Open Acess in Bioinformatics & Genomics

Roderic Guigó Serra

Centre de Regulació Genòmica (CRG)

roderic.guigo@crg.cat

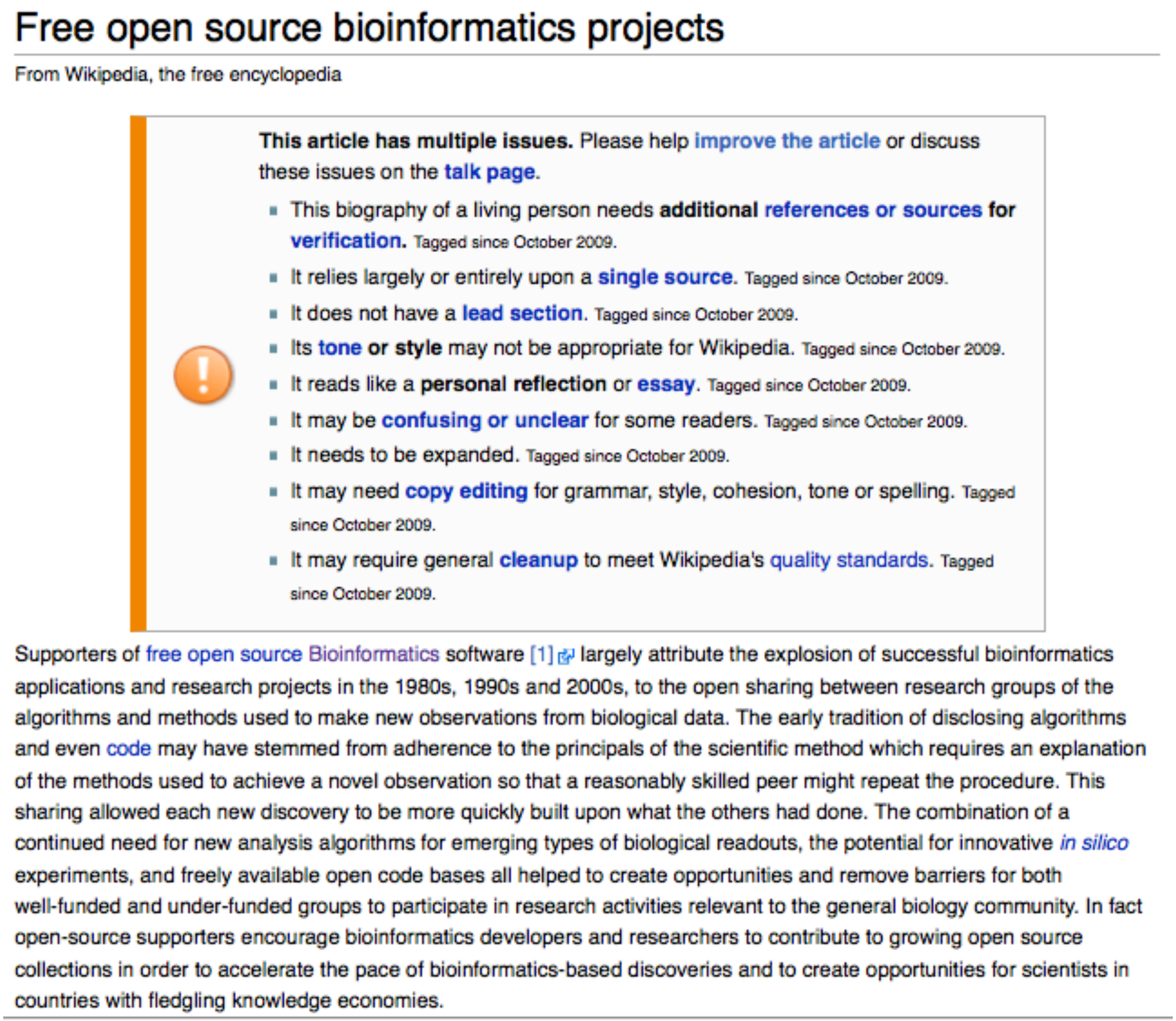
Tres ideas

- Open source in Bioinformatics Software
- Open Access to Genome Information
  - The case for individual genomes
- (Open Access to Scientific Publication)

- Nosaltres som un grup de desenvolupament en Bioinformatica. Desenvolupem programes per l’analisi de sequencies genomiques.
- Quan varem tenir les primeres implementacions dels nostres programes varem decidir posar-les en el domini public

Bioinformatics & Unix

- Early development of Bioinformatics strongly linked to the UNIX operating System, and GNU tools become very popular among bioinformatic programers and developers: emacs, gcc, bash, gawk



