Introduction to Machine Learning for Genomics

Research examples

Director and Professor
Department of Evolutionary Biology and Environmental Studies
University of Zurich
Kentaro Shimizu
kentaro.shimizu@uzh.ch

Two published examples of genomic researches using machine learning

- Genome-wide association studies and genomic prediction of biodiversity effects
  - LASSO regression both for choosing interpretable variables and for prediction

- Hardware and software PlantServation for time-course image analysis of polyploid species in field
  - Deep learning, random forest

Examples of NGS analysis

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

NGS analysis: assembly and mapping

Assembly

DNA: genome assembly

Keyword: contig, coverage

Mapping to a reference genome/transcriptome

DNA: polymorphism analysis and genome-wide association studies

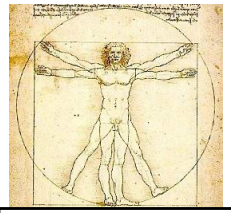
DNA methylation, ChIP-seq

RNA: RNA-seq

28 Oct 2010: 179 human genomes were reported. How are they meaningful?

Why do genome-wide polymorphisms matter?

Projects for medical, agricultural, ecological & evolutionary biology

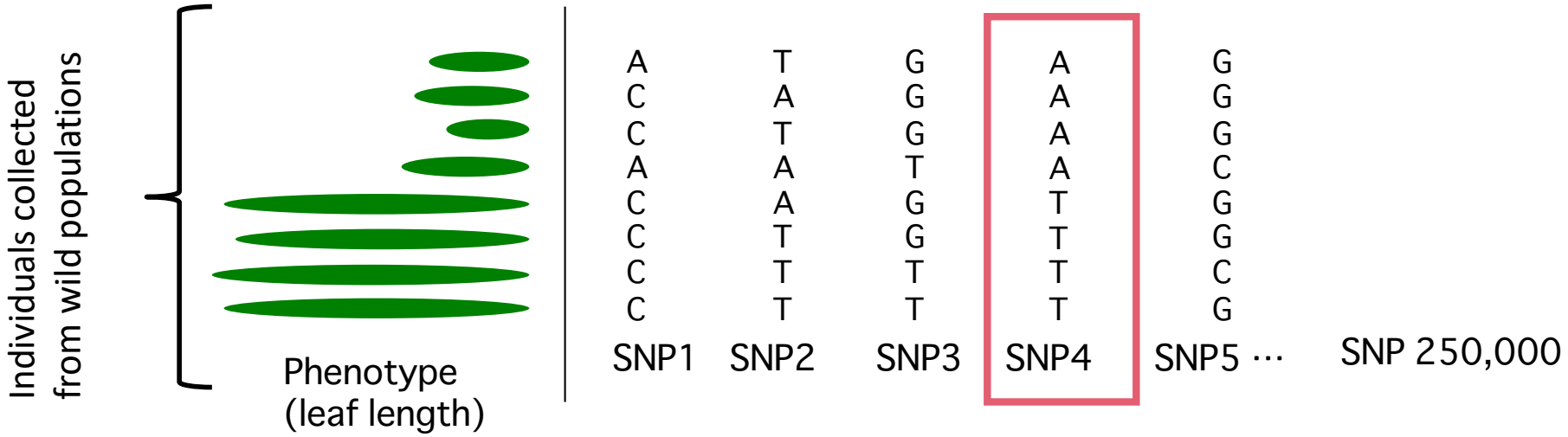


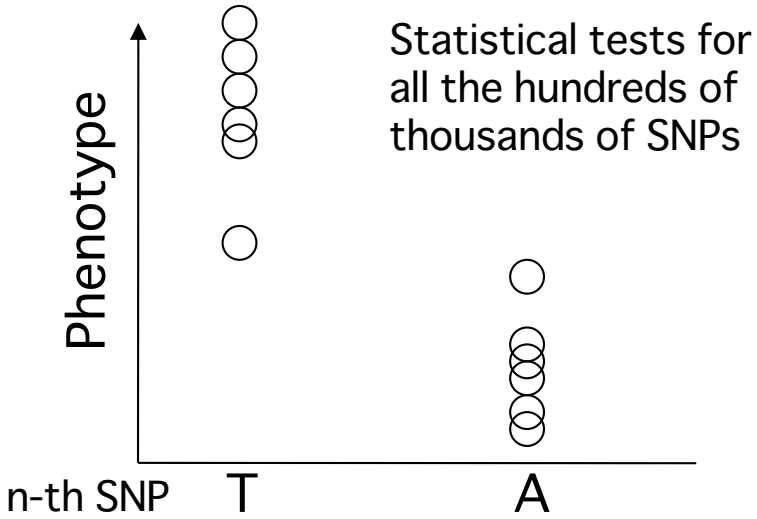

	Homo sapiens	*Drosophila melanogaster*	*Arabidopsis thaliana*	*Triticum aestivum* (bread wheat)
Size	3 Gb	180 Mb	130 MB	16 Gb, allohexaploid (6*x*)
genome	2001	2000	AGI, *Nature* 2000	IWGSC, *Science* 2018
SNP (single nucleotide polymorphism)	HapMap *Nature* 2005;2007		96 accessions, 876 loci, Sanger Nordborg et al. *PLoS Biol* 2005 250 k SNPs, 191 accessions Kim et al. *Nat Genet* 2007, Atwell et al. *Nature* 2010	
re-sequencing or de novo	C. Venter J. Watson 1000 genomes	192 genomes DGRP *Nature* 2012	20 genomes, Clark et al., *Science* 2007 1001 genomes (Gan et al. *Nature* 2011, *Cell* 2016)	10+ genomes (de novo) (Walkowiak et al. *Nature* 2020; Shimizu et al. *PCP* 2021)
related species	Chimp Neanderthal	Drosophila 12 Genomes	*A. lyrata* (Hu et al. Nature Genetics 2011), *Cardamine hirsuta* (Gan et al. *Nature Plants* 2016), *A. halleri* (Akama et al. *NAR* 2014; Briskine et al. 2016) Phylogenomics consortium (Novikova et al. *Nature Genet* 2016)	4x wheat (Avni et al. *Science* 2017) and many others

Genome-Wide Association Study

Developed for human disease studies

Looks for the genes responsible for the phenotypic variations based on simple associations between phenotypes and genotypes (e.g. SNPs)

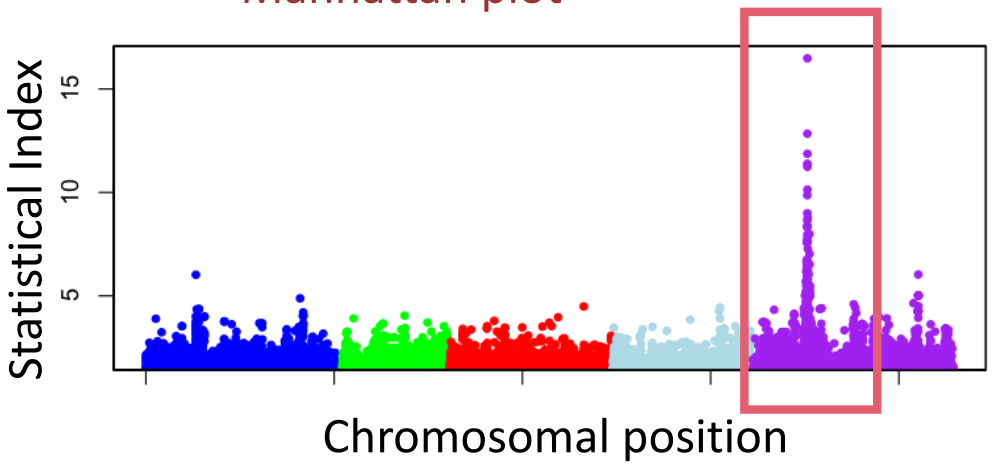




Statistical tests for all the hundreds of thousands of SNPs

Closely linked to a causal gene?

Manhattan plot



GWAS important for medical studies

Weedon et al. Nature Genet 39:1245, 2007

Human height is a classic, highly heritable quantitative trait. To begin to identify genetic variants influencing height, we examined genome-wide association data from 4,921 individuals. Common variants in the *HMGA2* oncogene, exemplified by rs1042725, were associated with height ($P = 4\ 10\text{-}8$). *HMGA2* is also a strong biological candidate for height, as rare, severe mutations in this gene alter body size in mice and humans, so we tested rs1042725 in additional samples. We confirmed the association in 19,064 adults from four further studies ($P = 3\ 10\text{-}11$, overall $P = 4\ 10\text{-}16$, including the genome-wide association data). We also observed the association in children ($P = 1\ 10\text{-}6$, $N = 6{,}827$) and a tall/short case-control study ($P = 4\ 10\text{-}6$, $N = 3{,}207$). We estimate that rs1042725 explains 0.3% of population variation in height (0.4 cm increased adult height per C allele). There are few examples of common genetic variants reproducibly associated with human quantitativetraits; these results represent, to our knowledge, the first consistently replicated association with adult and childhood height.

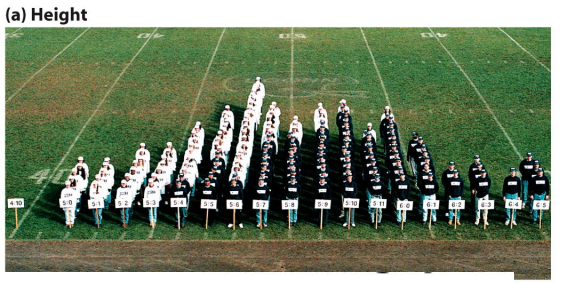


Figure 9-1 Evolutionary Analysis, 4/e
© 2007 Pearson Prentice Hall, Inc.

Wood et al. Nature Genet 46:1173, 2014

Defining the role of common variation in the genomic and biological architecture of adult human height

Using genome-wide data from 253,288 individuals, we identified 697 variants at genome-wide significance that together explained one-fifth of the heritability for adult height. By testing different numbers of variants in independent studies, we show that the most strongly associated ~2,000, ~3,700 and ~9,500 SNPs explained ~21%, ~24% and ~29% of phenotypic variance. Furthermore, all common variants together captured 60% of heritability. The 697 variants clustered in 423 loci were enriched for genes, pathways and tissue types known to be involved in growth and together implicated genes and pathways not highlighted in earlier efforts, such as signaling by fibroblast growth factors, WNT/β-catenin and chondroitin sulfate–related genes. We identified several genes and pathways not previously connected with human skeletal growth, including mTOR, osteoglycin and binding of hyaluronic acid. Our results indicate a genetic architecture for human height that is characterized by a very large but finite number (thousands) of causal variants.

HMGA2 gene:
mutation in mouse results in *pygmy*:

Vol 447|28 June 2007|doi:10.1038/nature05887 nature

ARTICLES

Genome-wide association study identifies novel breast cancer susceptibility loci

Douglas F. Easton[1], Karen A. Pooley[2], Alison M. Dunning[2], Paul D. P. Pharoah[2], Deborah Thompson[1], Dennis G. Ballinger[3], Jeffery P. Struewing[4], Jonathan Morrison[2], Helen Field[2], Robert Luben[5], Nicholas Wareham[5], Shahana Ahmed[2], Catherine S. Healey[2], Richard Bowman[6], the SEARCH collaborators[2*], Kerstin B. Meyer[2], Christopher A. Haiman[8], Laurence K. Kolonel[9], Brian E. Henderson[8], Loic Le Marchand[9], Paul Brennan[10], Suleeporn Sangrajrang[11], Valerie Gaborieau[10], Fabrice Odefrey[10], Chen-Yang Shen[12], Pei-Ei Wu[12], Hui-Chun Wang[12], Diana Eccles[13], D. Gareth Evans[14], Julian Peto[15], Olivia Fletcher[16], Nichola Johnson[16], Sheila Seal[17], Michael R. Stratton[17,18], Nazneen Rahman[17], Georgia Chenevix-Trench[19], Stig E. Bojesen[20], Børge G. Nordestgaard[20], Christen K. Axelsson[21], Montserrat Garcia-Closas[22], Louise Brinton[22], Stephen Chanock[23], Jolanta Lissowska[24], Beata Peplonska[25], Heli Nevanlinna[26], Rainer Fagerholm[26], Hannaleena Eerola[26,27], Daehee Kang[28], Keun-Young Yoo[28,29], Dong-Young Noh[28], Sei-Hyun Ahn[30], David J. Hunter[31,32], Susan E. Hankinson[32], David G. Cox[32], Per Hall[33], Sara Wedren[33], Jianjun Liu[34], Yen-Ling Low[34], Natalia Bogdanova[35,36], Peter Schürmann[36], Thilo Dörk[36], Rob A. E. M. Tollenaar[37], Catharina E. Jacobi[38], Peter Devilee[39], Jan G. M. Klijn[40], Alice J. Sigurdson[41], Michele M. Doody[41], Bruce H. Alexander[42], Jinghui Zhang[4], Angela Cox[43], Ian W. Brock[43], Gordon MacPherson[43], Malcolm W. R. Reed[44], Fergus J. Couch[45], Ellen L. Goode[45], Janet E. Olson[45], Hanne Meijers-Heijboer[46,47], Ans van den Ouweland[47], André Uitterlinden[48], Fernando Rivadeneira[48], Roger L. Milne[49], Gloria Ribas[49], Anna Gonzalez-Neira[49], Javier Benitez[49], John L. Hopper[50], Margaret McCredie[51], Melissa Southey[50], Graham G. Giles[52], Chris Schroen[53], Christina Justenhoven[54], Hiltrud Brauch[54], Ute Hamann[55], Yon-Dschun Ko[56], Amanda B. Spurdle[19], Jonathan Beesley[19], Xiaoqing Chen[19], kConFab[57*], AOCS Management Group[19,57*], Arto Mannermaa[58,59], Veli-Matti Kosma[58,59], Vesa Kataja[58,60], Jaana Hartikainen[58,59], Nicholas E. Day[5], David R. Cox[3] & Bruce A. J. Ponder[2,7]

