Genome scanning for non model forest species ~linking a gap between genetic and phenotypic diversity~

Japan International Research Center for Agricultural Sciences

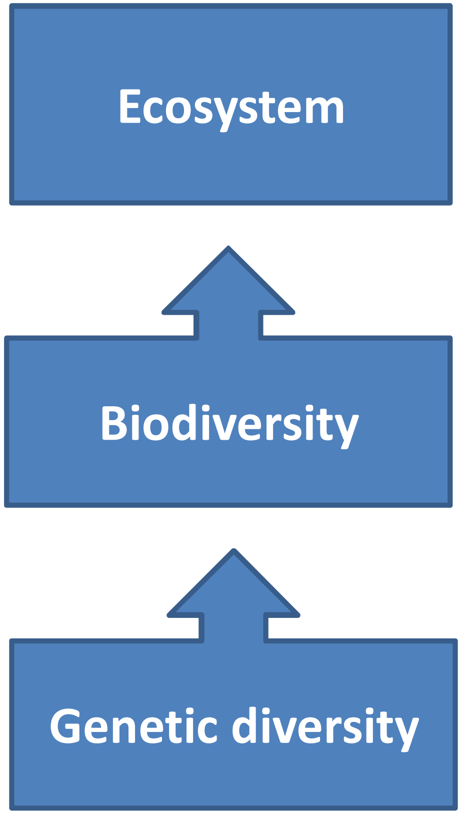
University of Tsukuba

Naoki TANI

The 11th AP-BON Workshop @ Double trees by Hilton, KL

2019-6-21

How genetic diversity supports biodiversity?



We believe that larger genetic diversity demonstrates higher resilience.

HOW?

Selection works against phenotypes, not for all genotypes.

For example,
- ✓ Drought causes less survival for water demand individuals
- ✓ Diseases cause less survival for susceptible individuals

Limited number of genetic markers show neutral against genetic diversity

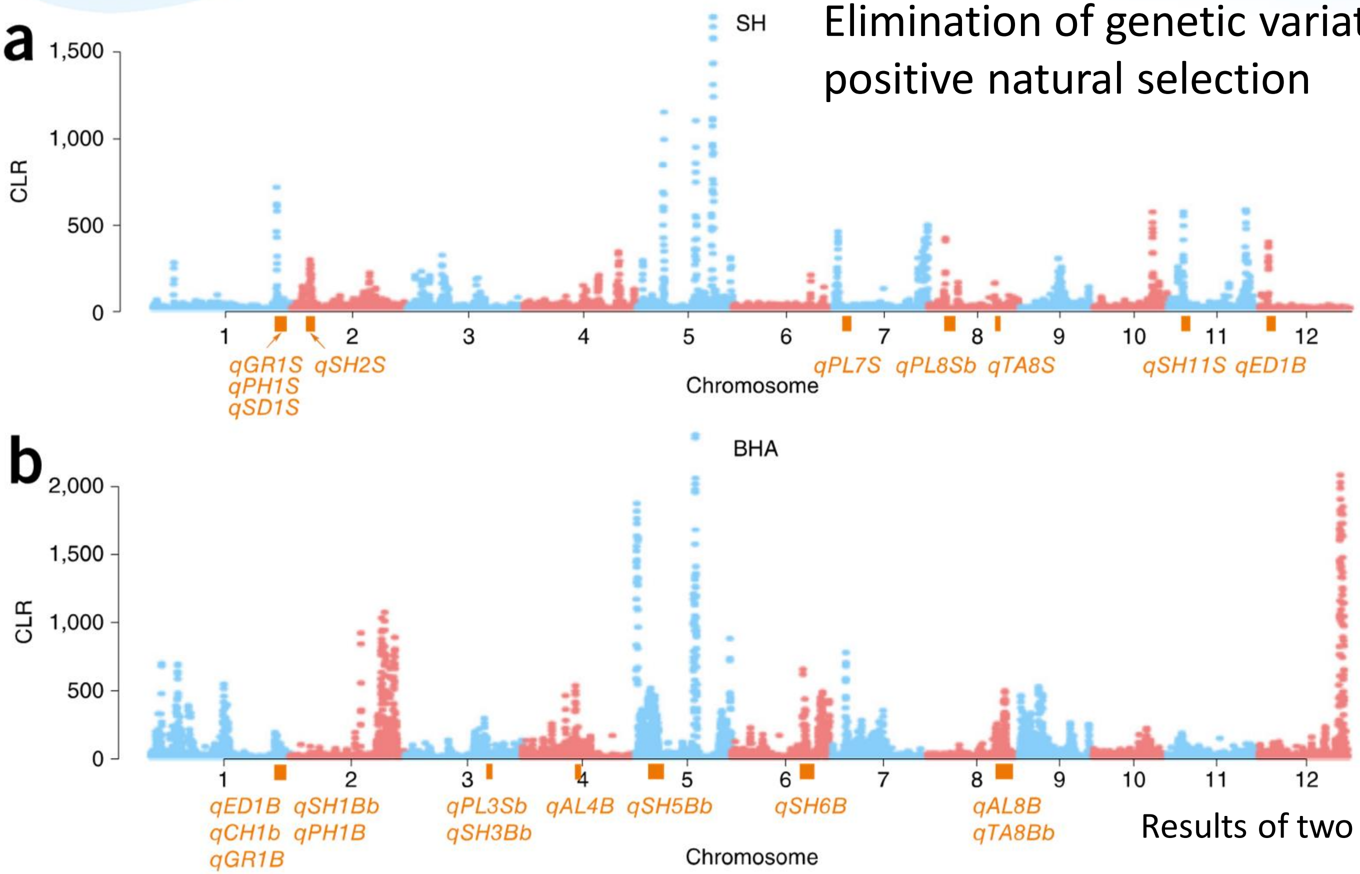
⬇

Linking between genotype and phenotype also require for biodiversity conservation

