Current Plant Biology 21 (xxxx) xxxx

Contents lists available at ScienceDirect

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

In silico quality assessment of SNPs—A case study on the Axiom® Wheat genotyping arrays

Thomas M. Lange[a], Felix Heinrich[a], Matthias Enders[b], Markus Wolf[c], Armin O. Schmitt[a,d,*]

[a] Breeding Informatics Group, University of Göttingen, Germany
[b] NPZ Innovation GmbH, Hohenlieth, Germany
[c] German Seed Alliance GmbH, Cologne, Germany
[d] Center for Integrated Breeding Research (CiBreed), University of Göttingen, Germany

ARTICLE INFO

Keywords:
Triticum aestivum
Single nucleotide polymorphism
SNP classification
Affymetrix
Wheat Breeder's
Wheat HD

ABSTRACT

Genotyping arrays proved to be an exemplary tool for the simultaneous analysis of a multitude of single nucleotide polymorphisms (SNPs), a special case of genomic variants. By the example of SNPs represented on the Axiom® Wheat HD genotyping array as well as on the Axiom® Wheat Breeder's genotyping array, we applied a three way classification system to assess the quality of SNPs in bread wheat (*Triticum aestivum* L.) and subsequently the quality of these genotyping arrays. Class 1 SNPs could be aligned uniquely to the reference genome and did not show any genomic variants in their flanking sequence. Class 2 SNPs could also be aligned uniquely to the reference genome but showed genomic variants in their flanking sequence. The remaining SNPs were assigned to class 3. To determine the number of genomic variants in a SNP's flanking sequence, we used all currently available SNPs in the Ensembl Plants database. From the 819,571 SNPs on the Axiom® Wheat HD genotyping array, we assigned 24,343 to class 1 and from the 35,143 SNPs on the Axiom® Wheat Breeder's genotyping array we classified 2295 SNPs as class 1. We show that class 1 SNPs of the Axiom® Wheat HD genotyping array result in an equidistant coverage of the reference genome. We make the classification table as well as R-scripts available to give breeders and researchers the possibility to reproduce our analysis in an easy way. Moreover, we discuss the possibilities and limitations of such an *in silico* analysis of genotyping arrays as well as future research possibilities for this approach.

1. Introduction

1.1. Bread wheat and its need for genetic improvement

Due to the growing world population which is projected to reach 8.9 billion people until 2050 [1] paired with diet shifts and increasing biofuel consumption [2], agricultural production must increase tremendously in the next decades. For bread wheat (*Triticum aestivum* L.) which is one of the three most important crops in the world for both human consumption and livestock feed [3], the production target is to double the yield until 2050 [4] in spite of changing climate conditions which are most likely to increase both evaporation [5] and the time period for insect herbivores to grow and reproduce [6]. Using improved genomic research capabilities and acquisition of genetic diversity could support the genetic gain in breeding which is a key to tackle these challenges [7].

1.2. Recent advances in genetic analyses of bread wheat

Wheat is an allohexaploid plant with $2n = 6x = 42$ chromosomes [8]. The structure of the wheat genome with its three sub-genomes as well as its huge size of more than 14 billion base pairs (bp) and the high amount of repetitive sequences is the reason why both the sequencing and the annotation of a reference genome as well as the discovery of genomic variants like single nucleotide polymorphisms (SNPs) was challenging [9–11]. Nevertheless, at the beginning of this century, marker-assisted selection (MAS) based on microsatellite markers could successfully be implemented into wheat breeding programs, providing a huge advantage for breeding disease resistant varieties in wheat [12]. Breeding strategies have been adapted to the newly developed tool in the form of marker-assisted backcrossing [13] and MAS involving selection in a double haploid population [14]. Thus, the demand for SNP markers in wheat breeding programs is obvious as to increase the exploitation of the whole wheat genome [15] and to reduce costs per

* Corresponding author at: Breeding Informatics Group, Margarethe von Wrangell-Weg 7, 37075 Göttingen, Germany.
E-mail address: armin.schmitt@uni-goettingen.de (A.O. Schmitt).

https://doi.org/10.1016/j.cpb.2020.100140
Received 30 September 2019; Received in revised form 13 February 2020; Accepted 13 February 2020

Abbreviations

BLAST	Basic Local Alignment Search Tool
bp	base pairs
IWGSC	International Wheat Genome Sequencing Consortium
MAS	marker-assisted selection
SNP	single nucleotide polymorphism

marker [16].

