Brigham Young University
BYU ScholarsArchive

All Theses and Dissertations

2009-12-02

Genetic Dissection of Triterpenoid Saponin Production in Chenopodium quinoa Using Microarray Analysis

Derrick James Reynolds
Brigham Young University - Provo

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

 Part of the Animal Sciences Commons

BYU ScholarsArchive Citation

Reynolds, Derrick James, "Genetic Dissection of Triterpenoid Saponin Production in Chenopodium quinoa Using Microarray Analysis" (2009). *All Theses and Dissertations*. 1974.
https://scholarsarchive.byu.edu/etd/1974

This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.

Genetic Dissection of Triterpenoid Saponin Production in *Chenopodium quinoa*

Using Microarray Analysis

Derrick James Reynolds

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Peter J. Maughan
Joshua A. Udall
Steven G. Wood

Department of Plant & Wildlife Sciences

Brigham Young University

December 2009

Copyright © 2009 Derrick James Reynolds

All Rights Reserved

ABSTRACT

Genetic Dissection of Triterpenoid Saponin Production in *Chenopodium quinoa*

Using Microarray Analysis

Derrick James Reynolds

Department of Plant & Wildlife Sciences

Master of Science

Quinoa (*Chenopodium quinoa* Willd.) is an important food crop for subsistence farmers in the Altiplano (high plains) of Peru, Bolivia, and Argentina. Saponins are part of a diverse family of secondary metabolites that are found in high concentrations in the pericarp of many varieties of quinoa. Due to their bitter taste and anti-nutritive properties, saponins must be removed before the quinoa grain is consumed. There are 'sweet' varieties of quinoa that have significantly reduced levels of saponin. Previous research suggests saponin production is controlled by a single locus. The major objective of this research was to elucidate the genetic components in the saponin biosynthesis pathway. Thus, we report the development and annotation of the first large scale expressed sequence tag (EST) collection for quinoa based on Sanger and 454 pyrosequencing of maturing seed tissue expressing saponins. Sanger sequencing produced 18,325 reads with an average read length of 693 nucleotides, while 454 GS-FLX pyrosequencing generated 295,048 reads with an average read length of 202 nucleotides. A hybrid assembly of all sequences generated 39,366 unigenes, consisting of 16,728 contigs and 22,638 singletons. Repeat sequence analysis of the unigene set identified 291 new microsatellite markers. From the unigene set, a custom microarray was developed and used to assay transcriptional changes in

developing seeds of saponin-containing and saponin-free quinoa lines. The microarray consisted of 102,834 oligonucleotide probes representing 37,716 sequences of the unigenes set. Three different statistical comparisons, based on comparisons of 'sweet' vs. 'bitter' seed tissue at two developmental stages, were assayed on the custom array. Using a p-value cutoff threshold of 0.01, we identified a list of 198 significantly differentially expressed candidate genes common to all three comparisons. We also identified a list of candidate genes (p-value ≤ 0.05) that are known to be associated with identified triterpenoid (saponin) biosynthetic pathways that were differentially expressed in all three comparisons. Included in this list are candidate genes that share homology to cytochrome P450s (20), cytochrome P450 monooxygenases (10), and glycosyltransferases (49) suggesting that transcriptional differences in the saponin biosynthesis pathway possibly responsible for the absence or presence of saponin in quinoa are determined after the formation of the β-amyrin skeleton. These candidate genes are suggested for use in future studies in the production of saponin in quinoa.

Keywords: Chenopodium quinoa, saponins, EST assembly, microarray, 454 sequencing, SSRs

ACKNOWLEDGEMENTS

I would like to express my appreciation to all those who assisted me in this project. Thanks to Dave Elzinga for being a great friend and assistant and for the incredible amount of time spent in front of a computer screen, teaching me the secrets of bioinformatics. I give my thanks to all who spent many laborious hours in the greenhouse harvesting individual immature seeds the size of pin heads including, Jimena Alvarez, Chris Nye, Ana Eguiluz, Amalia Vargas, Rozaura Hall and Marcus Soliai. I would also like to thank for their encouragement and help with side-projects, in addition to the aforementioned persons, Caryn Hattingh and Kedra Foote among many others. Special thanks to my graduate advisor, Dr. Maughan, for all of the time and effort he has put forth to help me and this project succeed including subtle reminders to "get to work." I would also like to thanks my committee members, Dr. Udall and Dr. Wood, whose academic knowledge and advice were great resources to me. I would also like to thank my family and friends for being a support to me throughout my educational experience. I would especially like to thank my beautiful and wonderful wife, Rachel, who always believed in me, assisted me whenever possible and encouraged me to push past failed reactions, and without whom I never would have considered applying to this program and would have missed out on all of the great experiences and relationships I have gained here in the BYU Genetics Lab. Grants from the McKnight Foundation, Holmes Family Foundation, and Ezra Taft Benson Agriculture and Food Institute supported this research.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

LIST OF FIGURES

Chapter 1

CHAPTER 1: GENETIC DISSECTION OF TRITERPENOID SAPONIN PRODUCTION IN *CHENOPODIUM QUINOA* USING MICROARRAY ANALYSIS

Introduction

Quinoa (*Chenopodium quinoa* Willd) is a putative allotetraploid (2n = 4x = 36) member of the family Amaranthaceae (alt. Chenopodiaceae) which contains the economically important plant species spinach (*Spinacia oleracea* L.) and sugar beet (*Beta vulgaris* L.). It is an important crop for subsistence farmers in the Altiplano (high plains) of Peru, Bolivia, and Argentina. Anciently, quinoa was honored and cultivated extensively throughout the Incan Empire (D'Altroy and Hastorf, 1984). The Spanish conquest of the Americas led to the suppression of quinoa cultivation due to its cultural and religious importance (Cusack, 1984). As a result, quinoa production declined significantly following the Spanish conquest of the Americas. Recently, quinoa has seen a revival in interest and usage due in part to recent studies that call attention to the nutritive properties of quinoa, including an excellent balance of carbohydrates, lipids, and proteins as well as an ideal balance of essential amino acids for human nutrition (Chauhan et al., 1992; Coulter and Lorenz, 1990). In addition to its high nutritional value and in light of global climate change, quinoa has the potential to be an future crop of global importance as it is also well-adapted to many abiotic stresses (Prado et al., 2000; Vacher, 1998). For example, salares ecotypes are adapted to the highly saline and drought affected soils of the salares (salt flats) region of the Bolivian Altiplano. Indeed, few plant species, particularly cultivated ones, can rival quinoa's combination of resistance to drought, frost, and soil salinity (Jacobsen et al., 2003; Risi and Galwey, 1984; Sanchez et al., 2003)

Saponins are a major family of secondary metabolites that occur in a wide range of plant species. Saponins are usually triterpenoid glycoalkaloid molecules with one or more sugar chains (Fenwick et al., 1991). They are commonly characterized as soap-like substances that exhibit a wide range of properties and therefore are regarded as important biological compounds. The

multiplicity of properties and functions of saponins are due to the variety of backbone and sugar side chain components (Dini et al., 2001). Saponins are believed to play an important role as a natural pesticide; acting as bitter compounds that deter insects and avian predation in quinoa (Risi and Galwey, 1984). Unfortunately, the same properties also have anti-nutritional properties in humans. Saponin molecules easily complex with sterols in lipid membranes resulting in loss of membrane integrity (Morrissey and Osbourn, 1999; Osbourn, 2003). This disruption of membranes interferes with molecule and protein transport, as well as the proper absorption of essential minerals and nutrients (Modgil and Mehta, 1993; Onning et al., 1996).

Madl et al. (2006) reported that at least 87 different triterpene saponins are present on the quinoa seed. There are two main seed types of quinoa, namely 'sweet' and 'bitter' types. The 'bitter' quinoas produce saponin on their seed coats and require an additional saponin-removal step during seed processing prior to human consumption. The 'sweet' varieties of quinoa have significantly reduced levels of saponin, are non-bitter and do not decrease palatability (Masterbroek et al., 2000). The presence and concentration of saponin can be measured by an afrosimetric test developed by Koziol (1991) in which quinoa seeds are agitated in deionized water and the resultant characteristic foam is measured. Preliminary afrosimetric testing of 'bitter' quinoa immature seed places the beginning of measurable saponin production sometime after the 'aqueous' (~14 days post-anthesis (dpa)) stage of development but before the 'milky' (~21dpa) stage. Ricks et al. (2005) showed that the production of 'bitter' saponins in quinoa is controlled by a single dominant locus. The absence of saponin, while normally detrimental to crop yield due to insect and avian predation, is a desirable characteristic on the southern Altiplano where avian predation is not a concern. While there is no effect on the nutritional quality of quinoa after saponin removal (Chauhan et al., 1999), the removal process requires large amounts

of clean water and/or machinery – both of which are resources that are not readily available to the average subsistence farmers who grow quinoa on the Southern Altiplano.

While a potentially important new crop, very little DNA sequence information and thus few genomic tools (e.g. genetic markers, dense linkage maps, microarrays, etc.) are currently available to help facilitate genomic research and modern breeding of quinoa. Only a single previous effort to sequence transcribed genes has been reported in quinoa and it resulted in the depositing of only 424 Expressed Sequence Tags (ESTs) in the publically accessible NCBI GenBank database (Coles et al., 2005). EST sequences are partial sequences from transcribed cDNA sequences that reflect expressed genes in a given tissue type at a specific point of development. Made publically available, EST sequences facilitate gene discovery, genetic marker development, and homology searches with sequences from other organisms. Collections of these sequences can also provide researchers with a rapid and cost effective tool to analyze transcriptome changes via DNA microarray analysis. Since a major objective of our research is to elucidate the genetic components in the saponin biosynthesis pathway, we report here the i) development and annotation of the first large scale EST collection for quinoa based on Sanger and 454 sequencing technologies; ii) development of a custom microarray to assay gene expression in developing seeds of quinoa; and iii) the transcriptional variation between ‘sweet’ and ‘bitter’ quinoa varieties at two different stages of development. From this research, we identified a narrowed list of candidate genes that may be specifically associated with the saponin biosynthetic pathways and therefore represent candidate genes for future studies of the genetic underpinnings of saponin biosynthesis in quinoa.

MATERIALS AND METHODS

Plant Material and RNA isolation. A cDNA library was developed from seed tissue of the 'bitter' Peruvian valley quinoa breeding line '0654', obtained from A. Bonifacio at the Foundation for the Promotion and Investigation of Andean Products (PROINPA), La Paz, Bolivia. All '0654' plants were grown at 25 °C with 16-h day lengths in greenhouses at Brigham Young University, Provo, Utah. '0654' seed tissue was harvested at five distinct developmental stages, (8 days post anthesis (dpa), 16dpa, 24dpa, 32dpa, and 40dpa) and was immediately frozen in liquid nitrogen and stored at −80 °C. Total RNA was extracted from frozen plant tissue using LiCl precipitation (Puissant and Houdebine, 1990). Total RNA quantity was measured on a NanoDrop® ND-1000 spectrophotometer v. 3.30 (NanoDrop® Technologies Inc, Wilmington, DE, USA) and RNA integrity was verified using an Agilent 2100 Bioanalyzer and RNA NanoChip with 2100 Expert software (Agilent Technologies, Santa Clara, CA).

Sequencing of ESTs

Sanger Sequencing. A bulked sample of '0654' seed RNA was created by adding equimolar amounts of total RNA from each of the five developmental seed tissues (see above). The RNA bulk was used for double-stranded cDNA synthesis and amplification using a Clontech SMART First-Strand cDNA Synthesis kit (Clontech Laboratories, Mountain View, CA). The resultant double-stranded cDNA was normalized using a double-stranded nuclease kit (Evrogen, Moscow, Russia) prior to cloning. Ten thousand recombinant clones were picked robotically, plasmid extracted, and sequenced bi-directionally via standard Big-dye cycle sequencing at the Arizona Genomics Institute (Tucson, AZ).

454 Sequencing. In addition to Sanger sequencing as described above, 454 pyrosequencing of the seed transcripts was also performed on a Genome Sequencer FLX (454 Life Sciences, Branford,

CT), at The Genome Center at Washington University (St. Louis, MO). Total RNA representing equimolar concentration of the five seed developmental stages (see above) of the 'bitter' breeding line'0654' were bulked and shipped to The Genome Center at Washington University where a 454 pyrosequencing amenable cDNA library was produced and sequenced. Briefly, total RNA was reverse transcribed using a Clontech SMART First-Strand cDNA kit (Clontech Laboratories, Mountain View, CA) with modified adaptors to allow for *Mme*I excision (5' Smart Oligo {5'- AAGCAGTGGTAACAACGCATCCGACGCrGrGrG-3'}; 3' Oligo dT SmartIIA {5'- AAGCAGTGGTAACAACGCATCCGACTTTTTTTTTTTTTTTTTTTTTTVN-3'}; NEW SmartIIA {5'-Biotin-TEG-AGCAGTGGTAACAACGCATCCGAC -3'}. The cDNA was normalized using a double-stranded nuclease (DSN) kit (Evrogen, Moscow, Russia). Lastly a *Mme*I digestion of the normalized cDNA was performed to excise the 5' and 3' modified SMART adaptors. The resultant cDNAs were sequenced according to standard 454 protocols using a Roche-454 GS FLX instrument and FLX reagents (Branford, CT).

EST Assembly and Annotation. The Sanger sequences, the 424 quinoa EST sequences previously deposited in GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) and the 454-pyrosequencing sequences were assembled *de novo* using the Roche Newbler assembler (v. 2.0.00.20; Branford, CT) with the minimum overlap length set to 40 bp, the minimum overlap identity set to 90% and the Large Genome and Iterative Assembly options turned on. All reads were trimmed for vector contamination prior to assembly using a trim file containing the SMART modified adaptor sequences. The resulting assembly of contigs and singletons, collectively referred to as the 'unigene' set, was analyzed for gene sequence homology and microsatellite content. Putative gene homologies were assigned to the unigene set using BlastX (build 2.2.18) searches against the NCBI non-redundant protein database (subset: *Viridi Plantae; E-value >1e-05*). Unigenes without

a significant BlastX homology to the *Viridi Plantae* subset database were then compared to the entire non-redundant (nr) database using BlastX (*E-value >1e-05*). ESTScan was used to identify the coding frames and to generate putative protein sequences (Lottaz et al., 2003). The matrix for ESTScan was based on *Arabidopsis thaliana* coding and non-coding sequences. Custom PERL scripts were used to place the unigenes into sense-strand orientation based on the combined results of the BlastX searches and ESTScan scores (Stajich et al., 2002). Scores were calculated for each possible coding frame, with ESTScan results being weighted twice as heavily as BlastX hits. Gene ontologies were added to the unigenes using Blast2GO, a program for high-throughput functional annotation and data mining (Gotz et al., 2008). GO accessions were mapped to GO terms according to the non-redundant classifications of molecular function, biological process and cellular component (http://www.geneontology.org/).Microsatellites were identified using MISA (Thiel et al., 2003). Microsatellites were selected if they met a minimum motif repeat threshold of 8, 6, 5 for the di- tri- and tetra-nucleotide motifs, respectively. Microsatellites that were separated by less than 100 bp were classified as compound microsatellites.

