Supporting Information – Statistical phylogeography from individual, *de novo* genome assemblies

Jack Hearn[1], Graham N. Stone[1], Lynsey Bunnefeld-McInnes[1], James A. Nicholls[1], Nick H. Barton[2] and Ko

[1]Institute of Evolutionary Biology

University of Edinburgh

Kings Buildings

Edinburgh EH9 3JT, UK

[2]Institute of Science and Technology

Am Campus 1

A-3400 Klosterneuburg

Austria

Table S1: Sampling and rearing information of individuals used for genome sequencing.

Species	code	sex	country	locality	Lat/Long	col. date	oak host
Belizinella gibbera	Bgib 15	f	Russia	Khasan Lake, Primorsky Krai	42.45 N, 130.65 E	26/09/08	*Q. dentata*
	Bgib 18	f	Russia	Khasan Lake, Primorsky Krai	42.45 N, 130.65 E	26/09/08	*Q. dentata*
Biorhiza pallida	Wa (Bpal 1398)	m	Spain	Embalse de Garcia, Extramadura	39.17 N, 5.22 W	12/04/05	*Q. faginea*
	Wb (Bpal 2)	m	Spain	Mairena, Granada	37.37 N, 5.75 W	06/05/09	*Q. faginea*
	Ca (Bio 4)	m	Hungary	Szokolya	47.87 N, 19.02 E	15/05/98	*Q. petraea/robur*
	Cb (1613)	m	Croatia	Ze Medvedgrad, Zagreb	45.86 N, 15.94 E	16/05/11	*Q. petraea*
	E (1560)	m	Iran	Bane, Kordestan	35.99 N, 45.90 E	01/04/11	*Q. infectoria*
	UK, pool	f	UK	Silwood Park	51.41 N, 0.64 W	01/04/11	*Q. robur*

Table S2: Raw short-read data generated for each individual. For *B. pallida* and *B. gibbera* all reads regardless of type (PE= paired-end, SE=single end), read length or individual were combined to generate respective in-and out-group meta-assemblies.

Individual	Type	Read length	Raw reads	Q-trimmed pairs	Single reads
Wa	PE	100	57,929,835	54,246,080	3,614,584
Wb	PE	100	21,314,072	20,072,304	1,050,353
Wb	PE	50	11,964,815	11,110,836	577,455
Ca	PE	100	20,394,203	18,792,799	1,666,698
Ca	PE	50	51,190,824	47,239,071	2,982,296
Ca	SE	80	20,298,581	N/A	18,683,882
Cb	PE	100	43,171,168	41,505,059	1,627,413
E	PE	100	41,005,941	39,277,818	1,690,764
UK, pool	PE	50	N/A	227,859,76	57
Bgib A	PE	100	36,377,936	34,386,347	1,949,865
Bgib B	PE	100	22,848,965	21,548,722	1,102,887
Bgib B	PE	50	47,084,801	83,829,015	8,786,606
Bgib B	SE	80	23,753,570	N/A	23,230,193

Table S3: Proportion of individual reads re-aligned to the *B. pallida* meta-assembly.

Individual	% Reads mapped	% Pairs mapped	% Pairs properly mapped
Wa	98.1	97.0	38.5
Wb	97.6	96.1	60.1
Ca	97.4	95.9	58.8
Cb	98.0	96.7	51.9
E	97.9	96.7	52.5

Table S4: The total number of blocks observed with each topologically resolved mutational configuration (in the 1kb *WaCaE* data). For simplicity, mutational configurations are defined here only in terms of number of mutations on the internal branch (1–3, left to right) and the number of mutations on the two shorter external branches (0–3, top to bottom) (so ignoring the longer external branch). The theoretical expectations given the best-fitting model (see Table 3) (with mutational heterogeneity) are given in brackets. Note that most blocks are topologically unresolved (74.9% observed, 72.5 % expected). For this class (last column), we give the number of blocks containing a particular total number of mutations (S).

