A Catalog of Genome Content Variation in *Arabidopsis thaliana*

Christopher J. Fiscus & Daniel Koenig

@fiscuscj

koenig lab

UC RIVERSIDE UNIVERSITY OF CALIFORNIA

Background

- Genome content varies both between and within species
- Genome content differences are driven by both genetics & the environment

Motivating questions

1. What factors affect the rate of change in genome content?
2. Which sequences exhibit copy number variation?
3. What is the genetic basis of copy number variation?

Methods

Assay genome content

WGS reads from 1142 genomes [1,2,3]

Distance to ancestor ← Count K-mers → PCA

Estimate copy number of annotated sequences

Contribution of environment

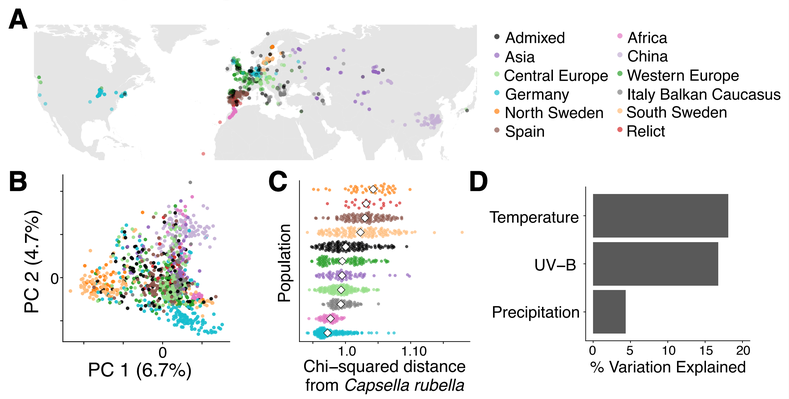
Calculate correlations between environmental variables and distance from ancestor

Map genetic basis

Treat copy number as molecular phenotype for GWAS

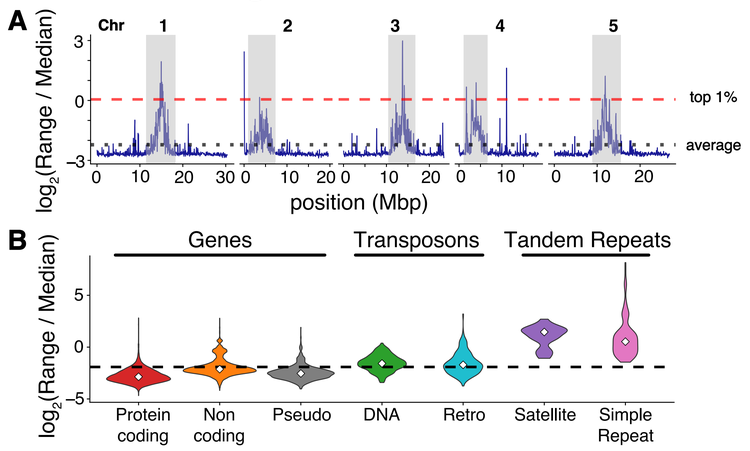
Meta-analysis of GWAS studies

Population structure and climate partially explain genome content differences



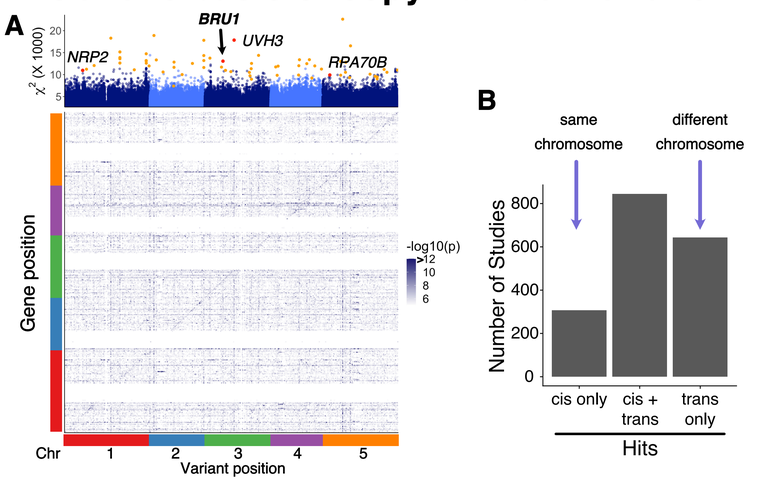
A) Collection sites of accessions. Colors indicate population groups.
B) Principal component analysis of K-mer profiles.
C) Chi-squared distance from *Capsella rubella*. Diamonds are medians.
D) Variation in distance from ancestor explained by environmental variables.

Patterns of genome content variation



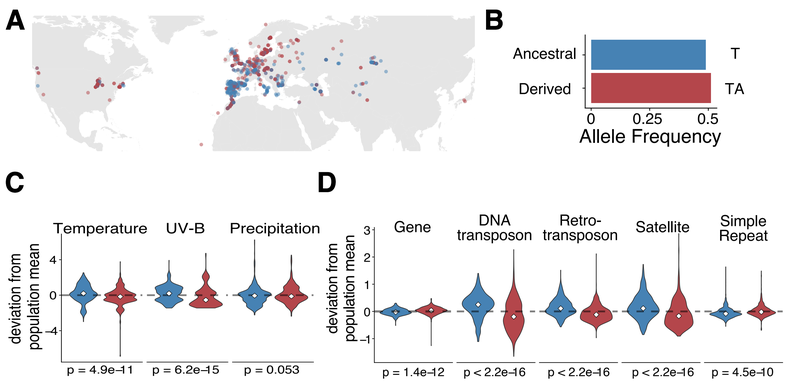
A) Genome content variation in non-overlapping 50 KB windows across the TAIR10 reference genome. Shaded regions indicate centromeres.
B) Sequence copy number variability by class. Dotted black line indicates threshold to call copy number variants. Diamonds are medians.

Genetic basis of copy number variation



A) Meta-analysis of 2868 GWAS for gene copy number. Highlighted variants were found in all meta-analyses across sequence classes.
B) Gene GWAS studies partitioned by location of significant variants.

BRU1 (TSK) affects recombination rates and putatively regulates genome stability [4]



Global distribution (A) and frequency (B) of *BRU1* alleles.

BRU1 is associated with environment (C), sequence copy number (D), and the distance to ancestor (not shown). Dotted black line indicates population means and diamonds indicate group medians. All p values are from Wilcox-signed-rank tests.

Citations

1. 1001 Genome Consortium. *Cell* **166**, 481-491 (2016).
2. Durvasula, A *et al.* PNAS. **114**, 5213-5218 (2017).
3. Zou, Y. P. *et al. Genome Biol.* **18**, 239 (2017).
4. Ohno, Y *et al. Genetics.* **189**, 83–95 (2011).

Acknowledgements

We thank Keely E. Brown, Jacob B. Landis, and Danelle K. Seymour for their helpful comments on this work. Computations were performed using the computer clusters and data storage resources of the High-Performance Computing Center at UC Riverside, which were funded by grants from NSF (MRI-1429826) and NIH (1S10OD016290-01A1).