# Comparison of methods:
# Genotyping-by-sequencing and exome capture applied in genomic selection of Norway spruce (*Picea abies*)

Jin Pan
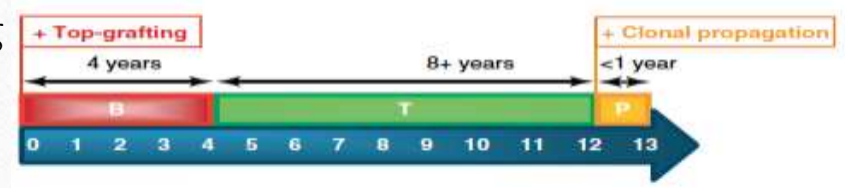
S L U
Umeå Plant Science Centre

2017-10-23

# Outline

- Background
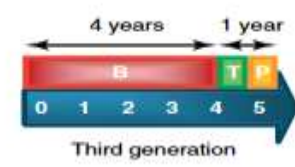
- Project aims

- Materials and Methods

- Results

- Summary
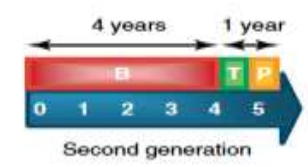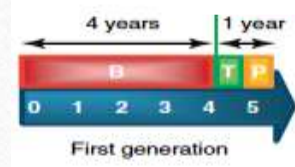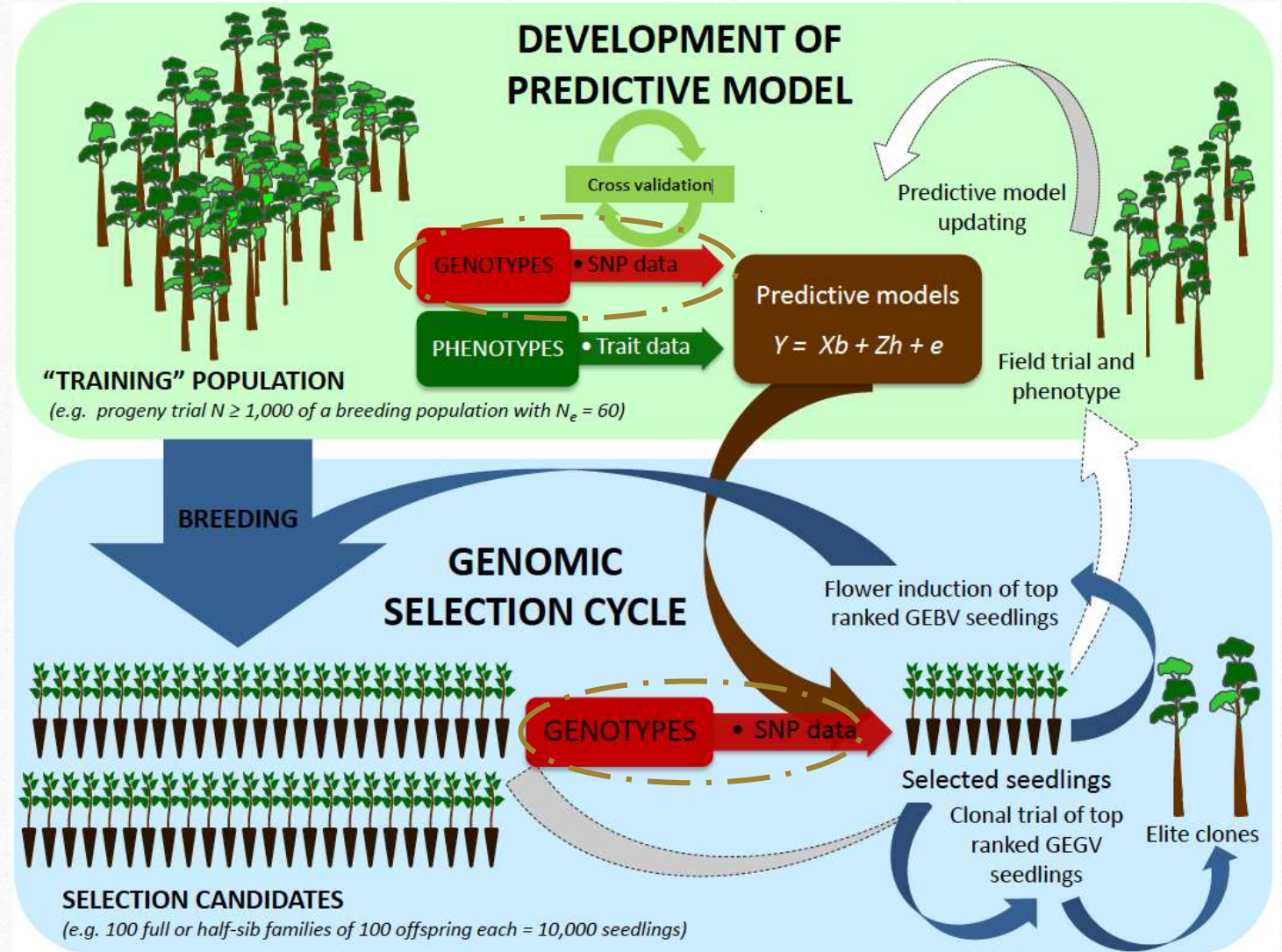
# Molecular breeding



