

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

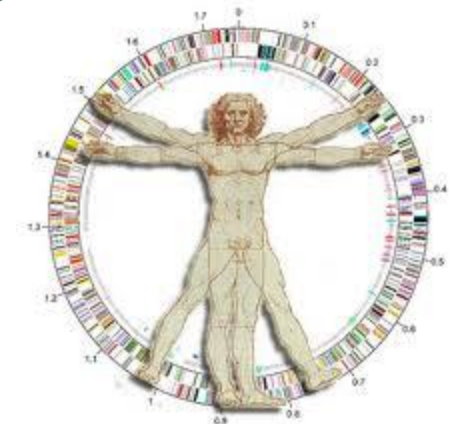
Genome science challenges

The foreseeable future of genome science

Readings



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



Genomics Landmarks

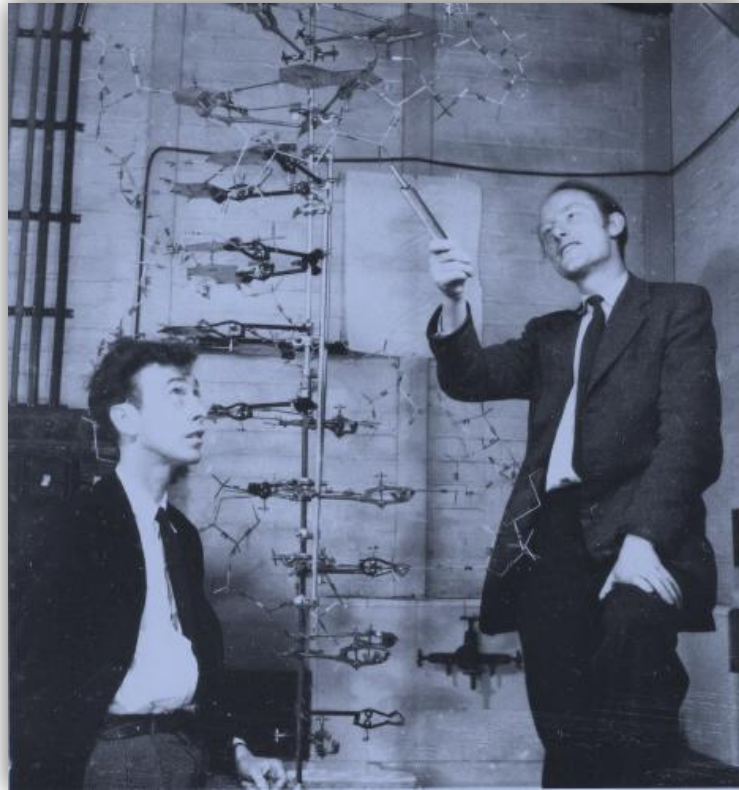
2001 and 2004

The publication of the draft and complete human sequence

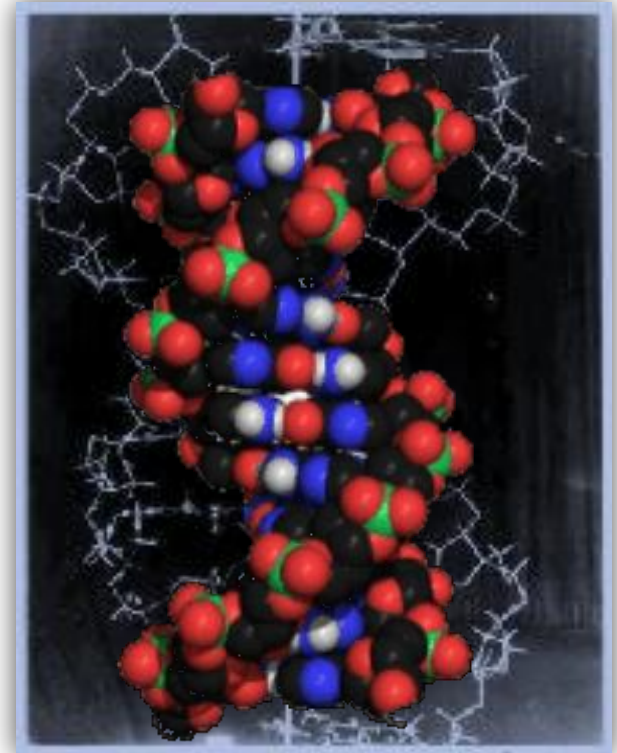
The emergence of

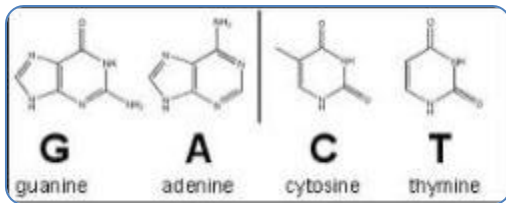
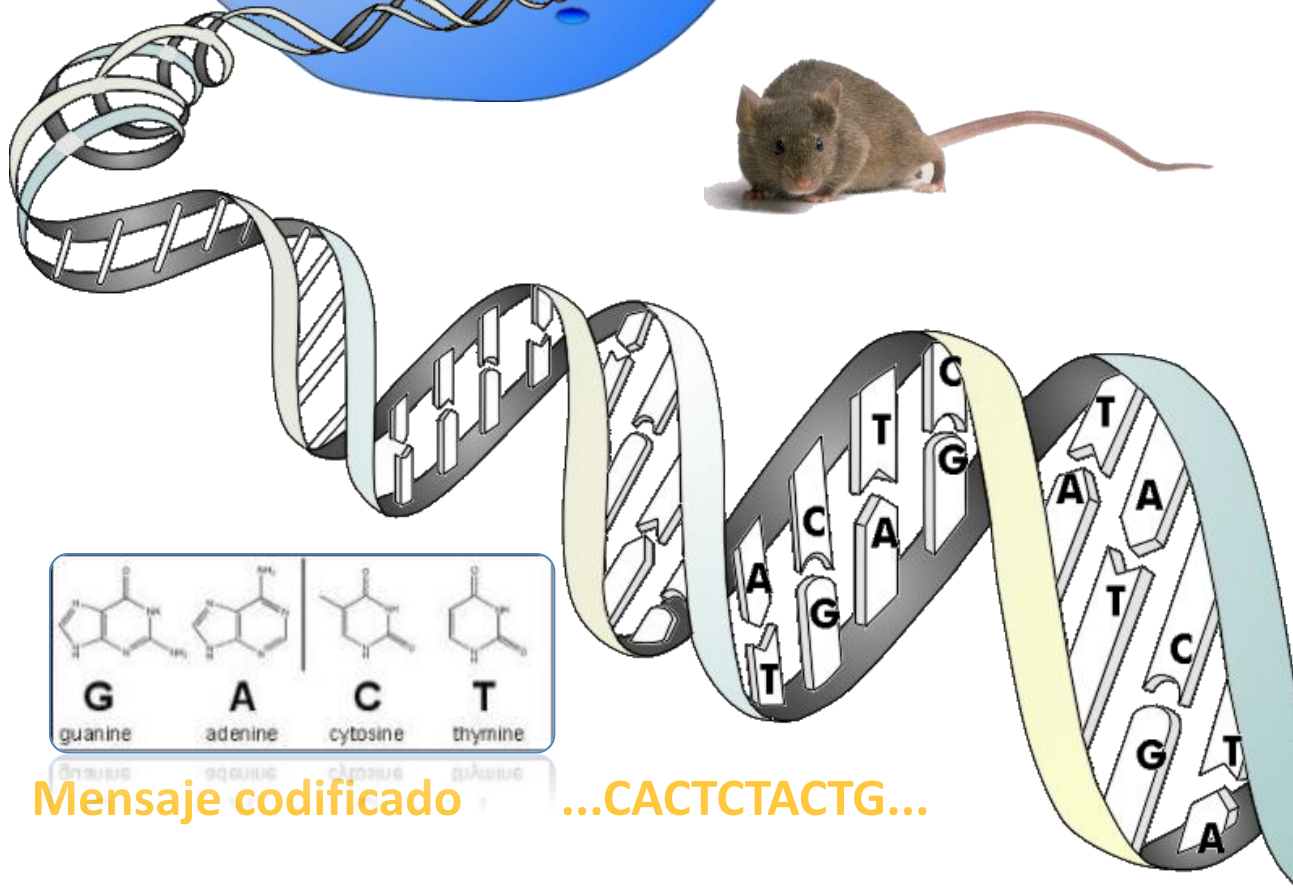
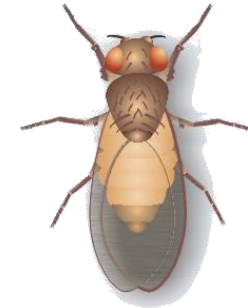
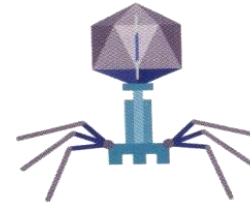
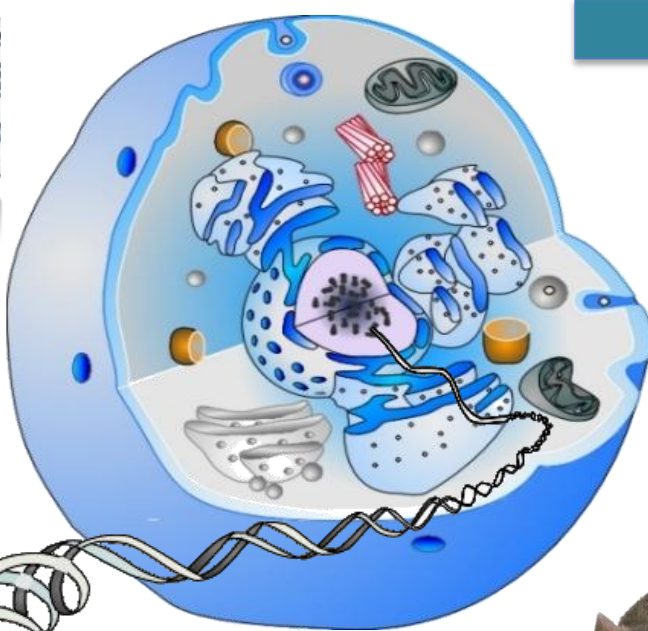


DNA Doble Helix: the secret of life



1953

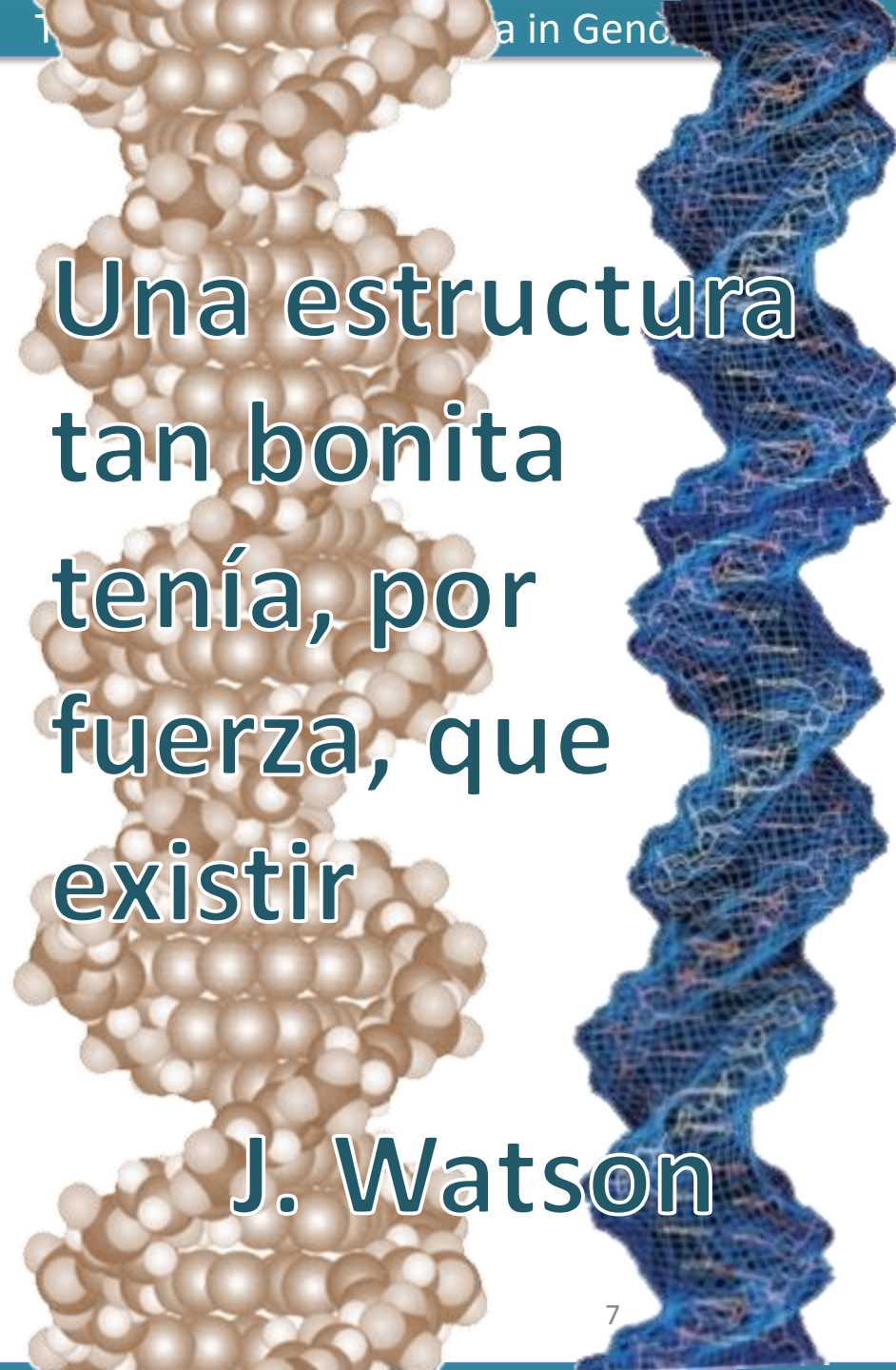
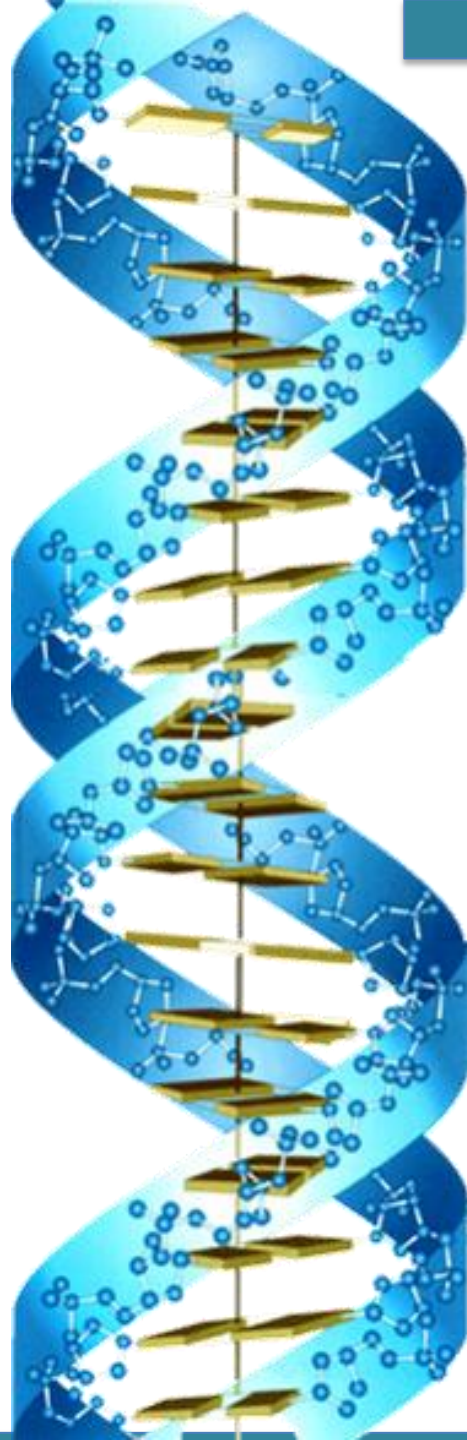
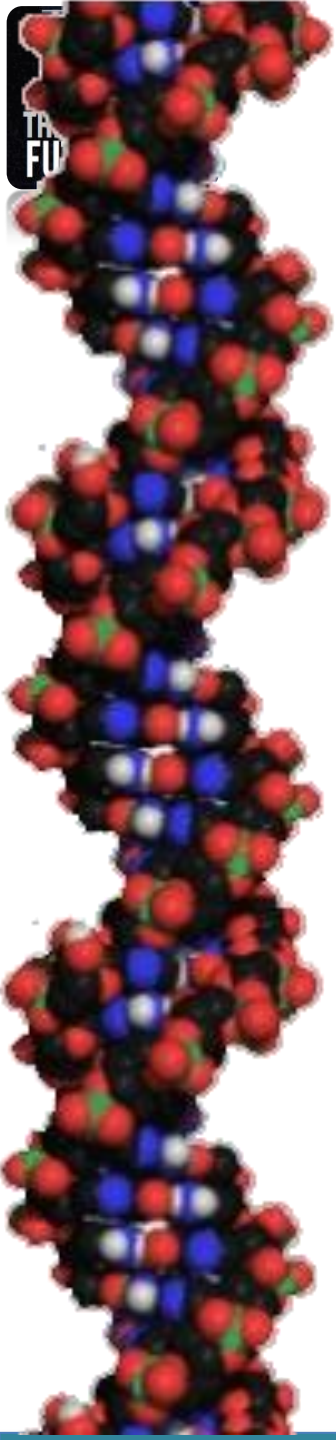




Mensaje codificado

...CACTCTACTG...





