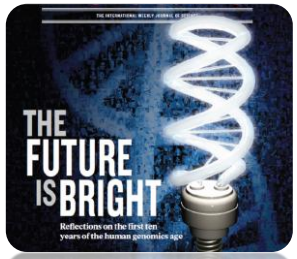


The challenges of big data in Genomics

Antonio Barbadilla

Group Genomics, Bioinformatics & Evolution

Institut Biotecnologia I Biomedicina

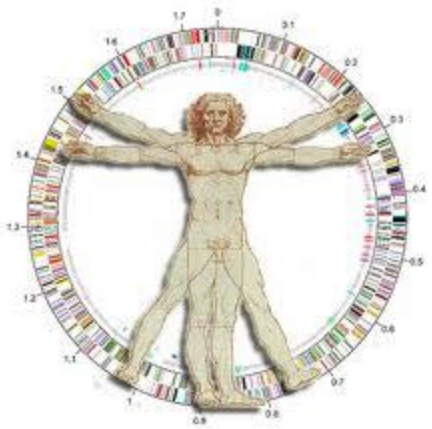
Departament de Genètica i Microbiologia

UAB

Outline

- Genome science: the HGP, *a new starting point*
- The essence of Genomics
- Genome sequencing
- Steps of genome analysis
- The technological explosion: Genome sequences as commodity
- The triumphal march of genomics
- Genome science challenges
- The foreseeable future of genome science
- Readings

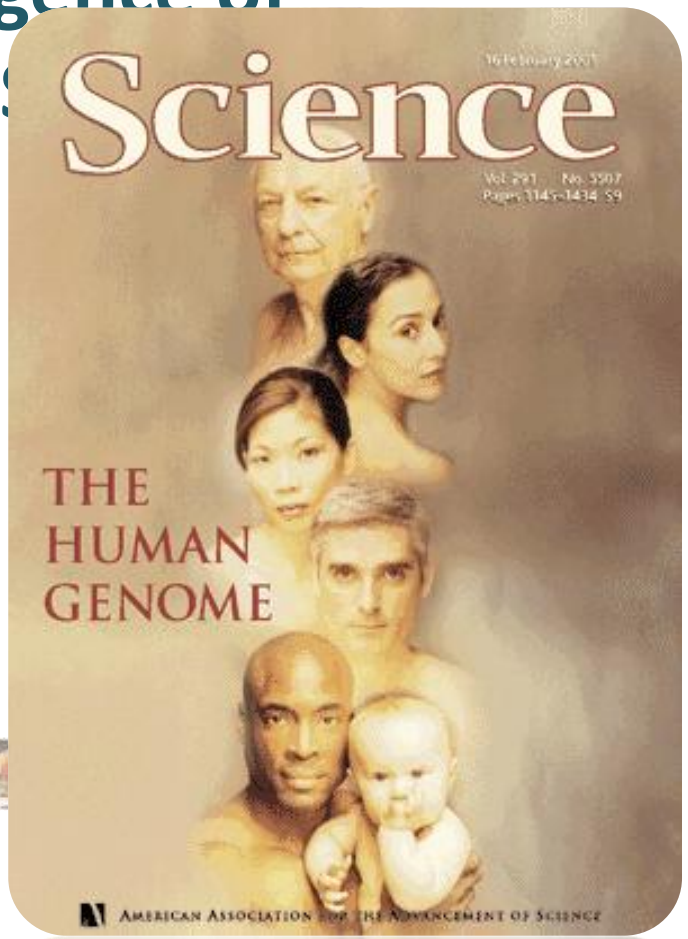
Genome science: the HGP, *a new starting point*



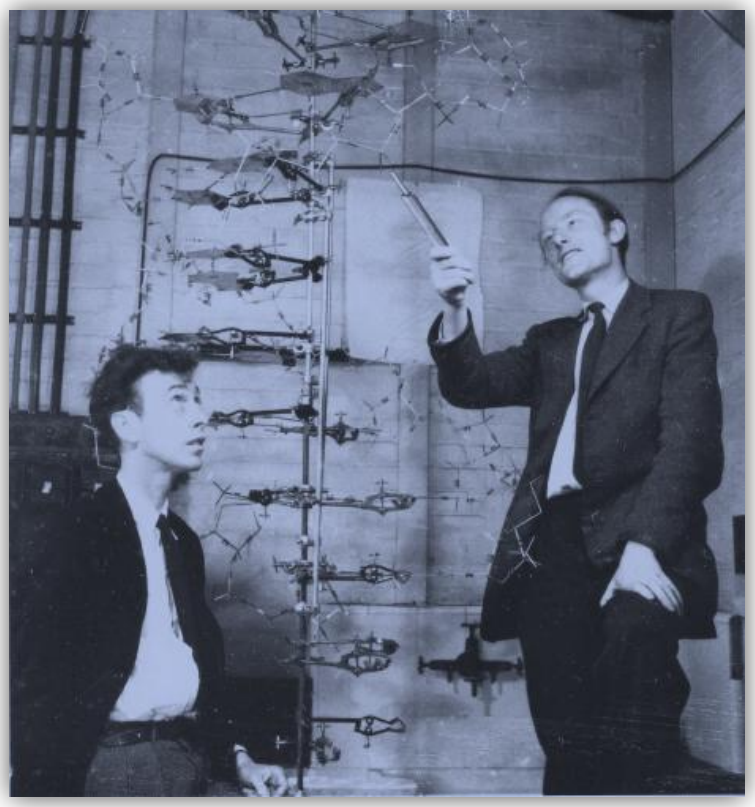
Genomics Landmarks 2001 and 2004

The publication of the draft and complete human sequence

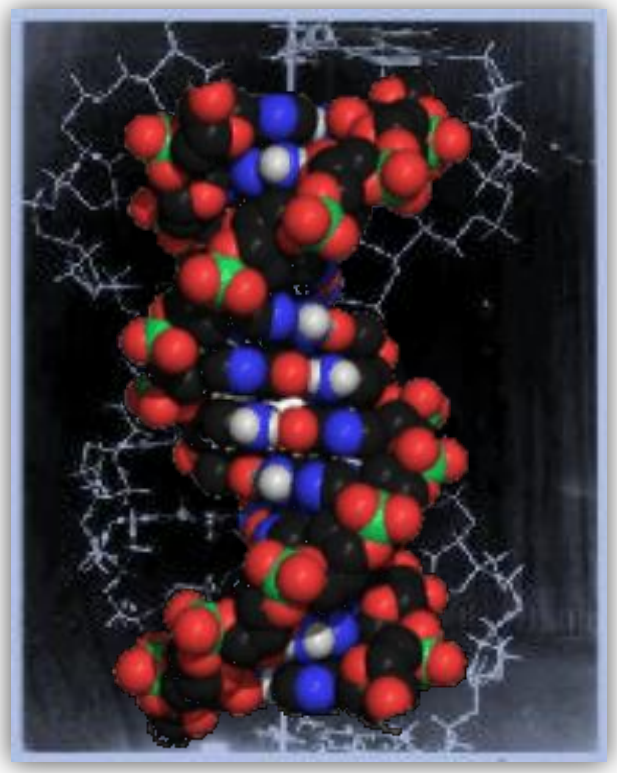
The emergence of

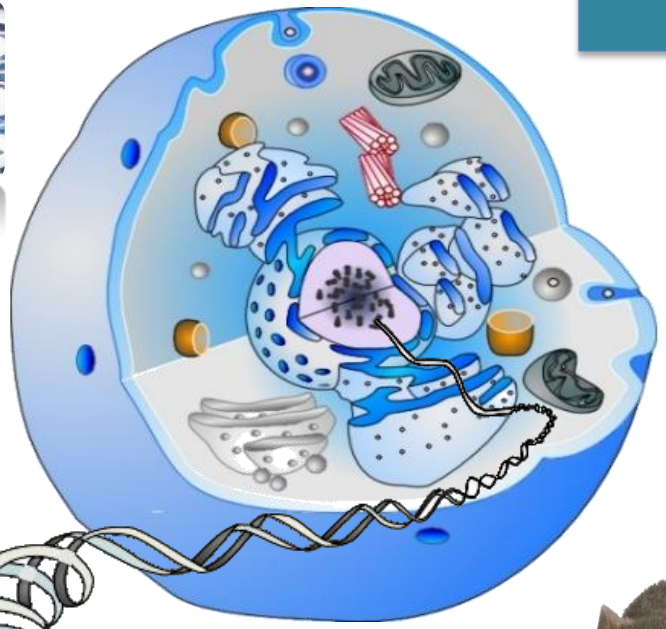


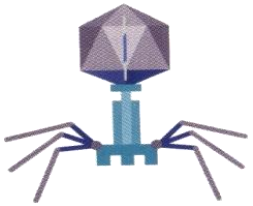
DNA Doble Helix: the secret of life

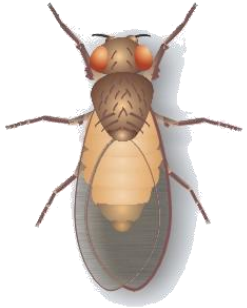


1953

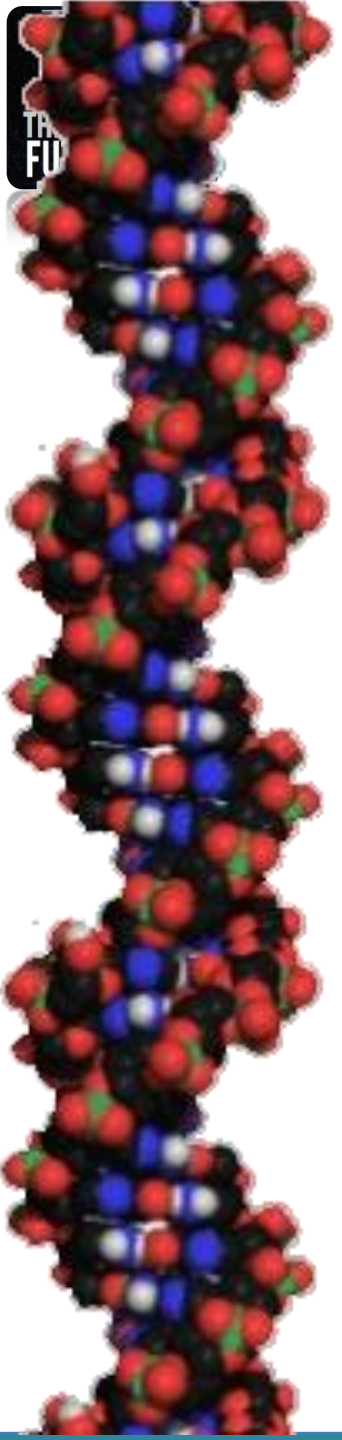


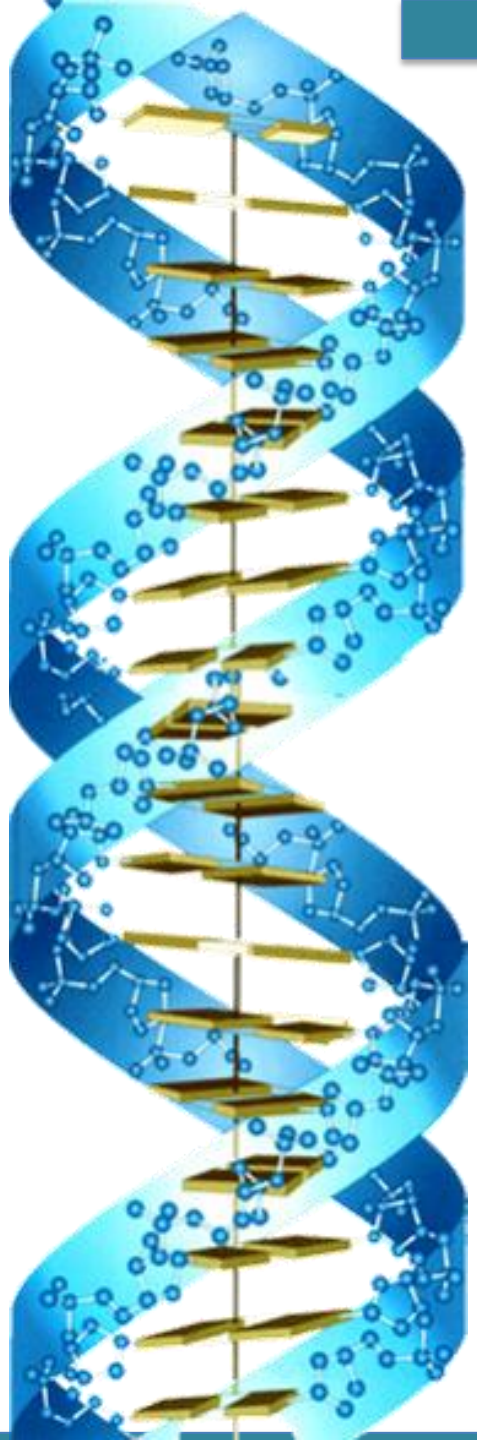


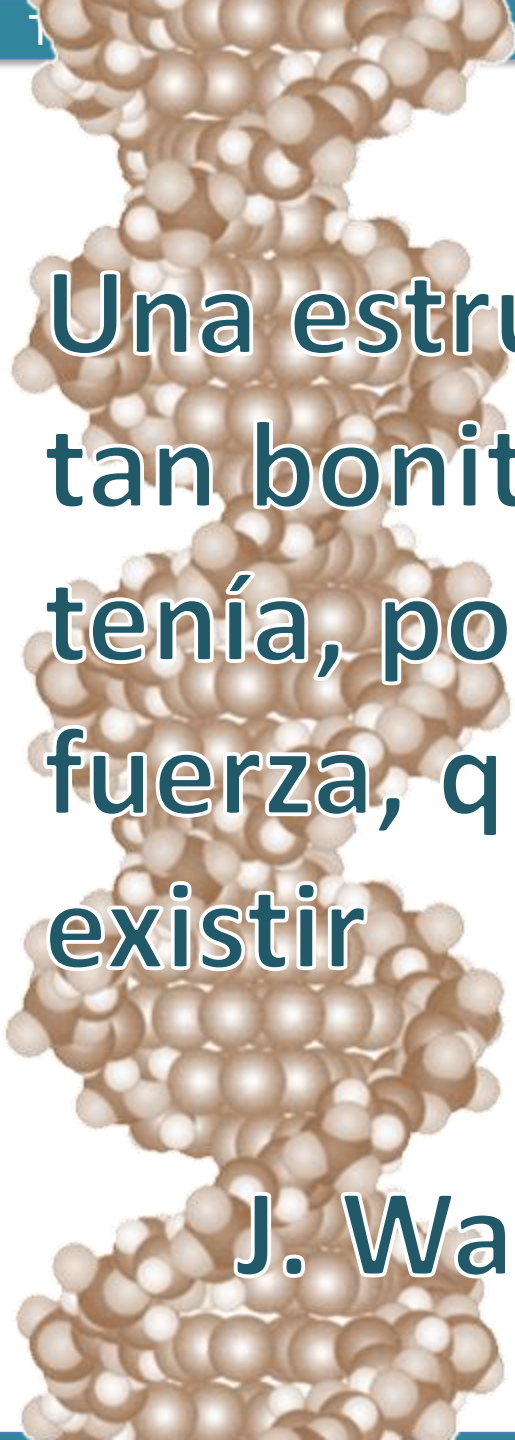




Mensaje codificado **...CACTCTACTG...**





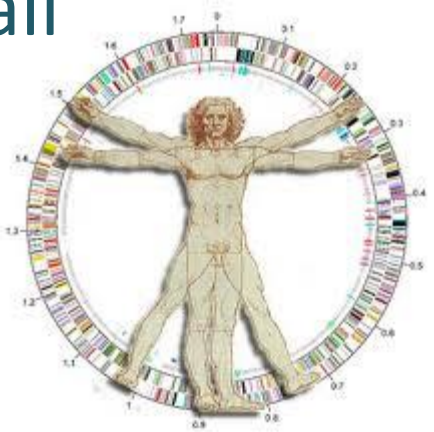


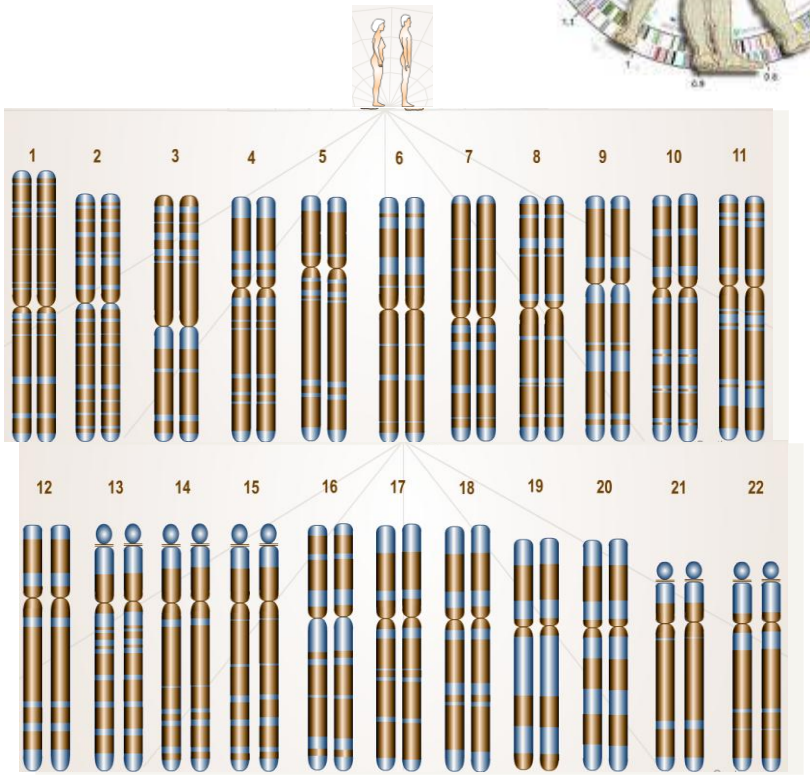
Una estructura tan bonita tenía, por fuerza, que existir

