Gene duplication in soybean and polymorphism in the steroid biosynthesis pathway indicate highly conserved gene sequences

DeCaire, J., Brumett, S., Payne, J., Oakley, J., Ray, J., Smith, J., Atkins, D., Bailey, D., Bell, C., Bordelon, C., Bostick, T., Butler, E., Cefalu, J., Chatham, B., Colvin, L., Dejean, K., Dobbins, S., Faul, B., Flowers, D., Goode, J., Haburne, H., Hamilton, D., Hermes, A., Hightower, J., Keator, A., Kile, M., May, B., Mcgee, P., Parker, K., Pinton, K., Pyles, B., Richardson, M., Roberson, M., Roberts, R., Rodriguez, J., Roe, R., Scott, C., Shook, L., Stokes, K., Stowell, R., Taylor, C., Thompson, C., Vincent, M., Walker, T., Patel, S., Shultz, J.

Jeffery Ray and James Smith are from the Crop Genetics Production and Research Unit, USDA, Stoneville, MS 38756. All other authors are from the School of Biological Sciences, Louisiana Tech University, Ruston, LA 71272

Corresponding author: Jeffry L. Shultz, Telephone: 318-257-2753, jlshultz@latech.edu

Abstract

The steroid biosynthesis pathway is responsible for producing several important biochemicals and precursor molecules. The KEGG database renders an authoritative view of the interactions in this pathway. We used this and other online resources to design PCR primers based primarily on soybean expressed sequence tags (ESTs). A total of 50 gene-based primer pairs (38 soybeans, 12 alternatives) were tested for function and four polymorphism states on two soybean breeding lines. A total of nine genes exhibited a single, monomorphic amplification product and four amplified a single polymorphic fragment. The production of 13 multiple monomorphic products and ten multiple products that included at least one polymorphism was also observed. A total of 14 primer pairs failed to produce a clear amplification product. The sequences used to design these primers were tested against an initial scaffold build of soybean genomic sequences consisting of 950 Mbp of DNA. These data indicate that directed analysis of a biochemical pathway can yield multiple opportunities to identify specific genic polymorphisms in soybean, while taking into consideration the duplicated nature of the genome.

Introduction

Soybean (*Glycine max* (L.) Merr.) is the second most valuable crop in the U.S., accounting for up to $17 billion in annual revenue [1]. At an estimated 1.1 Bbp [2], the soybean genome is up to 10 times the size of model plants such as *Arabidopsis thaliana*, *Medicago truncatulata* or *Lotus japonicus.* Although diploid, the soybean genome shows evidence of a paleopolyploid origin with gene-rich islands that appear to have been conserved following duplication [3].

Ongoing sequencing of the soybean genome [4] provides a potent tool for genomic analysis of this crop, however, the ability to assign function to genomic regions using only sequence is hampered due to the repetitive nature of the genome. In other words, if multiple copies of a gene exist, how do these copies interact with each other to produce a phenotype? The most effective method of assigning an agronomically important phenotype to a genomic region continues to be the utilization of genetic maps and association of these maps to experimental results using segregating populations. Simple Mendelian traits when measured in a population offer confirmation of function to a region. When a trait appears to be quantitative, the use of a segregating population allows the designation of several putative regions as contributing to a particular phenotype using quantitative trait loci (QTL) analysis.

The marker types used to generate molecular maps in soybean include restriction fragment length polymorphism (RFLP), amplified fragment length polymorphism (AFLP), randomly amplified polymorphic DNA (RAPD), simple sequence repeat (SSR) and single nucleotide polymorphisms (SNP). Of these types, RAPDS, SSRs and SNPs continue to be reported. These marker technologies are PCR-based and can be employed in high-throughput applications.

The dominant nature and poor repeatability of RAPDs severely restrict this technology from marker assisted selection. SSRs are based on polymorphism generated when micro-satellite sequences mismatch during cell division, resulting in a reliable, sequence-length difference. Because the marker is based on repetitive sequence, the likelihood of it being located within a gene is low and multiple amplicons are common. SNPs suffer from the opposite problem, in that they are often derived from highly conserved genic sequences [5, 6]. SNP detection is usually a two step process, in which the genomic region is amplified by PCR, followed by detection of the SNP within this PCR product (Choi et al., 2007; Lee et al., 2004; Zhu et al., 2003). The duplicated nature of soybean complicates SNP detection during the pre-detection PCR step, due to homologous sequence amplification, which creates a heterogenous PCR product. This has forced the design of these pre-detection primers based on the less conserved 3’ un-translated region of a gene [5], increasing specificity, but reducing the detection of gene-coding sequence in the process.

