Supporting Information

Casteleyn et al. 10.1073/pnas.1001380107

DNA Extraction and Amplification of Microsatellite Loci. Cells from a 50-mL growing batch culture were harvested by centrifugation and stored at −80 °C. Genomic DNA was extracted from pelleted cells using a DNeasy Plant Mini Kit (Qiagen) following the manufacturer's protocol. Six nuclear microsatellite loci (PP1, PP2, PP3, PP4, PP5, and PP6) were amplified in separate PCR reactions with fluorescently labeled primers (6FAM, VIC, NED; Applied Biosystems) following PCR conditions as described in Evans and Hayes (1) and modified as outlined in Casteleyn et al. (2). PCR products were analyzed on an ABI 3130xl genetic analyzer (Applied Biosystems) according to Casteleyn et al. (2). To reduce genotyping errors, positive and negative controls were used and scoring of alleles was performed by the same person. Genotypes were scored using Genemapper v4.0 software (Applied Biosystems).

Analysis of Genetic Diversity: Methods. Seven predefined populations were considered, based on the geographical area sampled: Belgium, Denmark, Ireland, Atlantic Canada, Pacific United States, Japan, and New Zealand. Previous studies (2, 3) showed that temporal variation does not exceed spatial variation for the *P. pungens* clade I. Therefore populations sampled in different seasons or years from the same geographic area were pooled. Possible scoring errors due to stuttering or large allele dropout were assessed using the program MICROCHECKER v2.2.3 (4). CONVERT v1.31 (5) was used to format the MLG data and to identify private alleles for each population. To identify matching MLGs, the Excel add-in MS TOOLS v3 (6) was used. Identical MLGs from the same sampling location and sampling date were removed from the dataset because isolations were not always done within 24 h after sampling. This is a potential problem because asexual (i.e., clonal) reproduction may take place between the time of field sampling and the moment of isolating the diatom cells, leading to a potential underestimation of diversity. Therefore, we did not estimate clonal diversities (number of unique genotypes relative to sample size) per population. GIMLET v1.3.3 (7) was used to calculate the probability of identity, $P_{(ID)}$, and the probability of identity for siblings, $P_{(ID)sib}$, for each locus and over all loci. Waits et al. (8) suggested that $P_{(ID)}$ or $P_{(ID)sib}$ values between 0.01 and 0.0001 indicate that the markers possess sufficient resolving power to distinguish unrelated individuals. Numbers of alleles, allele frequencies, number of genotypes, observed heterozygosity (H_o), and expected heterozygosity (H_e), were calculated per population for each locus using GENEPOP v4.0 (9). To compare the number of alleles between populations with different sample sizes, the software ARES v1.2.2 (10) was used. Unlike other tools based on rarefaction, ARES extrapolates beyond the sample size. Departures from HWE at each locus in every population were tested using an exact test (11) with a Markov chain method to estimate exact P values (dememorisation number: 10,000, 300 batches with 5,000 iterations per batch; to keep the SEs < 0.01). F_{IS} was used as an indicator of heterozygote excess or deficit (12). Fisher's exact tests were performed to test across loci within each population. HWE tests were done in GENEPOP v4.0, which was also used for testing LD between all pairs of loci per population using exact tests as outlined above. Multiple test problems were dealt with by calculating corrected P values using the sequential Bonferroni technique (13).

Analysis of Genetic Diversity: Results. The number of alleles per locus across the seven populations ranged from 9 (locus PP5) to 39 (locus PP3). In total 118 alleles were recovered in 242 isolates. Results of the allelic richness analysis showed that Japan had the highest and the United States the lowest numbers of alleles (Fig. 2). Private alleles (Table S2) in populations were nearly always due to single isolates. However, private alleles at PP3 and PP6 in Japan were found in two to four isolates. The same was true for private alleles at PP2, PP3, PP5, and PP6 in Canada. In New Zealand, private alleles at PP2, PP3, PP4, and PP6 were shared by multiple isolates with up to 17 isolates for PP4 and 39 isolates for PP6.

