2017

Developing specialty corn for niche markets in the public sector: A story of tradition and innovation

Hannah Worral
Iowa State University

Developing specialty corn for niche markets in the public sector: A story of tradition and innovation

by

Hannah Marie Worral

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Genetics and Genomics

Program of Study Committee:
Marvin Paul Scott, Co-Major Professor
Jianming Yu, Co-Major Professor
Jode Edwards

Iowa State University

Ames, Iowa

2017

DEDICATION

To my loving parents Bernie and Carolyn, for always encouraging me in my quest for knowledge, and for giving me a firm foundation upon which to build my life.

Also, to my dearest Jon, for lending a patient ear and an encouraging word when I needed them the most, and for his endless support through all the highs and lows of grad school.

Finally, in memory of Paula Newhouse, whose infectious smile could light up the room and whose heart warmed all those around her. You are dearly missed.

TABLE OF CONTENTS

NOMENCLATURE

o2	*opaque2*
fl2	*floury2*
QPM	Quality Protein Maize
CIMMYT	International Maize and Wheat Improvement Center
CL	Commercial line
Ga1-S	*Gametophytic Incompatibility Factor – Strong*
NCBI	National Center for Biotechnology Information
SRA	Sequence Read Archive
GUI	Graphical User Interface
TASSEL	Trait Analysis by Association, Evolution and Linkage

ACKNOWLEDGMENTS

For their patience and genuine investment in my future, I would like to offer thanks to my major professor Paul Scott, co-major professor Jianming Yu and committee member Jode Edwards. Gratitude is also due my friends and colleagues both in Agronomy Hall and across the Interdepartmental Genetics and Genomics program for their comradery, understanding, and support. Finally, I wish to thank the various farmers who grow acre upon acre of coffee trees, the hired hands who harvest the fruits, and everyone involved along the way to bring fresh ground coffee to the local grocery stores in Ames. Without their dedication to agriculture, this thesis may have taken a lot longer to write!

ABSTRACT

Humans derive more than half their dietary protein from cereal grain sources with maize comprising nearly 40% of the total cereal grain tonnage. Despite the market prevalence of this crop and the improvements we have seen in yields since hybrid production of maize started in the early 20th century, maize remains a rather poor protein source, providing only one-fifth the amount of lysine – the most limiting amino acid in maize – required for optimal human nutrition. The effect of this lower protein quality is moderated by the consumption of other high-protein sources, but in regions like sub-Saharan Africa where maize constitutes over 20% of the daily energy intake (DEI) and alternative protein sources are not readily available, there is a high risk of protein malnutrition, especially in countries where the DEI exceeds 50% like Lesotho, Zambia, and Malawi. Following the discovery of *opaque2* (*o2*) – a chalky endosperm mutant with high levels of lysine and tryptophan relative to common maize – and modifier genes that mitigated the negative effects associated with *o2*, the International Maize and Wheat Improvement Center (CIMMYT) began the development of quality protein maize (QPM) in the 1960s to help combat undernutrition in developing countries. Maintenance of the preferential amino acid profile in QPM, however, requires that the endosperm be homozygous for *o2*. Genetic purity is essential to QPM production as foreign pollen from neighboring fields that lacks the *o2* allele can contaminate a QPM field, resulting in heterozygous seeds lacking the preferential amino acid profile. Thus, we are pursuing avenues for the incorporation of a gametophytic cross-incompatibility (GCI) system into

future QPM lines. *Gametophyte factor1-Strong* (*Ga1-S*), currently the most well-understood of these systems in maize, has been used by the popcorn industry since its discovery in the early 1900s. To deepen our understanding of these systems and their genetic and biochemical underpinnings, we have developed a suite of bioinformatics tools that capitalizes on the sequence read archive (SRA) and facilitates ancillary analyses of data, which will allow us to incorporate archived sequence data alongside newly generated data to expand our ability to call genotypes across diverse samples.

THESIS ORGANIZATION

Chapter One gives a brief introduction to maize and its nutritional quality, the use of gametophytic-incompatibility systems to maintain genetic purity in specialty corn, and an overview of DNA sequencing technologies and the potential for use of archived data sets in future studies.

Chapter Two is published in the *Journal of Plant Registrations* and documents the development of 9 quality protein maize (QPM) lines for use in the Midwest. I was responsible for the statistical analysis of the agronomic data collected over the period of development for these lines, choosing the top lines for release, and continuing to collaborate with Drs. Hallauer and Scott on the advancement of these and other BQPM lines including making field selections for advancement and generating hybrid combinations for yield testing.

Chapter Three introduces a suite of tools that I developed for the analysis of archived sequence data files in conjunction with data that is currently being generated. This was developed in order to offer a more user-friendly alternative to sequence processing, and makes use of mainstream data processing tools to enable the user to automate their data analysis to a greater extent than is currently possible.

CHAPTER 1. INTRODUCTION: THE IMPORTANCE OF MAIZE

Humans derive around half of their dietary protein from cereal grains and in developing countries that percentage can jump up to the 70% mark (Lutz et al., 2001). Maize commands a large proportion of the cereal grains market – nearly 40% – with 8.2 billion tons produced worldwide in 2014 alone (FAO, 2014). Despite being a staple crop for more than 200 million people, maize like other cereal grains lacks sufficient amounts of lysine and tryptophan which are essential for human and animal nutrition (Nuss & Tanumihardjo, 2010; Gibbon & Larkins, 2005).

Nutritional Quality of Maize

Despite its status as a staple crop and its prevalence as a major cereal grain – responsible for providing a large proportion of the dietary protein consumed by humans worldwide – common maize is a poor source of protein. The nutritional value of a protein source can be estimated by comparing it to milk casein with a protein efficiency ratio (PER) of 100%, and common maize falls well below the 50% mark while rice hovers just below 80% (Bressani, 1992; FAO, 1992). Total protein has been decreasing at an average rate of 0.3% per 10 years since the 1920s, but a greater concern is that the amino acid composition has not improved in this time: lysine and tryptophan remain limiting amino acids for monogastric animals, with lysine alone only present at one-fifth the amount required for optimal human nutrition (Obsborne & Mendel, 1914; Baker et al., 1969; Lewis et al., 1982; Young et al., 1998; Duvick et al., 2004; Scott et al., 2006). Efforts to

improve the PER of maize have focused on increasing these limiting amino acids (Frey, 1951; Gibbon & Larkins, 2005).

opaque 2 and Quality Protein Maize (QPM)

The discovery that the *opaque2* (*o2*) mutation increases levels of lysine and tryptophan in the endosperm, precipitated a number of studies that investigated the impact of this mutation on the nutritive value of maize (Mertz et al., 1964; Nelson et al., 1965; Wolf et al., 1969; Robutti et al., 1974; Geetha et al., 1991; Habben et al., 1993). The reduction in the zein (prolamine) protein fraction of the endosperm observed in *o2/o2* genotypes leads to an increase in the fraction of proteins containing lysine and tryptophan as the zein fraction contains lower levels of these amino acids (Mertz et al., 1964; Habben et al., 1993; Huang et al., 2005). While incorporation of the *o2* mutation into common maize lines led to higher levels of lysine and tryptophan in the endosperm, it also resulted in lower yields and chalky endosperm that was more susceptible to insect, fungal, and mechanical damage (Brown et al., 1988; Bjarnason & Vasal, 1992; Vasal, 2001; Ignjatovi-Micic et al., 2009). The discovery of *o2* modifier genes, however, enabled plant breeders led by the International Maize and Wheat Improvement Centre (CIMMYT) to develop maize – called quality protein maize (QPM) (Paez, et al., 1969; Vasal, et al., 1980; Prasanna, et al., 2001). Inheritance of these modifiers is complex and little is known about their mode of action, however several studies have shown a connection between higher levels of 27-kDa ɣ-zeins and the hard endosperm observed in QPM (Geetha, et al., 1991; Lopes, 1991). Additionally, the increased yields and improved protein quality observed in QPM are genetically distinct from the *o2* modifiers

(Vasal, et al., 1980; Wessel-Beaver, 1985; Pixley, 2002). Several promising candidate loci have been identified, but the identity of the *o2* modifiers remains elusive (Gibbon, 2005; Holding, 2008).

Regardless, these breeding efforts saw an increase in PER from ~40% in common maize to over 80% in *o2* maize and QPM (FAO, 1992; Nuss & Tanumihardjo, 2010). These efforts continue today as researchers investigate the genetic and biochemical mechanisms of *o2* and explore new avenues for its use and incorporation in breeding programs around the world (Zarkadas et al., 2000; Bhatnagar et al., 2004; Krivanek et al., 2007; Ngaboyisonga et al., 2009). It has been my privilege to be involved in one such effort – the development of 9 QPM lines adapted for growth in the Midwest for use in niche markets (Worral et al., 2015).

Pollen Contamination

As the improved balance of amino acids in QPM is contingent on the homozygosity of the recessive allele of *opaque2*, it is important to maintain genetic purity in the production field to ensure that the seed being produced contains the altered protein composition. Should pollen from a foreign source lacking the *o2* allele land on a silk and produce a successful fertilization event, the resulting progeny will lack the characteristic amino acid profile that is quintessential to QPM. Genetic purity can be accomplished by isolating the field or performing pollinations by hand, but as distances close between plots of cropland this becomes more difficult.

