Theor Appl Genet (2005) 110: 1419–1428
DOI 10.1007/s00122-005-1968-x

ORIGINAL PAPER

**Renate Horn · Anne-Claire Lecouls · Ann Callahan
Abhaya Dandekar · Lilibeth Garay · Per McCord
Werner Howad · Helen Chan · Ignazio Verde
Doreen Main · Sook Jung · Laura Georgi
Sam Forrest · Jennifer Mook · Tatyana Zhebentyayeva
Yeisoo Yu · Hye Ran Kim · Christopher Jesudurai
Bryon Sosinski · Pere Arús · Vance Baird · Dan Parfitt
Gregory Reighard · Ralph Scorza · Jeffrey Tomkins
Rod Wing · Albert Glenn Abbott**

Candidate gene database and transcript map for peach, a model species for fruit trees

Received: 9 August 2004 / Accepted: 15 November 2004 / Published online: 22 April 2005

Abstract Peach (*Prunus persica*) is a model species for the Rosaceae, which includes a number of economically important fruit tree species. To develop an extensive *Prunus* expressed sequence tag (EST) database for identifying and cloning the genes important to fruit and tree development, we generated 9,984 high-quality ESTs from a peach cDNA library of developing fruit mesocarp. After assembly and annotation, a putative peach unigene set consisting of 3,842 ESTs was defined. Gene ontology (GO) classification was assigned based on the annotation of the single "best hit" match against the Swiss-Prot database. No significant homology could be found in the GenBank nr databases for 24.3% of the sequences. Using core markers from the general *Prunus* genetic map, we anchored bacterial artificial chromosome (BAC) clones on the genetic map, thereby providing a framework for the construction of a physical and transcript map. A transcript map was developed by hybridizing 1,236 ESTs from the putative peach unigene set and an additional 68 peach cDNA clones against the peach BAC library. Hybridizing ESTs to genetically anchored BACs immediately localized 11.2% of the ESTs on the genetic map. ESTs showed a clustering of expressed genes in defined regions of the linkage groups. [The data were built into a regularly updated Genome Database for Rosaceae (GDR), available at (http://www.genome.clemson.edu/gdr/).]

Introduction

Peach [*Prunus persica* (L.) Batsch] has become a model species for genetic studies within the Rosaceae (Abbott

Communicated by H.C. Becker

R. Horn · A.-C. Lecouls · L. Garay · D. Main · S. Jung
L. Georgi · S. Forrest · J. Mook · C. Jesudurai
J. Tomkins · A. G. Abbott (✉)
Department of Genetics,
Biochemistry and Life Science Studies,
Clemson University, Clemson, SC 29634, USA
E-mail: aalbert@clemson.edu
Tel.: +1-864-656-3060
Fax: +1-864-656-6879

A. Callahan · R. Scorza
USDA Appalachian Fruit Research Station,
Kearneysville, WV 25430, USA

A. Dandekar · H. Chan · D. Parfitt
Department of Pomology, University of California Davis,
Davis, CA 95616, USA

P. McCord · B. Sosinski
Department of Horticultural Science,
North Carolina State University,
Raleigh, NC 27695, USA

W. Howad · P. Arús
Departament de Genètica Vegetal,
Institut de Recerca I Tecnologia Agroalimentàries,
08348 Cabrilis, Spain

I. Verde
Istituto Sperimentale per la Frutticoltura (ISF),
00040 Rome, Italy

D. Main · S. Jung · C. Jesudurai · J. Tomkins
Clemson University Genomics Institute,
Clemson University,
SC 29634, USA

T. Zhebentyayeva · V. Baird · G. Reighard
Horticulture Department,
Clemson University,
Clemson, SC 29634,
USA

Y. Yu · H. R. Kim · R. Wing
Plant Sciences Department,
University of Arizona,
Tucson, AZ 85721-0036, USA

et al. 2002). The diploid genome ($2n = 16$) (Jelenkovic and Harrington 1972), the small genome size of 300 Mb (Baird et al. 1994) and the relatively short generation time for a fruit tree (2–3 years until flowering) facilitate genetic studies in this species compared to any of the polyploid species of this family.

Several genomic maps have been constructed for peach (Chaparro et al. 1994; Dirlewanger and Bodo 1994; Rajapakse et al. 1995; Dirlewanger et al. 1998; Lu et al. 1998; Shimada et al. 2000), almond (Viruel et al. 1995; Joobeur et al. 2000), interspecific crosses between almond and peach (Foolad et al. 1995; Joobeur et al. 1998; Bliss et al. 2002; Aranzana et al. 2003) and for other closely related fruit trees like apricot (Hurtado et al. 2002) and cherry (Wang et al. 1998). From these maps, a general *Prunus* genetic map that was developed (Joobeur et al. 1998) based on the interspecific cross between almond and peach serves as a reference for genome analysis in the genus *Prunus*. Recently, 96 simple sequence repeat (SSR) markers have been added to this general map (Joobeur et al. 1998), and 24 single-locus SSRs highly polymorphic in peach and covering the whole genome have been proposed for a "genotyping set" useful as a reference for fingerprinting, pedigree and comparative genetic analyses (Aranzana et al. 2003). SSR markers developed by several other groups have also been shown to be useful in characterization and genetic diversity studies across species in the genus *Prunus* (Cipriani et al. 1999; Sosinski et al. 2000; Dirlewanger et al. 2002; Wang et al. 2002). In addition to the marker and map development, large-insert genomic libraries have been constructed for peach (Wang et al. 2001; Georgi et al. 2002; Georgi et al. unpublished). The bacterial artificial chromosome (BAC) library developed from the DNA of the peach rootstock Nemared represents a theoretical eight- to ninefold haploid genome equivalent (Georgi et al. 2002), while the BAC library from the haploid peach cultivar Lovell covers a nine- to tenfold haploid genome equivalent (Georgi et al. unpublished). Wang et al. (2001) used the traditional cultivar Jingyu to develop a BAC library with an average insert size of 95 kb and a sevenfold coverage.