Amount of genetic variation is NOT equal throughout genome

Selective sweep

Elimination of genetic variation due to positive natural selection

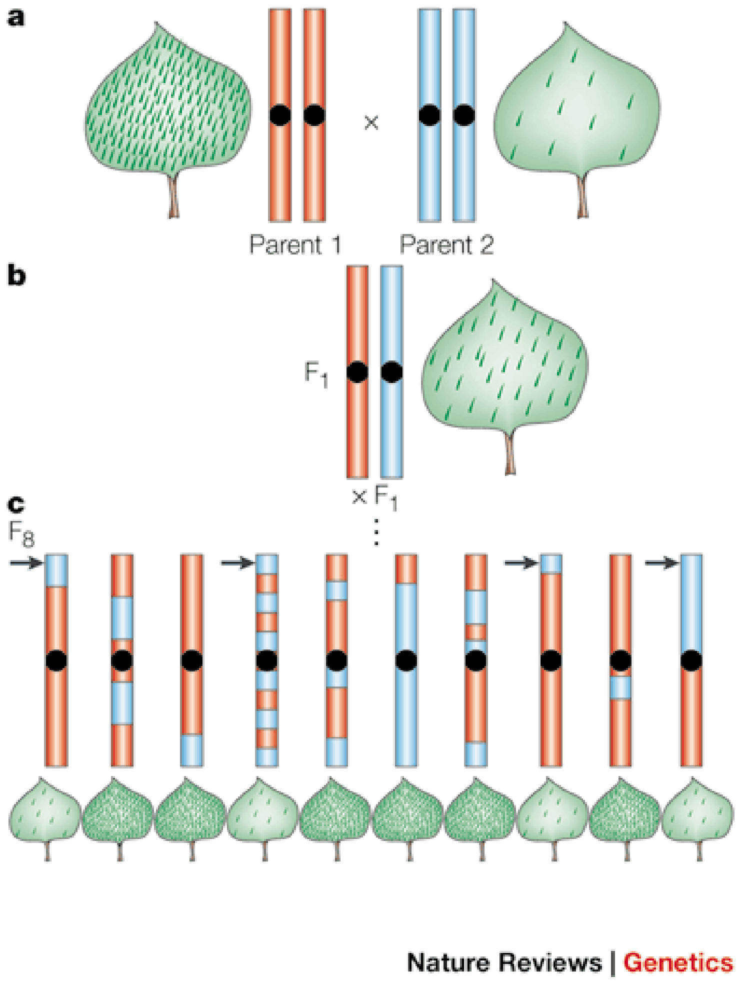


Results of two weed rice

Li et al. 2017 Nature Genetics

Phenotype is positively selected by environments, and genetic variation of the genes and flanking regions is eliminated.

Conventional methods to identify phenotype – genotype relationship



Mauricion 2001, Nature Genetics

Quantitative traits loci (QTL) mapping

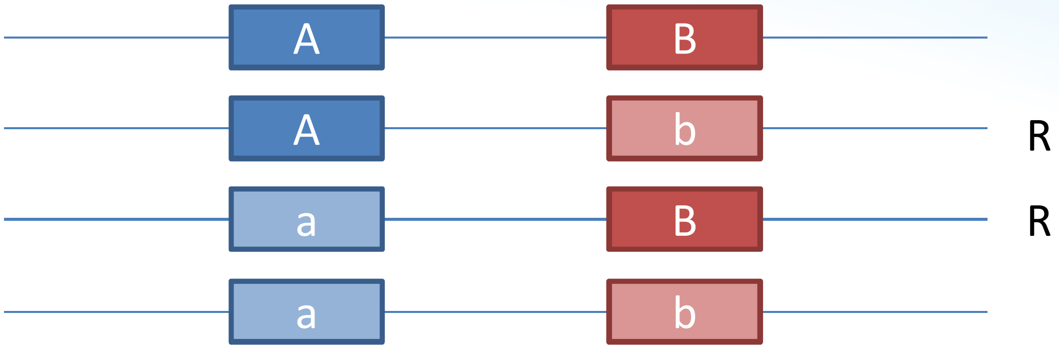
- ✓ At least 2 generations, ideally 3 generations are required.
- ✓ Segregation of quantitative phenotypic traits is analyzed on linkage map composed by genetic markers.

Limitation of QTL mapping

- ✓ Such as for long-lived woody plant, it is very difficult to make 2 or 3 generation pedigree.
- ✓ Identification of QTL is based on recombination of F1 individuals. Non-segregated QTLs are not surveyed.

Linkage disequilibrium (LD)

Consider only two SNPs

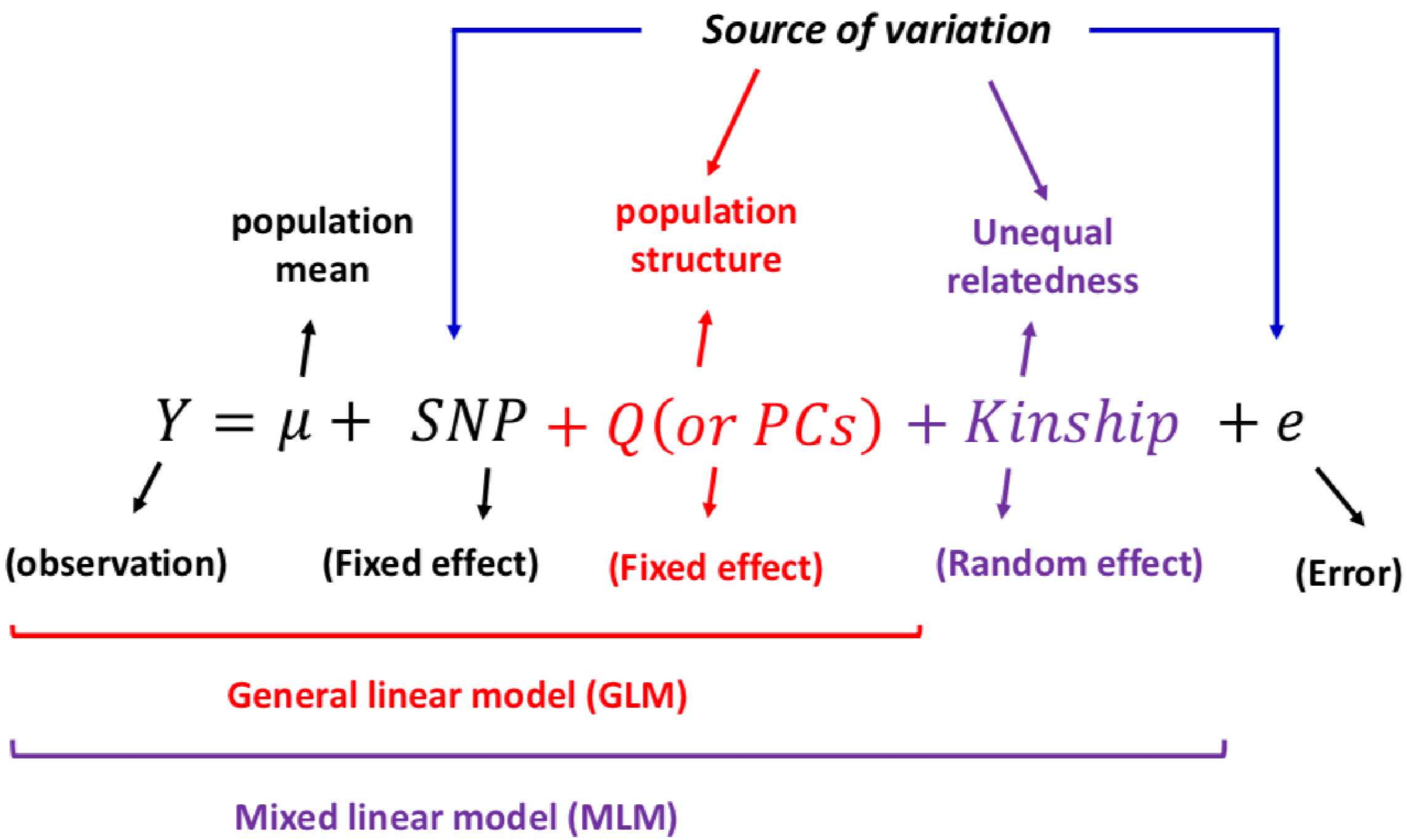


Probabilities for each haplotype in population

	B	b
A	$p_A q_B$	$p_A(1-q)_B$
a	$(1-p)_a q_B$	$(1-p)_a(1-q)_b$

- ✓ We can estimate distance between two loci on chromosome from frequency of two multi-locus genotypes
- ✓ Multi-generation pedigrees are not required!

