Methods

*Samples*

The sampling is focused on 12 sites that Dave is surveying as part of his thesis. These represent 3 elevational transects with four locations along each elevational transect. We also include 24 samples from other regions (Table 1 for summary of all locations surveyed and samples). The list of all DNA extractions that are currently housed in the new -80 (bottom shelf) is “/Users/Maggie/Dropbox/Dave Chris Maggie/ Good\_Samples\_19April2016.xlsx”

We decided to separate our sequencing efforts into batches of samples. Our first batch is comprised of 96 samples, 24 of which are samples from regions other than Dave’s survey (Table 2). 24 are from each of the 3 transects, 12 from 2014 and 12 from 2015. The sample submission form for Laval is located “/Users/Maggie/Dropbox/Dave Chris Maggie/Laval/ Rhinathus\_Set\_form\_sequence\_en.xls”

For each of the 12 sites, we will survey SNPs in 15 samples in each of 2 years (2014 and 2015) for a total of 360 samples. (Note: we have DNA samples for 361 samples of Dave’s survey).

15 samples per site for each of 12 sites in 2 years = 360 samples.

The 12 sites:

NkS: Nakiska South : High+ Mid-High+ Mid-Low +low

Kd: Kidd : High + Mid-High+ Mid-Low +Low

NkN: Nakiska North: High + Mid-High+ Mid-Low +Low

Samples are labelled as follows:

Year-Mountain-Site-plant-id

e.g., 2014-NkS-H-15-4 (Nakiska South at High site plant 15-4)

Table 1. Summary of broad sampling of *R. minor.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Region |  | Site\_ID | Year | Sample number |
| 1 |  | Dave’s survey | Kd | 2014 | 60 |
| 2 |  | Dave’s survey | Kd | 2015 | 61 |
| 3 |  | Dave’s survey | NkN | 2014 | 60 |
| 4 |  | Dave’s survey | NkN | 2015 | 60 |
| 5 |  | Dave’s survey | NkS | 2014 | 59 |
| 6 |  | Dave’s survey | NkS | 2015 | 61 |
| 7 |  |  | APK | 2015 | 2 |
| 8 |  |  | APR | 2015 | 2 |
| 9 |  |  | HB | 2012 | 4 |
| 10 |  |  | Kob | 2015 | 2 |
| 11 |  |  | Pho | 2015 | 2 |
| 12 |  |  | SP | 2015 | 6 |
| 13 |  |  | SS | 2015 | 6 |

Note: there were typos in the original dataset NkS, NKS, NKS, Nks, NkS were all present. I assumed it was the same site and standardized site name to capital letter, lowercase, capital.

We have also included 24 samples from sites at other regions that should be representative of the species broader genetic diversity (Table 2).

Table 2. Samples representing broader genetic diversity of *R. minor*. These samples were part of the first 96 that were sequenced.

|  |  |  |  |
| --- | --- | --- | --- |
| Region | Location | Elevation | Sample ID |
|  |  |  | 2015KobHi01 |
|  |  |  | 2015KobHi14 |
|  |  |  | 2015PhoMH20 |
|  |  |  | 2015PhoMH21 |
|  |  |  | 2015APRMd09 |
|  |  |  | 2015APKHi23 |
|  |  |  | 2015SSLo09 |
|  |  |  | 2015SSLo02 |
|  |  |  | 2015SSML06 |
|  |  |  | 2015SSMH22 |
|  |  |  | 2015SSHi13 |
|  |  |  | 2015SSHi19 |
|  |  |  | 2015SPLo16 |
|  |  |  | 2015SPLo15 |
|  |  |  | 2015SPMd03 |
|  |  |  | 2015SPMd23 |
|  |  |  | 2015SPHi04 |
|  |  |  | 2015SPHi17 |
|  |  |  | 2012HBLo01 |
|  |  |  | 2012HBLo02 |
|  |  |  | 2012HBMd02 |
|  |  |  | 2012HBMd01 |
|  |  |  | 2015APRMd20 |
|  |  |  | 2015APKHi08 |

*DNA extraction*

DNA was extracted from silica dried tissue using a Qiagen DNeasy plant mini kit with slight modifications to the protocol. The volume of the extraction buffer AP1 was increased to 500uL (and all other buffer volumes were increased accordingly). The initial spin at step 10 was increased to 7-9minutes. The final incubation period was increased to 7-10 min. We evaluated DNA quality based on agaraose gel visualization and the spectral properties of the sample using Nanodrop One/Onec UV-Vis spectrophotometer. Only samples with 260/280 ratios of 1.8 -1.85 and 260/230 ratios greater than 1.5 were used for library construction. DNA concentration was estimated using Qubit 3.0 flourometric quantitation. All samples were standardized to 20ng/uL +/2ng/uL.

*GBS library construction*

GBS library was constructed using double digest with Pstl and Mspl and was prepared by Brian Boyle at Laval University’s IBIS centre (Institut de Biologie Integrative et des Systemes). GBS relies on reducing genome complexity to ensure sufficient overlap in sequence coverage. In this protocol genome complexity is reduced by restriction enzymes (RE). Choosing appropriate REs repetitive regions can be avoided and lower copy regions can be targeted with two to three-fold efficiency. Barcodes are included in one of the adaptor sequences (Elshire et al 2011 *PlosOne*). The barcode is just upstream of the RE cut-site. GBS is similar to RAD but the barcode composition and length results in fewer sequence phasing errors (phase errors is the sequencing errors that usually appear more frequently towards ends of the reads -Sleep et al. 2013 *BMC Bioinformatics*).

The first GBS library comprised of 96 samples was sent to McGill by Brian Boyle around May 12, 2016. Brian sent one GBS library, which will be sequenced on 2 lanes of Hi Seq 100bp sequencing lane. Single-end genomic sequencing of GBS libraries was conducted at the Genome Quebec Innovation Centre at McGill University using the Illumina HiSeq 100bp platform.

*Do I need to trim off adapaters and RE sites in the reads?*

*Processing of reads from Elshire*

In the original Elshire et al (2013) publication they analyze qseq files “since the filtereing process that produces fastq files sometimes discards good reads that aligned perfectly well to the reference genome for at least 64 bases” (they sequence 86 bp).

1. Filtered reads that perfectly matches one of the barcodes and the expected four-base remnant of the APeKI cut site.
2. Were not adaptor/adaptor dimers (Adapter dimers are the result of self-ligation of the adapters without a library insert sequence).
3. contained no ‘‘Ns’’ in their first 72 bases

The reads are then sorted into separate files according to their barcode, the the barcode removed and remainder of the sequence trimmed to 64 bases including the initial CWCG (cut-site of the enzyme).

Then they made a genotype table with reads with a minimum Q of 10 across the first 72 bases and that occurred at least twice were kept. This forms the reference tags set. Sequence “tags” containing radon sequencing errors should not occur multiple times in multiple samples and should not map genetically so these should be filtered out. It seems they take the reference genome of maize and then *in silico* digest it and then this expected 64 bp tags from this are then “added” to the reference set –genotype table. Then they take their DNA samples reads and map these to the genotype table. Any tags that occurred in 10 or fewer DNA samples were removed. They argue this process should remove all sequencing errors.

Then they map this filtered set of reads to the Maize reference genome.