The 8th International Purdue Symposium on Statistics
Session: Interactions Between Omics and Statistics: Analyzing High Dimensional Data
Sunday, June 24, 11:00am - 11:30am

Opportunities and Challenges of Statistical Genetics in Genome-wide Association Studies

Jianming Yu,
Department of Agronomy, Kansas State University

Outline

- GWAS: Opportunities and challenges
- Rare allele (CR-GWAS)
  - Arabidopsis
- Genomic distribution of trait-associated SNPs
  - Maize
- Examples of functional haplotypes
  - Sorghum

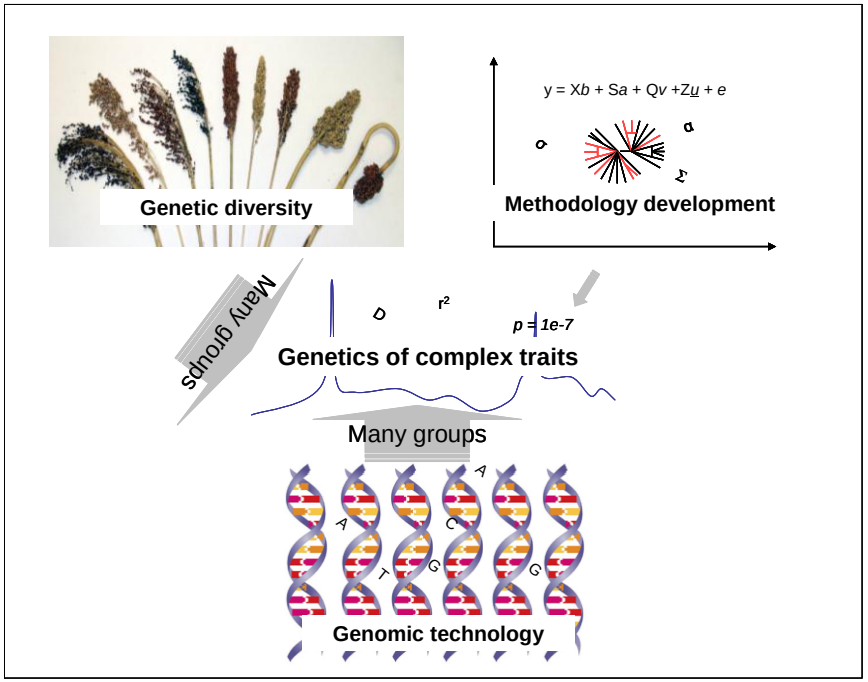
Geno→Pheno & Geno↔Pheno

- Genome-wide association study (GWAS)
  - Gene identification
  - Finding association "**signals**" with a large set of SNPs across diverse germplasm
  - Array-based genotyping & resequencing

- Genome-wide selection (GS)
  - "**Prediction**", breeding, genetic gain
  - Selection of individuals based on predicted phenotypic values using all markers, rather than only significant markers linked to QTL
  - Integration of genomic technology with plant breeding
    - Bernardo and Yu 2007, Crop Science 47:1082-1090

Opportunities of GWAS

- An additional strategy/tool in gene identification
  - Initiation/validation QTL cloning
  - Hypothesis of new functions of known genes
  - New genes/pathways
- The ability of rapidly nailing down genes for human diseases with relatively simple inheritance is impressive!
- Appreciation of the complexity and beauty of the natural variation
- Has the potential to offer a global view of **genetic architecture of complex traits**



Challenges of GWAS

- Population structure and relative kinship
  - Human genetics, plant and animal genetics
    - Structure; Mixed model QK; Dimension determination/model testing; Accuracy of variance-covariance matrix, R^2_{LR} for mixed model

Yu et al, 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38, 203-208

$$y = X\beta + S\alpha + Qv + Zu + e$$

y: Trait values

X: Environments, etc.

S: Candidate SNP effects

Q: Subpopulation effects

Z: Background Genetic effects, Var(*u*) = 2KV_A

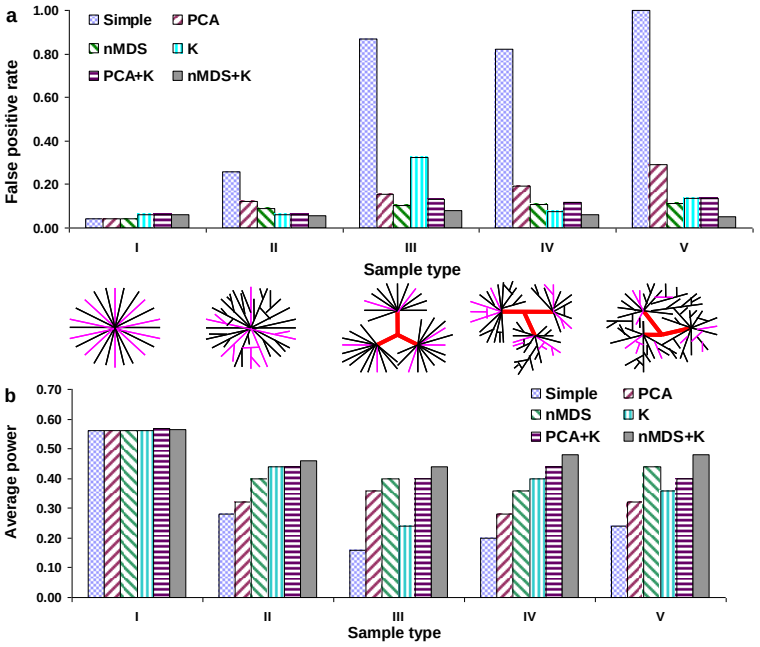
e

Q model

K model

QK model

Simple model



Zhu and Yu 2009, Genetics 182:875-888.

Challenges of GWAS

- Computational demand for mixed model
  - P3D, Compression, EMMA, EMMAX, and Fast-LMM
  - GEMMA, MLMM (Online first Nature Genetics)
- Multiple testing issue and significance threshold
- Missing heritability
- **Rare allele** and epistasis
- Genome, genetics, gene, coding region, allele, hapoltype, SNP, structural variation, etc.

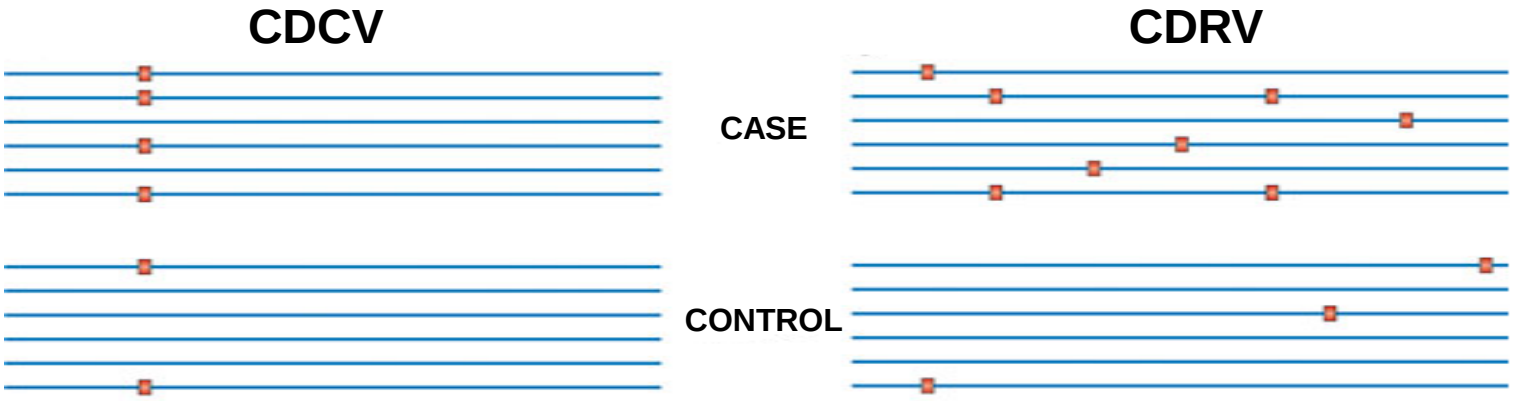
Blame "complex biology and genetics"