	$(W,(E,C))$			$(C,(E,W))$			$(E,(C,W))$			unresolved
	1	2	3	1	2	3	1	2	3	S
0	69 (72)	21 (25)	5 (7.7)	43 (52)	3 (12)	4 (3)	23 (30)	5 (7.3)	1 (1.8)	292 (253)
1	83 (79)	19 (28)	9 (8.6)	50 (47)	8 (12)	4 (2.9)	32 (31)	6 (7.6)	1 (1.9)	405 (411)
2	38 (47)	17 (17)	7 (5.1)	23 (24)	5 (5.9)	1 (1.5)	18 (17)	0 (4.3)	0 (1.1)	353 (365)
3	13 (21)	9 (7.2)	1 (2.2)	16 (8.7)	0 (2.2)	0 (0.6)	10 (6.8)	1 (1.7)	1 (0.4)	252 (237)
Total	291 (319.2)			160 (171.7)			98 (111.2)			
Proportion	0.157 (0.171)			0.086 (0.09)			0.053 (0.06)			

Table S5: Support (ΔlnL) relative to the best model for alternative histories of refugial populations of *B. pallida* estimated from the *b* dataset (Model B in Fig. 2 has highest support and is shown in bold). The labelling of populations (1–3) and of models (A–F) corresponds to that in Fig. 2; all scenarios involving unidirectional admixture were assessed for each of the three possible orders of population divergence (columns 1–3). Models of strict divergence without admixture between two (2 pop., i.e. $T_1 = 0$) or three (3 pop.) populations were fitted assuming either a single or two different N_e for ancestral populations. Parameters for which the maximum likelihood estimate is 0 (i.e. the model reduces to a simpler nested model) are indicated in brackets (f^* refers to complete admixture, i.e. $f = 1$).

Model	k			
Panmixia	1	-484.8		
Polytomy	2	-60.3		
Topology		$(W_1,(C_2,E_3))$	$(C_1,(E_2,W_3))$	$(E_1,(C_2,W_3))$
A), $2 \rightarrow 1$	5	-14.9, (T_1)	-21.1	-33.2, (f^*)
B), $3 \rightarrow 1$	5	**0**	-59.9, (T_1)	-59.4, (T_2, T_{gf})
C), $2/3 \rightarrow 1$	5	-14.3	-59.9	-60.3, (T_{gf}, f^*)
D), $1 \rightarrow 2$	5	-18.0	-19.4, (T_1)	-19.4, (T_1)
E), $1 \rightarrow 3$	5	-18.0	-60.0, (f)	-60.0, (f^*)
F), $1 \rightarrow 2/3$	5	-33.2, (f^*)	-49.7	-14.4, (T_{gf})
2 pop.	2	-265.3	-293.6	-386.7
3 pop.	2	-33.2	-60.0	-60.3,(T_2)
2 pop. N_e	3	-46.1	-60.0	-64.7
3 pop. N_e	4	-31.0,	-60.0	-60.3, (T_2)

Table S6: Fisher information for 1kb *WaCaE* dataset for model B (W, (C,E))

Parameter	T_{gf}	T_1	T_2	f
I	836.4	1198.5	45.1	3596.4
E[I]	880.0	1334.1	49.2	3762.5
E[SD]	0.0337	0.0274	0.143	0.0163
ML estimate	1.04	1.21	3.34	0.76

Given the parameters of the best supported model estimated for the 1kb *WaCaE* dataset (bottom row), the observed I, $E[I]$ and $E[SD]$ are shown based on 2231 loci. θ = 0.69.

Table S7: Support (ΔlnL relative to the best model) for alternative divergence scenarios for three refugial populations of *B. pallida* without admixture or with unidirectional admixture (A–F) for alternative block lengths (500b and 2kb). All possible scenarios (the labelling of populations (1–3) and of models (A–F) corresponds to Fig. 2) were assessed for the three possible orders of population divergence (columns 1–3). Parameters for which the maximum likelihood estimate is 0 (i.e. the model reduces to a simpler nested model) are indicated in brackets (f^* refers to complete admixture, i.e. $f = 1$). The model with highest support is shown in bold.

