BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 4 2015, pages 599–601
doi:10.1093/bioinformatics/btu691

Sequence analysis Advance Access publication October 21, 2014

Simple, rapid and accurate genotyping-by-sequencing from aligned whole genomes with ArrayMaker

Cali E. Willet[1,*], Bianca Haase[1], Michael A. Charleston[2] and Claire M. Wade[1]

[1]Faculty of Veterinary Science and [2]School of Information Technologies, University of Sydney, Sydney, New South Wales 2006, Australia

Associate Editor: Janet Kelso

ABSTRACT

Summary: Whole-genome sequencing has revolutionized the study of genetics. Genotyping-by-sequencing is now a viable method of genotyping, yet the bioinformatics involved can be daunting if not prohibitive for some laboratories. Here we present ArrayMaker, a user-friendly tool that extracts accurate single nucleotide polymorphism genotypes at pre-defined loci from whole-genome alignments and presents them in a standard genotyping format compatible with association analysis software and datasets genotyped on commercial array platforms. Using this tool, geneticists with only basic computing ability can genotype samples at any desired list of markers, facilitating genome-wide association analysis, fine mapping, candidate variant assessment, data sharing and compatibility of data sourced from multiple technologies.

Availability and implementation: ArrayMaker is licensed under The MIT License and can be freely obtained at https://github.com/cw2014/ArrayMaker/. The program is implemented in Perl and runs on Linux operating systems.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: cali.willet@sydney.edu.au

Received on March 6, 2014; revised on September 25, 2014; accepted on October 16, 2014

1 INTRODUCTION

The rapid evolution of whole-genome sequencing (WGS) technologies and decline in sequencing costs in recent years has seen an increase in utilization of WGS in genetic experiments. Wide ranging research questions can be asked of such datasets, but for identifying genes underlying phenotypes of interest, genome wide association analysis (GWAS) remains a crucial first step in localizing candidate regions for sequence investigation. Commercially available single nucleotide polymorphism (SNP) arrays have been dedicated to this task, genotyping samples at carefully selected informative markers. From WGS data, genotyping-by-sequencing (GBS) enables us to create array-type data, yet a simple tool to extract SNP genotypes at informative sites into array format is currently lacking. In light of this, we developed ArrayMaker tool to extract SNP genotypes from mammalian WGS alignments at pre-defined loci with the resultant output in a GWAS-ready format.

Producing array-type files at known informative markers enables GWAS to be carried out on WGS datasets. It also offers an interface for compatibility between WGS data and samples that have been genotyped on commercial SNP array platforms. ArrayMaker can also genotype samples at bespoke lists of SNP markers, facilitating fine mapping and providing a means of assaying candidate loci across multiple samples without the need for extensive post-alignment steps. The sequencing platform and alignment protocols used do not preclude compatibility of SNP genotypes called from various datasets, enabling a platform-independent common language amongst samples sourced from diverse studies.

2 METHODS

ArrayMaker operates on alignments in binary alignment/map (BAM) format, and utilizes SAMtools (Li *et al*., 2009) to create a pileup of sequence information at each locus in a user-defined list. The SAMtools SNP calling algorithm was recently shown to have favorable accuracy when compared with other popular callers (Cheng *et al*., 2014). The resultant variant call format file is converted into PLINK (Purcell *et al*., 2007) transposed PED (tped) format, with genotypes possible in the top or forward calling orientation. Genotypes can be extracted for a single sample or many samples simultaneously. ArrayMaker can be applied to alignments from any species for which a reference genome exists.

3 RESULTS

3.1 Genotyping rate

ArrayMaker performance was tested on equine and canine Illumina HiSeq2000 datasets (Table 1). Genotyping rate (GR) was influenced by depth of sequence coverage (X) and experiment-specific run conditions. The highest GR of 98.32% was obtained from a sample sequenced to 14X. A 7X sample showed a GR of 90.2%, while a sample from a previous run on the same machine demonstrated 79.6% GR despite greater than 12X. These results demonstrate the impact of sequence quality on genotyping potential.

3.2 Concordance with commercial genotyping arrays

To perform high-throughput validation of ArrayMaker SNP calls, genotypes from Illumina SNP arrays for the same samples that were sequenced were used as the ‘truth-dataset’ (Table 1). Mean concordance ranged between 98.48 and 99.46%.

*To whom correspondence should be addressed.

Table 1. Genotyping rates, array concordance rates and error structure of three datasets genotyped using ArrayMaker

Run	Array GR (%)[a]	Str.[b]	Ch.[c]	GR (%)[d]	Cc (%)[e]	Mis. (%)[f]	Alt. hom. (%)[g]	Het. over. (%)[h]	Het. under. (%)[i]
1 ($n = 2$, $X = 13.5$)	99.16	H	Y	97.11	99.24	0	0.07	0.21	0.49
		H	N	97.16	99.19	0.03	0.08	0.22	0.49
		L	Y	97.69	99.28	0	0.09	0.28	0.35
		L	N	97.73	99.24	0.02	0.10	0.29	0.35
2 ($n = 2$, $X = 12.2$)	97.67	H	Y	92.54	98.50	0	0.05	0.11	1.32
		H	N	92.57	98.48	0.01	0.05	0.12	1.32
		L	Y	93.18	98.90	0	0.05	0.15	0.90
		L	N	93.21	98.87	0.01	0.06	0.16	0.89
3 ($n = 4$, $X = 15.1$)	98.27	H	Y	89.30	99.41	0	0.05	0.06	0.48
		H	N	89.32	99.39	0.01	0.05	0.06	0.48
		L	Y	89.79	99.46	0	0.05	0.09	0.38
		L	N	89.81	99.44	0.01	0.06	0.10	0.38

