communications biology

ARTICLE

https://doi.org/10.1038/s42003-023-05748-4 OPEN

High quality genomes produced from single MinION flow cells clarify polyploid and demographic histories of critically endangered *Fraxinus* (ash) species

Steven J. Fleck[1✉], Crystal Tomlin[1], Flavio Augusto da Silva Coelho[1], Michaela Richter[1], Erik S. Danielson[2], Nathan Backenstose[1], Trevor Krabbenhoft[1], Charlotte Lindqvist[1] & Victor A. Albert[1✉]

Communications Biology **7**, Article number: 54 (2024) | Cite this article

Red list IUCN (2017) - Oleaceae

- **White ash (Fraxinus americana)**
- **Black ash (Fraxinus nigra)**
- **Green ash (Fraxinus pennsylvanica)**
- Pumpkin ash (Fraxinus profunda)
- Blue ash (Fraxinus quadrangulata)

- The three ash species were sequenced using a single ONT MinION flow cell each.

- Advancements in homology-based gene prediction allowed us to produce high quality genome annotations without needing to produce RNA-seq reads for our samples.

Uma flow cell de MinION seria suficiente para anotação de genomas?

Table 1 Assembly and annotation statistics for *Fraxinus* species after running Purge Haplotigs.

Assembly	***F. americana***	***F. nigra***	***F. pennsylvanica***	***F. pennsylvanica* v1.4**
# contigs	4364	2544	6764	110
Largest contig	3,688,385	6,590,250	6,486,905	56,547,140
Est. Total length	875 Mbp	829 Mbp	869 Mbp	869 Mbp
Total length	851,858,478	776,258,169	841,639,580	756,791,283
GC (%)	35.26	34.76	35.2	34.40%
N50	507,263	1,081,099	244,798	33,221,578
L50	450	206	928	10
# N's per 100 kbp	0.80	0.48	0.62	12,120.86
Complete BUSCOs	1561 (96.7%)	1561 (96.7%)	1553 (96.2%)	1576 (97.6%)
Complete single-copy BUSCOs	1281 (79.4%)	1302 (80.7%)	1271 (78.7%)	1308 (81.0%)
Complete duplicated BUSCOs	280 (17.3%)	259 (16.0%)	282 (17.5%)	268 (16.6%)
Fragmented BUSCOs	37 (2.3%)	25 (1.5%)	37 (2.3%)	26 (1.6%)
Missing BUSCOs	16 (1.0%)	28 (1.8%)	24 (1.5%)	12 (0.8%)
Total BUSCOs searched	1614	1614	1614	1614
Annotation	***F. americana***	***F. nigra***	***F. pennsylvanica***	***F. pennsylvanica* v1.4**
gene model/mRNA count	41,093/46,464	37,496/42,777	40,446/45,696	35,470/35,470
Complete BUSCOs	96.9% (1564)	97.3% (1571)	97.6% (1575)	82.5% (1332)
Complete single-copy BUSCOs	78.7% (1271)	80.3% (1296)	79.4% (1281)	70.8% (1142)
Complete duplicated BUSCOs	18.2% (293)	17.0% (275)	18.2% (294)	11.8% (190)
Fragmented BUSCOs	0.9% (15)	0.9% (14)	1.0% (16)	2.4% (38)
Missing BUSCOs	2.2% (35)	1.8% (29)	1.4% (23)	15.1% (244)
Total BUSCOs searched	1614	1614	1614	1614

Species names are italicized.
F. pennsylvanica v1.4 is the reference assembly from Huff et al.[17].

- **Fraxinus americana e F. pennsylvanica**:
  - Contiveram muitas regiões haplotípicas não colapsadas
  - amostras altamente heterozigóticas.

