Haplotyping interspecific hybrids by dual alignment to both parental genomes

Patrick Brown[1]

[1]University of California Davis

January 18, 2023

Abstract

Sequencing-based genotyping of heterozygous diploids requires sufficient depth to accurately call heterozygous genotypes. In interspecific hybrids, alignment of reads to both parental genomes simultaneously can generate haploid data, potentially eliminating the problem of heterozygosity. Two populations of interspecific hybrid rootstocks of walnut (Juglans) and pistachio (Pistacia) were genotyped using alignment to the maternal genome, paternal genome, and dual alignment to both genomes simultaneously. Downsampling was used to examine concordance of imputed genotype calls as a function of sequencing depth. Dual alignment resulted in datasets essentially free of heterozygous genotypes, simplifying the identification and removal of cross-contaminated samples. Concordance between full and downsampled genotype calls was always highest after dual alignment. Nearly all SNPs in dual alignment datasets were shared with the corresponding single-parent datasets, but 60-90% of single-parent SNPs were private to that dataset. Private SNPs in single-parent datasets had higher rates of heterozygosity, lower levels of concordance, and were enriched in fixed differences between parental genomes ("homeo-SNPs") compared to shared SNPs in the same dataset. In multi-parental walnut hybrids, the paternal-aligned dataset was ineffective at resolving population structure in the maternal parent. Overall, the dual alignment strategy effectively produced phased, haploid data, increasing data quality and reducing cost.

