Next Generation Sequencing: Applications

Anna De Grassi
-- European Institute of Oncology - Milan --
- F. Ciccarelli group -

BITS - March 20, 2009 - Genoa
Several Flavours of Throughput...

- Genome sequencing
- Metagenomics
- Amplicon sequencing
- UltraDeep sequencing
- Transcriptome Analysis
- Chip-seq
- Structural Variations
- SNPs and Point Mutations
- Nucleosome positioning
"metagenomics is the application of modern genomics techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species."

Kevin Chen and Lior Pachter (University of California, Berkeley)

>99% of all microbes cannot be cultured

Soil - Sea - Air - ancient DNA - body parts
Metagenomics

**Selection of microbial cells**
- DNA extraction

**Shotgun sequencing:**
- cloning in plasmid library
- 3730xl capillary sequencer

**454 sequencing:**
- nebulization, ligation, fixed to beads and emulsion PCR
- GS20 pyrosequencer

<table>
<thead>
<tr>
<th>Microbiome</th>
<th>Average read length</th>
<th>Number of reads</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>lean1 (GS20)</td>
<td>90.9</td>
<td>1,046,611</td>
<td>94,913,476</td>
</tr>
<tr>
<td>ob1 (GS20)</td>
<td>96.4</td>
<td>677,384</td>
<td>65,370,448</td>
</tr>
<tr>
<td>lean1 (3730xl)</td>
<td>765</td>
<td>10,752</td>
<td>8,227,047</td>
</tr>
<tr>
<td>lean2 (3730xl)</td>
<td>782</td>
<td>11,136</td>
<td>8,705,876</td>
</tr>
<tr>
<td>lean3 (3730xl)</td>
<td>706</td>
<td>10,752</td>
<td>7,590,528</td>
</tr>
<tr>
<td>ob1 (3730xl)</td>
<td>735</td>
<td>11,136</td>
<td>8,185,880</td>
</tr>
<tr>
<td>ob2 (3730xl)</td>
<td>771</td>
<td>8,832</td>
<td>6,811,035</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>-</td>
<td><strong>1,776,603</strong></td>
<td><strong>199,804,290</strong></td>
</tr>
</tbody>
</table>

Abbreviations: GS20, pyrosequencer; 3730xl, capillary sequencer

Draft genome of the most common bacterium (*E. rectale*):
- overlap generation
- contig layout
- consensus generation

**Metagenomics Analyses:**
- BLASTX (e<10^-5)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Contigs</th>
<th>Average contig length</th>
<th>Contiged bases^1^</th>
<th>Largest Assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>lean1 (GS20)</td>
<td>102,299</td>
<td>117</td>
<td>11,966,580</td>
<td>2,793</td>
</tr>
<tr>
<td>ob1 (GS20)</td>
<td>56,425</td>
<td>116</td>
<td>6,518,469</td>
<td>2,174</td>
</tr>
<tr>
<td>lean1 (3730xl)</td>
<td>167</td>
<td>1527</td>
<td>254,985</td>
<td>5,500</td>
</tr>
<tr>
<td>lean2 (3730xl)</td>
<td>407</td>
<td>1598</td>
<td>650,499</td>
<td>5,522</td>
</tr>
<tr>
<td>lean3 (3730xl)</td>
<td>224</td>
<td>1528</td>
<td>342,172</td>
<td>3,281</td>
</tr>
<tr>
<td>ob1 (3730xl)</td>
<td>320</td>
<td>1393</td>
<td>445,814</td>
<td>3,225</td>
</tr>
<tr>
<td>ob2 (3730xl)</td>
<td>269</td>
<td>1644</td>
<td>442,210</td>
<td>4,186</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Microbiome</th>
<th>Total NR EGTs</th>
<th>Total COG EGTs</th>
<th>Total KO EGTs</th>
<th>Total EGTs</th>
<th>Percent unassigned</th>
</tr>
</thead>
<tbody>
<tr>
<td>lean1 (GS20)</td>
<td>48,625</td>
<td>51,481</td>
<td>28,359</td>
<td>56,599</td>
<td>94.6</td>
</tr>
<tr>
<td>ob1 (GS20)</td>
<td>33,360</td>
<td>32,819</td>
<td>18,308</td>
<td>39,058</td>
<td>94.2</td>
</tr>
<tr>
<td>lean1 (3730xl)</td>
<td>7,973</td>
<td>7,970</td>
<td>2,810</td>
<td>8,462</td>
<td>21.3</td>
</tr>
<tr>
<td>lean2 (3730xl)</td>
<td>7,309</td>
<td>7,687</td>
<td>2,723</td>
<td>8,170</td>
<td>26.6</td>
</tr>
<tr>
<td>lean3 (3730xl)</td>
<td>7,042</td>
<td>7,119</td>
<td>2,562</td>
<td>7,616</td>
<td>29.2</td>
</tr>
<tr>
<td>ob1 (3730xl)</td>
<td>7,331</td>
<td>7,299</td>
<td>2,639</td>
<td>7,859</td>
<td>29.4</td>
</tr>
<tr>
<td>ob2 (3730xl)</td>
<td>6,008</td>
<td>6,016</td>
<td>2,053</td>
<td>6,425</td>
<td>27.3</td>
</tr>
</tbody>
</table>