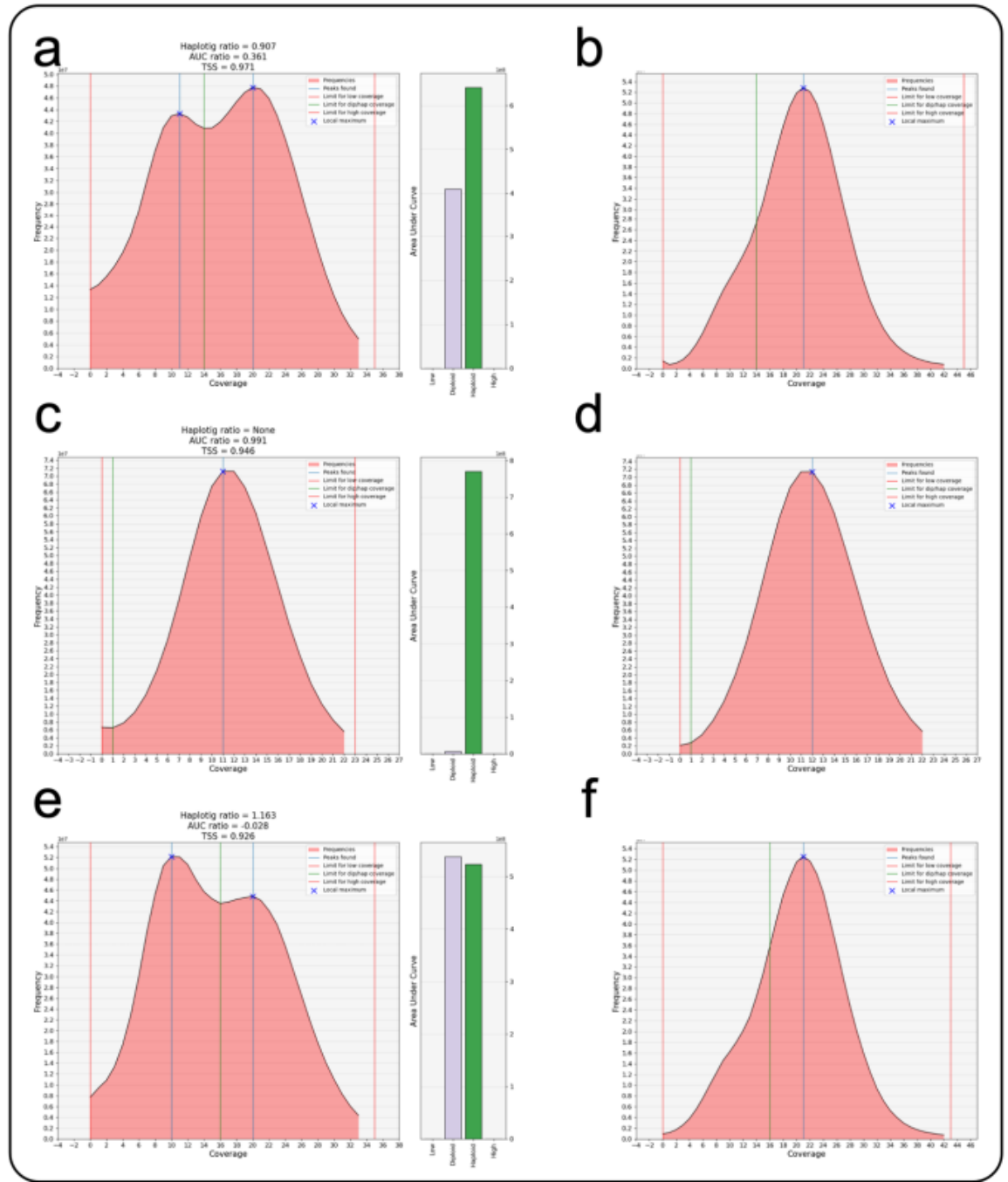
Research article | Open access | Published: 05 June 2021

Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms

Nadège Guiglielmoni, Antoine Houtain, Alessandro Derzelle, Karine Van Doninck & Jean-François Flot

BMC Bioinformatics **22**, Article number: 303 (2021) | Cite this article

Figure S2: Haploidy assessment of *Fraxinus* genome assemblies before and after running Purge Haplotigs.



We therefore generated largely haploid assemblies by post-processing using Purge Haplotigs[31], resulting in similarly high levels of complete BUSCOs of 96.7%, 96.7%, and 96.2% in *Fraxinus americana*, *F. nigra*, and *F. pennsylvanica*, respectively (Table 1). HapPy detected only haploid peaks for all three purged assemblies (Fig. S2b, d, f), and self-vs-self Ks frequency