Model	Admixture	k	500b			2kb		
Panmixia	No	1	-263.2			-948.7		
Polytomy	No	2	-59.3			-111.6		
Topology			$(W_1,(E_2,C_3))$	$(C_1,(E_2,W_3))$	$(E_1,(C_2,W_3))$	$(W_1,(E_2,C_3))$	$(C_1,(E_2,W_3))$	$(E_1,(C_2,W_3))$
2 pop.	No	3	-109.6	-197.3	-239.8	-446.4	-601.6	-692.2
3 pop.	No	4	-19.2	-59.3, (T_2)	-59.3, (T_2)	-46.2	-111.6, (T_2)	-111.6, (T_2)
A)	$2 \rightarrow 1$	5	-12.8, (T_1)	-5.5	-19.2, (f^*)	-16.6, (T_1)	-46.2 (f^*)	-28.3
B)	$3 \rightarrow 1$	5	**0**	-59.3, (T_1, f^*)	-59.3, (T_2, f^*)	**0**	-111.6, (T_2,f)	-111.6, (T_2 f^*)
C)	$2/3 \rightarrow 1$	5	-12.5	-59.3, (T_{gf}, f^*)	-59.3, (T_{gf}, T_2)	N/A	-111.6, (T_{gf}, f^*)	-111.6, (T_{gf}, f^*)
D)	$1 \rightarrow 2$	5	-19.2, (f)	-8.3, (T_1)	-19.2, (f^*)	-30.1, (T_1)	-46.2, (f^*)	-30.1
E)	$1 \rightarrow 3$	5	-8.2	-59.3, (T_1, T_2)	-59.3, (T_2, f)	-46.2, (f)	-111.6, (T_1, T_2)	-111.9, (T_2 f)
F)	$1 \rightarrow 2/3$	5	-19.2, (f^*)	-57.5	-13.4, (T_{gf})	-45.8,	-40.4, (T_{gf})	-100.4

Table S8: Maximum likelihood parameter estimates under the best supported model (see Table 2) for the *WaCaE* alignment and three different block lengths: 500b, 1kb (as in Table 1) and 2kb. Both effective population size and divergence time parameters are scaled relative to the rate of coalescence, i.e. in $2N_e$ generations. Absolute values calibrated using the direct, genome-wide *Drosophila* mutation rate of Keightley et al. (2009) and assuming two generations per year are given in brackets.

dataset	μ het.	lnL	f	θ (N_e)	T_{GF} (t_{GF})	T_1 (t_1)	T_2 (t_2)
WaCaE, 500b	no	-6560.4	0.67	0.39 (59,200)	0.65 (38KY)	0.93 (54KY)	2.44 (144KY)
WaCaE, 1kb	no	-9269.3	0.76	0.69 (52,000)	1.04 (54KY)	1.21 (63KY)	3.34 (173KY)
WaCaE, 2kb	no	-10713.2	0.69	1.34 (52,900)	1.04 (53KY)	1.23 (62KY)	2.73 (138KY)

Figure S1: Distribution of read coverage in the *de novo* meta-assemblies for *B. pallida* (left) and *B. gibbera* (right). The red dashed lines indicate mean coverage and show that modal coverage is a better summary of the distribution given the long tail. The tails of the distributions stop at the thresholds chosen (75 fold for *B. pallida* and 30 fold *B. gibbera*) as they were likely to indicate remaining unfiltered collapsed repeats whose sequences had been amalgamated during assembly.

