GWAS analysis report

SNP data preparation

The original SNP data is stored in a single VCF file and was generated according to the following metrics:

- SNPs with minor allele frequency less than 5% were filtered out and
- SNPs with maximum missing frequency higher than 10% were removed.

The filtered VCF file from KnowPulse needed further conversion as the chromosome information was missing. Thus, contigs/scaffolds needed to be mapped to genome after the file was generated. However, there were still certain number of contigs/scaffolds unmapped after the conversion. The total number of SNPs is 20811, while the numbers of contigs and scaffolds are 2418 and 694, respectively. Notably, all the names of contigs are preceded with chromosome numbers. While it was neither safe to directly replace such contigs with chromosome numbers nor was it wise to ignore all of them, such contigs were mapped to faked chromosomes for future analysis. For instance, if a SNP comes from a contig whose chromosome number is 1, we mapped the contig to 8, i.e., 1 + 7, given that there are 7 chromosomes in lentil genome. The remaining unmapped scaffolds were mapped to Chromosome 15.

On top of that, PLINK was applied to transform the data format into bed/bim/fam/ped/map, which could be accepted by the mainstream GWAS software packages, *e.g.*, PLINK, FaST-LMM, TASSEL, *etc.*

Since samples coming from a same family yet growing in different regions have only one copy of SNPs, duplication of SNPs for samples is needed. By doing so, we assumed that the genetic materials among samples within one family were equivalent (needs citation here).

Covariant data preparation

Other than the two different types of environmental data, *i.e.*, precipitation and day length, seed weight was also taken into account. To comply with file format required by GWAS packages, *i.e.*, PLINK and FaST-LMM, PLINK was employed to aggregate the genotype, covariate and phenotype data.

GWAS analysis via PLINK

PLINK is the gold standard to do GWAS analysis as it is fast and is equipped with multiple functionalities (cite Alarcon *et al.*), *e.g.*, data management, summary statistics, population stratification and association analysis (cite Purcell *et al.*). In our case, other than the data management functionality that was applied to transform the VCF into other file formats, summary statistics was applied to describe the genome coverage, whereas the association analysis tool was employed to examine how genes interact with phenotypes, *i.e.*, days to emergence and growing degree days.

Since the purpose of GWAS is to tell which SNPs are playing significant roles in our observed traits, association analyses of both quantitative and binary traits were conducted in this study. The distributions of the two phenotypic datasets are shown in Fig 1.

The medians of DTE and GDD are 19 and 191.4025, respectively. To do case-control studies over the two traits, samples with DTE less than 19 were grouped into the affected and otherwise the unaffected, whereas samples with GDD less than 191.4025 were grouped into the affected and otherwise the unaffected. The results are shown in Fig 2.

Table 1 gives the top 6 significant SNP information from PLINK quantitative association study on days to emergence.