H_o averaged over all loci for each population ranged from 0.55 (United States) to 0.75 (Japan) and was similar to H_e except in Japan, where lower H_o values were found (Table S2). Hence HWE tests revealed only significant deviations at PP3 and PP4 in the Japanese population, which could probably be attributed to the higher allelic richness at these loci compared with other populations. In the New Zealand population there was also a significant HWE deviation at PP1 (Table S2). There was no evidence that scoring errors due to large allele drop-out or stutter contributed to the observed HWE deviations. Exact tests for LD yielded eight significant values out of 105 pairwise tests (7.6%), six of them were found in the Japanese population (five of them involving PP3 and PP4). This suggests that LD is not a particularly important characteristic of this dataset.

STRUCTURE Analysis: Methods. The Bayesian clustering program STRUCTURE v2.2 (14) was used to infer population structure without predefined population subdivision. STRUCTURE divides sampled individuals into a number of clusters (K) independent of locality information (i.e., based only on MLGs), so as to minimize deviations from Hardy-Weinberg and linkage equilibrium. Individuals are probabilistically assigned to one cluster or more than one cluster if they are genetically admixed. The most likely number of populations (K) was estimated by performing 10 independent runs for each value of K from 2 to 10 with a burn-in and run length of 100,000 repetitions and using the model with correlated allele frequencies, noninformative priors and assuming admixture. The estimated "Ln Probability of Data [Ln P(D)]" given K was used as a criterion to select the most likely number of populations (K) represented by our data by looking for either a maximum value or a plateau for increasing K. The lower value of K showing such behavior was considered as representative of the most appropriate clustering model (15). Using this criterion, the most likely number of populations was estimated to be $K = 6$ (Fig. S2). For the selected K value, we evaluated the individual membership coefficient (q_{ind}) to the inferred clusters. Individuals with a proportion of membership to each cluster $q_{ind} < 0.90$ (admixed individual) were assigned to multiple clusters, whereas individuals with $q_{ind} \geq 0.90$ were assigned to a single cluster. CLUMPP v1.1.1 (16) was used to line up the cluster labels across runs and to estimate the degree of congruence between independent runs, by calculating the Symmetric Similarity Coefficient (SSC) for pairs of runs at each K value, resulting in an average pairwise similarity measure named H (the nearer H is to one, the higher the degree of congruence between independent runs). The Fullsearch algorithm was used for K=2 and K=3 and the Greedy algorithm for greater values of K (testing a predefined number of random sequences: 100,000 for K=4; 10,000 for K=5 and 1,000 for K>6).

STRUCTURE Analysis: Results. Using the six-cluster model (Fig. S2), we found three clusters that showed little admixing and three clusters with higher levels of admixing (Fig. 3). Isolates from the Pacific United States, Japan, and New Zealand could be readily assigned to the clusters corresponding to their geographic origin except in a few rare cases. The North Atlantic clusters showed

more admixing. For Canadian isolates, 18 out of 26 isolates had a $q_{ind} > 0.9$ with respect to a "Canadian" cluster. The northeastern Atlantic isolates showed higher levels of admixture. Of the Belgian isolates, 23 out of 50 could be assigned to the "purple" cluster with a probability > 0.9. The majority of Belgian isolates had the highest probability of belonging to the "purple" cluster ($q_{ind} > 0.6$). Only five Danish isolates had a $q_{ind} > 0.9$ of belonging to one cluster (purple) and almost an equal amount of isolates had highest q_{ind} (> 0.6) in the purple or the green cluster. Of the Irish isolates, almost half of the isolates could be readily assigned to the green cluster ($q_{ind} > 0.9$) and 42 isolates had the highest probability ($q_{ind} > 0.6$) in that cluster.