I. Composite Resequencing-based GWAS (CR-GWAS)

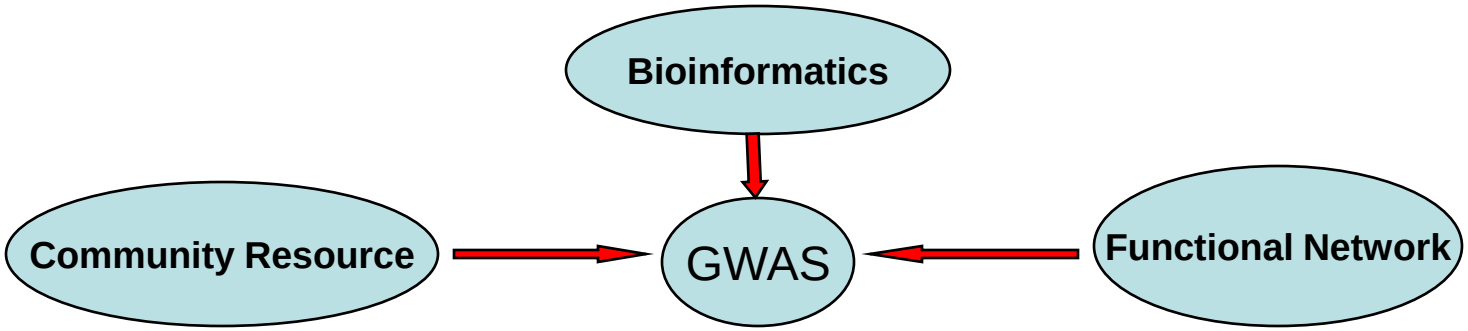
Hypothesis of GWAS: CDCV or CDRV

Rare allele: 4% from Genotyping versus 50% from Resequencing

Single-locus test after controlling for population stratification

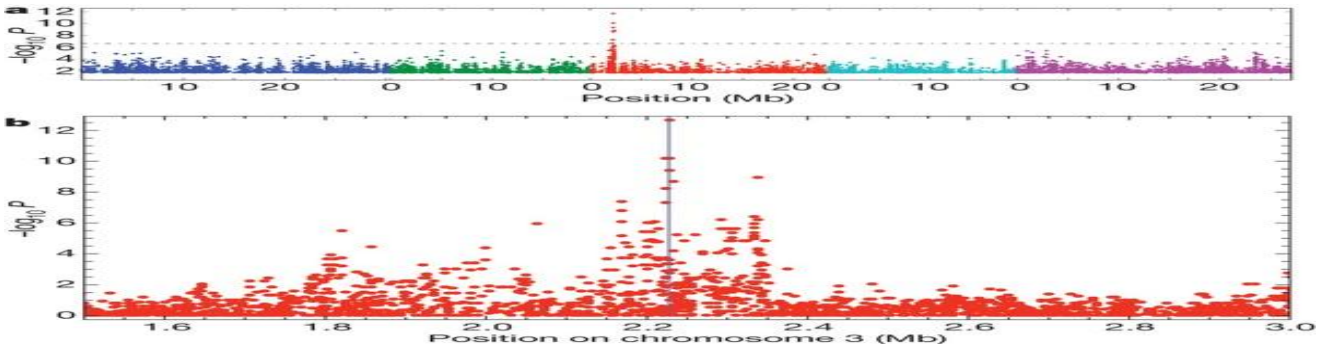


Nat Rev Genet **11, 773-785**

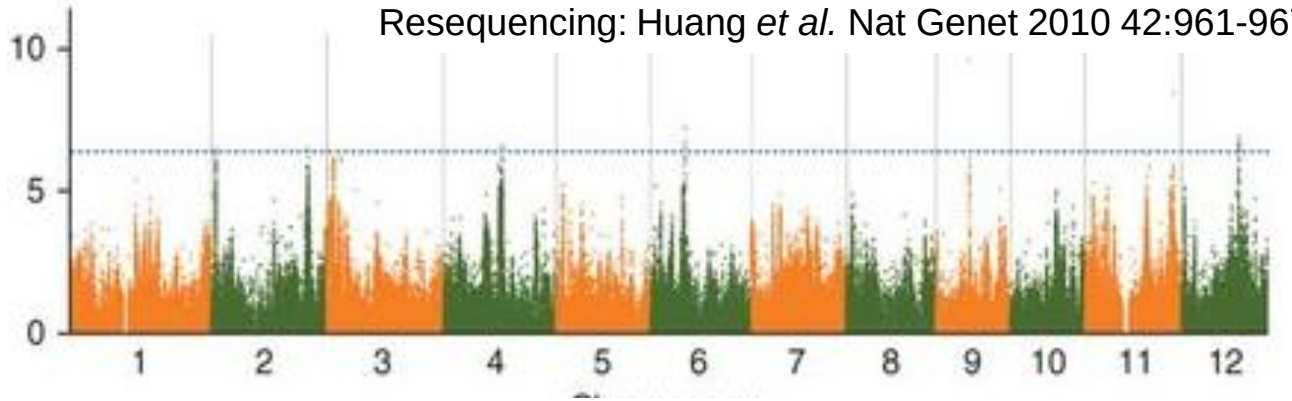


GWAS in Plants

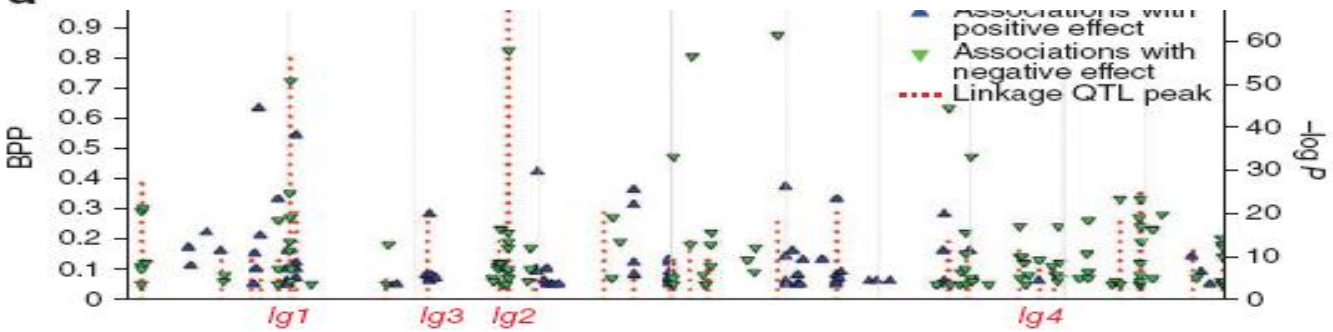
Genotyping: Atwell *et al.* Nature 2010 465:627-631
Resequencing: Zhao *et al.* PLoS Genet 2007 3:e3



Resequencing: Huang *et al.* Nat Genet 2010 42:961-967



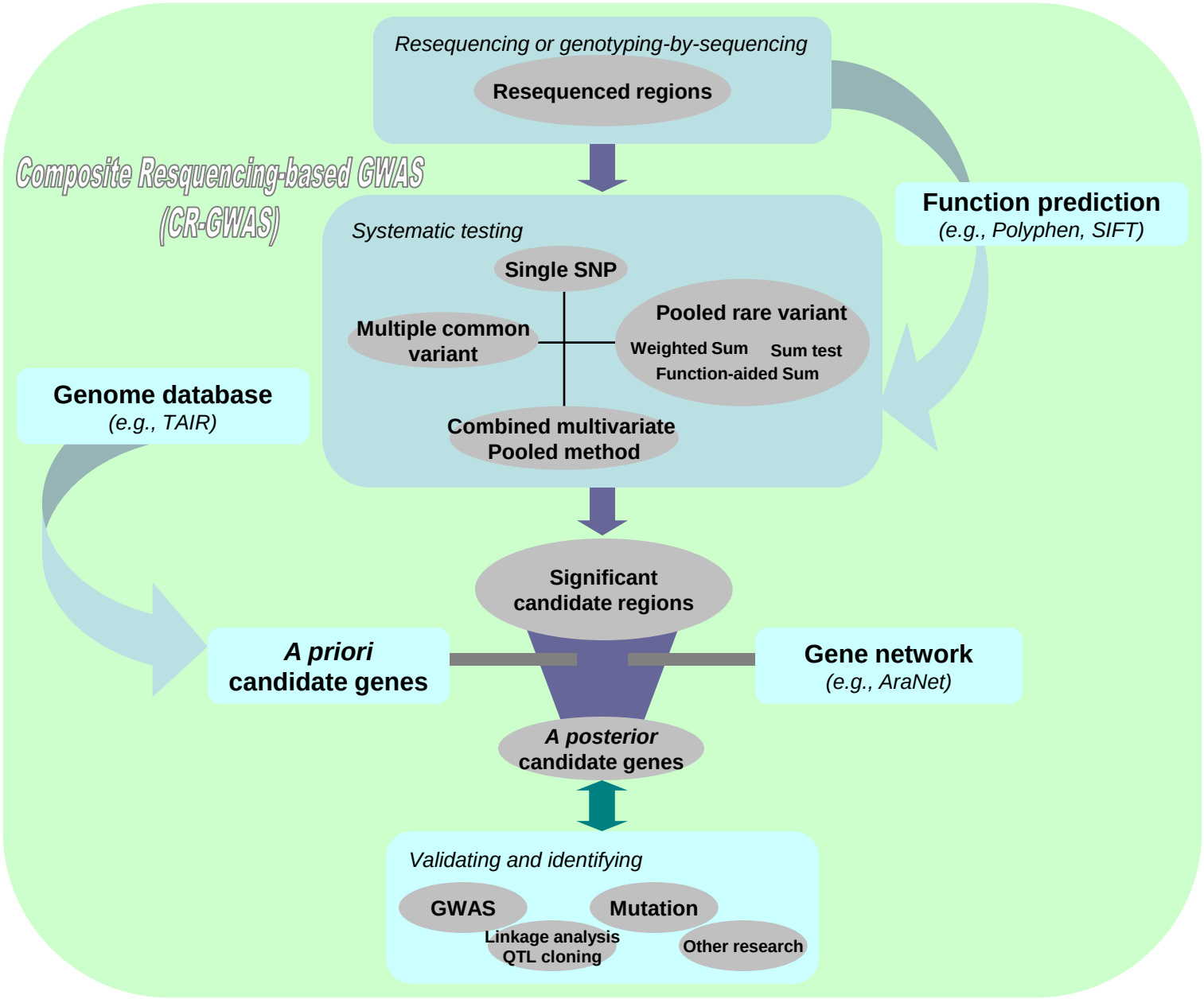
Genetic design + Resequencing: Tian *et al.* Nat Genet 2011 43:159-162
Genetic design + Resequencing: Kump *et al.* Nat Genet 2011 43:163-168