Una estructura
tan bonita
tenía, por
fuerza, que
existir

J. Watson

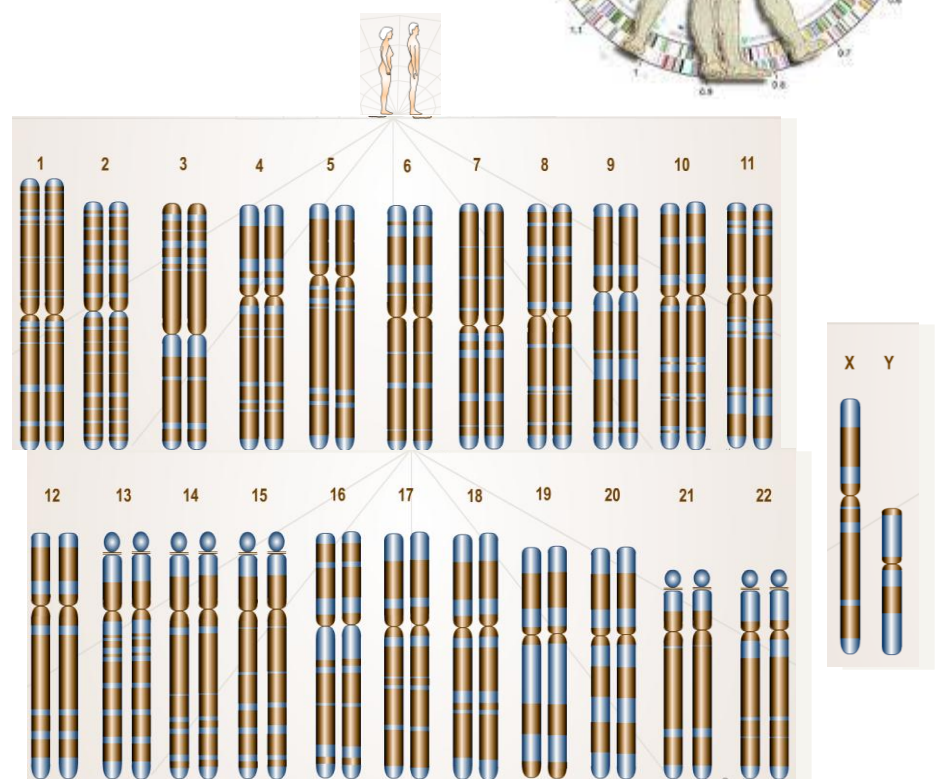
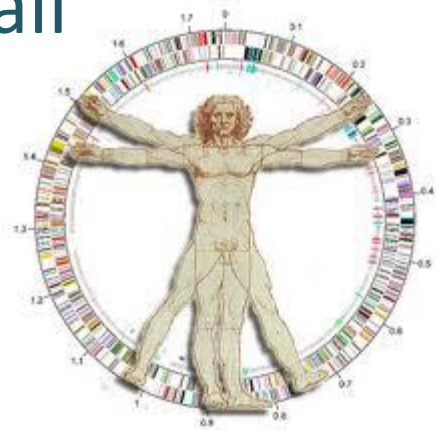
The human genome: The Holy Grail

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, 315000000
 /cromosoma "1-22, X, Y"
 /nota «Libro de la vida, santo grail, mapa humano»

INICIO

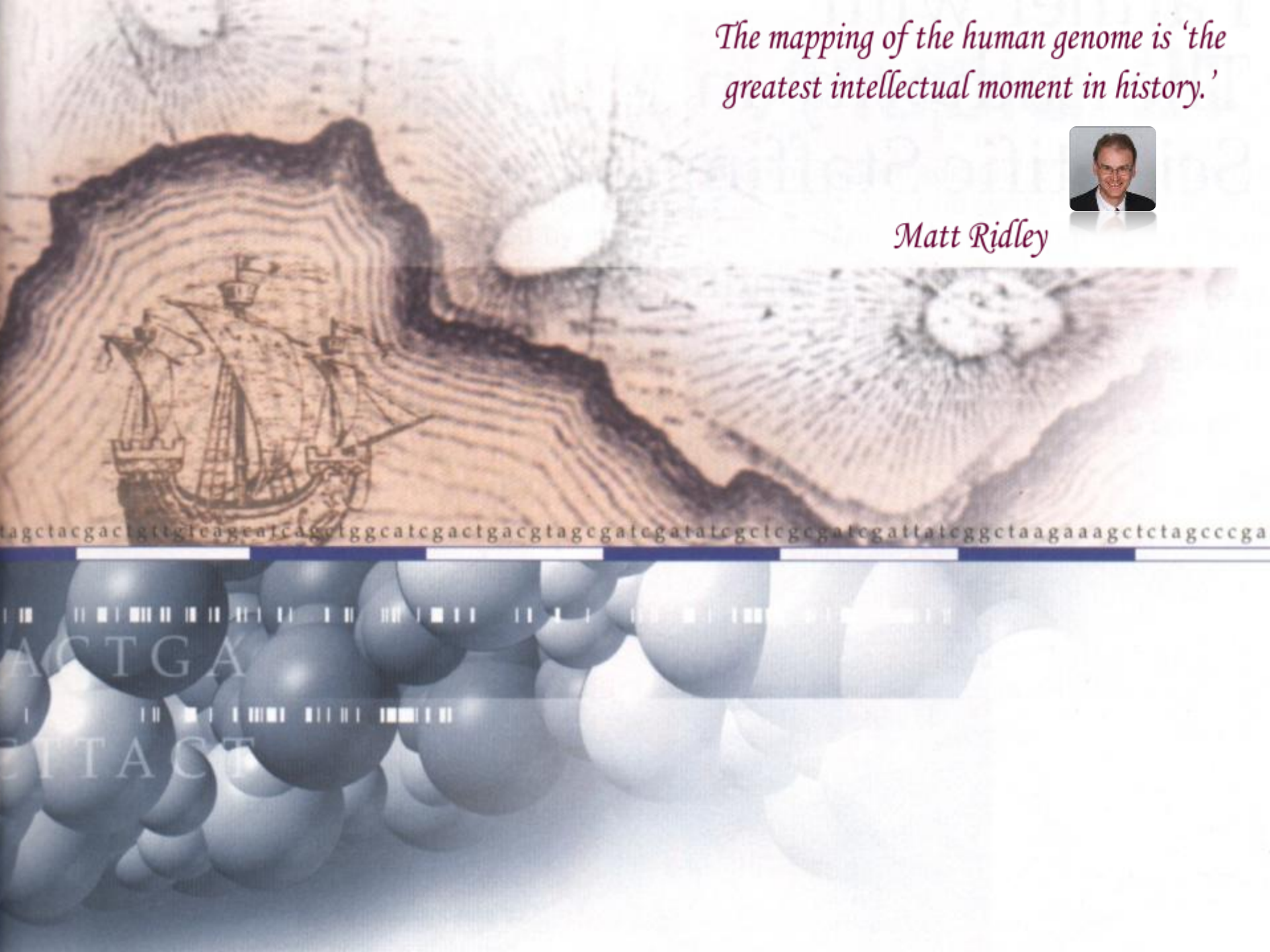
1	agctcgctga	gacttetgg	accccgacc	aggctgtggg	gtttctcaga	taactgggcc
61	cctgcgctca	ggagccttc	accctctgct	ctgggtaaag	ttcattggaa	cagaagaaga
121	tggatttacc	tgctcttcgc	gttgaagaag	tacaaaaatg	cattaatgct	atgcagaaaa
181	tcttagagtg	tccctctctg	ctggagtgta	tcaaggaacc	tgctctccaca	aagtgtgacc
241	acatattttg	caaattttgc	atgctgaaac	ttctcaacca	gaagaaaggg	ccttcacagt
301	gtcctttatg	taagaatgat	ataaccaaaa	ggaggcctaca	agaaagtacg	agatttagtc
361	aacttggtga	agagctattg	aaaatcattt	gtgcttttca	gcttgacaca	ggtttggagt
421	atgcaaacag	ctataatttt	gcaaaaaaag	aaaaaactc	tcctgaaact	ctaaaagatg
481	aagtttctat	catccaaagt	atgggctaca	gaaaccgtgc	caaaagactt	ctacagagt
541	aacccgaaaa	ctcttccttg	caggaacca	gtctcagtgt	ccaactctct	aaccttggaa
601	ctgtgagaac	tctgaggaca	aagcagcgga	tacaacctca	aaagacgtct	gtctacattg
661	aattgggagc	tgattcttct	gaagataccg	ttaataaggc	aacttattgc	agtggtggag
721	atcaagaatt	gttacaatc	accctcaag	gaaccagga	tgaaatcagt	ttggattctg
781	caaaaaagcg	tgcttgtgaa	ttttctgaga	cggatgtaac	aaatactgaa	catcatcaac
841	ccagtaataa	tgatttgaac	accactgaga	agcgtgcagc	tgagagccat	ccagaaaagt
901	atcagggtag	ttctgtttca	aactgcatg	tggagccatg	tggcacaat	actcatgcca
961	gctcattaca	ctcagagaac	agcagtttat	tactcaacta	agacagaatg	aatgtagaaa
1021	aggctgaatt	ctgtaataaa	agcaaacagc	ctggcttagc	aaggagccaa	cataacagat
1081	ggcctggaag	taaggaaca	tgtaatgata	ggcggactcc	cagcacagaa	aaaaaggtag
1141	atctgaatgc	tgatcccctg	tgtgagagaa	aagaatggaa	taagcagaaa	ctgcccagct
1201	cagagaatcc	tagagatact	gaagatgttc	cttgataac	actaaaatgc	agcattcaga
1261	aagttaatga	gtggttttcc	agaagtgatg	aactgttagg	ttctgatgac	tcacatgatg
1321	gggagctcga	atcaaatgcc	aaagtagctg	atgtattgga	cgttctaaat	gaggtagatg
1381	aatattctgg	ttcttcagag	aaaaatagat	tactggccag	tgatccctcat	gagccttaa
1441	tatgtaaaag	tgaaagagtt	cactccaat	cagtagagag	taatattgaa	gacaaaat
1501	ttgggaaaac	ctatcggag	aaggcaagcc	tcaccaact	aagccatgta	actgaaaatc
1561	taattatag	agcatttgtt	actgagccc	agataataca	agagcgtccc	ctcacaaa
1621	aattaaagcg	taaaaagcga	cctacatcag	ctctctacc	tgaggatttt	atcaagaag
1681	cagatttggc	agttcaaaa	actcctgaaa	tgataaatca	gggaactaac	caaacggagc
1741	agaatggtca	agtgatgaat	attactaata	gtggtcatga	gaataaaaca	aaaggtgatt
1801	ctattcagaa	tgagaaaaat	cctaacccea	tagaatcact	cgaaaaagaa	tctgcttca
1861	aaacgaaagc	tgaacctata	agcagcagta	taagcaatat	ggaactcgaa	ttaaataatc
1921	acaattcaaa	agcacctaaa	aagaatagc	tgaggaggaa	gtcttctacc	aggcatatc
1981	atgcgcttga	actagtagtc	agtagaaatc	taagcccacc	taattgtact	gaattgcaa
2041	ttgatagttg	ttctagcagt	gaagagataa	agaaaaaaa	gtacaccaa	atgccagtc

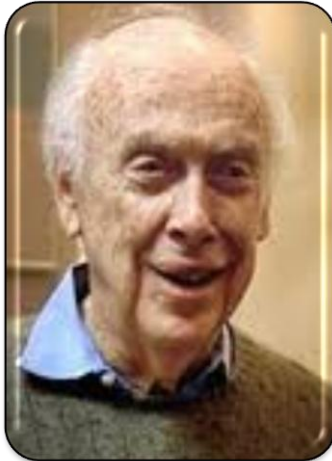


The mapping of the human genome is 'the greatest intellectual moment in history.'



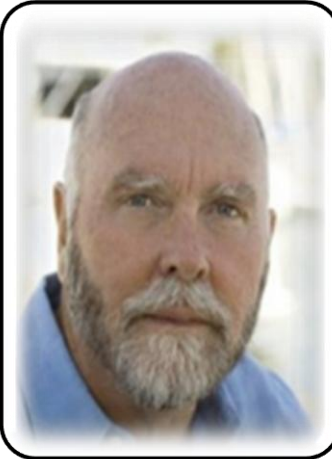
Matt Ridley





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

- ❑ 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)



International Human Genome Sequencing Consortium*

Genome Sequencing Centres (Listed in order of total genomic sequence contributed, with a partial list of personnel. A full list of contributors of each centre is available as Supplementary Information.)

Wellcome Trust Centre for Human Genetics, Oxford
 Michael C. Zody, Jennifer R. Kidd, Gill Bevan, Kim Devor, Michael Boyle, William Watkins, Bob Ferrel, Steve Daly, Kendra Harsh, Andrew Birdall, John Rowland, Lisa Kiani, Jessica Leachley, Susan LeVine, Paul McEvoy, Keith Morrison, James Mulrow, Jill Weston, Cher Woods, William Wright, Jerome Taylor, Christina Payne, Mark Ross, Ralph Samra, Andrew Stratton, Corie Szymanski, Nicola Stoneking, Wade Stephens, Arvid Sjöström & Sander Vermeulen

The Sanger Centre, Hinxton, UK
 Richard A. Harte, Stephen Beck, David Bentley, John Barker, Christopher Chen, Hugh Carter, Alan Coulson, Barbara Davidson, Frances Dickinson, Andrew Durrant, Ian Dunham, Richard Durbin, Lisa French, Guyan Gough, Steve Gregory, Tim Hubbard, Sean Humphrey, Adrienne Hunt, Matthew James, Christine Lloyd, Amanda McMurtry, Les McWhirter, Simon Munro, Sarah Murrell, James C. Mullikin, Andrew Murray, Robert Platts, Mark Ross, Reta Shewmaker & Sarah Sims

Washington University Genome Sequencing Center, St. Louis, MO
 Robert A. Hackett, Richard E. Wilson, LaShawn M. Hillier, John D. McPherson, Marco A. Maruy, Elaine S. Marder, Lorraine A. Pulley, Jeff T. Chikowski, Kyriakos N. Pappas, Warren R. Gish, Susumu I. Okazaki, Michael G. Wood, Kim E. Oelschlaeger, Tracy L. Moore, Andrew Osherson, Jason R. Kravet, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Hillier & Sandra W. Gilman

DSB Joint Genome Institute, Walnut Creek, CA
 Eileen E. Eichler, Paul Frock, Paul Richardson, Sarah Wessinger, Tam Shook, Norman Duggan, Jun-Pang Cheng, Anne Oliva, Susan Lucas, Christopher Ebbert, Edward Shtatfner & Marvin Pflieger

Baylor College of Medicine Human Genome Sequencing Center, Houston, TX
 Michael A. Elder, Corey M. Murrey, Steven C. Scherer, John B. Boockvar, Erika J. Subirana, Kim C. Worley, Catherine M. Hines, James R. Gerton, Michael L. Metzker, Susan L. Nayler, Raju S. Kuchenthan, David L. Nelson, & George M. Weinstock

NIH Human Genome Sequencing Center, Bethesda, MD
 Anne F. Beck, William Skarnes, Francisco Antequera, Philippe Broder, Thomas Brubaker, Eric Pfeiffer, Catherine Robert & Patrick Narelik

NYC Sequencing Center, New York, NY
 Lynn Zhuravskaya, Marc Stubbins, Keith Weinstock, Hong Mei Luo & John DeRubeis

Department of Genome Analysis, Institute of Molecular

Methodology: Andre Kosman, Matthias Ploner, Gerald Hoekstra, Doris Teichler & Andrea Ramp

Beijing Genomics Institute/Human Genome Center: Haoning Yang, Jun Tu, Jun Wang, Guying Huang & Jun Gu

Wellcome Trust Genome Sequencing Center, The Institute for Systems Biology, Leroy Hood: Leroy Hood, Anup Madan & Shihon Cho

Stanford Genome Technology Center: Ronald W. Davis, Nancy A. Fedorov, A. Flo Alcala & Michael J. Proctor

Stanford Human Genome Center: Richard M. Myers, Jeremy Schumacher, Mark Dickson, Jane Grimwood & David H. Cox

University of Washington Genome Center: Myron D. Olson, Rajinder Kaur & Christopher Raymond

Department of Molecular Biology, Yale University School of Medicine: Indira V. Sreekumar, Kazuhiko Kawasumi & Shinya Mizuno

University of Texas Southwestern Medical Center at Dallas: Greg A. Evans, Malik Akhavanfar & Roger Schultz

University of Oklahoma's Advanced Center for Genome Technology: Bruce A. Roe, Ting Chen & Joseph Fox

Max Planck Institute for Molecular Genetics: Jochen Fussler, Hans Lehrach & Richard M. Krauss

