Plant Biotechnology Journal

Plant Biotechnology Journal (2013), pp. 1–11 doi: 10.1111/pbi.12081

Comparative genomics analysis in Prunoideae to identify biologically relevant polymorphisms

Tyson Koepke[1,2,†], Scott Schaeffer[1,2,†], Artemus Harper[1], Federico Dicenta[3], Mark Edwards[4], Robert J. Henry[5], Birger L. Møller[6], Lee Meisel[7], Nnadozie Oraguzie[8], Herman Silva[9], Raquel Sánchez-Pérez[3,6,*] and Amit Dhingra[1,2,*]

[1]*Department of Horticulture, Washington State University, Pullman, WA, USA*

[2]*Molecular Plant Sciences Graduate Program, Washington State University, Pullman, WA, USA*

[3]*Department of Plant Breeding, CEBAS-CSIC, Murcia, Spain*

[4]*Southern Cross University, Lismore, NSW, Australia*

[5]*Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St Lucia, Qld, Australia*

[6]*Plant Biochemistry Laboratory, Department of Plant and Environmental Sciences, University of Copenhagen, Copenhagen, Denmark*

[7]*INTA-Universidad de Chile, Santiago, Chile*

[8]*IAREC, Department of Horticulture, Washington State University, Prosser, WA, USA*

[9]*Laboratorio de Genómica Funcional & Bioinformática, Departamento de Producción Agrícola, Facultad de Ciencias Agronómicas, Universidad de Chile, La Pintana Santiago, Chile*

Received 12 December 2012;
revised 27 March 2013;
accepted 8 April 2013.
**Correspondence* (fax 509 335 8690;
email rasa@life.ku.dk; adhingra@wsu.edu)
†The authors wish it to be known that the first two authors would like to be regarded as co-first authors.

Keywords: genomics, SNPs, *Prunus*, Rosaceae, fruit ripening, missense mutations.

Summary

Prunus is an economically important genus with a wide range of physiological and biological variability. Using the peach genome as a reference, sequencing reads from four almond accessions and one sweet cherry cultivar were used for comparative analysis of these three *Prunus* species. Reference mapping enabled the identification of many biological relevant polymorphisms within the individuals. Examining the depth of the polymorphisms and the overall scaffold coverage, we identified many potentially interesting regions including hundreds of small scaffolds with no coverage from any individual. Non-sense mutations account for about 70 000 of the 13 million identified single nucleotide polymorphisms (SNPs). Blast2GO analyses on these non-sense SNPs revealed several interesting results. First, non-sense SNPs were not evenly distributed across all gene ontology terms. Specifically, in comparison with peach, sweet cherry is found to have non-sense SNPs in two 1-aminocyclopropane-1-carboxylate synthase (ACS) genes and two 1-aminocyclopropane-1-carboxylate oxidase (ACO) genes. These polymorphisms may be at the root of the nonclimacteric ripening of sweet cherry. A set of candidate genes associated with bitterness in almond were identified by comparing sweet and bitter almond sequences. To the best of our knowledge, this is the first report in plants of non-sense SNP abundance in a genus being linked to specific GO terms.

Introduction

Genetic and genomic diversity arises through multiple mechanisms including whole-genome duplication, gene copy and transposable elements. However, in closely related species, and especially within varieties, single nucleotide polymorphisms (SNPs) play a large role in contributing to genetic variation. The SNP differences in closely related species and varieties determine the phenotypic diversity observed in these plants. While large-scale rearrangements, duplications and deletions contribute to genetic changes, SNPs as well as insertions and deletions (indels) can have a direct effect on gene expression and function. SNPs and indels can be rapidly assessed through high-throughput sequencing and re-sequencing efforts and are becoming widely used as genetic markers in breeding programmes (Ahmad *et al.*, 2011; Ganal *et al.*, 2009; Hyten *et al.*, 2010; Kulheim *et al.*, 2009).

While most previously identified polymorphisms have been the result of intraspecific analyses, the genetic changes contributing to the phenotypic variation across different species of a genus are also of interest. *Prunus*, a diverse genus in the Rosaceae family with economically important ornamentals, fruits, seeds and wood-based products, is a good candidate genus for this type of analysis. *Prunus* contains species that are diploid with $n = x = 8$ and have estimated genome sizes between 225 and 300 Mb (International_Peach_Genome_Initiative, 2013; Shulaev *et al.*, 2008; Zhebentyayeva *et al.*, 2008), relatively small for the Rosaceae family. Peach has also been established as a reference genotype for *Prunus* due to the vast genomic resources available for peach including many ESTs, DNA markers and linkage maps (Zhebentyayeva *et al.*, 2008). The recently completed draft genome sequence of peach is 220–230 Mb (International_Peach_Genome_Initiative, 2013).

Production of peach, almond and sweet cherry was collectively valued at over 3.6 billion dollars in the US in 2010 (National Agricultural Statistical Services, 2011), demonstrating the economic importance of this genus and the value of understanding

Please cite this article as: Koepke, T., Schaeffer, S., Harper, A., Dicenta, F., Edwards, M., Henry, R.J., Møller, B.L., Meisel, L., Oraguzie, N., Silva, H., Sánchez-Pérez, R. and Dhingra, A. (2013) Comparative genomics analysis in Prunoideae to identify biologically relevant polymorphisms. *Plant Biotechnol. J.*, doi: 10.1111/pbi.12081

the genomic structure of these species. Each of these is crops in the *Prunus* genus that produce stone fruits, have a perennial growth habit and have a prolonged juvenility stage that has hindered the rate of progress of conventional breeding and genetic analyses.

While these species are closely related, they have differences in economically important traits that are important to production. In almond, the primary trait of interest is the difference between bitter and sweet almonds, although flowering time and shell hardness are also important. Bitterness in almonds is driven by the production of amygdalin, a cyanogenic diglucoside and its degradation products benzaldehyde and cyanide (Sánchez-Pérez *et al.*, 2008, 2012). This trait has been found to be controlled by a single, dominant gene called *Sweet kernel* (*Sk*) that produces sweet almonds (Dicenta and García, 1993; Dicenta *et al.*, 2007). SSR markers have placed *Sk* on linkage group 5 of the 'R1000' and 'Desmayo Largueta' almond maps (Sánchez-Pérez *et al.*, 2010). The position of these SSRs on the almond 'Texas' × peach 'Earlygold' (T × E) places the *Sk* locus between 11 and 14.6 Mb on the peach scaffold 5 (Sánchez-Pérez *et al.*, 2010). Several targets for DNA markers in sweet cherries are fruit size, firmness, pedicel-fruit retention force (PFRF) and powdery mildew resistance. As the peach genome is available and intra-specific polymorphism analyses have already been concluded in peach (Ahmad *et al.*, 2011), this work focuses on the genomic differences of almond and cherry with respect to peach.

Here, a reference mapping approach using the peach genome v1.0 as the reference genome and high-throughput sequencing from four almond accessions and one sweet cherry cultivar were used to identify regions of increased and decreased conservation in *Prunus*. Detailed analysis of SNPs and indels was completed to build a resource for future inquiries into these species. Additionally, preliminary analysis of the *Sk* locus in almond was completed, identifying 228 SNP candidates associated with the *Sk* gene. The collective polymorphism data set provides several regions of interest that have lower polymorphism rates and may be essential to the shared characteristics of these *Prunus* species.