Microarray design. Microarrays were designed based on the custom gene expression 2X105K platform from Agilent (Santa Clara, CA). Each slide consisted of the 2 arrays, each with 105,072 possible features, including 1,325 Agilent controls and 103,747 user-defined 45-60 base pair oligonucleotide probes. The negative and positive controls include spike-in control probes for an external RNA reference. All quinoa seed unigene sequences (including those without a BlastX hit or ESTScan data) were submitted to eArray v5.4 (Agilent Technologies, Santa Clara, CA) for probe design. Probes were selected using a probe melting temperature specificity of 80 °C and a length of 45-60 nucleotides optimized to the melting temperature. Up to three probes were selected using the Best Distribution Methodology option in eArray, a methodology that favors

even distribution of the probes across the sequence. The quinoa seed unigenes were also used as the reference transcriptome file that is used to define most or all transcripts within the target transcriptome. eArray compares each newly designed probe to this file to ensure a maximum amount of unique probes, and warn of any potential cross-hybridization that may occur due to probes with high sequence similarity.

Microarray sample preparation and hybridization. An F_2 population segregating for saponin production was created from a cross of a 'sweet' (non-saponin) Bolivian breeding line 'LP' and a 'bitter' (saponin-containing) Peruvian breeding line '0654'. F_2 individuals were advanced by single seed decent to the $F_{2:3}$ generation. The saponin content of 12 $F_{2:3}$ progeny plants were used to determine the genotype of each F_2 plant using a previously described afrosimetric method (Koziol, 1991). F_2 individuals that did not segregate for saponin content in the $F_{2:3}$ generation were classified as homozygous dominant ($saponin^+$; 22 F_2 individuals) or homozygous recessive ($saponin^-$; 21 F_2 individuals), respectively, at the 'bitter' saponin production (BSP) locus. F_2 individuals that segregated for the presence of saponin content in the $F_{2:3}$ generation were classified as heterozygous (49 F_2 individuals). For our microarray analysis, we collected immature seed tissue from the $F_{2:3}$ generation of only the homozygous individuals ($saponin^+$ and $saponin^-$) at two distinct developmental stages, specifically at the aqueous stage (~14 dpa), and at the milky stage (~21 dpa). Immature seeds were dissected from the seed head and were immediately frozen in liquid nitrogen and subsequently stored at −80 °C.

Equal amounts of frozen seed tissue from each homozygous $F_{2:3}$ plant was randomly chosen and assigned to a bulk RNA extraction based upon its genotypic designation (homozygous $saponin^-$ or homozygous $saponin^+$) and development phase (aqueous or milky). Thus, four experimental treatments were derived, specifically: i) $Saponin^+$/aqueous, ii) $Saponin^-$/aqueous, iii)

Saponin^{+}/milky and iv) Saponin^{-}/milky. Total RNA was extracted from bulked seed tissue for each treatment using a Qiagen RNeasy Plant extraction kit (Chatsworth, CA). In order to create biological replicates for statistical analyses and to reduce batch effects, RNA was extracted independently and simultaneously four times for all four sample types resulting in a total of 16 RNA extractions. RNA integrity was verified using an Agilent 2100 Bioanalyzer and Agilent 2100 Expert software (Agilent Technologies, Santa Clara, CA).

Total RNA (1.25µg-2.5µg) and external control RNA (Agilent Two-Color RNA Spike-In Kit) were reverse-transcribed to create cDNA from which cRNA was simultaneously synthesized and labeled using an Agilent Quick Amp Labeling Kit according to Agilent's recommended protocols with Cyanine-5-CTP (Cy5)and Cyanine-3-CTP (Cy3) (Agilent Technologies, Santa Clara, CA). cRNA quantity and the efficiency of the labeling was estimated by calculating the Cy-3 (550nm) and Cy-5 (650nm) fluorescence specific activity as measured on a NanoDrop® ND-1000 spectrophotometer v. 3.30 (NanoDrop® Technologies Inc, Wilmington, DE, USA). Microarray hybridizations were performed according Agilent recommendations, where 750 ng of each of two fluorescently (Cy3 and Cy5) labeled cRNA samples were hybridized to the quinoa seed microarray in a rotisserie hybridization oven set at 65 °C for 17h using the Agilent GE Hybridization Kit. The arrays were washed using Agilent's Gene Expression Wash protocols, including the optional acetonitrile and Agilent Stabilization and Drying Solution washes to prevent ozone-mediated fluorescent signal degradation. Arrays were scanned with an Agilent Microarray Scanner G2505B (Agilent Technologies, Santa Clara, CA). Array spot intensities and quality control features were determined using the Extended Dynamic Range option and Agilent's Feature Extraction Software (v 10.5.1.1). Array quality was determined by analysis of control features as well as spike-in controls (Agilent Two-Color RNA Spike-In Kit). Agilent's Feature

Extraction software automatically normalizes within arrays, subtracting background fluorescence, and flagging any outliers.

Statistical analysis. An analysis of variance (ANOVA) was performed on the means for each treatment group using the signal intensities processed for each probe by Agilent Feature Extraction Software (v. 10.5.1.1). The following model was applied to each microarray probe:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

Where *y* is the gene expression and μ is the group mean for the ij^{th} factor group, and ε indicates the random error. Subscripts indicate the factor level, where *i* indicates the saponin level (saponin^{+} or saponin^{-}), *j* indicates the developmental stage (milky or aqueous) and *k* indicates the observation number per group (for example, if there were 4 saponin^{+} milky observations, the first one would have a *k* index of 1, etc.). Using this model, four different comparisons were calculated: **1)** [SM - SA] vs. [NSM - NSA] ; **2)** SM vs. NSM; **3)** SM vs. [SA + NSM + NSA], where S, NS, A, and M denote Saponin^{+}, Saponin^{-}, Aqueous and Milky, respectively.

RESULTS & DISCUSSION

EST sequencing and assembly. Sanger sequencing of the '0654' quinoa seed tissue resulted in 18,325 reads with an average read length of 693 nucleotides and a total length of 12.7Mb with 86.6% of the bases having a quality score greater than 20. The percentage of bases called as "N" was 0.002%. 454 pyrosequencing of the seed tissue produced 295,048 reads with an average length of 204 nucleotides and a total length of 60.2Mb with 93.7% of the bases having a quality score greater than 20. The percentage of bases called as "N" was 0.03%. Trimming of the 454 sequences reduced the average read length to 202 nucleotides and reduced the total number of bases sequenced to 59.7Mb. Most of the trimmed base pairs in 454 sequences were a result of

incomplete digestion and removal of SMART oligonucleotide adapter sequences that were incorporated into the cDNA in preparation for sequencing.

De novo assembly of the 454 sequences, Sanger sequences (including the 424 quinoa ESTs previously deposited in GenBank) resulted in the identification of 39,366 unigenes, consisting of 16,728 contigs and 22,638 singletons. Of the 295,048 454 reads, 273,117 (92.6%) assembled into contigs. Similarly, 90.9% (16,668) of Sanger reads also assembled into contigs. The average contig length was 472 nucleotides and while the average number of reads per contig was 17.3, many contigs consisted of more than 100 reads (Table 1). Some contigs were constructed of both types of reads (3,891); however, several contigs were composed solely of Sanger (1,643) or 454 sequences (11,194). There are a number of possible reasons that might explain these observations, including the sheer volume of 454 reads used in the assembly when compared to number of Sanger reads. Additionally, Sanger sequences exhibit a 5' and a 3' bias in sequencing, whereas cDNA sequence by 454 pyrosequencing is randomly sheared via nebulization allowing for sequencing of all regions of a transcript. Differences in sample preparation could also be a contributing factor. Indeed, while both the 454 and Sanger sequences were prepared from the same total RNA sample, the reverse transcription and normalization procedures varied between the two cDNA library preparations. Singletons consisted of 2,078 and 20,560 Sanger and 454 sequences, respectively. For many of the same reasons listed above, we expected to see the disparity between the numbers of singletons unique to each method of sequencing.

Functional annotation of unigenes by BlastX. Putative gene homologies were assigned to the unigene set using BlastX searches against the NCBI non-redundant protein database. Homologous sequences were found for 45% of all unigenes. Of these unigenes, 59% of all contigs had

homology to sequences in the nr database. Homology was assigned to 54% of contigs made up of only 454 reads, while 58% of contigs composed of only Sanger sequences had BlastX. Contigs composed of both 454 and Sanger sequences had a much higher percentage (74%) of BlastX hits than contigs composed of reads from a single sequencing technology. Singletons had a much lower percentage (37%) of homologous sequences, with 72% of Sanger singletons and 34% of 454 singletons having BlastX hits. Functional annotation of quinoa unigenes also revealed significant homology to other plant species, the most common being rice (*Oryza sativa* L.), thale cress (*Arabidopsis thaliana* L.), grape (*Vitis vinifera* L.) and corn (*Zea mays* L.) (Fig. 1) - a result based likely on the volume of DNA sequences available for these species in the NCBI databases and not shared phylogeny.

GO annotation of unigenes. Gene Ontology (GO) is a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data. The Blast2GO suite, a program for high-throughput functional annotation and data mining (Gotz et al., 2008) was used to assign GO annotations to the assembly using the BlastX information, producing GO annotations for 9,536 unigenes (24.2%). A large percentage (75.8%) of the unigenes is entirely unique to quinoa with no sequence homology to other reported sequences. Unigenes were placed into three categories, including biological process (BP), molecular function (MF), and cellular component (CC). The most common GO terms, as determined by the number of sequences, in the biological process category (Fig. 2), were cellular metabolic process (18.7%), primary metabolic process (17.4%), macromolecule metabolic process (12.4%) and biosynthetic process (6.6%). GO terms for cellular component category (Fig. 3) consisted of intracellular (18.7%), intracellular part (18.5%), intracellular organelle (17.3%) and membrane-bounded organelle (16.2%). The most numerous GO terms in the molecular function category (Fig. 4) were transferase activity (12.9%),

nucleotide binding (11.2%), protein binding (10.6%) and hydrolase activity (10.3%). The GO annotated unigenes cover a broad range of GO categories suggesting that the collection of unigenes is representative of the overall *Chenopodium quinoa* seed transcriptome.

Identification of Microsatellite Markers. The utility of EST sequencing goes beyond gene expression studies. Indeed, microsatellites found within the EST sequences can serve as valuable genetic markers. In the quinoa unigene collection, a total of 291 microsatellites were identified, of which 194 are suitable for primer design (i.e., contain sufficient upstream and downstream sequence information for primer design; Table 2). Previous studies for microsatellite development in quinoa have yielded 402 microsatellites, with the most common repeats observed being GA (49%), CAA (35.6%), and AAT (12.9%) (Jarvis et al., 2008; Mason et al., 2005). In the unigene-based microsatellites, AAT (13.4%) is the most frequent repeat, followed closely by AC (12.7%), AAG (11.3%), CAA (11.0%) and GA (10.0%). The repeats in these unigene-based microsatellites appear to be much more diverse than the genomic microsatellites previously reported, however this is most likely due to the specific creation of GA, CAA, and AAT enriched libraries by Jarvis (2008) and Mason (2005) more so than an actual difference in repeat frequency throughout the genome. EST-based microsatellites in cotton (*Gossypium hirsutum* L.) has also been reported with similar diversity of repeats (Han et al., 2006). These new microsatellites markers represent a potential 48% increase in total number of microsatellites now available for quinoa. The current genetic marker linkage map of quinoa consists of 275 genetic markers, including 200 microsatellite markers, spread across 38 linkage groups (Jarvis et al., 2008). The fact that 38 linkage groups were identified, while there are only 18 chromosome pairs in quinoa suggests that additional markers are needed to refine the map and coalesce linkage groups with chromosome number.

An additional advantage of EST-based microsatellites is the high transferability between related species. One study in tall fescue (*Festuca arundinacea* Schreb.) tested 157 EST-microsatellite primer pairs on seven related grass species with nearly 92% of the primer pairs producing characteristic simple sequence repeat (SSR) bands in at least one of the species tested (Saha et al., 2004). Thus, the transportability of these EST-based markers makes them potentially valuable for numerous other under researched crop and weed species related to quinoa such as cañahua (*C. palidicale* Heller), fat hen (*C. album* L.), taak, bithus or khan (*C. giganteum* D.Don) and the cross compatible Nuttal's goosefoot, huautzontle or quelite (*C. berlandieri* Moq. var. *nuttaliae*) (Sederberg, 2008).

Microarray Design. Strand selection of 454 pyrosequencing of cDNA is mostly random, while microarray analysis is dependent on the hybridization of cRNAs to reverse complementary sequences. Thus the correct coding strand must be identified prior to oligonucleotide production. Using BlastX and ESTscan results we determined the coding frame for 15,550 unigenes. The directionality of 8,347 ESTs was either confirmed or reoriented based on the entire nr database BlastX results, while the remaining 15,379 ESTs were entirely unique with no BlastX or ESTScan results.

Using the sense-strand oriented unigene set and Agilent's eArray v5.4 we designed 100,443 probes from 38,124 of the 39,366 unigenes. An additional 138 probes were designed from 45 genes from GenBank related to previously described saponin pathways in other plants, and a nearly full-length cDNA sequence of a *C. quinoa* β-amyrin synthase gene that had been previously sequenced in our laboratory. Up to three probes were able to be designed for 14,842 of the unigenes without BlastX or ESTScan sense strand data. One of these three probes was reverse complemented to ensure at least one probe per unigene was arrayed in sense-strand orientation.

Additionally, if only one unique probe was designed from the unigenes without BlastX or ESTScan sense data, the single probe was also arrayed in reverse complementation. The reverse complemented probes were compared to the probes originally designed by eArray to ensure that none of the reverse complemented probes were identical to the original eArray designed probe set. The quinoa seed microarray design was completed with 104,159 total features, including 102,834 total probes designed from unigenes and 1,325 Agilent control features.

