Next Generation Sequencing 2 Advanced Course: Transcriptomes, Variant Calling and Biological Interpretation

RNA-seq examples

Kentaro Shimizu

kentaro.shimizu@ieu.uzh.ch

Examples of NGS analysis

- *de novo* genome assembly
- genome-wide polymorphism, and genome-wide association mapping (GWAS)
- RNA-seq (expression analysis, transcriptome)
  - cDNA assembly
- metagenomics
- small RNA
- ChIP-Seq

Two usages of 'mapping'
1. pairwise alignment to a reference genome (resequencing, RNA-seq, Chip-seq, etc.)
2. linkage in genetic crosses or in pedigrees

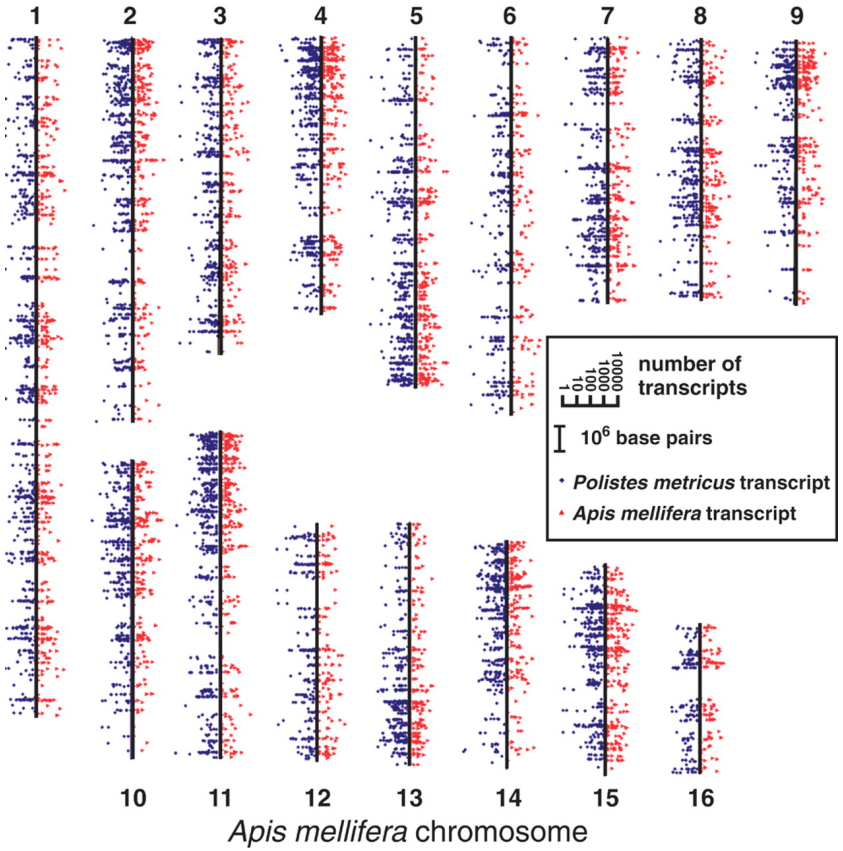
Strategies for transcriptome of non-model species

- Mapping to related model species
- *de novo* assembly of RNA-seq data (long read or short read), and annotation

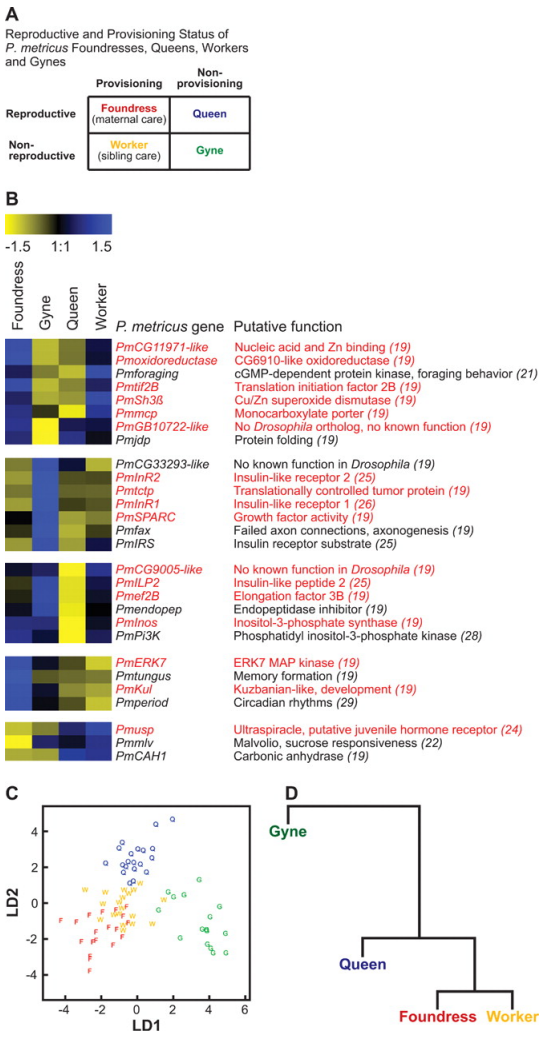
Transcriptome of nonmodel species

Whole genome sequence of honey bee was available, but wasp is distantly related to it.
-> Roche 454 fragments of wasp cDNA are long enough for homology search against honey bee.

Polistes metricus
www.wikipedia.org



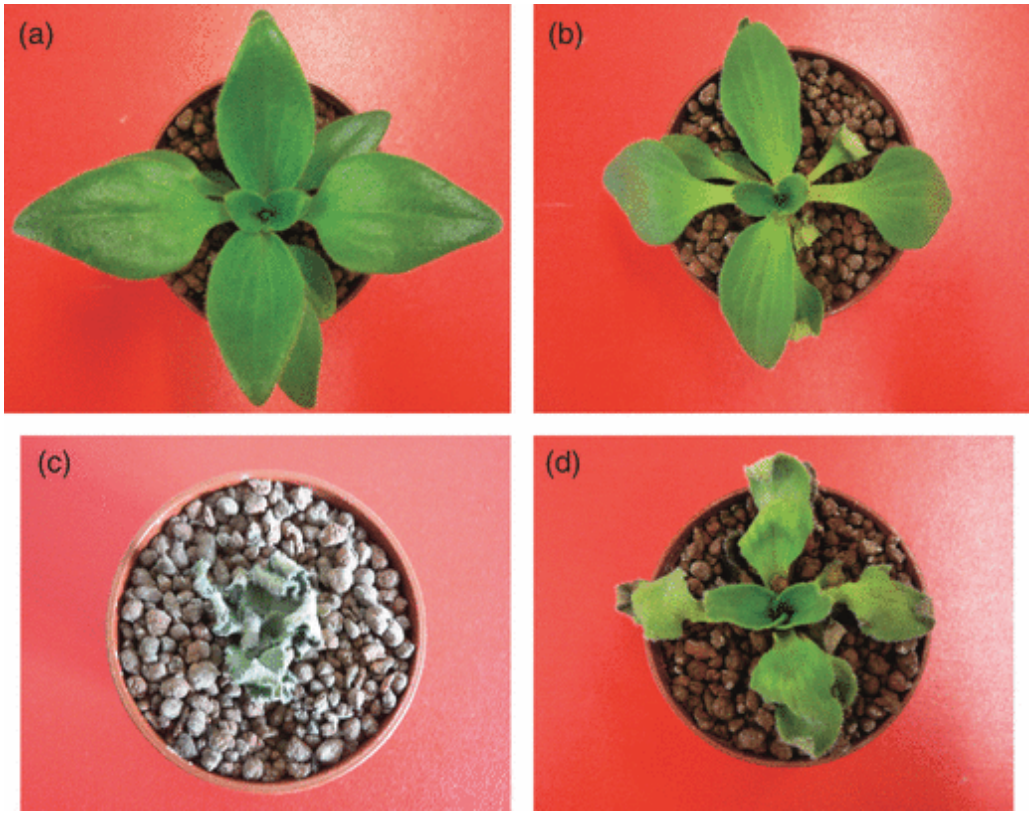
A representation of P. metricus brain transcripts overlaid on a honey bee genome template (16) shows wide coverage and similar transcript abundance for P. metricus relative to known honey bee transcripts
A. L. Toth et al., Science 318, 441 -444 (2007)



P. metricus wasp brain gene expression analysis tests the prediction that maternal and worker (eusocial) behavior share a common molecular basis

A. L. Toth et al., Science 318, 441 -444 (2007)

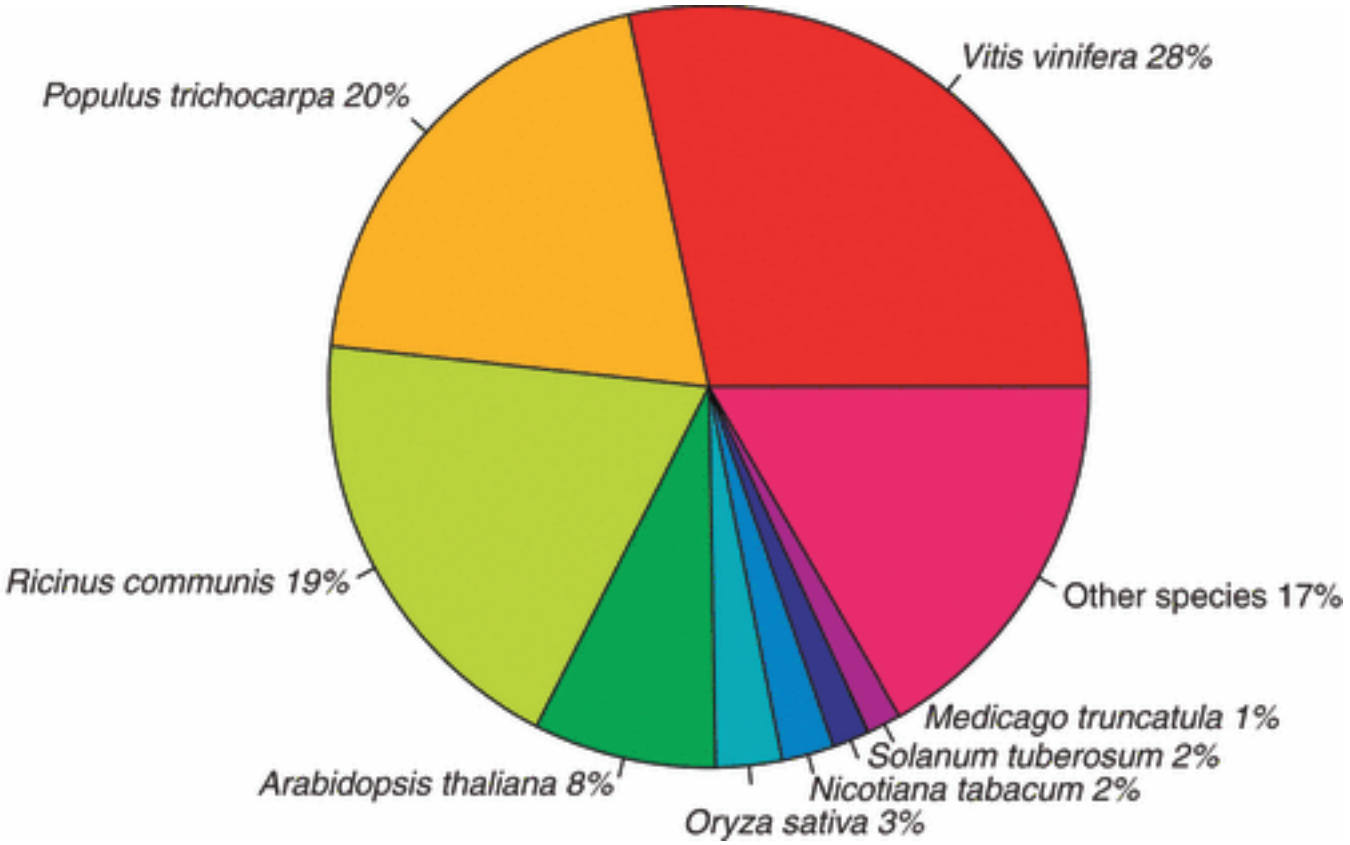
Unique nonmodel plant: desiccation-tolerant resurrection plant