J. Watson

The human genome: The Holy Grail

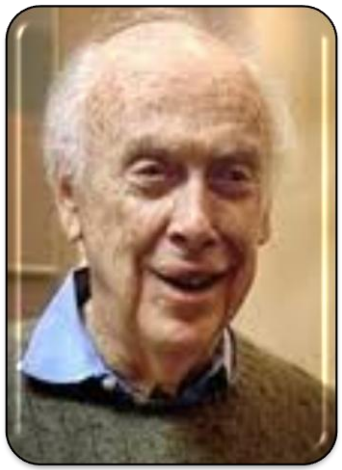
```
DEFINICIÓN     Genoma humano
LOCUS          HSGP010101      3.158 Gb       ADN
FECHA          04-25-03
VERSIÓN        Ensamblaje 1.0
ORGANISMO      «Homo sapiens»
   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

TÍTULO         La secuencia
/fuente        1, 3150000000
/cromosoma     "1-22, X, Y"
/nota          «Libro de la vida, santo grial, mapa humano»

INICIO
1    agctcgctga gacttctgg  accccgcacc aggctgtggg gtttctcaga taactgggcc
61   cctgcgctca ggaggccttc accctctgct ctgggtaaag ttcattggaa cagaaagaaa
121  tggatttatc tgctcttcgc gttgaagaag tacaaaatgt cattaatgct atgcagaaaa
181  tcttagagtg tcccatctgt ctggagttga tcaaggaacc tgtctccaca aagtgtgacc
241  acatattttg caaattttgc atgctgaaac ttctcaacca gaagaaaggg ccttcacagt
301  gtcctttatg taagaatgat ataaccaaaa ggaggcctaca agaaagtacg agatttagtc
361  aacttgttga agagctattg aaaatcattt gtgcttttca gcttgacaca ggtttggagt
421  atgcaaacag ctataatttt gcaaaaaagg aaaataactc tcctgaacat ctaaaagatg
481  aagtttctat catccaaagt atgggctaca gaaaccgtgc caaaagactt ctacagagtg
541  aacccgaaaa tccttccttg caggaaacca gtctcagtgt ccaactctct aaccttggaa
601  ctgtgagaac tctgaggaca aagcagcgga tacaacctca aaagacgtct gtctacattg
661  aattgggatc tgattcttct gaagataccg ttaataaggc aacttattgc agtgtgggag
721  atcaagaatt gttacaaatc acccctcaag gaaccaggga tgaaatcagt ttggattctg
781  caaaaaaggc tgcttgtgaa ttttctgaga cggatgtaac aaatactgaa catcatcaac
841  ccagtaataa tgatttgaac accactgaga agcgtgcagc tgagaggcat ccagaaaagt
901  atcagggtag ttctgtttca aacttgcatg tggagccatg tggcacaaat actcatgcca
961  gctcattaca gcatgagaac agcagtttat tactcactaa agacagaatg aatgtagaaa
1021 aggctgaatt ctgtaataaa agcaaacagc ctggcttagc aaggagccaa cataacagat
1081 gggctggaag taaggaaaca tgtaatgata ggcggactcc cagcacagaa aaaaaggtag
1141 atctgaatgc tgatcccctg tgtgagagaa aagaatggaa taagcagaaa ctgccatgct
1201 cagagaatcc tagagatact gaagatgttc cttggataac actaaatagc agcattcaga
1261 aagttaatga gtggttttcc agaagtgatg aactgttagg ttctgatgac tcacatgatg
1321 gggagtctga atcaaatgcc aaagtagctg atgtattgga cgttctaaat gaggtagatg
1381 aatattctgg ttcttcagag aaaatagact tactggccag tgatcctcat gaggctttaa
1441 tatgtaaaag tgaaagagtt cactccaaat cagtagagag taatattgaa gacaaaatat
1501 ttgggaaaac ctatcggaag aaggcaagcc tccccaactt aagccatgta actgaaaatc
1561 taattatagg agcatttgtt actgagccac agataataca agagcgtccc ctcacaaata
1621 aattaaagcg taaaaggaga cctacatcag gccttcatcc tgaggatttt atcaagaaag
1681 cagatttggc agttcaaaag actcctgaaa tgataaatca gggaactaac caaacggagc
1741 agaatggtca agtgatgaat attactaata gtggtcatga gaataaaaca aaaggtgatt
1801 ctattcagaa tgagaaaaat cctaacccaa tagaatcact cgaaaaagaa tctgctttca
1861 aaacgaaagc tgaacctata agcagcagta taagcaatat ggaactcgaa ttaaatatcc
1921 acaattcaaa agcacctaaa aagaataggc tgaggaggaa gtcttctacc aggcatattc
1981 atgcgcttga actagtagtc agtagaaatc taagcccacc taattgtact gaattgcaaa
2041 ttgatagttg ttctagcagt gaagagataa agaaaaaaaa gtacaaccaa atgccagtca
```





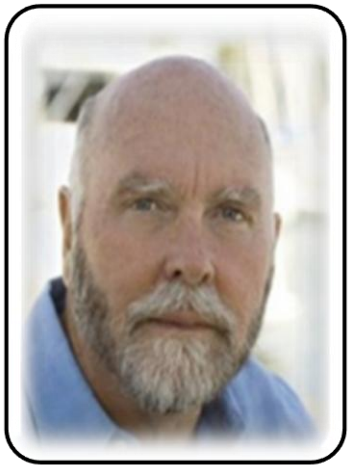
The mapping of the human genome is 'the greatest intellectual moment in history.'

Matt Ridley



It's a giant resource that will change mankind, like the printing press

James Watson



This period is a very historic time, a new starting point

Craig Venter

The achievements of the HGP has radically changed the practice of biomedical research

Distinguishing characteristics of Genomics

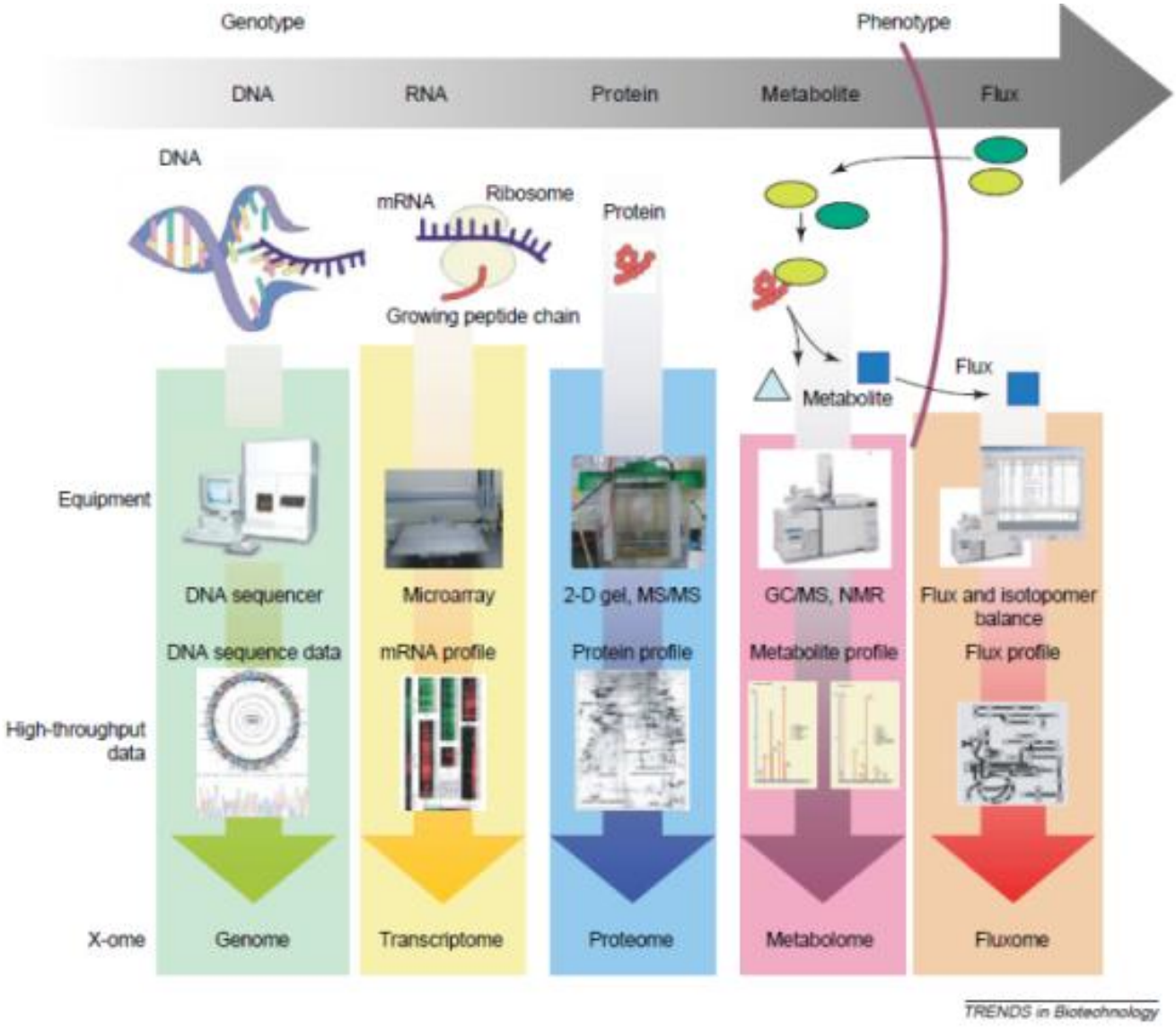
International Human Genome Sequencing Consortium*

- Big teams -> Multidisciplinary and international teams
- New high-throughput technologies for large-scale data production (Omics)
- “Discovery science” or “Data driven” approach vs. “hypothesis driven” approach
- Computational intensity and expertise
- High standard for data quality
- Rapid data release
- Attention to societal implications

Barcelona Computing Center (BCS)


THE
FUTURE

Omics


Genotype
Phenotype
DNA
RNA
Protein
Metabolite
Flux
DNA
mRNA
Ribosome
Protein
Growing peptide chain
Metabolite
Flux
Equipment
DNA sequencer
Microarray
2-D gel, MS/MS
GC/MS, NMR
Flux and isotopomer balance
DNA sequence data
mRNA profile
Protein profile
Metabolite profile
Flux profile
High-throughput data
X-ome
Genome
Transcriptome
Proteome
Metabolome
Fluxome
TRENDS in Biotechnology

The achievements of the HGP has radically changed the practice of biomedical research

Distinguishing characteristics of Genomics

International Human Genome Sequencing Consortium*

- Big teams -> Multidisciplinary and international teams
- New high-throughput technologies for large-scale data production (Omics)
- "Discovery science" or "Data driven" approach vs. "hypothesis driven" approach
- Computational intensity and expertise
- High standard for data quality
- Rapid data release
- Attention to societal implications

Barcelona Computing Center (BCS)

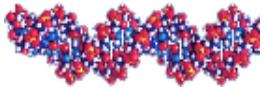
BIG DATA SCIENCE

The Fundamental Question

Clarividencia René Magritte

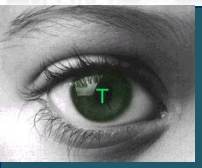
Fenotipo

DNA



¿Cómo se decodifica el mensaje genético para formar el fenotipo?

The HGP has had a profound consequence in the conceptualization of biological systems

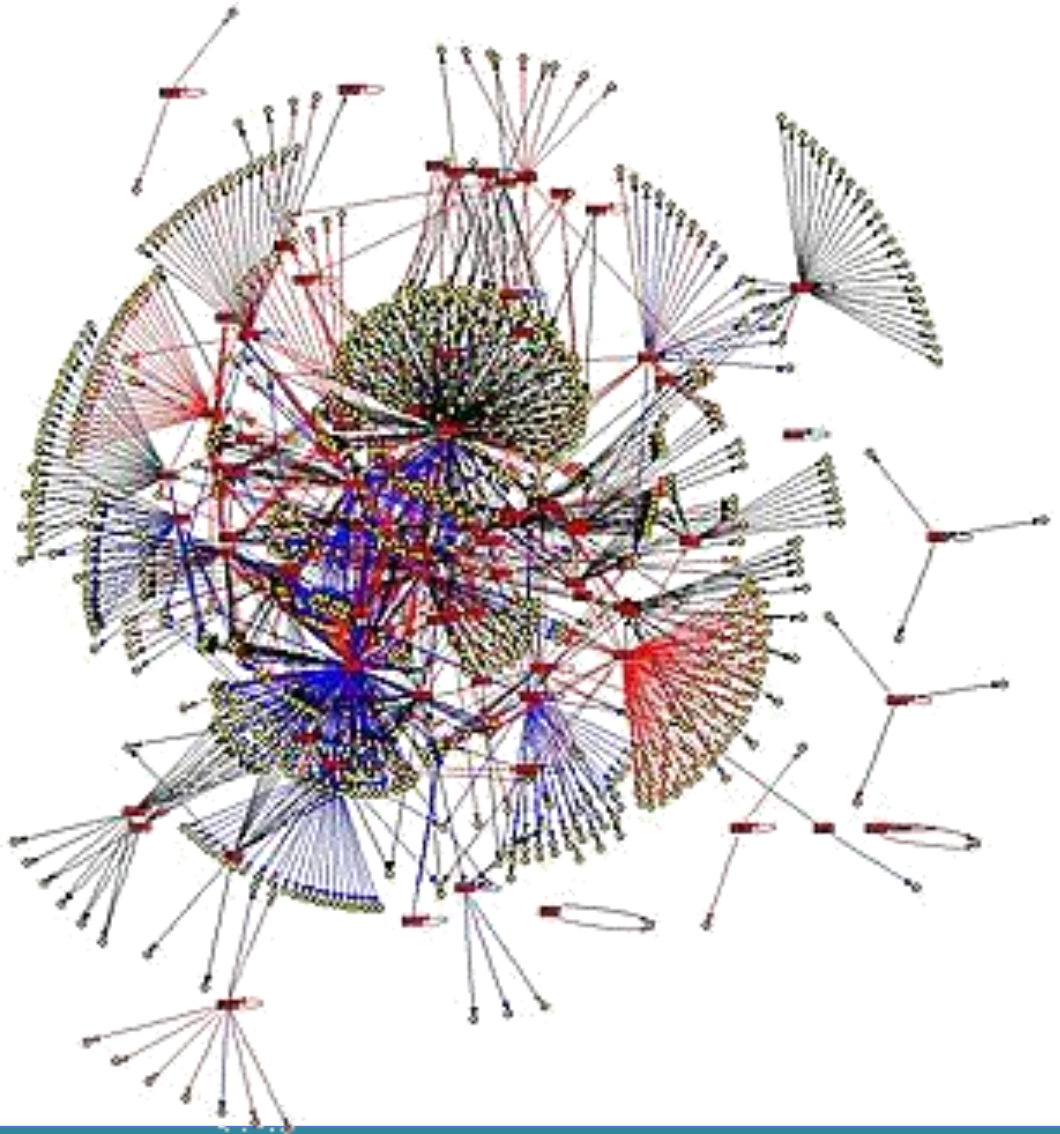


The HGP has changed the way we conceptualize molecular biology

- The analysis of biological systems in terms of the storage, transmission and transformation of the information coded in the genomes

