Microbial Oceanomics using High-Throughput DNA Sequencing

Ramiro Logares

Institute of Marine Sciences, CSIC, Barcelona

9th RES Users'Conference – 23 September 2015
Importance of microbes in the sunlit ocean

- Phytoplankton: 50% primary production of the Earth (Field et al. 1998)
- Microplankton crucial for the marine food chain
- Biogeochemical cycling
- Large phylogenetic and metabolic diversity

Chlorophyll concentration by SeaWiFS
September 1997 – August 2000
Microbial phylogenetic diversity

Brock “Biology of Microorganisms”
Prokaryotes

Multicellular

Protists

Viruses

Body Size [µm]

Source: C. de Vargas
Our typical questions

Who is in there?

How are they doing it?

What are they doing?
High throughput DNA sequencing technologies: a new lens for viewing the microbial world
DNA sequencing: an accelerating revolution

<table>
<thead>
<tr>
<th>Year</th>
<th>Landmark</th>
</tr>
</thead>
<tbody>
<tr>
<td>1953</td>
<td>Discovery of the double helix</td>
</tr>
<tr>
<td>1977</td>
<td>First DNA genome (bacteriophage)</td>
</tr>
<tr>
<td>1977</td>
<td>F. Sanger publishes “chain-terminator” method for DNA sequencing</td>
</tr>
<tr>
<td>1987</td>
<td>First commercial sequencing machine (ABI 370)</td>
</tr>
<tr>
<td>1995</td>
<td>First genome of a free-living organism (bacteria). WGSS initial use</td>
</tr>
<tr>
<td>1996</td>
<td>Nygren &amp; Ronaghi publish “pyrosequencing”</td>
</tr>
<tr>
<td>2001</td>
<td>First draft human genome (3 billion US$)</td>
</tr>
<tr>
<td>2004</td>
<td>454 pyrosequencing commercialized</td>
</tr>
<tr>
<td>2009</td>
<td>Illumina 50 K US$ per human genome</td>
</tr>
<tr>
<td>2010</td>
<td>Single molecule real time sequencing (SMRT) commercialized</td>
</tr>
<tr>
<td>2011</td>
<td>Human genome for 8000 US$. About 30 K human genomes sequenced</td>
</tr>
<tr>
<td>2015</td>
<td>1,000 US$ - Human genome (Illumina X10; 18,000 per year-machine)</td>
</tr>
<tr>
<td>Year</td>
<td>Landmark</td>
</tr>
<tr>
<td>-------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1953</td>
<td>Discovery of the double helix</td>
</tr>
<tr>
<td>1977</td>
<td>First DNA genome (bacteriophage)</td>
</tr>
<tr>
<td>1977</td>
<td>F. Sanger publishes “chain-terminator” method for DNA sequencing</td>
</tr>
<tr>
<td>1992</td>
<td>Sequencing machine (370)</td>
</tr>
<tr>
<td>1995</td>
<td>WGSS initial use</td>
</tr>
<tr>
<td>1995</td>
<td>Establish “pyrosequencing”</td>
</tr>
<tr>
<td>1995</td>
<td>454 Life Sciences (later 2010)</td>
</tr>
<tr>
<td>1998</td>
<td>Human genome</td>
</tr>
<tr>
<td>2000</td>
<td>Whole Genome shotgun sequencing (WGS)</td>
</tr>
<tr>
<td>2007</td>
<td>Sequencing by synthesis (SBS)</td>
</tr>
<tr>
<td>2008</td>
<td>HM4 sequencing (SMRT) commercialized</td>
</tr>
<tr>
<td>2011</td>
<td>Human genome for $1000 US$. About 30 K human genomes sequenced</td>
</tr>
</tbody>
</table>

62 years
Sequencing platforms evolution

1st Generation
Still used for smaller projects or when high quality is needed
Quality reference

2nd Generation
Widely used in most sequencing projects

3rd Generation
Not widely used yet, some devices still not in the market
Next Generation Genomics: World Map of High-throughput Sequencers

http://omicsmaps.com/hts/centres/imppc/
Amount of HTS DNA data produced now in the world:

- ≈ 7,389 functional HTS machines
- ≈ 35 x 10^{15} bases / year == 35 PETAbases
- 250,000 human genomes per year
Computing power
Moore’s Law

Computer power tends to double every two years.

Kryder’s Law

Storage capacity doubles annually.
DNA Sequencing Is Now Improving Faster Than Moore's Law!

A "worldwide genomics revolution" is upon us.