At present, a total of 394,370 EST sequences are available for soybean [7]. Placing these genes on molecular marker maps may allow the identification of genes controlling a particular trait and provide a greater understanding of the mechanisms resulting in the expression of the trait. Such an understanding could lead to the development of improved cultivars, through the use of gene-specific molecular markers.

A method of primer design based on expressed sequence tags within a specific pathway has previously been described [8]. In brief, selected genes from within the purine metabolism pathway were used to design PCR primers, which were tested on mapping population parents for polymorphism. In this example, several different sources were used to derive the tested PCR

primer sequences. This procedure, although effective, was not amenable to high throughput primer design.

We had four objectives for this research. The development of an increased throughput procedure for pathway-based primer design and testing was our first priority. Our second objective was the identification of immediately map-able polymorphism(s). Third, we sought to identify the effects of duplication within soybean genomic DNA for the genes within the pathway of interest. Our final goal was the visual presentation of these data to allow intuitive analysis of the pathway.

Materials and Methods

Two soybean plant introductions were used in this survey (PI567743 and PI87623). PI567743 is a maturity group IV accession collected from Jiangsu, China in 1993. This accession has reported heterogeneity to phytopthora rot races 1 and 3, and susceptibility to Phytopthora races 7, 17 and 25 and Brown stem rot. PI87623 is a maturity group IV accession collected from Japan in 1930, with resistance to earworm and susceptibility to SDS, pythium rot, PMV, phytopthora and SCN races 3, 4, and 5 [9]. DNA from these parents was extracted from leaf tissue using a tissue pulverizer/liquid nitrogen procedure as outlined by Shultz et al. [8].

Primer Design

All soybean primers were designed using a two-step process. First, expressed sequence tag (EST) sequences attributed to specific gene function (via EC#) were identified using the KEGG database (green-filled enzymes, Figure 1). When soybean sequences were not annotated for a gene (white filled enzymes, Figure 1), primers were designed based on an alternate eukaryote whenever possible, followed by a prokaryote if necessary. Once a sequence was selected from KEGG, primers were created using the Primer 3 program [10]. The default settings were an 18-27 bp oligo length, T_m = 57-63^o C, and GC% of 20-80%. An excluded region was set for 50 bp from start of sequence to 50 bp from the end of that sequence.

PCR Conditions

Components in each reaction included 10 ul 2X PCR mix (cat # M7122, Promega, Madison, WI), 3 ul H_2O, 5 ul target DNA and 1 ul each of 10 mM forward and reverse primers. Polymerase chain reaction (PCR) amplification conditions consisted of an initial 94^o C denaturation for 5 min followed by 36 cycles of a denaturing step (94^o C for 60 s), an annealing step (50^o C for 75 s) and primer extension (72^o C for 45 s). After a final extension of 5 minutes at 72^o C, the product was loaded into a 1.5% agarose (cat # BP160, Fisher Scientific, Waltham, MA) prepared with ethidium bromide (0.1 ul:20ml gel matrix). Electrophoresed was performed @ 4v/cm for 3-4 hours. Amplicon banding patterns were documented with a UVP Gel documentation system (Upland, CA).

Sequence Comparisons

Comparisons between EST sequences and assembled genomic DNA were made using the BLAST sequence utility at http://www.phytozome.net/soybean.php (June, 2008). The default settings of BLOSUM62, 11 word length, expect greater than 0.01 and filter on were used. Only similarities with e-values greater than e-^{09} were used to assign ESTs to genomic DNA scaffolds. Sequence comparisons between ESTs were performed using the BLAST two sequences utility at NCBI [11].

