Systematic Approach to Identify Genes Controlling Capsaicinoid Content in *Capsicum* spp.

May 25, 2019
Seoul National University
Byoung-Cheorl Kang
Overview of my talk

• Introduction

• Identification of a novel Pungency Controlling locus (*Pun3*) in *Capsicum*

• QTL mapping and GWAS for capsaicinoid content
Capsicum (hot and sweet peppers)

- Domestication: about 6,000 years ago in Mexico
- The most important vegetable crop in Asian countries
  (The 3rd most important vegetable crop in the world)
- Consists of approximately 22 wild species and five domesticated species
  (C. annuum, C. chinense, C. frutescens, C. pubescens, C. baccatum)
- Use: spices, vegetables, ornamental, pharmaceutical and industrial applications
- A member of the family Solanaceae that includes tomato, potato, eggplant, tobacco, and petunia
- Genome size: about 3.5 Gb (x=12)
- Molecular markers are available for major traits
- Genome sequence and high density maps are available
Capsicum spp. reference genomes

Capsicum annuum ‘CM334’ (Nature Genetics, 2014)
  • Resistant to *P. capsici*
  • Resistant to nematode (*Me7*)
  • Resistant potyviruses (*Pvr4*), CMV, PVX

Capsicum annuum ‘Zunla-1’ (PNAS, 2014)
Capsicum annuum ‘Chitepin’

Capsicum chinence ‘PI159236’ (Genome Biology, 2017)
  • Resistant to viruses

Capsicum baccatum ‘PBC81’
  • Resistant to anthrancnose

Capsicum annuum ‘CM334’ (Hortic Research, 2018)
Capsicum annuum ‘Dempsey’ (Unpublished)
**Disease resistance**
Phytophthora, CMV, TSWV, Potyvirus, Nematode, Geminivirus, Anthracnose

**Secondary metabolites**
Capsaicinoid, carotenoids, anthocyanin

**Genetic resources**
RIL, mutant, CC, MAGIC

**Gene cloning & QTL mapping**
Fruit size, yield, secondary metabolites

**Genomics tool**
GBS, Targeted sequencing, RENseq, whole genome sequencing

**GWAS & GS**
Fruit size, yield, secondary metabolites
Capsaicinoid

- Pungency causing substance in pepper (*Capsicum* spp.)
- Unique characteristic only in the genus *Capsicum*
- Pharmacological effects:
  - Anticancer
  - Antioxidant
  - Pain relief
  - Obesity control

![Chemical structures of capsaicinoids](image)

- **Capsaicin**
- **Dihydrocapsaicin**
- **Nordihydrocapsaicin**
- **n-vanillylnonanamide**

*TRPV1* painful, burning sensation
Accumulation of capsaicinoid

• Capsaicinoid is accumulated in **placenta and interlocular septum**

Overview of my talk

• Introduction

• Identification of a novel Pungency Controlling locus (Pun3) in Capsicum

• QTL mapping and GWAS for capsaicinoid content
Structural genes controlling pungency

• Mutations in the Pun1 gene (capsaicin synthase) result in nonpungency
  (Stewart & Kang 2005, Lee et al., 2005; Aza-González et al, 2011; Stellari et al, 2010)

• Mutations in the p-AMT gene greatly reduce pungency level
  (Lang et al., 2009; Tanaka et al., 2018)

✓ The positive correlation of expression of structural genes and the level of pungency → regulated and coordinated at by transcription factors.
Non-pungent accession ‘YCM334’

- *C. annuum* ‘YCM334’ × *C. annuum* ‘Tean’ (165 RILs)
  - Both parents have the *Pun1/Pun1* genotype

![Red peppers](image1.png)

*C. annuum* ‘YCM334’ × *C. annuum* ‘Tean’
(Derived from CM334 and Yolo Wonder)

![Graph of Capsaicinoid content](image2.png)

Number of samples

- YCM334
- Tean
Sequence and expression of *Pun1*

- No sequence differences between ‘YCM334’ and ‘Tean’
- No expression of *Pun1* in ‘YCM334’

(Han et al, TAG 2019)
Mapping the *Pun 3* locus

**MYB transcription factors**

- One of the largest families in plants
- Regulates different biological processes:
  - primary and secondary metabolism
  - biotic and abiotic stresses
  - developmental process
  - hormonal responses

(Han et al, TAG 2019)
Candidate genes for *Pun3*

Expression of candidate genes from ‘Zunla-1’ and ‘PI159236’.