Microarray hybridizations. Initial afrosimetric tests of seed tissue at different developmental stages indicated that 'bitter' saponins do not appear until after the aqueous phase of seed development (data not shown). Thus, immature seed samples were taken for the microarray analysis at two distinct developmental stages, specifically aqueous (~14 dpa) and milky (~21 dpa) from $F_{2:3}$ plants of a population segregating for the presence and absence of saponins. To determine the genotypic state at the 'bitter' saponin locus for each $F_{2:3}$ family, we determined saponin content for 12 F_3 plants for each F_2 individual using a afrosimetric method (Koziol 1991). F_2 individuals that produced F_3 progeny that were only saponin-containing were designated as homozygous dominant (22 F_2 individuals) (Fig. 5) and classified as saponin$^+$. Similarly, F_2 individuals that produced F_3 progeny that were only saponin-containing were designated as homozygous recessive (21 F_2 individuals) and were classified as saponin$^-$. F_2 individuals that segregated for the presence of saponin content in the $F_{2:3}$ generation were classified as heterozygous or saponin$^{+/-}$ (49 F_2 individuals). We note that the population segregated as expected for a single gene (1:2:1; $p \leq 0.01$). Plants classified as heterozygous (saponin$^{+/-}$) were excluded from all subsequent analyses.

Using the seed developmental stage (Aqueous or Milky) and *BSP* locus genotypic state (saponin$^+$ or saponin$^-$), each sample was assigned to the one of following treatments: Saponin

Milky (SM); Saponin Aqueous (SA); Non-Saponin Milky (NSM); and Non-Saponin Aqueous (NSA). RNA was extracted from each sample and prepared for amplification and labeling with Cy3 and Cy5 for microarray hybridization. Each of the four (SM, SA, NSM, NSA) bulked $F_{2:3}$ seed RNA samples were extracted, amplified and labeled four times; each sample twice with Cy3 and twice with Cy5, creating an intentional dye swap to account for dye bias, and hybridized to eight quinoa seed microarrays as shown in Table 3. Two (Array 2_2, Array 3_2) of the eight microarrays were flagged by Agilent Feature Extraction Software (v 10.5.1.1) as arrays to discard from any further analysis due to wash artifacts.

Across all samples and arrays, 64.9% of all probes were flagged as significantly expressed above background. A more conservative approximation of significantly expressed probes is given by the "WellAboveBG" flag, which only includes probes that are 2.6 standard deviations above background. "WellAboveBG" estimated 48.4% of all probes as significantly expressed above background. These values are consistent, although somewhat lower, than a similar microarray platform in three-spine stickleback (*Gasterosteus aculeatus* L.) that reported 71% of all probes as significantly expressed and 57% of all probes as "WellAboveBG" (Leder et al., 2009). The decreased number of significantly expressed probes is likely due to the fact that many of the probes on the quinoa microarray are potentially in reverse complementation (as a result of no ESTScan or BlastX result) and thus are not expected to accommodate hybridization. Additionally, the quinoa microarray was designed from an EST library containing expressed sequences from five stages of seed development of the breeding line '0654', while the hybridized samples were from only two stages of seed development. An array-by-array list of significantly expressed probes is given in Table 4.

Microarray data analysis. Statistical analysis was performed on the signal intensities processed by Agilent Feature Extraction Software (v. 10.5.1.1). Three different comparisons were calculated (results summarized in Table 5):

1) [SM - SA] vs. [NSM - NSA]. This comparison was made to screen for genes differentially expressed between the aqueous and milky stages of saponin and non-saponin quinoa. Since saponin is not detected by the afrosimetric shake test until the milky stage in saponin producing quinoa, this comparison tests for genes that are 'turned on' (upregulated) or 'turned off' (downregulated) between the aqueous and milky stages that are possible candidates for the BSP locus. A total of 1,389 probes were significant (p-value ≤ 0.01). Of these, 883 were upregulated while 506 were downregulated. This comparison had the most significant probes of all the comparisons. Probes for 86 unigenes were found in duplicate (190 probes). These unigenes with duplicated results for multiple probes are very good candidate genes for components either responsible for the saponin^{+} or saponin^{-} genotype, or genes affected downstream by the mutation due to feedback.

2) SM vs. NSM; this comparison was made to identify differences in gene expression between saponin milky and non-saponin milky quinoa. This comparison is based on the concept that the gene responsible for saponin production or lack of saponin production should be detected as differentially expressed in the saponin milky quinoa and non-saponin quinoa. 531 probes were found to be significant (p-value ≤ 0.01) for this comparison, 322 being upregulated and 209 downregulated. Probes for 17 unigenes were found in duplicate (35 probes).

3) SM vs. [SA + NSM + NSA]. This comparison was made to identify any genes that are differentially expressed in the only sample that produces saponin, saponin milky stage

quinoa. There were 427 probes significant (p-value ≤ 0.01) in this comparison, 243 upregulated and 184 downregulated. Probes for 4 unigenes were found in duplicate (8 probes).

Using a p-value cutoff threshold of 0.01, we identified a list of 198 significantly differentially expressed candidate unigenes common to all three comparisons. Of these, 151 unigenes were upregulated and 47 were downregulated using fold-change averaged across all three comparisons. which ranged from 13.9 to 1.4 in upregulated unigenes, and 0.03 to 0.74 in downregulated genes where a fold-change of 1 is defined as equally expressed between samples in the comparison. Ninety-four unigenes found in this candidate gene list were entirely unique to quinoa with 81 being upregulated and 13 downregulated. These sequences found only in quinoa could represent genes that are unique to the biosynthetic pathway in quinoa.

Saponin biosynthetic pathway related unigenes. Saponins are synthesized from mevalonic acid via the isoprenoid pathway where they are derived from triterpenoid or steroid cyclization of 2,3-oxidosqualene (Fig. 6) (Kuljanabhagavad and Wink, 2009). Functionally annotated unigenes that correspond to the hypothetical saponin pathway in quinoa were of particular interest in this study. This process proceeds with geranyl pyrophosphate and isopentenyl pyrophosphate being converted into farnesyl pyrophosphate by farnesyl pyrophosphate synthase. Squalene synthase then connects two farnesyl pyrophosphates via tail-to-tail linkage to form squalene (Holstein and Hohl, 2004). Oxidation of squalene by squalene monooxygenase yields 2,3-oxidosqualene. β-amyrin synthase catalyzes the cyclization of 2,3-oxidosqualene converting it to β-amyrin, with β-amyrin then being modified by cytochrome P450s to form sapogenin aglycones which are glycosylated by various glycosyltransferase enzymes to synthesize many different saponins (Suzuki et al., 2002).

Functionally annotated unigenes related to saponin biosynthetic pathways were found throughout the array including geranyl diphosphate synthase (18 probes), farnesyl diphosphate synthase (8 probes), squalene synthase, (9 probes), squalene monooxygenase (14 probes), β-amyrin synthase (12 probes), cytochrome P450 (192 probes), cytochrome P450 monooxygenase (65 probes), glycosyltransferases and other enzymes involved in sugar transport and linkage (312 probes). Not surprisingly, none of the probes for the first four enzymes listed in the saponin pathway (geranyl diphosphate synthase, farnesyl diphosphate synthase, squalene synthase, squalene monooxygenase), showed any significant differential gene expression between 'bitter' and 'sweet' quinoa. These results were expected as each of the products of these enzymes are required for pathways essential to plant survival, such as the sesquiterpenoid pathways (farnesyl diphosphate) and the brassinosteroid biosynthesis pathways (squalene, 2,3-oxidosqualene). Interestingly, β-amyrin synthase showed no significant differential gene expression. As the first committed step in triterpenoid saponin biosynthesis, β-amyrin was considered a prime candidate to control the production of saponin in quinoa. However, several probes with homology to cytochrome P450s (20), cytochrome P450 monooxygenases (10), and glycosyltransferases (49) were found to be significantly (p-value ≤ 0.05) differentially expressed in at least one of the three comparisons (Table 6). These results suggest that the differences in the saponin biosynthesis pathways between saponin producing and sweet varieties of quinoa arise following the formation of the β-amyrin skeleton. The significant probes represent candidate genes that may catalyze the formation of saponins (upregulated) or inhibits the production of saponins (downregulated). We note that genes shown to be downregulated in saponin containing samples may actually be upregulated genes in non-saponin samples, producing gene products capable of blocking the saponin biosynthesis pathway by inhibiting oxidation of β-amyrin to oleanolic acid, or by the

inhibition of further oxidation, esterification or glycosylation of oleanolic acid-derived aglycones (Kuljanabhagavad and Wink, 2009) prohibiting the linking of sugar moieties or other side chains that give the fully synthesized saponins their characteristic properties.

Conclusions and Future Work. We report the development and annotation of the first large scale EST collection for quinoa containing 39,366 unigenes and the development of a custom microarray to assay gene expression in developing seeds of quinoa. These resources can be used to help facilitate genomic research in quinoa. In addition we report several candidate genes that could be involved in the production of 'bitter' saponin in quinoa. Additional efforts should focus on the development of primers for the sequencing of candidate unigenes between 'bitter' quinoa and 'sweet' quinoa types and searching for Single Nucleotide Polymorphisms (SNPs). Subsequent segregation analysis of these candidate genes (via SNP analysis) in the $F_{2:3}$ population should reveal the gene(s) responsible for saponin production.

The entire concept of finding the gene responsible for 'bitter' saponin production using microarray analysis is dependent on the presence or absence of 'bitter' saponin production being controlled at the transcriptional level. If the lack of 'bitter' saponins is the result of a mutation that is manifest post-transcriptionally then it would be impossible to find the genetic component responsible for saponin production using microarray analysis. One possible scenario involving the post-transcriptional control of saponin production in quinoa is a mutation in the DNA sequence which does not affect transcription of the gene. Instead the mutant transcript would hybridize to a microarray with the same efficiency of the wild-type gene (especially if the mutation is not located in the probe sequence). However, upon translation of the mutated gene, the mutation could: 1) change an amino acid which causes the protein to misfold, resulting in the ubiquination of the misfolded protein and subsequent degradation; 2) change an amino acid in a

function-specific domain, while not affecting protein folding, it greatly reduces enzymatic efficiency or even renders the enzyme non-functional; 3) due to insertion or deletion, create a frameshift in the sequence completely altering the functionality of the enzyme. The extraction and isolation of enzymes involved in the saponin biosynthesis pathway and testing the respective quantities and enzymatic efficiencies between 'bitter' and 'sweet' quinoa could possibly elucidate the mutated enzyme if the mutation is indeed manifest post-transcriptionally.

It has been reported that some fungi subvert saponin-base plant defense systems by producing a saponin-detoxifying enzyme (Bouarab et al., 2002; Bowyer et al., 1995). This is most likely accomplished by deglycosylation of the saponins by β-glucosidases (Faure, 2002). Interestingly, three of the contigs that have a large amount of read depth have homology to 'glucan endo-1,3-β-D-glucosidase' (contig15038 (228 reads), contig02663 (105 reads), and contig00951 (97 reads)). Contigs with a large amount of read depth are likely the result of the transcript being very highly expressed in the tissue, in spite of the normalization process. It is currently unknown how quinoa protects itself from the toxicity of the saponins it produces; however, these findings could provide a clue to the as to the source of its immunity.

A study is currently underway using mass spectrometry protocols previously described (Kuljanabhagavad et al., 2008; Madl et al., 2006) to characterize differences in saponin content and quantity between 'sweet' and 'bitter' quinoa. It is hoped that the characterization of the structural differences in the saponin content of 'bitter' and 'sweet' varieties of quinoa will provide key clues in discovering the how the recessive mutation affects the saponin biosynthetic pathway in quinoa.

LITERATURE CITED

Bouarab K., Melton R., Peart J., Baulcombe D., Osbourn A.J.N. (2002) A saponin-detoxifying enzyme mediates suppression of plant defences 418:889-892.

Bowyer P., Clarke B.R., Lunness P., Daniels M.J., Osbourn A.E. (1995) Host Range of a Plant Pathogenic Fungus Determined by a Saponin Detoxifying Enzyme. Science 267:371-374.

Chauhan G.S., Eskin N.A.M., Tkachuk R. (1992) Nutrients and antinutrients in quinoa seed. Cereal Chemistry 69:85-88.

Chauhan G.S., Eskin N.A.M., Mills P.A. (1999) Effect of saponin extraction on the natritional quality of quinoa (*Chenopodium quinoa* Willd.) Proteins. Journal of Food Science and Technology 36:123-126.

Coles N.D., Coleman C.E., Christensen S.A., Jellen E.N., Stevens M.R., Bonifacio A., Rojas-Beltran J.A., Fairbanks D.J., Maughan P.J. (2005) Development and use of an expressed sequenced tag library in quinoa (Chenopodium quinoa Willd.) for the discovery of single nucleotide polymorphisms. Plant Science 168:439-447. DOI: DOI: 10.1016/j.plantsci.2004.09.007.

Coulter L., Lorenz K. (1990) Quinoa--composition, nutritional value, food applications. Journal of Food Science and Technology 23:203-207.

Cusack D.F. (1984) Quinua:Grain of the Incas. The Ecologist 14:21-31.

D'Altroy T., Hastorf C.J.A.A. (1984) The distribution and contents of Inca state storehouses in the Xauxa region of Peru 49:334-349.

Dini I., Schettino O., Simioli T., Dini A. (2001) Studies on the constituents of Chenopokium quinoa seeds: Isolation and characterization of new triterpene saponins. Journal of Agricultural and Food Chemistry 49:741-746.

Faure D.J.A.a.E.M. (2002) The family-3 glycoside hydrolases: from housekeeping functions to host-microbe interactions 68:1485.

Fenwick G.R., Price K.R., Tsukamota C., Okubo K. (1991) Saponins, in: J. P. D'Mello, et al. (Eds.), Toxic substances in crop plants, The Royal Society of Chemistry, Cambridge, UK. pp. 285-327.

Gotz S., Garcia-Gomez J., Terol J., Williams T., Nagaraj S., Nueda M., Robles M., Talon M., Dopazo J., Conesa A.J.N.A.R. (2008) High-throughput functional annotation and data mining with the Blast2GO suite.

Han Z., Wang C., Song X., Guo W., Gou J., Li C., Chen X., Zhang T. (2006) Characteristics, development and mapping of Gossypium hirsutum derived EST-SSRs in allotetraploid cotton. TAG Theoretical and Applied Genetics 112:430-439.

Holstein S., Hohl R. (2004) Isoprenoids: Remarkable diversity of form and function. Lipids 39:293-309.

