



Dr. Susan C. Miyasaka

# Use of pseudo-reference genomes to improve genotyping-by-sequencing of taro (*Colocasia esculenta*)

Thomas K. Wolfgruber<sup>1</sup>, Martin Helmkampf<sup>2</sup>, Michael B. Kantar<sup>1</sup>, M. Renee Bellinger<sup>2</sup>, Susan C. Miyasaka<sup>1</sup>, and Michael Shintaku<sup>2</sup>



<sup>1</sup>University of Hawai'i at Manoa, Dept. of Tropical Plant & Soil Sciences and <sup>2</sup>University of Hawai'i at Hilo

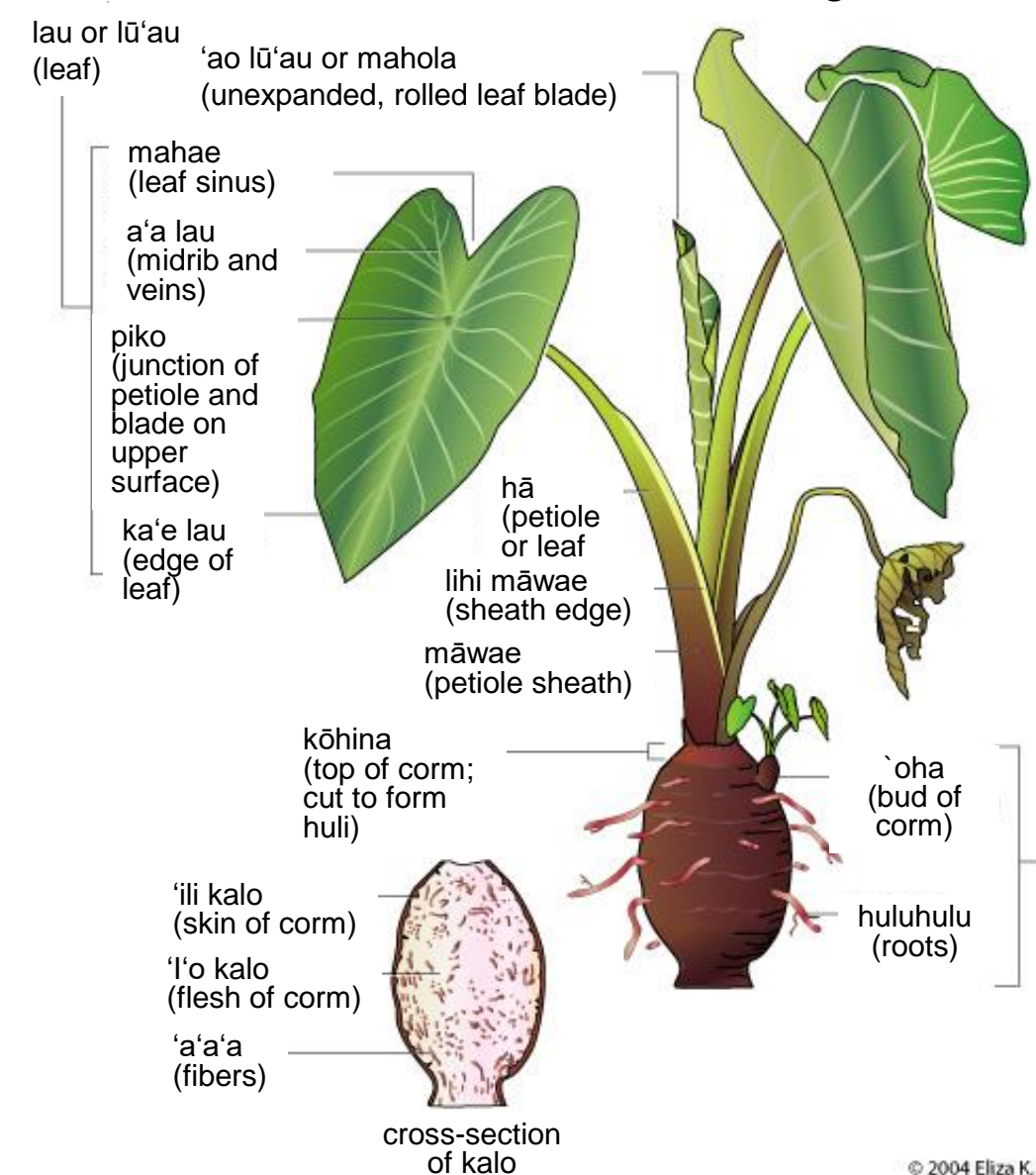
## INTRODUCTION

**Taro** (*Colocasia esculenta*) is a food staple in the Pacific. In Hawaii, 84 cultivars of taro have been classified using morphological features (Whitney, et al. 1939), but genetic relationships between cultivars are unknown. Here, Hawaiian and Pacific cultivars are analyzed for single nucleotide polymorphisms using genotyping-by-sequencing (GBS) protocols (Elshire, et al. 2011). Since taro lacks a reference genome sequence, we employed pseudo-reference genomes to evaluate SNP diversity. We found that our pseudo-reference genome methods yielded a greater number of SNPs relative to the reference-free UNEAK (Lu, et al. 2013) GBS method. Results for 70 "pure" (expected non-hybrid) taro are used to explore relationships between morphological and genomic characteristics.



Taro grown in Waipio Valley, Hawaii Island

### The corm of taro is eaten after cooking



### Corm color differences between siblings

(identical numbers are siblings from the same parents)



## METHODS

### Taro samples and GBS

- Sampled 189 non-hybrid ("pure") and hybrid taro individuals
- GBS done with *Pst*I endonuclease, single-end Illumina sequencing, and TASSEL (Bradbury, et al. 2007)
- Visualization 3D plot made with R (<https://www.r-project.org/>) and all 189 samples
- SNP counts / population analysis done on 70 "pure" samples

### UNEAK vs. pseudo-reference GBS runs

- UNEAK reference-free GBS pipeline run through TASSEL and used all 189 samples; results for a subset of 70 "pure" samples was used for SNP counts / population analysis
- Reference-based GBSs were done with TASSEL using bowtie2 (Langmead and Salzberg 2012) and a pseudo-reference
- GBS pseudo-reference: GBS sequencing reads from 95 samples (single GBS plate including the "pure" 70) were assembled using IDBA-UD (Peng, et al. 2012)
- RAD+GBS pseudo-reference: RAD-Seq (Davey and Blaxter 2010) paired-end reads for 2 samples and GBS reads from the 189 samples were assembled using IDBA-UD
- Transcriptome pseudo-reference: paired-end Illumina RNA-Seq available through NCBI (You, et al. 2015) were assembled using Trinity (Grabherr, et al. 2011); only longest isoforms were used