Gametophytic Cross-incompatibility Systems

Gametophytic cross-incompatibility (GCI) systems – which prevent cross-pollinations from sources containing opposing genotypes – have the potential to be applied to specialty crops like QPM to ensure their genetic purity. One such GCI system is *Gametophyte factor1-Strong* (*Ga1*-S). First noted in South American-derived popcorns, it has since been mapped to the short arm of chromosome 4 and is used extensively in popcorn breeding to maintain purity of type (Mangelsdorf & Jones, 1926; Thomas, 1955; Ziegler & Ashman, 1994; Bloom & Holland, 2011; Zhang et al., 2012; Liu et al., 2014). Three haplotypes – *Ga1-S Ga1-M*, and *ga1* – function in concert with one another or discordantly depending on the interaction. This is a dual-function system wherein the female function presents a barrier to prevent successful pollination when off-type pollen is present, and the male function overcomes this female barrier to achieve a successful pollination (Nelson, 1952; Kermicle & Evans, 2005; Kermicle, 2006). Plants homozygous for *Ga1-S* display both male and female functions. Thus, pollinations arising from *Ga1-S* pollen – whether self-pollination or cross-pollination – will be successful while pollinations initiated by *ga1* pollen will be unsuccessful. Plants homozygous for *ga1*, on the other hand, contain neither the female nor the male functions and so pollinations arising from both *ga1*, *Ga1-S*, and *Ga1-M* pollen will be successful. A plant homozygous for *Ga1-M*, on the other hand, only possesses the male function. Thus, it can pollinate and be successfully pollinated by any of the three haplotypes – *Ga1-S*, *Ga1-M*, or *ga1* (Jimenez & Nelson, 1964; Kermicle & Evans, 2010).

The Sequence Read Archive (SRA)

Established in 2009 by the International Nucleotide Sequence Database Collaboration (INSDC) in response to the growing number of large data sets being generated following the next-generation sequencing revolution, the SRA is maintained by the DNA Data Bank of Japan (DDBJ), the European Bioinformatics Institute (EBI), and the National Center for Biotechnology Information (NCBI). Since its debut, it has grown from just 6.5 terabases in 2009 to nearly 9.7 petabases in 2017 (Kodama et al., 2012; NCBI, 2017). Sequencing approaches and technologies continue to improve and offer more cost-effective options to researchers investigating the genetic underpinnings of biological mechanisms (Sanger & Coulson, 1975; Craig et al., 2008; Elshire et al., 2008; Rothberg et al., 2008; Illumina, 2015; Rhoads and Au, 2015). Access to computational resources commensurate to the collection, management, and analysis of the growing body of sequence data then becomes a limiting factor in the advancement of our understanding of the underlying genetic components of living organisms (Oyler-McCance, 2016; Muir, 2016; Nakazatoet al., 2013). As researchers begin to realize the utility of resources like the SRA, the need for proper computational infrastructure is only amplified. Whether reusing previously-generated sequence data to improve sequencing coverage, improving the efficiency of calling SNPs, or expanding the utility of current data sets and presenting them in a format that is more readily usable for additional analysis, bioinformatic tools and access to computational platforms enable researchers to apply their data sets to far more than the initial investigation that led to the generation of those sequence reads in the first place (Schlautman, 2017; Boyles, 2016; Torkamaneh, 2017; Clark, 2016). By

improving accessibility to the SRA and developing tools for the analysis of archived data in conjunction with newly-generated data, we can broaden the impact of each sequence file that is generated, all while cutting costs and reducing time and effort spent on data generation and analysis. We propose CONIFER – a suite of tools for the bioinformatic analysis of SRA-derived data files and their incorporation into ongoing genetic and genomic studies – as one such innovation to be implemented in the application of archived sequence data to a bevy of ancillary investigations that expand our knowledge of the natural world and the organism that life within it.

References

Baker, D., Becker, D., Norton, H., Jensen, A., & Harmon, B. (1969). Lysine imbalance of corn protein in the growing pig. *J. Anim. Sci., 28*, 23.

Bhatnagar, S., Betran, F., & Rooney, L. (2004). Combining abilities of quality protein maize inbreds. *Crop Sci., 44*, 1997-2005. doi:10.2135/cropsci2004.1997

Bjarnason, M., & Vasal, S. (1992). Breeding of quality protein maize (QPM)). *Plant Breed. Rev., 9*, 181-216.

Bloom, J., & Holland, J. (2011). Genomic localization of the maize cross-incompatibility gene, Gametophyte factor 1 (ga1). *Maydica*, 379-387.

Boyles, R. E. (2016). Genome-Wide Association Studies of Grain Yield Components in Diverse Sorghum Germplasm. *Plant Genome, 9*(2). doi:doi:10.3835/plantgenome2015.09.0091

Bressani, R. (1992). Nutritional value of high-lysine maize in humans. In E. Mertz (Ed.), *Quality protein maize* (pp. 204-224). St. Paul, MN: Am. Assoc. Cereal Chem.

Brown, W., Bressani, R., Glover, D., Hallauer, A., Johnson, V., & Qualset, C. (1988). *Quality-protein maize: report of an ad hoc panel of the advisory committee on technology innovation, Board on Science and Technology for International Development National Research Council, in cooperation with the Board on Agriculture National Research Co.* Washington, D.C.: National Academy Press.

Clark, L. V. (2016). TagDigger: user-friendly extraction of read. *Source Code for Biology and Medicine, 11*(11). doi:10.1186/s13029-016-0057-7

Craig, D. P. (2008). Identification of genetic variants using barcoded multiplexed sequencing. *Nat Methods, 5*(10), 887-893. doi:10.1038/nmeth.1251

Duvick, D., Smith, J., & Cooper, M. (2004). Long-term selection in a commercial hybrid maize breeding program. *Plant Breed. Rev., 24*, 109-151.

Elshire, R. G. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS). *PLoS ONE, 6*(5), e19379. doi:10.1371/journal.pone.0019379

FAO. (1992). Chemical composition and nutritional value of maize. In *Maize in human nutrition.* Rome. Retrieved from http://www.fao.org/docrep/T0395E/T0395E00.htm

FAO. (2014). *FAOSTAT, Production*. Retrieved April 12, 2017, from http://www.fao.org/faostat/en/#data/QC

Frey, K. (1951). The interrelationships of proteins and amino acids in corn. *Cereal Chem., 28*, 123-132.

Geetha, K., C. L., Lopes, M., Wallace, J., & Larkins, B. (1991). opaque-2 modifiers increase gamma-zein synthesis and alter its. *Plant Cell., 3*, 1207-1219.

Gibbon, B., & Larkins, B. (2005). Molecular genetic approaches to developing quality protein maize. *Trends in Genetics, 21*, 227-233.

Habben, J., Kirleis, A., & Larkins, B. (1993). The origin of lysinecontaining. *Plant Mol. Biol., 23*, 825-838. doi:10.1007/BF00021537

Huang, S., Kruger, D., Frizzi, A., D'Ordine, R., Florida, C., & W.R. Adams. (2005). High-lysine corn produced by the combination of enhanved lysine biosynthesis and reduced zein accumulation. *Plant Biotechnol. J., 3*, 555-569. doi:10.1111/j.1467-7652.2005.00146.x

Ignjatovi-Micic, D., Markovic, K., Ristic, D., Drinic, S., Stankovic, S., Lazic-Janic, V., & Denic, M. (2009). Variability analysis of normal and opaque2 maize inbred lines. *Genetika-Belgrade, 41*(1), 81-93. doi:10.2298/GENSR0901081I

Illumina. (2015) A wide variety of library prep methods derived from the scientific literature. Retrieved May 20, 2017. http://www.illumina.com/techniques/sequencing/ngslibrary-prep/library-prep-methods.html.

Jimenez, J., & Nelson, O. (1964). A fourth chromosome gametophyte locus in maize. *Journal of Heredity*, 259-263.

Kermicle, J. (2006). The gametophyte-1 locus and reproductive isolation among Zea mays subspecies. *Maydica*, 219-225.

Kermicle, J., & Evans, M. (2005). Pollen-pistil barriers to crossing in maize and teosinte result from incongruity rather than active rejection. *Sexual Plant Reproduction, 18*, 187-194. doi:10.1007/s00497-005-0012-2

Kermicle, J., & Evans, M. (2010). The Zea mays sexual compatibility gene ga2: Naturally occurring alleles, their distribution, and role in reproductive isolation. *Journal of Heredity*.

Kodama, Y., Shumway, M., & Leinonen, R. (2012). The sequence read archive: explosive growth of sequencing data. *Nucl. Acids Res., 40*, D54056. doi:10.1093/nar/gkr854

Krivanek, A., De Groote, H. G., Diallo, A., & Friesen, D. (2007). Breeding and disseminating quality protein maize (QPM) for Africa. *Afr. J. Biotechnol., 6*, 2312-324.

Lewis, A., Barnes, M., Grosbach, D., & Peo, E. (1982). Sequence in which the amino acids of corn (Zea mays) become limiting for growing rats. *J. Nutri., 112*, 782-788.

Liu, X., Sun, H., Wu, P., Tian, Y., Cui, D., Xu, C., . . . Chen, H. (2014). Fine mapping of the maize cross-incompatibility locus Gametophytic factor 1 (ga1) using a homogenous population. *Crop Science, 54*, 873-881.

Lopes, M., & Larkings, B. (1991). Opaque-2 modifiers increase gamma-zein synthesis and alter its spatial-distribution in maize endosperm. *Plant Cell 31*, 1655-1662

Lutz, W., Sanderson, W., & Scherbov, S. (2001). The end of world population growth. *Nature, 412*, 543-545.

Mangelsdorf, P., & Jones, D. (1926). The expression of Mendelian factors in the. *Genetics*, 423-455.

Mertz, E., Bates., L., & Nelson Jr, O. (1964). Mutant gene that changes protein composition and increases lysine content of maize endosperm. *Science, 145*, 279-280. doi:10.1126/science.145.3629.279

Muir, P. L. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology, 17*(53). doi:10.1186/s13059-016-0917-0

Nakazato, T., Ohta, T., & Bono, H. (2013). Experimental design-based functional mining and characterization of hight-throughput sequencing data in the seuence read archive. *PLoS ONE, 8*(10), e77910. doi:10.1371/journal.pone.0077910

NCBI. (2017). *SRA Growth Infographics*. (U. N. Medicine, Producer) Retrieved May 19, 2017, from National Center for Biotechnology Information: https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/

Nelson, O. (1952). Non-reciprocal cross-sterility in maize. *Genetics*, 101.