Genome-Wide Association Study (GWAS)



Source of variation

population mean

population structure

Unequal relatedness

$$Y = \mu + SNP + Q(or\ PCs) + Kinship + e$$

(observation)

(Fixed effect)

(Fixed effect)

(Random effect)

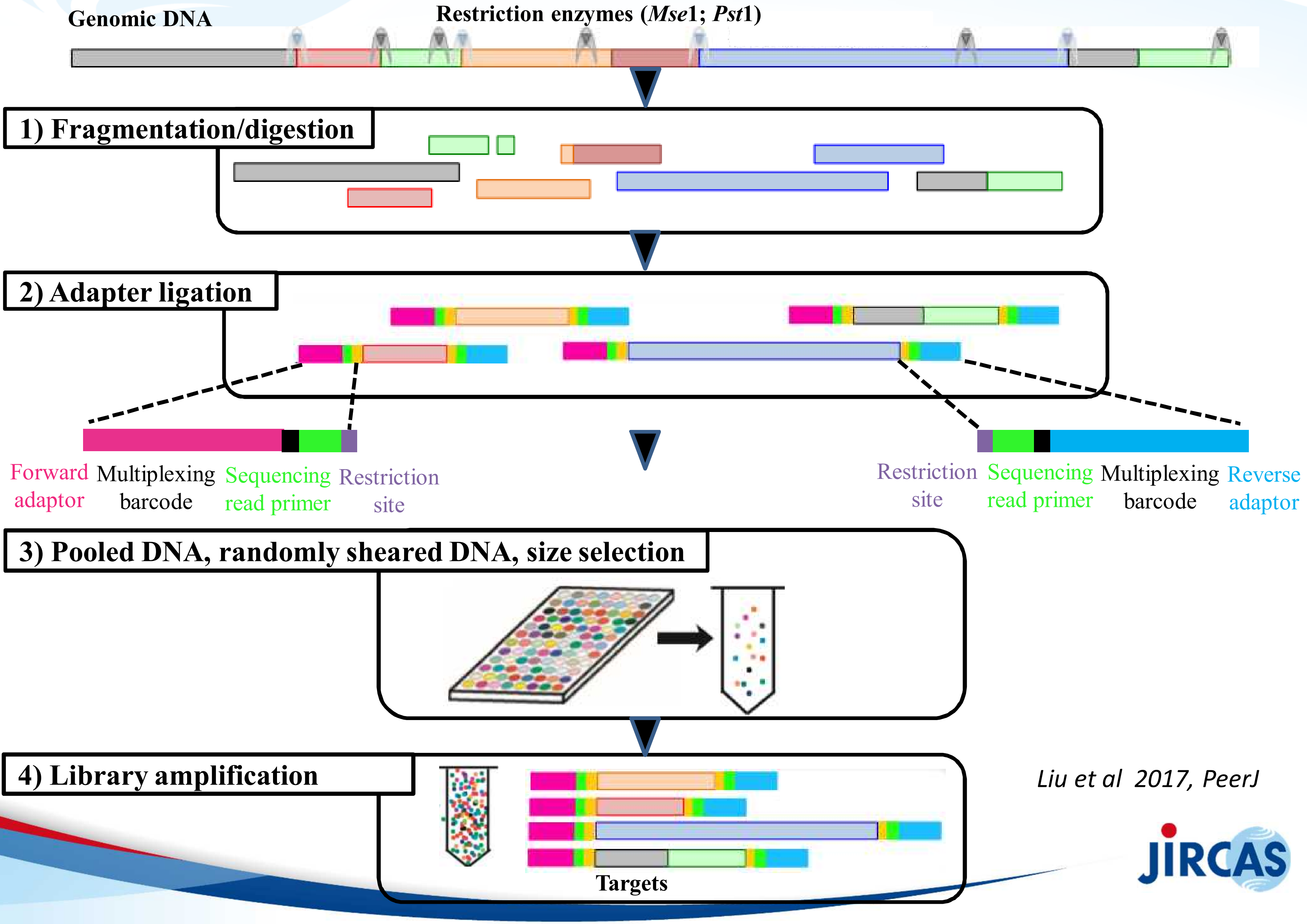
(Error)

General linear model (GLM)

Mixed linear model (MLM)

- ✓ Association of observed phenotypes is assessed by linear model.
- ✓ Now, large amount of SNP loci is required.

Obtaining many SNP loci by ddRADseq technology



Genomic DNA

Restriction enzymes (*Mse*1; *Pst*1)

