UNIVERSITAT AUTÒNOMA DE BARCELONA

Facultat de Biociències

“Genetic architecture of agronomic traits in peach [*Prunus persica* (L.) Batsch]: subacid, flat shape and nectarine”

Elena López Girona

Barcelona, noviembre 2014

PhD Thesis

"Genetic architecture of agronomic traits in peach [*Prunus persica* (L.) Batsch]: subacid, flat shape and nectarine"

presented to

Universitat Autònoma de Barcelona

Facultat de Biociències

Dept. Biologia Animal, Biologia Vegetal i Ecologia

Estudis de Doctorat en Biologia i Biotecnologia Vegetal

by

Elena López Girona for the degree of Doctor of Biology and Plant Biotechnology by Universitat Autònoma de Barcelona (UAB)

Thesis director	Tutor	PhD candidate
Dra María José Aranzana Civit	Dra María Carmen Espunya Prat	Elena López Girona

Barcelona, noviembre 2014

This thesis has been entirely carried out in the department of Plant Genetics of the Centre de Recerca en Agrigenòmica (CRAG) CSIC-IRTA-UAB-UB, Barcelona.

A mi primo,

por compartir mutua admiración por la naturaleza.

"A las aladas almas de las rosas
del almendro de nata te requiero,
que tenemos que hablar de muchas cosas,
compañero del alma, compañero."

La Elegía a Ramón Sijé, Miguel Hernández (1936)

ACKNOWLEDGEMENTS

Llegó el momento de agradecer a todas las personas con las que he compartido estos maravillosos cuatro años y que de un modo u otro han hecho posible que hoy esté escribiendo estas líneas. Siempre tendré un especial recuerdo de mis días en Catalunya gracias a vosotros!

En primer lugar me gustaría agradecer a mi directora de tesis; Txosse, por darme la oportunidad de realizar la tesis con ella. Gracias por tu empeño en hacerme ver las cosas desde el lado positivo y por haberme enseñado que con paciencia, trabajo y esperanza "todo se puede".

Gracias a Pep Casacuberta, Concepción Royo y Santiago Vilanova por aceptar formar parte del tribunal de mi tesis. Gracias a Maria Carmen Martínez Gómez por ser mi tutora de tesis en la universidad.

I will be always in debt to Daniel Sargent who believed in me as a good candidate for his research assistant position at EMR. He provided me the best thing anyone could give me ever: 'encouragement'. This source of willpower allowed and still allows me to reach and achieve many important things (I would need a whole thesis to list all of them) despite this thesis, which would never has happened without your training, help, advice, friendship and especially because of our hard work during the 2009 year! I will always be grateful to all the people I worked there with and all the people I met during that worthy period of my life.

Gracias a los "caseteros" de Cabrils por todos los buenos momentos pasados bajo ese techo. A los inquilinos aventajados cuando llegue: Ali, Eudald, Xen, Julio, Claudio, Álvaro y Juan. En especial a los tres "itos": Claudito, Juanito y Julito por enseñarme el verdadero espíritu becario! Gracias por vuestros consejos, ánimos y buen humor. Gracias especialmente a 'La Puri', nuestra mami particular! los miércoles sin ti créeme no son los mismos!

Y llegó el párrafo más duro de escribir de toda la tesis. Gracias a todos mis compañeros de despacho del CRAG. En el bando vegetal muy especialmente Cèlia, saps que vals molt, has sigut un gran suport durant tot aquest temps. Un luxe haver-te tingut de companya d'escriptori en els moments de rialles i en els difícils també. Els teus detalls amb tots, et fan especial, no canviïs mai. A María Urrutia y Pablo Rivers por vuestra ayuda, nuestros piques y por todo lo que nos hemos reído!

Gracias también al bando cerdito. A Yuli y Jordi Corominas por ser una gran fuente de motivación y buena compañia durante las horas extra y los findes infinitos de despacho. A Anita, siempre dispuesta a ayudar y a reivindicar la valía femenina del despachito! A Rayner por apoyarme cuando mi ausencia o presencia en algunos eventos no era evidenciada. A Sebas por su gran humor, por sus grandes consejos y su apoyo a la figura del becario. Gràcie mille a Erica e Ivan por compartir muchos buenos ratos de sobremesa y sofá. Erica, ese terremoto de positividad y determinismo! A Sarai, tan simpática como sus rizos. A Betlem porque es la caña y a Ana Mercadé por su amabilidad y preocupación.

No me olvido de mis compis del zulo; José Manuel, siempre tan educado y atento (aunque se olvidara de mencionarme en sus agradecimientos, sé que no era su intención). A Pableras, siempre tan *tuanis* y sus historias para no dormir... Suerte en esta etapa al amable y nuevo prunero, Octavio. A la guapa Vero, por su simpatía.

Guardo un especial recuerdo de mi Andrea, creo que no me equivoco cuando digo que eres una de las personas más especiales que he conocido. Gracias por tu gran amistad, generosidad y por la paz que transmites. Aún me emociono cuando recuerdo tu despedida. Gracias también a Montse Saladie, no sé por dónde empezar, porque eres increíble. Gracias por toda tu ayuda profesional y personal. Gracias por mantener siempre abierta la puerta de tu despachito! A Gisela, porque ser tan motivadora y activa. Pero sobre todo por nuestras quedadas mesoneras, cómo las echo de menos… Gracias Raúl por tu gran paciencia conmigo, tu amable disponibilidad siempre, por todo lo *Bioinformaticamente* aprendido y tu buen humor. Gracias a Marc por sus 'trending topics', que siempre nos entretenían durante la comida o el café y por toda tu ayuda y apoyo. Gracias Walter y Jose Ramón por vuestra disponibilidad y generosidad. Gracias Carme, siempre es bueno tenerte cerca, eres calma y serenidad! Gracias Micheletti por todo lo vivido en 'Villa Sebastiana' y por las travesías por 'Collserola', tu inmensa paciencia conmigo, tus grandes consejos-lecciones y por tus insuperables tuppers!!! Gracias a Jason y Elena de Castro por confiar en nosotros como padres adoptivos de Tessa. Sois una pareja ejemplo para nosotros. Grandes personas.

Gracias a Ibo, mi prunero de referencia. Gracias por enseñarme que: "Life is not easy my friend" y que "la paciència és la mare de la Ciència". Ha sido un lujo trabajar y compartir poyata contigo. Siempre serás un 'viejoven' muy especial para mí. Gracias también a Aurora, la mejor anfitriona. Gracias mi otro prunero de referencia, Werner, por todos los consejos y aportaciones. Gracias también a los doctores meloneros: Juan, Michael, Diego, Pani, Jordi

Morata y a Ana Giner por vuestros consejos sabios y todos los jueves! Gracias también a Marta Pujol por estar siempre disponible y por mantener el órden en el labo. A Montse Martín por aportar una visión diferente a los experimentos pruneros y por su tarea al frente de la cafetera!

Gracias Amparo por confiar en mí como becaria para el IRTA, sin tu consideracion creo que ahora no estaria escribiendo estas lineas. Una pena no haber trabajado para ti, pero un orgullo haber compartido estos años. Gracias por tus consejos durante las comisiones de doctorado.

Adentrándonos en el "labo" me gustaría agradecer en primer lugar a Ángel, por ser como es, "el alma del labo" y por ponerte siempre a mi altura! A Dani, el guapetón del labo, el rey de los disfraces y la mejor compañía "hacienda puntas". A Esteve porque detrás de la careta de lobo se esconde un corderillo, por hacerme amenos los ratos de poyata y estar siempre dispuesto a ayudar con la informática! En el lado femenino, me gustaría dar las gracias a Joaneta por sus grandes ánimos y su motivación sin límites, eres grande y sé que siempre estarás ahí! A Ana Sesé, por todo el trabajo compartido pero sobre todo por nuestros sudores al poner la ultracentrífuga!! A Fuensi, siempre dispuesta a echarte una mano. A Vane, "la alegría de la huerta". Gracias a Celine, luchadora donde las haya! Gracias por ser tan cercana y por cuidar de los "pececitos"!

Un especial agradecimiento a Cris Vives, por el gran trabajo realizado durante su master y por toda tu ayuda en los últimos momentos de esta tesis. Aprendimos muchas cosas juntas y lo seguiremos haciendo.

A toda la gente del departamento. Gracias a Noemí y Tania por vuestra simpatía, buen humor y vuestra gran tarea. Gracias a los jefes Pere y Jordi por ser tan cercanos y hacer del IRTA una gran familia. Gracias también a los compañeros de "in vitro", sobre todo a Elena y Victoria por compartir buenos ratos y desayunos.

A toda la gente fuera del CRAG, pero que han sido un gran apoyo durante todos estos años. Gracias a Patrick por media vida juntos y lo que queda! Por alegrarme los días y seguirme en la andadura! Gracias a mi familia, ellos se merecen esta tesis más que nadie. Aunque en la distancia siempre – lo siento! Estáis en mi corazón y cabeza.

Thanks to all the people in The James Hutton Institute for their consideration during this time. Especially to Glenn Bryan and to the "potato team". Gracias a mi albaceteña favorita, Almudena, sin ti los días en Dundee serían muy grises. A todos los que me habéis dado ánimos durante este tiempo: Estela, Maria, Carmen, Vane, Guille, Mikel, Nora, Stephen, Ferran, Carla, etc.

No me gustan las despedidas, me gustan los hasta luegos! Mil gracias a todos y hasta siempre!

INDEX

INDEX OF CONTENTS

CHAPTER I. Development of diagnostic markers for selection of the subacid trait in peach

CHAPTER II. A candidate gene for fruit shape

INDEX OF FIGURES

GENERAL INTRODUCTION

CHAPTER I. Development of diagnostic markers for selection of the subacid trait in peach

CHAPTER II. A candidate gene for fruit shape

CHAPTER III. Somatic variability between peach to nectarine sport mutants and its implication in the *G* locus

INDEX OF TABLES

GENERAL INTRODUCTION

CHAPTER I. Development of diagnostic markers for selection of the subacid trait in peach

CHAPTER II. A candidate gene for fruit shape

CHAPTER III. Somatic variability between peach to nectarine sport mutants and its implication in the *G* locus

SUMMARIES

SUMMARY

The aim of current breeding programs is to provide new fruit varieties adapted to the local agronomic conditions and, at the same time, to satisfy the requirements of the consumers. This last fact implies to improve the fruit quality. The strategy followed by most breeding programs is based on performing controlled crosses to select those individuals showing the target traits. Although this approach has succeed in the production of most of the varieties available today, it is time consuming and costly due the time required to obtain fruits (2-3 years) and the resources needed to keep the seedlings in the field during the evaluation and selection processes. The objective of this thesis was to develop molecular markers useful in marker assisted selection (MAS) for three important agronomical traits in peach fruits: low acidity, flat shape and glabrous skin (nectarine trait). In the first two chapters of this document we have used region-based association analysis to study the architecture of the locus responsible for either subacid (D) and flat fruits (*S*). In both cases the study has provided markers (SSR and SNPs) ready to be applied for MAS in peach breeding programs. The study of the length of the subacid haplotype, which is maintained more than 24Kbp long, allowed to hypothesize about a unique origin of this trait and to identify candidate genes. Similarly, the analysis of the *S* locus allowed the identification of two linked INDELs in the second exon of the gene ppa025511m highly associated with the flat shape of the fruit. The association was tested in a broad panel of varieties and in the offspring of a crossing population between two flat peaches. The sequencing analysis of the whole gene allowed the identification of a big deletion, of about 9Kbp, affecting its 5' UTR, its first exon and its intron. The function of the gene was validated in a round sport mutant from a flat peach (UFO-4). This mutant was chimeric; the mutation only affected cells of the second layer (LII) of the meristerm, which generates the flesh of the fruit. A PCR amplification of the gene and the use of specific primers for the INDEL revealed a mutation in the flat allele in the flesh of the chimeric mutant, which produced the reversion to the round shape. The obligated heterozygosis of the flat allele and the reversion to the wild shape suggest a dominant negative (DN) mechanism.

In the third chapter we sequenced the whole genome of 5 peach varieties and 6 sport nectarines derived from them. The sequence data was used to estimate the overall somatic variability and to identify the causal mutation from hairy fruit (peach) to glabrous (nectarine). Standard pipelines for reads processing and SNP calling revealed an excess of false variants that was especially evident in the analysis of the sport mutants. One of the main causes for the false variants was the misalignments of repetitive regions. The use of more restrictive SNP calling filters reduced the excess of false variants. The nucleotide diversity ($\pi = 4.91x10^{-4}$) and heterozygosity (H_o=24.12%) of the varieties was similar to the one previously reported for

peach (Aranzana *et al.,* 2012; Verde *et al.,* 2013). The analysis of the variations in the *G* locus region showed lower π and higher Ho. To look for the causal allele for the nectarine trait we postulated two possible causes for the new mutation. The analysis of the sequences according to these two working hypothesis provided several candidate genes involved in the cell wall development, however none of them was the gene PpeMYB25, where a big insertion of 7Kb in its second exon has recently been described as linked with the trait (Vendramin *et al.,* 2014). This was probably due to an insufficient sequencing coverage in this genomic region.

RESUMEN

El objetivo actual de los programas de mejora genética del melocotón es generar variedades de frutos adaptados a las condiciones agronómicas locales y al mismo tiempo satisfacer los requerimientos del consumidor. Esto último implica mejorar la calidad del fruto. La estrategia seguida por muchos mejoradores se basa en la selección de descendientes de cruzamientos donde se espera segregación para determinados caracteres fenotípicos. Aunque mediante este procedimiento se han obtenido la mayoría de las variedades cultivadas actualmente, se trata de un método costoso tanto en tiempo como en dinero debido a que el melocotonero tiene un período de juvenilidad de 2-3 años y también a los recursos que supone el mantener las plántulas en el campo durante el proceso de evaluación y selección. El objetivo de esta tesis fue el desarrollo de marcadores moleculares para su aplicación en la selección asistida por marcadores (SAM) de tres caracteres de fruto importantes como son la subacidez, fruto plano y piel glabra (carácter nectarina). En los dos primeros capítulos de este documento hemos estudiado la arquitectura del locus responsable del carácter subácido (*D*) y del carácter fruto plano (*S*) y realizado análisis de asociación en esas regiones genómicas. Para ambos caracteres se han generado y validado marcadores moleculares (SSRs y SNPs) que pueden ser directamente aplicados a SAM en los programas de mejora del melocotonero. A partir del estudio de la longitud del haplotipo subácido (24kb) proponemos que existe un único origen para el alelo subácido. El análisis de la región también nos permitió identificar varios genes candidatos.

De la misma manera, el análisis del locus *S* nos permitió identificar dos INDELs altamente asociados con el carácter fruto plano. Estos polimorfismos se observaron en región codificante del gen ppa025511m, concretamente en el segundo exón del gen. Dicha asociación fue evaluada en un amplio set de variedades y en una población obtenida a partir del cruzamiento de dos parentales cuyos frutos eran planos. El análisis de la secuencia completa de este gen permitió la identificación la supresión de un fragmento de 9Kb que afecta la región 5'UTR del gen así como al primer exón, el intrón y a una pequeña parte del segundo exón. La función de dicho gen fue validada en un mutante tipo sport obtenido de manera natural en un árbol de la variedad plana 'UFO4'. Se trata de un mutante quimérico en el que la mutación sólo afecta a las células de la segunda capa meristemática (LII). Esta capa genera la pulpa del fruto. La amplificación de este gen mediante PCR y el uso de cebadores específicos para el INDEL identificado revelaron una mutación en el alelo plano en la pulpa del mutante quimérico que producía la reversión al fenotipo redondo.

El comportamiento genético de este carácter es siempre heterocigoto. Este hecho junto con la reversión a la forma redonda del alelo plano en el mutante sugieren que este alelo actúa como dominante negativo.

En el tercer capítulo secuenciamos el genoma de 5 variedades de melocotón y sus respectivos mutantes nectarina. Los datos de secuencia fueron utilizados para estimar la variabilidad somática y la mutación causal de la piel glabra. La metodología bioinformática empleada para el procesamiento de las lecturas y la identificación de pequeños polimorfismos presentó un exceso de falsos polimorfismos debido a problemas de alineamiento en las secuencias repetitivas del genoma. Mediante el uso de un método de filtrado más restrictivo se redujo el exceso de falsos polimorfismos. Los valores de diversidad nucleotídica (π= $4.91x10^{-4}$) y de heterocigosidad (H_o=24.12%) de las variedades analizadas fueron similares a los observados previamente por Aranzana *et al.,* 2012 y Verde *et al.,* 2013. El análisis del *G* locus mostró una baja π y una alta H_o. Para la búsqueda del alelo causal del carácter nectarina asumimos dos posibles causas para la aparición de la nueva mutación. Encontramos varios genes candidatos que presentaban funciones relacionadas con el desarrollo de la pared celular, sin embargo ninguno de ellos resultó ser PpeMYB25. Una inserción en el segundo exón de este gene de 7Kb ha sido descrita como la causa del carácter nectarina (Vendramin *et al.,* 2014). La insuficiente cobertura de secuenciación en la región genómica del locus *G* puede haber sido la causa para la no identificación de este gen en nuestros datos

RESUM

L'objectiu dels programes de millora genética del préssec és generar varietats de fruits adaptats a les condicions agronòmiques locals i al mateix temps satisfer els requeriments del consumidor. Això últim implica millorar la qualitat del fruit. L'estratègia seguida per molts milloradors es basa en la selección de descendents de creuaments on s'espera segregació per a determinats caràcters fenotípics. Encara Encara que mitjançant aquest procediment s'han obtingut la majoria de les varietats comercialitzades actualmente, es tracta d'un mètode costós tant en temps com en diners a causa del període de juvenilitat del presseguer (2-3 anys) i també als recursos que suposa el mantener les plàntules en el camp durant el procés d'avaluació i selecció.

L'objectiu d'aquesta tesi va ser el desenvolupament de marcadors moleculars per a la seva aplicació en la selecció assistida per marcadors (SAM) de tres caràcters del fruit: baixa acidesa (fruits subàcids), fruit pla (paraguaians) i pell glabra (caràcter nectarina). En els dos primers capítols d'aquest document hem estudiat l'arquitectura del locus responsable del carácter subàcido (D) i el fruit pla (S) i hem realitzat l'anàlisis d'associació en les seves respectives regions genòmiques. Per a ambdós caràcters s'han generat i validat marcadors moleculars (SSRs i SNPs) que poden ser directament aplicats a SAM presseguer. L'estudi de l'extensió de l'haplotipus subàcid (de més de 24kb) ens va permetre identificar diversos gens candidats. L'existència d'un únic haplotipus en un panell de varietats genèticament distants ens suggereix l'existència d'un únic origen de l'al·lel subàcido. De la mateixa manera, l'anàlisi del locus S ens va permetre identificar dues INDELs altament associats amb el caràcter fruit pla. Aquests polimorfismes es van observar en regió codificant del gen ppa025511m, concretament en el segon exò del gen. Aquesta associació va ser avaluada en un ampli panell de varietats i en una població obtinguda a partir del creuament de dues parentals de fruits plans. L'anàlisi de la seqüència completa d'aquest gen va permetre la identificació de la supressió d'un fragment de 9Kb que afecta la regió 5'UTR del gen així com al primer exò, a l'intró i a una petita part del segon exò. La funció d'aquest gen va ser validada en un mutant tipus "sport" generat espontàneament en un arbre de la varietat plana 'UFO4'. Es tracta d'un mutante quimèric mb una mutació que només afecta a les cèl·lules meristemàtica (LII). Aquesta capa genera la polpa del fruit. L'amplificació per PCR de l'INDEL d'aquest gen en la polpa del mutant rodó va revelar un canvi a l'al·lel pla. Malgrat que l'alelo pla és dominant, els fruits plans han de presentar-lo en hetericigosis per a ser viables. Aquest fet juntament amb la reversió a la forma rodona del mutant d'UFO4 suggereixen que aquest al·lel pot actuar com dominant negatiu.

En el tercer capítol seqüenciem el genoma de 5 varietats de préssec i els seus respectius mutantes amb fenotip nectarina. Les dades de se seqüència van ser utilitzats per a estimar la variabilitat somática i la mutació causal de la pell glabra. La metodología bioinformática empleada per al processament de les lectures i la identificació dels petis polimorfismes va generar un excés de falsos polimorfismes, possiblement causats per alineaments erronis de les seqüències repetitives del genoma. Mitjançant l'ús d'un mètode de filtrat més restrictiu es va reduir l'excés de falsos polimorfismos. Els valors de diversitat nucleotídica (π= 4,91x10-4) i d'heterozigositat (Ho=24,12%) de les varietats analitzades van ser similars als observats prèviament per Aranzana *et al.,* (2012) i Verd *et al.,* (2013). L'analisi del locus *G* va mostrar una menor π i una major Ho. Per a la recerca de l'al·lel causal del caràcter nectarina vam postular dues causes posibles per a l'aparició de la nova mutació. Sota aquestes dues possibles hipótesis vam identificar diversos gens candidats que presentaven funcions relacionades amb el desenvolupament de la paret cel·lular. No obstant això cap d'ells va resultar ser PpeMYB25 on recentment s'ha descrit una inserció de 7Kb en el seu segon exòn asssociada al carácter nectarina (Vendramin *et al.,* 2014). La insuficient cobertura de seqüenciació en la regió genómica del locus G pot haver estat la causa de la no identificació d'aquest polimorfisme en les nostres seqüències.

ABBREVIATIONS

A: Adenine
AB: Applied Biosystems
AFLP: Amplified Fragment Length Polymorphism
Alt: Alternative
ASF: Agro Sélection Fruits
Asn: Asparagine
ASPE: Allele Specific Primer Extension
ATP: Adenosine triphosphate
BAC: Bacterial Artifitial Chromosome
BC: Before Christ
BLAST: Basic Local Alignment Search Tool
BLOSUM: Blocks of Amino Acid Substitution Matrix
Bp: base pair
C: Cytosine
CCD: Charge Coupled Device
cDNA: complementary DNA
CDS: Coding DNA Sequence
CG: Candidate Gene
CIV: Consorzio Italiano Vivaisti
cM: centimorgan
CRA: Consiglio per la Ricerca e la Sperimentazione in Agricoltura
CRAG: Centre for Research in Agricultural Genomics
CTAB: Cetyl Trimethylammonium Bromide
cv: cultivar
ddNTP: dideoxy nucleotide triphosphate
DNA: Deoxyribonucleic acid
dNTP: deosynucleotide triphosphate
DOFI: Horticultural Department of Florence University
DZ: dehiscence sone
EST: Expressed Sequence Tag
EU: European Union
F: Forward
F_1: First Fillial Generation
F_2: Second Fillial Generation
FAOSTAT: Statistics division of the FAO (Food and Agriculture Organization)
G: Guanine
GBS: Genotyping by Sequencing
GDR: Genome Database Rosaceae
GO: Gene Onthology
GS: Genome Sequencer
GS-FLX: Genome Sequencer FLX (flexible) system
GWA: Genome Wide Approach
HRM: High Resolution Melting
IFC: Integrated Fluidic Circuit
INDEL: Insertion/Deletion polymorphism
INRA: Institut National de la recherche Agronomique
IPSA: Institute for Post Graduate Studies in Agriculture

IPSC: International Peach SNP Consortium
IPTG: Isopropyl-b-D-1-tiogalactopiranoside
IRTA: Institut de Recerca i Tecnologia Agroalimentàries
ISF: Instituto Sperimentale per la Frutticoltura Roma
Kb: Kilobase
Kpb: Kilobase pair
Kv: Kilovat
L: Layer
LD: Linkage Disequillibrium
Leu: Leucine
LG: Linkage Group
LRR: Leucine Rich Repeat
MAFFT: Multiple Alignment using Fast Fourier Transform
MAS: Marker Assisted Selection
Mb: Megabase
MEGA: Molecular Evolutionary Genetics Analysis
meq/L: Milliequivalents per Liter
ML: Maximum likelihood
MSA: Multiple Sequence Alignment
N: Eq/L equivalent per litre, Normality
NaOH: Sodium Hydroxide
NCBI: National Center of Biotechnology Information
NGS: Next Generation Sequencing
NJ: Neighbour Joining
nr: non-redudant
pacBio: Pacific Biosciences
PAGE: PolyAcrylamide Gel Electrophoresis
PCA: Principal Components Analysis
PCR: Polymerase Chain Reaction
PGM: Personal Genome Machine
pH: power of Hydrogen
PSB: Vegetal production company, Murcia, Spain
R: Reverse
RAPD: Random Amplified Polymorphic DNA
RFLP: Restriction Fragment Length
RLK Receptor Like Kinase
RNA: Ribonucleic acid
RNAi: RNA interference
rpm: Revolutions per minute
RU-NJ: Rutgers University New Jersey
SAM: Sentrix Array Matrix
SBE: Single Base Extension
SBS: Sequencing by Synthesis
SMS: Single Molecule Sequencing
SNP: Single Nucleotide Polymorphism
SOLiD: Sequencing by Oligonucleotide Ligation and Detection
SSC: Soluble Solid Concentration

STMS:	Sequencing Tagged Microsatellite Sites
STR:	Short Tandem Repeats
T:	Thymine
TA:	Titratable Acidity
TC:	Técnica Comercial frutas, Barro, Spain
TILLING:	Targeting Induced Local Lesions in Genomes
Tyr:	Tyrosine
UCD:	University of California, Davis
USA:	United States of America
v/v %:	[volume of solute]/[volume of solution] * 100 %
WGS:	Whole Genome Shotgun
YAC:	Yeast Artifitial Chromosome

GENERAL INTRODUCTION

I.1. PEACH

I.1.1 Peach taxonomy

Peach [*Prunus persica* (L.) Batsch] is a diploid (2n = 2x = 16) fruit tree species and belongs to the *Prunus* genus which comprises more than 430 species from subtropical to temperate regions (Rehder, 1940). Based on fruit type, *Prunus* genus and other small genera were traditionally classified into the *Prunoideae* (drupe) subfamily of the *Rosaceae* family together with *Spiroideae* (follicle or capsule)*, Rosoideae* (achene) and *Maloideae* (pome) subfamilies. However, recent molecular phylogenetic studies based on analysis of sequences from multiple chloroplast and nuclear genes (Morgan *et al*., 1994; Potter *et al*., 2007; Potter *et al*., 2002) divide *Rosaceae* family into three subfamilies: *Dryadoideae* (*Cercocarpus*, *Dryas* and *Purshia*; x=9), *Rosoideae* (*Fragaria*, *Potentillia, Rosa*, *Rubus* and others; x=7) and *Spiraeoideae*, which has been corrected to *Amygdaloideae* based on recent changes on the International Code of Nomenclature for Algae, Fungi and Plants (*Kerria*, *Spiraea* and others; x=8, 9, 15 or 17) (McNeill *et al.*, 2012). *Prunus* genus is included in the *Amygdaloideae* subfamily (Morgan *et al,* 1994; Potter *et al*., 2007; Potter *et al*, 2002). (**Fig. I.1**).

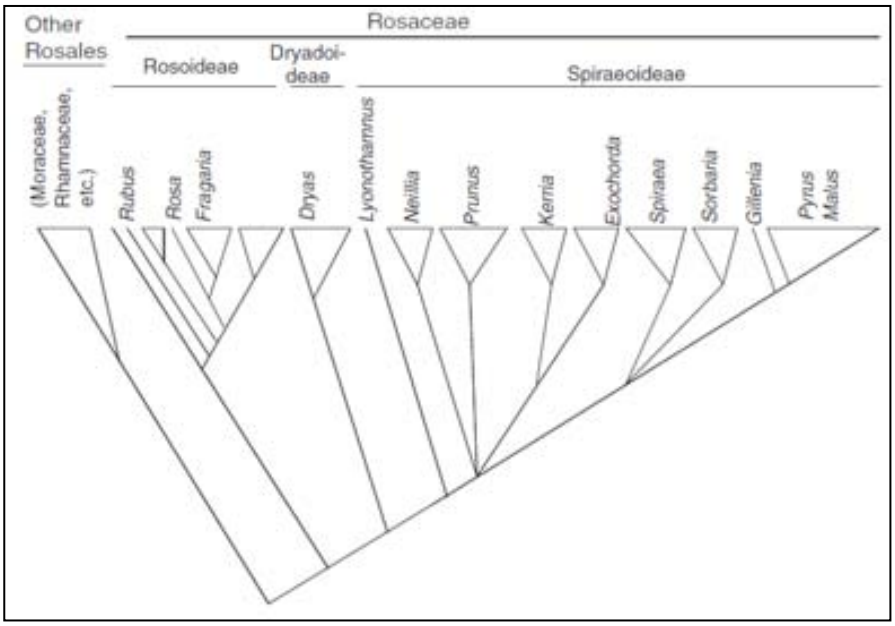


Figure I.1. Phylogenetic relationship in *Rosaceae* from Potter, 2007, with the circumscriptions of the three subfamilies in their infrafamilial classification indicated. Polytomies indicate cases in which analyses to date have not been able to resolve the branching order among lineages. .

The most widely accepted infrageneric classification of *Prunus* genus is the one by Rehder 1940 which consists of five subgenera : *Amygdalus* (peaches and almonds), *Cerasus* (cherries), *Prunus* (plums), *Laurocerasus* (evergreen laurel-cherries), and *Padus* (deciduous bird-cherries). Thus, the Prunus systematic classification is the following:

Kingdom*:* ***Plantae***
Division: ***Magnoliophyta***
Class: ***Magnoliopsida***
Order: ***Rosales***
Family: ***Rosaceae***
Subfamily: ***Amygdaloideae***
Tribe: ***Amygdaleae***
Genus*:* ***Prunus***
Subgenus: ***Amygdalus***
Section: ***Euamygdalus***

Prunus genus consists of over 200 species of deciduous and evergreen trees and shrubs with several members that are economically important stone fruit and nut crops in addition to peach such as: almond (*P. dulcis* (Mill.) D.A Webb), apricot (*P. armeniaca* (L.)), prune plum (*P. domestica* (L.)), Japanese plum (*P. salicina* (Lindl.)), sour cherry (*P. cerasus* (L.)) and sweet cherry (*P. avium* (L.)). Other closely related species within *Prunus* genus are: *P. mira* (Koehne.), *P. daviniana* ((Carr). Franch), *P. ferganensis* (Kostfina and Rjablov), *Prunus kansuensis* (Rehder) and the recently discovered *P. pananesis* (Chen *et al*., 2013). These wild relative species are sexually compatible with peach producing fertile hybrids (Moing, 2003).

I.1.2 Peach origin and distribution

Peaches were originated in China, probably in Tarim basin north of Kun Lun mountains, where they were cultivated for at least 4000 years and where still exits the greatest genetic diversity. The most ancestral peach form reported from China may be the Mao Tao (hairy peach) wild peach (Rieger, 2006).

Peach spread to the western hemisphere through the trade routes from China to Persia (actual Iran) in the 2nd to 1st century BC, from where takes its name *P. persica* (Hedrick *et al.,* 1917). Then, peach traveled from Persia to the Mediterranean region. It is not clear if there was a unique arrival or if there were two independent arrivals of this tree to Europe. Thus,

peach could have arrived to Italy in the 1st century BC, or it could have arrived independently and almost simultaneously to France along the Danube river and the Black sea region (Werneck, 1956). The introduction of peach into America had to wait until the 16th century when Spaniards brought it and spread it along the eastern and northern region, where they started to cultivate and propagate peaches by seeds (Byrne *et al.,* 2012; Hedrick *et al.,* 1917). A second peach introduction in the western North America occurred directly from China in the mid-1850s, with few varieties (Chin *et al.,* 2014). One of them was 'Chinese Cling', which is considered one of the founders of the current peach commercial varieties in Occidental countries (Scorza & Sherman, 1996). 'Elberta', a peach variety originated from an open pollination of 'Chinese Cling', become the most famous variety in the USA and in the most important peach growing countries, with big fruits and a good firmness. This variety, among others, was intensively used as parental line in breeding programs. The massive use of few progenitors in breeding programs produced a bottleneck, that together with the self-compatibility of peach are the principal reasons of the low levels of genetic variability in occidental peach varieties.

In consequence, nowadays, Chinese germplasm and local varieties (i.e. varieties not obtained in breeding programs) may constitute the main source of diversity for modern occidental breeding programs (Li *et al.,* 2013; Xie *et al.,* 2010).

I.1.3 Production and Economic importance

Peach is grown in the both hemispheres, especially in the temperate zone between 30° and 45° latitude (Scorza & Sherman, 1996), with mild winters and with few cold hours, which are needed to brake bud dormancy. However, production is also found throughout the subtropics and tropical regions (Byrne *et al.,* 2000).

Peach is in the tenth place between all fruits produced world-wide (excluding melons) (**Fig. I.2**) world-wide. Its production has increased in the last five years in 27.36%. In 2011 the surface of peach crop was 157,188,039 ha, producing 2,151,018,000 t of fruits. The main producer countries are: China (11,529,719 t), Italy (1,636,753 t), Spain (1,336,362 t) and EE.UU. (1,176,610 t) (**Fig.I.3**).

In 2011, peaches were the fourth most produced fruit in EU after grapes, apples and oranges with a total production of 4,329,917 t in 284,149 ha, distributed principally in Italy, Spain and Greece. Spain is the fourth producer worldwide and the second in EU. Furthermore, Spain is the first exporter worldwide, exporting 657,976 t in 2011, which represents the 49.23% of its production (1,336,362 t) (FAOSTAT, 2014) (**Fig. I.3**).

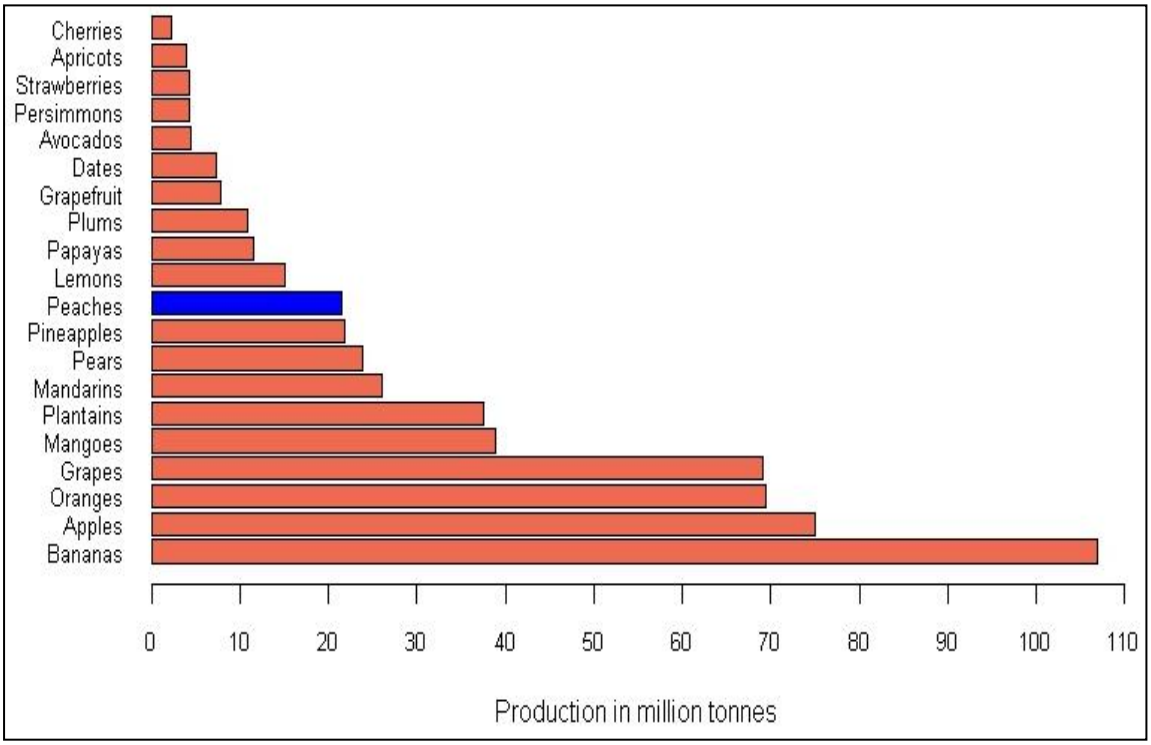


Figure I.2. The first ten most produced fruits world-wide in 2011 (FAOSTAT 2014).

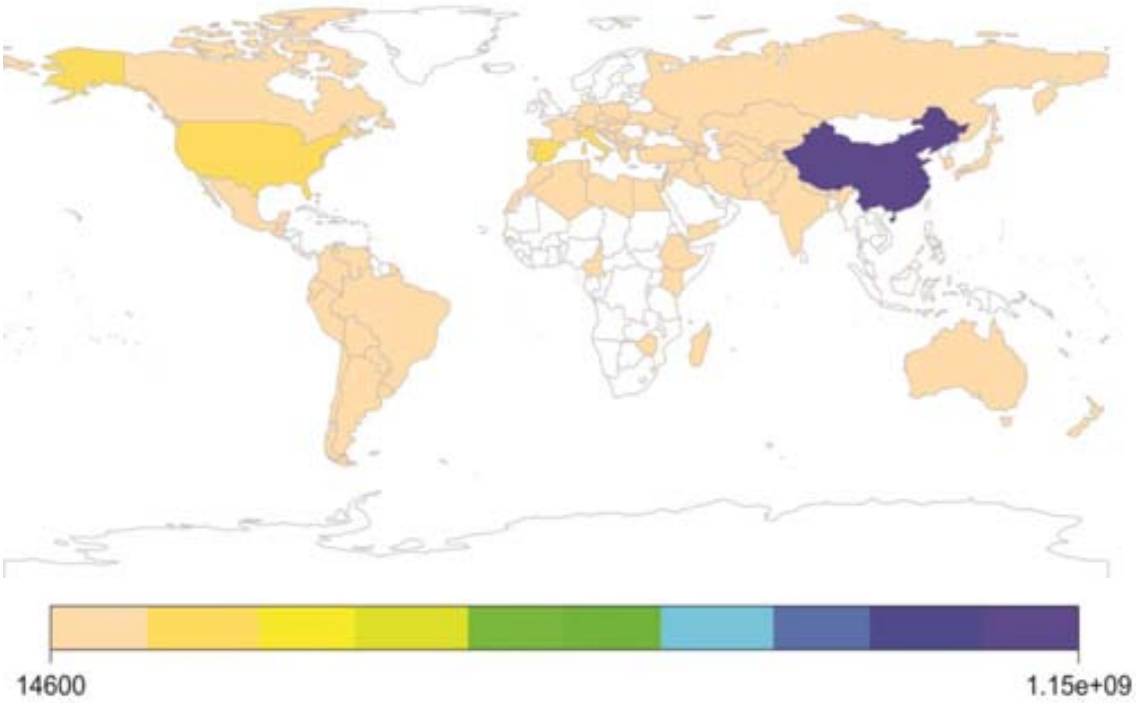


Figure I.3. Worldwide peach production in 2011 (FAOSTAT 2014)

Peach is the most important fruit species in Spain with a cultivated area of 81,374 ha in 2011, followed by apple, pear and cherry. Within Spain, the biggest peach production is located in Valle del Ebro (Cataluña (417,760 t) and Aragón (401,277 t)), followed by Murcia (116,000 t), Extremadura (115,520 t), Andalucía (79,000 t) and Comunidad Valenciana (22,000 t) (Reig *et al*., 2013).This high diversity in the production areas in Spain has provided a wide calendar for harvest which covers the period between middle of April and the end of October.

In the last forty years there has been an increase in the peach cultivated area in Spain (**Fig. I.4**) mainly due to the high varietal dynamism and the use of well adapted rootstocks which allow a quick establishment of the most adapted varieties to the climate conditions, pathogens, consumer requirements and demands (Iglesias & Casals 2013; Llácer 2005).

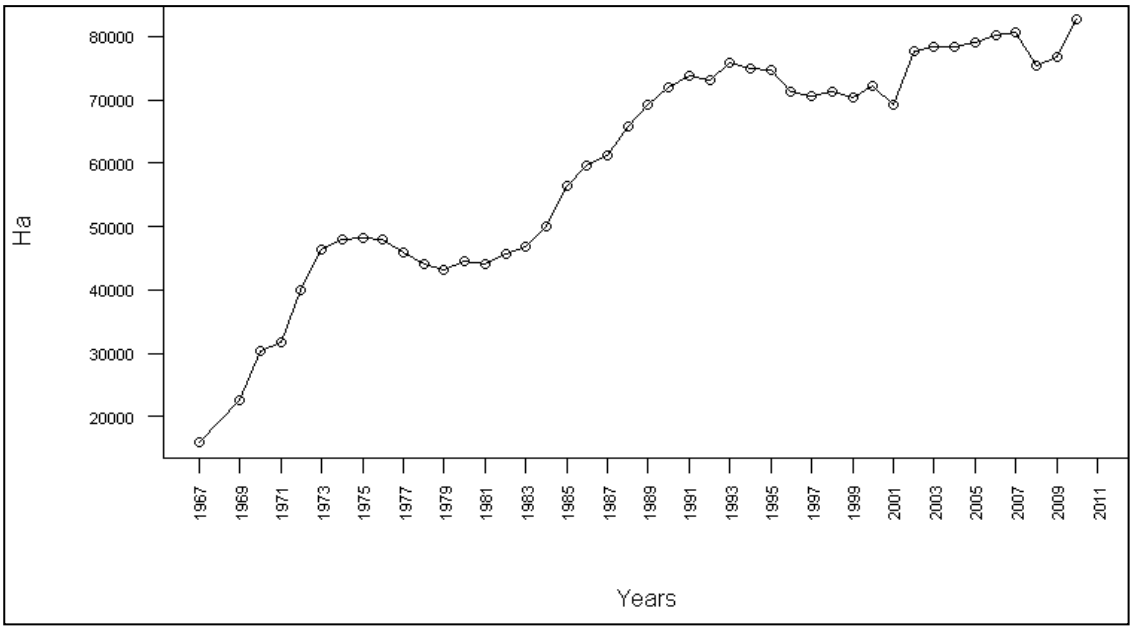


Figure I.4. Expansion of peach cultivated surface in Spain between 1967 and 2011. (Source:FAOSTAT 2014)

I.1.4 Peach Genetics

Peach has long been one of the genetically best characterized species in the *Rosaceae* (Arús *et al.,* 2012), and is considered together with *Malus* × domestica Borkh and *Fragaria vesca* L. a model species for the development of genetic studies due to several advantageous characteristics. It is a diploid species contrarily to other fruit crops, such as European plum, sour cherries, apple and pear that are polyploid. Its genome is divided up into eight chromosomes (2n = 2x = 16) (Jelenkovic & Harrington 1972) and its size is small (≈227Mb) (Verde *et al.*, 2013) compared with *Fragaria vesca* (Shulaev *et al.,* 2011), *Malus x domestica* (Velasco *et al.,* 2010) and the recently sequenced *Prunus mume* (Zhang *et al.,* 2012), but about twice that of *Arabidopsis* (The Arabidopsis Genome Initiative, 2000). The reference peach genome released by the International Peach Genome Initiative (Verde *et al.,* 2013) has supposed a valuable tool for the improvement of genetic studies for this species and other relative species .

Other important characteristic of peach is that despite being self-compatible and, thereafter, it is mainly autogamous, it can also be cross pollinated which is a possible mechanism to introduce new genetic variability (Byrne, 1990). Furthermore, it has a relatively short juvenile period of 2–3 years compared to most other fruit tree species that require 5–10 years. For all of these reasons, the inheritance of many major genes is already known (Monet *et al.,* 1996).

I.2. GENETIC MARKERS

I.2.1 Definition, history and classification

Genetic markers are biological characteristics established by the genetic variants between individual organisms or species and, if they are located in genes or are closely linked to them, can be used as 'signs', 'flags', 'probes' or 'tags' of such genes. The first genetic markers used were the morphological ('classical' or 'visible') markers which themselves are phenotypic characters or variants. They were the ones used in the early plant breeding. The first biochemical markers used were the isozymes, i.e. the genetic variants of a specific enzyme. The utility of such markers was limited due to their small numbers of potential marker loci, low levels of polymorphism between closely related individuals and their not always consistent expression. Isozymes were replaced by DNA markers in the early 1980s like RAPDs (Random Amplified Polymorphic DNA), RFLPs (Restriction Fragment Length Polymorphisms), AFLPs (Amplified Fragment Length Polymorphisms), SSRs (Simple Sequence Repeats) and SNPs (Single Nucleotide Polymorphisms). Such DNA markers are based on variants in the DNA sequence such as point mutations produced by single nucleotide substitutions, insertions or deletions of more or less big DNA fragments produced by errors in replication of tandemly repeated DNA fragments (Paterson, 1996).

DNA markers advantages are their abundance, their high polymorphism and the availability of evaluation at any developmental stage. Moreover their detection is not influenced by environmental factors (Winter & Kahl 1995).

DNA markers can be classified (1) based on the methodology for their detection (1.a) southern or hybridization-based, (1.b) polymerase chain reaction (PCR)-based and (1.c) DNA sequence based; (2) based on their dominant or codominant polymorphism and (3) based on their location respect to a gene, in which they can be classified into (3.1) random molecular markers (anonymous or neutral markers), (3.2) gene targeted markers and (3.2) functional markers. Random markers are distributed all across the genome while gene targeted markers are found within genes not necessarily involved in phenotypic variation, e.g. un-translated regions (UTRs) of EST sequences (Aggarwal *et al*., 2007). Functional markers are located in the polymorphism causally associated with a phenotypic trait variation, so they are totally linked to the allelic forms in the locus and the functional motifs (Andersen & Lübberstedt, 2003). Random and gene targeted markers can be used to tag functional variations if QTL studies establish an association between marker and trait, however and unlike functional markers, the association can be broken by recombination.

The choice of one DNA marker will depend on the research goal. A comparison between the most widely-used DNA markers is shown in **Table I.1**. In peach, RFLPs, RAPD and AFLPs markers have been used for genetic diversity studies (Warburton & Bliss, 1996; Bouhadida & Martín, 2007; Nagaty *et al*., 2011;), for synteny studies (Dirlewanger *et al*., 2002; Illa *et al*., 2011; Vilanova *et al.*, 2008), for cultivar identification (Aranzana *et al.,* 2003; Han *et al.,* 2014; Lu *et al.,* 1998; Rojas *et al.,* 2008) and for construction of linkage maps (Boudehri *et al.,* 2009; Dhanapal *et al.,* 2012; Martinez-Garcia *et al.,* 2013; Pirona *et al.,* 2013; Salazar *et al.,* 2014; Verde *et al.,* 2005).

Table I.1. Comparison of the five most widely used DNA markers in plants**. A.** Mutation at enzyme restriction or PCR priming site, **B.** Insertion or deletion between enzyme restriction or PCR priming sites, **C.** Change of tandem repeat units between enzyme restriction or PCR banding sites and **D**. Single nucleotide mutation.

Characteristics	DNA markers				
	RFLPs	RAPDs	AFLPs	SSR	SNPs
Methodology	Southern blot	PCR	PCR	DNA-sequence	DNA-sequence
Molecular basis	A, B	B	A, B	C	D
Genomic Coverage	Low copy coding region	Whole genome	Whole genome	Whole genome	Whole genome
Inheritance	Codominant	Dominant	Dominant	Codominant	Codominant
Polymorphism	Medium	Medium	High	Very High	Medium
No. Loci / marker	1-3	1-20	10-100	1-2	1
No. Alleles / locus	Multiallelic	2	2	Multiallelic	2
DNA quantity (ug) / reaction	5-10	0,02	0,5-1	0,05	0,05
DNA quality	High	Moderate	High	Moderate	High
Reproducibility	Very High	Moderate	High	Very High	Very High
Cost development / analysis	High/High	Low/Low	Low/Low	Low/Moderate	High/Low
Automation	Low	Medium	High	High	High
Suitable utility in diversity genetics and breeding	Genetics	Diversity	Diversity and Genetics	All purposes	All purposes
Type of probes/primers	Low copy DNA or cDNA clones	Usually 10 bp random nucleotides	Specific sequence	Specific sequence	Allele-specific PCR primers
Effective multiplex ratio	Low	Medium	High	High	Medium to high

Nowadays the SSR and SNP markers are the most widely used and are the ones used in this research work.

I.2.2 Microsatellites or SSRs

Microsatellites or SSRs (Single Sequence Repeats), STRs (Short Tandem Repeats) or STMS (Sequence Tagged Microsatellite Sites) are tandemly repeated units of short nucleotide motifs, which are flanked by very conservative sequences (Buschiazzo *et al*., 2006; Morgante & Olivieri, 1993; Zane *et al*., 2002), which are used as a template for the development of PCR primers to amplify the region covering the SSR repeats. The length of a single repeated motif is usually 1-6bp long. Normally, the shorter motifs have more repeats than longer motifs. These motifs can be called 'perfect motifs' (if it is a single motif) or 'compound' (when the motif is compound by two or more

motifs). The longer and perfect SSR, the greater allelic variability it exhibits (Buschiazzo & Gemmell, 2006; Kelkar & Tyekucheva, 2008). SSRs are characterized to be hyper variable mainly due to the predominant mutation mechanism that generates them. This mechanism is the slipped-strand mispairing of the DNA polymerase during DNA replication, which results in the gain or loss of one or more repeat motifs depending on whether the newly synthesized DNA chain or the template chain loops out (Coenye & Vandamme, 2005; Schlöterrer, 2002). The rate of mutation depends on several factors including the number of repeats, the class of repeat (di-, tri-, etc.), and the chromosomal location in relation to a gene and on the GC content. The mutation rate (μ) per generation per locus is an important parameter in models of population genetics as it permits to estimate the timing of evolutionary divergence between species (Schötterer, 2000; Wehrhahn, 1975), and the effective population size of the species (Slatkin 1995; Vigouroux *et al.,* 2002). Mutation rates of microsatellites have long been estimated in numerous studies. One of the most important observations was that the mutation rate largely varies in several orders of magnitude among different species, ranging from 5×10^{-6} in Drosophila (Schug *et al.,* 1997; Vazquez *et al.,* 2000) to 10^{-3} in humans (Brinkmann *et al.,* 1998; Xu *et al.,* 2000).

Other properties that make SSRs a good DNA marker are their reproducibility, codominant nature, locus specificity, random dispersion across genomes (in coding or non-coding regions although more abundant in the last one) and their transferability between close related species SSR markers can be easily analyzed by PCR and electrophoresis. They can be multiplexed and their genotyping can be semi-automated by using end-labeling primers enabling the visualization of length variants on automated DNA sequencer.

When SSRs were initially used in genetic studies their identification was done by screening sequences or expressed sequence tag (ESTs) in databases or libraries of clones (Edwards *et al*., 1996; Kantety *et al.*, 2002; Santana *et al*., 2009) when available, or alternatively they were obtained *de novo* by constructing genomic libraries enriched for a few targeted motifs.

Currently, the most efficient option for SSR discovery is by *in silico* search across next-generation sequencing (NGS) data (Zalapa *et al*., 2012). There are different algorithms used for SSR detection (Cavagnaro *et al*., 2010). Some of the most widely used programs for SSR identification are: 'mreps', able to find imperfect repeats (Kolpakov, 2003); 'MIcroSAtellite' (MISA; Thiel *et al*., 2003); 'SSR locator', which is Windows-based (Da Maia *et al*., 2008); 'WebSat', which has an interactive visualization (Martins *et al*., 2009) and 'GMATo' for large genomes, providing statistic distribution of microsatellites through genome (Wang *et al.*, 2013).

I.2.3 Single Nucleotide Polymorphisms or SNPs

SNPs are based on a change or substitution in a single base pair in the genomic DNA sequence. The different sequence alternatives (alleles) can be A, T, C or G and the least frequent allele has to be present in at least 1% in the population to consider this variation as a SNP (Brooks, 1999) They are considered the ultimate form of molecular marker because a nucleotide base is the smallest unit of inheritance and they are the most abundant genetic markers in all organisms. The 90% of human genetic variation is due to SNPs, with one SNP every 100-300 base pairs (Wang *et al*., 1998). SNP variability in peach is lower; with an estimated average of 1 SNP every 598 base pairs (Aranzana *et al.,* 2012).

SNPs can be divided in transition and transversions, according to the nucleotide substitution. Transitions consist on the substitution of one purine by other purine (C/T) or of one pyrimidine by another pyrimidine (G/A), while transversions consist on the substitution of one pyrimidine by a purine or vice versa. Transitions are more abundant than transversions in humans and plants, where 67% of the total SNPs are transitions (Edwards *et al*., 2007). Within transitions, two out three SNPs are based on a substitution from a C to a T (Yu *et al*., 2005).

SNPs can fall within coding sequences of genes, non-coding regions or in intergenic regions at different frequencies in different chromosome regions (Li & Sadler, 1991; Schmid *et al*., 2003). Those SNPs within a coding region may or may not change the amino acid sequence (non-synonymous or synonymous, respectively). Although the SNPs producing non-synonymous mutations are normally located in coding regions, those falling in non coding regions could have consequences on the expression of a gene by producing changes in splicing events, in the binding of transcription factors or in the sequence of non-coding RNA.

I.2.3.1 SNP discovery techniques

The discovery of novel SNPs can be achieved by several approaches. The conventional and direct method for the identification of new SNPs is the sequencing of DNA PCR products (by Sanger method) from different accessions or individuals. In general, the amplicons sequenced can be coding regions (genes of interest or ESTs), but selecting non-coding regions normally increases the frequency of polymorphism found (Zhu & Perry, 2005).

Other sources of SNPs are the EST and the genomic sequence libraries prepared from diverse set of individuals, which can be screened *in silico.* The drawback of this methods is that the

SNPs must be validated by re-sequencing or by other genotyping method (Batley *et al*., 2003; Bonet *et al*., 2009; Chagné *et al*., 2008; Dantec *et al*., 2004; Georgi *et al*., 2002).

In the last ten years the huge advance in next generation sequencing (NGS) technologies has represented a true revolution in the discovery of novel SNPs, producing a massive amount of nucleotide reads per run from either genomic DNA or cDNA that once assembled to a reference genome, provides genome-wide SNPs (Chan, 2009).This strategy has been used for example in strawberry (Celton *et al*., 2010), potato (Anithakumari *et al*., 2010), flax (Kumar *et al*., 2012), olive (Kaya *et al*., 2013), chickpea (Gaur *et al*., 2012), eucalyptus (Hendre *et al.,* 2012), melon (Blanca *et al*., 2012) and oat (Oliver *et al*., 2011) and peach (Verde *et al.*, 2013) among other species. Sequencing technologies are described in section I.3 of this introduction.

I.2.3.2 SNP genotyping techniques

SNP genotyping assays can be classified in: (1) allele-specific hybridization methods; (2) enzyme based methods and (3) post amplification methods based on physical properties of the DNA.

The allele-specific hybridization methods interrogate SNPs by hybridizing complementary DNA probes to the SNP site. The challenge of this approach is reducing cross-hybridization between the allele-specific probes. This challenge is generally overcome by manipulating the hybridization stringency conditions. One of the most currently widely used approaches based on hybridization are the SNP microarrays in which hundreds of thousands of probes are arrayed on a small chip, allowing for many SNPs to be interrogated simultaneously. Because SNP alleles only differ in one nucleotide and because it is difficult to achieve optimal hybridization conditions for all probes on the array, the target DNA has the potential to hybridize to mismatched probes. This is addressed somewhat by using several redundant probes to interrogate each SNP. Probes are designed to have the SNP site in several different locations as well as containing mismatches to the SNP allele. By comparing the differential amount of hybridization of the target DNA to each of these redundant probes, it is possible to determine specific homozygous and heterozygous alleles (Heller, 2002).

The enzyme based methods use a broad range of enzymes including DNA ligase, DNA polymerase and nucleases to generate high-fidelity SNP genotypes. An example of this methodology are those approaches based on primer extension which consist in the specific addition of a unique nucleotide to an extension reaction from a template DNA (Sokolov, 1990). The identification of the incorporated nucleotide is done by fluorescence like SnaPshot® (Applied Biosystems, CA) or pirosequencing (Ronaghi *et al.,* 1996) or mass spectrometry by MALDI-TOF (Matrix Assisted Laser

Desorption Ionization Time-of Flight (Braun *et al.,* 1997). Illumina Incorporated's Infinium assay is an example of a whole-genome genotyping pipeline that is based on primer extension method.

Within the post-amplification methods based on the physical properties of the DNA, the High Resolution Melting analysis has been one of the SNP genotyping methodologies used in this thesis. The method is based on detecting small differences in PCR melting (dissociation) curves. It is enabled by improved dsDNA-binding dyes used in conjunction with real-time PCR instrumentation that has precise temperature ramp control and advanced data capture capabilities. The region of interest within the DNA sequence is first amplified using the polymerase chain reaction. During this process, special saturation dyes are added to the reaction, that fluoresce only in the presence of double stranded DNA. Such dyes are known as intercalating dyes. During PCR, the amplicons of interest is amplified. As the amplicon concentration in the reaction tube increases the fluorescence exhibited by the double stranded amplified product also increases. After the PCR process the HRM analysis begins. In this process the amplicon DNA is heated gradually from around 50°C up to around 95°C. As the temperature increases, at a point the melting temperature of the amplicon is reached and the sample DNA denatures and the double stranded DNA melts apart. Due to this the fluorescence fades away. This is because in the absence of double stranded DNA the intercalating dyes have nothing to bind to and they only fluoresce at a low level. This observation is plotted showing the level of fluorescence vs the temperature, generating a melting curve. Since different genetic sequences melt at slightly different rates, they can be viewed, compared, and detected using these curves.

Up to this moment, the most widely used array-based platforms in plants are the GoldenGate and Infinium assays based on BeadArray technology of Illumina ® (**Table I.2**). The Illumina's Infinium shows higher throughput than the GoldenGate and the choice between them will depend on the number of SNPs and samples to study. The existence of commercially validated Infinium chips in some species is an advantage for those related species, because the use of these pre-made arrays could involve a reduction in the cost but obviously the number of valid markers will depend on the relationship between the reference and the studied specie.

There are other good arrays in terms of throughput like Beckman Coulter's GenomeLAb SNPstream (Bell *et al.,* 2002) which can process up to three million genotypes in 384 samples per day per instrument. The widely used Affimetrix GeneChip system allows the detection of hundreds of thousands of SNPs per array and can be used for SNP discovery as well by hybridization (Wang *et al.,* 1998). More recently ultra-high throughput nano-arrays or nano-chips were released for the

screening of human genome (Chen & Li, 2007). These small chips have been already improved in their sensitivity by the incorporation of semiconductor fluorescent nanocrystals (Ioannou & Griffin, 2010).

Currently, the newest genotyping array is the Ion Torrent Chip, with very high throughput because is based on semiconductor technology which uses fluidics and micromachining. Hence, the Ion Torrent 314 Sequencing Chip supports up to 1.3 million DNA testing wells and it has been extensive used in lots of species already (Sarris *et al*., *2013;* Whiteley *et al*., 2012; Zhang *et al*., 2013) and the 318 chip is starting to be used in bacteria (Whiteley *et al*., 2012) and human genotyping (Lu *et al.*, 2013).

High-density SNP genotyping arrays have been designed for several domestic animals including cattle (Matukumalli *et al.,* 2009), pig (Ramos *et al.,* 2009) and chicken (Groenen *et al.,* 2011); arrays are being developed in several plant species including apple (Chagné *et al.,* 2012), maize (Ganal *et al.,* 2011), tomato (Sim *et al.,* 2012), potato (Felcher *et al.,* 2012) and cherry (Peace *et al.,* 2012). In peach, was developed a moderate-density high-throughput Infinium®genotyping platform relevant for worldwide peach breeding germplasm utilizing SNPs discovered using next generation sequencing platforms. The SNP detection was done by whole genome re-sequencing of 56 peach breeding accessions using Illumina and Roche/454 sequencing technologies. A total of 1,022,354 SNPs were detected and a subset of them was validated with the Illumina Golden Gate® assay, verifying 75% of genic (exonic and intronic) SNPs while only about a third of intergenic SNPs were verified. After several filtering steps, a total of 8,144 SNPs were introduced in The International Peach SNP Consortium (IPSC) 9K SNP array v1. These SNPs were distributed over the eight chromosomes separated by 26.7 kb. A total of 6,869 polymorphic SNPs were found using the Infinium ® genotyping assay in 709 accessions divided in two independent evaluation panels; one panel form European Union (EU) consisting in 229 peach cultivars and 3 wild related Prunus species or their hybrid with peach and the other one form USA (US) composed by 1479 samples including pedigree-linked cultivars, breeding lines and seedlings (Verde *et al.,* 2012).

Table I.2. The most used micro-array-based high throughput SNP genotyping systems.

Features	Illumina		Beckman coulter	Affymetrix	
	GoldenGate	Infinium	SNPstream	MIP	GeneChip or oligonucleotide arrays
Array type	Tag array on beads	Specific probe primers on beads	Tag array on glass	Tag array on glass	Oligonucleotide array on glass
Reaction	ASPE	ASPE	SBE	SBE	Allele-specific hybridization
Labelling and detection	2-colour fluorescence	Biotin-avidin, single colour fluorescence	2-colour fluorescence	2 o 4-colour fluorescence	Biotin-avidin, single colour fluorescence
Multiplexing	from 384-1536	from 10 000 to hundreds of thousands	from 12-48 SNps in 384 samples/array	12 000 SNPs	—
SNP, sample size	3027 SNPs per array (≈ 110 000 SNPs / SAM)	Up to 500000 SNPs	Tens of SNPs, hundreds of samples per plate	10 000 SNPs	Up to 500 000 SNPs

Abbreviations: ASPE, allele-specific primer extension; SAM, Sentrix Array Matrix; SBE, single-base extension. Table modified from Gupta *et al*., (2008).

These array based platforms are under constant improvement and used for high throughput variant discovery and genotyping, but the low cost of NGS technologies may replace these array based marker systems as it is already happening. Moreover one of the main drawbacks of SNPs arrays is that the development of new markers requires significant investment and usually they are developed in specific populations, resulting in an allelic bias that can be highly problematic when applying the array to divergent populations. However, through the sequencing of a large number of individuals within the same specific species it is possible simultaneously discover sequence variations and scoring the genotype. This new approach is called genotyping-by-sequencing (GBS) and allows the simultaneous rapid and direct study of the species diversity and the mapping of a trait or an interesting mutation. An extended and detailed lecture about the existing arrays based platforms can be found on (Gupta *et al*., 2008; Gupta *et al*., 2013; Ragoussis, 2009).

I.3. SEQUENCING TECHNOLOGIES

DNA sequencing technologies have suffered enormous improvement during the last thirty years, becoming in a faster, more accurate, easier to manage and cheaper technology.

I.3.1 First generation sequencing

Looking back on the history of sequencing technology we find the first generation Sanger or dideoxy sequencing technique (Sanger *et al*., 1977). This method is based on DNA chain terminators or dideoxy nucleotide triphosphates (ddNTPs) where the fragments obtained in four reactions (one for each base) are separated by size using electrophoresis gels. Later on, in the 1990s Walter Gilbert included an improvement on the technique, consisting on the incorporation of different colored fluorescent dyes, emitting light at different wavelengths, to label each ddNTP terminator allowing to obtain the DNA sequencing fragments in a single reaction (Prober *et al*., 1987; Hunkapiller *et al*., 1991). Later, PAGE (PolyAcrylamide Gel Electrophoresis) was replaced by capillaries (Swerdlow *et al*., 1990), increasing read lengths. As a result, the combination of dye terminators sequencing, capillary separation and computer driven laser detection of DNA succeeded (Madabhushi, 1998). Since then the improvement in machinery has been constant. Then, the invention of automated sequencing instruments led to the initial sequencing of the human genome genome project in 1998 (Lander *et al.,* 2001). Nowadays, the 3500 Genetic Analyzer from Applied Biosystems (AB) with up to 24 capillaries produces read lengths of 1000bp.

Along these last two decades numerous methods have been developed to improve the high throughput sequencing pipelines, such us whole genome shotgun (WGS) approach or strategies of subgenome sample pooling of YAC, BAC and cosmid based on physical maps of individual loci and entire chromosomes (this strategy was mainly used by the International Human Genome Project team). Despite the fact that Sanger methodology is still considered as 'the gold standard' for sequencing and it is still widely applied, shows several limitations. The main one is the cost. Even for a relatively small genome, the cost would be very high. Others limitations are the very low-throughput and the excessive time needed, the difficult analysis of allele frequencies and finally the difficulty of the novo assembly of repeats without high resolution physical maps (Men *et al*., 2008). Although this last limitation is also present when using the next generations sequencing technologies (NGS).

I.3.2 Next generation sequencing (NGS)

Currently 5 second generation and 4 third generation platforms are available. The 454 sequencer from 454 Life Sciences was created in 2005 as the first commercial NGS platform, later on, in 2007 the company was acquired by Roche. It uses a picotiter plate where each well can hold a single bead with a single DNA molecule attached that would be amplified via emulsion PCR. The picotiter plate can hold millions of beads, which will be sequenced in parallel by pyrosequencing (Margulies *et al*., 2005). In October 2013, Roche announced that it will shut down 454, and stop supporting the platform by mid-2016. Solexa (acquired by Illumina) released the second NGS commercial platform. The technology that Illumina follows is the Sequencing by Synthesis (SBS), which is explained in detail in the next section. Then, the third platform was SOLiD developed by Invitrogen, which was acquired by Applied Biosystems (AB), forming Life Technologies. It uses ligation as sequencing technology. Helicos developed HeliScope, being the first commercial single-molecule sequencer, but currently it survives as service center due to the high cost of its machinery. Ion Torrent was released in 2010; it is based on semiconductor sequencing by synthesis. It consists in something like 454 technology but in this case hydrogen ions are detected instead of pyrophosphate. It uses microchips with different output data capabilities. No laser, cameras or fluorescent dyes are needed, so the cost is very low. Also in 2010 PacBio developed the first platform that allows sequencing single DNA molecule in real time. It uses microscope slides where individual DNA polymerases are bounded. Individual DNA strands are determined because each dNTP has a unique fluorescent label that is detected prior to being cleaved off. StartLight is quite similar to PacBio but uses quantum dots for single-molecule sequencing. The main advantage is that DNA polymerases can be replaced when they have lost their activity (Karrow, 2010).

The first three platforms mentioned before are the preferred choices for whole genome sequencing due to their cost-efficiency. These platforms have split their focus between long reads in the case of 454 or more short reads in the case of Illumina and SOLiD. Longer reads will be useful for the novo assembly of genome and transcriptome characterization while a higher amount of shorter reads will be suitable for re-sequencing and for frequency method analysis. An overall comparison between the currently most used next generation sequencing platforms (second and third generation) is shown in **Table I.3**.

Table I.3. Comparison of the main next generation sequencing platforms. Modified from *(Liu et al.*, 2012) and complemented with data from (Glenn, 2011; Kircher & Kelso, 2010; Moorthie *et al*., 2011). Labeled by an asterisk are the sequencer considered as the third generation platforms.

Sequencer	Sequencing mechanism	Read Lenght (pb)	Time/run	Advantages	Disadvantages	Released Year
Roche 454 GS-FLX	Pyrosequencing	700	24h	Read length	Homopolymer error, high cost, low throughput	2008
Illumina HiSeq2000	SBS	50 SE, 50PE, 101PE	3-10d	cheap, high throughput	Short read assembly	2010
SOLiD 5500xl	AB Ligation and two base coding	2x50 MP 50x25PE	MP 11d PE 12d	High throughput	Short read assembly Homopolymer error	2007
Polonator* G.007	Ligation and two base coding	26	5d	Cheap, open source software	Short read lenght	2008
Life Technologies Ion Proton	semiconductor SBS	200	2-4h	Low cost instrument, low cost, low run time	Homopolymer error rate	2012
Helicos BioScience HeliScope	SBS-SMS	25	8d	No PCR, less bias and error	Low throughput, error rate	2008
Complete* Genomics (CG)	Probe anchor hybridization and Ligation	2x35	ND	High throughput, accuracy	Not commercialized	2009
Pacific BioScience* PacBio RS	SMS-RT fluorescent signal	10000	2h	No PCR, read lenght, speed	High error rate low throughput	2010
Oxford* Nanopore	SMS-RT electric current signal	2x50000	24h tens of Gb	No PCR, long reads speed, not optics	High systematic error indel error	2012

I.3.3 More advanced sequencing technologies

The new generation sequencing is based on single molecule sequencing (SMS) as the performed by PacBio or StartLight described above. This technology offers many advantages over the past and current technologies: higher throughput, faster run times, longer read lengths, higher accuracy, and small amount of starting material and low cost.

The SMS technology can be divided in four categories (Schadt *et al*., 2010): (a) Sequencing by synthesis (SBS) in which a single molecule of DNA is imaged as the molecule is synthesized; PacBio is an example of this sequencing technology. (b) Real time DNA sequencing by fluorescence resonance energy transfer. This is the promising approach of Life Technologies which hopes to improve Helicos technology. (c) Tunneling and transmission-electron-microscopy for DNA sequencing, which can be done by a direct capture of images of DNA using electron microscopy and direct imaging of DNA sequencing using scanning tunneling microscope tips. Halcyon Molecular is the first company approaching this method. (d) DNA sequencing with Nanopores with *Mycobacterium smegmatis* PorinA (MspA) or with optical readout or direct electrical detection with transmission-mediated DNA sequencing.

In conclusion, the current and future main goal of next generation sequencing technology is the production of entire genome sequences in less time and at reasonable cost and the increase of their applications in the biological and biomedical science.

I.3.4 Sequencing by Synthesis: Illumina Technology

In this section is described the sequencing technology used in this thesis to study genome-wide somatic variability (see chapter III). The sequencer used was Illumina HiSeq2000, which uses the technology of sequencing by synthesis (SBS). The workflow starts by library preparation (**Fig. 5**) in which genomic DNA is fragmented into 100-500 base pairs fragments by sonication. This creates flayed DNA ends which must be blunted or repaired, ending up with 5'-phosporilated ends. Then, adenine is added to each 3'end, pair ends adapters are ligated to each end of the A-tailed DNA fragment. There are two types of libraries depending on the sequencing strategy; adapter configurations will be specific of each kind of library (**Fig.I.6**). For those reads sequenced just from one end, the library must be single end, while when reads are sequenced from both ends, the library must be paired ends. Furthermore, there is the option of multiplexing, and for that an additional barcode or index is included in the adapter, called P7. Fragments of 200-600bp are size selected by gel electrophoresis; this method is labor intensive and lacks reproducibility. An alternative of gel size

selection it is the use of solid-phase reversible immobilization beads but they have the limitation that can result in a broad fragment size range. The ultimate alternative for size selection is the use of semi-automated preparative DNA electrophoresis systems such as Calipser Labchip XT (PerkinElmer) or Pippin Prep (Sage Science) (Quail *et al.,* 2012). Following size selection and clean up, libraries are amplified by PCR to enrich for properly ligated template strands. Then, quantification of the library is necessary in order to add an adequate concentration of each library that will result in a right density of clusters that will provide enough yield of data (Bronner *et al., 2*014). Then, before sequencing step, the library with fixed adaptors is denatured to single strands by sodium hydroxide. These adaptors have flow cell binding sites, P5 and P7, which allow the library fragment to attach to the flow cell surface (**Fig. I.6a**). Furthermore, adapters contain several other primer binding sites, depending on the library that is going to be used in the Illumina SBS process (**Fig. I.5**).

The flow cell oligos act as primers and a strand complementary to the library fragment is synthesized by 3' extension using a high fidelity DNA polymerase (**Fig. 6b**). The original strands are denatured, leaving behind fragments copies that will be covalently bounded to the flow cell surface in different orientations (**Fig. I.6c**).

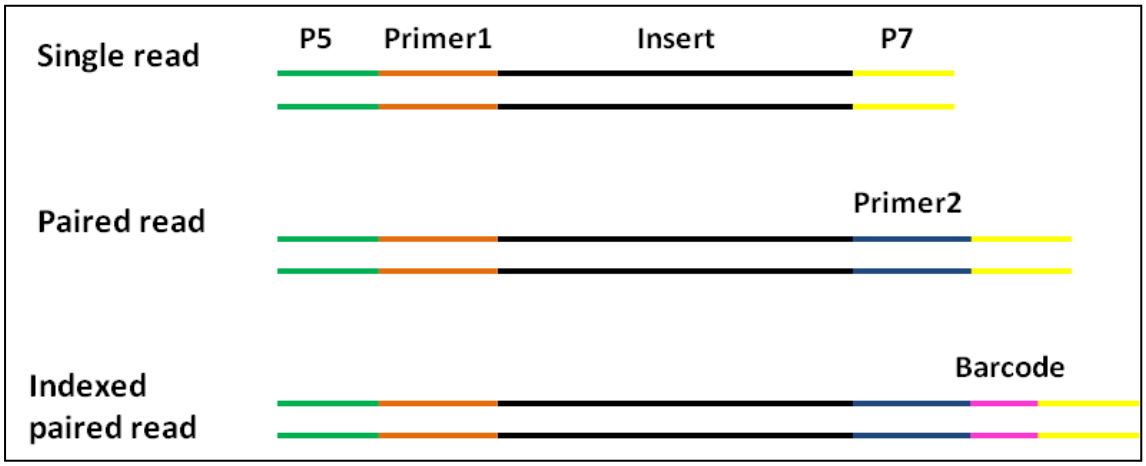


Figure I.5. Types of Illumina libraries. P5 and P7 are the Illumina adapters. In black the insert reads.Primer1: first sequencing primer and Primer2: second sequencing primer, just used when performing paired end sequencing. Barcode is a 6-bp index sequence sequenced when applying multiplexing.

Unlabeled nucleotides and enzymes will be then added to initiate solid-phase bridge amplification. DNA polymerase copies the templates from the hybridized oligonucleotides forming dsDNA bridges which are denatured to form two ssDNA strands (**Fig. I.6d**). These two strands loop over and hybridize to adjacent oligonucleotides and are extended again to form two new dsDNA loops. The process is repeated on each template by cycles of amplification and denaturation to

generate clusters containing 2000 molecules (**Fig. I.6e**). Each cluster of dsDNA bridges is denatured and the reverse strand is removed by specific base cleavage, leaving the forward DNA strand (fragments which are attached by P7 end) to ensure that all copies are sequenced in the same direction. The 3' ends of flow cell bound oligonucleotides and DNA strands are blocked to avoid any interference during sequencing reaction (**Fig. I.6f**). Then, sequencing primer is hybridized to P5 fragment end allowing for the sequencing by synthesis process (**Fig.I.6g**).

It has to be highlighted that there is the possibility of multiplexing, in which samples are uniquely tagged with short identifying sequences known as barcodes, pooled and then sequenced together in a single line. When the first read is finished, it is removed and an index primer is added, which anneals at the P7 end of the fragment and sequences the barcode (**Fig. I.6h**).

In the case of paired ends sequencing, the workflow that will perform the second read will form clusters by bridge amplification as in read one, leaving fragment copies bounded to the flow cell (**Fig. 6i**). In this method is the adaptor P7 the one that gets cut, producing clusters containing only fragments attached to P5 region (**Fig. I.6j**).This ensures that all copies are sequenced in the same direction (opposite to read one). Then, the sequencing primer anneals to the P7 region and sequences the other end of the template (**Fig. I.6k**).

During sequencing by synthesis all four labeled terminators and DNA polymerase enzyme are added. Only one base can be incorporated at a time, and each time the laser will excite the fluorescent tags and the images will be captured via CCD camera. The first base in each cluster is recorded and then the fluorescent tag is removed. In subsequent cycles, the process of adding sequencing reagents, removing unincorporated bases and capturing the signal of the next base to identify is repeated. Once the top surface of the flow cell channel has been scanned, the imaging step is repeated on the bottom surface enabling twice the number of reads compared to single surface imaging.

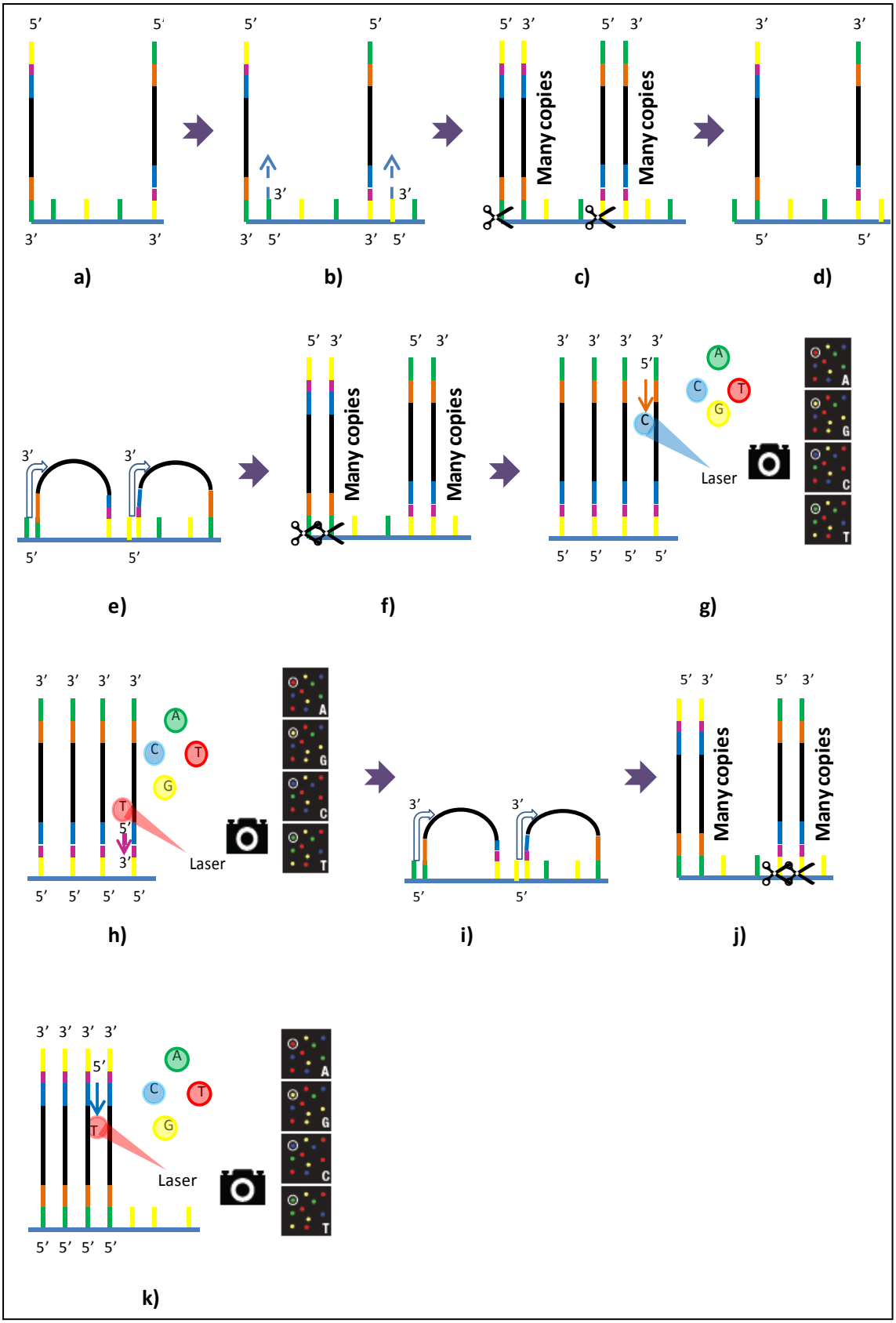


Figure I.6. Sequencing by synthesis Illumina technology workflow of a paired end library. The horizontal blue line represents the surface of the flow cell. The colors of the fragments are the same than previously mentioned in **Fig. I.5**. The process is explained step by step in the text.

I.4. MARKER ASSISTED SELECTION (MAS)

Plant breeding began with the domestication of crop plants. Since the development of agriculture 11,000 years ago in south-western Asia (Xu, 2010), plant domestication responded to the increase of population's size and changes in the exploitation of local resources. Men have been for thousands of years adapting plants and animals to their own needs producing the domestication of species. The successive selection of specific plants carrying traits or qualities desirable for consumers has been the usual procedure of the conventional plant breeding which has changed the genetic composition of the crop under consideration. In the intensive breeding, these evaluations are time consuming, especially in fruit trees since they have to overcome a juvenile period. Moreover they can be strongly influenced by: the environment, tissue sampled, developmental stage of the plant, the heritability of the trait, the number of genes involved, their effects and the way these loci interact. The use of molecular markers in (or linked to) the gene responsible for the trait in the selection process of the seedlings is known as marker assisted selection (MAS) and may overcome these limitations. One of the main advantages of MAS in fruit trees is that marker genotypes can be scored at early stage of the plant development.

Although the ideal marker is the one designed in the causal genetic variant (functional marker), they are difficult to obtain. Alternatively we can use markers in linkage with the allelic variant, which are more likely to find. The idea is that, once that linkage-based association between trait and marker has been proved to be accurate, the marker genotypes can be used as predictors of the phenotype. The establishment of those associations is the key question in MAS. So far, finding markers associated to major genes, with one or few causal alleles in one locus, is more straightforward than those associated to QTLs, where more than one locus interact.

The establishment of the association between marker and traits is usually done by QTL studies, through either linkage or association mapping (also called linkage disequilibrium LD mapping) or combining both strategies. Ideally the markers should work in all germplasm, however this is not always the case and they are population (or breeding program) driven, thereafter it is necessary to confirm that the associated markers are polymorphic and linked to the causal trait in the parental lines used in the breeding program and, consequently, in the offspring before they are used in MAS. This normally occurs for QTLs, which usually are controlled by many genes, but also for major genes where more than one causal allele may be responsible for the trait. In some cases the use of haplotypes of marker alleles is more efficient than single markers.

I.4.1 MAS applied to peach breeding

The existence of the densely covered reference map for *Prunus* (Joobeur *et al.,*1998) as well as the rest of the linkage maps available in peach (Aranzana *et al.,* 2003b; Blenda *et al.,* 2007; Chaparro *et al.,* 1994; Dettori *et al.,* 2001; Dirlewanger & Bodo 1994; Dirlewanger *et al.,* 1998, 2006; Lu *et al.,* 1998; Mammadov *et al.,* 2012) made possible the localization of many traits and have been an important tool for the application of MAS in this species. Additionally, the peach genome sequence has been released (Verde *et al*., 2013), consisting of a high quality whole-genome shotgun assembly of a double haploid genotype of peach cultivar 'Lovell' (Toyama, 1974) with an estimated size of 227Mb and 27,852 protein coding sequences. The genome sequence has provided high amount of useful information for genetics in peach. Furthermore, several peach varieties that have been sequenced (Ahmad *et al*., 2011) or are in the process of being sequenced. This, together with the application of the 9K SNP array in breeding germplasm (Verde *et al*., 2012) and all the transcriptomics (Chan *et al*., 2007; Nilo *et al*., 2012; Renaut *et al*., 2008) and metabolomics (Borsani *et al*., 2009; Lara *et al*., 2009) analysis conducted in peach are enabling the identification of polymorphisms between different varieties that may be associated with quality traits or the identification of candidate genes, of signal transduction and metabolic pathways that play an important role in fruit quality and production (Carrasco *et al*., 2013).

