# Development and application of a bioinformatics pipeline for genotyping-by-sequencing (GBS) of autotetraploid potato

Schuyler D. Smith, Jeffrey B. Endelman

Department of Horticulture, University of Wisconsin, Madison, WI, USA

## INTRODUCTION

Genotyping-by-sequencing (GBS) is being widely used in diploid crops as an efficient technology for identifying genome-wide markers to assist breeding. Both genome-wide association studies and genomic selection have benefited from the availability of GBS markers (Morris et al. 2012; Poland et al. 2012). One of the challenges with GBS is how to handle missing data, but for diploids there are many accurate and accessible options for marker imputation (Rutkoski et al. 2013; Swarts et al. 2014).

For autotetraploid crops, such as potato (Solanum tuberosum), the cost effectiveness of GBS is less certain due to the higher read depths (50–60X) needed to accurately differentiate the three heterozygous genotypes. There is also less information published about imputation methods and accuracy. Our objectives were to (1) develop a bioinformatics pipeline for GBS variant discovery, genotype calling, and imputation in autotetraploids, (2) apply the pipeline to a panel of elite potato breeding lines and varieties from across North America, and (3) interpret the results in the context of a competing marker technology: the potato Infinium SNP array (Douches et al. 2014).

## MATERIALS AND METHODS

Ninety-one elite tetraploid russet potato lines were selected from the National Fry Processing Trial (NFPT) for GBS. Based on the results of an enzyme optimization study, the methylation-insensitive, 6 bp restriction enzyme EcoT22I was used for the DNA digestion. DNA fragments were barcoded and pooled into one library at the Cornell Biotech Center, following the protocol of Elshire et al. (2011). The library was sequenced on two lanes of an Illumina HiSeq2000, producing 100 bp single-end reads. A bioinformatics pipeline was developed to demultiplex (Cutadapt, Martin 2011) and align (BWA mem, Li 2010) reads to the potato reference genome (v4.03, Sharma et al. 2013), and then variants were called using the Genome Analysis Toolkit (GATK, McKenna, et al. 2010), without removing duplicates. BAM and VCF files were analyzed using bedtools and VCFtools, respectively.

Infinium SNP array data for 88 of the 91 lines from the GBS study, plus 73 additional NFPT lines (total N = 161), were used to test the accuracy of three imputation methods: k-Nearest Neighbors (kNN), Random Forest (RF, Breiman 2001; Liaw and Wiener 2002), and a polyploid Hidden Markov Model (HMM, Su et al. 2008). For all three methods, the imputed values were categorical rather than numeric. The SNP dataset contained 2277 markers with tetraploid genotype calls, minor allele frequency ≥ 0.1, and fraction of missing calls ≤ 0.05. For each marker tested, 30 randomly chosen samples were masked, imputed, and then compared to the true genotypes to calculate the percent error.

## RESULTS AND DISCUSSION

The number of SNPs and number of aligned base pairs were determined at progressively higher minimum depth (Figure 1). The ratio of these two quantities, which equals the genomic SNP density, declined rapidly as depth increased from one, reaching a quasi-constant value of 1 SNP per 42 bp above 100X per SNP. Using exome capture with a panel of 83 European tetraploid varieties, Uitdewilligen et al. (2013) reported intron and exon SNP densities of 1 SNP per 15 and 24 bp, respectively.

**Acknowledgments**

We thank the breeders involved in the National Fry Processing Trial (NFPT) for contributing germplasm, as well as the financial sponsors of the NFPT. Primary funding for this project was provided by USDA-NIFA-AFRI Grant No. 2014-67013-22418.

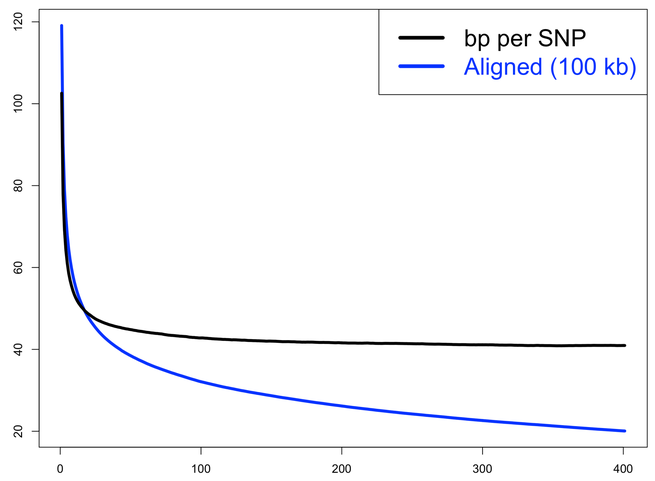


Figure 1. Estimating genomic SNP density. Above 100X across the population, the number of aligned base pairs per SNP is approximately constant at 42.