Phylogenetic tree based on CDS sequences of candidate genes from ‘Zunla-1’ and ‘PI159236’.

**YCM334 pun3 allele**

- **133 bp**
- **130 bp**
- **484 bp**
- **249 aa**

(Han et al, TAG 2019)
Germplasm survey of the *pun3* allele

<table>
<thead>
<tr>
<th>Species</th>
<th>Accessions</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>C. annuum</em></td>
<td>218</td>
</tr>
<tr>
<td><em>C. baccatum</em></td>
<td>47</td>
</tr>
<tr>
<td><em>C. chinense</em></td>
<td>46</td>
</tr>
<tr>
<td><em>C. frutescens</em></td>
<td>25</td>
</tr>
<tr>
<td>Others</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>351</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pun1 genotype</th>
<th>Pun3 genotype</th>
<th>Pungency</th>
<th># accessions</th>
<th>Capsaicinoids contents (ug/gDW)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Pun1</em></td>
<td>-</td>
<td>P</td>
<td>131</td>
<td>926.2</td>
</tr>
<tr>
<td><em>Pun3</em></td>
<td>NP</td>
<td>29</td>
<td></td>
<td>29.1</td>
</tr>
<tr>
<td><em>pun3</em></td>
<td>NP</td>
<td>1(1)</td>
<td></td>
<td>ND (-)</td>
</tr>
<tr>
<td><em>pun1</em></td>
<td>-</td>
<td>NP</td>
<td>52</td>
<td>4.3</td>
</tr>
</tbody>
</table>
RNA-Seq analysis of ‘YCM334’ and ‘Tean’

Key structural genes (*BCAT*, *KasI*, *FatA*, *Ca4H*, and *Pun1*) for capsaicinoid biosynthesis were down regulated in YCM334.

(Han et al, TAG 2019)
**Pun3 and cap7.2**

- **Cap7.2**: QTL controlling capsaicinoid content in two interspecific populations
- **Pun2**: a locus controlling presence of capsaicinoid in *C. chacoense* (Stellari et al., 2010)
Expression patterns of capsaicinoid biosynthetic genes Kas, AT3, Comt, Ca4H, and pAMT in placental tissue from California Wonder (nonpungent) were greatly reduced.

Gene silencing of CaMYB31 resulted in reduction of structural gene expression and capsaicinoid content in fruits of hot pepper.

Natural variations in the MYB transcription factor MYB31 determine the evolution of extremely pungent peppers.
Summary of part 1

- Non-pungent *C. annuum* ‘YCM334’ has functional *Pun1* sequence, but the gene was not expressed. We named this gene *Pun3*.

- *Pun3* was mapped on chromosome 7 near *Pun2* and the QTL locus *cap7.2*.

- *Pun3* encodes R2R3 MAB transcription factor and a dysfunctional allele of *Pun3* down-regulated most of key capsaicinoid biosynthetic genes in YCM334.
Overview of my talk

• Introduction

• Identification of a novel Pungency Controlling locus (Pun3) in Capsicum

• QTL mapping and GWAS for capsaicinoid content
Variation of pungency in *Capsicum* spp.

