

IBERS

Sefydliad y Gwyddorau Biolegol, Amgylcheddol a Gwledig
Institute of Biological, Environmental and Rural Sciences

Prospects for using genomic selection in plant breeding

Leif Skøt

Plant breeding

- Success depends on
 - **Clear objectives**
 - **Genetic variation for target traits that can be easily measured**
 - **Obtaining lots of high quality data**
 - **Minimising generation time**
 - **Effective pollen control**

IBERS Grass Breeding

Ryegrasses = over 80% of UK grass seed market

Type

Italian ryegrass

Festulolium

Hybrid ryegrass

Perennial ryegrass
(intermediate heading)

Perennial ryegrass
(late heading)

Recommended use

Late sowing after cereals

Late sowing after cereals

Silage and with red clover

Grazing and/or silage*

Grazing and/or silage*

* With or without white clover



**GERMINAL
HOLDINGS**

www.germinal.com

IBERS Grass Breeding

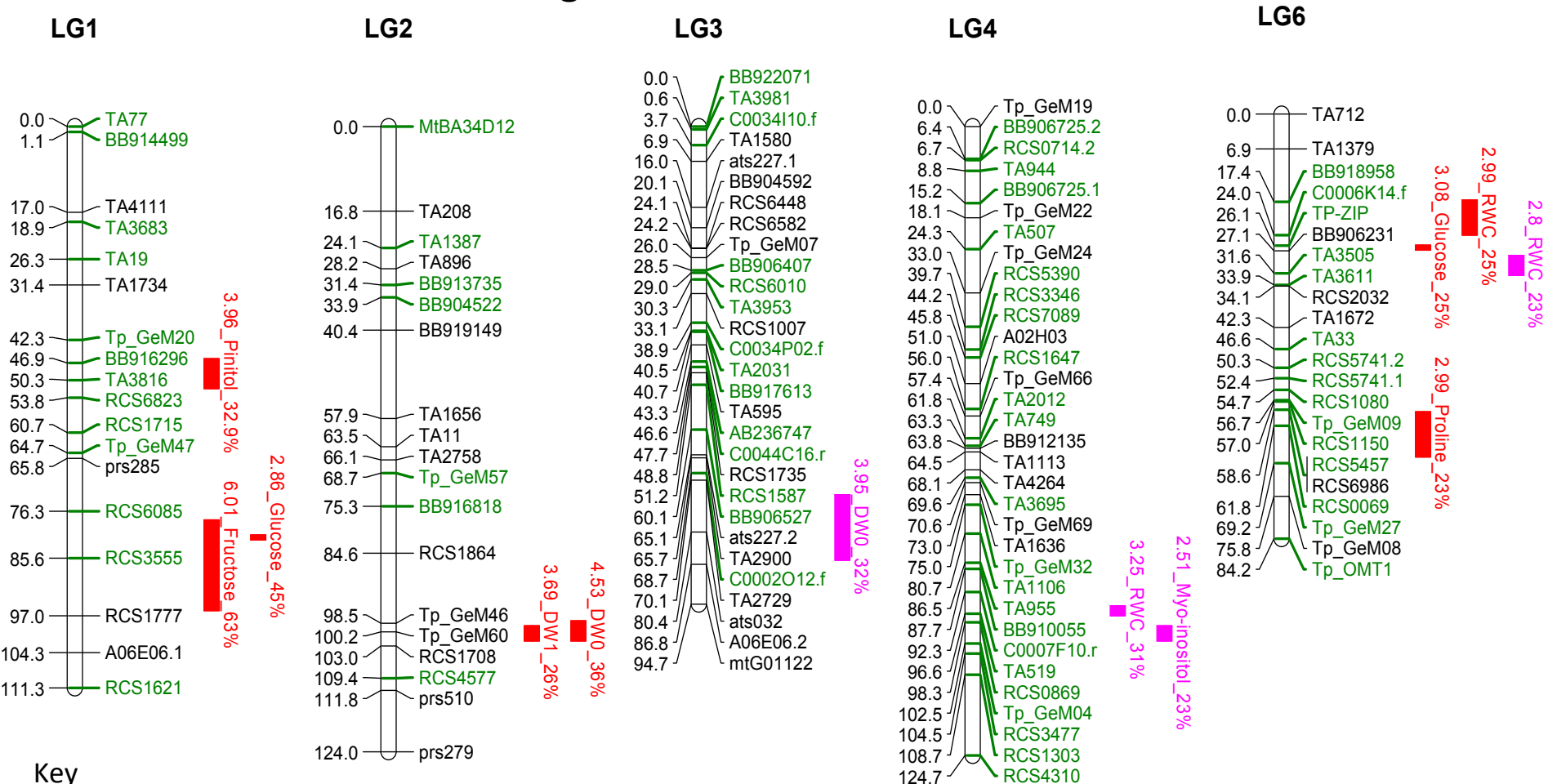
- One generation per four year cycle of half-sibling recurrent selection
- Most traits are assessed **phenotypically** as spaced plants or progeny plot trials
- 12 or more years to market a variety from the current cycle of selection



Molecular markers as a tool in plant breeding

- **Mapping of QTL**
- **Marker assisted backcrossing**
- **Gene pyramiding**
- **Pedigree breeding**
- **Recurrent selection**

Drought-related QTL in red clover



Key

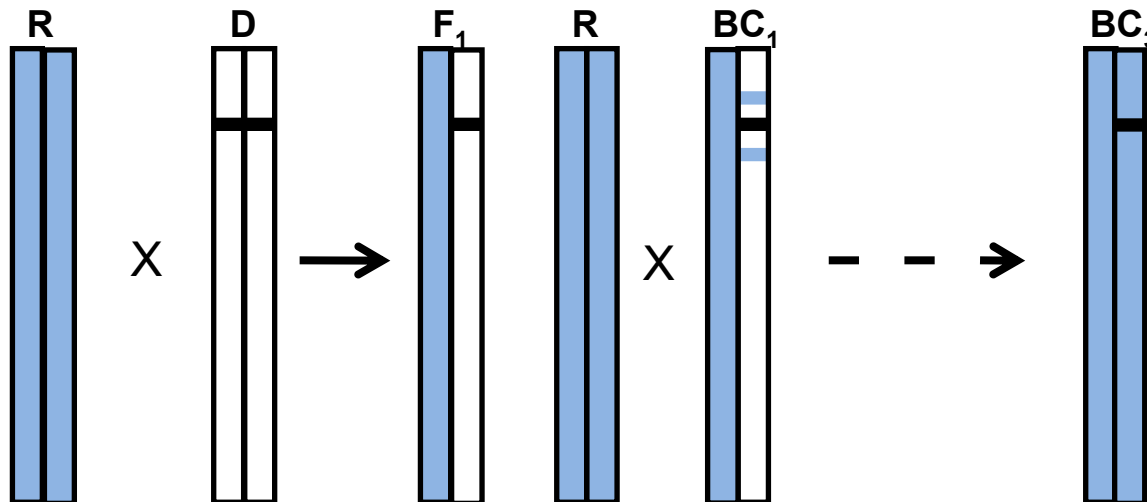
Red = 1st experiment (early summer 2010)

Pink = 2nd experiment (late summer 2010)

