# Transcriptome analysis of drough-tolerant CAM plants, *Agave deserti* and *Agave tequilana*

Stephen M. Gross[1,2], Jeffrey A. Martin[1,2], June Simpson[3], Zhong Wang[1,2], and Axel Visel[1,2]

1. DOE Joint Genome Institute, Walnut Creek, CA
2. Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA
3. CINVESTAV, Irapuato, MX

## ABSTRACT

Agaves are succulent monocotyledonous plants native to hot and arid environments of North America. Because of their adaptations to their environment, including crassulacean acid metabolism (CAM, a water-efficient form of photosynthesis) and existing technologies for ethanol production, agaves have gained attention both as potential lignocellulosic bioenergy feedstocks and models for exploring plant responses to abiotic stress. However, the lack of comprehensive *Agave* sequence datasets limits the scope of investigations into the molecular-genetic basis of *Agave* traits. Here, we present comprehensive, high quality de novo transcriptome assemblies of two *Agave* species, *A. tequilana* and *A. deserti*, from short-read RNA-seq data. Our analyses support completeness and accuracy of the *de novo* transcriptome assemblies, with each species having approximately 35,000 protein-coding genes. Comparison of agave proteomes to those of additional plant species identifies biological functions of gene families displaying sequence divergence in agave species. Additionally, we use RNA-seq data to gain insights into biological functions along the *A. deserti* juvenile leaf proximal-distal axis. Our work presents a foundation for further investigation of agave biology and their improvement for bioenergy development.

## CAM PHOTOSYNTHESIS, ARID ENVIRONMENTS, AND BIOENERGY

*Agave* species are adapted to their native habitat in arid regions of Mexico and the United States. *Agave* thus holds promise as a biofuel feedstock [1,2], capable of growing on marginal lands where other proposed bioenergy plants cannot. The ability of agaves to withstand hot and arid conditions relies upon crassulacean acid metabolism (CAM)—a specialized form of photosynthesis allowing agaves to keep leaf stomata (pores) closed during the hot day, minimizing water loss through evapotranspiration.

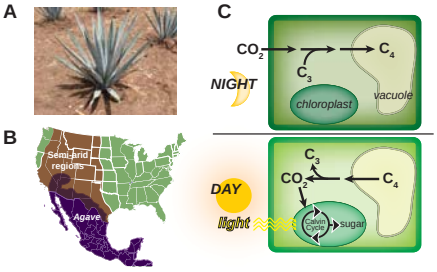



**FIGURE 1: Agaves and CAM biology**

(A) *Agave tequilana* cultivated in Mexico.

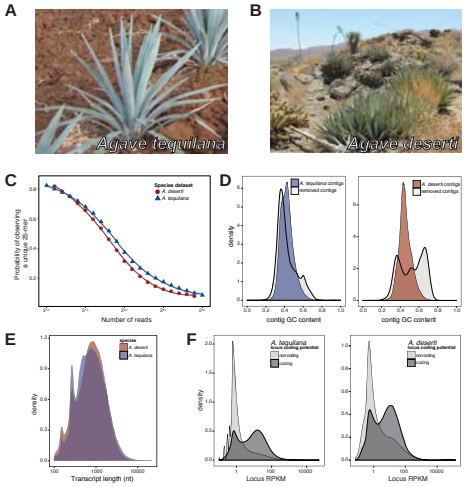
(B) Semi-arid regions of the United States (brown) are unsuitable for cultivation of other bioenergy plants, which require more temperate regions (green). Most *Agave* species are adapted to semi-arid regions in Mexico and the extreme southwestern USA (purple).

