Conserving the Molecular Profiles of California's Endangered Plants

Flint Mitchell; Weilun Tan; Hollie Mickelson; Chris Baysdorfer; Ana M.R. Almeida; Brian Perry; Joanna Garaventa; Bharti Parihar; Randall Davis

Department of Biological Sciences, California State University, East Bay, Hayward, CA

Abstract: California is home to over 6,000 plant species, a large portion of which are endemic to the state. As of November 2023, 386 plant species are at risk of extinction (ranked 1B.1) according to the California Native Plant Society due to factors such as climate change and urbanization driven habitat loss. Meaningful efforts in preserving these remaining populations requires foundational knowledge of the species' ecology, evolution, and increasingly- the plants genomic characteristics, in order to guide conservation and management strategies. The Green Biome Institute (GBI) at California State University East Bay (CSUEB) aims to contribute to the conservation of California's rare and endangered plant species by creating molecular profiles of over 300 rare plants by 2026 and making them freely available to the public and scientific community. These profiles will eventually include each plant's genomic sequence, transcriptome, universally recognized barcode regions, epigenome, metabolome, and leaf microbiome. To date, the GBI has produced approximately 30X coverage of short-read genome sequence data for over 100 species, along with genome assemblies for around 80 of these rare plants. Additional data generation and conservation efforts include long-read sequencing, genome size estimations by k-mer analysis, metabarcoding of endophytic foliar communities, as well as germplasm and propagation efforts. In addition, the GBI genome profiling program has created unique opportunities for engagement in meaningful research of a diverse student body at the High School and College undergraduate and graduate levels. Altogether, we believe these collective efforts will contribute towards data-driven conservation of California's plant diversity while growing tomorrow's science-based, conservation leaders.

Genome Assembly and Annotation of 80 Rare and Endangered California Plants

80 rare and endangered California plants were sequenced and analyzed using local and Amazon Web Services (AWS) hosted compute with the following results:

- Genome size estimates using Jellyfish (Marcais & Kingsford, 2011) and GenomeScope (Vurture et al., 2017):
  - Average genome size: 742 Mb
  - Largest: 3.35 Gb (***S. hallii 113***)
  - Smallest: 218 Mb (*Cercocarpus traskiae*)
- Draft assemblies for 80 species (Table 1):
  - Average BUSCO (Simão *et al.* 2015):
    - Complete (single and double-copy): 62.3%
    - Fragmented: 25.1%
    - Missing: 12.6%
  - QUAST (Mikheenko et al., 2018) metrics:
    - Average NG50: ~8.7 Kb
    - Average LG50: ~75.2 Kb

Please reach out to a GBI scientist for access to our sequencing data and to collaborate!

Table 1 - Draft assemblies for short-read data using Abyss 2.0 (Jackman et al. 2017)