1. Evans KM, Hayes PK (2004) Microsatellite markers for the cosmopolitan marine diatom *Pseudo-nitzschia pungens*. *Mol Ecol Notes* 4:125–126.
2. Casteleyn G, et al. (2009) Lack of population genetic structuring in the marine planktonic diatom *Pseudo-nitzschia pungens* (Bacillariophyceae) in a heterogeneous area in the Southern Bight of the North Sea. *Mar Biol* 156:1149–1158.
3. Evans KM, Kuhn SF, Hayes PK (2005) High levels of genetic diversity and low levels of genetic differentiation in North Sea *Pseudo-nitzschia pungens* (Bacillariophyceae) populations. *J Phycol* 41:506–514.
4. Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: Software for identifying and correcting genotyping errors in microsatellite data. *Mol Ecol Notes* 4:535–538.
5. Glaubitz JC (2004) CONVERT: A user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. *Mol Ecol Notes* 4:309–310.
6. Park SDE (2001) *Trypanotolerance in West African cattle and the population genetic effects of selection.* PhD thesis (Univ of Dublin, Ireland).
7. Valiere N (2002) GIMLET: A computer program for analysing genetic individual identification data. *Mol Ecol Notes* 2:377–379.
8. Waits LP, Luikart G, Taberlet P (2001) Estimating the probability of identity among genotypes in natural populations: Cautions and guidelines. *Mol Ecol* 10:249–256.
9. Rousset F (2008) GENEPOP'007: A complete re-implementation of the GENEPOP software for Windows and Linux. *Mol Ecol Resour* 8:103–106.
10. Van Loon EE, Cleary DFR, Fauvelot C (2007) ARES: Software to compare allelic richness between uneven samples. *Mol Ecol Notes* 7:579–582.
11. Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* 48:361–372.
12. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358–1370.
13. Rice WR (1989) Analyzing tables of statistical tests. *Evolution* 43:223–225.
14. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.
15. Pritchard JK, Wen X, Falush D (2009) STRUCTURE 2.2.
16. Jakobsson M, Rosenberg NA (2007) CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23:1801–1806.