(C) Crassulacean Acid Metabolism (CAM). $CO_2$ enters plant cells at night, joins with a 3-carbon molecule ($C_3$) and is stored in the vacuole as a 4-carbon molecule ($C_4$). During the day, $C_4$ molecules diffuse out of the vacuole, and $CO_2$ is released and assimilated into sugar in the chloroplast.

| Feedstock | Inputs | | | Outputs | |
|---|---|---|---|---|---|
| | Water (cm yr$^{-1}$) | Drought tolerance | Nitrogen (kg ha$^{-1}$ yr$^{-1}$) | Dry biomass (Mg ha$^{-1}$ yr$^{-1}$) | Ethanol (liters yr$^{-1}$) |
| Corn grain | 50–80 | low | 90–120 | 7–10 | 2900 |
| Corn stover | | | | 3–6 | 900 |
| *Miscanthus* | 75–120 | low | 0–15 | 15–40 | 4600–12,400 |
| Poplar coppice | 70–105 | moderate | 0–50 | 5–11 | 1500–3400 |
| ***Agave* spp.** | **30–80** | **high** | **0–12** | **10–34** | **3000–10,500** |

Comparison of inputs (water and nitrogen) and outputs (biomass and ethanol) of agaves and other biofuel feedstock species. Though agaves are harvested at several years of age, their annualized growth rate is on par with *Miscanthus*. Table is modified from reference [2].

## *AGAVE* TRANSCRIPTOME ASSEMBLIES FROM DEEP RNA-seq




To provide sequence resources for the *Agave* research community, we built de novo transcriptomes of *Agave tequilana* and *Agave deserti* from deep Illumina RNA-seq data. Sequences were assembled by Rnnotator [3], a *de novo* transcriptome assembly pipeline.

**FIGURE 2: *A. tequilana*, *A. deserti*, and their respective transcriptomes**

(A) Cultivated *A. tequilana* in Jalisco, Mexico.

(B) *A. deserti* (foreground) in natural habitat, Riverside County, California, USA.

(C) Plot of the fraction of unique 25-mers over indicated read depth (log2 scale).

(D) Density plot of GC content of agave transcript contigs vs. contigs from contamination and commensal organisms.

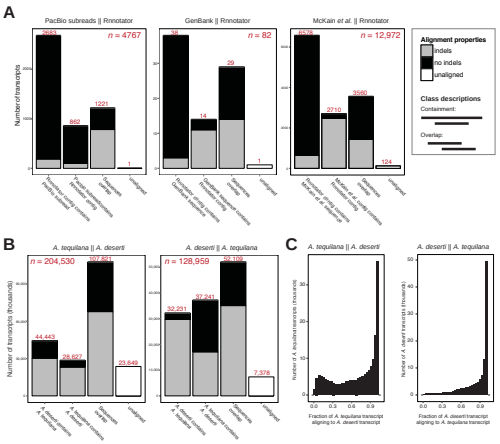
(E) Density plots of *A. deserti* and *A. tequilana* transcript lengths. Note $\log_{10}$ scale. Peaks at 150 and 250 nt represent single reads or paired-end reads, respectively, that were not assembled into larger contigs.

(F) Density plot of locus RPKM values for coding (dark shading) and non-coding (light shading) loci.

OVERVIEW OF *AGAVE* TRANSCRIPTOME ASSEMBLIES

| Species | Total Sequencing | No. of loci | No. transcript contigs | N50 length | Sum assembled length | No. protein-coding loci |
|---|---|---|---|---|---|---|
| *A. tequilana* | 293.5 Gbp | 139,525 | 204,530 | 1387 bp | 204.9 Mbp | 34,870 |
| *A. deserti* | 184.7 Gbp | 88,718 | 128,869 | 1323 bp | 125.0 Mbp | 35,086 |

## COMPARISON OF *AGAVE DE NOVO* ASSEMBLIES




Analysis of assembled contigs suggest the *Agave de novo* assemblies are comprehensive and accurate.