Plant name	Genome size (Gb)	K value	BUSCO (viridiplantae_odb10)	# contigs	largest contig (bp)	Total length (bp)	N50 (bp)	NG50 (bp)	L50 (bp)	LG50 (bp)
Dudleya densiflora	0.223	127	C:99.1%[S:76.0%,D:23.1%],F:0.5%,M:0.4%,n:425	22118	883805	244178256	61470	68958	996	833
Castilleja mollis	0.48	125	C:96.7%[S:80.5%,D:16.2%],F:2.4%,M:0.9%,n:425	118807	126357	341535715	8298	3955	10454	22470
Galium buxifolium	0.57	128	C:96.0%[S:91.5%,D:4.5%],F:2.8%,M:1.2%,n:425	162041	137295	523971121	12460	10716	10836	12828
Carex xerophila	0.397	111	C:93.9%[S:90.1%,D:3.8%],F:4.7%,M:1.4%,n:425	122267	117176	425770334	9846	10738	11918	10519
Panicum acuminatum var. thermale	0.628	111	C:93.4%[S:88.9%,D:4.5%],F:5.6%,M:1.0%,n:425	92658	196237	604545650	21331	20342	7964	8528
Astragalus pycnostachyus var. lanosissimus	1.09	121	C:93.4%[S:41.2%,D:52.2%],F:6.4%,M:0.2%,n:425	227658	150584	686622053	9576	2593	19581	57827
Bensoniella oregona	0.48	128	C:93.0%[S:91.1%,D:1.9%],F:5.6%,M:1.4%,n:425	84137	179798	400585194	18192	13733	5980	8487
Castilleja hololeuca	0.43	111	C:93.0%[S:77.2%,D:15.8%],F:5.6%,M:1.4%,n:425	120123	110332	301916462	6258	2926	12595	27433
Eriogonum truncatum	0.66	121	C:92.9%[S:89.4%,D:3.5%],F:5.9%,M:1.2%,n:425	74891	517010	611146981	43345	38587	3693	4291
Agrostis hooveri	0.679	119	C:92.9%[S:27.3%,D:65.6%],F:6.1%,M:1.0%,n:425	112461	153094	620366228	16481	14727	10999	12880
Cercocarpus traskiae	0.218	119	C:92.7%[S:80.0%,D:12.7%],F:5.9%,M:1.4%,n:425	54908	221847	316238731	18032	27161	4578	2337
Rosa woodsii var. glabrata	0.39	109	C:92.5%[S:86.4%,D:6.1%],F:6.4%,M:1.1%,n:425	126619	82494	381864196	7857	7624	12930	13456
Hoita strobilina	0.46	121	C:91.0%[S:88.2%,D:2.8%],F:7.1%,M:1.9%,n:425	170195	231052	454173044	8275	8068	11050	11407
Suaeda esteroa	0.862	128	C:89.2%[S:30.6%,D:58.6%],F:10.4%,M:0.4%,n:425	203052	224697	829290328	13915	13143	15994	17203
Frangula purshiana ssp. ultramafica	0.25	125	C:85.6%[S:81.6%,D:4.0%],F:11.1%,M:3.3%,n:425	94399	179193	241289329	8146	7104	7032	8262
Ceanothus confusus	0.461	111	C:85.4%[S:76.5%,D:8.9%],F:11.8%,M:2.8%,n:425	180636	74643	480123937	6639	7024	19459	18058
Gambelia speciosa	0.746	111	C:84.7%[S:50.8%,D:33.9%],F:12.7%,M:2.6%,n:425	229880	161425	734295146	8611	8389	20783	21471
Clarkia imbricata	0.404	111	C:83.7%[S:62.1%,D:21.6%],F:13.2%,M:3.1%,n:425	258999	101656	535601701	4477	6461	29957	17675
Ceanothus maritimus	0.422	111	C:83.0%[S:73.6%,D:9.4%],F:14.1%,M:2.9%,n:425	129235	475159745	422000000	6967	8164	17250	13726
Lonicera subspicata var. subspicata	0.741	111	C:82.9%[S:78.4%,D:4.5%],F:14.8%,M:2.3%,n:425	209692	159280	744629424	11063	11150	16352	16189
C. amcali	0.54	111	C:82.1%[S:48.0%,D:34.1%],F:17.2%,M:0.7%,n:425	435006	2045981	978271211	5086	11004	44420	14463
Ceanothus roderickii	0.406	115	C:80.7%[S:61.9%,D:18.8%],F:16.9%,M:2.4%,n:425	185053	102347	477747148	6050	7299	21967	16559
Crocanthemum greenei	1.06	111	C:80.5%[S:23.8%,D:56.7%],F:17.2%,M:2.3%,n:425	280567	46155	682558520	5721	2074	35416	88507
Ceanothus divergens	0.381	111	C:80.0%[S:71.5%,D:8.5%],F:17.6%,M:2.4%,n:425	149607	202090	413426427	6873	7603	16240	13996
Calystegia stebbinsii	0.59	63	C:79.8%[S:77.9%,D:1.9%],F:17.2%,M:3.0%,n:425	108075	71549	228437500	3765	x	15677	x
Arctostaphylos virgata	0.54	119	C:77.6%[S:56.2%,D:21.4%],F:20.0%,M:2.4%,n:425	147468	235952	588541566	8676	9522	18882	16211
Ceanothus foliosus var. vineatus	0.48	121	C:75.8%[S:61.2%,D:14.6%],F:20.9%,M:3.3%,n:425	287522	73542	580685839	4170	5270	37150	26398
Eriogonum crocatum	1.281	111	C:75.3%[S:73.2%,D:2.1%],F:20.7%,M:4.0%,n:425	413421	104763	1047476677	6253	4234	41422	64164
Abronia alpina	0.418	111	C:75.3%[S:72.0%,D:3.3%],F:20.2%,M:4.5%,n:425	142907	108027	398469044	7628	7083	13546	14874
Berberis nevinii	0.97	121	C:75.1%[S:58.6%,D:16.5%],F:21.4%,M:3.5%,n:425	229870	252390	914528213	16874	14769	12106	13863
Arctostaphylos stanfordiana ssp. decumbens	0.514	105	C:74.8%[S:61.4%,D:13.4%],F:21.6%,M:3.6%,n:425	144371	155307	517392572	8040	8112	17463	17253
Arctostaphylos confertiflora	0.506	105	C:74.4%[S:49.9%,D:24.5%],F:23.1%,M:2.5%,n:425	143676	204438	589193574	8922	10364	18991	14660
Arctostaphylos silvicola	0.478	105	C:74.3%[S:48.7%,D:25.6%],F:23.3%,M:2.4%,n:425	147371	137592	614221075	8823	11157	20503	13463
Arctostaphylos morroensis	0.48	121	C:74.1%[S:49.6%,D:24.5%],F:22.8%,M:3.1%,n:425	149392	116555	621879961	9213	11487	19815	13393
Arctostaphylos pumila	0.456	111	C:74.1%[S:42.1%,D:32.0%],F:21.6%,M:4.3%,n:425	140187	102342	614525665	9508	12400	19066	11744