Fruit quality determines the economic value of peach. Traits such as peach/nectarine (Chaparro et al. 1994; Rajapakse et al. 1995; Bliss et al. 2002), melting flesh/stony hard flesh (Warburton et al. 1996), freestone/clingstone (Warburton et al. 1996; Dettori et al. 2001), polycarpel (Bliss et al. 2002) and flesh color (Rajapakse et al. 1995; Warburton et al. 1996; Bliss et al. 2002) have been mapped. Quantitative trait loci (QTLs) for traits such as soluble sugars (sucrose, fructose, glucose and sorbitol) and organic acids (malic, citric and quinic acid), which determine fruit quality in peach, have been localized on maps (Dirlewanger et al. 1999; Etienne et al. 2002). A candidate gene approach has been used to identify the genes placed within these QTL intervals. However, progress in identifying candidate genes controlling many important fruit tree characters is hampered by the lack of comprehensive transcript and physical maps.

Therefore, we initiated a peach expressed sequence tag project with the goal of developing an extensive peach EST database for identifying and cloning genes important to fruit and tree development. We sequenced 9,984 high-quality peach ESTs and assembled and annotated these into contigs and singletons to define the first putative unigene set of peach. ESTs represent a valuable source for developing markers either by directly using them as probes (Wu et al. 2002) or by developing microsatellite markers (Kantety et al. 2002; Thiel et al. 2003) or markers based on single nucleotide polymorphisms (SNPs)(Bundock et al. 2003; Neuhaus and Horn 2004) from the sequences, which can then be genetically mapped. Their genetic map location may be determined by hybridizing them against previously marker-anchored contigs or BACs (i.e. placing them on an integrated physical/genetic map). We applied this approach to localize ESTs on the *Prunus* general genetic map and to develop a transcript map for peach.

We report here the development of a peach EST database and the construction of a transcript map using the peach unigene set as probes. To facilitate gene identification and functional studies in peach, we annotated the fruit EST set using the structured vocabulary provided by the Gene Ontology Consortium (2001). The peach EST database provides a very valuable resource that will considerably enhance the isolation and characterization of agronomical important genes in Rosaceae.

Materials and methods

Fruit cDNA libraries

RNA was extracted according to Callahan et al. (1992) to develop the peach unigene set (PP_LE) from fruit collected at the mature picking (8/10) and at the eating ripe (8/17) stages from a doubled haploid peach selection, P-21-5-2N. This doubled haploid peach line is homozygous at all loci, so similar sequence ESTs should represent different genes (i.e., family members) and not different alleles of the same gene. Extracted RNA was treated with DNase using a DNA-free kit (Ambion, Austin, Tex.) according to the manufacturer's directions. PolyA + RNA was isolated using a PolyATtract IV kit according to manufacturer's directions (Promega, Madison, Wis.). Five micrograms of polyA + RNA, as measured by absorbance at 260 nm, was used to construct a cDNA library in the Uni-ZAP XR vector l using the Lambda cDNA construction kit (Stratagene, La Jolla, Calif.) according to manufacturer's directions. This vector permits directional cloning of the cDNA into the *Xho*I and *Eco*RI sites and excision as pBluescript SK. The titer of the un-amplified library was approximately 2.5×10^5 plaques per milligram vector DNA. Twenty individual phage plaques were placed individually in 5 ml of NZY broth (Gibco, Gaithersburg, Md.) and grown overnight. Following chloroform

treatment and a brief centrifugation to pellet the bacterial debris, five ml of each culture was subjected to PCR (AmpliTaq; Applied Biosystems, Foster City, Calif.) with M13 forward and reverse primers using the following cycle conditions: one cycle at 95°C for 10 min, one cycle at 80°C for 30 min, then, after addition of Taq, 30 cycles at 95°C for 1 min, 55°C for 1 min, and 72°C for 1 min. When the amplified inserts were resolved on a gel the average insert size was greater than 500 bp. The library was then converted to plasmids. Following mass excision of 100 ml of the primary library according to manufacturer's instructions (Stratagene), cells were infected with the resulting phagemids, yielding over 10,000 individual plasmid containing clones.

A second cDNA library (LF) was constructed in the same way from fruit of the peach cultivar Loring, which had 21–40 N firmness (nearly ripe) and 41–60 N firmness (early ripe).

Sequencing

Plasmid preparations were performed in a 96-well format according to the protocol given at the web site http://www.genome.arizona.edu/information/protocols/prep_web.ppt. Plasmids were sequenced by ABI Prism BigDye Terminator Cycle Sequencing (Applied Biosystems, Foster City, Calif.) using T7 primer. Sequence reactions were analyzed on an ABI Prism 3700 Sequencer.

EST processing and annotation

EST data processed at Clemson University Genomic Institute (CUGI) utilizes publicly available software incorporated into a fully automated in-house developed script (PROCEST). The processing occurs in three stages

In Stage I, which consists of trace file processing, sequence trace files are converted into fasta files and quality-score files using the PHRED base-calling program (Ewing et al. 1998). Vector and host contamination are identified and masked using the sequence comparison program CROSS_MATCH (Gordon et al. 1998). Vector trimming excises the longest non-masked sequence, and further trimming removes low-quality bases (PHRED score less than 20) at both ends of a read. Sequences are discarded if they have more than 5% ambiguous bases, more than 40 PolyA or PolyT bases, or fewer than 100 high-quality bases (minimum phred score of 20). At this stage of processing the script generates an overall summary report file, clone report tables, a Genbank submission file and FASTA-formatted library files of the high-quality trimmed sequences and associated quality values. The FASTA library is further filtered to remove reads having significant similarity with mitochondrial, rRNA, tRNA, or snoRNA sequences downloaded from the Genbank nucleotide database.

In Stage II processing, which consists of the assembly of high-quality sequences, the filtered library file is assembled using the contig assembly program CAP3 (Huan and Madan 1999). More stringent parameters (-p 95, -d 60) are typically used to prevent over-assembly and help identify potential paralogs.