Results

Sequencing data acquisition

Four almond accessions were chosen for sequencing including two sweet cultivars, Ramillete and Lauranne, and two bitter selections of CEBAS-CSIC, D05-187 and S3067. Shotgun sequencing of these four almond genotypes produced 142 million 76-base Illumina reads. Each of the individual almond genotypes was sequenced at 8–13× coverage, or 2.1–3.3 Gb of sequence, and combined to yield a 10.8 Gb data set or 43× coverage (Table 1). The sweet cherry cultivar Stella was chosen for genomic sequencing, and through 454 single-end reads, 454 paired-end reads and Illumina paired-end reads, 1.6 Gb of sequence or roughly 7× coverage was acquired. Transcriptome sequencing of sweet cherry produced an additional 460 Mb of sequence of single-end 454 reads from Bing and Rainier cultivars. The raw data were submitted to NCBI SRA (accession number SRP020000).

Assembly

A reference-based assembly of the reads onto the v1.0 of the peach genome (International_Peach_Genome_Initiative, 2013) was completed to identify regions of conservation and divergence in the *Prunus* genus. Out of all the combined Illumina reads, 56% mapped to the peach nuclear genome. Ninety-nine per cent of the remaining reads (44% of the total reads) mapped to the peach chloroplast genome. Only 0.2% of the total reads did not map to either. This confirmed that the mapping was efficient and the chloroplast-mapped reads were not analysed further. The eight primary scaffolds of peach were covered between 0.4 and 6.3× as shown in 'Data S1' which contains the coverage statistics for each scaffold and data set. These scaffolds were covered an average of 4.94× for the combined cherry data and a 3.14× average for the almond genotypes. Overall, 162 of the 334 scaffolds contained zero reads from cherry or almond, while an additional 24 were not mapped by the cherry data. Also, mapping data show that 96%–99% of peach genes were mapped to with these data sets (Table 2).

Polymorphism analyses

Overall, 13 126 567 initial polymorphisms were identified between each individual genotype and peach. The raw SNP report is made available from authors upon request. Potential polymorphisms were initially identified and parsed to 9 751 035 after filtering to retain only sites with at least three reads supporting the difference as previously described (Deschamps and Campbell, 2010; Hyten *et al.*, 2010; Koepke *et al.*, 2012; Kulheim *et al.*, 2009). These polymorphisms were then further identified based on their position revealing a total 6 138 404 polymorphic sites.

Polymorphism type and region identification

Based on the reference genome annotations (Data S2) from GDR (Jung *et al.*, 2008), the polymorphisms passing the filtering criteria were classified by their location (Table 3) yielding an average of 260 000 polymorphisms in the coding sequence (CDS) for the almond accessions and >300 000 polymorphisms in sweet

Table 1 Raw sequencing data. Total data acquired for one sweet cherry cultivar and four almond accessions. Only genomic data was used for almond genotypes

	Sweet cherry			Almond			
	Transcriptome	Genomic					
Data type	454	454	Illumina	Bitter1 Illumina	Bitter2 Illumina	Sweet1 Illumina	Sweet2 Illumina
Total sequences	1 225 030	3 742 780	977 713	29 202 304	43 522 066	42 403 474	27 607 822
Total bases (Mb)	467	1020	557	2219	3307	3222	2098
Mean read length	381	272	57	76	76	76	76
Approximate genome coverage	2.1×	4.5×	2.5×	8.9×	13.2×	12.9×	8.4×

Table 2 Non-sense mutation analysis statistics

	Sweet cherry	Bitter 1 almond	Bitter 2 almond	Sweet 1 almond	Sweet 2 almond
Total number of peach genes represented	27 576 (96.20%)	28 332 (99.33%)	28 420 (99.64%)	28 488 (99.38%)	27 590 (96.73%)
Total genes with predicted non-sense mutation(s)	5384	4016	5110	5302	5529
Total genes with predicted non-sense mutation(s) unique to each Genotype	2535	190	409	467	624

Table 3 Classifications of polymorphisms identified in each data set based on their location relative to peach genes. Locations are denoted based on the annotations as found in Data S2. Potential synonyms for the mutations are listed in parentheses

	Cherry			Almond			
Mutation location (Synonym)	Cherry transcripts	Cherry 454	Cherry illumina	Almond sweet 1	Almond sweet 2	Almond bitter 1	Almond bitter 2
Gene total	247 818	701 534	39 338	678 587	774 780	483 905	646 641
mRNA Total	266 340	739 110	40 978	707 564	808 159	504 028	673 423
CDS – Total	194 279	305 314	22 881	274 904	290 793	207 911	266 252
CDS – Sense	99 947	12 702	149 960	132 747	156 693	196 849	133 753
CDS – Mis-sense	80 131	10 688	127 288	118 544	151 659	106 056	138 853
CDS – Non-sense (Stop Gain)	3261	443	20 791	32 810	8109	5446	7452
CDS – Read-through (Stop loss)	113	23	312	355	404	257	354
CDS – Deletions	5114	6415	107	2415	3399	3549	4028
CDS – Insertions	4820	9533	100	2299	3295	3412	3761
3′ UTR	32 317	42 486	1083	26 924	30 330	19 015	25 398
5′ UTR	10 788	13 967	402	13 309	14 687	10 102	12 891
Intergenic	43 137	797 699	58 654	1 789 323	2 263 108	1 208 734	1 734 295

cherry. Polymorphisms in the exons of cDNAs (Data S3) of the almond genotypes average 52.1% (155 010), 43.3% (128 778) and 4.5% (13 454) for sense, mis-sense and non-sense mutations, respectively (Table 3). Additionally, 0.1% (342) of the CDS SNPs are read-through mutations, mutations modifying a stop codon into an amino acid yielding C terminus extension also termed 'stop loss mutations' (Zirn *et al.*, 2005). Sweet 1, however, had a much higher rate of non-sense mutations (10.6%), while the other three genotypes had fewer non-sense mutations (2%–3%). The insertions and deletions in the exons averaged ~3000 each for the four almond genotypes. In the cherry genomic data set, exonic SNPs were 50.4% (162 662) sense, 42.8% (137 976) mis-sense, 6.6% (21 234) non-sense and 0.1% (335) read-through mutations along with 16 155 indels (Table 3).

Polymorphism depth analyses

The passed filtering data set was also used to analyse the occurrence patterns of the polymorphisms. For scaffold 1 (Data S4), it is clear that there are several regions of interest containing significantly more or less than the average number of polymorphic sites. Similar mapping of the number of genes in these regions of the peach scaffold reveal low gene density regions with high polymorphism rates. Statistical analyses reveal 346 sections that significantly differ from the mean number of polymorphisms in each 50-kb region on each individual scaffold (Data S5). 95 of these sections combine to make 31 regions that are >100 kb in length with the longest region containing significantly higher polymorphisms being a 600-kb block in almond from 20.45 to 30 Mb on scaffold 1. This region in cherry contains two 50-kb blocks and one 100-kb block that are also significantly higher in polymorphism rate. These genomic regions may potentially be the regions responsible for phenotypic divergence from other members of Prunoideae.

Analysis of *Sk* locus

Further filtering of the almond polymorphisms around the *Sk* locus was completed to identify putative candidates for the *Sk* gene and causative mutations for the bitter/sweet phenotype. Using the BPPCT017 (11 Mb) and BPPCT038 (14.6 Mb) markers flanking the *Sk* locus as the boundaries reduced the 311 497 polymorphisms identified on scaffold 5–56 155 located between the SSR markers that have been reported previously as flanking the *Sk* locus (Table 4). Subsequent reduction in this data set was completed by removing polymorphisms that were not homozygous in both sweet cultivars and within both bitter accessions. Also, the homozygous polymorphisms were required to be different between the sweet and bitter accessions yielding 6304 polymorphisms of which 228 caused codon-changing mutations. These mis-sense, non-sense and read-through SNPs, as well as the indels, comprise the reduced set of putative candidates for future screening and analysis.