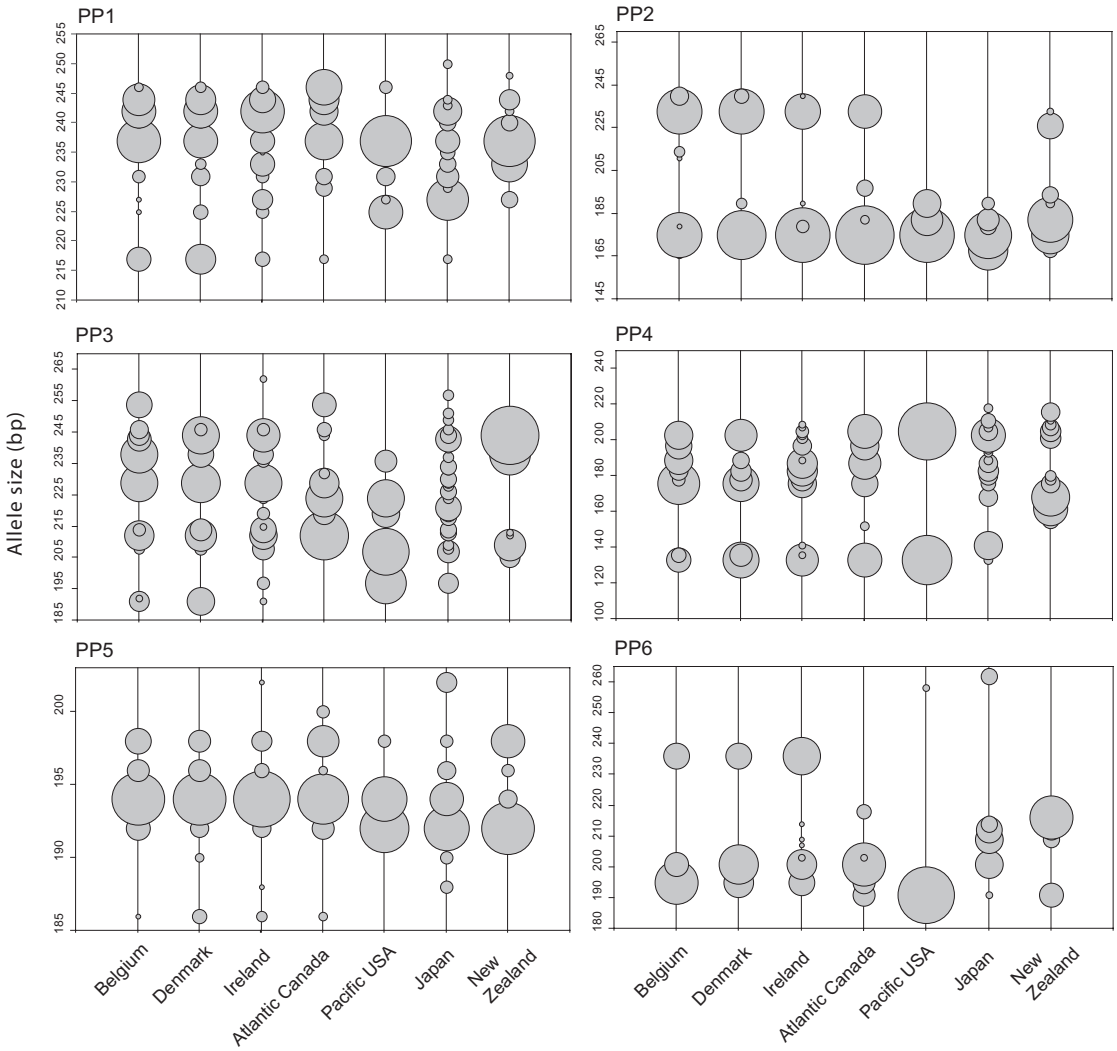

Fig. S1. Allele frequencies for the seven predefined populations: Belgium, Denmark, Ireland, Atlantic Canada, Pacific United States, Japan and New Zealand. Individual circles are centered at a specific allele size. The area of a particular circle is proportional to the frequency of that allele.

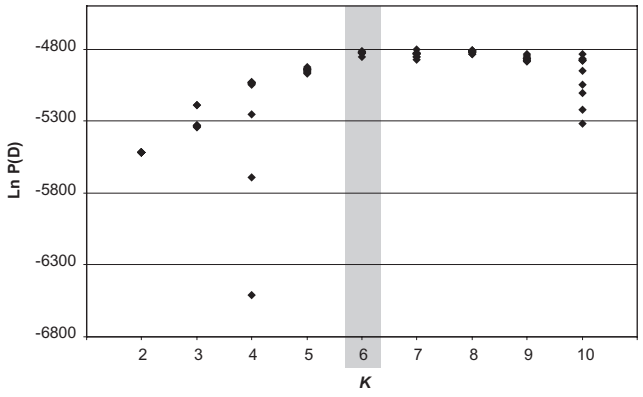

Fig. S2. Estimation of the most likely number of populations (*K*) by performing 10 independent STRUCTURE runs for each value of *K* from 2 to 10. Estimated Ln Probability of Data [Ln P(D)] over 10 runs as a function of *K* is shown.

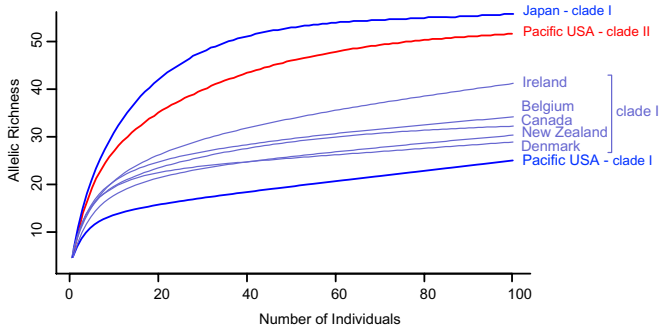

Fig. S3. Comparison of allelic richness between predefined populations of *P. pungens* clade II (northeastern Pacific) (red line), and clade I from both sides of the northern Pacific (dark blue lines) and other regions (light blue lines). Allelic richness was inferred from multilocus genotypes [four microsatellite loci analyzed by Adams et al. (1): PP2, PP3, PP5, PP6], and extrapolated beyond the sample sizes using ARES (2). No microsatellite data are yet available for *P. pungens* clade III populations.

