CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Opposing patterns of intraspecific and interspecific differentiation in sex chromosomes and autosomes
Authors	Moran, Peter A.;Pascoal, Sonia;Cezard, Timothee;Risse, Judith E.;Ritchie, Michael G.;Bailey, Nathan W.
Publication date	2018-05-22
Original Citation	Moran, P. A., Pascoal, S. , Cezard, T. , Risse, J. E., Ritchie, M. G. and Bailey, N. W. (2018) 'Opposing patterns of intraspecific and interspecific differentiation in sex chromosomes and autosomes', Molecular Ecology. doi:10.1111/mec.14725
Type of publication	Article (peer-reviewed)
Link to publisher's version	10.1111/mec.14725
Rights	© 2018, John Wiley & Sons Inc. This is the peer reviewed version of the following article: Moran, P. A., Pascoal, S. , Cezard, T. , Risse, J. E., Ritchie, M. G. and Bailey, N. W. (2018) 'Opposing patterns of intraspecific and interspecific differentiation in sex chromosomes and autosomes', Molecular Ecology. doi:10.1111/mec.14725, which has been published in final form at https://doi.org/10.1111/mec.14725. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.
Download date	2026-09-09 16:35:16
Item downloaded from	https://hdl.handle.net/10468/6207

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

MOLECULAR ECOLOGY

Supplemental Information for:

Opposing patterns of intraspecific and interspecific differentiation in sex chromosomes and autosomes

Peter A. Moran[1] | Sonia Pascoal[2] | Timothee Cezard[3] | Judith E. Risse[4] | Michael G. Ritchie[5] | Nathan W. Bailey[5]

Table of Contents:

S.1 Additional sampling information	Page 2
S.2 Comparison of three different *de novo* assemblies for SNP-calling	Page 3
S.3 Putative autosomal and X-linked loci	Page 5
S.4 Demographic models	Page 12
S.5 Comparison of filtering approaches and diversity estimates	Page 13
S.6 Formulas	Page 16

MOLECULAR ECOLOGY

S.1 | Additional sampling information

Table S1 Sampled sites in eastern Australia, population codes, species identity, geographic coordinates and sample sizes for genomic data.

Population	Pop	Species	Latitude	Longitude	Female DNA	Male DNA
Altona Meadows, VIC	**AM**	*T. com*	-37.8827	144.7834	16	14
Bairnsdale, VIC	**BN**	*T. com*	-37.8264	147.6378	17	17
Cooma Creek, NSW	**CC**	*T. com*	-36.2353	149.118	17	17
Moss Vale, NSW	**MV**	*T. com*	-34.4867	150.3736	16	14
Bluey's Beach, NSW	**BL**	*T. com*	-32.3497	152.5357	17	17
Coff's Harbour, NSW	**CH**	*T. com*	-30.295	153.1167	16	17
Brisbane, QLD	**UQ**	*T. com*	-27.4951	153.0123	10	5
Maleny, QLD	**SV**	*T. com*	-26.7528	152.8472	17	17
Hervey Bay, QLD	**HB**	*T. o / T. c mix*	-25.3008	152.8631	16	14
Tannum Sands, QLD	**TS**	*T. o / T. c mix*	-23.9627	151.332	17	17
Rockhampton, QLD	**RH**	*T. o / T. c mix*	-23.3835	150.5067	17	15
Yeppoon, QLD	**YP**	*T. o / T. c mix*	-23.1444	150.7635	16	17
Mount Pleasant, QLD	**PL**	*T. oc*	-21.1129	149.1558	17	17
Townsville, QLD	**JC**	*T. oc*	-19.3288	146.7599	17	17
Daintree, QLD	**DV**	*T. oc*	-16.249	145.3223	16	9
Daintree (KH), QLD	**KH**	*T. marini*	-16.249	145.3223	8	6
					250	**230**
					All = 480	

MOLECULAR ECOLOGY

S.2 | Comparison of three different *de novo* assemblies for SNP-calling

If the species examined are highly divergent, there may be a large discrepancy in the number of SNPs returned for the three different assembly methods. In particular, species-specific or rare variants may be lost which may affect downstream analysis. This is an important consideration for merging intraspecific and interspecific studies without a reference genome. Table S2 compares the number of SNPs pre- and post-filtering derived from a *de novo* assembly constructed from data from both *Teleogryllus* species, and *de novo* assemblies constructed from data for each species separately.

Venn diagrams of this data (Figure S1) indicate that for species comparisons, the three approaches result in a similar number of shared SNPs, while for intraspecific analysis the choice of assembly has a greater effect on the number of species specific SNPs obtained. A higher number of SNPs unique to each species group were obtained when using the corresponding species specific assembly. The combined assembly may be slightly biased towards *T. commodus* due to the larger number of these individuals sampled. *T.marini* shares a low number of variants with both species. The large number of species-specific variants may represent either fixed species differences or sites that are missing (or due to allelic drop out) or were removed during filtering as they were at a low frequency.

Table S2 Comparison of the number of SNPs obtained from the three different assemblies at different thresholds of missing data allowed. The first three rows indicate SNP numbers when filtering is applied across data sets containing all populations (individuals from the KH population were removed due to presence of the third species *T.marini*). The bottom six rows indicate number of SNPs returned after subsetting into species specific groups. Filtered SNPs indicates the number of SNPs obtained after removing sites which were missing across more than 80% (0.8), 50% (0.5), 20% (0.2) of individuals.