In Stage III processing, which consists of annotation, both the filtered library and the contig consensus library file are compared pairwise against the GenBank nr protein database using the FASTX3.4 algorithm (Pearson and Lipman 1988). The most significant matches ($< 1\times10^{-9}$) for each contig and individual clones in the library are recorded. The script generates a web page, which displays the best protein match for each contig and singleton. The unigene data set was derived by selecting the clone that best represented each contig, and the singletons that had either unique protein matches ($< 1\times10^{-9}$) or no significant matches. The sequence, assembly and homology data were stored in an Oracle relational database management system, facilitating efficient data querying and display. From our newly created Genome Database for Rosaceae resource (http://www.genome.clemson.edu/gdr/), users can view contig assembly, clones and annotation, download the library and unigene sequence libraries and search their sequences against the EST database using our BLAST/FASTA server facility.

High-density BAC filters

The BAC library, prepared from leaves of the peach rootstock Nemared (Georgi et al. 2002), was used to make high-density BAC filters. This BAC library, which consisted of 44,160 clones (eight- to ninefold coverage of the genome), was spotted onto three filters using a 4×4 pattern. Colonies on the filters were incubated for 16 h at 37°C on LB agar plates containing 12.5 μg/ml chloramphenicol. Filters were denatured for 7 min in denaturing solution (1.5 *M* NaCl, 0.5 *M* NaOH), followed by 7 min in neutralization solution (1.5 *M* NaCl, 0.5 *M* Tris/HCl pH 7.2, 1 m *M* EDTA) and finally were washed for 1 min in 2× SSC.

Hybridization of the ESTs to the high-density BAC filters

Bacteria were grown for 4 h at 37°C in shaking culture at 280 rpm. The insert was amplified by PCR from 1 μl of these cultures using T7 and T3 primers in 1 U AmpliTaq polymerase, 10 pmol of each primer, and 0.25 m *M* dNTP in 1× PCR buffer (Applied Biosystems, containing 1.5 m *M* $MgCl_2$). After a 2-min denaturation at 2 min 94°C, DNA amplification was performed for 30 cycles of 1 min at 94°C (denaturing), 2 min at 55°C (annealing) and 2 min 72°C (polymerization), with a final extension for 4 min at 72°C. PCR products were separated on a 2% agarose gel to estimate the amount of product. The PCR product in a 20-μl volume corresponding to 150 ng was denatured for 5 min at 94°C and cooled down on ice before 10 μl of the labeling mix was added to each sample. The labeling mix contained 6.0 μl

OLB [Sol A:Sol B:Sol C = 1:2.5:1; Sol O contained1.25 *M* Tris/HCl pH 8.0; Sol A contained 1 ml solution O, 18 µl *β*-mercaptoethanol, 5 µl each of dTTP (0.1 *M*), dGTP (0.1 *M*), dATP (0.1 *M*); Sol B contained 2 *M* HEPES/NaOH; Sol C contained pd(N)6 random hexamer at 90 OD units/ml), 1.2 µl BSA (1 mg/ml), 0.3 µl Klenow (5 U/ µl), 1 µl α-[^{32}P]dCTP (10 μCi) and 1.5 µl H_2O. Samples were incubated at 37°C for 1 h. Labeled probes were separated from unincorporated nucleotides on Sephadex G50 spin columns.

BAC filters were prehybridized in 30 ml of hybridization buffer [0.25 *M* Na phosphate buffer pH 7.2, 7% sodium dodecyl sulfate (SDS)] at 65°C for at least 1 h. The 8–16 labeled PCR products from individual ESTs were bulked and denatured for 10 min at 94°C before being added to the hybridization buffer. Hybridization was performed overnight at 65°C in a hybridization oven. The filters were washed twice for 40 min each time at 65°C in 2× SSC containing 0.1% SDS in the tubes and then once for 10 min at 65°C in 1× SSC containing 0.1% SDS. The filters were exposed to X-ray films for 3–4 days. For rehybridizations, BAC filters were stripped by boiling in 0.5% SDS solution for 20 min.

Positive BAC clones were verified and assigned to individual probes by rehybridization to colony dot blots. Detected BACs were grown in a 96-well format (100 µl LB medium, 12.5 µg/ml chloramphenicol) overnight. From these plates, bacterial clones were stamped onto a nylon membrane (up to 48 BAC clones/membrane) and grown overnight on LB agar plates (12.5 µg/ml chloramphenicol) at 37°C. Bacteria were lysed for 7 min in denaturing solution (1.5 *M* NaCl, 0.5 *M* NaOH), then for 7 min in neutralization solution (1.5 *M* NaCl, 0.5 *M* Tris/HCl pH 7.2, 1 m *M* EDTA). Dot blots were briefly washed in 2× SSC solution and dried. After UV crosslinking the dot blots were hybridized to individual probes as described for the BAC filters.

Results

Generation and assembly of fruit ESTs

We generated a cDNA library from the developing fruit mesocarp of peach to provide a resource of expressed genes important for fruit development and to establish a transcript map for peach based on the *Prunus* general genetic map (Joobeur et al. 1998).

The insert size of the clones for the peach cDNA library ranged from 0.3 kb to 2.4 kb, estimated from PCR amplification of 156 clones. The average sequence length was 1.2 kb. All 13,331 cDNA clones were 5′-sequenced with an overall success rate of 75%, calculated after the removal of poor-quality and vector sequences. This resulted in 9,984 successful reads with an average length of 502 bp. These sequences were submitted to NCBI GenBank dbEST (accession nos. BU039022 through BU49005). Using the assembly program CAP3, the ESTs were assembled into 1,309 contigs and 3,500 singletons. Annotation consisted of a pairwise comparison of both the filtered library and the contig consensus library against the GenBank nr protein database using the FASTX3.4 algorithm. Of the 1,309 contigs, 1,200 were found to have significant matches ($< 1\times10^{-9}$), while 109 had no significant matches and were annotated as putatively unknown. The 3,500 singletons identified by the assembly process were further filtered out by removing those with a similar protein match to either contigs, the clones comprising the contigs or other singletons. This resulted in 2,533 unique singletons, of which 1,708 had significant matches ($< 1\times10^{-9}$), and 825 had no matches. The tentative unigene set for peach was derived by selecting the clones that best represented the contigs and the unique singletons. This annotated set consists of 3,842 putative unique genes arrayed in 96-well microtiter plates, which are publicly available at the Clemson University Genomics Institute web site (http://www.genome.clemson.edu). It is notable that 24.3% of the ESTs of the putative peach unigene set had no significant homology in the NCBI Genbank. These sequences are of special interest as some may be unique to fruiting tree species, thus worthy of future study.

