

國立臺灣大學生物資源暨農學院農藝學系暨研究所

博士論文



Department of Agronomy  
College of BioResources and Agriculture  
National Taiwan University  
Doctoral Dissertation

醋栗番茄全基因體分子標誌之開發與  
探究控制番茄雄蕊長度之候選基因

Development of Genome-Wide High-Density SNP Markers in  
*Solanum pimpinellifolium* and Investigation of Candidate Loci  
of Stamen Length in Tomato

林亞平

YA-PING LIN

指導教授：陳凱儀 博士

Advisor: KAI-YI CHEN, Dr.

中華民國 108 年 04 月

April 2019

## 謝辭



在博士班的求學生活中經歷許多挫折與自我懷疑，我想在論文完成之際，謝謝那些一直在身旁支持我、且願意相信我的人。首先，我想感謝我的指導老師陳凱儀博士，陳老師對研究的熱誠不僅開拓我的研究視野，進而讓我跟進當代國際研究水平，尤其老師在學術研究上認真仔細與堅持，更影響了我追求學問的態度。

其次，我要感謝我的論文指導小組及口試委員，包括胡凱康博士、常玉強博士、鍾國芳博士、董致韡博士、林耀正博士及李承叡博士，謝謝他們對這篇論文的批評與指教，任何一項建議都促成這篇論文更加精進。我更要感謝系上許多關心我、支持我的老師，老師們不僅樂於與我交流學術上的意見，也願意在我迷惘的時候鼓勵我，使我常能保持對研究生生活的積極與樂觀。我也要感謝系辦的行政人員，除了貼心提醒我申請獎學金補貼生活外，也在我畢業之際幫忙處理許多行政瑣事，讓我心無旁騖地準備口試。我要謝謝台大農場的大哥們，提供我許多田間操作的建議，讓我從田間菜鳥變成半個田間專家。此外，我也要感謝平常關心我的學長姐、同學及學弟妹們，他們提供我許多學業上的協助，更不吝於分享他們的生活目標或日常趣事，為外人眼中枯燥無味的研究生活增添許多樂趣。我還要感謝那些非學術圈的朋友們，讓我能在研究室以外的時間看到時代的潮流，懂得享受非學術研究的生活。

最後，我要感謝一路以來陪伴我的家人，他們都覺得「博士們」不好相處，但即便如此，仍全力支持我選擇這條路，願我往後能謹記他們的叮嚀，做一個謙遜有禮的讀書人。




## 摘要

自達爾文提出演化論後，異型花的遺傳機制一直是植物學家感興趣的議題之一。古典的研究認為控制自交不親和性與異花型的基因緊密連鎖，稱之為 S 基因座，這些調控相關性狀且緊密連鎖的基因稱為超級基因。一般認為植物從異交演化成自交的過程中會先失去自交不親和性，而後在超級基因內發生重組，使得異型花變成同型花，以確保自交成功的機率。然而伴隨著分子技術的進步，現今的分子證據卻顯示同型花可能是由半合子造成，而非傳統上認為的罕見重組。在農業上，研究此議題可以了解作物在馴化過程中，受到強烈選拔壓力後對基因體造成的改變。此外，也可控制作物的自交不親和性與花型，如此不僅能有效地生產雜交種子，也能藉由提高授粉率而增加產量。

醋栗番茄為野生番茄的一種，原生於秘魯與厄瓜多沿岸，是栽培番茄的近親。因為醋栗番茄具有許多抗病性狀，且可與栽培種番茄相互雜交，故為重要的番茄種原之一。目前醋栗番茄已提供番茄育種工作上一些抗病基因座，也應用於農藝性狀相關的全基因體關聯性定位中。前人研究發現醋栗番茄可分成異交、自交與中間型三種交配系統，異交的醋栗番茄不僅具有較高的遺傳歧異度，且具有較突出的雌蕊。由於醋栗番茄在交配系統與花的形態上都具有多型性，故很適合利用於自交不親和性與異型花的研究。

現今分子標誌已廣泛地應用於作物育種上，伴隨著次世代定序的成本下降，開發重要種原的全基因體分子標誌已成為基礎的育種工作。本研究針對 99 個醋栗




番茄收集系進行 *Pst*I 限制酶關聯性定序，定序範圍涵蓋 12,790 個基因。我們一共得到 24,330 個單一核苷酸多型性分子標誌，其中 16,365 個分子標誌座落於 7,383 個基因上。我們觀察到定序範圍與基因的分佈類似，顯示使用 *Pst*I 限制酶來篩選基因體片段的策略適合應用於尋找候選基因的研究上。此外，該族群可以分成三個先祖次族群及四個混合基因體的次族群。主成份分析、成對  $F_{st}$ 、AMOVA 皆支持這樣的分群，顯示這組高密度分子標誌可穩定估計族群結構。接著，我們估計該族群整體的連鎖衰變在 18 千鹼基對，意味著此族群可以在全基因體關聯性分析得到精密的解析度，甚至可以定位到單一基因。然而要滿足這樣的解析度，至少需要 50,000 個分子標誌。

在雄蕊長度的全基因體關聯性定位中，我們利用 98 個醋栗番茄收集系進行混合線性模式分析，定位到三個候選基因座，但這三個基因座皆為高度錯誤發現率。由於全基因體關聯性定位的檢定力及錯誤發現率皆與研究樣本的族群大小有關，我們建議兩種增加樣本的方式，一是在各個次族群中均勻地增加取樣數目，這個方法也可能使罕見對偶基因變成一般對偶基因，故也可能增加分子標誌的數目。另一個方法是在秘魯北部增加取樣數目，因為此處是醋栗番茄的發源地，遺傳歧異度大，也可能增加對偶基因數。

另一方面，前人研究顯示 *style2.1* 下游附近有兩個控制雄蕊長度的基因 *stamen2.2* 及 *stamen2.3*，我們利用轉錄體組定序來挑選這兩個候選基因，使用的材料是栽培種番茄品系 M82 及其滲透系 TA3178，TA3178 在 *style2.1* 附近是野生番茄（潘那利番茄）的染色體片段。我們藉由單一核苷酸多型性的數量差異來界定

滲透片段的範圍。接著，依據本研究室之前的結果，我們篩選從標誌 cLED19A24 到 CT9 該區間的 18 個候選基因，比較這些候選基因的表現量及多型性後，發現 Solyc02g087960.2、Solyc02g087970.1 與 Solyc02g088070.2 可能是 *stamen2.2* 及 *stamen2.3* 的候選基因。

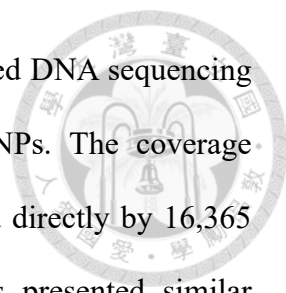
## Abstract



Botanists have been fascinated by the genetic mechanism of heterostyly since Darwin's theory of evolution. It was believed that the genes controlling self-incompatibility and floral morphology were linked tightly, so-called *S-locus*. According to the classical evolutionary studies, when a plant evolved from outcrossing to selfing, it was necessary to lose self-incompatibility and then adjusted the positions of male and female floral organs through the rare recombination within the *S-locus*. However, new evidence suggested that homostyly resulted from hemizygote rather than the rare recombination. In agriculture, studying the genetic mechanism of self-incompatibility and heterostyly can understand the changes of crop genomes under the selection forces during domestication processes. Additionally, it can accelerate the production of hybrid seeds or ensure the pollination to increase yield.

*Solanum pimpinellifolium* is a wild tomato originated from the coastal region of Peru and Ecuador. It serves as an important germplasm in tomato breeding programs because it displays many resistant traits and can freely cross to cultivated tomatoes. Previous studies classified this species as complete or near complete allogamy, complete autogamy and intermediate type based on its mating system. In addition, allogamous accessions displayed higher genetic diversity and more exertion of stigma than autogamous ones. Because *S. pimpinellifolium* contains the variations of outcrossing rate and floral morphology within its own species, it could be an ideal material to study the genetic mechanism of self-incompatibility and heterostyly.

Nowadays, molecular markers have been applied to crop breeding extensively. Accompanying by the cost down of next generation sequencing, the development of genome-wide high-density markers for germplasm becomes essential in breeding



programs. In this research, we performed the *Pst*I-digested associated DNA sequencing for 99 accessions of *S. pimpinellifolium*, resulting in 24,330 SNPs. The coverage extended to 12,790 genes, and a total of 7,383 genes were targeted directly by 16,365 SNPs. Besides, the sequencing regions and the annotated genes presented similar distributions through each chromosome. This suggested that *Pst*I-digested associated DNA sequencing was an appropriate strategy to investigate candidate genes. This collection was divided into three subpopulations of single-ancestral genome and four subpopulations of mix-ancestral genome by ADMIXTURE. Principle component analysis, pairwise  $F_{st}$  and AMOVA all supported the subpopulations, implying this set of high-density markers was capable to estimate the subpopulations stably. Moreover, the overall LD decay was within 18 Kb, suggesting a fine resolution in genome-wide association study even to a single-gene level. However, to achieve such fine resolution, at least 50,000 markers were required.

Three candidate loci controlling stamen length were identified via the mixed linear model in genome-wide association study of 98 *S. pimpinellifolium* accessions, but all three loci presented high false discovery rate. Since the power and false positive rate of genome-wide association study depend on the sample size of a studying population, we suggest two approaches to increase sample size. One is to increasing samples in each subpopulation evenly. This approach can potentially make rare alleles to common alleles by increasing the allele frequency. The other is to sampling more individuals in the northern Peru because the accessions in the northern Peru present more genetic diversity. This approach can also increase both rare alleles and common alleles.

On the other hand, following the previous studies, *stamen2.2* and *stamen2.3* were located in the downstream interval next to *style2.1*. We performed a RNA sequencing

experiment of M82 and TA3178. TA3178 is an introgression line of M82 and contains a segment of *Solanum pennellii* near *style2.1*. We identified this introgression region by comparing the difference of SNPs between these two lines. Afterwards, following the previous work in our team, we screened 18 candidate genes from marker cLED19A24 to CT9 by comparing the fold change and cDNA polymorphism between M82 and TA3178. This result suggested that Solyc02g087960.2, Solyc02g087970.1 and Solyc02g088070.2 should be the candidates of *stamen2.2* and *stamen2.3*.





## Contents

謝辭.....	II
摘要.....	IV
<b>ABSTRACT .....</b>	<b>VI</b>
<b>LIST OF FIGURES.....</b>	<b>XIII</b>
<b>LIST OF TABLES.....</b>	<b>XIV</b>
<b>LIST OF SUPPLEMENTARY DATA .....</b>	<b>XV</b>
<b>CHAPTER 1 INTRODUCTION .....</b>	<b>1</b>
1.1 HETEROSTYLY .....	1
1.1.1 Evolution of heterostyly.....	1
1.1.2 Heterostyly in tomato species.....	2
1.2 SOLANUM PIMPINELLIFOLIUM .....	3
1.2.1 The mating systems and flower characters in <i>S. pimpinellifolium</i> .....	3
1.2.2 <i>S. pimpinellifolium</i> is a diverse and attractive tomato germplasm.....	4
1.2.3 The population differentiation of <i>S. pimpinellifolium</i> .....	4
1.2.4 The genetic diversity of <i>S. pimpinellifolium</i> .....	5
1.3 GENOME-WIDE ASSOCIATION STUDY.....	6
1.3.1 The concept of GWAS.....	6
1.3.2 LD determines the resolution of GWAS.....	7
1.3.3 Population structure and kinship cause confounding effects in GWAS .....	8
1.4 NEXT GENERATION SEQUENCING (NGS) TECHNOLOGY .....	10
1.4.1 Restriction-site associated DNA sequencing.....	10
1.4.2 RNA sequencing .....	11
1.5 DEVELOPMENT OF STAMEN.....	12
1.5.1 MADS box genes determine stamen differentiation .....	12
1.5.2 Phytohormones regulate the stamen development .....	12
1.6 CONCLUSION .....	13

1.7 REFERENCE .....	14
---------------------	----

**CHAPTER 2 ASSESSMENT OF POPULATION DIFFERENTIATION AND LINKAGE DISEQUILIBRIUM IN *SOLANUM PIMPINELLIFOLIUM* USING GENOME-WIDE HIGH-DENSITY SNP MARKERS.....25**

2.1 PURPOSE.....	25
2.2 MATERIAL AND METHOD .....	25
2.2.1 <i>Plant materials</i> .....	25
2.2.2 <i>RAD sequencing</i> .....	26
2.2.3 <i>SNP calling</i> .....	26
2.2.4 <i>Population differentiation</i> .....	27
2.2.5 <i>Isolation by distance</i> .....	28
2.2.6 <i>Estimate of genetic variation and LD</i> .....	28
2.2.7 <i>Analysis of SolCAP array data of <i>S. pimpinellifolium</i></i> .....	28
2.3 RESULT .....	29
2.3.1 <i>Identification of 24,330 SNPs from PstI-digested DNA libraries</i> .....	29
2.3.2 <i>A similar distribution between genes and SNPs was identified in the vicinity of PstI cutting site throughout the genome</i> .....	31
2.3.3 <i>Genetic differentiation of <i>S. pimpinellifolium</i> was corresponding to the geographic area</i> .....	32
2.3.4 <i>Meta-analysis of SolCAP genotyping array resulted in 15 subpopulations</i> ...35	
2.3.5 <i>Rapid LD decay</i> .....	36
2.3.6 <i>Heterogeneity of genetic recombination within each chromosome</i> .....	37
2.4 DISCUSSION.....	38
2.4.1 <i>Subpopulations clustering from north to south are expected due to the high correlation between genetic distance and geographic distance</i> .....	38
2.4.2 <i>Discrepancy of genetic clustering in SolCAP meta-analysis</i> .....	40
2.4.3 <i>More markers are required to cover through the genome of <i>S. pimpinellifolium</i></i> .....	41
2.5 REFERENCE .....	42
2.6 SUPPLEMENTARY DATA .....	47

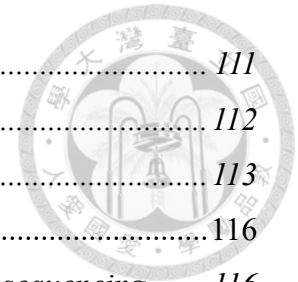
**CHAPTER 3 GWAS OF THE CANDIDATE GENES CONTROLLING STAMEN LENGTH IN *SOLANUM PIMPINELLIFOLIUM* .....73**

3.1 PURPOSE.....	73
3.2 MATERIAL AND METHOD .....	73
3.2.1 <i>Plant material and phenotyping</i> .....	73
3.2.2 <i>GWAS</i> .....	74
3.2.3 <i>Haplotype block</i> .....	74
3.3 RESULT .....	75
3.3.1 <i>SSL2.50ch06_45620556 is significant among all the GLM and MLM analysis</i> .....	75
3.3.2 <i>The LD patterns of these significant loci</i> .....	78
3.4 DISCUSSION.....	78
3.4.1 <i>QTL on chromosome 2, 3 and 7</i> .....	78
3.4.2 <i>Large sample size is essential for GWAS</i> .....	79
3.4.3 <i>r<sup>2</sup> or D' as an indicator for LD</i> .....	81
3.4.4 <i>A gap between the estimation of r<sup>2</sup> in different softwares</i> .....	82
3.4.5 <i>Insufficient coverage makes the build of haplotypes unsuccessful</i> .....	83
3.4.6 <i>More markers or more individuals</i> .....	84
3.5 REFERENCE .....	84
3.6 SUPPLEMENTARY DATA .....	88

**CHAPTER 4 THREE CANDIDATE GENES CONTROLLING STAMEN LENGTH REVEALED VIA THE TRANSCRIPTOME PROFILES OF M82 AND ITS INTROGRESSION LINE TA3178.....108**

4.1 PURPOSE.....	108
4.2 MATERIAL AND METHOD .....	109
4.2.1 <i>RNA sequencing</i> .....	109
4.2.2 <i>The boundary of introgression segment in TA3178</i> .....	109
4.2.3 <i>Differential expression analysis</i> .....	110
4.2.4 <i>The cDNA polymorphisms of the genes from cLED19A24 to CT9</i> .....	110
4.3 RESULT .....	111
4.3.1 <i>The summary of RNA-seq</i> .....	111

4.3.2	<i>The 1.1 Mb introgression segment of S. pennellii</i>	111
4.3.3	<i>Only two DEGs in the introgression segment</i>	112
4.3.4	<i>Three candidate genes of stamen2.2 and stamen2.3</i>	113
4.4	DISCUSSION	116
4.4.1	<i>M82 presented more SNPs than TA3178 due to its deeper sequencing</i>	116
4.4.2	<i>Lacking biological replications may underestimate DEGs</i>	116
4.4.2	<i>Transcription profiles and polymorphisms in the introgression segment</i>	117
4.4.4	<i>Narrow down the candidate genes of stamen2.2 and stamen2.3</i>	118
4.5	REFERENCE	119
4.6	SUPPLEMENTARY DATA	121



## List of Figures

FIGURE 2.1 THE DISTRIBUTIONS OF ITAG2.4 GENE MODEL, <i>Pst</i> I CUTTING SITES AND SNPs THROUGH WHOLE GENOME.....	32
FIGURE 2.2 ANCESTRY AND GEOGRAPHIC DISTRIBUTION OF 98 <i>SOLANUM PIMPINELLIFOLIUM</i> ACCESSIONS FROM THE TOMATO GENETICS RESOURCE CENTER.....	34
FIGURE 2.3 VISUALIZATION FOR LD. A) THE 50 KB INTERVAL OF OVERALL LD DECAY. B) THE LOCAL LD OF CHROMOSOME 1.....	37
FIGURE 3.1 THE DISTRIBUTION OF STAMEN LENGTH.....	376
FIGURE 3.2 THE GEOGRAPHIC DISTRIBUTION OF THE STAMEN CHARACTERS AMONG 98 ACCESSIONS.....	76
FIGURE 4.1 THE SNPs IN THE INTROGRESSION SEGMENT IN M82 AND TA3178.....	112



## List of Tables

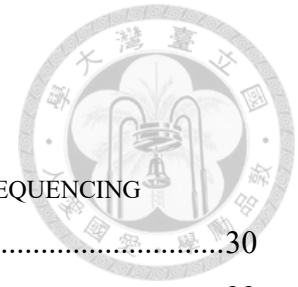


TABLE 2.1 SUMMARY OF THE MARKERS DEVELOPED WITH THE RAD SEQUENCING STRATEGY AND THE SEQUENCED GENES AS WELL.....	30
TABLE 2.2 GENETIC VARIATION OF EACH SUBPOPULATION. ....	33
TABLE 2.3 THE LOCAL LD PROFILES OF INDIVIDUAL CHROMOSOMES. ....	36
TABLE 3.1 SIGNIFICANT LOCI FOR STAMEN LENGTH IN TASSEL AND GEMMA.....	77
TABLE 3.2 THE TWO SIGNIFICANT LOCI BASED ON $P = G + Q + E$ MODEL. ....	78
TABLE 4.1 THE SUMMARY OF RNA-SEQ .....	111
TABLE 4.2 THE NUMBER OF DEGs IN EACH CHROMOSOME. ....	113
TABLE 4.3 THE SUMMARY OF THE CANDIDATE GENES FROM cLED19A24 TO CT9.....	115
TABLE 4.4 THE EXPRESSED GENES AND THE SNP DENSITY THROUGH EACH CHROMOSOME .....	116

## List of Supplementary data

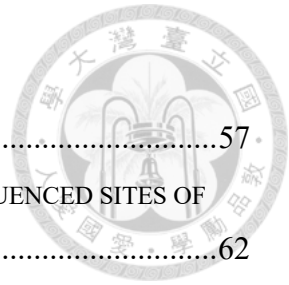


### SUPPLEMENTARY FIGURE

S_FIG 2.1 THE CROSS-VALIDATION ERROR OF K VALUE IN ADMIXTURE. ....	47
S_FIG 2.2 PAIRWISE ISOLATION BY DISTANCE OF 98 ACCESSIONS. ....	47
S_FIG 2.3 THE PCA OF SOLCAP META-ANALYSIS. A) THE PCA PLOT OF BI-ALLELIC SNPs. B) THE PCA PLOT AFTER REMOVING THOSE SNPs OF REVERSE-COMPLEMENT ALLELE DESIGNATION. ....	48
S_FIG 2.4 THE CROSS-VALIDATION ERROR OF SOLCAP META-ANALYSIS. ....	49
S_FIG 2.5 THE GENOME PATTERNS OF 214 SAMPLES IN SOLCAP META-ANALYSIS. ....	49
S_FIG 2.6 PAIRWISE ISOLATION BY DISTANCE OF SOLCAP META-ANALYSIS. ....	50
S_FIG 2.7 LD DECAY OF THE WHOLE GENOME. ....	50
S_FIG 2.8 50 KB INTERVAL LD DECAY OF EACH CHROMOSOME. ....	52
S_FIG 2.9 THE LOCAL LD OF EACH CHROMOSOME. ....	56
S_FIG 3.1 THE Q-Q PLOTS OF MIXED LINEAR MODELS. ....	88
S_FIG 3.2 THE MANHATTAN PLOTS OF TASSEL AND GEMMA RESULTS. ....	89
S_FIG 3.3 THE HEATMAP OF LD FOR EACH SIGNIFICANT LOCUS IN GWAS. ....	94
S_FIG 3.4 THE HEATMAP OF $R^2$ AND $D'$ FROM SSL2.50CH03_56790852 TO SSL2.50CH03_56903592. ....	95
S_FIG 3.5 THE DIFFERENCE OF $R^2$ BETWEEN TASSEL AND PLINK BASED ON 206,375 PAIR-WISE LD. ....	96
S_FIG 3.6 THE OVERALL LD DECAY BASED ON TASSEL. ....	96
S_FIG 4.1 THE DIFFERENCE OF SNP NUMBER BETWEEN M82 AND TA3178 THROUGH EACH CHROMOSOME. ....	123
S_FIG 4.2 THE DISTRIBUTION OF DEGS. ....	124

SUPPLEMENTARY TABLE

S_TAB 2.1 THE DETAILED INFORMATION ON EACH ACCESSION. ....	57
S_TAB 2.2 THE STATISTICAL SUMMARIES OF EXPECTED SITES AND SEQUENCED SITES OF <i>Pst</i> I, THE SITES TARGETED BY SNP AND THE SEQUENCED GENES. ....	62
S_TAB 2.3 THE INFORMATION ON 24,330 SNPs. ....	65
S_TAB 2.4 PAIRWISE $F_{ST}$ OF SUBPOPULATIONS. ....	66
S_TAB 2.5 THE LOCATIONS AND GENOTYPES OF 214 SAMPLES OF SOLCAP GENOTYPING ARRAY. ....	67
S_TAB 2.6 THE REMOVED 627 SNPs WITH REVERSE-COMPLEMENT ALLELE DESIGNATION. .....	68
S_TAB 2.7 THE IDENTITY OF 2,307 SNP MARKERS WITHIN ACCESSIONS. ....	72
S_TAB 3.1 THE STAMEN LENGTH OF EACH ACCESSION. ....	97
S_TAB 3.2 THE DIFFERENCE OF PAIR-WISE $R^2$ BETWEEN TASSEL AND PLINK, TAKING SSL2.50ch03_56790852 TO SSL2.50ch03_56903592 (THE SNPs IN S_FIG 3.4) FOR EXAMPLE .....	98
S_TAB 3.3 THE HAPLOTYPE BLOCKS ESTIMATED BY THE 24,330 SNPs. ....	106
S_TAB 3.4 THE HAPLOTYPE BLOCKS ESTIMATED BY ABOUT 68,000 SNPs.....	107
S_TAB 4.1 THE DETAIL INFORMATION OF 159 GENES IN THE INTROGRESSION SEGMENT. ....	125
S_TAB 4.2 THE DETAIL INFORMATION OF THE DEGs BASED ON THE 99.9 <sup>TH</sup> PERCENTILE METHOD. ....	128
S_TAB 4.3 THE DETAIL INFORMATION OF THE DEGs BASED ON THE DESEQ ANALYSIS. ....	130





# Chapter 1 Introduction



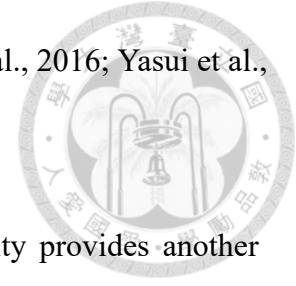
## 1.1 Heterostyly

### 1.1.1 Evolution of heterostyly

Heterostyly is a fascinating theme that draws deep interest of many botanists. Two morphs of *Primula* were appreciated by Charles Darwin for its evolutionary meaning: long-styled flowers promote outcrossing and short-styled flowers tend to occur self-fertilization (Charles Darwin, M.A., P.B.S., F.L.S. &c., 1862). Darwin proposed that heterostyly with self-incompatibility promoted the selective advantages of outcrossing because it could increase both male and female fitness through pollen transfer between inter-morph individuals, preventing pollen waste, and reducing progenies of inbreeding depression. Even in the case with self-compatibility, heterostyly could still reduce the disadvantage of producing less-fit selfing progenies (Darwin, 1877; Ganders, 1979; Keller, Thomson, & Conti, 2014).

In dimorphic heterostyly plants, long-styled flowers (pin flowers) show an elongated style at the mouth of flowers and anthers are located within a floral tube. Short-styled flowers (thrum flowers), on the contrary, show a short style within a floral tube and anthers are exposed at a flower mouth (Darwin, 1862). The genetic mechanism of heterostyly with self-incompatibility was established as a single locus (*S* locus) that consisted of several functionally related genes, so-called the supergene. The *S* locus contained at least three genes that controlled the style length (*G*), pollen size (*P*) and anther length (*A*) (Muenchow, 1981). However, recent studies have revealed that the occurrence of self-fertile non-heterostyly flower may result from the mutation of

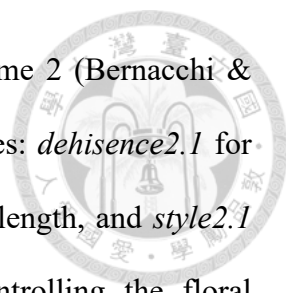
hemizygote, not the rare recombination within the supergene (Li et al., 2016; Yasui et al., 2016).



The genetic mechanism of heterostyly with self-incompatibility provides another application in agriculture. The common buckwheat (*Fagopyrum esculentum*) is a heteromorphic self-incompatible crop. Through whole genome sequencing of buckwheat, a segment of at least 5.4 Mb was identified as the short-styled specific allele. Nearly 75% of this hemizygous segment contained the sequences of transposon elements and the rest was annotated as 32 genes (Yasui et al., 2016). Deciphering the connection between self-incompatibility and heterostyly could increase the yield by removing the self-incompatibility and designing a homomorphic flower to guarantee self-fertilization and increase cereal crop production.

### **1.1.2 Heterostyly in tomato species**

Tomato is a perfect material to study the relationship between mating system and floral morphology because it displays both various mating systems and floral characters (Bedinger et al., 2011; Moyle, 2008; Spooner, Peralta, & Knapp, 2005). For example, *S. pennellii* is self-incompatible and has a more exerted style while *S. lycopersicum* is self-compatible and has a recessed style (Chen, Cong, Wing, Vrebalov, & Tanksley, 2007; Spooner et al., 2005). The quantitative trait loci (QTL) of self-incompatibility and floral morphology have been mapped by different tomato crosses (Bernacchi & Tanksley, 1997; Fulton et al., 1997; Georgiady, Whitkus, & Lord, 2002; Tanksley & Loaiza-Figueroa, 1985). According to those studies, the *S* locus and QTL of floral characters were not located on the same chromosome. The *S* locus was mapped on chromosome 1 through different tomato populations (Bernacchi & Tanksley, 1997; Tanksley & Loaiza-Figueroa, 1985). Meanwhile, *se2.1*, which was responsible for the



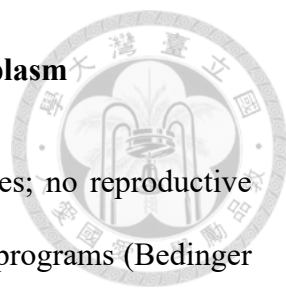
recessed stigma of cultivated tomatoes, was mapped on chromosome 2 (Bernacchi & Tanksley, 1997; Chen & Tanksley, 2004). *se2.1* contained five genes: *dehiscence2.1* for anther dehiscence, *stamen2.1*, *stamen2.2* and *stamen2.3* for anther length, and *style2.1* for the style length (Chen & Tanksley, 2004). Other QTL controlling the floral characters included: *stg2.1* and *stg2.9* for stigma exsertion, *sty8.1* for style length, *ant3.2*, *atl2.1*, and *atl7.1* for anther length (Fulton et al., 1997; Georgiady et al., 2002; Grandillo & Tanksley, 1996). Following Darwin's theory, the heterostyly in tomato clade is supposed to prevent from producing less-fit selfing progenies because *S* locus and the QTL of floral characters are not associated.

## 1.2 *Solanum pimpinellifolium*

### 1.2.1 The mating systems and flower characters in *S. pimpinellifolium*

*S. pimpinellifolium* is a perennial wild tomato native to Ecuador and Peru. Charles M. Rick utilized *S. pimpinellifolium* to illustrate the relationship between mating systems and floral characters and their impacts on genetic diversity (Rick, Fobes, & Holle, 1977; Rick, Holle, & Thorp, 1978). Three mating type were found within this wild tomato: complete autogamy, nearly complete allogamy and intermediate mating types (Rick et al., 1977). Because the exsertion of stigma interfered self-fertilization, both the floral morphology and the outcrossing rate were correlated to the genetic diversity (Rick et al., 1977, 1978). In addition, a F<sub>2</sub> population derived from LA1237 (a selfing *S. pimpinellifolium* accession) crossing to LA1581 (an outcrossing accession) revealed QTL related to floral characters, *ant3.2* and *sty8.1* (Georgiady et al., 2002). In this case, the QTL controlling anther length is not associated with that of style length, suggesting that floral characters are not always inherited as a single compressed unit.

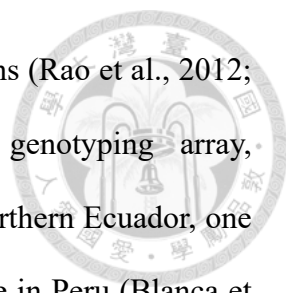
### 1.2.2 *S. pimpinellifolium* is a diverse and attractive tomato germplasm



*S. pimpinellifolium* is the closest relative to cultivated tomatoes; no reproductive barrier with cultivated tomatoes makes it advantageous in breeding programs (Bedinger et al., 2011; Moyle, 2008; Spooner et al., 2005). Several desired traits, such as abiotic and biotic resistances, have been revealed in some *S. pimpinellifolium* accessions. For example, *Ph1*, *Ph2*, *Ph3* and *Ph5*, the QTL for late blight resistance, were identified in *S. pimpinellifolium*. Among them, the most effective *Ph3* was further designed as DNA markers to screen the major resistance gene in tomato breeding programs (Jung et al., 2015; Panthee, Gardner, Ibrahim, & Anderson, 2015). Recently, World Vegetable Center has developed a core collection of *S. pimpinellifolium* in order to conserve and utilize this germplasm efficiently (Rao, Kadirvel, Symonds, Geethanjali, & Ebert, 2012). In addition, *S. pimpinellifolium* was involved in genome-wide association studies (GWAS) to increase the genetic diversity of the studying populations and to maintain the allele balance (Bauchet et al., 2017).

### 1.2.3 The population differentiation of *S. pimpinellifolium*

*S. pimpinellifolium* was originated from the northern Peru and then migrated to Ecuador and the southern Peru (Rick et al., 1977). The facultative allogamous *S. pimpinellifolium* was separated from the originated allogamous ones because the new environments might not be suitable to outcrossing (Rick et al., 1977). These regions present gradient temperature and precipitation changes from Ecuador towards southern Peru: western Ecuador is equatorial winter dry; northern Peru is a hot, arid desert; southern Peru is a cold, barren desert (Kottek, Grieser, Beck, Rudolf, & Rubel, 2006). The selfing and adaptation to different environments created several subpopulations (Rick et al., 1977; Zuriaga et al., 2009). Previous studies have showed the Ecuadorian



and the Peruvian accessions were genetically different subpopulations (Rao et al., 2012; Zuriaga et al., 2009). Recently, with the aid of SolCAP genotyping array, *S. pimpinellifolium* was divided into three subpopulations: one in northern Ecuador, one in the mountains of Ecuador extending to the north of Peru, and one in Peru (Blanca et al., 2012; Blanca et al., 2015). Since the genetic distance of these subpopulations was correlated to the major climatic parameters, such as temperature and humidity, special genetic characters could be selected and maintained in differential subpopulations (Blanca et al., 2015; Zuriaga et al., 2009).

#### **1.2.4 The genetic diversity of *S. pimpinellifolium***

*S. pimpinellifolium* presents intermediate genetic diversity when comparing with other wild tomatoes (Moyle, 2008). However, this species still provides many attractive genetic variations, especial in resistant genes. For example, at least 26 alleles of *Cf-2*, a R gene resistant to *Cladosporium fulvum*, were identified in a set of 138 natural individuals (Caicedo & Schaal, 2004). Previous studies support its relatively high diversity when comparing to cultivated tomatoes (Blanca et al., 2012; Blanca et al., 2015). In addition, the higher outcrossing rate maintained the higher genetic diversity; therefore, the genetic diversity declined from the northern Peru to the south (Blanca et al., 2015; Caicedo, 2008; Rick et al., 1977; Zuriaga et al., 2009). The outcrossing could break the linkage disequilibrium of *S. pimpinellifolium*, suggesting faster LD decay. The LD decay of *S. pimpinellifolium* ranged from 73 to 2,035 Kb, implying a finer resolution in GWAS in comparison with that from 3,178 to 15,554 Kb in *S. lycopersicum* (Bauchet et al., 2017).

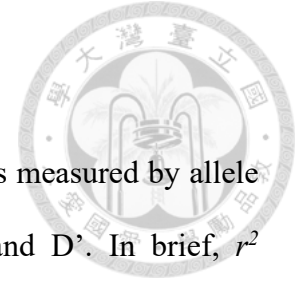
## 1.3 Genome-wide association study



### 1.3.1 The concept of GWAS

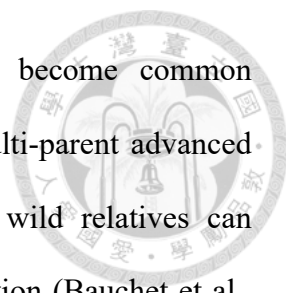
GWAS is basically the association mapping of a germplasm but with markers through whole genome. A significant marker is identified when the phenotypes between different genotypes are statistically different, usually examined by t-test or ANOVA. In this process, no linkage map is required. Once a significant marker is revealed, the QTL should be located within the LD interval of this marker. That is to say, GWAS utilizes markers through whole genome to examine which markers are associated with a studying phenotype. Comparing to a bi-parental cross population, GWAS involves more alleles because a germplasm accumulates mutations and recombinant events through its whole history. Together with the cost down of sequencing that makes the genotyping of a natural population much redundant, an explosive growth of GWAS in plants is now happening (Huang & Han, 2014; Soto-Cerda & Cloutier, 2012; Zhu, Gore, Buckler, & Yu, 2008). Following the concept of GWAS, number of markers and the LD between markers and QTL in a given population will determine the GWAS result (Korte & Farlow, 2013). More markers and more individuals mean more detectable recombinant events between markers and QTL, suggesting more precise estimations of LD and QTL effects. Unfortunately, QTL controlled by small-effect alleles and/or rare alleles could not be detected in a small population due to the limitation of statistical methods (Ingvarsson & Street, 2011; Korte & Farlow, 2013; Visscher et al., 2017). Despite many statistical models were proposed to rescue the problem, the fundamental solution would be a population of large sample size.

### 1.3.2 LD determines the resolution of GWAS



LD is the non-random assortment between pairwise alleles; it is measured by allele frequency and recombination using generally two statistics,  $r^2$  and  $D'$ . In brief,  $r^2$  summarizes the recombinant events and mutations, while  $D'$  presents only the information of recombination. A main concern for  $D'$  is that it is affected heavily by allele frequency, especially for a small population, because it is less possible to find a genotypic combination containing a rare allele. Meanwhile,  $r^2$  has a relatively small bias in a small population and additionally, it can reflect the correlation between markers and QTL. Therefore,  $r^2$  is utilized much more common in GWAS (Flint-Garcia, Thornsberry, & Buckler, 2003). Since allele frequency and recombination determine LD, any factor that affects these two factors may have an influence on LD and consequently GWAS results. (Flint-Garcia et al., 2003; Slatkin, 2008). In population history, allele frequency serves as an essential parameter; therefore, migration, mutation, selection and populations with or without subdivision all reflect on LD. Generally, migration and mutation that provide new genetic materials to a population would increase genetic diversity and consequently decrease LD. Strong selection force or genetic bottleneck would decrease genetic diversity and then create LD in a population (Flint-Garcia et al., 2003; Slatkin, 2008).

Recombination is basically determined by mating system in a natural population. In selfing genomes, generally an extensive region of LD would be observed because alleles tend to be fixed after selfing (Huang et al., 2012; Yano et al., 2016). In addition, great selection force during domestication process made LD extending to hundreds Kb, leading rough resolution in GWAS (Bauchet et al., 2017; Sauvage et al., 2014). To overcome the natural disadvantages of selfing plants, discovering new materials of high



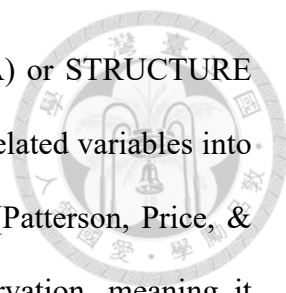
genetic diversity or designing diverse population panels have become common strategies. The population that consists of hybrid genomes, the multi-parent advanced generation intercross population or the population involving in wild relatives can increase genetic diversity and consequently improve GWAS resolution (Bauchet et al., 2017; Crowell et al., 2016; Huang et al., 2012; Ranc et al., 2012). In addition, more markers for a world-wide collection could also detect higher diversity, resulting in a better resolution as well (Kim et al., 2007).

### **1.3.3 Population structure and kinship cause confounding effects in GWAS**

Any factor contributing to LD can inflate the significance of GWAS result because the associations between markers and phenotypes determine the results of GWAS (Huang & Han, 2014; Soto-Cerda & Cloutier, 2012; Korte & Farlow, 2013). The confounding is created when LD is formed by only different allele frequency among families or among subpopulations. Two main confounding effects are the population structure, the distant common ancestry of a population, and the kinship, the existence of relatedness in a relatedness-unknown population (Astle & Balding, 2010). So far, the mixed linear model is a standard procedure to correct both confounding factors (Astle & Balding, 2010; Korol, Ronin, Itskovich, Peng, & Nevo, 2001; Yu et al., 2006; Zhang et al., 2010). However, population structure and kinship actually reflect a part of the genetic nature in a studying population rather than a problem. Simply using any correction could underestimate the genetic factors (Vilhjálmsón & Nordborg, 2013). Therefore, the correction would be strongly recommended when performing a candidate gene research but would be optional when investigating the genetic architectures of a given trait (Korte & Farlow, 2013).

The most practical method to correct population structure into GWAS would be the





integrations of the matrix from principal component analysis (PCA) or STRUCTURE and/or ADMIXTURE. PCA transforms a large data of possibly correlated variables into a smaller set of linearly-uncorrelated principal components (PCs) (Patterson, Price, & Reich, 2006). The first PC has the largest variance of the observation, meaning it accounts for the largest variation, and the succeeding PCs have the largest variance in a condition of orthogonal to the former components. By reducing the variables, PCs could reflect the main pattern of the genotypic data and distinguish the genetic difference among samples. Therefore, PCA is widely applied to cluster subpopulations of a studying population and PCs are added as a matrix of fixed effect into GWAS (Price, Zaitlen, Reich, & Patterson, 2010). On the other hand, STRUCTURE and/or ADMIXTURE is an algorithm that using the posterior probability to estimate the best number of subpopulations (K) (Pritchard, Stephens, & Donnelly, 2000). It identifies the simplest haplotypes among individuals and then assigns the individuals into subpopulations as probabilities. The best K can be determined by the natural logarithm of the probability of K or delta K (Evanno, Regnaut, & Goudet, 2005; Pritchard et al., 2000). Once K is determined, the probability of each individual assigned to each subpopulation can also reflect the portion of different genomes for each individual. And this probability matrix can be added as a fixed effect in GWAS.

Kinship refers to the degree of genetic relatedness and traditionally is estimated by identical by descent (IBD) while pedigree information is well informed (Jacquard, 1972). When incorporating to a pedigree-unknown germplasm, two identical alleles are considered as IBD or random sampling from a gene pool. Hence, the kinship can be modified by allele frequency and treated as the correlation coefficient of pairwise individuals (Anderson & Weir, 2007). Generally, kinship would be a random effect in GWAS because traditionally the relatedness is used to estimate the variance of heritable

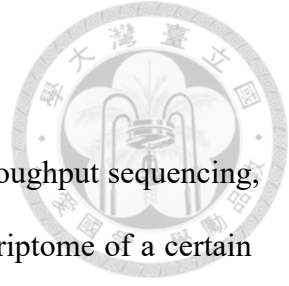
components (Yu *et al.* 2006; Astle & Balding 2009; Zhang *et al.* 2010).



## 1.4 Next generation sequencing (NGS) technology

### 1.4.1 Restriction-site associated DNA sequencing

So far, the genetic characteristics for *S. pimpinellifolium* accessions were mainly investigated based on SSR markers and the SolCAP array what were developed based on many genetic backgrounds (Blanca *et al.*, 2012; Blanca *et al.*, 2015; Rao *et al.*, 2012; Zuriaga *et al.*, 2009). Although the SolCAP array contains 7,720 SNPs derived from cDNA and functional markers and indeed accelerates the genotyping, more SNPs are desired in GWAS (Bauchet *et al.*, 2017; Sim *et al.*, 2012). In reality, limited resource makes it a dilemma to choose higher marker density or greater population size. Restriction-site associated DNA sequencing (RADseq) is one of the genome-wide genotyping techniques that applies NGS technology in a selective way (Davey & Blaxter, 2010). The advantage of RADseq is to force the sequencing resource on the vicinity of restriction enzyme cutting sites. Therefore, it provides the flexibility of experimental design regarding to the trade-off between budget saving and marker density. Choosing restrict enzymes depends on the number of cutting sites or special purposes. One can predict the sites via reference genomes to estimate the reduced coverage of a genome (Shirasawa, Hirakawa, & Isobe, 2016). And, one can also use methylation-sensitive restriction enzymes, such as *Pst*I, to concentrate the sequencing resource on gene-rich regions, preventing the resource from large heterochromatin on plant genomes (Bhakta, Jones, & Vallejos, 2015; Chen *et al.*, 2014; Hohenlohe *et al.*, 2010).



### 1.4.2 RNA sequencing

RNA can be converted into cDNA libraries to perform high-throughput sequencing, so-called RNA sequencing (RNA-seq). RNA-seq profiles the transcriptome of a certain tissue or organ in a certain development process through two major evaluations: the differentially expressed genes (DEGs) between groups and the polymorphisms in the coding sequences (Wang, Gerstein, & Snyder, 2009). However, the relative high cost of RNA-seq makes researchers struggle in the experiment design: more technical replications, more sequencing depth or more biological replications? First of all, it is recommended to prepare RNA-seq with technical replications in a balanced block design, to multiplex bar-coding samples in a single lane, because it can eliminate the confounding lane effect and simultaneously create technical replications (Auer & Doerge 2010). Second, increasing depth can produce greater power to detect DEGs but with a reduced feedback when passing over a threshold (Liu, Zhou, & White, 2014; Robles et al., 2012). Surprisingly, reducing depth as low as 15% did not affect false positive or true positive rates (Robles et al., 2012). Finally and most importantly, biological replications can increase power and the percentage of differentiated expressed (Robles et al., 2012). Therefore, to prepare biological replication is more essential than to increase sequencing depth. In tomato, two biological replications were often prepared and the reads ranged from 10 to 70 million per sample (Li et al., 2016; Tan et al., 2015; Zhang et al., 2016; Zhu et al., 2015; Zouari et al., 2014). This implied the quantities of reads heavily depended on the sequencing resources from case to case.



