Molecular markers in plant breeding

Jumbo MacDonald et al.,

MAIZE BREEDERS’ COURSE
Palace Hotel
Arusha, Tanzania
4 Sep to 16 Sep 2016
Key areas to look at

- Molecular Markers
- QTL Mapping
- Association mapping
- GWAS
- Genomic Selection
- Mapping populations
- Linkage
  - Linkage disequilibrium
Genetic Markers

• Genetic differences between organisms
• Neural sites of variations at DNA sequence level
• They act as signs or flags
• Markers in close proximity to genes can be referred to as gene tags
• Such markers themselves do not affect the phenotype of the trait of interest because they are located only near or ‘linked’ to genes controlling the trait.
• All genetic markers occupy specific genomic positions within chromosomes
Types of genetic markers

There are three major types of genetic markers:

• Morphological (also ‘classical’ or ‘visible’) markers which themselves are phenotypic traits or characters;
• Biochemical markers, which include allelic variants of enzymes called isozymes; and
• DNA (or molecular) markers, which reveal sites of variation
• Morphological markers are usually visually characterized phenotypic characters such as flower colour, seed shape, growth habits or pigmentation.
• Isozyme markers are differences in enzymes that are detected by electrophoresis and specific staining.
• The major disadvantages of morphological and biochemical markers are that they may be limited in number and are influenced by environmental factors or the developmental stage of the plant.
DNA Markers

• DNA markers are the most widely used type of marker predominantly due to their abundance.
• They arise from different classes of DNA mutations such as substitution mutations (point mutations), re-arrangements (insertions or deletions) or errors in replication of tandemly repeated DNA.
• Unlike morphological and biochemical markers, DNA markers are practically unlimited in number and are not affected by environmental factors and/or the developmental stage of the plant.
DNA markers

- Apart from the use of DNA markers in the construction of linkage maps, they have numerous applications in plant breeding such as assessing the level of genetic diversity within germplasm and cultivar identity.

- DNA markers may be broadly divided into three classes based on the method of their detection:
  - hybridization-based;
  - polymerase chain reaction (PCR)-based and
  - sequence-based.
DNA markers

• DNA markers may reveal genetic differences that can be visualized by using a technique called gel electrophoresis and staining with chemicals (ethidium bromide or silver) or detection with radioactive or colourimetric probes.

• DNA markers are particularly useful if they reveal differences between individuals of the same or different species.

• These markers are called polymorphic markers, whereas markers that do not discriminate between genotypes are called monomorphic markers.
Polymorphic markers may also be described as codominant or dominant.

This description is based on whether markers can discriminate between homozygotes and heterozygotes (Figure 1).

Codominant markers indicate differences in size whereas dominant markers are either present or absent.

The different forms of a DNA marker (e.g. different sized bands on gels) are called marker ‘alleles’.

Codominant markers may have many different alleles whereas a dominant marker only has two alleles.
• Comparison between (a) codominant and (b) dominant markers. Codominant markers can clearly discriminate between homozygotes and heterozygotes whereas dominant markers do not.

• Genotypes at two marker loci (A and B) are indicated below the gel diagrams.
Types of Markers

• **Hybridization-based molecular markers**
  – RFLP is the most widely used hybridization-based molecular marker.
• Digestion of the DNA with one or more restriction enzyme(s).
• b) Separation of the restriction fragments in agarose gel.
• c) Transfer of separated fragments from agarose gel to a filter by Southern blotting.
• d) Detection of individual fragments by nucleic acid hybridization with a labeled probe(s)
RFLPs

• Restriction enzymes (endonucleases) are bacterial enzymes (e.g., MseI, EcoRI, PstI, etc.) that recognize specific four, six or eight base pair (bp) sequences in DNA, and cleave double-stranded DNA whenever these sequences are encountered. For example,

• EcoRI has six bp recognition sequence and it cuts between G and A whenever the sequences 5’…GAATTC…3’ or 3’…CTTAAG…5’ exist.