<table>
<thead>
<tr>
<th>Scoville heat units (SHU)</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>15,000,000</td>
<td>Pure Capsaicin</td>
</tr>
<tr>
<td>2,000,000-5,300,000</td>
<td>Carolina Reaper, Pepper X</td>
</tr>
<tr>
<td>1,500,000-2,000,000</td>
<td><strong>Trinidad Moruga Scorpion</strong></td>
</tr>
<tr>
<td>855,000-1,463,700</td>
<td>Infinity Chilli, <strong>Bhut Jolokia</strong> chili pepper</td>
</tr>
<tr>
<td>350,000-580,000</td>
<td>Red Savina habanero</td>
</tr>
<tr>
<td>100,000-350,000</td>
<td><strong>Habanero</strong>, Scotch bonnet pepper, Datil pepper</td>
</tr>
<tr>
<td>50,000-100,000</td>
<td>Bird's eye chili (<strong>Thai Chili</strong>), Malagueta pepper</td>
</tr>
<tr>
<td>30,000-50,000</td>
<td>Cayenne pepper, Tabasco pepper, Cumari pepper</td>
</tr>
<tr>
<td>10,000-23,000</td>
<td>Serrano pepper, Peter pepper, Aleppo pepper</td>
</tr>
<tr>
<td>3,500-8,000</td>
<td>Espelette pepper, Jalapeño pepper, Chipotle</td>
</tr>
<tr>
<td>1,000-2,500</td>
<td>Anaheim pepper, Poblano pepper, Rocotillo pepper</td>
</tr>
<tr>
<td>100-900</td>
<td>Pimento, Banana pepper</td>
</tr>
<tr>
<td>No significant heat</td>
<td><strong>Bell pepper (Early Calwonder: ECW)</strong>, Aji Dulce</td>
</tr>
</tbody>
</table>

Genes controlling capsaicinoid content are largely unknown!

(wikipedia.org/wiki/Scoville_scale)
QTL mapping in peppers

Importance of quantitative traits
- Molecular breeding has been focused on monogenic traits
- Agriculturally important traits including yield, contents of secondary metabolites and disease resistance are under the control of quantitative genes
- Studies on quantitative traits are very limited

Opportunities & challenges
- Diverse molecular markers have been developed
- Whole-genome sequences are available
- Difficulties in improving quantitative traits
# QTL studies for capsaicinoid content

<table>
<thead>
<tr>
<th>Cross (Species)</th>
<th>Population (No. of plants)</th>
<th>No. of QTL</th>
<th>QTL location</th>
<th>Marker type</th>
<th>R² (%)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maor (<em>C. annuum</em>) × BG 2816 (<em>C. frutescens</em>)</td>
<td>$F_2$ (242)</td>
<td>1</td>
<td>Chr. 7</td>
<td>400 RAPD</td>
<td>34-38</td>
<td>Blum et al, 2003</td>
</tr>
<tr>
<td>NuMex RNaky (<em>C. annuum</em>) × BG2814-6 (<em>C. frutescens</em>)</td>
<td>$F_3$ (234)</td>
<td>6</td>
<td>Chr. 3, 4, 7</td>
<td>728 SSR, AFLP, and RFLP</td>
<td>9-25</td>
<td>Ben-Chaim et al, 2006</td>
</tr>
<tr>
<td>NuMex RNaky (<em>C. annuum</em>) × BG2814-6 (<em>C. frutescens</em>)</td>
<td>RIL (105)</td>
<td>12</td>
<td>Chr. 3, 4, 5, 6, 7, 10, 11</td>
<td>16,188 unigenes</td>
<td>9-26</td>
<td>Yarnes et al, 2013</td>
</tr>
<tr>
<td>NB1 (<em>C. annuum</em>) × Bhut Jolokia (<em>C. chinense</em>)</td>
<td>$F_2$ (175)</td>
<td>4</td>
<td>Chr. 2, 3, 6</td>
<td>234 HRM, SSR, CAPS, and gene-based markers</td>
<td>9-19</td>
<td>Lee et al, 2016</td>
</tr>
</tbody>
</table>

- Improving capsaicinoid content in pepper varieties is very challenging

- Comparison of QTLs from different studies was difficult
- Markers were far from QTLs due to low-density genetic map
- No candidate genes under QTLs are known.

