Notes regarding Bromeliad Bait Design

Gil Yardeni, Univie

May 31st, 2019

summary

Here's a report describing the preparation of a bait set for target capture in Bromeliaceae*Tillandsia* based on *Ananas comosus* genes. Choice of genes and filtering criteria were largely based on (Harpe et al., 2018) as well as other publications and feedback from the group and community.

Genes were filtered according to measures of heterozygosity and evolutionary rate in order to achieve a set relevant to family-level research and according to exon size and copy-number variation to fit the technical properties of bait capture (exon size greater than 120bp and single or low gene copy numbers). In addition, only genes that mapped to a linkage group rather than to a scaffold on the *A. comosus* genome were included. A lengthy description of filtering criteria can be found in section 1.

Hundreds of additional genes with putatively interesting biological function and/or previously used as markers for phylogenomic inference in Bromeliacea were added to the set. An comprehensive description of the genes and functions can be found in section 2.

Finally, to avoid homologous and possibly duplicated genes, all exonic sequences within the potential set were BLASTed against the *A. comosus* genome, as well as against a draft *Tillandsia leiboldiana* genome and a *Tillandsia* plastid assembly. Genes with a high 'match' were removed as described in section 3.

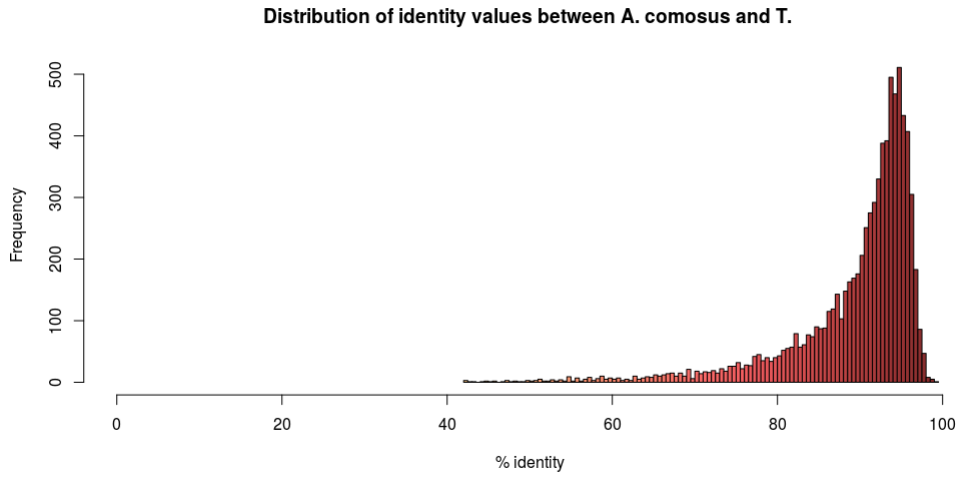
The final set of genes for bait design contains 1776 genes with 10204 exons and is described in section 4.

1 Filtering Criteria

To choose genes from which to design capture probes, *A. comosus* genes were filtered according to criteria outlined below.

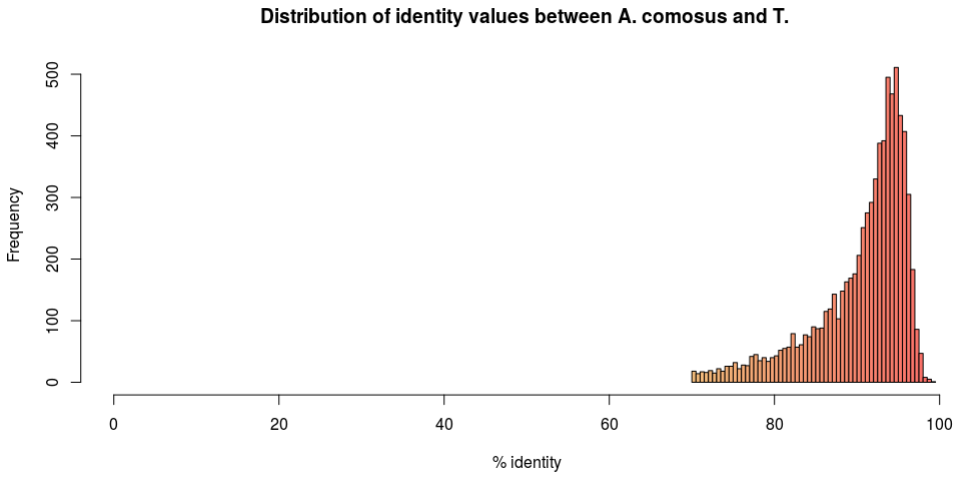
- Total number of *A. comosus* genes I started with: 31991
- Genes that passed evolutionary rate filter: 31691
- Genes that passed heterozygosity filtering: 6038
- Genes that passed exonic size filtering: 1343
- Genes that passed copy number filtering: 1243

Evolutionary rate As a proxy for evolutionary rate, I looked into the divergence between *A. comosus* and *Tillandsia sphaerocephalia*. Average identity was quite high, as is presented in the next figure:



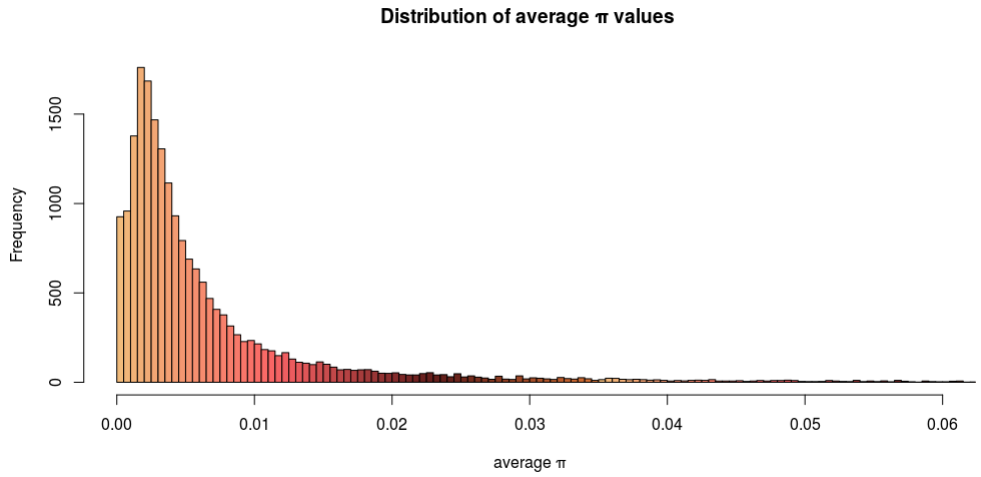
Identity between A. comosus and T. sphaerocephalia. Colors represent nothing and serve for entertaining purposes only.

In order to guarantee high capture efficiency, all genes with >70% identity were retained. Genes that lacked information on evolutionary rate were retained as well. [1] The distribution of identity values then lost its left 'tail'.



Heterozygosity Used as a proxy for the ability to recover intraspecific variation. This was done based on Jacky's π calculations within *A. comosus* and 4 Tillandsia species (*T. australis*, *T. fasciculata*, *T. floribunda* & *T. sphaerocephala*). I calculated average π values for the filtering step. The next figures shows the distribution of average π values:

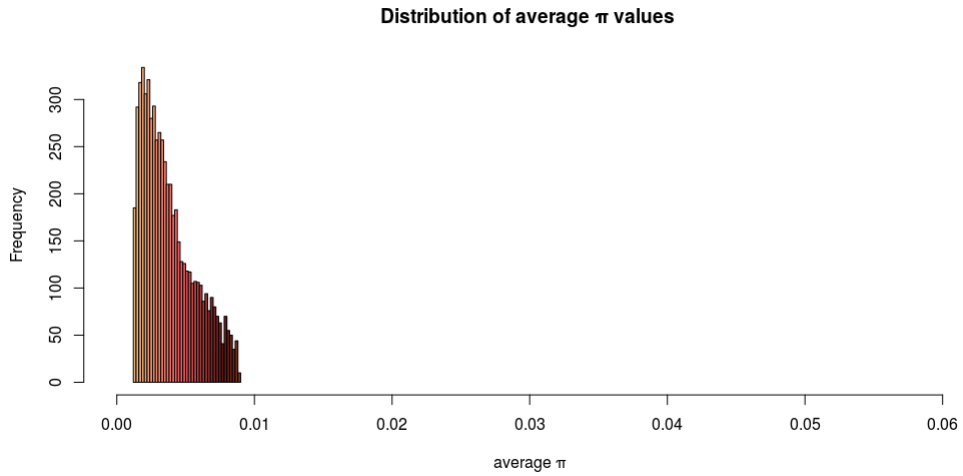
[1] Out of 32k genes, ∼7800 had information on evolutionary rate. However, the majority of these genes were filtered in other stages. A summary of the dataset properties is in section 4



The genes with highest and lowest 10% of the distribution were removed - low values were removed to recover variation and high values were removed to avoid potential duplication.

