


RESEARCH ARTICLE

Open Access



Genotyping of whole genome amplified reduced representation libraries reveals a cryptic population of *Culicoides brevitarsis* in the Northern Territory, Australia

Maria G. Onyango^{1,2}, Nicola C. Aitken³, Cameron Jack⁴, Aaron Chuah⁴, James Oguya⁵, Appolinaire Djikeng^{5,6}, Steve Kemp⁵, Glenn A. Bellis^{7,8}, Adrian Nicholas⁹, Peter J. Walker¹ and Jean-Bernard Duchemin^{1*} 

Abstract

Background: The advent of genotyping by Next Generation Sequencing has enabled rapid discovery of thousands of single nucleotide polymorphism (SNP) markers and high throughput genotyping of large populations at an affordable cost. Genotyping by sequencing (GBS), a reduced representation library sequencing method, allows highly multiplexed sequencing of genomic subsets. This method has limitations for small organisms with low amounts of genomic DNA, such as the bluetongue virus (BTV) vectors, *Culicoides* midges.

Results: This study employed the GBS method to isolate SNP markers *de novo* from whole genome amplified *Culicoides brevitarsis* genomic DNA. The individuals were collected from regions representing two different Australian patterns of BTV strain distribution: the Northern Territory (NT) and the east coast. We isolated 8145 SNPs using GBS. Phylogenetic analysis conducted using the filtered 3263 SNPs revealed the presence of a distinct *C. brevitarsis* sub-population in the NT and this was confirmed by analysis of mitochondrial DNA. Two loci showed a very strong signal for selection and were unique to the NT population. Bayesian analysis with STRUCTURE indicated a possible two-population cluster.

Conclusions: The results suggest that genotyping vectors with high density markers in combination with biological and environmental data is useful. However, more extensive sampling over a wider spatial and temporal range is needed. The presence of sub-structure in populations and loci under natural selection indicates the need for further investigation of the role of vectors in shaping the two Australian systems of BTV transmission. The described workflow is transferable to genotyping of small, non-model organisms, including arthropod vectors of pathogens of economic and medical importance.

Keywords: *Culicoides brevitarsis*, GBS, SNPs, Bluetongue virus, Australia

Background

Advances in next generation sequencing technology have provided access to studies of whole genome variation. This technological jump has pushed the fields of population genetics and phylogeography to a higher level of marker density than was conceivable in the PCR era. Significant reduction of costs of whole genome

sequencing [1] has led to a predictable shift of gold standard methods towards whole genome sequencing approaches, either whole genome resequencing or *de novo* DNA sequencing. This trend has already been applied to research on arthropod vectors in which knowledge of population structure, dispersal and gene flow, especially in the framework of vector control, is of paramount importance. To date, most studies using whole genome sequencing methods have targeted mosquitoes which serve as malaria [2] or arbovirus [3] vectors. In contrast, despite their role as vectors for

* Correspondence: Jean-Bernard.Duchemin@csiro.au

¹CSIRO Health and Biosecurity, Australian Animal Health Laboratory, 5 Portalington Road, Geelong 3220, VIC, Australia
Full list of author information is available at the end of the article



arboviruses of veterinary and economical importance, and being widely spread in much of the world [4], research on *Culicoides* has not yet benefited from this technological jump. *Culicoides* genome size is about 200 Mb, similar to *Anopheles* mosquitoes [5].

However, as biting midges are very small in size (1–3 mm adult length) and only few species have been reared in the laboratory, genomics resources are rare for *Culicoides* species.

Culicoides brevitarsis was first described in Australia in 1917 [6] and it is known to have a wide geographical distribution across the Oriental and Australasian regions [7]. In Australia, *C. brevitarsis* appears to be the principal vector of both bluetongue virus (BTV) and Akabane virus which cause economically significant infections of livestock [8, 9]. Since the initial detection of BTV in Australia in 1977, 12 serotypes (1, 2, 3, 5, 7, 9, 12, 15, 16, 20, 21, 23) have been isolated from the central northern region of Australia, in the Northern Territory (NT) [10, 11]. In contrast, only four serotypes (1, 2, 15 and 21) have been detected in the east of the continent with serotypes 1 and 21 being detected since the 1980s along the east coast down to the New South Wales, while serotypes 2 and 15 have been detected only since 2010, and only in Queensland (Qld) (Fig. 1)

[10, 12, 13]. This suggests the existence of two Australian BTV systems of transmission (northern and eastern), including virus, vectors, potential hosts and environments, or episystems [14]. The reasons for these two distinct episystems in Australia are unknown.

The distribution of *C. brevitarsis* encompasses the northern part of Western Australia and the Northern Territory (NT), across the northern and east coastal regions of Queensland (Qld), to the narrow coastal strip of the northern half of New South Wales (NSW) [9]. Previous studies have shown that *C. brevitarsis* is able to migrate and maintain gene flow between distant populations, such as between Indonesia and Australia [15, 16]. Knowing if this vector species encounters barriers to its movement that could explain the BTV serotype distribution would be of great importance in understanding the risk of incursion of new serotypes into the eastern zone.

The Restriction Associated DNA sequencing (RADseq) technique, first described by Baird et al. in 2008 [17], is the most popular of reduced-representation library sequencing methods [18, 19]. It allows reduction of the complexity of genomes leading to deep sequence coverage of the fragments adjacent to the restriction site, subsequently leading to detection of SNPs [20]. It has several

