GENETIC CHARACTERIZATION OF BUCKWHEAT ACCESSIONS THROUGH GENOME-WIDE ALLELE FREQUENCY FINGERPRINTS

GENETSKA KARAKTERIZACIJA VZORCEV AJDE Z ODTISI FREKVENCE ALELOV V GENOMU

Michelle M. NAY[1], Stephen L. BYRNE[2], Eduardo A. PÉREZ[3], Achim WALTER[3], Bruno STUDER[1]

http://dx.doi.org/10.3986/fbg0063

ABSTRACT

Genetic characterization of buckwheat accessions through genome-wide allele frequency fingerprints

Genomics-assisted breeding of buckwheat (*Fagopyrum esculentum* Moench) depends on robust genotyping methods. Genotyping by sequencing (GBS) has evolved as a flexible and cost-effective technique frequently used in plant breeding. Several GBS pipelines are available to genetically characterize single genotypes but these are not able to represent the genetic diversity of buckwheat accessions that are maintained as genetically heterogeneous, open-pollinating populations. Here we report the development of a GBS pipeline which, rather than reporting the state of bi-allelic single nucleotide polymorphisms (SNPs), resolves allele frequencies within populations on a genome-wide scale. These genome-wide allele frequency fingerprints (GWAFFs) from 100 pooled individual plants per accession were found to be highly reproducible and revealed the genetic similarity of 20 different buckwheat accessions analysed in our study. The GWAFFs cannot only be used as an efficient tool to precisely describe buckwheat breeding material, they also offer new opportunities to investigate the genetic diversity between different buckwheat accessions and establish variant databases for key material. Furthermore, GWAFFs provide the opportunity to associate allele frequencies to phenotypic traits and quality parameters that are most reliably described on population level. This is the key to practically implement powerful genomics-assisted breeding concepts such as marker-assisted selection and genomic selection in future breeding schemes of allogamous buckwheat.

Key words: Buckwheat (*Fagopyrum esculentum* Moench), genotyping by sequencing (GBS), population genomics, genome-wide allele frequency fingerprints (GWAFFs)

IZVLEČEK

Genetska karakterizacija vzorcev ajde z odtisi frekvence alelov v genomu

Genomsko podprto žlahtnjenje ajde (*Fagopyrum esculentum* Moench) je odvisno od robustnih metod genotipiziranja. Genotipiziranje s spremljanjem sekvenc (genotyping by sequencing, GBS) se je razvilo kot fleksibilna in razmeroma poceni metoda, ki se jo uporablja pri žlahtnjenju rastlin. Uporabnih je več virov GBS za genetsko karakterizacijo posamičnih genotipov, toda te metode niso primerne za predstavitev genetske raznolikosti vzorcev ajde, ki jih vzdržujemo v heterozigotni obliki, kar velja za odprto oplodne populacije. Tu poročamo o razvoju GBS metode, ki, namesto prikazovanja bi-alelnega polimorfizma posameznih nukleotidov (single nucleotide polymorphisms, SNPs), pokaže frekvence alelov v populaciji na nivoju genoma. Ta prikaz frekvence alelov na nivoju genoma (genome-wide allele frequency fingerprints, GWAFFs) z združenimi sto posameznimi rastlinami vsakega vzorca se je pokazal kot visoko ponovljiv in je prikazal genetsko podobnost 20 različnih vzorcev ajde, ki smo jih analizirali v naši raziskavi. Metoda GWAFFs ni uporabna samo kot učinkovito orodje za natančen opis materiala za žlahtnjenje ajde, ponuja tudi možnosti raziskave genetskih razlik med različnimi vzorci ajde in omogoča zbirke podatkov. Nadalje, metoda GWAFFs omogoča povezovanje frekvenc alelov s fenotipskimi lastnostmi in kvalitativnih parametrov, ki so najbolj zanesljivo opisani na nivoju populacij. To je ključ za praktično uporabo z genomiko podprtega žlahtnjenja, kot je z genskimi markerji podprta selekcija in genomska selekcija z GWAFFs.