• The choice between of the enzymes depends on the resolution needed
RFLPs

- These are then separated by electrophoresis through agarose or polyacrylamide gels. The choice between agarose and polyacrylamide is based on the restriction enzymes chosen.
- Four-cutters produce fragments too small to be resolved by agarose gels; hence, polyacrylamide gels are required.
- Conversely, polyacrylamide gels can not normally be used to resolve the fragments produced by six-cutters so agarose gels must be used.
- These considerations have led to most workers use six-cutter enzymes, as agarose gels are much easier to handle.
PCR based Markers

The various PCR-based techniques are of two types depending on the primers used for amplification:

• 1) Arbitrary or semi-arbitrary primed PCR techniques developed without prior sequence information (e.g., AP-PCR, DAF, RAPD, AFLP, ISSR).

• 2) Site-targeted PCR techniques developed from known DNA sequences (e.g., EST, CAPS, SSR, SCAR, STS).
Types of Markers

A number of factors need to be considered in choosing one or more of the various molecular marker types:

- Marker system availability
- Marker should be informative (Polymorphic Information content-PIC)
- Simplicity of the technique and time availability.
- Anticipated level of polymorphism in the population.
- Quantity and quality of DNA available.
- Transferability between laboratories, populations, pedigrees and species.
- The size and structure of the population to be studied
- Availability of adequate skills and equipment
- Cost per data-point and availability of sufficient funding.
- Marker inheritance (dominant versus codominant) and the type of genetic information sought in the population
Marker type by application

• Foreground markers
  – *Foreground selection*, in which the breeder selects plants having the marker allele of the donor parent at the target locus. The objective is to maintain the target locus in a heterozygous state (one donor allele and one recurrent parent allele) until the final backcross is completed. Then, the selected plants are self-pollinated and progeny plants identified that are homozygous for the donor allele.
Marker type by application

• Background markers
  – *Background selection*, in which the breeder selects for recurrent parent marker alleles in all genomic regions except the target locus, and the target locus is selected based on phenotype. Background selection is important in order to eliminate potentially deleterious genes introduced from the donor. So-called 'linkage drag', the inheritance of unwanted donor alleles in the same genomic region as the target locus, is difficult to overcome with conventional backcrossing, but can be addressed efficiently with the use of markers.
Polymorphic alleles

Alleles

142
170
178
184
193

Genotypes

142/178 Heterozygote
142/170 Heterozygote
170/170 Homozygote
178/178 Homozygote
184/184 Homozygote
193/193 Homozygote

Adopted from Dr. Kassa
## SSRs

<table>
<thead>
<tr>
<th>Code</th>
<th>nc130</th>
<th>phi014</th>
<th>phi029</th>
<th>phi031</th>
</tr>
</thead>
<tbody>
<tr>
<td>TL2012-1</td>
<td>139:142</td>
<td>428:431</td>
<td>152:152</td>
<td>221:221</td>
</tr>
<tr>
<td>TL2012-2</td>
<td>139:139</td>
<td>428:428</td>
<td>152:152</td>
<td>185:189</td>
</tr>
<tr>
<td>TL2012-8</td>
<td>139:139</td>
<td>428:428</td>
<td>148:152</td>
<td>189:221</td>
</tr>
<tr>
<td>SubjectID</td>
<td>csu1171</td>
<td>d8_2</td>
<td>d8_3</td>
<td>lac1_3</td>
</tr>
<tr>
<td>-----------</td>
<td>---------</td>
<td>------</td>
<td>------</td>
<td>--------</td>
</tr>
</tbody>
</table>
## GBS SNP calls - Lots of Missing Data

