Supporting Information for online publication

Journal: Molecular Ecology Resources

Title: A genome-wide SNP genotyping resource for tropical pine tree species

Authors:

Colin Jackson[1]*, Nanette Christie[2]*, Melissa Reynolds[2], Christopher Marais[2], Yokateme Tii-kuzu[2], Madison Caballero[3], Tamanique Kampman[2], Erik A. Visser[2], Sanushka Naidoo[2], Dominic Kain[4], Ross W. Whetten[1], Fikret Isik[1], Jill Wegrzyn[3], Gary Hodge[1], Juan Jose Acosta[1#], Alexander A. Myburg[2#]

1. Department of Forestry & Environmental Resources, North Carolina State University, Raleigh, NC, USA. 2. Department of Biochemistry, Genetics and Microbiology, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, 0002, South Africa. 3. Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, USA. 4. HQPlantations Pty Ltd, North Lakes, LD, 4509, Australia.

* These authors contributed equally. # Co-corresponding authors

Corresponding authors:

Juan Jose Acosta – jjacosta@ncsu.edu

Alexander A. Myburg – zander.myburg@fabi.up.ac.za

Table of Contents

Supplementary Methods

Supplementary Method 1. Altering of the *P. taeda* genome assembly (v2.01) to create the tropical pine reference genome.

Using SNP calls generated from a pilot study of the current study (unpublished data), v2.01 of the *P. taeda* genome assembly was modified to create the modified reference used in this study. The modified reference was created by switching the alternative and reference alleles at any location where the alternative allele appeared in more than 60% of the total number of observations. The alternative allele coverage was calculated per SNP across all provenances by [AO/(AO+RO)], where AO is the number of alternative observations and RO is the number of reference observations. Approximately 25K bases were changed. These changes were made in order to more accurately reflect what a "tropical pine" genome might look like. This effectively removed a number of reference SNPs that could interfere with the read mapping phase of the pipeline and reduced noise in the dataset.

Supplementary Method 2. Filtering of capture-seq SNP variants prior to probe design.

Using two methods, the set of raw capture-seq derived SNPs were filtered into two lists. The top-down procedure filtered SNPs across all samples based on the following criteria: an overall alternative allele fraction (AAF) > 0.1 and overall coverage > 100X. The bottom-up procedure used the following criteria: maximum within-sample AAF > 0.2 and maximum within-sample alternative observation (AO) > 4. If a SNP was in either post-filter dataset, it was selected for probe design.

Supplementary Method 3. Filtering of SNP probe sets for inclusion on the screening array.

SNP probe sets included on the screening array were prioritized based on Thermo Fisher's *in sillico* scoring recommendations along with other scoring metrics and sequencing statistics. All probe sets were scored multiple times, i.e. each probe set was scored against each reference transcriptome assembly and the tropical pine reference genome. SNP probe sets were placed into nine categories based on our selection emphasis, with lower numbered categories being priority for inclusion on the array. Category 1 probe sets were probes derived from targeted capture-seq data that were scored as recommended or neutral when scored against the genome reference. Category 2 probe sets were transcriptomic probes that were scored recommended or neutral when scored against their respective transcriptome reference and genome reference. Category 3 probe sets were transcriptome derived probes that were recommended when scored against their transcriptome reference but not recommended when scored against the genome. Category 4 probe sets were transcriptome derived probes, where the scaffold of origin was not present in category 2, that were recommended when scored against their transcriptome reference and not against the genome. Category 5 probe sets were targeted capture-seq probes derived from *P. caribaea* samples. These were not recommended when scored against the genome, but were selected if they had a homology count < 10, global depth > 100, and had an alternative allele fraction (AAF) between 0.05 and 0.95. Category 6 probe sets were targeted capture-seq derived probes selected to be ancestry informative. These were not recommended when scored against the genome reference, but were selected if they had a homology count < 20, an AAF > 0.9 at the subdivision level with an AAF < 0.05 in all other species/subdivisions, and an alternative observation (AO) > 4. Category 7 probe sets were probes that had confirmed utility in *P. elliottii*. Category 8 were transcriptome derived probes that had their scaffold origins represented in category 2 and were recommended when scored against their transcriptome reference. Category 9 probe sets were targeted capture-seq derived probes that were not recommended when scored against the genome reference, but had a homology count < 10, global depth > 100, and had an AAF between 0.05 and 0.95 in at least one subdivision.

Supplementary Tables

Supplementary Table 1. Geographic and climatic information for the 81 provenances representing the natural ranges of six tropical pine species in Mexico and Central America. A total number of 567 trees were sampled and pooled per provenance. Targeted capture sequencing was performed on the pooled samples for SNP discovery.

Species / Subdivision	Provenance	Country	Lat (N)	Long (W)	Elevation (m)	Rainfall (mm)	Number of trees in pooled samples
P. oocarpa	Dipilto	Nicaragua	13° 43'	86° 32'	1075 - 1320	--	8
	San Luis Jilotepeque	Guatemala	14° 37'	89° 46'	950 - 1010	895	8
	Camotán	Guatemala	14° 49'	89° 22'	740 - 960	926	6
	El Castaño (Bucaral)	Guatemala	15° 01'	90° 09'	930 - 1330	900	8
	San Sebastian Coatlán	Mexico	16° 11'	96° 50'	1750 - 1750	598	6
	El Jícaro	Mexico	16° 32'	94° 12'	1000 - 1000	1684	8
	San Pedro Solteapán	Mexico	18° 15'	94° 51'	602 - 602	1812	8
	Taretan	Mexico	19° 25'	102° 04'	1610 - 1610	1622	6
	Huayacocotla	Mexico	20° 30'	98° 25'	1190 - 1410	1711	4
	La Petaca	Mexico	23° 25'	105° 48'	1560 - 1710	1155	8
	Chinipas	Mexico	27° 18'	108° 35'	1140 - 1780	822	5
P. patula var. patula	Potrero de Monroy	Mexico	20° 24'	98° 25'	2320 - 2480	1350	8
	Corralitla	Mexico	18° 38'	97° 06'	2000 - 2230	2500	8
	Conrado Castillo	Mexico	23° 56'	99° 28'	1500 - 2060	1012	8
	Tlacotla	Mexico	19° 40'	98° 05'	2750 - 2915	1097	8
	Pinal de Amoles	Mexico	21° 07'	99° 41'	2380 - 2550	1350	8