Blast2GO global analysis

A global comparison of putative non-sense mutations within cherry and the four selected genotypes of almond reveals a similar distribution of mutations across various gene ontology terms. This can be seen in GO terms relating to biological process, molecular function as well as cellular component (Data S6). Response to stress, protein modification process, catabolic process and transport each comprised at least 10% each of the total biological process GO terms for each tested data set. With

Table 4 *Sk* locus analysis demonstrating the effect of the various parameters on the reduction in potential targets related to bitterness in almond

	Number of target polymorphisms
A. Chromosome 5	311 497
B. A + 11-14.6 MB	56 155
C. B + fitting genetic patterns	6304
D. C + with codon change	228

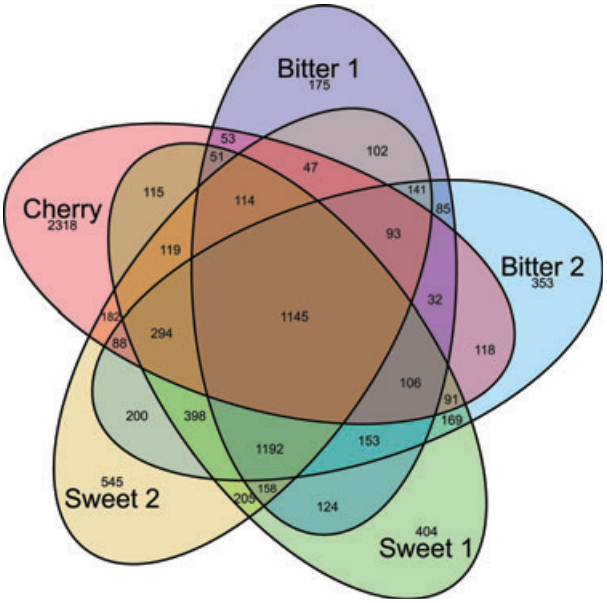

Figure 1 Venn diagram displaying the presence of non-sense single nucleotide polymorphisms (SNPs) present within the five investigated data sets mapped against peach predicted genes. A comparison of the composition of putative non-sense SNP-containing genes between the four investigated genotypes of almond and the combined cherry data set reveals the presence of a large set, 1191, of non-sense containing homologues across all members. Additionally, each sample has a unique set of genes containing putative non-sense SNPs, most notably cherry with 2535 genes.

respect to molecular function, over 35% of annotated genes containing non-sense SNPs are involved in nucleotide binding with approximately 15% having kinase activity and slightly fewer than 15% having DNA binding activity. Finally, with respect to cellular component, about 25% of all annotated genes were predicted to be localized to the plastid, with both the mitochondrion and plasma membrane comprising 15% of all annotated genes. As there appeared to be little variation in the GO term composition of the five data sets, Blast2GO analysis of the entire peach gene set was performed and compared with data sets mapping back to non-sense-SNPs. A chi-square test revealed that several GO terms have statistically higher or lower GO terms than predicted (Data S7). Non-sense mutations map back to a total of 133 unique KEGG pathways, with Bitter 1 mapping to 121, Bitter 2–119, Sweet 1–124, Sweet 2–127 and Cherry to 127 KEGG pathways, respectively (Data S8). The cherry non-sense SNP-containing data set contains members participating in atrazine degradation, chlorocyclohexane and chlorobenzene degradation, fluorobenzoate degradation, synthesis and degradation of ketone bodies and toluene degradation, while none of the investigated almond accessions did. Conversely, all four almond genotypes contained predicted non-sense mutations within genes involved in butirosin and neomycin biosynthesis, D-alanine metabolism, D-arginine and D-ornithine metabolism. In the almond accessions, non-sense mutations were also found in genes assigned in databases as involved in glucosinolate biosynthesis all of which lacked participating genes with putative non-sense mutations in cherry. As members of the Rosaceae family do not produce glucosinolates but synthesize cyanogenic glucosides using similar gene families, the assignment of such genes to glucosinolate biosynthesis is obviously erroneous (Conn, 1969; Sánchez-Pérez *et al.*, 2008).

Comparison of the non-sense-containing genes within the five data sets reveals that a large subset of the genes, 1191 in total, is shared between all members (Figure 1). Additionally, some non-sense SNPs are unique to individual samples. The largest of these sets, 2535 genes, are the non-sense mutations unique to cherry. One thousand two hundred and seventy-six genes containing putative non-sense SNPs are present within each individual almond genotype and absent from the cherry analysis.

Blast2GO targeted pathway analysis

The most abundant biological process gene ontology (GO) term represented in the data sets 'Response to stress' was selected as a GO of interest to further investigate. Further breakdown of this category reveals that its members are involved in a total of 92 KEGG pathways within the five investigated data sets (Data S9). While all data sets contain genes with putative non-sense SNPs in numerous pathways, only sweet cherry contains putative non-sense SNPs related to stress in C5-branched dibasic acid metabolism, chlorocyclohexane and chlorobenzene degradation, indole alkaloid biosynthesis, isoquinoline alkaloid biosynthesis, naphthalene biosynthesis, N-glycan biosynthesis, nicotinate and nicotinamide metabolism, primary bile acid biosynthesis, retinol metabolism, steroid degradation, steroid hormone biosynthesis, toluene degradation, tropane, piperidine and pyridine alkaloid biosynthesis and valine, leucine and isoleucine biosynthesis. Alternately, all four accessions of almond contain potential non-sense SNPs in alanine, aspartate and glutamate metabolism, benzoate degradation, caprolactam degradation, fatty acid elongation, geraniol degradation, monoterpenoid biosynthesis, sulphur metabolism and vitamin B6 metabolism, while cherry lacks non-sense mutation in these pathways.

Further investigation into non-sense mutations present within members of the genes involved in cyanogenic glucoside metabolism was performed as these biosynthetic and catabolic pathways lead to amygdalin synthesis and catabolism, respectively. Blast2GO analysis performed through searching for the keywords 'Prunasin' and 'Amygdalin' revealed the presence of four isoforms of peach prunasin beta-glucosidase and amygdalin beta-glucosidase with non-sense mutations in cherry (ppa003891m, ppa016583, ppa003856m and ppa003831m), one in Bitter 1 (ppa003831m), three in Bitter 2 (ppa003856m, ppa003891m and ppa003831m), four in Sweet 1 (ppa003891m, ppa016583, ppa003856m and ppa003831m) and four in Sweet 2 (ppa003891m, ppa016583, ppa003856m and ppa003831m). Based upon the annotations, numerous other members within this pathway contained putative non-sense mutations and additional members with potential prunasin beta-glucosidase or

amygdalin beta-glucosidase activity were detected (Figure 2 and Table 5).

These data were further probed to identify differences in non-sense SNPs involved in ripening, another biological process of interest with respect to the *Prunus* genus. Within the 25 genes in all peach sequences mapping to the ripening GO term, cherry surprisingly contains eight that are predicted to contain putative non-sense SNPs (two putative multidrug resistance genes, two 1-aminocyclopropane-1-carboxylate synthase (ACS) genes, two 1-aminocyclopropane-1-carboxylate oxidase (ACO) genes, one polygalacturonase gene and one glycosyl hydrolase family nine protein gene; Data S10). Bitter 1 and Bitter 2 almond data sets each contain two non-sense SNP-containing genes (one ACO gene and one putative multidrug resistance gene), while Sweet 1 and Sweet 2 almonds each contain three (two ACO gene and one putative multidrug resistance gene). Further analysis of genes containing SNPs in the KEGG pathway 'cysteine and methionine biosynthesis' reveals additional non-sense SNPs that may affect ethylene synthesis, but are not mapped to the ripening GO term (Figure 3 and Table 6).