### Population assignments using Structure

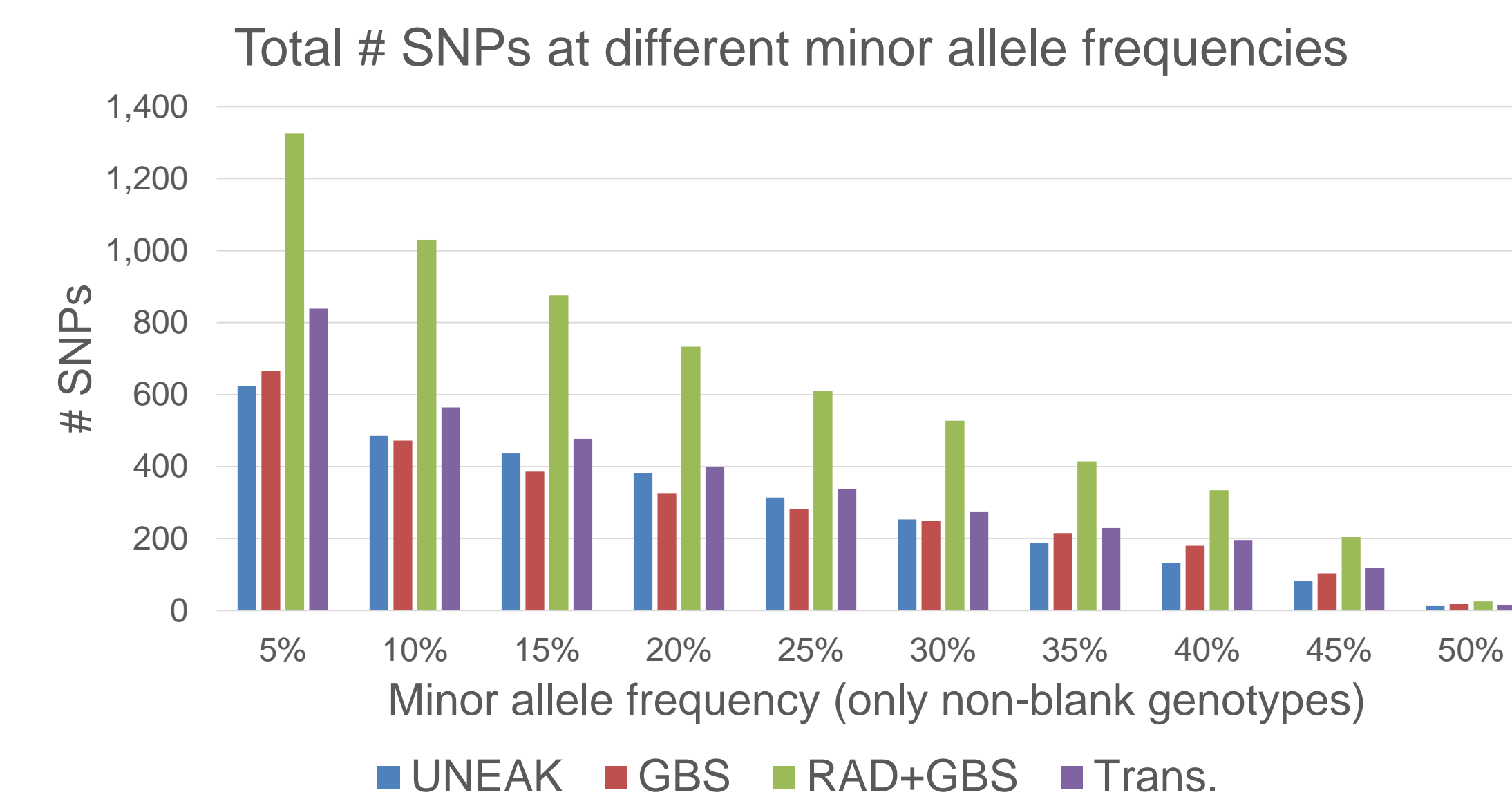
- Structure (Pritchard, et al. 2000) was run on the 70 "pure" samples over K clusters 2 through 9
- Structure Harvester (Dent and vonHoldt 2012, Evanno, et al. 2005) was used to select the median/mode optimal K clusters
- Clumpak (Kopelman, et al. 2015) was used to visualize population assignments and admixture
- Initial group assignments in visualizations set to morphological classifications made by Whitney, et al. 1939

### Trees

- Phylogenetic trees were constructed for each dataset with RAxML (Stamatakis 2014)
- Based on the GTR+G model and 250 bootstrap replicates

