MOLECULAR ECOLOGY RESOURCES

Supplemental Information for:

Inference of the distribution of fitness effects of mutations is affected by SNP filtering methods, sample size and population structure

Bea Angelica Andersson[1], Wei Zhao[1], Benjamin C. Haller[2], Åke Brännström[3,4,5], Xiao-Ru Wang[1]

[1]Department of Ecology and Environmental Sciences, Umeå University, Sweden

[2]Department of Computational Biology, Cornell University, USA

[3]Department of Mathematics and Mathematical Statistics, Umeå University, Sweden

[4]Advancing Systems Analysis Program, International Institute for Applied Systems Analysis, Laxenburg, Austria

[5]Complexity Science and Evolution Unit, Okinawa Institute of Science and Technology Graduate University, Japan

Table of Contents:

Table S1. Sequence information of the *Arabidopsis lyrata* samples included in this study. Source codes are NCBI accession IDs.

Population	Source	Coverage (Mb, >=5x)	Mean depth (>=5x)	Median depth (>=5x)	No. reads
Norway	ERR3397904	147.26	25.71	16	63201441
Norway	ERR3397905	144.61	21.83	14	52859026
Norway	ERR3397906	143.18	22.77	13	54633798
Norway	ERR3397907	143.9	22.27	14	53839619
Norway	ERR3397908	146.44	23.74	15	59507152
Norway	ERR3397909	146.45	23.69	15	59465532
Norway	ERR3397910	145.45	24.72	14	62069317
Norway	ERR3397911	148	26.05	15	66516823
Norway	ERR3397912	145.1	23.77	14	59254019
Norway	ERR3397913	143.16	21.8	13	53951341
Norway	SRR5124977	151.64	70	38	127927590
Norway	SRR5124983	144.92	33.26	22	59142085
Norway	SRR5124985	135.63	22.97	13	34814708
Norway	SRR5124997	153.66	55.35	35	102114289
Norway	SRR5124998	149.93	49.79	30	88534336
Norway	SRR5124999	139.45	27.54	16	42830997
Austria	ERR3514864	130.18	17.02	11	20727733
Austria	ERR3514865	141.4	23.75	15	31042548
Austria	ERR3514866	140.31	21.29	15	27625035
Austria	ERR3514869	156.87	49.12	37	72515100
Austria	ERR3514870	145.82	25.47	18	34493579
Austria	ERR3514871	147.85	26.79	19	36994924
Austria	ERR3514872	130.12	17.95	11	22086024
Austria	ERR3514873	136.53	19.89	13	25584056
Austria	ERR3514874	141.42	26.26	16	34646984
Austria	ERR3514875	128.41	16.94	10	20457660
Austria	ERR3514876	148.58	29.17	22	39961518
Austria	ERR3514877	130.4	16.15	11	19708718
Austria	ERR3514878	155.98	42.5	33	62109925
Austria	ERR3514879	141.14	21.85	14	29584688
Austria	ERR3514880	153.3	38.58	27	54727972
Austria	ERR3514883	130	19.1	11	23003063
Austria	ERR3514884	130.2	17.75	11	21314626
Austria	ERR3514885	141.2	27.63	16	37910237
Austria	ERR3514886	124.05	15.02	10	17565928
Austria	ERR3514887	144.2	26.05	17	35825064
Austria	ERR3514888	142.67	26.2	17	34966847
Austria	ERR3514889	141.14	24.65	16	32554512
Austria	ERR3514892	138.26	19.26	14	24924196
Austria	ERR3514893	142.71	23.43	16	30871081
Austria	ERR3514895	148.06	30.95	20	42713949
Austria	ERR3514896	149.81	31.97	23	44511357
Austria	ERR3514897	135.69	22.12	13	27521061
Austria	ERR3514898	145.12	32.64	21	44216326
Austria	ERR3514899	144.58	29.89	19	39745319

Figure S1. Principal component analysis in the 45 individuals of *Arabidopsis lyrata* sampled from Austria and Norway, based on 3,921,575 SNPs.