EGS = enviromental gene tags
Capillary Pros:
- more confident gene calling

454 Pros:
- less time consuming
- higher sequence coverage
- not affected by cloning bias

only 454 for metagenomics applications
Ultra-deep sequencing

Re-sequencing a region several times to detect non-common variants

Sanger

Only consensus sequence: ATCGT

NGS

ATCGT
ATCGT
ATAGT
ATCGT
ATAGT
ATCGT
ATAGT
Ultra-deep sequencing

Detection of rare sub-clonal mutations in cancer cells

**Samples:**
- Blood of 24 patients affected by CCL (chronic lymphocytic leukemia)
- Renal cell of 1 patient

- PCR amplification
- Equimolar pool of amplicons
- One 454 run

- Somatic hypermutations
  - IGH VDJ locus
  - ~300bp

- 385,000 reads, ~250bp per read (>95% aligned to the reference)

Campbell, PJ PNAS - 105, 13081 – 13086 2008
Ultra-deep sequencing

ERROR PROCESSING: Analysis of the control locus
all the variations from the reference sequence are artifacts

Sequencing errors:
• polyN > 4
• many indels (sequence ends)
• few substitutions (throughout the sequence)

DNA polymerase errors:
• not associated to polyN
• typical substitution pattern
  e.g (G:C->A:T) most common
Ultra-deep sequencing

Filter to detect “real” rare variants in 24 samples by excluding:
- poor quality reads
- indels and substitutions in polyN tracts > 4bp
- expected from the distribution of polymerase errors
- only in forward and reverse

Sub-clonal mutations can be detected down to a frequency of 1/5000 reads

Phylogenetic analysis:
- clustalW
- maximum parsimony
- 1000 bootstrap

Campbell, PJ *PNAS* - 105, 13081 – 13086 2008
Protein-DNA binding sites

Chip-chip limits:
- low resolution
- incorrect hybridizations
- *a priori* knowledge of potential binding sites
- no information on the sequence

Protein-DNA binding sites

**Protein: NRSF (neuron-restrictive silencer factor)**
- known "gold standard" target genes
- known DNA motif
- high-quality antibody

**DNA samples:**
- NRSF enriched Chip sample
- control of chromatin not immuno-enriched

**Sequencing and Mapping:**
- 2-5M reads, 25nt
- 50% maps on unique locations
- <3 mismatches allowed

**Detection of binding sites:**
- >= 13 reads per sequence
- 5 fold enrichment vs control

Protein-DNA binding sites

Benchmark:
• compare with known positive and negative binding sites
• sensitivity = 87%
• specificity = 98%

Variation of DNA motifs at the binding site:
• 100bp from the “best” 10% segments screened by a motif-finding algorithm
• 75% have the known canonical motif
• detection of novel non canonical motifs

Canonical

Non canonical
Single-stranded RNA molecules of 21-23nt long that regulate gene expression

Samples:
- Pluripotent human embryonic stem cells (hESCs)
- Differentiated cells: embryonic bodies (EBs)

RNA preparation and sequencing:
- Extraction of small RNAs
- Libraries of single-stranded cDNA
- Illumina sequencing

Filter and Mapping on the genome:
- Unfiltered reads: 6M, 25nt
- Perfect alignments to the genome (no indels): ~4M (70%) reads and ~0.75M unique sequences
- Only sequences observed > 3 reads

Overlap with DBs of known sequences: 5% sequences
Qualitative analysis (known microRNAs):
- detect the variability between reads of the same microRNA sequence: cleavage positions and post-translational modifications
**Quantitative analysis:**
- reads count per sequence is an index of the expression level (digital expression)
- detect the differential expression of microRNAs between samples