The final biological process of interest, abscission, has very few members containing non-sense mutations compared with the other GO terms, in fact only 23 genes in the entire peach gene set map to this term. Sweet 1 and Sweet 2 almonds are predicted to have a single gene related to abscission containing a non-sense SNP in a gene encoding btb/poz ankyrin repeat protein. Bitter 1 almond sequences had no non-sense mutations in any predicted gene sequences, while Bitter 2 almond has a non-sense mutation in a gene encoding a probable adp-ribosylation factor gtpase-activating protein (agd5-like). Cherry has non-sense SNPs present

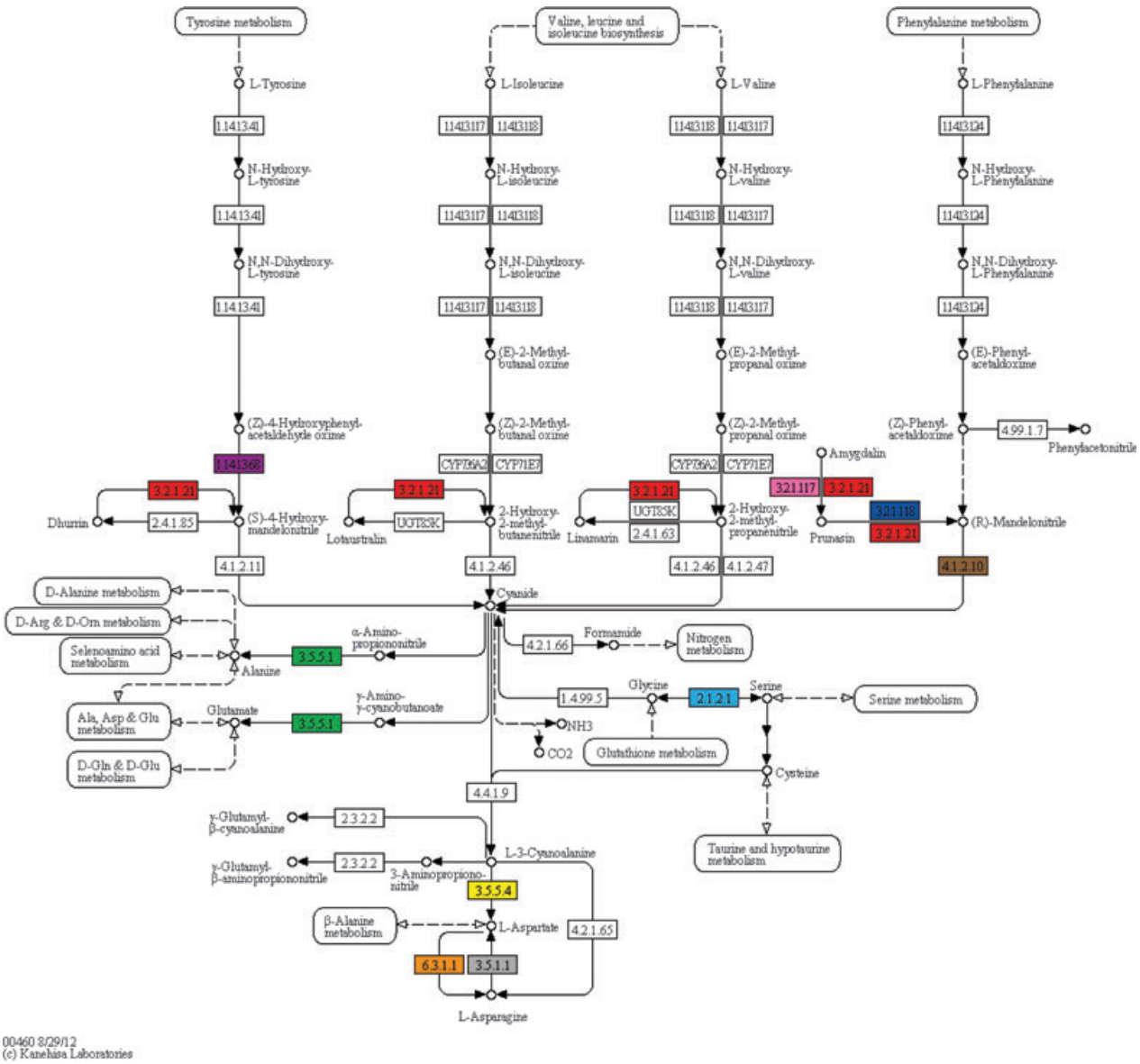

Figure 2 KEGG Pathway of genes containing non-sense mutations within cherry and almond samples involved in Cyanogenic glucoside metabolism. Table 5 lists the EC numbers, their identities, as well as the predicted coding sequence (CDS) from *Prunus persica* which contain a SNP for each respective species. Coloured boxes correspond to the colour code found in the first column of Table 5.

Table 5 *Prunus persica* predicted CDS IDs containing non-sense mutations for putative members in cyanoamino acid metabolism (Figure 2)

EC number	Enzyme	Cherry	Bitter 1	Bitter 2	Sweet 1	Sweet 2
3.2.1.21 (Red)	Beta-glucosidase	ppa018777m	ppa015619m	ppa015619m	ppa015619m	ppa015619m
		pp a015330m	pp a001675m	pp a001675m	pp a014607m	pp a015239m
		pp a004484m	pp a004484m	pp a004484m	pp a018777m	pp a014607m
		pp a001692m	pp a006167m	pp a001656m	pp a019582m	pp a014605m
		pp a006142m	pp a006142m	pp a006167m	pp a001675m	pp a018777m
		pp b011574m	pp a023264m	pp a007195m	pp a001656m	pp a019582m
		pp a023763m	pp a023763m	pp a001692m	pp a006167m	pp a001675m
		pp a001675m		pp b021184m	pp a006142m	pp a004484m
				pp a026252m	pp b021184m	pp a001656m
				pp a024207m	pp a026252m	pp a006167m
				pp a021476m	pp a024207m	pp a007195m
				pp a023264m	pp a020836m	pp b021184m
				pp a023763m	pp a023264m	pp a024207m
					pp a023763m	pp a020836m
						pp a023264m
						pp a023763m
3.5.5.4 (Yellow)	Cyanoalanine nitrilase	ppa008090m	ppa008090m	ppa008090m	ppa008090m	ppa008090m
3.5.5.1 (Green)	Nitrilase	ppa008583m	ppa008767m	ppa008767m	ppa008767m	ppa008767m
						pp a007102m
6.3.1.1 (Orange)	Aspartate-ammonia ligase	0	ppa015268m	0	0	ppa015268m
4.1.2.10 (Brown)	(R)-mandelonitrilelyase	ppa016983m	ppa003595m	ppa003595m	ppa003595m	ppa003414m
		pp a003595m	pp a003414m	pp a003414m	pp a003414m	pp a003422m
		pp a003414m		pp a003422m	pp a003422m	pp a020579m
		pp a003422m		pp a020579m	pp a020579m	pp a022916m
		pp a004308m		pp a022916m	pp a022916m	
		pp a020579m				
3.2.1.118 (Blue)	Prunasin beta-glucosidase	ppa017484m	ppa017484m	ppa017484m	ppa017484m	ppa017484m
		pp a019137m	pp a019137m	pp a018933m	pp a016583m	pp a019137m
		pp a019262m	pp a018933m	pp a015970m	pp a015970m	pp a018933m
		pp a015721m	pp a018404m	pp a019262m	pp a019262m	pp a016583m
		pp a003718m	pp a003831m	pp a015161m	pp a016757m	pp a015970m
		pp a003891m		pp a003718m	pp a015161m	pp a016757m
		pp a021158m		pp a003856m	pp a003718m	pp a015161m
		pp a020817m		pp a003891m	pp a003856m	pp a004108m
		pp a026358m		pp a003831m	pp a003891m	pp a003718m
		pp a016583m		pp a022831m	pp a003831m	pp a003856m
		pp a003856m		pp a020817m	pp a022831m	pp a003891m
		pp a003831m		pp a020368m	pp a025067m	pp a003831m
				pp a026358m	pp a020368m	pp a021158m
				pp a027189m	pp a026358m	pp a020839m
					pp a027189m	pp a026358m
						pp a027189m
3.2.1.117 (Pink)	Amygdalin beta-glucosidase	ppa017484m	ppa017484m	ppa017484m	ppa017484m	ppa017484m
		pp a019573m	pp a019573m	pp a019573m	pp a019573m	pp a019573m
		pp a019137m	pp a019137m	pp a018933m	pp a016583m	pp a019137m
		pp a019262m	pp a018933m	pp a015970m	pp a015970m	pp a018933m
		pp a015721m	pp a018404m	pp a019262m	pp a019262m	pp a016583m
		pp a003718m	pp a004380m	pp a015161m	pp a016757m	pp a015970m
		pp a003891m	pp a003831m	pp a004380m	pp a015161m	pp a016757m
		pp a021158m		pp a003718m	pp a004380m	pp a015161m
		pp a020817m		pp a003856m	pp a003718m	pp a004380m
		pp a020067m		pp a003891m	pp a003856m	pp a004108m
		pp a026358m		pp a003831m	pp a003891m	pp a003718m
		pp a016583m		pp a022831m	pp a003831m	pp a003856m
		pp a003856m		pp a020817m	pp a022831m	pp a003891m
		pp a003831m		pp a020067m	pp a020067m	pp a003831m
				pp a020368m	pp a025067m	pp a021158m

