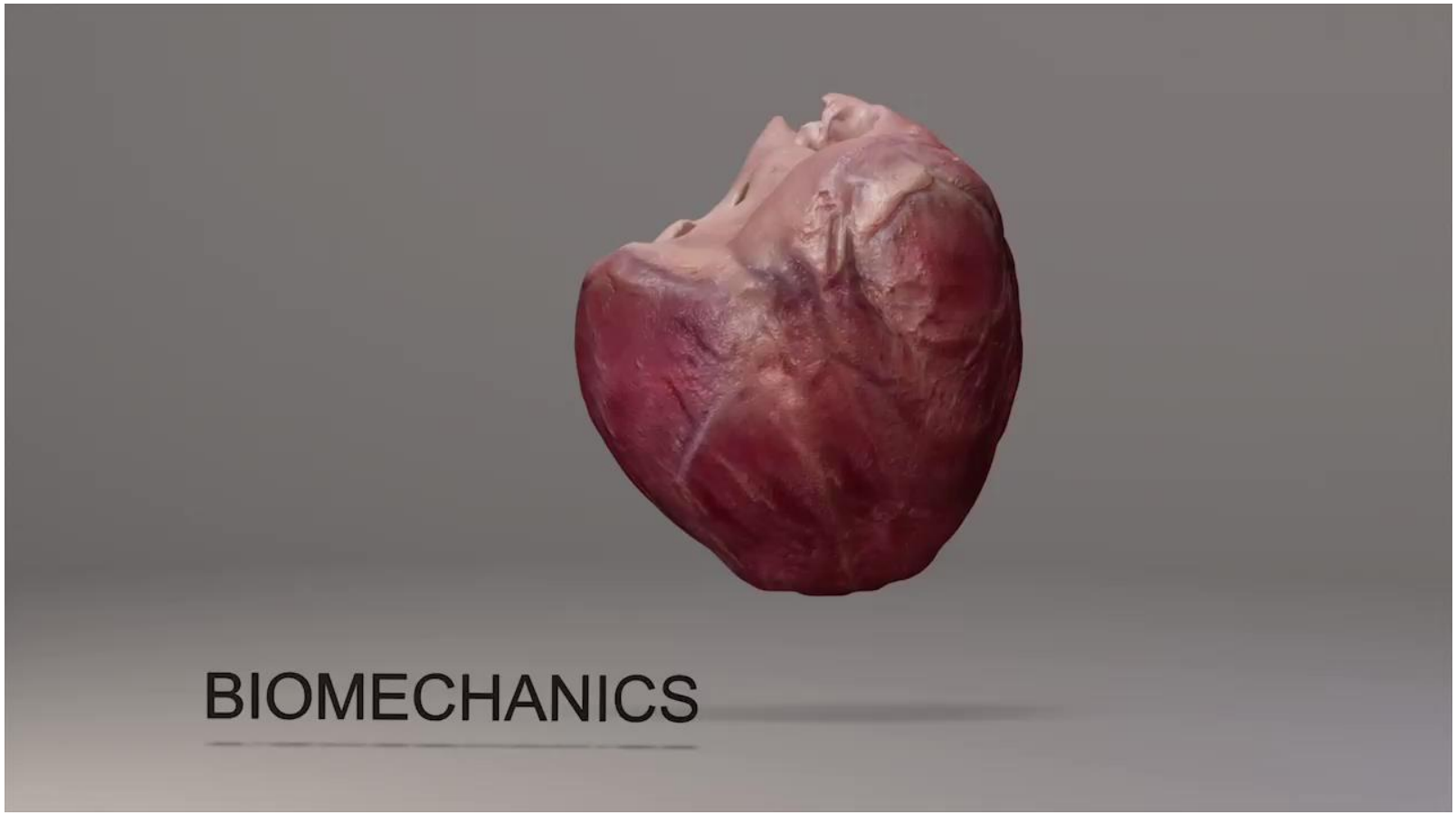
Description	Tested		Identified		Not identified		P-value
	Direct	Indirect	TP	FP	FN	TN	
All interactions	**687**	**56**	**189**	**2**	498	54	**1.01E-05**