In 2011, Allen et al. were able to discover the first SNPs in wheat and published 14,078 SNPs [17]. These SNPs as well as other newly discovered SNPs were stored in the online database CerealsDB [18,19]. In 2014, the first genotyping array for bread wheat was created by Wang et al. containing 81,587 SNPs [20]. This array was commercially distributed as Illumina's iSelect genotyping array. One year later, Winfield et al. published a high definition SNP genotyping array for wheat with a total of 819,571 SNP markers [21] which are commercially available on the Axiom® Wheat HD genotyping array. Another year later, Allen et al. used this genotyping array to select 35,143 markers for the Axiom® Wheat Breeder's genotyping array [22]. In 2017, Krasileva et al. were able to discover more than 10 million SNPs in bread wheat [23]. Since these SNPs were detected using exome capturing followed by targeted re-sequencing [21], mainly SNPs in the coding regions of the genome could be discovered [24]. In 2018, around 3.3 million new SNPs were discovered, mostly in the non-coding regions of the wheat genome [24]. Moreover, the International Wheat Genome Sequencing Consortium (IWGSC) was able to publish a fully annotated reference genome for bread wheat in August 2018 despite the above mentioned challenges [25]. Information about SNPs in plants are publicly available in the Ensembl Plants database (http://plants.ensembl.org). Currently this database contains 16,448,754 SNPs for bread wheat (release 44).

Results of the application of both genotyping arrays in breeding procedures were published by Brinton et al. who used the Axiom® Wheat HD genotyping array to identify markers for increased pericarp cell length and grain weight [26]. Moreover, the Axiom® Wheat HD genotyping array was used to analyze the genetic diversity of bread wheat [7] as well as of close relatives [27]. The later analysis was then used to integrate resistance against African stem rust from *Aegilops sharonensis* L. into *Triticum aestivum* L. [27]. Furthermore, the Axiom® Wheat Breeder's genotyping array was used to identify resistance genes against wheat stripe rust [28], showing that both the Axiom® Wheat HD and the Wheat Breeder's genotyping array can be beneficial in breeding processes.

1.3. Quality assessment of genotyping arrays

Although the Axiom® genotyping arrays have proven to be beneficial for the introgression of certain traits in wheat breeding programs, an objective quality assessment of these arrays has been missing. Quality constraints for the analysis of SNPs have been expressed [29,30] but a combination of those constraints was not yet adapted to control the quality of SNPs and genotyping arrays. With the availability of a reference genome as well as of a database with a huge set of SNPs in bread wheat, we assessed the quality of SNPs in the 820 K (Wheat HD) and 35 K (Wheat Breeder's) genotyping arrays. In this way, we present and discuss the possibility to assess the quality of a genotyping array *in silico* if a reference genome is available. Therefore, we aligned the flanking sequences of the SNPs to the reference genome and counted the number of perfect matches to the reference genome for each SNP. For those SNPs which could be aligned one time perfectly to the reference genome, we counted the number of genomic variants in their flanking sequences. We provide our findings in two supplementary tables (**additional file 1** and **additional file 2**) and publish the R-scripts necessary to reproduce the analysis (**additional file 3** and **additional file 4**).

2. Material and methods

2.1. Material

We downloaded both annotation files for the 820 K genotyping array (`Axiom_BristolW_A_Annotation.r2.csv` and `Axiom_BristolW_B_Annotation.r2.csv`) as well as the annotation file for the 35 K genotyping array (`Axiom_WhtBrd-1_Annotation.r3.csv`) from the Affymetrix support site (www.affymetrix.com) on August 9, 2019. Next, we downloaded a data set containing SNPs in bread wheat in gzipped variant call format (`triticum_aestivum.vcf.gz`) as well as the wheat reference genome (`Triticum_aestivum.IWGSC.dna.toplevel.fa`) from Ensembl Plants (ftp://ftp.ensemblgenomes.org/pub/release-44/plants/) on August 9, 2019.

The complete reference genome of bread wheat consists of 14,547,261,565 bp. The three subgenomes contain seven chromosomes each. The lengths of the chromosomes vary from 473,592,718 bp to 830,829,764 bp. An additional unassigned chromosome (`ChrUn`) consisting of 480,980,714 bp represents sequences which could not be assigned to one of the chromosomes [25].