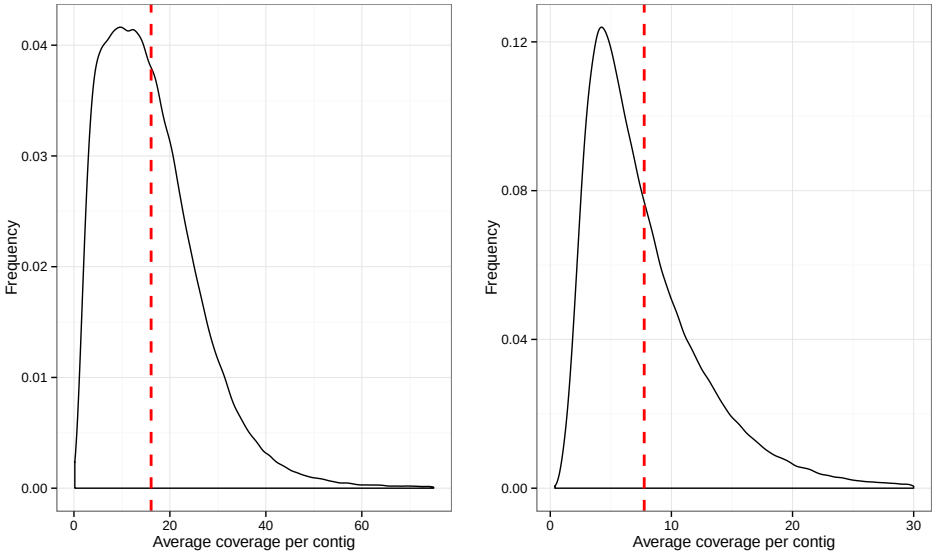

Figure S2: SNP frequencies before (Q0) and after (Q20) quality filtering

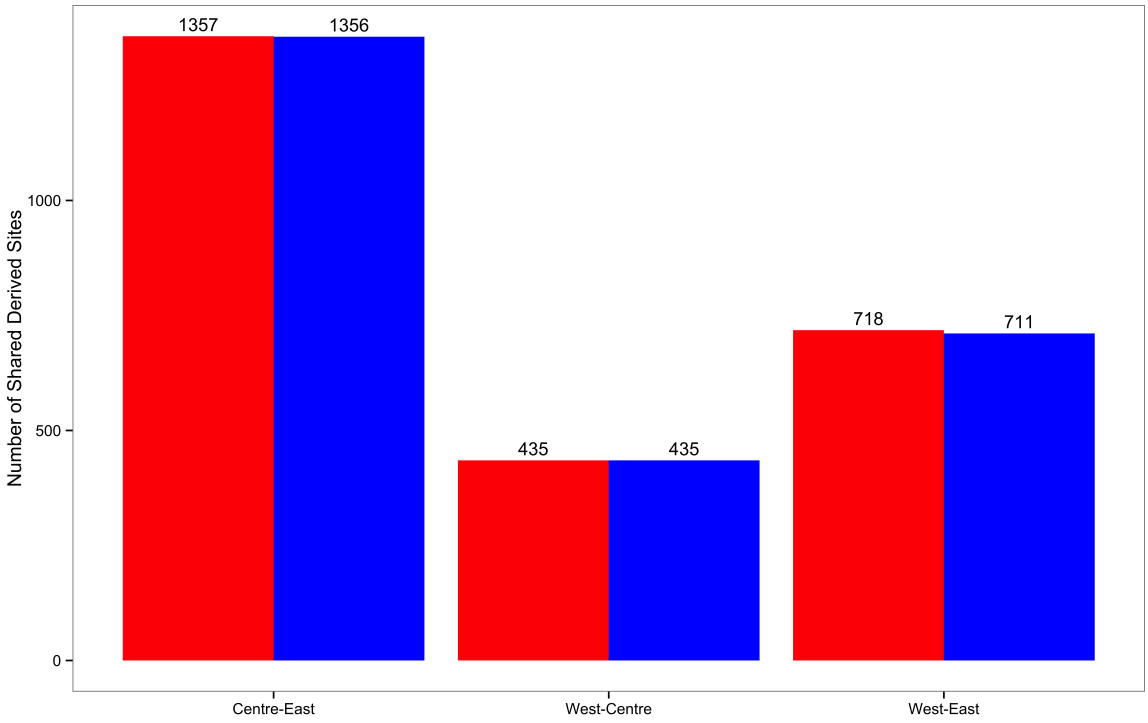

Figure S3: Expected information in parameters as a function of block size (θ). In each plot, the horizontal dashed line is the expected information in a single SNP for each parameter (other parameters held at their maximum likelihood estimate from the 1kb *WaCaE* dataset). The vertical dashed line is the value of θ that gives, on average, one SNP per block.

