Long reads sequencing technology to solve complex genomic regions assembly in plants

Hélène BERGES
Director of the Plant Genomic Center

The French Plant Genomic Center
Created in 2004 by INRA (French National Institute for Agricultural Research)
- Depository of genomic libraries for the scientific community
  - BAC libraries
- A dedicated structure to assist plant genomic programs
  - Distribute the genomic resources at the international level
  - Provide high quality research material and efficient tools and services for studying plant genomes
  - Develop innovative solutions
  - Develop genomic projects in collaboration
  - Host scientists in the frame of collaborations

The challenges - The expectations

Genome
- Size / diversity
- Reduce the complexity

Chromosome
- Focus / Target / Markers
- Target specific Genomic regions
- Manage diversity

Genomic Libraries available at CNRGV

- > more than 430 different Genomic Libraries
  - > 20 M unique clones

Interactions with laboratories around the world

- 2 617 240 BAC clones distributed during the last 3 years (2012-2014)

Services at CNRGV
The Next-generation DNA sequencing technologies

Complex plant genomes sequencing projects became possible (many billions of bases per day for hundreds or thousands of dollars per gigabase instead of millions or billions of dollars per gigabase)

but de novo assembling of plant genomes remains challenging
- Size of the reads
- Gene families are difficult to assemble and may collapse into a mosaic
- Repeat elements
- How to assemble these genomes accurately?

The progress made with the NGS technologies we still don’t have enough reference plant genomes with high confident data (false conclusions ?)
- Third-generation sequencing technologies ?

Sequencing strategies for complex genomes

Two main sequencing strategies :
- Using BAC libraries
- Whole Genome Shotgun

Genome exhibits high levels of diversity

Repeat elements (>80 % in maize, barley, sunflower, wheat etc …)
- Polyploidy

The Sugarcane genome Sequencing initiative (SUGESI): Strategies for Sequencing a Highly Complex Genome

The Sugarcane Genome Sequencing Initiative (SUGESI) was envisaged to provide the resources to assemble repeat sequences of BAC libraries and to carry out the assembly of whole genome sequences.

A. thaliana (2x) Rice (2x) Maize (2x) Barley (2x) Wheat (6x)
Genome sequencing strategies

BAC by BAC

1. Construction of BAC libraries
2. Physical map
3. MTP selection
4. BACs sequencing

Anchoring using genetic maps

SHOTGUN

1. Genome fragmentation
2. Paired end Sequencing
3. Sequences Assembling by contigs
4. Combined data to assemble contigs into scaffolds

BAC libraries are still of interest

BAC libraries related to the project

High genome coverage 15-20X -> comprehensiveness
Whole genome sequencing project

Low genome coverage 2X
Targeted Genomic Region on specific genotypes

Medium insert size 120kb
X: genome equivalent coverage
P: Probability to find any region in the genome

The low coverage BAC library

BAC library with 1-2X genome coverage
Specific genotypes of interest – specific species

1 – Establishment of physical maps on locus of interest
If no genotypes with no BAC libraries available (ex: resistant plant)
2 – Filling of residual gaps on physical map obtained with classical BAC libraries
3 – Sequencing of a target zone or genes in several different genotypes / species for syntenic analysis

Possibility to increase the coverage step by step
- Arrayed Libraries for small genomes
- Non gridded BAC Libraries for large genomes

Comparison of PacBio RS II and Roche-454 technologies
For BAC pool sequencing

Pool of nine BAC clones, originated from six different plant species:
- maize (2)
- wheat (2)
- strawberry
- sunflower
- barley
- sugarcane (2)

Roche-454: BAC clones individually tagged prior to be mixed in the same sequencing run.
PacBio RS II: BAC clones pooled without tags

The non-gridded BAC library strategy

- Non gridded library from specific genotype
- Pools of BAC transformants

- Genetic map
- Physical map established on specific genotype
- Markers definition