Traditional breeding

Molecular breeding

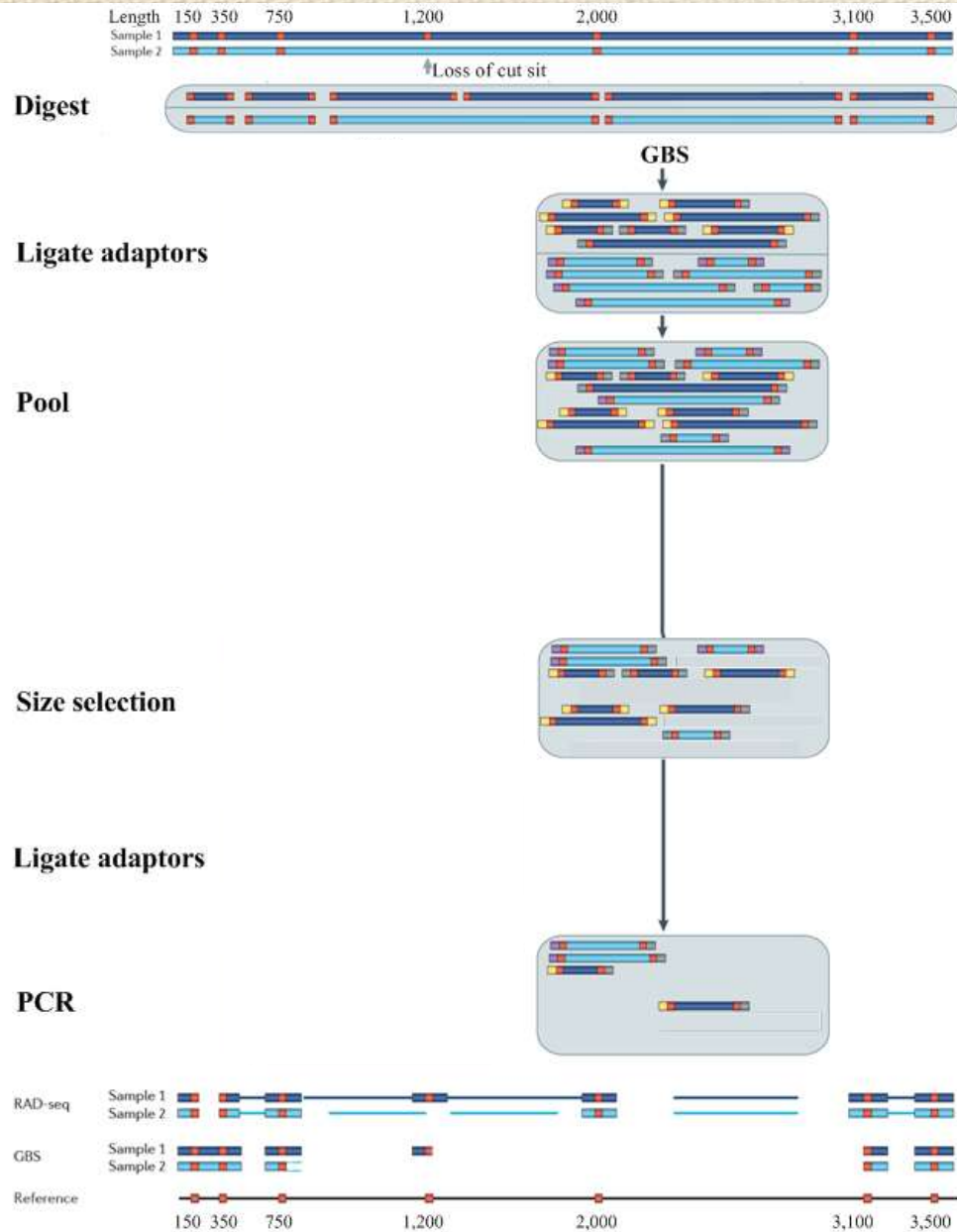B: **B**reeding
T: **T**esting
P: **P**ropagate

# Genomic selection



Background

Target enrichment  Exome capture

Transcriptome sequencing

Genome DNA
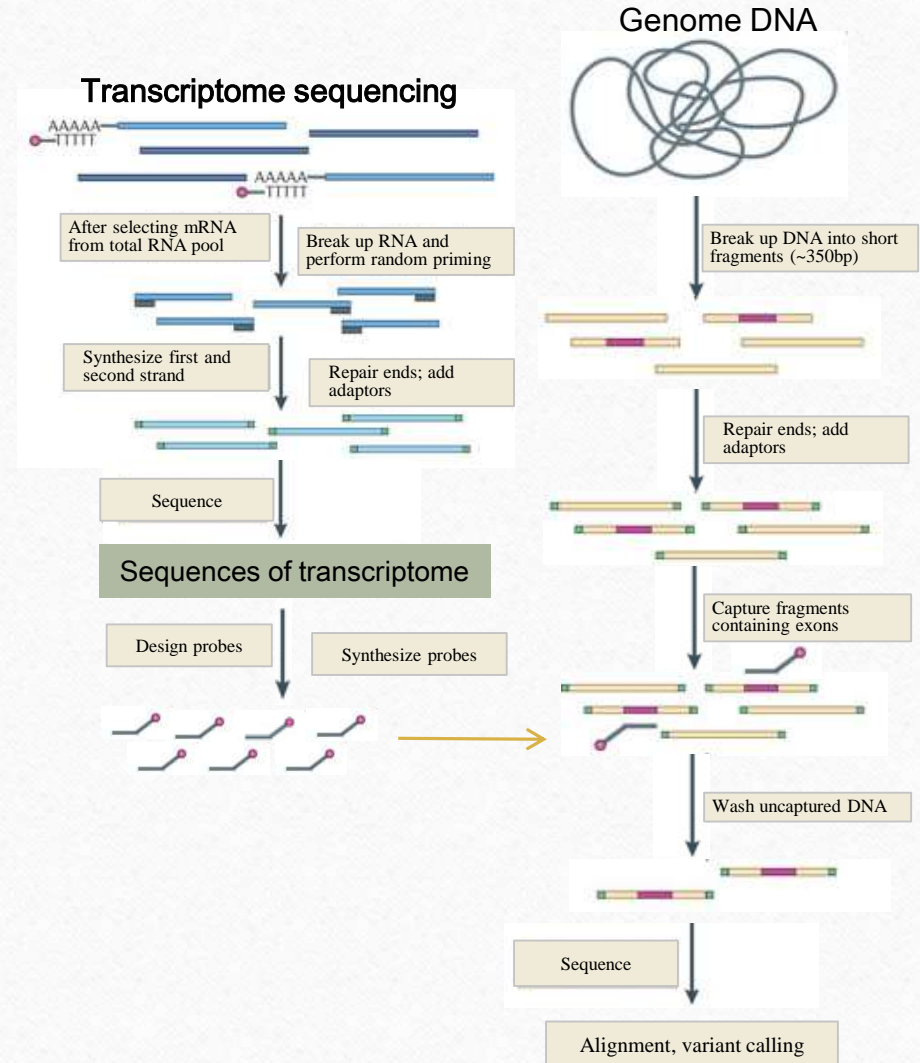
# Disadvantage

GBS

- A large percentage of uncalled genotypes (missing)

- Under-calling of heterozygous genotypes (false homozygous genotypes)

# Project aims

- Optimize SNP-calling

    - Estimate the rate of false genotypes and the rate of discordant genotypes in replicated samples

- Evaluate the potential of GBS and EC in genomic selection of Norway spruce

# Materials

## I Genotyping

- 376 samples from 24 ful-sib families (31 parental individuals)
- ~100 endospurms and 5 mother trees
- 5 (GBS) or 21 (EC) individauls sequenced repeatedly (2-8×)

## II Phenotyping

4 growth and wood quality traits:

- Tree height
- Wood density (Pilodyn and Hitman)
- Wood stiffness (Modulus of elasticity - MOE)



Hädanberg

# Methods

➢Genotyping-by-sequencing (GBS)

- *Pst*I & 230 - 430bp restriction fragments

- ~ 20,000 restriction fragments

➢Exome capture (EC)