	Zacualtipán	Mexico	20° 39'	98° 40'	1980 - 2200	2047	8
	Llano de las Carmonas	Mexico	19° 48'	97° 54'	2530 - 2880	1097	8
P. patula var. longipedunculata	El Manzanal	Mexico	16° 06'	96° 33'	2350 - 2660	1348	8
	El Tlacuache	Mexico	16° 44'	97° 09'	2300 - 2620	2000	8
	Ixtlán	Mexico	17° 24'	96° 27'	2600 - 2870	1750	8
	Santa María Papalo	Mexico	17° 49'	96° 48'	2270 - 2720	1100	8
P. tecunumanii High Elevation	Celaque	Honduras	14° 33'	88° 40'	1540 - 2030	1273	4
	Chanal	Mexico	16° 42'	92° 23'	2010 - 2350	1238	8
	Chempil	Mexico	16° 45'	92° 25'	2020 - 2220	1146	8
	Chiul	Guatemala	15° 20'	91° 04'	2440 - 2680	1996	8
	El Carrizal	Mexico	15° 24'	92° 18'	2130 - 2280	2000	8
	Jitotol	Mexico	17° 02'	92° 51'	1660 - 1750	1701	4
	Juquila	Mexico	16° 15'	97° 13'	2090-2260	1325	8
	Las Trancas	Honduras	14° 07'	87° 49'	2075 - 2185	1579	5
	Montebello	Mexico	16° 06'	91° 45'	1660 - 1750	1909	8
	Napite	Mexico	16° 34'	92° 19'	2070 - 2350	1350	8
	Pachoc	Guatemala	14° 51'	91° 16'	2000 - 2500	1350	8
	Rancho Nuevo	Mexico	16° 41'	92° 35'	2280 - 2340	1238	8
	San Jerónimo	Guatemala	15° 03'	90° 18'	1620 - 1850	1200	8
	San José	Mexico	16° 42'	92° 41'	2245 - 2400	1252	7
	San Lorenzo	Guatemala	15° 05'	89° 40'	1900 - 2100	1700	5
	San Vicente	Guatemala	15° 05'	90° 07'	1690 - 2200	1700	4
P. tecunumanii Low Elevation	Cerro Cusuco	Honduras	15° 29'	88° 12'	1350 - 1630	2287	8
	Cerro la Joya	Nicaragua	12° 25'	85° 59'	940 - 1160	1394	4

	Culmí	Honduras	15° 08'	85° 36'	400 - 950	1491	8
	Gualaco	Honduras	15° 03'	86° 08'	600 - 800	1491	8
	Los Planes	Honduras	14° 48'	87° 53'	1100 - 1650	2287	5
	San Francisco	Honduras	14° 57'	86° 07'	900 - 1590	1491	8
	San Rafael del Norte	Nicaragua	13° 14'	86° 08'	910 - 1170	1366	8
	Villa Santa	Honduras	14° 12'	86° 17'	800 - 1000	1302	8
	Yucul	Nicaragua	12° 55'	85° 47'	1080 - 1330	1394	8
P. greggii North	El Madroño	Mexico	21° 16'	99° 10'	1500 - 1660	1100	8
	Jamé	Mexico	25° 21'	100° 37'	2500 - 2590	650	8
	Laguna Seca	Mexico	21° 02'	99° 10'	1750 - 1900	820	8
	Las Placetas	Mexico	24° 55'	100° 11'	2370 - 2520	750	6
	Las Placetas	Mexico	24° 55'	100° 11'	2370 - 2520	750	4
	Los Lirios	Mexico	25° 22'	100° 29'	2300 - 2400	650	8
	Jamé	Mexico	25° 21'	100° 37'	2500 - 2590	650	5
	Cerro El Potosí	Mexico	24° 54'	100° 12'	2430 - 2500	750	4
	Ojo de Agua	Mexico	24° 53'	100° 13'	2115 - 2400	750	8
	La Tapona	Mexico	24° 37'	100° 10'	2090 - 2350	650	8
P. greggii South	Magueyes	Mexico	--	--	2250 - 2350	1200	8
	El Madroño	Mexico	21° 16'	99° 10'	1500 - 1660	1100	8
	Laguna Atezca	Mexico	20° 49'	98° 46'	1250 - 1420	1642	4
	Laguna Seca	Mexico	21° 02'	99° 10'	1750 - 1900	820	8
	Valle Verde	Mexico	21° 29'	99° 10'	1150 - 1250	1400	8
	San Joaquín	Mexico	20° 56'	99° 34'	2130 - 2350	1109	8
	Carrizal Chico	Mexico	20° 26'	98° 20'	1360 - 1770	1855	8
P. caribaea	Gualjoco	Honduras	14° 55'	88° 14'	240 - 355	1200	8

	Santa Cruz de Yojoa	Honduras	14° 53'	86° 56'	530 - 720	2758	7
	Limón	Honduras	15° 51'	85° 23'	20 - 85	2452	5
	Trincheras	Guatemala	15° 27'	89° 03'	150 - 500	2000	4
P. maximinoi	Cobán	Guatemala	15° 28'	90° 24'	1300 - 1440	2075	8
	San Jerónimo	Guatemala	15° 03'	90° 15'	1280 - 1860	970	8
	San Juan Sacatepéquez	Guatemala	14° 41'	90° 38'	1580 - 2000	1138	8
	Dulce Nombre de Copán	Honduras	14° 50'	88° 51'	1100 - 1300	1386	8
	Marcala	Honduras	14° 10'	88° 01'	1600 - 1800	1670	5
	Tatumbla	Honduras	14° 02'	87° 07'	1400 - 1600	1153	8
	San Jerónimo CH	Mexico	17° 03'	92° 08'	940 - 1020	1417	4
	San Jerónimo OA	Mexico	16° 10'	97° 00'	1220 - 1480	1350	5
	Candelaria	Mexico	16° 00'	96° 31'	1370 - 1480	1117	8
	Las Compuertas	Mexico	17° 10'	99° 59'	1050 - 1200	1400	5
	El Portillo	Honduras	14° 28'	89° 01'	1400 - 1600	1325	8
	Yuscarán	Honduras	13° 50'	86° 55'	1500 - 1700	1300	8
	La Lagunilla	Guatemala	14° 42'	89° 57'	1540 - 1860	1017	4

Supplementary Table 2. Summary of the diagnostic SNP probe sets (Set C2 in Figure 3) included on Pitro50K. For the query species (column 1), SNP probe sets with a MAF higher than the MAF cut-off in column 2 and a MAF lower than the MAF cut-off in column 4 for the species (or group of species) in column 3 were identified. The aim was to identify more or less 200 diagnostic SNP probe sets per species, where possible. A total set of 1,564 unique SNP probe sets were identified as diagnostic.