1) Fragmentation/digestion

2) Adapter ligation

Forward adaptor
Multiplexing barcode
Sequencing read primer
Restriction site

Restriction site
Sequencing read primer
Multiplexing barcode
Reverse adaptor

3) Pooled DNA, randomly sheared DNA, size selection

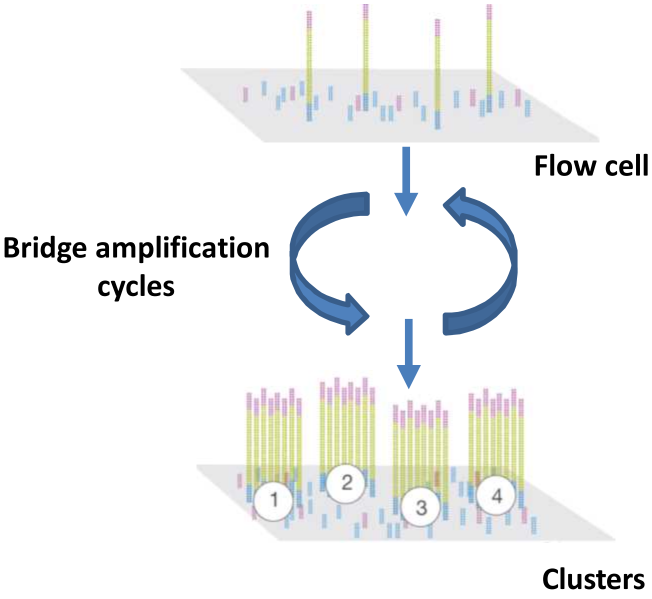
4) Library amplification

Targets

Liu et al 2017, PeerJ

Illumina HiSeq sequencing

1) Cluster Amplification

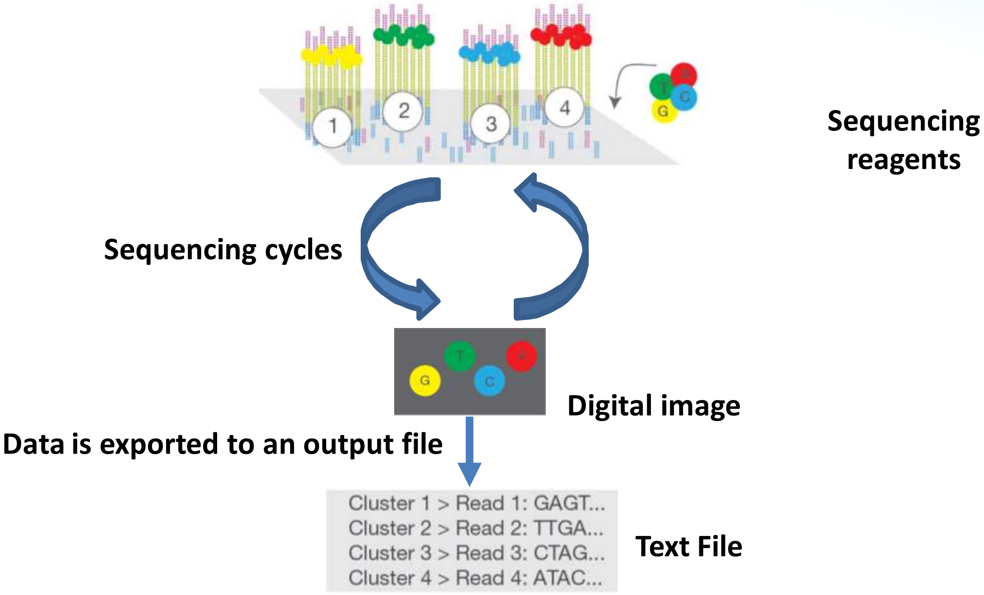


Flow cell

Bridge amplification cycles

Clusters

2) Sequencing



Sequencing reagents

Sequencing cycles

Digital image

Data is exported to an output file

Text File

```
Cluster 1 > Read 1: GAGT...
Cluster 2 > Read 2: TTGA...
Cluster 3 > Read 3: CTAG...
Cluster 4 > Read 4: ATAC...
```

3) Alignment and Data Analysis

Reads

```
  ATGGCATTGCAATTTGACAT
   TGGCATTGCAATTTG
AGATGGTATTG
 GATGGCATTGCAA
     GCATTGCAATTTGAC
  ATGGCATTGCAATT
AGATGGCATTGCAATTTG
```

Reference genome

```
AGATGGTATTGCAATTTGACAT
```

Liu et al 2017, PeerJ

Merits and demerits of RADseq and MIGseq

- ✓ RADseq generates more SNP loci than MIGseq
- ✓ RADseq require larger amount and higher quality DNA than MIGseq

Applications of GWAS

Phenotypes should be measured from uniform environment with some genetic diversity.

Experimental crop field

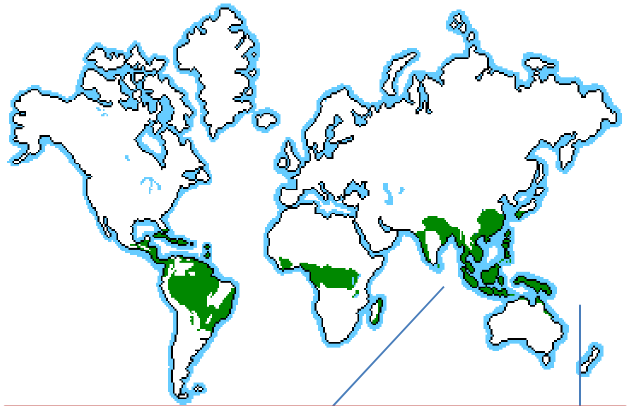
Fish farming

Tree clonal test

Cattle faming

Applications of GWAS for adaptation to climate change

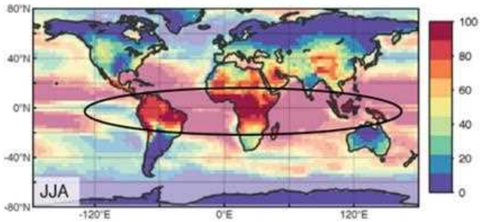
Selection of ecologically important species



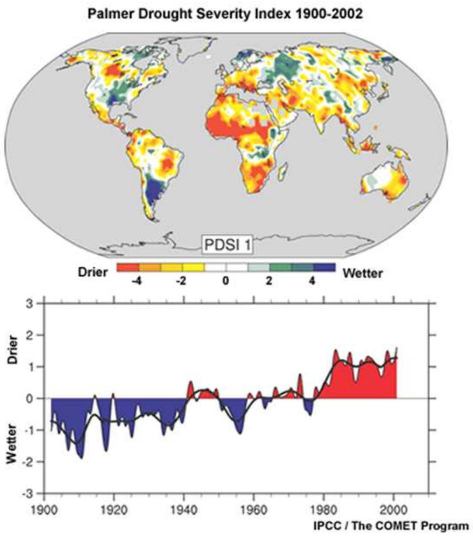
Tropical forest occupies only 7-10 % of land,
But it stores 40-50 % of carbon of land vegetation.

In South-east Asia, dipterocarps occupies nearly 80% of canopy trees.

Climate change threaten tropical forest



Increasing high temperature stress



Increasing drought stress

Utilize genetic resources to adapt climate change

Phenotype:
Growth traits
Physiological traits

Selection of genetic resources to adapt

Mitigation

Phenotype measurement

Difficulty of GWAS in tropical forest

Q: Can phenotypes be measured from uniform environment in tropical forest?

A: No, such test forest is very limited. But we found some.

Common garden experiment in FRIM
Species: *Shorea leprosula*

Progeny trial in central Kalimantan
Species: 5 species
S. leprosula, S. platyclados, S. Macrophylla etc.

