Techniques for genomic analysis

Mario Cáceres
ICREA Research Professor
Institut de Biotecnologia i de Biomedicina
Universitat Autònoma de Barcelona

The genomics revolution

In the last years advances in genomics have generated an extraordinary amount of information:

• Whole genome sequences of tens of species and individuals (and more to come!)

• Gene expression profiling in multiple conditions, tissues, individuals and species

• Mapping of functional regions in the genome

Best time for genomic studies!!!
Sequenced genomes

Over the past years the genomes of some of the most important model organisms have been sequenced:

Main technical developments

This has been possible by a parallel increase in the power and throughput of genomic techniques:

- DNA microarrays (DNA chips)
- Next-generation sequencing methods (ultrasequencing)
Overview

1. DNA microarrays
2. 2nd generation sequencing systems
3. 3rd generation sequencing systems
4. Examples of some applications

Common features of microarrays

- Parallelism/high-throughput  
  (thousands of genes analyzed simultaneously)
- Miniaturization  
  (small feature and chip size)
- Automation  
  (chip manufacture, processing and analysis)
- Principle similar to Southern/Northern  
  (based on nucleic-acid hybridization and base pairing)
  - Probes fixed on a glass slide
  - Labeling of target samples
Increase of microarrays use

![Graph showing the increase of microarrays use from 1995 to 2004 with a sharp rise in publications for gene expression and microarrays]

First description of oligonucleotide arrays (Lockhart et al. 1996)
First description of cDNA microarrays (Schena et al. 1995)

Classes of microarrays

- Custom/spotted/two-color microarrays (cDNAs, BACs)
- High-density oligonucleotide arrays (GeneChip, Affymetrix)
- Long oligonucleotide microarrays
  - Agilent (25-60 bases)
  - Illumina (50 bases)
  - Nimblegen (50-75 bases)
Custom microarrays technology

1. Isolation or PCR amplification of DNA fragments (~1-200 Kb) of the genes or regions of interest

2. Spotting of DNAs at high density onto a glass microscopy slide (5000-10000 spots per slide) and cross-linking to the glass surface

3. Two independent mRNA or DNA samples are fluorescently labeled with Cy3 (green) or Cy5 (red)

4. The two labeled populations are combined in equal amounts and hybridized to the array

5. A laser scans the slide and calculates the ratio of fluorescence intensities between the two samples
Affymetrix oligonucleotide arrays
Affymetrix oligonucleotide arrays

The array elements are a series of 25-mer oligos designed from known sequence and synthesized directly on the surface.

The entire array is formed by >500,000 cells, each containing a different oligo.

Affymetrix GeneChip array design

1. Each gene is represented by a unique set of different probe pairs (11-20).

2. Each probe pair has two almost identical oligos, with one base change.

<table>
<thead>
<tr>
<th>Probe pair</th>
<th>PM (Perfect match: ref. seq.)</th>
<th>MM (Mismatch: background and cross-hybridization control)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTAACCCAGTCTCCTGAGGATACAC</td>
<td>TTAACCCAGTCTCCTGAGGATACAC</td>
<td></td>
</tr>
</tbody>
</table>

3. Gene quantification is the average difference between the fluorescence intensity in the PM and the MM over all the probe pairs.
Agilent DNA microarray platform

- Single long oligonucleotide probes (25-60 bases)
- High density (244,000 features per chip)
- In situ synthesis printing process (highly consistent SurePrint technology)
- Supports 1 or 2 samples hybridization
- Great design flexibility (rapid design of custom arrays using eArray)
- Possibility of multiplexing (as many as 8 arrays in one slide)

Comparison of microarray platforms

<table>
<thead>
<tr>
<th></th>
<th>Custom microarrays</th>
<th>Oligonucleotide arrays</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probes</td>
<td>cDNAs/BACs</td>
<td>25-60 mer oligos</td>
</tr>
<tr>
<td>Gene density</td>
<td>5,000-10,000</td>
<td>10,000-40,000</td>
</tr>
<tr>
<td>Representation</td>
<td>moderate</td>
<td>high</td>
</tr>
<tr>
<td>Hybridization</td>
<td>two color</td>
<td>single color</td>
</tr>
<tr>
<td>Quantification</td>
<td>relative</td>
<td>~absolute</td>
</tr>
<tr>
<td>Variability</td>
<td>higher</td>
<td>lower</td>
</tr>
<tr>
<td>Replicability</td>
<td>lower</td>
<td>higher</td>
</tr>
<tr>
<td>Specificity</td>
<td>moderate</td>
<td>high</td>
</tr>
<tr>
<td>Comparison design</td>
<td>a priori</td>
<td>a posteriori</td>
</tr>
<tr>
<td>Organisms</td>
<td>any</td>
<td>model (sequenced)</td>
</tr>
<tr>
<td>Cost</td>
<td>lower (&lt;$100)</td>
<td>higher ($500)</td>
</tr>
</tbody>
</table>
Main applications of microarrays