LOD score_Trait_% variance accounted for

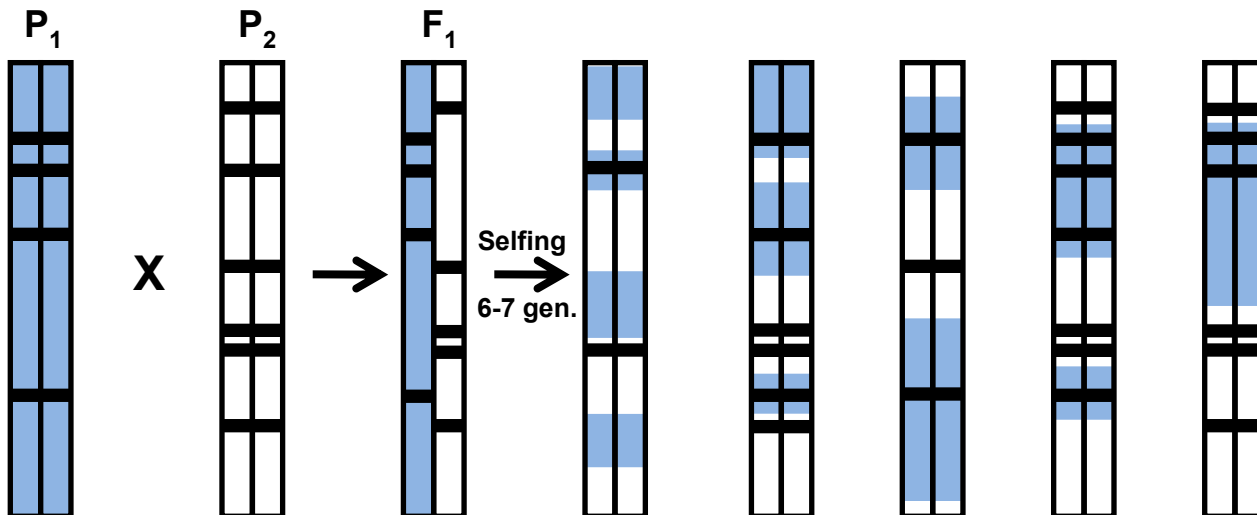
Marker assisted selection

- Marker assisted backcross breeding (foreground and background selection)



Marker assisted selection in plant breeding

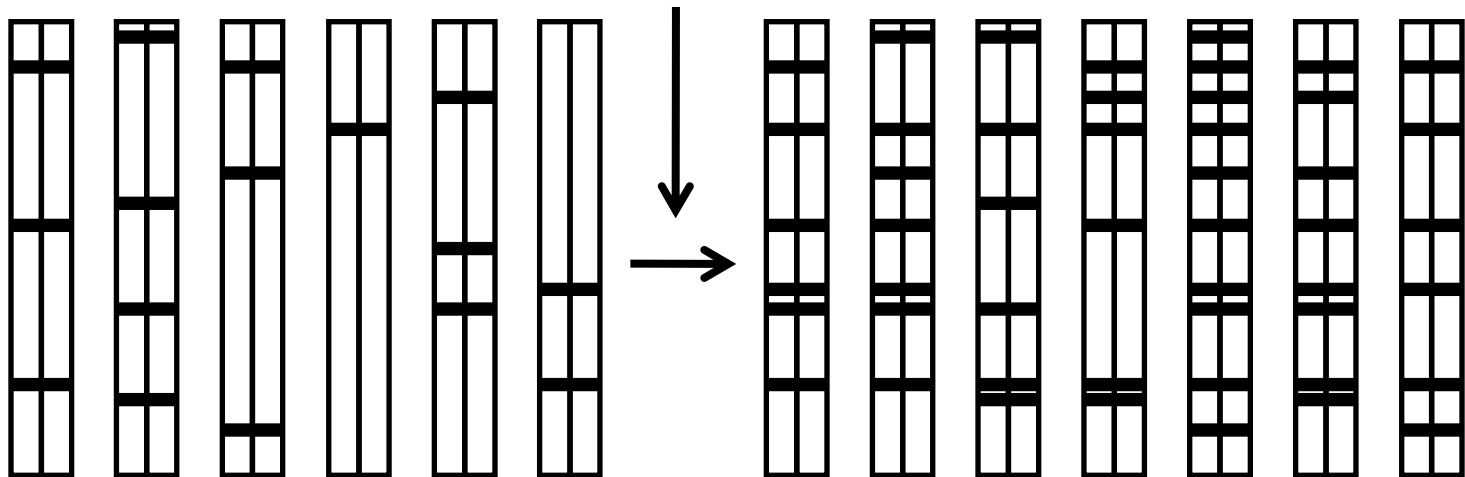
- Pedigree breeding



Marker assisted selection in plant breeding

- Recurrent selection

Selection and intermating



Selection index:

$$M_j = \sum b_i X_{ij}$$

A few selective examples of marker assisted selection

- **Barley**
 - **Disease resistance**
 - **Malt quality**
- **Rice**
 - **Disease and pest resistance**
 - **Yield**
 - **Morphological traits**
- **Wheat**
 - **Pest and disease resistance**
 - **Abiotic stress**
 - **Quality traits**
- **Pearl millet**
 - **Drought**

Limitations of current MAS

- **QTL detection in bi-parental mapping population**
- **Only allelic variation present in the two parents captured**
- **Best suited for traits controlled by few QTL with major effect (high heritability)**
- **Most current target traits are multi-genic, i.e. controlled by many genes with small effects each. Difficult to get statistical power to detect these**
- **LD or association mapping sidesteps the need for bi-parental mapping populations and can be used to associate genotype with trait in breeding or other populations which have been extensively phenotyped.**

Linkage disequilibrium

□ Non-random distribution of alleles at two loci

$$D_{AB} = p_{AB} - p_A p_B \neq 0$$

Association mapping

□ Search for LD due to linkage between molecular marker and phenotype

Linkage disequilibrium

Non-random association of alleles at two loci

Two loci with two alleles each

Locus 1: A_1A_2 ; Locus 2: B_1B_2

$$D = p(A_1B_1) \cdot p(A_2B_2) - p(A_1B_2) \cdot p(A_2B_1)$$

$$r^2 = D^2 / ((pA_1) \cdot (pB_1) \cdot (pA_2) \cdot (pB_2))$$

Factors affecting LD

Increase

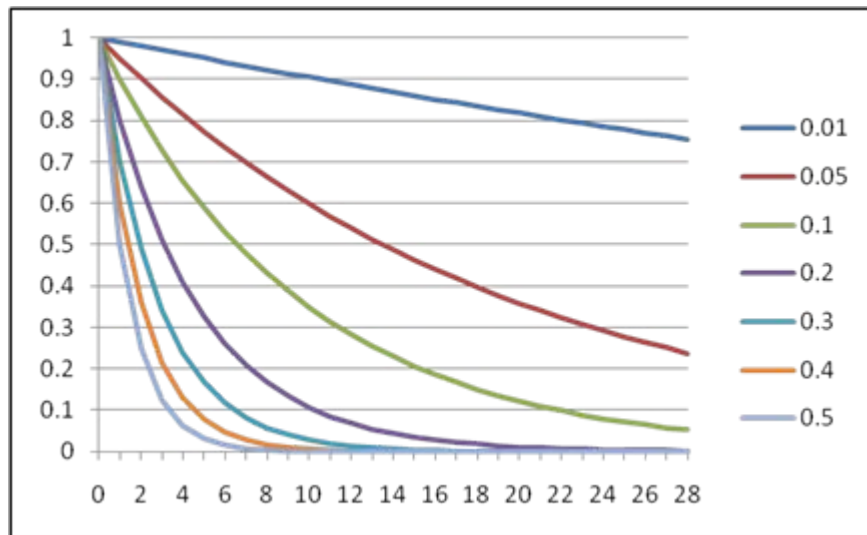
- Small effective population size
- Inbreeding
- Genetic isolation
- Population subdivision
- Low recombination rate
- Genetic drift
- Population admixture