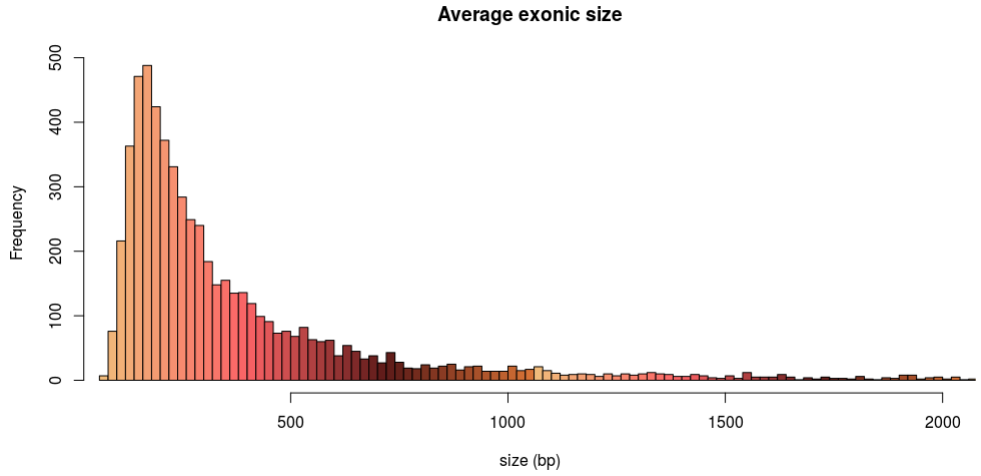
10%	90%
0.001249152	0.008857841

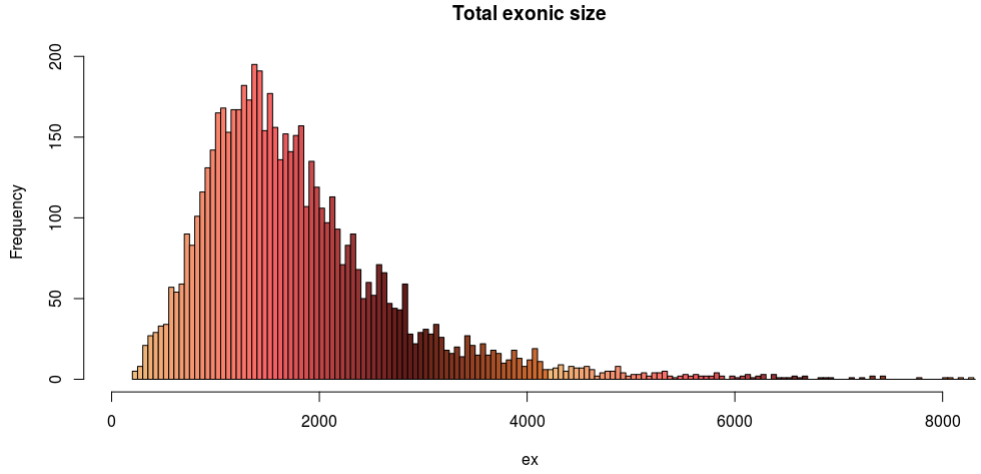
The resulting distribution simply lost its left and right 'tails':

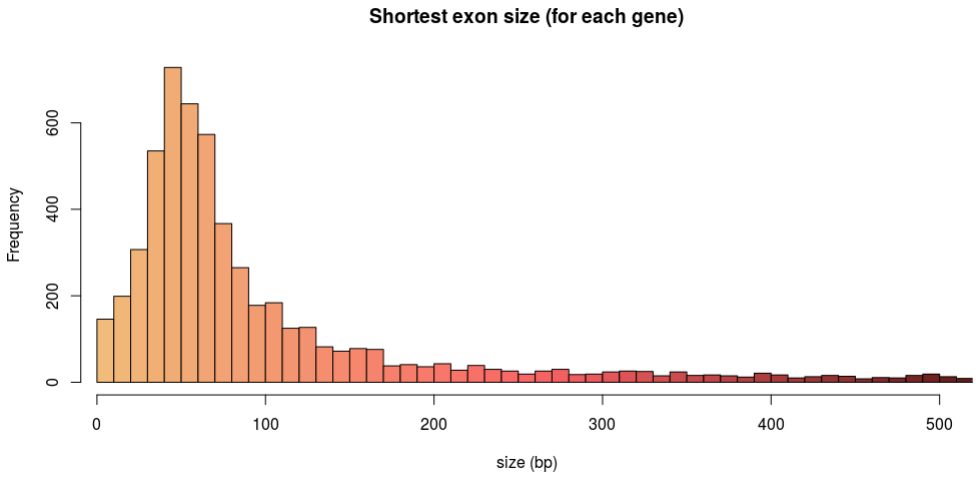


Exonic size I calculated average and total exonic size and also looked at the size of the shortest exon in each gene. Exons shorter than the bait insert size of 120bp can be submitted for design and 'filled' with intergenic sequences, however they might result in lower capture efficiency, especially at exon edges. Some papers report the difference in capture efficiency is not significant (for example, see (Schott et al., 2017; Portik, Smith, & Bi, 2016)

some histograms of the data before filtering:



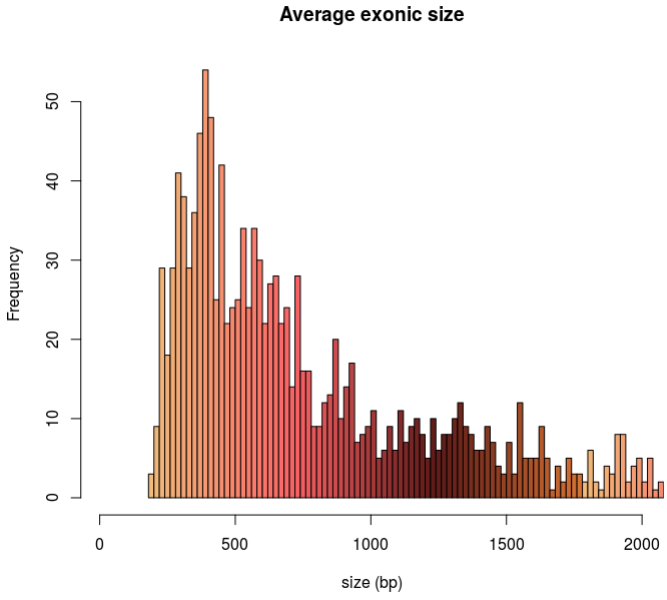


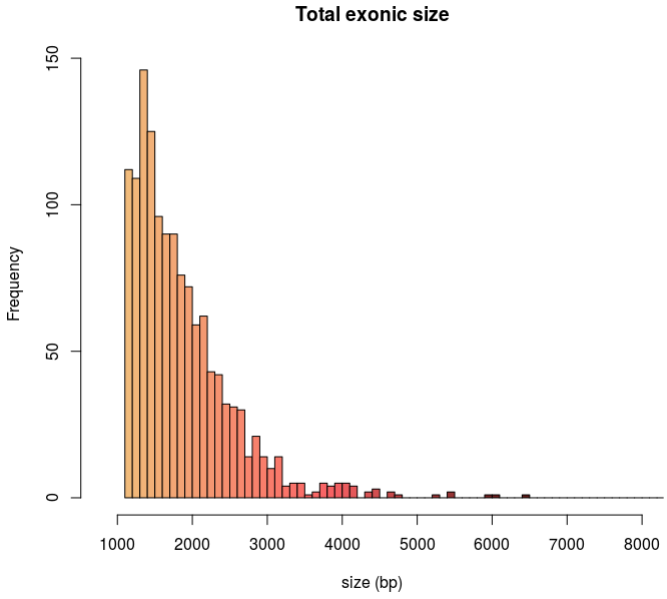


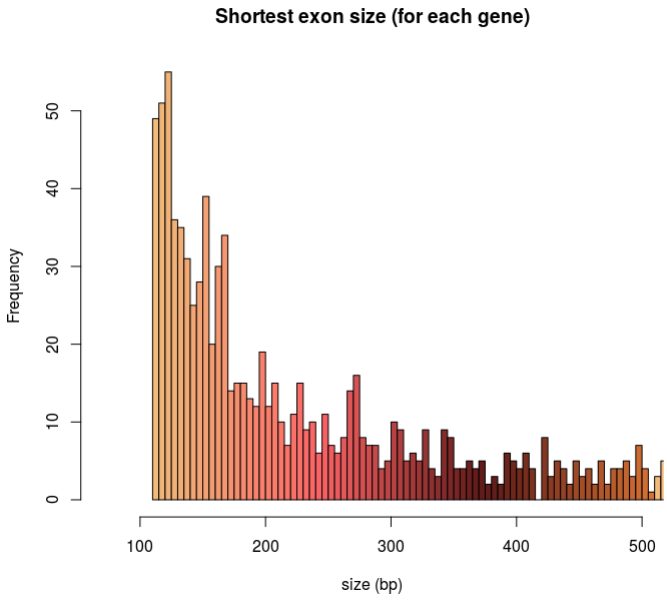
Avoiding small exons laid the greatest constraints on the data set and resulted in rigorous filtering, as the vast majority of genes contain at least one exon shorter than 110bp.

The filtering setting I applied were as following: retained genes in which shortest exon $>$ 110bp and total exonic size $>$ 1100bp.

some histograms of how the exon size distribution looks after filtering. The greatest different was for average exonic size, as the filter imposed on total exonic size effected the distribution of exon sizes


Average exonic size
Frequency
0
10
20
30
40
50
0
500
1000
1500
2000
size (bp)


Total exonic size
Frequency
0
50
100
150
1000
2000
3000
4000
5000
6000
7000
8000
size (bp)



Copy-number [2] filtering was based on Jacky's CNV analysis. 126 low copy genes were removed.

2 Collection of interesting genes

To the bait set above I added genes of interest, such as genes involved in CAM/C3 shifts, genes useful for the Bromeliaceae community, etc. Several resources were used, details below.

- Genes that match to the Paftol bait set (AKA Angiosperm353) in the Ananas genome (Johnson et al., 2018)
- Genes significantly differentially expressed between CAM/C3 species, as found in Marylaure's RNA-seq experiements (Harpe et al., 2018); 190 genes [3]

[2] single copy genes = 1 copy, low copy genes ≤ 5, high copy >5

[3] counting only single copy and low-copy genes - high-copy genes and genes without CNV analysis data were filtered out

- Genes that show positive selection (dN/dS) along CAM shifts, again as indicated in (Harpe et al., 2018); 22 genes
- Genes that belong to gene families with putative association to CAM/C3 correlated trait syndrome (Harpe et al., 2018); 57 genes
- Genes annotated with interesting functions ('special annotations'), curated by Marylaure; 572 genes
- Genes with potential functional interest related to photosynthesis or reproductive systems, based on literature; 599 genes
- Markers previously used for phylogenetic inference in Bromeliacea; 17 genes

Genes in the set above definitely overlap occasionally.

special annotations Details of the genes annotated by Marylaure:

1. Putative CAM-related genes from pineapple genome assembly publication Ming et al., 2015, (Ming et al., 2015); 31 genes
2. Genes related to stomata opening and closing based on (Winter & Holtum, 2014) and (Christin et al., 2014); 32 genes
3. Genes related to aquaporin regulation (VERA-ESTRELLA, Barkla, AMEZCUA-ROMERO, & Pantoja, 2012); 21 genes
4. Genes related to malate transferase in the vacuole, assimilation of inorganic carbon and compartmentation of carbohydrates according to (Cosentino et al., 2013); 32 genes
5. Genes related to drought-resistance (Xiao, Huang, Tang, & Xiong, 2007); 52 genes
6. Circadian clock genes (McClung, 2006); 8 genes
7. Genes related to glucogenesis and starch synthesis according to (Cushman, Tillett, Wood, Branco, & Schlauch, 2008; Antony et al., 2008; Wada & Murata, 2009); 288 genes

Genes with potential functional interest, based on literature:

1. Gene families found to be under positive selection within the portullugo clade (Caryophyllales), a lineage that contain multiple evolutionary origins of all known photosynthesis types (Goolsby, Moore, Hancock, De Vos, & Edwards, 2018); 86 genes

2. transcripts of genes that encode key enzymes in the flavonoid and anthocyanin biosynthesis pathways. Isolated from transcriptomes of two Pitcairnia species (Palma-Silva, Ferro, Bacci, & Turchetto-Zolet, 2016); 119 genes

3. Genes with circadian expression or key roles in metabolism used to examine dynamics of CAM photosynthesis by (Wai et al., 2017) Circadian oscillators (11 genes), fructose, hexose, malate transporters (9 genes), proton pumps (11 genes), genes related to stomatal movement, etc; 41 genes.

4. Genes for anthocyanin biosynthesis in *A. comosus* or *A. bracteatus*; 17 genes and candidate genes for self incompatibility in *A. comosus*; 4 genes, identified in (Ming et al., unpublished)

Genes recommended by Michael Barfuss, previously used for phylogenetic inference within Bromeliaceae:

1. PHYC (phytochrome C - Ananas comosus)

2. PRK (phosphoribulokinase - Ananas comosus

3. MS (malate synthase)

4. NIA (nitrate reductase 1, NADH)

5. PGIC (glucose-6-phosphate isomerase, cytosolic)

6. RPB2 (RNA polymerase II, beta subunit) [4]

7. XDH (xanthine dehydrogenase)

8. Floricaula/Leafy genes, used for phylogenetics in Alcantarea (Versieux et al., 2012)

[4] this gene was marked as high-copy in CNV analysis & is worth discussing if this is an artifact

9. AGT1 (Serine–glyoxylate aminotransferase) used for nuclear barcoding (Heller, Leme, Paule, Koch, & Zizka, 2017)
10. PEPC (Phosphoenolpyruvat carboxylase)
11. PHYA (phytochrome A) at LG13 (Ananas genome)
12. PHYE/B (phytochrome E/B) at LG18 (Ananas genome)

The pipeline for adding all these genes to the table was as follows: the sequences for candidate genes was obtained (a FASTA file, usually released with the corresponding publication. Some genes were annotated and can be found in NCBI or the Phytozome portal). Each gene was then blasted against the Ananas comosus protein database[5] and the highest match in blast was used to 'mark' the Ananas accession as an interesting gene. Candidate genes regularly 'matched' to several Ananas accessions and vice versa - certain Ananas accessions would match to different candidate genes. This could be the result of both gene duplications and homology. In any case, when possible, I included the highest blast match only. When the two highest matches had the same score (usually at values around 0), I retained both. Eliminating exons that might be duplications was performed later as is discussed in section 3.

The following table presents the number of genes from each category (naturally, some genes overlap, fitting into several categories):

[5]in cases where Ananas genes already respond to the list of accessions we used, specifically (Ming et al., 2015) and (Wai et al., 2017), these genes weren't - simply marked

Category	total	single-copy	low-copy	high-copy
Ming et al., 2015	38	13	18	0
Stomatal	66	15	25	7
Aquaporin	34	4	17	0
ATPase	26	8	13	0
Oxygen evolving enhancer	14	5	6	0
Drought resistance	70	18	34	5
Circadian clock	8	6	1	0
Glucogenesis	388	130	158	36
Glycolysis	61	26	22	10
Angiosperm (paftol) set	281	224	54	2
Phylogenomic markers	13	2	7	1
CAM-related, Goolsby et al.,	109	33	53	22
Flavonoid & anthocyanin pathways	125	80	39	7
Genes from Wai et al.,	51	19	22	9
Differential expression CAM/C3	236	102	88	22
Positive selection	22	8	9	4
Genes families CAM/C3 associated	57	23	27	0
Ananas anthocyanin pathway & SI genes	21	7	11	0
Total (?)	1625	723	592	125

With the addition of specially-picked genes and removal of genes that are present on Ananas scaffolds, the gene set includes 2664 genes.

3 'Exon Match'

To exclude loci that are potentiality present in multiple copies in the genome due to homology, all exons were blasted again against the *A. comosus* nucleotide database as well as against several *Tillandsia* chloroplast assembly and our draft *Tillandsia* genome, an approach used in several key publications of bait design (Weitemier et al., 2014; Li, Hofreiter, Straube, Corrigan, & Naylor, 2013). Any targets (=exons) 1. sharing $> 80\%$ similarity with another sequence 2. match E-score $\leq 10^{-5}$ 3. matching against a sequence (matched alignment length) > 100bp & 4. **not already marked as multi-copy** were removed. I scanned the list further manually for 'interesting genes' and made exceptions where important genes didn't look like a 'true' duplication (only a small number of exons match against short alignments, or E-score is very different between duplication or 'true' match, etc.).