Results and Discussion

In order to simplify the procedure and increase throughput of primer design, the single online resource at KEGG [12, 13] was used to develop a panel of PCR primers based on ESTs in the biosynthesis of steroids pathway. This pathway is fed by pyruvate and the glycolysis product D-Glyceraldehyde-3-phosphate (Figure 1). This pathway leads to terpenoids, ketone bodies, carotenoids, bile acid and brassinosteroids. Porphyrin, chlorophyll and steroid hormone metabolic activities are supplied with products from this pathway. A total of 50 primer pairs were designed, of which 37 soybean-annotated genes were utilized, with the remaining 13 based on non-soybean sequences. These sequences include 10 eukaryotes (four human, three rice, one arabidopsis and two *S. cerevisiae*) and 3 prokaryotes (*Y. pestis*, Synechococcus and *P. torridus*). The utilization of this single source instead of multiple sources along with a standardized procedure for EST selection allowed primer design in less than 10 minutes for each gene.

There were five possible outcomes for each tested primer pair. First, and most common among primers designed from alternative genomes, was no detected amplification product. Second was the amplification of multiple, monomorphic products Third was the amplification of multiple products including *clear* polymorphisms (bright, reproducible band). A fourth possibility was the amplification of a single, monomorphic product. The amplification of a single, polymorphic product was the fifth possibility. The various phenotypes are indicated in Table 1 and in Figure 1. If a very bright single amplification product was produced with minor secondary amplification product(s), it was considered a single amplification product (see Figure 1b, CYP710A).

The failure of 14 primer pairs is most likely due to the DNA sequence used for primer design. A total of nine of these failures were primers based on alternative genomes. Because each sequence utilized for primer design is a transcribed EST, the probability that each corresponding genomic sequence contains introns is very high. The attempted amplification of complete genomic DNA which includes intronic sequence may very well increase the PCR product size beyond that which could reasonably be amplified (>2000 bp, including EC2.5.1.21, 4.6.1.12, and 5.3.3.5). In addition, these failed reactions may also be represented by enough duplicated genes that PCR of these sequences would yield an essentially failing reaction (EC1.1.1.34).

The amplification of multiple, monomorphic amplicons is likely due to the duplicated nature of soybean. The total number of soybean scaffold matches was determined by BLASTing the full NCBI GI sequence of the EST used for primer design against the preliminary sequence assembly at http://www.phytozome.net/soybean.php. All matches above e-^{09} are indicated as scaffold sequence matches and are considered evidence of the number of copies of this sequence

in the soybean genome. The average number of sequences for all soybean-annotated sequences tested was 2.9, with a range of 0 – 17 copies (Table 1).

A useful product of PCR was the amplification of multiple amplicons, including at least one clear polymorphism. This amplification pattern was immediately useful for both identifying the gene as putatively duplicated and for potentially mapping the identified polymorphism.

A single monomorphic band indicates either a unique sequence or a highly conserved, duplicated region. The importance of a single-copy gene lies in the ease to which it could be transferred via breeding from one line to another (Mendelian inheritance). If, however, the single amplification product reflects multiple loci, an argument could be made that the conservation of these sequences may reflect the importance of the gene in the genome. The requirement for a single amplicon for SNP detection has previously been reported (Choi et al., 2007; Lee et al., 2004; Shultz et al., 2008). This pre-amplification step can be carried out using primers with this profile.

Finally, a single polymorphic amplification product has immediate utility for mapping the genomic location of the gene (after sequence verification) and for marker assisted selection. There were four of these patterns identified for this pathway.

We recognize that this survey of polymorphism is a highly specific snapshot of the biosynthesis of steroids pathway in soybean. A multitude of variables make this a “moving target”. These variables include the number of soybean ESTs that are annotated within the pathway, the number of soybean lines tested, the thoroughness of primer testing, the number of copies in the genome and human error.

With over 390,000 soybean ESTs available, it is not surprising that the average number of soybean ESTs for soybean-annotated genes was 28. There are two enzymes (EC 1.1.4.1 and 2.5.1.21) that have only one EST sequence and eight that have less than six sequences reported. A conservative assumption would be that additional soybean sequences would be annotated within this pathway if additional EST sequencing were to be performed. These additional sequences would most certainly be distributed over the more fully represented genes, but it is also probable that sequences with limited or no representation could also increase in representation within this pathway.

The current study used only two soybean lines. It is expected that the number of single monomorphic amplicons would decrease from 11 to none if all of the 18,000+ available soybean germplasm lines [9] were tested and the method of detection were changed from visual detection using agarose to a more accurate fluorescent-labeled capillary electrophoresis. In addition, all multiple, monomorphic amplification patterns would also be assumed to show polymorphism as additional lines were tested.

