High throughput crop genome genotyping by a combination of pool next generation sequencing and haplotype-based data processing

Michael Schneider
Institute of Crop Science and Resource Conservation

Asis Shrestha
University of Bonn Institute of Crop Science and Resource Conservation

Agim Ballvora
University of Bonn

Jens Leon (✉️ j.leon@uni-bonn.de)
University of Bonn  https://orcid.org/0000-0002-6514-3112

Research

Keywords: Pool sequencing, crops, allele frequency estimation, SNP, haplotype, Hordeum vulgare

DOI: https://doi.org/10.21203/rs.3.rs-415602/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Background

The identification of environmentally specific alleles and the observation of evolutilional processes is a goal of conservation genomics. By generational changes of allele frequencies in populations, questions regarding effective population size, gene flow, drift, and selection can be addressed. The observation of such effects often is a trade-off of costs and resolution, when a decent sample of genotypes should be genotyped for many loci. Pool genotyping approaches can derive a high resolution and precision in allele frequency estimation, when high coverage sequencing is utilized. Still, pool high coverage pool sequencing of big genomes comes along with high costs.

Results

Here we present a reliable method to estimate a barley population's allele frequency at low coverage sequencing. Three hundred genotypes were sampled from a barley backcross population to estimate the entire population's allele frequency. The allele frequency estimation accuracy and yield were compared for three next generation sequencing methods. To reveal accurate allele frequency estimates on a low coverage sequencing level, a haplotyping approach was performed. Low coverage allele frequency of positional connected single polymorphisms were aggregated to a single haplotype allele frequency, resulting in two to 271 times higher depth and increased precision. We compared different haplotyping tactics, showing that gene and chip marker-based haplotypes perform on par or better than simple contig haplotype windows. The comparison of multiple pool samples and the referencing against an individual sequencing approach revealed whole genome pool resequencing having the highest correlation to individual genotyping (up to 0.97), while transcriptomics and genotyping by sequencing indicated higher error rates and lower correlations.

Conclusion

Using the proposed method allows to identify the allele frequency of populations with high accuracy at low cost. This is particularly interesting for conservation genomics in species with big genomes, like barley or wheat. Whole genome low coverage resequencing at 10x coverage can deliver a highly accurate estimation of the allele frequency, when a loci-based haplotyping approach is applied. Using annotated haplotypes allows to capitalize from biological background and statistical robustness.

Full Text

This preprint is available for download as a PDF.

Tables
Due to technical limitations, Tables 1-4 are only available as a download in the supplemental files section.

**Supplementary Table 1 is not available with this version**

**Figures**

**Figure 1**

Gene extension algorithm scheme of gene- and marker-based haplotype calculations. The raw SNP based allele frequency (2nd row; “Raw Frequency”) for identified SNPs (1st row) are identified at a given read depth (3rd row; “Read depth”). Annotated genes (4th row; “Genes”) and markers (6th row; “Markers”) are extended in size up- and downstream (“Genes extended”; “Marker extended”) to associate SNPs in the particular region to the genes / markers. By the extension, more SNPs can be annotated to a gene / marker than without (dashed lines and arrows below the 3rd row indicate relationships). Reported frequency in the last two rows of markers and genes are the calculated haplotype frequency for the wild donor parent in the population. The marker and gene-based haplotype calculation is illustrated in the methods section. The figure should only illustrate the model of SNP information aggregation and does not contain real data.
Figure 2

Read count per haplotype for the three applied sequencing methods with their mean value as a vertical line and the distribution of the haplotype window size for gene-based (GH) and marker-based (MH) haplotypes. A – Count of reads per GH for the three applied sequencing methods. B – same as for A, but based on MH. C – Size of the GH in base pairs. The dotted line indicates the median value, the dashed line represents the mean value. D – same as C for MH.
Figure 3

The correlation of pool obtained allele frequency for three tested sequencing strategies (WGrS, GBS, MACE). Three different methods were compared on their accuracy for pool sequencing of large pools. A – SNP-based pool AF compared to the KASP assay detected allele frequency for WGrS. B – Gene haplotype-based pool AF compared to KASP assay for WGrS. C – SNP-based pool AF to KASP for GBS. D - Gene haplotype-based pool AF compared to KASP assay for GBS. E – SNP-based pool AF to KASP for MACE RNAseq. F - Gene haplotype-based pool AF compared to KASP assay for MACE RNAseq. The dotted line indicates the optimal match of individual (KASP assay) and pool sequencing. Each point represents a single locus – for A, C and E, a locus is a SNP; for B, D and F, a locus is a gene, related to a KASP marker. The colour of the points presents the read coverage of the locus. The red line is a regression through all points.
Figure 4

Genome wide allele frequency on a genetic map for MH. The donor allele frequency is plotted in % (y-axis) against the genomic position (x-axis), split by chromosome and illustrated in centiMorgan. Each dot represents a MH and the color is related to the read coverage. The orange line indicates the expected allele frequency in the BC2F1. A – MACE RNA sequencing output, B – WGrS, C – GBS.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Tables.pdf
- SupplementaryMaterial.pdf