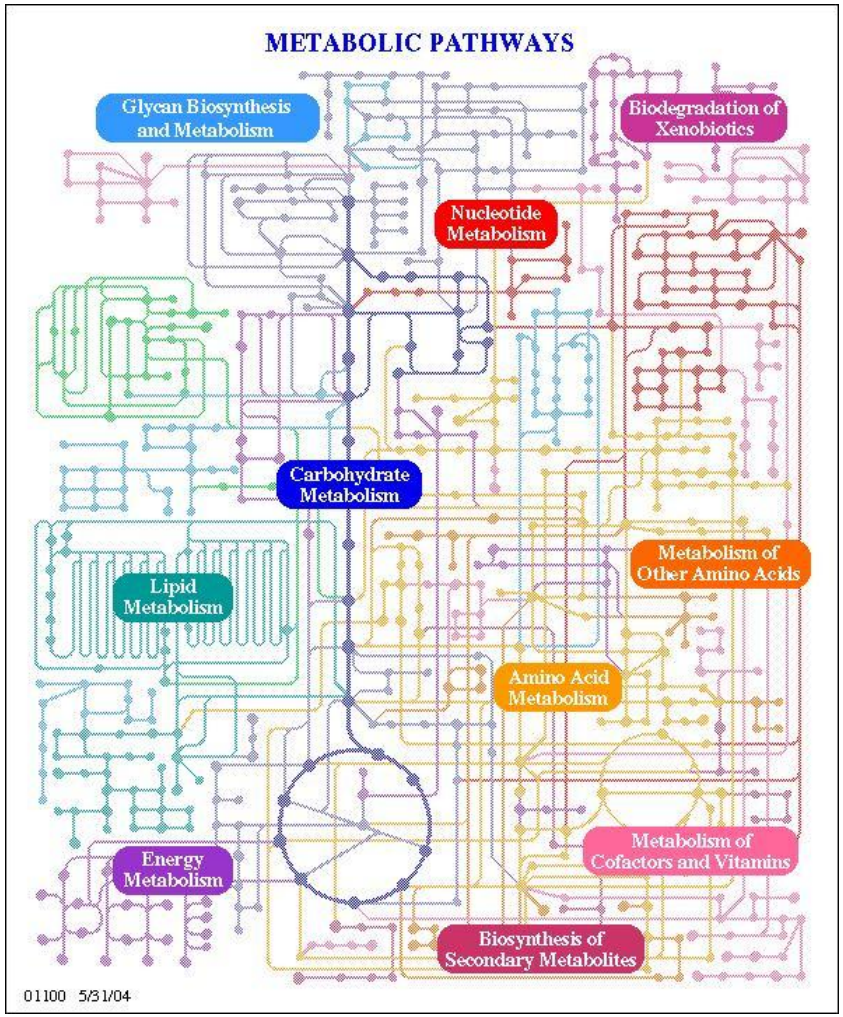
Integrative Biology

From the new paradigma, biological systems are complex networks of myriad of pathways, many of them interconnected



Integrative Biology

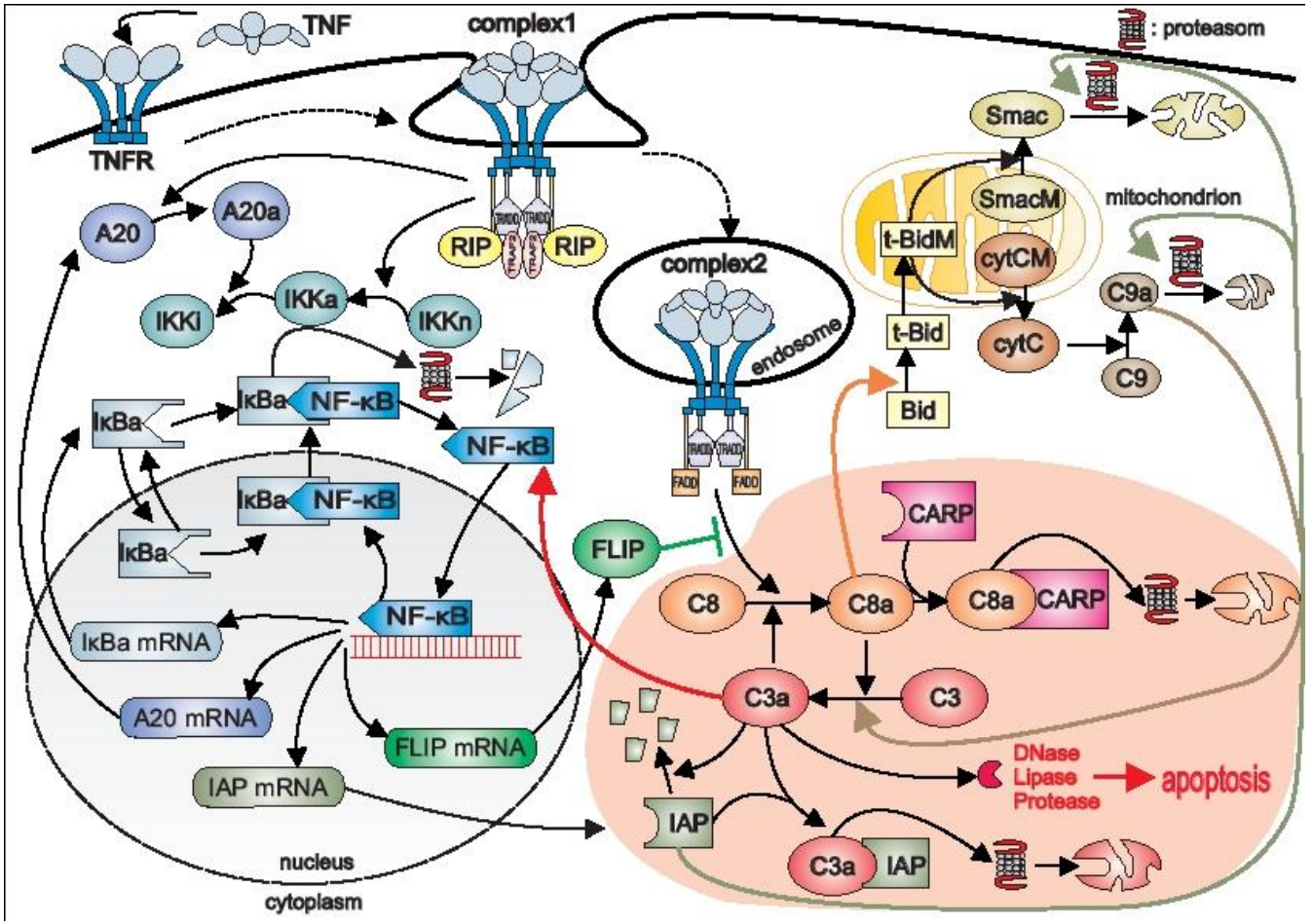
Biosynthesis pathways,



Metabolic pathway

Integrative Biology

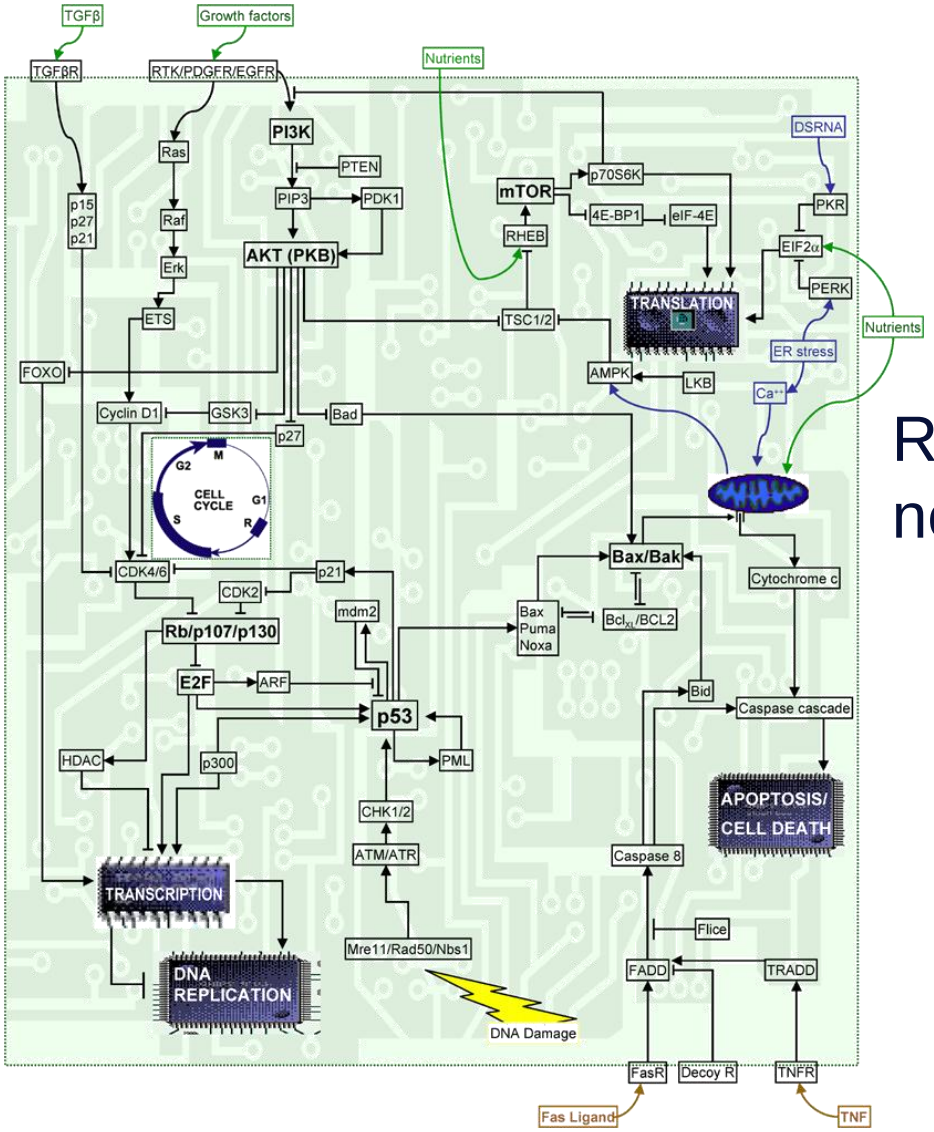
Signal transduction pathways,



Signal transduction pathways

Integrative Biology

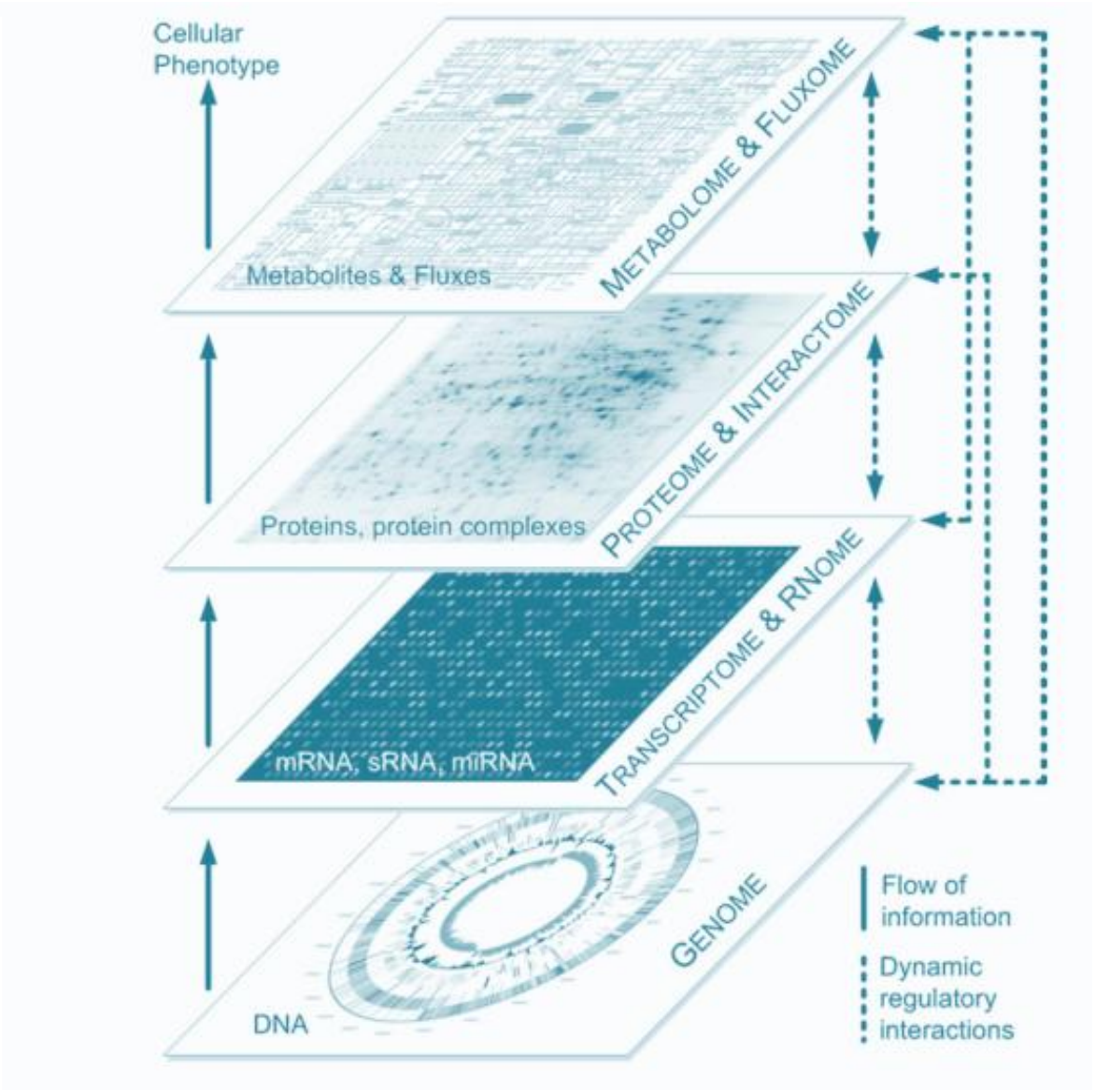
Regulatory networks.



Regulatory networks

Kohlsted et al. 2010

Integrative Biology



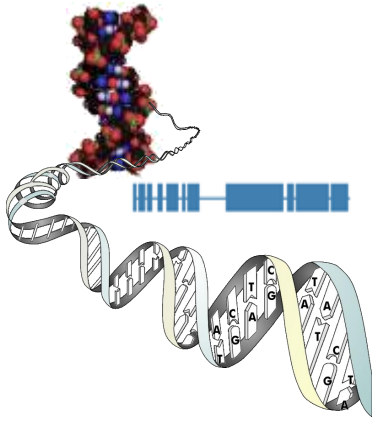
The HGP has had a profound consequence in the **conceptualization of biological systems**

Era of the organism reconstruction: synthetic approach, interdisciplinary and big research teams to explain the emergent properties of biological systems

Biological System (Organism)

20th century biology

Reductionist approach (Experiments)



Construction blocks (Genes/Molecules)

Synthetic and interdisciplinary approach

(Bioinformaticians, Biologists, Statisticians, Mathematicians, Biochemists, Physicist, Medical Doctors,…)

21st century biology

Genome sequencing

Why sequencing a genome?

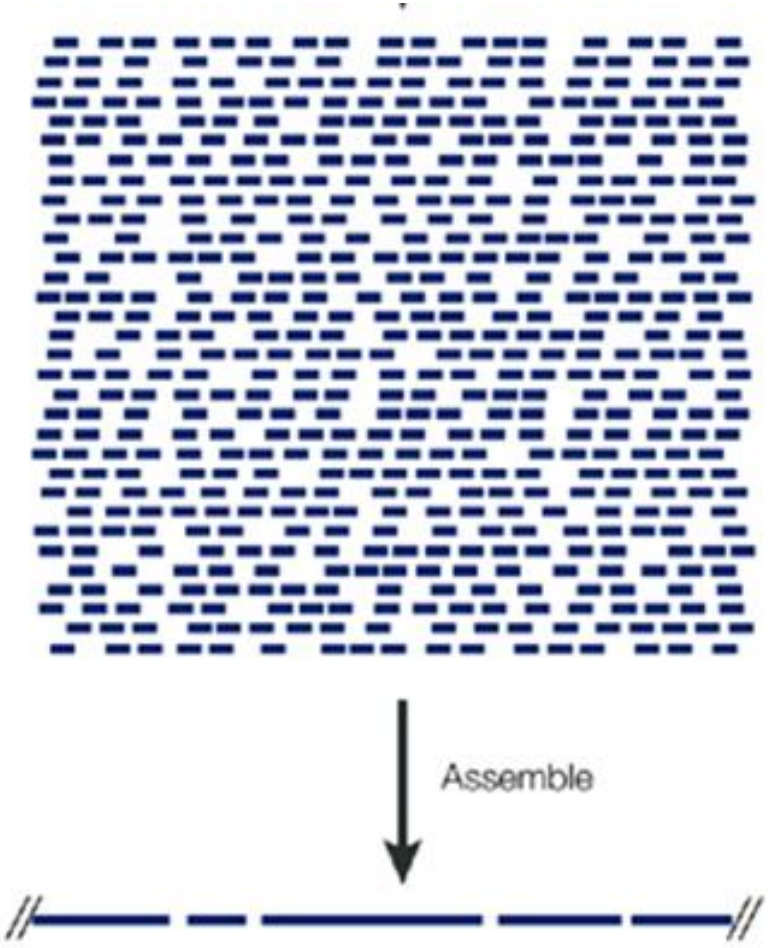
Each genome sequence is a **treasure trove** containing an endless and invaluable source of biological information

- Knowledge of the number, structure of genes and regulatory (functional) regions
- Basic Principles on the organization of the organism (functional classes, ...)
- Learn basic functions of genes conserved in different species (molecular biology lexicon)
- Chromosomal organization
- Genome evolution (conservation of gene order, sequence evolution)
- Genome variation (Population genomics)
- Association studies
- Expression analysis
- Integrative genomics (System biology)
- Applied genomics (Personalized medicine, Pharmacogenomics, Nutrigenomics, Agrigenomics, Conservation biology, Bioremediation,…)
- New areas of enquiry (Metagenomics, gene regulation, life evolution, human diaspora, medical enquiry, ethics, law,..)