- BAC-Pool Sequencing (454, PacBio)
- 35 to 100 x coverage
- PacBio Technology
  1 BAC: 1 contig
  MTP of BACs: 1 contig

- Screening of the pools with specific markers
- Identification of positive BAC clones
- BAC clone characterisation (BES, Insert size)

Workflow Assembly of PacBio RS II reads and related metrics

Nathalie RODDE – P0228

Stéphane CAUET
**Comparison of assembly quality**

<table>
<thead>
<tr>
<th>Clone name</th>
<th>Estimated insert size (kb)</th>
<th>Roche-454 contigs</th>
<th>PacBio RS II contigs</th>
<th>Roche-454 size (bp)</th>
<th>PacBio RS II size (bp)</th>
<th>Roche-454/PacBio RS II size ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frg-55210</td>
<td>90</td>
<td>2</td>
<td>1</td>
<td>90248</td>
<td>90557</td>
<td>0.99659</td>
</tr>
<tr>
<td>Heli-37708</td>
<td>170</td>
<td>6</td>
<td>1</td>
<td>175471</td>
<td>175563</td>
<td>0.98808</td>
</tr>
<tr>
<td>Hord-155N13</td>
<td>155</td>
<td>10</td>
<td>1</td>
<td>155553</td>
<td>155889</td>
<td>0.97984</td>
</tr>
<tr>
<td>Secx2-24H10</td>
<td>150</td>
<td>17</td>
<td>1</td>
<td>142576</td>
<td>152547</td>
<td>0.94600</td>
</tr>
<tr>
<td>Secx2-27N020</td>
<td>100</td>
<td>1</td>
<td>1</td>
<td>108594</td>
<td>108581</td>
<td>1.00003</td>
</tr>
<tr>
<td>TIR-12NP13</td>
<td>120</td>
<td>2</td>
<td>1</td>
<td>124789</td>
<td>125174</td>
<td>0.95264</td>
</tr>
<tr>
<td>TIR-11J10</td>
<td>110</td>
<td>17</td>
<td>1</td>
<td>121533</td>
<td>131314</td>
<td>0.94077</td>
</tr>
<tr>
<td>Zeam-34K24</td>
<td>135</td>
<td>10</td>
<td>1</td>
<td>128036</td>
<td>133231</td>
<td>0.96108</td>
</tr>
<tr>
<td>Zeam-30K1L</td>
<td>85</td>
<td>1</td>
<td>1</td>
<td>86647</td>
<td>86651</td>
<td>0.99955</td>
</tr>
</tbody>
</table>

PacBio RS II reads assembly performed following HGAP workflow.

- Assembly of PacBio RS II sequences of pool of untagged BAC clones led to one contig per BAC assigned with BAC-end sequences.