**FIGURE 3: Comparison of the de novo Agave transcriptome assemblies**

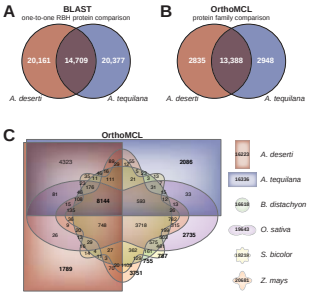
(A) Comparisons of the *A. tequilana de novo* Rnnotator assembly to error corrected Pacific Biosciences subreads, 82 GenBank *A. tequilana* sequences, and an additional *A. tequilana* dataset from McKain *et al.* 2012. [4]

(B) Comparisons between the *A. tequilana* and *A. deserti de novo* Rnnotator assemblies.

(C) Histograms of the fraction of aligned sequence lengths between *A. deserti* and *A. tequilana*.

Symbol || separates query sequence dataset from subject sequence dataset. Total number of sequences (*n*) is noted in each bar chart, total number of sequences in alignment classes are noted above bar.

## PROTEOMIC ANALYSES SUPPORT COMPREHENSIVE *AGAVE* TRANSCRIPTOME ASSEMBLIES




Proteome comparisons between *Agave* species and additional monocot species suggest the majority of *Agave* proteins are conserved across taxa. We can also identify protein families specific to agaves.

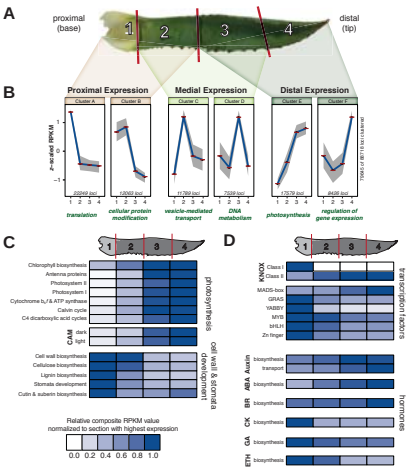
**FIGURE 4: Proteomic comparison of agaves to other plant species**

(A) Venn diagram of BLASTP-based one-to-one reciprocal best hit proteins shared between *A. deserti* and *A. tequilana*.

(B) Venn diagram of OrthoMCL-defined protein families shared between agaves.

(C) Edwards-Venn diagram of OrthoMCL-defined plant orthologous-group protein families (Plant OGs) shared between agave and 4 additional monocotyledonous plant species. Shape and color used for each species is at the right with the total number of Plant OGs within each species.

## PROFILING OF THE *A. DESERTI* LEAF HIGHLIGHTS REGIONS CRITICAL TO DEVELOPMENT AND PHOTOSYNTHESIS




Agaves spend the majority of their lives as compact rosettes, thus leaves are important organs in which to study *Agave* developmental and bioenergetic processes.

**FIGURE 5: Transcriptomic analysis of the *A. deserti* leaf proximal-distal axis.**

(A) One of the *A. deserti* leaves used for analysis, indicating proximal-distal (PD) sections 1–4.

(B) Six major *K*-means clusters of gene expression along the PD axis. Clusters are manually grouped by highest expression in proximal, medial, or distal tissues. Blue lines connect mean *z*-scaled RPKM values, shaded areas represent the 25th and 75th percentiles, red lines indicate standard error at each mean. Green text beneath each cluster denotes the description of the most significantly enriched GO term in each cluster.

(C, D) Heatmaps of composite gene expression for indicated biological processes along the leaf PD axis.

## ACKNOWLEDGEMENTS AND CITATIONS

This work performed at the U.S. Department of Energy Joint Genome Institute was supported in part by the Office of Science of the U.S. Department of Energy under contract DE-AC02-05CH112.

[1] Davis, A. S. *et al.* The global potential for *Agave* as a biofuel feedstock. GCB Bioenergy 3, 68–78, (2011).
[2] Somerville, C. *et al.* Feedstocks for lignocellulosic biofuels. Science 329, 790-2, (2010).
[3] Martin, J. *et al.* Rnnotator: an automated *de novo* transcriptome assembly pipeline from stranded RNA-seq reads. BMC Genomics 11, 663, (2010).
[4] McKain, M. *et al.* Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotates in Agavoideae (Asparagaceae). Am J Bot 99:2, 397–406.