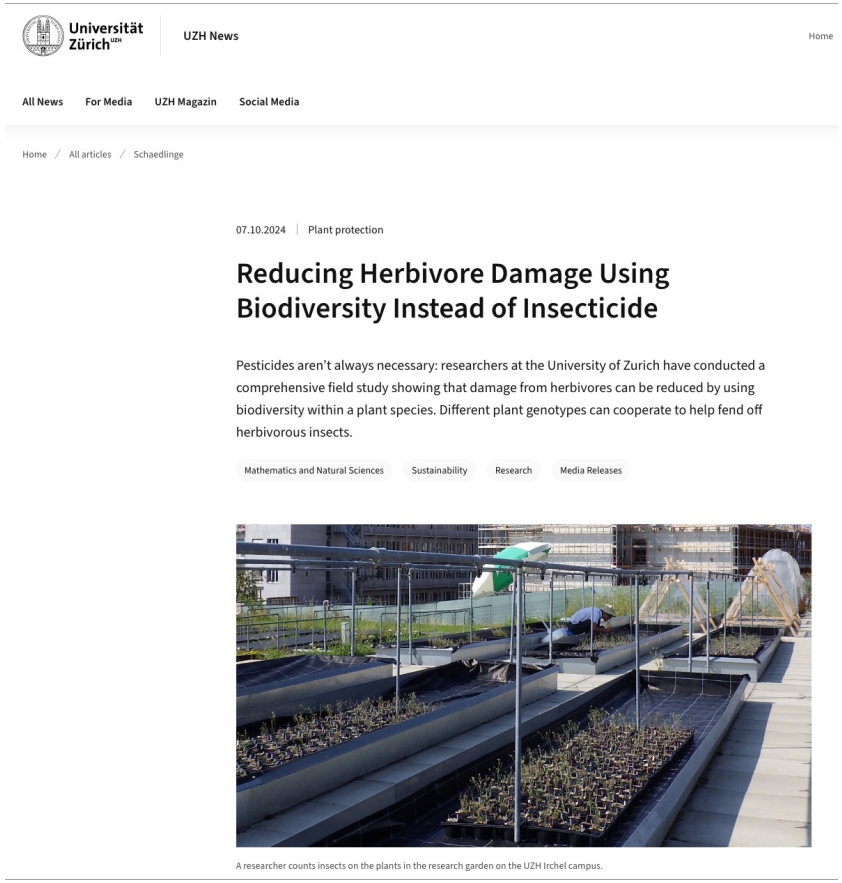
Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels

Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes for BioMedical Research*†

Science **316**, 1331 (2007)

Genome-wide association studies and genomic prediction of biodiversity effect using machine learning

Sato, Shimizu-Inatsugi, Takeda, Schmid, Nagano, Shimizu
Nature Communications 15: 8467, 2024



Universität Zürich[UZH] | UZH News

Home

All News | For Media | UZH Magazin | Social Media

Home / All articles / Schaedlinge

07.10.2024 | Plant protection

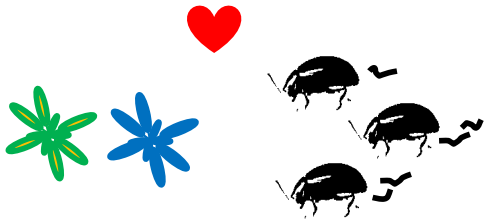
Reducing Herbivore Damage Using Biodiversity Instead of Insecticide

Pesticides aren't always necessary: researchers at the University of Zurich have conducted a comprehensive field study showing that damage from herbivores can be reduced by using biodiversity within a plant species. Different plant genotypes can cooperate to help fend off herbivorous insects.

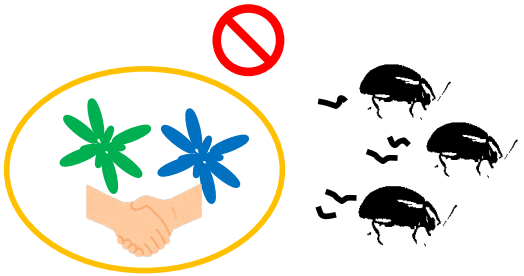
Mathematics and Natural Sciences | Sustainability | Research | Media Releases

Similar to genomic prediction:
Polygenic Risk Score (PRS) in medicine

Health of animals and plants depends on neighbors



Associational susceptibility



Associational resistance
(positive biodiversity effect)

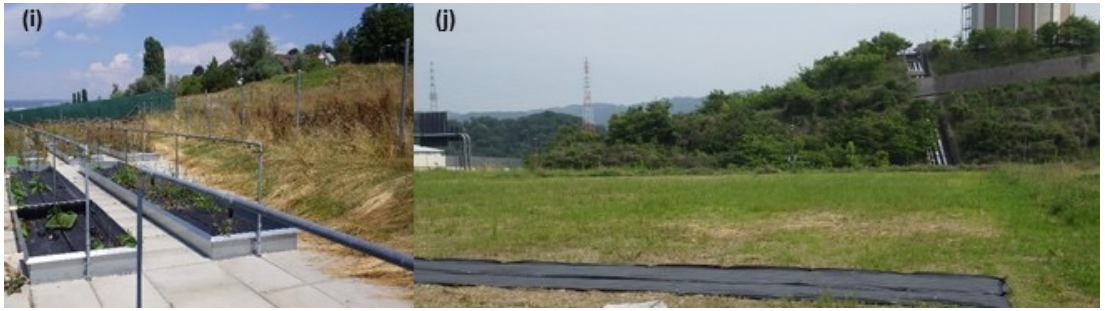
Having right or wrong neighbors?
Mixed stands are less or more damaged by herbivores (Jactel et al. 2021)

Arabidopsis interact with other plants and natural enemies in the field

Lab (solitary)

Outdoor (group)

Pest insect community on *Arabidopsis thaliana* in Irchel



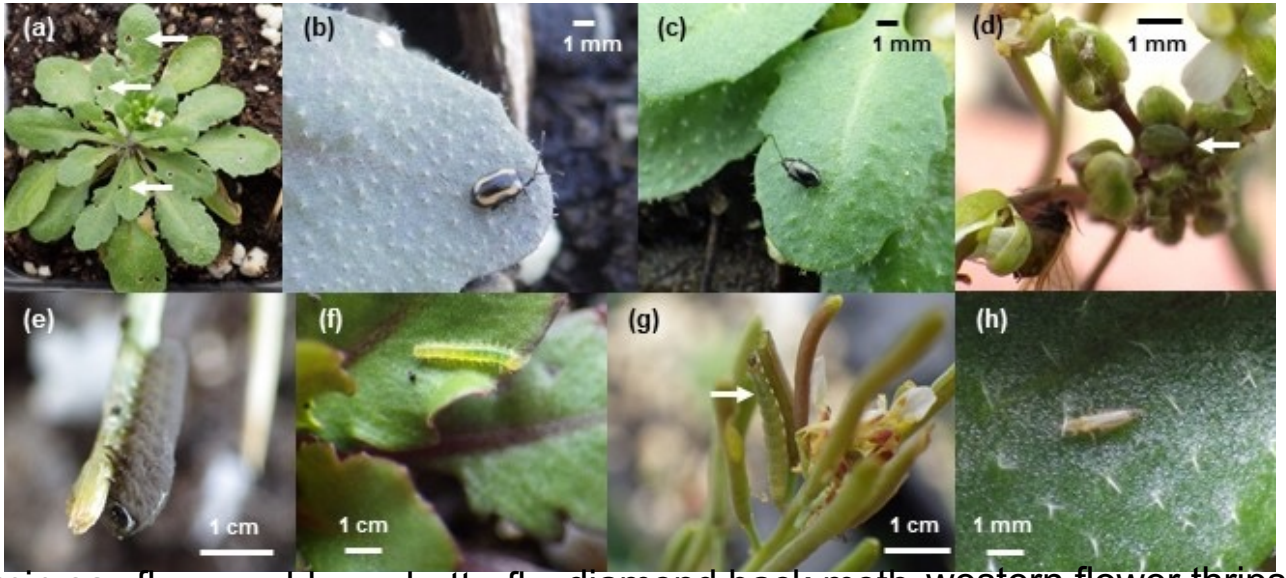
Zurich (natural range) Shiga, Japan (naturalized range)

leaf holes made by flea beetles

a striped flea beetle *Phyllotreta striolata*

turnip flea beetle *Phyllotreta atra*

mustard aphids *Lipaphis erysimi*



turnip sawfly *Athalia rosae*

cabbage butterfly *Pieris rapae*

diamond back moth *Plutella xylostella*

western flower thrips *Frankliniella occidentalis*

Sato et al., *BMC Plant Biol* 2019

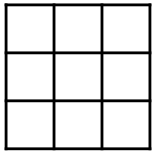
Aim: Genomic prediction of positive plant-plant interactions

Arabidopsis thaliana

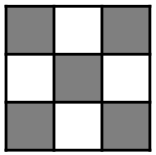
Randomized mixture

Insect herbivory

(1) Picking up genotypes pairs that gain associational resistance



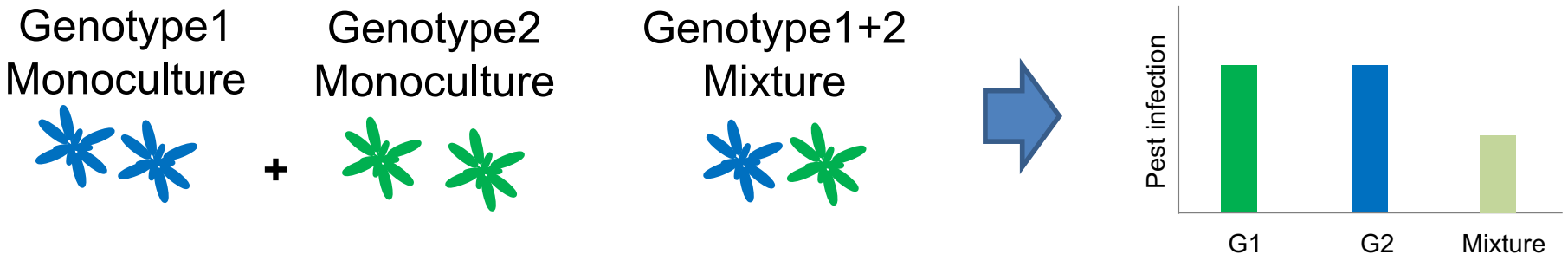
Monoculture vs. Mixture



(2) Verifying key pairs with associational resistance

Sato et al. *Nature Communications* 15: 8467, 2024

When does genetic diversity matter? Finding beneficial pairs for mixed planting (Wuest et al. 2021)



$_{200}C_2 = \frac{200*199}{2} =$ **19900** *pairs from 200 genotypes…*

Q. How can we find better pairs out of many possible pairs?

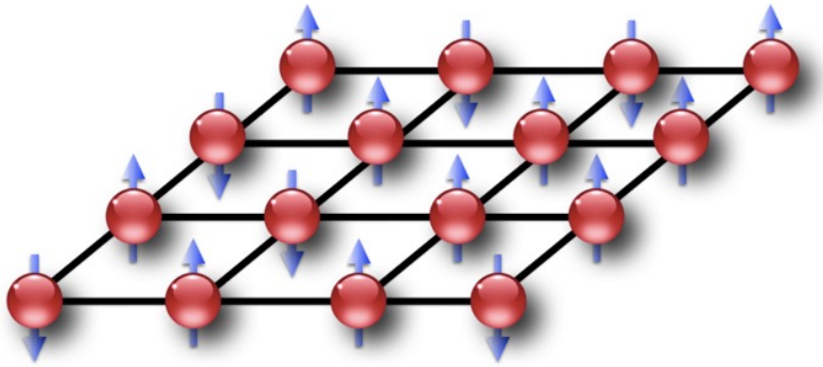
Sato et al. *Nature Communications* 15: 8467, 2024

Ising model to study neighbor interactions

Ernst Ising (1900-1998)

Taroni, *Nature Physics* 2015

mathematical model of ferromagnetism
in statistical mechanics



https://nedo-quantum.aist.go.jp/ising-machine.html

New method Neighbor GWAS: Detecting neighbor interactions in randomized mixtures of genotypes

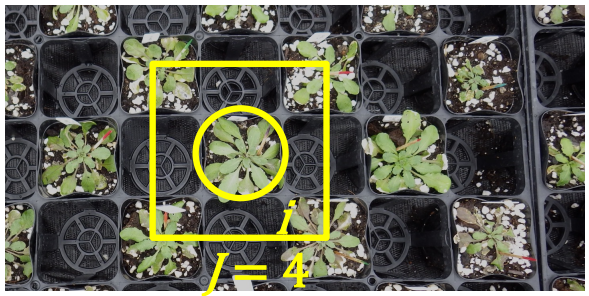
Normal GWAS: **Phenotype = Genotype + Environment**

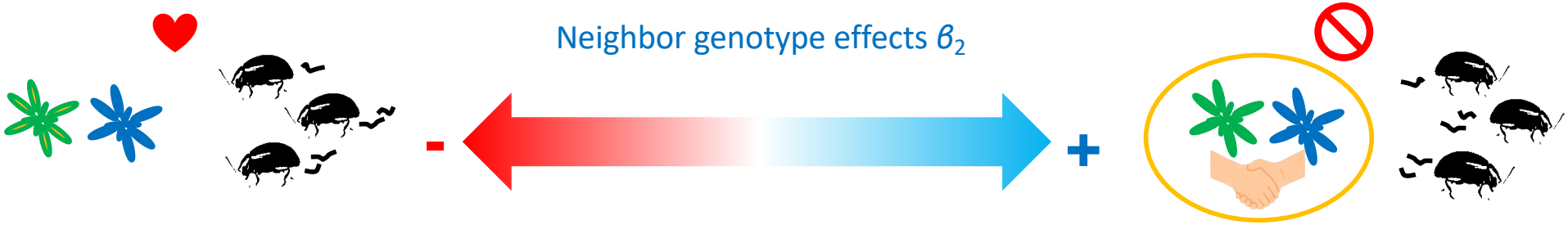
Neighbor GWAS: **P = G + E + G×E (=G×NeighborG)** (Sato et al. 2021 *Heredity*)

$$y_i = \beta_0 + \boxed{\beta_1 g_i} + e_i + \boxed{\beta_2 \frac{\sum_{j=1}^{J} g_i g_j}{J}}$$

focal G

G×NeighborG

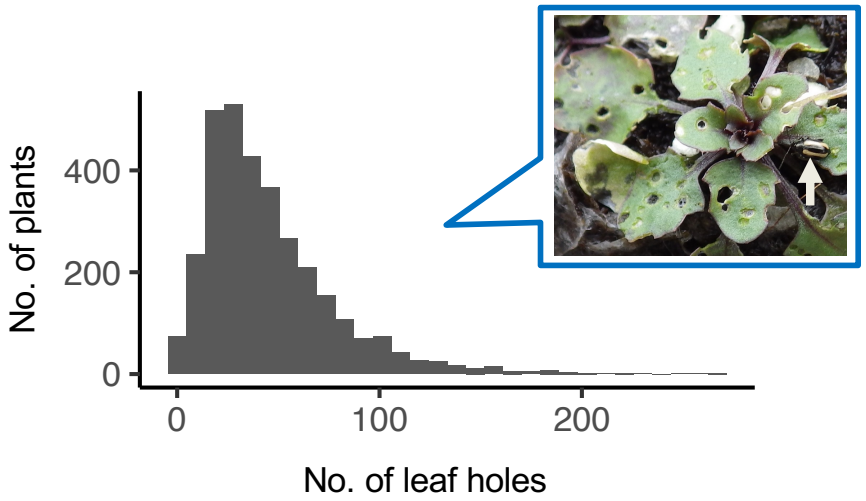




Sato et al. *Nature Communications* 15: 8467, 2024

Large dataset by field study in Irchel

During July at Zurich, Switzerland

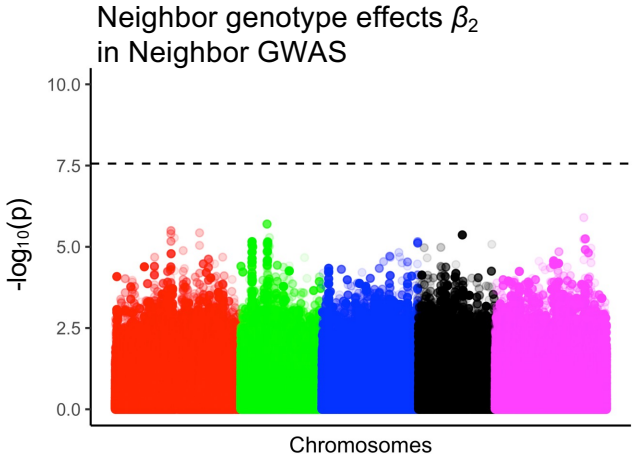


- 200 genotypes randomly assigned within a block
× 8 blocks × 7 survey × 2 years × 2 sites = **44,800** data points