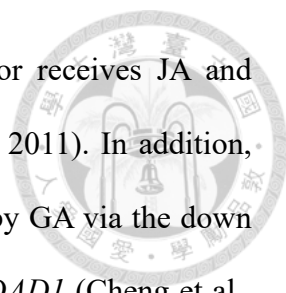
## 1.5 Development of stamen

### 1.5.1 MADS box genes determine stamen differentiation

Two main types of genes control flower development: one identifies floral organ differentiation, so-called ABC model genes; the other generally regulated by phytohormones participates in organ initiation or later development processes (Haughn & Somerville, 1988; Song, Qi, Huang, & Xie, 2013). In the ABC model, B- and C-class genes are responsible for stamen differentiation. Mutations of these genes can cause abnormal stamens. The B-class mutant of *Tomato MADS* gene 6 (*TM6*) and *TOMATO APETALA 3* (*TAP3*) showed carpelloid stamen and sepaloid petal (de Martino, 2006). The C-class mutant of *TOMATO AGAMOUS 1* (*TAG1*) displayed not only petaloid stamen but also abnormal carpels (Pnueli, 1994). Since B and C genes all belong to the MADS box, these MADS box transcription factors are heavily responsible for stamen development (Smaczniak, Immink, Angenent, & Kaufmann, 2012).

### 1.5.2 Phytohormones regulate the stamen development

Previous studies have reviewed that auxin, gibberellin (GA), jasmonate (JA), brassinosteroid (BR) and cytokinin regulate the stamen development in different stages (Cardarelli & Cecchetti, 2014; Mandaokar et al., 2006; Song et al., 2013). Therefore, genes participating in phytohormone biosynthesis and/or regulated by phytohormones affect stamen development. For example, mutants of auxin synthesis (*yuc2 yuc6*) and auxin response factor (*arf6 arf8*) display non-elongated or shorter stamen (Cheng, Dai, & Zhao, 2006; Nagpal et al., 2005). Meanwhile, phytohormones contribute to stamen development in crosstalk manners. Taking JA-regulated mechanism for example, the jasmonate zim-domain proteins release R2R3-type MYB transcription factors (MYB21



and MYB24) to participate stamen development when JA receptor receives JA and recruits jasmonate zim-domain proteins for degradation (Wu et al., 2011). In addition, JA biosynthesis is triggered not only by ARF6 and ARF8 but also by GA via the down regulation of DELLA, which suppresses the JA biosynthesis gene *DADI* (Cheng et al., 2009; Ishiguro, Kawai-Oda, Ueda, Nishida, & Okada, 2001; Nagpal et al., 2005; Tabata et al., 2010). The complicated mechanism of stamen development implies that many genes of small effect may be involved in the stamen length.

## 1.6 Conclusion

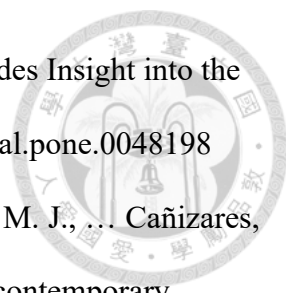
The natural variation of outcrossing rate and floral morphology within *S. pimpinellifolium* made it an ideal material to study the relationship between self-incompatibility and heterostyly via GWAS. In this research, we intended to identify the QTL or candidate genes controlling stamen length with different tomato materials. In chapter 2, we developed a set of genome-wide high-density SNP markers for a collection of 99 *S. pimpinellifolium* accessions through RADseq. Afterwards, population differentiation was investigated via this SNP set. In addition, LD analysis revealed the advantage and the weakness of this collection in GWAS. In chapter 3, we performed a GWAS to map the QTL controlling stamen length with the same *S. pimpinellifolium* population. We checked the false discovery rate (FDR) of the candidates and made some suggestions to reduce the high FDR. Finally, in chapter 4, a RNA-seq experiment was performed for M82 and its introgression line TA3178, which contained a segment of *S. pennellii* near *style2.1*. Based on the previous work in our team, *stamen2.2* and *stamen2.3* were located in the interval from marker cLED19A24 to CT9. This interval was annotated as 18 candidate genes. We narrowed the candidate list of *stamen2.2* and *stamen2.3* by comparing the expression level and cDNA polymorphisms between M82

and TA3178.



## 1.7 Reference

- Anderson, A. D., & Weir, B. S. (2007). A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. *Genetics*.  
<https://doi.org/10.1534/genetics.106.063149>
- Astle, W., & Balding, D. J. (2010). Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science*. <https://doi.org/10.1214/09-sts307>
- Auer, P. L., & Doerge, R. W. (2010). Statistical design and analysis of RNA sequencing data. *Genetics*. <https://doi.org/10.1534/genetics.110.114983>
- Bauchet, G., Grenier, S., Samson, N., Bonnet, J., Grivet, L., & Causse, M. (2017). Use of modern tomato breeding germplasm for deciphering the genetic control of agronomical traits by Genome Wide Association study. *Theoretical and Applied Genetics*. <https://doi.org/10.1007/s00122-017-2857-9>
- Bedinger, P. A., Chetelat, R. T., McClure, B., Moyle, L. C., Rose, J. K. C., Stack, S. M., ... Royer, S. (2011). Interspecific reproductive barriers in the tomato clade: Opportunities to decipher mechanisms of reproductive isolation. *Sexual Plant Reproduction*. <https://doi.org/10.1007/s00497-010-0155-7>
- Bernacchi, D., & Tanksley, S. D. (1997). An interspecific backcross of *Lycopersicon esculentum* X *L. hirsutum*: Linkage analysis and a QTL study of sexual compatibility factors and floral traits. *Genetics*.
- Bhakta, M. S., Jones, V. A., & Vallejos, C. E. (2015). Punctuated distribution of recombination hotspots and demarcation of pericentromeric regions in *Phaseolus vulgaris* L. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0116822>
- Blanca, J., Cañizares, J., Cordero, L., Pascual, L., Diez, M. J., & Nuez, F. (2012).

- 
- Variation Revealed by SNP Genotyping and Morphology Provides Insight into the Origin of the Tomato. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0048198>
- Blanca, J., Montero-Pau, J., Sauvage, C., Bauchet, G., Illa, E., Díez, M. J., ... Cañizares, J. (2015). Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. *BMC Genomics*. <https://doi.org/10.1186/s12864-015-1444-1>
- Caicedo, A. L. (2008). Geographic diversity cline of R gene homologs in wild populations of *Solanum pimpinellifolium* (Solanaceae). *American Journal of Botany*. <https://doi.org/10.3732/ajb.95.3.393>
- Caicedo, A. L., & Schaal, B. A. (2004). Heterogeneous evolutionary processes affect R gene diversity in natural populations of *Solanum pimpinellifolium*. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 17444–17449. <https://doi.org/10.1073/pnas.0407899101>
- Cardarelli, M., & Cecchetti, V. (2014). Auxin polar transport in stamen formation and development: how many actors? *Frontiers in Plant Science*. <https://doi.org/10.3389/fpls.2014.00333>
- Charles Darwin M.A., P.B.S., F.L.S., &c. (1862). On the two forms, or dimorphic conditions in the species of *Primula*, and on their remarkable sexual relations. *The Journal of Linn Soc Lond Bot*, 6, 77–69.
- Chen, A. L., Liu, C. Y., Chen, C. H., Wang, J. F., Liao, Y. C., Chang, C. H., ... Chen, K. Y. (2014). Reassessment of QTLs for late blight resistance in the tomato accession L3708 using a restriction site associated DNA (RAD) linkage map and highly aggressive isolates of *Phytophthora infestans*. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0096417>
- Chen, K. Y., Cong, B., Wing, R., Vrebalov, J., & Tanksley, S. D. (2007). Changes in regulation of a transcription factor lead to autogamy in cultivated tomatoes.

*Science*. <https://doi.org/10.1126/science.1148428>

Chen, K. Y., & Tanksley, S. D. (2004). High-resolution mapping and functional analysis of *se2.1*: A major stigma exertion quantitative trait locus associated with the evolution from allogamy to autogamy in the genus *lycopersicon*. *Genetics*. <https://doi.org/10.1534/genetics.103.022558>

Cheng, H., Song, S., Xiao, L., Soo, H. M., Cheng, Z., Xie, D., & Peng, J. (2009). Gibberellin acts through jasmonate to control the expression of *MYB21*, *MYB24*, and *MYB57* to promote stamen filament growth in *Arabidopsis*. *PLoS Genetics*. <https://doi.org/10.1371/journal.pgen.1000440>

Cheng, Y., Dai, X., & Zhao, Y. (2006). Auxin biosynthesis by the YUCCA flavin monooxygenases controls the formation of floral organs and vascular tissues in *Arabidopsis*. *Genes and Development*. <https://doi.org/10.1101/gad.1415106>

Crowell, S., Korniliev, P., Falcão, A., Ismail, A., Gregorio, G., Mezey, J., & McCouch, S. (2016). Genome-wide association and high-resolution phenotyping link *Oryza sativa* panicle traits to numerous trait-specific QTL clusters. *Nature Communications*. <https://doi.org/10.1038/ncomms10527>

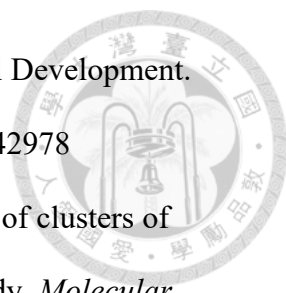
Darwin, C. (1862). On the Two Forms, or Dimorphic Condition, in the Species of *Prumla*, and on their remarkable Sexual Relations. *Journal of the Proceedings of the Linnean Society, Botany*, 6, 77–96.

Darwin, C. (1877). *The different forms of flowers on plants of the same Species*. *The Different Forms of Flowers on Plants of the Same Species*. <https://doi.org/10.1017/CBO9780511731419>

Davey, J. W., & Blaxter, M. L. (2010). RADseq: Next-generation population genetics. *Briefings in Functional Genomics*. <https://doi.org/10.1093/bfgp/elq031>

de Martino, G. (2006). Functional Analyses of Two Tomato *APETALA3* Genes



- 
- Demonstrate Diversification in Their Roles in Regulating Floral Development.  
*THE PLANT CELL ONLINE*. <https://doi.org/10.1105/tpc.106.042978>
- Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*. <https://doi.org/10.1111/j.1365-294X.2005.02553.x>
- Flint-Garcia, S. A., Thornsberry, J. M., & Buckler, E. S. (2003). Structure of Linkage Disequilibrium in Plants. *Annual Review of Plant Biology*.  
<https://doi.org/10.1146/annurev.arplant.54.031902.134907>
- Fulton, T. M., Beck-Bunn, T., Emmatty, D., Eshed, Y., Lopez, J., Petiard, V., ... Tanksley, S. D. (1997). QTL analysis of an advanced backcross of *Lycopersicon peruvianum* to the cultivated tomato and comparisons with QTLs found in other wild species. *Theoretical and Applied Genetics*.  
<https://doi.org/10.1007/s001220050639>
- Ganders, F. R. (1979). The biology of heterostyly. *New Zealand Journal of Botany*.  
<https://doi.org/10.1080/0028825X.1979.10432574>
- Georgiady, M. S., Whitkus, R. W., & Lord, E. M. (2002). Genetic analysis of traits distinguishing outcrossing and self-pollinating forms of currant tomato, *Lycopersicon pimpinellifolium* (Jusl.) Mill. *Genetics*.
- Grandillo, S., & Tanksley, S. D. (1996). QTL analysis of horticultural traits differentiating the cultivated tomato from the closely related species *Lycopersicon pimpinellifolium*. *Theoretical and Applied Genetics*.  
<https://doi.org/10.1007/BF00224033>
- Haughn, G. W., & Somerville, C. R. (1988). Genetic control of morphogenesis in *Arabidopsis*. *Developmental Genetics*, 9(2), 73–89.  
<https://doi.org/10.1002/dvg.1020090202>

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., & Cresko, W.

A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*.

<https://doi.org/10.1371/journal.pgen.1000862>



Huang, X., & Han, B. (2014). Natural Variations and Genome-Wide Association Studies in Crop Plants. *Annual Review of Plant Biology*.

<https://doi.org/10.1146/annurev-arplant-050213-035715>

Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., ... Han, B. (2012).

Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nature Genetics*.

<https://doi.org/10.1038/ng.1018>

Ingvarsson, P. K., & Street, N. R. (2011). Association genetics of complex traits in plants. *New Phytologist*. <https://doi.org/10.1111/j.1469-8137.2010.03593.x>

Ishiguro, S., Kawai-Oda, a, Ueda, J., Nishida, I., & Okada, K. (2001). The

*DEFECTIVE IN ANTHER DEHISCENCE* gene encodes a novel phospholipase A1 catalyzing the initial step of jasmonic acid biosynthesis, which synchronizes pollen maturation, anther dehiscence, and flower opening in *Arabidopsis*. *Plant Cell*, 13(10), 2191–2209. <https://doi.org/10.1105/tpc.010192>

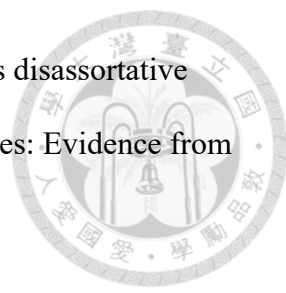
Soto-Cerda, B. J., and S. Cloutier, 2012 Association mapping in plant genomes, in *Genetic Diversity in Plants*, edited by C. Mahmut. InTech, Rijeka.


Jacquard, A. (1972). Genetic Information Given by a Relative. *Society*.

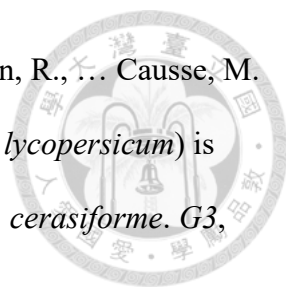
<https://doi.org/10.2307/2528643>

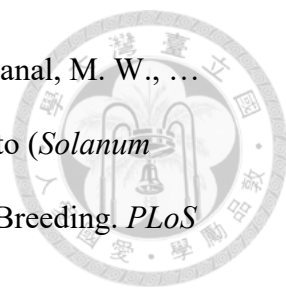
Jung, J., Kim, H. J., Lee, J. M., Oh, C. S., Lee, H. J., & Yeam, I. (2015). Gene-based molecular marker system for multiple disease resistances in tomato against Tomato yellow leaf curl virus, late blight, and verticillium wilt. *Euphytica*.

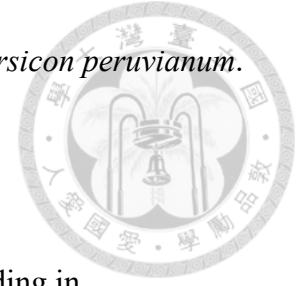
<https://doi.org/10.1007/s10681-015-1442-z>

- 
- Keller, B., Thomson, J. D., & Conti, E. (2014). Heterostyly promotes disassortative pollination and reduces sexual interference in Darwin's primroses: Evidence from experimental studies. *Functional Ecology*.  
<https://doi.org/10.1111/1365-2435.12274>
- Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R. M., Ossowski, S., ... Nordborg, M. (2007). Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics*. <https://doi.org/10.1038/ng2115>
- Korol, A. B., Ronin, Y. I., Itskovich, A. M., Peng, J., & Nevo, E. (2001). Enhanced efficiency of quantitative trait loci mapping analysis based on multivariate complexes of quantitative traits. *Genetics*.  
<https://doi.org/10.1534/genetics.107.080101>
- Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: A review. *Plant Methods*. <https://doi.org/10.1186/1746-4811-9-29>
- Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*, *15*(3), 259–263. <https://doi.org/10.1127/0941-2948/2006/0130>
- Li, J., Cocker, J. M., Wright, J., Webster, M. A., McMullan, M., Dyer, S., ... Gilmartin, P. M. (2016). Genetic architecture and evolution of the *S* locus supergene in *Primula vulgaris*. *Nature Plants*. <https://doi.org/10.1038/nplants.2016.188>
- Li, J., Tao, X., Li, L., Mao, L., Luo, Z., Khan, Z. U., & Ying, T. (2016). Comprehensive RNA-seq analysis on the regulation of tomato ripening by exogenous auxin. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0156453>
- Liu, Y., Zhou, J., & White, K. P. (2014). RNA-seq differential expression studies: More sequence or more replication? *Bioinformatics*.  
<https://doi.org/10.1093/bioinformatics/btt688>

- 
- Mandaokar, A., Thines, B., Shin, B., Markus Lange, B., Choi, G., Koo, Y. J., ...  
Browse, J. (2006). Transcriptional regulators of stamen development in  
*Arabidopsis* identified by transcriptional profiling. *Plant Journal*.  
<https://doi.org/10.1111/j.1365-313X.2006.02756.x>
- Moyle, L. C. (2008). Ecological and evolutionary genomics in the wild tomatoes  
(*Solanum* Sect. *Lycopersicon*). *Evolution*.  
<https://doi.org/10.1111/j.1558-5646.2008.00487.x>
- Nagpal, P., Ellis, C. M., Weber, H., Ploense, S. E., Barkawi, L. S., Guilfoyle, T.  
J., ...Reed, J. W. (2005). Auxin response factors ARF6 and ARF8 promote  
jasmonic acid production and flower maturation. *Development*.  
<https://doi.org/10.1242/dev.01955>
- Panthee, D. R., Gardner, R. G., Ibrahim, R., & Anderson, C. (2015). Molecular  
Markers Associated with *Ph- 3* Gene Conferring Late Blight Resistance in Tomato.  
*American Journal of Plant Sciences*, 6, 2144–2150.  
<https://doi.org/10.4236/ajps.2015.613216>
- Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis.  
*PLoS Genetics*. <https://doi.org/10.1371/journal.pgen.0020190>
- Pnueli, L. (1994). Isolation of the Tomato *AGAMOUS* Gene *TAG1* and Analysis of Its  
Homeotic Role in Transgenic Plants. *The Plant Cell Online*, 6(2), 163–173.  
<https://doi.org/10.1105/tpc.6.2.163>
- Price, A., Zaitlen, N., Reich, D., & Patterson, N. (2010). New approaches to population  
stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7),  
459–463. <https://doi.org/10.1038/nrg2813>
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure  
using multilocus genotype data. *Genetics*.

- 
- Ranc, N., Munos, S., Xu, J., Le Paslier, M. C., Chauveau, A., Bounon, R., ... Causse, M. (2012). Genome-wide association mapping in tomato (*Solanum lycopersicum*) is possible using genome admixture of *Solanum lycopersicum* var. *cerasiforme*. *G3*, 2(8), 853–864. <https://doi.org/10.1534/g3.112.002667>
- Rao, E. S., Kadirvel, P., Symonds, R. C., Geethanjali, S., & Ebert, A. W. (2012). Using SSR markers to map genetic diversity and population structure of *Solanum pimpinellifolium* for development of a core collection. *Plant Genetic Resources: Characterisation and Utilisation*. <https://doi.org/10.1017/S1479262111000955>
- Rick, C. M., Fobes, J. F., & Holle, M. (1977). Genetic variation in *Lycopersicon pimpinellifolium*: Evidence of evolutionary change in mating systems. *Plant Systematics and Evolution*. <https://doi.org/10.1007/BF00984147>
- Rick, C. M., Holle, M., & Thorp, R. W. (1978). Rates of cross-pollination in *Lycopersicon pimpinellifolium*: Impact of genetic variation in floral characters. *Plant Systematics and Evolution*. <https://doi.org/10.1007/BF00988982>
- Robles, A., Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J., Taylor, J. M., ... Taylor, J. M. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics*, 13(1), 484. <https://doi.org/10.1186/1471-2164-13-484>
- Sauvage, C., Segura, V., Bauchet, G., Stevens, R., Do, P. T., Nikoloski, Z., ... Causse, M. (2014). Genome-Wide Association in Tomato Reveals 44 Candidate Loci for Fruit Metabolic Traits. *Plant Physiology*, 165(3), 1120–1132. <https://doi.org/10.1104/pp.114.241521>
- Shirasawa, K., Hirakawa, H., & Isobe, S. (2016). Analytical workflow of double-digest restriction site-associated DNA sequencing based on empirical and in silico optimization in tomato. *DNA Research*. <https://doi.org/10.1093/dnares/dsw004>

- 
- Sim, S. C., van Deynze, A., Stoffel, K., Douches, D. S., Zarka, D., Ganai, M. W., ... Francis, D. M. (2012). High-Density SNP Genotyping of Tomato (*Solanum lycopersicum* L.) Reveals Patterns of Genetic Variation Due to Breeding. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0045520>
- Slatkin, M. (2008). Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*. <https://doi.org/10.1038/nrg2361>
- Smaczniak, C., Immink, R. G. H., Angenent, G. C., & Kaufmann, K. (2012). Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. *Development*. <https://doi.org/10.1242/dev.074674>
- Song, S., Qi, T., Huang, H., & Xie, D. (2013). Regulation of stamen development by coordinated actions of jasmonate, auxin, and gibberellin in *Arabidopsis*. *Molecular Plant*. <https://doi.org/10.1093/mp/sst054>
- Spooner, D. M., Peralta, I. E., & Knapp, S. (2005). Comparison of AFLPs with other markers for phylogenetic inference in wild tomatoes [*Solanum* L. section *Lycopersicon* (Mill.) Wettst.]. *Taxon*, 54(1), 43–61.
- Tabata, R., Ikezaki, M., Fujibe, T., Aida, M., Tian, C. E., Ueno, Y., ... Ishiguro, S. (2010). *Arabidopsis* AUXIN RESPONSE FACTOR6 and 8 regulate jasmonic acid biosynthesis and floral organ development via repression of class 1 *KNOX* genes. *Plant and Cell Physiology*. <https://doi.org/10.1093/pcp/pcp176>
- Tan, G., Liu, K., Kang, J., Xu, K., Zhang, Y., Hu, L., ... Li, C. (2015). Transcriptome analysis of the compatible interaction of tomato with *Verticillium dahliae* using RNA-sequencing. *Frontiers in Plant Science*. <https://doi.org/10.3389/fpls.2015.00428>
- Tanksley, S. D., & Loaiza-Figueroa, F. (1985). Gametophytic self-incompatibility is



controlled by a single major locus on chromosome 1 in *Lycopersicon peruvianum*.  
*Proceedings of the National Academy of Sciences*.

<https://doi.org/10.1073/pnas.82.15.5093>

Vilhjálmsón, B. J., & Nordborg, M. (2013). The nature of confounding in  
genome-wide association studies. *Nature Reviews Genetics*.

<https://doi.org/10.1038/nrg3382>

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., &  
Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and  
Translation. *American Journal of Human Genetics*.

<https://doi.org/10.1016/j.ajhg.2017.06.005>

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for  
transcriptomics. *Nature Reviews Genetics*. <https://doi.org/10.1038/nrg2484>

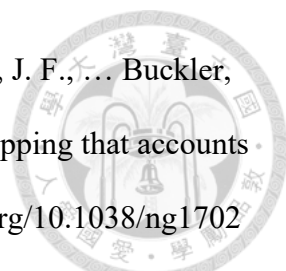
Wu, D., Liu, Y., Song, S., Chang, C., Peng, J., Peng, W., ... Qi, T. (2011). The  
Jasmonate-ZIM Domain Proteins Interact with the R2R3-MYB Transcription  
Factors MYB21 and MYB24 to Affect Jasmonate-Regulated Stamen Development  
in *Arabidopsis*. *The Plant Cell*, 23(3), 1000–1013.

<https://doi.org/10.1105/tpc.111.083089>

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P. C., Hu, L., ... Matsuoka, M.  
(2016). Genome-wide association study using whole-genome sequencing rapidly  
identifies new genes influencing agronomic traits in rice. *Nature Genetics*, 48(8),  
927–934. <https://doi.org/10.1038/ng.3596>

Yasui, Y., Hirakawa, H., Ueno, M., Matsui, K., Katsube-Tanaka, T., Yang, S. J., ...  
Mori, M. (2016). Assembly of the draft genome of buckwheat and its applications  
in identifying agronomically useful genes. *DNA Research*.

<https://doi.org/10.1093/dnares/dsw012>

- 
- Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., ... Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*. <https://doi.org/10.1038/ng1702>
- Zhang, S., Xu, M., Qiu, Z., Wang, K., Du, Y., Gu, L., & Cui, X. (2016). Spatiotemporal transcriptome provides insights into early fruit development of tomato (*Solanum lycopersicum*). *Scientific Reports*. <https://doi.org/10.1038/srep23173>
- Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., ... Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*. <https://doi.org/10.1038/ng.546>
- Zhu, B., Yang, Y., Li, R., Fu, D., Wen, L., Luo, Y., & Zhu, H. (2015). RNA sequencing and functional analysis implicate the regulatory role of long non-coding RNAs in tomato fruit ripening. *Journal of Experimental Botany*. <https://doi.org/10.1093/jxb/erv203>
- Zhu, C., Gore, M., Buckler, E. S., & Yu, J. (2008). Status and Prospects of Association Mapping in Plants. *The Plant Genome*, 1(1), 5–20. <https://doi.org/10.3835/plantgenome2008.02.0089>
- Zouari, I., Salvioli, A., Chialva, M., Novero, M., Miozzi, L., Tenore, G. C., ... Bonfante, P. (2014). From root to fruit: RNA-Seq analysis shows that arbuscular mycorrhizal symbiosis may affect tomato fruit metabolism. *BMC Genomics*. <https://doi.org/10.1186/1471-2164-15-221>
- Zuriaga, E., Blanca, J. M., Cordero, L., Sifres, A., Blas-Cerdán, W. G., Morales, R., & Nuez, F. (2009). Genetic and bioclimatic variation in *Solanum pimpinellifolium*. *Genetic Resources and Crop Evolution*. <https://doi.org/10.1007/s10722-008-9340-z>



# Chapter 2 Assessment of population differentiation and linkage disequilibrium in *Solanum pimpinellifolium* using genome-wide high-density SNP markers



## 2.1 Purpose

Before performing a GWAS, the population structure and the LD should be investigated to understand the genetic nature of a studying population. First of all, a *Pst*I-digested RADseq of 99 accessions was conducted to develop a genome-wide high-density SNP set. The population differentiation was examined by different approaches, including ADMIXTURE, PCA, pair-wise  $F_{st}$  and AMOVA. Afterwards, the LD and the marker density were evaluated to reveal the advantage and the potential weakness of this collection in GWAS. This chapter is modified based on the published paper on G3; Genes/Genomes/Genetics: Assessment of Genetic Differentiation and Linkage Disequilibrium in *Solanum pimpinellifolium* Using Genome-Wide High-Density SNP Markers (<https://doi.org/10.1534/g3.118.200862>).

## 2.2 Material and Method

### 2.2.1 Plant materials

All plant materials and their information were obtained from TGRC (S\_Tab 2.1; <http://tgrc.ucdavis.edu/>). A total of 12 accessions from Ecuador and 87 accessions from Peru were utilized in this study. According to their mating types, 43 accessions were facultative self-compatible (FSC), and 56 accessions were autogamous self-compatible (ASC). Seeds were propagated by self-pollination for two generations using the method

of single-seed descent in a greenhouse. Young leaves collected from plants of these single-seed descendent seeds were used for DNA extraction.

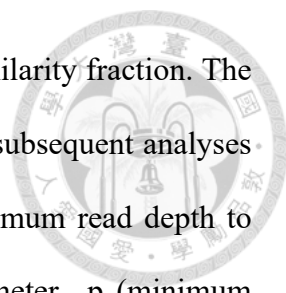


### 2.2.2 RAD sequencing

Total genomic DNA was extracted from young leaves using a modified CTAB method (Fulton, Chunwongse, & Tanksley, 1995) and purified with a DNeasy Blood & Tissue Kit (QIAGEN, Venlo, Netherland) following the manufacturer's instructions. We chose *Pst*I to select the sequencing regions because *Pst*I is a methylation-sensitive restriction enzyme and it may cut more frequently in euchromatin regions than heterochromatin regions (Dobritsa & Dobritsa, 1980). *Pst*I-digested DNA libraries were prepared following the protocol of Etter *et al.* (Etter, Bassham, Hohenlohe, Johnson, & Cresko, 2011). Four RADseq libraries were constructed, and each was sequenced in one lane of an Illumina HiSeq2000 flow cell (100 bp single-end reads) (Illumina Inc., San Diego, CA, USA). All the sequences of RADseq were submitted to the NCBI SRA database, and the BioProject Number is PRJNA358110.

### 2.2.3 SNP calling

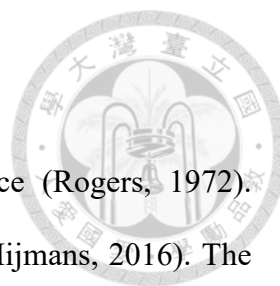
Reads were analyzed with Stacks version 1.37 (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013) and with CLC Genomics Workbench software version 6.5.1 (QIAGEN, Venlo, Netherlands). First, the *process\_radtags* command in Stacks filtered out low-quality reads with Q scores less than 20. The remaining reads were mapped to the tomato reference genome SL2.50 (Fernandez-Pozo *et al.*, 2015) using the "Map Reads to Reference" tool in the CLC Genomics Workbench software. Considering that genetic variation between the tomato reference genome *S. lycopersicum* and *S. pimpinellifolium* is larger than genetic variation within *S. lycopersicum*, mapping



parameters were set as 0.5 for the length fraction and 0.9 for the similarity fraction. The reads of the same individual in different lanes were merged. In the subsequent analyses using Stacks, the *ref\_map.pl* command set the parameter *-m* (minimum read depth to create a stack) as 10, and the *populations* command set the parameter *-p* (minimum number of populations a locus must be present) as 75. SNPs with a minor allele frequency of less than 0.05 were further excluded, and a set of 24,330 SNP markers was obtained. This set of 24,330 SNP markers was utilized for the analyses of genetic variation, LD,  $F_{st}$  and AMOVA. Another SNP set without ‘redundant SNP markers’ was used to conduct the principal component analysis (PCA) and ADMIXTURE because these two matrices are expected to correct the structure in GWAS. To remove ‘redundant SNP markers’, we defined a sequencing unit as a sequencing region surrounding a *PstI* site, usually 186 bp long, which has at least one SNP with a minor allele frequency greater than 0.05 in the *S. pimpinellifolium* population. If more than one SNPs are located in a sequencing unit and they are in complete LD ( $r^2 = 1$ ), only the first SNP is kept. This process resulted in a total of 19,993 SNP markers. ITAG2.4 gene model from SGN was used as the reference gene annotation.

#### **2.2.4 Population differentiation**

PCA was performed in TASSEL5.0 (Bradbury et al., 2007). ADMIXTURE was completed following the manual; the best K was determined following the procedure of cross-validation in the manual (Alexander, Novembre, & Lange, 2009). Pairwise  $F_{st}$  (Weir & Cockerham, 1984) and analysis of molecular variance (AMOVA) (Excoffier, Smouse, & Quattro, 1992) were conducted in the R package StAMPP (Pembleton, Cogan, & Forster, 2013).



### **2.2.5 Isolation by distance**

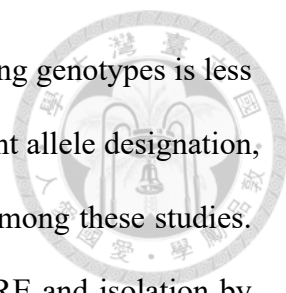
Pairwise genetic distance was measured by Rogers' distance (Rogers, 1972). Geographic distance was calculated by the R package geosphere (Hijmans, 2016). The significance of the correlation between pairwise genetic distance and geographic distance was examined by the Mantel test in the R package adegenet with 1,000 permutations (Jombart, 2008).

### **2.2.6 Estimate of genetic variation and LD**

Genetic variation within overall accessions and within each of the seven groups was assessed based on observed heterozygosity and the within-population gene diversity (expected heterozygosity) using the R package hierfstat (Goudet & Jombart, 2015). Pairwise  $r^2$  values between SNP markers were calculated to assess overall extent of LD via plink1.9 within a 1-Mb window (Gaunt, Rodríguez, & Day, 2007) and fit by non-linear regression (Remington et al., 2001). The baseline of the  $r^2$  value was set at 0.1 (Bauchet et al., 2017). The local LD along each chromosome was assessed as following: for each pair of consecutive sequencing units (defined in the section of SNP calling), the average  $r^2$  was calculated between two SNPs in different sequencing units and plotted along the left *PstI* cutting site based on the physical position. The heterochromatin regions were marked according to the genetic map of EXPIM 2012 and the physical map of the tomato reference genome (Sim et al., 2012).

### **2.2.7 Analysis of SolCAP array data of *S. pimpinellifolium***

The SolCAP data of 214 samples of *S. pimpinellifolium* were downloaded from previous studies (Blanca et al., 2012; Blanca et al., 2015; Sim et al., 2012). A set of 2,934 bi-allelic polymorphic SNPs was extracted after filtered with the criteria that



minor allele frequency is more than 0.05 and the proportion of missing genotypes is less than 25%. We dropped 627 SNP markers that are reverse-complement allele designation, resulting in a set of 2,307 SNPs with consistent allele designation among these studies. This set of 2,307 SNPs was utilized in the analyses of ADMIXTURE and isolation by distance following the same procedures described in the sections of population differentiation and isolation by distance. Meanwhile, because some accessions were genotyped in more than one SolCAP studies, different suffixes—“\_2012S,” “\_2012B,” and “\_2015B,”—were added to the sample name to indicate their original references Sim *et al.* 2012a, Blanca *et al.* 2012, and Blanca *et al.* 2015, respectively. Also, the percentage of identical SNP genotypes of the same accessions were calculated based on the 2,307 SNP genotypes without missing values.

## 2.3 Result

### 2.3.1 Identification of 24,330 SNPs from *PstI*-digested DNA libraries

A total of 655,973,270 short DNA reads were obtained from four lanes of the Illumina HiSeq2000 flow cell and were divided into 99 parts according to barcode sequences. Each part was derived from the DNA of a *S. pimpinellifolium* accession and contained at least 3.7 million DNA reads, except for LA2647 (S\_Tab 2.1). Among the 82,814 *PstI* sites in the tomato reference sequence SL2.50, only 23,988 *PstI* sites were covered by the sequenced DNA reads (S\_Tab 2.2). The sequenced regions included 0.54% of the SL2.50 reference sequences and 12,790 annotated genes (Table 2.1). Interestingly, approximately 84% of the sequenced *PstI* sites were located in the euchromatic regions (S\_Tab 2.2). Besides, the proportion of sequenced genes in euchromatin (43.13%) were about twice as that in heterochromatin (19.75%) (S\_Tab 2.2).

Table 2.1 Summary of the markers developed with the RAD sequencing strategy and the sequenced genes as well.

Chr.	SNPs	Genes in sequenced region	Genes with SNPs	SNPs in gene regions
0	147	62	25	57
1	3,222	1,742	1,029	2,374
2	2,401	1,400	803	1,661
3	2,522	1,389	812	1,779
4	2,121	1,054	611	1,328
5	1,680	783	437	1,049
6	2,179	1,195	673	1,422
7	1,756	902	535	1,174
8	1,929	952	599	1,304
9	1,670	877	507	1,192
10	1,616	812	444	954
11	1,563	834	466	1,054
12	1,524	788	440	1,017
<b>Total</b>	<b>24,330</b>	<b>12,790</b>	<b>7,381</b>	<b>16,365</b>

Two criteria were set to ensure the accuracy of SNP calling and genotype calling: one was that the read depth aligning to the reference sequence was equal to or greater than 10, and the other was that at least 75% of the accessions showed genotypes associated with a defined SNP marker. A total of 67,804 SNPs were identified in the sequenced regions of 99 *S. pimpinellifolium* accessions, and 24,330 of them had the minor allele frequency higher than 0.05. In the genotypic dataset of the 24,330 SNP markers (S\_Tab 2.3), the missing proportion of each accession ranged from 0.72% to 15.92%, except for LA2647 of which the value was 65.68% due to a low number of sequencing reads (S\_Tab 2.1). Regarding the location of these 24,330 SNPs, 16,365 SNPs were found in 7,383 annotated genes (Table 2.1), and the remaining SNPs were in the intergenic regions. Concerning the proportion of sequenced *Pst*I sites that contained SNPs, there is no significant difference between those sites in euchromatin (68.85%) and those in heterochromatin (60.59%) (S\_Tab 2.2). Meanwhile, the genotypic data of the LA0411 accession was dropped because the observed heterozygosity of LA0411

was inconsistent with its mating type (S\_Tab 2.1).



### **2.3.2 A similar distribution between genes and SNPs was identified in the vicinity of *Pst*I cutting site throughout the genome**

The observation that 67.26% (16,365 to 24,330) of the SNPs were located in the annotated gene regions (Table 2.1) implied a correlation between the distribution of the identified SNPs in the current study and the distribution of the annotated genes. Additional observations in the current study indicated a preference for genomic DNA digestion by the *Pst*I restriction enzyme in the euchromatic regions: only 28.97% (23,988 to 82,814) of *Pst*I sites were found in the deep sequencing regions, and 83.55% (20,043 to 23,988) of the deep sequencing regions were located in the euchromatic region (S\_Tab 2.2). It is worth noting that the current RADseq protocol did produce low coverage of sequencing reads in some *Pst*I sites (with a read depth less than 10), and these *Pst*I sites were filtered by the criteria of SNP and genotype calling; therefore, the deep sequencing regions indicated that their read depths were no less than 10. Incidentally, because SNPs can be identified only in the sequenced regions, it is a reasonable deduction that most SNPs found in the current study are located in the euchromatic regions. Figure 2.1 confirms clearly that the annotated tomato genes (A layer), the *Pst*I sites in the deep sequencing regions (C layer), and identified SNPs (D layer) are mainly located in the euchromatic regions.

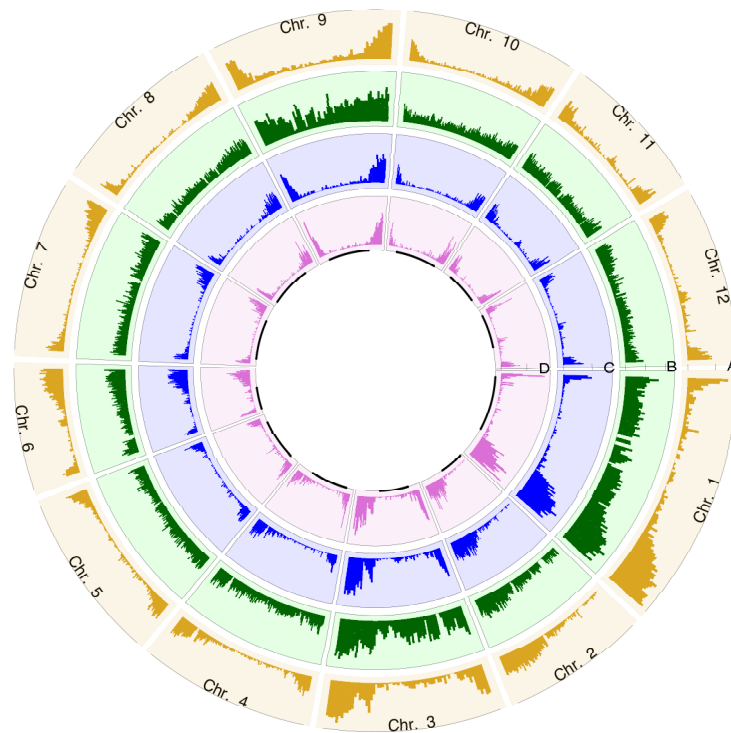


Figure 2.1 The distributions of ITAG2.4 gene model, *PstI* cutting sites and SNPs through whole genome. Each section referred to one chromosome, labeling on the circumference. A, B, C and D circles indicated the distribution of ITAG2.4 genes, expected *PstI* cutting sites, *PstI* cutting sites in the deep sequencing regions and RADseq SNPs, respectively. The black lines in the inner of D layer indicated the heterochromatic regions.

### 2.3.3 Genetic differentiation of *S. pimpinellifolium* was corresponding to the geographic area

The collection of 98 *S. pimpinellifolium* accessions was divided into three single-ancestral subpopulations and four mixed-ancestral subpopulations by the ADMIXTURE software (Figure 2.2A; S\_Fig 2.1). We named the red, blue, and green single-ancestral subpopulations POP S1, POP S2, and POP S3, respectively (Table 2.2). Meanwhile, the red-blue, blue-green, red-green, and red-blue-green mixed-ancestral subpopulations were named as POP M1, POP M2, POP M3, and POP M4, respectively



(Table 2.2). POP S1, POP S2, and POP S3 were clustered separately in the PCA plot, in which the first and the second principal components counted for 16.04% and 8.00% of the variance, respectively (Figure 2.2B). Moreover, pairwise  $F_{st}$  confirmed the genetic differentiation (S\_Tab 2.4), and AMOVA revealed that the variance between subpopulations was 41.96% (p-value < 0.001).

Table 2.2 Genetic variation of each subpopulation.

Subpopulation ID <sup>a</sup>	Genome pattern in ADMIXTURE	Sample size	Missing (%)	$H_o^b$	$H_s^c$
<b>Total</b>		<b>98</b>	<b>5.72</b>	<b>0.0761</b>	<b>0.2786</b>
POP S1	Red group	7	6.14	0.0660	0.1856
POP S2	Blue group	15	4.87	0.0558	0.1947
POP S3	Green group	21	6.70	0.0451	0.1549
POP M1	Red-Blue group	33	6.57	0.0948	0.2714
POP M2	Blue-Green group	15	3.63	0.0779	0.1913
POP M3	Red-Green group	4	4.78	0.1188	0.2133
POP M4	Red-Blue-Green	3	4.45	0.1468	0.1850

<sup>a</sup>: POP S indicates single ancestral subpopulation; POP M indicates mixed ancestral subpopulation.

<sup>b</sup>:  $H_o$  indicates the observed heterozygosity.

<sup>c</sup>:  $H_s$  indicates the within-population gene diversity (or “expected heterozygosity”).

The within-population gene diversity was calculated to compare genetic variation within each subpopulation. POP S2 and POP M1 showed the highest genetic variation among the single-ancestral subpopulations and the mixed-ancestral subpopulations, respectively (Table 2.2). Both subpopulations were in northern Peru, which indicated that northern Peru is the origin of *S. pimpinellifolium*.