Craterostigma plantagineum, Scrophlariaceae

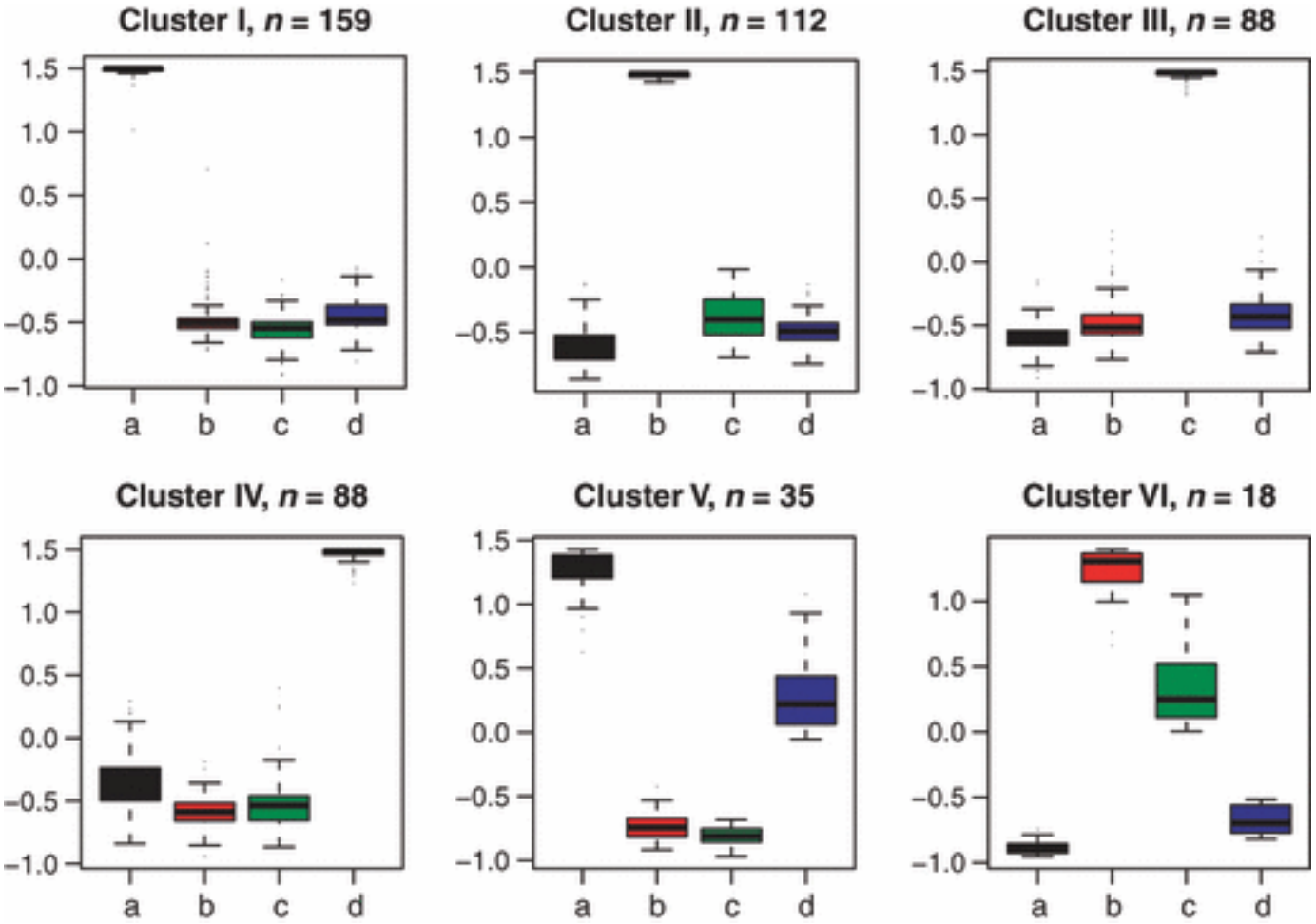
Rodriguez et al. 2010 Plant Journal 63:212-228

- four cDNA libraries
- 182 Mb using Roche 454
- Assembly: 29400 contigs (genes?)
- BLAST



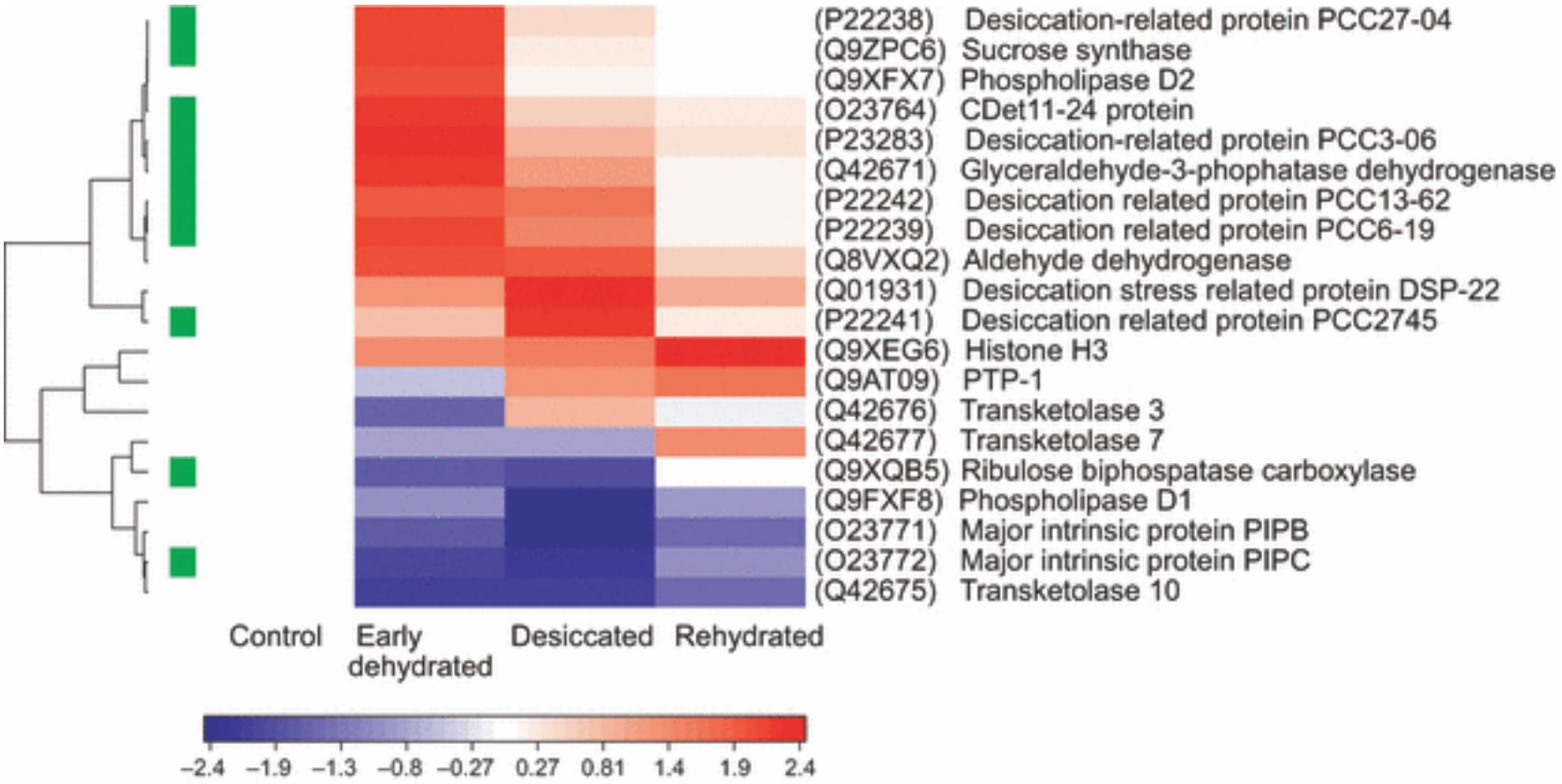
GO (Gene Ontology)

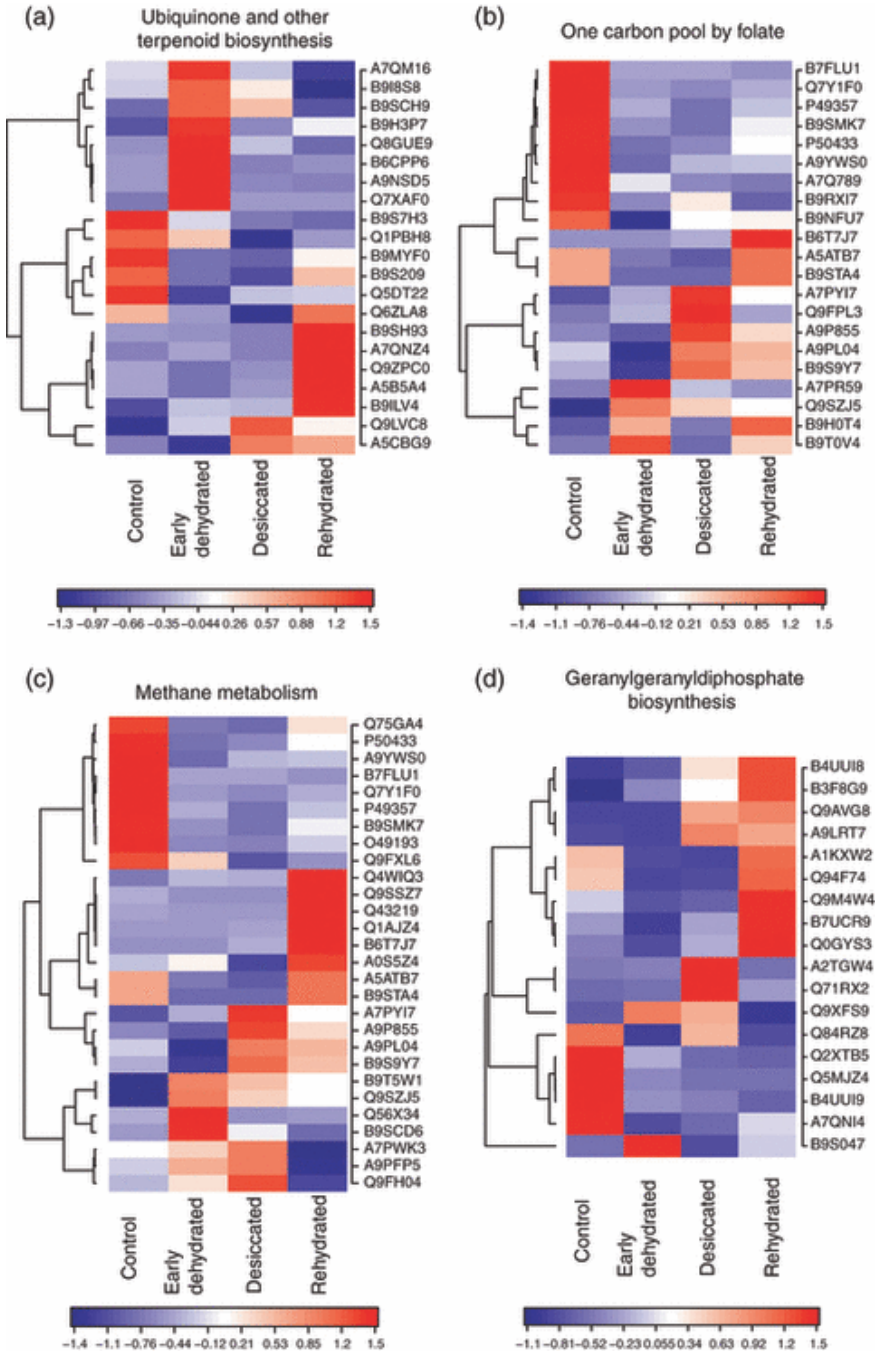
Thiamin biosynthetic process
(protection from oxidative stress)



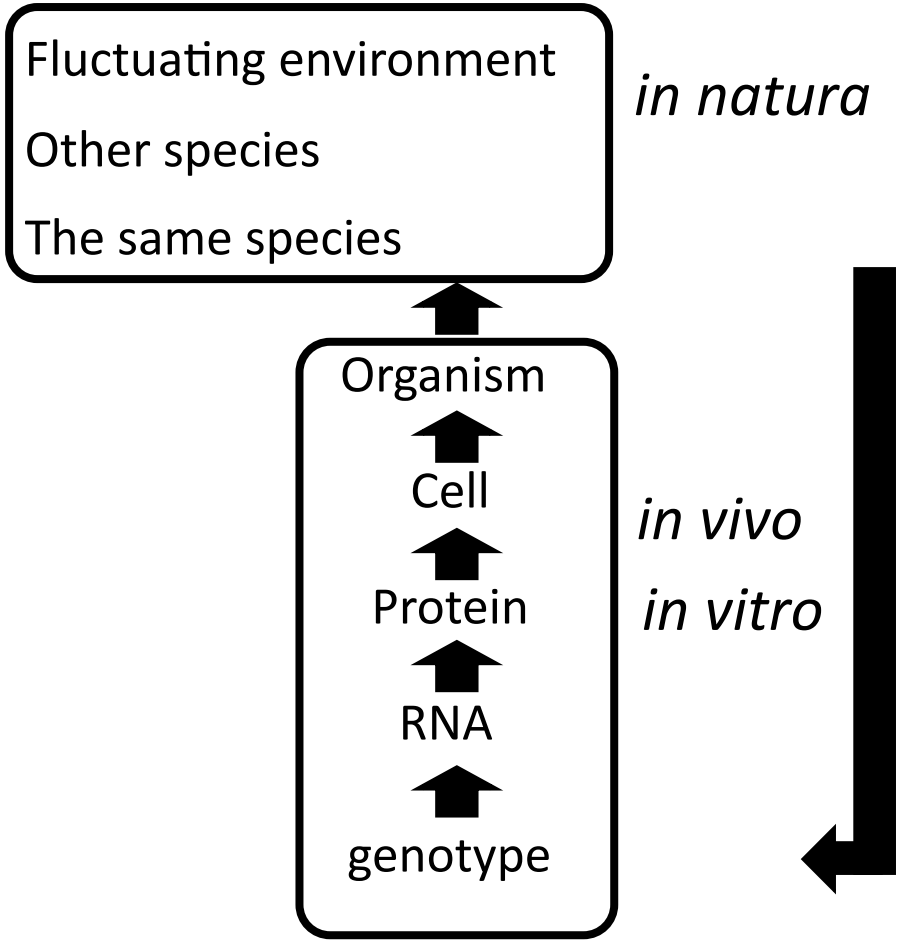
Vitamin K biosynthetic process
(protection from oxidative stress)

Genes already known




(a)
Ubiquinone and other terpenoid biosynthesis
A7QM16
B9I8S8
B9SCH9
B9H3P7
Q8GUE9
B6CPP6
A9NSD5
Q7XAF0
B9S7H3
Q1PBH8
B9MYF0
B9S209
Q5DT22
Q6ZLA8
B9SH93
A7QNZ4
Q9ZPC0
A5B5A4
B9ILV4
Q9LVC8
A5CBG9
Control
Early dehydrated
Desiccated
Rehydrated
−1.3 −0.97 −0.66 −0.35 −0.044 0.26 0.57 0.88 1.2 1.5
(b)
One carbon pool by folate
B7FLU1
Q7Y1F0
P49357
B9SMK7
P50433
A9YWS0
A7Q789
B9RXI7
B9NFU7
B6T7J7
A5ATB7
B9STA4
A7PYI7
Q9FPL3
A9P855
A9PL04
B9S9Y7
A7PR59
Q9SZJ5
B9H0T4
B9T0V4
Control
Early dehydrated
Desiccated
Rehydrated
−1.4 −1.1 −0.76 −0.44 −0.12 0.21 0.53 0.85 1.2 1.5
(c)
Methane metabolism
Q75GA4
P50433
A9YWS0
B7FLU1
Q7Y1F0
P49357
B9SMK7
O49193
Q9FXL6
Q4WIQ3
Q9SSZ7
Q43219
Q1AJZ4
B6T7J7
A0S5Z4
A5ATB7
B9STA4
A7PYI7
A9P855
A9PL04
B9S9Y7
B9T5W1
Q9SZJ5
Q56X34
B9SCD6
A7PWK3
A9PFP5
Q9FH04
Control
Early dehydrated
Desiccated
Rehydrated
−1.4 −1.1 −0.76 −0.44 −0.12 0.21 0.53 0.85 1.2 1.5
(d)
Geranylgeranyldiphosphate biosynthesis
B4UUI8
B3F8G9
Q9AVG8
A9LRT7
A1KXW2
Q94F74
Q9M4W4
B7UCR9
Q0GYS3
A2TGW4
Q71RX2
Q9XFS9
Q84RZ8
Q2XTB5
Q5MJZ4
B4UUI9
A7QNI4
B9S047
Control
Early dehydrated
Desiccated
Rehydrated
−1.1 −0.81 −0.52 −0.23 0.055 0.34 0.63 0.92 1.2 1.5

Can phenotypes and gene functions be understood and predicted in laboratory environments?