Query Species / Subdivision	**MAF[†] more than:**	**Species / Subdivision**	**MAF[†] less than:**	**Number of SNPs**
P. oocarpa	0.61	All other species	0.1	104
P. oocarpa	0.75	*P. patula, P. tecunumanii*	0.1	43
P. oocarpa	0.75	*P. tecunumanii*	0.1	64
P. oocarpa	0.89	*P. patula*	0.1	101
P. patula	0.6	All other species	0.1	63
P. patula var. *longipedunculata*	0.5	All other species	0.3	5
P. patula var. *patula*	0.6	*P. patula* var. *longipedunculata*	0.4	11
P. patula var. *longipedunculata*	0.5	*P. patula* var. *patula*	0.3	14
P. tecunumanii	0.6	All other species	0.3	50
P. tecunumanii High Elevation	0.5	All other species	0.3	30
P. tecunumanii Low Elevation	0.5	All other species	0.2	113
P. tecunumanii High Elevation	0.7	*P. tecunumanii* Low Elevation	0.25	50
P. tecunumanii Low Elevation	0.73	*P. tecunumanii* High Elevation	0.2	37
P. greggii	0.95	All other species	0.1	87
P. greggii North	0.5	All other species	0.1	34
P. greggii North	0.8	*P. greggii* South	0.2	16
P. greggii South	0.65	All other species	0.2	48
P. greggii South	0.73	*P. greggii* North	0.2	47
P. caribaea	0.86	All other species	0.1	118
P. elliottii	0.79	All other species	0.1	116
P. caribaea	0.93	*P. elliottii*	0.07	128
P. elliottii	0.95	*P. caribaea*	0.05	130
P. maximinoi	0.85	All other species	0.1	108
P. maximinoi	0.9	*P. pseudostrobus*	0.1	94
P. pseudostrobus	0.85	*P. maximinoi*	0.15	162
P. pseudostrobus	0.79	All other species	0.1	116

[†] At the subdivision level, the minor allele frequency (MAF) was calculated by tracking the global minor allele across all species clustering runs, and then calculating the MAF on a scale from 0 to 1 for each marker per subdivision.

Supplementary Table 3. Sequencing metrics generated from targeted capture sequencing. Raw reads underwent quality control trimming, requiring a base quality score of 30 and minimum read length of 50bp. Average read length decreased slightly from the 150bp raw reads generated from sequencing to ~143bp.

Species	Raw reads (millions)	Raw bases (millions)	Trimmed reads (millions)	Trimmed bases (millions)
P. caribaea	4.6	693	4	588
P. greggii	5.3	795	4.58	652
P. maximinoi	4.92	737	4.2	598
P. oocarpa	4.83	724	4.12	586
P. patula	5.1	776	4.38	622
P. tecunumanii	5.27	791	4.52	641

Supplementary Table 4. Summary of the numbers of SNP markers identified and SNP probe sets successfully designed from RNA sequencing per species.

Metrics	*P. greggii*	*P. maximinoi*	*P. oocarpa*	*P. patula*	*P. tecunumanii*
SNPs identified	680 K	1,110 K	1,030 K	1,390 K	1,110 K
Probes designed	175 K	293 K	301 K	290 K	259 K

Supplementary Table 5. Annotation of each of the 49,694 SNP probe sets tiled on the Pitro50K genotyping array. This table also indicates whether each SNP probe set (i) has membership in the respective Sets defined in Figure 3 and (ii) is informative in each of the species or subdivisions.

External file: < Suppl_Table5.xlsx >

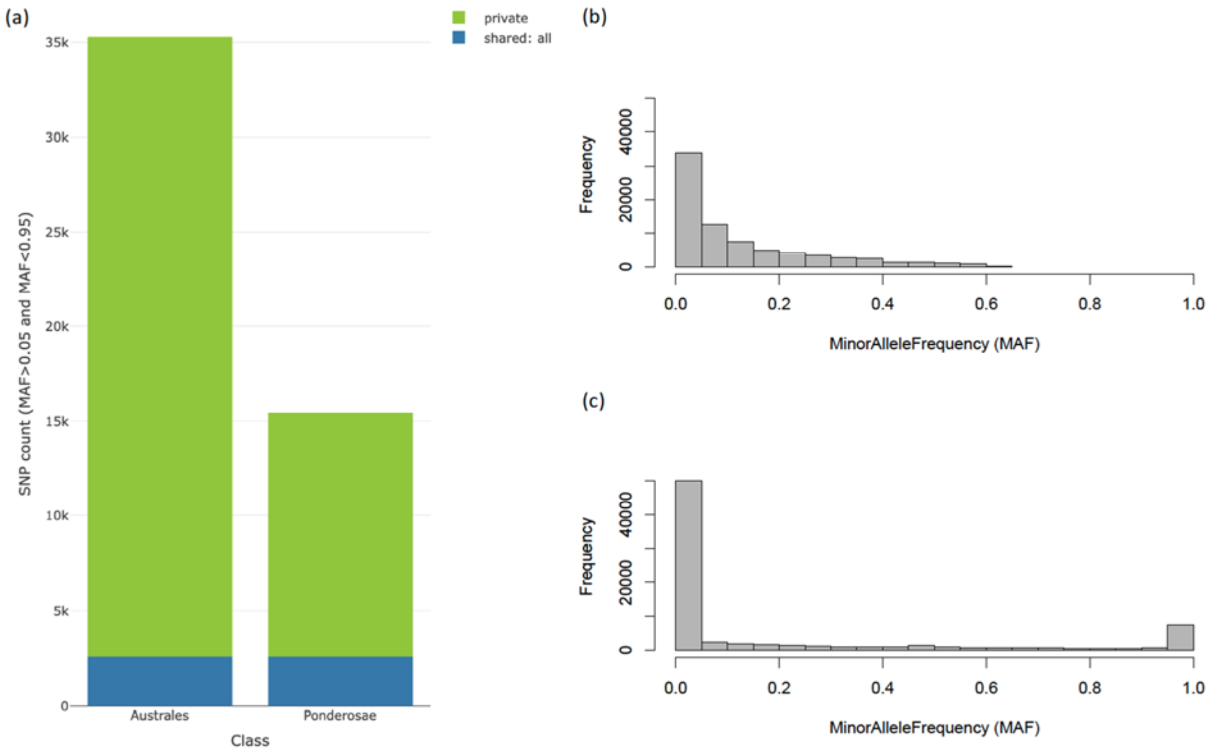
Supplementary Table 6. SNP & Variation Suite input file for the 49,694 SNP probe sets (rows; Supplementary Table 5) across 538 samples (columns; Supplementary Table 7). Per-species clustering was used to guide genotype calling in the Axiom Analysis Suite and the forward strand base call genotypes were exported. The resulting genotype files from different species were merged using functions from the tidyverse package in R.