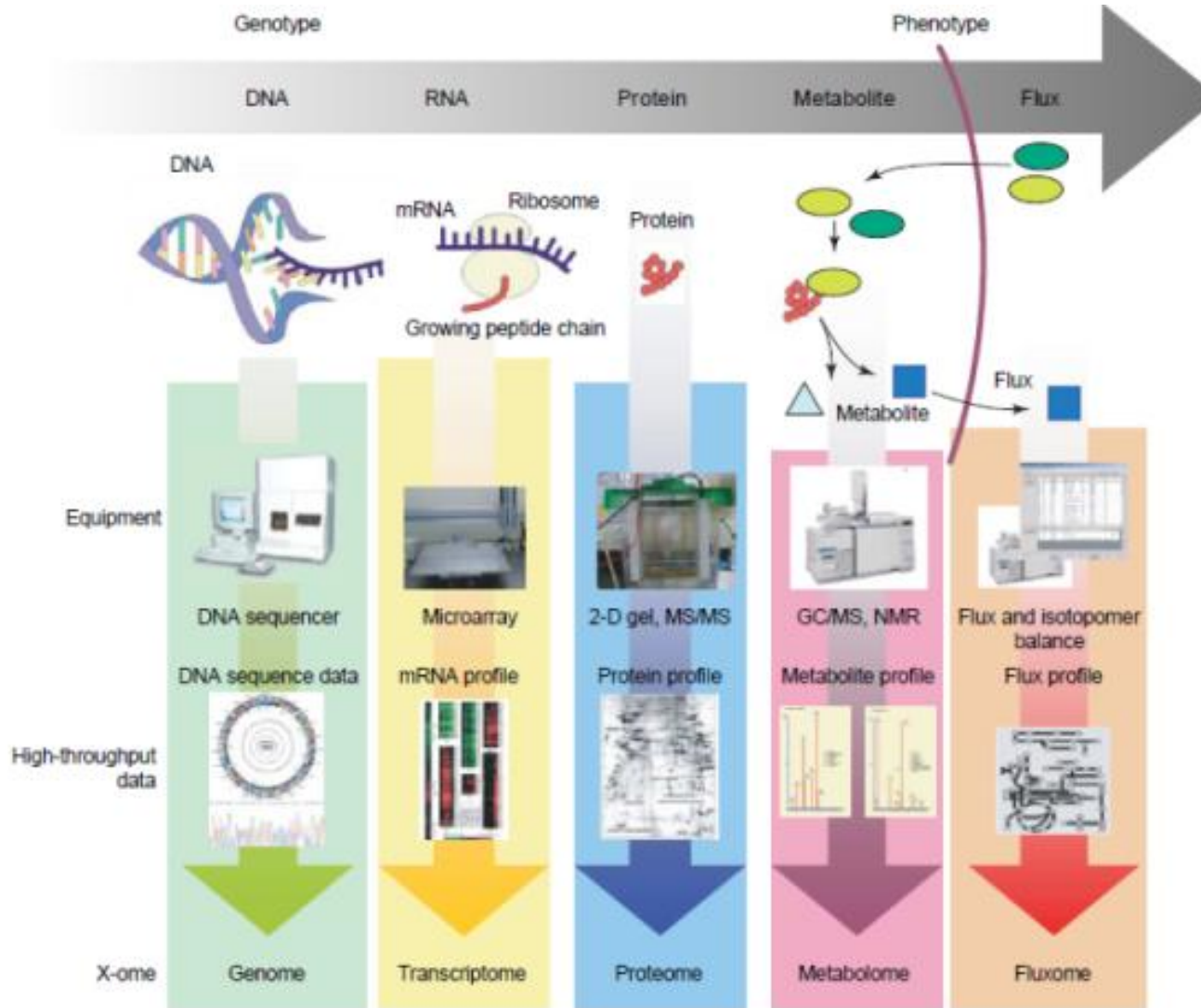
Cold Spring Harbor Laboratory, Uta Ansorge: Uta Ansorge, Center for Molecular Biology, Melissa de la Bastide & Hedy Shtut

EMBL—German Research Centre for Biotechnology: Helmut Moritz, Bruce Henrichsen & Gabriele Novak

Genome Analysis Group (listed in alphabetical order, also includes individuals listed under other headings): Rishi Agarwal, I. Anwar, Jeffrey S. Bailey, Rick Bakker, Sander Bateman, Owen Brown, Peter Bork, Daniel S. Brown, Christopher B. Burge, Lawrence David, Hui Chen Chen, Diana Church, Michele Chang, Richard R. Copley, Tobias Doering, Sean R. Eddy, Devin E. Dickler, Terence S. Flory, James Galagan, James G. H. Gibby, Cyrus Gharizadeh, Yoshihide Hayashizaki, David Hesselton, Jennifer Hertzberg, Karsten Hokeness, Huijun Jiang, Li Dong Johnson, Thomas A. Jones, Steven Kauf, Anil Kapany, Scott Kennedy, N. James Kent, Paul Kim, Eugene K. Kozlov, Jui-Kui, David Kulp, Doron Lancet, Todd M. Livan, Arsh Malik, Toshiyuki Mikabe, John V. Moran, Shosh Mulkar, Victor J. Palfrey, Chris P. Pandey, Greg Poulos, Jing Shih, Jay Shier, Aron P. A. Smith, Ella Stankovic, Joseph Staudacher, Danielle Therny May, John Therny May, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Paul J. Work, Kenneth H. Wolfe, Didiya-Ping Yang & Fu-Fang Yen

Scientific management: National Human Genome Research Institute, US National Institutes of Health: Francis Collins, Mark S. Guyer, Jane Peterson, Adam Peterson & Eric A. Werthevsky; Office of Science, US Department of Energy: Arvid Sjöström; The Wellcome Trust: Michael J. Morgan

Omic



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

Distinguishing characteristics of Genomics

- ❑ 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)



International Human Genome Sequencing Consortium*

Genome Sequencing Centres (listed in order of total genomic sequence contributed, with a partial list of personnel. A full list of contributors of each centre is available as Supplementary Information.)

Wellcome Institute for Biomedical Research, Center for Genome Research: Eric S. Lander¹, Lauren M. Linton¹, Bruce A. Boyer¹, Chad Nusbaum¹, Michael C. Zody¹, Jennifer A. Hillier¹, Karl Dunham¹, Kim DeWitt¹, Michael Boyle¹, William Watkins¹, Neil Hall¹, Steve Searl¹, Kendra Harshbarger¹, Andrew Richardson¹, John Newkome¹, Lisa Kiani¹, Jessica Leachley¹, Susan LeVine¹, Paul McEwen¹, Amir M. Mirzaei¹, James Mulrow¹, Jill F. Mesirov¹, Cher Mearns¹, William Morris¹, Jerome Naylor¹, Christina Pasovici¹, Mark Powell¹, Ralph Samra¹, Andrew Sridharan¹, Corrie Sougnez¹, Nicole Stange-Thomson¹, Wade Stephens¹, Arvind Subramanian¹ & Sudley Whyte¹

The Sanger Centre: Jane Rogers², John Sebatian², Richard Ammend², Stephen Beck², David Bentley², John Barker², Christopher Chiu², Nigel Carter², Alan Coulson², Rebecca Davidson², Frances Dickinson², Andrew Durrant², Ian Dunham², Richard Durbin², Lisa French², Clayton Gaulton², Simon Gregory², Tim Hubbard², Sean Humphrey², Adrienne Hunt², Matthew James², Christine Lloyd², Amanda McWhirter², Lucy Millward², Simon Murray², Sarah Murrell², James C. Mullin², Andrew Murray², Robert Platts², Mark Ross², Rana Sankaranarayanan² & Sarah Sims²

Washington University Genome Sequencing Center: Robert A. Waterston³, Richard A. Wilson³, LaShana H. Miller³, John D. McPherson³, Marcos A. Maruy³, Elaine S. Marder³, Loretta A. Pulley³, Jeff T. Chinwalla³, Kyuminho H. Park³, Warren R. Gish³, Souhaila L. Chikara³, Michael G. Wood³, Kim E. Oelschlaeger³, Tracy L. Miner³, Andrew Oelschlaeger³, Jason B. Kravetz³, Lisa L. Cook³, Robert S. Fulton³, Douglas L. Johnson³, Patrick J. Blair³ & Sandra W. Gilman³

DSB Joint Genome Institute: Trevor Hastie⁴, Ehab Issacovich⁴, Paul Frenk⁴, Paul Richardson⁴, Derek Winters⁴, Tam Shook⁴, Norman Eggert⁴, Jun-Pang Cheng⁴, Jesse Oliver⁴, Susan Lucas⁴, Christopher Elser⁴, Edward Blattner⁴ & Marvin Pflieger⁴

Baylor College of Medicine Human Genome Sequencing Center: Michael A. Gillet⁵, Corey M. Murphy⁵, Steven C. Scherer⁵, John R. Bowick⁵, Erika J. Goldberg⁵, Gao C. Wu⁵, Catherine M. Hines⁵, James R. Gerton⁵, Michael L. Metzker⁵, Susan L. Naylor⁵, Raju S. Kacharathala⁵, David L. Nelson⁵ & George M. Weinstock⁵

NIH Human Genome Sequencing Center: Yoshiyuki Sakaki⁶, Asoo Fujimasa⁶, Naohiko Katayama⁶, Tetsuaki Fudo⁶, Masahito Teramitsu⁶, Takahiko Bito⁶, Chiharu Kaneko⁶, Mitsuo Watanabe⁶, Yasuaki Takai⁶ & Tadayuki Toyoda⁶

Genome and CNV (GNC) Center: Joon Hong Kim⁷, Seung-Ho Park⁷, William Skaruppa⁷, Francisco Antequera⁷, Philippe Broder⁷, Thomas Brubaker⁷, Eva Pellegrini⁷, Catherine Robert⁷ & Patrick Winick⁷

NYC Sequencing Center: Douglas R. Smith⁸, Lynn Zhuravskiy-Stamen⁸, Marc Stubbins⁸, Keith Weinstock⁸, Hong Mei Luo⁸ & Jochen Drenth⁸

Department of Genome Analysis, Institute of Molecular Technology: Andrei Kosarev⁹, Matthias Pflüger⁹, Gerald Hoekstra⁹, Stefan Trautwein⁹ & Andreas Rump⁹

Beijing Genomics Institute/Human Genome Center: Haoning Yang¹⁰, Jun Tu¹⁰, Jun Wang¹⁰, Guyang Huang¹⁰ & Jun Gu¹⁰

Wellington Genome Sequencing Center, The Institute for Systems Biology: Leroy Hood¹¹, Leo Hewitt¹¹, Anup Madan¹¹ & Shown Chen¹¹

Stanford Genome Technology Center: Ronald W. Davis¹², Nancy A. Fedorov¹², A. Flo Alcala¹² & Michael J. Proctor¹²

Stanford Human Genome Center: Richard H. Myers¹³, Jeremy Schumacher¹³, Mark Dickson¹³, Jose Gutwood¹³ & David H. Eick¹³

University of Washington Genome Center: Myron D. Dvoretzky¹⁴, Rajinder Kaur¹⁴ & Christopher Raymond¹⁴

Department of Molecular Biology, Yale University School of Medicine: Mohyeddin Shmida¹⁵, Kazuhiko Kawano¹⁵ & Shinsuke Mizushima¹⁵

University of Texas Southwestern Medical Center of Dallas: Gao A. Evans¹⁶, Maria Athanasiou¹⁶ & Roger Schultz¹⁶

University of Oklahoma's Advanced Center for Genome Technology: Bruce A. Roe¹⁷, Hong Chen¹⁷ & Joseph Fox¹⁷

Max Planck Institute for Molecular Genetics: Juliane Rauscher¹⁸, Hans Lehrach¹⁸ & Richard Reinhardt¹⁸

Gold Spring Harbor Laboratory, Uta Ansorge Human Genome Center: W. Richard McCombs¹⁹, Melissa de la Bastide¹⁹ & Heiko Imhof¹⁹

EMBL—German Research Centre for Biotechnology: Helmut Moritz²⁰, Bruce Henningher²⁰ & Gabriele Nordberg²⁰

Genome Analysis Group (listed in alphabetical order, also includes individuals listed under other headings): Rishi Agarwal²¹, I. Anwar²¹, Jeffrey S. Bailey²¹, Alok Barman²¹, Sankar Chatterjee²¹, Owen Chen²¹, Peter Cook²¹, Daniel C. Drews²¹, Christopher B. Rump²¹, Lawrence Smith²¹, Hui-Chen Chen²¹, Dennis Chubb²¹, Michele Ciampi²¹, Richard R. Copley²¹, Tobias Doering²¹, Sean R. Eddy²¹, Evan E. Eichler²¹, Terrence S. Ferry²¹, James Galagan²¹, James G. H. Hillier²¹, Cyrus Hootan²¹, Fumihiko Iwatazaki²¹, David Keane²¹, Hanming Nengping²¹, Karsten Harkany²¹, Huijie Jiang²¹, Li-Dong Johnson²¹, Thomas A. Jones²¹, Steven Kang²¹, Anik Kanyavsky²¹, Sui-Kwan Lee²¹, James Kim²¹, Paul Kim²¹, Eugene K. Koozekan²¹, Jui-Kui²¹, David Kuy²¹, David Lacerot²¹, Todd M. Livan²¹, Anitha Malyan²¹, Tejraj Mankabhan²¹, John V. Moran²¹, Shouk Mulkar²¹, Victor J. Palfrey²¹, Chris P. Pandey²¹, Greg Schatz²¹, Jing Shih²¹, Jay Sitter²¹, Aron F. A. Sim²¹, Ella Singsper²¹, Joseph Sussangkarn²¹, Danielle Therny May²¹, John Therny May²¹, Lukas Wagner²¹, John Wallis²¹, Raymond Wheeler²¹, Alan Williams²¹, Farid A. Wolf²¹, Kenneth H. Wolfe²¹, Didiya-Ping Yang²¹ & Fu-Fang Yin²¹

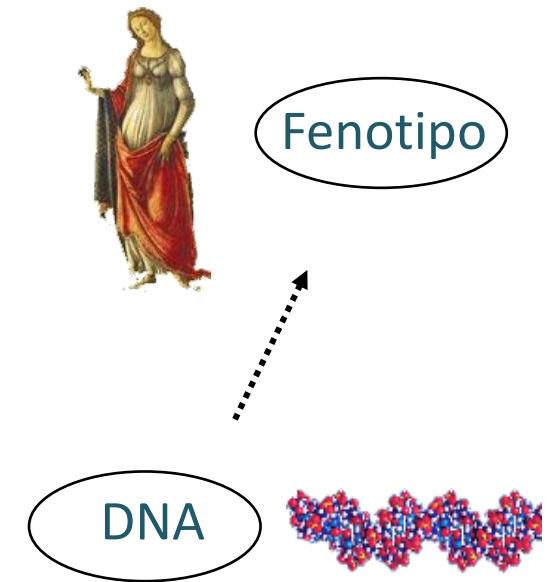
Scientific management: National Human Genome Research Institute, US National Institutes of Health: Francis Collins²², Mark S. Guyer²², Jane Peterson²², Adam Peterson²² & Eric A. Workalemichael²²; Office of Science, US Department of Energy: Arlinda Patricia²³; The Wellcome Trust: Michael J. Morgan²³