Observation of leaf damage, and 52,007 insects

Sato et al. *Nature Communications* 15: 8467, 2024

Genome-wide association studies (GWAS): no significant peak suggesting polygenic nature



- Neighbor effects are unlikely attributed to a few loci and likely have a polygenic basis

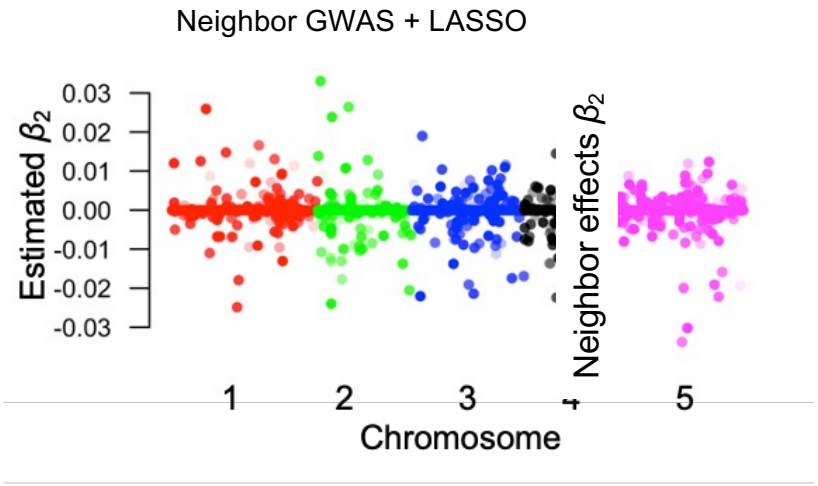
Sato et al. *Nature Communications* 15: 8467, 2024

Machine learning (LASSO regression)

both mechanisms and prediction

$Y \sim a_1 * x_1 + \sim a_2 * x_2 + ... + a_n * x_n$

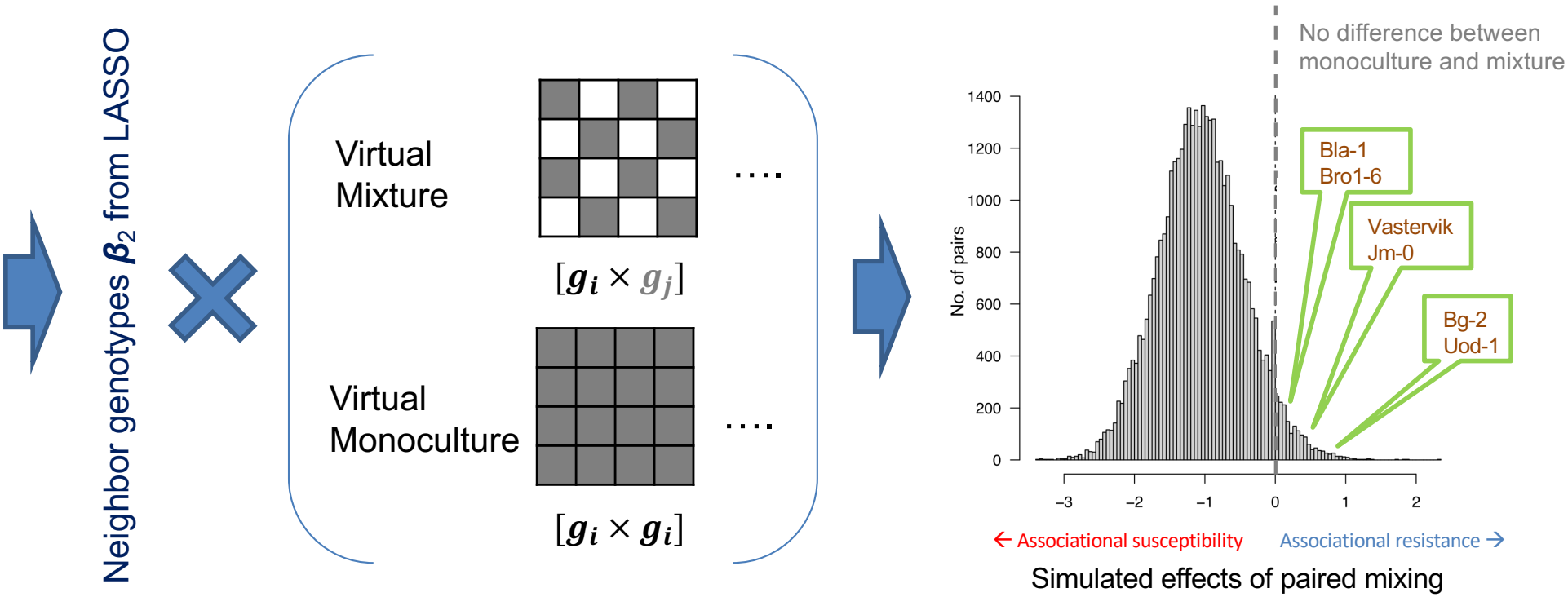
Most of a_i to be zero



- LASSO regression narrowed down from 1.2 million to 756 SNPs

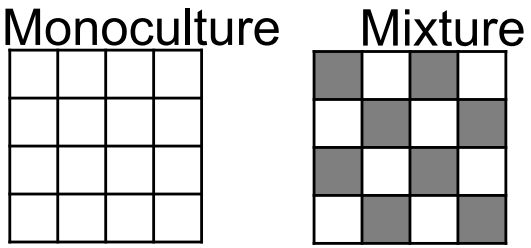
Sato et al. *Nature Communications* 15: 8467, 2024

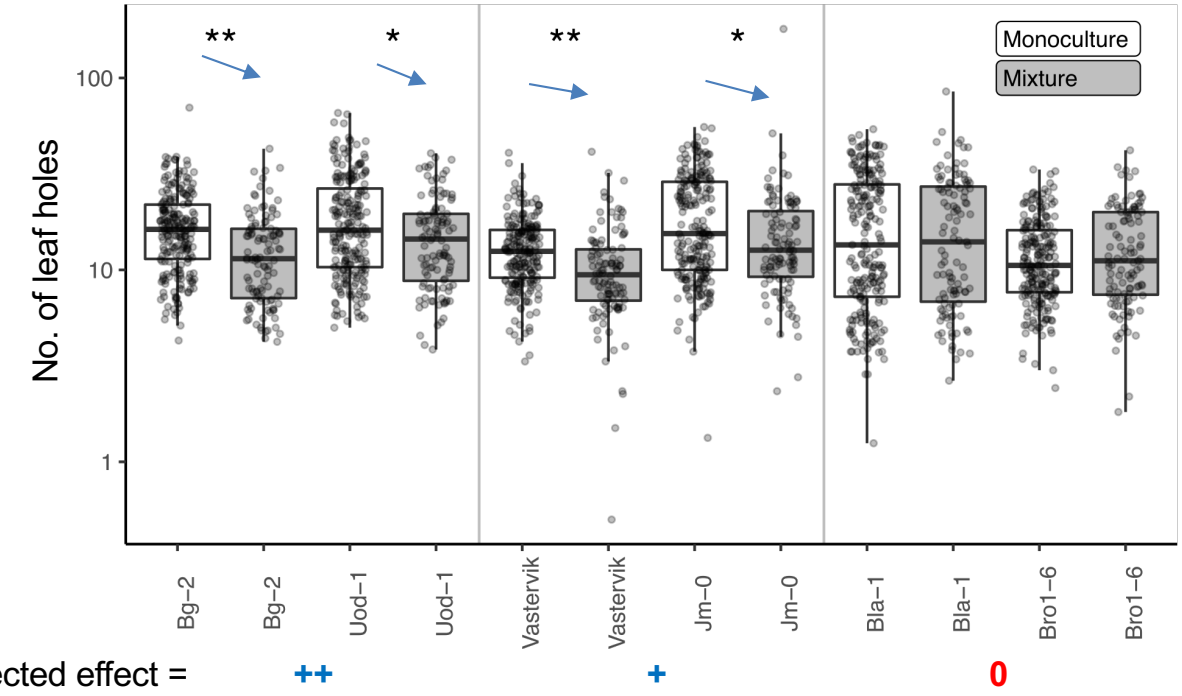
Genomic prediction of herbivore damage under virtual mixture and monoculture



Sato et al. *Nature Communications* 15: 8467, 2024

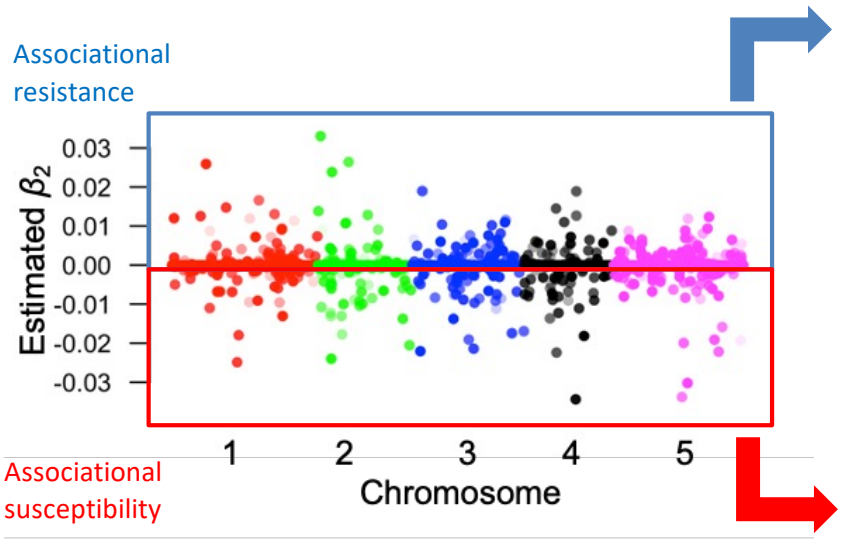
Validating the prediction by LASSO regression: two genotype pairs indeed mitigated herbivory in mixed planting





Sato et al. *Nature Communications* 15: 8467, 2024

Potential mechanism? JA-related GOs are enriched in SNPs associated with associational resistance (β_2>0)



GO:0009695: jasmonic acid biosynthetic process
FDR < 10^{-4}

LIPOXIGENASE 2 (*LOX2*) & *LOX6*

GO:0009695: jasmonic acid biosynthetic process
FDR > 0.05

Sato et al. *Nature Communications* 15: 8467, 2024

Aim: Genomic prediction of positive plant-plant interactions

Arabidopsis thaliana

Randomized mixture

Insect herbivory

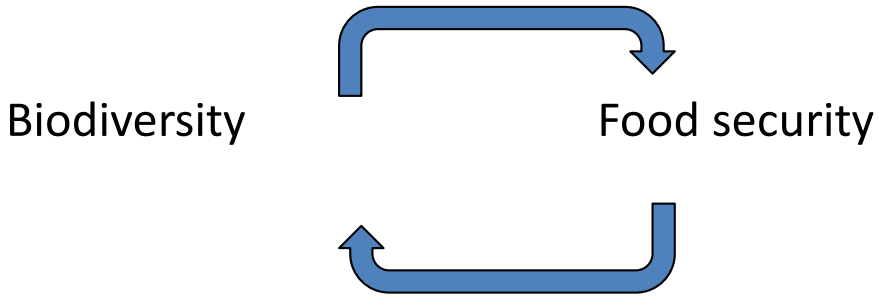
(1) Picking up genotypes pairs that gain associational resistance

Predicting genotype pairs with positive biodiversity effects

(2) Verifying key pairs with associational resistance

Insecticide:
Food security and environment

SUSTAINABLE DEVELOPMENT GOALS



sdgactioncampaign.org

Take-home message

1. Neighbor GWAS + LASSO found key genotype pairs out of randomized cultivation in field-grown plants
→ Any plant genomes in randomized mixtures

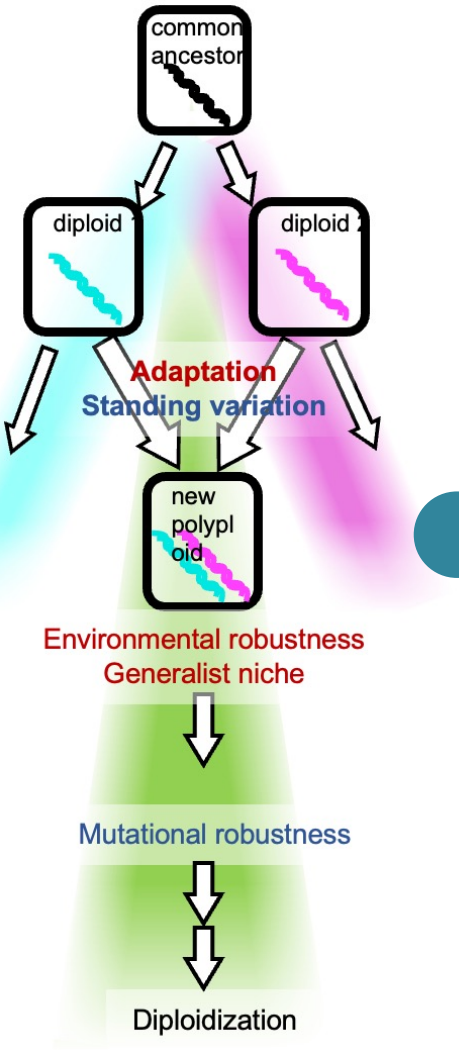
2. This approach is useful even when a target trait is highly polygenic
→ Any quantitative traits of ecological and agricultural interest

Two published examples of genomic researches using machine learning

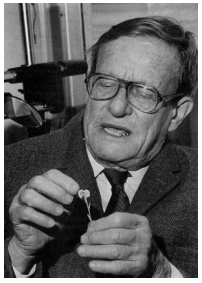
- Genome-wide association studies and genomic prediction of biodiversity effects
  - LASSO regression both for choosing interpretable variables and for prediction

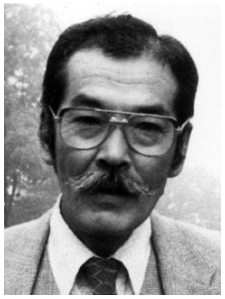
- Hardware and software PlantServation for time-course image analysis of polyploid species in field
  - Deep learning, random forest

Is genome duplication advantageous? Genomics and machine learning



”genome shock”, “gigas” effect





Stebbins (1950, 1971)
Polyploidy may retard adaptation

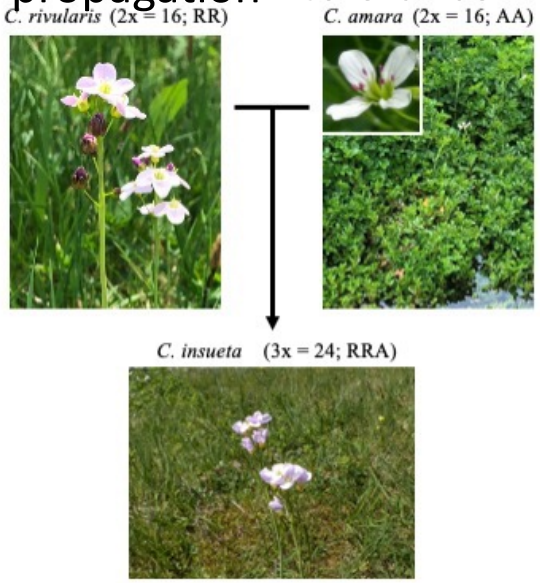
Ohno (1970)
Long-term evolutionary advantage
‘Evolution by gene duplication’

Why can polyploid species adapt to distinct or broader ecological niches?