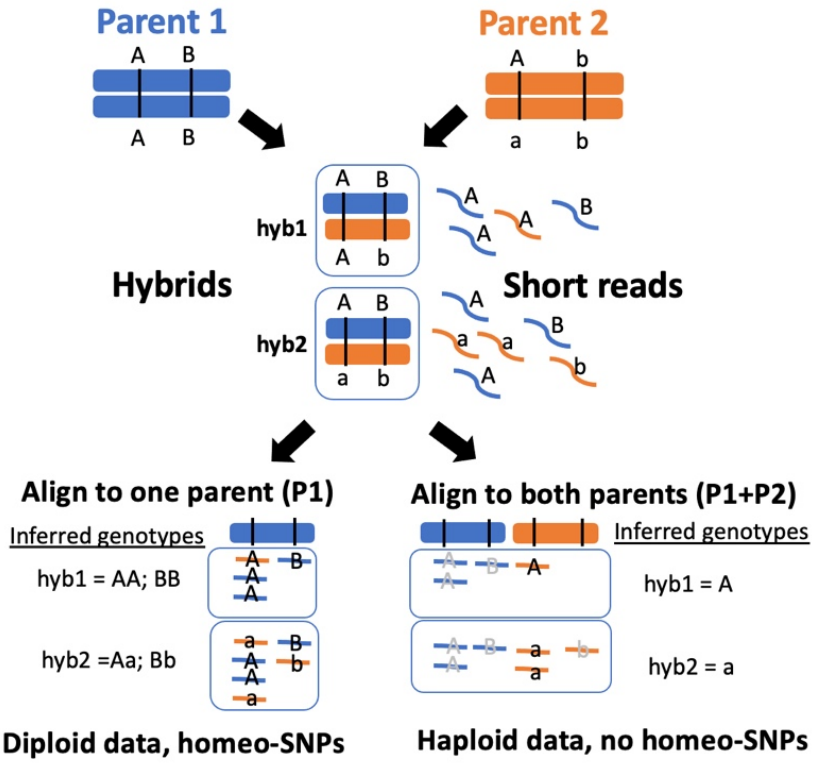
Introduction

The prodigious throughput of short-read sequencing technology has revolutionized quantitative genetics by allowing multiplexed genome-wide genotyping of large numbers of individuals with minimal ascertainment bias (Davey et al., 2011; Andrews et al., 2016). A major technical challenge to this approach is accurate calling of heterozygous genotypes at low sequencing depth. To circumvent this problem, reduced-representation libraries are often generated using restriction enzymes (Baird et al., 2008; Elshire et al., 2011) or sequence capture (Gnirke et al., 2009; Ali et al., 2016), increasing sequencing depth across a subset of the genome. However, haploid or inbred individuals can generally be genotyped and imputed much more accurately and inexpensively than heterozygous individuals (Swarts et al., 2014). A second challenge to genotyping with low-depth short-read data is the possibility of "homeo-SNPs" arising from alignment of reads from homeologous regions of the genome (Tinker et al., 2014; Hulse-Kemp et al., 2015). These false polymorphisms can often be identified by their excess heterozygosity relative to Hardy-Weinberg equilibrium, but homeo-SNPs that escape filtering may interfere with imputation and estimation of relatedness between individuals. Homeo-SNPs are particularly problematic in polyploids and interspecific hybrids.

Polyploidy and interspecific hybridization are common features of plant evolution that are exploited in plant breeding to generate novelty, increase vigor, and stack desirable alleles from different species (Alix et al., 2017). Tree and vine crops often rely on interspecific hybrid rootstocks to increase vigor and resilience to biotic or abiotic stresses without affecting fruit or nut quality in the grafted scion (Warschefsky et al., 2015). In California, for example, production of almonds *(Prunus dulcis)* (Ledbetter and Sisterson, 2008), walnuts*(Juglans regia)* (Ramasamy et al., 2021), and pistachios*(Pistacia vera)* (Ferguson et al., 2002) relies

on rootstocks that are interspecific hybrids. Each of these nut crops has a mating system that can be exploited to generate large numbers of hybrid progeny (self-incompatibility, monoecy, and dioecy, respectively), and superior hybrid genotypes can be propagated clonally. However, genetic gain in tree breeding programs is generally slow due to the time and space required for evaluation, as well as the difficulty of genotyping highly heterozygous material.

This study evaluates different methods for generating genotype data from elite populations of interspecific hybrid pistachio (*P. atlantica X P. integerrima* ; n=725) and walnut (*J. microcarpa X J. regia* ; n=228) rootstocks. Short read sequencing was performed on reduced-representation libraries for each species. A typical workflow would be to align the resulting reads against either the maternal (P1) or paternal (P2) genome (Figure 1). Because interspecific hybrids are composed of one haploid gamete from each parent, we expected that alignment to both parental genomes simultaneously (P1+P2) would result in haploid data, avoiding depth thresholding and greatly increasing genotyping efficiency for heterozygous germplasm.



Materials and Methods

Plant material

Pistachio interspecific hybrid rootstocks consisted of 768 seedlings derived from a single cross of *P. atlantica X P. integerrima*provided by Foundation Plant Services (Davis, CA) and were sequenced in 2 Illumina HiSeq SR100 lanes at the UC Davis Genome Center. Walnut interspecific hybrid rootstocks consisted of 228 clones derived from three crosses (*J. microcarpa* individuals 31.01, 31.09. and 29.11 crossed with pollen from *J. regia* cv "Serr") (Ramasamy et al., 2021) and were sequenced in a single Illumina NextSeqHigh 75 lane.

Genotyping, imputation, and downsampling

GBS libraries were prepared as previously described (Poland et al., 2012) using simultaneous restriction-ligation with HindIII-HF, MseI, and T4 DNA Ligase (NEB). Following the TASSEL GBS pipelineV2 (Glaub-

itz et al., 2014), BWA (Li and Durbin, 2009) was used to align 64 bp tags to reference assemblies for either the maternal parent species (P1), the paternal parent species (P2), or both maternal and paternal reference assemblies simultaneously (P1+P2). Reference assemblies for pistachio species were obtained from Palmer et al. (2022), for *J. microcarpa* from Zhu et al. (2019), and for *J. regia* from Marrano et al. (2020). Only tags that aligned uniquely (MAPQ>=20) were retained. The SNPQualityProfilerPlugin in TASSEL was used to remove candidate SNPs with low depth (log(depth)< -1) and low inbreeding coefficient (F < -0.05 for P1 and P2 alignments; F < 0.9 for P1+P2 alignments) before SNP calling. Vcftools (Danecek et al., 2011) was used to remove taxa with >90% missing data, and for depth thresholding (–minDP 5) of P1 and P2 datasets only. Imputation with Beagle 5.4 (Browning et al., 2018) was performed with no reference panel and a window size and walk speed of 12 and 4 Mb respectively. Downsampling (50%) was performed using the reformat.sh command in bbmap (Bushnell, 2014).