2.2. Methods

The flanking sequences were extracted from the annotation files and written into a fasta file using our custom R-script `PreparingFlanks.R` (**additional file 3**). The SNPs were coded in their respective symbols of the IUPAC nucleotide code ([A/G] = R, [C/T] = Y, [G/C] = S, [A/T] = W, [G/T] = K, [A/C] = M). The length of the flanking sequences varies from a total length (flanking sequences plus SNP) of 37 bp to 71 bp. The majority of flanking sequences (98%) had a length of 71 bp. Various IDs were used in the fasta file to assure unambiguous attribution of the data in downstream analyses. In this way, the probe set ID (named Affymetrix code in CerealDB), the Affymetrix SNP ID and the customer ID (termed Bristol Affy code in CerealDB) were used as identifiers in the created fasta file. To export the result in a fasta file, the R-package `ape` was used [31].

To locate the positions of the SNPs, we aligned the complete flanking sequence of the SNPs to the reference genome using the Basic Local Alignment Search Tool (BLAST) [32]. First, we created a BLAST database from the reference genome using the `makeblastdb` tool of the `blast+` suite (//ftp.ncbi.nlm.nih.gov/blast/executables/blast+/ in the version 2.9.0) with standard parameters [33]. Subsequently, we aligned the flanking sequences to the wheat reference genome with the `blastn` tool from the `blast+` suite using standard parameters [33] and the tabular output format (`-outfmt 6`).

After the flanking sequences were aligned to the reference genome, we filtered the results by considering only perfect alignments in further analyses. An alignment was considered perfect, if it stretched over the whole length of the flanking sequence without gaps and with the SNP itself as the only mismatch since the reference genome does not contain the ambigious IUPAC symbols.

We wrote an R-script (`ClassfiyMarkers.R`, **additional file 4**) which classifies SNPs depending on their genomic location. This script uses the R-packages `Biostrings` for reading in fasta files [34] and `vcfR` for reading in vcf files [35]. The counting of SNPs in the flanking sequences was done with the packages `tidyr` [36], `plyr` [37], `tibble` [38] and `seqinr` [39]. We considered SNPs with a unique position in the genome as most informative. As a second criterion we counted genomic variants in the flanking sequence of each uniquely matching SNP, following the quality criteria for SNPs as described by Ganal et al. [30]. In this way, we assigned all SNPs into one of three classes:

- Class 1 for SNPs which are perfectly matching to a single genomic locus and which have no other SNP in their flanking sequence.
- Class 2 for SNPs which are perfectly matching to a single genomic locus but with other SNPs in their flanking sequence.
- Class 3 for all other SNPs.

The results of the classification were stored in `ClassificationTable_820_WheatHD.csv` (**additional file 1**) and `ClassificationTable_35_WheatBrd.csv` (**additional file 2**). These tables contain the above mentioned identifiers of the SNPs, the position of the SNP as a combination of chromosome (`Chrom`) and position on the chromosome (`bp`), the number of times a SNP could be perfectly aligned to the genome (`FreqPerfectAlign`) and the number of SNPs in the flanking region of the SNP under investigation (`SNPInVicinity`). SNPs without unique perfect alignment to the reference genome are marked with `NA` for the position of the SNP as well as for the number of SNPs in vicinity (`SNPInVicinity`).

2.3. SNP classification

The SNPs from the genotyping arrays were located by the alignment of their flanking sequences to the reference genome. The SNPs which were downloaded from Ensembl on the other hand already contained information about their position but no flanking sequences which could have been aligned in the same way as the SNPs from the genotyping arrays.

Unfortunately, the SNP identifiers used in Ensembl and in the annotation file are incompatible such that the SNP position given by Ensembl was the only piece of information that could be used to determine if a SNP is located in the flanking region.

Therefore, the classification of SNPs that are perfectly aligned to the reference genome, i.e. which have just one mismatch in the alignment, and which have one Ensembl SNP in their flanking sequence require special consideration. Two cases can be distinguished:

(1) The SNP is mapped via BLAST to the same position as the corresponding SNP in the Ensembl database and no other SNP is found in the flanking sequence. Then this SNP belongs to class 1.
(2) The SNP that was mapped is not represented in the Ensembl database. Then the Ensembl SNP is necessarily a different SNP in the flanking region and, thus, the mapped SNP is assigned to class 2.