Rolland et al. *Cell* 2014

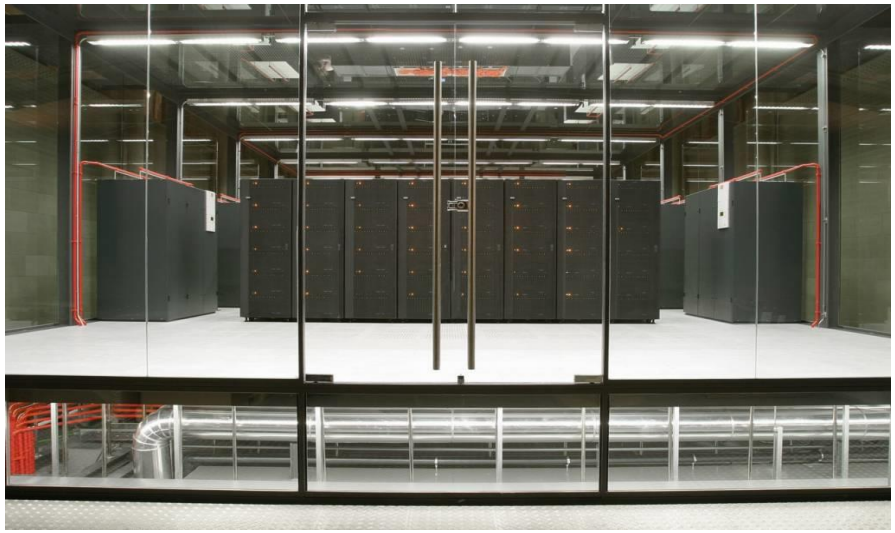
HEARTH SIMULATOR



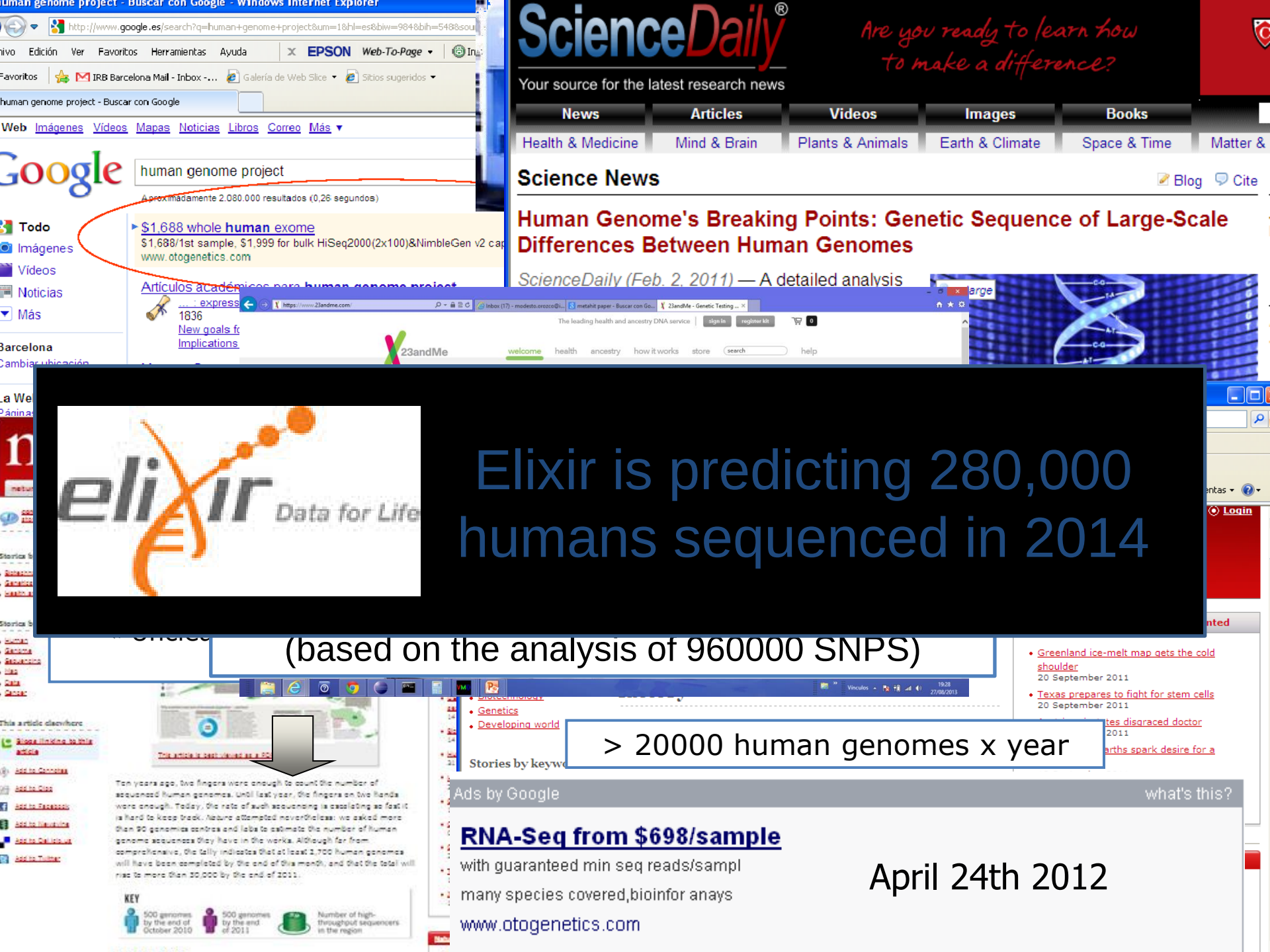
J.M.Cela (BSC)

The dual nature of computers

- Ordenador: a machine to manage data.
- Computador: a machine to do maths.