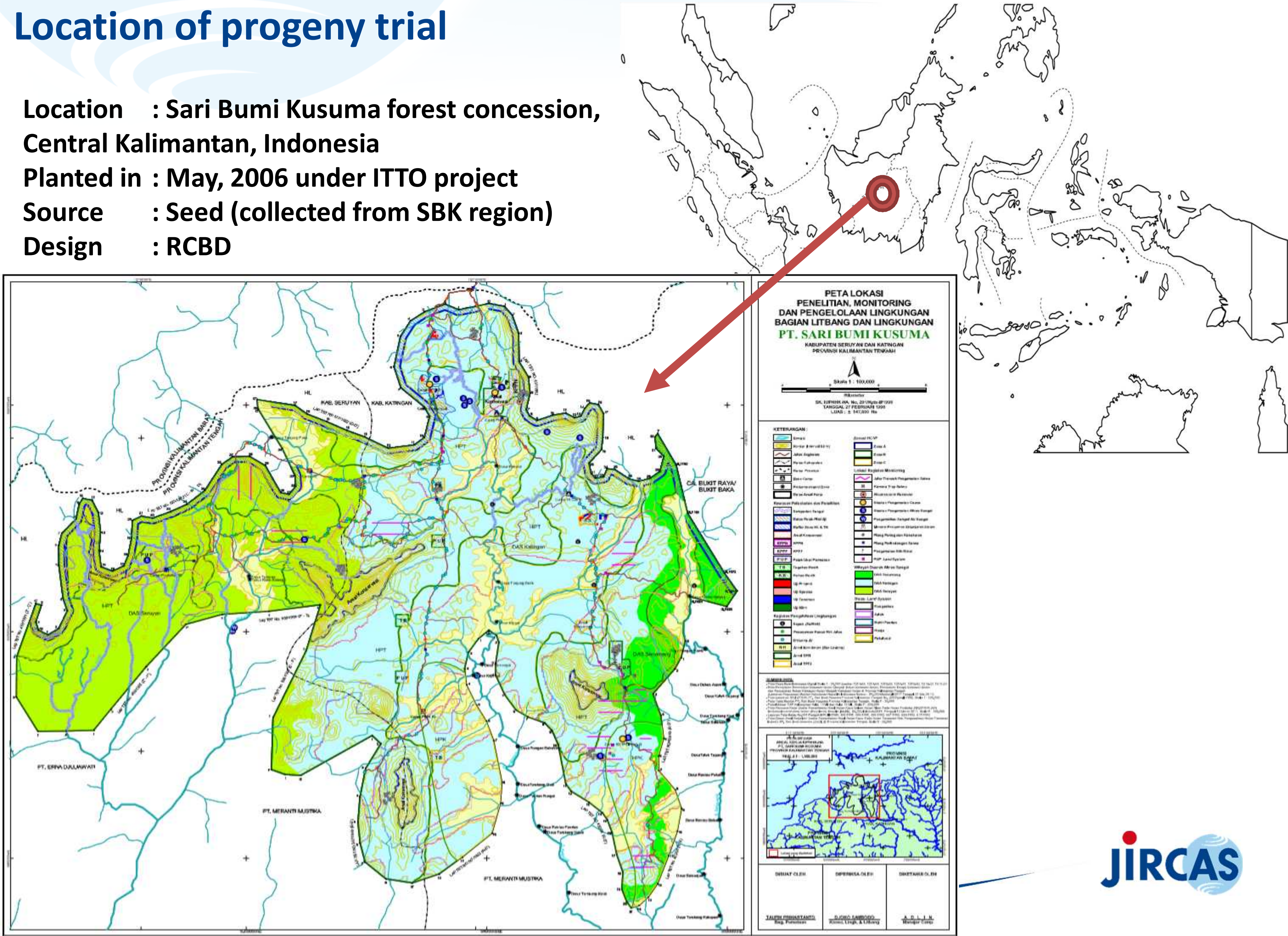
Location of progeny trial

Location : Sari Bumi Kusuma forest concession, Central Kalimantan, Indonesia

Planted in : May, 2006 under ITTO project

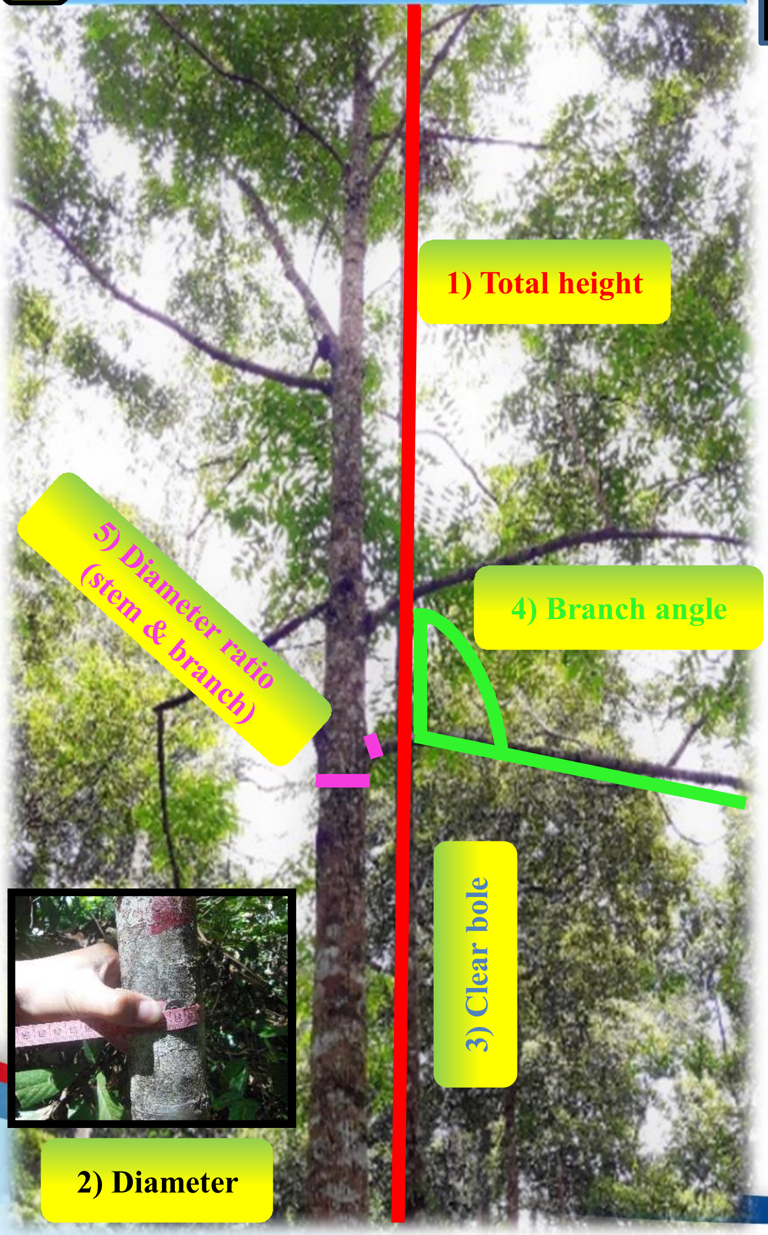
Source : Seed (collected from SBK region)

Design : RCBD



Plant materials and phenotype mearsurment

1 Phenotype data assessment



384 individuals from 77 families (mother trees)

