Local signatures of founding populations confound examination of adaptive divergence in invasive populations.

Adam P. A. Cardilini[1†], Katarina C. Stuart[2†], Phillip Cassey[3], Mark F. Richardson[4,5], William Sherwin[2], Lee A. Rollins[1,2]*, Craig D.H. Sherman[4]*

[1]Deakin University, School of Life and Environmental Sciences, Waurn Ponds, VIC, 3216, Australia

[2]Evolution and Ecology Research Centre, UNSW Sydney, Sydney NSW 2051 AUSTRALIA

[3]Centre for Applied Conservation Science, and School of Biological Sciences, University of Adelaide, SA 5005, Australia

[4]Deakin University, Centre for Integrative Ecology, School of Life and Environmental Sciences, Waurn Ponds, VIC, 3216, Australia

[5]Deakin University, Genomics Centre, School of Life and Environmental Sciences, Waurn Ponds, VIC, 3216, Australia

[†]Joint first Authors

*Joint last authors

Corresponding Authors: Craig Sherman (craig.sherman@deakin.edu.au); Lee Ann Rollins (l.rollins@unsw.edu.au)

Abstract

A detailed understanding of population genetics in non-native populations helps us to identify drivers of successful introductions. However, separating adaptive change from local signatures of founding populations represents a conceptual and technical difficulty when dealing with invasive populations. The history of introductions, as well as the process of range expansion, can confound interpretation of putative adaption in response to a novel environment. Here, we investigate putative signals of selection in Australian populations of introduced common starlings, Sturnus vulgaris, by examining population wide Single Nucleotide Polymorphisms (SNPs) and identifying SNP outliers associated with environmental variables. We determine that geographic distance plays a strong role in the genetic structure of populations, and that this is most likely strongly influenced by genetic differences in the founding populations, as well as modern day population connectivity. Examining candidate SNPs under putative selection indicated that local adaption has likely occurred, however, strong patterns in genetic variation, likely from founding populations, were visible in SNPs that were strongly associated with environmental variables. When examining putative adaption in invasive populations, we encourage critical interpretation of signatures of selection. Even strongly associated loci and environmental variables, when examined closely, may contain distinct footprints of invasion history or invasion expansion gradients, confounding analysis of the history of selection in these populations.

Keywords: *Sturnus vulgaris*, common starling, invasive species, local adaption, population genetics

Introduction

Understanding the proximate molecular mechanisms of natural selection is a central goal in evolutionary biology. Invasive species have long presented an opportunity to elucidate evolutionary mechanisms underlying selection (Baker and Stebbins, 1965). Rapid adaption within invasive ranges is commonplace in invasive populations (Lee, 2002) and is proposed to be fundamental to the long-term success of many invasive populations (Dlugosch and Parker, 2008). The short evolutionary timescale over which this adaption must take place provides opportunity to characterise the ecological and evolutionary processes that contribute to successful establishment and spread (Lee, 2002). Understanding these proximate mechanisms enable us to predict how evolutionary adaptation will influence range distribution, or how populations (both native and invasive) may respond to environmental change (Bay et al., 2018).

Investigation into the genetics underlying evolutionary change is facilitated by recent cost-effective genomic approaches, which enable the characterisation of hundreds to thousands of polymorphic markers across the genome. These markers provide unprecedented power for determining taxonomic identities, sources of introduction, invasion pathways, and understand a range of demographic and adaptive processes shaping native populations (e.g. bottlenecks, population expansions and adaptation) (reviewed in (Sherman et al., 2016)). Further, this information helps us understand invasion dynamics (Tepolt et al., 2009) and may illuminate the adaptive response of species to a novel or changing environment (Riquet et al., 2013). Such genomic approaches make it possible to investigate the occurrence of patterns of adaptive variation within invasive ranges, shedding light on selective pressures that may have influenced the success of non-native species. Using these data, we can identify signatures of molecular evolution in non-native species (Ashburner et al., 2000, Franks and Munshi-South, 2014), and identify loci putatively under

selection or those that are significantly associated with environmental variables (Foll and Gaggiotti, 2008, Gunther and Coop, 2013, Loh et al., 2015) .

While these genomic tools provide unprecedented sensitivity and power for testing for signatures of selection, caution should be applied when interpreting results for invasive species, or populations that have recently undergone rapid expansion, as these can mimic signature of selection (Lotterhos and Whitlock, 2014). For example, human introduced populations often have discrete (but potentially numerous and repetitious) introductory sites, which will have variable numbers of founding individuals from potentially different source populations (Dlugosch and Parker, 2008). What results is an amalgamation of subpopulations that, depending on the age of the populations and degree of connectivity, are genetically differentiated geographically due to legacy genotypes. Invasive species will often experience a bottleneck during establishment, impacting genetic diversity indices such as Tajima's D (Tajima 1989) in ways that resemble selection. Also, neutral loci can show false signatures of selection, due to stochastic genetic drift during a bottleneck, especially if rapid range expansion causes repeated bottlenecks – called 'allelic surfing' (Riquet et al., 2013). Bottlenecks encourage stochastic evolution through processes such as genetic drift resulting in low genetic diversity, and during rapid range expansion allelic surfing can create signatures at neutral loci that resemble selection (Hoban et al., 2016, Lotterhos and Whitlock, 2015) (see review in Sherman et al. 2016). Characterising adaptive genetic variation in invasive populations therefore presents a conceptually and technically difficult challenge and as such, care must be taken when interpreting signals of selection, particularly for invasive species, and the backdrop of invasion history must be considered.

The common starling (*Sturnus vulgaris*) is a highly successful invasive species that has established populations around the world (Higgins et al., 2006). Starlings were introduced into Australia from the UK during the 1850's onwards at several locations, including

Melbourne (Victoria), Brisbane (Queensland), Adelaide (South Australia) and Hobart (Tasmania), with the total individuals introduced numbering in the hundreds (Higgins et al., 2006, Jenkins, 1977). They quickly spread to occupy a large range in south-eastern Australia (Long 1981) across habitats with significant environmental gradients (Woolnough et al., 2005). Previous genetic surveys based on microsatellite markers showed that starlings in Australia form four distinct genetic clusters, largely reflecting the geographical separation, and complex nature, of the historic introductions (Rollins et al., 2009). The genetic cluster representing the western-most portion of the starlings' range edge (Western Australia) contained novel genetic variants, and evidence of genetic drift, selection, and admixture (Rollins et al., 2011, Rollins et al., 2009). This stands in contrast to the starling populations of North America, also introduced in the late 19th Century, for which high dispersal and continental wide interbreeding appears to have led to the genetic homogenisation and minimal population substructure evident in microsatellite datasets (Cabe, 1999). Furthermore, the Australian invasion is characterised by lower genetic diversity than that of the native range (Rollins et al., 2011). Despite this, there is clear evidence of morphological variation in response to temperature and rainfall gradients across the invasive range in Australia (Cardilini et al., 2016). Given the wide range of environments that Australian starlings inhabit and the approximately 180 years since introduction, the invasive starling provides an excellent model for investigating the population dynamics and adaptive variation of a highly dispersive non-native species.