The Fundamental Question



Clarividencia

René Magritte




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


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

New paradigm

Biology as a
Informational Science

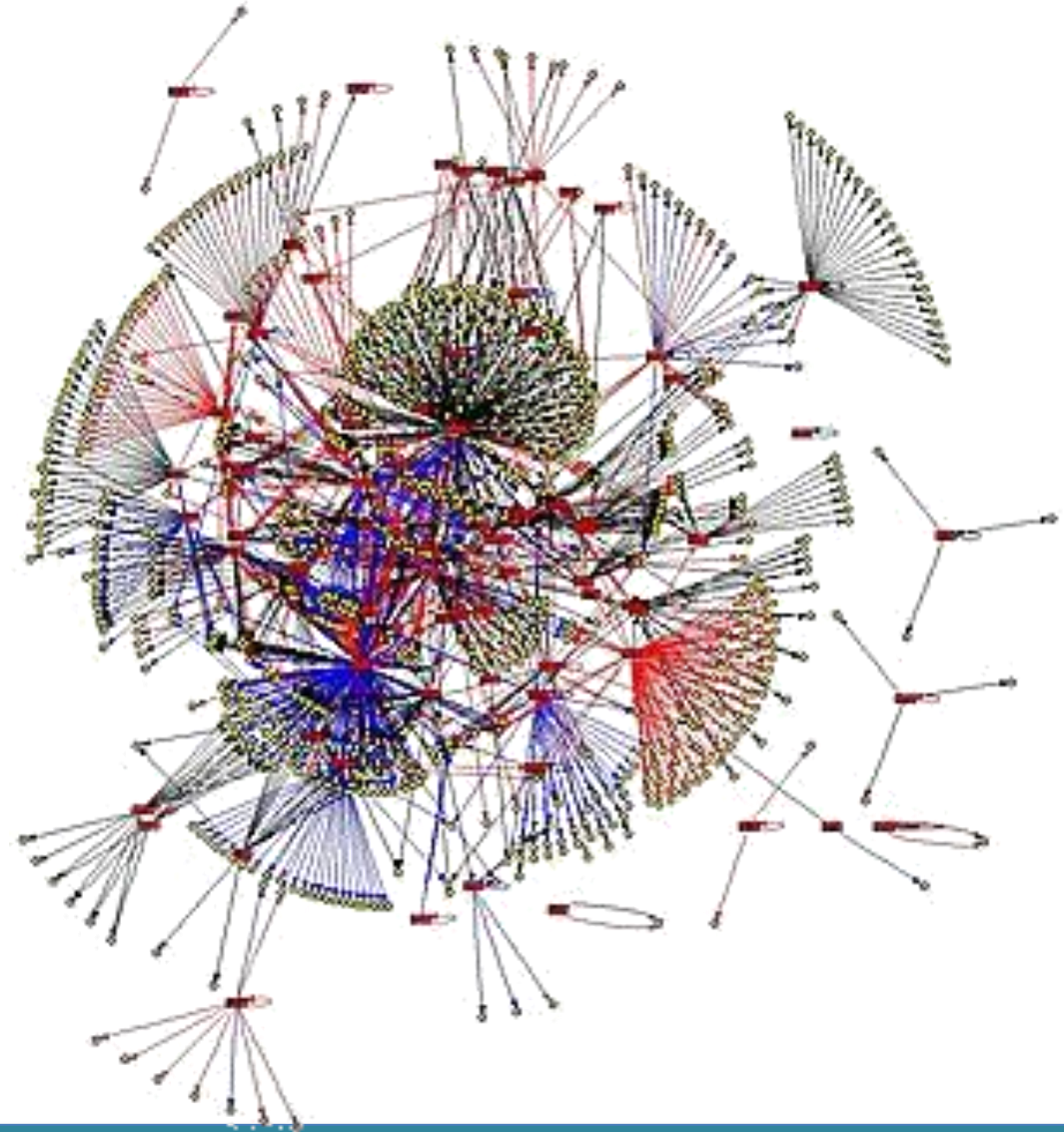


The HGP has
changed the
way we
conceptualize
molecular
biology

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

Integrative Biology

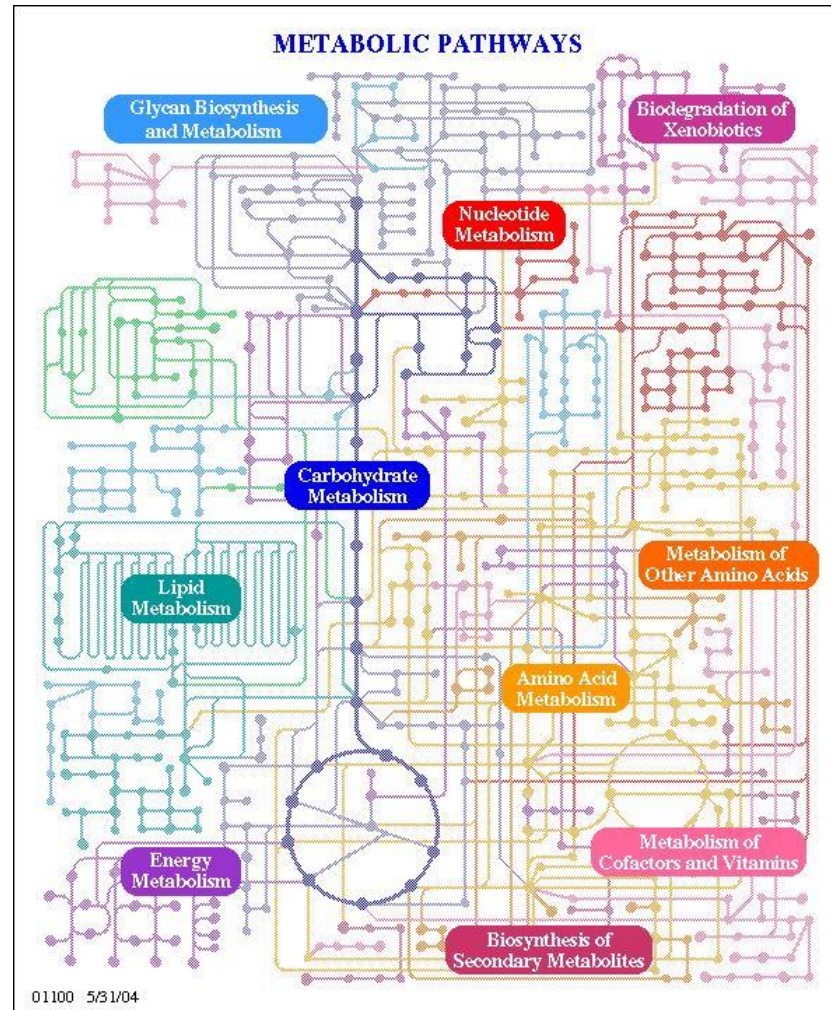
From the new paradigm, biological systems are complex networks of myriad of pathways, many of them interconnected



Integrative Biology



Biosynthesis pathways,

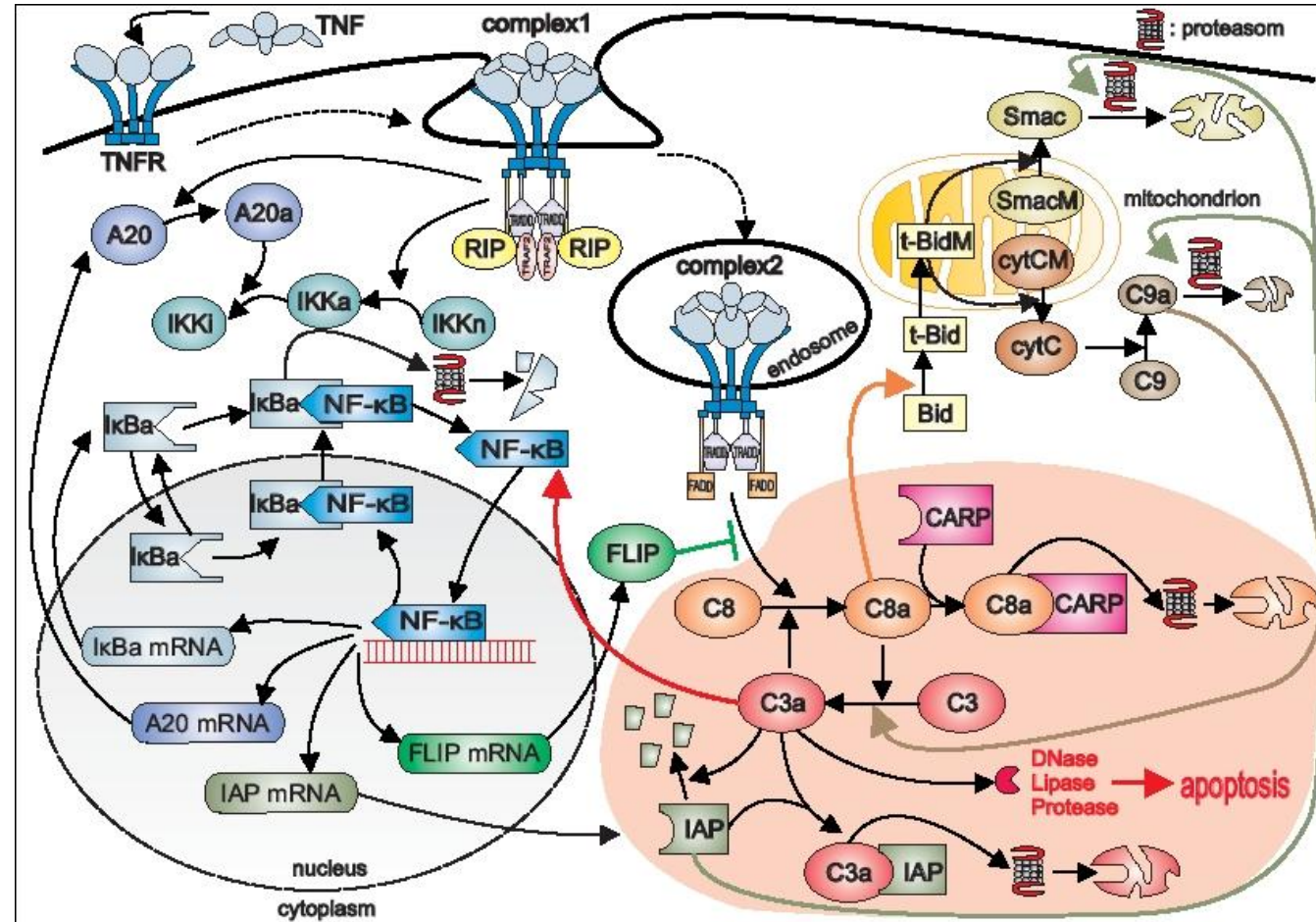


Metabolic pathway

Integrative Biology



Signal transduction pathways,

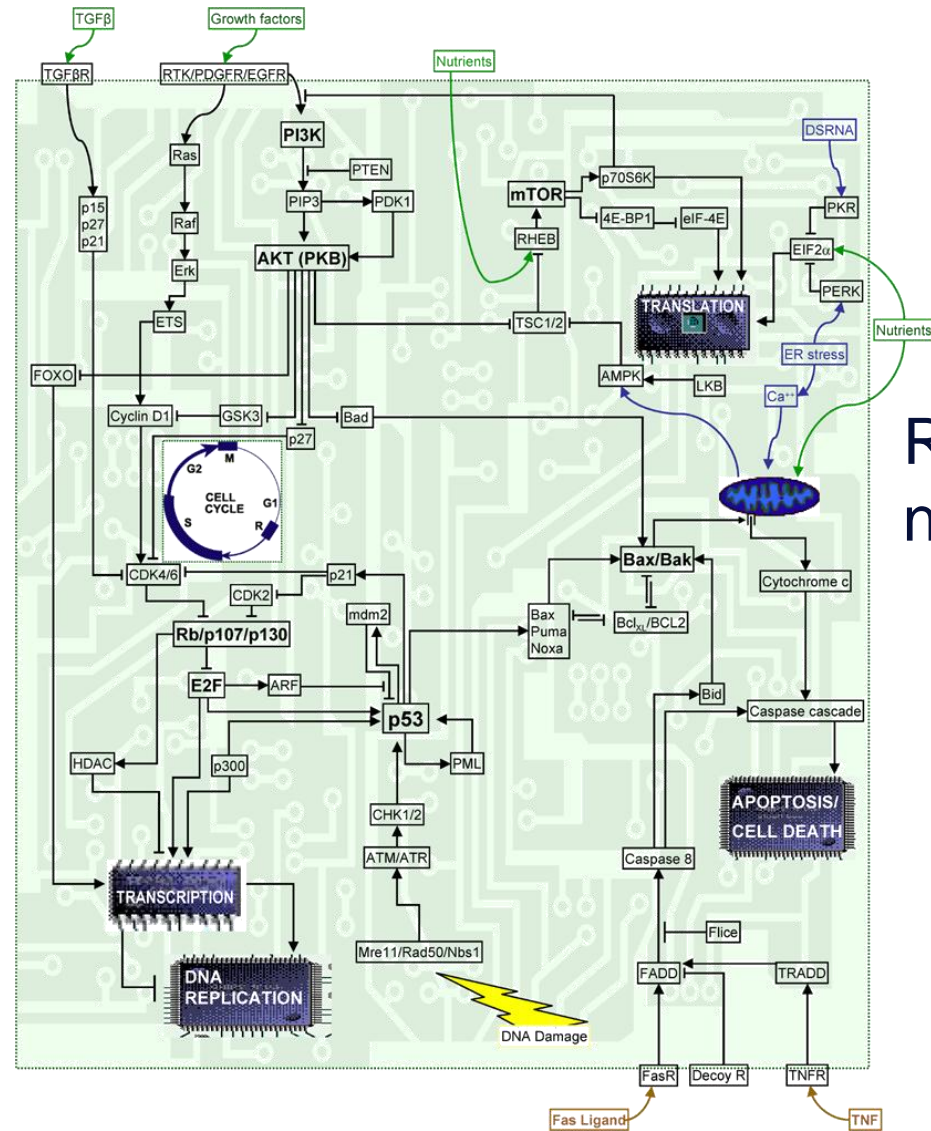


Signal transduction pathways

Integrative Biology



Regulatory networks.

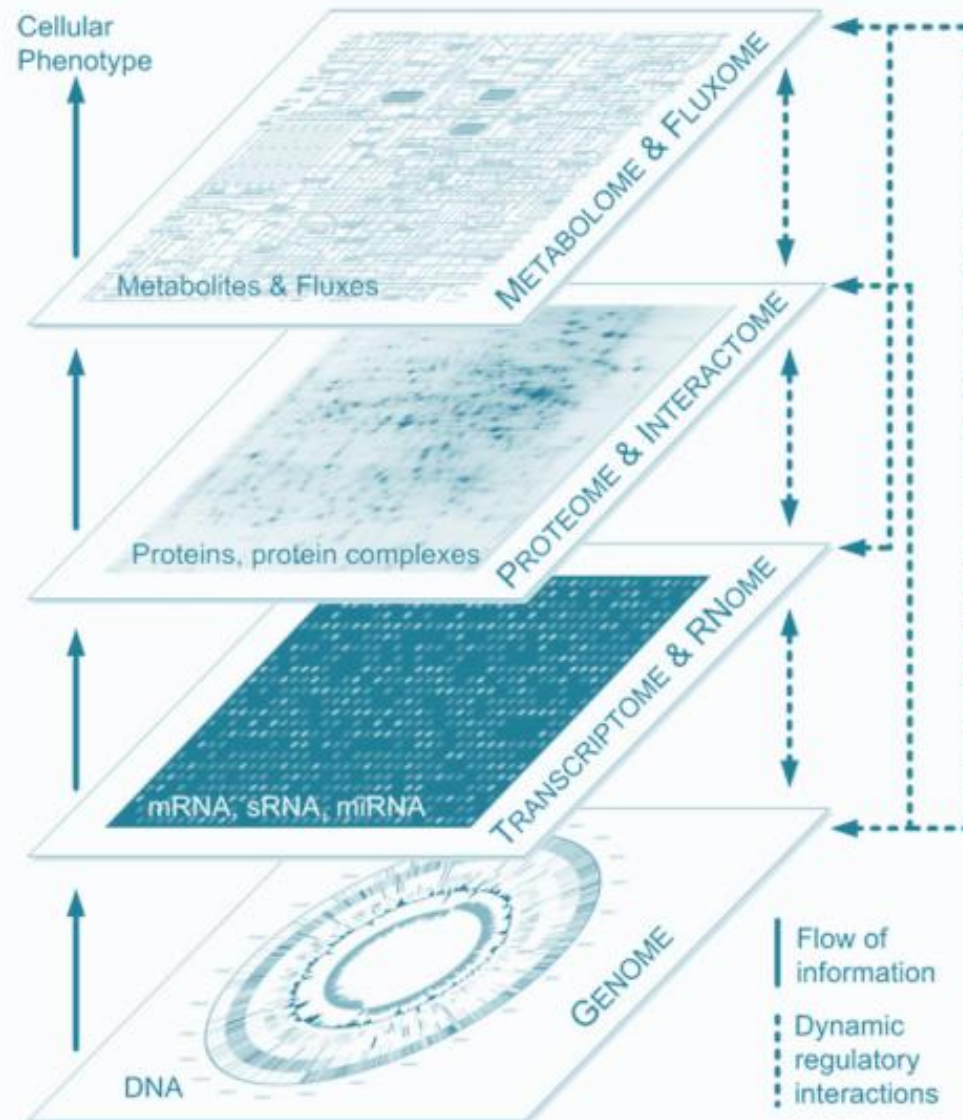


Regulatory networks

Kohlsted et al. 2010

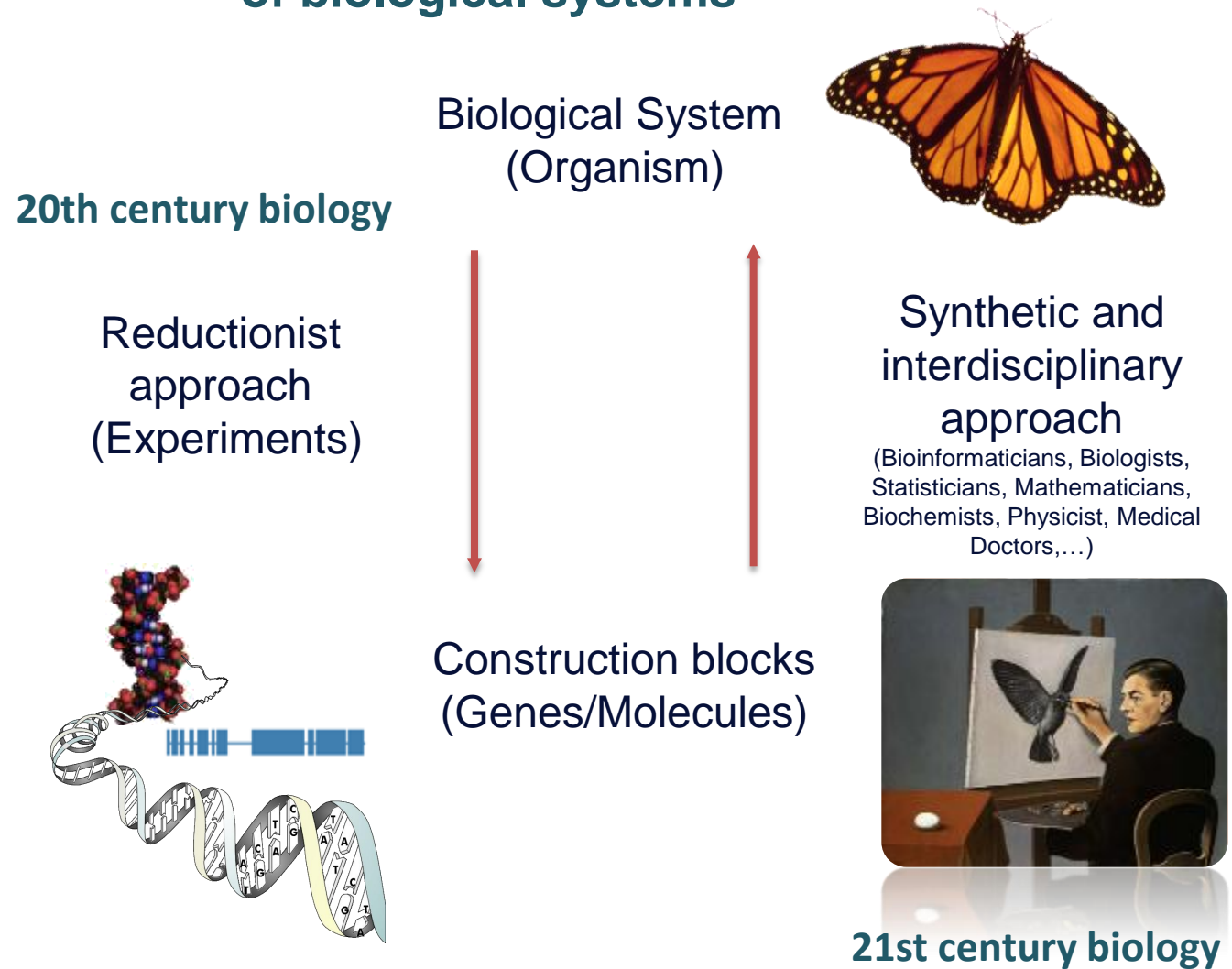
I
n
t
e
g
r
a
t
i
v
e

B
i
o
l
o
g
y



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





Genome sequencing

Why sequencing a genome?

