GENETIC STUDIES OF CANDIDATE GENES IN THE GLYCOALKALOID BIOSYNTHETIC PATHWAY OF POTATO

Norma Constanza Manrique-Carpintero

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Horticulture

Richard Veilleux - Chairman

James G. Tokuhisa

Bingyu Zhao

M. A. Saghai Maroof

December 11th 2012
Blacksburg, VA

Keywords: Solanaceae, wild potato species, SNPs, allelic mining

GENETIC STUDIES OF CANDIDATE GENES IN THE GLYCOALKALOID BIOSYNTHETIC PATHWAY OF POTATO

Norma Constanza Manrique-Carpintero

Abstract

Potato (*Solanum tuberosum* L) is an outcrossing, highly heterozygous cultivated in which the elucidation of the genetic basis of quantitative traits, is more complex than in self-pollinated crops. Both a candidate gene approach and a whole genome SNP genotyping analysis were used to assess allelic variation and to identify loci associated with biosynthesis and accumulation of steroidal glycoalkaloids (SGAs). SGAs are secondary metabolites produced in *Solanum* species as defense against insects and pathogens. Fragments of genomic DNA coding for regions of five SGA biosynthetic candidate genes were amplified, cloned and sequenced [3-hydroxy-3-methylglutaryl coenzyme A reductase 1 and 2 (*HMG1, HMG2);* 2,3-squalene epoxidase (*SQE*); solanidine galactosyltransferase (*SGT1*); and solanidine glucosyltransferase (*SGT2*)]. A germplasm panel of six wild potato species [*Solanum chacoense* (*chc* 80-1), *S. commersonii* subsp. *commersonii*, *S. demissum*, *S. sparsipilum*, *S. spegazzinii*, and *S. stoloniferum*] and a cultivated clone *S. tuberosum* Group Phureja (*phu* DH) was used in an allelic variation analysis. A segregating interspecific F_2 population *phu* DH × *chc* 80-1 was screened to assess association with SGAs. Sequence diversity analysis showed a tendency of purifying selection and increased frequency of rare alleles in most of the candidate genes. Genes of primary metabolism (*HMG1*, *HMG2* and *SQE*) had stronger selection constraints than those in secondary metabolism (*SGT1* and *SGT2*). Sequence polymorphism in *HMG2*, *SQE*, *SGT1* and *SGT2* separated either the *phu* DH clone which produced no SGAs, or *chc* 80-1, the greatest SGA accumulator, from other accessions in the panel. Segregation analysis of the F_2 population revealed that allelic sequences of *HMG2* and *SGT2* derived from *chc* 80-1 were significantly associated with the greatest SGA accumulation. In the whole genome analysis, SNP genotyping and cluster analysis based on putative association with SGA accumulation in the germplasm panel, allowed identification of eight informative SNPs that can be used in future studies. In the segregating F_2 population, loci located on five pseudochromosomes were associated with

SGA synthesis. Loci on pseudochromosomes 1 and 6 explained segregation ratios of synthesis for α-solanine and α-chaconine, the most common SGAs in most potato species. In addition, loci on seven pseudochromosomes were associated with accumulation. New candidate genes, putatively affecting synthesis and accumulation of SGAs, were identified in adjacent genomic regions of significant SNPs. This research demonstrates how the newly available genome sequence of potato and associated biotechnological tools accelerates the identification of genetic factors underling complex traits in a species with a difficult breeding structure.

Dedication

I dedicate my effort and perseverance to attain this goal in my professional and personal life to my family. I love my family and hope this is the begging of a better future. I feel pain for the time that I could have shared and enjoyed with my family. But I have improved my education, learned about other cultures, and joined people that taught me different points of view about the life and the world. Today I love and value my family more than never. I do not hesitate that it is necessary to fly and try hard to be someone and accomplish dreams. I feel pleased that I am one of many Colombians who fight every day to be better, that respect laws and human diversity rights. I want to be the inspiration for my little brother. I love my country Colombia and in spite of the problems, I know that if we just try to be better, help each other, improve our education and avoid complain about ourselves, we will have progress and piece in that tropical paradise. We have to be together because giants like the International Court of Justice, governments from other countries and people are trying to keep us in the poverty, injustice and division faking reasons of equality.

Acknowledgements

I want to thank my advisor, Dr. Richard Veilleux who gave me the opportunity to work in his distinguished research team. I admire and value Dr. Veilleux because is a smart and brilliant scientist and also a great human and generous person.

I thank my professors and committee members; I learned different s aspects of plant science and to be a dedicated and persistent scientist from you. Again I have to give a special acknowledgment to Dr. Veilleux my advisor here.

I found in Dr. Veilleux laboratory and other laboratories at Latham Hall people that taught me a learned from me. Team work is the basic element in any research project. I really appreciate that we could discuss about plant science, learn techniques, how to operate some software and share equipment. Special thanks to Suzanne Piovano, Juan Ruiz-Rojas, Alice Mweetwa, Tatiana Boluarte, Luciana Rosso, Kendall Teran Uphan, Kerri Mills, Janeth Donahue, Guillaume Pilot, Réjane Pratelli, Alexandra Jamie Weisberg**,** Maichel Miguel Aguayo Bustos, Dr. Holliday and people in Latham Hall that were part of this learning process.

Thank you to the faculty and staff from the Horticulture Department.

Friends are very important to me, thank you so much for your support, advice, plans for the free time, happiness, a numberless of things that we share. Some of them were crucial for my life in this town, Juan Jairo Ruiz-Rojas, Adelina Miranda, Raul Saucedo, Diego Cortes and Mihaela Babuceano, thank you so much. Dr. Holzman and Ms Ronna, thank you very much for let me live in your house and for being so nice with me.

Thanks to the BARD project, Horticulture Department and Language and Culture Institute for the financial support to do my Ph. D. and English Language studies.

Table of Contents

Abstract ii

Dedication iv

Acknowledgements v

1. Introduction 1

References 4

2. Allelic Variation in Genes Contributing to Glycoalkaloid Biosynthesis in a Diploid Interspecific Population of Potato 8

Authors contribution 8
Abstract 9
Introduction 11
Materials and methods 16
Results 23
Discussion 32
Acknowledgements 38
References 39
Tables and figures 45

3. Sequence Diversity in Coding Regions of Candidate Genes in the Glycoalkaloid Biosynthetic Pathway of Wild Potato Species. 68

Authors contribution 68
Abstract 69
Introduction 71
Materials and methods 75
Results 81
Discussion 87
Acknowledgements 92
References 93
Tables and figures 100

4. Conclusions 114

List of Tables

2. Allelic Variation in Genes Contributing to Glycoalkaloid Biosynthesis and a Diploid Interspecific Population of Potato……………………………………..8

Table 1 GenBank accessions used for primer design, and primer sequences used to amplify fragments of four candidate genes in the potato glycoalkaloid biosynthetic pathway 46

Table 2 Allele-specific primers for SNP, intron and indel genotyping of four candidate genes in a diploid F_2 population.. 47

Table 3 Summary statistics of analyzed allelic sequence sites (excluding sites with gaps on the total alignment), and sequence polymorphism found in four candidate genes of the glycoalkaloid biosynthetic pathway... 48

Table 4 Mean, range, standard deviation (SD), quartiles, and covariance (CV) of steroidal glycoalkaloid estimations in a diploid F_2 potato population *phu* DH × *chc* 80-1 49

Table 5 Chi-square (χ^2) test of phenotypic segregation ratios for synthesis of response variables in a diploid F_2 population *phu* DH × *chc* 80-1 50

Table 6 Chi-square (χ^2) test of genotypic segregation ratios of four candidate genes in a F_2 population *phu* DH × *chc* 80-1 and homogeneity test.. 51

Table 7 Mean SGAs for groups of individuals within an F_2 population derived from a cross between *phu* DH × *chc* 80-1 having the same genetic structure 52

Table 8 Effect of *HMG2* and *SGT2* genetic structure interaction on the accumulation of total SGAs, α-solanine and α-chaconine in a segregating F_2 population.......................... 53

Table 9 Effect of family pedigree on the accumulation of total SGAs, α-solanine, α-chaconine, leptine I and leptine II in a segregating F_2 population................................... 54

Table 10 Number of significant SNPs identified at different pseudochromosomes in the potato genome associated with α-solanine, α-chaconine, leptine I and leptine II accumulation in a segregating F_2 population.. 55

Table 11 SNPs from candidate genes in the 8303 Infinium array associated with leptine accumulation 56

3. Sequence Diversity in Coding Regions of Candidate Genes in the Glycoalkaloid Biosynthetic Pathway of Wild Potato Species………………………68

Table 1 Selected accessions with contrasting levels of SGA...................................... 101

Table 2 GenBank accessions used for primer design, and primer sequences used to amplify fragments of five candidate genes 102

Table 3 Number of allelic sequences identified in six wild and one cultivated potato species for five candidate genes within the glycoalkaloid biosynthetic pathway........... 103

Table 4 Summary of polymorphic sites and SNPs discovered in five candidate genes at the glycoalkaloid biosynthetic pathway.. 104

Table 5 Estimates of codon and nucleotide diversity of sequences of segments in coding regions for five candidate genes in wild potato species... 105

Table 6 Total SGA accumulation in leaf tissue of twelve accessions determined by HPLC 106

Table 7 SNPs found within exons of candidate genes putatively associated with SGA accumulation in leaf tissue of potato species.. 107

Table 8 Significant SNPs associated with SGA accumulation in a whole genome SNP chip analysis of wild and cultivated potato species 109

List of Figures

2. Allelic Variation in Genes Contributing to Glycoalkaloid Biosynthesis and a Diploid Interspecific Population of Potato……………………………………………..8

Figure 1 Potato steroidal glycoalkaloid biosynthetic pathway. Dashed line arrows indicate step with multiple enzymatic reactions 57

Figure 2 Pedigree scheme of *phu* DH clone and segregating F_2 population of *phu* DH × *chc* 80-1. 58

Figure 3 Physical position of significant superscaffolds identified on pseudochromosomes of *Solanum phureja* DM 1-3 516 R44... 60

3. Sequence Diversity in Coding Regions of Candidate Genes in the Glycoalkaloid Biosynthetic Pathway of Wild Potato Species……………………….68

Figure 1 Potato steroidal glycoalkaloid biosynthetic pathway. Dashed line arrows indicate step with multiple enzymatic reactions. 111

Figure 2 Haplotypes of informative SNPs identified in a whole genome analysis of wild potato species with different levels of SGA accumulation.................................... 112

Figure 3 Comparison of hierarchical clusters based on 3841 polymorphic SNPs (A) and eight SGA informative SNPs (B) from whole genome analysis of germoplam panel. 113

1. Introduction

Potato, a world food security crop, was initially domesticated by indigenous Andean populations in South America about 8000 years ago. The crop was later adapted and cultivated by European farmers and from there adopted across the globe (FAO 2008b). Potato is rich in carbohydrates and mineral nutrients, it is a low fat food mainly consumed fresh and is also industrially used to produce processed food, starch and alcohol (Lutaladio and Castaldi 2009). Since 2005 the world potato production has shifted from developed to developing countries positioning China as the largest producer (FAO 2008a). Intensive cultivation increases pest and disease problems that limit the production and agricultural development of this crop. For this reason, constant technical advances in the field and development of improved varieties are required. The most important agronomic characters for insect and disease resistance as well as tuber quality and yield correspond to quantitative traits (Gebhardt et al. 2011).

Underlying genetic factors involved in complex phenotypic traits have been addressed by different population genetic studies along the breeding history of crop science. First linkage analysis has been used to detect loci associated with a trait of interest in segregating populations, followed by positional cloning to characterize functional genes. Secondly, modern tools including association mapping, based on linkage disequilibrium analysis of unstructured populations, have provided an alternative approach (Zhu et al. 2008). In potato, an outcrossing and polyploid species, breeding schemes are based on intercrosses of multiple heterozygous parents followed by a series of recurrent selections, while mapping populations consist of diploid interspecific crosses or tetraploid backcross populations (Gebhardt et al. 2007). These facts make difficult to transfer mapped quantitative trait loci (QTL) markers to a breeding population because of the diverse genetic origin of the markers and the lack of mapping resolution of populations used in potato linkage analyses (Kloosterman et al. 2010). Therefore, the identification of genes responsible for the trait variation is crucial to understand and use the available natural variation. The candidate gene approach offers an alternative to disclosing genes associated with QTL traits beside positional cloning and insertional mutagenesis (Pflieger et al. 2001). This approach involves co-localization of genes,

potentially involved in cellular function mediating phenotype, with QTLs already associated to the trait (Gebhardt et al. 2007). The candidate genes are known functional genes that could be assumed that control a trait based on evaluation of linkage proximity of genes to fine mapped regions, comparative genomics, knowledge of related metabolic pathways, and differential expression analysis (Pflieger et al. 2001).

In this study we have used the candidate gene approach to elucidate genetic factors associated with biosynthesis and accumulation of steroidal glycoalkaloids (SGAs) in potato. Since the discovery of these compounds as deterrents of Colorado potato beetle *Leptinotarsa decemlineata* Say in wild potato species (Pierzchatski and Werner 1958; Prokoshev et al. 1952a), and the first attempts to incorporate them in cultivated potato (McCollum and Sinden 1979; Prokoshev et al. 1952b; Ross 1966; Sanford and Sinden 1972); it was suggested that a few major genes and polygenic inheritance could be acting on the synthesis and accumulation of SGAs. Characterization of insecticidal and pathogenic effects of SGAs as well as environmental factors influencing their accumulation enriched knowledge about phenotype of this trait (Maga 1994; Sinden et al. 1984; Tingey 1984). Then the studies were concentrated in the identification of multiple QTL based on different mapping populations developed from crosses of wild species or by introgression to cultivated potato (Hutvágner et al. 2001; Ronning et al. 1999; Sagredo et al. 2006; Sagredo et al. 2011; Sørensen et al. 2008; Yencho et al. 1998). In general, loci on chromosomes 1, 2, 4, 8 and 11 were mainly associated with synthesis and accumulation of SGAs with solanidine aglycone. The complexity of this trait also has shown that the interaction of interspecific genetic background, from crosses and somatic hybrids, generated progenies that synthesized SGAs no present in the parental lines (Carputo et al. 2010; Laurila et al. 1996; Prokoshev et al. 1952b; Väänänen et al. 2005). Even though many genetic studies have been done, neither the identification of functional genes nor the incorporation of SGA insect resistance to the cultivated potato has been successful. α-Solanine and α-chaconine are the most common SGAs in wild and cultivated potato (Deahl et al. 1993). In addition other SGAs (solasonine, solamargine, tomatine, leptine I, leptine II, commersonine and demissine) are produced in various *Solanum* species and have been associated with pest toxicity (Güntner et al. 2000; Kuhn and Löw 1961; Sanford et al. 1996; Sinden et al. 1986). The deleterious effects of SGAs

are mediated not only by the aglycone structure, but also by the fundamental influence of the glycoside moiety in the compound activity and final effect (Güntner et al. 2000). Chacotriose and lycotetraose are the most deterrent and toxic oligosaccharide residues in SGAs. Furthermore, some particular SGAs have been associated with specific pest resistances, e.g. leptines from *S. chacoense* with Colorado potato beetle *Leptinotarsa decemlineata* Say (Sinden et al. 1986), tomatine with potato leafhopper *Empoasca fabae* Harris (Sanford et al. 1996), and chaconine, solamargine, demissine and tomatine with potato aphid *Macrosiphum euphorbiae* (Güntner et al. 2000). The synthesis of SGAs occurs in the sterol branch of the mevalonic acid pathway, where some enzymes at primary and secondary metabolism have been associated with synthesis and accumulation of either sterols or SGAs (Arnqvist 2007; Choi et al. 1992; Krits et al. 2007; McCue et al. 2007a; McCue et al. 2007b; McCue et al. 2005; Moehs et al. 1997; Nahar 2011; Wentzinger et al. 2002). While genes of primary metabolism could regulate synthesis of metabolites at different branches within the metabolic pathway, the genes of secondary metabolism may modulate structural diversity of the final products (Chaturvedi et al. 2011; Ginzberg et al. 2012). The candidate gene approach provides an alternative to elucidate genetic factors associated with the synthesis and accumulation of SGAs using genes at the mevalonic acid pathway as a source for candidates.

Five candidate genes coding for 3-hydroxy-3-methylglutaryl coenzyme A reductase (*HMGR1* and *HMG2*), squalene epoxidase (*SQE*), solanidine galactosyltransferase (*SGT1*), and solanidine glucosyltransferase (*SGT2*) were selected to identify allelic variation in six wild (*Solanum chacoense* (*chc* 80-1)*, S. commersonii subsp. commersonii, S. demissum, S. sparsipilum, S. spegazzinii, S. stoloniferum*) and one cultivated potato *S. tuberosum* Group Phureja (*phu* DH) species with contrasting levels of SGAs. Then the association of allelic sequences with SGA synthesis and accumulation was evaluated in a segregating interspecific population between *phu* DH × *chc* 80-1. Sequence diversity and cluster analysis based on SGA accumulation level were also assessed for wild potato species. Finally whole genome analysis using the Potato SolCAP SNP array (Felcher et al. 2012) was done to identify loci associated with SGA accumulation and new candidate genes.

References

Arnqvist L (2007) Plant sterol metabolism with emphasis on glycoalkaloid biosynthesis in potato. Dept of Plant Biology and Forest Genetics. Swedish University of Agricultural Sciences, Uppsala, p 50

Carputo D, Savarese S, Andolfi A, Aversano R, Cimmino A, Frusciante L, Evidente A (2010) Glycoalkaloid profile in potato haploids derived from *Solanum tuberosum-S. bulbocastanum* somatic hybrids. Chem Biodivers 7:1885-1892

Chaturvedi P, Misra P, Tuli R (2011) Sterol glycosyltransferases-the enzymes that modify sterols. Appl Biochem Biotechnol 165:47-68

Choi D, Ward BL, Bostock RM (1992) Differential induction and suppression of potato 3-hydroxy-3-methylglutaryl coenzyme A reductase genes in response to *Phytophthora infestans* and to its elicitor arachidonic acid. Plant Cell 4:1333-1344

Deahl K, Sinden S, Young R (1993) Evaluation of wild tuber-bearing *Solanum* accessions for foliar glycoalkaloid level and composition. Am J Potato Res 70:61-69

FAO (2008a) The global potato economy. International year of the potato 2008. http://www.potato2008.org/en/potato/economy.html

FAO (2008b) Why potato? International year of the potato 2008. http://www.potato2008.org/en/aboutiyp/index.html

Felcher KJ, Coombs JJ, Massa AN, Hansey CN, Hamilton JP, Veilleux RE, Buell CR, Douches DS (2012) Integration of two diploid potato linkage maps with the potato genome sequence. PLoS One 7:e36347

Gebhardt C, Li L, Pajerowska-Mukthar K, Achenbach U, Sattarzadeh A, Bormann C, Ilarionova E, Ballvora A (2007) Candidate gene approach to identify genes underlying quantitative traits and develop diagnostic markers in potato. Crop Sci 47:S-106-111

Gebhardt C, Urbany C, Li L, Stich B, Paulo J, Draffehn A, Ballvora A (2011) Molecular diagnostics for complex pest and disease resistance and tuber quality traits: concept, achievements and perspectives. Potato Res 54:313-318

Ginzberg I, Thippeswamy M, Fogelman E, Demirel U, Mweetwa AM, Tokuhisa J, Veilleux RE (2012) Induction of potato steroidal glycoalkaloid biosynthetic pathway by overexpression of cDNA encoding primary metabolism HMG-CoA reductase and squalene synthase. Planta 235:1341-1353

Güntner C, Vázquez Á, González G, Usubillaga A, Ferreira F, Moyna P (2000) Effect of *Solanum* glycoalkaloids on potato aphid, *Macrosiphum euphorbiae*: Part II. J Chem Ecol 26:1113-1121

Hutvágner G, Bánfalvi Z, Milánkovics I, Silhavy D, Polgár Z, Horváth S, Wolters P, Nap JP (2001) Molecular markers associated with leptinine production are located on chromosome 1 in *Solanum chacoense*. Theor Appl Genet 102:1065-1071

Kloosterman B, Oortwijn M, Uitdewilligen J, America T, de Vos R, Visser RGF, Bachem CWB (2010) From QTL to candidate gene: Genetical genomics of simple and complex traits in potato using a pooling strategy. BMC Genomics 11(158)

Krits P, Fogelman E, Ginzberg I (2007) Potato steroidal glycoalkaloid levels and the expression of key isoprenoid metabolic genes. Planta 227:143-150

Kuhn R, Löw I (1961) Zur Konstitution der leptine. Chem Ber 94:1088-1095

Laurila J, Laakso I, Valkonen JPT, Hiltunen R, Pehu E (1996) Formation of parental-type and novel glycoalkaloids in somatic hybrids between *Solanum brevidens* and *S. tuberosum*. Plant Sci 118:145-155

Lutaladio N, Castaldi L (2009) Potato: The hidden treasure. J Food Compos Anal 22:491-493

Maga JA (1994) Glycoalkaloids in Solanaceae. Food Rev Int 10:385-418

McCollum G, Sinden S (1979) Inheritance study of tuber glycoalkaloids in a wild potato, *Solanum chacoense* Bitter. Am J Potato Res 56:95-113

McCue KF, Allen PV, Shepherd LV, Blake A, Maccree MM, Rockhold DR, Novy RG, Stewart D, Davies HV, Belknap WR (2007a) Potato glycosterol rhamnosyltransferase, the terminal step in triose side-chain biosynthesis. Phytochemistry 68:327-334

McCue KF, Allen PV, Shepherd LVT, Blake A, Rockhold DR, Novy RG, Stewart D, Davies HV, Belknap WR (2007b) Manipulation and compensation of steroidal glycoalkaloid biosynthesis in potatoes. Acta Hort 745:343-349

McCue KF, Shepherd LVT, Allen PV, Maccree MM, Rockhold DR, Corsini DL, Davies HV, Belknap WR (2005) Metabolic compensation of steroidal glycoalkaloid biosynthesis in transgenic potato tubers: using reverse genetics to confirm the in vivo enzyme function of a steroidal alkaloid galactosyltransferase. Plant Sci 168:267-273

Moehs CP, Allen PV, Friedman M, Belknap WR (1997) Cloning and expression of solanidine UDP-glucose glucosyltransferase from potato. Plant J 11:227-236

Nahar N (2011) Regulation of sterol and glycoalkaloid biosynthesis in potato (*Solanum tuberosum* L.) – Identification of key genes and enzymatic steps. Department of Plant Biology and Forest Genetics. Swedish University of Agricultural Sciences, Uppsala, p 66

Pflieger S, Lefebvre V, Causse M (2001) The candidate gene approach in plant genetics: a review. Mol Breeding 7:275-291

Pierzchatski T, Werner E (1958) Changes in glycoalkaloid contents in leaves of cultivated and wild potatoes and their hybrids during growth and their influence on the Colorado potato beetle (*Leptinotarsa decemlineata* Say) Part 1 (in Polish). Hodowla Roslin, Aklimat i Nasienn 2:157-180

Prokoshev SM, Petrochenko EI, Baranova VZ (1952a) Glycoalkaloids of tuber-bearing forms of *Solanum* in connection with their resistance to Colorado potato beetle (in Russian). Dokl Akad Nguk USSR 82:955-958

Prokoshev SM, Petrochenko EI, Baranova VZ (1952b) Inheritance of glycoalkaloids by interspecific hybrids of the potato (in Russian). Dokl Akad Nguk USSR 83:457-460

Ronning CM, Stommel JR, Kowalski SP, Sanford LL, Kobayashi RS, Pineada O (1999) Identification of molecular markers associated with leptine production in a population of *Solanum chacoense* Bitter. Theor Appl Genet 98:39-46

Ross H (1966) The use of wild *Solanum* species in German potato breeding of the past and today. Am Potato J 43:63-80

Sagredo B, Lafta A, Casper H, Lorenzen J (2006) Mapping of genes associated with leptine content of tetraploid potato. Theor Appl Genet 114:131-142

Sagredo B, Lorenzen J, Casper H, Lafta A (2011) Linkage analysis of a rare alkaloid present in a tetraploid potato with *Solanum chacoense* background. Theor Appl Genet 122:471-478

Sanford L, Domek J, Cantelo W, Kobayashi R, Sinden S (1996) Mortality of potato leafhopper adults on synthetic diets containing seven glycoalkaloids synthesized in the foliage of various *Solanum* species. Am J Potato Res 73:79-88

Sanford L, Sinden S (1972) Inheritance of potato glycoalkaloids. Am J Potato Res 49:209-217

Sinden S, Sanford L, Webb R (1984) Genetic and environmental control of potato glycoalkaloids. Am J Potato Res 61:141-156

Sinden SL, Sanford LL, Cantelo WW, Deahl KL (1986) Leptine glycoalkaloids and resistance to the Colorado potato beetle (Coleoptera: Chrysomelidae) in *Solanum chacoense*. Environ Entomol 15:1057-1062

Sørensen K, Kirk H, Olsson K, Labouriau R, Christiansen J (2008) A major QTL and an SSR marker associated with glycoalkaloid content in potato tubers from *Solanum tuberosum* × *S. sparsipilum* located on chromosome I. Theor Appl Genet 117:1-9

Tingey WM (1984) Glycoalkaloids as pest resistance factors. Am J Potato Res 61:157-167

Väänänen T, Ikonen T, Rokka VM, Kuronen P, Serimaa R, Ollilainen V (2005) Influence of incorporated wild *Solanum* genomes on potato properties in terms of starch nanostructure and glycoalkaloid content. J Agr Food Chem 53:5313-5325

Wentzinger LF, Bach TJ, Hartmann MA (2002) Inhibition of squalene synthase and squalene epoxidase in tobacco cells triggers an up-regulation of 3-hydroxy-3-methylglutaryl coenzyme a reductase. Plant Physiol 130:334-346

Yencho GC, Kowalski SP, Kobayashi RS, Sinden SL, Bonierbale MW, Deahl KL (1998) QTL mapping of foliar glycoalkaloid aglycones in *Solanum tuberosum* x *S. berthaultii* potato progenies: quantitative variation and plant secondary metabolism. Theor Appl Genet 97:563-574

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. Plant Genome 1:5-20

2. Allelic Variation in Genes Contributing to Glycoalkaloid Biosynthesis in a Diploid Interspecific Population of Potato

Manrique-Carpintero NC[1], Tokuhisa JG[1], Ginzberg I[2] and Veilleux RE[1](🖂)

(🖂) Author for correspondence

Tel: 540 231-5584

Fax: 540 231-3083

E-mail: potato@vt.edu

Authors contribution

Dr Richard Veilleux, Dr James Tokuhisa and Dr Idit Ginzberg were principal investigator in the project "Identification and allelic variation of genes involved in the potato glycoalkaloid biosynthetic pathway" funded by United States - *Israel* Binational Agricultural Research and Development (BARD). Dr Richard Veilleux directed this investigation and with Dr Tokuhisa and Dr Idit Ginzberg edited the article. I conducted the laboratory experiments, optimized the steroidal glycoalkaloid (SGAs) extraction protocol, collected data, did statistical analysis and wrote the first draft of this document as well as the corrections.

---

[1]Department of Horticulture, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

[2]Department of Vegetable Research, The Volcani Center, Agricultural Research Organization, Bet Dagan, Israel

Abstract

Steroidal glycoalkaloids (SGAs) are secondary metabolites produced in potato as a chemical defense. Identification of genetic factors associated with their biosynthesis and accumulation was assessed by a candidate gene approach and whole genome SNP genotyping. Allelic sequences spanning coding regions of four candidate genes expressing [3-hydroxy-3-methylglutaryl coenzyme A reductase 2 (*HMG2);* 2,3-squalene epoxidase (*SQE*); solanidine galactosyltransferase (*SGT1*); and solanidine glucosyltransferase (*SGT2*)] in the glycoalkaloid biosynthetic pathway were obtained from two potato species differing in SGA composition: *Solanum chacoense* clone USDA 8380-1 (*chc* 80-1) and *S. phureja* clone DH OT-B × N-B P5/AR2 (*phu* DH). Then an F_2 population between these plant lines was genotyped for foliar SGAs. F_2 plants with *chc* 80-1 alleles for *HMG2* or *SGT2* showed significantly greater accumulation of leptines or total SGAs, respectively, compared to plants with *phu* DH alleles. There was also a significant interaction between *HMG2* and *SGT2* such that plants homozygous for *chc* 80-1 alleles at both loci expressed the greatest levels of total SGAs, α-solanine and α-chaconine. *SGT1* and *SQE* allelic states did not affect SGA accumulation nor did they interact with other candidate genes for all the response variables. The concentrations of α-solanine, α-chaconine, leptine I, leptine II and total SGAs varied broadly among F_2 individuals among the four compounds. Regression analysis showed a positive correlation between α-solanine and α-chaconine accumulation as well as for leptine I and leptine II. Regarding biosynthesis of SGAs, Chi square analysis confirmed a pattern of 15:1 presence/absence ratios of synthesis for α-solanine and α-chaconine and a 3:1 presence/absence ratio for synthesis for leptines I and II. A whole genome single nucleotide polymorphism (SNP) genotyping analysis of an F_2 subsample verified the importance of *chc* 80-1 alleles at *HMG2* and *SGT2* for SGA synthesis and accumulation and suggested several additional candidate genes. SolCAP SNPs within a homolog of our *SQE* candidate were significantly associated with *chc* 80-1 alleles favoring accumulation of leptines. Loci on five potato pseudochromosomes were associated with synthesis of SGAs and loci on seven pseudochromosomes were associated with accumulation of SGAs. Two loci, on pseudochromosome 1 and 6, explained phenotypic segregation of α-solanine and α-chaconine synthesis. The whole genome study implicated genes that had

been previously associated with sterol and SGA production. These results contribute to our understanding of the genetic factors influencing SGA production in potato such that they may be more easily manipulated in breeding for pest resistance.

