

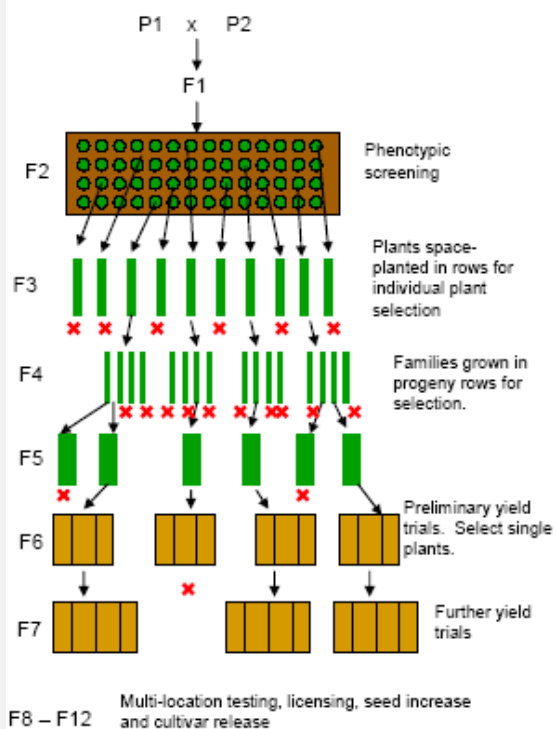
# **ASSOCIATION MAPPING (GWAS) IN PLANTS**

**Sujan Mamidi**

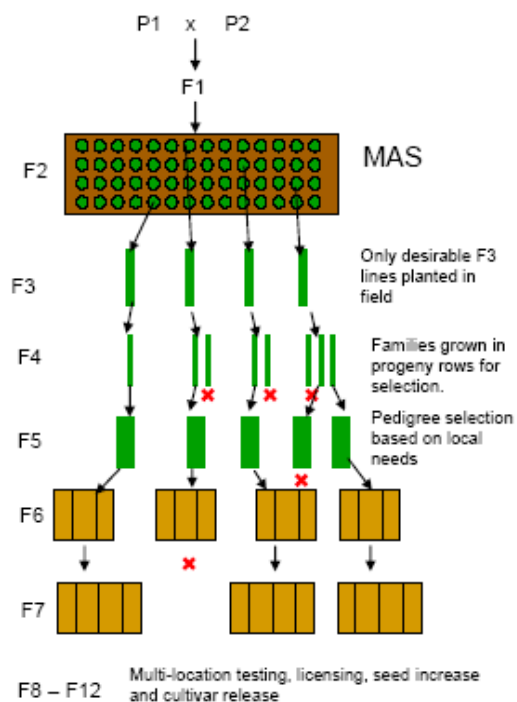
## MARKER ASSISTED SELECTION (MAS)

- Simpler/low cost compared to phenotypic screening
- Selection may be carried out at seedling stage
- Single plants may be selected with high reliability
- Accelerated line development in breeding programs
  - At least two but possibly three or even four backcross generations can be saved by using markers
- High throughput
- Makes pyramiding of genes easier
- Achieved through identification of markers/QTL/genes
  - QTL Mapping
  - Association Mapping

## PEDIGREE METHOD



## EARLY GENERATION SELECTION MARKER ASSISTED SELECTION



## ASSOCIATION MAPPING

- **Association mapping**, also known as "linkage disequilibrium mapping", is a method of mapping quantitative trait loci (QTLs) that takes advantage of linkage disequilibrium (LD) to find associations between phenotypes to genotypes

# LINKAGE DISEQUILIBRIUM

- LD refers to the non random association between two markers or two genes/QTLs

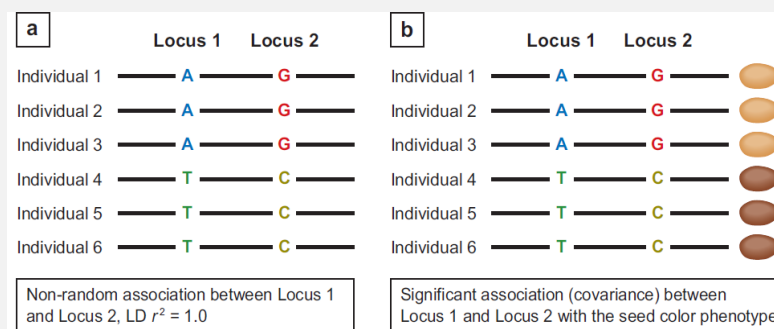


Fig. 1. Principles of linkage disequilibrium and association mapping. a. Linkage disequilibrium. Locus 1 and Locus 2 present an unusual pattern of association between alleles A-G and T-C, which deviate from Hardy-Weinberg expectations, but without any statistical correlation with a phenotype. b. Association mapping. Locus 1 and Locus 2 are in LD. Significant covariance with the seed colour phenotype is considered evidence of association.

- LD can occur between more distant sites or sites located in different chromosomes

## LD MEASUREMENT

- $D'$  - Only reflects the recombination history
  - If two loci are in linkage equilibrium, then  $D = 0$
  - If the two loci are in linkage disequilibrium, then  $D \neq 0$
- $r^2$  - Summarizes both recombinational and mutational history
  - Value between 0 and 1
  - $r^2 = 0$ , Loci are in complete linkage equilibrium
  - $r^2 = 1$ , Loci are in complete linkage disequilibrium
  - **Correlation squared**

## **PARTIAL LD**

- LD calculation in presence of covariate
- Partial correlation because unlinked loci can be in LD simply because of population structure and/or kinship

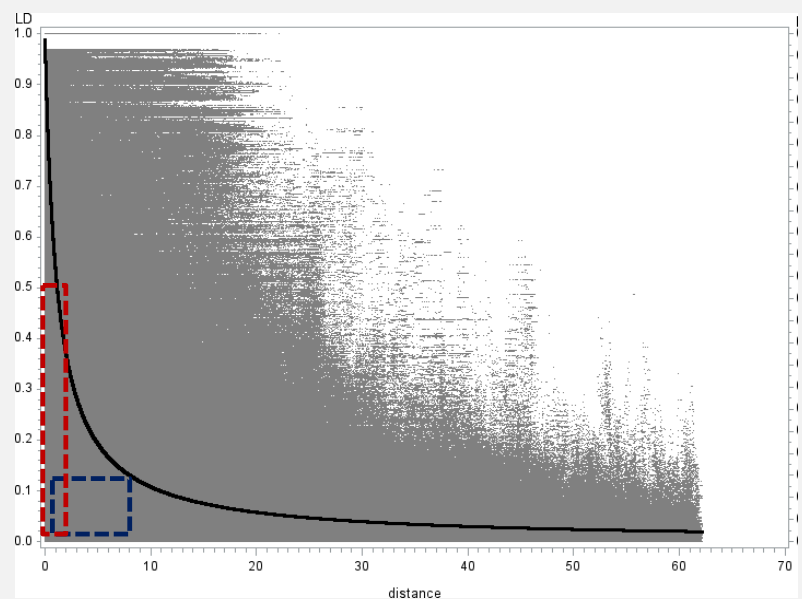
## LD DECAY & LD HEAT PLOTS

- **LD Decay plots:** The distance over which LD persists will **determine the number and density of markers**, and experimental design needed to perform an association analysis.
- Long distance LD
  - Mapping at the centimorgan cM/Mbp distances
- Short distance LD
  - Mapping at the base pair (gene) distance



