SCIENTIFIC REPORTS

OPEN

Genetic analysis of Indian tasar silkmoth (*Antheraea mylitta*) populations

Saikat Chakraborty[1,2,*], M Muthulakshmi[1,*], Deena Vardhini[1], P Jayaprakash[3], J Nagaraju[1,‡] & K. P. Arunkumar[1]

Indian tasar silkmoth, *Antheraea mylitta* is an economically important wild silkmoth species distributed across India. A number of morphologically and ethologically well-defined ecotypes are known for this species that differ in their primary food plant specificity. Most of these ecotypes do not interbreed in nature, but are able to produce offspring under captive conditions. Microsatellite markers were developed for *A. mylitta*, and out of these, ten well-behaved microsatellite loci were used to analyze the population structure of different ecoraces. A total of 154 individual moths belonging to eight different ecoraces, were screened at each locus. Hierarchical analysis of population structure using Analysis of MOlecular VAriance (AMOVA) revealed significant structuring ($F_{ST} = 0.154$) and considerable inbreeding ($F_{IS} = 0.505$). A significant isolation by distance was also observed. The number of possible population clusters was investigated using distance method, Bayesian algorithm and self organization maps (SOM). The first two methods revealed two distinct clusters, whereas the SOM showed the different ecoraces not to be clearly differentiated. These results suggest that although there is a large degree of phenotypic variation among the different ecoraces of *A. mylitta*, genetically they are not very different, and the phenotypic differences may largely be a result of their respective ecology.

Populations of several species are further classified by taxonomists into subspecies, races, demes, clines and so on, of which only cline and deme have non-arbitrary definitions[1]. Ecotypes consist of genetically distinct subsets of populations in a species, that are specialized to particular environments[2,3]. More generally, ecotypes can be defined as subsets of populations within a species that have different fundamental niches[4]. The Indian tasar silkworm, *Antheraea mylitta* is a natural fauna of tropical India, distributed in different geographical locations and habitats in this country. Possibly, because of the distinct ecological conditions prevailing in these different localities, several morphological variants, traditionally called ecoraces, have been identified in *A. mylitta*[5]. As high as 44 ecoraces are reported in this species, which feed primarily on *Terminalia* (Family: Combretaceae) species and *Shorea robusta* (Family: Dipterocarpaceae), and also on a number of secondary food plants[5]. The genus *Terminalia* and *Shorea* are quite far phylogenetically. Although both are Rosids, *Terminalia* belong to order Myrtales and *Shorea* to Malvales[6]. The ecoraces are uni, bi or trivoltine depending upon the geo-ecological conditions and differ from each other in several qualitative and quantitative traits[7], such as cocoon weight and colour, larval colour, and so on. Although most of these ecotypes do not interbreed in nature, some of them produce offspring when mated in captivity[8]. Tasar cocoons are reported to be the largest among all the silk-producing insects in the world[9]. The silk fibre produced by *A. mylitta* has its own distinctive colour, and is coarser than *Bombyx mori* silk. However, the tasar silk has higher tensile strength, elongation, and

[1]Centre of Excellence for Genetics and Genomics of Silkmoths, Laboratory of Molecular Genetics, Centre for DNA Fingerprinting and Diagnostics, Tuljaguda, Nampally, Hyderabad 500001, India. [2]Centre for Ecological Sciences, Indian Institute of Science, Bangalore 560012, India. [3]Muga silkworm seed organization and Eri silkworm seed organization, Central Silk Board, Guwahati 781006, India. *These authors contributed equally to this work. ‡Deceased. Correspondence and requests for materials should be addressed to K.P.A. (email: arun@cdfd.org.in)

stress-relaxation values than the silk secreted by the domesticated silkworm *B. mori.* These properties have made tasar silk as competent and desirable as *B. mori* silk[10,11].

Since, *A. mylitta* primarily inhabits forested habitats, it is expected that with the gradual depletion of forest cover due to surge in the human activities the habitat lost its continuity and resulted in geographic isolation. This isolation might have allowed the populations to continue separately for generations, resulting in different ecoraces. There is quite a bit of ambiguity in naming of these ecoraces. As the boundaries between the ecoraces are often fuzzy and the races do not obey the concept of static boundaries, the racial identity seems to be arbitrary at times. Quite surprisingly, two crops of the same population are sometimes named as two different ecoraces[12].

There is a lack of well-characterized molecular markers to study *A. mylitta* ecoraces. Although few studies have been carried out with RFLP[13], RAPD[14], SCAR[15], ISSR[12,16] and other DNA markers[17], except for RFLP, the rest are dominant markers, and hence the estimation of allele frequencies are based on the assumption that the loci are in Hardy-Weinberg equilibrium. Also, barring a couple of reports on *Antheraea assama*[18–20], there are no reported studies that describe the detailed population genetics of saturniid silkmoths, using SSR markers.

The pronounced phenotypic and behavioral variation of *A. mylitta* ecoraces have made it difficult for researchers to identify ecorace specific phenotypic markers[8]. Therefore, there is a need to identify genetic markers for a specific phenotype to differentiate ecoraces. Are these ecoraces genetically distinct from each other? Do the ecoraces form structured population? Are any of these ecoraces in decline? These questions are of considerable importance to the biology of *A. mylitta*. With these points in mind, we developed 32 microsatellite markers and screened eight *A. mylitta* ecoraces collected from different geographical locations across India to obtain insights into the population genetics of these ecoraces (Fig. 1 and Table 1).

Results

Genetic variation. *Descriptive statistics.* Considerable variation was observed at all microsatellite loci. Except Bhandara, which was monomorphic at the locus Amysat013, all other populations showed polymorphism at all the 10 loci (Supplementary data 1). Number of alleles ranged from 1 (Locus: Amysat013, Population: Bhandara) to 15 (Locus: Amysat023, Population: Modal) across the ten microsatellite loci studied among the eight ecoraces. The average number of alleles per locus ranged between 2.5 ± 0.76SD (Amysat013) and 11.75 ± 2.05SD (Amysat023). The minimum average number of alleles for a population across all 10 loci was 3.7 ± 1.77SD for Sukinda and the maximum was 6.6 ± 4.24SD for Modal. Average observed heterozygosity per locus (H_o) taking all the populations together ranged from 0 (Amysat013) to 0.81 ± 0.12 (Amysat023). The minimum average H_o across all the loci was observed in the population Daba Trivoltine (0.25 ± 0.32) and maximum in Sukinda (0.37 ± 0.34).

HWE. Hardy-Weinberg exact tests were done for all the ten loci and inferences were drawn based on the sequential Bonferroni corrected α (Supplementary data 2, Table 2). No test could be performed for locus Amysat013 for the Bhandara population, as the locus was monomorphic in this population. Also GENEPOP did not return any result for the locus Amysat037 in case of the populations Modal and Daba Bivoltine. Closer inspection of the data revealed that in these two populations, the expected (H_e) and observed heterozygosities (H_o) for this locus are exactly the same. Generally, the more similar H_e and H_o are, the higher is the probability that the locus is in HWE. This is supported by the observations in all other loci. A similar exact test[21,22] was also performed in ARLEQUIN 3.0[23], where the p-value of the HWE test in both these cases was 1.0. Hence, these two scenarios were inferred with a p-value of 1. Six of the ten loci were in HWE in at least two populations. Amysat026 was in equilibrium in all the eight populations and Amysat001 was in equilibrium in seven populations. Both Amysat023 and Amysat037 were in equilibrium in six populations. Only Amysat013 was not in equilibrium in any of the populations. Seven out of ten loci were in equilibrium for the Sukinda population, and for Sarihana, only two of the loci were in equilibrium. Among other reasons, departures from HWE expectations may occur in a case where a locus contains null alleles. MICROCHECKER was used to evaluate the effect of null alleles on the results. Except four loci (Amysat001, Amysat023, Amysat026, Amysat037), all other loci were observed to have significant homozygote excess indicating the presence of possible null alleles.