Here, we use a reduced representation sequencing approach to generate a SNP data set to assess population genomic structure and describe patterns of genome-wide and putatively adaptive genetic variation in starlings across the Australian invasive range. Specifically, we aim to i) determine neutral patterns of population structure and gene flow across Australia using genome-wide loci, and *ii*) identify putative loci under selection, particularly those associated with environmental variation across the invasive range. Lastly, (iii) we assess

whether these apparent signatures of selection could also have been caused by the demographic history of an evolutionarily young invasive populations.

Methods

Environmental data collection

For each collection locality, we extracted climatic variables from Bioclim data sets using the RASTER package in R (Hijmans, 2016). Climatic variables included: mean diurnal range (Bio02), isothermality (Bio03), temperature seasonality (Bio04), maximum temperature of the warmest month (Bio05), minimum temperature of the coldest month (Bio06), temperature annual range (Bio07), precipitation of the wettest month (Bio13), precipitation of the driest month (Bio14) and precipitation seasonality (Bio15). We extracted aridity index (AI) data from the CGIAR-CSI Global-Aridity and Global-PET Geospatial Database (Zorner et al., 2008). We downloaded monthly average normalised difference vegetation index (NDVI) values from the Australian Bureau of Meteorology for the period 1997-2011. We averaged monthly datasets across years (e.g. mean NDVI January 1997-2011) and then across all months to create a mean NDVI raster file from which we extracted mean NDVI (mnNDVI) for each sample. We calculated variability in day length (varDL) per sample using the 'daylength()' function in the R package GEOSPHERE (Hijmans et al., 2015). We defined the position of each collection locality as the central longitude and latitude of all samples collected from the same collection locality. Using an R function that we wrote, we extracted elevation data (Elev) for each sample collection locality from Google maps. We calculated the average of each environmental variable at each collection locality and used these for further analysis. We calculated geographic distance variables as Euclidean distance (km) of each locality to the nearest coastline and Euclidean distance of each locality to nearest known site of introduction, and then calculated pairwise distances as the greater circle distances between collection localities using the R package SP (Bivand et al., 2013).

To identify a reduced set of variables that explained environmental variation across the introduced range, we conducted a principle components analysis (PCA) including all environmental variables and run in R using the PRCOMP function (RCoreTeam, 2015) . We calculated the pairwise distance between points on a plot of PC1 and PC2 to calculate pairwise environmental distances between collection localities, which were subsequently used in the analysis of isolation by environment (Leydet et al., 2018, Wang et al., 2013).

Sample collection and sexing

We collected a total of 568 starling samples from 24 localities across the invasive range in Australia (Fig. 1). Sample collection occurred in two separate phases. First, we used a subset of pre-extracted DNA samples (n = 150) collected by Rollins *et al.* (2009), from localities across South Australia and Western Australia from 2003 – 2007. Second, we collected 418 samples between May 2011 and October 2012 from 18 locations across the south-eastern portion of the Australian range (Fig. 1). These birds were collected in three ways: *i)* by trapping, *ii)* as carcasses collected from 3rd party hunters or land owners that humanely killed birds on their property, or *iii)* chicks collected directly from nests. In the latter case, only one individual per nest (adult or chick) was included in the dataset to avoid the inclusion of closely related individuals. We recorded Global Position System (GPS) coordinates upon collection. When 3rd party hunters did not record GPS coordinates, the closest corresponding address was recorded and GPS coordinates were identified using Google maps. We took a muscle sample from the thigh or breast and stored it in 70% ethanol prior to DNA extraction. We extracted DNA with a Gentra PureGene Tissue Kit, following the manufacturer's protocol. All samples were sexed using molecular methods (Griffiths et al., 1998).

Library construction, sequencing and SNP calling

The Cornell University Institute of Biotechnology Genomics Facility conducted library construction and Genotyping-by-Sequencing (GBS) (Elshire et al., 2011), using the restriction enzyme *PstI*. Each individual received a unique barcode before multiplexing 96 individuals per lane (Illumina HiSeq 2000; 6 size lanes, 100 bp, singled-end reads).

We processed raw sequence data using TASSEL 3.0 and called SNPs for all individuals using the UNEAK pipeline (Bradbury et al., 2007, Glaubitz et al., 2014, Lu et al., 2013). We used default parameters and putative tags that were recorded at least five times across all samples were retained. We compared putative tags using network filtering and considered those with a 1 bp mismatch to be candidate SNPs, using an error tolerance rate of 0.03 (Lu et al., 2013). After these initial filtering steps, 6,102,279 reads remained and we used these to create a pseudo-chromosome, which acted as a reference. SNPs were called for each individual and we identified a total of 291,189 SNPs. When calling SNPs against the reference pseudo-chromosome, we only considered tags recorded in at least three individuals. We used a maximum genotype depth of 60 (~10X the average genotype depth of the dataset) for all data sets to filter out genotypes that were possibly overrepresented (Almeida *et al.* 2014). We removed any locus that was missing in more than 90% of samples and, subsequently, removed all samples that had less than 10% of the total number of loci. We tested our data for conformity to Hardy Weinberg Equilibrium (HWE) (VCFTOOLS, v0.1.12b) (Danecek et al., 2011) and, after correcting *p*-values for false discovery rates, removed any SNPs that were not in equilibrium.

We examined a range of filtering metrics (detailed methods in Supplementary Material, Fig. S1) and chose a final dataset with a minor allele frequency (MAF) of 0.05, a minimum genotype depth of 3 and maximum level of missingness of 50%, resulting in the inclusion of 16,177 SNP loci across 499 individuals. We chose this dataset because including tags with

large proportions of missing data can obscure signal, while stringent missingness filtration by definition reduces diversity, leading to an underestimation of population structure (Huang and Knowles, 2016). Regarding MAF, higher thresholds are proven to inhibit population substructure detection, a problem that also persists when singletons are retained (i.e. low MAF threshold) (Linck and Battey, 2019). Hereafter this SNP dataset will be called the genome-wide dataset.

Population structure analysis

Population structure based on genome-wide SNP loci

We assessed patterns of connectivity and gene flow across the geographic range in Australia. Global and collection locality (location vs all others) pairwise F_{ST} values were estimated in the R package HIERFSTAT (Goudet, 2005), as were observed heterozygosity [H_O] and within population gene diversity [H_S]. Upper and lower confidence intervals (95%) were obtained for F_{ST} from 10,000 bootstraps, *p*-values were corrected for multiple comparisons using FDR 0.05 (Benjamini and Hochberg, 1995). F_{ST} vs. All (F_{ST} of one population against all other individuals) and Tajima's D (Tajima, 1989) values between all collection localities were calculated in the R package POPGENOME (Pfeifer et al., 2014).