To date 28 major genes and 30 QTLs have been located on a single map for the *Prunus* genus (Dirlewanger *et al*., 2004), plus four additional genes recently mapped (Falchi *et al.,* 2013; Pascal *et al.,* 2010 ; Shen *et al.,* 2013, **Table I.4**). Most of the mapped traits are important for both the consumer and breeder preferences. Consumer acceptance relays on fruit quality traits such as flavor, texture, color and shape while growers are more focused on the acquisition of productive cultivars resistant to diseases and to get varieties with different harvest dates with long period of storability (Byrne *et al.*, 2012) . Although some markers linked to the major genes have been developed and are successfully applied in peach breeding, to date no MAS activities have been reported for any of the QTLs mapped in peach or other *Prunus.*

Currently public and commercial breeding programs apply MAS for few monogenic peach characters (Arús *et al*., 2012) like flesh softening (melting/non-melting, *M/m*) and flesh adhesion (freestone/clingstone, *F/f*), both controlled by two copies of the endogalacturonase gene (Peace & Norelli, 2009) located in the distal end of chromosome 4; fruit acidity, in which dominant allele *D* determines low acidity (*D*/*d* locus), flat shaped fruit controlled by a single gene (*Sh*/*sh*) where flat peach individuals have heterozygous genotype, fruit flesh color (yellow/white, *Y/y*) and skin glabrousness (*G/g*). For each of the four last traits SSR linked markers are available, however the

markers are not completely linked with the trait and they need to be validated in each breeding program.

Table I.4. Peach major genes affecting morphological or agronomic characters that have been mapped on the *Prunus* reference map.

Characters	LG	Gene	References
Affecting flower traits			
Double Flower (single/double)	2	*Dl*	(Chaparro *et al*., 1994)
Anther color (yellow/anthocyanic)	3	*Ag*	(Joobeur *et al*., 1998)
Flower color(pink/red)	4	*B*	(Jauregui,. 1998)
Flower color (pale pink/pink)	3	*Fc*	(Yamamoto *et al*., 2001)
Male sterility (fertile/sterile)	6	*Ps*	(Rodriguez *et al*., 1994; Scott.& Weinberger, 1944)
Flower morphology (showy/non-showy)	8	*ShF*	(Bailey &French, 1942; Fan *et a*l., 2010)
Affecting fruit traits			
Flesh color (white/yellow)	1	*Y*	(Falchi *et al*., 2013, Connors, 1920)
Flesh color around the stone (red/white)	3	*Cs*	(Yamamoto *et al*., 2001)
Recessive blood flesh	4	*Bf*	
Dominant blood flesh	5	*DBF*	(Shen *et al.,* 2013)
Polycarpel pistil (mono/poly)	3	*Pcp*	(Bliss *et al*., 2002)
Flesh adhesion (clingston/freestone)	4	*F*	(Yamamoto *et al*., 2001)
Non-acid	5	*D*	(Monet, 1979)
Skin hairiness (nectarine/peach)	5	*G*	(Blake,1932)
Kernel taste (bitter/sweet)	5	*Sk*	(Bliss *et al*., 2002, Werner & Creller, 1997)
Fruit shape (flat/round)	6	*S**	(Lesley, 1939)
Fruit skin color	6 and 8	*Sc*	(Yamamoto *et al*., 2001)
Maturity day (early/intermediate/late)	4	MD	(Pirona *et al*., 2013)
Affecting leaf traits			
Evergrowing (annual/perennial)	1	*Evg*	(Rodriguez *et al*.,1994; Wang *et al*., 2002)
Leaf color (red/yellow)	6 and 8	*Gr*	(Blake,. 1932)
Leaf gland (reniform/globose/eglandular)	7	*E*	(Dettori *et al*., 2001,Connors, 1920)
Leaf shape (normal/dwarf)	6	*Nl*	(Yamamoto *et al*.,2001)
Conferring resistance			
to root-knot nematode			
M.incognita	1 and 2	*Mi*	(Weinberger *et al*., 1943)
M.javanica	2	*Mj*	(Sharpe *et al*., 1970)
to powdery mildew			
S.panosa	7	*Sf*	(Dabov, 1983)
S.panosa	6 and 8	*Vr3*	(Pascal *et al*.,2010)
Affecting plant structure			
Plant height (normal/dwarf)	6	*Dw*	(Yamamoto *et al*., 2001)
Broomy (or pillar) growth habit	2	Br	(Scorza *et al*., 2002)

Despite of these few cases, MAS has a low impact in peach breeding, even though it is one of the fruit tree species where more conventional genetic improvement has been done and one of the fruit with the greatest annual varietal dynamism. The low availability of markers tightly linked to quantitative trait seems to be one of the main reasons of the low use of MAS in peach. Another reason is the still high cost of using MAS, which will be reduced with higher throughput markers like SNPs. Additionally, there is a low flow of information between researchers and breeders or in others words, between the conventional plant breeding and the molecular breeding. Two big projects, RosBreed (*www.*rosbreed*.org*) in USA and FruitBreedomics (http://www.fruitbreedomics.com) in Europe, are addressing this issue aiming to bridge the gap between research and breeding.

I.5. IDENTIFICATION OF CANDIDATE GENES (CGs)

A candidate gene (CG) is a gene thought to be the causal of phenotypic variation. Thus, a candidate gene can be any structural or regulator gene implicated in a metabolic pathway involved in the expression of trait being studied (Pflieger *et al.,* 2001). Such effect is usually deduced by its known biological function (functional CG approach) or by its proximity to markers or DNA fragments associated to the phenotypic trait variation (positional CG approach). This strategy is applied in three steps: selection, screening and validation.

The choice of candidate genes is done by their role in the phenotypic variation. Normally, when the biochemical and/or physiological pathways related to the trait are known, any gene involved in the pathways could be a good CG (functional CG). Few years ago, the limiting factor of this approach was the low availability of gene sequences. When this is the case, an option is to search collections of ESTs coming from cDNA libraries of different relative species or coming from different developmental stages or tissues, etc. Fortunately, in *Rosaceae* family there is available a well annotated database assembled to the sequenced genomes (Jung *et al.,* 2013) with 236,191 genes from which 27,864 are from *Prunus persica.*

Nowadays, computational, statistical approaches and omics data are used for inferring gene function in plants with an emphasis on network-based inference. Thus, *in silico* methods are developed for assistance in elucidating and annotating gene function. These methods are based on the integration of different kinds of omics data (mRNA expression, protein-protein interactions, genome sequences and genetic interactions) to build up co-function networks that are useful for inferring biological processes. This methodology is broadly known as Systems Biology. For further details read Rhee & Mutwil (2014).

Alternatively, candidate genes can be detected from QTL analysis by either linkage, comparative or LD mapping. The size of the QTL will determine the accuracy of the method since it will determine the number of candidate genes involved. Thereafter, reducing the confidence interval of the QTL will reduce the number of putative genes. This will require the increase of the sample size of the population.

Once the candidate genes have been identified and genetically or physically localized, further experiments will be needed to select the most probable CGs. Today, positional cloning relies on genotyping and phenotyping large numbers of progeny to detect chromosome recombination events that break linkage between the trait of interest and flanking molecular markers following meiosis. Nowadays, this strategy is no longer limited by the availability of high-density molecular markers but rather by the slow and intensive labour that implies the development of large segregating populations and their phenotyping and genotyping to detect rare recombination events in a narrow chromosome block flanking the target gene of interest (Lukowitz *et al.,* 2000). To overcome this drawback, correlation analyses between the phenotypic segregation and the polymorphisms within the CG can be established analyzing unrelated individuals (e.g germplasm collections). The studies based in this last procedure are called association mapping studies and represent a complementary approach to the classical QTL mapping. These studies relay on the extent of linkage disequilibrium (LD) which determines the marker density required for association mapping. Or in other words, the minimum number of loci required to scan the genome depends on the extent of the LD. The association between the phenotype and the genotype by genome wide association (GWA) approaches are applied when the LD declines slowly and/or large number of markers are available. In plant species the first application of this method was done in *Beta vulgaris* ssp. Maritime, a wild form of sugar beet (Hansen *et al.,* 2001). Recently in peach, GWAs was applied to analyze the association (Micheletti *et al.,* in preparation).

The use of collection of cultivars instead of bi-parental crosses for assaying the genotype-phenotype correlations shows some advantages. Firstly, broader genetic variation and genetic background more representative of the crop breeding potential will be available. This implies that maker and trait data is not limited to the one found between the two parents of a progeny. Secondly, LD-base mapping provides more resolution, because of the use of all the meiosis accumulated. This technique has been already applied in many crop plants (Huang & Han 2014) and it is expected to be extensive applied due to the increase of improvement and low cost in the high-throughput genotyping and sequencing technologies.

Even when it exits a statistic correlation between a gene and a QTL it is also necessary to perform complementary experiments to validate the participation of the CG in the variation of the trait. The complexity of these experiments will depend on the nature of the trait (mono or polygenic). Genetic transformation, virus induced gene silencing (VIGS), RNA-mediated interference (RNAi), insertional mutagenesis mediated by virus or transposons, fast neutron mutagenesis or chemical mutagenesis and TILLING are reverse genetic techniques that can validate the exact function of the CG by the disruption or modification of the gene or the gene product and measuring the phenotype (Gilchrist & Haughn 2010).

In *Prunus* the CG approach has been used to: identified the self-incompatibility locus in almond (Ushijima *et al.,* 2003), to identify genes linked to aroma volatiles in peach (Sánchez *et al.,* 2013), or linked to flowering time in almond (Silva *et al.*, 2005), candidate genes and QTLs for sugar and organic acid content in peach (Etienne *et al.,* 2002), candidate for evergrowing locus in peach (Bielenberg *et al.,* 2008) among many others. In the second chapter of this thesis it is explained a CG approach applied to peach fruit shape.

I.6. GENETIC DIVERSITY STUDIES IN PEACH

Several genetic diversity studies in peach have been addressed to preserve and to manage the available genetic resources in this species and to provide information and useful tools for breeders. Such studies have revealed low variability in commercial varieties in both Occidental (Aranzana *et al*., 2003a; Aranzana *et al*., 2010) and Oriental collections (Cao *et al*., 2012; Li *et al*., 2013) due to peach self-compatibility (Arulsekar *et al*., 1986; Byrne & McMahon, 1991) and because of the reduced parental material used in breeding programs. In peach the number of alleles per SSR locus in commercial varieties range from 2.9 to 7.3, with observed heterozygosity (H_o) between 0.21 and 0.46 (Carrasco *et al.,* 2013). These values contrast with the ones observed in other *Prunus* species, which are self-incompatible. For example more than 53-74% of SSR loci in plum and sweet cherry were heterozygous, with an average of 4.1 to 12.1 alleles per locus (Carrasco *et al.,* 2013).

Most of the currently commercialized occidental peach varieties are in some extent related to those obtained in the early USA breeding programs which rely on few high quality varieties, producing a variability bottleneck. New sources of variability can be found on germplasm from different origins or from wild relatives. One of these valuable resources is the Chinese material, because as a center of origin should have greater levels of variability (Vavilov, 1926). In addition, a

recent analysis based on the comparison of genetic diversity, population structure and LD between Oriental and Occidental accessions using the same molecular markers has revealed that although Chinese landraces have greater levels of variability, the varieties obtained in breeding programs have suffered a reduction of variability similar to the one occurred in Occidental countries although with a different genetic background, suggesting that they both can complement each other (Li *et al*., 2013). Crosses between peach elite varieties and other close related *Prunus* species can also be used to increase peach variability and at the same time to introduce genes not described in peach like some resistance genes (Foolad *et al.,* 1995).

Additionally, peaches present considerable large levels of somatic variability producing, in some cases, observable phenotypes. This is the case of the yellow flesh peach 'Redhaven' a somaclonal mutant of the white flesh peach 'Redhaven Bianca' (Brandi *et al.,* 2011), or the case of nectarines mutants from peaches like 'Yuval' from 'Oded' (Dagar *et al.,* 2011) among many other cases reported in bibliography (Mase *et al.,*2007; Scorza & Sherman 1996; Shamel 1938; Stoner 1948; Yamamoto *et al.,* 2003).

In these two cases fruits with different phenotypes grow at the same time in the same plant. Plants showing adjacent cells with more than one genotype are called chimeric. The mode of spreading and spatial arrangement of the mutant cell lineage results from the layered structure of the shoot apical meristem and ordered orientation of cell division. In a typical angiosperm shoot meristerm, the core tissue or corpus is covered by two tunica layers. The number of tunica layers may vary (from only one to three or even more) among species and at different stages of development on the same species (Schmidt, 1924). In, peach like most woody plants, the meristerm is composed by three histogenic layers. The L-I layer gives rises a single-layered epidermal tissue, but also produces several layers of cells at the suture of the ovary wall (formed by a wide band of cells from the L-I and L-II layers), seven or eight cell layers of the nucellus at the micropylar end of the ovule and almost all the integuments. The L-II layer which is one or two-cell thick layers, is located bellow the L-I and it produces subepidermal tissues such us: the outer cortex and part of the vascular cylinder, also the petals, anthers and ovules. The L-III produces the inner cortex, vascular cylinder and pith (Dermen & Stewart 1972). But, L-III also participates in the formation of pistil, contributing to the central region of the ovary, and to the stylar region but not the stigma. The controlled pattern of cell divisions in the tunica results in the maintenance of discrete layers which organization is retained in leaves, lateral buds and fruits. Layer LII and LIII produce cells both by anticlinal and periclinal mitosis, while, LI only shows anticlinal divisions. Moreover, cells originating

from different layers are distinguished not only by their division plan, but also by size, vacuolization and proliferative speed (Szymkowiak & Sussex 1996).

The spontaneous somatic mutations observed in nature can have three possible chimeric conditions 1) periclinal chimeras, when the mutation occurs in one layer and through mitosis the mutant cells are gradually driving out of wild type, this is to be expected only when the mutation is advantageous compared with wild type. More frequently, periclinal chimeras develop, when a lateral bud originates from within the sector bearing the mutated tissue layer producing that one entire meristematic layer(s) is different from the other two 2) mericlinal chimera, when only one part of the layer contains mutant cells, and 3) sectorial chimera, when the mutation can affect sections of the apical meristem extending through all the cell layers. Mericlinal chimeras are unstable and tend to lose the mutated tissue or develop into stable periclinal chimeras. Because they may appear phenotypically similar, mericlinal chimeras are sometimes confused with sectorial chimeras. Periclinal mutations are relatively stable and can be vegetatively propagated while the sectorial ones are unstable and can give rise to shoots and leaves which are not chimeras (Geier, 2012).

OBJECTIVES

The broad goal of this thesis was the identification of genetic markers and genes associated to interesting agronomic traits in peach.

The specific objectives were:

1. Conduct candidate-region association analysis approach to:

    1.1. Identify molecular markers, especially SNPs, linked to low acidity and flat shape in peach.
    1.2. Identify candidate genes for low acidity and flat shape in peach.

2. Identify and validate the causal mutation for the flat shape trait in peach.

3. Analyze whole genome sequences of 6 pairs of peach sport mutants to:

    3.1. Obtain a first estimate of qualitative and quantitative peach somatic variability.
    3.2. Identify the causal mutation(s) responsible for the glabrous phenotype (nectarine) in peach fruits.

This thesis contains the following sections: a general introduction, objectives, three chapters (formatted as scientific paper each with introduction, material and methods, results and discussion), a general discussion, conclusions and, finally, the references. The first chapter has been already published in the journal “Tree Genetics and Genomes” (DOI: 10.1007/s11295-014-0789-y) and the second will be submitted (with some additional analysis) for publishing.

CHAPTER I: Development of diagnostic markers for selection of the subacid trait in peach

I. Eduardo*, E. López-Girona*, I. Batlle, G. Reig,
I. Iglesias, W. Howad, P. Arús and M. J. Aranzana

Tree Genetics & Genomes, DOI 10.1007/s11295-014-0789-y, *published online: 30 August 2014*

http://link.springer.com/article/10.1007%2Fs11295-014-0789-y

*I. Eduardo and E. López-Girona contributed equally to this work

CI.1 ABSTRACT

Peaches with low acidity are preferred in the market and this trait is usually selected in commercial breeding programs. A major gene (*D*/*d*) has been described for this character located on linkage group 5 of peach, where the low acid character is determined by the dominant *D* allele. In this paper, we analyze a collection of 231 varieties and 542 offspring to identify diagnostic markers for this character. The CPPCT040 single sequence repeat (SSR) is known to be tightly linked to *D*. We found that one of its alleles (193) is diagnostic for the subacid character and identified with high probability individuals with low acidity (titratable acidity <5.5 mg/l). The region around CPPCT040 was explored using 13 DNA fragments for a total of 5,297 bp, covering a length of 70.4 kbp of the peach genome. The sequenced fragments detected 19 single nucleotide polymorphisms (SNPs) and five INDELs. All subacid individuals shared a large haplotype (>24 kb) around CPPCT040, a region with higher than average SNPs between acid and subacid varieties. The CPPCT040 marker plus one of the SNPs identified (DH875) were used to genotype a collection of 542 seedlings, from different crosses expected to segregate for this character, which were phenotyped by tasting the fruit in the field. Data provided by both markers were always consistent and only 24 plants (4%) did not fit the expectations. These markers and others that can be obtained from the haplotype identified can be readily used for marker-assisted selection in peach breeding.

CI.2 KEYWORDS

Peach breeding, Acidity, Subacid trait, Marker-assisted selection and Peach variability

CI.3 INTRODUCTION

Peach [Prunus persica (L.) Batsch)] cultivars exhibit considerable phenotypic variability in tree phenology, production, and fruit morphology and quality, despite sharing a narrow genetic background due to the self-compatible mating behaviour and to a bottleneck produced by the use of few parents in the early breeding programs (Scorza *et al.,* 1985; Faust and Timon 1995).

Many new cultivars of peaches and nectarines are released every year (Sansavini *et al.,* 2006; Iglesias 2013), to cover the widest production period and provide a large diversity of fruit appearance, texture, and flavor and to allow for the maximum possible shelf life. To meet consumer's acceptance with respect to flavor, new released peaches and nectarines have a wide range of soluble solid concentration (SSC) and titratable acidity (TA) at harvest (Iglesias *et al.,* 2005;

Crisosto and Valero 2008). Studies have reported a high relationship between the sugar-to-acid (SSC/TA) ratio and consumer acceptance (Crisosto and Crisosto 2005; Iglesias and Echeverria 2009). Consumer acceptability increased with higher values of SSC in low acid varieties and increased with SSC initially but reached a plateau in acid varieties (Crisosto and Crisosto 2005), while low acid nectarines were always preferred by consumers irrespective of their SSC content (Iglesias and Echeverria 2009).

Peach acidity is mainly determined by the content of malic (the most abundant), citric and quinic acid (Reig *et al.,* 2013) and is usually measured as pH and TA, which are negatively correlated (Cantín *et al.,* 2009; Abidi *et al.,* 2011). Many of the most successful new peach cultivars have low levels of acidity (Iglesias and Echeverría, 2009), also called subacid or non-acid, as low acidity produces a higher sugar-to-acid ratio, resulting in greater consumer satisfaction (Iglesias and Echeverría 2009). The TA value to classify cultivars into acid and subacid class has not been clearly established. In Spain, Italy and France the commercial classification includes five groups based on TA (meq/l or g acid malic/l): subacid, <50/<3.3; sweet/semisweet, 50-90/3.3-6.0; balanced, 90-120/6.0-8.0; acid, 120-150/8.0-10, and very acid, >150/>10 (Iglesias and Echevarria 2009). Boudheri *et al.,* (2009), after analysis of 1,718 genotypes according their pH and TA levels established a threshold strategy for these two parameters preventing misclassification of individuals (either D/d or d/d) to positionally clone the D gene.

The subacid character is inherited as a single dominant gene *D/d* (D from 'doux', the French word for sweet), the dominant allele of which, *D*, determines subacid fruit with pH>4.0 (Yoshida 1970, Monet 1979). The *D/d* gene was first mapped at the proximal end of linkage group 5 (Dirlewanger *et al.,* 1998; Dirlewanger *et al.,* 2006). Further results in a large set of seedlings allowed fine-mapping of the *D* locus to a region of 0.4cM (Lambert *et al.,* 2009; Boudehri *et al.,* 2009). The CPPCT040 marker in this region has been developed in our lab and is known from previous data to be associated with the acidity trait (Lambert *et al.,* 2009; Boudehri *et al.,* 2009).

The aim of this work is to develop diagnostic markers (i.e., markers for which the presence of a certain allele or alleles in any individual allows prediction of a phenotype with high probability) for marker assisted selection (MAS) for the subacid gene. We used a large collection of peach cultivars and progenies phenotyped for TA to evaluate the single sequence repeat (SSR) CPPCT040 and then, using information from the peach whole genome sequence (http://www.rosaceae.org/species/prunus_persica/genome_v1.0), we developed a set of single-nucleotide

polymorphisms (SNPs) in this region. A large conserved DNA fragment associated to the subacid character was identified, compatible with a recent introgression of this character from a single origin. The subacid character is typically selected in commercial breeding programs, so having diagnostic markers such as those we developed and validated in this paper may have an immediate application for cross design and early seedling selection.

CI.4 MATERIAL AND METHODS

CI.4.1 Plant materials and DNA extraction

A collection of 231 peach cultivars (**Table CI.1**) and 542 seedlings derived from 34 peach crosses (involving at least 44 different parents) from the IRTA-ASF peach breeding program was used in this study. For most of the crosses, the male genitors were either unknown or not confirmed. Female genitors were included in the list of cultivars. Three trees of each cultivar were grown at the IRTA Experimental Station in Lleida (Spain) on GF-677 INRA rootstock, trained as central axis and with a spacing of 4.5 m×2.5 m. Seedlings were grown for 3 years in selection plots (3.5 m×0.8 m) before fruiting and then phenotyped according to breeding aims. Records were taken from unselected families.

DNA of each individual was extracted from young leaf tissue following the Doyle and Doyle (1990) protocol adapted to 96-well plates (DNeasy 96 Plant mini Kit, Qiagen, Valencia, CA, USA).

CI.4.2 Acidity phenotyping

From each plot of three trees, two trees per cultivar were selected, based on uniformity of tree size and crop load. For each cultivar, titratable acidity (TA) was measured in the juice of a sample of 28 fruits collected from the periphery of the tree canopy at 1.5-2.0 m above ground level and representative of the cultivar at maturity (fruit firmness from 4 to 5 kg using 8- mm-diameter plunger tip penetrometer). TA was measured by titrating 10 ml of the juice with 0.1 N NaOH to pH 8.2 with 1 % (v/v) phenolphthalein, and the results were recorded as grams of malic acid per litre. TA values were obtained over 1 to 12 years, in the period between 1997 and 2010. The 542 seedlings from the breeding families were phenotyped by taste, as is normal in breeding programs, classifying the individuals as subacid or acid.

CI.4.3 SSR and SNP genotyping

CI.4.3.1 SSRs

All 773 individuals were genotyped with the CPPCT040 SSR marker. PCR reactions and fragment separation with the ABI/Prism 3130xl automated sequencer (PE/Applied Biosystems) were as described in Aranzana *et al.,* (2003).

CI.4.3.2 Sequencing

Using the peach genome sequence produced by the International Peach Genome Initiative (Verde *et al.,* 2013), (http://www.rosaceae.org/species/prunus_persica/genome_v1.0, http://www.phytozome.net/peach), we selected 13 DNA fragments corresponding to coding and non-coding regions of a 96.3kb chromosomal fragment flanking CPPCT040. Most of the fragments were chosen close to CPPCT040 as we considered that this was the most probable location of the D gene and only a few in the extremes of this interval. To visualize the relative position of the fragments, we used DNAplotter software (Carver *et al.,* 2009).

Specific primer pairs were designed for each region using Primer3 software (Untergrasser *et al.,* 2012; http://bioinfo.ut.ee/primer3-0.4.0/) to amplify fragments of about 450 bp, avoiding amplification of SSR motifs. The primers were first tested in six peach varieties, three subacid, with TA≤3.7 g/l ('Paraguayo Delfin', 'Douceur', and 'Gratia'), and the other three acid, with TA≥6.9 g/l ('Dolores', 'Glenna', and 'August Red').

Sequencing reactions were carried out as in Aranzana *et al.,* (2012) and visualized and manually edited with Sequencher 4.8 software (Gene Codes Corporation, Ann Arbor, MI, USA). Fragment ends were trimmed to remove low-quality sequence. Among the analyzed sequences, the nine primers yielding high-quality, unique, and polymorphic sequences were selected (**Table CI.2**) and used in 32 additional varieties. SNP genotypes were graphically visualized with Flapjack software (Milne *et al.,* 2010). The 38 peach varieties used for SNP detection were additionally genotyped with 17 SSRs (**Appendix CI.1**) for population structure analysis. The SSRs were selected for being highly polymorphic in peach germplasm (Li *et al.,* 2013) and distributed along the 8 *Prunus* linkage groups separated about 10–20 cM to prevent from being in linkage disequilibrium.

CI.4.3.3 High-resolution melting

Two primers were designed to genotype one of the SNPs found to be linked with the trait in a larger set of 63 varieties, using high-resolution melting (HRM) (**Table CI.1**). These primers, 233-0875 F (5'-AGACGAGTGATATATCAGAT-3') and DF0875R (see **Table CI.2**), were used to amplify a single product of 106 bp containing the variant "A" in the acid allele and "C" in the subacid one. PCR was in a total volume of 10 µl containing 20 ng of template DNA, 2.5 mM MgCl2, 300 nM forward and reverse primers, and 1× HRM master mix (Roche Applied Science). Both PCR and HRM were performed using a Roche LightCycler® 480 (Roche Applied Science). For the PCR parameters, we used an initial denaturation step of 95 °C for 10 min, followed by 45 cycles of 95 °C for 10 s, 57 °C for 15 s, and 72 °C for 15 s. Following amplification, the samples were heated to 95 °C for 1 min and then cooled to 40 °C for 1 min. Melting curves were generated with continuous fluorescence acquisition during a final slope from 65 to 95 °C at 1.1 °C/s, and the resultant fluorescence data were processed using the LightCycler480® software (version 1.5.0.39, Roche Applied Science).

CI.4.3.4 Linkage analysis

Linkage between CPPCT040 and D was evaluated in 542 seedlings of the breeding populations using the Kosambi mapping function $d=-1/2\ln(1-2p)$ $(0\leq p\geq 0.5)$, where p is the observed recombination fraction and d is the genetic distance. For some of the crosses, the male parental was either unknown or not confirmed; consequently, we treated the seedlings as open pollinated.

CI.4.3.5 Population structure

Population structure was studied with the Structure v.2 software (Pritchard *et al.,* 2000), running the program under the admixture model assumption with correlated alleles. Five independent repeats of each assumed number of subpopulation (K), ranging from 1 to 15, were run using 1,000,000 interactions after a burn-in of 100,000. The final number of populations was assessed using the ad hoc statistic ΔK based on the rate of change in the log probability of data between the successive K values (Evanno *et al.,* 2005). Varieties were assigned to a subpopulation when their membership coefficient was higher than 0.8.

CI.4.3.6 Association test

The association of CPPCT040 alleles with TA levels was evaluated through logistic regression using the generalized linear model (GLM) procedure in R (R Core Team 2013). The coefficients given by the model were used to calculate the probability of finding CPPCT040 alleles at

different TA levels using the formula $P\ (\mathrm{CPPCT040}^{\mathrm{allele}}/\mathrm{TA}) = e^{\beta_0 - \beta_1 * \mathrm{TA}}/1 + e^{\beta_0 - \beta_1 * \mathrm{TA}}$, where β_0 and β_1 are the estimated regression coefficients.

Association between the CPPCT040 marker and its alleles with the acid and subacid traits was further confirmed using Pearson's χ^2 test, where expected counts of the contingency table are determined under the assumption of independence between genotype and trait.

CI.5 RESULTS

CI.5.1 Association of SSR marker CPPCT040 with TA levels

A collection of 231 peach varieties was evaluated for titratable acidity (TA) and genotyped with the CPPCT040 marker. TA results and CPPCT040 genotypes are shown in **Table CI.1**. The 231 cultivars had six different alleles, with frequencies ranging from 0.002 (for the unique allele CPPCT040195 in 'Babygold-7') to 0.68, with an average frequency of 0.17. Allele sizes differed in 2 or a multiple of 2 bp, being compatible with the 2 bp repeated motif (CT) of this SSR marker. The alleles were combined in 13 different genotypes (**Fig. CI.1**).

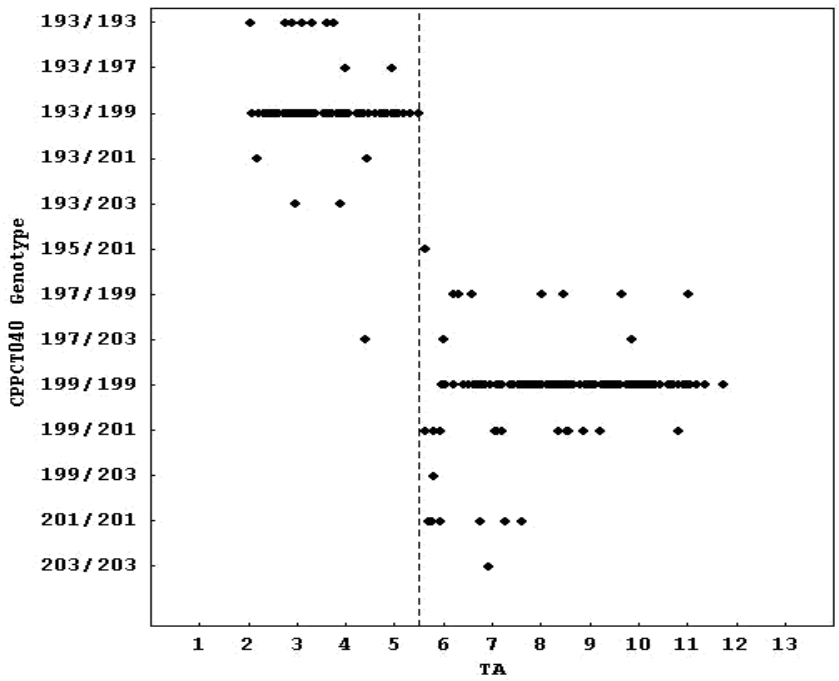


Figure CI.1 CPPCT040 genotypes observed in 231 peach varieties and their corresponding TA (g/l) values.

TableCI.1 Cultivar information and CPPCT040 genotype of the 231 cultivars analyzed.

Cultivar([1])	Fruit type	Origen	Mean Acidity ± SD (g/l)	n([2])	CPPCT040 genotype
ASF 04-92*	PWMF	Maillard, France	2.03 ± 0.32	4	193/193
ASF 02-27	NWMR	Maillard, France	2.08 ± 0.00	1	193/199
Redwing	PWMR	Armstrong Nurseries, USA	2.17 ± 0.44	6	193/201
Paraguayo Delfin*	PWMF	TC, Spain	2.22 ± 0.49	6	193/199
Royal Prince*	PYMR	Zaiger's Genetics, USA	2.31 ± 0.49	6	193/199
UFO7	PYMF	CRA-Roma, Italy	2.35 ± 0.14	3	193/199
ASF 04-06	NYMR	Maillard, France	2.41 ± 0.34	2	193/199
ASF 04-81*	PYMF	Maillard, France	2.41 ± 0.28	3	193/199
UFO4	PWMF	CRA-Roma, Italy	2.44 ± 0.50	6	193/199
ASF 02-87*	PWMF	Maillard, France	2.48 ± 0.23	5	193/199
ASF 04-94	PWMF	Maillard, France	2.49 ± 0.28	3	193/199
ASF 02-86	PWMF	Maillard, France	2.50 ± 0.39	6	193/199
ASF 05-56	PWMR	Maillard, France	2.52 ± 0.62	4	193/199
UFO3	NWMF	CRA-Roma, Italy	2.54 ± 0.63	6	193/199
ASF 04-52	PWMR	Maillard, France	2.55 ± 0.62	3	193/199
Kevina*	PWMR	Zaiger's Genetics, USA	2.59 ± 0.51	6	193/199
PG3/719*	PYMR	A. Minguzzi, Italy	2.59 ± 0.66	4	193/199
Platibelle	PWMF	INRA-QN, France	2.61 ± 1.01	3	193/199
UFO8	PYMF	CRA-Roma, Italy	2.62 ± 0.06	3	193/199
ASF 05-93	PWMF	Maillard, France	2.73 ± 0.41	4	193/199
ASF 03-64	PWMR	Maillard, France	2.74 ± 0.44	3	193/199
ASF 06-88	NWMF	Maillard, France	2.75 ± 0.00	1	193/199
White Lady*	PWMR	Zaiger's Genetics, USA	2.75 ± 0.47	5	193/199
ASF 02-80*	PWMF	Maillard, France	2.76 ± 0.76	6	193/193
Grenat	PYMR	Monteux-Caillet, France	2.80 ± 0.31	7	193/199
M-104	PYMR	Zaiger's Genetics, USA	2.87 ± 0.34	4	193/199
Gratia*	PWMR	Zaiger's Genetics, USA	2.90 ± 0.00	1	193/193
IFF0331	PWMR	CRA-Forli, Italy	2.93 ± 0.58	4	193/199
Fidelia*	PWMR	Zaiger's Genetics, USA	2.93 ± 0.48	5	193/199
Extreme Sweet	NWMR	Cabal, Spain	2.94 ± 0.33	4	193/199
UFO9	PWMF	CRA-Roma, Italy	2.96 ± 0.81	3	193/203
ASF 04-13	NWMR	Maillard, France	3.02 ± 0.79	4	193/199
Luciana*	NYMR	PSB, Spain	3.06 ± 0.20	5	193/199
Honey Glo*	NYMR	Zaiger's Genetics, USA	3.08 ± 0.28	5	193/193
Sweetlove	PWMR	Maillard, France	3.09 ± 0.76	4	193/199
Royal Glory*	PYMR	Zaiger's Genetics, USA	3.10 ± 0.60	5	193/199
IFF1233	PYMR	CRA-Forli, Italy	3.13 ± 0.21	4	193/199
ASF 01-81	PWMF	Maillard, France	3.15 ± 0.71	6	193/199
Extreme July	PYMR	Cabal, Spain	3.18 ± 1.47	4	193/199
ASF 04-14	NYMR	Maillard, France	3.20 ± 0.23	3	193/199
ASF 05-20*	NYMR	Maillard, France	3.27 ± 0.00	1	193/199
Platifun	PWMF	INRA-QN, France	3.30 ± 0.28	3	193/193
EP 93.06	PWMF	Maillard, France	3.32 ± 0.47	5	193/199
ASF 04-93	PWMF	Maillard, France	3.34 ± 0.55	2	193/199
ASF 03-28	NWMR	Maillard, France	3.35 ± 0.34	2	193/199
IFF1180	PWMF	CRA-Forli, Italy	3.52 ± 1.10	4	193/199
ASF 01-03	NYMR	Maillard, France	3.53 ± 0.92	3	193/199
ASF 02-08	NYMR	Maillard, France	3.55 ± 0.41	6	193/199
ASF 05-03	NYMR	Maillard, France	3.59 ± 0.34	3	193/199
ASF 06-12	NYMR	Maillard, France	3.60 ± 0.39	2	193/193
ASF 03-81*	PWMF	Maillard, France	3.62 ± 0.30	5	193/199

Cultivar([1])	Fruit type	Origen	Mean Acidity ± SD (g/l)	n([2])	CPPCT040 genotype
ASF 04-04	NYMR	Maillard, France	3.63 ± 0.53	4	193/199
Douceur*	PWMR	Maillard, France	3.71 ± 1.02	6	193/199
PG3/1312	NYMR	A. Minguzzi, Italy	3.72 ± 0.33	4	193/199
ASF 06-90	PWMF	Maillard, France	3.75 ± 0.39	3	193/193
ASF 04-71	NWMF	Maillard, France	3.82 ± 0.36	5	193/199
ASF 07-78	PWMF	Maillard, France	3.86 ± 0.00	1	193/199
Honey Royale	NYMR	Zaiger's Genetics, USA	3.86 ± 0.91	5	193/203
NG4/720	NYMR	A. Minguzzi, Italy	3.96 ± 0.26	4	193/199
M-110	PYMR	Zaiger's Genetics, USA	3.97 ± 0.88	3	193/197
ASF 04-30	NWMR	Maillard, France	3.99 ± 0.42	5	193/199
ASF 03-21	NWMR	Maillard, France	3.99 ± 0.51	6	193/199
ASF 01-04	NYMR	Maillard, France	3.99 ± 0.58	4	193/199
ASF 99-02	NYMR	Maillard, France	4.01 ± 0.89	6	193/199
Subirana[b]	NWMF	Agromillora, Spain	4.03 ± 0.35	3	193/199
Fidelia Ruth	NWMR	IRTA, Spain	4.04 ± 0.35	5	193/199
ASF 06-71[b]	NWMF	Maillard, France	4.05 ± 0.84	2	193/199
Mesembrine	NYMF	INRA-Bordeaux, France	4.23 ± 1.12	6	193/199
ASF 01-05	NYMR	Maillard, France	4.28 ± 1.49	5	193/199
ASF 04-10[b]	NYMR	Maillard, France	4.36 ± 0.07	2	193/199
Feraude	PYNR	INRA, France	4.40 ± 0.35	5	197/203
Gartairo[b]	NYMR	PSB, Spain	4.43 ± 0.86	5	193/201
Big Top[a,b]	NYMR	Zaiger's Genetics, USA	4.44 ± 0.74	12	193/199
ASF 05-25	NWMR	Maillard, France	4.61 ± 0.49	3	193/199
Extreme Red	NYMR	Cabal, Spain	4.71 ± 1.36	4	193/199
ASF 02-22	NWMR	Maillard, France	4.73 ± 0.42	4	193/199
ASF 06-07[b]	NYMR	Maillard, France	4.80 ± 0.49	3	193/199
ASF 04-26	NWMR	Maillard, France	4.82 ± 0.48	5	193/199
EP 97.48	NYMR	Maillard, France	4.93 ± 0.51	6	193/199
ASF 02-23	NWMR	Maillard, France	4.95 ± 0.42	6	193/199
Garcica[b]	NWMR	PSB, Spain	4.95 ± 0.62	5	193/197
ASF 05-15	NYMR	Maillard, France	4.96 ± 2.57	4	193/199
Jesca	PYNR	TC, Spain	5.02 ± 0.16	3	193/199
ASF 05-08[b]	NYMR	Maillard, France	5.03 ± 0.20	2	193/199
ASF 05-19	NYMR	Maillard, France	5.06 ± 0.51	3	193/199
Nectarreve[b]	NWMR	Maillard, France	5.08 ± 0.30	3	193/199
ASF 04-23[b]	NWMR	Maillard, France	5.17 ± 0.15	3	193/199
ASF 04-27[b]	NWMR	Maillard, France	5.19 ± 0.35	3	193/199
ASF 05-01	NYMR	Maillard, France	5.32 ± 0.49	3	193/199
Honey Fire[b]	NYMR	Zaiger's Genetics, USA	5.48 ± 0.13	3	193/199
Magique[b]	NWMR	Maillard, France	5.48 ± 0.85	7	193/199
Babygold7[a]	PYNR	RU-NJ - USA	5.60 ± 0.00	1	195/201
Niagara[a]	NYMR	USA	5.62 ± 1.04	5	199/201
Calabacero	PYNR	TC, Spain	5.70 ± 1.60	6	201/201
MB-3	PWMR	IRTA, Spain	5.72 ± 0.90	5	201/201
Calante[b]	PYNR	Local variety, Spain	5.75 ± 1.29	3	201/201
Agabés	PYNR	Local variety, Spain	5.76 ± 0.00	1	201/201
Tirrenia	PYNR	ISF-Roma, Italy	5.78 ± 0.57	6	199/201
Ferlot	PYNR	INRA, France	5.78 ± 0.80	5	199/203
Canongí	PYMR	TC, Spain	5.91 ± 0.14	2	201/201
Evaisa[b]	PYNR	TC, Spain	5.91 ± 0.20	2	199/201
Maria Delizia	PWMR	DOFI, Italy	5.94 ± 1.71	6	199/201

Cultivar([1])	Fruit type	Origen	Mean Acidity ± SD (g/l)	n([2])	CPPCT040 genotype
Tardibelle[b]	PYMR	Maillard, France	5.96 ± 0.67	5	199/199
Fercluse	PYNR	INRA, France	5.98 ± 0.86	5	197/203
IFF0962	PYMR	CRA-Forli, Italy	6.04 ± 1.02	4	199/199
Christalrose	NWMR	Escande - France	6.18 ± 0.94	6	199/199
Voluptia	PWMR	ISF-Roma, Italy	6.18 ± 1.81	6	197/199
Romea[b]	PYNR	ISF-Roma, Italy	6.30 ± 1.42	12	197/199
ASF 02-55[b]	PYMR	Maillard, France	6.39 ± 1.21	4	199/199
Red Coast	PYMR	C.L.C. Ferrara, Italy	6.41 ± 0.89	6	199/199
Maria Emilia	NYMR	DOFI, Italy	6.52 ± 0.68	6	199/199
Catherina[b]	PYNR	RU-NJ - USA	6.58 ± 1.21	6	197/199
Zee Lady[a]	PYMR	Zaiger's Genetics, USA	6.62 ± 1.45	6	199/199
O'henry[a,b]	PYMR	G. Merril, USA	6.64 ± 1.29	11	199/199
Summer Lady	PYMR	Visalia (California), USA	6.67 ± 1.12	5	199/199
Surprise[b]	PWMR	INRA, France	6.69 ± 0.89	6	199/199
Lucie	PYMR	Bradford, USA	6.70 ± 0.84	5	199/199
ASF 02-65	PWMR	Maillard, France	6.71 ± 0.44	3	199/199
IFF0813	NYMR	CRA-Forli, Italy	6.75 ± 0.71	4	201/201
Tendresse	PWMR	Maillard, France	6.76 ± 1.28	5	199/199
ASF 02-52	PYMR	Maillard, France	6.86 ± 1.22	5	199/199
Dolores[a,b]	PWMR	Zaiger's Genetics, USA	6.90 ± 0.00	1	203/203
Fire Red	PYMR	UCD, USA	6.95 ± 0.87	5	199/199
Alexandra	PYMR	Zaiger's Genetics, USA	7.05 ± 0.98	6	199/201
Maycrest	PYMR	Minami, Reedley, California, USA	7.08 ± 0.47	4	199/199
Glenna[a]	PWMR	Zaiger's Genetics, USA	7.10 ± 0.00	1	199/201
ASF 04-42[b]	PYMR	Maillard, France	7.16 ± 0.67	3	199/199
Queen Crest	PYMR	Balakian Reedley (California), USA	7.17 ± 1.24	4	199/199
Rome Star	PYMR	ISF-Roma, Italy	7.18 ± 0.78	6	199/199
Sweetprim[b]	PWMR	Maillard, France	7.19 ± 1.21	3	199/201
Summersun[b]	PYNR	Visalia (California), USA	7.26 ± 0.86	4	201/201
Isabella d'Este	PWMR	Lodi, Ferrara, Italy	7.36 ± 0.98	6	199/199
ASF 02-46	PYMR	Maillard, France	7.39 ± 1.03	4	199/199
EP 94.20	PYMR	Maillard, France	7.40 ± 1.02	5	199/199
Bolero	PYMR	Bologna (ICA-CMVF), Italy	7.44 ± 0.60	4	199/199
Symphonie	PYMR	Maillard, France	7.52 ± 1.67	6	199/199
ASF 03-62	PWMR	Maillard, France	7.57 ± 0.97	5	199/199
Weinberger	NYMR	ISF-Roma, Italy	7.57 ± 1.02	3	199/199
Latefair[b]	NYMR	Zaiger's Genetics, USA	7.59 ± 1.03	6	201/201
Red Valley	PYMR	C.I.V. Ferrara. Italy	7.61 ± 0.75	6	199/199
Top Lady	PYMR	Merril, USA	7.63 ± 2.26	6	199/199
Sensation	PYMR	Maillard, France	7.68 ± 0.79	6	199/199
Etoile	PYMR	Maillard, France	7.75 ± 1.27	5	199/199
June Crest	PYMR	Zaiger's Genetics, USA	7.76 ± 1.22	6	199/199
ASF 04-09[b]	NYMR	Maillard, France	7.81 ± 3.72	2	199/199
Fantasie	PYMR	Fresno, USA	7.84 ± 0.98	6	199/199
ASF 02-48	PYMR	Maillard, France	7.86 ± 0.79	3	199/199
ASF 04-53[b]	PWMR	Maillard, France	7.92 ± 0.23	3	199/199
Silver Rome	NWMR	FaViFrut, Italy	7.97 ± 1.54	6	199/199
Crimson Lady[a]	PYMR	Bradford, USA	7.97 ± 1.03	7	199/199
EP 94.28	PYMR	Maillard, France	7.99 ± 1.06	6	199/199
John Henry	PYMR	California, USA	8.00 ± 1.20	5	199/199
Villa Giulia[a,b]	PYNR	ISF-Roma, Italy	8.00 ± 0.00	1	197/199

Cultivar([1])	Fruit type	Origen	Mean Acidity ± SD (g/l)	n([2])	CPPCT040 genotype
Early Maycrest	PYMR	Toeus, Ridley, California, USA	8.10 ± 1.10	6	199/199
Summer Rich	PYMR	Zaiger's Genetics, USA	8.10 ± 0.82	7	199/199
Big Sun[a]	PYMR	Maillard, France	8.11 ± 1.01	5	199/199
Elegant Lady	PYMR	G. Merril, USA	8.12 ± 0.82	6	199/199
August Queen	NWMR	IPSA, Italy	8.19 ± 1.36	5	199/199
Early Rich[b]	PYMR	Zaiger's Genetics, USA	8.24 ± 0.72	6	199/199
Merril June Lady[a]	PYMR	Red Bluff (California), USA	8.30 ± 0.00	1	199/199
IFF1230	PWMR	CRA-Forli, Italy	8.33 ± 0.57	4	199/199
Spring Bright	NYMR	Bradford, USA	8.35 ± 2.12	6	199/201
NG187[b]	NYMR	A. Minguzzi, Italy	8.35 ± 2.60	4	199/199
Maria Bianca	PWMR	DOFI, Italy	8.41 ± 1.38	6	199/199
Rich Lady	PYMR	Zaiger's Genetics, USA	8.42 ± 1.18	12	199/199
Corine	PYMR	Escande - France	8.45 ± 0.56	6	197/199
Sweet Red	NYMR	Convi, Italy	8.49 ± 1.12	6	199/199
Morsiani 51	NYMR	P.L. Morsiani i Sciutti, Italy	8.52 ± 1.08	6	199/199
Duchessa d'Este	PWMR	Scanavini, Italy	8.52 ± 1.27	5	199/201
PI2/84[a]	PYNR	A. Minguzzi, Italy	8.53 ± 1.28	4	199/199
Silver Late	NWMR	Zaiger's Genetics, USA	8.54 ± 0.35	4	199/201
Armking	NYMR	Armstrong Nurseries, USA	8.60 ± 0.91	6	199/199
ASF 03-63[b]	PWMR	Maillard, France	8.61 ± 1.01	5	199/199
Vista Rich	PYMR	Zaiger's Genetics, USA	8.62 ± 1.06	7	199/199
Red Fair	NYMR	Zaiger's Genetics, USA	8.65 ± 0.89	4	199/199
ASF 05-26	NWMR	Maillard, France	8.66 ± 0.61	4	199/199
Azurite	PYMR	Monteux-Caillet, France	8.66 ± 0.54	7	199/199
Weinberger 5199	PYMR	Italy	8.80 ± 0.60	3	199/199
Red Moon	PYMR	C.I.V. Ferrara. Italy	8.84 ± 0.64	6	199/201
Queen Ruby	NWMR	Zaiger's Genetics, USA	8.90 ± 1.38	6	199/199
Seduction	PYMR	Maillard, France	8.90 ± 0.86	6	199/199
Spring Lady	PYMR	Merril, USA	8.91 ± 0.69	6	199/199
Festina	NWMR	Escande, France	8.96 ± 0.83	5	199/199
Rich May	PYMR	Zaiger's Genetics, USA	8.99 ± 1.14	5	199/199
Snow Queen	NWMR	Armstrong Nurseries, USA	9.00 ± 1.18	7	199/199
Flavour Queen	NWMR	Zaiger's Genetics, USA	9.05 ± 1.17	5	199/199
Super Queen	NWMR	IPSA, Italy	9.08 ± 0.88	6	199/199
Spring Red	NYMR	Bradford, USA	9.11 ± 2.56	6	199/199
Diamond Brigth[a]	NYMR	Bradford, USA	9.19 ± 0.78	6	199/201
Royal Gem[a]	PYMR	Zaiger's Genetics, USA	9.23 ± 1.00	4	199/199
ASF 02-83[b]	NWMF	Maillard, France	9.31 ± 2.37	4	199/199
IFF1190	PYMR	CRA-Forli, Italy	9.37 ± 1.53	4	199/199
Amiga[b]	NYMR	A. Minguzzi, Italy	9.44 ± 1.43	6	199/199
IFF0800	NYMR	CRA-Forli, Italy	9.46 ± 0.76	4	199/199
Perfect Delight[a]	NYMR	Zanzi, Ferrara, Italy	9.46 ± 1.21	5	199/199
Ruby Rich	PYMR	Zaiger's Genetics, USA	9.47 ± 0.79	6	199/199
Sweet Lady	NYMR	Convi, Italy	9.49 ± 1.13	5	199/199
ASF 03-02[b]	NYMR	Maillard, France	9.53 ± 1.41	3	199/199
Diamond Ray[b]	NYMR	Plantas Sevilla S.L., Spain	9.62 ± 1.28	6	199/199
Red Silver	NWMR	Zaiger's Genetics, USA	9.65 ± 2.03	6	197/199
IFF1182	NWMR	CRA-Forli, Italy	9.73 ± 1.75	4	199/199
Dellys	NWMR	Escande - France	9.80 ± 1.33	6	199/199
Early Top[b]	NYMR	Zaiger's Genetics, USA	9.83 ± 1.44	6	199/199
Fire Top[b]	NYMR	Zaiger's Genetics, USA	9.84 ± 1.27	5	197/203

Cultivar([1])	Fruit type	Origen	Mean Acidity ± SD (g/l)	n([2])	CPPCT040 genotype
Venus	NYMR	ISF-Roma, Italy	9.86 ± 1.79	5	199/199
Fairlane	NYMR	U.S.D.A. Fresno, California, USA	9.87 ± 0.93	10	199/199
Onyx[b]	PWMR	Monteux-Caillet, France	9.96 ± 1.57	7	199/199
Snow Red	NWMR	Escande, France	10.02 ± 1.39	5	199/199
Silver King	NWMR	Prim, France	10.05 ± 0.50	6	199/199
Alice[a]	NYMR	Vivai Giuseppe Battistini, Italy	10.10 ± 1.03	6	199/199
Royal Giant	NYMR	Zaiger's Genetics, USA	10.12 ± 1.99	6	199/199
ASF 01-29[b]	NWMR	Maillard, France	10.14 ± 1.43	5	199/199
Queen Giant	NWMR	Zaiger's Genetics, USA	10.22 ± 1.82	6	199/199
Big bel	NWMR	Maillard, France	10.27 ± 1.31	3	199/199
Early Sun Grand	NYMR	Bradford, USA	10.29 ± 0.29	4	199/199
Ruby Gem	NWMR	Zaiger's Genetics, USA	10.33 ± 0.66	6	199/199
August Red[a]	NYMR	Bradford, USA	10.41 ± 2.14	6	199/199
Superstar[a]	NYMR	Sun World International - USA	10.60 ± 0.00	1	199/199
Maria Aurelia	NYMR	DOFI, Italy	10.62 ± 0.57	6	199/199
September Queen	NWMR	IPSA, Italy	10.67 ± 1.62	5	199/199
Garaco[b]	NWMR	PSB, Spain	10.71 ± 0.53	5	199/199
Autumn Free	NYMR	Bradford, USA	10.79 ± 1.03	4	199/199
Delice	PYMR	Maillard, France	10.81 ± 1.06	6	199/201
Silver Belle	NWMR	Zaiger's Genetics, USA	10.90 ± 1.73	6	199/199
Silver Ray	NWMR	FaViFrut, Italy	10.91 ± 1.72	6	199/199
Royal Moon	PYMR	Zaiger's Genetics, USA	10.96 ± 1.58	5	199/199
Flavor Gold[a]	NYMR	Zaiger's Genetics, USA	11.00 ± 0.00	1	197/199
Maria Laura	NYMR	DOFI, Italy	11.03 ± 1.33	5	199/199
Red Diamond	NYMR	Bradford, USA	11.18 ± 0.74	4	199/199
Silver Star	NWMR	FaViFrut, Italy	11.34 ± 2.48	6	199/199
Carolina	NYMR	University of Florida, USA	11.71 ± 1.06	5	199/199

[a] Varieties used in sequence analysis

[b] Varieties genotyped with the SNP DS875 using HRM

[c] First letter:P peach, N nectarine; second letter:W white,Y yellow; third letter:N non-melting flesh, M melting flesh; fourth letter F flat or R round

[d] Number of years with phenotypic data

1 Name cultivar

2 number of years with data available

The relationship between CPPCT040 alleles and TA values was analyzed through a logistic regression test. Two of the alleles, CPPCT040193 and CPPCT040^{199}, were associated with TA values ($p=3.1\times10^{-4}$ and $p=5.4\times10^{-3}$, respectively). According to the logistic regression model for CPPCT040193 ($\beta_0=30.46$ and $\beta_1=5.61$, $p\leq0.001$), the probability of finding the allele CPPCT040193 in varieties with TA≤5 g/l was high (more than 91%) while this probability decreased rapidly with increasing TA (**Appendix CI.4**). The probability for the CPPCT040199 allele was higher than 60% at all TA levels (β0=0.66 and β1=0.24, p≤0.5) and increased with TA values, but this was due to its high prevalence in the population: 92% of cultivars with TA>5 g/l carried this allele, and 72% of them in homozygosis.

As shown in **Fig. CI.1** and **Table CI.1**, all but one cultivar with TA below 5.5 g/l ('Feraude') had the CPPCT040193 allele, seven of them in homozygosis, while this allele was not observed at higher TA values.

Based on these results, cultivars were classified into subacid (TA<5.5 g/l) and acid (TA≥5.5 g/l) to conduct a χ^2 test, which confirmed the association of the marker ($p=2.56\times 10^{-35}$). The allele CPPCT040^{193} was found to be the only one contributing to the subacid phenotype ($p=2.60\times10^{-25}$). No association of CPPCT040^{199} or the other alleles with TA was observed at p≤0.01.

Although a wide distribution of TA levels was observed in the sample, no additional associations could be established between CPPCT040 alleles and acidity. Moreover, no additive effect of the allele CPPCT040^{193} was observed, i.e., two copies of this allele did not produce lower TA.

The marker CPPCT040 was also tested in 542 peach offspring (**Appendix CI.3**). Linkage analysis of marker and trait in the populations placed D at 0.048 cM from CPPCT040.

The CPPCT040^{193} allele was diagnostic for the subacid versus acid trait in all but 21 cases (4 %): 13 with this allele were classified as acid and 8 without were classified as subacid.

CI.5.2 SNP detection in the *D* region

To identify SNPs located in the region around the CPPCT040 SSR marker, and to verify the association between these SNPs and TA levels, 13 fragments in a region of 96.3 kbp flanking CPPCT040 were sequenced in six varieties, three described as subacid ('Paraguayo Delfin', 'Douceur', and 'Gratia') and three as acid ('Dolores', 'Glenna', and 'August Red'). The 13 primer pairs yielded 5,897 bp of good quality (**Table CI.2**), with 60.3%corresponding to coding DNA according with the peach genome sequence (http://www.rosaceae.org/species/prunus_persica/genome_v1.0).

Table CI.2 PCR primers of 13 amplicons positioned in peach Scaffold5: 943323..1039611 flanking CPPCT040 microsatellite marker (Scaffold5:993617..994035)

Fragment	Forward primer sequence	Reverse primer sequence	Expected amplicon size	Amplicon physical position		SNPs	INDELs
				Start	End		
HL-38*	TGACCATAAAAGTTAGGTGACTGG	TCGAGGGAGGATGAACTGTC	645	943323	943968	3	0
DF0875*	GCAGATGAACGTTATCAGACAGC	TAGGGCAAGACAAAGTCAGAGG	424	989492	989915	3	2
DF1652*	TGGCTTCAAGTCTGAGTGTGC	TTTTCACACCAAGCCCAAGG	427	990269	990695	5	3
DF2044	ATGTAGATGCTCATGCCTTGG	ATGTCCGCAATGGAAAAAGC	364	990661	991025	0	0
DF4607*	CAGGAAACGGTGATTCTTGC	GAATTCGGTGAAGCATATGACG	400	993224	993624	2	0
DF6331*	TTAGGAAGCTGCTCTTTCTTCC	GCTTATAGGGCTAGGTCAAGTCG	399	994948	995347	1	0
DF7617*	GAGTAATCTCTTGCCACAAAGG	GGTGTCCTTCAGTAAATTGTGG	394	996234	996628	1	0
DF7589*	CTAAACAGACCCCAGATTTTCC	CGCGATGTTATAATGACCAACC	504	1001206	1001709	1	0
DF9128	AAATAGGGCAGCGTAAATTCTG	TTCGTTTTGCGTATCTTCTCTG	650	1002745	1003394	0	0
DF11052*	ATGGAGGTTTTGGTTGATCG	TTGAAGACTTTCTGGGAACTGG	707	1004669	1005375	2	0
DF19433*	ACGGGAATAGTCTCAGAACTGG	TTCACGCTAAACAGGTACATCC	637	1013050	1013686	1	0
DF35167	CAAGACGCAGAGACAACTTCAG	CGGGAAATATTGAGGAAATCAG	665	1028784	1029448	0	0
DF45552	AAACTACAACAGGTTCGGTTCG	CAATGTTGTCACACGGTTGC	443	1039169	1039611	0	0

Polymorphisms (18 SNPs and five INDELs) were observed in nine of the fragments (69. 2 %), which were subsequently sequenced in 32 additional peach varieties. In total, 38 varieties were analyzed, 19 of them acid and 19 subacid, covering a broad range of TA values. In these additional 32 varieties, just one new polymorphism was observed, in 'Flavor Gold' and 'Villa Giulia' (in fragment DF4607).

In total, we observed 19 SNPs (10 transitions and 9 transversions) representing 1 SNP every 310 bp, with the number of SNPs per fragment ranging from 1 to 5, with an average of 2 SNPs per fragment. When accounting for coding and non-coding DNA, SNP frequencies varied between one SNP every 356 bp and 260 bp for coding and non-coding DNA, respectively. Additionally, five INDELs of sizes ranging from 1 to 25 bp were observed in two of the fragments (DF0875 with two INDELs of 2 and 25 bp, respectively, and DF1062 with three INDELs of 1, 2, and 3 bp).

The nine polymorphic fragments spanned 70.4 kb (**Appendix CI.5)**. The alignment of the sequences of fragments DF0875 and DF1062 was not legible after the INDELs when they were in heterozygosis. In the initial set of six varieties, 'Gratia', 'Dolores', 'Glenna', and 'August Red' were homozygous for the INDELs in both fragments. In them, most of the polymorphisms detected revealed two long haplotypes, each at least 24,194 bp long (**Fig. CI.2**). 'Paraguayo Delfin' and 'Douceur' were heterozygous for the INDELs and, consequently, we could not read the haplotype; however, we observed that SNPs flanking the fragments with INDELs (i.e., 24 kb apart) were still linked. One of the haplotypes (A) was only observed in the subacid varieties in both homozygosis and heterozygosis, while the other haplotype (B) was observed in heterozygosis in the subacid and in homozygosis in the acid varieties. Haplotype A was linked to CPPCT040^{193} while B was indistinctly observed with the other CPPCT040 alleles amplified (199, 201, and 203). When looking at the additional 32 varieties, 'Flavor Gold' and 'Villa Giulia' contained an exclusive SNP; these were the only two cultivars carrying CPPCT40^{197}. In these two varieties, the B haplotype was broken at some point in a region 1.3 kb downstream of CPPCT040.

To obtain the aligned sequence of the two chains in the fragments with INDELs, additional primers were designed to sequence the regions. Due to the distribution of the INDELs in the fragments, this strategy was only possible for DF0875. The maintenance of the two haplotypes observed in homozygous varieties was confirmed.

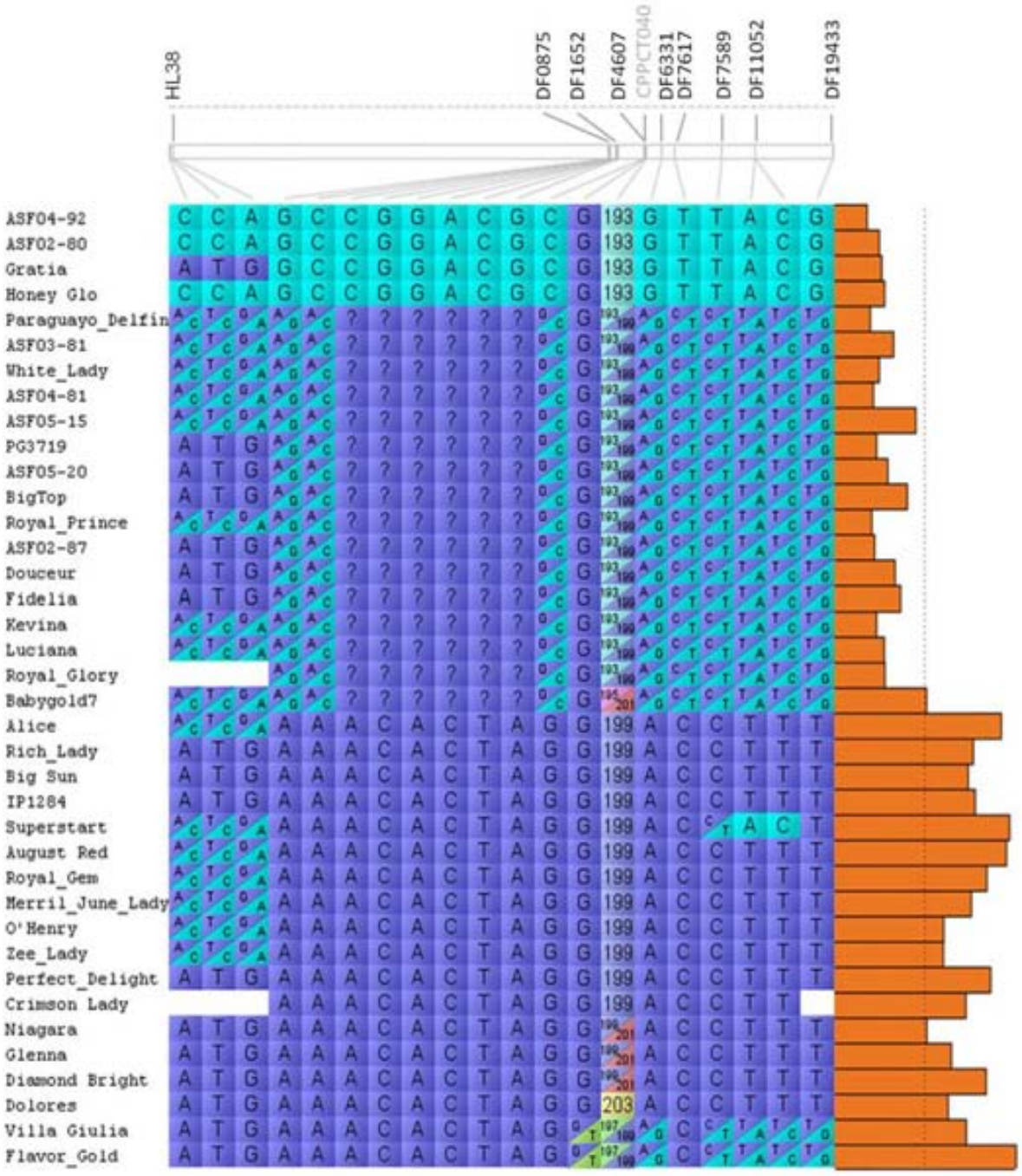


Figure CI.2 Graphical visualization of the polymorphisms obtained in 38 cultivars sequenced with nine fragments flanking CPPCT040 and spanning 70.4Kb. *Orange bars* represent titratable acidity values. The *dotted line* represents the TA level criteria to group the accessions as acid (≥5.5 g/l) or subacid (≤5.5 g/l).

Thus, after sequencing all nine fragments in 38 varieties and assuming that the haplotypes in DF1062 in the heterozygous varieties remained as in the homozygous ones, we consider that the subacid haplotype (A) including the CPPCT040^{193} allele is unique and ≥24 kbp long in all subacid varieties, while haplotype B broke through recombination downstream CPPCT040 in some acid varieties. Both A and B haplotypes recombined at some point between 4.1 and 49.6 kb upstream of CPPCT040, meaning that SNPs upstream of this point broke their linkage disequilibrium (LD) with CPPCT040 alleles.

As observed in the initial sample, haplotype A was observed in all varieties with the allele CPPCT040^{193} and vice-versa, and those heterozygous for A were also heterozygous for CPPCT040^{193}. This haplotype was also observed in 'Babygold-7' (TA=5.6 g/l), the only variety with CPPCT040^{195}.

By analyzing for structure with 17 SSRs, the sample of 38 varieties was subdivided into four subpopulations (**Fig. CI.3**). All populations but one (the red population in **Fig. CI.3** with all four subacid varieties developed in the same breeding program) included both acid and subacid varieties, excluding spurious association due to population structure.

With the aim of applying these results in breeding programs through MAS, one of the SNPs was converted to an HRM marker (DS875), based on the sequence of DF0875. The polymorphism of the amplified region of 106 bp was a C/A substitution, allele C being the one linked to the low-acidity trait (in both homozygosis and heterozygosis). The efficiency of the marker was tested in 64 varieties: In all cases, the SNP genotype corresponded to the expected result according to the TA and CPPCT040 genotype.

DS875 was also tested in the 21 seedlings of the crosses where CPPCT040 alleles were not predictive of the field phenotype. In all cases, the allele A of DS875 was in coupling with CPPCT040^{193}.

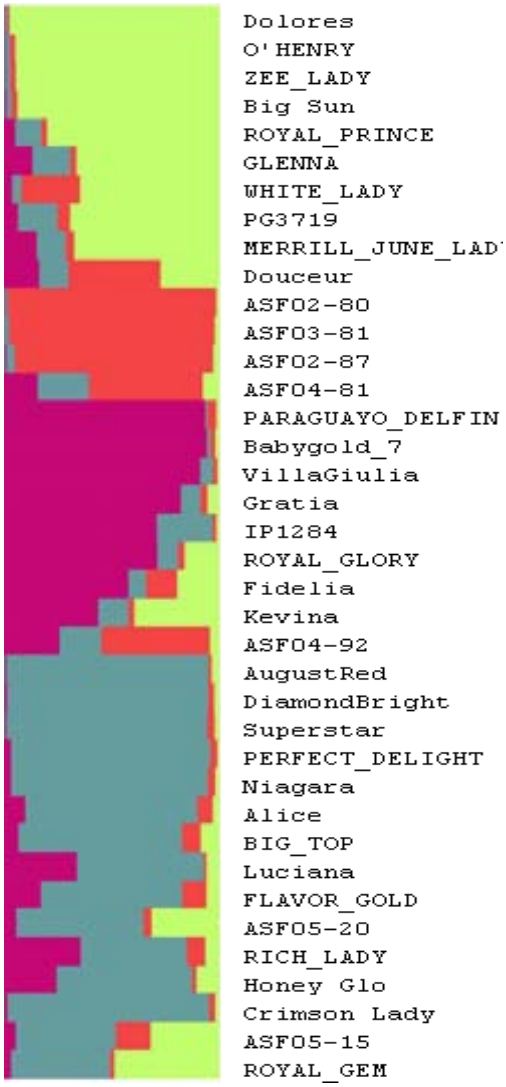


Figure CI.3 Pupulation structure of the 38 cultivars sequenced calculated with 17 SSRs unlinked and genome-wide distributed. Each cultivar is represented by a *colored horizontal bar*, each color being the percentage of membership to each of the four populations detected.

CI.6 DISCUSSION

CI.6.1 Association of the molecular markers and TA levels

In this study we evaluated the use of the SSR marker CPPCT040 in marker-assisted selection for the subacid trait in peach. In a sample of 231 varieties, this marker amplified six alleles, and the presence of one of them, the allele CPPCT040^{193} 18 either in homozygosis or heterozygosis, resulted in a TA lower than 5.5 g/l, which is consistent with the dominant nature of the subacid trait (Monet 1979). 'Feraude' (TA=4.4 g/l) was the only variety without CPPCT040^{193} 21 with a TA under 5.5 g/l. Despite our TA data 'Feraude' is often classified as acid in peach catalogs, as are its siblings 'Fercluse' and 'Fergold'. These varieties have non-melting fruits and their hard consistency allows them to be

left longer on the tree. In some acid varieties, acidity decreases after harvest maturity (Cascales *et al.,* 2005), so delaying the harvest time may produce fruits with low TA and acceptable consistency and postharvest life. This may have prevented a correct classification of the acidity trait here.

The regression analysis also shows a slight association of the allele CPPCT040^{199} not confirmed by the Pearson's χ_2-test. This spurious association is probably caused by the low frequency of the other alleles in the collection.

This is the first time that the subacid trait has been associated to a threshold of TA levels based on a wide collection of cultivars. Previously, Boudehri *et al.,* (2009) fixed the threshold between acid and subacid peaches at TA equal to 4.02 g/l and pH equal to 4.0, based on the distribution of TA and pH values and genotypes in 1,718 seedlings from 7 populations derived from three parentals. Here we conclude that subacid varieties carry an exclusive allele, CPPCT040^{193}, always linked to TA values lower than 5.5 g/l.

The classification of varieties into acid and subacid classes is clear, but the variability of TA also results in a wide range of levels of acidity. We have not found association between CPPCT040 and TA other than that leading to this classification. Part of the observed variability may be due to environmental factors (Etienne *et al.,* 2013), to other QTLs controlling a minor part of the trait (Etienne *et al.,* 2002; and Quilot *et al.,* 2004) or to the interactions between different alleles at this locus, including the subacid allele (Boudheri *et al.,* 2009).

A detailed analysis of the genomic region flanking the SSR marker linked to the trait provides valuable knowledge of the genetic structure of the region as well as SNPs useful in MAS. Sequences in a region of 70.4 Kb were more variable than the estimations of the average variability genome wide. Here we observed one SNP every 310 pb and one INDEL every 1.2 Kpb. This contrasts with data reported by Aranzana *et al.,* (2012), where one SNP was observed every 598 bp and one INDEL every 4 Kbp, and with the average nucleotide diversity of one SNP every 900 Kbp found by Verde *et al.,* (2013) after comparing European and North American peach varieties.

The number of SNP-haplotypes was lower than the number of CPPCT040 alleles, which can be explained by the high mutation rate of SSRs compared to SNPs.

Haplotype A (24 Kb long) was linked to the CPPCT040^{193} 18 allele. 'Babygold-7' was the only variety with haplotype A lacking CPPC040^{193} but carrying, instead, CPPCT040^{195}.In a wide analysis of peach cultivars with SSRs we observed this allele only in the 'Babygold' series and descendants (data not shown). In a study of 434 *Prunus* accessions, most of them peaches from China, allele CPPCT040^{195} was also rare (frequency 0.8%) in both homozygosis and heterozygosis. All four

accessions carrying the CPPCT040^{195} 24 allele were subacid landraces; three of them clustered together in a UPGMA dendrogram while the fourth was close (Li *et al.,* 2013). In our study we classified 'Babygold-7' as acid from one year of TA data (TA = 5.6 g/l). However, the TA evaluated in 'Babygold-7' over a period of 11 years at another IRTA research station was equal to 5.3 ± 1.1 (J. Carbó, personal communication). 'Babygold-7' is a non-melting variety and the wide standard deviation observed could be due to differences in maturity at harvest time in different years. A hypothesis compatible with these results is that 'Babygold-7' is a subacid variety, in which case the CPPCT040^{195} allele could be a recent mutation of CPPCT040^{193} that would also be associated to the subacid trait.

We have found that the subacid haplotype is longer and clearly different than the acid one. This, together with the low SSR variability observed in the subacid varieties could indicate that this trait was introduced recently in our collections from a unique germplasm source. The most likely hypothesis is that it comes from China, where peach originated and spread to the rest of the world and where the subacid trait is largely preferred by consumers. In our sample we have analyzed the subacid varieties 'White Lady' and 'Fidelia' both obtained by Zaiger Genetics. These two cultivars have in their pedigree the subacid variety 'Sam Houston', probably the donor of the allele. 'Sam Houston' was created in Texas A&M University College Station were Honey peaches, a group of white-fleshed fruit and honey-sweet flavor varieties coming from southern China, were intensively used in breeding programs (Cullinan 1937). The subacid allele is also present in the original flat peaches from Chinese origin used by US breeders and additionally characterized by their white skin and very sweet flesh (Cullinan 1937). The same haplotype around CPPCT040 was observed in 'Paraguayo Delfin', a Spanish local variety with a subacid taste and flat shape. This variety usually clusters with 'Chinese Cling' and is genetically similar to the Chinese flat peach landrace Yu Lu Pantao (Aranzana *et al.,* 2010; Li *et al.,* 2013). Although peach genetic variability is large in Chinese germplasm, Chinese breeding efforts have reduced diversity in the same way than the Occidental ones did (Xie *et al.,* 2010). It is likely that the Chinese materials used in the early US breeding programs carried a single allele that was later transferred to the modern varieties of US and Europe.

CI.6.2 Implications for MAS

Here we provide a tool to identify the subacid trait independently of environmental conditions and stage of maturity, representing a useful tool for breeders.

We show that, using the CPPCT040 SSR marker we can predict the subacid trait with high probability. This high association was proved first in a broad set of cultivars and then in the

descendants of a breeding program. In the latter case, using allele CPPCT040^{193} as diagnostic of the subacid trait, we would have chosen 97.5% of the seedlings that the breeder would have selected as subacid, and others (13; 2.4% of the total) that would be classified as acid by taste. This means that the use of the marker in the early selection of the seedlings would have resulted in selecting 2.4% false positives and 2.5% false negatives according to breeder field decisions. The low false positive and negative rate could be due to recombinations between the marker and the trait locus, but also to phenotypic, sampling or genotyping errors.

Independently of the reason, this level of accuracy is highly acceptable in breeding programs, mainly in those of fruit trees species with a juvenile phase and which require large surface areas and resources to maintain the seedlings until they fruit.

Boudehri *et al.,* (2009) developed SCAR and CAP markers in a region of 4.8 cM flanking locus D. The closest SCAR marker was at 0.4 cM from CPPCT040 and was based on a SSR. The two CAP markers detected SNPs at 195 Kb and 317 Kb, respectively, upstream of CPPCT040, and consequently they may not 1 be in LD with CPPCT040 in peach germplasm.

In *Prunus* several candidate genes have been mapped (Horn *et al.,* 2005; Ogundiwin *et al.,* 2009; Illa *et al.,* 2011) in the D locus region and since 2010 the *Prunus* genome annotation is available. The haplotype 24 kb long conserved in all subacid varieties contains 3 annotated genes, a homeodomain-like (ppa011225m), a homoserine dehydrogenase (ppa013023m) and a NAD(P)-binding domain (ppa012176m). To our knowledge none of these genes have been reported to have a role in fruit acidity.

Recently, several projects have focused on obtaining SNPs associated to interesting agronomic traits for use in early character diagnosis of parents or progenies. The main advantages of SNPs are their stability and their inclusion in multiplexing platforms for high-throughput genotyping. Here we present eight linked SNPs useful for this purpose and developed primers to genotype one of them using HRM, a cheaper option than standard SSR genotyping methods, providing a robust tool for MAS.

CI.7 CONCLUSIONS

The subacid flesh taste is one of the main traits under selection in commercial breeding programs. Usually breeders select for this trait by tasting the fruits at maturity or by measuring the titratable acidity (TA). Several minor QTLs have been previously identified for variability in TA levels;

however the subacid trait is controlled by a major gene. Although epistatic interaction between QTLs may occur here our data confirmed the high association between the subacid trait, measured as TA and field taste, one of the alleles of the CPCCT040 SSR marker and a set of SNP and INDEL markers on a long haplotype of at least 24 Kb around CPPCT040. This haplotype was conserved in all subacid varieties tested and our results suggest that it could have been introgressed from a single source into the European-North American commercial germplasm. The markers provided can be used in breeding programs as diagnostic for the character in peach both for parents and derived progenies. Our results also suggest that the TA value of 5.5 g/l is the cut-off point to distinguish between varieties which do or do not carry the subacid allele.

CI.8 ACKNOLEDGEMENTS

Funding for this research was partly provided from project AGL2012-40228-C02-01 from the Spanish Ministry of Economy and Knowledge. We thank Christian Fontich for providing data and plant material from the peach progenies 1 of ASF-IRTA breeding program supported by Fruit Futur.

CI.9 DATA ARCHIVING STATEMENT

FASTA sequences of the subacid variety 'Honey Glo' and the acid variety 'Glenna' have been submitted to the NCBI GeneBank using the BankIt tool, with accession numbers KJ023869-KJ023894. Both varieties are homozygous at all loci. A table with the full list of accession numbers is presented in **Appendix CI.2**.

CI.10 REFERENCES

Abidi W, Jimenez S, Moreno MA, Gogorcena Y (2011) Evaluation of antioxidant compounds and total sugar content in a nectarine [Prunus persica (L.) Batsch]Int J Mol Sci 12:6919–6935

Aranzana MJ, Carbó J, Arús P (2003) Microsatellite variability in peach [Prunus persica (L.) Batsch]: cultivar identification, marker mutation, pedigree inferences and population structure. Theor Appl Genet 106:1341–1352

Aranzana M, Abbassi E-K, Howad W, Arus P (2010) Genetic variation, population structure and linkage disequilibrium in peach commercial varieties. BMC Genet 11:69

Aranzana M, Illa E, Howad W, Arus P (2012) A first insight into peach [Prunus persica (L.) Batsch] SNP variability. Tree Genet Genomes 8:1359–1369

Boudehri K, Bendahmane A, Cardinet G, Troadec C, Moing A, Dirlewanger E (2009) Phenotypic and fine genetic characterization of the D locus controlling fruit acidity in peach. BMC Plant Biol 9:59

Cantín CM, Gogorcena Y, Moreno MA (2009) Analysis of phenotypic variation of sugar profile in different peach and nectarine [Prunus persica (L.) Batsch] breeding progenies. J Sci Food Agric 89:1909–1917

Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J (2009) DNAPlotter: circular and linear interactive genome visualization. Bioinformatics 25:119–120

Cascales AI, Costell E, Romojaro F (2005) Effects of the degree of maturity on the chemical composition, physical characteristics and sensory attributes of peach (Prunus persica) cv. Caterin. Food Sci Technol Int 11:345–352

Crisosto CH, Crisosto GM (2005) Relationship between ripe soluble solids concentration (RSSC) and consumer acceptance of high and low acid melting flesh peach and nectarine (Prunus persica (L.) Batsch) cultivars. Postharvest Biol Tec 38:239–246

Crisosto CH, Valero D (2008) Harvesting and postharvest handling of peaches for the fresh market. In: Layne, Bassi (eds) The Peach: Botany, Production and Uses, pp 575-596

Cullinan FP (1937) Improvement of Stone Fruits (Peaches). USDA Yearbook of Agriculture. pp 665-702

Dirlewanger E, Pronier V, Parvery C, Rothan C, Guye A, Monet R (1998) Genetic linkage map of peach [Prunus persica (L.) Batsch] using morphological and molecular markers. Theor Appl Genet 97:888–895

Dirlewanger E, Cosson P, Boudehri K, Renaud C, Capdeville G, Tauzin Y, Laigret F, Moing A (2006) Development of a second-generation genetic linkage map for peach [Prunus persica (L.) Batsch] and characterization of morphological traits affecting flower and fruit. Tree Genet Genomes 3:1–13

Doyle JJ, Doyle JI (1990) Isolation of plant DNA from fresh tissue. Focus 12:13–15

Etienne C, Rothan C, Moing A, Plomion C, Bodénès C, Svanella-Dumas L, Cosson P, Pronier V, Monet R, Dirlewanger E (2002) Candidate genes and QTLs for sugar and organic acid content in peach [Prunus persica (L.) Batsch]. Theor Appl Genet 105:145–159

Etienne A, Genard M, Lobit P, Mbeguie-A-Mbeguie D, Bugaud C (2013) What controls fleshy fruit acidity? A review of malate and citrate accumulation in fruit cells. J Exp Bot 64:1451–1469

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620

Faust M, Timon B (1995) Origin and dissemination of peach. Hortic Rev 17:331–379

Horn R, Lecouls AC, Callahan A, Dandekar A, Garay L, McCord P, Howad W, Chan H, Verde I, Main D, Jung S, Georgi L, Forrest S, Mook J, Zhenbentyayeva T, Yu Y, Kim HR, Jesudurai C, Sosinski B, Arús P, Baird V, Parffit D, Reighard G, Scorza R, Tomkins J, Wing R, Abbott AG (2005) Candidate gene database and transcript map for peach, a model species for fruit trees. Theor Appl Genet 110:1419–1428

Iglesias I, 2013 Peach production in Spain: current situation and trends, from production to consumption. Proceedings of the 4th Conference, Innovations in Fruit Growing, 75-96. Ed.: D. Milatovic, Belgrad University (Belgrard, Serbia)

Iglesias I, Echeverria G (2009) Differential effect of cultivar and harvest date on nectarine colour, quality and consumer acceptance. Sci Hortic-Amst 120:41–50

Iglesias I, Carbó J, Bonany J, Casals M, Dalmau R, Montserrat R (2005) Innovación varietal en melocotonero: especial referencia a las nuevas variedades de nectarina. Frutic Profesional 152:6–36

Illa E, Eduardo I, Audergon J, Barale F, Dirlewanger E, Li X, Moing A, Lambert P, Le Dantec L, Gao Z, Poëssel J-L, Pozzi C, Rossini L, Vecchietti A, Arus P, Howad W (2011) Saturating the Prunus (stone fruits) genome with candidate genes for fruit quality. Mol Breed 28:667–682

Lambert P, Dirlewanger E, Laurens F (2009) La sélection assistée par marqueurs (SAM) chez les arbres fruitiers: une approche prometteuse au service de l'innovation variétale. Innov Agronomiques 7:139–152

Li XW, Meng XQ, Jia HJ, Yu ML, Ma RJ, Wang LR, Cao K, Shen ZJ, Niu L, Tian JB, Chen MJ, Xie M, Arús P, Gao ZS, Aranzana MJ (2013) Peach genetic resources: diversity, population structure and linkage disequilibrium. BMC Genet 14:84

Milne I, Shaw P, Stephen G, Bayer M, Cardle L, Thomas WTB, Flavell AJ, Marshall D (2010) Flapjack-graphical genotype visualization. Bioinformatics 26:3133–3134

Monet R (1979) Transmission génétique du caractère "fruit doux" chez le pêcher. Incidence sur la sélection pour la qualité. Eucarpia Fruit Section, Tree Fruit Breeding, Angers, France, INRA, pp 273-276

Ogundiwin E, Peace C, Gradziel T, Parfitt D, Bliss F, Crisosto C (2009) A fruit quality gene map of Prunus. BMC Genomics 10:587

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Quilot B, Wu BH, Kervella J, Génard M, Foulongne M, Moreau K (2004) QTL analysis of quality traits in an advanced backcross between Prunus persica cultivars and the wild relative species P. davidiana. Theor Appl Genet 109:884–897

R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/

Reig G, Iglesias I, Gatius F, Alegre S (2013) Antioxidant capacity, quality, and anthocyanin and nutrientcontents of several peach cultivars [Prunus persica (L.) Batsch] grown in Spain. J Agr Food Chem 61:6344–6357

Sansavini S, Bassi D, Gamberini A (2006) Miglioramento varietale del pesco: genetica e genomica per nuove tipologie di frutto. Tendenze in California, Francia e Italia . Rivista di Frutticoltura 7-8

Scorza R, Mehlenbacher SA, Lightner GW (1985) Inbreeding and coancestry of freestone peach cultivars of the eastern United States and implications for peach germplasm improvement. J Am Soc Hortic Sci 110:547–552

Untergrasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3—new capabilities and interfaces. Bioinformatics 23:1289–1291

Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, Zhebentyayeva T, Dettori MT, Grimwood J, Cattonaro F, Zuccolo A, Rossini L, Jenkins J, Vendramin E, Meisel LA, Decroocq V, Sosinski B, Prochnik S, Mitros T, Policriti A, Cipriani G, Dondini L, Ficklin S, Goodstein DM, Xuan P, Fabbro CD, Aramini V, Copetti D, Gonzalez S, Horner DS, Falchi R, Lucas S, Mica E, Maldonado J, Lazzari B, Bielenberg D, Pirona R, Miculan M, Barakat A, Testolin R, Stella A, Tartarini S, Tonutti P, Arus P, Orellana A, Wells C, Main D, Vizzotto G, Silva H, Salamini F, Schmutz J, Morgante M, Rokhsar DS (2013) The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet 45:487–494

Xie R, Li X, Chai M, Song L, Jia H, Wu D, Chen M, Chen K, Aranzana M, Gao Z (2010) Evaluation of the genetic diversity of Asian peach accessions using a selected set of SSR markers. Sci Hortic-Amst 125:622–629

Yoshida M (1970) Genetical studies on the fruit quality of peach varieties. 1. Acidity. Bull Fruit Trees Res Stn Ser A 9:1–15

CHAPTER II: A candidate gene for fruit shape in peach

CII.1 INTRODUCTION

Fruits are the mature ovary formed after fertilization and show a huge variability in morphology. Fruits may be simple, aggregate or composite, fleshy or dry, with one or more seeds, with many different shapes. Fruits may be flat, round, elongated, ribbed and ovate with different dimensions. Many genetic studies have aimed to unravel the genetic bases of fruit pattering, especially in the model species *Arabidopsis*.