Jacobsen S.-E., Mujica A., Jensen C. (2003) The resistance of quinoa (Chenopodium quinoa Willd.) to adverse abiotic factors. Food Reviews International 19:99-109.

Jarvis D.E., Kopp O., Jellen E., Mallory M., Pattee J., Bonifacio A., Coleman C., Stevens M., Fairbanks D., Maughan P. (2008) Simple sequence repeat marker development and genetic mapping in quinoa (Chenopodium quinoaWilld.). Journal of Genetics 87:39-51.

Koziol M.J. (1991) Afrosimetric estimationi of threshold saponin concentration for bitterness in quinoa. Journal of the Science of Food and Agriculture 54:211-219.

Kuljanabhagavad T., Wink M.J.P.R. (2009) Biological activities and chemistry of saponins from Chenopodium quinoa Willd 8:473-490.

Kuljanabhagavad T., Thongphasuk P., Chamulitrat W., Wink M.J.P. (2008) Triterpene saponins from Chenopodium quinoa Willd.

Leder E.H., Merilä J., Primmer C.R. (2009) A flexible whole-genome microarray for transcriptomics in three-spine stickleback (Gasterosteus aculeatus). BMC Genomics 106:426-433.

Livak K.J., Schmittgen T.D. (2001) Analysis of relative gene expression data using real-time PCR and the 2-ΔΔCt method. Methods 25:402-408.

Lottaz C., Iseli C., Jongeneel C., Bucher P.J.B.-O. (2003) Modeling sequencing errors by combining Hidden Markov models 19:103-112.

Madl T., Sterk H., Mittelbach M., Rechberger G.N. (2006) Tandem mass spectrometric analysis of a complex triterpene saponin mixture of Chenopodium quinoa. Journal of The American Society for Mass Spectrometry 17:795-806.

Mason S.L., Stevens M.R., Jellen E.N., Bonifacio A., Fairbanks D.J., Coleman C.E., McCarty R.R., Rasmussen A.G., Maughan P.J. (2005) Development and use of microsatellite markers for germplasm characterization in quinoa (Chenopodium quinoa Willd.). Crop Science 45:1618 - 1630. DOI: doi:10.2135/cropsci2004.0295.

Masterbroek H.D., Limburg H., Gilles T., Marvin H.J.P. (2000) Occurrence of sapogenins in leaves and seeds of quinoa (*Chenopodium quinoa* Willd). Journal of the Science of Food and Agriculture 80:152-156.

Modgil R., Mehta U. (1993) Antinutritional factors in pulses as influenced by different levels ofCallosobruchus chinensis L. (Bruchids) infestation. Plant Foods for Human Nutrition (Formerly Qualitas Plantarum) 44:111-117.

Morrissey J.P., Osbourn A.E. (1999) Fungal resistance to plant antibiotics as a mechanism of pathogenesis. Microbiology and molecular biology reviews 63:708-724.

Onning G., Wang Q., Weström B.R., Asp N.G., Karlsson B.W. (1996) Influence of oat saponins on intestinal permeability in vitro and in vivo in the rat. The British Journal of Nutrition 76:141-151.

Osbourn A. (2003) Molecules of interest: saponins in cereals. Phytochemistry 62:1-4.

Prado F.E., Boero C., Gallardo M., Gonzalez J.A. (2000) Effect of NaCl on germination, growth, and soluble sugar content in Chenopodium quinoa Willd. seeds. Botanical Bulletin Of Academia Sinica 41:27-34.

Puissant C., Houdebine L.J.B. (1990) An improvement of the single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction 8:148.

Risi C.J., Galwey N.W. (1984) The Chenopodium grains of the Andes: Inca crops for modern agriculture. Advances in Applied Biology 10:145-216.

Rozen S., Skaletsky H.J.M.M.B. (2000) Primer3 on the WWW for general users and for biologist programmers 132:365-386.

Saha M., Mian M., Eujayl I., Zwonitzer J., Wang L., May G.J.T.T.a.A.G. (2004) Tall fescue EST-SSR markers with transferability across several grass species 109:783-791.

Sanchez H., Lemeur R., Damme P., Jacobsen S.-E. (2003) Ecophysiological analysis of drought and salinity stress of quinoa (Chenopodium quinoa Willd.). Food Reviews International 19:111-119.

Sederberg M.C. (2008) Physical mapping of ribosomal RNA genes in new world members of the genus Chenopodium using fluorescence in situ hybridization, Department of Plant and Wildlife Science, Brigham Young University, Provo. pp. 53.

Stajich J., Block D., Boulez K., Brenner S., Chervitz S., Dagdigian C., Fuellen G., Gilbert J., Korf I., Lapp H.J.G.r. (2002) The Bioperl toolkit: Perl modules for the life sciences 12:1611.

Suzuki H., Achnine L., Xu R., Matsuda S., Dixon R. (2002) A genomics approach to the early stages of triterpene saponin biosynthesis in Medicago truncatula. The Plant Journal 32:1033–1048.

Thiel T., Michalek W., Varshney R., Graner A. (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theoretical and Applied Genetics 106:411-422.

Vacher J.J. (1998) Responses of two main Andean crops, quinoa (Chenopodium quinoa Willd) and papa amarga (Solanum juzepczukii Buk.) to drought on the Bolivian Altiplano: Significance of local adaptation. Agriculture, Ecosystems and Environment 68:99-108. DOI: DOI: 10.1016/S0167-8809(97)00140-0.

TABLES

TABLE 1. ESTs per Contig.

# of Reads	# of Contigs
1	26
2-5	3,988
6-10	3,543
11-20	3,758
21-30	2,202
31-50	2,010
51-100	1,073
100+	128
Total	16,728

TABLE 2. EST-SSRs	
Total number of sequences examined	39,366
Total size of examined sequences (bp)	13,362,222
Total number of identified SSRs	291
Number of SSR containing sequences	278
Number of sequences containing more than 1 SSR	10
Number of SSRs present in compound formation	12
Number of SSRs suitable for primer design	194

TABLE 3. Microarray Experimental Design.

Array	Cy3	Cy5
1_1	NSM-3	SM-3
1_2	SA-4	NSA-4
2_1	SA-1	SM-1
2_2	SM-2	SA-2
3_1	NSA-2	NSM-2
3_2	SM-4	NSM-4
4_1	NSM-1	NSA-1
4_2	NSA-3	SA-3

TABLE 4. Probe Hybridizations Across Microarrays

	g	r	gIsPosAndSignif	rIsPosAndSignif	gIsWellAboveBG	rIsWellAboveBG
array1_1	NSM-3	SM-3	71,213	79,106	53,901	62,325
array1_2	SA-4	NSA-4	48,406	55,174	28,101	35,508
array2_1	SA-1	SM-1	66,357	81,734	49,746	65,897
array2_2	SM-2	SA-2	x	x	x	x
array3_1	NSA-2	NSM-2	65,439	78,003	50,237	60,686
array3_2	SM-4	NSM-4	x	x	x	x
array4_1	NSM-1	NSA-1	67,701	73,865	50,686	56,467
array4_2	NSA-3	SA-3	52,384	61762	37,710	45,810

TABLE 5. Significant Probes Across Statistical Comparisons

p-value = 0.01	(SM-SA) vs (NSM-NSA)	SM vs NSM	SM vs (SA+NSM+NSA)
Total probes	1,389	531	427
Upregulated probes	883	322	243
Downregulated probes	506	209	184

TABLE 6. Probes Related to the Saponin Biosynthetic Pathway								
Gene products related to saponin biosynthetic pathway	Total related probes on array	Total significant probes	(SM-SA) vs (NSM-NSA)		SM vs NSM		SM vs (SA+NSM+NSA)	
			up	down	up	down	up	down
geranyl diphosphate synthase	18	0	0	0	0	0	0	0
farnesyl diphosphate synthase	8	0	0	0	0	0	0	0
squalene synthase	9	0	0	0	0	0	0	0
squalene monooxygenase	14	0	0	0	0	0	0	0
β-amyrin synthase	12	0	0	0	0	0	0	0
cytochrome P450	192	20	5	7	2	3	1	2
cytochrome P450 monooxygenase	65	10	3	5	0	1	0	1
glycosyltransferase	312	49	13	8	9	5	6	6

FIGURES

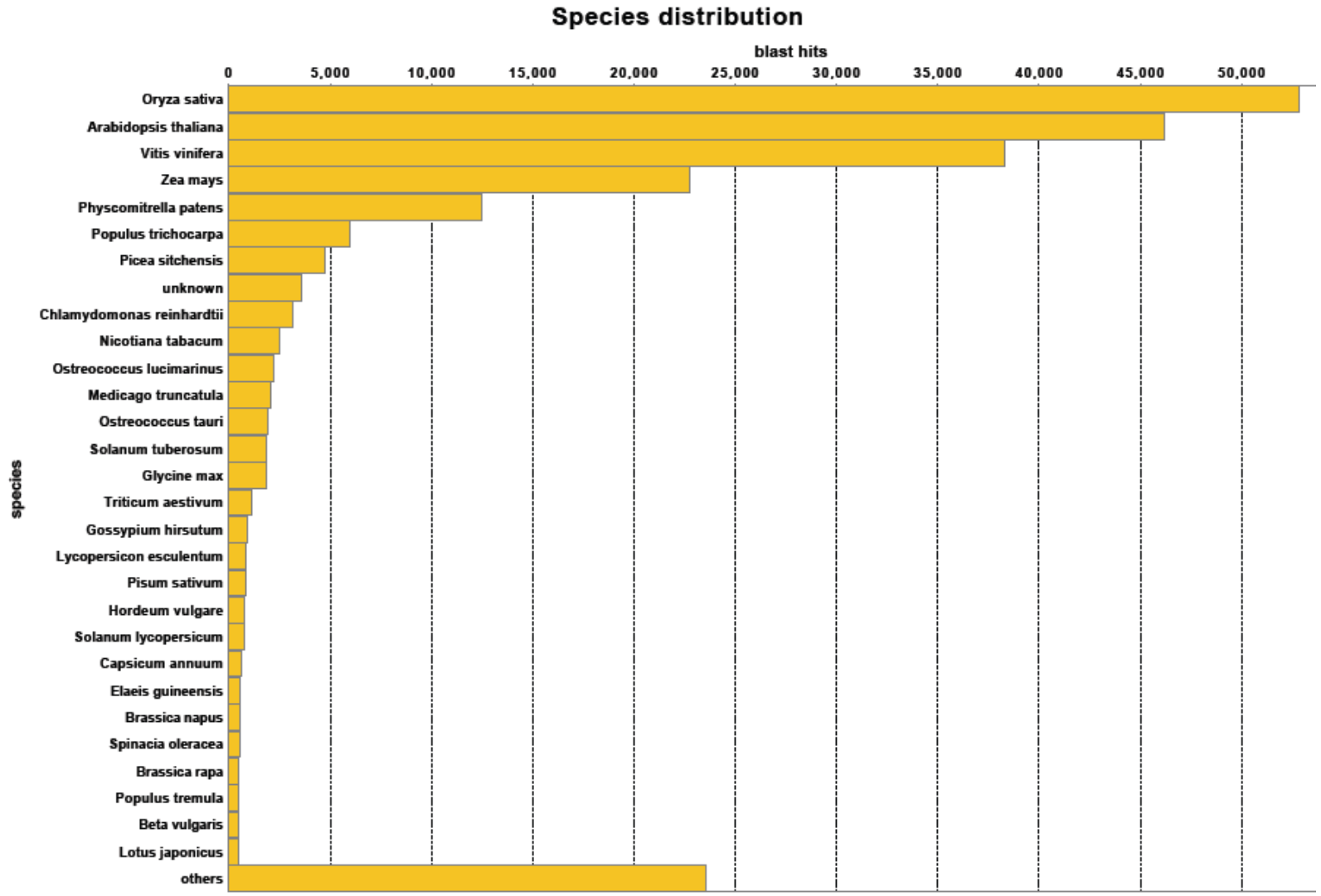


Figure 1. Species Distribution of Blast hits on *Chenopodium quinoa* unigenes

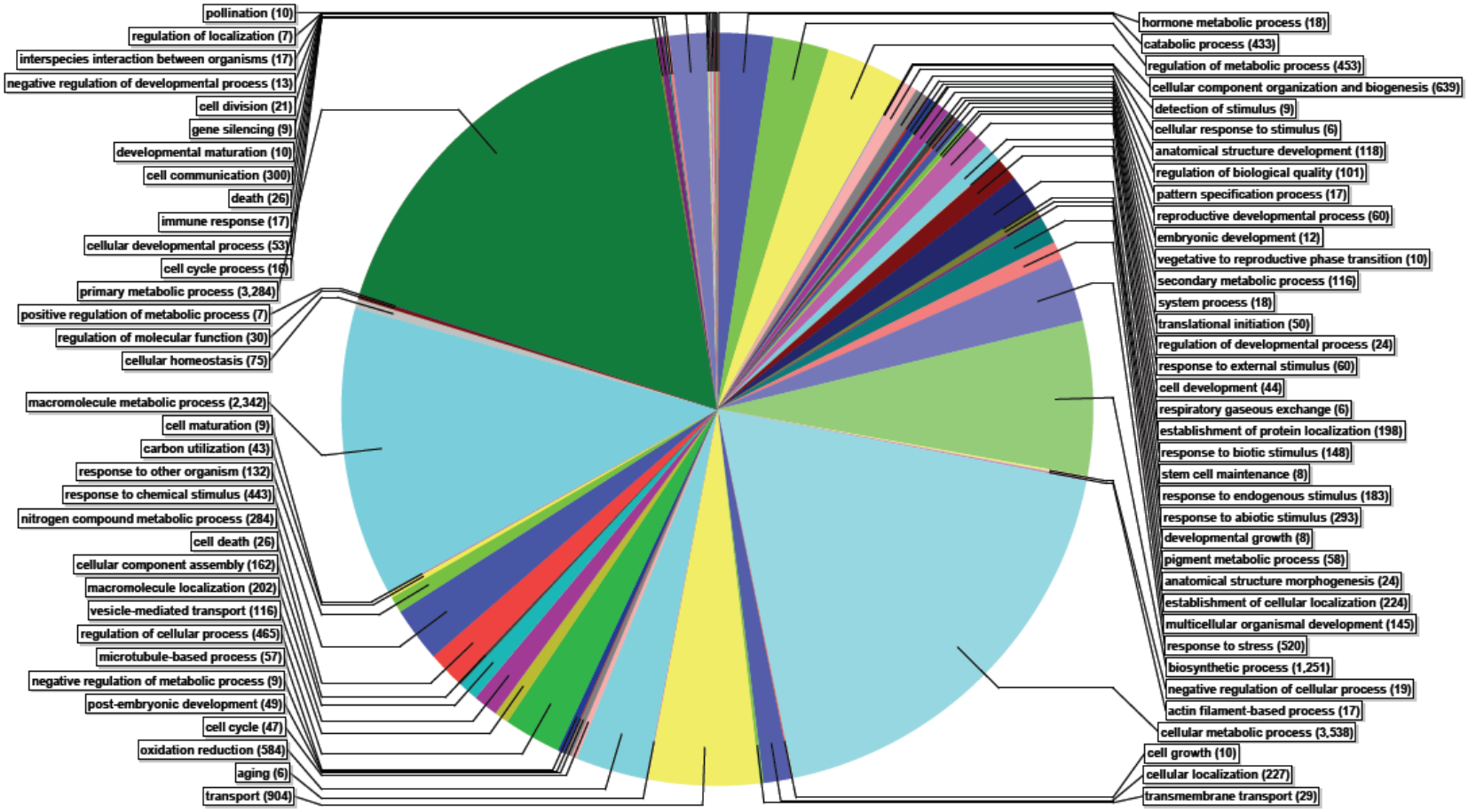


Figure 2. Blast2GO Functional annotation of all *Chenopodium quinoa* unigenes for Biological Process (Level 3).