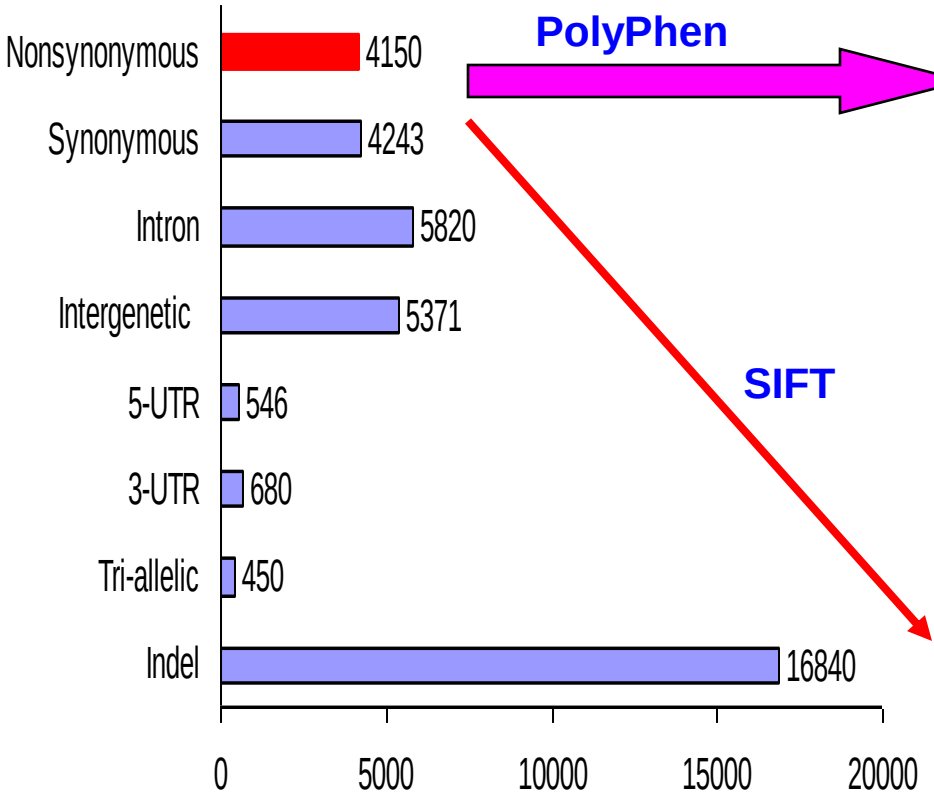


Critical Advances

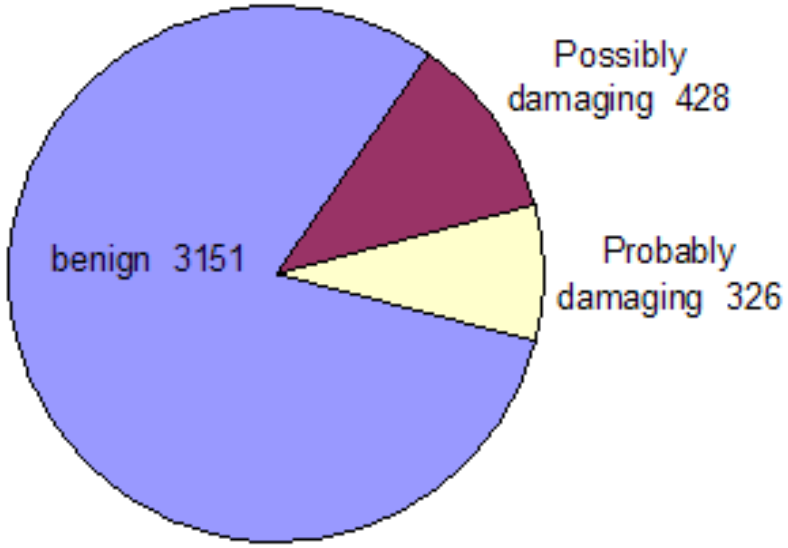
- With next generation sequencing technologies, **exome sequencing or whole genome resequencing** is now possible (*Shendure and Ji, 2008; Ansorge, 2009; Ng et al., 2010*).
- **Biological functions** of nucleotide polymorphisms can be **predicted** with the context sequence of genes (*Ramensky et al., 2002; Kumar et al., 2009*).
- Attention has been given to the **rare allele** issue (*Cohen et al., 2004; Bodmer and Bonilla, 2008; Nejentsev et al., 2009*) and some specific **statistics** have been developed to assess the significance of rare variants (*Morgenthaler and Thilly, 2007; Li and Leal, 2008; Madsen and Browning, 2009; Morris and Zeggini, 2010*).
- **Genome databases** and **gene networks** have been developed to aid the search and confirmation processes of gene-trait associations (*Lee et al., 2008; Lee et al., 2010; Lee et al., 2010*).

Composite Resequecing-based GWAS (CR-GWAS) integrates function prediction, genome database and gene network information, and common and rare variant testing

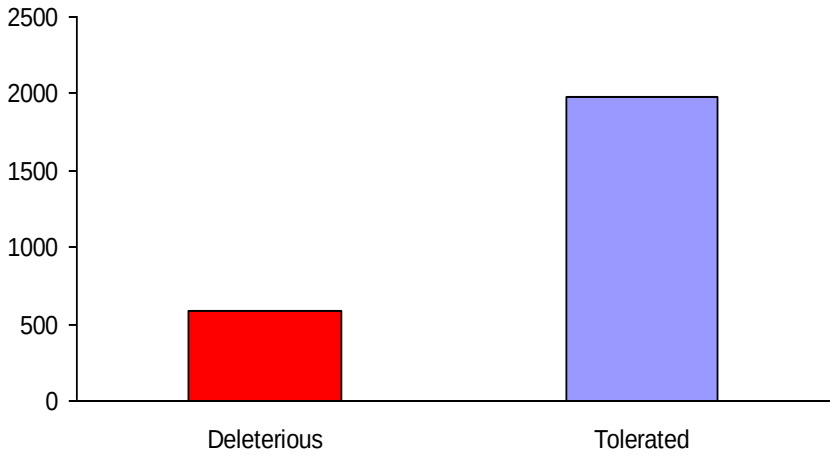




PolyPhen

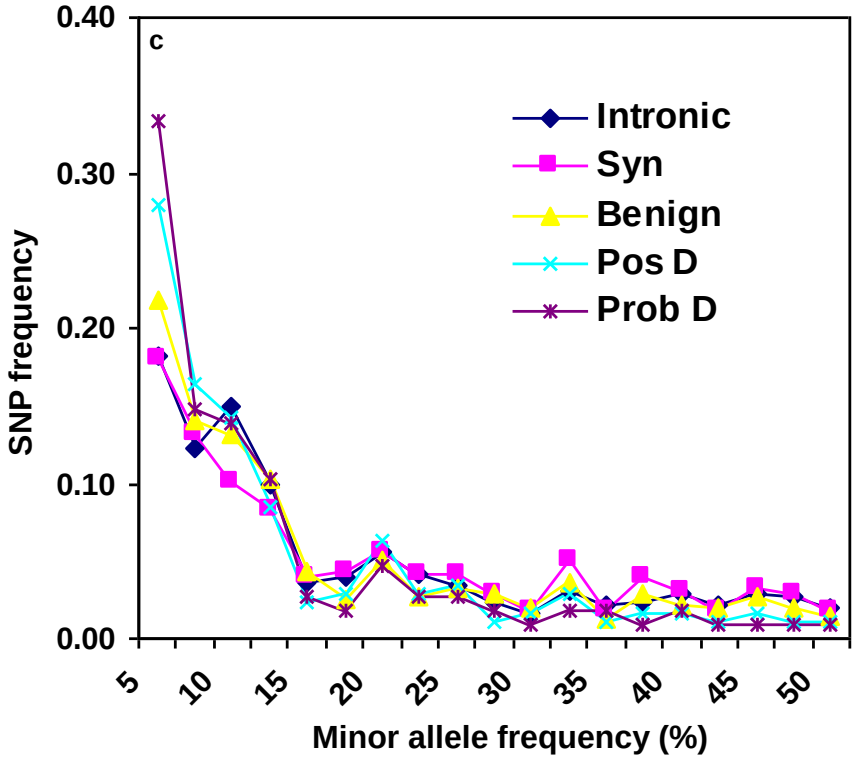


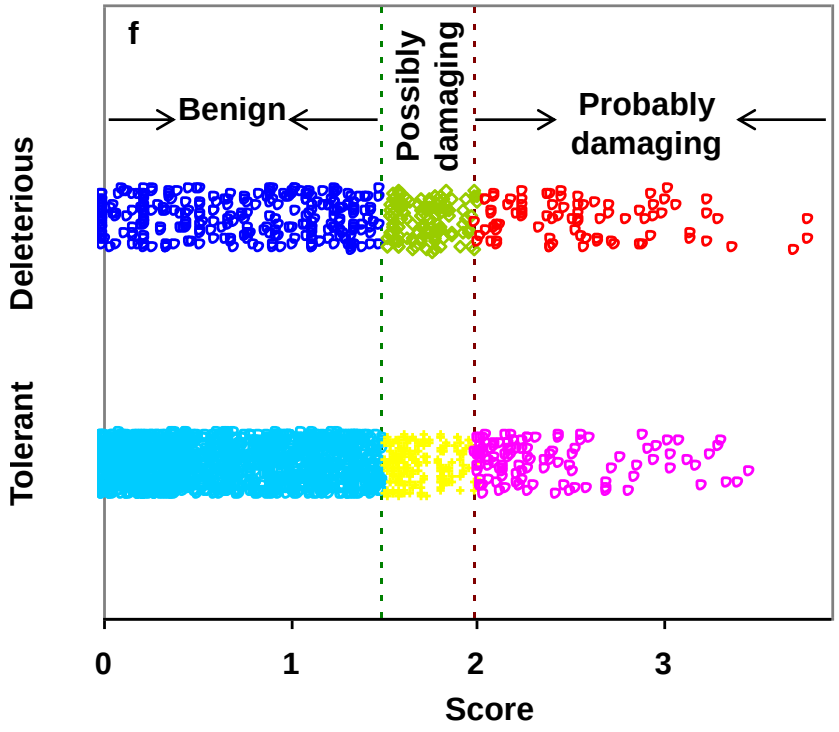
SIFT



Distribution of SNPs with different function prediction across different minor allele frequency

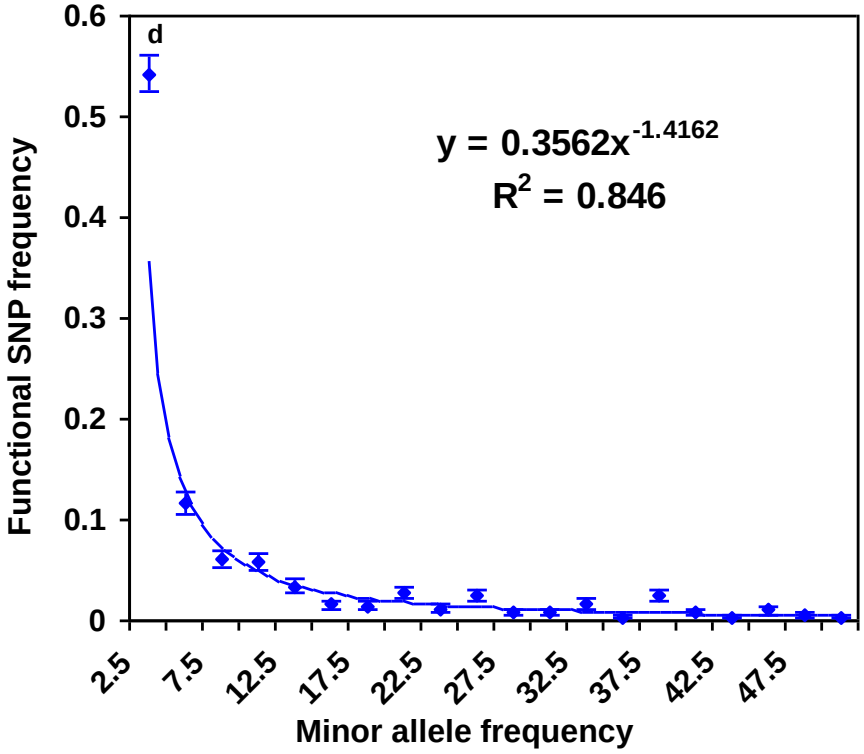
SIFT-predicted function class, deleterious or tolerant, and Polyphen-predicted score value.