## LD DECAY EXAMPLE

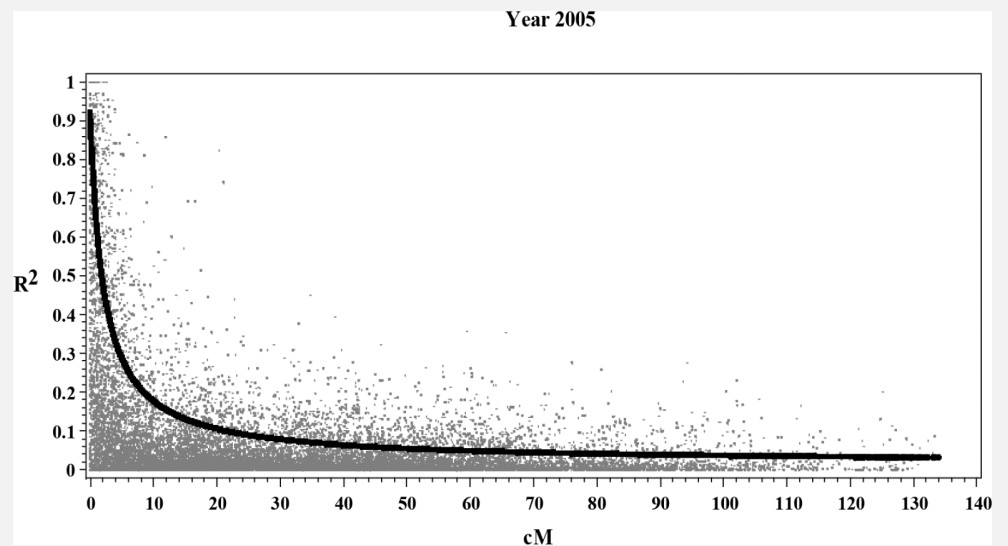
- Two populations (~130) of Advanced Breeding lines in Soybean
- ~ 35000 markers
- Non Linear Regression



Mamidi et al. (2012)

## LD DECAY – GENETIC DISTANCE

- ~ 1500 markers
- ~130 advanced soybean breeding lines

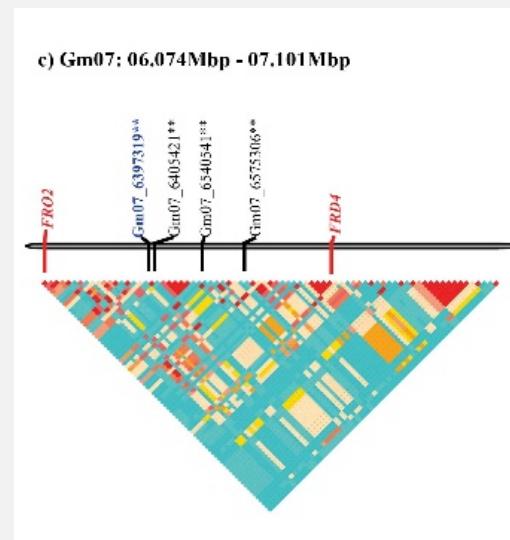


Mamidi et al. (2010)

- LD will tend to decay with distance
- LD decays by one-half with each generation of random mating
- LD declines as the number of generations increases, so in old populations LD is limited to small distances
  - Landraces/Wildtypes – Limited LD
  - Cultivars – long LD

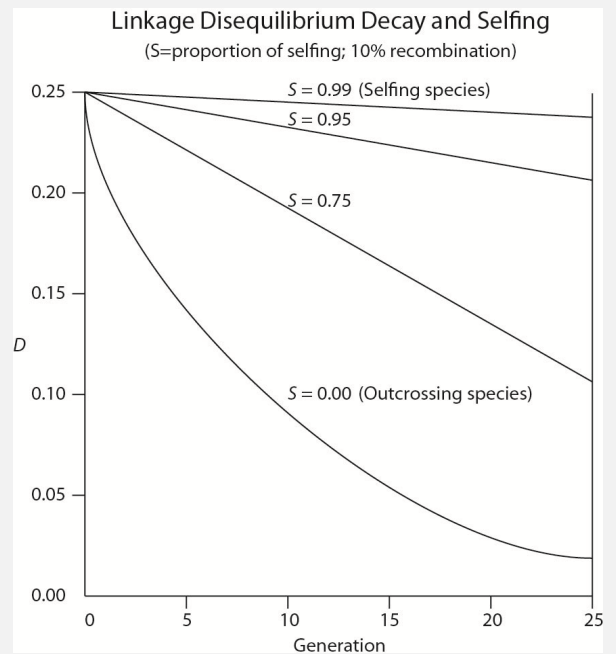
## LD HEAT MAPS

- Disequilibrium matrices or **LD heat maps**:
- Useful for visualizing the linear arrangement of LD between polymorphic sites within a gene or chromosome.



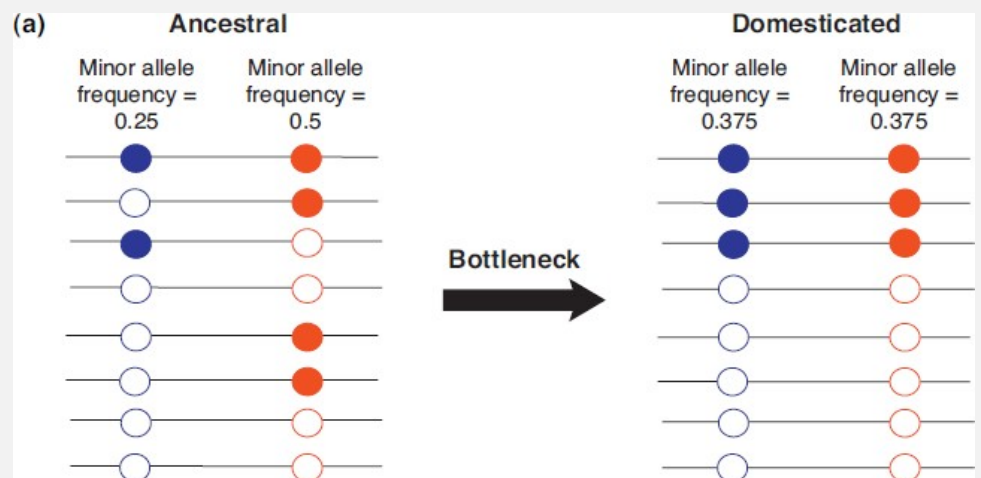
## FACTORS INCREASING LD

- **Mating system**
  - Selfing reduces opportunities for effective recombination



## Domestication:

- After a bottleneck, some haplotypes will be lost
- Resulting in increased LD.

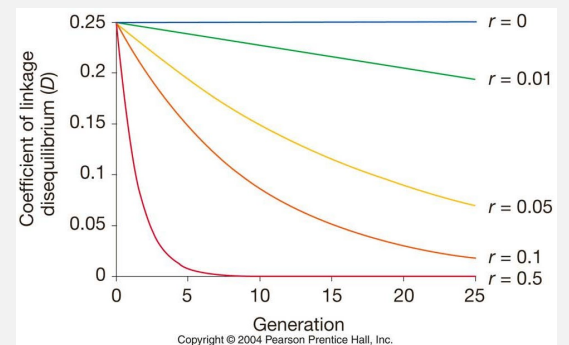


- **Genetic drift**
  - The effect of genetic drift in a small population results in the consistent **loss of rare allelic combinations**
- **Gene flow**
  - Gene flow **introduces new individuals** or gametes with different ancestries and allele frequencies among populations.
- **Selection**
  - **Positive selection will increase LD** between and in the vicinity of the selected loci, a phenomenon known as genetic hitchhiking
  - If both loci are maintained by balancing selection, then LD can persist indefinitely
- **Population structure:**
  - Intentional or unintentional mixing of individuals with different allele frequencies creates LD.