We look at the forest, not a particular tree

The pervasive assembly puzzle

How to map millions or billions of short read fragments onto a genome?

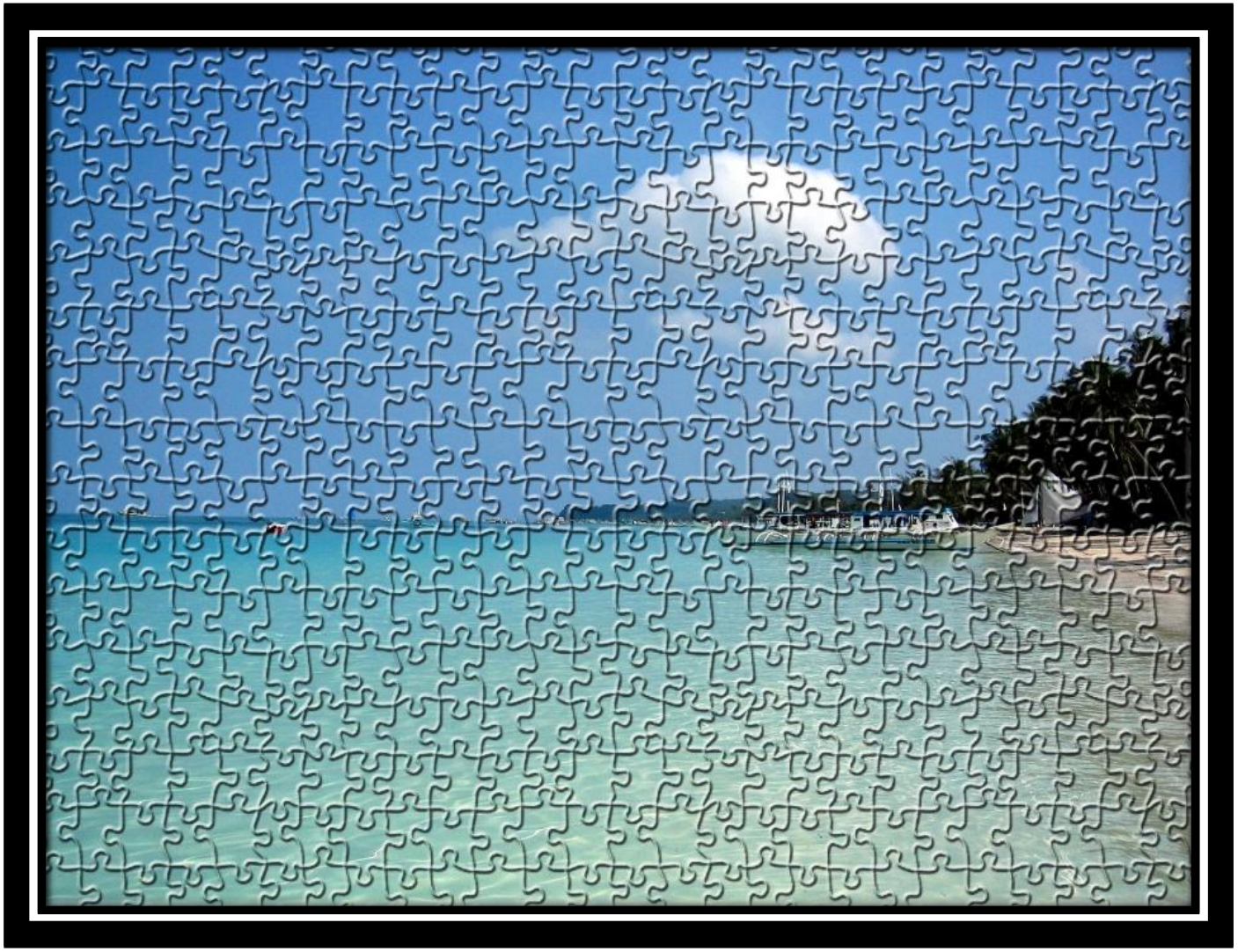


The assembly puzzle

Complex pattern (prokaryote genomes)

The assembly puzzle

Simple pattern (genomes with high amount of repetitive sequence)



Genome sequence metrics

Redundancy = Fold coverage

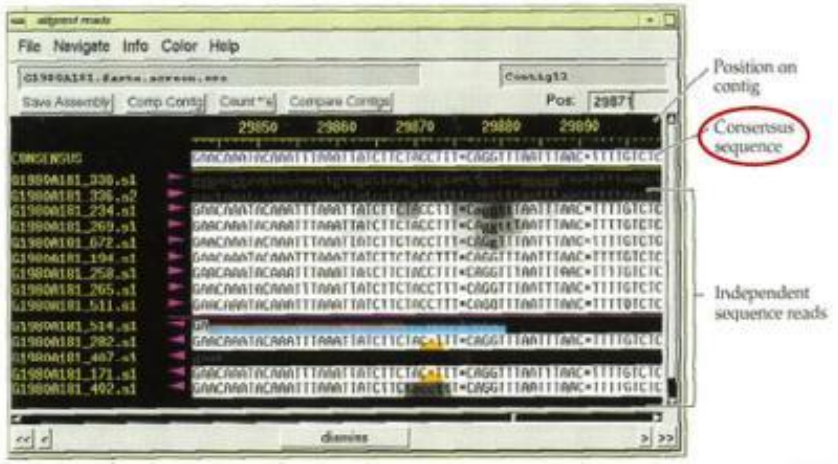
$$FC = \frac{N \cdot L}{G}$$

N = number of reads
L = mean read length
G = genome size
10 x is considered high quality

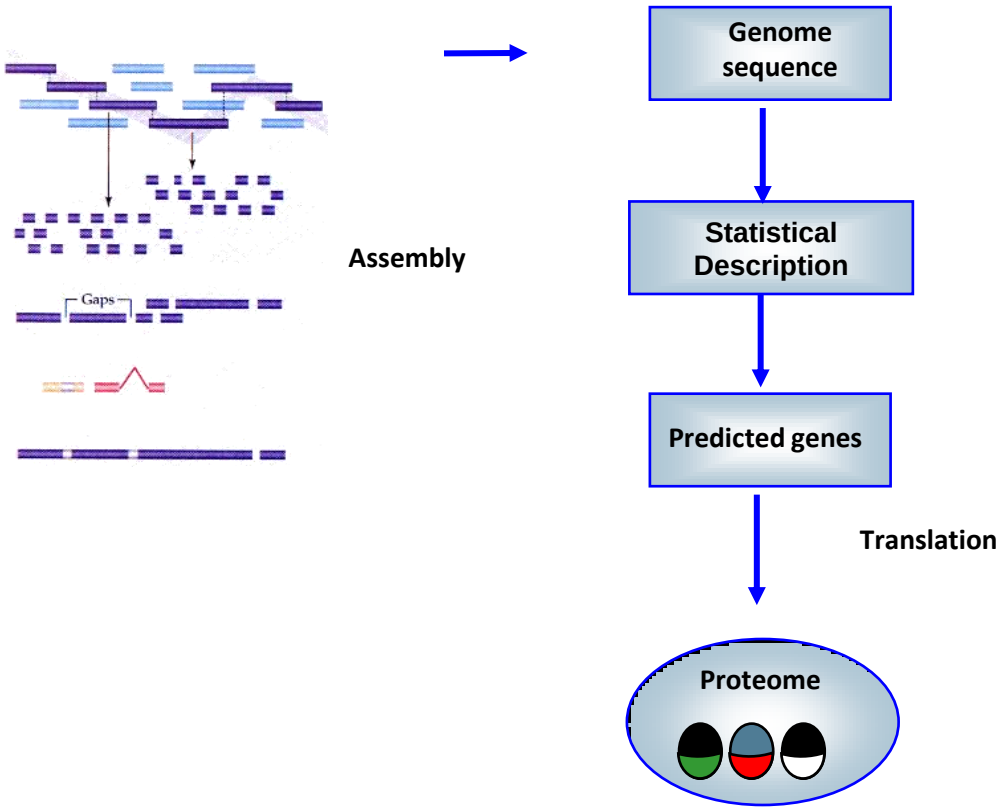
Base quality => phred score (Q)

$$Q = -10 \log_{10} P$$

P = Probability of calling a wrong base
Q = 20 draft sequence (P = 0.01)
Q = 40 finished sequence (P = 0.0001)



Steps of genome analysis



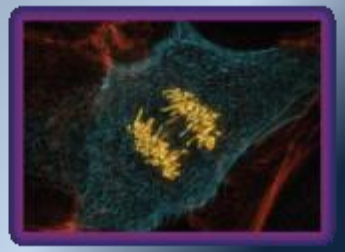
- DNA annotation and functional genomics
- Expression
- Proteins
- Molecular Evolution
- Genome variations and genotype-phenotype association studies
- System Biology

The technological

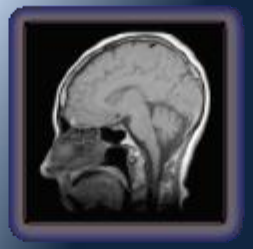
explosion

Technology Advances Drive Science


Astronomy

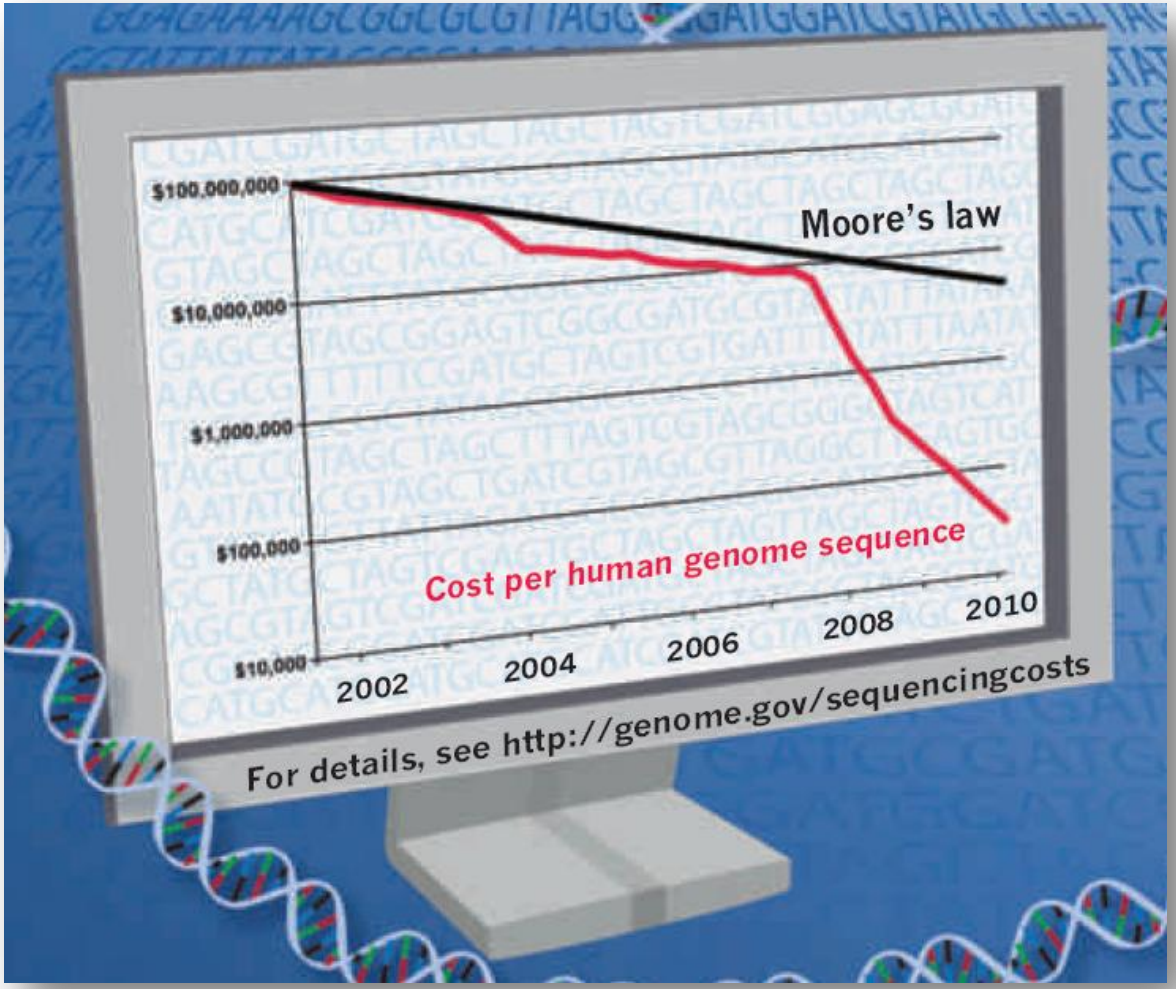
Cell Biology

Radiology

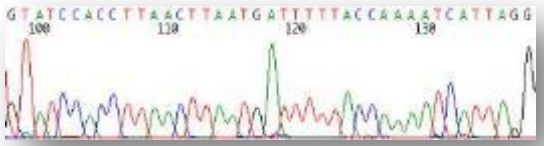

Genomics

The technological explosion



Sequencing techologies

- Sanger (the *Gold Standard*) (http://www.wiley-vch.de/books/sample/3527320903_c01.pdf) Capillary electroforesis. Reads (trace files) of 500-700 bases.



- Next Generation Technologies (starting in 2005):

Massively parallel sequencing
($1000 genome' program NHGRI)

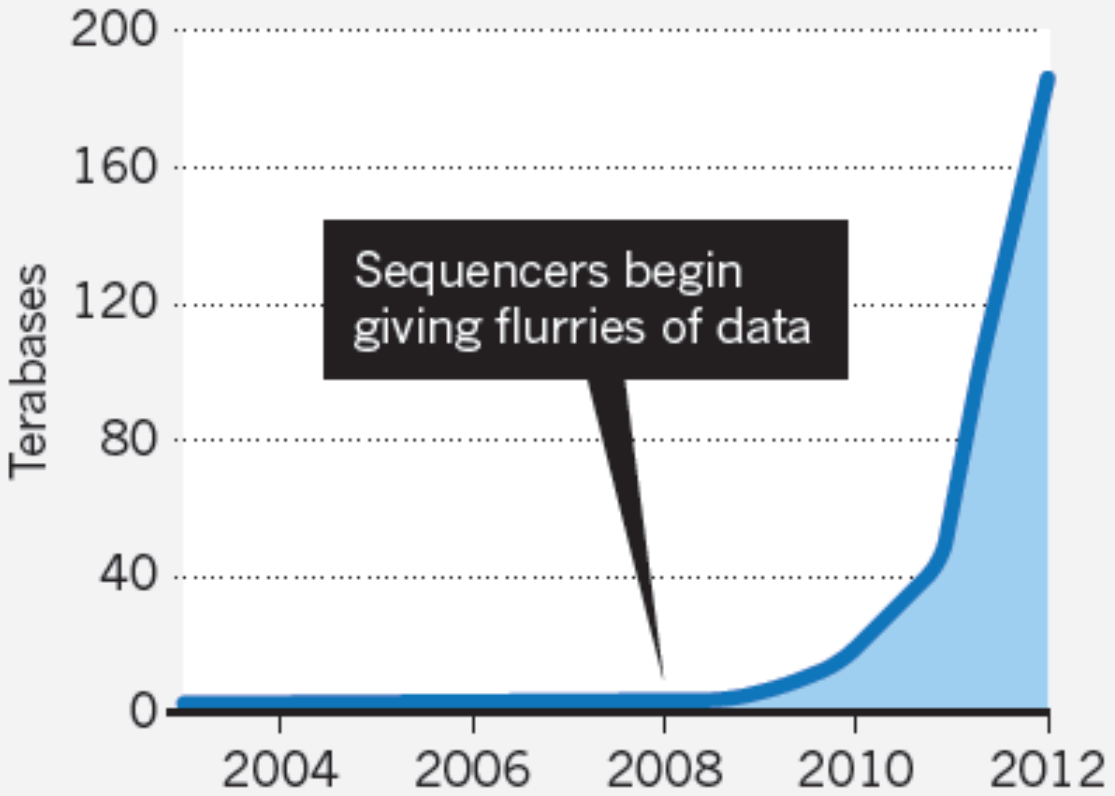
- Dramatic increase of sequence output
- Significant decrease in the length of reads
- Decrease in the accuracy of calling bases -> Very different error profiles depending on the technological platform

↓

Analytical difficulties (profound changes in data analysis pipelines) → Revitalization of bioinformatics

DATA EXPLOSION

The amount of genetic sequencing data stored at the European Bioinformatics Institute takes less than a year to double in size.



