Genetic Mechanisms of Genome Changes

Jinyu Wang[1], Xianran Li[1], Yanhai Yin[2] and Jianming Yu[1]

[1]Department of Agronomy, [2]Department of Genetics, Development, and Cell Biology, Iowa State University, Ames, IA

Abstract

Progress of sequencing technologies has generated a huge amount of genomic sequences, which provide a great opportunity to address fundamental questions about the genome itself. By investigating the Base Composition across SNPs (BCS) within species, our research group has discovered a conserved pattern of BCS among individuals, *i.e.*, modern crops all have significantly higher [AT] value than their ancestors. This consistent pattern across multiple species implies BCS is an ideal indicator of genome changes accumulated through evolution. We hypothesize that the genome divergence is under genetic control.

Arabidopsis is a feasible system to test the hypothesized genetic mechanism. By analyzing the 297 *Arabidopsis* genomes, we identified one locus at chromosome 5 strongly associated with the BCS variation through Genome Wide Association Studies (GWAS). To pinpoint the underlying gene, a set of Near-Isogenic Lines (NILs) are used for functional assays. These NILs were developed from two parents harboring different alleles at the GWAS signals. Research discoveries from this genetic mechanism will broadly enrich our understanding of genomics, genetics, and evolution.

Objective

Unravel genetic mechanisms of *Arabidopsis'* genome changes

- Develop a measurement to capture the systematic difference of DNA polymorphisms.
- Conduct GWAS to identify loci associated with the genome divergence.
- Identify the functional gene underlying the GWAS signal in defined genetic background.
- Construct co-expression network for unravelling pathways contributing to genome divergence.

Base Composition across SNPs (BCS)

Acc	SNP matrix (6 × 16)	Base composition across SNPs [A]	[C]	[G]	[T]
1	AGACTCTCGCCGGGGC	2/16=0.13	6/16=0.38	6/16	2/16
2	AGACTCTAGCCGGGGC	3/16=0.19	5/16=0.31	6/16	2/16
3	CGACTCTAGCCGAGGC	3/16=0.19	6/16=0.38	5/16	2/16
4	AGACTATAGACTGTGC	5/16=0.31	3/16=0.19	4/16	4/16
5	AGTTCACAGAATGTGA	6/16=0.38	2/16=0.13	4/16	4/16
6	AAATCATATAAGGTTC	7/16=0.44	2/16=0.13	2/16	5/16

Table 1. An example of how to calculate BCS for 6 accessions across 16 SNP sites. Each accessions has 4 values corresponding to 4 different types of nucleotides.

BCS captures the overall genome divergence

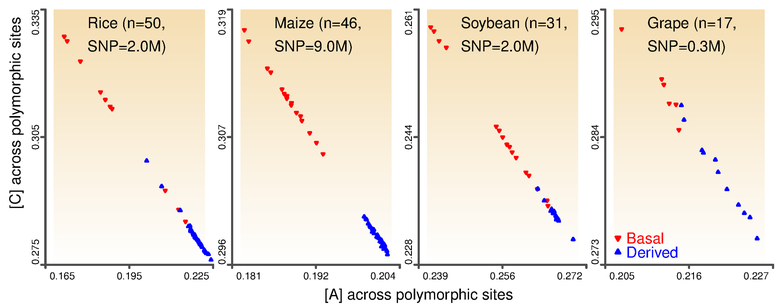

Fig. 1. Domesticated crops have different BCS values (higher [A]) than their wild relatives. Individual strand base equality rule (Parity Rule 2) also holds in SNP sites: [A] ≈ [T], [C] ≈ [G], thus, [A] + [C] ≈ 0.5, as all dots are aligned along the off-diagonal line. Composition of other three nucleotides can be fairly inferred from one based on this equality rule. **Red dots denote wild accessions, blue for domesticated accessions. [A] is on X-axis and [C] is on Y-axis.**