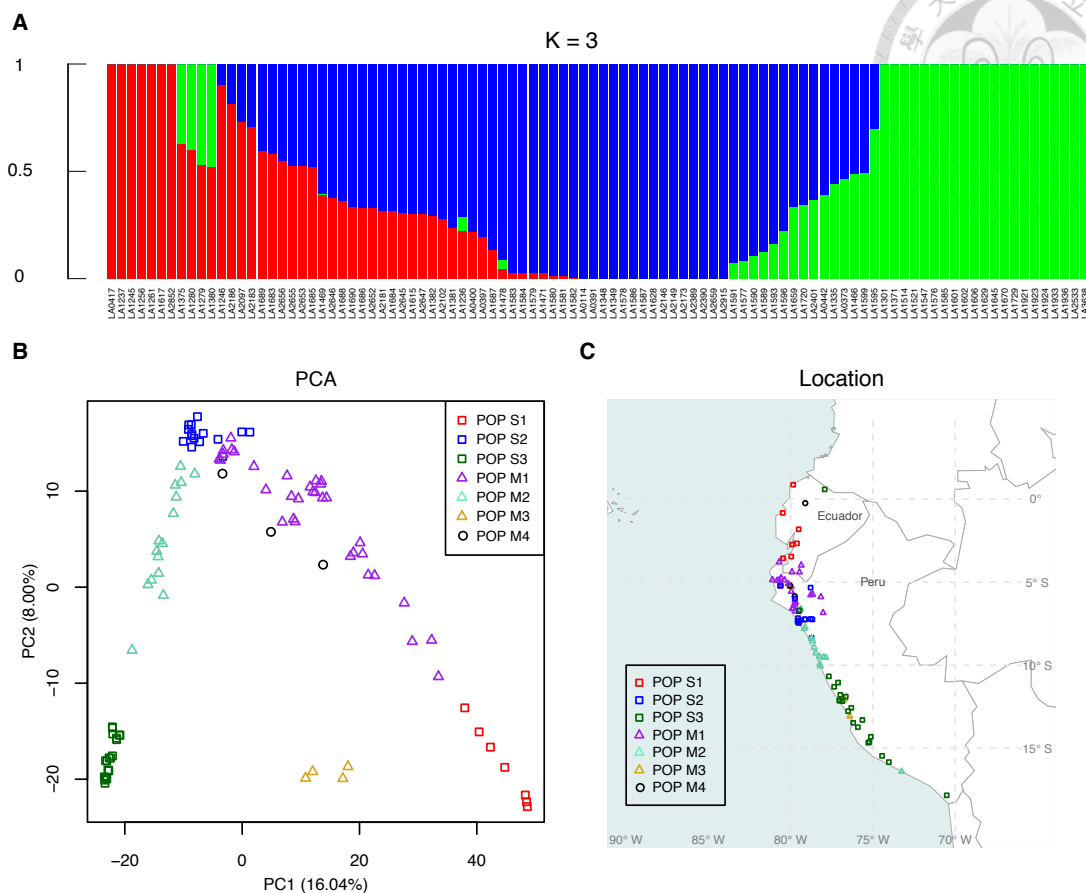


Figure 2.2 Ancestry and geographic distribution of 98 *Solanum pimpinellifolium* accessions from the Tomato Genetics Resource Center. A) Model-based ancestry for each accession. B) Principle component analysis of the *S. pimpinellifolium* population. C) Geographical distribution of the 98 *S. pimpinellifolium* accessions. Symbol and color codes are as follows: square symbols with red, blue and green colors indicate three single ancestral subpopulations corresponding to the same colors in the ancestry plot (POP S1, POP S2 and POP S3, respectively); triangle symbols with purple, aquamarine and goldenrod colors present the POP M1, POP M2 and POP M3, respectively; black circle symbols were the POP M4.

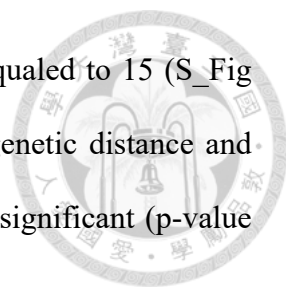
Interestingly, most accessions in the same subpopulation were in the same vicinity of their collection sites (Figure 2.2C). Also, POP S1, POP S2, and POP S3 spread in somewhat distinct geographic areas along the coastline from Ecuador to southern Peru (Figure 2.2C). The geographic distribution of these subpopulations appeared in the following order from north to south: POP S1, POP M1, POP S2, POP M2, and POP S3 (Figure 2.2C). This geographic distribution showed a trend in which the mixed-ancestral

subpopulations were located between their corresponding single-ancestral subpopulations. For the analysis of isolation by distance (IBD) using all pairs of samples, the correlation coefficient between the genetic distance and geographic distance was 0.34, and this correlation was statistically significant ( $p$ -value  $< 0.001$ ) (S\_Fig 2.2).

### 2.3.4 Meta-analysis of SolCAP genotyping array resulted in 15 subpopulations

To compare with our analysis of the genetic differentiation of *S. pimpinellifolium* in the current study, we performed a meta-analysis of the genetic differentiation of *S. pimpinellifolium* using combined SNP-marker genotypic data of SolCAP array from the previous studies. We downloaded the genotypes of 214 samples representing 126 accessions from three previous studies (Blanca et al., 2012; Blanca et al., 2015; Sim et al., 2012) and conducted the meta-analysis using our workflow (please see details in the “Materials and Methods” section) (S\_Tab 2.7). Initially, we extracted a marker set of 2,934 bi-allelic SNPs to investigate genetic diversity between samples from different studies but tagged the same name. The samples in Blanca *et al.*, 2012 separated from those of the other two studies in the PCA plot (S\_Fig 2.3A), while most of the accessions in Blanca *et al.*, 2012 were involved in the study of Blanca *et al.*, 2015 (S\_Tab 2.5). It suggested that the batch effect occurred when these datasets merged. Considering the SolCAP genotyping array is an Illumina bead array, which uses the TOP/BOT strand and A/B allele designation to assign the actual polymorphism of samples, data merging might introduce reverse-complement allele designation (Illumina, 2014). We resolved the problem of the batch effect after we removed the markers with inconsistent SNP assignment among these three datasets (S\_Fig 2.3B). The genotypic data of 2,307 SNPs in 214 samples was remained (S\_Tab 2.5 and S\_Tab 2.6) and used

to conduct further analyses. ADMIXTURE suggested the best K equaled to 15 (S\_Fig 2.4 and S\_Fig 2.5). Also, the correlation coefficient between the genetic distance and geographic distance was 0.55, and this correlation was statistically significant (p-value < 0.001) (S\_Fig 2.6).



### 2.3.5 Rapid LD decay

LD decay was estimated for the mapping resolution in GWAS. In this population, the non-linear regression curve dropped very quickly (S\_Fig 2.7). Following the non-linear regression curve, the overall LD decay was within 18 Kb when the baseline of the  $r^2$  value was set at 0.1 (Table 2.3; Figure 2.3A). The fastest LD decay was within 10 Kb on chromosome 9 while the slowest LD decay was within 30 Kb on chromosome 4 (Table 2.3; S\_Fig 2.8).

Table 2.3 The local LD profiles of individual chromosomes.

Chr.	LD decay (Kb)	For paired flanking sequencing units		Proportion of LD for paired flanking sequencing units (%)
		Number of $r^2 \geq 0.1$	Number of $r^2 < 0.1$	
1	14	632	1,130	35.87
2	12	475	881	35.03
3	15	460	927	33.17
4	30	423	687	38.11
5	21	309	514	37.55
6	20	428	750	36.66
7	21	397	581	40.59
8	28	401	618	39.35
9	10	280	617	31.22
10	19	330	525	38.60
11	19	310	535	36.69
12	17	253	539	31.94
<b>Total</b>	<b>18</b>	<b>4,698</b>	<b>8,304</b>	<b>36.13</b>

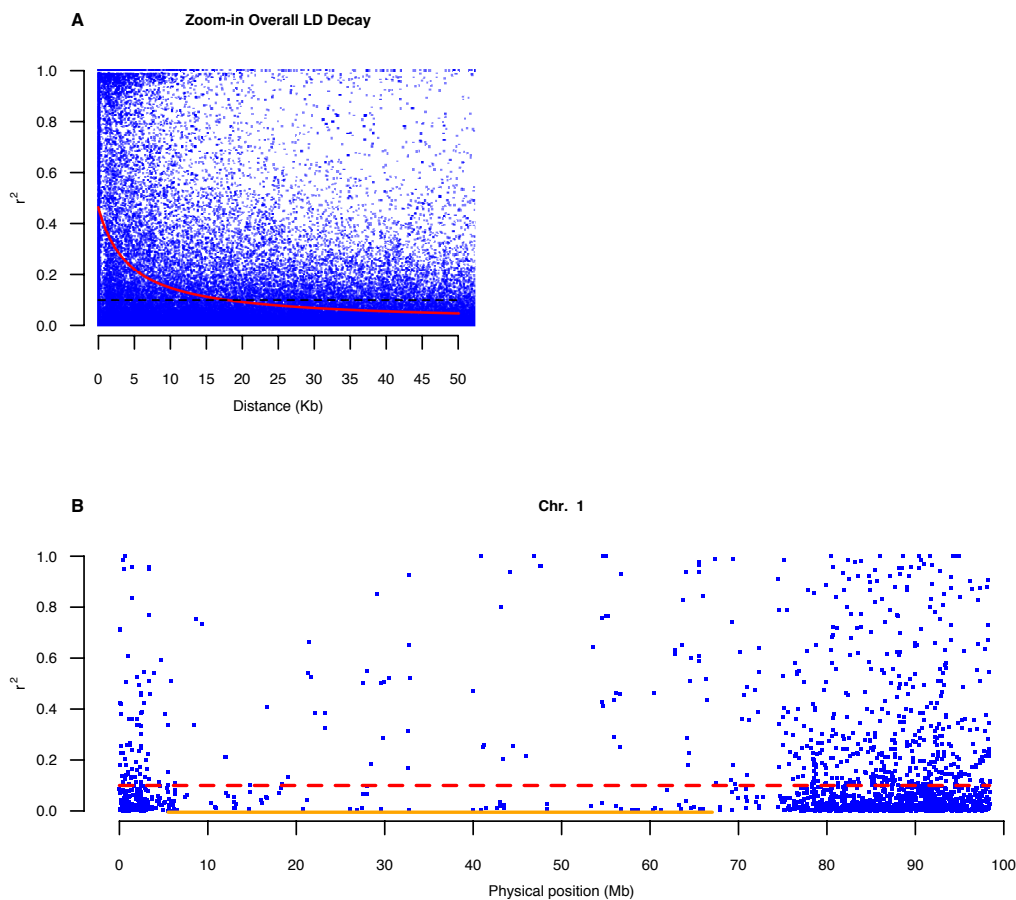


Figure 2.3 Visualization for LD. A) The 50 Kb interval of overall LD decay. The red curve indicated non-linear regression and black dotted line referred to the baseline of  $r^2$  at 0.1. B) The local LD of chromosome 1. The red dotted line was the baseline of  $r^2$  and the orange line indicated the heterochromatic region.

### 2.3.6 Heterogeneity of genetic recombination within each chromosome

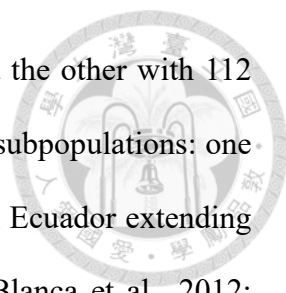
LD decay of individual chromosomes was insufficient to capture the local variations of historically accumulated recombination events because the tomato genome comprises more than 75% heterochromatin which usually suppresses recombination events (Sim et al., 2012). We assessed the local LD profile of individual chromosomes based on the average  $r^2$  value of flanking sequencing units that contained at least one SNP marker. We observed two main trends: marker density in the heterochromatic

regions was lower than that in the euchromatic regions (Figure 2.3B; S\_Fig 2.9), and approximately two-thirds of the  $r^2$  values were less than 0.1 (Table 2.3). The latter observation indicated that these flanking SNP markers were not in a state of linkage disequilibrium.

## 2.4 Discussion

### 2.4.1 Subpopulations clustering from north to south are expected due to the high correlation between genetic distance and geographic distance

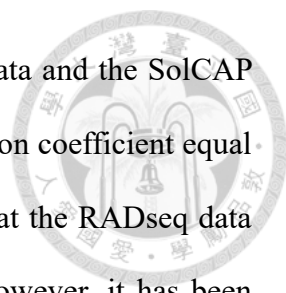
The genetic differentiation revealed in this study should be similar to previous findings because the collection sites of this collection cover most of recorded habitats of *S. pimpinellifolium*. One previous study for the genetic diversity of *S. pimpinellifolium* assessed 213 accessions with the genotypes of 10 SSR markers. It suggested the existence of Peruvian and Ecuadorian subpopulations (Zuriaga et al., 2009). Another study investigated a collection of 190 *S. pimpinellifolium* accessions using 48 SSR markers (Rao et al., 2012). It evaluated 120 accessions collected from Peru and 31 accessions from Ecuador, and divided these accessions into two single-ancestral subpopulations and one mixed-ancestral subpopulation. One of the single-ancestral subpopulations contained 93 accessions from Peru and 3 Ecuadorian accessions. These three Ecuadorian accessions were the only Ecuadorian accessions that were grouped into this single-ancestral subpopulation that contained mainly the Peruvian accessions, and the duplicated entries with the same names of these Ecuadorian accessions (LA0411, LA1246, LA1261) in the same study were grouped into the other two subpopulations. Despite of these three confounded Ecuadorian accessions, this study still inferred strong correlation between genetic diversity and geographic distance between Peruvian and Ecuadorian subpopulations (Rao et al., 2012). With the aid of SolCAP array, two



consecutive studies, one with 63 *S. pimpinellifolium* accessions and the other with 112 *S. pimpinellifolium* accessions, sorted *S. pimpinellifolium* into three subpopulations: one in northern Ecuador, another in the mountainous area from southern Ecuador extending to northern Peru and the third in the low-altitude areas of Peru (Blanca et al., 2012; Blanca et al., 2015). Our study also supports three single-ancestral subpopulations: one in Ecuador, one in northern Peru, and another in southern Peru. Among all the aforementioned studies, two ancestral subpopulations are confident: one includes the accessions in Ecuador; the other includes the accessions in southern Peru. The different grouping among these studies for those accessions from southern Ecuador to northern Peru may result from different markers and different genetic diversity in each study.

Previous studies suggested that genetic differentiation of *S. pimpinellifolium* correlated to the climatic variation (Rick *et al.* 1977; Zuriaga *et al.* 2009; Blanca *et al.* 2012, 2015). The analysis of genetic differentiation based on the RADseq data in the current study supported the same conclusion: most POP S1 accessions are in hot and humid Ecuador; most POP M1 scatter in northern Peru, along the western Andean slopes, in which is a warm desert; most POP S2 are located in the Andean Mountains; most POP M2 are in a warm semi-arid region; most POP S3 spread along the coastal region from central to southern Peru, in which is a relatively cold desert (S\_Tab 2. 1 and Figure 2.2C). Since these subpopulations are located in the environments with different climates, and  $F_{st}$  as well as AMOVA support these subpopulations (S\_Tab 2.4), the genetic differentiation of *S. pimpinellifolium* is observed evidently with the aid of RADseq SNP markers.

Isolation by distance (IBD) is a common tool to access genetic differentiation that expect a positive correlation between genetic variation and geographic distance (Wright,



1943). We conducted this analysis for both datasets, the RADseq data and the SolCAP array data, and made comparisons. The former data had the correlation coefficient equal to 0.34, and the latter one was 0.55 (S\_Fig 2; S\_Fig 6). It seems that the RADseq data showed less genetic differentiation than the SolCAP array data. However, it has been argued that IBD test can be severely biased in two situations: unequal migration among all populations in a system, and the detection of loci under selection (Meirmans, 2012). We do not know whether the investigated accessions were equally migrated, but we do know that the SolCAP array was designed mainly on the SNP sites of coding sequences within cultivated tomatoes or between cultivated tomato and wild tomatoes (Sim et al., 2012). Therefore, the SNPs on the SolCAP array had higher chances under selection in domestication. Under this premise, the comparisons of the IBD test between the RADseq data and the SolCAP array data could be confounded by the differences in selection strength.

#### **2.4.2 Discrepancy of genetic clustering in SolCAP meta-analysis**

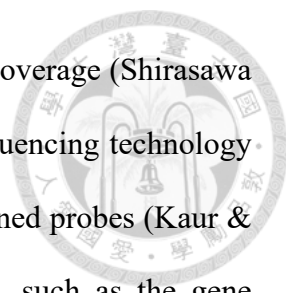
Our meta-analysis concluded that the genetic compositions of 214 samples came from 15 ancestral populations. This conclusion is different from the conclusion of Blanca et al. (2012) and our RADseq data, both of which suggested that there were three ancestral populations of *S. pimpinellifolium*. It implied an unclear structure; especially the cross validation error has an ambiguous minimal value (S\_Fig 2.4). It is possible that genetic diversity between wild tomatoes are underestimated because the polymorphisms of SolCAP array are selected between cultivars and wild tomatoes (Sim et al., 2012). We noticed that two samples of LA0373 with 76% identity display different genome patterns in ADMIXTURE, while two samples of LA1478 with 71% identity present different patterns as well (S\_Tab 2.7; S\_Fig 2.5). Since two samples of



the same accession demonstrate dissimilar genome patterns, the SolCAP may be less appropriate to quantize the population structure of *S. pimpinellifolium*, especially when more samples are involved. Also, for the same reason, we cannot validate the genetic differentiation in the SolCAP meta-analysis by  $F_{st}$  or AMOVA nor achieve a stable estimation of genetic differentiation in a scenario of more accessions via the SolCAP meta-analysis.

### **2.4.3 More markers are required to cover through the genome of *S. pimpinellifolium***

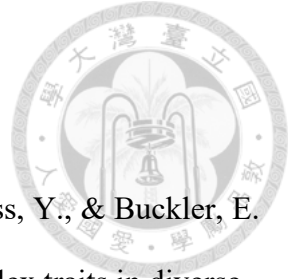
The observed and expected heterozygosity of this population were 0.0761 and 0.2786, respectively, slightly higher than those in previous researches (Blanca et al., 2012; Blanca et al., 2015). Since *S. pimpinellifolium* was detected with up to a 40% outcrossing rate (Rick et al., 1977) and demonstrated high genetic variation, it is expected to cause rapid LD decay. In this study, LD decay was within 18 Kb throughout the genome, which was much shorter than cultivated tomatoes (Bauchet et al., 2017; Sim et al., 2012). However, to put at least one SNP marker within each of 18 Kb intervals in this genome, the 900-Mb tomato genome would require at least 50,000 markers to fulfill QTL detection in GWAS. Therefore, acquiring many SNPs using different methods is essential to conduct a GWAS in the *S. pimpinellifolium* population. Here, we proposed three possible approaches to increase markers. One is to increase the sample size evenly for each subpopulation (Brachi, Morris, & Borevitz, 2011). Since approximately 64% of alleles were rare in this population, the augmentation of the subpopulation size may adjust rare alleles to common alleles, potentially increasing the SNPs without extending coverage. One is to construct DNA libraries with a frequently cutting restriction enzyme. This approach can be simulated and optimized *in silico* to



balance sequencing resource between sample sizes and sequencing coverage (Shirasawa et al., 2016). Another is exome sequencing, a selective genome sequencing technology that selects desired sequencing regions by the hybridization of designed probes (Kaur & Gaikwad, 2017). Based on tomato genome sequence information, such as the gene model or EST database, one could design different sets of probes to limit sequencing regions (Ruggieri et al., 2017). Given the approximately 110 Mb total gene length in the ITAG2.4 gene model, the potential coverage could reach 12% and all target the gene region. This exome sequencing strategy may be able to increase SNPs without increasing population size.

## 2.5 Reference

- Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*.  
<https://doi.org/10.1101/gr.094052.109>
- Bauchet, G., Grenier, S., Samson, N., Bonnet, J., Grivet, L., & Causse, M. (2017). Use of modern tomato breeding germplasm for deciphering the genetic control of agronomical traits by Genome Wide Association study. *Theoretical and Applied Genetics*. <https://doi.org/10.1007/s00122-017-2857-9>
- Blanca, J., Cañizares, J., Cordero, L., Pascual, L., Díez, M. J., & Nuez, F. (2012). Variation Revealed by SNP Genotyping and Morphology Provides Insight into the Origin of the Tomato. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0048198>
- Blanca, J., Montero-Pau, J., Sauvage, C., Bauchet, G., Illa, E., Díez, M. J., ... Cañizares, J. (2015). Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. *BMC Genomics*. <https://doi.org/10.1186/s12864-015-1444-1>
- Brachi, B., Morris, G. P., & Borevitz, J. O. (2011). Genome-wide association studies in



plants: The missing heritability is in the field. *Genome Biology*.

<https://doi.org/10.1186/gb-2011-12-10-232>

Bradbury, P. J., Zhang, Z., Koon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E.

S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btm308>

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013).

Stacks: An analysis tool set for population genomics. *Molecular Ecology*.

<https://doi.org/10.1111/mec.12354>

Dobritsa, A. P., & Dobritsa, S. V. (1980). DNA protection with the DNA methylase M ·

*BbvI* from *Bacillus brevis* var. GB against cleavage by the restriction endonucleases *PstI* and *PvuII*. *Gene*.

[https://doi.org/10.1016/0378-1119\(80\)90128-6](https://doi.org/10.1016/0378-1119(80)90128-6)

Etter, P. D., Bassham, S., Hohenlohe, P. A., Johnson, E. A., & Cresko, W. A. (2011).

SNP Discovery and Genotyping for Evolutionary Genetics Using RAD Sequencing. *Molecular Methods for Evolutionary Genetics*, 772(2), 1–19.

<https://doi.org/10.1007/978-1-61779-228-1>

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance

inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics*.

Fernandez-Pozo, N., Menda, N., Edwards, J. D., Saha, S., Teclé, I. Y., Strickler, S.

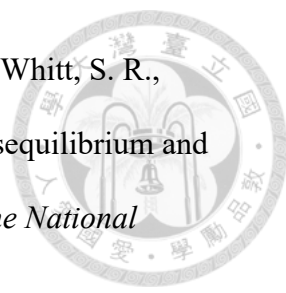
R., ... Mueller, L. A. (2015). The Sol Genomics Network (SGN)-from genotype to phenotype to breeding. *Nucleic Acids Research*.

<https://doi.org/10.1093/nar/gku1195>

Fulton, T. M., Chunwongse, J., & Tanksley, S. D. (1995). Microprep protocol for

extraction of DNA from tomato and other herbaceous plants. *Plant Molecular*

- Biology Reporter*, 13(3), 207–209. <https://doi.org/10.1007/BF02670897>
- Gaunt, T. R., Rodríguez, S., & Day, I. N. M. (2007). Cubic exact solutions for the estimation of pairwise haplotype frequencies: Implications for linkage disequilibrium analyses and a web tool “CubeX.” *BMC Bioinformatics*. <https://doi.org/10.1186/1471-2105-8-428>
- Goudet, J., & Jombart, T. (2015). hierfstat: Estimation and Tests of Hierarchical F-Statistics.
- Hijmans, R. J. (2016). geosphere: Spherical Trigonometry. R package version 1.5-5.
- Illumina. (2014). Infinium ® Genotyping Data Analysis. *Technical Note*, 10p. <https://doi.org/10.1111/j.1532-950X.2005.00092.x>
- Jombart, T. (2008). Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btn129>
- Kaur, P., & Gaikwad, K. (2017). From Genomes to GENE-omes: Exome Sequencing Concept and Applications in Crop Improvement. *Frontiers in Plant Science*. <https://doi.org/10.3389/fpls.2017.02164>
- Meirmans, P. G. (2012). The trouble with isolation by distance. *Molecular Ecology*. <https://doi.org/10.1111/j.1365-294X.2012.05578.x>
- Pembleton, L. W., Cogan, N. O. I., & Forster, J. W. (2013). StAMPP: An R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Molecular Ecology Resources*. <https://doi.org/10.1111/1755-0998.12129>
- Rao, E. S., Kadirvel, P., Symonds, R. C., Geethanjali, S., & Ebert, A. W. (2012). Using SSR markers to map genetic diversity and population structure of *Solanum pimpinellifolium* for development of a core collection. *Plant Genetic Resources: Characterisation and Utilisation*. <https://doi.org/10.1017/S1479262111000955>

- 
- Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J. F., ... Buckler, E. S. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.201394398>
- Rick, C. M., Fobes, J. F., & Holle, M. (1977). Genetic variation in *Lycopersicon pimpinellifolium*: Evidence of evolutionary change in mating systems. *Plant Systematics and Evolution*. <https://doi.org/10.1007/BF00984147>
- Rogers, J. S. (1972). Measures of similarity and genetic distance. In *In Studies in Genetics VII* (pp. 145–153). Austin, Texas: University of Texas Publication 7213.
- Ruggieri, V., Anzar, I., Paytuyi, A., Calafiore, R., Cigliano, R. A., Sanseverino, W., & Barone, A. (2017). Exploiting the great potential of Sequence Capture data by a new tool, SUPER-CAP. *DNA Research : An International Journal for Rapid Publication of Reports on Genes and Genomes*. <https://doi.org/10.1093/dnares/dsw050>
- Shirasawa, K., Hirakawa, H., & Isobe, S. (2016). Analytical workflow of double-digest restriction site-associated DNA sequencing based on empirical and *in silico* optimization in tomato. *DNA Research*. <https://doi.org/10.1093/dnares/dsw004>
- Sim, S. C., Durstewitz, G., Plieske, J., Wieseke, R., Ganal, M. W., van Deynze, A., ... Francis, D. M. (2012). Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0040563>
- Sim, S. C., van Deynze, A., Stoffel, K., Douches, D. S., Zarka, D., Ganal, M. W., ... Francis, D. M. (2012). High-Density SNP Genotyping of Tomato (*Solanum lycopersicum* L.) Reveals Patterns of Genetic Variation Due to Breeding. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0045520>

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38(6), 1358–1370.

Wright, S. (1943). Isolation by distance. *Genetics*, 28, 114–138.

<https://doi.org/10.5194/isprs-Archives-XLII-5-W1-419-2017>

Zuriaga, E., Blanca, J. M., Cordero, L., Sifres, A., Blas-Cerdán, W. G., Morales, R., & Nuez, F. (2009). Genetic and bioclimatic variation in *Solanum pimpinellifolium*.

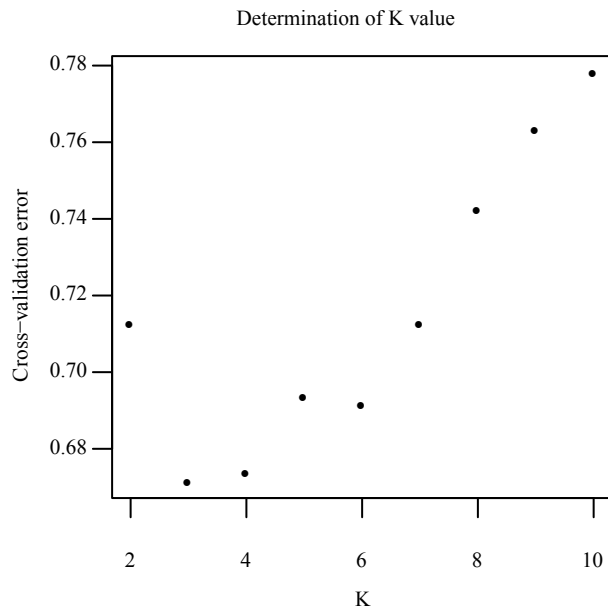
*Genetic Resources and Crop Evolution*.

<https://doi.org/10.1007/s10722-008-9340-z>

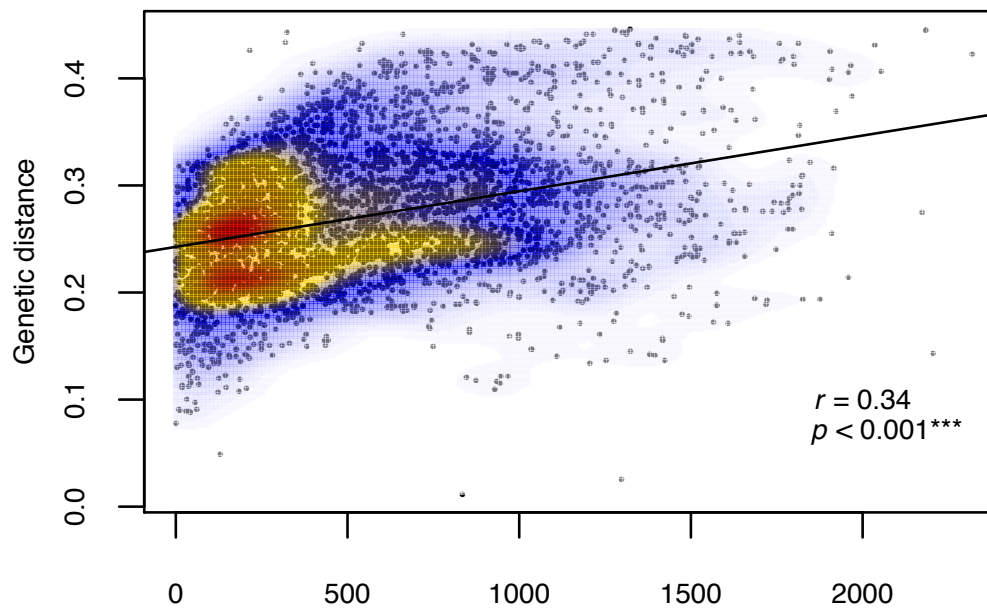




## 2.6 Supplementary data

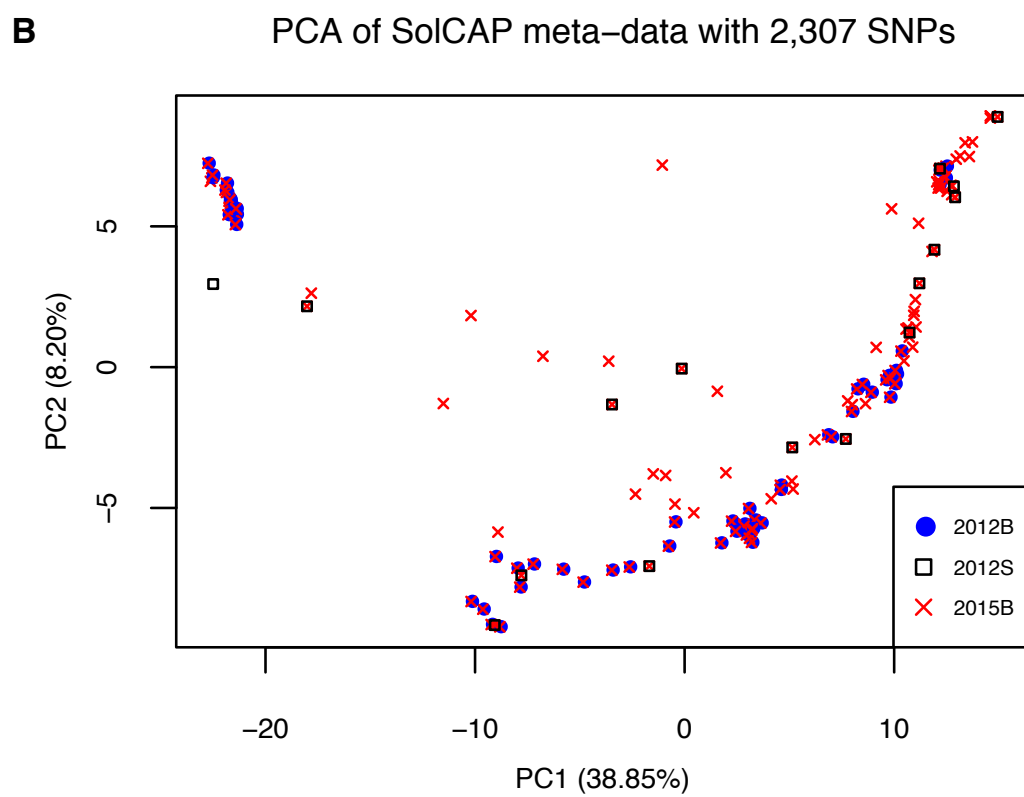
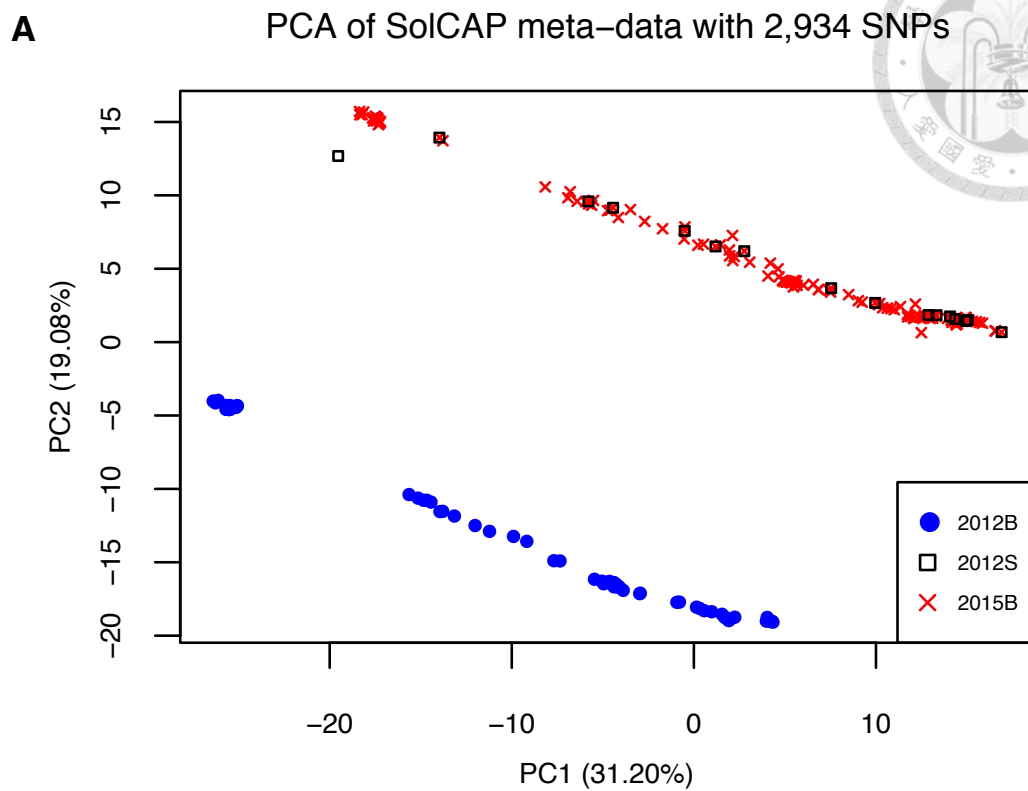


S\_Fig 2.1 The cross-validation error of K value in ADMIXTURE.



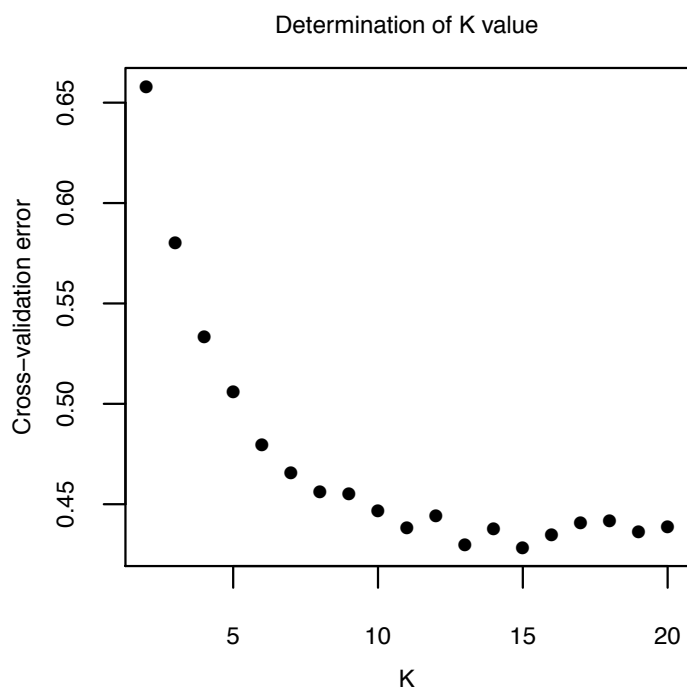
S\_Fig 2.2 Pairwise isolation by distance of 98 accessions.

Colors present the density from low (blue) to high (red).

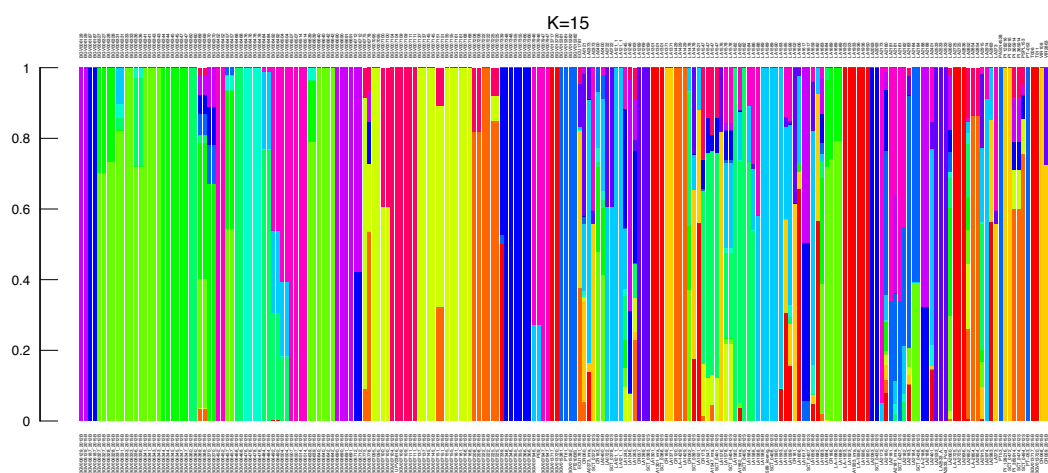


S\_Fig 2.3 The PCA of SolCAP meta-analysis. A) The PCA plot of bi-allelic SNPs. B) The PCA plot after removing those SNPs of reverse-complement allele designation.



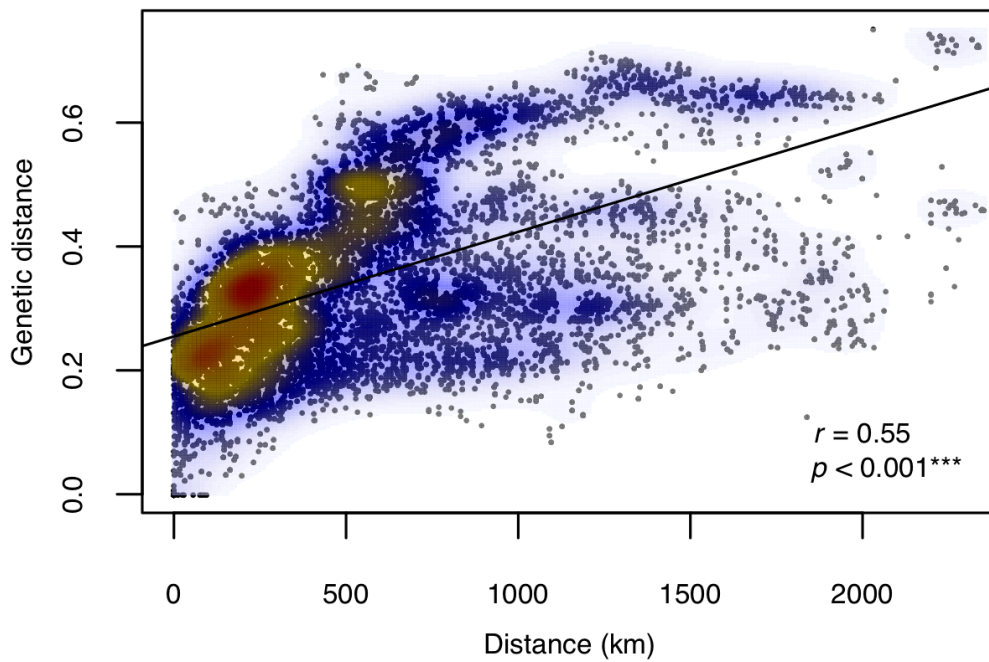


S\_Fig 2.4 The cross-validation error of SolCAP meta-analysis



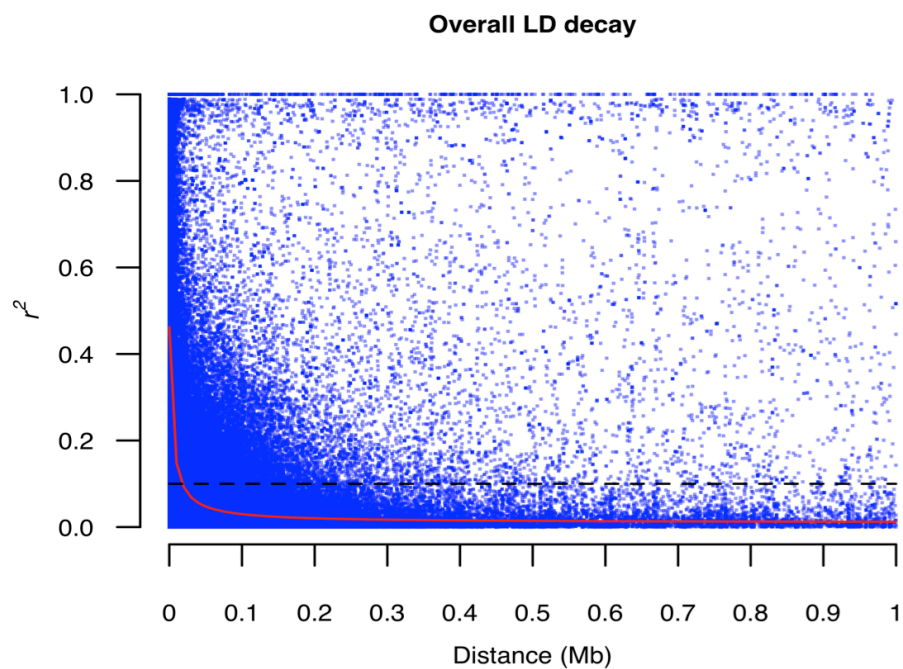
S\_Fig 2.5 The genome patterns of 214 samples in SolCAP meta-analysis.

The labels on the top indicate the accessions; the labels on the bottom indicate the sample ID in this meta-analysis.



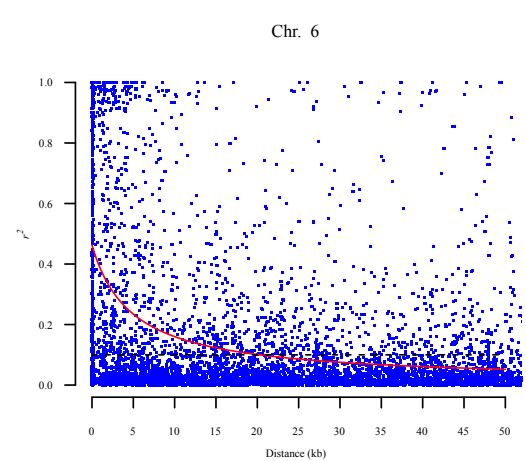
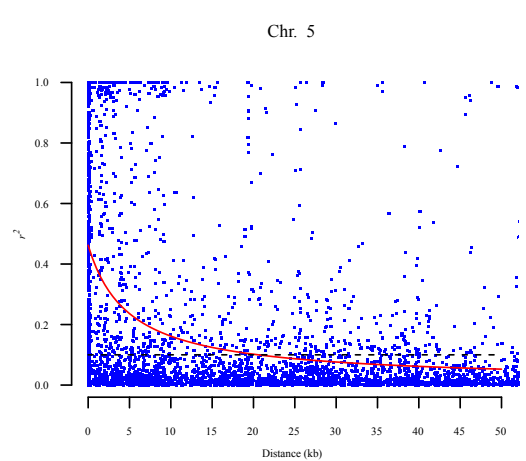
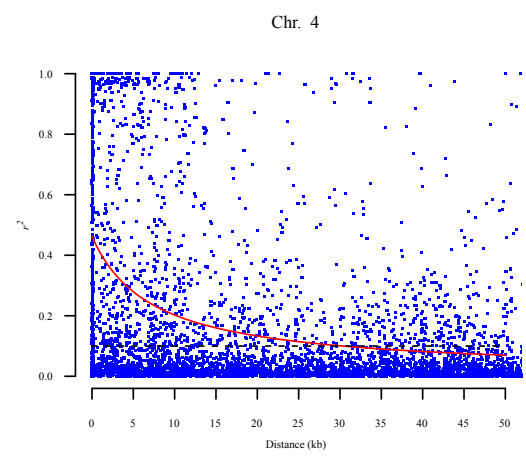
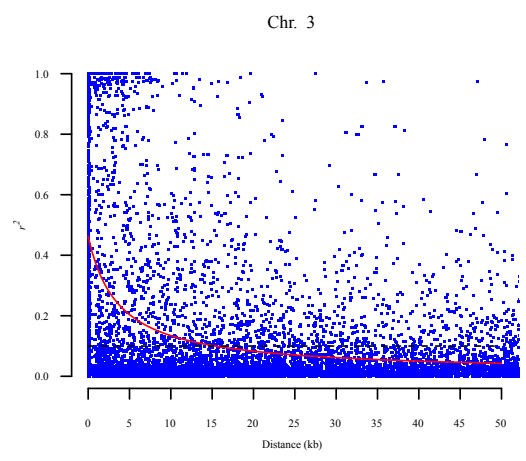
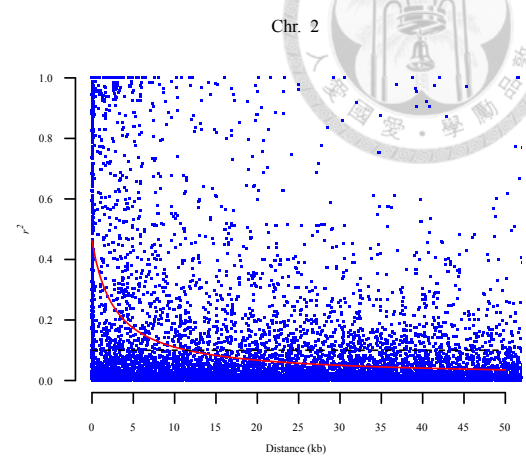
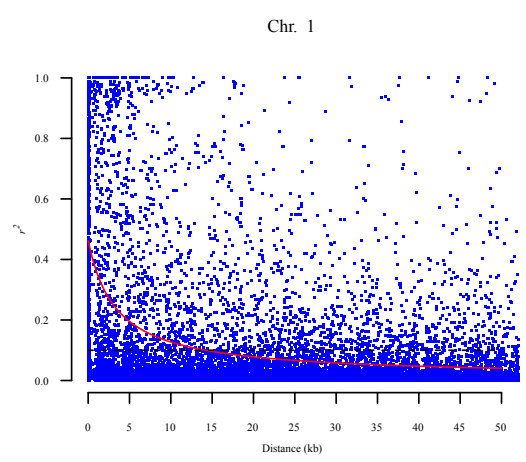
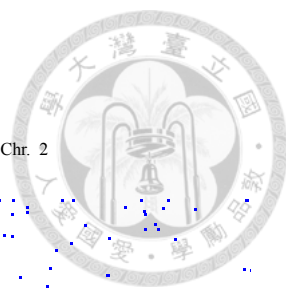
S\_Fig 2.6 Pairwise isolation by distance of SolCAP meta-analysis.

