Discovering Common Sequence Variation in *Arabidopsis thaliana*

G. Zeller,* G. Schweikert,* R. Clark, S. Ossowski, C. Toomajian, N. Warthmann, K. Frazer, J. Ecker, M. Nordborg, D. Weigel, B. Schölkopf, and G. Rätsch
(* = contributed equally)

In order to characterize natural sequence variation in 20 strains of *Arabidopsis thaliana*, whole-genome resequencing with high-density oligonucleotide arrays was performed in collaboration with Perlegen Sciences Inc. Array data were analyzed with a combination of existing model-based (MB) (1) and novel machine learning (ML) methods.

For the identification of single nucleotide polymorphisms (SNPs) we developed an algorithm based on support vector machines. Training and evaluation was done on published alignments (2). Confidence levels allow to calibrate false discovery rates (FDR) for the ML SNP predictions. At the same FDR as the MB algorithm, the ML algorithm identifies significantly more true SNPs, especially in regions of high polymorphism density and/or low hybridization quality. The union of SNP predictions from both methods contains on average 143,572 SNPs per strain at a FDR of 2.8% (648,570 non-redundant SNPs).

Furthermore, a machine learning algorithm was developed to detect polymorphic *regions* containing insertions, deletions and variational hotspots, where SNP detection algorithms typically fail to identify *individual* SNPs. It discovers the approximate location of a substantial additional proportion of polymorphisms ($\sim$54% of deleted nucleotides and $\sim$33% of insertion, additional $\sim$42% of SNPs).

We examined the patterns of and forces shaping sequence variation in *Arabidopsis* (3): e.g. significant differences were observed between gene families, and genes mediating interaction with the biotic environment harbor exceptional polymorphism levels.

[1] Hinds D, Stuve L, Nilsen G, Halperin E, Eskin E, et al. (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079.

[2] Nordborg M, Hu T, Ishino Y, Jhaveri J, Toomajian C, et al. (2005) The Pattern of Polymorphism in Arabidopsis thaliana. PLoS Biology 3:1289–1299.

[3] Clark R, Schweikert G, Toomajian C, Ossowski S, Zeller G, et al. (2007) Common sequence polymorphisms shaping genetic diversity in arabidopsis thaliana. Under revision.