The genomics industry marked a new milestone on Tuesday. As Forbes' Matthew Herper reported in three separate posts and nearly 100 related Tweets, the two leading manufacturers of DNA sequencing instruments announced almost simultaneously at an investors' conference that they would introduce new machines this year capable of sequencing an entire human genome in a single day. Life Technologies said its forthcoming Ion Proton machine, which processes DNA on a semiconductor chip, will do it for a cost of $1,000 per genome.

These advances are not just big news for biotech and medicine, but exciting for all Technologists. They're proof that the pace of advances in genome sequencing technology has exceeded Moore's Law. The speed of genome sequencing has far better than doubled every two years since 2003, when the

Forbes Magazine, 2012

Worldwide sequencing capacity is growing at about 2-3 times per year
- Only one HiSeq2500 produces about 3 TeraBytes of data per month

- Data processing costs should be considered
  - Electricity
  - Costs for data-admin, and reparation

- Amortization (value decrease) of equipment (3yrs CSIC)

- Data storage cost is not trivial

- What to do with used data? And backups? (maybe cheaper re-sequencing than storing?)
Data processing and computation
## Minimum needed computer power

<table>
<thead>
<tr>
<th>Generation</th>
<th>SANGER</th>
<th>454 Roche</th>
<th>Illumina HiSeq</th>
<th>PacBio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cores</td>
<td>1-2</td>
<td>16-32</td>
<td>&gt;64 (128)</td>
<td>&gt;16 (32)</td>
</tr>
<tr>
<td>Mem</td>
<td>1-4 GB</td>
<td>32-64</td>
<td>&gt;64 (128)</td>
<td>&gt;32 (64)</td>
</tr>
<tr>
<td>Disk</td>
<td>0.2Tb</td>
<td>&gt;2 Tb</td>
<td>&gt; 10 Tb</td>
<td>&gt;1Tb</td>
</tr>
</tbody>
</table>
Welcome to MareNostrum III

- All home directories are in GPFS and quotas are enabled
- Applications are located at /apps
- To change password, please login from your home directory to:
  dl01.bsc.es
- Active Archive and transfer management is available at:
  dt01.bsc.es
- For further information read MareNostrum III documentation:
  http://www.bsc.es/support/MareNostrum3

BSC SUPPORT COMMANDS:

See `man bsc` for more information

Location:

CSIC CIBER-BBVA

MareNostrum III (Barcelona)
Question in microbial community ecology

1. Who is in there? (METAGENETICS)
2. What can they do? (METAGENOMICS)
3. What are they doing? (METATRANSCRIPTOMICS)

Sequencing effort
Community DNA / RNA

Targeted amplification & HTS

Shotgun HTS (mRNA enrichment)

Amplicon library

OR

Taxonomic markers 16/18S rDNA

Extraction

Metagenome / Metatranscriptome

Metabolic potential or actual transcription of the community

METAGENETICS

METAGENOMICS / METATRANSCRIPTOMICS

Functional Gene
Alternatively: Single Cell Genomics
Some results using MareNostrum
Malaspina 2010 expedition
1) Deep-MalaspinOmics (4,000m samples)

Sample characteristics: 120 L filtered through two fractions:
✓ 0.2-0.8 μm (free-living bacteria and archaea)
✓ 0.8-20 μm (protists and attached bacteria)

62 metagenomes,
55 metatranscriptomes
Corresp. 18/16S-rRNA i-454Tags
2) MalaspinOmics (0 to 4,000m samples)

Sample characteristics:

- 0.2-3 μm (picoplankton)
- 3-20 μm (nanoplankton and attached bacteria)

Vertical profiles: 7 depths from surface to 4000m

- 88 Metagenomes
- 15 Metatranscriptomes
-(Corresp. 18/16S-rRNA iTags )

picoplankton

13 profiles extracted: large dots
11 profiles sequenced

nanoplankton

2 profiles sequenced
Metagenomes

metaG (HiSeq)

Assembly

Binning

Fragment Recruitment Analysis

Gene Prediction

Gene Annotation

Aim: Genomic characterization of dominant protists

(single / multiple metaG)
Metagenomes
4,000m samples

Assembly per sample ➔ eukaryotic few contigs

…but, we knew the same OTUs were found in several samples

Co-Assembling all samples together (58 metagenomes) should generate longer contigs