- 40,019 probes located in 24,919 scaffolds

# Results

1. Optimization of SNP-calling

2. Comparison of GBS- and EC-SNP data

# Results

1. Optimization of SNP-calling


2. Comparison of GBS and EC data

# GBS-SNP-calling pipeline
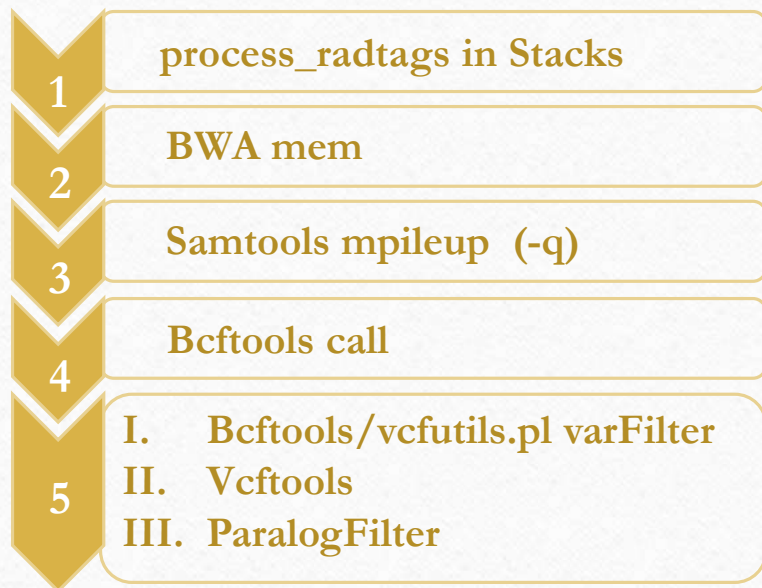
**1** — **process_radtags in Stacks** — - Barcode identification and quality trimming of raw illumina reads

**2** — **BWA mem** — - Align sample reads to reference genome

**3** — **Samtools mpileup  (-q)** — - Align all samples together

**4** — **Bcftools call** — - Call SNPs and indels

**5** —
I. **Bcftools/vcfutils.pl varFilter**
II. **Vcftools**
III. **ParalogFilter**
— - Filter SNPs

### bcftools/vcfutils.pl varFilter

```
-Q INT     minimum RMS mapping quality for SNPs [10]
-d INT     minimum read depth [2]
-D INT     maximum read depth [10000000]
-a INT     minimum number of alternate bases [2]
-w INT     SNP within INT bp around a gap to be filtered [3]
-W INT     window size for filtering adjacent gaps [10]
-1 FLOAT   min P-value for strand bias (given PV4) [0.0001]
-2 FLOAT   min P-value for baseQ bias [1e-100]
-3 FLOAT   min P-value for mapQ bias [0]
-4 FLOAT   min P-value for end distance bias [0.0001]
-e FLOAT   min P-value for HWE (plus F<0) [0.0001]
```

### vcftools

```
SNP-site-filtering options
--max-missing    FLOAT    Minimum calling rate
--max-meanDP     INT      Maximum mean read depth
--maf            FLOAT    Minimum allele frequency
Genotype-filtering options
--minQ           INT      Minimum base quality
--minGQ          INT      Minimum genotype quality
--minDP          INT      Minimum read depth
```

### ParalogFilter

```
options
-F  FLOAT   Minimum heterozygous rate based on genotypes of haplotypes
```

# Optimization of GBS-SNP calling

- GQ (genotype quality)