1. Adams NG, et al. (2009) Genetic population structure of *Pseudo-nitzschia pungens* (Bacillariophyceae) from the Pacific Northwest and the North Sea. *J Phycol* 45:1037–1045.
2. Van Loon EE, Cleary DFR, Fauvelot C (2007) ARES: Software to compare allelic richness between uneven samples. *Mol Ecol Notes* 7:579–582.

Table S1. Isolates of *Pseudo-nitzschia pungens* clade I used in this study

Water basin	Predefined population	Sampling location	Latitude	Longitude	Sampling date	Isolated by	GenBank accession number: ITS	GenBank accession number: *rbc*L	n	n_{tot}
Atlantic Ocean										
North Sea	Belgium	VLIZ 120	51.185°	2.701°	4 May 2007	V. Chepurnov	FN823050		50	50
Limfjord	Denmark		56.8629°	8.94292°	Aug-Sep 1997	N. Lundholm	FN823051	same as FM207548†	20	20
Irish Sea	Ireland	Belfast Harbor, HAB6	54.738°	−5.653°	7 July 2007	G. Casteleyn			29	
		Belfast Harbor, HAB7	54.728°	−5.68°	7 July 2007	G. Casteleyn			20	
		Belfast Harbor, HAB8	54.713°	−5.723°	7 July 2007	G. Casteleyn		same as FM207548†	3	52
Northwestern Atlantic	Canada	Cardigan River, PEI	46.2231°	−62.5708°	Sep 2002	C. Léger	AM778786		1	
		Miramichi Bay, NB	47.1153°	−65.1659°	Sep-Oct 2002	C. Léger	AM778787		3	
		Bay of Fundy, NB	44.9944°	−65.7945°	Oct 2002	C. Léger	AM778788		1	
		Miramichi Bay, NB	47.1153°	−65.1659°	Sep 2003	C. Léger	AM778789		2	
		Malpeque Bay, PEI	46.533°	−63.8°	Oct 2003	C. Léger	AM778790		2	
		Brudenell River, PEI	46.2232°	−63.4842°	Sep 2004	C. Léger	AM778791		2	
		Cardigan River, PEI	46.2231°	−62.5708°	Sep 2004	C. Léger	AM778792		2	
		Boughton Bay, PEI	46.1656°	−62.4769°	Sep 2004	C. Léger	AM778794	same as FM207548†	2	
		Boughton Bay, PEI	46.1656°	−62.4769°	Sep 2005	C. Léger			3	
		Malpeque Bay, PEI	46.533°	−63.8°	Nov 2005	C. Léger			2	
		New London Bay, PEI	46.4906°	−63.4601°	Nov 2005	I. Sahraoui		same as FM207548†	1	
		Miramichi Bay, NB	47.1153°	−65.1659°	Sep 2007	C. Léger		same as FM207548†	5	26
Pacific Ocean										
Northeastern Pacific	United States	BS	48.148°	−125.248°	Oct 2004	N. Adams	FM207591	same as FM207548†	3	
		E3	48.2°	−125.883°	Sep 2005	N. Adams	same as FM207591*	same as FM207548†	5	
		E1	48.039°	−125.616°	Sep 2005	N. Adams	same as FM207591*	same as FM207548†	5	
		E2	48.131°	−125.72°	Sep 2005	N. Adams	same as FM207591*	FM207548	6	
		E4	48.074°	−126.141°	Sep 2005	N. Adams	same as FM207591*	same as FM207548†	8	27
Northwestern Pacific	Japan	Ofunato Bay	39.0617°	141.732°	7 Aug 2000	Y. Kotaki	AM778812	same as FM207548†	1	
					25 Jun 2001	Y. Kotaki	AM778813		1	
					3 Dec 2001	Y. Kotaki	AM778814		1	
					13 Nov 2006	Y. Kotaki		same as FM207548†	21	24
Southwestern Pacific	New Zealand	Steels Reef, North Island	−37.9442°	177.058°	Oct 2000	L. Rhodes	AM778815		1	
		Big Glory Bay, Stewart Island	−46.9824°	168.1°	Feb 2003	L. Rhodes	AM778816		1	
		Taylor's Mistake, South Island	−43.5788°	172.79°	Aug 2004	L. Rhodes	AM778817	same as FM207548†	1	
		Collingwood, South Island	−40.686°	172.69°	Oct 2004	L. Rhodes			1	

Table S1. Cont.