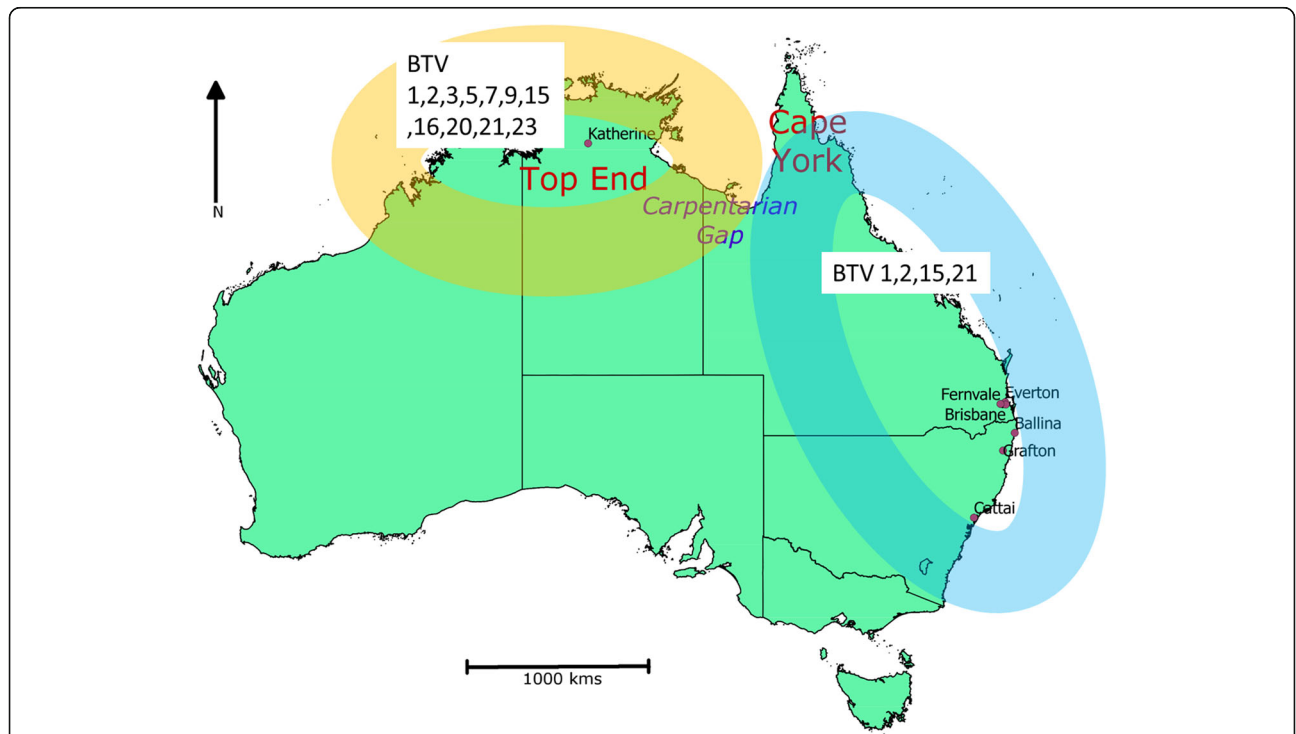


Fig. 1 A map showing the collection sites and the distribution of BTV serotypes in Australia. The distribution of the BTV serotypes depicts the two BTV epidemiological systems: one in the north of Northern Territory (pale orange) and the other in the eastern states of Queensland and New South Wales (pale blue). The administrative limits of the Australian states and territories are marked with a continuous black line. The BTV serotypes are tagged with their respective distribution zones. The sites of collection of *Culicoides* are indicated as purple dots. Note the Carpentarian Gap (in blue) located between the two tropical zones of Cape York (northern Queensland) and the Top End (northern NT), at the junction of the BTV serotype distribution zones

advantages over previous marker discovery tools such as restriction fragment length polymorphisms (RFLPs), amplified fragment length polymorphisms (AFLPs) and random amplified polymorphic DNA (RAPD) due to its ability to identify, verify and score markers concurrently. The RADseq technique is suitable for organisms without an existing reference genome [21]. It has been applied previously in phylogeography and population genetics of diverse organisms [19], including mosquitoes [2, 3, 22]. Among the techniques for detecting DNA sequence variation for conservation applications, and by extension to population genetics and phylogeography, RADSeq is one of the more economical [20]. Briefly, genomic DNA from several samples of interest is digested using a restriction enzyme of choice. Adapters containing barcodes that can identify uniquely the specific samples and an overhang corresponding to the restriction enzyme cut site are ligated to the digested fragments. The ligated fragments are sheared and bands in the size range of 300–700 bp are selected. These are subsequently amplified using polymerase chain reaction (PCR) before sequencing on an Illumina platform [23–25]. While the significance of the RADseq technique in population genetics studies cannot be overstated, a number of methods with simpler and cheaper library preparation steps have been described.

Genotyping by sequencing (GBS), another reduced representation library method [1], allows highly multiplexed sequencing of genomic subsets. The reduction strategy is similar to RADSeq, with the restriction site length defining the degree of reduction [23], and the potential use of double restriction for a higher reduction [26]. In the GBS method, the genomic DNA, the restriction enzyme of choice and adapters are added to the same well. The ligated fragments are amplified by PCR and subsequently sequenced on an Illumina platform. Compared to RADSeq, the shearing and size selection steps are eliminated in the GBS procedure, reducing the amount of hands-on time [6]. This method has been successfully used to study diversity in several species: black cottonwood (*Populus trichocarpa*) [27], wheat [10], a moth, the European corn borer (*Ostrinia nubilalis*) [28] and the bumble bee (*Bombus bifarius*) [29].

The hypothesis to be tested in the present study was that deeper sequence coverage and isolation of SNPs from *C. brevitarsis* DNA might lead to a better understanding of the gene flow pattern of this important BTV vector in Australia. Resolving this may help in understanding the role played by the migration and population structure of this vector in sustaining the epidemiological pattern of serotypes across the continent. Both RADseq and GBS have a significant limitation for organisms with scarce amounts of genomic DNA, such as *Culicoides* [4]. Pooling of individual samples limits SNP discovery to those with very high allele frequency in the general population and rare alleles are lost

[30]. The first objective of this study was to overcome the limitations posed by pooling of samples by using whole genome amplified (WGA) DNA of *C. brevitarsis* individuals to isolate SNP markers generated by GBS. The second objective was to apply these SNP markers for population genetic studies on *C. brevitarsis* collected from the two Australian BTV episystems. Because of its useful properties for taxonomic resolution and previous use for several *Culicoides* species [7, 31–33], sequence variation in the mitochondrial marker cytochrome oxidase subunit I (COI) was chosen for comparison with population structure detected using the GBS generated SNPs.

The present study describes a simple workflow that allows the isolation of genetic markers from samples of individuals with limited yields of genomic DNA and the further analysis of population genetics. This simplified workflow has great potential for application in ecology, evolution and conservation.

Methods

Insect sampling and DNA preparation

This study did not require any ethics approval because the midges were collected from animal sheds within farms and with the consent of the farm owners. Minimal contact with livestock animals took place without any disturbance.