Decrease

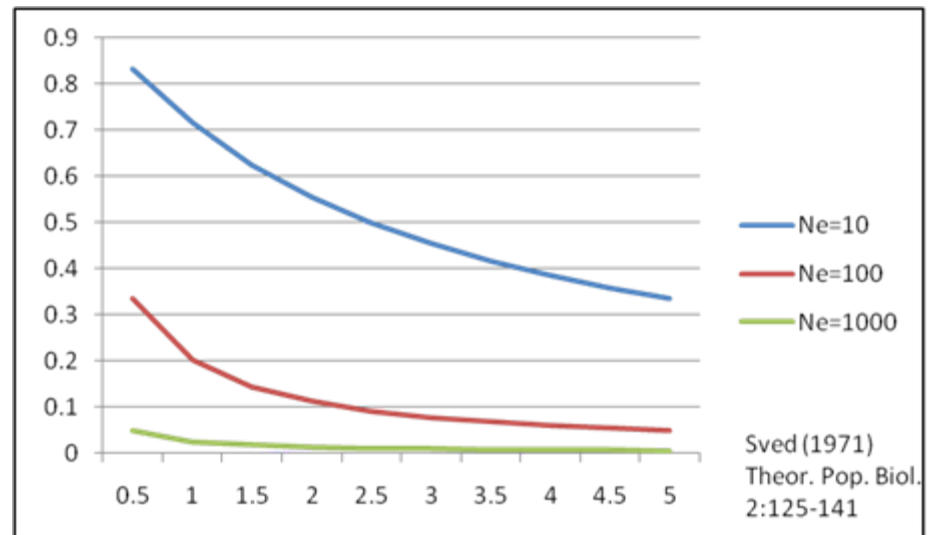
- Outcrossing
- High recombination rate
- High mutation rate

Influence of recombination fraction and Size of N_e on LD

Effect of recombination fraction



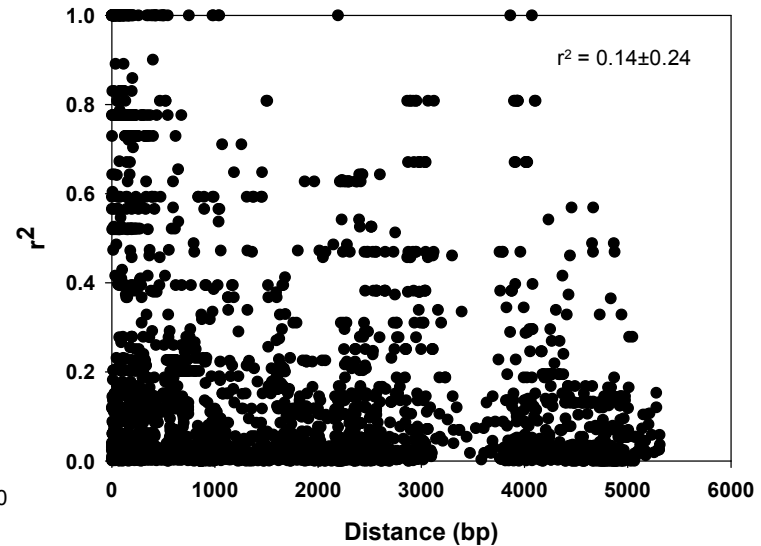
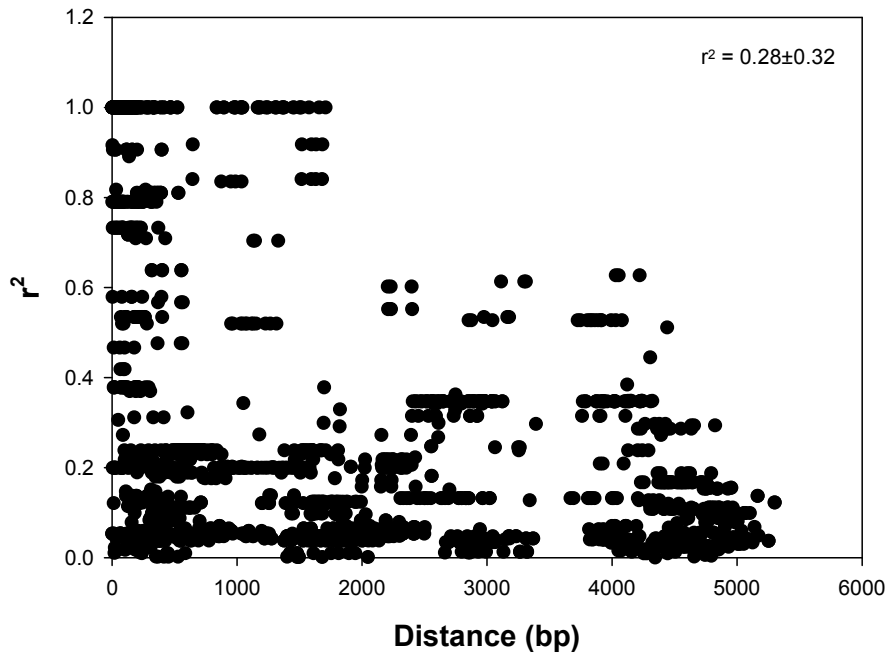
Effect of N_e



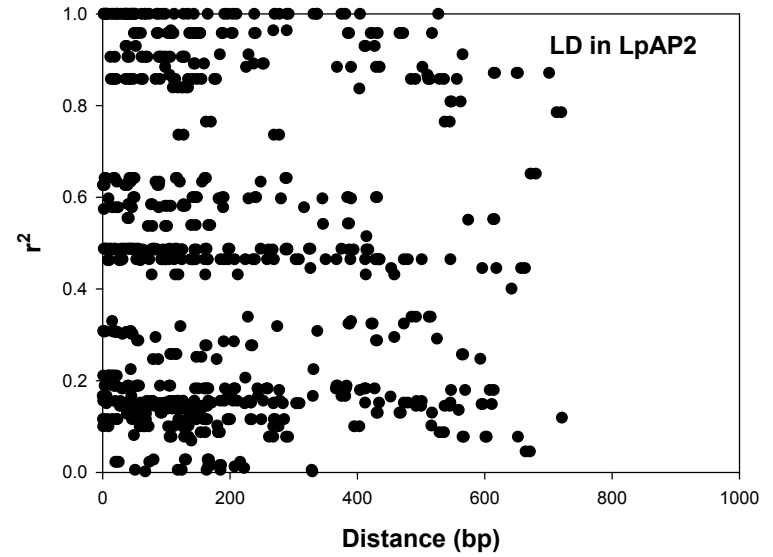
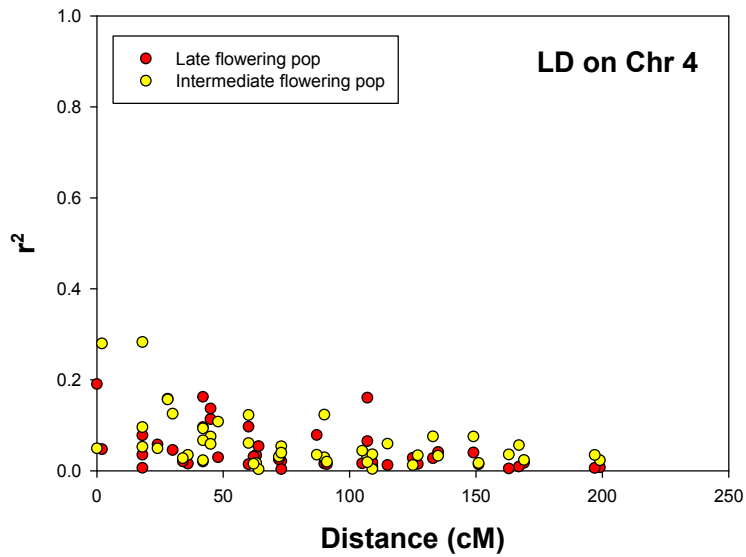
Sved (1971)
 Theor. Pop. Biol.
 2:125-141