Quintana-Murci et al. *Nature Immunology* 2007

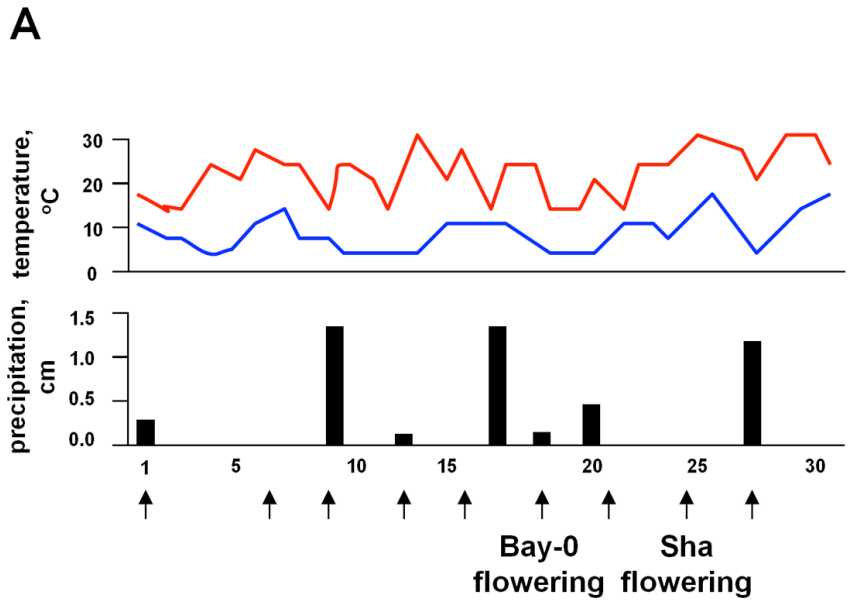
Aikawa et al. *PNAS* 2010, Shimizu, Kudoh, Kobayashi, *Ann Bot* 2011

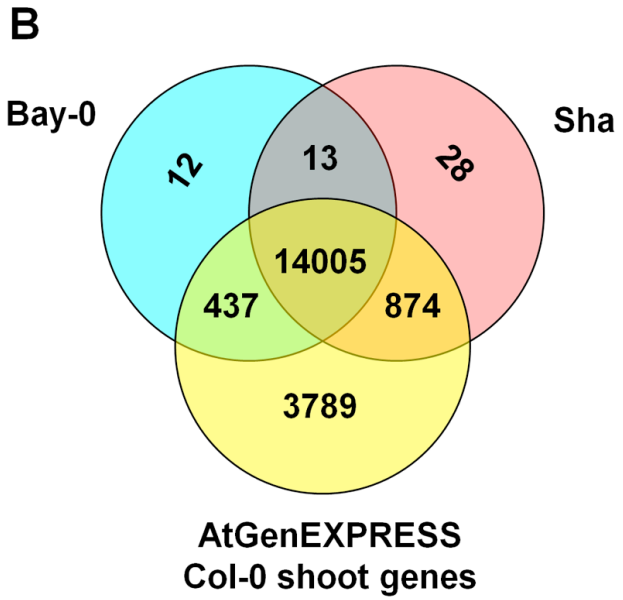
Advantage of plants: they won't escape

Genetic basis of phenotypic plasticity

Example 1: *Arabidopsis thaliana* outside

microarray data at the Cold Spring Harbor Laboratory, NY
9 timepoints x 2 accessions x 3 replicates

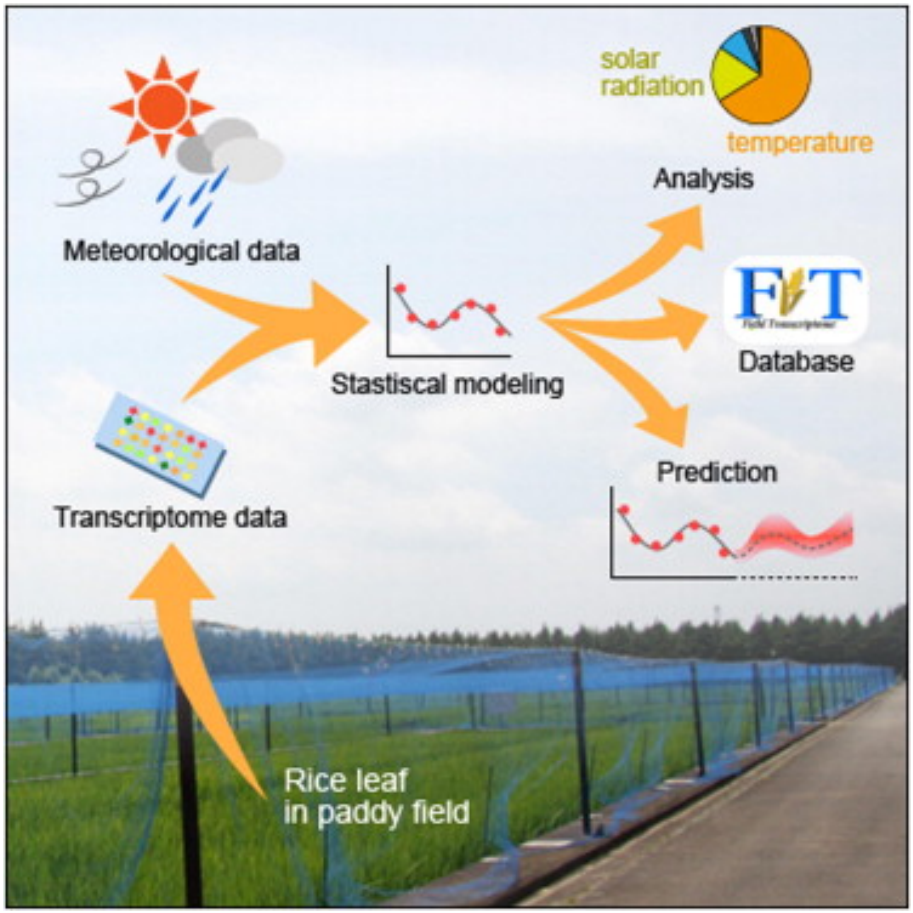




Richards et al. 2012
PLoS Genet 8: e1002662

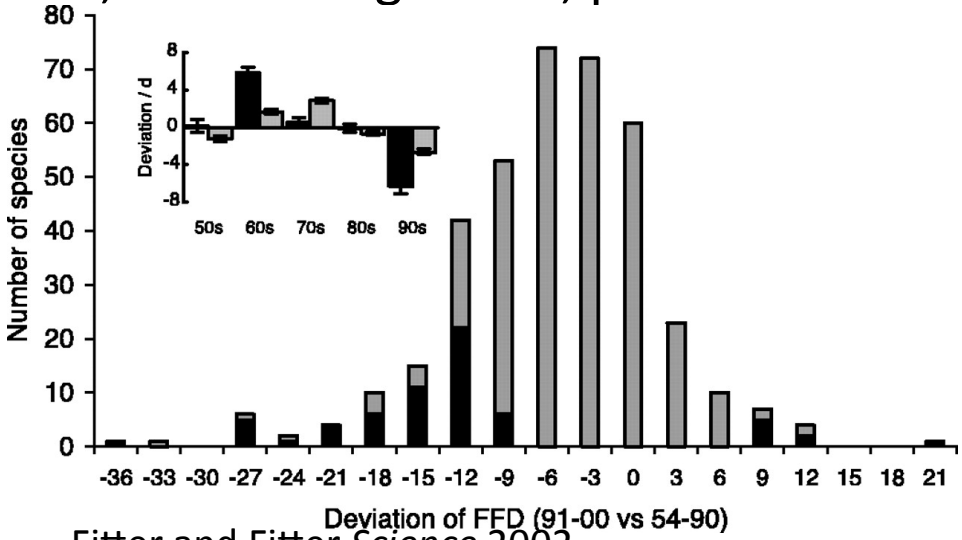
Study 2: rice in agricultural fields

Rice (*Oryza sativa*)
461 microarray data x 27201 genes



Nagano et al. 2012
Cell 151: 1358-1369

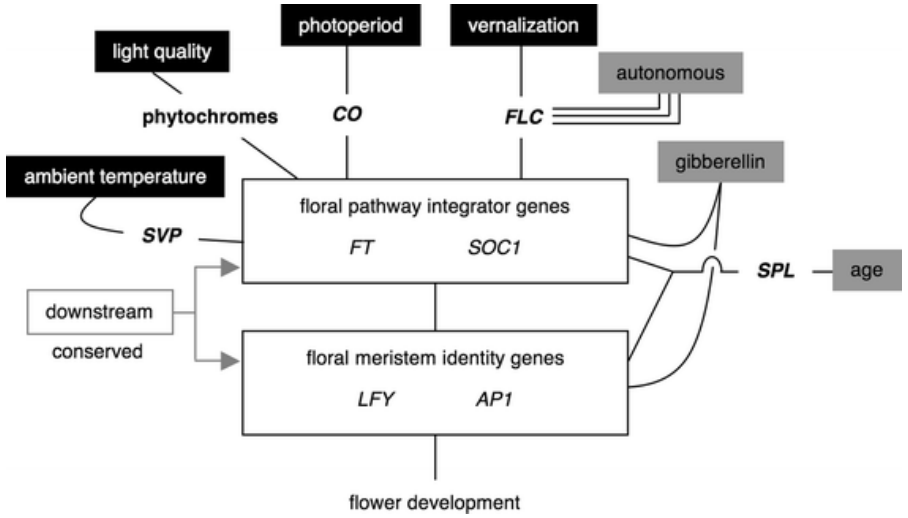
Flowering time as a critical decision for plant reproduction: appropriate environment, outcrossing mates, predators



Fitter and Fitter *Science* 2002.
Rapid changes in flowering time in British plants.

- plant species are beginning to respond to global climate change
  through plasticity
- difficulty in identifying environmental cues and prediction
based on morphological phenotype

Chuine et al. 2003 *in Phenology*

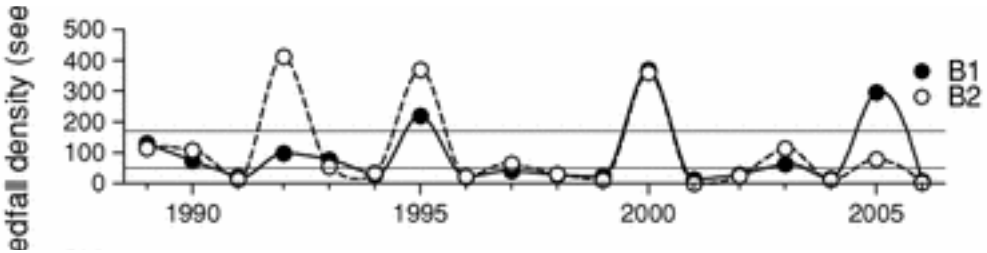


Kobayashi & Shimizu *Ecol Res* 2011

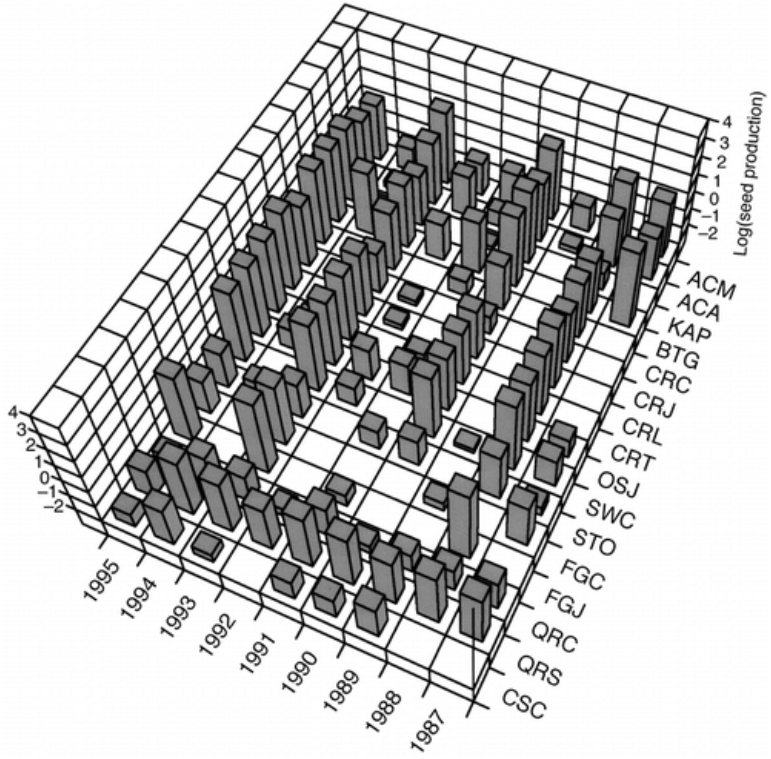
Mass flowering (and mast seeding)