20 petabytes ~ 6 millones HG

The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.

Explore the EBI:

Search

Examples: blast, keratin, bfl1...

Press release

Bioinformatics embraces Semantic Web technologies

EMBL-EBI's new Resource Description Framework (RDF) platform provides access to bioinformatics resources that support Semantic Web technologies.

Functional genetic variation in humans: comprehensive map published

GEUVADIS project presents the largest-ever dataset linking human genomes to gene activity at the level of RNA.

It's not just noise

EMBL scientists discern key gene expression patterns in the human genome

Research infrastructures

EMBL-EBI is a pivotal partner in ELIXIR, the European life sciences infrastructure for biological information, as part of the European Strategy on Research Infrastructures (ESFRI) process. On

Popular

- Services
- Research
- Training
- News
- Jobs
- Visit us
- EMBL
- Contacts

Events

From Genome Sequencing to Gene Function - Izmir, Turkey
Oct 26 2013
Registration deadline: Oct 25 2013

Managing and Exploring Next Generation Sequencing Data
Oct 29 2013 -Oct 30 2013
Registration deadline: Oct 25 2013

Understanding 'omics data (joint dixa - Genedata workshop) - Basel, Switzerland
Nov 19 2013 -Nov 21 2013
Registration deadline: Oct 27 2013

Mosquito Informatics (INFRAVEC)
Feb 5 2014 -Feb 6 2014
Registration deadline: Nov 30 2013

Agricultural-Omics
Feb 17 2014 -Feb 21 2014
Registration deadline: Dec 20 2013

Next Generation Sequencing Workshop
Mar 3 2014 -Mar 6 2014
Registration deadline: Jan 3 2014

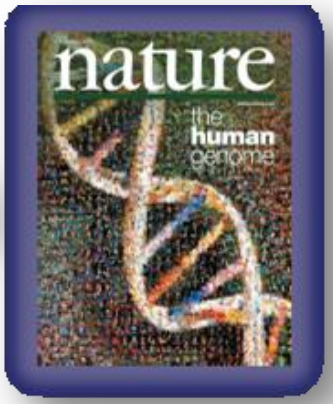
EMBO Practical Course on Metabolomics Bioinformatics for Life Scientists
Mar 17 2014 -Mar 21 2014
Registration deadline: Jan 17 2014

Micro B3 Marine Metagenomics Bioinformatics

http://www.ebi.ac.uk/

The technological explosion

~$1,000,000,000



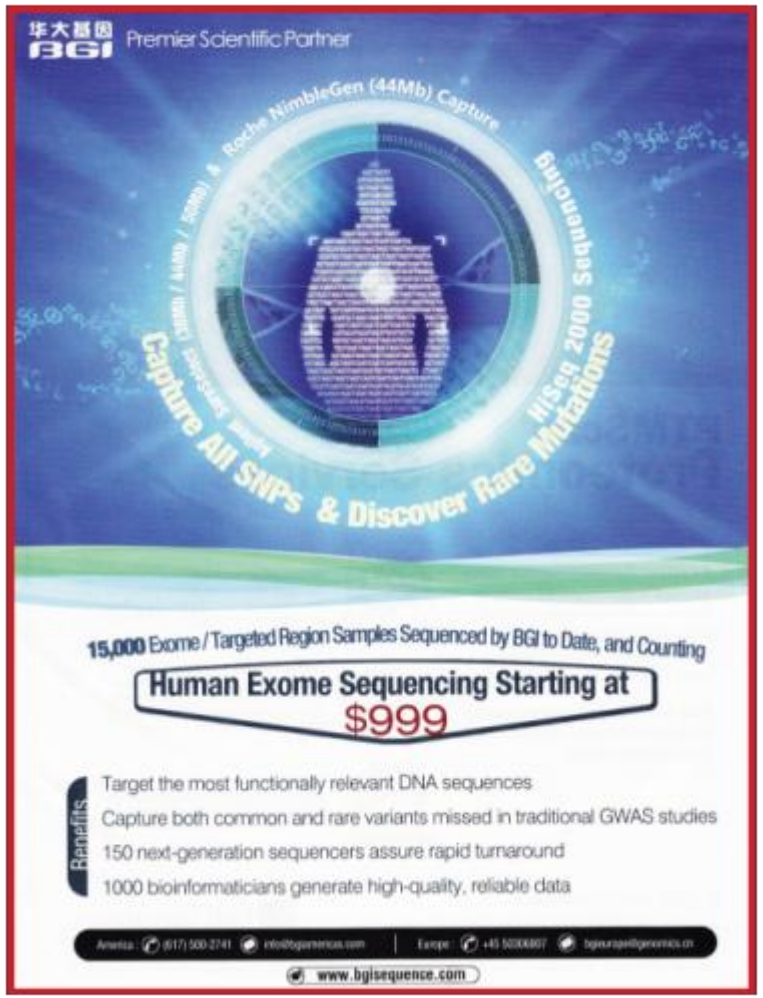
Cost: 1 million fold lower!!!

Today

~$1,000

"The $1000 Genome"

Genome Sequencing as a "Commodity"

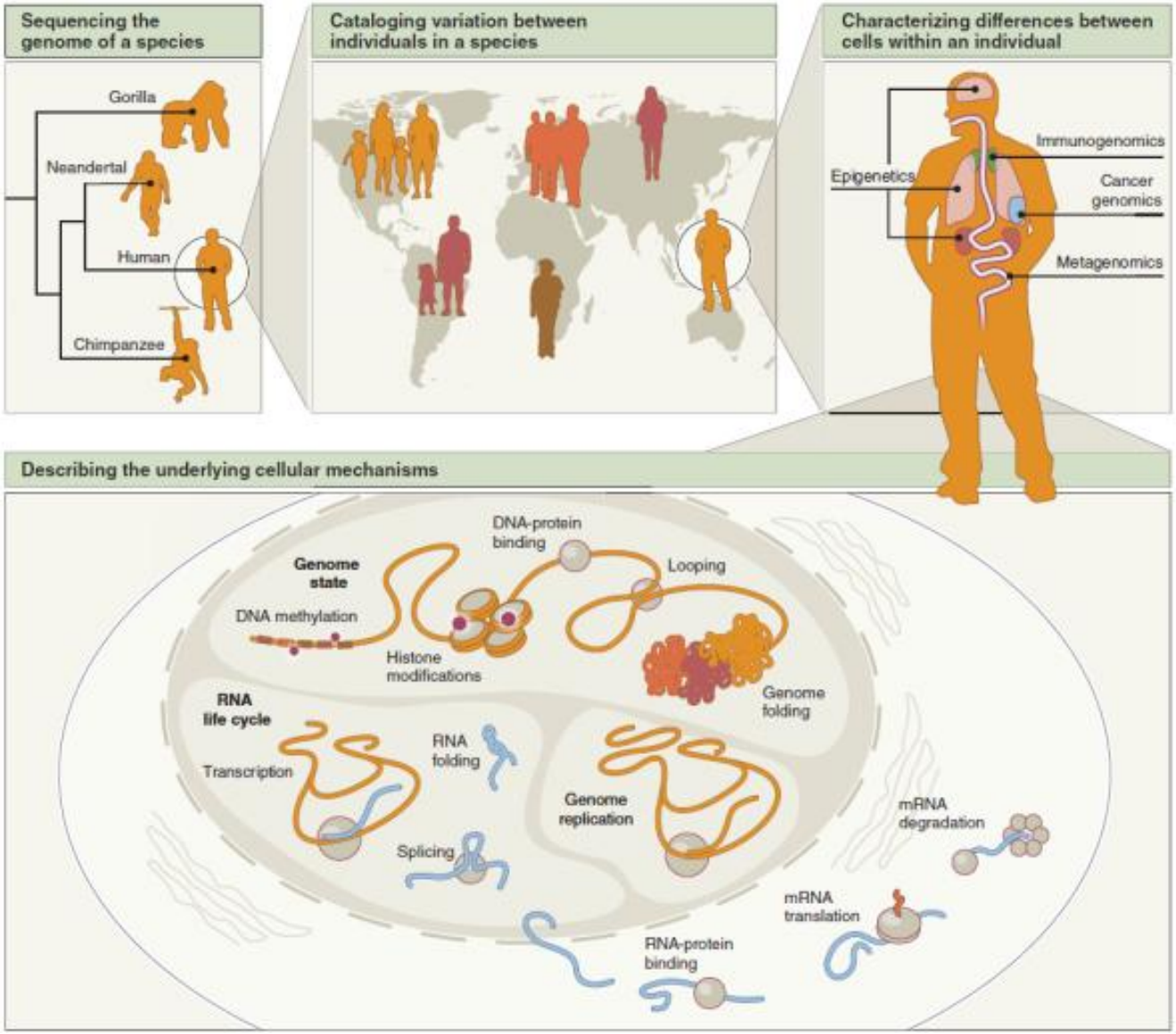


The expanding scope of DNA sequencing

Jay Shendure[1] & Erez Lieberman Aiden[2–4]

VOLUME 30 NUMBER 11 NOVEMBER 2012 NATURE BIOTECHNOLOGY

Road map of sequencing science



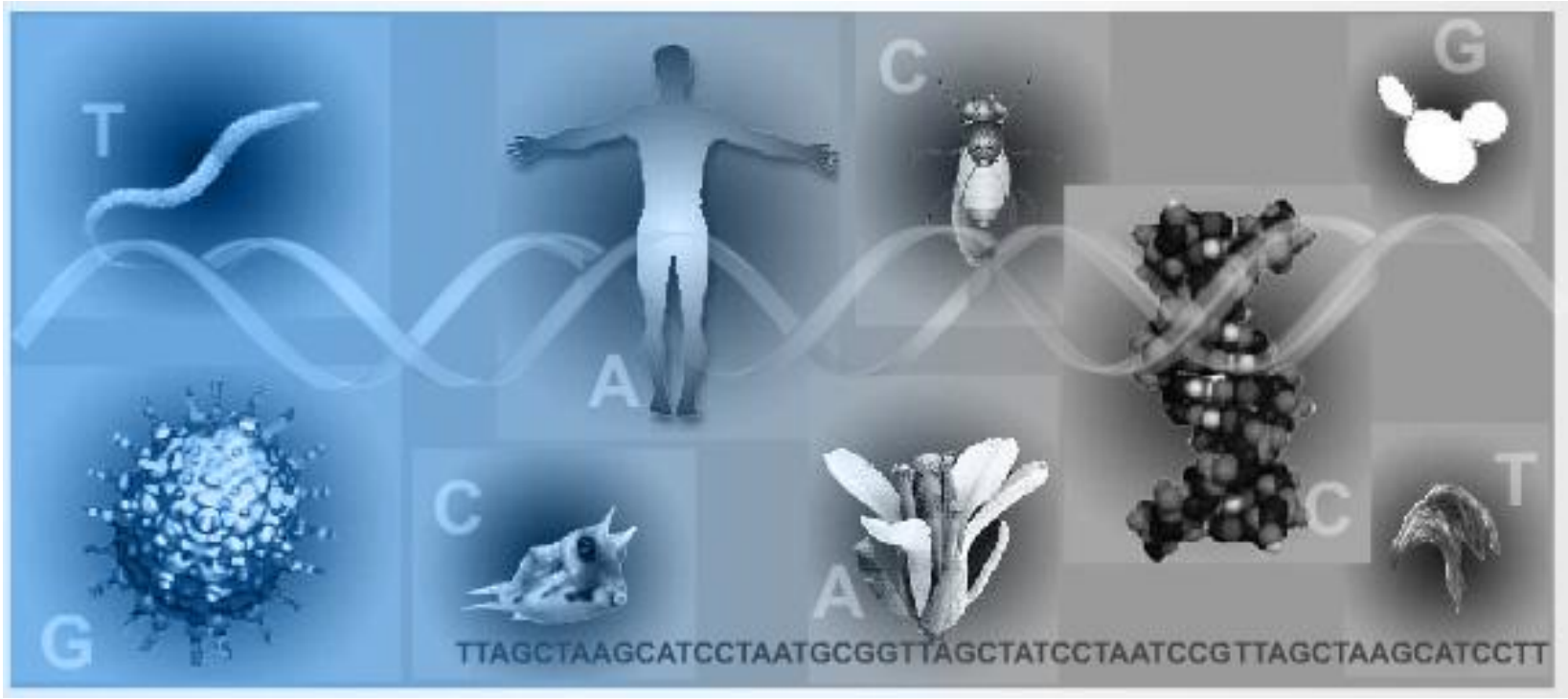
GOLD

Completed and ongoing genome projects

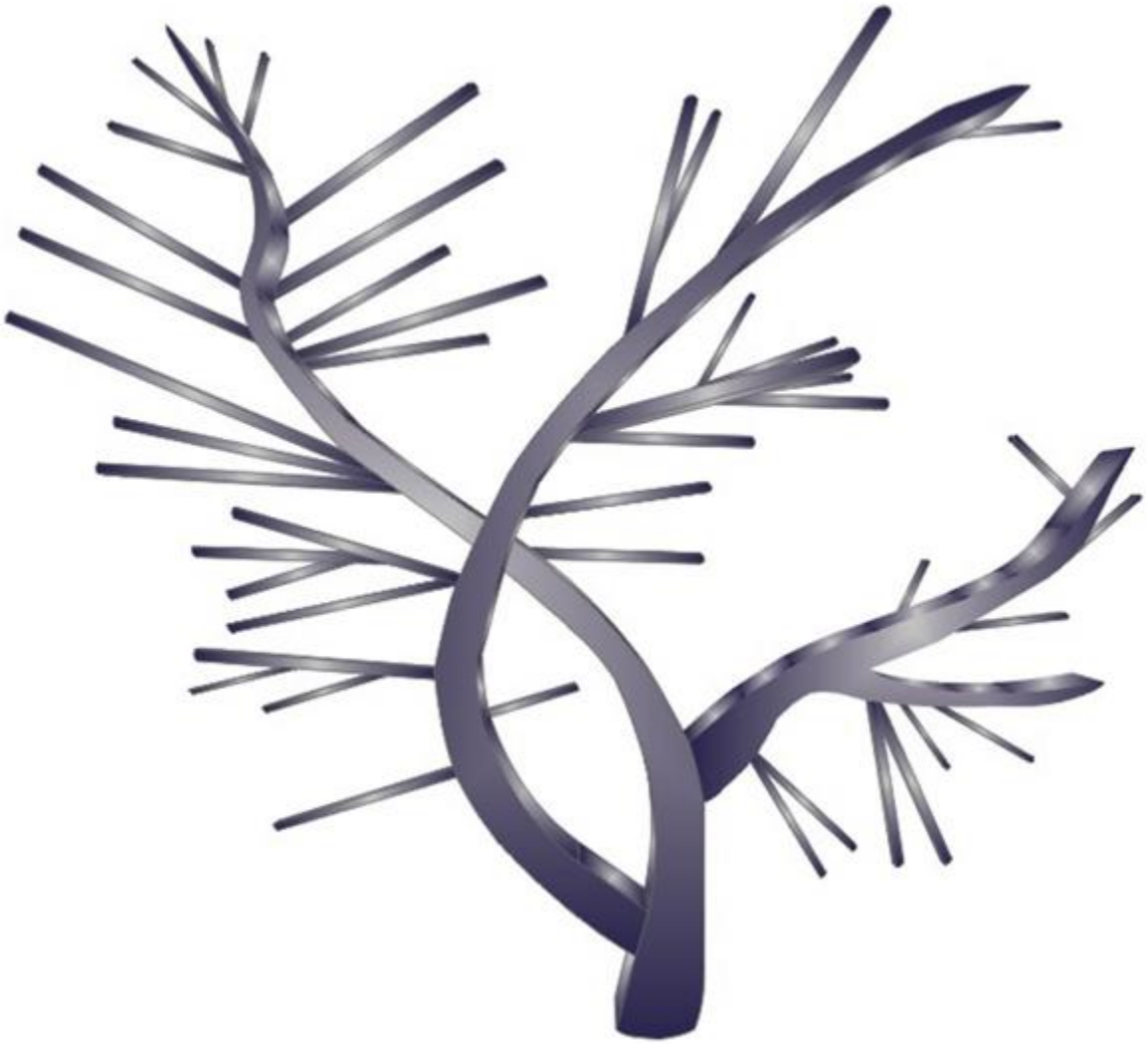
Complete Genome Projects: 7396

Archaeal: 234 Bacterial: 6851 Eukaryal: 311

Finished: 2649 Permanent Draft: 4747



Genomas de especies



Association Studies

Variación fenotípica individual

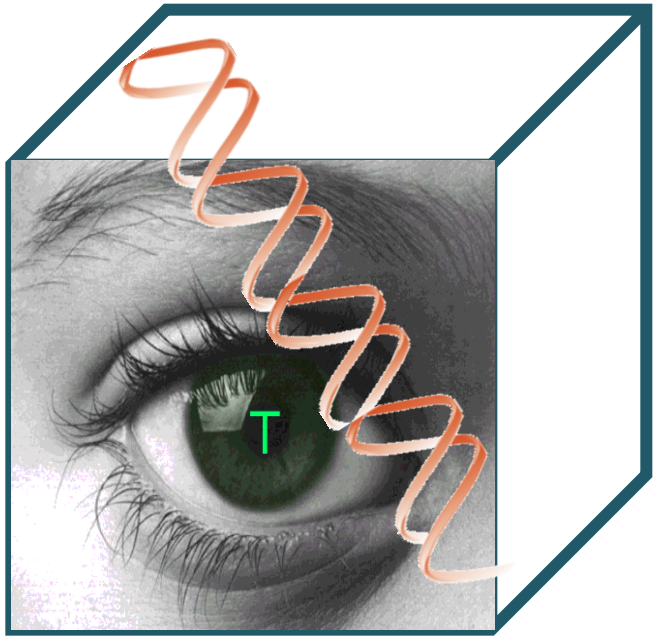
(incluye la predisposición a enfermedades)