Key words:

Candidate gene approach, Infinium 8303 Potato Array, Solanaceae, *Solanum tuberosum*, *Solanum chacoense*, single nucleotide polymorphism- SNPs

Introduction

Potato (*Solanum tuberosum*) originated some 8,000 years ago in the altiplano of Peru has become a staple food for hundreds of millions of people throughout the world (FAO 2006). Intensive potato cultivation has resulted in many pest and disease problems. For this reason, it is necessary to develop new resistant potato varieties, which would reduce chemical control while increasing production (FAO 2008a). One characteristic of cultivated and wild potatoes is the production of biologically active secondary metabolites called steroidal glycoalkaloids (SGAs). These compounds evolved in potato species as part of their defense mechanism against pathogens and insects (Friedman 2006; Nema et al. 2008). Several types of SGAs have been reported in *Solanum*, the most common of which are the triose glycosides of solanidine aglycone, α-chaconine and α-solanine (Distl and Wink 2009; Ginzberg et al. 2009). However, leptines which have been found only in rare accessions of *S. chacoense* are known for conferring resistance to Colorado potato beetle (*Leptinotarsa decemlineata* Say), one of the most serious pests of the potato crop (FAO 2008b; Sinden et al. 1986; Sinden et al. 1980). Besides their role in plant resistance, SGAs can have toxic effects in humans due to their anticholinesterase and membrane disruption activities (Morris and Lee 1984). For that reason, an upper safe limit of 200 mg of total SGAs per kg of fresh weigh of potato tubers is a requirement for the release of commercial cultivars (Valkonen et al. 1996). For wild potato species to be used as a source of SGA-based resistance, it is fundamental to characterize the genes involved in their biosynthesis and regulation, to understand the variation in structural types and accumulation of SGAs (Ginzberg et al. 2009; Yencho et al. 1998).

SGAs are accumulated in all parts of the potato plant, depending on developmental, genetic and environmental factors. The concentration of SGAs varies with the type of organ and developmental stage, aerial and regenerative tissues (flowers, leaves, stolons and buds) as well as younger tissues exhibiting the greatest levels (Distl and Wink 2009; Friedman and Dao 1992; Friedman and McDonald 1997; Kolbe and Stephan-Beckmann 1997; Mweetwa 2009; Nema et al. 2008). The fact that each organ is capable of producing different amounts and types of SGAs and that there is no evidence of upward transport, suggests that biosynthesis and catabolism of SGAs occur at the

organ level (Friedman and McDonald 1997; Mweetwa 2009). The distribution and content of SGAs depend not only on tissue type, but also on genetic factors (Distl and Wink 2009; Friedman 2006). Interspecific crosses and somatic hybrids have given rise to progeny with unusual types and unexpected levels of SGAs compared to the parents (Laurila et al. 1996; Sanford et al. 1996; Yencho et al. 1998). Molecular markers associated with genomic regions hosting dominant genes with additive effects on SGA biosynthesis as well as recessive genes controlling high content of SGAs have been identified (Boluarte-Medina et al. 2002; Hutvágner et al. 2001; Ronning et al. 1999; Sagredo et al. 2006; Van Dam et al. 2003; Yencho et al. 1998). Furthermore, environmental factors also influence the level of SGAs. Growing under stress conditions such as drought or high temperature, mechanical injury during harvest, postharvest management and exposure of tubers to light are conditions that increase the abundance of SGAs in potato (Bejarano et al. 2000; Choi et al. 1994; Dale et al. 1993; Lafta and Lorenzen 2000).

Even though molecular markers and quantitative trait loci (QTL) have been associated with synthesis and accumulation of SGAs, neither the identification of underlying genes nor marker assisted selection (MAS) has been possible in potato breeding for this complex quantitative trait. Genetic heterogeneity and sampling of limited numbers of progeny have been the main obstacles to unravel the genetic architecture of complex traits in plants using QTL mapping (Holland 2007). The continuous phenotypic variation among progeny due to the effects of multiple genes and their interactions, the high level of heterozygosity in cultivated and wild potato species, and polyploidy are factors that have interfered with mapping genes responsible for SGAs. Besides linkage analysis followed by positional cloning and insertional mutagenesis, the candidate gene approach offers an alternative for discovery of functional genes associated with complex traits. Candidate genes are known-function genes that are presumed to control the trait of interest based on previous studies (Pflieger et al. 2001). Genetic polymorphism in candidate genes can be identified and then correlated with phenotypic trait variation.

The candidate gene approach can be used to discover structural and regulatory

genes involved in the biosynthesis and accumulation of SGAs. The recently sequenced potato genome (The Potato Genome Sequencing Consortium 2011) and the Illumina single nucleotide polymorphism (SNP) genotyping platform developed using elite potato germplasm (Hamilton et al. 2011) enable the identification of candidate genes within targeted genomic regions. Knowledge of genes involved in the glycoalkaloid biosynthetic pathway, genic regions surrounding QTL, markers associated with SGAs, and a whole genome scan focused on coding regions of SGAs in a segregating diploid population provide important resources for the identification of candidate genes.

The general biosynthetic pathway for SGAs has been partially defined (Figure 1). In the mevalonic acid pathway, twelve enzymes catalyze eleven biosynthetic reactions that lead to the formation of primary steroidal backbones. Some of these enzymes are members of gene families that may be associated with specialized biochemical or physiological functions (Suzuki and Muranaka 2007). Some reactions within primary isoprenoid metabolism have been associated with biosynthesis and accumulation of SGAs in potato. The enzyme 3-hydroxy-3-methylglutaryl coenzyme A reductase (HMGR) catalyzes the formation of mevalonic acid. Even though the *HMGR* family of genes plays an important role in biosynthesis of sterols and triterpenoids basic for plant development, they also influence the production of sesquiterpenoid phytoalexin defense compounds (Suzuki and Muranaka 2007). In potato *HMG2* and *HMG3* genes induced by wounding and pathogen inoculation have increased sesquiterpenoid production, and *HMG1* has been associated with SGA accumulation after wounding (Choi et al. 1994; Choi et al. 1992; Krits et al. 2007) Squalene synthase (SQS) condenses two farnesyl diphosphate molecules to form squalene. The transcript levels of *SQS* have been associated with high SGA levels in potato species (Ginzberg et al. 2009; Krits et al. 2007; Yoshioka et al. 1999). Squalene epoxidase (SQE) catalyzes the oxidation of squalene to 2,3-oxidosqualene. Inhibition of this enzyme activity decreased sterol biosynthesis (Wentzinger et al. 2002). Cholesterol is the precursor of SGAs (Arnqvist 2007); hence, SQE may have a regulatory role in SGA biosynthesis. Lanosterol synthase (LAS) and cycloartenol synthase (CAS) convert 2,3-oxidosqualene into lanosterol and cycloartenol, respectively (Ohyama et al. 2009; Suzuki and Muranaka 2007). The products of the lanosterol pathway may be secondary metabolites related to plant defense (Ohyama et al.

2009). Lanosterol and cycloartenol are metabolized further to produce cholesterol, campesterol and sitosterol; these are essentially the end-point sterols found in plants that are used as precursor of complex sterols as well as in the secondary metabolism (Ginzberg et al. 2009; Schaller 2004).

The secondary metabolism that leads the formation of SGAs begins with the modification of cholesterol. A hypothetical pathway from cholesterol to solanidine has been proposed; however, all of the enzymes involved have not been identified (Ginzberg et al. 2009). Solanidine is the aglycone precursor of α-chaconine and α-solanine (Canonica et al 1977, Tschesche and Hulpke 1997, and Heftmann 1983 cited by Arnqvist (2007)). Three enzymes that catalyze the final glycosylation reactions of the solanidine skeleton have been identified in potato (McCue et al. 2007a; McCue et al. 2007b; McCue et al. 2005; Moehs et al. 1997). Solanidine galactosyltransferase (SGT1) and solanidine glucosyltransferase (SGT2) catalyze the glycosylation of solanidine to γ-solanine and γ–chaconine, respectively. The final reaction for formation of α-solanine and α–chaconine is mediated by rhamnosyltransferase (SGT3). Functional genomic analysis using antisense constructs of *SGT1* and *SGT2* resulted in alteration of the α-solanine/α-chaconine ratio, but with unchanged total SGA levels (McCue et al. 2007b). This suggests that only the specific structures of SGAs are defined at this step, and that the total level of SGA may be controlled by genes linked to upstream metabolism in the pathway. Support for this hypothesis has come from the overexpression of C24-methytransferase type I (*SMT1*) which increased metabolic flux into the brassinosteroid pathway and reduced cholesterol and SGA levels in potato (Arnqvist 2007). These results also support the notion that cholesterol is the main precursor of glycoalkaloid biosynthesis, and that this pathway can be down-regulated by overexpression of genes of alternative routes in this network.

The main goal of this study was to identify allele-specific sequence polymorphism at four candidate genes involved in the biosynthesis of SGAs (*HMG2, SQE, SGT1* and *SGT2*) for *S. chacoense* clone *chc* 80-1 and *S. phureja* clone *phu* DH. A segregating F_2 population from a cross between these two clones (*phu* DH × *chc* 80-1) was analyzed to detect association of allele-specific markers with SGA biosynthesis

and/or accumulation. A second part of this study was to identify new candidate genes by whole genome SNP genotype analysis of selected individuals within our F_2 population that exhibited dissimilar levels of accumulation of SGAs.

Materials and methods

Plant material

Two clones, *Solanum chacoense* USDA 8380-1 (*chc* 80-1) and *S. phureja* clone DH OT-B × N-B P5/AR2 (*phu* DH), and an F_2 population from a cross between them were used in this study. *Solanum chacoense* produces many different types of SGAs, some at high levels, whereas *phu* DH does not accumulate SGAs. *Phu* DH is a dihaploid derived from the sexual progeny of two intermonoploid somatic hybrids from *S. phureja* BARD clones, selected for tuberization under long photoperiod (Haynes 1972; Johnson et al. 2001; Lightbourn and Veilleux 2007; Veilleux 1990) (Figure 2). In addition two *chc* 80-1 and 11 *phu* DH anther-derived monoploids were used to identify the allele-specific sequences of SGA candidate genes in the parental clones. With only 12 chromosomes the monoploids eliminated confusion between allelic differences and duplicate genes.

Due to self-incompatibility in the F_1 hybrids between *phu* DH × *chc* 80-1, we obtained the F_2 population from crosses between sibling F_1 plants. F_1 seeds were treated with 1,000 mg l^{-1} GA_3 solution overnight, and then sown in cell packs with Fafard® super fine germination mix (Farfard, Agawan, MA). The seedlings were transferred to D-40 Deepots™ (Hummert International, Earth City, MO) with Premier Horticulture® Pro-mix BX 15 days after sowing. The plants were grown under controlled environment (Conviron, Winnipeg) set to 60% relative humidity, 14 h photoperiod, 250 µmol $m^{-2}s^{-1}$ light intensity and day/night temperatures of 20 and 16 °C, respectively. Plants were fertilized with MiracleGro All Purpose® (Scotts Co, Marysville, OH) at the rate of 1 g l^{-1} every 15 days. At 77 days after germination leaves 4-6 from the shoot apex were harvested for SGA extraction and shoot apical meristems were collected for DNA extraction. Then the F_1 plants were transferred to the greenhouse and pollinations between them were done to develop the F_2 population. The total level of SGA quantified in the F_1 plants was used to select specific crosses. The nine plants in the F_1 population were classified as: low (L) <5 µg g^{-1} dry weight leaf tissue (DW), intermediate (I) 6-11 µg g^{-1} DW or high (H) level of SGA >11 µg g^{-1} DW. Crosses between plants with similar

and contrasting levels were made. F_1 plants #1, #2 and #6 were classified as L SGAs; #3, #4 and #9 as I; and #5, #7 and #8 as H. The F_2 population was started from 200 seeds of five different families derived from different cross combinations (family 3 L×H, family 5 L×L, family 7 I×H, family 8 L×H, and family 9 H×L) with 40 seed per family (Figure 2). The F_2 population was handled as described for the F_1 population. The plants were grown under a 16 h photoperiod in growth chamber conditions and discarded after SGA and DNA samples were taken at 45 days after transplant to Deepots (Steuwe & Sons, Tangent, OR).

DNA extraction and PCR amplification

Total DNA was extracted using a cetyltrimethylammonium bromide (CTAB) protocol (Doyle and Doyle 1987; Murray and Thompson 1980; Stewart and Via 1993). Homologous nucleotide sequences from solanaceous species for *HMG2*, *SQE*, *SGT1* and *SGT2* were obtained from the available data at GenBank database between 2008 and 2010. These sequences were aligned to design primers in conserved coding regions (Table 1). The DNA fragment of *HMG2* coded for a 134 amino acid (aa) sequence with catalytic, NADP(H) binding, substrate binding and tetramerization residues. The 134 aa region of *SQE* spanned more than one functional domain. A multidomain region of 277 aa was amplified for *SGT1* (single exon), and a segment of 189 aa with active site conserved domains and TDP-(thymidine diphosphate) binding site residues for *SGT2* (single exon). The last is a binding site in the interdomain cleft of glycosyltransferase family to form a complex with TDP-glucose in the biosynthesis of vancomycin (Mulichak et al. 2004). All primers were designed using DNASTAR® Lasergene 9 core suite software for sequence analysis and assembly. PCR was performed in 25 µl of 10 mM Tris-HCl, pH 8.43, 50 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM of each dNTP, 0.64 µM of each primer, 0.5 units of Takara Taq™ and 100-200 ng of genomic DNA template. Standard cycling conditions were 7 min of initial denaturation at 94 °C, 0.5-1 min at 94 °C, 0.5-1 min of annealing at the appropriate T_m and 0.5-1 min of extension time for 30-40 cycles. The reactions were finished by a 5 min incubation at 72 °C. PCR products were gel-purified (Qiagen Gel Extraction Kit, Hilden, Germany) and used for cloning or direct sequencing. PCR products from diploid heterozygous *chc* 80-1 and *phu* DH were

excised from agarose gels, purified and cloned in the pGEM-T vector system (Promega, Madison, WI) following the supplier's instructions. At least five colonies from the cloned PCR products were sequenced in both directions (Tufts University Core Facility, Boston, MA). PCR products from monoploid lines were sent for direct sequencing in both strands for each sample. PCR conditions for *SGT1* were adjusted due to the presence of PCR recombinant artifact sequences. Reconditioning PCR was used (Judo et al. 1998; Lenz and Becker 2008; Thompson et al. 2002). In this case the final PCR product was derived from two sequential PCR reactions of 20 cycles each. In the second reaction 2 µl of product from the first reaction were used as template instead of genomic DNA. Three different reconditioned PCR reactions were done per DNA template from monoploid lines. PCR products from *phu* DH monoploids were excised from agarose gels, purified and sequenced. PCR products from *chc* 80-1 monoploids were cloned and sequenced. At least three colonies from each cloned PCR product were sequenced in both directions. In addition to allelic sequencing, flanking sequences neighboring *SGT1* in *chc* 80-1 monoploids were studied by high-efficiency thermal asymmetric interlaced PCR (hiTAIL-PCR) according to (Liu and Chen 2007).

Identification of allelic sequences and genotyping

Forward and reverse nucleotide sequences from the same PCR sample were aligned to identify consensus sequences. Visual inspection of peak quality in chromatograms was used to resolve conflicting sites in the sequence analysis. The set of consensus sequences identified for each genotype and gene was aligned to identify haplotype alleles. A minimum alignment of at least three identical consensus sequences was used to determine allelic sequences. Sequence errors detected when consensus sequences following the same pattern were aligned were also corrected. The sequences were confirmed by comparing with candidate entries in the Genebank. Sequence alignment and editing were done using DNASTAR® Lasergene 9 core suite software. Sequence nucleotide polymorphisms were analyzed using the DNA Sequence Polymorphism (DnaSP5) software version 5.0 (Librado and Rozas 2009). The allelic sequence alignments per gene used for analysis in DnaSP5 were done by CLUSTAL W method (Thompson et al. 1994) using Molecular Evolutionary Genetics Analysis

(MEGA5) software version 5.0 (Tamura et al. 2011).

TaqMan® technology and electrophoresis separation of allele-specific PCR products were used to genotype *chc* 80-1 and *phu* DH alleles identified in the four candidate genes. Allele-specific primers were designed for tag SNP and polymorphic sites detected in the allelic sequences (Table 2). An allelic genotyping strategy was defined for each gene:

HMG2, Even though tag SNPs that differentiated *HMG2* alleles were identified, the intron size difference between *phu* DH alleles (P1 and P2) and conformational configuration due to a SNP in the *chc* 80-1 allele (C1) were used to screen the F_2 population. DNA fragment analysis of simple PCR products, using primers flanking the *HMG2* intron, was done in the high-resolution QIAxcel system® that discriminated the three alleles.

SQE, TaqMan® SNP genotyping assay and simple PCR-based DRR-GL marker system (Ramkumar et al. 2010) were used to screen *SQE* alleles (P1, C1 and C2) at two SNP positions. The TaqMan® assay for one *SQE* SNP was designed by Applied Biosystems® to differentiate the C2 allele from C1 and P1. PCR and genotyping, by fluorescent allelic discrimination, was performed in an Applied Biosystems real time 7300 fast instrument following the kit instructions. The PCR-based DRR-GL marker system was used to detect the P1 allele from C1 and C2. Two allele specific primers were designed at the SNP, each with a polymorphism at the last base position from the 3’ end. The primers were used in forward and reverse orientation from the SNP position. The tag SNP was located at a position that divided the gene fragment into two different size portions. In a multiplex reaction, using the regular flanking primers of the gene fragment and the allele specific primers, we could discriminate a co-dominant amplification pattern for the target SNP.

SGT1, Multiple polymorphic sites in the sequence and a promoter region were used to design allele specific primers for five alleles at *SGT1* (C1, C2, C3, P1 and P2). Single PCR reactions with allele-specific primers at forward and reverse position were necessary to detect C3 and P2 alleles. Another reaction with flanking primers and one

allelic primer was used to screen for C1. Finally a multiplex reaction differentiated C2 and P1 alleles.

SGT2, To screen *SGT2* alleles (C1 and P1), a region containing a 9 bp deletion and a segment with two SNPs were used to design allele specific primers. A multiplex reaction with flanking primers and the allele specific primers discriminated C1 and P1.

The robustness and optimum annealing temperature for the allele specific reactions were adjusted through gradient PCR and longer extension time in the parental lines. PCR reactions were performed in 15 µl of 1X Immomix^{TM} (Bioline, Taunton, Mass.), 2.5 mM $MgCl_2$, 0.2 µM of each primer, and 50-100 ng of genomic DNA template. Standard cycling conditions were 10 min of initial denaturation at 95 °C, 0.5 min at 94 °C, 0.5 min of annealing at the appropriate T_m and 1-2 min of extension time for 30-40 cycles. The reactions were finished by 5 min incubation at 72 °C.

SGA extraction and quantification

After 72 h of lyophilization (Labconco, Kansas City, MO), freeze-dried leaf tissue from two parental lines, nine F_1 plants and 148 F_2 plants was ground and SGAs extracted following a modified procedure (Edwards and Cobb 1996). Leaf powder (30 mg) and 1 ml of extraction buffer (5% v/v acetic acid, 0.02 M heptane sulfonic acid) were mixed by vortex and ultrasonication for 2 sec (Digital sonifier® cell disruptor, Branson Ultrasonic corporation, NY, USA with 20% amplitude). The extract was maintained for 15 min at 1,200 rpm in a microtube thermal-mixer, and centrifuged 3-5 min at 16,000 rpm. The supernatant was sieved in a 50 µm filter plate and the precipitate used for a second cycle of extraction. SGAs were concentrated and cleaned by solid phase extraction using using Sep-pak® classic C_{18} cartridge columns (250 mg) (Waters, Milford, MA, USA). The columns were activated with 5 ml of methanol (MeOH), and equilibrated with 5 ml of extraction buffer. Leaf extract was applied and a series of washes: 7.5 ml of water, 5 ml of 50 mM ammonium bicarbonate (NH_4 HCO_3), 5 ml of 50 mM NH_4 HCO_3:MeOH (1:1 v/v), and 7.5 ml of water. SGAs were eluted with 1.2 ml of elution buffer (80% v/v MeOH, 0.5% v/v formate).

SGAs were separated by high performance liquid chromatography (HPLC, Agilent HP 1200 Series, Santa Clara, CA) on a C-18 reverse-phase column (Agilent Eclipse XDB-C18, 5 µm pore size and 4.6 × 150 mm) and quantified using a photodiode array detector. SGAs were eluted using a binary gradient system consisting of Solvent A (30% acetonitrile, 6 mM Tris-HCl, pH 8.0) and Solvent B (80% acetonitrile, 6 mM Tris-HCl, pH 7.6) at a flow rate of 0.3 ml/min at 25 °C column temperature. The gradient elution was: 0-0.5 min, 0% B 0.5-8.5 min, 0-30% B; 8.5-12 min, 30-100% B; 12-16 min 100% B, 16-16.5 min, 100-0% B; 16.5-21 min, 0% B. Eluent was monitored by diode array detection at 202 nm. Quantification of SGAs was based on peak absorbance area at A_{202nm}, which was converted to µg using a response factor (3.4 ng/mAU) for all four SGAs analyzed (α-solanine, α-chaconine, leptine I and leptine II) as determined by Mweetwa et al. (2012). Concentrations were expressed in µg of SGAs per mg of dry weight leaf tissue (µg mg^{-1} DW) for further analysis. Calibration curves for α-solanine and α-chaconine (Sigma) were generated with purified standards.

SNP chip analysis

The SolCAP 8303 Illumina Infinium potato SNP chip was used to identify new candidate genes. Genomic DNA of a subsample of 43 F_2 plants was used for this study. The subsample was selected based on F_2 individuals with sufficient viable stored tubers. Different numbers of plants coming from families 3, 5, 7, 8 and 9 were available (2, 1, 5, 23 and 12 respectively). The SNP genotyping facility at Michigan State University processed the samples on an Illumina iScan Reader utilizing the Infinium® HD Assay Ultra (Illumina, Inc., San Diego, CA) and the Infinium 8303 Potato Array. The 8,303 SNP data were filtered and used for contingency table analysis and ANOVA based on synthesis and accumulation of α-solanine, α-chaconine, leptine I and leptine II. SNPs that were monomorphic for all individuals, and SNPs with a no-call rate >10% (greater than five progeny with missing genotypes) were eliminated from the initial data set. From 3,196 analyzed SNPs per compound, significant SNPs in the contingency tables were selected based on P-value < 0.02, F_2 expected genotypic segregation ratios of 1:2:1 or 1:1 based on Chi-square tests, and strong association of homozygous alleles with synthesis of each compound (0 individuals). In addition, significant SNPs in ANOVAs were selected

based on $R^2 > 0.2$, P-value < 0.02, F_2 expected genotypic segregation ratios of 1:2:1 or 1:1 based on Chi square tests, and SNPs with superscaffold and pseudomolecule information on the potato genome browser (http://solanaceae.plantbiology.msu.edu/cgi-bin/gbrowse/potato/). Finally, superscaffolds with significant SNPs were mapped using the megabase (Mb) physical distances published by Felcher et al. (2012). These physical distances used as reference the recently sequenced potato genome (The Potato Genome Sequencing Consortium 2011). Superscaffolds with significant SNPs were BLAST searched on the potato genome browser to identify possible candidate genes related to biosynthesis and/or regulation of SGAs. The new putative candidate genes were listed using the potato genome gene model references. These names start with the initials of potato genome sequencing consortium (PGSC) followed by Assembly Version 3 of DM (double monoploid) *S. tuberosum* Group Phureja DM1-3 516 R44 (CIP801092) and the gene number.

Statistical analysis

A completely randomized design was used for F_2 and parental plants in the growth chamber and for SGA analysis. The amounts of α-solanine, α-chaconine, leptine I and leptine II were estimated in three subsamples of the leaf tissue collected per plant. The average of the three subsamples was used for statistical analysis because the subsampling covariance was less than 20% for 90% of samples. Three F_2 samples with highly variable subsamples were taken out of the analyses. Finally, subsampling analysis did not show major differences with ANOVA based on averages. ANOVA of SGA response variables using as source allelic states for each candidate gene or for SNPs from whole genome screening, and interactions for pairs of genes were conducted. Allelic sequences per gene, identified in the segregating F_2 population, were consolidated in a unique code from *phu* DH (PP), *chc* 80-1 (CC) and heterozygous (CP). Chi-square analyses were done to confirm phenotypic (presence/absence) and genotypic segregating ratios in the F_2 population. All statistical analyses were done using JMP® 9. SAS Institute Inc., Cary, NC, USA.

Results

Allelic sequences and sequence polymorphism

Allelic sequences from coding region fragments of four candidate genes at various parts of the SGA biosynthetic pathway were obtained. PCR amplification, cloning and sequencing from *chc* 80-1 and *phu* DH as well as direct sequencing of monoploid PCR products were used to identify the sequences of *SQE*, *HMG2*, *SGT1* and *SGT2* occurring in heterozygous diploid plants. However, a different strategy: PCR amplification, cloning and sequencing from *chc* 80-1 monoploids, was necessary to identify *SGT1* allelic sequences. Haplotypes defined the alleles for *phu* DH and *chc* 80-1. Thirteen allelic sequences were found for four candidate genes. A single allele was found for *SQE* and *SGT2* in *phu* DH, and for *HMG2* and *SGT2* in *chc* 80-1. Two alleles for *HMG2* and *SGT1* in *phu* DH, and for *SQE* in *chc* 80-1. PCR recombinant artifacts for *phu* DH and *chc* 80-1 cloned sequences were originated because of the heterozygous diploid condition of these plants, and also because *SGT1* occurs within a family of highly similar sequences (McCue et al. 2007a). Analysis of 20 different cloned monoploid sequences confirmed the presence of multiple alleles for *SGT1* in *chc* 80-1 monoploids due to possible gene duplication. In two different monoploids of *chc* 80-1 we found the same three allelic sequences, suggesting that there were three copies of this gene in the *chc* 80-1 parent. In fact allelic-specific primers designed for each of the three copies amplified fragments that always co-segregated in both *chc* 80-1 monoploids and the F_2 individuals. In addition, we used hiTAIL PCR analysis to identify two different promoter regions and a specific *chc* 80-1 sequence for the 3'untranslated region (UTR) adjacent these sequences.

Alignment of allelic sequence fragments revealed synonymous, non-synonymous, and indel features (Table 3). In 13 allelic sequences we found 106 SNPs generating 108 mutations. An average of 687 nucleotide sites was analyzed per gene. The sequenced fragments for *HMG2* and *SQE* were within regions with one and three introns, respectively. The proportion of SNPs was greater in the introns than in the exons for these genes (61.5% and 86.4% each). Regarding the frequency of SNPs only in the exons for the four candidate genes, *HMG2* (1/40 bp), *SQE* (1/134 bp), *SGT1* (1/16 bp) and

SGT2 (1/63 bp), the greatest frequency of SNPs was in *SGT1* and the lowest in *SQE*. Genes in primary metabolism, *HMG2* and *SQE*, showed more silent (synonymous and non-coding region) SNPs than non-synonymous SNPs. On the other hand, genes in secondary metabolism, *SGT1* and *SGT2*, had more non-synonymous than synonymous SNPs. Other sequence polymorphisms found were indels, three in intronic regions and three in exonic regions. Two of the indels were located in the exon of one *chc* 80-1 allele; these shifted the reading frame and produced stop codons.

F_2 population phenotypic analysis

Total SGA, α-solanine, α-chaconine, leptine I and leptine II content was determined for parental lines and their F_2 progeny (Table 4). The foliar extract of the *phu* DH female parent did not exhibit any detectable SGAs, whereas the *chc* 80-1 male parent produced all four of the SGAs, with greater levels of leptine I and II. The accumulation of SGAs in the 148 F_2 individuals was highly variable. Total SGAs ranged from 0-126.9 µg mg^{-1} DW, α-solanine from 0-46.4 µg mg^{-1} DW, α-chaconine from 0-84.9 µg mg^{-1} DW, leptine I from 0-44.7 µg mg^{-1} DW, and leptine II from 0-18.5 µg mg^{-1} DW. Regression analyses showed a positive correlation between α-solanine and α-chaconine accumulation (n=140, r^2=0.84, P<0.001*) as well as for leptine I and leptine II (n=38, r^2=0.86, P<0.001*). However, there was no correlation between the accumulation of leptines and α-solanine or α-chaconine. The content of α-solanine and α-chaconine in 35 and 85 plants, respectively, exceeded the parental line *chc* 80-1. However, there was not a single plant with more leptine I and leptine II than *chc* 80-1. Quantile distribution showed that the greatest levels of accumulation of α-solanine and α-chaconine were concentrated in 10% of total of the F_2 population analyzed data and intermediate levels in 65% of the population. Leptine I and leptine II accumulated to the greatest levels in 2.5% of the evaluated population and were intermediate in 15%.

