Section 6 (Texas Traditional) Report Review

Form emailed to FWS S6 coordinator (mm/dd/yyyy): 1/3/2017

TPWD signature date on report: 8/31/2016

Project Title: Conservation genetics and genomics, pollination biology and phenology of Cryptantha crassipes I. M. Johnst., Terlingua Creek cat's-eye"

Final or Interim Report? Final

Grant #: Grant No. TX E-160-R (F13AP00689)

Reviewer Station: Austin ESFO

Lead station concurs with the following comments: NA (reviewer from lead station)

Interim Report (check one):

- ☐ Acceptable (no comments)
- ☐ Needs revision prior to final report (see comments below)
- ☐ Incomplete (see comments below)

Final Report (check one):

- ☒ Acceptable (no comments)
- ☐ Needs revision (see comments below)
- ☐ Incomplete (see comments below)

Comments:

FINAL PERFORMANCE REPORT

As Required by

THE ENDANGERED SPECIES PROGRAM

TEXAS

Grant No. TX E-160-R

(F13AP00689)

Endangered and Threatened Species Conservation

Conservation genetics and genomics, pollination biology and phenology of Cryptantha crassipes I. M. Johnst., Terlingua Creek cat's-eye"

Prepared by:

Dr. James Cohen

Carter Smith
Executive Director

Clayton Wolf
Director, Wildlife

31 August 2016

FINAL REPORT

STATE: ____Texas_____ GRANT NUMBER: ___TX E-160-R-1__

GRANT TITLE: Conservation genetics and genomics, pollination biology and phenology of Cryptantha crassipes I. M. Johnst., Terlingua Creek cat's-eye".

REPORTING PERIOD: ____1 September 2013 to 31 August 2016_

OBJECTIVE(S). The proposed project will involve the investigation of the conservation genetics and genomics, pollination biology, and phenology of *Cryptantha crassipes* in order to understand the best manner in which to conserve and grow the species.

Segment Objectives:

Task 1: March 2014. Students and the Principal Investigator (PI) will travel to Brewster Co., TX to visit populations of *Cryptantha crassipes*. During this one-week trip, we will accomplish multiple objectives, including phenological observations, pollination studies, and the collection of leaves and flowers.

Task 2: April and May 2014. During one four-day trip in April and one week-long trip in May, students and the PI will revisit populations of *C. crassipes* to observe and record the phenology of individuals. Students and the PI will observe the plants following the same procedure as in Task 1.

Task 3: May 2014 and ongoing. The PI will design a website for the research efforts on *C. crassipes*. This website will include images of the plants at various stages of development, pollinators, and information on genetics and diversity. The PI will continuously update the website to reflect the status of the project.

Task 4: October 2014. Students and the PI will use a modified CTAB method (Doyle and Doyle, 1990) to extract DNA from leaf samples of the 400 individuals.

Task 5: August 2014. During a four-day trip, students and the PI will revisit populations of *C. crassipes* to observe and record the phenology of individuals. Should the plants be flowering, students and the PI will observe the plants following the same procedure as in Task 1.

Task 6. November 2014 – May 2015. After the 10 microsatellite loci are identified, students will use polymerase chain reaction (PCR) to amplify the loci for the 400 DNA isolations of *C. crassipes*.

Task 7. March – May 2015. During a one-week trip in March, a four-day trip in April, and a one-week trip in May, students and the PI will revisit populations of *C. crassipes* and observe the phenology of the individuals.

Task 8. June – November 2015. Genotypes of individuals will be scored with GeneMapper software (ABI), and genetic diversity and population structure will be studied subsequently.

Task 9. August 2015. During a four-day trip, students and the PI will revisit populations of *C. crassipes* to observe and record the phenology of individuals.

Task 10. December 2015 – February 2016. Students and the PI will interpret results and prepare manuscripts.

Significant Deviations:

None.

Summary Of Progress:

Please see Attachment A, and data residing at following websites:

- Raw SNP data - https://www.dropbox.com/sh/iaohy8hlyfmzena/AACas2UW_W4KI0lVdU9AJ2r6a?dl=0 Note: The files are quite big, around 140 GB total.
- Raw microsatellite data - https://www.dropbox.com/sh/hijndou3p2099x7/AADKpohHdo2qCwSIuXvZSgvJa?dl=0

Location: Fizzle Flat Lentil geologic formation, Brewster County, Texas, USA.

Cost: ___Costs were not available at time of this report, they will be available upon completion of the Final Report and conclusion of the project.__

Prepared by: _Craig Farquhar_____ Date: _31 August 2016

Approved by: _____ Date:_____31 August 2016_
C. Craig Farquhar

ATTACHMENT A

Conservation genetics and genomics, pollination biology, and phenology of *Cryptantha crassipes* I. M. Johnst., Terlingua Creek Cat's-eye

James Cohen
Kettering University, 1700 University Ave., Flint, MI 48504
810-249-4383
jcohen@kettering.edu

Abstract

Oreocarya crassipes (formerly *Cryptantha crassipes*) is an endangered plant species endemic to the area just north of Big Bend National Park. While the ecology of the plant has been well-studied, the breeding system, life history, and population genetics and genomics have not been examined. *Oreocarya crassipes* exhibits the breeding system heterostyly, which involves multiple floral morphs within a population. To better understand the breeding system of the species, floral-morph ratios, the extent of herkogamy, and controlled crosses were conducted. Results suggest that morph ratios are near one in the four studied populations and that even though style length was relatively continuous between the two morphs, two distinct anther heights were observed. Also, for self-, intra-, and intermorph crosses, either compatible pollen tube growth or seed set was seen, suggesting that multiple types of crosses are compatible. Single nucleotide polymorphism data from tunable genotyping-by-sequencing and microsatellite loci were amplified. Genetic diversity within and among four populations was studied, and two to three genetic populations were identified. The populations exchange migrants, with populations that are geographically closer to each other having greater rates of migration. The results provide appropriate information for conservation management of the species

Introduction

Oreocarya crassipes (I. M. Johnst.) Hasenstab & M. G. Simpson (formerly *Cryptantha crassipes* I. M. Johnst.) is a species in the plant family Boraginaceae endemic to the area just north of Big Bend National Park in West Texas. Indeed, the common name of the plant, Terlingua Creek Cat's-Eye, provides the small geographic region in which the species is centered. Ten known populations have been recognized (USFWS, 1993), all of which are on private property. While many property owners have, in general, been quite amenable to research and conservation measures on the plants (e.g., building fences around populations), the small number of populations and individuals, as well as possible disease (Warnock, 2012), make the long-term viability of the species questionable. Consequently, studies on the the breeding system, life history, genetic diversity, and past and current demographics were undertaken.

The species is restricted to a type of habitat known as moonscape due to the barren conditions of the ecosystem (Fig. 1A). While the soil of the moonscape habitat contains gypsum, Warnock (2012) identified even higher levels of gypsum in the vicinity of individuals of *O. crassipes* suggesting that the species is an obligate gypsophile, at least under natural conditions. Additionally, *O. crassipes* produces pyrrolizidine alkaloids, a secondary compound common in species of Boraginaceae. These pyrrolizidine alkaloids may confer herbivore and microbe resistance and adaptation to xeric habitats, allowing the plants to survive under the harsh conditions in the region (Warnock, 2012).

Along with the edaphic specialization in this species, *O. crassipes* exhibits heterostyly, a complex and elegant breeding system. Heterostyly is characterized by two or three floral morphs in a population. In the simplest case, distyly, which is the type of heterostyly in *O. crassipes*, two floral morphs are present. In one, the long-style (LS) morph, anthers are situated below the stigmas (Fig. 1B), and in the other, the short-style (SS) morph, anthers are positioned above the stigmas (Fig. 1C). The anthers of one morph are at the same height as the stigmas in the other morph, a condition known as reciprocal herkogamy (Cohen, 2010). Along with the morphological component of heterostyly, there also usually exists a self- and intramorph incompatibility mechanism, which results in only sexual organs at the same height producing offspring, although this is not always the case (e.g., Casper, 1985). Additionally, there often are micromorphological differences between morphs, such as in pollen size and epidermal cell lengths of the style and corolla. Heterostyly in *O. crassipes* has not been well-characterized, and variation in floral organ lengths and whether or not the species is self- and intramorph

Figure 1. Images of moonscape habitat (A) in which *Oreocarya crassipes* grows, and long-style (LS) morph (B) and short-style (SS) morph of species.

incompatible can have significant consequences for the viability of the species. For example, if only LS morph individuals are present in the population and the species is self- and intramorph incompatible, sexual reproduction will be challenging. Pollinators and floral visitors also remain understudied, although Warnock (2012) identified some floral visitors.

The relationship among the populations of the species is unknown, and the manner in which the populations diverged and exchange genetic material has not been studied. Because the species has a restricted geographic range and a small number of populations, determining genetic diversity within populations as well as the extent to which genetic material is being exchanged among populations can help to establish appropriate conservation measures for the species. Therefore, investigating this variation with multiple types of molecular markers from throughout the genome can result in a clear understanding of current and past demographic patterns of *O. crassipes*.

The presented study builds on the comprehensive ecological work of Warnock (2012) and others, with the goal of better understanding the life history, breeding system, and demographics of *O. crassipes*.

Objective

The proposed project will involve the investigation of the conservation genetics and genomics, pollination biology, and phenology of *Cryptantha crassipes* in order to understand the best manner in which to conserve and grow the species.

Location

Four populations of the Terlingua Creek Cat's-Eye were sampled. One from the Field Lab (29.546483° N and

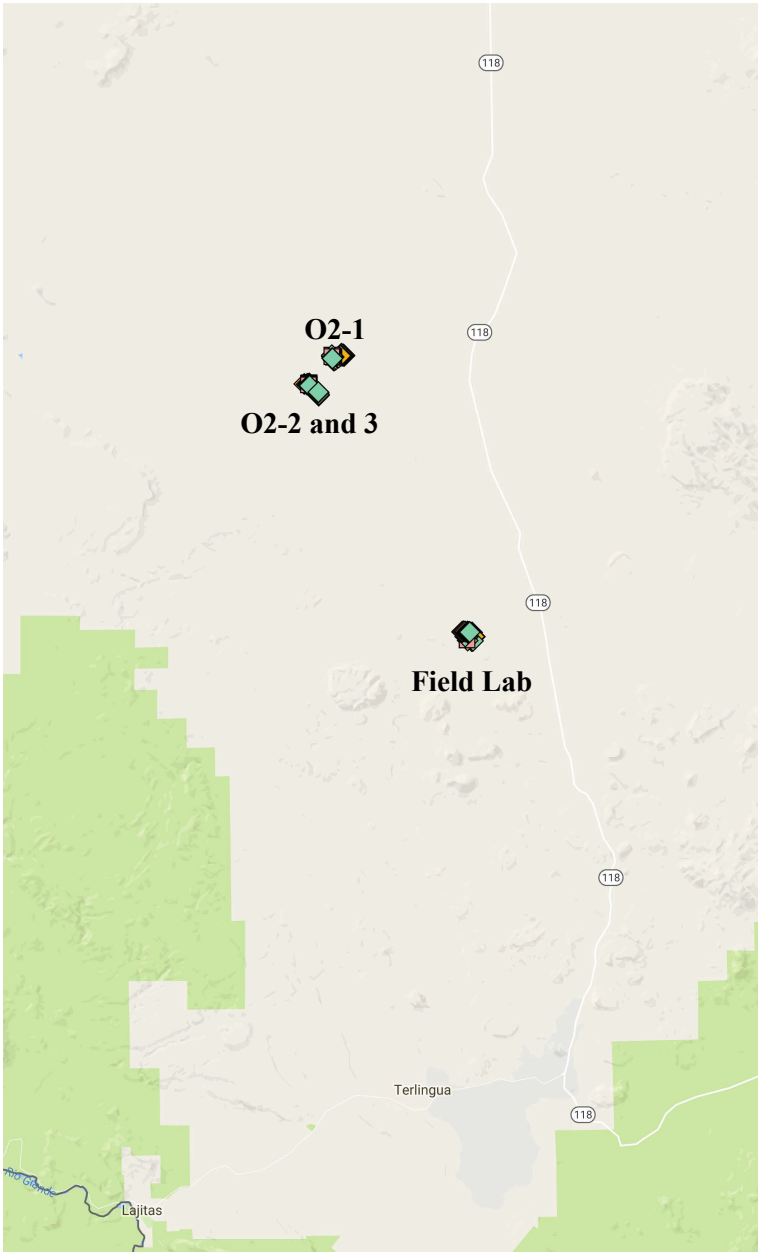


Figure 2. Map of four sampled populations of *Oreocarya crassipes*, shape and color denote morph, more detailed map available at http://tinyurl.com/hguq2cc)

103.587267° W), and three from the O2 Ranch (29.687367° N and 103.662133° W, 29.686933° N and 103.667817° W, and 29.668467° N and 103.675583° W). Figure 2 is a map showing the locations of the populations with morphs, and a more detailed map of the sampling locations can be found online at http://tinyurl.com/hguq2cc.

Methods

Task 1. March 2014. Students and the Principal Investigator (PI) will travel to Brewster Co., TX to visit populations of *Cryptantha crassipes*. During this one-week trip, we will accomplish multiple objectives, including phenological observations, pollination studies, and the collection of leaves and flowers.

For phenological observations, we will identify the timing of flowering and fruiting. This will involve observing, documenting, and imaging species and tagging them via a global positioning system (GPS). This digital tagging will make it possible to revisit the same individual to observe its phenology throughout the season and in subsequent years.

Three types of breeding system studies will be conducted during this initial trip. One will be to observe pollinators visiting the flowers of *C. crassipes*. Students and the PI will observe the plants of *C. crassipes* for one-hour periods in the early morning, late morning, early afternoon, late afternoon, and evening. During these times, we will take images of and collect floral visitors. These collections will later be identified in the research laboratory of the PI at Texas A&M International University (TAMIU). The second type of study will involve the determination of legitimate and illegitimate crosses within *C. crassipes*. Using mature flowers, we will perform 10 replicates each of intraindividual, intramorph, and intermorph controlled crosses via manual pollinations. The flowers of the short-style morph will be emasculated prior to anther dehiscence, but this is not necessary for the flowers of the long-style morph due to the position of the stigma above the anthers. After each manual pollination, flowers will be bagged and observed to determine if fruits are produced. The gynoecium of flowers that do not produce fruit will later be investigated with aniline blue staining, following the protocol of Ruzin (1999), to determine the site of incompatibility. The third type of breeding system study will involve identifying the morph of at least 50 randomly chosen plants in each population in order to determine the ratio of LS to SS plants in the populations and species.

Students and the PI will collect leaves and flowers, which will be used for subsequent study, at TAMIU, on the conservation genetics and genomics and micromorphology of *C. crassipes*. Leaves from 400 individuals across the 10 populations (USFWS, 1993; Warnock, 2012) will be collected and placed in bags with silica gel. DNA for each individual will be isolated at TAMIU. Mature flowers from 10 individuals of each morph will be collected and stored in FAA (Ruzin, 1999) for subsequent observation with scanning electron microscopy to identify micromorphological differences between the two morphs.

Deviation - The field trip was shortened to four days, but during this time, plant material was collected and breeding system studies were conducted.

Task 2. April and May 2014. During one four-day trip in April and one week-long trip in May, students and the PI will revisit populations of *C. crassipes* to observe and record the phenology

of individuals. Students and the PI will observe the plants following the same procedure as in Task 1.

Task 3. May 2014 and ongoing. The PI will design a website for the research efforts on *C. crassipes*. This website will include images of the plants at various stages of development, pollinators, and information on genetics and diversity. The PI will continuously update the website to reflect the status of the project.

Task 4. May – October 2014. Students and the PI will use a modified CTAB method (Doyle and Doyle, 1990) to extract DNA from leaf samples of the 400 individuals. DNA will be sent to the Savannah River Ecology Laboratory (http://www.srel.edu/microsat/Microsat_DNA_Development.html) for microsatellite design, and at least 48 microsatellites loci will be identified. Students will screen these microsatellite loci to identify 10 to amplify for the project. The 10 microsatellite loci will be determined by two criteria: 1) ability to consistently amplify, and 2) demonstrated intraspecific variation.

During this time, students and the PI will use scanning electron microscopy to investigate micromorphological differences between the long-style and short-style morphs of *C. crassipes*. This includes pollen size, epidermal cell length, stigma papillae, and other features. Additionally, pollinators of *C. crassipes* will be identified with the use of keys and, when necessary, the consultation of experts.

Deviation - It was only possible to collect leaf material from 244 samples across four populations. The Savannah River Ecology Laboratory was not involved in microsatellite design as it was possible to utilize microsatellites developed for related species in the genus *Oreocarya* Greene (Bresowar and McGlaughlin, 2014). Consequently, funding for this part of the project was moved to other types of sequencing endeavors. Light microscopy has been used instead of scanning electron microscopy to gain an understanding of the micromorphology and development of the two morphs of *O. crassipes*.

Task 5. August 2014. During a four-day trip, students and the PI will revisit populations of *C. crassipes* to observe and record the phenology of individuals. Should the plants be flowering, students and the PI will observe the plants following the same procedure as in Task 1.

Deviation - This field trip did not take place because the PI had moved institutions.

Task 6. November 2014 – May 2015. After the 10 microsatellite loci are identified, students will use polymerase chain reaction (PCR) to amplify the loci for the 400 DNA isolations of *C. crassipes*. After amplification, diluted PCR product will be mixed with 0.2 µL of GeneScan 500 LIZ size standard (Applied Biosystems [ABI]), followed by the addition of Hi-Di Formamide (ABI) to a final volume of 15 µL. This mixture will be run on an Applied Biosystems (ABI) 3500 DNA analyzer, at TAMIU, to identify the length of each amplified DNA region.

In addition to the amplification of microsatellite loci, the PI will sequence the genome of 10 individuals, each from a separate population (USFWS, 1993). This in-depth genome

sequencing will be conducted with 100 base-pair paired-end reads over two lanes run on an Illumina HiSeq 2000. This approach will generate over 15 gigabases of sequence data per individual.

Deviation - Microsatellite loci were analyzed in May of 2016 instead of during the latter portion of 2014 and the early part of 2015. Leaf material from 184 individuals of *O. crassipes* and primer sequences were sent to Eurofins STA Laboratories, and 10 loci were amplified prior to analysis by the PI. Additionally, leaf material for 192 individuals of *O. crassipes* was sent to data2bio for tunable genotyping-by-sequencing (tGBS), a method that involves the identification of thousands of single nucleotide polymorphisms (SNPs). This approach was utilized rather than the originally proposed whole genome sequencing of 10 samples in order to better understand the population genomics of the sampled individuals because large quantities of the genome were surveyed for almost 20 times as many individuals as originally proposed. After sequencing, data2bio identified SNPs and constructed datasets that included various quantities of SNPs and missing data. The PI was provided with raw and aligned sequences and SNP data.