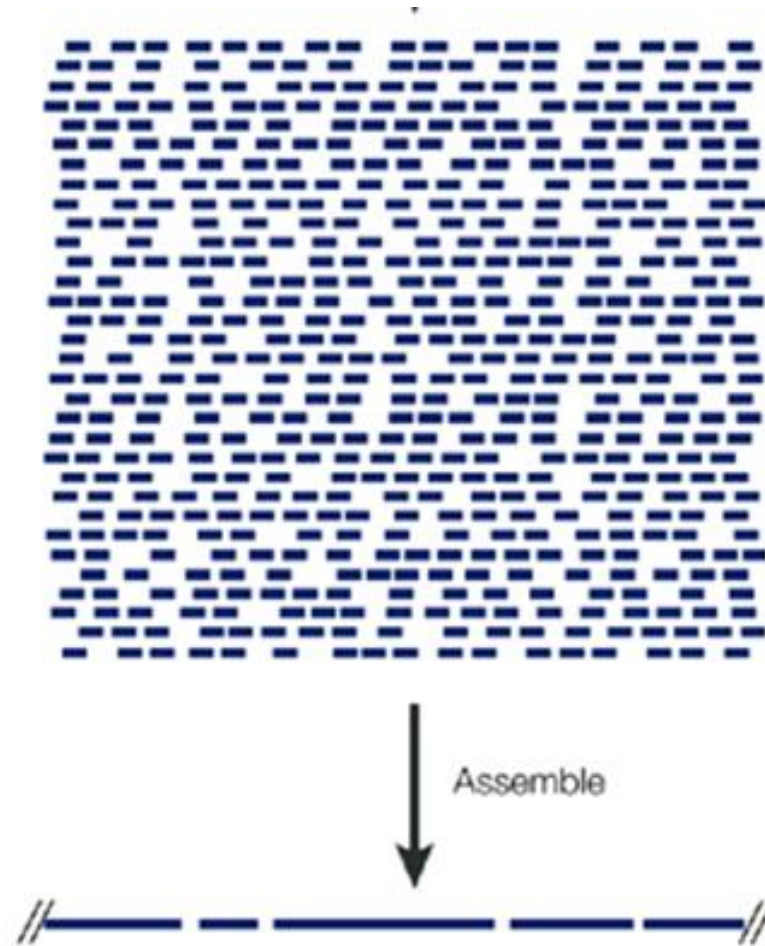
Each genome sequence is a **treasure trove** containing an endless and invaluable source of biological information

- 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?



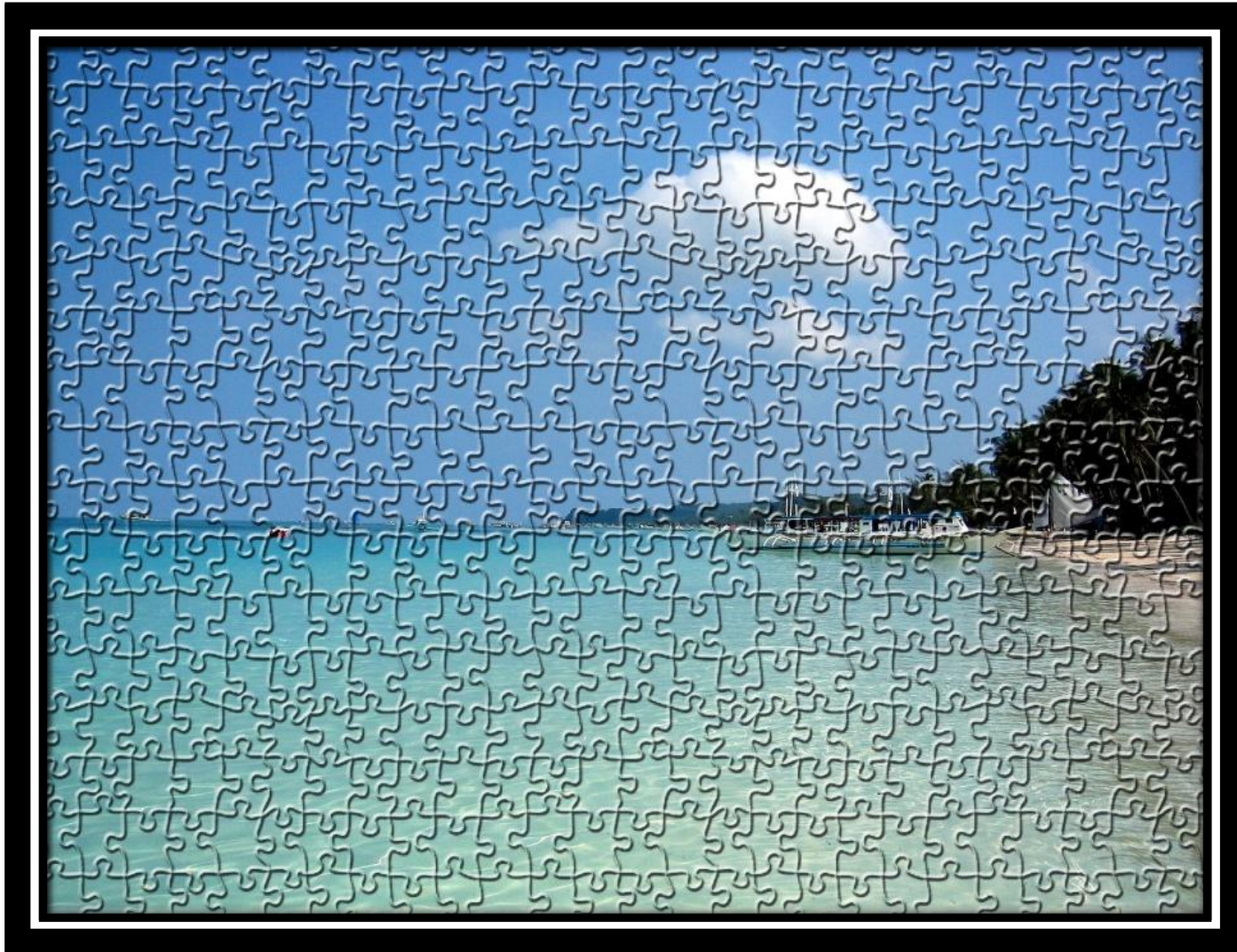
The assembly puzzle

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

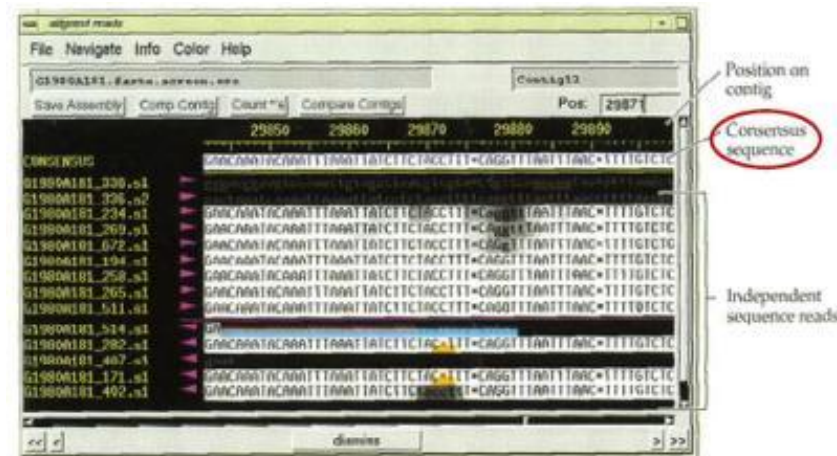
Base quality => phred score (Q)

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

P = Probability of calling a wrong base

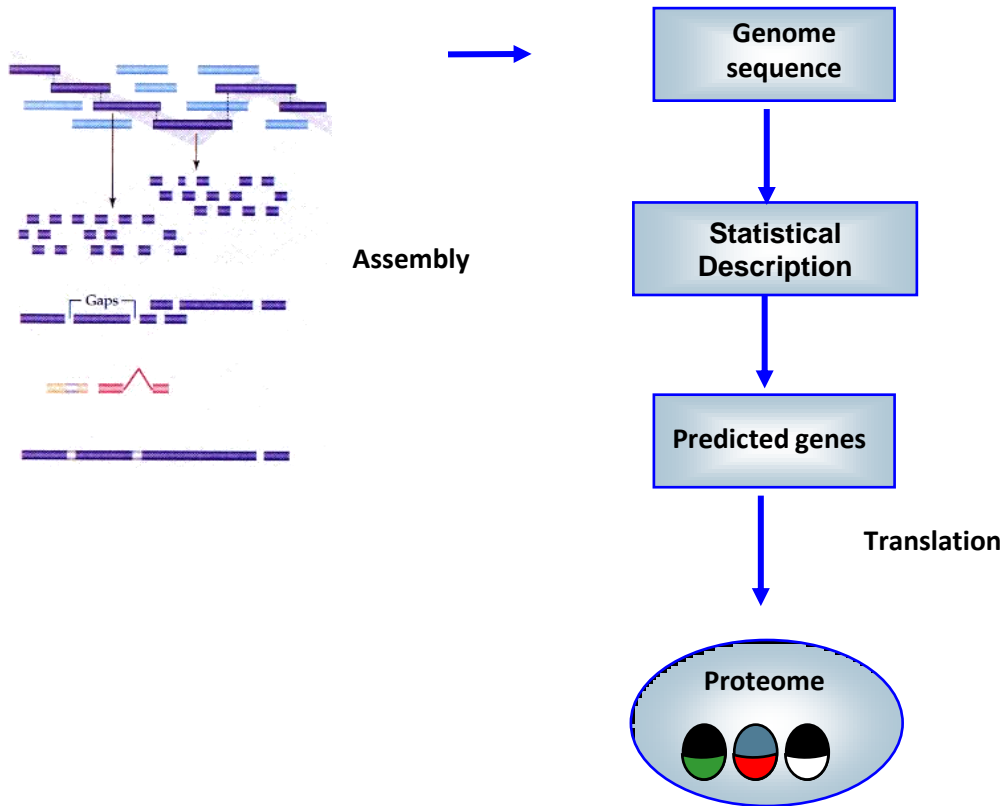
Q = 20 draft sequence (P = 0.01)

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

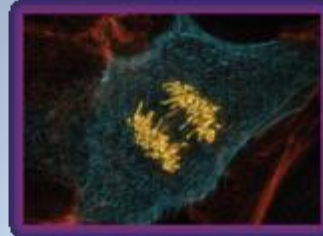


The technological explosion

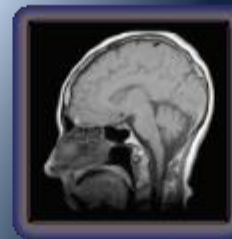
Technology Advances Drive Science



Astronomy



Cell Biology

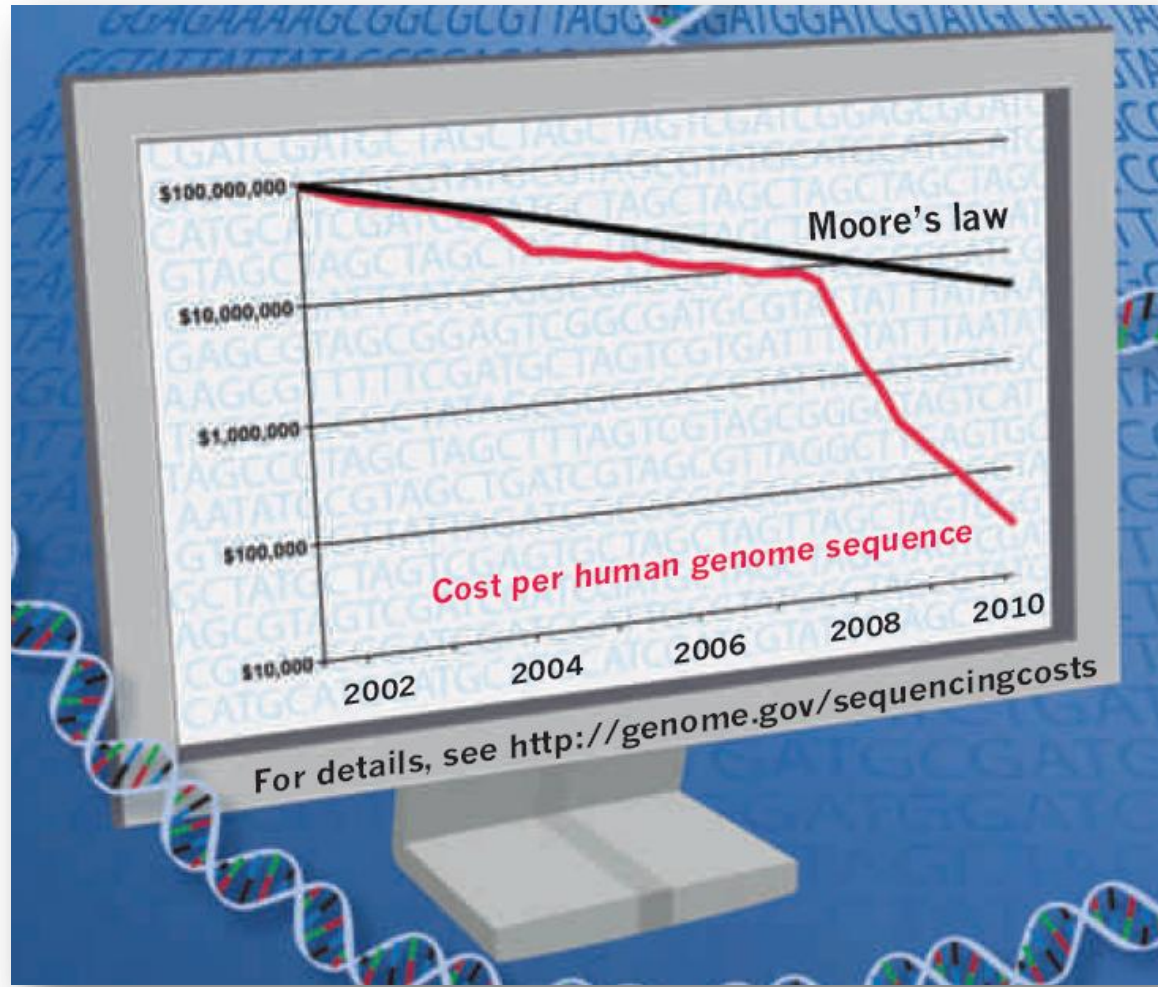


Radiology



Genomics

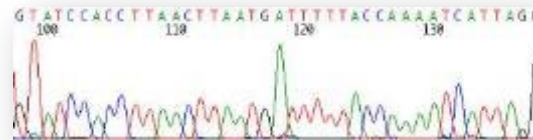
The technological explosion



Sequencing technologies



- Sanger (the *Gold Standard*) (http://www.wiley-vch.de/books/sample/3527320903_c01.pdf) Capillary electrophoresis. 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

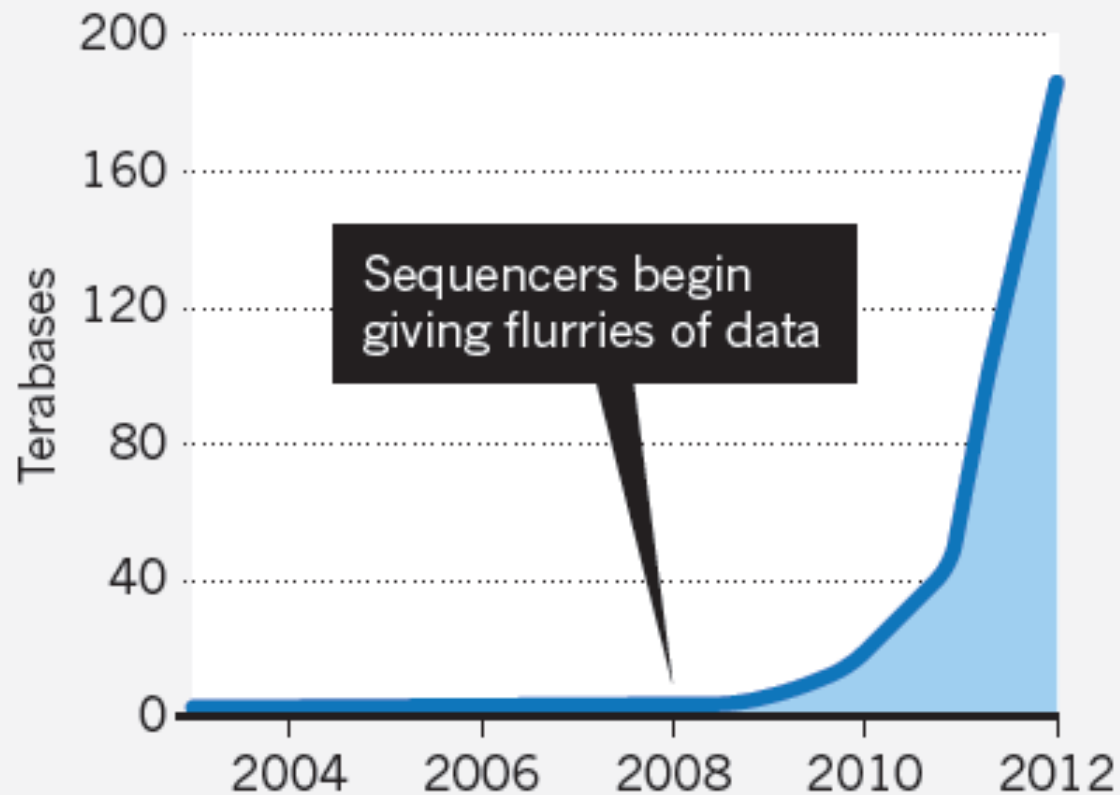


Analytical difficulties (profound changes in data analysis pipelines) → Revitalization of bioinformatics



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:

Examples: [blast_keratin_bf1](#)

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

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 \(joint dixia - Genedata workshop\) - 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](#)

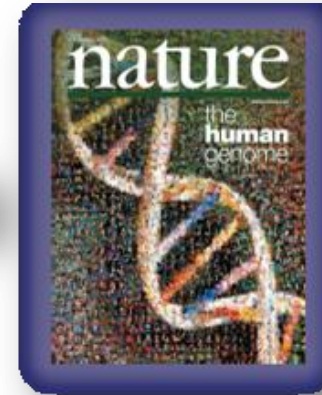
Research infrastructures

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 (ESRI) process. On



The technological explosion

~\$1,000,000,000



Cost: 1 million fold lower!!!