## FACTORS DECREASING LD

- **Recombination**

- LD is broken down by recombination
- GC-rich sequences may be associated with higher rates of recombination and/or mutation





- **Population**

- In general, the **larger the genetic variation**, the faster the LD decay, a direct consequence of the broader historical recombination
- Ex: In corn,
  - LD decays within 1 kb in landraces
  - ~ 2kb in diverse inbred lines and
  - Several hundred kb in commercial elite inbred lines

- **Mutation**

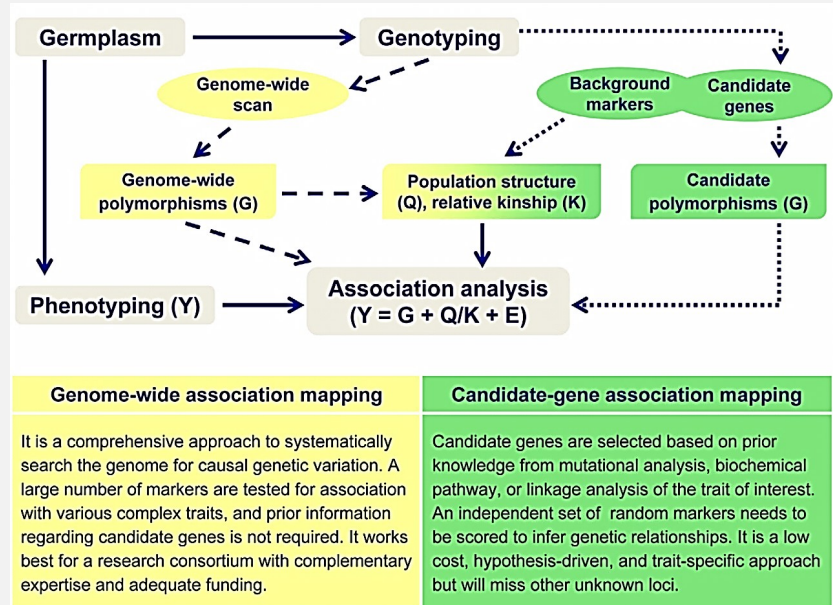
- **Gene conversion**

## INCREASED OR DECREASED LEVEL OF LD

- Populations with either rapid or slow LD decay can be useful in AM, depending on the purposes of the study.
- Populations with narrow genetic diversity and **long extent of LD** are amenable to **coarse mapping** with fewer markers – Few Markers needed
- In more genetically diverse populations, and **short LD decay** for **fine mapping** – More Markers needed

# ASSOCIATION MAPPING

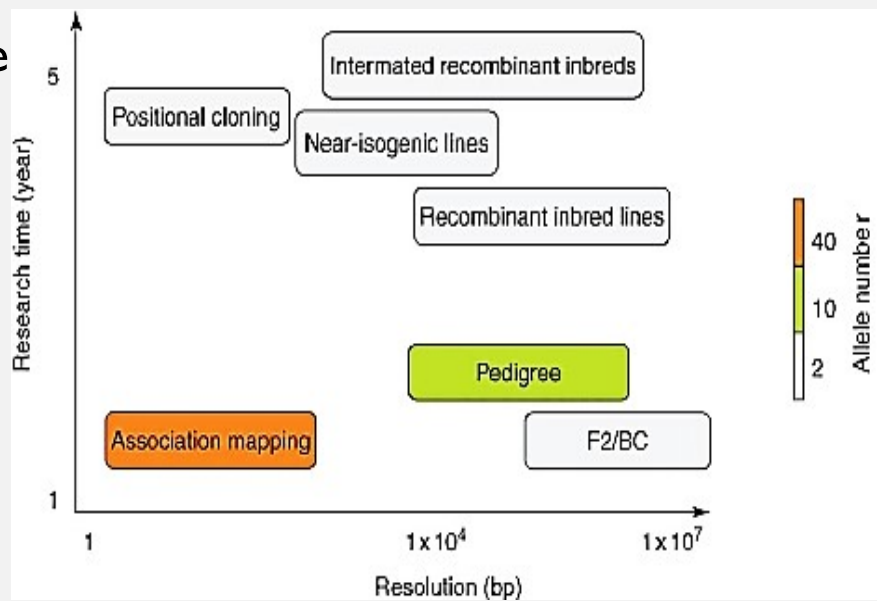
- Genome wide Association Mapping (GWAS)
- Candidate gene Association Mapping



Zhu et al. 2008

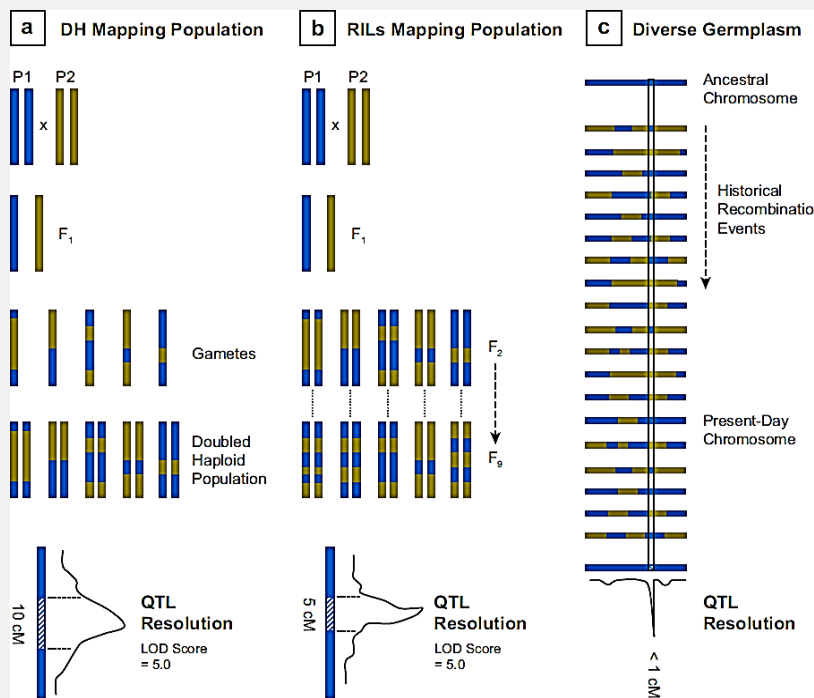
## ADVANTAGES OF AM

- Time saved in developing the population
- Higher resolution













(Yu and Buckler, 2006)

# RESOLUTION



## ASSOCIATION EXAMPLE

Genotype Data			Phenotype Data	
Genotyped Low LD SNP	NOT Genotyped Functional SNP	Genotyped High LD SNP	Berry Number	
G	T	C		15
A	T	C		14
G	T	C		13
A	T	T		12
A	T	C		11
G	A	T		10
G	A	C		9
A	A	T		8
G	A	T		7
A	A	T		6

ASSOCIATION RESULTS						
Low LD SNP		Functional SNP		High LD SNP		
G	A	T	A	C	T	Alleles
10.8	10.2	13.0	8.0	12.4	8.6	Mean Berry Number
0.77		0.0011		0.037		P value of association test
0.04		1		0.36		R <sup>2</sup> - LD with functional SNP

Myles et al. 2009

PROBLEMS OF AM:

I) Population structure

	Sub population 1										Sub population 2									
Height	10	10	12	11	13	9	11	10	13	12	4	6	5	7	6	6	4	5	9	5
Disease Resistance	S	S	S	T	S	S	S	S	T	S	T	S	T	T	T	T	T	S	T	T
SNP1	T	T	T	G	T	T	T	T	G	T	G	G	G	G	G	G	T	G	T	G

2) Kinship  
Family Relatedness

## DIFFERENCES FROM QTL MAPPING

QTL mapping	Association Mapping
Uses bi-parental populations Ex: F <sub>2</sub> , BC, RIL	Uses natural variation of the crop Ex: Landraces, Cultivars
Limited # of recombination events	Many recombination events from multiple lineages
Low resolution (QTL covers many cM)	High resolution (up to casual SNP)
Useful for discovering rare alleles	Common alleles



## COMPARISON

- Linkage - **Correlated inheritance of loci** through the physical connection on a chromosome,
- In QTL mapping, LD is generated by the **mating design**
- In a mapping population, LD is influenced only by **recombination** in the absence of segregation distortion,
- LD - refers to the **correlation** between alleles in a population
- In AM, LD is a reflection of the **germplasm collection** under study.
- In AM, LD is influenced by selection, mutation, mating system, population structure as well as by recombination.