Task 7. March – May 2015. During a one-week trip in March, a four-day trip in April, and a one-week trip in May, students and the PI will revisit populations of *C. crassipes* and observe the phenology of the individuals. Students and the PI will observe the plants following the same procedure as in Task 1. The phenology and pollinators during this early season will be compared to that of the previous season. Additionally, manual pollinations experiments will be conducted again, using the same procedure described in Task 1.

Task 8. June – November 2015. Genotypes of individuals will be scored with GeneMapper software (ABI), and genetic diversity and population structure will be studied subsequently. Students and the PI will use HP- Rare 1.1 (Kalinowski, 2005) and Arlequin 3.5 (Excoffier, Laval, and Schneider, 2005) to determine gene and allelic diversity, observed heterozygosity, and departures from expected Hardy-Weinberg equilibrium within *C. crassipes*. InStruct (Gao, Williamson, and Bustamante, 2007) will be utilized to identify the population structure of the species. Using sequence data generated from the Illumina HiSeq 2000 run, the PI will identify single nucleotide polymorphisms (SNPs), small variants in the DNA of different individuals. In order to do so, sequence data from all individuals will be pooled and, using Trinity (Grabherr et al., 2011), assembled into a reference genome. With Bowtie (Langmead et al., 2009), a short-read alignment software, sequence data from each individual will be aligned to the reference genome. After alignment, the PI will use FreeBayes (Garrison and Marth, 2012) to identify SNPs, which will be compared among the 10 individuals.

This tiered approach to the conservation genetics and genomics of *C. crassipes* will provide information on genetic variation at two different levels. Broad sampling will allow for an understanding of genetic variation among many individuals, while deep sequencing will provide evidence of genomic variation within the species. By using both of these approaches, it will be possible to compare microsatellite and SNP heterozygosity (cf., Väli et al., 2008) in order to estimate total genomic variation among individuals of the populations of *C. crassipes*.

Deviation - While InStruct was employed to investigate the population genetic and genomic data, other methods of data analysis were also utilized, including Arlequin (Excoffier, Laval, and Schneider, 2005), Genodive (Meirmans and Van Tienderen, 2004), GenePop (Raymond and Rousset, 1995), and Hierfstat (Goudet, 2005) to identify basic genetic diversity statistics within and among populations, Bayescan (Foll, 2012) to identify loci under selection, Bottleneck (Piry, Luikart, and Cornuet, 1999) to test for loci that are the result of recent genetic bottlenecks, SNPrelate (Zheng et al., 2012), fastSTRUCTURE (Raj, Stephens, and Pritchard, 2014), STRUCTURE (Pritchard, Stephens, and Donnelly, 2000), and Geneland (Guillot, Mortier, and Estoup, 2005) to examine population structure, and Migrate-N (Beerli and Palczewski, 2010), LAMARC (Kuhner, 2006), and IMa2 (Hey and Nielsen, 2007) to use coalescent methodology to investigate the demographic history of sampled populations. For the submitted final report, the results of *F* statistics from Hierfstat (Goudet, 2005), outlier loci identified via Bayescan (Foll, 2012), potential previous bottleneck events from Bottleneck (Piry, Luikart, and Cornuet, 1999), population structure analyses from STRUCTURE (Pritchard, Stephens, and Donnelly, 2000), fastSTRUCTURE (Raj, Stephens, and Pritchard, 2014) and SNPrelate (Zheng et al., 2012), and demographic history from Migrate-N (Beerli and Palczewski, 2010) and IMa2 (Hey and Nielsen, 2007) are presented. Please see deviation above concerning microsatellite and tGBS sequence data.

Task 9. August 2015. During a four-day trip, students and the PI will revisit populations of *C. crassipes* to observe and record the phenology of individuals. Should the plans be flowering, students and the PI will observe the plants following the same procedure as in Task 1.

Deviation - This field trip was not taken as it was not deemed necessary given a field trip in June 2015.

Task 10. December 2015 – February 2016. Students and the PI will interpret results and prepare manuscripts.

Results and Discussion

Tasks 1, 2, 5, 6, and 9. During the four field trips, plants of *O. crassipes* were observed and material was collected from one hundred and six plants at the Field Lab and 121 plants were collected from three populations from the O2 Ranch. Plants were observed in flower in mid-March and late April, and pollinators were collected at the Field Lab location. Bees, wasps, and flies were found to visit the plants, but it could not be definitively determined if these insects were pollinators or just floral visitors.

At the Field Lab, pollination studies were conducted, with the following crosses LS pollen X SS stigma, LS intramorph, LS self, SS pollen X LS stigma, SS intramorph, and SS self. While these pollination studies were abbreviated due to time constraints from the PI (due to his relocation to Flint, MI), preliminary data collected, based on these crosses, from pollen tube staining and seed set suggest that *O. crassipes* is compatible not only between morphs but also within morphs and within an individual plant and flower. Not only does pollen tube staining provide evidence of self- and intramorph compatibility (Fig. 3) but also preliminary intramorph

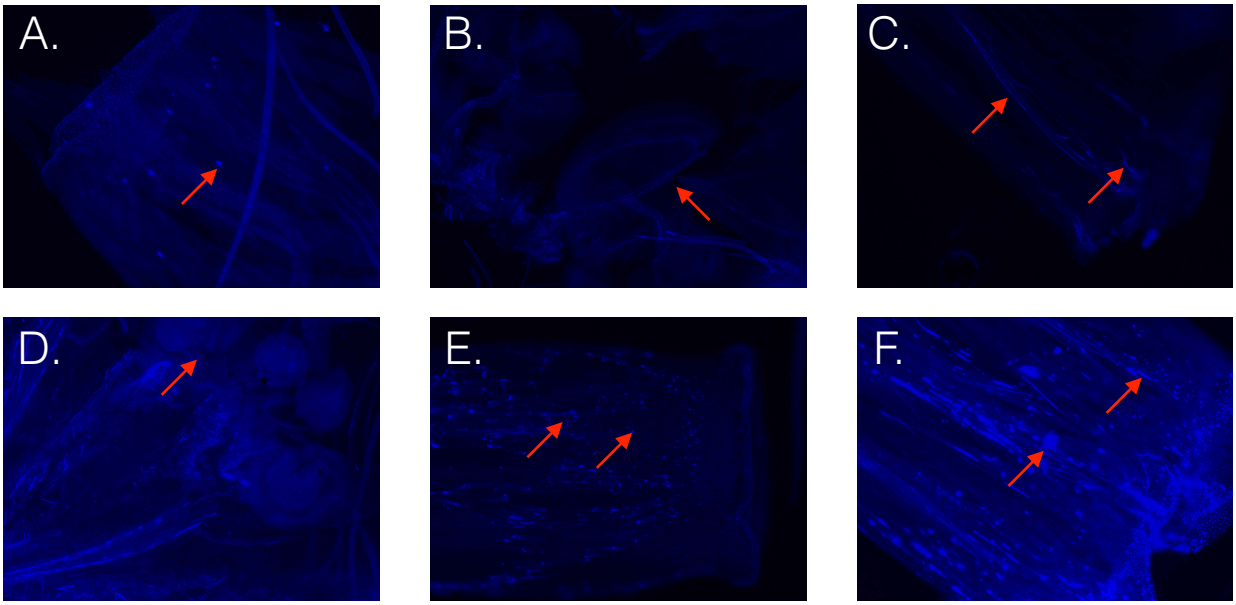


Figure 3. Images of pollen tube staining for various crosses of *Oreocarya crassipes*. A. is long-style (LS) self (100X), B. is LS intramorph (40X) showing pollen tube growing into ovule, C. LS stigma X short-style (SS) pollen (100X), D. SS self displaying pollen tube growing into ovule (40X), E. SS intramorph (100X), and F. SS stigma X LS pollen (100X), arrows point to some of the developing pollen tubes.

crosses resulted in seed set, demonstrating that these crosses can produce seed, which was collected in May and June. During this time, individuals of *O. crassipes* were still growing, but had finished flowering.

The results of these crosses are consistent those from *Oreocarya flava* A. Nelson (Casper, 1985), a species that also demonstrates self- and intramorph incompatiblity. This compatibility is uncommon for heterostylous species in Boraginaceae and throughout the angiosperms; however, as noted by Casper (1985) and is seen in *O. crassipes*, the LS:SS morph ratios are close to one (if not one [Table 1]), suggesting that other factors apart from pollen tube incompatibility may influence viability of self- and intramorph crosses.

The ratio of the two floral morphs was determined for each population, from arbitrarily collected flowers, as well as from the total number of collected flowers (Table 1). While morph ratios varied among the four populations, the long-style to short-style morph ratio was surprisingly one (1) for the four combined populations. Additionally, separation of anther and stigma heights from collected flowers provides evidence that the species is distylous. While stigma height is continuous, anthers are at two distinct heights, depending on morph (Fig. 4), which is similar to

Population	Long-style morph	Short-style morph	LS:SS ratio
Field Lab	66	56	1.18
O2 - 1	28	31	0.90
O2 - 2	12	20	0.60
O2 - 3	10	9	1.11
Total	116	116	1.00

Table 1. Morph counts and ratios of four sampled populations of *O. crassipes*.

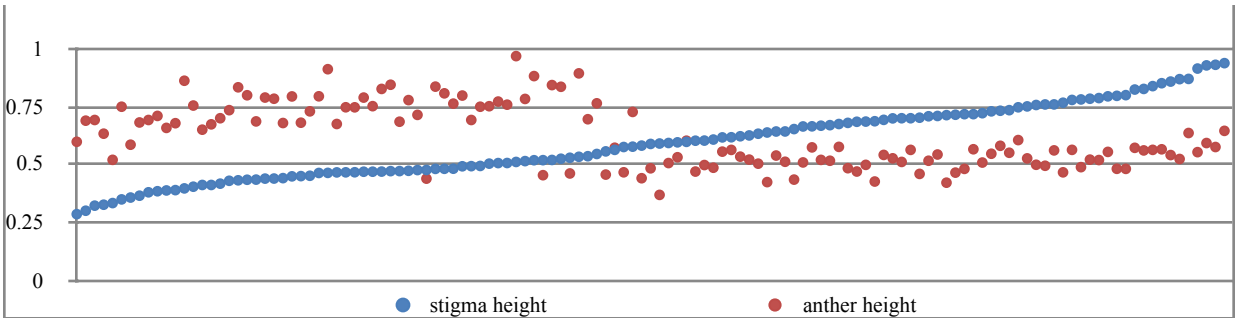


Figure 4. Stigma and anther height (in mm) of sampled flowers of *O. crassipes*, stigma height in blue and anther height in red.

other heterostylous species (e.g., Nishihiro et al., 2000). Anther-stigma separation is also variable, but there are two distinct groupings of the LS and SS morphs based on this separation (Fig. 5), providing additional evidence that the species is distylous.

The Wilcoxon/Kruskal-Wallis test was undertaken in JMP v12.1 (SAS Institute, 2009) to investigate differences, between morphs and populations, in anther height, stigma height, anther-stigma separation, corolla length, corolla tube length, corolla width, and corolla tube width. Stigma height, anther height, stigma-anther separation, and corolla tube width significantly differed ($Z < 0.0001$) between the long-style (LS) and short-style (SS) morphs of *O. crassipes*. The SS morph had a wider corolla tube compared to that of the LS morph. Corolla length, corolla tube length, and corolla width did not significantly differ between morphs. The flowers of the SS morph differed in anther height, stigma height, corolla length, corolla width, and corolla tube width among various combinations of the three O2 Ranch populations and the Field Lab population. In general, the greatest differences were between O2-2 and the Field Lab and O2-3 and the Field Lab. The flowers of the LS morph only differed in stigma height, for O2-2 and the Field Lab and O2-3 and the Field Lab, and in corolla tube width for O2-1 and the Field Lab. Given the closer geographic proximity of the three O2 Ranch populations to each other compared to any to the Field Lab population (Fig. 2), it is unsurprising that the flowers of the

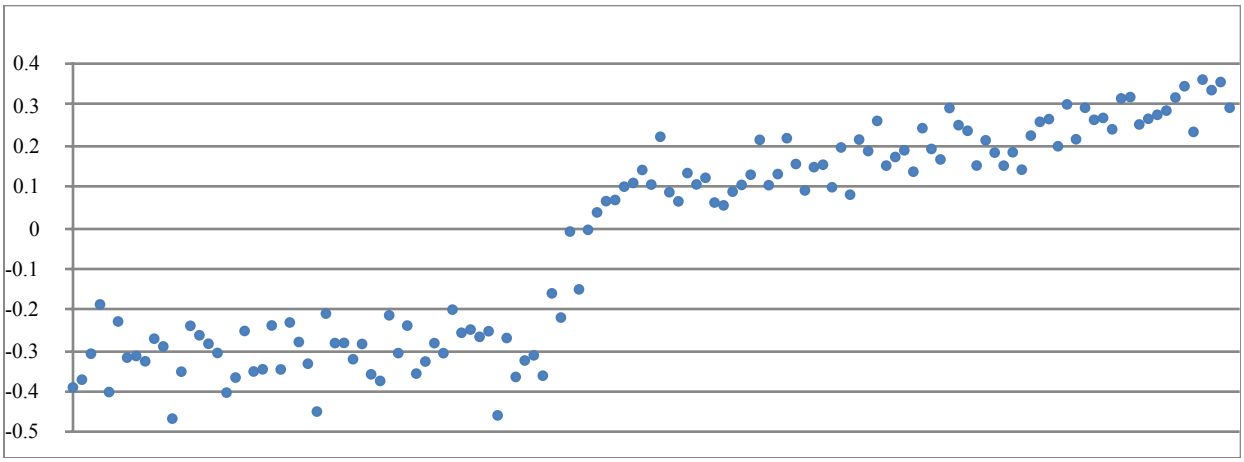
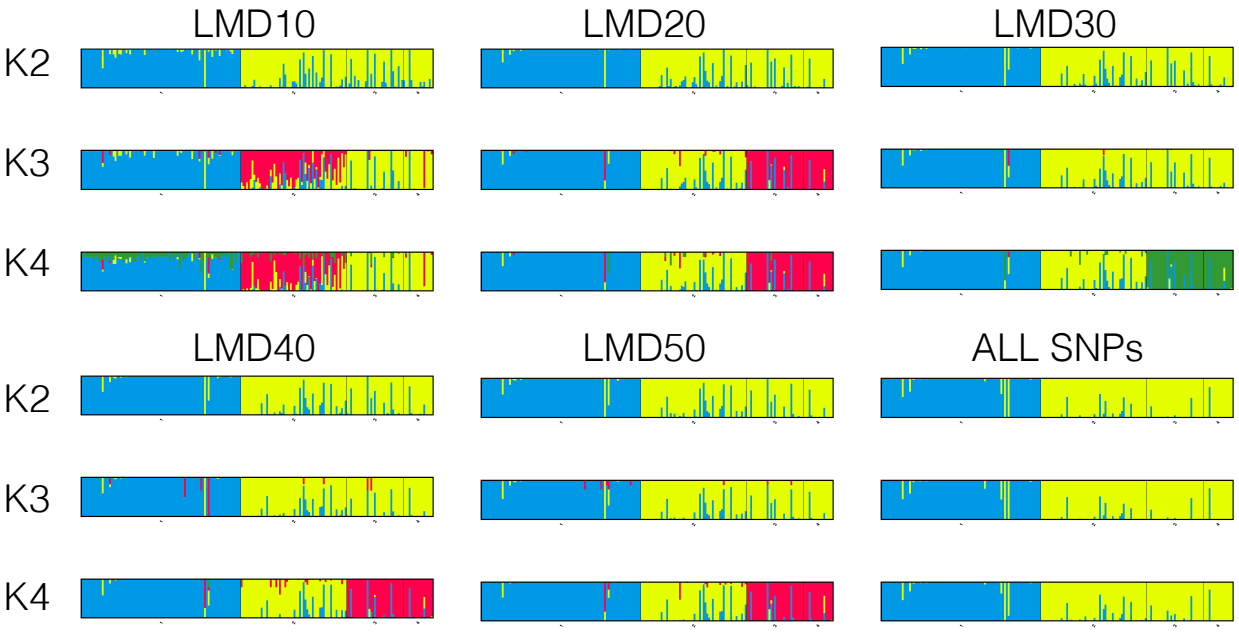


Figure 5. Anther-stigma separation (in mm) of short-style and long-style morphs, negative values denote anthers above stigmas, and positive values show anthers below stigmas, with zero being anthers and stigmas at same height.



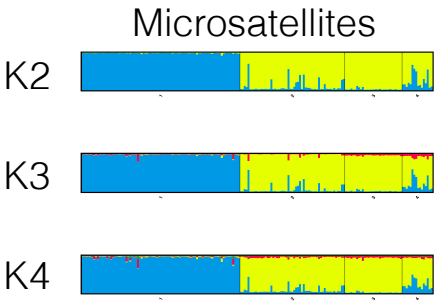
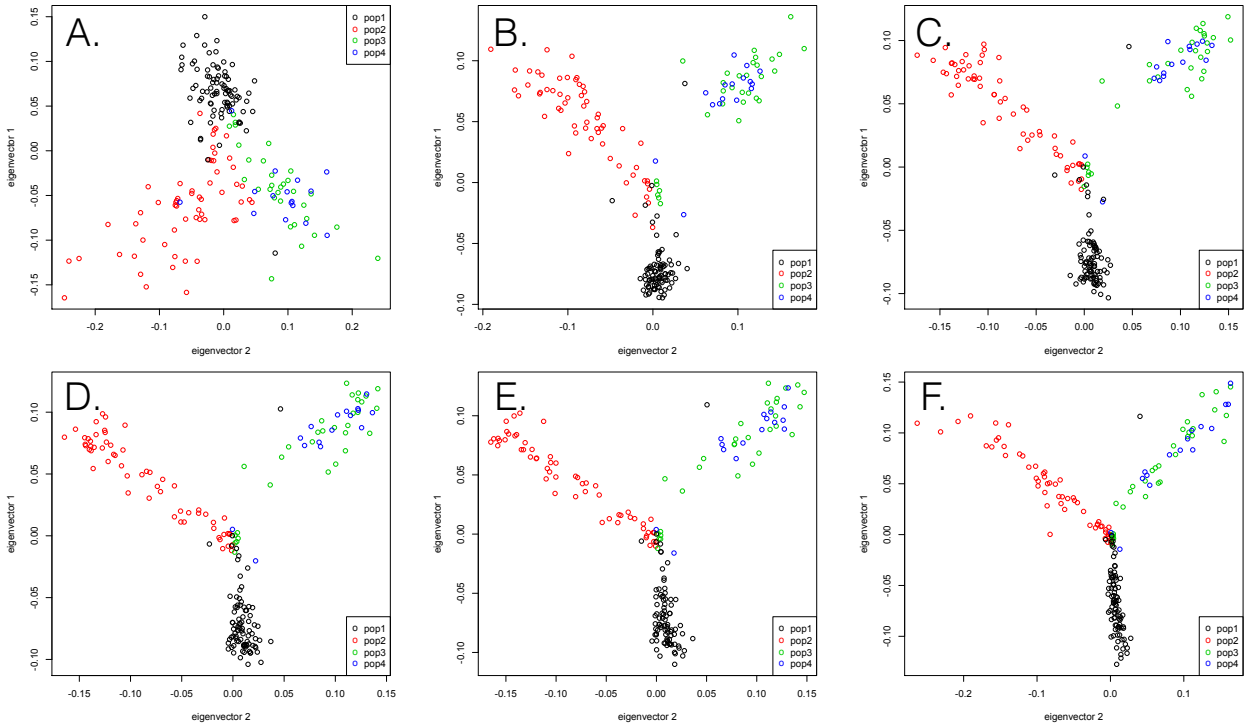


Figure 6. Bar graphs from fastSTRUCTURE and STRUCTURE analyses of various SNP and microsatellite datasets with two to four potential clusters (K) identified for four populations of *Oreocarya crassipes*.