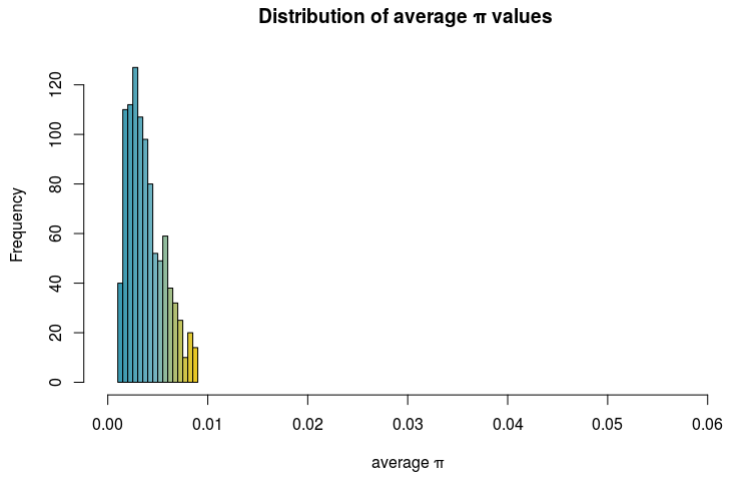
Out of 2664 genes, 455 matched in blast against the *Ananas*, 467 matched in blast against *T.leiboldiana* draft genome and 14 were removed following

blasts against plastid assemblies of various *Tillandsia*. 106 genes were NOT excluded after detailed, manual examination. 696 genes were excluded[6], resulting in 1776 genes in the bait set.

4 Properties of bait set

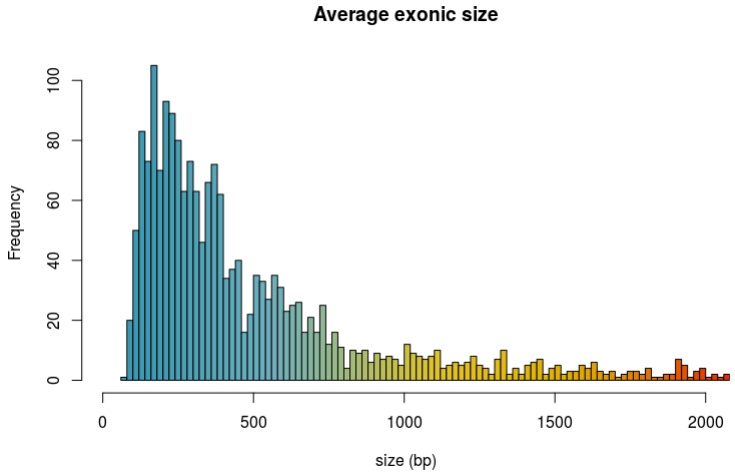
The designed bait set contains 1776 genes with 10204 exons. If we count an addition 353 Paftol baits the number of genes is 2129. Properties of the bait set are described below:

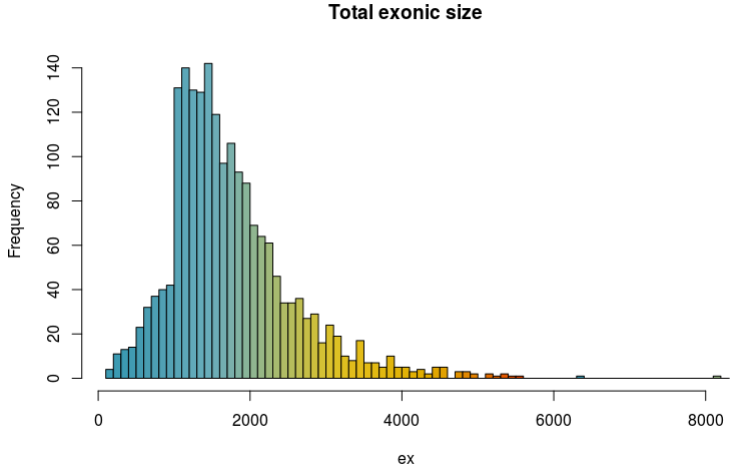
heterozygosity

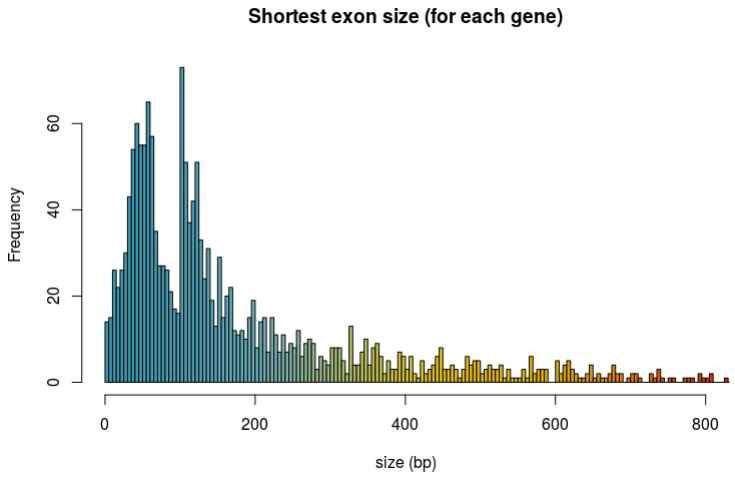


[6]as there was overlap between genes that matched in different blast searches.