Nelson, O., Mertz, E., & Bates, L. (1965). Second mutant gene affecting the amino acid pattern of maize endosperm proteins. *Science, 150*, 1469-1470. doi:10.1126/science.150.3702.1469

Ngaboyisonga, C., Njoroge, K., Kirubi, D., & Guthiri, S. (2009). Effects of low nitrogen and drought on genetic parameters of grain yield and endosperm hardness of quality protein maize. *Asian J. Agric. Res., 3*, 1-10. doi:10.3923/ajar.2009.1.10

Nuss, E., & Tanumihardjo, S. (2010). Maize: A paramount staple crop in the context of global nutrition. *Comprehensive Reviews in Food Sci. and Food Safety, 9*, 417-436.

Obsborne, T., & Mendel, L. (1914). Amino acids in nutrition and growth. *J. Biol. Chem., 17*, 325.

Oyler-McCance, S. O. (2016). A field ornithologist's guide to genomics: Practical considerations for ecology and conservation. *The Auk, 133*(4), 626-648.

Paez, A., Helm, J., & Zuber, M. (1969). Lysine content of opaque-2 maize kernels having different phenotypes. *Crop Sci., 9*, 251. doi:10.2135/cropsci1969.0011183X0009000020045x

Pixley, K., & Bjarnason, M. (2002). Stability of grain yield, endosperm modification, and protein quality of hybrid and open-polinated quality protein maize (QPM) cultivars. *Crop Sci 42*, 1882-1890

Prasanna, B., Vasal, S., Kassahun, B., & Singh, N. (2001). Quality protein maize. *Curr. Sci., 81*, 1308-1319.

Rhoads, A., & Au, K. (2015). PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics, 13*(5), 278-289. https://doi.org/10.1016/j.gpb.2015.08.002

Robutti, J., Hoseney, R., & Deyote, C. (1974). Modified opaque-2 corn endosperms. I. Protein distribution and amino acid composition. *Cereal Chem., 51*, 163-172.

Rothberg, J., & Leamon, J. (2008). The development and impact of 454 sequencing. *Nature Biotechnology, 26*(10), 1117-1124. doi:10.1038/nbt1485

Sanger, F.. (1975). A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase. *J. Mol. Biol., 94*, 441-448.

Schlautman, B. C.-P.-G. (2017). Construction of a High-Density American Cranberry (Vaccinium macrocarpon Ait.) Composite Map Using Genotyping-by-Sequencing for Multi-pedigree Linkage Mapping. *G3: Genes, Genomes, Genetics, 7*(4), 1177-1189. doi:https://doi.org/10.1534/g3.116.037556

Scott, M., Edwards, J., Bell, C., Schussler, J., & Smith, J. (2006). Grain composition and amino acid content in maize cultivars representing 80 years of commercial maize varieties. *Maydica, 51*, 417-423.

Thomas, W. (1955). Transferring the Gas factor for dent incompatibility to dent compatible lines of popcorn. *47*, 440-441.

Torkamaneh, D. L. (2017). Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPd from genotyping-by-sequencing data. *BMC Bioinformatics, 18*(5). doi:10.1186/s12859-016-1431-9

Vasal, S. (2001). High quality protein corn. In A. Hallauer (Ed.), *Specialty corns* (pp. 85-129). Boca Raton, FL: CRC Press.

Vasal, S., Vilegas, E., Bjarnason, M., Gelaq, B., & Goertz, P. (1980). Genetic modifiers and breeding strategies in developing hard endosperm opaque-2 materials. In W.

Pollmers, & R. Philips (Eds.), *Quality traits of maize for grain and silage use* (pp. 37-73). London: Martinus Nijhoff.

Wessel-Beaver, L. (1985). Genetic variability and correlations in a modified endosperm texture opaque-2 maize population. *Crop Sci. 25*, 129-132.

Wolf, M., Khoo, U., & Seckinger, H. (1969). Distribution and subcellular structure of enodsperm protein in varieties of ordinary and high-lysine maize. *Cereal Chem., 46*, 253-263.

Worral, H., Hallauer, A., & Scott, M. (2015). Registration of temperate quality protein maize (QPM) lines BQPM9, BQPM10, BQPM11, BQPM12, BQPM13, BQPM14, BQPM15, BQPM16, and BQPM17. *J. of Plant Reg., 9*(3), 371-375.

Young, V., Scrimshaw, N., & Pellett, P. (1998). Significance of dietary protein source in human nutrition: animal or plant proteins? In J. Waterlow, D. Armstrong, L. Fowden, & R. Riley (Eds.), *Feeding a world population of more than eight billion people: a challenge to science* (pp. 205-221). New York: Oxford University Press.

Zarkadas, C., Hamilton, R., Yu, Z., Choi, V., Khanizadeh, S., Rose, N., & Pattinson, P. (2000). Assessment of the protein quality of 15 new northern adapted cultivars of quality protein maize using amino acid analysis. *J. Agric. Food Chem., 48*, 5351-5361. doi:10.1021/jf000374b

Zhang, H., Liu, X., Zhang, Y., Jiang, C., Cui, D., Liu, H., Li, D., Wang, L., Chen, T., Ning, L., Ma, X., & Chen, H. (2012). Genetic analysis and fine mapping of the Ga1-s gene region conferring cross-incompatibility in maize. *Theoretical and Applied Genetics*, 459-465.

Ziegler, K., & Ashman, B. (1994). Popcorn. In A. Hallauer, & A. Hallauer (Ed.), *Specialty Corns* (pp. 189-223). Boca Raton: CRC Press.

CHAPTER 2. REGISTRATION OF TEMPERATE QUALITY PROTEIN MAIZE (QPM) LINES BQPM9, BQPM10, BQPM11, BQPM12, BQPM13, BQPM14, BQPM15, BQPM16, AND BQPM17

Article published in *Journal of Plant Registrations* 9(3):371-375 · September 2015

Hannah M. Worral[1,2], M. Paul Scott[3], and Arnel R. Hallauer[2]

Abstract

The discovery of the *opaque2* (*o2*) mutation and *o2* modifier genes in maize (*Zea mays* L.) has resulted in the development of Quality Protein Maize (QPM) lines with increased lysine and tryptophan content. The QPM lines BQPM9 (Reg. No. GP-584, PI 671795), BQPM10 (Reg. No. GP-585, PI 671796), BQPM11 (Reg. No. GP-586, PI 671797), BQPM12 (Reg. No. GP-587, PI 671798), BQPM13 (Reg. No. GP-588, PI 671799), BQPM14 (Reg. No. GP-589, PI 671800), BQPM15 (Reg. No. GP-592, PI 673348), BQPM16 (Reg. No. GP-590, PI 671801), and BQPM17 (Reg. No. GP-591, PI 671802) were developed jointly by Iowa State University and the USDA-ARS to address the lack of QPM lines adapted to the US Corn Belt. These lines originated from crosses made between two QPM lines from the International Maize and Wheat Improvement

[1] Primary Author

[2] Department of Agronomy, Iowa State University, Ames, IA 50011

[3] USDA-ARS, Corn Insects and Crop Genetics Research, Iowa State University, Ames, IA 50011

Center (CIMMYT) (CLQ06901 and CLRQ00502) and six inbred lines released by Iowa State University (B91, B97, B98, B99, B100, and B113). Increased lysine and tryptophan content, characteristics associated with the presence of the *o2* mutation, and agronomic performance were used as selection criteria in the development of the nine BQPM lines released herein.

Abbreviations

CIMMYT, International Maize and Wheat Improvement Center; CL, commercial line; *fl2*, *floury2*; *o2*, *opaque2*; QPM, Quality Protein Maize

Introduction

The maize community has sought to improve the nutritional quality of maize (*Zea mays* L.), one of the world's staple food crops, for more than a century. While their deficiencies in maize have been documented for nearly 100 years, lysine and tryptophan continue to be limiting amino acids in the utilization of maize as a balanced source of protein for both human and animal consumption (Osborne and Mendel, 1914; Baker et al., 1969; Lewis et al., 1982). Various strategies have been applied effectively to the problem, including recurrent selection for higher amino acid concentrations (Choe et al., 1976; Scott et al., 2008), use of transgenic techniques to increase specific limiting amino acids (Lai and Messing, 2002; Huang et al., 2005; Houmard et al., 2007; Bicar et al., 2008; Tang et al., 2013), and supplementation of normal maize with soybean [*Glycine max* (L.) Merr.], synthetic methionine, and synthetic lysine. The latter is the simplest but also possibly the costliest option to achieve an optimal balance of amino acids in the diet.

The cost reduction seen when Quality Protein Maize (QPM) is substituted for normal maize is appealing not only for large farms and corporations looking to minimize input expenses but also for small farms that may already rely on maize as the sole feed component (López-Pereira, 1993; Nyanamba et al., 2003). Similarly, QPM developed for human consumption has the potential to positively affect many countries where maize is a staple of the diet (Krivanek et al., 2007; Gunaratna et al., 2010).

Naturally occurring mutations, such as *opaque2* (*o2*) and *floury2* (*fl2*), decrease the lysine and tryptophan-poor zein (prolamine) protein fraction present in the endosperm of a mature kernel, resulting in proportionately higher levels of these limiting amino acids (Mertz et al., 1964; Nelson et al., 1965; Geetha et al., 1991; Munck, 1992; Habben et al., 1993). Following the discovery of *o2* and *fl2* in the 1960s, several nutritional studies were conducted to investigate the nutritional value of these mutants compared with normal maize (Beeson et al., 1966; Pickett, 1966; Cromwell et al., 1967). Despite the higher amino acid levels, the soft, chalky endosperm characteristic of these mutations was more susceptible to fungal ear rots, lower yielding, and unappealing to maize growers (Bjarnason and Vasal, 1992; Vasal, 2001; Ignjatovi-Micic et al., 2009). With the discovery of *o2* modifier genes, however, maize breeders were able to produce higher yielding, lysine-rich germplasm that lacked the characteristic opaque endosperm and is now designated QPM (Paez et al., 1969; Vasal et al., 1980; Mertz, 1992; Prasanna et al., 2001).