The above explained method was used to classify all SNPs of the 820 K genotyping array (**additional file 1**). Subsequently, the classes of SNPs in the 35 K genotyping array were extracted from this table to analyze how many of the class 1 and class 2 SNPs from the 820 K genotyping array were transferred into the 35 K genotyping array (**additional file 2**).

3. Results

The alignment of the 819,571 SNPs to the reference genome resulted in 2,442,786 possible positions for the SNPs under investigation. The majority of flanking sequences in the 820 K genotyping array (87.2%) could be mapped to the reference genome without gaps or mismatches (except for those resulting from the SNP itself in IUPAC symbols), resulting in 104,859 SNPs (12.8%) from the 820 K genotyping array which could not be matched perfectly to the reference genome and were thereby assigned to class 3. 62.3% of the SNPs under investigation could be mapped to a unique position on the reference genome, whereas the remaining 24.9% were mapped two or more times to the reference genome. The SNPs on the 35 K genotyping array show similar results (Table 1).

Both arrays show the least number of SNPs assigned to class 1. Nevertheless, the 35 K genotyping array shows a considerably higher proportion of SNPs in class 1 than the 820 K genotyping array. While from the SNPs on the 820 K genotyping array 3% could be assigned to class 1, 6.5% of SNPs in the 35 K genotyping array could be classified as class 1. In both genotyping arrays, most SNPs were classified as class 2.

After aligning the SNPs from the 820 K genotyping array to the reference genome, 6067 SNPs were located on the unassigned chromosome (`ChrUn`). From those SNPs, 1927 SNPs were classified as class 1 and the remaining 4140 SNPs as class 2. From the 35 K genotyping array, 382 SNPs were located at the unassigned chromosome. 141 of those SNPs were classified as class 1 and the remaining 241 SNPs as class 2.

Fig. 1 shows the distribution of SNPs present in the 35 K genotyping array which were classified as class 1 (red dots) and the class 1 SNPs on the 820 K genotyping array (blue dots). As one can see in Fig. 1, class 1 SNPs of the 820 K genotyping array are evenly distributed across every chromosome of the wheat genome whereas the class 1 SNPs of the 35 K genotyping array show substantial gaps at several locations.

4. Discussion

Through our analysis, 24,343 SNPs of the 820 K genotyping array and 2295 SNPs from the 35 K genotyping array were assigned to class 1. As one can see in Fig. 1, the class 1 SNPs of the 820 K genotyping array show an equidistant distribution and a full coverage of the whole wheat genome despite the limited number of SNPs in this class. The distribution of class 1 SNPs from the 35 K genotyping array on the other hand shows at certain places an accumulation and at other locations a depletion of SNPs. Considering the distribution of class 1 SNPs on the 820 K genotyping array, it would have been possible to select SNPs from the 820 K genotyping array to fill the locations with lower density on the 35 K genotyping array. Using this resource, our analysis of the SNPs in both the 820 K as well as the 35 K genotyping array shows the potentiality of a reference genome for reliable SNP selection and quality assessment of genotyping arrays.

As described by Ganal et al., the number of SNPs on a genotyping array is limited due to the additional costs per marker, setting an economical limitation to the number of SNPs per genotyping array [30]. But when considering every SNP on a genotyping array as a single test, the number of SNPs on a genotyping array increases the probability of finding at least one of them to be statistically significant just by chance. That is the reason why it is common to use a correction of α such as the Bonferroni correction when analyzing associations between genotype and phenotype [40]. There is a variety of more sophisticated methods to overcome the inflation of p-values like the usage of mixed models that account for population structure [41], usage of permutation algorithms like PRESTO or usage of a multivariate normal distribution-based correction like SLIDE [42]. Apparently, in this context it also makes sense to reduce the inflation of p-values by reducing the number of SNPs and hence, the number of tests used in a genomic analysis.

Moreover, if for example a SNP with association to a phenotypic characteristic could be analyzed through a genotyping array but the same SNP aligns simultaneously to several genomic loci, the effects of the other locations might be in conflict with the actual effect and, hence, lower the probability of detecting this SNP as significantly associated to the trait. Therefore, it could be an interesting approach to use our classification system before performing an association study

Table 1
Number of SNPs classified in the two analyzed genotyping arrays. SNPs in vicinity are SNPs found in the flanking sequence other than the SNP under investigation.