In order to guarantee homogeneity of comparisons, PCR troubleshooting was not allowed. The use of gradient and/or long-range PCR would have certainly reduced the number of failures and would have increased the number of high-quality PCR products observed during this study, but it would not allow strict comparison between the results obtained herein and future surveys.

The repeat nature of the soybean genome allows the designation of each of the polymorphic alleles as one of possibly many genes with similar sequence. As can be seen in figure 1b, the quality of PCR was affected by the number of copies of a gene. A single amplification product was typically very strong, but as the number of priming sites increased, it was clear that amplification was reduced for each individual loci. A contradictive result to this is

shown for CYP710A, which has 17 significant matches within soybean genomic DNA, yet has only one main amplification product.

Although given separate designations in the KEGG pathway representation, four pairs of enzymes (ERG6 and SMT1, sterol 24-C-methyltransferase; 1.3.1.72 and DWF1, delta24-sterol reductase; 5.3.3.5 and HYD1, cholestenol delta-isomerase; 1.3.1.21 and DWF5, 7-dehydrocholesterol reductase) are considered equivalent for this study. These enzymes have identical soybean EST and KEGG annotation, and are connected via pink dashed lines in Figure 1. The shared function between EC # 5.3.3.5 and HYD1 may indicate a mechanism of control for this pathway. Down-regulation of these identical enzymes could lead to an increase in ergosterol production. Further, down-regulation of ERG6/SMT1 sterol 24-C-methyltransferase activity would effectively lead to increases in bile acid synthesis and/or steroid hormone metabolism.

The location of EC #'s 1.1.1.34 and 2.2.1.7 in close proximity to each other on sequence scaffold 30 (6,549,564 and 6,964,329 bp respectively) and their positions as the first enzymes in this pathway indicate that a putative regulatory role for this pathway may lie in this genomic region. These sequences are not significantly similar to each other.

Mapping genes involved in the biosynthesis of steroid pathway may provide specific insights into steroid production in soybean and indicate ways to moderate the level of steroid production in soybean. However, this is a major undertaking with 390,000+ ESTs sequenced. The method that we present reduces the number of ESTs to test to the number of genes in the pathway, with specific rules allowing a comparison of results across pathways and germplasm.

Data generated from this project identifies functional primer sets for further use in molecular mapping of biochemical pathways. The identification of genomic regions with putative control over the pathway can lead to utilizing these regions for MAS and for identifying potentially additive quantitative trait loci (QTL). Finally, each of the single amplification products can be further tested for polymorphism within a larger set of mapping parents.

Conclusion

We have combined molecular pathway information with genetic diversity data within soybean varieties in order to create a resource for the study of enzymes specific to a biological process. As a first step, we have designed and tested a series of primers within the Biosynthesis of Steroids pathway. We have created a visual tool that shows gene duplication within a genome, potential MAS targets and duplication of enzymes within a pathway.

Author Contributions

J. DeCaire - J. Oakley order based on number of primers designed and tested; J. Ray and J. Smith provided editing of manuscript, equipment loan as well as DNA material; D. Atkins – T. Walker listed alphabetically, with equivalent research input; J. Shultz conceived of the study, led research and wrote the manuscript.

Acknowledgements

We would like to acknowledge the Louisiana Tech University College of Applied and Natural Sciences and the Student Technology Fee Board for their support.