Limitations
Overview of our approaches

Pungency in placenta
- Two RIL populations
- Genotyping
  - whole genome resequencing
  - genotype-by-sequencing
- Phenotyping
  - multi-year & multi location
  - Placenta tissues
  - HPLC

Pungency in pericarp
- Two F2 populations
- Genotyping and expression
  - genotype-by-sequencing
  - transcriptome sequencing
- Phenotyping
  - multi-year & multi location
  - Pericarp tissues
  - HPLC
(Park et al., TAG, 2018)

Pungency in whole fruit
- GWAS population
- Genotyping
  - genotype-by-sequencing
- Phenotyping
  - Whole fruits
  - HPLC

(Han et al., Plant Biotech J, 2018)
(Han et al., TAG, 2019)
QTL mapping in *Capsicum*

**Experimental population**
- Biparental pop
- Core collection

**Genotype**
- Proper sequencing method for pepper
- GBS

**Phenotype**
- Interesting traits useful for breeding
- Plant architecture
- Pungency
- Fruit color
- Plant disease

**Association mapping**
- Diverse model approach
- CIM, MIM

**Validation**
- Using bi-parental population and previous researches
- General Linear Model (GLM)
- Mixed Linear Model (MLM)

**TASSEL 5.0**
## RIL populations

<table>
<thead>
<tr>
<th>Cross combination</th>
<th>Generation</th>
<th># of lines</th>
<th>Trait</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>C. annuum</em> ‘Long sweet’ x <em>C. annuum</em> ‘AC2212’</td>
<td>F$_{6-7}$</td>
<td>207</td>
<td>Vit C and E Fruit color</td>
</tr>
<tr>
<td><em>C. annuum</em> ‘35001’ x <em>C. annuum</em> ‘35009’</td>
<td>F$_{7-9}$</td>
<td>176</td>
<td>Carotenoid content</td>
</tr>
<tr>
<td><em>C. annuum</em> ‘Perennial’ x <em>C. annuum</em> ‘Dempsey’</td>
<td>F$_{6-11}$</td>
<td>166</td>
<td>Pungency Morphological traits</td>
</tr>
<tr>
<td><em>C. annuum</em> ‘TF68’ x <em>C. chinense</em> ‘Habanero’</td>
<td>F$_{10-11}$</td>
<td>92</td>
<td>Pungency Yield</td>
</tr>
<tr>
<td><em>C. annuum</em> ‘YCM334’ x <em>C. annuum</em> ‘Tean’</td>
<td>F$_{9-11}$</td>
<td>176</td>
<td>Pungency <em>Phytophthora</em> resistance</td>
</tr>
<tr>
<td><em>C. annuum</em> ‘CM334’ x <em>C. annuum</em> ‘ECW30R’</td>
<td>F$_{6-7}$</td>
<td>250</td>
<td>Disease resistance</td>
</tr>
</tbody>
</table>
Genotype-by-sequencing

- **Sequencing**
  - Whole-genome skim sequencing
    - Sequencing $18\times$ for parents and $1\times$ for RILs
  - Genotyping-by-sequencing
    - Restriction enzyme digestion ($PstI/MseI$ & $EcoRI/MseI$)

- **Genotyping**
  - SNP calling
    - *C. annuum* ‘CM334’ reference genome
    - BWA-mem, SAMtools, Picard, and GATK UnifiedGenotyper
    - SNPs with minimum QUAL 30 and depth 3

- **Bin map**
  - Sliding window approach
    - Imputation of missing data
    - Window length: physical length (PD RIL) or SNP number (TH RIL)

- **Association analysis**
  - Identification of loci controlling traits
  - Composite interval mapping
    - WinQTL Cartographer
Imputation by sliding window approach

Sequencing / genotyping error

Missing data

Bin map

Increased accuracy

Useful for QTL analysis

(Huang et al, 2009)
Whole-genome skim sequencing

Bin map of ‘120 PD RILs’