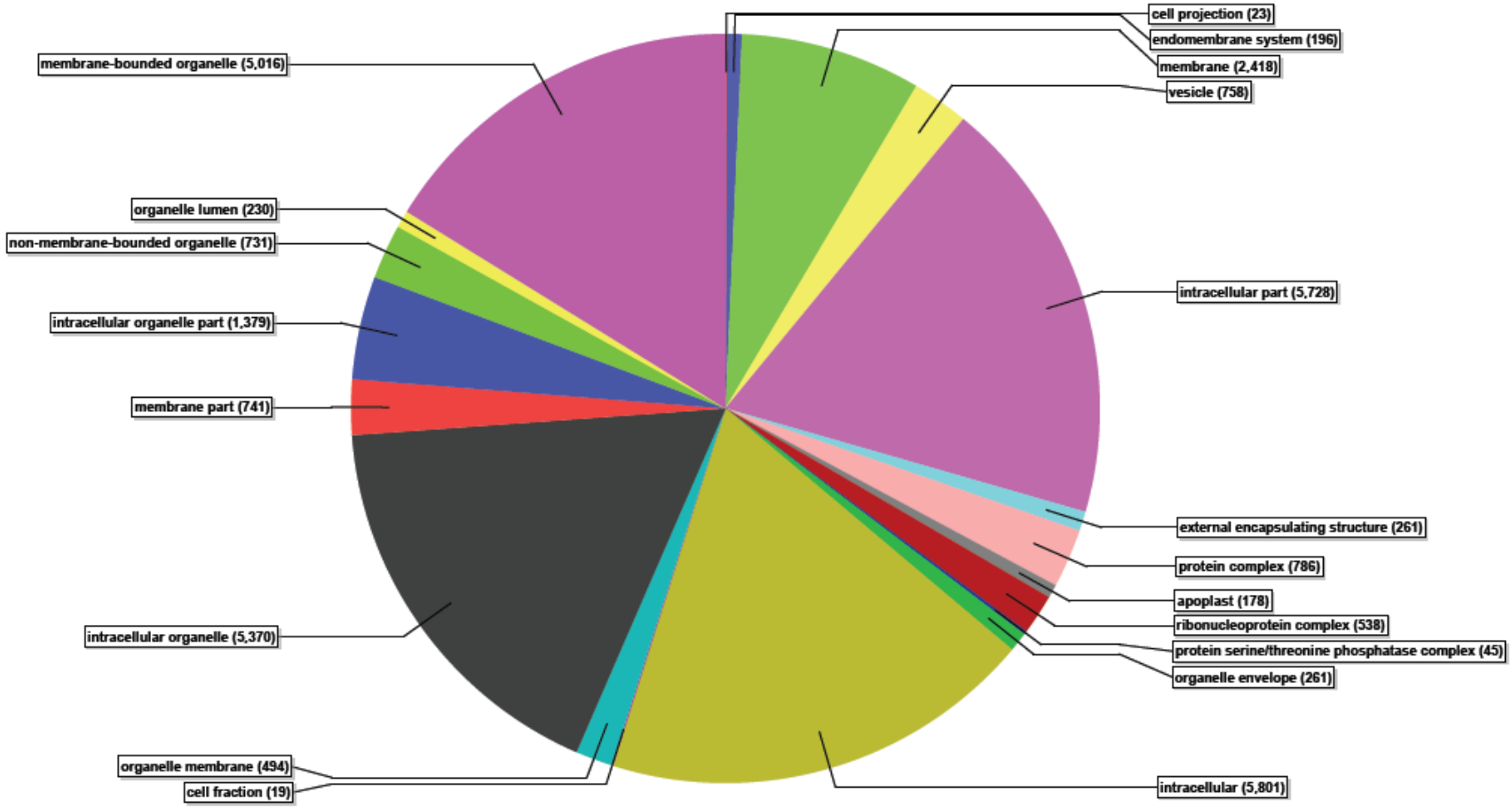


Figure 3. Blast2GO Functional annotation of all *Chenopodium quinoa* unigenes for Cellular Component (Level 3).

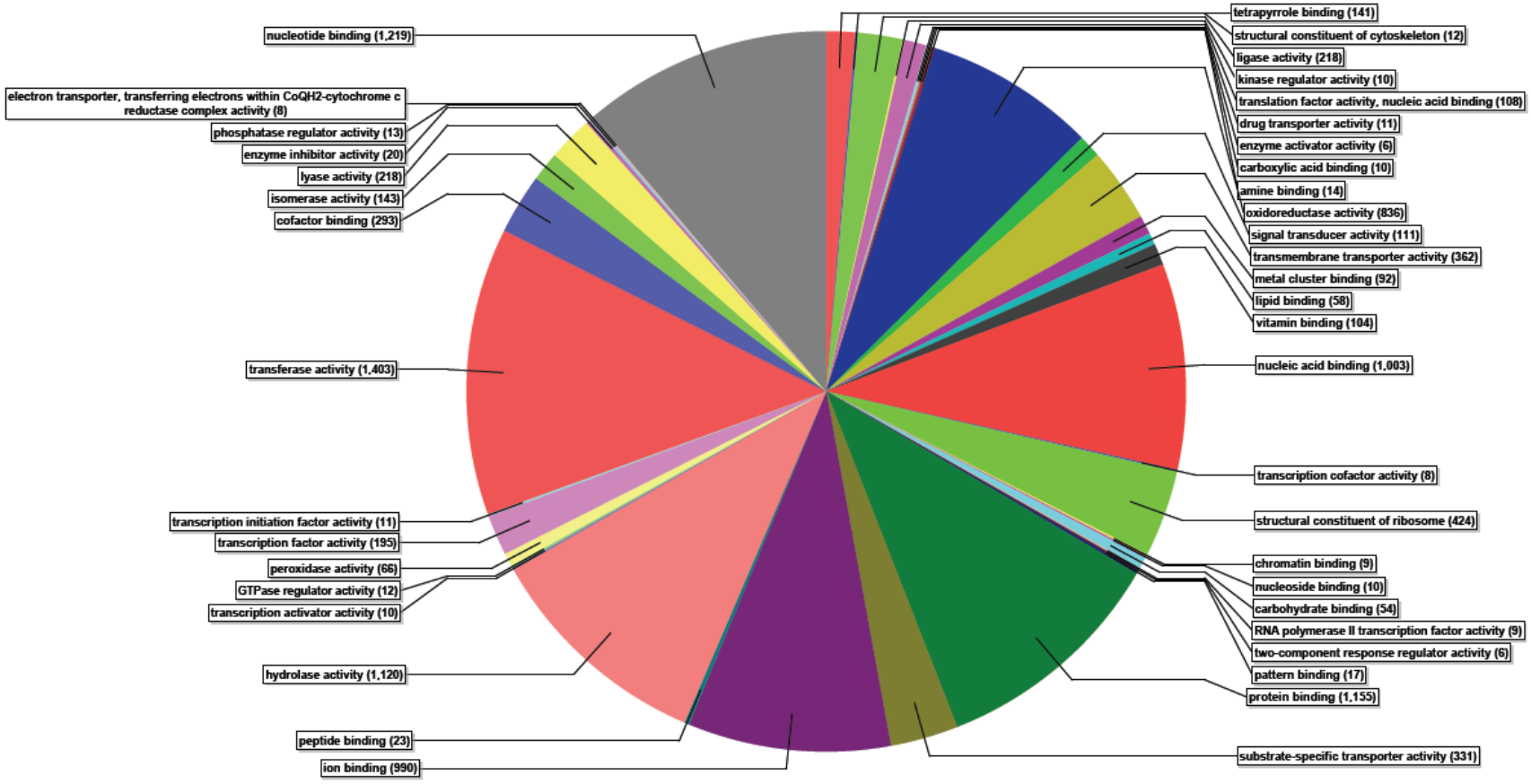


Figure 4. Blast2GO Functional annotation of all *Chenopodium quinoa* unigenes for Molecular Function (Level 3).

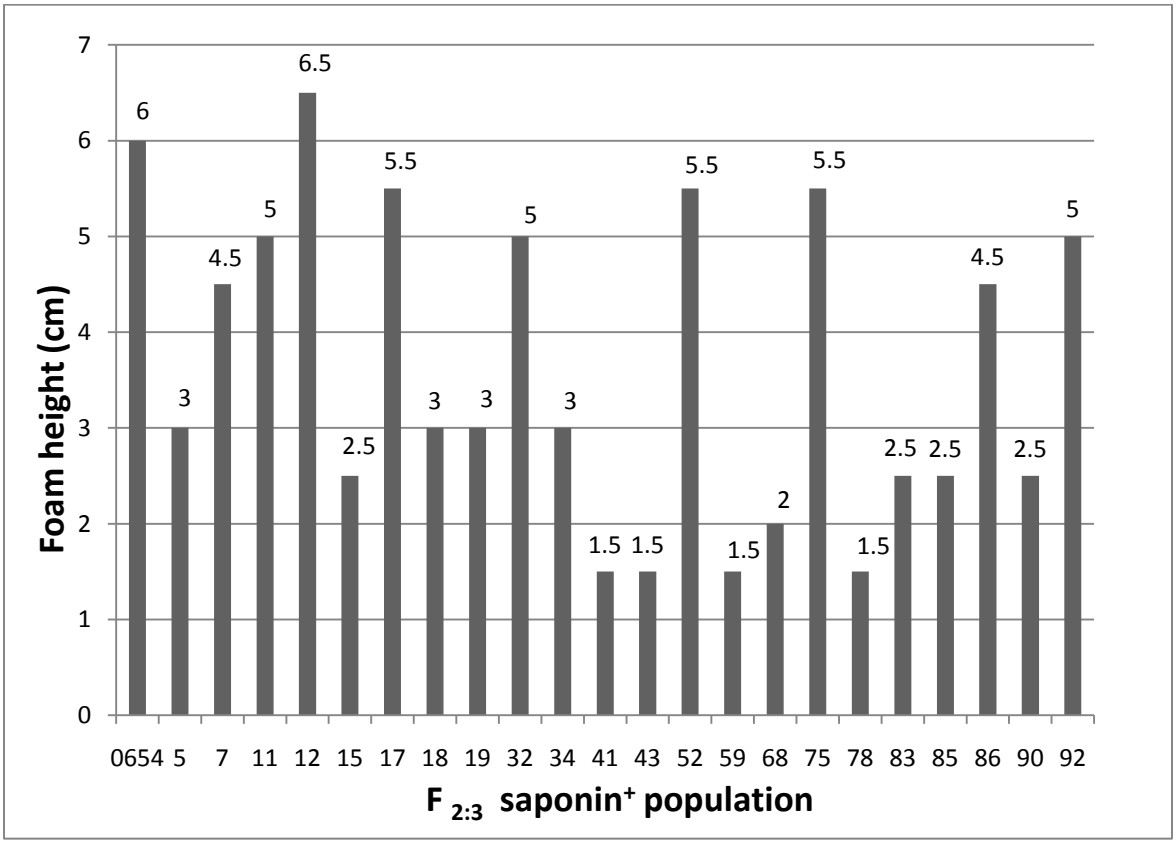


Figure 5. Afrosimetric shake test for saponin content in quinoa homozygous 'bitter' population (saponin^{+}). '0654' is the parent, with progeny identified by arbitrary number assigned during determination of homo- or heterozygosity. ('sweet' population not shown – all individuals registered zero cm.)