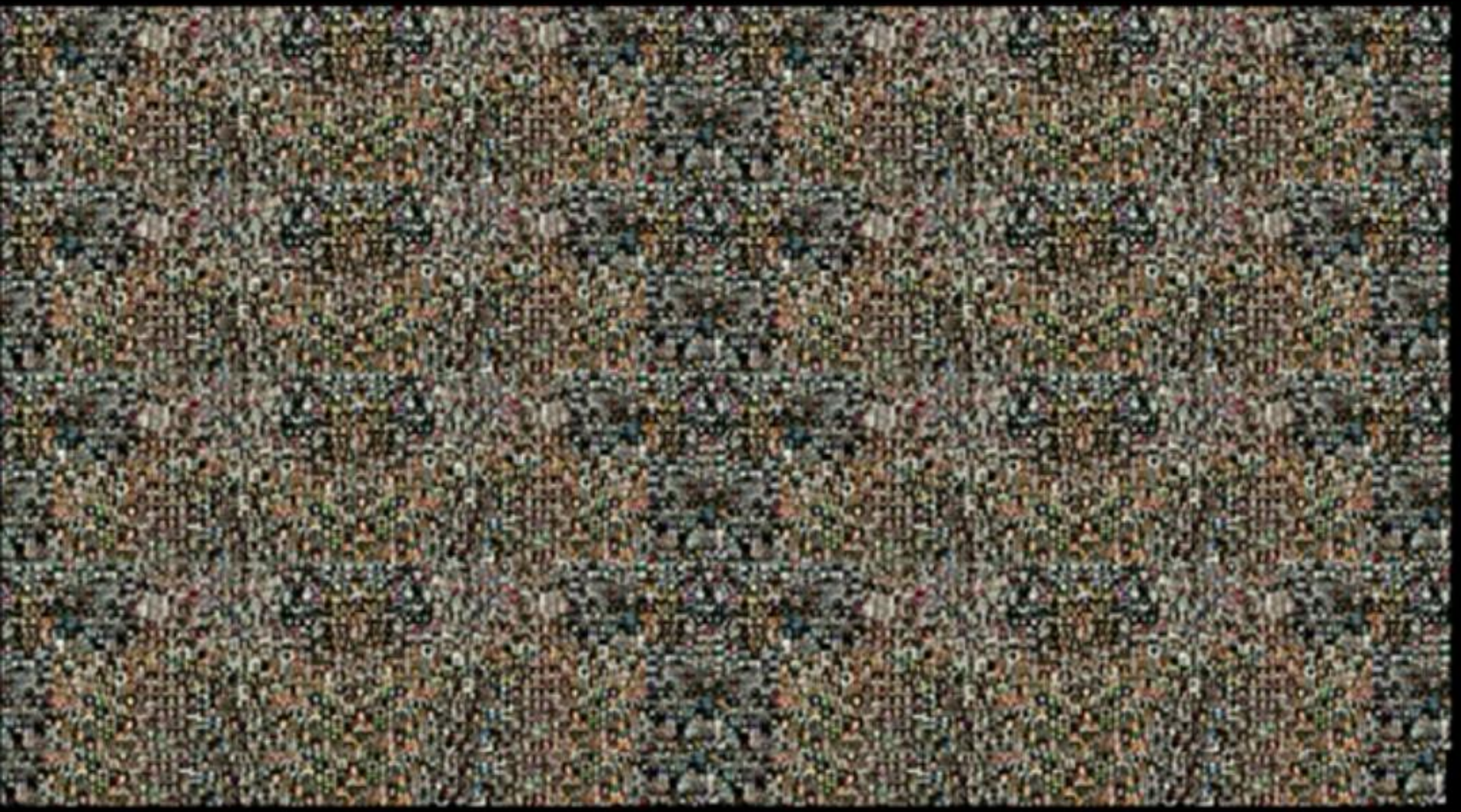
GENOMICS

Fenotipo = Genotipo + Ambiente

The empirical space of human genetic variation:

Data desideratum





Association studies: Phenotpyic effect of SNPs

Human genetic & phenotypic diversity database

	Genotype				Phenotype	
	SNP_1	SNP_2	SNP_3	…	Disease 1	Trait i
Secuence individual 1	A/A	G/C	G/T		Healthy	x_1
Secuence individual 2	A/C	C/C	T/T		Cervical Cancer	x_2
…					…	…

Estimation phenotypic effect

BioBanks: Studies of cohorts at a great scale

The UK Biobank — A study of genes, environment and health

USA

- deCODE (Islandia)
- Estonia
- Germany
- Canada
- Japan
- China

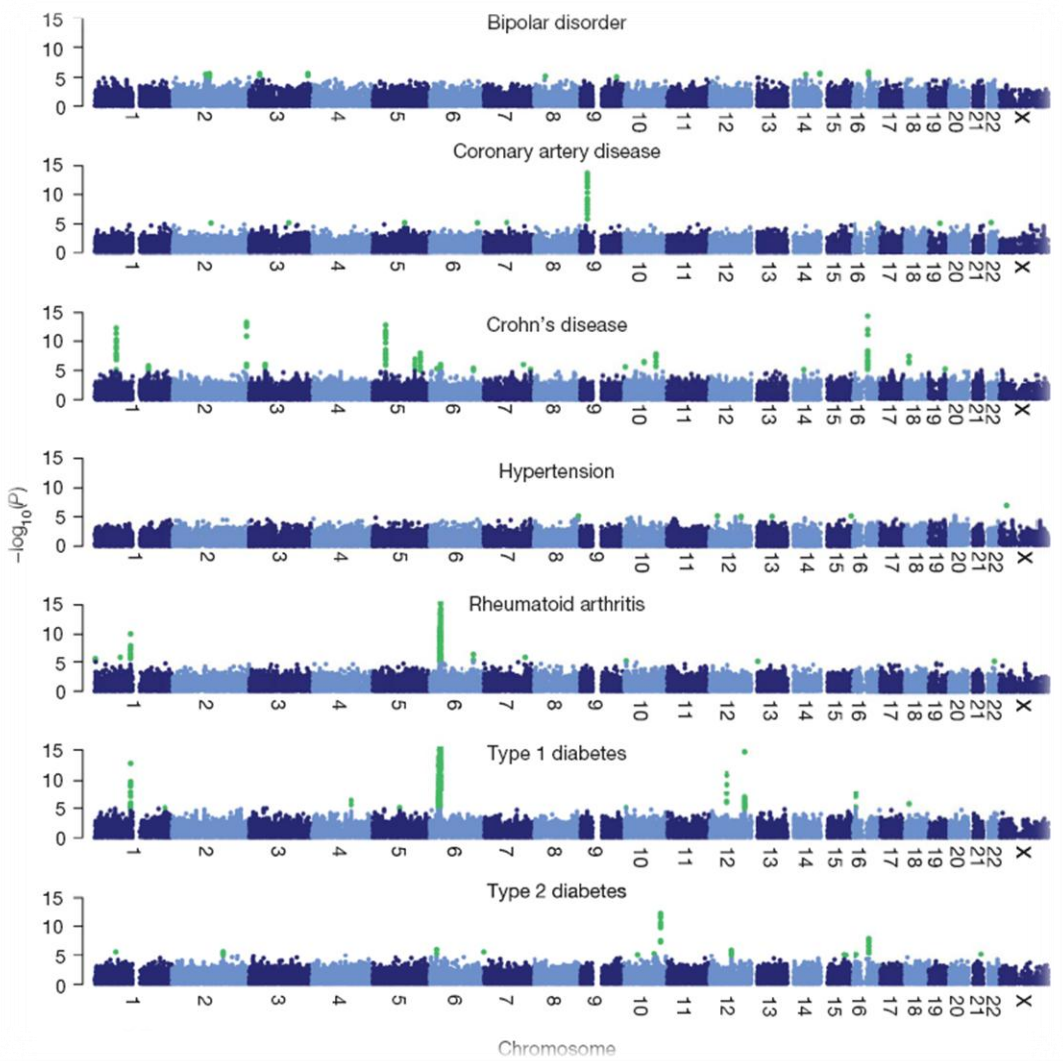
Vol 447|7 June 2007|doi:10.1038/nature05911 nature

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

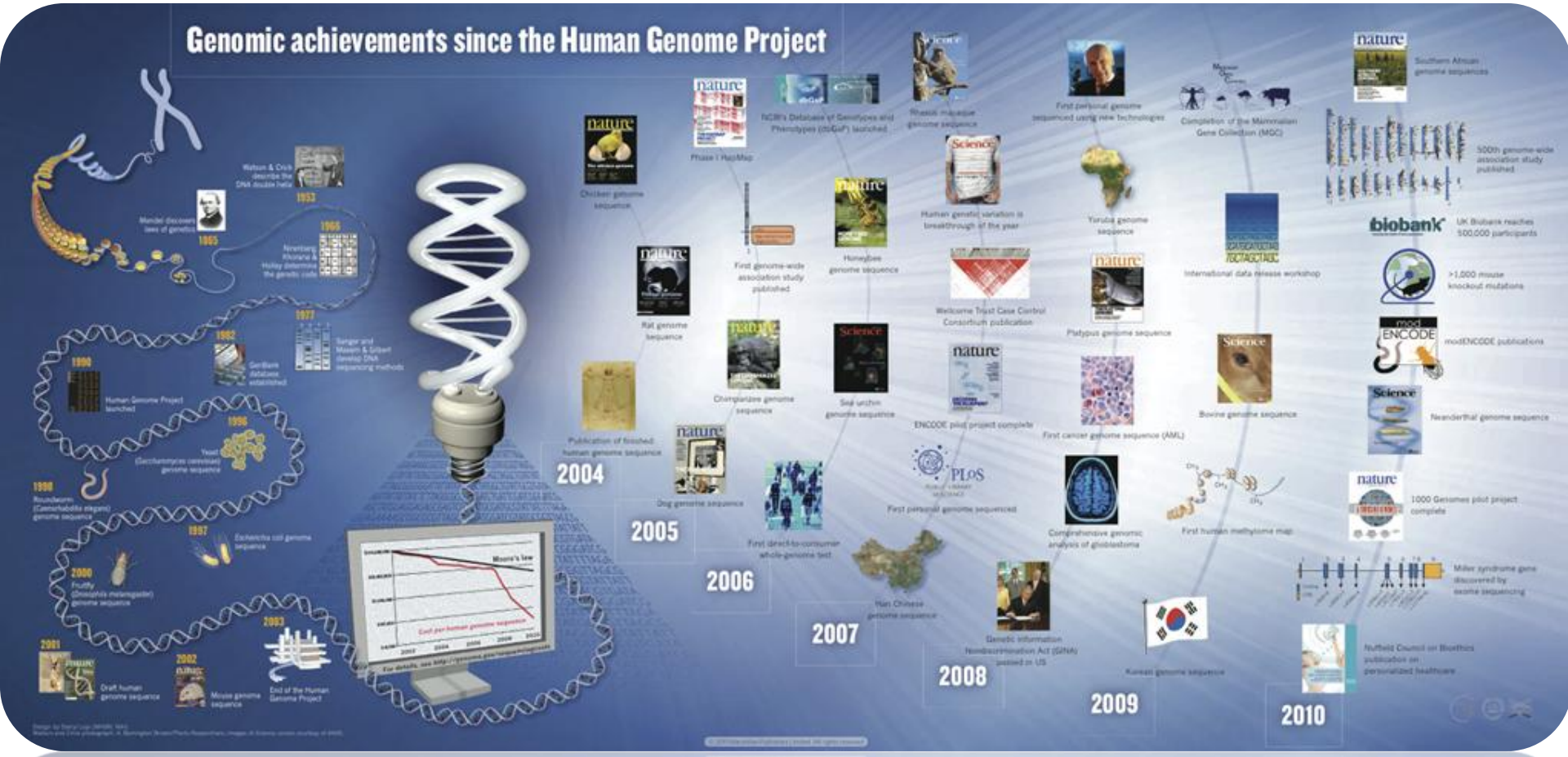
GWA Big Data problems

- Data bias
- False positives
- Data structure
- Security issues



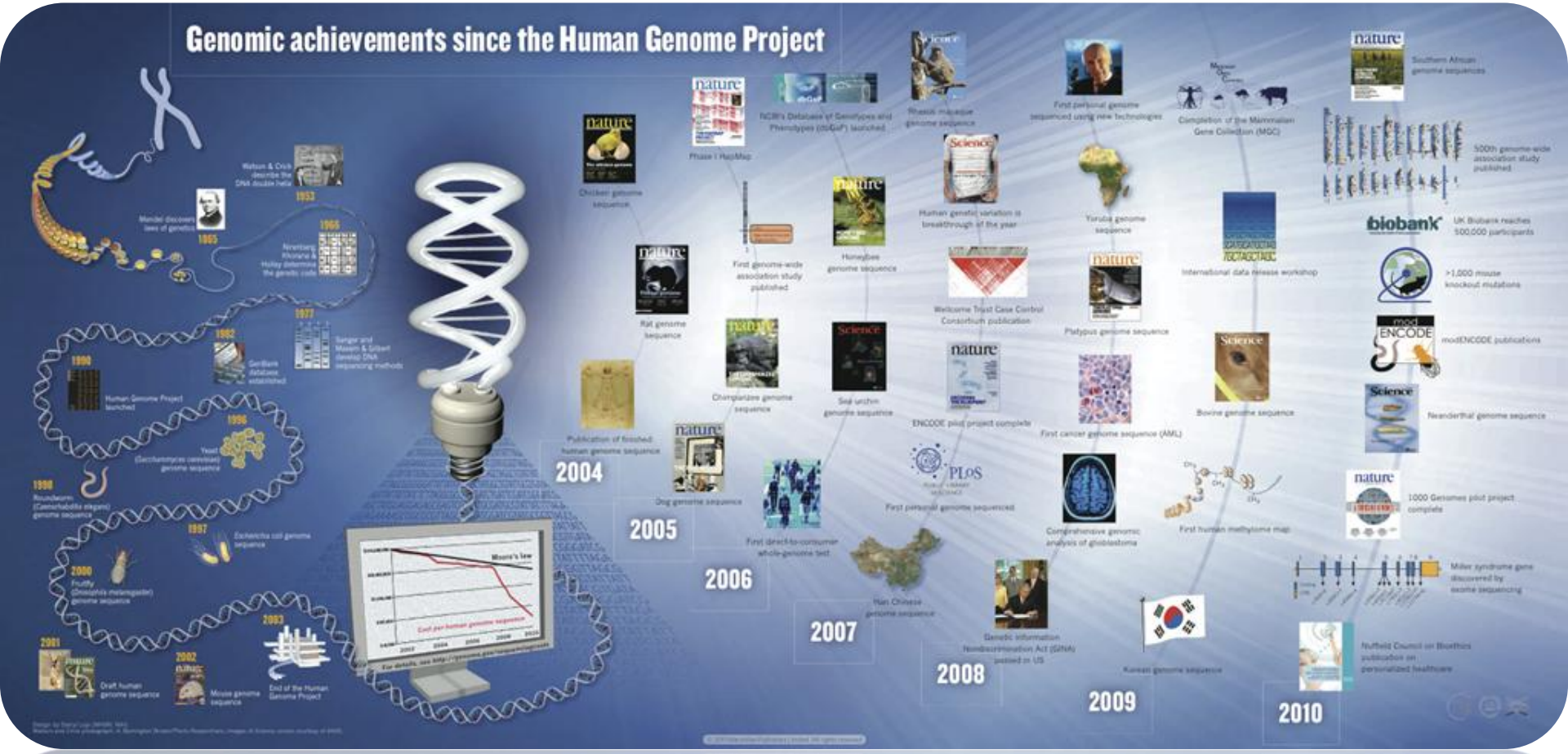
The triumphal march of genomics

The triumphal march of genomics: from the human genome to the 1000 human genomes



ED Green *et al. Nature* **470**, 204-213 (2011) doi:10.1038/nature09764

Genomic achievements since the Human Genome Project



ED Green *et al. Nature* **470**, 204-213 (2011) doi:10.1038/nature09764

nature

Genome science challenges

doi:10.1038/nature09764

The challenges of big data in Genomics

BOX 3

Bioinformatics and computational biology

The major bottleneck in genome sequencing is no longer data generation—the computational challenges around data analysis, display and integration are now rate limiting. New approaches and methods are required to meet these challenges.