Colors present the density from low (blue) to high (red).

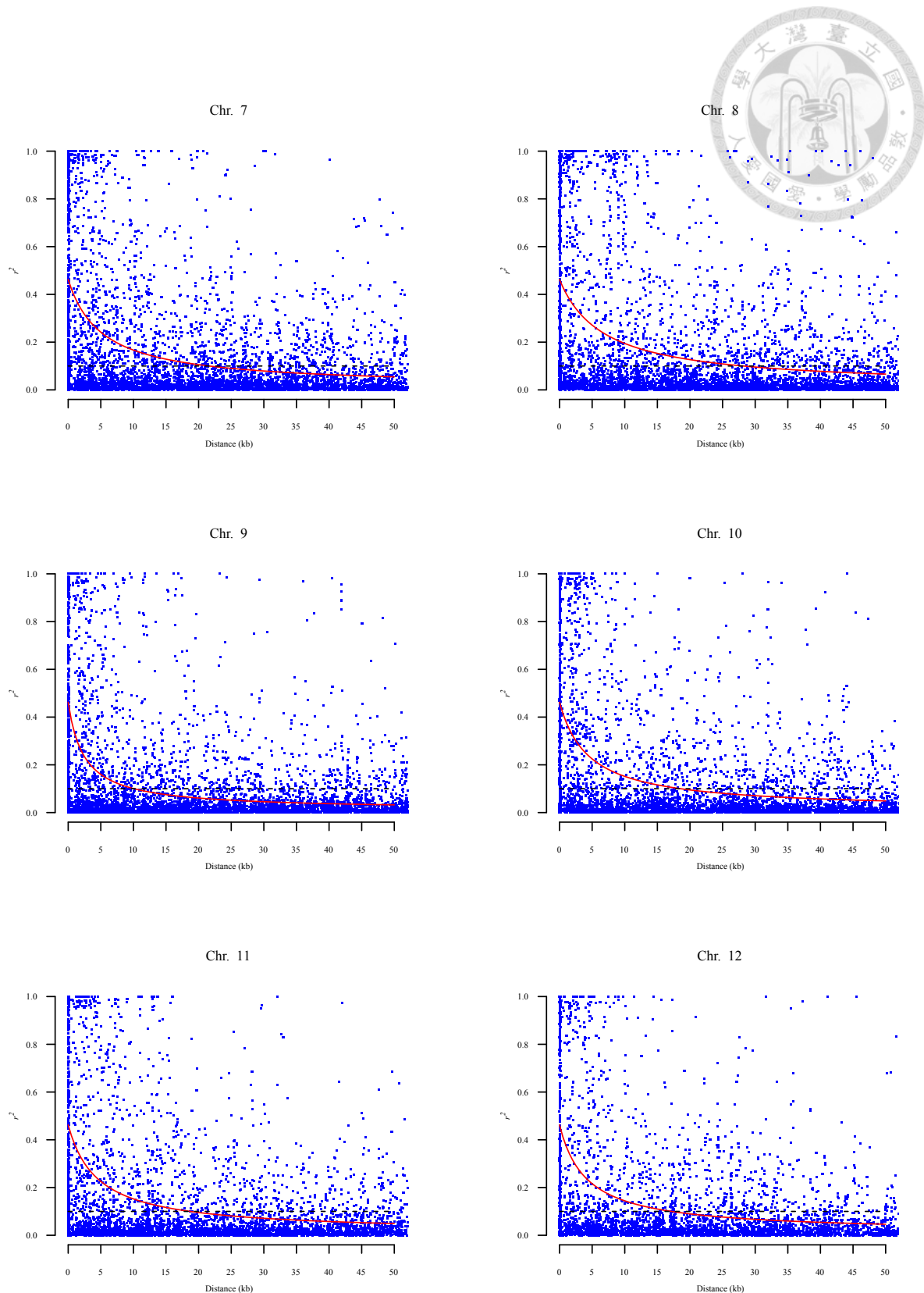


S\_Fig 2.7 LD decay of the whole genome.

The red curve indicates non-linear regression. The dotted line indicates the fixed  $r^2$  on 0.1.

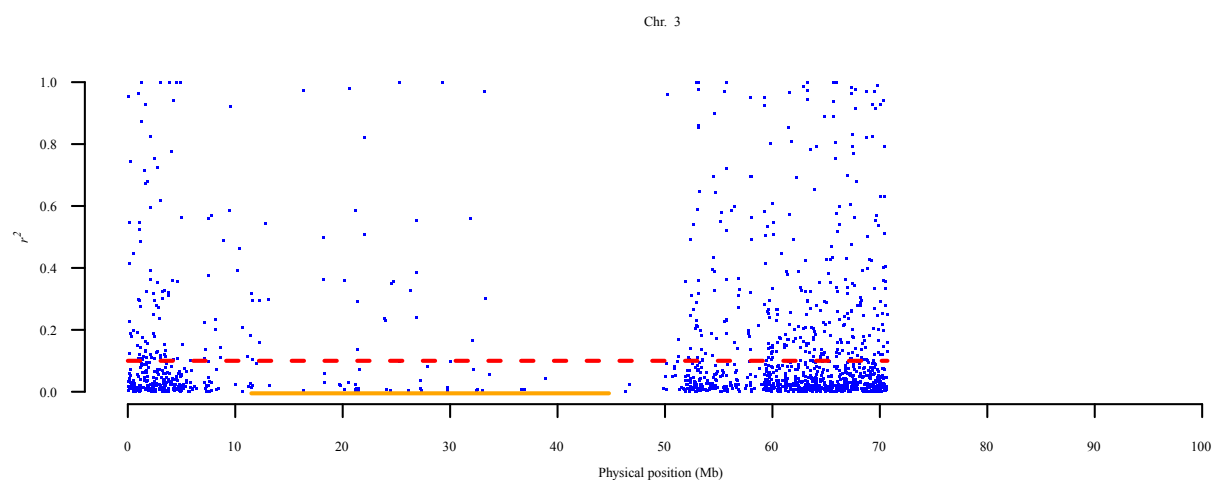
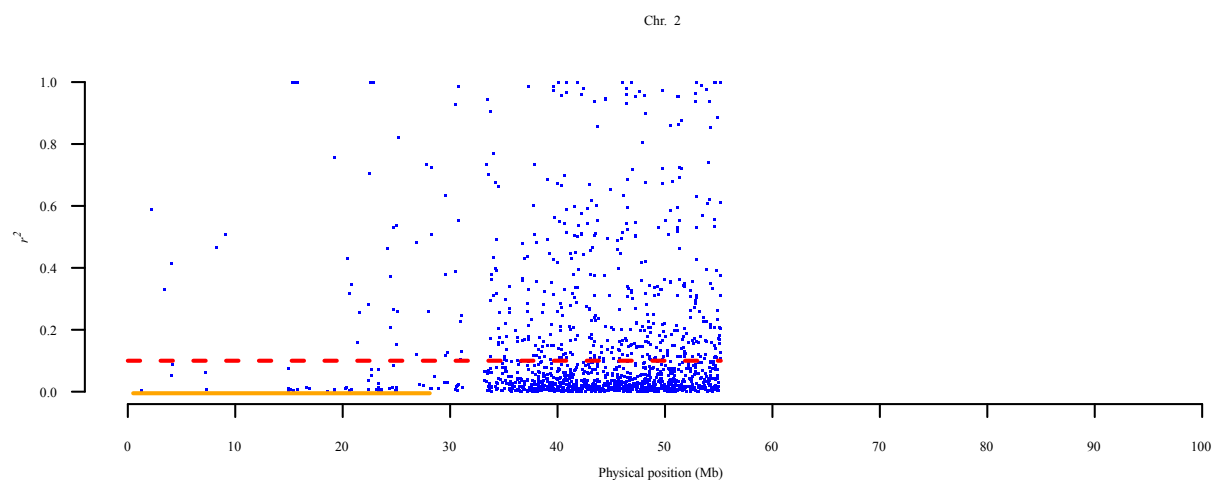
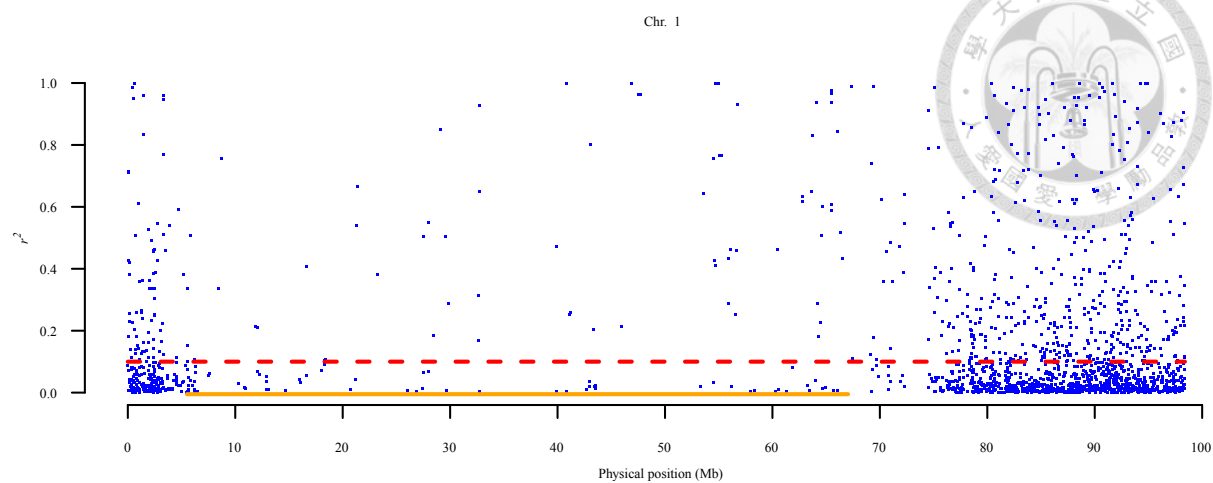
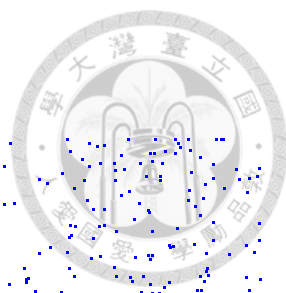


S\_Fig 2.8 (page 1/2)

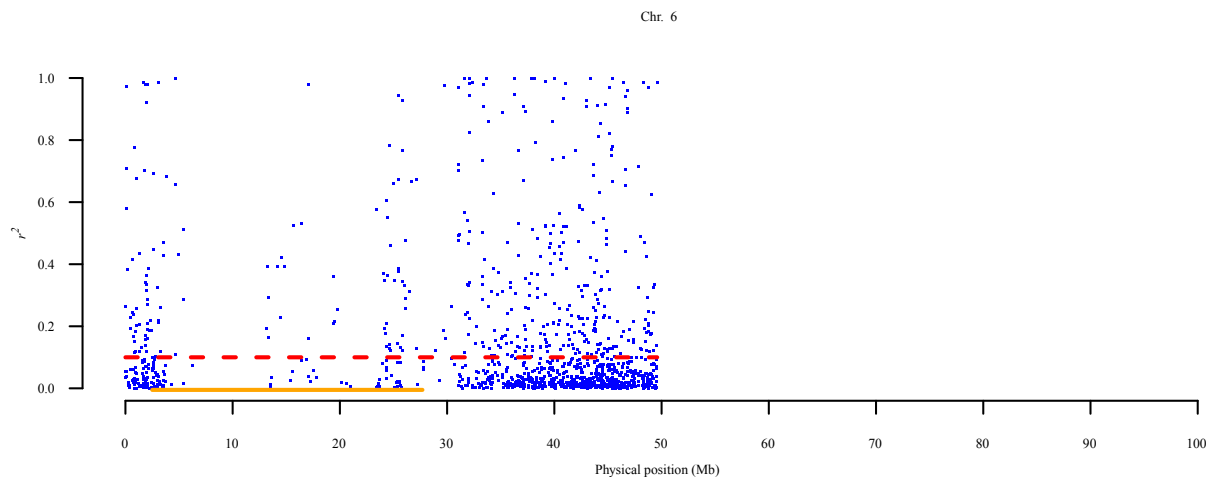
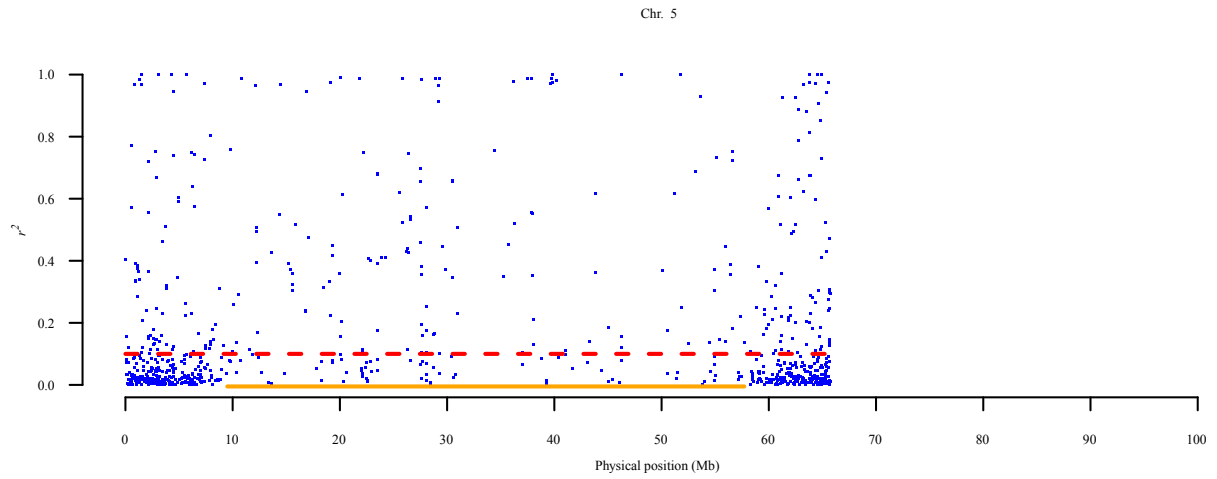
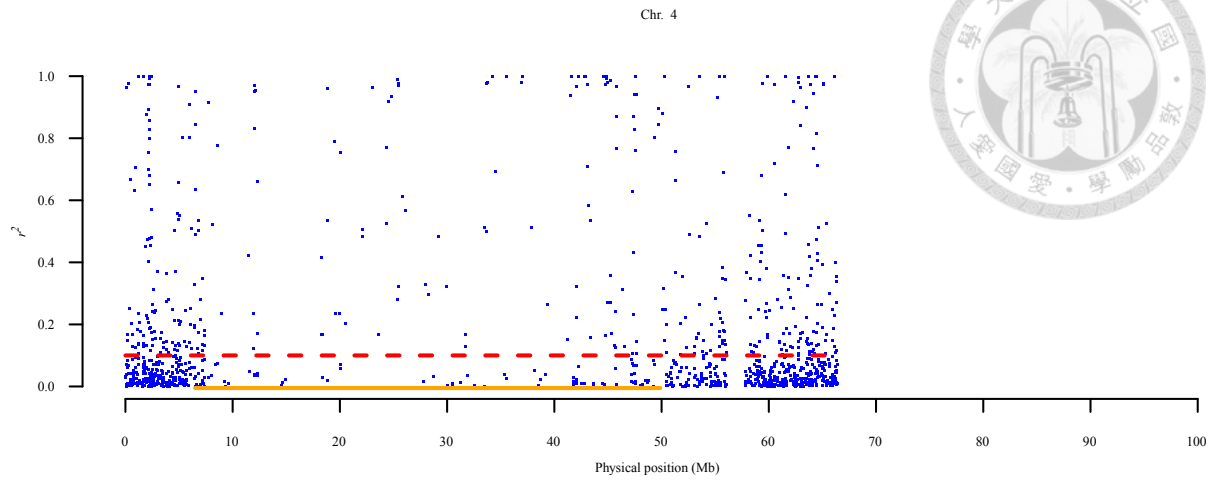


S\_Fig 2.8 50 kb interval LD decay of each chromosome.

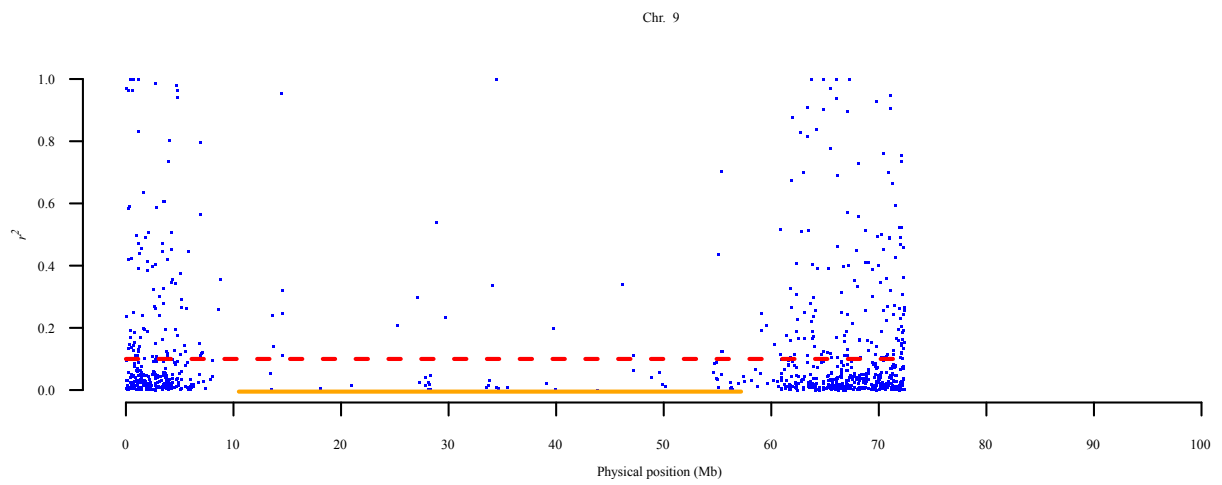
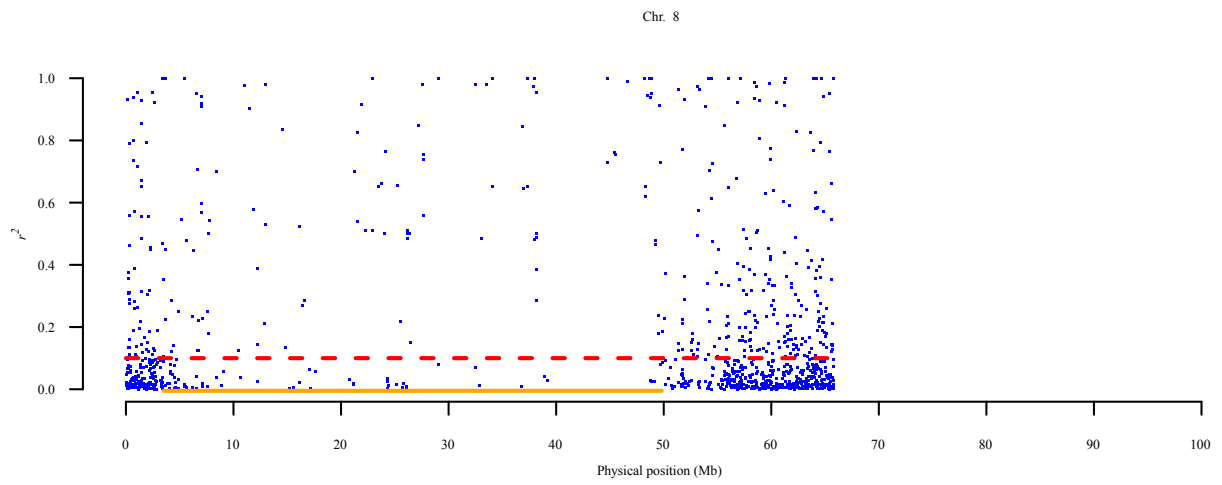
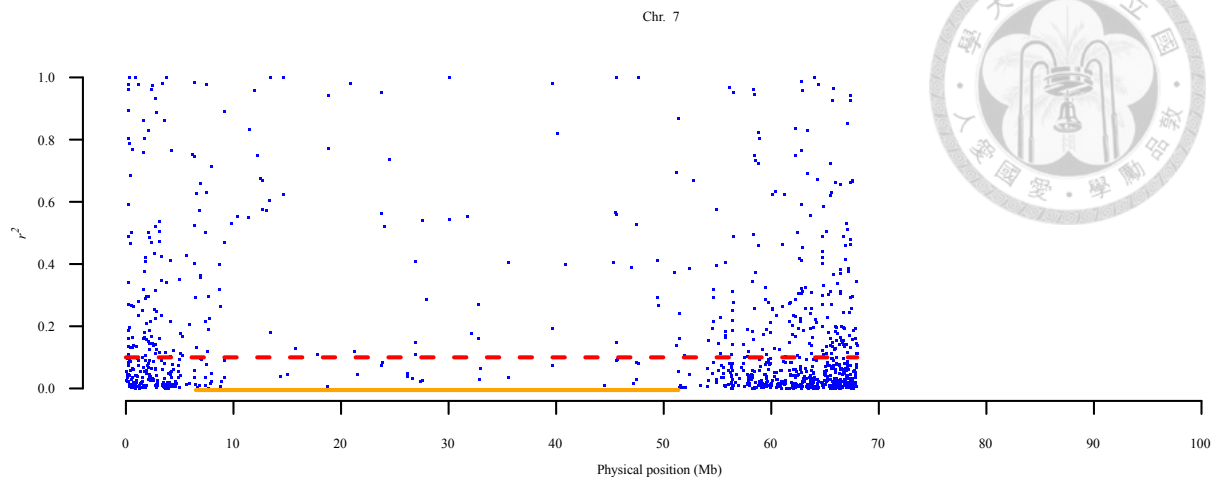
The red curves indicate non-linear regression. Black dotted lines indicate the fixed  $r^2$  on 0.1.



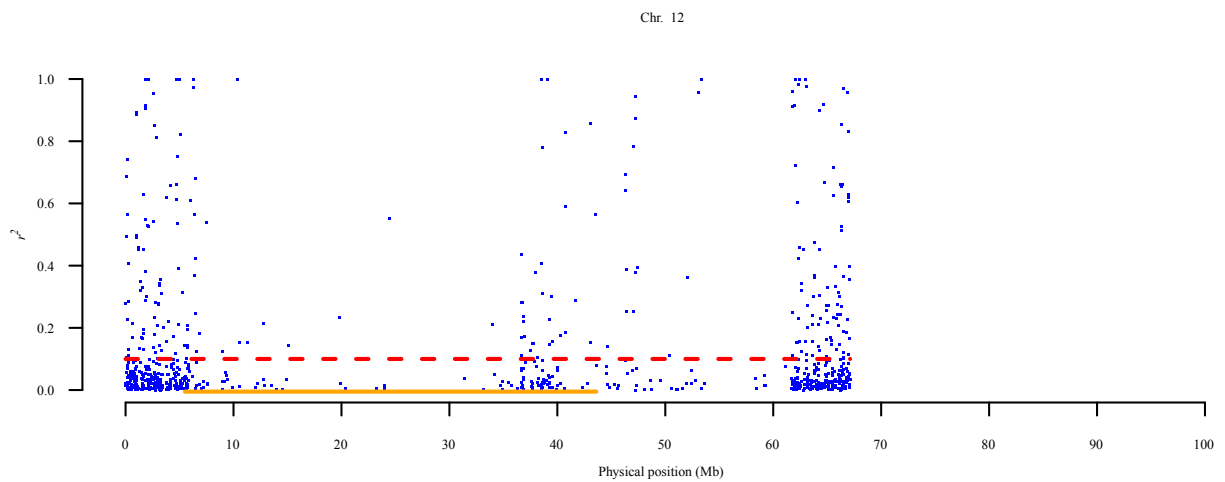
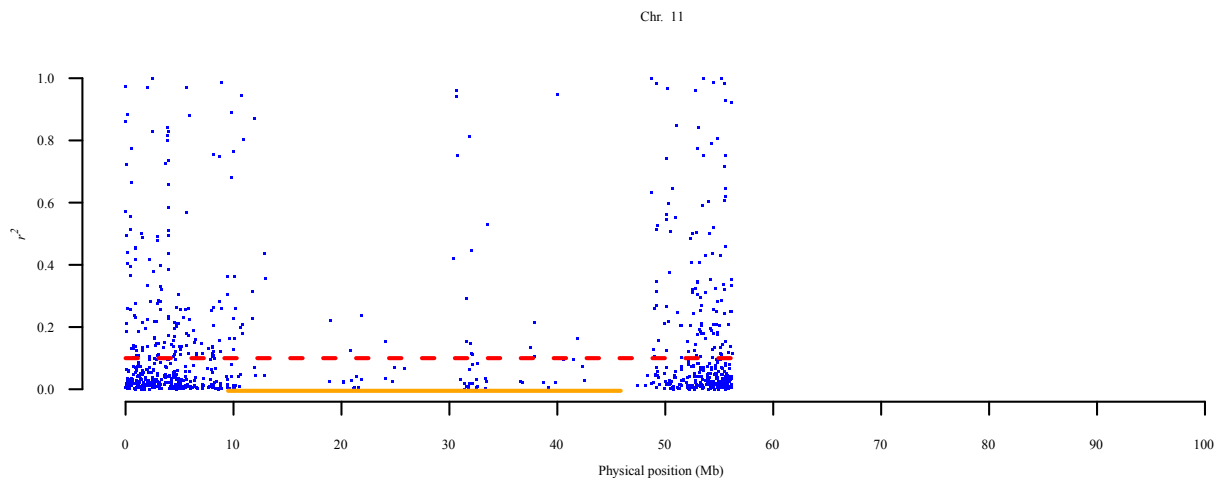
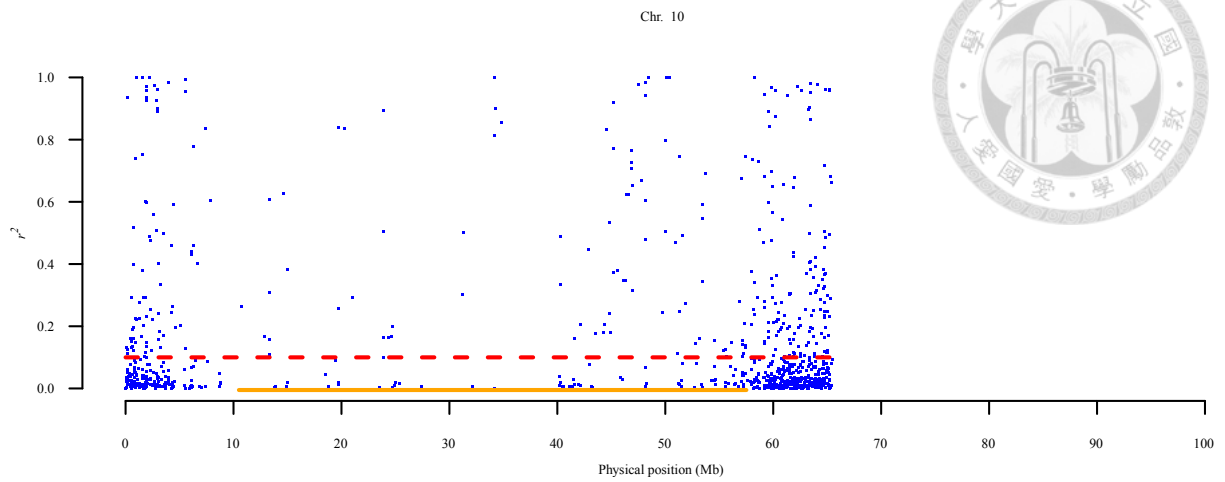
S\_Fig 2.9 (page 1/4)



S\_Fig 2.9 (page 2/4)



S\_Fig 2.9 (page 3/4)



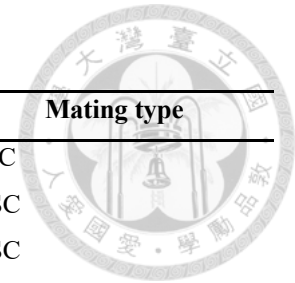
S\_Fig 2.9 The local LD of each chromosome.

The red dotted line was the baseline of  $r^2$  and the orange line indicated the heterochromatic region.



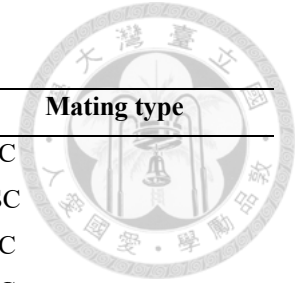
S\_Tab 2.1 The detailed information on each accession.

Accession	Reads	Missing proportion	Heterozygosity	Latitude	Longitude	Province/Department	Country	Mating type
LA0114	5,349,688	0.0313	0.0500	-7.4000	-79.5667	La Libertad	Peru	FSC
LA0373	7,025,393	0.0247	0.0509	-9.9400	-78.2300	Ancash	Peru	ASC
LA0391	5,003,706	0.0438	0.0471	-7.2442	-78.6817	Cajamarca	Peru	ASC
LA0397	3,814,682	0.0805	0.0647	-6.7500	-79.7167	Lambayeque	Peru	FSC
LA0400	11,180,073	0.0156	0.1039	-5.2608	-79.9642	Piura	Peru	FSC
LA0411	8,561,463	0.0197	0.4025	-1.1000	-79.4833	Los Rios	Ecuador	ASC
LA0417	3,975,053	0.1014	0.1043	-2.7333	-79.9167	Guayas	Ecuador	ASC
LA0442	10,903,757	0.0148	0.0494	-9.4817	-78.2592	Ancash	Peru	FSC
LA1236	7,911,580	0.0237	0.0649	-0.2500	-79.1500	Pichincha	Ecuador	ASC
LA1237	4,280,319	0.1362	0.0262	0.8667	-79.8500	Esmeraldas	Ecuador	ASC
LA1245	12,636,210	0.0163	0.2044	-3.4583	-79.9667	El Oro	Ecuador	ASC
LA1246	6,278,827	0.0386	0.0412	-3.9900	-79.3600	Loja	Ecuador	ASC
LA1256	6,005,267	0.0433	0.0397	-2.6667	-79.6167	Guayas	Ecuador	ASC
LA1261	6,385,191	0.0413	0.0444	-1.8167	-79.5167	Los Rios	Ecuador	ASC
LA1279	4,454,138	0.0611	0.0380	-12.1333	-76.8167	Lima	Peru	ASC
LA1280	6,781,216	0.0312	0.2068	-12.0333	-76.7167	Lima	Peru	ASC
LA1301	6,577,016	0.0303	0.0516	-13.7333	-75.9167	Ica	Peru	ASC
LA1335	5,784,776	0.0330	0.2336	-16.4000	-73.2500	Arequipa	Peru	ASC
LA1348	7,482,737	0.0273	0.0514	-7.4500	-79.5000	La Libertad	Peru	FSC
LA1349	8,285,652	0.0264	0.0543	-6.7436	-79.4997	Lambayeque	Peru	ASC
LA1371	7,521,506	0.0866	0.0465	-11.8894	-76.6539	Lima	Peru	ASC
LA1375	11,219,496	0.0189	0.1784	-13.0747	-76.4025	Lima	Peru	ASC



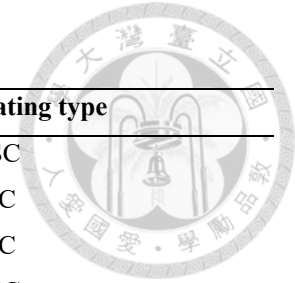
S\_Tab 2.1 (Continued)

Accession	Reads	Missing proportion	Heterozygosity	Latitude	Longitude	Province/Department	Country	Mating type
LA1380	3,781,668	0.0801	0.0698	-5.2525	-80.0506	Piura	Peru	FSC
LA1381	6,410,608	0.0241	0.0667	-5.5667	-79.9667	Lambayeque	Peru	ASC
LA1382	4,068,859	0.0714	0.0968	-6.8449	-78.0293	Amazonas	Peru	FSC
LA1466	4,210,180	0.0581	0.2282	-6.6333	-79.3833	Lambayeque	Peru	FSC
LA1469	4,448,552	0.0550	0.1639	-5.8600	-79.7900	Lambayeque	Peru	ASC
LA1471	5,445,081	0.0433	0.0782	-6.3167	-79.7500	Lambayeque	Peru	FSC
LA1478	4,860,773	0.0547	0.2351	-5.2167	-80.0833	Piura	Peru	FSC
LA1514	4,219,105	0.0634	0.0500	-11.0453	-77.1189	Lima	Peru	ASC
LA1521	10,311,260	0.0323	0.0470	-12.7647	-76.5053	Lima	Peru	ASC
LA1547	7,508,178	0.0200	0.0479	0.5833	-77.9333	Carchi	Ecuador	ASC
LA1576	6,922,139	0.0228	0.0573	-12.1667	-76.8667	Lima	Peru	ASC
LA1577	4,269,250	0.0492	0.0485	-7.8100	-79.1800	La Libertad	Peru	FSC
LA1578	3,874,621	0.0742	0.0522	-7.3333	-79.5833	La Libertad	Peru	FSC
LA1579	6,100,130	0.0275	0.1024	-6.5900	-79.8700	Lambayeque	Peru	FSC
LA1580	3,870,654	0.0630	0.1883	-6.5900	-79.8700	Lambayeque	Peru	FSC
LA1581	11,660,990	0.0163	0.1130	-6.6000	-79.8900	Lambayeque	Peru	FSC
LA1582	7,891,018	0.0258	0.0495	-6.1500	-79.7333	Lambayeque	Peru	FSC
LA1583	7,797,420	0.0182	0.1547	-6.2300	-79.7200	Lambayeque	Peru	FSC
LA1584	10,578,595	0.0162	0.1131	-6.3700	-79.7900	Lambayeque	Peru	FSC
LA1585	9,000,344	0.0186	0.0511	-6.6922	-79.4664	Lambayeque	Peru	FSC
LA1586	5,352,344	0.0334	0.0848	-8.3600	-78.7300	La Libertad	Peru	FSC
LA1587	5,799,426	0.0277	0.0992	-7.4333	-79.5167	La Libertad	Peru	FSC



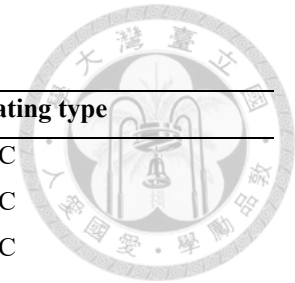
S\_Tab 2.1 (Continued)

Accession	Reads	Missing proportion	Heterozygosity	Latitude	Longitude	Province/Department	Country	Mating type
LA1589	9,790,673	0.0134	0.0468	-8.3900	-78.7400	La Libertad	Peru	ASC
LA1590	9,428,513	0.0164	0.1004	-8.3700	-78.7300	La Libertad	Peru	FSC
LA1591	8,445,761	0.0261	0.0508	-7.7167	-79.1167	La Libertad	Peru	FSC
LA1593	7,892,684	0.0197	0.0535	-8.5400	-78.6700	La Libertad	Peru	ASC
LA1595	3,896,054	0.0593	0.0701	-9.2700	-78.4700	Ancash	Peru	ASC
LA1596	3,909,064	0.0653	0.0544	-8.9250	-78.5667	Ancash	Peru	FSC
LA1599	4,024,711	0.0594	0.0486	-10.0583	-78.1833	Ancash	Peru	ASC
LA1601	4,457,135	0.0497	0.0867	-10.6700	-77.6800	Lima	Peru	ASC
LA1602	8,336,516	0.0234	0.0492	-11.7833	-76.9833	Lima	Peru	ASC
LA1606	10,023,328	0.0178	0.0529	-13.4667	-76.2000	Ica	Peru	FSC
LA1615	3,757,408	0.0720	0.0687	-5.2333	-80.6333	Piura	Peru	ASC
LA1617	7,144,146	0.0411	0.0363	-3.5667	-80.4667	Tumbes	Peru	FSC
LA1628	8,967,730	0.0422	0.0704	-7.1667	-79.5500	La Libertad	Peru	ASC
LA1629	3,880,195	0.0993	0.0426	-12.1167	-77.0333	Lima	Peru	ASC
LA1645	3,880,379	0.0963	0.0466	-12.1314	-77.0333	Lima	Peru	ASC
LA1659	7,656,966	0.0233	0.0533	-9.5467	-77.8586	Ancash	Peru	ASC
LA1670	4,000,691	0.0705	0.0475	-17.8333	-70.5167	Tacna	Peru	ASC
LA1683	9,277,274	0.0241	0.1313	-4.8700	-81.1100	Piura	Peru	FSC
LA1684	9,046,167	0.0191	0.1035	-5.1000	-80.1500	Piura	Peru	FSC
LA1685	3,951,431	0.1156	0.0530	-4.8867	-80.6975	Piura	Peru	FSC
LA1686	10,666,902	0.0177	0.1231	-5.0700	-80.6200	Piura	Peru	FSC
LA1687	15,271,160	0.0072	0.2843	-5.0700	-80.6200	Piura	Peru	FSC



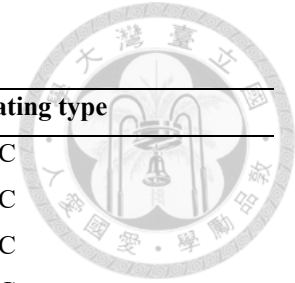
S\_Tab 2.1 (Continued)

Accession	Reads	Missing proportion	Heterozygosity	Latitude	Longitude	Province/Department	Country	Mating type
LA1688	7,693,836	0.0341	0.0435	-4.8833	-80.3750	Piura	Peru	FSC
LA1689	11,245,688	0.0280	0.0749	-5.1764	-80.6175	Piura	Peru	FSC
LA1690	11,506,994	0.0143	0.0545	-5.1764	-80.6175	Piura	Peru	FSC
LA1720	5,217,210	0.0476	0.0652	-9.5167	-78.0000	Ancash	Peru	ASC
LA1729	6,124,422	0.0362	0.0455	-13.2969	-75.6406	Ica	Peru	ASC
LA1921	4,154,202	0.0889	0.0439	-14.3119	-75.1272	Ica	Peru	ASC
LA1923	4,531,572	0.1001	0.0391	-14.6667	-75.2833	Ica	Peru	ASC
LA1924	4,373,364	0.0823	0.0442	-14.6289	-75.2142	Ica	Peru	ASC
LA1933	4,069,640	0.1389	0.0401	-15.4564	-74.4458	Arequipa	Peru	ASC
LA1936	4,271,265	0.1305	0.0385	-15.8336	-74.0325	Arequipa	Peru	ASC
LA2097	5,501,138	0.0719	0.1790	-4.3939	-79.9181	Loja	Ecuador	ASC
LA2102	4,149,108	0.0970	0.2470	-4.4017	-79.4675	Loja	Ecuador	ASC
LA2146	5,948,231	0.0490	0.0521	-7.3019	-79.4161	La Libertad	Peru	ASC
LA2149	7,302,190	0.0281	0.0467	-7.2181	-78.7878	Cajamarca	Peru	ASC
LA2173	6,722,975	0.0319	0.0538	-5.3307	-78.7905	Cajamarca	Peru	ASC
LA2181	5,770,182	0.0473	0.0932	-5.7758	-78.7831	Cajamarca	Peru	ASC
LA2183	4,367,799	0.0852	0.0413	-5.7400	-78.6700	Amazonas	Peru	ASC
LA2186	4,491,690	0.0850	0.0516	-5.8917	-78.1667	Amazonas	Peru	ASC
LA2389	3,835,029	0.1225	0.0603	-7.2500	-79.1333	Cajamarca	Peru	FSC
LA2390	4,768,009	0.0945	0.0517	-7.2333	-79.1417	Cajamarca	Peru	ASC
LA2401	6,288,238	0.0340	0.0617	-9.5083	-78.2278	Ancash	Peru	ASC
LA2533	4,043,132	0.0857	0.0419	-11.3000	-77.3600	Lima	Peru	ASC



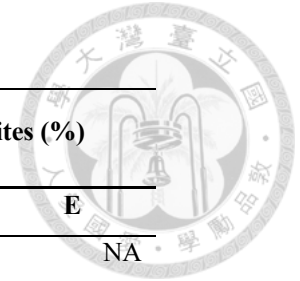
S\_Tab 2.1 (Continued)

Accession	Reads	Missing proportion	Heterozygosity	Latitude	Longitude	Province/Department	Country	Mating type
LA2645	9,438,437	0.0245	0.0964	-5.1667	-80.1833	Piura	Peru	FSC
LA2646	6,525,332	0.1592	0.1357	-5.0500	-79.8000	Piura	Peru	FSC
LA2647	615,835	0.6568	0.0416	-5.1750	-79.9833	Piura	Peru	FSC
LA2652	7,698,150	0.0281	0.0752	-4.9031	-80.6842	Piura	Peru	FSC
LA2653	4,012,161	0.0730	0.0453	-4.7500	-80.5833	Piura	Peru	FSC
LA2655	9,642,583	0.0409	0.0856	-4.9083	-80.8250	Piura	Peru	ASC
LA2656	7,178,813	0.0270	0.1070	-3.8000	-80.7000	Tumbes	Peru	FSC
LA2659	9,031,421	0.0253	0.0533	-5.2167	-80.6250	Piura	Peru	FSC
LA2852	6,145,708	0.0499	0.0308	-0.8333	-80.4833	Manabi	Ecuador	ASC
LA2915	5,019,946	0.0737	0.0521	-5.9847	-79.7453	Lambayeque	Peru	FSC
LA3638	4,059,256	0.1141	0.0410	-12.5667	-76.3167	Lima	Peru	ASC



S\_Tab 2.2 The statistical summaries of expected sites and sequenced sites of *PstI*, the sites targeted by SNP and the sequenced genes.