# **GWAS METHODOLOGY**

## **GWAS METHODOLOGY**

1. Population and phenotyping
2. Genotyping (marker information)
3. Imputation/ MAF Filtering
4. Covariates - Population Structure/Kinship
5. Regression – Multiple models
6. Model testing
7. QTL and significant markers
  1. Significant markers for MAS
  2. Stepwise Regression
  3. Candidate genes
8. Validation

## **I. POPULATION**

- Population with overall diversity
  - Ex: Landraces through out world
- Regional Population with local diversity
  - Cultivars used in ND

## **PHENOTYPING**

- Data from multiple locations
  - Analysis for individual location
  - Analysis for combined locations
    - Using adjusted entry means (LS means)
      - Analysis is fast
    - Using location, replication etc as covariates in the linear model step
      - Data multiplied and analysis is slow

## **ADJUSTED ENTRY MEANS (LSMEANS)**

- Example: 143 genotypes, 3 locations, 4 reps, 2 ratings, scale 1-5
- Significant difference between locations, has G X E interaction

Genotype	Mean	Adjusted entry mean
ID1	2.77	2.76
ID10	3.27	3.25
ID101	1.64	1.46
ID105	2.54	2.51
ID107	3.13	3.07

**Adjusts for significant effects**

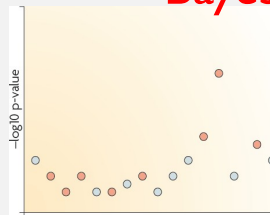
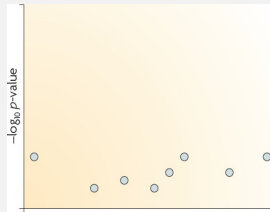
## 2. GENOTYPING

- Low density
  - Ex: 200 (few) SSRs
- Medium density
  - Golden gate assay – 1536 markers
- High Density
  - Arabidopsis 250,000 SNPs on Affymetrix chip
  - Next generation sequencing: thousands/millions
    - Example: ~ 1x coverage in soybean, 79000 snp
    - ~ 3x coverage/individual in a wild bean pool – 10 million snp



### 3. IMPUTATION

- Predicting missing genotypes based on likelihood
- Increases power of AM
- All markers considered in mixed model, since mixed model does not run with missing data variables
- Facilitates Meta analysis
- Before and after Imputation
- Likelihood based – PHASE, fastPHASE, BEAGLE
- Bayesian based – BIMBAM



Marchini and Howie 2010



#### A) Raw data

Ind1	A	T	T	G	C	?	C	T	G	C
Ind2	G	C	C	A	T	T	G	A	A	?
Ind3	A	?	T	G	C	G	C	T	G	C
Ind4	G	?	C	A	T	T	G	A	A	C
Ind5	G	?	C	A	T	T	G	A	?	C

#### B) Make Haplotypes

Ind1	A	T	T	G	C	?	C	T	G	C	}
Ind3	A	?	T	G	C	G	C	T	G	C	
Ind2	G	C	C	A	T	T	G	A	A	?	}
Ind4	G	?	C	A	T	T	G	A	A	C	
Ind5	G	?	C	A	T	T	G	A	?	C	

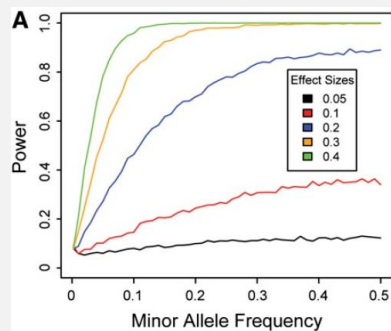
#### c) Impute Missing and estimate likelihood

Ind1	A	T	T	G	C	<b>G</b>	C	T	G	C
Ind3	A	<b>T</b>	T	G	C	G	C	T	G	C
Ind2	G	C	C	A	T	T	G	A	A	<b>C</b>
Ind4	G	<b>C</b>	C	A	T	T	G	A	A	C
Ind5	G	<b>C</b>	C	A	T	T	G	A	?	C

Repeated 'n' times with different starts and finally outputs one with highest likelihood

## MINOR ALLELE FREQUENCY (MAF)

- Sequencing errors
- Leads to biased LD calculations
- Rare variants have little effect to discover QTL

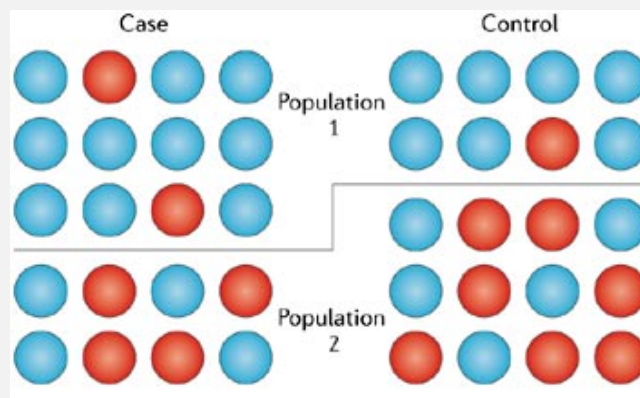


Myles et al. 2009, The Plant Cell

## 4. POPULATION STRUCTURE/KINSHIP

- Estimate covariates that control for population structure
- Structure (software)
  - Assumes Hardy Weinberg
  - Lots of time
- PCA
  - Easy and most used

- To avoid false positives





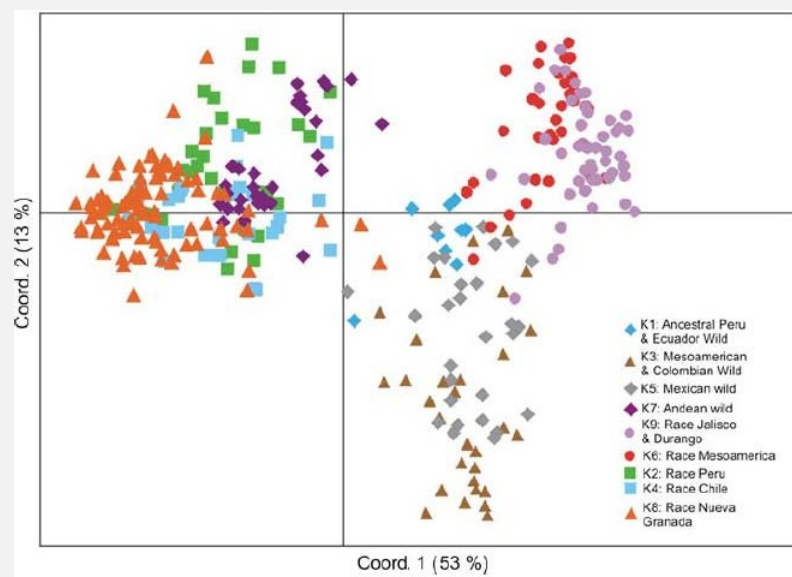
- Use sub population membership as covariates

Genotype	Subpop 1	Subpop 2	Subpop 3
Ind 1	0.9	0.02	0.08
Ind 2	0.2	0.7	0.1
Ind 3	0.1	0.85	0.05

- Estimation:
  - Using Bayesian approaches like STRUCTURE, Instruct, BAPS, Admixture
    - Assume Hardy-Weinberg equilibrium
    - Time consuming
    - Subset of markers selected
      - Based on LD or a random set

## PRINCIPAL COMPONENT ANALYSIS (PCA)

- Infer continuous axes of genetic variation.
- Data reduced to a small number of dimensions
- Summarizes the overall variability among individuals, which includes both
  - the divergence between groups (*i.e.*, structured genetic variability), and
  - the variation occurring within groups ('random' genetic variability).



Kwak et al. 2009



## NUMBER OF PRINCIPAL COMPONENTS

- Based on amount of variation explained
  - 25 %
  - 50 %
- Statistically
  - Based on P-value from Tracy Widom distribution (Patterson et al. 2006, Plos genetics)
  - Velicers Minimum Average Partial test (MAP) (Shriner 2011, Heredity)
- Others
  - Top 10
  - Top 3
- Recently using stepwise to determine which PCs
  - PC1, PC3, PC7 .....

## KINSHIPS

- A random effect in mixed model
- Used to calculate Genetic variance and residual variance
- Population structure describes remote common ancestry of large groups of individuals,
- Relatedness refers to recent common ancestry among smaller groups (often just pairs) of individuals.