External file: < Suppl_Table6.csv >

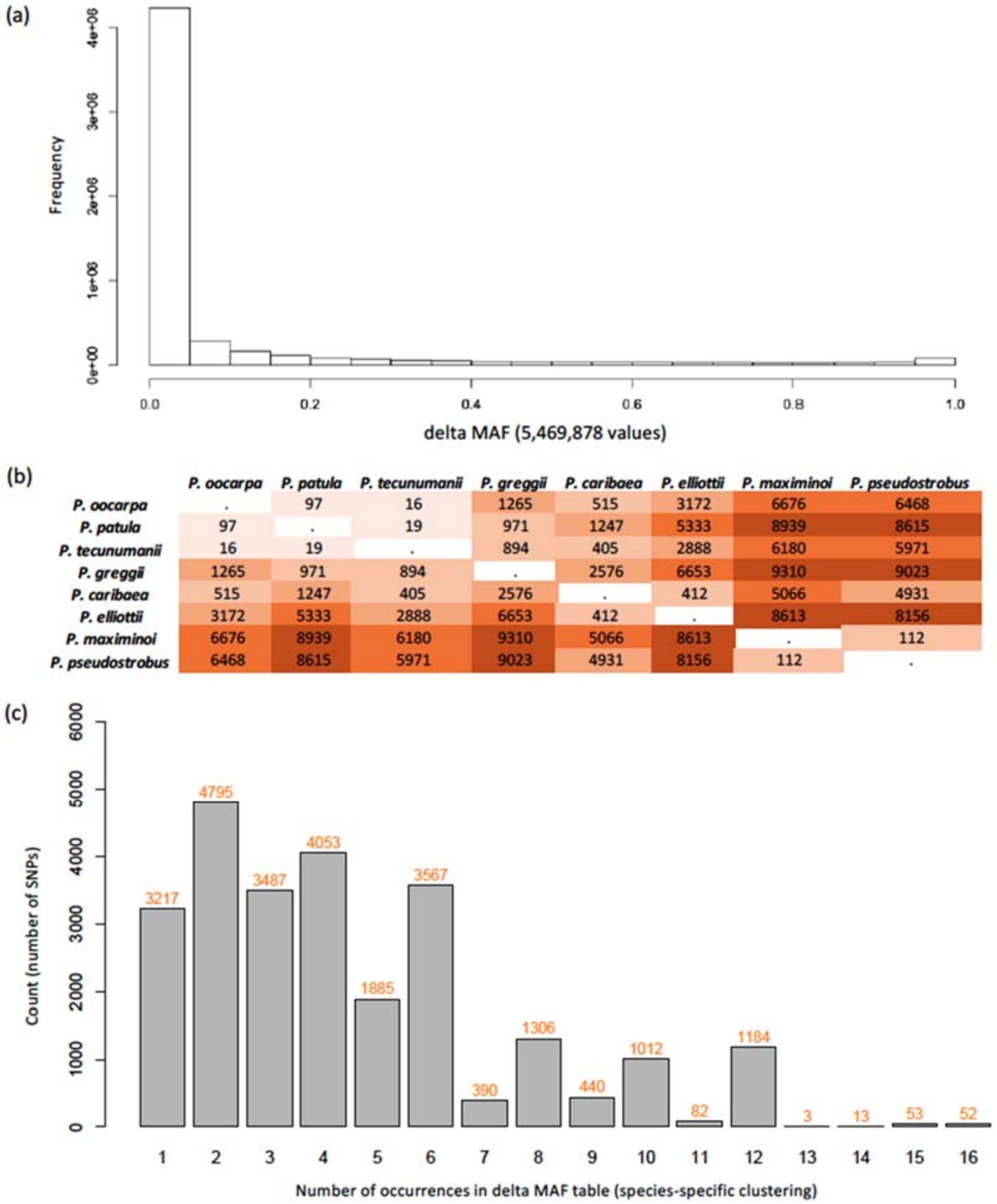
Supplementary Table 7. Sample metadata accompanying the structure plots (Figure 5b and Supplementary Figure 7). The samples are ordered within each species per provenance by country and then latitude (from South to North).

External file: < Suppl_Table7.xlsx >

Supplementary Figures

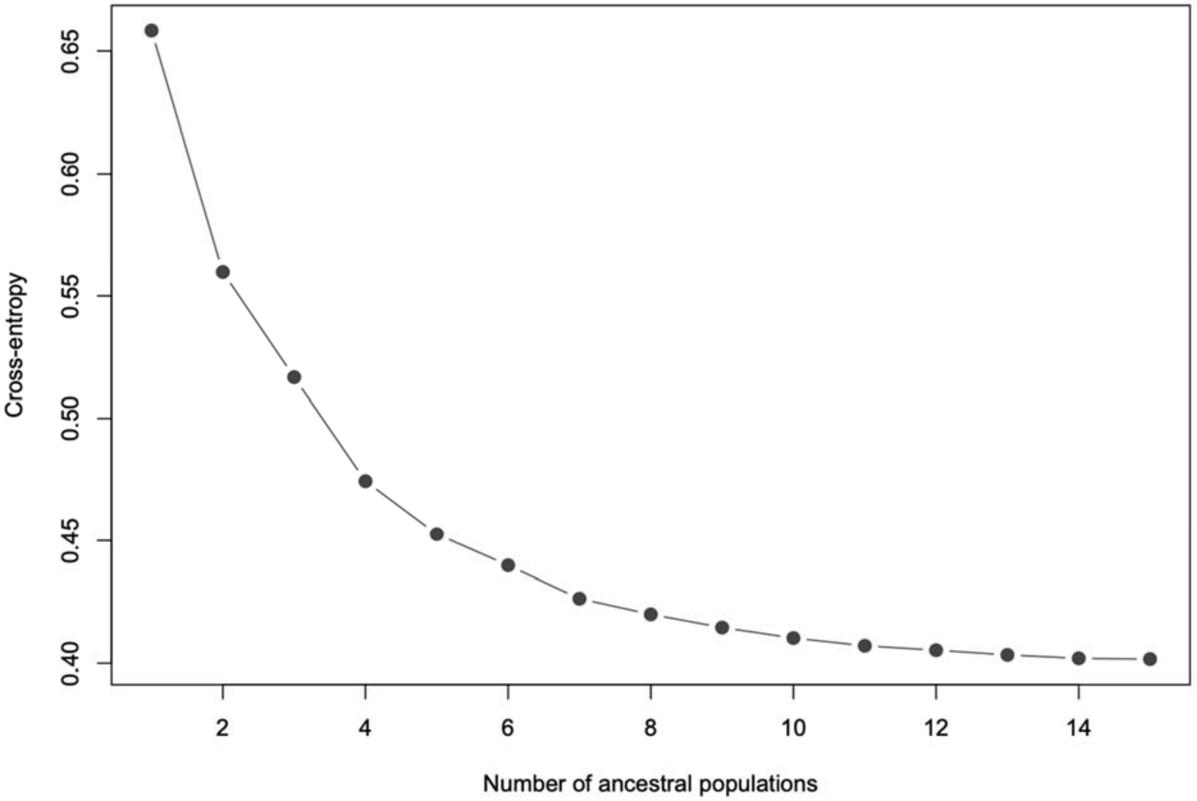


Supplementary Figure 1. Stable assay SNPs at the subsection level using global clustering. Set A1 (see Figure 3) consists of 2,615 SNP probe sets, where the probe set is informative in both subsections: Australes (*P. caribaea, P. elliottii*, *P. oocarpa*, *P. patula, P. tecunumanii*, *P. greggii*) and Ponderosae (*P. maximinoi, P. pseudostrobus*). (a) Stacked barplot giving a summary of the number of informative SNP probe sets per class, shared by both classes (blue) or unique to the class (green); (b) Minor allele frequency (MAF) histogram of all SNP probe sets informative in the Australes subsection; (c) MAF histogram of all SNP probe sets informative in the Ponderosae subsection.