- Synchronization in each species
- irregular interval (one to several years)
- many "keystone" species in temperate and tropical forests
- adaptation: predator satiation hypothesis

Fagus crenata (Masaki et al. *Pop Ecol* 2008)

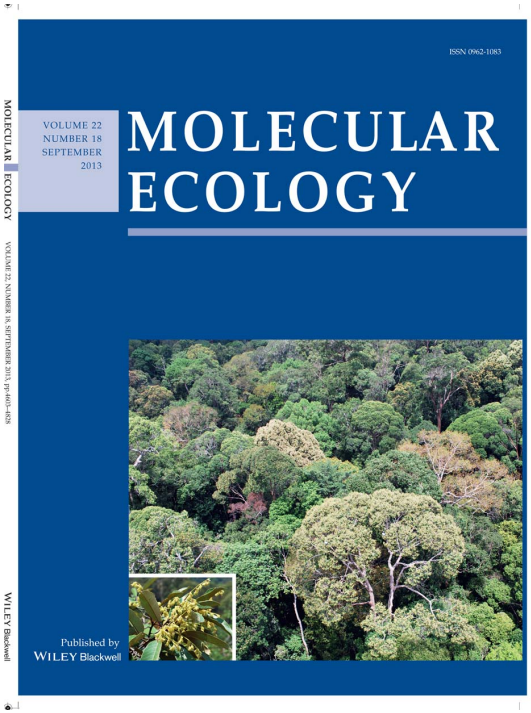


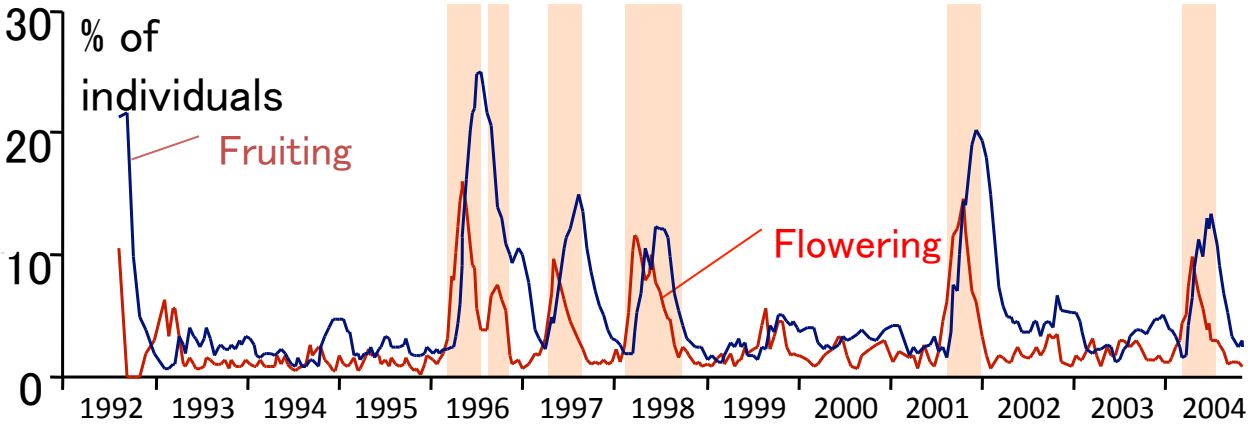
16 principal species in Japanese temperate forest (Shibata et al. *Ecology* 2002)



"General flowering" in Asian tropical forests

- “most spectacular and mysterious phenomena in tropical biology"
- hundreds of species (Dipterocarpaceae and others)
- Ecosystem-wide event: predator satiation and pollinator attraction?

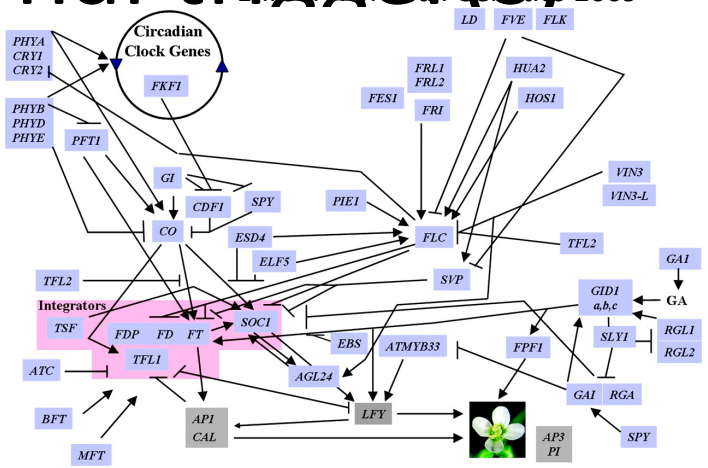




Sakai et al. *Am. J. Bot.* 2002 and unpubl.

Biological synchronization: internal clock and external trigger(s)

Circadian clock
- internal clock: molecular negative feedback such as transcription
- external trigger: light, temperature



Engelhorn et al. *Genetics* 2009

Mass flowering
- Internal clock: resource budget model

(Isagi et al. *J. Theor. Biol.* 1997, Satake & Iwasa *J. Ecol.* 2002, etc.)

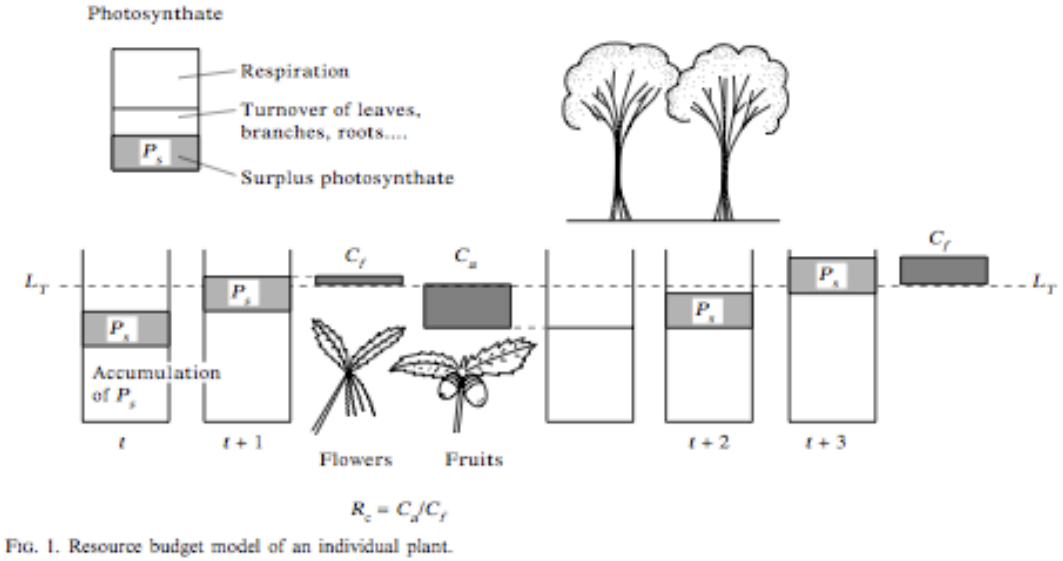


FIG. 1. Resource budget model of an individual plant.

Pollen coupling in temperate forests

What is the external trigger of synchronization?

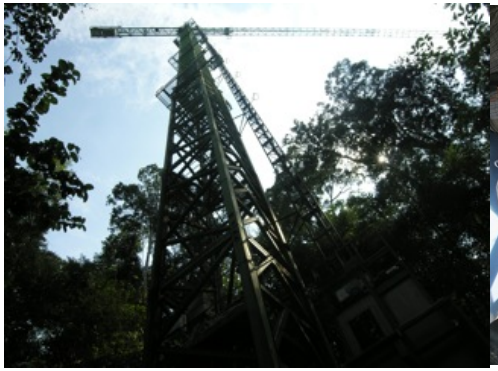
Environmental factors that occur intermittently

- prolonged drought associated with El Niño Southern Oscillation.
(Medway 1972; Appanah 1985; Sakai et al. 2006; Brearley et al. 2007)
- increased sunshine hours due to less cloudy conditions (Wright & Vanschaik 1994)
- increase or decrease in mean air temperature (Appanah 1985)
- falls in minimum air temperature (Ashton et al. 1988; Yasuda et al. 1999)

Which is the trigger of the flowering?
What kind of genes are up- and downregulated?
Would the forest be sustained in facing the climate change?

Lambir Hills National Park

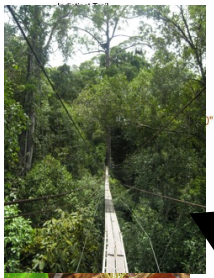
- a highest diversity of tree species on the earth
- 30 minutes from Miri airport

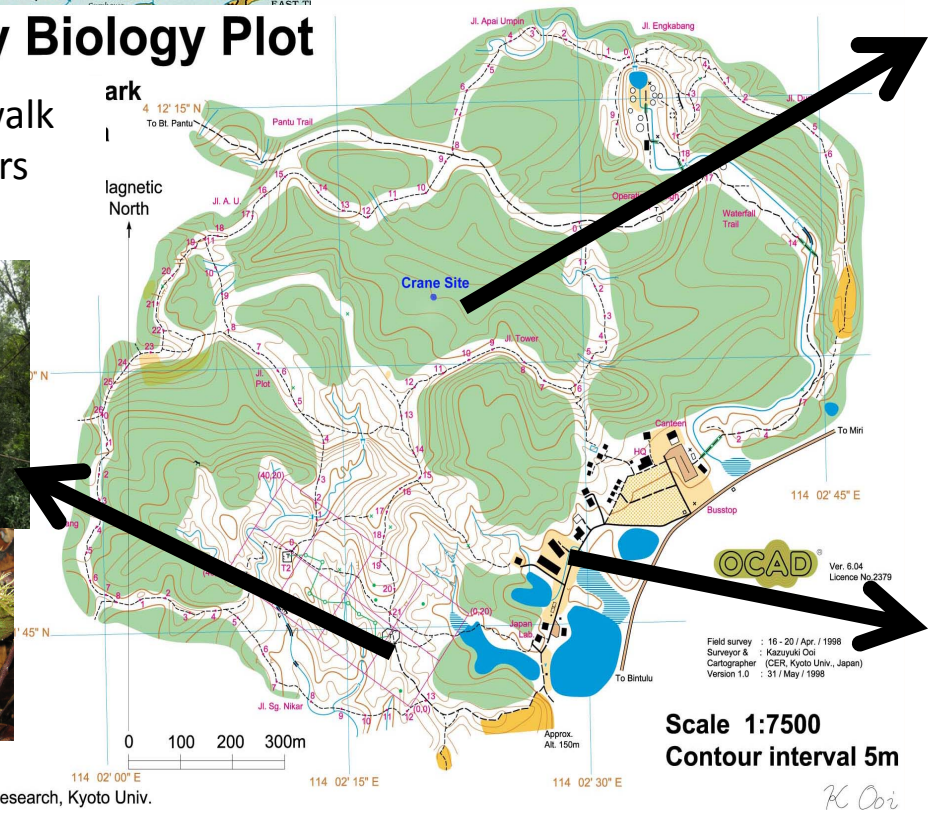


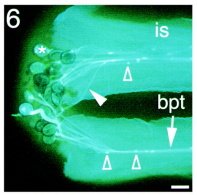
Liebherr Crane
80 m height, 80 m radius, 4 ha plot

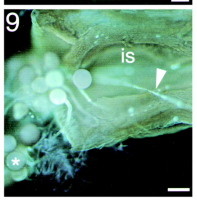
Canopy Biology Plot

Canopy walk
and towers
8ha plot







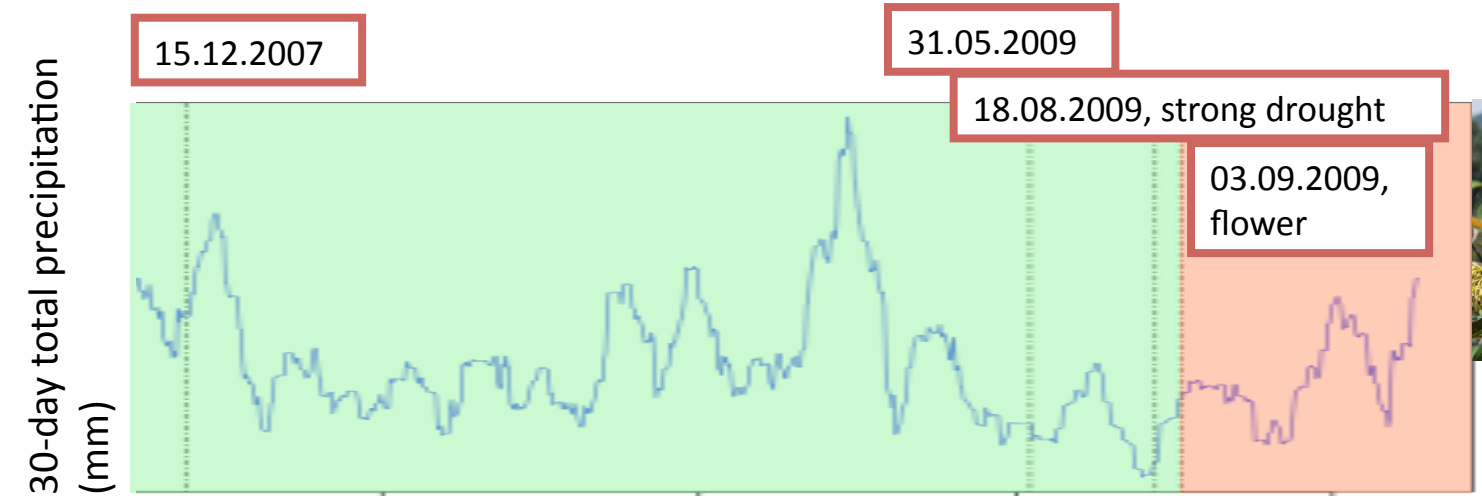


Kenta, Shimizu et al.
Self-incompatibility
of *Dipterocarpus tempehes*
Am. J. Bot. 2002