A total of 96 midges was obtained from seven sites in Australia [NT (1 site), Qld (3 sites) and NSW (3 sites)] (Fig. 1) (Table 1). The midges were collected using green LED light traps [34] set 1 h prior to sunset and collected at approximately 0800 h the following morning. The collected samples were transported to the laboratory in 70 % ethanol. Specimens were identified to species level based on the wing pattern [7] observed using a binocular microscope. Species identification was verified using genetic methods as described in Bellis et al. [7, 32] to ensure isomorphic species of *Culicoides* were not included in analyses. Total genomic DNA was extracted from individual specimens using the DNeasy blood and tissue kit (Qiagen, Valencia, USA) according to the manufacturer's protocol. The genomic DNA was quantitated using a Qubit fluorometer using Qubit dsDNA HS Assay Kit (Life Technologies, Invitrogen, California, USA).

Table 1 A summary of the sites of collections of *Culicoides*

Region	Site	Year	<i>n</i>	Latitude	Longitude
NT	Katherine	2012	40	−14.24411782	132.4565168
Queensland	Fernvale	2013	4	−27.4559047	152.6532798
	Brisbane	2013	2	−27.4710107	153.0234489
	Everton park	2013	2	−27.4071662	152.9910401
NSW	Cattai	2013	21	−33.5599283	150.9074577
	Ballina	2012	21	−28.8684827	153.560001
	Grafton	2013	3	−29.6911226	152.9331993

Whole genome amplification of *C. brevitarsis*

To increase the genomic DNA yield from individual midges (<100 ng), multiple displacement amplification (MDA)-based WGA was conducted on each individual using the Repli-g ultrafast mini kit (Qiagen, Valencia, USA) according to the manufacturer's protocol [15, 33]. The amount of DNA used for each midge was about 1 ng. The resulting DNA was quantitated using a Qubit fluorometer and a Qubit dsDNA BR Assay Kit (Life Technologies, Invitrogen, California, USA) and visualised on 1 % agarose gel at 7.40 V/cm.

GBS library preparation

GBS libraries were constructed in 96-plex using custom adapters and barcodes. Approximately 500 ng of DNA generated by WGA was combined with 2.25 ng adapter that included a *Pst*I restriction overhang. The methodology for library construction was essentially that of Elshire et al. [1], except pairwise barcoding was used to enable multiplexing. Briefly, the whole genome amplified DNA of each individual was digested with *Pst*I (CTGCAG) for genome complexity reduction and ligated with one of 96 unique pairs of barcoded sequencing adaptors. The barcoded samples were then PCR amplified using MyTaq HS 2× Mix (Bioline) according to the manufacturer's specifications. Samples were individually quantitated and pooled in an equimolar manner. Library amplicons, 250–600 bp in length, were extracted and sequenced on an Illumina HiSeq2000 using a 100 bp Paired End protocol at the Biomolecular Resource Facility at the Australian National University, Canberra ACT.

Analysis of Illumina raw reads using the UNEAK GBS pipeline

The raw reads of the sequenced GBS libraries were analyzed using the Universal Network Enabled Analysis Kit (UNEAK) GBS pipeline which is designed for taxa without a reference genome and is part of the TASSEL 3.0 bioinformatics analysis package [35, 36]. In this method, a tag (haplotype) is considered to be a unique sequence representing a group of reads. The default parameters in the UNEAK pipeline were used.

Reads were retained if they contained a barcode, a cut site, and had no indeterminate bases within the first 64 bp after the barcode. From the raw reads, each sequence was trimmed to 64 bp in length. Reads were first merged into individual taxon tag count files and then merged into a 'master' tag to include all the tags from the same taxon, keeping tags with total reads count greater than or equal to 5 per sample. Globally, tag pairs that differed by a single nucleotide were retained as SNPs.

Post-UNEAK pipeline analysis

Isolated SNPs were analysed using R script [37] that we developed in this study (Additional file 1). The script filtered the *Culicoides* individuals with more than or equal to 5 % of the total SNPs isolated (8145 SNPs) and the SNPs present in more than or equal to 10 % of the total individuals (96 individuals). Once the SNPs and the samples selected, the script was used to calculate a genetic distance matrix and obtain a hierarchical cluster dendrogram. To limit the ordering bias of SNPs and samples in the data matrix, Pvcust [38] was used for assessing the uncertainty in hierarchical cluster analysis. *P* values are calculated through multiscale bootstrap resampling, by randomly shuffling the samples and/or SNPs a number of times (here 1000 times). Pvcust-generated dendrogram has two types of *p* values—AU (Approximately Unbiased) and BP (Bootstrap Probability) value. AU is computed by multiscale bootstrap resampling; this is obtained by looking at the changes of frequencies of fallen replicates along changing sample sizes, at each topography of the tree [39]. BP is computed by normal bootstrap resampling, with a constant sample size. In this study, the default setting of 1000 bootstraps was used.

To infer population structure from the SNP data, multilocus genetic distance estimates, with *F*_{st} [40, 41] were calculated between two population pairs corresponding to the BTV epistystems (NT and East Coast) using GenePop [42] and Arlequin [41]. Permutation tests (100 replications) were used to determine the significance of the population structure estimates.

Deviation from the Hardy-Weinberg equilibrium (HWE) was estimated by using GenAlex v6.502 [43]. The observed number of heterozygotes and homozygotes for each locus in each population was tested against expected values using a chi-square test.

Detection of loci under natural selection

To detect any loci under natural selection, BayeScan v2.1 [44] was applied. It defines two alternative models; one includes the effect of selection and the other excludes it. Bayesian inference utilises a likelihood function that results in a quantity called the posterior probability. The posterior probability of being under the effect of selection was estimated for each given locus using a reversible-jump Markov Chain Monte Carlo approach (MCMC) which simulates random processes. Its computational algorithm repeatedly samples randomly in order to obtain numerical results. This statistical method allows control against false positives. A *q*-value (the false discovery rate analogue of the *p*-value) of 0.1 is considered stringent. A Bayes factor of 32–100 corresponding to a posterior probability of 0.97–1.00 is considered strongly indicative of loci under natural selection.

Inferring population structure using a Bayesian model

To infer the presence of population structure, assign individuals to populations and identify admixed individuals, a Bayesian model-based approach in STRUCTURE v2.3 [45] was utilised. STRUCTURE assigns individuals a probability to belong to either a population. Selecting the 'Admixture model' option, allows the possibility for individuals, if their genotype is considered admixed, to be assigned to more than one population. A parameter set consisting of standard values of 100,000 'burnin' period and 100,000 MCMC Reps after burnin was used. Within STRUCTURE, 'Burnin' refers to the practice of discarding the initial portion of an MCMC run. To estimate the optimal number of populations (K) fitting the data, Genodive v2.0 [46] was used. Individuals were clustered using analysis of molecular variance (AMOVA), and checked against a range of 1–10 theoretical clusters (or populations). Clumpak [47] was used to collate all the results obtained from these Genodive iterations and import them in the STRUCTURE software.