Water basin	Predefined population	Sampling location	Latitude	Longitude	Sampling date	Isolated by	GenBank accession number		n	n_{tot}
							ITS	*rbc*L		
		Nydia Bay, South Island	−41.1653°	173.786°	Aug 2005	L. Rhodes			1	
		Cannon Bay, North Island	−41.0987°	174.895°	Sep 2005	L. Rhodes			1	
		Gisborne Harbor, North Island	−38.7°	178°	18 Apr 2007	L. Rhodes		same as FM207548[†]	37	43
										242

NB: New Brunswick; PEI: Prince Edward Island; n: number of isolates; n_{tot}: total number of isolates per population.
*Sequence identity determined by direct sequencing (1).
[†]Sequence identity determined by Denaturing Gradient Gel Elecrophoresis (DGGE) (1).

1. Casteleyn G, et al. (2009) Natural hybrids in the marine diatom *Pseudo-nitzschia pungens* (Bacillariophyceae): genetic and morphological evidence. *Protist* 160:343–354.

Table S2. Number of *Pseudo-nitzschia pungens* clade I isolates genotyped (*n*), size ranges of the alleles (bp), number of alleles (N_A), number of private alleles (N_{priv}), number of unique genotypes observed (G_o), observed (H_o) and expected (H_e) heterozygosity, F_{IS} according to Weir and Cockerham (1) and probabilities of identity ($P_{(ID)}$ and $P_{(ID)sib}$) across loci for every population and overall populations

	Populations							
locus	Belgium	Denmark	Ireland	Canada	United States	Japan	New Zealand	Total
PP1								
n	50	20	52	26	27	24	43	242
bp	217–246	217–246	217–246	217–246	225–246	217–250	227–248	217–250
N_A	9	9	10	7	6	12	9	15
N_{priv}	0	0	0	0	0	2	1	
G_o	17	12	25	12	9	15	15	
H_o	0.76	1.00	0.81	0.77	0.78	0.75	0.63	
H_e	0.78	0.84	0.81	0.81	0.70	0.84	0.70	
F_{IS}	0.03	−0.18	0.01	0.05	−0.12	0.11	**0.11**	
PP2								
n	50	20	52	26	27	24	43	242
bp	166–240	168–240	168–240	175–233	166–190	168–190	168–233	166–240
N_A	9	5	6	4	4	5	7	13
N_{priv}	2	0	0	1	0	0	1	n/a
G_o	13	6	9	6	5	8	11	28
H_o	0.70	0.65	0.52	0.50	0.74	0.67	0.81	
H_e	0.68	0.62	0.56	0.49	0.58	0.68	0.71	
F_{IS}	−0.03	−0.04	0.07	−0.02	−0.28	0.01	−0.14	
PP3								
n	50	20	52	26	27	24	43	242
bp	191–254	191–246	191–262	212–254	196–236	197–257	205–244	191–262
N_A	12	9	15	9	7	26	8	39
N_{priv}	1	0	2	1	2	11	2	n/a
G_o	30	15	36	12	13	21	9	97
H_o	0.86	0.75	0.92	0.96	0.63	0.83	0.81	
H_e	0.87	0.87	0.89	0.81	0.79	0.97	0.69	
F_{IS}	0.02	0.14	−0.04	−0.19	0.20	**0.14**	−0.19	
$P_{(ID)}$								
$P_{(ID)sib}$								
PP4								
n	50	20	52	25	27	24	43	241
bp	133–203	133–203	133–209	133–205	133–205	133–218	155–216	133–218
N_A	10	9	16	10	3	18	13	29
N_{priv}	0	0	0	2	0	5	3	n/a
G_o	26	15	28	17	4	17	26	93
H_o	0.74	0.80	0.92	0.80	0.63	0.71	0.86	
H_e	0.84	0.86	0.88	0.87	0.51	0.93	0.87	
F_{IS}	0.12	0.07	−0.05	0.08	−0.23	**0.23**	0.01	
PP5								
n	50	20	52	26	27	24	43	242
bp	186–198	186–198	186–202	186–200	192–198	188–202	192–198	186–202
N_A	5	6	7	6	3	7	4	9
N_{priv}	0	0	0	1	0	0	0	n/a
G_o	9	10	10	8	5	10	7	21
H_o	0.60	0.55	0.42	0.62	0.48	0.83	0.40	
H_e	0.60	0.62	0.48	0.61	0.54	0.74	0.56	
F_{IS}	0.01	0.11	0.11	−0.01	0.11	−0.13	0.30	
PP6								
n	50	20	52	26	27	24	43	242
bp	195–236	195–236	195–236	191–218	191–258	191–262	191–216	191–262
N_A	3	3	7	5	2	7	5	13
N_{priv}	0	0	0	1	1	1	1	n/a
G_o	6	6	10	7	2	14	8	35
H_o	0.54	0.70	0.58	0.65	0.04	0.71	0.74	
H_e	0.57	0.65	0.69	0.60	0.04	0.83	0.63	
F_{IS}	0.06	−0.08	0.16	−0.09	0.00	0.14	−0.18	