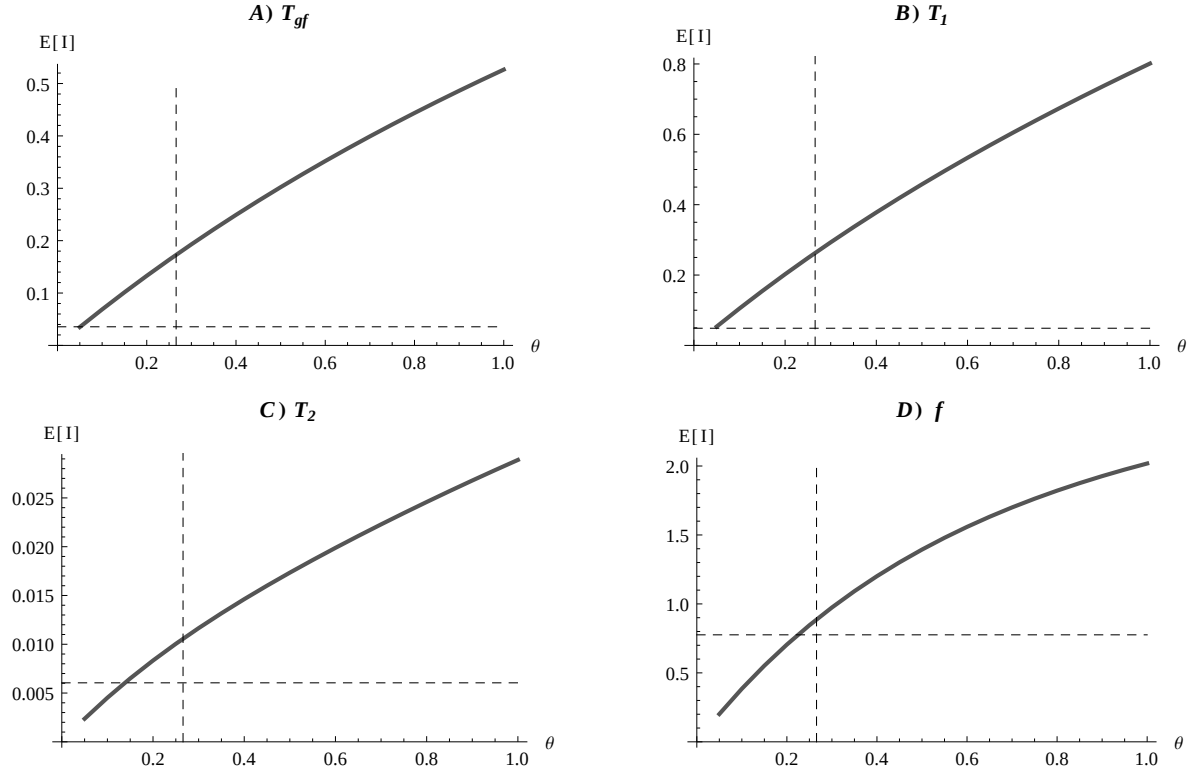

Figure S4: Correlation between contig length and mean per site divergence between *B. pallida* and *B. gibbera*.

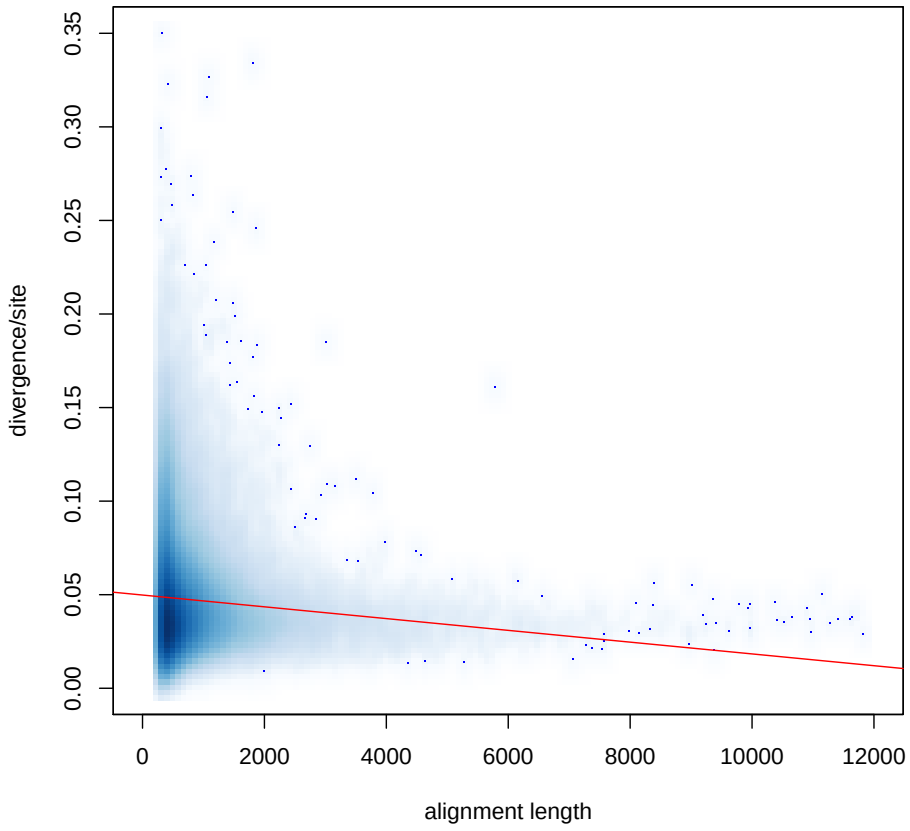

Figure S5: There is a negative correlation (Spearman Rank $\rho = -0.51$, $p < 0.001$) between the number of divergent sites (between *B. pallida* and *B. gibbera*) and the % of coding sequence (cds) in 2-kb blocks as determined by BLAST against the *B. pallida* transcriptome.

