PEANUT SCIENCE

The Journal of the American Peanut Research and Education Society

SUPPLEMENTAL MATERIAL

A Disillusioned Note on Biofortification Breeding: Supplementary Materials & Methods

K.R. Fritz[1]; R.J. Andres[1]; C.S. Newman[1]; A.T. Oakley[1]; R.R. Alvarez[2]; L.L. Dean[3]; J.C. Dunne[1*]

[1]Department of Crop and Soil Sciences, North Carolina State University, Raleigh NC, 27695 2Department of Molecular and Structural Biochemistry, North Carolina State University, Raleigh NC, 27695;3USDA, ARS, Food Science and Market Quality and Handling Research Unit, Raleigh, NC 27695-7624.

ENZYME-LINKED IMMUNOSORBENT ASSAY (ELISA)

The General Tetrahydrofolic Acid (THFA) Enzyme-Linked Immunosorbent Assay (ELISA) Kit from MyBioSource (San Diego, CA, MBS2000078) was tested on seed cores taken from five replicates each of a low (Georgia Green), medium (PI 476025), and high (PI 371521) folate accession (Dean *et al.*, 2009). Seed cores were taken with a 3.0mm biopsy pen (World Precision Instruments, Sarasota, FL, WP3030), placed in a 2mL micro-centrifuge tube along with a 4mm stainless steel grinding bead (SPEX SamplePrep, Metuchen, NJ, #2150) and shaken for 1min at 1,500rpm in a SPEX 1600 MiniG. Seed cores were highly uniform in size, and all came from plants harvested, dried, and shelled in October 2020 using our standard program protocols. Seeds were stored at room temperature until the ELISAs were performed in three replicates on Nov 9, Nov17, and Dec 16 of 2020. Samples were assayed per the manufacturer's instructions after being lysed with the accompanying 'Cell Lysate' procedure. Standards used were 100, 50, 25, 12.5, 6.25, 3.125, 1.5625, and 0 ng/mL. The standard curve was fit using an exponential decay model and the best-fit trend line used to calculate folate concentrations for the replicated check samples.

GENOME WIDE ASSOCIATION STUDY

Both marker sets were pruned for linkage disequilibrium (LD) at an r^2 value of 0.99 using Plink and VCFtools (Danecek *et al.*, 2011). Folate phenotypic data (Dean *et al.*, 2009) was then paired with both marker sets to conduct a genome-wide association study (GWAS) in TASSEL 5 (Bradbury *et al.*, 2007). Any taxa with >5% heterozygosity was removed from its respective data set. In addition, sites were filtered out based on the following criteria: heterozygosity >0%, minor allele frequency (MAF) <5%, and/or missing data >5%. Population structure and kinship matrices were developed using TASSEL 5 to control for population substructure and relatedness in the GWAS model, respectively (Kang *et al.*, 2008). To reduce the complexity of the GWAS model, four subpopulations were used reflecting the prior population structure analysis of the mini core (Otyama *et al.*, 2019). A mixed linear model (MLM) with 1,000 permutations was fit using principal components accounting for population structure, the kinship matrices accounting for covariance among mini core accessions, and marker data to identify SNPs associated with folate content. Initially, any SNP with a permutation p-value ≤0.05 was considered significant. Physical locations of significant SNPs were then determined by aligning corresponding probe sequences to the Tifrunner gnm2 reference genome using Bowtie 2 (Langmead and Salzberg, 2012; Langmead *et al.*, 2019).