<table>
<thead>
<tr>
<th>Chr.</th>
<th>Number of SNPs</th>
<th>Number of bins</th>
<th>Physical length of bin (Mb)</th>
<th>Genetic distance of bin (cM)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>Total</td>
</tr>
<tr>
<td>1</td>
<td>82,966</td>
<td>390</td>
<td>0.7</td>
<td>272.6</td>
</tr>
<tr>
<td>2</td>
<td>80,141</td>
<td>207</td>
<td>0.8</td>
<td>171.1</td>
</tr>
<tr>
<td>3</td>
<td>87,793</td>
<td>279</td>
<td>0.9</td>
<td>257.9</td>
</tr>
<tr>
<td>4</td>
<td>54,657</td>
<td>224</td>
<td>1.0</td>
<td>222.5</td>
</tr>
<tr>
<td>5</td>
<td>82,413</td>
<td>201</td>
<td>1.2</td>
<td>233.4</td>
</tr>
<tr>
<td>6</td>
<td>107,015</td>
<td>234</td>
<td>1.0</td>
<td>236.9</td>
</tr>
<tr>
<td>7</td>
<td>84,339</td>
<td>180</td>
<td>1.3</td>
<td>231.9</td>
</tr>
<tr>
<td>8</td>
<td>24,383</td>
<td>224</td>
<td>0.6</td>
<td>144.8</td>
</tr>
<tr>
<td>9</td>
<td>275,842</td>
<td>179</td>
<td>1.4</td>
<td>252.7</td>
</tr>
<tr>
<td>10</td>
<td>230,360</td>
<td>160</td>
<td>1.5</td>
<td>233.6</td>
</tr>
<tr>
<td>11</td>
<td>252,765</td>
<td>202</td>
<td>1.3</td>
<td>259.7</td>
</tr>
<tr>
<td>12</td>
<td>68,540</td>
<td>233</td>
<td>1.0</td>
<td>235.7</td>
</tr>
<tr>
<td>Total</td>
<td><strong>1,431,214</strong></td>
<td><strong>2,713</strong></td>
<td>1.0</td>
<td><strong>2,752.8</strong></td>
</tr>
</tbody>
</table>

(DNA Resear Han et al., 2016)
## High density linkage maps in *Capsicum*

<table>
<thead>
<tr>
<th>Population</th>
<th>Sequencing</th>
<th>Number of SNP</th>
<th>Genetic map size (cM)</th>
<th>Average distance of markers (cM)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>C. annuum</em> ‘Perennial’ x <em>C. annuum</em> ‘Dempsey’ RIL</td>
<td>Whole genome resequencing</td>
<td>1,431,214</td>
<td>1,358</td>
<td>0.5</td>
</tr>
<tr>
<td><em>C. annuum</em> ‘MicroPep’ x <em>C. annuum</em> ‘Jeju’ F₂</td>
<td>GBS (PstI/Msel)</td>
<td>2,612</td>
<td>1,731</td>
<td>1.8</td>
</tr>
<tr>
<td><em>C. annuum</em> ‘TF68’ x <em>C. chinense</em> ‘Habanero’ RIL</td>
<td>GBS (PstI/Msel)</td>
<td>8,587</td>
<td>1,127</td>
<td>1.0</td>
</tr>
<tr>
<td><em>C. annuum</em> ‘YCM334’ x <em>C. annuum</em> ‘Tean’ RIL</td>
<td>GBS (PstI/Msel, EcoRI/Msel)</td>
<td>2,335</td>
<td>1,063</td>
<td>1.2</td>
</tr>
<tr>
<td><em>C. annuum</em> ‘ECW30R’ x <em>C. annuum</em> ‘CM334’ F₆</td>
<td>GBS (PstI/Msel)</td>
<td>4,621</td>
<td>1,537</td>
<td>2.7</td>
</tr>
<tr>
<td><em>C. chinense</em> ‘Habanero’ x <em>C. chinense</em> ‘Jolokia’ F₂</td>
<td>GBS (PstI/Msel)</td>
<td>217,152</td>
<td>3,654</td>
<td>3.9</td>
</tr>
</tbody>
</table>
Phenotype evaluation

• **Plant growth - more than two environments**

<table>
<thead>
<tr>
<th>Population</th>
<th>Year</th>
<th>No. of RILs</th>
<th>Geological location</th>
<th>Cultivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD RIL</td>
<td>2011</td>
<td>170</td>
<td>Anseong</td>
<td>Ground</td>
</tr>
<tr>
<td></td>
<td>2012a</td>
<td>162</td>
<td>Anseong</td>
<td>Ground</td>
</tr>
<tr>
<td></td>
<td>2012b</td>
<td>162</td>
<td>Suwon</td>
<td>Pot</td>
</tr>
<tr>
<td>TH RIL</td>
<td>2013</td>
<td>102</td>
<td>Anseong</td>
<td>Ground</td>
</tr>
<tr>
<td>GWAS</td>
<td>2014</td>
<td>104</td>
<td>Anseong</td>
<td>Ground</td>
</tr>
<tr>
<td></td>
<td>2016</td>
<td>351</td>
<td>Jeonjoo</td>
<td>Ground</td>
</tr>
<tr>
<td></td>
<td>2017</td>
<td>351</td>
<td>Jeonjoo</td>
<td>Ground</td>
</tr>
</tbody>
</table>