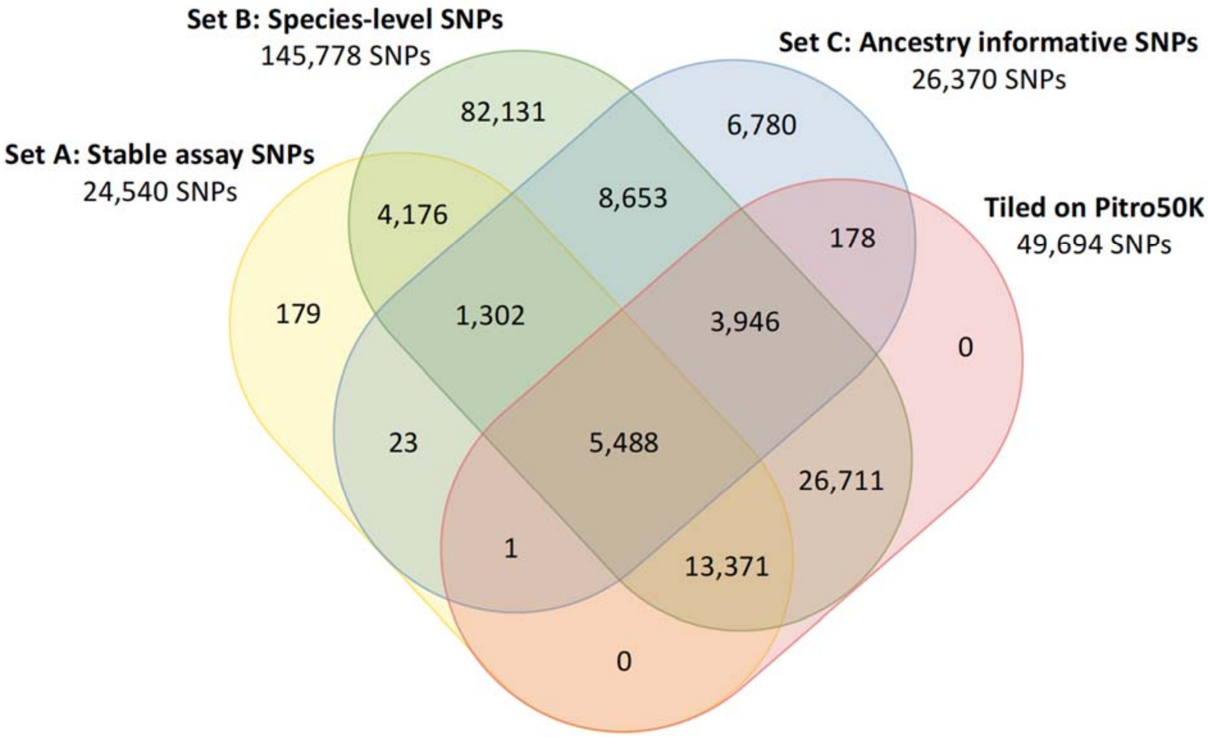


	P. oocarpa	*P. patula*	*P. tecunumanii*	*P. greggii*	*P. caribaea*	*P. elliottii*	*P. maximinoi*	*P. pseudostrobus*
P. oocarpa	.	97	16	1265	515	3172	6676	6468
P. patula	97	.	19	971	1247	5333	8939	8615
P. tecunumanii	16	19	.	894	405	2888	6180	5971
P. greggii	1265	971	894	.	2576	6653	9310	9023
P. caribaea	515	1247	405	2576	.	412	5066	4931
P. elliottii	3172	5333	2888	6653	412	.	8613	8156
P. maximinoi	6676	8939	6180	9310	5066	8613	.	112
P. pseudostrobus	6468	8615	5971	9023	4931	8156	112	.

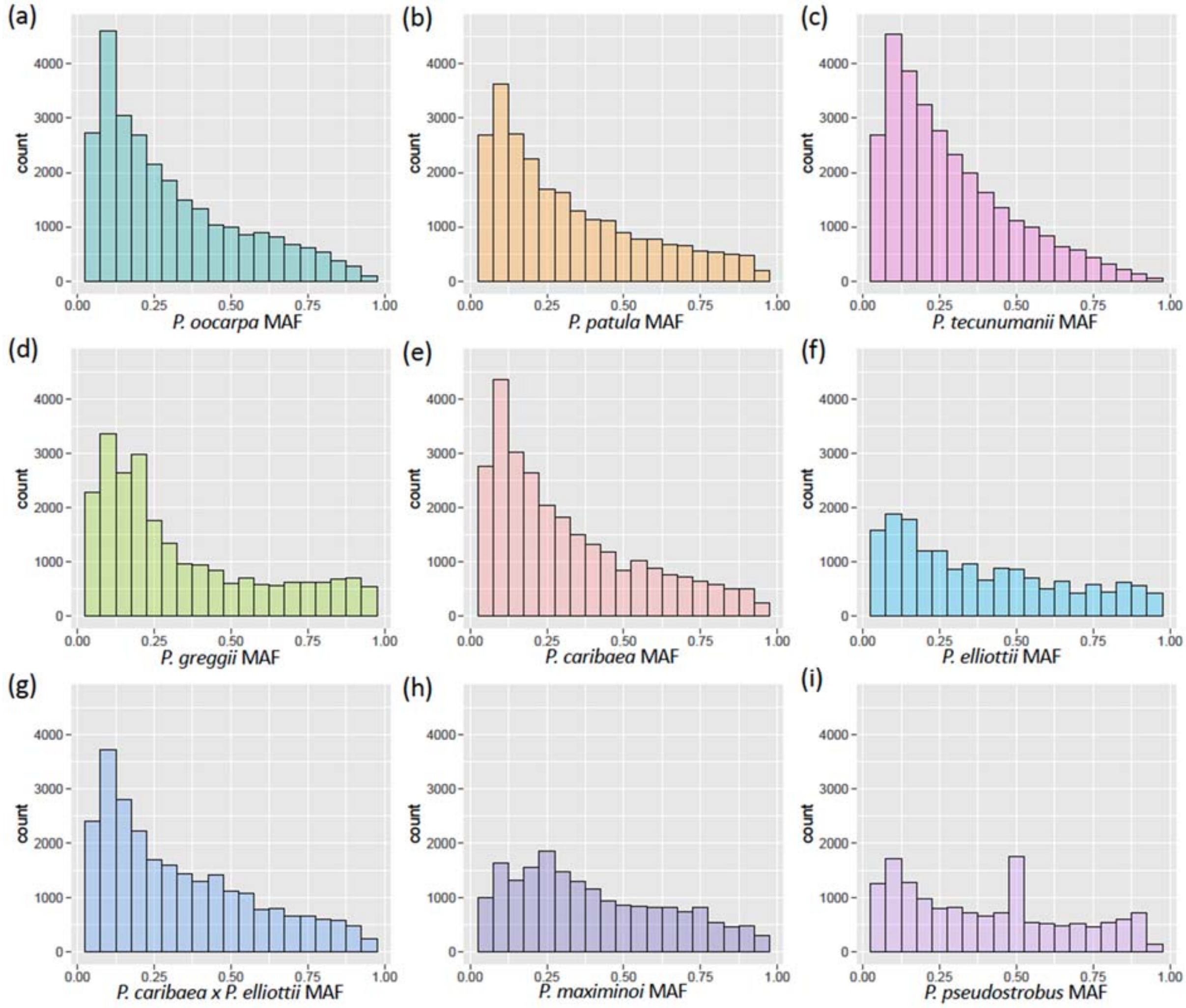
Supplementary Figure 2. SNPs with a large pairwise minor allele frequency (MAF) difference between species. SNP probe sets with a large pairwise difference (>0.9) in MAF between any two species were labeled *high delta MAF* SNPs (Set C1 in Figure 3). (a) Histogram of all the delta MAF values that were calculated for all SNPs across all pairwise species combinations (5.5 million values); (b) The number of SNPs per species combination where the calculated delta MAF was greater than 0.9; (c) A barplot of the number of SNP probe sets (y-axis) with a delta MAF > 0.9 for x pairwise species combinations (where x is the number on the x-axis).