Table 5 Continued

EC number	Enzyme	Cherry	Bitter 1	Bitter 2	Sweet 1	Sweet 2
				pp a026358m pp a027189m	pp a020368m pp a026358m pp a027189m	pp a020067m pp a020839m pp a026358m pp a027189m
3.5.1.1 (Grey)	Asparaginase	ppa008583m	0	ppa016260mppa008761m	ppa008583m	ppa016260m pp a008583m
1.14.13.68 (Purple)	4-Hydroxyphenylacetaldehyde oximemonooxygenase	0	0	0	ppa014661m	0
2.1.2.1 (Cyan)	Glycine hydroxymethyltransferase	ppa003640m	0	0	0	ppa004090m

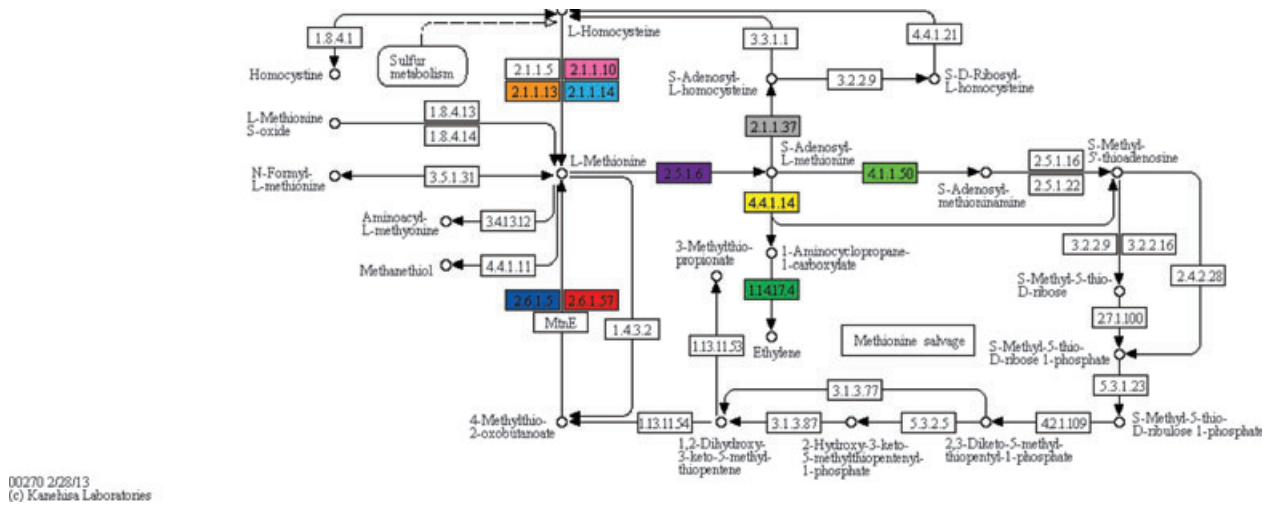

Figure 3 KEGG pathway of genes containing non-sense mutations within cherry and almond samples participating in cysteine and methionine metabolism. Genes containing non-sense mutations were mapped to EC numbers using Blast2GO and mapped on the cysteine and methionine metabolism KEGG map closest to ethylene production. Table 6 lists the identity of each enzyme, as well as the predicted CDS from *Prunus persica* which contain a SNP for each respective species. Only the lower half of the pathway was imaged. Coloured boxes correspond to the colour code found in the first column of Table 6.

in both the gene encoding adp-ribosylation factor gtpase-activating protein (agd5-like) and the btb/poz ankyrin repeat protein.

Discussion

Uneven distribution of sequencing reads

Reference-based assemblies are built upon the presumption that the sequenced genomes are highly similar to the reference. When differences exceed the threshold of the mapping software, reads from these highly divergent regions are not mapped. 7.5% of reads from *indica* rice cultivars did not map to the Nipponbare rice genome (Subbaiyan *et al.*, 2012). In our data, 0.2% of the Illumina reads were unmapped to the peach genome (International_Peach_Genome_Initiative, 2013), strengthening the assertions that peach is a credible reference genome for *Prunus*.

In addition to unmapped reads, the data produced in this work identified 162 and 186 scaffolds of peach that were not covered by any reads from almond and sweet cherry, respectively. One explanation for this is that these smaller, un-anchored scaffolds may be unique to the peach genome. Alternatively, as the peach genome was built using Sanger sequencing, these could be repetitive regions and the shorter reads used here were placed in the dual location during the mapping. In either case, these regions provide insights into genome structure differences that need to be further evaluated to fully understand the differences among these *Prunus* species.

Polymorphism analyses

Our analyses show 48% of the mutations in the CDS to be nonsynonymous, being either non-sense or mis-sense mutations, in almond and 50% in sweet cherry. This is comparable to the 57% found in rice cultivars (Subbaiyan *et al.*, 2012); however, it is interesting that this interspecies comparison identified a higher percentage of synonymous SNPs than the intraspecific comparison in rice. While it is possible that the heterozygous nature of the almond and sweet cherry genotypes caused polymorphisms to be screened out during filtering, it is unlikely that this would have significantly shifted the representation of synonymous and nonsynonymous mutations.