Functional annotation of fruit ESTs

We characterized the *Prunus persica* EST sequences with respect to functionally annotated genes in the Swiss-Prot database. Of the 1,552 sequences from the putative peach unigene set that had matches with the Swiss-Prot database, 1,439 could tentatively be assigned gene ontology (GO) classifications based on the annotation of the single "best hit" match against the Swiss-Prot database ($< 1\times10^{-9}$). Functional assignments of peach ESTs described here are at the "inferred from electronic annotation" (IEA) level of evidence (see The Gene Onotology Consortium 2001). Figure 1 summarizes the assignments of peach sequences to major molecular functions and biological processes. For the molecular functions, peach ESTs were assigned to nine GO terms, of which 76.6% are covered by three GO terms: 57.5% catalytic activity (GO: 0003824), 9.7% binding (GO: 0005488), and 9.4% transporter activity (GO: 0005215). For the biological processes, peach ESTs were assigned to three GO terms (Fig. 1b), with the GO term physiological process (GO: 0007582) representing 59.4% of the assigned ESTs. Annotation of peach EST sequences with respect to the GO terms molecular function, biological process and cellular component are regularly updated and can be accessed at http://www.genome.clemson.edu/gdr/.

Physical map framework

To anchor BACs onto the genetic map and to provide a framework for the physical map, we used 141 probes representing different marker types (RFLP, SSR and

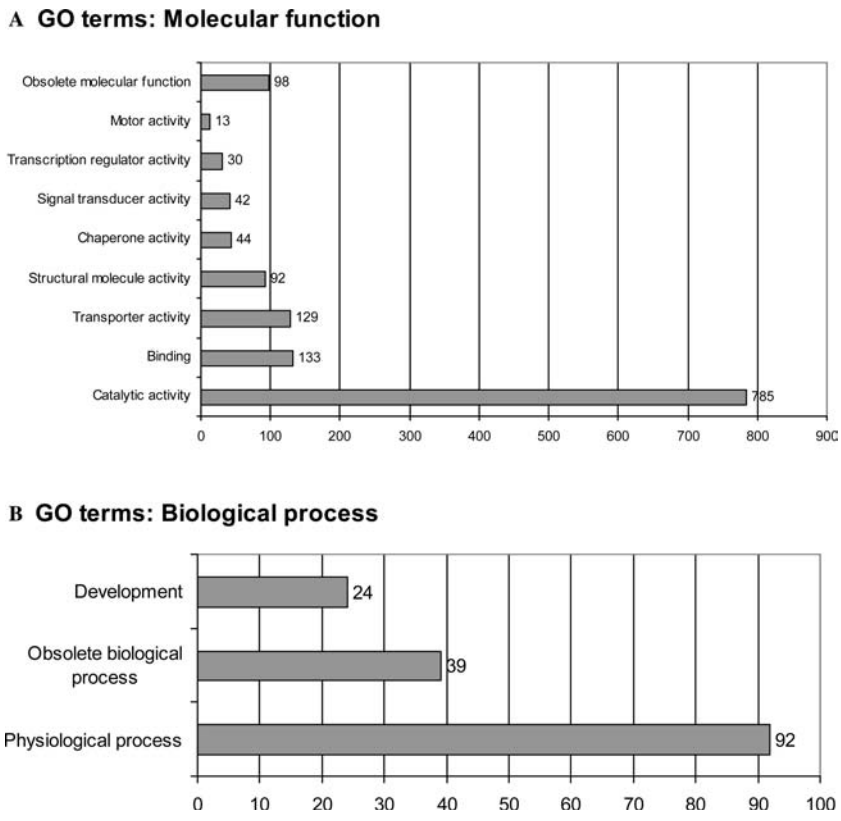

Fig. 1 General statistics for the number of proteins in the peach proteome that were assigned to high-level gene ontology (GO) terms from each of the two gene ontologies, molecular function and biological process. ESTs may be assigned to more than one GO term. Also note that child terms (not shown) may have more than one parent term (for example, ''hydrolase/hydrolyzing glycosyl compounds'' is a child of both ''enzyme'' and ''defense/immunity protein''). **a** Molecular function. **b** Biological process

AFLP markers from peach as well as RFLP markers from *Prunus ferganensis*, cherry and almond) to screen the *Hin*dIII peach BAC library (theoretical eight- to ninefold haploid genome equivalents). The majority of these markers was derived from the general *Prunus* genetic map based on the cross 'Texas' × 'Earlygold', which consists of 246 markers (Joobeur et al. 1998). This map covers 491 cM of the peach genome. In addition, we implemented a ''neighbors'' map approach in which we extrapolated locations of loci from other maps to their nearest neighbors in the general *Prunus* map. Shared loci on the two genetic maps defined an interval containing the locus of interest. The 141 probes represent 153 markers, with an average spacing of 4 cM between the markers if loci with two or more markers are counted as one (Fig. 2). About one-half of the core markers are single-copy, whereas the remaining probes are present in two or more copies in the genome. Sixty-nine single-copy core markers identified an average of 4.1 BACs with a range of 1–15 BACs per probe. As the library represents a theoretical eight- to ninefold haploid genome equivalent, each single-copy sequence should be present about eight times if all sequences are equally represented in the library. The lower number of BAC clones identified is probably due to the high stringency used in identifying positive clones, which might miss clones with weak hybridization signals but gives a higher reliability for the detected positive clones by reducing the number of false positives.

The remaining core markers represent sequences that are present in the genome in more than one copy. Seventy-two probes with two or more copies gave 1–30 positive BAC clones with an average of 5.5 hits per probe. The hybridization results confirm that the peach BAC library provides good genome coverage, since positive BAC clones were detected with all 153 core markers distributed all over the peach genome. In total, the core markers identified and anchored 679 BACs on the genetic map.