Despite progress in adapting these QPM lines to various environments, there is little documentation for QPM lines that are well-adapted to the US Corn Belt (Zarkadas

et al., 2000; Bhatnagar et al., 2004; Ngaboyisonga et al., 2009). Our goal was to develop temperate QPM lines that were well-adapted to the US Corn Belt. We used two QPM lines (CLRQ00502 and CLQ06901) from the International Maize and Wheat Improvement Center (CIMMYT) as donors of *o2* and endosperm modifier genes and public inbred lines released by Iowa State University to develop temperate QPM lines BQPM9 (Reg. No. GP-584, PI 671795), BQPM10 (Reg. No. GP-585, PI 671796), BQPM11 (Reg. No. GP-586, PI 671797), BQPM12 (Reg. No. GP-587, PI 671798), BQPM13 (Reg. No. GP-588, PI 671799), BQPM14 (Reg. No. GP-589, PI 671800), BQPM15 (Reg. No. GP-592, PI 673348), BQPM16 (Reg. No. GP-590, PI 671801), and BQPM17 (Reg. No. GP-591, PI 671802).

Line Development

BQPM9, BQPM10, BQPM11, BQPM12, BQPM13, BQPM14, BQPM15, BQPM16, and BQPM17 are QPM lines adapted to the US Corn Belt derived from two QPM sources (CLQ06901 and CLRQ00502, developed at CIMMYT) and six Iowa inbred lines (B91, B97, B98, B99, B100, and B113) (Table 1; Russell, 1989; Hallauer et al., 1994, 1995, 1997, 1998, 2001). The QPM × Iowa inbred F_1 generation and one backcross to each Iowa inbred were conducted at CIMMYT in Mexico. Backcrosses were then planted at the Ames, IA, nursery in spring 2002 and self-pollinated. Subsequent generations (S_1–S_7) were planted ear-to-row each season in the nursery at Ames. Codominant simple-sequence repeat (SSR) markers *phi057* and *umc1066* were used to confirm the presence of the *opaque2* gene, and the presence of the *o2* modifiers was maintained by scoring kernels for percentage opacity on a light box (Babu et al., 2005). Only those kernels with

opacity scores of 1 to 2 out of 5 (i.e., <25% opaque) were advanced to the next generation. Microbial amino acid assays were also used in some years to evaluate the lysine, methionine, and tryptophan concentrations of the selected kernels (Scott et al., 2004, 2009).

At the S_3 generation, a North Carolina design II mating design was used to produce hybrids to assess the combining ability of the developing temperate QPM lines (Comstock and Robinson, 1952); the results are presented in Scott et al. (2009). Yield trials of test cross hybrids were then conducted in 2007 and 2008 using two-row plots with 0.762-m spacing between rows and 4.57-m plot length (8.36-m^2 plot size) and planting densities similar to regional commercial production (~65,000 plants ha^{-1}). Planting occurred at two locations in 2007 (Crawfordsville and Carroll, IA) and three locations in 2008 (Ames, Crawfordsville, and Carroll, IA). Plants were evaluated for stalk and root lodging, grain yield, and moisture content. In addition, amino acid balance of the BQPM inbred lines per se was shown to be typical of QPM, with elevated levels of lysine and tryptophan (Scott et al., 2009).

To assess the yield potential of the BQPM lines, an additional yield trial was initiated in 2009 following the same plot design as previously. In this trial, nine commercial checks and 39 BQPM lines were used as the female parent, and test crosses of those lines were made using three commercial, non-QPM inbred lines as male testers. The resulting test-cross lines were assessed for stalk and root lodging and the yield and

moisture content of the grain (Table 2). Trials were conducted over a tristate area encompassing Iowa (Atlantic and Slater), Illinois (Mt. Pulaski), and Nebraska (York).

Characteristics

The nine BQPM lines herein released were selected on the basis of the presence of the *o2* and *o2* modifier genes and their agronomic performance when grown in the US Corn Belt. BQPM lines and commercial inbred checks were crossed to the same three testers so that BQPM hybrid performance could be compared with that of commercial hybrids. Across the nine BQPM lines released, there was an 11.5% ($p = 0.05$) drop in yield compared with the commercial checks. The top three yielding BQPM lines by tester were not significantly different in yield than the commercial checks ($p = 0.05$; Table 3). The average yield across the three testers for the BQPM lines was 6377.8 kg ha^{-1}, with BQPM9 hybrids having the highest average yield at 6685.4 kg ha^{-1} and BQPM16 hybrids having the lowest at 6041.4 kg ha^{-1} (Table 2). Grain density, stalk lodging, and root lodging were not significantly different between the BQPM hybrids and their commercial line (CL) hybrid counterparts when averaged across the testers ($p < 0.05$). Percentage moisture varied across both CL and BPQM hybrids. The top-yielding BQPM hybrids for each tester had similar yields ($p < 0.05$) to hybrids derived from commercial stock and crossed to the same testers (Table 3). These results coincide with similar results reported in Atlin et al. (2011) concerning the agronomic performance of QPM lines versus contemporary non-QPM lines, which demonstrate that there is little to no disadvantage associated with QPM traits pertaining to yield and other important agronomic traits.

Single samples for each of the nine BQPM lines were evaluated by the AOAC standard method for amino acid concentrations, and crude protein per line was determined via combustion analysis at the University of Missouri Experiment Station Chemistry Laboratory (Method 982.30 E(a,b,c), Horwitz, 2005a; Method 9990.03, Horwitz, 2005b) before release (Table 4). The results show a range of 0.28 to 0.51 for percentage (w/w) lysine and 0.06 to 0.09 for percentage (w/w) tryptophan. As described in Scott et al. (2009), ranges for normal inbred lines fall between 0.29 and 0.33% (w/w) lysine and 0.06% (w/w) tryptophan. In general, BQPM10 had the highest amino acid concentrations among the lines and also the highest crude protein content. BQPM14 had the lowest amino acid concentrations in general, which coincided with the lowest crude protein reading. Despite having the *o2* mutation, BQPM15 has demonstrated a relatively high methionine concentration in conjunction with a relatively low lysine concentration, which is not typical of a QPM line but may be of interest nonetheless.

Availability

Seed for each BQPM line in the amount of 50 kernels per line is available through the National Plant Germplasm System (NPGS) or by contacting Dr. Paul Scott at paul.scott@ars.usda.gov. We ask that proper recognition be given in the event that any of this germplasm is used in the development of a new cultivar, hybrid, or breeding line.

Conclusions

Since overcoming the negative pleiotropic effects of the *o2* mutation with *o2* modifiers, QPM has become a desirable crop to grow in terms of nutritive quality but

continues to suffer from poorer yields relative to elite normal hybrids grown in the US Corn Belt. Temperate QPM lines BQPM9, BQPM10, BQPM11, BQPM12, BQPM13, BQPM14, BQPM15, BQPM16, and BQPM17 have higher lysine and tryptophan content than currently available inbred lines and a capacity for use in high-yielding hybrid maize production, making them good candidates for nutritive enhancement of feed corn.

Acknowledgments

The authors wish to thank Dr. Slobodan Trifunovic for his work at CIMMYT, and Paul White and Scott Johnson for their technical assistance. This project was funded in part by the USDA National Institute of Food and Agriculture OREI grant number IOWW-2010-02363. Mention of trade names or commercial products in this report is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture. The USDA is an equal opportunity provider and employer.

Tables

Table 1. Origin of each of the BQPM lines.

Line†	Pedigree‡	Recurrent Parent	Origin	Source of QPM§
BQPM9	(B99 x CLQ 06901) x B99	B99(Hallauer et al., 1995)	Iowa Corn Borer Synthetic No 1 (R) C10	CLQ06901
BQPM10	(B99 x CLRQ 00502) x B99	B99 (Hallauer et al., 1995)	Iowa Corn Borer Synthetic No 1 (R) C10	CLRQ00502
BQPM11	(B100 x CLQ 06901) x B100	B100 (Hallauer et al., 1995)	(B85 x H99)H99	CLQ06901
BQPM12	(CLQ 06901 x B98) x B98	B98 (Hallauer et al., 1994)	Pioneer two-ear Composite (FR) C5	CLQ06901
BQPM13	(CLQ 06901 x B97) x B97	B97 (Hallauer et al., 1994)	Iowa Corn Borer Synthetic No 1 (R) C9	CLQ06901
BQPM14	(CLQ 06901 x B97) x B97	B97 (Hallauer et al., 1994)	Iowa Corn Borer Synthetic No 1 (R) C9	CLQ06901
BQPM15	(B91 x CLQ 06901) x B91	B91 (Russell 1989)	Iowa Corn Borer Synthetic No 1 (R) C7	CLQ06901
BQPM16	(CLQ 06901 xB98) x B98	B98 (Hallauer et al., 1994)	Pioneer two-ear Composite (FR) C5	CLQ06901
BQPM17	(CLQ 06901 xB113) x B113	B113 (Hallauer et al., 2001)	Pioneer two-ear Composite (FR) C9	CLQ06901

† Varieties developed in the state of Iowa are given names starting with B by convention.

‡ Pedigrees were shortened for simplicity.

§ The two QPM donor lines have little genetic relationship to one another. CLRQ00502 comes from the subtropical population 502 and CLQ06901 was derived from Templado Amarillo QPM population 69. The latter has a flinty yellow kernel phenotype and intermediate maturity.