Home › BMC Bioinformatics › Article

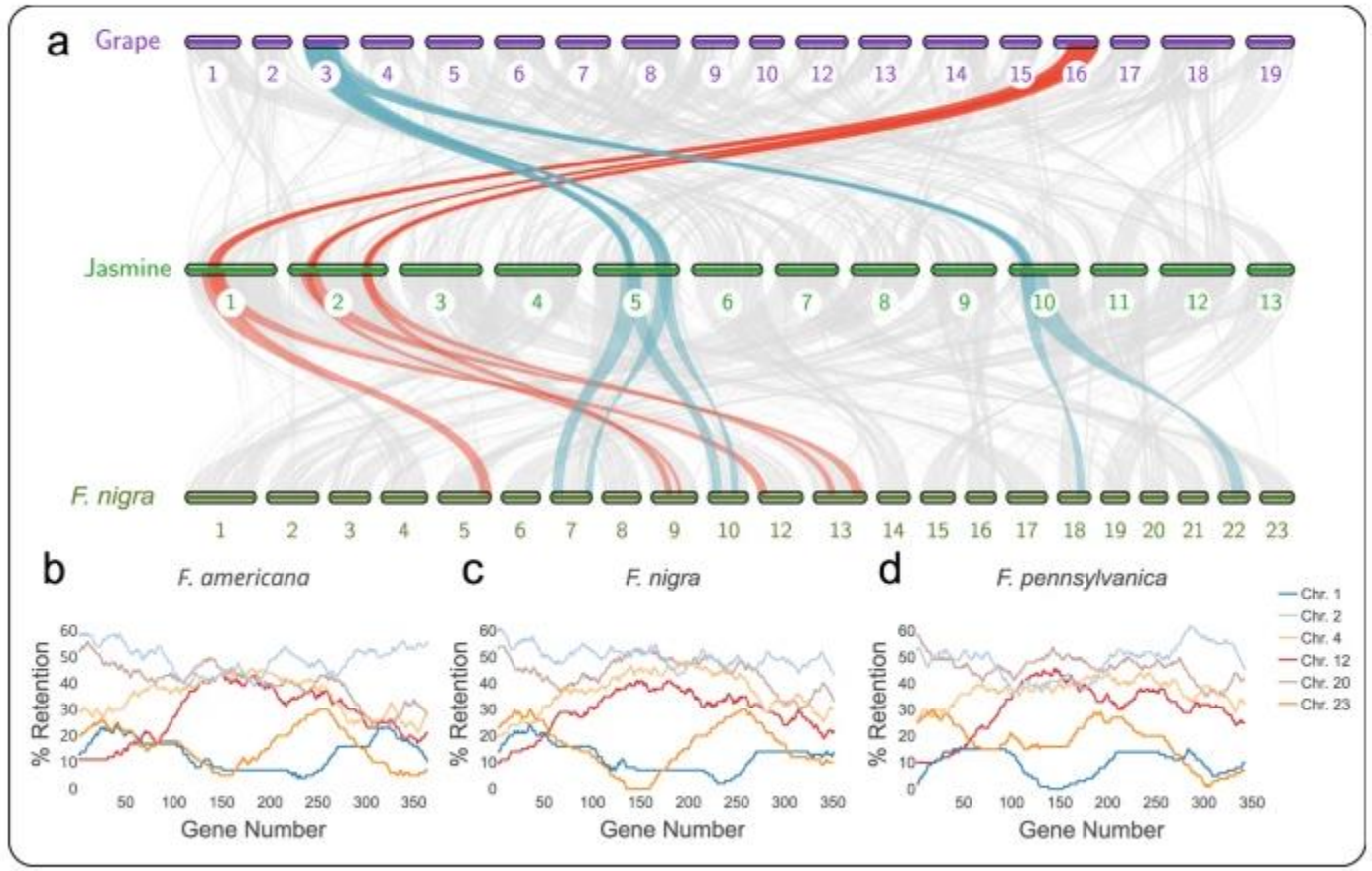
Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies

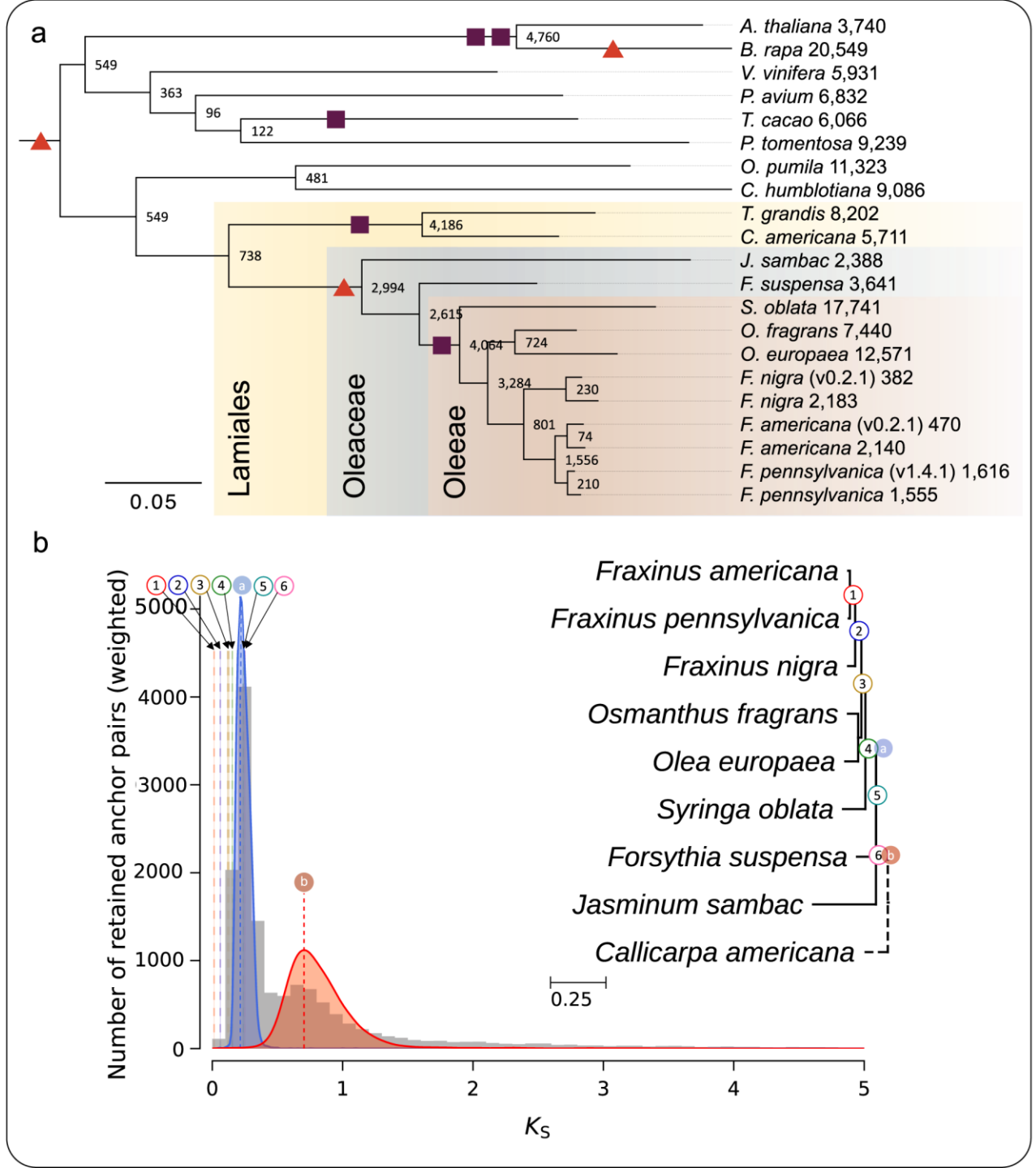
Software | Open access | Published: 29 November 2018

Volume 19, article number 460, (2018) Cite this article

Fraxinus and *Vitis vinifera*, which lacks any additional polyploidy events beyond the *gamma* hexaploidy event[35], each *Fraxinus* species showed a six-to-one ratio against *V. vinifera* (Fig. 1b,

Fig. 1: Polyploidy events in jasmine and ash since their divergence from grapevine.