MareNostrum Supercomputer
2,048 processors with Ray assembler

<table>
<thead>
<tr>
<th></th>
<th>One Sample</th>
<th>Sum of All Samples</th>
<th>Assembly All Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contigs &gt; 2Kb</td>
<td>1,055</td>
<td>102,705</td>
<td>152,175</td>
</tr>
<tr>
<td>Mean Coverage (&gt; 2Kb)</td>
<td>24.87</td>
<td>-</td>
<td>139.8</td>
</tr>
<tr>
<td>Contigs &gt; 10Kb</td>
<td>21</td>
<td>5,823</td>
<td>23,086</td>
</tr>
<tr>
<td>Largest Contig (bp)</td>
<td>40,779</td>
<td>207,037</td>
<td>925,604</td>
</tr>
<tr>
<td>% assembled reads (&gt; 2Kb)</td>
<td>~5%</td>
<td>-</td>
<td>~40%</td>
</tr>
</tbody>
</table>
| Largest Scaffold (bp)    | 40,779     | -                  | 1,275,015
Co-assembly of 1,500-4,000m samples:

- 58 | 4,000m (5Gb each)
- 29 | 1,500-4,000m (20-40 Gb each)

- Ray assembly with 2,048 threads @MN (18hs)

<table>
<thead>
<tr>
<th></th>
<th>4,000m</th>
<th>1,500 - 4,000m</th>
</tr>
</thead>
<tbody>
<tr>
<td># Contigs &gt; 2kb</td>
<td>152,175</td>
<td>339,898</td>
</tr>
<tr>
<td>Largest contig (bp)</td>
<td>925,604</td>
<td>1,119,237</td>
</tr>
</tbody>
</table>
Metagenomes: deep ocean fungi

4,000m samples

FRA of 2.6Mb of what seems to be a widespread fungus in the deep ocean

Pyrotag data

Recruitment of metagenomic reads by the 35 fungal contigs
Analyzed Pico–euk SAGs from TARA

Stations
23: Mediterranean (Adriatic sea)
41: Indian Ocean (Arabic sea)

Included here: 45 SAGs
Being analyzed: 75 SAGs

A large collection of SAGs is being generated. So far, 903 SAGs were identified, with 568 affiliating to Stramenopiles
# Co-assembly stats

<table>
<thead>
<tr>
<th></th>
<th>% Genome recovery (CEGMA)</th>
<th>Assembly size (Mb; contigs &gt;1,000bp)</th>
<th>Contigs (&gt;1,000bp)</th>
<th>Max. contig</th>
<th>N50 (&gt;1,000bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Co-assembly 14 SAGs MAST-4 clade A (SPAdes)</strong></td>
<td>89.1</td>
<td>47.5</td>
<td>14,564</td>
<td>57,905</td>
<td>4,563</td>
</tr>
<tr>
<td><strong>Co-assembly 14 SAGs MAST-4 clade A (MegaHit)</strong></td>
<td>80.6</td>
<td>42.5</td>
<td>15,158</td>
<td>51,080</td>
<td>3,475</td>
</tr>
<tr>
<td>*MAST-4 clade A single SAG assembly (mean</td>
<td>SD)*</td>
<td>20.6</td>
<td>10.2</td>
<td>9.1</td>
<td>4.5</td>
</tr>
<tr>
<td><strong>Co-assembly 9 SAGs MAST-4 clade E (SPAdes)</strong></td>
<td>68.5</td>
<td>32.3</td>
<td>5,677</td>
<td>104,912</td>
<td>9,991</td>
</tr>
<tr>
<td>*MAST-4 clade E single SAG assembly (mean</td>
<td>SD)*</td>
<td>14.3</td>
<td>5.5</td>
<td>6.2</td>
<td>2.4</td>
</tr>
</tbody>
</table>
Continuing analyses with the co-assembly

1) Gene prediction [Augustus]

2) Annotation (KEGG, KOG, Pfam, eggNOG, OMRGC, MMETSP)
General metabolic pathways

In Red: mapped MAST4 pathways/functions according to KO
Large phylogenetic analyses

- Deep ocean Bacteria [Malaspina]
- 3,500 sequences (16S rRNA)
- 60 samples
- About 1,000 processes

- Stramenopiles (protists)
- 3,835 sequences (>1,100bp) [18S]
- About 1,000 processes

G. Salazar
Simulations

Primary production vs. richness

Vallina et al., 2014. Nat. Comm
Summary of results with RES support since 2011

- 11 published papers
- 2 in revision

Contributing mostly
- Metagenomics
- Genomics
- Phylogenetics
- Modelling
Conclusions: Microbial ecology

- Massive amounts of DNA data need powerful computers as well as programs that can deal with them.
- Future developments require further integration with high-performance computers and quantitative methods.
- Analysis of large datasets will likely unveil patterns of genomic functioning as well as interactions between marine microbes.
microbial Malaspina@ICM
Single Molecule Real Time (SMRT) sequencing

Average read length: 4,200-8,500 bp (longest read 30Kbp)
- P4/C2: shorter reads, higher accuracy
- P5/C3: longer reads, lower accuracy
- 200-300 Mbp from each SMRT cell for 15-20kb insert size libraries
- 100-150 Mbp for >20Kbp libraries
- No multiplexing in genomics libraries (multiplexing in amplicons)
- Library preparation 400-1200€
- About 350 € per SMRT cell
- Signal: colors
What if you are not interested in the whole community but in one species?