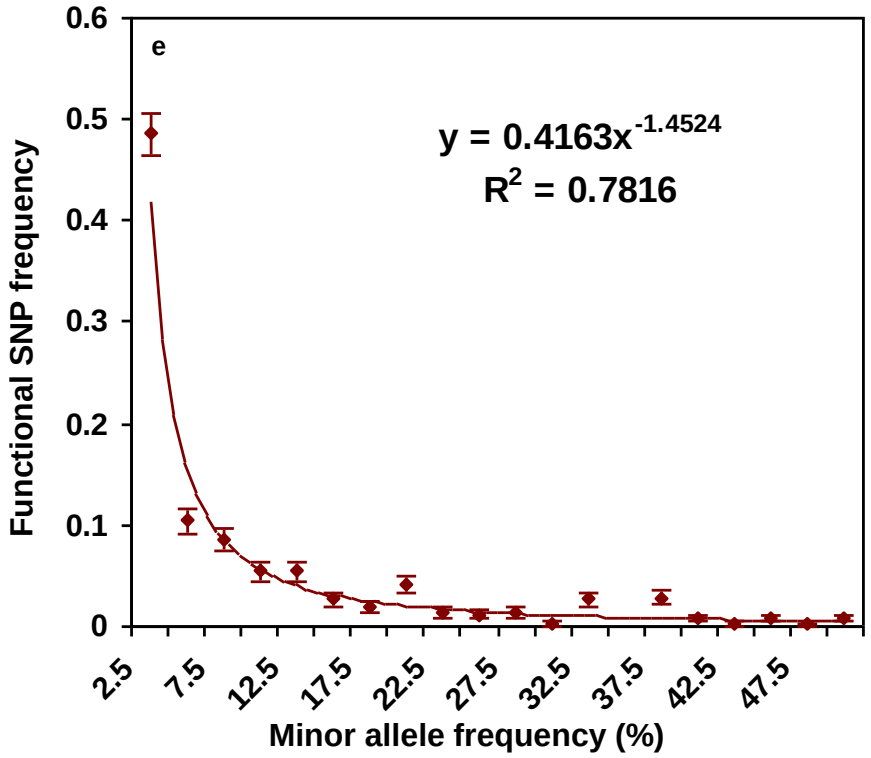


Functional SNP frequency & minor allele frequency

PolyPhen-predicted functional SNP frequency



SIFT-predicted functional SNP frequency



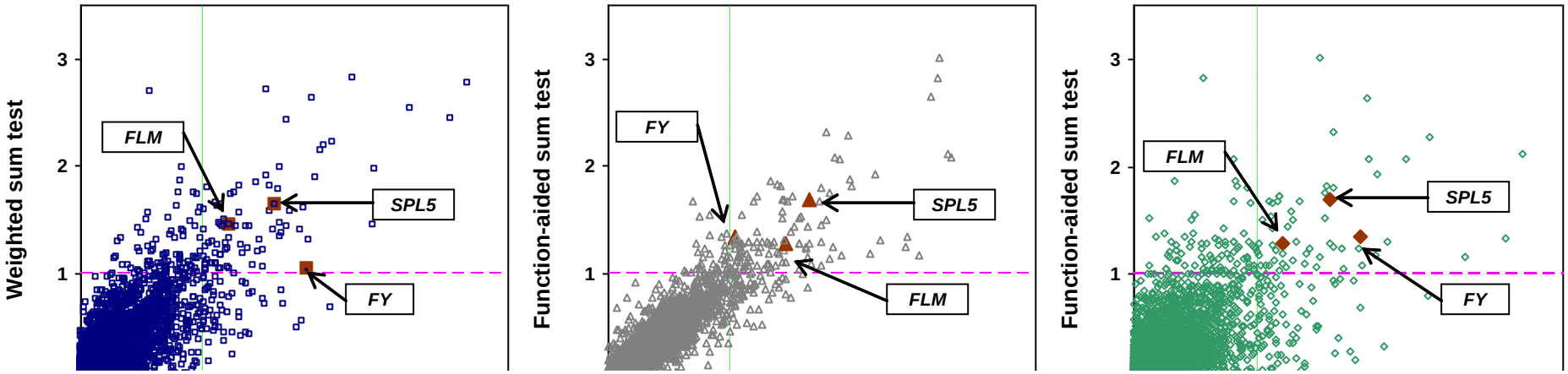
Rare allele statistics

$$y_i = \beta_0 + \beta_1 z_i + e_i$$

$$z_i = \sum_{j=1}^{m} \frac{x_{ij}}{m} \qquad z_i = \sum_{j=1}^{m} \frac{x_{ij}}{\sqrt{np_i(1-p_i)}} \qquad z_i = \sum_{j=1}^{m} S_j p_j^F x_{ij}$$

S = average probabilistic score of each class

$$p^F = 0.3562(p)^{-1.4162}$$



Candidate genes are over-represented among statistically significant associations.

Results for three genes (*FLM*, *SPL5*, and *FY*) with rare variants were consistent.

Genes with rare variants showing associations to flowering time

Gene & Gene ID	Sum test	Weighted sum	Function-aided sum	A priori candidate gene	Connected in AraNet	Supporting evidence
FLM AT1G77080	LD (3.32) JIC2W (4.78) JIC4W (3.23)	JIC2W (1.45)	JIC2W (3.12)	yes	yes	Scortecci *et al.* 2001 Werner *et al.* 2005
BAS1 AT2G26710	LD (4.19) SD (2.91) JIC2W (3.52)	LD (4.55) SD (3.47) JIC2W (3.69)	LD (4.33) SD (3.12) JIC2W (2.23)	yes	yes	Turk *et al.* 2005
SPL5 AT3G15270	JIC/USC (3.22)	JIC/USC (3.47)	JIC/USC (3.57)	yes	yes	Wu *et al.* 2009 Wu and Poethig 2006
FY AT5G13480	JIC2W (3.55) JIC8W (2.24)	JIC2W (4.05) JIC8W (2.31)	JIC2W (3.67)	yes	yes	Simpson *et al.* 2003 [Brachi *et al.* 2010]

Genes with common variants showing associations to flowering time

Gene & Gene ID	Single SNP test	Multiple common	Combined multivariate Pooled	A priori candidate Gene	Connected in AraNet	Supporting evidence [GWAS]
AP1 AT1G69120	JIC0W (2.91) FLC (2.85)	JIC0W (3.17) FLC (3.35) JIC4W (3.37) JIC8W (3.09)	JIC0W (3.28) FLC (3.72) JIC4W (3.48) VERN (3.77)	yes	yes	Gustafson-Brown *et al.* 1994 [Brachi *et al.* 2010] Mouradov *et al.* 2002
CR88 AT2G04030	JIC/USC (1.75)	JIC/USC (2.59)	JIC/USC (2.72)	yes	no	Cao *et al.* 2000
TIC AT3G22380	JIC4W (5.31)	JIC4W (1.87)	JIC4W (3.27)	yes	no	Ding *et al.* 2007
DCL2 AT3G03300	SDV (3.08)	SDV (3.55)	SDV (3.94)	no	yes	Henderson *et al.* 2006
FCA AT4G16280	±V(SD) (4.19)	±V(SD) (3.34)	±V(SD) (3.62)	yes	yes	Macknight *et al.* 1997 [Atwell *et al.* 2010] [Brachi *et al.* 2010] [Zhao *et al.* 2007]
FRI AT4G00650	FRI (14.78) FLC (4.13)	FRI (12.34) FLC (4.77)	FRI (9.43) FLC (3.68) JIC4W (4.23)	yes	yes	Johanson *et al.* 2000 Shindo *et al.* 2005 [Atwell *et al.* 2010] [Zhao *et al.* 2007]
FLC AT5G10140	SD/LD(V) (3.81)	SD/LD(V) (3.14) SDV (3.59)	SD/LD(V) (4.34) SDV (4.81)	yes	yes	Ratcliffe *et al.* 2001 [Atwell *et al.* 2010] [Zhao *et al.* 2007]

Summary

- We presented a CR-GWAS strategy to systematically exploit the collective biological information and analytical tools association analysis
- With the proposed strategy, we confirmed several well-known true positives and identified several new promising associations
- We identified AT3G03300 (*DCL2*) as a new target candidate in regulating flowering time in *Arabidopsis*
- We demonstrated that both common and rare variants contributed to the variation of flowering-time related traits

G3·Genes | Genomes | Genetics Volume 1 | August 2011 | 233

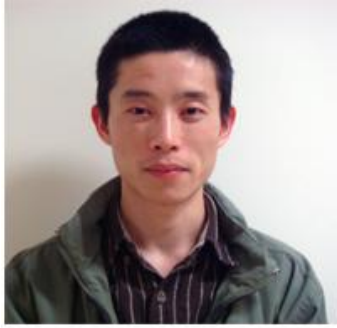
Integrating Rare-Variant Testing, Function Prediction, and Gene Network in Composite Resequencing-Based Genome-Wide Association Studies (CR-GWAS)