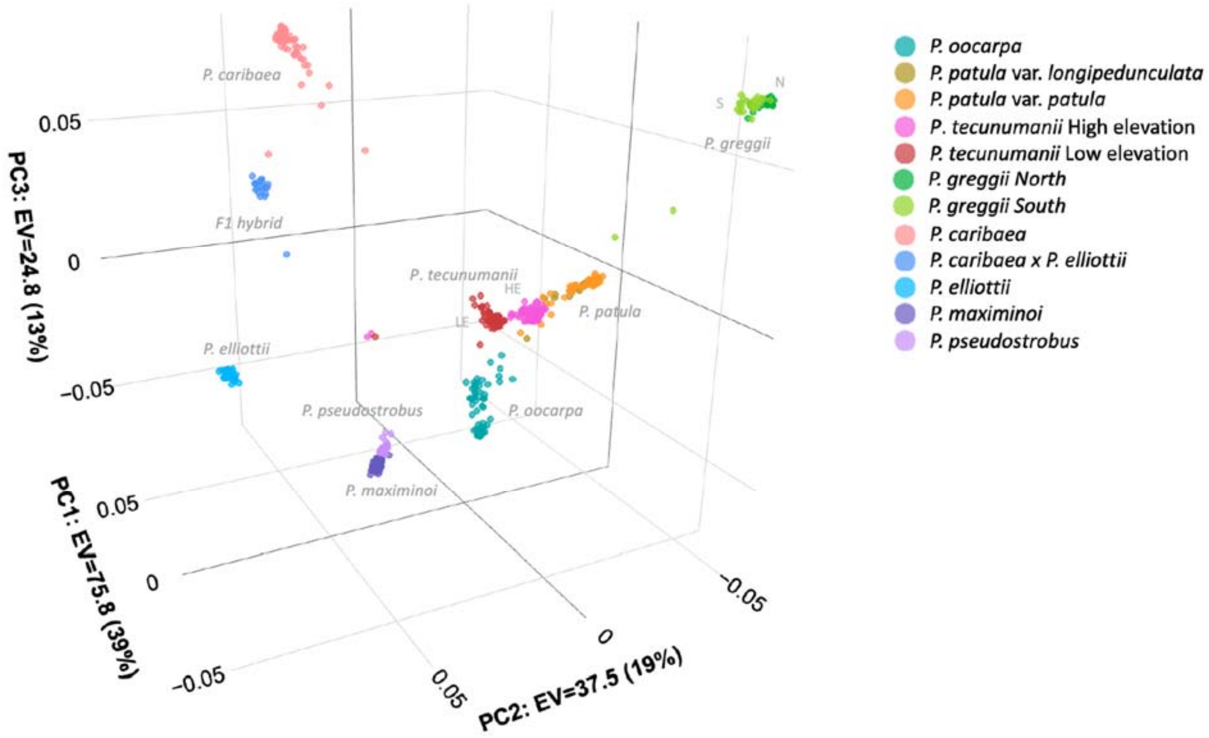
Supplementary Figure 3. Cross-entropy for pine species. Values of the cross-entropy criterion as a function of the number of factors in sNMF (non-negative matrix factorization algorithms) runs for K=1 to K=15 where K is the number of ancestral populations, across 538 individuals and a genetic space of 49,694 SNP probe sets.



Supplementary Figure 4. Venn diagram of how the sets identified after initial filtering of SNP probe sets on the screening array overlap with the selected 50K SNP probe sets. Set A represent stable assay SNP probe sets (yellow), Set B represent species level SNP probe sets (green) and Set C represent ancestry informative SNP probe sets (blue). Figure 3 gives details on how the SNP probe sets in sets A, B and C were obtained and Table 3 gives the strategy of how the final set of SNP probe sets was selected (red). Of the selected 50K SNP probe sets: 10.8% (5,488) appears in all three sets, 54% (26,711) in set B only, 26.5% (13,371) in both sets A and B, and 8.2% (3,946) in both sets B and C.