SGA quantitative data were converted to a binary format (1/0 = presence/absence) to study the genetics of SGA biosynthesis. Analysis of frequency showed that total SGAs, α-solanine, and α-chaconine had similar patterns of presence in the F_2 population; 94.6% synthesized α-chaconine and 93.2% both α-solanine and α-

chaconine. This phenotypic distribution corresponded to a segregation ratio expected for two independently inherited Mendelian genes. One sixteenth (9 of 148) of the F_2 population did not synthesize any SGAs. In comparison, the frequency distribution of leptine I and leptine II followed a single gene Mendelian model where 74.3% and 78.4% of F_2 plants did not synthesize leptine I and leptine II, respectively. Chi-square test (χ^2) showed no difference between observed and expected 15:1 and 3:1 ratios (Table 5). Contingency table analyses of presence/absence between either α-solanine, α-chaconine, leptine I or leptine II showed different patterns of association in the synthesis of these compounds. Strong association between synthesis of α-solanine and α-chaconine (n=148, Likelihood ratio χ^2_1=52.2, P=0.0001*) and leptine I and leptine II (n=148, Likelihood ratio χ^2_1=121.4, P=0.0001*). The synthesis of either α-solanine or α-chaconine was not associated with the synthesis of either leptine I or leptine II, but the synthesis of leptine I was weakly associated with synthesis of α-chaconine as well as for leptine II with α-solanine (n=148, Likelihood ratio χ^2_1=4.9, P=0.03*, and n=148, Likelihood ratio χ^2_1=5.1, P=0.02*).

The accumulation ratios of α-solanine/α-chaconine and leptine II/leptine I were analyzed in the subsamples of plants that synthesized these compounds (Table 4). The range from total subsample data varied broadly from 0.2-3.8 for α-solanine/α-chaconine and from 0.1-0.5 for leptineII/leptine I, compared with 1.2 and 0.5 in parent *chc* 80-1. However, when outliers were omitted, 65% of data were between 0.3-0.6 and 0.2-0.4, respectively. Even though the synthesis of leptines was partially associated with the synthesis of α-solanine and α-chaconine in the entire population, the regression analyses showed no correlation with the accumulation between these groups. However, there was a weak positive correlation between accumulation ratios of α-solanine/α-chaconine versus leptineII/leptine I in the subsample that synthesized all compounds (n=38, r^2=0.30, P<0.004*).

F_2 population genotyping analysis

Sequence polymorphisms from allelic sequences were used to screen the allelic structure at four candidate genes in a segregating F_2 population with 148 individuals. An

intron size difference among *HMG2* alleles and allele-specific primers targeting SNP and indel regions for *SQE*, *SGT1* and *SGT2* were the molecular markers used. The observed F_2 genotypic segregation ratios in the progeny did not significantly differ from the 1:2:1 expected based on Chi-square analysis for all genes (Table 6). In effect, the total data of five families were pool in an F_2 population because based on test of homogeneity (Strickberger 1968); the samples per family were more than 95% homogeneous for all genes. ANOVA using the allelic composition at each candidate gene as source and response variables concentrations of total SGAs, α-solanine, α-chaconine, leptine I and leptine II, showed significantly greater levels of accumulation of total SGAs in samples homozygous for *chc* 80-1 alleles for *SGT2* ($P<0.05$*), and for leptine I and leptine II when *HMG2 chc* 80-1 alleles were present ($P<0.03$* for both) (Table 7). The two-way ANOVA analysis of six different combinations of two genes and five response variables showed no interaction or significance of *SGT1* and *SQE* with any other gene for all the response variables. However, there was a significant interaction between *HMG2* and *SGT2,* with the greatest levels of total SGAs, α-solanine and α-chaconine expressed in plants homozygous for *chc* 80-1 alleles at both loci (Table 8).

The association of allelic states of each gene and synthesis of α-solanine, α-chaconine, leptine I and leptine II was studied in data converted to presence/absence format (1/0). For *SQE* and *SGT1* there was no association with the synthesis of any of the SGAs, while for *HMG2* and *SGT2* significant differences were detected. For *HMG2* the contingency table showed that α-solanine and α-chaconine were produced regardless of genotype, and that the few non-producer samples (10 and 8) had preferentially *HMG2* heterozygous alleles and not *chc* 80-1 homozygous alleles (n=147, Likelihood ratio χ^2_2= 10.0, P=0.01*, and n=147, Likelihood ratio χ^2_2=7.3, P=0.03*). In the case of leptine I and II there was strong association between presence of *HMG2 phu* DH homozygous alleles and non-synthesis, while either synthesis or not occurred when *chc* 80-1 alleles were present even though they were highly associated with synthesis (n=147, Likelihood ratio χ^2_2=30.2, P<0.0001*, and n=147, Likelihood ratio χ^2_2=24.2, P<0.0001*). A similar trend was found for *SGT2*; the contingency table analyses detected a weak association of *phu* DH homozygous alleles and non-synthesis, 3.2% and 2.6% (1 of 148) of plants with this genotype were able to synthesize leptine I and II (n=148, Likelihood ratio χ^2_2=14.1,

P=0.001*, and *n*=148, Likelihood ratio χ^2_2=10.6, P=0.005*). In addition, *chc* 80-1 alleles were more associated with synthesis of these compounds. For α-solanine and α-chaconine, no significant differences were detected with any *SGT2* genotypes.

The F_2 population is made up of five families from specific crosses between sibling F_1 plants with contrasting or similar levels of SGAs. Statistical analyses of SGA accumulation by family were done. Families 7, 8 and 9 had greater total SGAs, α-solanine and α-chaconine, and for leptine I and II, family 7 had significantly more accumulation than the other families (Table 9). The five families exhibited the expected F_2 genetic segregation ratio of 1:2:1 (PP:PC:CC) per gene based on Chi-square test, except for *SQE* in family 5 (Table 6). Besides the average level of SGA accumulation, analyses between allelic structure at each gene and level of SGA accumulation for the five response variables by families were done. These analyses did not show any pattern of accumulation in the families associated with specific allelic composition for all response variables as found in the entire population. Variation for SGA accumulation and the few homozygous lines for each gene, due to the 1:2:1 segregation ratio decreased the power of statistical analysis to detect differences among genetic structure by families. Instead of having similar results among families we found some point significances. *SQE chc* 80-1 alleles in family 9 had significantly more α–solanine (P=0.02*), *HMG2 chc* 80-1 alleles in family 8 and 9 had significant more leptine I as well as leptine II for family 9 (P=0.03*, 0.02* and 0.02*), and *SGT1 phu* DH alleles in family 8 had greater levels of total SGAs and α–chaconine (P=0.04* and 0.04*).

SNP chip analysis

A whole genome SNP genotype analysis of 43 F_2 plants was done to attempt to identify genomic regions and likely candidate genes related to synthesis and accumulation of SGAs. Association of significant SNPs with presence/absence of SGAs was studied by a contingency table analysis (Supplementary Table 1). Significant SNPs were identified, their positions observed in the potato genome browser (http://solanaceae.plantbiology.msu.edu/cgi-bin/gbrowse/potato/), and putative candidate genes located in neighboring regions were selected. For synthesis of α-solanine and α-

chaconine, we found 34 highly significant (P<0.02*) SNPs. Most of the SNPs (*n*=33) occurred on five superscaffolds clustered at the bottom of pseudochromosome 6, with a single significant SNP located at bottom of pseudochromosome 1. Eleven of the SNPs on pseudochromosome 6 were clustered in four superscaffolds and had the lowest P-value (P=0.005* for all 11) relative to the others (Figure 3). There was no recombination in our population for these 11 SNPs, so that they segregated as two haplotypes: *chc* 80-1 and *phu* DH. Effectively the haplotype segregation yielded a 10 PP:22 PC:11 CC ratio (χ^2=0.97 ns, for the expected 1:2:1 ratio). The four F_2 plants with *phu* DH alleles for the SNPs on both pseudochromosomes 1 and 6, either did not synthesize SGAs (*n*=3) or expressed a low level of α-solanine and α-chaconine (*n*=1). Since the locus on pseudochromosome 1 was homozygous in *phu* DH and heterozygous for *chc* 80-1 with one allele identical to *phu* DH, and the haplotypes for the 11 SNPs on pseudochromosome 6 were homozygous but different in both parents, the expected dihybrid segregation ratio would be 2:2:4:4:2:2 in the F_2 population for two independent loci, and the expected ratio did not differ from this (χ^2=0.56 ns). Possible candidate genes found in this region of pseudochromosome 1 included sterol desaturase (PGSC0003DMG400001676) and a cytochrome P450 on pseudochromosome 6 cytochrome P450 71D7 (PGSC0003DMG403020453).

Leptine I and leptine II shared 58 of the same SNPs for synthesis of SGAs, of the 59 and 60 that were found highly significant for leptine I and II, respectively (Supplementary Table 1). In the contingency table analyses for presence/absence of leptines, 18 superscaffolds with 55 highly significant SNPs were situated between 1.5-22.1 Mb of pseudochromosome 2 (Figure 3). There were four haplotypes in our F_2 population for this linkage block: ten individuals were homozygous for the *phu* DH haplotype, 13 homozygous for the *chc* haplotype, 18 heterozygous, and there were two individuals with recombinant haplotypes. None of the ten F_2 with only the *phu* DH haplotype synthesized leptines. One SNP (solcap_c1_15974) occurred within our *HMG2* candidate gene. Two significant SNPs, on pseudochromosomes 12 and 8 were also negatively associated with synthesis of leptines when F_2 individuals carried the *phu* DH allele. The SNP on pseudochromosome 8 was 1.6 Mb from our *SGT2* candidate gene, and demonstrated a similar segregation pattern with only two exceptions. The *chc* 80-1 alleles

for one SNP on pseudochromosome 1 and two on pseudochromosome 7 were positively associated with synthesis of leptine I and leptine II. The SNP on pseudochromosome 1 occurred on the same superscaffold identified for α-solanine and α-chaconine and harbored a putative candidate gene, sterol desaturase. Possible leptine biosynthetic and regulatory candidate genes occurring in regions nearby significant SNPs on pseudochromosomes 7, 2 and 8 included: cytochrome P450 with steroid hydroxylase activity CYP72A58 (PGSC0003DMG402012386) on pseudochromosome 7; *DWARF1/DIMINUTE* (PGSC0003DMG400011801), *HMG2* (PGSC0003DMG400003461), BRASSINAZOLE-RESISTANT 1 protein (PGSC0003DMG400004501), sterol-C5(6)-desaturase homolog (PGSC0003DMG400010415), *SQE* (PGSC0003DMG400003324) and N-acetyltransferase (PGSC0003DMG400013015) on pseudochromosome 2; *BRASSINOSTEROID INSENSITIVE 1-associated receptor kinase 1* (PGSC0003DMG400030075) and *SGT2* (PGSC0003DMG400017508) on pseudochromosome 8.

If we base our ANOVAs for allelic composition at each of 3,196 polymorphic SNPs on accumulation (amount of α-solanine, α-chaconine, leptine I and leptine II, as determined by HPLC) instead of synthesis, we identified 103 unique significant SNPs located on seven pseudochromosomes for the four different SGAs (Table 10 and Supplementary Table 2). For α-solanine and α-chaconine, 33 and four significant SNPs were found, respectively. Since three of the four significant SNPs for α-chaconine were also significant for α-solanine, we will consider them together. Of these 34 significant SNPs, 15 (45%) and 11 (32%) were located on pseudochromosomes 6 and 11 (Table 10); the other eight were scattered, one each on pseudochromosomes 1 and 7 and three each on pseudochromosomes 8 and 10. Clusters of significant superscaffolds surrounding a candidate gene would be expected due to linkage (Figure 3). Within each cluster, we might expect the P-value of the ANOVA to decrease as the SNPs approach a candidate gene. The 15 significant SNPs on pseudochromosome 6 occurred on seven superscaffolds spanning 10.7 Mb at the bottom arm and one superscaffold located at the top. Although all of the P-values indicated a high level of significance, a nadir P-value spanning three superscaffolds (PGSC0003DMB000000315, PGSC0003DMB000000227 and

PGSC0003DMB000000158) suggests the following candidate genes: CYP 72A58 (PGSC0003DMG400002351), CYP71D7 (PGSC0003DMG400034479), sesquiterpene synthase 2 (PGSC0003DMG400013034) and geraniol 10-hydroxylase (PGSC0003DMG400020452). Likewise, the 11 significant SNPs on pseudochromosome 11 covered four superscaffolds, with the lowest P-values occurring on superscaffold PGSC0003DMB000000148, implicating the candidate gene UDP-rhamnose:rhamnosyltransferase (PGSC0003DMG400016194). The superscaffolds bearing the significant SNPs on pseudochromosomes 1, 7, 8 and 10 did not have obvious candidate genes related to SGA metabolism.

Leptine I and leptine II shared 39 significant SNPs of the 48 and 62 that were found highly significant for accumulation of each compound (Table 10 and Supplementary Table 2). Since there is likely interconversion between the two leptines, we will consider only the 39 common significant SNPs. As with α-solanine and α-chaconine, pseudochromosomes 6 and 11 were implicated, accounting for ten (26%) and 23 (59%) of the significant SNPs. The other six were scattered, with one, two and three on pseudochromosomes 1, 7 and 8, respectively. Again, following the trend of selecting candidate genes from superscaffolds harboring SNPs with the lowest P-values, we suggest three (PGSC0003DMB000000087, PGSC0003DMB000000150 and PGSC0003DMB000000686) on pseudochromosome 6 and three (PGSC0003DMB000000131, PGSC0003DMB000000017 and PGSC0003DMB000000433) on pseudochromosome 11 as the most likely (Figure 3). On pseudochromosome 6 the candidate genes included protein farnesyltransferase/geranyltransferase (PGSC0003DMG400027009), 0-methyltransferase (PGSC0003DMG400020095) and C-8, 7 sterol isomerase (PGSC0003DMG400027684) and on pseudochromosome 11 UDP-glucose:glucosyltransferase (PGSC0003DMG400028670) and 3-hydroxy3-metylglutaryl coenzyme A reductase 1 (PGSC0003DMG400046343). Pseudochromosome 1 and 7 had one and two highly significant SNPs. The same superscaffold that was positively associated with synthesis of α-solanine, α-chaconine and leptines on pseudochromosome 1, as well as the same SNPs associated with synthesis of leptines on pseuchromosome 7 in the previous contingency table analysis were significant for accumulation of leptine. For both loci, homozygosity

for the *chc* 80-1 homozygous alleles was associated with high accumulation and synthesis of leptines. SNPs on pseudochromosome 8 did not indicate obvious candidate genes.

Regarding our four original candidate genes, three of them (*HMG2*, *SGT1* and *SGT2*) had SNPs that were included in the SolCAP 8303 Infinium array. We used available Solanaceae sequences to design and amplify the fourth candidate gene, *SQE*, since the potato genome sequence was not available at the time. Using the current potato genome browser, we can locate the *SQE* sequence amplified by our primers on pseudochromosome 4. We had nominated all of our candidate genes for inclusion in the SNP chip design; however a homolog of *SQE* on pseudochromosome 2 was inadvertently selected. Our candidate *SQE* found on pseudochromosome 4 was 73% similar at the nucleotide level with the *SQE* on pseudochromosome 2. Four (solcap_snp_c1_9695, 9696, c2_32413 and 32417) of six SNPs within this second *SQE* gene (PGSC0003DMG00003324) were significantly associated with leptine I and II, with heterozygotes intermediate in accumulation between the two homozygous states (Table 11). The SNPs occurred in exons 1, 5 and 6 of the *SQE* gene model; those in exons 5 and 6 were nonsynonymous. Homozygous states for the *chc* 80-1 allele for all SNPs were associated with greater leptine I and II accumulation. Of the five SNPs within the coding region of *HMG2* (PGSC0003DMG00003461) included on the SNP chip, three (solcap_snp_c1_15972, 15973 and 15974) were polymorphic in our population and their allelic states were significantly associated with leptine I and II accumulation. All three occurred within the first exon of *HMG2*, one (15973) was nonsynonymous, and another (15974) discriminated homozygotes with the *chc* allele expressing greater leptine accumulation. All five SNPs within the coding region of *SGT2* (PGSC0003DMG00017508) were not polymorphic in our population. For *SGT1* (PGSC0003DMG00011749), three of 12 SNPs included on the Illumina chip were polymorphic in our population but none of the three was significantly associated with any SGA accumulation. Among the 22 SNPs in our candidate genes that were included on the potato Illumina array, only two within *SGT1* (solcap_snp_c2_23384 and _23385) were within our sequenced fragments.

Discussion

Allelic sequences and sequence polymorphism

Sequencing and haplotype analysis of fragments of coding regions were used to identify alleles of four candidate genes in two potato species. Thirteen allelic sequences were found and their synonymous SNPs, nonsynonymous SNPs, and indel features identified. Fragments of genes in primary metabolism (*HMG2* and *SQE*) had more silent SNPs (synonymous and in non-coding region), whereas genes in secondary metabolism (*SGT1* and *SGT2*) had more nonsynonymous SNPs. In a metabolic pathway, upstream genes (primary metabolism) are expected to have major natural selection pressure on mutations that can affect the coded protein function than those downstream (Waxman and Peck 1998). Upstream enzymes are more pleiotropic, being essential to produce intermediary compounds used in synthesis of several primary and secondary metabolites whereas secondary metabolism is characterized by specialized gene function (Pichersky and Gang 2000). Gene duplication and subsequent mutations are part of the evaluative mechanisms to provide novel gene function (Todd et al. 1999). New SGA compounds are defined by structural modifications of either the glycosidic residues or aglycone unit. Genetic background due to specialized genes or allelic interactions could be related to synthesis of new SGAs, as suggested by interspecific crosses of potato with wild species (Sagredo et al. 2011; Väänänen et al. 2005). For *SGT1* we found three allelic sequences in the *chc* 80-1 monoploids; two had different promoter regions whereas all three had the same 3'UTR. This, in addition to the observed co-segregation of the three alleles, suggested gene duplication in the high SGA clone, *chc* 80-1.

F_2 population phenotypic analysis

The synthesis and accumulation of α-solanine and α-chaconine followed different distribution patterns in the segregating F_2 population than those for leptine I and leptine II. Firstly, even though the distribution of accumulation was highly variable in general; most of the population accumulated α-solanine and α-chaconine (±95%) whereas only ±25% accumulated leptines I and II. Secondly, there were positive correlations between accumulation of α-solanine and α-chaconine and also between leptine I and

leptine II. However, there was no correlation in the accumulation of α-solanine or α-chaconine with leptine I or leptine II. Thirdly, synthesis of leptines was weakly associated with synthesis of α-solanine and α-chaconine grouped by similar glycosidic moiety, but not vice versa. This suggests different genetic mechanisms that regulated synthesis and accumulation of these two groups of SGAs. Specialized genes controlling synthesis of SGAs based on aglycone structure were also suggested by Sanford et al. (1996). Similar to our results, they found high correlation in the accumulation of SGAs with the same aglycone, α-solanine-α-chaconine and leptine I-leptine II. The synthesis of SGAs followed some Mendelian segregation ratios, e.g., for the synthesis of α-solanine and α-chaconine, the F_2 population exhibited epistatic inheritance of two independent dominant genes with interaction of two or more genes for accumulation. On the other hand, synthesis of leptine I and leptine II followed a single gene Mendelian model in which a recessive gene was responsible for the synthesis of leptines. The variability in accumulation as well as the concentration of high levels of SGAs in a small proportion of the population that synthesized these compounds, suggest that accumulation is associated with interaction of several genes. Similar conclusions about a few dominant genes affecting biosynthesis of SGAs, recessive genes controlling leptines, as well as polygenetic control of accumulation of SGAs have been reached (Hutvágner et al. 2001; Ronning et al. 1999; Sagredo et al. 2006; Sanford et al. 1996; Van Dam et al. 1999; Yencho et al. 1998).

F_2 population genotyping analysis

Two of our four candidate genes segregated with accumulation of α-solanine and α-chaconine, and one with accumulation of leptine I and leptine II in our F_2 population. Genetic interaction between homozygous *chc* 80-1 alleles for *HMG2* and *SGT2* showed statistically significant association with the greatest levels of accumulation of total SGAs, α-solanine and α-chaconine. Besides this specific allelic combination, neither of the other allelic combinations explained genetic control of accumulation. Only the accumulation but not the synthesis of α-solanine and α-chaconine was explained by these genes. Therefore, *HMG2* and *SGT2* on pseudochromosomes 2 and 8 could be segregating with major genes that are associated with regulating the accumulation of these compounds;

since *HMG2* has been induced by wounding and pathogen inoculation (Choi et al. 1992). On the other hand, even though there was significant association between *HMG2 chc* 80-1 alleles in homozygous and heterozygous genotypes and the greatest levels of accumulation of leptine I and leptine II, the contingency table analysis showed that the presence of these alleles does not guarantee synthesis of these compounds. Only *HMG2 phu* homozygous alleles were strongly negatively associated with synthesis of leptines. Hence, *HMG2* is also associated with regulating the accumulation of leptine I and leptine II, but does not seem to play a major role in the synthesis of these compounds. Segregating populations with different genetic background have been used to elucidate genetic factors associated with SGA and leptine accumulation. Two loci on chromosomes 2 and 8 with complementary epistasis were associated with synthesis of leptinines and leptines, where the locus on chromosome 8 was directly responsible for leptine biosynthesis (Sagredo et al. 2006). *HMG2* is located on pseudochromosome 2 and *SGT2* on pseudochromosome 8; however the results of this study specifically showed association of these loci with accumulation of SGAs. Chromosome 1 and 8 could have genes that control synthesis of α-solanine-α-chaconine and leptines, respectively. Consistently, major loci on chromosome 1 have been associated with production of solanidine aglycone, total SGAs, α-solanine, α-chaconine and repression of leptines (Hutvágner et al. 2001; Ronning et al. 1999; Sørensen et al. 2008; Yencho et al. 1998). As an overall conclusion Sørensen et al. (2008) thought that the genes for biosynthesis of the solanidine aglycone from cholesterol occurred on this chromosome. The simpler genetics of our diploid population resulted in more easily understandable genotypic and phenotypic segregation patterns.

SNP chip analysis

The SNP chip contingency table analyses identified loci that accurately explained synthesis of different SGAs. For α-solanine and α-chaconine, the most common SGAs biosynthesized in potato species, several studies have reported that a few dominant genes direct their synthesis (Sørensen et al. 2008; Yencho et al. 1998). We also found that the phenotypic segregation of the F_2 population in this study could be explained by the epistatic interaction of two independent genes. In fact the SNP chip

analysis detected two loci strongly associated with the synthesis of α-solanine and α-chaconine. The enzymatic reactions catalyzed by putative candidate genes located near both loci suggest their role in the synthesis of α-solanine and α-chaconine. Our significant SNP found at 61 Mb on pseudochromosome 1 occurs between the two SSR markers (STM5136 at 53.3 Mb and STM2030 at 65.7 Mb; genetic distances were converted to physical distances by blast searching the primer sequences against the potato genome) defining the significant QTL mapped by Sørensen et al. (2008), which was thought to have a biosynthetic gene acting before the synthesis of solanidine aglycone. Sterol desaturase type enzymes are involved in the sterol biosynthesis in the post squalene segment of the pathway in *A. thaliana* (Benveniste 2004). For potato a candidate sterol desaturase on pseudochromosome 1 could play a role in the synthesis of cholesterol which is the main precursor of SGAs. In the same way, the candidate cytochrome P450 CYP71D7 on pseudochromosome 6, could catalyze the addition of hydroxyl groups in the transformation of cholesterol to solanidine. At least three major reactions take place in the conversion of solanidine: hydroxylation of cholesterol at C-26, oxidation at C-22 and amination at C-26 (Heftmann 1983; Kaneko et al. 1976; Kaneko et al. 1977). Cytochrome P450 CYP71D7 is a monooxygenase reductase, the kind of enzyme that incorporates hydroxyl groups.

Four loci on pseudochromosomes 1, 2, 7 and 8 were significantly associated with synthesis of leptine I and leptine II. A few recessive genes have been associated with the synthesis of leptines (Boluarte-Medina et al. 2002; Hutvágner et al. 2001; Ronning et al. 1999; Sagredo et al. 2006). In our study some loci were positively associated (pseudochromosome 1 and 7) and others negatively associated with synthesis of leptines (pseudochromosome 2 and 8), but could not explain the 1:3 phenotypic ratio found for leptine synthesis in this population. The function and importance to leptine synthesis of putative candidate genes identified in these loci were analyzed. The leptine significant SNP at 60.6 Mb of pseudochromosome 1 was nearby a sterol desaturase candidate gene that was also associated with synthesis of α-solanine and α-chaconine. CYP72A58, a cytochrome P450 with steroid hydroxylase activity was the candidate gene found on pseudochromosome 7. This enzyme could catalyze reactions in the conversion of cholesterol to leptines. Plant P450 monooxygenase protein family catalyzes a broad range

of chemical reactions involved in biosynthetic pathways of phenols, sterols, terpenes and flavonoids (Donaldson and Luster 1991). Specialized functions of these proteins have evolved to synthesize secondary metabolites related to plant defense compounds, antioxidants, phytoestrogens and compounds for detoxification of pollutants and herbicides (Rupasinghe and Schuler 2006), suggesting that cytochrome P450 enzymes could be involved in the unknown steps of SGA biosynthesis (Osman et al. 1987; Sagredo et al. 2011). Indeed, many CYP specialized cytochrome enzymes acting in plant sterol biosynthesis have been characterized (Ohnishi et al. 2009). The synthesis of phytosterols from cycloartenol precursor is carried out in three parallel pathways C8 (cholesterol), C9 (campesterol) and C10 (stigmasterol) (Benveniste 2004; Schaller 2004); where multiple enzymes are simultaneously involved (Arnqvist 2007). Regarding putative genes controlling leptine biosynthesis, *DWARF1/DIMINUTE* (*DWF1*), *HMG2*, *BRASSINAZOLE-RESISTANT 1 protein*, sterol-C5(6)-desaturase homolog, *SQE* and N-acetyltransferase were selected on pseudochromosome 2. In addition, *BRASSINOSTEROID INSENSITIVE 1-associated receptor kinase 1* and *SGT2* were proposed on pseudochromosome 8. Some of these genes have been involved with sterol and SGA biosynthesis and accumulation and others were selected because the reaction they perform occurs in the metabolic pathway. Expression analyses of wounded and light exposure potato plants identified eight strongly induced genes related to sterol and SGA accumulation: *HMG1*, squalene synthase 1 (*PSS1*), sterol-4α-methyl oxidase 1(*SMO1-1*), obtusifoliol 14α-demethylase (*CYP51*), $^{\Delta}$8-$^{\Delta}$7 sterol isomerase (*HYD1*), $^{\Delta}$24-reductase (*DWF1-L*), *SGT1* and rhamnosyltransferase (*SGT3*) (Nahar 2011). These genes catalyze reactions in the sterol biosynthetic pathway as well as in the SGA glycosylation. In addition, a phylogenetic analysis showed that two divergent *DWF1* genes were specific to solanaceous plants that produce SGAs (*DWF1* and *DWF1-L*). In the same way, the level of cholesterol and SGAs was reduced in antisense *StDWF1* and *StDWF1-L* potato plants. *DWARF1/DIMINUTE* and *HMG2* are tightly linked on pseudochromosome 2, future sequence and expression analyses of *chc* 80-1 could elucidate the specific role of *DWF1* in leptine biosynthesis since some aminoacid deletions were already identified in the cultivated potato by Nahar (2011). Brassinosteroids are part of the end-products of sterol metabolism, and many of the genes identified in the metabolic C9 pathway cycloartenol-

campesterol-brassinosteroids have an influence on metabolism of C10 sitosterol and sigmasterol as well as C8 cholesterol pathways. For this reason *BRASSINAZOLE-RESISTANT 1 protein* was postulated as a candidate gene on pseudochromosome 2 and *BRASSINOSTEROID INSENSITIVE 1-associated receptor kinase 1* on pseudochromosome 8.