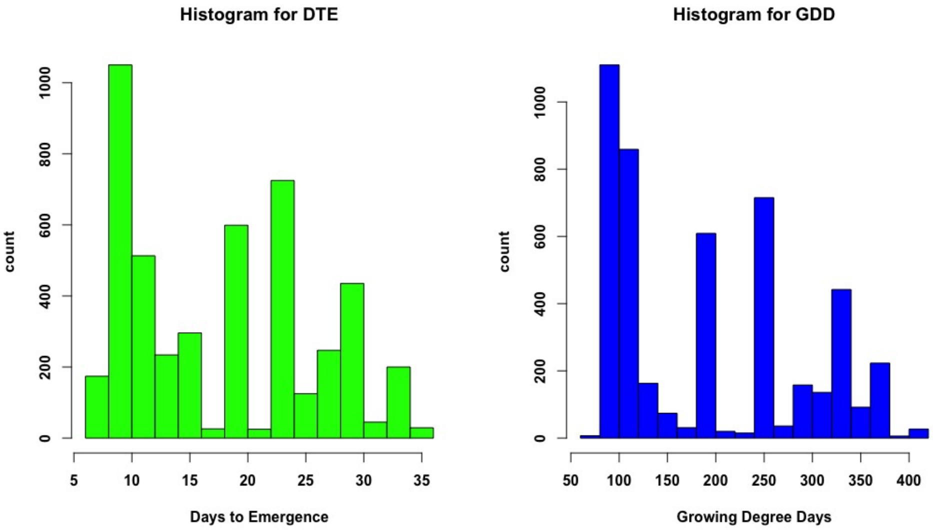


Figure 1: Histograms of DTE and GDD datasets

Chromosome	Physical Position	Regression Coefficient	Standard Error	Regression r^2	p-value*
4	163844531	07983	0.1473	0.006817	6.290e-8
1	291984120	0.9569	0.1827	0.006173	1.711e-07
1	192789217	0.9287	0.1839	0.005688	4.571e-07
4	243535564	0.6334	0.1257	0.006004	4.860e-07
4	243535565	0.6334	0.1257	0.006004	4.860e-07
1	291941752	0.8953	0.1779	0.005745	5.002e-07

* Wald Test asymptotic p-value

Table 1: Top SNPs from PLINK quantitative association study on days to emergence

Table 2 gives the top 6 significant SNP information from PLINK quantitative association study on growing degree days.

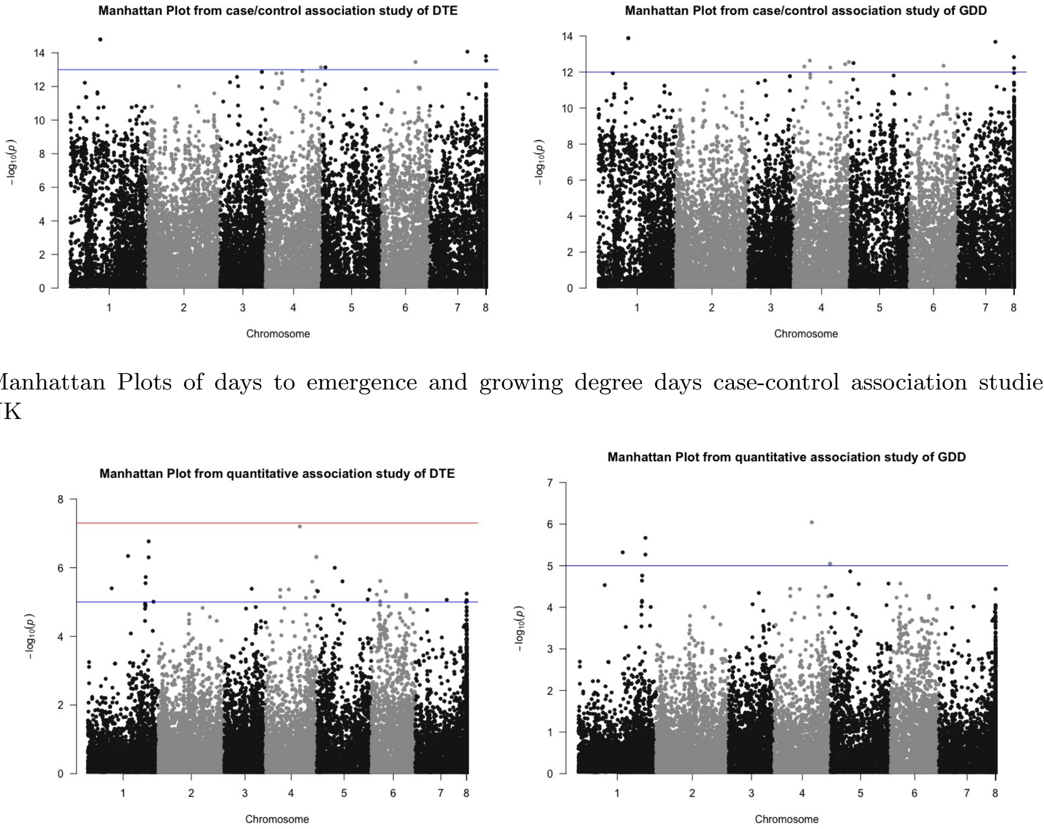
Chromosome	Physical Position	Regression Coefficient	Standard Error	Regression r^2	p-value*
4	163844531	8.597	1.748	0.005621	9.042e-07
1	291984120	10.300	2.170	0.005075	2.143e-06
1	192789217	10.000	2.183	0.004681	4.775e-06
1	291941752	9.622	2.113	0.004707	5.409e-06
4	243535564	6.637	1.493	0.004677	9.001e-06
4	243535565	6.637	1.493	0.004677	9.001e-06

* Wald Test asymptotic p-value

Table 2: Top SNPs from PLINK quantitative association study on growing degree days

Table 3 gives the top 6 significant SNP information from PLINK case-control association study on days to emergence.