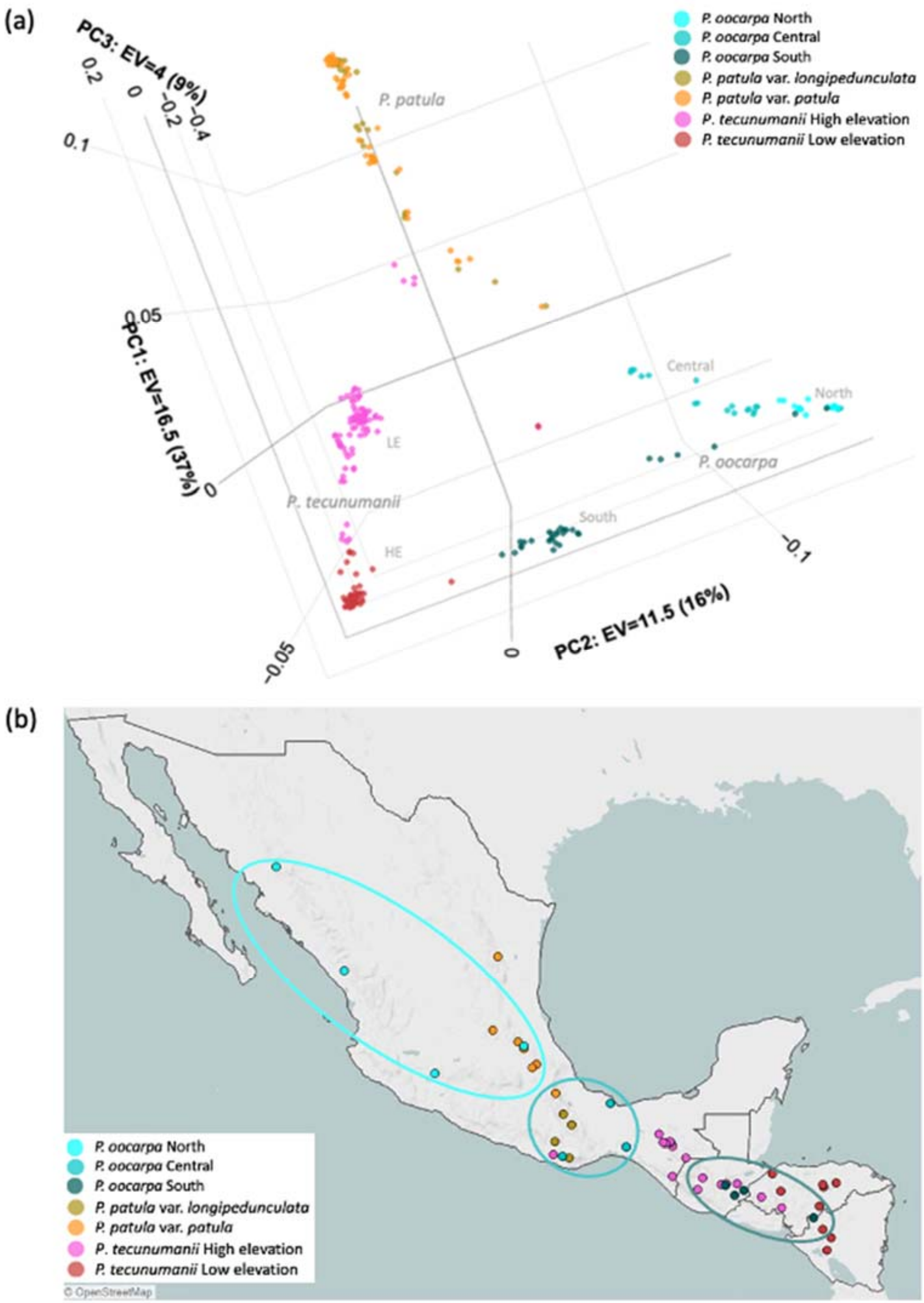
Supplementary Figure 5. Minor allele frequency (MAF) distributions of the 50K SNP probe sets per species. The MAF from 0.05 to 0.95 are displayed (i.e. the bottom 5% and the top 5% are excluded) for (a) *P. oocarpa*, (b) *P. patula,* (c) *P. tecunumanii*; (d) *P. greggi*, (e) *P. caribae*a, (f) *P. elliotti*, (g) *P. caribaea* x *P. elliottii*, (h) *P. maximino*i and (i) *P. pseudostrobus*.



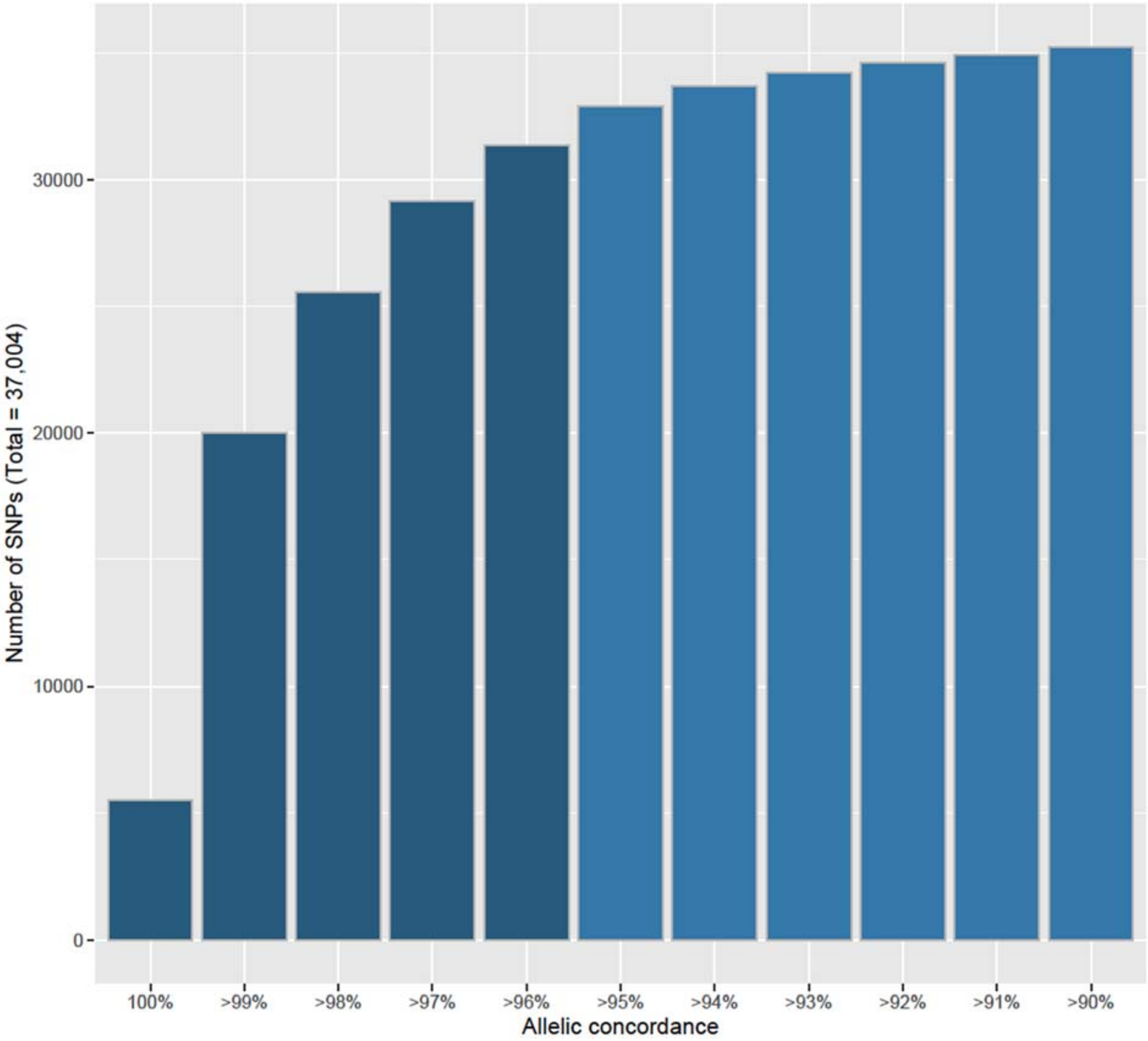
Supplementary Figure 6. Principal component (PC) analysis of the 1,564 diagnostic SNP probe sets across all 538 samples. The origin of this set of diagnostic SNP probe set is given in Figure 3 (Set C2) and Supplementary Table 2. PCs were calculated with SNP & Variation Suite (SVS) and visualized in 3D using the *plotly* R package. See the interactive 3D PCA plot: https://chart-studio.plotly.com/~nanettec/7/#/. The first PC explains 21% of the total variance, the second PC 19% and the third PC 16%.