Results and Discussion

Effects of alignment strategy on heterozygosity and minor allele frequency

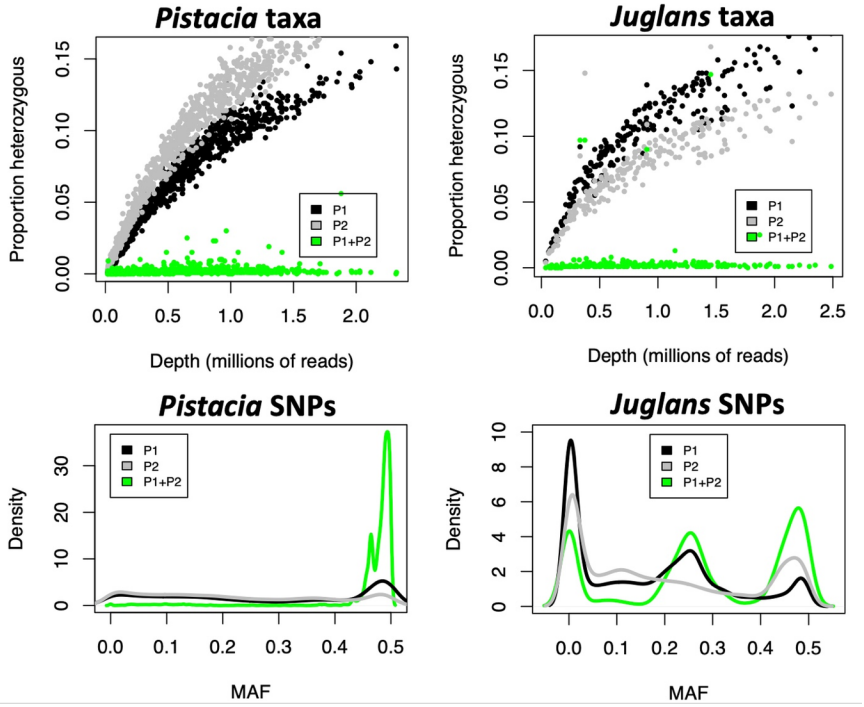
Three genotypic datasets each were generated for both *Pistacia*and *Juglans* , by aligning reads to the maternal genome only (P1), to the paternal genome only (P2), or to both maternal and paternal genomes simultaneously (P1+P2). Dual alignment to both parental genomes resulted in a higher proportion of reads aligning for hybrids of both genera (Table 1). SNPs with low coverage and excess heterozygosity were filtered from each dataset, and a depth threshold of 5 was applied to P1 and P2 datasets prior to imputation to minimize undercalling of heterozygotes. Results are summarized in Table 1. Note that*Juglans* and *Pistacia* have basic chromosome numbers of 16 and 15, and that dual alignment (P1+P2) results in alignment to 32 and 30 chromosomes respectively.

Table 1. Summary of *Juglans* and *Pistacia* datasets produced using three different alignment strategies.

Genus	Alignment strategy (% aligned)	Down-sampling	Taxa	SNPs	Chromo-somes	Avg. MAF	Heterozygote Frequency	Data retained after de thresho ing (%)
Juglans	*microcarpa* (P1; 63%)	none	228	114965	16	0.163	0.066	18.3
		50%	226	113380	16	0.147	0.048	10.2
	Regia (P2; 55%)	none	228	83355	16	0.191	0.093	16.4
		50%	228	83515	16	0.168	0.064	10.7
	Dual (P1+ P2; 76%)	none	224	65517	32	0.29	0.003	100
		50%	222	64607	32	0.28	0.003	100
Pistacia	*atlantica* (P1; 60%)	none	731	33602	15	0.274	0.071	9.5
		50%	714	23984	15	0.265	0.05	4.8
	integerrima (P2; 57%)	none	731	30202	15	0.22	0.094	10.1
		50%	714	21099	15	0.205	0.067	7.7