LD. Exact test for genotypic disequilibrium was performed using Markov chain simulations to identify significant association between pairs of loci. Tests for each population as well as global tests were carried out. As there are 10 loci in all, there can be 45 possible pairs. However, due to lack of polymorphism for Amysat013 in the Bhandara population, only 36 tests could be performed for this population. Out of these 351 total comparisons, significant LD, following sequential Bonferroni correction, could be observed in only three cases, i.e. Loci Amysat023 & Amysat032, Amysat026 & Amysat032 and Amysat023 & Amysat026, all in Sukinda population. In case of the global test, following sequential Bonferroni correction, only three pairs of loci (Amysat023 & Amysat032, Amysat023 & Amysat026 and Amysat026 & Amysat032), out of 45 comparisons were found to be under LD. All other loci segregated independently of each other (Supplementary data 3).
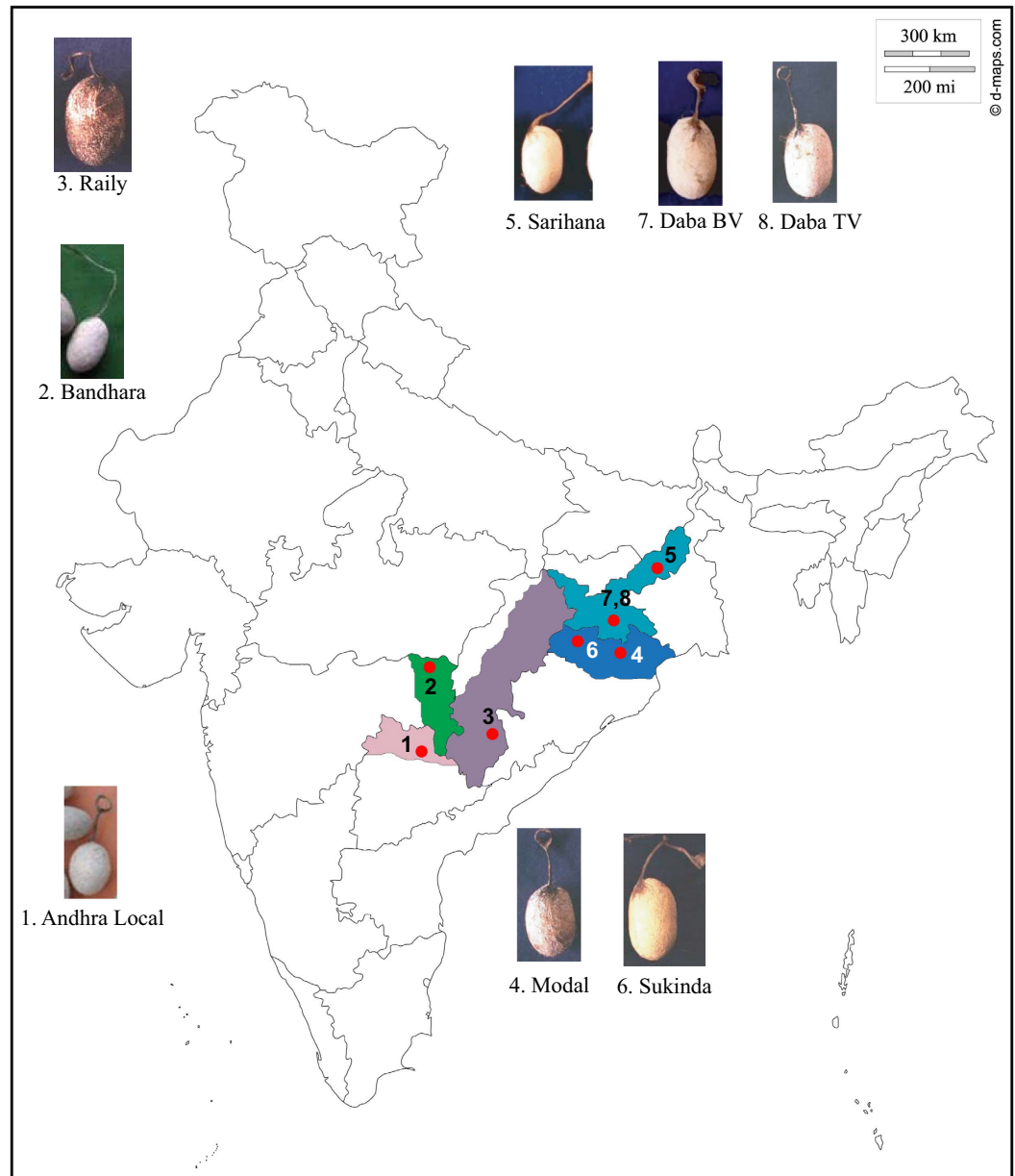
Figure 1. Spatial distribution of *A. mylitta* ecoraces used in this study. The map template was downloaded from d-maps.com[109] and the distribution of the different ecoraces were defined upon it using Adobe® Photoshop®. The photographs of the cocoons were taken by one of the authors, Dr. P. Jayaprakash, which were then included in the figure using Adobe® Illustrator®.

F-statistics and AMOVA. The locus specific fixation indices are shown in Table 2. Except Amysat026 and Amysat032, all other loci had a significant F_{ST} value. F_{IS} values were significant for all loci, except Amysat026. Result of the AMOVA, along with the fixation indices and their bootstrap confidence intervals are shown in Table 3. As the proportion of missing data was very less, the results of the haplotype as well as the locus by locus analysis were very similar. Here, results of only the locus by locus analysis are reported. Weighted average of the locus by locus F_{IS} value was considerably high (0.505), suggesting significant inbreeding within these populations. F_{ST} was also observed to be significant (0.154). In accordance with these values, 15.4% of the covariance was assigned to the across populations and 43% to the within population component. About 42% of the covariance could be assigned to the within individuals component.

Mantel test. The samples for Daba Bivoltine and Daba Trivoltine were collected from the same locality and hence geographic distance between them was zero. The maximum pairwise physical distance, i.e., 1034.4 km was between Sarihana and Andhra Local populations. Pairwise F_{ST} values ranged from 0.015

Serial No.	Ecotype	Collection site	State	Sampling coordinates	Primary food plant	Cocoon weight (g) (Mean ± SD)	Cocoon shell weight (g) (Mean ± SD)	Silk ratio (%) (Mean ± SD)	Average filament length (m) (Mean ± SD)
1	Andhra Local	Adilabad	Andhra Pradesh	19.66N/78.53E	*Terminalia. arjuna T. tomentosa*	9.45 ± 0.52	1.57 ± 0.714	16.61 ± 1.58	700 ± 36.62
2	Bandhara	Bhandara	Maharashtra	21.09N/79.42E	*T. arjuna T. tomentosa*	6.93 ± 0.61	1.24 ± 0.74	17.89 ± 1.39	613 ± 51.39
3	Raily	Jagdalpur	Chhattisgarh	19.08N/82.02E	*Shorea robusta*	12.77 ± 1.05	2.66 ± 0.79	20.83 ± 1.08	1232 ± 36.62
4	Modal	Keonjhar	Orissa	21.63N/85.58E	*S. robusta*	14.17 ± 0.88	3.64 ± 0.77	25.68 ± 0.68	1383 ± 66.36
5	Sarihana	Santhal Pargana	Jharkhand	24.26N/87.24E	*T. arjuna T. tomentosa*	9.75 ± 0.46	1.88 ± 0.749	19.28 ± 0.45	785 ± 47.05
6	Sukinda	Sukindagarh	Orissa	21.20N/85.52E	*T. arjuna T. tomentosa*	13 ± 0.84	1.80 ± 0.49	13.58 ± 0.75	845 ± 60.33
7	Daba Bivoltine	Singhbhum	Jharkhand	23.23N/85.23E	*T. arjuna T. tomentosa*	12.10 ± 0.86	2.80 ± 0.76	12.64 ± 0.99	750 ± 51.74
8	Daba Trivoltine	Singhbhum	Jharkhand	23.23N/85.23E	*T. tomentosa*	11.06 ± 0.54	1.28 ± 0.43	11.57 ± 0.68	700 ± 48.22

Table 1. *A. mylitta* ecoraces analyzed in this study.

Locus	Primer sequence	Allelic size range	F_{IS}	F_{ST}	F_{IT}
Amysat001	F-5′CGTACGTACTTTTGATGGATT 3′	161–231	0.256*	0.040*	0.286*
	R-5′AAGCGACATACCTGCGTT 3′				
Amysat013	F-5′TCACCACATGTCACTGACTGAA 3′	192–195	1*	0.315*	1*
	R-5′TCACTACAATCAGGCGCAAT 3′				
Amysat015	F-5′TACCAACAGACAGCCCTCCT3′	179–199	0.811*	0.204*	0.850*
	R-5′TCAAGGCTTTGACGTTGTATG3′				
Amysat019	F-5′CGCCATCTTGTTAGTTCTTCG 3′	241–264	0.632*	0.213*	0.711*
	R-5′TCGGGTGCTTTCCAAAGATA 3′				
Amysat021	F-5′CAAGATCGCGTTATCCTTTTT 3′	183–189	0.705*	0.196*	0.763*
	R-5′TGTTGTGAAGAACCCCCATT 3′				
Amysat023	F-5′TGCCCAGATAGTGTTTACCG 3′	106–175	0.088*	0.041*	0.125*
	R-5′GACCGGTCTGAACATAGTTGC 3′				
Amysat025	F-5′GTCTGGGGAAGCTGTTAAGAC 3′	228–250	0.739*	0.094*	0.763*
	R-5′CAATCGATATTTGAAACCGCTAA 3′				
Amysat026	F-5′CGCCACTTGCTCAGGAAT 3′	131–157	0.029	0.017	0.045
	R-5′CACGAGATGAGGTTGTGGTG 3′				
Amysat032	F-5′ACACCCCAAAATGTCAATG 3′	119–200	0.588*	0.024	0.597*
	R-5′ACCGGCACCATCACAT 3′				
Amysat037	F-5′CGCACGCGCACACT 3′	68–79	0.415*	0.576*	0.752*
	R-5′CCTATACACAAGCGAGTAGT 3′				

Table 2. **Microsatellite loci, primer sequences, allelic size range and F-statistics of 10 polymorphic loci in *A. mylitta*. Significant fixation indices are indicated by*.**

(between Sarihana and Daba Bivoltine populations) to 0.339 (between Bhandara and Sukinda populations). All pairwise genetic distance values were significant except the one between Sarihana and Daba Bivoltine populations ($F_{ST} = 0.015$, p = 0.44) (see Supplementary data 4). The Mantel test revealed significant positive correlation between genetic and physical distance r = 0.59, p = 0.026 (Fig. 2). The Mantel test between the residuals of a km-F_{ST} regression, however, was found not to be significant (r = 0.35, p = 0.055) suggesting a lack of migration-drift equilibrium (Fig. 3).