Table 6 *Prunus persica* predicted CDS IDs containing non-sense mutations with putative functions cysteine and methionine metabolism (Figure 3)

EC Number	Enzyme	Cherry	Bitter 1	Bitter 2	Sweet 1	Sweet 2
2.6.1.57 (Red)	Aromatic-amino-acid transaminase	ppa003908m	ppa004475m	ppa004475m	ppa004475m	ppa004475m
4.4.1.14 (Yellow)	1-Aminocyclopropane-1-Carboxylate synthase	ppa015636m	ppa015636m	ppa015636m	ppa015636m	ppa015636m
		ppa016458m	ppa004475m	ppa004475m	ppa004475m	ppa004475m
		ppa004774m		ppa003850m	ppa005521m	ppa003850m
		ppa003908m				
		ppa005521m				
1.14.17.4 (Green)	Aminocyclopropanecarboxylate oxidase	ppa008813m	ppa008813m	ppa008813m	ppa008813m	ppa008813m
		ppa008791m			ppa008791m	ppa008791m
2.1.1.13 (Orange)	Methionine synthase	0	ppa015268m	0	0	ppa015268m
4.1.1.50 (Lime Green)	Adenosylmethionine decarboxylase	ppa007732m	ppa007732m	0	0	ppa007294m
2.6.1.5 (Blue)	Tyrosine transaminase	ppa019805m	0	ppa019805m	ppa019805m	ppa019805m
						ppa018754m
2.1.1.10 (Pink)	Homocysteine S-methyltransferase	ppa008404m	0	0	0	ppa010310m
2.1.1.37 (Grey)	DNA (cytosine-5-)-methyltransferase	ppa019831m	ppa019831m	ppa015623m	ppa019831m	ppa000190m
		ppa000190m		ppa000190m	ppa015623m	ppa006086m
		ppa006086m		ppa006086m	ppa000190m	
					ppa006086m	
2.5.1.6 (Purple)	Methionine adenosyltransferase	ppa006915m	0	0	ppa025497m	0
2.1.1.14 (Cyan)	5-Methyltetrahydropteroyltriglutamate-homocysteine S-methyltransferase	0	ppa026306m	ppa021650m	ppa026306m	ppa021650m
			ppa021650m		ppa021650m	

It is important to note that the read-through mutations could be discussed as non-sense mutations of the almond gene in peach; therefore, discussion of read-through and non-sense mutations is limited by the perspective of the analysis which, in this case, is in respect to the peach reference genome. At first glance, the 0.1% generation rate of read-through mutations suggests that these mutations may be highly deleterious with strong selection against them as they occur at ~1/50th of the rate that non-sense mutations arise. A closer examination, however, reveals that while the probability of a stop codon mutation causing a read-through mutation is 85%, there is only one stop codon per protein. This contrasts significantly with the 4.2% chance of a random SNP causing a non-sense mutation multiplied by the 403 amino acids found in the average CDS in the peach genome. Calculating for the distribution of amino acids yields one polymorphism having a 4.18% or 0.21% chance of causing a non-sense or read-through mutation, respectively, in the average peach gene. The data from this work show a 2.5-fold change from expected providing intrigue but requiring further evaluation regarding the effect of these mutations on gene function.

Three hundred and forty six regions with higher and lower rates of polymorphism were identified in this work. Higher rates could result from genomic duplications or from low conservation yielding more divergence. Similarly, regions with lower average polymorphisms could be the result of either low divergence where few polymorphisms arose or very high amounts of differentiation preventing the mapping of the sequencing reads to these locations. Subbaiyan *et al.* (2012) revealed similar regions of lower polymorphism rates in six inbred lines of rice with several being >100 kb in length. The 600-kb region in almond is particularly interesting as it may represent a larger region of diversity between almond and peach and may contain genes related to the divergence of these two species.

Analysis of the *Sk* locus

The combination of the existing DNA markers, the reference sequence and genotype-specific sequencing yielded 228 candidate mutations for the *Sk* trait in almond. As this work was completed using only two bitter and two sweet genotypes, reductions in this candidate set would be expected if more genotypes were examined. However, whole-genome sequencing of further genotypes is not necessary at this time as site-specific testing of the genotypes for the identified mutations is expected to identify the allele responsible for the difference between these types of almonds. As the major and highly critical trait, developing a gene-based marker for the *Sk* gene will provide an important benefit to the almond community by rapidly identifying the undesirable bitter genotypes. As suggested by Michelmore *et al.* (1991), bulked segregant analysis can function in an obligate outcrossing species. The results shown here demonstrate the ability of the approach to produce a small candidate list from a large region of interest. Adding more individuals to the bulks in this work would allow the marker placement to be independently confirmed as well, although using two individuals of each phenotype was possible due to the previously developed markers for the *Sk* locus.

Blast2GO comparisons

The global distribution of GO terms within the non-sense SNP-containing genes was similar among all samples tested. This suggests two potential options regarding the presence of non-sense SNPs: (i) certain gene ontology terms have accrued non-sense SNPs at similar levels across species in *Prunus*, and (ii) non-sense SNPs simply occur randomly throughout the genome. Each gene ontology term contains a similar number of genes among the samples investigated. To assess this, comparison of the observed number of members for each GO term was performed against expected values generated from the entire

peach predicted gene set using a chi-square test. This test showed that numerous GO categories contained statistically significant higher or lower numbers of non-sense SNPs than expected (Data S7). This suggests that many GO terms are linked to an increased likelihood to generate non-sense SNPs in *Prunus*, while other GO terms appear to be more conserved in the genus supporting option one above. Interestingly, the GO terms associated with significantly higher non-sense SNPs (*P*-value $<1E^{-10}$) include the following: the biological processes: 'DNA metabolic process' (GO:0006259), 'cellular protein modification process'(GO: 0006464), 'signal transduction' (GO:0007165) and 'pollen-pistil interaction' (GO:0009875); the cellular components: 'mitochondrion' (GO:0005739), 'cytoskeleton' (GO:0005856) and 'plastid' (GO:0009536); and the molecular functions: 'nucleotide binding' (GO:0000166) and 'kinase activity' (GO:0016301). GO terms associated with significantly lower non-sense SNPs (*P*-value $<1E^{-10}$) include the biological processes 'response to biotic stimulus' (GO:0009607), 'response to abiotic stimulus' (GO: 0009628), 'anatomical structure morphogenesis' (GO:0009653) and 'response to endogenous stimulus' (GO:0009719); the cellular component 'cytosol' (GO:0005829); and the molecular functions 'chromatin binding' (GO:0003682), 'sequence-specific DNA binding transcription factor activity' (GO:0003700) and 'structural molecule activity' (GO:0005198). While a connection between GO term and occurrence of non-sense SNPs appears to exist, this does not disprove the option two stated above.

Concerning the GO term 'response to stress', there appears to be significant genetic variability with respect to non-sense SNPs. In fact, sequences containing detected non-sense SNPs mapped to this GO term more than any other GO term investigated in the biological process domain. This gene ontology is of high agricultural importance as breeding and genetic modification of plants resistant to both biotic and abiotic stresses is a large focus in both industry and academia. Previous studies have used gene-based SNPs detected through interspecific comparisons to identify, verify and attach function to SNPs which may be involved in stress response (Parida *et al.*, 2012). These putative non-sense SNPs represent a preliminary data set within *Prunus* which may be used in similar studies.

Basic differences exist in the ripening patterns of members of the *Prunus* genus. Peach, apricot and plum fruits are climacteric, meaning that a burst of ethylene occurs quickly followed by an increase in respiration. Cherry and almond, on the other hand, exhibit nonclimacteric ripening, outliers in the genus. The identification of non-sense mutations in several versions of ACS and ACO could significantly disrupt the ethylene production pathway in cherry rendering it nearly unable to provide the burst seen in other fruits in this genus. It has been suggested that in sweet cherry regulation of respiration may not be under the regulation of ethylene receptors (Gong *et al*., 2002).