Regarding SGA accumulation, 103 unique significant SNPs were found. These mapped to seven pseudochromosomes on the potato genome. Pseudochromosomes 6 and 11 held the most significant SNPs for α-solanine and 1, 6, 7 and 11 for leptine I and leptine II. Some putative candidate genes affecting biosynthesis or regulating the expression of SGAs located nearby those loci were proposed. Genes associated with primary metabolism, others from the cytochrome P450 family and some glycosyltransferases were found at different locations on pseudochromosome 6 and 11 for α-solanine and leptines. Association of SGA accumulation with multiple pseudochromosomes and loci within them agreed with the polygenic control that was proposed for this trait (Sagredo et al. 2006; Van Dam et al. 1999; Yencho et al. 1998). Interestingly significant SNPs associated with accumulation on pseudochromosome 1 and 7 for leptines were located on the same superscaffolds positively associated with synthesis of α-solanine-α-chaconine-leptines and leptines. Future studies will elucidate the genetic interactions between these two loci and the one on pseudochromosome 2 in the synthesis and accumulation of leptines. In the phenotypic analysis we found that synthesis of leptines was associated with synthesis of α-solanine and α-chaconine. The SNP chip analysis showed that pseudochromosome 1 is associated with syntheses of all SGAs, and pseudochrmosome 7 with biosynthesis of leptines.

Overall the results from cloning and sequencing of candidate genes and whole genome SNP analysis expose the genetic complexity of synthesis and accumulations of SGAs. The SNP chip analysis was able to identify two loci, on chromosome 1 and 6 with putative sterol biosynthetic genes that explained genotypic and phenotypic synthesis of α-solanine and α-chaconine. Both the candidate gene approach and the SNP chip associated allelic states of *HMG2* and *SGT2* and other putative candidate genes on pseudochromosome 2 and 8 with synthesis of leptines. In addition, the SNP chip also

identified two other loci positively associated with leptine synthesis on pseudochromosomes 1 and 7. None of the loci on pseudochromosomes 1, 2, 7 and 8 explained the 1:3 phenotypic segregation ratio of leptine synthesis. Future analysis of epistatic interactions and additive effects in a larger population could be used to understand synthesis of leptines. In the accumulation analysis, the detection of associated loci was more complex and suggested multiple loci and pseudochromosomes with no overlap between results from our two approaches. The greatest accumulation of α-solanine and α-chaconine was associated with interaction of *HMG2* and *SGT2* on pseudochromosomes 2 and 8 in the candidate gene approach; whereas the SNP chip analysis detected multiple loci on pseudochromosomes 6 and 11. For leptine accumulation, the candidate gene approach found *HMG2* on pseudochromosomes 2 and the SNP chip analysis loci on pseudochromosome 1, 6, 7 and 11. *Chc* 80-1 alleles were always associated with greatest levels of SGA accumulation on the different pseudochromosomes, and with synthesis for α-solanine and α-chaconine on pseudochromosomes 1 and for 6 and for leptines on pseudochromosome 1 and 7. Putative candidate genes that were associated with primary and secondary metabolism showed the high level of introgression of *S. chacoense* in the F_2 population. The candidate gene approach and whole genome SNP genotyping of this study contributed to consolidate results from several studies, and postulated putative candidate genes for future studies in the elucidations of genetic factor associated with synthesis and accumulation of SGAs.

Acknowledgements

This research was supported by Research Grant No. IS-4134-08 R from BARD, The United States – Israel Binational Agricultural Research and Development Fund. We thank Suzanne Piovano for technical assistance, Maichel Miguel Aguayo Bustos for programming help, Anne Ryan of the Laboratory for Interdisciplinary Statistical Analysis LISA for statistical consultation.

References

Arnqvist L (2007) Plant sterol metabolism with emphasis on glycoalkaloid biosynthesis in potato. Dept of Plant Biology and Forest Genetics. Swedish University of Agricultural Sciences, Uppsala, p 50

Bejarano L, Mignolet E, Devaux A, Espinola N, Carrasco E, Larondelle Y (2000) Glycoalkaloids in potato tubers: the effect of variety and drought stress on the α-solanine and α-chaconine contents of potatoes. J Sci Food Agric 80:2096-2100

Benveniste P (2004) Biosynthesis and accumulation of sterols. Annu Rev Plant Biol 55:429-457

Boluarte-Medina T, Fogelman E, Chani E, Miller AR, Levin I, Levy D, Veilleux RE (2002) Identification of molecular markers associated with leptine in reciprocal backcross families of diploid potato. Theor Appl Genet 105:1010-1018

Choi D, Bostock RM, Avdiushko S, Hildebrand DF (1994) Lipid-derived signals that discriminate wound- and pathogen-responsive isoprenoid pathways in plants: methyl jasmonate and the fungal elicitor arachidonic acid induce different 3-hydroxy-3-methylglutaryl-coenzyme A reductase genes and antimicrobial isoprenoids in *Solanum tuberosum* L. Proc Natl Acad Sci USA 91:2329-2333

Choi D, Ward BL, Bostock RM (1992) Differential induction and suppression of potato 3-hydroxy-3-methylglutaryl coenzyme A reductase genes in response to *Phytophthora infestans* and to its elicitor arachidonic acid. Plant Cell 4:1333-1344

Dale MFB, Griffiths DW, Bain H, Todd D (1993) Glycoalkaloid increase in *Solanum tuberosum* on exposure to light. Ann Appl Biol 123:411-418

Distl M, Wink M (2009) Identification and quantification of steroidal alkaloids from wild tuber-bearing *Solanum* species by HPLC and LC-ESI-MS. Potato Res 52:79-104

Donaldson RP, Luster DG (1991) Multiple forms of plant cytochromes P-450. Plant Physiol 96:669-674

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19:11-15

Edwards EJ, Cobb AH (1996) Improved high performance liquid chromatographic method for the analysis of potato (*Solanum tuberosum*) glycoalkaloids. J Agr Food Chem 44:2705-2709

FAO (2006) Buried treasure: the potato. International year of the potato 2008. http://www.fao.org/ag/magazine/pdf/0611-1.pdf

FAO (2008a) The global potato economy. International year of the potato 2008. http://www.potato2008.org/en/potato/economy.html

FAO (2008b) Potato pest and disease management. International Year of the Potato 2008. http://www.potato2008.org/en/potato/pests.html

Felcher KJ, Coombs JJ, Massa AN, Hansey CN, Hamilton JP, Veilleux RE, Buell CR, Douches DS (2012) Integration of two diploid potato linkage maps with the potato genome sequence. PLoS One 7:e36347

Friedman M (2006) Potato glycoalkaloids and metabolites: roles in the plant and in the diet. J Agric Food Chem 54:8655-8681

Friedman M, Dao L (1992) Distribution of glycoalkaloids in potato plants and commercial potato products. J Agr Food Chem 40:419-423

Friedman M, McDonald GM (1997) Potato glycoalkaloids: chemistry, analysis, safety, and plant physiology. Crit Rev Plant Sci 16:55-132

Ginzberg I, Tokuhisa JG, Veilleux RE (2009) Potato steroidal glycoalkaloids: biosynthesis and genetic manipulation. Potato Res 52:1-15

Hamilton JP, Hansey CN, Whitty BR, Stoffel K, Massa AN, Van Deynze A, De Jong WS, Douches DS, Buell CR (2011) Single nucleotide polymorphism discovery in elite North American potato germplasm. BMC Genomics 12:302

Haynes FL (1972) The use of cultivated diploid *Solanum* species in potato breeding. *In*: Prospects for the potato in the developing world. International Potato Center (CIP), Lima, Peru

Heftmann E (1983) Biogenesis of steroids in Solanaceae. Phytochemistry 22:1843-1860

Holland JB (2007) Genetic architecture of complex traits in plants. Curr Opin Plant Biol 10:156-161

Hutvágner G, Bánfalvi Z, Milánkovics I, Silhavy D, Polgár Z, Horváth S, Wolters P, Nap JP (2001) Molecular markers associated with leptinine production are located on chromosome 1 in *Solanum chacoense*. Theor Appl Genet 102:1065-1071

Johnson AAT, Piovano SM, Ravichandran V, Veilleux RE (2001) Selection of monoploids for protoplast fusion and generation of intermonoploid somatic hybrids of potato. Am J Potato Res 78:19-29

Judo MS, Wedel AB, Wilson C (1998) Stimulation and suppression of PCR-mediated recombination. Nucleic Acids Res 26:1819-1825

Kaneko K, Tanaka MW, Mitsuhashi H (1976) Origin of nitrogen in the biosynthesis of solanidine by *Veratrum grandiflorum*. Phytochemistry 15:1391-1393

Kaneko K, Tanaka MW, Mitsuhashi H (1977) Dormantinol, a possible precursor in solanidine biosynthesis, from budding *Veratrum grandiflorum*. Phytochemistry 16:1247-1251

Kolbe H, Stephan-Beckmann S (1997) Development, growth and chemical composition of the potato crop (*Solanum tuberosum* L.). II. Tuber and whole plant. Potato Res 40:135-153

Krits P, Fogelman E, Ginzberg I (2007) Potato steroidal glycoalkaloid levels and the expression of key isoprenoid metabolic genes. Planta 227:143-150

Lafta AM, Lorenzen JH (2000) Influence of high temperature and reduced irradiance on glycoalkaloid levels in potato leaves. J Am Soc Hort Sci 125:563-566

Laurila J, Laakso I, Valkonen JPT, Hiltunen R, Pehu E (1996) Formation of parental-type and novel glycoalkaloids in somatic hybrids between *Solanum brevidens* and *S. tuberosum*. Plant Sci 118:145-155

Lenz TL, Becker S (2008) Simple approach to reduce PCR artefact formation leads to reliable genotyping of MHC and other highly polymorphic loci--implications for evolutionary analysis. Gene 427:117-123

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451-1452

Lightbourn GJ, Veilleux RE (2007) Production and evaluation of somatic hybrids derived from monoploid potato. Am J Potato Res 84:425-435

Liu YG, Chen Y (2007) High-efficiency thermal asymmetric interlaced PCR for amplification of unknown flanking sequences. BioTechniques 43:649-656

McCue KF, Allen PV, Shepherd LV, Blake A, Maccree MM, Rockhold DR, Novy RG, Stewart D, Davies HV, Belknap WR (2007a) Potato glycosterol rhamnosyltransferase, the terminal step in triose side-chain biosynthesis. Phytochemistry 68:327-334

McCue KF, Allen PV, Shepherd LVT, Blake A, Rockhold DR, Novy RG, Stewart D, Davies HV, Belknap WR (2007b) Manipulation and compensation of steroidal glycoalkaloid biosynthesis in potatoes. Acta Hort 745:343-349

McCue KF, Shepherd LVT, Allen PV, Maccree MM, Rockhold DR, Corsini DL, Davies HV, Belknap WR (2005) Metabolic compensation of steroidal glycoalkaloid biosynthesis in transgenic potato tubers: using reverse genetics to confirm the in vivo enzyme function of a steroidal alkaloid galactosyltransferase. Plant Sci 168:267-273

Moehs CP, Allen PV, Friedman M, Belknap WR (1997) Cloning and expression of solanidine UDP-glucose glucosyltransferase from potato. Plant J 11:227-236

Morris SC, Lee TH (1984) The toxicity and teratogenicity of Solanaceae glycoalkaloids, particularly those of the potato (*Solanum tuberosum*): a review. Food Technol Aust 36:118-124

Mulichak AM, Lu W, Losey HC, Walsh CT, Garavito RM (2004) Crystal structure of vancosaminyltransferase GtfD from the vancomycin biosynthetic pathway: Interactions with acceptor and nucleotide Ligands. Biochemistry 43:5170-5180

Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res 8:4321-4325

Mweetwa AM, Hunter D, Poe R, Harich KC, Ginzberg I, Veilleux RE, Tokuhisa JG (2012) Steroidal glycoalkaloids in *Solanum chacoense*. Phytochemistry 75:32-40

Mweetwa MA (2009) Biosynthesis of steroidal glycoalkaloids in *Solanum chacoense* Bitter. Virginia Polytechnic Institute and State University, Blacksburg, p 118

Nahar N (2011) Regulation of sterol and glycoalkaloid biosynthesis in potato (*Solanum tuberosum* L.) – Identification of key genes and enzymatic steps. Department of Plant Biology and Forest Genetics. Swedish University of Agricultural Sciences, Uppsala, p 66

Nema PK, Ramayya N, Duncan E, Niranjan K (2008) Potato glycoalkaloids: formation and strategies for mitigation. J Sci Food Agric 88:1869-1881

Ohnishi T, Yokota T, Mizutani M (2009) Insights into the function and evolution of P450s in plant steroid metabolism. Phytochemistry 70:1918-1929

Ohyama K, Suzuki M, Kikuchi J, Saito K, Muranaka T (2009) Dual biosynthetic pathways to phytosterol via cycloartenol and lanosterol in *Arabidopsis*. Proc Natl Acad Sci USA 106:725-730

Osman S, Sinden SL, Deahl K, Moreau R (1987) The metabolism of solanidine by microsomal fractions from *Solanum chacoense*. Phytochemistry 26:3163-3165

Pflieger S, Lefebvre V, Causse M (2001) The candidate gene approach in plant genetics: a review. Mol Breeding 7:275-291

Pichersky E, Gang DR (2000) Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. Trends Plant Sci 5:439-445

Ramkumar G, Sivaranjani AKP, Pandey MK, Sakthivel K, Rani NS, Sudarshan I, Prasad GSV, Neeraja CN, Sundaram RM, Viraktamath BC, Madhav MS (2010) Development of a PCR-based SNP marker system for effective selection of kernel length and kernel elongation in rice. Mol Breeding 26:735-740

Ronning CM, Stommel JR, Kowalski SP, Sanford LL, Kobayashi RS, Pineada O (1999) Identification of molecular markers associated with leptine production in a population of *Solanum chacoense* Bitter. Theor Appl Genet 98:39-46

Rupasinghe S, Schuler MA (2006) Homology modeling of plant cytochrome P450s. Phytochem Rev 5:473-505

Sagredo B, Lafta A, Casper H, Lorenzen J (2006) Mapping of genes associated with leptine content of tetraploid potato. Theor Appl Genet 114:131-142

Sagredo B, Lorenzen J, Casper H, Lafta A (2011) Linkage analysis of a rare alkaloid present in a tetraploid potato with *Solanum chacoense* background. Theor Appl Genet 122:471-478

Sanford L, Kobayashi R, Deahl K, Sinden S (1996) Segregation of leptines and other glycoalkaloids in *Solanum tuberosum* (4x) × *Solanum chacoense* (4x) crosses. Am J Potato Res 73:21-33

Schaller H (2004) New aspects of sterol biosynthesis in growth and development of higher plants. Plant Physiol Biochem 42:465-476

Sinden SL, Sanford LL, Cantelo WW, Deahl KL (1986) Leptine glycoalkaloids and resistance to the Colorado potato beetle (Coleoptera: Chrysomelidae) in *Solanum chacoense*. Environ Entomol 15:1057-1062

Sinden SL, Sanford LL, Osman SF (1980) Glycoalkaloids and resistance to the Colorado potato beetle in *Solanum chacoense* Bitter. Am Potato J 57:331-343

Sørensen K, Kirk H, Olsson K, Labouriau R, Christiansen J (2008) A major QTL and an SSR marker associated with glycoalkaloid content in potato tubers from *Solanum tuberosum* × *S. sparsipilum* located on chromosome I. Theor Appl Genet 117:1-9

Stewart CN, Via LE (1993) A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR applications. BioTechniques 14:748-750

Strickberger MW (1968) Genetics. Macmillan, New York

Suzuki M, Muranaka T (2007) Molecular genetics of plant sterol backbone synthesis. Lipids 42:47-54

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731-2739

The Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. Nature 475:189-195

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680

Thompson JR, Marcelino LA, Polz MF (2002) Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'. Nucleic Acids Res 30:2083-2088

Todd AE, Orengo CA, Thornton JM (1999) Evolution of protein function, from a structural perspective. Curr Opin Chem Biol 3:548-556

Väänänen T, Ikonen T, Rokka VM, Kuronen P, Serimaa R, Ollilainen V (2005) Influence of incorporated wild *Solanum* genomes on potato properties in terms of starch nanostructure and glycoalkaloid content. J Agr Food Chem 53:5313-5325

Valkonen JPT, Keskitalo M, Vasara T, Pietilä L (1996) Potato glycoalkaloids: a burden or a blessing? Crit Rev Plant Sci 15:1-20

Van Dam J, Levin I, Struik PC, Levy D (1999) Genetic characterisation of tetraploid potato (*Solanum tuberosum* L.) emphasising genetic control of total glycoalkaloid content in the tubers. Euphytica 110:67-76

Van Dam J, Levin I, Struik PC, Levy D (2003) Identification of epistatic interaction affecting glycoalkaloid content in tubers of tetraploid potato (*Solanum tuberosum* L.). Euphytica 134:353-360

Veilleux RE (1990) *Solanum phureja*: anther culture and the induction of haploids in a cultivated diploid potato species., Y. P. S. Bajaj edn. Sringer-Verlag, New York

Waxman D, Peck JR (1998) Pleiotropy and the preservation of perfection. Science 279:1210-1213

Wentzinger LF, Bach TJ, Hartmann MA (2002) Inhibition of squalene synthase and squalene epoxidase in tobacco cells triggers an up-regulation of 3-hydroxy-3-methylglutaryl coenzyme a reductase. Plant Physiol 130:334-346

Yencho GC, Kowalski SP, Kobayashi RS, Sinden SL, Bonierbale MW, Deahl KL (1998) QTL mapping of foliar glycoalkaloid aglycones in *Solanum tuberosum* x *S. berthaultii* potato progenies: quantitative variation and plant secondary metabolism. Theor Appl Genet 97:563-574

Yoshioka H, Yamada N, Doke N (1999) cDNA cloning of sesquiterpene cyclase and squalene synthase, and expression of the genes in potato tuber infected with *Phytophthora infestans*. Plant Cell Physiol 40:993-998

Tables and figures

Table 1 GenBank accessions used for primer design, and primer sequences used to amplify fragments of four candidate genes in the potato glycoalkaloid biosynthetic pathway

Gene	Linkage Group	PGSC Gene Models[1]	GenBank Accessions	Primer Sequence	Product Size	Tm (°C)
HMG2	2	PGSC0003DMG400003461	M63642, AF110383, AB041031, U51985-6, L01400, AF110383	f- TGGTGTCCAAAGGTGTACAAAATG r- ACAGAAATATGGAGGTCCTTGCC	608 bp	59
SQE	4	PGSC0003DMG400004923	AY995182, BG123494, CU915722	f- TGGGGTTCGTTGCAGTTTTC r- CAGGGGATAAGAAAGACGTGTACTC	884 bp	60
SGT1	7	PGSC0003DMG400011749	U82367, DQ218276, DQ218277, DQ266437, AK323113, AB182385	f- TCCCTTGGACAGTAGATATTGCTG r- TTCCCAATCCCCTAACCTCG	834 bp	58
SGT2	8	PGSC0003DMG400017508	U82367, DQ218276, DQ218277, DQ266437, AK323113, AB182385	f- CCTGCGGATGAGAGGAATGC r- CACCAACGGCACCCCAGCG	567 bp	64

Primer direction: forward (f) and reverse (r). Melting temperature of primers = Tm. Linkage groups correspond with the location of these gene sequences in the potato genome.[1] Published gene models from Potato Genome Sequencing Consortium (PGSC) sequence.

Table 2 Allele-specific primers for SNP, intron and indel genotyping of four candidate genes in a diploid F_2 population

Gene	Primer Sequence 5' to 3'	Allele: Product Size	Tm (°C)
HMG2	Intron f- GGATGTCATCGGCATATCTGGTA Intron r- TTGTCCGAGCAAAAGTTCCCTGT	C1: 124 bp P1: 167 bp P2: 124 bp	68
SQE	C1 f- GGGACTACAGGAAGGGCGTCGTGTA**A** P1 r- GGCTCTGTCAAATCTCTTTCAATTACTT**G** C2 TaqMan® assay	C1: 355 bp P1: 583bp C2:	58
SGT1	C1 r- CTAAATTACCAACTGGCAACCACGAC**ATT** C2 f- TACAGCATTATGCACAACCCTAAC**CA** C3 Promoter f- **C**T**CGAAGG**A**G**T**GAA**T**CAAAAAGAGAAG** C3 r- C**A**CAGAAATTAGTTCCTTACTACGGAT P1 r- CGTTCCTCCGAATCTC**GT**A**C**TT**TA** P2 f- TTGCTCGAACAA**A**TTGGAGATTC P2 r- TAGTCTTGTCCTCTAAATTACCAAC**G**G	C1:627 bp C2: 760 bp C3:1203 bp P1: 274 bp P2:393 bp	63 58 59 58 66
SGT2	C1 indel f- GTGAACAAGGACCAATC**AGCAAAAA** P1 f- TGCCTACGCTGA**C**TACTATCA**A**AAGATG	C1: 189 bp P1: 454 bp	67

The *chc* 80-1 and *phu* DH alleles are defined as C or P with its arbitrary allelic numbers and the direction of each primer as forward (f) and reverse (r). Polymorphic sites in each primer are bold. Melting temperature of primers = Tm.

Table 3 Summary statistics of analyzed allelic sequence sites (excluding sites with gaps on the total alignment), and sequence polymorphism found in four candidate genes of the glycoalkaloid biosynthetic pathway

Gene	N	Total Sites (bp)	Total SNP Sites (S)	Total Number of Mutations	SNPs in Exonic Regions	SNPs in Intronic Regions	Silent[1] SNPs	Non-Synonymous SNPs	Indels (Location)
HMG2	3	490	26	26	10	16	23	3	2 (Intron)
SQE	3	884	22	22	3	19	21	1	1 (Intron)
SGT1	5	815	49	51	51	0	14	37	2 (Exon)
SGT2	2	558	9	8	8	0	3	5	1 (Exon)
Average	3,3	686.8	26.5	27	18.0	8.8	15.3	11.5	1.5
Total	13	2,747	106	108	72	35	61	46	6

[1] Synonymous and noncoding SNPs. N= number of allelic sequences.

Table 4 Mean, range, standard deviation (SD), quartiles, and covariance (CV) of steroidal glycoalkaloid estimations in a diploid F_2 potato population *phu* DH × *chc* 80-1

	Total SGA (µg/mg DW)	**α-Solanine (µg/mg DW)**	**α-Chaconine (µg/mg DW)**	**Leptine I (µg/mg DW)**	**Leptine II (µg/mg DW)**	**Sol/Chac Ratio**	**Leptine II/I Ratio**
Parents							
phu DH	0	0	0	0	0		
chc 80-1	190.7	10.5	8.8	113.0	58.4	1.2	0.5
F_2 Progeny						*n*=138	*n*=32
Mean	25.6	7.5	14.4	2.8	0.9	0.6	0.3
SD	25.8	8.0	14.0	8.2	2.7	0.3	0.1
Range	0-126.9	0-46.4	0-84.9	0-44.7	0-18.5	0.2-3.8	0.1-0.5
0.1-0.75 quartiles	2-37.8	0.8-10.0	1.4-19.9	0-0.7	0-0	0.3-0.6	0.2-0.4
CV	101	106	97	291	314	59	30

The amount of accumulation of SGAs in µg of compound per mg of dried weight leaf tissue (DW). F_2 progeny sample size = 148 for all calculations. Ratios were analyzed in subsamples (*n*) with presence of those compounds.

Table 5 Chi-square (χ^2) test of phenotypic segregation ratios for synthesis of response variables in a diploid F_2 population *phu* DH × *chc* 80-1

Variable	Presence	Absence	Total	Ratio	χ^2 P-value
Total SGA	140	8	148	15:1	0.664
α-Solanine	138	10	148	15:1	0.801
α-Chaconine	140	8	148	15:1	0.664
Leptine I	38	110	148	1:3	0.850
Leptine II	32	116	148	1:3	0.335

Table 6 Chi-square (χ^2) test of genotypic segregation ratios of four candidate genes in a F_2 population *phu* DH × *chc* 80-1 and homogeneity test

	CC	CP	PP	Total	Ratio	χ^2 P-value
Gene						
HMG2	38	70	39	147	1:2:1	0.841
SQE	45	75	28	148	1:2:1	0.134
SGT1	48	71	29	148	1:2:1	0.083
SGT2	31	87	30	148	1:2:1	0.100
Family 3 (6×5 = L×H)						
HMG2	8	15	5	28	1:2:1	0.656
SQE	6	20	3	29	1:2:1	0.071
SGT1	6	16	7	29	1:2:1	0.824
SGT2	5	17	7	29	1:2:1	0.549
Family 5 (2×1 = L×L)						
HMG2	8	12	6	26	1:2:1	0.802
SQE	13	8	5	26	1:2:1	0.022*
SGT1	9	10	7	26	1:2:1	0.439
SGT2	4	17	5	26	1:2:1	0.271
Family 7 (3×5 = I×H)						
HMG2	4	13	6	23	1:2:1	0.672
SQE	8	10	5	23	1:2:1	0.580
SGT1	8	13	2	23	1:2:1	0.1196
SGT2	3	14	3	23	1:2:1	0.347
Family 8 (5×2 = H×L)						
HMG2	8	16	14	38	1:2:1	0.271
SQE	9	21	8	38	1:2:1	0.786
SGT1	13	19	6	38	1:2:1	0.267
SGT2	9	21	8	38	1:2:1	0.786
Family 9 (2×5 = L×H)						
HMG2	10	14	8	32	1:2:1	0.696
SQE	9	16	7	32	1:2:1	0.882
SGT1	12	13	7	32	1:2:1	0.292
SGT2	10	18	4	32	1:2:1	0.206
Test of Homogeneity						
	Σ Fam χ^2	**Hom χ2**	**Hom DF**	**Probability**		
HMG2	3.098	2.258	8	>0.95		
SQE	2.341	2.207	8	>0.95		
SGT1	1.941	1.858	8	>0.95		
SGT2	2.160	2.060	8	>0.95		

The *chc* 80-1 and *phu* DH alleles are defined as C or P in the diploid allelic structure per gene. Fam = Family and Hom = Homogeneity

Table 7 Mean SGAs for groups of individuals within an F_2 population derived from a cross between *phu* DH × *chc* 80-1 having the same genetic structure

Gene	Total SGA ± SE (µg/mg DW)	α-Solanine ± SE (µg/mg DW)	α-Chaconine ± SE (µg/mg DW)	Leptine I ± SE (µg/mg DW)	Leptine II ± SE (µg/mg DW)
HMG2					
CC	28.5 ± 4.6	7.5 ± 1.2	14.6 ± 2.6	4.9 a ± 1.6	1.6 a ± 0.6
CP	26.1 ± 3.3	7.9 ± 1.1	14.0 ± 1.6	3.3 a ± 1.1	1.0 ab ± 0.3
PP	21.9 ± 3.1	7.0 ± 6.7	14.9 ± 2.0	0.0 b ± 0.0	0.0 b ± 0.0
P-value	0.526	0.875	0.952	0.026*	0.031*
SQE					
CC	29.9 ± 4.4	8.8 ± 1.4	15.6 ± 2.0	4.2 ± 1.6	1.21 ± 0.5
CP	23.0 ± 2.8	6.9 ± 0.8	13.6 ± 1.7	1.9 ± 0.7	0.6 ± 0.2
PP	25.8 ± 4.5	7.3 ± 1.5	14.5 ± 2.6	3.0 ± 1.6	1.0 ± 0.6
P-value	0.369	0.438	0.740	0.318	0.526
SGT1					
CC	27.5 ± 4.2	7.3 ± 1.2	15.5 ± 2.5	3.5 ± 1.3	1.1 ± 0.5
CP	22.8 ± 2.6	6.8 ± 0.8	12.8 ± 1.4	2.5 ± 0.9	0.7 ± 0.2
PP	29.3 ± 5.1	9.7 ± 1.8	16.3 ± 2.6	2.5 ± 1.6	0.9 ± 0.5
P-value	0.439	0.273	0.408	0.772	0.750
SGT2					
CC	35.1 a ± 6.3	10.4 ± 1.8	18.5 ± 3.5	4.6 ± 2.0	1.6 ± 0.8
CP	24.3 b ± 2.6	7.1 ± 0.8	13.4 ± 1.4	3.0 ± 0.9	0.9 ± 0.3
PP	19.7 b ± 2.5	6.1 ± 0.9	13.1 ± 1.6	0.4 ± 0.4	0.2 ± 0.2
P-value	0.049*	0.070	0.184	0.116	0.120

The amount of accumulation of SGAs in µg of compound per mg of dried weight leaf tissue (DW) followed by standard error of mean (±SE). *Means followed by the same latter are not significantly different at the 0.05 α level using student t mean separation analysis. The *chc* 80-1 and *phu* DH alleles are defined as C or P in the diploid genetic structure for *HMG2*, *SQE*, *SGT1* and *SGT2* genes. Each F_2 plant was characterized as to whether it had C, P or both alleles at each of four loci based on genetic polymorphism between C and P alleles. The numbers of individuals per mean are in Table 6.