A textbook example of new species during past 150 years

Clonal propagation

Submergence tolerance



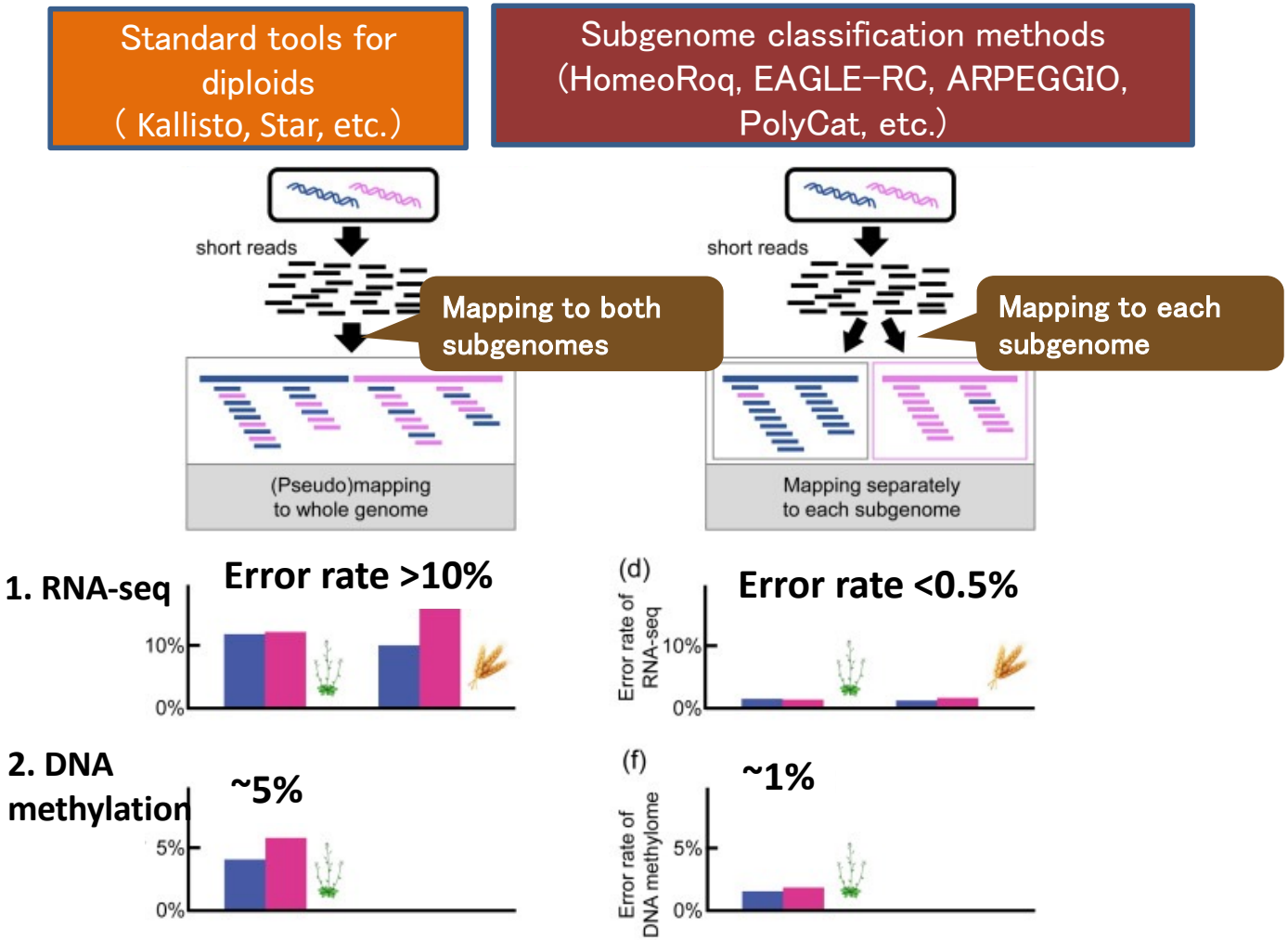
Caradamine insueta

https://www3.nhk.or.jp/nhkworld/en/tv/scienceview/

Jianqiang Sun, Rie Shimizu-Inatsugi, Hugo Hofhuis, **Kentaro Shimizu**, Angela Hay, **Kentaro K. Shimizu**, Jun Sese, *Front Genet*, 2020

Novel bioinformatic workflow for allopolyploid species

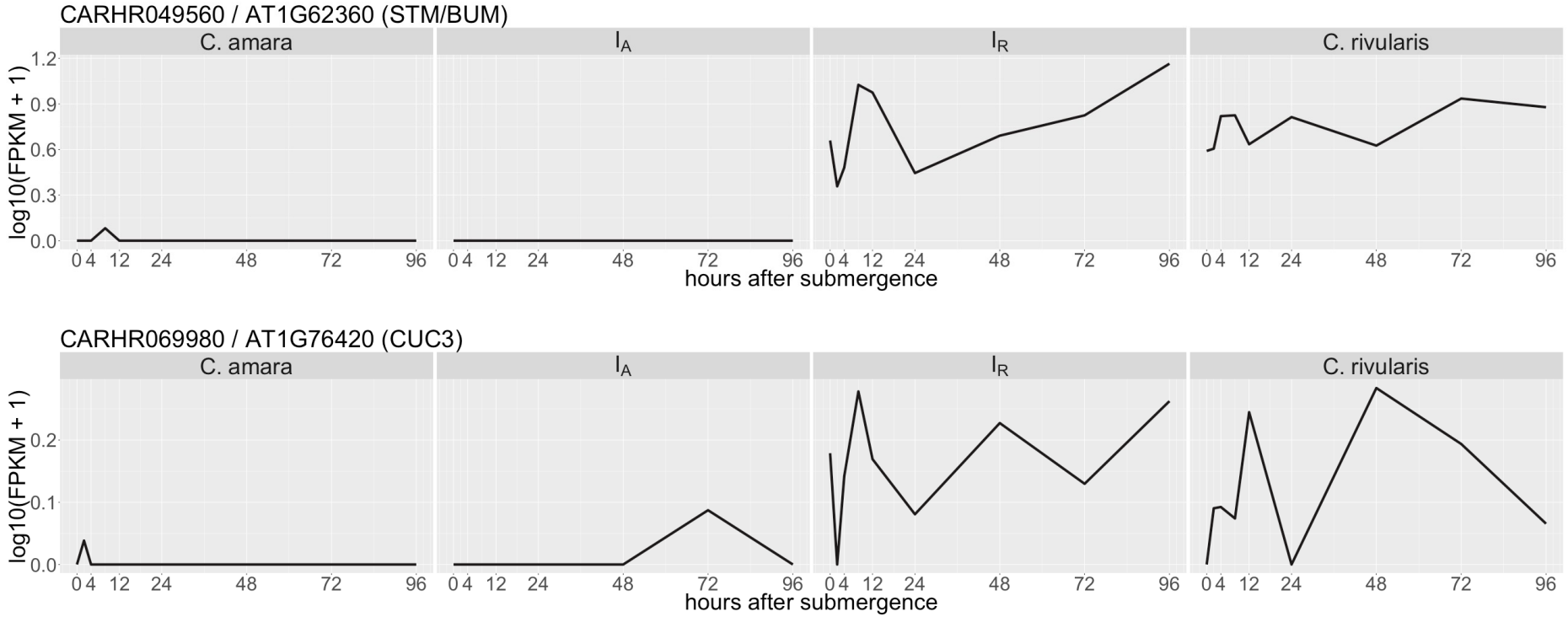
Shimizu, *Current Opinion in Plant Biology,* 69, 102292, 2022

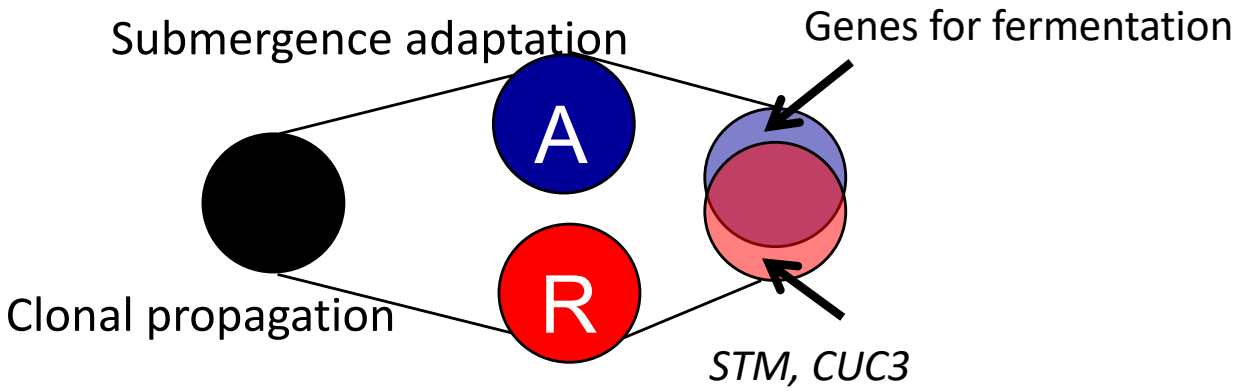


Akama *et al., NAR* 2014, Kuo *et al., Brief Bioinf* 2020, Milosavljevic et al. *BMC Genomics* 2022

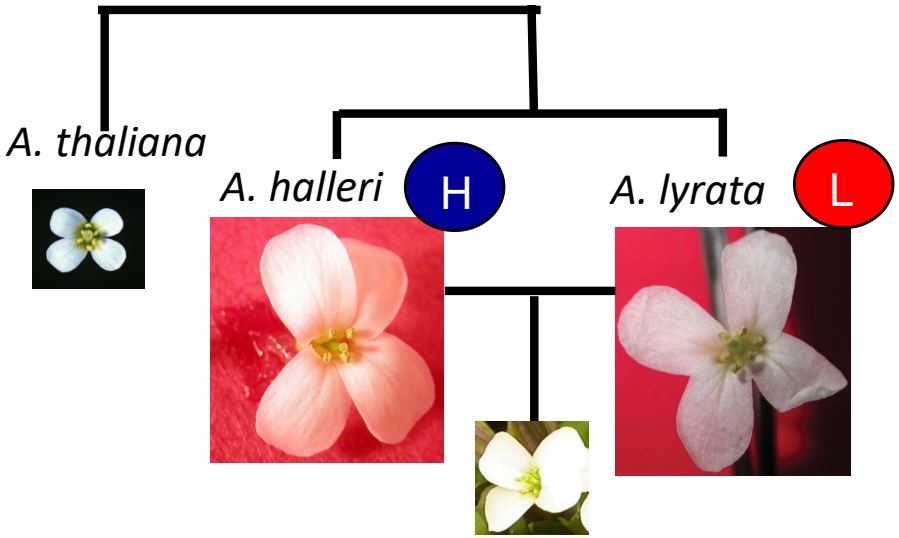
Assembly, Resequencing, RNA-seq, Epigenetics

Expression of key stem cell (meristem) gene was inherited



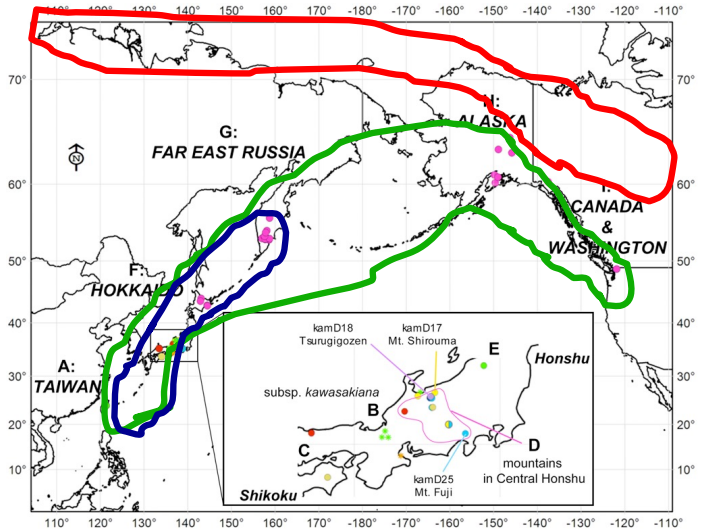


Model polyploid species : the simplest for bioinformatic development: Synthetic and natural *Arabidopsis kamchatica*