- **Identity by Descent (IBD)**

- Pedigree analysis
  - Pedigree is available for all genotypes
- Loiselle, Ritland coefficients
  - Genotypes are derived from a small subset of parents

- **Identity by State (IBS)**

- Similarity matrix
  - % of shared alleles or distance
- Descent separated from Similarity
  - Similarity can be due to descent
  - Multiple steps of 'k'

## 5. REGRESSION - LINEAR MODELS

- GLM – Only has fixed effects
- MLM – Has both fixed and random effects

$$y = X\alpha + Q\beta + K\gamma + \epsilon$$

$y$  is a vector for phenotypic observations

$\alpha$  is the fixed effects related to the SNP marker

$\beta$  is a vector of the fixed effects related to the population structure,

$\gamma$  is a vector of the random effects related to the relatedness among the individuals,

$\epsilon$  is a vector of the residual effects.  $X$  is genotypes of the SNP markers,  $Q$  is the matrix of the subpopulation,  $K$  is the kinship matrix

The variances of the random effects were estimated as  $\text{Var}(u) = 2KV_g$  and  $\text{Var}(e) = IV_R$ , where  $K$  is a kinship matrix,

$I$  is an identity matrix,

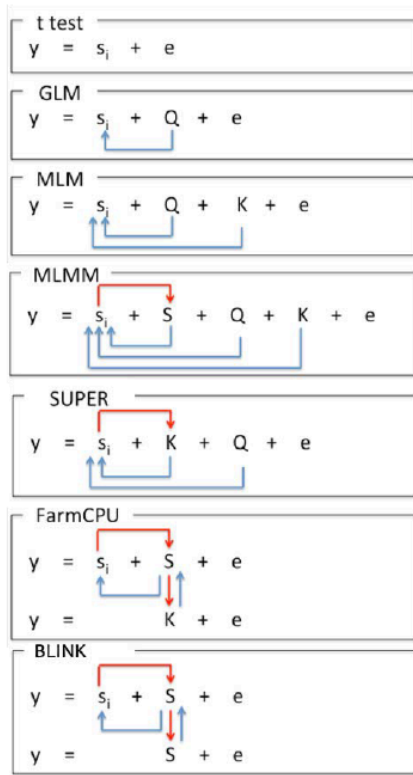
$V_g$  is the genetic variance,

$V_R$  is the residual variance.

- Regression analysis performed for one marker at a time
- $V_g$  and  $V_R$  calculated for each marker – **Exact Tests**
  - Implemented in SAS, EMMA, FAST-LMM, GEMMA, TASSEL
- If calculated only once to save computational burden –  
**Approximate tests**
  - Called as P3D/EMMAX algorithm
  - Implemented in TASSEL, EMMAX, GAPIT, GRAMMER

## **OTHER COMMON MODELS USED**

- Wilcoxon test
  - A non parametric t-test
  - No effects included
- Logistic regression
  - When phenotype is logical (0 or 1)
- Generalized linear model
  - When phenotype is discrete (Ex: scale of 1 -5) or logical
  - Can include both fixed and random effects



$s_i$ : Testing marker     $Q$ : Population structure     $K$ : Kinship  
 $S$ : Pseudo QTNs     $Q$ : Population structure    Arrow: adjustment

## MULTIPLE MODELS

- The best model is dependent of
  - Distribution of Phenotype
  - # of QTL governing Phenotype
  - Genotypes used
  - Sub population structure among individuals
  - Relatedness between individuals
    - Examples
      - Phenotype = Marker
      - Phenotype = Marker + Structure
      - Phenotype = Marker + Kinship
      - Phenotype = Marker + Structure + Kinship

**Multiple Models suggested for every phenotype and/or population**



## 6. MODEL SELECTION

- The residual errors follow a uniform distribution
  - MSD for each model for p-observed with p-expected
  - Q-Q Plot
- MSD for each model
  - For each value sort p-values
  - Assign a rank (1,2...n), where n is number of markers
  - Expected p-value = Divide rank with 'n'
  - Subtract Expected p-value from Observed p-value
  - Square the difference and calculate mean

## 7. MARKER SIGNIFICANCE

- Bonferroni correction
  - $0.05 / \# \text{ of markers}$
- Multiple correction (FDR, pFDR)
- Experimental error using permutations
  - A simple regression model with 'n' permutations
- Bootstrapping

## STEPWISE REGRESSION

- To account for variation explained by all markers
- Minimize # of QTL/markers for MAS
- Considers LD (interchromosomal and intra chromosomal)
- No need of covariates
- Gives combined effect of all markers

## EPISTASIS

Example:

- $m1$  and  $m2$  are significant
- In ANOVA include  $m1*m2$  interaction effect
- Ex: With 33 significant markers at 0.1 percentile, 528 two way interactions were tested with a population structure component, and five interactions were significant at  $p\text{-value} < 0.001$ .

## VARIATION EXPLAINED

- Regression  $R^2$
- Likelihood based including effects ( $R^2_{LR}$ ; Sun et al. 2010)

$$R^2_{LR} = 1 - \exp\left(-\frac{2}{n}(\log L_M - \log L_0)\right)$$

Log  $L_M$  is likelihood of Full model

Ex:  $y = m_I + \text{PCA} + \text{kinship} + \varepsilon$

Log  $L_0$  is likelihood of reduced model

Ex:  $y = \text{PCA} + \text{kinship} + \varepsilon$

## EFFECT OF ALLELE

- The additive effect of the variant allele is calculated as half the difference between mean of the variant allele and mean of the reference allele.
- Ref allele mean = 10
- Var allele mean = 16
- Additive effect = 3

## **SIGNIFICANT MARKER ALLELIC COMBINATIONS**

- Lets assume 3 significant markers
- 8 possible combinations (for SNP)
  - AAA 2.5
  - AAB 2.7
  - ABA 3.9
  - ABB 4.6
  - BAA 3
  - BAB 3.1
  - BBA 1.9
  - BBB 4.2

## FUNCTIONAL ASPECTS

- Annotate SNP - snpEff

### QTLs

- A QTL region is defined as the region around significant stepwise markers that has a partial LD ( $r^2$ )  $>0.6$  with the adjacent marker (corresponds to partial correlation of 0.77).
- Adjacent blocks were combined if the distance is less than 10 kb and includes up to four markers that are not in LD with either of the block.



Phenotype



Means/Adjusted Means

Markers



Imputation



MAF



PCA and Kinship

Regression (Multiple Models)

Model Selection (MSD)

Significant Markers (Bootstrap/Bonferroni/FDR)

Stepwise Regression

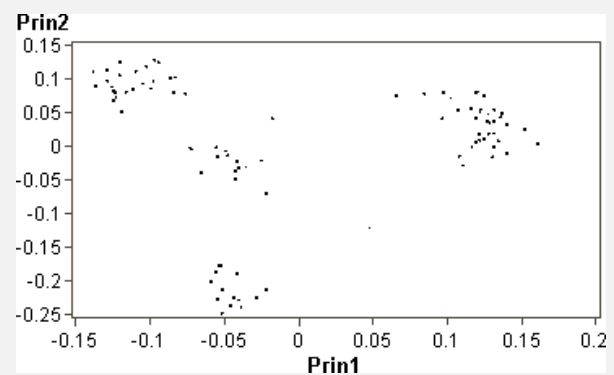
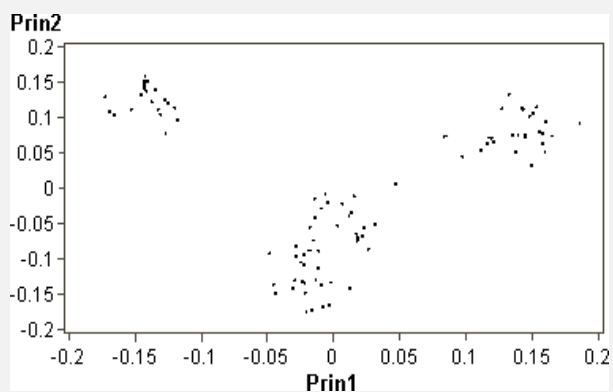
Allelic Effects; QTL regions; SNP Annotation