Figure 7. PCA graphs from SNPrelate for four populations of *Oreocarya crassipes*, Field Lab (Pop1) in black, O2-1 (Pop2) in red, O2-2 (Pop3) in green, and O2-3 (Pop4) in blue. A is LMD10, B. is LMD20, C. is LMD30, D. is LMD40, E. is LMD50, and F. is all SNPs.



three O2 Ranch populations are more similar to each other than they are to the Field Lab population. This knowledge of stigma and anther height and stigma-anther separation of the flowers of the two morphs can help better understand pollinators and pollen flow within this endangered species.

Task 3. The website on *O. crassipes* has been constructed and is constantly being updated based on new information. The website can be viewed at http://www.cohen.science/oreocarya-crassipes-1.

Tasks 4, 6, and 8.

Population structure - SNP data from tGBS resulted in six datasets that range from 238 SNPs with 10% missing data to 61,487 SNPs with no limit on the quantity of missing data (Table 2). All of these datasets were used for analyses of population structure and outlier loci, but only the LMD10 dataset (and microsatellite data) was utilized for investigating patterns of population demographics, which was primarily due to limitations on computational abilities of the program and/or resources available to the PI.

SNP data, in general, with the exception of all SNPs suggest that there is genetic structure for two or three populations, depending on the potential number of clusters (K). While the optimal number of clusters was identified as two by STRUCTURE Harvester (Earl, 2012) and fastSTRUCTURE (Raj, Stephens, and Pritchard, 2014), the SNP data provide evidence that the Field Lab population is distinct from those distributed across the O2 Ranch as well as differentiation between O2-1 and O2-2+3 (Fig. 6). These population groupings are echoed by the SNPrelate data that demonstrate that with increasing SNPs (along with increasing missing data) the population divisions become more evident (Fig. 7).

There are two notable aspects of the analyses of SNPrelate data. One being that while the three population groupings are differentiated with the LMD10 dataset (Fig. 7A), the separation among individuals does not become much greater after 5,545 SNPs from the LMD30 dataset (Fig. 7C). Additionally, despite the three groupings of the four populations, there are some individuals that are at the nexus of the three identified groups, and this can also be seen in the fastSTRUCTURE and STRUCTURE plots which show that some individuals in one population are resolved as more genetically similar to those from other populations (e.g., in Fig. 6 LMD10, individuals in the Field Lab population [1] that are yellow instead of blue, are more similar to individuals in O2 populations). The microsatellite data show similar results to those from the SNP data, providing further evidence of population structure (Fig. 6); although, the SNP data appear to be able to more finely distinguish population structure (i.e., identify of a third grouping composed of individuals from the O2-2 and O2-3 populations).

Fst values between various pairs of the O2 populations are less than those from between the Field Lab

Table 2. Datasets from tGBS SNP data used in population genetic analyses

Dataset	Number of SNPs	Percentage missing data
LMD10	238	10%
LMD20	1,888	20%
LMD30	5,545	30%
LMD40	10,321	40%
LMD50	17,143	50%
All SNPs	61,487	0%

LMD10	Field lab	O2-1	O2-2	O2-3
Field lab	-	0.054	0.067	0.070
O2-1	0.052	-	0.046	0.049
O2-2	0.066	0.048	-	0.029
O2-3	0.069	0.052	0.029	-
LMD20	Field lab	O2-1	O2-2	O2-3
Field lab	-	0.077	0.101	0.103
O2-1	0.076	-	0.062	0.065
O2-2	0.097	0.062	-	0.030
O2-3	0.101	0.067	0.031	-
LMD30	Field lab	O2-1	O2-2	O2-3
Field lab	-	0.080	0.101	0.100
O2-1	0.079	-	0.056	0.058
O2-2	0.100	0.056	-	0.029
O2-3	0.100	0.059	0.029	-
LMD40	Field lab	O2-1	O2-2	O2-3
Field lab	-	0.079	0.101	0.103
O2-1	0.078	-	0.057	0.061
O2-2	0.101	0.057	-	0.029
O2-3	0.104	0.062	0.029	-
LMD50	Field lab	O2-1	O2-2	O2-3
Field lab	-	0.078	0.099	0.097
O2-1	0.077	-	0.056	0.059
O2-2	0.099	0.056	-	0.029
O2-3	0.098	0.060	0.029	-
ALL SNPs	Field lab	O2-1	O2-2	O2-3
Field lab	-	0.068	0.084	0.080
O2-1	0.067	-	0.051	0.050
O2-2	0.085	0.052	-	0.023
O2-3	0.081	0.051	0.023	-
Microsatellites	Field lab	O2-1	O2-2	O2-3
Field lab	-	0.125	0.152	0.082
O2-1	0.124	-	0.007	0.028
O2-2	0.151	0.007	-	0.018
O2-3	0.080	0.028	0.017	-

Table 3. *Fst* values for sampled populations based on various SNP and microsatellite datasets, Weir and Cockerham's *Fst* above diagonal, and Nei's *Fst* below diagonal

and any of the O2 Ranch populations regardless of dataset (Table 3). The Field Lab population appears to be more differentiated from the O2 populations than any of them is to the other O2 populations, and this is especially the case for O2-2 and O2-3, which are geographically close, group together in the fastSTRUCTURE analyses, and show low *Fst* values between them.

Collectively, these data suggest moderate genetic differentiation between the Field Lab and O2 populations, and this is the case regardless of whether SNPs or microsatellites are used. Indeed, with microsatellites there are greater *Fst* values than with SNPs (Table 3). These data provide evidence that should individuals or populations be chosen for *in* or *ex situ* conservation efforts, it would be prudent to select individuals from, at minimum, the Field Lab and O2 populations, and, if possible, from the O2-1 and O2-2+3 populations.

The Field Lab has greater *Fis* values suggesting that there is more inbreeding within that population compared to within the other three, which is consistent with geographic distances among the four sampled populations (Table 4). Indeed, it would be expected that the O2 populations would be more likely to breed among themselves than with the Field Lab population. However, it should be noted that some individuals in the Field Lab population are more similar, genetically, to those from the O2 Ranch populations (Figs 6 and 7). Despite the geographic distance between the Field Lab and O2 populations (Fig. 2), migration occurs between and among the various populations (see Demographics below).

The *Fst* and *Fis* values are in line with those from other rare plant species, such as *Penstemon albomarginatus* M. E. Jones (Wolfe et al., 2016), even those that have a greater geographic range, which is interesting given the small number of populations of *O. crassipes* and the relatively small number of populations sampled.

Outlier Loci - While outlier loci (those under natural selection) were identified, this was a small number of

Dataset and Population	Ho	Hs	Ht	*Fis*
LMD10				
Total	0.090	0.089	0.093	-0.008
Field lab	0.078	0.081	0.081	0.045
O2-1	0.100	0.103	0.103	0.021
O2-2	0.091	0.085	0.085	-0.072
O2-3	0.090	0.087	0.087	-0.032
O2-(2+3)	0.090	0.087	0.087	-0.041
O2-(1+2+3)	0.095	0.095	0.097	-0.007
LMD20				
Total	0.130	0.135	0.143	0.039
Field lab	0.115	0.126	0.126	0.083
O2-1	0.133	0.141	0.141	0.055
O2-2	0.138	0.141	0.141	0.020
O2-3	0.133	0.134	0.134	0.003
O2-(2+3)	0.136	0.140	0.140	0.028
O2-(1+2+3)	0.135	0.141	0.145	0.041
LMD30				
Total	0.137	0.148	0.156	0.070
Field lab	0.126	0.144	0.144	0.125
O2-1	0.138	0.151	0.151	0.088
O2-2	0.143	0.148	0.148	0.034
O2-3	0.142	0.146	0.146	0.032
O2-(2+3)	0.143	0.149	0.149	0.046
O2-(1+2+3)	0.140	0.150	0.154	0.067
LMD40				
Total	0.136	0.148	0.157	0.081
Field lab	0.126	0.148	0.148	0.149
O2-1	0.135	0.151	0.151	0.105
O2-2	0.143	0.148	0.148	0.030
O2-3	0.140	0.146	0.146	0.036
O2-(2+3)	0.142	0.149	0.149	0.046
O2-(1+2+3)	0.139	0.150	0.154	0.075
LMD50				
Total	0.135	0.147	0.156	0.087
Field lab	0.125	0.148	0.148	0.153
O2-1	0.133	0.150	0.150	0.115
O2-2	0.141	0.146	0.146	0.036
O2-3	0.140	0.146	0.146	0.041
O2-(2+3)	0.140	0.148	0.148	0.050
O2-(1+2+3)	0.136	0.149	0.153	0.083
All SNPs				
Total	0.125	0.143	0.150	0.121
Field lab	0.114	0.144	0.144	0.212
O2-1	0.123	0.145	0.145	0.154
O2-2	0.132	0.138	0.138	0.046
O2-3	0.132	0.141	0.141	0.064
O2-(2+3)	0.132	0.138	0.138	0.046
O2-(1+2+3)	0.127	0.143	0.147	0.111
Microsatellites				
Total	0.357	0.387	0.409	0.078
Field lab	0.332	0.377	0.377	0.121
O2-1	0.339	0.380	0.380	0.106
O2-2	0.371	0.383	0.383	0.032
O2-3	0.385	0.408	0.408	0.055
O2-(2+3)	0.375	0.394	0.394	0.049
O2-(1+2+3)	0.332	0.377	0.377	0.121

Table 4. Ho, Hs, Ht, and *Fis* values, based on various SNP and microsatellite datasets, for each population and groupings of populations

the total number of loci and corresponding SNPs investigated. Unfortunately, the function of these loci is unknown; however, identifying the loci and SNPs sets the stage for further examination of these markers as the loci may function in adaptation to the particular environments or ecological conditions of the different populations (e.g., Kramer and Havens, 2009). Additionally, these loci can be examined among species of *Oreocarya* to determine if there are differences that might result in particular adaptation to the moonscape habitat of West Texas, or if these loci might help confer adaptation to xeric or gypsum environments.

Current and historical demographics - Investigations of the demographic history of the four sampled populations of *O. crassipes* provide evidence of patterns of migration between populations as well as changes in the size of the populations. The results of the multiple Migrate-N analyses suggest that the SNP data best fit a model of unidirectional gene flow from the O2 Ranch populations to the Field Lab, rather than one of bi- or multidirectional gene flow between or among the four populations or from the Field Lab to the O2 populations. Analyses with Migrate-N based on microsatellite data resolve the opposite pattern of migration, with a model of unidirectional gene flow from the Field Lab to the O2 Ranch populations being favored. Using microsatellite data with IMa2, results provide evidence that while migrants are occurring across all populations, the greatest rates of migration are occurring in a modified stepping-stone model, with migrants moving from the most northerly population, O2-1, to O2-2 and from O2-2 to O2-3 and the Field Lab and from O2-3 to the Field Lab (Fig. 8), with the Field Lab being the most southerly population (Fig. 2). Results suggest that the migrants move back north from the Field Lab population to the O2-3 population at approximately an equal rate as migration from O2-3

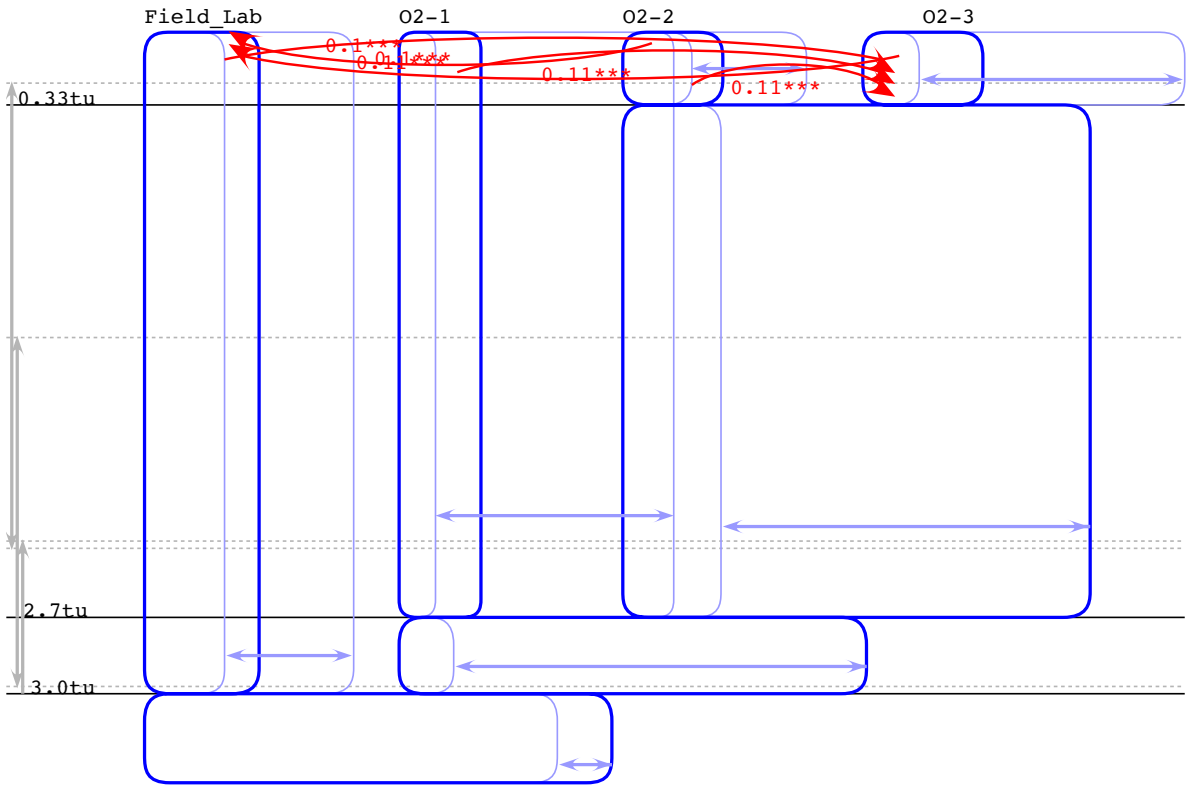


Figure 8. Demographic history of four sampled populations of *Oreocarya crassipes* populations over time. Population names at top, time units displayed in black on the side, population size in blue, with 95% confidence intervals in light blue, red arrows and numbers showing direction and 2NM values between populations. All displayed 2NM values statistically significant (p <0.001).

to the Field Lab (Fig. 8). As the results from IMa2 support a modified unidirectional pattern of gene flow in the sampled populations of *O. crassipes*, these results are not incongruent with those from SNP data (LMD10 dataset) or microsatellite analyzed in Migrate-N and nicely mesh the two conflicting results. The results of Migrate-N also provide evidence that migrants are moving at approximately equal rates among populations. Given the results of population structure and lack of homogenous genetic populations, it is unsurprising to identify migration among the populations; however, whether the migrants among populations are due to pollen flow or fruit dispersal remains unknown.

Population	model	2.50%	97.50%	95% confidence interval range
Field Lab	unidirectional	0.03973	0.0462	0.00647
	multidirectional	0.05207	0.05793	0.00586
O2-1	unidirectional	0.05313	0.05653	0.0034
	multidirectional	0.06287	0.06667	0.0038
O2-2	unidirectional	0.048	0.053	0.005
	multidirectional	0.03353	0.03747	0.00394
O2-3	unidirectional	0.04867	0.05207	0.0034
	multidirectional	0.04127	0.04487	0.0036

Table 5. Values of 95% confidence interval of Θ (theta), as determined by Migrate-N, for four sampled populations, using LMD10 dataset. Unidirectional model is from O2 populations to Field Lab populations.

The size of the four studied populations also appears to be relatively stable over time. The range of Θ (theta) for the four populations is relatively small in Migrate-N, regardless of model used for the SNP data, with 95% intervals

ranging from 0.0034 to 0.00647 (Table 5). These results are consistent with those from IMa2, although the 95% confidence intervals (represented by faint blue lines and boxes in Fig. 8) appear to be larger than those from Migrate-N analyses. The results of the patterns of migration as well as relatively stable population sizes provide evidence that while the species is rare, populations are not isolated nor are they shrinking in size. It should be noted, however, that these results do not take into account the most recent challenges facing the populations, such as off-road vehicles, disease, and the potential for development.

The program Bottleneck (Piry, Luikart, and Cornuet, 1999) was employed to test for recent bottlenecks in each of the four populations. Interestingly, different results were obtained depending upon the use of SNP or microsatellite data, which is similar to the use of these two types of data in Migrate-N. Using SNP data, each of the O2 Ranch populations was determined to have undergone a bottleneck, but the Field Lab population was not; however, the opposite results were found using the microsatellite data. Given the small number of microsatellites employed (nine polymorphic ones) compared to the 238 SNPs, the results from the SNP data seem to be more appropriate. Information on genetic bottlenecks is based on investigating patterns of heterozygote excess or deficiency. In all four studied populations, fewer loci have heterozygosity excess than expected, with many loci being heterozygote deficient. This could be due to inbreeding within the populations or low genetic diversity in general; however, as already noted, the *Fst* and *Fis* values are similar to those from other rare species and migration has recently occurred, so the identified bottlenecks do not appears to be an issue for the viability of the species. While the genetic bottleneck results provide additional data to suggest that the populations are distinct from each other, at the same time, the results should be treat with caution given those from other types of analyses as well as conflicting data from the two different types of molecular markers.

Task 10. Currently, the data is being interpreted, and manuscripts are being prepared. Two posters were presented, on the breeding system and conservation genetics of *O. crassipes*, at the Botany 2016 meeting in Savannah, GA.