Chengsong Zhu, Xianran Li, and Jianming Yu[1]
Department of Agronomy, Kansas State University, Manhattan, Kansas 66506

ABSTRACT High-density array-based genome-wide association studies (GWAS) are complemented by exome sequencing and whole-genome resequencing-based association studies. Here we present a composite resequencing-based genome-wide association study (CR-GWAS) strategy that systematically exploits collective biological information and analytical tools for a robust analysis. We showcased the utility

KEYWORDS
complex trait
dissection
association

II. Genomic Distribution of Trait-Associated SNPs (TAS)

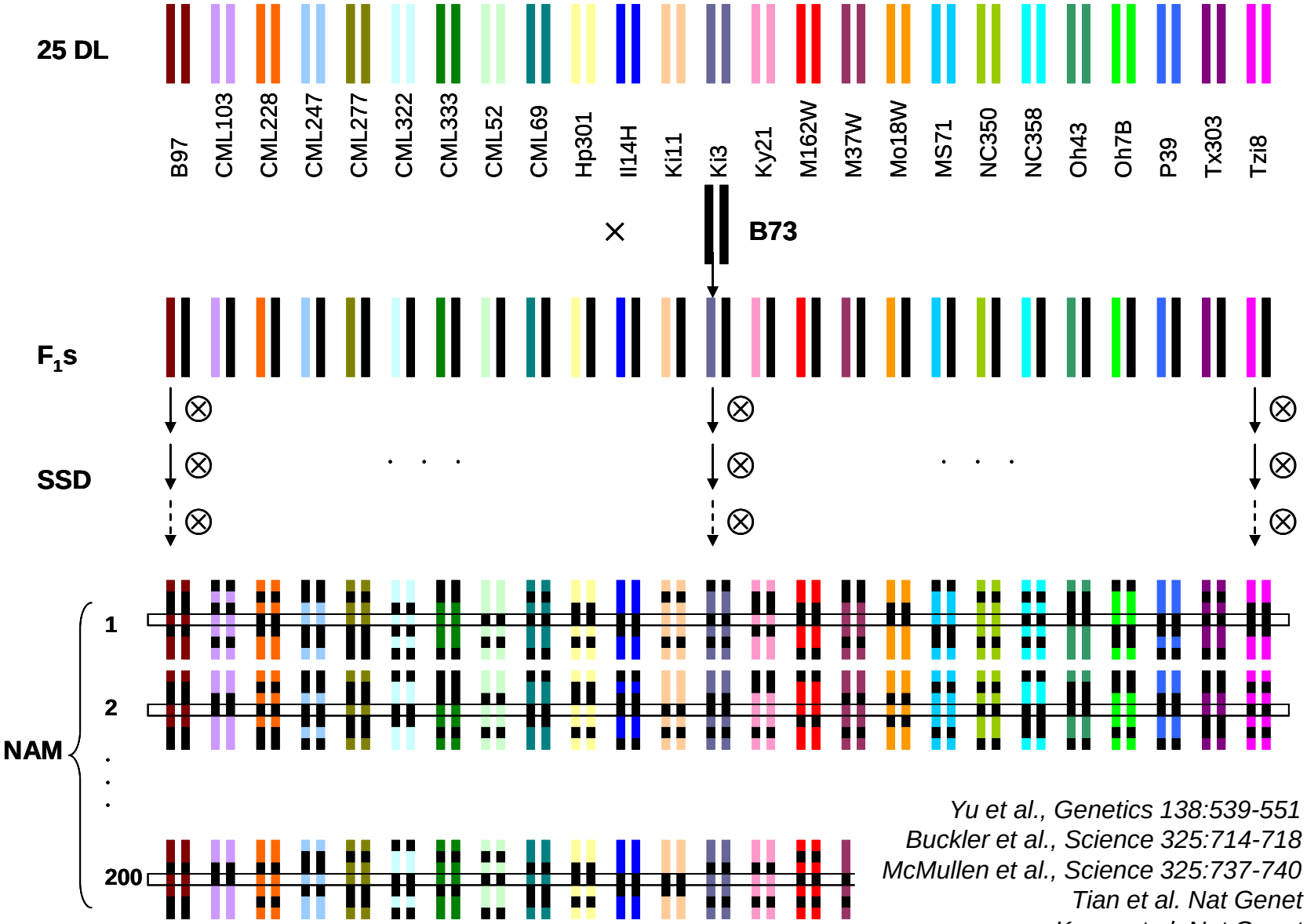


- Genetic Architecture of Complex Traits
  - **How many loci**?
  - What is the frequency of different alleles?
  - **Genomic location of these loci**?
  - **Genetic effects** (*e.g.,* homozygous, heterozygous, epistatic and pleiotropic effects) of these loci?
  - Gene x gene, gene x background, gene x environment interaction?

- Tabulation of previous QTL mapping results, information from QTL cloning experiments

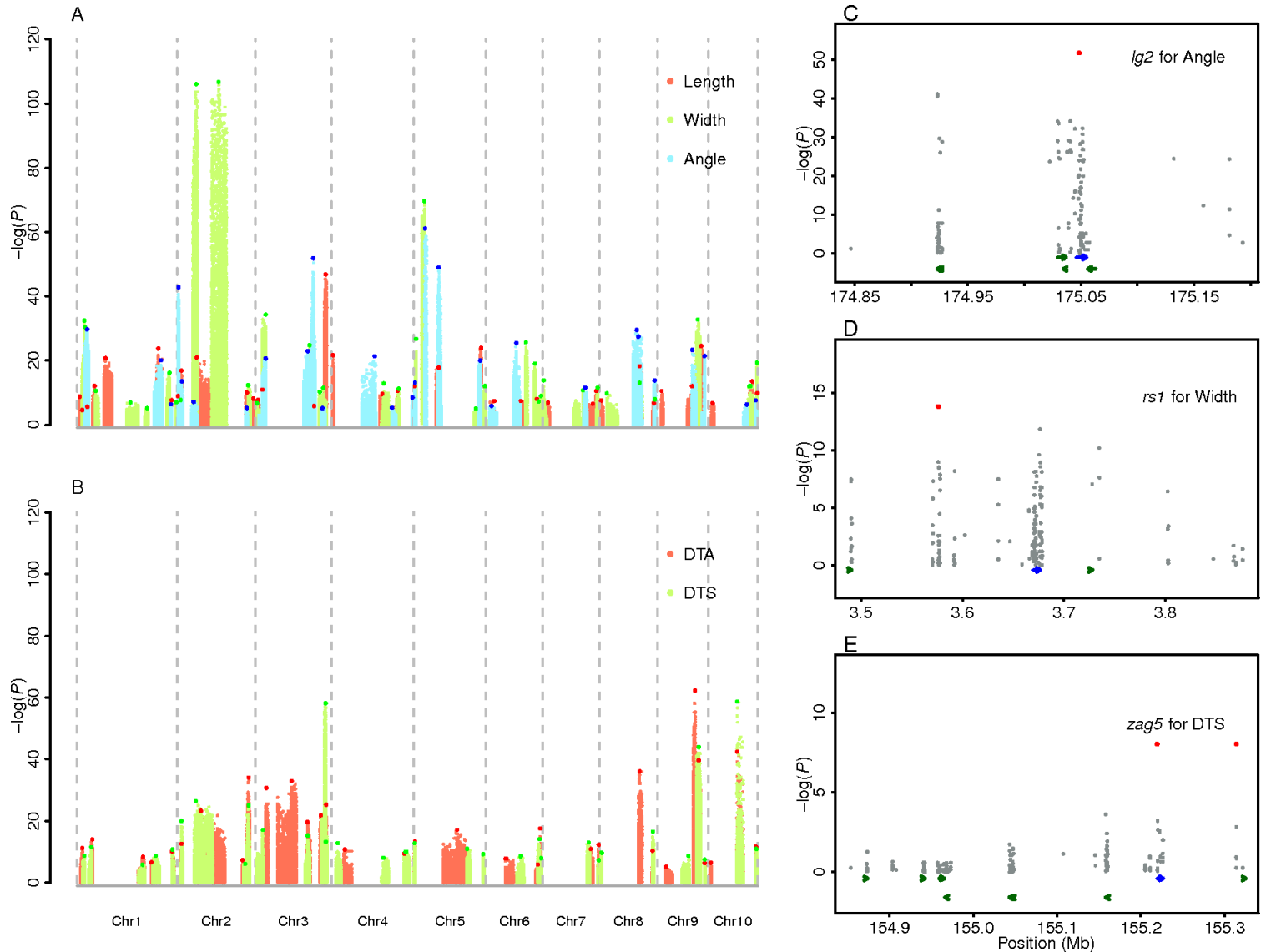
- Tabulation of Trait-Associated SNPs (TASs) from NAM-GWAS
  - Which part of the genome should be prioritized?