## WHAT IS NEXT AFTER AM

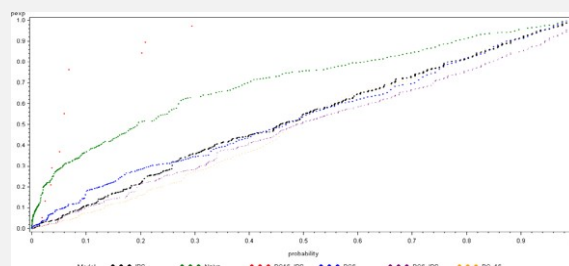
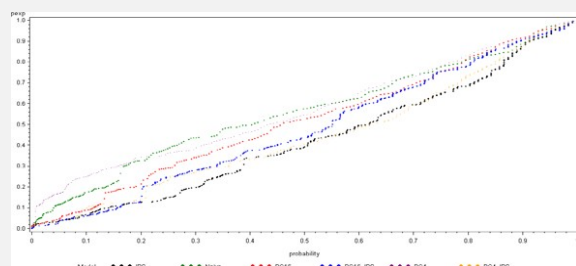
- **Evaluation:** Markers should be evaluated on several different genetic backgrounds
- **Verification:** The association reported should be verified either through re-evaluation in an independent population sample or through allelic silencing / knock-outs.
- **Breeding:** The best alleles obtained through the study can be used for MAS.

## I) CD ACCUMULATION IN DURUM WHEAT

- 96 near inbred lines (NILs) grown in the field in 2009 and 2010
- Screened for Illumina iSelect BeadChip platform for 9000 markers



2009		2010	
Model	MSD <sup>†</sup>	Model	MSD <sup>†</sup>
PC16	0.00093	IBS	0.00688
IBS	0.001166	Naive	0.005768
PC6+IBS*	0.000283	PC16	0.000971
PC6	0.001708	PC16+IBS*	0.000539
Naive	0.050978	PC4	0.006497
PC16+IBS	0.237218	PC4+IBS	0.004963

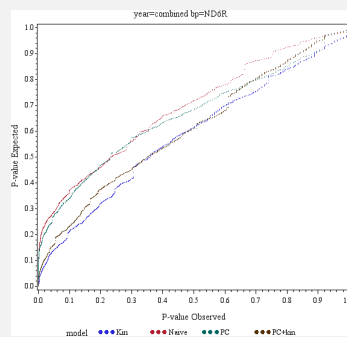
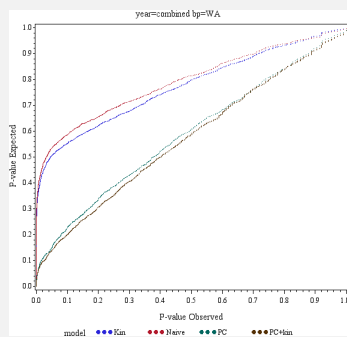
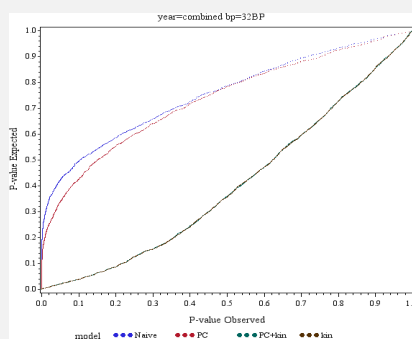


Year	SNP marker	Chrom	Genetic position (cM)	$-\log_{10}(p)$	R <sup>2</sup>	MAF
2009	Ex-c1996-3754394	2B	7.25	3.52	3.04	47.92
	Ra-rep-c106727-90434958	2B	7.25	3.52	3.04	47.92
	Ex_c2098_3934284	-	-	3.52	3.04	47.92
2010	Ex-c17754-26503892	5B	165.7	5.26	33.74	12.50
	Ex-c20019-29052512	5B	178.3	3.99	26.96	11.45

## **2) GERMINATION IN BARLEY:**

**DR.HORSLEY ,ANA MARIA, SUJAN**

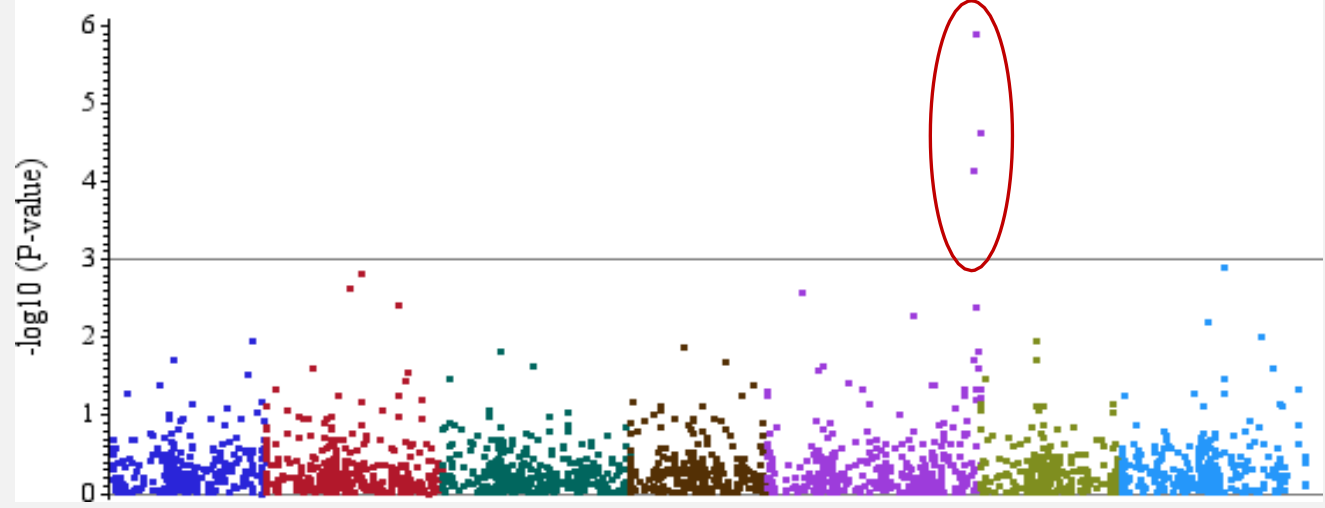
- 8 breeding programs
- 4 years
- 96 lines each
- 4 regression models

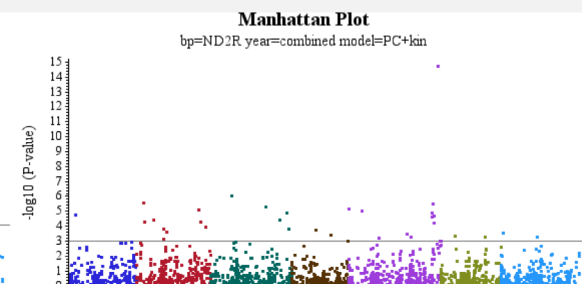
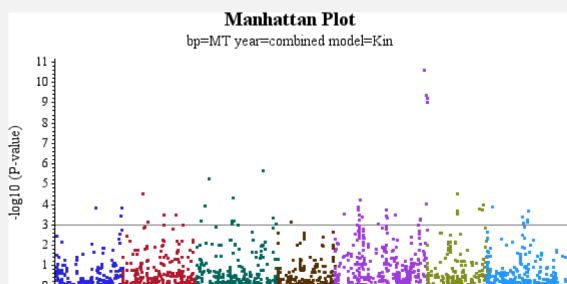
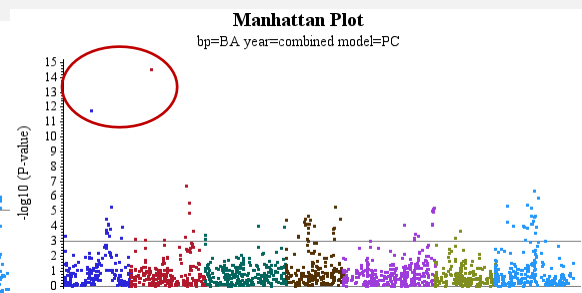
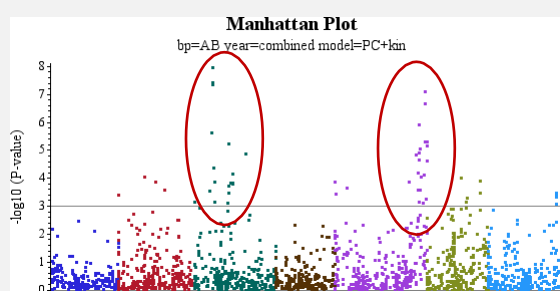