Table 2. Summary of agronomic traits for nine BQPM lines crossed to three testers at four (testers 1 and 3) and three (tester 2) locations. There were no significant differences between any of the BQPM lines for any of the agronomic traits at the 0.05 probability level.

Line	Grain				Lodging	
QPM	**Yield**	**Density**	**Moisture†**		**Stalk**	**Root**
	kg ha^{-1}	kg m^{-3}	———————		**%**	———————
BQPM9	6685.4	706.1	24.4		3.0	3.5
BQPM10‡	6564.1	716.2	24.0	A	1.7	3.2
BQPM11	6460.8	708.3	26.9		2.6	0.8
BQPM12	6447.6	697.1	27.3		0.6	2.5
BQPM13	6398.6	704.9	26.9		2.4	8.0
BQPM14	6107.1	698.0	27.3		0.8	0.1
BQPM15‡	6078.5	687.3	25.2		1.4	2.2
BQPM16	6041.4	685.1	28.5	B	0.1	1.7
BQPM17	6579.0	699.3	25.5		3.8	1.0
Mean§	6377.8	700.1	26.4		1.8	2.7
LSD (0.05)¶	782.3	55.2	4.3		8.6	7.9

† Means with different letters are significantly different at the 0.05 probability level

‡ BQPM10 and BQPM15 were only crossed to two testers rather than three.

§ This mean was calculated from the nine BQPM lines herein released and does not contain the nine commercial checks.

¶ 5% LSD calculated from the original 39 BQPM lines ($p < 0.05$).

Table 3. Summary of agronomic traits for 3 testers crossed to nine commercial check inbred lines and the three highest-yielding BQPM lines for each tester at four (testers 1 and 3) and three (tester 2) locations.

Line		Grain			Lodging	
Tester	**Check†**	**Yield**	**Density**	**Moisture**	**Stalk**	**Root**
		kg ha^{-1}	kg m^{-3}	———	**%**	———
T1	CL Alfa	6795.3	691.0	24.1	0.8	0.2
	CL Bravo	7243.3	678.2	24.8	1.0	4.2
	CL Charlie	6806.0	678.2	27.8	0.4	0.8
	CL Delta	7931.4	683.4	27.4	1.2	3.1
	BQPM11‡	**6981.6**	**708.3**	**26.9**	**2.6**	**0.8**
T2	CL Echo	7440.4	706.6	25.2	0.0	1.4
	CL Delta	7358.0	701.4	26.0	0.0	0.0
	CL Foxtrot	7031.8	680.8	26.7	0.6	1.4
	BQPM17‡	**6834.7**	**699.3**	**25.5**	**3.8**	**1.0**
T3	CL Golf	6895.6	707.9	22.8	1.0	0.0
	CL Hotel	7203.8	706.6	25.6	0.6	0.0
	CL India	7375.9	716.9	21.1	1.5	0.0
	BQPM9 ‡	**6634.0**	**706.1**	**24.4**	**3.0**	**3.5**
	CL Mean	7208.1	695.1	25.2	0.7	1.1
	BQPM Mean§	6377.8	700.1	26.4	1.8	2.7
	LSD (0.05) ¶	788.0	54.9	4.2	8.52	7.8
	Top BQPM mean#	6816.8	704.6	25.6	3.1	1.8
		———	———	**%**	———	———
	Checks vs. BQPM††	11.5*	NS	NS	NS	NS
	Checks vs. top BQPM	NS	NS	NS	NS	NS

† Commercial inbred lines used as checks (CL) are coded to protect confidential business information.

‡ Highest yielding of the 9 BQPM lines for the tester listed.

§ Calculated from Table 2.

¶ 5% LSD calculated from the original 39 BQPM lines by tester and the nine commercial checks by tester hybrid progeny ($p < 0.05$).

Calculated from the three highest-yielding BQPM lines listed above.

†† Percent difference between checks and BQPM

* Significant at the 0.05 probability level

NS Not significant at the 0.05 probability level

Table 4. Amino acid concentrations of the nine BQPM lines released.

Amino acid	BQPM9	BQPM10	BQPM11	BQPM12	BQPM13	BQPM14	BQPM15	BQPM16	BQPM17
Aspartic Acid	0.96†	1.56	0.97	0.92	0.74	0.83	0.68	1.12	1.07
Threonine	0.39	0.46	0.38	0.45	0.37	0.35	0.40	0.44	0.42
Serine	0.43	0.51	0.40	0.51	0.45	0.39	0.51	0.52	0.47
Glutamic Acid	1.74	2.24	1.63	2.40	1.84	1.51	2.25	2.41	1.94
Proline	0.94	1.18	0.93	1.24	0.94	0.79	1.13	1.15	1.03
Glycine	0.45	0.54	0.46	0.49	0.43	0.44	0.39	0.52	0.52
Alanine	0.64	0.79	0.58	0.85	0.69	0.55	0.86	0.80	0.69
Cysteine	0.27	0.34	0.25	0.30	0.24	0.24	0.25	0.29	0.31
Valine	0.52	0.63	0.53	0.59	0.51	0.48	0.53	0.59	0.56
Methionine	0.18	0.21	0.17	0.17	0.21	0.19	0.31	0.20	0.24
Isoleucine	0.33	0.39	0.32	0.39	0.33	0.29	0.40	0.39	0.37
Leucine	1.01	1.15	0.89	1.30	1.11	0.83	1.61	1.14	1.06
Tyrosine	0.30	0.35	0.27	0.40	0.30	0.26	0.39	0.35	0.30
Phenylalanine	0.43	0.49	0.40	0.53	0.45	0.37	0.58	0.50	0.46
Lysine	0.40	0.50	0.41	0.43	0.38	0.40	0.28	0.51	0.46
Histidine	0.39	0.46	0.39	0.44	0.36	0.36	0.34	0.43	0.41
Arginine	0.53	0.65	0.55	0.59	0.53	0.53	0.44	0.69	0.67
Tryptophan	0.07	0.09	0.06	0.08	0.06	0.07	0.06	0.09	0.09
Crude protein‡	11.27	13.93	10.77	13.02	10.97	9.93	11.55	13.61	12.46

† Values expressed as W/W% = g amino acid/100 g of sample

‡ Percentage N x 6.25

References

Atlin, G.N., N. Palacios, R. Babu, B. Das, S. Twumasi-Afriyie, D.K. Friesen, H. De Groote, B. Vivek, and K.V. Pixley. 2011. Quality protein maize: Progress and prospects. In: Janick, J., editor, Plant breeding reviews. Vol. 34. John Wiley & Sons, Hoboken, NJ. Chapt. 3.

Babu, R., S.K. Nair, A. Kumar, S. Venkatesh, J.C. Sekhar, N.N. Singh, G. Srinivasan, and H.S. Gupta. 2005. Two-generation marker-aided backcrossing for rapid conversion of normal maize lines to quality protein maize (QPM). Theor. Appl. Genet. 111:888–897. doi:10.1007/s00122-005-0011-6

Baker, D.H., D.E. Becker, H.W. Norton, A.H. Jensen, and B.G. Harmon. 1969. Lysine imbalance of corn protein in the growing pig. J. Anim. Sci. 28:23.

Bhatnagar, S., F.J. Betran, and L.W. Rooney. 2004. Combining abilities of quality protein maize inbreds. Crop Sci. 44:1997–2005. doi:10.2135/cropsci2004.1997

Beeson, W.M., R.A. Pickett, E.T. Mertz, G.L. Cromwell, and O.E. Nelson. 1966. Nutritional value of high-lysine corn. In: Proceedings of the Distillers Feed Research Conference, Cincinnati, OH. p. 70–77.

Bicar, E.H., W. Woodman-Clikeman, V. Sangtong, J.M. Peterson, S.S. Yang, M. Lee, and M.P. Scott. 2008. Transgenic maize endosperm containing a milk protein has improved amino acid balance. Transgenic Res. 17:59–71. doi:10.1007/s11248-007-9081-3

Bjarnason, M., and S.K. Vasal. 1992. Breeding of quality protein maize (QPM). Plant Breed. Rev. 9:181–216.

Choe, B.H., M.S. Zuber, G.F. Krause, and E.S. Hilderbrand. 1976. Inheritance of high lysine in maize. Crop Sci. 16:34–38. doi:10.2135/cropsci1976.0011183X001600010009x

Comstock, R.E., and H.F. Robinson. 1952. Estimation of the average dominance of genes. In: Gowen, J.W., editor, Heterosis. Iowa State College Press, Ames, IA. p. 494–518.

Cromwell, G.L., J.C. Rogler, W.R. Featherston, and R.A. Pickett. 1967. Nutritional value of *opaque-2* corn for the chick. Poult. Sci. 46:705–712. doi:10.3382/ps.0460705

Geetha, K.B., C.R. Lending, M.A. Lopes, J.C. Wallace, and B.A. Larkins. 1991. opaque-2 modifiers increase gamma-zein synthesis and alter its spatial distribution in maize endosperm. Plant Cell 3:1207–1219.

Gunaratna, N.S., H. De Groote, P. Nestel, K.V. Pixley, and G.P. McCabe. 2010. Food Policy 35:202–210. doi:10.1016/j.foodpol.2009.11.003

Habben, J.E., A.W. Kirleis, and B.A. Larkings. 1993. The origin of lysine-containing proteins in opaque-2 maize endosperm. Plant Mol. Biol. 23:825–838. doi:10.1007/BF00021537

Hallauer, A.R., K.R. Lamkey, W.A. Russell, and P.R. White. 1994. Registration of B97 and B98, two parental lines of maize. Crop Sci. 34:318–319. doi:10.2135/cropsci1994.0011183X003400010088x

Hallauer, A.R., K.R. Lamkey, W.A. Russell, and P.R. White. 1995. Registration of B99 and B100 inbred lines of maize. Crop Sci. 35:1714–1715. doi:10.2135/cropsci1995.0011183X003500060045x

Hallauer, A.R., K.R. Lamkey, W.A. Russell, and P.R. White. 1997. Registration of five inbred lines of maize: B102, B103, B104, B105, and B106. Crop Sci. 37:1405–1406. doi:10.2135/cropsci1997.0011183X003700040094x

Hallauer, A.R., K.R. Lamkey, and P.R. White. 1998. Registration of B107, B108, and B109 inbred lines of maize. Crop Sci. 38:1731. doi:10.2135/cropsci1998.0011183X003800060076x

Hallauer, A.R., K.R. Lamkey, and P.R. White. 2001. Registration of B110, B111, B113, and B114 inbred lines of maize. Crop Sci. 40:1518–1519.