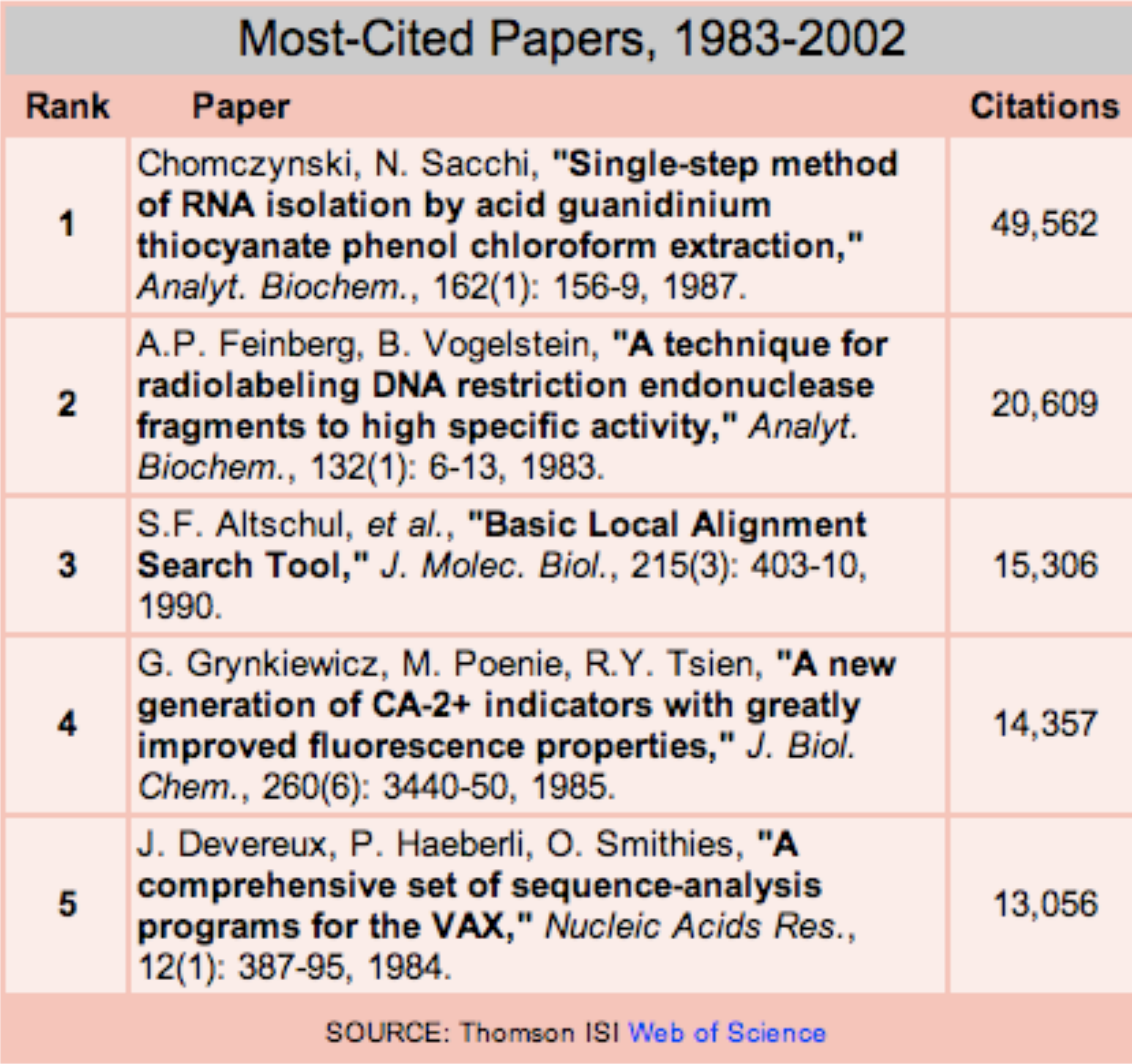
Free open source bioinformatics projects

From Wikipedia, the free encyclopedia

This article has multiple issues. Please help improve the article or discuss these issues on the **talk page**.

- This biography of a living person needs **additional references or sources for verification.** Tagged since October 2009.
- It relies largely or entirely upon a **single source.** Tagged since October 2009.
- It does not have a **lead section**. Tagged since October 2009.
- Its **tone or style** may not be appropriate for Wikipedia. Tagged since October 2009.
- It reads like a **personal reflection** or **essay**. Tagged since October 2009.
- It may be **confusing or unclear** for some readers. Tagged since October 2009.
- It needs to be expanded. Tagged since October 2009.
- It may need **copy editing** for grammar, style, cohesion, tone or spelling. Tagged since October 2009.
- It may require general **cleanup** to meet Wikipedia's quality standards. Tagged since October 2009.

Supporters of free open source Bioinformatics software [1] largely attribute the explosion of successful bioinformatics applications and research projects in the 1980s, 1990s and 2000s, to the open sharing between research groups of the algorithms and methods used to make new observations from biological data. The early tradition of disclosing algorithms and even code may have stemmed from adherence to the principals of the scientific method which requires an explanation of the methods used to achieve a novel observation so that a reasonably skilled peer might repeat the procedure. This sharing allowed each new discovery to be more quickly built upon what the others had done. The combination of a continued need for new analysis algorithms for emerging types of biological readouts, the potential for innovative *in silico* experiments, and freely available open code bases all helped to create opportunities and remove barriers for both well-funded and under-funded groups to participate in research activities relevant to the general biology community. In fact open-source supporters encourage bioinformatics developers and researchers to contribute to growing open source collections in order to accelerate the pace of bioinformatics-based discoveries and to create opportunities for scientists in countries with fledgling knowledge economies.

