Morus Rubra Identification Experiment Methodology
Summary by Nigel Daniels and Colin Kruse

DNA Extraction

Mulberry leaves were shipped to the laboratory, packaged in sealed plastic bags containing a moistened paper towel. Upon receipt, the plastic bags containing the leaves were placed in a -20°C freezer. Genomic DNA was extracted from the leaves using the Quick-DNA Miniprep Kit (Zymo Research, Irvine, California, USA) according to the manufacturer's instructions for solid tissue samples. Briefly, leaves were allowed to thaw on ice, before 25 mg of tissue was excised and placed in a 1.5 mL DNase- and RNase-free microfuge tube. The tissue was mechanically homogenized in 500 µL of Quick-DNA Genomic Lysis Buffer using an RNase, DNase and Pyrogen-free Disposable Pellet Pestle (Thermo Fisher Scientific, Waltham, Massachusetts, USA). Genomic DNA was bound to the Zymo-Spin™ IICR Column by centrifugation at 12,000 x g for one minute followed by centrifugation washes at the same conditions with 200 µL of DNA Pre-Wash Buffer and 500 µL of g-DNA Wash Buffer. Finally, genomic DNA was collected with 50 µL of DNA Elution Buffer prewarmed to 60°C by centrifugation at 21,000 x g for 30 seconds. Eluted genomic DNA was assessed for purity using a Nanodrop ND1000 Spectrophotometer (Thermo Fisher Scientific) and quantified using a Qubit 3.0 Fluorometer (Thermo Fisher Scientific) before being stored at -20°C until use. Samples numbered 1-16 were extracted previously, following this protocol, by researchers at UVA Wise.

Genomic Sequencing

Twelve mulberry leaf genomic DNA samples, selected by Weston Lombard as most likely to originate from *Morus rubra* leaves based on morphology, were assessed for fragment length by analysis using a Bioanalyzer High Sensitivity DNA Assay (Agilent Technologies, Santa Clara, California, USA). Two of the samples that contained the longest average fragments were selected for sequencing: #61 "Spillway, Lucky Pittman" and #91 "Roberts Farm #6, Female M.rubra". These two samples were sent to SeqCenter LLC (Pittsburgh, Pennsylvania, USA) for both short-read sequencing using an Illumina NovaSeq X Plus (San Diego, California, USA) and long-read sequencing using either an Oxford Nanopore MinION Mk1B or an Oxford Nanopore GridION (New York, New York, USA). The Illumina sequencing details, according to SeqCenter, were "Illumina sequencing libraries were prepared using the tagmentation-based and PCR-based Illumina DNA Prep kit and custom IDT 10bp unique dual indices (UDI) with a target insert size of 280 bp. No additional DNA fragmentation or size selection steps were performed. Illumina sequencing was performed on an Illumina NovaSeq X Plus sequencer in one or more multiplexed shared-flow-cell runs, producing 2x151bp paired-end reads. Demultiplexing, quality control and adapter trimming was performed with bcl-convert (v4.2.4)." The Oxford Nanopore sequencing details, according to SeqCenter, were "Sample libraries were prepared using the PCR-free Oxford Nanopore Technologies (ONT) Ligation Sequencing Kit (SQK-NBD114.24) with the NEBNext® Companion Module (E7180L) to manufacturer's specifications. No additional DNA fragmentation or size selection was performed. Nanopore sequencing was performed on an Oxford Nanopore a MinION Mk1B sequencer or a GridION sequencer using R10.4.1 flow cells in one or more multiplexed shared-flow-cell runs. Run design utilized the 400bps sequencing mode with a minimum read length of 200bp. Adaptive sampling was not enabled. Guppy (v6.5.7) was used for super-accurate basecalling (SUP), demultiplexing, and adapter removal (dna_r10.4.1_e8.2_400bps_modbases_5mc_cg_sup.cfg)." The total number of reads

generated for the samples were 98518146 Illumina reads and 968007 Nanopore reads for #61; and 91077715 Illumina reads and 1128209 Nanopore reads for #91.