# Manhattan Plot

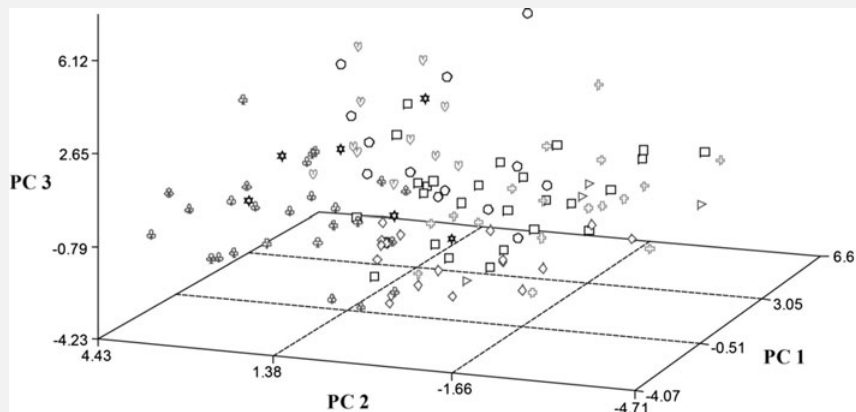
bp=32BP year=combined model=PC+kin





### 3) SEED WEIGHT IN LUPINUS

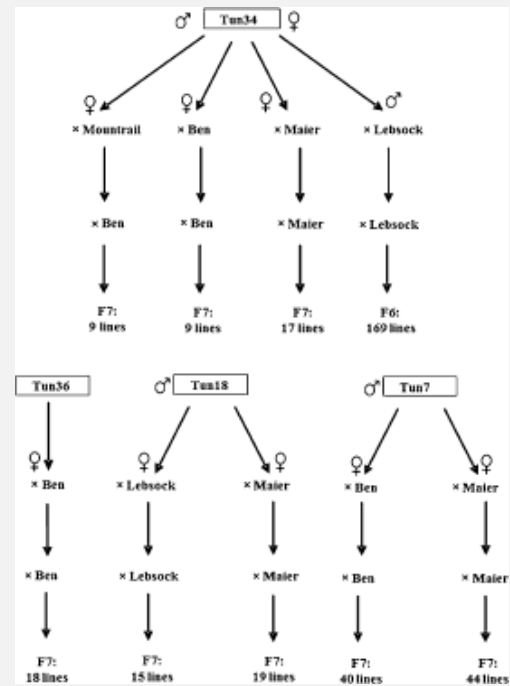
- 122 Plant introductions (PI) lines
- 2277 AFLP bands, 892 polymorphic



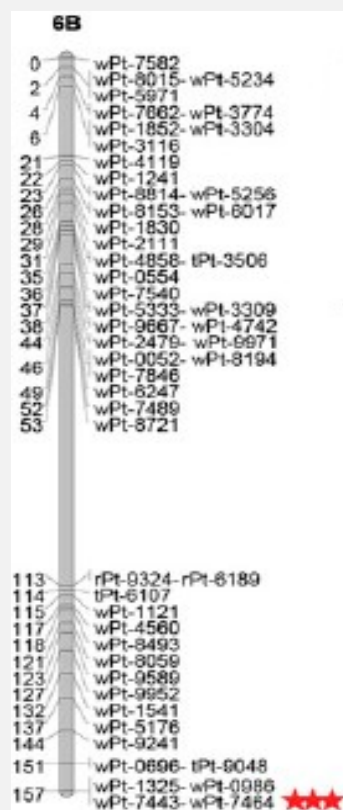
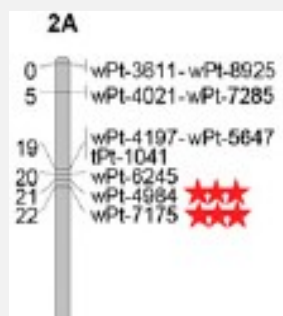
Marker	MAF	Minor allele mean	Major allele mean	p value	R <sup>2</sup>
ECAGMCGC76	0.148	24.65	30.65	1.53E-04	22.69
ECACMCGC105	0.459	27.99	31.3	2.30E-04	20.50

#### 4) FHB IN WHEAT

- 171 BC<sub>1</sub>F<sub>7</sub> and 169 BC<sub>1</sub>F<sub>6</sub> lines
- 2300 Dart markers



Gavami et al. 2012



## 5) IRON DEFICIENCY CHLOROSIS (IDC) IN SOYBEAN

- ~ 140 advanced breeding lines of ND
- 1536 golden gate assay

Model	Statistical model
Naïve	$y = X\alpha + \epsilon$
$K$	$y = X\alpha + K\nu + \epsilon$
$K^*$	$y = X\alpha + K^*\nu + \epsilon$
$Q$	$y = X\alpha + Q\beta + \epsilon$
PCA	$y = X\alpha + P\beta + \epsilon$
$Q + K$	$y = X\alpha + Q\beta + K\nu + \epsilon$
$Q + K^*$	$y = X\alpha + Q\beta + K^*\nu + \epsilon$
PCA + $K$	$y = X\alpha + P\beta + K\nu + \epsilon$
PCA + $K^*$	$y = X\alpha + P\beta + K^*\nu + \epsilon$

Mamidi et al. 2011

BARC SNP marker	Chromosome	2005							2006						
		$-\log_{10}(p)$	pFDR <sup>†</sup>	R <sup>2</sup> (%)	Minor allele frequency	Minor allele mean	Major allele mean		$-\log_{10}(p)$	pFDR	R <sup>2</sup> (%)	Minor allele frequency	Minor allele mean	Major allele mean	
BARC-029969-06762	2	2.707	0.046	11.8	49.0	3.0	2.7		7.531	0.000	24.3	40.4	2.9	2.5	
BARC-044603-08734	3	3.661	0.022	14.9	40.6	2.7	3.0		4.898	0.001	15.7	45.4	2.5	2.8	
BARC-060109-16388	3	3.321	0.024	16.3	46.2	2.7	3.0		3.781	0.003	13.9	47.5	2.5	2.8	
BARC-016535-02085	3	2.886	0.046	15.3	46.9	2.7	3.0		3.781	0.003	13.9	47.5	2.5	2.8	
BARC-010457-00640	6	2.185	0.076	0.1	39.9	2.8	2.9		4.018	0.002	2.6	44.7	2.6	2.7	
BARC-039383-07310	7	3.114	0.034	17.8	21.7	2.5	3.0		4.548	0.002	13.5	21.3	2.3	2.7	
BARC-025897-05144	13	2.309	0.067	15.2	38.5	3.1	2.7		4.008	0.002	15.9	37.6	2.9	2.5	
BARC-055499-13329	13	2.420	0.058	9.4	31.5	3.1	2.8		2.602	0.021	13.4	27.0	2.9	2.6	
BARC-059723-16418	19	3.415	0.022	15.4	44.8	2.7	3.0		4.124	0.002	10.7	31.2	2.4	2.8	