- ParalogFilter

9.2‰ replicate discordance rate
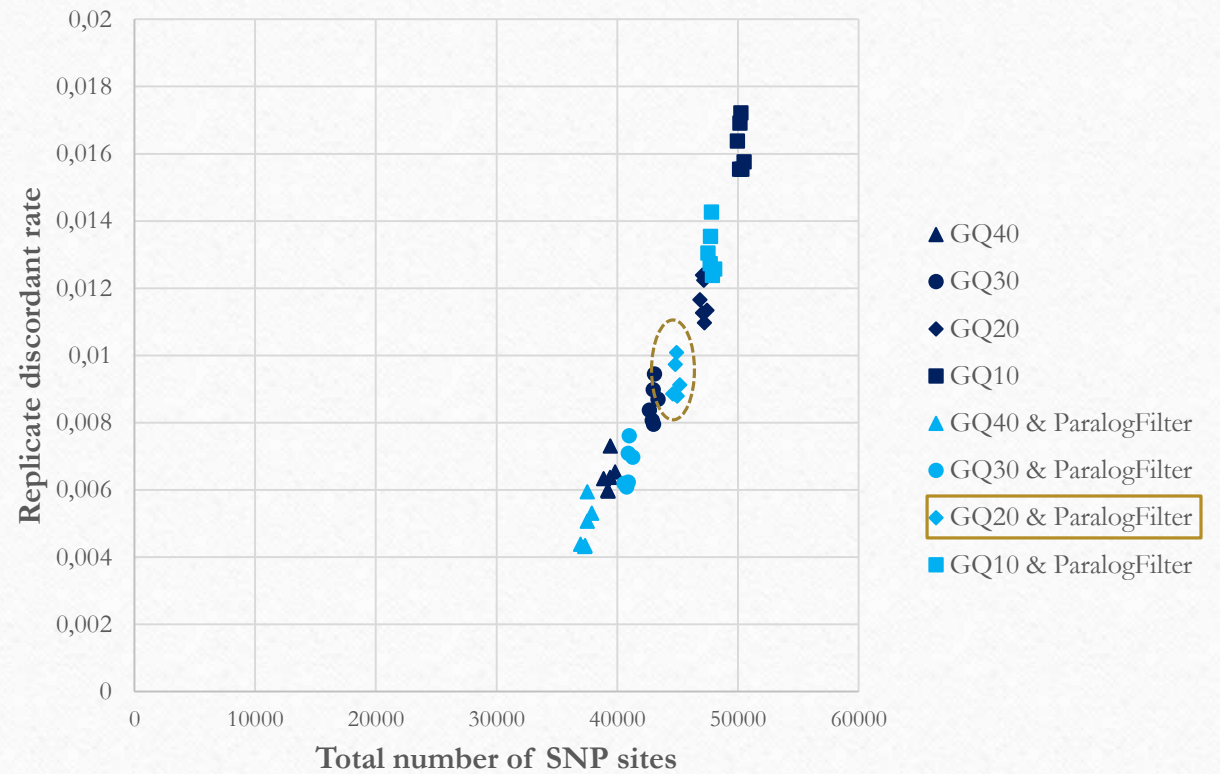
# EC-SNP-calling pipeline
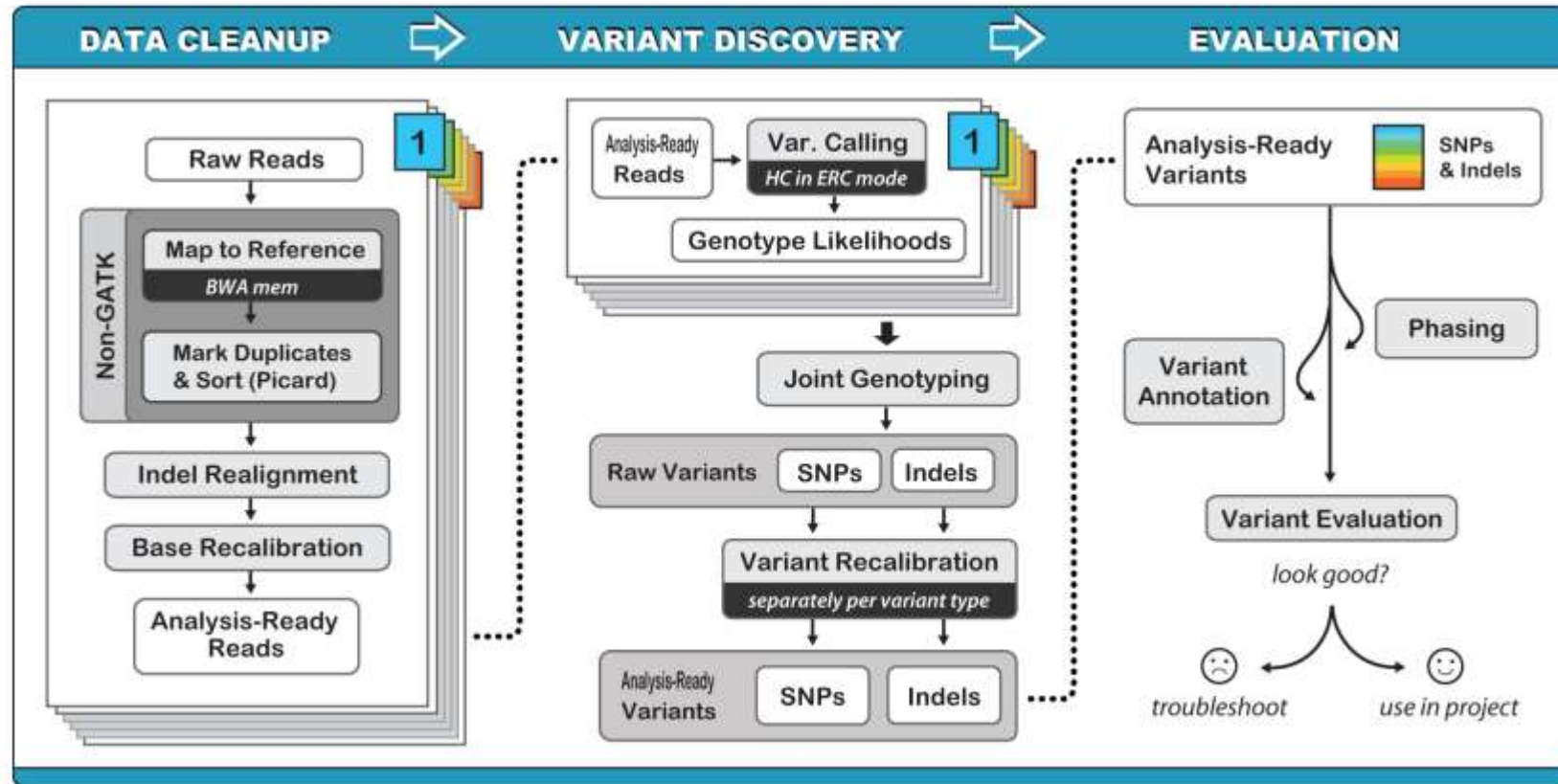


- **BWA mem**
- **Picard**
- **samtools**
- **GATK**

# Optimization of EC-SNP calling

- GQ (genotype quality)

- minDP (minimum read depth)

- ReplicateErrFilter
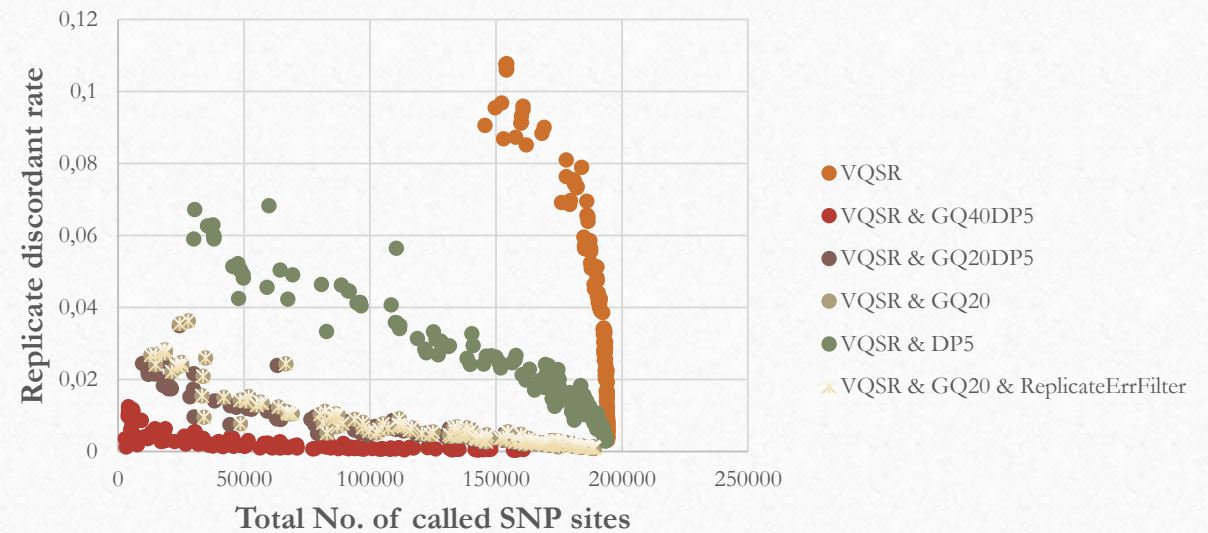


| Filtering stratgy | Mean replicate discordance rate | No. of SNPs with <0.5 missing rate |
|---|---|---|
| VQSR | 4% | 204726 |
| VQSR & minDP=5 | 2.5% | 204265 |
| VQSR & minGQ=20 | 0.95% | 194886 |
| VQSR & minDP=5 & minGQ=20 | 0.85% | 192635 |
| VQSR & minDP=5 & minGQ=40 | 0.28% | 106714 |
| VQSR & minGQ=20 & ReplicateErrFilter | 0.80% | 187496 |