Arabidopsis fruits are dehiscent pods (siliques) generated from two carpels fused. This fusion generates a cylinder of cells which grows apically to form the gynoecium. At the end of development the pods dry and open longitudinally. Genes from different families control their development. These families are: YABBIE genes, MADS-Box genes, and receptor like kinases (RLKs) which encode proteins containing conserved domains involved in DNA binding, protein-protein interactions or both. YABBIE genes such as: CRAB, CLAW, SPATULA gene or ETTIN factor have functions related to carpel morphogenesis (Bowman & Smyth, 1999; Sessions *et al.,* 1997), while genes belonging to MADS-box transcription factor family, such as: FRUITFULL (FUL; which belongs to APETALA 1 (AP1)/FULL clade), SHATTERPROOF 1 and 2 (SHP1 /SHP2) and *SEEDSTICK* (*STK*) in the AGAMOUS (AG) clade, are the first regulator genes in the pathway involved in the formation of the dehiscence zone (DZ), being necessary for fruit pattern organization, cell expansion and separation and lignin deposition (Gu *et al.,* 1998; Liljegren *et al.,* 2000). Genes of these two families are expressed in different tissues and stages of plant development. Receptor like kinases (RLKs) of the large leucine-rich repeat (LRR) group belong to the largest subfamily of transmembrane receptor-like kinases in plants, with over 200 members in *Arabidopsis* (Torii, 2004). LRR-RLKs have been reported to be involved in different developmental processes as well as in defense-related processes. Two of these LRR-RLKs, CLAVATA-1 (CLV1) and ERECTA (ER), show functional implications in the maintenance, size and shape of meristems (Mandel *et al.,* 2014; Torii *et al.,* 1996) and represent in conjunction with WUSCHEL (WUS) transcription factor the main regulatory network to maintain the homeostasis between the continuous development of stem cells and the cell recruitment for lateral organ formation (Uchida *et al.,* 2013).

Among cultivated species, tomato is the one where fruit shape has been more studied. Tomato fruits are berries which develop from the ovary after fertilization of the ovules. The wall of the ovary transforms into the pericarp and encloses the placenta and seeds. Four genes controlling tomato fruit shape have been cloned: SUN which encodes a protein of IQ domain family that is characterized by calmodulin-binding domain (Abel *et al.,* 2005; Xiao *et al.,* 2008), OVATE involved in fruit elongation by encoding a transcriptional repressor belonging to the Ovate Family protein

(OFP)(Hackbusch *et al.,* 2005; Liu *et al.,* 2002) and LOCULE NUMBER (LC) and FASCIATED (FAS) which determine locule number and flat shape (Rodríguez *et al.,* 2011) by encoding an orthologous of the *A. thaliana gene* WUS which regulates meristem size (Muños *et al.,* 2011) and a protein of the YABBY family which controls organ polarity (Cong *et al.,* 2008) respectively. In addition to these genes, several loci have been identified regulating fruit shape including two suppressors elements of the ovate mutation (Sov1 and Sov2) (Rodríguez *et al., 2013),* one in the mutant Self1 mapped on the long arm of chromosome 8 producing fruit elongation by increasing cell layers in the ovary (Chusreeaeom *et al.,* 2014) and the QTL fs8.1 which also controls fruit elongation (Paran & Van der Knaap, 2007) and has a size of 3Mb region containing 122 candidate genes.

SUN, OVATE, and fs8.1 act together in additive manner controlling fruit shape producing longer fruits than acting alone or in combination with other genes. Recently, Monforte *et al.*, (2014) investigated orthologous genes between tomato and melon in order to see if the molecular basis controlling morphology variation in *Solanaceae* family could explain the morphology variation in *Cucurbitaceae* family. They could physically localize on the melon pseudo-chromosomes 24 members of the SUN (CmSUN), 17 of the OFP (CmOFP), one of the YABBY (CmYABBY), nine of the CNR (CmCNR), five of the KLUH/CYP78A (CmCYP78A), and 10 of the WOX (CmWOX).

Peach fruits are drupes which develop from a single carpel. The calyx and the stamen of the flowers are fused into the hypanthium tissue forming a cuplike structure around the ovary. All peach tissues come from the ovary; the outer skin is the exocarp, the edible flesh from the mesocarp and the pit from the endocarp. Studies in peach have revealed crucial role of the PLENA-like (PpPLENA) gene during the transformation of the carpel into a ripe fleshy fruit (Causier *et al.,* 2005; Tadiello *et al.,* 2009). However differences between peach and *Arabidopsis* or between peach and tomato and melon in fruit development predicts that most of the previously mentioned genes won't be responsible for peach shape natural variation

Peach is one of the fruit species economically more important in temperate regions. Most of commercialized varieties are round and oval shaped, however commercial interest in flat shape fruits is increasing fast. Nowadays, only in Spain 3420 ha are cultivated with flat peaches, producing 51.000 tons (Iglesias, 2009).

Flat peaches originated in South China, where are known as "pentao" derived from the original Chinese "Pan Tao". In the mid-1800s several Chinese flat varieties were introduced in USA breeding programs as carriers of characters such us low chilling (Cullinan, 1937), but they were popular for a brief period of time. It is believed that the first bred flat peach was a variety called

'Saturn' by Starks Nursery in 1985. Few years later, in 1990s it began to be cultivated widespread (Bassi & Monet, 2008).

The flat shape of the peach fruit is determined by a single dominant gene *S* (for saucer-shaped) (Lesley, 1939) mapped in the distal part of chromosome 6 (Dirlewanger *et al.,* 1998). The flat allele is dominant over the non-flat one, however flat fruits with this allele in homozygosis (S/S) abort two months after anthesis (Dirlewanger *et al.,* 2006). Although the hypothesis of a single gene is the one most applauded, the abortion of young flat shaped fruits has also suggested the hypothesis of the existence of two dominant closely linked genes in repulsion. In this last case, *S*-/*Af*- would produce flat peaches, *S*-/*afaf* would determine aborting fruits while round fruit would have *ss*/*Af*- genotype (Dirlewanger *et al.,* 2006). Up to now several markers have been identified around the *S* locus, by either the analysis of mapping populations (Dirlewanger *et al.*, 2006; Picañol *et al.,* 2012) and the analysis of germplasm (Picañol *et al.,* 2012). One of the markers, the SSR UDP98-412 has been reported to be tightly linked to the *S* locus and works efficiently in MAS (Picañol *et al.,* 2012).

Although Horn *et al.,* (2005) mapped ESTs of 3,842 candidate genes for fruit quality in the *Prunus* reference map, no candidate genes for fruit shape have been identified so far close to this locus in peach. In this work, we find a LRR-kinase as the causal gene of the flat shape of peach varieties. We have validated the function of this gene in a sport mutant of a flat variety that reverts to the round shape.

CII.2 MATERIAL AND METHODS

CII.2.1 Plant material and DNA extraction

In total we studied 200 peach samples. Among them 129 corresponded to peach cultivars (67 round, 57 flat fruit cultivars and 4 with unknown fruit shape; see **Appendix C II.1**) sixty-nine were F1 seedlings from the cross between the two flat peaches'UFO-3' x 'Sweet cap' and a flat variety ('UFO4') plus its round shape sport mutant. All these samples were classified as round, flat or aborting in those cases where fruit set stopped their development few weeks after pollination.

DNA was extracted from young leaves using either the Doyle's method (Doyle & Doyle, 1987) or the Viruel's protocol (Viruel *et al.,* 1995). DNA from mutant was extracted from leaves, flesh fruit, skin fruit and stone using DNAsy Qiagen kit (*Qiagen*, Hilden, Germany).

CII.2.2 Genotyping

All samples were genotyped with the SSR marker UDP98-412 SSR using the conditions previously described in Picañol *et al.,* (2012). The primer forward was labeled with fluorochrome and products were separated by capillary electrophoresis using the ABI/Prism 3130xl (PE/Applied Biosystems) automatic sequencer (Aranzana *et al*., 2003).

Using the peach genome sequence available at the *Rosaceae* website (http://www.rosaceae.org/gb/gbrowse/prunus_persica) and the peach genome browser in the Genome Database for Rosaceae (Jung *et al.,* 2008) we designed 23 primer pairs to amplify fragments of 450-600 bp in a 388.6 Kb region (scaffold_6:24,389,857..24,778,479) including the UDP98-412 marker (**Table CII.1**). Fourteen of them were designed covering a 30 Kb region including the marker UDP98-412 (scaffold_6: 24,748,247..24,778,479) and the nine remaining in a 26,75kb region 337Kb upstream UDP98-412 (scaffold_6: 24,753,353..24,753,728). This region was the closest one to UDP98-412 with SNPs annotated (**Appendix CII.5**) in the peach genome browser. Primers were designed using Primer3 software (Rozen & Skaletsky, 1999) avoiding amplification of SSR regions.

Primers were first tested in six varieties, three of them flat ('Mesembrine', 'Paraguayo delfín' and 'Subirana') and three with round fruits ('Garcica', 'HoneyGlo' and 'Luciana'). PCR products amplifying a single band were purified with Exosap-it (GE HealthcareLife Science) in a single pipetting step and used as a template for sequencing reaction using BigDye^{TM}Terminator Cycle Sequencing Kit (Applied Biosystems, Foster City, CA, USA) and forward primers. The sequencing reaction profile included 25 cycles of 96ºC for 10s followed by 50ºC for 6s, and 60ºC for 4 min and it was carried out by ABI Prism 3130xl DNA Analyzer (Applied Biosystems, Foster City, California, CA, USA). Sequences were visualized and manually edited with Sequencher 5.0 software (Gene Codes Corporation; Ann Arbor, MI, USA). Fragment ends were trimmed to remove low-quality sequence.

The primer pairs Flatin 1F (5'-ATTATTCCCCCATGCTTGAC-3') and kinase-5R (**Table CII.2**) were used together to genotype flat and round varieties and the offsprings. The primer Flatin-1F was labelled with fluorochrome. PCR conditions, fragment separation and analysis were performed as previously described for the SSR marker.

Table CII.1 Primer pairs used to look for SNPs around UDP98-412.

Amplicon	Forward primer	Reverse primer	Length	Start	End	PCR	SNPs
Amplicon-1	cttgaatctcagtggttgttcg	ttctgaaaggtccacactgg	670	24389857	24390526	✓	–
Amplicon-2	ggttccctattgaaaactgtcc	attcaaggatgcaaggtagg	465	24391313	24391777	✓	–
Amplicon-3	tgtagattgtgtggtgacagagg	aggagacagaggaaacacaagc	611	24398063	24398673	✓	6
Amplicon-4	tatgttaaggagcgggttaagg	agtgttccaagttctggtctgg	627	24399129	24399755	✓	6
Amplicon-5	ggattactcaggcaaccatttc	tcccgcaataattgtatccag	646	24406396	24407041	✓	9
Amplicon-6	tcccctatcgattgtcaaattc	taatcccacgatggccagaa	551	24406996	24407546	✓	4
Amplicon-7	ggggataagttctctttctcagc	ggcctttaatctgattccttcc	473	24411848	24412320	✓	5
Amplicon-8	caatttggaaagacctcgaatc	gatagatcaagcacccgaagac	604	24414812	24415415	✓	–
Amplicon-9	tccctaacagaggtcaaattcc	gtaacctgggcttttgatatgc	516	24415828	24416343	✓	–
UDP-5106	ggggcatgcacaaacataatag	gcgtcatatatgctgggaagtc	356	24748247	24748603	✓	–
PY_1	gtgaataggtttggctctttcc	ccctttcatttacccttgtcc	226	24750026	24750252	✓	–
UDP-4070	atattacccctcttcgttggt	ctgggtataaaatggggcatct	446	24750498	24750944	✓	–
PY_2	acttgtaagccgaaagagatgg	agtttacttcacaggccaaagc	422	24750703	24751125	✓	–
PY_3	ttaattccactcctctctcatgc	tccctctcaacataaatgatcc	290	24751259	24751549	✓	–
PY-4	cagcaccactgactaagtgacc	cctaaccgcagctctttatacg	200	24752916	24753116	✓	–
UDP-3566	gccaactgaaaagtctctgtcc	tgccactagatgtgttctgagg	504	24756919	24757423	✓	–
UDP+6923	gagcttacatttcaggagttcg	ctgtaggacacgtttgttttgg	508	24760276	24760784	✓	–
UDP+9322	aatccaggagatgctgtaatgg	ctcttcatcttgtcagctctgg	541	24762675	24763216	✓	–
UDP+11962	aagtccaagtcaaaacgtaggc	gaatgttctccctcatggtagg	587	24765315	24765902	✓	–
UDP+15090	caagaagccaaatcacactgc	ctcatggagggtagatctgagg	677	24768443	24769120	✓	–
UDP+18630	gtcgcaagttgaccatgttacc	atcaaccacgagagtccatagg	680	24771983	24772663	✓	–
UDP+21817	atagcttcggtagggtacatgc	tagcctaccccaagaaaatacg	672	24775170	24775842	✓	–
UDP+24557	agctgctcaaggagaaagagg	ataactcgtgcgaatctcaagg	569	24777910	24778479	✓	–

CII.2.3 Cloning of PCR fragments

The PCR products were cloned into the pGEM T-easy vector (Promega) following the manufacture instructions. *Escherichia coli* DH5alpha electro competent cells (Invitrogen) were transformed with the ligated plasmid by electroporation in the Gene PulserXcel electroporation system (BIORAD) following the conditions: capacitance 25 µF; resistance 200 ohm and voltage 1,8kv. Transformed cells were shaken horizontally at 250 rpm and 37ºC for 1h and a half in 1ml liquid LB medium. Then, fifty microliters of transformed cells solution was pipetted onto 10 cm Luria-Bertani (LB) agar plates containing 50ug/ml ampicillin, 80ug/ml X-gal and 0,5mM isopropyl-β-D-1-tiogalactopiranósido (IPTG). Positive colonies were tooth picked from the LB plates for use as template DNA for colony PCR. Colonies were genotyped by PCR following the conditions described previously. Colonies carrying the desirable allele were grown in 5mL of LB liquid broth containing 50ug/ml of carbenicillin with overnight incubation at 37ºC in a shaking oven at 250rpm. Bacterial

cultures pellets were obtained by centrifugation at 3000rpm for 10 min. Plasmids were extracted from bacterial cells using a QIAprep miniprep spin-kit (Qiagen) according to the manufacturer's protocol and resuspended in 50 µl of sterile water. Then, 4ul of each extracted plasmid were sequenced with the vector specific primers, either T7 or SPS6 and following the same sequencing protocol previously described.

CII.2.4 Sequencing of ppa025511 gene

CII.2.4.1 Round allele amplification and sequencing

Using as a reference the peach genome sequence we designed 6 overlapping primers (**Table CII.2**) in ppa025511m (scaffold_6:24,405,493..24,407,745) to obtain the full sequence of the gene (**Fig. CII.1**). Primers were designed to amplify single fragments avoiding amplification of duplicated regions. Sequencing reactions and analysis were performed as described above. Primers were used in the same 6 varieties used to find polymorphism ('Mesembrine', 'Paraguayo delfín', 'Subirana', 'Garcica', 'Honeyglo', and 'Luciana') plus in 'aborting05' seedling.

Table CII.2 Overlapping primer pairs used for the whole sequencing of the candidate gene

PC*	Forward primer name	Sequence	Reverse primer name	Sequence	Start	End	Size (bp)
PC1	FullKinase6_F	ccaccacaacctttatttctc	Kinase6_827_R	gagactgcttgaatcgtcaatg	24405409	24405799	390
PC2	Kinase6_1128_F	gccttcaattttctcatgatcc	Kinase6_1128_R	atctggttttctgaaaggtcca	24405621	24406215	595
PC3	Kinase6_interno_F	tgacaacctacttgaggggagt	Kinase6_1701_R	accacctaactgatttccatcg	24406060	24406717	657
PC4	Kinase-5F	ggattactcaggcaaccatttc	Kinase-5R	tcccgcaataattgtatccag	24406396	24407041	646
PC5	Kinase6_2337_F	tccttgtgttgcggtccaaca	Kinase-6R	taatcccacgatggccagaa	24406830	24407546	716
PC6	Kinase6_2641_F	caccttgattgacttctcttgc	FullKinase6_R	taaagaaaaagatggccaggaa	24407134	24407819	685

* Primer combination

CII.2.4.2 Flat allele amplification and sequencing

We designed three primers to do a Long-Range PCR amplification of the flat and round alleles of ppa025511m. The forward primers 1F (5'-GGAGGTGTCCCTTTTTCCACT-3'), 14F (5'-TCCACCACGCCTTATCTGAC-3'), and 3F (5'-ATTTCTTGCAGGCACCGACT-3'), were designed 16,127bp, 10,072bp and 523bp upstream the gene respectively. The reverse primer 3R (5'-AGTCCATCTGTCGAGTTGGC-3'), was designed 558bp downstream the gene (**Fig. CII.4 D**).

Long-range PCR were performed with primer combinations 9F-3R using LongAmp® *Taq* Polymerase (New England BioLabs ®$_{INC}$). Each reaction contained 1x LongAmp reaction buffer, 0.3mM dNTP mix, 0.8μM each primer, 5% DMSO, 5 units of polymerase, 40ng of template DNA, and sterile Milli-Q water to a final volume of 25μl. The following PCR protocol was performed on a S-1000^{TM} Thermal Cycler (Bio-Rad Laboratories, Inc.Hercules, California, USA): 95ºC for 5 min; 35 cycles of 95ºC (30sec), 60ºC (30sec), 65ºC (17min); followed by a final step at 65ºC for 10 min. All PCR amplicons were checked on 1% agarose gel in TAE buffer. A standard ethidium bromide staining was used for band visualization.

The PCR band obtained with the combination 14F-3R was purified with the High Pure PCR product purification kit (Roche Diagnostic, Basel, Switzerland). Thirty nanograms of purified product were used as template to obtain the whole sequence of the amplicons in 4 sequencing reactions using primers 14F, Kinase-5R, Kinase-6R and 3R (see **Table CII.2**).

Table CII.3 Annotated transcripts found on the region where the 20 annotated SNPs (**Appendix CII.5**) are located on scaffold6 of peach genome

Transcript ID	Start	End	Length	Protein prediction
ppa015129m	24389492	24392166	2031	Leucine Rich Repeat
ppa024472m	24398087	24400912	2760	Reverse Transcriptase
ppa025511m	24405493	24407745	2223	Leucine Rich Repeat
ppa015767m	24409461	24413344	3162	Leucine Rich Repeat
ppa023752m	24413575	24416245	1407	Leucine Rich Repeat

CII.2.5 Functional prediction and phylogenetic tree construction

The protein sequence of the ppa025511m gene round allele was obtained from GDR webpage (Jung *et al.,* 2008). Similarity searches were performed on the NCBI web page (www.ncbi.nlm.nih.gov) against the nr (non-redundant collection of sequences in GenBank) and the UniProtKB/SwissProt databases, using the blastp and the Position-Specific iterated BLAST algorithm (Altschul *et al*., 1997). The quality of the pairwise sequence alignment was evaluated under a BLOSUM62 protein substitution matrix allowing a gap existence value of 11 and an extension value of 1.

Full-length amino acid sequences of thirty five receptor-like protein kinases with known functions (**Appendix CII.2**) representing most of the LRR-RLK genes in the *Arabidopsis thaliana* (L.)

Heynh. genome (Gou *et al.,* 2010) were obtained by searching a public database available at (NCBI, www.ncbi.nlm.nih.gov).

For multiple sequence alignment (MSA) and phylogenetic analysis, protein sequences were analyzed by using MAFFT online tool (www.ebi.ac.uk/Tools/msa/mafft/). The weighing matrix used for the MSA alignment was BLOSUM82 with the penalty of gap opening 3 and gap extension 0.2. The obtained MSA was used as an input file to construct the phylogenetic trees by the Neighbor-Joining (NJ) (**Figure CII.9**) method (Saitou & Nei, 1987) and Maximum likelihood (**Figure CII.8**) based on the JTT matrix-based model (Jones *et al*., 1992) using MEGA6.0 software (Tamura *et al*., 2013). The bootstrap (Felsenstein, 1985) consensus trees were inferred from 1000 random replicates.

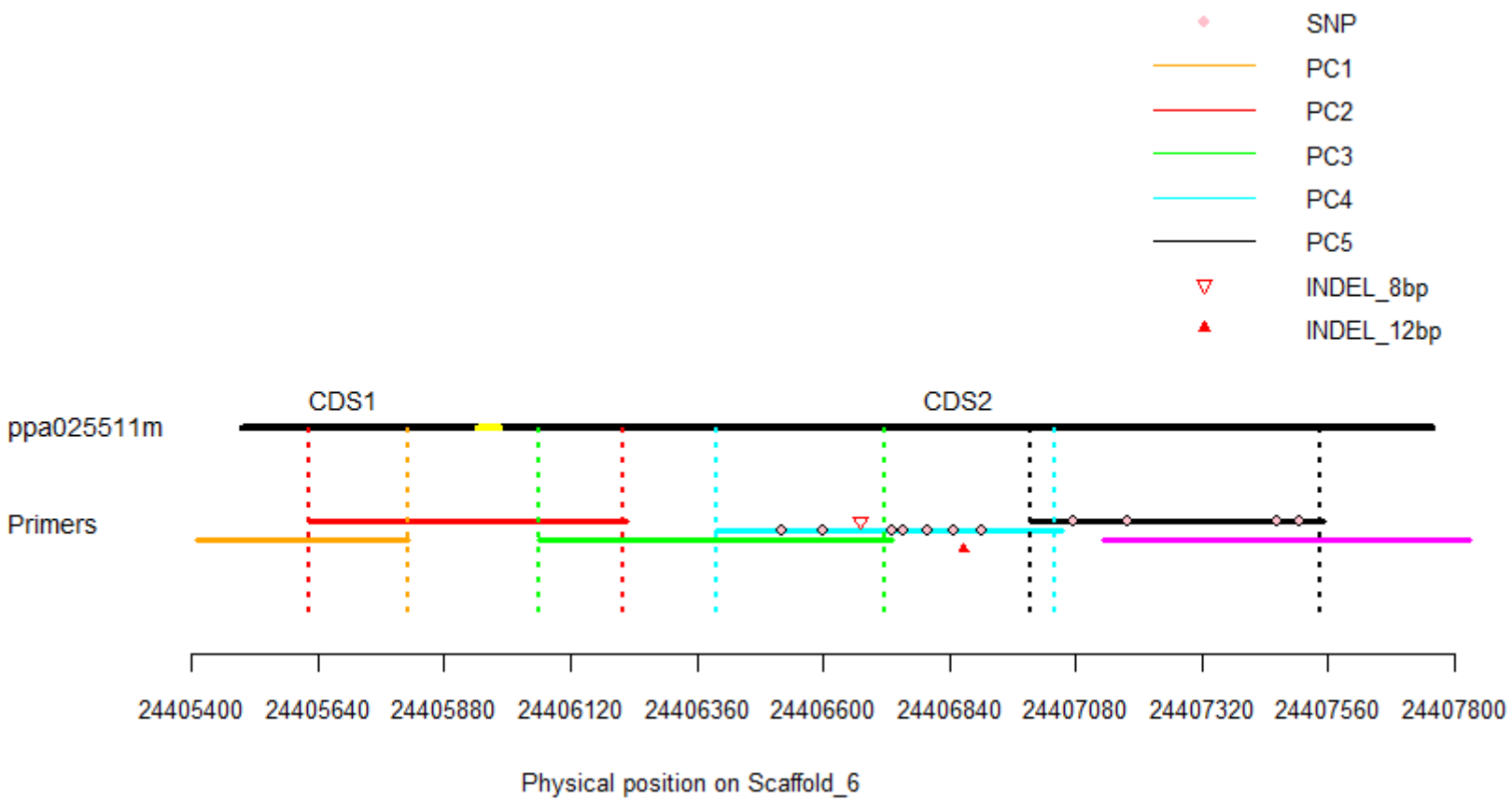


Figure CII.1 .Graphical representation of the overlapping amplicons for the whole sequencing of the candidate gene ppa025511m (LG6) and the variations found within the gene between the testing samples. Coding DNA sequences of ppa025511m are showed by CDS1 and CDS2. And in yellow its intron. Primer combinations (PC) details are shown in **Table CII.2.**

CII.3 RESULTS

CII.3.1 Gene discovery: search of SNPs associated to the flat shape trait

To find polymorphisms associated to the flat trait in peach, we explored a 30kb region flanking the SSR UDP98-412, previously reported to be highly linked to this trait (Picañol *et al.,* 2012). For this we designed 14 primer pairs to amplify fragments of 350-680bp along this region which were used in a small set of flat ('Mesembrine', 'Paraguayo Delfín' and 'Subirana') and round peaches ('Garcica', 'HoneyGlo' and 'Luciana'). Considering the large variability observed genome-wide between round and flat peaches (Aranzana *et al.,* 2002), the large extension of LD in peach (Aranzana *et al.,* 2013) and the dominant nature of the flat allele, which must be in heterozygosis in varieties with viable fruits, we expected to find a large level of heterozygosis the region flanking the marker associated to the trait, and subsequently close to the gene. Surprisingly no polymorphisms were observed in this region with these primers.

The SNPs closest to UDP98-412 reported in the peach genome occurred 337.5 kb upstream this marker, with 20 SNPs (**Appendix CII.5**) in a 26.75kb region (scaffold6: 24,392,166-24,416,245). All these SNPs occurred in coding regions of 5 annotated transcripts (**Table CII.3**). We confirmed the 20 *in silico* SNPs, plus 10 additional, in the same set of flat ('Mesembrine', 'Paraguayo Delfín' and 'Subirana') and round peaches ('Garcica', 'HoneyGlo' and 'Luciana') by sequencing nine amplicons (**Table CII.1**). Thirteen out of the 30 SNPs showed association with the flat phenotype in the small panel of cultivars. All these 13 SNPs occurred in two consecutive amplicons of the transcript ppa025511m (named here Amplicon 5 and Amplicon 6). The sequences of the two amplicons did not overlap but aligned 180bp apart and covered a region of 1150bp. In addition to the 4 SNPs identified in Amplicon 5 we also detected one INDEL in heterozygosis in flat varieties producing a not legible alignment. To confirm the association of the SNPs and the INDEL with the phenotype we sequenced the two regions (Amplicon 5 and Amplicon 6) in 98 varieties (46 round, 53 flat) and 3 aborting phenotypes from a F1 population of the two flat varieties 'UFO3'x 'SweetCap'. All round varieties were homozygous for eleven out of the 13 SNP alleles previously found while all the flat ones where heterozygous; the aborting seedlings where homozygous for the alternative allele, which is concordant with the genetics of the trait. The two SNPs not linked to the trait segregated among the round varieties and occurred in Amplicon-5 (**Appendix CII.1**). The alignment of the round and aborting sequences of Amplicon-5 discovered 2 INDELs instead of the one initially thought: an 8 bp deletion in round peaches and, few bases downstream, a 13 bp deletion in the aborting cultivars (**Fig. CII.2, C and D**). Forward and reverse sequences revealed that all flat varieties had both INDELs

in heterozygosis. By cloning the fragment in one flat variety ('UFO-8') we confirmed that the flat allele was coincident with the one occurring in the aborting individuals. The haplotypes observed in the varieties tested is shown in **Fig. CII.3**. Additionally, we used two primers (Flatin-1F and kinase-5R) flanking the two INDELs to genotype the 98 varieties.

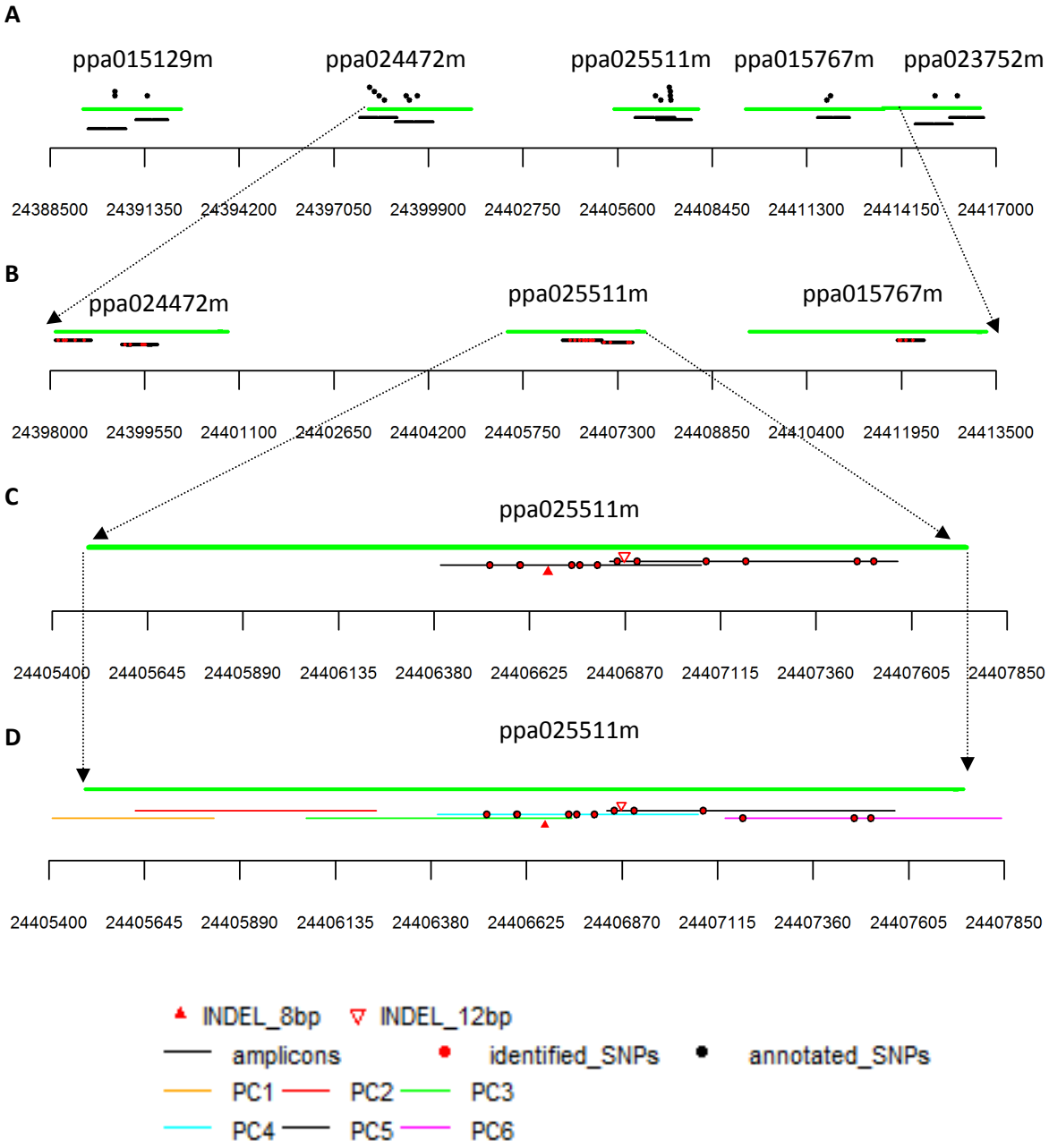


Figure CII.2 Strategy followed to find and sequence the candidate gene ppa025511m on scaffold 6. In green are represented the transcript found in the studied region; black dots represent the *in silico* SNPs; red dots represent SNPs validated or new discovered by sequencing. **A.** Region with annotated SNPs in databases (http://www.rosaceae.org/gb/gbrowse/prunus_persica/). **B.** The studied region was narrowed down to that conformed by amplicons containing SNPs. **C.** It represents the two amplicons containing associated variations to the flat trait and the candidate gene. **D.** Overlapping amplicons for the amplification and sequencing of the candidate gene.

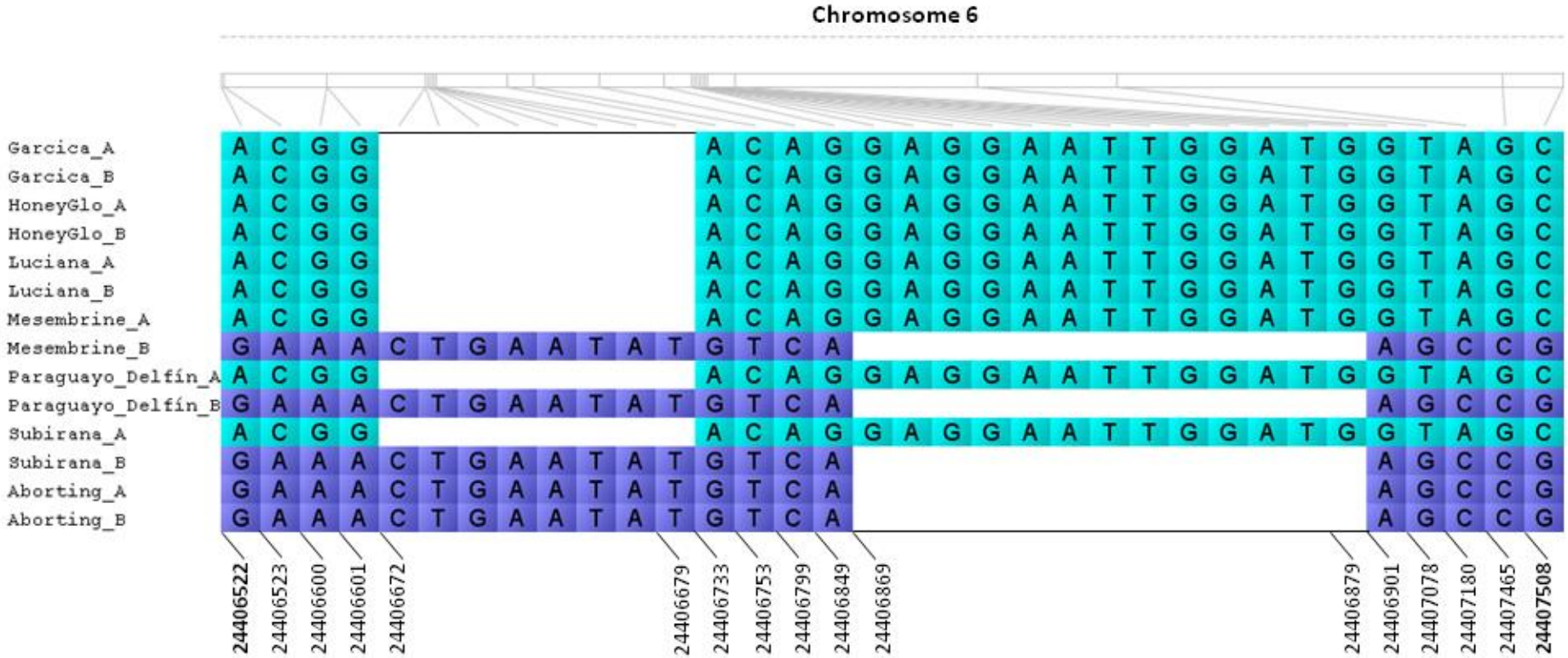


Figure CII.3 Graphical representation of the two haplotypes found for the flat trait in the candidate gene ppa025511m for the samples which constituted the testing set. In purple is shown the aborting haplotype and in turquoise the round haplotype. The flat varieties had both haplotypes in heterozygosis. All 13 associated SNPs and both INDELs are represented with their physical location

The size of the fragments confirmed that all round varieties presented the allele size corresponding to the round sequence (with a deletion of 8 bp and insertion of 13 bp) in homozygosis and all flat varieties presented such fragment in heterozygosis together with one alelle 5bp shorter (containing an insertion of 8 bp and a deletion of 13 bp respect to the round one).

Additionally, we confirmed the co-segregation of this polymorphism and the trait in 69 F1 progenies from the cross between the two flat peaches 'UFO3' and 'Sweet Cap'. The allele specific primers Flatin-1F and Kinase-5R amplified the flat and the round alleles in flat genotypes while the round and aborting seedlings where homozygous for the respective expected allele.

CII.3.2 Gene description: whole sequencing analysis

According to the genome annotation, Amplicon-5 and Amplicon-6 are part of coding regions of an annotated transcript 2,223bp long (ppa025511m; scaffold_6:24,405,493.. 24,407,745). The gene is 2,253 bp long and contains 2 exons of 449 and 1,774 bp long respectively, and an intron 30 bp long. This gene codifies a binding protein and contains Leucine Rich Repeat domains (LRR).

The gene ppa02551m is located in scaffold 6 of the peach genome (Verde *et al.,* 2013) in a cluster of 6 LRR-kinase genes covering all a region of 42.8 Kb. An alignment of this gene with the peach genome shows two partial hits with two peach LRR-kinases, one in scaffold 7 (ppa024468m_LG7:12,624,185..12,627,118; 912 bits, E=0.0) and the other in scaffold 8 (ppa022349m_LG8:6,209,344..6,212,744; 720 bits, E=0.0).

We designed nested primers to obtain the whole sequence of the flat and round alleles in two round ('Garcica' and 'Honey Glo') and two flat ('Paraguayo delfin' and 'Mesembrine') varieties. In total we obtained 2023pb of the region, and due to the high homology of this gene with other LRR kinase genes we could not obtain a unique sequence of the extreme 3' of the gene. The gaps produced by the INDELs in the sequence of the flat varieties were covered with the sequence of the aborting samples. With the nearly whole sequence of the gene we did not detect polymorphisms additional to the previously reported, all occurring in the second exon. The first two SNPs occurred at positions 1,030 (scaffold_6:24,406,522) and 1,031 (scaffold_6:24,406,523) of the peach reference genome v1 and consisted, respectively, in a transition (A in round peaches and G in the aborting ones) and

in a transversion (C in the round allele and A in the aborting one) producing an amino acid change Glu/Thr. The following two SNPs located at 1,108 (scaffold_6:24,406,600) and 1,109 (scaffold_6:24,406,601) with a G/A transition in both polymorphisms and produced an amino acid change from a Gly/Asn. The 8bp insertion in the aborting allele at position 1,180 (scaffold_6:2,446,672-2,446,680) consisted on a repeat of the 8 previous bases ('CTGAATATA') and produced an insertion of two amino acids in the protein (Leu and Asn) and a posterior shift in the reading frame changing the protein sequence leading to a STOP codon. The deduced protein sequence for round varieties contains 750 amino acids (**Appendix CII.3**)

No polymorphisms were found in the first exon of the candidate gene, which contrasts with the high variability observed in the second one. One hypothesis to explain such large variability was the possible amplification of two homologous regions. To confirm or discard this hypothesis we sequenced one aborting and one round sample with a forward primer placed in the first exon of the gene (kinasa 6F) and a reverse primer in the second exon (kinasa 5R). Surprisingly the two amplicons were identical to the one observed in the round allele while the 8bp deletion for the flat allele was missing. The lack of SNPs in the first exon of the gene and this unspecific amplification suggested that we were not able to amplify part of the flat allele when using primers designed in its first exon of the gene. To further confirm this hypothesis we used two primers flanking the gene (3F and 3R); only the round allele could be amplified (**Fig CII.4 A**) producing a band of the expected size (3.3 Kbp), while the flat did not amplify indicating a big polymorphism few nucleotides upstream the 8bp deletion affecting the the 5' UTR and the first intron of the gene.

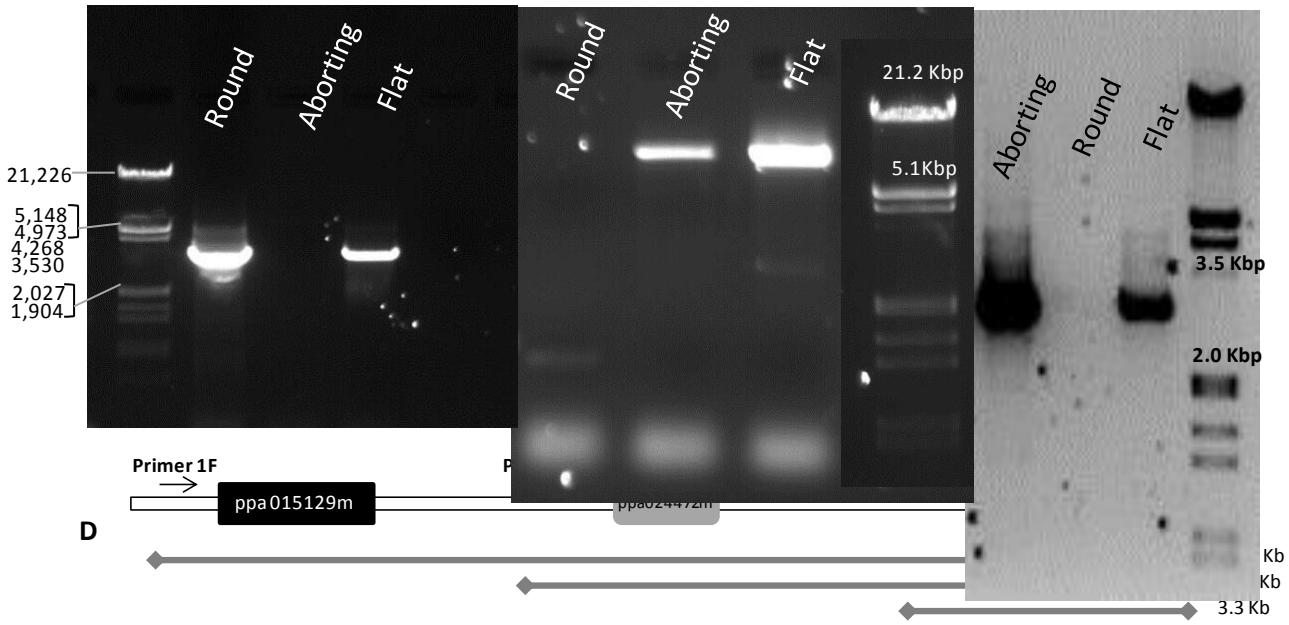


Figure CII.4. Variant discovery in three seedlings from 'UFO3'x'Sweet Cap': one with round fruits, one with aborting and one with flat peaches. **A**) Long-Range PCR with primers 3F-3R flanking ppa025511m produce band with the expected size in round and flat genotypes, but not in the aborting one. **B**) Long-Range PCR with primers 1F-3R covering a region of 18.9 Kb in the peach reference genome v.1. (Verde *et al.,* 2013). **C**) Long-Range PCR with primers 14F-3R covering a region of 12.9 Kb in the peach reference genome v.1. (Verde *et al.,* 2013). In both **B** and C we only could amplify the flat allele, which in both amplicons has a size about 10 Kbp shorter than expected revealing a big deletion. **D**) Position of the fragments in the peach genome.

CII.3.3 Flat allele cloning

To amplify the flat allele we designed primers 16.1Kb and 10.1kb upstream ppa025511m (primers 1F and 14F, respectively). Combining each of these two primers with primer 3R (558bp downstream the second CDS region of ppa02551m) we only obtained the flat allele while we were not able to amplify the round one. In both cases the flat allele had a size about 10Kb shorter than expected (**Fig. CII.4 B and C**). To obtain the whole sequence of the flat allele we sequenced the fragment obtained with 14F/3R (of about 2.8Kb) plus two primers inside the gene (kinase-5R and kinase-6R). In total we obtained a sequence of 2,912 nucleotides, which is 9,970bp less than the reference sequence obtained in round genotype. The polymorphisms respect to the reference genome consisted in the loss of a region starting 9,324 bp upstream of the CDS1 (scaffold_6: 24,396,169) of the gene and ending 693bp downstream the CDS1 (scaffold_6: 24,406,186) lacking all CDS1, the 30bp intron and 214 bp of CDS2.

CII.3.4 Functional prediction and phylogenetic tree construction

To predict the function of the protein we searched for similarity of the protein sequence of the round allele with other proteins in the non-redundant (nr) protein and in the UniprotKB/Swiss-Prot databases (**Appendix CII.4**). The first best hits with the nr protein database were other proteins containing LRR domains annotated on the peach genome. Two of them (ppa015129 and ppa015767) located on chromosome 6 and relatively close to ppa025511m (13.3Kb upstream and 1.7Kb downstream, respectively); one on chromosome 7 (ppa024468m) and one on chromosome 8 (ppa022349m). Additionally we obtained several hits with Receptor-like protein 12-like (RLP-12-like) in *Fragaria vesca* and *Glycine max*.

The best hits of the translated protein with other proteins in the Uniprot/Swiss-Prot database were with GASSHO1 (GSO1) and GASSHO2 (GSO2) proteins, essential for the normal development of epidermal surface of *Arabidopsis* embryos (Tsuwamoto *et al*., 2008). Hits were obtained also for other Leucine-rich repeat receptor-like protein kinases (LRR-RLKs) belonging to the same LRR-family; the LRR-XI family such as CLAVATA-1 (CLV-1) which encodes a putative LRR-RLK that controls shoot and floral meristem size and determines the balance between undifferentiated and differentiated shoot and floral meristem cells in *Arabidopsis* (Clark *et al.*, 1997); then also a LRR-RLK codified by BARELY ANY MERISTEM 1 (BAM1), which is necessary for male gametophyte development, as well as ovule specification and function, and it is also involved in cell-cell communication processes, required during early anther development and regulates cell division and differentiation to organize cell layers. Furthermore, BAM1 is required for the development of high-ordered vascular strands within the leaf and a correlated control of leaf shape, size and symmetry. Additionally it may regulate CLV-1dependent CLV-3-mediated signalling in meristems maintenance (Deyoung & Clark, 2008; DeYoung *et al.,* 2006; Hord *et al.,* 2006). In the same family of LRR-RLKs *Arabidopsis* proteins we also found a hit for an homologous to HAESA gen, which controls floral organ abscission (Jinn *et al*., 2000) and a putative LRR-RLK called PXL-1 which is very closely related to PXY (a receptor-like kinase essential for maintaining polarity during plant vascular-tissue development) and it seems to act synergistically with PXY (Fisher & Turner, 2007). On the other hand, we also found a hit of LRR-RLKs belonging to the *Arabidopsis* LRR X subfamily, encoded by EXCESS MICROSPOROCYTES-1 gene (EMS1/EXS) which controls somatic and reproductive cell fates in anther development. In seeds, it determines cell size and the rate of embryonic development (Zhao *et al*., 2002).

Our protein blasted also with LRR-kinases involved in pathogen response such us *Arabidopsis* FLS2 (Flagellin Sensing 2) (Nürnberger & Kemmerling, 2006) and PEPR1, an homolog of BAK1 (BRI1 brassinolide; BL steroid hormone associated receptor kinase 1) (Li *et al.,* 2002).

Phylogeny analysis of the round allele protein of ppa025511m with full length amino acid sequences of 35 LRR-RLK proteins with known biological function in *Arabidopsis* (**Appendix CII.2**) was performed by heuristic search (or the Neighbor-Joining (NJ) algorithm) and by Maximum likelihood (ML) (**Fig CII.8 and CII.9**) Both trees resulted in a similar topology, revealing a group of protein members of the same subfamily of LRR-RLK, the LRRII. This subfamily split in two well supported branches, one conformed by proteins involved in antiviral defense response, and the other branch by those LRR-RLK involved in BR signaling/male sporogenesis and pathogen response. We also obtained a big cluster of proteins belonging to the LRRX subfamily and also two proteins from LRRXIII subfamily (although its branch is not well supported), another cluster of proteins from the LRRXI subfamily and another one well supported that groups proteins of LRR XIII subfamily involved in organ growth and stomatal patterning differentiation.

The round protein is clustering with GSO2 protein, a LRR-RLK involved in epidermal surface formation during embryogenesis and also close to proteins involved with pathogen response and proteins such ERECTA, which determines organ shape in *Arabidopsis* (Torii *et al.,* 1996) and ERL1 and ERL2.

CII.3.5 Gene validation

We studied this gene in a round peach generated from a sport mutation of the flat variety 'UFO-4' (**Fig CII.5**). Eight highly polymorphic SSRs were used to confirm that they were clones (**Table CII.4**).

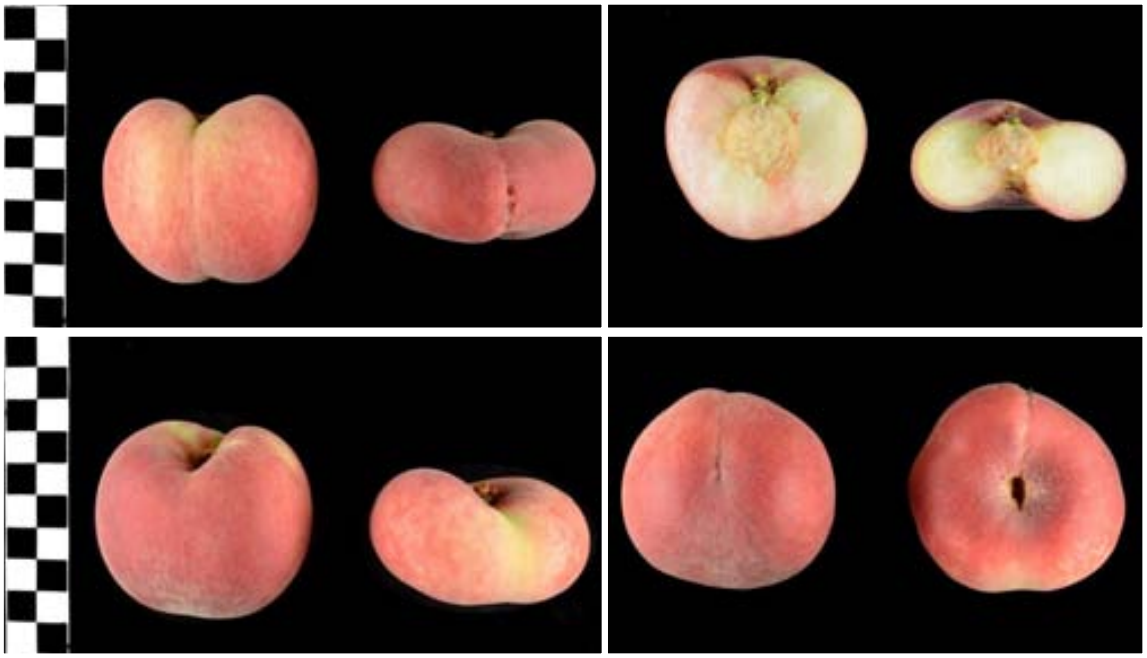


Figure CII.5 Differences in shape of 'UFO4' (right side of each picture) and its sport round mutant (on the left side of each picture).

Table CII.4 Details of the 16 SSRs used to validate that mutant 'UFO4' was a clon of'UFO4'.

SSR	MAP	LG	cM	Ta	Physical position	Fluorescence label	References
CPPCT042	TxE	1	38	62,5	scaffold_1:39307938	HEX	(Aranzana *et al*., 2002)
UDP96-005	TxE	1	29,2	57	scaffold_1:13903361	FAM	(Cipriani *et al*., 1999)
CPSCT021	TxE	2	39,4	42	scaffold_2:23734599	HEX	(Mnejja *et al*., 2004)
BPPCT001	TxE	2	20,9	57	scaffold_2:16134154	HEX	(Dirlewanger et al., 2002)
UDP96-008	TxE	3	36,4	57	scaffold_3:16946762	FAM	(Cipriani *et al.,* 1999)
BPPCT039	TxE	3	18	57	scaffold_3:5802709	NED	(Dirlewanger *et al.,* 2002)
CPSCT005	TxE	4	53,8	62	scaffold_4:29887942	NED	(Mnejja *et al.,* 2004)
UDP98-024	TxE	4	11,3	57	scaffold_4:3499623	FAM	(Yamamoto *et al.,* 2005)
BPPCT014	TxE	5	44	57	scaffold_5:16626108	FAM	(Dirlewanger *et al.,* 2002)
UDP97-401	TxE	5	11	57	scaffold_5:5940392	HEX	(Cipriani *et al*., 1999)
UDP98-412	TxE	6	72	57	scaffold_6:24753353	PET	(Vilanova *et al*., 2003)
UDP96-001	TxE	6	17,5	57	scaffold_6:7040757	VIC	(Cipriani *et al.,* 1999)
UDP98-408	TxE	7	23,7	57	scaffold_7:12216594	FAM	(Cipriani *et al.,* 1999)
CPPCT033	TxE	7	38,9	50	scaffold_7:16702195	FAM	(Aranzana *et al.,* 2002)
UDP98-409	TxE	8	44,5	57	scaffold_8:17783528	FAM	(Cipriani *et al.,* 1999)
UDP96-015	PxF	8	11,3	57	scaffold_8: 3336823	FAM	(Dettori *et al*., 2001)

An analysis of leaf's DNA with Flatin-1F and Kinase-5R showed a faint amplification of the flat allele in the mutated round cultivar compared with the strong signal observed in the original flat (**Fig. CII.6**). This differential amplification could be due to a mutation in one layer of the meristem (LI, LII and LIII) originated in the branch producing round peaches. To evaluate this possibility we used the same primers in DNA from fruit skin (LI), flesh (LII) and stone (LIII). The amplification showed that the flat allele was absent in the flesh mutated DNA while it was present in the skin DNA. Faint amplification of the flat allele was observed in the stone DNA of the mutant, which could be due to a chimeric mutation in LIII with LII cells.

The amplification of mutated flesh DNA with the primers flanking ppa025511m (P3F and P3R) produced only the round allele which was identical to the 'UFO4' round allele. On the other side, the combination of primers P14F/P3R in the mutated flesh did not amplify the flat allele. All these results reveal a mutation in the flat allele producing a reversion of the phenotype and confirm, thereafter, ppa025110 gene as the one responsible for the flat shape in peach.

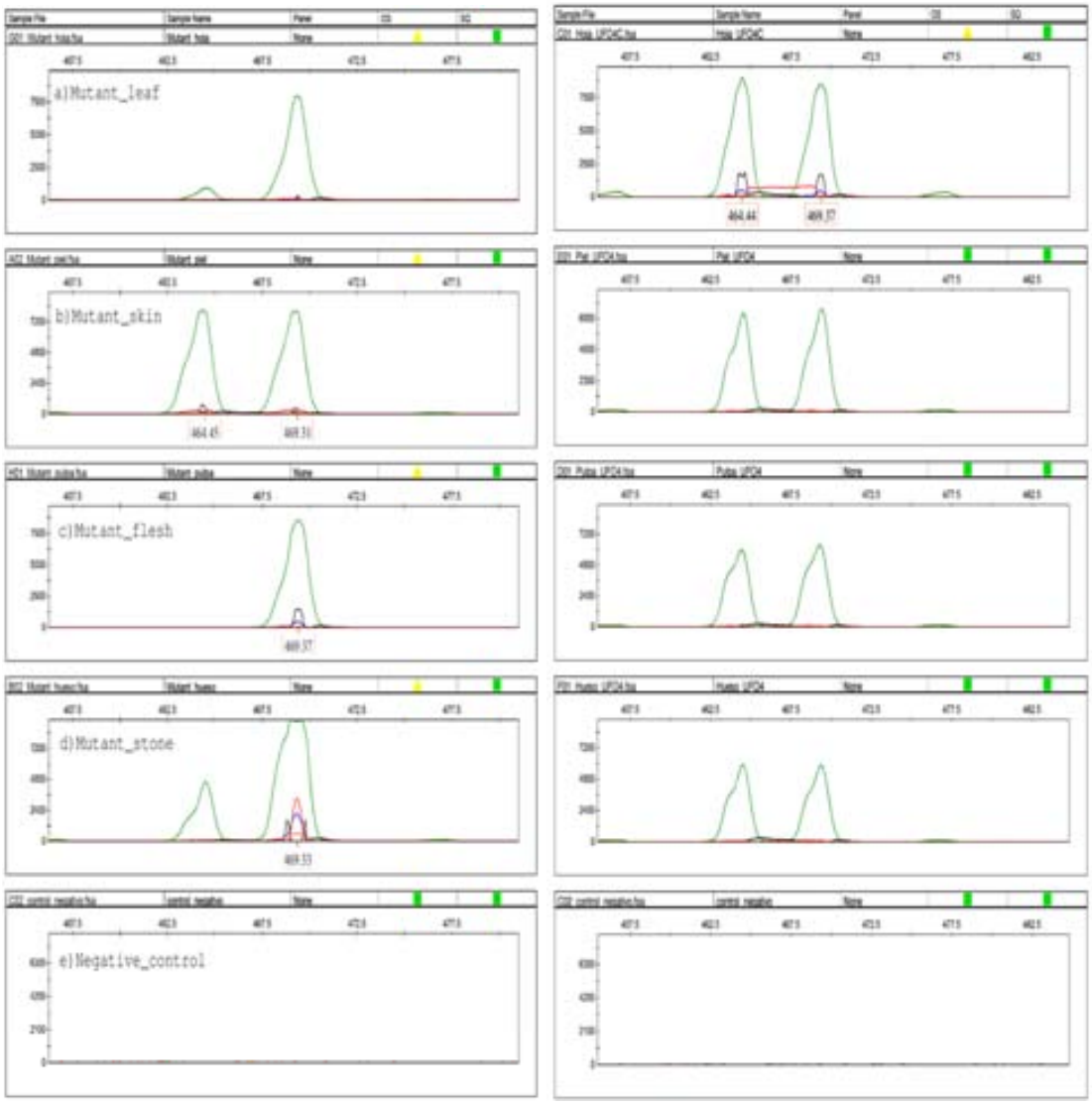


Figure CII.6 Allelic profile obtained with the amplification of UFO4 (right) and its round mutant (left) DNA with the allelic specific primers Flatin1F + kinase5R. DNA was extracted from: leaves; fruit skin; fruit flesh; fruit pit. The 464 bp fragment corresponds to the flat allele and the 469 fragment corresponds to the round allele.

Figure CII.7 Phylogenetic tree of *Arabidopsis* LRR-RLKs proteins with known functions and the predicted protein derived from the round allele of the candidate gene ppa02551m, inferred using the Maximum Likelihood method (Jones *et al.,* 1992). The tree with the highest log likelihood (-39837.7524) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbour-Join and Bio NJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. There were a total of 1561 positions in the final dataset. Evolutionary analyses were conducted in MEGA6.0 (Tamura *et al.,* 2013). Colors correspond with the LRR subfamily to which each protein belongs to **(Appendix II.2)**

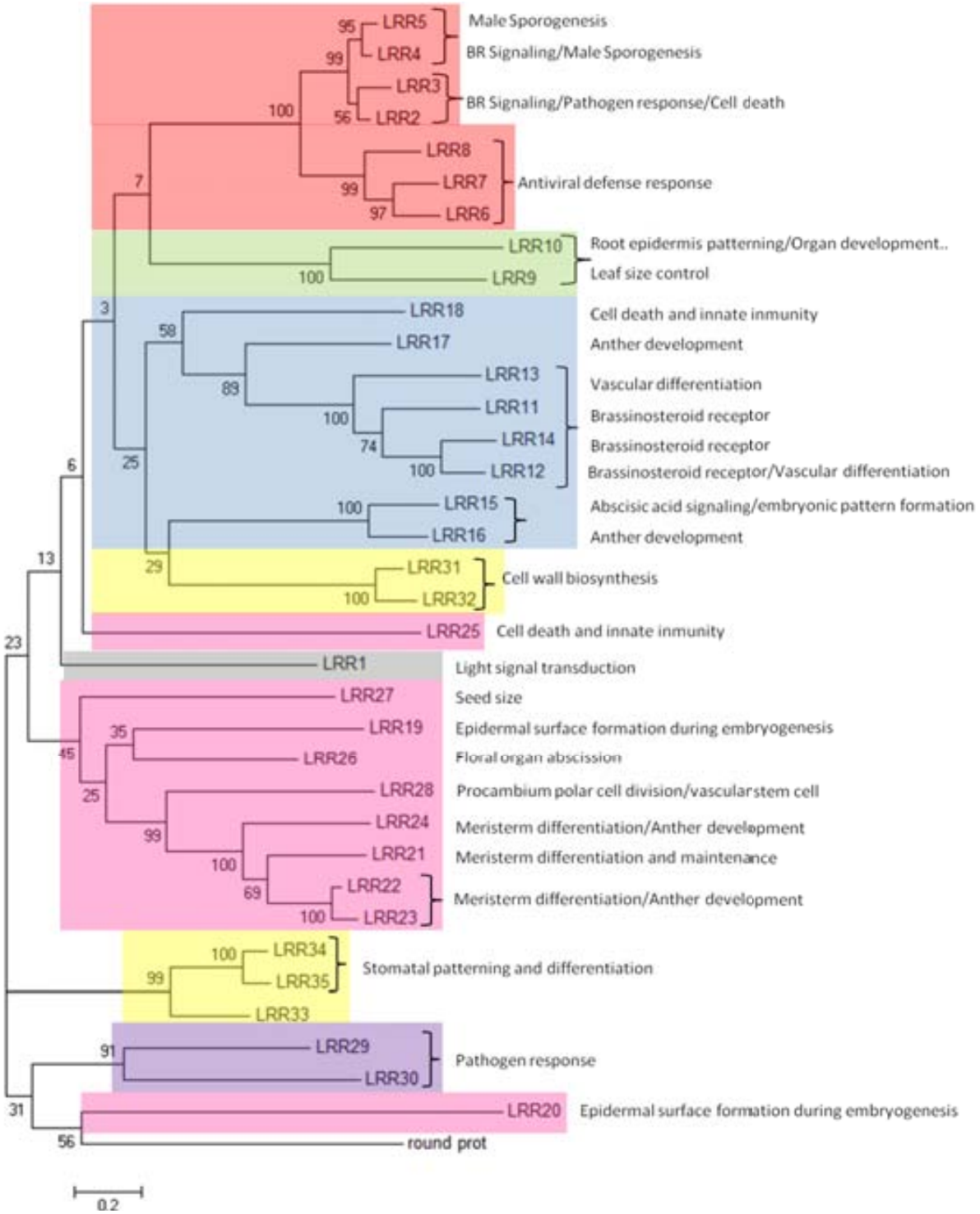
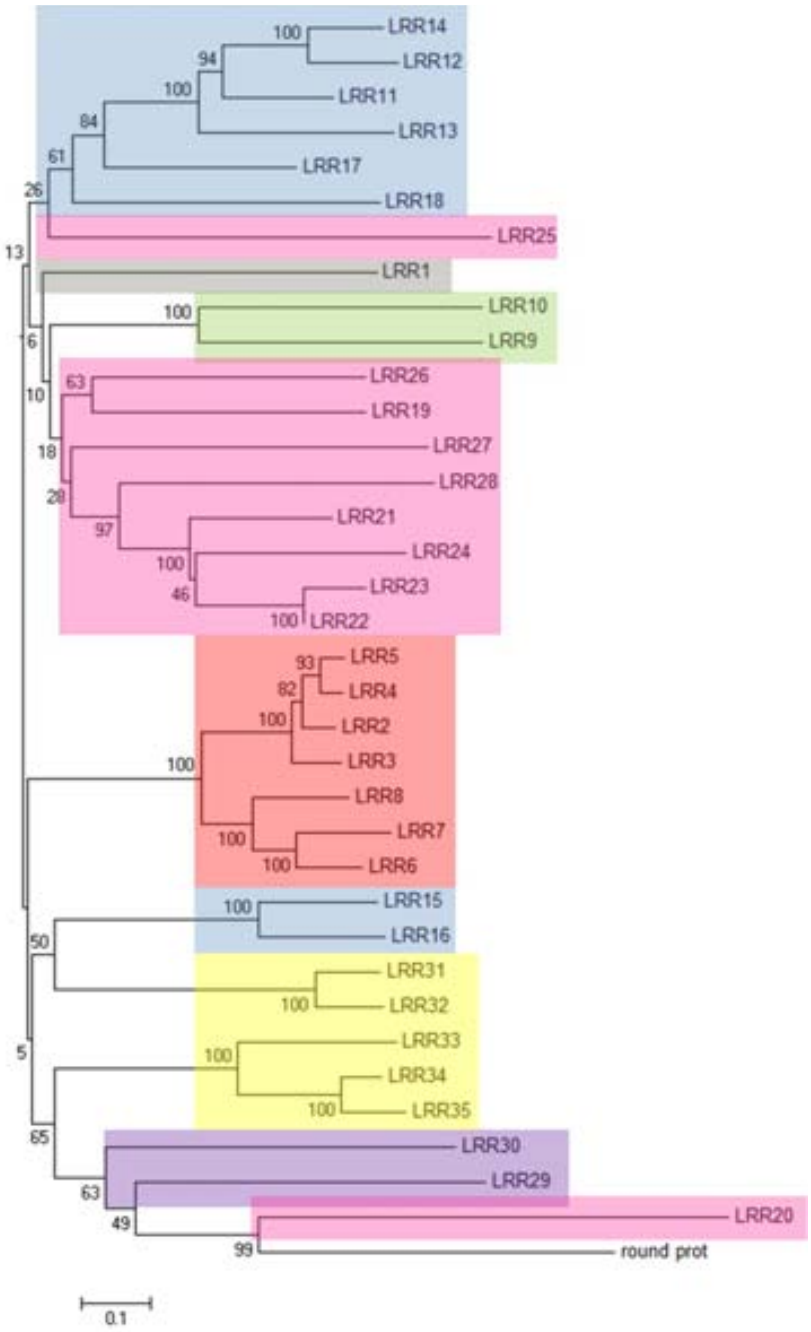


Figure CII.8 Phylogenetic tree of *Arabidopsis* LRR-RLKs proteins with known functions and the predicted proteins derived from the round and aborting allele of the candidate gene ppa02551m, inferred using the NJ method (Saitou & Nei, 1987). The optimal tree with the sum of branch length = 13.75479208 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (5000 replicates) are shown next to the branches (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method (Zuckerkandl & Pauling, 1965) and are in the units of the number of amino acid substitutions per site. The analysis involved 37 amino acid sequences. All ambiguous positions were removed for each sequence pair. There were a total of 1895 positions in the final dataset. Evolutionary analyses were conducted in MEGA6.0 (Tamura *et al.,* 2013).



CII.4 DISCUSSION

Here we have used association genetics to clone a gene responsible for the flat shape in peach. In literature there are several examples where this method, often in combination with linkage mapping, has identified genes responsible for the studied trait. This is the case for example of the major gene Scmv1; a positional candidate for the resistance to sugarcane mosaic virus (SCMV) in maize which was fine mapped in a segregating population and also mapped by association mapping using a panel of inbred lines with different levels of resistance to SCMV (Tao *et al.,* 2013). Here we explored the region with the closest SNPs to the SSR marker reported to be tightly linked to the flat shape in Occidental peaches (UDP98-412). The reasons for searching a long and highly polymorphic region were: i) the described genome-wide variability between flat and round peaches, ii) the obligated heterozygosity of the flat varieties and iii) long LD in peach (Aranzana *et al.,* 2010). This region 337.5Kb upstream to the cited SSR marker contained one SNP every 521bp, close to the SNP density of 1 SNP every 598 bp found by Aranzana *et al.,* (2012) after sequencing few genes in peach varieties, but much higher than the density of 1 SNP every 1076bp observed by (Cao *et al.,* 2014) in Chinese edible varieties.

After sequencing 23 amplicons in a panel of varieties we identified SNPs in two amplicons highly associated with flat, round and aborting phenotypes. The amplicons identified belong to the gene ppa025511m (scaffold_6:24,405,493-24,407,745) annotated as binding protein (GO:0005515) containing leucine-rich domains (leucine-rich repeat-containing N-terminal, type 2). The gene codifies 2 CDS (ppa025511.CDS1 and ppa025511.CDS2). By amplification, cloning and sequencing part of the gene we could identify 11 SNPs and two INDELs co-segregating with the trait, which allowed us to design an allelic specific marker diagnostic for this trait. This marker was validated in 98 varieties, including occidental and oriental varieties and some of those where UDP98-412 alleles escaped the association with the trait (Picañol *et al.,* 2012), as well as in 69 F1 progenies from the cross between two flat peaches 'UFO3'x'SweetCap'; in all cases the genotype obtained was in concordance with fruit shape phenotype. In consequence, we provide here a pair of primers able to amplify two fragments differing in 5bp useful for MAS. Additionally to this primer combination we make available here several SNPs that can be used for the same purpose.

Curiously all polymorphisms between flat and round cultivars were restricted to the ppa025511.CDS2. Further analysis identified a non-specific amplification when using primers from ppa025511.CDS1 in aborting individuals. Long-range PCR reactions detected a 9.97Kb

deletion in the flat allele involving most of the gene: the first exon and the intron, which together represent the first 479bp of the candidate gen. The lack of the first exon and the intron produce the loss of four LRR domains on the first 200 amino acids of the protein. Surprisingly we could not amplify the region in round varieties although the primers were designed from the available genome sequence, which belongs to a round peach.

The protein of the round wild allele is similar to some receptor-like kinases (RLKs) containing leucine-rich repeats. Receptor-like kinases are transmembrane proteins which have an amino-terminal extracellular domain that varies in structure, a single membrane-spanning region and a cytoplasmic protein kinase catalytic domain, which is activated by ligand binding to the extracellular receptor domain in order to produce a response through a signal transduction pathway to the extracellular information (Walker, 1994).

These proteins constitute ligand-receptor systems that control cell fate specification, mediate correct cell divisions and cell to cell communication which allow a correct generation of tissues and organs through growth and development of both animals and plants (Cock *et al.,* 2002). The general mechanism of RLKs starts by binding of an extracellular signal ligand which induces receptor dimerization, which allows the approximation of the intercellular kinase domain facilitating its autophosphorylation followed by its activation that will activate downstream signaling proteins to regulate a cellular response (Becraft, 2002).

Plant RLKs can be classified into 6 classes based on the structural feature of the extracellular domain. The largest class of plant RLKs is the LRR-RLKs class (700 in *Arabidopsis* and 1400 in rice) (Matsushima & Miyashita, 2012), proteins that contain leucine rich repeats, which are tandem repeats of approximately 24 amino acids with conserved leucines involved in protein-protein interactions. The protein of ppa025511 belongs to this LRR-RLK class. Most LRR-RLKs are involved in embryonic pattern formation, which suggests a putative role of our protein in the coordination of cell proliferation during embryogenesis and during morphogenesis of embryonic cells at meristems. In *Prunus*, as in most plants, the entire shoot system derives from post-embryonic development in shoot apical meristems, where similar developmental mechanisms to the embryonic cell to cell signaling are mediated through RLKs (Dodsworth, 2009).

All proteins obtained in the BLAST search were RLKs that mediate cell interactions in adult plants during post-embryonic development. However, the most similar RLK to our protein is the At5g44700 protein codified by GSO2, involved in the maintenance of the

epidermis at the beginning of the heart stage during embryogenesis in *Arabidopsis.* GSO2 is very similar to GSO1 which are functional redundant between the two. Double mutant gso1-1 gso2-1 homozygous show mutant embryos that expand laterally at heart torpedo transition stage of embryogenesis and in later stages such as at the late torpedo stage, the cotyledon adds to the endosperm peripheral tissue arrest ordinary vertical development and causes the embryo to bend in reverse. Thus, we could hypothesize that the LRR-kinase protein codified by the round allele of ppa025511 is involved in a possible cell signaling pathway during peach development that ensure a final round shape.

Gene function is usually validated by genetic transformation or by the screening of mutants. The main obstacle in validating candidate genes in peach through genetic transformation is the regeneration of transformed plantlets. Although some works have reported the transformation and regeneration of stable transgenic plantlets in peach (Hammerschlag *et al.,* 1989; Padilla *et al.,* 2006; Pérez-Clemente *et al.,* 2005) this is not a well resolved method yet. As in other species with similar limitation one of the alternatives is to modify their orthologous genes in other systems easily to transform like *Arabidopsis* or tomato. For example (An *et al.,* 2012) validated the role of the peach gene PpLFL in flower induction overexpressing it in *Arabidopsis*. Similarly (Cohen *et al.,* 2014) silenced orthologous of CmPH genes in tomato and cucumber to validate its role controlling melon acidity.

Although the screening of big collections of natural or induced mutants in vegetative species like *Arabidopsis* (Austin *et al.,* 2011; Martín *et al.,* 2009), melon or tomato (Minoia *et al.,* 2010) is becoming highly successful in gene function validation the generation and maintenance of such big populations in tree species is not viable. However the spontaneous natural generation of sport mutants in peach is frequent and some cases are reported in literature (Brandi *et al.,* 2011; Conte *et al.,* 1994; Dermen *et al.,* 1972; Dermen, 1953, 1956) The main problem of these mutations is that they are usually chimeric and, thereafter the mutation occurs only in some tissues and most of times are not sexually transmitted. Here we have been able to validate the role of ppa025511 gene in the control of fruit shape by studying a chimeric natural mutation occurring in the meristematic LII (producing the fruit flesh tissue) which reverted the flat to the round phenotype. Although we have not been able to obtain the sequence of the mutated flat allele yet, the analysis of flesh DNA with the allele specific primers for the ppa025511.CDS2 INDELS and with primers on regions flanking the gene reveal a new structural mutation in the flat allele, while the skin DNA shows the intact flat and round alleles.

The flat allele in flat varieties, which acts as dominant, lacks the promoter as well as a big portion of the wild allele which may cause a loss of function. This, together with the inheritance of this trait, resembles the mechanism of a haploinsufficient locus. Loss-of-function alleles at haploinsufficient loci are typically dominant because the level of gene function in a heterozygote is below the threshold to produce a wild-type phenotype and homozygotes typically exhibit more sever phenotypes, including early lethality (Meinke, 2013). The most common explanation is that these loci are involved in cellular processes sensitive to dosage effects and changes in protein concentration (Birchler & Veitia, 2010; Veitia & Birchler, 2010). In *Arabidopsis* only few cases of haploinsuficiency have been documented. Among them we find the case of ERECTA-family genes (ERECTA (ER) and its two paralogs ERECTA-LIKE 1 (ERL1) and ERL2 that encode for leucine-rich receptor-like kinases that act coordinating cell shape and inflorescence architecture. All three are functionally related having an overlapping but unique transcript expression pattern. In absence of functional ER and ERL1, Arabidopsis plants heterozygous at ERL2 exhibit female sterility because they develop an aberrant ovule growth and abortion of the embryo sac. Thus, a single copy of ERL2 is haploinsufficient for female sterility although sufficient for floral patterning and inflorescence elongation is still (Pillitteri *et al.,* 2007).

As we have reported here, a mutation in the flat allele can produce a reversion to the round shape. One hypothesis of mechanism is that the flat allele acts as a dominant-negative allele which would produce a mutant protein which could not bind properly to other proteins producing poisoned complex for the functionality of the cells or which stops the normal degradation of the proteins due to substitutions or the absence of protein interaction domains (Meinke, 2013). The function of this dominant-negative allele would be truncated in the sport mutation, recovering then the wild round phenotype. The dominant-negative mechanism has been reported in LRR genes in most of CLAVATA-1 (CLV1) alleles (Diévart *et al.,* 2003). The CLV pathway is the best known plant receptor-like kinase cascade which controls the size of the central stem cell pool in the shoot apical meristem and the differentiation at the shoot and flower apical meristem. In this pathway, the receptor-like protein CLV1 seems to dimerize with CLV2, while CLV3 is a dodecapeptide which acts as the ligand for CLV1 (De Smet *et al.,* 2009). clv1 mutants accumulate stem cells at the shoot and apical meristem, leading to enlarged meristems and additional floral organs. There is variability in the severity of these phenotypes depending on the location of the mutation within the CLV1 protein. The strongest phenotypes of these plants are found when the lesions are located at the extracellular domain, location

that it is very unusual for the dominant-negative mutations in protein kinases or in Tyr kinases of animals, which normally are clustered at the ATP binding site of the kinase domain that correspond with domain II of CLV1. However, no clv1 alleles have been identified that contain mutations in the catalytic regions of the kinase domains, which highlight the necessity of some catalytic activity within the CLV1 for the dominant-negative behavior. Furthermore, the extracellular domain seems to be important in this negative behavior since the chimeric receptor composed by the extracellular domain of CLV1 and the BRASSINOSTEROID INSENSITIVE (BRI1) kinase domain gives rise to phenotypes with the same level of severity than the clv1 mutants (Diévart *et al.,* 2003).

Another hypothesis for the gain of function compatible with the haploinsuficiency mechanism is the recombination of mutant flat allele with other of the LRR-Kinase present around ppa025511 LRR-kinase. In fact ppa025511 clusters with other LRR-Kinases and shows high homology with ppa015129, located 13.3Kb upstream from the wild round allele and 3.3Kb from the flat mutated allele. Thus, the existence of a new functional dominant recombinant allele at this locus could be explained by the genetic recombination with a LRR kinase located nearby our candidate one. There is no evidence of such chimeric kinase receptors in nature but as previously mentioned, chimeric kinase receptors made in the lab (Albert *et al.,* 2010; Diévart *et al.,* 2003) can drive expression as the endogenous genes, so the mutant flat allele could be fused by recombination with another kinase receptor composing a new round allele reverting the function of the receptor. In fact, sequence divergence, genetic recombination, duplication events and selective forces have been proved to be the main forces for the continuous RLK gene expansion and representing a specific plant adaptation that lead the production of variable cell surfaces and cytoplasmic receptors. One additional example of that are the genes that encode proteins containing a nucleotide-binding site (NBS) and C-terminal leucine-rich repeats (LRRs) which diverge more rapidly than the rest of the genome due to their pathogen response (Guo *et al.,* 2011).

Cloning the new mutated allele will provide information of the gene mechanism, thereafter next experiments will pursue to obtain the round allele as well as the mutant one.

CHAPTER III: Somatic variability between peach to nectarine sport mutants and its implication in the *G* locus

CIII.1. INTRODUCTION

Peach is a species with low levels of genetic variability, which represents an important handicap for its genetic improvement. Peach intraspecific variability is the major, but not the unique, source of variability used for commercial breeding and has been broadly studied with SSRs (Aranzana *et al.* 2003; Aranzana *et al.,* 2002; Li *et al.,* 2013) and recently with SNPs (Micheletti *et al.* in preparation;Cao *et al.,* 2014). Additionally, variability due to somatic or vegetative mutations, which arise naturally in many plant groups (Hartmann & Kester, 1975), sometimes represent a valuable tool for the development of new cultivars and for the identification of the causal gene responsible for the mutant trait (Falchi *et al.,* 2013). Contrary to intraspecific variability, the knowledge of the levels of somatic variability in peach is very limited. The only available data corresponds to the analysis of 28 sport mutants with 50 SSRs, estimating a mutation rate of $2.1x10^{-3}$ per allele (Aranzana *et al.,* 2010). Some interesting mutant phenotypes have been already resolved genetically in some species like grapefruit (Hartmann & Kester, 1975), banana (Simmonds, 1966) or potato (Howard, 1970). There are also some examples of somatic mutations in peach in flower shape, maturity day, flesh color and glabrous skin (Scorza & Sherman, 1996). Recently, the comparison of two peach sports showing a different flesh color (yellow and white) has been a successful strategy for the identification of a candidate gene for such trait (Brandi *et al.,* 2011).