Bottleneck analysis.   In the BOTTLENECK analysis, none of the populations showed significant excess of simulated gene diversity compared to the observed gene diversity either under IAM or TPM, suggesting lack of recent bottleneck in any of these eight populations.

Source of variation	Sum of squares	Variance component	Percentage variation	Fixation index	Bootstrap CI (95%)
Among populations	177.13	0.55	15.41	$F_{ST} = 0.154^*$	(0.078–0.251)
Among individuals within populations	652.32	1.52	42.68	$F_{IS} = 0.505^*$	(0.299–0.708)
Within individuals	227	1.49	41.91	$F_{IT} = 0.581^*$	(0.376–0.769)
Total	1056.45	3.55			

Table 3. Result for locus by locus Global AMOVA as a weighted average over all loci. The estimates of the respective fixation indices and their bootstrap intervals are also shown. Significant estimates are indicated by*.
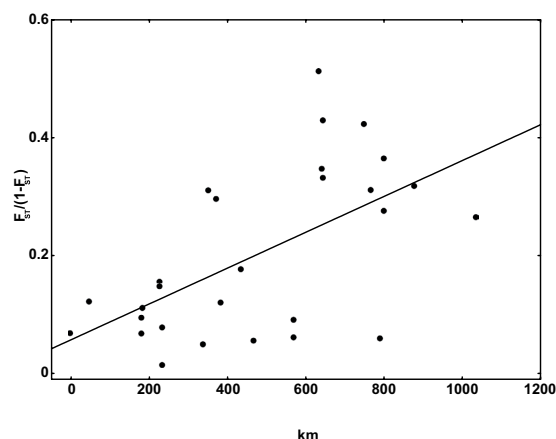

Figure 2. Relation between physical distance and genetic distance. The regression line represents the equation $F_{ST}/(1\ F_{ST}) = 0.0003^*km + 0.0571$, p < 0.001; $R^2 = 0.35$. Mantel test r = 0.59, p = 0.026.
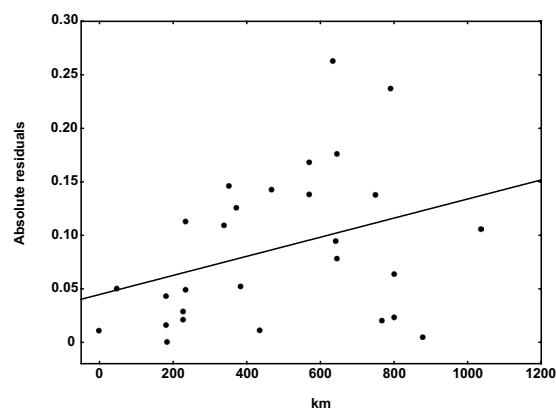

Figure 3. Relation between the residuals of the regression in Fig. 2 and physical distance. The regression line represents the equation $Residuals = 8.9251E\text{-}5\ ^*km + 0.0447$, p = 0.07; $R^2 = 0.12$. Mantel test r = 0.35, p = 0.055.

Population analysis using distance methods. The NJ tree constructed based on Nei *et al.*'s D_A distance, showed two major clusters (Figure S1), one with Andhra Local and Bhandara populations, and another included remaining six ecoraces (Raily, Modal, Sukinda, Sarihana, Daba Bivoltine and Daba Trivoltine). NJ tree did not show any specific relation among these six ecoraces. Although the two Daba ecoraces clustered together, the bootstrap value was found to be too less to make any strong conclusion.

Bayesian analysis of probable population clusters. The results of STRUCTURE analysis were very similar to that of the phylogenetic analysis. The modal value of ΔK was found to be at K = 2, suggesting that the data is best explained consisting of only two populations. One group containing two ecoraces (Andhra
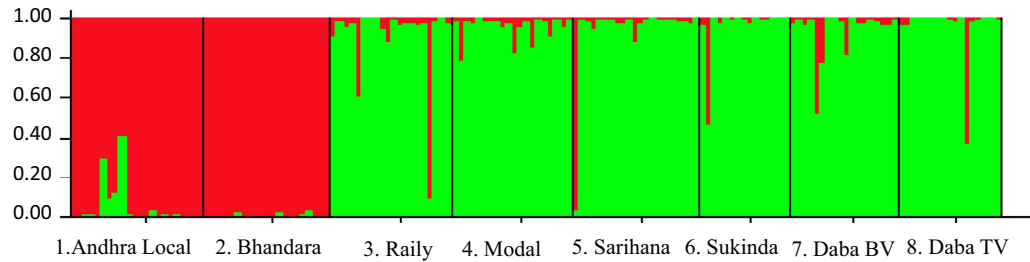
Figure 4. Population structure grouped by *A. mylitta* ecoraces using STRUCTURE at K = 2.
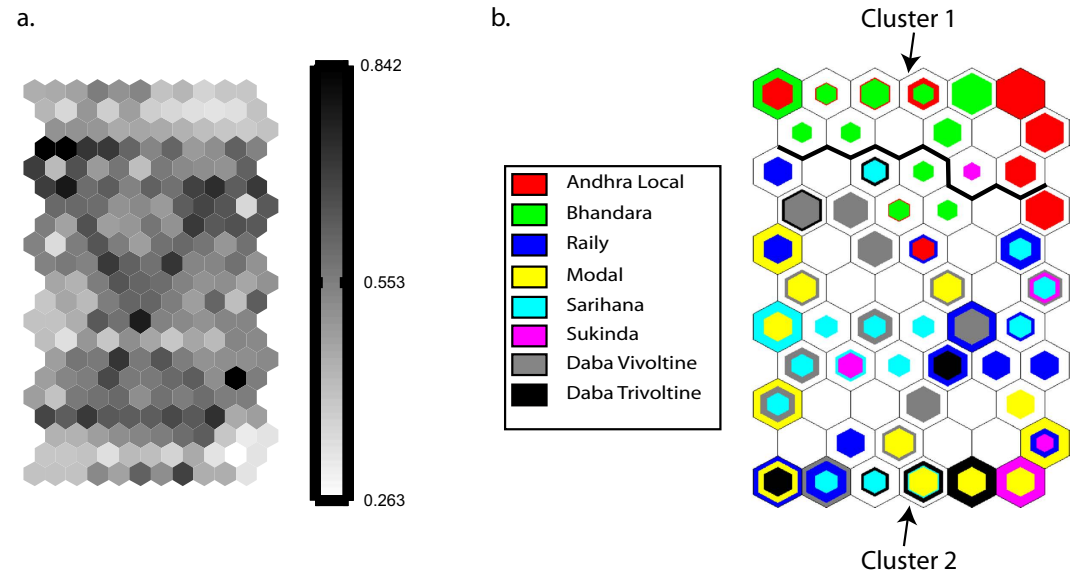

Figure 5. Self organization map. The u-matrix is shown in Fig. 5(a). The color bar represents distance between neighbouring neurons. The smaller the distance the closer are the neurons from each other. The respective individual labels on this map are shown in Figure S3. (b) shows the same map with different colour for each ecorace. The two main clusters obtained through cluster analysis are delineated with a solid black line.

Local and Bhandara) of southern and central India and the second group consisted of remaining six ecoraces (Raily, Modal, Sarihana, Sukinda, Daba Bivoltine and Daba Trivoltine) mainly from eastern India (Fig. 4).