Table S2. Cont.

locus	Populations							
	Belgium	Denmark	Ireland	Canada	United States	Japan	New Zealand	Total
Across loci								
n	50	20	52	26	27	24	43	242
N_A*	8.0	6.8	10.2	6.8	4.2	12.5	7.7	19.7
H_o	0.70	0.74	0.70	0.72	0.55	0.75	0.71	
H_e	0.72	0.75	0.72	0.70	0.53	0.83	0.69	
F_{IS}	0.03	0.00	0.03	−0.03	−0.04	**0.10**	**-0.02**	
$P_{(ID)}$	4.70E-07	8.48E-08	2.01E-07	6.04E-07	1.67E-04	5.57E-12	1.79E-06	3.75E-09
$P_{(ID)sib}$	5.26E-03	4.87E-03	5.24E-03	6.95E-03	2.87E-02	2.07E-03	7.21E-03	1.95E-03

In bold: significant deviation from HWE (after sequential Bonferroni correction). Totals per locus, where appropriate are shown in the last column.
*Average number of alleles across the six loci.

1. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358–1370.

Table S3. Representatives of *Pseudo-nitzschia* clade I [as presented in Lundholm et al. (1)] used in the phylogenetic analysis with indication of GenBank numbers

	*rbc*L	LSU rDNA	rDNA ITS
P. multistriata CM2	EF423505	AF416753	DQ990368
P. multistriata KoreaA			AY257843
P. multistriata CM1			DQ990367
P. americana	EF423504	U41390	EU523099
P. seriata f. *obtusa* T5			DQ062667
P. seriata Lynaes8		AF417653	DQ062666
P. seriata Nissum3			AY257841
P. australis PLYSt54B		AF417651	AY452528
P. australis au43			DQ062661
P. multiseries PM02		AF440772	EU302796
P. multiseries OFPm984			DQ062664
P. pungens (clade III) P24	EF423506	AF417650	AY257845
P. pungens (clade III) Mex18			AY257846
P. pungens (clade III) Viet			DQ166533
P. pungens (clade III) KBH2			DQ062665
P. pungens (clade II) US115	FM207547	AF440775	AM778804
P. pungens (clade II) US135			AM778811
P. pungens (clade I) V120			AM778747
P. pungens (clade I) Cn172	FM207548	EF642979	AM778786

1. Lundholm N, et al. (2006) Inter- and intraspecific variation of the *Pseudo-nitzschia delicatissima* complex (Bacillariophyceae) illustrated by rRNA probes, morphological data and phylogenetic analyses. *J Phycol* 42:464–481.