Amplification and sequencing of standard mitochondrial gene *cytochrome c oxidase subunit I (COI)*

As described previously [32], primers *Bc1 Culic Fm* and *JerR2m* were used to PCR amplify a 692 bp segment of the *COI* gene from 76 *C. brevitarsis* individuals with accession numbers [GenBank: KX247448-KX247523]. A total of four previously identified and published haplotypes was added with the following accession numbers: [GenBank: KJ162968, KJ162967, KJ162975 and KJ162957]. These sequences included one of *C. asiatica*, the most closely related species to *C. brevitarsis* [7]. The PCR amplicons were purified using QIAquick PCR purification kit (Qiagen) and 20 μ l was sequenced using the Sanger sequencing method (Macrogen, Geumchun-gu, Seoul).

Phylogenetic analysis of mitochondrial DNA (*mtDNA*)

COI gene sequences were manually edited using Bioedit v7.1.9 [48] and aligned using MUSCLE [49]. The haplotype network was constructed in PopART [50] using the TCS Java program, estimating genealogy by calculating probabilities of DNA pairwise differences (95 % connection limit). DnaSP v5 [51] was used to estimate haplotype diversity. Pair-wise F_{ST} value estimates of genetic distances, Tajima's D and Fu's F_s tests of neutrality for the *COI* data per population were performed using Arlequin v3.5 [41].

Results

SNP marker isolation

A total of 16,102,542 reads was obtained from the 96-plex *PstI* library resulting in an average of 167,734 reads per sample. Following filtration, the final genotype matrix contained genotypes of 3263 SNPs across 75 samples (Additional file 2).

Intra- and interpopulation genetic differences

Hardy-Weinberg equilibrium tests conducted for the filtered SNPs in the two populations (NT and east coast) indicated that they deviated significantly from the HWE in 15 % of cases (975 of 6526 of for the two populations: $6526 = 2 \times 3263$ SNPs) ($P < 0.001$) (see Additional file 3). The F_{ST} values were significantly ($P < 0.05$) low ($F_{ST} = 0.01$) between NT and the east coast populations suggesting that the populations are genetically panmictic.

Phylogenetic relationship

The phylogenetic relationships of all populations were assessed using a dendrogram. A total of 10 individuals from NT samples clustered distinctly from all other individuals with high bootstrap support (>90 %). The remaining individuals from the NT clustered with the east coast individuals while all the few individuals from Qld clustered with the east coast samples (Fig. 2).

Loci under natural selection

Two loci (TP 616 and TP 2560) showed very strong evidence for selection [P ($\alpha \neq 0$) = 0.97; q-value (0.03); \log_{10} (PO) = 1.5] (Additional file 4). The two loci were detected among NT samples and were absent among east coast individuals. In the case of loci TP 616, 75 % of the individuals were found to be heterozygous while 30 % of the individuals were heterozygous at the TP 2560 locus.

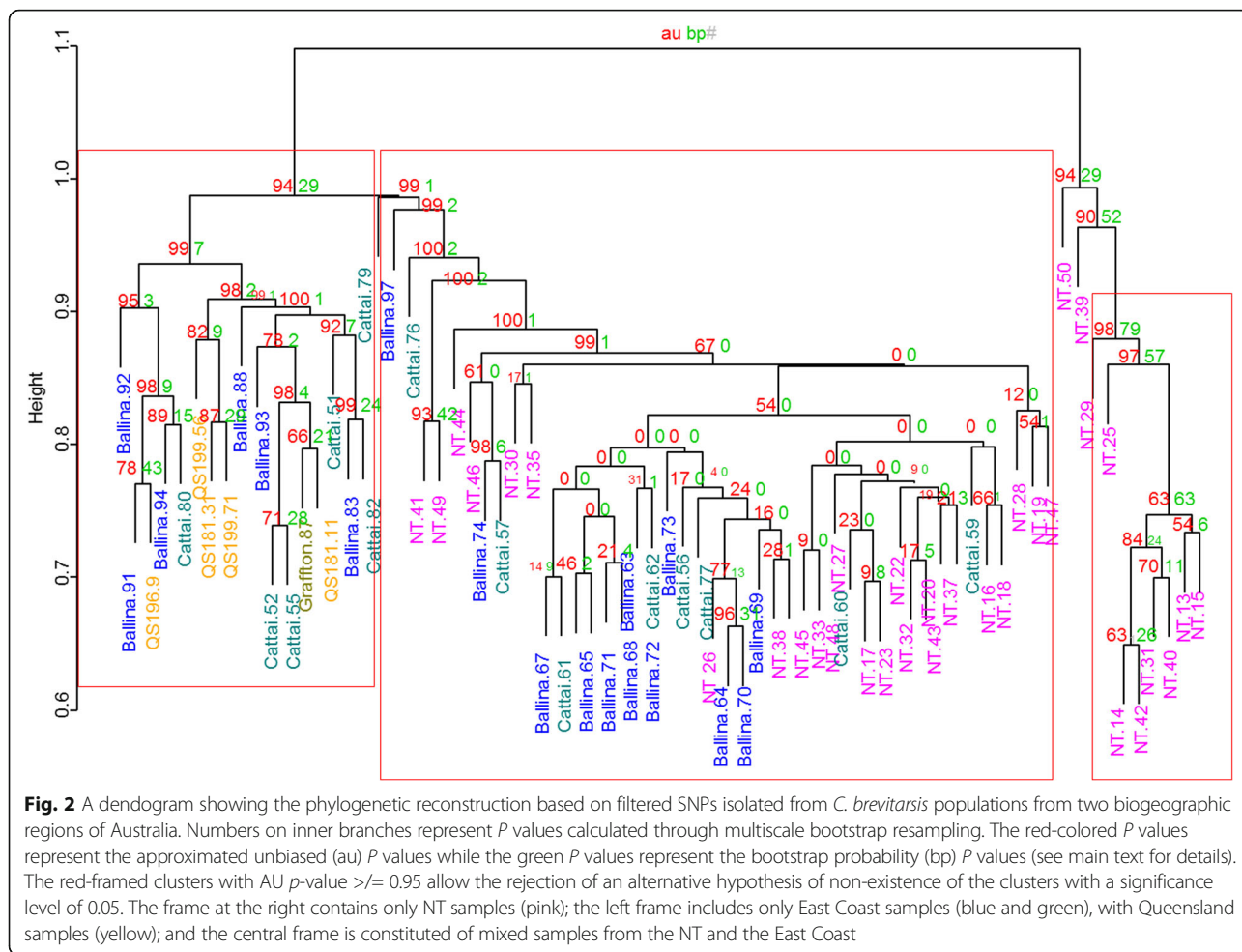
Population structure inference using a Bayesian modelling

The best clustering model of population number of the individuals was $K = 2$ (Fig. 3). The proportion of membership of each pre-defined population (NT and east coast) in each of the two clusters was: East coast (population 1 = 0.139; population 2 = 0.861) and NT (population 1 = 0.492; population 2 = 0.508).