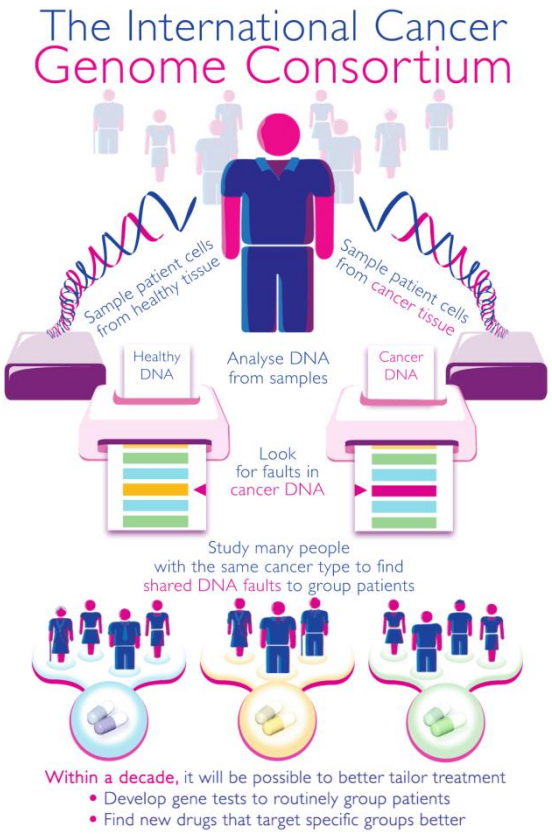


FLOPS → BYTE but also BYTE→FLOPS

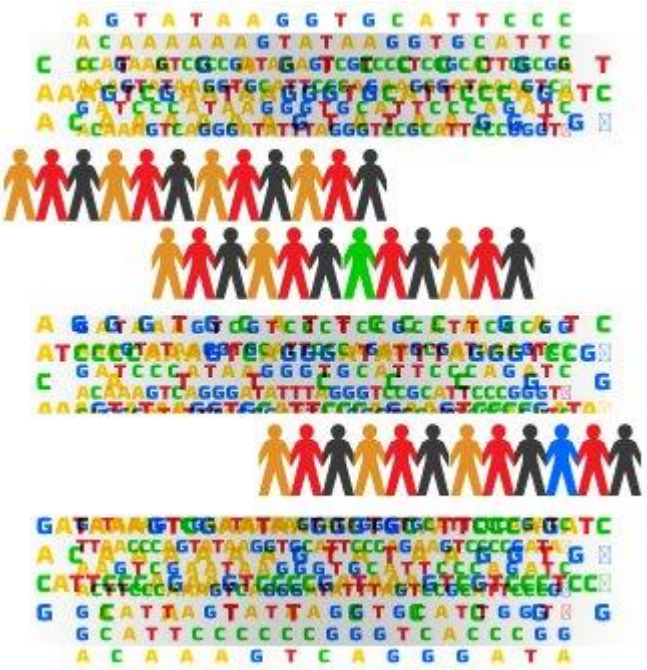

human genome project - Buscar con Google - Windows Internet Explorer
ScienceDaily
Your source for the latest research news
Are you ready to learn how to make a difference?
Science News
Human Genome's Breaking Points: Genetic Sequence of Large-Scale Differences Between Human Genomes
ScienceDaily (Feb. 2, 2011) — A detailed analysis
human genome project
$1,688 whole human exome
$1,688/1st sample, $1,999 for bulk HiSeq2000(2x100)&NimbleGen v2 cap
www.otogenetics.com
23andMe
elixir
Data for Life
Elixir is predicting 280,000 humans sequenced in 2014
(based on the analysis of 960000 SNPS)
> 20000 human genomes x year
Ads by Google
RNA-Seq from $698/sample
with guaranteed min seq reads/sampl
many species covered,bioinfor anays
www.otogenetics.com
April 24th 2012

Genomic Grand Challenges

ICGL



1000G



Genomics-UK 100,000 genomes

Inbox (74) - modesto.oro | Genomics England | 100,0

www.genomicsengland.co.uk

Genomics England

Home | About Genomics England | The 100,000 Genomes Project | GeCIP | Library and resources | News | Contact us

Genomics England, with the consent of participants and the support of the public, is creating a lasting legacy for patients, the NHS and the UK economy through the sequencing of 100,000 genomes: the 100,000 Genomes Project.

