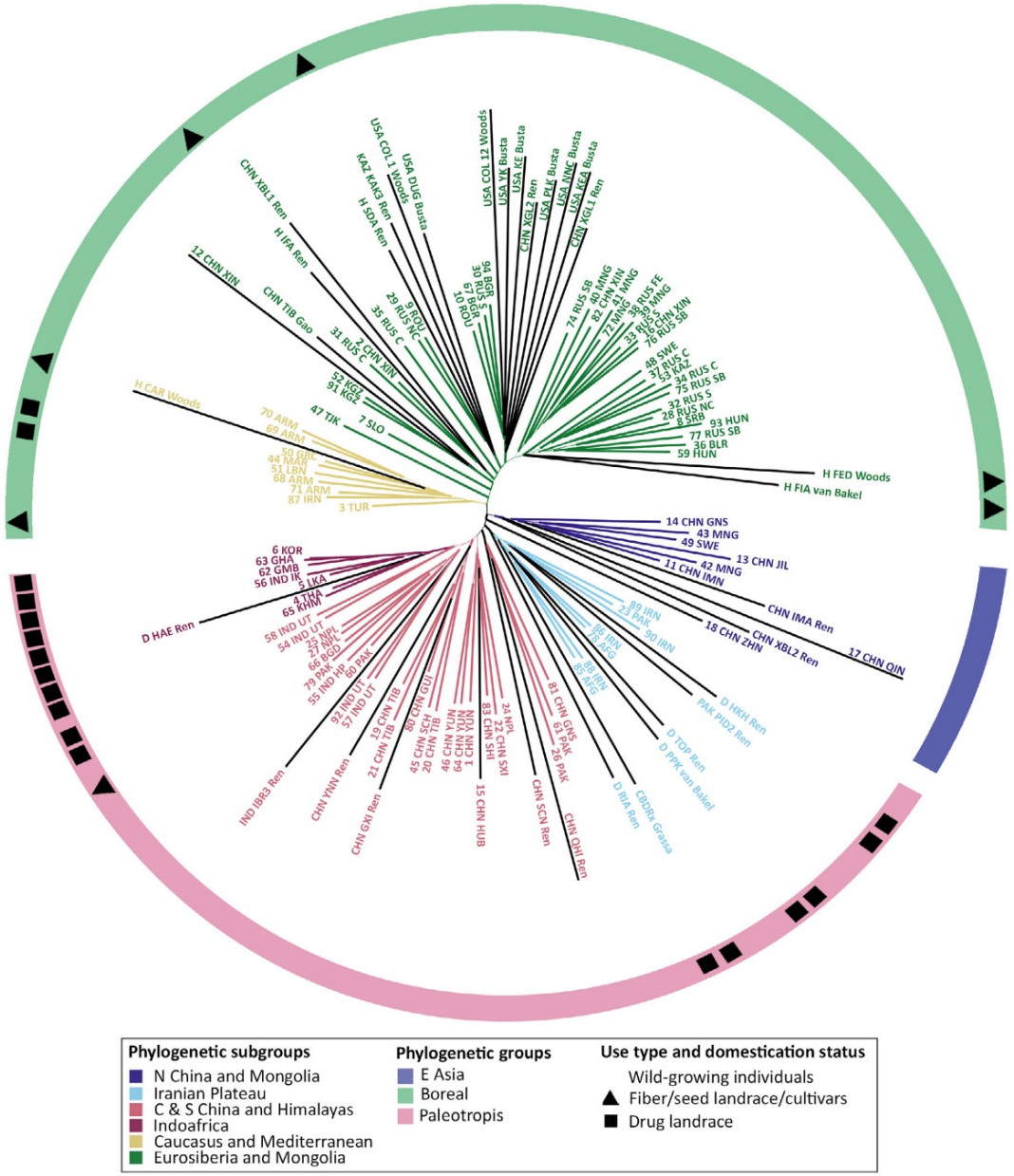


Manica Balant*, Daniel Vitales*, Zhiqiang Wang*, Zoltán Barina, Lin Fu, Tiangang Gao, Teresa Garnatje, Airy Gras, Muhammad Qasim Hayat, Marine Oganesian, Jaume Pellicer, Seyed A. Salami, Alexey P. Seregin, Nina Stepanyan-Gandilyan, Nusrat Sultana, Shagdar Tsooj, Magsar Urgamal, Joan Vallès, Robin van Velzen, Lisa Pokorny. 2025. "**Integrating target capture with whole genome sequencing of recent and natural history collections to explain the phylogeography of wild-growing and cultivated *Cannabis*"**. Plants People Planet. 1-18. https://doi.org/10.1002/ppp3.70043 [* = equal contributions]