External file: < Suppl_Figure7.pdf >

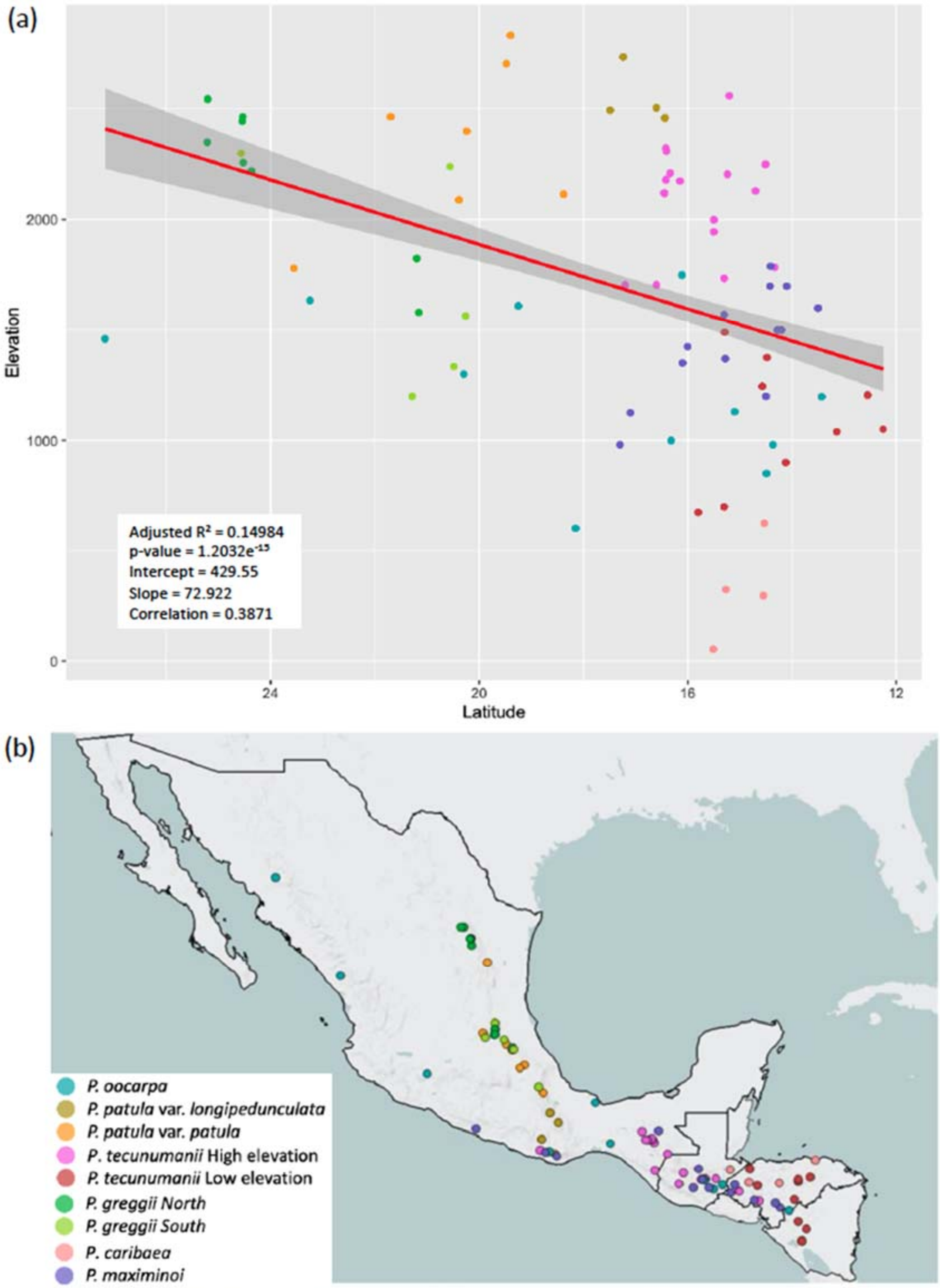
Supplementary Figure 7. High quality image of the structure plot for K=11. It is possible to zoom in and browse at the provenance level. The samples are ordered by species/subdivision and then by provenance from South to North (based on latitude). Supplementary Table 6 provides the metadata and full label for each provenance.



Supplementary Figure 8. Principal component (PC) analysis and geographic location of *P. oocarpa, P. patula* and *P. tecunumanii* samples. (a) PC analysis of the selected 50K SNP probe sets for all *P. oocarpa, P. patula* and *P. tecunumanii* samples. PCs were calculated with SNP & Variation Suite (SVS) and visualized using in 3D the *plotly* R package. See the interactive 3D PCA plot: https://chart-studio.plotly.com/~nanettec/13/#/. The first PC explains 36% of the total variance, the second PC 16% and the third PC 9%. (b) Geographic location of the relevant provenances, distinguishing between South, central and North provenances of *P. oocarpa*. The *P. oocarpa* subgroups are evident in the structure plot (Supplementary Figure 7).



Supplementary Figure 9. Allelic concordance between genotypes originating from global clustering versus per-species clustering runs. For each SNP probe set, the number of samples with exactly the same genotype call in the global vs the per-species clustering runs were calculated. A total of 37,004 SNPs (all selected SNP probe sets that were recommended in the global analysis) and 521 samples were included in the analysis. Approximately half of the SNP probe sets (20,024 probe sets) had a concordance of more than 99%; whereas 5,617 probe sets, corresponding to 500 out of the 521 samples, had a concordance of 95% or less (bars indicated in a lighter shade of blue).



Supplementary Figure 10. Elevation increases as latitude changes from South to North in Mexico and Central America. (a) Scatter plot of elevation vs latitude (correlation = 0.39) across the 81 provenances representing six tropical pine species (Supplementary Table 1) and (b) geographic location of the same provenances.