References

1. NASS: **Crop Values 2006 Summary**. *http://wwwnassusdagov/* 2007.
2. Arumuganathan K, Earle ED: **Nuclear DNA content of some important plant species**. *Plant Mol Biol Rep ISPMB* 1991, **9**(3):208-218.
3. Shoemaker RC, Polzin K, Labate J, Specht J, Brummer EC, Olson T, Young N, Concibido V, Wilcox J, Tamulonis JP *et al*: **Genome duplication in soybean (Glycine subgenus soja)**. *Genetics* 1996, **144**(1):329-338.
4. DOE: **Joint Genome Institute** *http://wwwjgidoegov/* 2008.
5. Choi IY, Hyten DL, Matukumalli LK, Song Q, Chaky JM, Quigley CV, Chase K, Lark KG, Reiter RS, Yoon MS *et al*: **A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis**. *Genetics* 2007, **176**(1):685-696.
6. Shultz JL, Ray JD, Lightfoot DA: **A sequence based synteny map between soybean and Arabidopsis thaliana**. *BMC Genomics* 2007, **8**(1):8.
7. NCBI: **National Center for Biotechnology Information**. 2008.
8. Shultz JL, Ray JD, Smith JR: **Mapping two genes in the purine metabolism pathway of soybean**. *DNA Seq* 2008, **19**(3):264-269.
9. GRIN: **Germplasm resources information network**. *http://wwwars-gringov* 2008.
10. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers.** Totowa, NJ: Humana Press; 2000.
11. Tatusova TA, Madden TL: **BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences**. *FEMS Microbiol Lett* 1999, **174**(2):247-250.
12. Kanehisa M, Goto S, Kawashima S, Nakaya A: **The KEGG databases at GenomeNet**. *Nucleic Acids Res* 2002, **30**(1):42-46.
13. **Kegg Database; Glycine max ESTs identified in the Biosynthesis of Steroids Pathway** [http://www.genome.jp/dbget-bin/get_pathway?org_name=egma&mapno=00100]

Table 1. Primer sequences designed from KEGG annotated sequences from the Biosynthesis of steroids biochemical pathway Amplicon patterns for 50 primer pairs tested on two soybean mapping population parents.

Gene (EC#)	Source organism and GI #	Primer Sequences 5'-3' Forward (top) Reverse (bottom)	Enzyme Copies*	Predicted Amplicon size**	Amplicon type***	G. max records in KEGG	Scaffold Sequence Matches****
1.1.1.34	*G. max* 10843386	CAGATCCCCGTGGGAGTAG CATTCCCATCGACTATGACAGA	1	496	F	72	30, 58, 182, 15
1.1.1.170	*G. max* 10237318	GATACTGCAACTGGCAACATTT CATGGGTGCAAAGGTATGAA	1	378	MM	5	137, 112, 44