<table>
<thead>
<tr>
<th>rs#</th>
<th>alleles</th>
<th>chrom</th>
<th>pos</th>
<th>d</th>
<th>RIL_1</th>
<th>RIL_10</th>
<th>RIL_100</th>
<th>RIL_101</th>
<th>RIL_102</th>
<th>RIL_103</th>
<th>RIL_104</th>
<th>RIL_105</th>
</tr>
</thead>
<tbody>
<tr>
<td>S10_13181</td>
<td>T/G</td>
<td>10</td>
<td>13181</td>
<td>+</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td></td>
<td>N</td>
<td>T</td>
<td>N</td>
<td>T</td>
</tr>
<tr>
<td>S10_13355</td>
<td>T/C</td>
<td>10</td>
<td>13355</td>
<td>+</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>N</td>
<td>T</td>
<td>N</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>S10_15605</td>
<td>A/G</td>
<td>10</td>
<td>15605</td>
<td>+</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>N</td>
</tr>
<tr>
<td>S10_15607</td>
<td>A/G</td>
<td>10</td>
<td>15607</td>
<td>+</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>N</td>
</tr>
<tr>
<td>S10_15619</td>
<td>A/G</td>
<td>10</td>
<td>15619</td>
<td>+</td>
<td>N</td>
<td>A</td>
<td>N</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>N</td>
</tr>
<tr>
<td>S10_15629</td>
<td>G/A</td>
<td>10</td>
<td>15629</td>
<td>+</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>S10_15685</td>
<td>C/G</td>
<td>10</td>
<td>15685</td>
<td>+</td>
<td>C</td>
<td>N</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>N</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>S10_15687</td>
<td>G/A</td>
<td>10</td>
<td>15687</td>
<td>+</td>
<td>G</td>
<td>N</td>
<td>G</td>
<td>G</td>
<td>N</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>S10_15699</td>
<td>G/C</td>
<td>10</td>
<td>15699</td>
<td>+</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>N</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>S10_15720</td>
<td>A/G</td>
<td>10</td>
<td>15720</td>
<td>+</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>A</td>
</tr>
<tr>
<td>S10_15721</td>
<td>G/C</td>
<td>10</td>
<td>15721</td>
<td>+</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>G</td>
</tr>
<tr>
<td>S10_15722</td>
<td>A/T</td>
<td>10</td>
<td>15722</td>
<td>+</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>S10_15723</td>
<td>T/C</td>
<td>10</td>
<td>15723</td>
<td>+</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>T</td>
</tr>
<tr>
<td>S10_16315</td>
<td>G/T</td>
<td>10</td>
<td>16315</td>
<td>+</td>
<td>N</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>N</td>
<td>N</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>S10_16419</td>
<td>A/G</td>
<td>10</td>
<td>16419</td>
<td>+</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>S10_16432</td>
<td>C/G</td>
<td>10</td>
<td>16432</td>
<td>+</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>N</td>
<td>N</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>S10_16439</td>
<td>A/G</td>
<td>10</td>
<td>16439</td>
<td>+</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>S10_16497</td>
<td>C/A</td>
<td>10</td>
<td>16497</td>
<td>+</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>N</td>
<td>C</td>
<td>C</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>S10_16498</td>
<td>A/G</td>
<td>10</td>
<td>16498</td>
<td>+</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>N</td>
<td>A</td>
<td>A</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>S10_16499</td>
<td>C/A</td>
<td>10</td>
<td>16499</td>
<td>+</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>N</td>
<td>C</td>
<td>C</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>S10_16573</td>
<td>T/A</td>
<td>10</td>
<td>16573</td>
<td>+</td>
<td>T</td>
<td>T</td>
<td>N</td>
<td>N</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>S10_17505</td>
<td>G/A</td>
<td>10</td>
<td>17505</td>
<td>+</td>
<td>G</td>
<td>G</td>
<td>N</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>S10_17518</td>
<td>G/T</td>
<td>10</td>
<td>17518</td>
<td>+</td>
<td>G</td>
<td>N</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>S10_17528</td>
<td>G/C</td>
<td>10</td>
<td>17528</td>
<td>+</td>
<td>G</td>
<td>G</td>
<td>N</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>S10_17533</td>
<td>C/G</td>
<td>10</td>
<td>17533</td>
<td>+</td>
<td>C</td>
<td>N</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>S10_17550</td>
<td>G/C</td>
<td>10</td>
<td>17550</td>
<td>+</td>
<td>G</td>
<td>N</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>S10_17551</td>
<td>A/C</td>
<td>10</td>
<td>17551</td>
<td>+</td>
<td>A</td>
<td>N</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>S10_17591</td>
<td>G/C</td>
<td>10</td>
<td>17591</td>
<td>+</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>N</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>S10_17593</td>
<td>T/C</td>
<td>10</td>
<td>17593</td>
<td>+</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>N</td>
<td>T</td>
</tr>
<tr>
<td>S10_17613</td>
<td>A/G</td>
<td>10</td>
<td>17613</td>
<td>+</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>N</td>
</tr>
</tbody>
</table>
What do we use the genotypic data for