Genomics England was set up by the Department of Health to deliver the 100,000 Genomes Project. Initially the focus will be on rare disease, cancer and infectious disease. The project is currently in its pilot phase and will be completed by the end of 2017.

Read more...

News

NHS Genomic Medicine Centres announced for 100,000 Genomes Project

NHS England has announced eleven Genomic Medicine Centres that will lead the way in delivering the 100,000 Genomes Project. This marks the start of the main phase of the

Understanding genomics

What is genomics?

Our Head of Engagement, Vivienne Parry, explains more about genomics in this film courtesy of our partners at Health Education England.

Tweets Follow

Genomics England @GenomicsEngland 7 Jan

Check out this Storify about the day our new Genomic Medicine Centres were revealed: bit.ly/GMCStorify #Genomes100k #100KGP

Show Media

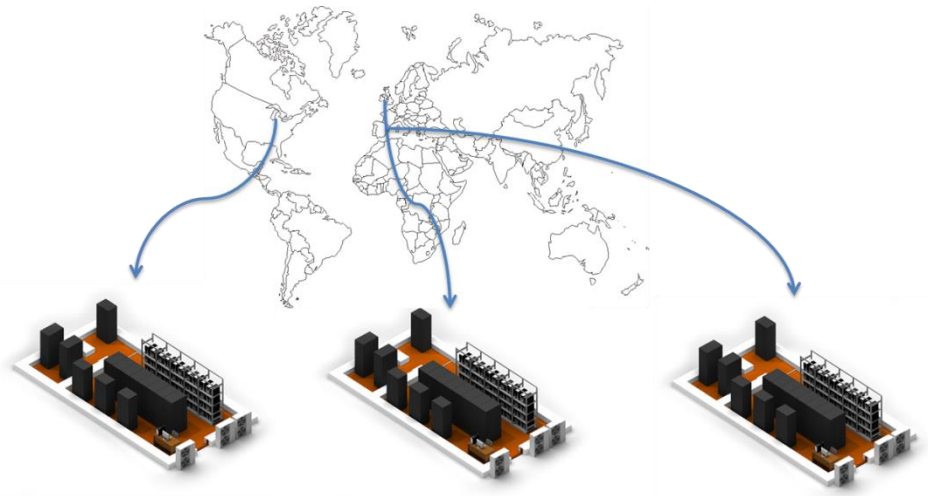
Department of Health 22 Dec

World-level genomic consortiums

European Genome-phenome Archive

International Cancer Genome Consortium

The Cancer Genome Atlas
Understanding genomics to improve cancer care



University of Chicago
NIH Trusted Partner
Holds TCGA Data

EBI
Data Manager for 1KGP
Hosts EGA

Barcelona Supercomputer Center
EGA Partner

THE UNIVERSITY OF CHICAGO MEDICINE

EMBL-EBI

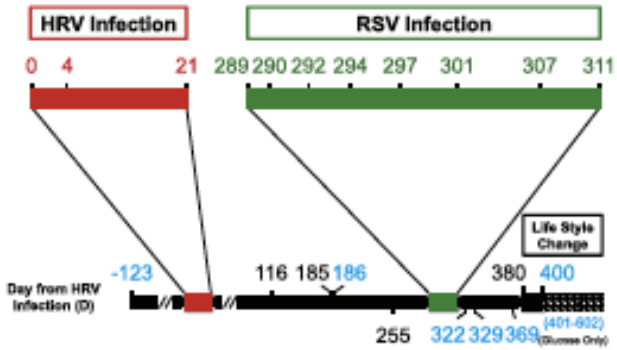
EXCELENCIA SEVERO OCHOA
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

- European Genome-phenome Archive (EGA) repository will allow us to explore datasets from numerous genetic studies
- Pan-cancer will produce rich data that will provide a major opportunity to develop an integrated picture of differences across tumor lineages