1.1.1.267	*G. max* 10235144	CCTTGTGCAGCCACCATTAT CCATCATGAATCCCAGCTCT	1	572	MP	165	35, 25, 16, 97
1.1.1.270	*H. sapiens* 7705421	GTTTTGATCACCGGGGCTA TGAGCCTGGCCTGATTATCT	1	997	F	0	none
1.1.4.1	*G. max* 7796718	TGCACTTACACTCTCATTCAATCAT AACCACCTTAGGATCCTCCAT	1	351	MM	1	54, 22
1.3.1.21 DWF5	*G. max* 15813693	TTCAACCCTACGATTGTTTATCA TGCAAAGAATGCCAGCTACT	3	538	F	26	113, 215
1.3.1.70	*G. max* 51337045	GGGAACATGACACTGTTACTTTTG ATACAGGGGTGGTGGCTCTT	1	765	SP	8	261, 200
1.3.1.72 DWF1	*G. max* 58017512	GGTCACTGGGATACACAATGG CTTCTGGAGGATGTTTGATGC	8	599	SM	38	53, 98
1.6.5.2	*H. sapiens* 70995356	GCACTGATCGTACTGGCTCA GGTTGTCAGTTGGGATGGAC	1	790	F	0	none
1.14.13.13	*H. sapiens* 74099700	CCTGGCAGAGCTTGAATTG TATCTGTCCAAAAACTGTAGGTTGA	1	150	F	3	none
1.14.13.70	*G. max* 21479923	AAATCACACCACAGATCTGCAC TCCAAAACTTGGCACATTGA	1	556	MM	34	73, 72
1.14.13.72	*G. max* 7926000	CCACTTCCCCACTCTCTCAC CTGAGCTGCAGGGGTGTTAT	1	249	MM	20	312, 104
1.14.99.7	*A. thaliana* 18406296	TCTGGCTTGGAGCCAAATTA CAGGTACAGTTGCTGGGAAAA	1	102	F	74	26, 4, 342, 154, 30, 74, 22, 70, 308
1.17.1.2	*G. max* 58020548	GATACTGCAACTGGCAACATTT CATGGGTGCAAAGGTATGAA	1	908	MM	17	112, 137, 99
1.17.4.3	*G. max* 19270772	GGAAAAATTTGTCTGATTTATCTGTG AGGGTTTTGCCTACAGCAAG	1	563	SP	5	none
1.142.1.6	*G. max* 7283699	GTCGAGGACACGGACTTGTA AAGGGGAGAGAGTGTTCTGTTT	1	493	MM	4	2, 7
2.2.1.7	*G. max* 6134201	CTCACTGTGGAGCATTTGACA CAAAGAAGTAACTGAAATATGGTTTGA	1	590	MM	74	25, 84, 30
2.5.1.1	*G. max* 19269369	CCCCCGTCTCCCTTTTAAT TTGAGTTAAGTATTTCACGCATTG	2	104	SM	6	76, 41, 124, 43
2.5.1.10	*G. max* 16284821	TCGTCCCACCTCTATGTTGA GTGCTAGAACGCGAAGGAGA	2	562	MP	70	53, 15, 13, 77, 111, 148, 82, 39
2.5.1.11	Synechococcus JA-3-3Ab 86606912	CCTGTAGAAGCGGATCTGGA CTCCAGCACATAGTCCACCA	1	187	F	0	none
2.5.1.21	*G. max* 16344125	CACCGGGAATTCCAACTTTA AATTGGAACCCTTGCCTTTT	2	622	F	1	53
2.5.1.29	*G. max* 13790752	GCTGCTCCTTTGATTGCTTT CTCCTGCCATTTTACTCAACTT	3	298	SM	2	148, 111
2.5.1.30	*Y. pestis* 145600809	GAGTTAACCGCGCCAGATA TGCACTGAGATATGTGCCAAG	1	938	F	0	none
2.5.1.33	*P. torridus* 48478271	CTGGCAGTACAGCATGAGGA CATCAAACCTGCCTCATCAA	1	967	MM	0	none
2.7.1.36	*G. max* 10847219	CATGAGGGCAAGAATTGGTT TTGAACCCAGAGCAAGTTCA	1	508	MM	4	41, 76, 116, 50