We determined the number of distinct genetic groups in Australia using a Bayesian assignment test in the program FASTSTRUCTURE (v1.0) (Raj et al., 2014) (detailed methods in Supplementary Material).

Population structure based on outlier SNP loci

We identified outlier loci in BAYESCAN (v2.1) (Fischer et al., 2011), which uses a Bayesian approach to compare the F_{ST} values of a locus across subpopulations (Foll and Gaggiotti, 2008). BAYESCAN was run for two specified subpopulations. On the first run, we allocated individuals to one of 24 subpopulations that related to their collection localities. On the

second run, we allocated individuals to one of three subpopulations that related to aridity (arid, semi-arid, and not-arid; aridity was identified as important in the environmental PCA) (Middleton and Thomas, 1997). BAYESCAN is prone to high levels of false positives; however, Lotterhos and Whitlock (2014) showed that this effect can be reduced by increasing the prior odds value. BAYESCAN was run with default parameters except that the prior odds value was set to 100. Only loci with a Bayes factor greater than three were kept as outliers because values above this indicate substantial evidence of selection (Jeffreys, 1961). We determined the number of distinct genetic groups as identified by the outlier loci dataset using FASTSTRUCTURE (detailed methods in Supplementary Material).

Genetic, environmental and geographic relationships

We used Multiple Matrix Regression Analysis (MMRR) to determine if geographic distance (between localities) predicts genetic distance (isolation by distance, IBD), or if environmental distance (using the environmental data variables) predicts genetic distance (isolation by environment; IBE). MMRR was implemented by comparing a pairwise genetic distance matrix with IBD and IBE distance matrices following Wang et al. (2013). This was conducted on both the genome-wide and the outlier SNP loci datasets. Additionally, we used regression to determine if geographic distance predicts environmental distance.

Detecting putative loci under selection

We investigated adaptive variation across the starling's range using candidate loci identified by BAYSCAN (F_{ST} outlier loci) as outlined above, and through two different methods of environmental association analyses.

We used BAYENV2 (v2.0; Gunther and Coop 2013) to test for association of minor allele frequencies with environmental variables. Bayenv2 calculates the correlation between allele frequencies and environmental variables and was used to tests for associations between all SNPs and three environmental variables (Coop et al., 2010, Gunther and Coop, 2013). The

three environmental variables that were tested included aridity, Bio05 and Bio15. These variables were chosen because temperature and precipitation have previously been shown to influence starling phenotype (Cardilini et al., 2016). We estimated the covariate matrix using a random subset of 2,000 loci. Runs were set at 100,000 iterations and we kept any locus with a Bayes factor greater than three as being potentially under selection. From the candidate SNPs identified by BAYENV2, we extracted the SNP that most highly correlated with each of Aridity, Bio05, and Bio15, and plotted allele frequency against the environmental variables score.

We used the R package vegan (v3) (Oksanen et al., 2018) to conduct redundancy analysis (RDA) to determine if any particular SNP locus is heavily loaded onto predictor axes, from which we may infer the occurrence of selection. Under the conditions of genotype-environment associations, constrained ordination techniques such as RDA represent parsimonious models that attempt to capture total population variation specifically explained by constraining predictor (e.g. environmental) variables. Constrained ordination methods are fast becoming a fundamental method when conducting genotype-environment associations with multivariate and multilocus datasets, boasting high detection power coupled with low false-positive and high true-positive rates (Forester et al., 2018), and as such are invaluable for investigative adaption (Capblancq et al., 2018), particularly in invasive populations where selection may be obscured by demographic signals. Thus RDA environmental association analysis was conducted alongside the traditional allele frequency association tests. The genome-wide SNP dataset was loaded against the climatic variables associated with our sampling sites. The SNP genotype dataset was processed into 012 format (count reflects the number of the minor allele) using VCFtools, and missing SNP genotypes were imputed based on probabilities for each given genotype at that genome locality. Of the complete list of environmental variables, we retained seven predictors (Elev, Bio03, Bio06, Bio13, Bio15, mn NDVI, and varDL) with relatively low variance inflation factors (range: 1.9—6.3) to reduce

multicollinearity. Once the SNP data were loaded against the RDA axes, candidates for selection were determined to be those that lay more than 3 standard deviations away from the mean.

We extracted the associated 64 bp tags for loci identified by at least one of the methods of BAYESCAN, BAYENV2, and RDA. These tags were searched against the annotated starling genome available on NCBI (GCF_001447265.1), following the protocol described in Richardson and Sherman (2015) (Richardson and Sherman, 2015). Blast matches were assigned names and functional descriptions.

Results

Population environmental distribution

A PCA analysis of environmental parameters grouped collection localities into three distinct clusters (Fig. 1; Supplementary Material, Fig. S3, Table S1). These clusters defined three distinct environmental regions: (i) Western Australian and South Australian collection localities (arid); (ii) all eastern Australia collection localities inland of The Great Dividing Range (semi-arid); and (iii) all collection localities on the coastal side of The Great Dividing Range (non-arid) (Supplementary Material, Table. S1).

Population structure analysis

Population structure based on genome-wide SNP loci

Analysis of the patterns of genetic diversity revealed that populations at the expansion front in Western Australia (Munglinup and Condingup) displayed relatively lower levels of genetic diversity compared to populations in the east of the introduced range (Table 1). Further, Munglinup was the only population to have a positive Tajima's D (low frequency of rare alleles) indicating a potential population contraction or action of balancing selection. The

global level of genetic differentiation across collection localities was low, but significant with an average F_{ST} of 0.027 (95% CI 0.025 – 0.030) (Supplementary Material, Table S2). Collection locality (location vs all others) pairwise F_{ST} values revealed that those populations with the lowest levels of genetic diversity also showed the greatest level of genetic differentiation to all other populations (Table 1).

The Bayesian structure analysis based on the full dataset of 16,177 SNPs identified four genetic clusters (Fig. 1; Munglinup, southern Australia, southwest New South Wales, northern Australia, Supplementary Material; Fig. 3Sa). The Munglinup and southwest New South Wales clusters were distinct while the two larger clusters were less well defined suggesting some admixture. Subsequent FASTSTRUCTURE analysis found no substructure within any of the four genetic clusters.

Population structure based on outlier SNP loci

Out of the 16,177 SNPs, BAYESCAN analyses identified 89 outlier loci (Table 2) as candidates for selection (log10(BF) > 3), while all other loci were considered neutral. Pairwise F_{ST} values calculated for outlier loci were considerably higher than those calculated for the genome-wide dataset, with an average F_{ST} of 0.122 (95% CI 0.101 – 0.130, non-overlapping with the CI for genope-wide SNP loci calculated F_{ST}) and most locality F_{ST} comparisons were significant (256/276; Supplementary Material, Table S2). FASTSTRUCTURE analyses of the outlier dataset identified two distinct genetic clusters (Western and Southern Australia, Eastern and Northern Australia; Supplementary Material, Figure S3b), whereas we found no genetic structure in the random subset of loci (Supplementary Material, Fig. S2). Subsequent FASTSTRUCTURE analysis found no substructure in the two main genetic clusters (data not presented).