Artificial neural network. The SOM output layer consisted of a 11 × 6 grid (Fig. 5). The map had a topographical error of 0. The quantization error was also reasonably small (1.792) suggesting the map quality to be good. The distances between neighbouring units on this map are shown in Fig. 6a. One can also visually inspect the level of clustering by comparing Fig. 6a and Figure S3. A better visualization can be obtained by looking at the hit histograms (Figure S2 in Supplementary file). From Fig. 5b, it appears that ecoraces Andhra Local (red) and Bhandara (green) are clustered together. However, this is a crude way of clustering the data as boundaries of each cluster are difficult to visualize in this way. Therefore, this SOM was further subjected to cluster analysis using Ward's linkage (minimum variance criterion). The resulting dendrogram is shown in Fig. 6 (A higher resolution version can be found in Supplementary file, Figure S4). The Davies-Boulding clustering index[24] indicated the best clustering to be 7 (Minimum value of the index 0.954 at level 7). The cluster boundaries are also indicated in Figure S3. If a cut-off is taken at squared Euclidean distance of 5, Group 1 and the rest forms two distinct clusters, Group 1 consist of the highest number of neurons among these 7 groups and is dominated by Andhra Local and Bhandara populations. However, these two ecoraces are not exclusive to this group and can be found in other groups as well.
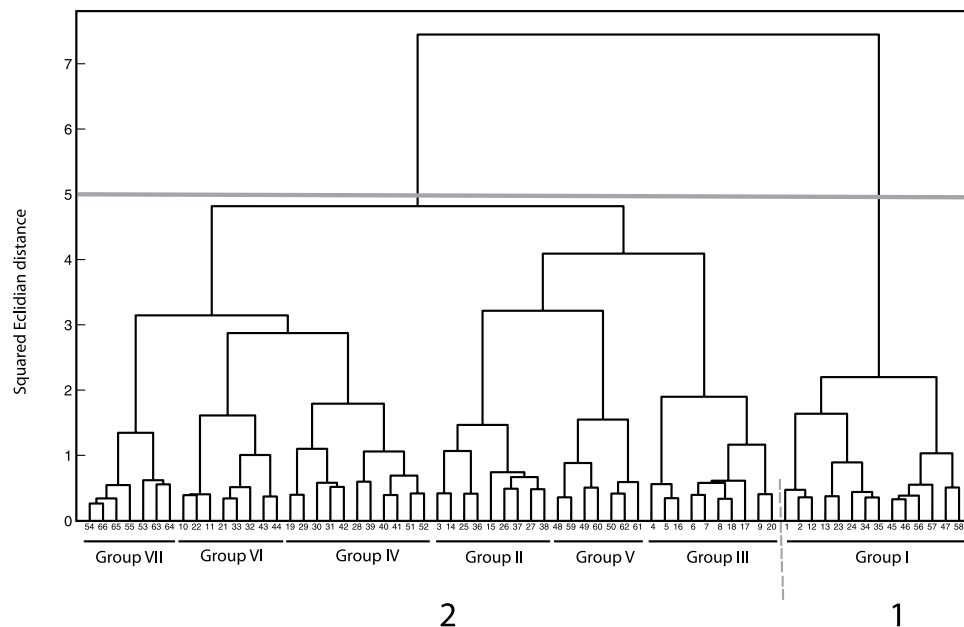
Figure 6. Dendrogram obtained following hierarchical cluster analysis of the SOM. The numbers on the tip of the trees are the same neuron numbers shown in Figure S3. Seven distinct clusters could be identified based on Davies-Boudin index. These have been labeled as Group I to VII. Moreover, taking a cut-off at squared Euclidian distance of 5 two main groups, 1 and 2 could be identified. A comparison of this dendrogram with Fig. 5b and Figure S3 would suggest that cluster 1 consisting only of Group I neurons is represented almost exclusively by individuals from Andhra Local and Daba ecoraces. (See text for more details).

Discussion

In this paper, we have characterized the microsatellites of *A. mylitta* and used them to study the genetic structure of its different ecoraces. Like most other species studied, we observed dinucleotide microsatellites to be the most abundant, followed by tri- and tetra nucleotide. Among the dinucleotide microsatellites, those with (CA) motifs were the most abundant. CA/GT repeats are generally the most common dinucleotide repeat in a wide variety of vertebrates and arthropods[25,26]. Insects[27,28], including lepidopterans[29–37] are no exception.

There have been studies investigating the nature and distribution of genetic variation in wild lepidopterans. However, these have mainly focused on threatened or declining species[38–40] or pests[41–44]. Among other silkworms the genetic structure of *A. assama*, a species with a very restricted distribution, has been studied and its populations were found to be reasonably differentiated[18,20]. Different strains and lines of the domesticated silkworm *B. mori* have also been subjected to this kind of analysis[32,45–47], however these studies are arguably not comparable to wild lepidopteran species.

A couple of studies using ISSR markers to analyze the intra-race diversity in Raily[16] and Daba[12] ecoraces of *A. mylitta* revealed considerable genetic differentiation across and within the populations of both the ecoraces. In another work, using RAPD markers, genetic variation among the different ecoraces of this moth were assessed[14]. However, none of the previous works did a detailed study of the genetic structure of the different *A. mylitta* ecoraces.

Here, we have studied the genetic structure of *A. mylitta* for a large part of its distribution. Eight ecoraces distributed as far as 1000 km apart were collected and their population genetics was investigated using ten polymorphic microsatellite loci. There was significant deviation from HWE in some of the loci in a few populations. This can be because of several reasons- 1) There was an overall excess of homozygotes than one would expect just by chance alone as reflected by the high locus specific F_{IS} values as well as its weighted average (0.505) across loci in the AMOVA (Table 3). As has been reported in the 'Methods' section, the sample cocoons were collected within about 500 m radius of forest area, and may consist of offspring of the same mother. This may be a possible reason of observing such a high F_{IS} value. 2) The sample sizes for each population was consistently low (Maximum N = 22, Minimum N = 15), and may not represent every genotype adequately, and 3) Presence of null alleles, as was found using MICROCHECKER, which in turn may result in the excess of homozygote in all the loci. Significant association was absent in almost all locus pairs in all the populations. Even globally, i.e., when all the populations were included together, there was significant association between only three of the possible 45 pairs of loci.

AMOVA revealed significant structuring of *A. mylitta* populations in the region under study ($F_{ST} = 0.154$). Although result of the Mantel test was significant ($r = 0.59$, $p = 0.026$), there was an absence of migration-drift equilibrium as reflected by the lack of significant correlation between the residuals of a km-F_{ST} regression and km ($r = 0.35$, $p = 0.055$). Excluding the localities with non-significant pairwise F_{ST} values, i.e., Daba Bivoltine and Sarihana, either sequentially or together, did not change this pattern, i.e., the km-F_{ST} correlation was consistently significant, but the km-residual Mantel was not significant. This suggests a lack of migration-drift equilibrium in the region studied. One may posit from the extent of scatter of individual points in Figs 2 and 3 that random genetic drift has been relatively more important than migration in shaping the current genetic structure of *A. mylitta*. However, as there are not too many points in the short distance range, this lack of equilibrium can also be an artifact of the sampling regime[48].

Both distance method and Bayesian analysis divided *A. mylitta* populations into two groups. One group consisted of one population each from southern and central India (Andhra Local and Bhandara). Another consisted of populations from eastern India. To investigate this pattern further, Andhra Local and Bhandara was taken as one group and the rest of the six populations were taken as another and an AMOVA was performed, giving 17% of the covariance assigned to the "Among groups" component and 5% to the "Among populations within groups" component, as compared to the overall "among populations" component of 15.41%, suggesting the possibility of substructuring. This is in full agreement with the STRUCTURE analysis. Remarkably, the two Daba ecoraces, though occurring sympatrically was found to be genetically distinct ($F_{ST} = 0.06452$). This is in line with the previous report[14], wherein they have used RAPD markers to differentiate 10 ecoraces of *A. mylitta*. In the NJ tree (Figure S1), except for the cluster consisting of Andhra Local and Bhandara ecoraces, all other bootstrap values were quite low (33 to 53). We further investigated this pattern using Kohonen maps. Although the Bayesian clustering algorithm and SOM have been shown to behave similarly for multilocus genotype data[49], in the present study, the results of the two methods were not exactly the same. A hit map of the eight ecoraces revealed samples from Andhra Local and Bhandara sharing the same or nearby neurons (Fig. 5b and Supplementary Figure S3). However in the hierarchical cluster analysis the pattern observed was not as clear as that of the phylogenetic and Bayesian analysis. Although many samples from Andhra Local and Bhandara ecoraces did cluster together [Cluster 1 (Group I) in Fig. 6], there were other clusters in which other samples from these ecoraces were distributed (Fig. 5b and Supplementary Figure S3).

Clustering of Andhra Local and Bhandara has also been shown previously using RAPD markers[14]. Andhra Local and Bhandara have adjacent distributions with no obvious geographic barrier between these two and the other six ecoraces (Fig. 1). This relationship shown in the phylogenetic tree is in line with its distribution geographically.