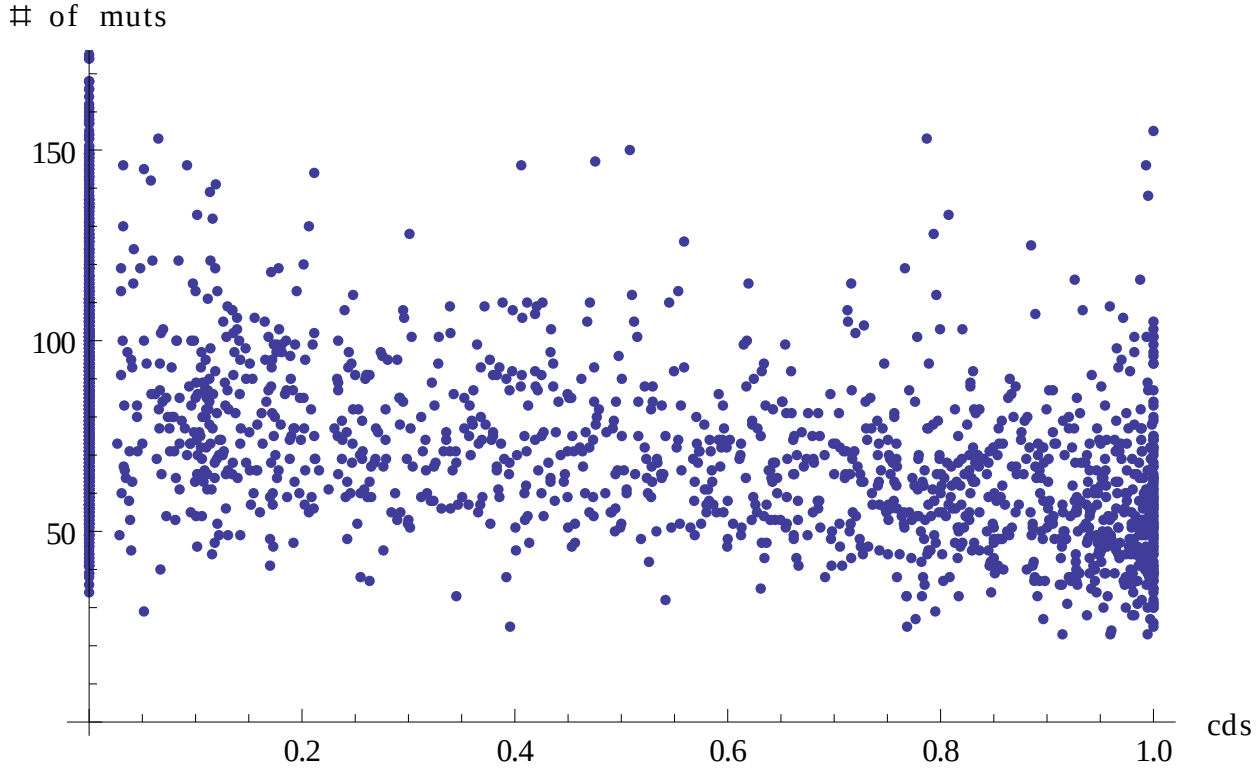

Figure S6: The effect of undetected recombination on model choice and parameter estimates. (A) ΔlnL from the best supported model (always model B, (W,(C,E)) from Fig. 2) to the next best supported models. Thick dashes - model A/C, thin dashes - model E. (B-D) Maximum likelihood parameter estimates for θ, divergence and admixture times (thick dashes - t_{gf}, thin dashes - t_1, solid - t_2) and the admixture proportion, *f*. In all plots, the horizontal dotted lines correspond to the ML parameters estimated from the 1kb *WaCaE* dataset. The vertical dotted line at 4.5 is a reasonable r/μ ratio for *Biorhiza pallida* (see text for more details)