INTRODUCTION TO GENOME SCIENCE

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

Scope of a genome sequencing project

Steps of genome analysis

Completed and ongoing genome projects

The triumphal march of genomics

The technological explosion: Genome sequences as commodity

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 Genomics
Landmarks 2001 and 2004
The publication of the draft and complete human sequence
DNA Doble Helix: the secret of life

1953
The human genome: The Holy Grail
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

BIG DATA SCIENCE
Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.
Genome Sequencing Centres (listed in order of total genomic sequence contributed, with a partial list of personnel. A full list of contributors at each centre is available as Supplementary Information.)

Whitehead Institute for Biomedical Research, Center for Genome Research: Eric S. Lander1, Lauren M. Linton1, Bruce Birren1, Chad Nusbaum1, Michael C. Zody1, Jennifer Baldwin1, Keri Devon, Ken Dewar, Michael Doyle1, William FitzHugh1, Roel Funke, Diane Gage, Katrina Harris, Andrew Heathford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim1, Jill P. Mesirov1, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramaniam, & Dudley Wyman


Biotechnology: André Rosenthal, Matthias Platter, Gerald Nyakatura, Stefan Taudien, & Andreas Rump

Beijing Genomics Institute/Human Genome Center: Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, & Jun Gu

Multigegbase Sequencing Center, The Institute for Systems Biology: Leroy Hood, Lee Rowen, Anup Madan & Shizhen Qin

Stanford Genome Technology Center: Ronald W. Davis, Nancy A. Federbergl, A. Pia Abola & Michael J. Proctor

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

University of Washington Genome Center: Maynard V. Olson, Rajinder Kaul & Christopher Raymond

Department of Molecular Biology, Keio University School of Medicine: Nobuyoshi Shimizu, Kazuhiko Kawasaki & Shinsei Minoshima

University of Texas Southwestern Medical Center at Dallas: Glen A. Evans, Maria Athanasiou & Roger Schultz

University of Oklahoma's Advanced Center for Genome Technology: Bruce A. Roe, Feng Chen & Huaqin Pan

Max Planck Institute for Molecular Genetics: Juliane Ramser, Hans Lehrach & Richard Reinhardt

Cold Spring Harbor Laboratory, Lita Annenberg Hazen Genome Center: W. Richard McCombie, Melissa de la Bastide & Neilay Dedhia

GBF—German Research Centre for Biotechnology: Helmut Bölcker, Klaus Hornischer & Gabriele Nordsiek

* Genome Analysis Group (listed in alphabetical order, also includes individuals listed under other headings): Dieter Anagnostopoulos, Jeffrey A. Bailey, Alex Bateman,
Omics
The Fundamental Question

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

Clarividencia

René Magritte

Curso Bioinformática Grau Genètica

Introduction to Genome Science

Fenotipo

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

New paradigm
Biology as an 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
From the new paradigm, biological systems are complex networks of myriad of pathways, many of them interconnected.
Biosynthesis pathways,

Metabolic pathway
Integrative Biology

Regulatory networks.
Kohlsted et al. 2010
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.

20th century biology

Reductionist approach (Experiments)

Biological System (Organism)

Construction blocks (Genes/Molecules)

21st century biology

Synthetic and interdisciplinary approach (Bioinformaticians, Biologists, Mathematicians, Biochemists, Physicist, Medical Doctors, …)
Scope of a genome sequencing project
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
A Genome Project is a complex scientific endeavor

The multiple components of a genome project

The car analogy of Elaine R. Mardis (Nature 2011 470:198-202)

- Sequence technology
- Raw material: DNA
- Reagents
- Robotics
- Bioinformatics
- Genome Center
- Biological understanding
  - Engine
  - Fuel
  - Spark
  - Mechanical parts
  - Direction
  - Driver
  - Destination
Major Genome-Sequencing Centers

- Baylor college of medicine human genome sequencing center
  - [http://www.hgsc.bcm.tmc.edu](http://www.hgsc.bcm.tmc.edu)
- Beijing Genomics Institute Bioinformatics
  - [http://www.genomics.cn](http://www.genomics.cn)
- The Broad Institute
  - [http://www.broadinstitute.org](http://www.broadinstitute.org)
- Genoscale
- US department energy joint genome institute
  - [http://www.jgi.doe.gov/](http://www.jgi.doe.gov/)
- Washington university genome center
  - [http://genome.wustl.edu/](http://genome.wustl.edu/)
- The Wellcome Trust Sanger Institute
  - [http://www.wellcome.ac.uk/](http://www.wellcome.ac.uk/)
- Centro Nacional de Análisis Genómico (CNAG, Spain)
  - [http://www.cnag.es/](http://www.cnag.es/)
Criteria for genome sequencing

- Non-human organism
  - Extensive prior biological knowledge (model species)
  - Human pathogen
  - Phylogenetic interest
  - Population genomics
  - Industrial application (agribusiness, pharmaceutical,...)