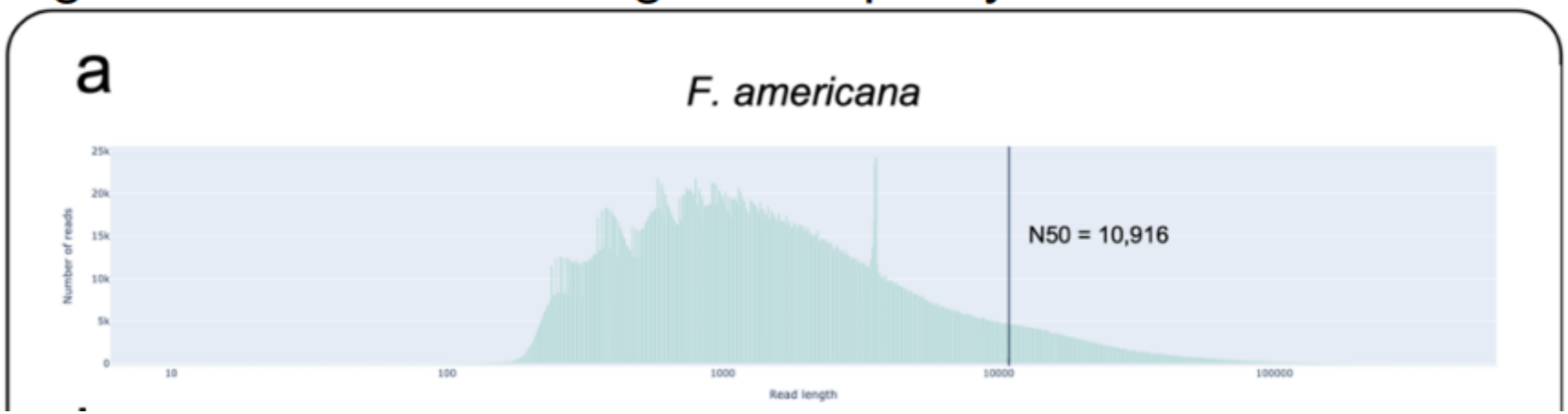

a
A. thaliana 3,740
B. rapa 20,549
V. vinifera 5,931
P. avium 6,832
T. cacao 6,066
P. tomentosa 9,239
O. pumila 11,323
C. humblotiana 9,086
T. grandis 8,202
C. americana 5,711
J. sambac 2,388
F. suspensa 3,641
S. oblata 17,741
O. fragrans 7,440
O. europaea 12,571
F. nigra (v0.2.1) 382
F. nigra 2,183
F. americana (v0.2.1) 470
F. americana 2,140
F. pennsylvanica (v1.4.1) 1,616
F. pennsylvanica 1,555
4,760
549
363
96
122
481
549
4,186
738
2,994
2,615
4,064
724
3,284
230
801
74
1,556
210
0.05
Lamiales
Oleaceae
Oleeae
b
Number of retained anchor pairs (weighted)
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
K_S
Fraxinus americana
Fraxinus pennsylvanica
Fraxinus nigra
Osmanthus fragrans
Olea europaea
Syringa oblata
Forsythia suspensa
Jasminum sambac
Callicarpa americana
0.25

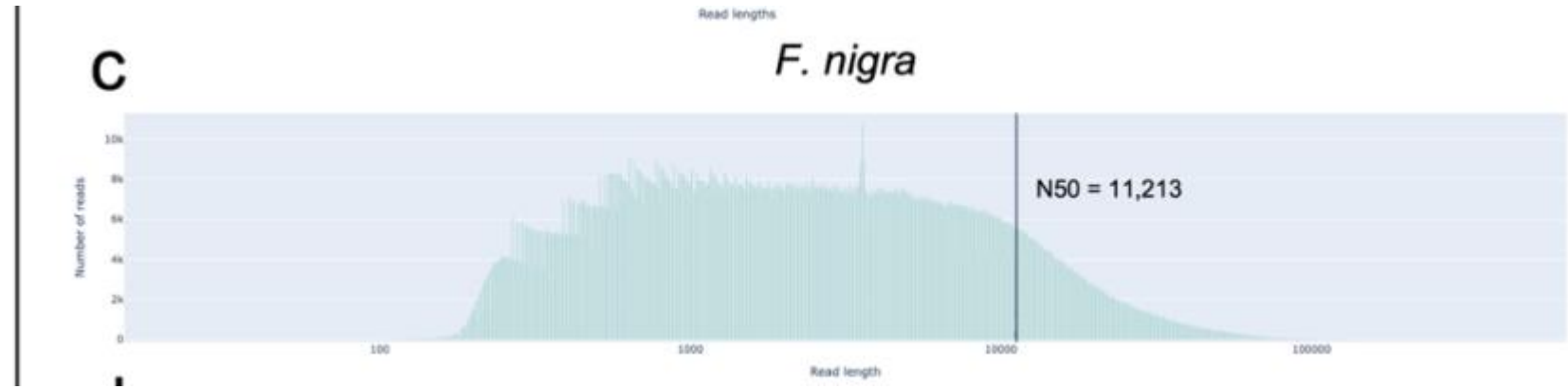
Methods

DNA extraction and sequencing

We followed the BioNano NIBuffer nuclei isolation protocol and the library preparation procedure from Pacific Biosciences – "Preparing Arabidopsis Genomic DNA for Size-Selected ~20 kb SMRTbellTM Libraries". Additionally, Circulomics Short Read Eliminator (SRE) was used to remove reads less than 25 kb in length with all samples. DNA sequencing was carried out on an ONT' GridION instrument utilizing MinION flowcells (version R9.4). Genomic DNA libraries were prepared with the ONT ligation kit SQK-LSK110. Each flow cell underwent two washes in order to maximize the sequencing output. Sequenced reads were basecalled with the high-accuracy model in Guppy version 5.0.11. Read quality was assessed with Nanostat version 1.5.0 and NanoPlot version 1.38.0[62] (Table S1, Fig. S1).

Figure S1: Raw read length and quality distributions for *Fraxinus* spp..