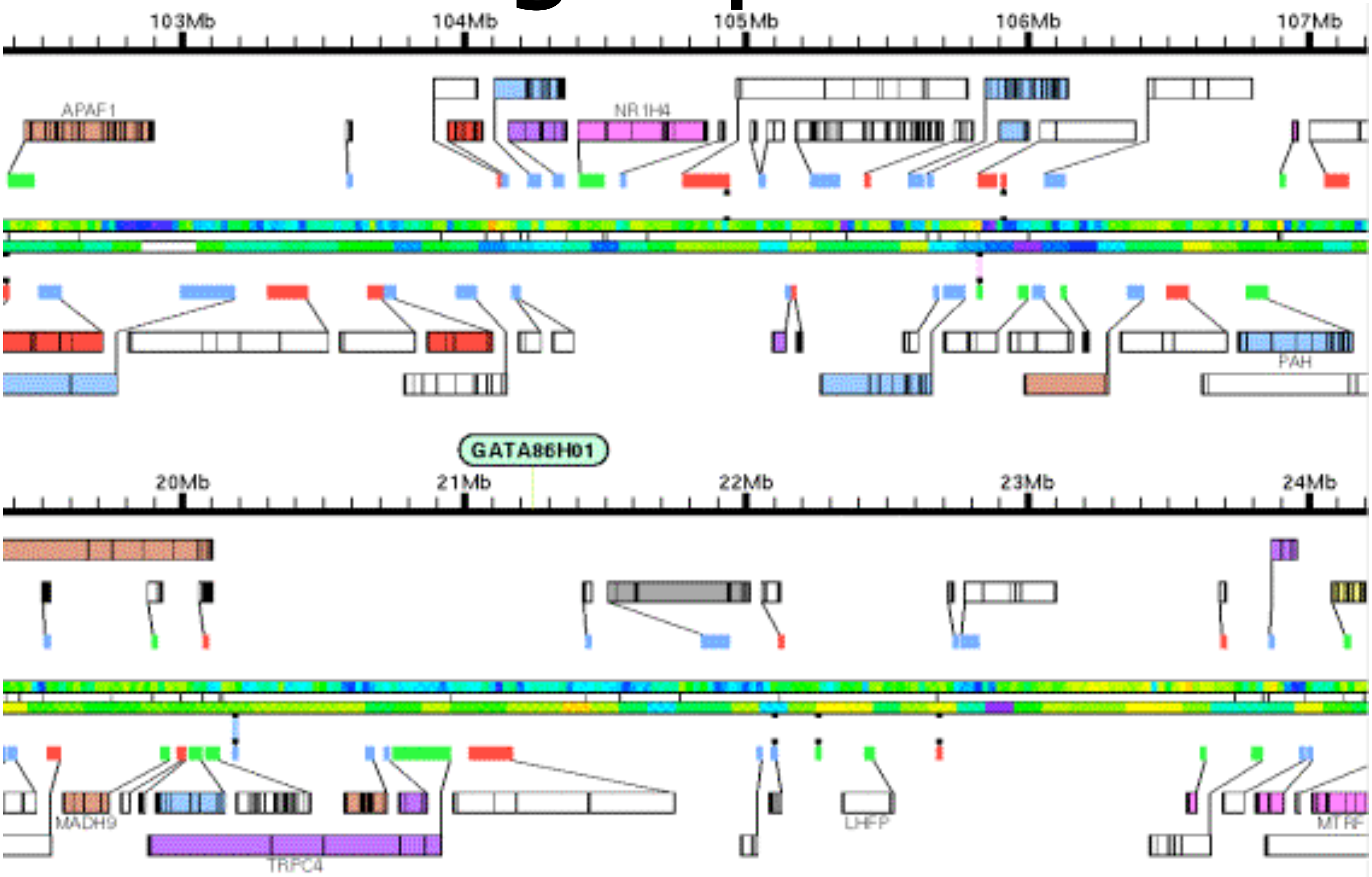


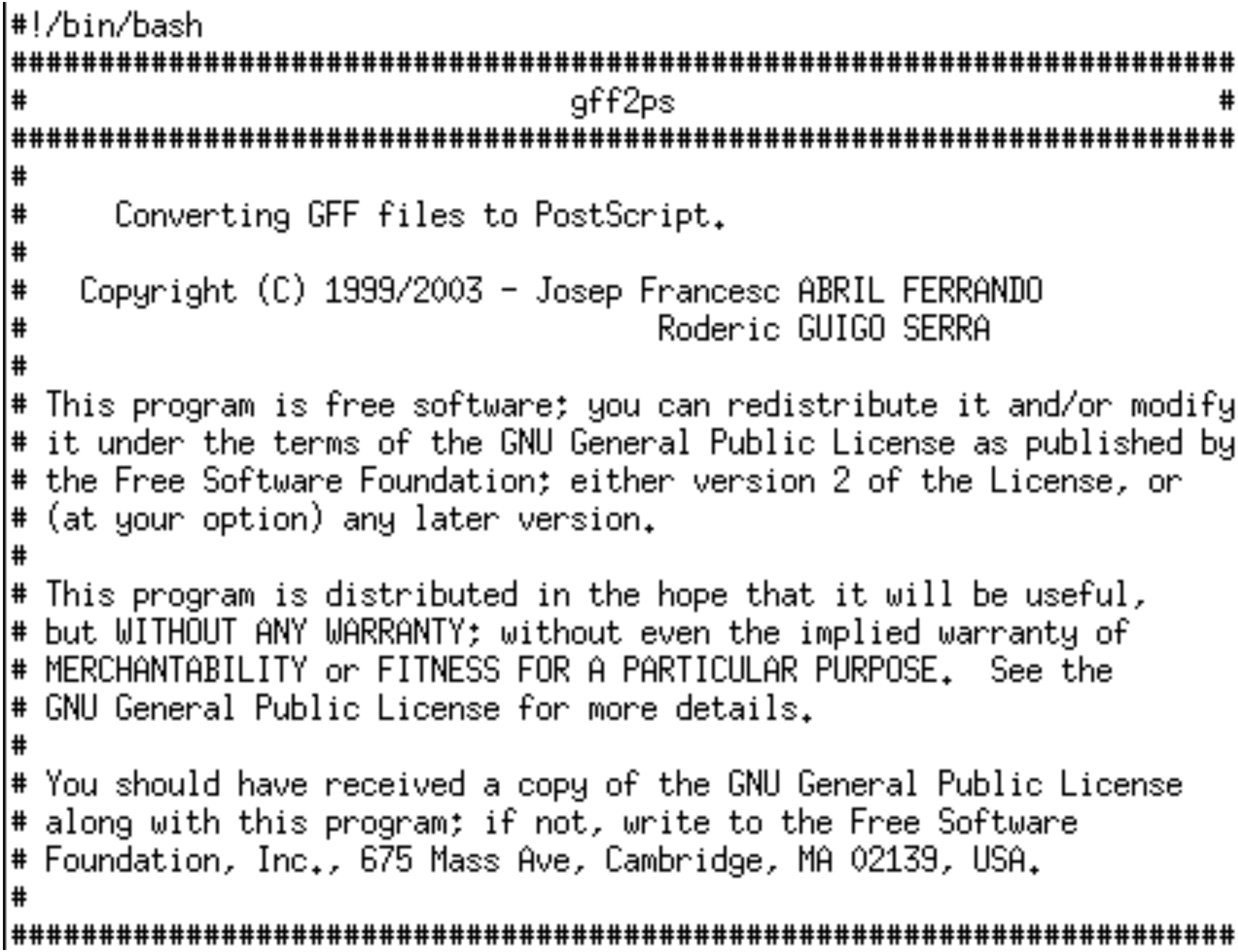
Most-Cited Papers, 1983-2002

Rank	Paper	Citations
1	Chomczynski, N. Sacchi, **"Single-step method of RNA isolation by acid guanidinium thiocyanate phenol chloroform extraction,"** *Analyt. Biochem.*, 162(1): 156-9, 1987.	49,562
2	A.P. Feinberg, B. Vogelstein, **"A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity,"** *Analyt. Biochem.*, 132(1): 6-13, 1983.	20,609
3	S.F. Altschul, *et al.*, **"Basic Local Alignment Search Tool,"** *J. Molec. Biol.*, 215(3): 403-10, 1990.	15,306
4	G. Grynkiewicz, M. Poenie, R.Y. Tsien, **"A new generation of CA-2+ indicators with greatly improved fluorescence properties,"** *J. Biol. Chem.*, 260(6): 3440-50, 1985.	14,357
5	J. Devereux, P. Haeberli, O. Smithies, **"A comprehensive set of sequence-analysis programs for the VAX,"** *Nucleic Acids Res.*, 12(1): 387-95, 1984.	13,056

SOURCE: Thomson ISI Web of Science

gff2ps


103Mb
104Mb
105Mb
106Mb
107Mb
APAF1
NR1H4
PAH
GATA86H01
20Mb
21Mb
22Mb
23Mb
24Mb
MADH9
TRPC4
LHFP
MTRF



```
#!/bin/bash
######################################################################
#                               gff2ps                               #
######################################################################
#
#     Converting GFF files to PostScript.
#
#   Copyright (C) 1999/2003 - Josep Francesc ABRIL FERRANDO
#                                    Roderic GUIGO SERRA
#
# This program is free software; you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation; either version 2 of the License, or
# (at your option) any later version.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
# GNU General Public License for more details.
#
# You should have received a copy of the GNU General Public License
# along with this program; if not, write to the Free Software
# Foundation, Inc., 675 Mass Ave, Cambridge, MA 02139, USA.
#
######################################################################
```

Science
ANNOTATION OF THE CELERA HUMAN GENOME A

Science
16 February 2001
Vol. 291 No. 5507
Pages 1145–1434 $9
THE HUMAN GENOME
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

bioinformàtica

Google search: X-informatics (11 juny, 2007)

bioinformatics	14,100,000
chemoinformatics	226,000
astroinformatics	195
neuroinformatics	364,000
socioinformatics	610
geoinformatics	506,000
meteoinformatics	48
econoinformatics	441
ecoinformatics	160,000

Open Access to Genome informatation

Sharing Data from Large-scale Biological Research Projects

1996: Bermuda principles

The goal of the agreement was to provide a basis for a free sharing of pre-published data on gene sequences among scientists.