- Diversity studies
- Quality control
- Mapping
  - QTL
  - Association mapping
  - GWAS
- Marker Deployment
  - Marker assisted backcrossing
  - Forwardbreeding
  - Marker assisted recurrent selection

- Genomic selection
Mapping populations

- Segregating populations  
  - F2s, F3s BCs (Temporary)
- Recombinant inbred lines (RILs) – permanent
- Doubled Haploid lines (permanent)
- Nested Association Mapping panels (NAM)
- Multi-parent advanced generation intercross (MAGC)
Linkage Analysis

• Single-marker analysis (also ‘single-point analysis’) is the simplest method for detecting QTLs associated with single markers.

• The simple interval mapping (SIM) method makes use of linkage maps and analyses intervals between adjacent pairs of linked markers along chromosomes simultaneously, instead of analyzing single markers.

• Composite interval mapping (CIM) has become popular for mapping QTLs.

• This method combines interval mapping with linear regression and includes additional genetic markers in the statistical model in addition to an adjacent pair of linked markers for interval mapping.
Composite interval mapping

3.0 LOD

2 cM interval

Linkage group 1

Cofactors
QTL analysis
GENOMIC SELECTION (GS)
Marker Assisted Selection (MAS)

- Benefits of MAS
  - Higher genetic gain per unit time
  - Increased Reliability
  - Not affected by environmental factors
  - Increased efficiency
  - Traits that come later in the development stage can be scored before
  - Reduced costs?

In case of multi-environment trials
Marker Assisted Selection (MAB)

- Benefits of MAB
  - Reduced Linkage Drag
    - Marker assisted back Crossing
  - Gene pyramiding
    - Resistance genes
  - Marker Assisted breeding of polygenic traits
  - Keeping tract of all genes involved in complex traits
  - Introduction of novel characters
    - Back Cross
  - Effective exploitation of exotic germplasm
How does the QTL work

• QTL-Based Marker Assisted Selection
  • QTLs localized to marker intervals, their effect sizes estimated
  • QTLs ranked by effect size. Those with largest effects declared significant
QTL-Based Marker Assisted Selection:

From Breeders Perspective has it delivered desired results?

- Precision problems in estimating QTL position, genetic effects, false positives and negatives
- Limited proportion of the total genetic variance is captured by the markers
- Bias of estimated effects (overestimation of selected effects—“Beavis effect”)
- Effects too small for detection—Ignoring some variation?
- Often lead to poor response
Genomic Selection (GS) - Concept

- GS is based on utilization of high-density marker application

- GS differs from QTL-based breeding approaches in that it uses all markers in a prediction of performance genomic estimated breeding value (GEBV)
Utilization of GS and its benefits

• GS has advantage of increasing genetic gain by reducing cycle time

• Reduce phenotyping cost by predicting GEBVs of untested lines

• Filtering bulk of lines in stage 1 trials before advancing them to next level

• More accuracy to capture variation by including alleles with minor effects apart from those alleles with major effect
Genomic Selection (GS)

Two steps:

• Estimation of the effects of chromosome segments in a reference population and,

• Prediction of Genomic Estimated Breeding Values (GEBVs) not in the reference population (selection candidates)

  • QTL are in Linkage Disequilibrium (LD) with a marker or haplotype of markers
LD: Non Random association of Alleles = $r^2$
Genomic Selection (GS)

1. In a **training population** (both genotypic and phenotypic data available), fit a large number of markers as random effects in a linear model to estimate all genetic effects simultaneously for a quantitative trait. The aim is to capture all of the additive genetic variance due to alleles with both large and small effects on the trait.