Table 8 Effect of *HMG2* and *SGT2* genetic structure interaction on the accumulation of total SGAs, α-solanine and α-chaconine in a segregating F_2 population

HMG2, SGT2	**Total SGA ± SE (µg/mg DW)**		*HMG2, SGT2*	**α-Solanine ± SE (µg/mg DW)**		*HMG2, SGT2*	**α-Chaconine ± SE (µg/mg DW)**
CC,CC	63.1 a ± 9.3		CC,CC	16.8 a ± 2.9		CC,CC	32.0 a ± 5.1
CP,CP	29.2 b ± 4.1		PP,PP	9.2 b ± 2.4		PP,PP	18.3 b ± 4.3
CP,CC	28.6 bc ± 5.5		CP,CC	9.1 b ± 1.7		CP,CP	15.1 b ± 2.2
PP,PP	27.6 bc ± 7.8		CP,CP	8.6 b ± 1.3		CP,CC	14.8 b ± 3.0
CC,CP	20.9 bc ± 5.0		PP,CP	6.4 b ± 1.5		PP,CP	13.8 b ±2.7
PP,CP	20.2 bc ± 4.9		CC,PP	5.9 b ± 2.9		PP,CC	13.2 b ± 6.8
CC,PP	20.1 bc ± 9.3		PP,CC	5.5 b ± 3.9		CC,PP	12.0 b ± 5.1
PP,CC	18.6 bc ± 12.3		CC,CP	5.2 b ± 1.6		CC,CP	10.2 b ± 2.8
CP,PP	13.5 c ± 6.8		CP,PP	3.7 b ± 2.1		CP,PP	9.8 b ± 3.8
P-value	0.005*		P-value	0.015*		P-value	0.015*

The amount of accumulation of SGAs in µg of compound per mg of dried weight leaf tissue (DW) followed by standard error of mean (±SE). *Means followed by same latter are not significantly different at the 0.05 α level using student t mean separation analysis. The *chc* 80-1 and *phu* DH alleles are defined as C or P in the diploid genetic structure for *HMG2* and *SGT2* genes in the F_2 population.

Table 9 Effect of family pedigree on the accumulation of total SGAs, α-solanine, α-chaconine, leptine I and leptine II in a segregating F_2 population

Family	**Total SGA ± SE (µg/mg DW)**		**α-Solanine ± SE (µg/mg DW)**		**α-Chaconine ± SE (µg/mg DW)**	**Leptine I ± SE (µg/mg DW)**		**Leptine II ± SE (µg/mg DW)**
Family 3 (6×5 = L×H)	19.0 b ± 2.9		5.9 b ± 0.8		10.5 bc ± 1.3	1.7 b ± 1.3		0.8 b ± 0.6
Family 5 (2×1 = L×L)	14.2 b ± 2.6		5.2 b ± 1.0		8.5 c ± 1.4	0.4 b ± 0.3		0.1 b ± 0.1
Family 7 (3×5 = I×H)	34.1 a ± 6.6		7.0 ab ± 1.3		14.1 abc ± 2.4	1.0 a ± 3.3		2.9 a ± 1.0
Family 8 (5×2 = H×L)	32.4 a ± 5.2		10.4 a ± 1.8		20.5 a ± 3.2	1.2 b ± 0.6		0.3 b ± 0.1
Family 9 (2×5 = L×H)	26.8 ab ± 4.4		7.9 ab ± 1.4		15.5 ab ± 2.4	2.5 b ± 0.9		0.7 b ± 0.2
P-value	0.015*		0.072		0.005*	<.0001*		0.002*

The amount of accumulation of SGAs in µg of compound per mg of dried weight leaf tissue (DW) followed by standard error of mean (±SE). *Means followed by same latter are not significantly different at the 0.05 α level using student t mean separation analysis. The F_2 families come from specific crosses between F_1 plants with contrasting or similar levels of SGAs high (H), intermedium (I) or low (L). The ID number of F_1 plants used by cross and their level of SGAs in parenthesis.

Table 10 Number of significant SNPs identified at different pseudochromosomes in the potato genome associated with α-solanine, α-chaconine, leptine I and leptine II accumulation in a segregating F_2 population

Pseudomolecule	α-Solanine	α-Chaconine	Leptine I	Leptine II	Total	No of Unique SNPs
chr 01	0	1	1	1	3	2
chr 02	0	0	2	1	3	3
chr 06	15	3	17	10	45	32
chr 07	1	0	2	2	5	3
chr 08	3	0	3	3	9	6
chr 10	3	0	0	0	3	3
chr 11	11	0	23	45	79	54
Total	33	4	48	62	147	103

For α-solanine most of the significant SNPs were located on pseudochromosomes (chr) 6 and 11, and leptine I and leptine II shared 39 significant SNPs on Chr 6, 7, 8 and 11.

Table 11 SNPs from candidate genes in the 8303 Infinium array associated with leptine accumulation

	CC	CP	PP	P-value
SQE				
Leptine I (µg/mg DW) ± SE	12.9a ± 4.3	8.0 ab ± 3.2	0.0 b ± 0.0	0.030*
Leptine II (µg/mg DW) ± SE	4.4 a ± 1.8	2.6 ab ± 1.0	0.0 b ± 0.0	0.048*
HMG2				
Leptine I (µg/mg DW) ± SE	14.0 a ± 4.5	8.1 ab ± 3.2	0.0 b ± 0.0	0.016*
Leptine II (µg/mg DW) ± SE	4.8 a ± 1.9	2.6 ab ± 1.0	0.0 b ± 0.0	0.027*

The amount of accumulation of SGAs in µg of compound per mg of dried weight leaf tissue (DW) followed by standard error of mean (±SE). *Means followed by same latter are not significantly different at the 0.05 α level using student t mean separation analysis. ANOVA analyses for solcap_snp_c1_9695, 9696, c2_32413 and 32417 SNPs from *SQE* and solcap_snp_c1_15974 from *HMG2*.

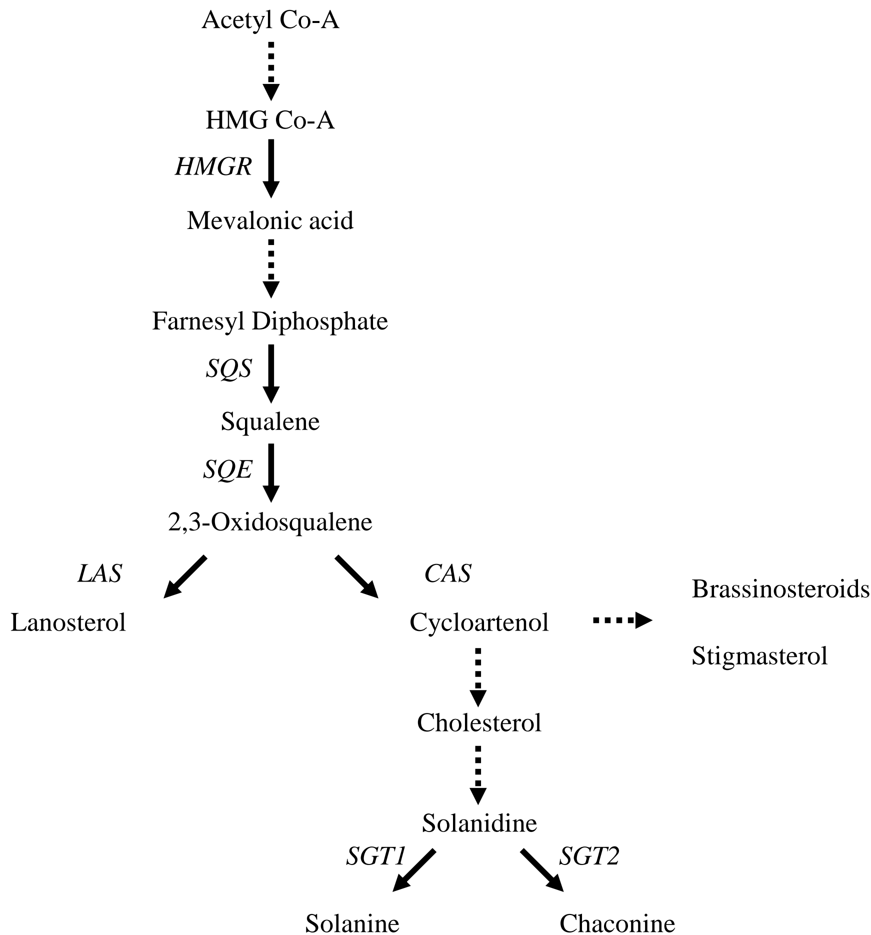


Figure 1 Potato steroidal glycoalkaloid biosynthetic pathway. Dashed line arrows indicate step with multiple enzymatic reactions. Continues line arrows have beside the abbreviation of gene performing the reaction. *HMGR* 3-hydroxy-3-methylglutaryl coenzyme A reductase, *SQS* squalene synthase, *SQE* squalene epoxidase, *LAS* lanosterol synthase and *CAS* cycloartenol synthase, *SGT1* solanidine galactosyltransferase and *SGT2* solanidine glucosyltransferase. Brassinosteroids and stigmaterol are the other end products of sterol biosynthesis in addition to cholesterol.

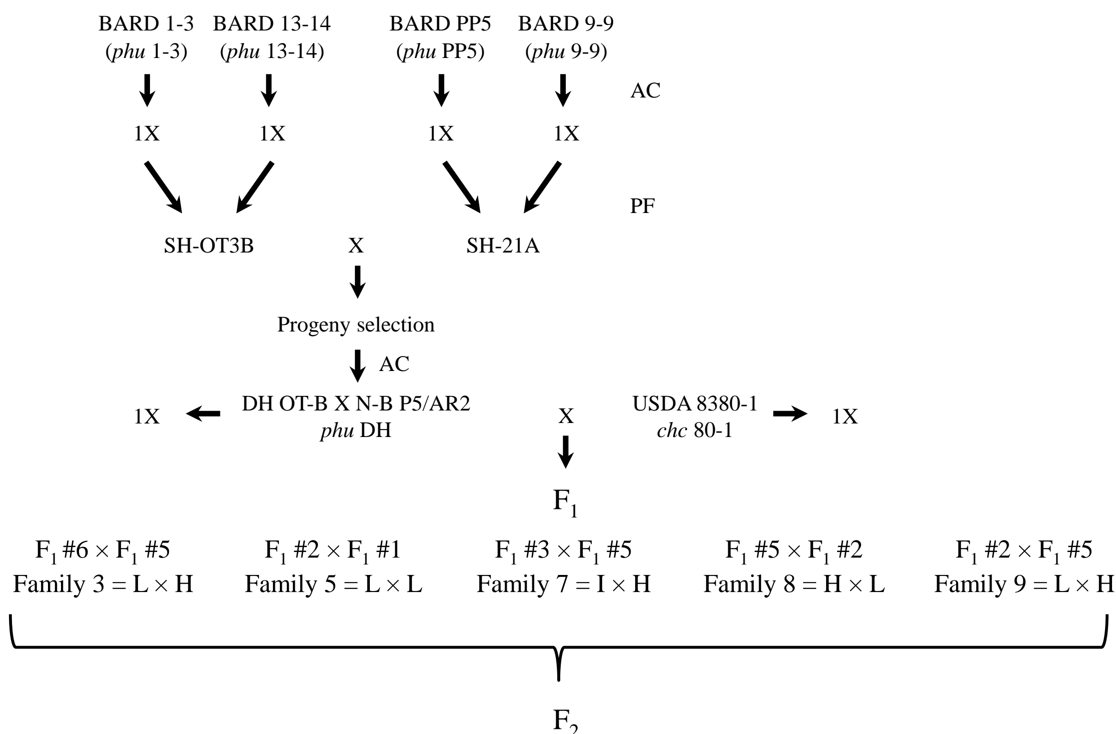
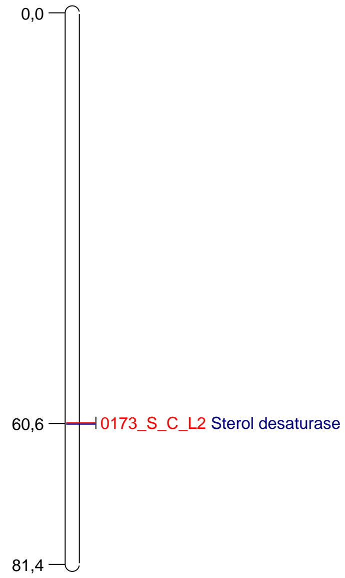
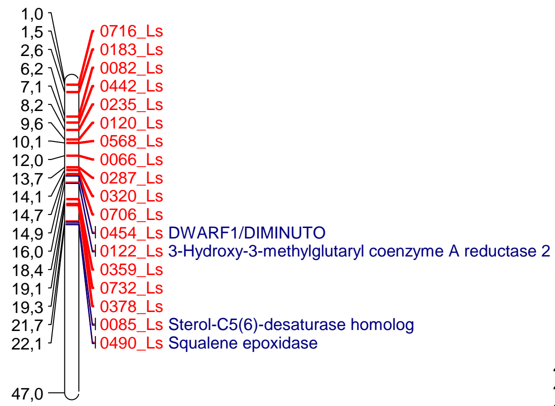
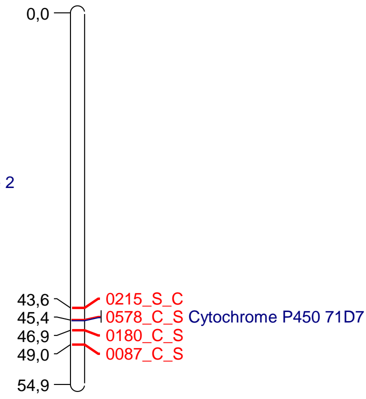
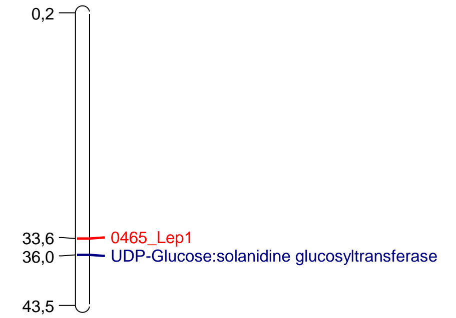


Figure 2 Pedigree scheme of *phu* DH clone and segregating F_2 population of *phu* DH × *chc* 80-1. Four heterozygous *S. phureja* BARD clones, selected for tuberization under long photoperiod, were used to generate monoploid lines (1X) by anther culture (AC). Then intermonoploid somatic hybrids were obtained by protoplast fusion (PF). Tetraploid somatic hybrids (SH-OT3B and SH-21A) generated for each fusion were crossed under greenhouse conditions. *Phu* DH is one of the di-haploids obtained by anther culture (AC) from the progeny of a tetraploid somatic hybrid cross. In this study F_1 plants from a cross between *phu* DH × *chc* 80-1 clones were used to generate an F_2 segregating population. Five families from crosses between F_1 plants with similar or contrasting SGA levels, low (L), intermediate (I) and high (H) formed the F_2 population. Monoploids (1X) from each parental line were used to sequence analysis.

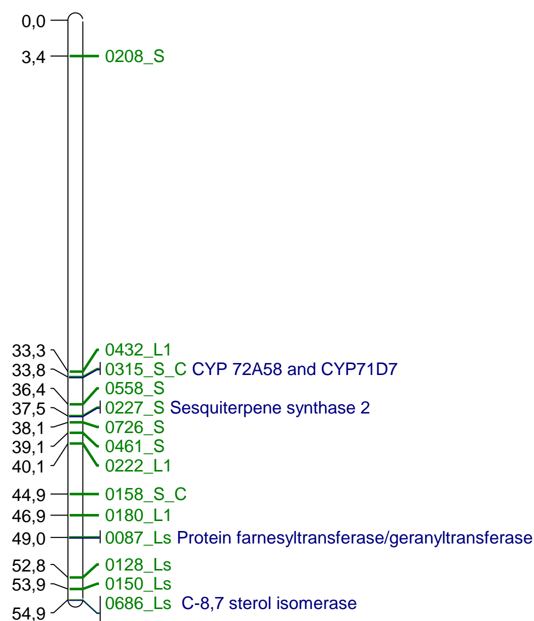


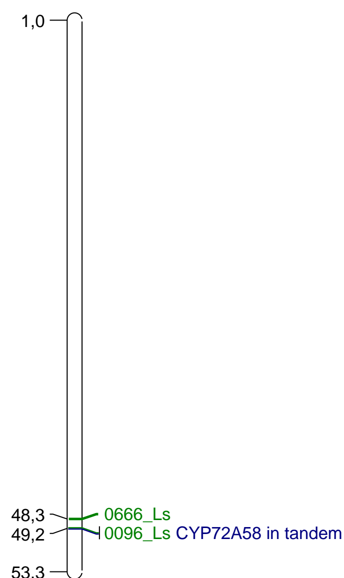


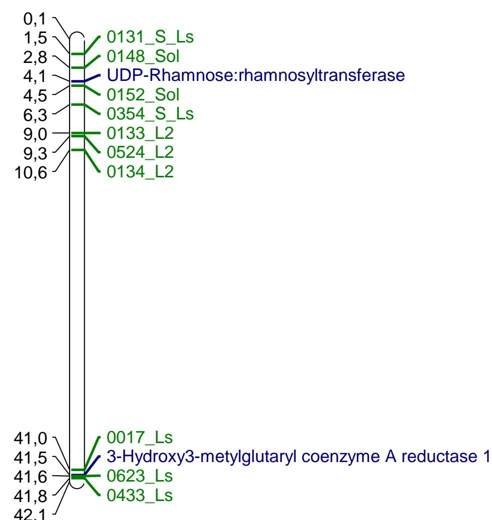




A







B

Figure 3 Physical position of significant superscaffolds identified on pseudochromosomes of *Solanum phureja* DM 1-3 516 R44. On the left of each pseudochromosome is the physical distances in Mb, and on the right superscaffold number (the last four numbers of each superscaffold PGSC0003DMB000000173 = 0173) and the compound associated (α-solanine = S, α-chaconine = C, leptine I = L1, leptine II = L2 or Leptines = Ls). Significant superscaffolds associated with synthesis (A) and accumulation (B). The location of putative candidate genes in the SGA biosynthetic pathway are shown in blue.

Supplementary Table 1 Significant SNPs associated with SGA synthesis in a segregating F_2 population

Superscaffold	Pseudomolecule	α-Solanine	α-Chaconine	Leptine I	Leptine II
PGSC0003DMB000000173	chr01	solcap_snp_c2_52492	solcap_snp_c2_52492		solcap_snp_c1_6518
PGSC0003DMB000000716	chr02			solcap_snp_c2_54579 solcap_snp_c1_15874	solcap_snp_c2_54579 solcap_snp_c1_15874
PGSC0003DMB000000183	chr02			solcap_snp_c1_4794 solcap_snp_c2_4515 solcap_snp_c2_4502	solcap_snp_c1_4794 solcap_snp_c2_4515 solcap_snp_c2_4502
PGSC0003DMB000000082	chr02			solcap_snp_c1_6462 solcap_snp_c2_20445	solcap_snp_c1_6462 solcap_snp_c2_20445
PGSC0003DMB000000442	chr02			solcap_snp_c2_41875	solcap_snp_c2_41875
PGSC0003DMB000000235	chr02			solcap_snp_c2_16362 solcap_snp_c2_54473 solcap_snp_c2_16341 solcap_snp_c2_16324	solcap_snp_c2_16362 solcap_snp_c2_54473 solcap_snp_c2_16341 solcap_snp_c2_16324
PGSC0003DMB000000120	chr02			solcap_snp_c1_3746 solcap_snp_c1_3753 solcap_snp_c2_11591	solcap_snp_c1_3746 solcap_snp_c1_3753 solcap_snp_c2_11591
PGSC0003DMB000000568	chr02			solcap_snp_c2_11587	solcap_snp_c2_11587
PGSC0003DMB000000066	chr02			solcap_snp_c2_48734 solcap_snp_c2_48725 solcap_snp_c2_4358 solcap_snp_c2_4354 solcap_snp_c2_4344 solcap_snp_c1_1504 solcap_snp_c2_4372	solcap_snp_c2_48734 solcap_snp_c2_48725 solcap_snp_c2_4358 solcap_snp_c2_4354 solcap_snp_c2_4344 solcap_snp_c1_1504 solcap_snp_c2_4372
PGSC0003DMB000000287	chr02			solcap_snp_c2_765 solcap_snp_c2_736 solcap_snp_c2_735 solcap_snp_c2_811	solcap_snp_c2_765 solcap_snp_c2_736 solcap_snp_c2_735 solcap_snp_c2_811
PGSC0003DMB000000320	chr02			solcap_snp_c2_796 solcap_snp_c2_791 solcap_snp_c1_242	solcap_snp_c2_796 solcap_snp_c2_791 solcap_snp_c1_242
PGSC0003DMB000000706	chr02			solcap_snp_c2_806 solcap_snp_c2_805 solcap_snp_c2_803	solcap_snp_c2_806 solcap_snp_c2_805 solcap_snp_c2_803

Superscaffold	Pseudomolecule	α-Solanine	α-Chaconine	Leptine I	Leptine II
				solcap_snp_c2_32244 solcap_snp_c2_32252 solcap_snp_c2_32253 solcap_snp_c2_32254	solcap_snp_c2_32244 solcap_snp_c2_32252 solcap_snp_c2_32253 solcap_snp_c2_32254
PGSC0003DMB000000454	chr02			solcap_snp_c1_9655	solcap_snp_c1_9655
PGSC0003DMB000000122	chr02			solcap_snp_c1_15974 solcap_snp_c2_41124	solcap_snp_c1_15974 solcap_snp_c2_41124
PGSC0003DMB000000359	chr02			solcap_snp_c1_237	solcap_snp_c1_237
PGSC0003DMB000000732	chr02			solcap_snp_c2_30160 solcap_snp_c2_30162	solcap_snp_c2_30160 solcap_snp_c2_30162
PGSC0003DMB000000378	chr02			solcap_snp_c2_17400	solcap_snp_c2_17400
PGSC0003DMB000000085	chr02			solcap_snp_c2_38022 solcap_snp_c2_32381 solcap_snp_c2_32400	solcap_snp_c2_38022 solcap_snp_c2_32381 solcap_snp_c2_32400
PGSC0003DMB000000490	chr02			solcap_snp_c1_9695 solcap_snp_c1_9696 solcap_snp_c2_32413 solcap_snp_c2_32415 solcap_snp_c2_32417 solcap_snp_c2_32440 solcap_snp_c2_50885 solcap_snp_c2_50878	solcap_snp_c1_9695 solcap_snp_c1_9696 solcap_snp_c2_32413 solcap_snp_c2_32415 solcap_snp_c2_32417 solcap_snp_c2_32440 solcap_snp_c2_50885 solcap_snp_c2_50878
PGSC0003DMB000000215	chr06	solcap_snp_c1_10109 solcap_snp_c2_33871 solcap_snp_c2_5836 solcap_snp_c2_5843 solcap_snp_c2_5867 solcap_snp_c1_2109 solcap_snp_c1_2117	solcap_snp_c1_10109 solcap_snp_c2_33871 solcap_snp_c2_5836 solcap_snp_c2_5843 solcap_snp_c2_5867 solcap_snp_c1_2109 solcap_snp_c1_2117		
PGSC0003DMB000000578	chr06	solcap_snp_c1_2060 solcap_snp_c1_2065 solcap_snp_c2_5772 solcap_snp_c2_5774 solcap_snp_c2_5775 solcap_snp_c2_5812	solcap_snp_c1_2060 solcap_snp_c1_2065 solcap_snp_c2_5772 solcap_snp_c2_5774 solcap_snp_c2_5775 solcap_snp_c2_5812		

Superscaffold	Pseudomolecule	α-Solanine	α-Chaconine	Leptine I	Leptine II
PGSC0003DMB000000550	chr06	solcap_snp_c2_5821	solcap_snp_c2_5821		
PGSC0003DMB000000180	chr06	solcap_snp_c1_2112 solcap_snp_c2_41406 solcap_snp_c1_13135 solcap_snp_c1_13155 solcap_snp_c2_8832 solcap_snp_c1_3001 solcap_snp_c2_8867 solcap_snp_c1_3011	solcap_snp_c1_2112 solcap_snp_c2_41406 solcap_snp_c1_13135 solcap_snp_c1_13155 solcap_snp_c2_8832 solcap_snp_c1_3001 solcap_snp_c2_8867 solcap_snp_c1_3011		
PGSC0003DMB000000087	chr06	solcap_snp_c2_22289 solcap_snp_c2_8664 solcap_snp_c2_8663 solcap_snp_c2_8662 solcap_snp_c2_8661 solcap_snp_c2_8660 solcap_snp_c1_2944 solcap_snp_c2_9002 solcap_snp_c2_8999 solcap_snp_c1_2978 solcap_snp_c2_8786	solcap_snp_c2_22289 solcap_snp_c2_8664 solcap_snp_c2_8663 solcap_snp_c2_8662 solcap_snp_c2_8661 solcap_snp_c2_8660 solcap_snp_c1_2944 solcap_snp_c2_9002 solcap_snp_c2_8999 solcap_snp_c1_2978 solcap_snp_c2_8786		
PGSC0003DMB000000666	chr07			solcap_snp_c2_28228	solcap_snp_c2_28228
PGSC0003DMB000000096	chr07			solcap_snp_c1_4029	solcap_snp_c1_4029
PGSC0003DMB000000465	chr08			solcap_snp_c2_51053	
PGSC0003DMB000000131	chr11				solcap_snp_c2_52749
PGSC0003DMB000000034	chr12			solcap_snp_c2_7995	solcap_snp_c2_7995

The superscaffolds are ordered by pseudochromosomes (chr) and then by their physical location from top to bottom on each chr. The superscaffolds are named with the initials of Potato Genome Sequencing Consortium (PGSC) followed by Assembly Version 3 of DM (double monoploid) *S. tuberosum* Group Phureja DM1-3 516R44 (CIP801092), PGSC Version 2.1.11 Pseudomolecule Sequences and superscaffold number.