While these results enable the identification of targets for gene-linked marker screening, it is important to realize the limitations of this project. First of all, non-sense SNPs do not necessarily equate to loss of function of a protein. Additionally, as these sequences were aligned to a predicted peach data set, the true sequence of genes of interest may be biased. Potential splice variants may have the 'non-sense' mutation in an exon that is not utilized in these species. Also, the presence of a single non-sense mutation may not be deleterious at all and could be sufficiently complemented by the other allele especially in a genus where very few self-compatible varieties exist leading to high amounts of heterosis. Gene duplications or those genes unique to almond or cherry may not be represented in these data; alternatively, they may be represented as SNPs, while they are actually different alleles.

Conclusions

Using reference-based assemblies of four almond accessions and one sweet cherry cultivar, we were able to begin interspecific comparative genomic analysis of Prunoideae. Over 99% of the raw reads mapped to the peach genome although nearly 44% mapped to the chloroplast. Identifying hundreds of smaller scaffolds in the peach genome that were not mapped to by either the almond or sweet cherry data finds many potentially peach-specific regions of interest for further investigation. The 6.1 million putative SNPs provide a resource for gene-based investigations. While many of the SNPs and indels are in noncoding regions, 250–300 thousand SNPs are located in the coding regions of annotated peach genes. These SNPs should prove to be useful in expanding our knowledge of genetics and genomics in these species through their use as molecular markers and gene-based interrogations. The coverage depth images revealed 31 regions that have significantly different amounts of SNPs.

A keystone goal of genomics is to identify genes responsible for specific traits. Here, we examined the bitterness trait of almond and identified 228 codon-changing mutations near the previously identified *Sk* locus. Additionally, to the best of our knowledge, we provide the first report in plants of non-sense SNP abundance in a genus being linked to specific GO terms. A global analysis of SNPs has also revealed several candidate mutations of interest for different physiological properties of these species including response to stress, ripening and abscission. Combined, these data should provide a foundation for further genomics and genetics research in Prunoideae.

Methods

Sequencing data acquisition

Almond

D05-187 (Bitter1) and S3067 (Bitter2) are homozygous bitter selections from the CEBAS-CSIC, and Ramillete (Sweet1) and Lauranne (Sweet2) are each homozygous sweet cultivars of almond. Using an estimated genome size of 250 Mb, approximately 10× coverage was obtained for each of the four genotypes with 76 bp Illumina paired-end reads.

Cherry

The sweet cherry genome project has developed roughly 7× coverage of Stella, an important parental cultivar based on a 225-Mb genome size estimation. These data were derived mostly through single-end 454 with some paired-end 454 and Illumina paired-end sequencing. Both 454 GS-FLX and 454 GS-FLX+ versions were used to acquire these sequences. Also, 454 transcriptome data from Bing and Rainier cultivars of sweet cherry were obtained and used in the analyses. These transcriptome data were utilized only for polymorphism analysis to compare to the gene annotations of peach and were not obtained in sufficient depth for expression-based analyses.

Peach genome

The peach genome version 1.0 (International_Peach_Genome _Initiative, 2013) was obtained from GDR (Jung *et al.*, 2008) for use as the reference sequence throughout this project. The

chloroplast and mitochondrial genomes were excluded from the assembly initially, and the chloroplast was later used to screen the unassembled reads.

Assembly

A reference-based assemblies of both the cherry genomic 454 and cherry transcriptomic 454 data were assembled using the NGen assembler (DNAStar, Madison, WI) version 3.1.0 with the peach genome version 1.0 as the reference and using the following 454 default parameters: mersize = 21, merSkipQuery = 3, minMatchPercent = 85, MaxGap = 15, minAlignedLength = 50. Similarly, all Illumina data from the four almond accessions and sweet cherry were assembled using the peach genome as a reference with the Illumina default parameters: mersize = 21, minMatchPercent = 93, mismatchPenalty = 20, MaxGap = 6, min Aligned Length = 35. For each assembly, the different genotypes were input separately to enable unique SNP information to be attained for each individual.

Polymorphism analyses

Assembled data were imported into SeqMan (DNAStar) where SNP reports were created. A custom script was used to remove polymorphisms with <3 reads confirming each nonreference call similar to previous SNP reporting works (Deschamps and Campbell, 2010; Hyten *et al.*, 2010; Koepke *et al.*, 2012; Kulheim *et al.*, 2009). These filtered SNPs were then imported into ArrayStar (DNAStar) to enable further analyses.

Polymorphism type and region identification

Custom computational comparisons of the base calls from the sequenced individuals against the peach genome were completed to determine the base changes involved. Similarly, polymorphism regions were identified by analysing the reference position against the annotation of the peach genome. These SNPs were classified as 5′ UTR, intron, exon, 3′ UTR or intergenic. Exonic polymorphisms were further classified as sense, non-sense, mis-sense or read-through mutations based on the resulting amino acid compared with the peach genome annotation. Read-through mutations were defined as the SNPs causing a stop codon to be changed into an amino acid thereby elongating the C terminus of the protein with respect to the peach gene (Zirn *et al.*, 2005).

Polymorphism depth analyses

To visualize the depth of the polymorphisms across the eight main scaffolds of the peach reference, the total polymorphisms in each discreet 50-kb window were analysed and displayed as a single-pixel-wide bar one pixel high for each 20 polymorphisms. The graphs for each individual were then compiled into a single image per scaffold. The composite polymorphism set, where each unique SNP was counted once for each species, was also analysed in this manner. The distribution of polymorphism counts per 50-kb window was analysed to identify regions of the peach reference that had a polymorphism depth >2 standard deviations from the mean of that scaffold.

Analysis of *Sk* locus

The total almond SNP report was filtered to retain only the polymorphic sites near the *Sk* locus. As the markers BPPCT017 and BPPCT038 are located at ~11–14.6 Mb on peach linkage group 5, respectively (Sánchez-Pérez *et al.*, 2010), they were used for the bounds around the *Sk* locus. All polymorphisms that were conserved within a group but contrasting between the two types were retained as both bitter and both sweet genotypes are homozygous for the trait. Further screening reduced the data set to only contain codon-changing polymorphisms that make up the candidate gene set.

Blast2GO comparisons

Nucleotide sequences for all predicted *Prunus persica* genes were imported into Blast2GO (Conesa *et al.*, 2005; Gotz *et al.*, 2008). Details of Blast2GO methods used are provided in Data S11z. A gene annotation file containing the information from this study was submitted to the Plant Ontology project. A chi-square test was performed to determine whether the observed GO distribution of non-sense SNP-containing genes was significantly different from the expected. Custom scripts were used to compare data sets to determine which contained unique or shared entries. Finally, KEGG pathway maps and corresponding information were downloaded from the KEGG Pathway Database through Blast2GO (http://www.genome.jp/kegg/pathway.html) (Kanehisa, 2002; Kanehisa *et al.*, 2012).

Author contributions

AD and HS led the sweet cherry genome data generation as part of the sweet cherry genome consortia with contribution from NO and LM. RPS led the almond genome data generation with contributions from FD, BLM, ME and RH. TK, SS and AD designed the analyses. TK and AH completed the reference mapping, mutation analyses and analysis of the *Sk* locus. SS performed the BLAST2GO analysis and processing. TK and SS performed statistical analyses. SS, TK and AD wrote the first draft of the manuscript. AD and RPS supervised the study. All authors contributed to, read and approved the final manuscript.