Intragenic LD in varieties vs natural populations

- Some evidence of more LD in varieties in LpAI gene



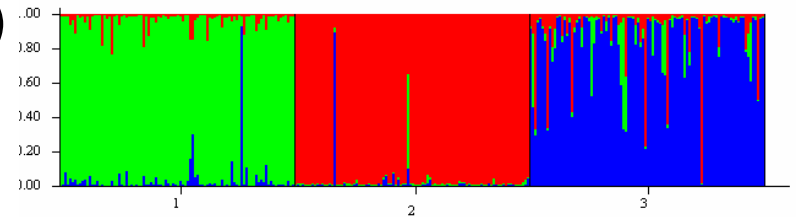
LD in breeding populations



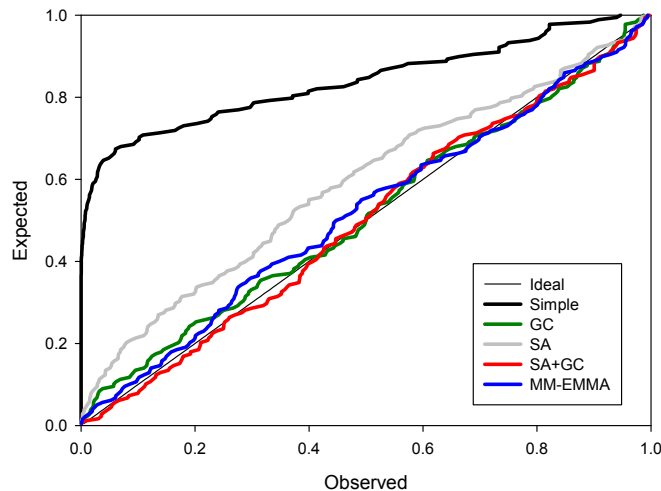
Average $r^2 = 0.47 \pm 0.31$

Correction for population structure and relationship between genotypes

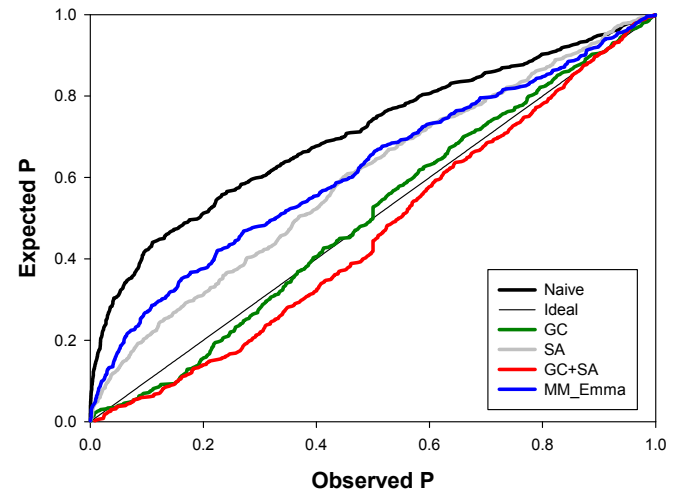
- **Structured association (STRUCTURE)**
- **Genomic control**
- **Mixed model with kinship**



Natural populations



Synthetic populations (varieties)

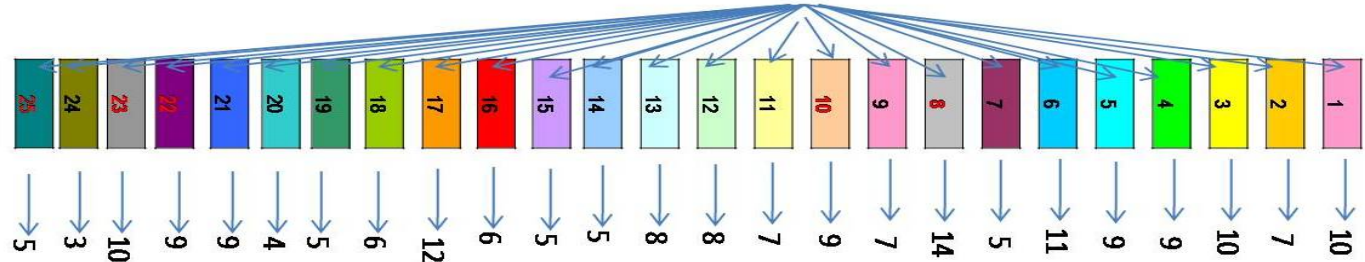


Rattan Yadav, Deepmala Sehgal
Tom Hash

1000 pearl millet accessions, landraces, cultivars assembled from Africa and Asia

↓ Analysed by 19 SSR markers

24 core clusters



+207 more accessions with specific traits such as drought tolerant, salt tolerant, productive tillers etc.

Selfed and testcrosses made on ms line 99022A using bulk pollen from 400 accessions



Selection of 'diversity panel' of 272 most informative genotypes grouped into 4 maturity groups

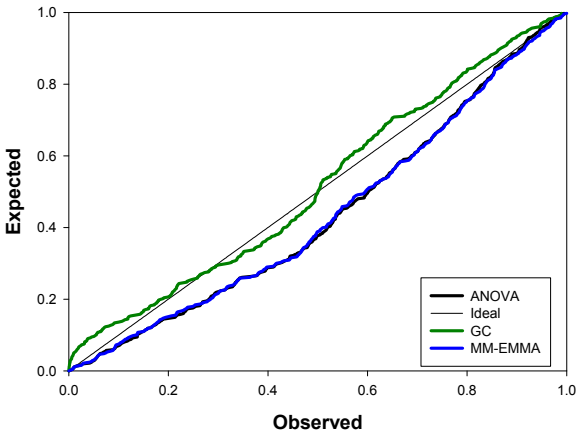
Two rounds of selfing

Testcross seed production

Candidate gene (mapping to LG2) sequencing in the diversity panel

Drought phenotyping

Association models



Observed

— ANOVA
— Ideal
— GC
— MM-EMMA

Genomic Control (GC)

- **About 50 genome wide markers needed**
- **Distribution of test statistic for association**
- **If test statistic is above expected value of 1 structure is present**
- **Null hypothesis: No association above what is expected as a result of population structure**
- **χ^2 (GC) = χ^2 (Obs)/ $\Sigma(\chi^2$ (Null markers)/n)**
- **Simple, but can overcorrect**

Flowering time in ryegrass

- **Fundamental to biology of grasses**
- **Impact on field performance (biomass)**
- **Quality traits**
- **Seed production**

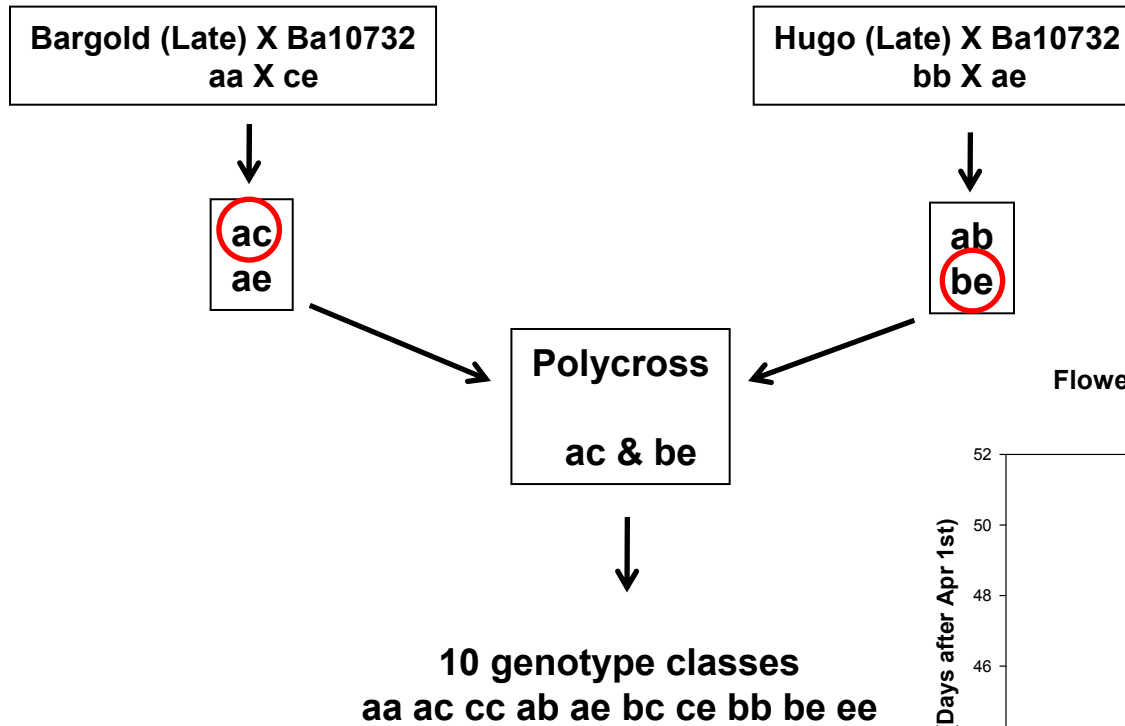
Association analysis of flowering time with *LpFT*