2. In a breeding population (only genotypic data available), use estimates of marker effects to predict breeding values and select individuals with the best GEBVs.
GS: Predicting Using Many Markers

Meuwissen et al. 2001 Genetics 157:1819-1829
Summary of GS Scheme

Genotyping by Sequencing (GBS)

1. DNA extraction
2. Sequencing (GBS)
3. Allele calls
4. SNPs
5. Imputation (depending on statistical model)
6. Statistical Models
7. Analysis
Linear Mixed Models & Bayesian estimation of many QTL effects, set as random effects, can be estimated simultaneously.

Simple basic model:
\[ Y = 1 \mu + Zg + e \]

- \( Y \) = Data vector
- \( 1 \) = vector of ones (\( n \) = records)
- \( Z \) = design matrix
- \( g \) = genetic effects to be estimated
- \( e \) = vector of residuals
Ridge Regression BLUP

Equal variance of marker effects:
- Overcomes the problem of over-estimation of segment effects by shrinking estimates towards the mean

Problem
- treats all effects equally across all loci, whereas in fact many markers have negligible effects
- However ridge regression may still perform reasonably well in the context of estimating genomic breeding values, as the effects are accumulated across many segments.
Genomic Selection (GS)-Linear Models

- **Bayesian methods**
  Different variance for each marker
  Captures prior knowledge that there are some chromosome segments containing QTL of large effects, some segments with moderate to small effects, and some segments with no QTL at all when estimating the effects of haplotypes (or single markers) within the chromosome segments.
Genomic Selection (GS)-Linear Models

Bayesian Shrinkage Regression-Bayes A (Meuwissen et al),

- Assumption: marker variance = inverse chi-square distribution.

Bayesian Variable Selection - Bayes B, Bayes Cpi.

- Assumptions: marker variance = inverse chi-square distribution. Some marker values are zero.

Mark E. Sorrells, Jessica Rutkoski, Elliot Heffner and Long-Xi Yu
Genomic Selection (GS)-Statistical Models

Kernel Regression & Reproducing Kernel Hilbert Spaces (RKHS) Regression (parameters control complexity of the distribution of the QTL effects) (Gianola et al) Model performance is based on correlation between GEBV and True Breeding Value (TBV)

G-BLUP method same as RKSH
Genomic Selection (GS)-Statistical Models

G-BLUP method—same as RKSH

- Equal variance for marker effects
- Model performance is based on correlation between GEBV and True Breeding Value (TBV)
- Uses genotypic data for G-matrix file used for prediction
- No need for imputation of genotypic data
Proof of Concept Experiments in Maize-GBS

• 2,300 $S_4$ lines were genotyped and their testcrosses phenotyped
  • Phenotypic testcross data from 154 trials was assembled
• 700 SYN$F_2$ lines (Group A & B) have been genotyped. Their testcrosses are being phenotyped
• 19 bi-parental Populations
Proof of Concept Experiments in Maize-GBS

- For stage 1 & 2 testcrosses, we are trying to analyze within tester
- Within Management (optimal, managed drought, managed low nitrogen and random drought)
- Stage 1 predict stage 2
- Cross validations within trials
- For 19 bi-parental Populations, we are trying Bayesian models
# G-BLUP CROSS VALIDATIONS IN MANAGED DROUGHT TRIALS