In this thesis we study sport mutants from peach to nectarine to i) estimate the overall intraclonal variation and ii) use somatic variability to identify the causal mutation from hairy fruit (peach) to glabrous (nectarine).

In total we analyze here six pairs of clones from five different peach varieties. The nectarine sport mutant 'Yuval' arose in 2002 within an Oded peach population from a commercial orchard in Israel. 'Oded' is a white, melting-flesh, cling-stone fruit, and an early season peach cultivar (Dagar *et al.,* 2011). Both 'Julyprince' and 'Flameprince' peach varieties were originated in the Agricultural Research Service-USDA Southeastern Fruit and Tree Nut Research Laboratory in Byron (Georgia) in 1993. Both varieties are part of the called 'prince varieties' in honor of the fruit breeder Vic Prince who made some of the crosses. 'Julyprince' and Flameprince were selected by W.R Okie when it first fruited in 1995 (Okie & Layne, 2008). Julyprince variety ripens at early to mid-July and it bears peaches characterized to be: round, large, yellow fleshed with some red in the stone cavity, freestone, melting, red skin with little pubescence, with a sweet acidic flavor. Julyprince and Flameprince varieties seem to be heterozygous for the stony hard gene due to its slow-softening profile (Okie & Layne, 2008). 'Flameprince' is a medium-large peach variety with a very firm and yellow flesh, melting, a

yellow-red skin color, freestone and with a ripening period in September. We have analyzed here two nectarine sport mutants from 'Flameprince', each originated independently from the other in different orchards. 'Flameprince Pearson' nectarine comes from the Pearson farm (Georgia), while 'Flameprince Ham' nectarine arose in Ham Orchard (Texas). 'Florida Glo' is a white fleshed, low acid, melting and self pollinating peach obtained by the University of Florida. Its sport nectarine mutant was called 'Gal-I'. And finally, 'Large White' is a large round white fleshed peach, with an acid flavor.

The peach pubescence is due to the presence of trichomes on their skin while nectarines are glabrous. Nectarines were originated in North West China around the Tarim basin north of the Kulum mountains (Hedrick *et al.,* 1917), the center of peach diversity (Vavilov, 1951). There is evidence of nectarine's existence in China for over 2,000 years (Yoon *et al.,* 2006). Its introduction into Europe is assumed to occur in parallel to the peach's introduction in the early 1800's. Thus, nectarine arrived at Persia from China, and then it was carried to Greece and Rome and spread into the temperate parts of Europe.

Parkinson, in 1629, was the first using the word "nectarine" in English language. More than one hundred years later, in 1737, Linnaeous classified nectarines as *Amygdalus persica* var. nucipersica L. Later on, William T. Aiton (1766-1849) called netarines as *Amygdalus nectarina.* It was not until the end of XIX century when European experts referred nectarines as subspecie calling them *Prunus persica* var. nectarine. First nectarines were introduced to North American States from England (Fairchild, 1938). Modern nectarine breeding started in the US in the middle of the 20th century, when Anderson introduced in the market the nectarine variety 'Le Grand', a descendant of the accession 'Quetta' discovered near the homonymous city in India (now part of Pakistan) in 1906 (Okie, 1998). Other known sources of the nectarine trait used in modern western breeding programs were 'Goldmine' and 'Lippiatt' discovered in New Zealand in 1900 and 1916, respectively (Okie, 1998). These latter three genotypes are acknowledged as donors of most of the current nectarine cultivars widespread in US and Europe. Modern Japanese breeding programs have extensively used two old European nectarines, 'Precoce di Croncels' and 'Lord Napier', and modern US cultivars (Konishi *et al.,* 1994).

The development of trichomes in peach starts first on the ovary about four weeks before anthesis (Creller & Werner, 1996). The glabrous trait is genetically controlled by a recessive and monogenic trait (*G*/g) (Blake, 1932) mapped in the distal part of the linkage group 5. Dirlewanger *et al.* (2006) placed this locus at position 81.4 cM in LG5 of the linkage

map based on the intraspecific F_2 population J ('Jalousia') x F ('Fantasia'), cosegregating with the AFLP eAC-CAA and flanked by the SSRs CPSCT030 (scaffold_5: 15,126,681..15,127,320) and CPSCT022 (scaffold_5:16,626,112..16,626,607), which corresponds to 14 cM (or 1.50 Mbp). Recently, Vendramin *et al.,* (2014) mapped the *G* locus within an interval of 1.1cM (corresponding to 635kb) in the F_2 progeny from 'Contender' (peach) x 'Ambra' (nectarine). The *G* locus fine mapping and the resequencing data of peach and nectarine varieties allowed the identification of the insertion of a Ty1-*copia* retrotransposon of about 7Kb within the third exon of the transcription factor gene *PpeMYB25* (*ppa023143m) at chromosome 5* (scaffold_5:15,897,836..15,899,002) in the nectarine allele. This insertion introduces an H112L substitution and a premature stop codon (TAA), resulting in a peptide of 112 aminoacids precisely truncated at the C-terminal end of the R3 MYB domain, and in consequence producing a non-functional form of the MYB transcription factor that normally promotes trichome formation in fuzzy peaches. In absence of the insertion, the new CDS encodes a peptide of 330 amino acids similar to the R2R3-MYB transcription factor GhMYB25 from the allotetraploid cotton *Gossypium hirsutum* (58.4% similarity) (Machado *et al.,* 2009) and MIXTA-like1 from *Antirrhinum* (AmMYBML1, 55.3% similarity) (Perez-Rodríguez *et al.,* 2005). *GhMYB25* is differentially expressed in the outer integument of ovules at fiber initiation between mutants and wild cotton lines (Lee *et al.,* 2006; Machado *et al.,* 2009; Wu *et al.,* 2006) and its modified expression affects trichome development in transgenic cotton. AmMYBML1 is involved in the trichome differentiation of the corolla tube of *the* Antirrhinum *flower* (Perez-Rodriguez *et al.,* 2005). The analysis of this insertion in a collection of nectarines has suggested that this is the unique allele responsible for the trait (Vendramin *et al.,* 2014).

In peach the frequency of spontaneous mutations from peach to nectarine is relatively high. Moreover some alleles show a higher predisposition to mutate. Although the origin of the mutated allele fixed in nectarine varieties seems to be unique, the nature of such spontaneous mutations is still unknown.

The candidate gene for the *G* gene was published when the experimental part of this thesis had already ended, and consequently this information was ignored when the experiment was designed and the data analysed. However the discovery has been taken into account in the discussion of the results presented herein.

CIII.2. MATERIAL AND METHODS

CIII.2.1. Plant materials

In this study we sequenced 11 genomes, five from peach varieties ('Flameprince', 'Julyprince', 'Oded', 'Large white' and 'FloridaGlo') and six from nectarine sport mutants derived from them: 'Flameprince Ham nectarine' and 'Flameprince Pearson nectarine' derived from 'Flameprince'; 'Julyprince nectarine' from 'Julyprince', 'Yuval' from 'Oded'; 'Large white nectarine' from 'Large white'; and 'Gal-I' from 'Florida Glo' (**Table CIII.1**).

Genomic DNA was isolated from leaf tissue with either the DNeasy® Plant Mini kit of (Qiagen, CA 91355 Valencia) or the Cesium Chloride density gradient protocol (Messeguer *et al.,* 1994). The concentration and quality of 1ul of the extracted DNA was quantified by spectrophotometer (NanoDrop, technologies, Wilmington, DE, USA) and confirmed by electrophoresis on 1% TBE agarose gel.

CIII.2.2. Genome analysis with SSRs

Clones were confirmed with 16 SSRs distributed along the 8 peach linkage groups (**Table CIII.2**). PCR products were obtained in a volume of 10µl using a 40ng of DNA of each cultivar as a template, primer pairs at 10uM (forward primer fluorescence labeled), 200mM of each dNTP, $MgCl_2$ 2.5 mM, 1.5 units of GoTaq® polymerase (Promega), 1x buffer *5x Colorless GoTaq*® (Promega). The amplification program used consisted in: 2 min at 94C, 35 cycles (25s at 94ºC, 20s at the annealing temperature of each primer pair (Ta) and 20s at 72ºC) followed by 5 min of extension at 72ºC. Then, 1µl of PCR product was mixed with 12µl of formamide (Applied Biosystems) and 0.4µl GeneScan™ 500 LIZ® Size Standard (Applied Biosystems). The amplified fragments were separated by capillary electrophoresis with the automatic sequencher ABI/Prism 310 (PE/Applied Biosystems), and fragment size scoring was done using GeneMapper v.4.0 (Applied Biosystems).

Table CIII.1.Peach and nectarine varieties sequenced. Isolation DNA methodology employed and library file names

Variety	DNA Isolation	Library file name
Flameprince_peach	CsCl	C16V7ACXX_6_3_1.fastq / ….2.fastq
Flameprince Ham_nectarine	CsCl	C16V7ACXX_6_4_1.fastq / ….2.fastq
Flameprince Pearson_nectarine	CsCl	C16V7ACXX_6_5_1.fastq / ….2.fastq
Flameprince Ham_nectarine2	CsCl	D1A9TACXX_6_4_1.fastq / ….2.fastq
Julyprince_Pearson_peach	CsCl	C16KRACXX_7_6_1.fastq / ….2.fastq
Julyprince_Pearson_peach2	CsCl	D1A9TACXX_6_6_1.fastq / ….2.fastq
Julyprince Pearson_nectarine	CsCl	D1A9TACXX_7_7_1.fastq / ….2.fastq
Oded_peach	Dneasykit	D1A9TACXX_8_12_1.fastq / ….2.fastq
Yuval_nectarine	Dneasykit	C172AACXX_7_13_1.fastq / ….2.fastq
Large White_peach	Dneasykit	D1A9TACXX_7_8_1.fastq / ….2.fastq
Large White_nectarine	Dneasykit	D1A9TACXX_7_9_1.fastq / ….2.fastq
FloridaGlo_peach	Dneasykit	D1A9TACXX_8_10_1.fastq / ….2.fastq
Gal-I_nectarine	Dneasykit	D1A9TACXX_6_11_1.fastq / ….2.fastq

Table CIII.2. Characteristics of 16SSRs used to verify the clonal identity of peaches and nectarines.

SSR	MAP	LG	cM	Ta (ºC)	Position in the peach reference genome (v1.1)	References
CPPCT042	TxE	1	38	62,5	scaffold_1:39,307,938	(Aranzana *et al.,* 2002)
UDP96-005	TxE	1	29,2	57	scaffold_1:13,903,361	(Cipriani *et al.,* 1999)
CPSCT021	TxE	2	39,4	42	scaffold_2:23,734,599	(Mnejja *et al.,* 2004)
BPPCT001	TxE	2	20,9	57	scaffold_2:16,134,154	(Dirlewanger *et al.,* 2002)
UDP96-008	TxE	3	36,4	57	scaffold_3:16,946,762	(Cipriani *et al.,* 1999)
BPPCT039	TxE	3	18	57	scaffold_3:5,802,709	(Dirlewanger *et al.,* 2002)
CPSCT005	TxE	4	53,8	62	scaffold_4:29,887,942	(Mnejja *et al.,* 2004)
UDP98-024	TxE	4	11,3	57	scaffold_4:3,499,623	(Yamamoto *et al.,* 2005)
BPPCT014	TxE	5	44	57	scaffold_5:16,626,108	(Dirlewanger *et al.,* 2002)
UDP97-401	TxE	5	11	57	scaffold_5:5,940,392	(Cipriani et al., 1999)
UDP98-412	TxE	6	72	57	scaffold_6:24,753,353	(Vilanova *et al.,* 2003)
UDP96-001	TxE	6	17,5	57	scaffold_6:7,040,757	(Cipriani *et al.,* 1999)
UDP98-408	TxE	7	23,7	57	scaffold_7:12,216,594	(Cipriani *et al.,* 1999)
CPPCT033	TxE	7	38,9	50	scaffold_7:16,702,195	(Aranzana *et al.,* 2002)
UDP98-409	TxE	8	44,5	57	scaffold_8:17,783,528	(Cipriani *et al.,* 1999)
UDP96-015	PxF	8	11,3	57	scaffold_8: 3,336,823	(Dettori *et al.,* 2001)

LG: linkage group; Ta: annealing temperature in °C; TxE = 'Texas' (almond) x'Earlygold' (peach)

CIII.2.3 Library preparation and sequencing

Ten micrograms of high quality DNA of each cultivar at a concentration of >200ng/ul, (OD 260/280 close to 1.8) resuspended in TE (EDTA=0.1mM) were delivered to CNAG ("Centre Nacional de Análisis Genómico", Barcelona-Spain) for Illumina/Solexa sequencing.

High quality Illumina TruSeq libraries were generated to obtain paired-end sequences for all 11 cultivars. The library preparation followed the standard Illumina workflow for paired-end library. Basically, this methodology starts with 1-5ug of genomic DNA as an input, and then, it is fragmented by nebulization to generate <800bp double–stranded fragments. Then, fragments are blunt ended and phosphorylated at 5' ends, while at 3'ends a single 'A' nucleotide is added in order to enable the ligation to an adapter which has a single-base 'T' overhang. Distinct sequence adapters are added at both ends of each strand in the genomic fragment. The products of this ligation reaction are purified and size-selected by agarose gel

electrophoresis. Then, it is produced an enrichment of the size-selected DNA fragments with adapters at both ends by PCR amplification. The library is purified, size-selected by agarose gel electrophoresis and quantified by Agilent 2100 Bioanalyzer (Agilent, Foster city, USA) for validation and prevention of possible presence of contaminants in the prepared library. Furthermore, during PCR amplification, SYBR green fluorescence detection is used as an additional quality check by comparing simultaneously the amplification efficiency of a previously sequenced library with the uncharacterized library. CNAG conducted the Illumina/Solexa flow cell sequencing by running 200 cycles on the Illumina HiSeq 2000. Each flow cell has eight lanes, and it is possible to mix several libraries individually barcoded on the same lane depending on the final sequencing coverage desired. In our case, 13 libraries were produced for the 11 genotypes because two of them were repeated ('Julyprince_peach' and 'Flameprince Ham nectarine' libraries) to reach a good starting DNA concentration for these two samples. The 13 libraries were placed on 3 lanes (first number indicated in the library's file name (**Table CIII.1**) after the code of each sample (i. e C16V7ACXX_**6**_3_1.fastq) and they were sequenced using a specific sequencing primer for each one).In total we received the 26 fastq format (Cock *et al.,* 2010) files containing the sequences.

CIII.2.4 Bioinformatics analysis

CIII.2.4.1 Quality assessment of raw data

As the first step, the quality of the data was evalated with FastQC v.0.10.0 (Andrews, 2010). FastQC generates summary tables and figures of broadly used indicators of the quality of the sequences, including information such as: basic statistics (total number of sequences, sequence length and overall GC%), per base sequence quality expressed as Phred score (Ewing *et al.,* 1998), per sequence quality scores, per base sequence content, per base GC content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels and overrepresented sequences.

Adapters and low quality reads were removed for further analysis. Adapters with a minimum match of 6 nucleotides were removed using *cutadapt*, and bases with low quality were removed by a command line provided by *fastx-Toolkit* (http://hannonlab .cshl.edu/fastx_toolkit/).

We set up the threshold at Phred-like base-calling accuracy score 30, which is equivalent to a probability of 1/1000 of assigning a wrong base. We also removed those reads that after preprocessing were shorter than 35 nucleotides. After quality and trimming (for details in the program code see 'Quality_and_Trimming.sh' in **Appendix CIII.1**) we obtained three output files for each of the 13 paired-end samples analyzed: a trimmed forward file containing paired reads, a trimmed reverse file containing paired reads and a third file containing all the single orphaned reads. Then, we paired forward and reverse reads (for details in the program code see 'Pairing_trimmed_reads.sh', **Appendix III.2**).

CIII.2.4.2 Mapping against the reference genome

Good quality sequences of each library were mapped against the peach reference genome constructed from the sequence of the peach variety 'Lovell' v.1.39 (Verde *et al.,* 2013) available at ("Phytozome"v3) using the Burrows-Wheeler Aligner (BWA) tool (Durbin et al., 2009). The first step in the mapping process performed with BWA was to index the reference genome. Aligned files were then refined by SAMtools (Durbin *et al.,* 2009) using the "sampe" option for each mate of pared-end data and the "samse" option for single data in order to generate alignments in SAM format, which is transformed into its binary representation format, the BAM format. BAM format can achieve a high compression level of the alignment data. Then, the two alignments of each sample are merged and sorted by leftmost coordinates. Next, to make sure that our alignments were fast and random accessible we indexed them again and we added the read group definition to the header of the bam files and to each read present in the alignment. Finally, we indexed the bam files again, removed the possible duplicates and indexed again the final alignment files. All the previous steps were performed following the command lines included in a shell script called 'Align_peach.sh' in **Appendix CIII.3**.

CIII.2.4.3 Mapping quality assessment

The quality assessment of the alignments was performed with three different programs: flagstat command included in samtools (Li *et al.,* 2009), SAMstat (Lassmann *et al.,* 2011) and Qualimap (García-Alcalde *et al.,* 2012), using as input all sorted bam files for each sample. Flagstat command provides various summary statistics from which we extracted the overall percentage of the mapped reads. SAMstat provided: the proportion of reads mapped in

each different mapping quality range. Qualimap v .0.7.1 was used to analyze the quality of the alignments using the option BAM QC in order to compare them with the results provided by flagstat and SAMstat and to evaluate others aspects of the quality not provided by the previous softwares such us the coverage distribution across the reference. In addition, the evaluation of quality provided by Qualimap was performed inside/outside the *Prunus persica*'s genes using the gene annotation gff file available at GDR (Jung *et al.,* 2008). The standard parameters provided by Qualimap were set up to perform the mapping quality analysis. We set up 400 windows to split the reference genome. This value is used for computing the graph that plot information across the reference. Each of these windows included 568bp considering the whole genome reference size.

CIII.2.4.4 Small variant calling

The small variant (including SNPs and small insertions and deletions) detection was performed using SAMtools 'mpileup' (Li *et al.,* 2009). We used a minimum read mapping quality of 20 as a command setting in Samtools 'mpileup'. The output generated by mpileup option checks for each position in the reference and for each sample whether each read mapping to that position has the same nucleotide (or the reserve complement) than the reference or a different one, as well as their qualities. The ouput is in BCF format. Then, the prior probability distribution and the data was used by bcftools (Li, 2011) which is packed in the SAMtools suite, to perform the actual variant calling, assigning the genotypes to each variant site. Thus, at the end we got a file in variant calling format (vcf), which includes the data in tabulated format that allows an easy and fast retrieval of specific sets of data ('Call.samtools.sh', **Appendix CIII 4**).

CIII.2.4.5 Variant filtering

The final vcf file needs to be filtered with the aim of extracting the desired and selective genotyping information. We applied a filter to this vcf file to remove variants with a Phred quality equal or lower than 20 and with a read depth lower than 10, using *vcfutils.pl varFilter*, which belongs to SAMtools suite. Then, we also applied a more restricted filter in the Phred-Likelihood (PL) field which shows Phred scaled Likelihoods of the given genotypes (0/0, 0/1, 1/1) separated by commas (AA, AB, BB). The most likely genotype is given in the GT field and the other likelihoods reflect their Phred-scaled Likelihoods relative to this most likely

genotype. When the most likely genotype was 0/0 (or A/A) we selected the variants that had the following Phred-scaled likelihoods: AA ≤ 10, AB ≥ 50, BB ≥ 50 and BB/AB ≤ 2. When the most likely genotype was 0/1 (or A/B), the selected variants were those having in the PL field: AA ≥ 50, AB ≤ 10, BB ≥ 50, AA/BB ≤ 2 or BB/AA ≤ 2. And if the most likely genotype was 1/1 the selected variants had PL field characterized by: AA ≥ 50, AB ≥ 50, BB ≤ 10, AA/AB ≤ 2 .

CIII.2.4.6 Variant annotation

The annotation of the variants was performed using SnpEff 3.4 software (Cingolani *et al.,* 2012). This software annotates the variants and calculates the effects they produce on genes present in the annotation of the reference genome sequence through an algorithm based on interval trees indexed by chromosome, which is implemented in Java programming language. SnpEff provides a list of binary databases to calculate the effects of each variant query by an efficient interval search on the specific loaded database. We used the peach available database at SnpEff which is based on peach v1.0 genome sequence. The ouput files (HTML and txt) summarize: the position of the SNP on the chromosome, the reference nucleotide, the alternative nucleotide, the type of change (transition/transversions) and the amino acid change between other parameters.

CIII.2.4.7 Nucleotide diversity and Heterozygosity calculation

General nucleotide diversity (π) was calculated as the average number of nucleotide differences (heterozygous or homozygous alternative) per site between each peach sequence and the reference genome sequence, assuming a genome size of 227Mb. Somatic nucleotide diversity was obtained as the average number of nucleotide differences per site between each peach sequence and its nectarine sport mutant assuming that both whole genome sequences have been equally covered by sequencing.

Heterozygosity (Ho) was calculated as the average number of heterozygous nucleotide differences per site between each peach sample sequence and the peach reference genome sequence divided by the total number of nucleotide differences obtained between all the studied samples. Somatoclonal heterozygosity was calculated in the same manner but accounting only for the heterozygous nucleotide differences between clones.

CIII.3 RESULTS AND DISCUSSION

CIII.3.1 Quality test of raw and trimmed sequences

In this chapter we analyze the whole genome variability of 5 peaches ('Flameprince', 'Julyprince', 'Oded', 'Large White' and 'Florida Glo') and 6 nectarine sport mutants ('Flameprince Ham nectarine', 'Flameprince Pearson nectarine', 'Julyprince nectarine', 'Yuval', 'Large White nectarine' and 'Gal-I'). Paired-end sequencing of their genomic DNA with Illumina HiSeq technology produced 13 libraries (for 'Flameprince Ham nectarine'and 'Julyprince' two libraries were required to ensure enough data) and 26 files of sequence data, which contained a total of 1,612,732,418 sequences of 101 nucleotides each (**Table CIII.3**). This corresponds to an average of 27.59 times the peach genome (size 227 Mbp).

Sequence quality was analyzed with FastQC software (Andrews, 2010). Quality scores per sequence were acceptable for all the files with an average Phred value of 37.5 (**table CIII.3**). Per base quality scores were also acceptable for all libraries but 'Julyprince' and 'Florida Glow', each with low quality after the site 99. Seven libraries showed an imbalance of the relative amount of each base in the 10 first sites, which is probably due to the primers and adapters used for sequencing. This fact produced a bias of the GC content in such sites in 'Oded', 'Yuval' and 'Florida Glow'. One of the libraries of 'Flameprince_Ham_nectarine' contained a large overrepresented adapter sequence belonging to the *Truseq* index 4 libarary kit.

The GC content of all the sequences followed normal distributions with mean values ranging from 37% to 40% (**Fig.CIII.1**). These values are close to the 38.71% obtained previously by Ahmad *et al.* (2011) and to the 37.6% reported by Fresnedo-Ramírez *et al.,* (2013) after sequencing 3 peach varieties in both cases. Sequence duplication level ranged from 28.3% in 'Large white nectarine" to 40.6% in 'Florida Glo' and showed slight variations between libraries (**Table CIII.3**). Also minor differencies were observed between each of the two paired-end sequences of each library. When considering all libraries the average level of duplication was 33.81%.

The reads were trimmed and filtered to remove low quality sequences. Only 0.26% of the sequences were removed leaving a final of 1,597,362,979 sequences in the 13 samples, which represent a coverage of 27.33 times the peach genome. In general, we observed a considerable degree of coverage variation within each pair of clones. Differences in the distribution of the coverage along the reference genome should not be ignored when

comparing variant calls between samples since one of the most important criteria for an accurate and sensitive variant calling from Illumina reads depends on an even coverage of sequence data genome wide (Tsai *et al.,* 2013).

In **Fig. CIII.2** we observed that quality scores of each site were always higher or equal to 30 and differencies when comparing peaches and their mutants where minor, which will be relevant for extracting somatic variability. After this step the percentage of duplications was reduced from 33.81% to 31.90%. This range of duplication is in agreement with previous observations reporting 27.33% of duplication in the peach genome due to transposable element sequences plus 7.54% of uncharacterized repeats (Verde *et al.,* 2013). Similarly to what occurred before trimming and filtering, we observed differences in the level of duplication between paired end-sequences of a library as well as between clones. Higher deviations where observed between 'Oded' and 'Yuval', mainly due to a larger proportion of trimmed sequences in the former variety.

CIII.3.2 Sequence alignment and mapping quality

Good quality sequences (Q ≥ 30) of each library were mapped against the peach reference genome (Verde *et al.,* 2013) obtaining 13 alignments. The quality of the alignments were evaluated with three methods: with SAMtools flagstat, SAMstat methods (Lassmann *et al.,* 2011) and with the Bam QC option of Qualimap software (García-Alcalde *et al.,* 2012). In general, the proportion of sequences properly aligned against the peach reference genome calculated by flagstat samtools was always lower than the one calculated by Qualimap (**Table CIII.5**), but smaller than the one calculated by SAMstat (**Fig. CIII.3**).

Using Qualimap we observed that in all libraries the number of paired sequences mapping at different chromosomes was lower than 1%. Integrating the peach annotation v.1 in Qualimap we analized the reads that mapped inside or outside genes. About 33%-35% of them mapped in genes (**Table CIII.4**) which is consistent with the peach genome gene content (36.2%, Verde *et al.,* (2013)). The majority of mate reads mapped correctly in pairs, which should allow the detection of the different types of small structural variations (SVs). However, a low percentage of them (0.06%) were singletons (i.e. pairs with only one of the mate reads mapping) (**Table CIII.4**).

The other method used to evaluate the quality of the alignments was SAMstat, which also provides the percentages of mapped reads within 6 quality ranges. The vast majority

(≥70%) of the reads mapped against the peach reference genome with high quality (MAPQ ≥30, which means that one out of 1000 mapped reads did it erroniously) (**Fig. CIII.3**). In all cases the quality of the alignments of the peach varieties were comparable with the ones of their sport mutants. All sequences from 'Flameprince' and its mutants 'Flamprince_Ham' and 'Flameprince_Pearson' could be mapped and the proportion of alignments with MAPQ ≤3 was low (0.2%). The proportion of reads unable to be mapped in the rest of libraries was also very low (0.2-0.3%), however the proportion of alignments with MAPQ ≤3 was much higher (about 20% of the mapped reads). This could indicate some problems of quality, but the consistency of the results between the peach varieties and their mutants also suggests possible rearrangements or genomic variants in some varieties as the cause of misalignments. Nevertheless, it is difficult to distinguish mapping errors caused by genomic variation from those introduced by sequencing errors and, more likely, by repetitive genomic sequences. Indeed NGS sequencing error rates are relatively low and their effects can often be mitigated with increased genomic coverage but repetitive sequences still create mapping ambiguity (English *et al.,* 2014). Mismatches and insertions were homogeniously distributed across the reads length in all libraries, indicating a good quality of the mapping.

It is particularly evident the valley floor of mapping coverage that occurs at the end of chromosome four for all samples (**Appendix CIII.5**). In addition, at the same position of the decrease of the coverage qualimap showed an increase in the GC content. In many cases the presence of CpG islands is the main impediment producing that some genome regions are less assessed by next generation sequencing methods (Wang *et al.,* 2011). Actually, at the end of chromosome 4, specifically between the positions 30,200,005 and 30,528,708, there is a long annotated repeat ('Repeat_94169') 329 Kbp long, whose sequence GC content is around 53.4 %, which represent a GC bias respect to what occurres at the whole genome level.

Table CIII.3. Description of the sequences obtained and a summary of their quality evaluated with FastQC software.

Library	Total Sequences[1]	Gbp obtained[2]	Coverage[3]	Quality		Per base sequence content[6]	GC Content			Per base N content[10]	% Sequence Duplication (each paired-end library)[11]	Over-represented sequences
				Per base[4]	Per sequence[5]		%[7]	Per base[8]	Per sequence[9]			
Flameprince_peach	53,483,206*2	13.1 *2	23.8	Ok	Ok	<10nt	39	Ok	Ok	Ok	32.6/31.5	Ok
Flameprince_Ham_necta1	70,781,189*2	17.4 *2	31.49	Ok	Ok	<10nt	38	Ok	Ok	Ok	32.2/30.9	Ok
Flameprince_Pearson_necta	51,054,109*2	12.5 *2	22.72	Ok	Ok	<10nt	40	Ok	Ok	Ok	34.9/33.1	Ok
Flameprince_Ham_necta2	61,847,645*2	17.4 *2	27.52	Ok	Ok	-	40	Ok	Ok	Ok	34/31.2	-
Julyprince_Pearson_peach	34,215,586*2	8.4 *2	15.22	>99nt	Ok	Ok	40	Ok	Ok	Ok	31.5/30.1	Ok
Julyprince_Pearson_peach_2	59,279,740*2	18.3 *2	32.79	Ok	Ok	Ok	38	Ok	Ok	Ok	36.8/34.8	Ok
Julyprince_Pearosn_necta	73,704,036*2	14.7 *2	26.38	Ok	Ok	Ok	38	Ok	Ok	Ok	33/31.6	Ok
Oded_peach	70,094,114*2	17.5 *2	31.19	Ok	Ok	<10nt	38	<10nt	Ok	Ok	39.2/37.7	Ok
Yuval_necta	70,941,338*2	17.4 *2	31.56	Ok	Ok	<10nt	37	<10nt	Ok	Ok	39.2/39.2	Ok
Large_White_peach	69,881,617*2	17.4 *2	31.09	Ok	Ok	Ok	37	Ok	Ok	Ok	31.6/30.1	Ok
Large_White_necta	62,030,099*2	15.4 *2	27.6	Ok	Ok	Ok	37	Ok	Ok	Ok	28.8/28.3	Ok
Florida Glow_peach	67,794,284*2	16.9 *2	30.16	>99nt	Ok	<10nt	38	<10nt	Ok	Ok	40.6/39	Ok
Gal-I_necta	61,256,246*2	15.3 *2	27.25	Ok	Ok	<10nt	38	Ok	Ok	Ok	33.6/33.6	Ok

1 Reads obtained per library at each of the 2 paired-end libraries
2 Amount of data in Gbp
3 Coverage considering a genome size of 227Mb
4 Quality values across all bases at each base position
5 Distribution of average qualities per sequence
6 Average percentage of each base at each position
7 Percentage of GC content
8 Average percentage of GC content at each base position
9 Distribution of average GC content, compared to a normal distribution best matching the data
10 Percentage of N at each base position

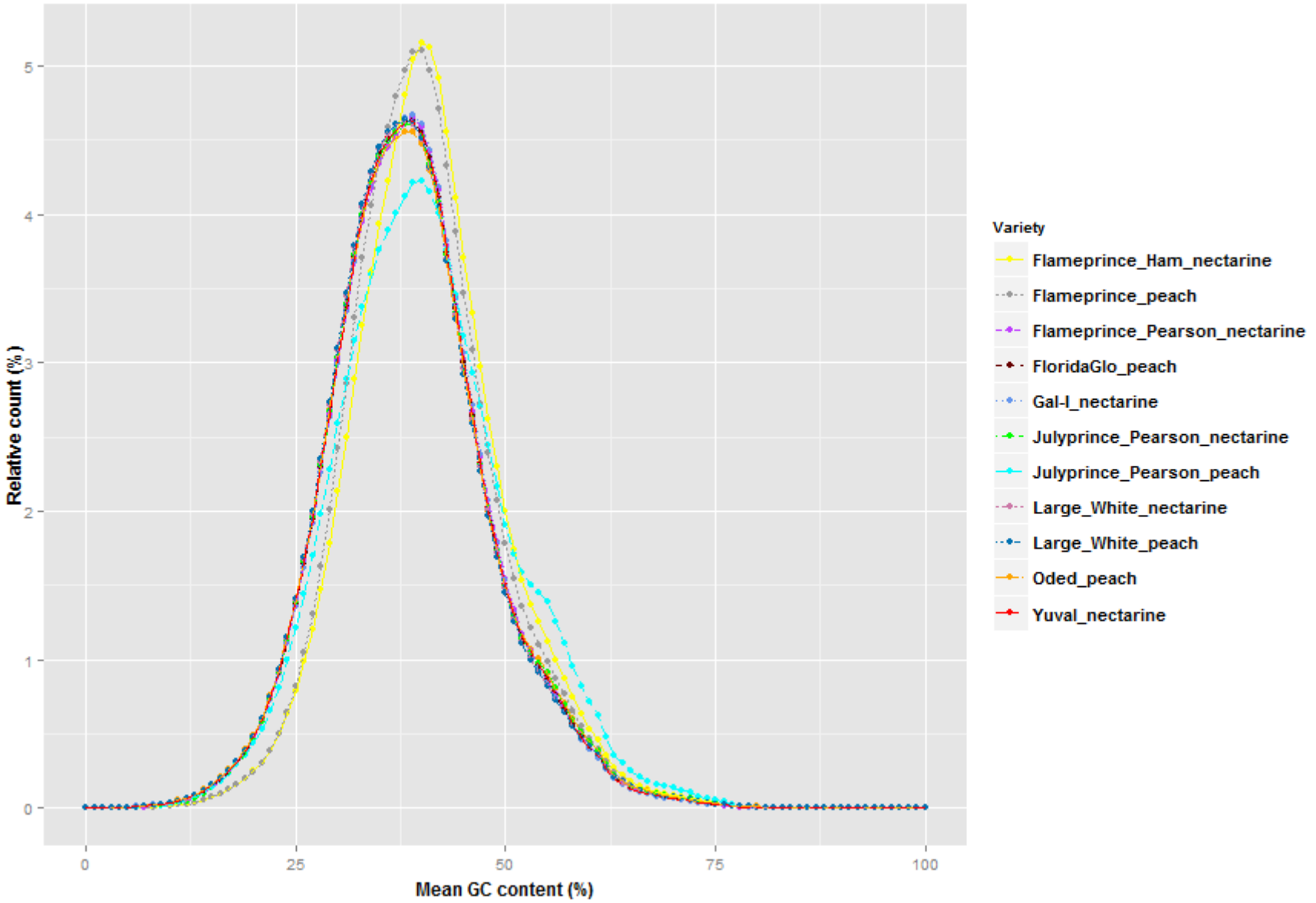


Figure CIII.1. Per sequence GC content. Graph was made in R using ggplot2.

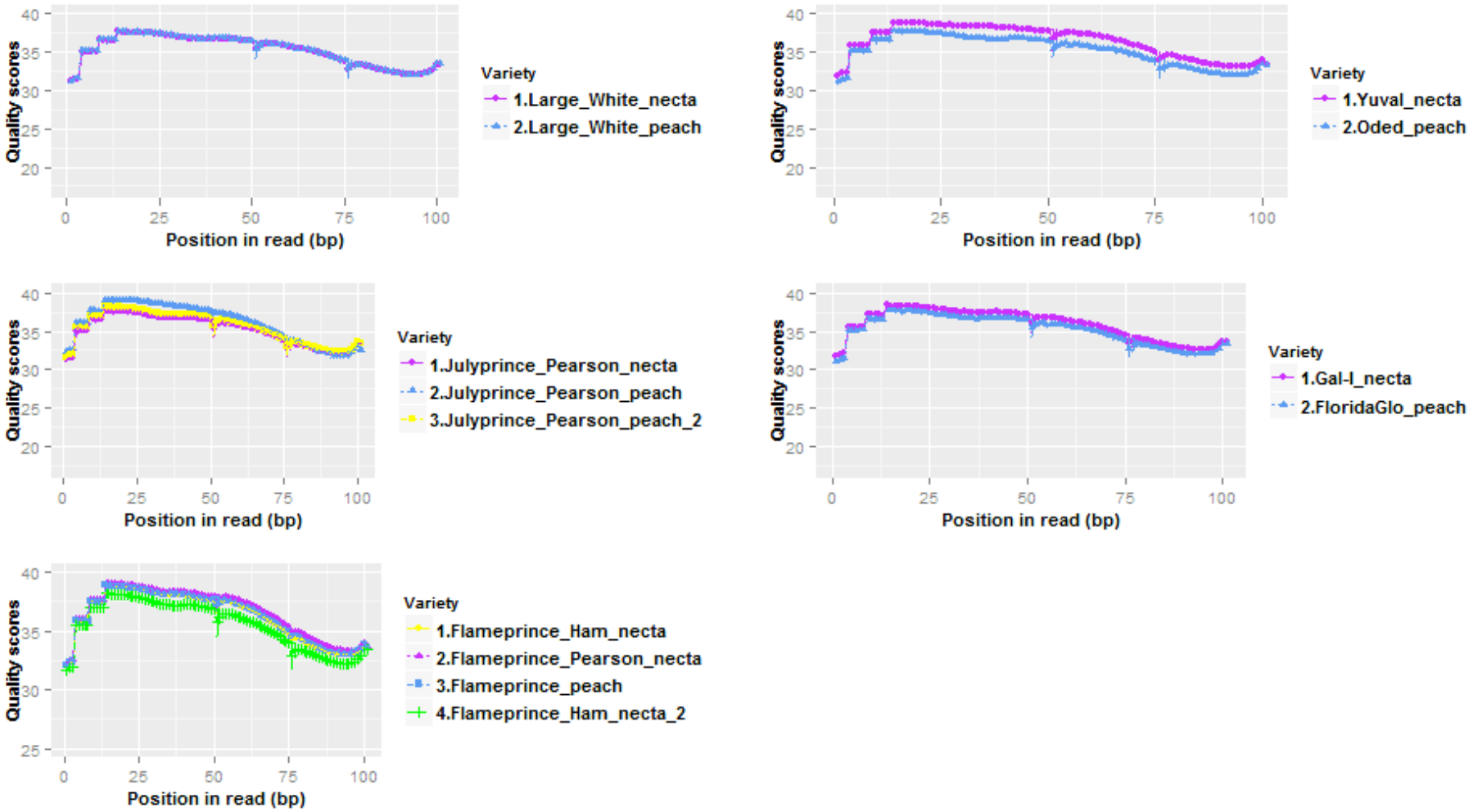


Figure CIII.2 Per base sequence quality analysis obtained by FastQC software after trimming and filtering. Graph was made in R using ggplot2.

Table CIII.4. Basic statistics of the aligment data using Qualimap software (García-Alcalde *et al.,* 2012).

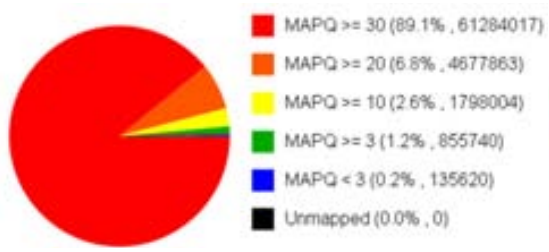
Library	No. Reads[1] (Q>30)	Mapped Reads[2]	Mapped in genes%[3]	Mapped out genes%[4]	Mapped both pairs in genes%[5]	Mapped both pairs out genes%[6]	Singletons in genes %[7]	Singletons out genes%[8]
Flameprince_peach	68,750,754	68,750,649	35.31	64.69	35.06	64.06	0.06	0.26
Flameprince_Ham_necta	92,307,679	92,307,565	33.7	66.3	33.46	65.71	0.06	0.23
Flameprince_Pearson_necta	62,570,483	62,570,364	33.93	66.07	33.54	65.17	0.13	0.4
Flameprince Ham_necta2	76,324,251	76,324,118	34.43	65.57	34.08	64.77	0.12	0.36
Julyprince_Pearson_peach	43,231,247	43,231,167	32.03	67.97	3175	67.27	0.1	0.37
Julyprince_Peason_peach2	86,486,510	86,486,385	32.85	67.15	32.55	66.40	0.09	0.32
Julyprince_Pearson_necta	73,722,467	73,722,367	34.95	65.05	34.70	64.42	0.04	0.18
Oded_peach	79,636,934	79,636,835	35.29	64.71	35.02	64.06	0.04	0.2
Yuval_necta	81,355,006	81,354,913	35.12	64.88	34.96	64.45	0.04	0.18
Large White_peach	91,113,755	91,113,755	35.34	64.66	35.1	64.09	0.03	0.18
Large White_necta	82,791,759	82,791,759	35.48	64.52	35.22	63.91	0.03	0.18
Florida Glo_peach	75,647,568	75,647,461	35.30	64.7	31.81	64.07	0.04	0.22
Gal-I_necta	75,405,553	75,405,437	35.45	64.55	35.22	63.96	0.04	0.21

1. Total amount of reads with quality higher than 30
2. Number of mapped reads
3. Number of mapped reads within genes
4. Number of mapped reads outside genes
5. Number of paired end reads for which both pairs mapped within genes
6. Number of paired end reads for which both pairs mapped outside genes
7. Number of paired end reads for which just a single read of the pair mapped within genes
8. Number of paired end reads for which just a single read of the pair mapped outside genes

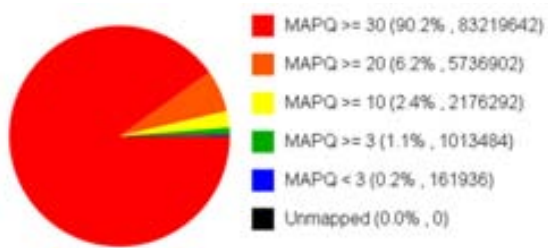
Table CIII.5. Amount of mapped reads against the reference genome using Flagstat command from SAMtools (Li *et al.,* 2009) and Qualimap (García-Alcalde *et al.,* 2012).

Variety	Total reads	Mapped reads using Flagstat	% mapped reads using Flagstat	Mapped reads using Qualimap	% mapped reads using Qualimap
Flameprince_peach	68,750,754	67,479,128	98.15	68,750,649	99.99
Flameprince Ham_nectarine	92,307,679	90,951,834	98.53	92,307,565	99.99
Flameprince Pearson_nectarine	62,570,483	60,619,276	96.88	62,570,364	99.99
Flameprince Ham_nectarine2	76,324,251	74,564,991	97.70	76,324,118	99.99
Julyprince_Pearson_peach	43,231,491	42,318,290	97.89	43,231,167	99.99
Julyprince_Pearson_peach2	86,487,060	84,732,350	97.97	86,486,385	99.99
Julyprince Pearson_nectarine	73,722,870	72,692,639	98.60	73,722,367	99.99
Oded_peach	79,637,358	78,484,662	98.55	79,636,835	99.99
Yuval_nectarine	81,355,445	80,354,520	98.77	81,354,913	99.99
Large White_peach	91,114,277	89,985,361	98.88	91,113,755	99.99
Large White_nectarine	82,792,077	81,706,528	98.69	82,791,759	99.99
FloridaGlo_peach	75,648,022	74,561,674	98.56	75,647,461	99.99
Gal-I_nectarine	75,406,017	74,325,363	98.57	75,405,437	99.99

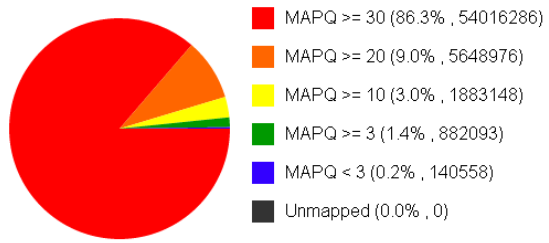
Flameprince_peach



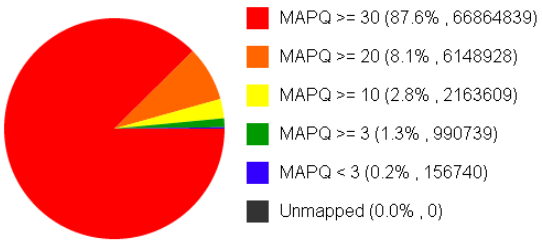
Flameprince_Ham_nectarine_1



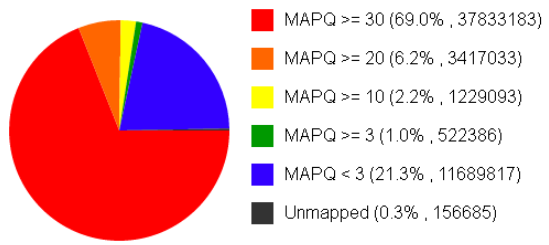
Flameprince_Pearson_nectarine



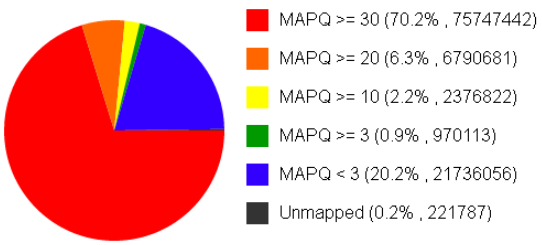
Flameprince_Ham_nectarine_2



Julyprince_Pearson_peach



Julyprince_Pearson_peach_2



Julyprince_Pearson_nectarine

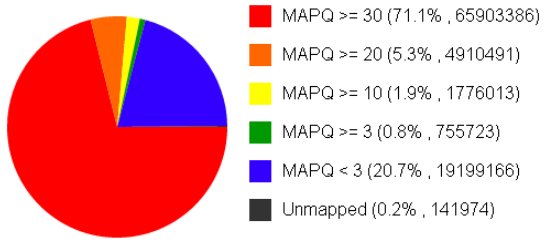
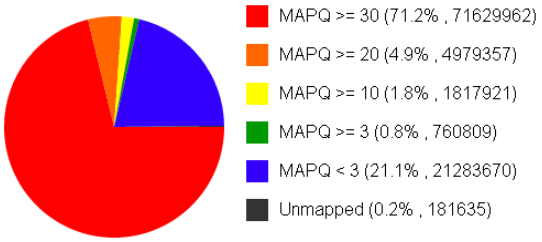
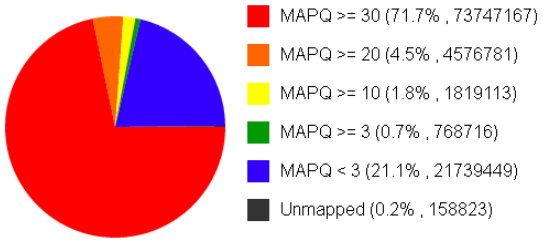


Figure CIII.3. Distribution of mapped reads in mapping quality ranges provided by SAMstat (Lassmann *et al.,* 2011): MAPQ<3, MAPQ≥3, MAPQ≥10, MAPQ≥20, MAPQ≥30, and unmapped reads figures in parenthesis indicate the proportion (in %) and the absolute number of reads.

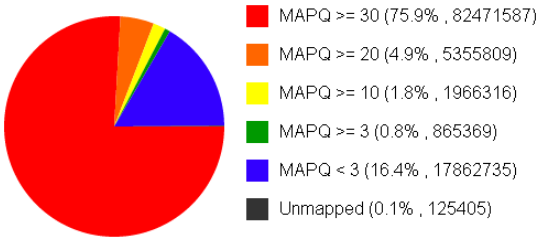
Oded_peach



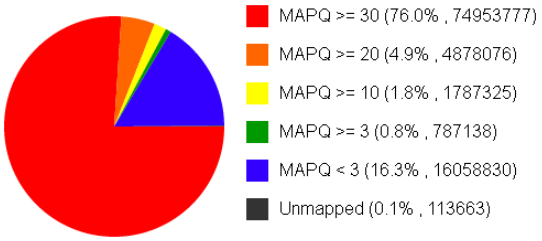
Yuval_nectarine



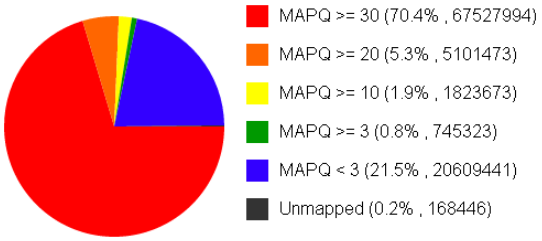
Large White _peach



Large White_nectarine



FloridaGlo_peach



Gal-I_nectarine

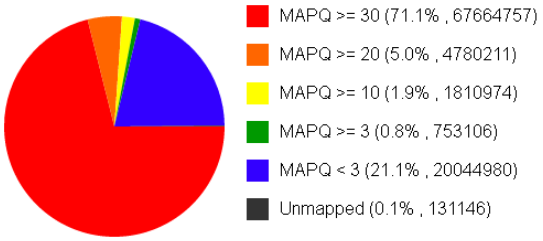


Figure CIII.3 Continued. Distribution of mapped reads in mapping quality ranges provided by SAMstat (Lassmann *et al.,* 2011): MAPQ<3, MAPQ≥3, MAPQ≥10, MAPQ≥20, MAPQ≥30, and unmapped reads figures in parenthesis indicate the proportion (in %) and the absolute number of reads.

CIII.3.3 Genetic variability of the varities: small variants

The small variants (SNPs and INDELs of 50 or less base pairs) between the alignments and the peach reference genome were detected using jointly all the 13 alignments with both MAPQ ≥30 and MAPQ ≥20. The joint analysis was done to reduce false positive SNPs due to low coverage of some libraries and to enhance the power of variant discovery. In general, the reduction of variants called when increasing the mapping quality from MAPQ ≥30 to MAPQ ≥20 was low (3%), while the probability of error decreased considerably (from 1 every 100 reads mapped incorrectly when MAPQ ≥20 to 1 every 1000 reads when MAPQ ≥30). Consequently, we conducted all further analysis with the alignments of higher quality (MAPQ ≥30).

Variability differed between cultivars. The cultivar with more variants when comparing with the reference sequence of peach was 'FloridaGlow_peach' followed by 'Oded', 'Julyprince_peach' and 'Large_White_peach, while the cultivar with less variants was 'Flameprince' which showed almost half of the variations showed by 'Florida Glo_Peach'(**Fig. CIII.4**). One of the alignments of 'Julyprince_peach' (from which two libraries were generated and sequenced) showed a much larger number of variants than the other (576,421in front of 378,774). This higher amount of variants was due to a higher amount of SNPs and INDELs with heterozygous genotypes, that could be suggesting a large amount of false positives explained by the low coverage of this library. Consequently the library C16KRACXX_7_6 was removed from further analysis.

In total, after removing this library, the pipeline used identified 805,506 small variants, which represents one every 282 bp. The most common variants were the SNPs (82.23%) occurring one every 351bp while one INDEL occurred every 1,440 bp. These figures contrast with the ones obtained by (Aranzana *et al.,* 2010) who found much lower density, with 1 SNP every 598 bp and 1 INDEL every 4,189 bp. Aproximately half of the INDELs were due to insertions and half to deletions (**Table CIII.6**)

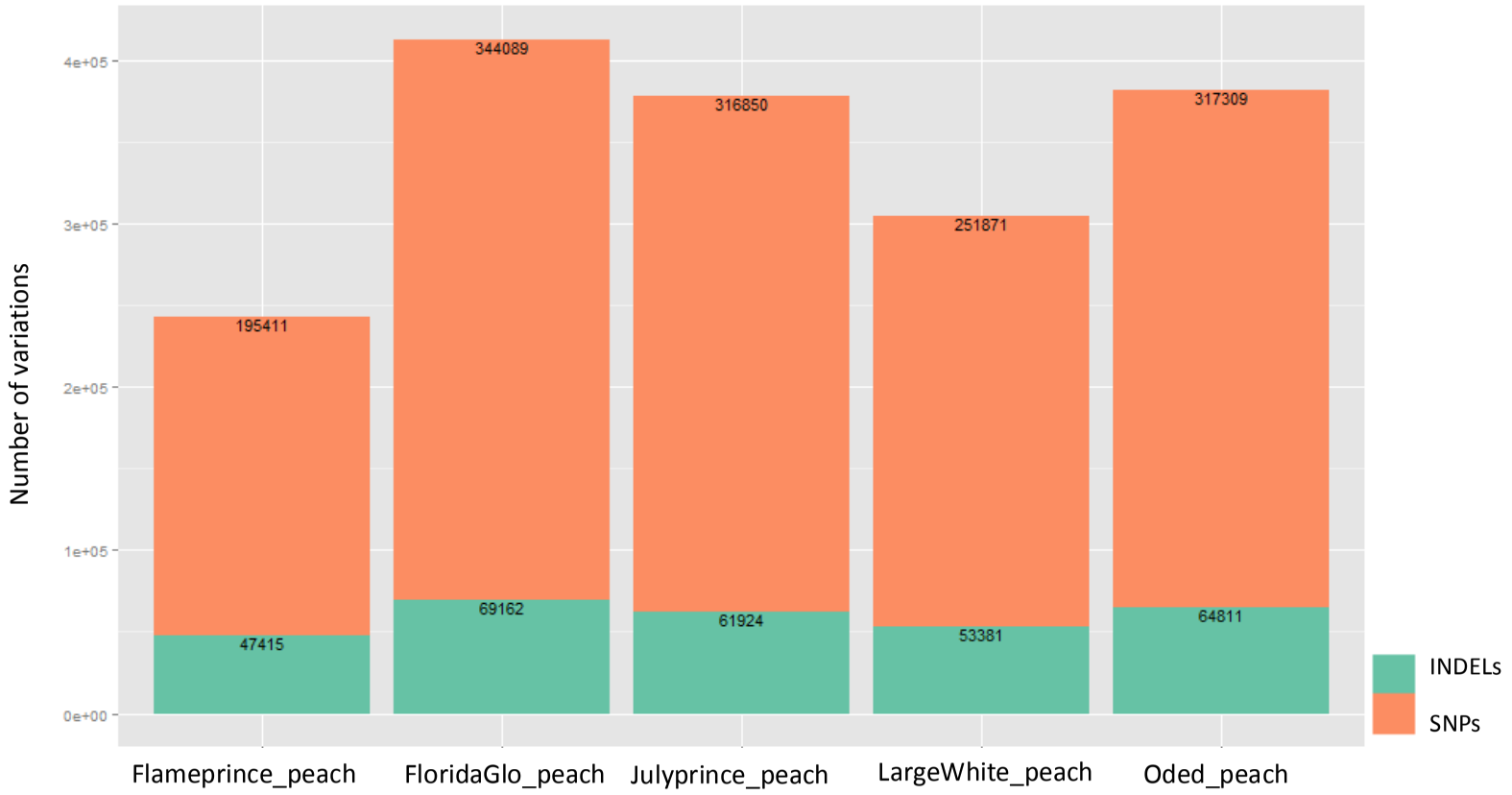


Figure CIII.4 General peach variability against the reference genome considering a depth equal or higher than ten reads per site, and a general genotype quality equal or higher than 20.

Table CIII.6. Total number of variants and their zygosity per sample considering a general depth equal or higher than ten reads per site, and a general genotype quality equal or higher than tweenty.

File	Total SNPs	SNPs		Total Deletions	Deletions		Total Insertions	Insertions		TOTAL		
		Homo	Het		Homo	Het		Homo	Het		π^1	Ho[2]
Flameprince_peach	**195,411**	74,948	120,463	**23,666**	9,401	14,265	**23,749**	9,744	14,005	**242,826**	0.0011	0.1846
Flameprince_Ham_necta1	**195,507**	75,789	119,718	**23,882**	9,467	14,415	**24,054**	9,803	14,251	**243,443**	0.0011	0.1842
Flameprince_Pearson_necta	**194,304**	74,342	119,962	**23,442**	9,382	14,060	**23,546**	9,733	13,813	**241,292**	0.0011	0.1835
Flameprince_Ham_necta_2	**198,073**	75,005	123,068	**23,608**	9,344	14,264	**23,593**	9,672	13,921	**245,274**	0.0011	0.1878
Julyprince_Pearson_peach	**316,850**	103,863	212,987	**30,928**	12,417	18,511	**30,996**	12,568	18,428	**378,774**	0.0017	0.3103
Julyprince_Pearson_necta	**329,701**	108,611	221,090	**31,391**	12,512	18,879	**31,518**	12,676	18,842	**392,610**	0.0017	0.3213
Oded_peach	**317,309**	83,151	234,158	**32,471**	11,199	21,272	**32,340**	11,125	21,215	**382,120**	0.0017	0.3434
Yuval_nectarine	**309,535**	81,522	228,013	**32,590**	11,193	21,397	**32,395**	11,080	21,315	**374,520**	0.0016	0.3361
Large_White_peach	**251,871**	61,082	190,789	**26,652**	8,000	18,652	**26,729**	8,028	18,701	**305,252**	0.0013	0.2832
Large_White_necta	**251,709**	60,873	190,836	**26,633**	7,986	18,647	**26,629**	7,984	18,645	**304,971**	0.0013	0.2832
FloridaGlo_peach	**344,089**	145,545	198,544	**34,478**	16,454	18,024	**34,684**	16,490	18,194	**413,251**	0.0018	0.2914
Gal-I_nectarine	**341,487**	144,458	197,029	**34609**	16421	18,188	**34,736**	16,511	18,225	**410,832**	0.0018	0.2898

1.Nucleotide diversity

2.Observed heterozygosity

Table CIII.7. Total number of variants and their zygosity per each sample considering a general depth equal or higher than ten reads per site, a general genotype quality equal or higher than 20 and applying to the genotypes the PL filter.

File	Total SNPs	SNPs		Total Deletions	Deletions		Total Insertions	Insertions		TOTAL		
		Homo	Het		Homo	Het		Homo	Het		π^1	Ho^2
Flameprince_peach	**57,224**	1,929	55,295	**8,924**	1,219	7,705	**8,877**	1,149	7,728	**75,025**	0.0003	0.1638
Flameprince_Ham_necta1	**71,466**	11,327	60,139	**11,577**	2,716	8,861	**11,690**	2,855	8,835	**94,733**	0.0004	0.1802
Flameprince_Pearson_necta	**63,712**	6,560	57,152	**8,823**	1,469	7,354	**8,786**	1,476	7,310	**81,321**	0.0004	0.1663
Flameprince_Ham_necta_2	**54,336**	1,923	52,413	**7,861**	992	6,869	**7,872**	996	6,876	**70,069**	0.0003	0.1532
Julyprince_Pearson_peach	**82,528**	335	82,193	**8,991**	698	8,293	**9,302**	648	8,654	**100,821**	0.0004	0.2295
Julyprince_Pearson_necta	**102,329**	3,005	99,324	**12,689**	2,072	10,617	**12,982**	2,050	10,932	**128,000**	0.0006	0.2798
Oded_peach	**120,842**	4,134	116,708	**15,131**	2,240	12,891	**15,213**	2,124	13,089	**151,186**	0.0007	0.3304
Yuval_nectarine	**119,893**	6,232	113,661	**15,652**	2,465	13,187	**15,690**	2,343	13,347	**151,235**	0.0007	0.3246
Large_White_peach	**105,490**	7,556	97,934	**13,792**	2,105	11,687	**26,975**	14,969	12,006	**146,257**	0.0006	0.2816
Large_White_necta	**98,505**	3,985	94,520	**12,933**	1,617	11,316	**13,293**	1,618	11,675	**124,731**	0.0005	0.2721
FloridaGlo_peach	**95,226**	4,871	90,355	**13,112**	2,955	10,157	**13,302**	2,833	10,469	**121,640**	0.0005	0.2569
Gal-I_nectarine	**94,041**	4,333	89,708	**13,242**	2,901	10,341	**13,473**	2,803	10,670	**120,756**	0.0005	0.2563

1.Nucleotide diversity

2.Observed heterozygosity

Nucleotidic diversity (π) and heterozygosity (Ho) was obtained by comparing each variety against the reference genome (**Table CIII.6**). Nucleotide diversity of peaches ranged from 1.10x10 $^{-3}$ to 1.8x10 $^{-3}$ (mean π = 1.4x10 $^{-3}$), which fits within the ranges of values reported in bibliography. For example (Aranzana *et al.,* 2012) obtained π ranging from 1.7x10^{-4} to 6.8x10^{-3} (average 2.7x10^{-3}) after sequencing 40 peach DNA fragments in 47 peach varieties. Similarly Verde *et al.,* (2013) obtained the average $\pi = 1.5\times10^{-3}$ at the whole-genome level.

The number of small variants in heterozygosis ranged from 34.3% (in 'Oded') to 18.4% (in 'Flameprince_Pearson_nectarine') with an average value of 26.7% (**Table CIII.6**). These percentages are concordant with the Ho values observed with SSR markers by Aranzana *et al.* (2010) and with SNPs by Micheletti *et al.,*(in preparation) and Aranzana *et al.,* (2012).

CIII.3.4 Somatic variability

To evaluate somatic variability we compared each of the 6 pairs of clones. The number of small variants detected ranged from 7,152 between 'Flameprince' and its nectarine mutant 'Flameprince_Pearson_nectarine' to 13,488 between 'Oded' and its mutant 'Yuval', with an average of 10,430. This represents 1 variant every 21.8 Kbp (or 45.9 variants per Mbp), which is close to 77 times less than the variation found between varieties in this work. Most of the somatic polymorphisms (80-84%) were due to (SNPs), with one every 17.8Kb in 'Flameprince' and 36.5Kb in 'Oded' (minimum and maximum, respectively). This ratio of variation is much higher than the observed between clones of other species like grape, where sport mutants occurred frequently. For example the study of sport mutants of the grape variety 'Pinot noir' have revealed 11.6 SNPs and 5.1 INDELs per Mbp (Carrier *et al.*, 2012) The variants detected fell in the 4 possible scenarios of polymosphisms: variant in the clone generating an heterozygous site (A_H); variant in homozygosis in the original variety and the alteranative allele in heterozygosis in the clone (B_H.) and a variant in heterozygosis in the original and in homozygosis in the clone (H_A and H_B). The number of polymorphisms (SNPs and INDELs) between pairs of clones is summarized in **Fig. CIII.5** and represented chromosome by chromosome in **Fig. CIII.6**. We did not find variants for the scenarios A_B and B_A.

The number and type of variations of polymorphisms per chromosome between the peach cultivar 'Flameprince' and its two nectarine sport mutants, which were generated from different mutation events in two different orchards, was very similar **(Fig. CIII.6)**. In both cases

chromosome 2 had the highest amount of variants (1,990 and 1,790 respectively) while chromosome 5 had the lowest (431 and 459 respectively). For the pairs of sports 'Oded'-'Yuval', 'Julyprince_peach'–'Julyprince_nectarine' and 'Large White_peach'-'Large-White nectarine', the greatest amount of polymorphisms was observed for chromosome 4, with a change rate of one every 9,333 bp, 10,006 bp and 9,024 bp respectively. Chromosome 5, wich contains the *G* locus, was the lowest variable in all but 2 sport pairs ('Oded'-'Yuval' and 'Florida Glo'-'Gal-I').

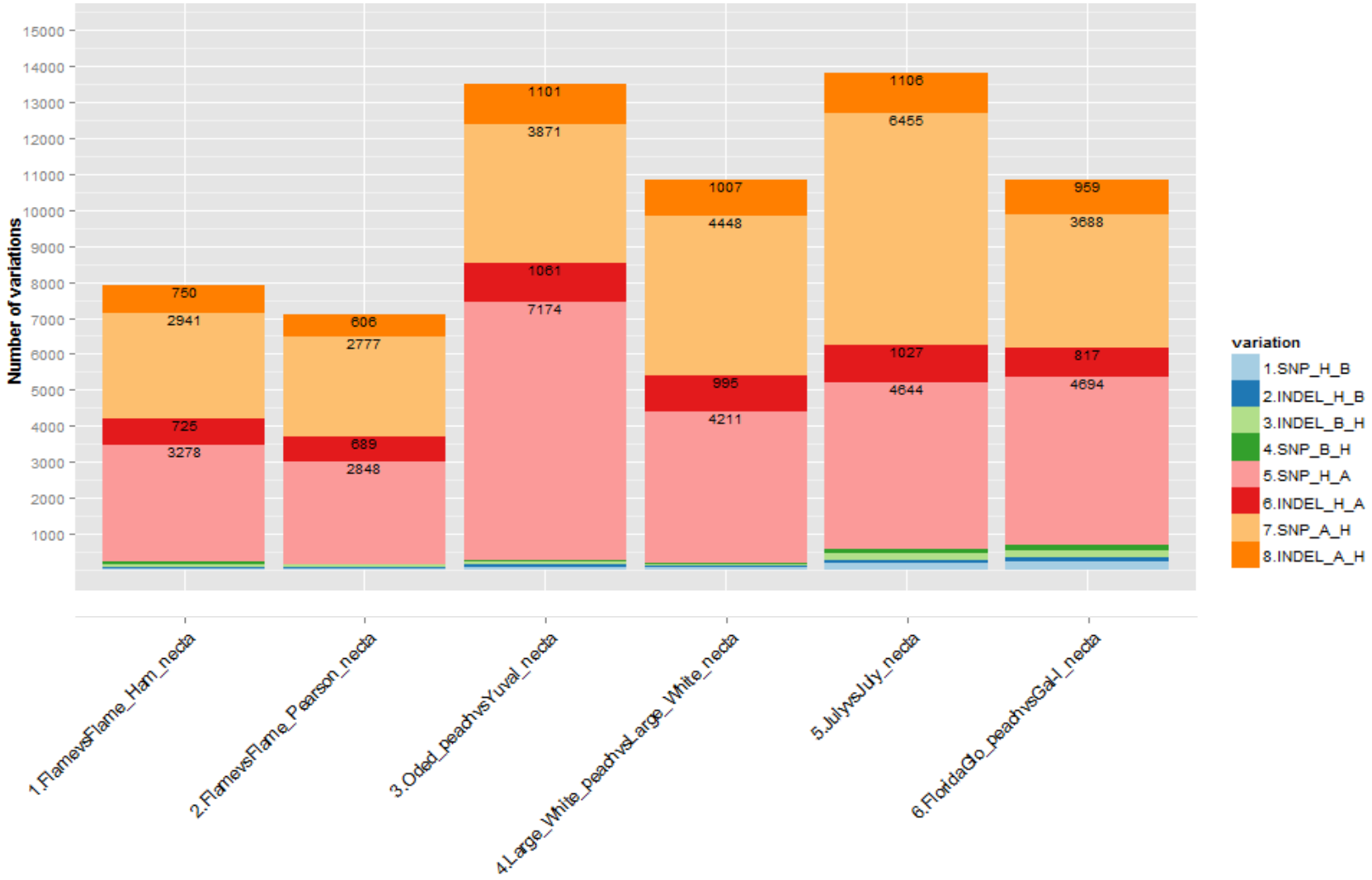


Figure CIII.5 Somatic variability split acrosst the different genotype scenarios. Flame: 'Flameprince'; July: 'Julyprince'; necta: nectarine.

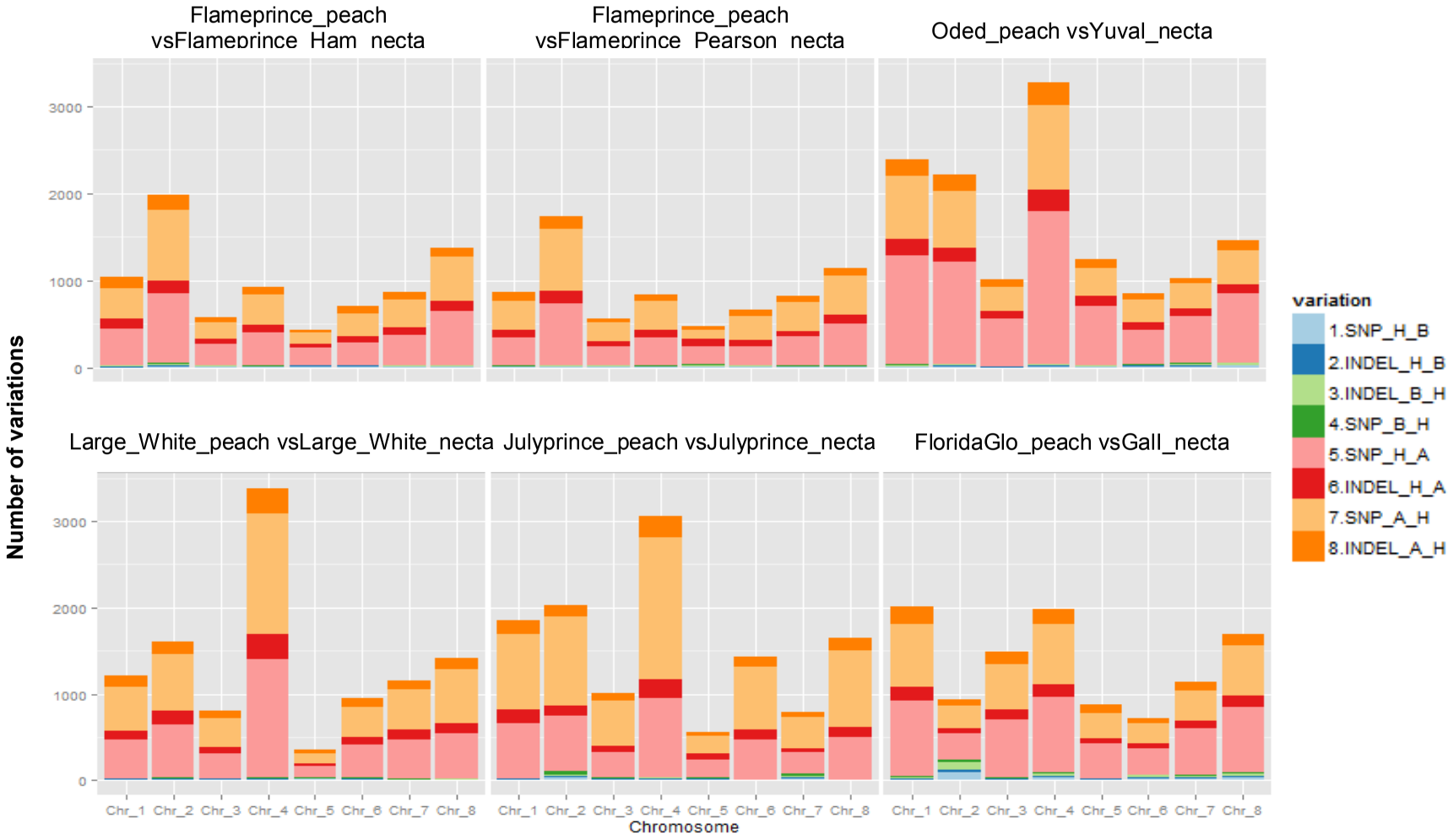


Figure CIII.6. Somatic variability per chromosomes at each pair of mutants considering a read depth equal or higher than ten and a genotype quality equal or higher than 20.

The most frequent variation between pairs of clones found genome-wide and also in each chromosome was a new SNP in heterozygosis in the clone (scenario A_H) which is consistent with the infinite site mutation model of SNPs (Kimura, 1969) wich assumes that multiple mutations never occur at the same sequence position. However the proportion of SNPs following the rest of scenarios was high (53.3%). These escenarions always involved a new mutation event occurring at a site mutated previously. Such differences between pairs of clones could be spurious caused by low coverage or due to a very broad-range variant calling, producing an excess of heterozygous sites. To confirm this hypothesis we randomly selected and visualized 20 SNP sites of each scenario for each pair. In an overall percentage of 94.5% of the cases the variants had low genotype quality even though they had a read depth of at least 10 and a general quality of at least 20.

The large levels of somatic and intraspecific variability detected suggest that the pipeline used called a large amount of false SNPs and INDELs, detecting especially large amount of heterozygous. After analyzing visually the quality parameters of a subset of SNPs we applied a new filter to the variants to reduce false positives. With this new filter we selected only those variants for which the Phred-Likelihood of the most probable genotype was lower than 10 , the Phred-Likelihood of each alternative genotypes higher than 50 and, at the same time, the most likely alternative genotype two times lower than the less likely alternative genotype (for more details see “materials and methods” section).

Using this new filter the number of variants identifyied genome-wide were 431,926 meaning that the reduction of variants was almost to half (53.62%). The average reduction was slightly higher for SNPs (67%) while insertions were reduced in 54% and deletions in 58%**.** This information is reported in **Table CIII.7**, which also shows the number variants found in each library. The peach varieties that experienced higher reduction in the number of variants were ‘Julyprince’, ‘Florida’ and ‘Flameprince’, which lost 70% of their variants. These varieties were followed by ‘Oded’ which lost 60% of variants and ‘Large White’ the 52% (**Fig. CIII.7**). It is remarkable that the general variability was reduced in 53.62% while the individual reduction per sample was around 65%, which suggests that the new filter removed those variants observed only in one variety,(variants not well supported likely to be false positives.).

After the new filter, π was reduced from 1.50×10^{-3} to 4.91×10^{-4}, which still fits within the range described in bibliographiy. The average number of small variants in heterozygosis decreased from 27.40% to 24.12%, ranging from 33% to 16% (**Table CIII.7**). These values are in concordance with previous ones observed for occidental varieties (Aranzana *et al.,* 2012;

Verde *et al.,* 2012; Micheletti *et al.,* in preparation), and much higher than the 1.55% obtained in the analysis of 84 Chinese accessions (wild, ornamental, landrace and cultivated peach varieties) using small variants markers (Cao *et al.,* 2014).

In general, the somatic variability was highly reduced (~99% of reduction) under the new filter conditions (**Fig.CIII.8**). The INDELs H_A and A_H were reduced in 97% while SNPs of both scenarios were reduced in 99%. All 4 possible scenarios of polymosphisms where also observed after the more stringent filtering conditions, being the most frequent those corresponding to A_H and H_A types. Although the H_A INDELs (i. e. the original genome presents an INDEL in heterozygosis lost in its clone and resembling the reference) can be easily explained by rearreangments, H_A SNPs are difficult to understand, indicating that we may have still allowed for false polymorphisms, which will explain also the very unlikely scenarios H_B and B_H (which implicate two mutational events at the same site) and removed true variability, obtaining variability levels concordant with those reported in bibliography.

The physical position of somatic variants and their genomic effect are summarized in **Table CIII.8**. Most variants were located in repeatitive regions of the genome. Repetitive regions tend to confound the alignments since the reads from regions with repetitive bases have a much higher probability of being aligned onto multiple locations (Yu *et al.,* 2012), thereafter some fragments could be wrongly aligned producing false positives. In conclusion, some of the variants detected as false (those involving two mutation events at the same site) could be due to the alignment rather to an erroneous SNP calling.

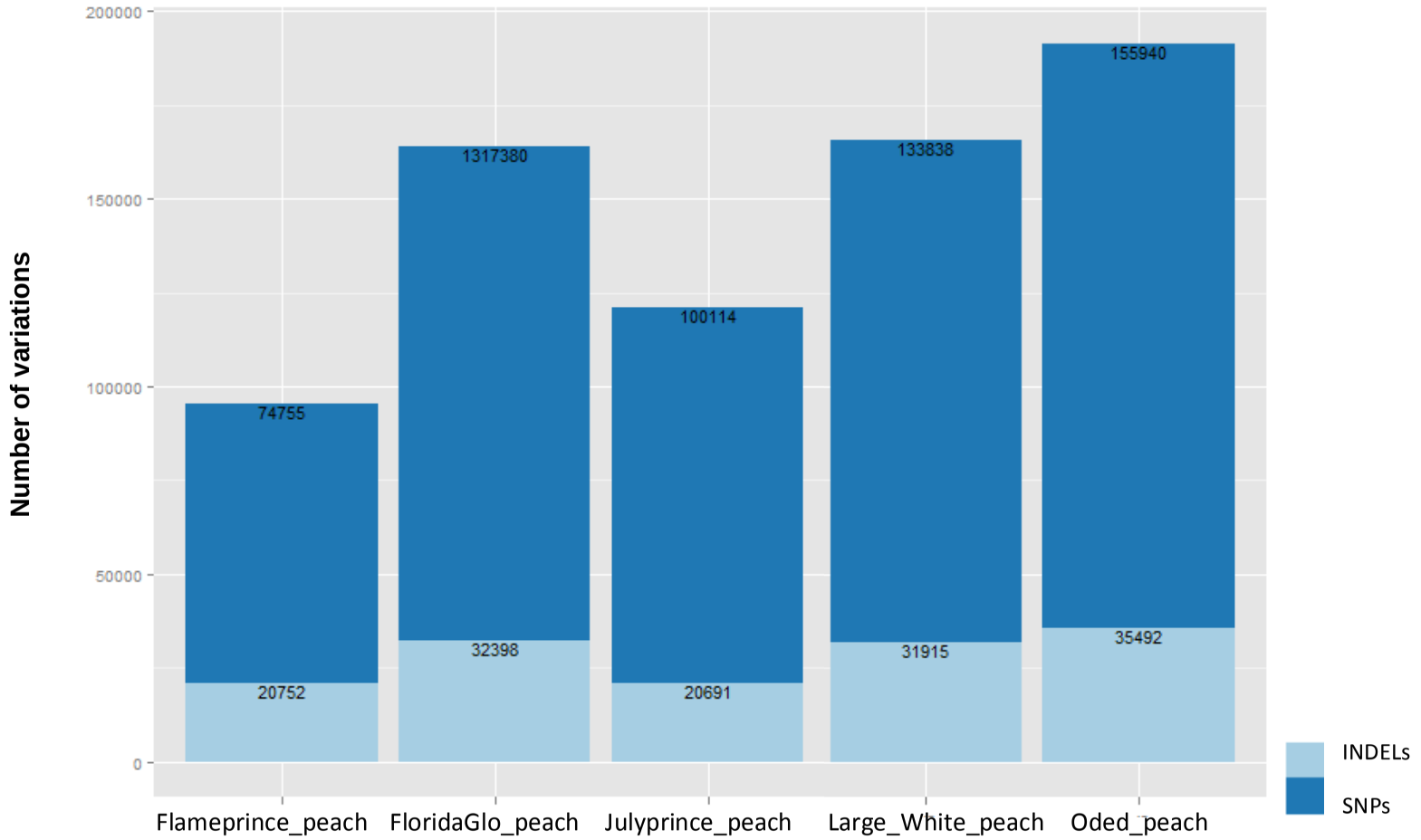


Figure CIII.7 Intraspecific variability of peaches after applying the Phred-Likelihood (PL) filter.

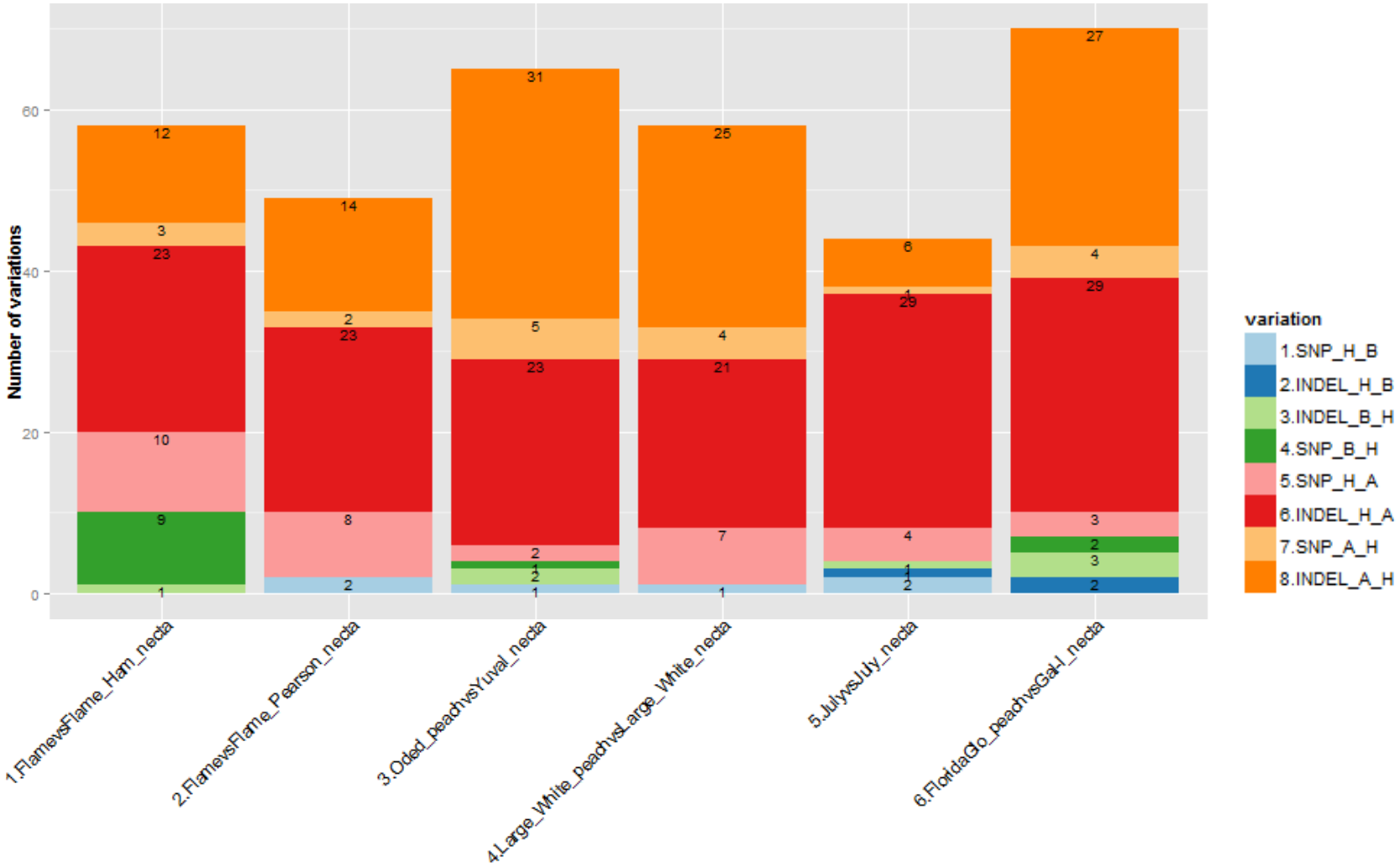


Figure CIII.8 Somatic variability after applying the Phred-Likelihood (PL) filter.