Table 1 displays the number and distribution of variants with at least 200 total reads. Of the 67K variants, nearly 63K were SNPs and 4K were indels. The vast majority of the SNPs were bi-allelic, with 930 tri-allelic and 9 tetra-allelic SNPs in the dataset. Genome-wide coverage was evenly distributed: the fewest variants were discovered on Chromosome 12, with this number still over 3000.

Table 1. Variant statistics and the distribution across chromosomes, with a minimum population depth of 200X.

| | Whole Genome | Chromosome | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Variants | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Bi-Allelic SNPs | 61,992 | 7566 | 4037 | 5459 | 7143 | 5654 | 4054 | 5812 | 5321 | 5956 | 3690 | 4448 | 2843 |
| Tri-Allelic | 930 | 150 | 59 | 81 | 136 | 86 | 51 | 106 | 82 | 98 | 34 | 6 | 41 |
| Tetra-Allelic | 9 | - | - | 3 | 1 | 1 | 1 | 1 | - | 1 | - | - | 1 |
| Total SNPS | 62,931 | 7716 | 4096 | 5546 | 7281 | 5742 | 4107 | 5920 | 5403 | 6056 | 3724 | 4454 | 2886 |
| Indels | 4,069 | 499 | 319 | 418 | 396 | 385 | 296 | 380 | 315 | 335 | 245 | 305 | 176 |
| Total | 67,000 | 8215 | 4415 | 5964 | 7677 | 6127 | 4403 | 6300 | 5718 | 6391 | 3969 | 4759 | 3062 |

The number of SNPs declined rapidly as the minimum average depth per sample was raised, from 14K SNPs with at least 20X per sample to only 3300 SNPs with at least 50X. These results, which correspond to an approximately 48-plex run, were compared against the results based on a single lane (i.e., 96-plex). With only one lane, around 600 variants were discovered with an average depth per sample of at least 50X (Figure 2). This means that reducing the multiplex level by a factor of two increased the number of useful variants for genotyping by a factor of six.

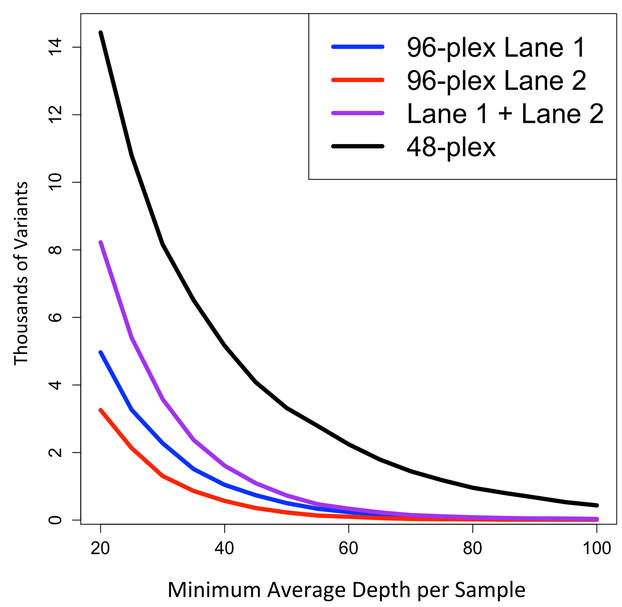


Figure 2. Effect of GBS multiplexing on variant discovery. The red and blue curves show the number of variants vs. depth when using a single Illumina lane (i.e., 96-plex). The purple curve is the sum of the red and blue curves, which should be compared with the combined analysis of both lanes (i.e., 48-plex) in black.