exonic size


Average exonic size
Frequency
0
20
40
60
80
100
0
500
1000
1500
2000
size (bp)

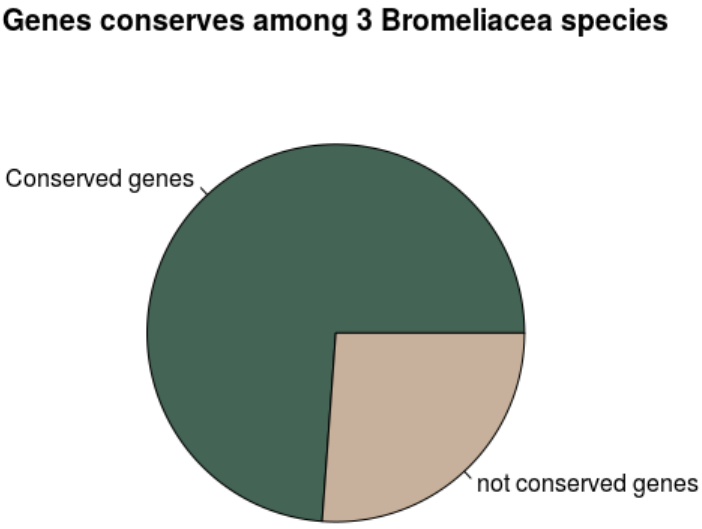

Total exonic size
Frequency
0
20
40
60
80
100
120
140
0
2000
4000
6000
8000
ex



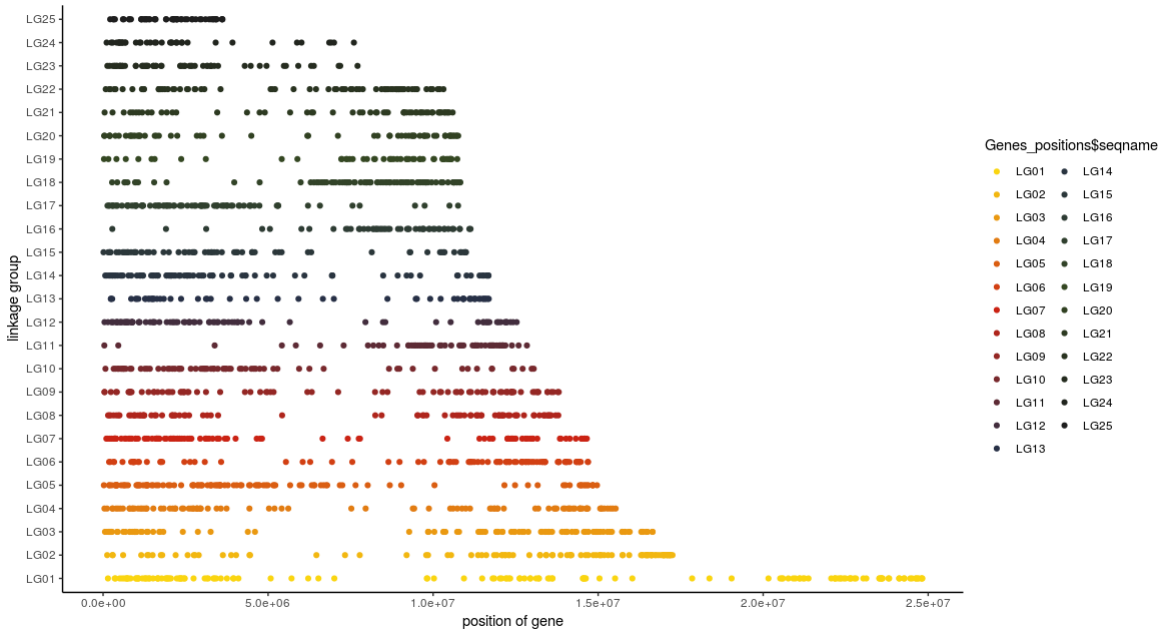
Sequences shared between Brochinnia, Tillandsia, Ananas Genes shared between the 3 species are more likely to serve as good markers for phylogenetic inference in Bromeliaceae in general. I did not filter the genes, just looked at the proportion of genes shared between the 3 species and found that the vast majority is shared between them:

conserved: 1389 not conserved: 492

amusing pie chart:



Representation along the genome I grpahically plotted the positions of genes in the dataset. Generally they're distributed on all chromosomes, with some gaps, i.e. in LG 16 & LG11.



5 Conclusions

.

References

Antony, E., Taybi, T., Courbot, M., Mugford, S. T., Smith, J. A. C., & Borland, A. M. (2008). Cloning, localization and expression analysis of vacuolar sugar transporters in the cam plant ananas comosus (pineapple). *Journal of Experimental Botany*, *59*(7), 1895–1908.