By anchoring BACs on the genetic map using core markers, a framework for the physical map has been established. This framework also provides a basis for assigning physical/genetic locations for our ESTs, thus, constructing a transcript map.

Transcript map of peach fruit ESTs

To develop a transcript map, 1,236 ESTs from the peach unigene set (PP_LE) and an additional 68 cDNA clones (LF) from a second fruit-specific cDNA library were hybridized to the three high-density peach BAC filters representing 44,160 BAC clones. As for the 141 core probes, hybridization of the BAC filters was performed using bulks of 8–16 probes. Positive BAC clones were then verified by hybridization to single EST probes by dot blot hybridizations. This represents a very economical way to perform the hybridizations (bulks) and

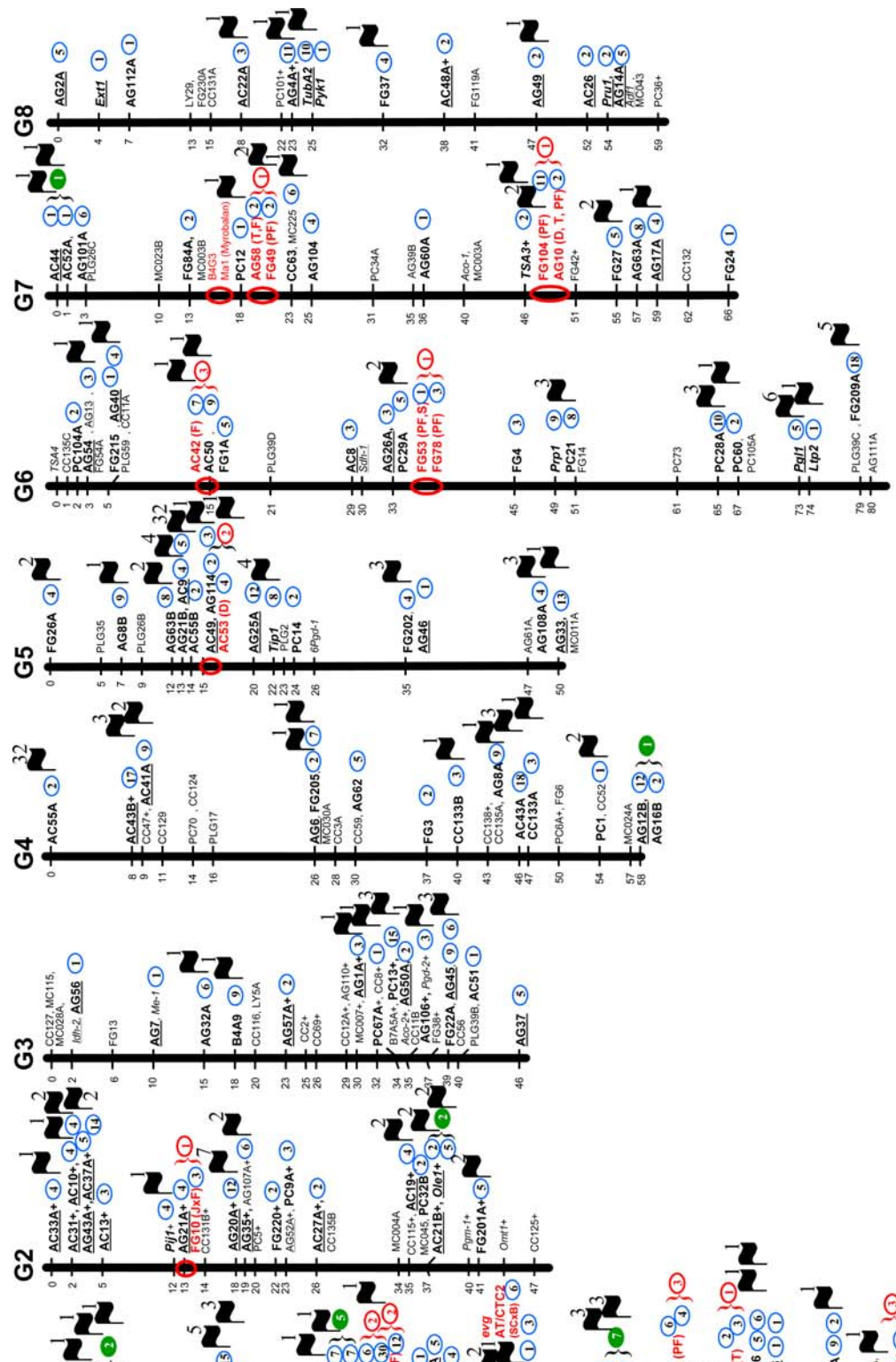

Fig. 2 Genetically anchored BACs on the general *Prunus* genetic map (Joobeur et al. 1998) and development of a transcript map. Markers depicted in *bold* have been used to anchor contigs of peach BACs on the genetic map. Markers followed by + represent markers that showed a strong hybridization signal and, in addition, a second weaker signal was obtained with this probe, which was either polymorphic or not. Markers *underlined* were used to join different maps. *Circles* on the linkage group mark areas where markers from other maps have been integrated. The number of positive BACs detected by a marker is shown as a *black number in circles* behind the markers. In addition, *brackets* followed by the corresponding number (*white numbers on gray background*) indicate BACs detected by two neighboring markers. *Black flags* behind the markers represent peach EST positions. The *number on top of a flag* gives the number of ESTs mapped to this marker. Abbreviations for integrated maps: *D P. davidiana* (Dirlewanger et al. 1996), *F* 'Ferragnes' (Viruel et al. 1995), *J* × *F* 'Jalousia' × 'Fantasia' (Dirlewanger et al. 1999), *P* × *A* 'Padre' × '54P455' (Bliss et al. 2002), *P* × *F* ('IF310828' × *P. ferganensis*) × 'IF310828' (Dettori et al. 2001), *SC* × *B* 'Suncrest' × 'Bailey' (Abbott et al. 1998), *S* 'Summergrand' (Dirlewanger et al. 1996), *T* 'Tuono' (Viruel et al. 1995)