Locus/allele	Year	<i>P</i>				% Marker effect
		Simple	SA	SA+GC	EMMA MM	
<i>FT-LD1a</i>	2004	< 10 ⁻¹⁷	4 x 10 ⁻¹²	3 x 10 ⁻¹⁰	4 x 10 ⁻²⁹	12.4
<i>FT-LD1b</i>	2004	NS	NS	NS	8 x 10 ⁻³	0.7
<i>FT-LD1c</i>	2004	< 10 ⁻¹⁷	6 x 10 ⁻¹²	5 x 10 ⁻¹⁰	3 x 10 ⁻³⁰	13.4
<i>FT-LD1e</i>	2004	5 x 10 ⁻⁸	NS	NS	1 x 10 ⁻³	1.0
<i>FT-LD2a</i>	2004	< 10 ⁻¹⁷	1 x 10 ⁻⁸	2 x 10 ⁻⁷	2 x 10 ⁻²¹	9.0
<i>FT-LD2b</i>	2004	6 x 10 ⁻¹⁵	1 x 10 ⁻²	2 x 10 ⁻²	7 x 10 ⁻⁵	1.6
<i>FT-LD2e</i>	2004	3 x 10 ⁻¹¹	7 x 10 ⁻³	1 x 10 ⁻²	3 x 10 ⁻⁵	1.8
<i>FT-LD3a</i>	2004	< 10 ⁻¹⁷	4 x 10 ⁻⁹	1 x 10 ⁻⁷	5 x 10 ⁻²⁰	8.5
<i>FT-LD3b</i>	2004	3 x 10 ⁻¹⁵	2 x 10 ⁻³	6 x 10 ⁻³	1 x 10 ⁻⁵	1.9
<i>FT-LD3e</i>	2004	3 x 10 ⁻⁸	4 x 10 ⁻²	NS	1 x 10 ⁻³	1.0
<i>FT-LD1a</i>	2005	< 10 ⁻¹⁷	< 10 ⁻¹⁷	2 x 10 ⁻¹⁰	8 x 10 ⁻⁴⁴	20.2
<i>FT-LD1b</i>	2005	3 x 10 ⁻⁵	2 x 10 ⁻³	NS	1 x 10 ⁻⁷	3.0
<i>FT-LD1c</i>	2005	< 10 ⁻¹⁷	< 10 ⁻¹⁷	< 10 ⁻¹⁷	2 x 10 ⁻⁷⁸	34.0
<i>FT-LD1e</i>	2005	NS	NS	NS	NS	0.2
<i>FT-LD2a</i>	2005	< 10 ⁻¹⁷	< 10 ⁻¹⁷	4 x 10 ⁻⁸	1 x 10 ⁻³⁴	16.1
<i>FT-LD2b</i>	2005	< 10 ⁻¹⁷	4 x 10 ⁻¹¹	4 x 10 ⁻⁵	1 x 10 ⁻¹⁰	5.0
<i>FT-LD2e</i>	2005	6 x 10 ⁻³	2 x 10 ⁻³	5 x 10 ⁻²	4 x 10 ⁻²	0.4
<i>FT-LD3a</i>	2005	< 10 ⁻¹⁷	< 10 ⁻¹⁷	2 x 10 ⁻⁸	5 x 10 ³⁴	15.7
<i>FT-LD3b</i>	2005	< 10 ⁻¹⁷	2 x 10 ⁻¹¹	3 x 10 ⁻⁵	6 x 10 ⁻¹¹	4.9
<i>FT-LD3e</i>	2005	NS	1 x 10 ⁻³	4 x 10 ⁻²	5 x 10 ⁻²	0.4

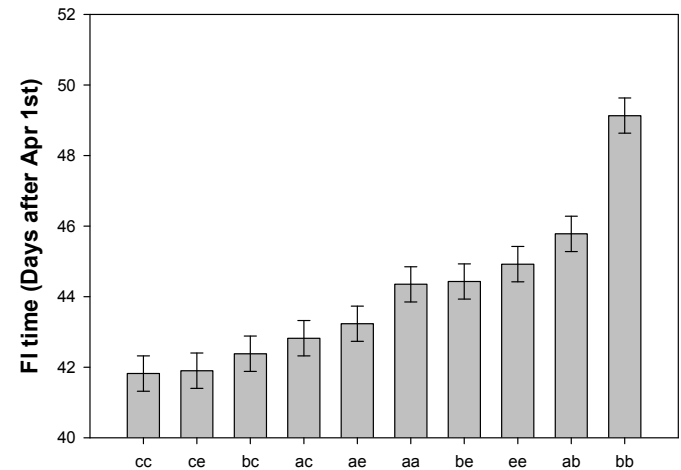
Association analysis of flowering time with *LpFT*

Locus/allele	Year	<i>P</i>				% Marker effect
		Simple	SA	SA+GC	EMMA MM	
<i>FT-LD1a</i>	2004	< 10 ⁻¹⁷	4 x 10 ⁻¹²	3 x 10 ⁻¹⁰	4 x 10 ⁻²⁹	12.4
<i>FT-LD1b</i>	2004	NS	NS	NS	8 x 10 ⁻³	0.7
<i>FT-LD1c</i>	2004	< 10 ⁻¹⁷	6 x 10 ⁻¹²	5 x 10 ⁻¹⁰	3 x 10 ⁻³⁰	13.4
<i>FT-LD1e</i>	2004	5 x 10 ⁻⁸	NS	NS	1 x 10 ⁻³	1.0
<i>FT-LD2a</i>	2004	< 10 ⁻¹⁷	1 x 10 ⁻⁸	2 x 10 ⁻⁷	2 x 10 ⁻²¹	9.0
<i>FT-LD2b</i>	2004	6 x 10 ⁻¹⁵	1 x 10 ⁻²	2 x 10 ⁻²	7 x 10 ⁻⁵	1.6
<i>FT-LD2e</i>	2004	3 x 10 ⁻¹¹	7 x 10 ⁻³	1 x 10 ⁻²	3 x 10 ⁻⁵	1.8
<i>FT-LD3a</i>	2004	< 10 ⁻¹⁷	4 x 10 ⁻⁹	1 x 10 ⁻⁷	5 x 10 ⁻²⁰	8.5
<i>FT-LD3b</i>	2004	3 x 10 ⁻¹⁵	2 x 10 ⁻³	6 x 10 ⁻³	1 x 10 ⁻⁵	1.9
<i>FT-LD3e</i>	2004	3 x 10 ⁻⁸	4 x 10 ⁻²	NS	1 x 10 ⁻³	1.0
<i>FT-LD1a</i>	2005	< 10 ⁻¹⁷	< 10 ⁻¹⁷	2 x 10 ⁻¹⁰	8 x 10 ⁻⁴⁴	20.2
<i>FT-LD1b</i>	2005	3 x 10 ⁻⁵	2 x 10 ⁻³	NS	1 x 10 ⁻⁷	3.0
<i>FT-LD1c</i>	2005	< 10 ⁻¹⁷	< 10 ⁻¹⁷	< 10 ⁻¹⁷	2 x 10 ⁻⁷⁸	34.0
<i>FT-LD1e</i>	2005	NS	NS	NS	NS	0.2
<i>FT-LD2a</i>	2005	< 10 ⁻¹⁷	< 10 ⁻¹⁷	4 x 10 ⁻⁸	1 x 10 ⁻³⁴	16.1
<i>FT-LD2b</i>	2005	< 10 ⁻¹⁷	4 x 10 ⁻¹¹	4 x 10 ⁻⁵	1 x 10 ⁻¹⁰	5.0
<i>FT-LD2e</i>	2005	6 x 10 ⁻³	2 x 10 ⁻³	5 x 10 ⁻²	4 x 10 ⁻²	0.4
<i>FT-LD3a</i>	2005	< 10 ⁻¹⁷	< 10 ⁻¹⁷	2 x 10 ⁻⁸	5 x 10 ³⁴	15.7
<i>FT-LD3b</i>	2005	< 10 ⁻¹⁷	2 x 10 ⁻¹¹	3 x 10 ⁻⁵	6 x 10 ⁻¹¹	4.9
<i>FT-LD3e</i>	2005	NS	1 x 10 ⁻³	4 x 10 ⁻²	5 x 10 ⁻²	0.4

