The evolution of weedy plants

Introduction

Weeds are the plant species that usually live in human disturbed habitats and sometimes cause great economical and environmental loss. While the ecology and phenotypes of weedy plants have been much studied, only few have started to investigate the evolution, genetics, and genomics of the origin of weedy plants.

The molecular model plant species *Arabidopsis thaliana* has long been regarded as a human-associated weed. Although this thought has been held for decades, recently through whole-genome sequencing of more than one thousand plants worldwide we have identified genetically distinct populations living in remote regions of natural habitats. Further analysis shoed that those ice age "relicts" once occupied the whole Eurasia before later rapidly replaced by the currently worldwide weed "non-relicts" (Figure 1 and 2). Since the non-relicts mostly inhabit human associated regions such as farms or roadsides and the expansion of non-relict population dates back to more than ten thousand years ago, it is thought that the expansion of human agriculture likely facilitates the worldwide spread of this weedy non-relict population.

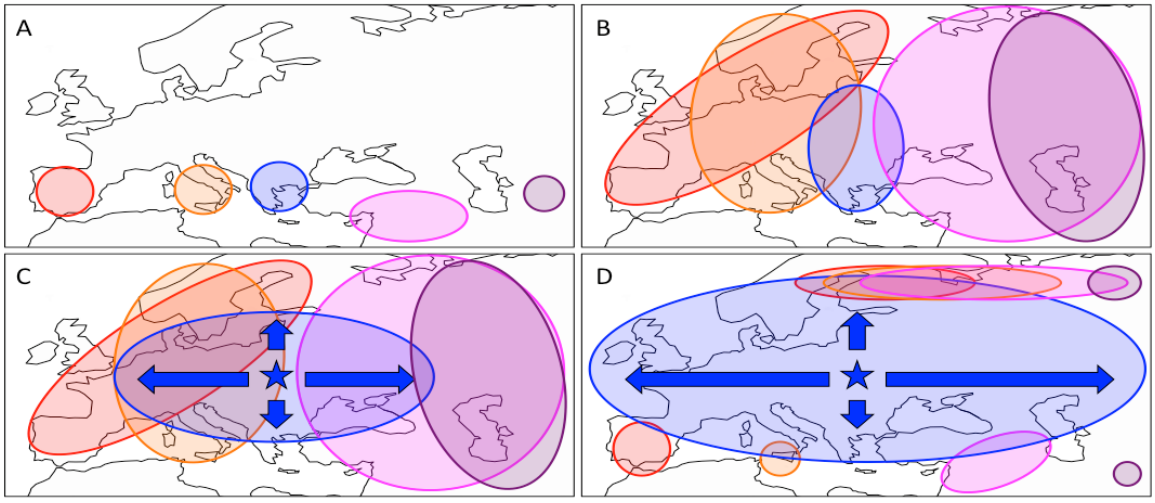

Figure 1. Possible demographic history creating the present-day pattern of spatial genetic variation in *Arabidopsis thaliana*. (A) During the last glaciation, populations in different refugia diverged into separate groups. (B) After ice age, each group expanded northwards. (C) Some time later, one group from eastern Europe (the non-relicts) quickly expanded east-west, generating the present-day pattern that (D) relict genomic regions are mainly found in the south and north of species range (From Lee *et al*. 2016).

We are interested in the genomic consequences of weed expansion and the genetic changes making populations weedy. Using next generation sequencing, we plan to apply population genomics methods and investigate the genomic patterns of hybridization and

adaptive introgression. With the combination of greenhouse work, molecular biology, and bioinformatics, we also plan to identify phenotypes and key genetic changes that enable plants to exhibit weedy phenotypes, such as rapid flowering, drought tolerance, producing more seeds, and longer seed dispersal. The existence of both weedy and natural populations in the same species *Arabidopsis thaliana* therefore enables careful dissection of the genetics of weedy plant evolution.