Assembly	Total data	SNPs	Samples	Filtered SNPs (0.8)	Filtered SNPs (0.5)	Filtered SNPs (0.2)
Combined		1,158,094	464	178,062	103,958	36,500
T. oceanicus		1,199,268	464	186,851	104,388	36,653
T. commodus		1,585,852	464	223,403	108,795	33,186
	Species subset					
Combined	*T. oc*		195			38,708
	T. com		269			41,757
T. oceanicus	*T. oc*		195			41,928
	T. com		269			35,972
T. commodus	*T. oc*		195			35,131
	T. com		269			38,121

MOLECULAR ECOLOGY

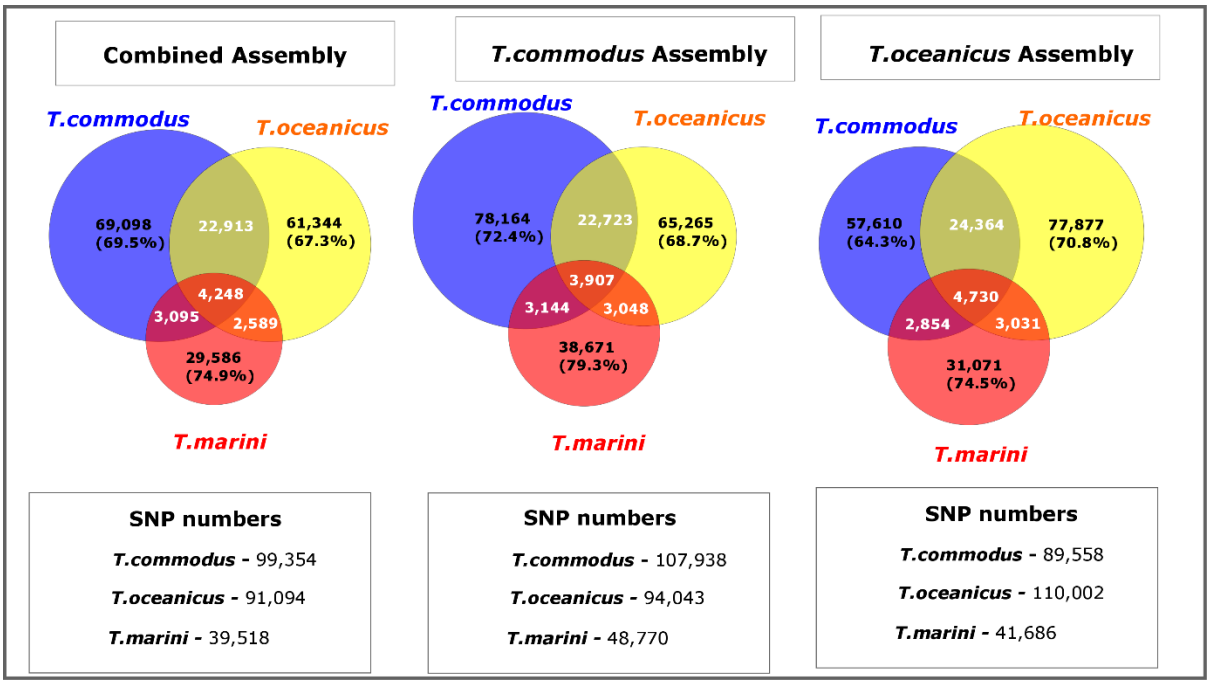
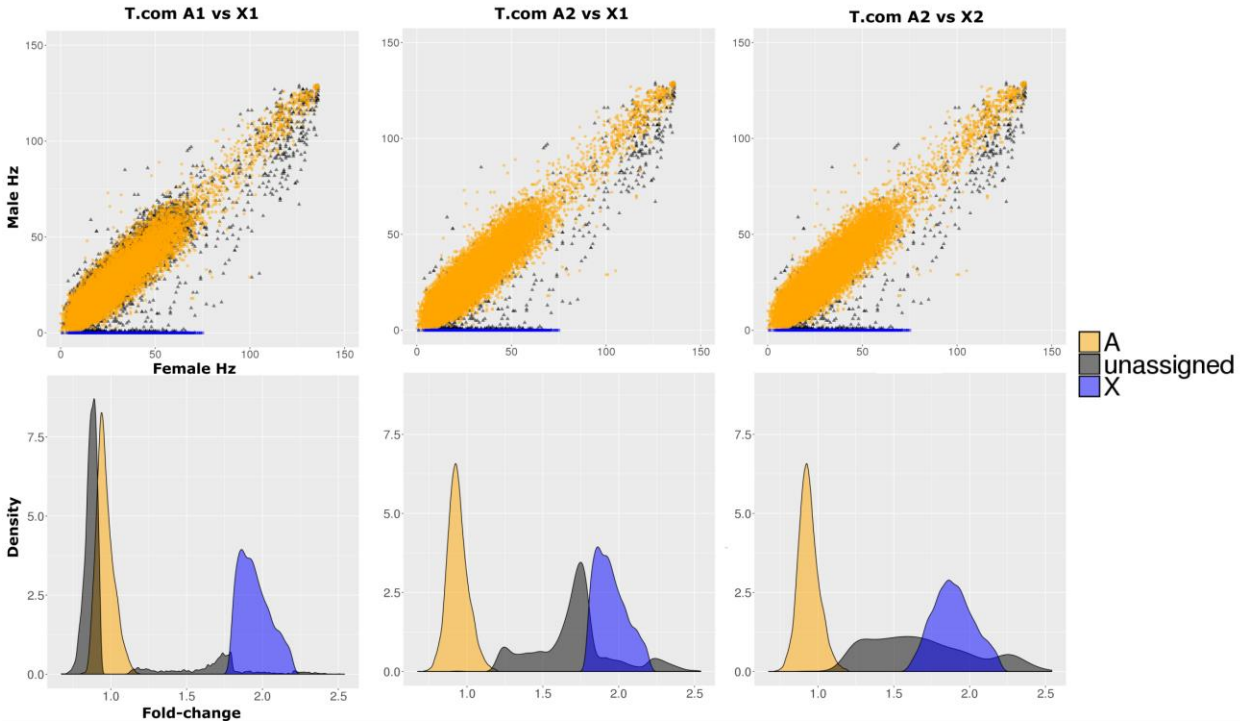


Figure S1 Venn diagram illustrating the overlap of variants amongst the three putative species for the three different assembly methods. Text box indicates the total number of SNPs per group (after filtering). Individuals were assigned to their species group based on their group assignment in the Bayesian clustering analysis. VCF files were filtered to remove variants with > 0.5 missing data and comparisons were implemented using bcftools.