Chr.	Expected sites			Sequenced sites			Proportion of sequenced sites (%)		
	All <sup>a</sup>	H <sup>a</sup>	E <sup>a</sup>	All	H	E	All	H	E
0	2,276	NA	NA	124	NA	NA	5.45	NA	NA
1	9,745	4,680	5,065	3,197	312	2,885	32.81	6.67	56.96
2	5,746	1,826	3,920	2,599	244	2,355	45.23	13.36	60.08
3	7,391	2,670	4,721	2,522	231	2,291	34.12	8.65	48.53
4	6,525	3,716	2,809	2,032	399	1,633	31.14	10.74	58.13
5	6,561	4,343	2,218	1,488	363	1,125	22.68	8.36	50.72
6	5,380	1,943	3,437	2,149	273	1,876	39.94	14.05	54.58
7	6,779	3,853	2,926	1,698	241	1,457	25.05	6.25	49.79
8	6,585	3,875	2,710	1,779	344	1,435	27.02	8.88	52.95
9	7,054	3,818	3,236	1,697	256	1,441	24.06	6.71	44.53
10	6,408	4,188	2,220	1,541	395	1,146	24.05	9.43	51.62
11	5,911	3,377	2,534	1,672	361	1,311	28.29	10.69	51.74
12	6,453	3,389	3,064	1,490	402	1,088	23.09	11.86	35.51
<b>Total</b>	82,814	41,678	38,860	23,988	3,821	20,043	28.97	9.17	51.58

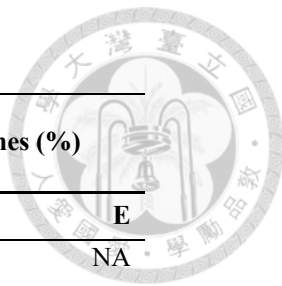


S\_Tab 2.2 (Continued)

Chr.	Sites containing SNP			Proportion of sites with SNP (%)		Expected genes in expected <i>Pst</i> I RADseq regions
	All	H	E	H	E	
0	76	NA	NA	NA	NA	216
1	2,178	198	1,980	63.46	68.63	1,914
2	1,737	127	1,610	52.05	68.37	1,501
3	1,665	113	1,552	48.92	67.74	1,526
4	1,354	246	1,108	61.65	67.85	1,152
5	999	231	768	63.64	68.27	889
6	1,435	170	1,265	62.27	67.43	1,290
7	1,182	160	1,022	66.39	70.14	1,020
8	1,195	208	987	60.47	68.78	1,077
9	1,121	122	999	47.66	69.33	1,003
10	1,062	247	815	62.53	71.12	952
11	1,142	217	925	60.11	70.56	941
12	1,045	276	769	68.66	70.68	936
<b>Total</b>	16,191	2,315	13,800	60.59	68.85	14,417



S\_Tab 2.2 (Continued)



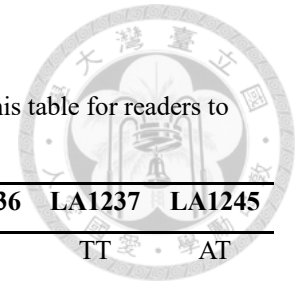
Chr.	Genes in sequenced regions			Genes with SNP			Proportion of sequenced genes (%)		
	All <sup>a</sup>	H <sup>a</sup>	E <sup>a</sup>	All	H	E	All	H	E
0	62	NA	NA	25	NA	NA	6.99	NA	NA
1	1,742	120	1,622	1,029	55	974	40.58	13.78	47.40
2	1,400	91	1,309	803	35	768	41.82	19.04	45.61
3	1,389	96	1,293	812	42	770	41.46	20.21	44.97
4	1,054	157	897	611	90	521	38.44	22.89	43.63
5	783	141	642	437	73	364	32.38	18.58	38.70
6	1,195	145	1,050	673	80	593	42.48	28.83	45.45
7	902	100	802	535	54	481	36.18	18.62	41.00
8	952	143	809	599	80	519	38.70	22.03	44.67
9	877	105	772	507	25	482	34.94	17.77	40.23
10	812	163	649	444	71	373	31.89	17.58	40.09
11	834	138	696	466	48	418	34.97	18.75	42.21
12	788	179	609	440	81	359	31.77	23.07	35.74
<b>Total</b>	12,790	1,578	11,150	7,381	734	6,622	36.83	19.75	43.13

<sup>a</sup>: All, H and E indicated each chromosome, the heterochromatin region and the euchromatin region.



S\_Tab 2.3 The information on 24,330 SNPs.

This supplementary material is a table of 24,330 rows (SNPs) x 98 columns (Accessions). We listed only the first 20 SNPs x 10 accessions of this table for readers to glimpse the data. The full table is published on <https://doi.org/10.1534/g3.118.200862>.



Marker	Major allele	Minor allele	LA0114	LA0373	LA0391	LA0397	LA0400	LA0417	LA0442	LA1236	LA1237	LA1245
SSL2.50ch00_1143661	T	A	TT	TT	TT	TT	TT	TT	TT	NN	TT	AT
SSL2.50ch00_3086004	C	T	NN	CC	CC	TT	TT	CC	CC	CC	NN	CC
SSL2.50ch00_3641105	A	T	AA	AA	AA	AA	AA	TT	AA	AA	TT	TT
SSL2.50ch00_4263006	C	T	NN	TT	CC	NN	CC	NN	NN	CC	NN	NN
SSL2.50ch00_4310217	G	T	GG	GG	TT	GG	GG	GG	GG	GG	GG	GG
SSL2.50ch00_4313972	T	C	CC	TT	CC	TT	CC	CC	TT	CC	CC	CC
SSL2.50ch00_4427214	C	G	CC	CC	CC	CC	GC	GC	CC	CC	CC	GC
SSL2.50ch00_4427220	C	A	CC	CC	CC	CC	CC	CC	CA	CC	CC	CC
SSL2.50ch00_4427223	C	T	CC	CC	TC	CC	TC	TC	TC	CC	TC	TC
SSL2.50ch00_4427226	T	C	CC	CC	TC	CC	TT	TT	TC	TT	TT	TT
SSL2.50ch00_4427229	A	G	AA	AA	GA	AA	GA	GA	GA	AA	GA	GA
SSL2.50ch00_4427230	G	A	GG	GG	GG	GG	GA	GG	GG	GG	GG	GA
SSL2.50ch00_4427233	C	T	TT	TT	TC	TT	CC	CC	TC	CC	CC	CC
SSL2.50ch00_4427239	T	G	TT	TT	TG	TT	TG	TG	TG	TT	TG	TG
SSL2.50ch00_4427250	G	A	AA	AA	GA	AA	GG	GG	GA	GG	GG	GG
SSL2.50ch00_4427255	C	T	CC	CC	TC	CC	TC	TC	TC	CC	TC	TC
SSL2.50ch00_4427265	C	T	CC	CC	CT	CC	CC	NN	CT	CC	CC	NN
SSL2.50ch00_6550092	A	G	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA
SSL2.50ch00_6556529	T	G	TT	TT	NN	GG	TT	TT	TT	TT	TT	TT
SSL2.50ch00_6556632	T	C	TT	CC	TT	TT	TT	TT	TT	TT	NN	TT

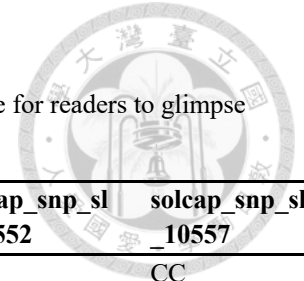
S\_Tab 2.4 Pairwise  $F_{st}$  of subpopulations.

Group Name	POP S1	POP S2	POP S3	POP M1	POP M2	POP M3
POP S2	0.0521***					
POP S3	0.0638***	0.0198***				
POP M1	0.0266***	0.0075***	0.0229***			
POP M2	0.0552***	0.0046***	0.0109***	0.0112***		
POP M3	0.0127***	0.0262***	0.0259***	0.0118***	0.0234***	
POP M4	0.0252***	0.0021***	0.0191***	-0.004	0.0020***	-0.0023



S\_Tab 2.5 The locations and genotypes of 214 samples of SolCAP genotyping array.

This supplementary material is a table of 214 rows (samples) x 2,312 columns (SNPs). We listed only the first 10 samples x 5 SNPs of this table for readers to glimpse the data. The full table is published on <https://doi.org/10.1534/g3.118.200862>.



ID <sup>a</sup>	Sample <sup>b</sup>	Accession <sup>c</sup>	Latitude	Longitude	solcap_snp_sl _10194	solcap_snp_sl _10195	solcap_snp_sl _10247	solcap_snp_sl _10552	solcap_snp_sl _10557
BGV006129_2012B	BGV006129	BGV006129	-3.9519	-79.4356	CC	TT	AA	GG	CC
BGV006129_2015B	BGV006129	BGV006129	-3.9519	-79.4356	CC	TT	AA	GG	CC
BGV006187_2012B	BGV006187	BGV006187	-3.3122	-79.6286	CC	TT	GG	TT	CC
BGV006187_2015B	BGV006187	BGV006187	-3.3122	-79.6286	CC	TT	GG	TT	CC
BGV006327_2012B	BGV006327	BGV006327	-4.8922	-80.3753	CC	TT	AA	GG	AA
BGV006327_2015B	BGV006327	BGV006327	-4.8922	-80.3753	CC	TT	AA	GG	AA
BGV006328_2012B	BGV006328	BGV006328	-4.9339	-80.5394	CT	TT	AA	GT	AC
BGV006328_2015B	BGV006328	BGV006328	-4.9339	-80.5394	CT	TT	AA	GT	AC
BGV006331_2012B	BGV006331	BGV006331	-4.9339	-80.5394	CC	TT	AA	GG	CC
BGV006331_2015B	BGV006331	BGV006331	-5.2872	-79.9581	CC	TT	AA	GG	CC

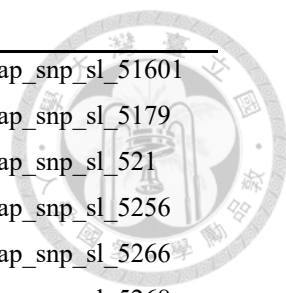
a: ID indicates the index, its original sample ID plus its original study, in this SolCAP meta-analysis.

b: Sample indicates the original sample ID in the original study.

c: Accession indicates the accession of each sample.

S\_Tab 2.6 The removed 627 SNPs with reverse-complement allele designation.

solcap_snp_sl_10196	solcap_snp_sl_20361	solcap_snp_sl_25267	solcap_snp_sl_3853
solcap_snp_sl_10236	solcap_snp_sl_20409	solcap_snp_sl_25270	solcap_snp_sl_38945
solcap_snp_sl_10246	solcap_snp_sl_20499	solcap_snp_sl_25277	solcap_snp_sl_38987
solcap_snp_sl_10377	solcap_snp_sl_20500	solcap_snp_sl_25278	solcap_snp_sl_3924
solcap_snp_sl_10516	solcap_snp_sl_20585	solcap_snp_sl_25283	solcap_snp_sl_39725
solcap_snp_sl_10563	solcap_snp_sl_20719	solcap_snp_sl_25296	solcap_snp_sl_3980
solcap_snp_sl_10569	solcap_snp_sl_20723	solcap_snp_sl_25297	solcap_snp_sl_39868
solcap_snp_sl_10596	solcap_snp_sl_20752	solcap_snp_sl_25304	solcap_snp_sl_39959
solcap_snp_sl_10686	solcap_snp_sl_20809	solcap_snp_sl_25305	solcap_snp_sl_3997
solcap_snp_sl_10796	solcap_snp_sl_20883	solcap_snp_sl_25313	solcap_snp_sl_4016
solcap_snp_sl_10904	solcap_snp_sl_20932	solcap_snp_sl_25322	solcap_snp_sl_4024
solcap_snp_sl_10928	solcap_snp_sl_20936	solcap_snp_sl_25336	solcap_snp_sl_4029
solcap_snp_sl_10946	solcap_snp_sl_20952	solcap_snp_sl_25362	solcap_snp_sl_4034
solcap_snp_sl_10961	solcap_snp_sl_20958	solcap_snp_sl_25414	solcap_snp_sl_4055
solcap_snp_sl_11221	solcap_snp_sl_20981	solcap_snp_sl_25429	solcap_snp_sl_4099
solcap_snp_sl_11232	solcap_snp_sl_20988	solcap_snp_sl_25485	solcap_snp_sl_4121
solcap_snp_sl_11509	solcap_snp_sl_21014	solcap_snp_sl_2565	solcap_snp_sl_42919
solcap_snp_sl_11532	solcap_snp_sl_21039	solcap_snp_sl_25696	solcap_snp_sl_42933
solcap_snp_sl_11539	solcap_snp_sl_21070	solcap_snp_sl_25735	solcap_snp_sl_42942
solcap_snp_sl_11569	solcap_snp_sl_21102	solcap_snp_sl_25745	solcap_snp_sl_42961
solcap_snp_sl_11670	solcap_snp_sl_21280	solcap_snp_sl_258	solcap_snp_sl_43
solcap_snp_sl_11736	solcap_snp_sl_21317	solcap_snp_sl_25879	solcap_snp_sl_43894
solcap_snp_sl_11751	solcap_snp_sl_21323	solcap_snp_sl_25918	solcap_snp_sl_43920
solcap_snp_sl_11805	solcap_snp_sl_21363	solcap_snp_sl_25951	solcap_snp_sl_44932
solcap_snp_sl_11982	solcap_snp_sl_21390	solcap_snp_sl_2604	solcap_snp_sl_4518
solcap_snp_sl_12101	solcap_snp_sl_21400	solcap_snp_sl_26129	solcap_snp_sl_47660
solcap_snp_sl_12135	solcap_snp_sl_21401	solcap_snp_sl_2614	solcap_snp_sl_48910
solcap_snp_sl_12261	solcap_snp_sl_21429	solcap_snp_sl_26438	solcap_snp_sl_48911
solcap_snp_sl_12268	solcap_snp_sl_21430	solcap_snp_sl_26551	solcap_snp_sl_4926
solcap_snp_sl_12289	solcap_snp_sl_21456	solcap_snp_sl_26780	solcap_snp_sl_4932
solcap_snp_sl_12372	solcap_snp_sl_21677	solcap_snp_sl_26791	solcap_snp_sl_49752
solcap_snp_sl_12414	solcap_snp_sl_21714	solcap_snp_sl_2686	solcap_snp_sl_5050
solcap_snp_sl_12501	solcap_snp_sl_2172	solcap_snp_sl_2691	solcap_snp_sl_5051
solcap_snp_sl_12664	solcap_snp_sl_21966	solcap_snp_sl_2695	solcap_snp_sl_50871
solcap_snp_sl_12718	solcap_snp_sl_21971	solcap_snp_sl_2701	solcap_snp_sl_5094
solcap_snp_sl_12769	solcap_snp_sl_22017	solcap_snp_sl_27107	solcap_snp_sl_5095
solcap_snp_sl_12841	solcap_snp_sl_221	solcap_snp_sl_27162	solcap_snp_sl_5103
solcap_snp_sl_12878	solcap_snp_sl_22130	solcap_snp_sl_27482	solcap_snp_sl_5113
solcap_snp_sl_12913	solcap_snp_sl_222	solcap_snp_sl_2797	solcap_snp_sl_5115



---

solcap_snp_sl_1295	solcap_snp_sl_22259	solcap_snp_sl_282	solcap_snp_sl_51601
solcap_snp_sl_13098	solcap_snp_sl_22594	solcap_snp_sl_28295	solcap_snp_sl_5179
solcap_snp_sl_13147	solcap_snp_sl_22604	solcap_snp_sl_28404	solcap_snp_sl_521
solcap_snp_sl_13193	solcap_snp_sl_22831	solcap_snp_sl_28407	solcap_snp_sl_5256
solcap_snp_sl_13200	solcap_snp_sl_22839	solcap_snp_sl_28409	solcap_snp_sl_5266
solcap_snp_sl_1325	solcap_snp_sl_22845	solcap_snp_sl_28425	solcap_snp_sl_5268
solcap_snp_sl_13398	solcap_snp_sl_22846	solcap_snp_sl_2879	solcap_snp_sl_52783
solcap_snp_sl_1345	solcap_snp_sl_22858	solcap_snp_sl_28826	solcap_snp_sl_5280
solcap_snp_sl_13455	solcap_snp_sl_22869	solcap_snp_sl_28914	solcap_snp_sl_53
solcap_snp_sl_13464	solcap_snp_sl_22877	solcap_snp_sl_29043	solcap_snp_sl_53173
solcap_snp_sl_13590	solcap_snp_sl_22878	solcap_snp_sl_29326	solcap_snp_sl_535
solcap_snp_sl_13594	solcap_snp_sl_22880	solcap_snp_sl_29332	solcap_snp_sl_53552
solcap_snp_sl_13604	solcap_snp_sl_22882	solcap_snp_sl_29351	solcap_snp_sl_53870
solcap_snp_sl_13621	solcap_snp_sl_22889	solcap_snp_sl_29357	solcap_snp_sl_53877
solcap_snp_sl_13842	solcap_snp_sl_22891	solcap_snp_sl_29388	solcap_snp_sl_54547
solcap_snp_sl_13958	solcap_snp_sl_22892	solcap_snp_sl_2939	solcap_snp_sl_55020
solcap_snp_sl_14155	solcap_snp_sl_22894	solcap_snp_sl_29394	solcap_snp_sl_55037
solcap_snp_sl_14354	solcap_snp_sl_22897	solcap_snp_sl_29398	solcap_snp_sl_55409
solcap_snp_sl_14415	solcap_snp_sl_22898	solcap_snp_sl_29506	solcap_snp_sl_5547
solcap_snp_sl_14428	solcap_snp_sl_229	solcap_snp_sl_29549	solcap_snp_sl_55475
solcap_snp_sl_14672	solcap_snp_sl_22906	solcap_snp_sl_29565	solcap_snp_sl_55514
solcap_snp_sl_14759	solcap_snp_sl_22911	solcap_snp_sl_2959	solcap_snp_sl_55837
solcap_snp_sl_14845	solcap_snp_sl_22916	solcap_snp_sl_2971	solcap_snp_sl_55906
solcap_snp_sl_14865	solcap_snp_sl_22917	solcap_snp_sl_2974	solcap_snp_sl_5791
solcap_snp_sl_14874	solcap_snp_sl_22924	solcap_snp_sl_298	solcap_snp_sl_5795
solcap_snp_sl_1499	solcap_snp_sl_22956	solcap_snp_sl_2984	solcap_snp_sl_5800
solcap_snp_sl_15039	solcap_snp_sl_22957	solcap_snp_sl_2990	solcap_snp_sl_5807
solcap_snp_sl_15173	solcap_snp_sl_22959	solcap_snp_sl_29911	solcap_snp_sl_58447
solcap_snp_sl_1519	solcap_snp_sl_22963	solcap_snp_sl_29920	solcap_snp_sl_5875
solcap_snp_sl_1527	solcap_snp_sl_22973	solcap_snp_sl_29932	solcap_snp_sl_58920
solcap_snp_sl_15289	solcap_snp_sl_22975	solcap_snp_sl_29934	solcap_snp_sl_59437
solcap_snp_sl_15417	solcap_snp_sl_22979	solcap_snp_sl_30046	solcap_snp_sl_5973
solcap_snp_sl_15446	solcap_snp_sl_22986	solcap_snp_sl_3008	solcap_snp_sl_6003
solcap_snp_sl_15515	solcap_snp_sl_22988	solcap_snp_sl_301	solcap_snp_sl_6022
solcap_snp_sl_15641	solcap_snp_sl_22994	solcap_snp_sl_30133	solcap_snp_sl_60360
solcap_snp_sl_15690	solcap_snp_sl_22996	solcap_snp_sl_3035	solcap_snp_sl_6038
solcap_snp_sl_15728	solcap_snp_sl_23004	solcap_snp_sl_30380	solcap_snp_sl_6051
solcap_snp_sl_15757	solcap_snp_sl_23010	solcap_snp_sl_30408	solcap_snp_sl_60513
solcap_snp_sl_15879	solcap_snp_sl_23011	solcap_snp_sl_306	solcap_snp_sl_6073
solcap_snp_sl_15885	solcap_snp_sl_23014	solcap_snp_sl_30819	solcap_snp_sl_60831

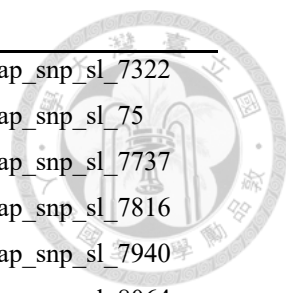
---



---

solcap_snp_sl_16096	solcap_snp_sl_23015	solcap_snp_sl_30911	solcap_snp_sl_6086
solcap_snp_sl_16099	solcap_snp_sl_23020	solcap_snp_sl_3094	solcap_snp_sl_6092
solcap_snp_sl_16133	solcap_snp_sl_23021	solcap_snp_sl_31119	solcap_snp_sl_6112
solcap_snp_sl_16141	solcap_snp_sl_23028	solcap_snp_sl_3112	solcap_snp_sl_61192
solcap_snp_sl_16162	solcap_snp_sl_23044	solcap_snp_sl_31275	solcap_snp_sl_6152
solcap_snp_sl_16196	solcap_snp_sl_23045	solcap_snp_sl_31277	solcap_snp_sl_6186
solcap_snp_sl_16421	solcap_snp_sl_23051	solcap_snp_sl_31280	solcap_snp_sl_6226
solcap_snp_sl_16424	solcap_snp_sl_23055	solcap_snp_sl_3130	solcap_snp_sl_62495
solcap_snp_sl_16499	solcap_snp_sl_23059	solcap_snp_sl_3159	solcap_snp_sl_6255
solcap_snp_sl_16501	solcap_snp_sl_23061	solcap_snp_sl_31671	solcap_snp_sl_62616
solcap_snp_sl_16576	solcap_snp_sl_23062	solcap_snp_sl_31687	solcap_snp_sl_62666
solcap_snp_sl_16579	solcap_snp_sl_23064	solcap_snp_sl_31723	solcap_snp_sl_62695
solcap_snp_sl_16584	solcap_snp_sl_23068	solcap_snp_sl_31730	solcap_snp_sl_6370
solcap_snp_sl_16642	solcap_snp_sl_23088	solcap_snp_sl_31775	solcap_snp_sl_63704
solcap_snp_sl_16650	solcap_snp_sl_23096	solcap_snp_sl_31777	solcap_snp_sl_6372
solcap_snp_sl_16840	solcap_snp_sl_23099	solcap_snp_sl_31884	solcap_snp_sl_64263
solcap_snp_sl_16920	solcap_snp_sl_23145	solcap_snp_sl_31953	solcap_snp_sl_64662
solcap_snp_sl_1701	solcap_snp_sl_23192	solcap_snp_sl_31971	solcap_snp_sl_6524
solcap_snp_sl_17063	solcap_snp_sl_23195	solcap_snp_sl_31973	solcap_snp_sl_65244
solcap_snp_sl_17239	solcap_snp_sl_23344	solcap_snp_sl_31978	solcap_snp_sl_6526
solcap_snp_sl_17289	solcap_snp_sl_234	solcap_snp_sl_32032	solcap_snp_sl_65262
solcap_snp_sl_17448	solcap_snp_sl_23453	solcap_snp_sl_32093	solcap_snp_sl_6568
solcap_snp_sl_17476	solcap_snp_sl_23561	solcap_snp_sl_32147	solcap_snp_sl_65880
solcap_snp_sl_17496	solcap_snp_sl_23591	solcap_snp_sl_32389	solcap_snp_sl_66569
solcap_snp_sl_17507	solcap_snp_sl_23608	solcap_snp_sl_32425	solcap_snp_sl_67010
solcap_snp_sl_17524	solcap_snp_sl_23702	solcap_snp_sl_32529	solcap_snp_sl_67119
solcap_snp_sl_17536	solcap_snp_sl_23734	solcap_snp_sl_32703	solcap_snp_sl_67772
solcap_snp_sl_17544	solcap_snp_sl_23763	solcap_snp_sl_330	solcap_snp_sl_67805
solcap_snp_sl_17563	solcap_snp_sl_23787	solcap_snp_sl_33136	solcap_snp_sl_6902
solcap_snp_sl_17581	solcap_snp_sl_23811	solcap_snp_sl_33139	solcap_snp_sl_69255
solcap_snp_sl_17643	solcap_snp_sl_23823	solcap_snp_sl_33547	solcap_snp_sl_69262
solcap_snp_sl_17645	solcap_snp_sl_23882	solcap_snp_sl_3355	solcap_snp_sl_69276
solcap_snp_sl_17649	solcap_snp_sl_23975	solcap_snp_sl_33642	solcap_snp_sl_6934
solcap_snp_sl_17717	solcap_snp_sl_24001	solcap_snp_sl_33736	solcap_snp_sl_69429
solcap_snp_sl_1772	solcap_snp_sl_24081	solcap_snp_sl_33817	solcap_snp_sl_7042
solcap_snp_sl_17751	solcap_snp_sl_24251	solcap_snp_sl_33822	solcap_snp_sl_7045
solcap_snp_sl_17839	solcap_snp_sl_24255	solcap_snp_sl_33830	solcap_snp_sl_7046
solcap_snp_sl_18055	solcap_snp_sl_2438	solcap_snp_sl_34143	solcap_snp_sl_70737
solcap_snp_sl_18057	solcap_snp_sl_24383	solcap_snp_sl_34165	solcap_snp_sl_70781
solcap_snp_sl_1815	solcap_snp_sl_24384	solcap_snp_sl_34177	solcap_snp_sl_7123

---



---

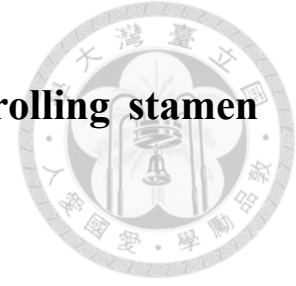
solcap_snp_sl_18185	solcap_snp_sl_24445	solcap_snp_sl_34186	solcap_snp_sl_7322
solcap_snp_sl_1819	solcap_snp_sl_24560	solcap_snp_sl_34221	solcap_snp_sl_75
solcap_snp_sl_18196	solcap_snp_sl_24562	solcap_snp_sl_3424	solcap_snp_sl_7737
solcap_snp_sl_1824	solcap_snp_sl_24604	solcap_snp_sl_34253	solcap_snp_sl_7816
solcap_snp_sl_18256	solcap_snp_sl_24609	solcap_snp_sl_34373	solcap_snp_sl_7940
solcap_snp_sl_1827	solcap_snp_sl_24755	solcap_snp_sl_34684	solcap_snp_sl_8064
solcap_snp_sl_18272	solcap_snp_sl_24787	solcap_snp_sl_34742	solcap_snp_sl_8120
solcap_snp_sl_18306	solcap_snp_sl_24973	solcap_snp_sl_34762	solcap_snp_sl_8121
solcap_snp_sl_18313	solcap_snp_sl_24987	solcap_snp_sl_3480	solcap_snp_sl_8464
solcap_snp_sl_18398	solcap_snp_sl_24990	solcap_snp_sl_35063	solcap_snp_sl_8514
solcap_snp_sl_18634	solcap_snp_sl_25015	solcap_snp_sl_35139	solcap_snp_sl_8526
solcap_snp_sl_18755	solcap_snp_sl_25082	solcap_snp_sl_35382	solcap_snp_sl_8659
solcap_snp_sl_18756	solcap_snp_sl_25150	solcap_snp_sl_35693	solcap_snp_sl_8697
solcap_snp_sl_18757	solcap_snp_sl_25167	solcap_snp_sl_357	solcap_snp_sl_8795
solcap_snp_sl_18943	solcap_snp_sl_25168	solcap_snp_sl_35757	solcap_snp_sl_8813
solcap_snp_sl_18944	solcap_snp_sl_25171	solcap_snp_sl_35777	solcap_snp_sl_9125
solcap_snp_sl_18949	solcap_snp_sl_2518	solcap_snp_sl_35779	solcap_snp_sl_9136
solcap_snp_sl_18995	solcap_snp_sl_25187	solcap_snp_sl_360	solcap_snp_sl_9235
solcap_snp_sl_19032	solcap_snp_sl_25188	solcap_snp_sl_36050	solcap_snp_sl_9260
solcap_snp_sl_19513	solcap_snp_sl_25195	solcap_snp_sl_36135	solcap_snp_sl_9292
solcap_snp_sl_19569	solcap_snp_sl_25201	solcap_snp_sl_36141	solcap_snp_sl_9447
solcap_snp_sl_19636	solcap_snp_sl_25207	solcap_snp_sl_36157	solcap_snp_sl_9512
solcap_snp_sl_19643	solcap_snp_sl_25208	solcap_snp_sl_36165	solcap_snp_sl_9513
solcap_snp_sl_19652	solcap_snp_sl_25210	solcap_snp_sl_36203	solcap_snp_sl_9531
solcap_snp_sl_19657	solcap_snp_sl_25211	solcap_snp_sl_36224	solcap_snp_sl_9533
solcap_snp_sl_19660	solcap_snp_sl_25213	solcap_snp_sl_36548	solcap_snp_sl_9536
solcap_snp_sl_19759	solcap_snp_sl_25220	solcap_snp_sl_36568	solcap_snp_sl_9546
solcap_snp_sl_19782	solcap_snp_sl_25232	solcap_snp_sl_36725	solcap_snp_sl_9550
solcap_snp_sl_19899	solcap_snp_sl_25236	solcap_snp_sl_37054	solcap_snp_sl_9558
solcap_snp_sl_19981	solcap_snp_sl_25242	solcap_snp_sl_37057	solcap_snp_sl_9560
solcap_snp_sl_20051	solcap_snp_sl_2525	solcap_snp_sl_37198	solcap_snp_sl_9690
solcap_snp_sl_20064	solcap_snp_sl_25251	solcap_snp_sl_3723	solcap_snp_sl_9751
solcap_snp_sl_20088	solcap_snp_sl_25255	solcap_snp_sl_37399	solcap_snp_sl_9752
solcap_snp_sl_2011	solcap_snp_sl_25256	solcap_snp_sl_37400	solcap_snp_sl_9798
solcap_snp_sl_20228	solcap_snp_sl_25258	solcap_snp_sl_3746	solcap_snp_sl_9814
solcap_snp_sl_20229	solcap_snp_sl_25260	solcap_snp_sl_37808	solcap_snp_sl_9816
solcap_snp_sl_20241	solcap_snp_sl_25261	solcap_snp_sl_38	solcap_snp_sl_9832
solcap_snp_sl_20256	solcap_snp_sl_25262	solcap_snp_sl_3849	

---

S\_Tab 2.7 The identity Of 2,307 SNP markers within accessions.

Accession	Individuals	Identity (%)	Accession	Individuals	Identity (%)
BGV006129	2	100	BGV007155	2	100
BGV006187	2	100	BGV007161	2	100
BGV006327	2	100	BGV007168	2	100
BGV006328	2	100	BGV007208	2	100
BGV006331	2	100	BGV007222	2	100
BGV006333	2	100	BGV007225	2	100
BGV006336	2	100	BGV007348	2	100
BGV006341	2	100	BGV007355	2	100
BGV006343	2	100	BGV007366	2	100
BGV006344	2	100	BGV007946	2	100
BGV006345	2	100	BGV007947	2	100
BGV006347	2	100	BGV015381	2	100
BGV006360	2	100	BGV015382	2	100
BGV006369	2	100	LA0373	2	76
BGV006370	2	100	LA0400	2	91
BGV006452	2	100	LA0722	2	100
BGV006457	2	100	LA121_1	2	100
BGV006468	2	100	LA1269	3	95
BGV006476	2	100	LA1301	3	99
BGV006478	2	100	LA1371	2	98
BGV006484	2	100	LA1429	3	97
BGV006492	2	100	LA1478	2	71
BGV006504	2	100	LA1547	4	56
BGV006507	2	100	LA1578	2	100
BGV006514	2	100	LA1582	3	81
BGV006639	2	100	LA1589	3	98
BGV006640	2	100	LA1617	2	100
BGV006642	2	100	LA1689	3	90
BGV006690	2	100	LA1923	2	100
BGV006691	2	100	LA1936	2	100
BGV006712	2	100	LA2093	2	100
BGV007095	2	100	LA2181	3	66
BGV007100	2	100	LA2184	2	100
BGV007104	2	100	LA2188	2	100
BGV007109	2	100	LA2533	3	99
BGV007111	2	100	LA2725	2	100
BGV007137	2	100	LA2854	2	100
BGV007145	2	100	PI 128216	2	100
BGV007151	2	100	PI 365914	2	100





## **Chapter 3 GWAS of the candidate genes controlling stamen length in *Solanum pimpinellifolium***

### **3.1 Purpose**

In this chapter, we performed a GWAS with the 98 *S. pimpinellifolium* accessions to identify the candidate genes controlling stamen length. Following the results in Chapter 2, we conducted three models in TASSEL using the set of genome-wide high-density SNPs from RADseq. The first model is the general linear model (GLM) with the correction of ADMIXTURE structure; the second one is the mixed linear model (MLM) with a matrix of kinship as a random effect; the third one is the MLM with the correction of both ADMIXTURE structure and kinship. In addition, Genome-wide Efficient Mixed Model Association (GEMMA) was also used to run MLM.

### **3.2 Material and Method**

#### **3.2.1 Plant material and phenotyping**

The collection of 98 *S. pimpinellifolium* accessions from TGRC was propagated by the way of single seed descent for two generations. Four individuals per accession were planted in the field by conventional agriculture practice from 2013 to 2014 in the farm of National Taiwan University, Taipei. Five flowers per plant were gathered in 2013 November, 2014 January and 2014 April. The stamen was scanned and measured by ImageJ software (Schneider, Rasband, & Eliceiri, 2012).



### 3.2.2 GWAS

GWAS was performed with the non-redundant dataset of 19,993 SNPs as we described in Chapter 2. Three models,  $P = G + Q + E$ ,  $P = G + K + E$  and  $P = G + Q + K + E$ , were completed in TASSEL (Bradbury et al., 2007). P, G, Q, K and E indicated phenotype, genotype, Q matrix of ADMIXTURE, kinship and error, respectively. The kinship was the only random effect. The  $P = G + K + E$  model was also performed with GEMMA following the manual (Zhou & Stephens, 2012). Q-Q plots and manhattan plots were presented by the R package qqman (Turner, 2014). For the GLM, the significant locus is determined by the permutation p value less than 0.05. For the MLM in TASSEL, the adjusted p value (FDR) was not provided; therefore R function p.adjust ('BH') was served as an alternative. A significant locus is determined if its p-value is less than 0.001. A candidate locus is defined if its false discovery rate (FDR) is less than 0.05 (Storey & Tibshirani, 2003).

### 3.2.3 Haplotype block

We built phased haplotypes with the 24,330 SNPs via BEAGLE and then estimated haplotype blocks via plink (Browning & Browning, 2007; Gaunt et al., 2007). The haplotype block was estimated by SNPs within 100-Mb interval considering there is a large proportion of heterochromatin in each chromosome. All the haplotypes are summarized without the data of chromosome 0. The LD heatmap was plotted by the R package LDheatmap (Shin, Blay, Graham, & McNeney, 2015).

### 3.3 Result

#### 3.3.1 SSL2.50ch06\_45620556 is significant among all the GLM and MLM analysis



The phenotype was the mean of stamen length among three measurements (Figure 3.1; S\_Tab 3.1). The accessions with long stamen were clustered in the north of Peru, which was the same as the previous finding (Figure 3.2) (Rick et al., 1977). We first observed the Q-Q plots, and the  $P = G + K + E$  model showed the least deviation from the expectation among the three models (S\_Fig 3.1); therefore, we continued a series of analyses following the result of  $P = G + K + E$ . In TASSEL, five significant loci were detected but with high FDR of 0.8382 (Table 3.1; S\_Fig 3.2); the heritability was 0.54. In GEMMA, 22 significant loci were detected also with high FDR; the heritability was 0.65 (Table 3.1; S\_Fig 3.2). SSL2.50ch01\_18302427, SSL2.50ch03\_70083752 and SSL2.50ch06\_45620556, were detected both in TASSEL and in GEMMA. Because high FDR suggested the high possibility of false association between these significant loci and the stamen length, more evidence is necessary to confirm these loci. In addition, we also listed the two significant loci based on the model of  $P = G + Q + E$ . The correction for only population structure was more reasonable because little kinship would occur in our sample (an individual standing for an accession) (Table 3.2; S\_Fig 3.2). As a result, SSL2.50ch06\_45620556 and SSL2.50ch12\_301545 were significant. SSL2.50ch06\_45620556 showed a conserved significance in  $P = G + K + E$  and  $P = G + Q + E$  models.

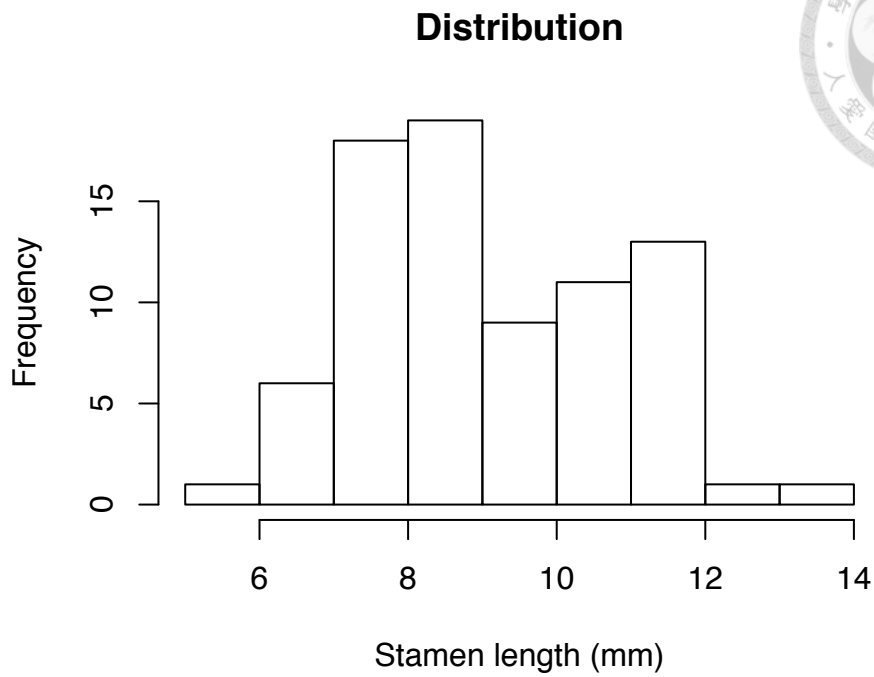


Figure 3.1 The distribution of stamen length.

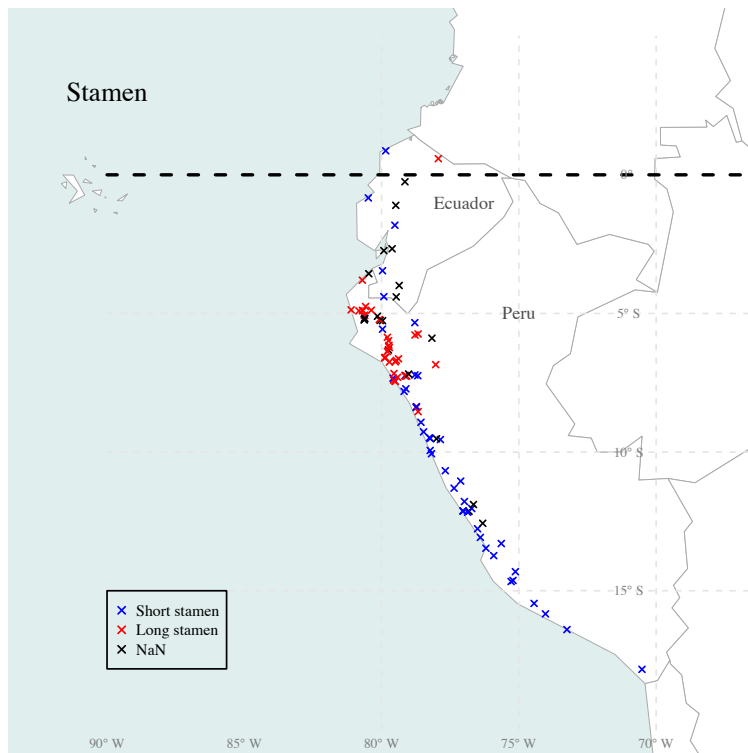


Figure 3.2 The geographic distribution of the stamen characters among 98 accessions. The black dotted line is the equator.

Table 3.1 Significant loci for stamen length in TASSEL and GEMMA.

Significant locus	Allele <sup>a</sup>	p value	FDR	R <sup>2</sup> <sup>b</sup>
<b>TASSEL</b>				
SSL2.50ch01_18302427 <sup>c</sup>	A/C	0.0003	0.8382	0.1998
SSL2.50ch03_70083752 <sup>c</sup>	T/C	0.0008	0.8382	0.2729
SSL2.50ch06_45620556 <sup>c</sup>	A/G	0.0003	0.8382	0.2394
SSL2.50ch08_2088583	C/T	0.0003	0.8382	0.2407
SSL2.50ch08_61447940	A/T	0.0009	0.8382	0.2000
<b>GEMMA</b>				
SSL2.50ch01_18302427 <sup>c</sup>	A/C	0.0002	0.5795	-
SSL2.50ch01_21314184	G/A	0.0002	0.5795	-
SSL2.50ch01_87989387	G/A	0.0009	0.7214	-
SSL2.50ch01_88015076	G/A	0.0003	0.6075	-
SSL2.50ch02_45035168	G/C	0.0002	0.5795	-
SSL2.50ch02_52704387	C/T	0.0002	0.5795	-
SSL2.50ch03_56810075	G/C	0.0005	0.7214	-
SSL2.50ch03_56828245	G/T	0.0002	0.5795	-
SSL2.50ch03_68538664	C/T	0.0008	0.7214	-
SSL2.50ch03_70083752 <sup>c</sup>	T/C	0.0005	0.7214	-
SSL2.50ch04_5128555	T/A	0.0008	0.7214	-
SSL2.50ch04_63808596	G/A	0.0009	0.7214	-
SSL2.50ch05_8922097	A/G	0.0003	0.6075	-
SSL2.50ch05_8922110	G/A	0.0002	0.5795	-
SSL2.50ch06_45620556 <sup>c</sup>	A/G	0.0002	0.5795	-
SSL2.50ch07_7730187	T/A	0.0001	0.7214	-
SSL2.50ch07_58342891	A/G	0.0005	0.5795	-
SSL2.50ch09_48896331	C/T	0.0010	0.7214	-
SSL2.50ch09_70184736	C/T	0.0005	0.7214	-
SSL2.50ch09_72338057	A/G	0.0008	0.7214	-
SSL2.50ch10_516364	C/G	0.0008	0.7214	-
SSL2.50ch11_54726695	G/A	0.0008	0.7214	-

<sup>a</sup>: Minor/ Major allele

<sup>b</sup>: Individual R<sup>2</sup> was not available in GEMMA.

<sup>c</sup>: The SNP was detected in both TASSEL and GEMMA.

Table 3.2 The two significant loci based on P = G + Q + E model.

Significant locus	Allele <sup>a</sup>	p value	Permutation p value	R <sup>2</sup>
SSL2.50ch06_45620556	A/C	3.3087*10 <sup>-7</sup>	0.0146	0.2099
SSL2.50ch12_301545	T/C	2.9833*10 <sup>-7</sup>	0.0133	0.2084

<sup>a</sup>: Minor/ Major allele

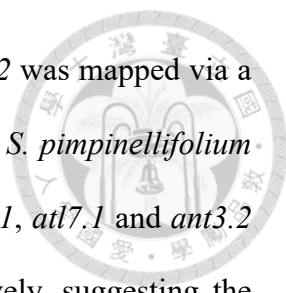
### 3.3.2 The LD patterns of these significant loci

We screened the LD patterns of these significant loci to confirm if the candidate loci were supported by their flanking significant SNPs. The LD blocks were defined in two methods; one was to utilize the significant SNPs as starting points and extended the LD decay of each chromosome to their upstream and downstream; this method resulted in decades-Kb LD blocks that contained 0 to 19 flanking SNPs (S\_Fig 3.3). The other was the haplotype blocks estimated by plink; this method revealed only six haplotype blocks and one of them extended to about 215 Kb (S\_Fig 3.3). However, none of the significant loci was supported by their flanking SNPs in these LD blocks, except for SSL2.50ch03\_56810075 and SSL2.50ch03\_56828245, which were supported by each other in haplotype block 2 (Table 3.1; S\_Fig 3.3). Because the haplotype block 2 extended only nearly 20 Kb and consisted of two significant markers, this locus may be a relatively confident candidate. Nevertheless, considering these significant SNPs presented high FDR, further investigation is necessary to confirm the candidate loci of stamen length.