Genetic, environmental and geographic relationships

Geographic distance did not predict environmental distance across the starling's Australian range (Fig. 2).

Similarly, when genetic distance was tested against geographic distance and environmental distance in an MMRR, only IBD was found to explain any variation in genetic differentiation observed between collection localities (Fig. 3a). IBE was not found to be a significant predictor of genetic distance (Fig. 2).

When the genetic distance of the outlier dataset was tested against geographic distance and environmental distance in an MMRR, only IBD was found to explain any variation between collection localities (Fig. 3b). IBE was not found to be a significant predictor of genetic distance (Fig. 2).

Detecting putative loci under selection

Using three different methods to identify SNPs putatively under selection, we found a total of 375 different SNPs. BAYESCAN analyses of F_{ST} identified 89 outlier loci (Table 2).

BAYENV2 identified positive associations between 245 SNPs and three environmental variables including Aridity (N = 28 SNPs), Bio05 (temperature) (N = 174 SNPs), and Bio15 (precipitation) (N = 43 SNPs) (Table 2). The minor allele frequency of the locus most strongly associated with aridity (ID = 68063014) showed a negative relationship (C = -0.272, SE = 0.070, t = -3.883, P < 0.001, r^2 = 0.380; Fig. 5a). The minor allele frequencies of the SNP (ID = 233190026) most strongly associated with Bio05 showed a strong positive relationship (C = 0.018, SE = 0.004, t = 4.116, P < 0.001, r^2 = 0.409; Fig. 5b). The relationship between Bio05 and the minor allele frequency for this SNP is closely aligned with geography, with the Munglinup cluster and almost all collection localities from the southern Australia cluster being fixed at this locus. Loci associated with Bio15 (ID =

196163030) showed a positive relationship (C = 0.008, SE = 0.002, t = 3.607, P = 0.002, r^2 = 0.343; Fig. 5c), with most individuals from the northern Australia cluster being fixed at this locus.

Genotypes within WA and SA were positively correlated with variables Bio03, Bio06, and Bio15 (Isothermality, Min Temperature of Coldest Month, and Precipitation Seasonality respectively), and were negatively associated with genotypes relating to Elev and mnNDVI (Elevation and mean vegetation cover) (Fig. 4). The converse of this was true for the genotypes of inland NSW starlings. Genotypes within VIC and TAS were positively correlated with varDL (Variability in Day Length), and negatively associated with genotypes relating to Bio13 (Precipitation of Wettest Month), while the converse of this was true for coastal NSW and QLD starlings.

RDA identified 111 candidate SNPs that showed strong association with the local environment (F = 1.2228, P = 0.001; Fig. 4). From these, a total of 25 proteins were identified when mapped to the starling reference genome, all of which came from SNPs identified through only one of the three candidate selection approaches (Supplementary Material, Table S3). Of these 25 mapped proteins, 7 loci were associated with BAYESCAN F_{ST} outliers, 11 were associated with BAYENV2 environmental variables, and 8 were associated with candidate loci identified through the RDA environmental association (Supplementary Material, Table S3). Among the annotated SNPs, we see a range of biological functions traditionally associated with adaptation in invasive species, including processes such as immune system responses (C1QC and STAB1) (Colautti et al., 2004), temperature tolerance (HSPA9), stress (TRAF1) (Zerebecki and Sorte, 2011), and dietary alterations (CTRL) (Spit et al., 2014).

Discussion

Our analysis of genome-wide SNP markers identified low but significant genetic structuring over the invasive range of starlings in Australia. Analyses of candidate SNPs that are putatively under selection indicate that temperature may be an important factor driving selective change across the starling's invasive range in Australia. However, our patterns of adaptive genetic variation across the invasive range is likely to be heavily influenced by the historic introduction regime, or from the process of rapid and recent range expansion. These demographic factors are not common in natural populations, and so are rarely considered in studies of adaptation in native species. When examining putative adaption in invasive populations, we encourage critical interpretation of signatures of selection. Even strongly associated loci and environmental variables, when examined closely, may contain distinct footprints of invasion history that are likely to confound analysis and interpretations of selection history.

Population structure

Population structure based on genome-wide SNP loci

In this study, our genome-wide SNP data set identified four distinct genetic clusters, two of which covered large geographic regions; the first encompassing southern Australia, from Western Australia to Tasmania (the southern Australia cluster), and the other encompassing New South Wales and southern Queensland (northern Australia cluster). Munglinup and southwest New South Wales represented two small distinct clusters, indicating higher population isolation in these areas. Additionally, Munglinup represented the only population that reported a positive Tajima's D (Table 1), which suggests demographic (e.g. a sudden population contraction) or selection effects. The identification of the two large genetic clusters (the southern Australia cluster and the northern Australia cluster) may be explained

by two factors, 1) geographic isolation and/or 2) historic genetic diversity from different source populations.

Starlings are predominantly found in agricultural landscapes relying on open fields to forage (Feare, 1985, Whitehead et al., 1995) and have an upper elevation limit of 2000m (Higgins et al., 2006). There are two possible barriers to movement between VIC & NSW. First, large contiguous areas of forest and elevated mountain ranges along the coast between the geographic regions of Victoria and New South Wales are likely to act as a barrier to starling movement, restricting gene flow (Fig. 1). High elevation is suspected to play an important role in shaping starling demography across multiple invasive populations, including New Zealand (Ross, 1983), and more recently in a parallel study of the North American starlings (Hofmeister et al. 2019). Second, the inland route (on the western side of The Great Dividing Range) between VIC and NSW localities is hot and dry, which may also restrict starling movement. These biogeographical barriers are therefore likely to have played a minor role in the genetic differentiation seen between the southern Australia and the northern Australia regions, despite the non-significant IBE result. This is suggestive that MMRR approaches to IBE may not be entirely appropriate for invasive populations, owing to the amalgamation of different demographic effects within such populations. The population genetic structure of Australian starlings identified with the genome-wide dataset of SNPs was similar to that identified in previous work (Rollins et al., 2011, Rollins et al., 2009).