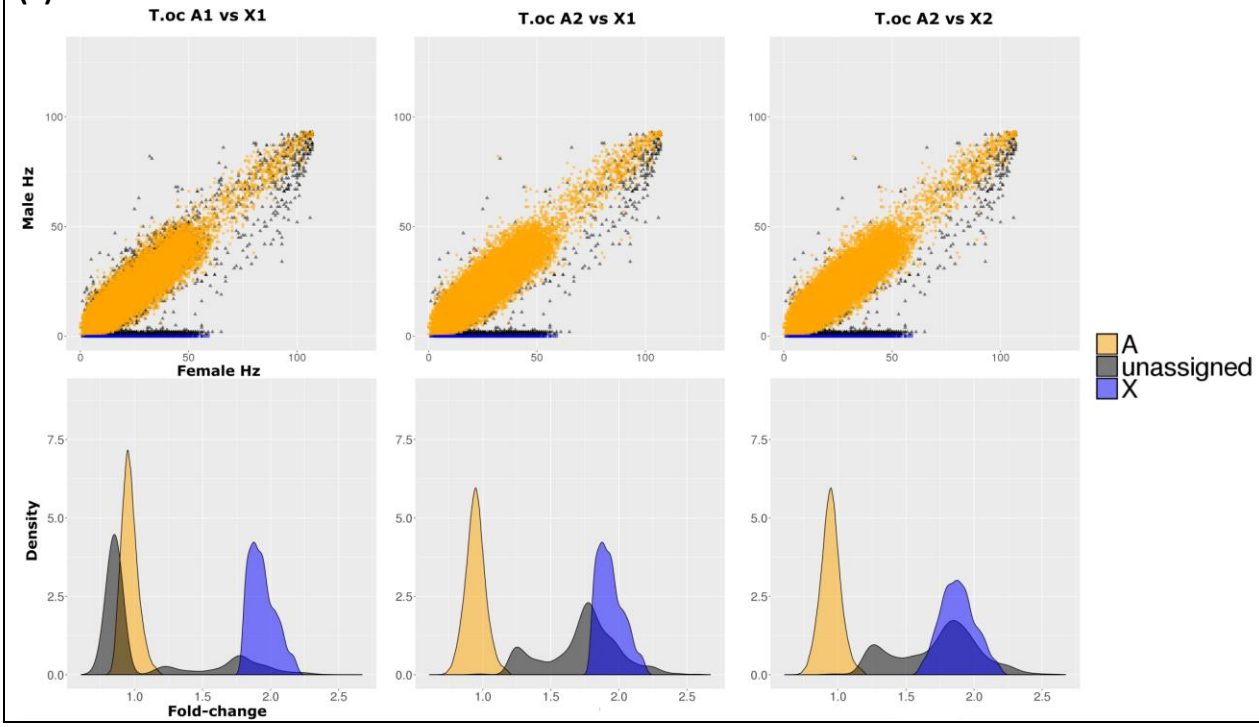
MOLECULAR ECOLOGY

S.3 | Putative Autosomal and X-linked loci

(i) *T. commodus*



(ii) *T. oceanicus*



MOLECULAR ECOLOGY

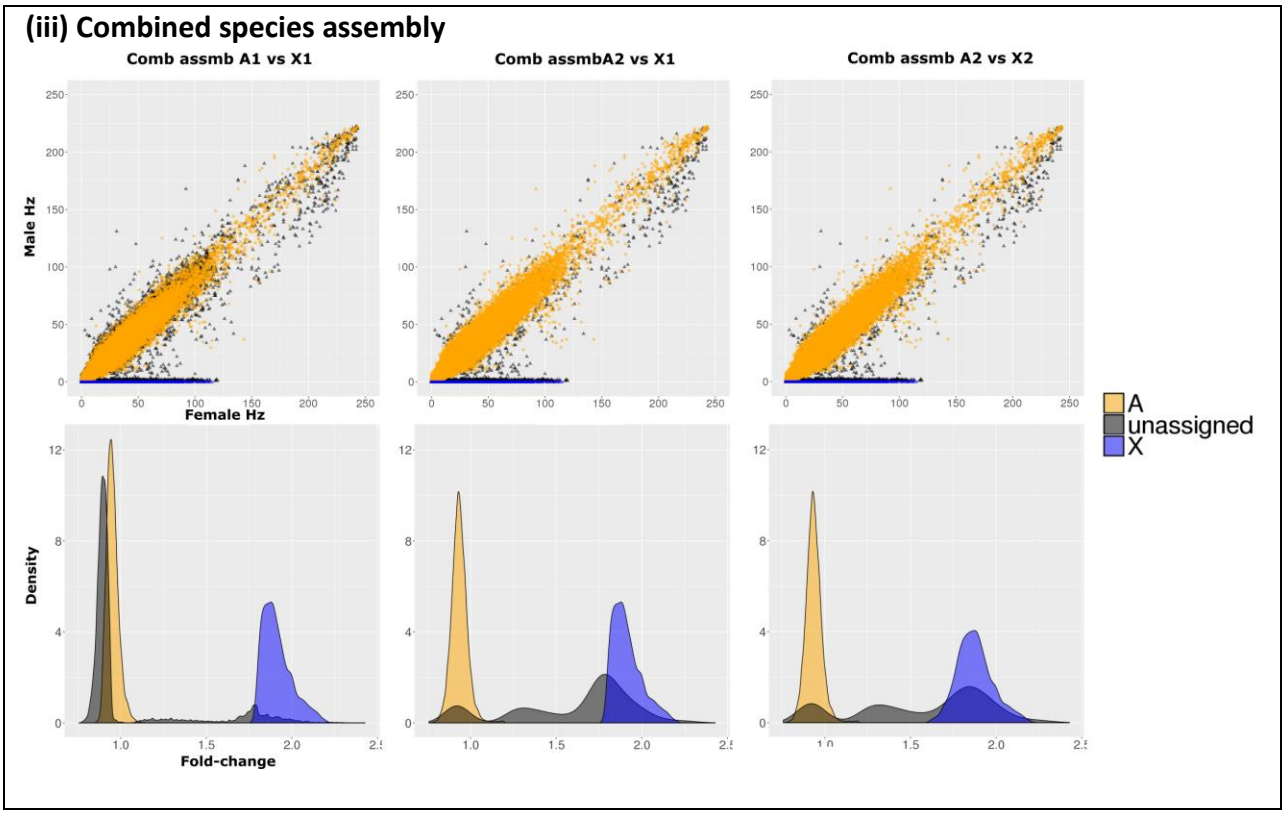


Figure S2. Heterozygosity and fold-change coverage for autosomal (A) and putative X-linked SNPs based on the different species assemblies: (i) T. commodus, (ii) *T. oceanicus* and (iii) the combined species assembly. Colours correspond to the SNP groups assigned based on the combination of autosomal and X filters applied (A:X 1-2; indicated above the figures). Orange represents autosomal SNPs, blue are X-linked SNPs and grey are unassigned SNPs.

