Possibilities and challenges of the potato genome sequence

Richard GF Visser, Wageningen UR Plant Breeding
EAPR 2014, July 6-11, Brussels
Overview

- A sequence is available what can we do with it?
- Sequencing broad or deep?
- Genotyping by Sequencing in Potato
  - Genome re-sequencing in Potato
  - Haplotype reconstruction in polyploids
- Sequencing and breeding
  - BreeDB in Potato
Two potatoes have been sequenced

- DM doubled monoploid has been sequenced
- Heterozygous diploid RH has been partly sequenced
- At best three alleles per gene known for a selection of traits

Is this a problem?

- No: Data sufficient to design SNP markers
- Yes: No info on maximum number of alleles per gene
Deliverables and direct use of sequence

- SNPs; useful for marker development on arrays
- Dense genetic map construction in every type of material (not always same SNPs!)
- Improved QTL mapping
- Enabling link of genetic to physical map
- Potential ID of candidate genes
- Forward and backward searches for gene/trait ID possible
Potato SNP array

- Infinium array ~18,000 SNPs
  - Selected from 129 K genomic DNA sequence variants over 83 cultivars
  - Stringent quality criteria
  - Low redundancy - one-per-gene
  - Including low-frequency SNP alleles
- Mapping populations not involved in choice of SNPs
  - Expecting: enough segregating SNPs for mapping
Samples hybridized with this array

- ~2500 potato samples:
  - Tetraploid mapping population: ~250 individuals
  - Parents 2, x 2 replicates
  - Grandparents (3), great-grandparent (1)
  - >500 diploid accessions
  - ~550 tetraploid cultivars and breeding lines (GWAS)
  - 1100 other tetraploid progenies
More spurious data of different potato genotypes has become available; ~30 tetraploids

Different micro arrays are constructed a.o. SOLCAP and Wageningen Infinium array

~100 million different SNP positions

Difference between exonic and intronic SNP density
Anywhere between 8 to 16 alleles per gene

Nucleotide diversity

- ~1 SNP/50 bp between two alleles
- ~1 SNP/16 bp across alleles
- On average 1 SNP/20 bp
‘Deep’ sequencing

- Necessary to get a ‘perfect’ (template) sequence, without sequencing mistakes
- Will improve the assembly (which is easier if sequence is better)
- Is important for a better gene identification and annotation of genes
- Calls for more ‘cleaning’ of data
‘Broad’ sequencing

- Necessary to get full insight in the number of alleles of a particular gene
- Will allow a better prediction of haplotypes
- Is currently done in a number of studies incl. all tuber bearing potato accessions for a selected number of genes incl. earliness
‘Broad’ sequencing

- Many SNP identified; ca 100 million
- Sequence of one genotype is not always indicative for the other sequence
- Potato = tomato =/= potato
- Analysis awaits further work but already many questions → how many, how deep, what technique?
Alignment of sequences shows large differences

200 kbp

RH BAC mtp 1280

DM reference sequence

RH BAC mtp 3424

32000000

34000000
Genotyping by Sequencing

Squeezing your genome is cheaper than lemonade
Genotyping by sequencing in tetraploids

- Traditional Sanger reads
  - Peak ratio
- Next generation reads
  - Read depth ratio
Sequencing of tetraploid potato

- 800 genes selected in 83 cultivars
- Distributed over the genome (all chromosomes)
- Capture technology (sure select)
- Sequenced > 2 Mb; seq depth 75-100 x
- 130,000 SNPs, 30,000 in coding regions
Haplotype diversity in cultivated potato germplasm

- Many rare alleles
- Many introgressed alleles
- High sequence divergence
- Average 3.2 per genotype

Sequence identity
95.6-99.9% within potato
94% tomato vs. potato

FIGURE 2. Dendogram of the 16 GWD haplotypes. The distances were computed using the Jukes-Cantor method and the tree inferred using the Neighbor-joining method. The percentage of replicate trees in which the associated haplotypes clustered together in the bootstrap test (1000 replicates) are shown next to the branches. For each allele the frequency and – when identified – the source is given. The tomato sequence was used as out-group to root the tree.
Potato Genome (Re-)Sequencing

Identify alleles underpinning phenotypic diversity across the entire genome and entire potato clade
Two different experiments:
1. use tetraploid varieties to assess the allelic variation in the gene encoding plant maturity using PCR cloning and Illumina sequencing.
2. Identify the sequence of the earliness locus CDF in tuber bearing wild species and landraces from north to South America using PACBIO
From 500 tetraploid varieties 180 selected and finally 15 for PCR cloning and sequencing of intron and exon region of ~600 bp

Outcome:

✓ 15 genotypes (varying from 3.5 till 8.3) show 3.4 different alleles
✓ 45 haplotypes
✓ Herald is 1278, Binella is 9,9,10,10
✓ 1 SNP per 8 bp!
✓ How many more haplotypes????