If you have a clonal culture:

Genomics and/or RNA-Seq

If you don’t have a clonal culture:

Single-Cell genomics
- 25x10^6 reads
- 15Gb/run
- 2x300bp

- 400x10^6 reads
- 120Gb/run
- 2x150bp

- 4x10^9 reads
- 1000 Gb/run
- 2x125bp

Signal: colors
1Gb=1x10^9 bases
El genoma de los mil dólares desborda a los científicos

Una empresa de EEUU crea una máquina capaz de secuenciar todo el ADN individuo por 740 euros en un solo día, un récord persiguido desde hace años. España, la falta de expertos en genética relativiza el logro.

Por unos 750 euros, se puede comprar un buen ordenador portátil, viajar cinco días a las playas de Miami, y, desde hace unos días, secuenciar un genoma humano. ¿Qué hace el mundo? ¿Qué hay que hacer para que España no se quede atrás en esta carrera hacia las mini-sequencers genéticas?

MinION USB stick gene sequencer finally comes to market

By John Hewitt on September 19, 2014 at 12:10 pm

The (Unmet) Potential Value of Cancer Genome Testing

As the cost of DNA sequencing plummets, the possibility of testing all cancer patients' tumor genomes is becoming a reality. For just $1000 or so, a doctor might submit most any malignant specimen for a complete genetic work-up. The sample might be a core
Oceanic microbial community

Includes all species occurring at a particular site and their abundances
Multiple metabolisms

Brock "Biology of Microorganisms"
## Comparison of platforms

<table>
<thead>
<tr>
<th>Platform</th>
<th>Run time</th>
<th>Mreads/run</th>
<th>Read length</th>
<th>Mb/run</th>
<th>€/Mb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sanger (3730xl)</td>
<td>2h</td>
<td>0.000096</td>
<td>400-900</td>
<td>0.06</td>
<td>1500</td>
</tr>
<tr>
<td>454 FLX Titanium</td>
<td>10h</td>
<td>1</td>
<td>400</td>
<td>400</td>
<td>15</td>
</tr>
<tr>
<td>454 FLX+</td>
<td>18-20 h</td>
<td>1</td>
<td>700</td>
<td>900</td>
<td>9</td>
</tr>
<tr>
<td>Ion Torrent</td>
<td>2h</td>
<td>80</td>
<td>400</td>
<td>32,000</td>
<td>1</td>
</tr>
<tr>
<td>PacBio</td>
<td>0.5-2h</td>
<td>0.005</td>
<td>4-8 K</td>
<td>300 (SMRT)</td>
<td>0.33-1</td>
</tr>
<tr>
<td>Illumina MiSeq</td>
<td>55h</td>
<td>25</td>
<td>2x300</td>
<td>15,000</td>
<td>0.1</td>
</tr>
<tr>
<td>Illumina GAIIx</td>
<td>14 days</td>
<td>320</td>
<td>2x150</td>
<td>96,000</td>
<td>0.12</td>
</tr>
<tr>
<td>Illumina HiSeq2500</td>
<td>1-11 days</td>
<td>4000</td>
<td>2x125</td>
<td>1,000,000</td>
<td>0.05</td>
</tr>
</tbody>
</table>

1Gb=1x10⁹ bases

Endocytosis

In Red: mapped MAST4 pathways/functions according to KO
GPU (graphics processing unit) computing

Serial part of an application runs on a CPU and the computationally-intensive part runs on a GPU

GPU Pipeline for HTS sequencing
Centro de Investigación Príncipe Felipe
http://docs.bioinfo.cipf.es/projects/ngs-gpu-pipeline/wiki
Cloud Computing...

- Purchase needed computer power
- Scalable (few to thousands of processors)
- No maintenance costs

- GALAXY (g2.bx.psu.edu)
- N3phele: HTS analyses at the EC2 with QIIME
REVEAL HIDDEN VARIATION

C1™ Single-Cell Targeted DNA Sequencing

The fastest and easiest workflow to discover and screen mutations in target genes from individual cells.
Molecular biology + computers + stats

The next 20 years of genome research

M. Schatz (2015)

http://biorxiv.org/content/early/2015/06/02/020289
General SAG construction strategy