- Measuring transcript abundance (expression arrays)
- Identification of transcribed sequences (tiling arrays)
- Analysis of alternative splicing (exon or exon-junction arrays)
- Re-sequencing and SNP genotyping (SNP oligonucleotide arrays and tiling arrays)
- Estimating DNA copy number (CGH on BAC or oligonucleotide arrays)
- Identifying protein binding sites (ChIP-chip on tiling or oligonucleotide arrays)
- Detection of epigenetic modifications (ChIP-chip on tiling or oligonucleotide arrays)
Classic genome sequencing methods

Sanger sequencing:
- Long reads (500-1000 bp)
- Low throughput (96 reactions/run)

Whole Genome Shotgun Sequencing Method

Large sequencing facilities and sequencing centers
The quest for the $1000 genome

Next-generation sequencing instruments can generate as much data in one day as several hundred Sanger-type DNA capillary sequencers, but are operated by a single person

Next generation sequencing systems

454 (Roche)  
Solexa (Illumina)  
SOLiD (Applied Biosystems)
**454 Genome Sequencer FLX (Roche)**

- Chemistry based on pirosequencing
- Sample amplified by emulsion PCR
- Read length 250-500 bp
- >1 million reads per run
- 400-600 Mb of sequence
- ~10 hours run

http://www.454.com/

---

**454 sequencing method**

1. DNA fragmentation and adapters ligation

2. Emulsion PCR within water-in-oil microdroplets
454 sequencing method

Solexa Genome Analyzer (Illumina)

- Chemistry based on reversible terminators
- Sample amplified by solid-phase amplification
- Read length 30-75 bp
- >100 million reads per run
- ~10 Gb of sequence
- 4-8 days run

http://www.solexa.com/
Solexa sequencing method

1. DNA fragmentation and adapters ligation

2. Solid-phase amplification and cluster generation by bridge PCR

3. Flowing of fluorescent reversible terminator dNTPs and incorporation of a single base per cycle

4. Read identity of each base of a cluster from sequencial images
SOLiD system (Applied Biosystems)

- Chemistry based on sequencing by ligation
- Sample amplified by emulsion PCR
- Read length 50-100 bp
- 100-500 million reads per run
- 50-100 Gb of sequence
- 4-8 days run

http://www3.appliedbiosystems.com
**SOLiD sequencing**

<table>
<thead>
<tr>
<th>Read length</th>
<th>Bases per machine run</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 bp</td>
<td>10 bp</td>
</tr>
<tr>
<td>1,000 bp</td>
<td>100 bp</td>
</tr>
<tr>
<td>100 bp</td>
<td>100 bp</td>
</tr>
<tr>
<td>10 Mb</td>
<td>100 bp</td>
</tr>
<tr>
<td>1 Gb</td>
<td>100 bp</td>
</tr>
<tr>
<td>100 Mb</td>
<td>100 bp</td>
</tr>
<tr>
<td>10 Gb</td>
<td>100 bp</td>
</tr>
<tr>
<td>1 Mb</td>
<td>100 bp</td>
</tr>
<tr>
<td>100 Gb</td>
<td>100 bp</td>
</tr>
<tr>
<td>1 Tl</td>
<td>100 bp</td>
</tr>
</tbody>
</table>

**Comparison of sequencing platforms**

**Predicted performance 2009-2010**

- **AB SOLiDv3**: 120Gb, 100 bp reads
- **Illumina GAII**: 90Gb, 175bp reads
- **454 GS FLX Titanium**: 0.4-0.6 Gb, 100-400 bp reads
- **ABI capillary sequencer**: 0.04-0.08 Mb, 450-800 bp reads

**Decoding**

- Possible dinucleotides
  - TA, AC, AA, CA, GC, CA, CC, TC
- Decoded sequence
  - A, T, G, C

**Double interrogation**

With 2 base encoding each base is defined twice.