All 9 markers combined explain 43.7 % in 2005 and 47.6 % in 2006



BARC marker	Chromosome	SNP <sup>†</sup> position (bp)	2005			2006			At <sup>‡</sup> gene	Gm <sup>§</sup> gene model	Start of model (bp)	Distance from SNP (bp)	E-value	Percent identity
			-log10 (p)	pFDR	R <sup>2</sup> (%)	-log10 (p)	pFDR	R <sup>2</sup> (%)						
BARC-060109-16388	3	45,391,018	3.32	0.02	16.26	3.78	0.00	13.94	NAS3	Glyma03g39050	45,279,921	111,097	5.00 E × 10 <sup>-109</sup>	62.9
BARC-053261-11776	5	937,302	2.21	0.08	0.72				AHA2	Glyma05g01460	960,820	23,518	0	81.1
BARC-021775-04203	5	41,114,078	0.08	0.53	0.07	2.47	0.03	9.30	BT12-ITP	Glyma05g37300	40,906,083	207,995	1.00 E × 10 <sup>-95</sup>	65.7
BARC-054331-12480	7	8,652,831				1.72	0.09	4.78	BT12-ITP	Glyma07g10870	9,082,076	429,245	2.00 E × 10 <sup>-45</sup>	36.8
BARC-049147-10810	9	35,895,343	0.44	0.38	0.73	2.07	0.05	1.50	YSL7	Glyma09g29410	36,298,317	402,974	0	54.6
BARC-062275-17736	11	38,020,165	0.02	0.56	1.64	2.63	0.02	0.18	FER4	Glyma11g35610	37,245,783	774,382	7.00 E × 10 <sup>-103</sup>	73.2
BARC-017917-02456	13	30,457,599	2.17	0.08	7.83				FRD3	Glyma13g27300	30,477,485	19,886	2.00 E × 10 <sup>-137</sup>	54.7
BARC-043041-08509	15	48,694,193	2.71	0.05	12.66	0.94	0.26	9.24	IRT1	Glyma15g41620	48,764,845	70,652	8.00 E × 10 <sup>-80</sup>	43.5
BARC-030595-06910	16	3,039,691	0.09	0.53	4.55	1.93	0.06	7.72	FRD2	Glyma16g03770	3,142,668	102,977	0	57.7
BARC-011625-00310	16	36,544,010	0.26	0.45	0.47	2.28	0.04	2.79	YSL7	Glyma16g33840	36,609,082	65,072	0	73.4
BARC-043087-08524	17	4,899,023	3.44	0.02	0.25	1.37	0.16	1.35	AHA2	Glyma17g06930	4,977,823	78,800	0	87.7
BARC-012289-01799	18	1,957,710				1.78	0.08	2.07	FER4	Glyma18g02800	1,821,344	136,366	6.00 E × 10 <sup>-102</sup>	75.1
BARC-016867-02359	18	56,429,447	1.26	0.19	2.77	2.11	0.05	6.63	FRD2	Glyma18g47060	56,712,622	283,175	0	52.2
BARC-059723-16418	19	40,357,687	3.42	0.02	15.41	4.12	0.00	10.68	OPT1	Glyma19g32400	40,140,993	216,694	3.00 E × 10 <sup>-146</sup>	47.4
BARC-042281-08231	20	343,106	0.22	0.47	2.33	1.96	0.06	8.58	YSL7	Glyma20g00690	418,225	75,119	0	62.8

## 6) IDC – 9 LOCI CONFIRMATION -SUJAN MAMIDI

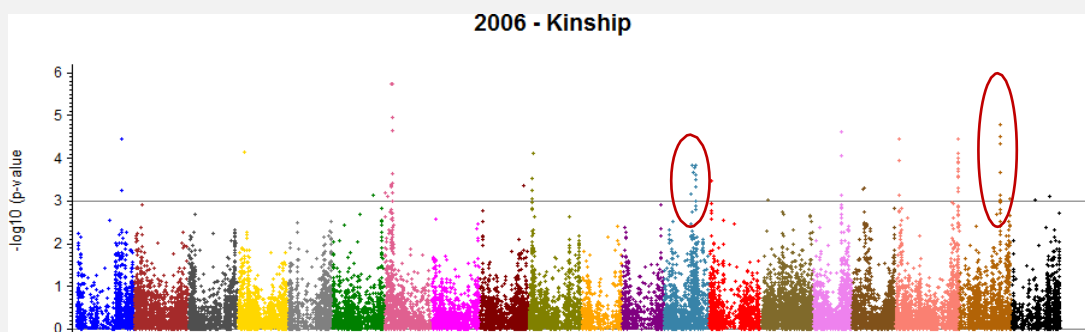
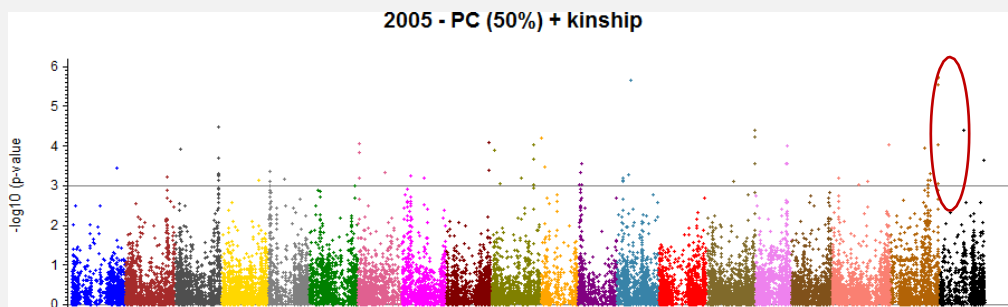
	Original Population		Confirmation Populations		
	2005	2006	2008	2009	2010
# of Genotypes	143	141	265	201	245
# of Locations	5	4	5	3	5
# of Reps	4	4	4	4	4
Range of IDC score	1.5-3.8	1.6-3.8	1.9-4.1	1.6-4.2	1.6-3.1
Average IDC Score	2.9	2.7	3.1	3	2.3

Marker		029969-06762	016535-02085	044603-08734	060109-16388	010457-00640	039383-07310	025897-05144	055499-13329	059723-16418
Ch		2	3	3	3	6	7	13	13	19
Mbp		2.45	45.42	45.00	45.39	45.28	7.15	27.14	31.47	40.35
2005	p-val	2.71	3.66	3.32	2.89	2.19	3.11	2.31	2.42	3.42
	R-sq	11.80	14.90	16.30	15.30	0.10	17.80	15.20	9.40	15.40
2006	p-val	7.53	4.90	3.78	3.78	4.02	4.55	4.01	2.60	4.12
	R-sq	24.30	15.70	13.90	13.90	2.60	13.50	15.90	13.40	10.70
2008	p-val	3.15	11.39	8.11	11.39	1.35	3.77	8.27	3.06	6.65
	R-sq	8.71	14.28	12.75	14.28	0.85	11.11	17.13	10.75	10.97
2009	p-val	0.75		2.68		0.26	0.77	1.95	1.12	1.74
	R-sq	6.74	9.54	3.78	9.54	0.59	5.93	4.86	4.26	6.88
2010	p-val	0.71	6.13	5.47	6.13	0.44	0.55	1.39	4.15	7.85
	R-sq	3.71	16.49	14.73	16.49	3.19	6.74	3.02	9.98	18.07

## **7) IDC – GBS DATA**

### **-PHILLIP MCCLEAN AND SUJAN MAMIDI**

- # of Genotypes: 132 and 138 in 2005 and 2006
- 79000 data points obtained
- 34,435 and 35185 SNP at 0.05 MAF



## REFERENCES:

- Abdurakhmonov and Abdukarimov 2008. International Journal of Plant Genomics:574927
- Astle and Balding. 2009. Statistical Science 24: 451-471
- Balding 2006. Nature Reviews Genetics 7:781
- Flint-Garcia et al. 2003. Annu. Rev. Plant Biol.54:357-74
- Gupta et al. 2005. Plant Molecular Biology 57:461-485
- Hamblin et al.2011. Trends in Genetics 27: 98
- Myles et al. 2009. The Plant Cell, 21: 2194-2202
- Rosyara UR and Joshi BK, 2012. Nepal Journal of Biotechnology, 2: 72-89
- Zhu et al. 2008. The Plant Genome 1:5-20