Class focus area: Phylogeography

EEB603: Brian O'Meara

All quotes and images from the above paper unless otherwise noted

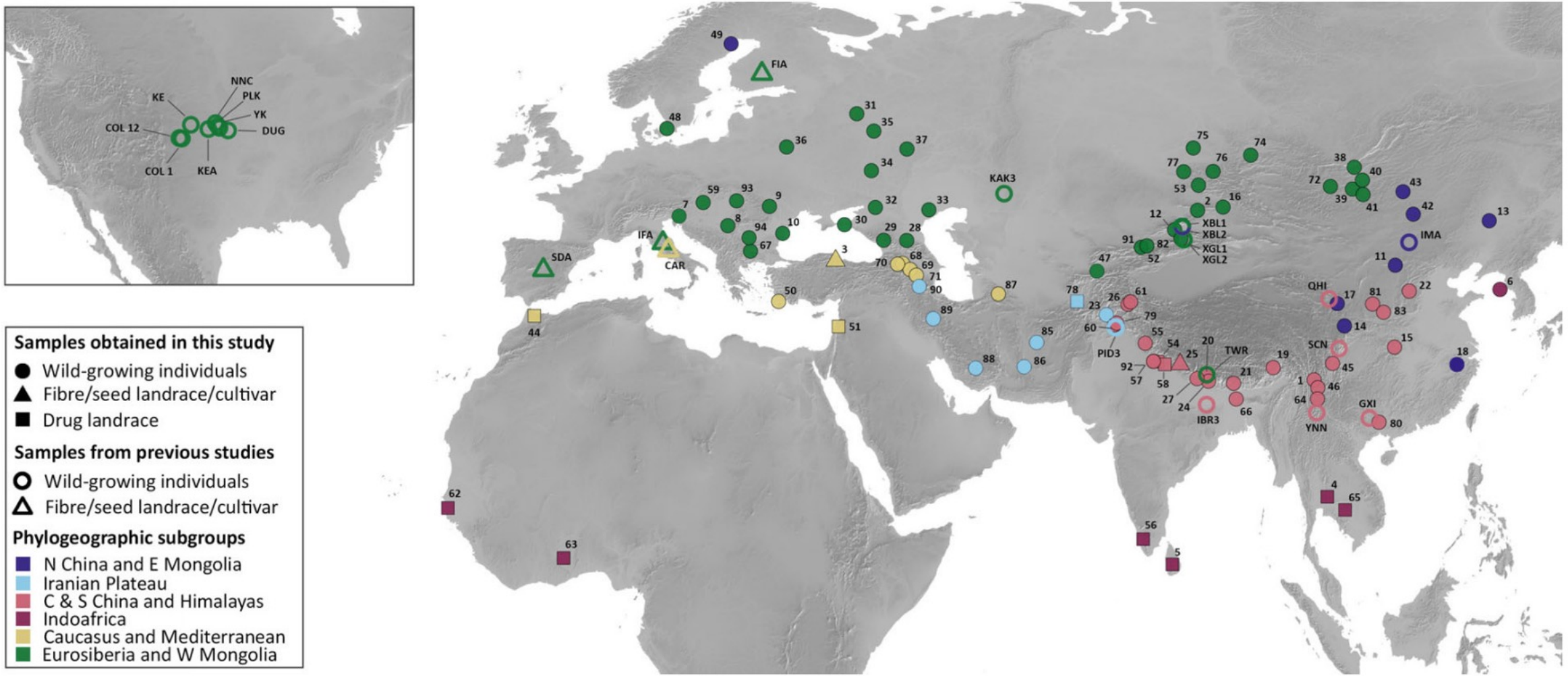