Molecular laboratory and housing
in front of the forest

Takeuchi et al. Bacterial community
PLoS One 2011 and in revision

Ecological transcriptome of 'non-model' species



Flower of *Shorea beccariana*

Masaki Kobayashi
RNA-seq using
Roche 454 Sequencer

546,212 sequences (x 251.7 bp on average)

>GJ8M0OH02GGVDD length=120
AAGCAGTGGTATCAACGCAGAGTACGCGGGGAACCAGATATTCGA
AAAACGAATTGCTTAATTACAGTATTGATTTTGGAATTTGTATATTTT
GCGGTAACTGCGTTGATACCACTGCTT

>GJ8M0OH02FSN3R length=169
ATCAGCATGTAGGCAGCATTTTGTTGATTCCTTCTCTTTTCTGTGGT
CCATGTATTTGCGTATTTTCGCTGTATAGCCATCCTCCTCGATTCCGA
GTATCCTTAGAACTTATGTTTGTAAACGTATTCTAGTATGCTGCTTTC
TAGAACAATCTATTTGGGTTTTTNTT

>GJ8M0OH02FYFJO length=168
ATCAGCATGTAGGCAGCATTTTGTTGATTCCTTCTCTTTTCTGTGGT
CCATGTATTTGCGTATTTTCGCTGTATAGCCATCCTCCTCGATTCCGA
GTATCCTTAGAACTTATGTTTGTAAACGTATTCTAGTATGCTGCTTTC
TAGAACAATCTATTTGGTTTTTNTT

homology search

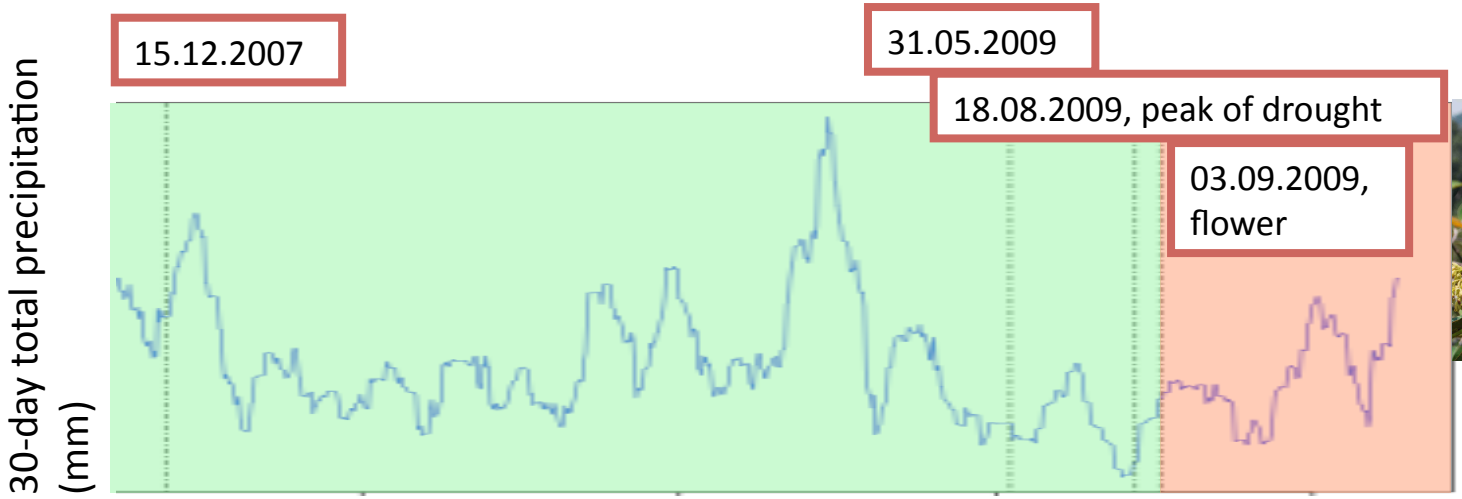
gene A

gene B

gene A

Kobayashi et al. *Mol Ecol* 2013

Ecological transcriptome of 'non-model' species



Flower of *Shorea beccariana*

Masaki Kobayashi
RNA-seq using
Roche 454 Sequencer

546,212 reads

↓ assembly using Newbler
BLASTX to Arabidopsis proteome

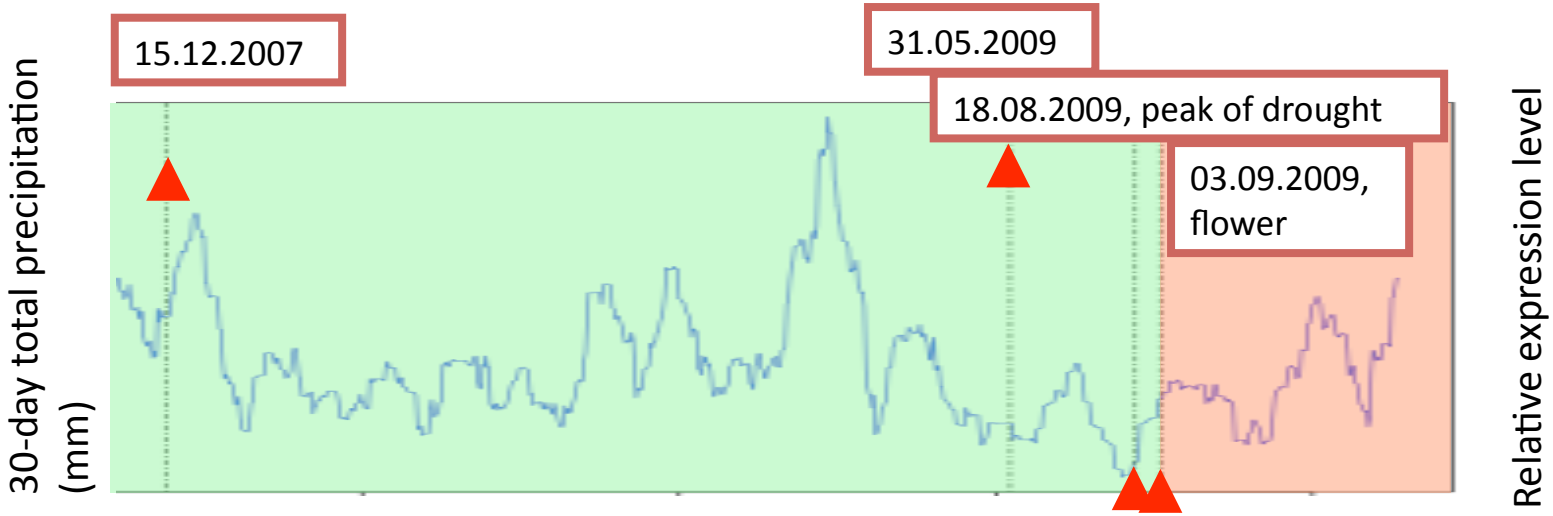
29,527 unigenes with homology to Arabidopsis

↓ DEGseq

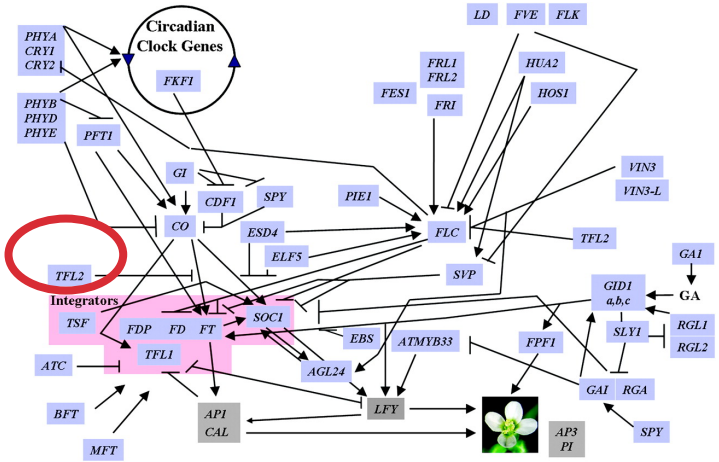
1,128 differentially expressed genes

Kobayashi et al. *Mol Ecol* 2013

Differentially expressed genes 1: flowering genes

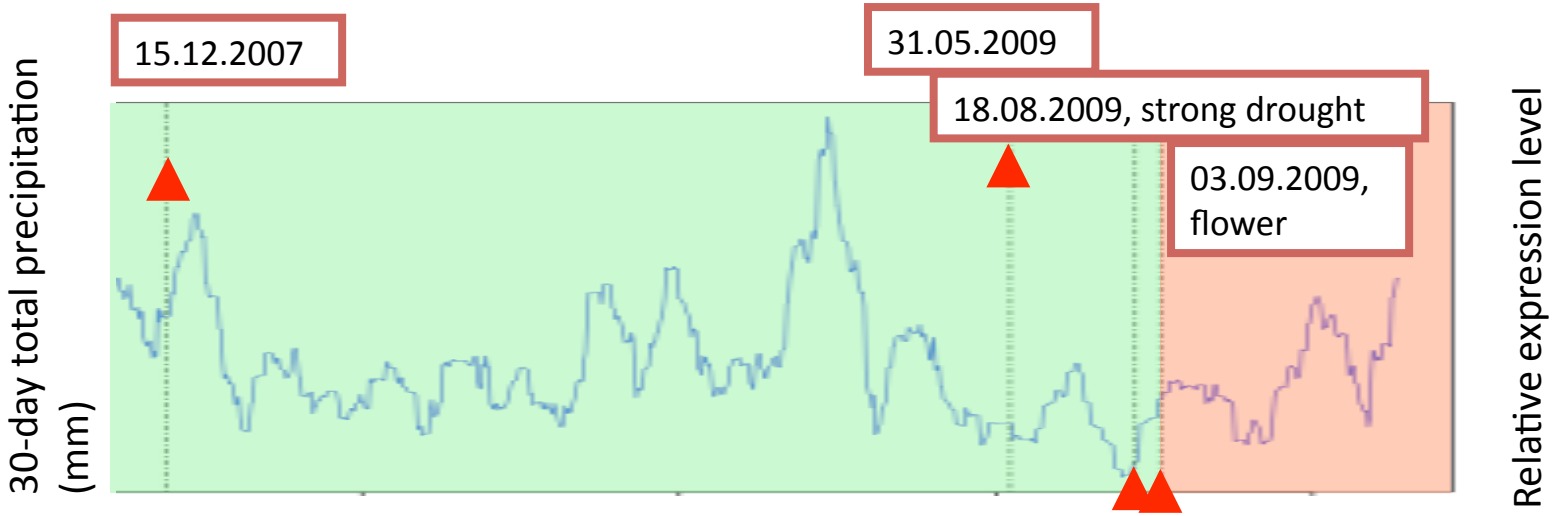


Homolog of *SVP (SHORT VEGETATIVE PHASE),* a MADS-box flowering gene

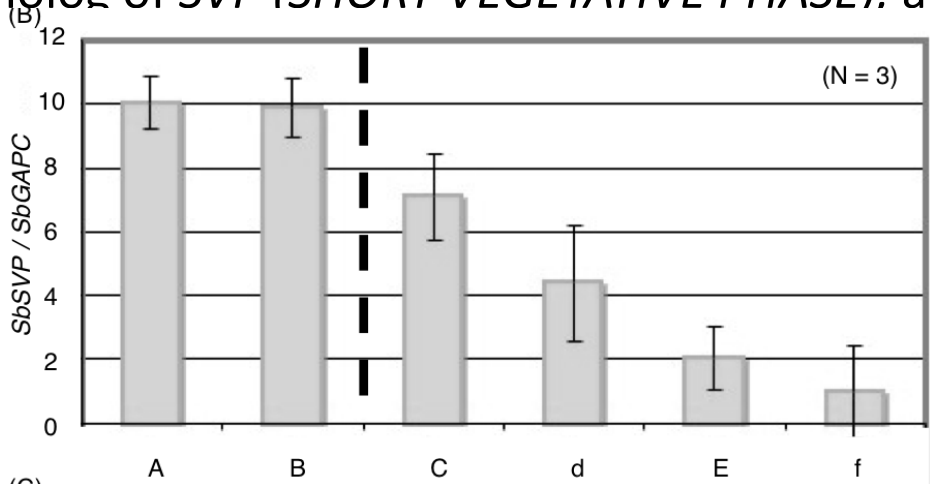


Ehrenreich et al. *Genetics* 2009

Experimental validation of differential expression and function



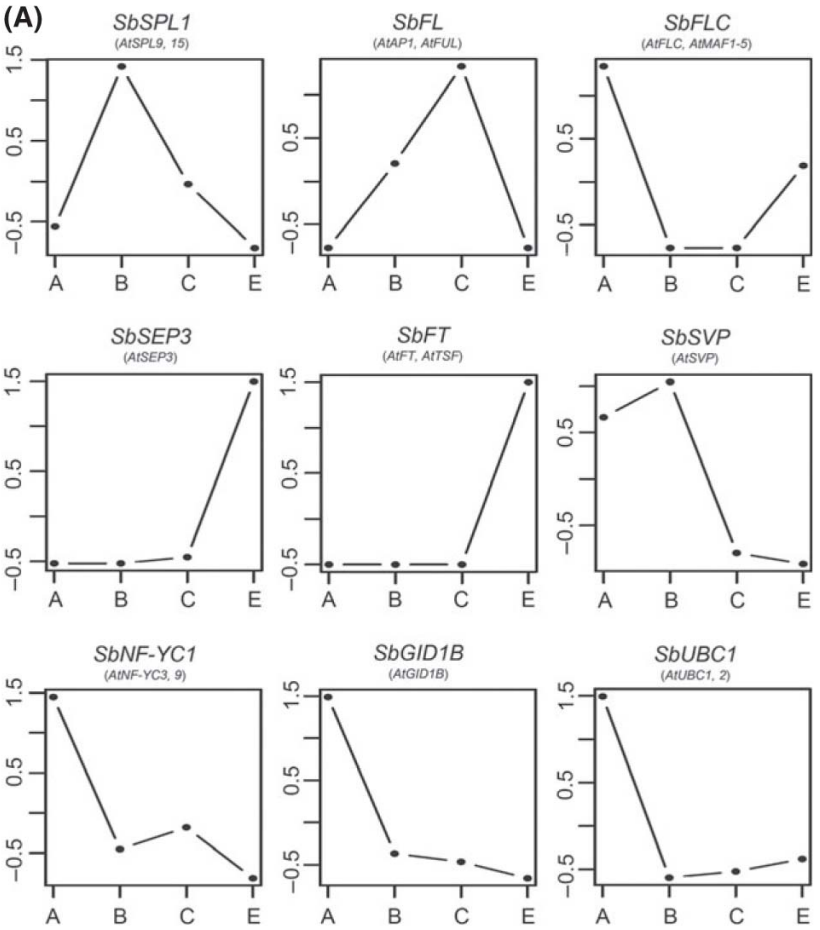
Homolog of *SVP (SHORT VEGETATIVE PHASE).* a MADS-box flowering gene