Table 4 gives the top 6 significant SNP information from PLINK case-control association study on growing degree days.



(a) Manhattan Plots of days to emergence and growing degree days case-control association studies from PLINK

(b) Manhattan Plots of days to emergence and growing degree days quantitative association studies from PLINK

Figure 2: Manhattan Plots of association studies from PLINK

Chromosome	Physical Position	A1[1]	F_A [2]	F_U[3]	A2[4]	χ^2	p-value
1	130505078	T	0.4709	0.3852	C	63.52	1.585e-15
1	130505103	C	0.4709	0.3852	T	63.52	1.585e-15
1	130505132	T	0.4709	0.3852	C	63.52	1.585e-15
7	165628331	C	0.3657	0.2869	T	60.22	8.464e-15
6	148586244	T	0.3954	0.3185	G	57.41	3.547e-14
5	16655279	G	0.4747	0.3937	A	56.01	7.228e-14

[1] Minor allele code
[2] Case allele frequency
[3] Control allele frequency
[4] Major allele code

Table 3: Top SNPs from PLINK case-control study on days to emergence

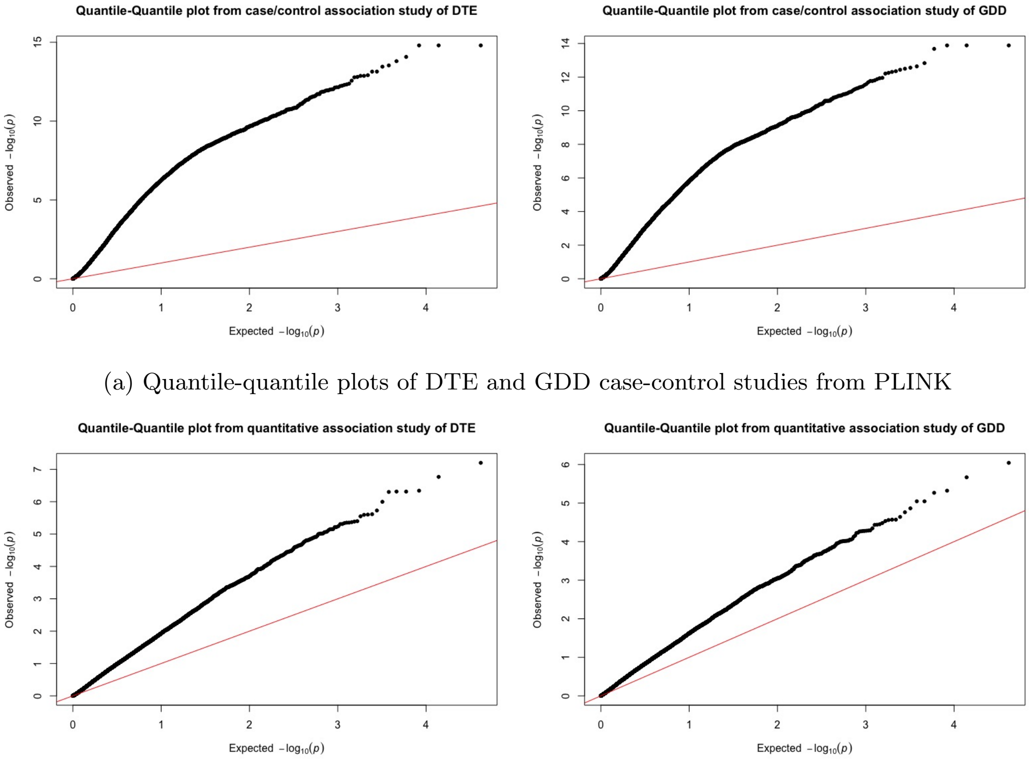
To examine how the observed p-values follow an expected pattern, quantile-quantile plots are shown in Fig 3.