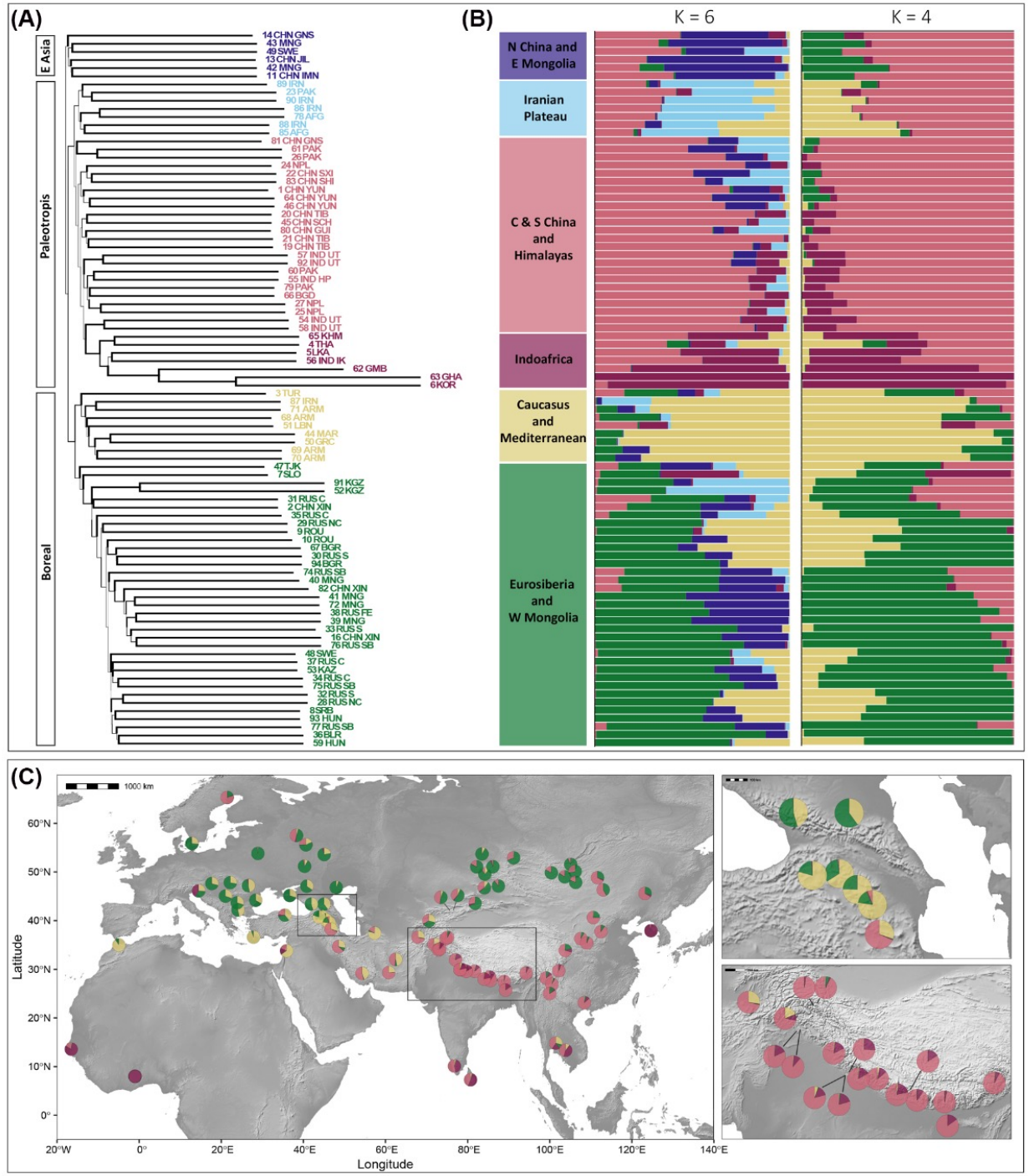
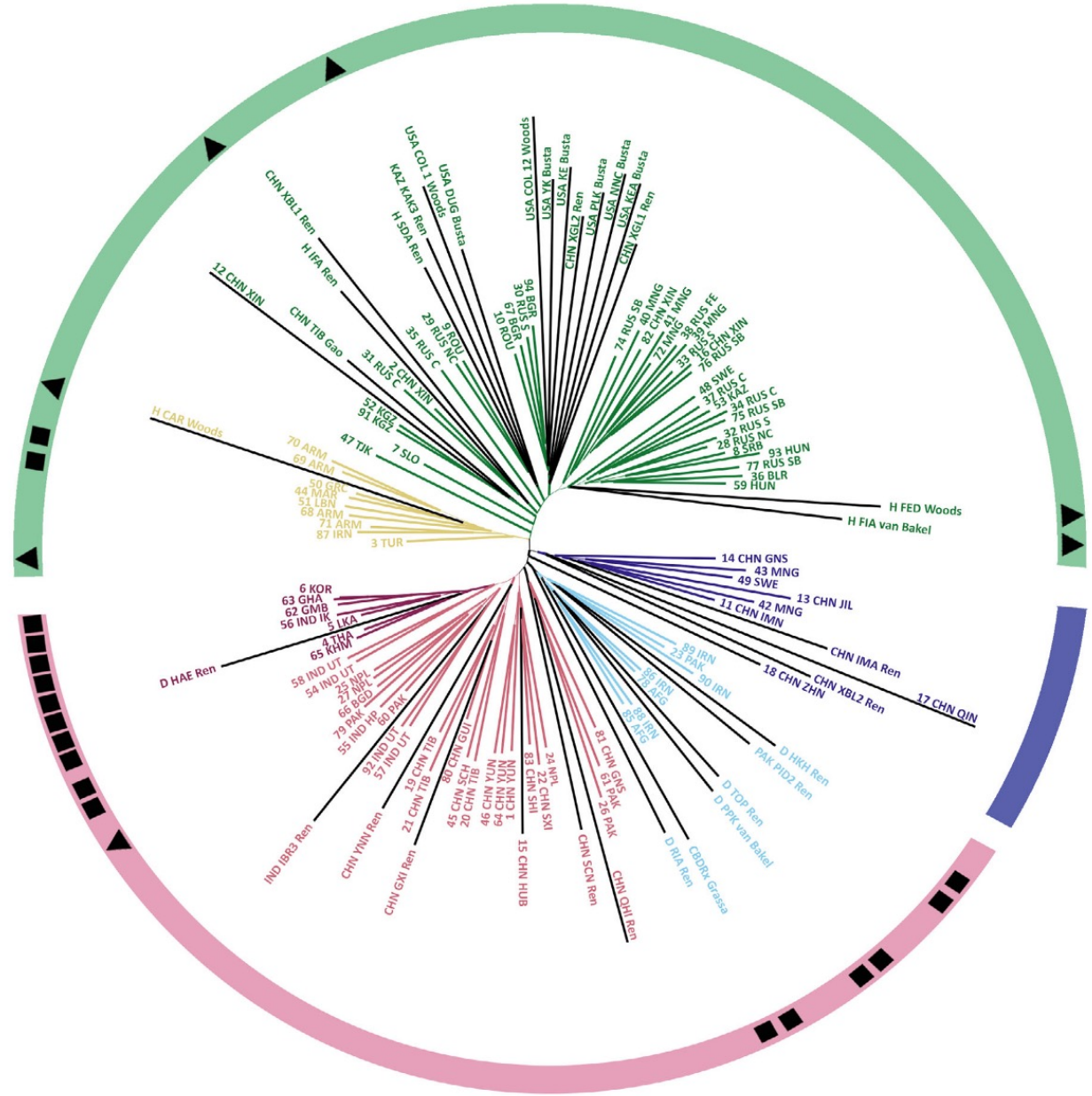