The *A. mylitta* ecoraces have considerably different niches. They have different host plants, are behaviourally different, have different cocoon colours and commercial characteristics[17]. The differences are so marked that they are popular with their unique local names in particular regions. Though there is a notion that each ecotype is specific to a particular region, overlapping distribution across the contact zones cannot be ruled out[50]. Experiments carried out to evaluate the effect of different food plants on cocoon characteristics revealed that food plants also have a considerable effect on these characteristics[51].

Though different morphologically identifiable ecoraces are found in different geographical regions of India, we speculate that these may be the result of a particular ecological condition prevailing in that region, and not because of their distinct genotype. Considering the lack of ecorace specific structure observed in this study, we opine that there may be less number of genetically distinct ecoraces than reported earlier. There seems to be a greater influence of environment such as host plants, climatic conditions, geographical locations etc., in shaping the phenotype of these ecoraces. Therefore, it is likely that most of these ecoraces are genetically similar but differ in certain morphology. *A. mylitta* is a highly polyphagous species, and displays a wide range of adaptation. Lepidopteran populations are known to be affected by habitat quality[52] and food plants[53]. Either or both of these factors could be responsible for the variation in phenotype observed among these ecoraces.

In this current work, we present a detailed genetic study of eight different ecoraces of a wild silkworm *A. mylitta*. However, one should keep in mind that we do not have replicates of the ecoraces in our dataset, i.e., each ecorace has been sampled only from a single locality and hence the effect of ecorace and locality might be confounded in the pattern we have observed.

Future work should concentrate on doing a more detailed behavioural and ecological study to understand the natural history of *A. mylitta* in greater detail. Cross feeding different ecoraces with the primary food plants of each other will help in confirming or ruling out the nature of host plant as a cause of phenotypic variation. Other ecological factors, similarly, can be tested, either sequentially or together, to get into the root cause of the origin of these ecoraces. Recent developments in statistical modeling such as the generalized linear model and the suit of different mixed models, allows one to carry out such investigations with great efficiency and cogency. Moreover, performing rigorous mating experiments may help us to understand the cause and nature of reproductive barrier among the different ecoraces. From the perspective of its genetics, identifying the hybrid zones of different ecoraces and studying the population structure including those regions may give a finer understanding of their population genetics. Development of additional microsatellite markers and employing a more extensive sampling regime

including samples at a finer distance scale would improve the resolution of the understood population genetic structure and, intra- and inter-race diversity of this species.

Methods

Collection of samples of different ecoraces. *A. mylitta* is an exceptionally versatile species as its ecoraces vary, in terms of phenotype, voltinism, distribution and also in their choice of food plants, and some of the ecoraces do not freely interbreed[54]. For example, the ecoraces Andhra Local, Bhandara and Sarihana are notable for their smaller cocoons than the others. On the other hand, the ecoraces Raily and Modal produce silk with significantly longer filament length than the other ecoraces. Wild populations of Daba can be both bi and trivoltine in nature, while semi-domestic counterparts are exclusively either bivoltine or trivoltine. Ecoraces Modal and Raily are wild in nature and are found exclusively in natural habitats[12]. The cocoon characteristics, sampling locations and information on primary food plants of all the eight ecoraces (Fig. 1) reported in the presented study are given in Table 1.

Live pupae were collected from their natural habitats in eight geographic locations representing eight ecoraces (Table 1 and Fig. 1). At each sampling site, wild cocoons were collected from the respective host plants. All sampling at each site was done within ~500 m radius of forest area.

Construction of repeat enriched genomic library and sequencing. For the construction of genomic library, DNA was isolated from *A. mylitta* pupae using the method described earlier[55]. Genomic DNA was enriched using an oligonucleotide mix of $(ATT)_8$, $(GAGT)_2$, $(CA)_{10}$, $(GA)_{10}$, $(GATA)_{10}$, $(CAC)_7$, and $(AGC)_7$, following previously reported protocol[56]. In brief, DNA was digested with *Rsa*I and *Xmn*I restriction enzymes, ligated to double stranded superSNX linkers, hybridized with biotinylated microsatellite oligonucleotides and captured on streptavidin coated magnetic beads. Unhybridized DNA was washed away, and captured DNA was recovered by polymerase chain reaction (PCR) using the single stranded superSNX-F as primer. The PCR products were ligated into the pCR®4-TOPO® TA vector (Invitrogen) and transformed into the XL1-Blue competent cells. Colony PCR was performed to select the amplicons of size 500–1000 bp. In total, 155 of the 240 colonies screened contained inserts of size more than 500 bp, which were sequenced using M13 forward and reverse primers, on ABI Prism 3100 Genetic Analyzer (Applied Biosystems). Among these, 46 had non-redundant microsatellite repeats, out of which a majority were found to harbor dinucleotide repeat motifs (65.2%), followed by tri- (17.4%), tetra- (6.5%) and complex (10.9%) repeat motifs. Among dinucleotide repeat motifs CA was found to be more abundant (60%) followed by GA (30%) and AT (10%). No CG motif was observed. A total of 16 clones were eliminated because the microsatellite sequences did not have sufficient flanking regions and so primer sequences could not be designed for amplification. Therefore, primers were designed for the rest of the 30 genomic microsatellite loci. Out of which, eight loci either did not yield expected product size or produced multiple non-specific bands. After the initial PCR assays, 22 microsatellite loci could be successfully amplified.

EST SSRs. *A. mylitta* unigene EST sequences[57] were downloaded from WildSilkBase. Out of 720 non-redundant EST sequences 50 were found to contain microsatellite repeats and 11 loci had insufficient flanking sequences to design the primer pairs. A total of 14 loci were selected based on the length and type of repeat motifs (di, tri, tetra and penta) for primer synthesis.

PCR amplification and genotyping of SSRs. We used Primer3 program[58] to design primers flanking SSRs in the EST clusters and microsatellite enriched sequences. PCR was carried out in a Mastercycler Gradient (Eppendorf, Germany) in a 10 µl reaction containing 1X PCR buffer (MBI Fermentas), 100 µM dNTPs, 1.0 to 3.0 mM $MgCl_2$, 5 pmole of each primers, 10 ng of genomic DNA as template and 0.5 U *Taq* polymerase (MBI Fermentas). The PCR conditions were initially at 94 °C for 3 minutes as initial denaturation and 35 cycles of: 94 °C denaturation for 30 seconds, appropriate annealing temperature (established empirically) for 30 seconds and 72 °C extension for 45 seconds and 72 °C for 10 minutes as final extension. Of the 14 EST-SSRs and 30 genomic SSRs, 10 and 22 respectively, gave amplification of exact size (Remaining were with multiple bands) as assessed by 1.5% agarose gel electrophoresis (Supplementary Table S1).

To identify microsatellite loci that are polymorphic, we randomly selected two individuals from each of the eight ecoraces (Fig. 1 and Table 1) and screened all the 32 microsatellite markers (Table 1 in Supplementary file). Out of these, we finally selected 10 microsatellite marker loci based on polymorphism and absence of non-specific PCR amplifications. In all, 154 individuals from the eight ecoraces were genotyped using 10 microsatellite markers. For all the amplifications we used FAM-labeled fluorescent forward primer and unlabeled reverse primer. The PCR products were run on an ABI PRISM 3730 DNA analyzer, with Pop-7 as sieving matrix, HiDi-Formamide as single-stranded DNA stabilizer and GeneScan 500 ROX as a size standard. Subsequently, ABI PRISM GeneMapper software version 3.0 was used to size the alleles. Data on allele sizes in each individual was tabulated for genetic analysis.

Statistical analysis. *Descriptive statistics and Hardy-Weinberg equilibrium.* For each locus, number of alleles (A), number of alleles in each polymorphic locus (Ap) (polymorphism cut-off = 99%), expected

heterozygosity as Nei's unbiased estimates of genetic diversity (H_E)[59], observed heterozygosity (H_O) and Weir and Cockerham's[60] estimate of inbreeding coefficient (f) were assessed using Genetic Data Analysis 1.0[61]. Exact test for Hardy-Weinberg equilibrium (HWE) with the explicit alternative hypothesis of heterozygote deficiency[62,63] was performed for each locus in each population using GENEPOP 4.0.5.3[64,65]. Locus specific global tests taking all the populations together were also performed with the same software. All tests were performed with the default parameters of 1000 dememorization steps and 20 batches per locus, with 5000 iterations per batch. All inferences were drawn following sequential Bonferroni correction[66]. The probability of null allele occurrence was estimated using MICROCHECKER[67] in which null alleles were considered to occur at a locus if an overall significant excess of homozygotes is seen, distributed evenly across the homozygote classes.

Linkage disequilibrium. Exact test for genotypic composite linkage disequilibrium (LD)[68] for each locus pair in each population was performed in GENEPOP with the default settings of 1000 dememorization steps, 20 batches per test and 5000 iterations per batch[64,65]. The null hypothesis of an equilibrium was accepted or rejected following sequential Bonferroni correction.