Mitochondrial DNA haplotypes

A total of 9 haplotypes were identified (haplotype diversity $H_d = 0.4478$) (Fig. 4). All substitutions were synonymous and no stop codon was present. The first, most prevalent, haplotype (H1) consisted of 58 sequences mixed from the NT ($n = 21$), east coast ($n = 35$), East Timor ($n = 1$) and Solomon Island ($n = 1$) while the second most prevalent H2 consisted of 14 sequences almost exclusively from the NT ($n = 13$), and Cattai ($n = 1$). The rest of the haplotypes had 1 sequence each except H6 that had 2 sequences from East Timor and NT. H9 was an outgroup sequence from the closely related species *C. asiatica*.

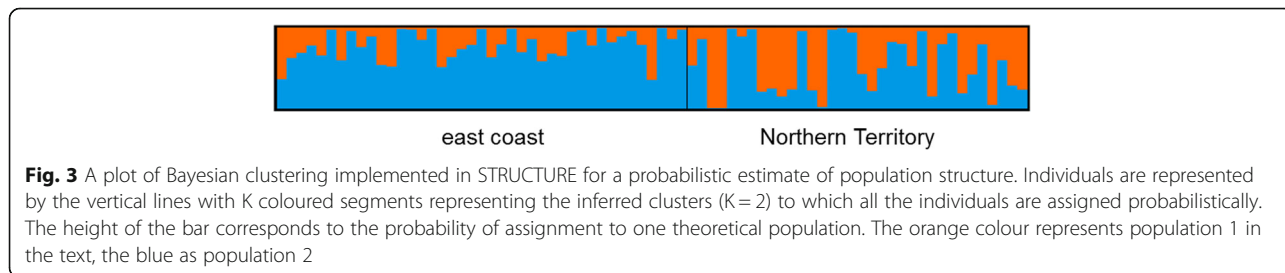
To detect selection, tests of neutrality were estimated and gave significant negative values ((Tajima's $D = -2.66699$) (Fu's F test = -6.8522]) $P < 0.02$).

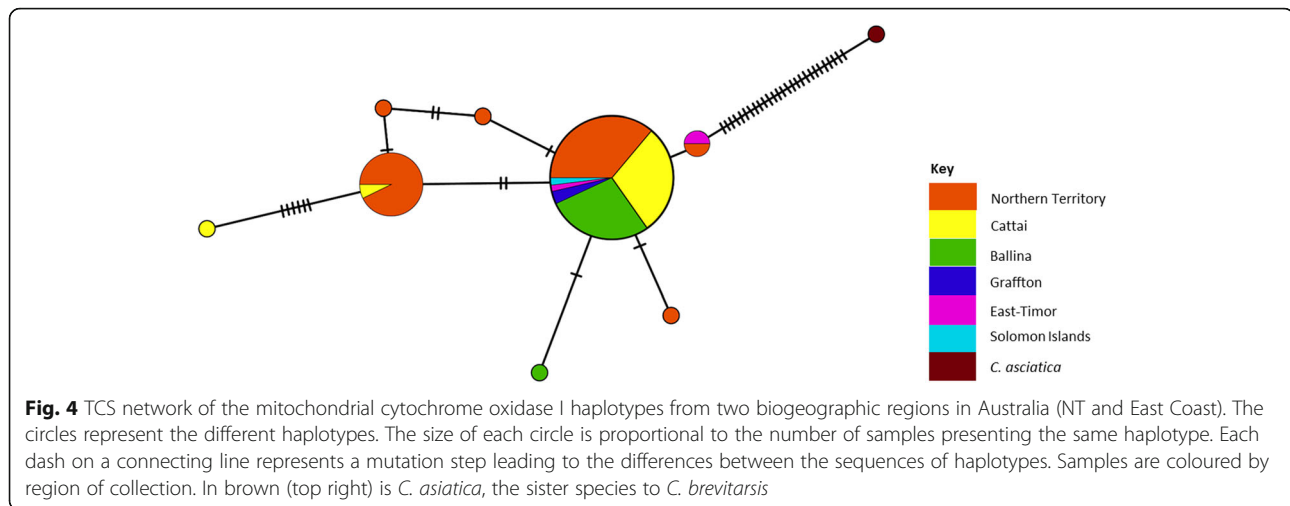


Discussion

The present study utilised GBS to investigate the phylogenetic relationships among *C. brevitarsis* populations across two biogeographic regions in Australia. The study aimed at providing clues to the geographic basis of the two epistystems. Genotyping by sequencing, as a reduced representation library method [1], allows highly multiplexed sequencing of genomic subsets. The increase in sampling depth of the genomes (hundreds to thousands of loci) provides a powerful tool for studies of evolutionary, demographic and adaptive mechanisms at a population level [52]. Small sized insects like mosquitoes have

been studied by reduced-representation library sequencing techniques such as RADSeq, either by pool [22] or individually [2, 3]. Despite their power, the application of NGS methods can be hampered by the limited amount of genomic DNA. Blair et al. [53] successfully isolated SNPs using RADseq after whole genome amplification of grey mouse lemur DNA with no significant genomic bias due to the previous genomic enrichment. The present study has isolated SNP markers using the GBS method from WGA DNA and used the resultant SNPs to genotype individual arthropods from two BTV epistystems in Australia, in order to analyze their genetic





structure. The approach did not require a previous reference genome.

Specimens from the same locations analyzed by microsatellite markers in a previous study revealed a panmictic population [15]. However, phylogenetic analysis conducted using both the isolated SNP markers and *mtDNA* in this study has revealed a sub population of *C. brevitaris* in the NT. This sub-population was resolved as a separate cluster with the remainder of the NT samples gathering with samples from east coast. Using *mtDNA* (among other molecular markers) Tay et al. [54] demonstrated the existence of genetic discontinuity between the NT and eastern populations of *C. brevitaris*. The discrepancy in terms of population structure between results obtained by microsatellite markers and SNPs and *mtDNA* could be expected. The mutation rates of SNPs (10^{-8} – 10^{-9}) [55] and *mtDNA* (10^{-8}) [56] are lower than that of microsatellites (range of 10^{-3} – 10^{-5}) [28, 57]. These differences in mutation rates may reveal a difference in the targeted time scale. A higher mutation rate, as occurs for microsatellite markers, would detect more recent events than markers with slower mutation rates.