**Possible dinucleotides encoded by each color**

- A
- C
- G
- T

**Template sequence**

- TA, AC, AA, CA, GC, CA, CC, TC

**Color space sequence**

- Decoded sequence
  - A, T, G, C

**Base space sequence**

- ABI capillary sequencer
- 0.04-0.08 Mb, 450-800 bp reads
- Illumina GAII
- 90Gb, 175bp reads
- 454 GS FLX Titanium
- 0.4-0.6 Gb, 100-400 bp reads
- AB SOLiDv3
- 120Gb, 100 bp reads
## Comparison of sequencing platforms

### Quicker, Smaller, Cheaper

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Time taken (start to finish)</td>
<td>13 years</td>
<td>4 years</td>
<td>4.5 months</td>
</tr>
<tr>
<td>Number of scientists listed as authors</td>
<td>≥ 2,800</td>
<td>31</td>
<td>27</td>
</tr>
<tr>
<td>Cost of sequencing (start to finish)</td>
<td>$2.7 billion</td>
<td>$100 million</td>
<td>&lt; $1.5 million</td>
</tr>
<tr>
<td>Coverage</td>
<td>8-10 x</td>
<td>7.5 x</td>
<td>7.4 x</td>
</tr>
<tr>
<td>Number of institutes involved</td>
<td>16</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Number of countries involved</td>
<td>6</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

### Table 2: Sequencing statistics on personal genome projects

<table>
<thead>
<tr>
<th>Personal Genome</th>
<th>Platform</th>
<th>Genomic template libraries</th>
<th>No. of reads (millions)</th>
<th>Read length (bases)</th>
<th>Base coverage (fold)</th>
<th>Assembly coverage (%)</th>
<th>SAs in millions (allignment tool)</th>
<th>No. of runs</th>
<th>Estimated cost (US$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>J-Craig Venter</td>
<td>Automated Sanger</td>
<td>MP from SAMs, karyochromosomes</td>
<td>33.9</td>
<td>800</td>
<td>75</td>
<td>Denovo N/A</td>
<td>3.21</td>
<td>~340,000</td>
<td>70,000,000</td>
</tr>
<tr>
<td>James D. Watson</td>
<td>Roche/454</td>
<td>Frags 500 bp</td>
<td>93.1</td>
<td>259</td>
<td>24</td>
<td>N/A</td>
<td>3.32 (BILIA)</td>
<td>214</td>
<td>1,000,000</td>
</tr>
<tr>
<td>Japanese male</td>
<td>Illumina/Solexa</td>
<td>93% MP: 200 bp</td>
<td>3.41 **</td>
<td>35</td>
<td>46.9</td>
<td>N/A</td>
<td>3.84 (MACS)</td>
<td>40</td>
<td>25,000 **</td>
</tr>
<tr>
<td>Korean male (AKI)</td>
<td>Illumina/Solexa</td>
<td>65% Frags: 150-250 bp</td>
<td>1.91 **</td>
<td>35</td>
<td>36</td>
<td>N/A</td>
<td>3.03 (SOAP)</td>
<td>35</td>
<td>500,000 **</td>
</tr>
<tr>
<td>Korean male (SK)</td>
<td>Illumina/Solexa</td>
<td>21% Frags: 110 bp</td>
<td>39 **</td>
<td>39</td>
<td>22.8</td>
<td>N/A</td>
<td>3.45 (SNAP)</td>
<td>30</td>
<td>200,000 **</td>
</tr>
<tr>
<td>Japanese male (NA18507)</td>
<td>Illumina/Solexa</td>
<td>MP: 300 bp, 2000 bp</td>
<td>1.647 **</td>
<td>35.74</td>
<td>29.0</td>
<td>N/A</td>
<td>3.44 (MACS)</td>
<td>15</td>
<td>250,000 **</td>
</tr>
<tr>
<td>Venter female</td>
<td>Life/454</td>
<td>9.9 Frags: 100-500 bp</td>
<td>21.7</td>
<td>50</td>
<td>11.9</td>
<td>N/A</td>
<td>3.82 (Corona-lite)</td>
<td>9.5</td>
<td>60,000 ***</td>
</tr>
<tr>
<td>Stephen R. Oosterhuis</td>
<td>Illumina/BIOSciences</td>
<td>Frags 500 bp</td>
<td>2.715 **</td>
<td>32</td>
<td>28</td>
<td>N/A</td>
<td>2.81 (inconsol)</td>
<td>4</td>
<td>48,000 **</td>
</tr>
<tr>
<td>AML female</td>
<td>Illumina/Solexa</td>
<td>Frags: 150-200 bp</td>
<td>2.760 **</td>
<td>32</td>
<td>27.7</td>
<td>N/A</td>
<td>3.83 (MACS)</td>
<td>98</td>
<td>1,000,000 **</td>
</tr>
<tr>
<td>AML male</td>
<td>Illumina/Solexa</td>
<td>Frags: 150-200 bp</td>
<td>1.851 **</td>
<td>35</td>
<td>13.9</td>
<td>N/A</td>
<td>2.54 (MACS)</td>
<td>94</td>
<td>500,000 **</td>
</tr>
<tr>
<td></td>
<td>MP: 200-250 bp</td>
<td>18.00m **</td>
<td>35</td>
<td>23.4</td>
<td>N/A</td>
<td>3.40 (MACS)</td>
<td>11.1</td>
<td>500,000 **</td>
<td></td>
</tr>
</tbody>
</table>
Challenges of existing NGS methods