Acknowledgements

AD and NO would like to acknowledge the support received from WSU ARC startup and Hatch funds for this project. Washington Tree Fruit Research Commission support to AD and NO is gratefully acknowledged. TK and SS acknowledge support received from NIH Protein Biotechnology Training Program T32GM008336 and ARCS fellowships. AH was supported in part by a US Department of Agriculture National Research Initiative (USDA-NRI) grant 2008 - 35300-04676 to AD. RSP would like to thank 'Séneca Foundation' for the project 'Molecular Biology of Cyanogenesis in Almonds' and 'MINECO' for the project 'Mejora Genetica del Almendro'. RSP is also grateful for her postdoctoral contracts by CSIC (JAE Doc) and MINECO. HS is supported by CONICYT, FONDECYT/Regular No1120261 and Innova CORFO (07CN13 PBT-167). LM is supported by CONICYT, FONDECYT/Regular No1121021 and Innova CORFO (07CN13 PBT-167). BLM would like to thank the Villum research center Pro-Active Plants and the Novo Nordisk Foundation Center for Bio-Sustainability for financial support.

Conflict of interest

The authors declare no conflict of interest.

References

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

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M. and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Conn, E.E. (1969) Cyanogenic glycosides. *J. Agric. Food Chem.* **17**, 519–526.

Deschamps, S. and Campbell, M. (2010) Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Mol. Breed.* **25**, 553–570.

Dicenta, F. and García, J.E. (1993) Inheritance of the kernel flavor in almond. *Heredity*, **70**, 308–312.

Dicenta, F., Ortega, E. and Martínez-Gómez, P. (2007) Use of recessive homozygous genotypes to assess genetic control of kernel bitterness in almond. *Euphytica*, **153**, 221–225.

Ganal, M.W., Altmann, T. and Roder, M.S. (2009) SNP identification in crop plants. *Curr. Opin. Plant Biol.* **12**, 211–217.

Gong, Y., Fan, X. and Mattheis, J. (2002) Responses of 'Bing' and 'Rainier' sweet cherries to ethylene and 1-MethylCyclopropene. *J. Amer. Soc. Hort. Sci.* **127**, 831–835.

Gotz, S., Garcia-Gomez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talon, M., Dopazo, J. and Conesa, A. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435.

Hyten, D.L., Cannon, S.B., Song, Q.J., Weeks, N., Fickus, E.W., Shoemaker, R.C., Specht, J.E., Farmer, A.D., May, G.D. and Cregan, P.B. (2010) High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics*, **11**, 38 pp.

International_Peach_Genome_Initiative. (2013) *The high quality draft genome of peach (Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution.

Jung, S., Staton, M., Lee, T., Blenda, A., Svancara, R., Abbott, A. and Main, D. (2008) GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. *Nucleic Acids Res.* **36**, D1034–D1040.

Kanehisa, M. (2002) The KEGG database. *Novartis Found. Symp.* **247**, 91–103.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. and Tanabe, M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114.

Koepke, T., Schaeffer, S., Krishnan, V., Jiwan, D., Harper, A., Whiting, M., Oraguzie, N. and Dhingra, A. (2012) Rapid gene-based SNP and haplotype marker development in non-model eukaryotes using 3′ UTR sequencing. *BMC Genomics*, **13**, 18 pp.

Kulheim, C., Yeoh, S.H., Maintz, J., Foley, W.J. and Moran, G.F. (2009) Comparative SNP diversity among four *Eucalyptus* species for genes from secondary metabolite biosynthetic pathways. *BMC Genomics*, **10**, 452 pp.

Michelmore, R.W., Paran, I. and Kesseli, R.V. (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis – a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl Acad. Sci. USA*, **88**, 9828–9832.

National Agricultural Statistical Services (2011) Noncitrus Fruits and Nuts: 2010 Summary. United States Department of Agriculture Bethesda, MD.

Oliveros, J.C. (2007) *VENNY. An interactive tool for comparing lists with Venn Diagrams*.

Parida, S.K., Mukerji, M., Singh, A.K., Singh, N.K. and Mohapatra, T. (2012) SNPs in stress-responsive rice genes: validation, genotyping, functional relevance and population structure. *BMC Genomics*, **13**, 426.

Sánchez-Pérez, R., Jørgensen, K., Olsen, C.E., Dicenta, F. and Moller, B.L. (2008) Bitterness in almonds. *Plant Physiol.* **146**, 1040–1052.

Sánchez-Pérez, R., Howad, W., Garcia-Mas, J., Arus, P., Martinez-Gomez, P. and Dicenta, F. (2010) Molecular markers for kernel bitterness in almond. *Tree Genet. Genom.* **6**, 237–245.

Sánchez-Pérez, R., Belmonte, F.S., Borch, J., Dicenta, F., Moller, B.L. and Jorgensen, K. (2012) Prunasin hydrolases during fruit development in sweet and bitter almonds. *Plant Physiol.* **158**, 1916–1932.

Shulaev, V., Korban, S.S., Sosinski, B., Abbott, A.G., Aldwinckle, H.S., Folta, K.M., Iezzoni, A., Main, D., Arus, P., Dandekar, A.M., Lewers, K., Brown, S.K., Davis, T.M., Gardiner, S.E., Potter, D. and Veilleux, R.E. (2008) Multiple models for Rosaceae genomics. *Plant Physiol.* **147**, 985–1003.

Subbaiyan, G.K., Waters, D.L.E., Katiyar, S.K., Sadananda, A.R., Vaddadi, S. and Henry, R.J. (2012) Genome-wide DNA polymorphisms in elite indica rice inbreds discovered by whole-genome sequencing. *Plant Biotechnol. J.* **10**, 623–634.

Zhebentyayeva, T.N., Swire-Clark, G., Georgi, L.L., Garay, L., Jung, S., Forrest, S., Blenda, A.V., Blackmon, B., Mook, J., Horn, R., Howad, W., Arus, P., Main, D., Tomkins, J.P., Sosinski, B., Baird, W.V., Reighard, G.L. and Abbott, A.G. (2008) A framework physical map for peach, a model Rosaceae species. *Tree Genet. Genom.* **4**, 745–756.

Zirn, B., Wittmann, S. and Gessler, M. (2005) Novel familial WT1 read-through mutation associated with Wilms tumor and slow progressive nephropathy. *Am. J. Kidney Dis.* **45**, 1100–1104.

Supporting information

Additional Supporting information may be found in the online version of this article:

Data S1 Excel file of the mapping coverage for each scaffold of the peach genome for each sample. Blank entries are the result of no mapping.

Data S2 GFF annotation file (International_Peach_Genome_Initiative, 2013) of the peach genome used in these analyses as downloaded from GDR (Jung *et al.*, 2008).

Data S3 Fasta file of peach cDNA sequences.

Data S4 PPT file with a compressed bar graph depicting polymorphism rate in each 50kb window for each sample.

Data S5 Excel file of the 50 kb regions with significantly higher or lower polymorphism depth.

Data S6 GO-term composition of nonsense SNP-containing datasets separated by molecular function, biological process, and cellular component. Blast2GO was used to assign function to sequences predicted to have nonsense mutations. GO-terms were separated by percent composition for each dataset including the entire peach dataset. Comparison to the entire peach identifies GO-terms which may have higher or lower frequencies of developing nonsense-SNPs.

Data S7 Chi-square test of observed Gene Ontology distribution amongst datasets.

Data S8 KEGG pathways with members predicted to have non-sense SNPs.

Data S9 KEGG pathways with members in 'Response to Stress' gene ontology.

Data S10 Venn diagram of Peach genes containing non-sense mutations detected within the four investigated genotypes of almond. Sequences corresponding to mutations in the peach predicted genes were recorded for each almond genotype (Oliveros 2007).

Data S11 Blast2GO methods used in this study.