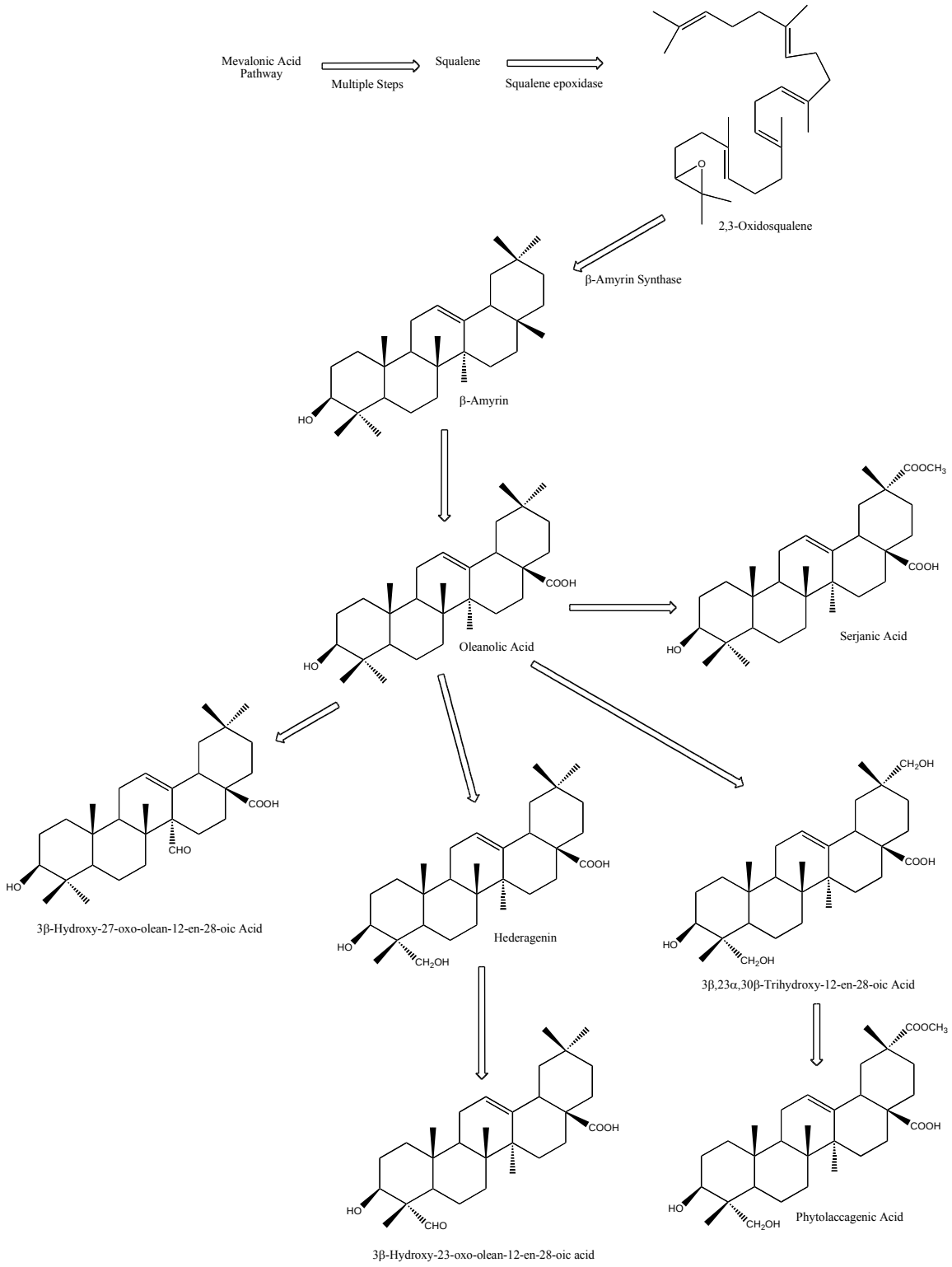


Figure 6. Proposed biosynthetic pathway of saponins in quinoa. The mevalonic acid pathway produces squalene which is oxidized by squalene monooxygenase to form 2,3-Oxidosqualene. Cyclization of 2,3-Oxidosqualene by β-amyrin synthase results in β-amyrin. β-amyrin is oxidized, presumably by cytochrome P450s to form Oleanolic Acid. Further oxidation of Oleanolic Acid produces the many different saponin aglycones in quinoa. Different sugar moieties are attached at various carbons (most commonly C-3 and C-28 among others) by glycosyltransferases (not shown).

CHAPTER 2: LITERATURE REVIEW

Introduction

Quinoa (Chenopodium quinoa Willd) is an important crop for subsistence farmers in the Altiplano (high plains) of Peru, Bolivia, and Argentina. Anciently, quinoa was honored and cultivated in the Incan Empire. The successive Spanish conquest led to the possible suppression of quinoa due to its cultural importance (Cusack 1984). As a result, quinoa production had been in a 400-year decline, grown only by the Altiplano descendants of the Incas.

Recently quinoa has seen a revival in interest and usage. This newfound attention to the pseudo-cereal comes as a consequence of recent studies that call attention to the vast nutritive properties of quinoa. In addition to its high nutritional value, quinoa has the potential to be an effective crop for many temperate and highland-tropical regions. This is due to its ability to thrive in drought, saline and high-altitude conditions (Vacher 1998; Prado, Boero et al. 2000). A joint effort between Bolivian researchers and Brigham Young University is currently working to improve quinoa, much in the same manner that other crops such as corn, rice, wheat etc. have been developed using biotechnological tools in plant breeding. This effort aims to provide consistent high quality yields of quinoa to the Altiplano region, enabling exportation and economic stability.

A major obstacle to this goal is the presence of saponin in the seed coat of many varieties of quinoa. Saponins are part of a diverse family of secondary triterpenoid metabolites that occur in a wide range of plant species. Due to their bitter taste and anti-nutritive properties, saponins must be removed before they are consumed. This is a process that requires large amounts of clean water and/or machinery – both of which are resources not available to the average subsistence farmers who grow quinoa.

However, there are some varieties of quinoa ('sweet quinoa') that have reduced levels of saponin. It has been previously demonstrated, using segregating populations and molecular markers, that the production of bitter saponin in quinoa is controlled by a single dominant locus (Ward 2001; Ricks 2005). In spite of this knowledge, the gene responsible for bitter saponin production remains unknown.

This study seeks to identify the BSP (Bitter Saponin Production) gene via microarray analysis. Microarrays are a tool used to decipher transcriptionally regulated responses. Thousands of genes can be analyzed simultaneously by measuring the transcription levels in controlled experimental treatments. A microarray is a glass microscope slide printed with partial gene sequences as probes to detect transcriptional changes in mRNA levels. Transcript variation of specific mRNAs from controlled experimental treatments, in this case 'sweet' vs. 'bitter', can identify function specific candidate genes, target the BSP locus and help to elucidate the associated biosynthetic pathways.

Nutrition properties of quinoa

The nutritional value of quinoa has been well documented. In the Altiplano region, quinoa is one of the principal protein sources and is used as substitute for the lack of animal protein in their diet (Repo-Carrasco et al., 2003). Several studies have shown that quinoa grain has an excellent balance of carbohydrates, lipids, and proteins and provides an ideal balance of essential amino acids for human nutrition (Chauhan et al., 1992; Coulter and Lorenz, 1990).

The protein content in quinoa grain is about 15% and starch content is about 60% (Ruales and Nair, 1993). The proteins in quinoa are important because of their quality of composition, which is very similar to that of casein, the protein of milk (Repo-Carrasco et al., 2003). These milk-like

protein properties make quinoa an excellent food to help curb malnutrition in children in low-income families. Ruales and Nair (1993) report that an infant food made from quinoa showed an increased level of insulin-like growth factor-1(which plays an important role in childhood growth) in the plasma of children who consumed the food. Taxonomy and nutrition analysis suggest that quinoa is safe to include in a gluten-free diet (Thompson 2000).

The content of tryptophan and lysine in quinoa protein is three times higher than that in whole wheat. Methionine content is at least two times higher in quinoa than in wheat (Ruales and Nair, 1992). This is important because lysine, tryptophan and methionine are essential amino acids, which means that humans cannot synthesize them; hence they must be ingested. Quinoa also contains about 9% fat (Ruales and Nair, 1993), with 50.2% of the oil being Omega 6 (linoleic acid), which makes it a candidate for oil extraction (Repo-Carrasco et al., 2003). Quinoa offers other health advantages. One of the inherent benefits of the quinoa grain is the 11% dietary fiber content (Ruales and Nair, 1993),which has many positive health effects, like the lowering of cholesterol levels and improved digestion (Repo-Carrasco et al., 2003).

Synthesis and structure of Saponin

Saponins are a major family of secondary metabolites that occur in a wide range of plant species. Saponins can be triterpenoid, steroid or steroidal glycoalkaloid molecules with one of more sugar chains (Fenwick et al. 1991). They are commonly characterized as soap-like substances that exhibit a wide range of properties and therefore are regarded as important biological compounds.

Saponins are synthesized from mevalonic acid via the isoprenoid pathway where they are derived from triterpenoid or steroid cyclization of 2,3-oxidosqualene (Osbourn, 2003). This process normally begins when 2,3-oxidosqualene is converted to β-amyrin by β-amyrin synthase,

with β-amyrin then being modified by cytochrome P450s to form sapogenins which are modified by glycosyltransferase enzymes to synthesize many different saponins (Suzuki et al., 2002). This process in Medicago truncatula is shown in Figure 1.

The multiplicity of properties and functions of saponins are due to the variety of backbone and sugar side chain components (Dini et al., 2001). These traits are taken advantage of commercially to manufacture a variety of products including as drugs and medicines, precursors for hormone synthesis, foaming agents, sweeteners, taste modifiers and cosmetics (Osbourn, 2003).

Saponin as a natural pesticide

Saponins, due to their triterpenoid chemical structures, have very potent antifungal properties. Because they are naturally occurring, it is believed that saponins are a critical part of the evolution of plant disease resistance. Their primary mode of action involves the formation of complexes with membrane sterols present in eukaryotes, resulting in loss of membrane integrity (Osbourn 1996). Papadopoulou et al. (1999) identified saponin deficient mutant-types in oats. These mutants were exposed to a variety of fungal pathogens. Most of them either died or suffered extensive damage, while the wild-type saponin varieties were unaffected (Papadopoulou, 1999).

Two acylated bisglycoside saponins (Acaciaside A and B) originally isolated from the funicles of Earleaf Acacia (Acacia auriculiformis), were shown to have antifungal and antibacterial activities. Complete inhibition of fungi (Aspergillus ochraceous and Curvularia lunata) and the inhibition of the growth of bacteria (Bacillus megaterium, Salmonella typhimurium and

Pseudomonas aeruginosa) were reported. Interestingly, the fungal and bacterial inhibitions were the result of a different mechanism of action (Mandal et al. 2005).

Triterpene saponins in particular have an important role in protecting some plants from predation(Dixon and Sumner, 2003). Some saponins found in quinoa act as a natural pesticide for the plant by producing bitter compounds that deter insects and birds (Zhu et al., 2002). In the constant arms race between pathogens and plants, some fungal pathogens produce secreted, saponin hydrolyzing enzymes, conferring resistance to saponin-based plant defense mechanisms. (Loria et al. 2006).

Saponins in quinoa

Quinoa grain also has a seed coating consisting of various saponins (Fig. 2). There are two principal types of saponin in quinoa: (1) a rare acid and neutral saponin group more commonly associated with white quinoas and (2) a more common type found in yellow quinoa cultivars (Johnson and Ward 1993). These saponins have been identified in both 'sweet' and 'bitter' varieties of quinoa (Dini et al., 2001; Woldemichael and Wink, 2001; Zhu et al., 2002).

This pattern suggests that the bitterness in quinoa is not caused by one particular saponin, but perhaps by the quantity and combination of saponins produced by the plant. This idea is supported by the following example. Woldemichael and Wink (2001) demonstrated that a 50μg/mL concentration of total quinoa saponins strongly inhibited the growth of Candida albicans, a fungus; but the individual saponins did not significantly affect the fungus even at concentrations as high as 500μg/mL.

Additionally, many 'sweet' varieties of quinoa produce low levels of saponin but are non-bitter and do not decrease palatability (Masterbroek et al., 2000). Identification of the gene responsible

for bitter saponin production in quinoa would greatly increase the success and decrease the time required in quinoa breeding programs.

Removal of saponins

Due to their bitter taste and anti-nutritive properties, saponins must be removed before they are consumed. There is no effect on the nutritional quality of quinoa after saponin extraction (Chauhan et al., 1999), and the removal of saponins does not have any negative effect on the digestibility of proteins in quinoa (Ruales and Nair, 1992).There are two main methods to remove saponin from quinoa: (1) washing or (2) dry polishing (Mujica et al., 2003). The wet methods are those traditionally used by subsistence farmers. The grains are washed while being rubbed with the hands or scrubbed with a stone. The dry method is an abrasive dehulling method where machinery is used to dry-polish the grains to remove the saponins.

Yet, not all of the saponins are removed by this process. The effectiveness of dry polishing can be increased if the grain is burnished more forcefully, but this may result in the loss of some of the proteins of the outer layers of the grain. An alternative and potentially more effective method involves quickly dry polishing the seeds and then briefly rinsing them just before cooking (Repo-Carrasco et al., 2003). Due to the slightly acidic nature of saponins, washing in slightly alkaline water might more effectively remove saponins (Zhu et al., 2002). This is due to rather simple acid/base chemistry where the acidic saponins bind to more alkaline water molecules effectively stripping the saponins from the seed.

Bitter saponin locus in quinoa

Bitter saponin, a major seed coating component found in quinoa, is responsible for bitterness and inhibits nutrient uptake in humans (Masterbroek et al., 2000). Breeding a high quality, pest-

resistant, bitter saponin-free quinoa variety with the other desired traits - high yield, short growing season, etc. - through traditional breeding can be a long process, possibly taking years to accomplish. A breeding program assisted through genetic knowledge of the inheritance of bitter saponins could potentially shave years off the process. Unfortunately, the gene responsible for bitter saponin production in quinoa is unknown.

The saponin levels in quinoa are both qualitatively and quantitatively controlled. It has been reported that saponin production in quinoa requires at least one dominant allele at the bitter saponin locus; quinoa with a fully recessive allele at the bitter saponin locus had no detectable amounts of saponin (Ward, 2001). However, the amount of saponin is determined by an unknown number of QTLs (Ward, 2001).

Additionally, the bitter saponin locus has not been tightly (< 5 cM) linked to molecular markers, making marker assisted selection very difficult (Ricks, 2005). Genetic mapping has produced some linked markers, the most tightly linked being an AFLP marker linked in coupling 9.4 cM from the bitter saponin locus (Ricks, 2005). However, the exact nature of the gene responsible for bitter saponin in quinoa remains unknown.

Functional Genomics– Analyzing the Transcriptome

In the central dogma of biology, DNA is transcribed into mRNA which is then translated into protein. By measuring the levels of mRNA from specific tissue, the amount of proteins being synthesized in the tissue can theoretically be determined. The expression of the gene ultimately determines the expression of the protein that the gene encodes. This is done through analysis of mRNAs transcribed (transcripts) from the genomic DNA. " The complete set of transcripts and their relative levels of expression in a particular cell or tissue type under defined conditions" is

defined as the transcriptome (Gibson and Muse, 2004). The analysis of gene expression is an essential part of learning the functions of genes and how they are involved in biological pathways. This study of transcription and gene expression is known as functional genomics. The general methodology of transcriptome analysis by microarray is illustrated in Fig 3.

Initially gene expression was measured on a gene by gene basis using Northern blot analysis. Many different methods have since been designed and used to study the expression of both known and unknown genes. Some of the methods, such as reverse transcriptase-polymerase chain reaction (RT-PCR), serial analysis of gene expression (SAGE), and cDNA- amplified fragment length polymorphism (cDNA-AFLP) are used to study differential expression between two sets of conditions (Kozian and Kirschbaum, 1999; Rishi et al., 2002). These methods are still used but are limited in their ability to measure transcription because they can only analyze a few samples at a time, and reliance on gel electrophoresis and repetition of data to verify the results. These limitations are easily overcome by DNA microarrays, with their ability to simultaneously measure genome-wide changes in gene expression.