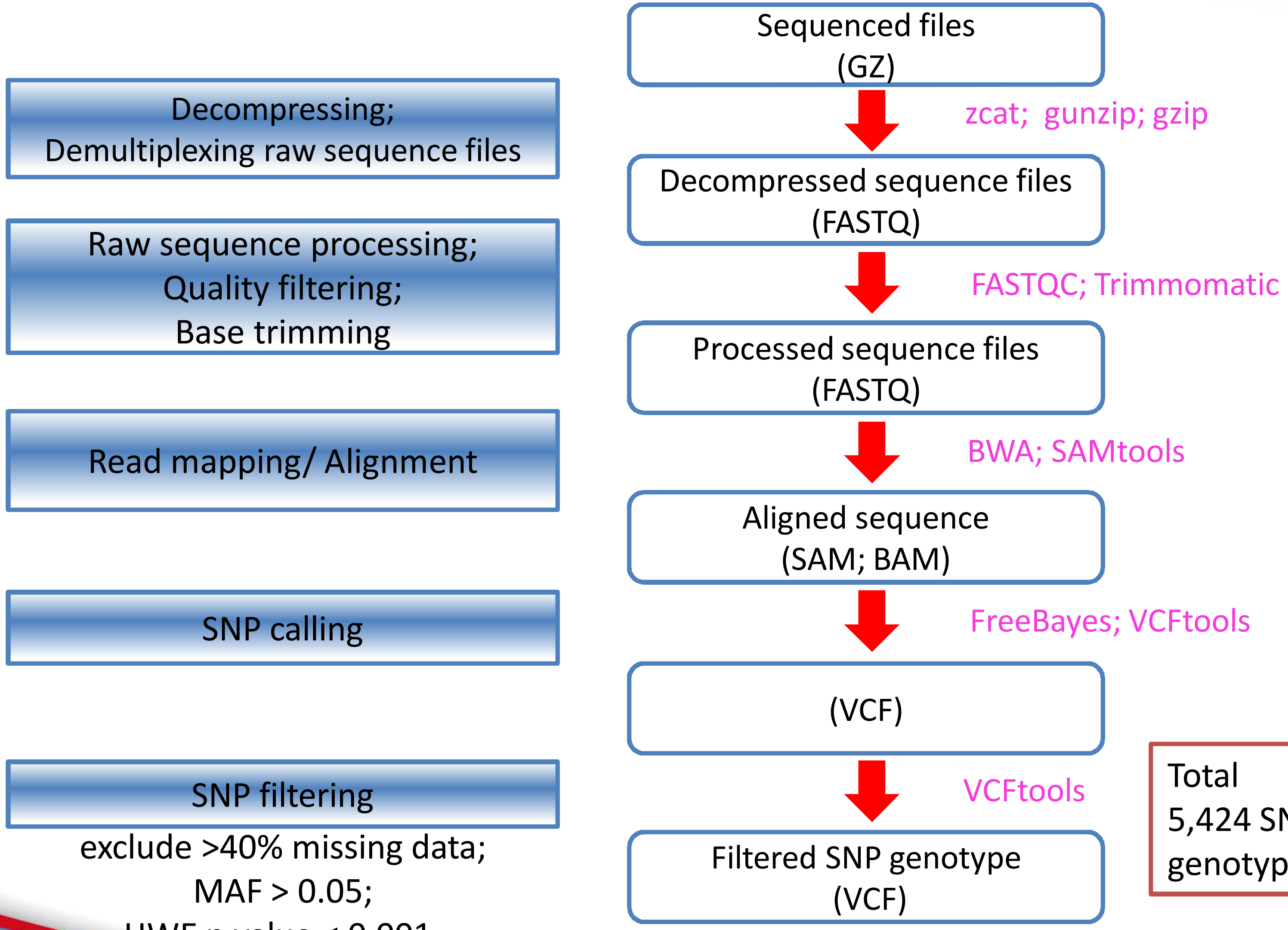
2 Leaf sampling collection

Dry leaf samples
(kept in silica gel during the transp. carried to Japan)

Genomic DNA extraction

SNP discovery after NGS

We use single flow cell of illumine Hiseq X, which outputs 800 – 900 Gb, 2.6 -3.0 billion reads.



Sequenced files
(GZ)

Decompressing;
Demultiplexing raw sequence files

zcat; gunzip; gzip

Decompressed sequence files
(FASTQ)

Raw sequence processing;
Quality filtering;
Base trimming

FASTQC; Trimmomatic

Processed sequence files
(FASTQ)

Read mapping/ Alignment

BWA; SAMtools

Aligned sequence
(SAM; BAM)

SNP calling

FreeBayes; VCFtools

(VCF)

SNP filtering

VCFtools

Filtered SNP genotype
(VCF)

exclude >40% missing data;
MAF > 0.05;
HWE p value < 0.001;
thin 1000 bp

Total
5,424 SNP loci were genotyped

Statistical pipeline after NGS

ANALYSIS	TOOLS
GENETIC POPULATION STRUCTURE **Probabilistic Principal Component Analysis (PPCA)** **Principal Coordinate Analysis (PCoA)** **Principal Component Analysis (PCA)**	**R packages** **GenAlex v6.5** **(**Peakall & Smouse, 2012**)** **TASSEL 5.0** **(**Bradbury et al., 2007**)**
LINKAGE DISEQUILIBRIUM (LD)	**TASSEL 5.0 (**Bradbury et al., 2007**)** **LOESS** in **R packages** (Locally weighted polynomial regression) (Cleveland, 1979)
ASSOCIATION ANALYSIS (MLM (Q+K model)) **(**incorporated both population structure and marker based kinship)	**rrBLUP in R packages** (Endelman, 2011) **qqman R package** (Turner, 2014)
GENE FUNCTION ANNOTATION	**Genome annotation tools**
GENOMIC PREDICTION **1) Ridge regression (RR)** **2) Least Absolute Shrinkage and Selection Operator (LASSO)** **3) Elastic Net**	**Beagle package ver. 4.1** (Browning et al., 2015) **glmnet in R packages** (Friedman et al., 2010)

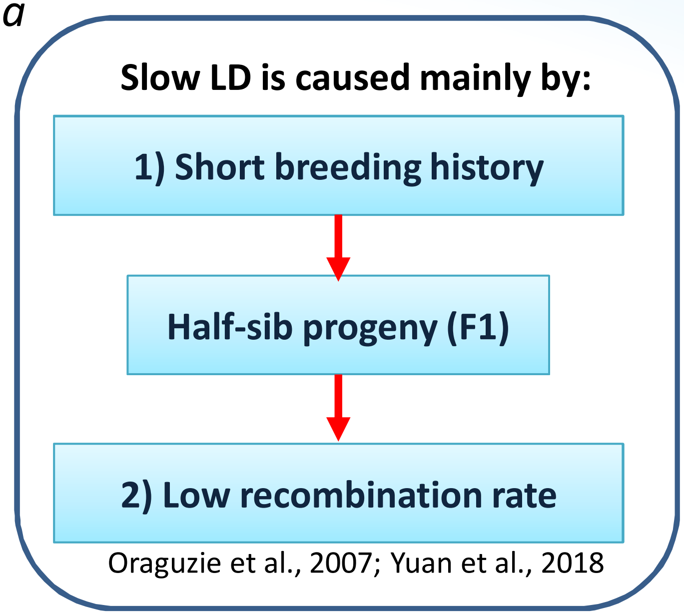
Genetic structure

Probabilistic Principal Component Analysis (PPCA) showed weak genetic structure.