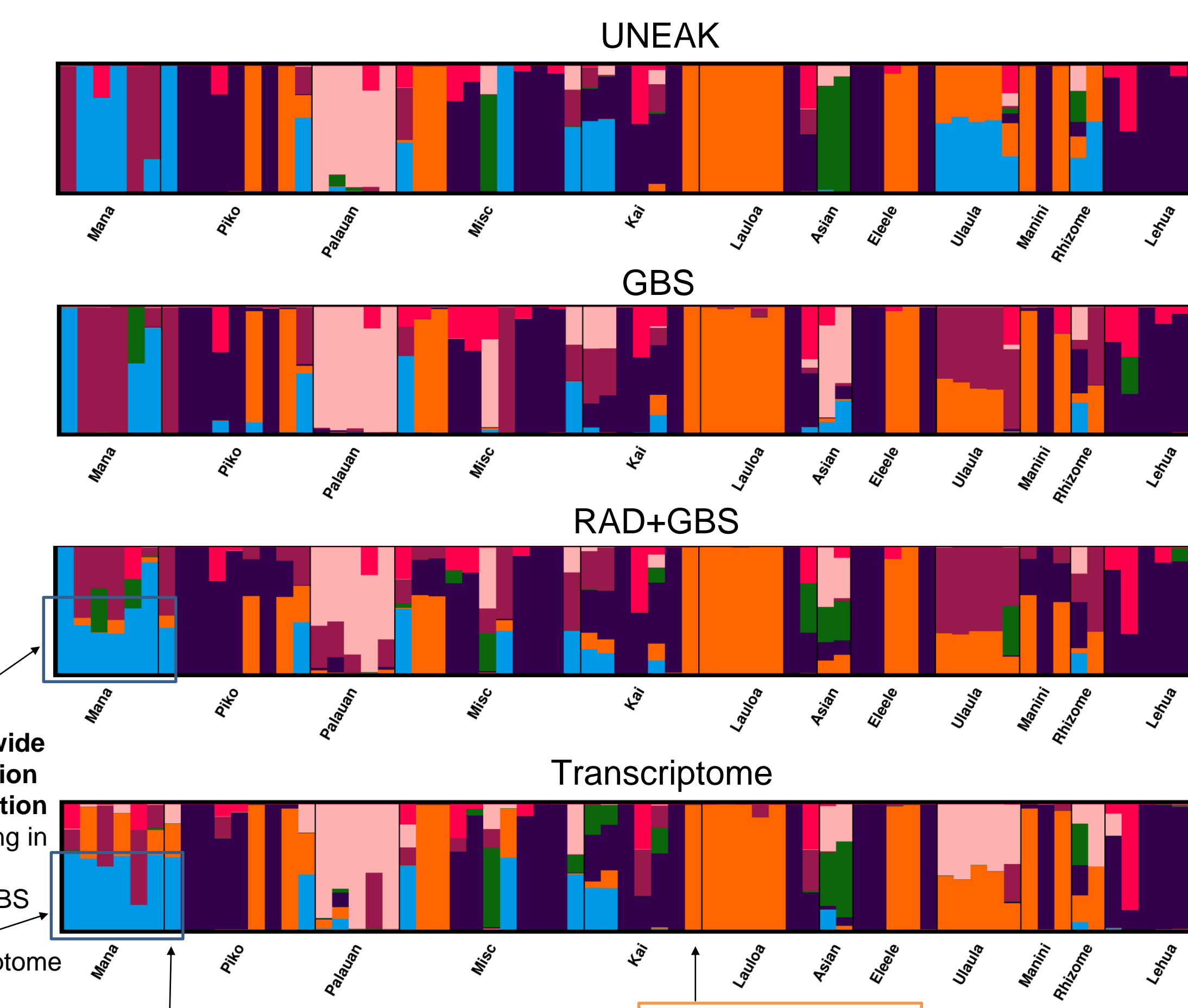
## RESULTS

### Higher # of SNPs with RAD+GBS pseudo-reference



### Population classifications

(minor allele frequency  $\geq$  10%; K clusters = 7; groups as described in Whitney et al., 1939)