Table CIII.8 Physical location, genomic effect and available annotation of somatic small variants sorted by each observed genotype scenario. Continued

Scenario	Pair	Change	variant	Inter	Gene	Location	Effect	Annotation
A-H	FlamevsFlame_Ham_necta	T-G	SNP		ppa026879m	LG1:13,709,236		Protein kinase, serine threonine, catalytic domain
A-H	FlamevsFlame_Ham_necta	A(T)3-A(T)4	INDEL			LG4:24,189,522		Repeat_90138
A-H	FlamevsFlame_Ham_necta	A-C	SNP			LG5:7,285,082		Repeat_99421
A-H	FlamevsFlame_Ham_necta	C-T	SNP		ppa026862m.g	LG6:23,630,613	(cGc/cAc)R631H)	Aminotransferase_like plant mobile domain
A-H	FlamevsFlame_Ham_necta	TC-T				LG7:4,892,591		Repeat_90138
A-H	FlamevsFlame_Pearson_necta	A(T)6-A(T)7	INDEL		ppa006198m.g	LG1:226,939	Upstream_1866	Glycosyl transferase, family4
A-H	FlamevsFlame_Pearson_necta	C(T)7-C(T)10	INDEL	x		LG1:25,769,176		
A-H	FloridaGlovsGal-I	C(A)6-C(A)7	INDEL			LG1:25,356,017		Repeat_18747
A-H	FloridaGlovsGal-I	G(A)5-GAGC(A)5	INDEL			LG2:517,2219		Repeat_41840
A-H	FloridaGlovsGal-I	G(A)7-g(A)6	INDEL			LG3:6,292,395		Repeat_61520
A-H	FloridaGlovsGal-I	GAA-GA	INDEL		ppa007365m.g	LG4:12,073,728		mRNA capting enzyme subunit metyltransferase
A-H	FloridaGlovsGal-I	C(T)2-C(T)4	INDEL	x		LG4:12,343,935		
A-H	FloridaGlovsGal-I	C-T	SNP			LG4:20,917,575		Repeat_8792
A-H	FloridaGlovsGal-I	A(T)7-A(T)6	INDEL	x		LG5:2,087,972		
A-H	FloridaGlovsGal-I	C-T	SNP	x		LG6:10,322,895		
A-H	FloridaGlovsGal-I	C-G	SNP			LG6:6,486,623		Repeat_111234
A-H	FloridaGlovsGal-I	C(A)7-C(A)9	INDEL			LG6:8,946,712		Repeat_113160
A-H	FloridaGlovsGal-I	G(A)4-G(A)3	INDEL			LG7:6778176		Repeat_132256
A-H	FloridaGlovsGal-I	AAT-A	INDEL		ppa023417m.g	LG8:5,656,979	Upstream_1414	Leucine_rich_repeat_N-terminal_type2

Table CIII.8 Physical location, genomic effect and available annotation of somatic small variants sorted by each observed genotype scenario. Continued.

Scenario	Pair	Change	variant	Inter	Gene	Location	Effect	Annotation
A-H	Large_WhitevsLarge_White_necta	G-A	SNP			LG1:18,994,683		Repeat_14426
B-H	FlamevsFlame_Ham_necta	A-G	SNP			LG1:13,803,534		Repeat_10720
B-H	FlamevsFlame_Ham_necta	T-C	SNP			LG2:13,362,267		Repeat_47555
B-H	FlamevsFlame_Ham_necta	TTATAT-(TTATAT)4	INDEL			LG5:6,629,969		Repeat_99250
B-H	FlamevsFlame_Ham_necta	A-G	SNP			LG7:4,761,964		Between repeat_130963-130964
B-H	FlamevsFlame_Ham_necta	A-C	SNP			LG8:4,768,993		Repeact_147220
B-H	FlamevsFlame_Ham_necta	T-C	SNP			LG8:5,352,318		Repeat_147771
B-H	FlamevsFlame_Ham_necta	A-C	SNP			LG8:5,352,365		Repeat_147771
B-H	JulyvsJuly_necta	C(T)7-C(T)8	INDEL	x		LG2:1,060,493		Close to Repeat_38908
H-A	FlamevsFlame_Ham_necta	T(A)6 -T(A)7	INDEL		ppb022959m.g	LG1:21,628,214	Upstream_675	Unknwon
H-A	FlamevsFlame_Ham_necta	AAG-	INDEL		AT1G11340.1	LG2:10,150,289		Repeat_ /S-locus lectin protein kinase
H-A	FlamevsFlame_Ham_necta	A(T)8-A(T)10	INDEL	x		LG3:14,485,863		
H-A	FlamevsFlame_Ham_necta	TA-T	INDEL			LG7:13,665,996		Repeat_137689
H-A	FlamevsFlame_Ham_necta	C-T	SNP	x		LG8:14,478,292		
H-A	FlamevsFlame_Pearson_necta	TG-T	INDEL		ppa020282m	LG7:8,837,962		Reverse transcriptase/Zing finger/
H-A	FlamevsFlame_Pearson_necta	AC-C	INDEL			LG8: 20,498,085		
H-A	FloridaGlovsGal-I	T-TAG	INDEL		ppb024849m.g	LG1:26,807,604	Downstream_815	unknown
H-A	FloridaGlovsGal-I	CA C	INDEL		ppa011741m.g	LG3:2,384,176	Frame_shift_158	Ankyrin repeat domain_prot-prot_interaction
H-A	FloridaGlovsGal-I	A-T	SNP		ppa006140m	LG3:4,578,760	INTRON	Zing finger_RING_type
H-A	FloridaGlovsGal-I	C(T)8-C(T)10	INDEL			LG4:28,182,544		Repeat_20794
H-A	FloridaGlovsGal-I	A- A(T)2	INDEL		ppa022863m.g	LG5:1,339,618	Upstream_1024	unknown

Table CIII.8 Physical location, genomic effect and available annotation of somatic small variants sorted by each observed genotype scenario.

Scenario	Pair	Change	variant	Inter	Gene	Location	Effect	Annotation
H-A	FloridaGlovsGal-I	T(A)6-T(A)7	INDEL		ppa0011989m	LG6:17,134,074	Upstream_21	Lipoxygenase:LH2
H-A	FloridaGlovsGal-I	C-CGAACA	INDEL		ppa007036m	LG6:23,699,753		Fructose-1,6-bisphosphatase class 1
H-A	FloridaGlovsGal-I	C(A)4-C(A)3	INDEL			LG6:3,576,422		Repeat_26002
H-A	FloridaGlovsGal-I	T(A)2-TA	INDEL	x		LG7:2,406,835		
H-A	FloridaGlovsGal-I	A-ATATATAT	INDEL			LG8:4,332,282		Repeat_147771
H-A	JulyvsJuly_necta	A(T)7-A(T)6	INDEL			LG1:23,475,481		Repeat_17384
H-A	JulyvsJuly_necta	ACCC-ACC	INDEL		ppa026846m	LG2:780,240	Frame_shift_153	Disease resistance prot(NB-ARC)
H-A	JulyvsJuly_necta	CCAC-CC	INDEL		ppa026938m.g	LG4:20,948,557	Upstream_279	Leucine-rich repeat
H-A	JulyvsJuly_necta	C(A)6-C(A)7	INDEL		ppa014197m.g	LG4:6,506,035	Upstream_1423	Unknown
H-A	JulyvsJuly_necta	C(A)8-C(A)10	INDEL			LG6:10,635,224		Repeat_114473
H-A	JulyvsJuly_necta	G(A)5-G(A)6	INDEL		ppa023987m.g	LG7:16,068,946	Upstream_1151	methyl esterase 3
H-A	Large_WhitevsLarge_White_necta	T(A)6 T(A)7	INDEL	x		LG2:1,285,766		
H-A	Large_WhitevsLarge_White_necta	T TG	INDEL		ppa022423m.g	LG2:1,630,369	Upstream_1320	Unknown
H-A	Large_WhitevsLarge_White_necta	C(A)8 C(A)9	INDEL	x		LG2:3,078,252		
H-A	Large_WhitevsLarge_White_necta	G(T)6-G(T)7	INDEL			LG4:12,688,209		Repeat_81697
H-A	Large_WhitevsLarge_White_necta	C(A)8-C(A)10	INDEL		ppb013565m.g	LG6:10,635,224	Downstream_149	Unknown
H-A	Large_WhitevsLarge_White_necta	T-TATA	INDEL	x		LG7:7,863,050		
H-A	Large_WhitevsLarge_White_necta	G-GC	INDEL		ppa010660m.g	LG8:1,882,481	Upstream_1986	Homologous:pathogenesis-related family
H-A	Large_WhitevsLarge_White_necta	G-A	SNP		ppa009536m	LG8:14,284,875	INTRON	TRAM/LAG7CLN8-Acetyl-CoA synthesis
H-B	Large_WhitevsLarge_White_necta	TGC-T	INDEL	x		LG4:28,318,174		

As a summary, in the previous sections we have described the variability found first with a broadly used pipeline for calling single nucleotide variants and small insertions and deletions (Jia *et al.*, 2012) and later applying a filter based on the Phred-likelihood (PL) values of the genotypes. The rationale for applying the filter was the detection of an excess of false variants. After visualizing a number of alignments containing some putatively false variants and the quality values of such variants, we removed those where the Phred-likelihood of the most frequent genotype was only slightly higher than the Phred-likelihood of the alternative genotypes and, consequently, leaving only those for which the genotype calling was evident. However this filtering may have eliminated true variants, as discussed above. For this reason in the following sections of this chapter we will work with the variants detected before the PL filter. As we should assume that the errors in variants are equally distributed genome-wide and affect equally all genomic regions, we expect that they won't disrupt the comparison of variability occurring genome-wide with the one in the region containing the *G* locus. Moreover, the study of the genomic effect of the variants and the search of a possible causal allele will be more wrongly altered with the elimination of true variants than with the inclusion of false positives.

CIII.3.5 Analysis of *G* locus region

At the starting of this work, the peach/nectarine locus (*G*) was mapped at the end of LG5 of the *Prunus* referencce map (TxE) between the SSRs CPSCT030 (scaffold_5:15,126,681..15,127,320) and CPSCT022 (scaffold_5:16,626,112..16,626,607) which corresponds to a genetic distance of 16 cM (or 1.50 Mbp in physical distance). Other works in our lab placed the locus in a slightly bigger window of 2 Mbp between the SSR markers BPPCT038 (scaffold_5:14,658,198..14,658,198) and CPSCT022. We considered this last wider window to analyze more deeply somatic variability and scrutinize for possible causal alleles for the nectarine phenotype.

Among the 805,506 variants detected genome-wide, 1,224 occurred in this 2Mbp region (**Table CIII.9**) which represents 1 variant every 1,633bp (one SNP every 2,301bp and one INDEL every 5,633bp). These values are lower than the ones observed for the whole genome (one variant every 281bp). Nucleotide diversity in this region was, consequently, smaller ($3.1x10^{-4}$ in front of $1.5x10^{-3}$). Contrary, the mean heterozygosity was higher (43% in front 26.7%). The peaches studied here mutated to nectarine, which is a recesive trait. Thereafter we can assume that these peaches carry the nectarine allele in heterozygosis and is the peach

wild allele the one that changed to produce the glabourness fruits. This explains the high heterozygosity of this region. Recent studies have reported a unique necarine allele as the causant of the trait in all modern varieties (Vendramin *et al.,* 2014). The short age of the allele and strong selection towards it in breeding programs may have generated a large and low variable region (as detected here) flanking the causal allele share by all nectarines and identical by descent.

CIII.3.5.1 Genomic effect of the variants

With the sofware SnpEff 3.4 (Cingolani *et al.,* 2012) we quantified and predicted the genomic effects of the variants found on the *G* locus and compared them with those occurring in the whole genome. Most of the calculated effects (1,746,611) produced by the variants (805,506) occurred in intergenic regions, which represents the 37.36% of the total changes (**Table III.12)**, while in the *G* locus most of variants occurred 5Kbp up and downstream genes (36.17% and 30,84%, respectively). Curiously no genomic effects were observed in exons of the *G* region, while those represented 3,25% of the effects in the whole genome. Thus, most of the variations occurred within non coding regions, as expected due to the DNA prevention function against disruptive changes (Goode *et al.,* 2010) and because it is calculated that just ~1.5% of the genome of species is composed by coding regions (Thomas & Touchman, 2003). In peach, it has been stimated that there are 1.22 genes every 10Kb. In total 27,852 protein-coding genes and 28,689 protein-coding transcripts were predicted (Verde *et al.*, 2013). In consequence, the majority of variants will not be transcribed into proteins.

The genomic effects of the variants had an impact classified by SNpeff as a modifier; only 0.2% of them had a high impact. This proportion was slightly higher in the *G* locus region (0,35%) (**Table CIII.10**) with variants producing a high impact on the structure of the protein altering either the ORF or the amino acid transcript sequence. These high impact variants produced: frame shifts (1,911 genome-wide and 8 in the G region), stop gained codons (847/1), stop lost codons (100/0), splice site acceptors (210/1), splice site donors (286/2) and start lost codons (84/1) (**Table CIII.12)**.

The effects can be also classified by their function as missense (non-synonymous), nonsense (stop codon gained) and silent (synonymous). The percentages for missense and silent changes were 56.91% and 41.59% respectively for the whole genome and 52.83% and 46.22% for the *G* locus. The nonsense changes were lower than 2% in both cases (**Table CIII.11)**. The observed values of missense and silent variantions are relevant. Missense changes produce a molecule chemically different that may disturb the structure or function of the

protein. On the other hand, although silent mutations do not change the amino acid sequence and generally are considered selectively neutral (Gorlov *et al.,* 2006), in some cases they can also change the structure or function of the protein throught the mRNA splicing or transport disturbance (Johnson *et al.,* 2011; Polony *et al.,* 2003; Shabalina *et al.,* 2013). The observed ratio between non-synonymous and synonymous changes (N/S) was higher than 1 in both, the whole genome and in the *G* region, indicating an excess of synonymous changes over the non-synonymous. This ratio was slightly lower in the *G* locus region.

All these data indicate that, despite the *G* locus showed lower nucleotide diversity, the type and effect of the polymorphisms in this region is simmilar to the ones observed genome-wide.

.

Table CIII.9 Total number of variants, type and zygosity for each sample file across the region: scaffold_5: 14650000..16650000 obtained from the multiple-sample calling performed by SAMtools mpileup (Li *et al.,* 2009) .In bold letters are shown the total amount (SNPs + INDELs), the total amount of SNPs (Homo and Het), the total amount of INDELs (Insertions + Deletions).

File	SNPs		INDELs				Change rate	π^1	Ho^2
			Deletions		Insertions				
	Homo	Het	Homo	Het	Homo	Het			
Flameprince_peach	71	405	26	81	27	93	2,885	0.000352	0.473039
Flameprince_Ham_necta	71	390	26	79	25	96	2,911	0.000344	0.461601
Flameprince Pearson_necta	69	409	26	82	25	94	2,836	0.000353	0.477941
Julyprince_peach	69	433	29	79	28	96	2,724	0.000367	0.496732
Julyprince Pearson_necta	69	417	27	79	25	96	2,805	0.000357	0.48366
Oded_peach	8	418	15	86	12	86	3,200	0.000313	0.482026
Yuval_necta	8	394	17	86	12	85	3,322	0.000301	0.461601
Large White_peach	67	416	24	78	26	72	2,928	0.000342	0.462418
Large White_necta	67	418	24	78	24	68	2,946	0.00034	0.460784
Florida Glo_peach	5	179	12	47	12	45	6,666	0.00015	0.221405
Gal-I_necta	5	181	13	45	11	42	6,734	0.000149	0.218954
TOTAL	**869**		**355**		**1224**				
Mean								**0.000313**	**0.434955**

1 Nucleotide diversity

2 Heterozygosity

Table CIII.10. Impact of changes evaluated from the peach annotation reference genome. The genomic effect classification into four impact categories described in **Appendix CIII.6**.

	Whole genome		***G* locus**	
Type	**Count**	**Percent (%)**	**Count**	**Percent (%)**
High	3,438	0.197	13	0.348
Low	25,830	1.479	54	1.444
Moderate	31,445	1.8	65	1.738
Modifier	1,685,898	96.524	3,607	96.47

Table CIII.11. Effects of variations per functional class.

	Whole genome		**G locus**	
Type	**Count**	**Percent (%)**	**Count**	**Percent (%)**
Missense	30,593	5.919	56	52.83
Nonsense	799	1.487	1	0.943
Silent	22,356	41,.594	49	46.226
Missense/silent ratio	1.3684		1.1429	

Table CIII.12. Effect per genomic region produced by the small variants. Some minor effects are not shown.

Region	Note	Whole Genome		*G* locus	
		Count	Percent (%)	Count	Percent (%)
Downstream	Downstream of a gene (5Kb)	443,211	25.38	1,150	30.84
Exon		56,749	3.25	_	_
Intergenic	Not transcript, between a gene	652,558	37.36	880	23.54
Intron	No exon in the transcript	93,577	5.39	198	5.29
Splice site aceptor	Two bases before exon starts. except for the first exon	210	0.01	1	0.03
Splice site donor	Two bases after coding exon and expect for the last exon	286	0.02	2	0.05
Upstream	Upstream of a gene (5Kb)	488,250	27.95	1,351	36.13
UTR-3'		4,759	0.27	14	0.37
UTR-5'		3,188	0.18	11	0.29
Splice site region	Either within 1-3 bases of the exon or 3-8 bases of the intron	3,130	0.18	5	0.13
Intragenic	No transcript within the gene	355	0.02	_	_
Start lost		84	0.01	1	0.03
Stop gained		847	0.05	1	0.03
Stop lost		100	0.01	-	-
Synonymous coding		22,321	1.28	49	1.31
Codon insertion		293	0.02	6	0.16
Frame shift		1911	0.11	8	0.22

CIII.3.5.2 Search of small variants responsible for the nectarine trait

Glabrourness is a monogenic recessive trait and, as the peaches studied here mutated to nectarine, we can assume that they have both the peach and the nectarine allele. This hypothesis is supported, as explained above, by the higher heterozygosis observed in the region. To investigate the causal mutation we worked with two hypotheses, the first considers that the peach allele (G) of the peaches studied here mutated to nectarine through the same mutation mechanism as occurred anciently, i.e. the new nectarine allele is the same as the one fixed in the cultivated varieties. This means that there will be just a single possible nectarine allele (g1) and, in consequence, the nectarine sport mutants will have it in homozygosis (g1/g1). The second hypothesis considers that all (or most of) the nectarine sport mutants analyzed here were generated through a mutation in the G allele different from the existing one. Consequenlty, in our sample there will be two different alleles g (g1 and g2), peaches will be G/g1 and their sports g1/g2. Assuming that SNPs arise throught the infinite site model, the occurrence of two independent mutations in the same site of the gene is unlikely, and thereafter the two hypotheses should apply only for INDELs. Other plausible hypotheses like i) a mutation in a gene other than the one with the mutation fixed, or ii) each of the sport nectarines have a different mutation, were discarded. Sport mutants in peach are frequently observed although only few have been studied genetically (López-Girona in preparation; Falchi *et al.,* 2013). In those cases the new mutation occurred in the same gene as the one fixed in the cultivated varieties. Multiple independent mutation events in the same gene, generating different alleles, have been also reported in peach (Brandi *et al.,* 2011). Despite this evidence we have discarded here the hypothesis of different new alleles. For some of the studied peach varieties here ('Flameprince', 'Julyprince' and 'Large White') it has been observed a large tendency to mutate to nectarine through different mutation events (for example we have included in this study two sport mutants of 'Flamprince' produced in two different orchards). Moreover, seems to be hereditary, 'Large White' offsprings produce also nectarine sports (personal communication from the peach breeder M. Ortiz). All peach varieties studied here come from Florida orchards and, although their pedigree is not available, they could share the same *G* allele with one site 'prone' to mutate, and thereafter the mutations may involved always the same sites. However this is just a hypothesis that simplifies the analysis but other mechanisms can be behind the high genetic predisposition of some varieties to mutate.

According to the two working hypothesis, in the region of chromosome 5 we selected the variants heterozygous for peach and homozygous alternative to the genome reference for the nectarine (H_B; hypothesis 1); and heterozygous variants for each peach and nectarine pair

(H_H) in the pair and homozygous as the reference for the peach but heterozygous for the nectarine (A_H) less than 1Kb apart and variants A_H, i.e. homozygous as the reference for the peach but heterozygous for the nectarine (hypothesis 2).

In total, 22 small variants in the *G* locus region were identified between mutants (**Table CIII.13**) compatible with the first hypothesis (i.e. with genotype H_B). Sixteen of them were INDELs and surprisingly 6 SNPs. Two of the variants (INDELs) occurred in intergenic regions while the 20 remaining affected a total of 12 genes, two of them in intronic regions. Most of the variants were located either upstream or downstream of the affected genes, which produce a very low effect. Four of the six sport pairs identified the gene ppa004540, a peptatricoopeptide repeat, which contains a domain repetated at least 5 times and thereafter difficult to aling. One of the SNPs between 'Flameprince' and its mutant 'Flameprince Pearson' occurred in the gene ppb020487, which codify for a protein directly involved in the cell wall development. Surprisingly most of the variants occurred between 'Flameprine' and its two mutants, and non of them occurred a in the other clones, rejecting the working hypothesis.

Althought when this thesis started the *G* gene was unknown, in 2014 Vendramin *et al.,* (2014) reported apolymomrphisms in a MYB25 gene strongly associated with the nectarine trait pointing this gene (ppa023143m.; scaffold_5: 15897836..15899002) as a strong candidate for the trait. The polymorphism reported consists in a close to 7Kb insertion of a trasposable element in the 2nd exon, producing the nectarine phenotype. The tool used here to call variants detected only short polymorphisms, and therafter this big deletion was not discovered; the closest small variant to the gene was an INDEL located at 119Kb apart from this gene, and it was just identified in two pairs of samples.

Under the second hypothesis we assume the presence of a recessive allele (g1) in heterozygosis in both peaches and nectarines, and a second allele (g2) in heterozygosis only in the mutants and at a relatively close distance from polymorphism causing g1. Therafter we looked for genotypes A_H close to H_H. In total, we identified 163 small variants, among them 30 were located in intergenic regions (data not shown). The rest (133) modified 60 genes (**Table CIII.14**). All pairs of mutants showed variation in the pentatricopeptide gene ppa004540m named earlier when describing variants compatible with hypothesis one. This can be explained by an excess of variants due to missaligments. As for the previously, here we didn't detect hypothesis two polymorphisms in the candidate gene MYB25.

Interestingly, the gene ppa010308m.g was affected by small variants identified in all of the sample pairs studied. These variants are on the promoter region of this gene, which is a MADS box gene. These genes are of ancient origin and are found in animals, fungi, and plants. All identified MADS box genes encode a highly conserved N-terminal DNA binding domain 55 to 60 amino acids in length named the MADS domain, which originated from the DNA binding subunit A of topoisomerases II subunit A (Gramzow *et al.,* 2010). Plant MADS box genes were first identified as regulators of floral organ identity and have been reported to control additional developmental processes such as: the determination of meristem identity of vegetative inflorescence, and floral meristems, root growth, ovule and female gametophyte development, flowering time, development of vascular tissue and seed and fruit formation, growth, ripening, and dehiscence (Buchner & Boutin, 1998; Colombo *et al.,* 2008; Giovannoni, 2004; Liljegren *et al.,* 2000; Ng & Yanofsky. 2001; Zhang *et al.,* 2010; Whipple *et al.,* 2004; Zhang, 1998).

Table CIII.13. Somatic small variants with an heterozygous genotype for peach and homozygous genotype for nectarine (hypothesis 1) and genomic regions where they occurred.

Pair	Variant	Position	REF	ALT	DP	MQ	Type	Gene or Repeat	Annotation
FlamevsFlame_Ham_necta	INDEL	16,561,465	GT		180	44	UP_3287	ppa004540m.g	Pentatricopeptide repeat
FlamevsFlame_Pearson_necta	INDEL	16,561,465	GT		180	44	UP_3287	ppa004540m.g	Pentatricopeptide repeat
JulyvsJuly_necta	INDEL	16,561,688	G		361	39	UP_3061	ppa004540m.g	Pentatricopeptide repeat
OdedvsYuval_necta	INDEL	16,563,860	A		402	56	UP_882	ppa004540m.g	Pentatricopeptide repeat
FlamevsFlame_Ham_necta	SNP	16,634,399	A	G	199	30	UP_260	ppa010308m.g	TF, MADS-box
FlamevsFlame_Pearson_necta	SNP	16,634,399	A	G	199	30	UP_260	ppa010308m.g	TF, MADS-box
FlamevsFlame_Pearson_necta	SNP	16,634,314	C	T	263	30	UP_345	ppa010308m.g	TF, MADS-box
FlamevsFlame_Ham_necta	INDEL	15,436,844	$(A)_4$		332	50	UP_329	ppa018776m.g	Porin, eukaryotic type
FlamevsFlame_Pearson_necta	INDEL	15,436,844	$(A)_4$		332	50	UP_329	ppa018776m.g	Porin, eukaryotic type
FlamevsFlame_Ham_necta	INDEL	15,778,965		AT	353	59	UP_1708	ppa024260m.g	ZF-HD homeobox protein, Cys/His-rich dimerisation; floral development
FlamevsFlame_Pearson_necta	INDEL	15,778,965		AT	353	59	UP_1708	ppa024260m.g	ZF-HD homeobox protein, Cys/His-rich dimerisation; floral development
FlamevsFlame_Ham_necta	INDEL	14,775,755	AT		404	58	DOWN_3956	ppb016228m.g	Domain of unknown function DUF2828
FlamevsFlame_Pearson_necta	INDEL	14,775,755	AT		404	58	DOWN_3956	ppb016228m.g	Domain of unknown function DUF2828
FloridaGlovsGall	INDEL	15,431,857	A		424	59	UP_46	ppa009625m.g	Stomatin, Band7 protein
FlamevsFlame_Pearson_necta	SNP	14,719,520	C	T	354	58	DOWN_3866	ppb020487m.g	Alpha-amylase, Glycosyl hydrolase, family 13, all-beta
Large_WhitevsLarge_White_necta	INDEL	15,393,249	$(T)_6$		303	56	UP_2580	ppa024635m.g	Oligopeptide transporter
OdedvsYuval_necta	SNP	14,700,174	A	G	300	60	INTRON	ppa010625m.g	lipid metabolic process,hydrolase activity
Large_WhitevsLarge_White_necta	INDEL	15,597,046	G		159	58	INTRON	ppa012417m.g	TB2/DP1/HVA22-related protein
Large_WhitevsLarge_White_necta	INDEL	15,509,039	AT		380	60	UP_1066	ppa003256m.g	IQ motif (protein kinase), EF-hand binding site
OdedvsYuval_necta	INDEL	15,181,949		CTT	431	58	UP_3194	ppa004023m.g	HEAt domain03;Armadillo-like helical domain protein-protein interaction
JulyvsJuly_necta	SNP	16,280,987	T	C	219	32	INTERGENIC		
FlamevsFlame_Pearson_necta	INDEL	15,661,952		TA	406	60	INTERGENIC		

DP: combined depth across samples ;MQ: mapping quality

Several genes have been reported to have a role in trichome formation. These genes are SPL transcription factors (Shikata *et al.,* 2009; Yu *et al.*, 2010) acting directly over MYB factors and MAD-box genes.

Apart of ppa010308m.g with polymorphisms in all pairs of sports, we found also an INDEL in an intron of the MADs box genes ppa015857 m.g, SNPS and one INDEL in non coding regions of ppa010391 m.g (also a MADs-box gene) between 'Flameprince' and its sport mutants and between 'Julyprince' and 'Julyprince_nectareine), one INDEL upstream the MYB factor ppa010908m.g in the 'Oded'-'Yuval' pair.

Another interesting candidate gene to be involved in the trichome development of these sport mutants would be the ppa024172m.g which codifies an extracellular glycosyl-phosphatidyl inositol-anchored protein which belongs to the COBRA protein family (Brady *et al.,* 2007). This protein family is involved in cell expansion in *Arabidopsis* playing an important role in cellulose deposition. This protein could be associated with the trichome development since trichomes are expansions of epidermal cells (Smith & Oppenheimer, 2005).

There are also two pairs of samples showing small variants in a gene codifying a protein that contains a WD-40 repeat (also known as WD or beta-transducin repeats) which are short ~40 amino acid motifs, often terminating in a Trp-Asp (W-D) dipeptide. These WD-repeat proteins are a large family found in all eukaryotes and are implicated in a variety of functions ranging from signal transduction and transcription regulation to cell cycle control and apoptosis. The WD40 repeats serve as platforms for the assembly of protein complexes or as a site for protein-protein interactions. On the other hand, we also observed several genes (ppa020940m.g. ppa018631m.g. ppa010295m.g and ppa026684m.g) that will transcribe proteins containing Helix-loop-helix DNA-binding sites. These two proteins and the previous MYB transcription factor mentioned above could be playing in peach a similar role than in other plants by conforming a regulatory complex which will comprise a R2R3-MYB transcription factor, a basic helix-loop-helix (bHLH) domain protein and a WD40 repeat protein that would regulate the production of anthocyanins and also it would control the formation of trichomes (Baudry *et al.,* 2004; Broun*,* 2005*;* Serna & Martin, 2006). In *Arabidopsis*, these proteins are encoded by *Transparent Testa2* (*TT2,Myb*),*Transparent Testa8* (*TT8*, *HLH*) and *Transparent Testa Glabrous1* (*TTG1*, *WD40 repeat*), which together regulate the late flavonoid pathway (Baudry *et al.,* 2004). Their lost of function leads to a lack of anthocyanin pigmentation in foliar tissue and a loss of proanthocyanidins (Pas) in the seed coat (Nesi *et al.,* 2000; Nesi *et al.,* 2001). The presence of TTG1 is essential in this complex for anthocyanin

biosynthesis, trichome formation, seed mucilage production and root hair formation (Walker *et al.,* 1999). Several other WD40 repeat proteins functionally orthologous to TTG1 have been described from other species such as petunia (*Petunia hybrida*), cotton (*Gossypium hirsutum*) and maize (*Zea mays*); a mutation of some of them affect both anthocyanin/PA and trichome phenotypes, whereas mutation of others only affects the anthocyanin/PA phenotype (Carey *et al.,* 2004; de Vetten *et al.,* 1997; Humphries *et al.,* 2005; Lloyd *et al.,* 1992; Sompornpailin *et al.,* 2002).

All these genes could be involved in the trichome development pathway but a deep molecular study is needed in order to elucidate their possible function in the trait.

Table CIII.14. Small variants with an homozygous genotype as the reference for peach and an heterozygous genotype for nectarine (hypothesis 2) and genomic regions where they occured. Continued.

Pair	Variant	Position	REF	ALT	DP	MQ	Type	Gene or Repeat	Predicted protein
Large_WhitevsLarge_White_necta	SNP	14,660,634	A	C	530	34	UP_532	ppa011616m.g	Mitochondrial inner membrane translocase
Large_WhitevsLarge_White_necta	SNP	14,660,644	G	A	516	31	UP_522	ppa011616m.g	Mitochondrial inner membrane translocase
FloridaGlovsGall_necta	SNP	14,682,064	A	C	412	60	DOWN_144	ppa009314m.g	UVR domain nucleotide-excision repair
FlamevsFlame_Ham_necta	SNP	14,750,787	T	C	414	59	UP_1134	ppa015648m.g	Zinc finger. Dof-type. DNA binding. zinc ion binding
FlamevsFlame_Pearson_necta	SNP	14,750,787	T	C	414	59	UP_1134	ppa014277m.g	Unknown
JulyvsJuly_necta	INDEL	14,827,085		TCTC	373	58	UTR_5'	ppa007712m.g	RST domain of plant C-terminal
FlamevsFlame_Pearson_necta	SNP	14,835,391	A	G	445	60	UP_2555	ppa014277m.g	Unknown
FlamevsFlame_Ham_necta	SNP	14,837,542	T	C	483	59	UP_404	ppa014277m.g	Unknown
FlamevsFlame_Ham_necta	SNP	14,845,700	G	T	526	60	INTRON	ppa004183m.g	Cytochrome P450. E-class. group I
OdedvsYuval_necta	INDEL	14,855,073	AGAG		333	56	UP_ 1598	ppa010908m.g	Myb transcription factor
FlamevsFlame_Ham_necta	SNP	14,865,036	T	A	496	59	INTRON	ppa001986m.g	Tetratricopeptide-like helical
JulyvsJuly_necta	INDEL	14,875,429		AGACCCCAAAAG	402	58	UP_971	ppa007935m.g	Tyrosine-protein kinase. catalytic domain
FlamevsFlame_Ham_necta	SNP	14,879,012	A	G	525	60	DOWN_56	ppa009176m.g	Uncharacterised protein family UPF0497. trans-membrane plant
FlamevsFlame_Pearson_necta	SNP	14,891,737	T	A	442	60	UP_785	ppa020940m.g	Helix-loop-helix DNA-binding
FlamevsFlame_Ham_necta	SNP	14,904,850	C	T	517	59	UP_3361	ppa020940m.g	Helix-loop-helix DNA-binding
FlamevsFlame_Ham_necta	SNP	14,905,364	T	C	507	60	UP_2847	ppa020940m.g	Helix-loop-helix DNA-binding
FlamevsFlame_Pearson_necta	SNP	14,905,364	T	C	507	59	UP_2847	ppa020940m.g	Helix-loop-helix DNA-binding
FlamevsFlame_Ham_necta	SNP	14,905,495	A	C	477	60	UP_2716	ppa020940m.g	Helix-loop-helix DNA-binding
FlamevsFlame_Pearson_necta	SNP	14,905,495	A	C	477	60	UP_2716	ppa020986m.g	HNH endonuclease
FlamevsFlame_Ham_necta	SNP	14,914,814	G	A	419	59	DOWN_1793	ppa026684m.g	Helix-loop-helix DNA-binding
FlamevsFlame_Pearson_necta	INDEL	14,929,444	GA		318	59	INTRON	ppa021908m.g	RAG1- protein-1-related. activation-expression of recombination activation genes
FloridaGlovsGall_necta	INDEL	14,929,444	GA		318	59	INTRON	ppa021908m.g	RAG1- protein-1-related. activation-expression of recombination activation genes
JulyvsJuly_necta	INDEL	14,929,444	GA		318	59	INTRON	ppa021908m.g	RAG1- protein-1-related. activation-expression of recombination activation genes

Table CIII.14. Small variants with an homozygous genotype as the reference for peach and an heterozygous genotype for nectarine (hypothesis 2) and genomic regions where they occured. Continued.

Pair	Variant	Position	REF	ALT	DP	MQ	Type	Gene or Repeat	Predicted protein
FlamevsFlame_Ham_necta	SNP	14,930,578	T	A	443	60	UP_826	ppa021908m.g	RAG1- protein-1-related. activation-expression of recombination
JulyvsJuly_necta	INDEL	14,950,284		AA	408	59	UP_445	ppa017088m.g	Spermine synthase
FlamevsFlame_Ham_necta	SNP	14,956,871	C	T	462	59	INTRON	ppa010209m.g	Unknown
FlamevsFlame_Ham_necta	INDEL	14,961,048	AG		394	59	UP_3417	ppa010209m.g	Unknown
FlamevsFlame_Pearson_necta	INDEL	14,961,048	AG		394	59	UP_3417	ppa010209m.g	Unknown
JulyvsJuly_necta	INDEL	15,195,809	ATAT		378	59	INTRON	ppa024758m.g	F-box domain. cyclin-like: present in numerous protein
Large_WhitevsLarge_White_necta	SNP	15,225,312	A	T	438	59	INTRON	ppa000926m.g	Armadillo-like helical
Large_WhitevsLarge_White_necta	SNP	15,226,792	A	T	441	60	INTRON	ppa000926m.g	Armadillo-like helical
OdedvsYuval_necta	SNP	15,369,739	T	C	483	59	UP_2274	ppa000934m.g	ATPase. P-type. K/Mg/Cd/Cu/Zn/Na/Ca/Na/H-transporter
Large_WhitevsLarge_White_necta	SNP	15,370,357	T	A	428	59	UP_1656	ppa000934m.g	ATPase. P-type. K/Mg/Cd/Cu/Zn/Na/Ca/Na/H-transporter
OdedvsYuval_necta	SNP	15,383,505	T	C	475	60	UP_2874	ppa003007m.g	Major facilitator superfamily domain. general substrate transporter
JulyvsJuly_necta	INDEL	15,386,106	C		294	60	INTRON	ppa003155m	Oligopeptide transporter
OdedvsYuval_necta	INDEL	15,387,175	A		549	60	UP_793	ppa003155m	Oligopeptide transporter
OdedvsYuval_necta	SNP	15,392,867	C	T	463	58	UP_2189	ppa024635m.g	Oligopeptide transporter
Large_WhitevsLarge_White_necta	SNP	15,392,921	G	A	426	59	UP_2243	ppa024635m.g	Oligopeptide transporter
FlamevsFlame_Ham_necta	SNP	15,393,874	C	A	516	60	UP_2418	ppa025181m.g	Cytochrome P450. E-class; monooxygenase activity.electron carrier activity
OdedvsYuval_necta	SNP	15,444,360	C	T	487	59	UP_645	ppa018365m.g	HSP20-like chaperone
OdedvsYuval_necta	SNP	15,460,858	C	T	467	59	UP_315	ppa001544m.g	Unknown
Large_WhitevsLarge_White_necta	SNP	15,460,867	A	C	503	58	UP_324	ppa001544m.g	Unknown
FloridaGlovsGall_necta	INDEL	15,461,445		TC	299	56	UP_953	ppa001544m.g	Unknown
JulyvsJuly_necta	INDEL	15,461,445		TC	299	56	UP_953	ppa001544m.g	Unknown
OdedvsYuval_necta	INDEL	15,461,445		TC	299	56	UP_ 953	ppa001544m.g	Unknown
Large_WhitevsLarge_White_necta	INDEL	15,469,492		AG(*7)	334	54	UP_1543	ppa027096m.g	Domain of unknown function DUF292. eukaryotic

Table CIII.14. Small variants with an homozygous genotype as the reference for peach and and a heterozygous genotype for nectarine (hypothesis 2) and genomic regions where they occured. Continued.

Pair	Variant	Position	REF	ALT	DP	MQ	Type	Gene or Repeat	Predicted protein
FloridaGlovsGall_necta	INDEL	15,492,028	CT		433	58	UP_344	ppa021195m.g	sequence-specific DNA binding transcription factor activity
JulyvsJuly_necta	INDEL	15,492,028	CT		433	58	UP_344	ppa021195m.g	sequence-specific DNA binding transcription factor activity
FloridaGlovsGall_necta	INDEL	15,559,872	GA		337	58	UP_392	ppa018631m.g	Helix-loop-helix DNA-binding
OdedvsYuval_necta	INDEL	15,559,872	GA		337	58	UP_ 392	ppa018631m.g	Helix-loop-helix DNA-binding
FlamevsFlame_Ham_necta	INDEL	15,605,596		TA	341	58	UP_4459	ppa010295m.g	Helix-loop-helix DNA-binding
JulyvsJuly_necta	INDEL	15,605,596		TA	341	58	UP_4459	ppa010295m.g	Helix-loop-helix DNA-binding
FlamevsFlame_Ham_nectaD	INDEL	15,685,034	AT		365	60	UP_3224	ppa014170m.g	Zinc finger. PARP-type
FloridaGlovsGall_necta	INDEL	15,685,034	AT		365	60	DOWN_1302	ppa012031m.g	Unknown
FlamevsFlame_Ham_necta	INDEL	15,821,800		A	478	60	UP_1855	ppa005339m.g	Zinc finger. Dof-type_DNA binding
FlamevsFlame_Pearson_necta	INDEL	15,821,800		A	478	60	UP_1855	ppa005339m.g	Zinc finger. Dof-type_DNA binding
OdedvsYuval_necta	INDEL	15,828,078		TCTC	271	58	DOWN_ 2065	ppa005339m.g	Zinc finger. Dof-type_DNA binding
FlamevsFlame_Ham_necta	SNP	15,831,000	A	G	477	59	DOWN_914	ppa020986m.g	HNH endonuclease
JulyvsJuly_necta	INDEL	15,836,800	G		475	59	UP_1996	ppa009955m.g	Biotin/lipoate A/B protein ligase. protein modification process.
JulyvsJuly_necta	INDEL	15,842,144	A		380	59	INTRON	ppa015857m.g	Transcription factor. MADS-box.protein dimerization activity
Large_WhitevsLarge_White_necta	SNP	15,849,291	G	T	506	59	UP_499	ppa010443m.g	Expansin/pollen allergen. DPBB domain
Large_WhitevsLarge_White_necta	INDEL	15,862,567	G		494	59	F_SHIFT(-56)	ppa011784m.g	Protein of unknown function DUF962
FloridaGlovsGall_necta	SNP	15,871,338	G	A	549	57	UP_1502	ppb024333m.g	Zinc finger. CCHC-type; nucleic acid binding. zinc ion binding
FlamevsFlame_Pearson_necta	INDEL	15,890,076	CT		425	59	DOWN_464	ppa007653m.g	No apical meristem (NAM) protein. transcriptional regulator
FlamevsFlame_Ham_necta	INDEL	15,941,494	AT		407	59	DOWN_2627	ppa003476m.g	galactoside 2-alpha-L-fucosyltransferase activity
OdedvsYuval_necta	INDEL	15,941,494	AT		407	59	DOWN_ 2627	ppa003476m.g	galactoside 2-alpha-L-fucosyltransferase activity
FloridaGlovsGall_necta	SNP	15,984,465	G	C	601	35	N_SYNO S600C	ppa021345m.g	Ribonuclease H-like domain. Reverse transcriptase
FloridaGlovsGall_necta	SNP	16,131,048	G	T	164	32	N_SYNO L126F	ppa003804m.g	Glycoside hydrolase. family 79

Table CIII.14. Small variants with an homozygous genotype as the reference for peach and an heterozygous genotype for nectarine (hypothesis 2) and genomic regions where they occured. Continued.

Pair	Variant	Position	REF	ALT	DP	MQ	Type	Gene or Repeat	Predicted protein
Large_WhitevsLarge_White_necta	INDEL	16,135,606	G		123	44	UP_291	ppa023293m	Alcohol dehydrogenase superfamily. zinc-type
OdedvsYuval_necta	INDEL	16,135,606	G		123	44	UP_ 291	ppa023293m	Alcohol dehydrogenase superfamily. zinc-type
FloridaGlovsGall_necta	SNP	16,141,466	G	T	157	31	N_SYNO L126F	ppa003813m.g	Glycoside hydrolase. superfamily
JulyvsJuly_necta	SNP	16,181,241	G	A	517	58	SYNO V301	ppa005788m.g	Zinc finger. C2H2-like
Large_WhitevsLarge_White_necta	INDEL	16,318,173		TC	329	57	UP_2235	ppa005034m.g	UDP-glucose/GDP-mannose dehydrogenase. dimerisation
FlamevsFlame_Pearson_necta	INDEL	16,417,314	AC		340	60	INTRON	ppa008674m.g	ATPase. F1 complex. gamma subunit conserved site
FlamevsFlame_Pearson_necta	SNP	16,458,821	G	C	322	32	UP_2125	ppa023669m.g	Known
FloridaGlovsGall_necta	INDEL	16,481,203		CT	334	58	UP_912	ppa014630m.g	Transcription factor TCP subgroup
Large_WhitevsLarge_White_necta	INDEL	16,481,203		CT	334	58	UP_912	ppa014630m.g	Transcription factor TCP subgroup
OdedvsYuval_necta	INDEL	16,507,507	GA		375	59	UP_1000	ppa002296m.g	Unknown
JulyvsJuly_necta	SNP	16,526,005	G	A	366	32	UP_2456	ppa023669m.g	Known
FlamevsFlame_Pearson_necta	SNP	16,526,044	C	G	299	32	UP_2495	ppa023669m.g	Known
JulyvsJuly_necta	SNP	16,526,044	C	G	322	32	UP_2495	ppa023669m.g	Known
FlamevsFlame_Pearson_necta	SNP	16,526,066	G	T	477	58	UP_2517	ppa004083m.g	Cytochrome P450. E-class. group I
JulyvsJuly_necta	SNP	16,526,066	G	T	299	32	UP_2517	ppa023669m.g	Known
Large_WhitevsLarge_White_necta	SNP	16,526,996	A	G	527	31	UP_3447	ppa023669m.g	Known
FlamevsFlame_Ham_necta	INDEL	16,527,042	C		267	36	UP_3494	ppa023669m.g	Known
FlamevsFlame_Ham_necta	SNP	16,558,145	C	T	477	55	DOWN_912	ppa004083m.g	Cytochrome P450. E-class. group I
FlamevsFlame_Pearson_necta	SNP	16,558,145	C	T	382	38	DOWN_912	ppa004540m.g	Pentatricopeptide repeat
OdedvsYuval_necta	INDEL	16,559,694		CT	367	40	DOWN_2475	ppa004083m.g	Cytochrome P450. E-class. group I
JulyvsJuly_necta	SNP	16,559,714	A	G	302	40	DOWN_2481	ppa004083m.g	Cytochrome P450. E-class. group I
Large_WhitevsLarge_White_necta	SNP	16,559,882	G	A	481	32	UP_4872	ppa004540m.g	Pentatricopeptide repeat

Table CIII.14. Small variants with an homozygous genotype as the reference for peach and an heterozygous genotype for nectarine (hypothesis 2) and genomic regions where they occured. Continued

Pair	Variant	Position	REF	ALT	DP	MQ	Type	Gene or Repeat	Predicted protein
OdedvsYuval_necta	SNP	16,559,906	C	A	557	35	UP_ 4848	ppa004540m.g	Pentatricopeptide repeat
OdedvsYuval_necta	SNP	16,560,329	T	C	412	37	UP_4421	ppa004540m.g	Pentatricopeptide repeat
JulyvsJuly_necta	SNP	16,560,989	G	A	342	31	UP_3765	ppa004540m.g	Pentatricopeptide repeat
FlamevsFlame_Ham_necta	SNP	16,562,325	T	C	382	58	N_SYNO L8F	ppa016508m.g	Serine-threonine/tyrosine-protein kinase catalytic domain
FlamevsFlame_Pearson_necta	SNP	16,562,325	T	C	478	59	UP_2429	ppa004540m.g	Pentatricopeptide repeat
JulyvsJuly_necta	SNP	16,562,325	T	C	382	38	UP_2429	ppa004540m.g	Pentatricopeptide repeat
Large_WhitevsLarge_White_necta	SNP	16,562,325	T	C	382	38	UP_2429	ppa004540m.g	Pentatricopeptide repeat
FlamevsFlame_Pearson_necta	SNP	16,563,464	A	C	509	60	UP_1290	ppa024172m.g	Glycosyl-phosphatidyl inositol-anchored; COBRA-like.
FloridaGlovsGall_necta	INDEL	16,563,797		AG(*5)	379	41	UP_933	ppa004540m.g	Pentatricopeptide repeat
JulyvsJuly_necta	SNP	16,565,866	T	C	413	59	SYNO CF344	ppa004540m.g	Pentatricopeptide repeat
FlamevsFlame_Ham_necta	INDEL	16,573,383		CTTT	415	58	UP_66	ppa024172m.g	Glycosyl-phosphatidyl inositol-anchored; COBRA-like.
FlamevsFlame_Pearson_necta	INDEL	16,573,383		CTTT	415	58	UP_66	ppa024172m.g	Glycosyl-phosphatidyl inositol-anchored; COBRA-like.
FlamevsFlame_Pearson_necta	SNP	16,576,381	A	G	265	30	INTRON	ppa014838m.g	Glycosyl-phosphatidyl inositol-anchored. COBRA-like
FlamevsFlame_Ham_necta	SNP	16,586,088	C	T	520	38	UP_2429	ppa004540m.g	Pentatricopeptide repeat
OdedvsYuval_necta	INDEL	16,586,190		A	365	51	INTRON	ppa016508m.g	Serine-threonine/tyrosine-protein kinase catalytic domain
FloridaGlovsGall_necta	SNP	16,586,204	T	A	357	48	INTRON	ppa016508m.g	Serine-threonine/tyrosine-protein kinase catalytic domain
FlamevsFlame_Ham_necta	SNP	16,592,422	G	C	265	54	UP_1993	ppa014838m.g	Glycosyl-phosphatidyl inositol-anchored. COBRA-like
FlamevsFlame_Pearson_necta	SNP	16,592,422	G	C	416	59	UP_1993	ppa010391m.g	TF K-box. MADs-BOX
FlamevsFlame_Ham_necta	INDEL	16,619,508	GAGAGC		362	56	UTR-3'	ppa014414m.g	Unknown
Large_WhitevsLarge_White_necta	SNP	16,620,461	G	C	445	60	UP_514	ppa014414m.g	Unknown
FloridaGlovsGall_necta	SNP	16,621,437	A	G	445	59	UP_1490	ppa014414m.g	Unknown

Table CIII.14. Small variants with an homozygous genotype as the reference for peach and an heterozygous genotype for nectarine (hypothesis 2)and genomic regions where they occured. Continued

Pair	Variant	Position	REF	ALT	DP	MQ	Type	Gene or Repeat	Predicted protein
FlamevsFlame_Ham_necta	SNP	16,623,025	A	C	500	30	UP_3078	ppa014414m.g	Unknown
JulyvsJuly_necta	SNP	16,625,445	T	A	441	59	UP_849	ppa010391m.g	TF K-box. MADs-BOX
JulyvsJuly_necta	INDEL	16,626,415		AG(*8)	300	58	UTR_5'	ppa010391m.g	TF K-box. MADs-BOX
FlamevsFlame_Ham_necta	SNP	16,627,679	C	A	455	59	INTRON	ppa010391m.g	TF K-box. MADs-BOX
FlamevsFlame_Pearson_necta	SNP	16,627,679	C	A	467	60	INTRON	ppa010391m.g	TF K-box. MADs-BOX
FlamevsFlame_Ham_necta	SNP	16,630,867	G	T	416	59	DOWN_191	ppa010391m.g	TF K-box. MADs-BOX
FlamevsFlame_Pearson_necta	SNP	16,630,867	G	T	416	59	DOWN_191	ppa008947m.g	Mitochondrial substrate/solute carrier
JulyvsJuly_necta	SNP	16,630,867	G	T	416	59	DOWN_191	ppa010391m.g	TF K-box. MADs-BOX
FlamevsFlame_Ham_necta	INDEL	16,632,424		GGTT	407	55	UP_2232	ppa010308m.g	Transcription factor. MADS-box. K-box
FlamevsFlame_Pearson_necta	SNP	16,633,341	T	G	431	59	UP_1318	ppa010308m.g	Transcription factor. MADS-box. K-box
FloridaGlovsGall_necta	SNP	16,633,341	T	G	431	59	UP_1318	ppa010308m.g	Transcription factor. MADS-box. K-box
JulyvsJuly_necta	SNP	16,633,341	T	G	431	59	UP_1318	ppa010308m.g	Transcription factor. MADS-box. K-box
Large_WhitevsLarge_White_necta	SNP	16,633,341	T	G	431	59	UP_1318	ppa010308m.g	Transcription factor. MADS-box. K-box
JulyvsJuly_necta	SNP	16,633,800	G	A	206	55	UP_859	ppa010308m.g	Transcription factor. MADS-box. K-box
JulyvsJuly_necta	SNP	16,634,196	A	T	550	32	UP_463	ppa010308m.g	Transcription factor. MADS-box. K-box
FloridaGlovsGall_necta	SNP	16,634,314	C	T	263	30	UP_345	ppa010308m.g	Transcription factor. MADS-box. K-box
JulyvsJuly_necta	SNP	16,634,314	C	T	263	30	UP_345	ppa010308m.g	Transcription factor. MADS-box. K-box
OdedvsYuval_necta	SNP	16,634,314	C	T	263	30	UP_ 345	ppa010308m.g	Transcription factor. MADS-box. K-box
FlamevsFlame_Pearson_necta	SNP	16,642,224	T	G	428	59	INTRON	ppa008947m.g	Mitochondrial substrate/solute carrier
FlamevsFlame_Pearson_necta	SNP	16,642,517	A	G	467	60	INTRON	ppa008947m.g	Mitochondrial substrate/solute carrier
FlamevsFlame_Ham_nectaD	INDEL	148,99,906	AT		488	59	DOWN_2331	ppa003003m.g	BURP domain
JulyvsJuly_necta	INDEL	165,83,657	A		437	60	UP_1179	ppa021191m.g	Glycosyl-phosphatidyl inositol-anchored. plant

GENERAL DISCUSSION

In this work I have used two genetic approaches to perform an exhaustive and valuable genetic analysis of three main agronomical traits in peach fruits, providing practical genetic tools for marker assisted selection.

One of the approaches consisted in region-based association mapping and was applied to examine the genetic variability and haplotype extension along two loci: the one producing subacid fruits and the one responsible the fruit flat shape. This approach allowed the identification of genetic markers suitable for MAS for both traits and, additionally, identified the allele mutation responsible for the flat shape in peach The second approach consisted in sequencing the whole genome of 6 peach varieties and their respective sport mutants with nectarine phenotype which has provided some clues for future studies of variability, specially somatic variability, in peach.

The three traits studied (subacidity, fruit shape and peach-nectarine fruit) are controlled by major genes, so they all have Mendelian inheritance. Another common aspect is that they are key traits in breeding programmes. Although some SSR markers are already available for MAS for these traits, the high renovation rate of peach varieties and the size of the current breeding programmes demand high throughput markers (HTM) to accelerate efficiently the development of new varieties. The SNPs associated with these quality traits and the candidate causal gen reported in this work will allow for MAS in peach for the quick release of new varieties every year.

MARKER ASSISTED SELECTION OF SUBACID TRAIT IN PEACH

One of the three Mendelian traits analysed here was the responsible of fruit low-acidity (D), previously mapped in the proximal end of LG5 (Dirlewanger et al., 1998b, 2006). The low acid allele is dominant and here we validated the use of the SSR marker (CPPCT040) in MAS. One allele of this marker ($CPPCT040^{193}$), present either in homozygosis or heterozygosis, is associated with TA values lower than 5.5 g/L. This information allowed us to establish a TA value as threshold between acid and subacid. This marker will suppose a real advantage to save time and surface in the field, because it can be used when peach are still in the seedling stage avoiding the growth those that don't carry the desired fruit phenotype.

We were also interested in the analysis of the genomic regions flanking the SSR marked linked to this trait to evaluate the length of the allele and discover high throughput markers (SNPs) for MAS. The sequence of a 70.4 Kbp region flanking CPPCT040 revealed high variability in this region respect to the one observed genome-wide. In total we found a density of 1 SNPs every 310 bp, which is almost double than the one observed by Aranzana *et al.,* (2010) after sequencing 23

fragments genome-wide distributed in 47 peach commercial varieties. The haplotype conserved around the CPPCT40 and linked to the subacid allele was longer than 24 Kbp. The haplotype was unique for all subacid varieties, reflecting a unique origin for the low acid phenotype. The most likely hypothesis is that in the beginning of the US breeding programs Chinese material carrying a single subacid allele was used and thereafter spread to the rest of the whole. One of the varieties ('Babygold 7') with a TA value close to the boundary between the two phenotypic classes (acid and subacid), presented a unique SSSR allele (CPPCT040^{195}) and conserved the subacid SNP haplotype linked to the CPPCT040^{193} allele. We hypothesize here that the new 195 was a recent step-wise mutation of the 193. Alternatively 'Babygold 7' could carry a new allele with SNP variability not detected within the fragments sequenced and in this case we could accept the existence of two different origins of the subacid allele. This variety is an old variety with genome-wide differences so we cannot discard this last hypothesis. Two acid varieties ('Villa Giulia' and 'Flavor Gold') presented a low frequent SSR allele (CPPCT040^{201}) which was linked to a haplotype with also low frequent SNPs.

The haplotype linked to the subacid allele contained 8 linked SNPs useful for diagnosis of this trait which can be included in any of the current available high throughput genotyping platform to provide a fast and accurate selection of varieties at the seedling stage. Here, as a prove of concept, we tested and validated one of them with High Resolution Melting (HRM) methodology.

The analysis of the length and association of haplotypes can be used to identify candidate genes for the studied trait. When a favourable mutation is positively selected, the variability close to it is also swept along. The extension of the haplotype with variants linked with the mutation is reduced through recombination during generations. Linkage disequilibrium (LD) extension in peach, i.e. length of haplotype with linked variants, is relatively high (Li *et al.,* 2013), and thereafter region-based and genome-wide association analysis can succeed in peach, especially since the availability of *Prunus* genome annotation. In *Prunus* several candidate genes have been mapped in the *D* locus. We found 3 annotated genes within the long conserved sub acid haplotype (24 Kb) but none of them have been reported to be involved in fruit acidity. In order to explore other candidate genes in this locus it would be necessary to expand the sequencing upstream and downstream the region. Some preliminary results, not included in this work, indicate that the subacid haplotype extends upstream CPPCT040 at least 80 Kb although more analysis is required. Within the upstream region there are two possible functional candidate genes. The gene ppa000751m is a calcium binding site, which could be involved in the modulation and content of protons in the fruit and therefore control fruit acidity levels. In sweet cherry a Ca post-harvest treatment results in a retarding TA loss associated with decreasing fruit metabolism, including respiration rate (Wang *et al.,* 2014). During the

respiration process the organic acid might be used as carbon source in the tri-carboxylic acid cycle, resulting in a decrease of TA concentration during fruit storage. The other candidate gene annotated in this region is ppa012357m, a glycoside hydrolase; these enzymes have been described in 29 families in rice and *Arabidopsis* and the majority of them play a role in cell wall polysaccharide metabolism. Other functions of glycoside hydrolases are the participation in the biosynthesis and re-modulation of glycans, mobilization of energy, defense, symbiosis, signaling, and metabolism of glycolipids. To test the implication of these candidate genes in fruit acidity, future analysis of the genomic region and validation experiments to confirm the possible associations are needed.

CLONING A CANDIDATE GENE FOR FLAT SHAPE IN PEACH

Flat shape in peach is controlled by a single dominant gene *S* (for saucer-shaped), (Lesley, 1939) mapped in the distal part of chromosome 6 (Dirlewanger *et al.,* 1998b). Due to its dominant behaviour, fruit carrying the flat allele in either homozygosis or heterozygosis should be flat, but just the heterozygous genotypes show this phenotype. Homozygous fruits, instead, abort two months after anthesis. This fact makes possible two alternative hypotheses for the genetic of this trait. The best supported hypothesis is the existence of a single gene, and the alternative one, is the existence of two dominant closely linked genes in repulsion. In this last case, *S*-/*Af*- would produce flat peaches, *S*-/*afaf* would determine aborting fruits while round fruit would have *ss*/*Af*- genotype (Dirlewanger *et al.,* 2006).

Our objective was to look for SNPs associated to flat shape for further application in MAS. Although no candidate genes had been identified for this trait, the availability of one SSR marker associated to the *S* locus in several mapping populations (Dirlewanger *et al.,* 2006; Picañol *et al.,* 2012) and in a wide range of germplasm (Picañol *et al.,* 2012) allowed to decide the starting point. The sequence analysis of this locus showed low heterozygosity which is in discordance with the obligated heterozygous genotype of the flat fruits. Accounting for the high level of LD in peach we moved our research area to a very polymorphic region 300Kb upstream the associated SSR marker. The SNPs in this region spread 26.7 Kb but we were able to narrow the associated region to 1Kb (split in two amplicons). This region was located on coding sequence annotated in the reference genome and corresponds to ppa025511m gene (scaffold_6:24,405,493-24,407,745) annotated as binding protein (GO:0005515) containing Leucine-rich domains (Leucine-rich repeat-containing N-terminal, type 2). The gene codifies 2 CDS (ppa025511.CDS1 and ppa025511.CDS2). The validation of this association was done by amplifying and sequencing both amplicons in a wide and diverse

sample comprising flat, round an aborting peaches. From this data we identified eleven associated SNPs and two INDELs; the first INDEL consisted in 8bp deletion in round peaches and the second one in 13bp deletion in aborting samples; flat peaches showed the SNPs and both INDELs in heterozygosis. As the two observed haplotypes appeared together in flat varieties, the sequence between the two INDELs was illegible; both haplotypes were confirmed by PCR cloning which also confirmed the SNPs between flat and round alleles.

As observed in the genome browser, the variability of this region was close to 3.5 times higher than the one observed genome-wide, with 1 SNP every 172 bp in the S locus while the variability genome-wide is about 1 SNP every 598 bp (Aranzana *et al.,* 2010). Contrary, the haplotype was shorter than what expected from peach LD extension. The posterior analysis of the region has revealed that at least the SNPs found in the sequence of kinasa-3 and kinasa-4, both amplicons of the reverse transcriptase gene ppa024472, were product of the sequence of two different genes, one in the region desired and one in chromosome 7.

From the sequence of the flat and round alleles we designed allele specific marker able to differenciate both of them, generating two bands with 5bp of difference. Theses markers will be already useful for MAS. In addition, any of the eleven associated SNPs identified are also useful for the same purpose.

These SNPs were located in the second CDS of the ppa025511 gene but no additional SNPs were found on the first CDS or in the small intron. Different long PCRs performed at 12-20 Kb upstream this gene confirmed the existence of a big deletion starting few nucleotides upstream the 8bp deletion, affecting the 5'UTR of this gene in flat varieties. The mutation consisted in the absence of a region starting 9,324 bp upstream of the CDS1 (scaffold_6: 24,396,169) of the gene and ending 693bp downstream the CDS1 (scaffold_6: 24,406,186) lacking all CDS1, the 30bp intron and 214 bp of CDS2.

To validate the ppa025511 gene role in the fruit shape it would have been ideal to perform its genetic transformation into a round variety, but the woody species are difficult to regenerate *in vitro* although in the past it has been a wide development of protocols for the regeneration in different species, such as: cherry (Tang *et al.,* 2002), pistachio (Tilkat *et al.,*2009), apricot (Petri *et al.*, 2008) or peach (Hammerschlag *et al.,* 1985). Peach is one of the most recalcitrant species (Padilla *et al.,* 2006) but some authors have achieved it by using immature material of seeds (Hammerschlag *et al.,* 1985; Mante *et al.,* 1989; Pooler & Scorza, 1995) which implies that the regeneration is done

from a material with an unknown genotype and phenotype. But there is a lack of a protocol based on adult tissues (Liu & Pijut, 2009).

Instead of validating the gene role through transgenic transformation we analyzed the natural mutation in a flat variety which reverted to round. The mutation was chimeric and occurred in the second meristematic layer, which generates the fruit flesh. Although until now, we have not been able to obtain the sequence of the flat allele in the mutant, the analysis of flesh DNA with the allele specific marker for the ppa025511.CDS2 INDELS reveals a new structural mutation in the flat allele for all the flat varieties tested, while the skin DNA shows the intact flat and round alleles.

The round protein codified by ppa025511 round allele is similar to some receptor-like kinases (RLPKs) containing leucine-rich repeats. These proteins are ligand-receptors that by phosphorylation and un-phosphorylation control cell fate specification, cell divisions and cell to cell communication which are important function in the development of both plants and animals (Matsushima & Miyashita, 2012). The most similar RLK to our round protein is the At5g44700 protein codified by GSO2, involved in the maintenance of the epidermis at the beginning of the heart stage during embryogenesis in *Arabidopsis*. Double mutants of gso1 and gso2 (a quasi-orthologous of gso1) in *Arabidopsis* have shown mutant embryos that expand laterally at heart torpedo transition stage of embryogenesis (Racolta *et al.,* 2014). We could hypothesize that the LRR-kinase protein codified by the round allele of ppa025511 is involved in a possible cell signalling pathway during peach development that ensure a final round shape.

The control of carpel and fruit development has been study deeply in *Arabidopsis* and tomato (Ferrándiz, *et al.,* 1999; Rodríguez *et al.,* 2011). From this knowledge we could imagine what it could be happening from the floral meristem to the fruit set in peach but it is difficult to extrapolate this information to peach because their fruits are quite different to peach drupes. In general in all angiosperms fruit development starts with the formation of a flower from the floral meristem. The floral meristem will give rise to four whorls: the sepals, petals, stamen and pistil. The stamen provides the male reproductive structures giving rise to pollen. The pistil provides the female reproductive structure giving raise the ovules within the ovary. After anthesis, pollen will land on the stigma of the pistil and germinate; the pollen tube will grow through the style towards the ovules. Fertilization of the ovules leads to fruit development and the production of the seeds ends the reproductive cycle. Then, fruit development generally follows the Gillaspy *et al.* (1993) model, in which cell division is the first stage, followed by cell expansionthat will define the final fruit size of the fruit, thereafter will start to ripen to end up as a mature fruit. All peach tissues come from the ovary; the outer skin is the exocarp, the edible flesh comes from the mesocarp and the pit from the

endocarp. Studies in peach have revealed a crucial role of the PLENA-like (PpPLENA) gene (a MADS-box gene) during the transformation of the carpel into a ripe fleshy fruit (Causier *et al.,* 2005; Tadiello *et al.,* 2009).

A dosage effect could be one possible reason for the flat shape or aborting fruits; however this won't explain the restoration of the phenotype when the flat allele is altered. One plausible hypothesis of mechanism compatible with this phenomenon is a dominant-negative (DN) effect of the flat allele. DN mutations lead to polypeptides that usually interact with the wild-type allele disrupting its activity, thereafter these mutations cause more severe effects that simple null alleles of the same gene (Read & Strachan, 2004) which would explain the abortion of the fruits in homozygous genotypes. This mechanism has been already observed in other receptor like kinases such as the case of CLAVATA loci in *Arabidopsis* (Diévart *et al.,* 2003) already explained in the discussion part of the second chapter, or the case of ERECTA family genes; ERECTA and two paralogous, ERECTA-LIKE-1 (ERL1) and ERL2, which evolved from a recent duplication, regulate the organ shape and the inflorescence architecture in *Arabidopsis*. The ER mutants that lack some of the genes but not all develop compact but normal inflorescences, but when all the ER-family genes are missing (triple mutants) the phenotypes observed are extreme like dwarf and sterile plants. The mechanism acts as following: in the absence of ER and ERL1, ERL2 is haploinsuffincient for female sterility producing aberrant ovule growth and abortion of embryo sac, whereas ERL1 is haplosufficient in the absence of ER and ERL2. On the other hand ERL2 is haplosufficient for inflorescence elongation and floral patterning (Pillitteri *et al.,* 2007).

Another hypothesis compatible with the reversion of the round shape in the mutant is the recombination of the mutant flat allele with other of the LRR-Kinase located around ppa025511 LRR-kinase. Thus, this recombination would complement the mutation of the flat allele recovering the functional activity of the receptor or would produce a new round allele. Recombinant receptors are not been observed in nature however some have been produced artificially and transformed into plants which have shown the functionality acquired from their fusion (Albert *et al.,* 2010; Diévart *et al.,* 2003; Zhang *et al.,* 2011).

Future experiments studying the mutation in the flesh DNA of this natural mutant will help to construct a hypothesis of the mechanism underlying the fruit shape in peach. Moreover, the study of the chimeric mutation could provide some clues to understand the genetic mosaicism that occur with a relatively high frequency in peach.

SOMATOCLONAL VARIABILITY BETWEEN PEACH-NECTARINE SPORT MUTANTS

Differently from animals, plants do not follow Weismann's doctrine, which proposes that a mutation that occurs outside of the germline (the cell lineage producing the gametes) cannot be inherited through gametes (Weismann, 1892)

The plant gametophytes (pollen and ovules) contain the gametes which are descendants of meristematic cell layers that have given rise not only to the gametophytes but also all the airborne tissues of the plant. Thus, gametes can be produced from cell lineages that may have undergone imperfect mitoses and errors during DNA replication that may be inherited through gametes. These somatic mutations occur naturally and accumulate producing mosaicism during plant growth (Gill *et al.,* 1995). Some of them can produce interesting phenotypes that derive in new cultivars. Some sport mutant examples have been reported already in nature, for grapefruit (Hartmann & Kester, 1975; Wegscheider *et al.,* 2009), banana (Simmonds, 1966), potato (Howard, 1970), flower shape, maturity day, flesh color and glabrous skin in peach (Scorza & Sherman, 1996). Recently, the comparison of two peach sports showing a different flesh color (yellow and white) has been a successful strategy for the identification of a candidate gene for such trait (Brandi *et al.,* 2011).

In the case of peach, this kind of variability was studied in 28 sport mutants with 50 SSRs, estimating a mutation rate of $2.1*10^{-3}$ per allele (Aranzana *et al.,* 2010). Here we provide the first insight regarding the somatic mutations between several pairs of peach to nectarine sport mutants using massively parallel genome sequencing.

Whole genome sequences of the peaches and their respective nectarine mutants were obtained through sequencing by synthesis in Illumina HiSeq platform (Bentley, 2006). Then, we aligned these sequences against the peach reference genome (Verde *et al.,* 2013) using one of the two major alignment algorithms used: the Burrows Wheeler transform (BWT)-based algorithm, method employed by BWA software (Li, 2009). Following the alignment we performed the single nucleotide variants/INDEL calling using mpileup SAM tools (Li *et al.,* 2009). The results were filtered to select the best well supported calls, which had at least a depth of 10 reads and a general single variant quality equal or major than 20. These cut-off values are generally applied to variant called by SAMtools mpileup (Jia *et al.,* 2012). There are few consistent filtering parameters including base quality, mapping quality and coverage supporting reads. Our data fulfilled all these quality thresholds however we found higher variability than expected between clones, which ranged from 8,000 to 13,500 suggesting a high rate of false positive small variants. Consequently, to be more conservatives in the variant calling, we used a more restraining filter based on the likelihoods of the

given genotypes field. Although there are not specific filtering rules with PL parameter, there are examples of analysis where PL filtering has been applied (Allen *et al.,* 2013; DePristo *et al.,* 2011; Durtschi *et al.,* 2013; Jia *et al.,* 2012). This new filter removed the 99% of the total variants between pairs of clones. However we believe that with the application of this new filter we removed true variants. Normally, somatic differences between generations will be sequenced at a very low frequency and thereafter will be removed from the analysis of the sequences only those accumulated during years will be identified. Thus, even though our pipeline includes the most common wide used software for whole genome sequencing analysis nowadays, their use in the identification of low-frequent, somatic mutations remain a major challenge. The identification of somatic mutations requires a high sequencing coverage and new massive parallel sequencing approaches that do not tend to discard low-abundant variants as potential errors.

Moreover, one of the main problems of all the current sequencing technologies in the identification of somatic mutations is production of sequencing errors. We used Illumina technology which has a base pair error rate of 0.05-1% (Kinde *et al.,* 2011; Quail *et al.,* 2008). This error rate is several orders of magnitude higher than the expected somatic base pair mutations and therefore these events are masked by them. Another challenging question is the identification of random somatic rearrangements when using paired-end sequencing because they can be miscalled with chimeric sequences, i.e., ligation of two genomic sequences to each other, during the library preparation (Quail *et al.,* 2008). During the library preparation, DNA samples are randomly fragmented and then end-polished and appended with an A-overhang, which promotes preferential annealing with T-overhang-containing sequencing adapters that excludes cross-ligation. However cross-ligation occurs at low frequency when the sequencing adapters attach to all DNA fragments. To overcome this problem a series of stringent gel-based size-selected fragments are applied after and before the ligation reaction during library preparation (Quail *et al.,* 2008). After the sequencing errors, alignment errors are another source of false positives calls. They are associated normally with repetitive or homologous sequence regions that can lead to single to several base substitutions, insertion, and deletion errors (Treangen & Salzberg, 2013). Sequencing and alignment errors can be associated with certain sequence motifs, and consequently they can be consistent between samples sequenced on the same instrument using the same sequencing chemistry and alignment methods, as demonstrated in several studies (Abnizova *et al.,* 2012; Bansal *et al.,* 2010; Margraf *et al.,* 2010; Muralidharan *et al.,* 2012). After the alignment errors, the next source of false positive is produced by the actual variant calling step. There are many variant callers available but two of them have dominated modern genotyping; SAMtools (Li & Durbin, 2009) and GATK (DePristo *et al.,* 2011). Both have been developed to include parameters that help identify and reduce false positive variants

(DePristo *et al.,* 2011; Li, 2011a). These parameters include the probabilistic base quality and alignment mapping quality score, the aligned read coverage for possible alleles, and, more recently, the base alignment quality score(BAQ) (Li, 2011). Under SAMtools mpileup -0.1.18-sl61 the variant calling takes into account these parameters. However, the aligned read coverage for the reference or alternative alleles in the forward and reverse strands that is provided by DP4 field in the vcf file generated from SAMtools mpileup is referred to the total of reads with high quality between all the samples analysed together, so it is not possible to know how many reads are supporting each possible allele in this site for each sample. Thus, variants showing an allelic imbalance (where one allele makes up a greater fraction of reads than the second allele) are not shown in the genotype field for each sample.

Alternatively to SAMtools, GATK program can be used for variant calling but it has been proved that its use in the step after read mapping and before small variant calling by SAMtools mpileup helps to produce lower false positive rates. This is achieved by two main functions, the recalibration of mapping scores, which reduces the base quality scores of specific homopolymer motifs, identifies small intra-read insertions and deletions and realign the reads at this loci using alignment algorithm that includes low penalties for insertions and deletions leading to a final cleaner variant calls. The inclusion of this software in our pipeline should help to reduce the amount of false variants we observed between clonal pairs and it would also provide a better mapping before the small variant calling to avoid the loss of some of these rare mutations that are in low frequency and are normally missed due to the high error background.

Together with the study of somatic variability, the aim of the third chapter of the thesis was to identify the gene responsible for the nectarine trait in peach fruits. We postulated 2 possible working hypotheses to look for causal polymorphisms. Both considered that the causal mutation consisted in a small variant, in the first hypothesis the nectarine new allele in the clones was identical to the one fixed in the commercial varieties. While in the second the new mutations were different from the one fixed but all occurring in the same gene. In this work we describe several candidate genes under each of the hypothesis, however none of them coincides with the one described recently by Vendramin *et al.,* (2014). The reason for this is that the mutation in the fixed nectarine allele consists in a large insertion of 7 Kbp instead of in a small variant. Although we used the program SVDetect (Zeitouni *et al.,* 2010) to call structural variants (data not shown in the corresponding chapter), we were not able to identified the causal insertion, which can be due to the pipeline used or the low depth of the sequences.

In summary, our results are the first insight of the whole genome somatic variability between pairs of sport mutants for nectarine trait in peach. The two lessons learnt here to reduce the huge amounts of false positive obtained by the current NGS technologies and the application of the most widely used bioinformatics pipelines are first try to reduce error rates experimentally (reducing polymerase errors during library preparation) or by applying filtering after sequencing or by decreasing machines sequencing errors, that are expected to decrease in the near future. However, there is currently an alternative solution to reduce the sequencing-related errors derived from the second generation technologies which is the singe molecule sequencing. The reduction of random errors is provided by the elimination of amplification of DNA templates and by successive passes of sequencing of the same molecule that improve the accuracy and additionally can sequence molecules with high GC content or secondary structures. There are three main single-molecule technologies: Helicos BioSciences (Harris *et al.,* 2008), Pacific Biosciences (Eid *et al.,* 2009) and Life Technologies (Hardin, 2008). Although depending on the technology there are differences in the ability of make use of some of the single molecule sequencing advantages (Gupta *et al.,* 2008; Pettersson *et al.,* 2009; Voelkerding *et al.,* 2009), all the systems provide more even coverage and thus do not require too much depth for proper detection of heterozygotes. Then, the latest alternative would be to make use of the third generation sequencing technologies such as Oxford Nanopore which will enable almost unlimited read lengths because it does not relay in exogenous labels but rely instead on the electronic or chemical structure of the different nucleotides being sequenced. This technology uses an exonuclease cleavage reaction and a protein nanopore to read individual cleaved bases by a unique electrical signature produced as they pass through the pore (Venkatesan & Bashir, 2011).

Another sequencing approach to avoid the sequencing errors as confounders in the identification of low-abundant mutations is single cell sequencing (Shapiro *et al.,* 2013). This strategy sequences the genomes of single cell instead of mixture of genomes from whole tissues. Even though that the possibility of massively sequence single cells has the potential to revolutionize cancer research and possibly, developmental biology and plant genomics, it still suffers from amplification bias, resulting in uneven coverages (Raghunathan *et al.,* 2005). This fact makes necessary the development of better error-correction algorithms that do not assume uniformity of coverage. The cost and the lack of good multiplexing are still drawbacks for this technology that without doubt will become the future of genome sequencing.

CONCLUSIONS

1. We have demonstrated here that region-based association analysis can be successfully used in peach to identify markers associated to agronomic interesting traits and identify candidate genes.

2. The allele 193 of the SSR CPPCT040 SSR has been validated as a diagnostic marker for the subacid trait in peaches and can be used for marker assisted selection in peach breeding programs.

3. The titratable acidity (TA) value of 5.5 mg/l has been established as an objective cut-off point to classify varieties as acid of subacid.

4. The subacid haplotype in the varieties studied was unique, longer and clearly different from the acid one. This suggests a recent unique origin of the subacid allele.

5. This haplotype contains at least eight SNPs linked to the subacid trait. Any of them can be included in high throughput genotyping platforms for more efficient MAS in large breeding programmes.

6. None of the genes contained in the conserved subacid haplotype has been reported to have a role in fruit acidity. A more extensive study of the genomic region is needed to find more plausible candidate genes for this trait.

7. We didn't find association between the SSR alleles nor the SNPs with levels of acidity within subacid and within acid groups suggesting the existence of additional major genes or QTLs with epistatic interactions.

8. The sequence analysis of the *S* locus identified 2 INDELs and 11 SNPs within the gene ppa025511m highly associated with the flat shape in peach.

9. Two primers flanking the two INDELS can be used as fragment-size markers for MAS, while any of the 11 SNPs identified can be included in high throughput genotyping platforms for more efficient MAS in large breeding programmes.
10. The polymorphisms are conserved in all flat varieties, suggesting a unique origin for the flat allele.

11. The sequence analysis of the upstream region of the gene reveals a big deletion of 9.97 Kbp in the flat allele, affecting the 5'-UTR, the first exon, the exon and part of the second exon, which can be considered the causal mutation for the flat shape.

12. A phylogenetic analysis of the round allele protein of ppa025511m gene with full amino acid sequences of 35 LRR-RLK proteins with known biological function in Arabidopsis suggests a possible role in meristem development and fruit shape.

13. We have demonstrated that sport mutants can be successfully used in gene validation.

14. The PCR amplification of the INDELs in a round peach chimeric variety derived from a flat peach identified a possible mutation in the second meristem layer (producing the flesh) of the flat allele that would produce a reversion of the phenotype.

15. The mechanism of the flat allele is compatible with a negative dominant allele, where the mutant protein interacts with the wild one producing a haploinsufficiency. The function of the dominant negative allele would be truncated in the sport mutation recovering the wild round phenotype. However more analysis in the chimeric mutant will provide more clues of the mechanism of the gene in fruit shape patterning.

16. Standard sequence analysis pipelines may produce an overestimation of the variability that can be detected comparing the sequence of pairs of clones.

17. The variant calling filter applied, based on genotype Phred-Likelihood parameters, to remove false variants selected only those with very high genotypic quality. However this filter was too restrictive and eliminated true variants.

18. Somatic variants can occur only in some meristematic layers and thereafter leaves can harbour chimeric DNA. The analysis of such DNA sequences with standard pipelines will remove true variants occurring only in one layer and, thereafter, with low frequency.

19. The analysis of the G locus in peach varieties with heterozygous genotype (G/g) has shown lower nucleotide diversity and higher heterozygosity than the ones observed genome-wide.

The mutations occurred in the G locus had similar effects than the one occurring in other regions of the genome.

20. The comparison of the sequences between pair of clones identified candidate genes for the nectarine phenotype, some involved in trichome development however a deep molecular study is needed to elucidate their role in the expression of this trait.

21. The pipeline used here didn't allow for the detection of the long polymorphism of 7Kbp strongly associated with the nectarine phenotype (Vendramin *et al.,* 2014). Similarly we didn't detect such polymorphism with a specific software for structural variants (SV) calling, indicating that more coverage is needed for SV studies.

BIBLIOGRAPHY

Abel, S., Savchenko, T., & Levy, M. (2005). Genome-wide comparative analysis of the IQD gene families in *Arabidopsis thaliana* and *Oryza sativa*. *BMC Evolutionary Biology*, *5*, 72.

Abnizova, I., Leonard, S., Skelly, T., Brown, A., Jackon, D., Gourtovaia, M., Qi, G., Te Boekhorst , R ., Faruque, N., Lewis K., & Cox, T. (2012). Analysis of context-dependent error for Illumina sequencing. *Journal of Bioinformatics and Computational Biology*, *10*(02), 1241005.

Aggarwal, R K., Hendre, P. S., Varshney, R. K., Bhat, P. R., Krishnakumar, V., & Singh, L. (2007). Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species.*Theoretical and Applied Genetics.*, *114*(2), 359–372.

Ahmad, R., Parfitt D.E., Fass, J., Ogundiwin,E., Dhingra, A., Gradziel, T. M., Lin, D., Joshi, N. A., Martinez-Garcia, P. J., & Crisosto, C. H. "Whole genome sequencing of peach (*Prunus persica* L.) for SNP identification and selection," *BMC Genomics*, vol. 12, p. 569, 2011.

Albert, M., Jehle, A. K., Mueller, K., Eisele, C., Lipschis, M., & Felix, G. (2010). *Arabidopsis thaliana* pattern recognition receptors for bacterial elongation factor Tu and flagellin can be combined to form functional chimeric receptors. *The Journal of Biological Chemistry*, *285*(25), 19035–42.

Albrecht, C., Russinova, E., Hecht, V., Baaijens' E., & de Vries, S. (2005). The *Arabidopsis thaliana* somatic embryogenesis receptor-like kinsases 1 and 2 control male sporogenesis. *The Plant Cell*, *17*(12), 3337–3349.

Allen, A. S., Berkovic, S. F., Cossette, P., Delanty, N., Dlugos, D., Eichler, and others. (Allen, A. S., Epi4K Consortium., Epilepsy Phenome/Genome Project., Berkovic, S.F., Cossette, P., Delanty, N., Dlugos, D., Eichler, E. E., Epstein, M. P., Glauser, T., Goldstein, D. B., Han, Y., Heinzen, E. L., Hitomi, Y., Howell, K. B., Johnson, M. R., Kuzniecky, R., Lowenstein, D. H., Lu, Y. F., Madou, M. R., Marson, A. G., Mefford, H. C., Esmaeeli-Nieh, S., O'Brien, T. J., Ottman, R., Petrovski, S., Poduri, A., Ruzzo, E. K., Scheffer, I. E., Sherr, E. H., Yuskaitis, C. J., Abou-Khalil, B., Alldredge, B. K., Bautista, J. F., Berkovic, S. F., Boro, A., Cascino, G. D., Consalvo, D., Crumrine, P., Devinsky, O., Dlugos, D., Epstein, M. P., Fiol, M., Fountain, N. B., French, J., Friedman, D., Geller, E.B., Glauser, T., Glynn, S., Haut, S. R., Hayward, J., Helmers, S. L., Joshi, S., Kanner, A., Kirsch, H. E., Knowlton, R. C., Kossoff, E. H., Kuperman, R., Kuzniecky, R., Lowenstein, D. H., McGuire, S. M., Motika, P. V., Novotny, E. J., Ottman, R., Paolicchi, J. M., Parent, J. M, Park, K., Poduri, A., Scheffer, I. E., Shellhaas, R. A., Sherr, E. H., Shih, J. J., Singh, R., Sirven, J., Smith, M. C., Sullivan, J., Lin Thio, L., Venkat, A., Vining, E. P., Von Allmen, G. K., Weisenberg, J. L., Widdess-Walsh, P., & Winawer, M. R. 2013). De novo mutations in epileptic encephalopathies. *Nature*, *501*(7466), 217–21.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.

