Genomics guided plant specialized metabolism study in *Catharanthus roseus*

03/04/15

Jeongwoon Kim

(Buell lab)
Plant specialized metabolites in daily life

Linalool (Lamiaceae family)

Basil

Mint

Humulone (Hop)
Acylsugars produced in tobacco are important for plant defense

Artemisinin produced in wormwood is used as anti-malarial drug.
Challenges to study plant specialized metabolism

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemical complexity and diversity</td>
<td>Complicated biosynthetic pathway</td>
</tr>
<tr>
<td>Species-specific, mostly non-model</td>
<td>Limited genomic resource available</td>
</tr>
<tr>
<td>Low abundance compounds</td>
<td>Low gene expression</td>
</tr>
<tr>
<td>Some enzymes from big gene families</td>
<td>High sequence resolution needed to distinguish homologs</td>
</tr>
</tbody>
</table>

⇒ Application of NGS!!
About Medicinal Plant Genomics Resource

Natural products from plants serve as rich resources for drug development with almost 100 plant-derived compounds in clinical trials in 2007. Plant derived natural products have had a profound and lasting impact on human health and include compounds successfully used for decades such as digitalis, Taxol, vincristine, and morphine isolated from foxglove, periwinkle, yew, and opium poppy, respectively. The enormous structural diversity and biological activities of plant-derived compounds suggest that additional, medicinally relevant compounds remain to be discovered in plants.

http://medicinalplantgenomics.msu.edu
Why *Catharanthus roseus*?

Madagascar periwinkle

$20\ M/kg$
Collaborative genome study

• Dr. Robin Buell group (MSU)  
  : bioinformatics analysis

• Dr. Sarah O’Connor group (The John Innes Center, UK)  
  : biochemistry, organic chemistry
Genome assembly

- Genomic DNA from *C. roseus* SunStorm™ Apricot
- A self-pollinating diploid (2x = 2n = 16)
- Estimated genome size: ~ 740 Mbp
Genome assembly

- TruSeq genomic DNA library (400 bp fragment size)
- A single lane Illumina HiSeq, 101 bp PE
- ~38 Gbp generated (~52x coverage)
- De novo assembly with Abyss, k-mer 71 (The Genome Analysis Center, UK)
**Genome assembly statistics**

<table>
<thead>
<tr>
<th>Scaffolds ≥ 1,000 bp</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>✔ Total assembled length</td>
<td>506,421,720 (Ns: 3.69 Mb)</td>
</tr>
<tr>
<td>✔ Number of scaffolds</td>
<td>41,176</td>
</tr>
<tr>
<td>✔ N50 scaffold size (bp)</td>
<td>27,337</td>
</tr>
<tr>
<td>✔ Minimum scaffold size (bp)</td>
<td>1,000</td>
</tr>
<tr>
<td>✔ Maximum scaffold size (bp)</td>
<td>250,200</td>
</tr>
<tr>
<td>✔ No. ESTs aligned at 90% identity/70% coverage (%)</td>
<td>18,681 (92.6%)</td>
</tr>
<tr>
<td>✔ No. ESTs aligned at 90% identity/50% coverage (%)</td>
<td>19,111 (94.7%)</td>
</tr>
<tr>
<td>✔ No. ESTs aligned at 90% identity/30% coverage (%)</td>
<td>19,319 (95.7%)</td>
</tr>
<tr>
<td>✔ No. ESTs aligned below 90% identity/30% coverage (%)</td>
<td>441 (2.2%)</td>
</tr>
<tr>
<td>✔ No. complete CEGMA proteins (%)</td>
<td>210 (84.7%)</td>
</tr>
<tr>
<td>✔ No. complete or partial CEGMA proteins (%)</td>
<td>242 (97.6%)</td>
</tr>
</tbody>
</table>
Genome annotation