← Today

~\$1,000

"The \$1000 Genome"

Genome Sequencing as a “Commodity”

Sherlock Holmes was an amateur.

SPECIAL PRICING \$4,998 Human Whole Genome Sequencing & Functional Interpretation (min. 10 genomes)

Investigating a genetic disease? We're the genome detectives to call. As experts in the functional interpretation of human genomes, we've built a state-of-the-art pipeline to richly annotate and thoroughly compare up to 300 whole genomes or exomes at once—to quickly track down the variants, genes, and pathways that govern disease. Starting with tissue samples, we deliver analyzed data, a shortlist of suspects, and powerful software to let you close the case in record time.

Knome
From DNA to Discovery

We can help you **identify the variants, genes, and pathways that characterize a genetic disease.** Visit www.knome.com/disease or call (857) 453-3895 to learn more.

华大基因 Premier Scientific Partner
BGI

Roche NimbleGen (44Mb) Capture
BGI Illumina HiSeq 2000 Sequencing
Capture All SNPs & Discover Rare Mutations

15,000 Exome / Targeted Region Samples Sequenced by BGI to Date, and Counting

Human Exome Sequencing Starting at \$999

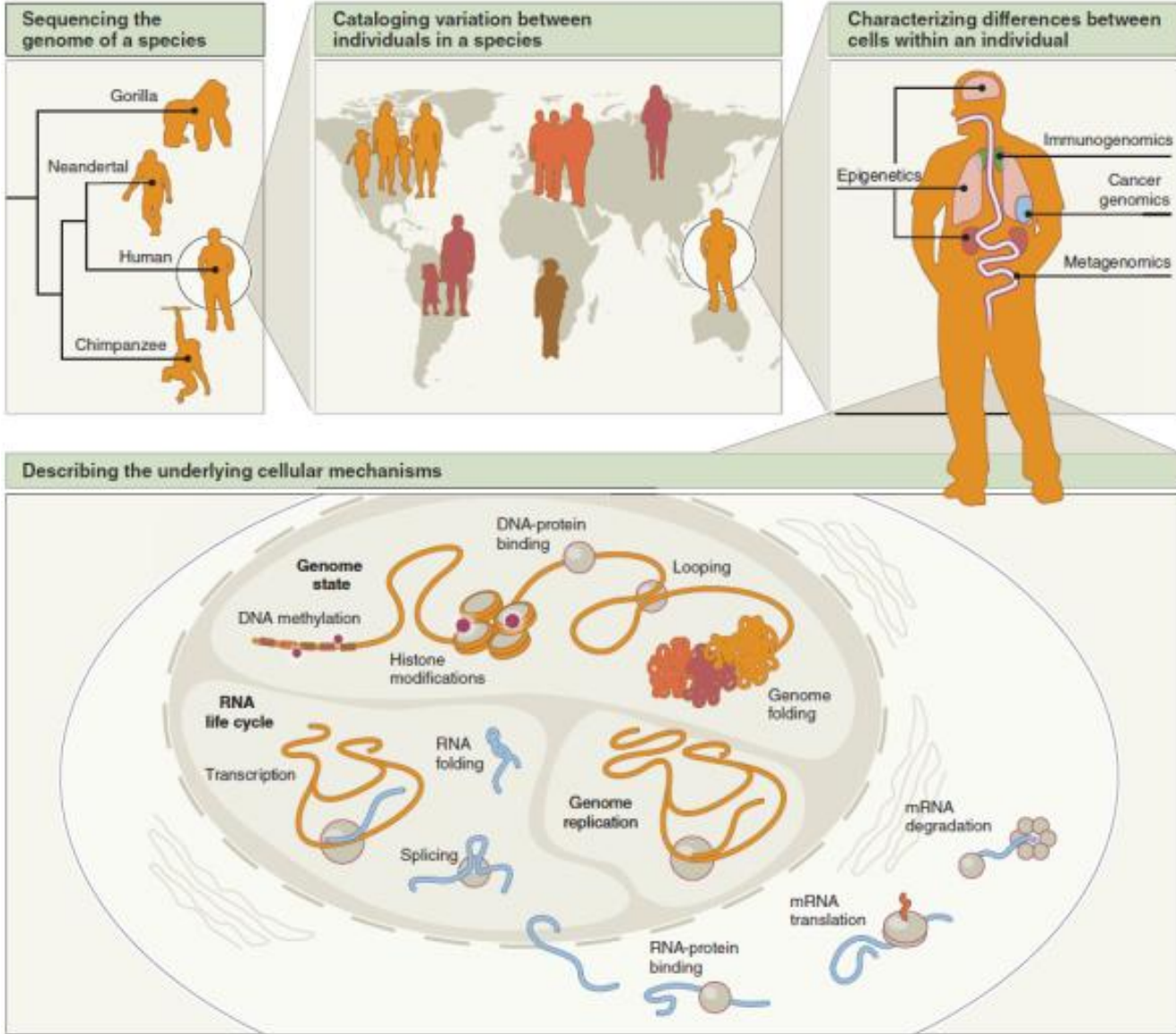
Benefits:

- Target the most functionally relevant DNA sequences
- Capture both common and rare variants missed in traditional GWAS studies
- 150 next-generation sequencers assure rapid turnaround
- 1000 bioinformaticians generate high-quality, reliable data

America: (617) 500-2741 | info@bgi-genomics.com | Europe: +41 52008007 | bgi.europe@genomics.cn

www.bgi-sequence.com

Road map of sequencing science



Completed and ongoing genome projects



Complete Genome Projects: 7396

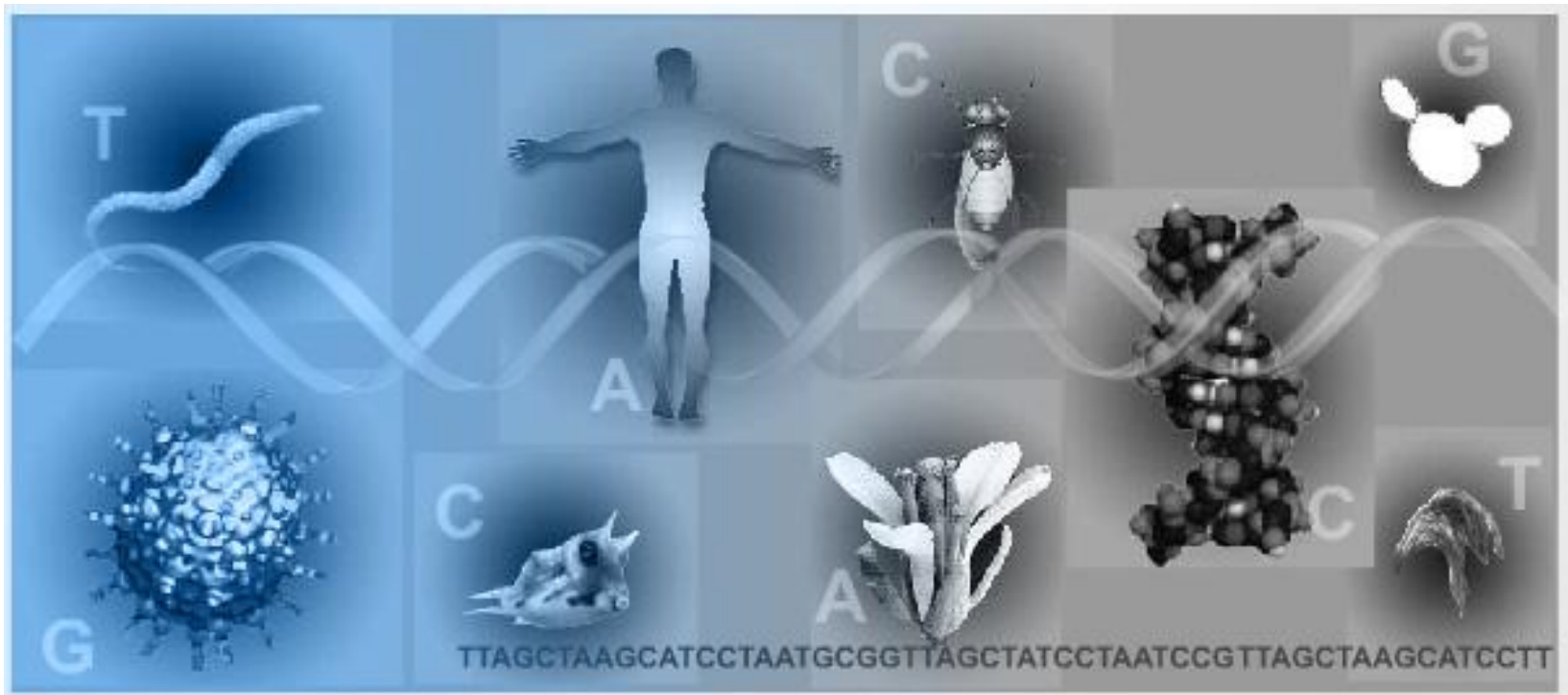
Archaeal: 234

Bacterial: 6851

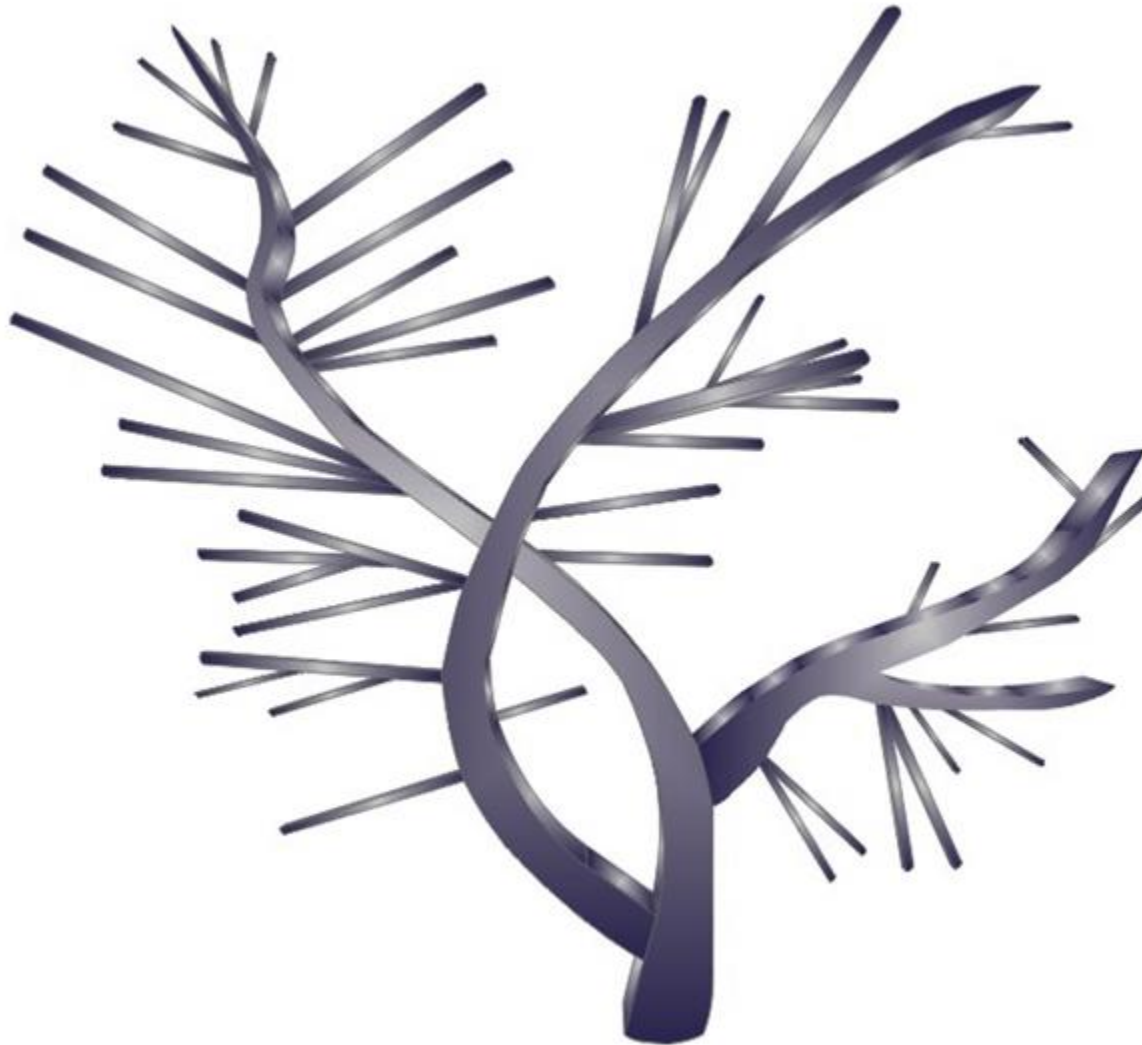
Eukaryal: 311

✓ Finished: 2649

Permanent Draft: 4747



Genomas de especies



Association Studies

Variación fenotípica individual

(incluye la predisposición a enfermedades)