An integrated mapping strategy, NAM

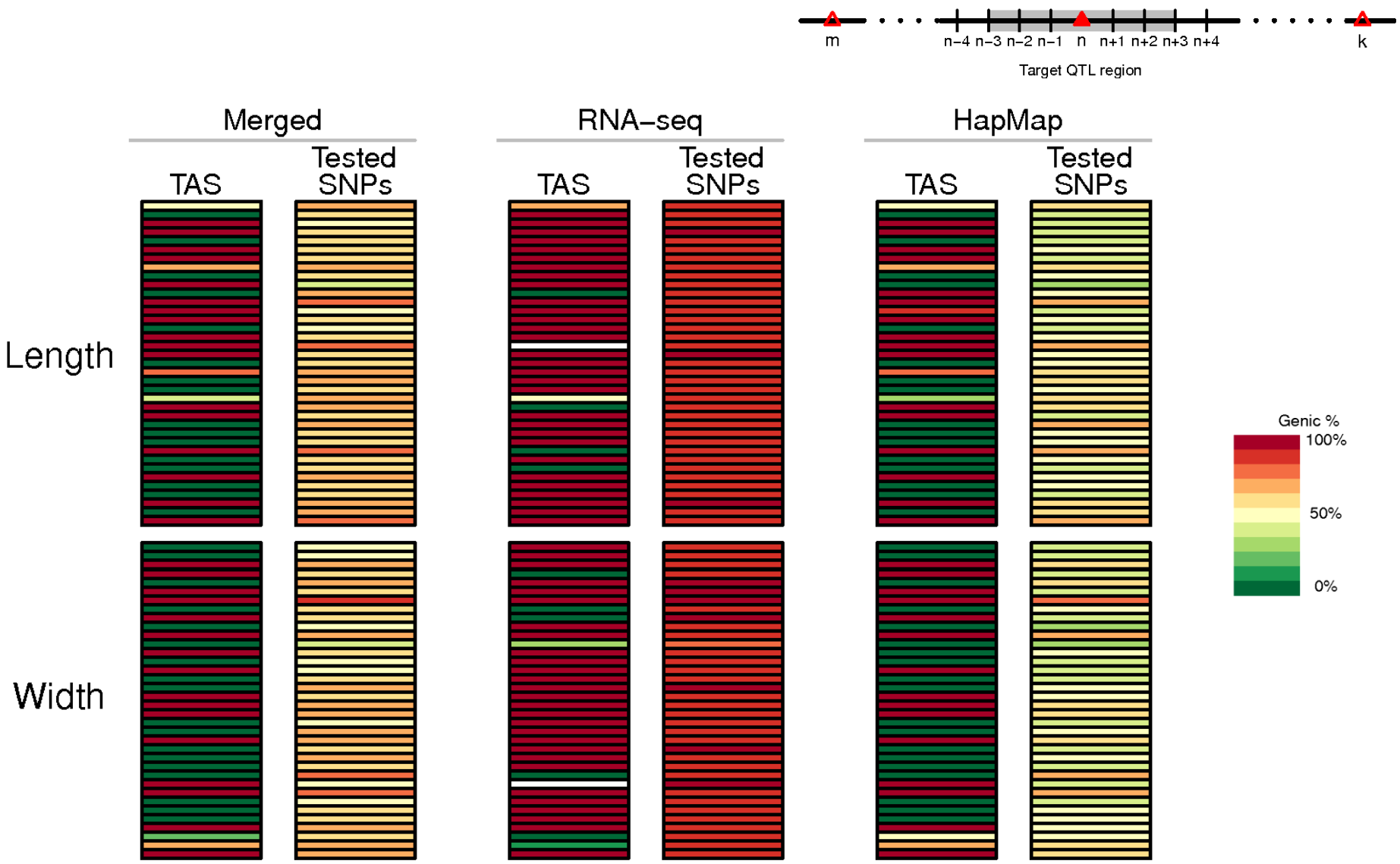


Yu et al., Genetics 138:539-551 (2008)
Buckler et al., Science 325:714-718 (2009)
McMullen et al., Science 325:737-740 (2009)
Tian et al. Nat Genet (2011)
Kump et al. Nat Genet (2011)

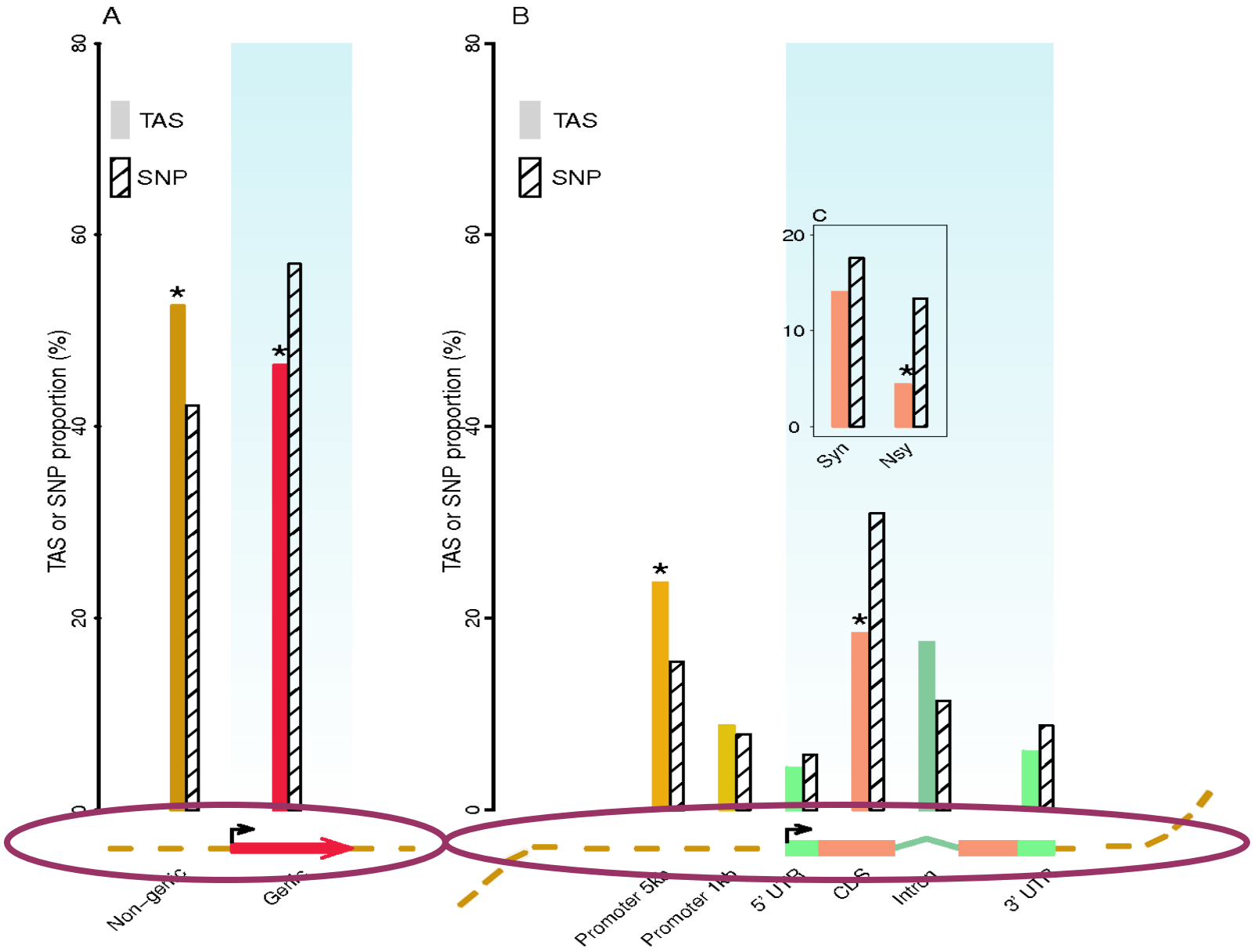
High Resolution of NAM-GWAS



Two-stage, targeted dissection genome scan



Genic + Promoter 5kb = 13% of maize genome, account for 71% of TASs



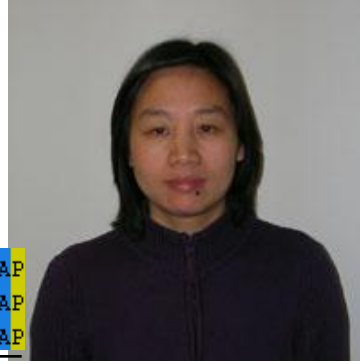
Summary

- Genic and non-genic TASs contribute approximately equally to phenotypic variation for maize quantitative traits

- Genic+promoter region, while comprises only 13% of maize genome, account for 71% of identified TASs and explain 79% of PVE of all TASs
  - *Evolutionary alterations in protein sequence appear to be quantitatively less important than changes in gene regulation in shaping the wide natural variation observed in maize.*

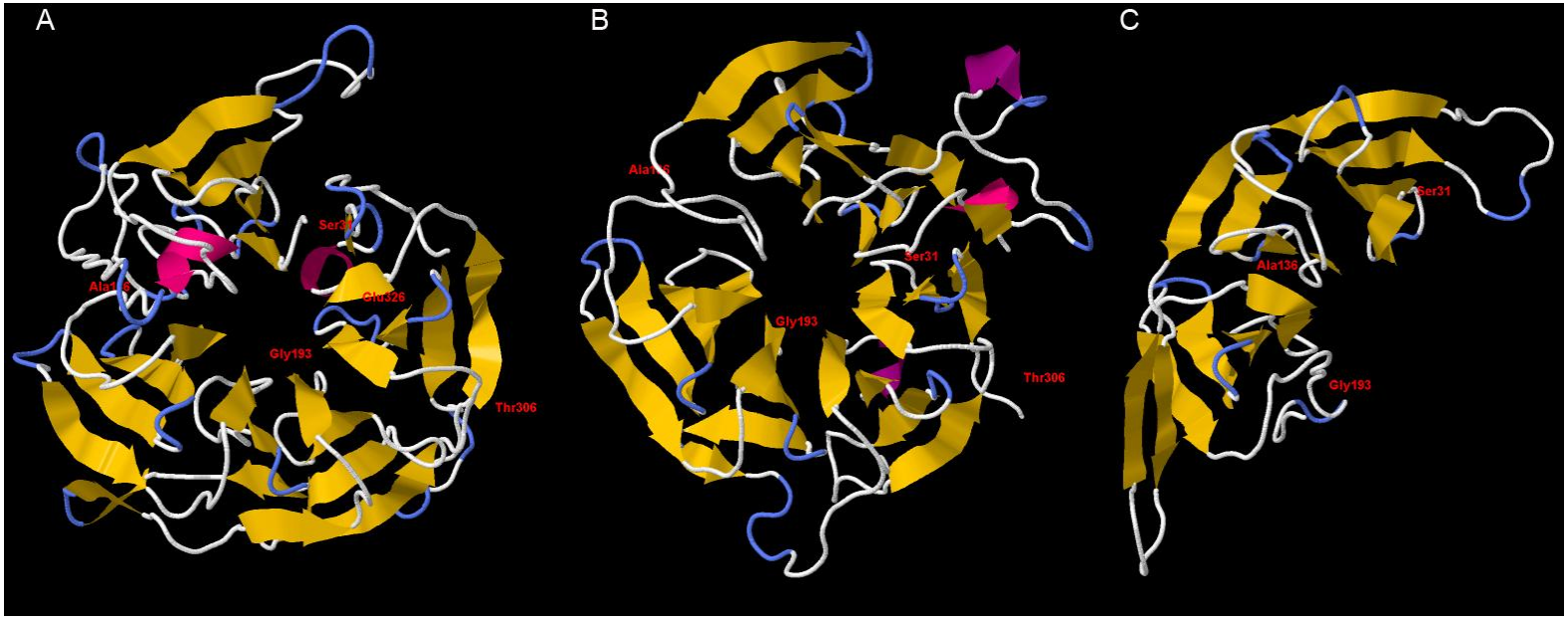
- Genotyping methods designed to discover SNPs in genes and their upstream regions would be the most economical approach for detecting genome-wide association signals
  - *The combination of RNA-seq and exome capture experiments using a long read (e.g., 454) and paired end (Illumina and 454) technologies*