In this study, we focussed our sampling efforts on the two extreme biogeographic regions, NT and NSW of Australia, with limited sampling carried out in Qld. It is noticeable that, in the SNP analysis, all samples from Qld clustered with the NSW samples. The Gulf of Carpentaria separates Australia's Top End, encompassing the northernmost section of the NT, from Cape York, a large peninsula at the far North of Queensland (Fig. 1). It is a biological barrier for many organisms (for a review, see [58]) including plants, birds, mammals [18] and insects [59]. For many organisms, the Cape York populations are closer to the Eastern region populations than to the northern Top End, with several examples of reduced gene flow between these populations [58]. We

recommend further studies involving a wider sampling of *C. brevitaris* from Queensland, including Cape York. This would shed light on the Carpentarian Gap as geographical barrier blocking gene flow between Qld and NT populations of *C. brevitaris*.

The subpopulation of the NT could be indicative of founder effects. There is a possibility that some genes present in the founder population of the NT may not have been dispersed to the east coast population. Arrival of a unique population in the NT, possibly from Timor-Leste [15, 16] but not yet dispersed to eastern Australia, could also explain the existence of the separate NT sub-population. However, our sampling included only one site in the NT and one time point.

The F_{ST} estimate value obtained in this study suggests significant gene flow between the two populations. This shows that contact between the NT and the east coast populations still exists despite the presence of some alleles in the former population not assorted with the latter population. The presence of few loci showing strong evidence of selection could be due to the strong selection for local adaptation of alleles present in these loci. Further studies of the SNPs under selection are recommended. The negative Tajima's D value could be indicative of either an expansion or a positive selection resulting from acquisition of a favourable trait.

The STRUCTURE results suggest that the individuals in this study had admixed genotypes and the individuals derived their ancestry from the two population clusters inferred in an unequal manner. The NT population seemed to derive its ancestry almost equally from both clusters while the east coast population seemed to derive its ancestry mainly from population cluster two. This could indicate that the NT is the ancestral population within Australia.

This study demonstrates that a combination of GBS to sample genomes densely [60], without need of reference genome, and the low bias of multiple displacement-based

amplification [61] allow the use of small amounts of DNA (such as might be obtained from non-invasive sampling) for population genetics studies. The technical workflow described in this study is easily translatable to other species and will facilitate understanding of the distribution of pathogens spread by tiny vectors such as sandflies, fleas, lice, ticks, aphids, psyllids and mites. The method also has application to the study of any biological materials with limited amounts of genomic DNA.

Conclusion

Although a heterogeneous population of *C. brevitarsis* (as shown in this study) could be a factor contributing to the presence of two separate BTV episystems in Australia (northern and eastern), these results should be taken with caution since individuals from the NT were sampled on only one occasion from a single site. More extensive sampling over a wider spatial and temporal range, including the northern part of Queensland (Cape York) would provide better resolution of the genetic connectivity of midge populations in Australia and shed light on the potential role of the Carpentarian Gap as barrier to gene flow. The epidemiological significance of the unique NT population, with loci under selection, also warrants investigation for vector competence studies comparing populations of *C. brevitarsis* from the NT and from eastern Australia. The described workflow is transferable for genotyping of small, non-model organisms, including arthropod vectors of pathogens of economic and medical importance.

Additional files

Additional file 1: R script used to perform filtering of the isolated SNPs. This file has to be read with R. (R 14 kb)

Additional file 2: The Matrix of data containing the list of SNPs selected after filtration crossed with selected samples. The 'N' tags the individuals where alleles are missing from the sequence reads. This could either be as a result of lack of a read at this site or a poor quality sequence, which cannot be called. (TXT 2744 kb)

Additional file 3: Results of estimates of Hardy-Weinberg equilibrium tests conducted for each locus in the NT and East Coast populations ($P < 0.01$). Within each population, the monomorphic loci are not informative. (XLSX 299 kb)

Additional file 4: BayeScan results for analysis for SNPs under natural selection. The first column is the tag of the selected reads. For each SNP, 'prob' represents the posterior probability of departure from neutrality; PO (posterior odds) is the ratio of posterior probabilities and is indicative of how more likely the model with selection is compared to the neutral model. The q value estimates the statistical significance of the deviation. The three lowest values are highlighted in yellow. The q value was used for ordering of the rows in the table. (XLSX 150 kb)

Abbreviations

BTV: Bluetongue virus; GBS: Genotyping by sequencing; MDA: Multiple displacement amplification; NSW: New South Wales; NT: Northern territory; PCR: Polymerase chain reaction; Qld: Queensland; RADseq: Restriction-associated DNA sequencing; UNEAK: Universal network enabled analysis kit; WGA: Whole genome amplification

Acknowledgement

M.G. Onyango was supported by a scholarship from the Australian Department of Foreign Affairs (AusAID). We would like to express our gratitude to Peter Mee for providing *C. brevitarsis* samples from Queensland. We would also like to acknowledge the critical review and comments of an early version of this manuscript by Drs David Boyle, Cadhla Firth and Kim Blasdel. We would also like to extend our gratitude to Justin O Borevitz for contributing resources towards GBS sequencing, Norman Warthman for initiating the collaboration and Doreen Busigye for helping with map construction. Laboratory work at AAHL were partly performed thanks to the National Collaborative Research Infrastructure Strategy of Australia. Thanks to Dr Prasad Paradkar for final reading.

Availability of data and materials

The process of submitting the data set supporting the results of this article to GenBank has been initiated and the unique identifier information will be made available immediately upon release. Once accepted, the data will be publicly available at the CSIRO Data Portal.

Authors' contributions

MGO: Designed, conducted the experiments, analysed results and drafted the paper. NCA: Participated in the experimental design of the GBS experiment and advice on experimental plans. CJ: Participated in the analysis of GBS data. AC: Participated in the analysis of GBS data and advice in installation of UNEAK pipeline in the server. JO: Participated in the experimental design of the GBS experiment and in the installation of UNEAK pipeline at ILRI's server. AD: Participated in the study design. SK: Participated in the study design. GB: Did the sampling and speciation of the midges. AN: Did the sampling and speciation of the midges. PJW: Involved in the study design, participated in the writing of the paper. JBD: Involved in the study design, participated in the writing of the paper. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

No human or vertebrate animal materials were involved in this work. The midges were captured during regular survey trapping sessions in farming environments by experienced teams from State or Territory administrations (NSW Department of Primary Industry and Northern Australian Quarantine Strategy).