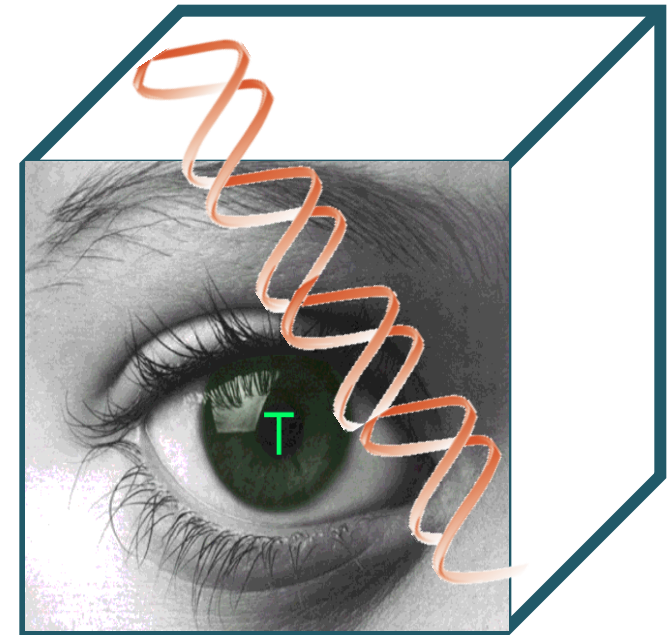
GENOMICS

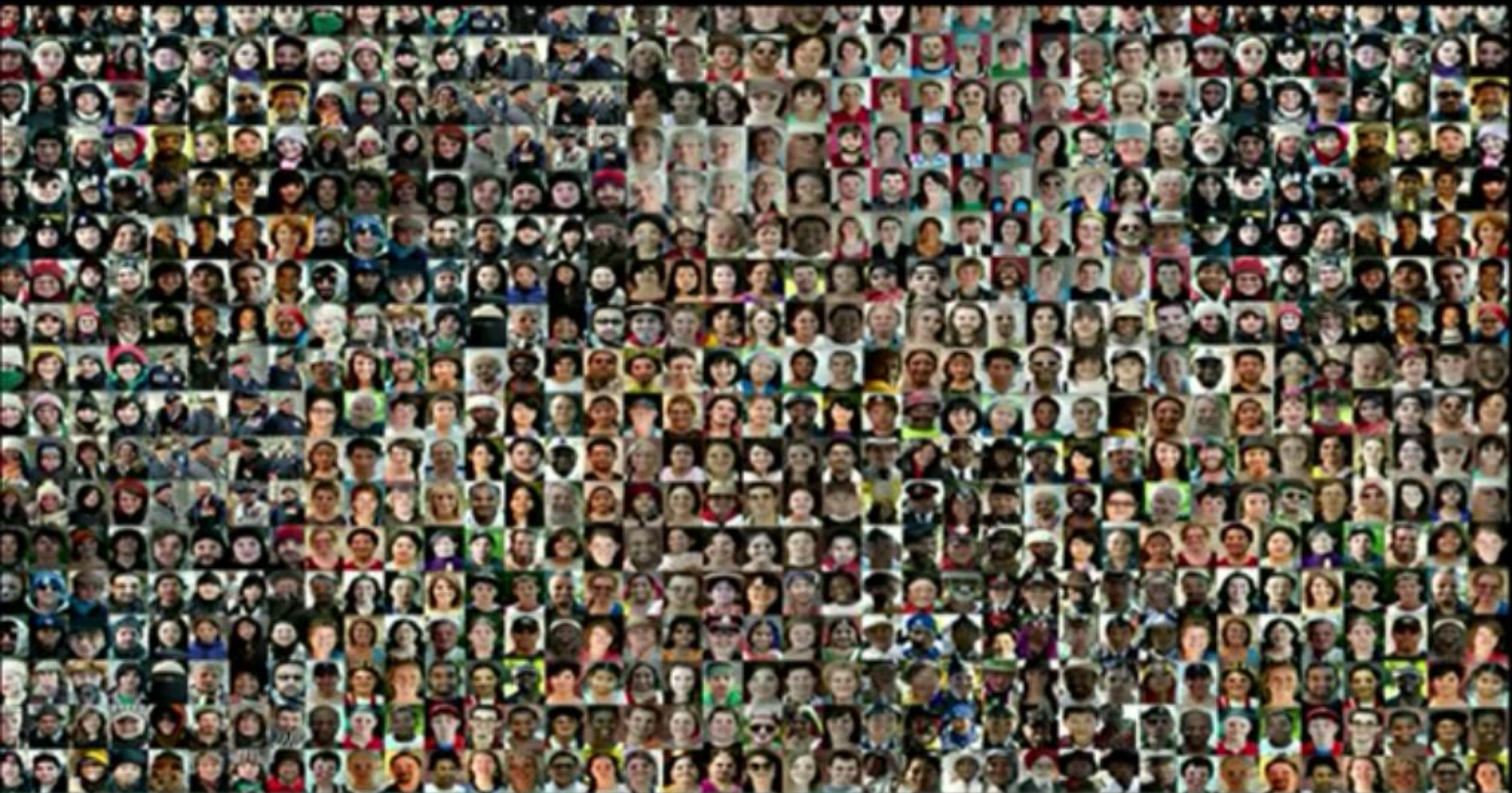
Fenotipo = Genotipo + Ambiente



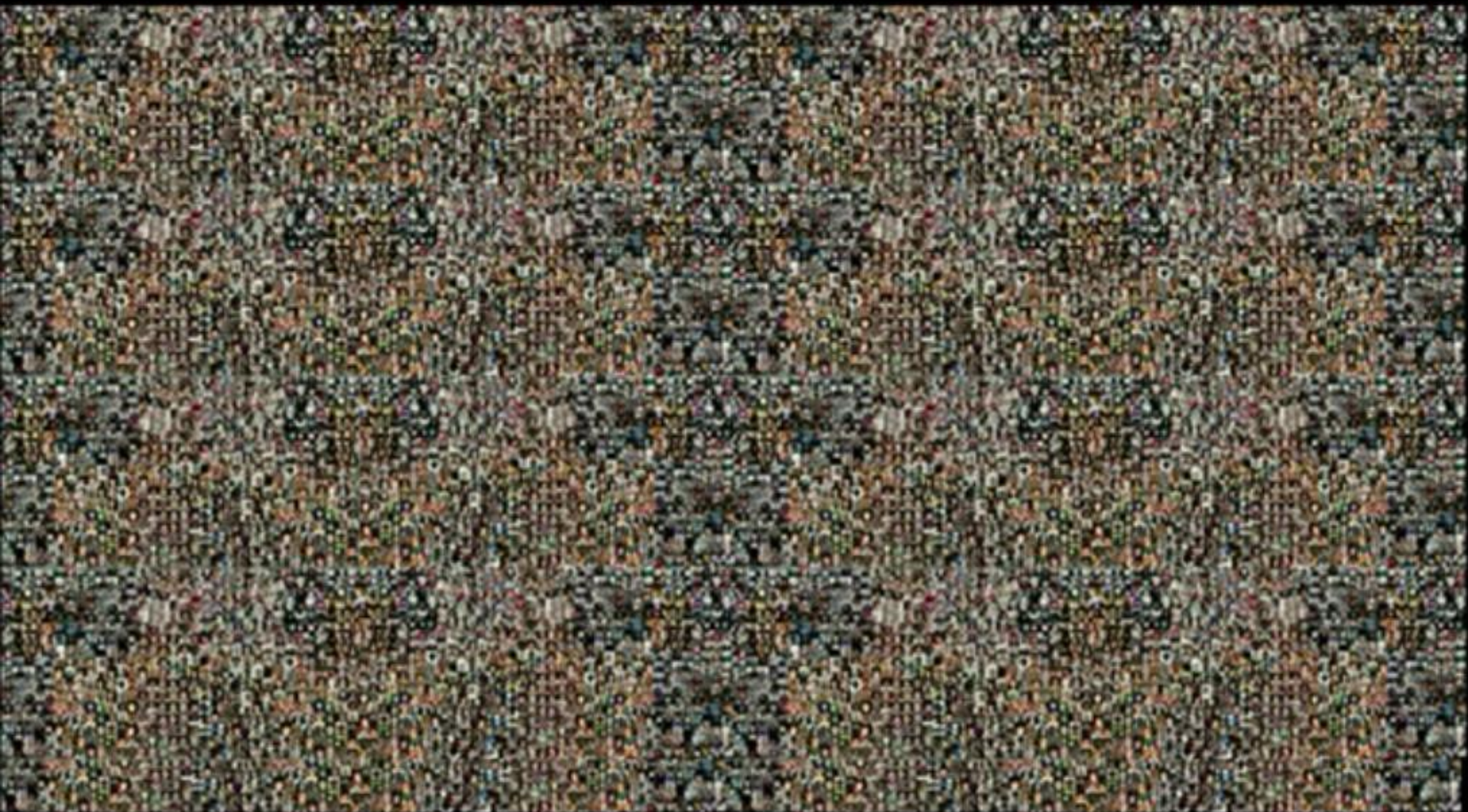
The empirical
space of human
genetic variation:

Data desideratum









Association studies: Phenotypic effect of SNPs

Human genetic & phenotypic diversity database

	Genotype				Phenotype	
	SNP ₁	SNP ₂	SNP ₃	...	Disease 1	Trait i
Sequence individual 1	A/A	G/C	G/T		Healthy	x ₁
Sequence 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



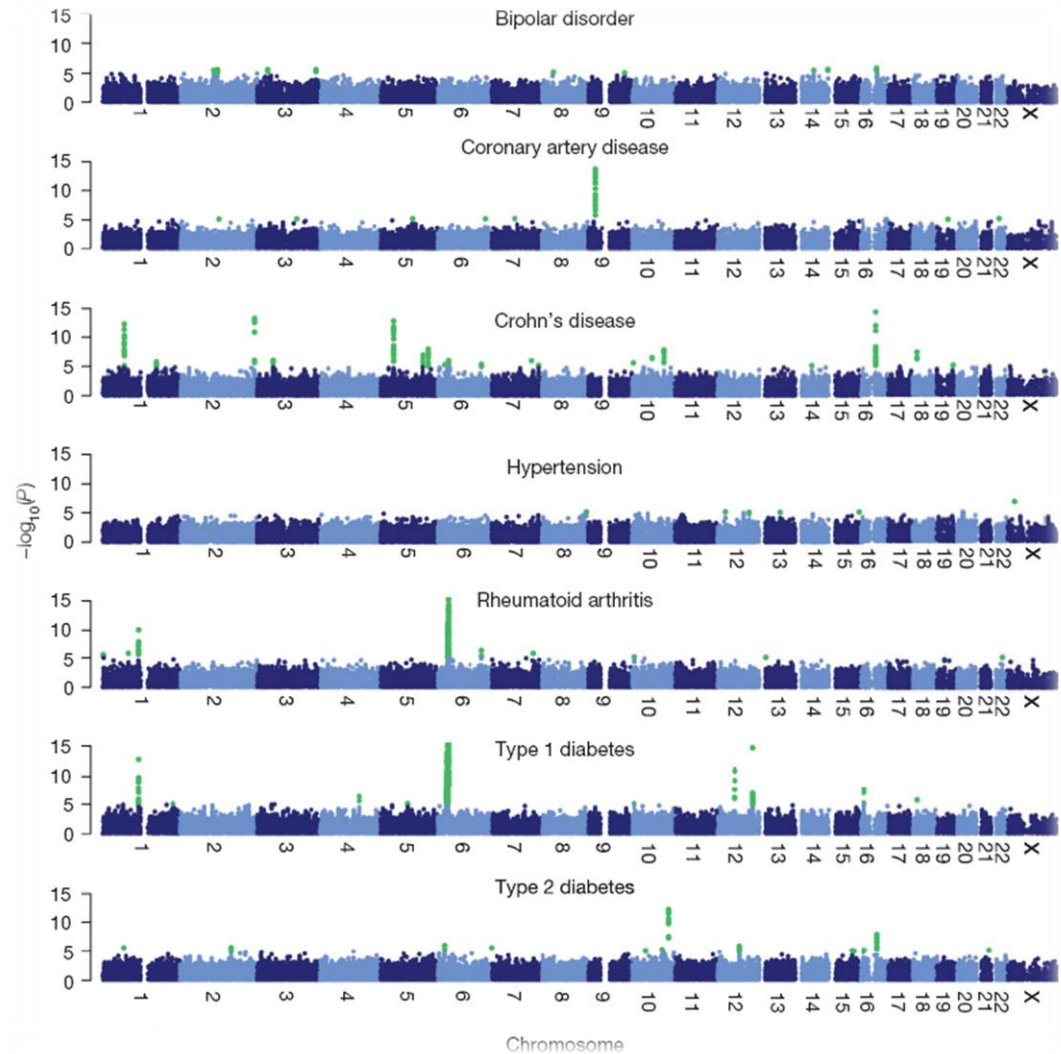
- deCODE (Icelandia)
- Estonia
- Germany
- Canada
- Japan
- China

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

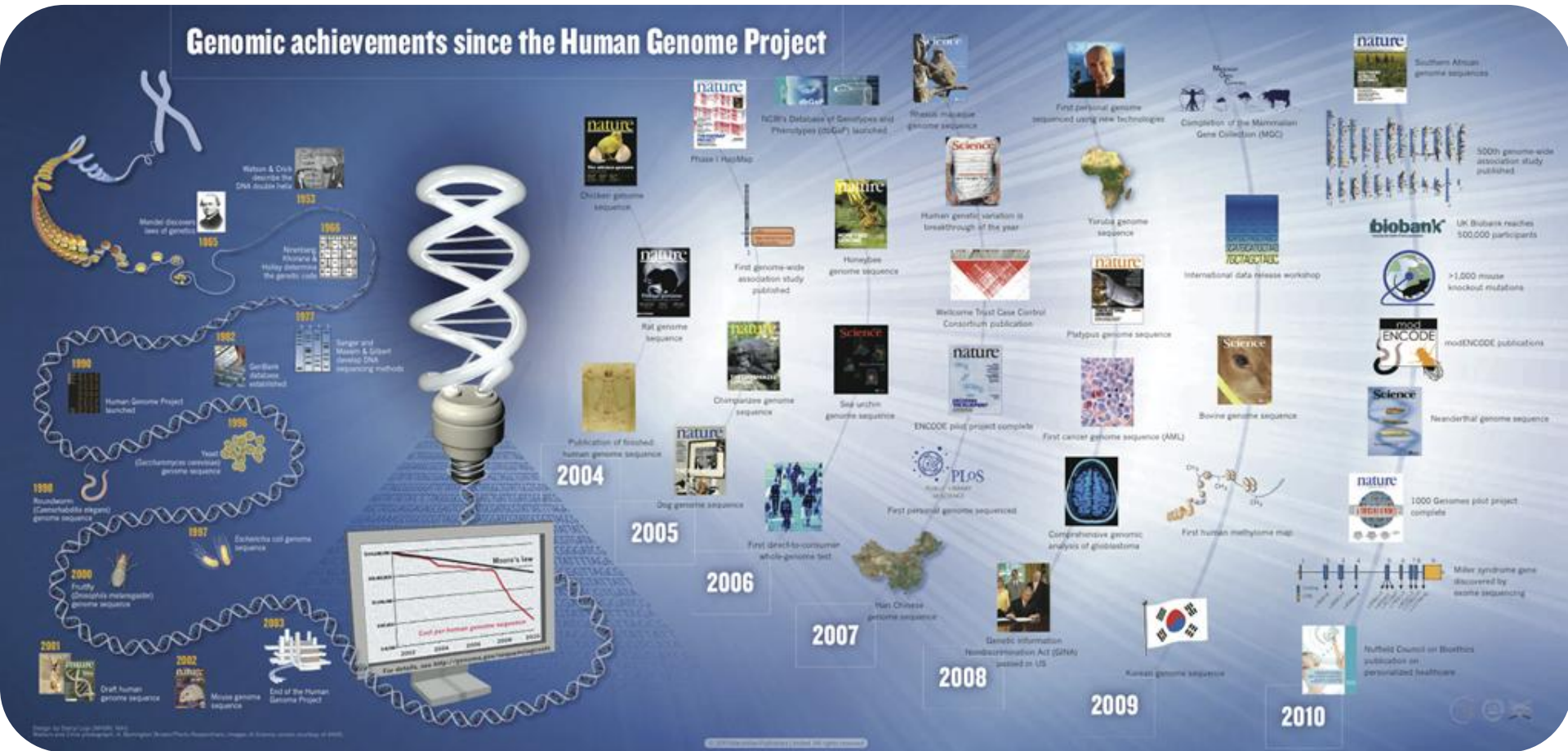
- Data bias
- False positives
- Data structure
- Security issues





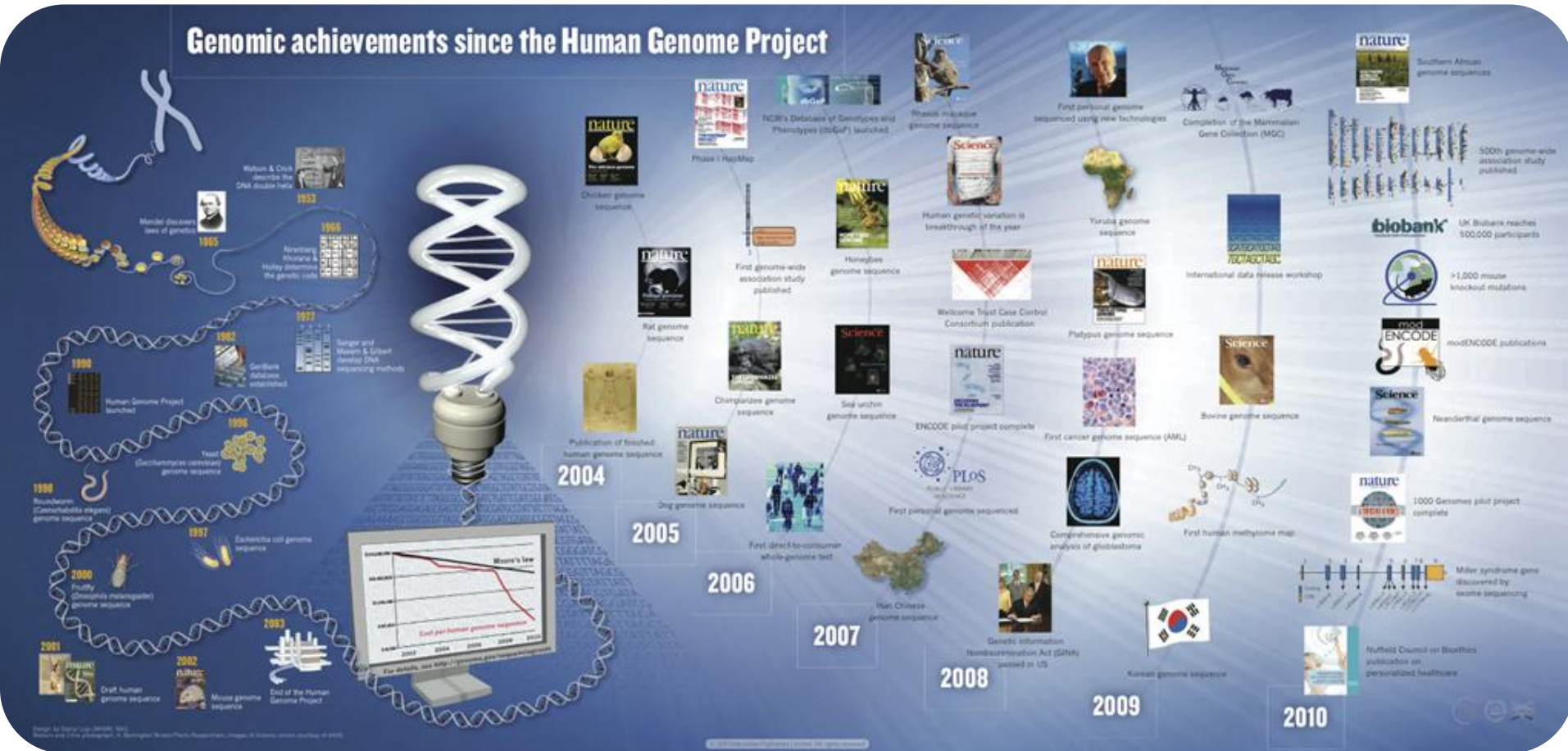
The triumphal march of genomics

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

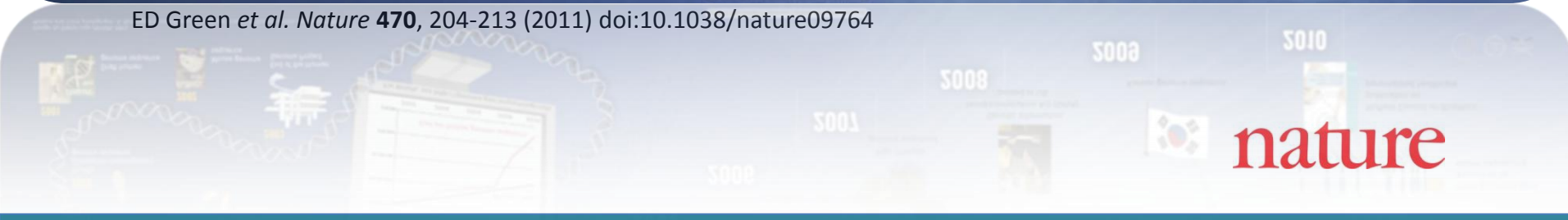


ED Green *et al.* *Nature* **470**, 204-213 (2011) doi:10.1038/nature09764

Genomic achievements since the Human Genome Project



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





Genome science challenges

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 developed¹⁰⁵.