Validation population

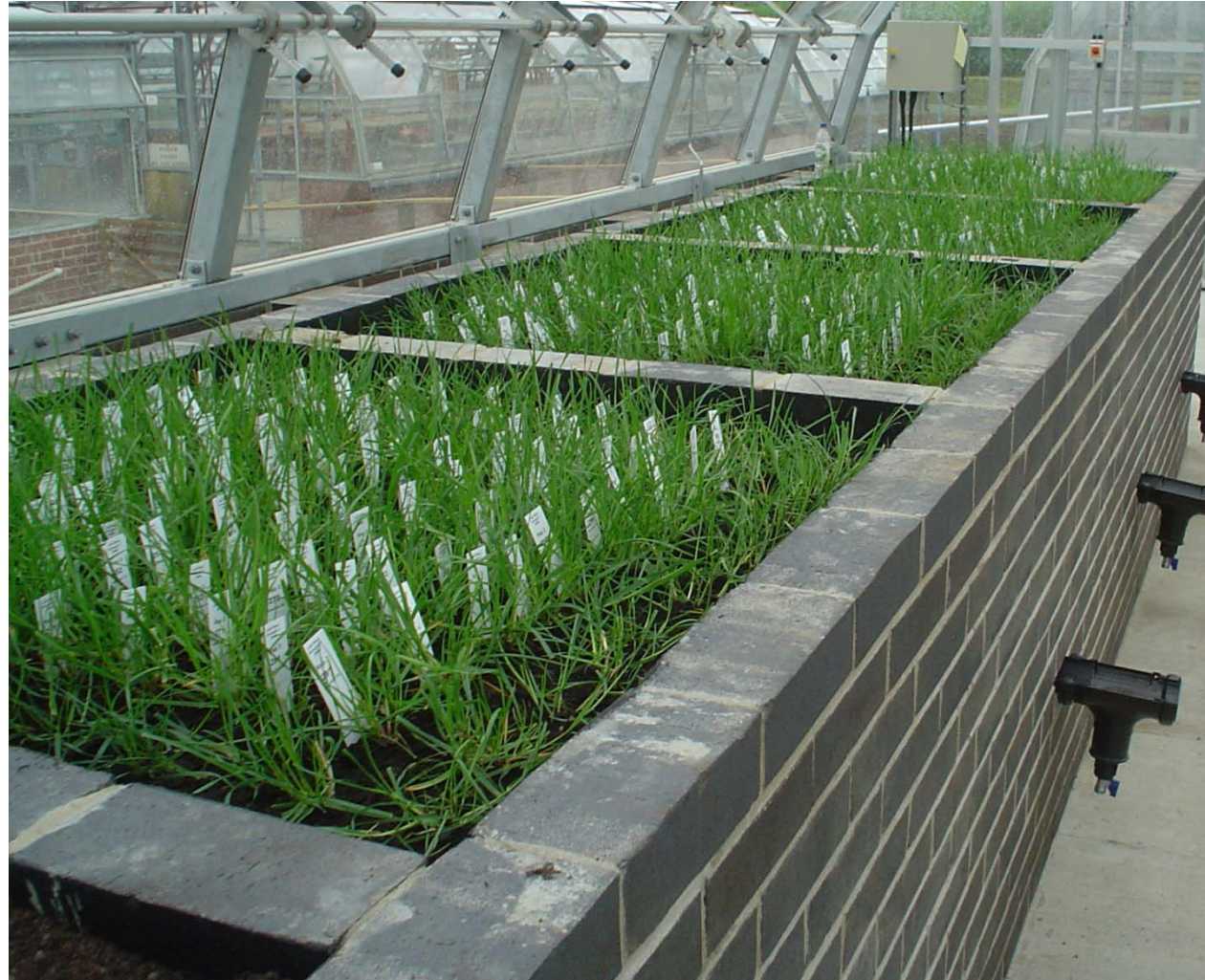


Flowering time in validation population



Drought experiment

- 3 current varieties
- 96 genotypes each
- 3 reps
- 2 years
- Biomass
- Survival
- Regrowth



Association analysis of drought experiment Mixed model – EMMA

- Most significant markers explain little of phenotypic variation

Marker effect (%)

2009	DW1	DW2	DW3	DW4	Leaf Ext	Tiller score	Tiller surv
G07_058	0.0	2.3	2.1	3.2	0.8	1.4	1.0
25ca1	0.2	0.7	1.8	1.5	2.8	0.0	0.9
2008							
G07_058	0.9	0.7	1.5	1.8	2.1	2.6	1.2
25ca1	1.7	2.2	4.5	3.0	2.5	1.1	1.4

Genomic selection

- **Association mapping still prone to biased effect estimates and inability explain all the variance of a trait**
- **Simultaneous estimation of all locus effects across the genome to calculate the genomic estimation of breeding value (Meuwissen et al. (2001) Genetics, 157: 1819-1829)**
- **Joint analysis of all markers genome-wide to try to explain the total variance. Ideally all QTL are in LD with at least one marker**
- **Training population of both genotyped and phenotyped individuals**
- **Use training set to develop a statistical prediction model for GEBV**
- **Calculate GEBV in test population or selection candidates**

Genomic selection and breeding value

- **Breeding value (BV) of an individual relates to the mean value of its progeny**
- **If an individual is crossed with a random selection of a population its BV relates to the deviation from the population mean**
- **BV is the sum of the average effect of the genes it carries**
- **Prediction models aim to calculate the effect of each gene**

Genomic selection

- Variance explained by a marker to a QTL = $r^2 \cdot h^2$
- If LD low to a single marker, haplotypes of several surrounding markers may explain more variance
- Consequence is even more explanatory variables per observation
- Scale the number of markers to size of N_e

Genomic selection prediction models

Stepwise regression

1. Select most significant segments or markers by forward or backward elimination
2. Estimate the effect of significant markers using multiple regression

Problems:

1. Since only significant markers are estimated not all variation explained
2. Risk of overestimation of significant effects