An, L., Lei, H., Shen, X., & Li, T. (2012). Identification and characterization of PpLFL, a homolog of Floricaula/Leafy in Peach (*Prunus persica*). *Plant Molecular Biology Reporter*, *30*(6), 1488–1495.

Andersen, J. R., & Lübberstedt, T. (2003). Functional markers in plants. *Trends in Plant Science*, *8*(11), 554–560.

Andrews, S. (2010). FastQC a quality control tool for high throughput sequence data.

Anithakumari, A M., Tang, J., van Eck, H. J., Visser, R. G. F., Leunissen, J. A. M., Vosman, B., & van der Linden, C. G. (2010). A pipeline for high throughput detection and mapping of SNPs from EST databases. *Molecular Breeding*, *26*(1), 65–75.

Aranzana, M.J., Illa, E., Howad, W., & Arus, P. (2012). A first insight into peach [*Prunus persica* (L.) Batsch] SNP variability. *Tree Genetics & Genomes*, *8*(6), 1359–1369.

Aranzana, M. J., Abbassi, E.-K., Howad, W., & Arús, P. (2010). Genetic variation, population structure and linkage disequilibrium in peach commercial varieties. *BMC Genetics*, *11*: 69.

Aranzana, M. J., Carbó, J., & Arús, P. (2003a). Microsatellite variability in peach [Prunus persica (L.) Batsch]: cultivar identification, marker mutation, pedigree inferences and population structure. *Theoretical and Applied Genetics*, *106*(8), 1341–1352.

Aranzana, M. J., Carbó, J., & Arús, P. (2003b). Using amplified fragment-length polymorphisms (AFLPs) to identify peach cultivars. *Journal of the American Society for Horticultural Science*, *128*(5), 672–677.

Aranzana, M. J., Garcia-Mas, J., Carbo, J., & Arus, P. (2002). Development and variability analysis of microsatellite markers in peach. *Plant Breeding*, *121*(1), 87–92.

Aranzana, M. J., Pineda, A., Cosson, P., Dirlewanger, E., Ascasibar, J., Cipriani, G., Ryder C.D., Testolin, R., Abbott, A., King, G.J., Iezzoni, A. F., & Arús, P. (2003). A set of simple-sequence repeat (SSR) markers covering the *Prunus* genome. *Theoretical and Applied Genetics*, *106*(5), 819–825.

Arulsekar, S., Parfitt, D. E., & Kester, D. E. (1986). Comparison of isozyme variability in peach and almond cultivars. *Journal of Heredity*, *77*(4), 272–274.

Arús, P., Verde, I., Sosinski, B., Zhebentyayeva, T., & Abbott, A. G. (2012). The peach genome. *Tree Genetics & Genomes*, *8*(3), 531–547.

Austin, R. S., Vidaurre, D., Stamatiou, G., Breit, R., Provart, N. J., Bonetta, D., Zhang, J., Fung, P., Gong, Y., Wang ,P. W., McCourt, P., & Guttman, D. S. (2011). Next-generation mapping of Arabidopsis genes. *The Plant Journal: For Cell and Molecular Biology*, *67*(4), 715–25.

Bailey, J. S., & French, A. P. (1942). The inheritance of blossom type and blossom size in peach. *Proceedings of the American Society for Horticultural Science*, (40), 248–250.

Bansal, V., Harismendy, O., & Tewhey, R. (2010). Accurate detection and genotyping of SNPs utilizing population sequencing data. *Genome Research*, 537–545.

Bassi, D., & Monet, R. (2008). Botany and taxonomy. *The Peach: Botany, Production and Uses*, 1–36.

Batley, J., Barker, G., O'Sullivan, H., Edwards, K. J., & Edwards, D. (2003). Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiology*, *132*(1), 84–91.

Baudry, A., Heim, M. a, Dubreucq, B., Caboche, M., Weisshaar, B., & Lepiniec, L. (2004). TT2, TT8, and TTG1 synergistically specify the expression of BANYULS and proanthocyanidin biosynthesis in *Arabidopsis thaliana*. *The Plant Journal: For Cell and Molecular Biology*, *39*(3), 366–380.

Becraft, P. W. (2002). Receptor kinase signaling in plant development. *Annual Review of Cell and Developmental Biology*, *18*, 163–192.

Bell, P. A., Chaturvedi, S., Gelfand, C. A., Huang, C. Y., Kochersperger, M., Kopla, R., Modica, F., Pohl, M., Varde, S., Zhao, R., Zhao, X., Boyce-Jacino, M.T., & Yassen A. (2002). SNPstream UHT: ultrahigh throughput SNP genotyping for pharmacogenomics and drug discovery. *BioTechniques*, (Supplemental)*74*, 70–72,74,76–77.

Bentley, D. R. (2006). Whole-genome re-sequencing. *Current Opinion in Genetics & Development*, *16*(6), 545–552.

Bielenberg, D. G., Wang, Y. E., Li, Z., Zhebentyayeva, T., Fan, S., Reighard, G. L., Gregory, L., Scorza, R., & Abbott, A. G. (2008). Sequencing and annotation of the evergrowing locus in peach [*Prunus persica* (L.) Batsch] reveals a cluster of six MADS-box transcription factors as candidate genes for regulation of terminal bud formation. *Tree Genetics & Genomes*, *4*(3), 495–507.

Birchler, J. A, & Veitia, R. A. (2010). The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *The New Phytologist*, *186*(1), 54–62.

Blake., M. A. (1932). The J. H. Hale peach as a parent in peach crosses. *Proceedings of the American Society for Horticultural Science*, (35), 49–53.

Blanca, J., Esteras, C., Ziarsolo, P., Pérez, D., Fernández-Pedrosa, V., Collado, C., Rodríguez de Pablos, R., Ballester, A., Roig, C., Cañizares, J., & Picó, B. (2012). Transcriptome sequencing for SNP discovery across *Cucumis melo*. *BMC Genomics*, *13*, 280.

Blenda, A. V, Verde, I., Georgi, L. L., Reighard, G. L., Forrest, S. D., Muñoz-Torres, M., & Baird, W. Abbott, A. G. (2007). Construction of a genetic linkage map and identification of molecular markers in peach rootstocks for response to peach tree short life syndrome. *Tree Genetics & Genomes*, *3*(4), 341–350.

Bliss, F. A., Arulsekar, S., Foolad, M. R., Becerra, V., Gillen, A. M., Warburton, M. L., Dandekar, A.M, Kocsisne, G. M & Mydin, K. K. (2002). An expanded genetic linkage map of *Prunus* based on an interspecific cross between almond and peach. *Genome*, *45*, 520–529.

Bonet, J., López-Girona, E., Sargent, D. J., Muñoz-Torres, M., Monfort, A., Abbott, A. G.., Arús, P., Simpson, D. W., & Davik, J. (2009). The development and characterisation of a bacterial artificial chromosome library for *Fragaria vesc*a. *BMC Research Notes*, *2*(1), 188.

Borsani, J., Budde, C. O., Porrini, L., Lauxmann, M. a, Lombardo, V. a, Murray, R., Andreo, C:S., Drincovich M. F., & Lara, M. V. (2009). Carbon metabolism of peach fruit after harvest: changes in enzymes involved in organic acid and sugar level modifications. *Journal of Experimental Botany*, *60*(6), 1823–1837.

Boudehri, K., Bendahmane, A., Cardinet, G. G., Troadec, C., Moing, A., & Dirlewanger, E. (2009). Phenotypic and fine genetic characterization of the *D* locus controlling fruit acidity in peach. *BMC Plant Biology*, *9*(1), 59.

Bouhadida, M., & Martín, J. (2007). Chloroplast DNA diversity in *Prunus* and its implication on genetic relationships. *Journal of the American Society for Horticultural Science*, *132*(5), 670–679.

Bowman, J. L., & Smyth, D. R. (1999). CRABS CLAW, a gene that regulates carpel and nectary development in Arabidopsis, encodes a novel protein with zinc finger and helix-loop-helix domains. *Development* , *126*(11), 2387–96.

Brady, S. M., Song, S., & Dhugga, K. S., Rafalski, J. A., & benfey, P. N (2007). Combining expression and comparative evolutionary analysis. The COBRA gene family. *Plant Physiology*, *143*(1), 172–187.

Brandi, F., Bar, E., Mourgues, F., Horvath, G., Turcsi, E., Giuliano, G., Liverani, A., Tartarini, S., Lewinsohn, E., & Rosati, C. (2011). Study of "Redhaven" peach and its white-fleshed mutant suggests a key role of CCD4 carotenoid dioxygenase in carotenoid and norisoprenoid volatile metabolism. *BMC Plant Biology*, *11*(1), 24.

Braun, A., Little, D. P., & Köster, H. (1997). Detecting CFTR gene mutations by using primer oligo base extension and mass spectrometry. *Clinical. Chemistry.*, *43*, 1151.

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., & Rolf, B. (1998). Mutation rate in human microsatellites: Influence of the structure and length of the tandem repeat. *The American Journal of Human Genetics*, *62*(6), 1408–1415.

Bronner, I., Quail, M., Turner., & D. J., Swerdlow, H (2014). Improved protocols for illumina sequencing. *Current Protocols in Human Genetics,* 79:18.2.1-18.2.42.

Brooks, A. J. (1999). The essence of SNPs. *Gene*, *234*(2), 177–186.

Broun, P. (2005). Transcriptional control of flavonoid biosynthesis: a complex network of conserved regulators involved in multiple aspects of differentiation in Arabidopsis. *Current Opinion in Plant Biology*, *8*(3), 272–279.

Buchner, P., & Boutin, J. P. (1998). A MADS box transcription factor of the AP1/AGL9 subfamily is also expressed in the seed coat of pea (*Pisum sativum*) during development. *Plant Molecular Biology*, *38*(6), 1253–1255.

Buschiazzo, E., & Gemmell, N. J. (2006). The rise, fall and renaissance of microsatellites in eukaryotic genomes. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *28*(10), 1040–1050.

Byrne, D. H. (1990). Isozyme variability in four diploid stone fruits compared with other woody perennial plants. *Journal of Heredity*, *81*(1), 68–71.

Byrne, D. H., Raseira, M. B., Bassi, D., Piagnani, M. C., Gasic, K., Reighard, G. L., Moreno, M. A., & Pérez, S. (2012). Peach. In M. L. Badenes & D. H. Byrne (Eds.), *Fruit Breeding*. Boston, MA: Springer US.

Byrne, D., Sherman, W., & Bacon, T. (2000). Stone fruit genetic pool and its exploitation for growing under warm winter conditions. In A. Erez (Ed.), *Temperate Fruit Crops in Warm Climates SE - 8* (pp. 157–230). Springer Netherlands.

Byrne, R. A., & McMahon, B. R. (1991). Acid-base and ionic regulation, during and following emersion, in the freshwater bivalve, *Anodonta grandis* simpsoniana (Bivalvia: Unionidae). *The Biological Bulletin*, *181*(2), 289–297.

Caño-Delgado, A., Yin, Y., Yu, C., Vafeados, D., Mora-García, S., Cheng, J.C., Nam, K. H., Li, J., & Chory, J. (2004). BRL1 and BRL3 are novel brassinosteroid receptors that function in vascular differentiation in *Arabidopsis*. *Development)*, *131*(21), 5341–5351.

Cao, K., Wang, L., Zhu, G., Fang, W., Chen, C., & Luo, J. (2012). Genetic diversity, linkage disequilibrium, and association mapping analyses of peach (*Prunus persica*) landraces in China. *Tree Genetics and Genomes*, *8*(5), 975–990.

Cao, K., Zheng, Z., Wang, L., Liu, X., Zhu, G., Fang, W., Cheng, S., Zeng, P., Chen, C., Wang, X., Xie, M., Zhong, X., Wang, X., Zhao, P., Bian, C., Zhu, Y., Zhang, J., Ma, G., Chen, C., Li, Y., Hao, F., Huang, G., Li, Y., Li, H., Guo, J., Xu, X., &Wang, J. (2014). Comparative population genomics reveals the domestication history of the peach, Prunus persica, and human influences on perennial fruit crops. *Genome Biology*, *15*(7), 415. 1

Carey, C. C., Strahle, J. T., Selinger, D. A., & Chandler, V. L. (2004). Mutations in the pale aleurone color1 regulatory gene of the Zea mays anthocyanin pathway have distinct phenotypes relative to the functionally similar TRANSPARENT TESTA GLABRA1 gene in *Arabidopsis thaliana*, *16* (February), 450–464.

Carrasco, B., Meisel, L., Gebauer, M., Garcia-Gonzales, R., & Silva, H. (2013). Breeding in peach, cherry and plum: from a tissue culture, genetic, transcriptomic and genomic perspective. *Biological Research*, *46*(3), 219–230.

Carrier, G., Le Cunff, L., Dereeper, A., Legrand, D., Sabot, F., Bouchez, O., Audeguin, L., & Boursiquot, J. M. (2012). Transposable elements are a major cause of somatic polymorphism in Vitis vinifera L. *PloS One*, *7*(3), e32973.

Causier, B., Castillo, R., Zhou, J., Ingram, R., Xue, Y., Schwarz-Sommer, Z., & Davies, B. (2005). Evolution in action: following function in duplicated floral homeotic genes. *Current Biology: CB*, *15*(16), 1508–12.

Cavagnaro, P. F., Senalik, D. a, Yang, L., Simon, P. W., Harkins, T. T., Kodira, C. D., Huang, S., & Weng, Y. (2010). Genome-wide characterization of simple sequence repeats in cucumber (*Cucumis sativus* L.). *BMC Genomics*, *11*(1), 569.

Celton, J.-M., Christoffels, A., Sargent, D., Xu, X., & Rees, D. J. (2010). Genome-wide SNP identification by high-throughput sequencing and selective mapping allows sequence assembly positioning using a framework genetic linkage map. *BMC Biology*, *8*(1), 155.

Chagné, D., Crowhurst, R. N., Troggio, M., Davey, M. W., Gilmore, B., Lawley, C., Vanderzande, S., Hellens, S. K., Cestaro, A., Velasco, R., Main, D., Rees, J. G., Iezzoni, A., Mockler, T., Wilhelm, L., Van de Weg, E., Gardiner, S.E., & Peace, C. (2012). Genome-Wide SNP detection, validation, and development of an 8K SNP array for apple. *PLoS ONE*, *7*(2), e31745.

Chagné, D., Gasic, K., Crowhurst, R. N., Han, Y., Bassett, H. C., Bowatte, D. R., Lawrence, T. J., Rikkerink, E. H., Gardiner, S. E., & Chagne, D. (2008). Development of a set of SNP markers present in expressed genes of the apple. *Genomics*, *92*(5), 353–358.

Chan, Z., Qin, G., Xu, X., Li, B., & Tian, S. (2007). Proteome approach to characterize proteins induced by antagonist yeast and salicylic acid in peach fruit. *Journal of Proteome Research*, *250*, 1677–1688.

Chaparro, J. X., Werner, D. J., O'Malley, D., & Sederoff, R. R. (1994). Targeted mapping and linkage analysis of morphological isozyme, and RAPD markers in peach. *Theoretical and Applied Genetics*, *87*(7), 805–815.

Chen, H., & Li, J. (2007). Nanotechnology. In J. Rampal (Ed.), *Microarrays SE - 22* (Vol. 381, pp. 411–436). Humana Press.

Chen, Z. L., Chen, W.J., Chen, H., Zhou, Y. Y., Tang, M. Q., Fu, M.Q., & Jin, X. F. (2013). *Prunus pananensis* (*Rosaceae*), a new species from Pan'an of central Zhejiang, China. *PloS One*, *8*(1), e54030.

Chin, S. W., Shaw, J., Haberle, R., Wen, J., & Potter, D. (2014). Diversification of almonds, peaches, plums and cherries - molecular systematics and biogeographic history of *Prunus* (*Rosaceae*). *Molecular Phylogenetics and Evolution*, *76*, 34–48.

Chusreeaeom, K., Ariizumi, T., Asamizu, E., Okabe, Y., Shirasawa, K., & Ezura, H. (2014). A novel tomato mutant, Solanum lycopersicum elongated fruit1 (Slelf1), exhibits an elongated fruit shape caused by increased cell layers in the proximal region of the ovary. *Molecular Genetics and Genomics: MGG*, *289*(3), 399–409.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effect of single nucleotide polymorphisms, SNPEff: SNPs in the genome of *Drosophila melanogaster* strain w (1118); iso-2; iso-3.*Fly* (Austin), 6(2), 80–92.

Cipriani, G., Lot, G., & Huang, W. G., Marrazzo, M. T., Peterlunger, E., Testolin, R. (1999). AC/GT and AG/CT microsatellite repeats in peach: isolation, characterisation and cross-species amplification in *Prunus*. *Theoretical and Applied Genetics*, 99 (1-2),65–72.

Clark, S. E., Williams, R. W., & Meyerowitz, E. M. (1997). The CLAVATA1 gene encodes a putative receptor kinase that controls shoot and floral meristem size in *Arabidopsis*. *Cell*, *89*(4), 575–585.

Clay, N., & Nelson, T. (2002). VH1, a provascular cell–specific receptor kinase that influences leaf cell patterns in *Arabidopsis*. *The Plant Cell*, *14*: 2707–2722.

Cock, J. M., Vanoosthuyse, V., & Gaude, T. (2002). Receptor kinase signalling in plants and animals: distinct molecular systems with mechanistic similarities. *Current Opinion in Cell Biology*, *14*(2), 230–236.

Cock, P. J. a, Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, *38*(6), 1767–1771.

Coenye, T., & Vandamme, P. (2005). Characterization of mononucleotide repeats in sequenced prokaryotic genomes. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, *12*(4), 221–233.

Cohen, S., Itkin, M., Yeselson, Y., Tzuri, G., Portnoy, V., Harel-Baja, R., Lev, S., Sa'ar,U., Davidovitz-Rikanati, R., Baranes, N., Bar, E., Wolf, D., Petreikov, M., Shen, S., Ben-Dor, S., Rogachev, I., Aharoni, A., Ast, T., Schuldiner, M., Belausov, E., Eshed, R., Ophir, R., Sherman, A., Frei, B., Neuhaus, H. E., Xu, Y., Fei, Z., Giovannoni, J., Lewinsohn, E., Tadmor, Y., Paris, H. S., Katzir, N., Burger, Y., & Schaffer, A. A.(2014). The PH gene determines fruit acidity and contributes to the evolution of sweet melons. *Nature Communications*, *5*.

Colcombet, J., Boisson-Dernier, A., Ros-Palau, R., Vera C. E., Schroeder, J. I. (2005). *Arabidopsis* SOMATIC EMBRYOGENESIS RECEPTOR KINASES 1 and 2 are essential for tapetum development and microspore maturation. *The Plant Cell*, 17:3350–3361.

Colombo, L., Battaglia, R., & Kater, M. M. (2008). Arabidopsis ovule development and its evolutionary conservation. *Trends in Plant Science*, *13*(8), 444–450.

Cong, B., Barrero, L. S., & Tanksley, S. D. (2008). Regulatory change in YABBY-like transcription factor led to evolution of extreme fruit size during tomato domestication. *Nature Genetics*, *40*(6), 800–4.

Connors, C. H. (1920). Some notes on the inheritance of unit characters in the peach. *Procidings of The American Society for Horticulticultural Science*, *16*, 24–36.

Conte, L., Della Strada, G., Fideghelli, C., Insero, O., Liverani, A., Moser, L., & Nicotra, A. (1994). Redhaven Bianca. *Monografia Di Cultivar Di Pesche, Nettarine E Percoche*, 94.

Creller, M. A., & Werner, D. J. (1996). Characterizing the novel fruit surface morphology of 'Marina' peach using ecanning electron microscopy. *Journal of the American Society for Horticultural Science*, *121*(2), 198–203.

Cullinan, F. P. (1937). Improvement of stone fruits. In *United States Department of Agriculture Yearbook of Agriculture.* (pp. 665–748). Washington.

Da Maia, L. C., Palmieri, D. A., de Souza, V. Q., Kopp, M. M., de Carvalho, F. I. F., & Costa de Oliveira, A. (2008). SSR locator: tool for Simple Sequence Repeat discovery integrated with primer design and PCR simulation. *International Journal of Plant Genomics*, 4: 363-374..

Dabov., S. (1983). Inheritance of peach resistance to powdery mildew III: Leaf resistance in F1 of 'J. H. Hale' × 'nectarine Ferganensis-2.' In *Genetics and Plant Breeding*, 16:146–150).

Dagar, A., Weksler, A., Friedman, H., Ogundiwin, A. E., Crisosto, C. H., Ahmad, R., & Lurie, S. (2011). Comparing ripening and storage characteristics of 'Oded' peach and its nectarine mutant 'Yuval'. *Postharvest Biology and Technology*, *60*(1), 1–6.

Dantec, L. Le, Chagné, D., Pot, D., Cantin, O., Garnier-Géré, P., Bedon, F., Frigerio, J. M., Caumeil, P., Léger, P., García, V., Laigret, F., de Daruvar, A., & Plomion, C. (2004). Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences. *Plant Molecular Biology*, *54*(3), 461–470.

de Smet, I., Voss, U., Jürgens, G., & Beeckman, T. (2009). Receptor-like kinases shape the plant. *Nature Cell Biology*, *11*(10), 1166–73.

de Vetten, N., Quattrocchio, F., Mol, J., & Koes, R. (1997). The an11 locus controlling flower pigmentation in petunia encodes a novel WD-repeat protein conserved in yeast, plants, and animals. *Genes & Development*, *11*(11), 1422–1434.

Deeken, R., & Kaldenhoff, R. (1997). Light-repressible receptor protein kinase: a novel photo-regulated gene from *Arabidopsis thaliana*. *Planta*, *202*(4), 479–486.

DePristo, M. A, Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., Philippakis, A.cA., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–8.

Dermen, H. (1953). Periclinal cytochimeras and origin of tissues in stem and leaf of peach. *American Journal of Botany*, *40*(3), 154–168.

Dermen, H. (1956). Histogenic factors in color and fuzzless peach sports. *Journal of Heredity*, *47*(63:76).

Dermen, H., Stewart, R. N. (1972). Ontogenetic study of floral organs of peach (*Prunus persica*) utilizing cytochimeral plants. *American Journal of Botany*, *60*, 283–291.

Dettori, M. T., Quarta, R., & Verde, I. (2001). A peach linkage map integrating RFLPs, SSRs, RAPDs, and morphological markers. *Genome*, *44*(5), 783–790.

DeYoung, B. J., Bickle, K. L., Schrage, K. J., Muskett, P., Patel, K., & Clark, S. E. (2006). The CLAVATA1-related BAM1, BAM2 and BAM3 receptor kinase-like proteins are required for meristem function in *Arabidopsis*. *The Plant Journal: For Cell and Molecular Biology*, *45*(1), 1–16.

Deyoung, B. J., & Clark, S. E. (2008). BAM receptors regulate stem cell specification and organ development through complex interactions with CLAVATA signaling. *Genetics*, *180*(2), 895–904.

Dhanapal, A. P., Martínez-García, P. J., Gradziel, T. M., & Crisosto, C. H. (2012). First genetic linkage map of chilling injury susceptibility in peach (*Prunus persica* (L.) Batsch) fruit with SSR and SNP markers. *Journal of Plant Science and Molecular Breeding*, *1*(1), 3.

Diévart, A., Dalal, M., & Tax, F.E., Lacey, A.D., Huttly, A., Li, J., & Clarck, S.E. (2003). CLAVATA1 dominant-negative alleles reveal functional overlap between multiple receptor kinases that regulate meristem and organ development. *The Plant Cell*, *15*(5), 1198–1211.

Dirlewanger, E., & Bodo, C. (1994). Molecular genetic mapping of peach. *Euphytica*, *77*(1-2), 101–103.

Dirlewanger, E., Cosson, P., Boudehri, K., Renaud, C., Capdeville, G., Tauzin, Y., Laigret, F., & Moing, A. (2006). Development of a second-generation genetic linkage map for peach [*Prunus persica* (L.) Batsch] and characterization of morphological traits affecting flower and fruit. *Tree Genetics & Genomes*, *3*(1), 1–13.

Dirlewanger, E., Cosson, P., Tavaud, M., Aranzana, M. J., Poizat, C., Zanetto, A., Arús, P., & Laigret, F. (2002). Development of microsatellite markers in peach [*Prunus persica* (L.) Batsch] and their use in genetic diversity analysis in peach and sweet cherry (*Prunus avium* L.). *Theoretical and Applied Genetics*, *105*(1), 127–138.

Dirlewanger, E., Graziano, E., Joobeur, T., Garriga-Caldere, F., Cosson, P., Howad, W., & Arús, P. (2004). Comparative mapping and marker-assisted selection in Rosaceae fruit crops. *Proceedings of The National Academy of Sciences USA*, *101*(26), 9891–9896.

Dirlewanger, E., Pronier, V., Parvery, C., Rothan, C., Guye, A., & Monet, R. (1998). Genetic linkage map of peach [*Prunus persica* (L.) Batsch] using morphological and molecular markers. *Theoretical and Applied Genetics*, *97*(5-6), 888–895.

Dodsworth, S. (2009). A diverse and intricate signalling network regulates stem cell fate in the shoot apical meristem. *Developmental Biology*, *336*(1), 1–9.

Doyle, J. J., Doyle, J. L. (1987). A rapid DNA isolation procedure from small quantities of fresh leaf tissues. *Phytochemistry Bulletin*, *19*, 11–15.

Durtschi, J., Margraf, R. L., Coonrod, E. M., Mallempati, K. C., & Voelkerding, K. V. (2013). VarBin, a novel method for classifying true and false positive variants in NGS data. *BMC Bioinformatics*, *14 Suppl 1*(Suppl 13), S2.

Edwards, D., Forster, J., Chagné, D., & Batley, J. (2007). What are SNPs? In N. Oraguzie, E. A. Rikkerink, S. Gardiner, & H. N. Silva (Eds.), *Association Mapping in Plants SE - 3* (pp. 41–52). Springer New York.

Edwards K. J, Barker J. H, Daly A, Jones C, Karp. A. (1996). Microsatellite libraries enriched for several microsatellite sequences in plants. *Biotechniques*, *20(5)*, 758–760.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., & Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., de Winter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holde, D., kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Roy, J., Sebra, R ; Shen, G; Sorenson, J ., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J ., Wu, D ., Yang, A ., Zaccarin, D ., Zhao, P., Zhong, F ., Korlach, J., & Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323 (5910), 133-138*.

English, A. C., Salerno, W. J., & Reid, J. G. (2014). PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics*, *15*, 180.

Etienne, Rothan, Moing, Plomion, Bodénès, Svanella-Dumas, L., Cosson, P., Pronier, V., Monet, R., & Dirlewanger, E. (2002). Candidate genes and QTLs for sugar and organic acid content in peach [*Prunus persica* (L.) Batsch].*Theoretical and Applied Genetics*, *105*(1), 145–159.

Ewing, B., Hillier, L. D., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, *8*(3), 175–185.

Eyüboglu, B., Pfister, K., Haberer, G., Chevalier, D., Fuchs, A., Mayer, K. F. X., & Schneitz, K. (2007). Molecular characterisation of the STRUBBELIG-RECEPTOR FAMILY of genes encoding putative leucine-rich repeat receptor-like kinases in *Arabidopsis thaliana*. *BMC Plant Biology*, *7*, 16.

Fairchild, D. (1938). Charles Scribner's Sons. In *The World Was My Garden* (p. 226).

Falchi, R., Vendramin, E., Zanon, L., Scalabrin, S., Cipriani, G., Verde, I., Vizzotto, G., Morgante, M. (2013). Three distinct mutational mechanisms acting on a single gene underpin the origin of yellow flesh in peach. *The Plant Journal: For Cell and Molecular Biology*,76(2), 175–187.

Fan, S., Bielenberg, D. G., Zhebentyayeva, T. N., Reighard, G. L., Okie, W. R., Holland, D., & Abbott, A. G. (2010). Mapping quantitative trait loci associated with chilling requirement, heat requirement and bloom date in peach (*Prunus persica*). *New Phytologist*, *185*(4), 917–930.

FAOSTAT. (2014). FAOSTAT. http//:www.faostat.fao.org.

Felcher, K. J., Coombs, J. J., Massa, A. N., Hansey, C. N., Hamilton, J. P., Veilleux, R. E., Buell, C. R., Douches, D. S. (2012). Integration of two diploid potato linkage maps with the potato genome sequence. *PLoS ONE*, *7*(4), e36347.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, *39*(4), 783–791.

Ferrándiz, C., Pelaz, S., Yanofsky, M. F.(1999). Control of carpel and fruit development in *Arabidopsis*. *Annual Review of Biochemistry*, 68: 321–354.

Fisher, K., & Turner, S. (2007). PXY, a receptor-like kinase essential for maintaining polarity during plant vascular-tissue development. *Current Biology: CB*, *17*(12), 1061–1066.

Fontes, E. P. B ., Santos, A. A., Luz, D. F., Waclawovsky, A. J., & Chory, J. (2004). The geminivirus nuclear shuttle protein is a virulence factor that suppresses transmembrane receptor kinase activity. *Genes &Development*,18:2545–2556.

Foolad, M. R., Arulsekar, S., Becerra, V., & Bliss, F. A. (1995). A genetic map of Prunus based on an interspecific cross between peach and almond. *Theoretical and Applied Genetics*, *91*(2), 262–269.

FresnedoRamírez, J., Martínez-García, P. J., Parfitt, D. E., Crisosto, C. H., & Gradziel, T. M. (2013). Heterogeneity in the entire genome for three genotypes of peach [*Prunus persica* (L .) Batsch] as distinguished from sequence analysis of genomic variants. *BMC Genomics, 14:750.*

Ganal, M. W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E. S., Charcosset, A., Clarke, J.D., Graner E. M., Hansen, M., Joets, J., Le Paslier, M.C., McMullen, M. D., Montalent, P., Rose, M., Schön, C.C., Sun, Q., Walter, H., Martin, O.C., Falque, M. (2011). A Large Maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS ONE*, *6*(12), e28334.

Gao, M., Wang, X., Wang, D., Xu, F., Ding, X., Zhang, Z., Dongling, Bi, Cheng J.T., Chen, S., Li, X., & Zhang, Y. (2009). Regulation of cell death and innate immunity by two receptor-like kinases in *Arabidopsis*. *Cell Host & Microbe*, *6*(1), 34–44.

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S.,Dopazo, J., Meyer, T. F., & Conesa, A. (2012). Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics (Oxford, England)*, *28*(20), 2678–2679.

Gaur, R., Azam, S., Jeena, G., Khan, A. W., Choudhary, S., Jain, M., Yadav, G., Tyagi, A. K., Chattopadhyay, D., & Bhatia, S. (2012). High-throughput SNP discovery and genotyping for constructing a saturated linkage map of chickpea (*Cicer arietinum* L.). *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, *19*(5), 357–373.

Geier, T. (2012). Chimeras: properties and dissociation in vegetatively propagated plants. In Q. Shu (Ed.), *Plant Mutation Breeding and Biotechnology* (p. 616). Japan: International Atomic Energy Agency.

Georgi, L. L., Wang, Y., Yvergniaux, D., Ormsbee, T., Inigo, M., Reighard, G., & Abbott, A. G. (2002). Construction of a BAC library and its application to the identification of simple sequence repeats in peach [*Prunus persica* (L.) Batsch]. *Theoretical and Applied Genetics*, *105*, 1151–1158.

Gilchrist, E., & Haughn, G. (2010). Reverse genetics techniques: engineering loss and gain of gene function in plants. *Briefings in Functional Genomics*, *9*(2), 103–10.

Gill, D. E., Chao, L., Perkins, S. L., & Wolf, J. B. (1995). Genetic mosaicism in plants and clonal animals. *Annual Review of Ecology and Systematics*, *26*(1), 423–444.

Gillaspy, G., Ben-David, H., & Gruissem, W. (1993). Fruits: A Developmental Perspective. *The Plant Cell*, *5*(10), 1439–1451.

Giovannoni, J. J. (2004). Genetic Regulation of Fruit Development and Ripening, Plant Cell, 16**:** S170–S180.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, *11*(5), 759–769.

Gómez-Gómez, L., & Boller, T. (2000). FLS2: an LRR receptor-like kinase involved in the perception of the bacterial elicitor flagellin in *Arabidopsis*. *Molecular Cell*, *5*(6), 1003–1011.

Goode, D. L., Cooper, G. M., Schmutz, J., Dickson, M., Gonzales, E., Tsai, M., Karra, K., Davydov, E., Batzoglou, S., Myers, R. M., & Sidow, A. (2010). Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. *Genome Research*, *20*(3), 301–310.

Gorlov, I. P., Kimmel, M., & Amos, C. I. (2006). Strength of the purifying selection against different categories of the point mutations in the coding regions of the human genome. *Human Molecular Genetics*, *15*(7), 1143–1150.

Gou, X., He, K., Yang, H., Yuan, T., Lin, H., Clouse, S. D., & Li, J. (2010). Genome-wide cloning and sequence analysis of leucine-rich repeat receptor-like protein kinase genes in *Arabidopsis thaliana*. *BMC Genomics*, *11*:19.

Gramzow, L., Ritz, M. S., & Theissen, G. (2010). On the origin of MADS-domain transcription factors. *Trends in Genetics: TIG*, *26*(4), 149–153.

Groenen, M., Megens, H.-J., Zare, Y., Warren, W., Hillier, L., Crooijmans, R. P. M. A., Vereijken, A., Okimoto, R., Muir, W. M., & Cheng, H. H. (2011). The development and characterization of a 60K SNP chip for chicken. *BMC Genomics*, *12*(1), 274.

Gu, Q., Ferrándiz, C., Yanofsky, M. F., & Martienssen, R. (1998). The FRUITFULL MADS-box gene mediates cell differentiation during *Arabidopsis* fruit development. *Development* , *125*(8), 1509–17.

Guo, Y. L., Fitz, J., Schneeberger, K., Ossowski, S., Cao, J., & Weigel, D. (2011). Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in *Arabidopsis*. *Plant Physiology*, *157*(2), 757–69.

Gupta, P. K., Rustgi, S., & Mir, R. R. (2008). Array-based high-throughput DNA markers for crop improvement. *Heredity*, *101*(1), 5–18.

Gupta, P. K., Rustgi, S., & Mir, R. R. (2013). *Cereal Genomics II*. (P. K. Gupta & R. K. Varshney, Eds.) (pp. 11–56). Dordrecht: Springer Netherlands.

Hackbusch, J., Richter, K., Müller, J., Salamini, F., & Uhrig, J. F. (2005). A central role of *Arabidopsis thaliana* ovate family proteins in networking and subcellular localization of 3-aa loop extension homeodomain proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(13), 4908–12.

Hammerschlag, F. A., Bauchan, G., & Scorza, R. (1985). Regeneration of peach plants from callus derived from immature embryos. *Theoretical and Applied Genetics*, *70*(3), 248–251.

Hammerschlag, F. A., Owens, L. D., & Smigocki, A. C. (1989). Agrobacterium-mediated transformation of peach cells derived from mature plants that were propagated in vitro. *Journal of American Society for Horticultural Sciencies*, *114*(3), 508–510.

Han, J., Wang, W. Y., Leng, X. P., Guo, L., Yu, M. L., Jiang, W. B., & Ma, R. J. (2014). Efficient identification of ornamental peach cultivars using RAPD markers with a manual cultivar identification diagram strategy. *Genetics and Molecular Research*, *13*(1), 32–42.

Hansen, M, Kraft, T, Ganestam, S, Säll, T, Nilsson, N, O. (2001). Linkage disequillibrium mapping of the bolting gene in sea beet using AFLP markers. *Genetical Research*, *77*(1), 61–66.

Hardin, S. H. (2008). Real-Time DNA Sequencing. In *Next Generation Genome Sequencing* (pp. 95–101). Wiley-VCH Verlag GmbH & Co. KGaA.

Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., DiMeo, J., Efcavitch, J. W., Giladi, E., Gill, J., Healy, J., Jarosz, M., Lapen, D., Moulton, K., Quake, S. R., Steinmann, K., Thayer, E., Tyurina, A., Ward, R; Weiss, H., & Xie, Z. (2008). Single-molecule DNA sequencing of a viral genome. *Science* , *320*(5872), 106–9.

Hartmann, H. T., & Kester, D. E. (1975). *Plant propagation*. *Pren*. New Jersey: Prentice-Hall.

He, K., Gou, X., Yuan, T., Lin, H., Asami, T., Yoshida, S., Russell, S. D., & Li, J. (2007). BAK1 and BKK1 regulate brassinosteroid-dependent growth and brassinosteroid-independent cell-death pathways. *Current Biology: CB*, *17*(13), 1109–1115.

Hedrick, U. P., Howe, G. H., Morehouse, T., & Burton, T. C. (1917). *The peaches of New York, by U.P. Hedrick, assisted by G.H. Howe, O.M. Taylor [and] C.B. Turbergen.* (p. 748). Albany,J. B. Lyon Company, printers.

Heller, M. J. (2002). DNA microarray technology: devices, systems, and applications. *Annual Review of Biomedical Engineering*, *4*(1), 129–153.

Hendre, P. S., Kamalakannan, R., & Varghese, M. (2012). High-throughput and parallel SNP discovery in selected candidate genes in *Eucalyptus camaldulensis* using Illumina NGS platform. *Plant Biotechnology Journal*, *10*(6), 646–656.

Hirakawa, Y., Shinohara, H., Kondo, Y., Inoue, A., Nakanomyo, I., Ogawa, M., Sawa, S., Ohashi-Ito, K., Matsubayashi, Y., & Fukuda, H. (2008). Non-cell-autonomous control of vascular stem cell fate by a CLE peptide/receptor system. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(39), 15208–15213.

Hong, S. W., Jon, J. H., Kwak, J. M., & Nam, H. G. (1997). Identification of a receptor-like protein kinase gene rapidly induced by abscisic acid, dehydration, high salt, and cold treatments in *Arabidopsis thaliana*. *Plant Physiology*, *113*(4), 1203–1212.

Hord, C. L. H., Chen, C., & DeYoung, B. J. (2006). The BAM1/BAM2 receptor-like kinases are important regulators of Arabidopsis early anther development. *The Plant Cell*, *18*(7), 1667–1680.

Horn, R., Lecouls, A. C., Callahan, A., Dandekar, A., Garay, L., McCord, P., Howard, W., Chan, H., Derde, I., Main, D., Jung, S., Georgi, L., Forrest, S., Mook, J., Zhebentyayeva, T., Yu, Y., Kim, H. R., Jesudurai, C., Sosinski, B., Arús, P., Baird, V., Parfitt, D., Reighard, G., Scorza, R., Tomkins, J., Wing, R., & Abbott, A. G. (2005). Candidate gene database and transcript map for peach, a model species for fruit trees. *Theoretical and Applied Genetics*, *110*, 1419–1428.

Howard, H. W. (1970). *Genetics of the potato* (Berlin Hei.). New York: Springer.

Huang, X., & Han, B. (2014). Natural variations and genome-wide association studies in crop plants. *Annual Review of Plant Biology*, *65*, 531–51.

Humphries, J., Walker, A., Timmis, J., & Orford, S. (2005). Two WD-repeat genes from cotton are functional homologues of the *Arabidopsis thaliana* TRANSPARENT TESTA GLABRA1 (TTG1) gene. *Plant Molecular Biology*, *57*(1), 67–81.

Hunkapiller, T., Kaiser, R. J., Koop, B. F., & Hood, L. (1991). Large-scale and automated DNA sequence determination. *Science*, *254*(5028), 59–67.

Iglesias, I. (2009). *Melocotón plano y nectarina plana las variedades de mayor interés*. (IRTA, Ed.) (p. 134).

Iglesias, I., & Casals, E. (2013). Exportación de melocotón en España. *Vida Rural*, *357*.

Illa, E., Sargent, D. J., López Girona E., Bushakra, J., Cestaro, A., Crowhurst, R., Pindo, M., Cabrera, A., Van der Knapp, E., Iezzoni, A., Gardiner, S., Velasco. R., Arús, P., Chagné, D., & Troggio, M. (2011) Comparative analysis of rosaceous genomes and the reconstruction of a putative ancestral genome for the family. *BMC Evol. Biol.* 11:9

Ioannou, D., & Griffin, D. K. (2010). Nanotechnology and molecular cytogenetics: the future has not yet arrived. *Nano Reviews*, *1*, 1–14.

Jauregui., B. (1998). *Identification of molecular markers linked to agronomic characters in an interspecific almond x peach progeny*. *University of Barcelona*. Barcelona.

Jelenkovic, G., & Harrington, E. (1972). Morphology of the pachytene chromosomes in *Prunus persica*. *Canadian Journal of Genetics and Cytology*, *14*(2), 317–324.

Jia, P., Li, F., Xia, J., Chen, H., Ji, H., Pao, W., & Zhao, Z. (2012). Consensus rules in variant detection from next-generation sequencing data. *PloS One*, *7*(6), e38470.

Jinn, T. L., Stone, J. M., & Walker, J. C. (2000). HAESA, an *Arabidopsis* leucine-rich repeat receptor kinase, controls floral organ abscission. *Genes & Development*,14(1), 108–117.

Johnson, A., Trumbower, H., & Sadee, W. (2011). RNA structures affected by single nucleotide polymorphisms in transcribed regions of the human genome. *Webmed Central BIOINFORMATICS*. 2(2):WMC001600

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*, *8*(3), 275–282.

Joobeur, T., Viruel, M. A. A., de Vicente, M. C., Jauregui, B., Ballester, J., Dettori, M. T., Verde, I., Truco, M. J, Messeguer, R., Battle, I., Quarta, R., Dirlewanger, E., & Arus, P. (1998). Construction of a saturated linkage map for *Prunus* using an almond x peach F_2 progeny. *Theoretical and Applied Genetics*, *97*(7), 1034–1041.

Jung, S., Ficklin, S. P., Lee, T., Cheng, C.-H., Blenda, A., Zheng, P., Yu, J., Bombarely, A., Cho, I., Ru, S., Evans, K., Peace, C., Abbott, A., Mueller, L. A., Olmstead, M. A., & Main, D. (2013). The Genome Database for Rosaceae (GDR): year 10 update. *Nucleic Acids Research, 42 (D1): D1237-D1244.*

Jung, S., Staton, M., Lee, T., Blenda, A., Svancara, R., Abbott, A., & Main, D. (2008). GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. *Nucleic Acids Research.*, *36*,(D1): D1034–1040.

Kantety, R., La Rota, M., Matthews, D., Sorrells, M. (2002).Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Molecular Biology,* 48(5):501–510.

Karlova, R., Boeren, S., & Russinova, E., Aker, J., Vervoort, J., & de Vries, S. (2006). The *Arabidopsis* somatic embryogenesis receptor-like kinase1 protein complex includes brassinosteroid-insensitive1. *The Plant Cell*, *18*(4), 626–638.

Karrow, J. (2010). Life Tech Details Real-Time Single-Molecule Tech at AGBT; Combines Qdots with FRET-Based Detection. *Genome Web in Sequence*. http://www.genomeweb.com

Kaya, H. B., Cetin, O., Kaya, H., Sahin, M., Sefer, F., Kahraman, A., & Tanyolac, B. (2013). SNP Discovery by Illumina-based transcriptome sequencing of the olive and the genetic characterization of Turkish olive genotypes revealed by AFLP, SSR and SNP markers. *PloS One*, *8*(9), e73674.

Kelkar, Y., & Tyekucheva, S., Chiaromonte, F., & Makova, K.D. (2008). The genome-wide determinants of human and chimpanzee microsatellite evolution. *GenomeResearch*,18(1): 30–38.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutation. *Genetics*, *61*(694), 893–903.

Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W., & Vogelstein, B. (2011). Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(23), 9530–5.

Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing concepts and limitations. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *32*(6), 524–536.

Kolpakov, R., Bana, G., & Kucherov, G. (2003). mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Research*, *31*(13), 3672–3678.

Konishi Iwahori, S., Kitagawa, H., Yakuwa, T., Organizing Committee XXIVth International Horticultural Congress., K. (1994). Horticulture in Japan. Tokyo: Asakura Pub. Co.

Kumar, S., You, F. M., & Cloutier, S. (2012). Genome wide SNP discovery in flax through next generation sequencing of reduced representation libraries. *BMC Genomics*, *13*(1), 684.

Lander, E. S., Consortium, I. H. G. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R.,Sheridan,A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston,J ., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French,L ., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, R., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, T. A., Pepin, K. H., Gish, W. R.,

Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L ., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S, W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E. J., Worley, K.C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L. Kucherlapati, R. S.,Nelson, D. L., Weinstock, J. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A.,Itoh, A., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, P., Nyakatura, G., Taudien, S., Rump, A., Yang, H. M., Yu, J., Wang,J., Huang, G. Y., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S. A., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R.M., Schmutz,J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M. v., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H.Q., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K Nordsiek, N., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C.B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D.,Hermjakob, H., Hokamp, K., Jang, W. H., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk,A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght,A., Mikkelsen, T., Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. R. Schultz, G. Slater, A. F. A. Smit, J. V., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R.F ., Collins, R., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., & Morgan, M. J. "Initial sequencing and analysis of the human genome. (2001)" *Nature*, vol. 409, no. 6822, pp. 860–921.

Lara, M. V, Borsani, J., Budde, C. O., Lauxmann, M. A, Lombardo, V. A, Murray, R., Andreo, C. S., & Drincovich, M. F. (2009). Biochemical and proteomic analysis of "Dixiland" peach fruit (*Prunus persica*) upon heat treatment. *Journal of Experimental Botany*, *60*(15), 4315–4333.

Lassmann, T., Hayashizaki, Y., & Daub, C. O. (2011). SAMStat: monitoring biases in next generation sequencing data. *Bioinformatics)*, *27*(1), 130–131.

Lee, J., Hassan, O. S., Gao, W., Wei, N., Kohel, R., Chen, X. Y., Payton, P., Sze, S. H., Stelly, D. M., & Chen, Z. J. (2006). Developmental and gene expression analyses of a cotton naked seed mutant. *Planta*, *223*(3), 418–432.

Lesley, J. W. (1939). A genetic study of saucer fruit shape and other characters in the peach. *Proceedings of the American Society for Horticultural Science*, *38*, 222.

Li, H. (2011a). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics,*, *27*(21), 2987–2993.

Li, H. (2011b). Improving SNP discovery by base alignment quality. *Bioinformatics,*, *27*(8), 1157–8.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760.

Li, J., & Chory, J. (1997). A putative leucine-rich repeat receptor kinase involved in brassinosteroid signal transduction. *Cell*, *90*(5), 929–938.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Li, J., Wen, J., Lease, K. A, Doke, J. T., Tax, F. E., & Walker, J. C. (2002). BAK1, an *Arabidopsis* LRR Receptor-like Protein Kinase, Interacts with BRI1 and Modulates Brassinosteroid Signaling. *Cell*, *110*(2), 213–222.

Li, W., & Sadler, L. (1991). Low nucleotide diversity in man. *Genetics*. 129(2):513-23.

Li, X., Meng, X., Jia, H., Yu, M.., Ma, R., Wang, L., Cao, K., Shen, Z. J., Nie, L., Tian, J. B., Chen, M. J., Xie, M., Arús,P., Gao, Z. S., & Aranzana, M. J. (2013). Peach genetic resources: diversity, population structure and linkage disequilibrium. *BMC Genetics*, *14*(1), 84.

Liljegren, S. J., Ditta, G. S., Eshed, Y., Savidge, B., Bowman, J. L., & Yanofsky, M. F. (2000). SHATTERPROOF MADS-box genes control seed dispersal in Arabidopsis. *Nature*, *404*(6779), 766–770.

Liu, J., Van Eck, J., Cong, B., & Tanksley, S. D. (2002). A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(20), 13302–6.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., & Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine & Biotechnology*, *2012*, 251364.

Liu, X., & Pijut, P. M. (2009). Agrobacterium-mediated transformation of mature *Prunus serotina* (black cherry) and regeneration of transgenic shoots. *Plant Cell, Tissue and Organ Culture (PCTOC)*, *101*(1), 49–57.

Llácer, G. (2005). *Situación actual del cultivo del melocotonero*. *Tendencias* (pp. 133–141). Agrícola Vergel.

Lloyd, a M., Walbot, V., & Davis, R. W. (1992). *Arabidopsis* and *Nicotiana* anthocyanin production activated by maize regulators R and C1. *Science*, *258*(5089), 1773–1775.

Lu, C., Wu, W., Xiao, J., Meng, Y., Zhang, S., & Zhang, X. (2013). Detection of pathogenic mutations in Marfan syndrome by targeted next-generation semiconductor sequencing. *Chinese Journal of Medical Genetics*, *30*(3), 301–304.

Lu, Z.-X., Sosinski, B., Reighard, G. L., Baird, W. V, & Abbott, A. G. (1998). Construction of a genetic linkage map and identification of AFLP markers for resistance to root-knot nematodes in peach rootstocks. *Genome*, *41*(2), 199–207.

Lukowitz, W., Gillmor, C. S., & Scheible, W. R. (2000). Positional cloning in *Arabidopsis*. Why it feels good to have a genome initiative working for you. *Plant Physiology*, *123*(3), 795–805.

Luo, M., Dennis, E. S,. Berger, F., Peacock, W.J., & Chaudhury, A. (2005). MINISEED3 (MINI3), a WRKY family gene, and HAIKU2 (IKU2), a leucine-rich repeat (LRR) KINASE gene, are regulators of seed size in *Arabidopsis. Proceedings of the National Academy of Sciences of The United States of America*, *102*(48), 17531–17536.

Machado, A., Wu, Y., Yang, Y., Llewellyn, D. J., & Dennis, E. S. (2009). The MYB transcription factor GhMYB25 regulates early fibre and trichome development. *The Plant Journal: For Cell and Molecular Biology*, *59*(1), 52–62.

Madabhushi, R. S. (1998). Separation of 4-color DNA sequencing extension products in noncovalently coated capillaries using low viscosity polymer solutions. *Electrophoresis*, *19*(2), 224–230.

Mammadov, J., Aggarwal, R., Buyyarapu, R., & Kumpatla, S. (2012). SNP markers and their impact on plant breeding. *International Journal of Plant Genomics*, 11 pages.

Mandel, T., Moreau, F., Kutsher, Y., Fletcher, J. C., Carles, C. C., & Eshed Williams, L. (2014). The ERECTA receptor kinase regulates Arabidopsis shoot apical meristem size, phyllotaxy and floral meristem identity. *Development*, *141*(4), 830–41.

Mante, S., Scorza, R., & Cordts, J. (1989). Plant regeneration from cotyledons of *Prunus persica*, *Prunus domestica*, and *Prunus cerasus*. *Plant Cell, Tissue and Organ Culture*, *19*(1), 1–11.

Margraf, R. L., Durtschi, J. D., Dames, S., Pattison, D. C., Stephens, J. E., Mao, R., & Voelkerding, K. V. (2010). Multi-sample pooling and illumina genome analyzer sequencing methods to determine gene sequence variation for database development. *Journal of Biomolecular Techniques: JBT*, *21*(3), 126–40.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J, Michael, E., Braverman, S., Chen, Y. J., Chen, Z., Dewell, S.B., Du, L, Fierro, J.M., Gomes, X. V, Godwin, B.C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvi, T. P., Jirage, K. B., Kim, J.B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade′ K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J .F., Simpson, J. W., Srinivasan, M., Tartaro, K. R.,Tomasz, A., Vogt, K. A.,Volkmer, G. A.,Wang, S. H.,Wang, Y., Weiner, M. P., Yu, P., Begley′ R. F., & Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380.

Martín, B., Ramiro, M., Martínez-Zapater, J. M., & Alonso-Blanco, C. (2009). A high-density collection of EMS-induced mutations for TILLING in Landsberg erecta genetic background of Arabidopsis. *BMC Plant Biology*, *9*, 147.

Martínez-García, P., Fresnedo-Ramírez, J., Parfitt, D., Gradziel, T., & Crisosto, C. (2013). Effect prediction of identified SNPs linked to fruit quality and chilling injury in peach [*Prunus persica* (L.) Batsch]. *Plant Molecular Biology*, *81*(1-2), 161–174.

Martins, W. S., César, D., Lucas, S., Fabricio, K., Neves, D. S., & John, D. (2009). WebSat: A web software for microsatellite marker development Bioinformation. *Bioinformation*, *3*(6), 282–283.

Mase, N., Iketani, H., & Sato, Y. (2007). Analysis of bud sport cultivars of peach (*Prunus persica* (L.) batsch) by simple sequence repeats (SSR) and restriction landmark genomic scanning (RLGS). *Journal of the Japanese Society for Horticultural Science*, *76*(1), 20–27.

Matsushima, N., & Miyashita, H. (2012). Leucine-Rich Repeat (LRR) domains containing intervening motifs in plants. *Biomolecules*, *2*(4), 288–311.

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., Heaton, M. P., O'Connell, J., Moore, S. S., Smith, T. P. L., Sonstegard, T. S., & Van Tassell, C. P. (2009). Development and characterization of a high density SNP genotyping assay for cattle. *PLoS ONE*, *4*(4), e5350.

McNeill, J., F.R.Barrie, W. R. Buck, V. Demoulin, W. Greuter, D. L. Hawksworth, P. S. Herendeen, S. Knapp, K. Marhold, J. Prado, W. F. Prud'Homme Van Reine, G. F. Smith, J. H. Wiersema, N. (2012). *International Code of Nomenclature for Algae, Fungi and Plants (Melbourne Code) 2012* (p. 240). Germany: Koeltz Scientific Books.

Meinke, D. W. (2013). A survey of dominant mutations in Arabidopsis thaliana. *Trends in Plant Science*, *18*(2), 84–91.

Men, A. E., Wilson, P., Siemering, K., & Forrest, S. (2008). Sanger DNA Sequencing. In *Next Generation Genome Sequencing* (pp. 1–11). Wiley-VCH Verlag GmbH & Co. KGaA.

Messeguer, R., Viruel, M. A., de Vicente, M. C., Dettori, M. T., & Quarta, R. (1994). RFLPs. In *Methods of molecular marker analysis in Prunus. IRTA Technical Report* (pp. 26–38).

Minoia, S., Petrozza, A., D'Onofrio, O., Piron, F., Mosca, G., Sozio, G., Cellini, F., Bendahmane, A., & Carriero, F. (2010). A new mutant genetic resource for tomato crop improvement by TILLING technology. *BMC Research Notes*, *3*, 69.

Mizuno, S., Osakabe, Y., Maruyama, K., Ito, T., Osakabe, K., Sato, T., Shinozaki, K., & Yamaguchi-Shinozaki, K. (2007). Receptor-like protein kinase 2 (RPK 2) is a novel factor controlling anther development in *Arabidopsis thaliana*. *The Plant Journal: For Cell and Molecular Biology*, *50*(5), 751–766.

Mnejja, M., Garcia-Mas, J., Howad, W., Badenes, M. L., & Arús, P. (2004). Simple-sequence repeat (SSR) markers of Japanese plum (*Prunus salicina* Lindl.) are highly polymorphic and transferable to peach and almond. *Molecular Ecology Notes*, *4*(2), 163–166.

Moing, A., Poëssel, J-L., Svanella-Dumas, L., Loonis, M., & Kervella, J. (2003). Biochemical basis of low fruit quality of *Prunus davidiana*, a pest and disease resistance donor for peach breeding. *Journal of the American Society for Horticultural Science*, *128*(1), 55–62.

Monet, R. (1979). *Transmission génétique du caractér "fruit doux" chez le pêcher. Incidence sur la sélection pour la qualiter*. *Eucarpia Fruit Section, Tree Fruit Breeding 1979; INRA, Angers, France* (pp. 273–279). INRA, Angers, France: Breeding, Proceedings of Eucarpia Fruit Section: Tree Fruit.

Monet, R., Guye, a., Roy, M., & Dachary, N. (1996). Peach mendelian genetics: a short review and new results. *Agronomie*, *16*(5), 321–329.

Monforte, A. J., Diaz, A. I., Caño-Delgado, A., & van der Knaap, E. (2014). The genetic basis of fruit morphology in horticultural crops: lessons from tomato and melon. *Journal of Experimental Botany*.2: 11.

Moorthie, S., Mattocks, C. J., & Wright, C. F. (2011). Review of massively parallel DNA sequencing technologies. *The HUGO Journal*, *5*(1-4), 1–12.

Morgan, D.R., Douglas, Soltis, D. E., & Robertson, K. R. (1994). Systematic and evolutionary implications of rbcL sequence variation in Rosaceae. *American Journal of Botany*, *81*(7), 890–903.

Morgante, M., & Olivieri, A M. (1993). PCR-amplified microsatellites as markers in plant genetics. *The Plant Journal: For Cell and Molecular Biology*, *3*(1), 175–182.

Muños, S., Ranc, N., Botton, E., Bérard, A., Rolland, S., Duffé, P., Carretero, Y., Le Parlier, M. C., Delanlance, C., Bouzayen, M., Brunel, D., Causse, M. (2011). Increase in tomato locule number is controlled by two single-nucleotide polymorphisms located near WUSCHEL. *Plant Physiology*, *156*(4), 2244–54.

Muralidharan, O., Natsoulis, G., Bell, J., Newburger, D., Xu, H., Kela, I., Ji., H., & Zhang, N. (2012). A cross-sample statistical model for SNP detection in short-read sequencing data. *Nucleic Acids Research*, *40*(1), e5.

Nagaty, M.A., El-Assal, S- E-D., Rifaat, M. M. (2011). Characterization of the genetic diversity of peach cultivars in taif by RAPD-PCR. *American Journal of Applied Sciences* 8 (7): 708-715.

Nam, K. H., & Li, J. (2002). BRI1/BAK1, a receptor kinase pair mediating brassinosteroid signaling. *Cell*, *110*(2), 203–212.

Nesi, N., Debeaujon, I., Jond, C., Pelletier, G., Caboche, M., & Lepiniec, L. (2000). The TT8 gene encodes a basic helix-loop-helix domain protein required for expression of DFR and BAN genes in *Arabidopsis* siliques. *The Plant Cell*, *12*(10), 1863–1878.

Nesi, N., Jond, C., Debeaujon, I., Caboche, M., & Lepiniec, L. (2001). The *Arabidopsis* TT2 gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed. *The Plant Cell*, *13*(9), 2099–2114.

Ng, M., & Yanofsky, M. F. (2001). Function and evolution of the plant MADS-box gene family. *Nat Rev Genet*, *2*(3), 186–195.

Nilo., R., Campos-Vargas, R., & Orellana, A. (2012). Assessment of *Prunus persica* fruit softening using a proteomics approach. *Journal of Proteomics*, *75*(5), 1618–1638.

Nodine, M. D., & Tax, F. E. (2008). Two receptor-like kinases required together for the establishment of *Arabidopsis* cotyledon primordia. *Developmental Biology*, *314*(1), 161–170.

Nodine, M. D., Yadegari, R., & Tax, F. E. (2007). RPK1 and TOAD2 are two receptor-like kinases redundantly required for arabidopsis embryonic pattern formation. *Developmental Cell*, *12*(6), 943–956. 3

Nürnberger, T., & Kemmerling, B. (2006). Receptor protein kinases–pattern recognition receptors in plant immunity. *Trends in Plant Science*, *11*(11), 519–522.

Okie, W., & Layne, D. (2008). "Early Augustprince"and "Augustprince"Peaches. *HortScience*, *43*(5), 1600–1602.

Okie, W. R. (1998). *Handbook of peach and nectarine varieties: performance in the southeastern United States and index of names*. U.S. Dept. of Agriculture, Agricultural Research Service.

Oliver, R., Lazo, G., Lutz, J., Rubenfield, M., Tinker, N., Anderson, J., Morehead, N. H. W., Adhikary, D., Jellen, E. N., Maughan, P.J., Guedira, G. L. B., Chao, S., beattie, A. D., Carson, M. L., Rines, H. W., Obert, D. E., Bonman, J. M., & Jackson, E. W. (2011). Model SNP development for complex genomes based on hexaploid oat using high-throughput 454 sequencing technology. *BMC Genomics*, *12*(1), 77.

Padilla, I. M. G., Golis, A., Gentile, A., Damiano, C., & Scorza, R. (2006). Evaluation of transformation in peach Prunus persica explants using green fluorescent protein (GFP) and beta-glucuronidase (GUS) reporter genes. *Plant Cell Tiss Org*, *84*(3), 309–314.

Paran, I., & van der Knaap, E. (2007). Genetic and molecular regulation of fruit and plant domestication traits in tomato and pepper. *Journal of Experimental Botany*, *58*(14), 3841–52.

Pascal, T., Pfeiffer, F., & Kervella, J. (2010). Powdery mildew resistance in the peach cultivar Pamirskij 5 is genetically linked with the Gr gene for leaf color. *HortScience*, *45*(1), 150–152.

Paterson, A. H. (1996). *Genome mapping in plants* (R.G. Lande., p. 330).

Peace, C., Bassil, N., Main, D., Ficklin, S., Rosyara, U. R., Stegmeir, T., Sebolt, A., Gilmore, B., Lawley, C., Mockler, T.C., Wilhelm, L., & Iezzoni, A. (2012). Development and Evaluation of a Genome-Wide 6K SNP Array for Diploid Sweet Cherry and Tetraploid Sour Cherry. *PLoS ONE*, *7*(12), e48305.

Peace, C., & Norelli, J. (2009). Genomics approaches to crop improvement in the Rosaceae. In F. K. Gardiner SE (Ed.), *Genetics and genomics of Rosaceae* (pp. 19–53). New York: Springer.

Pérez-Clemente, R. M., Pérez-Sanjuán, A., García-Férriz, L., Beltrán, J. P, & Cañas, L. A. (2005). Transgenic peach plants (*Prunus persica* L.) produced by genetic transformation of embryo sections using the green fluorescent protein (GFP) as an in vivo marker. *Molecular Breeding*, 419–427.

Pérez-Rodríguez, M., Jaffe, F. W., Butelli, E., Glover, B. J., & Martin, C. (2005). Development of three different cell types is associated with the activity of a specific MYB transcription factor in the ventral petal of *Antirrhinum majus* flowers. *Development*, *132*(2), 359–370.

Petri, C., Wang, H., Alburquerque, N., Faize, M., & Burgos, L. (2008). *Agrobacterium*-mediated transformation of apricot (*Prunus armeniaca* L.) leaf explants. *Plant Cell Reports*, *27*(8), 1317–24.

Pettersson, E., Lundeberg, J., & Ahmadian, A. (2009). Generations of sequencing technologies. *Genomics*, *93*(2), 105–11.

Pflieger, S., Lefebvre, V., & Causse, M. (2001). The candidate gene approach in plant genetics: a review. *Molecular Breeding*, *7*(4), 275–291.

Picañol, R., Eduardo, I., Aranzana, M. J., Howad, W., Batlle, I., Iglesias, I., Alonso, J. M., & Arús, P. (2012). Combining linkage and association mapping to search for markers linked to the flat fruit character in peach. *Euphytica*, *190*(2), 279–288.

Pillitteri, L. J., Bemis, S. M., Shpak, E. D., & Torii, K. U. (2007). Haploinsufficiency after successive loss of signaling reveals a role for ERECTA-family genes in *Arabidopsis* ovule development. *Development*, *134*(17), 3099–109.

Pirona, R., Eduardo, I., Pacheco, I., Da Silva Linge, C., Miculan, M., Verde, I., Tartarini, S., Dondini, J., Pea, G., Bassi, D., & Rossini, L. (2013). Fine mapping and identification of a candidate gene for a major locus controlling maturity date in peach. *BMC Plant Biology*, *13*(1), 166.

Polony, T. S., Bowers, S. J., Neiman, P. E., & Beemon, K. L. (2003). Silent point mutation in an avian retrovirus RNA processing element promotes c- myb associated short-latency lymphomas. *Journal of Virology*, *77*(17), 9378–9387.

Pooler, M. R., & Scorza, R. (1995). Regeneration of Peach [*Prunus persica* (L .) Batsch] Rootstock cultivars from cotyledons of mature stored seed. *HortScience*, *30*(2), 355–356.

Potter, D. (2011). Prunus. In Springer (Ed.), *Wild Crop Relatives Genomic and Breeding Resources: Temperate Fruits* (p. 263). Berlin.

Potter, D., Eriksson, T., Evans, R. C., Oh, S., Smedmark, J. E. E., Morgan, D. R., Kerr, Robertson, K. R., Arsenaul, m., Dickinson, T. A., & Campbell, C. S. (2007). Phylogeny and classification of *Rosaceae*. *Plant Systematics Evolution*, *266*(1-2), 5–43.

Potter, D., Gao, F., Bortiri, P. E., Oh, S. H., & Baggett, S. (2002). Phylogenetic relationships in *Rosaceae* inferred from chloroplast matK and trnL-trnF nucleotide sequence data. *Plant Systematics Evolution*, *231*: 77–89.

Prober, J., Trainor, G. L., Dam, R. J., Hobbs, F. W., Robertson, C. W., Zagursky, R. J., Cocuzza, A. J., jensen, M.A., Baumeister, K. (1987). A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science*, *238*(4543):336-41

Quail, M. A., Gu, Y., Swerdlow, H., & Mayho, M. (2012). Evaluation and optimisation of preparative semi-automated electrophoresis systems for Illumina library preparation. *Electrophoresis*, *33*(23), 3521–3528.

Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., Swerdlow, H., & Turner, D. J. (2008). A large genome center's improvements to the Illumina sequencing system. *Nature Methods*, *5*(12), 1005–1010.

Racolta, A., Bryan, A. C., & Tax, F. E. (2014). The receptor-like kinases GSO1 and GSO2 together regulate root growth in Arabidopsis through control of cell division and cell fate specification. *Developmental Dynamics: 243*(2), 257–278.

Raghunathan, A., Ferguson, H. R. J., Bornarth, C. J., Song, W., Driscoll, M., & Lasken, R. S. (2005). Genomic DNA Amplification from a Single Bacteri*um.* Applied and Environmental Microbiology *71*(6), 3342–3347.

Ragoussis, J. (2009). Genotyping technologies for genetic research. *Annual Review of Genomics and Human Genetics*, *10*, 117–133.

Ramos, A. M., Crooijmans, R. P. M. A., Affara, N. A., Amaral, A. J., Archibald, A. L., Beever, J. E., Bendixen, C., Churcher, C., Clark, R., Dehais, P., Hansen, M. S., Hedegaard, J., Hu, Z. L., Kerstens, H. H., Law, A. S., Megens, H. J., Milan, D., Nonneman, D. J., Rohrer, G. A., Rothschild, Smith, T. P., Scahnabel, R. D., Van Tassell, C. P., Taylor, J. F., Wiedmann, R. T., Schook, L. B., & Groenen, M. A. (2009). Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by Next Generation Sequencing Technology. *PLoS ONE*, *4*(8), e6524.

Read, T., & Strachan, A. (2004). *Human Molecular Genetics* (p. 696). Garland.

Rehder, A. (1940). *Manual of cultivated tress and shrubs hardy in North America. Exclusive of subtropical and warmer temperate regions*. (MacMillan, Ed.) (2nd ed., p. 996). New York.

Reig, G., Iglesias, I., Gatius, F., & Alegre, S. (2013). Antioxidant capacity, quality, and anthocyanin and nutrient contents of several peach cultivars [*Prunus persica* (L.) Batsch] grown in Spain. *Journal of Agricultural and Food Chemistry*, *61*(26), 6344–6357.

Renaut, J., Hausman, J. F., Bassett, C., Artlip, T., Cauchie, H. M., Witters, E., & Wisniewski, M. (2008). Quantitative proteomic analysis of short photoperiod and low-temperature responses in bark tissues of peach (*Prunus persica* L. Batsch). *Tree Genetics & Genomes*, *4*(4), 589–600.

Rhee, S., & Mutwil, M. (2014). Towards revealing the functions of all genes in plants. *Trends in Plant Science*, *19*(4), 212–21.

Rieger, M. (2006). Peach *(Prunus persica*). In *Introduction of fruit crops* (pp. 313–323). New York: The Haworth Press.

Rodríguez, G. R., Kim, H. J., & van der Knaap, E. (2013). Mapping of two suppressors of OVATE (sov) loci in tomato. *Heredity*, *111*(3), 256–64.

Rodríguez, G. R., Muños, S., Anderson, C., Sim, S.-C., Michel, A., Causse, M., Gardener, B. B., Francis, D., & van der Knaap, E. (2011). Distribution of SUN, OVATE, LC, and FAS in the tomato germplasm and the relationship to fruit shape diversity. *Plant Physiology*, *156*(1), 275–85.

Rodriguez-A, J., Sherman, W. B., Scorza, R., Wisniewski., & Okie, W. R. (1994). Evergreen peach, its inheritance and dormant behavior. *Journal of the American Society and Horticultural Sciences*, *119*(4), 789–792.

Rojas, G., Méndez, M. a., Muñoz, C., Lemus, G., & Hinrichsen, P. (2008). Identification of a minimal microsatellite marker panel for the fingerprinting of peach and nectarine cultivars. *Electronic Journal of Biotechnology*, *11*(5), 1–12.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., & Nyrén, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Analytical. Biochemistry.*, *242*, 84.

Rozen, S., & Skaletsky, H. (1999). Primer3 on the WWW for general users and for biologist programmers. *Bioinformatics Methods and Protocols*, *132*.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*(4), 406–425.

Salazar, J. A., Rasouli, M., Moghaddam, R. F., Zamani, Z., Imani, A., & Martínez-Gómez, P. (2014). Low-cost strategies for development of molecular markers linked to agronomic traits in Prunus. *Agricultural Sciences*, *5*(4), 430–439.

Sánchez, G., Venegas-Calerón, M., Salas, J. J., Monforte, A., Badenes, M. L., & Granell, A. (2013). An integrative "omics" approach identifies new candidate genes to impact aroma volatiles in peach fruit. *BMC Genomics*, *14*, 343.

Sanger, F., Nicklen, S., & Coulson, A. R., (1977). DNA sequencing with chain-terminating. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467.

Santana, Q., Coetzee, M., Steenkamp, E., Mlonyeni, O., Hammond, G., Wingfield, M., & Wingfield, B. (2009). Microsatellite discovery by deep sequencing of enriched genomic libraries. *BioTechniques*, *46*(3), 217–223.

Sarris, P. F., Trantas, E. A, Baltrus, D. A, Bull, C. T., Wechter, W. P., Yan, S., Ververidis, F., Almeida, N. F., Jones, C. D., Dangl, J. J., Panopoulos, N. J., Vinatzer, B. A., & Goumas, D. E. (2013). Comparative genomics of multiple strains of *Pseudomonas cannabina* pv. alisalensis, a potential model pathogen of both monocots and dicots. *PloS One*, *8*(3), e59366.

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, *19*(R2), R227–40.

Schlöterrer, C. (2002). A Microsatellite-Based Multilocus Screen for the Identification of local selective sweeps. *Genetics*, 160(2):753-63.

Schmid, K. J., Sorensen, T. R., Stracke, R., Torjek, O., Altmann, T., Mitchell-Olds, T., & Weisshaar, B. (2003). Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Research*, *13*(6A), 1250–1257.

Schmidt, A. (1924). Histologische studien am phanerogamen vegetationspunkten. *Botanisches Archives*, *8*, 345–404.

Schötterer, C. (2000). Evolutionary dynamics of microsatellite DNA. *Chromosoma*, *109*(6), 365–371.

Schug, M. D., Mackay, T. F. C., & Aquadro, C. F. (1997). Low mutation rates of microsatellite loci in *Drosophila melanogaster*. *Nature Genetics*, *15*(1), 99–102.

Scorza, R., Melnicenco, L., Dang, P., & Abbott, A. G. (2002). Testing a microsatellite marker for selection of columnar growth habit in peach (*Prunus persica* (L.)Batsch). *Acta Horticulturae*, (592), 285–289.

Scorza, R., & Sherman, W. B. (1996). Tree and tropical fruits. In J. W. & Sons (Ed.), *Fruit breeding* (pp. 325–340). New York.

Scott., D. H., & Weinberger., J. H. (1944). Inheritance of pollen sterility in some peach varieties. *Proceedings of the American Society for Horticultural Science*, (45), 229–232.

Serna, L., & Martin, C. (2006). Trichomes: different regulatory networks lead to convergent structures. *Trends in Plant Science*, *11*(6), 274–280.

Sessions, A., Nemhauser, J. L., McColl, A., Roe, J. L., Feldmann, K. A., & Zambryski, P. C. (1997). ETTIN patterns the *Arabidopsis* floral meristem and reproductive organs. *Development*, *124*(22), 4481–91.

Shabalina, S. A, Spiridonov, N. A, & Kashina, A. (2013). Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. *Nucleic Acids Research*, *41*(4), 2073–2094.

Shamel, A. (1938). A saucer peach bud variation. *Journal of Heredity*, 29 (7): 259.

Shapiro, E., Biezuner, T., & Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews. Genetics*, *14*(9), 618–30.

Sharpe, R. H., Hesse, C. O., Lownsberry, B. F., Perry, V. G., & Hansen, C. J. (1970). Breeding peaches for root-knot nematode resistance. *Journal of the American Society for Horticultural Science*, *94*, 209–212.

Shen, Z., Confolent, C., Lambert, P., Poëssel, J.-L., Quilot-Turion, B., Yu, M., Ruijuan, Ma., & Pascal, T. (2013). Characterization and genetic mapping of a new blood-flesh trait controlled by the single dominant locus DBF in peach. *Tree Genetics & Genomes*, *9*(6), 1435–1446.

Shikata, M., Koyama, T., Mitsuda, N., & Ohme-Takagi, M. (2009). *Arabidopsis* SBP-box genes SPL10, SPL11 and SPL2 control morphological change in association with shoot maturation in the reproductive phase. *Plant & Cell Physiology*, *50*(12), 2133–2145.

Shpak, E. D., McAbee, J. M., Pillitteri, L. J., & Torii, K. U. (2005). Stomatal patterning and differentiation by synergistic interactions of receptor kinases. *Science*, *309*(5732), 290–293.

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L.,Jaiswal, P., Mockaitis, K., Liston, A., Mane, S. P., Burns, P., Davis, T. D., Slovin, J. P., Bassil, N., Hellens, R. P., Evans, C., Harkins, T., Kodira, C., Desany, B., Crasta, O. R., Jensen, R. V., Allan, A. C., Michael, T. P., Setubal, J. C., Celton, J-M., Willians, K P., Holt, S. H., Ruiz-Rojas, J. J., Chatterjee, M., Liu, B., Silva, H., Meisel, L., Adato, A., Filichkin, S. A., Troggio, M., Ashman, T-L., Dharmawardhana, P., Elser, J., Raja, R., Priest, H. D., Bryant Jr, D. W., Fox, S. E., Givan, S. A., Wilhelm, L. J., Naithani, S., Christoffels, A., Salama, D. Y., López-Girona, E., Zdepski, A., Wang, W., Kerstetter, R.A., Schwab, W., Korban, S. S., Davik, J., Monfort, A., Denoyes-Rothan, B., Arús, P., Mittler, R., Flinn, Aharoni, A., Bennetzen, J. L., Salzberg, S. L., Dickerman, A. W., Velasco, R., Borodovsky, M., Veilleux, R. E., & Folta, K. M. (2011). The genome of woodland strawberry (*Fragaria vesca*). *Nature Genetics*, *43*(2), 109–116.

Silva, C., Garcia-Mas, J., Sánchez, a M., Arús, P., & Oliveira, M. M. (2005). Looking into flowering time in almond (*Prunus dulcis* (Mill) D. A. Webb): the candidate gene approach.*Theoretical and Applied Genetics*, *110*(5), 959–968.

Sim, S.C., Van Deynze, A., Stoffel, K., Douches, D. S., Zarka, D., Ganal, M. W., Chetelat, R. T., Hutton, S. F., Scott, J. W., Gardner, R.G., Panthee, D. R., Mutschler, M., Myers, J. R., & Francis, D. M. (2012). High-density SNP genotyping of tomato (*Solanum lycopersicum* L.) reveals patterns of genetic variation due to breeding. *PLoS ONE*, *7*(9), e45520.

Simmonds, N. W. (1966). *Bananas*. Second edition. London: Longmans.

Slatkin, M. (1995). A Measure of population subdivision based on microsatellite allele frequencies. *Genetics*, 139(1):457-62.

Smith, L. G., & Oppenheimer, D. G. (2005). Spatial control of cell expansion by the plant cytoskeleton. *Annual Review of Cell and Developmental Biology*, *21*, 271–295.

Sokolov, B. P. (1990). Primer extension technique for the detection of single nucleotide in genomic DNA. *Nucleic Acids Research*, *18(12)*: 3671.

Sompornpailin, K., Makita, Y., Yamazaki, M., & Saito, K. (2002). A WD-repeat-containing putative regulatory protein in anthocyanin biosynthesis in Perilla frutescens. *Plant Molecular Biology*, *50*(3), 485–495.

Stoner, R. L. (1948). Peach tree. Google Patents.

Swerdlow, H., Wu, S., Harke, H., & Dovichi, N. J. (1990). Capillary gel electrophoresis for DNA sequencing: Laser-induced fluorescence detection with the sheath flow cuvette. *Journal of Chromatography A*, *516*(1), 61–67.

Szymkowiak, E. J., & Sussex, I. M (1996). What chimeras can tell us about plant development. *Annual Review of Plant Physiology and Plant Molecular Biology*.47 (6):351-376.

Tadiello, A., Pavanello, A., Zanin, D., Caporali, E., Colombo, L., Rotino, G. L., Trainotti, L., & Casadoro, G. (2009). A PLENA-like gene of peach is involved in carpel formation and subsequent transformation into a fleshy fruit. *Journal of Experimental Botany*, *60*(2), 651–661.

Tang, H., Ren, Z., Reustlé, G., & Krczal, G. (2002). Plant regeneration from leaves of sweet and sour cherry cultivars. *Scientia Horticulturae*, *93* (3-4): 235-244.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular evolutionary Genetics analysis Version 6.0. *Molecular Biology and Evolution*, 1–12.

Tao, Y., Jiang, L., Liu, Q., Zhang, Y., Zhang, R., Ingvardsen, C. R., Frei, U. K., Wang, b., Lai, J., Lünbberstedt, T., & Xu, M. (2013). Combined linkage and association mapping reveals candidates for Scmv1, a major locus involved in resistance to sugarcane mosaic virus (SCMV) in maize. *BMC Plant Biology*, *13*, 162.

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature*, *408*(6814), 796–815.

Thiel, T., Michalek, W., Varshney, R. K., & Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics*, *106*(3), 411–422.

Thomas, J. W., Touchman, J. W., Blakesley, R. W., Bouffard, G. G., Beckstrom-Sternberg, S. M., Margulies, E. H., Blanchette, M., Siepel, A. C., Thomas, P. J., McDowell, J. C., Maskeri, B., Hansen, N. F., Schwartz, M. S., Weber, R. J., Kent, W. J., Karolchik, D., Bruen, T. C., Bevan, R., Cutler, D. J., Schwartz, S., Elnitski, L., Idol, J. R., Prasad, A. B., Lee-Lin, S. Q., Maduro, V. V., Summers, T. J., Portnoy, M. E., Dietrich, N. L., Akhter, N., Ayele, K., Benjamin, B., Cariaga, K., Brinkley, C. P., Brooks, S. Y., Granite, S., Guan, X., Gupta, J., Haghighi, P., Ho, S. L, Huang, M. C, Karlins, E., Laric, P. L., Legaspi, R., Lim, M. J., Maduro, Q. L., Masiello, C. A., Mastrian, S. D., McCloskey, J. C., Pearson, R., Stantripop, S., Tiongson, E. E., Tran J. T., Tsurgeon, C., Vogt, J. L., Walker, M. A., Wetherby K. D., Wiggins, L. S., Young, A. C., Zhang, L. H., Osoegawa, K., Zhu, B., Zhao, B., Shu, C. L., De Jong, P. J., Lawrence, C. E., Smit, A. F., Chakravarti, A., Haussler, D., Green, P., Miller, & W., Green, E. D. (2003). Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424(6950):788-93.

Tilkat, E., Onay, A., Yıldırım, H., & Ayaz, E. (2009). Direct plant regeneration from mature leaf explants of pistachio, *Pistacia vera* L. *Scientia Horticulturae*, *121*(3), 361–365.

Torii, K. U. (2004). Leucine-rich repeat receptor kinases in plants: structure, function, and signal transduction pathways. International Review of Cytology, 234:1-46.

Torii, K. U., Mitsukawa, N., Oosumi, T., Matsuura, Y., Yokoyama, R., Whittier, R. F., & Komeda, Y. (1996). The *Arabidopsis* ERECTA gene encodes a putative receptor protein kinase with extracellular leucine-rich repeats. *The Plant Cell*, *8*(4), 735–746.

Toyama, T. K. (1974). Haploidy in peach. *Hortscience*, (9), 187–188.

Treangen, T. J., & Salzberg, S. L. (2013). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*, *13*(1), 36–46.

Tsai, I. J., Hunt, M., Holroyd, N., Huckvale, T., Berriman, M., & Kikuchi, T. (2013). Summarizing Specific Profiles in Illumina Sequencing from Whole-Genome Amplified DNA. *DNA Research*: an International Journal for Rapid Publication of Reports on Genes and Genomes (3):243-54

Tsuwamoto, R., Fukuoka, H., & Takahata, Y. (2008). GASSHO1 and GASSHO2 encoding a putative leucine-rich repeat transmembrane-type receptor kinase are essential for the normal development of the epidermal surface in Arabidopsis embryos. *The Plant Journal: For Cell and Molecular Biology*, *54*(1), 30–42.

Uchida, N., Shimada, M., & Tasaka, M. (2013). ERECTA-family receptor kinases regulate stem cell homeostasis via buffering its cytokinin responsiveness in the shoot apical meristem. *Plant & Cell Physiology*, *54*(3), 343–51.

Ushijima, K., Sassa, H., Dandekar, A. M., Gradziel, R. T., Tao, R., & Hirano, H. (2003). Structural and transcriptional analysis of the self-incompatibility locus of almond: identification of a pollen-expressed F-box gene with haplotype-specific polymorphism. *The Plant Cell,* 3:771-781.

Vavilov, N. I. (1926). *Studies on the origin of cultivated plants* (pp. 1–245). (Russian) Bulletin of Applied Botany and Plant Breeding.