- Automatic release of sequence assemblies larger than 1 kb (preferably within 24 hours).
- Immediate publication of finished annotated sequences.
- Aim to make the entire sequence freely available in the public domain for both research and development in order to maximise benefits to society.

2003: Fort Lauderdale principles

- Large-insert clone-based projects: DNA sequence assemblies of 2 kb or greater are to be deposited in a public nucleotide sequence database (GenBank, EMBL or DDBJ) within 24 hours of generation. Sequence traces from these projects are to be deposited in a trace archive (NCBI Trace Repository or Ensembl Trace Server) within one week of production.

- Whole genome shotgun projects: Sequence traces from whole genome shotgun projects are to be deposited in a trace archive (NCBI Trace Repository or Ensembl Trace Server) within one week of production. Whole genome assemblies are to be deposited in a public nucleotide sequence database as soon as possible after the assembled sequence has met a set of quality evaluation criteria.

- **The deposited data should be available for all to use without restriction.**

2008: ENCODE Project

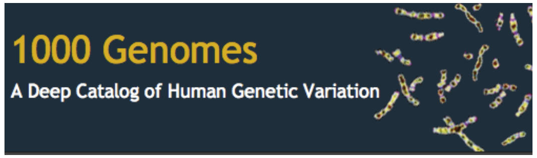
NHGRI has designated the Encyclopedia of DNA Elements (ENCODE) and model organism ENCODE (modENCODE) Projects as **community resource projects to accelerate access to and use of the data by the entire scientific community.** ENCODE/modENCODE resource producers will release data, as soon as they have been verified and prior to publication, to public databases. At the same time, until the data are published upon in a peer-reviewed journal, NHGRI asks <u>resource users to consider them to be unpublished and to follow standard scientific etiquette</u> regarding the use of unpublished data. Specifically, resource users are asked to respect the ability of the producers to publish an initial analysis of the data they have generated in a timely manner. To facilitate this compromise between unrestricted use of the data and unavailability of the data until publication, **NHGRI will promote observation of a 9-month period during which resource users may freely use the ENCODE/modENCODE data to design and carry out their own research programs, but not to submit publications that use unpublished ENCODE/modENCODE data without prior consent**. (NHGRI recognizes that there may be some exceptions to this blanket request; examples are discussed more fully below.) After the expiration of this 9-month period or publication of the data (whichever comes first), resource users should continue to properly acknowledge the ENCODE or modENCODE Project and resource producer(s) as the source of the data in any publication.