Population structure based on outlier SNP loci

Population structure analysis using the outlier loci dataset revealed only two population clusters. A restricted outlier dataset may allow for the detection of real adaptive variation that may be lost in the complete dataset (though prone to false positives unless physiologically checked), and has been used in many prior studies to resolve in finer detail structure than neutral loci (Benestan et al., 2015, Hess et al., 2013, Milano et al., 2014). This was not seen

in the common starling, with the neutral dataset identifying four genetic clusters while the outlier loci identified two genetic clusters (Supplementary Material; Fig. 3S). It is likely that population analysis using only the F_{ST} outliers may bias population structure to capture legacy genotypes, as loci may be fixed in one population and not in the other, either through genetic differences in founding individuals or through random processes such as allelic loss and allelic surfing (Excoffier and Ray, 2008). Historical records point to key mainland Australia starling introductions at Adelaide (SA), Melbourne (VIC), and Sydney (NSW) (Feare, 1985). If gene flow between starlings in the southern Australia cluster and the northern Australia cluster has been restricted by limited dispersal through the landscape, it is possible that a signature of the different source populations would remain from introduction and that further genetic differentiation (as seen in the genome-wide dataset) could have resulted from the effects of allele surfing, genetic drift, subpopulation admixture, and local adaptation. It is evident then that invasive population demographics are constrained by local signatures from historic founding populations, and possibly also by genetic gradients resulting from range expansion (discussed further below).

A strong genetic footprint resulting from founding populations is further supported by the fact that the environmental clustering of populations (Supplementary Material, Fig. S4) does not reflect genetic clustering in either the whole-genome or outlier dataset (Supplementary Material, Fig. S3**)**. Although analyses indicate clearly defined environmental regions, environmental parameters examined in this study explained little of the genome-wide genetic structure observed in Australian starling populations (Fig. 2). Geographic distance showed a strong relationship with, and explained a large amount of the variation in, genetic difference between collection localities. These findings are in contrast to similar analysis conducted on the North American starlings, which found geographic distance was a poor predictor of genetic differentiation across the species, and that environmental distance played a greater role in population similarity and clustering (Hofmeister et al. 2019).

Population structure based through isolation by distance

Despite being a highly dispersing species, we find a strong pattern of isolation by distance, with particular restricted gene flow between the VIC and NSW populations, possibly a result of the elevation in the Great Dividing Range as discussed above. Though collection localities within regions of Australia show relatively low levels of genetic differentiation, this is nevertheless higher than F_{ST} values reported in the comparatively larger (square km coverage) North American starling distribution (Hofmeister et al. 2019). This may suggest that there are greater constraints to gene flow across the Australian continent than in North America.

Putatively adaptive variation

There is an expectation that local adaptation in highly mobile species (such as starlings) will be rare (Kvistad et al., 2015). High levels of gene flow are often seen in highly mobile species, which acts to homogenise genetic differences between populations, while stabilising selection results in the loss of rare variants (North et al., 2011). However, locally adapted variants have been identified in species that are highly mobile (Benestan et al., 2015, Bourret et al., 2013, Hess et al., 2013, Limborg et al., 2012, Matala et al., 2014, Milano et al., 2014, Nielsen et al., 2009), and in introduced species (Chown et al., 2015, Rohfritsch et al., 2013), which suggests that local adaptation may indeed be common despite high levels of gene flow.

Several hundred unique candidate loci (375) for local adaptation were identified, and from this, a total of 25 proteins were identified (Supplementary Material, Table S3). Interestingly, within the BayEnv2 analysis, we identify many more SNPs that are associated with temperature (174) than precipitation (43) or aridity (28). We propose several biological reasons that temperature, specifically maximum temperature of the warmest month, may be exerting a greater evolutionary pressure on starlings in comparison to precipitation and

aridity. First, range expansion in invasive starling populations is strongly associated with suburban and rural areas (Zufiaurre et al., 2016), owing to a preference for grazing in open pasture land (e.g. agricultural areas, fields, and parks), and the excessive roosting opportunities provided by urban structures. The latitudinal spread and large population area that the Australian starling population covers ensures the species will experience large clinal variation in environmental factors. Urban structures undoubtedly provide supplementary access to water sources and may mitigate selection imposed by environmental factors such as aridity. However, we speculate that temperature and temperature extremes may exert selective pressure that is least mitigated by human-mediated environmental disturbances. Among the other proteins associated with Bio05 are a range of biological functions plausibly linked to temperature such as angiogenesis and cardiac conduction.

There was a particularly strong association between temperature and the candidate SNP associated with the genomic region that codes for Tcdd-Inducible Poly(Adp-Ribose) Polymerase (TIPARP). This protein is involved in pathways controlling androgen and estrogen metabolic process (and hence sexual characteristics), as well as the development of blood, skeleton and various organ structures (Supplementary Material, Table 3) (Ame et al., 2004). Given the large number of functions associated with TIPARP, there could be multiple reasons it is associated with temperature, including variation in exposure to xenobiotic chemicals across a temperature gradient, or because temperature serves as a good proxy for latitude and seasonal variation, both of which cue sexual maturation in starlings (Dawson, 2005).

We observed a more even distribution of candidate SNPs reported by the RDA across axis (Table 2), indicating a more even spread across environmental predictors and associated population clusters. The alignment of population clusters and environmental predictors reveals some interesting trends. For instance, the SNPs in populations from the lowest

latitude and hence more variable annual day-night cycles (VIC and TAS) are strongly positively associated with varDL, while the opposite is true for northern NSW and QLD populations, indicating variable light cycles are exerting a strong selective pressure on this crepuscular species. Similarly, SNPs found in the WA and SA populations are positively correlated with environmental variables related to higher aridity in these states, such as isothermality and precipitation seasonality.

However, closer examination of the plots of allelic frequency across environmental variables reveals that environmental associations may be confounded by local signatures of founding populations. Sampling sites that are fixed for a single allele may not be randomly distributed, but rather may appear exclusively (or predominantly) within a single genetic group. For example, the locus strongly associated with Bio15 (Fig.5c) is fixed in eight localities, six of which are in GC4. It is possible that the alternative allele is not present in those areas due to founding effects at introduction, limiting adaptive potential of those individuals. In fact, if the relationship between allele frequencies at this locus across populations GC1-3 were analysed, it would be even stronger. Therefore, there may be other loci linked to environmental variables whose relationship is obscured by the legacy of limited diversity in some founding populations. Nevertheless, it is still possible that differences between subpopulations (e.g. locus fixation) are a result of selective forces. For instance, Vandepitte *et al*. (2014) described the rapid adaptation of highly conserved photoperiodic flowering genes during the establishment phase (prior to spread), of an invasive population of Pyrenean Rocket, *Sisymbrium austriacum* subsp. c*hrysanthum* (Vandepitte et al., 2014). Characterising the native range genomic landscape for an invasive species would provide invaluable information to help discern the nature of polymorphisms within their invasive range.