W Horwitz. 2005a. AOAC official method 982.30 E(a,b,c). In: Official methods of analysis of AOAC International. 18th ed. AOAC International, Gaithersburg, Md.

W Horwitz. 2005b. Combustion analysis (LECO) AOAC Official Method 990.03 In: Official methods of analysis of AOAC International. 18th ed. AOAC International, Gaithersburg, Md.

Houmard, N.M., J.L. Mainville, C.P. Bonin, S. Huang, M.H. Luethy, and T.M. Malvar. 2007. High-lysine corn generated by endosperm-specific suppression of lysine catabolism using RNAi. Plant Biotechnol. J. 5:605–614. doi:10.1111/j.1467-7652.2007.00265.x

Huang, S., D.E. Kruger, A. Frizzi, R.I. D'Ordine, C.A. Florida, W.R. Adams, W.E. Brown, and M.H. Luethy. 2005. High-lysine corn produced by the combination of enhanced lysine biosynthesis and reduced zein accumulation. Plant Biotechnol. J. 3:555–569. doi:10.1111/j.1467-7652.2005.00146.x

Ignjatovi-Micic, D., K. Markovic, D. Ristic, S.M. Drinic, S. Stankovic, V. Lazic-Jancic, and M. Denic. 2009. Variability analysis of normal and opaque2 maize inbred lines. Genetika-Belgrade 41(1):81–93. doi:10.2298/GENSR0901081I

Krivanek, A.F., H. De Groote, N.S. Gunaratna, A.O. Diallo, and D. Friesen. 2007. Breeding and disseminating quality protein maize (QPM) for Africa. Afr. J. Biotechnol. 6:312–324.

Lai, J., and J. Messing. 2002. Increasing maize seed methionine by mRNA stability. Plant J. 30:395–402. doi:10.1046/j.1365-313X.2001.01285.x

Lewis, A.J., M.B. Barnes, D.A. Grosbach, and E.R. Peo. 1982. Sequence in which the amino acids of corn (*Zea mays*) become limiting for growing rats. J. Nutr. 112:782–788.

M.A López-Pereira. 1993. Economics of quality protein maize as a feed study. Agribusiness 9:557–568. doi:10.1002/1520-6297(199311)9:6<557::AID-AGR2720090603>3.0.CO;2–0

Mertz, E., L. Bates, and O.E. Nelson. 1964. Mutant gene that changes protein composition and increases lysine content of maize endosperm. Science 145:279–280. doi:10.1126/science.145.3629.279

E.T Mertz. 1992. Quality protein maize. American Association of Cereal Chemists, St. Paul, MN.

L Munck. 1992. The case of high-lysine barley breeding. In: Shewry, P., editor, Barley, genetics, biochemistry, molecular biology and biotechnology. CAB International, Wallingford, UK. p. 573–601.

Nelson, O.E., E.T. Mertz, and L.S. Bates. 1965. Second mutant gene affecting the amino acid pattern of maize endosperm proteins. Science 150:1469–1470. doi:10.1126/science.150.3702.1469

Ngaboyisonga, C., K. Njoroge, D. Kirubi, and S.M. Guthiri. 2009. Effects of low nitrogen and drought on genetic parameters of grain yield and endosperm hardness of quality protein maize. Asian J. Agric. Res. 3:1–10. doi:10.3923/ajar.2009.1.10

Nyanamba, T., H. De Groote, and R. Wahome. 2003. Quality protein maize for the feed industry in Kenya. Poster paper presented at the International Agriculture Economics Association, Durban, South Africa. 16–22 Aug.

Osborne, T.B., and L.B. Mendel. 1914. Amino acids in nutrition and growth. J. Biol. Chem. 17:325.

Paez, A.V., J.L. Helm, and M.S. Zuber. 1969. Lysine content of opaque-2 maize kernels having different phenotypes. Crop Sci. 9:251. doi:10.2135/cropsci1969.0011183X000900020045x

R.A Pickett. 1966. *Opaque-2* corn in swine nutrition. In: Proceedings of the High-Lysine Corn Conference. Corn Industries Research Foundation, Washington, DC. p. 19–22.

Prasanna, B.M., S.K. Vasal, B. Kassahun, and N.N. Singh. 2001. Quality protein maize. Curr. Sci. 81:1308–1319.

W.A Russell. 1989. Registration of B90 and B91 parental inbred lines of maize. Crop Sci. 29:1101–1102. doi:10.2135/cropsci1989.0011183X002900040079x

Scott, M.P., S. Bhatnager, and J. Betran. 2004. Tryptophan and methionine levels in quality protein maize breeding germplasm. Maydica 49:303–311.

Scott, M.P., A. Darrigues, T.S. Stahly, and K.R. Lamkey. 2008. Recurrent selection to control grain methionine content and improve nutritional value of maize. Crop Sci. 48:1705–1713. doi:10.2135/cropsci2008.01.0010

Scott, M.P., J.M. Peterson, and A.R. Hallauer. 2009. Evaluation of combining ability of quality protein maize derived from U.S. public inbred lines. Maydica 54:449–456.

Tang, M., X. He, Y. Luo, L. Ma, X. Tang, and K. Huang. 2013. Nutritional assessment of transgenic lysine-rich maize compared with conventional quality protein maize. J. Sci. Food Agric. 93:1049–1054. doi:10.1002/jsfa.5845

S.K Vasal. 2001. High quality protein corn. In: Hallauer, A.R., editor, Specialty corns. 2nd ed. CRC Press, Boca Raton, FL. p. 85–129.

Vasal, S.K., E. Vilegas, M. Bjarnason, B. Gelaq, and P. Goertz. 1980. Genetic modifiers and breeding strategies in developing hard endosperm opaque-2 materials. In: Pollmers, W.G., and Phillips, R.H., editors, Quality traits of maize for grain and silage use. Martinus Nijhoff, London. p. 37–73.

Zarkadas, C.G., R.I. Hamilton, Z.R. Yu, V.K. Choi, S. Khanizadeh, N.G. Rose, and P.L. Pattinson. 2000. Assessment of the protein quality of 15 new northern adapted cultivars of quality protein maize using amino acid analysis. J. Agric. Food Chem. 48:5351–5361. doi:10.1021/jf000374b

CHAPTER 3. CONIFER: UNRAVELING THE SEQUENCE READ ARCHIVE (SRA)

Article prepared for publication in PLoS One

Hannah M. Worral[4,5]

Abstract

As the Sequence Read Archive (SRA) continues to grow and data mining of that sequence data becomes more common, targeting the point of confluence between the repository and active research by developing tools that demonstrate interoperability between archival systems and analysis platforms becomes an increasingly important task for researchers. By improving the accessibility of the SRA for data analysis, the knowledge and understanding to be gained from those sequences expands, and we begin to build a better framework from which to investigate questions concerning the genetic underpinnings of organisms across the globe. Herein we describe a pseudo-graphical user interface (pseudo-GUI) for a suite of bioinformatics tools that facilitates the extraction of specific subsets of data from within the SRA and incorporation of archived sequence data into ongoing genetic and genomic studies. CONIFER (Converting Old to New: Individualized Fasta/q Extractor for Researchers) presents the scientific community with

---

[4] Primary author and researcher

[5] Department of Agronomy, Iowa State University, Ames, IA 50011

a user-friendly tool that facilitates the ancillary analysis of archived sequence data in isolation or in combination with newly-generated sequence data.

Introduction

As next-generation sequencing (NGS) technologies continue to improve and data sets become larger and more complex, the limiting factor becomes not the cost and availability of sequencing but the presence and accessibility of the computational infrastructure capable of storing these large data sets and analyzing them in an efficient manner (Metzker, 2010; Sboner, 2011; Muir, 2016; Oyler-McCance, 2016; Celesti, 2017). To address the growing masses of NGS data, the International Nucleotide Sequence Database Collaboration (INSDC) established the Sequence Read Archive (SRA) as a central repository from which users can submit and retrieve data (NCBI, 2010). This sequence repository currently houses nearly 9.7 petabases of data, up from 6.5 terabases in 2009, and will only continue to grow as sequencing costs decline and better computational tools for the analysis of that data become available (Davy, 2010; Poland, 2012; Edwards, 2013; Heffelfinger, 2014; NCBI, 2017). Thus, it is imperative that semantic interoperability between archival systems and analysis platforms is implemented in parallel. The finite number of barcodes available for use in genotyping-by-sequencing (GBS), for example, introduces ambiguity when combining data sets generated in different sequencing runs. In the event that these data sets were combined with the intention of bolstering sequencing coverage, preemptive action can be taken and barcodes assigned appropriately to the replicate samples (Boyles, 2016; Schlautman, 2017). Should the researcher desire to combine data sets where there is ambiguity in the

barcodes, an additional tool is required to sort out the sequence reads for specific samples.