Alignment	Other SNPs in vicinity	35 K	820 K	Class
Unique	0	2295	24,343	1
Unique	≥ 1	20,304	486,041	2
None	NA	4235	104,859	3
Ambiguous	NA	8309	204,328	3

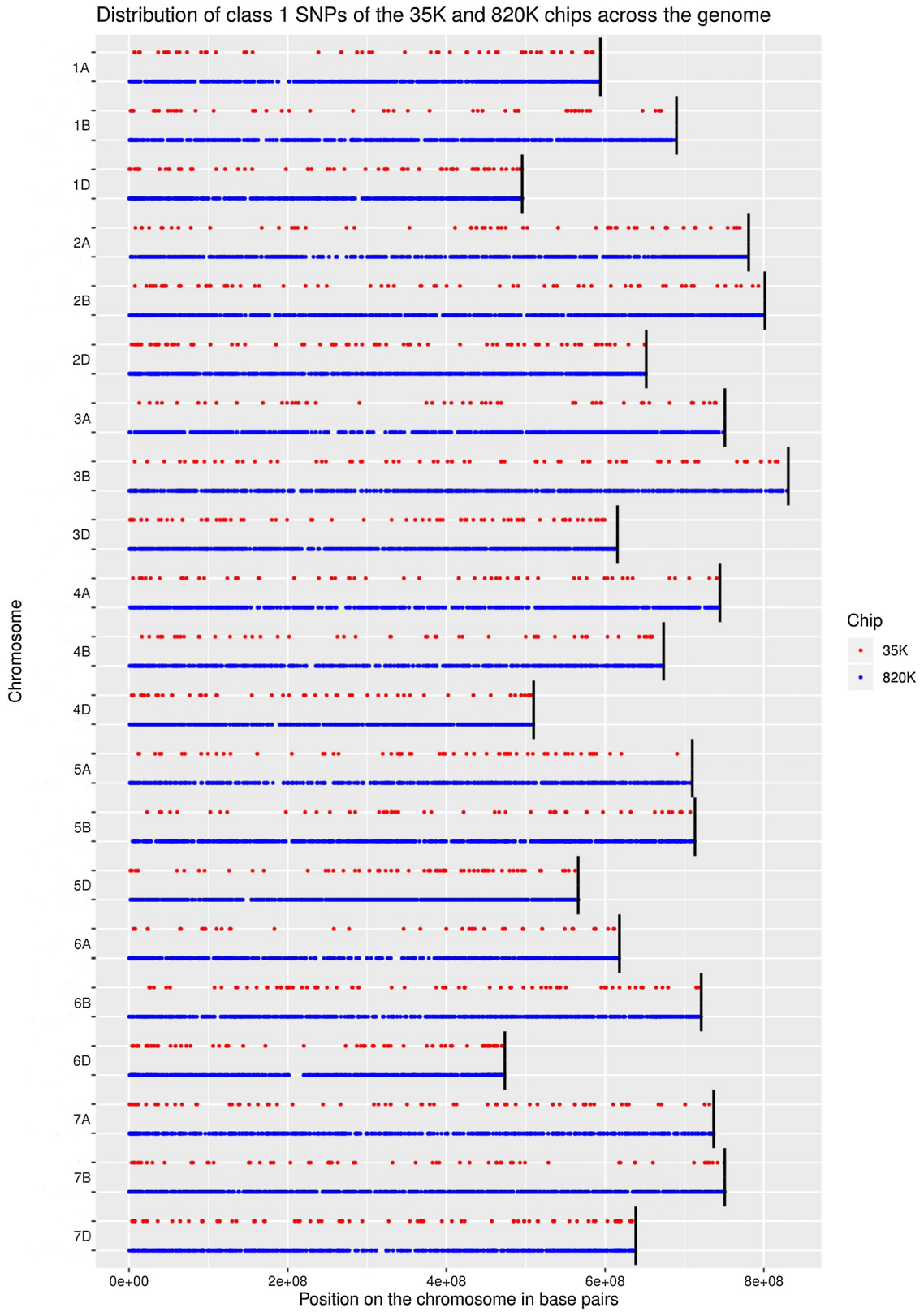

Fig. 1. Graphical display of class 1 SNPs on the 35 K (red dots) and 820 K genotyping array (blue dots). The *y*-axis displays the chromosome, the *x*-axis displays the position on the chromosome. The length of each chromosome is visualized by a vertical black line.

and then using only SNPs in certain classes in the following analysis. Nevertheless, it must be pointed out that this argument is of a theoretical nature. In future analyses it would be interesting to apply our classification system after performing an association study to test the number of SNPs which are associated with a phenotypic trait in different classes.