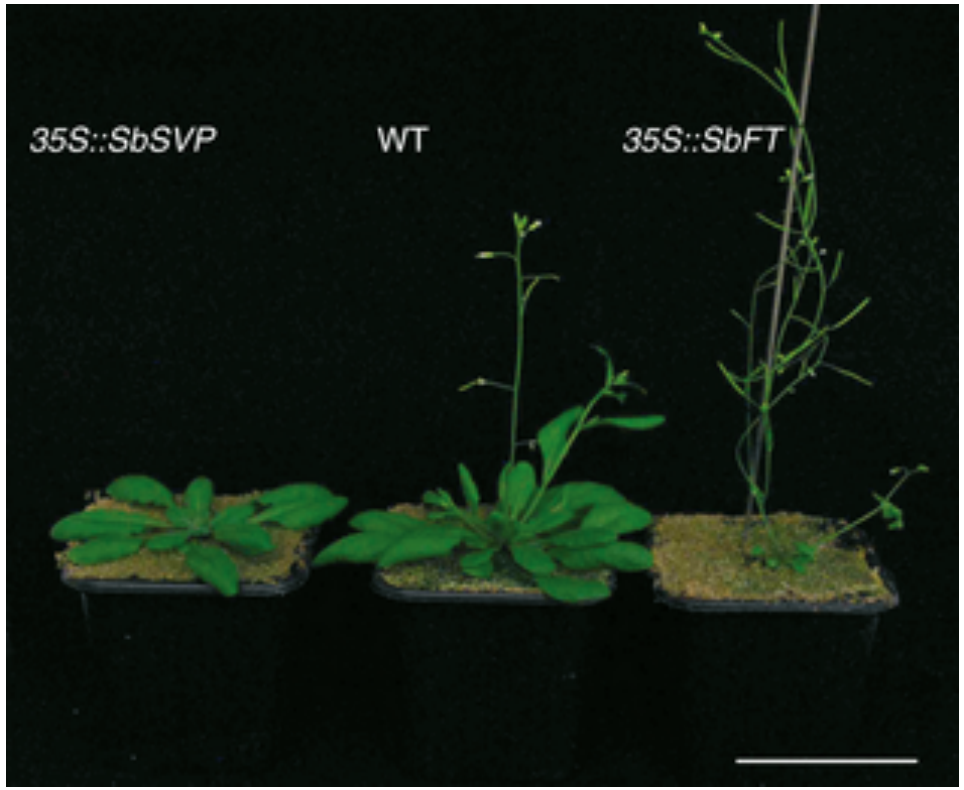


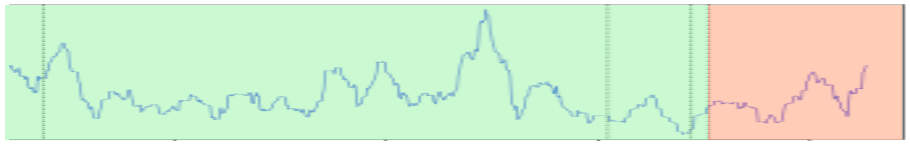
Real-time PCR

overexpression control

60% (98/181) of the homologs of Arabidopsis flowering genes were expressed in the four samples



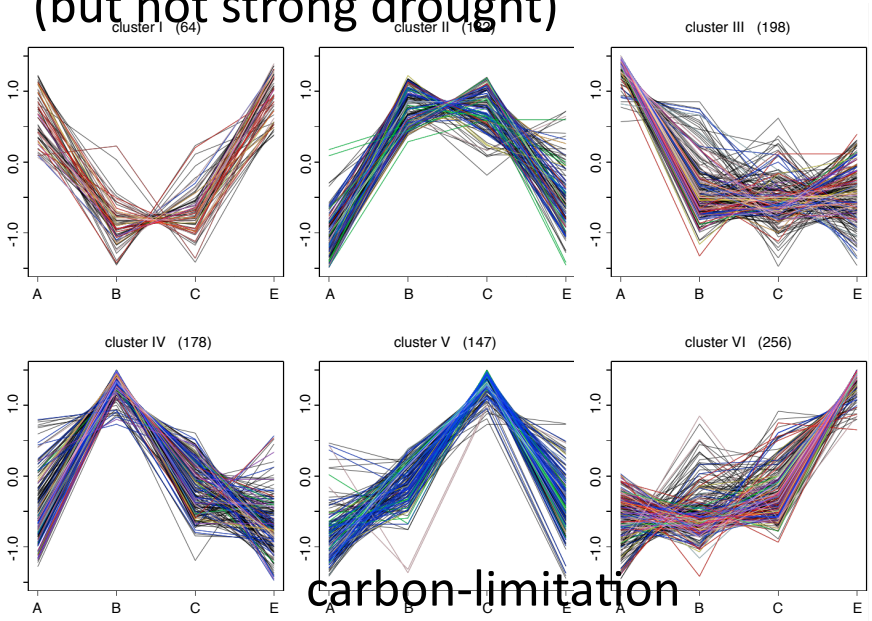




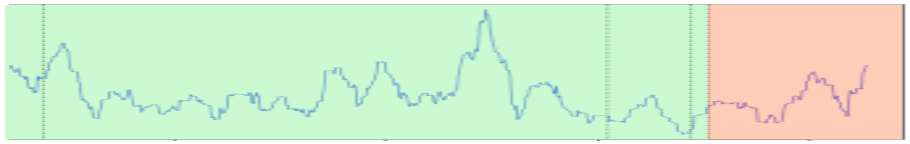
Differentially expressed genes 2: stress-responsive genes

- Data of this non-model species were compared with *Arabidopsis* public transcriptome data (33 datasets),
- complementary to Gene Ontology Enrichment

downregulated
by mild drought
(but not strong drought)



carbon-limitation
temperature



	differentially expressed in Borneo (975 genes)	not differentially regulated in Borneo
Downregulated by mild prolonged drought in Arabidopsis	55	125
Other genes	1073	7314

Fisher's exact test p < 2.68E–08

-> drought as a trigger,
not as a severe stress
-> Nutrient-limitation?

Summary

- Ecological genomics / Ecological transcriptome
- Analysis of gene expression pattern *in natura* combined with meteorological data is a powerful tool to understand and predict biological responses to environmental changes
- For conservation: a major difficulty in restoring Southeast Asian rain forests is the difficulty to obtain enough seeds. The gene expression level may be useful to prepare for seed collection by predicting flowering.
- Severity and frequency of drought by El Nino cycles is predicted to be enhanced in Borneo, which would hinder the maintenance of the forests

Weston et al. (2013) *Mol Ecol*
As illustrated by Kobayashi et al. (2013), the inclusion of molecular biology, genomics and bioinformatics has the potential to shed light on long-standing questions of ecological concern

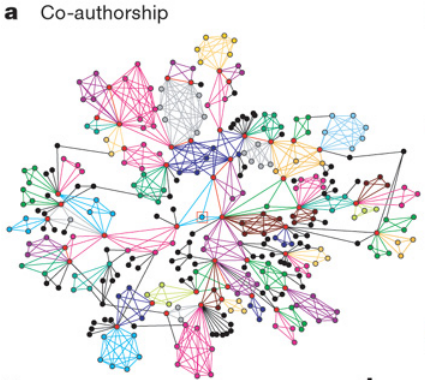
Environmental responses of polyploids

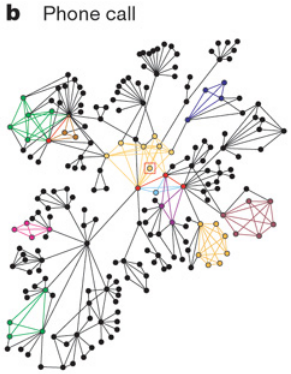
- Genome duplication (polyploidization) is prevalent in animals, fungi and plants including crops

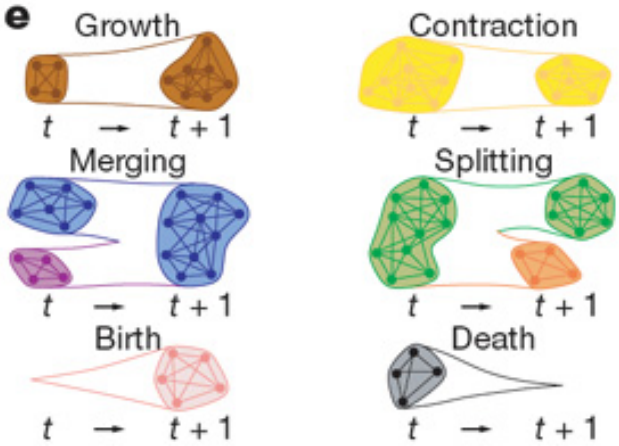
Ohno 1970, Swalla, *Heredity* 2006

Leitch & Leitch
Science 2008

Why can polyploid species often adapt to distinct or broader ecological niches?
Is there a short-term advantage of allopolyploidization?







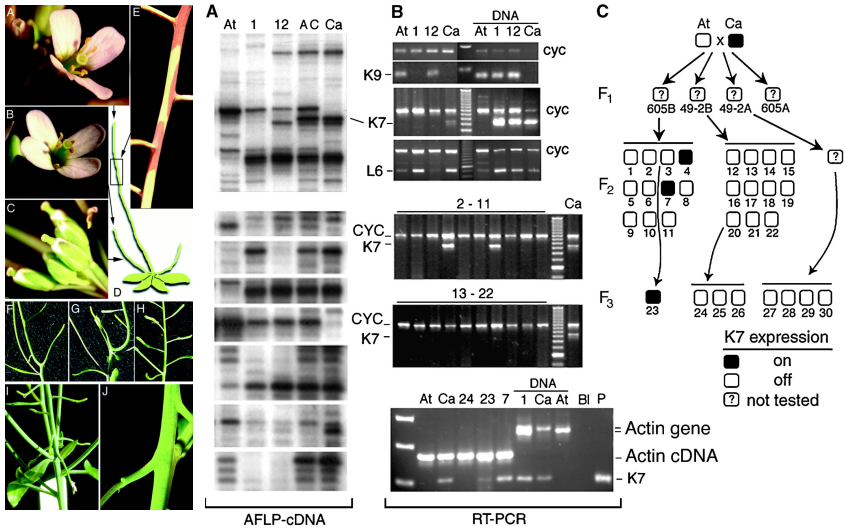
Pella, Barabasi, Vicsek (2007) *Nature*

HFSP

Jun Sese, AIST

Angela Hay, MPI

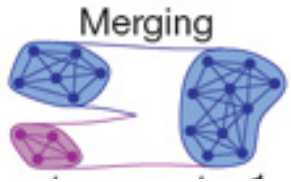
Stochastic variation in synthetic polyploids?



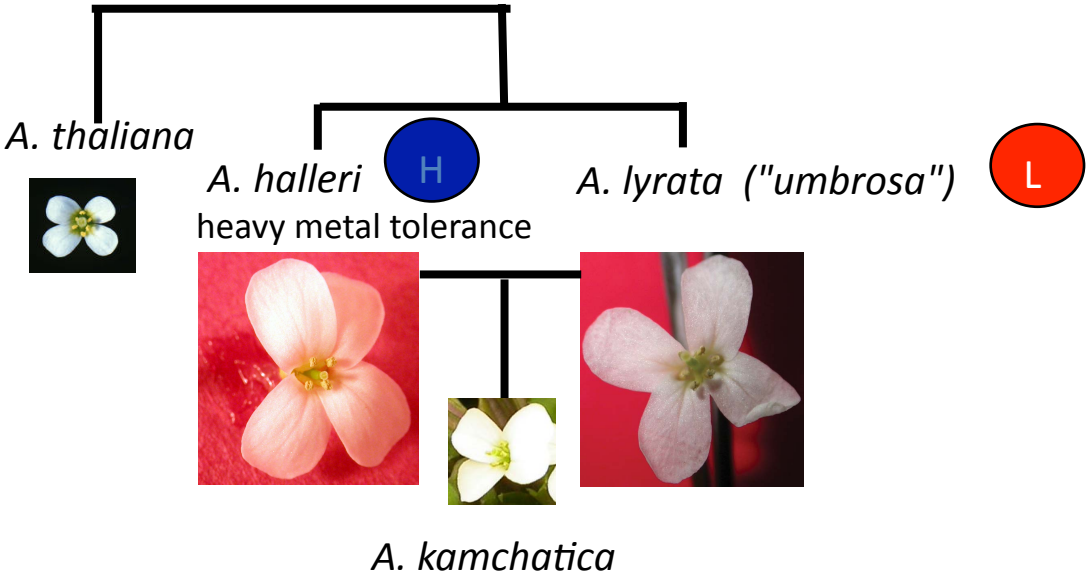
Comai et al. *Plant Cell* 2000

Synthetic polyploids of *Arabidopsis suecica*, wheat, cotton, etc.