Ključne besede: ajda (*Fagopyrum esculentum* Moench), genotipizacija s sekvenciranjem (GBS), populacijska genomika, GWAFFs

[1] Molecular Plant Breeding, Institute of Agricultural Sciences, ETH Zurich, Universitaetstrasse 2, 8092 Zurich, Switzerland, Bruno.studer@usys.ethz.ch

[2] Teagasc Crop Science Department, Oak Park, Carlow, R93XE12, Ireland

[3] Crop Science, Institute of Agricultural Sciences, ETH Zurich, Universitaetstrasse 2, 8092 Zurich, Switzerland

1 INTRODUCTION

Plant breeding plays a key role in order to meet the increasing demand for food of the world's growing population. Sophisticated breeding strategies have been developed depending on the reproductive strategy of crop species, but they are all based on one common principle: Through initial crossings, genes and alleles are reshuffled and among the hundreds or thousands of variants produced in the cross, variants outperforming the parents or combining desirable characteristics may be chosen (Fehr 1987; Stoskopf et al. 1999). To select the best variants, plant breeding depended for centuries on the trained breeder's eye. With the establishment of genomics-assisted plant breeding techniques, breeders were given an aid to support the selection process. Genomics-assisted plant breeding incorporates knowledge of the genetic determinant of the trait of interest and allows to select superior variants based on genetic data (Moose & Mumm 2008). This technique allows to select for multiple traits and usually within a fraction of time needed to measure them. Through the steep price drop for next-generation sequencing in the last decades (Wetterstrand 2019), genomics-assisted breeding became not only feasible for major crops but is becoming increasingly popular in orphan crops.

The orphan crop buckwheat (*Fagopyrum esculentum* Moench) is a desired food crop since its gluten-free grains are high in antioxidants and essential amino acids (Li & Zhang 2001). Agronomically, buckwheat has shown beneficial effects in crop rotations and is an attractive bee crop (Falquet et et al. 2015; Teboh & Franzen 2011), but its low seed-set, indeterminate growth and susceptibility to abiotic stresses have hindered wide adoption of buckwheat as a cash-crop in Europe (Lichtenhahn & Dierauer 2000). Buckwheat breeding is conducted in several programs around the world, but is complicated by the heterostylous self-incompatibility system (Ueno et al. 2016). Several attempts have been made to transfer the self-compatibility of its sister species *Fagopyrum homotropicum* Ohnishi (Matsui et al. 2003), while this was successful, the resulting lines often suffered from inbreeding depression (Campbell 1997). Hence, most buckwheat grown today is of outcrossing nature and accessions are maintained as diverse populations. This renders it difficult to fix beneficial alleles in the population, because the 'superior' plants selected at harvest time are already pollinated by the 'inferior' neighbour plants.

Genomics-assisted breeding offers opportunities to select plants containing beneficial alleles based on genetic data and known marker-trait associations. A requirement for this are reliable genotypic data of breeding germplasm. For buckwheat with a haploid genome size of around 1.3 Gb (2n=16), several genomic resources have become available recently (Logacheva et al. 2011; Mizuno & Yasui 2019; Nagano et al. 2000; Yasui et al. 2016). A widely used genotyping method that has evolved as highly flexible is genotyping by sequencing (GBS) (Elshire et al. 2011). In the GBS workflow, genomic DNA is cut by a restriction enzyme and sequencing adaptors are ligated to the cutting sites. After a PCR multiplication step, the fragments are short-read sequenced, which results in repeated coverage of thousands of genetic loci (Elshire et al. 2011). Through alignment of the sequencing reads to a reference genome, a genotyping matrix can be derived that allows for further downstream analysis to compare genotypes or conduct genetic studies. Since buckwheat accessions are populations of genetically distinct individuals, standard genotyping and variant calling pipelines tailored for genotyping single individuals or inbred lines (Li et al. 2009; Mckenna et al. 2010), are of limited use. As an alternative, a large number of single plants can be genotyped and instead of determining the allele present at a certain genetic location, allele frequencies can be calculated. A shortcut and budget-friendly option represents the pooling of multiple individuals before sequencing, which proved to be a highly accurate method in perennial ryegrass (*Lolium perenne*) (Byrne et al. 2013).

The main objective of this study was to find a reliable genotyping method for detailed genetic analyses of buckwheat accessions. Specifically, we adapted the GBS and analysis protocol reported by Byrne et al. (2013) to calculate genome-wide allele frequency fingerprints (GWAFFs) and tested their accuracy in a replicated set of twenty diverse accessions.