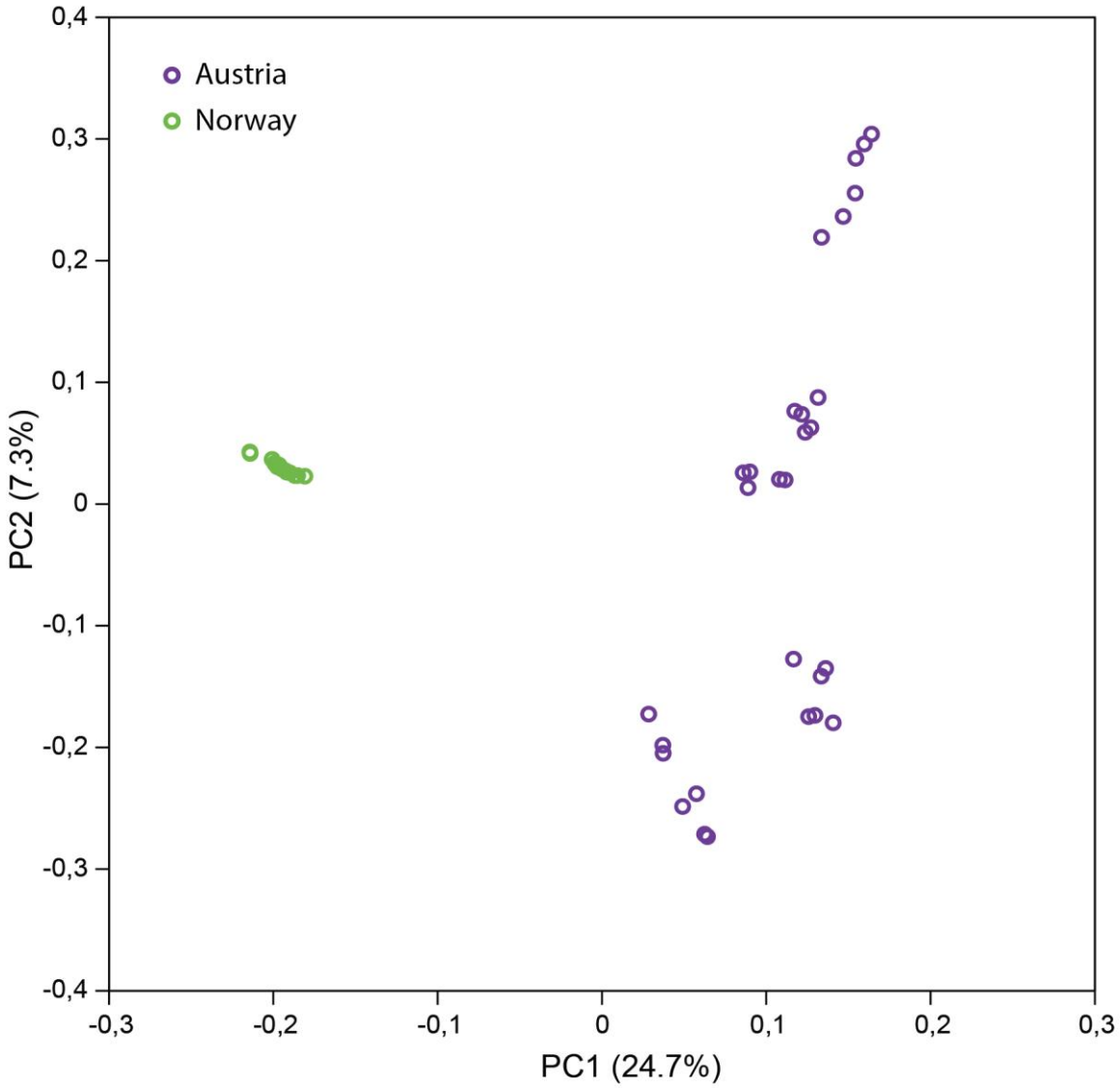

Table S2: Number of sites and SNPs in different simulated datasets, together with their respective estimated mean (*Es*) and shape (β) parameters of the DFE. Sites in VCF include all sites regardless of degeneracy, while sites in SFS include only zerofold and fourfold degenerate sites. Earth Mover's Distance (EMD) signifies the estimation error of each estimated gamma distribution of DFE compared to the DFE used in the simulation (shape β: 0.1, mean *Es*: -100). Lower EMD values indicate a closer fit to the known DFE.