- Increase read length
- Improve sequence accuracy
- Single-molecule sequencing (no amplification)
- De-novo assembly of complex genomes

<table>
<thead>
<tr>
<th>Method</th>
<th>Cost per megabase</th>
<th>Cost per instrument</th>
<th>Paired ends?</th>
<th>1st error modality</th>
<th>Read-length</th>
</tr>
</thead>
<tbody>
<tr>
<td>454</td>
<td>~$60</td>
<td>$500,000</td>
<td>Yes</td>
<td>Indel</td>
<td>250 bp</td>
</tr>
<tr>
<td>Solexa</td>
<td>~$2</td>
<td>$430,000</td>
<td>Yes</td>
<td>Subst.</td>
<td>36 bp</td>
</tr>
<tr>
<td>SOLiD</td>
<td>~$2</td>
<td>$591,000</td>
<td>Yes</td>
<td>Subst.</td>
<td>35 bp</td>
</tr>
<tr>
<td>Polonator</td>
<td>~$1</td>
<td>$155,000</td>
<td>Yes</td>
<td>Subst.</td>
<td>13 bp</td>
</tr>
<tr>
<td>HeliScope</td>
<td>~$1</td>
<td>$1,350,000</td>
<td>Yes</td>
<td>Del</td>
<td>30 bp</td>
</tr>
</tbody>
</table>

3rd generation sequencing methods

[Images of various sequencing methods]
Pacific biosystems technology

Real time monitoring of single-molecule sequencing with an immobilized DNA polymerase

Oxford nanopore technology

Base identification through differences in conductance of nucleic acid molecules driven through a nanopore
### Targeted DNA region enrichment

#### Array hybridization
- All exons
- Transcribed sequences
- Candidate regions
- Structural variants

#### Solution hybridization
ChIP-chip or ChIP-seq

Chromatin immunoprecipitation (ChIP)

- Transcription factor binding sites
- DNA methylation
- Histone modifications

ChIP-chip

ChIP-seq

- Isolate genomic DNA with formaldehyde
- Sonicate DNA to produce sheared, soluble chromatin
- Add protein-specific antibody
- Immunoprecipitate and purify immunocomplexes
- Reverse cross-links, purify DNA and prepare for sequencing

---

The impact of next-generation sequencing technology on genetics

Elaine R. Mandis
Washington University School of Medicine, Genome Sequencing Core, St. Louis, MO 63110, USA

It has become clear that the fundamental principle guiding scientific investigation is to determine the genotypes that explain phenotypes. The recent introduction of high-throughput DNA sequencing technologies has provided an opportunity to refine this approach by unraveling the complex regulatory networks that determine an organism’s phenotype. In this review, I will discuss the impact of these technologies on our understanding of the genome and the processes that control genetic variation. The ability to generate large amounts of sequence data from any species has allowed for the systematic analysis of the genome, providing insight into the mechanisms that underlie biological processes.

Sequencing technologies — the next generation

Michael L. Metzker*

Abstract: DNA has never been cheaper for revolutionary technologies that deliver fast, inexpensive and accurate genome information. This challenge has catalysed the development of next-generation sequencing (NGS) technologies. The expensive production of large volumes of sequence data is the primary advantage over conventional methods. Here, I present a technical review of template preparation, sequencing and imaging, genome alignment and assembly approaches, and recent advances in current and next-term commercially available NGS instruments. I also outline the broad range of applications for NGS technologies, in addition to providing guidelines for platform selection to address biological questions of interest.

* Correspondence: MMetzker@wustl.edu

Trends in Genetics Vol. 24 No. 3