The reference genome of bread wheat was published only recently in the first version and 480,980,714 bp were aligned to an unassigned

chromosome (ChrUn). 6067 SNPs on the 820 K genotyping array and 382 SNPs on the 35 K genotyping array were mapped to the unassigned chromosome. These SNPs might be classified differently if the contigs of the unassigned chromosome are assigned to the correct positions in the chromosomes in a future version of the wheat reference genome. Moreover, wheat breeders who have an individual reference genome could repeat the analysis since the results of this analysis are depending on the used reference genome.

Furthermore, this analysis was impaired due to limited information about the physical probes on these arrays. We were only able to use the complete flanking sequences at both sides of the SNPs while the physical probes on the array are likely to be shorter. On the one hand, this could lead to a perfect alignment of probes which we could not align perfectly to the reference genome when using the whole flanking sequence as query sequence. On the other hand, this might lead to ambiguous locations of SNPs which were aligned uniquely to the reference genome when mapping the whole flanking sequence.

5. Conclusion

Based on the assumptions about quality criteria of SNPs stated by Ganal et al. [30], we created a pipeline to classify SNPs into three classes depending on quality features. In this way, we provide a possibility to assess the quality of genotyping arrays *in silico* if information about the flanking sequences as well as a reference genome are available. Moreover, we used the pipeline to assess the quality of SNPs on the Axiom® 820 K and 35 K genotyping arrays. We showed that most SNPs of both genotyping arrays could be aligned uniquely to the reference genome and visualized the class 1 SNPs of both genotyping arrays to show that the classification criteria dissolve clusters of SNPs to some extent and, in the case of the 820 K genotyping array, provide an even coverage of the whole genome.

In future research, it would be interesting to test our classification system with data about the association of genotype and phenotype to control the utility of our pipeline as a possibility to reduce the number of SNPs on a genotyping array and thereby reducing the number of tests and the inflation of *p*-values in a genomic analysis. Moreover, it would be beneficial to reproduce the pipeline when a new version of the reference genome of bread wheat is available perhaps without unassigned contigs, although only few SNPs of both arrays were aligned to this chromosome. Furthermore, the mapping of SNPs was impaired due to limited information about the actual probe sequences which are physically represented on the genotyping arrays. The results could change if the actual probe sequences are used to map the SNPs on the reference genome. We are certain that it would be a benefit for all stakeholders if more precise information about the physical probe sequences on the genotyping arrays was publicly available.

Author contributions

AOS has conceptualized the study, TML, ME and MW have written the manuscript, FH has contributed to the writing of the analysis scripts.

Conflict of interests

The authors declare that there is no conflict of interests.

Acknowledgements

The authors would like to thank Dr. Dr. Johannes Martini from the International Maize and Wheat Improvement Center (CIMMYT) for professional advice in the writing process. The authors also thank the NPZ Innovation GmbH and the German Seed Alliance GmbH for providing support for data analysis. We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the University of Göttingen.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.cpb.2020.100140.