RIL POPULATION GENOTYPING

Young leaf tissue was harvested from a single plant of each of the four parental accessions and flash frozen in liquid nitrogen. DNA was isolated from the young leaf tissue with the Qiagen DNeasy Plant Mini Kit. Whole genome sequencing (WGS) library prep was performed with the Illumina Truseq Nano DNA kit and sequenced on an Illumina NovaSeq 6000. WGS data was cleaned by fastp v. 0.20.1 (Chen *et al.*, 2018) and aligned to the Tifrunner gnm2 genome (Bertioli *et al.*, 2019) using Bowtie 2 . The resulting VCF file was filtered in TASSEL to contain only biallelic SNPs where the common parent

Georgia Green carried the minor allele and all three high folate parents carried the same major allele (i.e. every SNP was informative in all three RIL populations). This list of filtered SNPs was then sent to Tecan Allegro® Targeted Genotyping to design a custom set of 10,000 probes to genotype the RIL populations.

Five seeds of all 557 RILs were germinated under greenhouse conditions. Two reps (five seeds each) were planted for each of the four parental accessions. This parental seed was descended from the individual plant that was whole-genome sequenced. In addition, two reps (five seeds each) were planted from a second source (USDA-GRIN) for each of the four parental accessions. Including three negative controls, the total number of samples entering the Tecan Allegro® Targeted Genotyping V2 protocol was 576.

Once germinated seedlings reached ~10cm in height, ~40mg of fresh leaf tissue was bulked amongst the germinated seedlings of each RIL or parental rep and flash frozen in liquid nitrogen. DNA was isolated using the Qiagen DNeasy 96 Plant Kit, quantified via PicoGreen, and normalized to a concentration of 10ng/uL. Normalized DNA was used in the Tecan Allegro® Targeted Genotyping V2 protocol per the manufacturer's instructions and sequenced on an Illumina NovaSeq 6000 SP 100bp SR flow cell by the NCSU Genomic Sciences Lab.

Markers were called using a bioinformatics pipeline (Supplementary File 1) run on the Windows Subsystem for Linux (Ubuntu version 20.04.2 LTS). Bioinformatics tools were installed and run through an Anaconda virtual environment (Version 4.5.11). Virtual environment was activated and channel dependencies installed. Files were extracted and file names subsequently shortened. Sequence reads within FASTA files were aligned to the Tifrunner gnm2 genome using the local setting in Bowtie 2. The aligned SAM file was sorted, converted to BAM form and multi-mapping reads removed using Samtools (Li *et al.*, 2009). SNPs were called using Freebayes (Garrison and Marth, 2012). VCFtools was used to filter off-target SNPs. The resulting VCF was filtered in TASSEL 5 to remove SNPs with >20% missing data as well was monomorphic SNPs.

REFERENCES

Bertioli, D.J., J. Jenkins, J. Clevenger, O. Dudchenko, D. Gao, et al. 2019. The genome sequence of segmental allotetraploid peanut Arachis hypogaea. Nat. Genet. 51: 877–884. doi: 10.1038/s41588-019-0405-z.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, et al. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics. 23(19): 2633–2635. doi: 10.1093/bioinformatics/btm308.

Chen, S., Y. Zhou, Y. Chen, and J. Gu. 2018. Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 34(17): i884–i890. doi: 10.1093/bioinformatics/bty560.

Danecek, P., A. Auton, G. Abecasis, C.A. Albers, E. Banks, et al. 2011. The variant call format and VCFtools. Bioinformatics. 27(15): 2156–2158. doi: 10.1093/bioinformatics/btr330.

Garrison, E., and G. Marth. 2012. Haplotype-based variant detection from short-read sequencing. http://arxiv.org/abs/1207.3907.

Kang, H.M., N.A. Zaitlen, C.M. Wade, A. Kirby, D. Heckerman, et al. 2008. Efficient control of population structure in model organism association mapping. Genetics. 178(3): 1709–1723. doi: 10.1534/genetics.107.080101.

Langmead, B., and S.L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 9(4): 357–359. doi: 10.1038/nmeth.1923.

Langmead, B., C. Wilks, V. Antonescu, and R. Charles. 2019. Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics. 35(3): 421–432. doi: 10.1093/bioinformatics/bty648.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25(16): 2078–2079. doi: 10.1093/bioinformatics/btp352.