BD & DS in ngs bioinformatics

jorge.amigo@usc.es
BD & DS in ngs bioinformatics

jorge.amigo@usc.es
DNA sequencing

50s
- 1953 Discovery of the structure of the DNA double helix.
- 1972 Development of recombinant DNA technology, which permits isolation of defined fragments of DNA; prior to this, the only accessible samples for sequencing were from bacteriophage or virus DNA.
- 1977 The first complete DNA genome to be sequenced is that of bacteriophage ϕX174.
- 1977 Allan Maxam and Walter Gilbert publish "DNA sequencing by chemical degradation." Frederick Sanger, independently, publishes "DNA sequencing with chain-terminating inhibitors."
- 1984 Medical Research Council scientists decipher the complete DNA sequence of the Epstein-Barr virus, 170 kb.
- 1986 Leroy E. Hood's laboratory at the California Institute of Technology and Smith announce the first semi-automated DNA sequencing machine.
- 1987 Applied Biosystems markets first automated sequencing machine, the model ABI 370.
- 1990 The U.S. National Institutes of Health (NIH) begins large-scale sequencing trials on Mycoplasma capricolum, Escherichia coli, Caenorhabditis elegans, and Saccharomyces cerevisiae (at US$0.75/base).
- 1991 Sequencing of human expressed sequence tags begins in Craig Venter's lab, an attempt to capture the coding fraction of the human genome.
- 1995 Craig Venter, Hamilton Smith, and colleagues at The Institute for Genomic Research (TIGR) publish the first complete genome of a free-living organism, the bacterium Haemophilus influenzae. The circular chromosome contains 1,830,137 bases and its publication in the journal Science marks the first use of whole-genome shotgun sequencing, eliminating the need for initial mapping efforts.
- 1996 Pál Nyrén and his student Mostafa Ronaghi at the Royal Institute of Technology in Stockholm publish their method of pyrosequencing.
- 1998 Phil Green and Brent Ewing of the University of Washington publish "phred" for sequencer data analysis.
- 2000 Lynx Therapeutics publishes and markets "MPSS" - a parallelized, adapter/ligation-mediated, bead-based sequencing technology, launching "next-generation" sequencing.
- 2001 A draft sequence of the human genome is published.
- 2004 454 Life Sciences markets a parallelized version of pyrosequencing. The first version of their machine reduced sequencing costs 6-fold compared to automated Sanger sequencing, and was the second of a new generation of sequencing technologies, after MPSS.
human genome project

1990 - started
2001 - first draft
2003 - "ended"
2006 - really ended
sanger sequencing

direct method: 1 sample = 1 sequence

1Gb ~ 1 year

1000b ~ $0.10

Human Genome => 3 years, $300,000
ngs sequencing
massive representation of random events

1Gb ~ 1 day

1000b ~ $0.02

Human Genome => 3 days, $60,000
**sanger sequencing**

Direct method: 1 sample = 1 sequence

- 1Gb ~ 1 year
- 1000b ~ $0.10

Human Genome => 3 years, $300,000

**ngs sequencing**

Massive representation of random events

- 1Gb ~ 1 day
- 1000b ~ $0.02

Human Genome => 3 days, $60,000
primary analysis
secondary and tertiary analysis
Figure 2. Cost of 1 MB of DNA sequencing. Decreasing cost of sequencing in the past 10 years compared with the expectation if it had followed Moore's law. Adapted from [11]. Cost was calculated in January of each year. MB, megabyte.
Figure 1. Contribution of different factors to the overall cost of a sequencing project across time. Left, the four-step process: (i) experimental design and sample collection, (ii) sequencing, (iii) data reduction and management, and (iv) downstream analysis. Right, the changes over time of relative impact of these four components of a sequencing experiment. BAM, Binary Sequence Alignment/Map; BED, Browser Extensible Data; CRAM, compression algorithm; MRF, Mapped Read Format; NGS, next-generation sequencing; TAR, transcriptionally active region; VCF, Variant Call Format.
Commercial (LifeScope)

Freely available (GATK, Picard tools,...)
Commercial (LifeScope)

mapping .bam

variant calling .gff
Freely available (GATK, Picard tools,...)

.bam *

variant calling

.vcf *
Commercial (LifeScope)  Freely available (GATK, Picard tools, ...)

mapping .bam  .bam *

variant calling .gff

variant calling .vcf *

annotation .txt
NGS results

alignments (GB)
and variants (MB)
BD & DS in ngs bioinformatics

jorge.amigo@usc.es
## Infrastructure Capacities

### Sequencers

<table>
<thead>
<tr>
<th>Number</th>
<th>Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>137</td>
<td>Illumina/HiSeq 2000</td>
</tr>
<tr>
<td>17</td>
<td>LifeTech/SOLiD 4</td>
</tr>
<tr>
<td>20</td>
<td>AB/3730xl + 110 MegaBACEs</td>
</tr>
<tr>
<td>2</td>
<td>Illumina iScan</td>
</tr>
</tbody>
</table>

### Multiple Supercomputing Centers

<table>
<thead>
<tr>
<th>Capacity</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>212</td>
<td>TB</td>
</tr>
<tr>
<td>37.2</td>
<td>TB</td>
</tr>
<tr>
<td>17</td>
<td>PB</td>
</tr>
</tbody>
</table>

## Data Production

- 6 Tb / day
- ~2000X of human genome / day
1000 Plant & Animal (1000P&A) Reference Genomes Project

The BGI-initiated project focuses on “1,000 economically and scientifically important plant/animal species.” Fifty have been completed, another 100 are in progress.
Genome 10K Project (G10K)

The G10K Project will sequence some 16,203 vertebrate genomes, a "genomic zoo" representing at least one member of each vertebrate genus. The first 101 species have already been announced.
Ten Thousand Microbial (10K M) Genomes Project

This BGI-led project is sequencing microbes from habitats as diverse as earth, air, glaciers, and hot springs. Their goal is the development of a genomic encyclopedia of microorganisms in China. Over 1,200 have been completed to date.
5,000 Insect (i5K) and Other Arthropod Genome Initiative

The Arthropod Genomic Consortium will target 5,000 insects with agricultural, medical, or research significance. So far, 76 have been proposed.
Building a database of extensive genetic information
Ensuring food security
Promoting medical applications.

- Improving ecological conservation
- Developing new forms of energy
Exploring the microbial ecosystem

Accelerating microbiological application development for human health, circular economies and ecology
Million Human Genomes Project

- Constructing a detailed map of human genetic variation.
- Exploring the origin and evolution of human.
- Promoting the genomic research for disease to accelerate application in health field.
Number of Bases Submitted To The EBI Short Read Archive

472,867,592,823,406

>100,000x human genome

from 1 sample to n samples
>100,000x human genome

from 1 sample to n samples
Creating a Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data

June 3, 2013

An initial draft of this White Paper was prepared for the January 26th meeting, and has since been revised substantially based on discussions at and since the meeting. A list of contributors and participants is provided at the end of this document.
June 3, 2013

Broad Institute, MIT, Harvard University, Stanford University, MD Anderson Cancer Center, Wellcome Trust, NIH, BGI,...

Amazon Web Services, Google, Microsoft
BD & DS in ngs bioinformatics

jorge.amigo@usc.es