Population differentiation and Mantel test. Pairwise F_{ST} values and permutation test to estimate their significance was performed in ARLEQUIN 3.5 using 10,000 permutations. Pairwise physical distance in kilometers using the GPS coordinates of the sampling sites was estimated using GEOGRAPHIC DISTANCE MATRIX GENERATOR[69]. Correlation between Slatkin's linearized estimate of F_{ST}[70] and genetic distance, i.e. isolation by distance (IBD), was tested with Mantel test[71] in Mantel Tester, which is the program zt, with a user friendly GUI[72]. Mantel test was inferred following 10,000 permutations. In addition, a Mantel test was done between the absolute residuals of a regression of $F_{ST}/(1- F_{ST})$ upon physical distance (km), and physical distance, to check for the existence of migration-drift equilibrium, one of the basic assumptions of IBD[48,73].

Bottleneck analysis. The BOTTLENECK program[74] was used to determine if any signal of past bottleneck could be detected. The program was run under infinite alleles model (IAM) and two-phase mutation model (TPM)[75] with probabilities of single step and multi step mutations to be 0.9 and 0.1 respectively[76]. When a population undergoes reduction in population size it consequently loses its genetic diversity. This is reflected in both loss in number of alleles as well as observed gene diversity. In a population that has experienced such a reduction recently, the former parameter is lost faster than the latter. Therefore in such a population one would expect observed gene diversity to be higher than expected gene diversity estimated from the number of alleles (k)[77,78]. BOTTLENECK is equipped with three different statistical methods to test for this, i.e., i) sign test, ii) standardized difference test and iii) Wilcoxon sign-rank test. The first suffers from very low statistical power, the second can only be performed if the number of loci is more than 20, the third is recommended for mot situations where there are reasonable numbers of markers and samples. Therefore Wilcoxon sign-rank test was used to compare the simulated gene diversity with the observed gene diversity.

Analysis of MOlecular Variance. Hierarchical population structure was evaluated with the AMOVA[60,68,79] using the software ARLEQUIN 3.0. The analysis was performed choosing two main hierarchical structures; 1) Taking all the eight populations together and including the individual level and 2) Taking the Andhra Local and Bhandara populations together in one group and the rest six populations in another group (Rationale for doing so is in the 'Discussion' section). As the frequency of missing data was very low (maximum frequency of missing data was 3% for Amysat025), both haplotype and locus by locus analyses were performed and their results compared. "Pairwise differences" was chosen as the distance method and 10,000 permutations were used to test for the significance of covariance components and fixation indices. ARLEQUIN also returned bootstrap percentile values for the fixation indices. More than 20,000 bootstraps were carried out for each fixation index.

Population analysis using distance methods. There are several methods of estimating distances among populations from microsatellite allele frequencies and construct a tree from it. There are traditional distance estimates such as Nei's standard genetic distance D_{ST}[80], Cavalli-Sforza and Edward's chord distance D_C[81] etc., which are designed for loci mutating following IAM. Along with these, there are measures of genetic distance that have been explicitly defined to estimate distances based on microsatellite data, such as Goldstein et al.'s $\delta\mu$[81,82] and Shriver et al.'s D_{SW}[83]. All these distance measures have their own advantages and disadvantages. However, it has been shown using both computer simulation[84] as well as with real data[85] that for constructing phylogenetic trees, the efficiency of $(\delta\mu)^2$ or D_{SW}, is not very good. Rather, it was shown that the probability of the traditional distance methods in returning the correct tree topology is more. In these studies, the measure D_A[86] was found to perform better than the traditional measures as well as the microsatellite specific measures. And it was found to be efficient with both classical markers as well as microsatellites. Therefore, D_A was selected as the distance measure to construct a dendrogram for the microsatellite data of these eight populations. Neighbour joining (NJ) tree with 10000 bootstrap resampling of the loci[87] was constructed using POPTREE2[88].

Bayesian analysis of probable population clusters. Phylogenetic clustering method using individual genotype data although widely used suffers from several limitations. The actual tree topology is dependent on the genetic distance used and obtaining a reliable confidence interval with a few number of loci is not easy[89]. Therefore, we examined the number of distinct groups into which individuals from these eight populations can be clustered using the Bayesian clustering program STRUCTURE 2.3.1[89–91]. STRUCTURE uses a model based clustering technique to estimate the number of potential genetic clusters into which the individuals in a sample can be grouped. Given a series of potential clusters (K), STRUCTURE estimates the likelihood of each K to the data, which is achieved by estimating the number of groups that are in HWE and linkage equilibrium (LE). The K at which the likelihood reaches a plateau is taken as the number of optimal cluster number for the dataset. STRUCTURE was run assuming admixture and correlated allele frequency models. The populations were assumed to be in LE. The sampling locations were taken as putative population origin for the individuals. For the MCMC runs 100000 burn in period, followed by 100000 replications were used as recommended in the STRUCTURE manual. Value of K was taken from 1 to 15 and for each K, 20 iterations were performed. The most probable K was inferred from the modal value of rate of change of the LnP(D) value between successive runs (ΔK) as suggested by Evanno *et al.*[92]. This "Evanno method" has been implemented in the web based program STRUCTURE HARVESTER and it was used to estimate ΔK for each value of K and identify the mode[93].

Artificial neural networks. We also analysed the data using unsupervised neural networks in the form self-organizing maps (SOM) or Kohonen maps[94,95]. This method allows one to explore and visualize a multidimensional dataset into a two dimensional grid. It also allows one to look for possible clusters in the data[96]. It has been shown in a recent study that under moderate differentiation, like in case of our data, classical clustering techniques like NJ, are not perfect in finding the correct number of clusters in the data. In such cases, Bayesian clustering techniques like STRUCTURE and SOM were more efficient[49]. SOM has been used extensively in ecology[97–103], but there have been very few instances of its application to genetic data[44,53,104].

The network consists of two layers, -i) the input layer containing as many nodes as there are markers in a dataset, each of which is connected to all the individual genotypes, and ii) the output layer, which contains the neurons arranged in a two dimensional grid of hexagonal lattice. Each of these neurons is represented by weight vectors (also known as the codebook vector), whose dimension is same as the number of nodes in the input.

During the training of the map, distance between each input vector and all the weight vectors are calculated, and the hexagon closest to the input vector is called its best matching unit (BMU). Following this, all the weight vectors are update so that the BMU moves closer to the input vector in the input layer. The topological neighbours of this neuron are also dragged along. This updating is continued in iterations. At the end of this process, individuals with similar multilocus genotypes will be placed closer on the SOM. This then can be visualized using different techniques. Detailed account of the theoretical background as well as the algorithm to make these maps can be found elsewhere[95,104,105].

Our data consisted of 131 different alleles across the 10 loci, which formed the number of nodes in our input layer. For each of these 131 markers an individual was coded as 1, 0.5, or 0, if it was either homozygote, heterozygote or lacks this marker respectively[103]. Individuals who had missing data for a particular locus were coded as 0 for all alleles in that locus. We used a linear initialization of the SOM and trained it using a batch algorithm. The training was done in two phases–i) rough training phase with large neighbourhood radii and ii) fine-tuning phase with smaller radii. We further subjected this SOM to hierarchical cluster analyses using an agglomerative clustering method[106]. These analyses were performed using the SOM toolbox for MATLAB[107] and SOMVIS[108].

References

1. Mayr, E. & Ashlock, P. D. *Principles of systematic zoology.* (McGraw-Hill, 1991).
2. Turesson, G. The genotypical response of the plant species to the habitat. *Hereditas* **3,** 211–350 (1922).
3. Turesson, G. The species and variety as ecological units. *Hereditas* **3,** 100–113 (1922).
4. Begon, M., Townsend, C. R. & Harper, J. L. *Ecology: from individuals to ecosystems.* (Blackwell, 2006).
5. Jolly, M. S., Sen, S. K. & Ahsan, M. M. *Tasar Culture.* (Ambika Publishers, 1974).
6. Wang, H. *et al.* Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proc. Natl. Acad. Sci. USA* **106,** 3853–3858 (2009).
7. Srivastava, A. K. *et al.* In *Advances in Indian Sericulture Research* (eds Dandin, S. B. Gupta, V. P.). pp. 387–394. (Central Sericultural Research and Training Institute, Mysore, India, 2002).
8. Singh, B. M. K. & Srivastava, A. K. In *Current Technology Seminar on Non-mulberry Sericulture* Base paper. 6, pp. 1–39. (CTR&TI, Ranchi, India, 1997).
9. Akai, H. Cocoon filament characters and post-cocoon technology. *Int. J. Wild Silkmoth Silk* **5,** 255–259 (2000).
10. Iizuka, E. Physical properties of silk thread from cocoons of various wild silkmoths including domestic silk moth. *Int. J. Wild Silkmoth Silk* **5,** 266–269 (2000).
11. Rajkhowa, R. Structure property correlation of non-mulberry and mulberry silk fibres. *Int. J. Wild Silkmoth Silk* **5,** 287–298 (2000).
12. Kar, P. K. *et al.* Genetic Variability and genetic strucuture of wild and semi-domestic populations of tasar silkworm (*Antheraea mylitta*) ecorace Daba as revealed through ISSR markers. *Genetica* **125,** 173–183 (2005).
13. Mahendran, B., Padhi, B., Ghosh, S. K. & Kundu, S. Genetic variation in ecoraces of tropical tasar silkworm, *Antheraea mylitta* D. using RFLF technique. *Curr. Sci.* **90,** 100 (2006).