Genus	Alignment strategy (% aligned)	Down-sampling	Taxa	SNPs	Chromo-somes	Avg. MAF	Heterozygote Frequency	Data retained after de thresho ing (%)
	Dual (P1+P2; 65%)	none	725	13361	30	0.468	0.002	100
		50%	725	11642	30	0.46	0.002	100

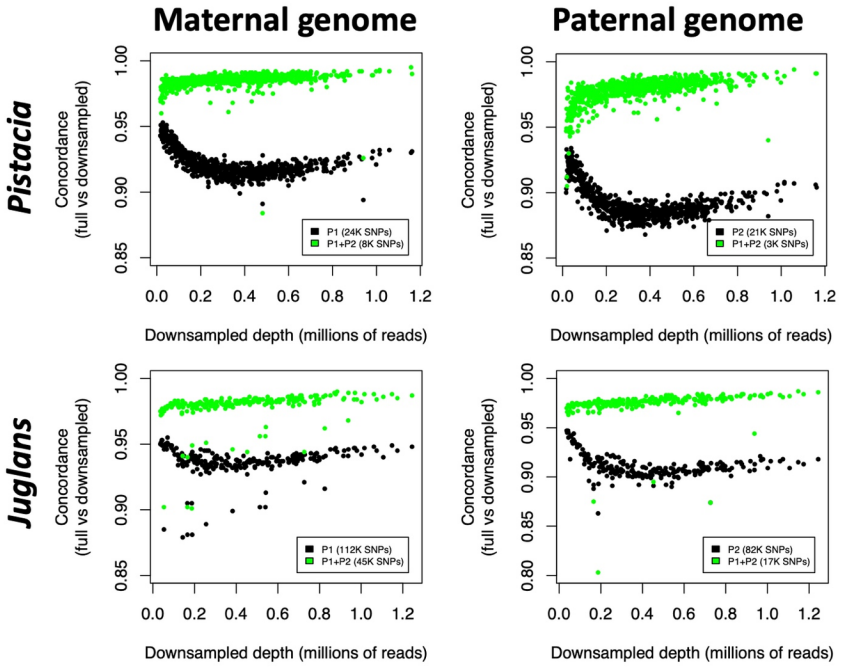
Heterozygote frequencies in *Juglans* and *Pistacia* P1+P2 datasets are 0.003 and 0.002 respectively, 20-50X lower than in their corresponding P1 and P2 datasets (Table 1). The proportion of heterozygous genotypes in P1 and P2 datasets increases strongly with increasing read depth, whereas the proportion in P1+P2 datasets is not affected by read depth (Figure 2). Most heterozygous genotypes in P1+P2 datasets are concentrated in a few individuals, probably due to cross-contamination during DNA extraction or library preparation (Figure 2). Therefore, the dual alignment strategy for interspecific hybrids results in effectively haploid datasets and enables simple detection and removal of cross-contaminated samples.



Average minor allele frequency (MAF) in P1 and P2 datasets is approximately half of that in the corresponding P1+P2 datasets (Table 1). In the *Pistacia* P1+P2 dataset derived from a single cross, almost all SNPs have MAF values close to 0.5. MAF frequencies in the*Juglans* dataset are more variable since this population is derived from three families. Two of these families are at frequencies close to 0.5, and the third is at less than 0.05 frequency. In the*Juglans* P1+P2 dataset, three distinct peaks in the MAF distribution are observed close to 0.5, 0.25, and 0.05. The abundance of SNPs with MAF ˜0.125 in P1 and P2 (but not P1+P2) datasets likely indicates segregation of diploid SNPs in one of the two more frequent Juglans crosses.