$$\underset{\text{(observation)}}{\underset{\swarrow}{Y}} = \mu + \underset{\text{(Fixed effect)}}{\underset{\downarrow}{SNP}} + \underset{\text{(Fixed effect)}}{\underset{\downarrow}{Q(or\ PCs)}} + \underset{\text{(Random effect)}}{\underset{\downarrow}{Kinship}} + \underset{\text{(Error)}}{\underset{\searrow}{e}}$$

Linkage disequilibrium (LD)

LD decay at 0.1 R^2 is ?? kbp in progeny trial forest of *Shorea platyclados*



Slow LD is caused mainly by:

1) Short breeding history

Half-sib progeny (F1)

2) Low recombination rate

Oraguzie et al., 2007; Yuan et al., 2018

If rapid LD decay observed,
Cover more candidate genes possibly related to the traits

LD decay of *S. platyclados* using the ???? SNP markers dataset

Slower LD decay help us to detect genotype – phenotype relationship with limited SNP loci.

Spatial analysis of phenotypes before GWAS

Progeny trials

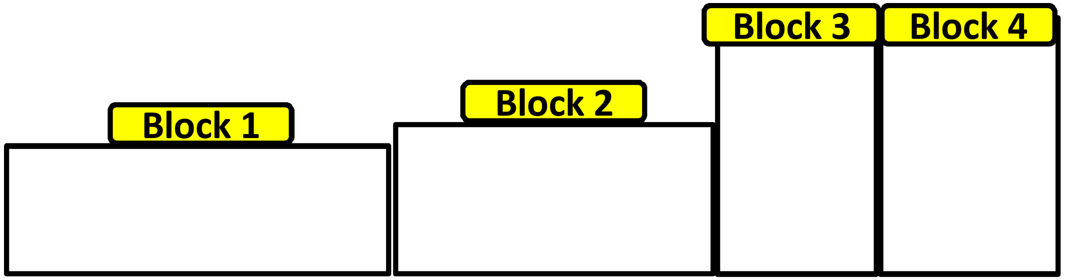
Cover large area:

- ❖ Large size of testing population
  (77 families with multiple individuals (4 trees plot/family)
- ❖ Large field space needed for individual trees
  (plant spacing of *S.platyclados*: 6 m x 3 m)

**Large physical area exhibits:
heterogeneous global and micro-environmental variation**

RCB Design

Site is divided into several blocks (replications) separating environmental variation from genetic variation



Spatial analysis

Improve the accuracy of breeding value prediction

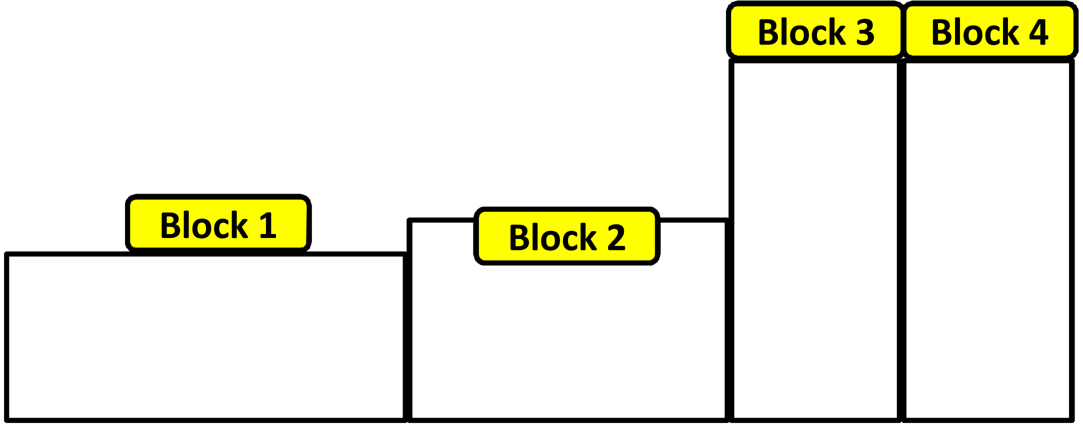
- ❖ Covering variance of microsite environmental factor (ex: soil depth, nutrition) in the local gradient variations within blocks (patches) and global gradient trend along the row and column of the trial between blocks (Bian et al., 2017; Chen et al., 2018)

Spatial analysis on breedR package

Diameter of *S. platyclados* before and after spatial correction

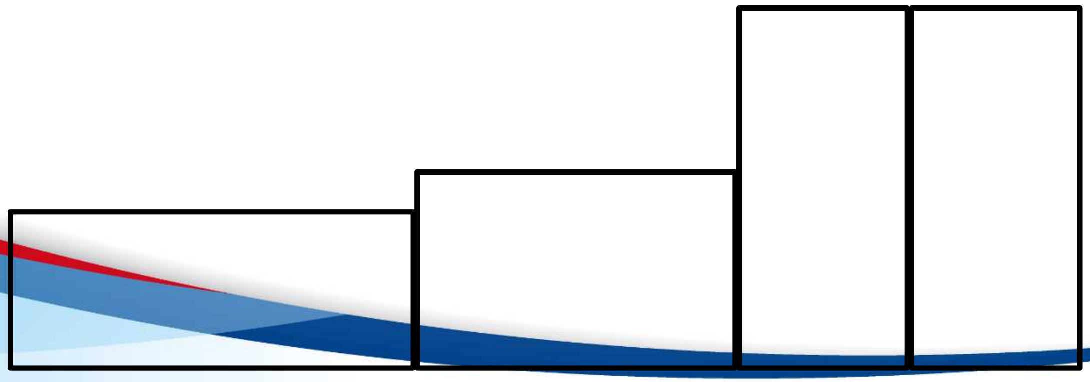
Before spatial analysis

Environmental heterogeneity :
High variance within block and between block



After spatial analysis

Spatial analysis reduce the error variance



Summary of phenotype

Phenotypic variation of the progeny trial of *Shorea platyclados*

Distribution of diameter and height at 12 years old

Genome Wide Association Study (GWAS)

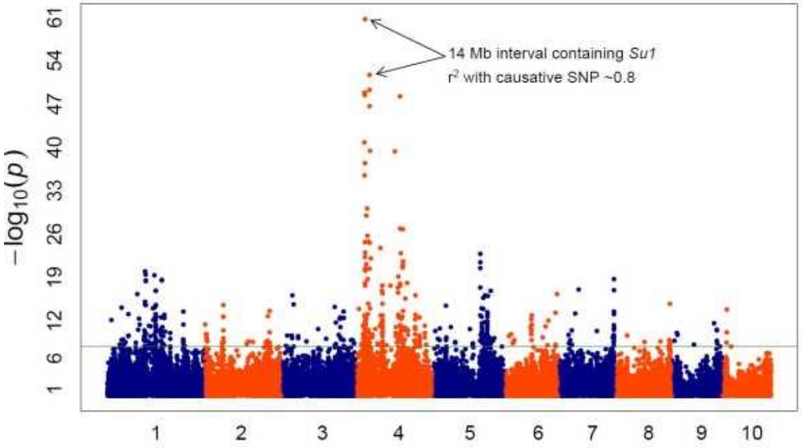
Association analysis based on MLM (Q+K model)

Incorporated both of population structure & marker based kinship

MLM approach, greatly reduced false positives, as shown in quantile-quantile plots

- ✓ Many SNP loci are associated with diameter (phenotype). However, -$\log_{10}(2)$ (0.01%) level can not distinguish between association and false positive, because of the large number of SNP loci.
- ✓ Multiple comparison is required. However, after controlling P value by Bonferroni and FDR, no significant SNP loci is detected.
- ✓ Many small effect loci may associate the diameter growth.