Table S3. X-A filtering parameters. Three different fold-change ranges (female coverage/male coverage) were examined for X and A markers (numbered 1-3 below). The fold-change ranges are based on the expectations that an X-linked SNP should have twice the coverage in females than males, whereas autosomal SNPs should exhibit equal coverage between the sexes. For A2 and A3 filters the p-value check (Student's t-test and Bonferroni correction) was removed as it was too restrictive. A1 and X1 rows are shaded as these are the filtered data we present a full analyses of in the main text.

Filters	Fold-change range (fc)	t-test (p-value check)
A1	0.8 – 1.2	>0.05
A2	0.6 – 1.2	No p-value check
A3	0.3 – 1.2	No p-value check
X1	1.8 – 2.2	< 0.05
X2	1.6 – 2.2	< 0.05
X3	1.3 – 2.2	< 0.05

MOLECULAR ECOLOGY

Table S4. The numbers of SNPs assigned to autosomal and X groups based on the combination of filters applied. SNPs that failed to be assigned to either group are referred to as unassigned ("Un") and were omitted from further analyses. X/A indicates the ratio of X to autosomal loci. A1 and X1 rows are shaded as these are the filtered data we present a full analyses of in the main text.

Species - Filter	Total SNPs	A	X	Un	X/A
T.com	40728				
A1 – X1		26447	2405	11876	9.09
A2 – X1		36342	2405	1981	6.62
A2 –X2		36342	3229	1157	8.89
A3 – X3		36342	3293	1093	9.06
T.oc	44941				
A1 – X1		34010	1288	9643	3.79
A2 – X1		41061	1288	2592	3.14
A2 –X2		41061	1783	2097	4.34
A3 – X3		41061	1810	2070	4.41
Comb Assmb	39388				
A1 – X1		23411	1838	14139	7.85
A2 – X1		34663	1838	2887	5.30
A2 –X2		34663	2397	2328	6.92
A3 – X3		34663	2403	2322	6.93

Table S5. Summary statistics for the different X and A filtered datasets. *Ho* is the observed heterozygosity, H_S is the within-population gene diversity. F_{IS} is the inbreeding coefficient. A1 and X1 rows are shaded as these are the filtered data we present a full analyses of in the main text. The bottom two rows (A1 – HWE and X1-HWE) correspond to SNP datasets in which loci that significantly deviated from Hardy Weinberg equilibrium (HWE) were removed (see Table S7 for details on HWE tests). Summary statistics were calculated for females only to avoid the influence of male hemizygoisty.

Filters	T.com			T.oc		
	Ho	Hs	Fis	Ho	Hs	Fis
A1	0.207	0.227	0.069	0.219	0.243	0.084
A2	0.207	0.228	0.069	0.222	0.243	0.075
A3	0.207	0.228	0.069	0.222	0.243	0.075
X1	0.218	0.241	0.073	0.202	0.233	0.102
X2	0.206	0.231	0.081	0.192	0.224	0.109
X3	0.204	0.230	0.084	0.191	0.223	0.111
A1 - HWE	0.196	0.217	0.068	0.213	0.235	0.077
X1 - HWE	0.221	0.241	0.061	0.204	0.231	0.089

MOLECULAR ECOLOGY

Table S6. Mean pairwise population F_{ST} values with 95% confidence intervals (based on 1,000 bootstraps) for the different filtered sets of SNPs. In the combined species assembly ("Comb assmb") genetic differentiation is reported for the overall interspecific comparisons and also the specific geographic comparisons between allopatric (allo) and sympatric (sym) populations. A1 and X1 rows are shaded as these are the filtered data we present a full analyses of in the main text. The bottom two rows (A1 – HWE and X1-HWE) correspond to SNP datasets in which loci that significantly deviated from Hardy Weinberg equilibrium (HWE) were removed (see Table S7 for details on HWE tests).

	T.oc	T.com	Comb Assmb		
			Interspecific	T.com vs. T.oc sym	T.com vs. T.oc allo
A1	0.018 [0.017 – 0.02]	0.036 [0.034 – 0.037]	0.331 [0.324 - 0.336]	0.318 [0.311 - 0.324]	0.336 [0.330 - 0.341]
A2	0.018 [0.018 – 0.02]	0.037 [0.035 – 0.038]	0.349 [0.345 - 0.354]	0.336 [0.331 - 0.341]	0.355 [0.351 - 0.360]
A3	0.018 [0.018 – 0.02]	0.037 [0.035 – 0.038]	0.349 [0.345 - 0.354]	0.336 [0.331 - 0.341]	0.355 [0.351 - 0.360]
X1	0.001 [-0.002 – 0.005]	0.018 [0.013 – 0.022]	0.484 [0.461 - 0.507]	0.523 [0.499 - 0.546]	0.468 [0.445 - 0.490]
X2	0.002 [-0.001 – 0.005]	0.019 [0.015 – 0.022]	0.45 [0.431 - 0.468]	0.489 [0.468 - 0.509]	0.434 [0.415 - 0.452]
X3	0.002 [-0.001 – 0.005]	0.018 [0.014 – 0.022]	0.45 [0.432 - 0.473]	0.491 [0.471 - 0.514]	0.435 [0.416 - 0.456]
A1 - HWE	0.011 [0.011 – 0.013]	0.03 [0.027 – 0.032]			
X1 - HWE	0.002 [-0.002 – 0.006]	0.017 [0.011 – 0.023]			

MOLECULAR ECOLOGY

Table S7. Average nucleotide diversity (π) estimates per-site for X and A markers. π was calculated per-site using VCFtools. This measure represents an approximation based on variant sites only (π SNP). To account for sequence length including invariant sites, we standardized π by overall SNP density from each of the four combinations of species and chromosome type (π per-site/ (total sites/variable sites)). Both measures give a very similar estimate of overall X/A diversity.