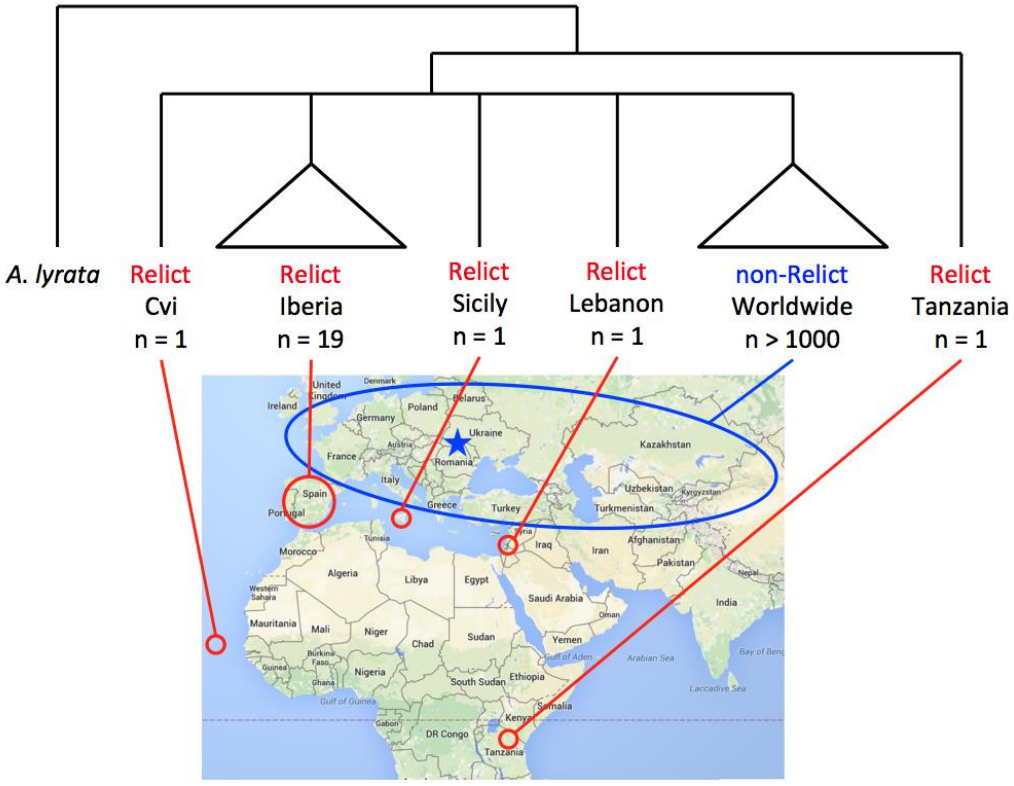

Figure 2. Distribution and phylogenetic relationship of all known native *A. thaliana* genetic groups. The blue star is the putative origin of non-relict expansion, and n represents number of accessions sequenced in each group.

Description of questions

We are interested in these important questions (also see Figure 3):

- Does local adaptation facilitate speciation?
- Why is there genetic variation, and what maintains it?
- What are the factors shaping geographical patterns of genomic variation?
- Some genomic regions are more polymorphic than others. Why?
- What are the genomic regions or genes controlling adaptive traits?
- Are they few genes each with large effect, or many genes each with small effect?
- Does the adaptive allele come from novel mutation or standing genetic variation?
- What are the genetic and genomic changes making a plant weedy?

- Is there gene flow between weedy non-relicts and natural relicts?
- Does such gene flow bring adaptive allele from relicts into non-relict genome?