Conduct GWAS for BCS variation in *Arabidopsis*

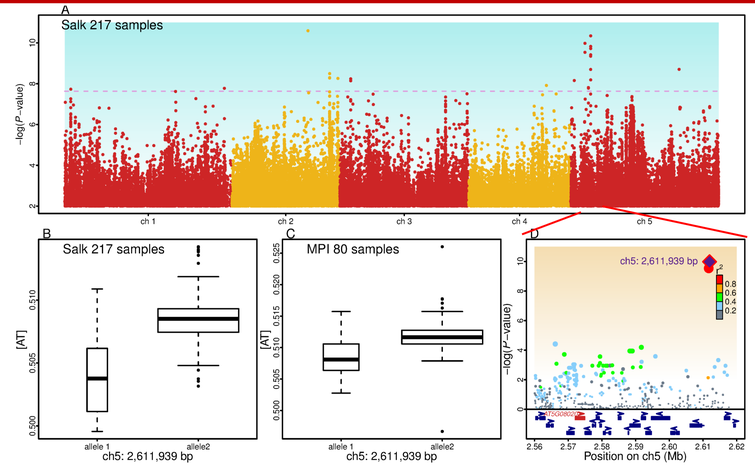

Fig. 2. One locus at chromosome 5 shows strong association with variation of BCS.

A. Manhattan plot of GWAS result with BCS using 217 Salk accessions.
B. Accessions carrying allele 1 (Cvi allele) at the association site have significantly less nucleotide A and T across the SNPs in the genome than those with allele 2 (L*er* allele).
C. The association site detected by the 217 Salk samples was further verified by an independent set (80 MPI samples).
D. Regional plot around the strongly association site at chromosome 5.

Connection between UV and *Arabidopsis* SNPs

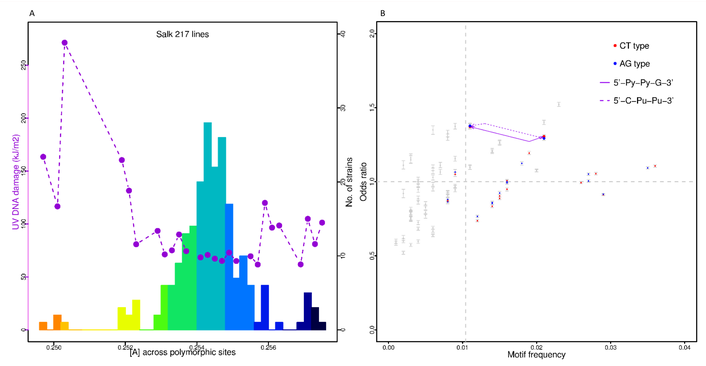

Fig. 3. Relationship between ultraviolet (UV) radiation and *Arabidopsis* SNPs.

A. Accessions with lower [A] are generally originated from habitats receiving higher UV radiation (Habitat of Cvi receives much higher UV radiation than the one of L*er*). Histogram shows the distribution of BCS. Dashed line indicated the average UV dosage for the habitats.
B. The enrichment test of motifs around SNPs suggested that *Arabidopsis* SNPs more likely arose from the potential solar UV signature. Solar UV induced mutations tend to occur at 5-methylcytosine (5-$_m$C)-containing dipyrimidine site (5′-P$_y$-$_m$C$_p$G-3′).

Identify the functional gene underlying the GWAS signal in defined genetic background