Resource Cell *148*, 1293–1307, March 16, 2012 ©2012 Cell

Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes

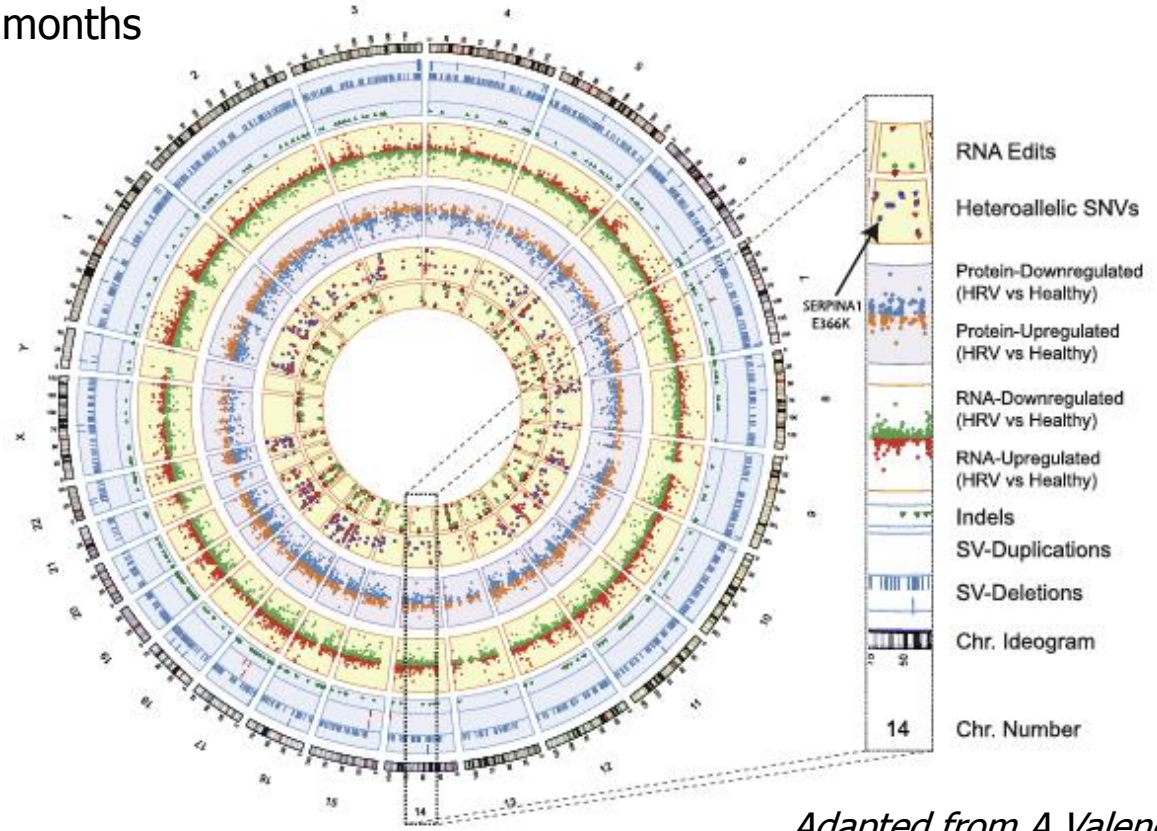
Rui Chen,[1,11] George I. Mias,[1,11] Jennifer Li-Pook-Than,[1,11] Lihua Jiang,[1,11] Hugo Y.K. Lam,[1,12] Rong Chen,[2,12] Elana Miriami,[1] Konrad J. Karczewski,[1] Manoj Hariharan,[1] Frederick E. Dewey,[3] Yong Cheng,[1] Michael J. Clark,[1] Hogune Im,[1] Lukas Habegger,[6,7] Suganthi Balasubramanian,[6,7] Maeve O'Huallachain,[1] Joel T. Dudley,[2] Sara Hillenmeyer,[1] Rajini Haraksingh,[1] Donald Sharon,[1] Ghia Euskirchen,[1] Phil Lacroute,[1] Keith Bettinger,[1] Alan P. Boyle,[1] Maya Kasowski,[1] Fabian Grubert,[1] Scott Seki,[2] Marco Garcia,[2] Michelle Whirl-Carrillo,[1] Mercedes Gallardo,[9,10] Maria A. Blasco,[9] Peter L. Greenberg,[4] Phyllis Snyder,[1] Teri E. Klein,[1] Russ B. Altman,[1,5] Atul J. Butte,[2] Euan A. Ashley,[3] Mark Gerstein,[6,7,8] Kari C. Nadeau,[2] Hua Tang,[1] and Michael Snyder[1,*]



1 Individual followed for 14 months

Personalized Medicine

Source: *New Yorker*



Adapted from A. Valencia

LETTER

doi:10.1038/nature11875

Towards practical, high-capacity, low-maintenance information storage in synthesized DNA

Nick Goldman[1], Paul Bertone[1], Siyuan Chen[2], Christophe Dessimoz[1], Emily M. LeProust[2], Botond Sipos[1] & Ewan Birney[1]