- Gene model prediction using MAKER annotation pipeline
- HMM prediction using SNAP, Augustus
- UniProt SwissProt, *Arabidopsis* TAIR10 protein evidence
- *C. roseus* EST/transcript evidence
- Maximum numbers of predicted gene models including *ab initio* models
- HMMScan for Pfam domain search
- Gene models predicted with transcript/protein evidence and/or Pfam domain
### Gene model annotation statistics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of protein coding genes</td>
<td>33,829</td>
</tr>
<tr>
<td>Average gene length (bp)</td>
<td>3,090</td>
</tr>
<tr>
<td>Average transcript length (bp)</td>
<td>1,159</td>
</tr>
<tr>
<td>Average CDS length (bp)</td>
<td>1,021</td>
</tr>
<tr>
<td>Average exon length (bp)</td>
<td>234</td>
</tr>
<tr>
<td>Average intron length (bp)</td>
<td>490</td>
</tr>
</tbody>
</table>

MAKER provides:
- PepNde/transcript fasta, gff files
- AnnotateEdit: Distance score (0: best, 1: no transcript/protein evidence)
Monoterpene indole alkaloid (MIA) genes in Catharanthus genome

Kellner* and Kim* et al.
Under review
Questions to address

• Known MIA genes identified from the genome assembly!
• Next steps
  - Identification of candidate genes in MIA pathway
  - Understanding evolution of MIA genes
  - Regulation of MIA pathway
Search for candidate genes #1

- Co-expression of genes in the same pathway
- No exposure of potentially toxic intermediates
- Efficient delivery of substrate to next enzyme


Co-expression of MIA genes in C. roseus

RNAseq (immature/mature leaf, stem, root, flower, seedlings, seedlings w/ MeJA)
TopHat/Cufflinks Heatmap (MeV)

- MIA gene expression induced by MeJA
- Co-expression of MIA genes from each part of pathway

Kellner* and Kim* et al. Under review
Search for candidate genes #2

- Physical clustering of plant specialized metabolism genes
  Ex. Terpene synthases in tomato
  Noscapine genes in Opium poppy

Winzer et al. (2012) Science 336, 1704-1708. – BAC study
Physical clustering of MIA genes on genome

- Physical location of known MIA genes – genome assembly and BAC
- 7 clusters, 8 novel candidate genes (white arrows)

Kellner* and Kim* et al. Under review
Expansion of MIA genes – STR

Ortholog search with OrthMCL
- C. roseus
- S. lycopersicum
- A. thaliana
- V. vinifera
- A. trichopoda

- In C. roseus, 12 STR domain encoding proteins
- Cro_006099: characterized STR
- 2 groups (ancient/recent duplication?)
- MeJA inducible (close circle) vs non-inducible (open)

Kellner* and Kim* et al., Under review
Expansion of ORCA3 TF

ORCA2 (Menke et al. 1999)
ORCA3 (van der Fits and Memelink 2000)

- Duplication of AP2 class TFs
- Differential gene expression of TFs in various tissues
Co-expression network for ORCA2 TF

- ORCA2 TF in MIA pathway (Menke et al. 1999)
- Mutual rank with FPKM matrix
- Top 200 genes co-expressed with ORCA2
- Cytoscape

Kellner* and Kim* et al. Under review
Transcriptional network by ORCA2

Kellner* and Kim* et al. Under review
Summary – *C. roseus* genome study

- 1 library, single lane Illumina pair-end sequencing
- ~52x coverage
- High quality genome assembly
  - ~70% genome assembled
  - N50 27.3kb
- Enriched gene space
- All known MIA genes identified, except SGD
Summary – *C. roseus* genome study

- Novel candidate gene discovery
  - Co-expression
  - Physical cluster (with BAC study)
- Gene expansion and neo-functionalization (ex. STR, ORCA TFs)
- Transcriptional regulatory network
- Release v1 soon!
- Release v2 will include mate pair library

*Genome guided plant specialized metabolism study!*
Acknowledgement

**Buell group (MSU)**
Robin Buell
John Hamilton
Kevin Childs
Brieanne Vaillancourt
Jason Cepela

**O’Connor group (JIC, UK)**
Sarah O’Connor
Franziska Kellner

**The Genome Analysis Center, UK**