### 100 microRNAs

<table>
<thead>
<tr>
<th>microRNA ID</th>
<th>Pre-miRNA arm (5-p or 3-p)</th>
<th>Most abundant sequence (isomiR)</th>
<th>hESC count</th>
<th>EB count</th>
<th>P-value</th>
<th>Fold change</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-199a</td>
<td>3-p</td>
<td>ACAGTACTGCTGCACATGATAGTTA</td>
<td>1110</td>
<td>13,163</td>
<td>0.00</td>
<td>11.86</td>
</tr>
<tr>
<td>miR-372</td>
<td>3-p</td>
<td>AAAGTGCTGCACATTTGACGCT</td>
<td>1388</td>
<td>13,653</td>
<td>0.00</td>
<td>9.84</td>
</tr>
<tr>
<td>miR-122</td>
<td>5-p</td>
<td>TGGAGTCTGACAATGGTGTGTTG</td>
<td>436</td>
<td>2565</td>
<td>0.00</td>
<td>5.88</td>
</tr>
<tr>
<td>miR-152</td>
<td>3-p</td>
<td>TCAGTCATGACAGAACTTGGG</td>
<td>622</td>
<td>3028</td>
<td>0.00</td>
<td>4.87</td>
</tr>
<tr>
<td>miR-10a</td>
<td>5-p</td>
<td>TACCCCTGATGATCCGAATTTT</td>
<td>948</td>
<td>3887</td>
<td>0.00</td>
<td>4.10</td>
</tr>
<tr>
<td>let-7a</td>
<td>5-p</td>
<td>TGAGGTAATAGGTTGTATAGTT</td>
<td>11,902</td>
<td>2951</td>
<td>0.00</td>
<td>4.03</td>
</tr>
<tr>
<td>miR-302a</td>
<td>5-p</td>
<td>TAAACGTGGATGTACCTGTTT</td>
<td>36,800</td>
<td>9917</td>
<td>0.00</td>
<td>3.71</td>
</tr>
<tr>
<td>miR-222</td>
<td>3-p</td>
<td>AGCTACATCCTGGCTAGGGTCTC</td>
<td>4719</td>
<td>1331</td>
<td>0.00</td>
<td>3.55</td>
</tr>
<tr>
<td>miR-340</td>
<td>5-p</td>
<td>TTATAAAGCAATGAGACTGATT</td>
<td>2247</td>
<td>7198</td>
<td>0.00</td>
<td>3.20</td>
</tr>
</tbody>
</table>
Samples:
- Pluripotent mouse embryonic stem cells (ES)
- Differentiated cells: embryonic bodies (EB)

- mRNA extraction
- Library generation (in triplicate per sample)
- Sequencing
Trascriptome profiling

Filter and Mapping strategy: 7 steps!!

1. Quality check or removal of 5nt
2. Clustering to unique tags
3. Mapping on the genome (<=2 mismatches)

Reads mapping on the genome:
- ~95M reads (60%)
  Multiple mapping is accepted (if less than 100 positions)

Gene expression (tag count):
- high reproducibility between replicates \((r > 0.95)\)
- good reproducibility between tag counts per gene and microarray signal

Differential expression between samples:
- tag counts per gene in ES and EB
  (35/50 ES markers were confirmed):
  70% sensitivity

Transcriptome profiling

Transcriptome discovery:
- ~33% of tags are in non-exonic sequences
- 20% of tags are in repeat elements (normally excluded from expression arrays)

Alternative splicing isoforms:
- high quality 35mers were clustered in a longer consensus (>50nt)
- BLAT on the genome
Discovery of expressed SNPs: Extensive filtering!

Only full length tags (35nt) and high quality

Mapping to the genome: (multi-mapping are excluded)

Filter by colour-space errors

Filter by error profile of tags: first 6nt, last 5nt and 26

Filter by proportion: 75% of tag are mutated: (heterozigous mutations are sistematically discarded)

- 2,000 putative SNPs in both samples
- 643 in Refseq (84% known SNPs)
- 8/10 non synonymous SNPs validated by PCR: specificity = 80%

## Summary

<table>
<thead>
<tr>
<th>Application</th>
<th>454</th>
<th>Illumina</th>
<th>SOLiD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome sequencing</td>
<td>Small Genomes</td>
<td>Small Genomes</td>
<td>No</td>
</tr>
<tr>
<td>Genome re-sequencing</td>
<td>Yes</td>
<td>Yes</td>
<td>Small genome</td>
</tr>
<tr>
<td>Metagenomics</td>
<td>Yes</td>
<td>Only virus</td>
<td>No</td>
</tr>
<tr>
<td>Amplicon sequencing</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Ultra-deep sequencing</td>
<td>Yes</td>
<td>Tested only for 100s reads</td>
<td>Tested only for 100s reads</td>
</tr>
<tr>
<td>Transcriptome Analysis</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Structural variations</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>SNPs and Point Mutations</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Chip-Seq</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Nucleosome positioning</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### Read length

- **454**
- **Illumina**
- **SOLiD**

### Number of reads

- Small Genomes
- Small genome
- Tested only for 100s reads
- Tested only for 100s reads