Supplementary Table 2 Significant SNPs associated with SGA accumulation in a segregating F_2 population

Superscaffold	Pseudomolecule	α-Solanine	α-Chaconine	Leptine I	Leptine II
PGSC0003DMB000000173	chr01			solcap_snp_c1_6518	solcap_snp_c1_6518
PGSC0003DMB000000290	chr01		solcap_snp_c1_11288		
PGSC0003DMB000000120	chr02			solcap_snp_c2_11591	
PGSC0003DMB000000528	chr02				solcap_snp_c2_45308
PGSC0003DMB000000004	chr02			solcap_snp_c2_47197	
PGSC0003DMB000000208	chr06	solcap_snp_c2_30508 solcap_snp_c2_30633			
PGSC0003DMB000000432	chr06			solcap_snp_c2_37611	
PGSC0003DMB000000315	chr06	solcap_snp_c2_51757	solcap_snp_c2_51757		
PGSC0003DMB000000558	chr06	solcap_snp_c2_43116 solcap_snp_c2_57018 solcap_snp_c1_16470 solcap_snp_c2_57017 solcap_snp_c2_57014 solcap_snp_c2_43135			
PGSC0003DMB000000227	chr06	solcap_snp_c2_54029			
PGSC0003DMB000000726	chr06	solcap_snp_c2_53053			
PGSC0003DMB000000461	chr06	solcap_snp_c2_37762			
PGSC0003DMB000000222	chr06			solcap_snp_c2_16863	
PGSC0003DMB000000158	chr06	solcap_snp_c2_25947 solcap_snp_c2_25926 solcap_snp_c1_8140	solcap_snp_c2_25947 solcap_snp_c2_25926		
PGSC0003DMB000000180	chr06			solcap_snp_c2_8832 solcap_snp_c1_3001 solcap_snp_c1_3011 solcap_snp_c1_3012	
PGSC0003DMB000000087	chr06			solcap_snp_c1_7040 solcap_snp_c1_6992	solcap_snp_c1_7040 solcap_snp_c1_6992

Superscaffold	Pseudomolecule	α-Solanine	α-Chaconine	Leptine I	Leptine II
				solcap_snp_c2_22301 solcap_snp_c2_8999 solcap_snp_c2_8786	solcap_snp_c2_22301 solcap_snp_c2_8999
PGSC0003DMB000000128	chr06			solcap_snp_c1_8659 solcap_snp_c1_16127 solcap_snp_c2_41223 solcap_snp_c2_41210	solcap_snp_c1_8659 solcap_snp_c1_16127 solcap_snp_c2_41223 solcap_snp_c2_41210
PGSC0003DMB000000150	chr06			solcap_snp_c2_9099	solcap_snp_c2_9099
PGSC0003DMB000000686	chr06			solcap_snp_c2_9255	solcap_snp_c2_9255
PGSC0003DMB000000251	chr07	solcap_snp_c2_15908			
PGSC0003DMB000000666	chr07			solcap_snp_c2_28228	solcap_snp_c2_28228
PGSC0003DMB000000096	chr07			solcap_snp_c1_4029	solcap_snp_c1_4029
PGSC0003DMB000000147	chr08	solcap_snp_c2_48951 solcap_snp_c2_30067 solcap_snp_c2_30037			
PGSC0003DMB000000723	chr08			solcap_snp_c2_44334	solcap_snp_c2_44334
PGSC0003DMB000000487	chr08			solcap_snp_c2_51374	solcap_snp_c2_51374
PGSC0003DMB000000306	chr08			solcap_snp_c2_45759	solcap_snp_c2_45759
PGSC0003DMB000000267	chr10	solcap_snp_c1_11801 solcap_snp_c1_11802 solcap_snp_c1_11804			
PGSC0003DMB000000131	chr11	solcap_snp_c2_52749		solcap_snp_c2_23976 solcap_snp_c2_52749	solcap_snp_c2_23976 solcap_snp_c2_52749
PGSC0003DMB000000148	chr11	solcap_snp_c1_4319 solcap_snp_c1_4322 solcap_snp_c1_4328 solcap_snp_c1_4336 solcap_snp_c2_13566			
PGSC0003DMB000000152	chr11	solcap_snp_c2_6001 solcap_snp_c2_5960 solcap_snp_c2_5957 solcap_snp_c1_2304			
PGSC0003DMB000000354	chr11	solcap_snp_c2_20947		solcap_snp_c2_20947	solcap_snp_c2_20947

Superscaffold	Pseudomolecule	α-Solanine	α-Chaconine	Leptine I	Leptine II
PGSC0003DMB000000133	chr11				solcap_snp_c2_12276 solcap_snp_c1_3996 solcap_snp_c2_24318 solcap_snp_c2_49312 solcap_snp_c2_49311 solcap_snp_c2_49310 solcap_snp_c2_53682 solcap_snp_c2_53684 solcap_snp_c1_15656 solcap_snp_c1_15658
PGSC0003DMB000000524	chr11				solcap_snp_c2_32982 solcap_snp_c2_32989 solcap_snp_c2_33002 solcap_snp_c2_53273
PGSC0003DMB000000134	chr11				solcap_snp_c2_32954 solcap_snp_c2_56630 solcap_snp_c2_56627 solcap_snp_c2_56624 solcap_snp_c2_56623 solcap_snp_c2_32341 solcap_snp_c2_29096 solcap_snp_c2_29112
PGSC0003DMB000000017	chr11			solcap_snp_c2_30300 solcap_snp_c2_30297 solcap_snp_c1_9204 solcap_snp_c2_39912 solcap_snp_c2_31487	solcap_snp_c2_30300 solcap_snp_c2_30297 solcap_snp_c1_9204 solcap_snp_c2_39912 solcap_snp_c2_31487
PGSC0003DMB000000623	chr11			solcap_snp_c2_43886 solcap_snp_c1_12896 solcap_snp_c2_43860	solcap_snp_c2_43886 solcap_snp_c1_12896 solcap_snp_c2_43860
PGSC0003DMB000000433	chr11			solcap_snp_c2_34229 solcap_snp_c1_10255 solcap_snp_c1_10256 solcap_snp_c2_34191 solcap_snp_c2_34193 solcap_snp_c2_34194	solcap_snp_c2_34229 solcap_snp_c1_10255 solcap_snp_c1_10256 solcap_snp_c2_34191 solcap_snp_c2_34193 solcap_snp_c2_34194

Superscaffold	Pseudomolecule	α-Solanine	α-Chaconine	Leptine I	Leptine II
				solcap_snp_c2_34196 solcap_snp_c2_34197 solcap_snp_c2_34198 solcap_snp_c2_34202 solcap_snp_c2_34203 solcap_snp_c2_34219	solcap_snp_c2_34196 solcap_snp_c2_34197 solcap_snp_c2_34198 solcap_snp_c2_34202 solcap_snp_c2_34203 solcap_snp_c2_34219

The superscaffolds are ordered by pseudochromosomes (chr) and then by their physical location from top to bottom on each chr. The superscaffolds are named with the initials of Potato Genome Sequencing Consortium (PGSC) followed by Assembly Version 3 of DM (double monoploid) *S. tuberosum* Group Phureja DM1-3 516 R44 (CIP801092), PGSC Version 2.1.11 Pseudomolecule Sequences and superscaffold number.

3. Sequence Diversity in Coding Regions of Candidate Genes in the Glycoalkaloid Biosynthetic Pathway of Wild Potato Species.

Manrique-Carpintero NC[2], Tokuhisa JG[1], Ginzberg I[2], Holliday JA[3] and Veilleux RE[1](✉)

(✉) Author for correspondence

Tel: 540 231-5584

Fax: 540 231-3083

E-mail: potato@vt.edu

Authors contribution

Dr Richard Veilleux, Dr James Tokuhisa and Dr Idit Ginzberg were principal investigator on the project "Identification and allelic variation of genes involved in the potato glycoalkaloid biosynthetic pathway" funded by United States - *Israel* Binational Agricultural Research and Development (BARD). Dr Richard Veilleux directed this investigation and with Dr Holliday, Dr Tokuhisa and Dr Idit Ginzberg edited the article. I conducted the laboratory experiments, optimized the steroidal glycoalkaloid (SGAs) extraction protocol, collected data, did statistical analysis and wrote the first draft of this document as well as the corrections.

[2]Department of Horticulture, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

[2]Department of Vegetable Research, The Volcani Center, Agricultural Research Organization, Bet Dagan, Israel

[3]Department of Forest Resources and Environmental Conservation, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

Abstract

The genetic resources of wild *Solanum* section Petota species are a crucial source of heritable diversity for breeding tolerance to abiotic and resistance to biotic stress in potato. Natural variation in five candidate genes of the steroidal glycoalkaloid (SGA) metabolic pathway and whole genome SNP genotyping were studied in six wild (*Solanum chacoense* (*chc* 80-1), *S. commersonii* subsp. *commersonii*, *S. demissum*, *S. sparsipilum*, *S. spegazzinii*, *S. stoloniferum)* and one cultivated *S. tuberosum* Group Phureja (*phu* DH) species of potato with contrasting levels of SGAs. One and two genomic sequences of partial coding regions were obtained for five candidate genes [3-hydroxy-3-methylglutaryl coenzyme A reductase 1 and 2 (*HMG1, HMG2);* 2,3-squalene epoxidase (*SQE*); solanidine galactosyltransferase (*SGT1*); and solanidine glucosyltransferase (*SGT2*)]. A total of 337 single nucleotide polymorphisms (SNPs) that produced 354 variations was detected among a total of 3.7 Kb of sequenced DNA. Greater frequencies of polymorphisms were found in intron regions and in genes at the secondary metabolism level. While no significant deviation from neutrality was found, dN/dS ratios < 1 and negative values of Tajima's D test suggested purifying selection and genetic hitchhiking in the analyzed gene fragments. In addition, patterns of dN/dS ratios across the SGA pathway support the possibility that this pathway is constrained by natural selection. Specifically, comparison of nucleotide diversity estimates and dN/dS ratios showed stronger selective constraints acting on genes at the primary metabolism level (*HMG1, HMG2* and *SQE*) than at secondary metabolism (*SGT1* and *SGT2*). Informative SNPs putatively associated with SGA levels were selected from candidate genes. Twenty four SNPs with an exclusive genotype for either *phu* DH or *chc* 80-1 with the least and the greatest SGA production, respectively, were identified for *HMG2, SQE*, *SGT1* and *SGT2*. The Illumina Potato SNP chip was used for whole genome genotyping of our studied germplasm. Cluster analysis of SNPs putatively associated with SGA accumulation was performed. Eight informative SNPs on six pseudochromosomes with homozygous and heterozygous genotypes that discriminated high, intermediate and low levels of SGA accumulation were determined. These results can be used to evaluate SGA accumulation trait on segregating or association mapping populations.

Key words: Nucleotide diversity, dN/dS ratio, *Solanum*, Infinium 8303 Potato Array

Introduction

Steroidal glycoalkaloids (SGAs) are secondary metabolites mainly produced in solanaceous species. These compounds function in potato species as one defense against pathogens and insects (Friedman 2006; Nema et al. 2008; Wink 2003). The SGA structure consists of a hydrophobic C_{27}-steroidal alkaloid skeleton (aglycone) containing a nitrogen atom as a secondary or tertiary amine, and a hydrophilic glycosidic moiety of three or four sugars attached to the 3-OH position of the aglycone (Bushway et al. 1980; Milner et al. 2011). This chemical structure has cytotoxic properties, such as the inhibition of acetylcholinesterase activity and the disruption of cell membrane function (Keukens et al. 1995; Orgell et al. 1958; Roddick et al. 1990). Several types of SGAs have been reported in cultivated potato (*Solanum tuberosum* L. Group Tuberosum) and wild relative species, the most commonly accumulated are the triose glycosides of solanidine, α-chaconine and α-solanine (Distl and Wink 2009; Shakya and Navarre 2008). Some of the SGAs that have been identified as the most effective poisonous compounds against different potato pests are demissine, commersonine, dehydrocommersonine and the leptines (Distl and Wink 2009). Several potato breeding programs have attempted to incorporate the resistance or tolerance to insects associated with these compounds into the cultivated potato by introgression (Lorenzen et al. 2001; Sanford et al. 1996; Sørensen et al. 2008; Thompson et al. 2008). However, tissue specific accumulation of SGAs in leaves rather than overall production is required to breed potato resistant to its major pests, in order to avoid undesirable characteristics in the tubers.

Wild potato species have an enormous genetic potential for potato breeding, both with regard to tuber quality and resistance to insects and pathogens. The tuber-bearing *Solanum* section *Petota* Dumont., with about 100 wild species and four cultivated species (Spooner 2009), is distributed from southwestern USA to Chile, Argentina and Uruguay. Taxonomically they are a closely related group due to sexual compatibility among many species, introgression, interspecific hybridization, auto- and allopolyploidy, a mixture of sexual and asexual reproduction, possible recent species divergence, phenotypic plasticity, and morphological similarity that make it difficult to distinguish

species and series (Spooner 2009; Spooner and Salas 2006). Fourteen wild species have been used to incorporate resistance to viral, fungal, and bacterial diseases, as well as to insect and nematode pests of potato (Spooner and Salas 2006). Presence of glycoalkaloids, dense hairs, and glandular trichomes are characteristics that have been identified in the wild potato species associated with resistance to major potato insect pests (Flanders et al. 1992). Introgression of resistance traits into potato cultivars implies several generations of backcrosses and recurrent selection (Gebhardt and Valkonen 2001). The breeding process could be limited by the introduction of undesirable traits associated with the resistance, genetic complexity of traits, or differences in endosperm balance number and incompatibility between the parental lines (Grafius and Douches 2008). Breeding for resistance mediated by SGAs is complicated by polygenic inheritance, by the high correlation in SGA content between foliage and tubers, and because SGA accumulation is influenced by environmental factors and crop management activities (Tingey 1984). Identification and cloning of genes associated with synthesis and accumulation of SGAs, present an alternative to either marker assisted selection or genetic transformation. Traditionally, the identification of genes associated with traits of interest has been by linkage analysis followed by positional cloning and insertional mutagenesis. However, polyploidy, heterozygosity, and inbreeding depression observed after self-pollination, are factors that limit fine mapping and identification of genes underling quantitative traits in potato (Gebhardt et al. 2007; Kloosterman et al. 2010). Genomic sequences of cultivated species or their relatives as well as elucidation of metabolic pathways have become a source of putative candidate genes associated with complex traits. Candidate genes are sequenced genes of known biological structural or regulatory function that could be involved with major loci controlling the trait of interest (Pflieger et al. 2001). Where a biochemical pathway is well characterized, as is the case of sterol biosynthesis, the candidate gene approach can be a powerful tool to determine relationships between genetic variation and trait variation.

The biosynthetic genes and the genetic factors that regulate the expression of SGAs are not fully understood. SGA biosynthesis begins as a part of primary metabolism in the cytoplasm mevalonic/isoprenoid pathway (Figure 1). In this part, acetyl-CoA is used to generate C_5 isoprene units. The isoprene is converted by different enzymes to

synthesize dimethylallyl pyrophosphate (DMAPP), geranyl pyrophosphate (GPP), and farnesyl pyrophosphate (FPP). These three compounds are precursors of primary and secondary isoprenoid metabolites, hemiterpenes, monoterpenes, sesquiterpenes, and steroids. In steroid metabolism, SGA synthesis commences with the formation of the most rudimentary sterol, cycloartenol and from this cholesterol. Little is known about further metabolic steps involved in the conversion of cholesterol to SGAs.

Specialized gene function is required for the synthesis of specific isoprenoids (Suzuki and Muranaka 2007). In relation to SGA biosynthesis some genes perform a highlighted function in the regulation of this pathway. At primary metabolism, 3-hydroxy-3-methylglutaryl coenzyme A reductase (HMGR) catalyzes the formation of mevalonic acid. *HMGR* is a family of genes where some members are involved in biosynthesis of sterols and triterpenoids mainly for plant development as is the case of *HMG1*, whereas others including *HMG2* are involved in the production of defense compounds (Suzuki and Muranaka 2007). In potato *HMG1* has been associated with SGA accumulation after wounding, and *HMG2* and *HMG3* transcripts are up-regulated following wounding and pathogen inoculation and associated with increasing sesquiterpenoid production (Choi et al. 1994; Choi et al. 1992; Krits et al. 2007). Downstream in this primary pathway, squalene synthase (SQS) condenses two farnesyl diphosphate molecules to form squalene. Increased transcription of *SQS* has been associated with high SGA levels in potato species (Ginzberg et al. 2009; Krits et al. 2007; Yoshioka et al. 1999). Squalene epoxidase (SQE) mediates the epoxidation of squalene to 2,3-oxidosqualene. Inhibition of this enzyme decreased sterol levels in tobacco *Nicotiana tabacum* cv. Bright Yellow-2 suspension cells (Wentzinger et al. 2002). A family of seven *SQE* genes has been reported in Arabidopsis which may imply a more regulated role of this enzyme in sterol biosynthesis (Suzuki and Muranaka 2007). Cycloartenol synthase (CAS) converts 2,3-oxidosqualene into cycloartenol which is used to produce cholesterol, campesterol and sitosterol, the main sterols found in plants (Ginzberg et al. 2009; Schaller 2004).

A second segment of SGA biosynthesis is the conversion of cholesterol to SGAs, which is considered as secondary metabolism involving particular enzymes which

are unique to the liliaceous and solanaceous species that produce SGAs (Heftmann 1983). Reactions that convert cholesterol into solanidine have been proposed (Kaneko et al. 1977); but the enzymes involved have not been identified (Ginzberg et al. 2009). Solanidine is the precursor of α–chaconine and α–solanine, the most common SGAs in potato species and (Tschesche and Hulpke (1967), Canonica et al. (1977), (Heftmann 1983) cited by Arnqvist (2007)). Glycosyltransferase enzymes catalyze the final reactions in the biosynthesis of SGAs where different sugar moieties are added to the solanidine aglycone. Three genes have been identified in potato: solanidine galactosyltransferase (*SGT1*), solanidine glucosyltransferase (*SGT2*), and rhamnosyltransferase (*SGT3*) (McCue et al. 2007a; McCue et al. 2007b; McCue et al. 2005; Moehs et al. 1997). *SGT1* and *SGT2* initiate the glycosylation of solanidine to γ-solanine and γ–chaconine, respectively; and *SGT3* catalyzes the conversion to α-solanine and α–chaconine.

Genetic studies of wild and cultivated potato species have concluded that multiple genetic factors interact for the production of SGAs (Sanford et al. 1996; Sanford and Sinden 1972; Sørensen et al. 2008; Yencho et al. 1998). The characterization of biosynthetic genes and their regulatory elements is necessary to develop potato cultivars with low SGAs in tubers and high levels in leaves to protect the crop against pathogens and pests. The objective of this work was to analyze the nucleotide diversity of coding region fragments of five candidate genes involved in biosynthesis of SGAs (*HMG1, HMG2, SQE, SGT1* and *SGT2*) in six wild potato species with different levels of SGAs (*Solanum chacoense*, *S. commersonii* subsp. *commersonii*, *S. demissum*, *S. sparsipilum*, *S. spegazzinii*, *S. stoloniferum)* compared to the cultivated *S. tuberosum* Group Phureja. The biosynthetic candidate genes involved in primary and/or secondary metabolism were selected based on previous studies, which implied a possible association of these genes with the accumulation of SGAs. We assessed levels of genetic variation and single nucleotide polymorphisms (SNPs) association to determine the possible role of these genes in governing accumulation of SGAs. A whole genome SNP genotyping analysis was also carried out to identify new genomic regions putatively associated with the accumulation of these compounds in potato species.

Materials and methods

Plant material

Two accessions with contrasting SGA content (based on SGA data available in the National Research Project 6-NRSP-6 United States Potato Genebank) were selected for five wild potato species from the NRSP-6 collection (Table 1). In addition, a single clone each of *Solanum chacoense* USDA 8380-1 (*chc* 80-1) and *S. tuberosum* Group Phureja DH OT-B × N-B P5/AR2 (*phu* DH) were used. Sinden et al. (1980) reported high levels of accumulation of SGAs in the leaves of *chc* 80-1, and the production of leptines associated with resistance to Colorado potato beetle (*Leptinotarsa decemlineata* Say). By contrast, *phu* DH with low levels of SGAs is a dihaploid derived from an intermonoploid somatic hybrid (Johnson et al. 2001; Lightbourn and Veilleux 2007; Veilleux 1990). Short names were assigned to each accessions, the standard abbreviation for S*olanun* species base on Simmonds (1963) followed by a random number (Table 1).

Five seeds per accession were treated with 1,000 mg/l GA_3 solution overnight, and then sown in cell packs with Fafard® super fine germination mix (Farfard, Agawan, MA). A single seedling per accession was transferred to D-40 Deepots™ (Hummert International, Earth City, MO) with Premier Horticulture® Pro-mix BX 15 days after sowing. Simultaneously, after 1 week of acclimation to grow in soil, *in vitro* grown plants of *chc* 80-1 and *phu* DH clones were transferred to D-40 Deepots™. The plants were grown under controlled environment (Conviron, Winnipeg) set to 60% relative humidity, 14 h photoperiod, 250 $\mu mol/m^2/s$ light intensity and day/night temperatures of 20 and 16 °C, respectively. Plants were fertilized with MiracleGro All Purpose® (Scotts Co, Marysville, OH) at the rate of 1 g/L every 15 days. Three biological replications were established per individual plant from each accession. At 15 day intervals serial apical cuttings were taken to generate the three temporal replications from 1 month-old established plants into D-40 Deepots™. At 55 days after cutting, leaves 4-6 from the shoot apex were harvested for SGA extraction. Shoot apical meristems were collected for DNA extraction from the starter plants.

DNA extraction and PCR amplification

Genomic DNA was extracted from leaf tissue of 12 accessions by a modified cetyltrimethylammonium bromide (CTAB) protocol (Doyle and Doyle 1987; Murray and Thompson 1980; Stewart and Via 1993). Homologous nucleotide sequences for each candidate gene from various solanaceous species were obtained from the GenBank database between 2008 and 2010. These sequences were aligned to design primers in conserved coding regions (Table 2). For *HMG1* the sequence from the draft genome of DM 1-3 516 *S. tuberosum* Group Phureja potato was also used (The Potato Genome Sequencing Consortium 2011). The scaffold PGSC0003DMS000003141 had the greatest similarity score in a BLAST search with LO1400 HMG1 GenBank sequence of *S. tuberosum*. For *HMG1* (PGSC0003DMG400013663, 35185 and 46343 gene model), *SGT1* (PGSC0003DMG400011749) and *SGT2* (PGSC0003DMG400017508) the fragment was located exclusively in a single exonic region, while the PCR products for *HMG2* (PGSC0003DMG400003461) and *SQE* (PGSC0003DMG400004923) captured some intronic sequences. The amplified region of *HMG1* coded for a 292 amino acid (aa) sequence with a catalytic and tetramerization interface domains. For *HMG2* the fragment covered 134 aa with catalytic, NADP(H) binding, substrate binding and tetramerization residues. The 134 aa segment of *SQE* spanned more than one domain of its gene family. A region of 277 aa with glycosyltransferase multidomain was amplified for *SGT1*, and a segment encoding 189 aa with active site conserved domains and TDP (thymidine diphosphate)-binding site residues was isolated for *SGT2*. The primers were designed using DNASTAR® Lasergene 9 core suite software for sequence analysis and assembly. PCR was performed in 25 µl of 1× *Ex Taq*™ polymerase buffer, 0.2 mM of each dNTP, 0.24 µM of each primer, and 0.2 units of high-fidelity TaKaRa *Ex Taq*™ DNA polymerase (Takara Biotechnology, Shiga, Japan) and 100 ng of genomic DNA template. Standard cycling conditions were 5 min initial denaturation at 95 °C followed by 30 cycles of 0.5 min at 94 °C, 0.5 min annealing at the appropriate T_m and a 2 min extension time (Table 2). The reactions were finished by a 5 min incubation at 72 °C. Reconditioning PCR conditions were used for *HMG1* and *SGT1* to avoid recombinant PCR products (Judo et al. 1998; Lenz and Becker 2008; Thompson et al. 2002). The final PCR product was derived from two PCR reactions of 20 cycles each, where 2 µl of product from the first reaction were used as a template in the second reaction.

Identification of allelic sequences

PCR products were gel-purified with QIAquick Gel Extraction Kit (Qiagen, Hilden, Germany) and cloned into pJET1.2/blunt vector (CloneJET^{TM} PCR Cloning Kit, Fermentas, Thermo Fisher Scientific Inc., Waltham, USA) using JM109 competent cells (Promega, Madison, WI). At least five colonies from each cloning reaction were sent for sequencing using both forward and reverse primers (Virginia Bioinformatics Institute Core Facility at Virginia Tech, Blacksburg, VA). Forward and reverse sequences were aligned to identify consensus sequences. Sequencing errors were corrected by visual inspection of peak quality in chromatograms. The set of consensus sequences per accession was aligned to identify alleles. A minimum alignment of at least three identical consensus sequences defined an allelic sequence from the pool of sequences. In some cases, additional sequence errors were detected when consensus sequences following the same pattern were aligned. These were corrected, and the sequences used for identification of allelic sequences. Sequence alignments and editing were done using the different applications included in the DNASTAR® Lasergene 9 core suite software for sequence analysis and assembly and Sequence Scanner Software v1.0 from Applied Biosystems®.

Sequence polymorphism and diversity analysis

Sequence polymorphisms were analyzed using the DNA Sequence Polymorphism (DnaSP5) software version 5.0 (Librado and Rozas 2009) and Molecular Evolutionary Genetics Analysis (MEGA5) software version 5.0 (Tamura et al. 2011). The ratio of nonsynonymous (dN) and synonymous (dS) substitution rates was calculated per gene using CODEML software from Phylogenetic Analysis by Maximum Likelihood version 4.5 (PAML4.5) package of programs (Yang 2007). This ratio provides insights of selective pressures acting on protein-coding regions, and allows identifying positive selection ($dN/dS > 1$) or purifying selection ($dN/dS < 1$). Likelihood analyses were calculated on six codon substitution site models implemented in CODEML. The difference between two models was estimated by likelihood ratio test statistics (LRT), twice the log likelihood difference between the two compared models ($2\Delta l$) compared

against χ^2 distribution. Multiple allelic sequences of each candidate gene were aligned by the CLUSTAL W method (Thompson et al. 1994) using MEGA5. Then a phylogenetic tree was constructed per gene according to maximum likelihood method in the analysis interface of MEGA5. The alignments and phylogenetic trees were used for analyses in MEGA5, DnaSP5 and PAML4.5. Besides the allelic sequences identified per gene in the wild potato species selected for this study, available GenBank cDNA sequences of *S. tuberosum* and genomic sequences from the sequenced potato genome *S. tuberosum* Group Phureja DM 1-3 516 R44 were added to the analysis of each gene. Nucleotide diversity for the entire population per gene was calculated using Tajima-Nei's π estimator (Tajima and Nei 1984) using MEGA5. π is the average number of nucleotide differences per site between two homologous sequences. MEGA5 calculates π as the average of the values for all pairwise comparisons. The deviation of nucleotide variation patterns from the neutral theory of molecular evolution (Kimura 1983), that assumes all mutations in a DNA region are selectively neutral, was tested by Tajima's D statistical analysis using MEGA5. The Tajima's D test compares whether π and θ values are significantly different (Tajima 1989). Exon delimitations were assigned in MEGA5 and DnaSP5 following GenBank cDNA sequences as a pattern. Numbers of synonymous and nonsynonymous substitutions were estimated using DnaSP5 conservative criteria. In some complex cases with sites segregating for several codons (highly variable regions), the synonymous and nonsynonymous substitutions were estimated manually.

SNP chip analysis

A whole genome SNP genotyping using the SolCAP 8303 Illumina Infinium potato SNP chip was done in twelve accessions (Hamilton et al. 2011). The SNP genotyping facility at Michigan State University processed the genomic DNA samples on an Illumina iScan Reader utilizing the Infinium® HD Assay Ultra (Illumina, Inc., San Diego, CA) and the Infinium 8303 Potato Array. The 8303 SNP data were filtered and used for Analysis of Variance (ANOVA) with average of total SGA accumulation amounts per accession. SNPs that were monomorphic for all individuals, and SNPs with a no-call rate >25% (greater than three samples with missing genotypes) were eliminated from the initial data set. From 5,392 analyzed SNPs, significant ones were selected based

on $R^2 > 0.2$, P-value < 0.05, minimum of two samples per SNP genotype, and minimum difference of 5 logarithm units between the means of the most significant different SNP genotypes, and SNPs with superscaffold and pseudomolecule information on the potato genome browser (http://solanaceae.plantbiology.msu.edu/cgi-bin/gbrowse/potato/). Significant SNPs were used for cluster analysis to identified informative SNPs putatively associated with SGA levels. The physical position in the potato genome of informative SNPs was obtained from Felcher et al. (2012).

SGA extraction and quantification

Leaf tissue harvested from each of three biological replications of the 12 accessions, was freeze-dried for 72 h in a lyophilizer (Labconco, Kansas City, MO), then ground and SGAs extracted using a modified procedure (Edwards and Cobb 1996). An initial extract, from 30 mg of leaf powder mixed by vortex with 1 ml of extraction buffer (5% v/v acetic acid, 0.02 M heptane sulfonic acid), was ultrasonicated for 2 sec (Digital sonifier® cell disruptor, Branson Ultrasonic Corporation, NY, USA with 20% amplitude). The extract was set on a microtube thermal-mixer for 15 min at 1,200 rpm, centrifuged 3-5 min at 16,000 rpm, and a clear solution recovered after it was sieved in a 50 µm filter plate. The precipitate was used for a second cycle of extraction. SGAs were concentrated and purified by solid phase extraction using Sep-pak® classic C_{18} cartridge columns (Waters, Milford, MA, USA). Five ml of methanol (MeOH) followed by 5 ml of extraction buffer were added to the columns to activate and equilibrate them. Then the leaf extract was applied followed by a sequence of washes: 7.5 ml of water, 5 ml of 50 mM ammonium bicarbonate (NH_4 HCO_3), 5 ml of 50 mM NH_4 HCO_3:MeOH (1:1 v/v), and 7.5 ml of water. The final SGA extract was eluted with 1.2 ml of elution buffer (80% v/v MeOH, 0.5% v/v formate).

A high performance liquid chromatography (HPLC, Agilent HP 1200 Series, Santa Clara, CA) on a C-18 reverse-phase column (Agilent Eclipse XDB-C18, 5 µm pore size and 4.6 × 150 mm) procedure was used to separate the SGAs, and a photodiode array detector to quantified them. Elution of SGAs was attained by using a binary gradient system consisting of Solvent A (30% acetonitrile, 6 mM Tris-HCl, pH 8.0) and Solvent B

(80% acetonitrile, 6 mM Tris-HCl, pH 7.6) at a flow rate of 0.3 ml/min at 25 °C column temperature. The gradient elution was: 0-0.5 min, 0% B; 0.5-8.5 min, 0-30% B; 8.5-12 min, 30-100% B; 12-16 min, 100% B; 16-16.5 min, 100-0% B; and 16.5-21 min, 0% B. Eluent was monitored at 202 nm. Purified standards solutions of α-solanine and α-chaconine (Sigma) were used to generate calibration curves. Peak absorbance area at A_{202nm} was used to quantify the different SGAs. Total SGA was estimated per sample in mAU per mg of dry weight leaf tissue (mAU/mg DW) and used for further analysis.

Statistical analysis

A completely randomized design was used for plants in the growth chamber and for SGA analysis. The amount of total SGAs was estimated in leaf tissue collected from each biological replication per accession. Analyses of variance were conducted for the total SGA response variable using as source either accessions or allelic states of polymorphic SNPs within candidate genes and the SNP array. The average of repetitions was used for statistical analysis of SNPs. SGA data were transformed with logarithm for all analyses. All statistical analyses were done using JMP® 9 (SAS Institute Inc., Cary, NC, USA).