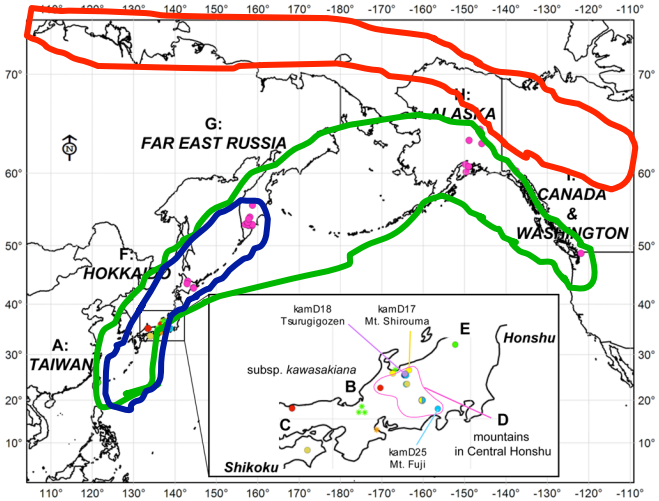
- the same DNA sequence
- high variations in phenotype, gene expression and DNA methylation
- considered stochastic (but not genome-wide)



Arabidopsis kamchatica
a broadest ecological niche in the genus *Arabidopsis*



Precipitation and temperature,
Hoffmann, *Evolution* 2005

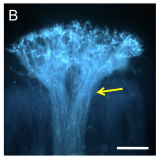


Shimizu-Inatsugi et al. *Mol Ecol* 2009
Tsuchimatsu et al. *PLoS Genet* 2012

ssp. *kamchatica*

ssp. *kawasakiana*

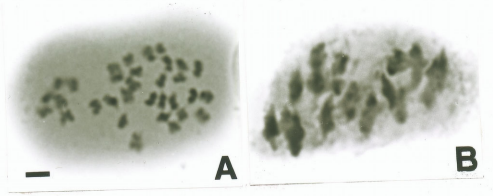
self-compatibility

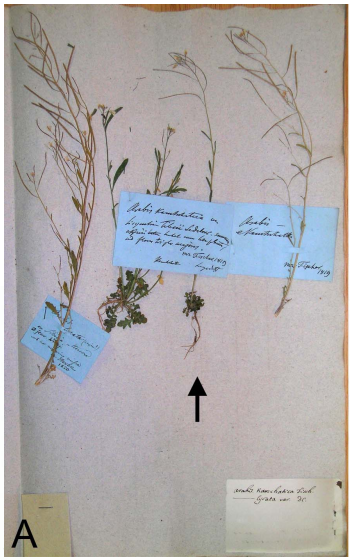


Taxonomic revision
species vs. subspecies

Arabidopsis kamchatica (Fisch. ex DC.) K. Shimizu & Kudoh, **comb. nov.**
Basionym. *Arabis lyrata* var. *kamchatica* Fisch. ex DC., Syst. Nat. 2: 231 (1821).
Type: Kamchatka, *Fisher s. n.* (lectotype, G-DC).

Arabis kamchatica (Fisch. ex DC.) Ledeb (1841).
Arabis lyrata subsp. *kamchatica* (Fisch. ex DC.) Hultén (1937).
Cardaminopsis kamchatica (Fisch. ex DC.) O. E. Schulz (1936).
Arabidopsis lyrata subsp. *kamchatica* (Fisch. ex DC.) O'Kane & Al-Shehbaz (1997).
Arabis lyrata var. *occidentalis* Wats (1895).
Arabis occidentalis (Wats.) Nelson (1937).
Arabis ambigua DC. var. *intermedia* DC. (1821).
Arabis lyrata var. *intermedia* (DC.) Farwell (1917).
Arabis kamchatica var. *intermedia* (DC.) N. Busch (1926).
Arabis ambigua DC. var. *glabra* DC. (1821).
Arabis lyrata var. *glabra* (DC.) Hopkins (1937).
Arabis kamchatica var. *glabra* (DC.) N. Busch (1926).

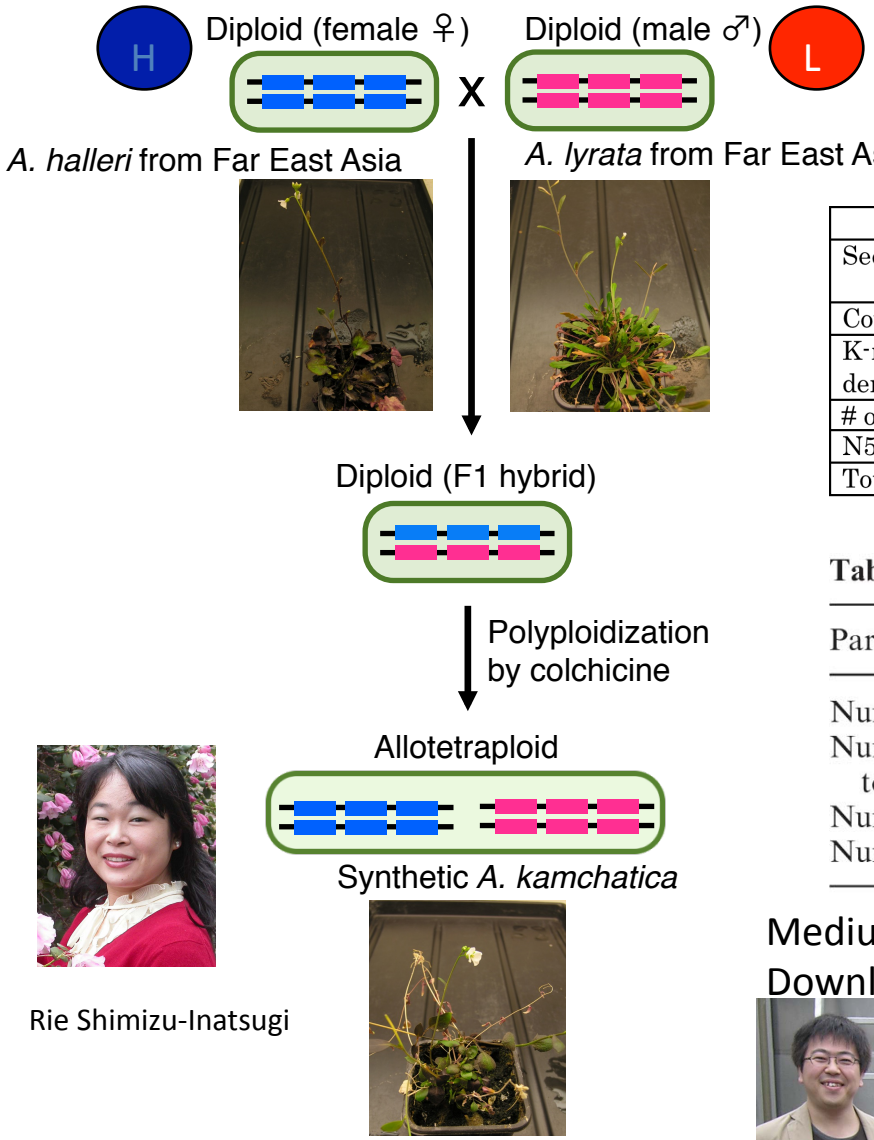




Type from Kamchatka,
Fischer 19 (meaning 1819)
to de Candolle in Geneva

Shimizu et al. *Acta Phytotax Geobot* 2005

To tackle the complex polyploid genome: synthetic allopolyploid as the simplest model



Rie Shimizu-Inatsugi

selfed
5 times

	A. halleri	*A. lyrata*
Sequenced nucleotides	36,942,496,896 bp (200, 500 and 800bp)	61,445,605,532 bp (200, 500 and 800bp)
Coverage (220Mbp)	167.92x	279.30x
K-mer's K in SOAP denovo	73	83
# of contigs	282,453	281,536
N50	17,686	7,848
Total length	221.14Mbp	202.97Mbp

Table 1. Statistics of parental species' genes

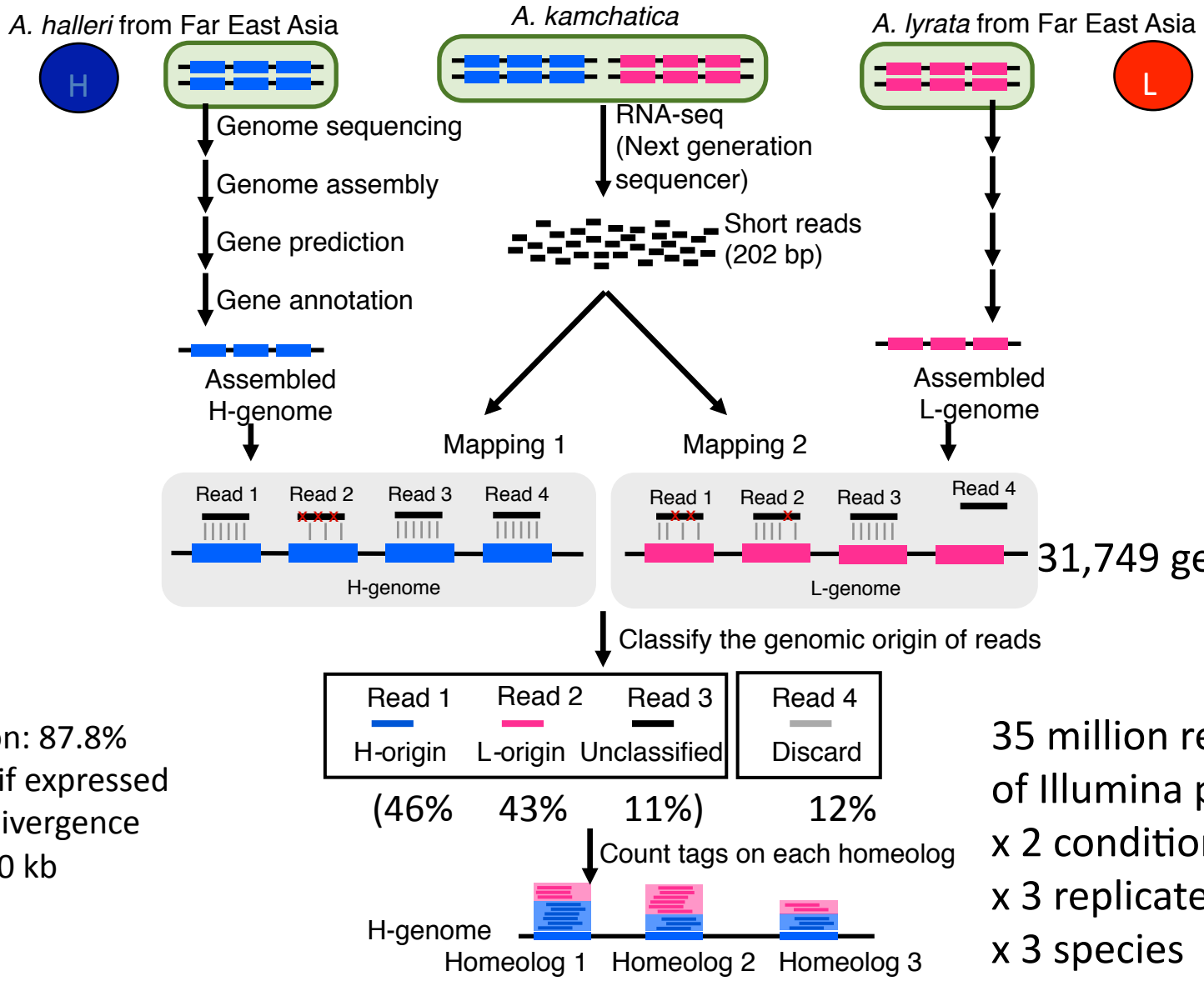
Parental species	*A. halleri*	*A. lyrata*
Number of predicted genes	36 737	35 392
Number of genes related to *A. thaliana*	21 263 (57.9%)	21 166 (59.8%)
Number of homeologs	31 749 (86.4% based on *A. halleri*)	
Number of expressed homeologs	18 928 (59.6%)	19 186 (60.4%)

Medium quality is enough for this purpose
Download and link: http://seselab.org/homeoroq/

Jun Sese, Satoru Akama, AIST

Akama et al. *Nuc Acids Res* 2014

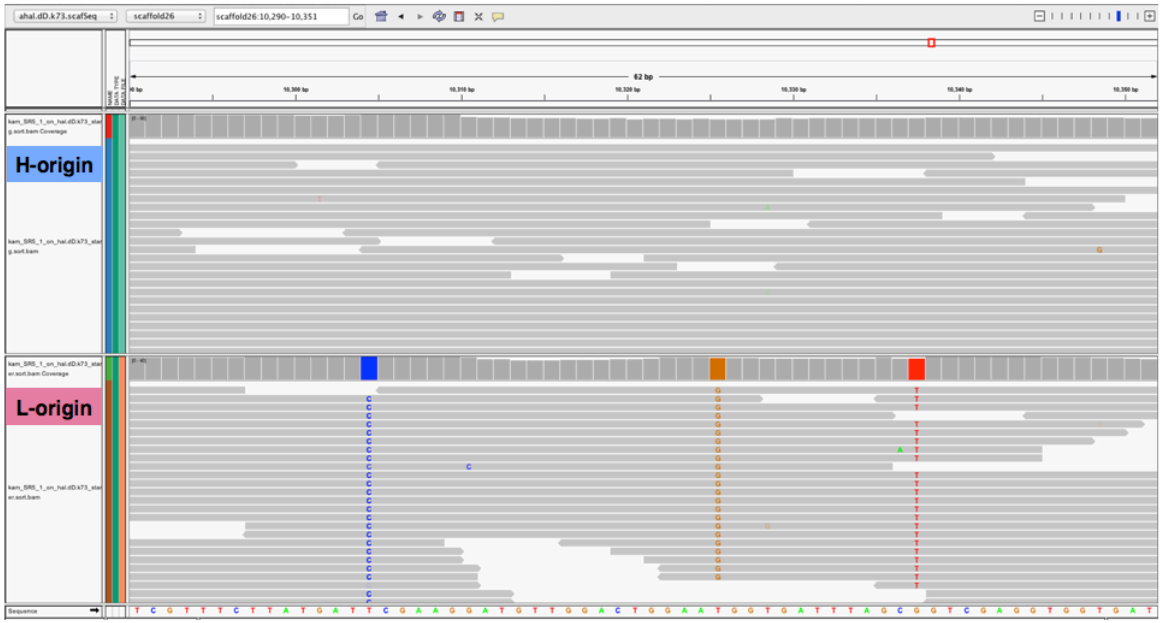
HomeoRoq (Homeolog Ratio and Quantification): bioinformatic workflow using NGS