In addition to several other applications, microarray technology is primarily used in various ways to study transcriptomes, or gene expression applied in novel ways to answer questions. Microarrays are being used to generate expression profiles, unravel gene function, identify and characterize transcriptional factors and promoter elements, diagnose disease and cancer, drug discovery and crop improvement among others (Albertson, Pinkel 2003). In plants gene expression profiles have been developed to study the effects of abiotic and biotic stresses, plant development and the associated metabolic pathways (Rabban et al. 2003).

Principles of Microarray Technology

Microarrays decipher gene expression by analyzing the transcriptome across two conditions or treatments. The basic principle of microarray technology is the hybridization of complementary single stranded nucleic acid sequences of the probe and the target (Kozian and Kirschbaum, 1999). Generally, thousands of gene specific sequences, or probes, are affixed to a glass slide.

The target mRNAs are labeled with fluorescent dyes and then co-hybridized to the microarray slide. The dyes used are usually Cy3-dCTP which is yellow-orange (~550 nm excitation, ~570 nm emission), while Cy5-dCTP is fluorescent in the red region (~650/670nm) (Jackson ImmunoResearch 2008). The labeled and hybridized microarray slide is then placed in a laser scanner which detects the intensity of the fluorescent dyes on each probe spot.

These scanned images are then given a computer-aided false-coloring of green(Cy3) and red (Cy5) respectively. This allows for the detection of differences in the relative mRNA levels between two treatment groups. If a spot fluoresces red, then only the treatment group labeled with Cy5 expresses the gene identified by the spot. If a spot fluoresces green then only the treatment group labeled with Cy3 expresses the gene. If a spot fluoresces yellow then both treatment groups express the gene. Very dark spots or no fluorescence detected indicates that the spotted gene is not expressed very highly or at all in either treatment group. The statistical analysis of the slide uses the numerical intensity readings from the laser scanner to determine the expression identity of the probe.

Types of Microarrays

There are two main types of microarray slide, cDNA amplicon-spotted and synthetic oligonucleotide-spotted, each with advantages and disadvantages. cDNA amplicon microarray

slides are PCR-amplified cDNA fragments, also called ESTs (Expressed Sequence Tags) spotted onto a microscope slide or filter paper. Which clones are to be spotted is determined by the annotation of the cDNA library. Ideally each cDNA is sequenced and then unique genes are spotted. Random clones can also be spotted, but this leads to overrepresentation of highly expressed genes (Gibson and Muse, 2004).

An advantage of cDNA microarrays is the lower cost associated with making them, allowing researchers to perform a large number of experiments without as much cost. This is because the researcher can make the slide themselves, allowing for greater versatility. However, mismatching of cDNA clones and ESTs can be problematic, due to tracking errors and ease of contamination. cDNAs microarrays also have trouble discriminating related genes, multigene families and differentially spliced genes (Lee et al., 2004).

Synthetic oligonucleotide microarrays utilize 50-70 basepair sequences spotted onto the glass microscope slide. This method differs from cDNA microarrays in that the entire gene sequence is not on the slide; just a short unique segment, referred to as a probe. The design of these probes can prove to be difficult. The best probes must distinguish between the intended target and all other targets in the mRNA pool. They must be able to detect differences in concentration under hybridization conditions with the least amount of variation (Nielsen et al., 2003).

Advantages of synthetic oligonucleotide microarrays include the elimination of error associated with tracking cDNA clones and ESTs, uniform probe sequence length which allows for uniform hybridization, and high hybridization specificity which allows related genes, multigene families and differentially spliced genes to be identified (Lee et al., 2004). Unfortunately, the high cost of

synthetic oligonucleotide microarrays may limit the amount of experiments researchers are able to perform under a tight budget.

Lee et al. (2004) determined that there is not a difference in the ability of either type of microarray to detect expression changes, and suggested that much more experimental variance results from dye-labeling. Variance in dye-labeling was later confirmed; Cy5 labeling is highly susceptible to ozone degradation, even at low levels (5-10 ppb for 10-30 seconds), and greatly affects microarray data quality, while Cy3 data quality remains unaffected at much higher concentrations (>100ppb) (Fare et al. 2003).

This means that there is no clear-cut "better" microarray technique, but that the best microarray is determined by the type of experimental question the researcher is trying to answer. Conversely, there is a clear shift toward the usage of synthetic oligonucleotide microarrays; this may be due to the ease, accuracy and reproducibility that the synthetic microarrays offer.

Agilent Microarray Technology

Agilent Technologies have designed the next generation of synthesized oligonucleotide microarrays. The key features are the accuracy of oligonucleotide printing and the density of probes on the slide. Agilent's proprietary ink-jet-based in situ fabrication method allows a single base to be incorporated onto the nucleotide sequence. This process is repeated 60 times to make 60-mer oligonucleotide probe sequences and ensures accurate and uniform probes (http://www.agilent.com).

This technology, based on solid-phase phosphoramidite chemistry (Fig. 4), is the replacement of the 5'-dimethoxytrityl blocking group with an aryloxycarbonyl and the use of N-dimethoxytrityl protection for the exocyclic amines of adenine and cytosine (Sierzchala et al.,

2003). This allows the coupling of a single 2'-deoxynucleoside 3'-phosphoramidite to the growing oligonucleotide that is anchored to the microarray slide. Washing with peroxy anions removes the carbonate protecting group while oxidizing the phosphate internucleotide linkage creating an accurate two-step synthesis process (Sierzchala et al., 2003).

Agilent is currently producing single-array and multiple-array microarrays with 1 X 244,000, 2 X 105,000, 4 X 44,000 and 8 X 15,000 features on standard 1" x 3" glass slides. The 1 X 244,000 microarray offers the highest sensitivity and allows for very intricate genome scanning. The 2 X 105,000 is designed to offer above average sensitivity or multiple probe per sequence analysis of a single treatment and deeper coverage. The 4 X 44,000 microarray is more versatile and is optimized for efficiency and coverage. Finally, the 8 X 15,000 microarray is best for targeted profiling of a large number of samples. This represents unparalleled density, sensitivity and flexibility in the microarray industry (Matlow, 2006).

Limitations of Microarray Technology

Although microarray technology is gaining popularity, and the field of functional genomics seems to be rising, it is also at the mercy of the limitations imposed on microarrays. Because microarray technology only measures gene expression at the mRNA level, post-transcriptional regulation cannot be determined. Proteomics will have to be incorporated in order to correctly assign functions to genes (Kislinger et al. 2006). Additionally, EST libraries may only represent 25-50% of the genes in a genome (Lee et al., 2004).

Even with all of the technological advancements that have been made, microarrays are still very expensive to perform. In addition to the cost of materials, the analysis also requires specialized

equipment and programs. This makes it difficult for small labs to use microarrays (Gibson and Muse, 2004)

Sometimes the amount of RNA that can be extracted from tissue is a limiting factor in microarray analyses. This is especially apparent in developmental tissues, where the amount of sample may be very small. To alleviate this problem, RNA can be amplified. RNA amplification by in vitro transcription is the most common amplification method. Generating a dsDNA template can be done two ways: 1) reverse transcription of mRNA followed by a second-strand cDNA synthesis; 2) a combination of the switch mechanism at the 5' end of RNA templates followed by PCR (Wang et al., 2003). There is no difference in the RNA quality as far as microarray results, however, more amplified RNA can be obtained using conventional second-strand cDNA synthesis than from the combination of SMART and PCR (Wang et al., 2003)

While these limitations can seem restrictive, in the future microarrays combined with bioinformatics and proteomics will accelerate the discovery and annotation of genes in breeding programs (Rishi et al., 2002).

Microarray Analysis

Perhaps the most important part of microarray technology is the ability to analyze the volumes of data that can be generated by a single microarray scan. There are several different programs designed to explore microarray data, but they all have similar features that aim to normalize the data and make it relevant. A typical statistical analysis of microarray data involves calculating a test statistic and determining the significance, or p-value, of the observed statistic (Slonim 2002). Statistical tools to detect significant change between multiple measurements of a single treatment

or probe can also be used; for example a t-test or the F statistic can be applied to multiple groups via Analysis of Variance (ANOVA) (Slonim 2002).

These statistical methods are very useful with one caveat; microarray data is inherently complex and subject to variation. Microarray data thus needs to be normalized before standard analyses can be performed. One of the major obstacles in microarray data is the variability of the microarray probes themselves, called probe effects. Another is reproducibility and the ability to compare experiments performed at different times and under different conditions, or batch effects.

Two color arrays are sensitive to probe effects, especially GC content; the higher GC probes tend to display higher intensity (red, yellow or green) than probes with lower GC content (Song et al. 2007). Copy number of probes and cross-hybridization of similar sequences are complications normally associated with microarray experiments; however, these are not concerns with newer synthetic arrays which are designed to (1) exclude repeated regions and (2) longer probes allowing more stringent washings to minimize cross-hybridization effects (Song et al. 2007).

Batch effects, or non-biological experimental variation can be the result of many factors, such as the time of day of the assay, the reagents used in the assay, the batch of amplification and a myriad of other factors (Johnson 2007). Ozone levels are highly correlated with batch effects in microarray data, due to the susceptibility of Cy5 to ozone levels above 5-10 ppb (Fare et al. 2003). Several statistical approaches have been proposed to normalize batch effects; Johnson et al. (2007) suggest that parametric and non-parametric empirical Bayes frameworks are effective in correcting for batch effects.

LITERATURE CITED

Achnine L, Huhman DV, Farag MA, Sumner LW, Blount JW and Dixon RA (2005) Genomics-based selection and functional characterization of triterpene glycosyltransferases from the model legume Medicago truncatula The Plant Journal 41:875–887.

Albertson DG, Pinkel D. (2003) Genomic microarrays in human genetic disease and cancer. Human Molecular Genetics, 12:R145-R152

Chauhan, G. S., N. A. M. Eskin, et al. (1999). "Effect of saponin extraction on the natritional quality of quinoa (Chenopodium quinoa Willd.) Proteins." Journal of Food Science and Technology 36(2): 123-126.

Chauhan, G. S., N. A. M. Eskin, et al. (1992). "Nutrients and antinutrients in quinoa seed." Cereal Chemistry 69: 85-88.

Clarke JD, Zhu T (2006) Microarray analysis of the transcriptome as a stepping stone towards understanding biological systems: practical considerations and perspectives. The Plant Journal 45:630-650.

Coulter, L. and K. Lorenz (1990). "Quinoa--composition, nutritional value, food applications." Journal of Food Science and Technology 23: 203-207.

Cusack, D. F. (1984). "Quinua:Grain of the Incas." The Ecologist 14(1): 21-31.

Dini, I., O. Schettino, et al. (2001). "Studies on the constituents of Chenopokium quinoa seeds: Isolation and characterization of new triterpene saponins." Journal of Agricultural and Food Chemistry 49: 741-746.

Dixon, R. and L. Sumner (2003). "Legume Natural Products: Understanding and Manipulating Complex Pathways for Human and Animal Health." Plant Physiology 131: 878–885.

Fare TL, Coffey EM, Dai H, He YD, Kessler DA, Kilian KA, Koch JE, LeProust E, Marton MJ, Meyer MR, Stoughton RB, Tokiwa GY, Wang Y (2003) Effects of atmospheric ozone on microarray data quality. Anaytical Chemistry, 75:4672-5.

Fenwick, G.R., Price, K.R., Tsukamota, C., and Okubo, K. 1991. Saponins. In Toxic substances in crop plants. Edited by J.P. D'Mello, C.M. Diffins, and J.H. Duffus. The Royal Society of Chemistry, Cambridge, UK. pp. 285–327.

Gibson, G. and S. Muse (2004). A Primer of Genome Science. Sunderland, MA, Sinauer Associates Inc.

Jackson ImmunoResearch."Cyanine Dyes (Cy2, Cy3, & Cy5)".Retrieved on 2008-12-08.

Johnson DL, Ward SM. (1993) Quinoa. p. 219-221. In: J. Janick and J.E. Simon (eds.), New crops. Wiley, New York.

Johnson WE, Rabinovic A, Li C (2007). Adjusting batch effects in microarray expression data using Empirical Bayes methods. Biostatistics 8:118-127.

Johnson WE, Li C (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 8:118–127.

Kislinger T, Cox B, Kannan A, Chung C, Hu P, Ignatchenko A, Scott MS, Gramolini AO, Morris Q, Hallett MT, Rossant J, Hughes TR, Frey B, Emili A (2006) Global Survey of Organ and Organelle Protein Expression in Mouse: Combined Proteomic and Transcriptomic Profiling. Cell 125:173-186.

Kozian, D. and B. Kirschbaum (1999). "Comparative gene-expression analysis." Trends in Biotechnology 17: 73-78.

Lander, E. S. (1999) Array of Hope. Nat Genet, 21:3-4.

Lee, H.-S., J. Wang, et al. (2004). "Sensitivity of 70-mer oligonucleotides and cDNAs for microarray analysis of gene expression in Arabidopsis and its related species." Plant Biotechnology Journal 2: 45-57.

Loria R, Kers J, Joshi M (2006) Evolution of Plant Pathogenicity in Streptomyces. Annual Review Phytopathology 44:469–87.

Mandal P, Sinha Babu SP, Mandal NC(2005) Antimicrobial activity of saponins from Acacia auriculiformis. Fitoterapia76:462-465.

Masterbroek, H. D., H. Limburg, et al. (2000). "Occurrence of sapogenins in leaves and seeds of quinoa (Chenopodium quinoa Willd)." Journal of the Science of Food and Agriculture 80: 152-156.

Matlow, S. (2006). " Agilent Technologies launches next-generation DNA microarray manufacturing process to drive emerging applications." from http://www.agilent.com/about/newsroom/presrel/2005/27oct-ca05062.html.

Mujica, A., M. Saturning, et al. (2003). "Current production and potential of quinoa (Chenopodium quinoa Willd.) in Peru." Food Reviews International 19(1-2): 149-154.

Nielsen, H., R. Wernersson, et al. (2003). "Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays." Nucleic Acids Research 31: 3491–3496.

Osbourn, A. (2003). "Molecules of interest: saponins in cereals." Phytochemistry 62: 1-4.

Osbourn A. 1996. Saponins and plant defense—a soap story. Trends Plant Sci. 1:4–9

Papadopoulou, K., Melton, R. E., Leggett, M., Daniels, M. J. & Osbourn, A. E. (1999). "Compromised disease resistance in saponin-deficient plants." Proceedings of the National Academy of Sciences of the United States of America 96: 12923-12928.