Digital production, transmission and storage have revolutionized how we access and use information but have also made archiving an increasingly complex task that requires active, continuing maintenance of digital media. This challenge has focused some interest on DNA as an attractive target for information storage[1] because of its capacity for high-density information encoding, longevity under easily achieved conditions[2–4] and proven track record as an information bearer. Previous DNA-based information storage approaches have encoded only trivial amounts of information[5–7] or were not amenable to scaling-up[8], and used no robust error-correction and lacked examination of their cost-efficiency for large-scale information archival[9]. Here we describe a scalable method that can reliably store more information than has been handled before. We encoded computer files totalling 739 kilobytes of hard-disk storage and with an estimated Shannon information[10] of 5.2×10^6 bits into a DNA code, synthesized this DNA, sequenced it and reconstructed the original files with 100% accuracy. Theoretical analysis indicates that our DNA-based storage scheme could be scaled far beyond current global information volumes and offers a realistic technology for large-scale, long-term and infrequently accessed digital archiving. In fact, current trends in technological advances are reducing DNA synthesis costs at a pace that should make our scheme cost-effective for sub-50-year archiving within a decade.

Although techniques for manipulating, storing and copying large amounts of existing DNA have been established for many years[11–13], one of the main challenges for practical DNA-based information storage is the difficulty of synthesizing long sequences of DNA *de novo* to an exactly specified design. As in the approach of ref. 9, we represent the information being stored as a hypothetical long DNA molecule and encode this *in vitro* using shorter DNA fragments. This offers the benefits that isolated DNA fragments are easily manipulated *in vitro*[11,13], and that the routine recovery of intact fragments from samples that are tens of thousands of years old[14,15] indicates that well-prepared synthetic DNA should have an exceptionally long lifespan in low-maintenance environments[3,4]. In contrast, approaches using living vectors[6–8] are not as reliable, scalable or cost-efficient owing to disadvantages such as constraints on the genomic elements and locations that can be manipulated without affecting viability, the fact that mutation will cause the fidelity of stored and decoded information to reduce over time, and possibly the requirement for storage conditions to be carefully regulated. Existing schemes used for DNA computing in principle permit large-scale memory[1,16], but data encoding in DNA computing is inextricably linked to the specific application or algorithm[17] and no practical storage schemes have been realized.

As a proof of concept for practical DNA-based storage, we selected and encoded a range of common computer file formats to emphasize the ability to store arbitrary digital information. The five files comprised all 154 of Shakespeare's sonnets (ASCII text), a classic scientific paper[18] (PDF format), a medium-resolution colour photograph of the European Bioinformatics Institute (JPEG 2000 format), a 26-s excerpt from Martin Luther King's 1963 'I have a dream' speech (MP3 format) and a Huffman code[10] used in this study to convert bytes to base-3 digits (ASCII text), giving a total of 757,051 bytes or a Shannon information[10] of 5.2×10^6 bits (see Supplementary Information and Supplementary Table 1 for full details).

The bytes comprising each file were represented as single DNA sequences with no homopolymers (runs of ≥2 identical bases, which are associated with higher error rates in existing high-throughput sequencing technologies[19] and led to errors in a recent DNA-storage experiment[9]). Each DNA sequence was split into overlapping segments, generating fourfold redundancy, and alternate segments were converted to their reverse complement (see Fig. 1 and Supplementary Information). These measures reduce the probability of systematic failure for any particular string, which could lead to uncorrectable errors and data loss. Each segment was then augmented with indexing information that permitted determination of the file from which it originated and its location within that file, and simple parity-check error-detection[10]. In all, the five files were represented by a total of 153,335 strings of DNA, each comprising 117 nucleotides (nt). The perfectly uniform fragment lengths and absence of homopolymers make it obvious that the synthesized DNA does not have a natural (biological) origin, and so imply the presence of deliberate design and encoded information[2].