Gene (EC#)	Source organism and GI #	Primer Sequences 5'-3' Forward (top) Reverse (bottom)	Enzyme Copies*	Predicted Amplicon size**	Amplicon type***	G. max records in KEGG	Scaffold Sequence Matches****
2.7.1.148	*G. max* 7926403	CGAGGAAGAAAGAACAAAAAGC AGCGCCTTAATAATCAAATTTCT	1	590	MP	17	177, 116, 31
2.7.4.2	*G. max* 19935159	GGCGAGGTTGTTAGTTCAGC AGCAAAGACTGCATCAAATCC	1	384	MP	18	65, 214
2.7.7.60	*O. sativa* 115441507	GACCAGAGCGCACACATCTA AGCCAGGAGCAAGTCATCAG	1	756	MM	0	none
4.1.1.33	*G. max* 33388894	ACGGGCTCATGTAAACGAA GAGTAGCATGGGCTTCCTTG	1	627	MP	8	1, 52
4.6.1.12	*O. sativa* 115447923	CGCCTTACCGCAAATACTTC TGAGCAGCTATGCTCCTGTTT	1	784	F	0	none
5.3.3.2	*G. max* 23056874	GTCCTCCTCGATCGTCGTTA TGGTACATCTTCGGCAACAA	1	553	MM	54	75, 21, 24, 9, 15, 150
5.3.3.5 HYD1	*G. max* 16346028	GGCACTCTCTATTCCCATTCA ATAGCTGGCCCAAAGAAATG	3	490	F	16	30, 100
5.4.99.7	*H. sapiens* 47933397	AGGAGAGAGGAGTCCGGTGT CAGGGTACAGCTGGGAGAAG	1	165	MP	11	none
5.4.99.8	*O. sativa* 115444137	ATAATGGAAGCCAGCTGTGG GGGCCATGCACCTTTAGATA	1	185	F	54	174
5.5.1.9	*G. max* 33390606	GTTCCATTGCTGCTCGTTG AGACTGGCCTTGACATTATTATCC	1	466	MP	13	47, 157
CYP51G1	*G. max* 21479923	AAATCACACCACAGATCTGCAC TCCAAAACTTGGCACATTGA	1	556	MP	34	73, 72
CYP710A	*G. max* 10844665	GACTTTGGCCCTCACCTTC CCCCATTTACCTAGATACAAATTCC	1	495	SP	18	47, 70, 71, 55, 48, 17, 9, 19, 21, 59, 75, 30, 319, 32, 199, 38, 18
ERG2	*S. cere.* 6323858	TCCCACTCCTTTTGTTGATTG TACCCATGTCCCTGGCAGT	1	630	F	0	none
ERG3	*G. max* 51336732	AAGAACGAGTGTGGACCAAT TCACCACAAGTGGATCATGC	1	700	F	2	none
ERG4	*G. max* 14206104	CCCTCAGTTATCCGTGGAAAG ATAGGAACATTTTGATAAAATCTGTGG	1	100	SM	4	111, 307
ERG5	*S. cere.* 6323657	CCCCTAACTATACCGCACCA GTGGACCACAACCAAAAACC	1	174	MM	0	none
ERG6 SMT1	*G. max* 15813168	CAACTCCTCCTTCAGGATCG ATAGCCATTTGCAGGGTCAC	2	225	MP	111	210, 253, 213, 159, 52, 127
FK	*G. max* 51337045	GGGAACATGACACTGTTACTTTTG ATACAGGGGTGGTGGCTCTT	1	732	SP	8	261, 200
GGCX	*G. max* 7686004	CGACCCTCAGTCATCATCCT ATTCGGGCAATTTTGTTGAT	1	378	SM	176	104
GTMT	*G. max* 7283465	GGATCCCAACAATCTCATGC GCATTTGCTCTTTGAGCTTG	2	490	SM	45	98, 138
HPT	*G. max* 5677140	GCACGAGGATAATTGGAGGT TTTTTGGATTTTCCTATTGTGC	1	470	SM	67	84, 25
SMT2	*G. max* 16347195	TCAACATGGGTTGCAGTGTT TCGTCATCCTCGTACTCGTG	1	194	SM	91	121, 151
STE1	*G. max* 7283699	GTCGAGGACACGGACTTGTA AAGGGGAGAGAGTGTTCTGTTT	2	493	MM	4	2, 7
VTE1	*G. max* 11412155	AAGCCAAGCTCTGGGAGTCT CCAGAAGAATTGTGATTCAGGA	2	458	SM	10	26, 69