14. Saha, M., Mahendran, B. & Kundu, S. C. Development of Random Amplified Polymorphic DNA markers for tropical tasar silkworm *Antheraea mylitta*. *J Econ. Entomol.* **101,** 1176–1182 (2008).

15. Saha, M. & Kundu, S. C. Molecular Identification of tropical tasar silkworm (*Antheraea mylitta*) ecoraces with RAPD and SCAR markers. *Biochem. Genet.* **44,** 75–88 (2006).

16. Srivastava, A. K., Kar, P. K., Sinha, R., Sinha, M. K. & Vijayaprakash, N. B. Assessment of genetic diversity in different populations of Raily ecorace of Indian tasar silkworm, *Antheraea mylitta* using ISSR marker. *Int. J. Indust. Entomol.* **19,** 249–253 (2009).

17. Mahendran, B., Acharya, C., Dash, R., Ghosh, S. K. & Kundu, S. C. Repetitive DNA in tropical tasar silkworm *Antheraea mylitta*. *Genetica* **370,** 51–57 (2006).

18. Arunkumar, K. P. *et al.* Genetic diversity and population structure of Indian golden silkmoth (*Antheraea assama*). *PLoS One* **7,** e43716, D:10.1371/journal.pone.0043716 (2012).

19. Arunkumar, K. P., Kifayathullah, L. & Nagaraju, J. Microsatellite markers for the Indian golden silkmoth, *Antheraea assama* (Saturniidae: Lepidoptera). *Mol. Ecol. Resour.* **9,** 268–270, doi: 10.1111/j.1755-0998.2008.02414.x (2009).

20. Singh, Y. T. *et al.* Genetic variation within native populations of endemic silkmoth *Antheraea assamensis* (Helfer) from Northeast India indicates need for *in situ* conservation. *PLoS One* **7,** e49972, doi: 10.1371/journal.pone.0049972 (2012).

21. Guo, S. W. & Thompson, E. A. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*, **48,** 361–372 (1992).

22. Levene, H. On a matching problem arising in genetics. *Ann. Math. Stat.* **20,** 91–94 (1949).

23. Excoffier, L. & Lischer, H. E. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.*, **10,** 564–567 (2010).

24. Davies, D. L. & Bouldin, D. W. A Cluster Separation Measure. *IEEE Trans. Pattern. Anal. Mach. Intell.* **1(2),** (1979).

25. Beckmann, J. S. & Weber, J. L. Survey of human and rat microsatellites. *Genomics* **12,** 627–631 (1992).

26. Tóth, G., Gáspári, Z. & Jurka, J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* **10,** 967–981 (2000).

27. Schug, M. D. *et al.* The mutation rates of di-, tri- and tetranucleotide repeats in *Drosophila melanogaster*. *Mol. Biol. Evo.l* **15,** 1751–1760 (1998).

28. Butcher, R., Hubbard, S. & Whitfield, W. Microsatellite frequency and size variation in the parthenogenetic parasitic wasp *Venturia canescens* (Gravenhorst) (Hymenoptera: Ichneumonidae). *Insect Mol. Biol.* **9,** 375–384 (2000).

29. Reddy, K. D., Abraham, E. & Nagaraju, J. Microsatellites in the silkworm, *Bombyx mori*: abundance, polymorphism, and strain characterization. *Genome* **42,** 1057–1065 (1999).

30. Luque, C., Legal, L., Staudter, H., Gers, C. & Wink, M. ISSR (inter simple sequence repeats) as genetic markers in Noctuids (Lepidoptera). *Hereditas* **136,** 251–253 (2002).

31. Kumar, L. S., Sawant, A. S., Gupta, V. S. & Ranjekar, P. K. Comparative analysis of genetic diversity among Indian populations of *Scirpophaga incertulas* by ISSR-PCR and RAPD-PCR. *Biochem. Genet.* **39,** 297–309 (2001).

32. Reddy, K. D., Nagaraju, J. & Abraham, E. G. Genetic characterization of the silkworm *Bombyx mori* by simple sequence repeat (SSR)-anchored PCR. *Heredity* **83,** 681–687 (1999).

33. Bogdanowicz, S., Mastro, V., Prasher, D. & Harrison, R. Microsatellite DNA variation among Asian and North American gypsy moths (Lepidoptera: Lymantriidae). *Ann. Entomol.Soc. Am.* **90,** 768–775 (1997).

34. Meglécz, E. *et al.* Microsatellite flanking region similarities among different loci within insect species. *Insect Mol. Biol.* **16,** 175–185 (2007).

35. Hundsdoerfer, A. K. & Wink, M. New source of genetic polymorphisms in Lepidoptera? *Z. Naturforsch. C* **60,** 618–624 (2005).

36. Meglécz, E. & Solignac, M. Microsatellite loci for *Parnassius mnemosyne* (Lepidoptera). *Hereditas* **128,** 179–180 (1998).

37. Palo, J., Varvio, S. L., Hansk, I. & VäinÖlä, R. Developing Microsatellite Markers for Insect Population Structure: Complex Variation in a Checkerspot Butterfly. *Hereditas* **123,** 295–300 (1995).

38. Keyghobadi, N., Roland, J. & Strobeck, C. Influence of landscape on the population genetic structure of the alpine butterfly *Parnassius smintheus* (Papilionidae). *Mol. Ecol.* **8,** 1481–1495 (1999).

39. Meglécz, E., Nève, G., Pecsenye, K. & Varga, Z. Genetic variations in space and time in *Parnassius mnemosyne* (L.) (Lepidoptera) populations in north-east Hungary: implications for conservation. *Biol. Conserv.* **89,** 251–259 (1999).

40. Packer, L. *et al.* Population biology of an endangered butterfly, *Lycaeides melissa samuelis* (Lepidoptera; Lycaenidae): genetic variation, gene flow, and taxonomic status. *Can. J. Zool.* **76,** 320–329 (1998).

41. Bourguet, D., Bethenod, M. T., Pasteur, N. & Viard, F. Gene flow in the European corn borer *Ostrinia nubilalis*: implications for the sustainability of transgenic insecticidal maize. *Proc. R. Soc. B* **267,** 117–122 (2000).

42. Zhou, X., Faktor, O., Applebaum, S. W. & Coll, M. Population structure of the pestiferous moth *Helicoverpa armigera* in the Eastern Mediterranean using RAPD analysis. *Heredity* **85,** 251–256 (2000).

43. Martel, C., Réjasse, A., Rousset, F., Bethenod, M. & Bourguet, D. Host-plant-associated genetic differentiation in Northern French populations of the European corn borer. *Heredity* **90,** 141–149 (2003).

44. Roux, O. *et al.* ISSR-PCR: Tool for discrimination and genetic structure analysis of *Plutella xylostella* populations native to different geographical areas. *Mol. Phylogenet. Evol.* **43,** 240–250 (2007).

45. Nagaraju, J., Reddy, K. D., Nagaraja, G. M. & Sethuraman, B. N. Comparison of multilocus RFLPs and PCR-based marker systems for genetic analysis of the silkworm, *Bombyx mori*. *Heredity* **86,** 588–597 (2001).

46. Furdui, E. M. *et al.* Genetic Characterization of *Bombyx mori* (Lepidoptera: Bombycidae) Breeding and Hybrid Lines With Different Geographic Origins. *J. Insect Sci.* **14(1)** doi: 10.1093/jisesa/ieu073 (2014).

47. Pereira, N. *et al.* Biological and molecular characterization of silkworm strains from the Brazilian germplasm bank of *Bombyx mori*. *Genet. Mol. Res.* **12,** 2138–2147 (2013).

48. Hutchison, D. W. & Templeton, A. R. Correlation of pairwise genetic and geographic distance measures: inferring the relative influences of gene flow and drift on the distribution of genetic variability. *Evolution* **53,** 1898–1914 (1999).

49. Peña-Malavera, A., Bruno, C., Fernandez, E. & Balzarini, M. Comparison of algorithms to infer genetic population structure from unlinked molecular markers. *Stat. Appl. Genet. Molec. Biol.* **13,** 391–402 (2014).