# Results

1. Optimization of GBS-SNP-calling

2. Comparison of GBS and EC data

# 1 - Mapping results

| | GBS  (Median ± SD) | EC  (Median ± SD) |
|---|---|---|
| Total No. | 1,879,903 ± 5,104,604 reads | 4,897,552 ± 2,597,783 reads |
| Mapped reads | 1,313,744 ± 4,064,215 reads | 4,821,132 ± 2,584,842 reads |
| properly_paired | 1,156,043 ± 3,569,836 reads | NA |
| MQ0 reads | 94,272 ± 249,546 reads | 504,494 ± 253,716 reads |
| | | |
| Mean covered bases (5×) | 2,037,017 ± 1,104,354 bp | 19,249,840 ± 6,473,049 bp |
| Mean coverage depth (5×) | ×58 ± 72 | ×16 ± 5 |

# 2 – Coverage bias
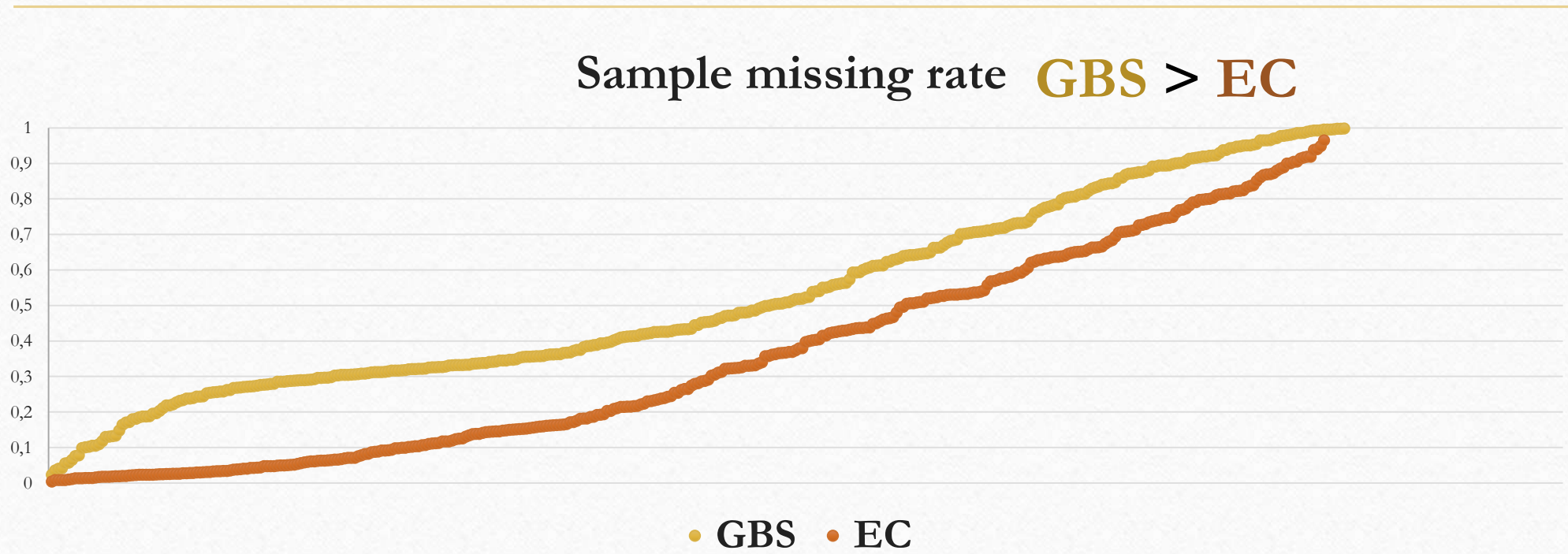
# 3 - Replicate discordance

After optimization of SNP-calling

- GBS ~9.2‰ replicate discordance rate

- EC ~ 8 ‰ replicate discordance rate

# 4 - SNP data

Individual Filter

- Maximum missing rate: 90%

SNP site Filter

- Minimum calling rate: 50%

- Minimum allele frequency: 0.05

Genotype Filter

- Minimum read depth: 5

- Minimum geotype quatily: 20

| | GBS | EC |
|---|---|---|
| Sample No. | 334 | 367 |
| SNP No. | 19,603 | 105,209 |
| Overlapped SNP No. | 572 | |
| Scaffold No. | 6,002 | 21,547 |
| Overlapped scaffold No. | 2,279 | |
| Mean SNP No. per scaffold | 3 ± 3 | 5 ± 4 |
| SNP density (per 1000bp) | ~10 | ~5 |
| Proportion of missing | 27.3% | 26.5% |

# 5 – SNP annotation

GBS

**Number of effects by functional class**

| Type (alphabetical order) | Count | Percent |
|---|---|---|
| MISSENSE | 1,683 | 34.014% |
| NONSENSE | 48 | 0.97% |
| SILENT | 3,217 | 65.016% |

Missense / Silent ratio: **0.5232**
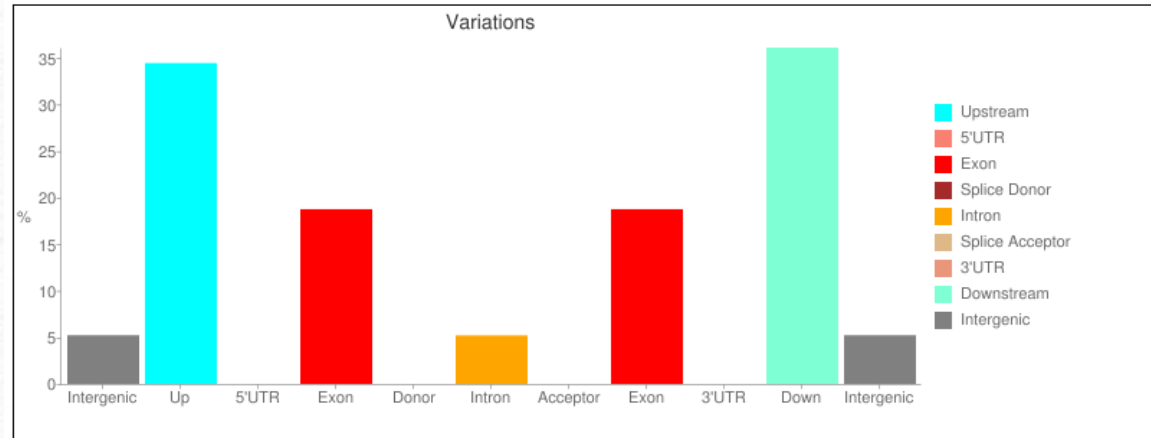
EC

**Number of effects by functional class**

| Type (alphabetical order) | Count | Percent |
|---|---|---|
| MISSENSE | 29,835 | 60.793% |
| NONSENSE | 968 | 1.972% |
| SILENT | 18,273 | 37.234% |