- Human
  - Medical interest
  - Human evolution
  - Association studies
  - Tumor – normal genome comparisons
  - Orphan (rare) diseases
Proposal for a genome project

- The consortium
- The white paper
- A white paper is a report detailing a chosen plan of action or summing up arguments for a subject of importance
- Examples of white papers in genomics
  - Bovine Sequence initiative
  - Proposal to Sequence a Drosophila Genetic Reference Panel
  - NHGRI White Papers for the 2008-2011 Planning Process
- The funding agency
  - Public
  - Private
  - Public-Private

Antonio Barbadilla
Sequencing technologies

- Sanger (the Gold Standard) ([http://www.wiley-vch.de/books/sample/3527320903_c01.pdf](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) (Lesson 2):

  **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 (Lesson 3)
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)
Assembly strategies

Hierarchical Shotgun Sequencing (a) vs Shotgun Sequencing (b)
Strategy for whole-genome shotgun sequencing assembly

Figure 13-6
Introduction to Genetic Analysis, Ninth Edition
© 2008 W.H. Freeman and Company
Draft DNA sequence: Sequence of a DNA with less accuracy than a finished sequence. In a draft sequence, some segments are missing or are in the wrong order or are oriented incorrectly.

- General overview
- Ongoing projects
- Mapping assembly to a reference genome (nucleotide population genomics)

Finished DNA sequence: A DNA sequence in which the bases are identified to an accuracy of no more than 1 error in 10,000 and are placed in the right order and orientation along a chromosome with almost no gaps.

- de-novo assembly
- Structural variation

Exome sequence: DNA sequence of the coding regions of the genome

Cheap and rapid screening for mutations in rare diseases
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 (\( P = 0.0001 \))

![Sequence alignment](alignment_image.png)
Steps of genome analysis

- DNA annotation and functional genomics
  - Gene number, exact locations, and functions
  - Predicted vs experimentally determined gene function
  - Noncoding DNA types, amount, distribution, information content, and functions
  - Gene regulation
  - DNA sequence organization
  - Chromosomal structure and organization
- Expression
  - Coordination of gene expression, protein synthesis, and post-translational events
- Proteins
  - Interaction of proteins in complex molecular machines
  - Proteomes (total protein content and function) in organisms
- Molecular Evolution
  - Evolutionary conservation among organisms
  - Protein conservation (structure and function)
- Genome variations and genotype-phenotype association studies
  - Nucleotide and structural variation
  - Correlation of SNPs (single-base DNA variations among individuals) with health and disease
  - Disease-susceptibility prediction based on gene sequence variation
  - Genes involved in complex traits and multigene diseases
- System Biology
  - Complex systems biology including microbial consortia useful for environmental restoration
  - Developmental genomics
Comparative genomics (Vista Browser)

Human -

Chimp

Mouse

Dog

Chicken

Frog

Zebra fish
Gene Ontology (GO)

The Gene Ontology project

major bioinformatics initiative for standardizing the representation of gene and gene product attributes across species and databases

**Molecular Function Ontology**
The tasks performed by individual gene products; examples are *carbohydrate binding* and *ATPase activity*

**Biological Process Ontology**
Broad biological goals, such as *mitosis* or *purine metabolism*, that are accomplished by ordered assemblies of molecular functions

**Cellular Component Ontology**
Subcellular structures, locations, and macromolecular complexes; examples include *nucleus*, *telomere*, and *origin recognition complex*
Gene Ontology (GO)

GO network of a typical plant gene
Genomics and reverse genetics

Direct genetics
From phenotype to the GENE

Reverse genetics
From GENE (genome) to phenotype
Completed and ongoing genome projects
Completed and ongoing genome projects

www.genomesonline.org

1. Register
Register your project information and Metadata in Genomes Online Database

2. Annotate
Annotate your microbial genome or metagenome with IMG/ER or IMG/MER

3. Publish
Publish your genome or metagenome in open access standards-supportive journal.
**Completed and ongoing genome projects**

**Complete Genome Projects**: 7396

- **Archaeal**: 234
- **Bacterial**: 6851
- **Eukaryal**: 311

**Finished**: 2649  
**Permanent Draft**: 4747
Welcome to the NCBI Entrez Genome Project database. This searchable database is a collection of complete and incomplete large-scale sequencing, assembly, annotation, and mapping projects for cellular organisms. The database is organized into organism-specific overviews that function as portals from which all projects in the database pertaining to that organism can be browsed and retrieved. Read more...

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

The technological explosion
Technology Advances Drive Science

Astronomy

Cell Biology

Radiology

Genomics
The technological explosion

For details, see http://genome.gov/sequencingcosts
DATA EXPLOSION

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

Terabases

Sequencers begin giving flurries of data
20 petabytes ~ 6 millones HG

http://www.ebi.ac.uk/
The technological explosion

~$1,000,000,000

~$1,000

Cost: 1 million fold lower!!!

“$1000 Genome”

Today
Genome Sequencing as a “Commodity”
The expanding scope of DNA sequencing

Road map of sequencing science
Genome science challenges
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.