The accuracy of three marker imputation methods was compared using SNP array data: k-nearest neighbors (kNN), Random Forest (RF), and a polyploid HMM algorithm. The first two do not rely on a genetic or physical map, while the HMM algorithm does. For all three methods, the error rate increased with minor allele frequency, as expected (Figure 3). The best method proved to be RF (19% error on average), followed by HMM (46%), and then kNN (54%). We had expected the HMM method to outperform the other two because it exploits map information, but further research is needed to realize this goal.

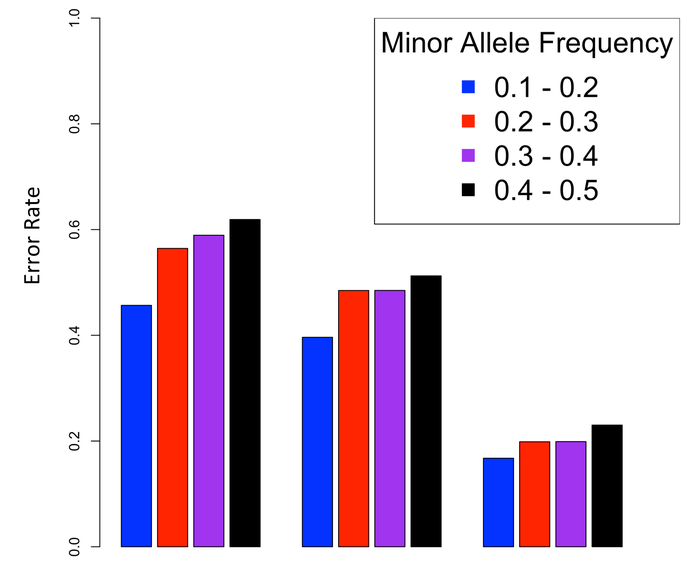


Figure 3. Comparison of the error rate for three marker imputation methods. (kNN = k-Nearest Neighbors, HMM = Hidden Markov Model, RF = Random Forest)

## CONCLUSIONS

GBS was originally developed for inbred lines, where only a single read is needed to infer the genotype of a sample (Elshire et al. 2011). For crops such as maize, barley, and rice, it is now practical to include 384 samples in one library (Spindel et al. 2013). For autotetraploid crops, such as potato, the optimal multiplex level is less clear. Lower multiplexing improves the number of variants with high read depth, which is needed to accurately differentiate between the three heterozygous states, but the additional cost per sample means fewer genotyped individuals at a fixed budget. From the binomial distribution, we expect a duplex call error rate of 10% at 50–60X. Uitdewilligen et al. (2013) suggested an even higher threshold of 60–80X. Even at 48-plex, only 3300 SNPs had an average depth per sample of at least 50X, which is less than the number of SNPs (≈5000) with high quality tetraploid calls on the potato Infinium array (Douches et al. 2014; Schmitz Carley et al. unpublished). Given that our GBS cost per sample at 48-plex is comparable to the array, at present there is little incentive to use GBS for routine use in the University of Wisconsin potato breeding program. This conclusion is predicated on our use of the EcoT22I enzyme and current sequencing technology. Improvements in either aspect, or the availability of a polyploid HMM for genotype calling (analogous to the diploid method of Fragoso et al. 2016), could make GBS more cost-effective for tetraploid potato.

### References

Breiman. 2001. Machine Learning. 45:5-32.
Douches et al. 2014. Potato Research.
Elshire et al. 2011. PLoS One. (6)5.
Fragoso et al. 2016. Genetics. (203)2:
Li et al. 2010. Bioinformatics. 26(5):589-595.
Liaw and Wiener. 2002. R News. 2(3):18-22.
Martin. 2011. EMBnet.journal. 17(1):10-12.
McKenna et al. 2010. Genome. Res. 20:1297-1303.
Morris et al. 2012. Nat. Ass. Sci. 110(2):453-458.
Poland et al. 2012. Plant Genome. 5(3):103-113.
Rutkoski et al. 2013. G3. (3)3:427-439.
Sharma et al. 2013. G3. 3(11):2031-2047.
Spindel et al. 2013. Theor. Appl. Genet. 126:2699-2716
Su et al. 2008. Bioinformatics. 24(7):972-978.
Swarts et al. 2014. The Plant Genome. (7)3:1-12.
Uitdewilligen et al. 2013. PLoS One. 8(5).