## 3.4 Discussion

### 3.4.1 QTL on chromosome 2, 3 and 7

Previous studies have revealed three QTL controlling the stamen length in *S. pimpinellifolium*: *atl2.1* and *atl7.1* were mapped via a backcross population derived

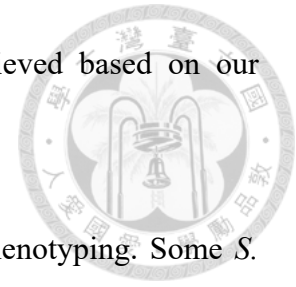


from *S. lycopersicum* crossing to *S. pimpinellifolium* LA1589; *ant3.2* was mapped via a F<sub>2</sub> population derived from *S. pimpinellifolium* LA1237 crossing to *S. pimpinellifolium* LA1581 (Georgiady et al., 2002; Grandillo & Tanksley, 1996). *atl2.1*, *atl7.1* and *ant3.2* explained 6.5%, 17.5% and 35.2% phenotypic variation, respectively, suggesting the confidence intervals extended from 10 to 32 cM (Darvasi & Soller, 1997). However, due to the large confidence intervals from those previous studies, we could not conclude if they were the same QTL as ours, though we indeed identified the QTL on chromosome 2, 3 and 7 as well.

### 3.4.2 Large sample size is essential for GWAS

GWAS is never successful unless a marker is linked with the real QTL contributed to a studying trait. Hence, saturated marker density seems to be the most important factors in GWAS. However, a small population size can lead to the decrease of marker density because many alleles will be excluded due to allele frequency if a studying population is not large enough. Rare alleles were removed in GWAS because of insufficient detecting power (Hamblin, Buckler, & Jannink, 2011; Ingvarsson & Street, 2011; Visscher et al., 2017). This phenomenon was also observed in this study, resulting in only one third SNPs passed the threshold of minor allele frequency. Another reason for large sampling is to detect the loci with small effect size. For a complex trait, generally controlled by many genes with small effects, a larger sample can increase detecting power and decrease FDR (Ingvarsson & Street, 2011; Korte & Farlow, 2013). In this study, all the significant loci revealed high FDR, suggesting a larger sample size was required. We estimated the obligatory sample size based on the heritability and detecting power. The heritability of the stamen length was 0.54 and 0.65 in TASSEL and GEMMA, respectively. To achieve 80% power, at least 1,400 samples are required

(Visscher et al., 2014). Unfortunately, only 5% power was achieved based on our sample size.



Another factor also affects the sample size is the successful phenotyping. Some *S. pimpinellifolium* accessions postponed the flowering in long-day condition (Soyk et al., 2017). It is possible that some accessions in this study are sensitive to photoperiod, resulting in only 79 accessions phenotyped (S\_Tab 3.1). Even with a larger sample size, this sensitivity of photoperiod is still an uncertain factor until a full phenotyping survey for each *S. pimpinellifolium* accession is completed.

It should be noted that increasing population size generally affects GWAS results because allele frequency and population structure are evaluated based on studying populations (Brachi et al., 2011). Although a wild germplasm is generally utilized in GWAS, it should be treated with caution that this material may introduce more rare alleles rather than common alleles, potentially limiting the detection of rare SNPs. We recommend two sampling strategies to increase sampling size for this study. One is to gather more facultative autogamous accessions in the northern Peru because those are higher outcrossing rate and highly diverse (Rick et al., 1977). The higher outcrossing rate can accelerate the breakdown of population structure and maintain genetic diversity simultaneously. This is supported by our result that several highly diverse accessions and also admixture genomes are revealed in the northern Peru. Moreover, the accessions in the north of Peru would maintain as many alleles as their ancestries because *S. pimpinellifolium* originated in the northern Peru. The other is to sample evenly in each subpopulation because it could increase the frequency of rare alleles (Brachi et al., 2011). Although this strategy would probably increase population structure due to the differential subpopulations (S\_Tab 2.6), it could focus on the genes related to local



adaption via several statistical methods designed to eliminate the structure (Brachi et al., 2015; Fournier-Level et al., 2011; Korte & Farlow, 2013).



### 3.4.3 $r^2$ or $D'$ as an indicator for LD

Generally  $r^2$  or  $D'$  are used to describe the LD of a given population for different purposes (Flint-Garcia et al., 2003; Soto-Cerda & Cloutier, 2012).  $r^2$  incorporates the history of recombination and mutation. Scientists utilize  $r^2$  to present LD in GWAS because it presents the correlation between markers and QTL. However,  $r^2$  is easily inflated by mutations or genetic heterogeneity (Korte & Farlow, 2013). To observe only the recombinant events,  $D'$  was also estimated (S\_Fig 3.4). According to S\_Fig 3.4, the interval from SSL2.50ch03\_56799394 to SSL2.50ch03\_56828279 did not form a clear LD block if based on the heatmap of  $r^2$ . SSL2.50ch03\_56799394 is randomly associated with SSL2.50ch03\_56809044, SSL2.50ch03\_56809050 and SSL2.50ch03\_56810075; SSL2.50ch03\_56810088 is randomly associated with SSL2.50ch03\_56828146, SSL2.50ch03\_56828151 and SSL2.50ch03\_56828153. However, considering this interval from SSL2.50ch03\_56799394 to SSL2.50ch03\_56828279 is less than 30 Kb, and a haplotype is detected within this interval, this interval should be a LD block. The contrast may result from the overestimation of  $r^2$  that includes all the mutations. This contrast diminished when  $D'$  served as the indicator of LD; most of  $D'$  in this interval were greater than 0.5 (S\_Fig 3.4).  $D'$  accounts only the recombinant history, making it appropriate to build LD blocks that reflect inherited units (Flint-Garcia et al., 2003). Despite  $D'$  actually supports the LD decay and the haplotype block in our example, we prefer  $r^2$  because it describes the genetic diversity of this collection. Besides, the estimation of  $r^2$  could be more stable when sample size and marker increase; consequently it can build haplotype blocks more precisely (Gaunt et al., 2007).

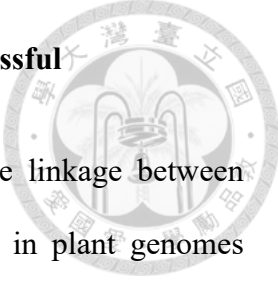
Therefore, we expect an estimation of  $r^2$  corresponding to LD decay or haplotypes in an experiment with a larger sample size and more markers.



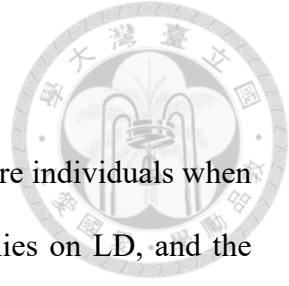
#### 3.4.4 A gap between the estimation of $r^2$ in different softwares

When we demonstrated the pattern of  $D'$  and  $r^2$  for the interval in S\_Fig 3.4, we used TASSEL to estimate both because the calculation of genome-wide pairwise  $D'$  is more practical in TASSEL than in plink. However, we noticed the  $r^2$  in TASSEL was not always equal to that in plink for the same pair SNPs, such as the example listed in S\_Tab 3.2. In plink, haplotype frequency is first estimated and then applied it to the standard  $r^2$  calculation. For a locus with a small sample size or rare allele frequency, the haplotype frequency has multiple solutions, which implies an unstable estimation of  $r^2$  (Gaunt et al., 2007). Meanwhile, in TASSEL, heterozygous genotypes are removed at the beginning and the residual genotypes are applied to the standard  $r^2$  calculation (Bradbury et al., 2007). Therefore, for the loci with high heterozygosity, many genotypes are removed. Among the  $r^2$  of 206,375 pairwise SNPs, a total of 8,397 pairs (4.07%) present equal  $r^2$  in TASSEL and in plink. A total of 1,984 pairs (0.96%) have the difference of  $r^2$  more than 0.1 (S\_Fig 3.5). Considering only the LD decay and required markers were estimated by  $r^2$  in this study, the consequence of the different  $r^2$  could be shown. The overall LD decay based on TASSEL was 22 Kb (S\_Fig 3.6), resulting in about 40,900 markers to cover through the whole tomato genome. Nevertheless, the  $r^2$  in plink was preferred because the elimination of heterozygosity in TASSEL could cause an uncertain bias, especially when accessions with high outcrossing rates were involved in this study.

### 3.4.5 Insufficient coverage makes the build of haplotypes unsuccessful



Haplotype-block based GWAS that takes the advantage of the linkage between nearby alleles and the significant loci has been proved successful in plant genomes (N'Diaye et al., 2017; Qian et al., 2017; Yano et al., 2016). Haplotype has a better biological interpretation than SNP in GWAS because it is inherited as an unit in the same chromosomal block in a giving population (Qian et al., 2017). Additionally, it provides another solution for rare alleles that lack statistic detecting power in GWAS via the formation of haplotype blocks with common alleles (Slatkin, 2008). Despite our common SNPs could not cover the whole genome, we tried to estimate the haplotype blocks using 24,330 SNPs. As a result, a total of 2,928 blocks were built by 11,729 SNPs (S\_Tab 3.3). These haplotypes extended to an average interval of 3 Kb. Since the haplotype could link rare alleles with common ones, about 68,000 SNPs that include the rare alleles in this population were also used to build haplotype blocks. A total of 3,148 blocks were formed by 11,872 SNPs, also with an average interval of 3 Kb (S\_Tab 3.4). These two datasets shared the same 2,385 haplotypes created by 8,510 SNPs. This result suggested that most of the rare alleles assembled haplotypes unsuccessfully. Two reasons may be responsible for the unsuccessful haplotype building. One is the insufficient coverage of these SNPs through the whole genome because the rare SNPs are still at the vicinity of *PstI* cutting site (Browning & Browning, 2007). The other is the overestimation of  $r^2$  due to the small sample size and the genetic heterogeneity (Gaunt et al., 2007; Korte & Farlow, 2013). More local sampling may provide a more stable estimation of  $r^2$  and also haplotypes (Korte & Farlow, 2013).



### 3.4.6 More markers or more individuals

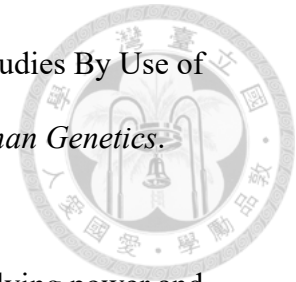
Researchers usually struggle to accomplish more markers or more individuals when designing a limited-budget GWAS. Indeed, the result of GWAS relies on LD, and the LD is affected by marker density. Insufficient markers are not capable to cover through whole genome, resulting in undetectable regions. However, in our case, the confidence of the significant loci and overestimation of  $r^2$  could be improved by a larger sample size. For a complex trait involved in many genes, if we dedicate on more markers, we may detect more loci but with less confidence. If we focus on more sample sizes, we may detect fewer loci but with higher confidence. Therefore, a larger sample size should take priority over a greater number of markers for a reliable GWAS result, especially when a complex trait is involved.

### 3.5 Reference

- Brachi, B., Meyer, C. G., Villoutreix, R., Platt, A., Morton, T. C., Roux, F., & Bergelson, J. (2015). Coselected genes determine adaptive variation in herbivore resistance throughout the native range of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.1421416112>
- Brachi, B., Morris, G. P., & Borevitz, J. O. (2011). Genome-wide association studies in plants: The missing heritability is in the field. *Genome Biology*. <https://doi.org/10.1186/gb-2011-12-10-232>
- Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btm308>
- Browning, S. R., & Browning, B. L. (2007). Rapid and Accurate Haplotype Phasing

and Missing-Data Inference for Whole-Genome Association Studies By Use of  
Localized Haplotype Clustering. *The American Journal of Human Genetics*.

<https://doi.org/10.1086/521987>



Darvasi, A., & Soller, M. (1997). A simple method to calculate resolving power and  
confidence interval of QTL map location. *Behavior Genetics*.

<https://doi.org/10.1023/A:1025685324830>

Flint-Garcia, S. A., Thornsberry, J. M., & Buckler, E. S. (2003). Structure of Linkage  
Disequilibrium in Plants. *Annual Review of Plant Biology*.

<https://doi.org/10.1146/annurev.arplant.54.031902.134907>

Fournier-Level, A., Korte, A., Cooper, M. D., Nordborg, M., Schmitt, J., & Wilczek,  
A. M. (2011). A map of local adaptation in *Arabidopsis thaliana*. *Science*,

334(6052), 86–89. <https://doi.org/10.1126/science.1209271>

Gaunt, T. R., Rodríguez, S., & Day, I. N. M. (2007). Cubic exact solutions for the  
estimation of pairwise haplotype frequencies: Implications for linkage  
disequilibrium analyses and a web tool “CubeX.” *BMC Bioinformatics*.

<https://doi.org/10.1186/1471-2105-8-428>

Georgiady, M. S., Whitkus, R. W., & Lord, E. M. (2002). Genetic analysis of traits  
distinguishing outcrossing and self-pollinating forms of currant tomato,

*Lycopersicon pimpinellifolium* (Jusl.) Mill. *Genetics*.

Grandillo, S., & Tanksley, S. D. (1996). QTL analysis of horticultural traits  
differentiating the cultivated tomato from the closely related species *Lycopersicon*  
*pimpinellifolium*. *Theoretical and Applied Genetics*.

<https://doi.org/10.1007/BF00224033>

Hamblin, M. T., Buckler, E. S., & Jannink, J. L. (2011). Population genetics of  
genomics-based crop improvement methods. *Trends in Genetics*.

<https://doi.org/10.1016/j.tig.2010.12.003>

Ingvarsson, P. K., & Street, N. R. (2011). Association genetics of complex traits in plants. *New Phytologist*. <https://doi.org/10.1111/j.1469-8137.2010.03593.x>

Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: A review. *Plant Methods*. <https://doi.org/10.1186/1746-4811-9-29>

N'Diaye, A., Haile, J. K., Cory, A. T., Clarke, F. R., Clarke, J. M., Knox, R. E., & Pozniak, C. J. (2017). Single marker and haplotype-based association analysis of semolina and pasta colour in elite durum wheat breeding lines using a high-density consensus map. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0170941>

Qian, L., Hickey, L. T., Stahl, A., Werner, C. R., Hayes, B., Snowdon, R. J., & Voss-Fels, K. P. (2017). Exploring and Harnessing Haplotype Diversity to Improve Yield Stability in Crops. *Frontiers in Plant Science*, 8, 1–11.

<https://doi.org/10.3389/fpls.2017.01534>

Rick, C. M., Fobes, J. F., & Holle, M. (1977). Genetic variation in *Lycopersicon pimpinellifolium*: Evidence of evolutionary change in mating systems. *Plant Systematics and Evolution*. <https://doi.org/10.1007/BF00984147>

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*. <https://doi.org/10.1038/nmeth.2089>

Shin, J. H., Blay, S., Graham, J., & McNeney, B. (2015). LDheatmap : An R Function for Graphical Display of Pairwise Linkage Disequilibria Between Single Nucleotide Polymorphisms . *Journal of Statistical Software*.

<https://doi.org/10.18637/jss.v016.c03>

Slatkin, M. (2008). Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*.

<https://doi.org/10.1038/nrg2361>

Soto-Cerda, B. J., & Cloutier, S. 2012 Association mapping in plant genomes, in *Genetic Diversity in Plants*, edited by C. Mahmut. InTech, Rijeka.

Soyk, S., Müller, N. A., Park, S. J., Schmalenbach, I., Jiang, K., Hayama, R., ...

Lippman, Z. B. (2017). Variation in the flowering gene *SELF PRUNING 5G* promotes day-neutrality and early yield in tomato. *Nature Genetics*.

<https://doi.org/10.1038/ng.3733>

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*.

<https://doi.org/10.1073/pnas.1530509100>

Turner, S. D. (2014). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. <https://doi.org/10.1101/005165>

Visscher, P. M., Hemani, G., Vinkhuyzen, A. A. E., Chen, G. B., Lee, S. H., Wray, N. R., ... Yang, J. (2014). Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples. *PLoS Genetics*.

<https://doi.org/10.1371/journal.pgen.1004269>

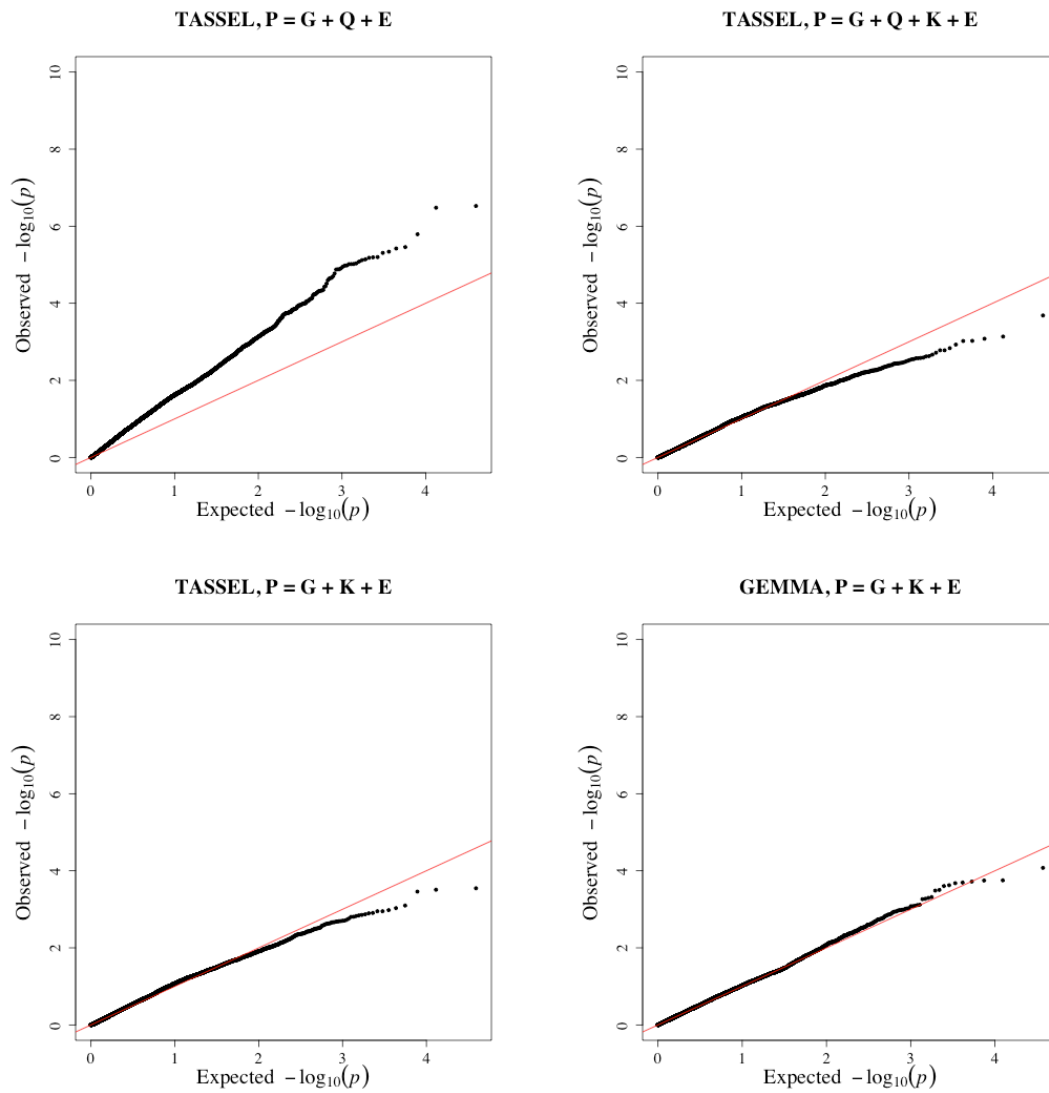
Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*.

<https://doi.org/10.1016/j.ajhg.2017.06.005>

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P. C., Hu, L., ... Matsuoka, M. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nature Genetics*, 48(8), 927–934. <https://doi.org/10.1038/ng.3596>

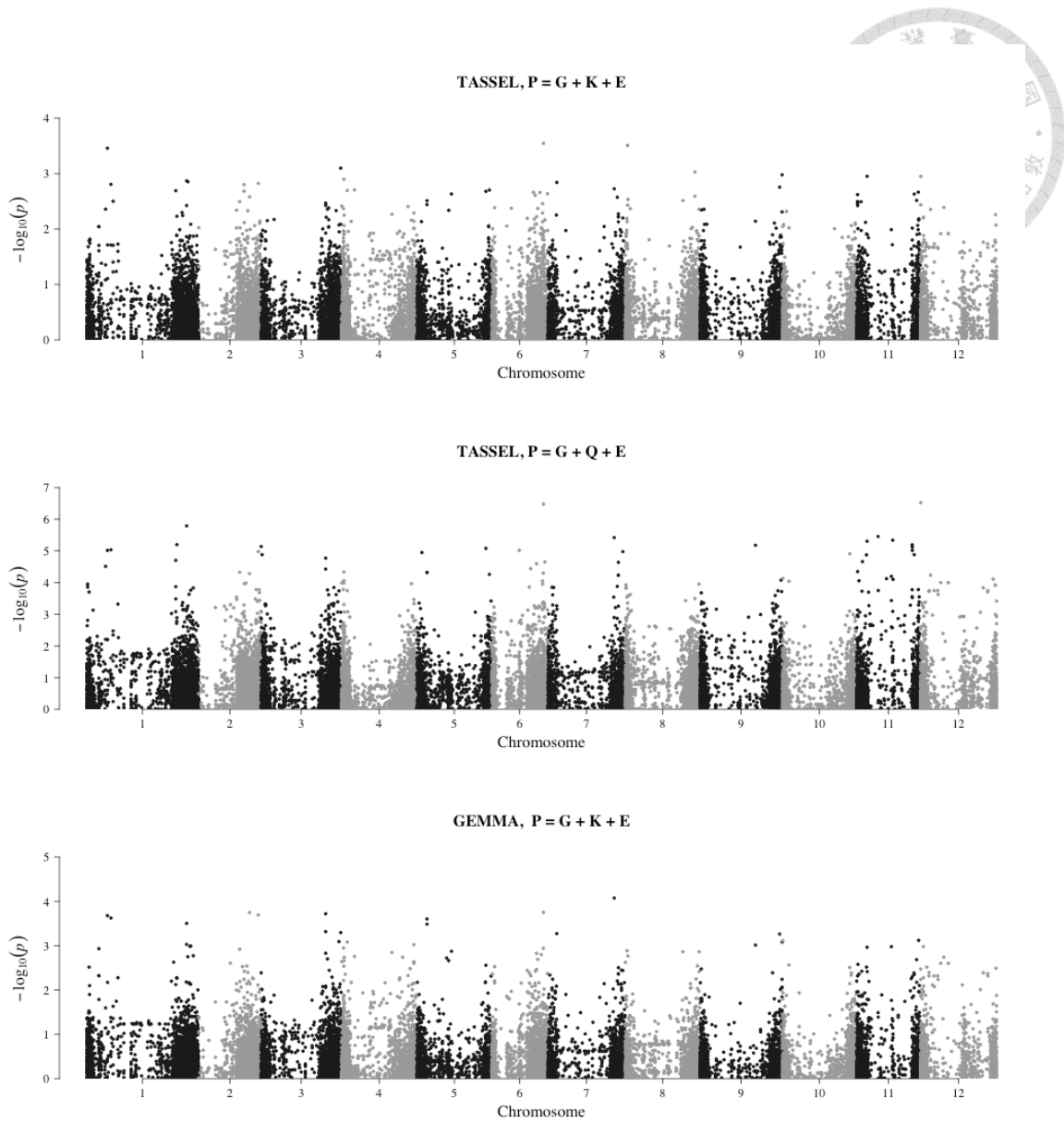
Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*. <https://doi.org/10.1038/ng.2310>

### 3.6 Supplementary Data



S\_Fig 3.1 The Q-Q plots.

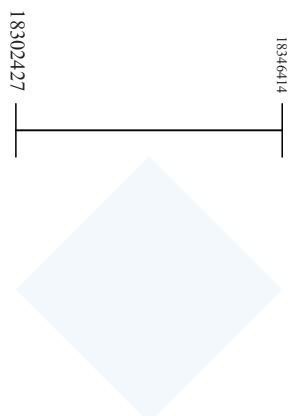




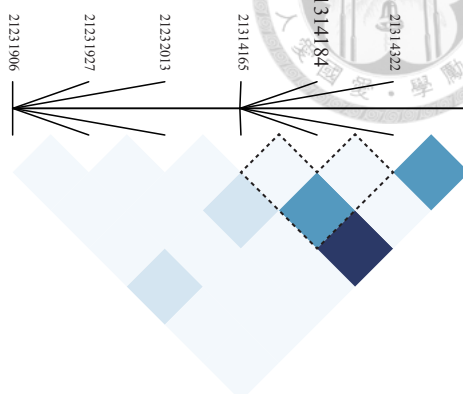
S\_Fig 3.2 The manhattan plots.



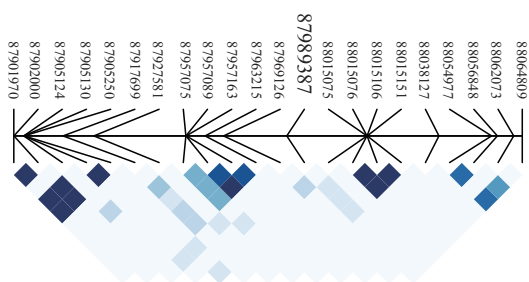
SSL2.50ch01\_18302427



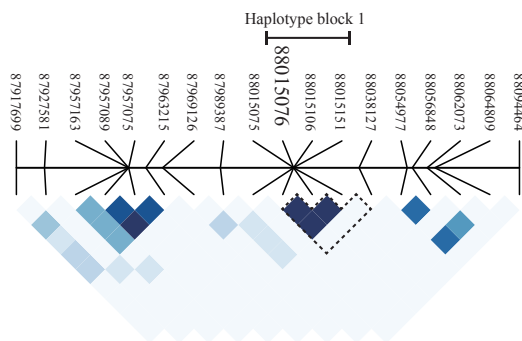
SSL2.50ch01\_21314184



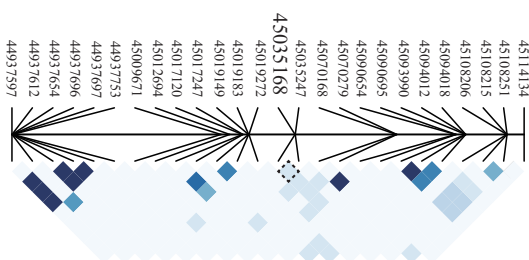
SSL2.50ch01\_87989387



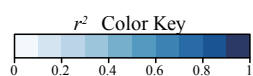
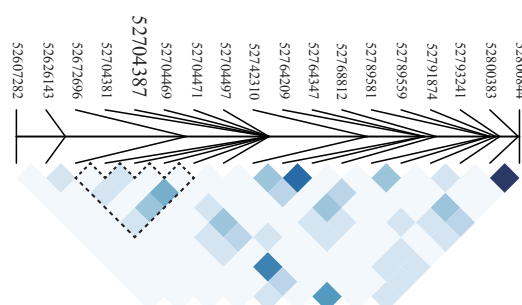
SSL2.50ch01\_88015076



SSL2.50ch02\_45035168

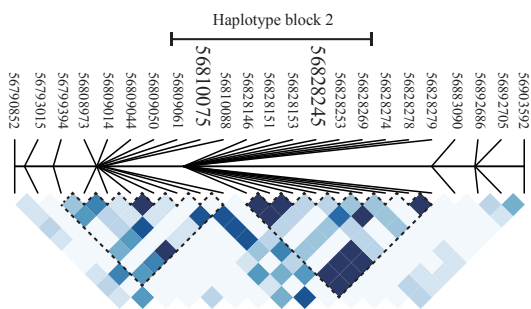


SSL2.50ch02\_52704387

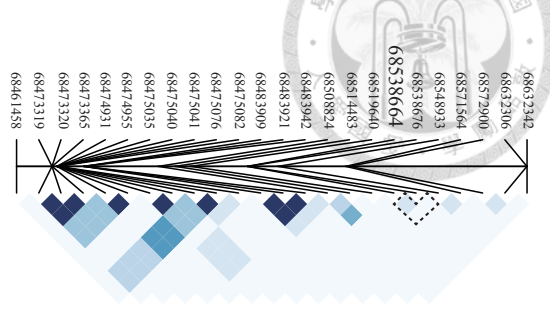


S\_Fig 3.3 (page 1/4)

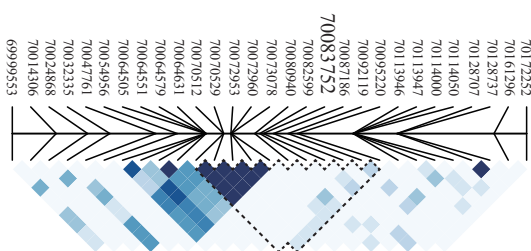
SSL2.50ch03\_56810075 & SSL2.50ch03\_56828245



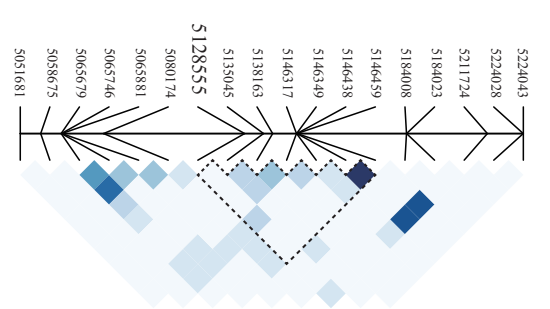
SSL2.50ch03\_68538664



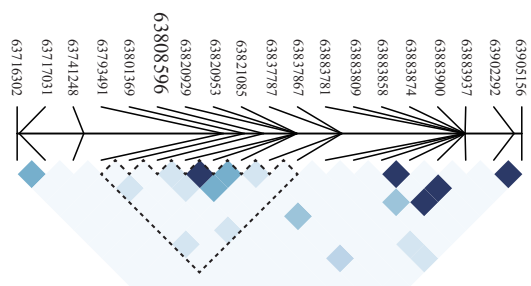
SSL2.50ch03\_70083752



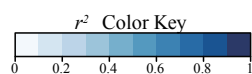
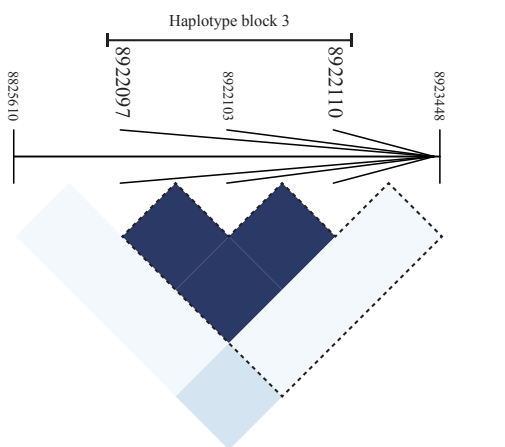
SSL2.50ch04\_5128555



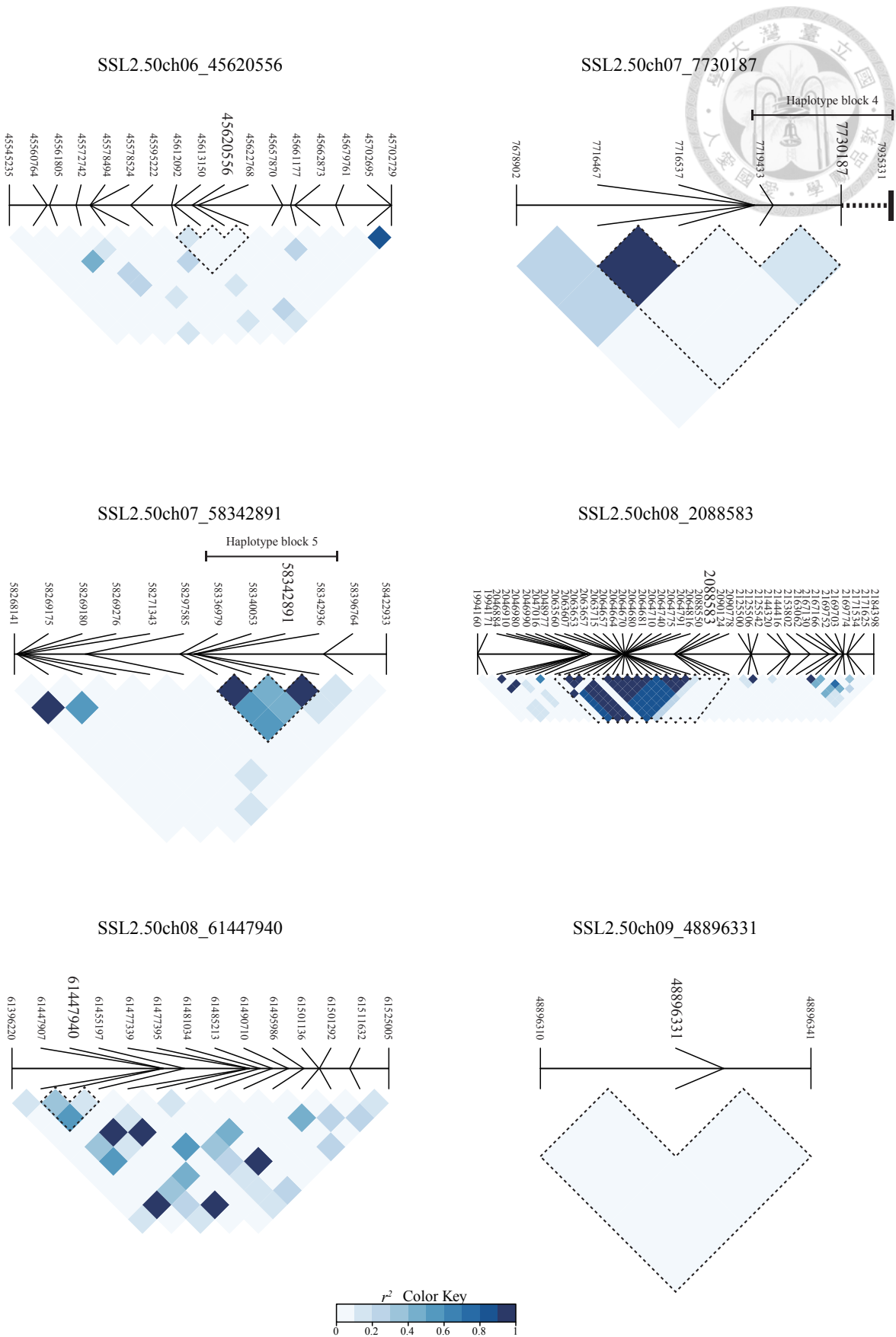
SSL2.50ch04\_63808596



SSL2.50ch05\_8922097 & SSL2.50ch05\_8922110



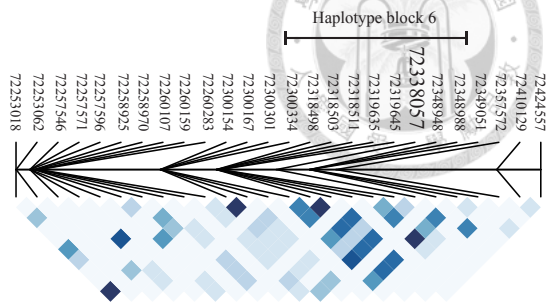
S\_Fig 3.3 (page 2/4)



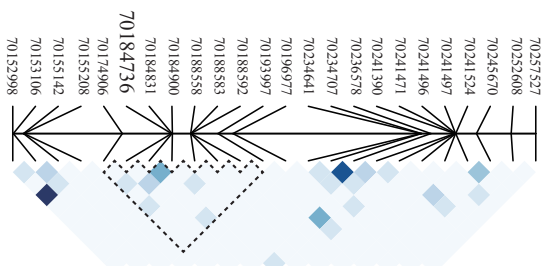
S\_Fig 3.3 (page 3/4)



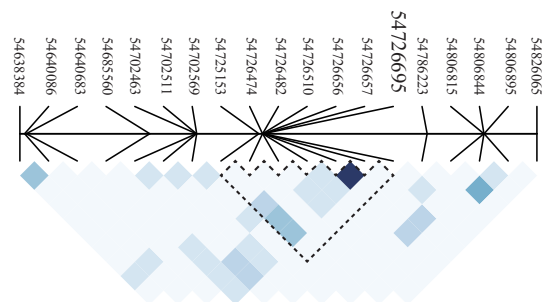
SSL2.50ch09\_72338057



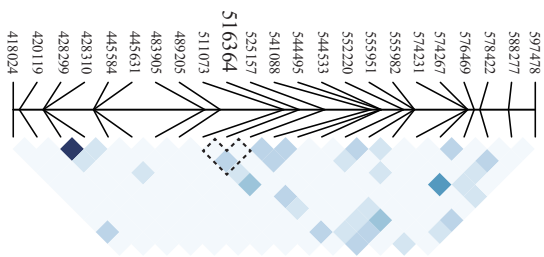
SSL2.50ch09\_70184736



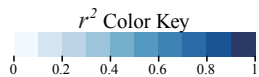
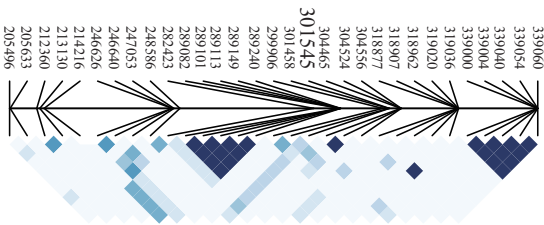
SSL2.50ch11\_54726695

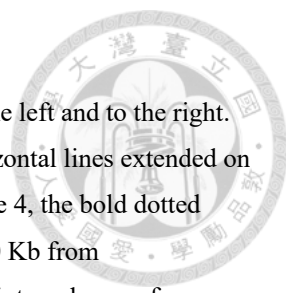


SSL2.50ch10\_516364



SSL2.50ch12\_301545





S\_Fig 3.3 The heatmap of LD for each significant locus in GWAS.

For each locus, the heatmap shows the  $r^2$  of the flanking SNPs within 100 Kb to the left and to the right.

The dotted lines indicate the intervals of LD decay of each chromosome. The horizontal lines extended on the top of the SNPs indicate the haplotype blocks predicted by plink. For haplotype 4, the bold dotted

lines labelled on the physical map indicateds this haplotype extends more than 100 Kb from

SSL2.50ch07\_7730187. Taking SSL2.50ch01\_21314184 for example, its 100-Kb interval spans from

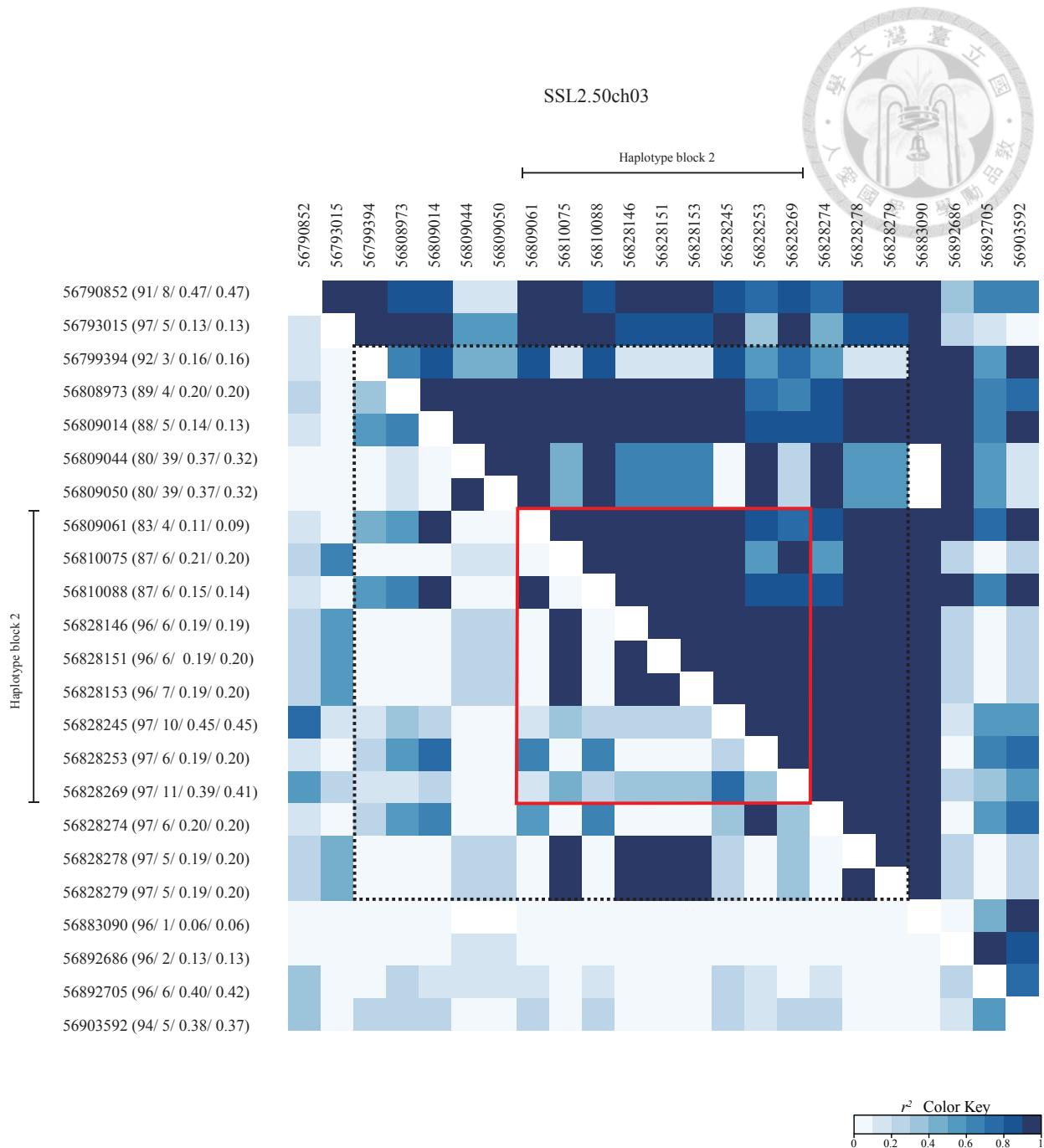
21214185 to 21414184. Therefore, all SNPs within this interval are included in this heatmap, resulting in

only seven SNPs, SSL2.50ch01\_21231906, SSL2.50ch01\_21231927, SSL2.50ch01\_21232013,

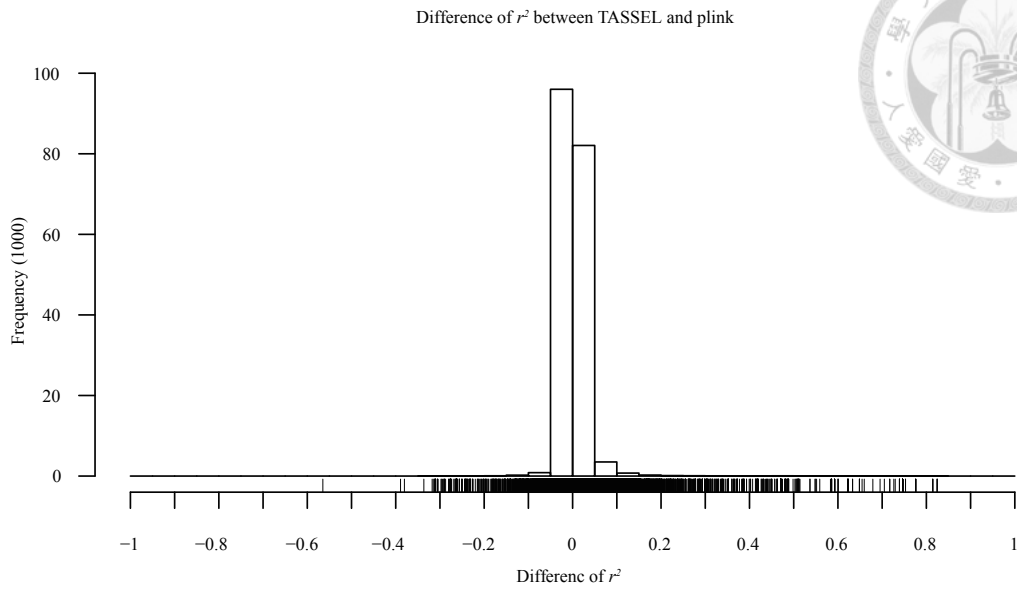
SSL2.50ch01\_21314165, SSL2.50ch01\_21314184, SSL2.50ch01\_21314322 and

SSL2.50ch01\_21397188. The LD decay of chromosome 1 is 14 Kb. Hence, from 21300185 to 21328184,

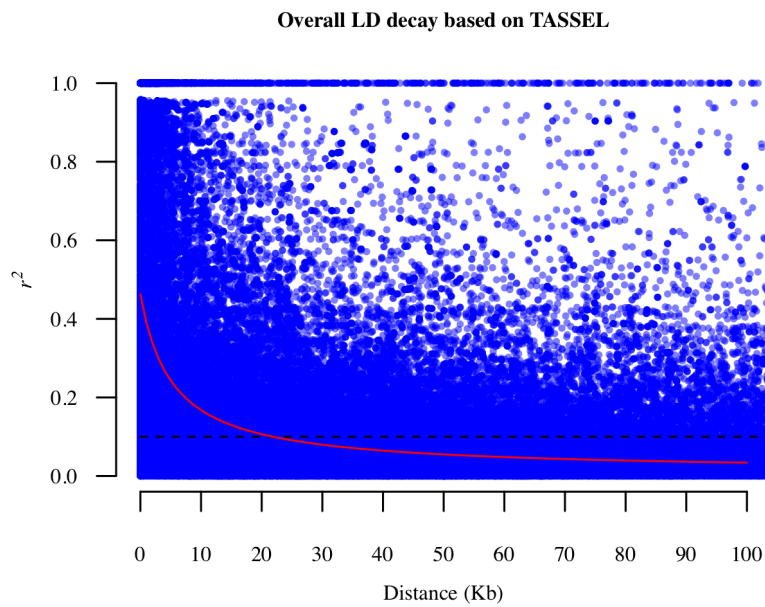
there are three pair-wise LD labelled by the dotted lines.