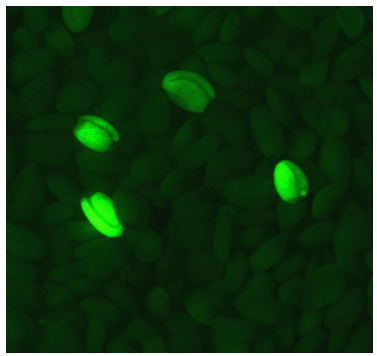


A. kamchatica (Fisch. ex DC.) K. Shimizu & Kudoh 2005

Precipitation and temperature,
Hoffmann, *Evolution* 2005



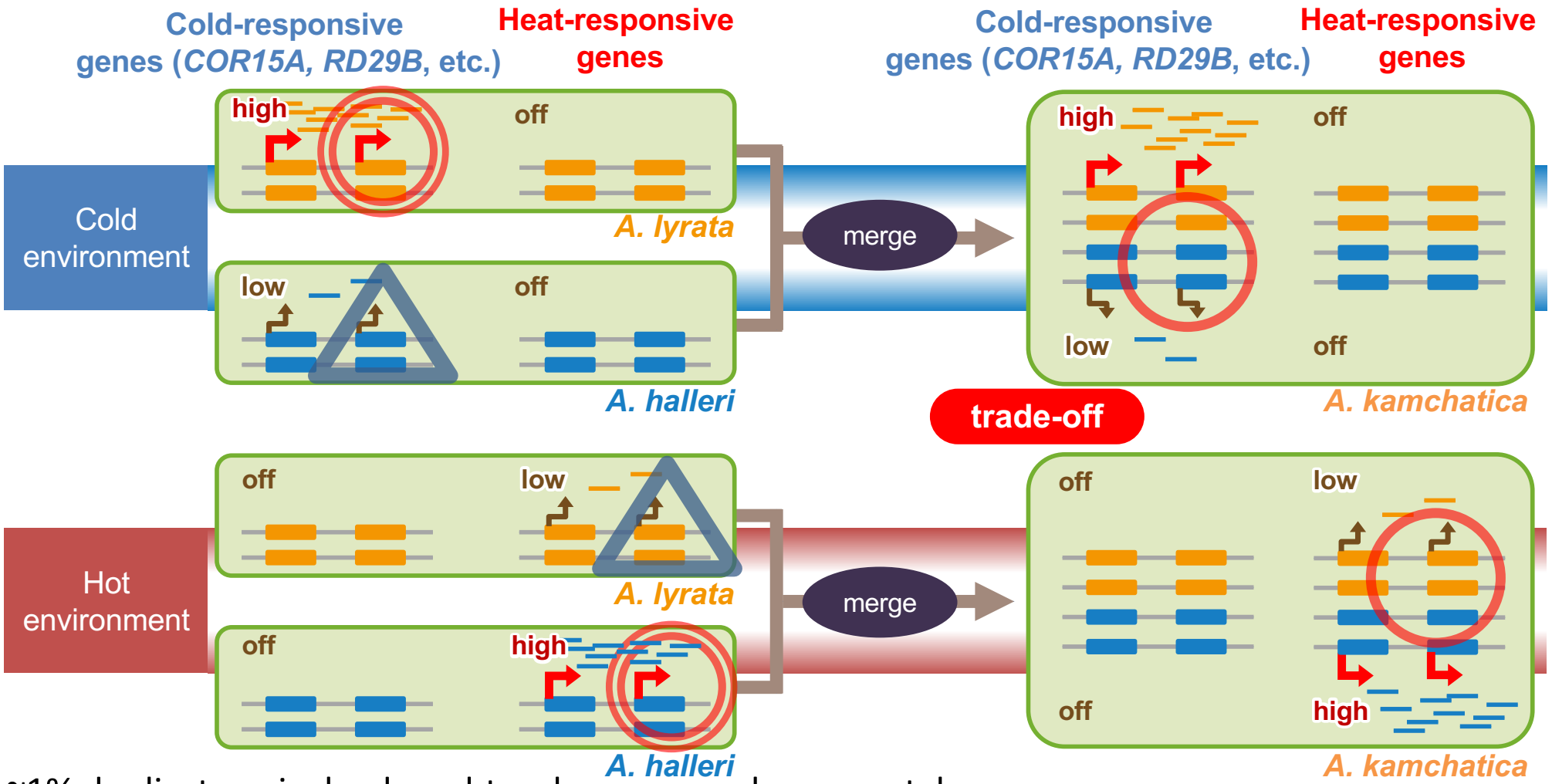
0-3000 m alt in Japan



Transgenic technique
Yew et al. *J Plant Res* 2018

Fujimoto et al. *PLoS Genet* 2008
Shimizu-Inatsugi et al. *Mol Ecol* 2009
Tsuchimatsu et al., *PLoS Genet* 2012

Model of environmental robustness by inheriting and merging parental traits

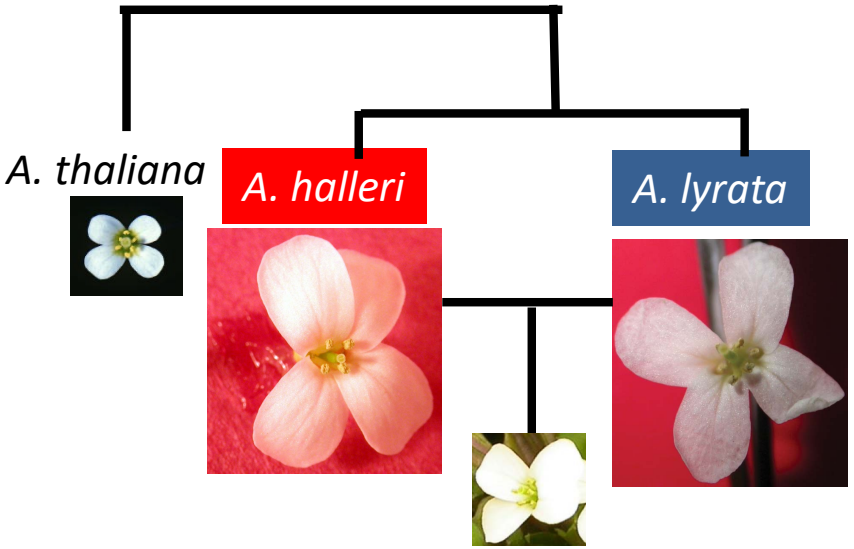


~1% duplicate pairs by drought, submergence, heavy metal

Shimizu, *Current Opinion in Plant Biology,* 69, 102292, 2022

Toward synthetic biology to confer robustness

Phenomics using machine learning
Often combined with genomics

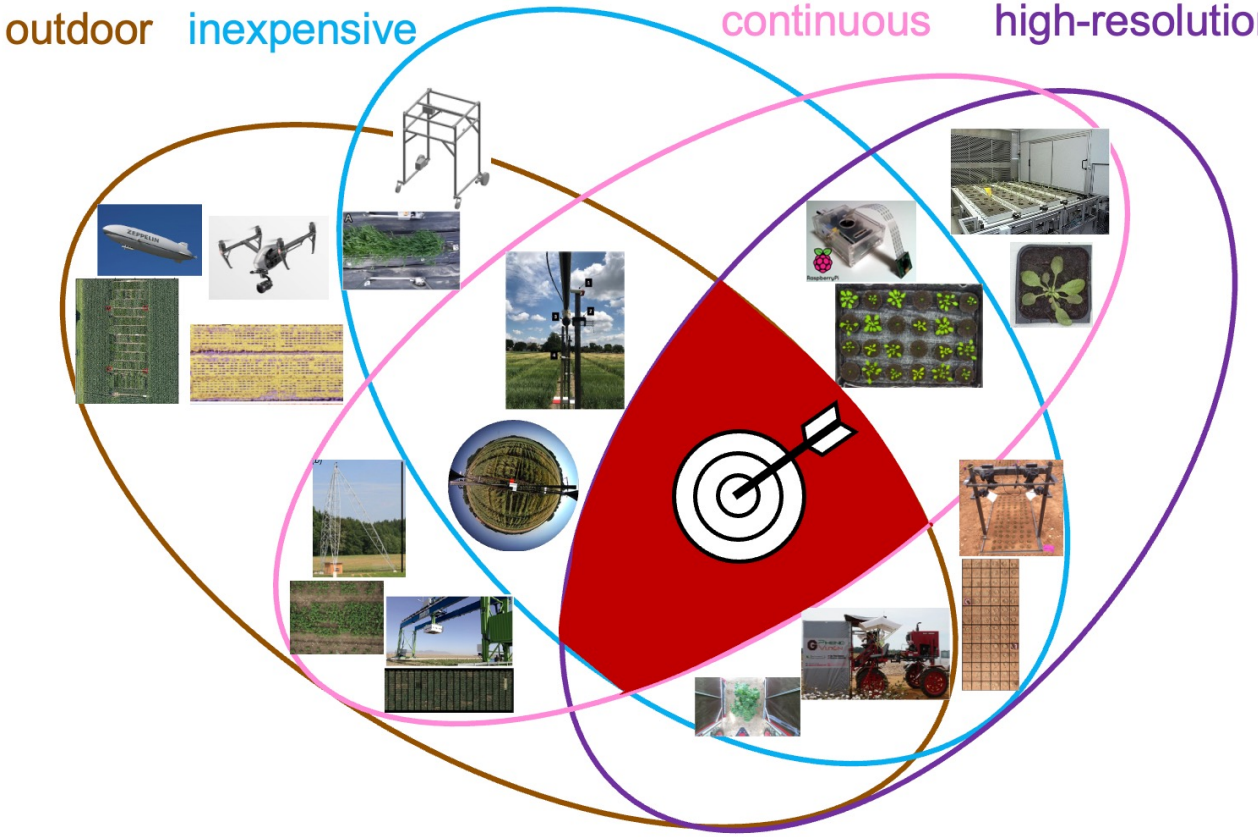


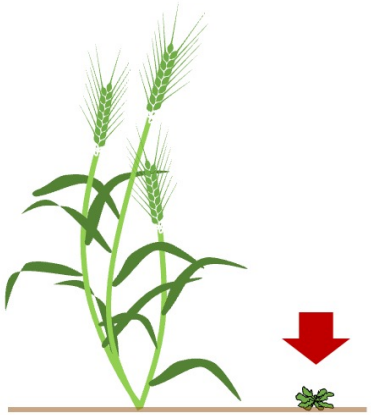
A. kamchatica (Fisch. ex DC.) K. Shimizu & Kudoh

Stress marker of light, temperature, etc.
1. Did synthetic polyploids combine parental responses?
2. Can synthetic polyploids recapitulate natural speciation?

Gap in monitoring hardware

Niche in remote sensing methods for accurate monitoring



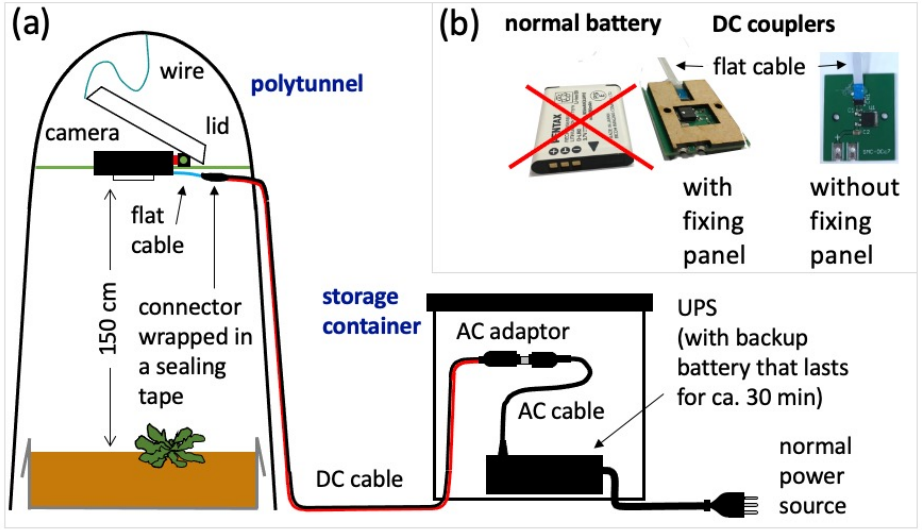


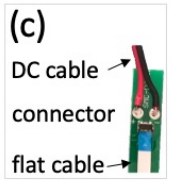
Araus et al. 2018 Trends Plant Sci 23: 451-466
Toda et al. 2021 The Plant Genome 14: e20157
https://www.dji.com/ch/inspire-1
Liebisch et al. 2015 Plant Methods 11: 9
Susko et al. 2018 HardwareX 4: e00029
Kuroki et al. 2022 Breeding Science 72: 66–74
Minervini et al. 2017 Plant J 90: 204-216
Humplík et al. 2015 Plant Methods 11: 29
An et al. 2016 Comp El Agr 127: 376-394
Jiang et al. 2018 Sci Rep 8: 1213
Burnette et al. 2018 Proc Practice Experience Ad Res Comp 27: 1-7
Kirchgessner et al. 2017 Func Plant Biol 44: 154-168

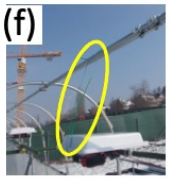
Figure by Reiko Akiyama

New methodology of imaging analysis

PlantServation:
cost-efficient hardware and software







http://www.nation.co.ke

Reiko Akiyama

Takao Goto

Toshiaki Tameshige

Rie Shimizu-Inatsugi

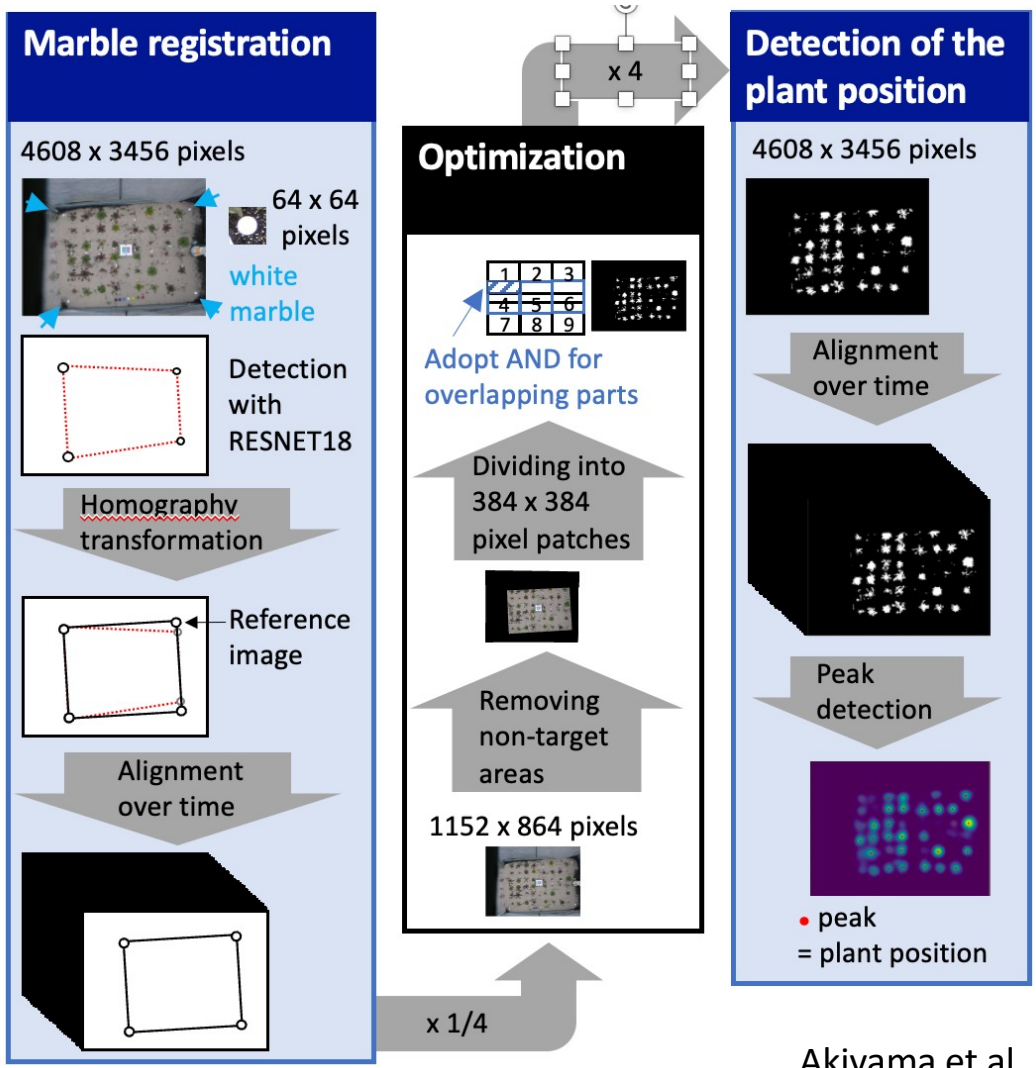
Akiyama et al., *Nature Commun* 14:5792, 2023