All of our best assemblies are analyzed using the below pipeline in order to do structural and functional annotation of the genome and chloroplast as well!

Figure 1 - N50 vs L50 with associated BUSCO completeness and genome size. Please refer to (**Kersey, 2019**) for a similar chart of assemblies submitted to the The International Nucleotide Sequence Database Collaboration.

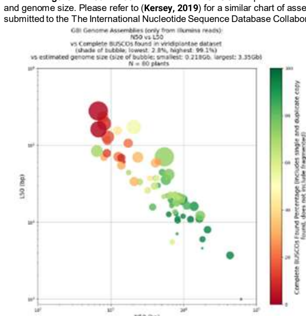
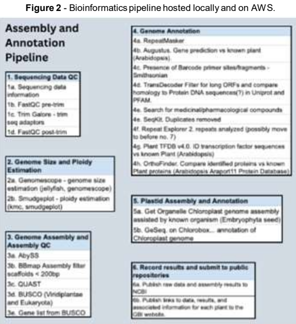

GBI Genome Assemblies (only from Illumina reads):
N50 vs L50
vs Complete BUSCOs found in viridiplantae dataset
(shade of bubble; lowest: 2.8%, highest: 99.1%)
vs estimated genome size (size of bubble; smallest: 0.218Gb, largest: 3.35Gb)
N = 80 plants
N50 (bp)
L50 (bp)

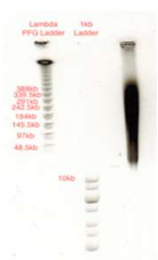
Figure 2 - Bioinformatics pipeline hosted locally and on AWS.


Assembly and Annotation Pipeline
1. Sequencing Data QC
1a. Sequencing data information
1b. FastQC pre-trim
1c. Trim Galore - trim seq adaptors
1d. FastQC post-trim
2. Genome Size and Ploidy Estimation
2a. Genomescope - genome size estimation (jellyfish, genomescope)
2b. Smudgeplot - ploidy estimation (kmc, smudgeplot)
3. Genome Assembly and Assembly QC
3a. AbySS
3b. BBmap Assembly filter scaffolds < 200bp
3c. QUAST
3d. BUSCO (Viridiplantae and Eukaryota)
3e. Gene list from BUSCO
4. Genome Annotation
4a. RepeatMasker
4b. Augustus. Gene prediction vs known plant (Arabidopsis)
4c. Presence of Barcode primer sites/fragments - Smithsonian
4d. TransDecoder Filter for long ORFs and compare homology to Protein DNA sequences(?) in Uniprot and PFAM.
4e. Search for medicinal/pharmacological compounds
4e. SeqKit. Duplicates removed
4f. Repeat Explorer 2. repeats analyzed (possibly move to before no. 7)
4g. Plant TFDB v4.0. ID transcription factor sequences vs known Plant (Arabidopsis)
4h. OrthoFinder. Compare identified proteins vs known Plant proteins (Arabidopsis Araport11 Protein Database)
5. Plastid Assembly and Annotation
5a. Get Organelle Chloroplast genome assembly assisted by known organism (Embryophyta seed)
5b. GeSeq. on Chlorobox... annotation of Chloroplast genome
6. Record results and submit to public repositories
6a. Publish raw data and assembly results to NCBI
6b. Publish links to data, results, and associated information for each plant to the GBI website.