Read lengths
c
F. nigra
N50 = 11,213
Number of reads
Read length

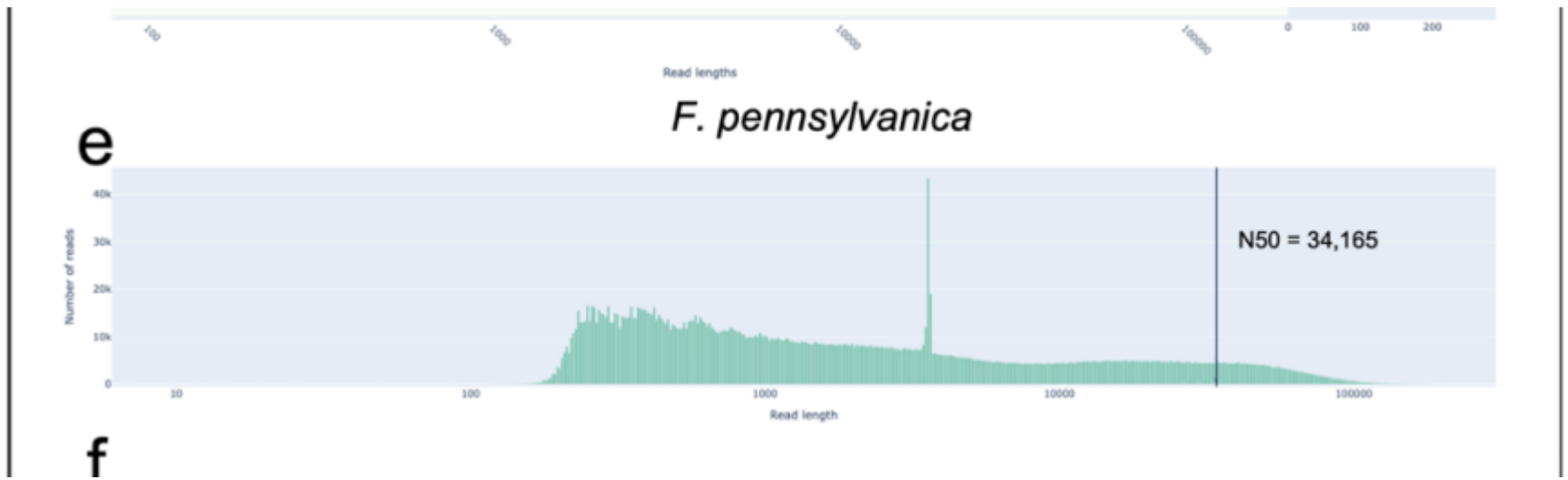

Read lengths
e
F. pennsylvanica
N50 = 34,165
Number of reads
Read length

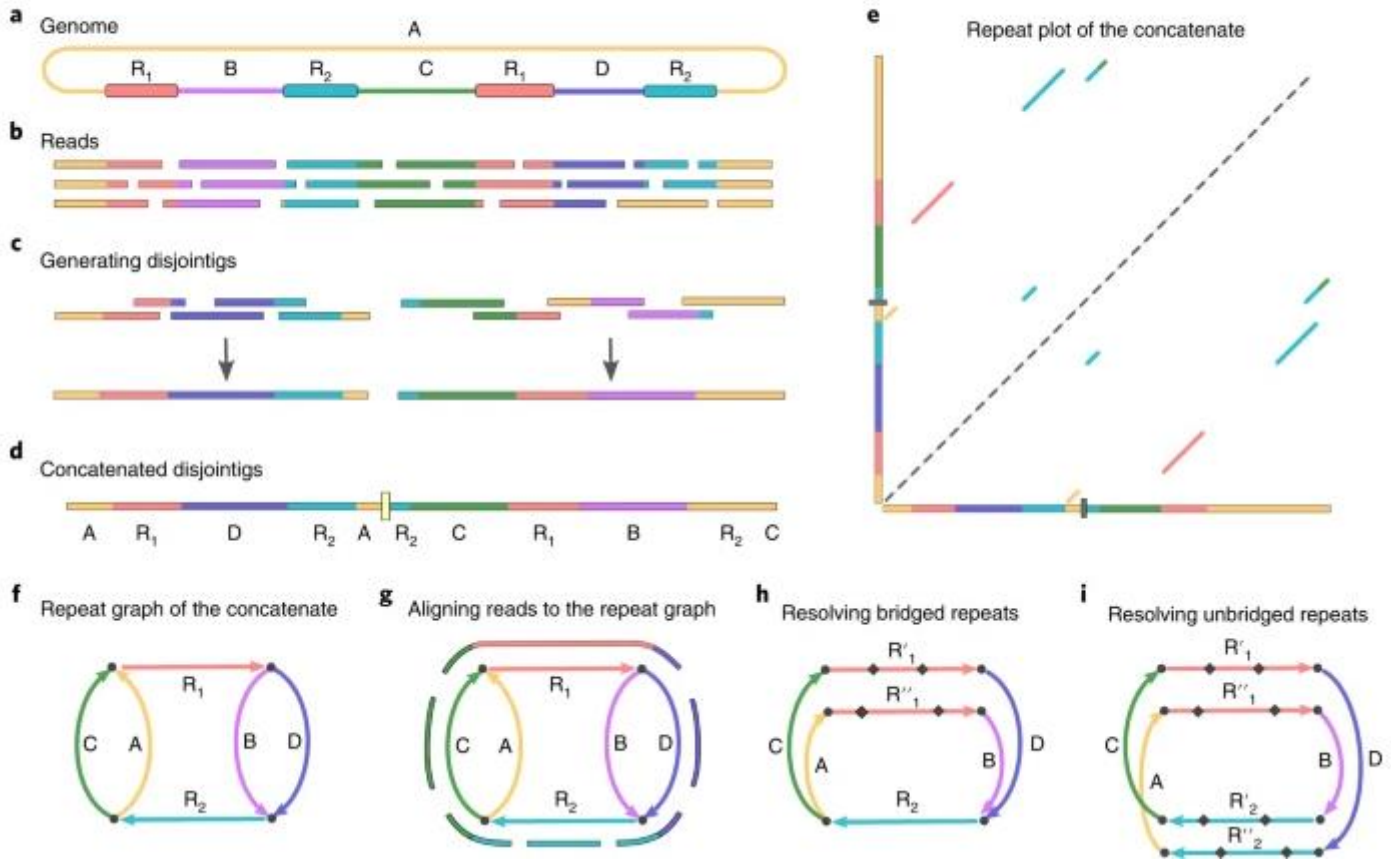
Table S1: Nanoplot sequencing stats for *Fraxinus* spp.