Sequence Analysis

A number of assembly and assembly-by-reference strategies were employed to assemble the red mulberry genome. These efforts are ongoing at Los Alamos National Laboratory (LANL) to produce a draft-quality genome despite the complexity of this plant genome. To develop reliable PCR probes to identify the genetic composition of mulberry samples, a variant analysis approach was ultimately the most successful. Samples #61 and #91 were compared against the white mulberry genome (GCA_012066045.3) with a LANL proprietary variant caller for potential primer sites. The LANL variant caller includes a mix of a k-mer based approach and a deep learning variant caller, DeepVariant (Poplin et al., 2018). K-mers are substrings of a nucleotide sequence of k length that elucidate differences that may not be detectable in the full length of the sequence. To ensure no cross-species alignment artifacts caused miscalled variants, a separate DeepVariant analysis using BWA-based alignments was performed. DeepVariant is a variant caller developed by Google and utilizes a convolutional neural network to identify and report variants. Additionally, in the pipeline, the samples were sent through a quality control step using fastp and aligned to the white mulberry reference genome with bwa mem (Chen, 2023, Li, 2013). The two samples were compared against each other in multiple steps of the variant calling process, both before and after the variant calling, to minimize the effect of using an imperfect reference genome as well as find the variants that are shared between the two samples. The variant list was then filtered for insertions or deletions (indels) over 100 base pairs. This variant calling and filtering identified 45 variants found in both samples #61 and #91. These variants were found across 15 mulberry contigs. A total of 36 of the 45 variants were called using the BWA-based variant calling and were selected for primer design. Primers were designed to verify the presence/absence of each indel using the Geneious implementation of Primer3 (Geneious Prime 2025.1).

Quantitative PCR

Primers targeting the genomic regions unique to *M. rubra* and *M. alba* were designed using Primer3.

Table 1. Quantitative PCR primers targeting unique genomic regions in *M. rubra* and *M. alba*.

Primer name	Primer sequence
Rub236-5F	TCCTTGTTGGAGATGGATGTTAG
Rub232-28F	AAGTCTGGTTGAAAGAATTTATAGTGG
Rub236-90R	AAGATCAGCGCCTACACCTG
Rub232-102R	TCTTCATGGCTTAAAAAGACTCATAAT
Alb233-79F	CTTACATAAAGTCACATCTCAACTCG
Alb233-165R	CACGCACCAACTTTAAATAAAAAGTA

Quantitative PCR (qPCR) reactions were prepared using the KAPA SYBR FAST qPCR Master Mix (2X) Kit (Roche Diagnostics, Indianapolis, Indiana, USA) according to manufacturer's instructions. Reactions of 10 µL total volume were analyzed using an AriaMx Real-time PCR System (Agilent Technologies) with a final concentration of 200 nM for each primer and 10 ng of genomic DNA. Thermocycling conditions were as follows: 95°C for 3 mins, 40 cycles of 95°C for 3 seconds then 60°C for 20 seconds with a fluorescence scan at the end of each cycle. QPCR results are presented as an average of technical duplicates, analyzed using Aria Real-Time PCR Software (Agilent

Technologies) to produce a Threshold Cycle (Ct) value. This is the cycle number at which the fluorescence signal crosses a predetermined threshold above background fluorescence.

Results

Ct values for each sample were compared for differences between the detection of the species-specific genomic regions labeled Rub232, Rub236, and Alb233. Differences of ~10 cycles between specific-region Ct values were used as an indication of whether a sample was likely *M. rubra*, *M. alba*, or a hybrid. A lower Ct value indicates a higher quantity of genomic DNA containing the region targeted by that primer pair present in the reaction. Thus, values of ~18-21 for the Rub regions and >~28 for the Alb region indicate a likely *M.rubra* species identification. Values opposite to this indicate likely *M.alba,* whereas ~18-21 values for any combination of Rub and Alb regions indicate a likely hybrid.

Table 2. Summary of Quantitative PCR Threshold Cycle (Ct) values for the species-specific genomic regions Rub232, Rub236, and Alb233 produced from Mulberry leaf genomic DNA samples. "Indicated Species" denotes the species of the sample based on the regions tested, but does not absolutely guarantee the absence of DNA from other species in untested regions.

Sample ID	Rub232 Ct	Rub236 Ct	Alb233 Ct	Indicated Species
1	19.02	18.99	28.37	Rubra
2	24.48	24	34.35	Mostly Rubra
3	20.56	20.99	36.5	Rubra
4	19.28	19.17	29.85	Rubra
5	21.98	21.95	21.78	Hybrid
6	20.61	20.15	20.15	Hybrid
7	19.42	20.02	19.94	Hybrid
8	20.45	19.91	27.44	Mostly Rubra
9	19.84	19.32	28.12	Rubra
10	19.98	19.44	31.97	Rubra
11	24.26	24.27	37.52	Rubra
12	Duplicate sample			
13	20.17	20.51	25.84	Hybrid
14	20.84	20.11	20.15	Hybrid
15	21.85	21.61	32.48	Mostly Rubra
16	21.73	21.67	23.87	Hybrid
17	19.1	19.03	30.02	Rubra
18	None	35.73	19.94	Alba
19	18.83	18.71	30.47	Rubra
20	18.92	18.81	30.08	Rubra
21	18.82	18.82	29.72	Rubra
22	19.1	18.94	29.68	Rubra
23	19.52	19.06	31.31	Rubra
24	18.41	17.77	28.36	Mostly Rubra
25	None	None	None	Unknown
26	18.91	18.42	29.25	Rubra
27	18.87	18.58	29.8	Rubra
28	Duplicate sample			
29	18.73	18.3	29.38	Rubra