Author details

¹CSIRO Health and Biosecurity, Australian Animal Health Laboratory, 5 Portalington Road, Geelong 3220, VIC, Australia. ²School of Medicine, Deakin University, 75 Pigdons Road, Waurn Ponds 3216, VIC, Australia. ³Research School of Biology, Australian National University, Canberra ACT 2601, Australia. ⁴The John Curtin School of Medical Research, Australian National University, Canberra ACT 2601, Australia. ⁵International Livestock Research Institute (ILRI), P.O. Box 3070900100 Nairobi, Kenya. ⁶Biosciences eastern and central Africa—ILRI Hub (BecA-ILRI Hub), ILRI, PO Box 3070900100 Nairobi, Kenya. ⁷Northern Australia Quarantine Strategy, 1 Pederson Road, Marrara 0812, NT, Australia. ⁸Research Institute for the Environment and Livelihoods, Charles Darwin University, Darwin 0909, NT, Australia. ⁹NSW Department of Primary Industries, Biosecurity, 4 Marsden Park Road, Calala 2340, NSW, Australia.

Received: 3 May 2016 Accepted: 26 September 2016

Published online: 30 September 2016

References

1. Elshire R, Glaubitz J, Sun Q, Poland J, Kawamoto K, Buckler E, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*. 2011;6:1–10.
2. O'Loughlin S, Magesa S, Mbogo C, Mosha F, Midega J, Lomas S, et al. Genomic analyses of three malaria vectors reveals extensive shared polymorphism but contrasting population histories. *Mol Biol Evol*. 2014;31(4):889–902.