It is possible that some of the above problems could be avoided if genetic groups resulting from historic introduction regimes were analysed separately, however, even in a scenario with only one introduction site, the artefacts of range expansion still need to be contended with. Historic starling introduction sites represent relatively favourable conditions for the species (Zufiaurre et al., 2016), and during range expansion, the starlings have slowly colonised less favourable areas, moving into the arid regions of inland and Western Australia. As a result, the expansion trajectory between introduction sites and range edge may co-occur with environmental clines. Further, it is long established that the process of range expansion in invasive species may result in a genetic gradient due to random processes or spatial sorting, and that often range-edge populations exhibit less genetic diversity than those near introductions sites (Rollins et al., 2009, Phair et al., 2018). Evidence may be taken from the functional nature of a polymorphism e.g. (Dudaniec et al., 2018, Leydet et al., 2018), however this is still only circumstantial. White *et al.* (2013) recognise this problem and propose replicated SNP analysis across different sections of the invasive range, suggesting that similar clines across multiple regions are less likely to be due to random processes such as drift (White et al., 2013). However, such an approach is made difficult for some invasive populations, particularly in the case of overlaying multiple introductions. As such, while we may consider that a high environmental association is indicative of selection, it is extremely difficult to separate evolutionary processes in invasive populations due to their rapid range expansion.

Our study demonstrates the necessity of using multiple approaches when searching for loci putatively under selection. We report minor overlap between candidate loci identified through the three methods we used. However, the majority of the loci were reported for only one of the exploratory methods, and all of the annotated proteins identified were unique to their identification method. Additionally, different candidate SNP identification methods will be detecting different aspects of genetic substructure. Population substructure (as determined

by outlier SNP loci) of the Australian starlings strongly reflects introductory site patterns, hence it is likely that SNPs identified through F_{ST} outlier methods may capture legacy genotype differences or signals created during range expansion particular to that subpopulation (Vitti et al., 2013, White et al., 2013). The confounding effects from founder populations are likely to impact, to varying degrees, genetic analysis of invasive species whose range covers multiple introduction sites.

In Australian starling populations, we showed that geographic distance plays a strong role in the genetic structure of populations, and that environmental barriers are likely reinforcing the observed patterns of genetic structure. Our genetic evidence also suggests that the contemporary Australian population of common starlings is most likely made up of two major and spatially distinct populations that likely resulted from different founding populations. Environmental associations for individuals loci are highly suggestive of adaptive genetic variation, and doubtless these selective pressures will continue to drive population differentiation. However, this study highlights the challenges of separating selective and neutral evolutionary processes shaping invasive populations.

Acknowledgements

PC was supported by an ARC Future Fellowship (FT0991420). Thank you to Katherine L. Buchanan and Katie Hyma for their assistance with this project and manuscript.

References

AME, J. C., SPENLEHAUER, C. & DE MURCIA, G. 2004. The PARP superfamily. *Bioessays,* 26**,** 882-893.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M., SHERLOCK, G. & GENE ONTOLOGY, C. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics,* 25**,** 25-29.

BAKER, H. G. & STEBBINS, G. L. 1965. *The Genetics of Colonizing Species,* New York., Academic Press.

BAY, R. A., HARRIGAN, R. J., LE UNDERWOOD, V., GIBBS, H. L., SMITH, T. B. & RUEGG, K. 2018. Genomic signals of selection predict climate-driven population declines in a migratory bird. *Science,* 359**,** 83-86.

BENESTAN, L., GOSSELIN, T., PERRIER, C., SAINTE-MARIE, B., ROCHETTE, R. & BERNATCHEZ, L. 2015. RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (Homarus americanus). *Molecular Ecology,* 24**,** 3299-3315.

BENJAMINI, Y. & HOCHBERG, Y. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology,* 57**,** 289-300.

BIVAND, R. S., EDZER, P. & VIRGILIO, G. R. 2013. *Applied Spatial Data Analysis with R,* New York, Springer

BOURRET, V., KENT, M. P., PRIMMER, C. R., VASEMAGI, A., KARLSSON, S., HINDAR, K., MCGINNITY, P., VERSPOOR, E., BERNATCHEZ, L. & LIEN, S. 2013. SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (Salmo salar). *Molecular Ecology,* 22**,** 532-551.

BRADBURY, P. J., ZHANG, Z., KROON, D. E., CASSTEVENS, T. M., RAMDOSS, Y. & BUCKLER, E. S. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics,* 23**,** 2633-2635.

CABE, P. R. 1999. Dispersal and population structure in the European Starling. *Condor,* 101**,** 451-454.

CAPBLANCQ, T., LUU, K., BLUM, M. G. B. & BAZIN, E. 2018. Evaluation of redundancy analysis to identify signatures of local adaptation. *Molecular Ecology Resources,* 18**,** 1223-1233.

CARDILINI, A. P. A., BUCHANAN, K. L., SHERMAN, C. D. H., CASSEY, P. & SYMONDS, M. R. E. 2016. Tests of ecogeographical relationships in a non-native species: what rules avian morphology? *Oecologia,* 181**,** 783-793.

CHOWN, S. L., HODGINS, K. A., GRIFFIN, P. C., OAKESHOTT, J. G., BYRNE, M. & HOFFMANN, A. A. 2015. Biological invasions, climate change and genomics. *Evolutionary Applications,* 8**,** 23-46.

COLAUTTI, R. I., RICCIARDI, A., GRIGOROVICH, I. A. & MACISAAC, H. J. 2004. Is invasion success explained by the enemy release hypothesis? *Ecology Letters,* 7**,** 721-733.

COOP, G., WITONSKY, D., DI RIENZO, A. & PRITCHARD, J. K. 2010. Using Environmental Correlations to Identify Loci Underlying Local Adaptation. *Genetics,* 185**,** 1411-1423.

DANECEK, P., AUTON, A., ABECASIS, G., ALBERS, C. A., BANKS, E., DEPRISTO, M. A., HANDSAKER, R. E., LUNTER, G., MARTH, G. T., SHERRY, S. T., MCVEAN, G., DURBIN, R. & GENOMES PROJECT ANAL, G. 2011. The variant call format and VCFtools. *Bioinformatics,* 27**,** 2156-2158.

DAWSON, A. 2005. The effect of temperature on photoperiodically regulated gonadal maturation, regression and moult in starlings - potential consequences of climate change. *Functional Ecology,* 19**,** 995-1000.

DLUGOSCH, K. M. & PARKER, I. M. 2008. Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. *Molecular Ecology,* 17**,** 431-449.

DUDANIEC, R. Y., YONG, C. J., LANCASTER, L. T., SVENSSON, E. I. & HANSSON, B. 2018. Signatures of local adaptation along environmental gradients in a range-expanding damselfly (Ischnura elegans). *Molecular Ecology,* 27**,** 2576-2593.

ELSHIRE, R. J., GLAUBITZ, J. C., SUN, Q., POLAND, J. A., KAWAMOTO, K., BUCKLER, E. S. & MITCHELL, S. E. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *Plos One,* 6.

EXCOFFIER, L. & RAY, N. 2008. Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution,* 23**,** 347-351.

FEARE, C. J. 1985. *The Starling*, Issue 7 of Shire natural history.

FISCHER, M. C., FOLL, M., EXCOFFIER, L. & HECKEL, G. 2011. Enhanced AFLP genome scans detect local adaptation in high-altitude populations of a small rodent (Microtus arvalis). *Molecular Ecology,* 20**,** 1450-1462.