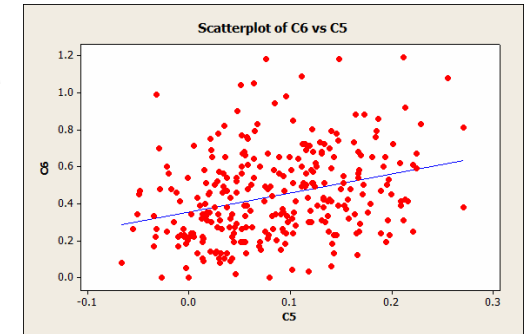
Genomic selection prediction models

Ridge regression BLUP

1. All marker effects assumed to have equal variance

2. $\hat{\mathbf{g}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1} \mathbf{X}'\mathbf{y}$

\mathbf{X} : Design matrix for genotypes
 λ : σ_e^2 / σ_g^2
 \mathbf{y} : Vector of phenotypes



Model can also (usually does) include a term for random effect (\mathbf{Zu}) such as kinship as determined by marker data.

$$\mathbf{y} = \mathbf{Xg} + \mathbf{Zu} + \mathbf{e}$$

Problem:

Some markers (QTL) explain a larger effect than others

Genomic selection prediction models

Bayes regression

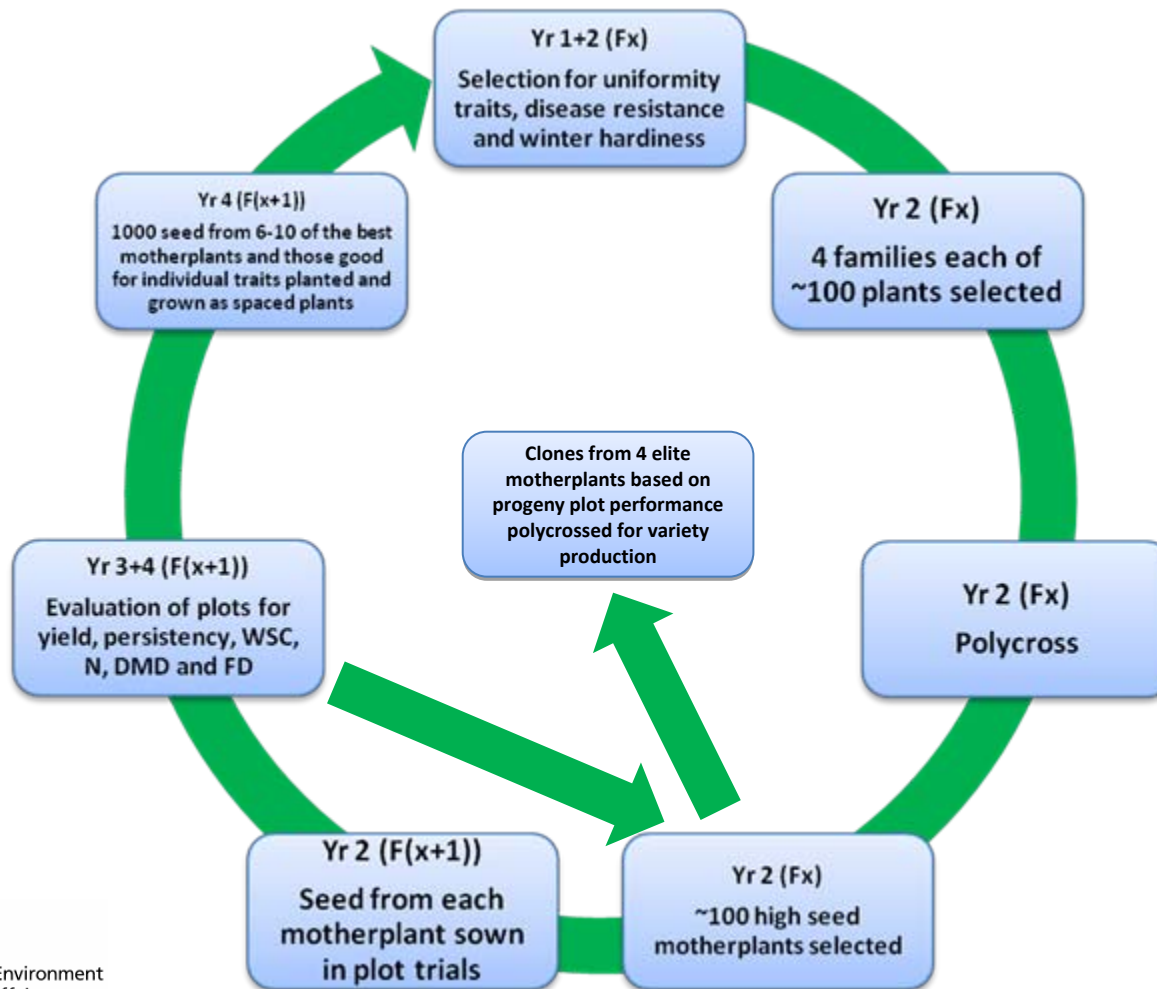
1. Marker variance treated more realistically
2. All marker effects are > 0 (Bayes A)
3. Some marker effects can be $= 0$ (Bayes B)

Meuwissen et al. (2001)

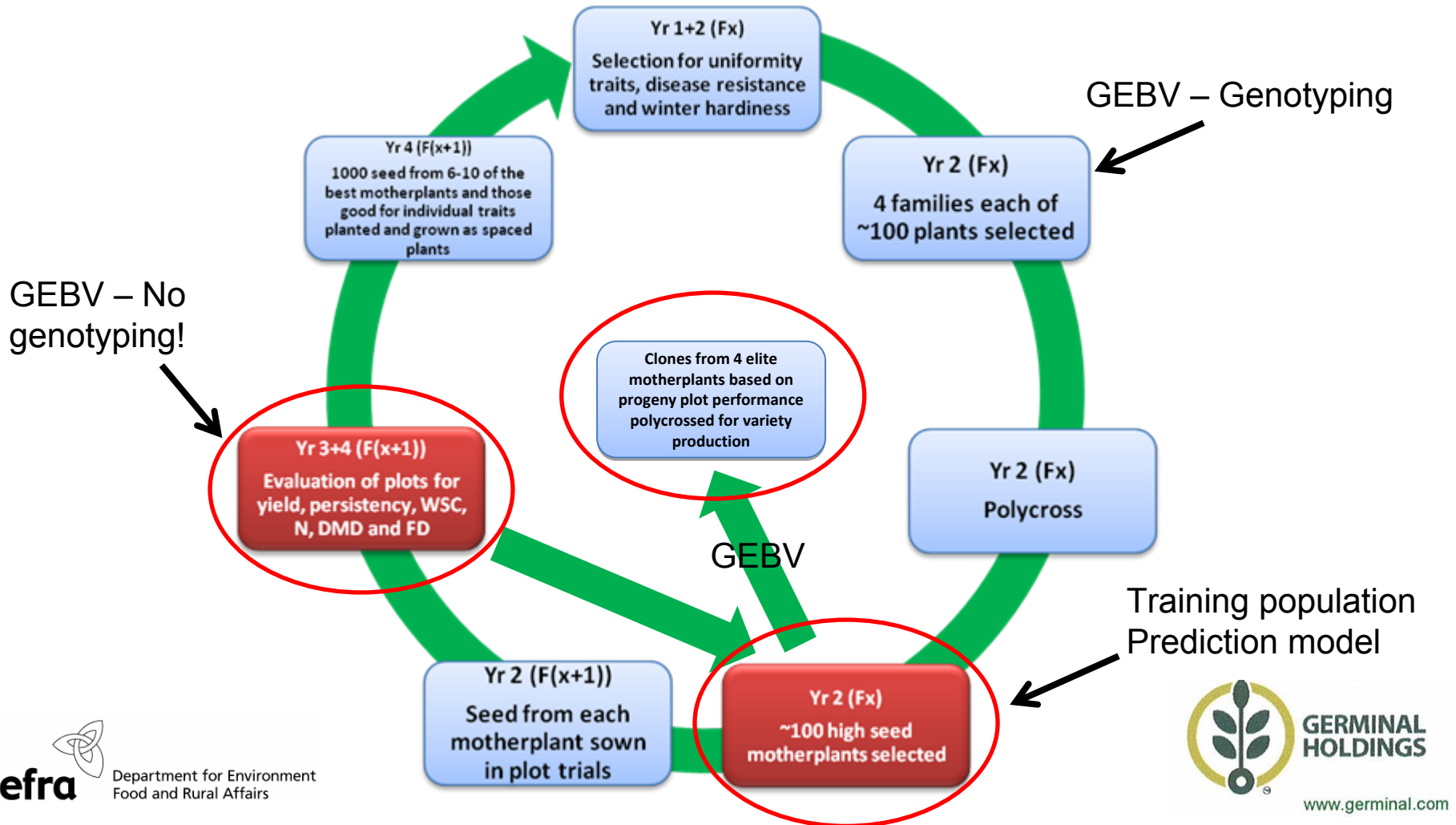
Other potential genomic selection prediction models