Data analysis. Computational tools are quickly becoming inadequate for analysing the amount of genomic data that can now be generated, and this mismatch will worsen. Innovative approaches to analysis, involving close coupling with data production, are essential.

Data integration. Genomics projects increasingly produce disparate data types (for example, molecular, phenotypic, environmental and clinical), so computational approaches must not only keep pace with the volume of genomic data, but also their complexity. New integrative methods for analysis and for building predictive models are needed.

Visualization. In the past, visualizing genomic data involved indexing to the one-dimensional representation of a genome. New visualization tools will need to accommodate the multidimensional data from studies of molecular phenotypes in different cells and tissues, physiological states and developmental time. Such tools must also incorporate non-molecular data, such as phenotypes and environmental exposures. The new tools will need to accommodate the scale of the data to deliver information rapidly and efficiently.

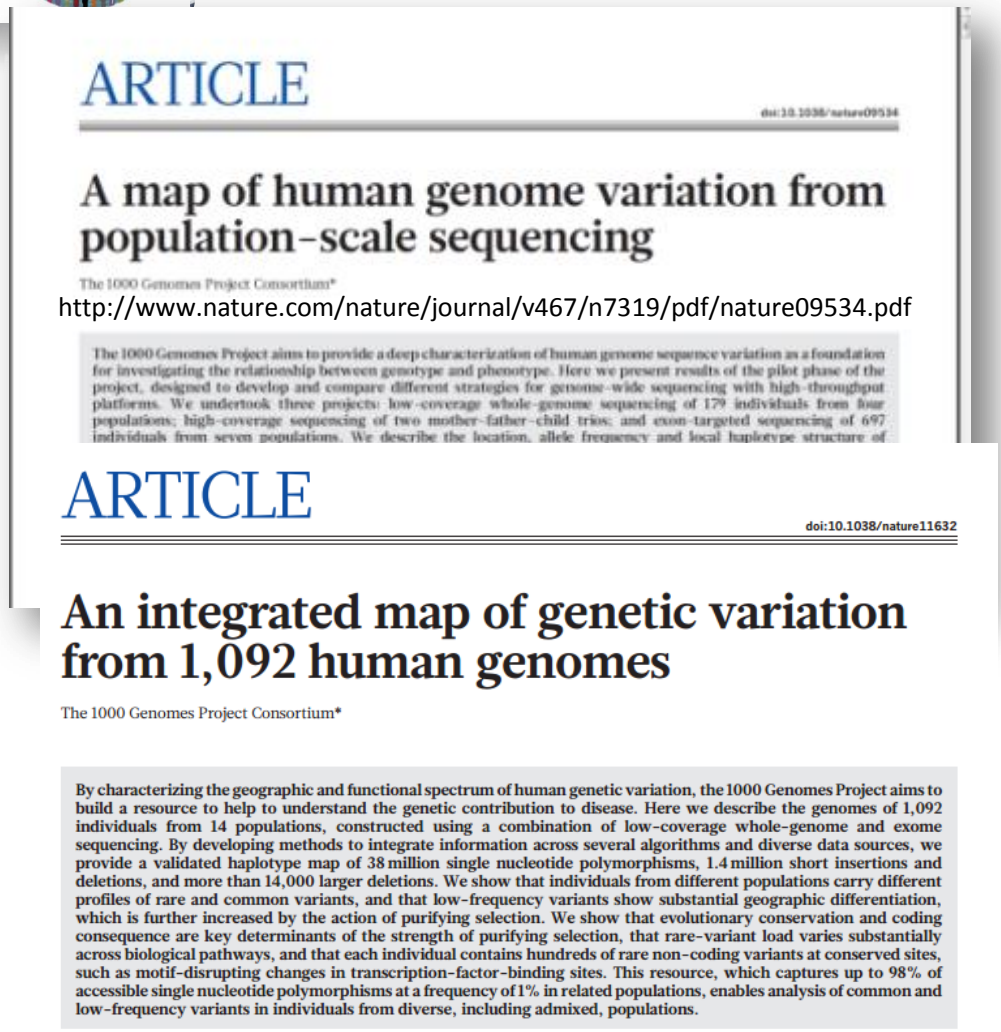
Computational tools and infrastructure. Generally applicable tools are needed in the form of robust, well-engineered software that meets the distinct needs of genomic and non-genomic scientists. Adequate computational infrastructure is also needed, including sufficient storage and processing capacity to accommodate and analyse large, complex data sets (including metadata) deposited in stable and accessible repositories, and to provide consolidated views of many data types, all within a framework that addresses privacy concerns. Ideally, multiple solutions should be developed[105].

Training. Meeting the computational challenges for genomics requires scientists with expertise in biology as well as in informatics, computer science, mathematics, statistics and/or engineering. A new generation of investigators who are proficient in two or more of these fields must be trained and supported.

The **major bottleneck** in genome sequencing is computational challenges around data analysis, display and integration

- **Data analysis**: Keep pace with the volume of genomic data
- **Data integration**: Keep pace with the complexity of genomic data.
- **Visualization**: New visualization tools will need to accommodate the multidimensional data from studies of molecular phenotypes in different cells and tissues, physiological states and developmental time.
- **Computational tools and infrastructure**:
  - Robust, well-engineered software that meets the distinct needs of genomic and non-genomic scientists.
  - Adequate computational infrastructure: sufficient storage and processing capacity to accommodate and analyse large, complex data sets deposited in stable and accessible repositories.
- **Training:** . A new generation of investigators proficient in two or more of fields of biology, informatics, computer science, mathematics, statistics and/or engineering must be trained and supported.

The 1000 Genomes Project



ARTICLE

doi:10.1038/nature09534

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

http://www.nature.com/nature/journal/v467/n7319/pdf/nature09534.pdf

The 1000 Genomes Project aims to provide a deep characterization of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype. Here we present results of the pilot phase of the project, designed to develop and compare different strategies for genome-wide sequencing with high-throughput platforms. We undertook three projects: low-coverage whole-genome sequencing of 179 individuals from four populations; high-coverage sequencing of two mother–father–child trios; and exon-targeted sequencing of 697 individuals from seven populations. We describe the location, allele frequency and local haplotype structure of

ARTICLE

doi:10.1038/nature11632

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. We show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, that rare-variant load varies substantially across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved sites, such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% in related populations, enables analysis of common and low-frequency variants in individuals from diverse, including admixed, populations.

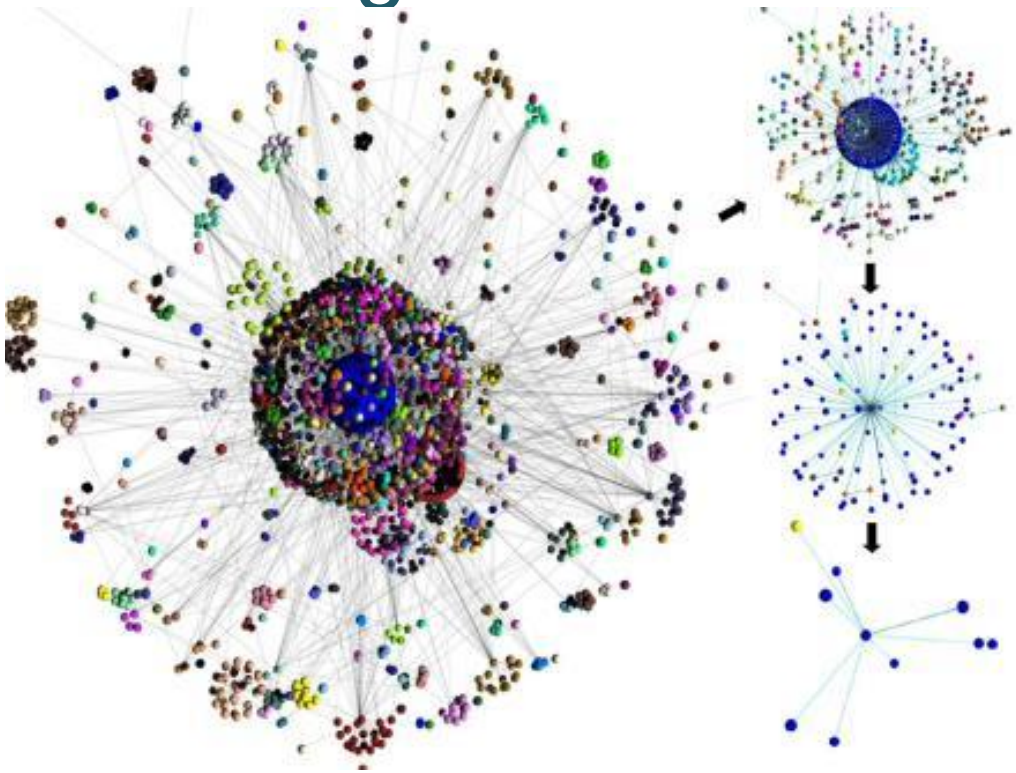
1000 genome project: The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65

1000 genomes Big Data problems

- Data storage (local vs cloud)
- Data computing
- Data delivery
- Security issues

The foreseeable future of genome science

¿Qué gran panorámica emergerá de la montaña de datos biológicos



?

a. La complejidad no es reducible
b. Nuevos principios generales de organización de lo biológico

NHGRI Plan: Charting a course for genomic medicine from base pairs to bedside

genome.gov
National Human Genome Research Institute
National Institutes of Health

SEARCH

Research Funding | Research at NHGRI | Health | Education | Issues in Genetics | Newsroom | Careers & Training | About | For You

Home › About › Long-Range Planning › Charting a course for genomic medicine from base pairs to bedside: The Strategic Plan

Long-Range Planning

- **Charting a course for genomic medicine from base pairs to bedside: The Strategic Plan**
- Event: A Decade with the Human Genome Sequence: Charting a Course for Genomic Medicine
- Past Long-Range Planning
- Topics: NHGRI 2008-2011 Planning Process
- White Papers: The 2008-2011 Planning Process
- Workshops: NHGRI 2008-2011 Planning Process

Share Print

The Strategic Plan

Charting a course for genomic medicine from base pairs to bedside

On February 10, 2011, *Nature* magazine published the National Human Genome Research Institute's (NHGRI) strategic plan for the future of human genome research called *Charting a course for genomic medicine from base pairs to bedside*. This strategic vision was developed in consultation with leading genome researchers over more than two years and is intended to inspire many to contribute to advancing genomic understanding, especially as other National Institutes of Health (NIH) institutes and centers focus genomic technologies on the diseases they study.

To celebrate the 10th anniversary of the first analysis of the draft human genome, and the launch of the new strategic vision for the field of genomics, NHGRI sponsored a symposium with leading thinkers in the field of genome research — including all three directors in the history of the National Human Genome Research Institute — who gathered on the campus of the National Institutes of Health on February 11, 2011, to consider the future of their field.

Follow the links below to the full strategic plan, symposium information, the planning process that developed the plan and related information.

- **Strategic Plan**: *Charting a course for genomic medicine from base pairs to bedside* PDF
  February 10, 2011
- **Symposium**: *A Decade with the Human Genome Sequence: Charting a Course for Genomic Medicine*
- **Strategic Plan Press Release**: NHGRI charts course for the next phase of genomics research
- NHGRI Long-Range Planning

To view the PDF document(s) on this page, you will need Adobe Reader. Adobe Reader

Top of page

Last Updated: March 23, 2012

NHGRI Director
Eric Green

http://www.genome.gov/Pages/About/Planning/2011NHGRIStrategicPlan.pdf

PERSPECTIVE

doi:10.1038/nature09764

Charting a course for genomic medicine from base pairs to bedside

Eric D. Green[1], Mark S. Guyer[1] & National Human Genome Research Institute*

There has been much progress in genomics in the ten years since a draft sequence of the human genome was published. Opportunities for understanding health and disease are now unprecedented, as advances in genomics are harnessed to obtain robust foundational knowledge about the structure and function of the human genome and about the genetic contributions to human health and disease. Here we articulate a 2011 vision for the future of genomics research and describe the path towards an era of genomic medicine.

Since the end of the Human Genome Project (HGP) in 2003 and the publication of a reference human genome sequence[1,2], genomics has become a mainstay of biomedical research. The scientific community's foresight in launching this ambitious project[3] is evident in the broad range of scientific advances that the HGP has enabled, as shown in Fig. 1 (see rollfold). Optimism about the potential contributions of genomics for improving human health has been fuelled by new insights about cancer[4–7], the molecular basis of inherited diseases (http://www.ncbi.nlm.nih.gov/omim and http://www.genome.gov/GWAStudies) and the role of structural variation in disease[8], some of which have already led to new therapies[9–11]. Other advances have already changed medical practice (for example, microarrays are now used for clinical detection of genomic imbalances[12] and pharmacogenomic testing is routinely performed before administration of certain medications[13]). Together, these achievements (see accompanying paper[14]) document that genomics is contributing to a better understanding of human biology and to improving human health.

As it did eight years ago[15], the National Human Genome Research Institute (NHGRI) has engaged the scientific community (http://www.genome.gov/Planning) to reflect on the key attributes of genomics (Box 1) and explore future directions and challenges for the field. These discussions have led to an updated vision that focuses on understanding human biology and the diagnosis, prevention and treatment of human disease, including consideration of the implications of those advances for society (but these discussions, intentionally, did not address the role of genomics in agriculture, energy and other areas). Like the HGP, achieving this vision is broader than what any single organization or country can achieve—realizing the full benefits of genomics will be a global effort.

quickly. Although genomics has already begun to improve diagnostics and treatments in a few circumstances, profound improvements in the effectiveness of healthcare cannot realistically be expected for many years (Fig. 2). Achieving such progress will depend not only on research, but also on new policies, practices and other developments. We have illustrated the kinds of achievements that can be anticipated with a few examples (Box 2) where a confluence of need and opportunities should lead to major accomplishments in genomic medicine in the coming decade. Similarly, we note three cross-cutting areas that are broadly relevant and fundamental across the entire spectrum of genomics and genomic medicine: bioinformatics and computational biology (Box 3), education and training (Box 4), and genomics and society (Box 5).

Understanding the biology of genomes

Substantial progress in understanding the structure of genomes has revealed much about the complexity of genome biology. Continued acquisition of basic knowledge about genome structure and function will be needed to illuminate further those complexities (Fig. 2). The contribution of genomics will include more comprehensive sets (catalogues) of data and new research tools, which will enhance the capabilities of all researchers to reveal fundamental principles of biology.