III. Multiple functional haplotypes at a single locus

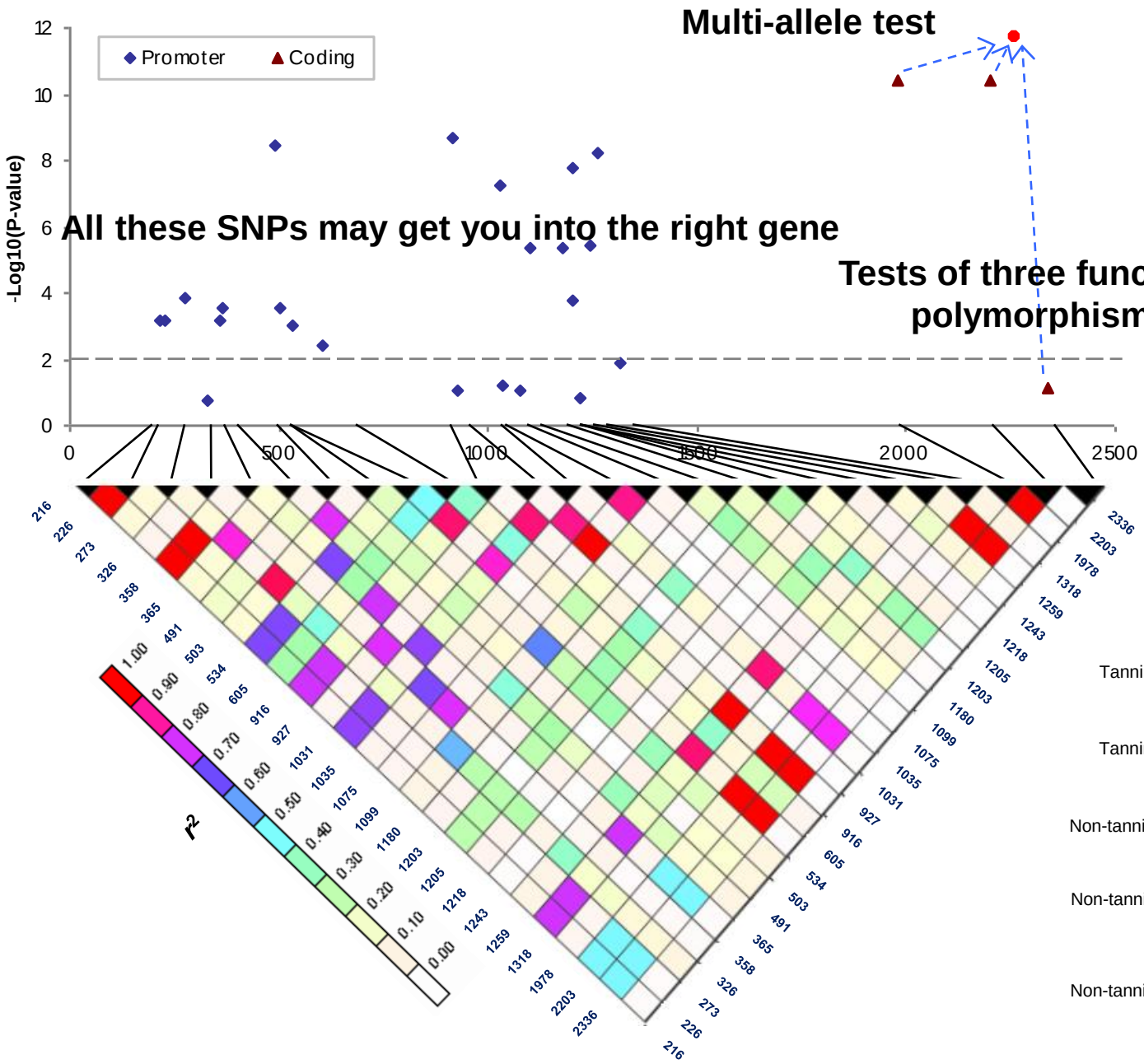


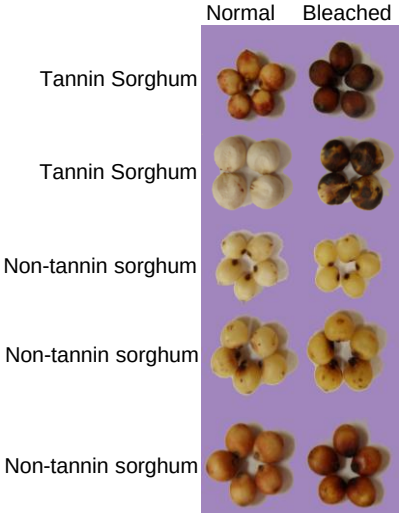
Tannin1 Gene Cloning, Wu *et al., 2012 PNAS (online first)*

```
Tan1     MDLPKPPSTAASSSGAETPNPHAFTCELPHSIYALAFSPGAPVLASGSFLEDLHNRVSLLSFDPVRPSAASFRALPALSFDHPYPPTKLQFNPRAAAP
tan1-a   MDLPKPPSTAASSSGAETPNPHAFTCELPHSIYALAFSPGAPVLASGSFLEDLHNRVSLLSFDPVRPSAASFRALPALSFDHPYPPTKLQFNPRAAAP
tan1-b   MDLPKPPSTAASSSGAETPNPHAFTCELPHSIYALAFSPGAPVLASGSFLEDLHNRVSLLSFDPVRPSAASFRALPALSFDHPYPPTKLQFNPRAAAP

Tan1     SLLASSADTLRIWHAPLDDLSATASAPELRSVLDNRKAASEFCAPLTSFDWNEVEPRRIGTASIDTTCTVWDIDLGVVETQLIAHDKAVHDIAWGEA
tan1-a   SLLASSADTLRIWHAPLDDLSATASAPELRSVLDNRKAASEFCAPLTSFDWNEVEPRRIGTASIDTTCTVWDIDLGVVETQLIAHDKAVHDIAWGRP
tan1-b   SLLASSADTLRIWHAPLDDLSATASAPELRSVLDNRKAASEFCAPLTSFDWNEVEPRRIGTASIDTTCTVWDIDLGVVETQLIAHDKAVHDIAWGEA

Tan1     VFASVSADGSVRVFDLRDKEHSTIVYESHRPDTPLLRLAWNRSDLRYMAALLMDSSAVVVLDIRAPGVPVAELHRHRACANAVAWAPQATRHLCSAGD
tan1-a   SSPPCRPTAPSASSTSGTRNTPPSSTRAPAPTRRSSGWRGTALTSAIWPRCSWTAAPSSCSTYVRPGCRLPSCTGTGRAPTQSRGRRRPLGTSARLGT
tan1-b   VFASVSADGSVRVFDLRDKEHSTIVYESHRPDTPLLRLAWNRSDLRYMAALLMDSSAVVVLDIRAPGVPVAELHRHRACANAVAWAPQATRHLCSAGD

Tan1     DGQALIWELPETAAAVPAEGIDPVLVYDAGAEINQLQWAAAHPDWMAIAFENKVQLLRV   353 aa
tan1-a   TGRH                                                          298 aa
tan1-b   DGQALIWELPETAAGSGGCARRGD                                      318 aa
```