2 MATERIAL AND METHODS

2.1 Plant material and DNA extraction

Twenty accessions from Austria (AT), France (FR), Germany (DE), Russia (RU), Slovenia (SI), Ukraine (UA), Czech Republic (CZ) and Switzerland (CH) were grown with a sowing density of 180 seeds/m2 in field plots of 3 x 4m at the ETH Research Station for Plant Sciences in Lindau-Eschikon, Switzerland (47.449N, 8.682E, 520 m a.s.l.). The following accessions were used: Bamby (AT), Billy (AT), Buchsa (CH), Carolin (FR), Carte Noir (FR), Darja (SI), Devyatka (RU), Dialog (RU), Dikul (RU), Drollet (F), Kerntner Hadn (AT), La Harpe (F), Lileja (UA), MinI (DE), Orphe (F), Pyra (CZ), Rosa (CH), Temp (RU), Theophani (DE), Tussi (DE).

DNA was extracted from leaf material cut out with an apple corer to ensure tissues are of approximately the same size. For each accession, pooled samples of 100 randomly selected plants were taken in triplicate. The plant material was flash frozen in liquid nitrogen and milled using mortar and pistil. DNA was extracted using the DNeasy Plant Mini Kit (Quiagen, Hilden, DE) according to the manufacturer recommendation.

2.2 GBS library preparation and DNA sequencing

GBS library was prepared with the restriction enzyme combination PstI and ApeKI at LGC (LGC Ltd, Teddington, UK) according to their in house protocol (Arvidsson et al. 2016). The libraries were 150bp single-end sequenced on an illumina HiSeq (Illumina, Inc., San Diego CA, USA) machine at a depth of approximately 400 million reads for 60 samples.

2.3 Genome-wide allele frequency fingerprints

Demultiplexed fastq files of the GBS data were used for the analysis. The reads were mapped to the available genome assembly FES_r1.0 (Yasui et al. 2016) using BWA (Li & Durbin 2010). A single nucleotide polymorphism (SNP) database was developed by combining read data of the three sample replicates of each accession. At each site where the minimum read depth (RD) of 30 was achieved, the allele frequency of the variant allele was calculated for each cultivar and sites where more than 25 percent of samples had missing data (RD < 30) were removed. The average variant allele frequency at each site was determined [frequency of variant allele / (frequency of reference allele + frequency of variant allele)] and used to filter out sites where the average minor allele frequency was less than 0.01. Allele frequencies were also determined for each sample replicate and a reduced SNP database was generated that included only variant sites with a minimum read depth of at least 30 in all samples. This dataset was used to analyse the similarity of replicates and accessions with R (R CORE TEAM 2008) using the libraries 'psych' and 'pheatmap'. To calculate homozygosity and similarity between accessions, the mean GWAFFs over the three replicates was calculated and loci were considered homozygous if the allele frequency was larger than 0.975 or lower than 0.025.

3 RESULTS

3.1 Allele frequency calling and distribution

Genotyping by sequencing of the 20 buckwheat accessions in triplicate resulted in 3.5-9.5 million reads per pooled sample. Mapping them to the available draft genome (Yasui et al. 2016) resulted in a database containing 40,696 SNPs after filtering. Calculation of the allele frequencies for each sample separately revealed 15,726 high-quality loci after filtering. These loci were distributed on 3363 out of the 387,594 scaffolds reported in the draft genome sequence.

3.2 Reproducibility of GWAFFs

Replicated sampling of the populations resulted in highly comparable GWAFFs within the replicates (Figure 1). The Pearson correlation between replicates of the same accession ranged between 0.971 and 0.999, while between pooled samples of different accessions it ranged between 0.320 and 0.983. All but the accessions Tussi, Theophanu and MinI showed an allele frequency distribution skewed towards the right, indicating that the alternative alleles were present at a low frequency. For Tussi, Theophanu and MinI the allele frequencies were distributed around the extremes (1 or 0,

data for Tussi shown in figure 1), implying that the accessions were highly homozygous, which is a consequence of their self-pollinating reproduction system transferred from *F. homotropicum (F.J. Zeller, personal communication).*

3.3 Homozygosity within buckwheat accessions

The buckwheat accessions showed little homozygosity, with the exception of the self-compatible acces-