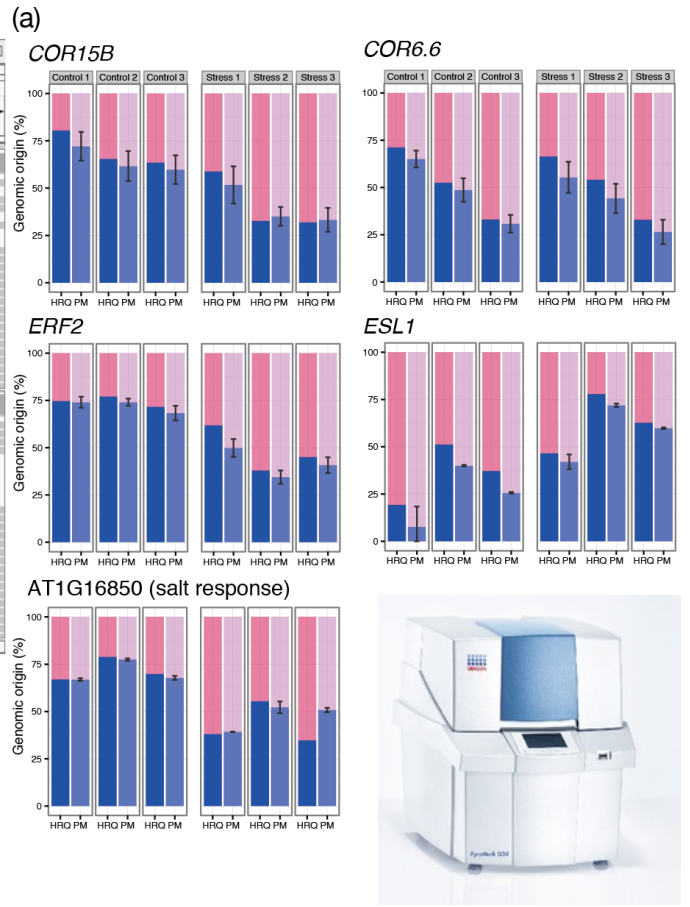


Simulation: 87.8%
of genes if expressed
- 2.44% divergence
- N50 ~ 10 kb

35 million reads
of Illumina paired-end
x 2 conditions
x 3 replicates
x 3 species

Visualization and experimental validation:
Free from potential bias in SNP-based methods

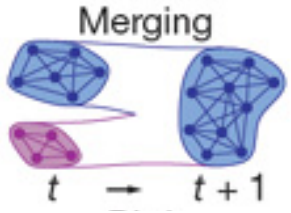




HomeoRoq quantification (left bars)
verified experimentally by PyroMark (right bars)

Akama et al. *Nuc Acids Res* 2014

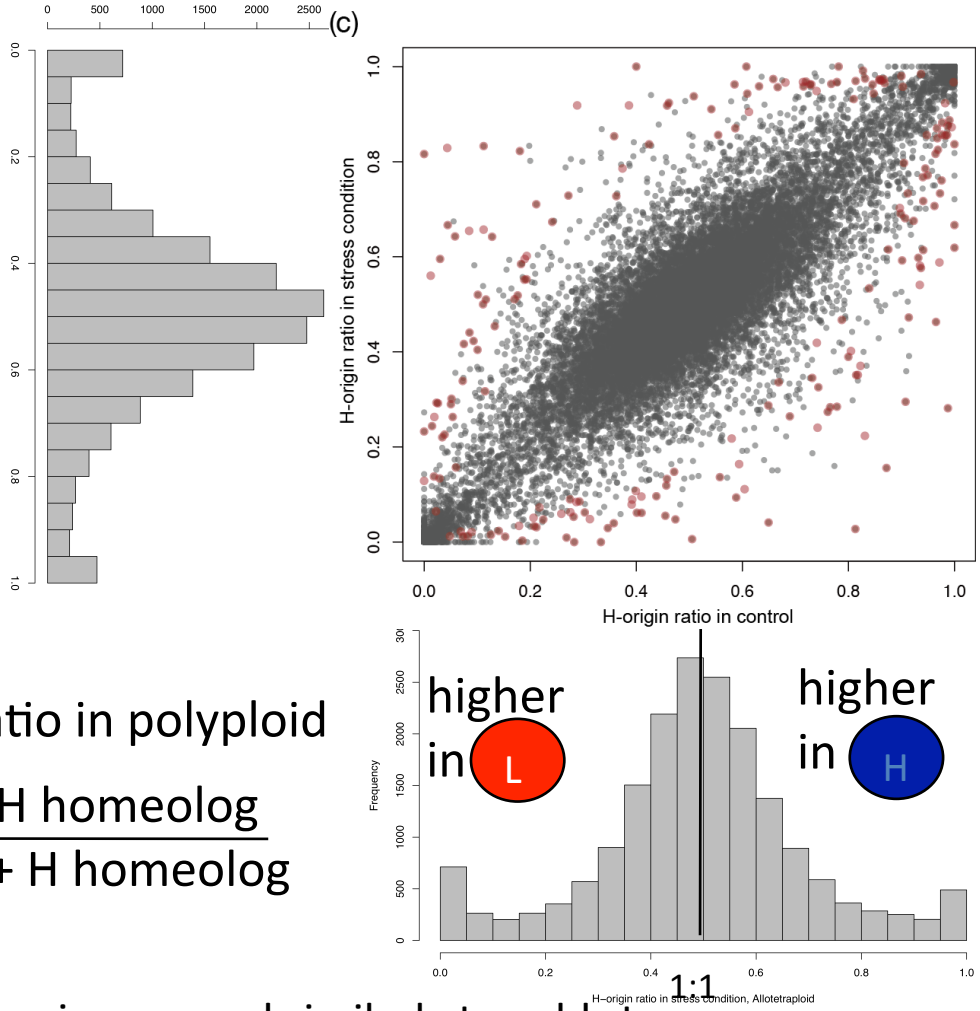
Questions in the gene expression in polyploids



Q1. Are changes in homeolog ratio in different conditions stochastic or regulated?
Q2. How can we find genes with significant ratio change?
Q3. Can parent-specific responses be maintained in polyploids to confer a broader niche?
Q4. Is the pattern of gene silencing common in natural and domesticated crop species?

Q1. Are changes in homeolog ratio in different conditions stochastic or regulated ?

after cold treatment (4˚C 7 day)



high correlation of the ratio (r=0.87)

H-ratio in polyploid

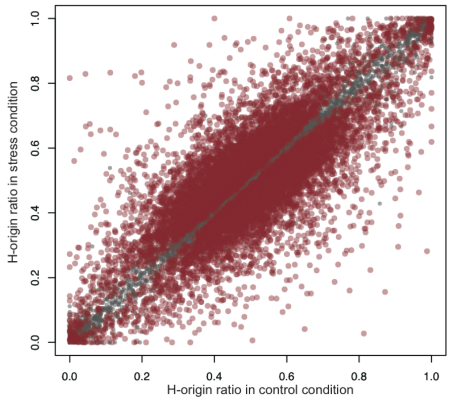
$$= \frac{\text{H homeolog}}{\text{L+ H homeolog}}$$

before cold treatment

Most duplicated pairs respond similarly to cold stress:
highlights non-stochastic regulation in polyploids

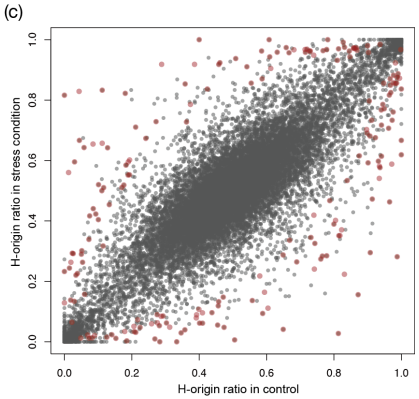
Q2. How to find genes with significant ratio change? – overdispersion

Standard Fisher's exact test identified too many genes (36%) with significant ratio change



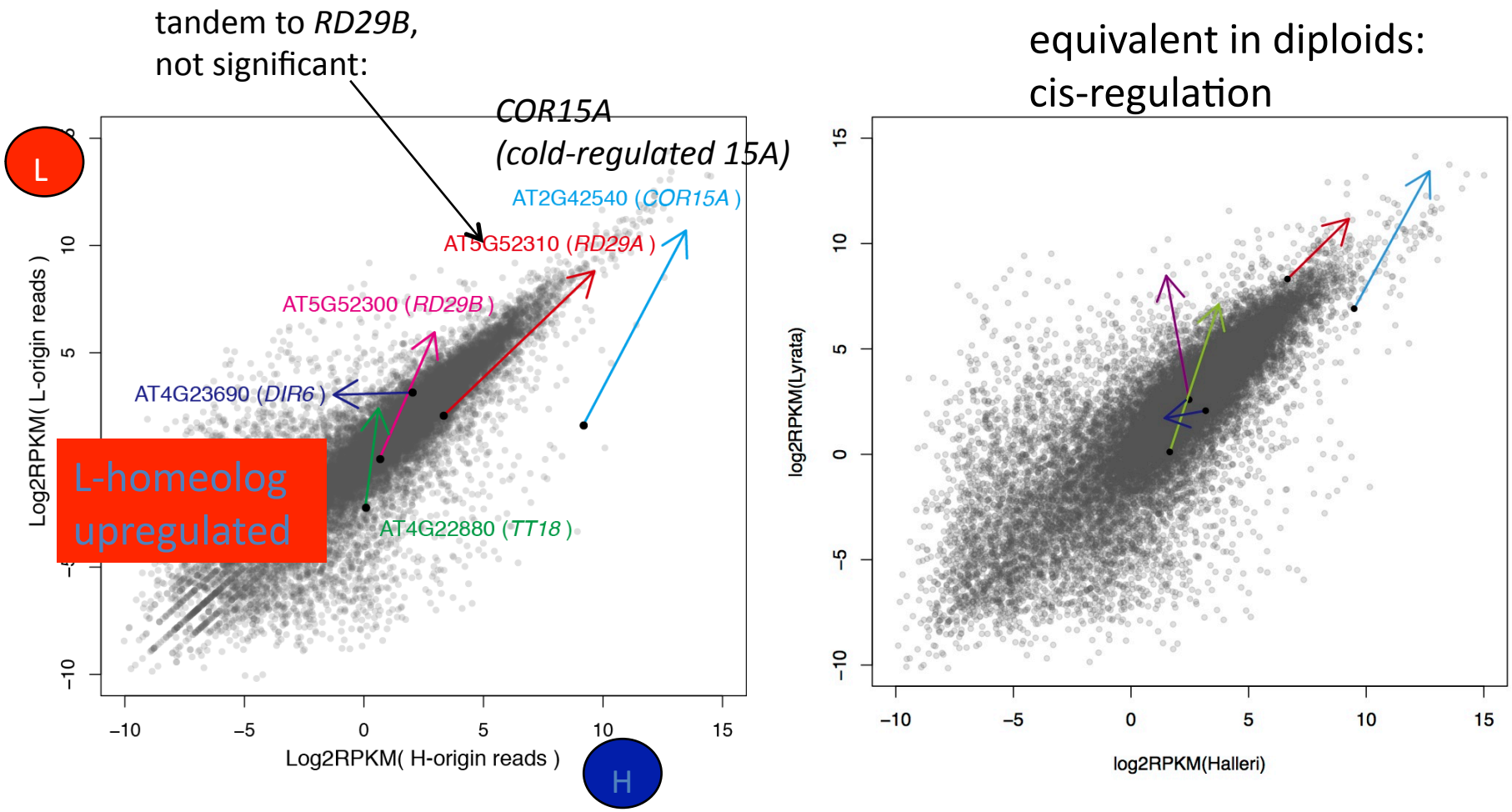
Overdispersion: variance is greater than expected from a simple Poisson distribution (incorporated in edgeR or DESeq, level change in ~10%)

New test incorporating overdispersion identified
226 genes
(1.11% of 20,369 expressed homeologs)



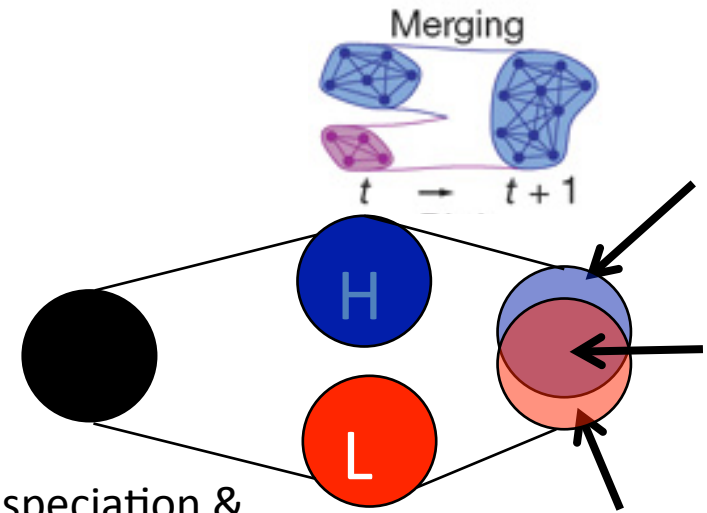
Akama et al. *Nuc Acids Res* 2014

Q3. The genes with ratio changes enriched with stress response genes; cis difference in polyploid



Akama et al. *Nuc Acids Res* 2014 and unpublished

Broader niche by network merging?



speciation &
cold adaptation

Heavy metal response specific to *A. halleri*?

99% of the 20,369 expressed homeologous pairs
-> non-stochastic regulation of ratio
(Coffee, Combes et al. *New Phytol* 2013)

Cold response specific to *A. lyrata*
(cis-regulation of *RD29B, COR15*, etc)

- Hypothesis: most are redundant, while parent-specific stress response networks confer immediate advantage of wider environmental niches