"Mana" group-wide population association appearing in the RAD+GBS and transcriptome results

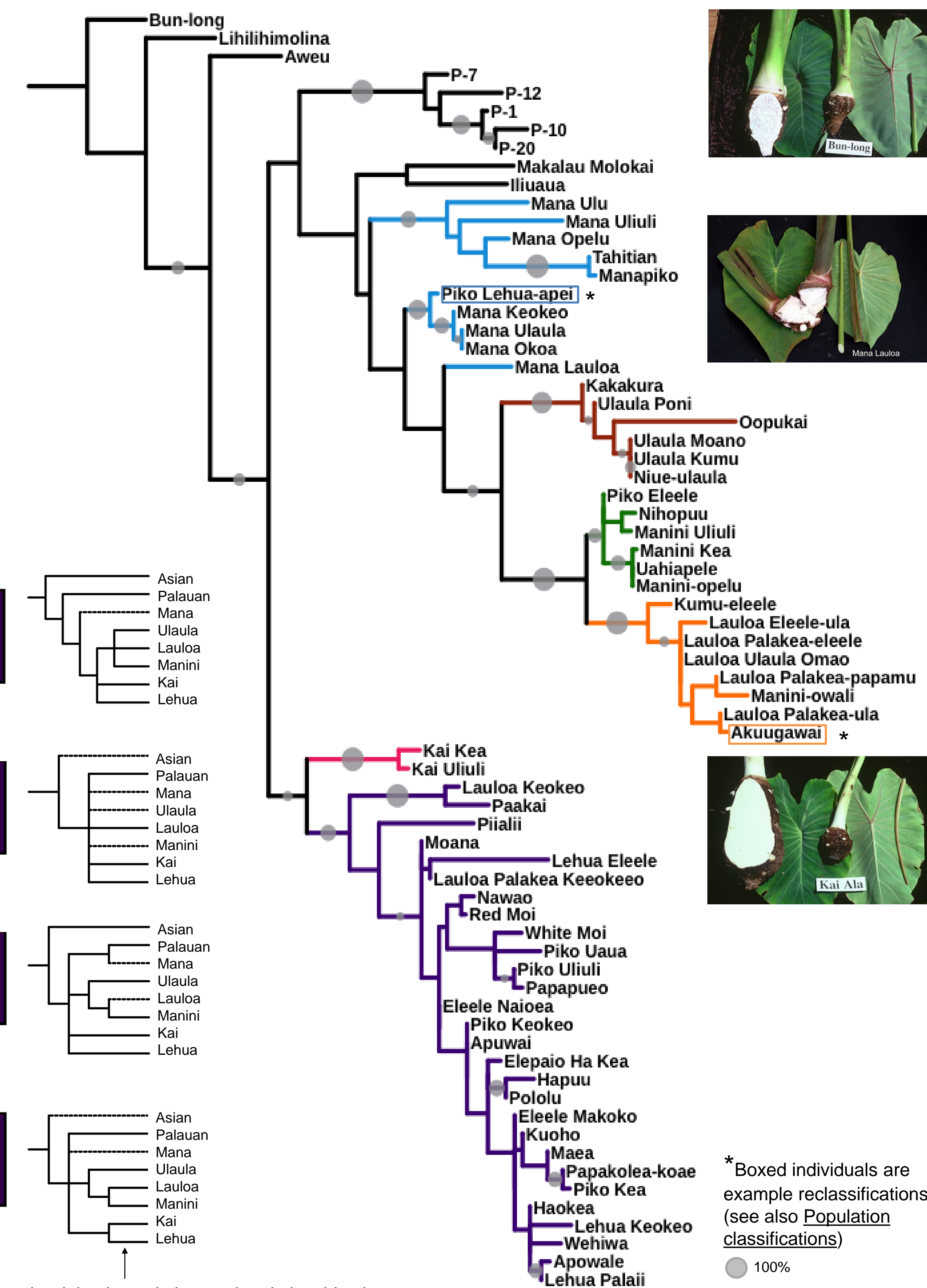
"Piko" individual that appears to be "Mana" (Piko Lehua-apei)