Vavilov, N. I. (1951). *Origin, Variation, Immunity and Breeding of Cultivated Plants: Phytogeographic Basis of Plant Breeding*. Redwood City Seed Co.

Vazquez, F., Perez, T., Albornoz, J., & Domínguez, A. (2000). Estimation of microsatellite mutation rates in *Drosophila melanogaster*. *Genetics Research*, *76*(03), 323–326.

Veitia, R., & Birchler, J. (2010). Dominance and gene dosage balance in health and disease: why levels matter! *The Journal of Pathology*, (2), 174–185.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S. K., Troggio, M., Pruss, D., Salvi, S., Pindo, M., Baldi, P., Castelletti, S., Cavaiuolo, M., Coppola, G., Costa, F., Cova, V., Dal Ri, A., Goremykin, V., Komjanc, M., Longhi, S., Magnago, P., Malacarne, G., Malnoy, M., Micheletti, D., Moretto, M., Perazzolli, M., Si-Ammour, A., Vezzulli, S., Zini, E., Eldredge, G., Fitzgerald, L. M., Gutin, N., Lanchbury, J., Macalma ,T., Mitchel, J. T., Reid, J., Wardell, B., Kodira, C., Chen, Z., Desany, B., Niazi, F., Palmer, M., Koepke, T., Jiwan, D., Schaeffer, S., Krishnan,V., Wu, C., Chu, V. T., King, S. T., Vick, J., Tao, Q., Mraz, A., Stormo, A., Stormo, K., Bogden ,R., Ederle, D., Stella, A., Vecchietti, A., Kater, M. M., Masiero, S., Lasserre, P., Lespinasse, Y., Allan, A. C., Bus, V., Chagné, D., Crowhurst, R. N., Gleave, A. P., Lavezzo, E., Fawcett, J. A., Proost, S., Rouzé, P., Sterck, L., Toppo, S., Lazzari, B., Hellens, R. P., Durel, C. E., Gutin, A., Bumgarner, R. E., Gardiner, S. E., Skolnick, M., Egholm, M., Van de Peer, Y., Salamini, F., & Viola, R. (2010). The genome of the domesticated apple (Malus × domestica Borkh.). *Nature Genetics*, *42*(10), 833–839.

Vendramin, E., Pea, G., Dondini, L., Pacheco, I., Dettori, M. T., Gazza, L., Scalabrin, S., Strozzi, F., Tartarini, S., Bassi, D., & Rossini, L. (2014). A unique mutation in a MYB Gene cosegregates with the nectarine phenotype in peach. *PloS One*, *9*(3), e90574.

Venkatesan, B. M., & Bashir, R. (2011). Nanopore sensors for nucleic acid analysis. *Nature Nanotechnology*, *6*(10), 615–24.

Verde, I., International Peach Genome Initiative., Abbott, A. G., Scalabrin, S., Jung, S., Shu, S., Marroni, F., Zhebentyayeva, T., Dettori, M. T., Grimwood, J., Cattonaro, F., Zuccolo, A., Rossini, L., Jenkins, J., Vendramin, E., Meisel, L. A., Decroocq, V., Sosinski, B., Prochnik, S., Mitros, T., Policriti, A., Cipriani, G., Dondini, L., Ficklin, S., Goodstein, D. M., Xuan, P., Del Fabbro, C., Aramini, V., Copetti, D., Gonzalez, S., Horner, D. S., Falchi, R., Lucas, S., Mica, E., Maldonado, J., Lazzari, B., Bielenberg, D., Pirona, R., Miculan, M., Barakat, A., Testolin, R., Stella, A., Tartarini, S., Tonutti, P., Arús, P., Orellana, A., Wells, C., Main, D., Vizzotto, G., Silva, H., Salamini, F., Schmutz, J., Morgante, M., Rokhsar, D. S. (2013). The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome volution. *Nature Genetics*, *45*(5), 487–94.

Verde, I., Bassil, N., Scalabrin, S., Gilmore, B., Lawley, C. T., Gasic, K., Micheletti, D., Rosyara, U. R., Cattonaro, F., Vendramin, E., Main, D., Aramini, V., Blas, A. L., Mockler, T. C., Bryant, D. W., Wilhelm., L., Troggio, M., Sosnski, B., Aranzana, M. J., Arús, P., Iezzoni, A., Morgante, M., & Peace, C. (2012). Development and evaluation of a 9K SNP array for peach by internationally coordinated SNP detection and validation in breeding germplasm. *PloS One*, *7*(4), e35668.

Verde, I., Lauria, M., Dettori, M. T., Vendramin, E., Balconi, C., Micali, S., Wang, Y., Marrazzo., M. T., Cipriani, G., Hartings, H., Testolin, R., Abbott, A. G., Motto, M., & Quarta, R. (2005). Microsatellite and AFLP markers in the *Prunus persica* [L. (Batsch)] x *P. ferganensis* BC(1) linkage map: saturation and coverage improvement*. Theoretical and Applied Genetics.*, *111*(6), 1013–1021.

Vigouroux, Y., Jaqueth, J. S., Matsuoka, Y., Smith, O. S., Beavis, W. D., Smith, J. S. C., & Doebley, J. (2002). Rate and pattern of mutation at microsatellite loci in maize. *Molecular Biology and Evolution* , *19* (8), 1251–1260.

Vilanova, S., Romero, C., Abbott, a G., Llácer, G., & Badenes, M. L. (2003). An apricot (*Prunus armeniaca* L.) F_2 progeny linkage map based on SSR and AFLP markers, mapping plum pox virus resistance and self-incompatibility traits*. Theoretical and Applied Genetics*, *107*(2), 239–247.

Vilanova, S., Sargent, D. J., Arús, P., Monfort, A., & Arus, P. (2008). Synteny conservation between two distantly-related *Rosaceae* genomes: *Prunus* (the stone fruits) and *Fragaria* (the strawberry). *BMC Plant Biology*, *8*(1), 67.

Viruel, M. A., Messeguer, R., De Vicente, M. C., Garcia-Mas, J., Puigdomenech, P., Vargas, F. J., & Arús, P. (1995). A linkage map with RFLP and isozyme markers for almond. *Theoretical and Applied Genetics*, *91*(6-7), 964–971.

Voelkerding, K. V, Dames, S. a, & Durtschi, J. D. (2009). Next-generation sequencing: from basic research to diagnostics. *Clinical Chemistry*, *55*(4), 641–58.

Walker, a R., Davison, P. a, Bolognesi-Winfield, a C., James, C. M., Srinivasan, N., Blundell, T. L., Esch, J. J., Marks, M. D., & Gray, J. C. (1999). The TRANSPARENT TESTA GLABRA1 locus, which regulates trichome differentiation and anthocyanin biosynthesis in Arabidopsis, encodes a WD40 repeat protein. *The Plant Cell*, *11*(7), 1337–1350.

Walker, J. (1994). Structure and function of the receptor-like protein kinases of higher plants. *Plant Molecular Biology*, *26*(5), 1599–1609.

Wang, D. G ., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M.,

Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lipshutz, R., Chee, M., Lander, E. S (1998). Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science*, *280*(5366), 1077–1082.

Wang, W., Wei, Z., Lam, T.W., & Wang, J. (2011). Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. *Scientific Reports*, *1*, 55.

Wang, X., Kota, U., He, K., Blackburn, K., Li, J., Goshe, M. B., Huber, S. C & Clouse, S. D. (2008). Sequential transphosphorylation of the BRI1/BAK1 receptor kinase complex impacts early events in brassinosteroid signaling. *Developmental Cell*, *15*(2), 220–235.

Wang, X., Lu, P., & Luo, Z. (2013). GMATo: A novel tool for the identification and analysis of microsatellites in large genomes. *Bioinformation*, *9*(10), 541–544.

Wang, Y., Georgi, L. L., Reighard, G. L., Scorza, R., & Abbott, A G. (2002). Genetic mapping of the evergrowing gene in peach [*Prunus persica* (L.) Batsch]. *The Journal of Heredity*, *93*(5), 352–358.

Wang, Y., Xie, X., & Long, L. E. (2014). The effect of postharvest calcium application in hydro-cooling water on tissue calcium content, biochemical changes, and quality attributes of sweet cherry fruit. *Food Chemistry*, *160*, 22–30.

Warburton, M., & Bliss, F. (1996). Genetic diversity in peach (*Prunus persica* L. Batch) revealed by randomly amplified polymorphic DNA (RAPD) markers and compared to inbreeding coefficients. *Journal of the American Society for Horticultural Science*, *121*(6), 1012–1019.

Wegscheider, E., Benjak, A., & Forneck, A. (2009). Clonal Variation in Pinot noir Revealed by S-SAP Involving Universal Retrotransposon-Based Sequences. *American Journal of Enology and Viticulture* , *60* (1), 104–109.

Wehrhahn, C. F. (1975). The evolution of selectively similar electro-phoretically detectable alleles in finite natural populations, *Genetics* , *80* (2), 375–394.

Weinberger JH, Marth PC, S. D. H. (1943). Inheritance study of root knot nematode resistance in certain peach varieties. *Proceedings of the American Society for Horticultural Science*, (42), 321–325.

Weismann. (1892). Das Keimplasma. In Fischer Jena (Ed.), *Eine Theorie der Vererbung*. Germany.

Werneck, H. L. (1956). *Römischer und vorrömischer Wein-und Obstban in Österreichischen Donaraum*. (Wein, Ed.) (pp. 114–131).

Werner, D. J., & Creller, M. A. (1997). Genetic Studies in Peach: Inheritance of Sweet Kernel and Male Sterility. *Journal of the American Society for Horticultural Science*, *122*(2), 215–217.

Whipple, C. J., Ciceri, P., Padilla, C. M., Ambrose, B. a, Bandong, S. L., & Schmidt, R. J. (2004). Conservation of B-class floral homeotic gene function between maize and *Arabidopsis*. *Development*, *131*(24), 6083–6091.

Whiteley, A. S., Jenkins, S., Waite, I., Kresoje, N., Payne, H., Mullan, B., Allcock, R., & O'Donnell, A. (2012). Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) Platform. *Journal of Microbiological Methods*, *91*(1), 80–88.

Winter, P., & Kahl, G. (1995). Molecular marker technologies for plant improvement. *World Journal of Microbiology and Biotechnology*, *11*(4), 438–448.

Wu, Y., Machado, A. C., White, R. G., Llewellyn, D. J., & Dennis, E. S. (2006). Expression profiling identifies genes expressed early during lint fibre initiation in cotton. *Plant & Cell Physiology*, *47*(1), 107–127.

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J., & van der Knaap, E. (2008). A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science*, *319*(5869), 1527–30.

Xie, R., Li, X., Chai, M., Song, L., Jia, H., Wu, D., Chen, M., Chen, K., Aranzana, M. J., & Gao, Z. (2010). Evaluation of the genetic diversity of Asian peach accessions using a selected set of SSR markers. *Scientia Horticulturae*, *125*(4), 622–629.

Xu, S.L., Rahman, A., Baskin, T. I., & Kieber, J. J. (2008). Two leucine-rich repeat receptor kinases mediate signaling, linking cell wall biosynthesis and ACC synthase in Arabidopsis. *The Plant Cell*, *20*(11), 3065–3079.

Xu, X., Peng, M., Fang, Z., & Xu, X. (2000). The direction of microsatellite mutations is dependent upon allele length . *Nature Genetics*, *24*(4), 396–399.

Xu, Y. (2010). Molecular Plant Breeding. Oxfordshire (UK) & Cambridge (USA): CAB International.

Yadav, R. K., Fulton, L., Batoux, M., & Schneitz, K. (2008). The *Arabidopsis* receptor-like kinase STRUBBELIG mediates inter-cell-layer signaling during floral development. *Developmental Biology*, *323*(2), 261–270.

Yamamoto, N., Tsugane, T., Watanabe, M., Yano, K., Maeda, F., Kuwata, C., Torki, M., Ban, Y., Nishimura, S., & Shibata, D. (2005). Expressed sequence tags from the laboratory-grown miniature tomato (*Lycopersicon esculentum*) cultivar Micro-Tom and mining for single nucleotide polymorphisms and insertions/deletions in tomato cultivars. *Gene*, *356*(0), 127–134.

Yamamoto, T., Mochida, K., Imai, T., Haji, T., Yaegaki, H., Yamaguchi, M., Yamaguchi, M., Matsuta, N., Ogiwara, I., & Hayashi, T. (2003). Parentage analysis in Japanese peaches using SSR Markers. *Breeding Science*, *53*(1), 35–40.

Yamamoto, T., Shimada, T., Imai, T., Yaegaki, H., Haji, T., Matsuta, N., Yamaguchi, M., & Hayashi, T. (2001). Characterization of morphological traits based on a genetic linkage map in peach. *Breeding Science*, *51*, 271–278.

Yoon, J., Liu, D., Song, W., Liu, W., Zhang, A., & Li, S. (2006). Genetic diversity and ecogeographical phylogenetic relationships among peach and nectarine cultivars based on simple sequence repeat (SSR) markers. *Journal of the American Society for Horticultural Science*, *131*(4), 513–521.

Yu, J., Arbelbide, M., & Bernardo, R. (2005). Power of in silico QTL mapping from phenotypic, pedigree, and marker data in a hybrid breeding program. *Theoretical and Applied Genetics*, *110*(6), 1061–1067.

Yu, N., Cai, W.-J., Wang, S., Shan, C.-M., Wang, L.-J., & Chen, X.-Y. (2010). Temporal control of trichome distribution by microRNA156-targeted SPL genes in *Arabidopsis thaliana*. *The Plant Cell*, *22*(7), 2322–2335.

Yu, X., Guda, K., Willis, J., Veigl, M., Wang, Z., Markowitz, S., Adams, M. D., & Sun, S. (2012). How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? *BioData Mining*, *5*(1), 6.

Zalapa, J. E., Cuevas, H., Zhu, H., Steffan, S., Senalik, D., Zeldin, E., McCown, B., Harbut, R., & Simon, P. (2012). Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *American Journal of Botany*, *99*(2), 193–208.

Zane, L., Bargelloni, L., & Patarnello, T. (2002). Strategies for microsatellite isolation: A review. *Molecular Ecology*, *11*(1), 1–16.

Zeitouni, B., Boeva, V., Janoueix-Lerosey, I., Loeillet, S., Legoix-né, P., Nicolas, A., Delattre, O., & Barillot, E. (2011). SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics, 26,15:1895-1896*

Zhang, H., & Forde, B. G. (1998). An *Arabidopsis* MADS box gene that controls nutrient-induced changes in root architecture. *Science*, *279*(5349), 407–409.

Zhang, H., Cao, Y., Zhao, J., Li, X., Xiao, J., & Wang, S. (2011). A pair of orthologs of a leucine-rich repeat receptor kinase-like disease resistance gene family regulates rice response to raised temperature. *BMC Plant Biology*, *11*(1), 160.

Zhang, J., Arro, J., Chen, Y., & Ming, R. (2013). Haplotype analysis of sucrose synthase gene family in three Saccharum species. *BMC Genomics*, *14*, 314.

Zhang, Q., Chen, W., Sun, L., Zhao, F., Huang, B., Yang, W., Tao, Y., Wang, J., Yuan, Z., Fan, G., Xing, Z., Han, C., Pan, H., Zhong, X., Shi, W., Liang, X., Du, D., Sun, F., Xu, Z., Hao, R., Lv, T., Lv, Y., Zheng, Z., Sun, M., Luo, L., Cai, M., Gao, Y., Wang, J., Yin, Y., Xu, X., Cheng, T., & Wang, J. (2012). The genome of *Prunus* mume. *Nature Communications*, *3*(4), 1318.

Zhao, D.-Z., Wang, G., Speal, B., & Ma, H. (2002). The EXCESS MICROSPOROCYTES1 gene encodes a putative leucine-rich repeat receptor protein kinase that controls somatic and reproductive cell fates in the Arabidopsis anther. *Genes & Development*, *16*(15), 2021–2031.

Zhou, A., Wang, H., Walker, J. C., & Li, J. (2004). BRL1, a leucine-rich repeat receptor-like protein kinase, is functionally redundant with BRI1 in regulating Arabidopsis brassinosteroid signaling. *The Plant Journal: For Cell and Molecular Biology*, *40*(3), 399–409.

Zhu, C., & Perry, S. E. (2005). Control of expression and autoregulation of AGL15, a member of the MADS-box family. *The Plant Journal: For Cell and Molecular Biology*, *41*(4), 583–594.

Zipfel, C., Kunze, G., Chinchilla, D., Caniard, A., Jones, J. D. G., Boller, T., & Felix, G. (2006). Perception of the bacterial PAMP EF-Tu by the receptor EFR restricts *Agrobacterium*-mediated trasnformaction. *Cell,* 125(4):749-60.

Zuckerkandl, E., & Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In Evolving Genes and Proteins (1965), pp. 97-16.

APPENDICES CHAPTER I

Appendix CI.1 Details of the 17 SSRs used in the population structure analysis.

SSR	Allele length-bp	Ta [1] (ºC)	Origin	TxE LG (cM)[2]	TxE BIN [3]	Physical position[4]	Reference
BPPCT006	111;113;115;117;125;127;129;131;133;135;137	57	*P.persica*	G8(14.1)	19	8:5982783_5982783	Dirlewanger *et al.*,2002
BPPCT007	124;129;139;141;143;145;147	57	*P.persica*	G3(11.2)	12	3:2741939_2741939	Dirlewanger *et al.*,2002
BPPCT008	99;127;133;135;137;145;147;154,156;158;160	57	*P.persica*	G6(30.1)	39	6:10280088_10280088	Dirlewanger *et al.*,2002
BPPCT014	197;200;214;225;227	57	*P.persica*	G5(44)	46	5:16626108_16626635	Dirlewanger *et al.*,2002
BPPCT017	148;151;158;161;163;165;172;176;178	57	*P.persica*	G5(20.1)	21	5:11174442_1117442	Dirlewanger *et al.*,2002
BPPCT020	188;196;198;200;202;206	57	*P.persica*	G1(52.6)	52	1:33281418_33281615	Dirlewanger *et al.*,2002
BPPCT025	172;175;180;182;186;188;190;192;194;196;198	57	*P.persica*	G6(56.4)	56	6:21129947_21129947	Dirlewanger *et al.*,2002
CPPCT002	172;175;180;182;186;188;190;192;194;196;198	52	*P.persica*	G3(31.9)	37	3:16205250_16207665	Aranzana *et al.,*2002
CPPCT022	74;98;100	50	*P.persica*	G7(18.6)	25	7:10225365_10225583	Aranzana *et al.,*2002
CPPCT033	249;251;261;279;281;284;291;293;295;297	50	*P.persica*	G7(38.9)	41	7:16702195_16702488	Aranzana *et al.,*2002
UDP96-001	119;121:123;126;128;136	57	*P.persica*	G6(17.5)	25	6:7040897_7041018	Cipriani *et al.,*1999
UDP96-003	117;12;122;124;126;129;134;136;138;141	57	*P.persica*	G4(28.3)	28	4:8757479_8757621	Cipriani *et al.,*1999
UDP96-005	154;156;158;167;169;171;173	57	*P.persica*	G1(29.2)	29	Position not found	Cipriani *et al.,*1999
UDP96-013	181;183;188;198;200;206;208	57	*P.persica*	G2(27.8)	28	2:18895941_18896211	Cipriani *et al.,*1999
UDP98-024	104;109;119;123;125	57	*P.persica*	G4(11.3)	18	4:3499686_3499806	Cipriani *et al.,*1999
UDP98-025	113;128;132;134;136	57	*P.persica*	G2(9.6)	13	2:10872238-10872370	Cipriani *et al.,*1999
UDP98-409	116;122;124;146;148	57	*P.persica*	G8(44.5)	60	8:17783855_17783529	Cipriani *et al.,*1999

(1) Ta: annealing temperature;
(2) Linkage group and distance in centimorgans from the top of the linkage group as in the Prunus reference map;
(3) Bin of the Prunus reference map;
(4) Physical position in the peach genome sequence v.1.0 (http://www.rosaceae.org/species/ prunus_persica/ genome_v1.0)

Aranzana *et al.*, (2002) Plant Breeding 121:184-184; Cipriani *et al.*, (1999) Theor Appl Genet 99:65-72; Dirlewanger *et al.*, (2002) Theor Appl Genet 105:127 – 138

Appendix CI.2. List of GeneBank accession numbers for 'HonyeGlo' (HG at the end of the sequence name) and 'Glenna' (Gl at the end of the sequence name) sequences.

Sequence Name	Accession Number
BankIt1688764 DF_HL38HG	KJ023869
BankIt1688764 DF_HL38Gl	KJ023870
BankIt1688764 DF-45552Gl	KJ023871
BankIt1688764 DF-45552HG	KJ023872
BankIt1688764 DF-35167HG	KJ023873
BankIt1688764 DF-35167Gl	KJ023874
BankIt1688764 DF-19433HG	KJ023875
BankIt1688764 DF-19433Gl	KJ023876
BankIt1688764 DF-11052HG	KJ023877
BankIt1688764 DF-11052Gl	KJ023878
BankIt1688764 DF-9128HG	KJ023879
BankIt1688764 DF-9128Gl	KJ023880
BankIt1688764 DF-7617Gl	KJ023881
BankIt1688764 DF-7617HG	KJ023882
BankIt1688764 DF-7589HG	KJ023883
BankIt1688764 DF-7589Gl	KJ023884
BankIt1688764 DF-6331HG	KJ023885
BankIt1688764 DF-6331Gl	KJ023886
BankIt1688764 DF-4607HG	KJ023887
BankIt1688764 DF-4607Gl	KJ023888
BankIt1688764 DF-2044HG	KJ023889
BankIt1688764 DF-2044Gl	KJ023890
BankIt1688764 DF-1652HG	KJ023891
BankIt1688764 DF-1652Gl	KJ023892
BankIt1688764 DF-0875HG	KJ023893
BankIt1688764 DF-0875Gl	KJ023894

Appendix CI.3. CPPCT40 genotype and field characterization of the 542 seedlings analyzed. These seelings dereived from 34 peach crosses involving at least 44 different parents.

Seedling	CPPCT040	Field evaluation	Seedling	CPPCT040	Field evaluation
Field evaluation (acid/subacid)	genotype	(acid/subacid)	Field evaluation (acid/subacid)	genotype	(acid/subacid)
PN414-002	193/193	subacid	PN511-014	199/199	acid
PN414-003	193/193	subacid	PN511-015	199/199	acid
PN414-004	193/199	subacid	PN511-017	193/199	subacid
PN414-005	193/199	subacid	PN511-019	193/193	subacid
PN414-006	193/193	subacid	PN511-021	199/199	acid
PN414-009	193/193	subacid	PN511-023	193/199	subacid
PN414-010	193/199	subacid	PN511-024	199/199	acid
PN414-011	193/193	subacid	PN511-025	199/199	acid
PN414-013	193/193	subacid	PN511-026	199/199	acid
PN414-015	193/193	subacid	PN527-001	199/199	acid
PN414-016	193/193	subacid	PN527-003	193/199	subacid
PN414-017	193/199	subacid	PN527-004	193/199	subacid
PN414-018	193/193	subacid	PN527-006	193/199	subacid
PN414-021	193/193	subacid	PN527-007	199/199	acid
PN414-022	193/199	subacid	PN527-008	193/199	subacid
PN414-023	193/199	subacid	PN527-009	199/199	acid
PN414-024	193/193	subacid	PN527-010	193/199	subacid
PN414-026	193/193	subacid	PN527-011	193/199	subacid
PN414-027	193/199	subacid	PN527-012	193/199	subacid
PN414-028	193/193	subacid	PN527-014	199/199	acid
PN414-030	193/193	subacid	PN527-015	193/199	subacid
PN414-031	193/193	subacid	PN527-016	193/199	subacid
PN414-032	193/193	subacid	PN527-019	199/199	acid
PN414-033	193/193	subacid	PN527-020	199/199	acid
PN414-035	193/193	subacid	PN527-021	199/199	acid
PN414-037	193/199	subacid	PN527-022	193/199	subacid
PN511-001	193/199	subacid	PN527-023	193/199	subacid
PN511-002	199/199	acid	PN527-024	199/199	acid
PN511-003	199/199	acid	PN527-027	199/199	acid
PN511-004	193/193	subacid	PN527-029	199/199	acid
PN511-005	199/199	subacid	PN527-030	199/199	acid
PN511-006	193/193	subacid	PN527-034	199/199	acid
PN511-007	193/199	subacid	PN527-039	193/199	subacid
PN511-008	193/199	acid	PN594-002	193/199	subacid
PN511-009	193/199	subacid	PN594-003	199/199	acid
PN511-012	193/199	subacid	PN594-005	193/199	subacid
PN511-013	199/199	acid	PN594-006	193/199	subacid

Seedling Field evaluation (acid/subacid)	CPPCT040 genotype	Field evaluation (acid/subacid)	Seedling Field evaluation (acid/subacid)	CPPCT040 genotype	Field evaluation (acid/subacid)
PN594-007	193/199	subacid	PN502-009	193/193	subacid
PN594-008	199/199	acid	PN502-010	193/193	subacid
PN594-009	199/199	acid	PN502-011	193/193	subacid
PN594-011	199/199	acid	PN502-012	193/193	subacid
PN594-012	199/199	acid	PN502-013	193/193	subacid
PN594-013	199/199	acid	PN502-015	193/193	subacid
PN594-015	199/199	acid	PN502-016	193/193	subacid
PN594-016	199/199	acid	PN502-017	193/193	subacid
PN594-018	199/199	acid	PN502-018	193/193	subacid
PN594-019	199/199	acid	PN502-020	193/193	subacid
PN596-001	193/199	subacid	PN502-024	193/193	subacid
PN596-007	199/199	subacid	PN502-025	193/193	subacid
PN596-008	193/193	subacid	PN502-026	193/193	subacid
PN596-009	193/199	subacid	PN502-027	193/193	subacid
PN596-010	193/199	subacid	PN502-028	193/193	subacid
PN596-011	193/193	subacid	PN502-032	193/193	subacid
PN596-012	193/193	subacid	PN502-034	193/193	subacid
PN596-013	193/199	subacid	PN502-035	193/193	subacid
PN596-016	193/193	subacid	PN502-038	193/193	subacid
PN596-017	193/199	subacid	PN502-040	193/193	subacid
PN596-019	199/199	acid	PN502-041	193/193	subacid
PN596-020	193/199	subacid	PN502-042	193/193	subacid
PN596-022	193/193	subacid	PN502-043	193/193	subacid
PN596-026	193/193	subacid	PN502-044	193/193	subacid
PN596-027	193/193	subacid	PN502-045	193/193	subacid
PN596-028	193/193	subacid	PN502-046	193/193	subacid
PN596-029	193/199	subacid	PN502-047	193/193	subacid
PN603-001	193/193	subacid	PN503-001	193/193	subacid
PN603-005	193/193	subacid	PN503-002	193/199	subacid
PN603-012	193/193	subacid	PN503-003	193/193	subacid
PN603-019	193/193	subacid	PN503-004	193/193	subacid
PN603-022	193/193	subacid	PN507-002	193/199	subacid
PN603-023	193/193	subacid	PN507-004	193/199	subacid
PN603-024	193/199	subacid	PN507-005	193/193	subacid
PN605-006	193/199	subacid	PN507-007	193/193	subacid
PN605-008	199/199	acid	PN507-008	193/193	subacid
PN605-011	199/199	acid	PN507-009	193/193	subacid
PN605-013	199/199	acid	PN507-010	193/199	subacid
PN502-002	193/193	subacid	PN508-002	199/199	acid
PN502-004	193/193	subacid	PN508-004	193/193	subacid
PN502-008	193/193	subacid	PN508-005	193/199	subacid

Seedling Field evaluation (acid/subacid)	CPPCT040 genotype	Field evaluation (acid/subacid)	Seedling Field evaluation (acid/subacid)	CPPCT040 genotype	Field evaluation (acid/subacid)
PN508-006	199/199	acid	PN551-005	193/199	subacid
PN508-007	193/199	subacid	PN551-006	193/199	subacid
PN508-008	193/199	acid	PN551-008	193/199	subacid
PN508-009	193/199	subacid	PN551-009	193/199	subacid
PN508-012	199/199	acid	PN551-010	193/199	subacid
PN508-013	193/199	subacid	PN582-001	193/193	subacid
PN508-014	193/193	subacid	PN582-004	199/199	acid
PN508-016	199/199	acid	PN582-005	193/199	subacid
PN508-018	199/199	acid	PN582-006	193/199	subacid
PN508-023	193/199	subacid	PN582-007	193/199	subacid
PN508-024	193/199	subacid	PN587-002	193/199	subacid
PN508-025	193/199	subacid	PN587-003	199/199	acid
PN508-028	199/199	acid	PN587-005	199/199	acid
PN508-029	193/199	subacid	PN587-006	199/199	acid
PN508-030	199/199	acid	PN587-007	193/199	subacid
PN508-032	193/193	subacid	PN587-009	199/199	acid
PN534-001	199/199	acid	PN587-010	193/193	subacid
PN534-002	199/199	acid	PN587-011	193/199	subacid
PN534-009	199/199	subacid	PN587-012	193/193	subacid
PN534-013	199/199	acid	PN560-001	193/193	acid
PN534-014	199/199	acid	PN560-002	193/199	subacid
PN534-015	193/199	subacid	PN560-003	193/199	subacid
PN534-016	199/199	acid	PN560-004	193/199	acid
PN534-017	199/199	acid	PN560-005	199/199	acid
PN536-001	193/199	subacid	PN588-001	193/199	subacid
PN536-002	199/199	acid	PN588-002	193/199	subacid
PN536-003	193/193	subacid	PN588-003	193/199	subacid
PN536-005	193/193	subacid	PN588-004	193/199	acid
PN536-006	193/193	subacid	PN588-005	193/199	subacid
PN536-007	193/199	subacid	PN588-007	193/199	subacid
PN536-008	193/193	subacid	PN593-002	193/199	subacid
PN536-009	199/199	acid	PN593-004	199/199	acid
PN536-010	193/199	subacid	PN593-005	193/199	subacid
PN536-011	193/193	subacid	PN602-001	193/199	subacid
PN536-012	199/199	acid	PN602-004	193/199	subacid
PN536-013	193/193	subacid	PN604-001	193/199	acid
PN536-014	193/199	subacid	PN604-002	199/199	acid
PN537-001	193/199	subacid	PN604-003	193/199	subacid
PN537-002	193/199	subacid	PN604-004	193/193	subacid
PN537-006	193/199	subacid	PN397-001	193/199	subacid
PN551-001	193/199	subacid	PN397-002	193/199	subacid

Seedling Field evaluation (acid/subacid)	CPPCT040 genotype	Field evaluation (acid/subacid)	Seedling Field evaluation (acid/subacid)	CPPCT040 genotype	Field evaluation (acid/subacid)
PN397-003	197/199	acid	PN397-052	193/199	subacid
PN397-004	193/199	subacid	PN397-054	193/199	subacid
PN397-005	193/199	subacid	PN399-006	199/201	acid
PN397-006	199?	acid	PN399-011	199/201	acid
PN397-007	197/199	acid	PN399-012	199/201	acid
PN397-008	193/199	subacid	PN399-013	193/199	subacid
PN397-009	193/199	subacid	PN399-014	197/199	acid
PN397-010	193/199	subacid	PN399-016	197/199	acid
PN397-011	193/199	subacid	PN399-018	199/199	acid
PN397-012	197/199	acid	PN399-019	199/201	acid
PN397-013	193/199	subacid	PN399-020	197/199	acid
PN397-014	193/199	subacid	PN399-025	199/201	acid
PN397-015	197/199	acid	PN399-026	197/199	acid
PN397-017	193/199	subacid	PN399-027	197/199	acid
PN397-019	193/199	subacid	PN399-028	197/199	acid
PN397-020	199/199	acid	PN399-029	199/201	acid
PN397-021	193/199	subacid	PN399-030	197/199	acid
PN397-022	193/199	subacid	PN399-032	197/199	acid
PN397-023	197/199	acid	PN399-033	199/201	acid
PN397-024	199/199	acid	PN399-034	197/199	acid
PN397-025	199/199	acid	PN399-035	199/201	acid
PN397-026	197/199	acid	PN399-036	197/199	acid
PN397-027	199/199	acid	PN399-037	197/199	acid
PN397-028	193/199	acid	PN409-092	197/199	acid
PN397-029	199/199	acid	PN409-093	193/199	subacid
PN397-030	199/199	acid	PN409-094	199/199	acid
PN397-031	193/199	subacid	PN409-095	193/199	subacid
PN397-032	197/199	acid	PN409-096	199/199	acid
PN397-033	193/199	subacid	PN409-097	193/199	subacid
PN397-035	193/199	acid	PN409-098	193/199	subacid
PN397-036	197/199	acid	PN409-099	193/199	subacid
PN397-038	193/199	subacid	PN409-103	199/199	acid
PN397-040	193/199	subacid	PN409-104	199/199	acid
PN397-041	193/199	subacid	PN409-105	199/199	acid
PN397-044	193/199	subacid	PN409-106	193/199	subacid
PN397-045	193/199	subacid	PN409-107	193/199	subacid
PN397-046	199/199	acid	PN409-108	193/199	subacid
PN397-048	193/199	subacid	PN409-109	193/199	subacid
PN397-049	193/199	subacid	PN409-110	193/199	subacid
PN397-050	193/199	subacid	PN409-111	197/199	acid
PN397-051	197/199	acid	PN409-112	193/199	subacid

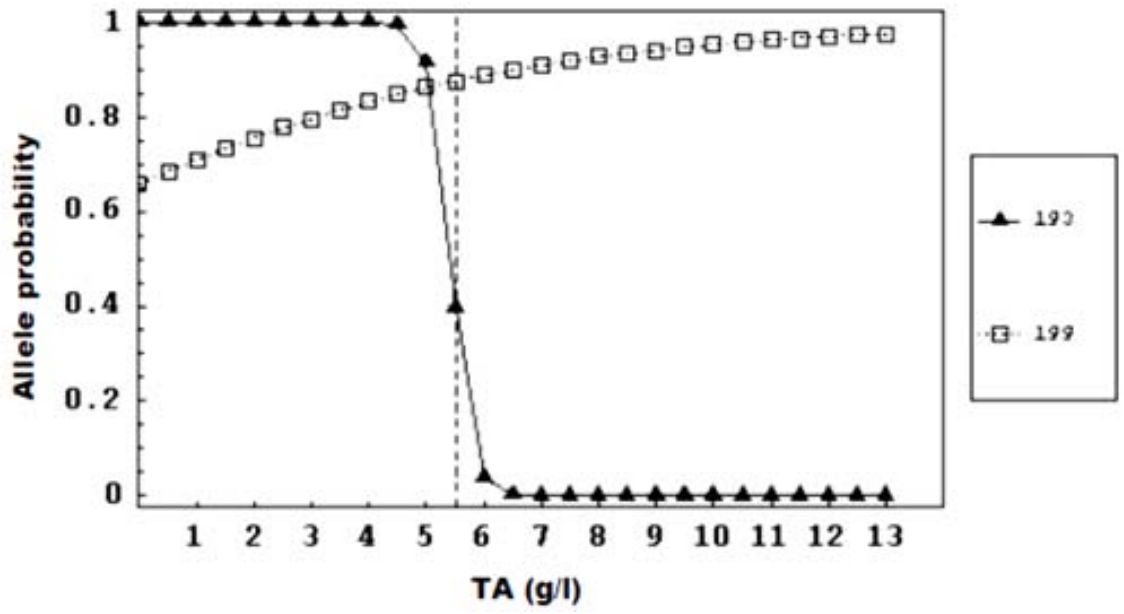
Seedling Field evaluation (acid/subacid)	CPPCT040 genotype	Field evaluation (acid/subacid)	Seedling Field evaluation (acid/subacid)	CPPCT040 genotype	Field evaluation (acid/subacid)
PN409-113	193/199	subacid	PN409-022	193/199	subacid
PN409-114	197/199	acid	PN409-024	193/199	subacid
PN409-115	197/199	acid	PN409-025	193/199	subacid
PN409-116	199/199	acid	PN409-026	197/199	acid
PN409-117	197/199	acid	PN409-027	193/199	subacid
PN409-118	193/199	subacid	PN409-028	199/199	acid
PN409-119	197/199	acid	PN409-031	193/199	subacid
PN409-121	199/199	acid	PN409-032	193/199	subacid
PN409-122	197/199	acid	PN409-033	193/199	subacid
PN409-123	193/199	subacid	PN409-034	197/199	acid
PN409-125	197/199	acid	PN409-035	193/199	subacid
PN409-126	197/199	acid	PN409-036	193/199	subacid
PN409-127	197/199	acid	PN409-037	199/199	acid
PN409-128	193/199	subacid	PN409-039	197/199	acid
PN409-129	197/199	acid	PN409-040	199/199	acid
PN399-038	197/199	acid	PN409-041	199/199	acid
PN399-040	197/199	acid	PN409-042	193/199	subacid
PN399-042	199/201	subacid	PN409-043	197/199	acid
PN399-043	199/201	acid	PN409-044	199/199	acid
PN399-044	197/199	acid	PN409-046	199/199	acid
PN399-045	197/199	acid	PN409-047	197/199	acid
PN399-046	199/201	acid	PN409-048	199/199	acid
PN399-051	199/201	acid	PN409-050	197/199	acid
PN409-001	197/199	acid	PN409-051	193/199	subacid
PN409-002	199/199	acid	PN409-052	197/199	acid
PN409-003	197/199	acid	PN409-053	193/199	subacid
PN409-004	199/199	acid	PN409-054	197/199	acid
PN409-005	193/199	subacid	PN409-055	197/199	acid
PN409-006	199/199	acid	PN409-056	199/199	acid
PN409-007	199/199	acid	PN409-057	193/199	subacid
PN409-008	199/199	acid	PN409-058	199/199	acid
PN409-009	193/199	subacid	PN409-059	199/199	acid
PN409-010	193/199	subacid	PN409-060	193/199	subacid
PN409-011	199/199	acid	PN409-061	197/199	acid
PN409-012	199/199	acid	PN409-062	197/199	subacid
PN409-015	193/199	subacid	PN409-063	197/199	acid
PN409-016	193/199	subacid	PN409-064	197/199	acid
PN409-017	193/199	subacid	PN409-065	197/199	acid
PN409-019	197/199	subacid	PN409-067	197/199	subacid
PN409-020	193/199	subacid	PN409-071	199/199	acid
PN409-021	199/199	acid	PN409-072	193/199	subacid

Seedling Field evaluation (acid/subacid)	CPPCT040 genotype	Field evaluation (acid/subacid)	Seedling Field evaluation (acid/subacid)	CPPCT040 genotype	Field evaluation (acid/subacid)
PN409-073	193/199	subacid	PN409-117	197/199	acid
PN409-074	193/199	subacid	PN409-118	193/199	subacid
PN409-075	193/199	subacid	PN409-119	197/199	acid
PN409-076	199/199	acid	PN409-121	199/199	acid
PN409-077	199/199	acid	PN409-122	197/199	acid
PN409-078	193/199	subacid	PN409-123	193/199	subacid
PN409-079	199/199	acid	PN409-125	197/199	acid
PN409-080	193/199	subacid	PN409-126	197/199	acid
PN409-081	197/199	acid	PN409-127	197/199	acid
PN409-082	193/199	subacid	PN409-128	193/199	subacid
PN409-083	199/199	subacid	PN409-129	197/199	acid
PN409-084	193/199	subacid	PN409-130	197/199	acid
PN409-085	199/199	acid	PN409-131	193/199	subacid
PN409-086	193/199	subacid	PN409-132	193/199	subacid
PN409-087	199/199	subacid	PN409-133	197/199	acid
PN409-088	199/199	acid	PN409-134	193/199	subacid
PN409-089	197/199	acid	PN409-135	199/199	acid
PN409-090	199/199	acid	PN409-136	193/199	subacid
PN409-091	197/199	acid	PN409-137	193/199	subacid
PN409-092	197/199	acid	PN409-138	199/199	acid
PN409-093	193/199	subacid	PN409-139	199/199	acid
PN409-094	199/199	acid	PN409-141	193/199	subacid
PN409-095	193/199	subacid	PN409-142	193/199	subacid
PN409-096	199/199	acid	PN409-143	197/199	acid
PN409-097	193/199	subacid	PN409-144	193/199	subacid
PN409-098	193/199	subacid	PN409-145	193/199	subacid
PN409-099	193/199	subacid	PN409-146	197/199	acid
PN409-103	199/199	acid	PN409-147	197/199	acid
PN409-104	199/199	acid	PN409-148	199/199	acid
PN409-105	199/199	acid	PN409-149	197/199	acid
PN409-106	193/199	subacid	PN434-002	193/193	subacid
PN409-107	193/199	subacid	PN434-003	193/199	subacid
PN409-108	193/199	subacid	PN434-004	193/193	subacid
PN409-109	193/199	subacid	PN434-005	193/193	subacid
PN409-110	193/199	subacid	PN434-006	193/193	subacid
PN409-111	197/199	acid	PN434-010	193/199	subacid
PN409-112	193/199	subacid	PN434-011	199/199	acid
PN409-113	193/199	subacid	PN434-013	193/199	subacid
PN409-114	197/199	acid	PN434-014	193/199	subacid
PN409-115	197/199	acid	PN434-015	193/199	subacid
PN409-116	199/199	acid	PN434-016	193/199	subacid

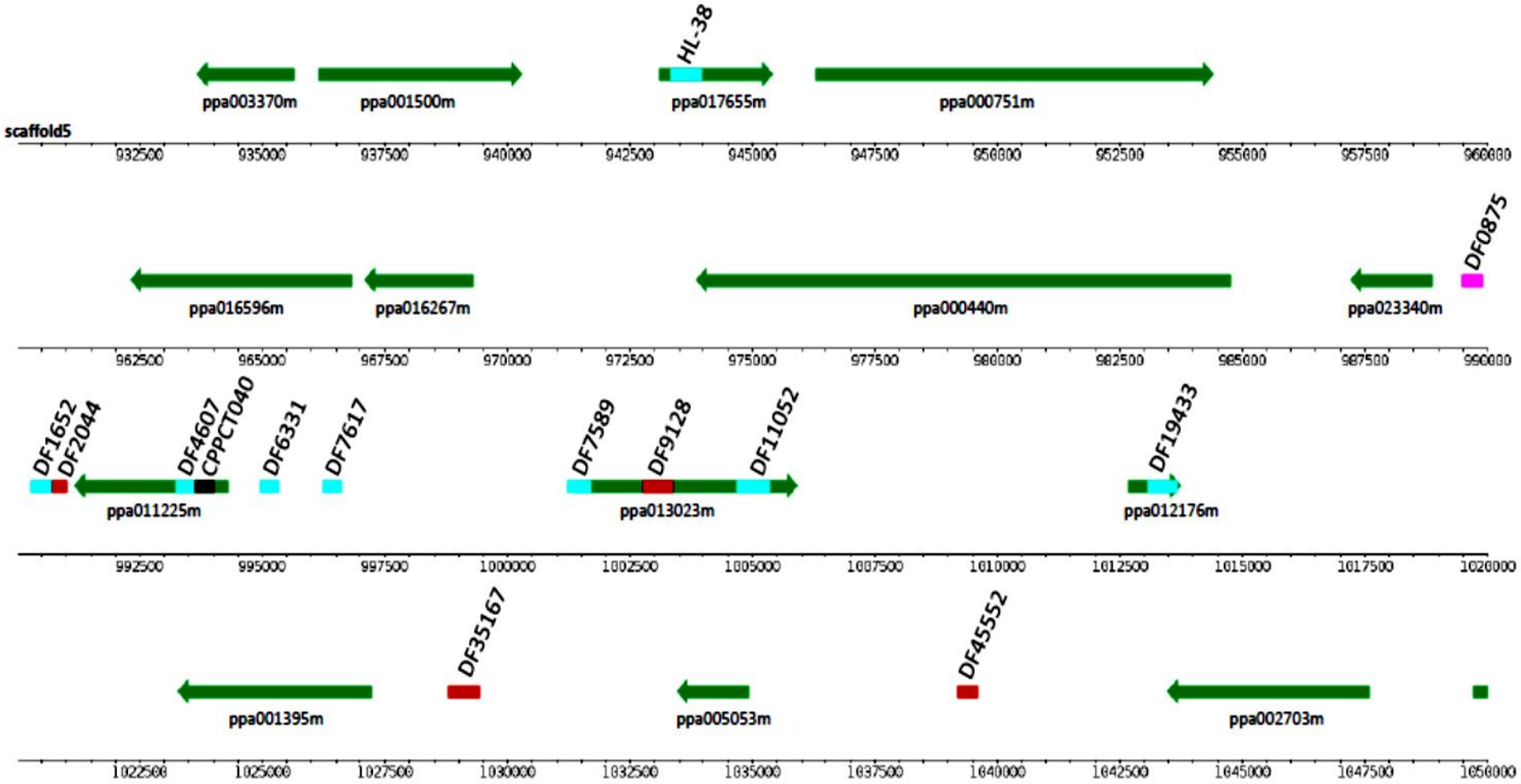
Seedling Field evaluation (acid/subacid)	CPPCT040 genotype	Field evaluation (acid/subacid)	Seedling Field evaluation (acid/subacid)	CPPCT040 genotype	Field evaluation (acid/subacid)
PN434-017	193/199	subacid	PN447-002	193/199	acid
PN500-001	193/201	subacid	PN447-003	193/193	acid
PN500-002	193/201	subacid	PN496-001	-	subacid
PN500-003	193/201	subacid	PN496-003	193/199	subacid
PN500-004	193/201	subacid	PN496-004	193/193	subacid
PN500-005	193/201	subacid	PN496-005	193/199	subacid
PN500-006	193/201	subacid	PN496-006	199/199	acid
PN500-007	193/201	subacid	PN496-007	193/199	subacid
PN500-008	193/201	subacid	PN496-008	193/199	acid
PN500-010	193/201	subacid	PN496-009	193/193	subacid
PN500-011	193/201	subacid	PN496-011	199/199	acid
PN500-012	193/201	subacid	PN496-012	193/199	subacid
PN500-013	193/201	subacid	PN398-001	193/199	subacid
PN500-014	193/201	subacid	PN398-002	199/199	acid
PN500-015	193/201	subacid	PN398-003	199/199	acid
PN500-016	193/201	subacid	PN398-004	199/199	acid
PN500-017	193/201	subacid	PN398-005	193/199	subacid
PN500-018	193/201	subacid	PN398-006	193/199	subacid
PN402-001	199/199	acid	PN398-007	193/193	acid
PN402-002	193/199	subacid	PN398-009	199/199	acid
PN402-003	193/199	subacid	PN398-010	199/199	acid
PN402-004	193/199	subacid	PN398-011	199/199	acid
PN402-005	199/199	acid	PN398-012	199/199	acid
PN402-006	193/199	subacid	PN398-014	193/193	subacid
PN402-008	199/199	acid	PN398-015	199/199	acid
PN402-009	193/199	subacid	PN398-016	199/199	acid
PN402-010	193/199	subacid	PN398-017	193/199	subacid
PN402-011	199/199	acid	PN398-018	193/199	subacid
PN402-012	193/199	subacid	PN398-020	193/199	subacid
PN402-013	193/199	acid	PN398-021	199/199	acid
PN402-014	193/199	subacid	PN398-022	199/199	acid
PN402-016	193/199	subacid	PN398-023	199/199	acid
PN403-001	193/193	subacid	PN398-024	199/199	acid
PN403-002	193/193	subacid	PN398-025	199/199	acid
PN403-003	193/193	subacid	PN398-026	199/199	acid
PN403-004	193/193	subacid	PN398-027	199/199	acid
PN407-001	199/199	acid	PN398-028	193/193	subacid
PN407-003	193/193	subacid	PN398-029	193/193	subacid
PN408-002	193/199	subacid	PN398-034	193/199	subacid
PN408-003	193/199	subacid	PN398-035	193/199	subacid
PN408-004	193/199	subacid	PN398-036	193/193	subacid

Seedling Field evaluation (acid/subacid)	CPPCT040 genotype	Field evaluation (acid/subacid)
PN398-037	193/193	subacid
PN398-039	193/199	subacid
PN398-041	193/199	subacid
PN398-042	193/199	subacid
PN398-045	193/193	subacid
PN398-050	193/199	acid
PN398-051	199/199	subacid
PN398-052	193/199	acid
PN398-053	193/199	acid

Appendix CI.4 Probability of finding CPPCT040^{193} (▲) and CPPCT040^{199} (□) alleles at different TA (g/l) values.



Appendix CI.5 Graphical summary of a 117.5 kbp region flanking the marker CPPCT040 (in black). Green arrows represent the transcripts annotated in the peach genome). The amplicons sequenced in 38 peach acid and subacid varieties are highlighted in red (monomorphic), blue (polymorphic) and pink (amplicon containing the SNP DS875 genotyped by HRM).



APPENDICES CHAPTER II

APPENDIX CII.1 Haplotypes in the candidate gene ppa022511mg and genotypes for UDP98-412 SSR and for the allelic specific primer pair Flatin1F+Kinase5R. P: peach; N: nectarine; W: white; Y:yellow; F: flat. SNPs: 1_24406522; 2_24406523; 3_24406600; 4_24406601; 5_INDEL_24406672-24406679; 6_24406733; 7_24406753; 8_24406799; 9_24406849; 10_INDEL_24406868-24406879; 11_24406901; 12_24407078; 13_24407180; 14_24407465; 15_24407508

					Amplicon-5											Amplicon-6			
Cultivar	Fruit type	Origin	UDP98-412	Flatin1F	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Almudi	PWF	Spain	131	464/469	G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
Almunia	PWF	Spain	127/131	464/469	G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 04-71	NWF	France	123/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 04-81	PYF	France	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 04-92	PWF	France	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 04-93	PWF	France	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 04-94	PWF	France	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 05-81	PWF	France	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 05-92	PWF	France	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 05-93	PWF	France	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 06-71	NWF	France	127/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 06-73	NWF	France	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 06-80	NWF	France	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 06-83	NWF	France	131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 06-87	NWF	France	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C

Cultivar	Fruit type	Origin	UDP98-412	Flatin1F	Amplicon-5											Amplicon-6			
					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ASF 06-88	NWF	France	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 06-90	PWF	France	127/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 06-91	PWF	France	127/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 06-96	PWF	France	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 06-97	PWF	France	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 06-99	PWF	France	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 07-73	NWF	France	127/131	464/469	G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 07-80	NWF	France	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
ASF 07-98	PWF	France	127/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
Caspe	PWF	Spain	131	464/469	G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
Donutnice	NWF	France	129/131	464/469	G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
Flatelate	PWF	France	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
–Yumyeong (DPRU 1612)	PWR	Soth Korea	131													T	A	G	C
																T	A	G	C
Xin Dai Jiu Bao (DPRU 2267)	PWR	China	125/131													T	A	G	C
																T	A	G	C
Kou Ho (DPRU 2268)	PWR	China	129/131													T	A	G	C
																T	A	G	C
Zin Dai Jiu Bao (DPRU 2363)	PWR	China	125/131													T	A	G	C
																T	A	G	C

Cultivar	Fruit type	Origin	UDP98-412	Flatin1F	Amplicon-5											Amplicon-6			
					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Kawanakajima Hakutou (DPRU 2466)	PWR	Japan	123/131													T	A	G	C
																T	A	G	C
Jing Yu (DPRU 2499)	NWR_	China	131													T	A	G	C
																T	A	G	C
Feng Bai (DPRU 2586)	PWR	China	129/131													T	A	G	C
																T	A	G	C
Okayama 11 (DPRU 2651)	PWR	Japan - China	123/131													T	A	G	C
																T	A	G	C
Wan Pan Tao (DPRU 2652)	PWF	China	129/131													T	A	G	C
																G	C	C	G
Green Pan Tao (DPRU 2661)	PWF	China	129/131													T	A	G	C
																G	C	C	G
Xiong Yue (DPRU 2662)	_	China	131													T	A	G	C
																G	C	C	G
Peregrine (DPRU 0654)	PWR	Unknown	129/131													T	A	G	C
	_															T	A	G	C
Platycarpa (DPRU 2169)	PWF	USA	131													T	A	G	C
																G	C	C	G
Mesembrine	NWF	France	123/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
Nebuly	PWF	Spain	131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
Niqui	PWF	Spain	127/131	464/469	G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
Ordigan	PYF	France	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
Oriane	PYF	France	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
Oriola	NYF	France	125/131	464/469	G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C

Cultivar	Fruit type	Origin	UDP98-412	Flatin1F	Amplicon-5											Amplicon-6			
					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Paraguayo Amarillo	PYF	Spain	125/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
Paraguayo Francia	PWF	Spain	125/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
Paraguayo Delfin	PWF	Spain	131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
Paraguayo Jota	PWF	Spain	127/131	464/469	G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
Parguayo B	PWF	Spain	123/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
Platibelle	PWF	France	123/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
Platifur	PWF	France	129/131	464/469	G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
PN665-01	PWR	China	120/131	464/469												T	A	G	C
																T	A	G	C
PN665-02	PWR	China	125/131	464/469												T	A	G	C
																T	A	G	C
PN665-010	PWR	China	120/131	464/469												T	A	G	C
																T	A	G	C
San Mateo	PWF	Spain	131	464/469	G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
Subirana	NWF	Spain	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
SweetCap	PWF	France	125/131	464/469	G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
T. Robert	PWF	Spain	125/131	464/469	G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
UFO 1	PWF	Italy	127/131	464/469	G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C

Cultivar	Fruit type	Origin	UDP98-412	Flatin1F	Amplicon-5											Amplicon-6			
					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
UFO 2	PWF	Italy	127/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
UFO 3	PWF	Italy	127/131	464/469	G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
UFO 4	PWF	Italy	127/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
UFO 5	PYF	Italy	123/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
UFO 6	PYF	Italy	127/131	464/469	G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
UFO 7	PYF	Italy	123/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
UFO 8	PYF	Italy	129/131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
UFO 9	PWF	Italy	125/131													G	C	C	G
																T	A	G	C
Vilamayor	PWF	Spain	131		G	A	A	A	CTGAATAT	?	?	?	?		?	G	C	C	G
					A	C	G	G		?	?	?	?	GAGGAATTGGATG	?	T	A	G	C
GEM090	_	Uzbequistan	129/131													T	A	G	C
																T	A	G	C
ASF 04-06	NYR	France	129		A	C	G	G		A	C	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	C	A	G	GAGGAATTGGATG	G	T	A	G	C
ASF 04-09	NYR	France	123/129		A	C	G	G		A	C	A	G	GAGGAATTGGATG	A	T	A	G	C
					A	C	G	G		A	C	A	G	GAGGAATTGGATG	G	T	A	G	C
ASF 04-23	NWR	France	123/129		A	C	G	G		A	C	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
ASF 04-27	NWR	France	129		A	C	G	G		A	C	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
ASF 04-30	NWR	France	123/129		A	C	G	G		A	C	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C

Cultivar	Fruit type	Origin	UDP98-412	Flatin1F	Amplicon-5											Amplicon-6			
					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ASF 04-42	PYR	France	129		A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
ASF 04-52	PWR	France	123/129		A	C	G	G		A	C	A	G	GAGGAATTGGATG	A	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
ASF 04-53	PWR	France	129		A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
ASF 05-08	NYR	France	127/129		A	C	G	G		A	C	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
ASF 05-25	NWR	France	127/129		A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
ASF 05-48	PYR	France	129	469	A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
ASF 06-07	NYR	France	127		A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
ASF04-04	NYR	France	129/129		A	C	G	G		A	C	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	C	A	G	GAGGAATTGGATG	G	T	A	G	C
Big bel	NWR	France	123/129	469	A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Big sun	PYR	France	129		A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Binaced	PWR	Spain	129	469	A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Calabacero	PYR	Spain	125/127	469	A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Calante	PYR	Spain	127	469	A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Catherina	PYR	USA	123/127		A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Elegant Lady	PYR	USA	125/127		A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C

Cultivar	Fruit type	Origin	UDP98-412	Flatin1F	Amplicon-5											Amplicon-6			
					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Evaisa	PYR	Spain	123/133	469	A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Extreme July	PYR	Spain	123/129		A	C	G	G		A	C	A	G	GAGGAATTGGATG	A	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Extreme Sweet	NWR	Spain	129	469												T	A	G	C
																T	A	G	C
Feraude	PYR	France	125/127	469	A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Fercluse	PYR	France	125/127		A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Ferlot	PYR	France	125/127	469	A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Garcica	NWR	Spain	127/129		A	C	G	G		A	C	A	G	GAGGAATTGGATG	A	T	A	G	C
					A	C	G	G		A	C	A	G	GAGGAATTGGATG	G	T	A	G	C
Indian Freestone (DPRU 1184)	PWR	USA	125/125													T	A	G	C
																T	A	G	C
Conservera 458 (DPRU 1990)	PYR	Brasil	123/123													T	A	G	C
																T	A	G	C
Honey royale	NYR	USA	123/129													T	A	G	C
																T	A	G	C
HoneyGlo	NYR	USA	129/129		A	C	G	G		A	C	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
IFF0331	PWR	Italy	129/133		A	C	G	G		A	C	A	G	GAGGAATTGGATG	A	T	A	G	C
					A	C	G	G		A	C	A	G	GAGGAATTGGATG	G	T	A	G	C
IFF0800	NWR	Italy	123		A	C	G	G		A	C	A	G	GAGGAATTGGATG	A	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
IFF0813	NYR	Italy	127		A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
IFF0962	PYR	Italy	129	469	A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C

Cultivar	Fruit type	Origin	UDP98-412	Flatin1F	Amplicon-5											Amplicon-6			
					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
IFF1182	NWR	Italy	127/129													T	A	G	C
																T	A	G	C
IFF1190	PYR	Italy	127/129													T	A	G	C
																T	A	G	C
IFF1230	PWR	Italy	123/129		A	C	G	G		A	C	A	G	GAGGAATTGGATG	A	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
IFF1233	PYR	Italy	123/129		A	C	G	G		A	C	A	G	GAGGAATTGGATG	A	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Jeronimo	PYR	Spain	120/129		A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Latefair	NYR	USA	123/125		A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Luciana	NYR	Spain	125/129		A	C	G	G		A	C	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Malatewhite	PWR	France	129/129		A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Maycrest	PYR	USA	120/129	469	A	C	G	G		A	C	A	G	GAGGAATTGGATG	A	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Nectabang	NYR	France	123/127		A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Nectajewel	NWR	France	127/129	469	A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Nectalady	NYR	France	129/129		A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Nectariane	NYR	France	120/129	469	A	C	G	G		A	C	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	C	A	G	GAGGAATTGGATG	G	T	A	G	C
Nectaross	NYR	Italy	127		A	C	G	G		A	C	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C
Nectarreve	NWR	France	127/129		A	C	G	G		A	C	A	G	GAGGAATTGGATG	G	T	A	G	C
					A	C	G	G		A	T	A	G	GAGGAATTGGATG	G	T	A	G	C

Appendix CII.2 Rich leucine repeat receptor-like protein kinases (LRR-RLKs) with known functions in *Arabidopsis thaliana* (Gou et al., 2010)

ID	Subfamily	Gene	Symbol	Functions	Reference
1	LRR I	At4g29990	LRRPK	Light signal transduction	(Deeken & Kaldenhoff, 1997)
2	LRR II	At4g33430	BAK1/AtSERK3	BR signalling/Pathogen response/Cell death	(He *et al*., 2007; Nam & Li, 2002)
3	LRR II	At2g13790	BKK1/AtSERK4	BR signaling/Pathogen response/Cell death	(He *et al.,* 2007)
4	LRR II	At1g71830	AtSERK1	BR signaling/ Male Sporogenesis	(Albrecht *et al*.,2005; Karlova *et al*., 2006)
5	LRR II	At1g34210	AtSERK2	Male Sporogenesis	(Albrecht *et al.,* 2005; Colcombet, 2005)
6	LRR II	At5g16000	NIK1	Antiviral defense response	(Fontes & Santos, 2004)
7	LRR II	At3g25560	NIK2	Antiviral defense response	(Fontes & Santos, 2004)
8	LRR II	At1g60800	NIK3	Antiviral defense response	(Fontes & Santos, 2004)
9	LRR V	At3g13065	SRF4	Leaf size control	(Eyüboglu *et al.,* 2007)
10	LRR V	At1g11130	Scrambled/	Root epidermis patterning/Organ development/	(Eyüboglu *et al.,* 2007; Yadav *et al.,* 2008)
11	LRR X	At4g39400	BRI1	Brassinosteroid receptor	(J Li & Chory, 1997)
12	LRR X	At1g55610	BRL1	Brassinosteroid receptor/Vascular differentiation	(Caño-Delgado *et al.,* 2004; Zhou *et al*., 2004)
13	LRR X	At2g01950	BRL2/VH1	Vascular differentiation	(Clay & Nelson, 2002)
14	LRR X	At3g13380	BRL3	Brassinosteroid receptor/Vascular differentiation	(Caño-Delgado *et al.,* 2004; Zhou *et al.,* 2004)
15	LRR X	At1G69270	RPK1/TOAD1	Abscisic acid signaling/embryonic pattern formation	(Hong *et al*., 1997; Nodine *et al*., 2007-2008)
16	LRR X	At3g02130	RPK2/TOAD2	Anther development/embryonic pattern formation	(Mizuno *et al*., 2007; Nodine & Tax, 2008)
17	LRR X	At5g07280	EMS1/EXS	Anther development	(Wang *et al*., 2008; Zhao *et al*.,2002)
18	LRR X	At5g48380	BIR1	Cell death and innate immunity	(Gao et al., 2009)
19	LRR XI	At4g20140	GSO1	Epidermal surface formation during embryogenesis	(Tsuwamoto *et al*., 2008)
20	LRR XI	At5g44700	GSO2	Epidermal surface formation during embryogenesis	(Tsuwamoto *et al*., 2008)
21	LRR XI	At1g75820	CLV1	Meristem differentiation and maintenance	(Clark *et al.,* 1997)
22	LRR XI	At5g65700	BAM1	Meristem differentiation/Anther development	(DeYoung *et al*., 2006; Hord *et al*., 2006)
23	LRR XI	At3g49670	BAM2	Meristem differentiation/Anther development	(DeYoung *et al*., 2006; Hord *et al*., 2006)
24	LRR XI	At4g20270	BAM3	Meristem differentiation/Anther development	(DeYoung *et al.,* 2006)
25	LRR XI	At2g31880	SOBIR1	Cell death and innate immunity	(Gao *et al.,* 2009)
26	LRR XI	At4g28490	HAESA	floral organ abscission	(Jinn *et al*., 2000)
27	LRR XI	At3g19700	IKU2	Seed size	(Luo & Dennis, 2005)
28	LRR XI	At5g61480	PXY/TDR	Procambium polar cell division/vascular stem cell fate	(Fisher & Turner, 2007; Hirakawa et al., 2008)
29	LRR XII	At5g46330	FLS2	Pathogen response	(Gómez-Gómez & Boller, 2000)
30	LRR XII	At5g20480	EFR	Pathogen response	(Zipfel *et al.,* 2006)
31	LRR XIII	At1g31420	FEI1	Cell Wall Biosynthesis	(Xu *et al*., 2008)
32	LRR XIII	At2g35620	FEI2	Cell Wall Biosynthesis	(Xu *et al*., 2008)
33	LRR XIII	At2g26330	ERECTA	Organ growth/ Stomatal patterning/differentiation	(Shpak *et al*., 2005; Torii *et al*., 1996)
34	LRR XIII	At5g62230	ERL1	Stomatal patterning and differentiation	(Shpak *et al*., 2005; Torii *et al*., 1996)
35	LRR XIII	At5g07180	ERL2	Stomatal patterning and differentiation	(Shpak *et al*., 2005; Torii *et al*., 1996)

Appendix CII.3 Nucleotide and deduced amino acid sequence of round allele for ppa025511m gene. Complete sequence of the full length of this gene was obtained from the peach genome browser. Translated amino acid sequence is also shown under nucleotide sequence. Numbers to the left of each row refer to nucleotide or amino acid position. The nucleotide and the amino acid variations between the round and aborting allele are highlighted by the grey shading. Stop codons are labelled in red and represented in the amino acid sequence by an asterisk.

A. Round nucleotide deduced amino acid sequence

```
1M  K  H  L  L  Q  Y  F  L  L  L  F  L  I  P  K  I  C  F  T
1ATGAAACATTTGCTCCAATATTTCTTGCTCCTATTCTTAATCCCTAAAATCTGTTTTACC
21    I  I  P  A  V  H  S  L  C  T  K  D  Q  Q  L  S  L  L  H  L
61    ATCATCCCTGCCGTTCACAGCCTCTGCACTAAAGACCAGCAACTATCATTGCTCCATTTG
41    K  K  S  L  Q  F  S  H  D  P  D  S  D  S  Y  P  T  K  V  I
121   AAGAAAAGCCTTCAATTTTCTCATGATCCTGATTCTGATTCATACCCAACCAAGGTTATA
61    S  W  N  S  S  T  D  C  C  S  W  L  G  V  N  C  S  S  D  G
181   TCTTGGAATTCAAGCACCGATTGTTGTTCTTGGCTTGGTGTTAATTGCAGTAGTGATGGG
81    H  V  V  G  L  D  L  S  S  E  A  I  N  D  G  I  D  D  S  S
241   CATGTCGTTGGTCTTGACCTTAGCAGCGAAGCTATCAACGATGGCATTGACGATTCAAGC
101   S  L  F  D  L  Q  H  L  Q  S  L  N  L  A  D  N  H  F  T  Y
301   AGTCTCTTCGATCTTCAACACCTTCAAAGCCTCAATTTGGCTGACAACCATTTTACCTAT
121   G  T  R  I  P  S  A  I  G  K  L  V  N  L  R  Y  L  N  L  S
361   GGTACTCGCATTCCATCTGCAATCGGAAAGCTTGTGAACTTGAGGTATCTAAATTTATCA
141   S  C  S  F  Y  G  S  I  P  K  S  I  A  N  L  T  Q  L  V  S
421   TCTTGCAGTTTCTATGGATCAATCCCAAAGTCAATAGCAAATCTAACACAATTGGTTAGT
161   L  H  L  G  L  N  T  F  S  G  S  I  D  S  I  S  W  E  N  L
481   TTGCATTTGGGATTAAATACGTTCAGTGGTTCAATTGATTCTATTAGCTGGGAAAACCTT
181   I  N  L  V  D  L  Q  M  D  D  N  L  L  E  G  S  I  P  S  S
541   ATTAATCTGGTAGACCTCCAGATGGATGACAACCTACTTGAGGGGAGTATTCCATCGTCT
201   L  F  Y  L  P  L  L  T  Q  L  V  L  S  R  N  Q  F  S  G  K
601   CTCTTTTATCTTCCCTTATTGACACAACTAGTACTTTCCCGCAATCAATTCTCTGGTAAA
221   L  H  A  F  S  N  T  S  S  D  L  E  Y  L  D  L  S  E  N  Q
661   CTTCATGCATTTTCTAACACCTCTTCCGACTTAGAATATTTGGACCTTTCAGAAAACCAG
241   I  Q  G  K  I  P  H  W  I  W  S  F  S  H  L  Y  Y  L  N  L
721   ATTCAAGGCAAGATACCCCATTGGATTTGGAGTTTCAGTCATCTTTATTACCTAAATCTT
```

261 S C N S L V T L E A P L Y N S S V S I V

781 TCTTGCAACTCTTTGGTAACTCTAGAAGCTCCTTTATATAATTCTAGTGTATCAATAGTT

281 D L H S N Q L Q G Q I P T F I P F G Y Q

841 GACCTTCATTCAAACCAACTCCAGGGTCAAATCCCAACTTTCATACCATTTGGTTACCAG

301 L D Y S G N H F N S I P S D I G Y F F T

901 CTGGATTACTCAGGCAACCATTTCAATTCTATACCATCTGACATTGGTTATTTCTTCACT

321 S T M F F S L S S N N L H G L I P A S I

961 TCCACAATGTTCTTCTCTCTTTCAAGCAATAACTTGCATGGGCTCATTCCGGCATCAATA

341 C N A T S F L M S L D L S N N F L S G I

1021 TGCAATGCGACAAGTTTTCTTATGAGTCTTGATCTGTCCAATAATTTTCTGAGTGGCATT

361 I P P C L T A M R G L R V L N L A R N N

1081 ATTCCCCCATGCTTGACTGCAATGCGCGGTCTCAGAGTACTTAATTTAGCAAGAAACAAC

381 L T G T I S N F Q V T E Y S L L E I L K

1141 CTCACTGGAACTATTTCTAATTTTCAAGTTACTGAATATAGTTTATTAGAAATTCTAAAG

401 L D G N Q L G G Q F P K S L G N C T Q L

1201 CTCGATGGAAATCAGTTAGGTGGTCAGTTTCCAAAATCTCTAGGTAACTGCATACAGTTA

421 Q V L N L G N N R I T D T F P C L L K N

1261 CAGGTTTTAAACTTGGGAAACAATCGTATAACAGATACATTTCCATGCTTGTTAAAAAAC

441 M S T L R V L V L R S N N F Y G G I G C

1321 ATGTCCACCTTGCGTGTCCTTGTGTTGCGGTCCAACAACTTCTATGGAGGAATTGGATGT

461 P N T Y G T W P V L Q I I H L A H N N F

1381 CCCAACACCTATGGCACCTGGCCAGTGCTTCAAATCATACACCTAGCTCACAACAATTTC

481 T G E I P G I F L T T W Q V M M A P E D

1441 ACTGGTGAAATACCGGGAATATTTTTGACAACATGGCAGGTAATGATGGCTCCCGAGGAT

501 G P L S I V K F Q L D T I I A G K S M L

1501 GGTCCCCTATCGATTGTCAAATTCCAACTGGATACAATTATTGCGGGAAAATCAATGTTG

521 I D Y S F N D R I T V T S K G L E M D L

1561 ATTGATTATTCTTTTAATGATCGTATAACAGTTACCAGCAAAGGGTTAGAGATGGATCTA

541 V R I L S I F T L I D F S C N N F S G P

1621 GTAAGGATTCTATCTATCTTCACCTTGATTGACTTCTCTTGCAACAACTTCAGTGGACCA

561 I P K E M G E F K S L H V L N L S R N S

1681 ATACCTAAGGAAATGGGAGAATTCAAATCACTACATGTCCTTAACTTGTCCAGAAATTCT

581 L T G E I P S S F G N M Q V L E S L D L

1741 TTGACAGGCGAAATCCCATCCTCATTTGGTAACATGCAGGTACTCGAGTCCTTGGACCTG

601 S Q N K L G G E I P Q Q L A K L T F L S

1801 TCACAGAACAAGTTGGGCGGGGAAATTCCACAACAGTTGGCAAAGCTTACTTTCCTTTCG

621 F L N I S Y N Q L V G R I P P S T Q F S

1861 TTCTTGAATATCTCATATAATCAACTGGTCGGCAGGATCCCACCCAGTACTCAGTTTTCA

641 T F P K D S F T G N K G L W G P P L T V

1921 ACATTTCCAAAAGACTCATTTACAGGAAACAAAGGACTATGGGGGCCTCCTTTGACAGTG

661 D N K T G L S P P P A L N G S L P N S G

1981 GATAACAAAACAGGATTATCACCACCACCAGCATTAAATGGAAGCCTTCCAAATTCTGGC

681 H R G I N W D L I S V E I G F T V G F G

2041 CATCGTGGGATTAATTGGGATCTGATCAGTGTTGAAATTGGATTTACAGTTGGCTTTGGA

701 A S V G S L V L C K R W S K W Y Y R A M

2101 GCTTCCGTTGGGTCACTTGTGTTGTGCAAGAGATGGAGTAAGTGGTATTACAGAGCTATG

721 Y R M V L K I F P Q L E E R I G I H R R

2161 TACAGGATGGTTCTTAAGATATTCCCACAGCTGGAGGAAAGAATTGGAATTCATCGAAGA

741 H V H I N R R W R R *****

2221 CATGTTCACATAAATCGAAGGTGGAGACGT**TGA**

Appendix CII.4 The forty best matches resulted from an iterated PSI-BLAST search using the protein codified by the round ppa025511mg allele as query against the nr database of NCBI and the UniprotKB/Swiss-Prot database.

	Accession number	Description	% identity	aligned length	evalue	bit score	Database
1	XP_007208681.1	PRUPE_ppa025511mg hypothetical protein [*Prunus persica*]	98.4	750	0	1488	nr NCBI
2	XP_007203972.1	PRUPE_ppa024468mg hypothetical protein [*Prunus persica*]	59.01	832	0	822	nr NCBI
3	XP_007208275.1	PRUPE_ppa015129mg hypothetical protein [*Prunus persica*]	57.2	708	0	712	nr NCBI
4	XP_007207537.1	PRUPE_ppa015767mg hypothetical protein [*Prunus persica*]	55.62	730	0	686	nr NCBI
5	XP_007199245.1	PRUPE_ppa022349mg hypothetical protein [*Prunus persica*]	58.35	641	0	665	nr NCBI
6	XP_004305545.1	PREDICTED: Probable LRR receptor-like protein kinase At1g35710-like [*Fragaria vesca*]	50.87	749	0	643	nr NCBI
7	XP_004305546.1	PREDICTED: receptor-like protein 12-like [*Fragaria vesca*]	52.74	675	0	625	nr NCBI
8	XP_004305135.1	PREDICTED: receptor-like protein 12-like [*Fragaria vesca*]	54.61	661	0	607	nr NCBI
9	XP_004308395.1	PREDICTED: receptor-like protein kinase BRI1-like 3-like [*Fragaria vesca*]	52.43	700	0	594	nr NCBI
10	XP_004305548.1	PREDICTED: receptor-like protein 12-like [*Fragaria vesca*]	51.05	715	0	602	nr NCBI
11	XP_007208498.1	PREDICTED: hypothetical protein PRUPE_ppa026755mg [Prunus persica]	51.74	690	0	593	nr NCBI
12	XP_004305110.1	PREDICTED: leucine-rich repeat receptor-like protein kinase PEPR1-like [*Fragaria vesca*]	52.73	660	0	564	nr NCBI
13	XP_004305547.1	PREDICTED: receptor-like protein 12-like [*Fragaria vesca*]	52.96	642	1,00E-176	549	nr NCBI
14	XP_002270356.2	PREDICTED: LRR receptor-like serine/threonine-protein kinase FLS2-like [*Vitis vinifera*]	47.74	643	3,00E-164	514	nr NCBI
15	XP_006374001.1	hypothetical protein POPTR_0016s12800g [*Populus trichocarpa*]	45.43	733	4,00E-162	507	nr NCBI
16	XP_007026631.1	LRR receptor-like serine/threonine-protein kinase GSO1, putative [*Theobroma cacao*]	46.58	672	1,00E-161	507	nr NCBI
17	XP_003632604.1	PREDICTED: LRR receptor-like serine/threonine-protein kinase GSO2-like [*Vitis vinifera*]	48.33	658	2,00E-159	501	nr NCBI
18	XP_002269481.2	PREDICTED: leucine-rich repeat receptor protein kinase EXS-like [*Vitis vinifera*]	48.06	643	3,00E-158	498	nr NCBI
19	XP_006574212.1	PREDICTED: receptor-like protein 12-like [*Glycine max*]	42.34	725	6,00E-156	491	nr NCBI
20	XP_004304727.1	PREDICTED: receptor-like protein 12-like [*Fragaria vesca*]	52.14	583	1,00E-151	483	nr NCBI

	Accession number	Description	% identity	aligned length	evalue	bit score	Database
21	Q9FIZ3.2	GASSHO 2: EMBRYO SAC DEVELOPMENT ARREST 23	24.85	817	7,00E-143	457	UniprotKB/Swiss
22	Q9LP24.1	Probable leucine-rich repeat receptor-like protein kinase At1g35710	25.58	774	2,00E-141	451	UniprotKB/Swiss
23	C0LGQ5.1	GASSHO 1: LRR receptor-like serine/threonine-protein kinase GSO1	26.26	754	8,00E-141	451	UniprotKB/Swiss
24	Q9SHI2.2	Leucine-rich repeat receptor-like serine/threonine-protein kinase At1g17230	24.06	748	5,00E-138	441	UniprotKB/Swiss
25	O49318.1	Probable leucine-rich repeat receptor-like protein kinase At2g33170	24.97	745	2,00E-137	440	UniprotKB/Swiss
26	P93194.2	Receptor-like protein kinase (Pharbitis nil)	25.13	748	6,00E-137	438	UniprotKB/Swiss
27	Q9FL28.1	FLAGELLIN-SENSITIVE 2: LRR receptor-like serine/threonine-protein kinase FLS2	24	800	3,00E-135	435	UniprotKB/Swiss
28	Q8VZG8.3	Probable LRR receptor-like serine/threonine-protein kinase At4g08850	27.71	682	4,00E-135	432	UniprotKB/Swiss
29	Q9FZ59.1	Leucine-rich repeat receptor-like protein kinase PEPR2	24.6	752	1,00E-133	429	UniprotKB/Swiss
30	Q9FRS6.1	PHLOEM INTERCALATED WITH XYLEM-LIKE 1: LRR receptor-like protein kinase PXL1	27.33	666	1,00E-131	422	UniprotKB/Swiss
31	O49545.1	Leucine-rich repeat receptor-like serine/threonine-protein kinase BAM1	28.62	650	1,00E-131	421	UniprotKB/Swiss
32	Q9ZPS9.1	Serine/threonine-protein kinase BRI1-like 2	26.12	739	1,00E-130	422	UniprotKB/Swiss
33	Q9LYN8.1	EXCESS MICROSPOROCYTES 1: Leucine-rich repeat receptor protein kinase EXS (EMS1)	25.1	721	2,00E-130	423	UniprotKB/Swiss
34	P0DL10.1	CLAVATA1-like protein:Leucine-rich repeat receptor-like kinase protein	26.32	646	7,00E-130	416	UniprotKB/Swiss
35	Q9SSL9.1	Leucine-rich repeat receptor-like protein kinase PEPR1	23.43	734	3,00E-129	418	UniprotKB/Swiss
36	Q9ZWC8.1	Serine/threonine-protein kinase BRI1-like 1	25.78	768	5,00E-128	416	UniprotKB/Swiss
37	Q9LVP0.1	Probable leucine-rich repeat receptor-like protein kinase At5g63930	26.5	732	7,00E-125	406	UniprotKB/Swiss
38	Q9SYQ8.3	Receptor protein kinase CLAVATA1	24.59	663	1,00E-124	403	UniprotKB/Swiss
39	Q9SGP2.1	HAESA-LIKE1:Receptor-like protein kinase HSL1	27.34	640	2,00E-124	402	UniprotKB/Swiss
40	Q5Z9N5.1	Leucine-rich repeat receptor-like kinase protein FLORAL ORGAN NUMBER1	25	430	6,00E-119	388	UniprotKB/Swiss

Appendix CII.5 Annotated SNPs and identified SNPs in the studied region of 26.75Kb onLG 6. SNP_IGA are already annotated SNPs in the peach genome. Asterisks are those identified in our samples.

SNP_ID	Position	Transcript
SNP_IGA-688382	24390454	ppa015129m
SNP_IGA-688383	24390455	ppa015129m
SNP_IGA-688386	24391430	ppa015129m
SNP_IGA-688412*	24398129	ppa024472m
SNP_ppa024472_1	24398217	ppa024472m
SNP_ppa024472_2	24398230	ppa024472m
SNP_ppa024472_3	24398262	ppa024472m
SNP_IGA-688415	24398263	ppa024472m
SNP_IGA-688416*	24398407	ppa024472m
SNP_IGA-688417*	24398563	ppa024472m
SNP_IGA-688419*	24399227	ppa024472m
SNP_IGA-688420*	24399316	ppa024472m
SNP_ppa024472_4	24399504	ppa024472m
SNP_ppa024472_5	24399505	ppa024472m
SNP_ppa024472_6	24399534	ppa024472m
SNP_ppa024472_7	24399561	ppa024472m
SNP_IGA-688424	24399562	ppa024472m
SNP_ppa025511_1	24406522	ppa025511m
SNP_ppa025511_2	24406523	ppa025511m
SNP_ppa025511_3	24406600	ppa025511m
SNP_ppa025511_4	24406601	ppa025511m
SNP_ppa025511_5	24406733	ppa025511m
SNP_IGA-688461	24406745	ppa025511m
SNP_ppa025511_6	24406753	ppa025511m
SNP_ppa025511_7	24406799	ppa025511m
SNP_ppa025511_8	24406849	ppa025511m
SNP_IGA-688463	24406892	ppa025511m
SNP_ppa025511_9	24406900	ppa025511m
SNP_ppa025511_10	24407078	ppa025511m
SNP_IGA-688466	24407160	ppa025511m
SNP_IGA-688467	24407178	ppa025511m
SNP_ppa025511_11	24407180	ppa025511m
SNP_IGA-688468	24407187	ppa025511m
SNP_IGA-688469	24407201	ppa025511m
SNP_ppa025511_12	24407465	ppa025511m
SNP_ppa025511_13	24407508	ppa025511m
SNP_IGA-688487*	24411905	ppa015767m
SNP_ppa015767_1	24411919	ppa015767m
SNP_ppa015767_2	24411935	ppa015767m
SNP_IGA-688490*	24412019	ppa015767m
SNP_ppa015767_3	24412132	ppa015767m
SNP_IGA-688514	24415159	ppa023752m

APPENDICES CHAPTER III

APPENDIX CIII.1 Script code written in Shell (Linux) language to perform the quality and trimming assessment.

```
#!/bin/bash -x

# # to submit sbatch, sinfo, scancel, squeue

# We name the job:

#SBATCH --job-name=Quality_and_trimming

#How many tasks we need

#SBATCH --ntasks-per-node=1

# # SBATCH --nodes=1

# Additional options:

# Limited working time.

# # SBATCH --time=24:45:0

# Self explanatory

#SBATCH --mem-per-cpu=40000M

# Needed space in /tmp

# # SBATCH --tmp=1000M

# #SBATCH --nodelist=node004

date

source /opt/Modules/3.2.9/init/Modules4bash.sh

module load FastQC-0.10.0

module load fastx-0.0.13-sl6

WORKFOLD=/projects/061-SECUENCIAS-Pd_Pp/Peach

RAWFOLD=/projects/061-SECUENCIAS-Pd_Pp/Peach/Raw_data

TRIMMEDFOLD=/projects/061-SECUENCIAS-Pd_Pp/Peach/Raw_data/Trimmed_q30

QUALTRIMMEDFOLD=/projects/061-SECUENCIAS-
Pd_Pp/Peach/Raw_data/Trimmed_q30/Quality_reports

CUTADAPT=/home/elopez/Software/cutadapt-1.2.1/bin/cutadapt

QTHR=30

MINL=35

cd $RAWFOLD

for file in *.fastq.gz

      do

      basename=`echo $file | sed 's/.fastq.gz//'`
```

```
      $CUTADAPT -b GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG -b
ACACTCTTTCCCTACACGACGCTCTTCCGATCT -b
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT -b
CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT -b
ACACTCTTTCCCTACACGACGCTCTTCCGATCT -b
CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT -O 6 -m $MINL --quality-base=33
$file | fastq_quality_trimmer -t $QTHR -l $MINL -Q 33 -v -z -o
$TRIMMEDFOLD/$basename.trimmed_q30.fastq.gz

done

cd $TRIMMEDFOLD

for trimmed in *.trimmed_q30.fastq.gz

      do

      fastqc --nogroup -o $QUALTRIMMEDFOLD/ -f fastq $trimmed

done

date

exit 0
```

APPENDIX CIII.2 Script code written in Shell (Linux) language to mate reads contained into each of the two trimmed fastq files for each sample.