~2,600 USD for 384 individuals

Suitable for researchers and small-scale cultivation

Difficulties *in natura*: movement of camera

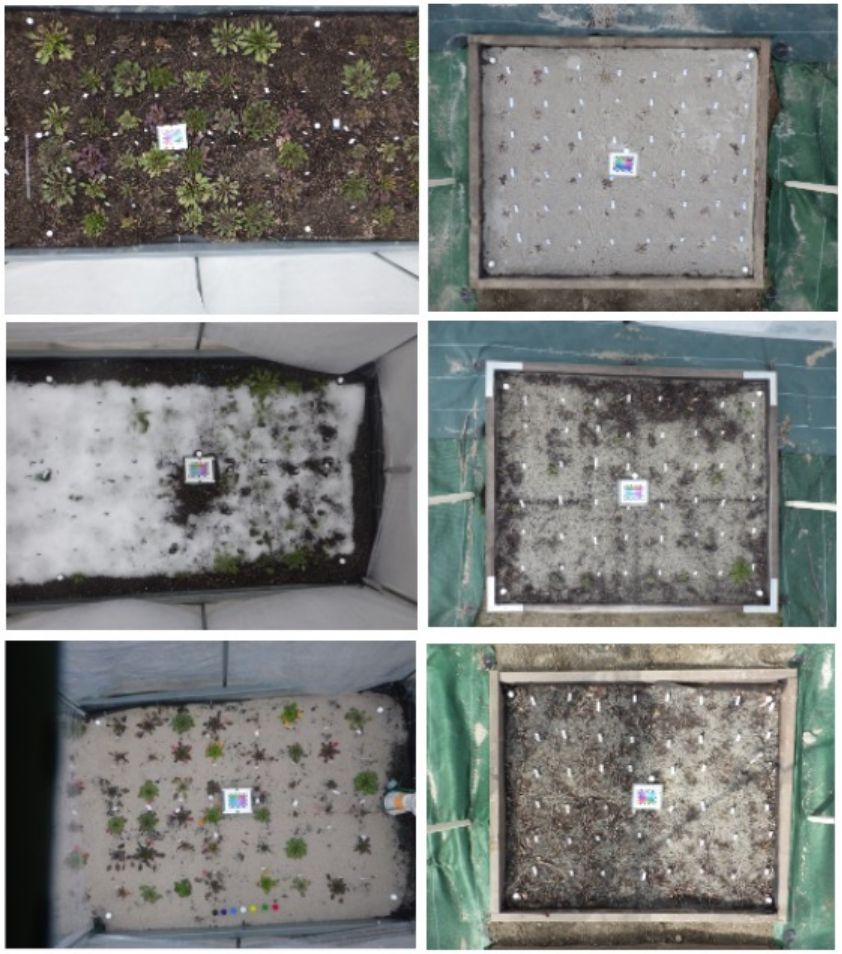
Wind, etc.

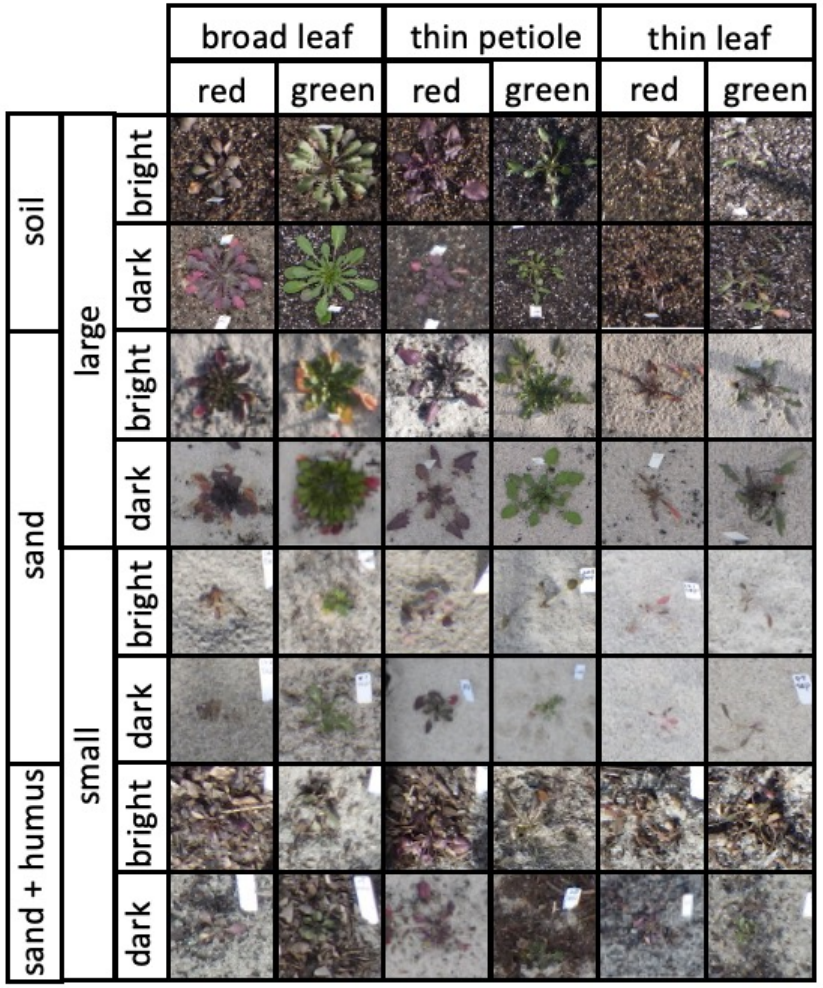


Akiyama et al., *Nature Commun* 14:5792, 2023

Identify plants using marbles

Difficulties *in natura:* broad variations

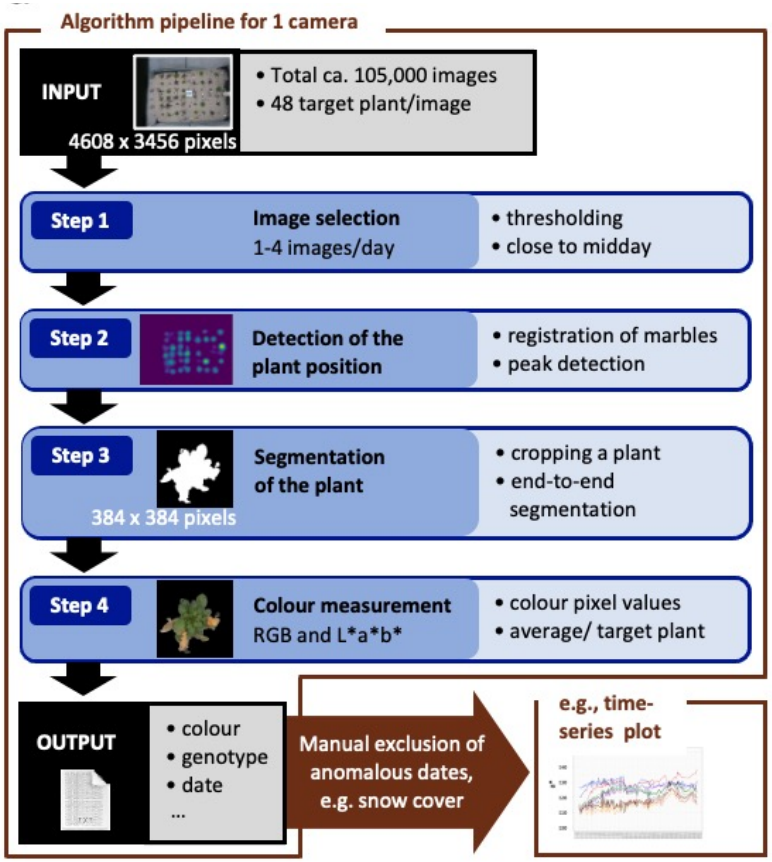


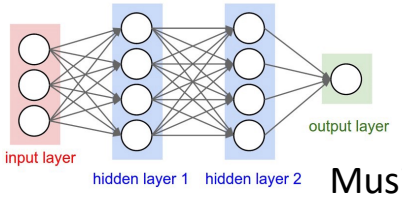


Akiyama et al., *Nature Commun* 14:5792, 2023

light condition and plants

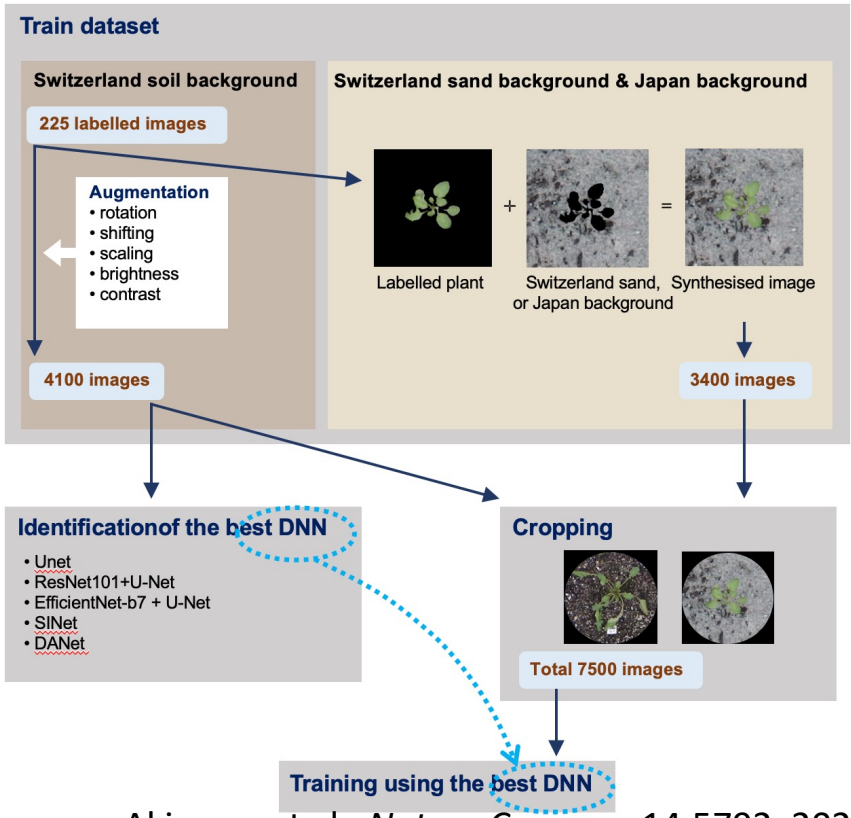
Deep learning: manual labeling and augmentation





Musiol 2016

Manual labeling
of 225 images
(4 weeks)

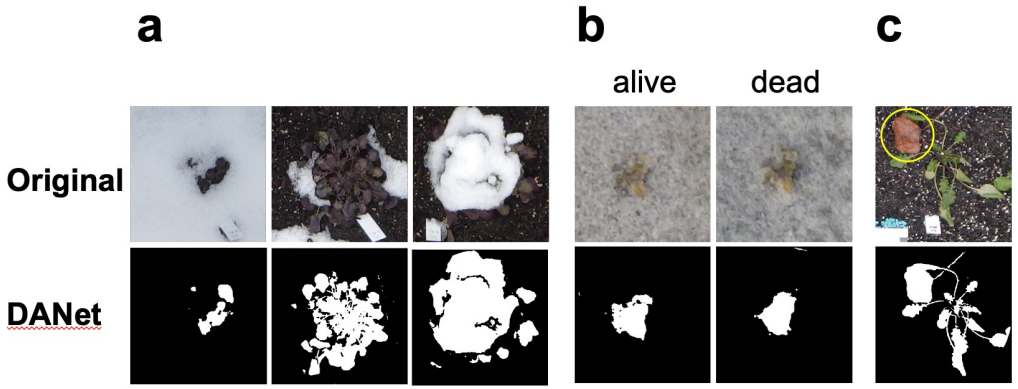


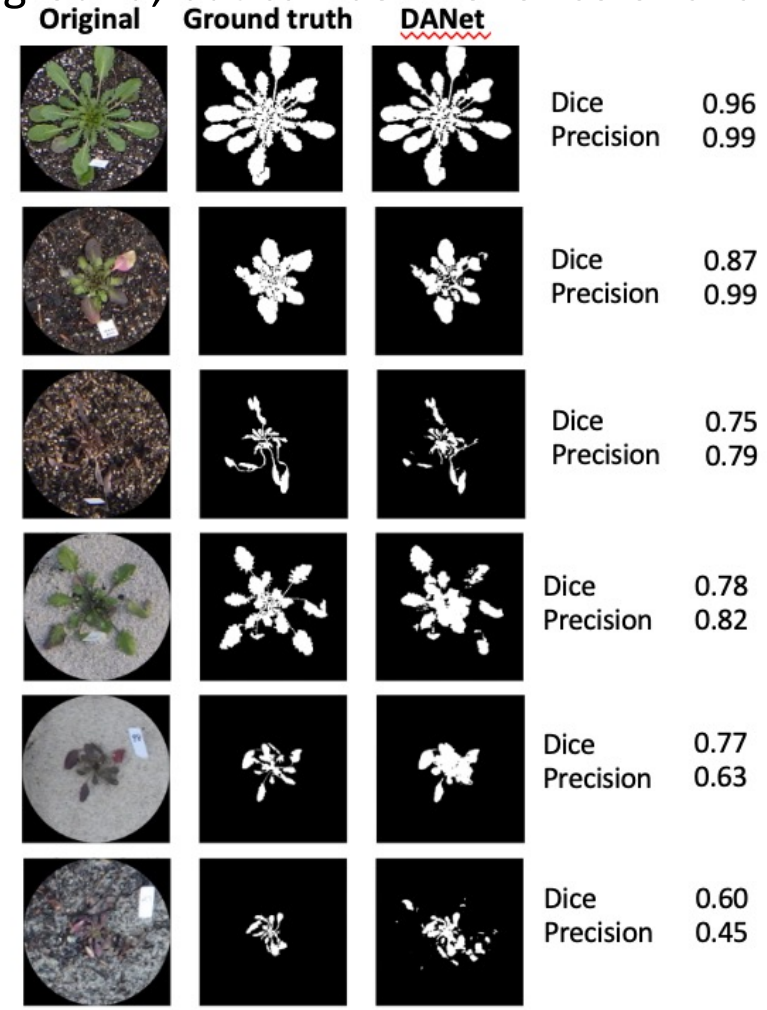
Akiyama et al., *Nature Commun* 14:5792, 2023