"Kai" individual that appears to be "Lauloa" (Akuugawai)

Cladograms on the right show phylogenetic relationships between groups based on each dataset. Relationships supported by less than 50% of bootstrap replicates are indicated by polytomies (between groups) and dashed lines (within groups). The RAD+GBS and transcriptome-based datasets provided better resolved trees, measured visually and by average bootstrap support.

### Phylogenetic analyses

(transcriptome dataset with minor allele frequency  $\geq$  10%; size of grey circles corresponds to bootstrap support values 50–100%)



### Reclassified groups

Asian (outgroup)

Palauan

Mana (here paraphyletic, monophyletic according to other datasets)

Ulaula

Manini

Lauloa

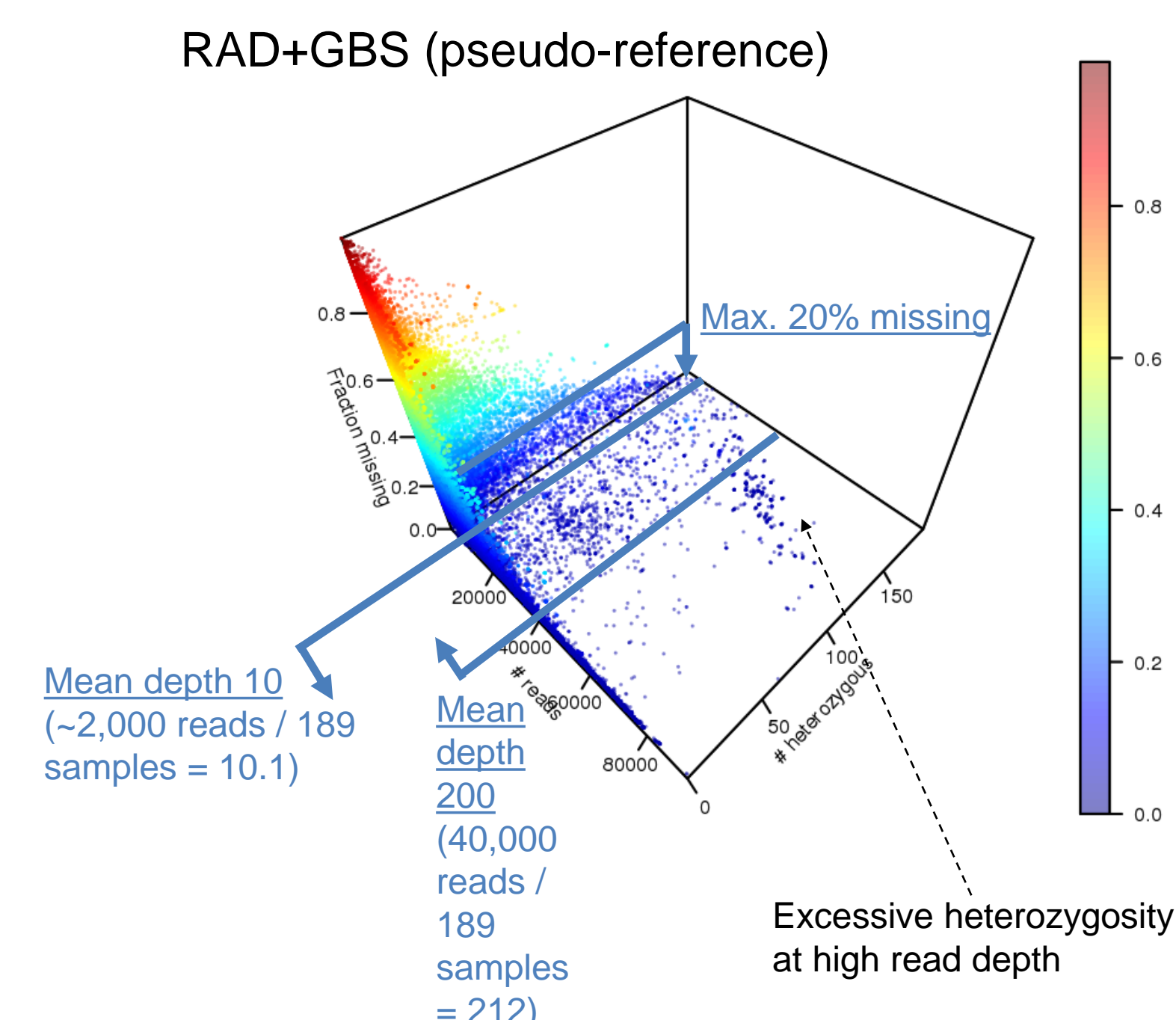
Kai

Lehua

\* Boxed individuals are example reclassifications (see also Population classifications)

<http://www.ctahr.hawaii.edu/STHTarot.aspx>

### Heterozygosity used as a proxy for repeat filtering



## CONCLUSIONS

- GBS with a pseudo-reference increases total # SNPs
- GBS with RAD+GBS or transcriptome pseudo-references improves quality of SNPs
- Hawaiian taro cultivars fall into 5 to 6 generally well supported clades
- Most clades are consistent with morphological classification (Ulaula, Manini, Lauloa, Kai)
- Some groups are not supported genetically (Piko) or more inclusive than suggested by morphology (Lehua)

## ACKNOWLEDGEMENTS

This project was partially funded by USDA National Institute of Food and Agriculture Hatch project #08029-H, managed by the College of Tropical Agriculture and Human Resources. In addition, this project was partially funded by USDA AFRI project #2015-05806. We would like to acknowledge Mr. Christopher Bernabe, Agricultural Technician, who conducted the conventional breeding of taro to produce the progeny on which GBS analysis was performed in one plate. This material is also based upon work supported by the National Science Foundation under Grant No. 1345247. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## REFERENCES

- Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* 23:2633-2635.
- Davey JW, Blaxter ML. 2010. RADSeq: next-generation population genetics. *Brief Funct Genomics*. Dec;9(5-6):416-23.
- Earl DA, vonHoldt BM. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* vol. 4 (2):359-61.
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*. May 4;6(5):e19379.
- Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* 14:2811-2820.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mucic E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman R, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011 May 15;29(7):644-52.
- Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. 2015. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol Ecol Resour*. Sep;15(5):1179-91.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. Mar 4;9(4):357-9.
- Lu F, Lipka AE, Glaubitz J, Elshire R, Cherny JH, Casler MD, Buckler ES, Costich DE. 2013. Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. *PLoS Genet* (9):1: e1003215.
- Peng Y, Leung HC, Yiu SM, Chin FY. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*. Jun 1;28(11):1420-8.
- Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics*. 2000 Jun;155(2):945-59.
- Stamatakis. 2014. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics*, 10.1093/bioinformatics/btu033.
- Whitney LD, Bowers FAI, Takahashi M. 1939. Taro varieties in Hawaii. *Bulletin No. 84*. Honolulu: Hawaii Agricultural Experiment Station, 86 p.
- You Y, Liu D, Liu H, Zheng X, Diao Y, Huang X, Hu Z. 2015. Development and characterisation of EST-SSR markers by transcriptome sequencing in taro (*Colocasia esculenta* (L.) Schott). *Molecular Breeding* June, 35:134.