[a]Illumina array genotyping rate (GR).
[b]ArrayMaker stringency, H = high, L = low.
[c]Genotypes checked against expected alleles, Y = yes, N = no.
[d]ArrayMaker GR.
[e]Mean concordance of ArrayMaker with Illumina array calls, as a total of all markers where a missing call was made by neither platform.
[f–i]Mean percentage of discordant call type, expressed as a total of all markers where a missing call was made by neither platform. [f]Mismatch, more than two alleles observed across both platforms.[g]Alternate homozygote, both platforms called a homozygote for a different allele. [h]Heterozygote overcall, Illumina array indicated homozygote and ArrayMaker called a heterozygote. [i]Heterozygote undercall, opposite of heterozygote overcall.

ArrayMaker has four genotyping options: high or low stringency calling, with genotypes checked or not checked against expected alleles. Low stringency includes reads not mapped in a proper pair. High stringency applies an extended base alignment quality calculation, downgrades mapping quality for excessive mismatches and excludes loci significantly affected by strand bias. Failure of a checked ArrayMaker genotype to concord with expected alleles after allowing for strandedness (mismatch error, ‘Mis.’ Table 1) downgrades a call to missing status. The greatest concordance rate was observed when low stringency was applied and genotypes were checked against expected alleles (Table 1).

The predominant GBS error type was undercalls of heterozygotes (false negatives; true genotype is polymorphic yet the call is homozygous reference or ‘missing’) (Table 1). Low stringency genotyping resulted in a lower proportion of heterozygote undercalls and an increase in apparently false heterozygotes (false positives; true genotype is homozygous reference yet the call is polymorphic) compared to high stringency. Of the increased number of calls made using low stringency parameters, 86.3% of these were concordant with the Illumina array call.

To further manipulate calling stringency, users can modify the minimum and maximum sequence coverage required at a locus to call a SNP. Decreasing the minimum coverage decreases the false negative rate while increasing the false-positive rate through undersampling of alleles. Increasing the minimum coverage ameliorates this undersampling and thus improves the false positive rate, but also increases the false-negative rate (data not shown). These results demonstrate how the choice of stringency and genotype checking should be selected according to what type of error structure is preferable, given the nature of the experiment.

3.3 Run time and resources used

Number of samples and depth of sequence coverage had a greater influence on run time than did total makers assayed (Supplementary Fig. S1). Eight minutes (min) and less than 120 MB of RAM from a single processor core at 3.07 GHz was required to genotype two samples at 6.5X at 5000 markers, compared to 28.5 min for the same number of markers and samples at 16X, or 56.5 min for 12 samples at 6.5X.

4 CONCLUSION

ArrayMaker provides a straightforward method of quickly extracting accurate SNP genotypes from BAM files at any combination of SNP loci. The tool can replicate existing and obsolete SNP array datasets, create GWAS-ready files from bespoke marker lists, and rapidly genotype candidate SNP loci in new samples without the need for complicated post-alignment bioinformatics. It requires minimal computing time and resources and is user-friendly, written in Perl to run on a Linux platform with simple command-line operation. Only an alignment, a reference genome sequence, a list of SNP markers and SAMtools are required. Additional steps such as formatting the marker list, including the conversion of legacy array datasets into current reference genome coordinates or the ascertainment of dataset-polymorphic SNPs for fine mapping or candidate variant analysis, may be required prior to the application of ArrayMaker using alternate tools such as LiftOver or SAMtools, respectively. Following its use, ArrayMaker data can be readily interrogated with PLINK (Purcell *et al.*, 2007) for the determination of population allele frequencies, GWAS and other useful analyses.

Given the shift towards WGS in genetic research projects, GBS is increasing in popularity and there are many existing tools that return accurate genotypes from whole-genome sequence data. However, calling and filtering variants can be a daunting computational task for those inexperienced in bioinformatics. ArrayMaker simplifies the task of extracting SNP genotypes at prescribed loci by packaging a popular calling algorithm with a light resource footprint into a user-friendly tool that emits genotypes in a human-readable, GWAS-ready format.

ACKNOWLEDGEMENTS

We are grateful for donation of data by Tosso Leeb, Stefan Rieder; Danika Bannasch, Zena Wolf (NIH NIDCR R01DE022532); Jonathan Early, Elizabeth Arnott, Paul McGreevy (MLA RIRDC PRJ-007806); Tracey Chew; Annie Pan, Peter Williamson; TGEN.

Funding: This research was funded by APA and Val Street scholarships. Donated data was funded in part by NIH and MLA (see Acknowledgements).

Conflicts of interest: none declared.

REFERENCES

Cheng,A.Y. *et al*. (2014) Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics*, **30**, 1707–1713.

Li,H. *et al*. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Purcell,S. *et al*. (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet*., **81**, 559–575.