30	Duplicate sample			
31	18.55	18.64	30.69	Rubra
32	19.08	19.04	31.11	Rubra
33	18.37	18.48	30.26	Rubra
34	20.11	20.02	19.89	Hybrid
35	18.22	18.2	29.55	Rubra
36	18.3	33.33	17.03	Hybrid
37	32.52	33.51	19.33	Alba
38	32.99	33.89	20.93	Alba
39	None	30.2	17.38	Alba
40	18.99	19.1	18.76	Hybrid
41	18.19	29.07	18.53	Hybrid
42	33.89	31.21	18.86	Alba
43	19.91	32.96	19.28	Hybrid
44	18.38	17.92	28.59	Mostly Rubra
45	18.37	18.13	25.79	Mostly Rubra
46	19.61	18.98	19.19	Hybrid
47	18.68	32.71	17.9	Hybrid
48	19.5	30.19	18.1	Hybrid
49	33.09	31.83	18.98	Alba
50	None	33.66	22.28	Alba
51	20.09	19.9	27.29	Mostly Rubra
52	33.14	30.55	19.26	Alba
53	33.7	33.87	18.13	Alba
54	33.17	32.25	18.36	Alba
55	18.5	28.89	18.26	Hybrid
56	19.24	18.77	30.52	Rubra
57	Duplicate sample			
58	Duplicate sample			
59	19.44	18.99	19.14	Hybrid
60	19.6	19.18	27.08	Mostly Rubra
61	19.17	18.07	29.57	Rubra
62	36.05	32.82	19.08	Alba
63	19.49	18.86	35.6	Rubra
64	Duplicate sample			
65	19.58	19.61	31.55	Rubra
66	33.6	21.84	20.41	Hybrid
67	18.8	18.53	29.92	Rubra
68	19.68	19.42	30.49	Rubra
69	20.58	32.86	19.89	Hybrid
70	19.33	19.23	31.04	Rubra
71	19.56	19.39	30.93	Rubra
72	20.89	20.39	32	Rubra
73	19.73	19.22	31.33	Rubra
74	18.79	18.43	28.94	Rubra
75	19.69	19.14	30.91	Rubra
76	20.34	19.7	31.53	Rubra
77	19.65	19.51	31.66	Rubra
78	20.27	20.38	29.57	Rubra

79	18.52	18.49	30.84	Rubra
80	18.77	18.7	30.66	Rubra
81	19.25	18.91	31.19	Rubra
82	18.85	18.77	29.68	Rubra
83	18.52	18.18	30.25	Rubra
84	19.89	32.95	18.88	Hybrid
85	20.72	31.06	19.51	Hybrid
86	19.97	19.45	22.6	Hybrid
87	19.18	18.54	21.96	Hybrid
88	20.7	20.22	20.52	Hybrid
89	19.01	18.87	30.74	Rubra
90	18.27	18.45	30.62	Rubra
91	18.92	18.92	31.18	Rubra
92	19.2	19.2	31.94	Rubra
93	18.98	18.93	31.49	Rubra
94	20.35	19.28	31.58	Rubra
95	21.35	30.7	20.94	Hybrid
96	19.66	19.44	33.1	Rubra
97	19.3	18.81	31.68	Rubra
98	20.62	19.2	32.65	Rubra
99	19.8	19.42	33.05	Rubra
100	20.19	19.59	33.29	Rubra
101	19.16	19	29.58	Rubra
102	19.07	19.08	30.63	Rubra

References

Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. Imeta. 2023 May 8;2(2):e107. doi: 10.1002/imt2.107. PMID: 38868435; PMCID: PMC10989850.

Li, Heng. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 1303.3997 (2013). https://arxiv.org/abs/1303.3997

Poplin, R., Chang, PC., Alexander, D. et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 36, 983–987 (2018). https://doi.org/10.1038/nbt.4235