Species/ Marker	Total sites	SNPs	SNP density	π SNP	π Corrected	X/A SNP	X/A Corrected
T.com A1	11404878	26447	431.24	0.225	0.00052	1.062	1.081
T.com X1	1018888	2405	423.65	0.239	0.00056		
T.oc A1	13508581	34010	397.19	0.242	0.00061	0.959	0.925
T.oc X1	530326	1288	411.74	0.232	0.00056		

Table S8. Hardy–Weinberg Equilibrium (HWE) test results (Bonferroni-corrected) reporting the total number of SNPs for each marker set, the number of loci which exhibited significant heterozygous excess or deficit and their relative proportions. The analysis was restricted to female samples to avoid male hemizygosity for X-linked loci.

Assmbl/ Filter	SNPs	Def/Exc	Def	Exc	% Def/Exc	% Def	% Exc
T.com_assmb_X1	2405	93	93	0	3.87	3.87	0
T.com_assmb_A1	26447	2202	1629	573	8.33	6.16	2.17
T.com_assmb_A2	36342	2630	1919	711	7.24	5.28	1.96
T.oc_assmb_X1	1288	49	49	0	3.8	3.8	0
T.oc_assmb_A1	34010	2352	1798	554	6.92	5.29	1.63
T.oc_assmb_A2	41061	2582	1911	671	6.29	4.65	1.63
Comb_assmb_X1	1838	768	768	0	41.78	41.78	0
Comb_assmb_A1	23411	8145	7516	629	34.79	32.1	2.69
Comb_assmb_A2	34663	11741	10879	862	33.87	31.39	2.49

MOLECULAR ECOLOGY

Table S9. Linkage Disequilibrium (LD) estimates (r2, D') for X and A SNPs for both species. The number of significant SNP pairs in LD (p-value <0.05) and their overall proportion are shown.

Species	Filter (SNPS)	r2	D'	Sig SNPs	Prop
T.oc	A1 (1,000)	0.009	0.171	84800	0.170
T.oc	X1 (437)	0.007	0.182	16654	0.175
X - A		-0.002	0.011		0.005
T.com	A1 (1,000)	0.007	0.159	92743	0.186
T.com	X1 (836)	0.008	0.175	78444	0.225
X - A		0.001	0.015		0.039

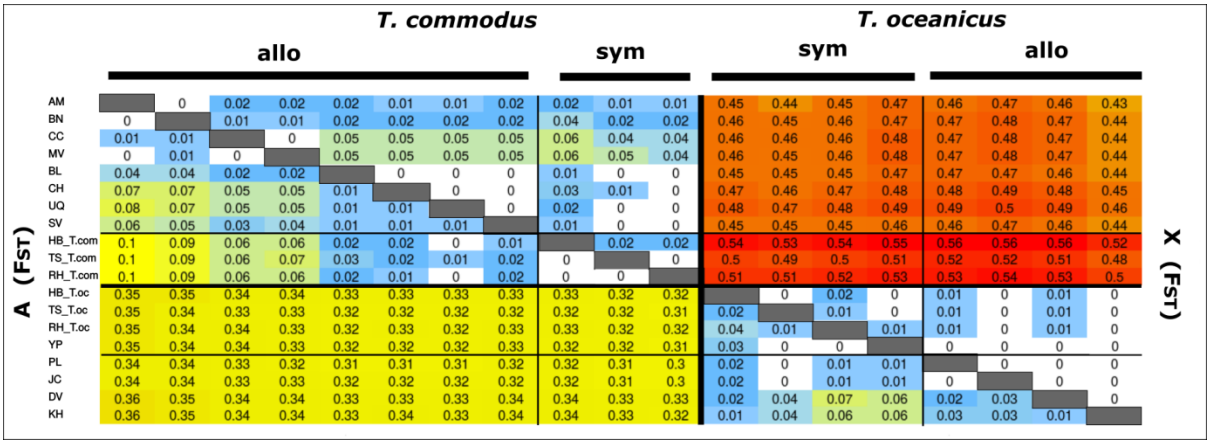


Figure S3. Pairwise population F_{ST} at autosomal loci below the diagonal and at X-linked markers above the diagonal. Colours correspond to increasing F_{ST} values (blue = low, red = high). F_{ST} values values < 0.001 are presented as 0. Populations are arranged in geographic order with AM the most southern population and DV the most northern. Black symbols indicate the geographic regions: ovals = allopatric *T. commodus*; diamonds = sympatry; squares = allopatric *T. oceanicus*.