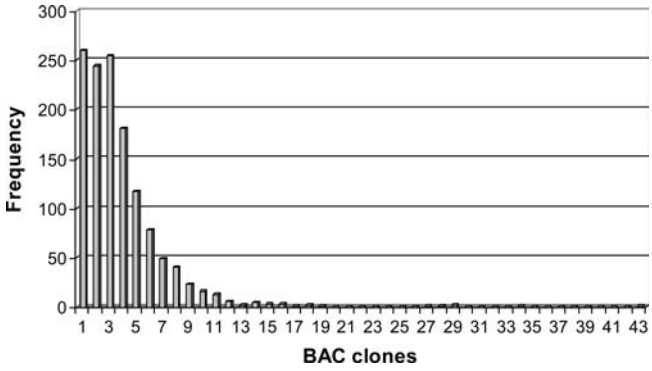

Fig. 3 Frequency of BACs detected by peach EST probes

to reliably identify positive BAC clones, as the dot blots represent a second hybridization. EST probes that hybridized to BACs that had been previously anchored by core markers, can be presumed to map to the same location on the general *Prunus* genetic map as the corresponding marker.

Each EST probe detected between 1 and 43 BAC clones (Fig. 3). On average, 3.8 BAC clones hybridized to each EST, resulting in a total number of 4,983 BAC clones attached to EST probes. In total, 147 EST probes shared BACs with previously marker-anchored BACs, which immediately localized the ESTs on the genetic map. The ESTs mapped to 79 marker locations corresponding to 73 core probes. The transcript map developed on this basis is shown in Fig. 2. Localization and annotation of these ESTs are given in the Electronic Supplementary Material (ESM) Table 1. The ESTs are sorted according to the location of the core markers on the linkage groups of the general *Prunus* genetic map.

Between 1 and 32 ESTs were mapped to single locations, with an average of two ESTs per map position (Fig. 4). ESTs showed a clustering in certain regions of the linkage groups. On linkage groups G4 and G5, 32 ESTs hybridized to two BACs (028F08 and 082I18), detected also by the restriction fragment length polymorphism (RFLP) probe AC55. As probe AC55 mapped to two locations (linkage groups G4 and G5) in the peach genome, it cannot be distinguished at this point whether the ESTs are divided between the two or if only one of them contains all the ESTs. To identify an overlap between 028F08 and 082I18, the two BAC clones were digested with *Hin*dIII and hybridized using one of the BAC clones as a probe, respectively. A hybridization signal of 7 kb was common to both BAC clones, demonstrating an overlap or duplication of 7 kb between the two BACs (data not shown). This likely explains why a number of EST probes hybridized to both BAC clones, whereas others hybridized to only one of them (ESM Table 1). The cluster of ESTs anchored to AC55 consists of ESTs that in most cases represent singletons with no match to the database, indicating that these regions might contain a cluster of genes specific to fruit trees or to peach. Five ESTs show different percentages of homology to the allergen protein from *Prunus armeniaca* and might represent members of a gene family in this region. Other clusters of genes containing three to seven ESTs were detected on all linkage groups except for G7 and G8. Hybridizing the ESTs to genetically anchored BACs immediately provided genetic map localizations for 11.2% of the ESTs and resulted in the first transcript map for peach.

All hybridization data are incorporated in the Genome Database for Rosaceae (http://www.genome.clemson.edu/gdr/), which is publicly available and regularly updated at Clemson University. The database can be searched for BACs, ESTs, markers and maps.

Discussion

By sequencing 13,331 cDNA from the developing mesocarp of peach fruit, we obtained 9,984 successful reads and submitted these to NCBI GenBank dbEST. After assembly and annotation, these 5′-sequences were used to define a putative peach unigene set of 3,842 ESTs. EST assembly tends to overestimate the actual number of genes represented because a failure of the ESTs to assemble can result from non-overlapping ESTs, alternative splicing, sequencing errors and sequence polymorphism. To reduce redundancy, we filtered the singletons for similar protein matches and only selected the clone with the most significant match to represent the gene.

GO terms were assigned to the "best hit" match against the Swiss-Prot database. Gene ontology helps to describe gene products in a standardized way and thus facilitates cross-species comparison (Camon et al. 2003).

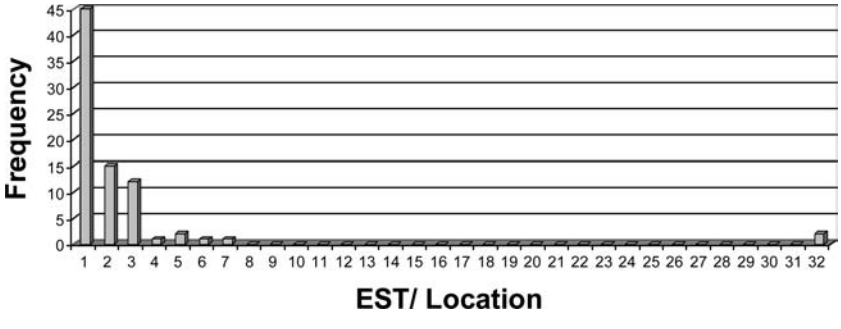

Fig. 4 Localization of EST probes on the general *Prunus* map. The frequency of localizing peach ESTs at a location anchored by a marker is shown

The number of ESTs assigned to each GO term was calculated to allow a comparison of the relative percentages of the ESTs assigned to each term. However, ESTs may be assigned to more than one GO term, therefore ESTs may be counted more than once. For molecular functions, the ESTs from the developing fruit cDNA library show the highest percentage for the category catalytic activity, which is consistent with the expected high metabolic activity in the developing tissue. For biological processes, the highest number of ESTs was assigned to the category physiological process. Deeper levels of GO terms are shown at the web site. The assignment of GO terms provides a powerful tool for researchers seeking to find candidate genes for the trait of interest.