Missense / Silent ratio: **1.6327**
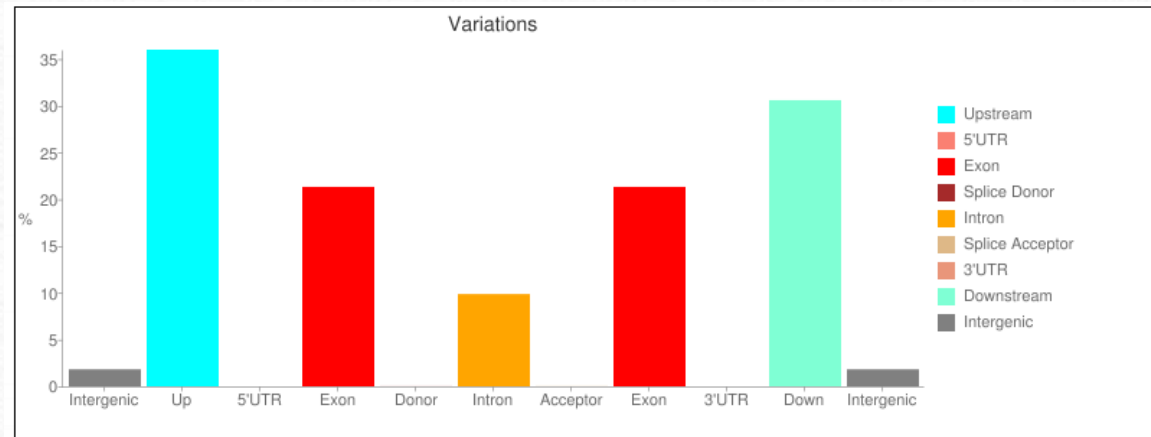
# 6 - Proportion of heterozygou genotypes
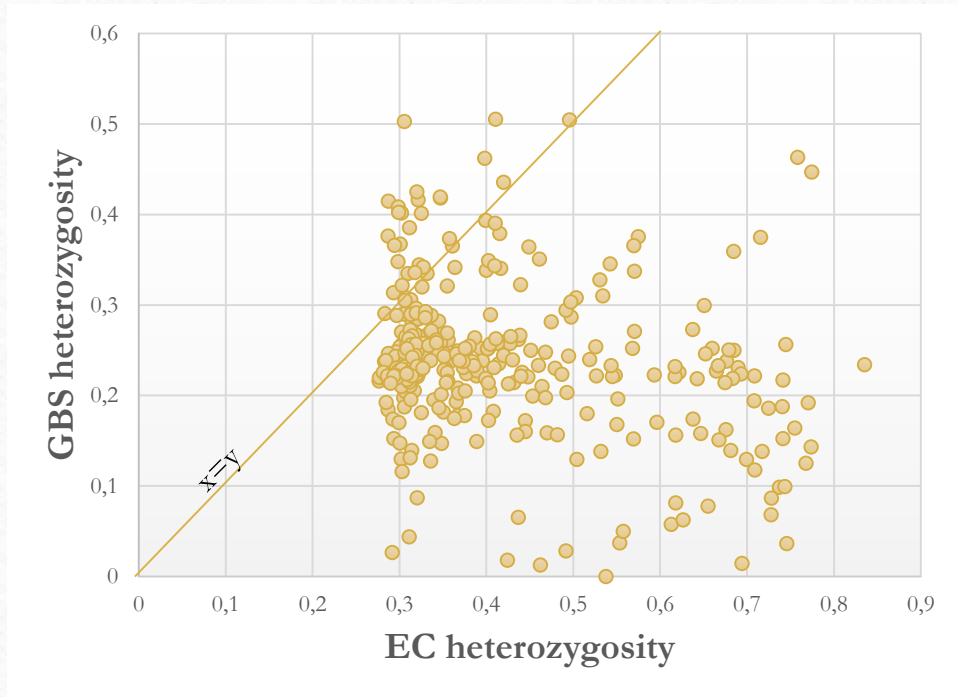


## EC > GBS

**Under-calling of *heterozygous* genotypes (false homozygous genotypes): ~8%**

# 7 – Relatedness analyses

- PCA

- Genetic distance

- Kinship　（under EM and LD kNNi imputation）
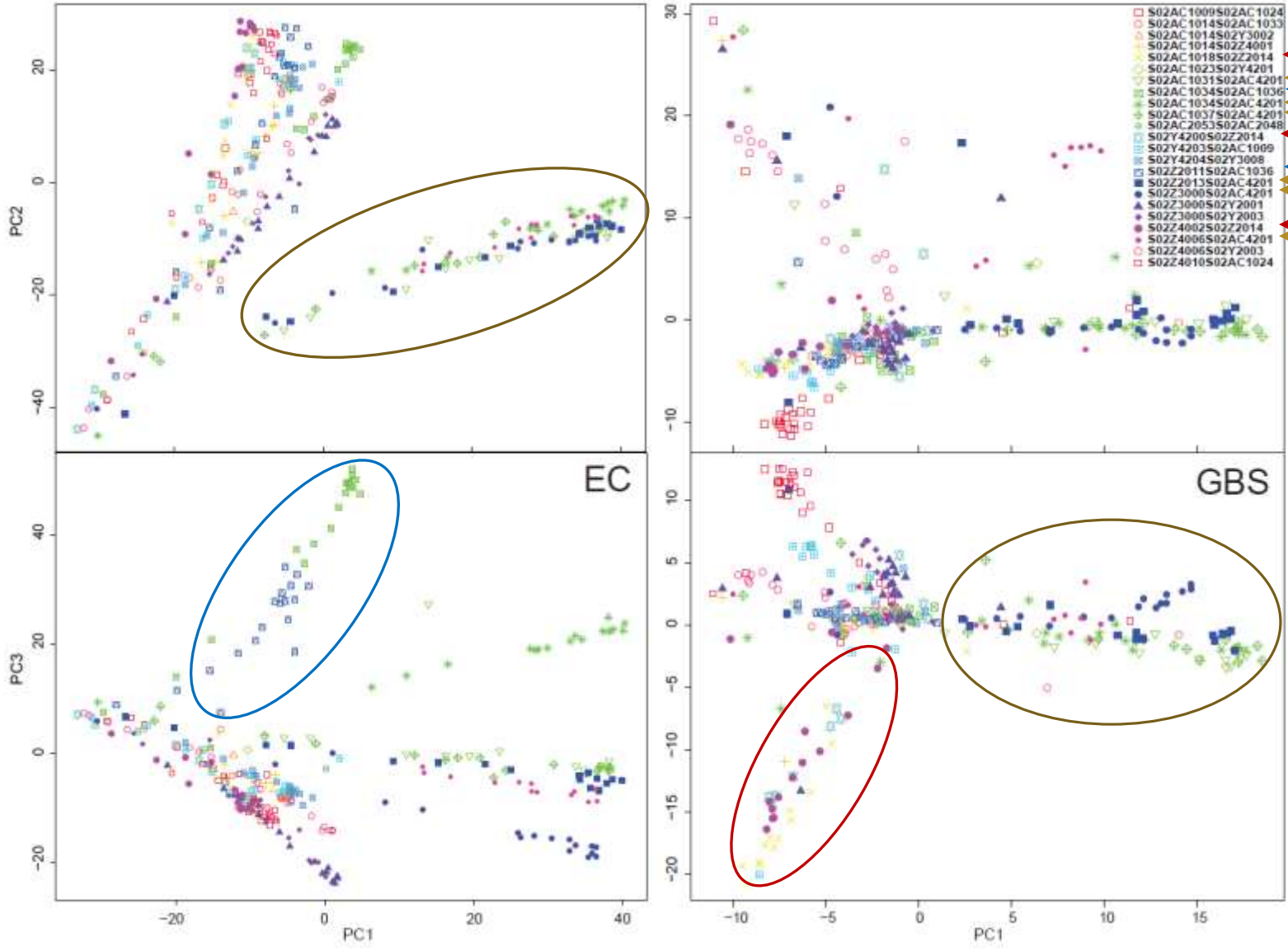
PCA analysis