<table>
<thead>
<tr>
<th>Trial Name</th>
<th>Number of Lines in Trial (Validation)</th>
<th>Number of Lines Training Set</th>
<th>Correlation Training Set</th>
<th>Correlation Validation Set</th>
<th>$H^2$</th>
<th>Correlation Validation/$\sqrt{H^2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILS2-TC-2-1</td>
<td>74</td>
<td>1435</td>
<td>0.8658</td>
<td>0.3516</td>
<td>0.39373</td>
<td>0.560337</td>
</tr>
<tr>
<td>ILS2-TC-3-1</td>
<td>64</td>
<td>1445</td>
<td>0.8661</td>
<td>0.037</td>
<td>0.11492</td>
<td>0.109145</td>
</tr>
<tr>
<td>ILS2-TC-4-1</td>
<td>85</td>
<td>1424</td>
<td>0.8694</td>
<td>0.1734</td>
<td>0.46992</td>
<td>0.252951</td>
</tr>
<tr>
<td>ILS2-TC-5-1</td>
<td>84</td>
<td>1425</td>
<td>0.863</td>
<td>0.0935</td>
<td>0.17252</td>
<td>0.225108</td>
</tr>
<tr>
<td>3WHYB-2010-15-1</td>
<td>183</td>
<td>1326</td>
<td>0.8528</td>
<td>0.131</td>
<td>0.70445</td>
<td>0.15608</td>
</tr>
<tr>
<td>TK-LXT-1-7</td>
<td>87</td>
<td>1422</td>
<td>0.8655</td>
<td>0.1066</td>
<td>0.37751</td>
<td>0.173497</td>
</tr>
<tr>
<td>DTMA-MARS-EVALTC-01-1</td>
<td>182</td>
<td>1327</td>
<td>0.8607</td>
<td>0.0581</td>
<td>0.05704</td>
<td>0.243269</td>
</tr>
<tr>
<td>DTMA-MARS-EVALTC-02-1</td>
<td>198</td>
<td>1311</td>
<td>0.8789</td>
<td>0.1982</td>
<td>0.28514</td>
<td>0.371171</td>
</tr>
<tr>
<td>DTMA-MARS-EVALTC-03-1</td>
<td>225</td>
<td>1284</td>
<td>0.8764</td>
<td>0.1425</td>
<td>0.29932</td>
<td>0.260464</td>
</tr>
<tr>
<td>EIHYB-2011-1-3</td>
<td>73</td>
<td>1436</td>
<td>0.8654</td>
<td>0.0998</td>
<td>0.33892</td>
<td>0.171428</td>
</tr>
<tr>
<td>EIHYB-2011-2-3</td>
<td>69</td>
<td>1440</td>
<td>0.8669</td>
<td>0.293</td>
<td>0.60457</td>
<td>0.376829</td>
</tr>
<tr>
<td>EIHYB-2011-3-3</td>
<td>72</td>
<td>1437</td>
<td>0.8678</td>
<td>0.1872</td>
<td>0.35835</td>
<td>0.312717</td>
</tr>
<tr>
<td>EIHYB-2011-4-3</td>
<td>63</td>
<td>1446</td>
<td>0.8659</td>
<td>0.438</td>
<td>0.49604</td>
<td>0.621893</td>
</tr>
<tr>
<td>3WHYB-2011-19-1</td>
<td>50</td>
<td>1459</td>
<td>0.8651</td>
<td>0.2198</td>
<td>0.44242</td>
<td>0.330453</td>
</tr>
<tr>
<td>Name of Trial</td>
<td>Number of Lines in Trial (Validation)</td>
<td>Number of Lines (Training Set)</td>
<td>Correlation Training Set</td>
<td>Correlation Validation Set</td>
<td>$H^2$</td>
<td>Correlation Validation/$\sqrt{H^2}$</td>
</tr>
<tr>
<td>------------------------</td>
<td>---------------------------------------</td>
<td>-------------------------------</td>
<td>----------------------------</td>
<td>----------------------------</td>
<td>-------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td>ILS2-TC-2-1</td>
<td>74</td>
<td>1435</td>