S\_Fig 3.4 The heatmap of  $r^2$  and  $D'$  from SSL2.50ch03\_56790852 to SSL2.50ch03\_56903592. The  $r^2$  and  $D'$  are plotted on the lower and upper triangle, respectively, sharing the same color key. The red line and dotted line indicate the interval of the haplotype 3 and that of the LD decay of chromosome 3, respectively. The parentheses labelled on the y-axis indicate, from left to right, the sample size, the heterozygous size, the minor allele frequency based on the sample size and that without heterozygosity.



S\_Fig 3.5 The difference of  $r^2$  between TASSEL and plink based on 206,375 pair-wise LD.



S\_Fig 3.6 The overall LD decay based on TASSEL.

The red curve indicates non-linear regression and black dotted line refers to the baseline of  $r^2$  at 0.1.



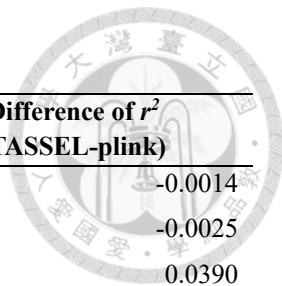
S\_Tab 3.1 The stamen length of each accession.

Accession	Stamen length (mm)	Accession	Stamen length (mm)	Accession	Stamen length (mm)	Accession	Stamen length (mm)	Accession	Stamen length (mm)
LA0114	9.3061	LA1380	NA	LA1587	10.1708	LA1686	13.1162	LA2390	8.6951
LA0373	7.6795	LA1381	8.3840	LA1589	8.7391	LA1687	11.4246	LA2401	8.0174
LA0391	7.6177	LA1382	11.0898	LA1590	8.8334	LA1688	12.3323	LA2533	7.0001
LA0397	11.0742	LA1466	11.5480	LA1591	8.1812	LA1689	8.7704	LA2645	10.3808
LA0400	NA	LA1469	10.4435	LA1593	9.4802	LA1690	NA	LA2646	10.2892
LA0417	NA	LA1471	NA	LA1595	7.9726	LA1720	NA	LA2647	NA
LA0442	7.5946	LA1478	9.5352	LA1596	7.9768	LA1729	7.1697	LA2652	10.2621
LA1236	NA	LA1514	7.0176	LA1599	7.4394	LA1921	6.9633	LA2653	9.3440
LA1237	6.7490	LA1521	6.9549	LA1601	7.6334	LA1923	7.0323	LA2655	10.3550
LA1245	8.9332	LA1547	11.0033	LA1602	6.7187	LA1924	7.9733	LA2656	10.9326
LA1246	NA	LA1576	7.4378	LA1606	7.9248	LA1933	5.9647	LA2659	NA
LA1256	NA	LA1577	8.3659	LA1615	NA	LA1936	6.6613	LA2852	8.5201
LA1261	7.7446	LA1578	8.8759	LA1617	NA	LA2097	8.9033	LA2915	11.7484
LA1279	NA	LA1579	11.0729	LA1628	9.3352	LA2102	NA	LA3638	NA
LA1280	8.4673	LA1580	11.4193	LA1629	7.4897	LA2146	10.6952		
LA1301	7.8352	LA1581	10.9539	LA1645	7.5196	LA2149	8.6963		
LA1335	6.5901	LA1582	11.0642	LA1659	8.8385	LA2173	8.5997		
LA1348	10.0666	LA1583	11.6203	LA1670	8.6164	LA2181	9.4307		
LA1349	11.1413	LA1584	11.2812	LA1683	9.5078	LA2183	9.2328		
LA1371	NA	LA1585	9.4693	LA1684	NA	LA2186	NA		
LA1375	8.2370	LA1586	10.2029	LA1685	11.2306	LA2389	8.4093		

S\_Tab 3.2 The difference of pairwise  $r^2$  between TASSEL and plink, taking SSL2.50ch03\_56790852 to SSL2.50ch03\_56903592 (the SNPs in S\_Fig 3.4) for example

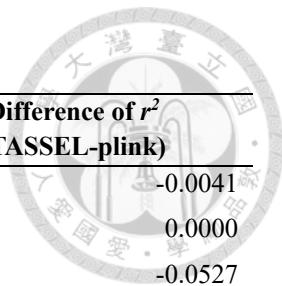
SNP_A	SNP_B	$r^2$ (plink)	$r^2$ (TASSEL)	D'	Difference of $r^2$ (TASSEL-plink)
SSL2.50ch03_56790852	SSL2.50ch03_56793015	0.1382	0.1291	1.0000	-0.0091
SSL2.50ch03_56790852	SSL2.50ch03_56799394	0.1792	0.1914	1.0000	0.0122
SSL2.50ch03_56790852	SSL2.50ch03_56808973	0.2527	0.2626	0.8787	0.0099
SSL2.50ch03_56790852	SSL2.50ch03_56809014	0.1588	0.1403	0.8140	-0.0185
SSL2.50ch03_56790852	SSL2.50ch03_56809044	0.0321	0.0079	0.1162	-0.0242
SSL2.50ch03_56790852	SSL2.50ch03_56809050	0.0321	0.0079	0.1162	-0.0242
SSL2.50ch03_56790852	SSL2.50ch03_56809061	0.1661	0.1543	1.0000	-0.0118
SSL2.50ch03_56790852	SSL2.50ch03_56810075	0.2465	0.2737	1.0000	0.0272
SSL2.50ch03_56790852	SSL2.50ch03_56810088	0.1680	0.1434	0.8111	-0.0246
SSL2.50ch03_56793015	SSL2.50ch03_56799394	0.0364	0.0254	1.0000	-0.0109
SSL2.50ch03_56793015	SSL2.50ch03_56808973	0.0422	0.0357	1.0000	-0.0065
SSL2.50ch03_56793015	SSL2.50ch03_56809014	0.0267	0.0216	1.0000	-0.0051
SSL2.50ch03_56793015	SSL2.50ch03_56809044	0.0386	0.0876	0.5632	0.0490
SSL2.50ch03_56793015	SSL2.50ch03_56809050	0.0386	0.0876	0.5632	0.0490
SSL2.50ch03_56793015	SSL2.50ch03_56809061	0.0192	0.0163	1.0000	-0.0029
SSL2.50ch03_56793015	SSL2.50ch03_56810075	0.6702	0.6735	1.0000	0.0033
SSL2.50ch03_56793015	SSL2.50ch03_56810088	0.0299	0.0230	1.0000	-0.0069
SSL2.50ch03_56793015	SSL2.50ch03_56828146	0.5048	0.5261	0.8811	0.0213
SSL2.50ch03_56799394	SSL2.50ch03_56808973	0.3019	0.3012	0.6400	-0.0007
SSL2.50ch03_56799394	SSL2.50ch03_56809014	0.5264	0.5697	0.8563	0.0433
SSL2.50ch03_56799394	SSL2.50ch03_56809044	0.0593	0.0210	0.4861	-0.0383
SSL2.50ch03_56799394	SSL2.50ch03_56809050	0.0593	0.0210	0.4861	-0.0383

S\_Tab 3.2 (Continued)



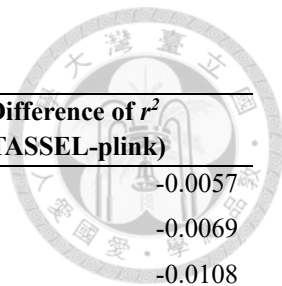
SNP_A	SNP_B	$r^2$ (plink)	$r^2$ (TASSEL)	D'	Difference of $r^2$ (TASSEL-plink)
SSL2.50ch03_56799394	SSL2.50ch03_56809061	0.4181	0.4167	0.8333	-0.0014
SSL2.50ch03_56799394	SSL2.50ch03_56810075	0.0043	0.0018	0.1889	-0.0025
SSL2.50ch03_56799394	SSL2.50ch03_56810088	0.5287	0.5677	0.8555	0.0390
SSL2.50ch03_56799394	SSL2.50ch03_56828146	0.0029	0.0011	0.1590	-0.0018
SSL2.50ch03_56799394	SSL2.50ch03_56828151	0.0029	0.0011	0.1590	-0.0018
SSL2.50ch03_56808973	SSL2.50ch03_56809014	0.6851	0.6751	1.0000	-0.0100
SSL2.50ch03_56808973	SSL2.50ch03_56809044	0.2287	0.1500	1.0000	-0.0787
SSL2.50ch03_56808973	SSL2.50ch03_56809050	0.2287	0.1500	1.0000	-0.0787
SSL2.50ch03_56808973	SSL2.50ch03_56809061	0.5106	0.5404	1.0000	0.0298
SSL2.50ch03_56808973	SSL2.50ch03_56810075	0.0758	0.0656	1.0000	-0.0102
SSL2.50ch03_56808973	SSL2.50ch03_56810088	0.6453	0.6723	1.0000	0.0270
SSL2.50ch03_56808973	SSL2.50ch03_56828146	0.0714	0.0533	1.0000	-0.0181
SSL2.50ch03_56808973	SSL2.50ch03_56828151	0.0714	0.0533	1.0000	-0.0181
SSL2.50ch03_56808973	SSL2.50ch03_56828153	0.0550	0.0507	1.0000	-0.0043
SSL2.50ch03_56809014	SSL2.50ch03_56809044	0.1242	0.0750	1.0000	-0.0492
SSL2.50ch03_56809014	SSL2.50ch03_56809050	0.1242	0.0750	1.0000	-0.0492
SSL2.50ch03_56809014	SSL2.50ch03_56809061	0.8639	1.0000	1.0000	0.1361
SSL2.50ch03_56809014	SSL2.50ch03_56810075	0.0497	0.0397	1.0000	-0.0100
SSL2.50ch03_56809014	SSL2.50ch03_56810088	0.9475	1.0000	1.0000	0.0525
SSL2.50ch03_56809014	SSL2.50ch03_56828146	0.0498	0.0323	1.0000	-0.0175
SSL2.50ch03_56809014	SSL2.50ch03_56828151	0.0498	0.0323	1.0000	-0.0175
SSL2.50ch03_56809014	SSL2.50ch03_56828153	0.0526	0.0333	1.0000	-0.0193

S\_Tab 3.2 (Continued)



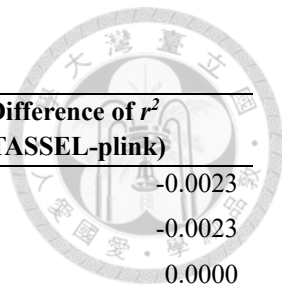
SNP_A	SNP_B	$r^2$ (plink)	$r^2$ (TASSEL)	D'	Difference of $r^2$ (TASSEL-plink)
SSL2.50ch03_56809014	SSL2.50ch03_56828245	0.2364	0.2323	1.0000	-0.0041
SSL2.50ch03_56809044	SSL2.50ch03_56809050	1.0000	1.0000	1.0000	0.0000
SSL2.50ch03_56809044	SSL2.50ch03_56809061	0.1282	0.0754	1.0000	-0.0527
SSL2.50ch03_56809044	SSL2.50ch03_56810075	0.0706	0.1667	0.4444	0.0961
SSL2.50ch03_56809044	SSL2.50ch03_56810088	0.1713	0.0933	1.0000	-0.0780
SSL2.50ch03_56809044	SSL2.50ch03_56828146	0.0990	0.2813	0.6224	0.1823
SSL2.50ch03_56809044	SSL2.50ch03_56828151	0.0990	0.2813	0.6224	0.1823
SSL2.50ch03_56809044	SSL2.50ch03_56828153	0.0878	0.2775	0.6190	0.1897
SSL2.50ch03_56809044	SSL2.50ch03_56828245	0.0097	0.0000	0.0000	-0.0097
SSL2.50ch03_56809044	SSL2.50ch03_56828253	0.0762	0.0717	1.0000	-0.0045
SSL2.50ch03_56809050	SSL2.50ch03_56809061	0.1282	0.0754	1.0000	-0.0527
SSL2.50ch03_56809050	SSL2.50ch03_56810075	0.0706	0.1667	0.4444	0.0961
SSL2.50ch03_56809050	SSL2.50ch03_56810088	0.1713	0.0933	1.0000	-0.0780
SSL2.50ch03_56809050	SSL2.50ch03_56828146	0.0990	0.2813	0.6224	0.1823
SSL2.50ch03_56809050	SSL2.50ch03_56828151	0.0990	0.2813	0.6224	0.1823
SSL2.50ch03_56809050	SSL2.50ch03_56828153	0.0878	0.2775	0.6190	0.1897
SSL2.50ch03_56809050	SSL2.50ch03_56828245	0.0097	0.0000	0.0000	-0.0097
SSL2.50ch03_56809050	SSL2.50ch03_56828253	0.0762	0.0717	1.0000	-0.0045
SSL2.50ch03_56809050	SSL2.50ch03_56828269	0.0004	0.0191	0.2047	0.0187
SSL2.50ch03_56809061	SSL2.50ch03_56810075	0.0289	0.0303	1.0000	0.0014
SSL2.50ch03_56809061	SSL2.50ch03_56810088	0.8659	1.0000	1.0000	0.1341
SSL2.50ch03_56809061	SSL2.50ch03_56828146	0.0301	0.0244	1.0000	-0.0057

S\_Tab 3.2 (Continued)



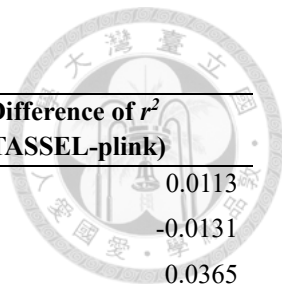
SNP_A	SNP_B	$r^2$ (plink)	$r^2$ (TASSEL)	D'	Difference of $r^2$ (TASSEL-plink)
SSL2.50ch03_56809061	SSL2.50ch03_56828151	0.0301	0.0244	1.0000	-0.0057
SSL2.50ch03_56809061	SSL2.50ch03_56828153	0.0320	0.0252	1.0000	-0.0069
SSL2.50ch03_56809061	SSL2.50ch03_56828245	0.1800	0.1692	1.0000	-0.0108
SSL2.50ch03_56809061	SSL2.50ch03_56828253	0.5275	0.6084	0.8401	0.0809
SSL2.50ch03_56809061	SSL2.50ch03_56828269	0.1513	0.1407	0.7872	-0.0106
SSL2.50ch03_56809061	SSL2.50ch03_56828274	0.4701	0.5297	0.8377	0.0596
SSL2.50ch03_56810075	SSL2.50ch03_56810088	0.0393	0.0361	1.0000	-0.0032
SSL2.50ch03_56810075	SSL2.50ch03_56828146	0.8856	0.9143	1.0000	0.0287
SSL2.50ch03_56810075	SSL2.50ch03_56828151	0.8856	0.9143	1.0000	0.0287
SSL2.50ch03_56810075	SSL2.50ch03_56828153	0.8647	0.9141	1.0000	0.0493
SSL2.50ch03_56810075	SSL2.50ch03_56828245	0.3240	0.3422	1.0000	0.0182
SSL2.50ch03_56810075	SSL2.50ch03_56828253	0.0194	0.0135	0.5515	-0.0059
SSL2.50ch03_56810075	SSL2.50ch03_56828269	0.4228	0.4556	1.0000	0.0328
SSL2.50ch03_56810075	SSL2.50ch03_56828274	0.0233	0.0171	0.5889	-0.0062
SSL2.50ch03_56810075	SSL2.50ch03_56828278	0.8665	0.9145	1.0000	0.0480
SSL2.50ch03_56810088	SSL2.50ch03_56828146	0.0442	0.0333	1.0000	-0.0109
SSL2.50ch03_56810088	SSL2.50ch03_56828151	0.0442	0.0333	1.0000	-0.0109
SSL2.50ch03_56810088	SSL2.50ch03_56828153	0.0471	0.0344	1.0000	-0.0127
SSL2.50ch03_56810088	SSL2.50ch03_56828245	0.2501	0.2408	1.0000	-0.0093
SSL2.50ch03_56810088	SSL2.50ch03_56828253	0.6531	0.6993	0.8836	0.0462
SSL2.50ch03_56810088	SSL2.50ch03_56828269	0.2225	0.2191	0.8479	-0.0034
SSL2.50ch03_56810088	SSL2.50ch03_56828274	0.5944	0.6289	0.8818	0.0345

S\_Tab 3.2 (Continued)



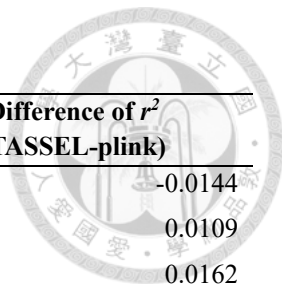
SNP_A	SNP_B	$r^2$ (plink)	$r^2$ (TASSEL)	D'	Difference of $r^2$ (TASSEL-plink)
SSL2.50ch03_56810088	SSL2.50ch03_56828278	0.0345	0.0323	1.0000	-0.0023
SSL2.50ch03_56810088	SSL2.50ch03_56828279	0.0345	0.0323	1.0000	-0.0023
SSL2.50ch03_56828146	SSL2.50ch03_56828151	1.0000	1.0000	1.0000	0.0000
SSL2.50ch03_56828146	SSL2.50ch03_56828153	0.9813	1.0000	1.0000	0.0187
SSL2.50ch03_56828146	SSL2.50ch03_56828245	0.2580	0.2650	1.0000	0.0070
SSL2.50ch03_56828146	SSL2.50ch03_56828253	0.0661	0.0473	1.0000	-0.0189
SSL2.50ch03_56828146	SSL2.50ch03_56828269	0.3580	0.3717	1.0000	0.0137
SSL2.50ch03_56828146	SSL2.50ch03_56828274	0.0704	0.0512	1.0000	-0.0192
SSL2.50ch03_56828146	SSL2.50ch03_56828278	0.9821	1.0000	1.0000	0.0179
SSL2.50ch03_56828146	SSL2.50ch03_56828279	0.9821	1.0000	1.0000	0.0179
SSL2.50ch03_56828146	SSL2.50ch03_56883090	0.0172	0.0127	1.0000	-0.0045
SSL2.50ch03_56828151	SSL2.50ch03_56828153	0.9813	1.0000	1.0000	0.0187
SSL2.50ch03_56828151	SSL2.50ch03_56828245	0.2580	0.2650	1.0000	0.0070
SSL2.50ch03_56828151	SSL2.50ch03_56828253	0.0661	0.0473	1.0000	-0.0189
SSL2.50ch03_56828151	SSL2.50ch03_56828269	0.3580	0.3717	1.0000	0.0137
SSL2.50ch03_56828151	SSL2.50ch03_56828274	0.0704	0.0512	1.0000	-0.0192
SSL2.50ch03_56828151	SSL2.50ch03_56828278	0.9821	1.0000	1.0000	0.0179
SSL2.50ch03_56828151	SSL2.50ch03_56828279	0.9821	1.0000	1.0000	0.0179
SSL2.50ch03_56828151	SSL2.50ch03_56883090	0.0172	0.0127	1.0000	-0.0045
SSL2.50ch03_56828151	SSL2.50ch03_56892686	0.0382	0.0471	0.2750	0.0089
SSL2.50ch03_56828153	SSL2.50ch03_56828245	0.2735	0.2761	1.0000	0.0027
SSL2.50ch03_56828153	SSL2.50ch03_56828253	0.0599	0.0473	1.0000	-0.0126

S\_Tab 3.2 (Continued)



SNP_A	SNP_B	$r^2$ (plink)	$r^2$ (TASSEL)	D'	Difference of $r^2$ (TASSEL-plink)
SSL2.50ch03_56828153	SSL2.50ch03_56828269	0.3604	0.3717	1.0000	0.0113
SSL2.50ch03_56828153	SSL2.50ch03_56828274	0.0642	0.0512	1.0000	-0.0131
SSL2.50ch03_56828153	SSL2.50ch03_56828278	0.9635	1.0000	1.0000	0.0365
SSL2.50ch03_56828153	SSL2.50ch03_56828279	0.9635	1.0000	1.0000	0.0365
SSL2.50ch03_56828153	SSL2.50ch03_56883090	0.0181	0.0130	1.0000	-0.0051
SSL2.50ch03_56828153	SSL2.50ch03_56892686	0.0451	0.0471	0.2750	0.0020
SSL2.50ch03_56828153	SSL2.50ch03_56892705	0.0002	0.0002	0.0438	0.0001
SSL2.50ch03_56828245	SSL2.50ch03_56828253	0.2754	0.2780	1.0000	0.0026
SSL2.50ch03_56828245	SSL2.50ch03_56828269	0.7626	0.7833	1.0000	0.0207
SSL2.50ch03_56828245	SSL2.50ch03_56828274	0.2965	0.3008	1.0000	0.0043
SSL2.50ch03_56828245	SSL2.50ch03_56828278	0.2655	0.2774	1.0000	0.0119
SSL2.50ch03_56828245	SSL2.50ch03_56828279	0.2655	0.2774	1.0000	0.0119
SSL2.50ch03_56828245	SSL2.50ch03_56883090	0.0467	0.0489	1.0000	0.0022
SSL2.50ch03_56828245	SSL2.50ch03_56892686	0.0083	0.0067	0.1876	-0.0015
SSL2.50ch03_56828245	SSL2.50ch03_56892705	0.2035	0.2159	0.5126	0.0124
SSL2.50ch03_56828245	SSL2.50ch03_56903592	0.2729	0.2966	0.5729	0.0237
SSL2.50ch03_56828253	SSL2.50ch03_56828269	0.3416	0.3565	1.0000	0.0149
SSL2.50ch03_56828253	SSL2.50ch03_56828274	0.9284	0.9252	1.0000	-0.0032
SSL2.50ch03_56828253	SSL2.50ch03_56828278	0.0653	0.0483	1.0000	-0.0170
SSL2.50ch03_56828253	SSL2.50ch03_56828279	0.0653	0.0483	1.0000	-0.0170
SSL2.50ch03_56828253	SSL2.50ch03_56883090	0.0149	0.0111	1.0000	-0.0038
SSL2.50ch03_56828253	SSL2.50ch03_56892686	0.0001	0.0006	0.0270	0.0005

S\_Tab 3.2 (Continued)



SNP_A	SNP_B	$r^2$ (plink)	$r^2$ (TASSEL)	D'	Difference of $r^2$ (TASSEL-plink)
SSL2.50ch03_56828253	SSL2.50ch03_56892705	0.1568	0.1424	0.6587	-0.0144
SSL2.50ch03_56828253	SSL2.50ch03_56903592	0.1681	0.1790	0.7562	0.0109
SSL2.50ch03_56828269	SSL2.50ch03_56828274	0.3695	0.3857	1.0000	0.0162
SSL2.50ch03_56828269	SSL2.50ch03_56828278	0.3669	0.3857	1.0000	0.0188
SSL2.50ch03_56828269	SSL2.50ch03_56828279	0.3669	0.3857	1.0000	0.0188
SSL2.50ch03_56828269	SSL2.50ch03_56883090	0.0444	0.0370	1.0000	-0.0073
SSL2.50ch03_56828269	SSL2.50ch03_56892686	0.0181	0.0201	0.2796	0.0020
SSL2.50ch03_56828269	SSL2.50ch03_56892705	0.0900	0.0903	0.3083	0.0002
SSL2.50ch03_56828269	SSL2.50ch03_56903592	0.1840	0.2146	0.5163	0.0306
SSL2.50ch03_56828274	SSL2.50ch03_56828278	0.0695	0.0522	1.0000	-0.0172
SSL2.50ch03_56828274	SSL2.50ch03_56828279	0.0695	0.0522	1.0000	-0.0172
SSL2.50ch03_56828274	SSL2.50ch03_56883090	0.0160	0.0121	1.0000	-0.0039
SSL2.50ch03_56828274	SSL2.50ch03_56892686	0.0000	0.0001	0.0137	0.0001
SSL2.50ch03_56828274	SSL2.50ch03_56892705	0.1313	0.1180	0.5753	-0.0133
SSL2.50ch03_56828274	SSL2.50ch03_56903592	0.1916	0.2046	0.7736	0.0130
SSL2.50ch03_56828278	SSL2.50ch03_56828279	1.0000	1.0000	1.0000	0.0000
SSL2.50ch03_56828278	SSL2.50ch03_56883090	0.0170	0.0130	1.0000	-0.0040
SSL2.50ch03_56828278	SSL2.50ch03_56892686	0.0269	0.0323	0.2242	0.0055
SSL2.50ch03_56828278	SSL2.50ch03_56892705	0.0000	0.0000	0.0077	0.0000
SSL2.50ch03_56828278	SSL2.50ch03_56903592	0.0086	0.0163	0.2398	0.0077
SSL2.50ch03_56828279	SSL2.50ch03_56883090	0.0170	0.0130	1.0000	-0.0040
SSL2.50ch03_56828279	SSL2.50ch03_56892686	0.0269	0.0323	0.2242	0.0055



S\_Tab 3.2 (Continued)

SNP_A	SNP_B	$r^2$ (plink)	$r^2$ (TASSEL)	D'	Difference of $r^2$ (TASSEL-plink)
SSL2.50ch03_56828279	SSL2.50ch03_56892705	0.0000	0.0000	0.0077	0.0000
SSL2.50ch03_56828279	SSL2.50ch03_56903592	0.0086	0.0163	0.2398	0.0077
SSL2.50ch03_56883090	SSL2.50ch03_56892686	0.0019	0.0035	0.0914	0.0016
SSL2.50ch03_56883090	SSL2.50ch03_56892705	0.0130	0.0093	0.4914	-0.0036
SSL2.50ch03_56883090	SSL2.50ch03_56903592	0.0336	0.0361	1.0000	0.0026
SSL2.50ch03_56892686	SSL2.50ch03_56892705	0.2284	0.2281	1.0000	-0.0002
SSL2.50ch03_56892686	SSL2.50ch03_56903592	0.1701	0.1810	0.8552	0.0109
SSL2.50ch03_56892705	SSL2.50ch03_56903592	0.4899	0.5275	0.7446	0.0375

S\_Tab 3.3 The haplotype blocks estimated by the 24,330 SNPs.

We listed the first 12 haplotypes for readers to glimpse the data and the full table is provided on the following link <https://goo.gl/8hUcy3>.



Chr.	Position 1	Position 2	Interval (Kb)	Number of SNPs	Lists of SNPs in this haplotype
1	32,607	52,465	19.859	5	SSL2.50ch01_32607 SSL2.50ch01_32987 SSL2.50ch01_50106 SSL2.50ch01_52415 SSL2.50ch01_52465
1	92,041	92,053	0.013	2	SSL2.50ch01_92041 SSL2.50ch01_92053
1	142,357	142,477	0.121	2	SSL2.50ch01_142357 SSL2.50ch01_142477
1	197,948	197,964	0.017	3	SSL2.50ch01_197948 SSL2.50ch01_197960 SSL2.50ch01_197964
1	243,659	243,824	0.166	5	SSL2.50ch01_243659 SSL2.50ch01_243669 SSL2.50ch01_243696 SSL2.50ch01_243716 SSL2.50ch01_243824
1	352,731	352,758	0.028	4	SSL2.50ch01_352731 SSL2.50ch01_352732 SSL2.50ch01_352733 SSL2.50ch01_352758
1	422,672	422,692	0.021	2	SSL2.50ch01_422672 SSL2.50ch01_422692
1	448,310	452,893	4.584	5	SSL2.50ch01_448310 SSL2.50ch01_448336 SSL2.50ch01_449409 SSL2.50ch01_452890 SSL2.50ch01_452893
1	505,564	507,033	1.470	2	SSL2.50ch01_505564 SSL2.50ch01_507033
1	530,449	530,453	0.005	2	SSL2.50ch01_530449 SSL2.50ch01_530453
1	591,036	597,515	6.480	8	SSL2.50ch01_591036 SSL2.50ch01_591098 SSL2.50ch01_596697 SSL2.50ch01_596724 SSL2.50ch01_596727 SSL2.50ch01_596736 SSL2.50ch01_596747 SSL2.50ch01_597515
1	612,723	612,796	0.074	9	SSL2.50ch01_612723 SSL2.50ch01_612728 SSL2.50ch01_612744 SSL2.50ch01_612750 SSL2.50ch01_612753 SSL2.50ch01_612776 SSL2.50ch01_612779 SSL2.50ch01_612794 SSL2.50ch01_612796

S\_Tab 3.4 The haplotype blocks estimated by about 68,000 SNPs

We listed the first 12 haplotypes for readers to glimpse the data and the full table is provided on the following link <https://goo.gl/8hUcy3>.



Chr.	Position 1	Position 2	Interval (Kb)	Number of SNPs	Lists of SNPs in this haplotype
1	32,607	52,465	19.859	5	SSL2.50ch01_32607 SSL2.50ch01_32987 SSL2.50ch01_50106 SSL2.50ch01_52415 SSL2.50ch01_52465
1	92,041	92,053	0.013	2	SSL2.50ch01_92041 SSL2.50ch01_92053
1	142,357	142,477	0.121	2	SSL2.50ch01_142357 SSL2.50ch01_142477
1	197,948	197,964	0.017	3	SSL2.50ch01_197948 SSL2.50ch01_197960 SSL2.50ch01_197964
1	243,659	243,824	0.166	5	SSL2.50ch01_243659 SSL2.50ch01_243669 SSL2.50ch01_243696 SSL2.50ch01_243716 SSL2.50ch01_243824
1	352,731	352,758	0.028	4	SSL2.50ch01_352731 SSL2.50ch01_352732 SSL2.50ch01_352733 SSL2.50ch01_352758
1	422,672	422,692	0.021	2	SSL2.50ch01_422672 SSL2.50ch01_422692
1	448,310	452,893	4.584	5	SSL2.50ch01_448310 SSL2.50ch01_448336 SSL2.50ch01_449409 SSL2.50ch01_452890 SSL2.50ch01_452893
1	530,449	530,453	0.005	2	SSL2.50ch01_530449 SSL2.50ch01_530453
1	591,036	597,515	6.480	8	SSL2.50ch01_591036 SSL2.50ch01_591098 SSL2.50ch01_596697 SSL2.50ch01_596724 SSL2.50ch01_596727 SSL2.50ch01_596736 SSL2.50ch01_596747 SSL2.50ch01_597515
1	612,723	612,796	0.074	9	SSL2.50ch01_612723 SSL2.50ch01_612728 SSL2.50ch01_612744 SSL2.50ch01_612750 SSL2.50ch01_612753 SSL2.50ch01_612776 SSL2.50ch01_612779 SSL2.50ch01_612794 SSL2.50ch01_612796
1	798,465	798,475	0.011	2	SSL2.50ch01_798465 SSL2.50ch01_798475

# Chapter 4 Three candidate genes controlling stamen length revealed via the transcriptome profiles of M82 and its introgression line TA3178



## 4.1 Purpose

The QTL responsible for the transmission of allogamy to autogamy in tomatoes, *se2.1*, contained five loci: *stamen2.1*, *dehiscence2.1*, *style2.1*, *stamen2.2* and *stamen2.3*. Among them, *style2.1* has been proved its regulation of style elongation due to the InDel in its promoter region. The two candidate genes for stamen length, *stamen2.2* and *stamen2.3*, were tightly linked to *style2.1*, and they have been mapped between marker cLED19A24 and CT9 (Chen & Tanksley, 2004).

M82 belongs to *S. lycopersicum* and has a thrum type flower. TA3178 is an introgression line of M82; it contains a segment of *S. pennellii* genome around *style2.1*. The stamen of homozygous TA3178 was as long as that of M82 because *stamen2.2* and *stamen2.3* had the opposite effect on the stamen length (Chen & Tanksley, 2004). In this chapter, we performed a RNA-seq experiment of M82 and TA3178 to investigate the differentially expressed genes (DEGs) in the introgression segment. The boundary of this segment is defined by comparing the SNP number between M82 and TA3178 because the later has an introgression segment of wild tomato. We expect to narrow down the candidate genes by comparing the expression level and cDNA polymorphism in the interval from cLED19A24 to CT9.



## 4.2 Material and Method

### 4.2.1 RNA sequencing

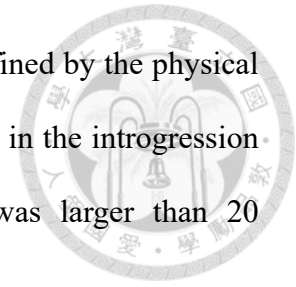
Plants were cultivated one plant per plot in nature-light greenhouses with day/night temperature 20°C/15°C. The RNA was extracted from flower buds using TRIzol® Reagent (Thermo Fisher Scientific, Waltham, MA, USA) following the manufacturer manual. The RNA library and RNA sequencing with the Illumina Hiseq2000 platform were conducted by the Genome Research Center at National Yang-Ming University, Taiwan.

The reads were trimmed and filtered by the R package ShortRead with a threshold of quality score greater than 20 and length longer than 20 bp (Morgan et al., 2009). We aligned the reads to the tomato reference genome SL2.50 by the *subjunc* function in the R package Rsubread (Liao, Smyth, & Shi, 2013). The counts were obtained by the *featureCounts* function in Rsubread with ITAG2.4 gene model and then normalized to reads per kilobase per million mapped reads (RPKM) by the R package edgeR (Robinson, McCarthy, & Smyth, 2009). A gene was expressed if its RPKM was greater than 1. Fold change was calculated by the formula  $\log_2(\text{RPKM}_{M82} + 1) - \log_2(\text{RPKM}_{TA3178} + 1)$  for visualization. All the sequences were uploaded to NCBI SRA database and this BioProject is PRJNA358109.

### 4.2.2 The boundary of introgression segment in TA3178

We used *exactSNP* function in Rsubread to obtain SNPs and filtered with quality score greater than 20. The differences of SNPs per 100 Kb were plotted along each

chromosome to identify the segment. The precise boundary was defined by the physical positions of the marginal SNPs in this segment. The effects of SNP in the introgression segment were detected by SNPeff if the SNP quality score was larger than 20 (Cingolani et al., 2012).



#### **4.2.3 Differential expression analysis**

Since there was no biological replication in this RNA-seq experiment, we identified the DEGs via two methods. The first one depended on the distribution of the fold changes of both lines; DEGs were defined as the genes whose fold changes were outside the 99.9<sup>th</sup> percentile. The other was based on the differentially expression analysis in the R package DEseq because it provided a differential expression analysis in a condition without any replication (Anders et al., 2010). The gene annotations and gene ontology terms of DEGs were obtained from ITAG2.4 gene model downloaded from SGN (Fernandez-Pozo et al., 2015).

#### **4.2.4 The cDNA polymorphisms of the genes from cLED19A24 to CT9**

For the genes from cLED19A24 to CT9, each consensus sequence was extracted by CLC Genomics Workbench version 10.0.1 (QIAGEN, Venlo, Netherlands) and aligned by MEGA7.0 (Kumar, Stecher, & Tamura, 2016). If the coverage of a gene is partially less than 50, the gene is defined as an undetermined polymorphism; if it is entirely less than 50, the gene is defined as no expression.



## 4.3 Result

### 4.3.1 The summary of RNA-seq

We obtained approximately 450 million raw reads and 91% of them were mapped to SL2.50 (Table 4.1). Following Table 4.1, M82 contained 22,741 SNPs and 282,877 InDels; TA3178 contained 17,614 SNPs and 181,551 InDels. Nearly 80% SNPs passed the quality control, leaving 17,802 and 13,963 SNPs in M82 and TA3178, respectively. A total of 19,225 genes were expressed in M82 while 15,794 genes in TA3178. In addition, 3,701 genes expressed only in M82 while 270 genes in TA3178.

Table 4.1 The summary of RNA-seq

RNA-seq	M82	TA3178
Original reads	249,647,418	201,368,484
Reads after trimming	239,854,220 (96.08%) <sup>a</sup>	194,310,543 (96.05%)
Reads after filtering	231,959,989 (92.92%)	187,837,372 (93.28%)
Mapped reads	227,643,933 (91.19%)	183,381,349 (91.07%)
Counts (34,725 genes)	208,594,221	167,330,172
Numbers of SNP	22,741	17,614
SNP quality score > 20	17,802	13,963
Numbers of InDel	282,877	181,551
Expressed genes	19,225	15,794
Uniquely expressed genes	3,701	270

<sup>a</sup>: The percentage was based on the original reads.

### 4.3.2 The 1.1 Mb introgression segment of *S. pennellii*

Since *S. pennellii* is relatively distant from the tomato reference genome, the introgression segment in TA3178 would be revealed more SNPs than that in M82. As a result, the SNP numbers were extremely different between these two lines at the position about 50 Mb in chromosome 2 (S\_Fig 4.1; Figure 4.1). Therefore, the marginal

SNPs of TA3178 in this interval, 49,946,234 bp and 51,013,830 bp, were marked as the boundary of the introgression segment (Figure 4.1). This 1.1 Mb segment contained 159 genes (S\_Tab 4.1).

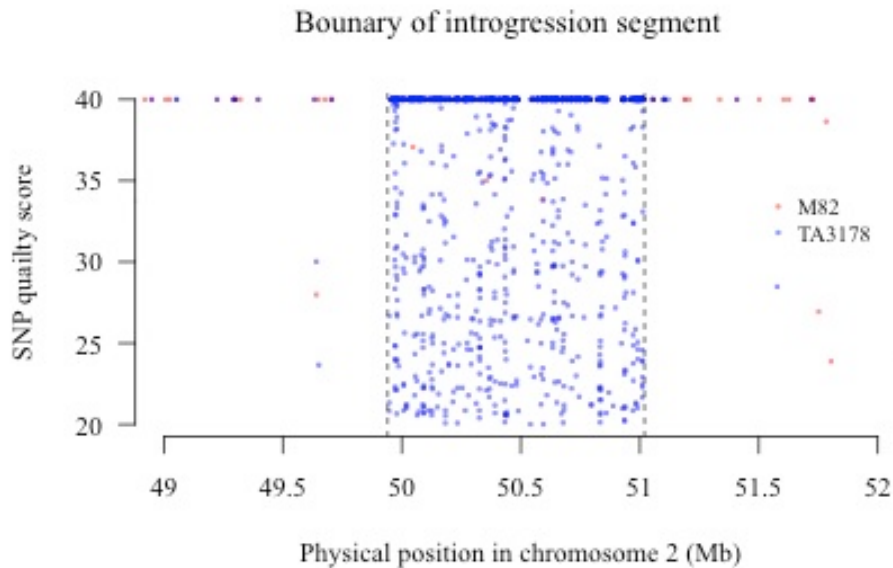
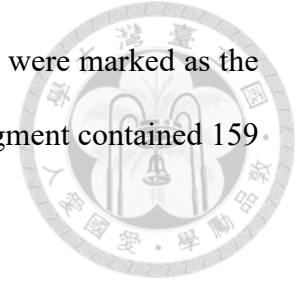


Figure 4.1 The SNPs in the introgression segment in M82 and TA3178.

In this introgression segment, M82 and TA3178 expressed 116 and 91 genes, respectively. The fold changes ranged from -1.97 to 4.27. A total of 134 genes contained SNPs (S\_Tab 4.1). Among them, Solyc02g087730.2, Solyc02g087900.2, Solyc02g088610.2, Solyc02g088620.2 and Solyc02g089050.2 contained high-effect SNPs causing different splicing patterns. Most of the other genes with low-effect SNPs would unlikely change their functions due to synonymous mutations (S\_Tab 4.1).

### 4.3.3 Only two DEGs in the introgression segment

According to the 99.9<sup>th</sup> percentile method, we obtained 324 DEGs whose fold changes ranged from -7.02 to 9.57 (S\_Tab 4.2). Meanwhile, based on DEseq analysis, a total of 140 DEGs were detected and their fold changes based on edgeR also ranged



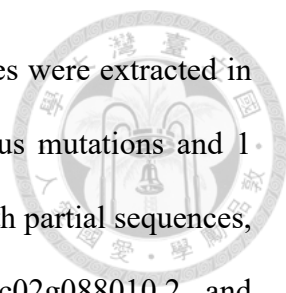
from -7.02 to 9.57, suggesting the most differentially expressed genes were detected by both methods (S\_Tab 4.3). Totally 90 DEGs were detected in both methods. It was interesting that the DEGs were distributed through the whole genome while the introgression segment was located on chromosome 2 (Table 4.2). In addition, among these 159 genes in the introgression segment, only Solyc02g087650.2 and Solyc02g088710.2 were identified as the DEGs (S\_Tab 4.1-3).

Table 4.2 The number of DEGs in each chromosome.

Chr.	99.9 <sup>th</sup> percentile method	DEseq analysis	Both
0	4	4	3
1	34	21	13
2	21	10	7
3	34	17	11
4	18	5	4
5	24	11	7
6	29	17	9
7	32	10	6
8	19	3	2
9	30	11	6
10	25	15	9
11	27	5	4
12	27	11	9
<b>Total</b>	<b>324</b>	<b>140</b>	<b>90</b>

#### 4.3.4 Three candidate genes of *stamen2.2* and *stamen2.3*

The result of the 18 candidate genes from cLED19A24 to CT9, including gene annotation, RPKM, fold changes and cDNA polymorphisms, were summarized in Table 4.3. First of all, Solyc02g087990.2, Solyc02g088020.1, Solyc02g088030.1, Solyc02g088050.1 and Solyc02g088060.1 expressed little in both lines (RPKM less than 1); therefore, they were removed from the candidates. For the other genes, we obtained 13 sequences in M82: 2 partial sequences and 11 full sequences without any



polymorphism when comparing to SL2.50. Meanwhile, 12 sequences were extracted in TA3178: 4 partial sequences, 7 full sequences with nonsynonymous mutations and 1 full sequence without any polymorphism. Because the five genes with partial sequences, Solyc02g087950.2, Solyc02g087970.1, Solyc02g087980.2, Solyc02g088010.2 and Solyc02g088080.1, all expressed higher in M82 and their RPKM were all greater than 1, we classified them as different expression. Solyc02g087930.2 displayed only different expression; Solyc02g087960.2 and Solyc02g088100.2 were detected only polymorphisms; the other five genes presented both differential expression and polymorphisms.