References

[1] J.E. Cohen, Human population: the next half century, Science 302 (5648) (2003) 1172–1175.
[2] D.K. Ray, N.D. Mueller, P.C. West, J.A. Foley, Yield trends are insufficient to double global crop production by 2050, PLOS ONE 8 (6) (2013) e66428.
[3] B. Leff, N. Ramankutty, J.A. Foley, Geographic distribution of major crops across the world, Global Biogeochem. Cycles 18 (1) (2004).
[4] A. Rasheed, A. Mujeeb-Kazi, F.C. Ogbonnaya, Z. He, S. Rajaram, Wheat genetic resources in the post-genomics era: promise and challenges, Ann. Bot. 121 (4) (2017) 603–616.
[5] K.W. Jaggard, A. Qi, E.S. Ober, Possible changes to arable crop yields by 2050, Philos. Trans. R. Soc. B: Biol. Sci. 365 (1554) (2010) 2835–2851.
[6] J.S. Bale, G.J. Masters, I.D. Hodkinson, C. Awmack, T.M. Bezemer, V.K. Brown, J.B. Whittaker, Herbivory in global climate change research: direct effects of rising temperature on insect herbivores, Global Change Biol. 8 (1) (2002) 1–16.
[7] M.O. Winfield, A.M. Allen, P.A. Wilkinson, A.J. Burridge, G.L. Barker, J. Coghill, K.J. Edwards, High-density genotyping of the a.e. watkins collection of hexaploid landraces identifies a large molecular diversity compared to elite bread wheat, Plant Biotechnol. J. 16 (1) (2017) 165–175.
[8] T. Marcussen, S.R. Sandve, L. Heier, M. Spannagl, M. Pfeifer, The International Wheat Genome Sequencing Consortium, K.S. Jakobsen, B.B.H. Wulff, B. Steuernagel, K.F.X. Mayer, O.-A. Olsen, Ancient hybridizations among the ancestral genomes of bread wheat, Science 345 (6194) (2014) 1250092.
[9] M.S. Röder, V. Korzun, K. Wendehake, J. Plaschke, M.H. Tixier, P. Leroy, M.W. Ganal, A microsatellite map of wheat, Genetics 149 (4) (1998) 2007–2023.
[10] M.O. Winfield, P.A. Wilkinson, A.M. Allen, G.L.A. Barker, J.A. Coghill, A. Burridge, K.J. Edwards, Targeted re-sequencing of the allohexaploid wheat exome, Plant Biotechnol. J. 10 (6) (2012) 733–742.
[11] P.J. Berkman, K. Lai, M.T. Lorenc, D. Edwards, Next-generation sequencing applications for wheat crop improvement, Am. J. Bot. 99 (2) (2012) 365–371.
[12] J. Dubcovsky, Marker-assisted selection in public breeding programs, Crop Sci. 44 (6) (2004) 1895.
[13] H. Kuchel, R. Fox, J. Reinheimer, L. Mosionek, N. Willey, H. Bariana, S. Jefferies, The successful application of a marker-assisted wheat breeding strategy, Mol. Breed. 20 (4) (2007) 295–308.
[14] G. Jia, P. Chen, G. Qin, G. Bai, X. Wang, S. Wang, B. Zhou, S. Zhang, D. Liu, QTLs for fusarium head blight response in a wheat DH population of wangshuibai/alondra's', Euphytica 146 (3) (2006) 183–191.
[15] P.K. Gupta, P. Langridge, R.R. Mir, Marker-assisted wheat breeding: present status and future possibilities, Mol. Breed. 26 (2) (2009) 145–161.
[16] H.M. William, R. Trethowan, E.M. Crosby-Galvan, Wheat breeding assisted by markers: CIMMYT's experience, Euphytica 157 (3) (2007) 307–319.
[17] A.M. Allen, G.L. Barker, S.T. Berry, J.A. Coghill, R. Gwilliam, S. Kirby, K.J. Edwards, Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (triticum aestivum l.), Plant Biotechnol. J. 9 (9) (2011) 1086–1099.
[18] P.A. Wilkinson, M.O. Winfield, G.L. Barker, A.M. Allen, A. Burridge, J.A. Coghill, K.J. Edwards, CerealsDB 2.0: an integrated resource for plant breeders and scientists, BMC Bioinform. 13 (1) (2012).
[19] P.A. Wilkinson, M.O. Winfield, G.L.A. Barker, S. Tyrrell, X. Bian, A.M. Allen, K.J. Edwards, CerealsDB 3.0: expansion of resources and data integration, BMC Bioinform. 17 (1) (2016).
[20] S. Wang, D. Wong, K. Forrest, A. Allen, S. Chao, B.E. Huang, EA, Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array, Plant Biotechnol. J. 12 (6) (2014) 787–796.
[21] M.O. Winfield, A.M. Allen, A.J. Burridge, G.L.A. Barker, H.R. Benbow, K.J. Edwards, High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool, Plant Biotechnol. J. 14 (5) (2015) 1195–1206.
[22] A.M. Allen, M.O. Winfield, A.J. Burridge, R.C. Downie, H.R. Benbow, G.L.A. Barker, K.J. Edwards, Characterization of a wheat breeders' array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (triticum aestivum), Plant Biotechnol. J. 15 (3) (2016) 390–401.
[23] K.V. Krasileva, H.A. Vasquez-Gross, T. Howell, P. Bailey, F. Paraiso, L. Clissold, J. Dubcovsky, Uncovering hidden variation in polyploid wheat, Proc. Natl. Acad. Sci. USA 114 (6) (2017) E913–E921.
[24] H. Rimbert, B. Darrier, J. Navarro, J. Kitt, F. Choulet, M. Leveugle, E. Paux, High throughput SNP discovery and genotyping in hexaploid wheat, PLOS ONE 13 (1) (2018) e0186329.
[25] IWGSC, Shifting the limits in wheat research and breeding using a fully annotated reference genome, Science 361 (6403) (2018).
[26] J. Brinton, J. Simmonds, F. Minter, M. Leverington-Waite, J. Snape, C. Uauy, Increased pericarp cell length underlies a major quantitative trait locus for grain weight in hexaploid wheat, New Phytol. 215 (3) (2017) 1026–1038.
[27] E. Millet, B.J. Steffenson, R. Prins, H. Sela, A.M. Przewieslik-Allen, Z.A. Pretorius, Genome targeted introgression of resistance to african stem rust from into bread