Results

Allelic sequences and sequence polymorphisms

Sequences of genomic DNA fragments corresponding to conserved coding regions of five candidate genes were obtained by PCR amplification, cloning, sequencing and sequence analysis. One or two allelic sequences per gene were detected per individual since the accessions derived from diploid heterozygous species (Table 3). In some cases a single allelic sequence was shared by accessions from the same species or different species. The total number of unique allelic sequences identified was 72, with a range of 12-18 unique sequences per gene. The identified allelic sequences of candidate genes from each of the 12 accessions, homologous sequences from the sequenced genome of potato and available GenBank sequences from *S. tuberosum* were used in further analysis. *HMG1*, *SGT1* and *SGT2* had respectively five, one and one sequences with SNP and/or indel polymorphisms that shifted the reading frame, creating stop codons or altering the protein sequence. These sequences were eliminated from the analysis to avoid distorting estimations. Sequence polymorphism analysis among the allelic sequences per gene detected a total of 337 variable sites (Table 4). The SNP sites produced 354 SNP mutations, of which 35.3% were nonsynonymous, and 64.7% were silent (synonymous or in noncoding regions). The frequency of transitions vs. transversions was greater for *HMG1* (3.8) and similar for all other genes (2.0-2.4). Genes involved in primary metabolism (*HMG1, HMG2, SQE*) had fewer SNPs in coding regions, 0.03, 0.05, and 0.05 SNPs/bp respectively, than those involved in secondary metabolism (*SGT1* and *SGT2*), with 0.13 and 0.10 SNPs/bp, respectively. However, *HMG2* and *SQE* fragments spanning regions with one and two introns had the greatest rates of SNPs, 0.26 and 0.14 SNPs/bp each. Indels with 1-3 codon deletions were found in *HMG1*, *SGT1* and *SGT2*, and various sized indels were identified in introns of *HMG2* and *SQE*. Codons with 2-3 SNP sites and several variants in each were detected in *SGT1* and *SGT2*; such variable codons would be expected to yield proteins with highly variable amino acids in the position.

Sequence diversity

Codon and nucleotide diversity were estimated for the pool of sequences of each gene (Table 5). In the codon based analysis, the dN/dS ratios were less than one in all gene fragments. The likelihood ratio test between unconstrained (M0) and constrained (M3) analysis were not significant in all cases; thus accepting the M0 null hypothesis of no evidence of selective constraints or adaptive evolution acting on those gene sequences. Tajima's test of neutrality was computed at the nucleotide level, where we found that all gene fragments had non-significant D-statistics values. This analysis confirms no evidence for natural selection at these genetic regions. A significant relationship between pathway position and the estimates of divergence (dN, dS and dN/dS ratios) was found (n=5, r^2=0.96, P<0.003*, n=5, r^2=0.85, P<0.026 and n=5, r^2=0.77, P<0.049*), with smaller ratios (0.29, 0.15 and 0.24) in the genes of primary metabolism (*HMG1*, *HMG2* and *SQE*) than in those (0.42 and 0.39) of secondary metabolism (*SGT1* and *SGT2*) respectively. The nucleotide diversity estimations followed the same pattern for exon regions, and increased greatly for *HMG2* and *SQE* when introns were taken into account for total estimations. Positions of the candidate genes in the metabolic pathway (1= *HMG1* and *HMG2*, 2= *SQE*, and 4= *SGT1* and *SGT2*), for regression analysis, were assigned based on the distance (in the number of reactions) between these genes. At least seven reactions (1 unit) could occur between *HMG1*/*HMG2* and *SQE* in the putative pathway of sterol biosynthesis proposed by Suzuki and Muranaka (2007). In the same way, there are around 13 reactions (2 units) from *SQE* to *SGT1*/*SGT1* using as reference the putative metabolic pathway of cholesterol suggested by Arnqvist (2007) and the three minimal reactions from cholesterol to solanidine aglycone proposed by (Kaneko et al. 1977).

SGA accumulation and association with SNPs at candidate genes

Two accessions with contrasting foliar SGA levels were selected for each of five wild potato species from the NRSP-6 United States Potato Genebank collection. Single seedlings were grown per accession and clonally propagated in three biological repetitions. Two additional clones known to have low (*phu* DH) and high (*chc* 80-1) SGA

levels were also cloned and grown under the same conditions. ANOVA of total SGAs estimated in the leaf tissue showed that there was a statistically significant difference in the SGA accumulation level among accessions (P-value = 0.0023* for nonparametric test and P-value <0.0001* in normal ANOVA). For the most part, the seedlings selected randomly from paired accessions within each species expected to be low or high for SGAs based on the data available from NRSP-6 differed dramatically from each other with regard to total SGAs (Table 6). Only the two *dms* accessions did not differ significantly. The high selections ranged from 5,966 to 55,611 mAU/mg DW with *dms* 78 significantly different from the other five. The low selections ranged from 0 to 5,857 mAU/mg DW with *phu* DH significantly different from the other five. There was some overlap between the lowest three (*dms* 78, *cmm* 26 and *spg* 55) of the high selections and the highest three (*cmm* 7, *dms* 54 and *spl* 16) of the low selections (Table 6).

ANOVAs were conducted on the logarithm of averages of total SGA levels estimated per accession using allelic variants for SNPs identified in exons of candidate gene fragments as the source of variation in order to determine potential associations between informative SNPs and SGA accumulation in the potato accessions (Table 7). Then haplotype clusters were built per SNP genotype, their accessions and total SGA levels, to identify clusters holding mainly low or high SGA accessions within a single SNP genotype. Since the significant SNPs identified by ANOVA were the most likely to be informative, we analyzed them first and used their pattern to identify other informative SNPs. For *HMG1* there were 19 polymorphic SNPs in our potato species, none of which was statistically significant for association with SGA accumulation, nor was there a SNP genotype cluster associated with high or low SGA accessions. From the 20 polymorphic SNPs in *HMG2* a set of three nonsynonymous SNPs could explain different levels of SGA accumulation. HMG2_snp_202 was an informative SNP identified by ANOVA (P<0.001*), which separated *phu* DH, which did not produce SGAs from the other individuals. The different alleles of HMG2_snp_202 would be expected to code for either serine, a polar amino acid, or alanine, a nonpolar amino acid. HMG2_snp_128 and 199 separated the highest SGA producer, *chc* 80-1, from the other samples, with codons specifying either arginine/lysine (both polar) or alanine/serine, with different polarity. One SNP (SQE_snp_220) out of 22 that were polymorphic in *SQE* was potentially

associated with SGA accumulation (P-value <0.001*). This SNP is nonsynonymous, specifying a change from lysine to glutamine, both polar amino acids, and separated *phu* DH from the other samples. *SGT1* had 106 polymorphic SNPs and for 11 the SGA accumulation levels showed significant differences among SNP genotypes with P-values <0.006. Seven were nonsynonymous, specifying amino acids with similar polarity, and four were synonymous. The 11 SNPs (SGT1_snp_171, 210, 249, 250, 255, 408, 415, 435, 612, 666 and 714) were mainly heterozygous genotype in *phu* DH and homozygous for all other individuals, with the exception of two that were also heterozygous in 1-2 other individuals. Three SNPs (SGT1_snp_210, 256 and 549) that were heterozygous for the greatest SGA producer, *chc* 80-1, separated it from all other individuals. SNP SGT1_snp_210 was detected by ANOVA, whereas the other two were found during the cluster analysis. All of them were nonsynonymous, specifying amino acids changes of similar polarity, lysine by asparagine, aspartic acid by glutamic acid and different polarity, glycine by arginine (SGT1_snp_256). In *SGT2* there were 53 polymorphic SNPs and seven were putatively associated with SGA accumulation. We observed only two homozygous genotypes for six of these SNPs in our germplasm panel whereas there were four possible different homozygotes for SGT2_snp_264. Four of these SNPs were nonsynonymous and three synonymous; two of the nonsynonymous changed from polar to non-polar and the other two kept same the polarity. Four SNPs (SGT2_snp_126, 264, 396 and 404) with an exclusive genotype for *phu* DH separated it from other accessions. These SNPs were detected by ANOVA, two were synonymous and two nonsynonymous specifying changes of glutamine by histidine and serine by leucine. Four SNPs (SGT2_snp_10, 11, 76 and 404) exhibited a specific genotype for *chc* 80-1. Two produced amino acid changes in *chc* 80-1 from glutamic acid to tryptophan and valine to isoleucine, and two were synonymous. In general the SNPs at candidate genes did not cluster accessions with similar levels of SGA accumulation, but were able to discriminate no synthesis in *phu* DH and the greatest accumulator of SGAs, *chc* 80-1. Some of these informative SNPs can be used as tag SNPs to screen segregating populations.

SNP chip analysis

A whole genome SNP chip analysis was done on the twelve potato accessions

representing seven species to identify genomic regions putatively associated with SGA accumulation. In this analysis, a monoploid line derived from *chc* 80-1 was used instead of *chc* 80-1 itself. In this analysis, we also used ANOVA to identify the most likely SNPs associated with SGA accumulation. We increased the stringency to identify significant SNPs that clustered at least two accessions with low or high SGA levels, with mean difference between SNP clusters equal to or greater than 5 logarithm units (Table 8). Then the allelic structure of significant SNPs was analyzed by accessions to identify informative SNPs. Thirty-four significant SNPs located on 10 pseudochromosomes were initially associated with total SGA accumulation (Table 8). Cluster analysis identified eight informative SNPs on six pseudochromosomes with homozygous and heterozygous genotypes that discriminated high, intermediate and low levels of SGA accumulation (Figure 2). Two accessions (*phu* DH and *spg* 55) were mostly representative in the cluster of low SGA levels. Of four SNPs located on pseudochromosome 1, one (solcap_snp_c1_5656) at 63.6 Mb mainly clustered in two SNP genotypes (AA and AG) six accessions with the greatest accumulation of SGAs. Five SNPs were found on pseudochromosome 2, of which solcap_snp_c2_30160 at 19.1 Mb grouped in two SNP genotypes (CC and TC) eight accessions with the greatest levels of accumulation, with a gradual decrease from the four with CC to the four with TC. Pseudochromosomes 3, 4 and 5 with two, four and one SNPs did not group the accessions into any logical arrangement. Of five SNPs on pseudochromosome 6, solcap_snp_5775 at 45.3 Mb was informative in grouping eight accessions with the greatest levels of SGA accumulation in homozygous (CC) and heterozygous (TC) genotypes similar to the distribution by solcap_snp_c2_30160 on pseudochromosome 2. Solcap_snp_c2_18573 at 53.2 Mb of five SNPs on pseudochromosome 7 again grouped the 8 accessions with the greatest levels of SGAs under similar patterns. Solcap_snp_c1_1512 one of the two SNPs on pseudochromosome 9, at 28.8 Mb clustered nine accessions; the five with the greatest levels of SGAs had the CC allele whereas the four with lower levels had the TC genotype. The SNP on pseudochromosome 10 did not have any defined cluster associated with high or low SGA accessions. Finally, three informative SNPs were identified out of five on pseudochromosome 11 (solcap_snp_c1_2304, solcap_snp_c2_57429 and solcap_snp_c2_49311) between 4.5 and 8.9 Mb. These SNPs have at least three

accessions with the greatest SGA level, three in intermediate levels and two with the lowest levels grouped in different SNP genotypes. The SNP on pseudochromosome 12 was not informative. The group of eight informative SNPs defined putative haplotypes for low, intermediate and high SGA accumulation (Figure 2). Comparison of hierarchical clustering trees, built in the statistical software JMP® using 3841 SNPs with non-missing data from the potato array and another with the eight informative SNPs, showed that the twelve accessions were grouped by species except for *spl* taxa in the first tree in contrast with the second where they were grouped by SGA levels (Figure 3). The neighbor-joining phylogenetic tree constructed in TASSEL 3.0 (Bradbury et al. 2007) (data not shown) using the 3841 SNPs followed the kinship between samples from the same species found in the analysis in JMP. *Phu* DH the only cultivated species in the germplasm panel was separated from all other samples. Three general clusters were generated based on informative SNPs that grouped by SGA levels for most of the accessions. Two accessions (*cmm* 7 and *sto* 61) with low levels of SGA were located in high and intermediate SGA cluster.

Discussion

Sequence polymorphisms and diversity

Between one and two genomic sequences of fragments in conserved coding regions of five candidate genes were identified for 11 wild potato species accessions and one cultivated accession. Analysis of polymorphisms in a total length of 3.7 Kb from segments of five candidate genes found 354 SNP variations. Nucleotide diversity and dN/dS ratio estimations showed that rates of alterations varied from lower to higher between exons and introns and between genes of primary and secondary metabolism. Even though no significant deviations from the neutral expectation were detected by either analyses, the negative values of Tajima's D test and dN/dS ratios smaller than 1 suggested a tendency toward purifying selection with less stringency in the genes of secondary metabolism. Small sample size, low divergence among lineages and strength of positive selection affect the power of this kind of analysis. The large and negative Tajima's D test in two genes (*HMG1* and *SGT1*) indicated an excess of rare nucleotide polymorphisms with low frequency compared with the expectation under neutral theory which could be explained by effect of genetic hitchhiking selection (Braverman et al. 1995). Nucleotide diversity analyses of genes associated with traits of interest in plants have reported different evolutionary constraints related to gene function and gene segments when comparing between and within gene sequences. Giordani et al. (2011) analyzed eight putative drought response genes in sunflower and found greater variability in the intron regions for one gene and less variability in a pool of genes coding for regulatory proteins than in those coding enzymes involved in cell metabolism. They compared different nucleotide diversity studies of plant genes to support the theory that sequence variability increased from upstream to downstream stress response genes, e.g., in Arabidopsis defense response genes involved in different signaling pathways displayed lower nonsynonymous levels of nucleotide diversity than actual R genes (Bakker et al. 2008). Transient balancing selection seemed to act on resistance genes to maintain high levels of protein variation in intermediate periods of time (Bakker et al. 2006). Evolutionary variation of genes involved in plant metabolic pathways also has been reported. Research of synonymous and nonsynonymous genetic distances of structural

genes in the anthocyanin pathway for species representing different divergent times (monocots and dicots) as well as within the genus *Ipomoea* showed that downstream genes exhibited statistically significant greater divergence rates than upstream genes (Lu and Rausher 2003; Rausher et al. 1999). Similar patterns of variation were found in four genes of the carotenoid biosynthetic pathway for six species (Livingstone and Anderson 2009). Significant positive rank correlation was also found between positions in the pathway and nonsynonymous substitution rates. Since upstream enzymes are intermediary of multiple end products, even slightly deleterious amino acid changes in upstream enzymes could have major deleterious fitness consequences. Greater constraint in upstream genes may also be explained because they are associated with pathway branches and may control pathway flux (Crabtree and Newsholme 1987). In fact a protein interaction network analysis showed that proteins with more interactions tend to evolve more slowly, because a greater portion of the protein is directly involved in its function, and connectivity is positively associated with pleiotropic effects on cellular function (Fraser et al. 2002; Hahn and Kern 2005; Promislow 2004; Rausher et al. 2008). The nucleotide and nonsynonymous variation found in SGA biosynthetic genes could be related not only with stronger purifying selection in upstream genes (*HMG1*, *HMG2* and *SQE*), but also with greater protein variation to direct synthesis of unusual SGAs since the glycoside chain structure adds different toxicity properties to the final compound (Rayburn et al. 1994). Regarding greater diversity found in introns compared with exons, contrasting effects of selection within a gene could explain differential levels of nucleotide diversity on introns. High levels of sequence conservation are expected if regulatory elements are present in the introns (Hare and Palumbi 2003). Enhancer and repressor regulatory element have been found on introns influencing gene expression (Le Hir et al. 2003). In maize domestication was associated with strong selection of non-transcribed gene regions carrying a regulatory element in *teosinte branched1* gene (Wang et al. 1999). In *HMG2* and *SQE* the greater sequence variation in intros was due to not only the number of SNPs but also more frequency of indels as it was found for an apoplastic invertase inhibitor gene in potato (Datir et al. 2012). The lack of regulatory elements in the introns and major constrains on exon region due to pleiotropic effect of these primary metabolism genes could explain greater polymorphism on introns.

Informative SNPs

Analysis of allelic variation of SNPs in candidate genes revealed 24 SNPs putatively associated with accumulation of SGAs. These SNPs mainly discriminated among the absence of SGAs in *phu* DH, the greatest accumulation of SGAs in *chc* 80-1, and other accessions. The whole genome analysis detected eight informative SNPs on six pseudochromosomes (1, 2, 6, 7, 9 and 11) with homozygous and heterozygous genotypes that discriminated high, intermediate and low levels of SGA accumulation. Hierarchical cluster analysis showed that the selected informative SNPs did not follow the phylogenetic pattern found when total polymorphic SNPs were used. In general, three haplotypes could describe each accumulation level. These haplotypes cluster most of the accessions except for *cmm* 7 and *sto* 61 that did not have the allelic structure found for low level of SGAs. Sequence and analysis of polymorphisms in candidate genes is a strategy to find variation associated with phenotypes of interest. Then informative SNPs should be tested in a segregating or association mapping population in order to elucidate the genetic control of allelic variation in the trait of interest. Draffehn et al. (2010) cloned and sequenced five sugar invertase candidate genes involved in cold-induced sweetening of potato tubers from six heterozygous genotypes of potato. The variation was used to screen an association-mapping population of 219 individuals. Allelic sequences associated with chip quality and tuber starch were successfully identified. The allelic polymorphisms of our candidate genes were used to screen a segregating F_2 population of *phu* DH × *chc* 80-1. Allelic sequences of *chc* 80-1 for *HMG2* and *SGT2* were significantly associated with greater levels of SGA accumulation (Chapter 2). A whole genome SNP genotyping of that population identified a locus at 61 Mb on pseudochromosome 1 that corresponded with a previously mapped QTL associated with SGA synthesis and accumulation (Sørensen et al. 2008). Our SNP chip cluster analysis also detected an informative SNP at 63.6 Mb on this pseudochromosome. Together these analyses highlight that major genes acting on synthesis and accumulation of SGA are located on pseudochromosome 1 and the potential of using this informative SNP. In addition, even though no informative SNPs associated with accumulation were detected within candidate gene *HMG2* for our germplasm panel, the SNP chip analysis detected one locus nearby this gene at 19.1 Mb on pseudochromosome 2. This chromosome has

been also associated with a QTL for synthesis of SGAs (Sagredo et al. 2006). Future studies will clarify the significance of the discovered informative SNP as a marker associated with SGA synthesis and accumulation.

Association of traits with allelic variation of SNPs in candidate genes could also lead to identification of sequence polymorphisms directly involved in the phenotypic variation or functional markers (Andersen and Lübberstedt 2003). The fragments analyzed in the candidate genes of this study spanned different protein domains, and the amino acid changes encoded by nonsynonymous SNPs could be directly involved in explaining phenotypes. We made a brief analysis about possible implications of amino acid changes of informative SNP in candidate genes. For *HMG2*, *SQE*, *SGT1* and *SGT2* we identified amino acids specific to *phu* DH and *chc* 80-1. In *HMG2* the amino acid changes were near a conserved tetramerization sequence. Alignment of multiple *HMG* homologues sequences showed that there were some variable base-pairs within an otherwise highly conserved region. Further studies will determine if regulatory sequences in the intron or nonsequenced gene and its contiguous regions could influence phenotype more than the amino acid changes studied here. The SNP of *SQE* was within the squalene monooxygenase conserved domain where multiple polar amino acids were found for homologous sequences. For *SGT1*, the nine amino acid changes located within the multidomain region of the glycosyltransferase family could play a role in the catalytic activity of this protein. However in the F_2 population analysis there was no association of this gene and SGA synthesis or accumulation. Both *phu* DH and *chc* 80-1 had one of two sequences where amino acid changes were concentrated, so the second allelic sequences could also neutralize part of amino acid changes effects. For *SGT2* homozygous sequences were found per clone, and the seven amino acid changes and deletion were within the active site of this protein and could influence its activity. Overall, we amplified fragments in coding regions of these candidate genes, where polymorphisms associated with SGA accumulation not only could alter the protein function but also could be tagging unsequenced segments. The analysis of the segregating population only detected *HMG2* and *SGT2* polymorphisms associated with accumulation of SGAs. *SGT1* had the greatest number of SNPs, nonsynonymous mainly, and we also found evidence of tandem duplication (unpublished data), but there was no association with SGA accumulation. A

previous study showed greater levels of expression of *SGT1* in *chc* 80-1 compared with a *phu* clone (Krits et al. 2007) Further analysis will elucidate if copy number variation of this gene could be affecting SGA accumulation. For *HMG2* the amino acid changes found in the analyzed fragment could not explain gene function effects on increased SGA accumulation of *chc* 80-1, but could tag neighboring sequences such as in the promoter.

Acknowledgements

This research was supported by Research Grant No. IS-4134-08 R from BARD, The United States – Israel Binational Agricultural Research and Development Fund. We thank Suzanne Piovano for technical assistance, Maichel Miguel Aguayo Bustos for programming help, Jonathan Stallings of the Laboratory for Interdisciplinary Statistical Analysis LISA for statistical consultation.

References

Andersen JR, Lübberstedt T (2003) Functional markers in plants. Trends Plant Sci 8:554-560

Arnqvist L (2007) Plant sterol metabolism with emphasis on glycoalkaloid biosynthesis in potato. Dept of Plant Biology and Forest Genetics. Swedish University of Agricultural Sciences, Uppsala, p 50

Bakker EG, Toomajian C, Kreitman M, Bergelson J (2006) A genome-wide survey of R gene polymorphisms in Arabidopsis. The Plant Cell 18:1803-1818

Bakker EG, Traw MB, Toomajian C, Kreitman M, Bergelson J (2008) Low levels of polymorphism in genes that control the cativation of defense response in *Arabidopsis thaliana*. Genetics 178:2031-2043

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140:783-796

Bushway RJ, Barden ES, Wilson AM, Bushway AA (1980) Analysis of potato glycoalkaloids by high performance liquid chromatography. J Food Sci 45:1088-1089

Canonica L, Ronchetti F, Russo G, Sportoletti G (1977) Fate of the 16β-hydrogen atom of cholesterol in the biosynthesis of tomatidine and solanidine. J Chem Soc, Chem Commun 0:286-287

Crabtree B, Newsholme EA (1987) The derivation and interpretation of control coefficients. Biochem J 247:113-120

Choi D, Bostock RM, Avdiushko S, Hildebrand DF (1994) Lipid-derived signals that discriminate wound- and pathogen-responsive isoprenoid pathways in plants: methyl jasmonate and the fungal elicitor arachidonic acid induce different 3-hydroxy-3-methylglutaryl-coenzyme A reductase genes and antimicrobial isoprenoids in *Solanum tuberosum* L. Proc Natl Acad Sci USA 91:2329-2333

Choi D, Ward BL, Bostock RM (1992) Differential induction and suppression of potato 3-hydroxy-3-methylglutaryl coenzyme A reductase genes in response to *Phytophthora infestans* and to its elicitor arachidonic acid. Plant Cell 4:1333-1344

Datir SS, Latimer JM, Thomson SJ, Ridgway HJ, Conner AJ, Jacobs JME (2012) Allele diversity for the apoplastic invertase inhibitor gene from potato. Mol Genet Genomics 287:451-460

Distl M, Wink M (2009) Identification and quantification of steroidal alkaloids from wild tuber-bearing *Solanum* species by HPLC and LC-ESI-MS. Potato Res 52:79-104

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19:11-15

Draffehn A, Meller S, Li L, Gebhardt C (2010) Natural diversity of potato (*Solanum tuberosum*) invertases. BMC Plant Biol 10:271

Edwards EJ, Cobb AH (1996) Improved high performance liquid chromatographic method for the analysis of potato (*Solanum tuberosum*) glycoalkaloids. J Agr Food Chem 44:2705-2709

Felcher KJ, Coombs JJ, Massa AN, Hansey CN, Hamilton JP, Veilleux RE, Buell CR, Douches DS (2012) Integration of two diploid potato linkage maps with the potato genome sequence. PLoS One 7:e36347

Flanders KL, Hawkes JG, Radcliffe EB, Lauer FI (1992) Insect resistance in potatoes - sources, evolutionary relationships, morphological and chemical defenses, and ecogeographical associations. Euphytica 61:83-111

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein Interaction network. Science 296:750-752

Friedman M (2006) Potato glycoalkaloids and metabolites: roles in the plant and in the diet. J Agric Food Chem 54:8655-8681

Gebhardt C, Li L, Pajerowska-Mukthar K, Achenbach U, Sattarzadeh A, Bormann C, Ilarionova E, Ballvora A (2007) Candidate gene approach to identify genes underlying quantitative traits and develop diagnostic markers in potato. Crop Sci 47:S-106-111

Gebhardt C, Valkonen JP (2001) Organization of genes controlling disease resistance in the potato genome. Annu Rev Phytopathol 39:79-102

Ginzberg I, Tokuhisa JG, Veilleux RE (2009) Potato steroidal glycoalkaloids: biosynthesis and genetic manipulation. Potato Res 52:1-15

Giordani T, Buti M, Natali L, Pugliesi C, Cattonaro F, Morgante M, Cavallini A (2011) An analysis of sequence variability in eight genes putatively involved in drought response in sunflower (*Helianthus annuus* L.). Theor Appl Genet 122:1039-1049

Grafius EJ, Douches DS (2008) The present and future role of insect-resistant genetically modified potato cultivars in IPM. In: Romeis J, Shelton AM, Kennedy GG (eds) Integration of insect-resistant genetically modified crops within IPM Programs. Springer Netherlands, pp 195-221

Hahn MW, Kern AD (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol 22:803-806

Hamilton JP, Hansey CN, Whitty BR, Stoffel K, Massa AN, Van Deynze A, De Jong WS, Douches DS, Buell CR (2011) Single nucleotide polymorphism discovery in elite North American potato germplasm. BMC Genomics 12:302

Hare MP, Palumbi SR (2003) High intron sequence conservation across three mammalian orders suggests functional constraints. Mol Biol Evol 20:969-978

Heftmann E (1983) Biogenesis of steroids in Solanaceae. Phytochemistry 22:1843-1860

Johnson AAT, Piovano SM, Ravichandran V, Veilleux RE (2001) Selection of monoploids for protoplast fusion and generation of intermonoploid somatic hybrids of potato. Am J Potato Res 78:19-29

Judo MS, Wedel AB, Wilson C (1998) Stimulation and suppression of PCR-mediated recombination. Nucleic Acids Res 26:1819-1825

Kaneko K, Tanaka MW, Mitsuhashi H (1977) Dormantinol, a possible precursor in solanidine biosynthesis, from budding *Veratrum grandiflorum*. Phytochemistry 16:1247-1251

Keukens EA, de Vrije T, van den Boom C, de Waard P, Plasman HH, Thiel F, Chupin V, Jongen WM, de Kruijff B (1995) Molecular basis of glycoalkaloid induced membrane disruption. Biochim Biophys Acta 1240:216-228

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge [Cambridgeshire]; New York

Kloosterman B, Oortwijn M, Uitdewilligen J, America T, de Vos R, Visser RGF, Bachem CWB (2010) From QTL to candidate gene: Genetical genomics of simple and complex traits in potato using a pooling strategy. BMC Genomics 11(158)

Krits P, Fogelman E, Ginzberg I (2007) Potato steroidal glycoalkaloid levels and the expression of key isoprenoid metabolic genes. Planta 227:143-150

Le Hir H, Nott A, Moore MJ (2003) How introns influence and enhance eukaryotic gene expression. Trends Biochem Sci 28:215-220

Lenz TL, Becker S (2008) Simple approach to reduce PCR artefact formation leads to reliable genotyping of MHC and other highly polymorphic loci--implications for evolutionary analysis. Gene 427:117-123

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451-1452

Lightbourn GJ, Veilleux RE (2007) Production and evaluation of somatic hybrids derived from monoploid potato. Am J Potato Res 84:425-435

Livingstone K, Anderson S (2009) Patterns of variation in the evolution of carotenoid biosynthetic pathway enzymes of higher plants. J Hered 100:754-761

Lorenzen JH, Balbyshev NF, Lafta AM, Casper H, Tian X, Sagredo B (2001) Resistant potato selections contain leptine and inhibit development of the Colorado potato beetle (Coleoptera: Chrysomelidae). J Econ Entomol 94:1260-1267

Lu YQ, Rausher MD (2003) Evolutionary rate variation in anthocyanin pathway genes. Mol Biol Evol 20:1844-1853

McCue KF, Allen PV, Shepherd LV, Blake A, Maccree MM, Rockhold DR, Novy RG, Stewart D, Davies HV, Belknap WR (2007a) Potato glycosterol rhamnosyltransferase, the terminal step in triose side-chain biosynthesis. Phytochemistry 68:327-334

McCue KF, Allen PV, Shepherd LVT, Blake A, Rockhold DR, Novy RG, Stewart D, Davies HV, Belknap WR (2007b) Manipulation and compensation of steroidal glycoalkaloid biosynthesis in potatoes. Acta Hort 745:343-349

McCue KF, Shepherd LVT, Allen PV, Maccree MM, Rockhold DR, Corsini DL, Davies HV, Belknap WR (2005) Metabolic compensation of steroidal glycoalkaloid biosynthesis in transgenic potato tubers: using reverse genetics to confirm the in vivo enzyme function of a steroidal alkaloid galactosyltransferase. Plant Sci 168:267-273