Prado, F. E., C. Boero, et al. (2000). "Effect of NaCl on germination, growth, and soluble sugar content in Chenopodium quinoa Willd. seeds." Botanical Bulletin Of Academia Sinica 41: 27-34.

Rabban MA, Maruyama K, Abe H, Khan MA, Katsura K, Ito Y, Yoshiwara K, Seki M, Shinozaki K, Yamaguchi-Shinozaki K (2003) Monitoring Expression Profiles of Rice Genes under Cold, Drought, and High-Salinity Stresses and Abscisic Acid Application Using cDNA Microarray and RNA Gel-Blot Analyses. Plant Physiology 133:1755-1767

Repo-Carrasco, R., C. Espinoza, et al. (2003). "Nutritional value and use of the Andean crops quinoa (Chenopodium quinoa) and kañiwa (Chenopodium pallidicaule)." Food Reviews International 19(1-2): 179-189.

Ricks, M. (2005). "Genetic Mapping of the Bitter Saponin Production Locus (BSP Locus) in Chenopodium quinoa Willd." MS Thesis Brigham Young University, Provo, UT

Rishi, A., N. Nelson, et al. (2002). "DNA Microarrays: Gene Expression Profiling in Plants." Reviews in Plant Biochemistry and Biotechnology 1: 81-100.

Ruales, J. and B. M. Nair (1992). "Nutritional quality of the protein in quinoa (Chenopodium quinoa Willd.) seeds." Plant Foods and Human Nutrition 42: 1-11.

Ruales, J. and B. M. Nair (1993). "Content of fat, vitamins and minerals in quinoa (Chenopodium quinoa, Willd) seeds." Food Chemistry 48: 131-136.

Sierzchala, A., D. Dellinger, et al. (2003). "Solid-Phase Oligodeoxynucleotide Synthesis: A Two-Step Cycle Using Peroxy Anion Deprotection." J. Am. Chem. Soc. 125(44): 13427-13442.

Slonim DK.(2002) From patterns to pathways: gene expression data analysis comes of age. Nature Genetics Supplement 32:502-508

Song JS, Johnson WE, Zhu X, Zhang X, Li W, Manrai, AK, Liu JS, Chen R, Liu XS (2007). Model-based analysis of two-color arrays (MA2C). Genome Biology, 8:R178.

Suzuki, H., L. Achnine, et al. (2002). "A genomics approach to the early stages of triterpene saponin biosynthesis in Medicago truncatula." The Plant Journal 32: 1033–1048.

Thompson T. (2000) Questionable Foods and the Gluten Free Diet Survey of Current Recommendations . Journal of the American Dietetic Association , Volume 100:463-465.

Vacher, J. J. (1998). "Responses of two main Andean crops, quinoa (Chenopodium quinoa Willd) and papa amarga (Solanum juzepczukii Buk.) to drought on the Bolivian Altiplano: Significance of local adaptation." Agriculture, Ecosystems and Environment 68: 99-108.

Wang, J., L. Hu, et al. (2003). "RNA Amplification Strategies for cDNA Microarray Experiments" BioTechniques 34: 394-400.

Ward, S. M. (2001). "A recessive allele inhibiting saponin synthesis in two lines of Bolivian quinoa (Chenopodium quinoa Willd.)." The Journal of Heredity 92(1): 83-86.

Woldemichael, G. M. and M. Wink (2001). "Identification and biological activities of triterpenoid saponins from Chenopodium quinoa." Journal of Agricultural and Food Chemistry 49: 2327-2332.

Zhu, N., S. Sheng, et al. (2002). "Triterpene Saponins from debittered quinoa (Chenopodium quinoa) seeds." Journal of Agricultural and Food Chemistry 50: 865-867.

FIGURES

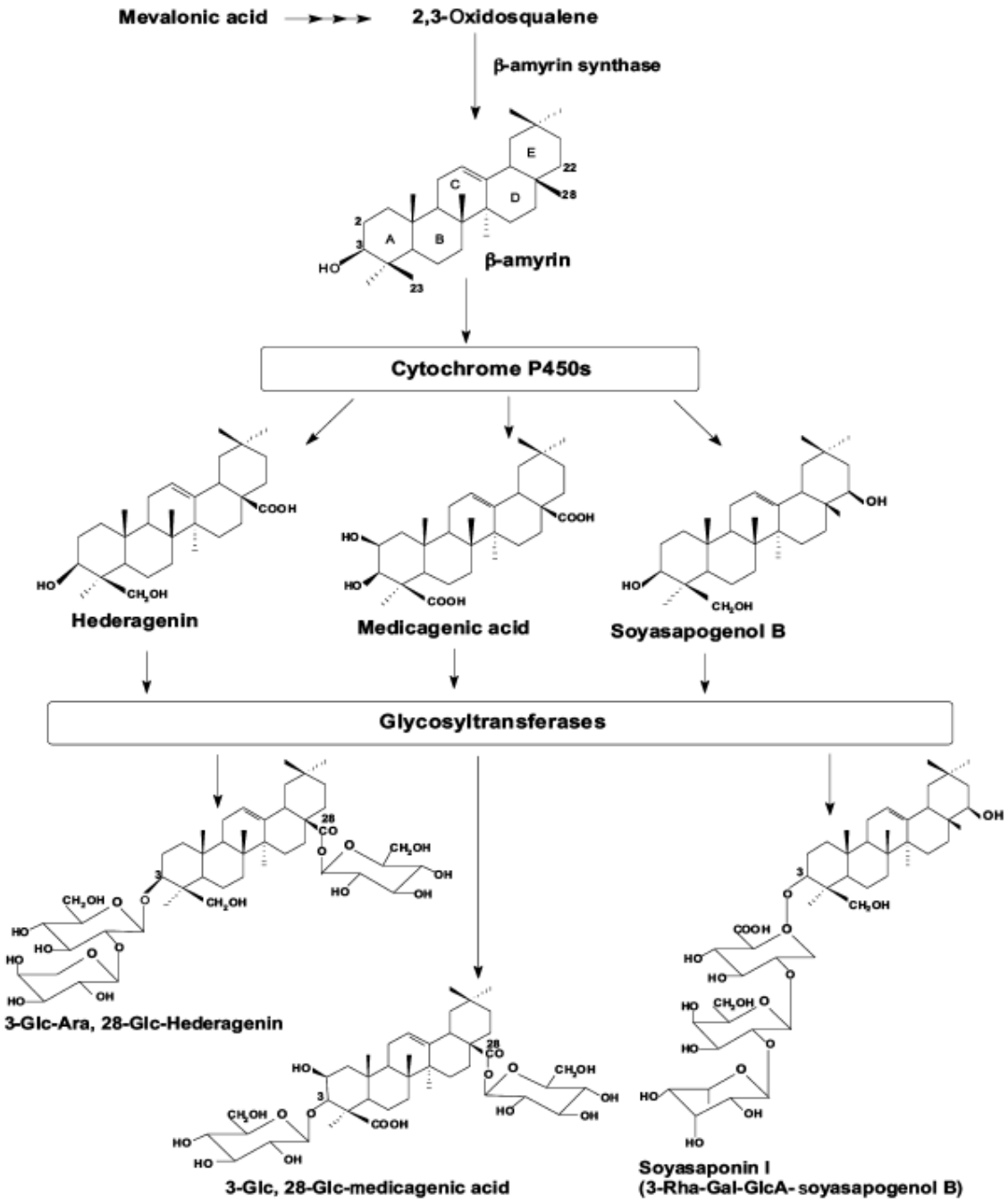
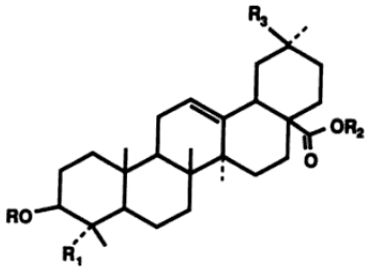


FIGURE 1. Example of the saponin biosynthetic pathway in Medicago truncatula. B-amyrin is converted to aglycones (three of which are shown), which are converted by glycosyltransferases to many different triterpene saponins (Achnine 2005).



R	R_1	R_2	R_3
Ara-	$-CH_2OH$	Glc-	$-CH_3$
Glc(1→3)ara-	$-CH_2OH$	Glc-	$-CH_3$
Glc(1→3)ara-	$-CH_2OH$	Glc-	$-CH_3$
Ara-	$-CH_2OH$	Glc-	$-COOCH_3$
Glc(1→3)ara-	$-CH_2OH$	Glc-	$-COOCH_3$
Glc(1→3)gal-	$-CH_2OH$	Glc-	$-COOCH_3$
Glc(1→2)glc(1→3)ara-	$-CH_3$	Glc-	$-COOCH_3$
H-	$-CH_3$	-H	$-COOCH_3$
Glc(1→2)glc(1→3)ara-	$-CH_3$	-H	-COCH
Glc(1→2)glc(1→3)ara-	$-CH_3$	Glc-	$-CH_3$
H-	$-CH_3$	-H	$-CH_3$
Glc(1→2)glc(1→3)ara-	$-CH_2OH$	-H	$-COOCH_3$
-H	$-CH_2OH$	-H	$-COOCH_3$
GlcUA-	$-CH_3$	-Glc-	$-CH_3$
GlcUA-	$-CH_2OH$	-Glc-	$-CH_3$
-H	$-CH_2OH$	-H	$-CH_3$
Xyl(1→3)glcUA-	$-CH_3$	-Glc-	$-CH_3$
Xyl(1→3)glcUA-	$-CH_2OH$	-Glc-	$-CH_3$

FIGURE 2. Typical saponin structures found in quinoa (Fenwick et al. 1991).

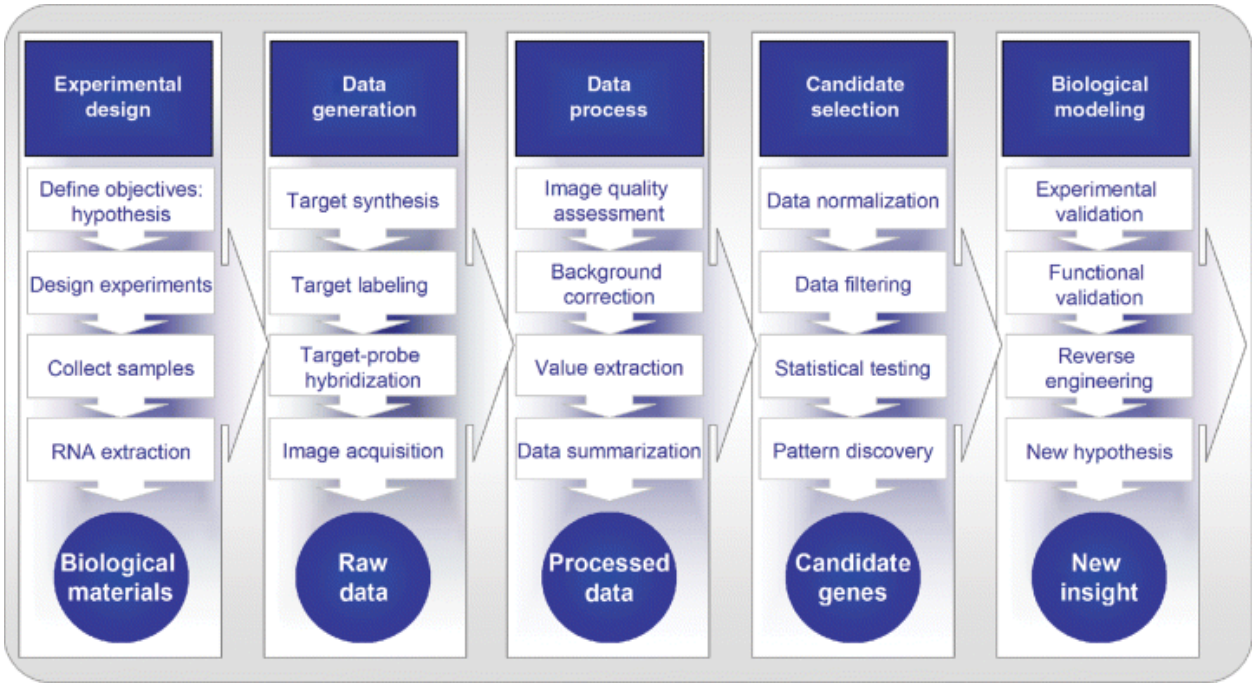


FIGURE 3 . An overview of microarray technology use in providing new understanding of biology concepts (Clarke, Zhu 2006).

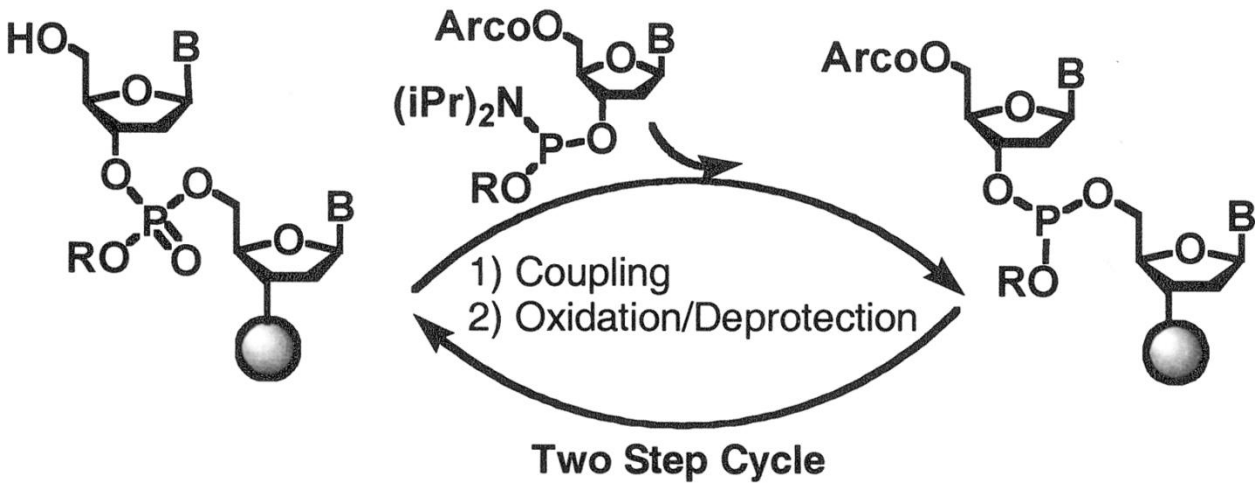


FIGURE 4. Illustration of solid phase phosphoramidite chemistry. (Image from Sierzchala, Dellinger et al. 2003).