Conclusion on the Conservation biology of O. crassipes

Collectively, the project on *O. crassipes* has involved the investigation of the morphology, breeding system, and conservation genetics and genomics of the species. The evidence suggests that, while heterostylous, this species is able to produce offspring via inter-morph crosses, which is the usual case of heterostylous species, as well as through self- and intramorph crosses. This latter situation is present in other heterostylous species of *Oreocarya*, such as *O. flava* (Casper, 1985), so it is not surprising that *O. crassipes* also is able to develop fruits through similar types of crosses. Additionally, while the style length appears to be continuous between the two morphs, there are two distinct anther heights and similar anther-stigma separation between the morphs, suggesting that heterostyly is functioning in effective pollen transfer between flowers of the morphs. Both morphs seem to have efficient male and female functions, but this could be better determined with more data on pollination among flowers, particularly with data related to pollen placement on pollinators. Because the species is able to produce offspring via all three types of crosses, the morph ratio (LS:SS) in populations

would not seem to be quite as important as it would if offspring could not be produced from particular crosses; however, given the morph ratios in the four sampled populations being close to one (and collectively being one [Table 1]), other forces may be impacting the number of LS and SS individuals, as has been suggested by Casper (1985). These forces are not currently known in the studied species.

Along with the data on the breeding system of *O. crassipes*, the population genetic and genomic analyses provides helpful information that can be utilized for conservation purposes. Utilizing SNP and microsatellite data, patterns of genetic diversity among the four populations were identified. While individuals from four populations were sampled, the genetic data provide evidence of two to three genetic populations, the Field Lab and O2 Ranch populations, at minimum. The O2 Ranch populations can be divided into the O2-1 and O2-2+3 populations, which were recognized based on SNP data, with microsatellite data only identifying the two populations (Fig. 6). The genetic data allow for the recognition not only of evolutionarily distinct groups but also of potential areas that can be used to conserve the species. Based on the present study, conservation efforts, either *in* or *ex situ*, should focus on individuals from the Field Lab and O2 Ranch, and if additional resources are available, trying to focus on individuals from the O2-1 population separate from O2-2+3 populations. This would ensure that a the greatest amount of genetic and genomic diversity is conserved, which is important for the preservation of the species. These efforts began years ago, and Warnock (2012) helped to fence off populations on the O2 Ranch in order to protect individuals *O. crassipes.* Unfortunately, it was not possible sample from other populations of the species, which would have provided further data to be able to better understand the population genetics and genomics of the species.

The demographic history analyses help inform potential conservation efforts of *O. crassipes*. Using these methods, it seems that the populations would be able to persist, and have persisted effectively, for generations. While population sizes have fluctuated over time, these sizes have remained relatively stable, particularly in recent history. Migration has also been occurring among populations, which is important for reducing inbreeding in this rare species. However, the efficiency of pollinators and the manner in which nutlets (fruits) are dispersed among populations and throughout the geographic region is understudied. This would be an interesting area of research in order to determine if pollinators or nutlet dispersal was having a greater influence on gene flow in the species, and if one or the other was more important for the persistence of the populations and species.

Because of the projects on *O. crassipes*, much more information is known regarding the population biology, phenology, geographic range, edaphic requirements, and breeding system of this rare species, but the long-term persistence of the population still remains questionable (Warnock, 2012). While in the field, a fungus appeared to be negatively impacting individuals in O2 Ranch populations, and off-road vehicles are frequently observed in areas around the plants. It is hoped that the presented project and others will provide sufficient information to allow for appropriate management of the species and the moonscape habitat.

Acknowledgements

I would like to thank the Texas Parks and Wildlife Department administering a grant (Section 6 TX E-160-R; TPWD #458186) awarded by U.S. Fish & Wildlife Service for the conducted research. John Wells and the O2 Ranch allowed access to their property to study the plants. Dr. Bonnie Warnock accompanied me and my students to the O2 Ranch. Diana Garcia and Jung-Mi Choi were unparalleled assistants in the field. Peter Beerli helped with the Migrate-N analyses, which were performed on the CIPRES portal. fastSTRUCTURE analyses were performed on the CyVerse atmosphere environment, and Veronica Moorman helped me set up these analyses. data2bio and Eurofins amplified the tGBS loci and microsatellites, respectively. Holly Hutcheson has conducted great work on the development of the two heterostylous morphs of *O. crassipes*. Anne Frey and Jodi Dorr were instrumental in the administrative aspects of the project. Dr. Craig Farquhar was helpful providing guidance throughout the project.

Literature Cited

Beerli, P., and M. Palczewski. 2010. Unified framework to evaluate panmixia and migration direction among multiple sampling locations. *Genetics* 185: 313-326.

Bresowar, G. E., and M. E. McGlaughlin. 2014. Characterization of microsatellite markers isolated from members of *Oreocarya* (Boraginaceae). *Conservation Genetics Resources* 6: 205-207.

Casper, B. B. 1985. Self-Compatibility in distylous *Cryptantha flava* (Boraginaceae). *New Phytologist* 99: 149-154.

Cohen, J. I. 2010. "A case to which no parallel exists": The influence of Darwin's *Different Forms of Flowers*. *American Journal of Botany*. 97: 701-716.

Doyle, J. J., and J. L. Doyle. 1990. A rapid total DNA preparation procedure for fresh plant tissue. *Focus* 12:13-15.

Earl, D. A. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* 4: 359-361.

Excoffier, L., G. Laval, and S. Schneider. 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary bioinformatics online* 1: 47.

Foll, M. 2012. BayeScan v2. 1 user manual. *Ecology* 20: 1450-1462.

Gao, H., S. Williamson, AND C. D. Bustamante. 2007. A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics* 176: 1635-1651.

Garrison, E., and G. Marth. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.

Goudet, J. 2005. Hierfstat, a package for R to compute and test hierarchical F−statistics. *Molecular Ecology Notes* 5: 184-186.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, et al. 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology* 29: 644.

Guillot, G., F. Mortier, and A. Estoup. 2005. GENELAND: a computer package for landscape genetics. *Molecular Ecology Notes* 5: 712-715.

Hey, J., and R. Nielsen. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences* 104: 2785-2790.

Kalinowski, S. T. 2005. hp−rare 1.0: a computer program for performing rarefaction on measures of allelic richness. *Molecular Ecology Notes* 5: 187-189.

Kramer, A. T., and K. Havens. 2009. Plant conservation genetics in a changing world. *Trends in plant science* 14: 599-607.

Kuhner, M. K. 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* 22: 768-770.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* 10: R25.

Meirmans, P. G., and P. H. Van Tienderen. 2004. GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes* 4: 792-794.

Nishihiro, , J., I. Washitani, J. D. Thomson, and B. A. Thomson. 2000. Patterns and consequences of stigma height variation in a natural population of a distylous plant, *Primula sieboldii*. *Functional Ecology*. 14: 502-512.

Piry, S., G. Luikart, and J.-M. Cornuet. 1999. BOTTLENECK: a program for detecting recent effective population size reductions from allele data frequencies. *Journal of heredity* 90: 502-503.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945-959.

Raj, A., M. Stephens, and J. K. Pritchard. 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* 197: 573-589.

Raymond, M., and F. Rousset. 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of heredity* 86: 248-249.

Ruzin, S. E. 1999. Plant microtechnique and microscopy. Oxford University Press New York.

SAS Institute. 2009. JMP 8 statistics and graphics guide, 2nd ed. Cary: SAS Publishing.

U.S. Fish and Wildlife Service. 1993. Draft Terlingua Creek Cat's-eye (*Crvptantha crassipes*) Recovery Plan. U.S. Fish and Wildlife Service, Austin, TX.

Väli, Ü., A. Einarsson, L. Waits, and H. Ellegren. 2008. To what extent do microsatellite markers reflect genome−wide genetic diversity in natural populations? *Molecular Ecology* 17: 3808-3817.

Warnock, B. 2012. Population biology, habitat description and delineation and conservation of Terlingua Creek Cat's-eye (*Cryptantha crassipes*). Texas Parks and Wildlife.

Wolfe, A. D., T. Necamp, S. Fassnacht, P. Blischak, and L. Kubatko. 2016. Population genetics of *Penstemon albomarginatus* (Plantaginaceae), a rare Mojave Desert species of conservation concern. *Conservation Genetics*.

Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie, and B. S. Weir. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28: 3326-3328.

Appendices

Appendix A. Report on tGBS data from data2bio

Appendix B. Figures accompanying report on tGBS data from data2bio

Appendix C. Microsatellite length information from Eurofins

Significant Deviations

Deviations from the original proposal were mentioned throughout the methods for each task. The majority of these deviations related to field trips that did not take place as well as sampling approaches for population genetics and genomics of *Oreocarya crassipes*. The field trip schedule changed as Cohen relocated from Laredo, TX to Flint, MI and to an institution with a modified academic calendar. After field trips during the first year of the project, the phenology of the plant became clearer, so field trips, particularly those in August, did not seem necessary. Therefore, field trips were concentrated in the spring and early summer.

Given the challenges of conducting field work on private land in Texas, it was unfortunately not possible to gain access to all populations of *O. crassipes*. While the project would have been improved with greater sampling, the number of populations sampled, which are the same as those in Warnock (2012), appear sufficient to understand the conservation genetics and genomics of the species. Additionally, Cohen decided to use tunable Genotyping-by-Sequencing methods to sample SNPs from throughout the genome for 192 samples rather than sequence genomes from 10 individuals. This approach seemed preferable to due to the increased coverage of the genome for almost 20 times the number of individuals. In general, all deviations from the original proposal either enhanced the project or did not negatively impact it.

Data2Bio, LLC
2079 Roy J. Carver Co-Laboratory
Ames, Iowa 50011-3650
questions@data2bio.com

Oreocarya crassipes (Boraginaceae) Tunable Genotyping-By-Sequencing (tGBS) Project

Data2Bio, LLC
(http://www.data2bio.com)

20 August 2015

Projects: 20150514_Jim_Cohen_192_DNA
Client: James Cohen, PhD.
1700 University Ave., Flint, MI 48504
Kettering University
Email: jcohen@kettering.edu
Phone: (810) 249-4383

Data2Bio, LLC
2079 Roy J. Carver Co-Laboratory
Ames, Iowa 50011-3650
questions@data2bio.com

TABLE OF CONTENTS

Data2Bio, LLC
2079 Roy J. Carver Co-Laboratory
Ames, Iowa 50011-3650
questions@data2bio.com

Context and Summary

The goal of the project was to SNP-type 192 heterozygous *Oreocarya crassipes* lines with Data2Bio's tunable Genotyping by Sequencing (tGBS®) technology and to construct a phylogenetic tree. The *Oreocarya crassipes* genome is estimated to be 1.3 Gb in size. The client provided 192 DNA samples and requested that we conduct a reference-free analysis, the workflow for which is shown in **slide 3** of the slide deck provided with this report.

Data2Bio sequenced the 192 samples using four (4) runs on an Ion Proton Instrument, generating ~307.6M raw reads (**slide 4**), which after processing resulted in ~349.2M reads (**slide 5**). After trimming low quality bases, ~319.4M reads remained.

Subsequently, the trimmed reads from each sample were aligned to a set of condensed/assembled tGBS reads (a surrogate for a reference genome), which consisted of 620,191 contigs with a total length of 74.7Mb (**slide 6**). Approximately 91.1% and 70.3% of the trimmed reads could be aligned non-uniquely and uniquely, respectively (**slide 7**).

SNP calling was conducted using the uniquely aligned reads after interrogating 1,165,875 bases that have ≥5 reads in at least 50% of the samples (**slide 8**). Ultimately, a set of 17,143 high quality SNPs, each of which exhibited less than 50% missing data among the 192 samples (**slide 17**) was used to create a phylogenetic tree (**slide 21**). On average, each SNP call in each sample was supported by 92 reads (**slide 20**), allowing Data2Bio to make confident genotyping calls.

NOTE: Sequence reads associated with this project will be automatically purged from Data2Bio's servers approximately sixty days (60) after the delivery of this report. Please keep in mind that having the data stored locally on our servers will be useful if you ask our staff to troubleshoot or answer any questions regarding your project. If you would like us to purge the data sooner or later than 60 days after report delivery, please send us your request via email at questions@data2bio.com.

Data2Bio, LLC
2079 Roy J. Carver Co-Laboratory
Ames, Iowa 50011-3650
questions@data2bio.com

Data Generation

Data2Bio conducted tGBS on the 192 heterozygous samples provided by client using four (4) Ion Proton runs, generating a total of ~307.6M raw reads (**slide 4**), which after processing resulted in ~349.2M reads (**slide 5**). The number of reads can increase at this step because individual reads can be split into two reads. A summary histogram and a Bar-plot as well as a table of minimum, maximum, average and median numbers of raw reads per sample are provided in **slide 4.**

Each sequenced read was scanned for low quality regions and bases with PHRED quality scores of <15 out of 40 (≤3% error rate) were trimmed. After trimming low quality bases, about 319.4M reads remained, i.e., 8.5% of raw reads were dropped and 85.9% of base pairs remained after trimming (**slide 5**).

Data2Bio generated consensus sequences that could be used *in lieu* of a reference genome for alignment and SNP calling. Quality trimmed reads from all samples were combined and normalized to a maximum of 50X coverage. The sequencing errors in the reads were then corrected using Fiona. Coverage-normalized and error-corrected reads were then condensed using CD-HIT-454 with >= 96% identity to form consensus clusters. Clusters with <10 component reads and <50bp in length were discarded. Finally, 620,191 consensus sequences were obtained with a total length of 74.7Mb and an average contig length of 120bp (left panel of **slide 6**). The distribution of contigs lengths is presented in the right panel of **slide 6**.

Subsequently, the trimmed reads from each sample were aligned to the assembled condensed reads using GSNAP (WU and NACU 2010). A summary of total, average, median numbers of reads that aligned (uniquely and non-uniquely) are provided in table of **slide 7**. Approximately 91.1% and 70.3% of the trimmed reads could be aligned non-uniquely and uniquely, respectively.

Data2Bio, LLC
2079 Roy J. Carver Co-Laboratory
Ames, Iowa 50011-3650
questions@data2bio.com

Analysis of Samples

tGBS differs from conventional GBS in that fewer sites are sequenced (i.e., it exhibits a higher "genome reduction level" or GRL). Hence, assuming that equal amounts of sequencing data are generated, tGBS provides greater read depth per sequenced site. Consequently, SNPs can be called at higher confidence and the need for imputation, which can introduce errors, is either eliminated or reduced. *All genotypes provided or referenced in this report were derived empirically (i.e., no imputation was employed).*

Data2Bio generates several sets of SNPs during the analysis of tGBS projects. The first set ("polymorphic sites") includes all sites that differ from the reference in at least one sample. This set is obtained after considering all reads that align to the reference genome (or consensus sequences if a reference genome is not available). We then examine, sample-by-sample, only those tGBS reads that align to these polymorphic sites to identify a set of SNPs we term "ALL SNPs". These SNPs are polymorphic within the population under analysis and meet certain other criteria listed in **slide 11**. Subsequently, additional cut-offs are applied (e.g., a minimum percentage of missing data, minimum and/or maximum heterozygosity rates, and/or minimum minor allele frequencies) to improve the utility of the selected SNPs (Low Missing Data or LMD) sets. These cut-offs are customized to identify a SNP set that best meets project needs. For example, defining an acceptable percentage of missing data per SNP across samples depends on project goals. In this project we used a missing data cut-off filter of ≤50%. This cut-off is not, however, cast in stone. One could easily define a set of SNPs for a given set of lines from the "ALL SNPs" table with a different missing data cut-off depending on project needs.

The numbers of detected SNPs are shown below:

	No. SNPs
Polymorphic Sites	373,702
ALL SNPs	61,487
LMD50 SNPs	17,143

Data2Bio, LLC
2079 Roy J. Carver Co-Laboratory
Ames, Iowa 50011-3650
questions@data2bio.com

Using the reads from the 192 samples that uniquely align to the assembled condensed reads Data2Bio identified 373,702 polymorphic sites (**slide 9**) after interrogating 1,165,871 bases that have ≥5 reads in at least 50% of the samples (**slide 8**). A subset of 61,487 were identified as ALL SNPs. The criteria for tGBS genotyping and ALL SNPs filtering are shown in **slides 10-11**. Distributions of various characteristics for the ALL SNP dataset, including quantity of missing data, minor allele frequency, heterozygosity and genotype number are summarized in **slide 12**. The numbers of ALL SNPs per sample that are homozygous for the "REF" (reference) allele, homozygous for the ALT (alternative) allele, heterozygous and missing are shown in the top panel of **slide 13**. To allow for comparisons among samples unbiased by varying levels of missing data among samples, the bottom panel of **slide 13** shows the proportions of the SNPs per sample that are homozygous for the REF allele, homozygous for the ALT allele, or heterozygous among the non-missing data. The average missing data rate per SNP site across samples is provided in left panel of **slide 14**. The right panel presents the minimum, maximum, average and median numbers of reads per SNP site per sample. Note that only samples with data were considered.

Distributions of missing data rate and heterozygosity of these samples are shown in **slide 15**. No sample were removed from subsequent analyses due to high missing rate of data or excessive amounts of heterozygosity.

Subsequently, Data2Bio filtered the ALL SNPs set to explore different missing data rates used for genotype across the 192 samples. The resulting number of SNPs remaining, missing data points and detected polymorphism rate for each LMD level are displayed in **slide 16**. Data2Bio selected LMD50 as the appropriate cut-off for this project. At this cut-off the % heterozygosity in this diversity panel is ~1.5%.

The criteria for tGBS genotyping and LMD50 SNPs filtering are shown in **slide 17**. The resulting LMD50 (low missing data, each of which was genotyped in at least 50% of the samples) SNP set contains 17,143 SNPs. Various characteristics of the LMD50 SNP dataset, including quantity of missing data, minor allele frequency, heterozygosity and genotype number are summarized in **slide 18**. In **slide 19** the numbers of LMD50 SNPs per sample that are homozygous for the REF allele, homozygous for the ALT allele,

Data2Bio, LLC
2079 Roy J. Carver Co-Laboratory
Ames, Iowa 50011-3650
questions@data2bio.com

heterozygous and missing are shown in the top panel. The bottom panel of **slide 19** shows the proportions of the SNPs per sample that are homozygous for the REF allele, homozygous for the ALT allele, or heterozygous among the non-missing data. The average missing data rate per LMD50 SNP site across samples is provided in the left panel in **slide 20**. Sequencing data support 65.1% of all possible SNP calls (No. samples x No. SNPs). The right panel presents the minimum, maximum, average and median numbers of reads per SNP per sample. Each SNP call is supported by an average of 92 tGBS sequence reads *per sample*, ensuring the accuracy of these non-imputed SNP calls.

Finally, a phylogenetic tree based on the 17,143 LMD50 SNPs data is presented in **slide 21**. A newick version (Oreocarya.LMD50SNPs.PhylogeneticTree.APE.newick.txt) of the tree is provided on the hard drive. The tree can be viewed using software ITOL (http://itol.embl.de/index.shtml) (LETUNIC and BORK 2006).