A. Shell script to mate trimmed fastq files

```
#!/bin/bash -x

# We name the job:

#SBATCH --job-name=Trimming_test

#How many tasks we need

#SBATCH --ntasks-per-node=1

# # SBATCH --nodes=1

# Limited working time.

# # SBATCH --time=24:45:0

# Self explanatory

#SBATCH --mem-per-cpu=45000

# # SBATCH --tmp=1000M

# #SBATCH --nodelist=node004

date

source /opt/Modules/3.2.9/init/Modules4bash.sh

module load FastQC-0.10.0
```

```
TRIMMEDUNPAIREDDIR=/scratch/061-SECUENCIAS-
Pd_Pp/Peach/Raw_data/Trimmed_q30

TRIMMEDPAIREDDIR=/scratch/061-SECUENCIAS-
Pd_Pp/Peach/Raw_data/Trimmed_q30/Trimmed_q30_and_paired

PAIRINGSCRIPT=/home/elopez/Software/Scripts/read_mater_light_db.pl

cd $TRIMMEDUNPAIREDDIR

perl $PAIRINGSCRIPT $1 $2 $3 $4 $5

date

exit 0
```

B. Perl script to perform the matting between the reads contained into two trimmed pair ends sequences and generates a file containing all the single reads that have not been possible to mate.

```
#!/usr/bin/perl

#read_mater_light_db.pl

#This script identifies mated reads in two independent files and
generates a file with single reads

use strict;

use warnings;

no warnings 'uninitialized';

use Benchmark ':hireswallclock';

use IO::Zlib;

## Input files

my $infile1=$ARGV[0];

my $infile2=$ARGV[1];

## Output files

my $outfile1=$ARGV[2];

my $outfile2=$ARGV[3];

my $single_reads=$ARGV[4];

chomp $infile1;

chomp $infile2;

chomp $outfile1;

chomp $outfile2;

chomp $single_reads;
```

```
my $t0=Benchmark->new;

my $mode=undef;

## Classification of Illumina sequence identifiers ##

if ($infile1 =~ /\.gz$/) {

         tie      *IN1,'IO::Zlib',$infile1, "rb” ;

      }

      else {

         open(IN1, "<$infile1") or die "Couldn't open $infile1"};

#open (IN1, "<$infile1") or die "Couldn't open $infile1";

my $first_line=<IN1>;

close IN1;

chomp $first_line;

our $is_illumina=IS_ILLUMINA($first_line);

## Hash1 population

my %hash1;

my $hash1_file=$infile1.'hash1';

dbmopen(%hash1, "$hash1_file", $mode);

my $file1_lc=0;

my $current_seq;

#open (IN1, "<$infile1") or die "Couldn't open $infile1";

if ($infile1 =~ /\.gz$/) {tie *IN1,'IO::Zlib',$infile1,"rb";} else
{open(IN1, "<$infile1") or die "Couldn't open $infile1"};

while ($current_seq=<IN1>) {

      $file1_lc++;

      chomp $current_seq;

      my $seq_id1=LIGHTID($current_seq);

      $hash1{$seq_id1}=$file1_lc;

      for (1..3) {

            my $void_var=<IN1>;

            $file1_lc++;

      }

}

close IN1;
```

```
my $t1=Benchmark->new;

my $td1=timediff($t1, $t0);

print "Hash1 generated in ",timestr($td1),"\n";

## Hash2 population

my %hash2;

my $hash2_file=$infile2.'hash2';

dbmopen(%hash2, "$hash2_file", $mode);

my $file2_lc=0;

#open (IN2, "<$infile2") or die "Couldn't open $infile2";

if ($infile2 =~ /\.gz$/) {tie *IN2,'IO::Zlib',$infile2,"rb";} else
{open(IN2, "<$infile2") or die "Couldn't open $infile2"};

while ($current_seq=<IN2>) {

$file2_lc++;

      chomp $current_seq;

      my $seq_id2=LIGHTID($current_seq);

      $hash2{$seq_id2}=$file2_lc;

      for (1..3) {

            my $void_var=<IN2>;

            $file2_lc++;

      }

}

close IN2;

my $t2=Benchmark->new;

my $td2=timediff($t2, $t1);

print "Hash2 generated in ",timestr($td2),"\n";

## Comparison of hashes

my %paired;

my @paired_positions1;

my @paired_positions2;

my @singles1;

my @singles2;

foreach my $key1 (sort {$hash1{$a} <=> $hash1{$b}} keys %hash1) {

      if (exits $hash2{$key1}) {
```

```
            push @paired_positions1, $hash1{$key1};
            push @paired_positions2, $hash2{$key1};
      } else {push @singles1, $hash1{$key1};}
}
my $t3=Benchmark->new;
my $td3=timediff($t3, $t2);
print scalar (@paired_positions1)," paired reads\n";
#print scalar (@paired_positions2)," paired reads\n";
print scalar(@singles1)," single reads in hash1\n";
print "Comparison1 of hashes generated in ",timestr($td3),"\n";
foreach my $key2 (sort {$hash2{$a} <=> $hash2{$b}} keys %hash2) {
      unless (exits $hash1{$key2}) {push @singles2, $hash2{$key2};}
}
my $t4=Benchmark->new;
my $td4=timediff($t4, $t3);
print scalar(@singles2)," single reads in hash2\n";
print "Comparison2 of hashes generated in ",timestr($td4),"\n";
%hash1=();
dbmclose(%hash1);
%hash2=();
dbmclose(%hash2);
my @sorted_paired_positions1=sort{$a<=>$b} @paired_positions1;
my @sorted_paired_positions2=sort{$a<=>$b} @paired_positions2;
my @sorted_single_positions1=sort{$a<=>$b} @singles1;
my @sorted_single_positions2=sort{$a<=>$b} @singles2;
## Output paired 1
print "Generating $outfile1\n";
#open (IN1, "<$infile1") or die "Couldn't open $infile1";
if ($infile1 =~ /\.gz$/) {tie *IN1,'IO::Zlib',$infile1,"rb";} else
{open(IN1, "<$infile1") or die "Couldn't open $infile1"};
#open (OUT1, ">$outfile1") or die "Couldn't save in $outfile1";
if ($outfile1 =~ /\.gz$/) {tie *OUT1,'IO::Zlib',$outfile1,"wb";} else
{open(OUT1, ">$outfile1") or die "Couldn't save in $outfile1"};
```

```
my $lc1=1;

my $line_to_print1=shift(@sorted_paired_positions1);

while (my $line_in_process1=<IN1>) {

      if ($lc1==$line_to_print1) {

            print OUT1 $line_in_process1;

            for (2..4) {

                  my $tmp_line=<IN1>;

                  print OUT1 $tmp_line;

            };

            $lc1=$lc1+4;

            $line_to_print1=shift(@sorted_paired_positions1);

      } else {$lc1++;}

}

close IN1;

close OUT1;

my $t5=Benchmark->new;

my $td5=timediff($t5, $t4);

print "$outfile1 generated in ", timestr($td5),"\n";

## Output paired 2

print "Generating $outfile2\n";

#open (IN2, "<$infile2") or die "Couldn't open $infile2";

if ($infile2 =~ /\.gz$/) {tie *IN2,'IO::Zlib',$infile2,"rb";} else
{open(IN2, "<$infile2") or die "Couldn't open $infile2"};

#open (OUT2, ">$outfile2") or die "Couldn't save in $outfile2";

if ($outfile2 =~ /\.gz$/) {tie *OUT2,'IO::Zlib',$outfile2,"wb";} else
{open(OUT2, ">$outfile2") or die "Couldn't save in $outfile2"};

my $lc2=1;

my $line_to_print2=shift(@sorted_paired_positions2);

while (my $line_in_process2=<IN2>) {

      if ($lc2==$line_to_print2) {

            print OUT2 $line_in_process2;

            for (2..4) {

                  my $tmp_line=<IN2>;
```

```
                  print OUT2 $tmp_line;

            };

            $lc2=$lc2+4;

            $line_to_print2=shift(@sorted_paired_positions2);

      } else {$lc2++;}
}

close IN2;

close OUT2;

my $t6=Benchmark->new;

my $td6=timediff($t6, $t5);

print "$outfile2 generated in ", timestr($td6),"\n";

## Output singles

print "Generating $single_reads\n";

#open (IN1, "<$infile1") or die "Couldn't open $infile1";

if ($infile1 =~ /\.gz$/) {tie *IN1,'IO::Zlib',$infile1,"rb";} else
{open(IN1, "<$infile1") or die "Couldn't open $infile1"};

#open (SINGLE, ">$single_reads") or die "Couldn't save in
$single_reads";

if ($single_reads =~ /\.gz$/) {tie
*SINGLE,'IO::Zlib',$single_reads,"wb";} else {open(SINGLE,
">$single_reads") or die "Couldn't save in $single_reads"};

$lc1=1;

$line_to_print1=shift(@sorted_single_positions1);

while (my $line_in_process1=<IN1>) {

      if ($lc1==$line_to_print1) {

            print SINGLE $line_in_process1;

            for (2..4) {

                  my $tmp_line=<IN1>;

                  print SINGLE $tmp_line;

            };

            $lc1=$lc1+4;

            $line_to_print1=shift(@sorted_single_positions1);

      } else {$lc1++;}

}
```

```
close IN1;

#open (IN2, "<$infile2") or die "Couldn't open $infile2";

if ($infile2 =~ /\.gz$/) {tie *IN2,'IO::Zlib',$infile2,"rb";} else
{open(IN2, "<$infile2") or die "Couldn't open $infile2"};

$lc2=1;

$line_to_print2=shift(@sorted_single_positions2);

while (my $line_in_process2=<IN2>) {

      if ($lc2==$line_to_print2) {

            print SINGLE $line_in_process2;

            for (2..4) {

                  my $tmp_line=<IN2>;

                  print SINGLE $tmp_line;

            };

            $lc2=$lc2+4;

            $line_to_print2=shift(@sorted_single_positions2);

      } else {$lc2++;}

}

my $t7=Benchmark->new;

my $td7=timediff($t7, $t6);

print "$single_reads generated in ",timestr($td7),"\n";

my $tdf=timediff($t7, $t0);

print "Overall process: ",timestr($tdf),"\n";

exit;

sub IS_ILLUMINA {

      my $first_header=shift;

      my @header=split(/:/,$first_header);

      my $illumina_assesment;

      if (scalar @header==5) {$illumina_assesment='1'}

      elsif (scalar @header==10) {$illumina_assesment='0'}

      else {die "Your sequence file is not in a valid format.\n$!\n"};

      return $illumina_assesment;

}

sub B2GB {
```

```
	my $result_in_bytes=shift;
	chomp $result_in_bytes;
	my $result_in_Gb=$result_in_bytes/(1024**3);
	return $result_in_Gb;
}
sub LIGHTID {
	my $id_to_process=shift;
	my @current_seq=split(/:/,$current_seq);
	my $light_id;
	if ($is_illumina=='1') {
		$current_seq[4]=substr($current_seq[4],0,-2);

	$light_id=$current_seq[2].':'.$current_seq[3].':'.$current_seq[4
];
	}
	elsif ($is_illumina=='0') {
		$current_seq[6]=substr($current_seq[6],0,-2);

	$light_id=$current_seq[4].':'.$current_seq[5].':'.$current_seq[6
];
	}
	else {die "Unexpected error. Check your sequences format\n"};
	return $light_id;
}
```

APPENDIX CIII.3 Script code written in Shell (Linux) language to map reads against the peach reference genome.

```
#!/bin/bash -x
# We name the job:
#SBATCH --job-name=Alingh_peach
#How many tasks we need
#SBATCH --ntasks-per-node=2
# #SBATCH --nodelist=node003
#Additional options
#SBATCH --mem-per-cpu=40G
```

```
# # SBATCH --partition=fatnodes

# # SBATCH --tmp=1000M

date

source /opt/Modules/3.2.9/init/Modules4bash.sh

module load /bwa/0.6.2

module load samtools-0.1.18-sl61

module load perl-libs-5.10

module load vcftools-0.1.7

WORKFOLD=/scratch/061-SECUENCIAS-Pd_Pp/Peach/Raw_data/Trimmed_q30/

TRIMEDPAIREDFOLDER=/scratch/061-SECUENCIAS
Pd_Pp/Peach/Raw_data/Trimmed_q30/Trimmed_q30_and_paired

BWA_DIR=/scratch/061-SECUENCIAS-
Pd_Pp/Peach/Raw_data/Trimmed_q30/BWA_output

GENOME=/projects/061-SECUENCIAS-
Pd_Pp/Reference/Prunus_persica.main_genome.scaffolds.fasta

#mkdir $BWA_DIR

#Create bwa index

#bwa index -a bwtsw $GENOME

#Perform the BWA algorithm to map reads on the reference genome

x = n° of differences

for x in 1 3 5 2 4;

do

base=`echo $1 | sed 's/.paired.sam//'`

basename1=`echo $1 | sed 's/.fastq.gz//'`

basename2=`echo $2 | sed 's/.fastq.gz//'`

basename3=`echo $3 | sed 's/.fastq.gz//'`

/opt/bwa/bwa aln -t 2 $GENOME $TRIMEDPAIREDFOLDER/$1 >
$BWA_DIR/$basename1.sai

/opt/bwa/bwa aln -t 2 $GENOME $TRIMEDPAIREDFOLDER/$2 >
$BWA_DIR/$basename2.sai

/opt/bwa/bwa aln -t 2 $GENOME $TRIMEDPAIREDFOLDER/$3 >
$BWA_DIR/$basename3.sai

#Run BWA for pair-end

#/opt/bwa/bwa sampe $GENOME $BWA_DIR/$basename1.sai
$BWA_DIR/$basename2.sai $TRIMEDPAIREDFOLDER/$1 $TRIMEDPAIREDFOLDER/$2
> $BWA_DIR/$base.paired.sam
```

```
#/opt/bwa/bwa samse $GENOME $BWA_DIR/$basename3.sai
$TRIMEDPAIREDFOLDER/$3 > $BWA_DIR/$base.single.sam

samtools view -Sb $BWA_DIR/$base.paired.sam >
$BWA_DIR/$base.paired.bam

samtools view -Sb $BWA_DIR/$base.single.sam >
$BWA_DIR/$base.single.bam

samtools view -Sbq 1 $BWA_DIR/$base.paired.sam >
$BWA_DIR/$base.paired.unique.bam

samtools view -Sbq 1 $BWA_DIR/$base.single.sam >
$BWA_DIR/$base.single.unique.bam

samtools merge $BWA_DIR/$base.bam $BWA_DIR/$base.paired.bam
$BWA_DIR/$base.single.bam

samtools merge $BWA_DIR/$base.unique.bam
$BWA_DIR/$base.paired.unique.bam $BWA_DIR/$base.single.unique.bam

samtools sort $BWA_DIR/$base.bam $BWA_DIR/$base.sorted

samtools sort $BWA_DIR/$base.unique.bam $BWA_DIR/$base.unique.sorted

samtools rmdup -S $BWA_DIR/$base.sorted.bam
$BWA_DIR/$base.sorted_rmdup.bam

samtools rmdup -S $BWA_DIR/$base.unique.sorted.bam
$BWA_DIR/$base.unique.sorted_rmdup.bam

samtools index $BWA_DIR/$base.sorted_rmdup.bam

samtools index $BWA_DIR/$base.unique.sorted_rmdup.bam
```

#Add RG

```
java -jar /opt/picard-tools-1.56/AddOrReplaceReadGroups.jar
I=$BWA_DIR/$base.sorted_rmdup.bam O=$BWA_DIR/$base.sorted_rmdup.RG.bam
ID="$base" LB=1 PL=illumina PU=1 SM="$base"
VALIDATION_STRINGENCY=SILENT

java -jar /opt/picard-tools-1.56/AddOrReplaceReadGroups.jar
I=$BWA_DIR/$base.unique.sorted_rmdup.bam
O=$BWA_DIR/$base.unique.sorted_rmdup.RG.bam ID="$base" LB=1
PL=illumina PU=1 SM="$base" VALIDATION_STRINGENCY=SILENT
```

#Index BAM

```
samtools index $BWA_DIR/$base.sorted_rmdup.RG.bam

samtools index $BWA_DIR/$base.unique.sorted_rmdup.RG.bam

samtools rmdup -S $BWA_DIR/62MF4AAXX_"$i".sort.bam/
$BWA_DIR/62MF4AAXX_"$i".sort_rmdup.bam

samtools index $BWA_DIR/62MF4AAXX_"$i".sort_rmdup.bam
```

#Create a bam file with all pair-end reads to SV

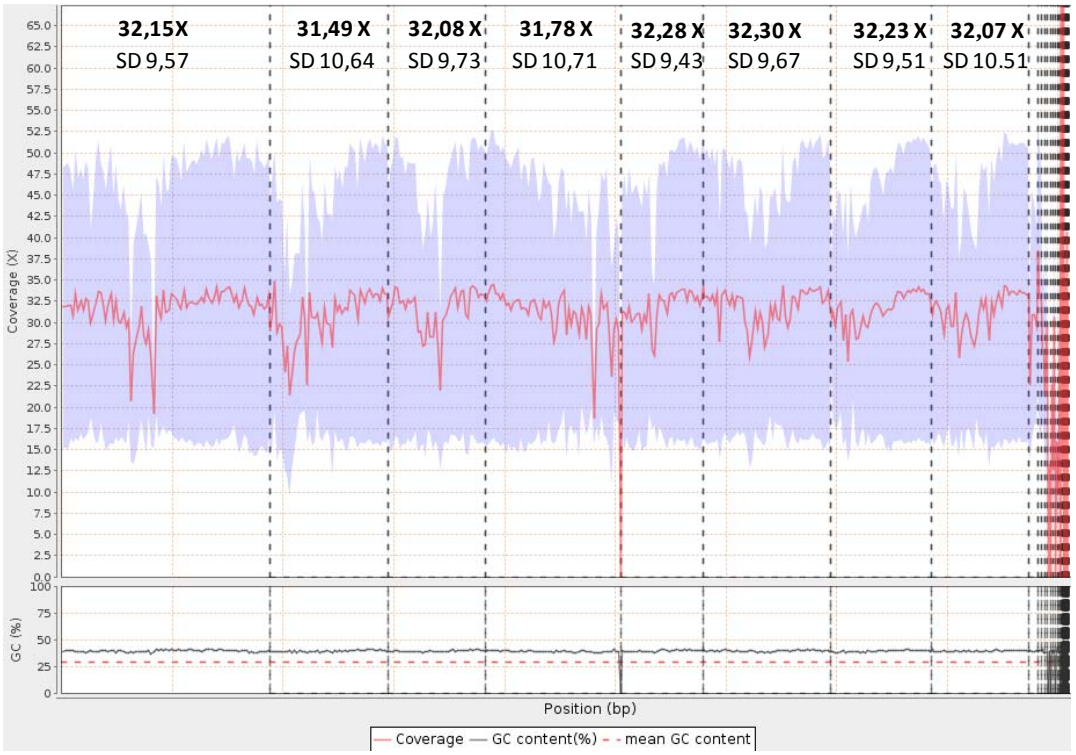
```
      egrep "(^@|XT:A:U)" $BWA_DIR/D0ACXX_"$i".sam >
$BWA_DIR/D0ACXX_"$i"_uniq.sam

      samtools view -Sb $BWA_DIR/D0ACXX_"$i"_uniq.sam >
$BWA_DIR/D0ACXX_"$i"_uniq_reads.bam

      samtools sort $BWA_DIR/D0ACXX_"$i"_uniq_reads.bam
$BWA_DIR/D0ACXX_"$i"_uniq_reads.sort

      samtools index $BWA_DIR/D0ACXX_"$i"_uniq_reads.sort.bam

      samtools rmdup -S $BWA_DIR/D0ACXX_"$i"_uniq_reads.sort.bam
$BWA_DIR/D0ACXX_"$i"_uniq_reads.sort.pcr_rem.bam

done;

date

exit 0
```

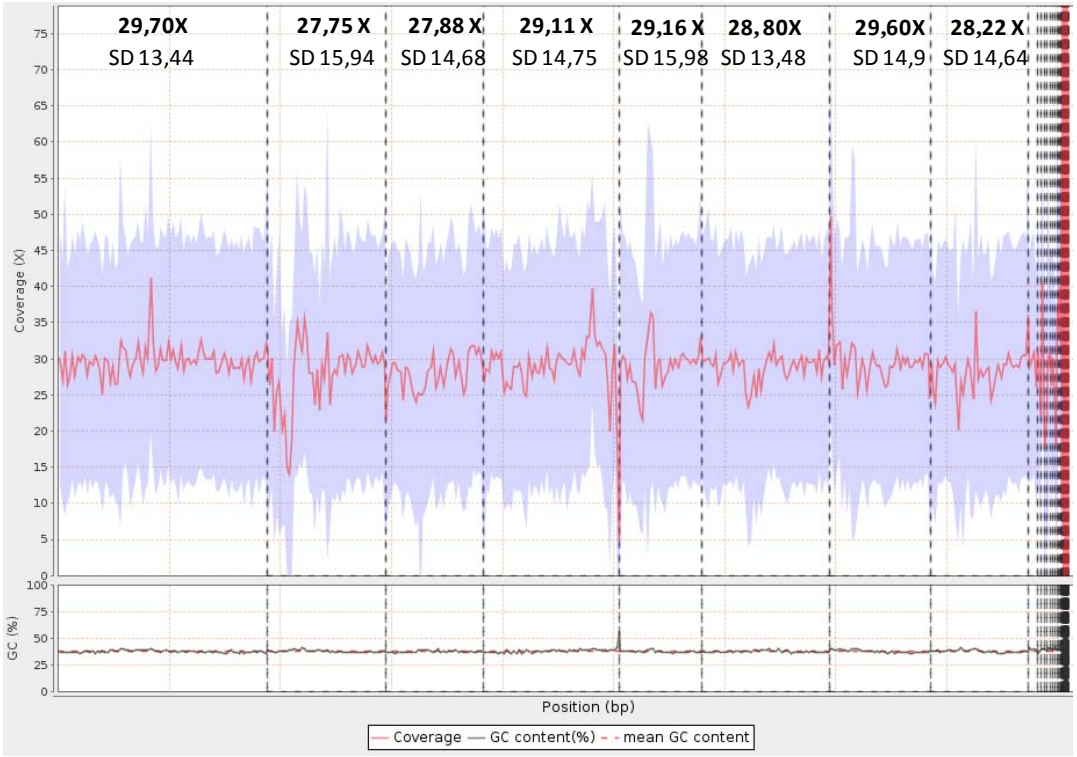
APPENDIX CIII.4 Script code written in Shell (Linux) language to perform the mpileup small variant calling.

```
#!/bin/bash -x

# We name the job:

#SBATCH --job-name=Alingh_peach

#How many tasks we need

#SBATCH --ntasks-per-node=2

# #SBATCH --nodelist=node003

#Additional options

#SBATCH --mem-per-cpu=40G

# # SBATCH --partition=fatnodes

# # SBATCH --tmp=1000M

date

source /opt/Modules/3.2.9/init/Modules4bash.sh

module load /bwa/0.6.2

module load samtools-0.1.18-sl61

module load perl-libs-5.10

module load vcftools-0.1.7

WORKFOLD=/scratch/061-SECUENCIAS-Pd_Pp/Peach/Raw_data/Trimmed_q30/

TRIMEDPAIREDFOLDER=/scratch/061-SECUENCIAS-
Pd_Pp/Peach/Raw_data/Trimmed_q30/Trimmed_q30_and_paired

BWA_DIR=/scratch/061-SECUENCIAS-
Pd_Pp/Peach/Raw_data/Trimmed_q30/BWA_output
```

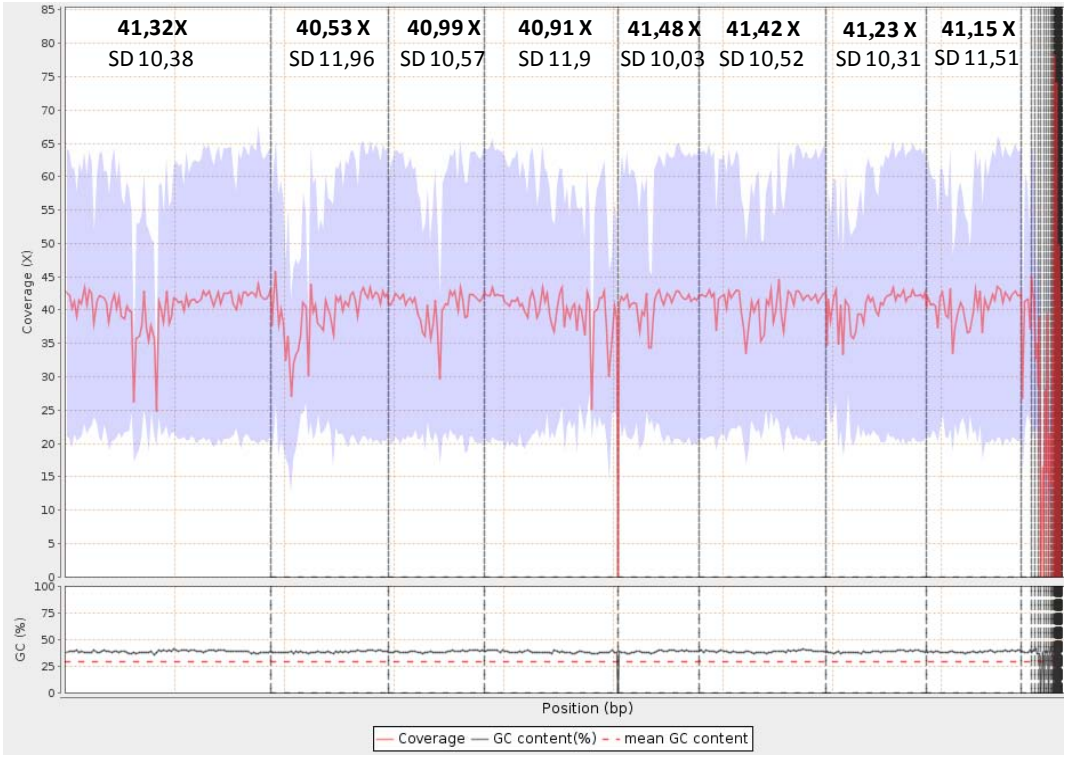
```
GENOME=/projects/061-SECUENCIAS-
Pd_Pp/Reference/Prunus_persica.main_genome.scaffolds.fasta

BCFTOOLFOLD=/opt/samtools/bcftools

base=`echo $1 | sed 's/.unique.sorted_rmdup.RG.bam//'`

#samtools mpileup -Q1 -uDf $WORKFOLD/$GENOME
$BWA_DIR/62MF4AAXX_"$x"_uniq_reads.sort.rmdup.RG.bam |
$BCFTOOLFOLD/bcftools view -Ncvg - > $BWA_DIR/62MF4AAXX_uniq_"$x".vcf

samtools mpileup -Q1 -uDf $GENOME $BWA_DIR/$1 | $BCFTOOLFOLD/bcftools
view -Ncvg - > $BWA_DIR/$base.uniqQ1vcf

date

exit 0
```

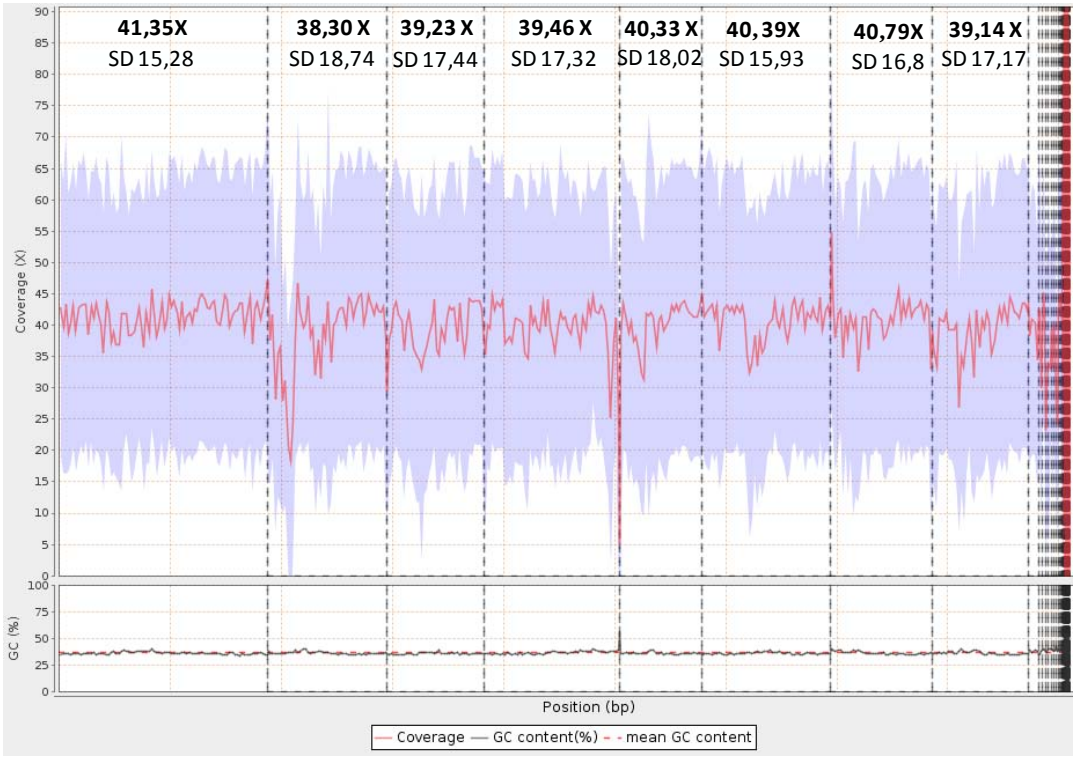
APPENDIX CIII.5 Coverage across each sample's alignment. Upper figures provide the coverage distribution (red line), coverage deviation across the reference sequence and the mean coverage at each chromosome with its standard deviation. The lower figures show the GC content across reference (black line) with its average value (red dotted line). The black vertical dotted line represents the chromosome limits. The first and second plots show the coverage distribution outside/inside gene regions across the reference respectively.



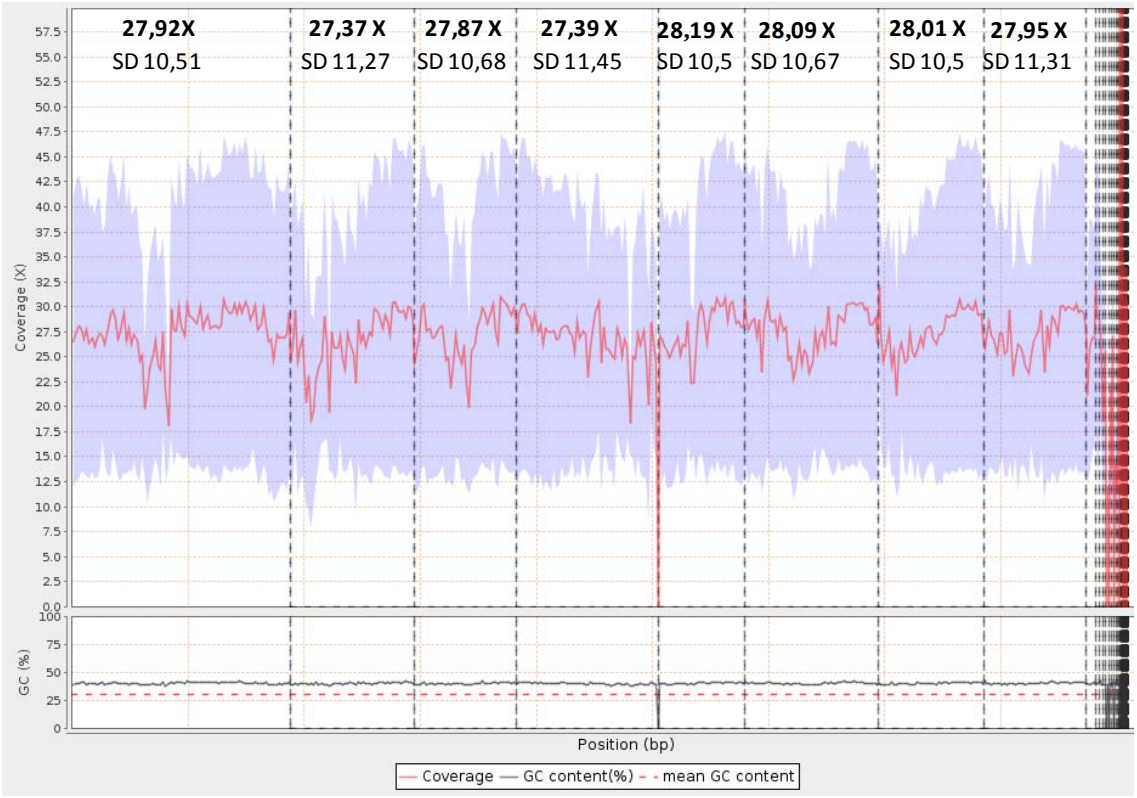


Flameprince_Pearson_peach

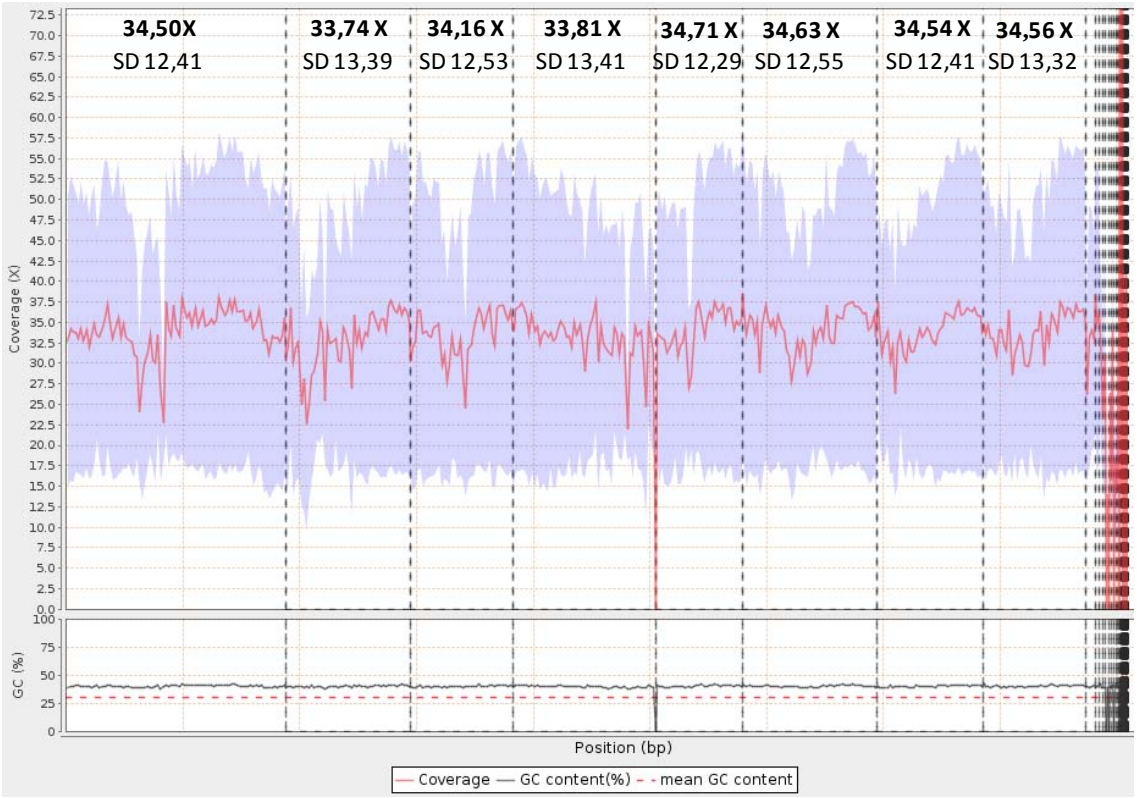


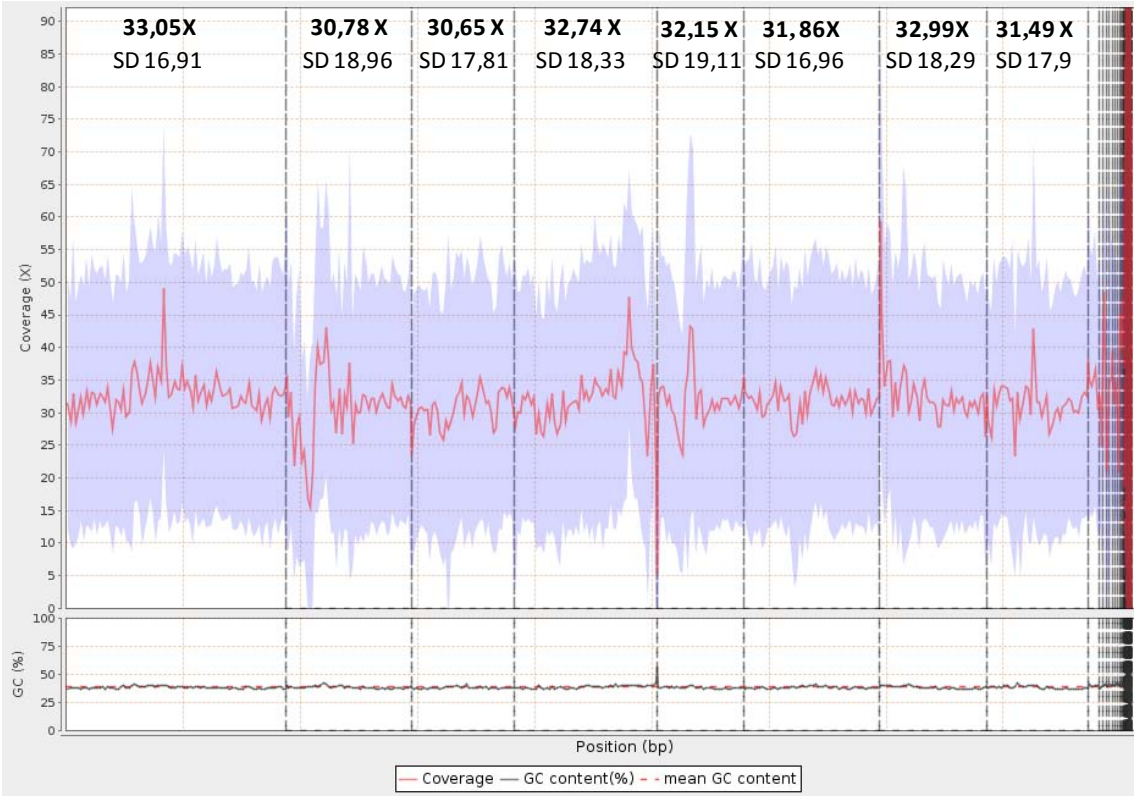


Flameprince_Ham_nectarine_1

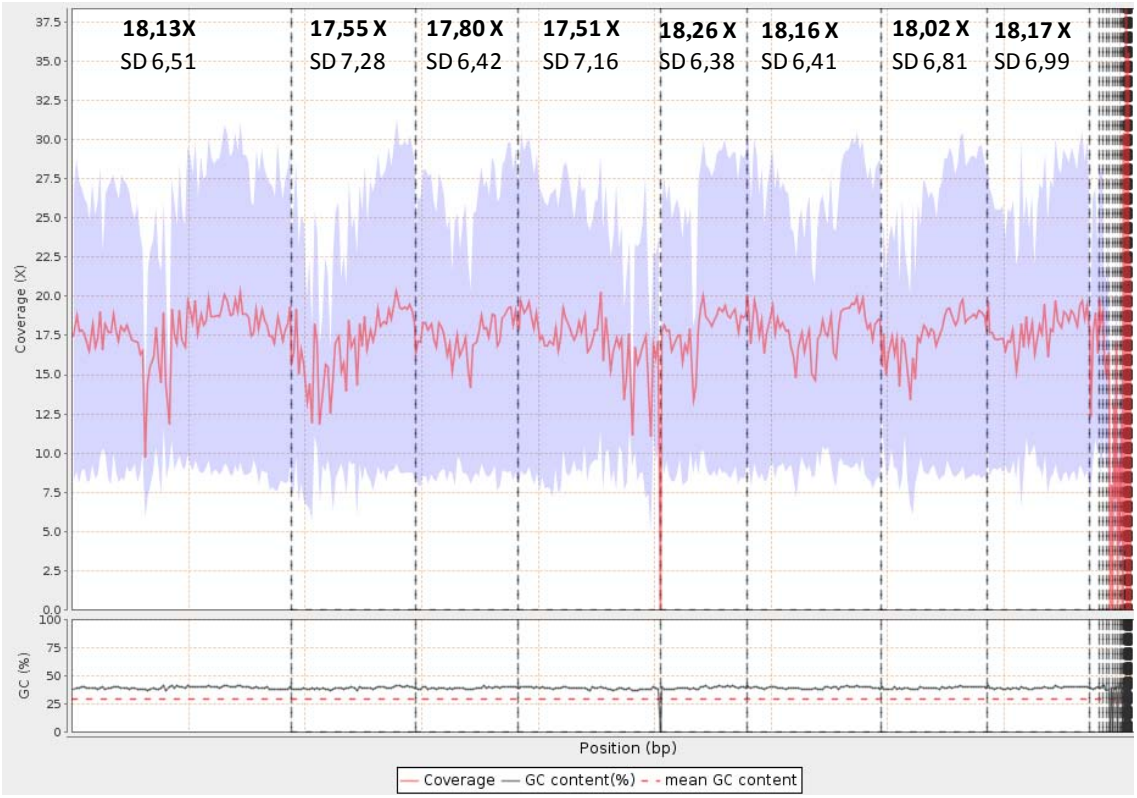


Flameprince_Pearson_nectarine

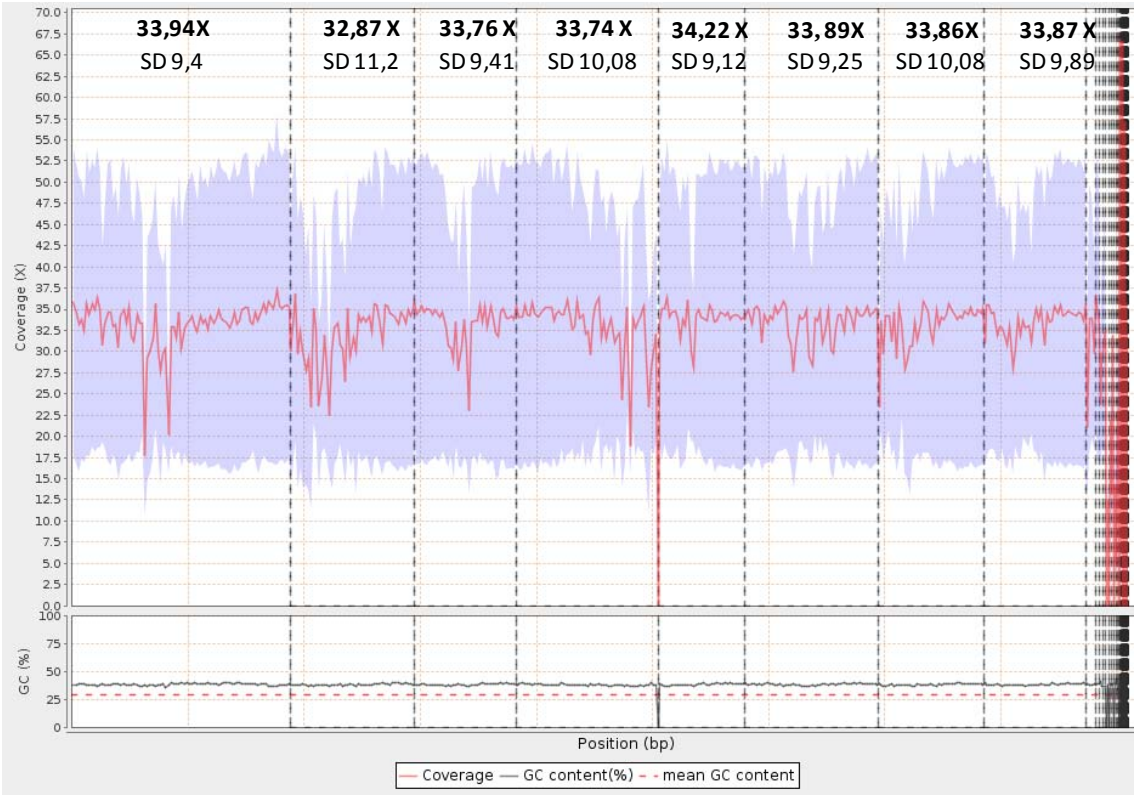


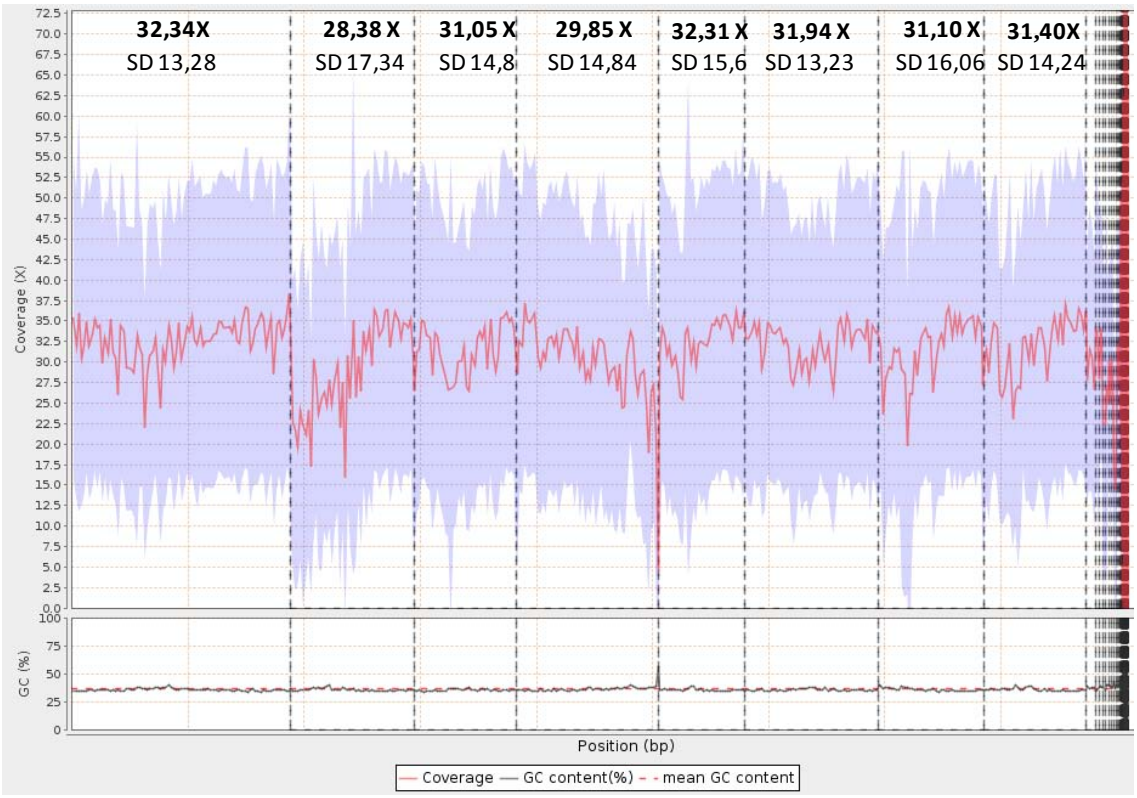


Flameprince_Ham_nectarine_2

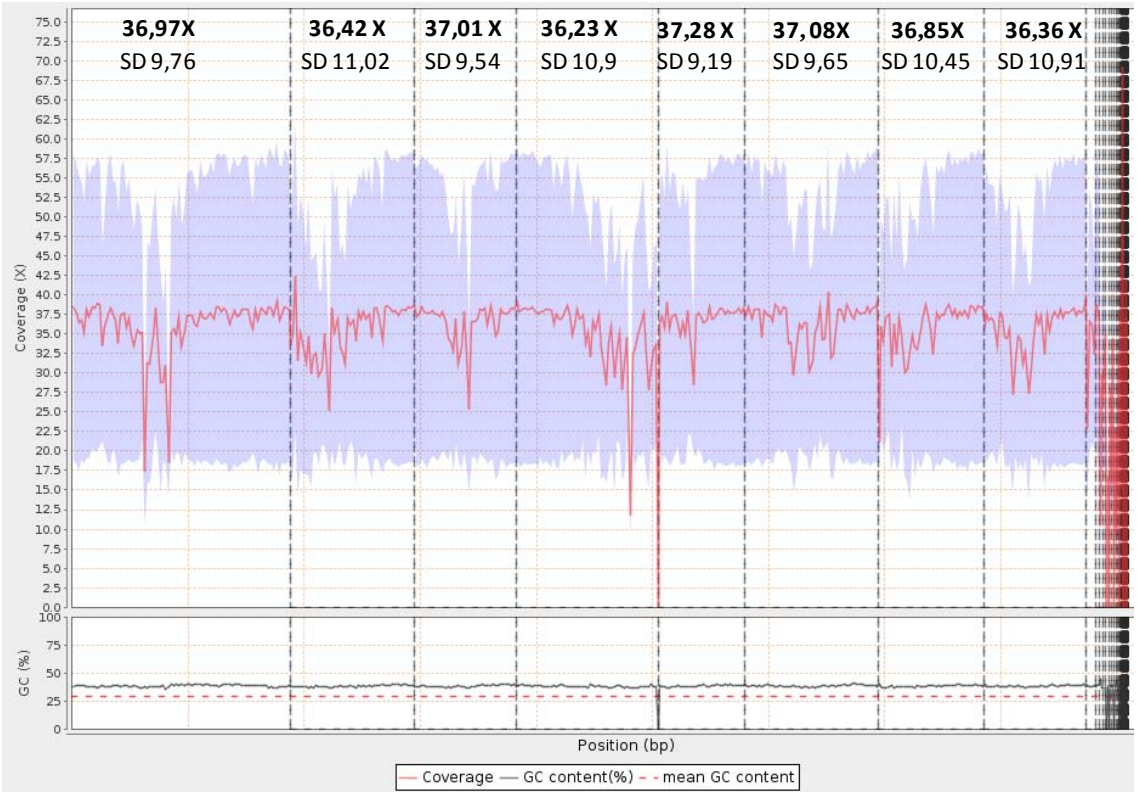


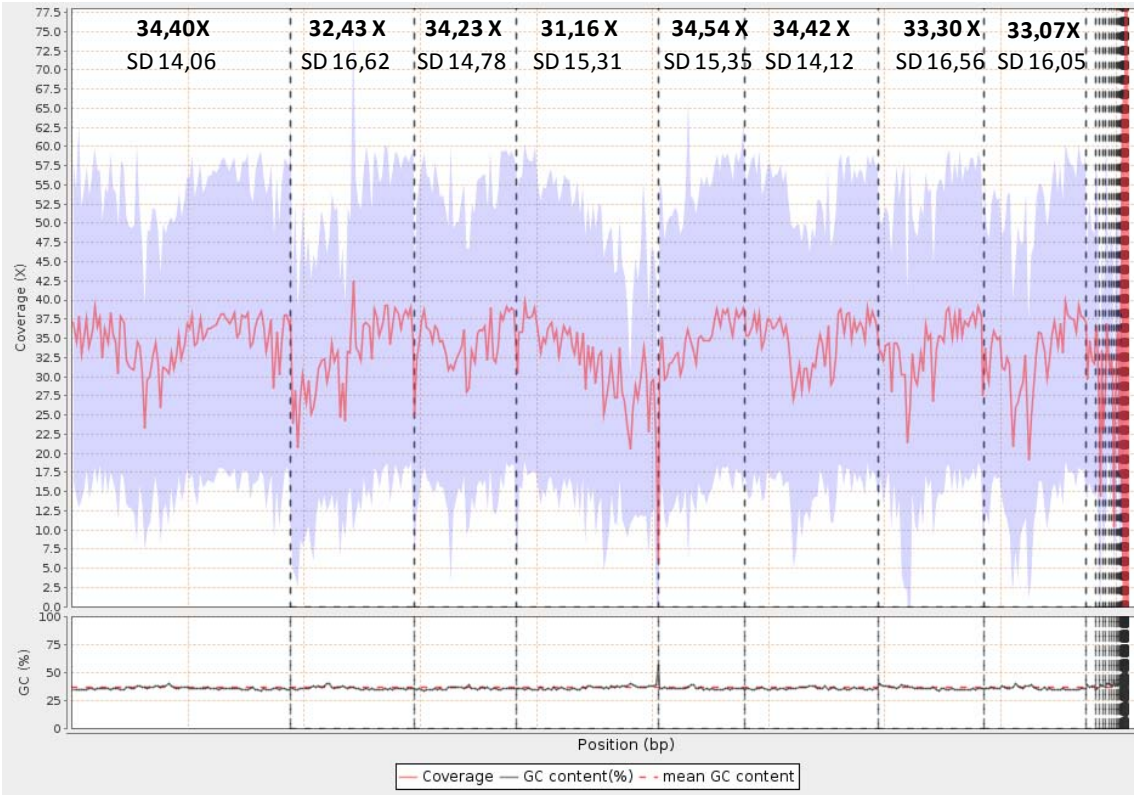
Julyprince_Pearson_peach



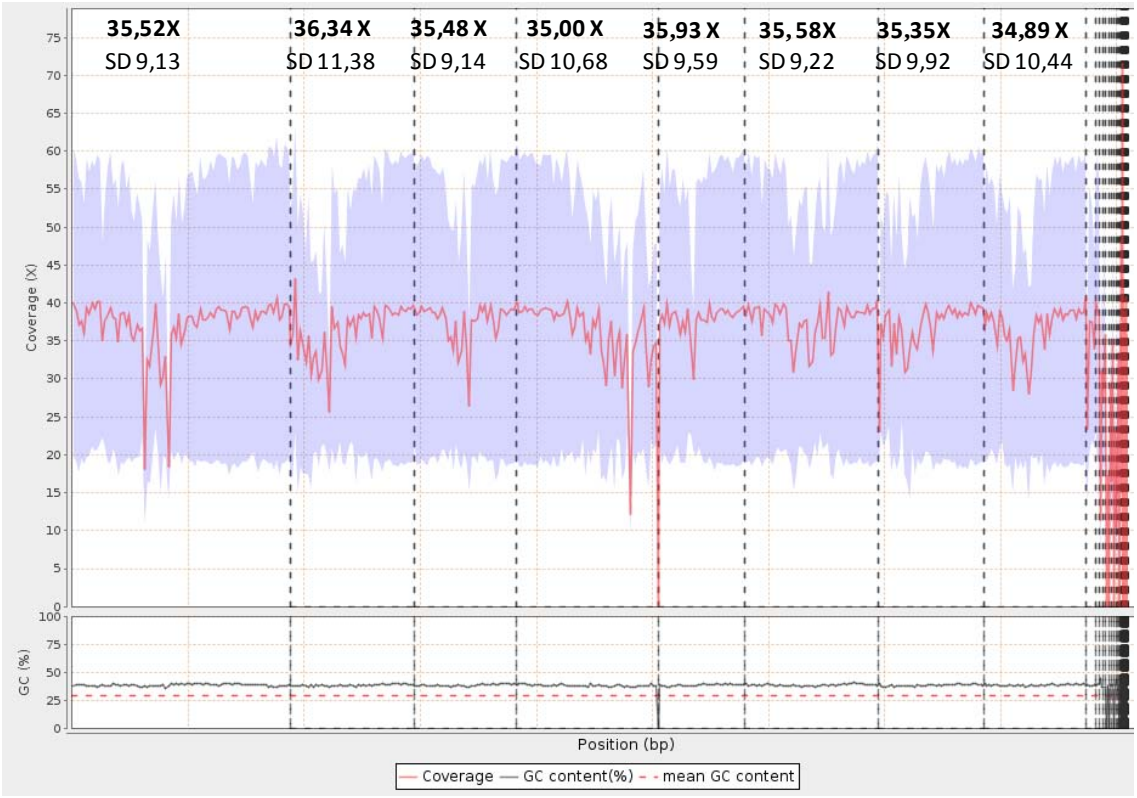


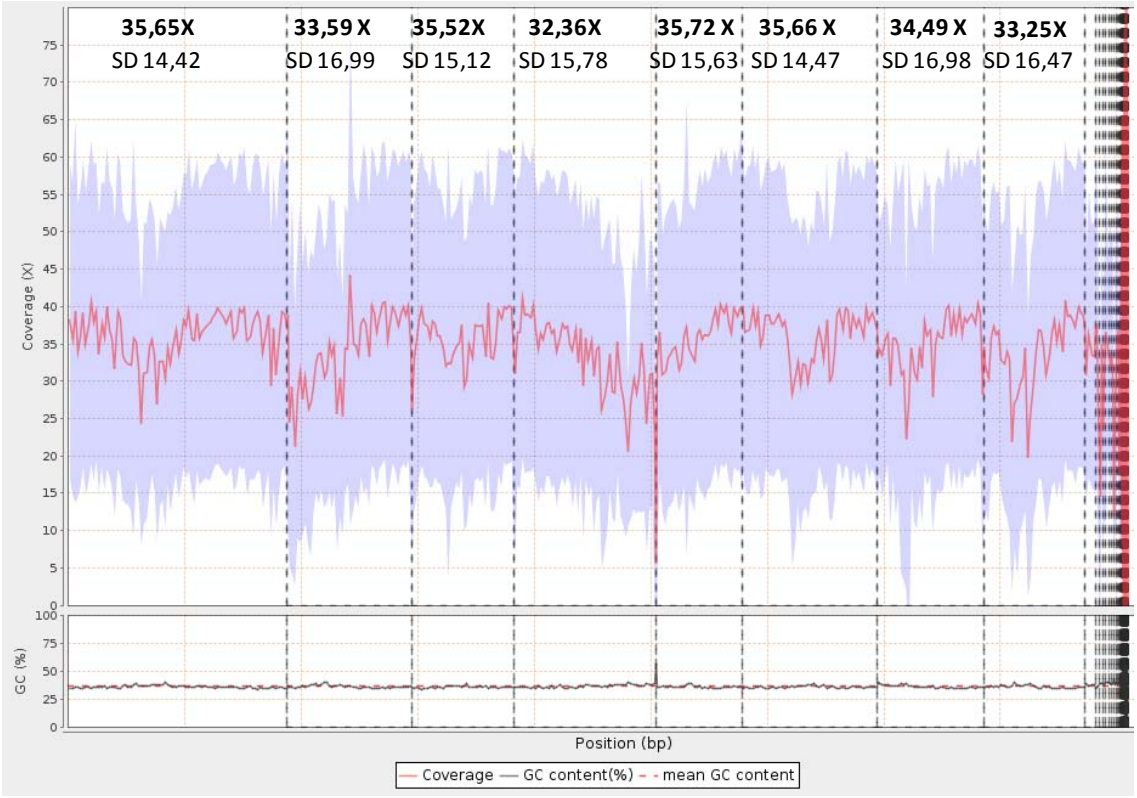
Julyprince_Pearson_nectarine


36,97X
SD 9,76
36,42 X
SD 11,02
37,01 X
SD 9,54
36,23 X
SD 10,9
37,28 X
SD 9,19
37,08X
SD 9,65
36,85X
SD 10,45
36,36 X
SD 10,91
Coverage (X)
GC (%)
Position (bp)
Coverage
GC content(%)
mean GC content

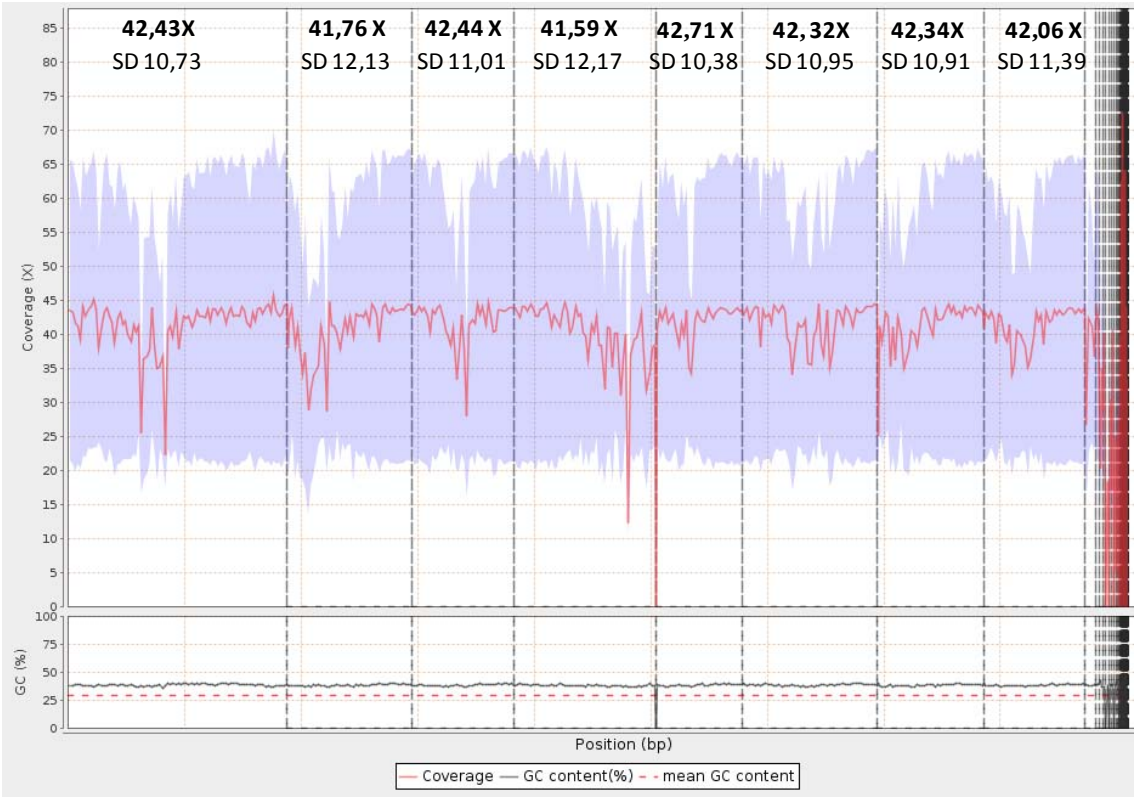

34,40X
SD 14,06
32,43 X
SD 16,62
34,23 X
SD 14,78
31,16 X
SD 15,31
34,54 X
SD 15,35
34,42 X
SD 14,12
33,30 X
SD 16,56
33,07X
SD 16,05
Coverage (X)
GC (%)
Position (bp)
Coverage
GC content(%)
mean GC content

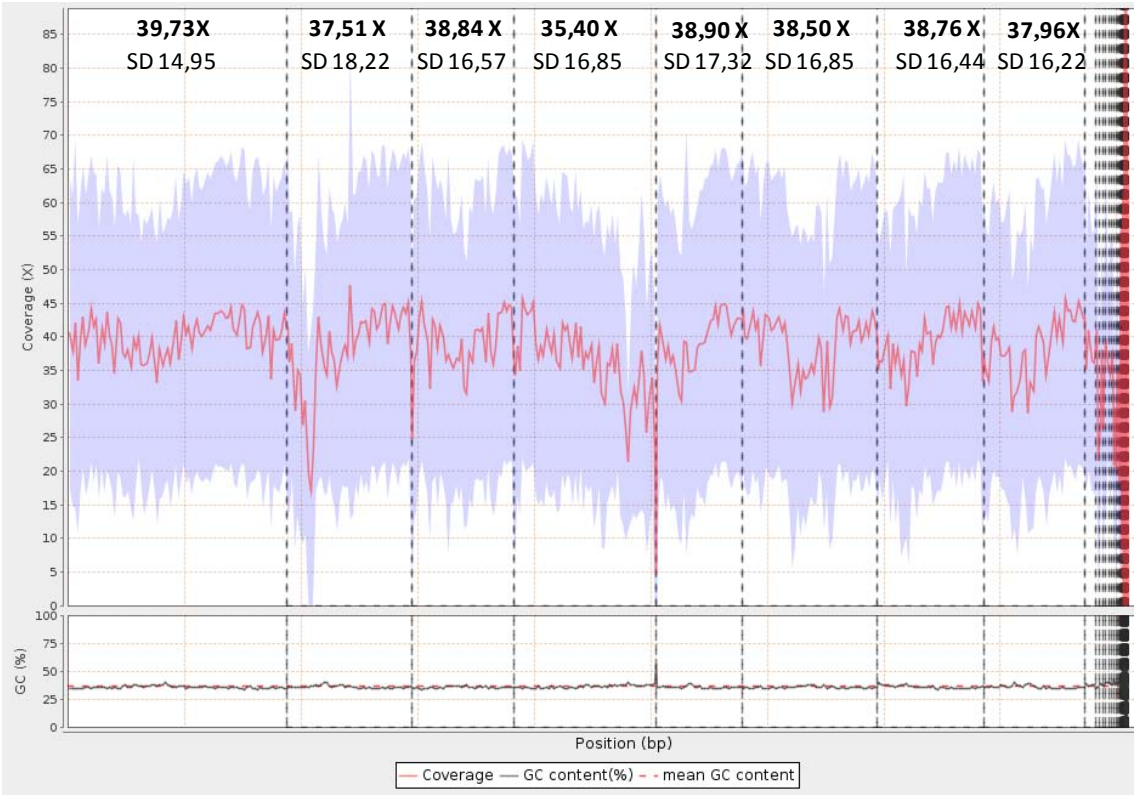
Oded_peach



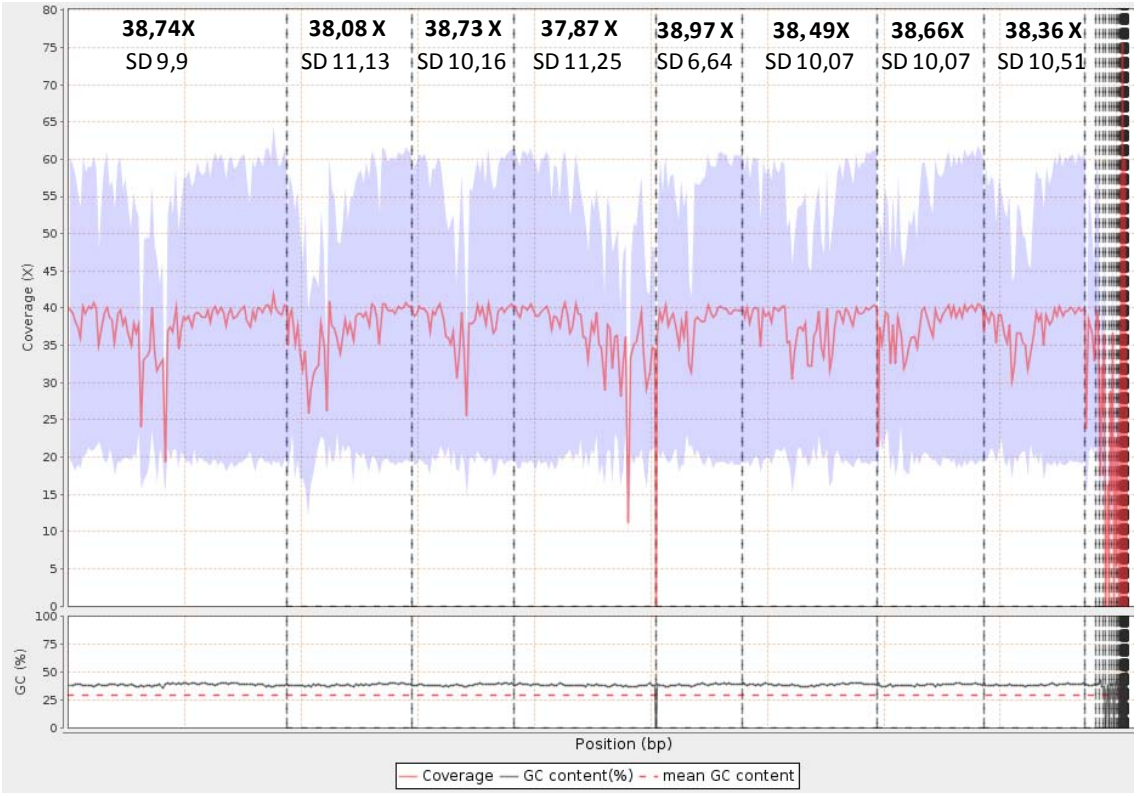


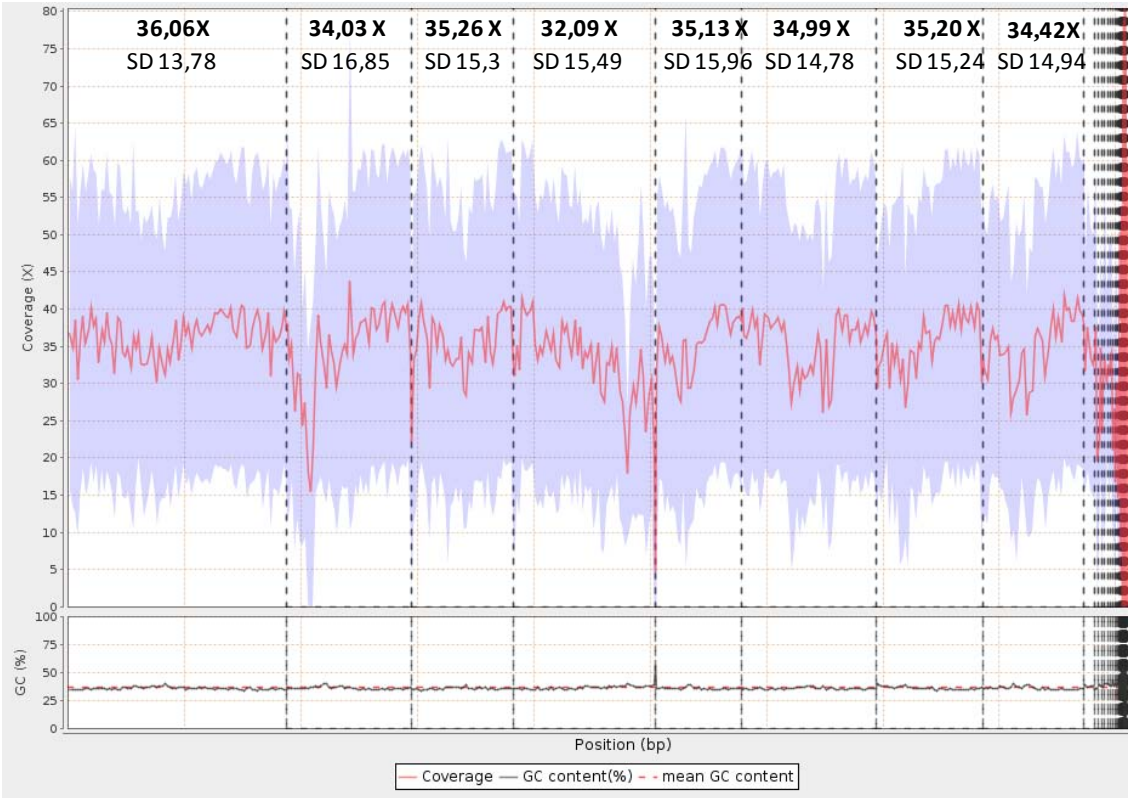
Yuval_nectarine



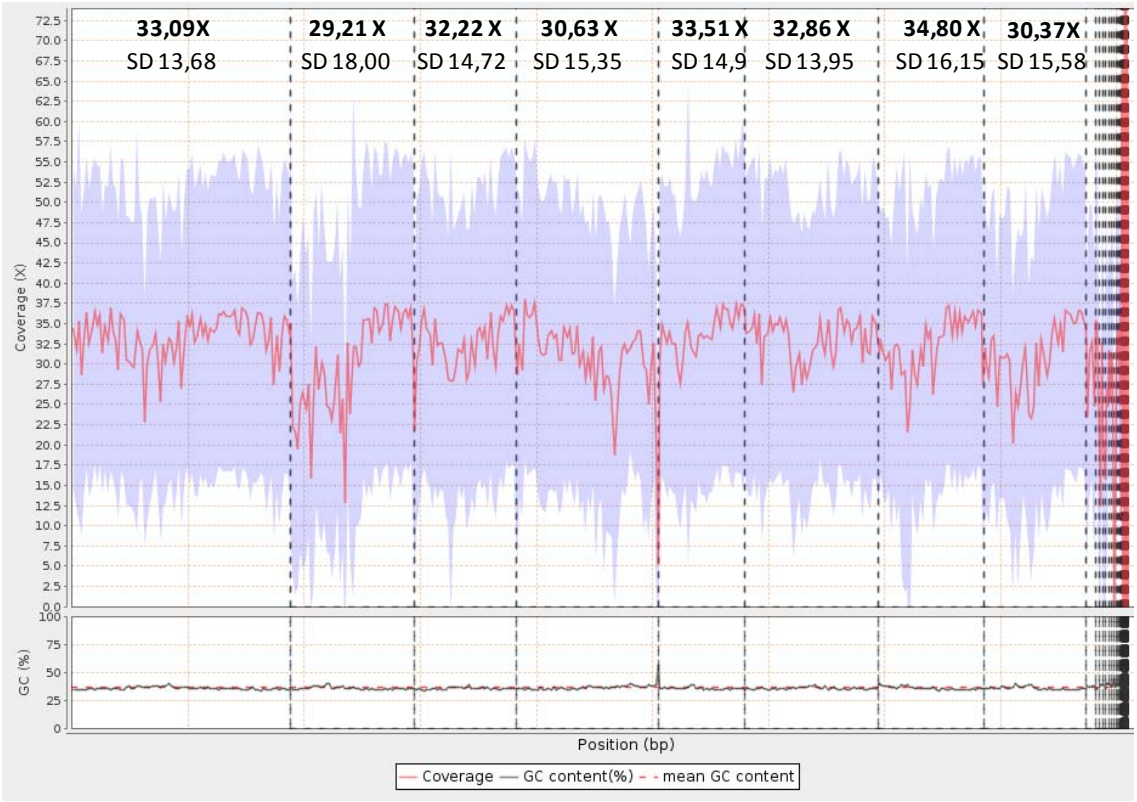


Large_White_peach

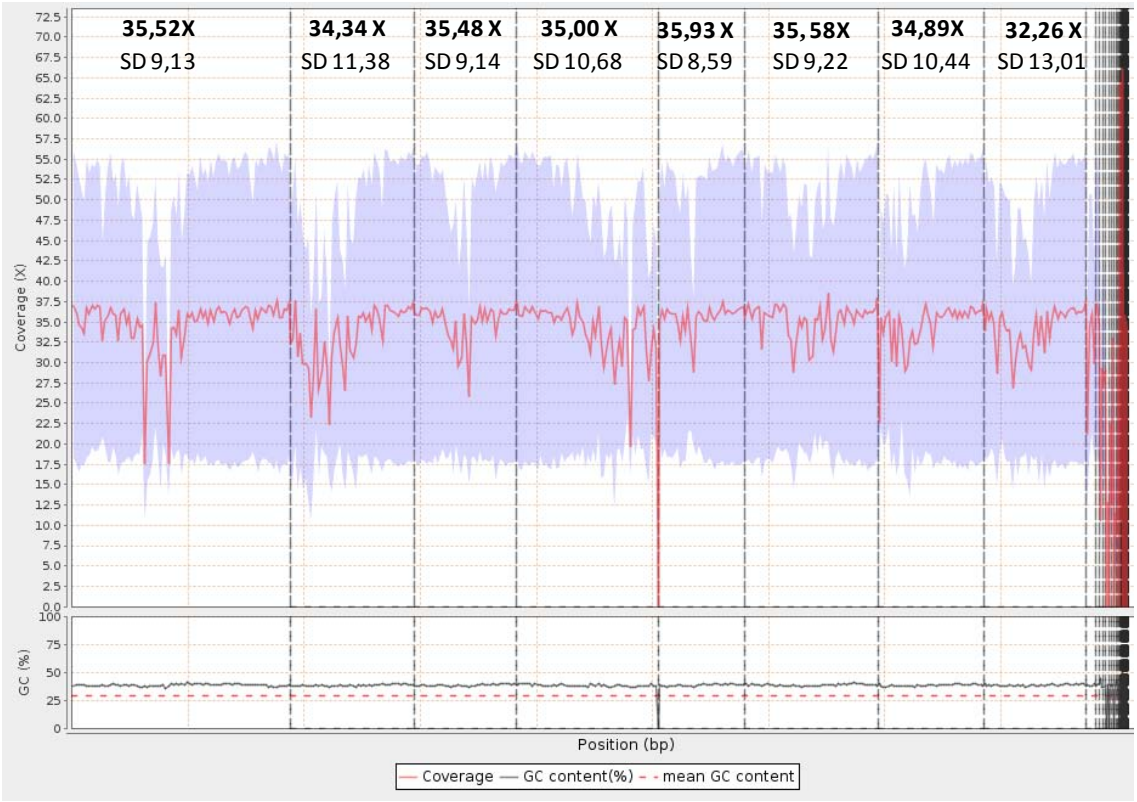




Large_White_nectarine



FloridaGlo_peach



GalI_nectarine

APPENDIX CIII.6 Possible genomic effects recognized by SnpEff, grouped by biological impact.

Effect Seq. Ontology	Effect Classic	Note & Example	Impact
Coding_sequence_variant	CDS	The variant hits a CDS.	Modifier
chromosome	Chromosome_Large_deletion	A large part (over 1%) of the chromosome was Deleted.	High
Coding_sequence_variant	Codon_Change	One or many codons are changed	Moderate
inframe_insertion	Codon_Insertion	One or many codons are inserted	Moderate
disruptive_inframe_insertion	Codon_Change_Plus Codon_Insertion	One codon is changed and one or many codons are inserted	Moderate
inframe_deletion	Codon_Deletion	One or many codons are Deleted	Moderate
disruptive_inframe_deletion	Codon_Change_Plus Codon_Deletion	One codon is changed and one or more codons are Deleted	Moderate
downstream_gene_variant	Downstream	Downstream of a gene (default length: 5K bases)	Modifier
exon_variant	Exon	The variant hits an exon.	Modifier
exon_loss_variant	Exon_Deleted	A deletion removes the whole exon.	High
frameshift_variant	Frame_Shift	Insertion or deletion causes a frame shift	High
gene_variant	Gene	The variant hits a gene.	Modifier
Intergenic_region	Intergenic	The variant is in an Intergenic region	Modifier

Effect Seq. Ontology	Effect Classic	Note & Example	Impact
Conserved_Intergenic_variant	Intergenic_Conserved	The variant is in a highly Conserved Intergenic region	Modifier
Intragenic_variant	Intragenic	The variant hits a gene, but no Transcripts within the gene	Modifier
Intron_variant	Intron	Variant hits and Intron. Technically, hits no exon in the Transcript.	Modifier
Conserved_Intron_variant	Intron_Conserved	The variant is in a highly Conserved Intronic region	Modifier
miRNA	Micro_RNA	Variant affects an miRNA	Modifier
Missense_variant	Non_Synonymous_Coding	Variant causes a codon that produces a different amino acid	Moderate
Initiator_codon_variant	Non_Synonymous_Start	Variant causes Start codon to be mutated into another Start codon (the new codon produces a different AA).	Low
Stop_retained_variant	Non_Synonymous_Stop	Variant causes Stop codon to be mutated into another Stop codon (the new codon produces a different AA).	Low
Rare_Amino_Acid_variant	Rare_Amino_Acid	The variant hits a rare amino acid thus is likely to produce protein loss of function	High
Splice_acceptor_variant	Splice_Site_Acceptor	The variant hits a splice acceptor site (defined as two bases before exon Start, except for the first exon).	High
Splice_donor_variant	Splice_Site_Donor	The variant hits a Splice donor site (defined as two bases after Coding exon end, except for the last exon).	High
Splice_region_variant	Splice_Site_Region	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the Intron.	Low
Splice_region_variant	Splice_Site_Branch	A variant affective putative (Lariat) branch point, located in the Intron.	Low

Effect Seq. Ontology	Effect Classic	Note & Example	Impact
Splice_region_variant	Splice_Site_Branch_U12	A variant affective putative (Lariat) branch point from U12 splicing machinery, located in the Intron.	Moderate
Stop_Lost	Stop_Lost	Variant causes Stop codon to be mutated into a non-Stop codon	High
5_Prime_UTR_premature Start_codon_gain_variant	Start_Gained	A variant in 5'UTR region produces a three base sequence that can be a Start codon.	Low
Start_Lost	Start_Lost	Variant causes Start codon to be mutated into a non-Start codon.	High
Stop_Gained	Stop_Gained	Variant causes a Stop codon	High
Synonymous_variant	Synonymous_Coding	Variant causes a codon that produces the same amino acid	Low
Start_retained	Synonymous_Start	Variant causes Start codon to be mutated into another Start codon.	Low
Stop_retained_variant	Synonymous_Stop	Variant causes Stop codon to be mutated into another Stop codon.	Low
Transcript_variant	Transcript	The variant hits a Transcript.	Modifier
Regulatory_region_variant	Regulation	The variant hits a known regulatory feature (non-Coding).	Modifier
Upstream_gene_variant	Upstream	Upstream of a gene (default length: 5K bases)	Modifier
3_Prime_UTR_variant	UTR_3_Prime	Variant hits 3'UTR region	Modifier
3_Prime_UTR_truncation + exon_loss	UTR_3_Deleted	The variant deletes an exon which is in the 3'UTR of the Transcript	Moderate
5_Prime_UTR_variant	UTR_5_Prime	Variant hits 5'UTR region	Modifier
5_Prime_UTR_truncation + exon_loss_variant	UTR_5_Deleted	The variant deletes an exon which is in the 5'UTR of the Transcript	Moderate