Although PLINK returned with significant SNPs, as shown in Fig 2a and Fig 2b, the observed p-values tend to be diverged from the expected pattern, as shown in Fig 3a and Fig 3b. Hence, further

Chromosome	Physical Position	A1[1]	F_A [2]	F_U[3]	A2[4]	χ^2	p-value
1	130505078	T	0.4702	0.3873	C	59.38	1.299e-14
1	130505103	C	0.4702	0.3873	T	59.38	1.299e-14
1	130505132	T	0.4702	0.3873	C	59.38	1.299e-14
7	165628331	C	0.3657	0.2881	T	58.45	2.081e-14
4	71465726	T	0.4738	0.3955	G	53.74	2.293e-13
4	242435083	G	0.5103	0.4320	T	53.38	2.742e-13

[1] Minor allele code
[2] Case allele frequency
[3] Control allele frequency
[4] Major allele code

Table 4: Top SNPs from PLINK case-control study on growing degree days



(a) Quantile-quantile plots of DTE and GDD case-control studies from PLINK

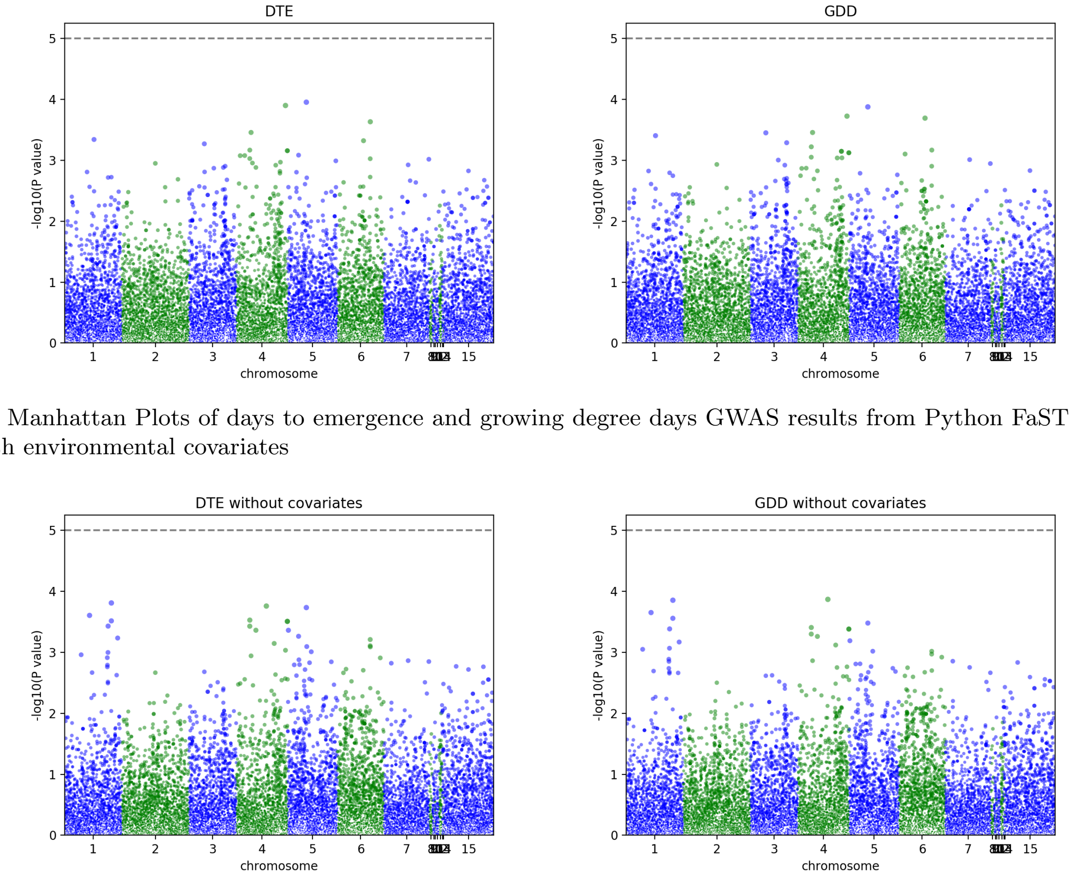
(b) Quantile-quantile plots of DTE and GDD quantitative association studies from PLINK

Figure 3: Quantile-quantile plots of association studies from PLINK

investigation on the SNPs needs to be done to determine whether or not they are truly significant in lentil DTE and GDD.