Slide 22 highlights in red those samples with more than 80% missing SNP calls. In **slide 23** we have distinguished among these samples those that have low number of tGBS reads (blue) and those that have high levels of missing SNPs despite have plenty of tGBS reads (red). In the case of the blue samples the missing SNPs are likely a consequence of low read number. Note that these samples cluster in the tree. It is at least possible that this clustering is an artifact of missing data. Hence, it may be appropriate to treat this clade should be some suspicion. Your understanding of the relationships among these samples may provide guidance. But something biological is likely responsible for the low SNP call rate among the red samples. We suspect that these samples are genetically quite distinct from the rest of the population, resulting in alignment problems, reducing SNP calling success.

There is other evidence for substantial heterogeneity among the samples in this diversity panel. Considering only the black dots on **slide 22**, samples fall on either the upper or lower curve. This is most obvious among the ALL SNPs, but is also apparent among the LMD50 SNPs. The samples in the upper curve differ from the samples in the lower curve in that they require more reads to yield the same SNP call rate. This is reminiscent of a project we conducted for another client that included both crop samples and samples

Data2Bio, LLC
2079 Roy J. Carver Co-Laboratory
Ames, Iowa 50011-3650
questions@data2bio.com

from the wild relative of that crop. What we believe is going on is that the "genome" that we created is based on the consensus sequence of the diversity panel. If some samples are quite distinct genetically, their sequences will be under-represented in this consensus "genome". Consequently, reads from these lines will experience reduced alignment success and consequently reduced SNP calling.

In the earlier project for the other client we addressed this problem in a follow-up project by constructing two separate "genomes" by separately condensing tGBS reads from the upper curve samples and lower curve samples and then repeating the SNP calling for each sample using the appropriate "genome". Let us know if you would like to discuss this possibility.

Provided data tables include genotyping calls for each SNP for each of the 192 samples and the numbers of reads supporting each genotype call and the numbers of reads that disagree with that call. Note that every data point in the delivered tales is supported by actual data-we did not conduct any imputation, thereby entirely avoiding imputation-induced errors in SNP calling. Data2Bio has also provided the tGBS DNA sequencing reads after the removal of proprietary sequences added during library preparation, and before and after trimming low quality bases.

Data2Bio, LLC
2079 Roy J. Carver Co-Laboratory
Ames, Iowa 50011-3650
questions@data2bio.com

Outputs of analyses

The numbers of reads obtained for each sample and genotyping calls are provided in tables provided along with this report. Listed below are descriptions of provided files in folder **tables**:

- **Oreocarya.all.snps.genotype.txt**: This much larger table contains all the SNPs identified by Data2Bio in the population and includes all the SNPs reported in the low missing data tables as well as many SNPs, which were genotyped in only a subset of the client's samples. The SNPs, genotype calls, and samples in this file were also formatted into Variant Calling Format (VCF) format and its included under the same directory with file extensions ending with *.vcf.

- **Oreocarya.all.snps.genotype.context_sequences.txt**: Context sequence with at most 100 bp upstream and downstream of each of the filtered SNP sites.

- **Oreocarya.LMD50.snps.genotype.txt:** This file contains markers for which data are available for at least 50% of the lines. This data set is a subset of ALL SNPs. The SNPs, genotype calls, and samples in this file were also formatted into Variant Calling Format (VCF) format and its included under the same directory with file extensions ending with *.vcf.

- **Oreocarya.LMD50.snps.genotype.context_sequences.txt**: Context sequence with at most 100 bp upstream and downstream of each of the filtered SNP sites.

- **Oreocarya.*.AlleleCounts:** Read counts per allele of each sample for each of the filtered SNP sites.

- **Oreocarya.LMD50SNPs.PhylogeneticTree.APE.newick.txt:** The Newick format of phylogenetic tree of the 192 samples constructed based on the LMD50 SNPs.

Data2Bio, LLC
2079 Roy J. Carver Co-Laboratory
Ames, Iowa 50011-3650
questions@data2bio.com

We have also provided DNA sequence reads, after trimming off proprietary sequences that are added during tGBS library preparation for each of the 192 samples.

- **raw:** Sequencing reads generated by Data2Bio after the removal of proprietary sequences.
- **trimmed:** Sequencing reads remaining after the removal of both proprietary sequences and sequences with low quality scores
- **ref.free:** generated assembled sequences in fasta format of samples.
- **SAM/BAM:** Alignment files in SAM and BAM output format (LI et al. 2009) (http://samtools.sourceforge.net/)
- **figures:** all figures presented in the slide deck.

Data2Bio, LLC
2079 Roy J. Carver Co-Laboratory
Ames, Iowa 50011-3650
questions@data2bio.com

Data visualization

IGV-based visualization of alignments of reads to the reference assembly

Data2Bio provides a Youtube video to explain how to use IGV for data visualization (http://www.youtube.com/watch?v=tOV47_ogPWY). Files for IGV visualization are included under the "BAM" folder in your Data2Bio output:

Brief instructions for IGV visualization:

- Download IGV here: http://www.broadinstitute.org/igv/download
- Install IGV
- Start IGV visualization program
- Copy aligment files (BAM/*bam and BAM/*bam.bai from your Data2Bio output) to your local computer
- Make sure you save the *bam and *bam.bai files under the same folder
- Load *bam files in the IGV. Now you can visualize the alignment results

Data2Bio, LLC
2079 Roy J. Carver Co-Laboratory
Ames, Iowa 50011-3650
questions@data2bio.com

Methods

Trimming of sequencing reads

Prior to alignment, the nucleotides of each raw read were scanned for low quality bases. Bases with PHRED quality value <15 (out of 40) (EWING and GREEN 1998; EWING *et al.* 1998), i.e., error rates of ≤3%, were removed by our trimming pipeline. Each read was examined in two phases. In the first phase reads were scanned starting at each end and nucleotides with quality values lower than the threshold were removed. The remaining nucleotides were then scanned using overlapping windows of 10 bp and sequences beyond the last window with average quality value less than the specified threshold were truncated. The trimming parameters were referred to the trimming software, Lucy (CHOU *et al.* 1998; LI and CHOU 2004).

Consensus reference sequence generation

Trimmed sequence reads from all samples were combined and normalized to a maximum of 50x coverage, using diginorm (BROWN et al. 2012). The sequencing errors in the reads were then corrected using Fiona (SCHULZ et al. 2014). The coverage-normalized and error-corrected reads were then condensed using CD-HIT-454 (FU et al. 2012) with >= 96% identity to form consensus clusters. Clusters with <10 component reads and <50bp in length were discarded in a final clean-up.

Alignment of reads to consensus reference sequence

Trimmed reads were aligned to the consensus reference sequence using GSNAP (WU and NACU 2010) and confidently mapped reads were filtered if it mapped uniquely (≤2 mismatches every 36 bp and less than 5 bases for every 75 bp as tails) and used for subsequent analyses.

Discovery of Polymorphic Sites

The coordinates of confident and single (unique) alignments to the consensus reference sequence that passed our filtering criteria were used for SNP

Data2Bio, LLC
2079 Roy J. Carver Co-Laboratory
Ames, Iowa 50011-3650
questions@data2bio.com

discovery. Polymorphisms at each potential SNP site were carefully examined and putative homozygous and heterozygous SNPs were identified in each sample separately using the following criteria:

- **Homozygous SNP calling**
  - The most common allele must be supported by at least 80% of all the aligned reads covering that position.
  - At least 5 unique reads must support the most common allele.
  - Polymorphisms in the first and last 3 bp of each read were ignored.
  - Each polymorphic base must have at least a PHRED base quality value of 20 (≤1% error rate)

- **Heterozygous SNP calling**
  - Each of the two most common alleles must be supported by at least 30% of all aligned reads covering that position.
  - At least 5 unique reads must support each of the two most-common alleles.
  - The sum of reads of the two most common alleles must account for at least 80% of all aligned reads covering that nucleotide position.
  - Polymorphisms in the first and last 3 bp of each quality-trimmed read were ignored.
  - Each polymorphic base must have at least a PHRED base quality value of 20 (≤1% error rate)

Any site that was deemed to be polymorphic (homozygous or heterozygous) as compared to the reference genome sequence in at least one sample was included in the set of polymorphic sites.

Data2Bio, LLC
2079 Roy J. Carver Co-Laboratory
Ames, Iowa 50011-3650
questions@data2bio.com

Criterion for tGBS genotyping

Homozygous vs. Heterozygous Calls

The criteria for tGBS genotyping are below:

A SNP site was called as homozygous in a given diploid sample if at least 5 reads supported the major common allele at that site and at least 90% of all aligned reads covering that site shared the same nucleotide at that site.

A SNP was called as heterozygous in a given diploid sample if at least 1 read supported each of at least two different alleles and each of the two allele types separately comprised more than 20% of the reads aligning to that site. And when the sum of the reads supporting those two alleles at least equal to 5 and comprised at least 90% of all reads covering the site.

Initial QC of SNP Quality for the ALL SNP Data Set

- Missing data rate ≤ 80%
- Allele number = 2
- Number of genotypes ≥ 2
- Minor Allele Frequency (MAF) ≥ 2/192
- Heterozygosity range: $(2 \text{ X Frequency}_{\text{Allele1}} \text{ X Frequency}_{\text{Allele2}}) \pm 0.2$

Defining LMD50 (low missing data) SNP Dataset

The ALL SNP sets were further filtered to define LMD50 SNP set.

- Missing data rate ≤ 50%
- Allele number = 2
- Number of genotypes ≥2
- Minor Allele Frequency (MAF) ≥ 2/192
- Heterozygosity range: $(2 \text{ X Frequency}_{\text{Allele1}} \text{ X Frequency}_{\text{Allele2}}) \pm 0.2$

Data2Bio, LLC
2079 Roy J. Carver Co-Laboratory
Ames, Iowa 50011-3650
questions@data2bio.com

Phylogenetic tree construction

Pairwise distances were estimated between genotyped individuals using an unbiased model of substitution frequencies. Distance estimates were then used to construct a phylogenetic tree using the Neighbor-Joining-like algorithm described by CRISCUOLO and GASCUEL (2008) and implemented in the njs module of the R APE package. Unlike conventional neighbor-joining methods, the njs algorithm is tolerant of missing data, enabling its use with GBS data. Relative branch lengths are proportional to the amount of divergence observed among individuals.

Data2Bio, LLC
2079 Roy J. Carver Co-Laboratory
Ames, Iowa 50011-3650
questions@data2bio.com

References

BROWN, C. T., A. HOWE, Q. ZHANG, A. B. PYRKOSZ, T. H. BROM, 2012 A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data, Available: arXiv:1203.4802. Accessed 2013 June 25.

CHOU, H. H., G. SUTTON, A. GLODEK and J. SCOTT, 1998 Lucy - A Sequence Cleanup Program, pp. in *Proceedings of the Tenth Annual Genome Sequencing and Annotation Conference (GSAC X)*, Miami, Florida.

CRISCUOLO, A., and O. GASCUEL, 2008 Fast NJ-like algorithms to deal with incomplete distance matrices. BMC Bioinformatics 2008, 9:166. doi:10.1186/1471-2105-9-166

EWING, B., and P. GREEN, 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res **8:** 186-194.

EWING, B., L. HILLIER, M. C. WENDL and P. GREEN, 1998 Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res **8:** 175-185.

FU, L., B. NIU, Z. ZHU, S. WU and W. LI, 2012 CD-HIT: accelerated for clustering the next generation sequencing data. Bioinformatics **28**: 3150-3152.

LETUNIC and BORK, 2006 Bioinformatics **23**(1):127-8

LETUNIC and BORK, 2011 Nucleic Acids Res doi: 10.1093/nar/gkr201

LI, S., and H. H. CHOU, 2004 LUCY2: an interactive DNA sequence quality trimming and vector removal tool. Bioinformatics **20:** 2865-2866.

SCHULZ, M. H., D. WEESE, M. HOLTGREWE, V. DIMITROVA, S. NIU, K. REINERT and H. RICHARD, 2014 Fiona: a parallel and automatic strategy for read error correction. Bioinformatics **30**(17): i356–i363. doi:10.1093/bioinformatics/btu440.

WU, T. D., and S. NACU, 2010 Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics **26:** 873-881.

Oreocarya crassipes (Boraginaceae) tGBS Analyses

James Cohen, PhD.
1700 University Ave., Flint, MI 48504
Kettering University
jcohen@kettering.edu

Data2Bio, LLC

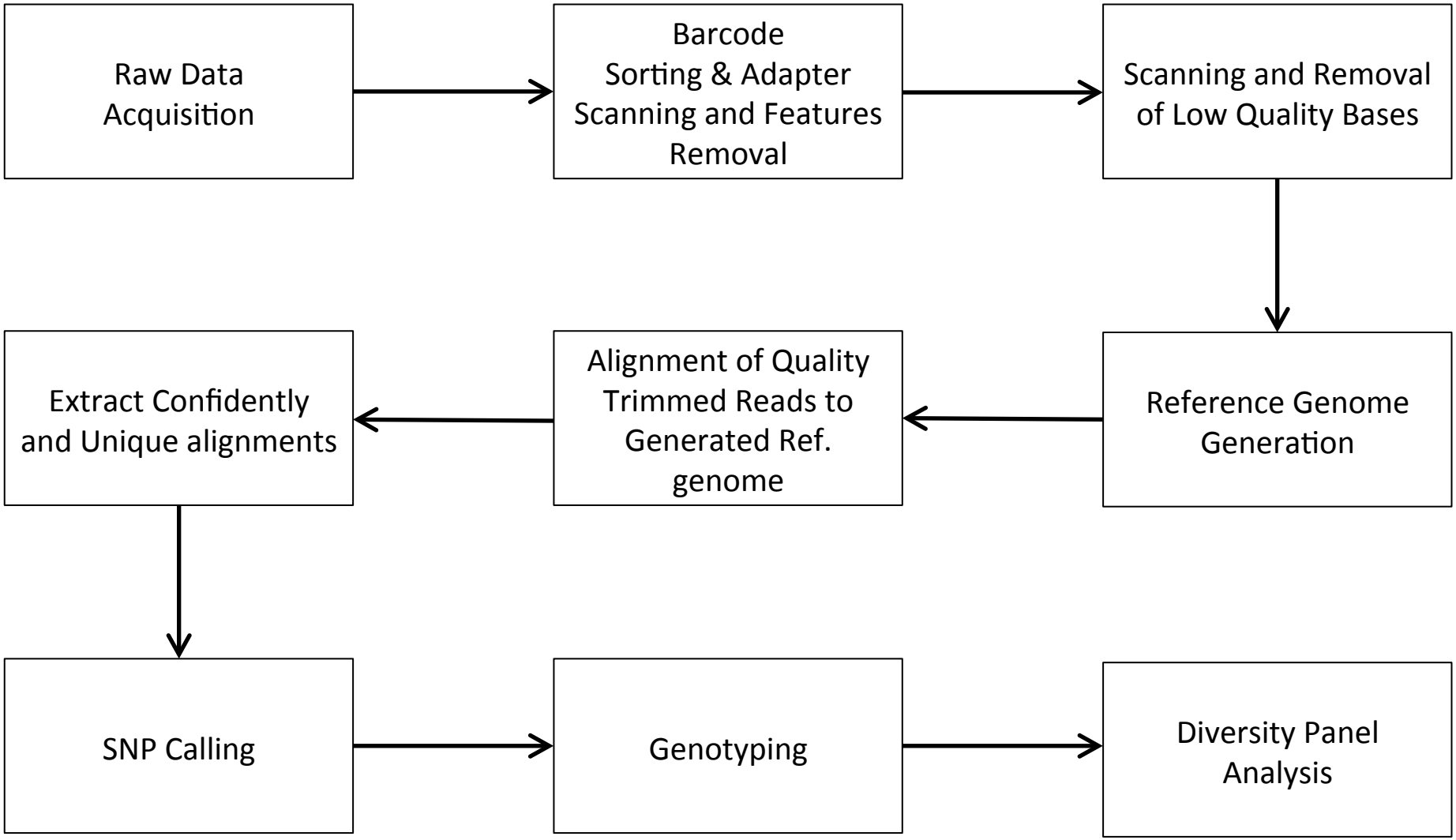
20 August 2015

DATA2BIO™

Services to be Provided

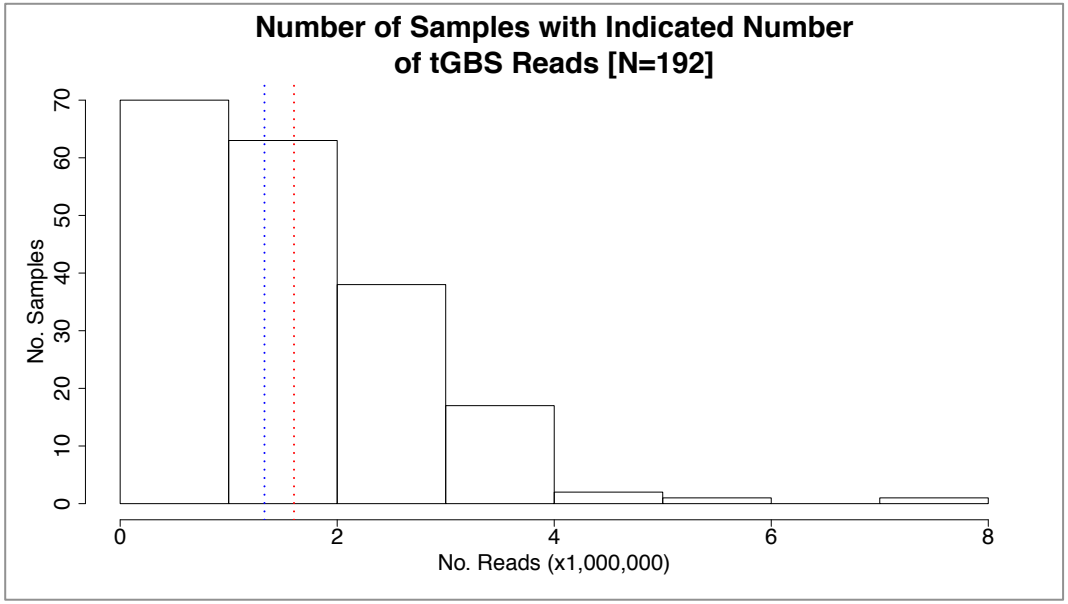
- Tunable Genotyping-by-Sequencing (tGBS®) of 192 heterozygous lines of *Oreocarya crassipes (Boraginaceae)* – Estimated genome size of ~1.3Gb
  - tGBS analysis (2bp selection) of 192 Heterozygous samples
  - QC and Trimming of reads
  - Reference-Free genome generation
  - Alignment of reads to reference-free genome
  - SNP calls/SNP-typing of each line
  - Diversity panel analyses