Filtering method	Individuals	Sites in VCF after filtering	Sites in SFS (min – max)	SNPs in SFS (min – max)	β (Median [min – max])	*Es* (Median [min – max])	EMD (10^{-7}) (Median	Mean	[min – max])
No missing dta (in 10 replicates)	4	50,000,000	12,162,648 – 12,162,869	5,156 – 5,448	0.0902 [0.0500 – 0.1675]	-682 [-6.36×10^6 – -4]	49.7	38.5	[6.7 – 56.5]
	8	50,000,000		6,840 – 7,205	0.1064 [0.0500 – 0.1318]	-127 [-5.19×10^6 – -21]	13.9	20.8	[5.8 – 56.5]
	12	50,000,000		8,002 – 8,342	0.1077 [0.0865 – 0.1326]	-103 [-1,044 – -18]	7.9	13.3	[2.5 – 30.5]
	16	50,000,000		8,832 – 9,182	0.1043 [0.0900 – 0.1308]	-148 [-714 – -22]	12.9	13.3	[3.2 – 26.5]
	20	50,000,000		9,420 – 9,859	0.1044 [0.0843 – 0.1220]	-134 [-1,293 – -37]	12.3	13.0	[3.7 – 32.5]
	24	50,000,000		10,019 – 10,336	0.1051 [0.0792 – 0.1291]	-126 [-2,773 – -24]	9.8	13.2	[2.6 – 40.5]
	50	50,000,000		12,617 – 12,818	0.1096 [0.0874 – 0.1235]	-94 [-821 – -36]	5.3	8.9	[3.4 – 27.5]
	100	50,000,000		15,357 – 15,702	0.1138 [0.0972 – 0.1280]	-79 [-276 – -33]	7.5	8.9	[3.5 – 21.5]
Downsampling (in 4 replicates)	4	28,420,524	6,911,406	2,959 – 3,099	0.1253 [0.0500 – 0.1411]	-39 [-1.23×10^7 – -11]	34.5	33,7	[9.4 – 57.5]
	8	28,420,524		4,010 – 4,075	0.0724 [0.0500 – 0.1001]	-8,392 [-3.15×10^6 – -147]	45.9	38,0	[4.8 – 55.5]
	12	28,420,524		4,621 – 4,694	0.0999 [0.0919 – 0.1162]	-198 [-420 – -52]	13.0	11,7	[2.2 – 19.5]
	16	28,420,524		5,085 – 5,122	0.0945 [0.0777 – 0.1082]	-336 [-3,711 – -80]	14.4	19,0	[4.9 – 42.5]
	20	28,420,524		5,446 – 5,539	0.0907 [0.0727 – 0.1096]	-456 [-6,829 – -78]	19.9	22,9	[5.5 – 46.5]
	24	28,420,524		5,795 – 5,827	0.0981 [0.0913 – 0.1030]	-229 [-417 – -136]	9.8	10,5	[3.7 – 19.5]
	50	28,420,524		7,165 – 7,216	0.0992 [0.0929 – 0.1088]	-248 [-391 – -88]	10.6	10,5	[3.0 – 18.5]
	85 (n=1)	28,420,524		8,323	0.1045	-139	4.0	4,0	
Imputation (in 4 replicates)	4	28,420,524	6,911,406	2,907 – 3,136	0.1140 [0.0500 – 0.1498]	-137 [-5.64×10^6 – -7]	38.3	36,4	[12.9 – 56.5]
	8	28,420,524		3,892 – 3,976	0.0998 [0.0642 – 0.1171]	-293 [-54,893 – -32]	19.6	24,2	[5.2 – 52.5]
	12	28,420,524		4,426 – 4,554	0.0848 [0.0815 – 0.1079]	-1,201 [-1,762 – -84]	30.0	24,9	[4.1 – 35.5]
	16	28,420,524		4,840 – 5,017	0.0866 [0.0794 – 0.1111]	-811 [-1,953 – -63]	26.8	24,6	[8.3 – 36.5]
	20	28,420,524		5,179 – 5,384	0.0894 [0.0760 – 0.0981]	-819 [-3,850 – -226]	22.2	24,4	[10.5 – 43.5]
	24	28,420,524		5,592 – 5,750	0.0793 [0.0729 – 0.1155]	-2,976 [-8,094 – -50]	36.5	33,2	[12.3 – 47.5]
	50	28,420,524		6,931 – 6,958	0.0948 [0.0814 – 0.1132]	-334 [-1,642 – -65]	15.4	18,5	[8.5 – 35.5]
	100 (n=1)	28,420,524		8,396	0.0986	-247	11.6	11,6	
Subsampling (in 4 replicates)	4	26,099,661	6,347,733	2,707 – 2,870	0.1279 [0.0733 – 0.1518]	-23 [-11,804 – -6]	37.2	36,5	[22.4 – 49.5]
	8	13,629,024	3,313,984	1,897 – 2,003	0.0702 [0.0500 – 0.0737]	-11,337 [-8.96×10^6 – -2,657]	48.7	48,3	[39.4 – 56.5]
	12	7,109,386	1,730,591	1,107 – 1,194	0.1040 [0.0696 – 0.1366]	-2,386 [-10,280 – -18]	37.0	37,5	[27.7 – 48.5]
	16	3,713,156	902,526	624 – 661	0.1109 [0.0782 – 0.1618]	-59 [-1,223 – -3]	26.3	27,6	[1.7 – 56.5]
	20	1,938,476	470,900	348 – 402	0.0907 [0.0500 – 0.1938]	-35,852 [-795,504 – -2]	53.1	47,1	[18.3 – 64.5]
	24	1,011,970	246,323	174 – 199	0.0540 [0.0500 – 0.5804]	-362,712 [-3.23×10^6 – 0]	55.1	63,6	[31.2 – 113.5]
	50	14,863	3,737	3 – 5	0.5172 [0.0500 – 0.5227]	0 [-2.16×10^{12} – 0]	113.2	102,3	[64.2 – 118.5]