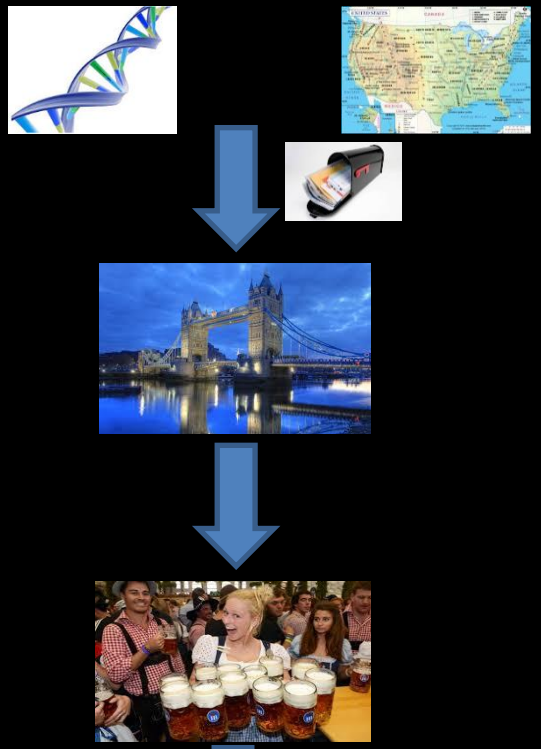
We synthesized oligonucleotides (oligos) corresponding to our designed DNA strings using an updated version of Agilent Technologies' OLS (oligo library synthesis) process[20], creating $\sim 1.2 \times 10^7$ copies of each DNA string. Errors occur only rarely (~1 error per 500 bases) and independently in the different copies of each string, again enhancing our method's error tolerance. We shipped the synthesized DNA in lyophilized form that is expected to have excellent long-term preservation characteristics[3,4], at ambient temperature and without specialized packaging, from the USA to Germany via the UK. After resuspension, amplification and purification, we sequenced a sample of the resulting library products at the EMBL Genomics Core Facility in paired-end mode on the Illumina HiSeq 2000. We transferred the remainder of the library to multiple aliquots and re-lyophilized these for long-term storage.

Our base calling using AYB[21] yielded 79.6×10^6 read-pairs of 104 bases in length, from which we reconstructed full-length (117-nt) DNA strings *in silico*. Strings with uncertainties due to synthesis or sequencing errors were discarded and the remainder decoded using the reverse of the encoding procedure, with the error-detection bases and properties of the coding scheme allowing us to discard further strings containing errors. Although many discarded strings will have contained information that could have been recovered with more sophisticated decoding, the high level of redundancy and sequencing coverage rendered this unnecessary in our experiment. Full-length DNA sequences representing the original encoded files were then reconstructed *in silico*. The decoding process used no additional information derived from knowledge of the experimental design. Full details of the encoding, sequencing and decoding processes are given in Supplementary Information.

Four of the five resulting DNA sequences could be fully decoded without intervention. The fifth however contained two gaps, each a run

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK. [2]Agilent Technologies, Genomics-LS SU, 5301 Stevens Creek Boulevard, Santa Clara, California 95051, USA.

All 154 of Shakespeare's sonnets (ASCII text), WC paper, a EBI photo, Martin Luther King's 1963 speech and Huffman code.