wheat, Plant Genome 10 (3) (2017) 0.
[28] J. Mu, S. Huang, S. Liu, Q. Zeng, M. Dai, Q. Wang, J. Wu, S. Yu, Z. Kang, D. Han, Genetic architecture of wheat stripe rust resistance revealed by combining QTL mapping using SNP-based genetic maps and bulked segregant analysis, Theor. Appl. Genet. 132 (2) (2018) 443–455.
[29] A.O. Schmitt, R.H. Bortfeldt, G.A. Brockmann, Tracking chromosomal positions of oligomers – a case study with illumina's bovinesnp50 beadchip, BMC Genomics 11 (1) (2010) 80.
[30] M.W. Ganal, A. Polley, E.-M. Graner, J. Plieske, R. Wieseke, H. Luerssen, G. Durstewitz, Large SNP arrays for genotyping in crop plants, J. Biosci. 37 (5) (2012) 821–828.
[31] E. Paradis, J. Claude, K. Strimmer, APE: analyses of phylogenetics and evolution in R language, Bioinformatics 20 (2) (2004) 289–290.
[32] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (3) (1990) 403–410.
[33] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T.L. Madden, BLAST+: architecture and applications, BMC Bioinform. 10 (1) (2009) 421.
[34] H. Pagès, P. Aboyoun, R. Gentleman, S. DebRoy, Biostrings: Efficient Manipulation of Biological Strings. R Package Version 2.52.0, (2019).
[35] B.J. Knaus, N.J. Grünwald, vcfr: a package to manipulate and visualize variant call format data in r, Mol. Ecol. Resour. 17 (1) (2016) 44–53.
[36] H. Wickham, L. Henry, tidyr: Tidy Messy Data. R Package Version 1.0.0, (2019).
[37] H. Wickham, The split-apply-combine strategy for data analysis, J. Stat. Softw. 40 (1) (2011) 1–29.
[38] K. Müller, H. Wickham, tibble: Simple Data Frames. R Package Version 2.1. (2019), p. 3.
[39] D. Charif, J. Lobry, SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. in: U. Bastolla, M. Porto, H. Roman, M. Vendruscolo (Eds.), Structural Approaches to Sequence Evolution: Molecules, Networks, Populations. Biological and Medical Physics, Biomedical Engineering, Springer Verlag, New York, 2007, pp. 207–232 ISBN: 978-3-540-35305-8.
[40] D.L. Streiner, G.R. Norman, Correction for multiple testing, Chest 140 (1) (2011) 16–18.
[41] A. Korte, A. Farlow, The advantages and limitations of trait analysis with GWAS: a review, Plant Methods 9 (1) (2013) 29.
[42] R.C. Johnson, G.W. Nelson, J.L. Troyer, J.A. Lautenberger, B.D. Kessing, C.A. Winkler, S.J. O'Brien, Accounting for multiple comparisons in a genome-wide association study (GWAS), BMC Genomics 11 (1) (2010) 724.