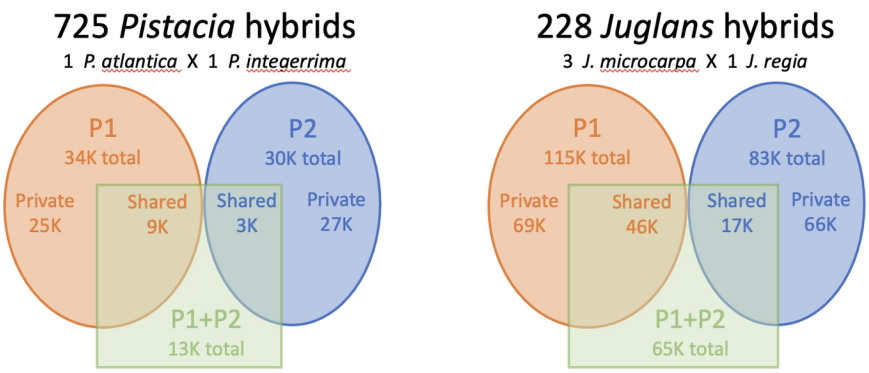
Concordance between genotypes before and after downsampling

To assess the robustness of each alignment strategy to reduced sequencing depth, raw Illumina files were downsampled by 50%, and genotype calls for each taxon from the imputed, downsampled datasets were compared to genotype calls from the full imputed datasets. Concordance between full and downsampled datasets is much higher using dual alignment (P1+P2; Figure 3). Downsampling also results in a one-third reduction in heterozygosity in P1 and P2 datasets but does not affect the proportion of heterozygous genotypes in P1+P2 datasets (Table 1). With increasing sequencing depth, concordance in P1+P2 datasets increases as expected, whereas concordance in P1 and P2 datasets is unexpectedly higher at low depth than at intermediate sequencing depth.



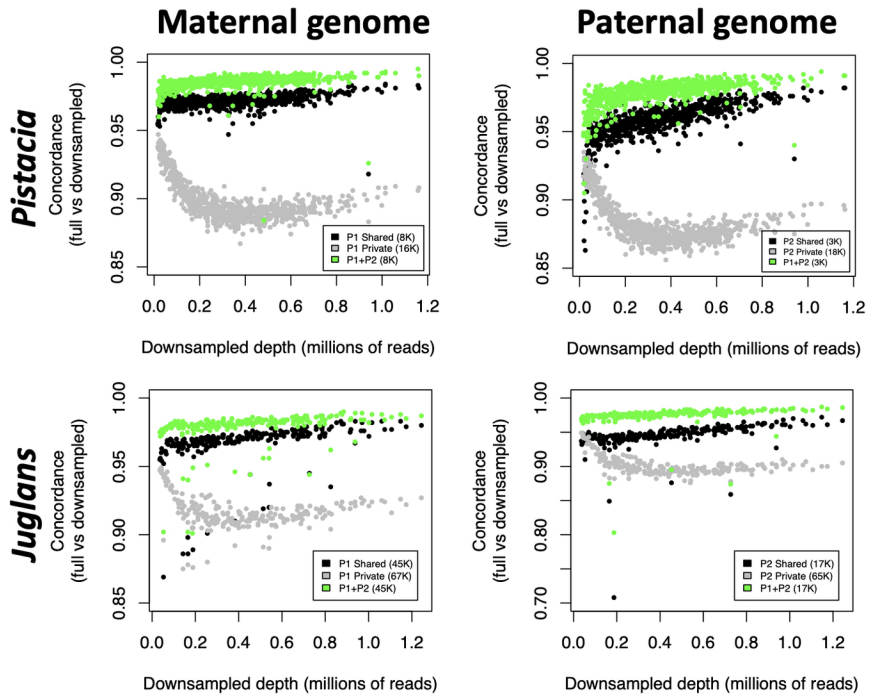
Shared and private SNPs across datasets

Virtually all the SNPs in the P1+P2 dataset for each genus are shared with either the corresponding P1 or P2 dataset (Figure 4). However, the much larger P1 and P2 datasets are composed primarily (60-90%) of private SNPs that are not present in the P1+P2 dataset. To investigate the hypothesis that these private SNPs are enriched with homeo-SNPs, we compared SAM files resulting from alignment of unique reads to P1, P2, and P1+P2 genomes, and associated each SNP with its underlying reads in the SAM files based on position. A multi-mapping index was defined for each SNP as the proportion of underlying unique reads that mapped to both parental genomes. Since homeo-SNPs arise from multi-mapped reads, we hypothesized that the reads underlying private SNPs would display a higher incidence of multi-mapping. Indeed, the mean multi-mapping index was ˜2X higher for private SNPs than for shared SNPs for both parents of both genera (Figure S1), suggesting that private SNPs are enriched with homeo-SNPs.