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Cultivars</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Herald</td>
</tr>
<tr>
<td>2</td>
<td>Herald</td>
</tr>
<tr>
<td>3</td>
<td>Ar96</td>
</tr>
<tr>
<td>4</td>
<td>Ar96</td>
</tr>
<tr>
<td>5</td>
<td>Early Rose</td>
</tr>
<tr>
<td>6</td>
<td>Ar96</td>
</tr>
<tr>
<td>7</td>
<td>Herald</td>
</tr>
<tr>
<td>8</td>
<td>Herald</td>
</tr>
<tr>
<td>9</td>
<td>Binella</td>
</tr>
<tr>
<td>10</td>
<td>Binella</td>
</tr>
<tr>
<td>11</td>
<td>IVP4X-144-2</td>
</tr>
<tr>
<td>12</td>
<td>IVP4X-144-2</td>
</tr>
<tr>
<td>13</td>
<td>IVP4X-144-2</td>
</tr>
<tr>
<td>14</td>
<td>IVP92-057-3</td>
</tr>
<tr>
<td>15</td>
<td>Alpha</td>
</tr>
<tr>
<td>16</td>
<td>Alpha</td>
</tr>
<tr>
<td>17</td>
<td>Wur038</td>
</tr>
<tr>
<td>18</td>
<td>Wur038</td>
</tr>
<tr>
<td>19</td>
<td>Wur038</td>
</tr>
<tr>
<td>20</td>
<td>Kepplestone kidney</td>
</tr>
<tr>
<td>21</td>
<td>Wur038</td>
</tr>
<tr>
<td>22</td>
<td>Aurora</td>
</tr>
<tr>
<td>23</td>
<td>Ar96</td>
</tr>
<tr>
<td>24</td>
<td>Ar96</td>
</tr>
<tr>
<td>25</td>
<td>Wur038</td>
</tr>
<tr>
<td>26</td>
<td>Karnico</td>
</tr>
<tr>
<td>27</td>
<td>IVP4X-061-1</td>
</tr>
<tr>
<td>28</td>
<td>Aurora</td>
</tr>
<tr>
<td>29</td>
<td>Early Rose</td>
</tr>
<tr>
<td>30</td>
<td>Civa</td>
</tr>
<tr>
<td>31</td>
<td>Civa</td>
</tr>
<tr>
<td>32</td>
<td>Civa</td>
</tr>
<tr>
<td>33</td>
<td>Alpha</td>
</tr>
<tr>
<td>34</td>
<td>Early Rose</td>
</tr>
<tr>
<td>35</td>
<td>Kerpondy</td>
</tr>
<tr>
<td>36</td>
<td>Alpha</td>
</tr>
<tr>
<td>37</td>
<td>Karnico</td>
</tr>
<tr>
<td>38</td>
<td>IVP92-057-3</td>
</tr>
<tr>
<td>39</td>
<td>Karnico</td>
</tr>
<tr>
<td>40</td>
<td>DM reference</td>
</tr>
<tr>
<td>41</td>
<td>Alpha</td>
</tr>
<tr>
<td>42</td>
<td>Kerpondy</td>
</tr>
<tr>
<td>43</td>
<td>Kerpondy</td>
</tr>
<tr>
<td>44</td>
<td>Karnico</td>
</tr>
<tr>
<td>45</td>
<td>IVP4X-061-1</td>
</tr>
<tr>
<td>46</td>
<td>Aurora</td>
</tr>
</tbody>
</table>
Difference in haplotype is often one SNP only, but more differences occur also.
PacBio results of CDF gene

- About 150 genotypes analysed
- Sequence from one tetraploid genotype
- 2.5 kb after cleaning 1.7 kb
- 2 haplotypes
PacBio results

- 200 bp fragment enlarged
- Still problems in aligning sequences, manual handling
Markers in breeding / research

- **Diploids**
  - Genetic maps; mapping qualitative / quantitative traits
  - Marker assisted introgression, marker-assisted selection, genetic distance/similarity, purity, pedigree relationships, etc.

- **Tetraploids**
  - Allopolyploids: inheritance as in diploid, markers/genes not always unique
  - Autopolyploids: much more complex genetics
Tetraploid analysis methods and software

- SNP genotyping: dosage scoring
  - 5 classes: nulliplex - quadruplex
  - fitTetra (Voorrips et al. 2011)

- Linkage analysis and QTL mapping
  - TetraploidMap (Hackett & Luo, 2003; Hackett et al. TAG 2014) limitations (a.o. max 50 markers / chromosome)
  - New high-throughput tools needed

- Software to simulate tetraploid progenies
  - PedigreeSim (Voorrips & Maliepaard, 2012) for development and validation of new tools
Why so complex?

- **Different dosages of alleles possible**
  - Extra problem if also null alleles are possible

- **Different alleles**: multiple allelism

- **More than 1 allele per locus** passed on to gamete
  - Presence of one allele ≠ absence of the other
Why so complex?