Slide from Paul Flicek. EBI Bioinformatics Advisory Council

Further Evolution of Large-scale Genome Sequencing

- 2000: Human genome working drafts
- Data unit of approximately 10x coverage of human
  - 10 years and cost about $3 billion

- 2008: Major genome centers can sequence the same number of base pairs **every 4 days**
  - 1000 Genome project launched
  - World-wide capacity dramatically increasing



- 2009: **Every 4 hours ($25,000)**
- 2010: **Every 14 minutes ($5,000)**
  - Illumina HiSeq2000 machine produces 200 gigabases per 8 day run (BGI have ordered have 128)

illumina

Complete Genomics | Powering large-scale human genome studies

For more information, contact:

Complete Genomics, Inc.	Waggener Edstrom Worldwide Healthcare
Jennifer Turcotte	Lisa Osborne
Vice President, Marketing	Account Director
(650) 943-2846	(202) 261-7806
jturcotte@completegenomics.com	lisao@waggeneredstrom.com

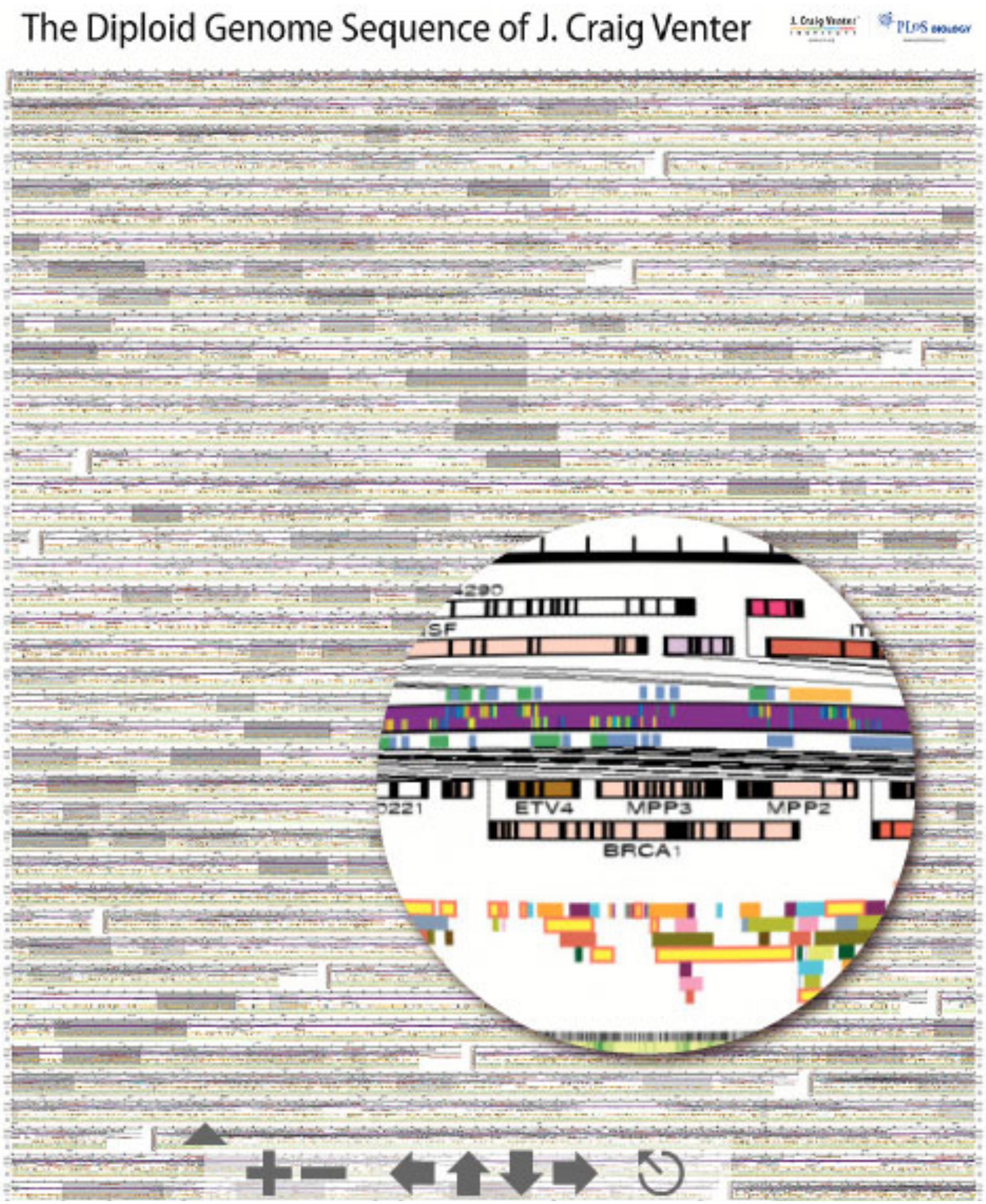
Embargoed until 8 a.m. ET on Monday, October 6, 2008

Complete Genomics Launches; Becomes World's First Large-scale Human Genome Sequencing Company