EC-NJtree
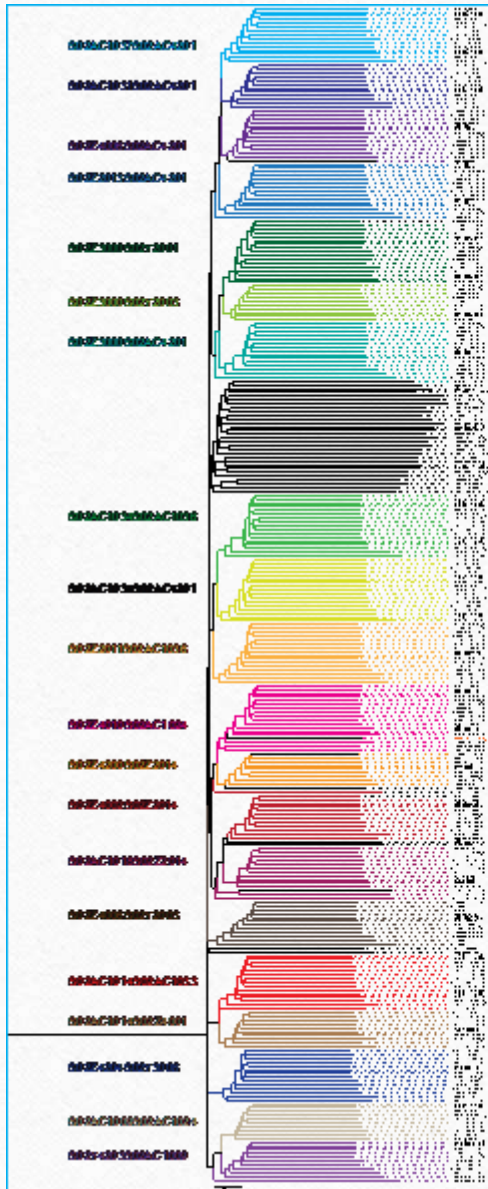
GBS-NJtree
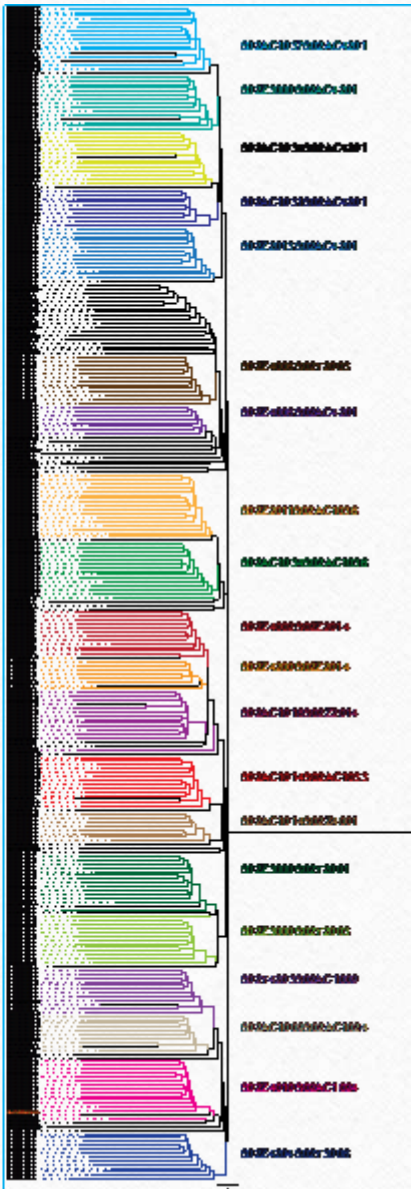
Genetic distance