Sample	*Fraxinus americana*	*Fraxinus nigra*	*Fraxinus pennsylvanica*
Mean read length	3,600.80	4,893.90	7,735.80
Mean read quality	10.9	11.2	11.4
Median read length	1,266	2,106	1,319
Median read quality	11.1	11.6	11.7
Number of reads	6,065,809	3,788,248	2,581,752
Read length N50	10,916	11,213	34,165
STDEV read length	7,857.10	7,447.30	15,418.20
Total bases	21,841,746,309	18,539,126,624	19,971,958,290
Number, percentage and megabases of reads above quality cutoffs			
>Q5	5,930,147 (97.8%) 21,585.6Mb	3,674,454 (97.0%) 18,219.9Mb	2,535,016 (98.2%) 19,830.2Mb
>Q7	5,497,205 (90.6%) 20,368.9Mb	3,325,444 (87.8%) 16,891.4Mb	2,357,593 (91.3%) 18,979.0Mb
>Q10	3,975,554 (65.5%) 15,529.1Mb	2,564,692 (67.7%) 13,685.5Mb	1,865,938 (72.3%) 15,784.9Mb
>Q12	2,194,311 (36.2%) 8,495.9Mb	1,706,471 (45.0%) 9,190.7Mb	1,170,738 (45.3%) 10,245.7Mb
>Q15	310,054 (5.1%) 755.8Mb	348,295 (9.2%) 1,526.5Mb	208,967 (8.1%) 796.1Mb