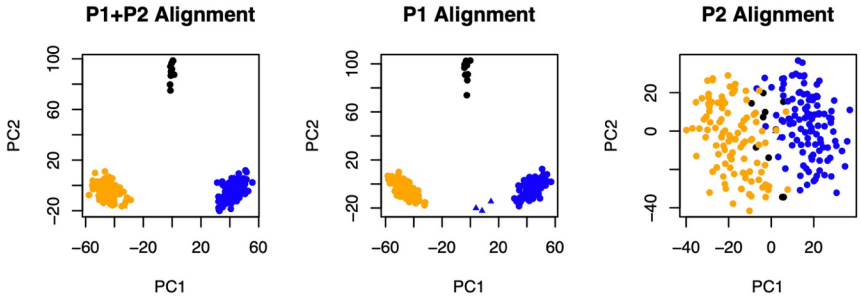
Concordance of shared and private SNPs in P1 and P2 datasets

P1 and P2 SNPs that are shared with the corresponding P1+P2 dataset have much higher levels of concordance than private SNPs in the P1 and P2 datasets (Figure 5). In addition, these shared SNPs show the expected increase in concordance with increasing sequence depth, similar to P1+P2 SNPs, whereas private SNPs show the unexpected increase in concordance at low sequencing depth previously noted in Figure 3. This phenomenon could be due to enrichment with homeo-SNPs, which are truly heterozygous in all individuals but may appear as homozygous when sequencing depth is limited. One limitation of our reduced representation library preparation protocol is that alleles underlying a locus may have different fragment sizes, leading to amplification bias (Davey et al., 2011). In samples with very low sequencing depth, underrepresented reads from the larger fragment may be entirely absent, leading to higher concordance at low depth in the private SNPs.



Alignment to a single parental genome is ineffective at resolving population structure in the other parent

Principal component analysis (PCA) was applied to P1, P2, and P1+P2 datasets to model population structure in the *Juglans* population of hybrids, which consists of three families derived from three maternal parents and a single paternal parent (Figure 6). Whereas P1+P2 and P1 datasets were effective at resolving this population structure, the P2 dataset was not. To interpret this result, we note that the Juglans P2 dataset contains far fewer SNPs at MAF˜0.25 (Figure 2), the frequency we would expect for a diploid SNP showing a fixed difference between the two common maternal parents. It is possible that a mapping quality threshold less stringent than the one used in this study (MAPQ>=20) might have been effective in retaining more genetic signal from the non-aligned parent. Although the PCA plots for P1 and P1+P2 look overall quite similar, the three samples with the fewest sequencing reads (shown as blue triangles in Figure 6) drift towards the origin in the P1 dataset but not the P1+P2 dataset, suggesting the latter is more robust to low sequencing depth.



Future work and similarity with other methods

Inbred or haploid genotypic datasets enjoy huge quality advantages over heterozygous datasets at comparable levels of sequencing depth. This study used a minimum depth threshold of 5 for P1 and P2 datasets, which should theoretically lead to 93.75% of truly heterozygous sites being called correctly (assuming no amplification bias) and which actually resulted in ˜80-90% of the raw data being discarded (Table 1). The luxury of relaxing or removing depth thresholds in inbred datasets results in retention of much more data, and summarizing heterozygosity by taxa or by SNP in inbred datasets simplifies the removal of cross-contaminated DNA samples and homeo-SNPs respectively. In this study, dual alignment of reads from interspecific hybrids to both parental genomes (P1+P2) resulted in effectively inbred datasets that enabled more rigorous quality control, displayed higher concordance following downsampling, and provided more robust estimation of population structure compared to standard alignment against a single reference genome. Although this study used Beagle imputation for purposes of comparing different alignment strategies, datasets resulting from dual alignment could also be imputed using FSFHap, an imputation method designed for inbred populations (Swarts et al., 2014), whereas P1 and P2 datasets could not. The practical conclusion of this study is that dual alignment allows interspecific hybrids to be genotyped and imputed as efficiently and inexpensively as inbreds.