Doable for biologists

Checking by eye

Difficult when the plant color is similar to the background, but can be fine for color analysis





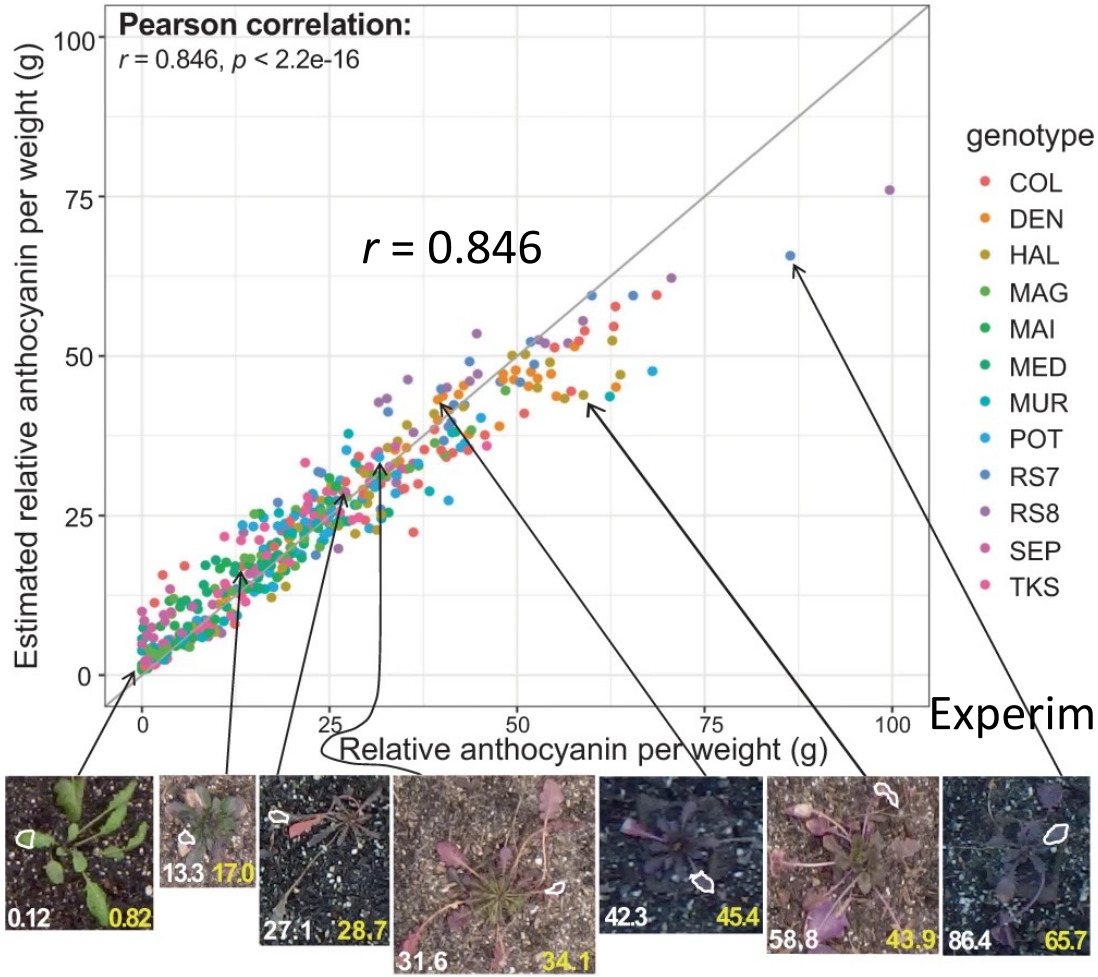
Akiyama et al., *Nature Commun* 14:5792, 2023

Remove snow season and outliners

Estimation and validation on anthocyanin concentration

Experimental (destructive) anthocyianin data of 451 samples

Fitting



Random forest
(a machine learning method)
used for lab samples by Askey et al. 2019
App Plant Sci 7: 11 e11301

Experimental data

Akiyama et al., *Nature Commun* 14:5792, 2023

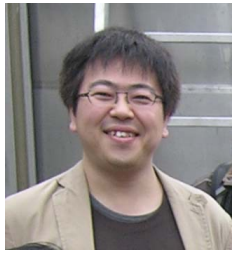
Automatic pipeline established

Machine learning: shape and color

Deep learning for segmentation

Manual labeling
of 225 images
(4 weeks)

Natsumaro Kutsuna, LPixel

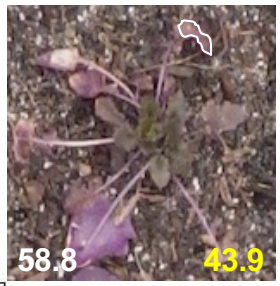


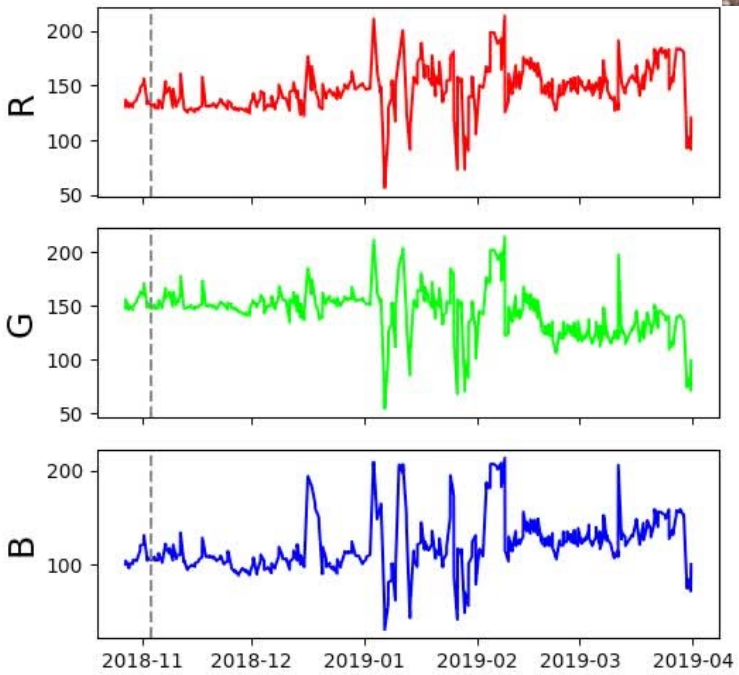
Jun Sese
Humanome

20181103_111011

Random forest for anthocyanin content

Experimental
Anthocyianin data of
451 samples



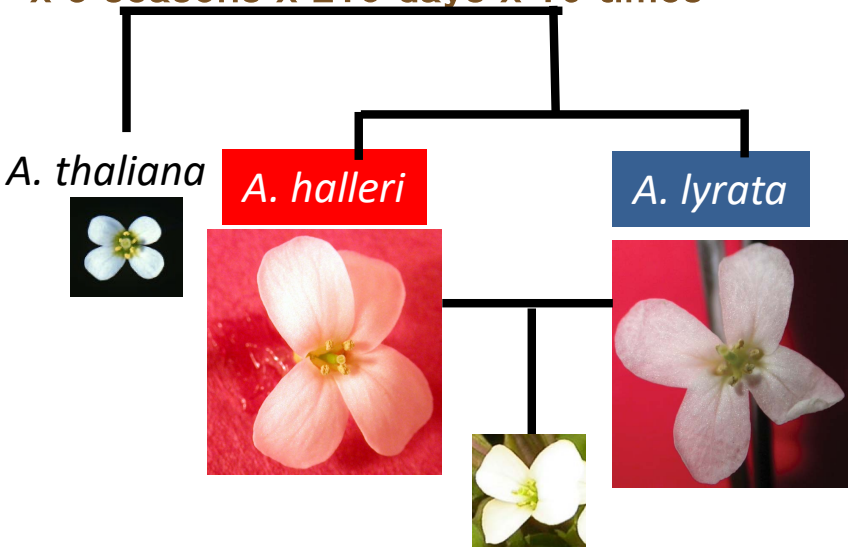


Akiyama et al., *Nature Commun* 14:5792, 2023

PlantServation pipeline established

Analysis of 3,870,000 images using PlantServation

12 genotypes x 16 replicated x 2 sites
x 3 seasons x 210 days x 16 times

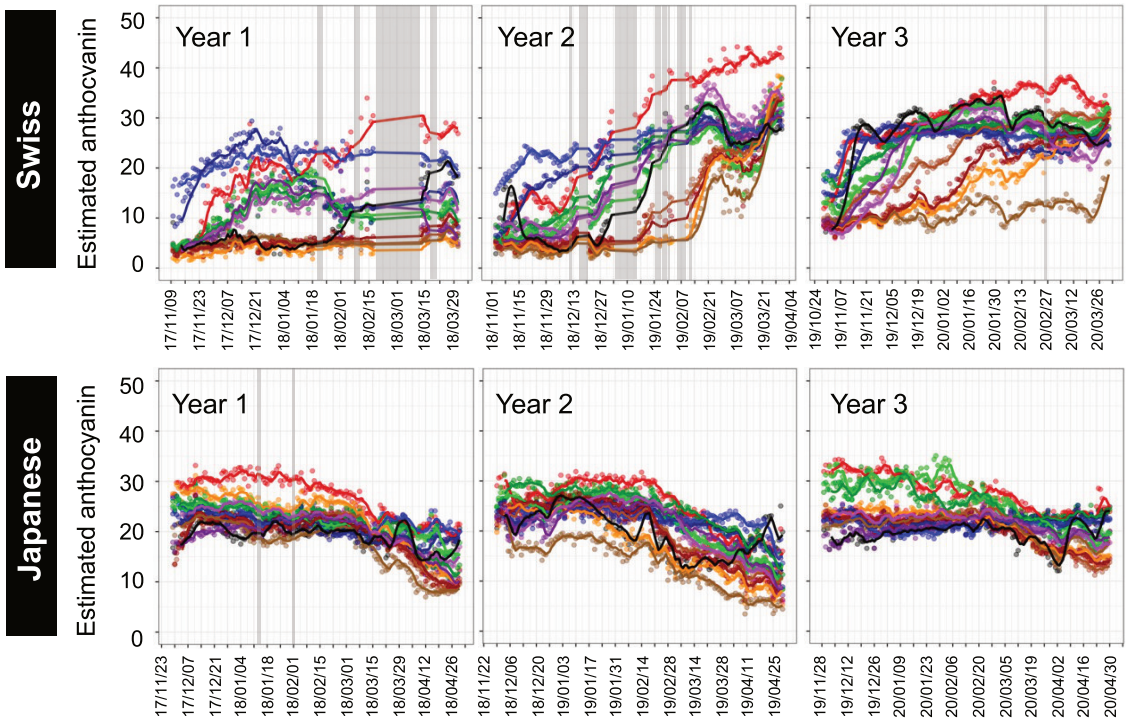


A. kamchatica
2 synthetic polyploids
2 natural (North)
4 natural (South)

Zurich, Switzerland

Kyoto University, Japan

Hiroshi Kudoh

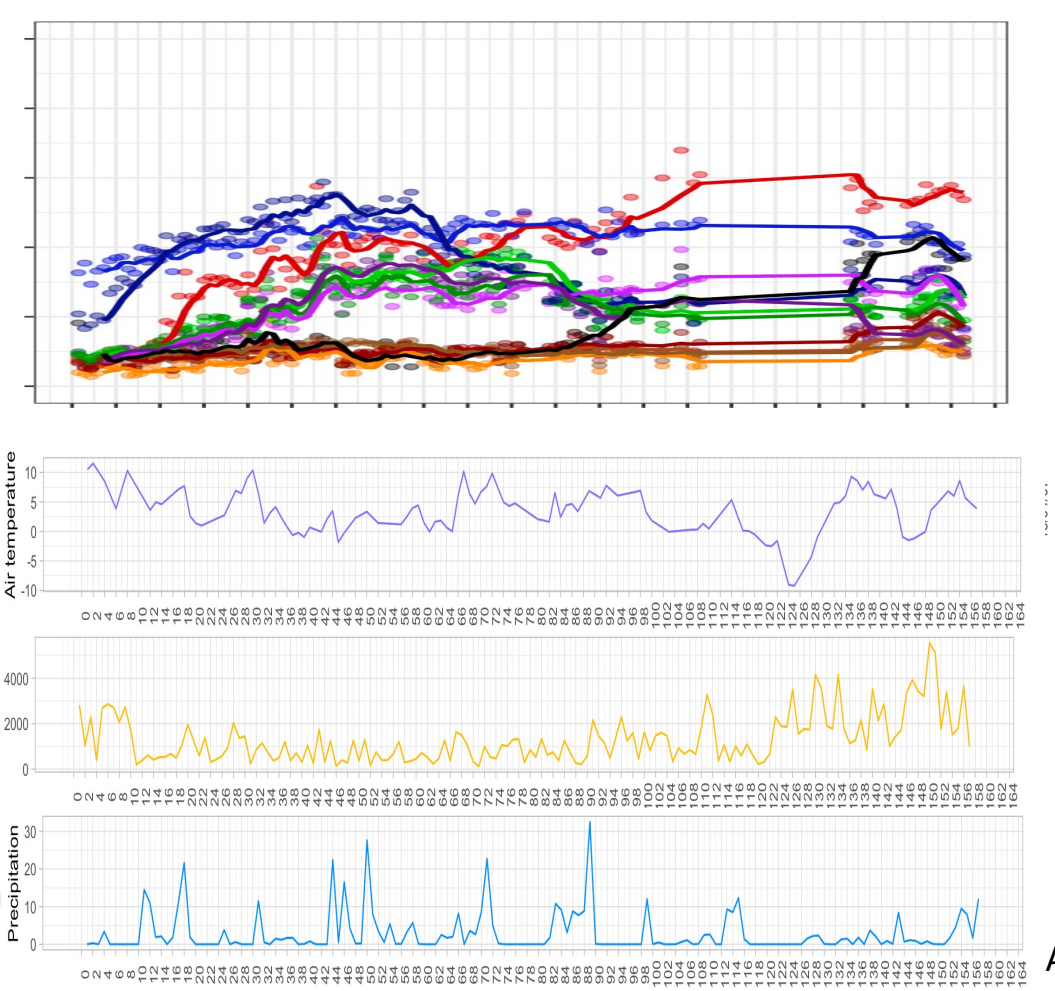


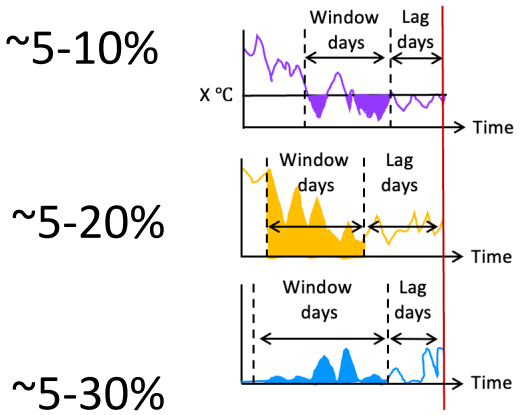
Akiyama et al., *Nature Commun* 14:5792, 2023