Milner SE, Brunton NP, Jones PW, O'Brien NM, Collins SG, Maguire AR (2011) Bioactivities of glycoalkaloids and their aglycones from *Solanum* species. J Agr Food Chem 59:3454-3484

Moehs CP, Allen PV, Friedman M, Belknap WR (1997) Cloning and expression of solanidine UDP-glucose glucosyltransferase from potato. Plant J 11:227-236

Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res 8:4321-4325

Nema PK, Ramayya N, Duncan E, Niranjan K (2008) Potato glycoalkaloids: formation and strategies for mitigation. J Sci Food Agric 88:1869-1881

Orgell WH, Vaidya KA, Dahm PA (1958) Inhibition of human plasma cholinesterase *in vitro* by extracts of Solanaceous plants. Science 128:1136-1137

Pflieger S, Lefebvre V, Causse M (2001) The candidate gene approach in plant genetics: a review. Mol Breeding 7:275-291

Promislow DEL (2004) Protein networks, pleiotropy and the evolution of senescence. Proc R Soc Lond B 271:1225-1234

Rausher M, Lu Y, Meyer K (2008) Variation in constraint versus positive selection as an explanation for evolutionary rate variation among anthocyanin genes. J Mol Evol 67:137-144

Rausher MD, Miller RE, Tiffin P (1999) Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. Mol Biol Evol 16:266-274

Rayburn JR, Bantle JA, Friedman M (1994) Role of carbohydrate side-chain of potato glycoalkaloids in developmental toxycity. J Agr Food Chem 42:1511-1515

Roddick JG, Rijnenberg AL, Weissenberg M (1990) Membrane-disrupting properties of the steroidal glycoalkaloids solasonine and solamargine. Phytochemistry 29:1513-1518

Sagredo B, Lafta A, Casper H, Lorenzen J (2006) Mapping of genes associated with leptine content of tetraploid potato. Theor Appl Genet 114:131-142

Sanford L, Kobayashi R, Deahl K, Sinden S (1996) Segregation of leptines and other glycoalkaloids in *Solanum tuberosum* (4x) × *Solanum chacoense* (4x) crosses. Am J Potato Res 73:21-33

Sanford L, Sinden S (1972) Inheritance of potato glycoalkaloids. Am J Potato Res 49:209-217

Schaller H (2004) New aspects of sterol biosynthesis in growth and development of higher plants. Plant Physiol Biochem 42:465-476

Shakya R, Navarre DA (2008) LC-MS analysis of solanidane glycoalkaloid diversity among tubers of four wild potato species and three cultivars (*Solanum tuberosum*). J Agr Food Chem 56:6949-6958

Simmonds N (1963) Abbreviations of potato names. Potato Res 6:186-190

Sinden SL, Sanford LL, Osman SF (1980) Glycoalkaloids and resistance to the Colorado potato beetle in *Solanum chacoense* Bitter. Am Potato J 57:331-343

Sørensen K, Kirk H, Olsson K, Labouriau R, Christiansen J (2008) A major QTL and an SSR marker associated with glycoalkaloid content in potato tubers from *Solanum tuberosum* × *S. sparsipilum* located on chromosome I. Theor Appl Genet 117:1-9

Spooner DM (2009) DNA barcoding will frequently fail in complicated groups: An example in wild potatoes. Am J Bot 96:1177-1189

Spooner DM, Salas A (2006) Structure, biosystematics, and genetic resources of potato. In: Gopal, J and Khurana, S M P Handbook of potato production, improvement

and post-harvest management. Haworth's Press, Binghampton, NY, USA pp p. 1-39

Stewart CN, Via LE (1993) A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR applications. BioTechniques 14:748-750

Suzuki M, Muranaka T (2007) Molecular genetics of plant sterol backbone synthesis. Lipids 42:47-54

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595

Tajima F, Nei M (1984) Estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol 1:269-285

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731-2739

The Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. Nature 475:189-195

Thompson AL, Farnsworth BL, Gudmestad NC, Secor GA, Preston DA, Sowokinos JR, Glynn M, Hatterman-Valenti H (2008) Dakota diamond: An exceptionally high yielding, cold chipping potato cultivar with long-term storage potential. Am J Potato Res 85:171-182

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680

Thompson JR, Marcelino LA, Polz MF (2002) Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by ‘reconditioning PCR’. Nucleic Acids Res 30:2083-2088

Tingey WM (1984) Glycoalkaloids as pest resistance factors. Am J Potato Res 61:157-167

Tschesche R, Hulpke H (1967) Biogenese von solanidin aus cholesterin (In German). Z Naturforsch B 22(7):971

Veilleux RE (1990) *Solanum phureja*: anther culture and the induction of haploids in a cultivated diploid potato species., Y. P. S. Bajaj edn. Sringer-Verlag, New York

Wang R-L, Stec A, Hey J, Lukens L, Doebley J (1999) The limits of selection during maize domestication. Nature 398:236-239

Wentzinger LF, Bach TJ, Hartmann MA (2002) Inhibition of squalene synthase and squalene epoxidase in tobacco cells triggers an up-regulation of 3-hydroxy-3-methylglutaryl coenzyme a reductase. Plant Physiol 130:334-346

Wink M (2003) Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective. Phytochemistry 64:3-19

Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591

Yencho GC, Kowalski SP, Kobayashi RS, Sinden SL, Bonierbale MW, Deahl KL (1998) QTL mapping of foliar glycoalkaloid aglycones in *Solanum tuberosum* x *S. berthaultii* potato progenies: quantitative variation and plant secondary metabolism. Theor Appl Genet 97:563-574

Yoshioka H, Yamada N, Doke N (1999) cDNA cloning of sesquiterpene cyclase and squalene synthase, and expression of the genes in potato tuber infected with *Phytophthora infestans*. Plant Cell Physiol 40:993-998

Tables and figures

Table 1 Selected accessions with contrasting levels of SGA

Accession No.	Species	Short Name[1]	Foliar SGA Level	Origin
PI 243503	*Solanum commersonii* subsp. *commersonii*	*cmm* (7)	Very high	Argentina
PI 320266	*Solanum commersonii* subsp. *commersonii*	*cmm* (26)	Low	Uruguay
PI 347760	*Solanum demissum*	*dms* (54)	Very low	Mexico
PI 186562	*Solanum demissum*	*dms* (78)	Very high	Mexico
PI 311000	*Solanum sparsipilum*	*spl* (16)	Very low	Peru
PI 473373	*Solanum sparsipilum*	*spl* (81)	High	Bolivia
PI 458334	*Solanum spegazzinii*	*spg* (55)	Very low	Argentina
PI 205394	*Solanum spegazzinii*	*spg* (74)	High	Argentina
PI 184773	*Solanum stoloniferum*	*sto* (40)	Very high	Mexico
PI 243458	*Solanum stoloniferum*	*sto* (61)	Very low	Mexico
	Solanum phureja	*phu* DH	Very low	
PI 458310	*Solanum chacoense*	*chc* 80-1	Very High	Argentina

Accessions number, SGA level and origin information from NRSP-6 United States Potato Genebank, the SGA level for the last two accessions by HPLC analysis in our laboratory.

[1] Short name assigned to each accession base on standard abbreviations for potato species and in parenthesis a random number.

Table 2 GenBank accessions used for primer design, and primer sequences used to amplify fragments of five candidate genes

Gene	Linkage Group	GenBank Accessions	Primer Sequence	Product Size	Tm (°C)
HMG1	11	L01400, AF110383, U60452, L40938, U51985	f- CGACCTGTTAAGCCTCTATACAC r- GCCACCAGAGACAAAGATAGCCT	877 bp	60
HMG2	2	M63642, AF110383, AB041031, U51985-6, L01400, AF110383	f- TGGTGTCCAAAGGTGTACAAAATG r- ACAGAAATATGGAGGTCCTTGCC	608 bp	59
SQE	4	AY995182, BG123494, CU915722	f- TGGGGTTCGTTGCAGTTTTC r- CAGGGGATAAGAAAGACGTGTACTC	884 bp	60
SGT1	7	U82367, DQ218276, DQ218277, DQ266437, AK323113, AB182385	f- TCCCTTGGACAGTAGATATTGCTG r- TTCCCAATCCCCTAACCTCG	834 bp	58
SGT2	8	U82367, DQ218276, DQ218277, DQ266437, AK323113, AB182385	f- CCTGCGGATGAGAGGAATGC r- CACCAACGGCACCCCAGCG	567 bp	64

Primer direction: forward (f) and reverse (r). Melting temperature of primers = Tm. Linkage group correspond with the location of these gene sequences at potato genome.

Table 3 Number of allelic sequences identified in six wild and one cultivated potato species for five candidate genes within the glycoalkaloid biosynthetic pathway

Gene	Total Unique Allelic Sequences per Locus	*cmm*	*dms*	*spg*	*spl*	*sto*	*chc*	*phu*
HMG1	15	2	2	2	4[1]	2[1]	2	2
HMG2	12	1	2[1]	3[1]	2	2[1]	2	2
SQE	16	3	2	2	3	3	2	1
SGT1	18	1	4[1]	4	3	2[1]	3	2
SGT2	12	1	1[1]	4	3	2[1]	1	1
Total	72	8	11	15	15	11	9	8

[1] These wild species share one identical allelic sequence. Potato species listed using standard abbreviations for potato species, see Table 2.

Table 4 Summary of polymorphic sites and SNPs discovered in five candidate genes at the glycoalkaloid biosynthetic pathway

Gene	N	Total Sites (bp)[1]	Total SNP Sites (S)	Total Number of Variations	SNPs in Noncoding Regions	Synonymous SNPs	Non-synonymous SNPs	Transitions/ Transversion	Indels (Location)
HMG1	22	876	22	22	0	12	10	3.8	1 (Exon)
HMG2	28	571	65	65	44	14	7	2.4	3 (Intron)
SQE	26	887	85	89	66	13	10	2.0	10 (Intron)
SGT1	28	834	106	113	0	49	64	2.3	1 (Exon)
SGT2	27	567	59	65	0	31	34	2.0	2 (Exon)
Average	26.2	754.2	67.4	70.8	22	23.8	25	2.5	3.4
Total	131	3,771	337	354	110	119	125		17

[1]Excluding sites with gaps on the total alignment. N: Total number of analyzed sequences

Table 5 Estimates of codon and nucleotide diversity of sequences of segments in coding regions for five candidate genes in wild potato species

Gene	N	dN	dS	dN/dS	(2Δl) M0 vs. M3 (df = 4)	π Exons	π Total	D-statistics
HMG1	22	0.02	0.06	0.29	-8.14 NS	0.005	0.005	-1.0329 NS
HMG2	28	0.02	0.15	0.15	0.00 NS	0.015	0.032	-0.0561 NS
SQE	26	0.03	0.14	0.24	0.00 NS	0.013	0.022	-0.8299 NS
SGT1	28	0.12	0.28	0.42	-22.91 NS	0.020	0.020	-1.5507 NS
SGT2	27	0.10	0.27	0.39	-18.90 NS	0.029	0.029	0.2592 NS

N: Total number of analyzed sequences; π: nucleotide diversity. NS: Not significant

Table 6 Total SGA accumulation in leaf tissue of twelve accessions determined by HPLC

Low Selection	*n*	Mean (mAU/mg DW)	SD	High Selection	*n*	Mean (mAU/mg DW)	SD	Difference
cmm 7	1	2,619 de		*cmm* 26	1	25,302 abc		22,683
dms 54	3	4,013 d	1,689	*dms* 78	3	5,966 d	1,689	1,953
spg 55	3	1,360 e	2,230	*spg* 74	3	19,892 ab	2,807	18,532
spl 16	3	5,857 cd	792	*spl* 81	3	42,095 ab	26,037	36,238
sto 61	3	1,290 e	283	*sto* 40	3	16,966 bc	5,964	15,676
phu DH	3	0 f	0	*chc* 80-1	3	55,611 a	4,937	

The accessions were named using the species abbreviation and a random number assigned in this study. The amount of accumulation of SGAs is in mili-absorbance units (mAU) of compound per mg of dried weight leaf tissue (DW). Means followed by the same letter are not significantly different at 0.01 α level using student t mean separation analysis. *n*= number of biological repetition per accession.

Table 7 SNPs found within exons of candidate genes putatively associated with SGA accumulation in leaf tissue of potato species

	SNP Genotypes				Number of Samples per Genotype				Mean per Genotype (Logarithmic Units)				
SNP_ID	**G1**	**G2**	**G3**	**G4**	n**1**	n**2**	n**3**	n**4**	**Mean 1**	**Mean 2**	**Mean 3**	**Mean 4**	**Type of SNP**
HMG2_snp_128	GG	AG			11	1			8	10.9			Nonsyn (R/K)
HMG2_snp_199	GG	TT			11	1			8	10.9			Nonsyn (A/S)
HMG2_snp_202	GG	GT			11	1			9	0			Nonsyn (S/A)
SQE_snp_220	AA	CC			11	1			9	0			Nonsyn(K/Q)
SGT1_snp_171	AA	GA			11	1			9	0			Nonsyn (I/M)
SGT1_snp_210	AG	GG	GT		1	10	1		0	8.8	10.9		Nonsyn (K/N/R)
SGT1_snp_249	AA	TA			11	1			9	0			Nonsyn (E/K/D)
SGT1_snp_250	AC	CC			1	11			0	9			Nonsyn (Q/K)
SGT1_snp_255	AT	TT			1	11			0	9			Nonsyn (V/I)
SGT1_snp_256	AG	CC	GC	GG	1	1	3	7	10.9	7.2	5.6	9.2	Nonsyn (A/G/R)
SGT1_snp_408	AA	GA			11	1			9	0			Syn
SGT1_snp_415	GT	TA	TT		1	2	9		0	8.5	9.2		Nonsyn (S/T/A)
SGT1_snp_535	GG	TG	TT		10	1	1		9.2	0	7.2		Nonsyn (A/S/D)
SGT1_snp_549	TA	TG	TT		1	1	10		10.9	7.2	8.1		Nonsyn (D/E/N)
SGT1_snp_612	AA	AC			11	1			9	0			Syn
SGT1_snp_666	CC	TT			11	1			9	0			Syn
SGT1_snp_714	TT	AA			11	1			9	0			Syn
SGT2_spn_10	GG	TT			11	1			8	10.9			Syn
SGT2_spn_11	AA	GG			11	1			8	10.9			Nonsyn (E/W)
SGT2_spn_76	AA	GG			1	11			10.9	8			Nonsyn (V/I)
SGT2_spn_126	AA	GG			1	11			0	9			Syn
SGT2_spn_264	AA	CC	GG	TT	2	1	2	7	9.005	0	9.1	9	Nonsyn (Q/H)
SGT2_spn_396	AA	CC			11	1			9	0			Nonsyn (S/L)

	SNP Genotypes				Number of Samples per Genotype				Mean per Genotype (Logarithmic Units)				
SGT2_spn_404	--	CC	TT		1	10	1		0	8.8	10.9		Syn

SNP ID: candidate gene and bp position in the sequenced fragment. Nature of SNP synonymous = Syn or nonsynonympus = Nonsyn, in parenthesis the amino acid changes in a standard amino acid abbreviation.

Table 8 Significant SNPs associated with SGA accumulation in a whole genome SNP chip analysis of wild and cultivated potato species

SNP_ID	P-value	R^2	G1	G2	G3	*n*1	*n*2	*n*3	Mean 1	Mean 2	Mean 3	Mean difference	Pseudomolecule	Mb Position
solcap_snp_c2_45058	0,023	0,568	AA	AC	CC	3	7	2	10,3	8,6	4,3	6	chr01	4,7
solcap_snp_c2_35520	0,023	0,568	CC	TC	TT	3	7	2	10,3	8,6	4,3	6	chr01	53,6
solcap_snp_c1_5656	0,018	0,592	AA	AG	GG	5	5	2	9,5	8,8	4,0	6	chr01	63,6
solcap_snp_c2_19956	0,028	0,433	TA	TT		9	2		9,2	4,7		5	chr01	65,7
solcap_snp_c2_30160	0,012	0,623	CC	TC	TT	5	5	2	9,8	8,5	4,0	6	chr02	19,1
solcap_snp_c2_41963	0,020	0,579	AA	AG	GG	6	4	2	9,3	8,9	4,0	5	chr02	26,2
solcap_snp_c2_17954	0,011	0,489	AA	GG		10	2		9,1	4,3		5	chr02	34,8
solcap_snp_c2_35212	0,005	0,595	CC	CG		9	2		9,3	4,0		5	chr02	36,3
solcap_snp_c1_2587	0,021	0,579	CC	TC	TT	6	4	2	9,3	8,9	4,0	5	chr02	40,6
solcap_snp_c2_20259	0,021	0,578	AA	AG	GG	3	7	2	9,4	9,0	4,0	5	chr03	20,9
solcap_snp_c2_29678	0,004	0,580	AG	GG		2	10		3,9	9,1		5	chr03	23,4
solcap_snp_c2_55709	0,004	0,573	CC	TC		10	2		9,1	4,0		5	chr04	19,4
solcap_snp_c2_45040	0,004	0,573	AG	GG		2	10		4,0	9,1		5	chr04	47,6
solcap_snp_c2_32550	0,009	0,593	AG	GG		2	8		3,9	9,2		5	chr04	54,4
solcap_snp_c2_34866	0,021	0,575	CC	TC	TT	2	3	7	4,0	9,0	9,2	5	chr04	60,1
solcap_snp_c2_11829	0,004	0,580	CC	TC		10	2		9,1	3,9		5	chr05	3,9
solcap_snp_c2_3108	0,021	0,575	GG	TG	TT	2	4	6	4,0	9,0	9,3	5	chr06	0,8
solcap_snp_c1_15371	0,011	0,489	CC	TT		10	2		9,1	4,3		5	chr06	41,8
solcap_snp_c2_5774	0,014	0,611	CC	TC	TT	6	4	2	9,6	8,5	4,0	6	chr06	45,3
solcap_snp_c2_5775	0,012	0,623	CC	TC	TT	2	5	5	4,0	8,5	9,8	6	chr06	45,3
solcap_snp_c2_29216	0,018	0,588	AA	AG	GG	2	2	8	4,0	8,5	9,3	5	chr06	50,4
solcap_snp_c2_36831	0,018	0,480	AA	GG		9	2		9,0	4,3		5	chr07	2,6
solcap_snp_c2_26003	0,011	0,489	AT	TT		2	10		4,3	9,1		5	chr07	46,8
solcap_snp_c2_12405	0,004	0,573	AG	GG		2	10		4,0	9,1		5	chr07	49,3
solcap_snp_c2_28855	0,004	0,573	AA	AG		2	10		4,0	9,1		5	chr07	52,3

SNP_ID	P-value	R^2	G1	G2	G3	*n*1	*n*2	*n*3	Mean 1	Mean 2	Mean 3	Mean difference	Pseudomolecule	Mb Position
solcap_snp_c2_18573	0,012	0,623	AA	AG	GG	5	5	2	9,8	8,5	4,0	6	chr07	53,2
solcap_snp_c1_4228	0,004	0,573	AG	GG		2	10		4,0	9,1		5	chr09	1,6
solcap_snp_c1_1512	0,013	0,622	CC	TC	TT	6	4	2	9,6	8,4	4,0	6	chr09	28,8
solcap_snp_c2_44210	0,028	0,433	AG	GG		2	9		4,7	9,2		5	chr10	3,8
solcap_snp_c2_13350	0,042	0,505	CC	TC	TT	4	6	2	9,5	8,8	4,3	5	chr11	2,0
solcap_snp_c1_2304	0,022	0,573	CC	TC	TT	2	4	6	4,0	9,2	9,1	5	chr11	4,5
solcap_snp_c2_57429	0,032	0,577	GG	TG	TT	2	4	5	3,9	8,9	9,1	5	chr11	5,0
solcap_snp_c2_49311	0,014	0,611	AA	TA	TT	6	4	2	9,6	8,5	4,0	6	chr11	8,9
solcap_snp_c2_32982	0,041	0,528	CC	GG		6	2		9,6	4,5		5	chr11	9,2

SNP ID: candidate gene and bp position in the sequence. SNP genotypes = G1-G3, number of samples per SNP genotype = *n*1-*n*3, and mean per SNP genotype in logarithmic units = Mean 1-3.

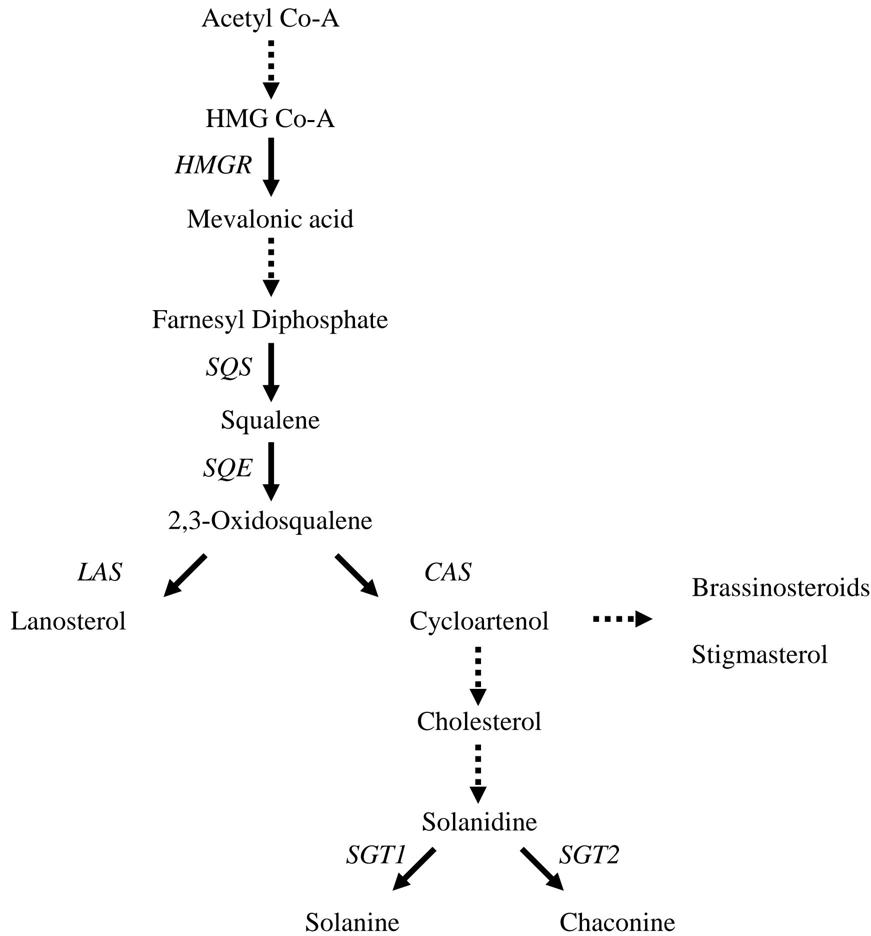


Figure 1 Potato steroidal glycoalkaloid biosynthetic pathway. Dashed line arrows indicate step with multiple enzymatic reactions. Continues line arrows have beside the abbreviation of gene performing the reaction. *HMGR* 3-hydroxy-3-methylglutaryl coenzyme A reductase, *SQS* squalene synthase, *SQE* squalene epoxidase, *LAS* lanosterol synthase and *CAS* cycloartenol synthase, *SGT1* solanidine galactosyltransferase and *SGT2* solanidine glucosyltransferase. Brassinosteroids and stigmaterol are the other end products of sterol biosynthesis in addition to cholesterol.

Accession	Total SGA (mAU/mg DW)	c1_5656	c2_30160	c2_5775	c2_18573	c1_1512	c1_2304	c2_57429	c2_49311
phu DH	0	GG	TT	CC	GG	TT	CC	GG	TT
*sto*61	1289	AG	TC	TC	AG	CC	TT	GG	TA
spg 55	1359	GG	TT	CC	GG	TT	CC	TT	TT
cmm 7	2619	AA	CC	TT	AA	TC	TT	TT	AA
dms 54	4013	AG	TC	TC	AG	TC	TC	TG	TA
spl 16	5856	AA	TC	TC	AG	TC	TT	TG	AA
dms 78	5966	AG	TC	TC	AG	TC	TC	TG	TA
sto 40	16966	AG	TC	TC	AG	CC	TC	TG	TA
spg 74	19277	AG	CC	TT	AA	CC	TC	TT	AA
cmm 26	25302	AA	CC	TT	AA	CC	TT	TT	AA
chc 80-1-4*	30571	AA	CC	TT	AA	CC	TT		AA
spl 81	42095	AA	CC	TT	AA	CC	TT	TT	AA
Pseudomolecule		chr01	chr02	chr06	chr07	chr09	chr11	chr11	chr11
Mb position		63.6	19.1	45.3	53.2	28.8	4.5	5.0	8.9

Figure 2 Haplotypes of informative SNPs identified in a whole genome analysis of wild potato species with different levels of SGA accumulation. Accessions are listed from low to high levels of accumulation. The bottom lines show the pseudochromosome and physical position of informative SNPs in the published potato genome sequence. Boarded by square line putativey low, intermediate and high SGA accumulation haplotypes.

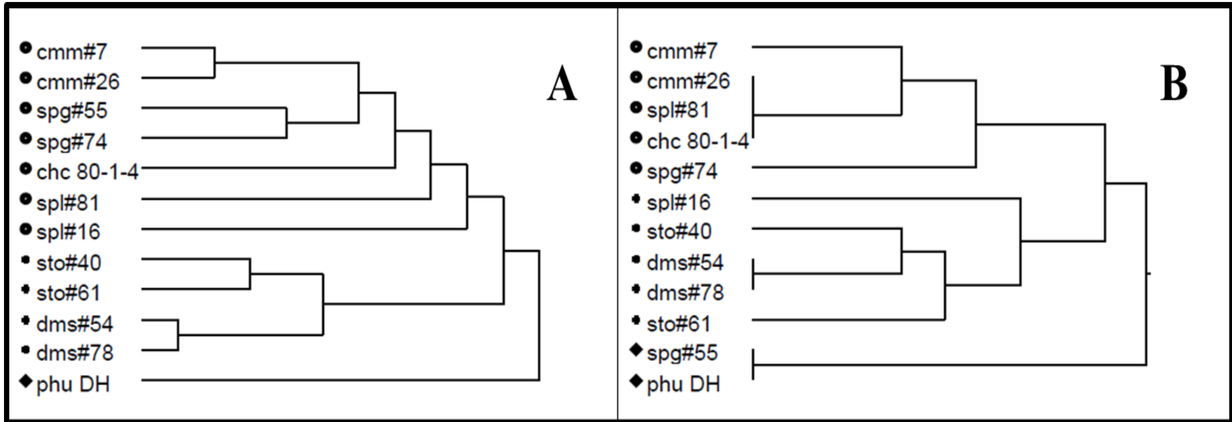


Figure 3 Comparison of hierarchical clusters based on 3841 polymorphic SNPs (A) and eight SGA informative SNPs (B) from whole genome analysis of germoplam panel. A neighbor-joining phylogenetic tree built in TASSEL classified species and accessions with similar patterns as A. Marks on the left of species name corresponded to defined clusters. Most of the accessions with high SGA levels cluster on big dot (•), intermediate levels on small dot, and low on diamond. However *cmm* 7 and *sto* 61 with low SGA levels were not in the right cluster.

4. Conclusions

A candidate gene approach and a whole genome analysis were successfully used to elucidate genes associated with synthesis and accumulation of SGAs as well to identify new putative candidate genes for future studies. *Chc* 80-1 alleles for *HMG2*, *SQE* and *SGT2* genes were associated with synthesis and accumulation of leptines, and the interaction of *HMG2* and *SGT2* with accumulation of α-solanine and α-chaconine in a segregating F_2 *phu* DH × *chc* 80-1 population. Multiple amino acid changes that altered the predicted protein sequence of *SGT2* were found, future analysis could explain their functional implication since they were located within the active site of this enzyme. For *HMG2* the amino acid changes were in variable sites suggesting little evidence for functional implications; the polymorphisms probably tagged unsequenced regions responsible for increasing SGA levels in *chc* 80-1. In the whole genome analysis, loci with putative candidate genes on pseudochromosomes 1 and 6 explained the synthesis of α-solanine and α-chaconine, the most common SGAs in potato species. Loci on pseudochromosomes 1, 2 ,7 and 8 were associated with synthesis of leptines, but did not explain the segregation ratios for synthesis of leptines in the F_2 population. Various candidate genes, on loci within seven pseudochromosomes associated with SGA accumulation, were postulated. Future genetic studies using these candidate genes and the eight informative SNPs found in the analysis of potato germplasm panel will help to identify functional genes in the biosynthesis and regulation of accumulation of SGAs.

Genetic diversity analysis in partial sequences of five candidate genes in seven potato species suggested a trend of purifying selection in the sequenced coding regions. Genes of primary metabolism (*HMG1*, *HMG2* and *SQE*) exhibited stronger selection pressure than those of secondary metabolism (*SGT1* and *SGT2*). This pattern was expected between genes with different regulatory, pleiotropic, and plasticity effects on the metabolic and stress response pathways. Future population genetic studies in a larger population could clarify the selective constraints that can be affecting SGA synthesis and accumulation in potato species.