* Number of copies of a particular enzyme listed in the pathway, including alternatively named enzymes

** Amplicon size predicted by the Primer3 program based on EST sequence.

***Amplicon types reported are F = failure; MM = multiple amplicons that are monomorphic; MP = multiple amplicons that contain a polymorphism; SP = single amplicon that is polymorphic; SM = single amplicon that is monomorphic

**** Sequence scaffold matches indicate significant match at e^{-10} or lower, with results in descending order of significance

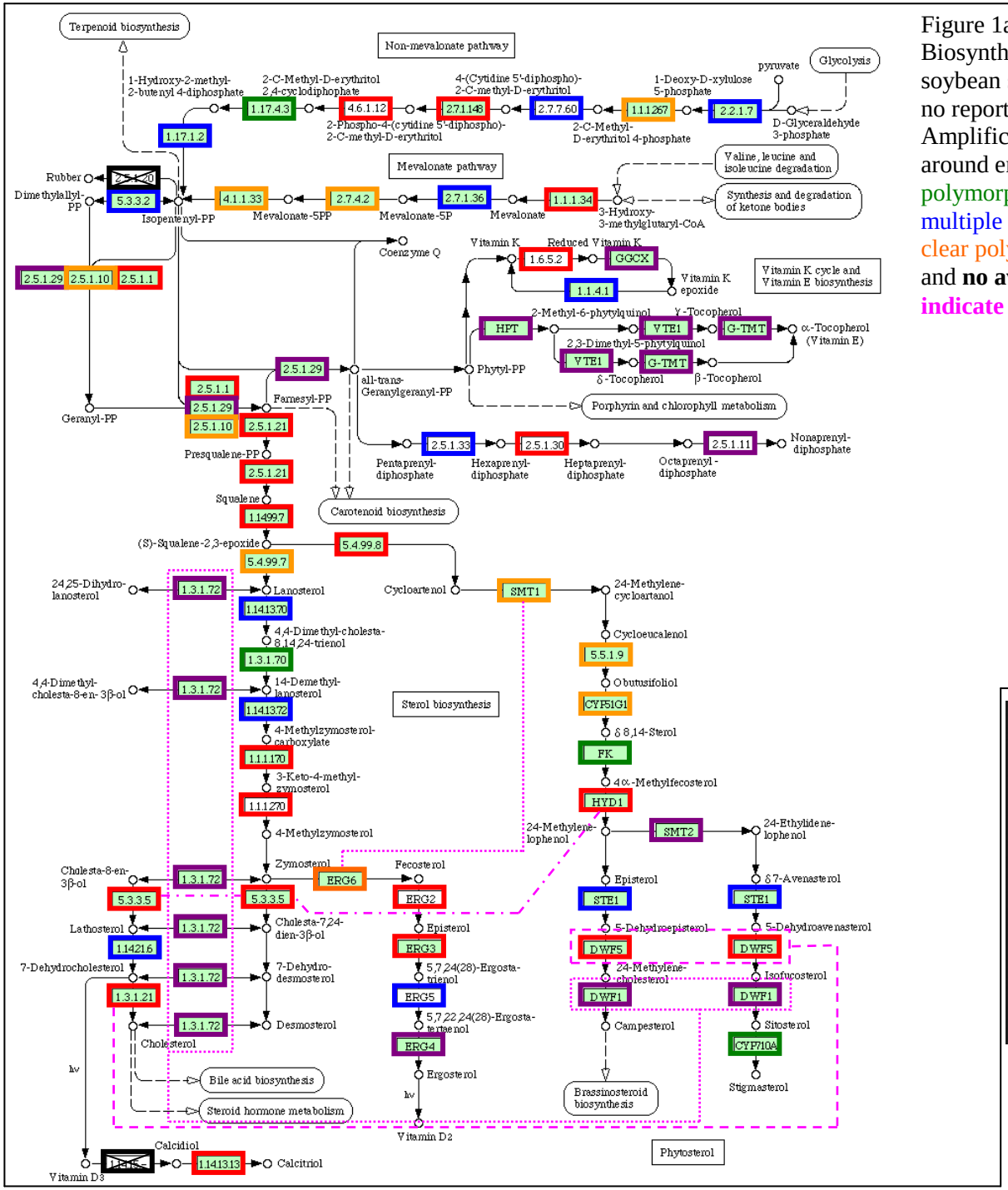
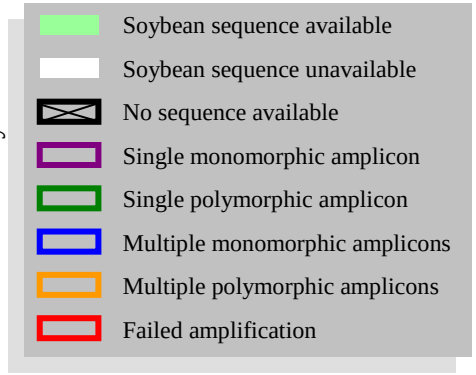


Figure 1a. KEGG pathway (April 2008) for the Biosynthesis of Steroids. Enzymes with representative soybean sequences are green-filled, while enzymes with no reported soybean sequences are un-colored. Amplification products are indicated by box color around enzyme name and are as follows: Single polymorphic (green), single monomorphic (purple), multiple monomorphic (blue), multiple, with at least one clear polymorphism (orange), failed amplification (red) and **no available sequence (black)**. **Dashed pink lines indicate equivalent enzymes**



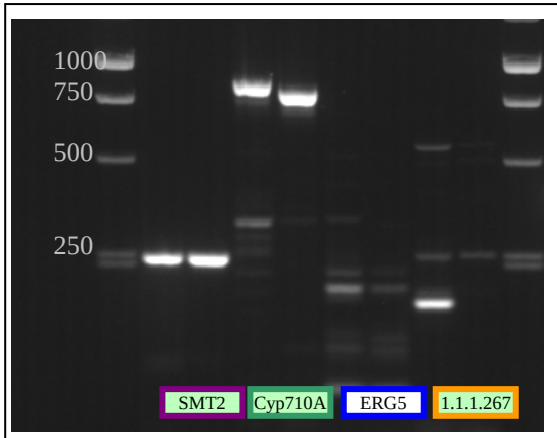


Figure 1b. Examples of amplicon profiles used to determine the duplicated nature of soybean genes when amplified using genomic DNA.