High variation among genotypes and seasons

Significant effect of temperature, sunlight, rainfall

Zurich
2017-2018



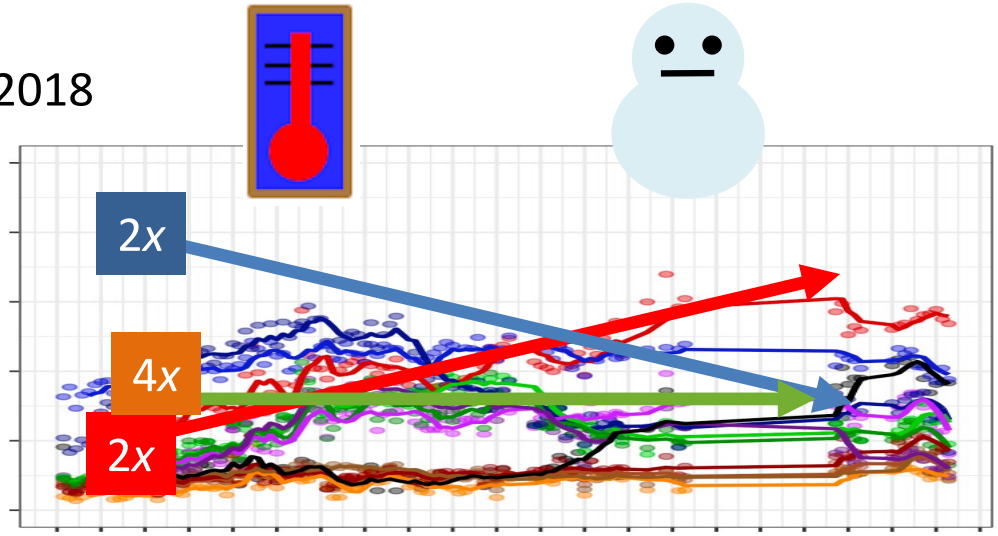


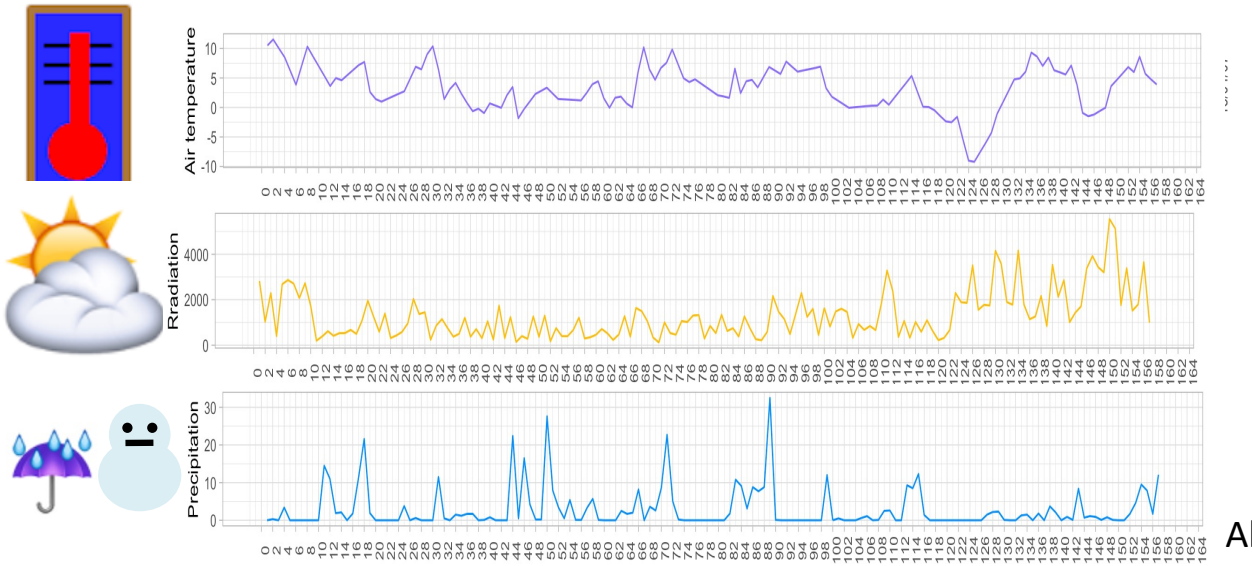
Akiyama et al., *Nature Commun* 14:5792, 2023

Majority of variation unexplained:
We know little about the environmental responses *in natura*

Genotypic differences

Zurich
2017-2018

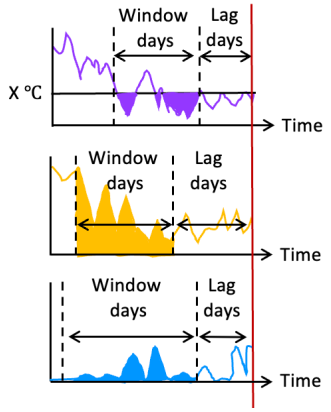




~5-10%

~5-20%

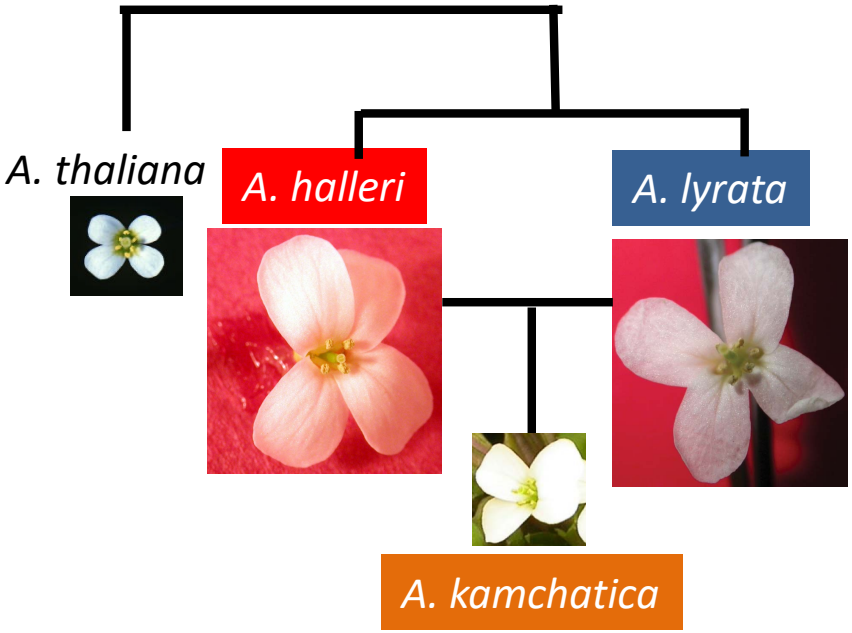
~5-30%



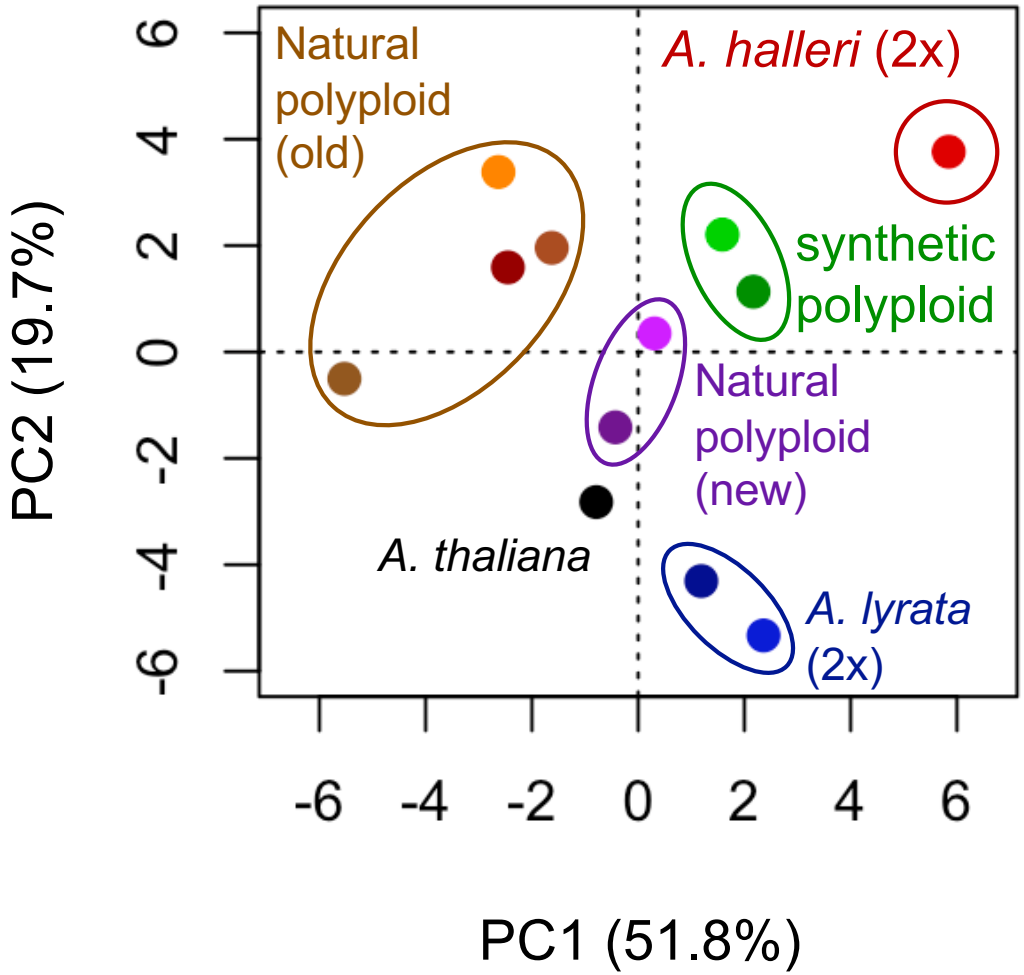
Akiyama et al., *Nature Commun* 14:5792, 2023

Polyploids were less stressed

Synthetic polyploids combined parental responses and recapitulated natural ones



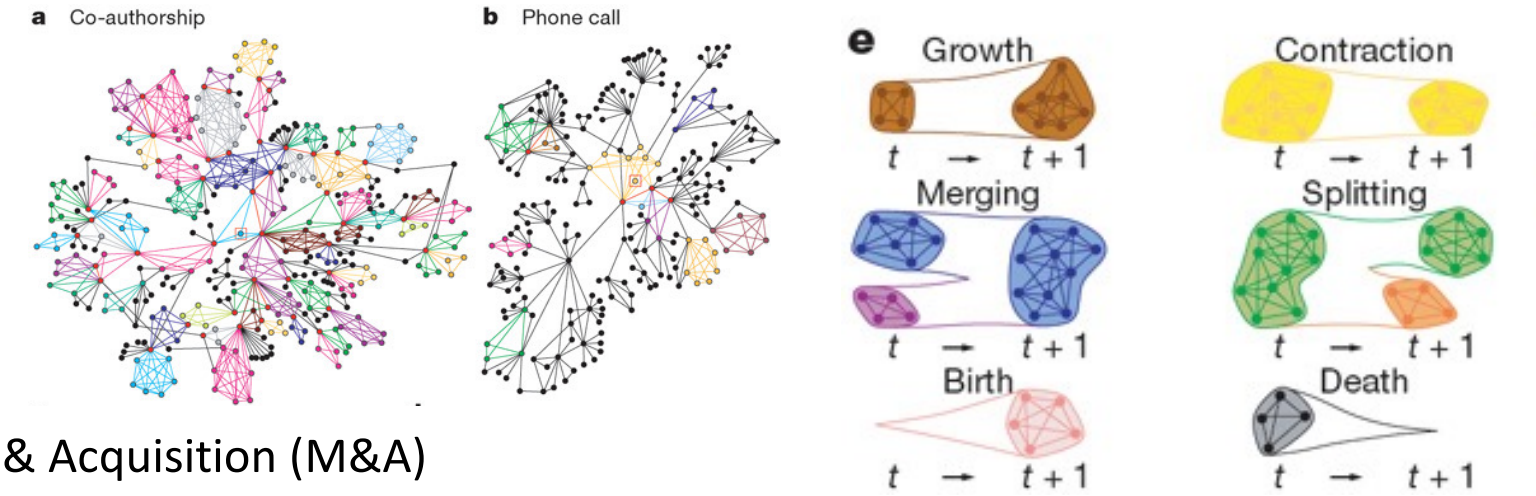
Principal Component Analysis



Akiyama et al., *Nature Commun* 14:5792, 2023

Did synthetic polyploids combine parental responses? YES
Can synthetic polyploids recapitulate natural speciation? YES

Discussion: generalist by a network merging



Merge & Acquisition (M&A)
Institute merging
Merge of gene networks by allopolyploidization

Pella, Barabasi, Vicsek, *Nature* 446;664, 2007

Diploids	Allopolyploids
Specialist	Generalist
Expert (geek)	Management (jack-of-all-trades and mater of none)
Niche industry	conglomerate

Shimizu, *Current Opinion in Plant Biology,* 69, 102292, 2022

Coexisting but agroecosystems may favor polyploids