• **Evaluation of capsaicinoid content**
  – Dissection of placenta (or whole fruits) from 5 fruits per each RIL
  – Measurement of capsaicin and dihydrocapsaicin contents by HPLC
Capsaicinoid content of RILs

- Transgressive segregation except 2012b of ‘PD RIL’
- Positive skew
  - Epistasis between QTLs
Common QTLs of both RILs

- Comparison physical locations of QTLs
- 5 QTLs on chromosome 1, 2, 3, 4, and 10 were detected commonly in both RILs

(Han et al., Plant Biotech J 2018)
Core collection

C. annuum accessions with one C. galapagoense mostly from Europe countries

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C. annuum</td>
<td>218</td>
</tr>
<tr>
<td>C. baccatum</td>
<td>47</td>
</tr>
<tr>
<td>C. chinense</td>
<td>46</td>
</tr>
<tr>
<td>C. frutescens</td>
<td>25</td>
</tr>
<tr>
<td>others</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>351</strong></td>
</tr>
</tbody>
</table>

Most C. baccatum, and C. frutescens accessions

C. baccatum, C. frutescens, C. pubescens, C. cardenasi, C. chacoense, C. eximium, C. praetermissum, C. tovarii accessions mostly from South America and Europe countries

Most C. chinense accessions

C. annuum accessions mostly from Asia countries

(BMC Genetics, Lee et al, 2016)
Capsaicin contents correlation plot (year/metabolite)

(Han et al., Plant Biotech J 2018)
Common loci from QTL and GWAS

**Horticultural Crops Breeding and Genetics Lab**

**pAMT**, putative aminotransferase; **CSE**, caffeoyl shikimate esterase; **4CL**, 4-coumaroyl-CoA ligase; **C4H**, cinnamate 4-hydroxylase; **FatA**, acyl-ACP thioesterase

---

**Chr1**: PD-dicap1.2, PD-dicap1.3, PD-total1.2, TH-dicap1.5, TH-total1.3

**Chr3**: pAMT (CA03g08530), CSE (CA03g24780)

**Chr6**: C4H (CA06g25930 / CA06g25940)

**Chr10**: S10_213596026 (Nimmakayala, 2016)

**Chr11**: 11R (Yarnes, 2013)
### Pun3 MYB transcription factor near the haplotype block (ref. ‘Zunla’)

<table>
<thead>
<tr>
<th>Population</th>
<th>Gene</th>
<th>Chromosome</th>
<th>Start pos (Mbp)</th>
<th>End pos (Mbp)</th>
<th>Length (bp)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>YT RIL</td>
<td><em>Pun3</em></td>
<td>7</td>
<td>~199</td>
<td>~200</td>
<td>Approx. 1M</td>
<td>Han et al. (2019)</td>
</tr>
</tbody>
</table>

(Hong et al., unpublished)
Box plots of capsaicinoid content by candidate genes

PD RIL

TH RIL

GWAS population

(Han et al., Plant Biotech J 2018)
Candidate genes for capsaicinoid QTL

(Stewart et al., 2007)
Summary of part 2

• High-density bin maps were constructed by WGS and GBS of RILs.

• Five common QTLs were detected on chromosome 1, 2, 3, 4, and 10.

• By comparison with GWAS results, 5 candidate genes in capsaicinoid biosynthetic pathway were identified.