MOLECULAR ECOLOGY

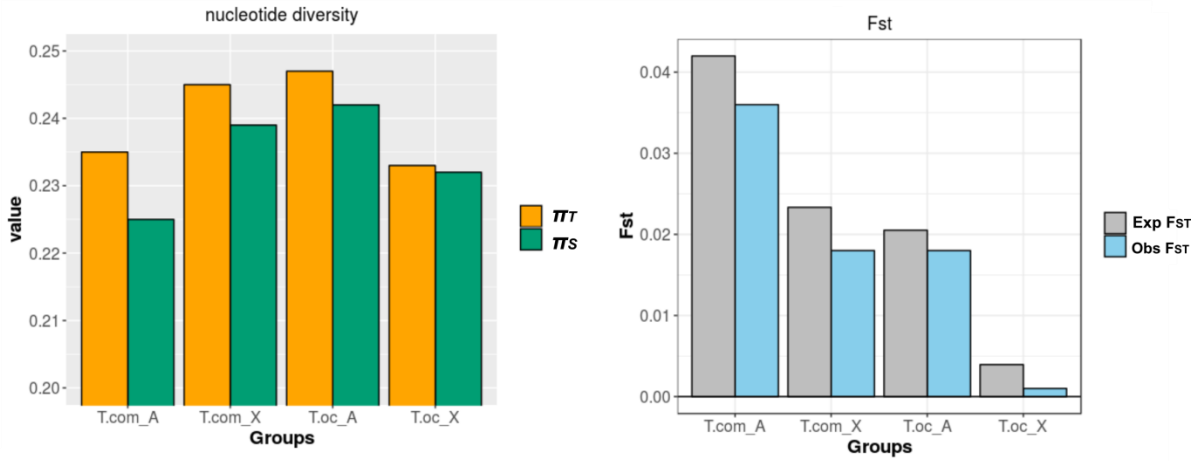


Figure S4. Mean Nucleotide diversity in each subpopulation (π_S) and in the combined subpopulation (π_T) for *T. commodus* and *T. oceanicus* populations (based on variants only). Fst observed (blue) and expected (grey) (expected Fst calculated using Hudson et al., (1992) eq. 3: Fst = $1 - \pi_S / \pi_T$). Using the above equation the proportion of observed within and between population diversity for X and A markers results in reduced X Fst. However, the estimates for expected Fst are higher than those observed. Although the relative difference in Fst between X and A is similar for both observed and expected (*T. commodus*: difference in expected Fst between A and X (A-X) = 0.019, observed Fst (A-X) = 0.018; *T. oceanicus*: expected Fst A-X = 0.017, observed Fst (A-X) = 0.017).

Table S10. Isolation by distance: species comparisons for the slopes and intercepts. Relationship between population genetic differentiation (mean F_{ST}) and geographic distance (Euclidean) at autosomal and X-linked markers (following the methods of Baselga, (2010) described in the Methods).

Species	Marker	intercept	slope
T.com	A	-8.29E-03	5.48E-05
	X	5.85E-03	1.48E-05
T.oc	A	4.51E-03	2.42E-05
	X	3.54E-03	-4.11E-06
T.com vs. T. oc	A	P = 0.002	P < 0.001
T.com vs. T. oc	X	P = 0.244	P < 0.001

MOLECULAR ECOLOGY

S.4 | Demographic Models

Table S10. Demographic model selection using AIC. Four demographic models were implemented in Fastsimcoal; population of constant size, decline, exponential growth and bottleneck. Most of the relative AIC weights were assigned to the model of population growth.

Species	Pop (SNPs)	Model	Log-likelihood	K	AIC	Delta AIC	AIC weight
T.com	AM (16,992)	Cons size	-64026.3	1	128054.6	2967.37	0.00
		Decline	-63742.3	3	127490.6	2403.30	0.00
		Expansion	-62540.6	3	125087.3	0.00	0.73
		Bottleneck	-62540.6	4	125089.2	1.98	0.27
	CH (16,370)	Cons size	-61460.1	1	122922.1	3479.13	0.00
		Decline	-61153.8	3	122313.7	2870.69	0.00
		Expansion	-59718.5	3	119443	0.00	0.70
		Bottleneck	-59718.3	4	119444.7	1.68	0.30
	MV (21,666)	Cons size	-81583	1	163168.1	3480.04	0.00
		Decline	-81255.9	3	162517.8	2829.77	0.00
		Expansion	-79841	3	159688	0.00	0.72
		Bottleneck	-79840.9	4	159689.9	1.85	0.28
	TS (18,731)	Cons size	-70412.8	1	140827.5	3192.19	0.00
		Decline	-70135.2	3	140276.5	2641.14	0.00
		Expansion	-68814.7	3	137635.3	0.00	0.60
		Bottleneck	-68814.1	4	137636.2	0.85	0.40
T.oc	HB (19,007)	Cons size	-71620.7	1	143243.4	2372.05	0.00
		Decline	-71413.2	3	142832.4	1961.06	0.00
		Expansion	-70432.7	3	140871.4	0.00	0.64
		Bottleneck	-70432.2	4	140872.5	1.12	0.36
	JC (18,865)	Cons size	-71011.8	1	142025.5	2307.12	0.00
		Decline	-70806.6	3	141619.3	1900.86	0.00
		Expansion	-69856.2	3	139718.4	0.00	0.70
		Bottleneck	-69856	4	139720.1	1.66	0.30
	YP (19,712)	Cons size	-74270.1	1	148542.3	2320.37	0.00
		Decline	-74064	3	148134.1	1912.15	0.00
		Expansion	-73108	3	146221.9	0.00	0.59
		Bottleneck	-73107.3	4	146222.7	0.76	0.41
	DV (14,793)	Cons size	-55084.6	1	110171.2	3308.85	0.00
		Decline	-54840.2	3	109686.5	2824.09	0.00
		Expansion	-53428.2	3	106862.4	0.00	0.72
		Bottleneck	-53428.1	4	106864.3	1.86	0.28

MOLECULAR ECOLOGY

Table S11. Demographic parameter point estimates from bottleneck model**.** Demographic parameter estimates are from the bottleneck model with the highest likelihood. NPOP is the current effective population size, NANC is the ancestral population size, NBOT is the effective population size during the bottleneck, and TBOT is time in generations ago that the bottleneck ended.

	NPOP	NANC	NBOT	TBOT	NPOP/NANC	NPOP/NBOT	NBOT/NANC
AM	358217	36693	30215	34993	9.76	11.86	0.82
CH	359228	28236	64838	39960	12.72	5.54	2.30
HB	300285	43789	31145	39574	6.86	9.64	0.71
JC	288241	43475	66431	41693	6.63	4.34	1.53
MV	352505	38897	56572	35262	9.06	6.23	1.45
TS	370628	38862	5879	35216	9.54	63.04	0.15
YP	309117	48396	28634	36969	6.39	10.80	0.59
DV	469848	34935	83888	35220	13.45	5.60	2.40

S.5 | Comparison of filtering approaches and diversity estimates

Table S12. To test whether the potential bias in X/A diversity is due to the unequal mix of samples from each species or the assembly used we randomly selected 200 individuals of each species, applied quality filtering (loci must be present in at least 80% of individuals, minor allele frequency > 0.05 and an overall genotype quality score >20), extracted X and A SNPS and then estimated diversity (π_S) for females only for the three different assemblies. Even with an equal mix of individuals from both species and the species specific assemblies the average X/A for *T. oceanicus* (T.oc) is reduced (~0.64); whereas for *T. commodus* (T.com) it is greater than 1 (~1.24).