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

1000 genomes Big Data problems

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.

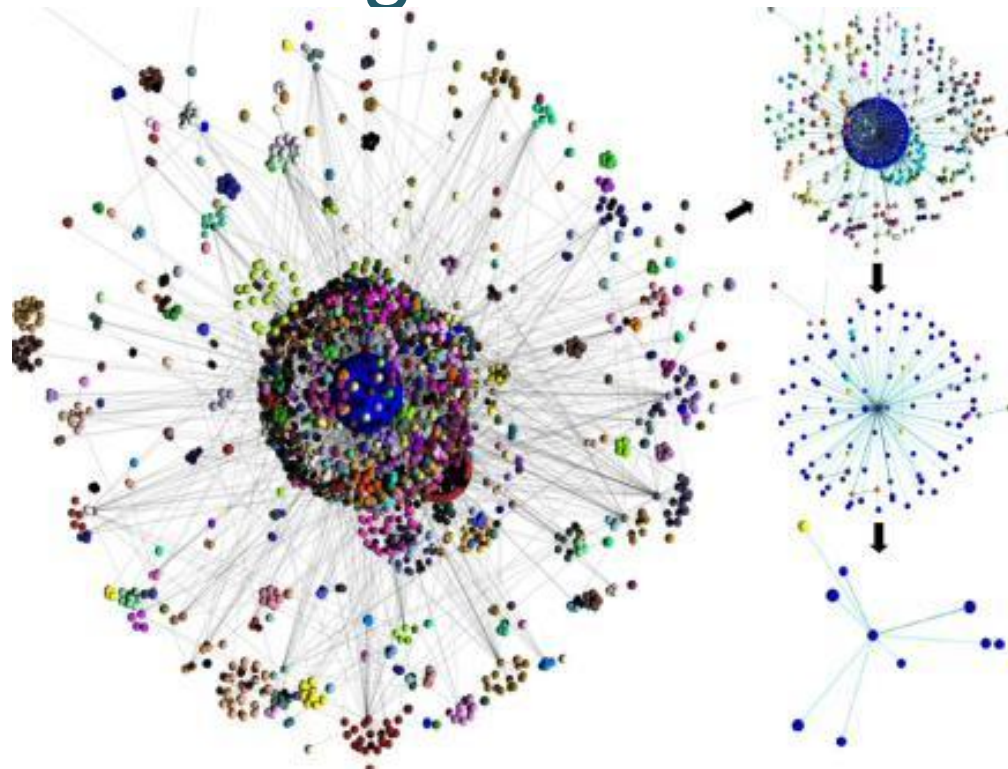
1000 genome project: The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. [Nature 491: 56-65](#)

- 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

Google™ Search 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

The Strategic Plan [Share](#) [Print](#)

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](#)
- **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. 

[Top of page](#)

Last Updated: March 23, 2012



NHGRI Director
Eric Green



PERSPECTIVE

doi:10.1038/nature09764

Charting a course for genomic medicine from base pairs to bedside

Eric D. Green¹, Mark S. Guyer¹ & National Human Genome Research Institute¹

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 project³ is evident in the broad range of scientific advances that the HGP has enabled, as shown in Fig. 1 (see rollout). Optimism about the potential contributions of genomics for improving human health has been fuelled by new insights about cancer^{4,5}, the molecular basis of inherited diseases (<http://www.ncbi.nlm.nih.gov/omim> and <http://www.genome.gov/GWASStudies>) and the role of structural variation in disease⁶, some of which have already led to new therapies⁷⁻¹². Other advances have already changed medical practice (for example, microarrays are now used for clinical detection of genomic imbalances¹³ and pharmacogenomic testing is routinely performed before administration of certain medications¹⁴). Together, these achievements (see accompanying paper¹⁵) document that genomics is contributing to a better understanding of human biology and to improving human health.

As it did eight years ago¹¹, 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.

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

Reflections on the first ten years of the human genomics age

February 2011

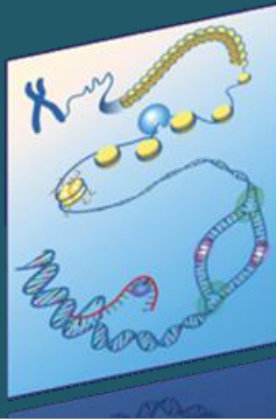
NHGRI Published New Vision for Genomics

Five domains of genomics research

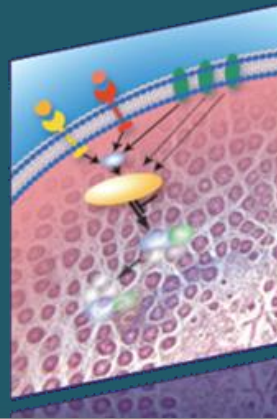
Understanding
the Structure of
Genomes



Understanding
the Biology of
Genomes



Understanding
the Biology of
Disease



Advancing
the Science of
Medicine



Improving the
Effectiveness
of Healthcare



**Basic
Science**

**Translational
Science**

**Implementation
Science**



Schematic representation of accomplishments across five domains of genomics research

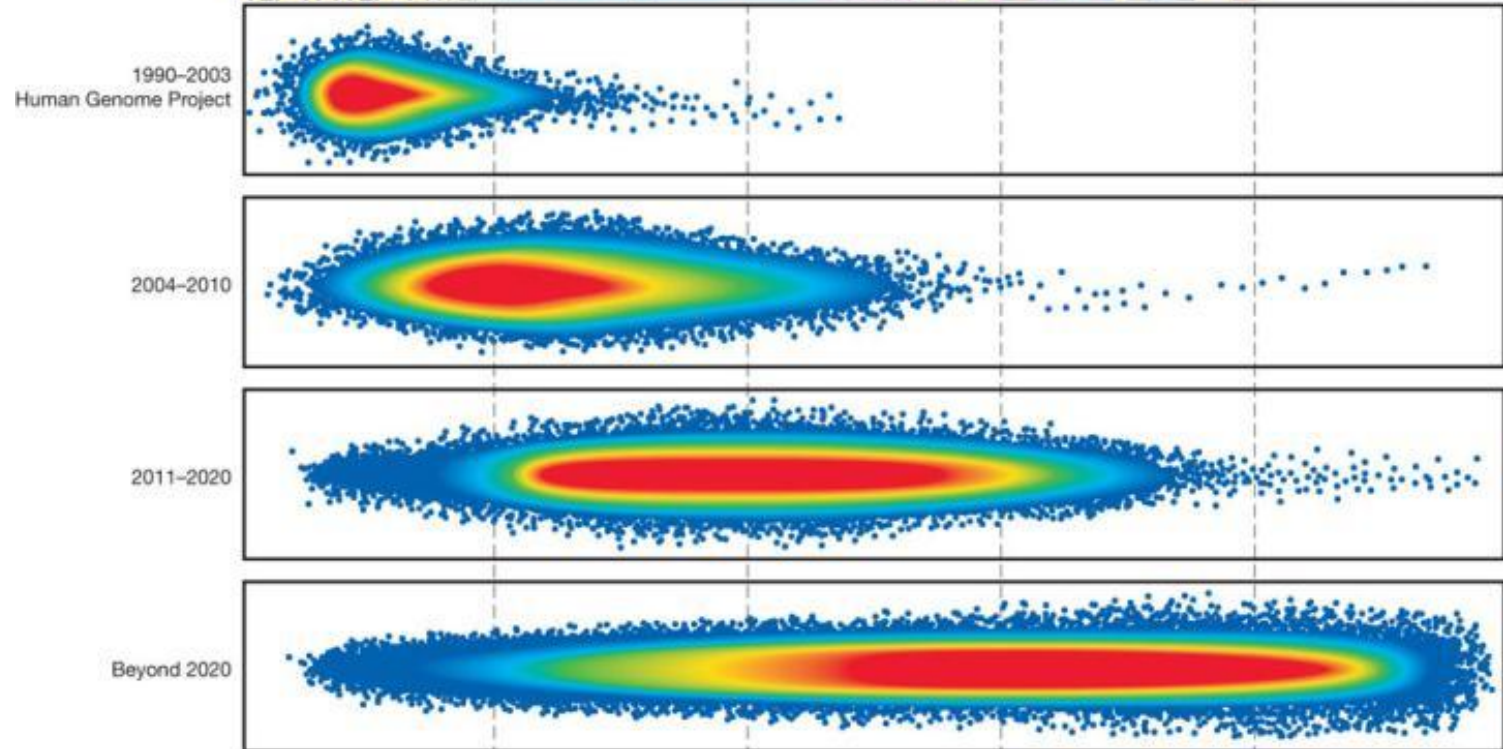
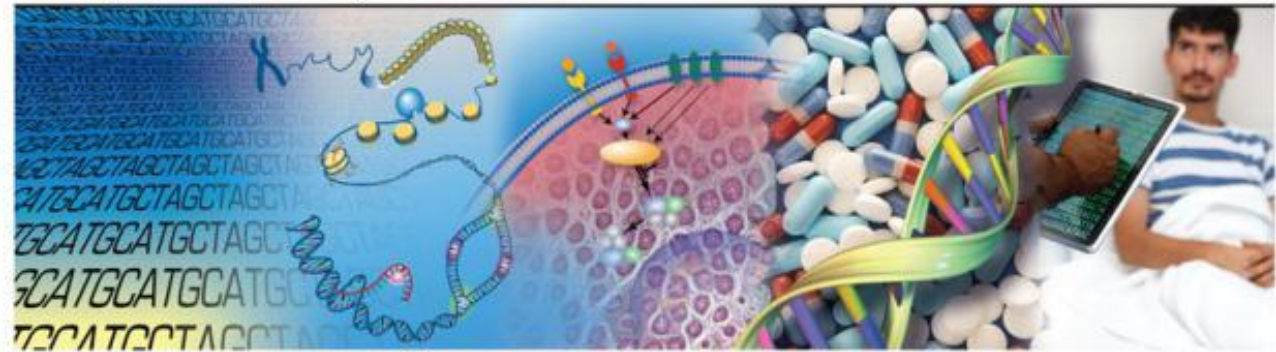
Understanding the structure of genomes

Understanding the biology of genomes

Understanding the biology of disease

Advancing the science of medicine

Improving the effectiveness of healthcare



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

nature

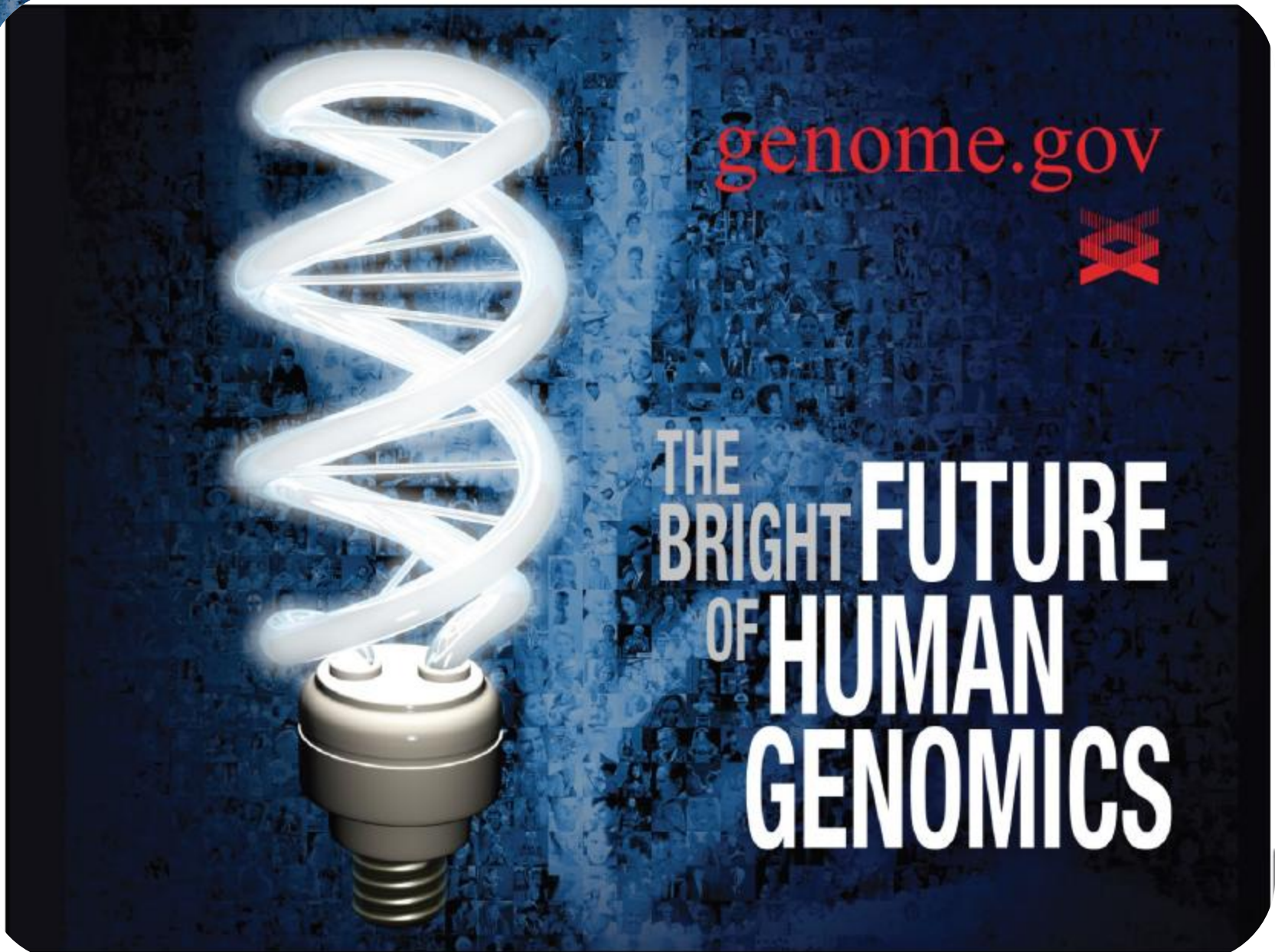
*This is the greatest
intellectual moment
in history*



agctaagacigtggaageatcagctggcagcgaactgacgtagegatcagatcgcctcgcgalegattatcggtcgaagaaagctctagccga



ACTGA
TTACT





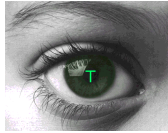
Readings & Videos

Readings

- Perspective: [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

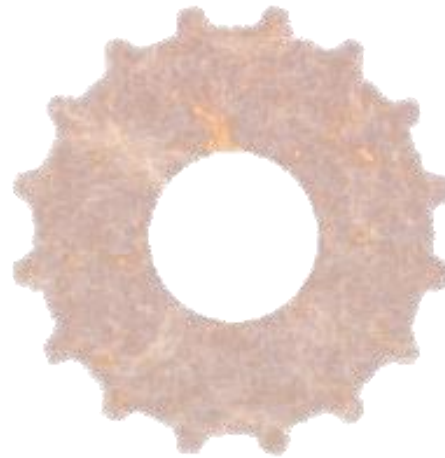
Videos

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

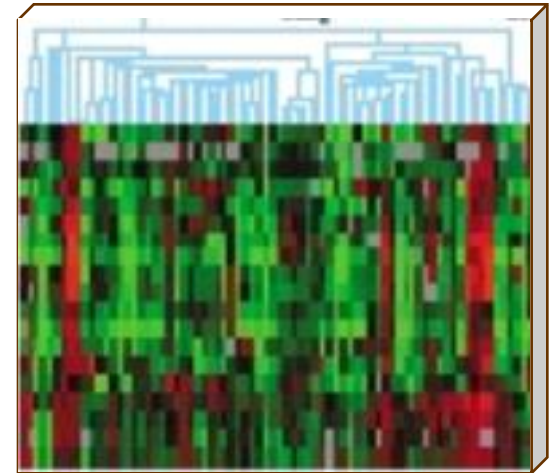


```
ATGTGCAATGCTT
CGTTACGGCTCAA
TATGCCGCAGTAA
GCTGCAGTATCCG
CCGCAGTAACTGG
GCCGCAG.....
```

Datos



Herramientas
bioinformáticas



Conocimiento