• Candidate genes and associated SNPs can be used for molecular breeding of highly pungent pepper cultivars.
In QTL and GWAS study, we revealed candidate structural genes can affect pungency. The Pun3 gene appears to be a master regulator of structural gene expression. In high pungent peppers, Ankyrin may work as a negative regulator.
Genomic selection model construction (Core Collection – PD RIL)

Flexible strategy

352 lines of core collection

Phenotyping for the interested traits

Genome-wide association study
• About 18,000 SNPs from GBS after filtering

Association mapping

Genomic prediction

Genome-wide prediction (Selection)
• About 18,000 SNPs from GBS after filtering
• Generated the common SNPs for target population

Association level genomic prediction
• Select the SNP markers based on the GWAS result
• Using with these decreased number of SNPs, conduct genomic prediction

Used data information

• **Training Population**
  - Pepper core collection 352 lines
  - 18,663 SNP markers derived from GBS

• **Test Population**
  - ‘PD RIL’ 122 lines (two lines of parents)
  - Genotyping based on low depth re-sequencing

• **Phenotype data**
  - 2012 - 2018 ‘Fruit weight’ and ‘Fruit length’, ‘Capsainoid content’
  - Three replicates, mean values
Acknowledgements

Company Partners: NongWoo Bio, FarmHannong, CJ, LGChem, Asia Seed, Hana Seed, EcoSeed, The Kiban, BioTong, East-West Seed, Enza Zaden, KeyGene

Collaborator: Doil Choi (SNU)

Min of Agriculture: 파프리카사업단, 돌연변이사업단, 채소병리검정사업단, 채소종자사업단, 채소육종센터

RDA: BG사업단, 농업유전자원센터, 차세대BG 식물분자육종사업단, 차세대BG 농생명계놈활용사업단

KRF: 식물분자육종센터, 작물유전체기능연구사업단, 신진학자지원

Horticultural Crops Breeding and Genetics Lab
Thank you for your attention
Genomic selection (GS)

Genomic Selection

- Genomic selection first proposed in early 2000s for animal breeding.
  Meuwissen, T.H.E. et al. (2001)

Dairy cattle

- Genomic selection was most successful for dairy cattle.
- Applied in the traits milk production, cattle body size, fertility.
  Hayes et al. (2009), Moser et al. (2009), Calus (2010)

Chicken

- Genomic selection for poultry.
- Applied in 16 traits related eggs and body.
  Wolc et al. (2011)

Beef cattle

- Genomic selection accuracy was not as good as dairy cattle.
- Applied in the traits daily gain and daily intake.
  Rolf et al. (2010), Mujibi et al. (2011), Calus (2010)

Especially, genomic selection showed good performance on complex traits like sex-limited trait and traits having difficulties to conduct phenotyping.
Cross validation Concept of 5-fold cross validation

Partition patterns 1

Fold 1  | Prediction 1 | Training 2 | Training3 | Training4 | Training5
Fold 2  | Training1   | Prediction 2 | Training3 | Training4 | Training5
Fold 3  | Training1   | Training 2  | Prediction 3 | Training4 | Training5
Fold 4  | Training1   | Training 2  | Training3 | Prediction 4 | Training5
Fold 5  | Training1   | Training 2  | Training3 | Training4 | Prediction 5

Conduct the training and prediction with different partition patterns

Comparison between predicted value and observed value

Proper model selection for the specific traits
## Models used for cross validation

<table>
<thead>
<tr>
<th>Method</th>
<th>Type</th>
<th>Liniearity</th>
<th>Reference</th>
</tr>
</thead>
</table>
Association level genomic prediction for fruit weight and fruit length (Core Collection – PD RIL)

**Used data information**

**Training Population**
- Pepper core collection 352 lines
- Top 100 SNP markers based on GWAS results
- Multi-variable linear regression

**Phenotype data**
- 2018 ‘Fruit weight’ and ‘Fruit length’
- Three replicates, mean values

**Test Population**
- ‘PD RIL’ 122 lines (two lines of parents)
- Genotyping based on low depth re-sequencing
Genomic selection model Construction (Core Collection – PD RIL)

Among the 10 models, chose the ‘gblupRR’, ‘RKHS’, ‘Random Forest’

- Fruit weight (g)
  - Common SNP 18,663 of pepper core collection (training pop.) and PD RIL (test pop.)
  - Three model trained with 284 lines with no missing data
  - Phenotype data was transformed by $\log_{10}$

- Fruit length (cm)
  - Common SNP 18,663 of pepper core collection (training pop.) and PD RIL (test pop.)
  - Three model trained with 285 lines with no missing data
Comparison of prediction values and observed values in test population (PD RIL)