- **More possibilities phase** of linked loci
  - Coupling and three times repulsion in a tetraploid

- **Recombination possible over more than two homologs**
  - Bivalents: preferential or random pairing homologs
  - Quadrivalents: possibility of double reduction
Haplotype inheritance for genes can be followed based on pedigree info

Markies has received two copies of haplotype 4 from Fianna: Gene is near the end of Chr.1 ← example of double reduction.
Linkage map construction

- Map the markers in each parental haplogroup
Research and breeding aims

- Understanding the genetic structure of potato germplasm
  - How many haplotypes are there in cultivars at different loci?
  - What is the best combination of alleles for a given trait?

- What loci are involved in (quantitative) quality traits?
  - Haplotypes with negative/positive effects
  - Haplotypes with dominant/recessive/additive effects
  - -> marker assisted selection

# Allele effects (kinship)

LogP_ kinship = 6.939

<table>
<thead>
<tr>
<th>allele</th>
<th>abs freq</th>
<th>rel freq</th>
<th>Effect_Starch_PO4_nmol</th>
<th>se_Starch_PO4_nmol</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>300</td>
<td>0.414</td>
<td>0.00</td>
<td>0.000</td>
</tr>
<tr>
<td>B</td>
<td>147</td>
<td>0.203</td>
<td>1.01</td>
<td>0.439</td>
</tr>
<tr>
<td>C</td>
<td>62</td>
<td>0.086</td>
<td>1.66</td>
<td>0.671</td>
</tr>
<tr>
<td>D</td>
<td>61</td>
<td>0.084</td>
<td>2.65</td>
<td>0.584</td>
</tr>
<tr>
<td>E</td>
<td>51</td>
<td>0.070</td>
<td>0.01</td>
<td>0.575</td>
</tr>
<tr>
<td>F</td>
<td>65</td>
<td>0.090</td>
<td>-0.20</td>
<td>0.506</td>
</tr>
<tr>
<td>G</td>
<td>4</td>
<td>0.006</td>
<td>0.06</td>
<td>2.213</td>
</tr>
<tr>
<td>H</td>
<td>15</td>
<td>0.021</td>
<td>6.10</td>
<td>1.061</td>
</tr>
<tr>
<td>I</td>
<td>2</td>
<td>0.003</td>
<td>-1.80</td>
<td>2.707</td>
</tr>
<tr>
<td>J</td>
<td>2</td>
<td>0.003</td>
<td>2.30</td>
<td>2.383</td>
</tr>
<tr>
<td>K</td>
<td>15</td>
<td>0.021</td>
<td>1.12</td>
<td>1.075</td>
</tr>
<tr>
<td>L</td>
<td>0</td>
<td>0.000</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
Breeding with databases: BreeDB

Genotyping
- Genotyping-by-sequencing
- SNP arrays

Dosage information
- fitTetra

Optional: Population Structure correction

LD analysis / genomic selection

Marker selection

Field trial
- Observations
- ~omics analysis

Research Focus: in silico cultivar design

Estimate genotypic means

Precision Breeding

BreeDB

Breeding with databases: BreeDB

Field trial
- Observations
- ~omics analysis

Genotyping
- Genotyping-by-sequencing
- SNP arrays

Dosage information
- fitTetra

Optional: Population Structure correction

LD analysis / genomic selection

Marker selection

BreeDB

Research Focus: in silico cultivar design
Precision Breeding

- Well characterized (elite) germplasm collections
- Design the desired genotype behind the computer
  - Breeding goals as input
- Obtain the most effective breeding strategy, starting with your current (elite) germplasm
- Cross (and if necessary genotype the offspring)
- Select candidates similar to the in silico designed line
  - Characterize and add to germplasm collection & Re-train predictive algorithms
Concluding

- Haplotyping is possible but should be automated
- Alignment of NGS sequences is a quite usable approach for haplotyping in tetraploid potato
  - Haplotypes of 1-1.5 kbp length were easily obtained in this project
  - For individual genes, spending time on 'manual' data analysis will significantly improve results
  - For genome-wide analyses, the automated haplotyping results are too irregular and fragmented ....... and it is necessary to use longer sequences → they are slowly comming eg PacBio
- Best allele combination predictions per trait are next challenge
Acknowledgements

**Wageningen UR Plant Breeding:**
Jose Abelenda, Christian Bachem, Jan de Boer, Theo Borm, Herman van Eck, Richard Finkers, Chris Maliepaard, María del Mar Pérez Nicolás, Roeland Voorrips, Peter Vos, Jan Uitdenwilligen, Annemarie Wolters

**Biometris WUR:** Joao Paulo, Fred van Eeuwijk,

**Potato breeding, growing & processing companies**
Agrico, Averis, HZPC, C. Meijer, KWS potato, VAVI, HPA
Sponsors

- Technology foundation STW
- Potato Breeding companies (2002-2012)
- Technological Top Institute - Green Genetics
- Ministry of Economic Affairs, Agriculture & Innovation
- Centre for BioSystems Genomics/NGI
- Wageningen University and Research Centre
Questions?