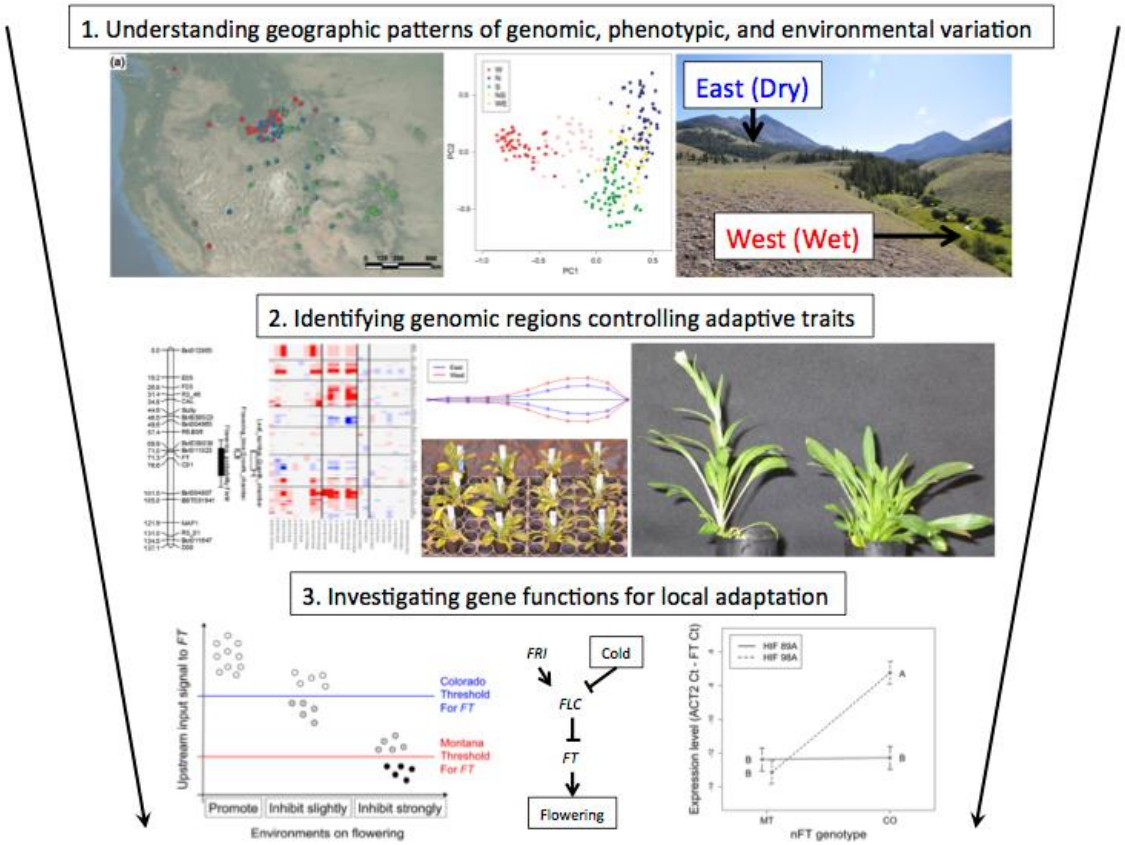

Figure 3. What we are doing. Our research ranges from the scale of landscape patterns genomic variation, to individual traits or genes affection evolution, to dissecting gene functions.

Methods

We mainly use next generation sequencing results and perform population genomics analyses. We mostly use third-party free software and our own custom code in R, perl, or python. Analytical steps involve:
1. Mapping reads from different individuals to a known genome
2. Identifying single nucleotide polymorphism from multiple aligned genomes
3. Analyzing the pattern of polymorphism across geographic regions or landscapes

Illumina reads will first be trimmed according to base quality, using SolexaQA, and trimmed reads will be mapped to the *A. thaliana* TAIR 10 reference genome using BWA. To put our samples in the global context of genomic variation in *A. thaliana*, we will combine our samples with representative relicts and non-relicts from the 1001 genomes project, and single nucleotide polymorphisms (SNPs) will be called jointly from

all samples following GATK best practice. This ensures SNPs from different experiments or populations were called by the same method, facilitating unbiased comparison of our samples with global patterns of genomic variation.

We will investigate the population structure using ADMIXTURE and principal component analysis (PCA). Given whole-genome SNP data and the number of genetic groups (the K value) pre-specified by the user, ADMIXTURE can estimate the proportional ancestry of each individual in the K genetic groups. The analysis will be run through several different K values, and the number of distinct genetic groups that best explain the data (the best K value) will be determined by comparing the cross-validation errors of K values.

In addition to genomic analysis, we will perform quantitative trait loci (QTL) mapping and genome-wide association study (GWAS), identifying loci controlling important quantitative traits.

In brief, GWAS goes though every bi-allelic SNP in the genome and estimates their association with quantitative traits. GWAS uses natural accessions, where historical recombination events through thousands of generations will break the spurious relationship between true causal variant and most unlinked loci, providing better resolution for genetic mapping. Since the whole-genomes of those accessions were already sequenced, even if the true causal variant is not a SNP (e.g. indel or copy number variation), nearby SNPs will still be in sufficient linkage disequilibrium to be detected.