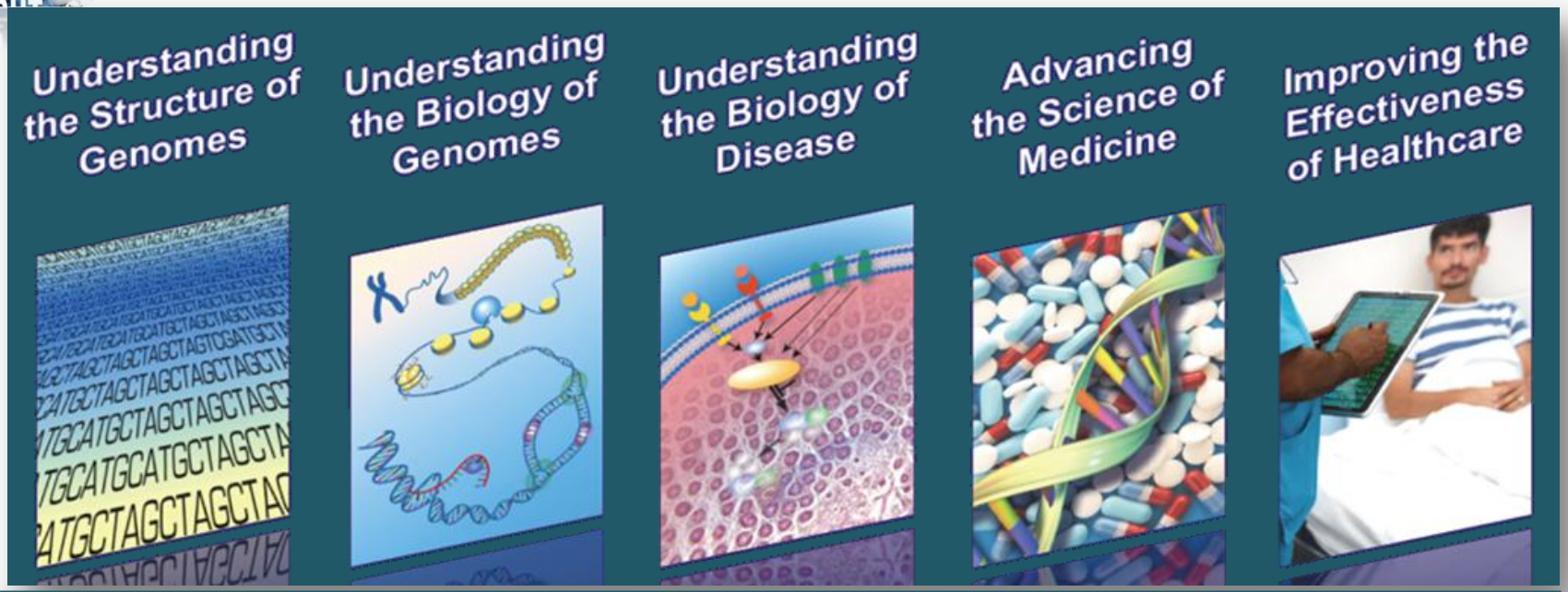
Comprehensive catalogues of genomic data

Comprehensive genomic catalogues have been uniquely valuable and widely used. There is a compelling need to improve existing catalogues and to generate new ones, such as complete collections of genetic variation, functional genomic elements, RNAs, proteins, and other biological

February 2011

NHGRI Published New Vision for Genomics

Five domains of genomics research

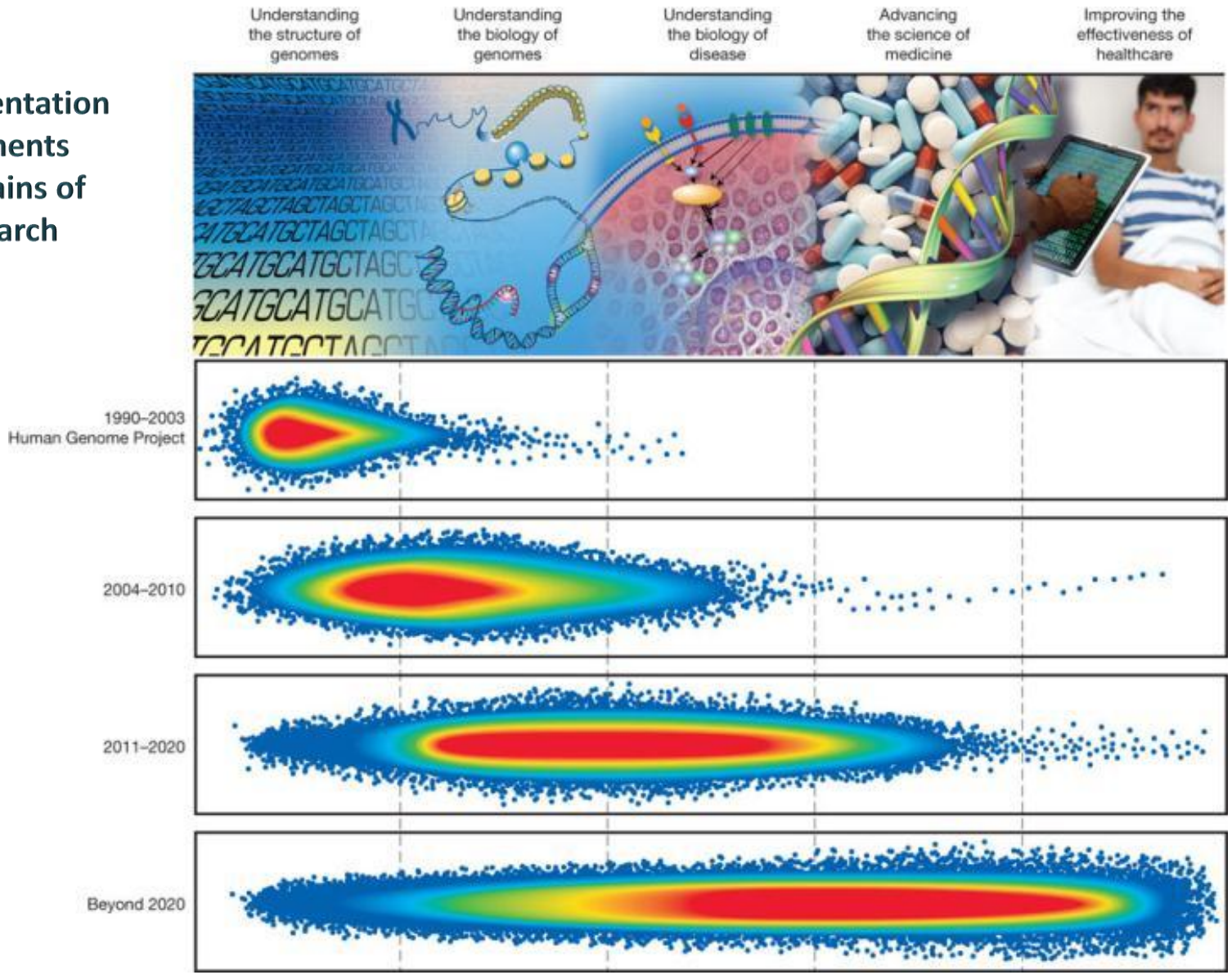


Basic Science

Translational Science

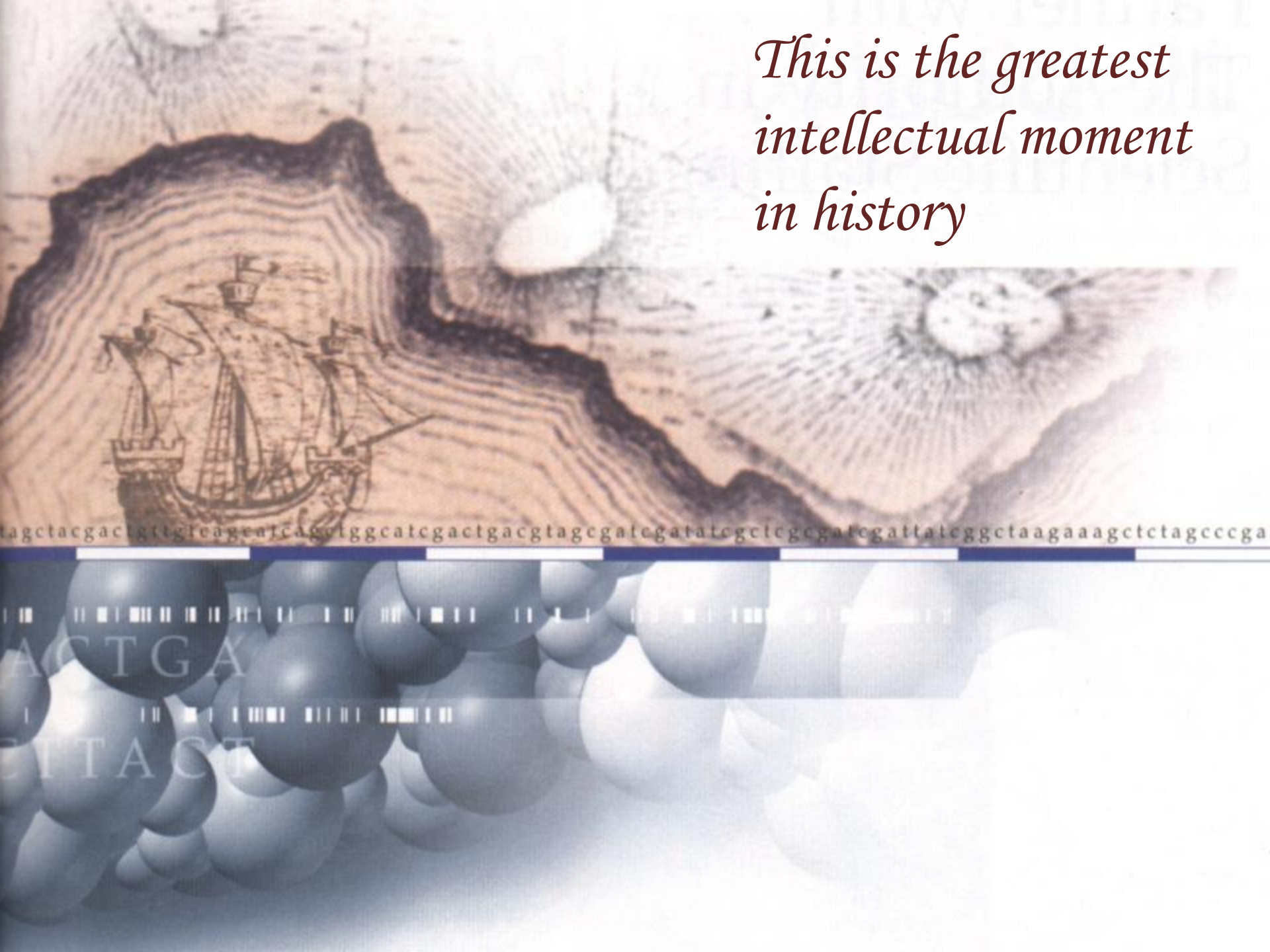
Implementation Science

Schematic representation of accomplishments across five domains of genomics research

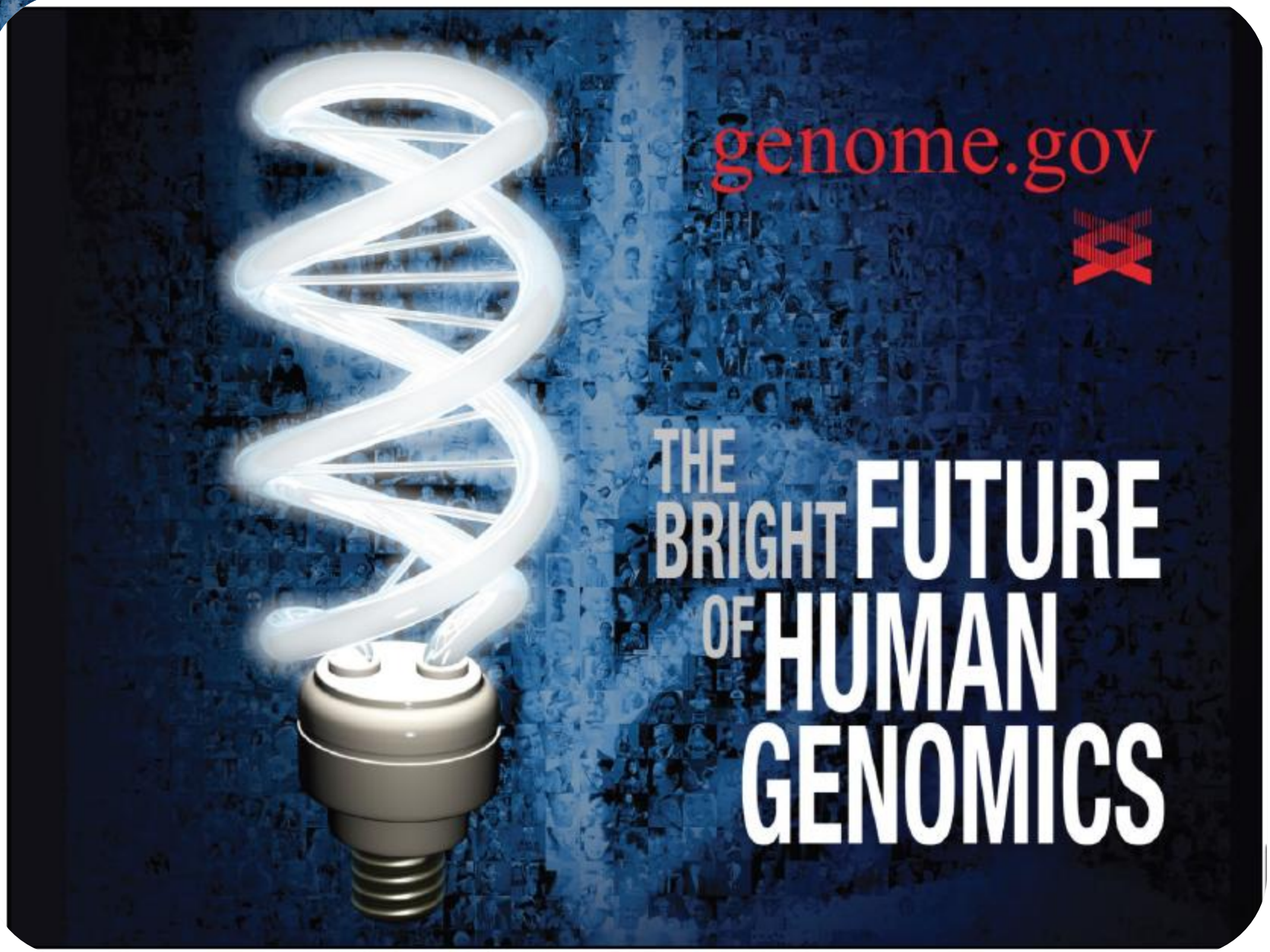


E D. Green *et al*. *Nature* **470**, 204-213 (2011) doi:10.1038/nature09764

nature


This is the greatest
intellectual moment
in history


THE
FUTURE


genome.gov
THE BRIGHT FUTURE OF HUMAN GENOMICS

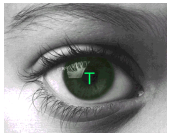
Readings & Videos

Readings

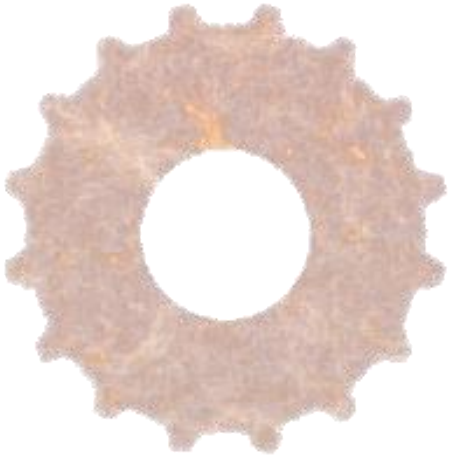
- Pespective: **Charting a course for genomic medicine from base pairs to bedside.** 2011. Eric D. Green, Mark S. Guyer & National Human Genome Research Institute. Nature 470, 204–213 (10 February 2011).
- Review: **Initial impact of the sequencing of the human genome**. Eric S. Lander. Nature 470, 187-197 (10 February 2011)
- Report economic study: **Economic Impact of the Human Genome Project.** 2011. May 2011. Battelle technology partnership practice.
- **The $1,000 genome, the $100,000 analysis?.** 2010. Elaine R Mardis. Genome Medicine 2010, 2:84

Videos

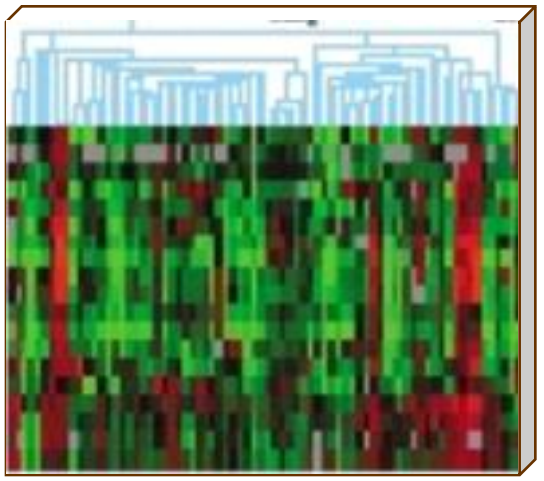
- **Nobel Week Dialogue 2012. The Genetic Revolution and its Impact on Society**
- **The Genomic Landscape *circa* 2012** *Eric Green, NHGRI*



Datos



Herramientas bioinformáticas



Conocimiento