Progress on Long reads Sequencing for California Rare and Endangered plants

BAYSDORFER LAB

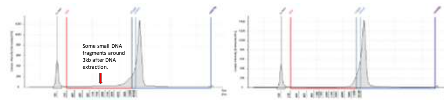
DNA Extraction Using CTAB



The DNA extracted via the CTAB protocol exhibits high molecular weight, with fragments from about 15kb to 388kb.

Size Selection using 2.5% PVP 360

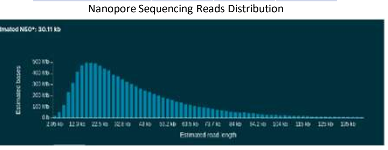
Before Size Selection (S.S.) / After Size Selection (S.S.)



The removal of short DNA fragments was achieved through the utilization of a 2.5% polyvinylpyrrolidone (PVP) 360 size selection buffer, resulting in an increased proportion of high molecular weight (HMW) DNA within the sample.

	Before S.S. (in 60µl volume)		After S.S. (in 40µl volume)	
Size range	Conc. (ng/µl)	Percentage	Conc. (ng/µl)	Percentage
250b -- > 60kb	53.5	100	55.1	100
15kb -- > 60kb	41.9	78.32	49.8	90.38
30kb -- > 60kb	35.9	67.10	40.8	74.05

Nanopore Sequencing Reads Distribution



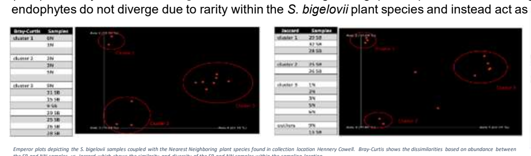
The N50 value for sequencing attained approximately 30 kb through the application of a ligation library preparation kit. Predominantly, sequencing coverage was centered on DNA fragments ranging from 15 kb to 20 kb, with the longest sequenced fragments exceeding 1 Mb in length.

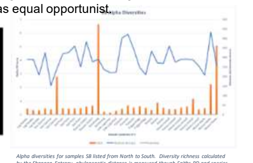
Analysis of the Foliar Endophytic Fungal Community of *Scoliopus bigelovii*, a native lily from the California Redwoods

PERRY LAB

Fungal endophytes are a highly diverse group of fungi that naturally reside within the tissues of plants and often exhibit a symbiotic relationship with the host individual. Previous research suggests symbionts could be important for maintaining rare species communities (Chung, Miller and Rudgers 2015) by contributing to overall fitness, stress tolerance and increased population growth (Rodriguez et al. 2009). Despite the growing amount of research on this topic, endophytic fungi remain understudied, undocumented, and poorly understood for most plant host taxa including *Scoliopus bigelovii*, a species of lily (Angiosperms, Liliaceae) endemic to the redwood habitats of coastal California. This flower dwells in the shady understory of the giant trees below 2,000 ft elevation, nestled between ferns, pacific trillium, and redwood sorrel. The rarity, restricted distribution, and uniqueness of the habitat of *S. bigelovii* drives the hypothesis that the endophytic community have a potential to also be rare and restricted throughout the endemic range. So far, no studies suggest that *S. bigelovii* inhabit any rare fungal species associated with a northern to southern distribution. Here we use Ion Torrent single end read sequences processed through Qiime2 to show no distinct pattern between *S. bigelovii* and nearest neighboring plant species as well as between sampling sites within the endemic range. We used α & β diversity indices to compare microbiomes though species richness, similarity, evenness, distribution, and dissimilarity to look for patterns associated with foliar fungal endophyte (FFE) diversity between *S. bigleovii* and nearest neighboring plant species. Concluding that the foliar fungal endophytes do not diverge due to rarity within the *S. bigelovii* plant species and instead act as equal opportunist.