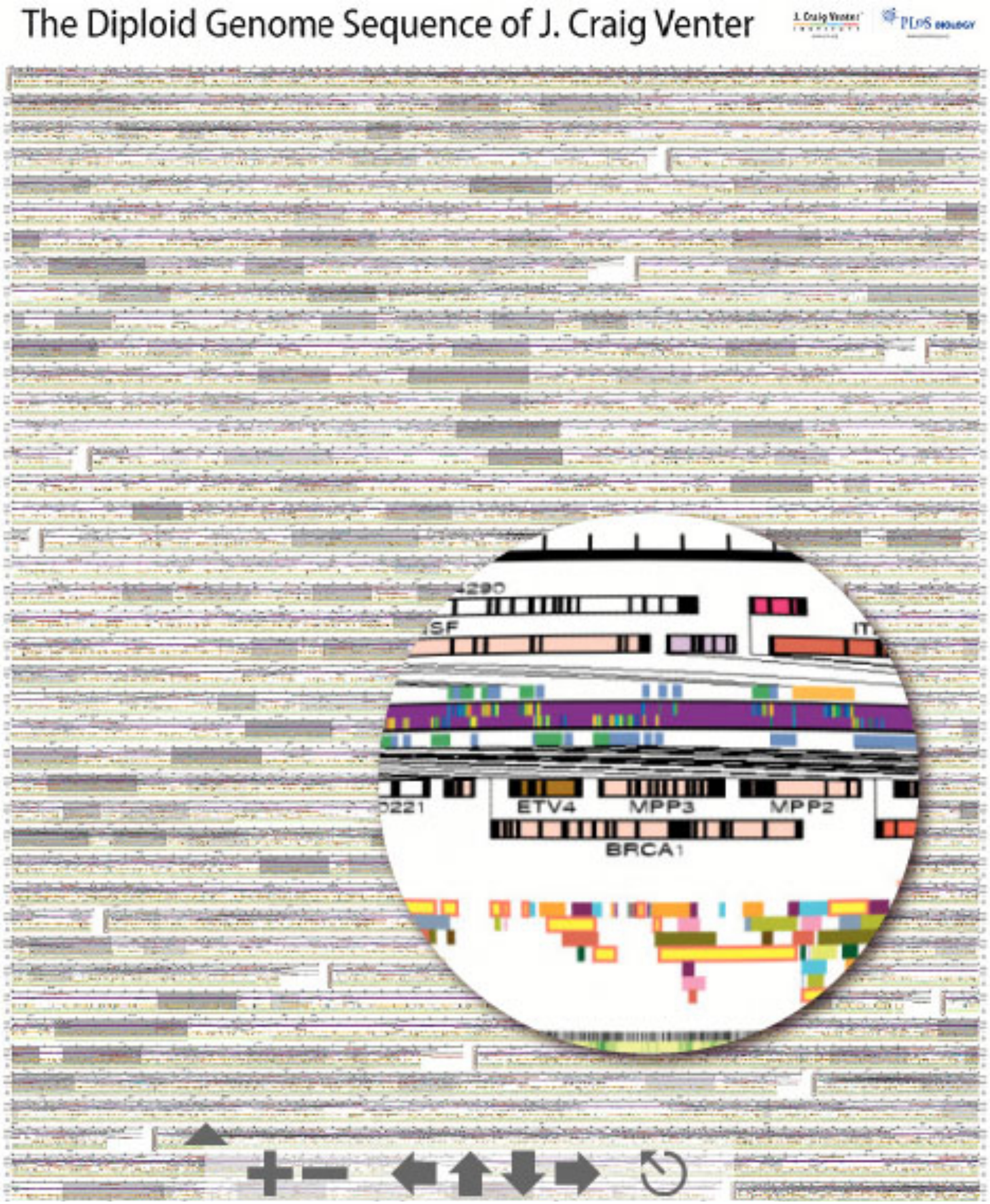
Company to sequence 1,000 human genomes in 2009 for $5,000 each

MOUNTAIN VIEW, CALIF. — Oct. 6, 2008 — Complete Genomics, Inc., a third-generation human genome sequencing company, today announced its formal launch as the world's first provider of large-scale human genome sequencing services.

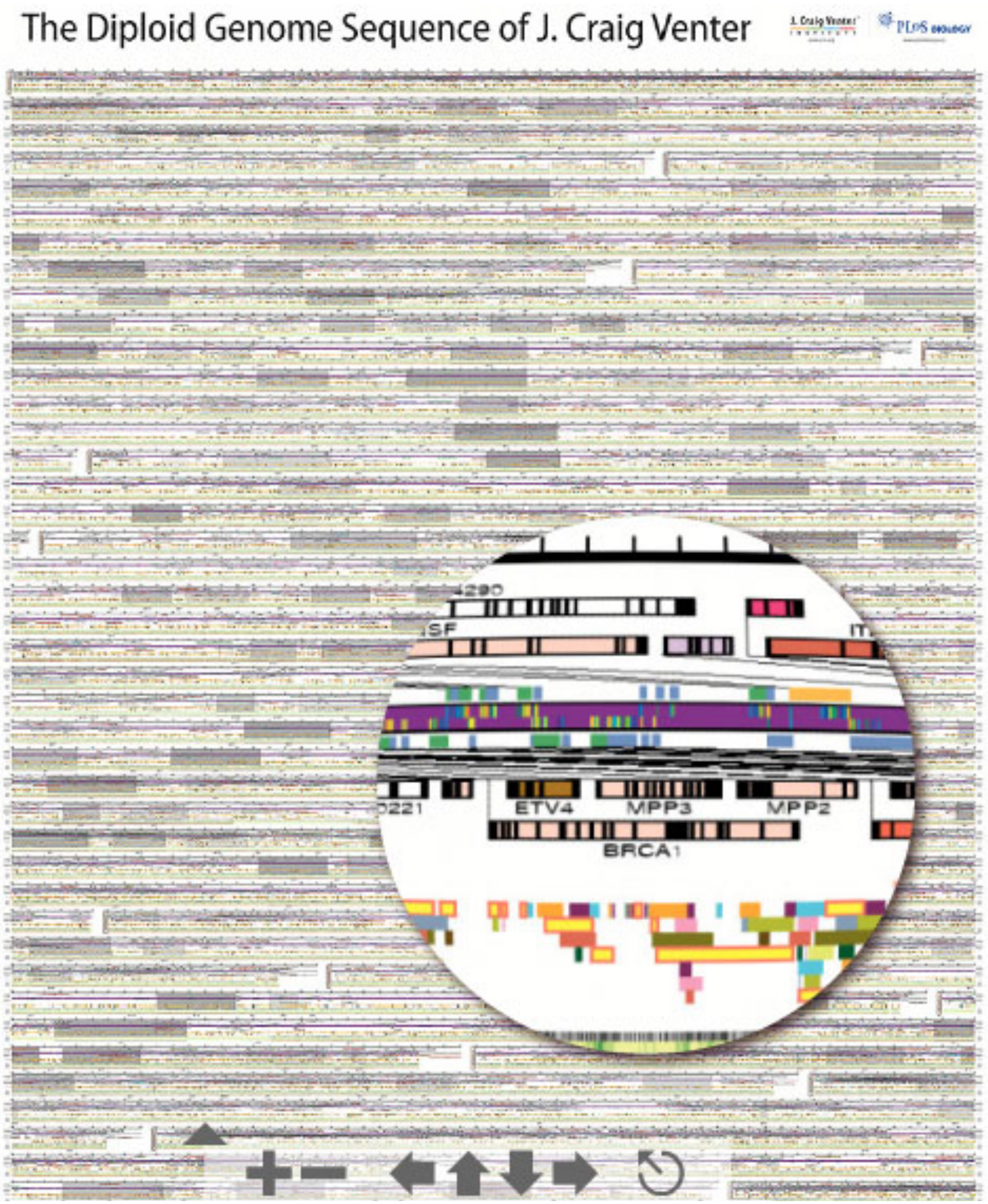
The company, which was established in March 2006, has been operating in "stealth mode" from a 32,000 sq. ft. facility near San Francisco, California. For the past 2.5 years, Complete Genomics has been reinventing the process of DNA sequencing based on pioneering technology invented by its founders and refined by a team of 100 employees with expertise in DNA engineering, molecular biology, instrumentation, semiconductors and high-performance computing. The convergence of these fields in Silicon Valley has allowed Complete Genomics to sequence its first complete human genome, providing proof-of-concept for its disruptive new technology and demonstrating the high accuracy and low cost of its approach.