FIGURE 1 Geographic distribution of samples included in this study, with individuals coloured according to the subgroups obtained in the phylogenomic analysis (see Figure 2). The shapes indicate cannabis accession types, them being, wild-growing (circles), fibre/seed (triangles) and drug (squares) types. Additionally, filled shapes are newly analysed Hyb-Seq samples, while empty shapes are NCBI sequence read archive (SRA) corresponding to WGS data mined for our Hyb-Seq targets. The inset shows United States wild-growing populations mined from NCBI SRAs. Drug cultivars mined are not shown. For more detailed information, see Table S1. The map was made with *Natural Earth* (free vector and raster map data @ naturalearthdata.com).


(A)
E Asia
Paleotropis
Boreal
14 CHN GNS
43 MNG
49 SWE
13 CHN JIL
42 MNG
11 CHN IMN
89 IRN
23 PAK
90 IRN
86 IRN
78 AFG
88 IRN
85 AFG
81 CHN GNS
61 PAK
26 PAK
24 NPL
22 CHN SXI
83 CHN SHI
1 CHN YUN
64 CHN YUN
46 CHN YUN
20 CHN TIB
45 CHN SCH
80 CHN GUI
21 CHN TIB
19 CHN TIB
57 IND UT
92 IND UT
60 PAK
55 IND HP
79 PAK
66 BGD
27 NPL
25 NPL
54 IND UT
58 IND UT
65 KHM
4 THA
5 LKA
56 IND IK
62 GMB
63 GHA
6 KOR
3 TUR
87 IRN
71 ARM
68 ARM
51 LBN
44 MAR
50 GRC
69 ARM
70 ARM
47 TJK
7 SLO
91 KGZ
52 KGZ
31 RUS C
2 CHN XIN
35 RUS C
29 RUS NC
9 ROU
10 ROU
67 BGR
30 RUS S
94 BGR
74 RUS SB
40 MNG
82 CHN XIN
41 MNG
72 MNG
38 RUS FE
39 MNG
33 RUS S
16 CHN XIN
76 RUS SB
48 SWE
37 RUS C
53 KAZ
34 RUS C
75 RUS SB
32 RUS S
28 RUS NC
8 SRB
93 HUN
77 RUS SB
36 BLR
59 HUN
(B)
K = 6
K = 4
N China and E Mongolia
Iranian Plateau
C & S China and Himalayas
Indoafrica
Caucasus and Mediterranean
Eurosiberia and W Mongolia
(C)
1000 km
60°N
50°N
40°N
30°N
20°N
10°N
0°
Latitude
20°W
0°
20°E
40°E
60°E
80°E
100°E
120°E
140°E
Longitude