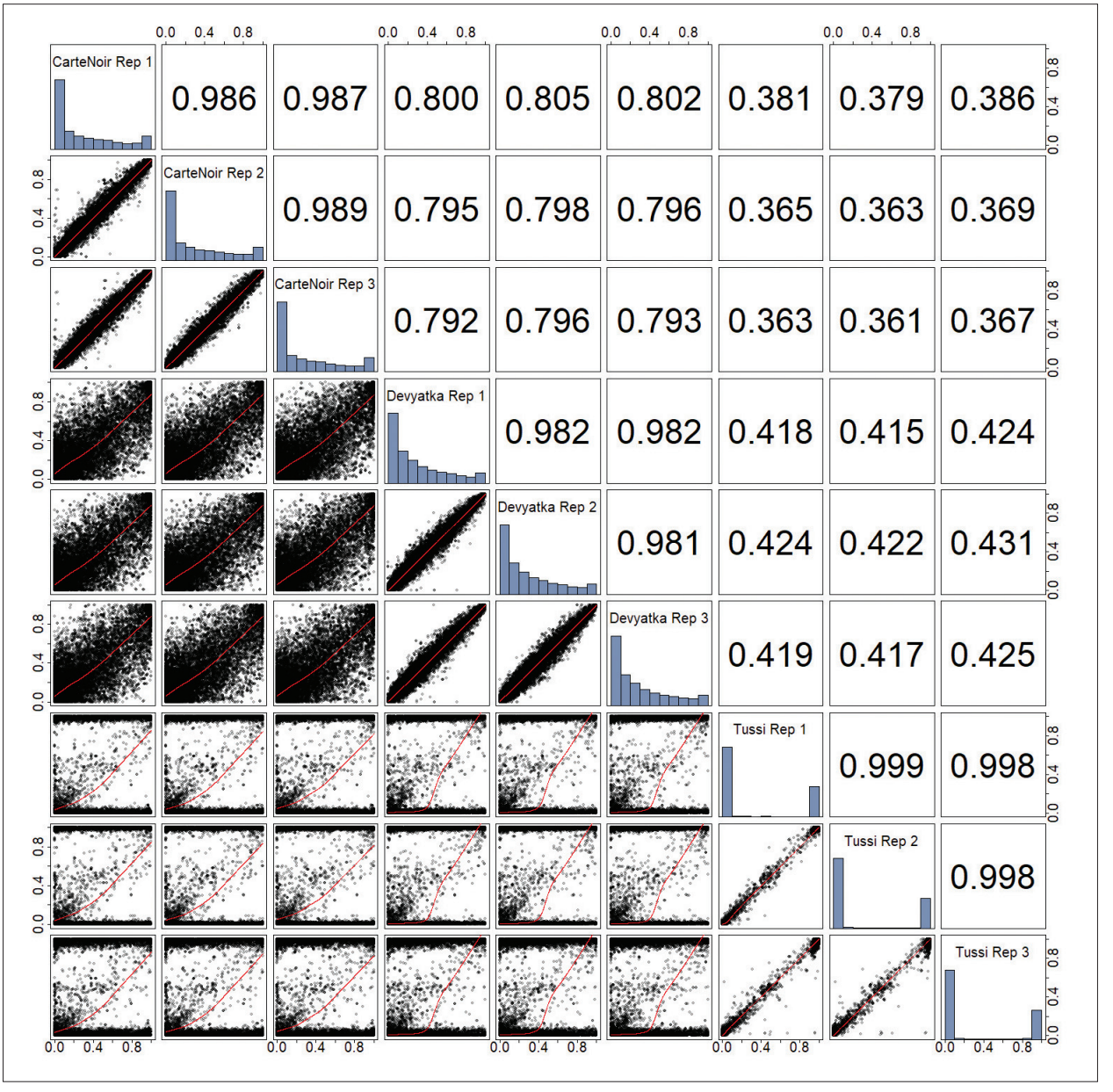

Figure 1: Correlation matrix of genome-wide allele frequency fingerprints within and between the replicated pooled samples of the accessions Carte Noir, Devyatka and Tussi. In the upper diagonal, Pearson correlations between the samples are shown. In the diagonal, histograms of the allele frequency distribution for each sample are shown. In the lower diagonal, the allele frequencies of the two samples are plotted against each other with the red line representing the LOESS (locally estimated scatterplot smoothing) line.
Slika 1: Korelacijska matrica frekvenc odtisov v genomu v in med združenimi vzorci akcesij Carte Noir, Devyatka in Tussi. V zgornji diagonali so prikazane Pearsonove korelacije med vzorci. V diagonali so prikazani histogrami razporeditve frekvenc alelov za vsak vzorec. Pod diagonal so prikazane frekvence alelov dveh vzorcev, rdeča črta označuje LOESS (locally estimated scatterplot smoothing) linijo.