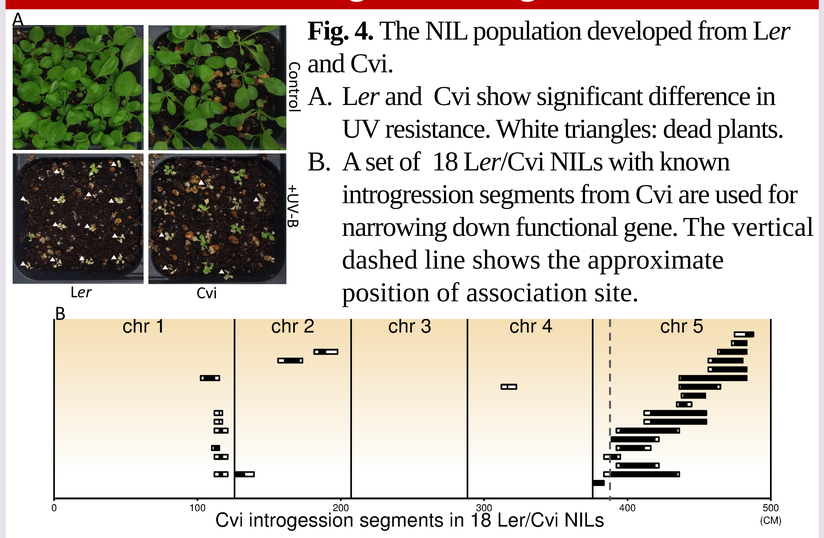

Fig. 4. The NIL population developed from L*er* and Cvi.

A. L*er* and Cvi show significant difference in UV resistance. White triangles: dead plants.
B. A set of 18 L*er*/Cvi NILs with known introgression segments from Cvi are used for narrowing down functional gene. The vertical dashed line shows the approximate position of association site.

Identify the functional gene underlying the GWAS signal in defined genetic background (Cont.)

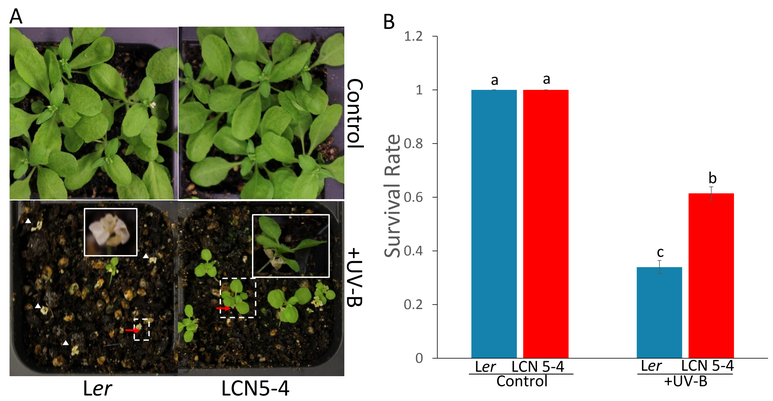

Fig. 5. Difference in UV resistance between L*er* and one NIL.

A. LCN5-4 carrying the Cvi segment around the association site has better resistance against UV radiation. Closeup view of UV caused damage on plant in dashed box is shown in the inset. White triangles: dead plants.
B. Significant difference in UV resistance between L*er* and LCN5-4. Bars with different letters indicate significant differences (P-value<0.01). Error bars indicate SE.

Future plan

Conduct fine-mapping experiment to narrow down functional gene by UV radiation

↓

Conduct UV induced mutation assays with L*er* and the best NIL

↓

Sequence genomes of M2 plants to compare BCS variation between L*er* and Cvi alleles

↓

RNAseq to construct co-expression network for unravelling pathways contributing to BCS variation

- A set of 18 NILs lines will be subjected to UV sensitivity test under the optimal UV dosage. The best NIL that carries the shortest introgression segment for co-localized QTL will be crossed back L*er* to narrow down the region. Next Generation Mapping (NGM) methods will be used to fine map the target gene.
- UV radiation will be applied to accumulate DNA mutations within the best NIL and L*er*. We aim to establish the connection between the functional gene and inside genome BCS variation based on the accumulated DNA mutations. Low and medium levels of UV dosage will be used for two generations. Genomes of M2 plants will be sequenced to identify polymorphisms and test the differences in BCS.
- Transcriptome analysis will be applied to address the question: compared to the L*er*-type function gene, does the Cvi-type tend to fix the error induced by natural UV radiation or to by-pass the error?

References

1. Li X, Scanlon MJ, & Yu J (2015) Evolutionary patterns of DNA base composition and correlation to polymorphisms in DNA repair systems. *Nucleic Acids Res* 43(7):3614-3625.
2. Schmitz RJ, *et al.* (2013) Patterns of population epigenomic diversity. *Nature* 495(7440):193-198
3. Keurentjes JJ, *et al.* (2007) Development of a near-isogenic line population of *Arabidopsis thaliana* and comparison of mapping power with a recombinant inbred line population. *Genetics* 175(2):891-905.