USA COL 12 Woods
USA YK Busta
USA KE Busta
CHN XGL2 Ren
USA PLK Busta
USA ANC Busta
USA KEA Busta
CHN XGL1 Ren
USA COL 1 Woods
USA DUG Busta
KAZ KAF3 Ren
H SDA Ren
CHN XBL1 Ren
H IFA Ren
12 CHN XIN
CHN TIB Gao
94 BGR
30 RUS S
67 BLR
10 BOU
29 RUS NC
9 BOU
35 RUS C
31 RUS C
2 CHN XIN
52 KGZ
91 KGZ
74 RUS SB
40 MNG
82 CHN XIN
41 MNG
72 MNG
38 RUS FE
5 MNG
33 RUS C
16 CHN XIN
76 RUS SB
48 SWE
37 RUS C
53 KAZ
34 RUS C
75 RUS SB
32 RUS S
28 RUS NC
8 SRB
93 HUN
77 RUS SB
36 BLR
59 HUN
H FED Woods
H FIA van Bakel
H CAR Woods
70 ARM
69 ARM
50 GRC
44 MAR
51 LBN
68 ARM
71 ARM
87 IRN
3 TUR
47 TJK
7 SLO
14 CHN GNS
43 MNG
49 SWE
13 CHN JIL
42 MNG
11 CHN IMN
CHN IMA Ren
18 CHN ZHN
CHN XBL2 Ren
17 CHN QIN
89 IRN
23 PAK
90 IRN
88 IRN
91 AFG
85 AFG
D HKH Ren
PAK PID2 Ren
D TOP Ren
D PPK van Bakel
CBDR+ Grassa
D RIA Ren
CHN SCN Ren
CHN QHI Ren
81 CHN GNS
61 PAK
26 PAK
24 NPL
22 CHN SXI
83 CHN SHI
15 CHN HUB
CHN HUN
64 CHN YUN
46 CHN YUN
1
45 CHN SCH
20 CHN TIB
80 CHN GUI
19 CHN TIB
21 CHN TIB
CHN GXI Ren
CHN YNN Ren
92 IND UT
57 IND UT
IND IBR3 Ren
6 KOR
63 GHA
62 GMB
56 IND IK
5 LKA
4 THA
65 KHM
D HAE Ren
58 IND UT
54 IND UT
25 NPL
27 NPL
66 BGD
79 PAK
55 IND HP
60 PAK
Phylogenetic subgroups
N China and Mongolia
Iranian Plateau
C & S China and Himalayas
Indoafrica
Caucasus and Mediterranean
Eurosiberia and Mongolia
Phylogenetic groups
E Asia
Boreal
Paleotropis
Use type and domestication status
Wild-growing individuals
Fiber/seed landrace/cultivars
Drug landrace