As the SRA continues to grow and interest in retrieving archival data from the repository deepens due to a growing interest in data mining, the development of tools that maximize the overall efficacy of the system and broaden the utility of the sequence data that is generated and stored there becomes paramount (Kodama et al., 2012; Nakazato, 2013; Zhu, 2013; Samuels, 2013). While tools for the extraction, conversion, and analysis of archived sequence data do exist, to our knowledge, no one tool exists for the seamless incorporation of archived sequence files into ongoing studies together with newly-generated sequence data. Two tools that have the potential to serve such a function, however, are the SRA toolkit and the TASSEL pipeline. The NCBI has a very useful SRA toolkit available for download (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc), but it only operates on one file at a time. As SRA files are downloaded one-per-subdirectory with multiple subdirectories present within the directory for the study of interest (SOI), this mode of action hinders efficient, automated conversion of archived data to formats preferred for genetic and genomic analyses. The TASSEL pipeline offers a flexible framework for the analysis of sequence data, executed via either a graphical user interface (GUI) or the command line (Glaubitz, et al., 2014). The command line offers more options than the GUI and has the potential to be automated, but may not be accessible to researchers with limited experience working within UNIX environments. In applying the above tools, we encountered a marked lack of ease in the attainment, conversion, and analysis of archived data files from the SRA. Pursuant to these observations, a line of inquiry to address these

issues was opened and a pipeline contrived to attenuate the problems encountered in this process. Within the larger context of improving the accessibility of the SRA for downstream applications, the following objectives were identified:

1. Automatize conversion of SRA files to FASTQ format for data files downloaded from the SRA for use in downstream applications with a focus on subsequent analysis using the TASSEL platform, but not limited to such analyses
2. Create a user interface that improves the accessibility of TASSEL for researchers unfamiliar or uncomfortable with the command line who are looking for a more flexible alternative to the TASSEL GUI.

We also decided to design a pseudo-graphical user interface (pseudo-GUI) that would echo the user-friendly nature of a traditional GUI while maintaining the flexibility of the command line. Built-in options include:

1. Conversion of SRA to FASTQ format and/or use of newly-generated FASTQ sequences
2. The ability to set and save parameters for each step of the analysis
3. Species-independent calling of genotypes (single nucleotide polymorphisms (SNPs) called against any specified reference genome)
4. Extraction of individual samples from larger data sets
5. The ability to increase coverage for ‘replicate’ samples

6. Application of uniform SNP-calling parameters to raw data acquired from various studies

By treating the SRA as an athenaeum rather than as a depository, the data stored within can be used to explore other questions pertaining to the genome than originally intended. By improving the accessibility of the SRA for data analysis and developing new platforms to diversify the types of analyses currently available for this type of data, the knowledge and understanding to be gained from these archived sequences expands beyond the limits of a singular study. In order to help facilitate this expansion, we have created a suite of tools for the analysis of archived sequence data that is flexible enough to be applied to sequence data outside the archive as well. We have given it the name CONIFER (CONversion and Individualized Fasta/q Extraction for Researchers).

Design Considerations

Accessibility/Ease of Use

In the design of CONIFER, an important consideration was overall accessibility and ease of use for researchers. This suite of tools is meant to be run in a UNIX environment and can be set up with little difficulty. Whether using a super cluster available for a fee or employing the use of your own computer, the programs and languages contained within the framework of CONIFER are readily available. We recommend downloading a command line interface such as GitBASH (https://git-scm.com/downloads) for those whose computers do not contain a native command line. To make CONIFER more accessible to a wider array of users, we implemented computer languages that are relatively straightforward and easy to understand, and we designed a

pseudo-graphical user interface (pseudo-GUI) to facilitate ease of use with a particular emphasis on downstream analysis via the TASSEL pipeline (Glaubitz, et al., 2014).

Efficiency

While the SRA toolbox found on the NCBI contains scripts for the conversion of the standard storage format (SRA) to FASTQ format, we observe an inefficiency of action due to the folder structure of the SRA files when downloaded and the limitation of the conversion tool of only modifying one file in a folder at a time in its execution. In the event that a researcher has several or more sequence files they wish to download, this singular inefficiency can cost them a significant amount of time in preparing the downloaded files for use in downstream applications. The first module of the CONIFER pipeline seeks to alleviate this problem.

Module I facilitates the rapid conversion of SRA files to FASTQ format for use in various downstream applications with an emphasis on their use within the TASSEL pipeline. It renames FASTQ files for use within TASSEL by making use of the SRARunInfo Excel file that is available for download along with the sequence data for each study from the SRA.

Once the SRA files have been converted to FASTQ format, the pseudo-GUI continues to walk the user through the TASSEL pipeline within the command line, allowing the investigator access to more memory and speed through a remote server or computer cluster than is possible using the native “point-and-click” GUI for TASSEL, all without forfeiting the ease of use associated with the GUI. This interface features prompts for the selection of parameters and input/output files, a printout of specified

parameters and the option to save custom settings for future use. For those choosing to opt-out of the option to convert SRA files to FASTQ files due to already having FASTQ files at their disposal, the pseudo-GUI will skip over Module I in favor of offering sequence subsetting and genotype calling through Modules II and III. As such, the pseudo-GUI allows the user access to all of TASSEL's features – raw sequence read processing, alignment to specified reference genome(s), and generation of HapMap genotype calls – and some additional options for subsetting and combining data sets available through Module II.

Flexibility

The data subsetting tool – Module II – found within the pseudo-GUI extracts genotype calls for individual samples from previously-generated HapMap files and subsets samples from large data sets into FASTA/Q files containing only those sequence reads pertinent to the selected sample. Raw sequence data from identical samples sequenced across multiple studies can be concatenated to improve sequencing coverage for calling SNPs, and those same individualized sequence lists can be used to query unassembled genomes for pertinent genetic information. Whether or not the user opts to subset their data with Module II, running their data through the pseudo-GUI gives them the flexibility to call genotypes against any user-supplied reference genome which greatly expands the utility of the original data sets. SNP-calling parameters for processing via TASSEL can be set and saved to be applied to future data sets to maintain consistency and enable meaningful conclusions to be drawn when comparing genotype calls made from sequences obtained from different studies.

Implementing the Pipeline

CONIFER can be downloaded as a singular directory containing all of the subdirectories and tools needed to execute all of the features described below. Once downloaded and moved to a local or remote site with sufficient space, the user will need to download the SRA toolkit and TASSEL so that they can be executed within the pseudo-graphical user interface (pGUI) of the CONIFER suite.

The user will then have the choice of using three different modules – either independently or in concert with one another. Module I will walk the user through the conversion of archived sequence data to useable FASTQ format and provides an automation of the conversion process that helps to minimize active time spend on this step. Module II continues to process the data within the CONIFER pseudo-GUI and will prompt the user to subset their data prior to generating SNP calls. Should the user decide *not* to subset the data, they will be directed to Module III which runs the full data set through to completion and generates HapMap files for further analysis. Figure 1 gives a general overview of the suite of tools and how each module can be used in a research project.

Herein, we describe the basic functions of each tool within the CONIFER suite and their benefits to researchers. Upon release of CONIFER, a user manual documenting the specific usage and function of each module will be available for download at https://github.com/HannahWorral.

Module I

To help facilitate the acquisition of data files from the SRA, Module I was created to improve the efficiency of data capture and conversion for use in ancillary capacities – that is, research beyond the scope of the original sequencing project. This module focuses on packaging the data in a manner suited for downstream analysis with TASSEL, but has the flexibility to offer alternative outputs suitable for other applications. In order to implement Module I, the user must first download their study of interest (SOI) from the NCBI SRA. The user will need to navigate to ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/ then, depending on the study accession number, move into one of three directories – DRP, ERP, or SRP. In the case of the data set used as a proof of concept for this release, the address was ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP/SRP021/SRP021921/. Transfer the SOI to a local site, e.g., C:/Deskop, then transfer to a remote working directory (unless the local site has sufficient space to handle the large data files). Once the directory containing the SOI has been transferred to its final location, execute Module I through the CONIFER suite.

Module II

Should the user elect to subset the data set for the selected SOI to the level of individualized lists of sequences for specific samples, the pseudo-GUI will process the data through the first couple of steps of the TASSEL pipeline, including sequence filtering and quality control steps. The subsequently-generated merged tags-by-taxa (TBT) file is formatted in such a way as to allow the extraction of sequence reads for individual samples as it is a matrix containing the list of sequences (tags) down the left-hand column and the samples (taxa) across the top with counts for the number of times

each sequence read occurs in each sample within the body of the matrix. In order to assign sequence reads to their respective sample, Module II merges ‘replicate’ samples and replaces the total counts within the matrix with presence/absence counts (1 or 0). Sequence reads are then assigned to sample-specific FASTA files based on presence counts (1). These FASTA files can then be converted to FASTQ format by the assignment of dummy quality scores (based on parameters specified in prior steps of the pipeline) to individual sequence reads through the use of a fasta-to-fastq Perl script (https://code.google.com/archive/p/fasta-to-fastq/).

By calling up Module II, the user is able to extract genotype calls for single samples within a larger study from previously-generated HapMap files by specifying subsets of samples in the barcode key files for those studies. This gives major and minor allele calls based on the whole set of samples rather than on the individual samples as is the case when running raw sequence data from the same study through the TASSEL pseudo-GUI wherein the investigator specifies the subset of samples prior to calling the SNPs. Should there be a desire to combine several ‘replicate’ samples across multiple studies to increase coverage for calling SNPs, new barcode key files would be generated to contain the barcodes assigned to each of those samples across the studies. In the event that more than one set of ‘replicate’ samples will be concatenated across multiple studies, and the pseudo-GUI will run each extraction and concatenation separately to avoid potential confusion over the use of the same barcode for two ‘replicate’ sets across the studies of interest.

Module III

Whether the FASTQ data set was stored in the SRA or was recently generated by the sequencing facility, Module III allows the user to make direct comparisons between samples from various studies by applying uniform parameters for filtering, quality control, and SNP calling to raw sequences generated in independent sequencing runs through the TASSEL pipeline and by way of the CONIFER pseudo-GUI. This makes use of the barcode key file that accompanies each sequencing run to select only those samples of interest and apply uniform parameters to each raw sequence data set. By making use of the archived sequence data and the TASSEL pipeline, new genotype calls can be made against additional reference genomes as they are released without requiring the resequencing of the samples of interest. The pseudo-GUI further hastens this analysis by offering a user-friendly, automated environment in which to process the data.