Genome Wide Prediction (Selection)



Sweet and starchy of corn kernel of 2,145 USA corn inbred line, Romay et al. (2013) Genome biology

SNP loci can be used for genetic markers for selection

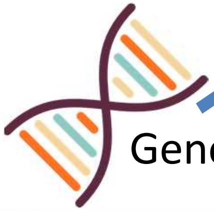
But, when there is no strong associated loci, how ?

DNA marker data

Phenotypic data

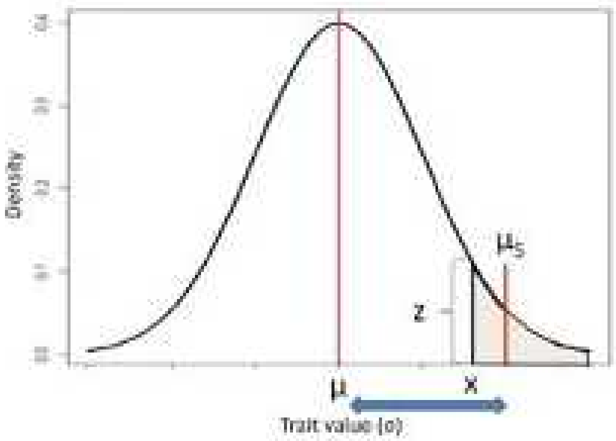
Model training

$$y = Xb + Zu + e$$

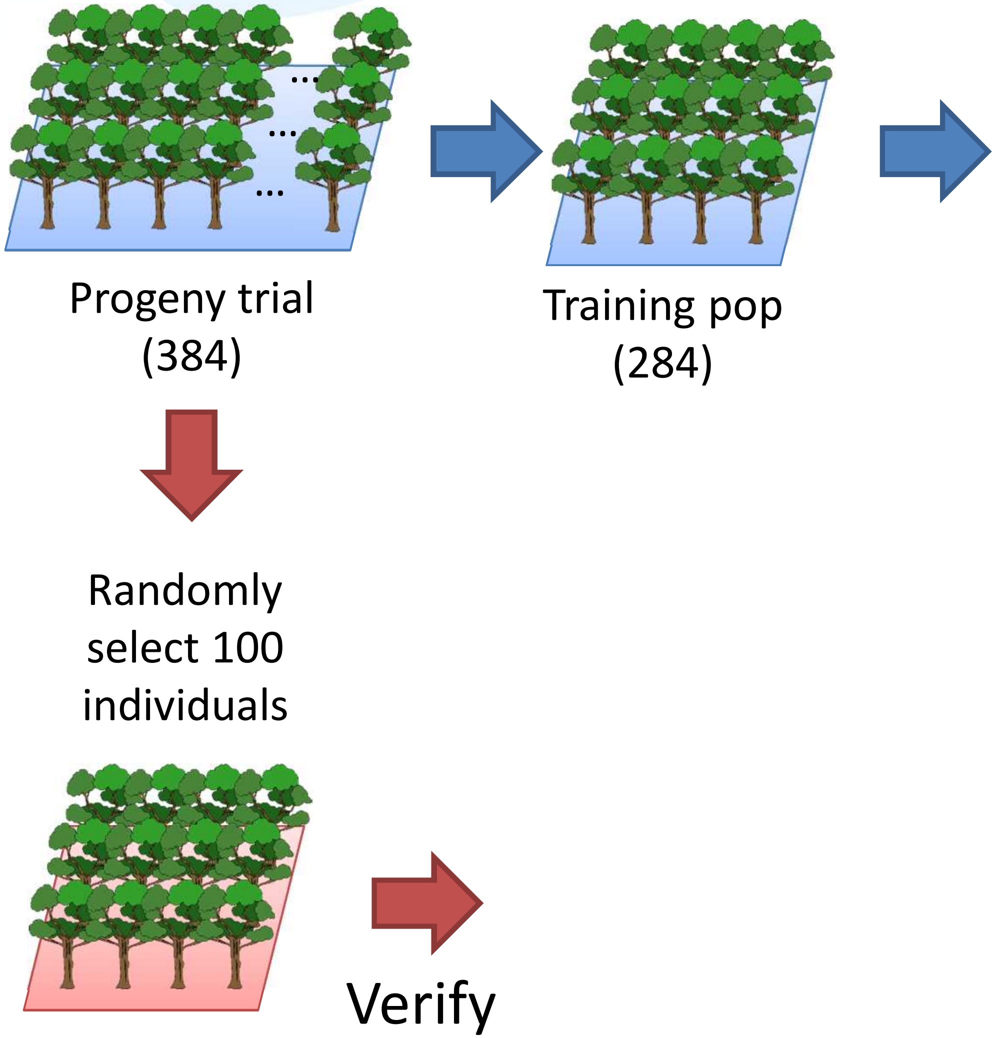


Genotyping

Predict and select



Genomic selection and enrichment planting



Progeny trial
(384)

Training pop
(284)

Model training

$$y = Xb + Zu + e$$

Randomly
select 100
individuals

Test pop
(100)

Verify

We can use SNP
genotypes for
selection!

Phenotyping in future

Our analysis is now ongoing…..

- ✓ Measured phenotypes are limited in growth related traits.
- ✓ Genomic selection can enhance growth of dipterocarp plantation, which can contribute to return of forestry, mitigation of climate change.

In near future, we measure physiological traits, which is highly related to environmental changes, such as climate change.

- ✓ Leaf flushing timing
- ✓ Photosynthesis activity
- ✓ Water potential in leaf
- ✓ Root catabolism speed

Etc.

We can efficiently utilize genetic resources through genomic selection and planting for mitigation and adaptation of climate change!

Research member and acknowledgement

U Gadja Mada	**Widiyatno, Indrioko S, Utomo S, Na’iem M**
U Tsukuba	**Sawitri, Tsumura Y,**
FFPRI	**Uchiyama K**
FRIM	**Lee SL, Ng KKS**
JIRCAS	**Oka H, Yamamoto Y, Tani N**

Supported by PT Sari Bumi Kusuma

www.jircas.go.jp