- Fruit weight (g)
  - Prediction data was generated by model ‘gblupRR’.
  - Prediction data was compared to previous observed data of test population (PD RIL) from 2011, 2012a, 2012b, 2014.
  - Mean correlation from four data set was 0.553
Comparison of prediction values and observed values in test population (PD RIL)

- Fruit weight (g)
  - Prediction data was generated by model ‘RKHS’.
  - Prediction data was compared to previous observed data of test population (PD RIL) from 2011, 2012a, 2012b, 2014.
  - Mean correlation from four data set was 0.558
Comparison of prediction values and observed values in test population (PD RIL)

- Fruit weight (g)
  - Prediction data was generated by model ‘Random Forest’.
  - Prediction data was compared to previous observed data of test population (PD RIL) from 2011, 2012a, 2012b, 2014.
  - Mean correlation from four data set was 0.486
Genomic prediction results of capsaicin contents in pepper core collection (10-fold cross validation)

Data information

- **Genotype data**
  - Pepper core collection 352 lines
  - 18,663 SNP markers derived from GBS

- **Phenotype data**
  - Capsaicin contents analyzed by HPLC
  - 242 lines without missing data

- **Validation of genomic selection**
  - 10-fold cross validation with nine genomic selection models
  - RKHS showed the highest accuracy with 0.689

<table>
<thead>
<tr>
<th>Models</th>
<th>Mean of accuracy</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>BayesB</td>
<td>0.649</td>
<td>0.004</td>
</tr>
<tr>
<td>BayesC</td>
<td>0.649</td>
<td>0.004</td>
</tr>
<tr>
<td>BLasso</td>
<td>0.668</td>
<td>0.005</td>
</tr>
<tr>
<td>EBLasso</td>
<td>0.566</td>
<td>0.016</td>
</tr>
<tr>
<td>ElasticNet</td>
<td>0.609</td>
<td>0.016</td>
</tr>
<tr>
<td>gblupRR</td>
<td>0.663</td>
<td>0.010</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.610</td>
<td>0.018</td>
</tr>
<tr>
<td>RandomForest</td>
<td>0.659</td>
<td>0.006</td>
</tr>
<tr>
<td>RKHS</td>
<td><strong>0.689</strong></td>
<td><strong>0.006</strong></td>
</tr>
<tr>
<td>RR</td>
<td>0.663</td>
<td>0.007</td>
</tr>
</tbody>
</table>
Association level genomic prediction for fruit weight and fruit length (Core Collection – PD RIL)

**Flexible strategy**

- **Training Population**
  - Pepper core collection 352 lines
  - Top 100 SNP markers based on GWAS results
  - Multi-variable linear regression
- **Phenotype data**
  - 2018 ‘Fruit weight’ and ‘Fruit length’
  - Three replicates, mean values

**Used data information**

- **Test Population**
  - ‘PD RIL’ 122 lines (two lines of parents)
  - Genotyping based on low depth re-sequencing

**Genomic prediction**

- **Genome-wide prediction (Selection)**
  - About 18,000 SNPs from GBS after filtering
  - Generated the common SNPs for target population

- **Association mapping**
  - Genome-wide association study
    - About 18,000 SNPs from GBS after filtering

- **Association level genomic prediction**
  - Select the SNP markers based on the GWAS result
  - Using with these decreased number of SNPs, conduct genomic prediction
Association level genomic prediction for fruit weight and fruit length (Core Collection – PD RIL)

Select the top 100 SNPs based on the GWAS results.

Fit the selected SNPs with multi-variable linear regression
Association level genomic prediction for fruit weight and fruit length (Core Collection – PD RIL)

Model fitting and prediction for core collection data

Prediction with constructed multi-variable linear model
Association level genomic prediction for fruit weight and fruit length (Core Collection – PD RIL)

Prediction with PD-RIL SNPs

- Fruit weight (g)
  - Data was transformed by $\log_{10}$
  - Prediction data was compared to previous observed data of test population (PD RIL) from 2011, 2012a, 2012b, 2014.
  - Mean correlation from four data set was 0.157