Table 1: Homozygosity rate of 20 buckwheat accessions based on genotyping by sequencing data of 15,726 genetic loci. Genetic loci were regarded as homozygous, if the allele frequency was higher than 97.5% or lower than 2.5%. Razpredelnica 1: Stopnje homozigotnosti 20 vzorcev ajde, zasnovane na genotipizaciji s pomočjo sekvenciranja 15.726 genetskih lokusov. Lokusi so bili upoštevani kot homozigotni, če je bila pogostnost alela višja od 97,5 % ali nižja od 2,5 %.

Accession	Homozygosity [%]
Bamby	22.7
Billy	13.4
Buchsa	17.7
Carolin	17.7
Carte Noir	33.3
Darja	12.0
Devyatka	19.4
Dialog	23.4
Dikul	18.5
Drollet	23.5

Accession	Homozygosity [%]
Kerntner Hadn	22.3
La Harpe	29.7
Lileja	15.4
MinI	91.8
Orphe	15.5
Pyra	15.4
Rosa	16.6
Temp	25.9
Theophanu	79.0
Tussi	85.7

sions Tussi, Theophanu and Min1 (Table 1). The range of heterozygosity excluding the self-compatible lines ranged from 66.7% (Carte Noir) to 88.0% (Darja) with a mean value of 80.3%.

3.4 Genetic similarity of accessions

The genetic similarity of the accessions was analysed based on a correlation analysis of the mean GWAFFs