Assembly	N (T.com, T.oc)	A SNPs	X SNPs	T.com X/A	T.oc X/A
Combined Species	200, 200	22,564	1701	1.228	0.644
T. oceanicus	200, 200	23,133	1779	1.197	0.647
T. commodus	200, 200	19,805	1614	1.307	0.637

MOLECULAR ECOLOGY

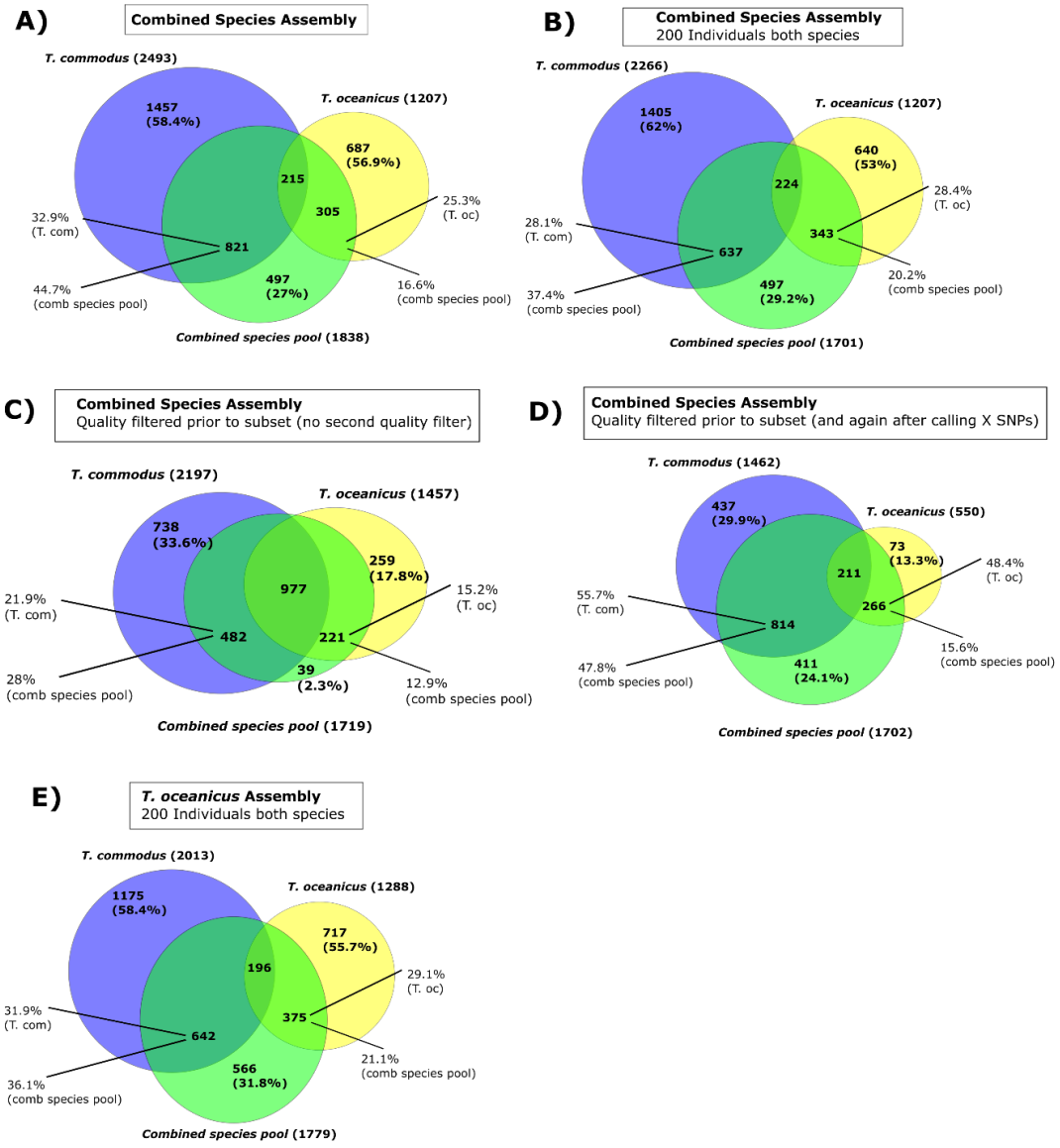


Figure S5 Overlap of putative X SNPs amongst: *T. commodus*, *T. oceanicus* and combined species pool. **A)** Combined species assembly - subset individuals into species specific groups, X SNPs called and then quality filtered (only sites with a minor allele frequency greater than or equal to 0.05 were kept and all loci that were not present in at least 80% of the individuals were excluded). **B)** Combined species assembly - randomly selected an equal mix of individuals of both species (200 per species), X SNPs called and quality filtered. **C)** Combined species assembly - quality filtered prior to subsetting individuals into species specific groups and then X SNPs called. **D)** Combined species assembly - quality filtered prior to subsetting individuals into species specific groups and then X SNPs called and quality filtered again. **E)** *T. oceanicus* assembly - subset individuals into species specific groups, X SNPs called and then quality filtered.