We will use at least five individuals per accession, and mean trait value of each accession will be used for GWAS. Mixed-model GWAS will be conducted with EMMAX to control for confounding effects from population structure or individual relatedness. Many genes have multiple haplotypes, each giving different trait value of quantitative traits. In the presence of such "allelic heterogeneity", studies have shown that using bi-allelic SNPs will not have sufficient power to detect those genes. We will therefore also LIMIX, using local kinship matrix of each genomic region instead of bi-allelic SNPs to account for allelic heterogeneity. I will also use a newly developed GWAS method, multi-trait mixed model (MTMM), which can jointly estimate the direct effect of each SNP to the trait and the magnitude of the genotype-by-environment interaction element of phenotypic plasticity. Once genomic regions controlling these quantitative traits are identified, following our previous study, we will estimate the magnitude of genome-wide introgression of relict alleles into non-relict genomic background and investigate whether SNPs affecting important quantitative traits also have higher magnitude of introgression.

We will use F2 QTL mapping not only to confirm GWAS results in Iberian accessions but also to map genomic regions controlling trait differences in other relict populations, where GWAS is not possible. QTL mapping complements GWAS in three ways: First, unless strong segregation distortion exists, the artificial cross between two parents ensures a roughly 1:1 allele frequency ratio in the mapping population, maximizing the power and minimizing the error to detect QTL. Second, recombination in

the artificial cross removes the complication of population structure. Third, in an artificial cross we could control which relict and non-relict accessions to study. The artificial cross, however, only has a few generations of recombination and thus lower precision to pinpoint the causal variant. Therefore GWAS and QTL mapping can complement each other and are both necessary for our study.

The genotyping step in traditional QTL mapping is tedious and laborious. One has to first identify at least a few hundred molecular markers (either SNPs, restriction enzyme cutting sites, or microsatellites) that differ between two parental accessions and then genotype these markers separately in all 200 F2 individuals. Illumina-sequencing of all F2 individuals avoids the labor of identifying polymorphic markers, but whole-genome sequencing is not cost effective. We will use a newly developed method, multiplexed shotgun genotyping (MSG), to genotype all individuals by sequencing them in Illumina platform. Briefly, MSG is a simplified version of restriction site associated DNA sequencing method (RAD), which uses a restriction enzyme to cut the genome and attaches specifically designed Illumina adaptors to the enzyme cutting sites. One therefore sequences only a subset of the genome – regions near the restriction enzyme cutting sites. With specifically designed barcodes, all 200 F2 individuals could be partially sequenced in one lane of Illumina HiSeq 2000, greatly reducing the labor and cost of genotyping.

Expected results

We expect to find genomic regions or even single genes controlling evolutionarily and ecologically important traits (Figure 4). The results help to answer, for example, what kind of genetic change makes plant weedy? Are they few genes each with large effect or many genes each with minor effect? The results may finally help us understand more about weeds and help combat the ecological and economical loss caused by weeds.

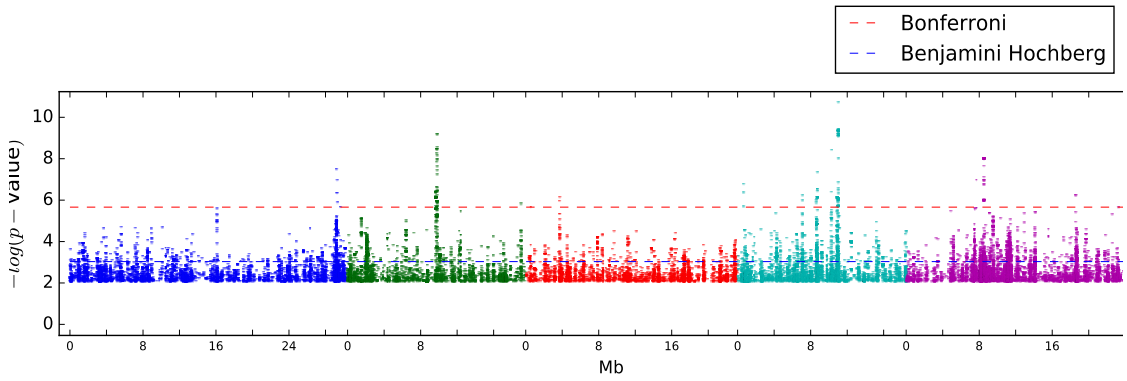

Figure 4. Exemplary genome-wide association study results as a "Manhattan plot". Horizontal axis is physical distance in mega-base-pairs across the genome, and vertical axis is statistical significance of the association between each single nucleotide polymorphism and biological traits. A peak means that genomic region contains candidate genes controlling the trait of interest.