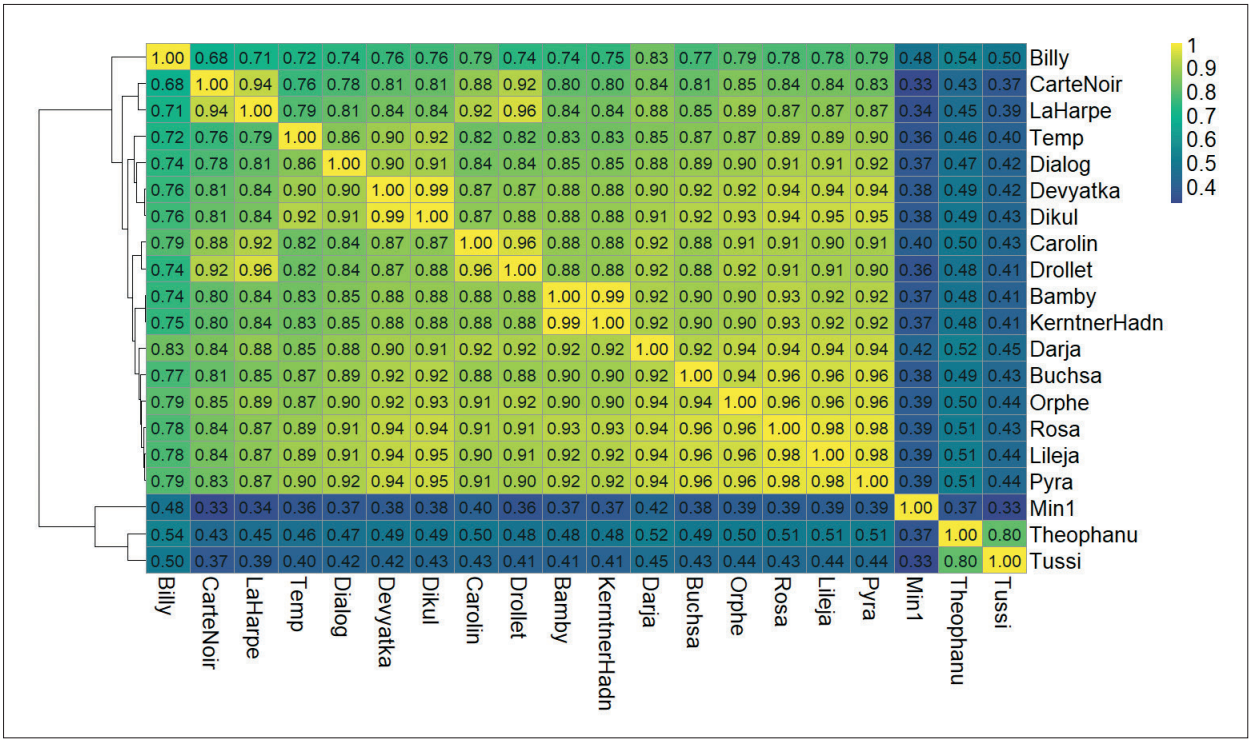

Figure 2. Correlation analysis of genome-wide allele frequency fingerprints of 20 buckwheat accessions. Pairwise Pearson correlations were calculated using the mean allele frequency of the three replicates per accession. On the left side, a hierarchical clustering analysis of the cultivars based on the correlation matrix is shown.
Slika 2. Korelacijska analiza odtisov alelnih frekvenc po celotnem genomu za 20 akcesij ajde. Pearsonove korelacije so izračunane po parih z uporabo srednjih frekvenc alelov treh ponovitev na vsako akcesijo. Na levi strain je prikazano hierarhično združevanje podatkov analiz kultivarjev, zasnovano na korelacijski matrici.

of the 20 accessions (Figure 2). The homozygous, self-compatible accessions Tussi, Theophanu and MinI clustered separately and were distinct from the remaining accessions with Pearson correlations of 0.33-0.54. Within the heterozygous, self-incompatible accessions, several were found to be highly similar, often clustering by the country of origin. A high genetic similarity was revealed between the Central European accessions Bamby, Kerntner Hadn, Darja, Buchsa, Orphe, Rosa, Lileja and Pyra, the Russian accessions Dikul, Devyatka, Dialog and Temp, and the French accessions Carte Noir, La Harpe, Carolin and Drollet (Figure 2).

4. DISCUSSION

4.1 Genome-wide allele frequency fingerprints allow precise genotyping of buckwheat accessions

In this study, we have shown that by pooling 100 individual plants of a buckwheat accession and subjecting them to GBS, representative and highly reproducible GWAFFs can be obtained. This allowed us to genetically characterize 20 buckwheat accessions at unprecedented precision. We identified genetically similar accessions and found that they often cluster by the region of their origin. The genetic similarity of accessions bred in the same country or region may be an indication of the narrow genetic base in each country with limited gene-flow between breeding programs. Analysis of further accessions would yield a better understanding of the buckwheat genetic resources worldwide and may allow to set exchange and conservation priorities.

4.2 Importance of robust genotyping method to implement genomics-assisted breeding in buckwheat

Accurate genotypic data, such as the GWAFFs presented in this work, are crucial to describe buckwheat breeding materials and investigate the genetic diversity between different buckwheat accessions. Furthermore, they enable to associate allele frequencies to plant phenotypic traits and nutrition quality parameters that are most reliably obtained for accessions rather than single plants.

In our study the sequencing reads were aligned to the publicly available reference genome (Yasui et al. 2016), which is still fragmented and does not assign chromosome numbers to the scaffolds. With the upcoming high-quality assembly by NRGene (NRGENE 2018), a better understanding of the genetic distances between polymorphisms and their density on the chromosome will be possible. Assigning genomic locations to the polymorphisms genotyped will increase the efficiency to select for superior germplasm in future crossing-experiments via marker-assisted or genomic selection, and may allow to find candidate genes for certain traits.

4.3 High heterozygosity within buckwheat accessions emphasizes the need to use population genetics approaches

This study was the first to genetically describe buckwheat materials using allele frequencies instead of bi-allelic SNPs. We found that on average 80.3% of the genetic loci covered by GBS were heterzygous. Hence, genotyping a single plant to represent the entire genepool of an accession would result in missing out a large part of the diversity. The accession-specific allele frequencies can, however, be dynamic; for example genetic drift may act if population sizes are small (e.g. seed multiplication from a small batch of seeds) or if certain genotypes within the population cope better with new climatic conditions and therefore contribute more seeds to the next generation (Wright 1931). How these dynamics have affected buckwheat populations in the past is not known, but documenting the changes in allele frequencies in the future may allow to better understand the genetic basis of adaption to new environmental conditions (Günther & Coop 2013).