About 24% of the ESTs from the peach unigene set were not homologous with sequences in the GenBank nr protein database. Also, in citrus, 17% of the ESTs, developed from 180-day-old whole immature seedlings failed to match with significance to any protein sequence found in public databases (Bausher et al. 2003). In *Vitis*, 12% of the ESTs showed no significant homology to any of the deposited sequences (Terrier et al. 2001). The percentage of sequences lacking homology will decrease as more sequence data are added to the public genome databases. Lack of homology might also be related to sequencing, post-processing and annotation errors. However, some of these sequences might represent genes unique to fruit trees, or sequences that might have evolved rapidly within the species. Therefore, some of these sequences might represent the most interesting for further investigations of fruit trees.

Using core markers from the general *Prunus* genetic map (Joobeur et al. 1998), we anchored BAC clones on the genetic map, thereby providing a framework for the physical and the transcript map. The development of a transcript map through the hybridization of ESTs to previously genetically anchored BAC clones proved to be an efficient approach for the construction of a transcript map. Using this approach, 11.2% of the ESTs were immediately assigned to locations on the general *Prunus* genetic map. In total, 147 ESTs were positioned, distributed over all linkage groups. These ESTs can provide additional SSR and SNP anchor loci for the map.

A cluster of 32 ESTs that mapped to the marker AC55 is of special interest because most of the ESTs were not homologous to sequences in the NCBI database. Five of the ESTs in that cluster showed different percentages of homology to the *P. armeniaca* putative allergen protein, indicating the presence of a gene family for this protein in the region. A number of ESTs from this region might be unique to fruit trees or might have rapidly evolved from a common ancestor to fulfill new functions in fruit trees. It will be interesting to investigate whether the genes representing these ESTs are also present in other fruit trees (e.g., apricot or apple) and whether they also cluster in one region. A clustering of ESTs, even though not that pronounced, was also observed on other linkage groups apart from linkage groups 7 and 8. Previous sequencing of single peach BAC clones also revealed concentrations of potential gene encoding regions followed by SSR-rich regions (Georgi et al. 2002).

The putative peach unigene set provides a valuable source for probes to be mapped on the general *Prunus* map, for candidate gene approaches in isolating genes, and for the development of microarrays for global differential gene expression analysis. To increase the number of possible anchors for the framework of the physical map, we are aiming at developing PCR-based genetic markers from the ESTs of the peach unigene set and mapping these on the general *Prunus* map. In peach, ESTs (Jung et al. 2004) as well as BAC clones (Georgi et al. 2002; Wang et al. 2002) have proven to be a valuable source material for the development of SSR markers. To extend the transcript map, additional ESTs from the peach unigene set will be hybridized on the BAC library filters. The development of new cDNA libraries for peach shoots, flower buds and disease-resistant roots are in progress to permit integration of transcribed sequences from other tissues into the transcript map. The established EST database for peach that we report here represents a unique resource for cloning agronomical important genes in Rosaceae using candidate gene approaches. This will accelerate the isolation of genes and their association with traits of interest and will facilitate the analysis of their differential expression. All our *Prunus* structural and functional genomics resources such as BAC libraries (peach, cherry, plum and apricot) and EST unigene libraries (peach and almond) are housed at the Clemson University Genome Institute and are publicly available through the online ordering system (http://www.genome.clemson.edu/orders).

Acknowledgements We would like to thank the USDA IFAFS program for supporting this research by the award of no. 2001-52100-11345, the NSF PGR program for supporting the Genome Database for Rosaceae award no. 0320544, the Clemson University for supporting the Bioinformatics analysis, the South Carolina Peach Council and the South Carolina Agricultural Experimental Station for grant SC-1700120 within the SCAFR program as well as the OECD for the fellowship to I. Verde at Clemson University under the OECD co-operative research program: "Biological resource management for sustainable agricultural systems".

References

Abbott AG, Rajapakse S, Sosinski B, Lu ZX, Sossey-Alaoui K, Gannavapura M, Reighard G, Ballard RE, Baird WV, Scorza R, Callahan A (1998) Construction of saturated linkage maps of peach crosses segregating for characters controlling fruit quality, tree architecture and pest resistance. Acta Hortic 465:41–49

Abbott A, Georgi L, Yvergniaux D, Inigo M, Sosinski B, Wang Y, Blenda A, Reighard G (2002) Peach: the model genome for Rosaceae. Acta Hortic 575:145–155

Aranzana MJ, Pineada A, Cosson P, Dirlewanger E, Ascasibar J, Cipriani G, Ryder CD, Testolin R, Abbott A, King GJ, Iezzoni

AF, Arús P (2003) A set of simple-sequence repeat (SSR) markers covering the *Prunus* genome. Theor Appl Genet 106:819–825

Baird WV, Estager AS, Wells J (1994) Estimating nuclear DNA content in peach and related diploid species using laser flow cytometry and DNA hybridization. J Am Soc Hortic Sci 119:1312–1316

Bausher M, Shatters R, Chaparro J, Dang P, Hunter W, Niedz R (2003) An expressed sequence tag (EST) set from *Citrus sinensis* L. Osbeck whole seedlings and the implications of further perennial source investigations. Plant Sci 165:415–422

Bliss FA, Arulsekar S, Foolad MR, Becerra V, Gillen AM, Warburton ML, Dandekar AM, Kocsisne GM, Mydin KK (2002) An expanded genetic linkage map of *Prunus* based on an interspecific cross between almond and peach. Genome 45:520–529

Bundock PC, Christopher JT, Eggler P, Ablett G, Henry RJ, Holton TA (2003) Single nucleotide polymorphisms in cytochrome P450 genes from barley. Theor Appl Genet 106:676–682

Callahan AM, Morgens PH, Wright P, Nichols KE Jr (1992) Comparison of pch313 (pTOM13 homology) RNA accumulation during fruit softening and wounding of two phenotypically different peach cultivars. Plant Physiol 100:482–488

Camon E, Magrane M, Barrell D, Binns D, Fleischmann W, Kersey P, Mulder N, Oinn T, Maslen J, Cox A, Apweiler R (2003) The gene ontology annotation (GOA) project: Implementation of GO in SWISS-PROT, TrEMBL and InterPro. Genome Res 13:662–672

Chaparro JX, Werner DJ, O'Malley D, Sederoff RR (1994) Targeted-mapping and linkage analysis of morphological, isozyme, and RAPD markers in peach. Theor Appl Genet 87:805–815