MOLECULAR ECOLOGY

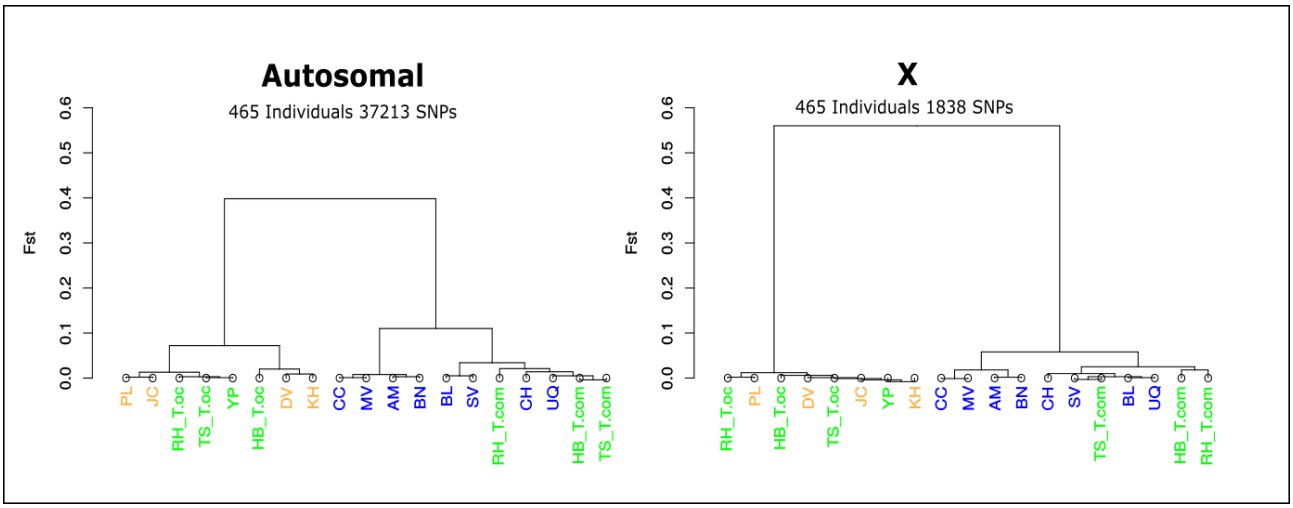


Figure S6 Population genetic structuring of populations at autosomal and X-linked loci. **Hierarchica**l clustering trees based on Pairwise F_{ST} at autosomal and X-linked markers. Populations are colour coded in accordance with Figure 1.

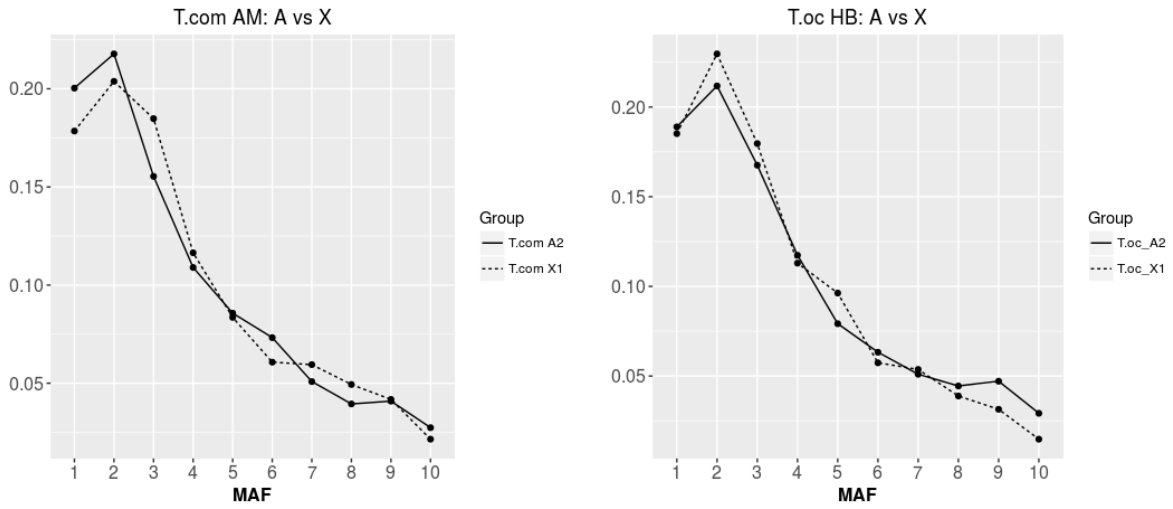


Figure S7 Folded site frequency spectrum (based on the minor allele frequency) for autosomal (A) (solid line) and X-linked (broken line) markers for populations of both species.

S.6 | Formulas

The within population component of genetic variation can be estimated as follows:
$W = 1 - F_{ST}$
For autosomal loci: $W_A = 1 - F_{ST}(_A)$
For X loci: $W_X = 1 - F_{ST}(_X)$
W_X can be predicted based on autosomal diversity (W_A) given a particular effective population size of males and females ($r = N_f / N_f + N_m$). When **r** = 0.5 the effective population size of males and females is equal and drift is assumed to proceed the same in both sexes. The formula below is from equation 7 in Ramachandran et al., (2004)

Suppl. Eq. 1. $$W_X = \frac{9W_A}{8(2-r) - W_A(7-8r)}$$

We applied Suppl. eq. (1) to our observed autosomal data to calculate the expected X F_{ST} given a range of different effective population sizes of males and females (r in eq.1). We used three r values encompassing the full range, (r = 0.5) an equal mix of males and females, (r = 0.9) an extreme female bias or (0.1) an extreme male bias. Two-tailed Wilcoxon sum rank tests were then used to test whether differences in the effective population sizes of males and females could account for discordance amongst our X and autosomal markers.
To account for potential sex differences in migration rates we applied equation 4 from Segurel et al., (2008) to our data. The effective number of females is given by Nf and N is the total effective numbers of males and females (N = Nf+Nm). (Nf/N) is similar to r in the formula proposed by Ramachandran et al., (2004)). The female-migration rate is mf. This model assumes an infinite island model.

Suppl. Eq. 2. $$F_{ST}(x) = \frac{4F_{ST}(A)}{4F_{ST}(A) - 3(F_{ST}(A)-1)\left(\frac{1+mF/m}{2-NF/N}\right)}$$

To estimate the expected diversity at X and autosomal (A) loci given a change in population size we applied equation 4 from Pool & Nielsen (2007).

Suppl. Eq. 3. $$\frac{\pi X}{\pi A} = \frac{h1}{h2}\frac{\mu 1}{\mu 2}\left(\frac{f-(f-1)(1-\frac{1}{2Nh1f})^g}{f-(f-1)(1-\frac{1}{2Nh2f})^g}\right)$$

The main parameters in the equation are: the inheritance factor (h) (h1 – X; h2 – A: under neutral assumptions of an equal number of breeding males and females: 0.75 and 1 respectively), the mutation rate (μ), the population size change factor (f), current population size (N) and the time in generations (g) ago that the size change occurred.