The divergence between parental genomes in this study is estimated at 38 million years for *Pistacia (P. atlantica* vs *P. integerrima)* (Xie et al., 2014) and 45 million years for *Juglans (J. microcarpa* vs *J. regia)* (Stevens et al., 2018). This study used 90 bp Illumina reads trimmed to 64 bp for speedier processing through the TASSEL GBS pipeline (Glaubitz et al., 2014), of which 65% and 76% mapped uniquely to the *Pistacia* and *Juglans* P1+P2 genomes respectively. Longer reads could be used to apply this strategy to hybrids with lower divergence, and perhaps even hybrids between heterotic groups within a species. Alternatively, strategies that make use of a "pan-genome", including the Practical Haplotype Graph (Bradbury et al., 2022), may achieve a similar result by including enough representative reference contigs to ensure that all reads align to a homologous (non-homeologous) sequence. The strategy described here could also be applied to transcriptome data of hybrids to investigate allele-specific or species-specific patterns of expression and co-expression.

Data Availability

Raw sequence data for this project have been submitted to the NCBI SRA under BioProject PRJNA909784. Twelve imputed VCF files representing two genera, three alignment strategies, and two levels of downsampling are included as supplemental material.

Acknowledgments

The author thanks Allison Krill for helpful discussion, and Steven Lee, Kristina McCreery, and Ilean Tracy for assistance with tissue collection and DNA extraction.

Funding

This work was supported in part by USDA NIFA-SCRI grant no. 2018-51181-28437, the California Walnut Board, and the California Pistachio Research Board.

Conflict of Interest

The author declares no conflict of interest.

References

Ali, O.A., S.M. ORourke, S.J. Amish, M.H. Meek, G. Luikart, et al. 2016. RAD Capture (Rapture): Flexible and Efficient Sequence-Based Genotyping. Genetics 202(2): 389–400. doi: 10.1534/genetics.115.183665.

Alix, K., P.R. Gérard, T. Schwarzacher, and J.S. (Pat) Heslop-Harrison. 2017. Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. Ann. Bot. 120(2): 183–194. doi: 10.1093/aob/mcx079.

Andrews, K.R., J.M. Good, M.R. Miller, G. Luikart, and P.A. Hohenlohe. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet. 17(2): 81–92. doi: 10.1038/nrg.2015.28.

Baird, N.A., P.D. Etter, T.S. Atwood, M.C. Currey, A.L. Shiver, et al. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS ONE 3(10). doi: 10.1371/journal.pone.0003376.

Bradbury, P.J., T. Casstevens, S.E. Jensen, L.C. Johnson, Z.R. Miller, et al. 2022. The Practical Haplotype Graph, a platform for storing and using pangenomes for imputation. Bioinformatics 38(15): 3698–3702. doi: 10.1093/bioinformatics/btac410.

Browning, B.L., Y. Zhou, and S.R. Browning. 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am. J. Hum. Genet. 103(3): 338–348. doi: 10.1016/j.ajhg.2018.07.015.

Bushnell, B. 2014. BBMap: A Fast, Accurate, Splice-Aware Aligner. https://escholarship.org/uc/item/1h3515gn (accessed 6 December 2022).

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.

Davey, J.W., P.A. Hohenlohe, P.D. Etter, J.Q. Boone, J.M. Catchen, et al. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12(7): 499–510.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, et al. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 6(5): e19379. doi: 10.1371/journal.pone.0019379.

Ferguson, L., J.A. Poss, S.R. Grattan, C.M. Grieve, D. Wang, et al. 2002. Pistachio Rootstocks Influence Scion Growth and Ion Relations under Salinity and Boron Stress. J. Am. Soc. Hortic. Sci. 127(2): 194–199.