Genome assembly

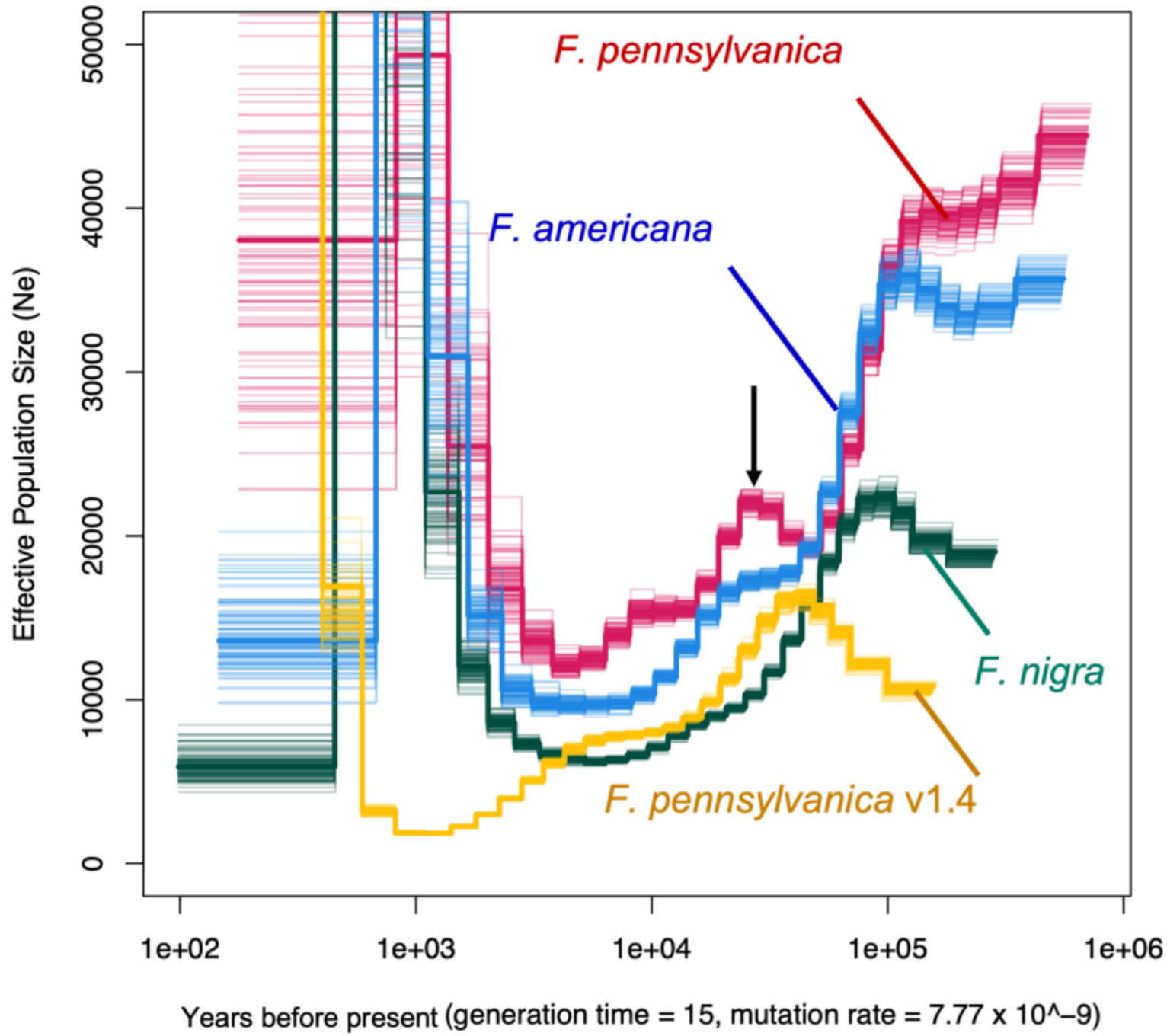
Flye[63] was utilized for de novo assemblies for the three *Fraxinus* genomes. Multiple Flye versions and assembly options were used to produce the most contiguous and sequence complete assemblies. Flye version 2.8.3 was used for *F. americana* and *F. nigra* and Flye version 2.9 was used for *F. pennsylvanica*. All genomes were assembled using a minimum overlap between reads of 10,000 bp and the scaffolding option. The *F. nigra* assembly benefitted most from three of Flye's internal polishing iterations, while those of *F. americana* and *F. pennsylvanica* benefited the most from one polishing iteration each. Mean genome size estimates were calculated using data from Whittemore et al.[24] Assembly stats were generated using Quality Assessment Tool for Genome Assemblies (QUAST) version 5.0.2[64], and assembly completeness was measured using Benchmarking Universal Single-Copy Orthologs (BUSCO) version 5.4.4 embryophyta_odb10[25] (Table S2).

Fig. 1: Flye outline.



The reads for each assembly were mapped onto their corresponding assembly using Minimap2 version 2.20

PSMC curves of closely related populations or species may differ in coalescence attributes because of inbreeding (i.e., heterozygosity) differences among accessions[51,79]. Initial PSMC output provides the capacity to estimate pi (π; nucleotide diversity), where $\pi = 4N_e\mu$. From this formulation, nucleotide diversity, effective population size, and mutation rate are linearly related. From initial PSMC, *F. americana*, *F. nigra*, and *F. pennsylvanica* had 3,458,091, 2,143,896, and 3,928,012 segregating sites, respectively


F. pennsylvanica
F. americana
F. nigra
F. pennsylvanica v1.4
Effective Population Size (Ne)
0
10000
20000
30000
40000
50000
1e+02
1e+03
1e+04
1e+05
1e+06
Years before present (generation time = 15, mutation rate = 7.77 x 10^–9)

Dúvidas?