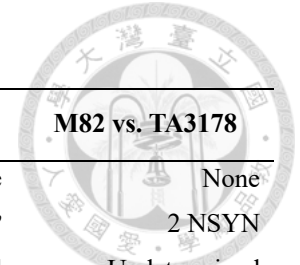
We emphasized on transcription factor since it played an important role on flower development. Solyc02g087960.2, Solyc02g087970.1, Solyc02g088030.1 and Solyc02g088070.2, which coded for MYB transcription factor, zinc finger protein, ring finger protein and Dof zinc finger protein, respectively, were transcription factors (Table 4.3). Solyc02g087960.2 expressed almost equally in both lines but was revealed five nonsynonymous mutations and one InDel. Solyc02g087970.1 expressed only in M82. Solyc02g088030.1 did not express in both lines and therefore was unlikely a candidate. Solyc02g088070.2 expressed slightly higher in TA3178 and was revealed 15 nonsynonymous mutations. Therefore, Solyc02g087960.2, Solyc02g087970.1 and Solyc02g088070.2 could be the candidates of *stamen2.2* and *stamen2.3*.

Table 4.3 The summary of the candidate genes from cLED19A24 to CT9

Gene ID	Gene Annotation	RPKM <sub>M82</sub>	RPKM <sub>TA3178</sub>	Fold Change	M82 vs. SL2.50	TA3178 vs. SL2.50	M82 vs. TA3178
Solyc02g087930.2	Ribosomal protein L34e	96.23	300.35	-1.63	None	None	None
Solyc02g087940.2	Unknown Protein	3.86	2.13	0.63	None	2 NSYN <sup>b</sup>	2 NSYN
Solyc02g087950.2	Unknown Protein	3.70	0.74	1.43	Undetermined	Undetermined	Undetermined
Solyc02g087960.2	MYB transcription factor	9.31	9.58	-0.04	None	5 NSYN & 1 InDel	5 NSYN & 1 InDel
Solyc02g087970.1	Zinc finger-homeodomain protein	4.36	0.38	1.96	Undetermined	-	Undetermined
Solyc02g087980.2	RecF/RecN/SMC protein	2.59	0.59	1.17	None	Undetermined	Undetermined
Solyc02g087990.2	Unknown Protein	0.07	0.05	0.03	- <sup>a</sup>	-	-
Solyc02g088000.2	Glycogen synthase	14.25	16.75	-0.22	None	11 NSYN	11 NSYN
Solyc02g088010.2	DCN1-like protein	4.59	0.97	1.50	None	Undetermined	Undetermined
Solyc02g088020.1	Unknown Protein	0.03	0.00	0.04	-	-	-
Solyc02g088030.1	RING finger protein, C3HC4 type	0.99	0.67	0.25	-	-	-
Solyc02g088040.1	Ribosomal protein L34e	7.94	14.45	-0.79	None	1 NSYN	1 NSYN
Solyc02g088050.1	ATPase, P-type	0.06	0.04	0.03	-	-	-
Solyc02g088060.1	ATPase, P-type	0.12	0.01	0.15	-	-	-
Solyc02g088070.2	Dof zinc finger protein	5.93	6.91	-0.19	None	15 NSYN	15 NSYN
Solyc02g088080.1	Unknown Protein	7.15	1.31	1.82	None	Undetermined	Undetermined
Solyc02g088090.1	Calmodulin-like protein	15.22	27.95	-0.84	None	3 NSYN	3 NSYN
Solyc02g088100.2	Pollen allergen/expansin	29.27	28.93	-0.02	None	2 NSYN	2 NSYN

<sup>a</sup>: - indicated no expression.

<sup>b</sup>: NSYN indicated nonsynonymous mutations.





## 4.4 Discussion

### 4.4.1 M82 presented more SNPs than TA3178 due to its deeper sequencing

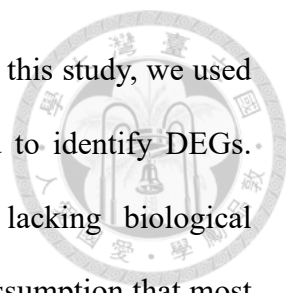
Theoretically TA3178 should have more polymorphisms than M82 due to its introgression segment of wild tomato. However, on the contrary, M82 presented more SNPs than TA3178 in this study (Table 4.1). The SNP density of chromosome 2 in TA3178 was actually the highest but those of the other chromosomes were higher in M82 except for chromosome 7 (Table 4.4). The higher SNP density in M82 and more uniquely expressed genes in M82 suggested more reads can reveal more polymorphisms and also more expressed genes (Table 4.1; Table 4.4).

Table 4.4 The expressed genes and the SNP density through each chromosome

Chr.	M82			TA3178		
	Gene	Gene with SNP	Density	Gene	Gene with SNP	Density
0	135	55	2.93	118	42	2.71
1	2,488	397	2.36	2,065	262	2.10
2	2,065	362	2.91	1,707	335	6.48
3	2,052	439	3.18	1,716	310	2.65
4	1,557	519	4.26	1,274	364	3.38
5	1,251	454	4.67	1,026	348	3.53
6	1,715	214	1.96	1,438	133	1.89
7	1,416	234	1.65	1,135	132	1.67
8	1,375	215	2.00	1,137	132	1.70
9	1,374	211	2.80	1,132	161	2.77
10	1,287	193	2.19	1,088	141	1.83
11	1,267	328	2.94	987	216	2.40
12	1,243	175	1.99	971	108	1.69
<b>Total</b>	<b>19,225</b>	<b>3,796</b>	<b>3.01</b>	<b>15,794</b>	<b>2,684</b>	<b>3.06</b>

### 4.4.2 Lacking biological replications may underestimate DEGs

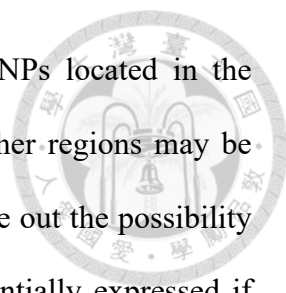
The most important factor in RNA-seq is the biological replication because the DEGs between samples can result from the comparison with the variation of the genes



within samples. Since the biological replication was not available in this study, we used the 99.9<sup>th</sup> percentile of the fold change distribution as a threshold to identify DEGs. Meanwhile, DEseq provides a test under a circumstance of lacking biological replications by treating different samples as replications under the assumption that most genes are not DEGs. In that case, the variance between samples would be greater than that between real replications, consequently underestimating DEGs (Anders et al., 2010). In our study, the 99.9<sup>th</sup> percentile method identified 324 DEGs while DEseq analysis obtained only 140 DEGs. In addition, the DEseq seemed to identify DEGs that presented higher fold changes (S\_Fig 4.2). These two observations supported the underestimation of DEGs in DEseq without real biological replications. One thing interesting was that *style2.1* (Solyc02g087860.2), which has been proved to regulate the style length via a different expression level, displayed -1.10 fold changes and it was not identified as the DEG in both differential expression analyses (S\_Tab 4.1). Despite high sequencing reads in this study, the 99.9<sup>th</sup> percentile method did not detect the known functional gene. This suggested that without biological replications, the expression level could not be compared properly even in a deep sequencing study, potentially resulting in underestimating DEGs as well. Therefore, the reliability of the differential expression analysis should be examined through biological replications.

#### **4.4.2 Transcription profiles and polymorphisms in the introgression segment**

In the introgression segment, two DEGs, Solyc02g087650.2 and Solyc02g088710.2, and five genes with high-effect SNPs, Solyc02g087730.2, Solyc02g087900.2, Solyc02g088610.2, Solyc02g088620.2 and Solyc02g089050.2, may affect or regulate the phenotypic difference between M82 and TA3178 (S\_Tab 4.1). It was very

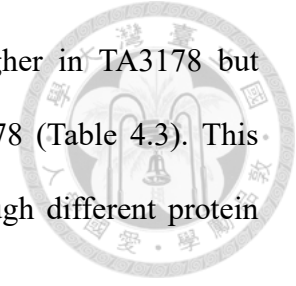


interesting that only a small number of DEGs and high-effect SNPs located in the introgression region (Table 4.2), suggesting that other DEGs in other regions may be regulated through complex mechanisms. However, we could not rule out the possibility of underestimation of DEGs because totally 45 genes were differentially expressed if the threshold was the fold change of *style2.1* (1.10).

#### 4.4.4 Narrow down the candidate genes of *stamen2.2* and *stamen2.3*

In the interval from cLED19A24 to CT9, 18 candidates were narrowed to 3 candidates via the transcription profiles and cDNA polymorphisms. We surveyed the related studies of these transcription factors. Solyc02g087960.2 belongs to R2R3-MYB transcription factor 94 and is involved in sequence-specific DNA binding. Previous studies have showed that it could be regulated by auxin response factor (SIARF3) or by DELLA-dependent GA mechanisms in tomato (Livne et al., 2015; Zhang et al., 2015). It might contribute to the stamen length in the crosstalk of auxin and GA since GA could promote cell elongation in a manner of degradation of DELLA, which inhibited ARF and then modulated the expression of ARF-targeted genes (Oh et al., 2014). In addition, Solyc02g087960.2 might contribute to the stamen length by the change of protein function because it expressed almost equally in both lines (Table 4.3). Solyc02g087970.1 is a mini zinc finger protein and involves in multiple phytohormone regulations (Hu & Ma, 2006). An overexpression line of its homolog in *Arabidopsis thaliana* showed the inhibition of cell elongation and shortened the stamen via the auxin, GA and brassinosteroid signaling (Hu & Ma, 2006). Since Solyc02g087970.1 only expressed in M82, it might potentially shorten the stamen length through inhibiting cell elongation. Solyc02g088070.2 is a Dof zinc finger protein but did not participate in

stamen development to our knowledge. It expressed slightly higher in TA3178 but displayed 15 nonsynonymous mutations between M82 and TA3178 (Table 4.3). This implied Solyc02g088070.2 might regulate the stamen length through different protein functions.



## 4.5 Reference

- Anders, S., Huber, W., Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., ... Salzberg, S. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(R106). <https://doi.org/10.1186/gb-2010-11-10-r106>
- Chen, K. Y., & Tanksley, S. D. (2004). High-resolution mapping and functional analysis of se2.1: A major stigma exertion quantitative trait locus associated with the evolution from allogamy to autogamy in the genus *lycopersicon*. *Genetics*. <https://doi.org/10.1534/genetics.103.022558>
- Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., ... Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*(2), 80–92. <https://doi.org/10.4161/fly.19695>
- Hu, W., & Ma, H. (2006). Characterization of a novel putative zinc finger gene MIF1: Involvement in multiple hormonal regulation of *Arabidopsis* development. *Plant Journal*. <https://doi.org/10.1111/j.1365-313X.2005.02626.x>
- Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*. <https://doi.org/10.1093/molbev/msw054>
- Liao, Y., Smyth, G. K., & Shi, W. (2013). The Subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*.

<https://doi.org/10.1093/nar/gkt214>

Livne, S., Lor, V. S., Nir, I., Eliaz, N., Aharoni, A., Olszewski, N. E., ... Weiss, D. (2015). Uncovering DELLA-Independent Gibberellin Responses by Characterizing New Tomato *procera* Mutants. *The Plant Cell*.

<https://doi.org/10.1105/tpc.114.132795>

Morgan, M., Anders, S., Lawrence, M., Aboyoun, P., Pagès, H., & Gentleman, R. (2009). ShortRead: A bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*.

<https://doi.org/10.1093/bioinformatics/btp450>

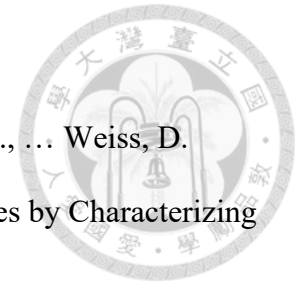
Oh, E., Zhu, J. Y., Bai, M. Y., Arenhart, R. A. ugusto, Sun, Y., & Wang, Z. Y. (2014). Cell elongation is regulated through a central circuit of interacting transcription factors in the *Arabidopsis* hypocotyl. *ELife*, 3, 1–19.

<https://doi.org/10.7554/eLife.03031>

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btp616>

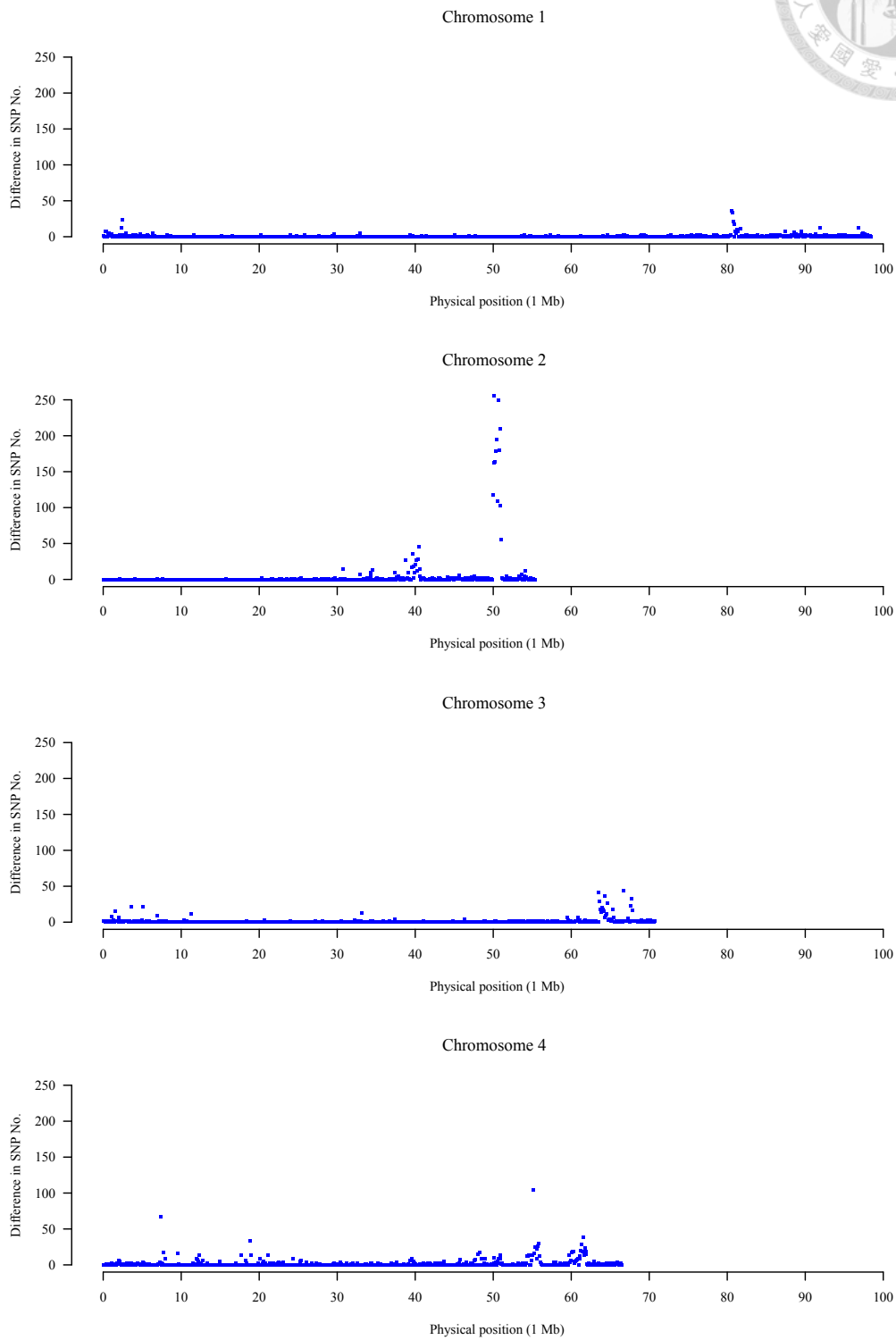
Zhang, X., Yan, F., Tang, Y., Yuan, Y., Deng, W., & Li, Z. (2015). Auxin Response Gene *SLARF3* Plays Multiple Roles in Tomato Development and is Involved in the Formation of Epidermal Cells and Trichomes. *Plant and Cell Physiology*.

<https://doi.org/10.1093/pcp/pcv136>

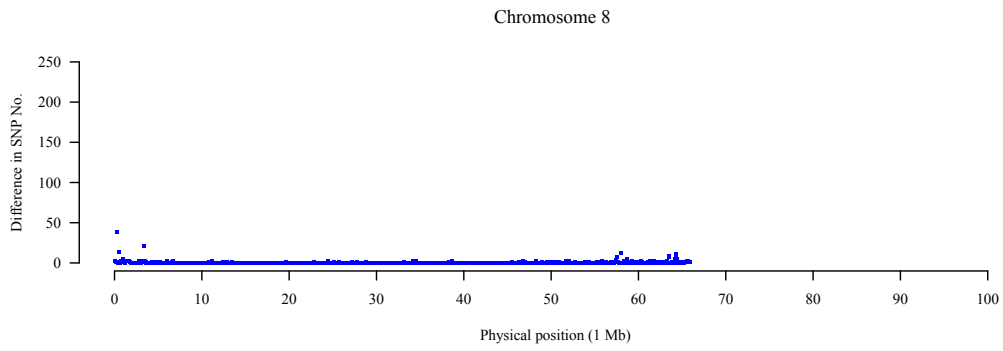
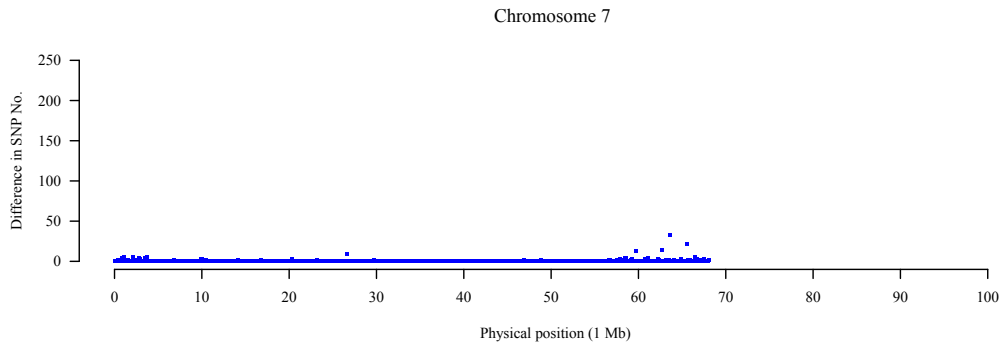
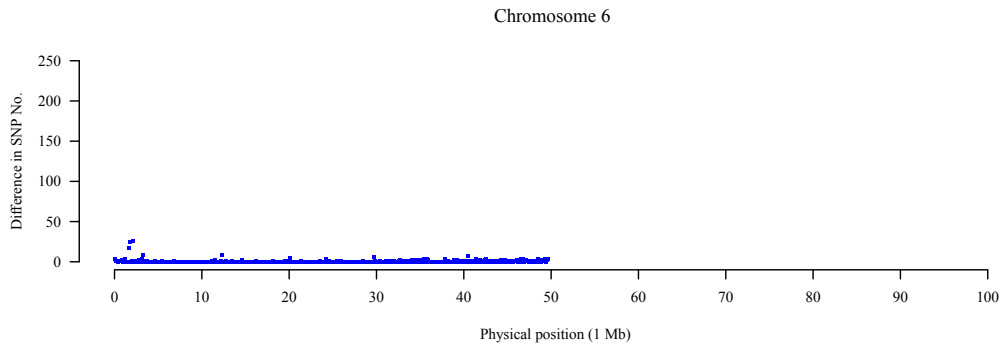
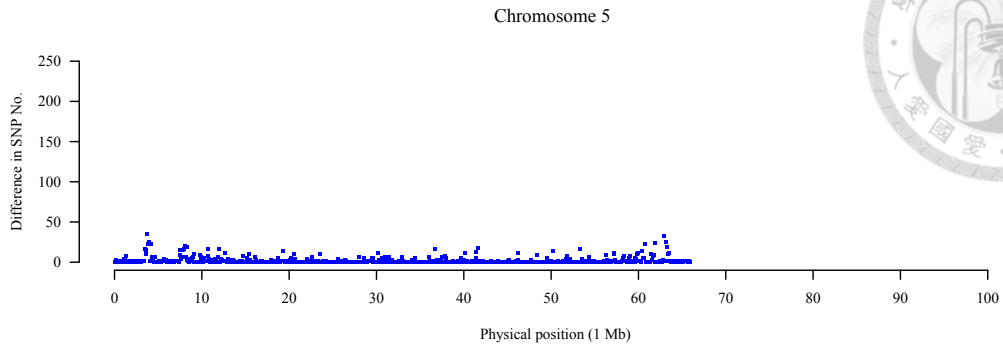




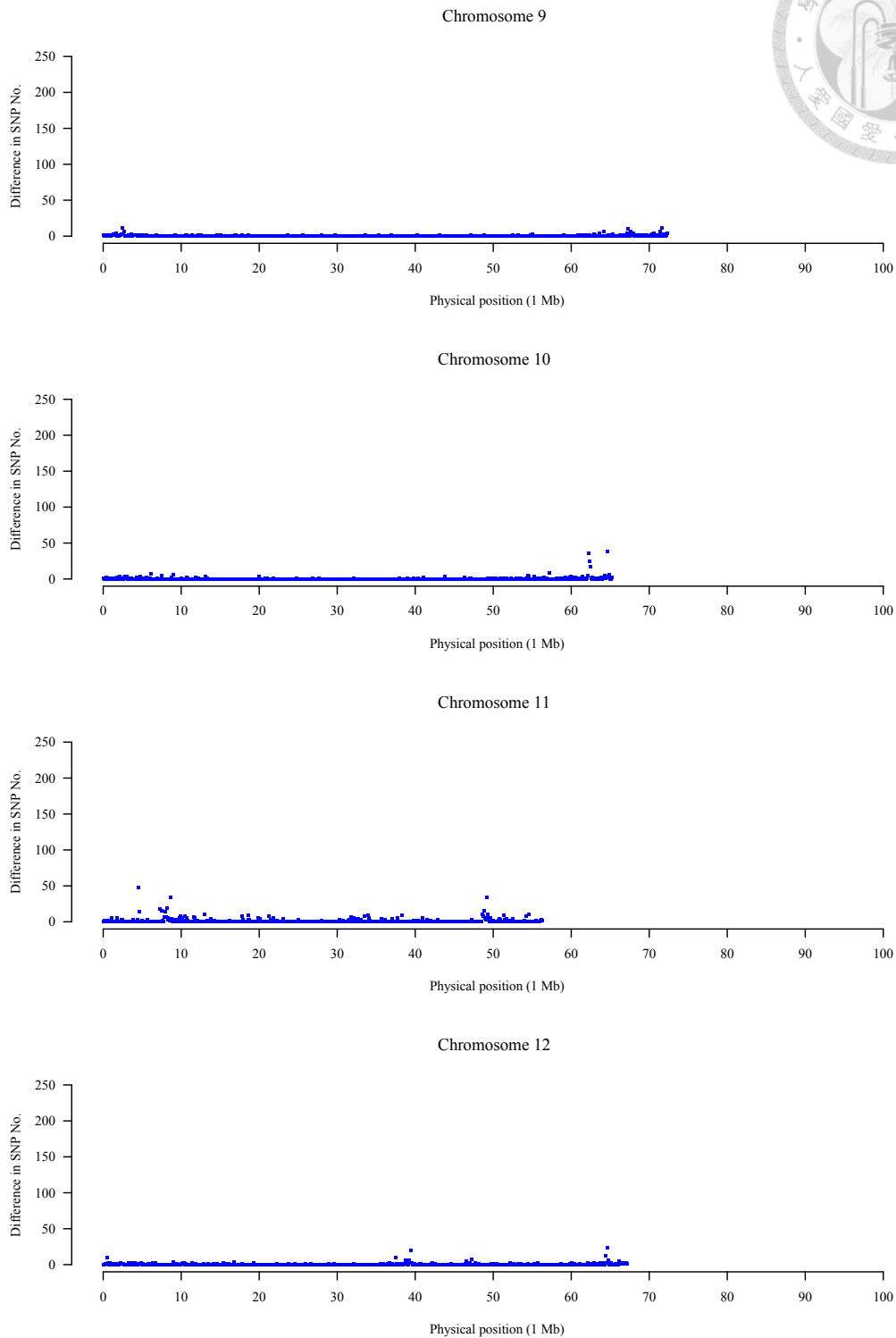
## 4.6 Supplementary data



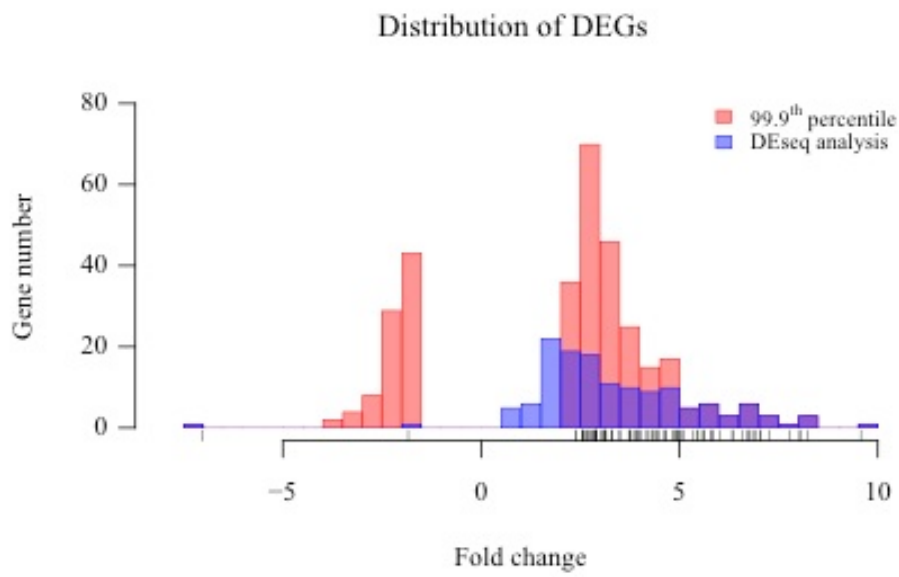
S\_Fig 4.1 (page 1/3)



S\_Fig 4.1 (page 2/3)



S\_Fig 4.1 The difference of SNP number between M82 and TA3178 through each chromosome.

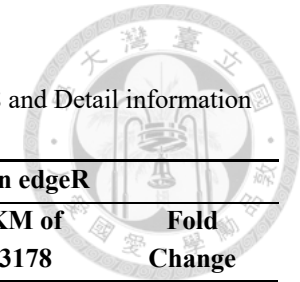


S\_Fig 4.2 The distribution of DEGs.

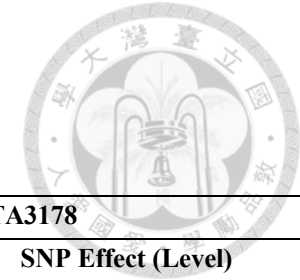
Red indicated the fold changes of DEGs from the 99.9<sup>th</sup> percentile method and blue indicated those from the DEseq analysis. The black ticks above the x-axis showed the common DEGs in these two methods.

S\_Tab 4.1 The detail information of 159 genes in the introgression segment.

This supplementary material contains Basic information, Result in edgeR, Result in DEseq, SNP number and effects in M82 as well as TA3178 and Detail information of 159 genes. We listed 20 genes for readers to glimpse the data and the full table is provided on the following link <https://goo.gl/8hUcy3>.

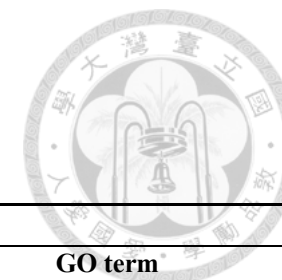


ID	Basic information					Result in edgeR		
	Start Position	End Position	Gene Width	Counts of M82	Counts of TA3178	RPKM of M82	RPKM of TA3178	Fold Change
Solyc02g087550.2	49,956,733	49,960,809	4,077	8,839	8,179	10.1395	11.6098	-0.1789
Solyc02g087560.1	49,964,967	49,966,964	1,998	3,017	1,223	7.0621	3.5424	0.8277
Solyc02g087570.1	49,967,581	49,967,952	372	28	30	0.3520	0.4667	-0.1175
Solyc02g087580.2	49,970,650	49,974,563	3,914	1,203	698	1.4375	1.0320	0.2625
Solyc02g087590.1	49,974,938	49,975,570	633	3	1	0.0222	0.0091	0.0185
Solyc02g087600.2	49,976,320	49,978,863	2,544	4,680	484	8.6036	1.1010	2.1925
Solyc02g087610.1	49,979,769	49,981,307	1,539	688	682	2.0908	2.5645	-0.2058
Solyc02g087620.2	49,985,370	50,005,010	19,641	13,650	2,453	3.2503	0.7228	1.3028
Solyc02g087630.2	50,005,692	50,006,645	954	9,142	19,239	44.8174	116.7073	-1.3612
Solyc02g087640.2	50,007,440	50,012,157	4,718	2,234	857	2.2145	1.0512	0.6481
Solyc02g087650.2	50,016,855	50,017,612	758	2,564	8,528	15.8199	65.1091	-1.9747
Solyc02g087660.2	50,018,064	50,020,570	2,507	541	320	1.0092	0.7387	0.2087
Solyc02g087670.2	50,023,499	50,025,492	1,994	1,416	87	3.3212	0.2525	1.7866
Solyc02g087680.1	50,026,966	50,027,599	634	6	1	0.0443	0.0091	0.0494
Solyc02g087690.1	50,027,724	50,027,930	207	3	0	0.0678	0.0000	0.0946
Solyc02g087700.2	50,027,852	50,028,463	612	3	3	0.0229	0.0284	-0.0077
Solyc02g087710.2	50,029,205	50,036,081	6,877	45,239	15,537	30.7658	13.0747	1.1744
Solyc02g087720.1	50,037,542	50,039,578	2,037	5,292	3,836	12.1502	10.8981	0.1443
Solyc02g087730.2	50,041,595	50,050,438	8,844	6,941	2,745	3.6705	1.7962	0.7401
Solyc02g087740.2	50,051,845	50,054,725	2,881	1,791	1,539	2.9074	3.0914	-0.0664



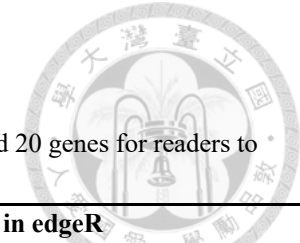
S\_Tab 4.1 (Continued)

ID	Result in DEseq		M82		TA3178	
	p value	p adjusted	SNP number	SNP Effect (Level)	SNP number	SNP Effect (Level)
Solyc02g087550.2	0.5111	1	0	-	21	Synonymous variant (low)
Solyc02g087560.1	0.8379	1	0	-	15	Synonymous variant (low)
Solyc02g087570.1	0.6275	1	0	-	0	-
Solyc02g087580.2	0.8678	1	0	-	35	Synonymous variant (low)
Solyc02g087590.1	1	1	0	-	15	-
Solyc02g087600.2	0.1104	1	0	-	11	Synonymous variant (low)
Solyc02g087610.1	0.4794	1	0	-	15	Synonymous variant (low)
Solyc02g087620.2	0.2904	1	1	-	35	Synonymous variant (low)
Solyc02g087630.2	0.1386	1	0	-	4	Splice region variant (low)
Solyc02g087640.2	0.7942	1	0	-	17	Splice region variant (low)
Solyc02g087650.2	0.0562	1	0	-	2	-
Solyc02g087660.2	0.8570	1	2	-	10	Synonymous variant (low)
Solyc02g087670.2	0.0435	1	0	-	0	-
Solyc02g087680.1	0.9341	1	0	-	3	-
Solyc02g087690.1	0.9646	1	0	-	0	-
Solyc02g087700.2	0.9453	1	0	-	0	-
Solyc02g087710.2	0.7017	1	1	Splice region variant & intron variant (low)	29	Splice region variant & synonymous variant (low)
Solyc02g087720.1	0.6870	1	0	-	16	Synonymous variant (low)
Solyc02g087730.2	0.8162	1	1	Splice donor variant & intron variant (high)	23	Synonymous variant (low)
Solyc02g087740.2	0.5657	1	0	-	32	Synonymous variant (low)



S\_Tab 4.1 (Continued)

Detail information		
ID	Gene Annotation	GO term
Solyc02g087550.2	Beta-1,3-galactosyl-O-glycosyl-glycoprotein beta-1,6-N-acetylglucosaminyltransferase 4	GO:0016757
Solyc02g087560.1	Pentatricopeptide repeat-containing protein	-
Solyc02g087570.1	LOB domain protein 4	GO:0005515
Solyc02g087580.2	Unknown Protein	-
Solyc02g087590.1	Serine/threonine protein kinase	GO:0005515; GO:0016301
Solyc02g087600.2	DTW domain-containing protein	-
Solyc02g087610.1	Pentatricopeptide repeat-containing protein	GO:0004519; GO:0008116
Solyc02g087620.2	Inositol hexakisphosphate and diphosphoinositol-pentakisphosphate kinase 2	GO:0033187
Solyc02g087630.2	Thioredoxin H	GO:0045454
Solyc02g087640.2	Protein midA homolog, mitochondrial	GO:0005515; GO:0008270
Solyc02g087650.2	Unknown Protein	-
Solyc02g087660.2	Auxin efflux carrier protein	GO:0009672
Solyc02g087670.2	Pectate lyase family protein	GO:0016829
Solyc02g087680.1	FACT complex subunit SSRP1	GO:0005634
Solyc02g087690.1	Unknown Protein	-
Solyc02g087700.2	FACT complex subunit SSRP1	GO:0005634
Solyc02g087710.2	FACT complex subunit SSRP1	GO:0042393
Solyc02g087720.1	At3g28720-like protein	GO:0032259
Solyc02g087730.2	Katanin p80 WD40-containing subunit B1	-
Solyc02g087740.2	Cupin RmlC-type	GO:0055114

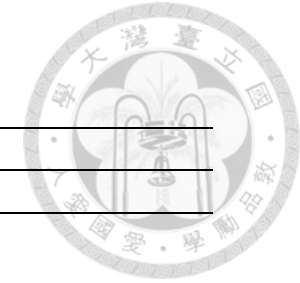


S\_Tab 4.2 The detail information of the DEGs based on the 99.9<sup>th</sup> percentile method.

This supplementary material contains Basic information, Result in edgeR, Result in DEseq and Detail information of 324 DEG genes. We listed 20 genes for readers to glimpse the data and the full table is provided on the following link <https://goo.gl/8hUcy3>.

ID	Basic information					Result in edgeR		
	Start Position	End Position	Gene Width	Counts of M82	Counts of TA3178	RPKM of M82	RPKM of TA3178	Fold Change
Solyc00g009070.1	8,775,151	8,775,537	387	1,658	5	20.0367	0.0748	4.2908
Solyc00g171710.1	18,049,090	18,050,010	921	4,902	4	24.8924	0.0251	4.6586
Solyc00g228260.1	19,569,553	19,569,864	312	22	190	0.3298	3.5242	-1.7665
Solyc00g257110.2	20,272,521	20,276,002	3,482	3,659	5	4.9146	0.0083	2.5523
Solyc01g005510.2	346,845	349,013	2,169	7,028	5	15.1539	0.0133	3.9947
Solyc01g006070.2	739,814	740,772	959	3,484	119	16.9908	0.7181	3.3884
Solyc01g006390.2	1,017,309	1,018,640	1,332	12,158	37,114	42.6886	161.2491	-1.8929
Solyc01g010390.2	5,254,174	5,260,994	6,821	23,427	546	16.0628	0.4632	3.5436
Solyc01g010530.1	5,548,702	5,550,171	1,470	5,959	67	18.9587	0.2638	3.9812
Solyc01g056310.2	53,293,891	53,295,912	2,022	49,001	34	113.3385	0.0973	6.7032
Solyc01g056360.2	53,737,447	53,737,872	426	1,898	0	20.8372	0.0000	4.4487
Solyc01g066620.2	74,760,560	74,766,184	5,625	13,544	70	11.2610	0.0720	3.5157
Solyc01g066810.2	74,989,661	74,991,710	2,050	2,181	9	4.9757	0.0254	2.5429
Solyc01g067350.2	75,812,003	75,813,413	1,411	9,193	69	30.4708	0.2830	4.6164
Solyc01g068080.2	77,121,171	77,124,462	3,292	51,579	4,156	73.2768	7.3060	3.1607
Solyc01g068110.2	77,233,578	77,234,566	989	1,386	1	6.5542	0.0059	2.9089
Solyc01g068120.2	77,234,467	77,235,354	888	1,233	1	6.4939	0.0065	2.8963
Solyc01g079890.2	79,070,877	79,072,892	2,016	38,155	196	88.5145	0.5626	5.8401
Solyc01g090350.2	84,055,544	84,056,304	761	61,031	95	375.0761	0.7224	7.7704
Solyc01g090600.2	84,269,539	84,271,141	1,603	66,309	3,689	193.4606	13.3180	3.7636





S\_Tab 4.2 (Continued)

ID	Result in DEseq		Detail information	
	p value	p adjusted	Gene Annotation	GO term
Solyc00g009070.1	0.0001	0.0174	Unknown Protein	-
Solyc00g171710.1	0.0000	0.0019	F-box associated type 1	-
Solyc00g228260.1	0.0181	1.0000	Unknown Protein	-
Solyc00g257110.2	0.0000	0.0035	ATPase, P type	GO:0008553
Solyc01g005510.2	0.0000	0.0013	Multicopper oxidase, type 3	GO:0055114
Solyc01g006070.2	0.0101	0.8811	Protein of unknown function DUF716	-
Solyc01g006390.2	0.0665	1.0000	Cysteine-rich extensin-like protein-4	-
Solyc01g010390.2	0.0034	0.3979	Glycoside hydrolase	GO:0005975
Solyc01g010530.1	0.0006	0.1025	Sugar/inositol transporter	GO:0016020; GO:0016021
Solyc01g056310.2	0.0000	0.0007	Multicopper oxidase, type 3	GO:0055114
Solyc01g056360.2	0.0000	0.0015	Unknown Protein	-
Solyc01g066620.2	0.0001	0.0206	3-hydroxyacyl-CoA dehydrogenase	GO:0006631; GO:0050662; GO:0008152; GO:0016491
Solyc01g066810.2	0.0001	0.0231	Universal stress protein	GO:0006950
Solyc01g067350.2	0.0002	0.0438	UDP-glucuronosyl/UDP-glucosyltransferase	GO:0016757
Solyc01g068080.2	0.0637	1.0000	NAD(P)-binding domain	GO:0044237; GO:0008152
Solyc01g068110.2	0.0000	0.0050	Unknown Protein	-
Solyc01g068120.2	0.0000	0.0066	Pectinesterase	GO:0005618
Solyc01g079890.2	0.0001	0.0188	Aquaporin	GO:0016020
Solyc01g090350.2	0.0000	0.0025	Plant lipid transfer protein and hydrophobic protein	GO:0006869
Solyc01g090600.2	0.0280	1.0000	Chalcone synthase 3 protein	GO:0005576; GO:0008415; GO:0008152; GO:0009058

S\_Tab 4.3 The detail information of the DEGs based on the DEseq analysis.

This supplementary material contains Basic information, Result in edgeR, Result in DEseq and Detail information of 140 DEG genes. We listed 20 genes for readers to glimpse the data and the full table is provided on the following link <https://goo.gl/8hUcy3>.

ID	Basic information					Result in edgeR		
	Start Position	End Position	Gene Width	Counts of M82	Counts of TA3178	RPKM of M82	RPKM of TA3178	Fold Change
Solyc00g009070.1	8,775,151	8,775,537	387	1,658	5	20.0367	0.0748	4.2908
Solyc00g058900.1	14,062,470	14,064,603	2,134	1,517	4	3.3246	0.0108	2.0970
Solyc00g171710.1	18,049,090	18,050,010	921	4,902	4	24.8924	0.0251	4.6586
Solyc00g257110.2	20,272,521	20,276,002	3,482	3,659	5	4.9146	0.0083	2.5523
Solyc01g005510.2	346,845	349,013	2,169	7,028	5	15.1539	0.0133	3.9947
Solyc01g005650.1	455,379	456,988	1,610	707	0	2.0537	0.0000	1.6106
Solyc01g008240.2	2,369,175	2,373,387	4,213	3,480	4	3.8631	0.0055	2.2740
Solyc01g010500.1	5,446,120	5,450,141	4,022	2,972	2	3.4559	0.0029	2.1516
Solyc01g011050.2	7,014,098	7,016,194	2,097	386	0	0.8609	0.0000	0.8960
Solyc01g056310.2	53,293,891	53,295,912	2,022	49,001	34	113.3385	0.0973	6.7032
Solyc01g056360.2	53,737,447	53,737,872	426	1,898	0	20.8372	0.0000	4.4487
Solyc01g066620.2	74,760,560	74,766,184	5,625	13,544	70	11.2610	0.0720	3.5157
Solyc01g066810.2	74,989,661	74,991,710	2,050	2,181	9	4.9757	0.0254	2.5429
Solyc01g067350.2	75,812,003	75,813,413	1,411	9,193	69	30.4708	0.2830	4.6164
Solyc01g068110.2	77,233,578	77,234,566	989	1,386	1	6.5542	0.0059	2.9089
Solyc01g068120.2	77,234,467	77,235,354	888	1,233	1	6.4939	0.0065	2.8963
Solyc01g079890.2	79,070,877	79,072,892	2,016	38,155	196	88.5145	0.5626	5.8401
Solyc01g087280.1	82,215,602	82,219,911	4,310	1,321	2	1.4334	0.0027	1.2791
Solyc01g090350.2	84,055,544	84,056,304	761	61,031	95	375.0761	0.7224	7.7704
Solyc01g094910.2	86,314,679	86,317,951	3,273	2,513	15	3.5909	0.0265	2.1610

S\_Tab 4.3 (Continued)

ID	Result in DEseq		Detail information	
	p value	p adjusted	Gene Annotation	GO term
Solyc00g009070.1	0.0001	0.0174	Unknown Protein	-
Solyc00g058900.1	0.0000	0.0157	GDSL esterase/lipase At2g31540	GO:0004091
Solyc00g171710.1	0.0000	0.0019	Unknown Protein	-
Solyc00g257110.2	0.0000	0.0035	H-ATPase	GO:0008553
Solyc01g005510.2	0.0000	0.0013	Laccase-2	GO:0055114
Solyc01g005650.1	0.0000	0.0106	Ariadne-like ubiquitin ligase	GO:0004842
Solyc01g008240.2	0.0000	0.0031	Solute carrier family 2, facilitated glucose transporter member	GO:0016020; GO:0016021
Solyc01g010500.1	0.0000	0.0023	Ein3-binding f-box protein 3	-
Solyc01g011050.2	0.0002	0.0492	LRR receptor-like serine/threonine-protein kinase, RLP	GO:0004675
Solyc01g056310.2	0.0000	0.0007	Laccase-2	GO:0055114
Solyc01g056360.2	0.0000	0.0015	Unknown Protein	-
Solyc01g066620.2	0.0001	0.0206	Fatty acid oxidation complex subunit alpha	GO:0006631; GO:0050662; GO:0008152; GO:0016491
Solyc01g066810.2	0.0001	0.0231	Universal stress protein	GO:0006950
Solyc01g067350.2	0.0002	0.0438	UDP-glucosyltransferase	GO:0016757
Solyc01g068110.2	0.0000	0.0050	Unknown Protein	-
Solyc01g068120.2	0.0000	0.0066	Pectinesterase	GO:0005618
Solyc01g079890.2	0.0001	0.0188	Aquaporin	GO:0016020
Solyc01g087280.1	0.0000	0.0103	Polygalacturonase A	GO:0004650
Solyc01g090350.2	0.0000	0.0025	Non-specific lipid-transfer protein	GO:0006869
Solyc01g094910.2	0.0002	0.0401	Ferric reductase oxidase	GO:0000293