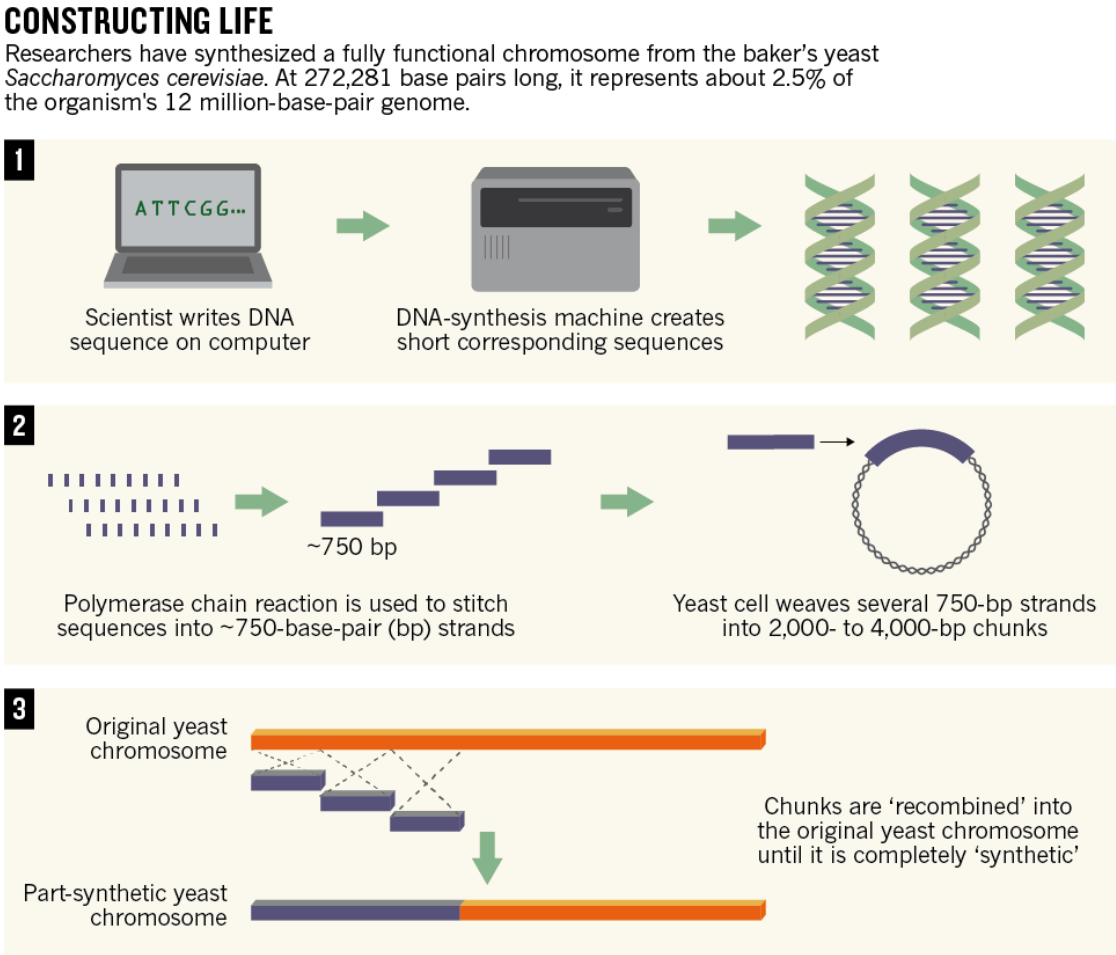


All 154 of Shakespeare's sonnets (ASCII text), WC paper, a EBI photo, Martin Luther King's 1963 speech and Huffman code.

Synthetic biology

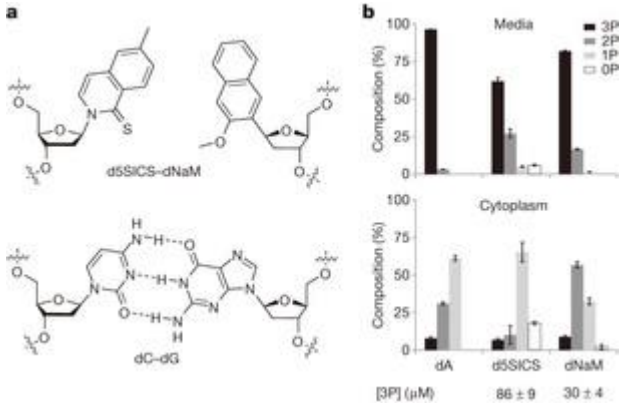
- *2010 -> parasite*
- *2014: the synthesis of a functional 272,871–base pair designer eukaryotic chromosome*
- 2.5% of the 12-million-base-pair *S. cerevisiae* genome

Science April 2014



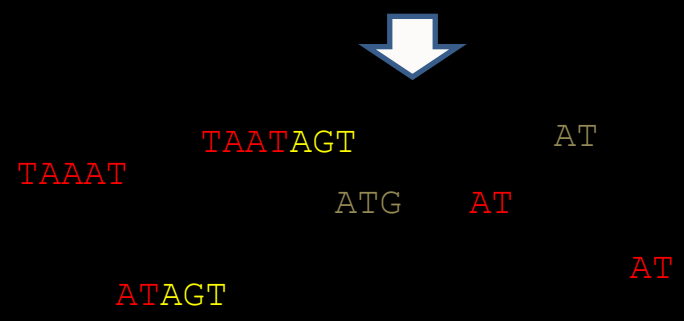
A semi-synthetic organism with an expanded genetic alphabet

- Neither the presence of the unnatural triphosphates nor the replication of the UBP introduces a notable growth burden.
- Lastly, we find that the UBP is not efficiently excised by DNA repair pathways. Thus, the resulting bacterium is the first organism to propagate stably an expanded genetic alphabet.



Nature 509,385–388 (15 May 2014)

Data is extremely noisy
deriving information from data is very costly



30 x to 200 x coverage needed

Final results are strongly dependent on processing
No accepted standards!
Processing is very slow

example just CLL processing → 25% of Marenostrum

Smufin a reference-free method for somatic mutation detection

Normal Tumor

Normal Tumor

Time x 1 cancer sample
Sequencing: < 7 hours
Variant Calling: 1 week on a 512 core cluster

Variant Calling Smufin: 7 hours on a 16 core cluster

either point mutations or
structural variants

- Need to map reads to reference genome and discard variants.
- Low sensitivity and/or specificicty.
- Detect either point mutation or structural vaiants
- Best detection of structural variants of certain sizes.

Point mutations AND
structural variants of any size
(inserions, deletions, inversions)

- Detection of somatic variations through direct read comparison without mapping to a reference genome.
- Hihg sensitivity and/or specificicty.
- Detects both point mutation or structural vaiants of any size

Nature Biotech 2014