<td>0.9793</td>
<td>0.3053</td>
<td>0.39373</td>
<td>0.48655</td>
</tr>
<tr>
<td>ILS2-TC-3-1</td>
<td>64</td>
<td>1445</td>
<td>0.9795</td>
<td>0.0296</td>
<td>0.11492</td>
<td>0.087316</td>
</tr>
<tr>
<td>ILS2-TC-4-1</td>
<td>85</td>
<td>1424</td>
<td>0.9861</td>
<td>0.181</td>
<td>0.46992</td>
<td>0.264036</td>
</tr>
<tr>
<td>ILS2-TC-5-1</td>
<td>84</td>
<td>1425</td>
<td>0.98</td>
<td>0.1812</td>
<td>0.17252</td>
<td>0.436253</td>
</tr>
<tr>
<td>3WHYB-2010-15-1</td>
<td>183</td>
<td>1326</td>
<td>0.9661</td>
<td>0.14</td>
<td>0.70445</td>
<td>0.166803</td>
</tr>
<tr>
<td>TK-LXT-1-7</td>
<td>87</td>
<td>1422</td>
<td>0.9811</td>
<td>0.0478</td>
<td>0.37751</td>
<td>0.077797</td>
</tr>
<tr>
<td>DTMA-MARS-EVALTC-01-1</td>
<td>182</td>
<td>1327</td>
<td>0.9735</td>
<td>0.1183</td>
<td>0.05704</td>
<td>0.49533</td>
</tr>
<tr>
<td>DTMA-MARS-EVALTC-02-1</td>
<td>198</td>
<td>1311</td>
<td>0.9848</td>
<td>0.1785</td>
<td>0.28514</td>
<td>0.334279</td>
</tr>
<tr>
<td>DTMA-MARS-EVALTC-03-1</td>
<td>225</td>
<td>1284</td>
<td>0.9791</td>
<td>0.135</td>
<td>0.29932</td>
<td>0.246755</td>
</tr>
<tr>
<td>EIHYB-2011-1-3</td>
<td>73</td>
<td>1436</td>
<td>0.9796</td>
<td>0.2164</td>
<td>0.33892</td>
<td>0.371714</td>
</tr>
<tr>
<td>EIHYB-2011-2-3</td>
<td>69</td>
<td>1440</td>
<td>0.9787</td>
<td>0.3325</td>
<td>0.60457</td>
<td>0.42763</td>
</tr>
<tr>
<td>EIHYB-2011-3-3</td>
<td>72</td>
<td>1437</td>
<td>0.9788</td>
<td>0.2832</td>
<td>0.35835</td>
<td>0.473085</td>
</tr>
<tr>
<td>EIHYB-2011-4-3</td>
<td>63</td>
<td>1446</td>
<td>0.9795</td>
<td>0.5104</td>
<td>0.49604</td>
<td>0.72469</td>
</tr>
<tr>
<td>3WHYB-2011-19-1</td>
<td>50</td>
<td>1459</td>
<td>0.9789</td>
<td>0.325</td>
<td>0.44242</td>
<td>0.488614</td>
</tr>
</tbody>
</table>
Factors Affecting the Accuracy of GEBVs

• Level and distribution of LD between markers and QTL
  – $R^2 > 0.2$ desirable, but more markers increase accuracy
  – Meuwissen 2009: Minimum number of markers for across family = $Ne^*L$ where $Ne$ is the effective population size and $L$ is the genome size in Morgans

Mark E. Sorrells, Jessica Rutkoski, Elliot Heffner and Long-Xi Yu
Factors Affecting the Accuracy of GEBVs

• Distribution of QTL effects
  – Many small effect QTL or low LD favor BLUP for capturing small effect QTL that may not be in LD with a marker
  – Prediction based on relationship decays faster than prediction based on LD (Habier et al 2007; Zhong et al 2009).

• Inbreeding Mendelian Sampling Term
  – Selection for favorable, low frequency alleles and against inbreeding

Mark E. Sorrells, Jessica Rutkoski, Elliot Heffner and Long-Xi Yu
Thank you for your interest!