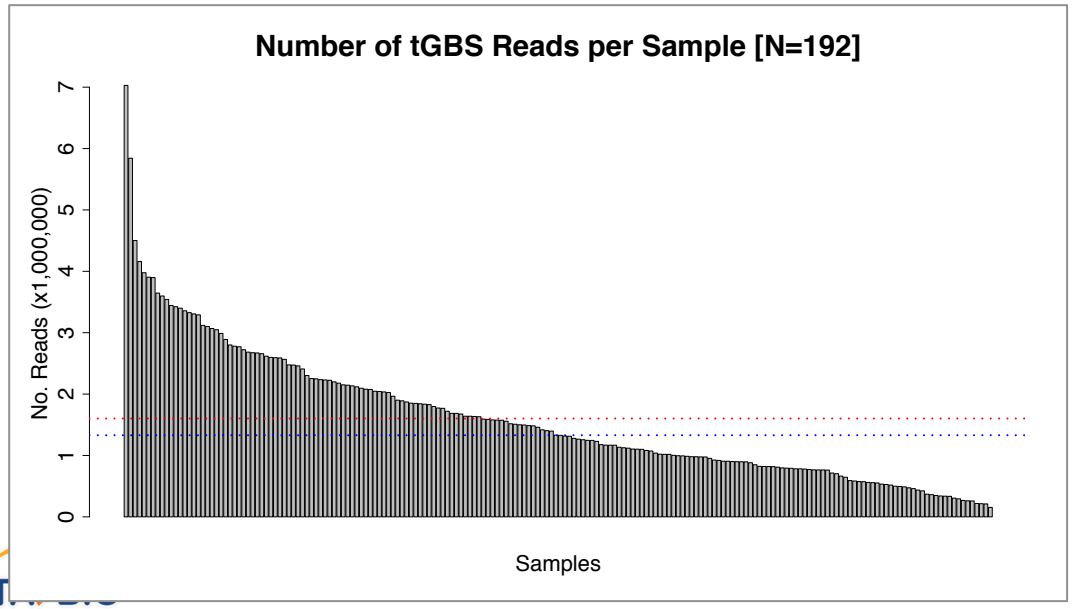
tGBS Workflow WITHOUT Reference Genome



Total Sequenced tGBS Reads Per Sample [N=192]

(Sequenced using 4 Runs on an Ion Proton Instrument)





DESCRIPTION	NO.
Number of DNA Samples	192
Total Reads*	**307,612,207**
Minimum Reads per Sample	153,638
Maximum Reads per Sample	7,028,811
Average Reads per Sample	**1,602,146**
Median Reads per Sample	**1,329,942**

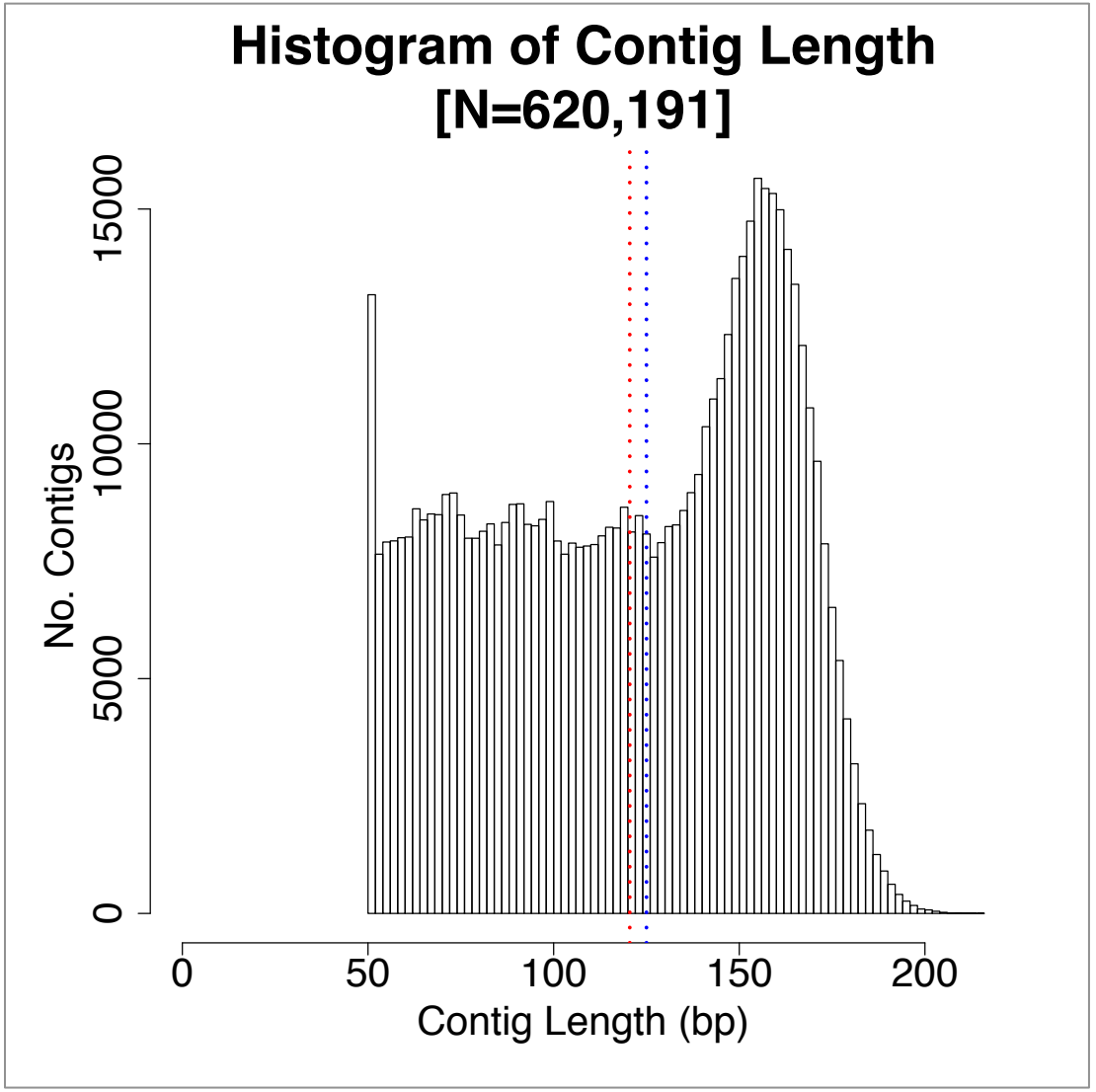
* Raw reads without any further processing

Quality Trimming of tGBS Reads

No.	Run ID	Processed Reads			Quality Trimmed Reads		
		No. Reads	Base Pairs (bp)	Read Length (bp)	No. Reads	Base Pairs (bp)	Read Length (bp)
1	G53_P1-1	116,633,396	14,150,935,317	123	110,715,577 (94.9%)	12,851,185,867 (90.8%)	117
2	G53_P2-1.redo	42,015,957	4,500,811,160	106	31,703,960 (75.5%)	2,791,814,654 (62.0%)	87
3	G53_ChP1-1	98,083,042	11,791,800,249	118	90,822,452 (92.6%)	10,280,513,544 (87.2%)	111
4	G53_ChP2-1	92,469,606	11,096,222,648	119	86,143,377 (93.2%)	9,770,120,282 (88.0%)	113
	SUM	**349,202,001**	**41,539,769,374**	**117**	**319,385,366 (91.5%)**	**35,693,634,347 (85.9%)**	**107**
	AVERAGE per Sample	**1,818,760**	**216,352,965**	**117**	**1,663,465 (91.5%)**	**185,904,345 (85.9%)**	**107**
	MEDIAN per Sample	**1,496,209**	**177,100,184**	**116**	**1,344,697 (89.9%)**	**148,141,810 (83.6%)**	**105**

Reference-Free Genome Generation

DESCRIPTION	NO.
No. Contigs	620,191
Total Bases	74,732,711
GC Content	41.5%
Minimum Length (bp)	50
Maximum Length (bp)	216
Average Length (bp)	120
Median Length (bp)	125
N50[1]	232,511
L50[2]	143



[1] **N50:** The largest number 'L' such that the combined length of all contigs of length >= 'L' is at least 50% of the total length of all contig
[2] **L50:** The length of contig/scaffolds at the N50 point

Alignment of Quality Trimmed Reads to Reference-Free Genome

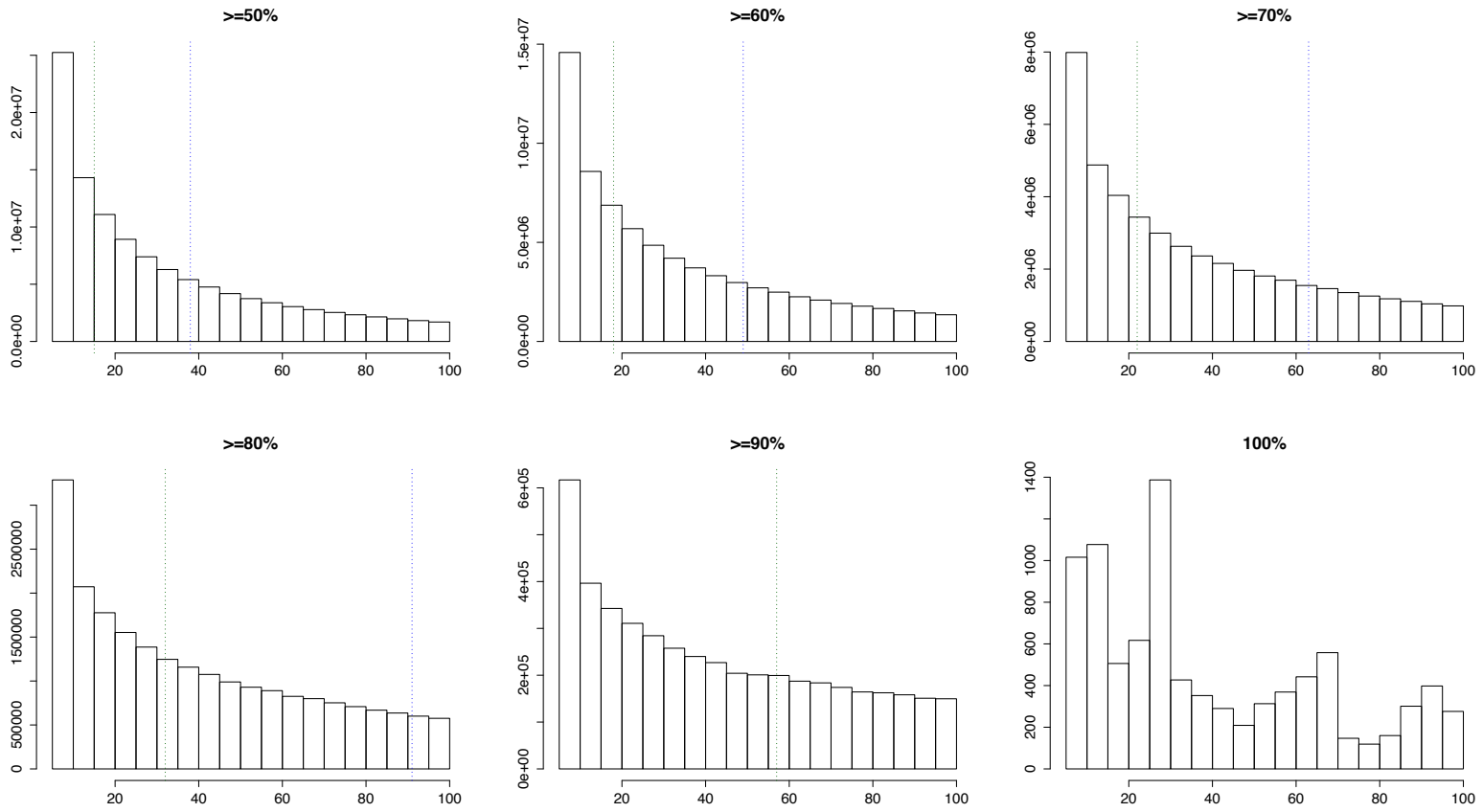
No.	Run ID	Quality Trimmed Reads			Alignment To Reference-Free Genome	
		No. Reads	Base Pairs (bp)	Read Length (bp)	Alignments (≥1 Location)	Unique Alignments (Single Location)
1	G53_P1-1	110,715,577	12,851,185,867	117	102,564,439 (92.6%)	79,979,828 (72.2%)
2	G53_P2-1.redo	31,703,960	2,791,814,654	87	25,899,146 (81.7%)	20,494,234 (64.6%)
3	G53_ChP1-1	90,822,452	10,280,513,544	111	83,150,992 (91.6%)	63,371,320 (69.8%)
4	G53_ChP2-1	86,143,377	9,770,120,282	113	79,261,561 (92.0%)	60,563,620 (70.3%)
	SUM	**319,385,366**	**35,693,634,347**	**107**	**290,876,138 (91.1%)**	**224,409,002 (70.3%)**
	AVERAGE per Sample	**1,663,465**	**185,904,345**	**107**	**1,514,979 (91.1%)**	**1,168,796 (70.3%)**
	MEDIAN per Sample	**1,344,697**	**148,141,810**	**105**	**1,228,103 (91.3%)**	**939,761 (69.9%)**

Only UNIQUE ALIGNMENTS were used in subsequent analyses

Number of Interrogated Bases

(Both polymorphic & non-polymorphic bases with ≥5 reads/sample in the indicated fraction of samples)

Read counts per interrogated base per sample in the indicated fraction of samples



Minimum % Samples Interrogated Per Base	Number Interrogated Bases	Percent Missing Data	25% Percentile Reads/Interrogated Base/Sample	Average Percentile Reads/Interrogated Base/Sample	Median Percentile Reads/Interrogated Base/Sample
≥50%	1,165,875	32.1%	15	118	38
≥60%	748,311	24.8%	18	147	49
≥70%	472,215	18.7%	22	185	63
≥80%	250,202	13.0%	32	259	91
≥90%	70,567	6.6%	57	498	168
100%	469	0.0%	305	1,508	805

SNP Discovery [N=192 Samples]

Total Polymorphic Sites Discovered: 373,702

- **Homozygous SNPs Criteria:**
  - Ignore first and last 3 bp of each read
  - Only consider polymorphic sites with PHRED quality ≥20 (≤1% error rate)
  - No. reads ≥ 5
  - Allele frequency among reads within a sample: ≥ 0.8
- **Heterozygous SNPs Criteria:**
  - Ignore first and last 3 bp of each read
  - Only consider polymorphic sites with PHRED quality ≥20 (≤1% error rate)
  - No. reads ≥ 5
  - Allele frequency per allele among reads within a sample ≥ 0.3
  - Overall allele frequency among reads within a sample ≥ 0.8

Criteria for tGBS genotyping

Homozygous call:

major alleles ≥ 5 reads & freq ≥ 0.9

Heterozygous call:

each allele ≥ 1 read & freq ≥ 0.2 and sum of the two alleles ≥ 5 total reads & total freq ≥ 0.9