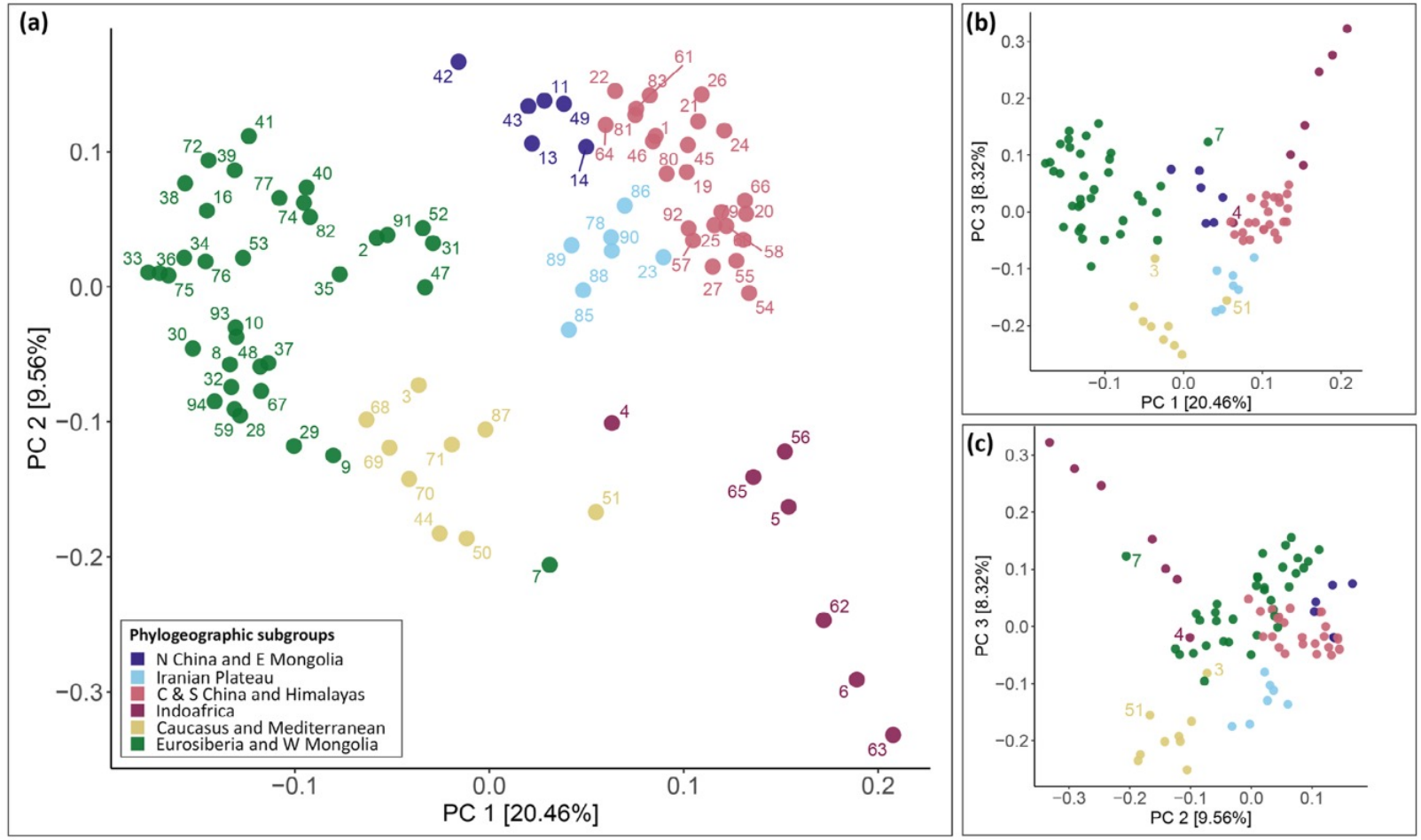


FIGURE 4 Principal component analysis (PCA; performed with PLINK) of *Cannabis* individuals for the 2875 (filtered and unlinked) single nucleotide polymorphisms (SNPs) called from the same 345 nuclear ortholog targets (comprising exons and their flanking regions) used to estimate the nuclear species tree (supercontig data matrix). Colours correspond to the six phylogeographic subgroups identified in the phylogenomic analysis (see Figure 2). (a) First and second PCA axes. (b) First and third PCA axes. (c) Second and third PCA axes.

TABLE 1 Pairwise fixation index (Hudson F_{ST}) values between phylogeographic subgroups.

Phylogeographic subgroup pairs		Hudson F_{ST}
Eurosiberia and W Mongolia	Indoafrica	0.155
Caucasus and Mediterranean	Indoafrica	0.136
N China and E Mongolia	Indoafrica	0.126
Indoafrica	Iranian plateau	0.120
Indoafrica	C and S China and Himalayas	0.090
Caucasus and Mediterranean	N China and E Mongolia	0.086
Eurosiberia and W Mongolia	C and S China and Himalayas	0.080
Eurosiberia and W Mongolia	Iranian plateau	0.077
Caucasus and Mediterranean	C and S China and Himalayas	0.076
Caucasus and Mediterranean	Eurosiberia and W Mongolia	0.061
N China and E Mongolia	Iranian plateau	0.060
N China and E Mongolia	Eurosiberia and W Mongolia	0.058
Caucasus and Mediterranean	Iranian plateau	0.048
Iranian plateau	C and S China and Himalayas	0.039
N China and E Mongolia	C and S China and Himalayas	0.036

Student questions

- They sent their DNA extraction samples to be sequenced by a private company. What are the risks and benefits of doing your own sequence work vs paying it for a company to do it for you?
- How important was human dispersal in the history?
- Potential issues with herbarium specimens?
- What's a paralog? SNP? Exon? Flanking regions? Linkage disequilibrium? F_{ST}?
- 68,212 SNPs -> 2,875 after filtering. Good idea? Bad?
- In addition, the STRUCTURE model assumes that markers are neutral, which--if they are located within genes--one can reasonably expect that they generally are not. The paper gave some phenomenal background information, but I'm not particularly comfortable with how they analyzed and interpreted their data.
- How might degraded herbarium samples affect results?
- RAXML-NG vs RAXML
- How would you analyze what changed with domestication?
- Why no time calibration?