# 7 – Relatedness analyses

- PCA

- Genetic distance

- Kinship （under EM and LD kNNi imputation）

# Kinship
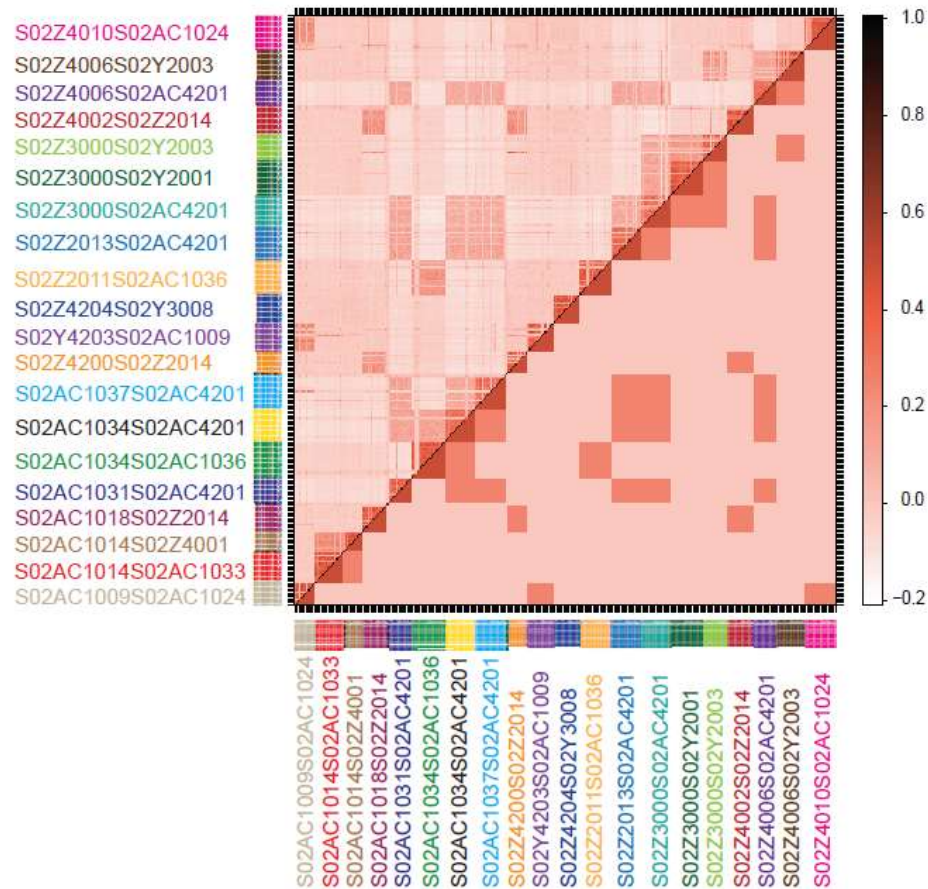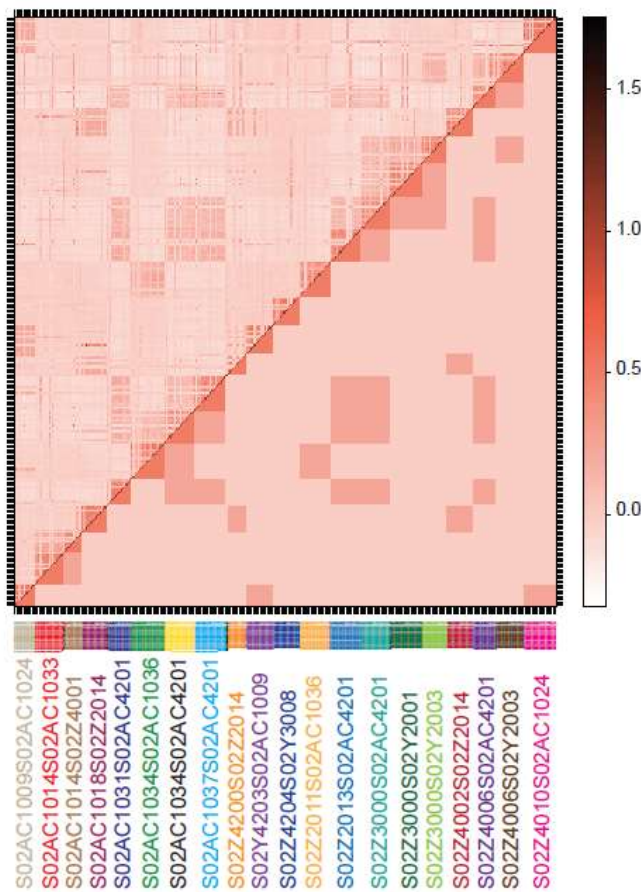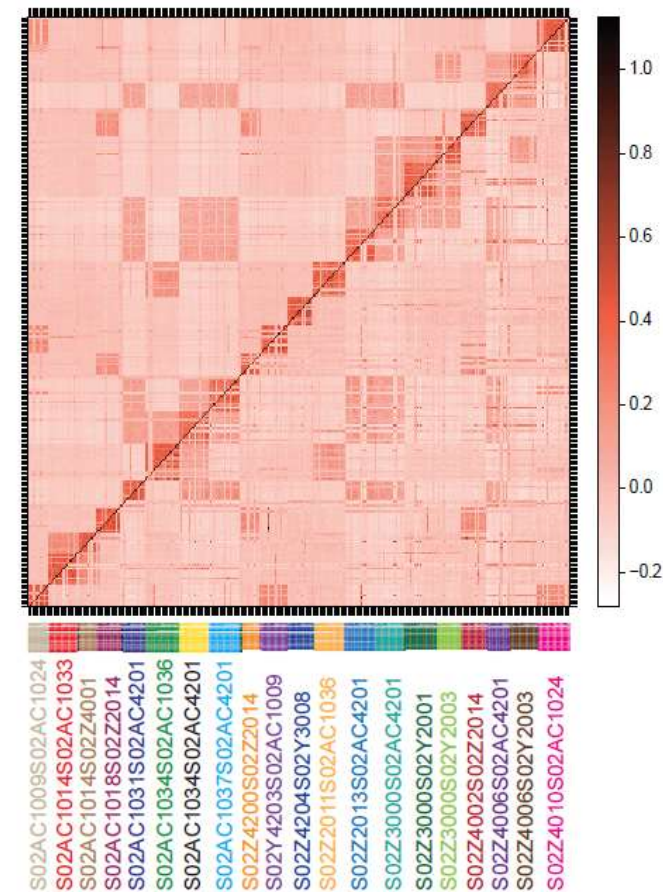


Imputation method: EM (Endelman and Jannink 2012)

The RMS error < 0.01

# Mean genomic relationship within and between family based on the EC-SNP data

Imputation method: **EM** (Endelman and Jannink 2012) with RMS error < 0.01 & **LD-kNNi** (Money et al., 2015)



Brown number: p-value <0.01 in paired t-test
Deep yellow: ful-sib family
Light yellow: half-sib family

EC
Mean relative difference: 0.147

GBS
Mean relative difference: 0.181

# Mean genomic relationship within and between family based on the EC- and GBS-SNP data

Imputation method: **EM** (Endelman and Jannink 2012) with RMS error < 0.01

Deep yellow: ful-sib family
Light yellow: half-sib family

ful-sib family (0.5):
GBS   0.377 – 0.141
EC    0.394 – 0.195
half-sib family (0.25):
GBS   0.147 – 0.036
EC    0.183 – 0.072

EM
Mean relative difference: 0.775

LD kNNi
Mean relative difference: 0.812

# GBS Summary & Next work

- Coverage bias

- ~ 8% underestimation of herterozygous genotypes

- Relatively high proportion of genetic variances in non-coding regions

- The number of covered bases is less than predicted

- Obvious difference between GBS- and EC- G matrixes

✓ Develop predictive model and evaluate predictive ability

✓ Estimate the impact of error genotypes in Genomic prediction

✓ … …

# Acknowledgement



**Forest Genetic Group in UPSC**

Thanks for your attention!