Glaubitz, J.C., T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, et al. 2014. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline (N.A. Tinker, editor). PLoS ONE 9(2): e90346. doi: 10.1371/journal.pone.0090346.

Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E.M. LeProust, et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 27(2): 182–189. doi: 10.1038/nbt.1523.

Hulse-Kemp, A.M., J. Lemm, J. Plieske, H. Ashrafi, R. Buyyarapu, et al. 2015. Development of a 63K SNP Array for Cotton and High-Density Mapping of Intraspecific and Interspecific Populations of Gossypium spp. G3 GenesGenomesGenetics 5(6): 1187–1209. doi: 10.1534/g3.115.018416.

Ledbetter, C.A., and M.S. Sisterson. 2008. Advanced generation peach-almond hybrids as seedling rootstocks for almond: first year growth and potential pollenizers for hybrid seed production. Euphytica 160(2): 259–266. doi: 10.1007/s10681-007-9569-1.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25(14): 1754–1760. doi: 10.1093/bioinformatics/btp324.

Marrano, A., M. Britton, P.A. Zaini, A.V. Zimin, R.E. Workman, et al. 2020. High-quality chromosome-scale assembly of the walnut (Juglans regia L.) reference genome. GigaScience 9(5): giaa050. doi: 10.1093/gigascience/giaa050.

Palmer, W., E. Jacygrad, S. Sagayaradj, K. Cavanaugh, R. Han, et al. 2022. Genome assembly and association tests identify interacting loci associated with vigor, precocity, and sex in interspecific pistachio rootstocks. : 2022.06.28.498047. doi: 10.1101/2022.06.28.498047.

Poland, J.A., P.J. Brown, M.E. Sorrells, and J.-L. Jannink. 2012. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach (T. Yin, editor). PLoS ONE 7(2): e32253. doi: 10.1371/journal.pone.0032253.

Ramasamy, R.K., M.-C. Luo, C.A. Leslie, D. Velasco, N. Ott, et al. 2021. Co-located quantitative trait loci mediate resistance to Agrobacterium tumefaciens, Phytophthora cinnamomi, and P. pini in Juglans microcarpa × J. regia hybrids. Hortic. Res. 8(1): 1–11. doi: 10.1038/s41438-021-00546-7.

Stevens, K.A., K. Woeste, S. Chakraborty, M.W. Crepeau, C.A. Leslie, et al. 2018. Genomic Variation Among and Within Six *Juglans* Species. G3amp58 GenesGenomesGenetics: g3.200030.2018. doi: 10.1534/g3.118.200030.

Swarts, K., H. Li, J.A. Romero Navarro, D. An, M.C. Romay, et al. 2014. Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. Plant Genome 7(3). https://dl.sciencesocieties.org/publications/tpg/abstracts/7/3/plantgenome2014.05.0023 (accessed 4 January 2015).

Tinker, N.A., S. Chao, G.R. Lazo, R.E. Oliver, Y.-F. Huang, et al. 2014. A SNP Genotyping Array for Hexaploid Oat. Plant Genome 7(3): plantgenome2014.03.0010. doi: 10.3835/plantgenome2014.03.0010.

Warschefsky, E.J., L.L. Klein, M.H. Frank, D.H. Chitwood, J.P. Londo, et al. 2015. Rootstocks: Diversity, Domestication, and Impacts on Shoot Phenotypes. Trends Plant Sci. doi: 10.1016/j.tplants.2015.11.008.

Xie, L., Z.-Y. Yang, J. Wen, D.-Z. Li, and T. Yi. 2014. Biogeographic history of Pistacia (Anacardiaceae), emphasizing the evolution of the Madrean-Tethyan and the eastern Asian-Tethyan disjunctions. Mol. Phylogenet. Evol. 77. doi: 10.1016/j.ympev.2014.04.006.

Zhu, T., L. Wang, F.M. You, J.C. Rodriguez, K.R. Deal, et al. 2019. Sequencing a Juglans regia × J. microcarpa hybrid yields high-quality genome assemblies of parental species. Hortic. Res. 6(1): 55. doi: 10.1038/s41438-019-0139-1.

Supplemental Figures.