

Association mapping in barley: two rows vs. six rows

Jerko Gunjača

Faculty of Agriculture,
University of Zagreb

Jean-Luc Jannink

USDA – ARS /

Cornell University, Ithaca



BARLEY



CAP

What is BarleyCAP?

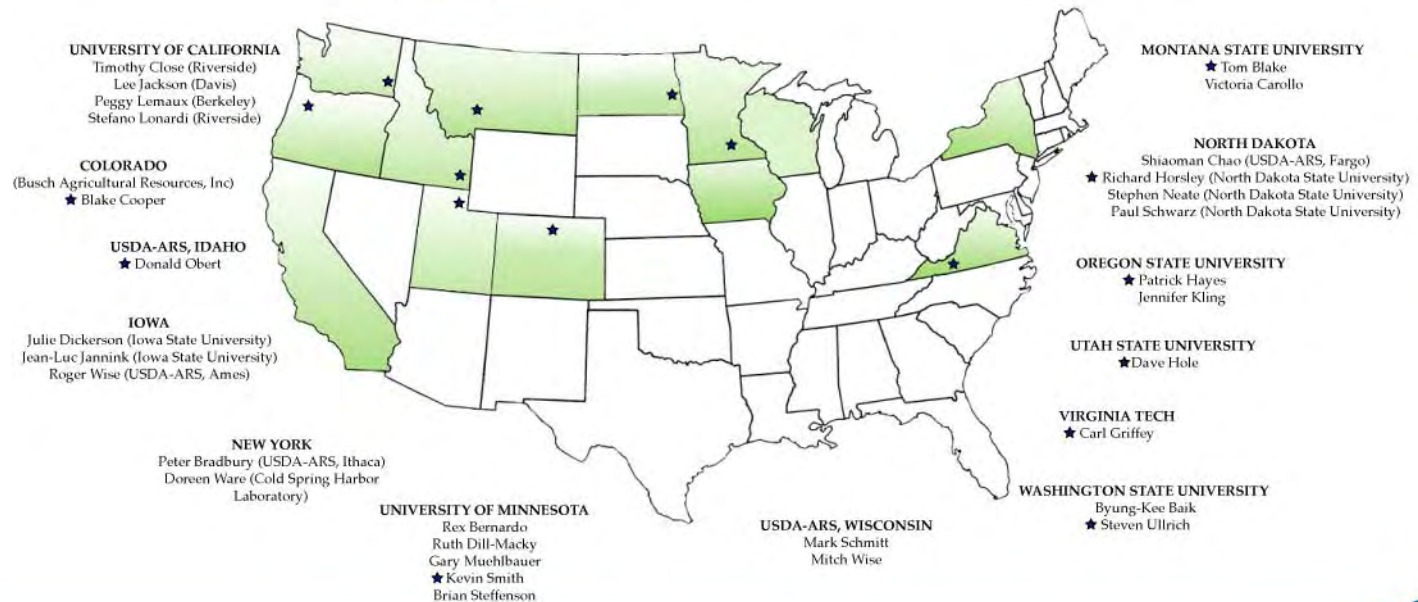
Coordinated Agricultural Project for Barley

This USDA/CSREES-funded project involves government, public and private laboratories, links laboratory and field research with education and outreach in an effort to use modern tools of genomics to facilitate classical plant breeding efforts to develop superior barley varieties.

Lead PI/Institution:

Gary Muehlbauer, University of Minnesota

★ denotes BarleyCAP breeders



Association mapping

- Fine mapping based on LD
- Search for markers associated (linked) to QTLs
- However, there are other mechanisms causing LD, thus generating false positives or false discoveries (FD)

Population structure

- Most important cause of LD besides linkage in populations which are not homogeneous
- In order to reduce FD rate, models for association mapping should include adjustment for population structure

Q + K adjustment

- Q – structure:
 - STRUCTURE (Pritchard et al. 2000)
 - PCA (Price et al. 2006)
- K – kinship:
 - Loiselle et al. 1995
 - Ritland 1996