5. REFERENCES

Arvidsson, S., Fartmann, B., Winkler, S. & Zimmermann, W., 2016: Efficient high-throughput SNP discovery and genotyping using normalised Genotyping-by-Sequencing (nGBS). https://biosearch-cdn.azureedge.net/assetsv6/efficient-high-throughput-snp-discovery-genotyping-ngbs-app-note.pdf (19.09.2019)

Byrne, S., Czaban, A., Studer, B., Panitz, F., Bendixen, C. & Asp, T., 2013: Genome Wide Allele Frequency Fingerprints (GWAFFs) of Populations via Genotyping by Sequencing. PLoS ONE 8 (3).

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S. & Mitchell, S. E., 2011: A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6 (5): 1–10.

Falquet, B., Gfeller, A., Pourcelot, M., Tschuy, F. & Wirth, J., 2015: Weed suppression by common buckwheat: A review. Environmental Control in Biology 53 (1): 1-6.

Fehr, W. R. (1987): Principles of cultivar development. Theory and technique. New York.

Günther, T. & Coop, G., 2013: Robust identification of local adaptation from allele frequencies. Genetics 195 (1): 205–220.

Li, H. & Durbin, R., 2010: Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26 (5): 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R., 2009: The Sequence Alignment/Map format and SAMtools. Bioinformatics 25 (16): 2078–2079.

Li, S. Q. & Zhang, Q. H., 2001: Advances in the development of functional foods from buckwheat. Critical Reviews in Food Science and Nutrition 41 (6): 451–464.

Lichtenhahn, M. & Dierauer, H., 2000: Buchweizen. Frick.

Logacheva, M. D., Kasianov, A. S., Vinogradov, D. V., Samigullin, T. H., Gelfand, M. S., Makeev, V. J. & Penin, A. A., 2011: De novo sequencing and characterization of floral transcriptome in two species of buckwheat (*Fagopyrum*). BMC Genomics 12,30.

Matsui, K., Tetsuka, T., Nishio, T. & Hara, T., 2003: Heteromorphic incompatibility retained in self-compatible plants produced by a cross between common and wild buckwheat. New Phytologist 159 (3): 701–708.

Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & Depristo, M. A., 2010: The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20 (9): 254–260.

Mizuno, N. & Yasui, Y., 2019: Gene flow signature in the S-allele region of cultivated buckwheat. BMC Plant Biology 19 (1): 1–9.

Moose, S. P. & Mumm, R. H., 2008: Molecular Plant Breeding as the Foundation for 21st Century Crop Improvement. Plant Physiology 147 (3): 969–977.

Nagano, M., Alp, J., Campbell, C. & Kawasakp, S., 2000: Genome size analysis of the genus *Fagopyrum*. Fagopyrum 39:35–39.

NRGENE, 2018: NRGene assembles first accurate buckwheat genome. https://www.nrgene.com/press-releases/nrgene-assembles-buckwheat-genome/ (21.08.2019)

R CORE TEAM, 2008: R: A language and environment for statistical computing.

Stoskopf, N. C., Tomes, D. T. & Christie, B. R., 1999: Plant breeding: theory and practice. Jodhpur, India.

Teboh, J. M. & Franzen, D. W., 2011: Buckwheat (*Fagopyrum esculentum* Moench) potential to contribute solubilized soil phosphorus to subsequent crops. Communications in Soil Science and Plant Analysis 42 (13): 1544–1550.

Ueno, M., Yasui, Y., Aii, J., Matsui, K., Ota., T., Ueno, M., Yasui, Y., Aii, J., Matsui, K., Sato, S. & Ota, T., 2016: Genetic analyses of the heteromorphic self-incompatibility (S) locus in buckwheat. In: ZHOU, M., I, KREFT., WOO, S.-H., CHRUNGOO, N. & WIESLANDER, G. Hrsg.: Molecular Breeding and Nutritional Aspects of Buckwheat. 411–421.

Wetterstrand, K. A., 2019: Data from the NHGRI Genome Sequencing Program (GSP). www.genome.gov/sequencingcostsdata (19.08.2019)

Wright, S., 1931: Evolution of mendelian populations. Genetics 16 (97): 97–159.

Yasui, Y., Hirakawa, H., Ueno, M., Matsui, K., Katsube-Tanaka, T., Yang, S. J., Aii, J., Sato, S. & Mori, M., 2016: Assembly of the draft genome of buckwheat and its applications in identifying agronomically useful genes. DNA Research 23 (3): 215–224.