FOLL, M. & GAGGIOTTI, O. 2008. A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics,* 180**,** 977-993.

FORESTER, B. R., LASKY, J. R., WAGNER, H. H. & URBAN, D. L. 2018. Comparing methods for detecting multilocus adaptation with multivariate genotype-environment associations. *Molecular Ecology,* 27**,** 2215-2233.

FRANKS, S. J. & MUNSHI-SOUTH, J. 2014. Go forth, evolve and prosper: the genetic basis of adaptive evolution in an invasive species. *Molecular Ecology,* 23**,** 2137-2140.

GLAUBITZ, J. C., CASSTEVENS, T. M., LU, F., HARRIMAN, J., ELSHIRE, R. J., SUN, Q. & BUCKLER, E. S. 2014. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. *Plos One,* 9.

GOUDET, J. 2005. HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes,* 5**,** 184-186.

GRIFFITHS, R., DOUBLE, M. C., ORR, K. & DAWSON, R. J. G. 1998. A DNA test to sex most birds. *Molecular Ecology,* 7**,** 1071-1075.

GUNTHER, T. & COOP, G. 2013. Robust Identification of Local Adaptation from Allele Frequencies. *Genetics,* 195**,** 205-220.

HESS, J. E., CAMPBELL, N. R., CLOSE, D. A., DOCKER, M. F. & NARUM, S. R. 2013. Population genomics of Pacific lamprey: adaptive variation in a highly dispersive species. *Molecular Ecology,* 22**,** 2898-2916.

HIGGINS, P. J., PETER, J. M. & COWLING, S. J. 2006. *Handbook of Australian, New Zealand & Antarctic Birds. Volume 7 Boatbill to Starlings.,* Melbourne, Oxford University Press.

HIJMANS, R. J. 2016. Raster Geographic Data Analysis and Modeling. . https://cran.r-project.org/web/packages/raster/index.html.

HIJMANS, R. J., WILLIAMS, E. & VENNES, C. 2015. Geosphere: spherical trigonometry. https://CRAN. R-project. org/package= geosphere.

HOBAN, S., KELLEY, J. L., LOTTERHOS, K. E., ANTOLIN, M. F., BRADBURD, G., LOWRY, D. B., POSS, M. L., REED, L. K., STORFER, A. & WHITLOCK, M. C. 2016.

Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. *American Naturalist,* 188**,** 379-397.

HOFMEISTER, N. R., WERNER, S. J., LOVETTE, I. J. 2019. Environment but not geography explains genetic variation in the invasive and largely panmictic European starling in North America. Biorxiv.

HUANG, H. T. & KNOWLES, L. L. 2016. Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences. *Systematic Biology,* 65**,** 357-365.

JEFFREYS, H. 1961. *Theory of probability,* Oxford, Oxford University Press.

JENKINS, C. F. H. 1977. *The Noah's ark syndrome,* Western Australia, The Zoological Gardens Board.

KVISTAD, L., INGWERSEN, D., PAVLOVA, A., BULL, J. K. & SUNNUCKS, P. 2015. Very Low Population Structure in a Highly Mobile and Wide-Ranging Endangered Bird Species. *Plos One,* 10.

LEE, C. E. 2002. Evolutionary genetics of invasive species. *Trends in Ecology & Evolution,* 17**,** 386-391.

LEYDET, K. P., GRUPSTRA, C. G. B., COMA, R., RIBES, M. & HELLBERG, M. E. 2018. Host-targeted RAD-Seq reveals genetic changes in the coral Oculina patagonica associated with range expansion along the Spanish Mediterranean coast. *Molecular Ecology,* 27**,** 2529-2543.

LIMBORG, M. T., HELYAR, S. J., DE BRUYN, M., TAYLOR, M. I., NIELSEN, E. E., OGDEN, R., CARVALHO, G. R., BEKKEVOLD, D. & CONSORTIUM, F. P. T. 2012. Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (Clupea harengus). *Molecular Ecology,* 21**,** 3686-3703.

LINCK, E. & BATTEY, C. J. 2019. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources,* 19**,** 639-647.

LOH, P. R., TUCKER, G., BULIK-SULLIVAN, B. K., VILHJALMSSON, B. J., FINUCANE, H. K., SALEM, R. M., CHASMAN, D. I., RIDKER, P. M., NEALE, B. M., BERGER, B., PATTERSON, N. & PRICE, A. L. 2015. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics,* 47**,** 284-290.

LOTTERHOS, K. E. & WHITLOCK, M. C. 2014. Evaluation of demographic history and neutral parameterization on the performance of F-ST outlier tests. *Molecular Ecology,* 23**,** 2178-2192.

LOTTERHOS, K. E. & WHITLOCK, M. C. 2015. The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology,* 24**,** 1031-1046.

LU, F., LIPKA, A. E., GLAUBITZ, J., ELSHIRE, R., CHERNEY, J. H., CASLER, M. D., BUCKLER, E. S. & COSTICH, D. E. 2013. Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. *Plos Genetics,* 9.

MATALA, A. P., ACKERMAN, M. W., CAMPBELL, M. R. & NARUM, S. R. 2014. Relative contributions of neutral and non-neutral genetic differentiation to inform conservation of steelhead trout across highly variable landscapes. *Evolutionary Applications,* 7**,** 682-701.

MIDDLETON, N. & THOMAS, D. 1997. *World Atlas of Desertification,* London: Edward Arnold, UNEP.

MILANO, I., BABBUCCI, M., CARIANI, A., ATANASSOVA, M., BEKKEVOLD, D., CARVALHO, G. R., ESPINEIRA, M., FIORENTINO, F., GAROFALO, G., GEFFEN, A. J., HANSEN, J. H., HELYAR, S. J., NIELSEN, E. E., OGDEN, R., PATARNELLO, T., STAGIONI, M., CONSORTIUM, F., TINTI, F. & BARGELLONI, L. 2014. Outlier SNP markers reveal fine-scale genetic structuring across European hake populations (Merluccius merluccius). *Molecular Ecology,* 23**,** 118-135.

NIELSEN, E. E., HEMMER-HANSEN, J., POULSEN, N. A., LOESCHCKE, V., MOEN, T., JOHANSEN, T., MITTELHOLZER, C., TARANGER, G. L., OGDEN, R. & CARVALHO, G. R. 2009. Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (Gadus morhua). *Bmc Evolutionary Biology,* 9.

NORTH, A., PENNANEN, J., OVASKAINEN, O. & LAINE, A. L. 2011. Local adaption in a changing world: The roles of gene-flow, mutation, and sexual reproduction. *Evolution,* 65**,** 79-89.

OKSANEN, J., GUILLAUME, B. F., FRIENDLY, M., KINDT, R., LEGENDRE, P., MCGLINN, D., MINCHIN, P. R., O'HARA, R. B., SIMPSON, G. L., SOLYMOS, P., HENRY, M., STEVENS, H., SZOECS, E. & WAGNER, H. 2018. vegan: Community Ecology Package. . https://CRAN.R-project.org/package=vegan.