Basic model

- Yu et al. (2006):

$$Y_{ip} = \mu + a_p + \sum_{u=1}^z Q_{iu} + g_i^* + e_{ip}$$

$$\text{Var}(g^*) = 2K\sigma_{g^*}^2$$

$$\text{Var}(e) = R\sigma_r^2$$

MET extension

- Stich et al. (2008):

$$Y_{ijknop} = \mu + a_p + \sum_{u=1}^z Q_{iu} + g_i^* + l_j + (al)_{pj} + \\ + (gl)_{ij} + t_{jk} + r_{jkn} + b_{jkno} + e_{ijknop}$$

$$Var(g^*) = 2K\sigma_{g^*}^2$$

$$Var(e) = R\sigma_r^2$$

Residual FD

- Zhao et al. (2007) – residual confounding after applying Q+K
- Adjustment for multiple testing?
 - FWER (Bonferroni, etc.)
 - FDR (Benjamini & Hochberg 1995, Benjamini & Yekutieli 2001, etc.)

Goal(s)

- Use field trial data and marker information for two row barleys only
- Perform association mapping and search for:
 - QTLs
 - QTL x E interactions

Available data

- Two row barley breeders' lines from: Washington, Idaho, Montana, North Dakota + checks (varieties)
- Field trials carried out on 1-6 sites per breeding program, 2006-2008
- Genotyped for ~ 3000 SNPs
- Weather data from public service

Field trials – set up

- Each breeding program used its own set of lines and sites:
 - Washington: 96 lines, 1 site
 - Idaho: 45 lines, 3 sites
 - Montana: 96 lines, 2(+1) sites
 - North Dakota: 96 lines, 6 sites
- Common checks!

Field trials – problems

- Means and replicate data available for all BP's except Idaho (means only)
- Scored for several traits, yield is the only one consistent over sites
- Complete data available only for 2006
- Different set of lines used each year
- Six row checks?

Analysis: stage one (FT)

- Analyze each trial within each site separately using the appropriate model matching the design
- Obtain estimates and standard errors for use in weighted analysis in stage two (Idaho: use residual variance to calculate standard errors for means)

Analysis: stage one (Q+K)

- Obtain Q using PCA rather than STRUCTURE
- Obtain K using Loiselle
 - Set negatives to zero
 - Alternative: Stich et al. (2008), following Bernardo (1993) proposed $K = \text{similarity}$ adjusted to average similarity between lines and unrelated varieties

Analysis: stage two (fit)

- Fit general model (no marker effects)
- Fit adapted Stich et al. (2008) model:

$$Y_{ijkp} = \mu + a_p + \sum_{u=1}^z Q_{iu} + g_i^* + l_j + (al)_{pj} + \\ + (gl)_{ij} + t_{jk} + e_{ijkp}$$