50. Pandey, S. K. *Silk Culture: A Biochemical Approach.* 254 (APH Publishing, 2005).

51. Kar, P. K., Srivastava, A. K. & Naqvu, A. H. Changes in voltinism in populations of *Antheraea mylitta*: Response to photoperiod. *Int. J. Wild Silkmoths* **5,** 176–178 (2000).

52. Clarke, R., Thomas, J., Elmes, G. & Hochberg, M. E. The effects of spatial patterns in habitat quality on community dynamics within a site. *Proc. R. Soc. B* **264,** 347–354 (1997).

53. Luque, C. *et al.* Apparent influences of host-plant distribution on the structure and the genetic variability of local populations of the Purple Clay (*Diarsia brunnea*). *Biochem. Syst. Ecol.* **37,** 6–15 (2009).

54. Arora, G. S. & Gupta, I. J. Taxonomic studies of some of the Indian non-mulberry silkmoths (Lepidoptera: Saturniidae: Saturniinae). *Memoirs Zool. Surv. India* **16,** 1–63 (1979).

55. Prasad, M. D. & Nagaraju, J. A comparative phylogenetic analysis of full-length mariner elements isolated from the Indian tasar silkmoth, *Antheraea mylitta* (Lepidoptera: saturniidae). *J. Bioscience* **28,** 443–453 (2003).

56. Glenn, T. C. & Schable, N. A. Isolating microsatellite DNA loci. *Method. Enzymol.* **395,** 202–222 (2005).
57. Arunkumar, K. P., Tomar, A., Daimon, T., Shimada, T. & Nagaraju, J. WildSilkbase: an EST database of wild silkmoths. *BMC Genomics* **9,** 338, doi: 10.1186/1471-2164-9-338 (2008).
58. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Method. Mol. Biol.* **132,** 365–386 (2000).
59. Nei, M. *Molecular Evolutionary Genetics* (Columbia University Press, New York, 1987).
60. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution* **38,** 1358–1370 (1984).
61. Lewis, P. O. & Zaykin, D. Genetic Data Analysis: Computer program for the analysis of allelic data. *Version 1.0 (d16c).* (2001).
62. Rousset, F. & Raymond, M. Testing heterozygote excess and deficiency. *Genetics* **140,** 1413–1419 (1995).
63. Raymond, M. & Rousset, F. An exact test for population differentiation. *Evolution* **49,** 1280–1283 (1995).
64. Rousset, F. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol. Ecol. Resour.* **8,** 103–106, (2008).
65. Raymond, M. & Rousset, F. Genepop (version 1.2), population genetics software for exact tests and ecumenicism. *J. Hered.* **86,** 248–249 (1995).
66. Rice, W. R. Analyzing tables of statistical tests. *Evolution* **43,** 223–225 (1989).
67. van Oosterhout, C., Hutchison, W. F., Wills, D. P. M. & Shipley, P. MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Mol. Ecol. Notes* **4,** 535–538 (2004).
68. Weir, B. S. *Genetic Data Analysis II: Methods for Discrete Population Genetic Data.* (Sinauer Associates, Inc., 1996).
69. Ersts, P. Geographic distance matrix generator (version 1.2. 3). *American Museum of Natural History, Center for Biodiversity and Conservation* (2012).
70. Slatkin, M. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139,** 457–462 (1995).
71. Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Res.* **27,** 209–220 (1967).
72. Bonnet, E. & Van De Peer, Y. zt: a software tool for simple and partial Mantel tests. *J. Stat. Software* **7,** 1–12 (2002).
73. Wright, S. Isolation by distance. *Genetics* **28,** 114 (1943).
74. Cornuet, J. M. & Luikart, G. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* **144,** 2001–2014 (1996).
75. Di Rienzo, A. *et al.* Mutational processes of simple-sequence repeat loci in human populations. *Proc. Natl. Acad. Sci. USA* **91,** 3166–3170 (1994).
76. Luikart, G., Allendorf, F., Piry, S. & Cornuet, J. Molecular genetic test identifies endangered populations. *Conserv. Biol.* **12,** 228–237 (1998).
77. Cornuet, J. M. & Luikart, G. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* **144,** 2001–2014 (1996).
78. Maruyama, T. & Fuerst, P. A. Population bottlenecks and nonequilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. *Genetics* **111,** 675–689 (1985).
79. Excoffier, L., Smouse, P. E. & Quattro, J. M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131,** 479–491 (1992).
80. Nei, M. Genetic disatance between populations. *Am. Nat.* **106,** 283–292 (1972).
81. Cavalli-Sforza, L. L. & Edwards, A. W. Phylogenetic analysis. Models and estimation procedures. *Am. J. Hum. Genet.* **19,** 233–257 (1967).
82. Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L. & Feldman, M. W. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc. Natl. Acad. Sci. USA* **92,** 6723–6727 (1995).
83. Shriver, M. D. *et al.* A novel measure of genetic distance for highly polymorphic tandem repeat loci. *Mol. Biol. Evol.* **12,** 914–920 (1995).
84. Takezaki, N. & Nei, M. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics* **144,** 389–399 (1996).
85. Takezaki, N. & Nei, M. Empirical tests of the reliability of phylogenetic trees constructed with microsatellite DNA. *Genetics* **178,** 385–392 (2008).
86. Nei, M., Tajima, F. & Tateno, Y. Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. *J. Mol. Evol.* **19,** 153–170 (1983).
87. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39,** 783–791 (1985).
88. Takezaki, N., Nei, M. & Tamura, K. POPTREE2: Software for constructing population trees from allele frequency data and computing other population statistics with Windows interface. *Mol. Biol. Evol.* **27,** 747–752, (2010).
89. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).
90. Hubisz, M., Falush, D., Stephens, M. & Pritchard, J. Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* **9,** 1322–1332 (2009).
91. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164,** 1567–1587 (2003).
92. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14,** 2611–20 (2005).
93. Earl, D. & vonHoldt, B. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4,** 359–361 (2012).
94. Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybernet.* **43,** 59–69 (1982).
95. Kohonen, T. *Self-Organizing Maps.* (Springer Berlin Heidelberg, 2001).
96. Kohonen, T. Essentials of the self-organizing map. *Neural Net.* **37,** 52–65 (2013).
97. Lek, S. & Guégan, J.-F. *Artificial neuronal networks: application to ecology and evolution.* (Springer, 2000).
98. Lek, S. *et al.* Application of neural networks to modelling nonlinear relationships in ecology. *Ecol. Model.* **90,** 39–52 (1996).
99. Giraudel, J. & Lek, S. A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. *Ecol. Model.* **146,** 329–339 (2001).
100. Park, Y. S., Chang, J., Lek, S., Cao, W. & Brosse, S. Conservation strategies for endemic fish species threatened by the Three Gorges Dam. *Conserv. Biol.* **17,** 1748–1758 (2003).
101. Park, Y.-S., Céréghino, R., Compin, A. & Lek, S. Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. *Ecol. Model.* **160,** 265–280 (2003).
102. Worner, S. & Gevrey, M. Modelling global insect pest species assemblages to determine risk of invasion. *J. Appl. Ecol.* **43,** 858–867 (2006).
103. Chon, T.-S., Park, Y. S., Moon, K. H. & Cha, E. Y. Patternizing communities by using an artificial neural network. *Ecol. Model.* **90,** 69–78 (1996).
104. Giraudel, J., Aurelle, D., Berrebi, P. & Lek, S. In *Artificial Neuronal Networks* (eds Lek, S. & Guégan, J. F.) Ch. 15, 187–202 (Springer, 2000).

105. Kohonen, T. The self-organizing map. *Neurocomputing* **21,** 1–6, (1998).
106. Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58,** 236–244 (1963).
107. Vesanto, J., Himberg, J., Alhoniemi, E. & Parhankangas, J. Self-organizing map in Matlab: the SOM Toolbox. in *Proceedings of the Matlab DSP conference, Espoo, Finland, November 16–17,* pp. 35–40 (1999).
108. Rauber, A. *Data mining with SOMVIS.* (2009) Available at: http://ifs.tuwien.ac.at/dm/somvis-matlab/index.html. (Accessed: 15th June 2015)
109. Dalet, D. d-maps.com (2007). Available at: http://d-maps.com/carte.php?num_car=24853&lang=en (Accessed: 5th October 2015).

Acknowledgements

Author Contributions

K.P.A. and J.N.: conceived and designed the project. S.C., M.M., D.V., P.J. and K.P.A.: performed the experiments and data analysis. K.P.A. and S.C. wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Chakraborty, S. *et al.* Genetic analysis of Indian tasar silkmoth (*Antheraea mylitta*) populations. *Sci. Rep.* **5**, 15728; doi: 10.1038/srep15728 (2015).