GWAS analysis via Python FaST-LMM

The Python FaST-LMM package does GWAS based on a different model (add more details of the FaST-LMM and cite the FaST-LMM paper). Currently, FaST-LMM has gained popularity in lentil GWAS analysis due to the small genome size of lentil. However, unlike PLINK, the Python FaST-LMM package does not output interaction results between genes and environmental factors (cite Alarcon *et al.*). The Manhattan Plot of SNPs with p-values is shown in Fig 4

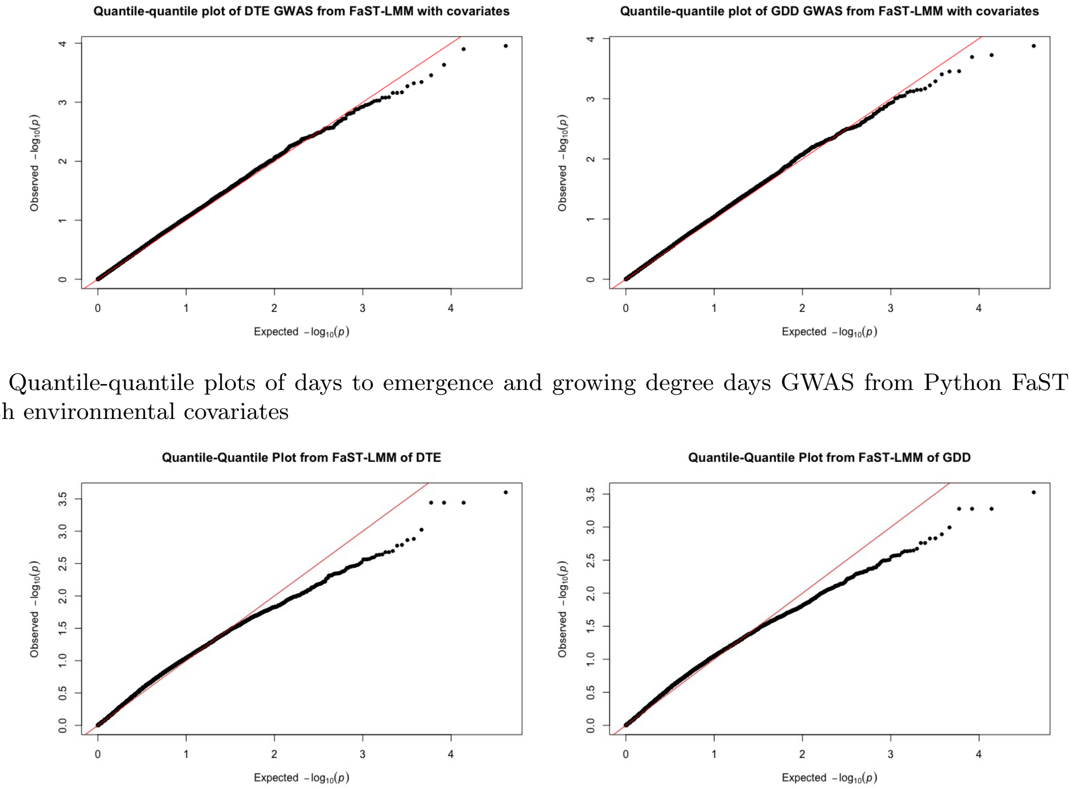


(a) Manhattan Plots of days to emergence and growing degree days GWAS results from Python FaST-LMM with environmental covariates

(b) Manhattan Plots of days to emergence and growing degrees days GWAS results from Python FaST-LMM without environmental covariates

Figure 4: Manhattan Plots of GWAS results from Python FaST-LMM

According to Fig 4a and Fig 4b, it can be told that there are basically no significant SNPs in either of the two traits, which is opposite to that from PLINK. Additionally, the quantile-quantile plots shown in Fig 5a and Fig 5b indicate the results from Python FaST-LMM package follow the expected pattern and should be trusted.



(a) Quantile-quantile plots of days to emergence and growing degree days GWAS from Python FaST-LMM with environmental covariates

(b) Quantile-quantile plots of days to emergence and growing degree days GWAS from Python FaST-LMM without environmental covariates

Figure 5: Quantile-quantile plots of GWAS results from Python FaST-LMM