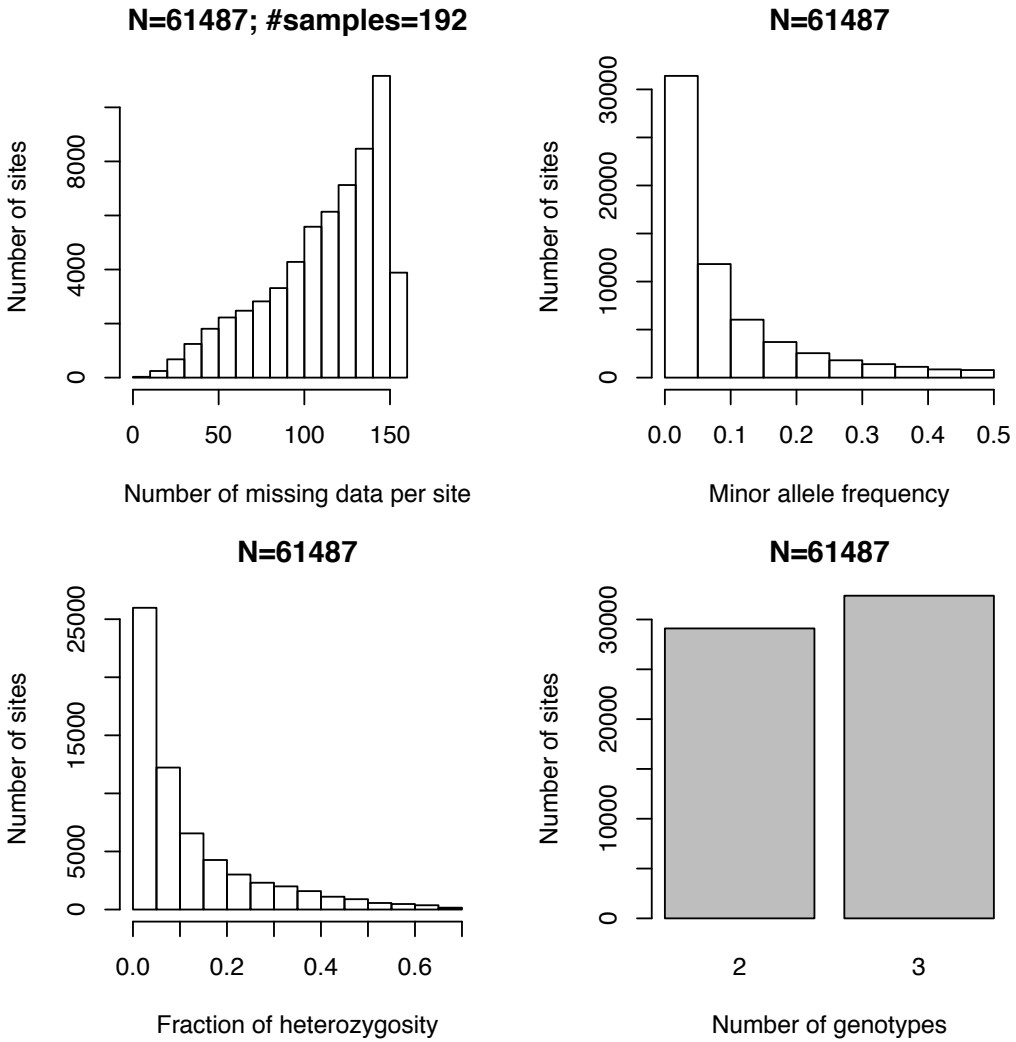
SNP filtering: ALL SNPs

61,487 remaining SNPs

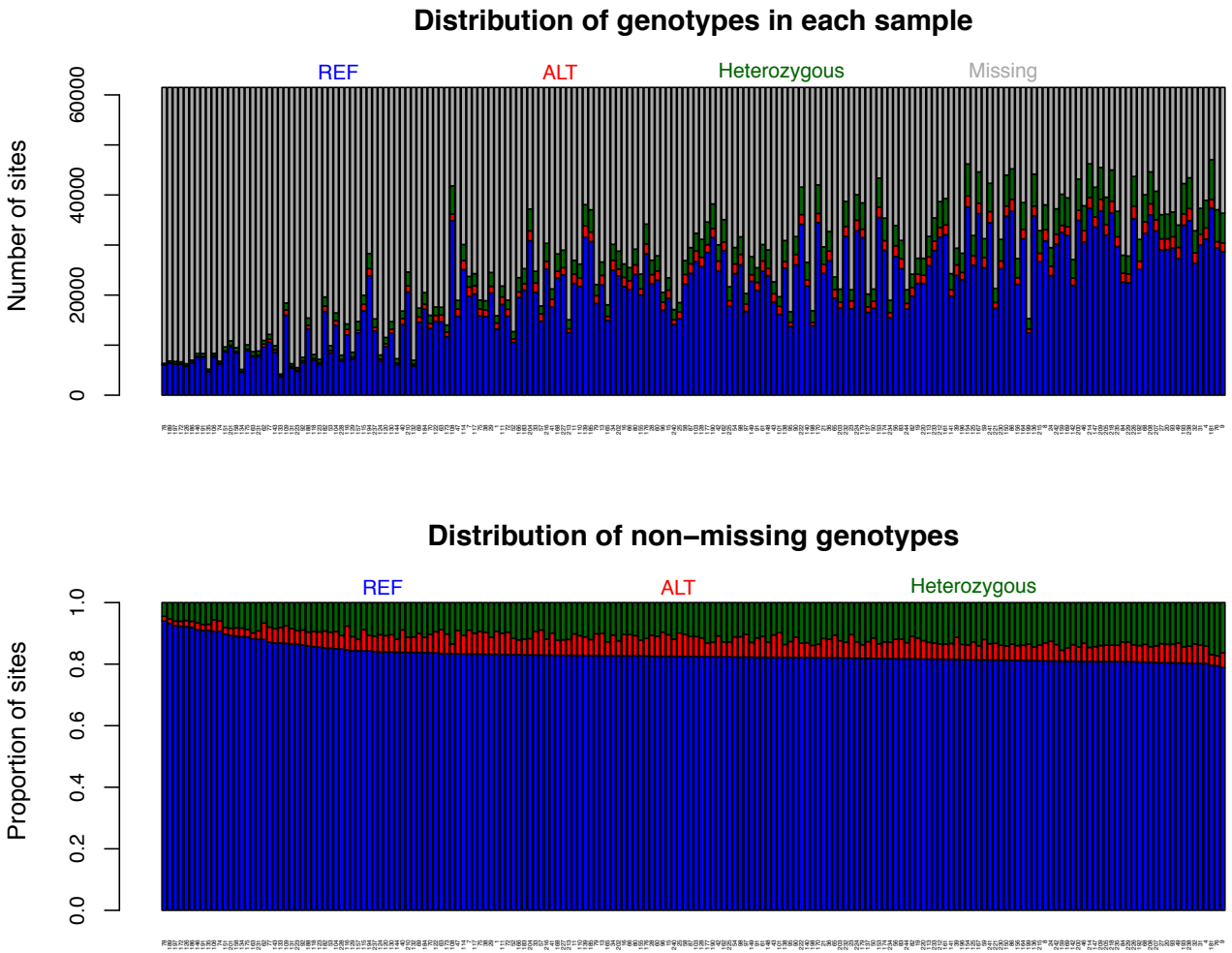
Filtering Criteria

- Missing data rate ≤ 80%
- Allele number = 2
- Genotype ≥ 2
- Minor Allele Frequency ≥ 2 / 192
- Heterozygosity range:
  - $(2 \text{ X } \text{Frequency}_{\text{Allele1}} \text{ X } \text{Frequency}_{\text{Allele2}}) \pm 0.2$

ALL SNP (N=61,487) Summary

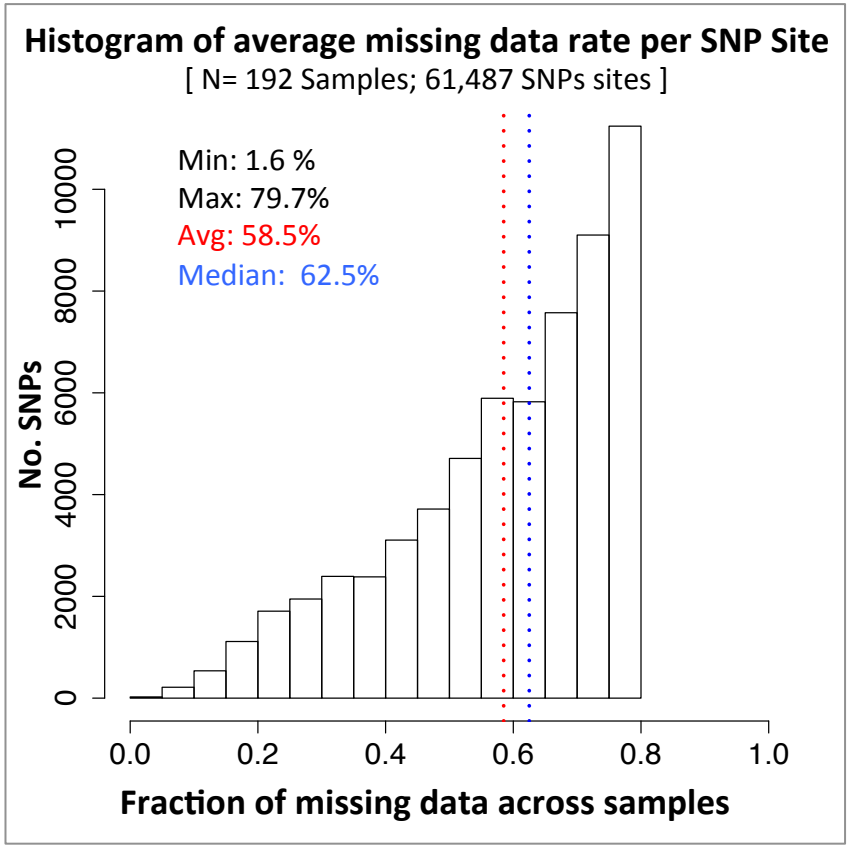


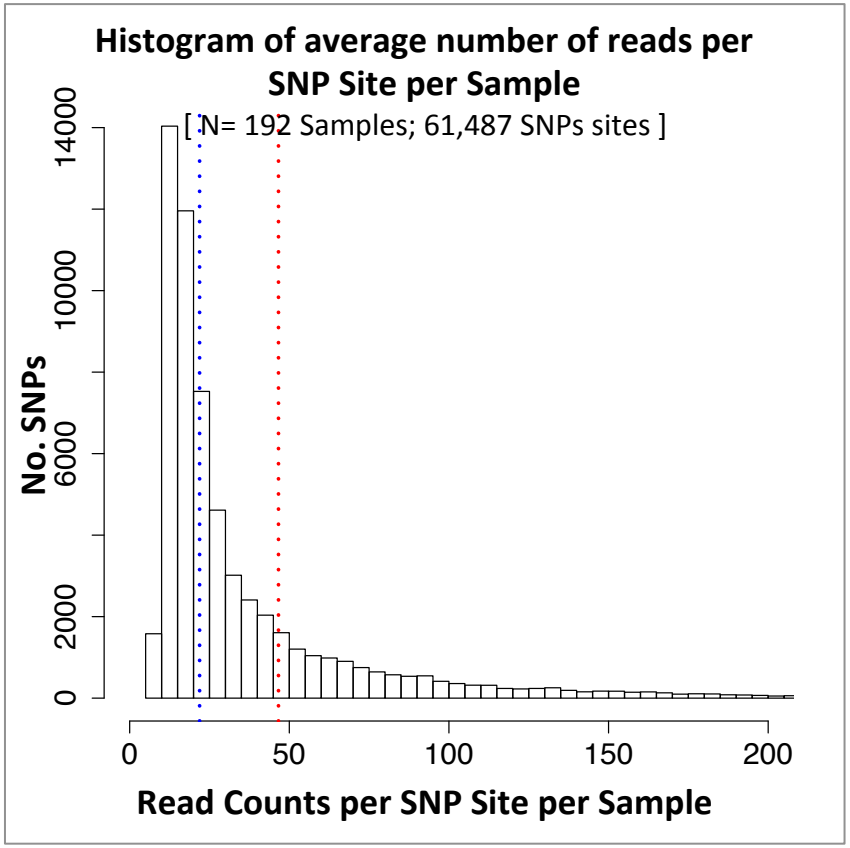
Sample (N=192) Summary for ALL SNPs



ALL SNPs : Average Missing Rate per SNP Site and read counts per SNP site per sample

(N= 192 Samples – 61,487 SNPs sites)





Overall Missing Data Points:

6,905,589 / 11,805,504 = **58.5%**

Min: 7 reads/SNP site/Sample*

Max: 32,706 reads/SNP site/Sample*

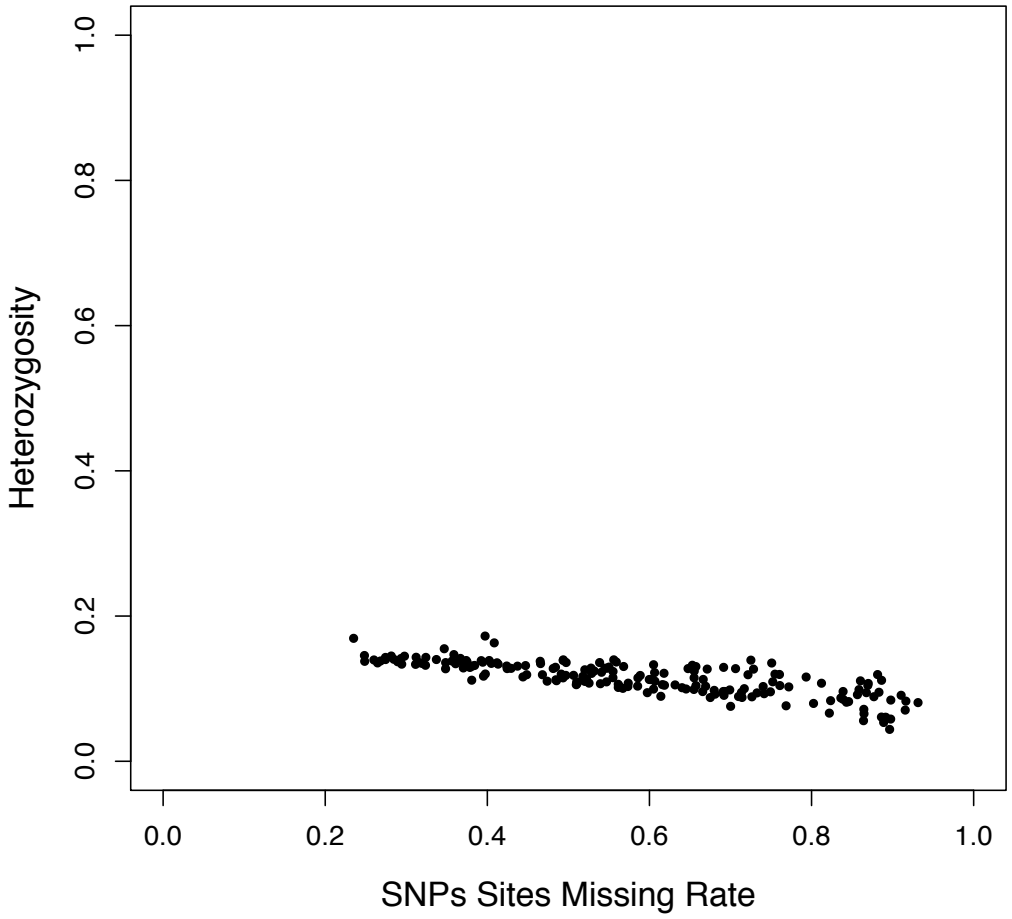
Avg: 47 reads/SNP site/Sample*

Median: 22 reads/SNP site/Sample*

* Only samples with data were considered

Missing Rate and Heterozygosity

Based on ALL SNPs



SNP Genotyping Summary [N=192]

Exploration of Different Missing Data Rates Used for Genotyping

LMDX: the set of nucleotides/SNPs that were interrogated/genotyped in at least X% of individuals

Low Missing Data Rate (LMD)	No. SNPs	Missing Data Points	% Polymorphisms*
LMD50	17,143	1,148,585 / 3,291,456 = 34.9%	17,143/1,165,875=1.47%
LMD40	10,321	554,432 / 1,981,632 = 28.0%	10,321/748,311=1.38%
LMD30	5,545	232,944 / 1,064,640 = 21.9%	5,545/472,215=1.17%
LMD20	1,888	54,495 / 362,496 = 15.0%	1,888/250,202=0.75%
LMD10	238	3,577 / 45,696 = 7.8%	238/70,567=0.34%

* Calculated as: (#SNPs/ #Interrogated bases) x 100)

DATA2BIO™

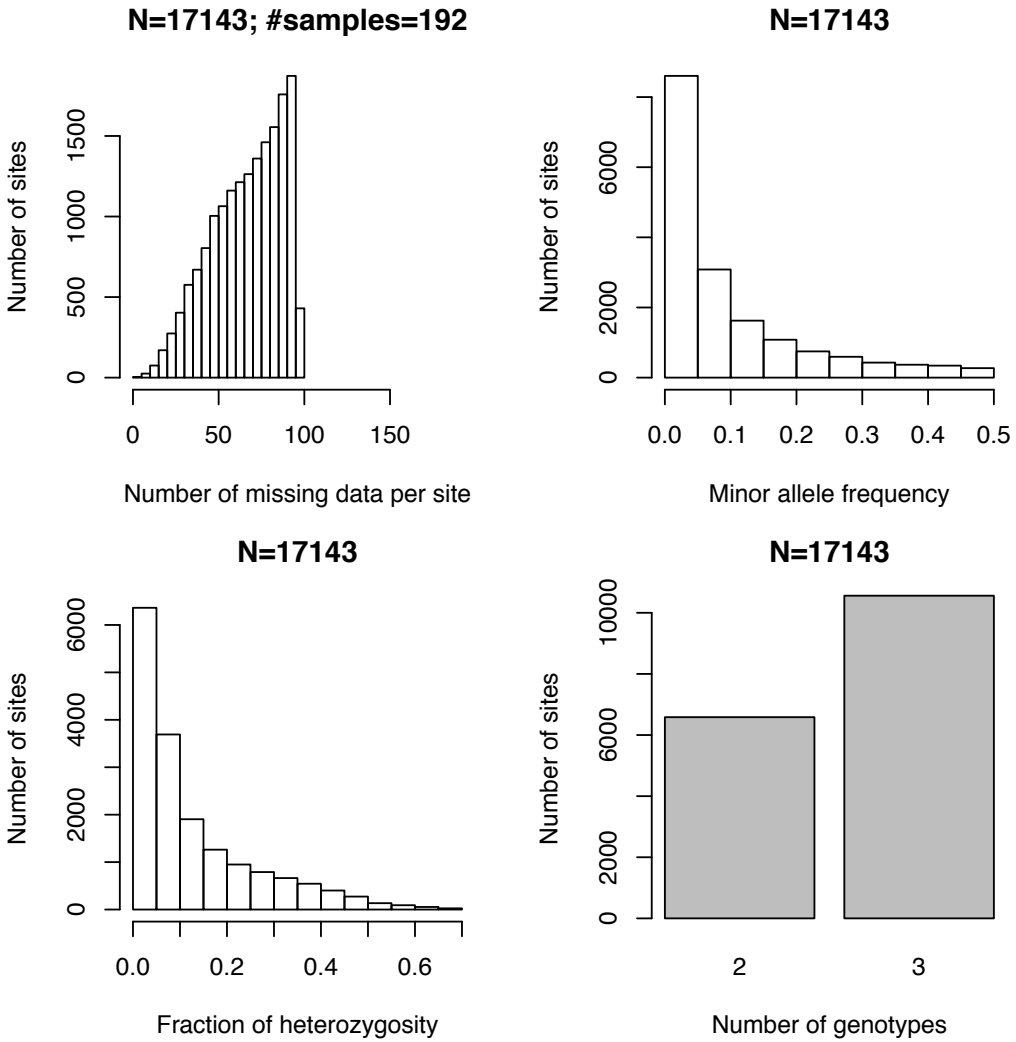
SNP filtering: LMD50 SNPs

17,143 remaining SNPs

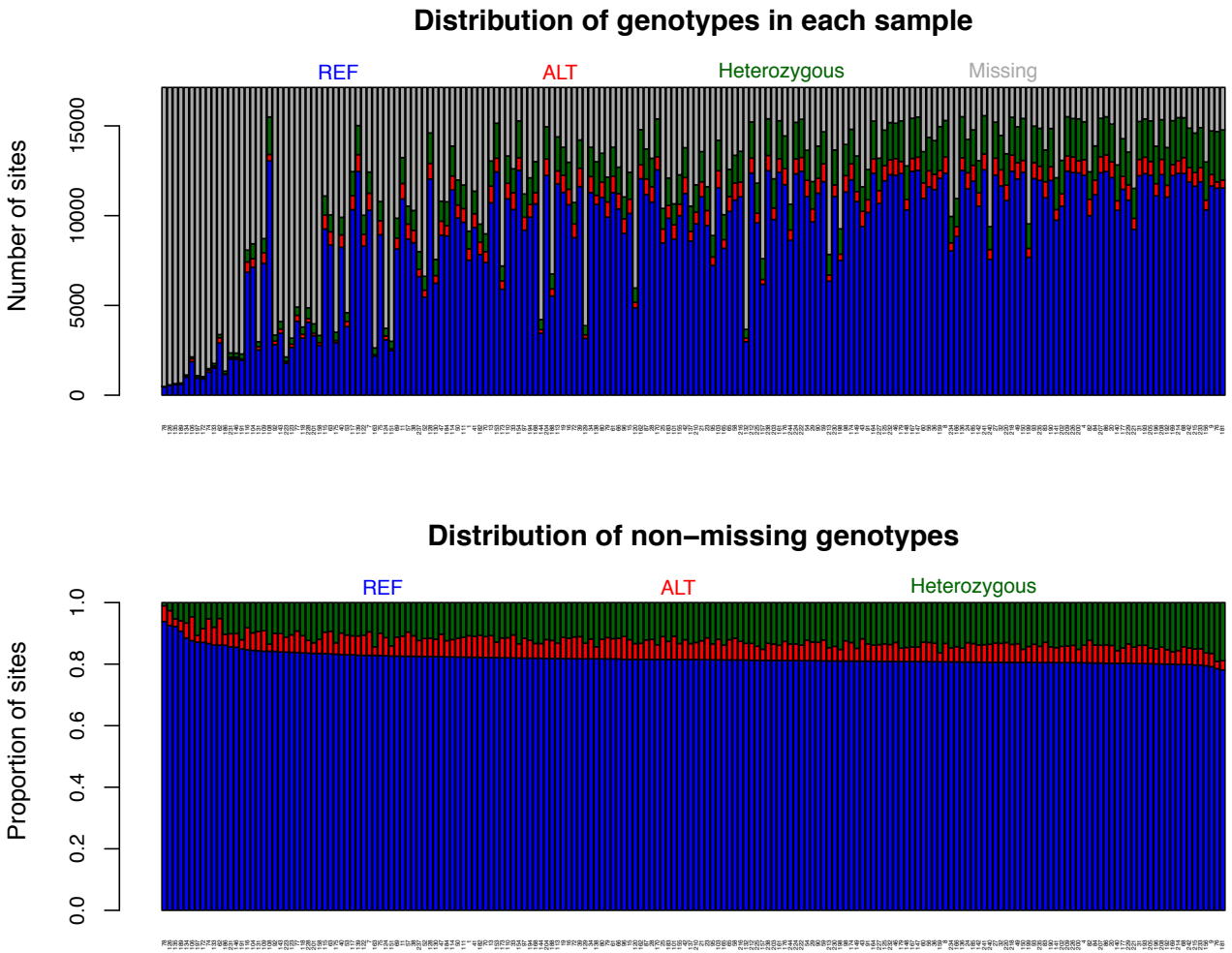
Filtering Criteria

- Missing data rate ≤ 50%
- Allele number = 2
- Genotype ≥ 2
- Minor Allele Frequency ≥ 2 / 192
- Heterozygosity range:
  - $(2 \text{ X } \text{Frequency}_{\text{Allele1}} \text{ X } \text{Frequency}_{\text{Allele2}}) \pm 0.2$