Conclusion

Strengths

For anyone planning on running their sequence data through TASSEL or interested in extracting sequence data from the SRA, CONIFER offers a user-friendly platform from which to do so. This suite of tools expands the utility of archived sequence data, enables researchers to draw meaningful comparisons between data from different studies, allows sequence lists for specific samples within a larger study to be generated, and shows promise for the concatenation of like samples across studies for improved sequencing coverage.

Limitations

TASSEL 4.0 (what we started working with) is riddled with bugs and TASSEL 5.0 changed the format and structure of several intermediary files within the pipeline, thus necessitating the use of three versions of TASSEL (3.0, 4.0., and 5.0) in CONIFER. We hope to remedy this prior to release, with each step optimized for use with TASSEL 5.0. Additionally, BASH 4.0 or higher is required to execute Module I to convert SRA files to FASTQ format. At the moment, CONIFER is can only provide uniform parameters to the analysis of data sets from different studies and cannot generate new data sets comprised of subsamples from different studies to run a single analysis. Since there are a limited number of barcodes available for use in sequencing, overlap across studies is possible. Therefore, joining the sequence data from two studies prior to extraction of sequence lists for individual samples is inadvisable as a strict concatenation of like barcodes could result in sequence data from random samples being combined. Concatenation of like samples from different studies following the generation of individualized FASTA/Q files for specific samples is possible, but not currently included in the CONIFER suite of tools

Future Directions

We plan to implement CONIFER in the generation of SNP calls for the 79 popcorn lines found in the Romay et al., 2013 Ames Diversity Panel against the W22 reference genome to compare to genotype calls generated against the B73 reference genome, to extract the sequence reads specific to each of those popcorn lines to query them for sequences related to *Ga1-S*, and to concatenate sequence information for 13 of the 16 Iowa Stiff Stalk Synthetic (BSSS) progenitors across two sequencing runs. Each of

these actions will serve as a proof of concept for three analyses CONIFER is capable of performing, and we plan to publish these investigations in a separate journal article upon completion of phenotypic work to be carried out by the next generation of graduate students in the Scott lab.

Figures

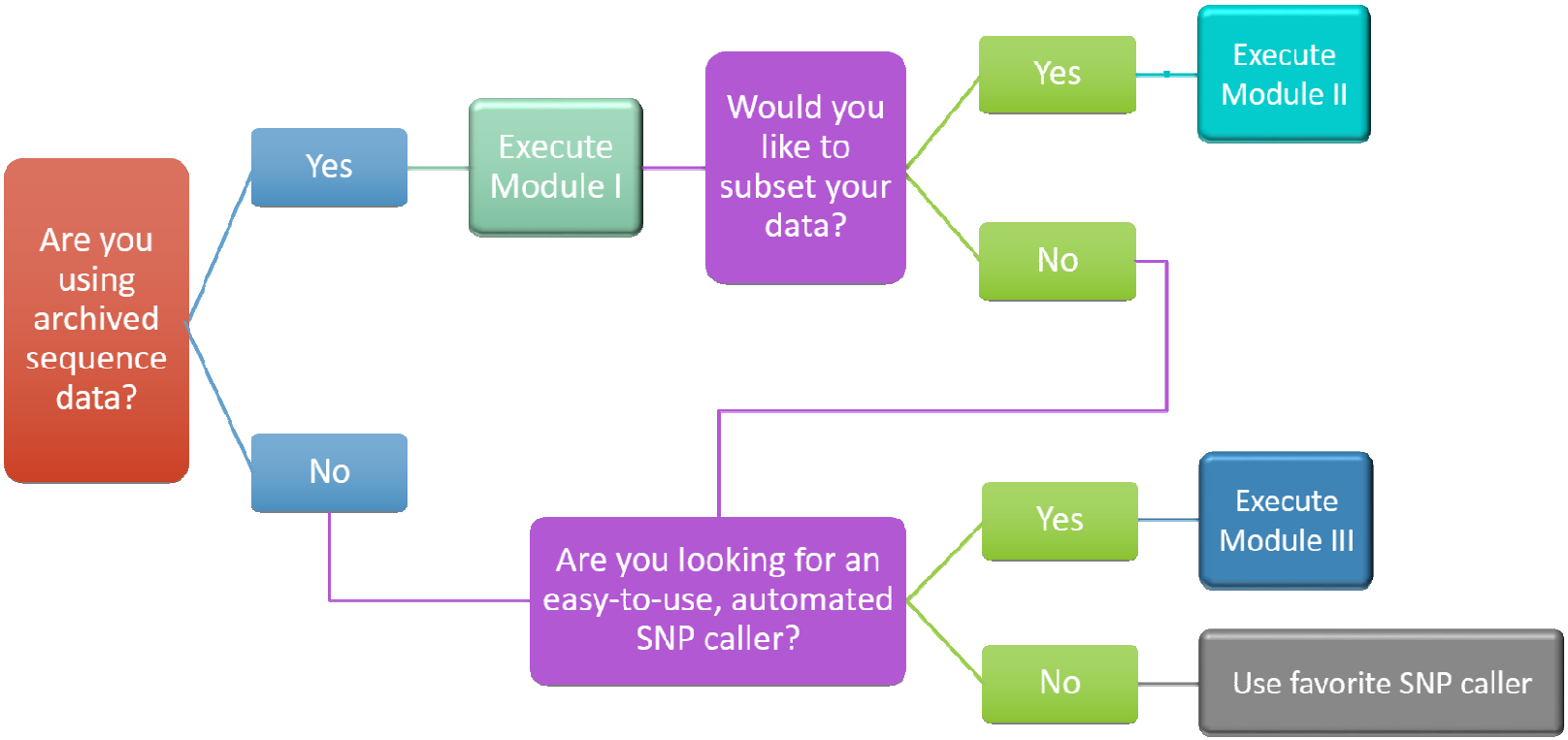

Figure 1. CONIFER decision tree

References

Boyles, R. E. (2016). Genome-Wide Association Studies of Grain Yield Components in Diverse Sorghum Germplasm. *Plant Genome, 9*(2). doi:doi:10.3835/plantgenome2015.09.0091

Celesti, A. C. (2017). Are next-generation sequencing tools ready for the cloud? *Trends in Biotech., 35*(6), 486-489.

Davy, J. D. (2010). RADSeq: next-generation population genetics. *Brief Funct. Genomics, 9*, 416-423.

Downloads. Git. Available: https://git-scm.com/downloads

Edwards, D. B. (2013). Accessing complex crop genomes with next-generation sequencing. *Theor. Appl. Genet.*, 1+11.

Elshire, R. G. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS). *PLoS ONE, 6*(5), e19379. doi:10.1371/journal.pone.0019379

Glaubitz, J., Casstevens, T., Lu, F., Harriman, J., Elshire, R., Sun, Q., & Buckler, E. (2014). TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. *PLOS ONE, 9*(2), e90346. doi:10.1371/journal.pone.0090346

Google Code Archive - Long-term storage for Google Code Project Hosting. Google. Google; Available: https://code.google.com/archive/p/fasta-to-fastq/

Heffelfinger, C. F. (2014). Flexible and scalable genotyping-by-sequencing strategies for population strategies. *BMC Genetics, 15*(979). doi:10.1186/1471-2164-15-979

Kodama, Y., Shumway, M., & Leinonen, R. (2012). The sequence read archive: explosive growth of sequencing data. *Nucl. Acids Res., 40*, D54056. doi:10.1093/nar/gkr854

Kumar, S. B. (2012). SNP discovery through next-generation sequencing and its applications. *Int. J. Plant Genome*. doi:10.1155/2012/831460

Metzker, M. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.*, 31-46. doi:10.1038/nrg2626

Muir, P. L. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology, 17*(53). doi:10.1186/s13059-016-0917-0

Nakazato, T. O. (2013). Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. *PLoS ONE, 8*(10), e77910. doi:10.1371/journal.pone.0077910

NCBI. (2010). *SRA Handbook.* Bethesda, MD: U.S. National Center for Biotechnology Information.

NCBI. (2017). *SRA Growth Infographics*. (U. N. Medicine, Producer) Retrieved May 19, 2017, from National Center for Biotechnology Information: https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/

Oyler-McCance, S. O. (2016). A field ornithologist's guide to genomics: Practical considerations for ecology and conservation. *The Auk, 133*(4), 626-648.

Poland, J. R. (2012). Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome J., 5*, 92-102.

Samuels, D. H. (2013). FInding the lost treasures in exome sequencing data. *Trends in Genetics, 29*(10), 593-599. doi:10.1016/j.tig.2013.07.006

Sboner, A. M. (2011). The real cost of sequencing: higher than you think! *Genome Biol., 12*(25). doi:10.1186/gb-2011-12-8-125

Schlautman, B., Covarrubias-Pazaran, G., Diaz-Garcia, L., Iorizzo, M., Polashock, J., Grygleski, E., Vorsa, Nocholi, & Zalapa, J. (2017). Construction of a High-Density American Cranberry (Vaccinium macrocarpon Ait.) Composite Map Using Genotyping-by-Sequencing for Multi-pedigree Linkage Mapping. *G3: Genes, Genomes, Genetics, 7*(4), 1177-1189. doi:https://doi.org/10.1534/g3.116.037556

Toolkit Documentation: Software: Sequence Read Archive: NCBI/NLM/NIH. National Center for Biotechnology Information. U.S. National Library of Medicine; Available: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc

Zhu, Y. S. (2013). SRAdb: query and use public next-generation sequencing data from within R. *BMC Bioinformatics, 14*, 19.