3. Rašić G, Filipović I, Weeks A, Hoffmann A. Genome-wide SNPs lead to strong signals of geographic structure and relatedness patterns in the major arbovirus vector, *Aedes aegypti*. *BMC Genomics*. 2014;15:275.
4. Mellor P, Boorman J, Baylis M. Culicoides biting midges: their role as arbovirus vectors. *Annu Rev Entomol*. 2000;45:307–40.
5. Nayduch D, Cohnstaedt L, Sasaki C, Lawson D, Kersey P, Fife M, et al. Studying *Culicoides* vectors of BTV in the post-genomic era: Resources, bottlenecks to progress and future directions. *Virus Res*. 2014;182:43–9.
6. Bishop A, Kirkland P, McKenzie H, Spohr L, Barchia I, Muller M. Distribution and seasonal movements of *Culicoides brevitarsis* Kieffer (Diptera: Ceratopogonidae) at the southern limits of its distribution in New South Wales and their correlation with arboviruses affecting livestock. *Aust J Entomol*. 1995;34:289–98.
7. Bellis G, Dyce A, Gopurenko D, Yanase T, Garros C, Labuschagne K, Mitchell A. Revision of the *Culicoides* (Avaritia) *Imicola* complex Khamala & Kettle (Diptera: Ceratopogonidae) from the Australasian region. *Zootaxa* 2014, 3768:401–427.
8. Standfast H, Dyce A, Muller M. Vectors of bluetongue virus in Australia. *Prog Clin Biol Res*. 1985;178:177–86.
9. St. George T, Baldock C, Bellis G, Bishop A, Cameron A, Doherty B, Ellis T, Gard G, Johnson S, Kirkland P et al. The History of Bluetongue, Akabane and Ephemeral Fever Viruses and Their Vectors in Australia 1975–1999. Canberra: Animal Health Australia; 2001. p 101.
10. Boyle D, Bulach D, Amos-Ritchie R, Adams M, Walker P, Weir R. Genomic sequences of Australian bluetongue virus prototype serotypes reveal global relationships and possible routes of entry into Australia. *J Virol*. 2012;86:6724–31.
11. St George T, Standfast H, Cybinski D, Dyce A, Muller M, Doherty R, et al. The isolation of a bluetongue virus from *Culicoides* collected in the Northern Territory of Australia. *Aust Vet J*. 1978;54:153–4.
12. Boyle D, Amos-Ritchie R, Broz I, Walker P, Melville L, Flanagan D, et al. Evolution of bluetongue virus serotype 1 in Northern Australia over 30 years. *J Virol*. 2014;88:13981–9.
13. Eagles D, Walker P, Zalucki M, Durr P. Modelling spatio-temporal patterns of long-distance *Culicoides* dispersal into northern Australia. *Prev Vet Med*. 2013;110:312–22.
14. Tabachnick W. Challenges in predicting climate and environmental effects on vector-borne disease epizootics in a changing world. *J Exp Biol*. 2010;213:946–54.
15. Onyango M, Beebe N, Gopurenko D, Bellis G, Nicholas A, Ogogo M, et al. Assessment of population genetic structure in the arbovirus vector midge, *Culicoides brevitarsis* (Diptera: Ceratopogonidae), using multi-locus DNA microsatellites. *Vet Res*. 2015;46:1–9.
16. Eagles D, Deveson T, Walker P, Zalucki M, Durr P. Evaluation of long-distance dispersal of *Culicoides* midges into northern Australia using a migration model. *Med Vet Entomol*. 2012;26:334–40.
17. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*. 2008;3:1–7.
18. Edwards S, Potter S, Schmitt C, Bragg J, Moritz C. Reticulation, divergence, and the phylogeography-phylogenetics continuum. *Proc Natl Acad Sci U S A*. 2016;113(29):8025–32.
19. Hohenlohe P, Amish S, Catchen J, Allendorf F, Luikart G. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Mol Ecol Resour*. 2011;11(Suppl):117–22.
20. Allendorf F, Hohenlohe P, Luikart G. Genomics and the future of conservation genetics. *Nat Rev Genet*. 2010;11(10):697–709.
21. Avise J, Arnold J, Ball R, Bermingham E, Lamb T, Neigel J, et al. Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annu Rev Ecol Syst*. 1987;18:489–522.
22. Emerson K, Merz C, Catchen J, Hohenlohe P, Cresko W, Bradshaw W, et al. Resolving postglacial phylogeography using high-throughput sequencing. *Proc Natl Acad Sci U S A*. 2010;107:16196–200.
23. Baird N, Etter P, Atwood T, Currey M, Shiver A, Lewis Z, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*. 2008;3:1–7.
24. Etter P, Bassham S, Hohenlohe P, Johnson E, Cresko W. SNP discovery and genotyping for evolutionary genetics using RAD sequencing. *Mol Methods Evol Genet*. 2011;772:157–78.
25. Davey J, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, Blaxter M. Special features of RAD Sequencing data: implications for genotyping. *Mol Ecol Resour*. 2013;22:3151–64.
26. Peterson B, Weber J, Kay E, Fisher H, Hoekstra H. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*. 2012;7:1–11.
27. Bishop A, Spohr L, Harris A, Collins D. Factors affecting the distribution of *Culicoides* spp. (Diptera: Ceratopogonidae) vectors of bluetongue virus (BTV) in Australia. *Austral Entomol*. 2015;54:385–401.
28. Strand M, Prolla T, Liskay R, Petes T. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature*. 1993;365:274–6.
29. Lozier J, Jackson J, Dillon M, Strange J. Population genomics of divergence among extreme and intermediate color forms in a polymorphic insect. *Ecol Evol*. 2016;6(4):1075–91.
30. Cutler D, Jensen J. To pool, or not to pool? *Genetics*. 2010;186:41–3.
31. Linto Y, Mordue Luntz A, Cruickshank R, Meiswinkel R, Mellor P, Dallas J. Phylogenetic analysis of the mitochondrial cytochrome oxidase subunit I gene of five species of the *Culicoides imicola* species complex. *Med Vet Entomol*. 2002;16(2):139–46.
32. Bellis G, Dyce A, Gopurenko D, Mitchell A. Revision of the *Immaculatus* group of *Culicoides latreillei* (Diptera: Ceratopogonidae) from the Australasian region with description of two new species. *Zootaxa* 2013, 3680:15–37.
33. Onyango M, Michuki G, Ogogo M, Venter G, Miranda M, Elissa N, et al. Delineation of the population genetic structure of *Culicoides imicola* in East and South Africa. *Parasit Vectors*. 2015;8:1–13.
34. Bishop A, Bellis G, McKenzie H, Spohr L, Worrall R, Harris A, et al. Light trapping of biting midges *Culicoides* spp. (Diptera: Ceratopogonidae) with green light-emitting diodes. *Aust J Entomol*. 2006;45:202–5.
35. Bradbury P, Zhang Z, Kroon D, Casstevens T, Ramdoss Y, Buckler E. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*. 2007;23:2633–5.
36. Lu F, Lipka A, Glaubitz J, Elshire R, Cherney J, Casler M, et al. Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. *PLoS Genet*. 2013;9:1–14.
37. Ihaka R, Gentleman R. R: A language for data analysis and graphics. *J Comput Graph Stat*. 1996;5:299–314.
38. Suzuki R, Shimodaira H. Pvcust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*. 2006;22:1540–2.
39. Shimodaira H. An approximately unbiased test of phylogenetic tree selection. *Syst Biol*. 2002;51(3):492–508.
40. Weir B, Cockerham C. Estimating F-Statistics for the Analysis of Population Structure. *Evol (N Y)*. 1984;38:1358–70.
41. Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evol bioinforma*. 2005;1:47–50.
42. Rousset F. GENEPOP'007: A complete re-implementation of the GENEPOP software for Windows and Linux. *Mol Ecol Resour*. 2008;8:103–6.
43. Peakall R, Smouse P. GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics*. 2012;28:2537–9.
44. Foll M, Gaggiotti O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics*. 2008;180:977–93.
45. Pritchard J, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000;155:945–59.
46. Meirmans P, Van Tienderen P. GENOTYPE and GENODIVE: Two programs for the analysis of genetic diversity of asexual organisms. *Mol Ecol Resour*. 2004;4:792–4.
47. Kopelman N, Mayzel J, Jakobsson M, Rosenberg N, Mayrose I. CLUMPAK: a program for identifying clustering modes and packaging population structure inferences across K. *Mol Ecol Resour*. 2015;15:1179–91.
48. Hall T. BioEdit: An important software for molecular biology. *GERF Bull Biosci*. 2011;2:60–1.
49. Edgar R. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792–7.
50. Leigh J, Bryant D. POPART: full-feature software for haplotype network construction. *Methods Ecol Evol*. 2015;6:1110–6.
51. Librado P, Rozas J. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*. 2009;25:1451–2.
52. Schilling M, Wolf P, Duffy A, Rai H, Rowe C, Richardson B, et al. Genotyping-by-Sequencing for *Populus* population genomics: An assessment of genome sampling patterns and filtering approaches. *PLoS One*. 2014;9:1–9.

53. Blair C, Campbell C, Yoder A. Assessing the utility of whole genome amplified DNA for next-generation molecular ecology. *Mol Ecol Resour.* 2015;15:1079–90.
54. Tay W, Kerr P, Jermiin L. Population genetic structure and potential incursion pathways of the bluetongue virus vector *Culicoides brevitarsis* (Diptera: Ceratopogonidae) in Australia. *PLoS One.* 2016;11:1–29.
55. Brumfield R, Beerli P, Nickerson D, Edwards SV. The utility of single nucleotide polymorphisms in inferences of population history. *Trends Ecol Evol.* 2003;18:249–56.
56. Haag-Liautard C, Coffey N, Houle D, Lynch M, Charlesworth B, Keightley P. Direct estimation of the mitochondrial DNA mutation rate in *Drosophila melanogaster*. *PLoS Biol.* 2008;6:1706–14.
57. Dallas J. Estimation of microsatellite mutation rates in recombinant inbred strains of mouse. *Mamm Genome.* 1992;3:452–6.
58. Bowman D, Brown G, Braby M, Brown J, Cook L, Crisp M, et al. Biogeography of the Australian monsoon tropics. *J Biogeogr.* 2010;37:201–16.
59. Braby M. Taxonomic status of *Delias aestiva smithersi* Daniels, 2012 (Lepidoptera: Pieridae) comb. nov. from the Gulf Country of northern Australia, with description of the female. *Rec Aust Mus.* 2012;66(5):241–6.
60. Narum S, Buerkle C, Davey J, Miller M, Hohenlohe P. Genotyping-by-sequencing in ecological and conservation genomics. *Mol Ecol Resour.* 2013;22:2841–7.
61. Dean F, Hosono S, Fang L, Wu X, Faruqi A, Bray-Ward P, et al. Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A.* 2002;99:5261–6.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