The Diploid Genome Sequence of J. Craig Venter
J. Craig Venter Institute
PLoS Biology
4290
SF
IT
0221
ETV4
MPP3
MPP2
BRCA1

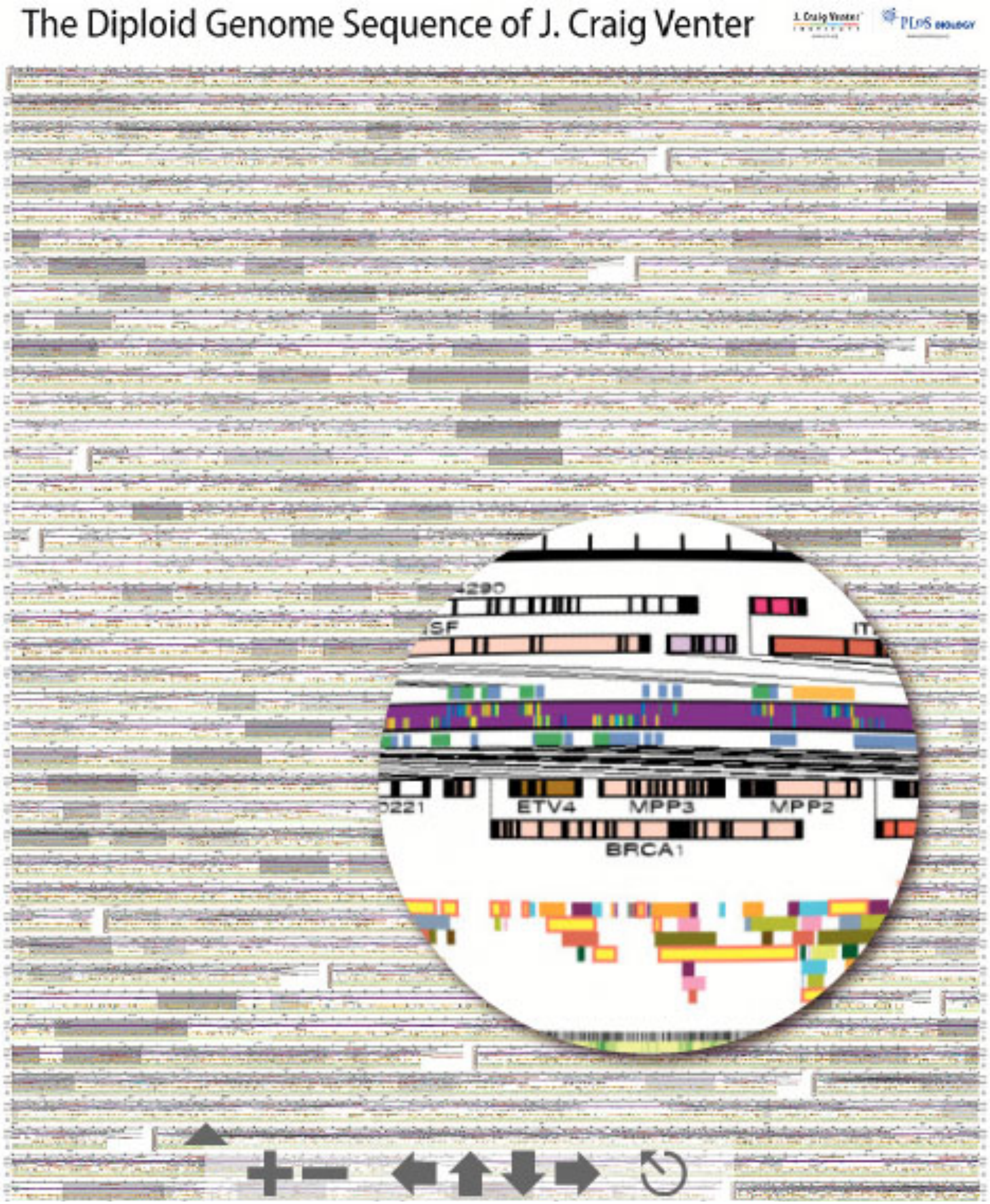
En pocos años dispondremos del genoma individual de cada uno de nosotros.