LMD50 SNP (N=17,143) Summary

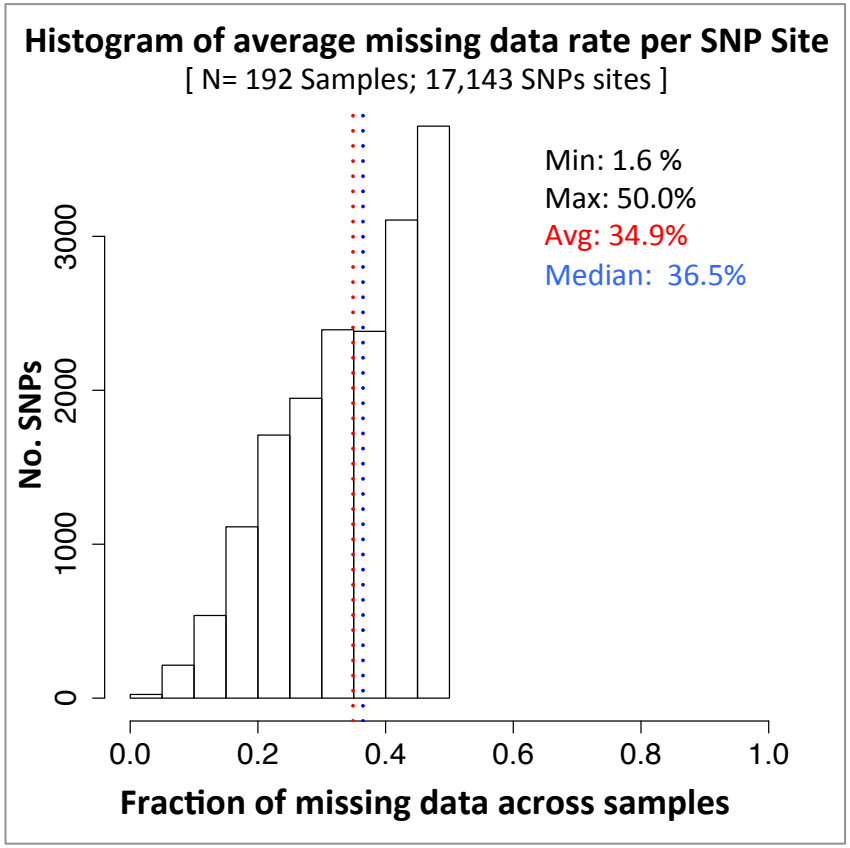


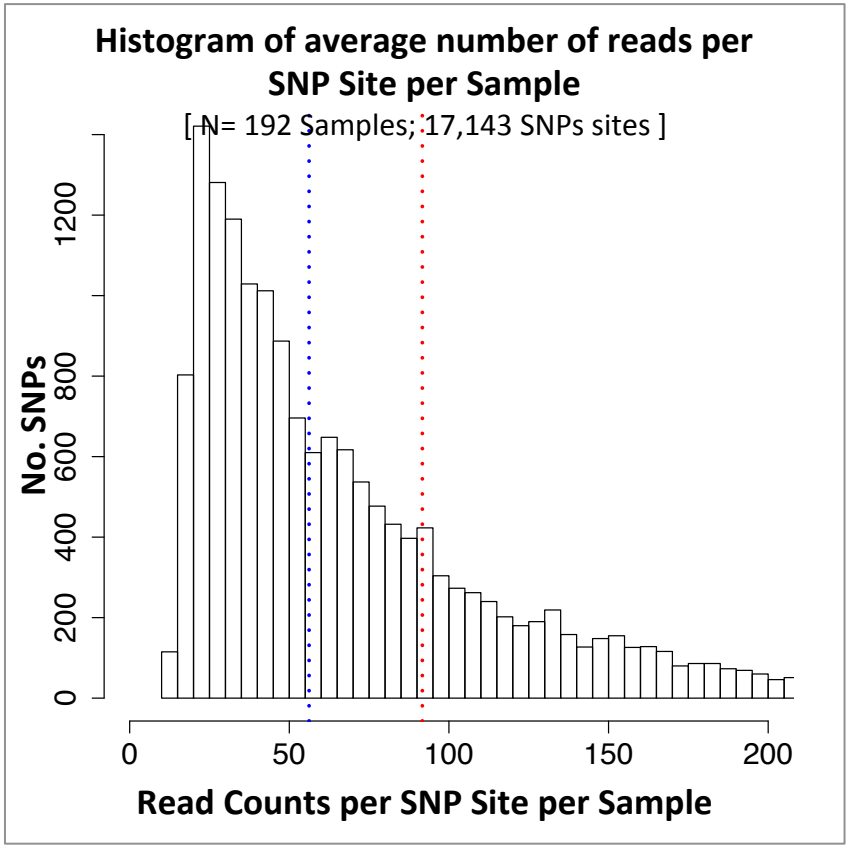
Sample (N=192) Summary for LMD50 SNPs



LMD50 SNPs : Average Missing Rate per SNP Site and read counts per SNP site per sample

(N= 192 Samples – 17,143 SNPs sites)





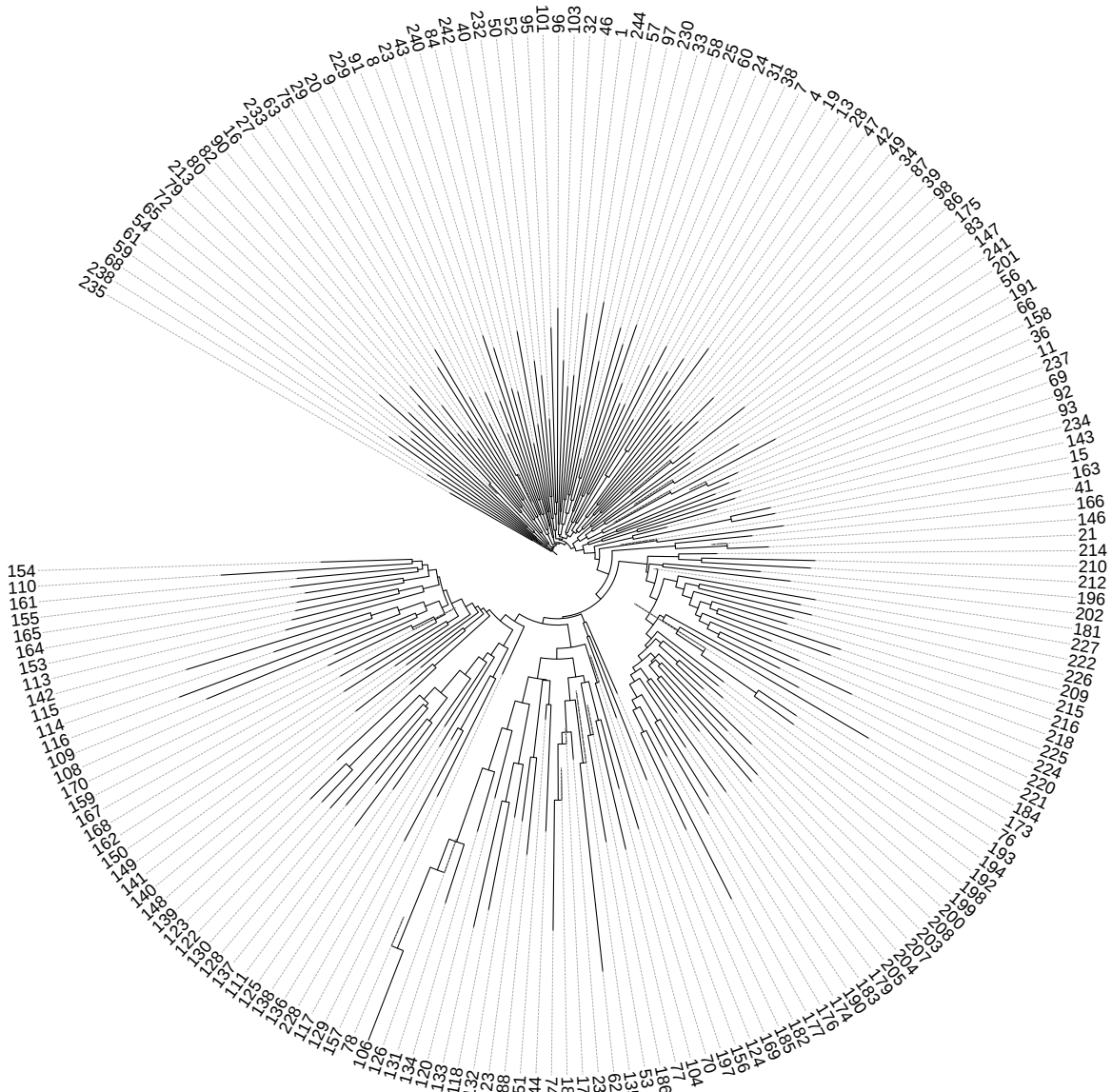
Overall Missing Data Points:

1,148,585 / 3,291,456 = **34.9%**

Min: 10 reads/SNP site/Sample*
Max: 32,706 reads/SNP site/Sample*
Avg: 92 reads/SNP site/Sample*
Median: 56 reads/SNP site/Sample*

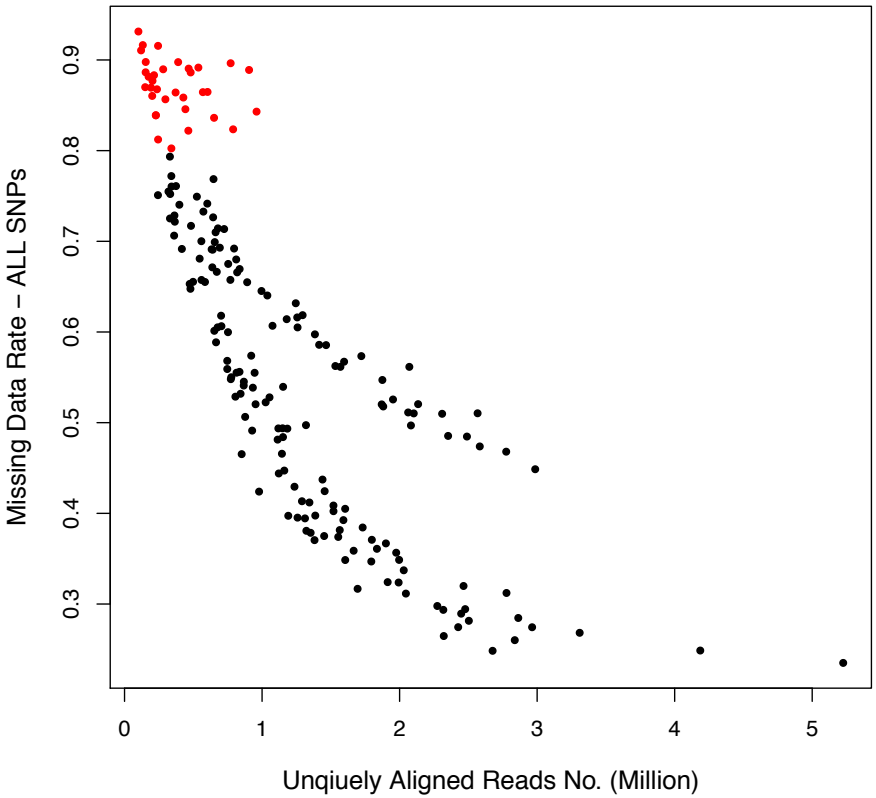
* Only samples with data were considered

Phylogenetic Tree using LMD50 SNPs

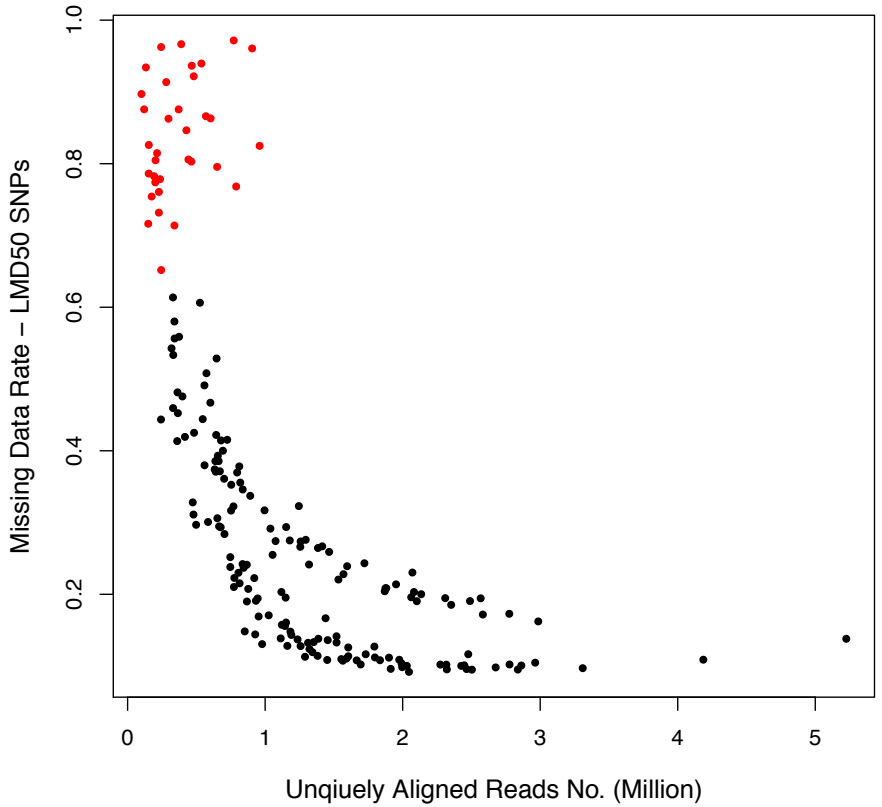


Uniquely Reads vs. Sample Missing Rate

ALL SNPs



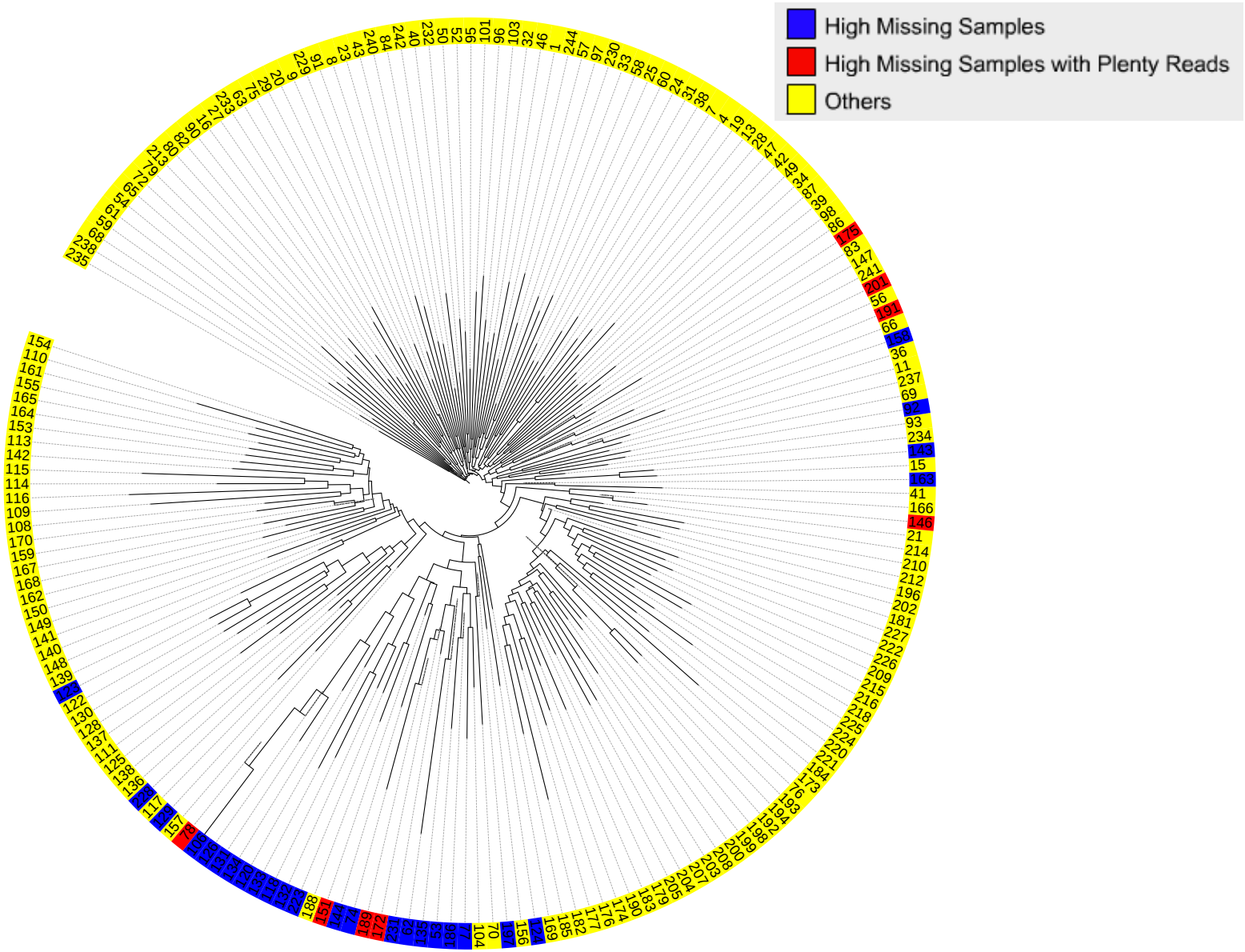
LMD50 SNPs



High missing samples (N=34)

DATA2BIO™

Phylogenetic Tree using LMD50 SNPs



DATA2BIO™