Christin, P.-A., Arakaki, M., Osborne, C. P., Bräutigam, A., Sage, R. F., Hibberd, J. M., ... others (2014). Shared origins of a key enzyme during the evolution of c4 and cam metabolism. *Journal of Experimental Botany*, *65*(13), 3609–3621.

Cosentino, C., Di Silvestre, D., Fischer-Schliebs, E., Homann, U., De Palma, A., Comunian, C., ... Thiel, G. (2013). Proteomic analysis of mesembryanthemum crystallinum leaf microsomal fractions finds an imbal-

ance in v-atpase stoichiometry during the salt-induced transition from c3 to cam. *Biochemical Journal*, *450*(2), 407–415.

Cushman, J. C., Tillett, R. L., Wood, J. A., Branco, J. M., & Schlauch, K. A. (2008). Large-scale mrna expression profiling in the common ice plant, mesembryanthemum crystallinum, performing c3 photosynthesis and crassulacean acid metabolism (cam). *Journal of Experimental Botany*, *59*(7), 1875–1894.

Goolsby, E. W., Moore, A. J., Hancock, L. P., De Vos, J. M., & Edwards, E. J. (2018). Molecular evolution of key metabolic genes during transitions to c4 and cam photosynthesis. *American journal of botany*, *105*(3), 602–613.

Harpe, M. d. L., Paris, M., Hess, J., Barfuss, M. H. J., Serrano-Serrano, M. L., Ghatak, A., ... Lexer, C. (2018). Genomic footprints of repeated evolution of CAM photosynthesis in tillandsioid bromeliads. *bioRxiv doi: https://doi.org/10.1101/495812*. (doi: https://doi.org/10.1101/495812)

Heller, S., Leme, E. M., Paule, J., Koch, M., & Zizka, G. (2017). Barcoding of bromeliaceae (poales). *Genome*, *60*(11), 943–945.

Johnson, M., Pokorny, L., Dodsworth, S., Botigue, L. R., Cowan, R. S., Devault, A., ... others (2018). A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *BioRxiv*, 361618.

Li, C., Hofreiter, M., Straube, N., Corrigan, S., & Naylor, G. J. (2013). Capturing protein-coding genes across highly divergent species. *Biotechniques*, *54*(6), 321–326.

McClung, C. R. (2006). Plant circadian rhythms. *The Plant Cell*, *18*(4), 792–803.

Ming, R., VanBuren, R., Wai, C. M., Tang, H., Schatz, M. C., Bowers, J. E., ... others (2015). The pineapple genome and the evolution of cam photosynthesis. *Nature genetics*, *47*(12), 1435.

Palma-Silva, C., Ferro, M., Bacci, M., & Turchetto-Zolet, A. (2016). De novo assembly and characterization of leaf and floral transcriptomes of the hybridizing bromeliad species (pitcairnia spp.) adapted to neotropical inselbergs. *Molecular ecology resources*, *16*(4), 1012–1022.

Portik, D. M., Smith, L. L., & Bi, K. (2016). An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (class: Amphibia, order: Anura). *Molecular ecology resources*, *16*(5), 1069–1083.

Schott, R. K., Panesar, B., Card, D. C., Preston, M., Castoe, T. A., & Chang, B. S. (2017). Targeted capture of complete coding regions across divergent species. *Genome biology and evolution*, *9*(2), 398–414.

VERA-ESTRELLA, R., Barkla, B. J., AMEZCUA-ROMERO, J. C., & Pantoja, O. (2012). Day/night regulation of aquaporins during the cam cycle in mesembryanthemum crystallinum. *Plant, cell & environment*, *35*(3), 485–501.

Versieux, L. M., Barbará, T., Wanderley, M. d. G. L., Calvente, A., Fay, M. F., & Lexer, C. (2012). Molecular phylogenetics of the brazilian giant bromeliads (alcantarea, bromeliaceae): implications for morphological evolution and biogeography. *Molecular Phylogenetics and Evolution*, *64*(1), 177–189.

Wada, H., & Murata, N. (2009). *Lipids in photosynthesis*. Springer.

Wai, C. M., VanBuren, R., Zhang, J., Huang, L., Miao, W., Edger, P. P., ... others (2017). Temporal and spatial transcriptomic and micro rna dynamics of cam photosynthesis in pineapple. *The Plant Journal*, *92*(1), 19–30.

Weitemier, K., Straub, S. C., Cronn, R. C., Fishbein, M., Schmickl, R., McDonnell, A., & Liston, A. (2014). Hyb-seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences*, *2*(9), 1400042.

Winter, K., & Holtum, J. A. (2014). Facultative crassulacean acid metabolism (cam) plants: powerful tools for unravelling the functional elements of cam photosynthesis. *Journal of Experimental Botany*, *65*(13), 3425–3441.

Xiao, B., Huang, Y., Tang, N., & Xiong, L. (2007). Over-expression of a lea gene in rice improves drought resistance under the field conditions. *Theoretical and Applied Genetics*, *115*(1), 35–46.