**High level of identity in the assembled sequences (with both long and short reads technologies)**

<table>
<thead>
<tr>
<th>Clone name</th>
<th>Matches</th>
<th>Substitution d'</th>
<th>Substitution ins/del</th>
<th>Substitution ins/del &amp; indels</th>
<th>Mismatches</th>
<th>Per-base identity</th>
<th>Per-event identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frg-55210</td>
<td>92048</td>
<td>0 (0)</td>
<td>23 (15)</td>
<td>0 (0)</td>
<td>23 (15)</td>
<td>0.99974</td>
<td>0.98834</td>
</tr>
<tr>
<td>Heli-37708</td>
<td>37471</td>
<td>0 (0)</td>
<td>80 (43)</td>
<td>0 (0)</td>
<td>80 (43)</td>
<td>0.99950</td>
<td>0.99718</td>
</tr>
<tr>
<td>Hord-155N13</td>
<td>155546</td>
<td>4 (4)</td>
<td>15 (11)</td>
<td>1 (1)</td>
<td>20 (16)</td>
<td>0.99971</td>
<td>0.99807</td>
</tr>
<tr>
<td>Secx2-24H10</td>
<td>142520</td>
<td>1 (1)</td>
<td>18 (15)</td>
<td>10 (10)</td>
<td>29 (24)</td>
<td>0.99979</td>
<td>0.99918</td>
</tr>
<tr>
<td>Secx2-27N020</td>
<td>108594</td>
<td>0 (0)</td>
<td>54 (32)</td>
<td>0 (0)</td>
<td>54 (32)</td>
<td>0.99966</td>
<td>0.99837</td>
</tr>
<tr>
<td>TIR-12NP13</td>
<td>124789</td>
<td>0 (0)</td>
<td>5 (5)</td>
<td>0 (0)</td>
<td>5 (5)</td>
<td>0.99960</td>
<td>0.99900</td>
</tr>
<tr>
<td>TIR-11J10</td>
<td>121533</td>
<td>3 (3)</td>
<td>28 (25)</td>
<td>4 (2)</td>
<td>48 (46)</td>
<td>0.99967</td>
<td>0.99903</td>
</tr>
<tr>
<td>Zeam-34K24</td>
<td>128036</td>
<td>0 (0)</td>
<td>7 (2)</td>
<td>0 (0)</td>
<td>7 (2)</td>
<td>0.99959</td>
<td>0.99915</td>
</tr>
<tr>
<td>Zeam-30K1L</td>
<td>86647</td>
<td>0 (0)</td>
<td>9 (3)</td>
<td>0 (0)</td>
<td>9 (3)</td>
<td>0.99958</td>
<td>0.99938</td>
</tr>
</tbody>
</table>

PacBio average error rates in raw reads ranging from 14.6% to 14.96% among BAC clones.

- Randomly distributed errors is corrected using high reads coverage to establish a consensus sequence that represents the biological reality.

**Interest of the long reads for complex genomic region**

Alignment of Roche-454 and PacBio RS II contigs of Zeam-34K24 BAC

- Two contigs (454) displaying strong homologies with two distinct regions (PacBio).
- Roche-454 reads coverage exhibited two spikes / PacBio RS II reads coverage is stable corresponding to misassembled data.

**Conclusion**

- Single molecule long reads technology resolves gaps and collapsing issues of shorter reads sequences assembly
  - Missing regions in short read technology
  - Collapsing of duplicated regions
- Interest of long reads technology in the assembly of genome sequences and consequently in the accuracy of the data generated
- Especially with complex genome such as plant genomes with duplicated genes, high level of repetitive elements or genome duplication but still some issues with polyploidy
- Combination of BAC and long read technology to solve the issues

**The strategy**

- **Specific Markers of genomic regions**
- **BAC-Pool Sequencing (PacBio)**
  - 35 to 100 x coverage
- PacBio Technology - BAC : 1 contig
  - MTP of BACs : 1 contig
- Essential and efficient tools for understanding the organization and function of specific genomic regions

- **Screening Identification of BAC clones**
- **BAC library from various genotypes**

**Alignment of Roche-454 and PacBio RS II contigs of Zeam-34K24 BAC**

- **Collapsing region corresponding to strong similar repeated element (LTR – copia superfamily)**

1/19/2016
Aknowledgements

Arnaud BELLEC  
Sonia VAUTRIN  
Genséric BEYDON  
William MARANDE  
Joëlle FOURMENT  
Elisa PRAT  
Nadine GAUTIER  
Nathalie RODDE  
Nadège ARNAL  
Audrey COURTIAL  
Stéphane CAUET  
David PUJOL  
Nicolas THERON  
Laetitia HOARAU  
Hélène BERGES  

http://cnrgv.toulouse.inra.fr/  
@CNRGV  

Clémentine Vitte  
Wolfgang Spielemeyer  
Kellye Eversole  
Béatrice Denoyes  

Nils Stein  
Anete Pereira de Souza / Danilo Augusto  

Béatrice Denoyes