Tannin1 Gene Cloning, Wu *et al.*, 2012 *PNAS (online first)*

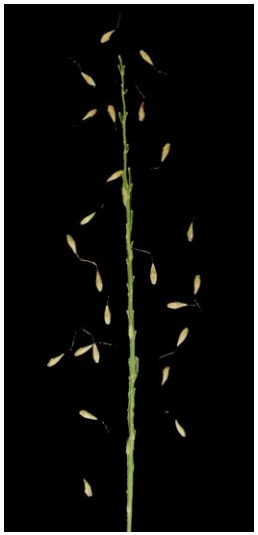




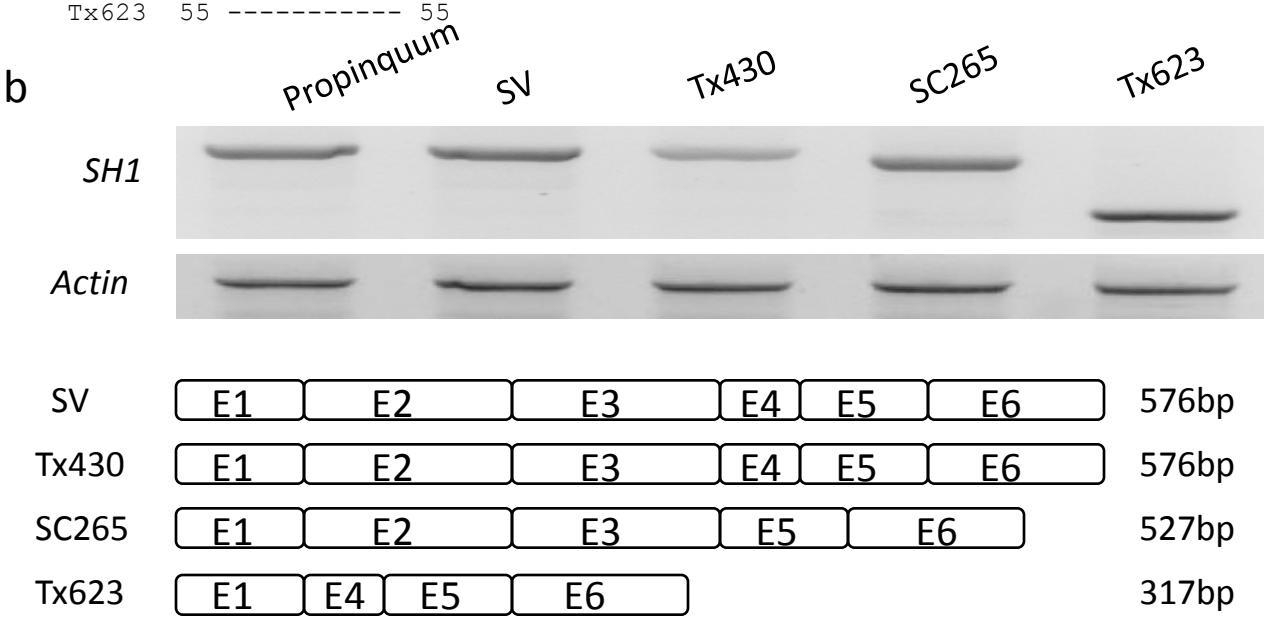
Shattering1 Gene Cloning, Lin *et al.*, 2012 *Nature Genetics 44:720–724*

a

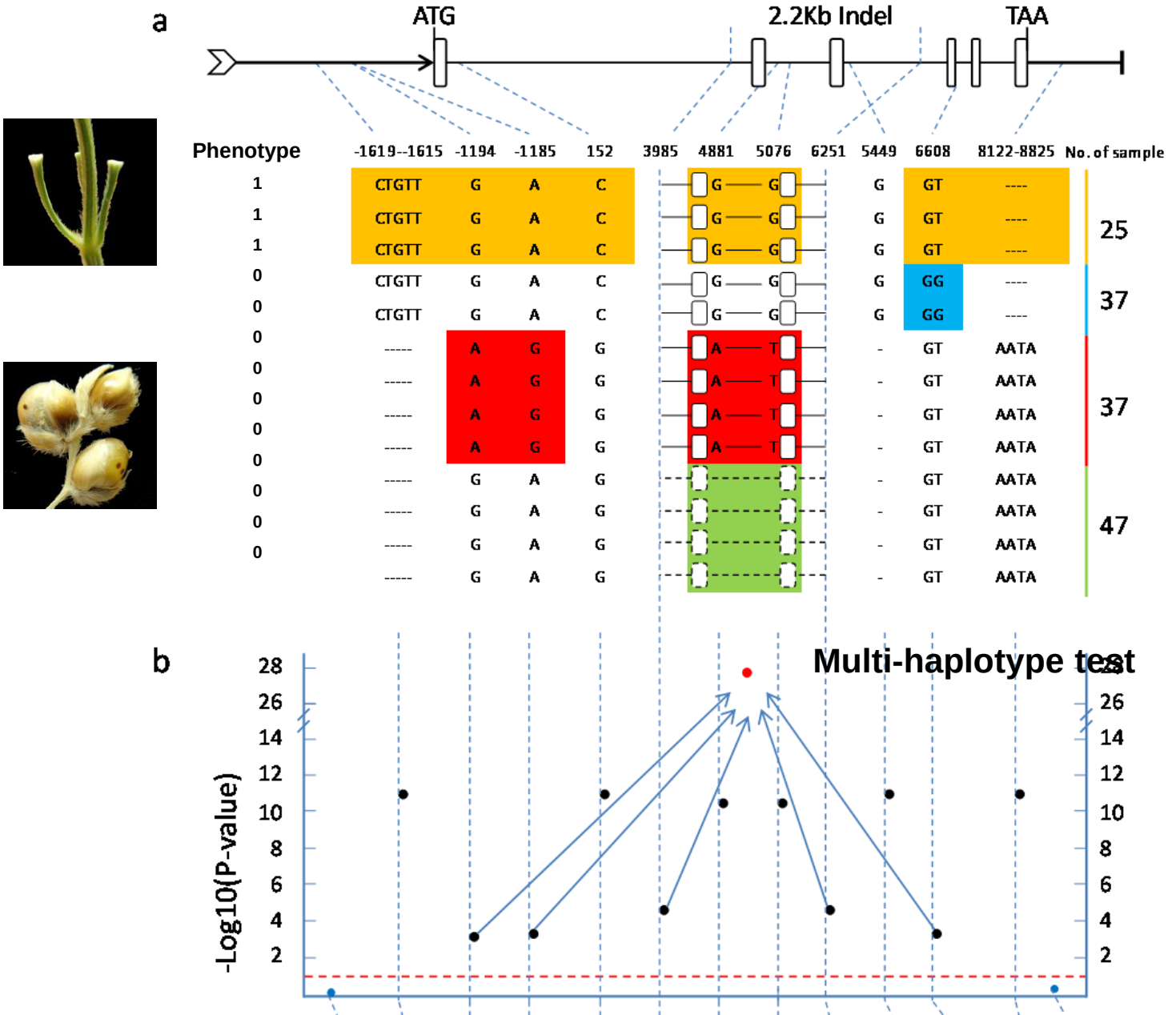
```
SV      1 MSAQQIAPVPEHVCYVHCNFCNTILAVSVPSHSMLNIVTVRCGHCTSLLSVNLRGLLQSL 60
Tx430   1 MSAQQIAPVPEHVCYVHCNFCNTILAVSVPSHSMLNIVTVRCGHCTSLLSVNLRGLLQSL 60
SC265   1 MSAQQIAPVPEHVCYVHCNFCNTILAVSVPSHSMLNIVTVRCGHCTSLLSVNLRGLLQSL 60
Tx623   1 MSAQQIAPVPEHVCYVHCNFCNTILALQRRGNVFLQHITDLLRKRYEGLKQATQT----- 55
          ************************.    .  *   *

SV     61 PVQNHYSQENNFKVQNFSFTENYPEYAPSSSKYRMPTMLSAKGDLDHMLHVRAPEKRQRV 120
Tx430  61 PVQNHYSQENNFKVQNFSFTENYPEYAPSSSKYRMPTMLSAKGDLDHMLHVRAPEKRQRV 120
SC265  61 PVQNHYSQENNFKVQNFSFTENYPEYAPSSSKYRMPTMLSAKGDLDHMLHVRGKRYEGLK 120
Tx623  55 ------------------------------------------------------------ 55

Sv    121 PSAYNRFIKEEIRRIKASNPDISHREAFSTAAKNWAHFPNIHFGLGPYESSNKLDEAIGA 180
Tx430 121 PSAYNRFIKEEIRRIKASNPDISHREAFSTAAKNWAHFPNIHFGLGPYESSNKLDEAIGA 180
SC265 121 QATQT------------------------------------------------------- 125
Tx623  55 ------------------------------------------------------------ 55

Sv    181 TGHPQKVQDLY 191
Tx430 181 TGHPQKVQDLY 191
SC265 125 ----------- 125
Tx623  55 ----------- 55
```



b



Shattering1 Gene Cloning, Lin *et al., 2012 Nature Genetics 44:720–724*



Overall Summary

- Traits are complex! Going genome-wide presents opportunities and challenges
- Many "**biological**" and "**statistical**" tools are available and it sure helps to use a combination of them
  - Geno→Pheno & Geno↔Pheno
- QTL → → TAS → → functional polymorphisms → → "**haplotypes**"
  - GWAS and GS
- Genic or non-genic region harbors functional polymorphisms?
  - Different functional polymorphisms/haplotypes for different allellic series, complex *versus* less complex traits
- How do we validate the findings of TAS/QTN through GWAS, with small-moderate effects, for complex traits?
  - ZFN, TALEs?

Communication

Statisticians	Biologists
These are the significant variables. All we can do so far.	Why my favorite genes are not tagged? No other choices?
These are the cut-off threshold that was used.	What if we changed that a bit? Will it make my genes significant?
These are the ones that are worth follow-up studies.	I only have one postdoc/student to chase after the strongest one.
We can look into other newly developed methods.	Well, I will have “HapMap X” data ready next week.
Statistical methods (generalized, less assumptions, higher power) Experimental design	Data collection (x, y, expression, protein, independent experiment, trangenic complementation, mutants) Biological questions

Acknowledgements

Mike Scanlon (Cornell)
Patrick Schnable (ISU)
Gary Muehlbauer (Cornell)
Marja Timmermans (CSHL)

Tesfaye Tesso (KSU)
Donghai Wang (KSU)
Scott Staggenborg (KSU)
Vara Prasad (KSU)
Ramasamy Perumal (KSU)
Scott Bean (USDA-ARS)

Lean Miller
Kallen Kershner
Chris Fontenot
Adeyanju Adedayo
Raymond Mutava

Guihua Bai (USDA-ARS)
MingLi Wang (USDA-ARS)
Gary Pederson (USDA-ARS)
Tom Clemente (UNL)
Harold Trick (KSU)

John Doebley (Wisconsin)
Ed Buckler (USDA-ARS)
James Holland (USDA-ARS)
Steve Kresovich (Cornell)
Mike McMullen (USD-ARS)
Zhiwu Zhang (Cornell)

Rex Bernardo (U of Minn.)
Ismail Dweikat (UNL)
Yiwei Jiang (Purdue)
Jianxin Ma (Purdue)

Min Zhang (Purdue)
Dabao Zhang (Purdue)
Weixing Song (KSU)
Hans-Peter Piepho (Hohenheim)

Frank White (KSU)
Clare Nelson (KSU)
Mitch Tuinstra (Purdue)

Targeted Excellence
K-STATE
Kansas State University

USDA
United States Department of Agriculture
National Institute of Food and Agriculture

Center for SORGHUM Improvement

Kansas GRAIN SORGHUM Commission

NSF

SAM
MAIZE SHOOT APICAL MERISTEM