- **Motherplants of breeding populations as training set**
- **Make predictions based on machine learning methodology such as “Knowledge Discovery in Databases” and “Active learning”**
- **Simulations to discover the effect of genetic architecture of trait**
- **Test predictive models on current breeding cycle and historical data (varieties)**

IBERS perennial ryegrass breeding programme



GS in IBERS ryegrass breeding programme?



What traits?

High sugar grasses

High sugar grass = grass with enhanced levels of water soluble carbohydrate.

Water soluble carbohydrates (WSC) - natural storage compounds – mainly fructans.

High sugar grasses - significantly higher WSC levels through the season

Still room for improvement if ryegrass is used for bioenergy

High throughput sequencing for transcriptomics analysis of WSC in ryegrass

- Extreme genotypes from F2 mapping family segregating for WSC
- (High light + water) or (sucrose + dark)
- Remobilisation
- 454 sequencing of *Lolium perenne* transcriptome
- Illumina sequencing of treated samples
- Differential expression and candidate gene discovery
- SNP discovery
- Genetic architecture of trait



Genomic selection

Some issues

- **Size of training population (differing estimates)**
- **Number of markers needed (scale to N_e)**
- **Risk of fast fixation of alleles (stronger selection = fewer individuals)**
- **Kinship – no individual can be predicted to have a higher breeding value than any which has already been phenotyped**
- **The closer the kinship, the more similar their breeding values**
- **Prediction accounts for less and less of genetic variance with each passing generation**

(Ian Mackay, NIAB)

Genomic selection

- **Has proved its value in animal breeding particularly dairy cattle**
(see e.g. Hayes & Goddard (2010))
- **GS and association mapping are complementary**
- **Still to prove its value over generations in crop plants**
- **Accuracy of predictions depends upon:**
 - **Trait heritability, Effective population size, Genome size, marker density, genetic architecture of trait, prediction model**
- **Simulation studies in plants suggest potential for improved gain per unit time**

(Jannink et al. (2010) Briefings Func. Genom. 9: 166-177

Heffner et al. (2009) Crop Sci. 49: 1-11

Bernardo & Yu (2007) Crop Sci. 47: 1082-1090)

Genomic selection

- In current breeding programmes phenotyping used for selection
- In GS main role of phenotyping is to calculate effect of markers

Future plans

Aims:

- **To utilise the ryegrass breeding populations for genome-wide association genetics and genomic selection. Provides a test bed for hypothesis testing regarding past history of selection and dissection of quantitative traits in addition to more efficient plant breeding**
- **Use NAM-type populations for association and QTL genetics with a view to discovering novel variation in traits of agronomic and biological importance**
- **Next generation sequencing technology for transcriptomics analysis of WSC and other quantitative traits**

Resource building to achieve the above aims

1. Physical map and genome sequence

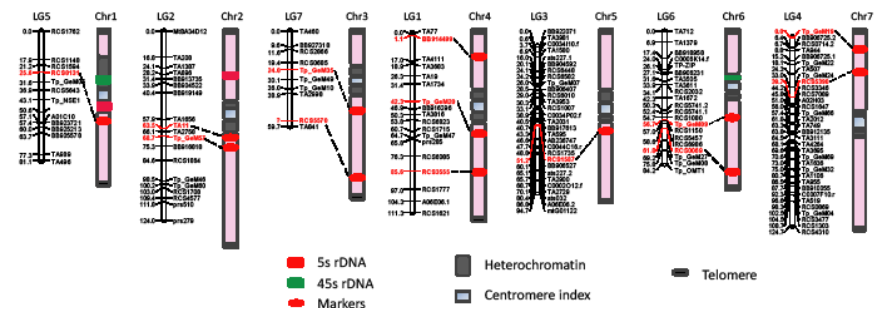
2. Genome-wide coverage with SNPs

3. Development of NAM-type population with motherplant from breeding population, and pollen donors from diverse germplasm

Red clover translational genomics

Red clover as a forage crop

1. High yield forage crop (10-15 T DM/Ha)
2. High protein fodder for silage for winter feed
3. Less need for N fertiliser due to N₂ fixation
4. Increasingly popular in low input agriculture
5. Need for better persistency
6. Need for biotic and abiotic stress tolerance

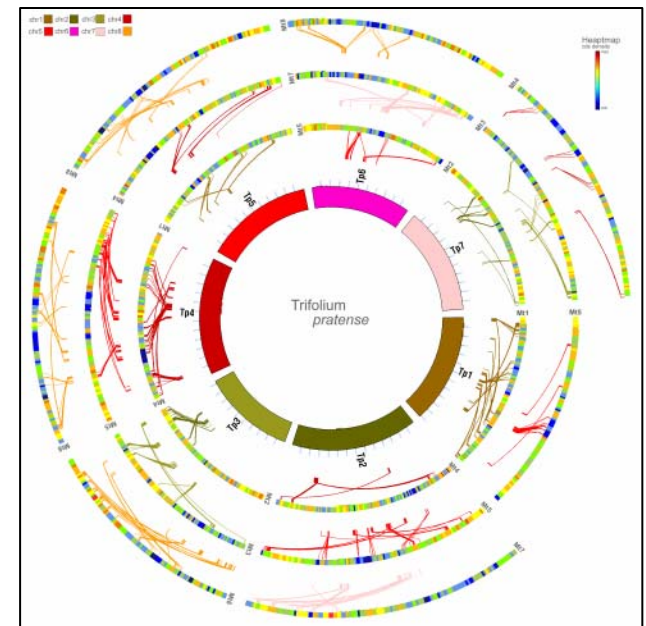


- Physical map of red clover linked to genetic map and aligned to *Medicago truncatula* genome

Lang *et al.* (2011) Morphology of *Trifolium pratense* L. pachytene chromosomes and integration of the cytogenetic and the genetic linkage map (to be submitted)

Skøt *et al.* (2011) Physical map of red clover (in preparation)

- Next generation sequencing of red clover genome in progress (with The Genome Analysis Centre)
- Prospects for genome scans to associate genotype with phenotype



Acknowledgements

Flowering time

Ruth Sanderson
Ian Armstead
Danny Thorogood
Kirsten Skøt
Ann Thomas
Galina Latypova
Torben Asp

GS

Richard Hayes
Ruth Sanderson
Matt Hegarty
Ross King
Ian Armstead
Ian Mackay
Wayne Powell

Drought & WSC

Joe Gallagher
Emma Timms-Taravella
Rhys Kelly
Ajoy Roy (IGFRI)
Lesley Turner
Justin Pachebat
Jenny Ashton
Sally O'Donovan
Mike Humphreys
Richard Hayes

ERANET

Michael Abberton
Charlotte Jones
Iain Donnison
Giles Oldroyd
Chunting Lang (Wageningen)
Rene Geurts (Wageningen)
Stephan Roessner (MIPS)
Klaus Mayer (MIPS)
Dave Kudrna et al (Arizona)