M.S. Student
Hollie Mickelson



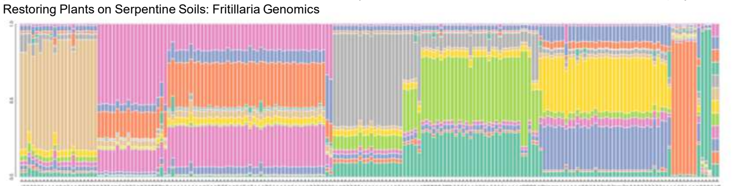


Population structure of rare and endangered California Fritillaria

BAYSDORFER LAB

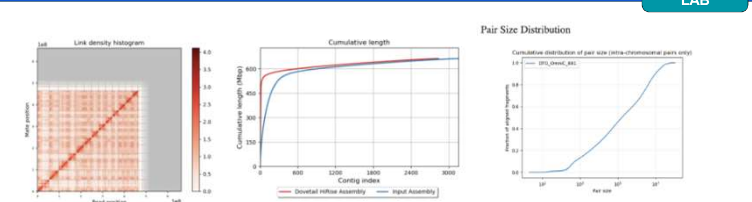
Over 200 plant leaf samples of various Fritillaria were obtained and DNA was isolated using the Qiagen Plant DNeasy protocol. DNA samples were quality checked by nanodrop and Qubit. 190 samples passed quality control and were normalized and sent to Floragenex (Beaverton Oregon) for ddRAD sequencing.

Analysis showed the presence of 11 groups. Four of these correspond to known Fritillaria species, these include F affinis, F pudica, F striata, and one population of F purdyi. Three groups correspond to F biflora with one including plants from LA, Riverside and San Diego counties, a second from SLO and Monterey counties and a third (close to the second) corresponding to plants from Santa Barbara County. One group corresponds to all plants labeled as F roderickii/grayana and F. liliacea from Nicasio reservoir. One group corresponds to plants labeled as F agrestis from Contra Loma and Diablo foothills and F liliacea from Jepson prairie. One group corresponds to plants labeled as F liliacea from Lake Chabot and F. agrestis from Brushy peak, Alameda watershed, Tesla and Greenville. One group corresponds to plants labeled as F liliacea from Limontour and Pulgas, F. agrestis from regional parks and all plants labeled as F. biflora ineziana. This work was funded by the San Francisco Public Utilities Commission as a project entitled: Restoring Plants on Serpentine Soils: Fritillaria Genomics



Draft Genome Assembly of Arctostaphylos Pallida

ALMEIDA LAB



References:

Jackman, SD; Vandervalk, BP; Mohamadi, H; Chu, J; Yeo, S; Hammond, SA; Jahesh, G; Khan, H; Coombe, L; Warren, RL; and Birol, I. (2017). **ABySS 2.0: Resource-efficient assembly of large genomes using a Bloom filter**. *Genome research*, 27(5): 768-777. doi:10.1101/gr.214346.116

Ramasamy, R.K., Ramasamy, S., Bindroo, B.B. *et al*. (2014) **STRUCTURE PLOT: a program for drawing elegant STRUCTURE bar plots in user friendly interface.** *SpringerPlus* **3**, 431. https://doi.org/10.1186/2193-1801-3-431

Alla Mikheenko, Andrey Prjibelski, Vladislav Saveliev, Dmitry Antipov, Alexey Gurevich, **Versatile genome assembly evaluation with QUAST-LG,** Bioinformatics (2018) 34 (13): i142-i150. doi: 10.1093/bioinformatics/bty266. First published online: June 27, 2018

Simão, F.A.; Waterhouse, RM; Ioannidis,P; Kriventseva, EV; Zdobnov, EM. (2015). **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**, *Bioinformatics*, 31(19): 3210–3212, https://doi.org/10.1093/bioinformatics/btv351

Guillaume Marcais and Carl Kingsford, **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.** Bioinformatics (2011) 27(6): 764-770 (first published online January 7, 2011) doi:10.1093/bioinformatics/btr011

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). **GenomeScope: fast reference-free genome profiling from short reads**. *Bioinformatics*, 33(14), 2202–2204.https://doi.org/10.1093/bioinformatics/btx153

Kersey, P. J. (2019). **Plant genome sequences: past, present, future**. *Current Opinion in Plant Biology, 48*, 1-8. https://doi.org/10.1016/j.pbi.2018.11.001

Acknowledgments:

East Bay Regional Park District
San Francisco Botanical Garden
University of California Botanical Garden at Berkeley
Vine Hill Preserve

***Green Biome Institute Contact Info*:**

Weilun Tan (Scientist)
wtan8@horizon.csueastbay.edu
Chris Baysdorfer, PhD (Director)
chris.baysdorfer@csueastbay.edu
Brian Perry, PhD (Associate Director)
brian.perry@csueastbay.edu
Ana M.R. Almeida, PhD (Associate Director)
ana.almeida@csueastbay.edu
Flint Mitchell (Bioinformatics Engineer)
flint@digimitch.com
Randy Davis (Advisor)
randydavis2018@gmail.com