- Drop $(al)_{pj}$, fit thereby reduced model

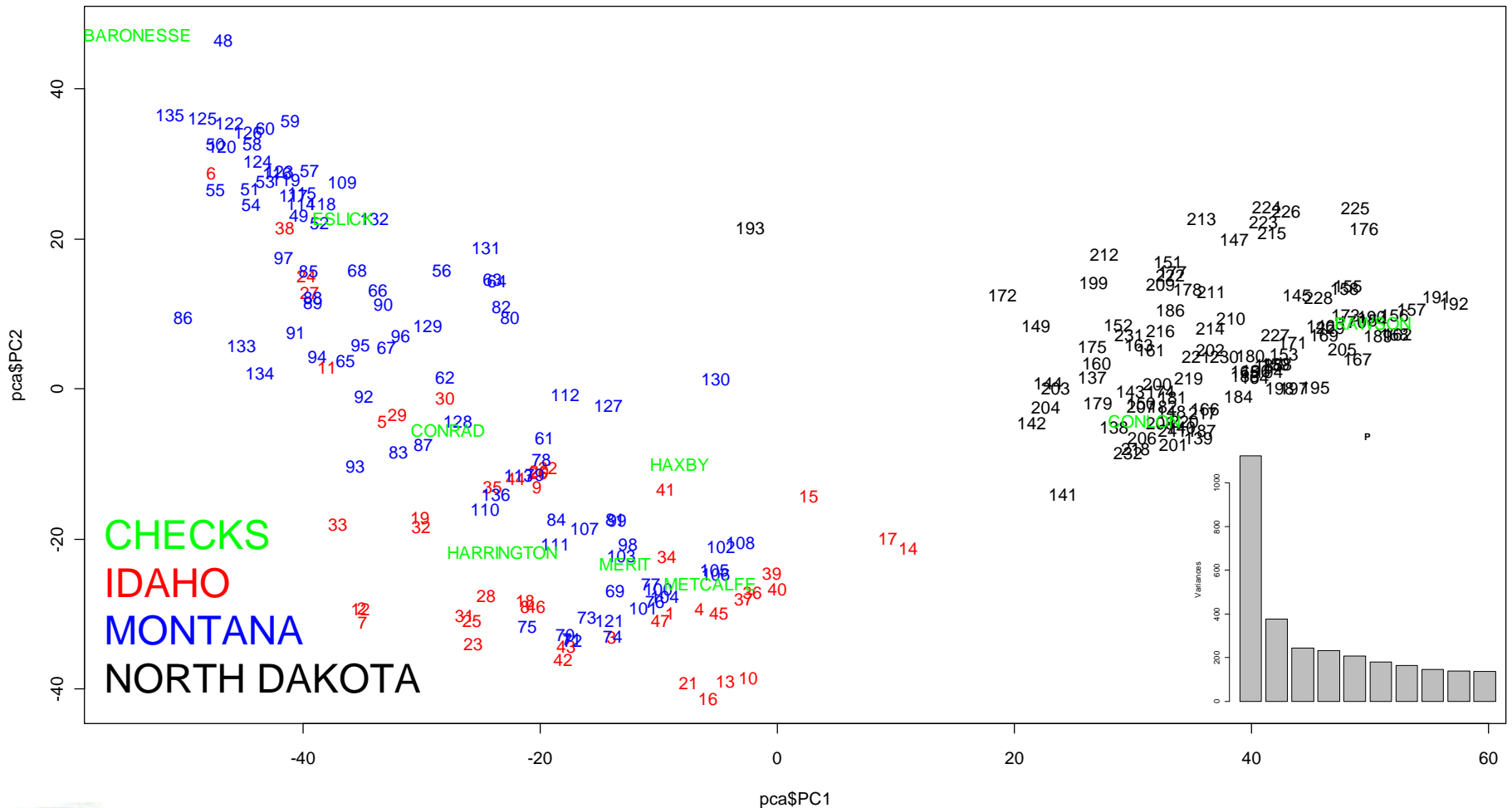
Analysis: stage two (test)

- Test for a_p : Wald F
- Test for $(aI)_{pj}$: LRT using full and reduced model
- Adjustment for multiple testing:
 - Benjamini and Yekutieli (2005)

Software

- Modelling: ASReml 3 alpha
- K estimates: SPAGeDi
- Everything else: R

Results: stage one (Q)



Results: stage one (K)

- Several incidences of $K > 1$ observed?
- Probable cause: rare alleles
- All SNPs with minor allele freq. $< 5\%$ removed, K's decreased but still > 1
- Average sim. with six row checks ~ 0.56
- Fine tuning around this value

Input for stage two

- Removed:
 - SNPs with minor allele frequency $< 5\%$
 - 6 identical lines
 - Whole set from Washington
 - Six row checks
- Remainder:
 - 242 genotypes, 12 sites, 1797 SNPs

Results: stage two (general)

- PCAs 2-3 significant, PCA1 not (but only by a margin)
- Best fit (highest likelihood) with K obtained by adjusting similarity to hypothetical average of 0.62
- K matrix not positive definite – but it works using appropriate qualifier in ASReml

Results: stage two (a_p & $(aI)_{pj}$)

- 37 significant SNPs
- Or 80 – from reduced model?
- However, there is nothing left when BY adjustment was applied ($q = 0.1$) in both cases
- 263 significant SNP x E
- After adjustment – 27 (with $q=0.1$)

Consequences (QTL)

- No yield QTLs to be detected?
 - Sample size?
- Check the map positions (if available) for the eventual grouping of significant SNPs (before adjustment)
 - No clear evidence of grouping?

Consequences (QTL x E)

- Too many significant tests for SNP x E compared to SNP (before BY adjustment)
- Population structure adjustment not applied when testing for SNP x E?
- Modeling of VCOV structure for G x E effect required?

Acknowledgements

- Fulbright program
- BarleyCAP – everybody at all labs
- ASReml code:
 - Randy Wisser
 - Arthur Gilmour
- Jannink's lab:
 - Tom Parker
 - Hiroyoshi Iwata
 - Peter Bradbury