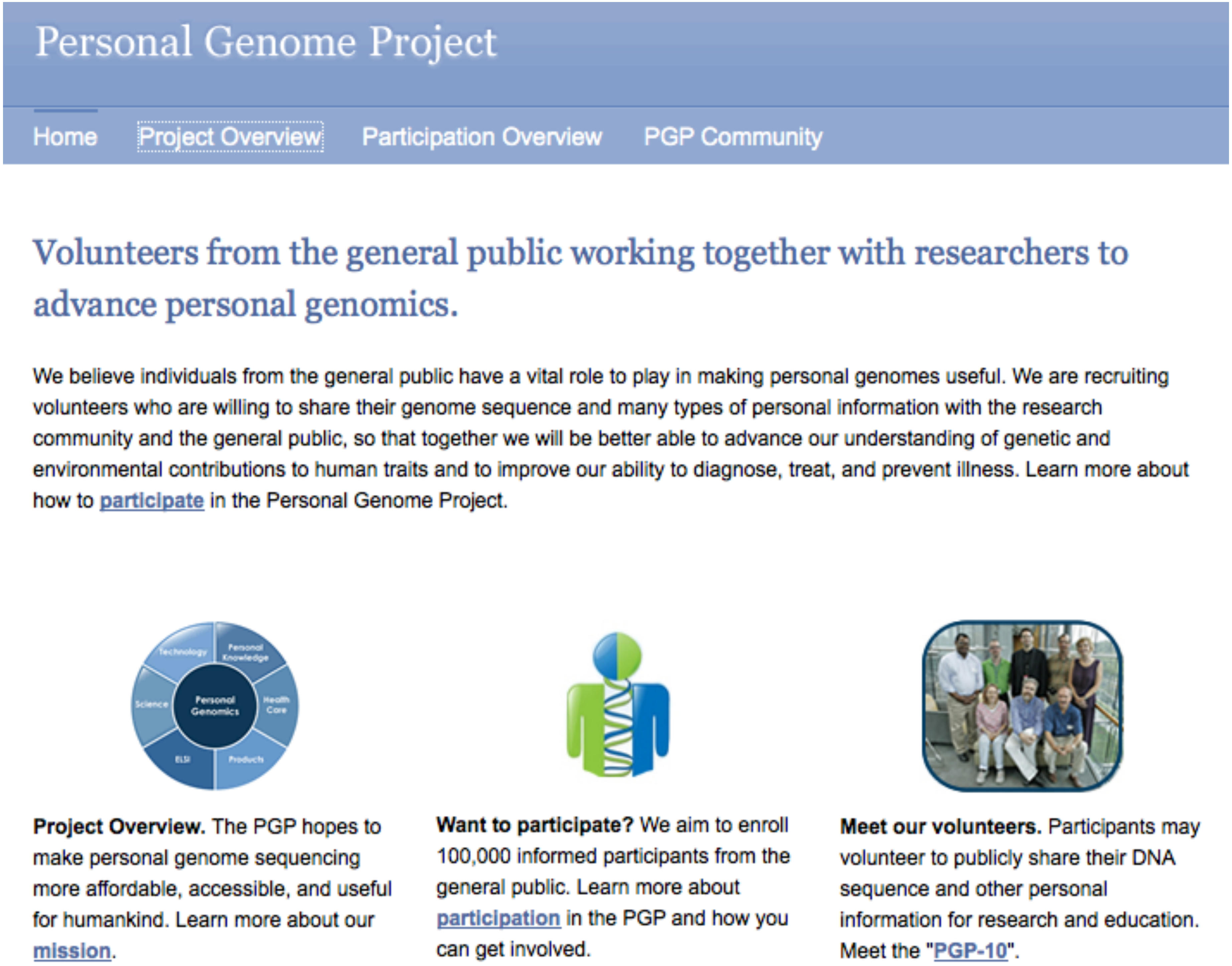
La información en nuestro genoma aislada de la información en el genoma de otros individuos, será de un valor limitado para cada uno de nosotros.



Sólo mediante la comparación de genomas de centenares de millones de individuos podremos empezar la comprender la relación entre nuestro genoma y nosotros (nuestras características biológicas)


The Diploid Genome Sequence of J. Craig Venter
J. Craig Venter Institute
PLoS Biology
4290
SF
IT
0221
ETV4
MPP3
MPP2
BRCA1

Deberíamos quizás asumir que nuestro genoma no es un patrimonio individual, sino un patrimonio común a la humanidad



Personal Genome Project

Volunteers from the general public working together with researchers to advance personal genomics.

We believe individuals from the general public have a vital role to play in making personal genomes useful. We are recruiting volunteers who are willing to share their genome sequence and many types of personal information with the research community and the general public, so that together we will be better able to advance our understanding of genetic and environmental contributions to human traits and to improve our ability to diagnose, treat, and prevent illness. Learn more about how to participate in the Personal Genome Project.

Project Overview. The PGP hopes to make personal genome sequencing more affordable, accessible, and useful for humankind. Learn more about our mission.

Want to participate? We aim to enroll 100,000 informed participants from the general public. Learn more about participation in the PGP and how you can get involved.

Meet our volunteers. Participants may volunteer to publicly share their DNA sequence and other personal information for research and education. Meet the "PGP-10".