Cipriani G, Lot G, Huang WG, Marrazzo MT, Peterlunger E, Testolin R (1999) AC/GT and AG/CT microsatellite repeats in peach [*Prunus persica* (L) Batsch]: isolation, characterization and cross-species amplification in *Prunus*. Theor Appl Genet 99:65–72

Dettori MT, Quarta R, Verde I (2001) A peach linkage map integrating RFLPs, SSRs, RAPDs, and morphological markers. Genome 44:783–790

Dirlewanger E, Bodo C (1994) Molecular genetic mapping of peach. Euphytica 77:101–103

Dirlewanger E, Pascal T, Zuger C, Kervella J (1996) Analysis of molecular markers associated with powdery mildew resistance genes in peach (*Prunus persica* (L.) Batsch) × *Prunus davidiana* hybrids. Theor Appl Genet 93:909–919

Dirlewanger E, Pronier V, Parvey 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, Moing A, Rothan C, Svanelle L, Pronier V, Guye A, Plomion C, Monet R (1999) Mapping QTLs controlling fruit quality in peach [*Prunus persica* (L.) Batsch]. Theor Appl Genet 98:18–31

Dirlewanger E, Cosson P, Tavaud M, Aranzana MJ, Poizat C, Zanetto A, Arús P, Liagret 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.). Theor Appl Genet 105:127–138

Etienne C, Rothan C, Moing A, Plomion C, Bodnes 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

Ewing B, Hiller L, Wendl M, Green P (1998) Base calling of automated sequence traces using PHRED I. Accuracy assessment. Genome Res 8:175–185

Foolad MR, Arulsekar S, Becerra V, Bliss FA (1995) A genetic linkage map of *Prunus* based on an interspecific cross between peach and almond. Theor Appl Genet 91:262–269

The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. Genome Res 11:1425–1433

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

Gordon D, Abanjian C, Green P (1998) CONSED: a graphical tool for sequence finishing. Genome Res 8:195–202

Huan X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9:868–877

Hurtado MA, Romero C, Vilanova S, Abbott AG, Llacer G, Badenes M (2002) Genetic linkage maps of two apricot cultivars (*Prunus armeniaca* L.) and mapping of PPV (Sharka) resistance. Theor Appl Genet 105:182–191

Jelenkovic G, Harrington E (1972) Morphology of the pachytene chromosomes in *Prunus persica*. Can J Genet Cytol 14:317–324

Joobeur T, Viruel MA, de Vicente MC, Jauregui B, Ballester J, Dettori MT, Verde I, Truco MJ, Messeguer R, Balle I, Quarta R, Dirlewanger E, Arús P (1998) Construction of a saturated linkage map for *Prunus* using an almond × peach F_2 progeny. Theor Appl Genet 97:1034–1041

Joobeur T, Periam N, de Vicente MC, King GJ, Arús P (2000) Development of a second generation linkage map for almond using RAPD and SSR markers. Genome 43:649–655

Jung S, Abbott A, Jesudurai C, Tomkins J, Main D (2004) Frequency, type, distribution and annotation of simple sequence repeats in Rosaceae ESTs. Funct Integr Genomics (in press)

Kantety RV, Rota ML, Matthewes DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol 48:501–510

Lu ZX, Sosinski B, Reighard G, Baird WV, Abbott AG (1998) Construction of a genetic linkage map and identification of AFLP markers for resistance to root-knot nematodes in peach rootstocks. Genome 41:199–207

Neuhaus G, Horn R (2004) Implications of single nucleotide polymorphisms for plant breeding. Prog Bot 65:55–71

Pearson JD, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448

Rajapakse S, Belthoff LE, He G, Estager AE, Scorza R, Verde I, Ballard RE, Baird WV, Callahan A, Monet R, Abbott AG (1995) Genetic linkage mapping in peach using morphological, RFLP and RAPD markers. Theor Appl Genet 91:964–971

Shimada T, Yamamoto T, Hayama H, Yamaguchi M, Hayashi T (2000) A genetic linkage map constructed by using an intraspecific cross between peach cultivars. J Jpn Soc Hortic Sci 69:536–542

Sosinski B, Gannavarapu M, Hager LD, Beck LE, King GJ, Ryder CD, Rajapakse S, Baird WV, Ballard RE, Abbott AG (2000) Characterization of microsatellite markers in peach [*Prunus persica* (L.) Batsch]. Theor Appl Genet 101:421–428

Terrier N, Ageorges A, Abbal P, Romieu C (2001) Generation of ESTs from grape berry at various developmental stages. J Plant Physiol 158:1575–1583

Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley. Theor Appl Genet 106:411–422

Viruel MA, Messeguer R, de Vicente MC, Garcia-Mas J, Puidomenech P, Vargas F, Arús P (1995) A linkage map with RFLP and isozyme markers for almond. Theor Appl Genet 91:964–971

Wang D, Karle R, Brettin TS, Iezzoni AF (1998) Genetic linkage map in sour cherry using RFLP markers. Theor Appl Genet 97:1217–1224

Wang Q, Zhang K, Qu X, Jia J, Shi J, Jin D, Wang B (2001) Construction and characterization of a bacterial artificial chromosome library of peach. Theor Appl Genet 103:1174–1179

Wang Y, Georgi LL, Zhebentyayeva TN, Reighard GL, Scorza R, Abbott AG (2002) High-throughput targeted SSR marker development in peach (*Prunus persica*). Genome 45:319–328

Warburton ML, Becerra-Velasquez VL, Goffreda JC, Bliss FA (1996) Utility of RAPD markers in identifying genetic linkages to genes of economic interest in peach. Theor Appl Genet 93:920–925

Wu J, Maehara T, Shimokawa T, Yamamoto S, Harada C, Takazaki Y, Ono N, Mukai Y, Koike K, Yazaki J, Fujii F, Shomura A, Ando T, Kono I, Waki K, Yamamoto K, Yano M, Matsumoto T, Sasaki T (2002) A comprehensive rice transcript map containing 6591 expressed sequence tag sites. Plant Cell 14:525–535