PFEIFER, B., WITTELSBURGER, U., RAMOS-ONSINS, S. E. & LERCHER, M. J. 2014. PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. *Molecular Biology and Evolution,* 31**,** 1929-1936.

PHAIR, D. J., LE ROUX, J. J., BERTHOULY-SALAZAR, C., VISSER, V., VAN VUUREN, B. J., CARDILINI, A. P. A. & HUI, C. 2018. Context-dependent spatial sorting of dispersal-related traits in the invasive starlings (Sturnus vulgaris) of South Africa and Australia. Biorxiv.

RAJ, A., STEPHENS, M. & PRITCHARD, J. K. 2014. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics,* 197**,** 573-U207.

RCORETEAM 2015. A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

RICHARDSON, M. F. & SHERMAN, C. D. H. 2015. De Novo Assembly and Characterization of the Invasive Northern Pacific Seastar Transcriptome. *Plos One,* 10.

RIQUET, F., DAGUIN-THIEBAUT, C., BALLENGHIEN, M., BIERNE, N. & VIARD, F. 2013. Contrasting patterns of genome-wide polymorphism in the native and invasive range of the marine mollusc Crepidula fornicata. *Molecular Ecology,* 22**,** 1003-1018.

ROHFRITSCH, A., BIERNE, N., BOUDRY, P., HEURTEBISE, S., CORNETTE, F. & LAPEGUE, S. 2013. Population genomics shed light on the demographic and adaptive histories of European invasion in the Pacific oyster, Crassostrea gigas. *Evolutionary Applications,* 6**,** 1064-1078.

ROLLINS, L. A., WOOLNOUGH, A. P., SINCLAIR, R., MOONEY, N. J. & SHERWIN, W. B. 2011. Mitochondrial DNA offers unique insights into invasion history of the common starling. *Molecular Ecology,* 20**,** 2307-2317.

ROLLINS, L. A., WOOLNOUGH, A. P., WILTON, A. N., SINCLAIR, R. & SHERWIN, W. B. 2009. Invasive species can't cover their tracks: using microsatellites to assist

management of starling (Sturnus vulgaris) populations in Western Australia. *Molecular Ecology,* 18**,** 1560-1573.

ROSS, H. A. 1983. Genetic differentiation of starling (Sturnus-vulgaris-aves) populations in New Zealand and Great Britain *Journal of Zoology,* 201**,** 351-362.

SHERMAN, C. D. H., LOTTERHOS, K. E., RICHARDSON, M. F., TEPOLT, C. K., ROLLINS, L. A., PALUMBI, S. R. & MILLER, A. D. 2016. What are we missing about marine invasions? Filling in the gaps with evolutionary genomics. *Marine Biology,* 163.

SPIT, J., ZELS, S., DILLEN, S., HOLTOF, M., WYNANT, N. & BROECK, J. V. 2014. Effects of different dietary conditions on the expression of trypsin- and chymotrypsin-like protease genes in the digestive system of the migratory locust, Locusta migratoria. *Insect Biochemistry and Molecular Biology,* 48**,** 100-109.

TAJIMA, F. 1989. Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics,* 123**,** 585-595.

TEPOLT, C. K., DARLING, J. A., BAGLEY, M. J., GELLER, J. B., BLUM, M. J. & GROSHOLZ, E. D. 2009. European green crabs (Carcinus maenas) in the northeastern Pacific: genetic evidence for high population connectivity and current-mediated expansion from a single introduced source population. *Diversity and Distributions,* 15**,** 997-1009.

VANDEPITTE, K., DE MEYER, T., HELSEN, K., VAN ACKER, K., ROLDAN-RUIZ, I., MERGEAY, J. & HONNAY, O. 2014. Rapid genetic adaptation precedes the spread of an exotic plant species. *Molecular Ecology,* 23**,** 2157-2164.

VITTI, J. J., GROSSMAN, S. R. & SABETI, P. C. 2013. Detecting Natural Selection in Genomic Data. *In:* BASSLER, B. L., LICHTEN, M. & SCHUPBACH, G. (eds.) *Annual Review of Genetics, Vol 47.*

WANG, I. J., GLOR, R. E. & LOSOS, J. B. 2013. Quantifying the roles of ecology and geography in spatial genetic divergence. *Ecology Letters,* 16**,** 175-182.

WHITE, T. A., PERKINS, S. E., HECKEL, G. & SEARLE, J. B. 2013. Adaptive evolution during an ongoing range expansion: the invasive bank vole (Myodes glareolus) in Ireland. *Molecular Ecology,* 22**,** 2971-2985.

WHITEHEAD, S. C., WRIGHT, J. & COTTON, P. A. 1995. Winter field use by the European starling Sturnus vulgaris - habitat preferences and the availability of prey *Journal of Avian Biology,* 26**,** 193-202.

WOOLNOUGH, A. P., GRAY, G. S., LOWE, T. J., KIRKPATRICK, W. E., ROSE, K. & MARTIN, G. R. 2005. *Distribution and abundance of pest animals in Western Australia: A survey of institutional knowledge,* Western Australia, Vertebrate Pest Research Section, Department of Agriculture.

ZEREBECKI, R. A. & SORTE, C. J. B. 2011. Temperature Tolerance and Stress Proteins as Mechanisms of Invasive Species Success. *Plos One,* 6.

ZORNER, R. J., TRABUCCO, A., BOSSIO, D. A. & VERCHOT, L. V. 2008. Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. *Agriculture Ecosystems & Environment,* 126**,** 67-80.

ZUFIAURRE, E., ABBA, A., BILENCA, D. & CODESIDO, M. 2016. Role of landscape elements on recent distributional expansion of European starlings (Sturnus vulgaris) in agroecosystems of the Pampas, Argentina. *Wilson Journal of Ornithology,* 128**,** 306-313.

Fig. 1 Map of Australia. Points denote collection localities. Point shape indicates the environmental cluster that the point belonged to, as identified by a PCA analysis of environmental variables; circle = arid, square = semi-arid, diamond = non-arid. Point colour indicates the genetic cluster that the population belonged to as identified by population Bayesian fastSTRUCTURE analysis; pink = Munglinup cluster, yellow = southern Australia cluster, orange = southwest New South Wales cluster, and blue = northern Australia cluster. The blue line represents The Great Dividing Range Mountains.

Fig. 2. Results of Multiple Matrix Regression Analysis (MMRR) capturing genetic, environmental and geographic relationships.

Fig. 3. Plot of an isolation by distance (IBD) mantel correlation. Each point represents a pairwise relationship between collection localities and the line represents the linear relationship between genetic distance and geographic distance. Plot a) represents IBD of the full dataset, while plot b) represents IBD of the outlier dataset. Both genetic distance and geographic distance were log_{10} transformed.

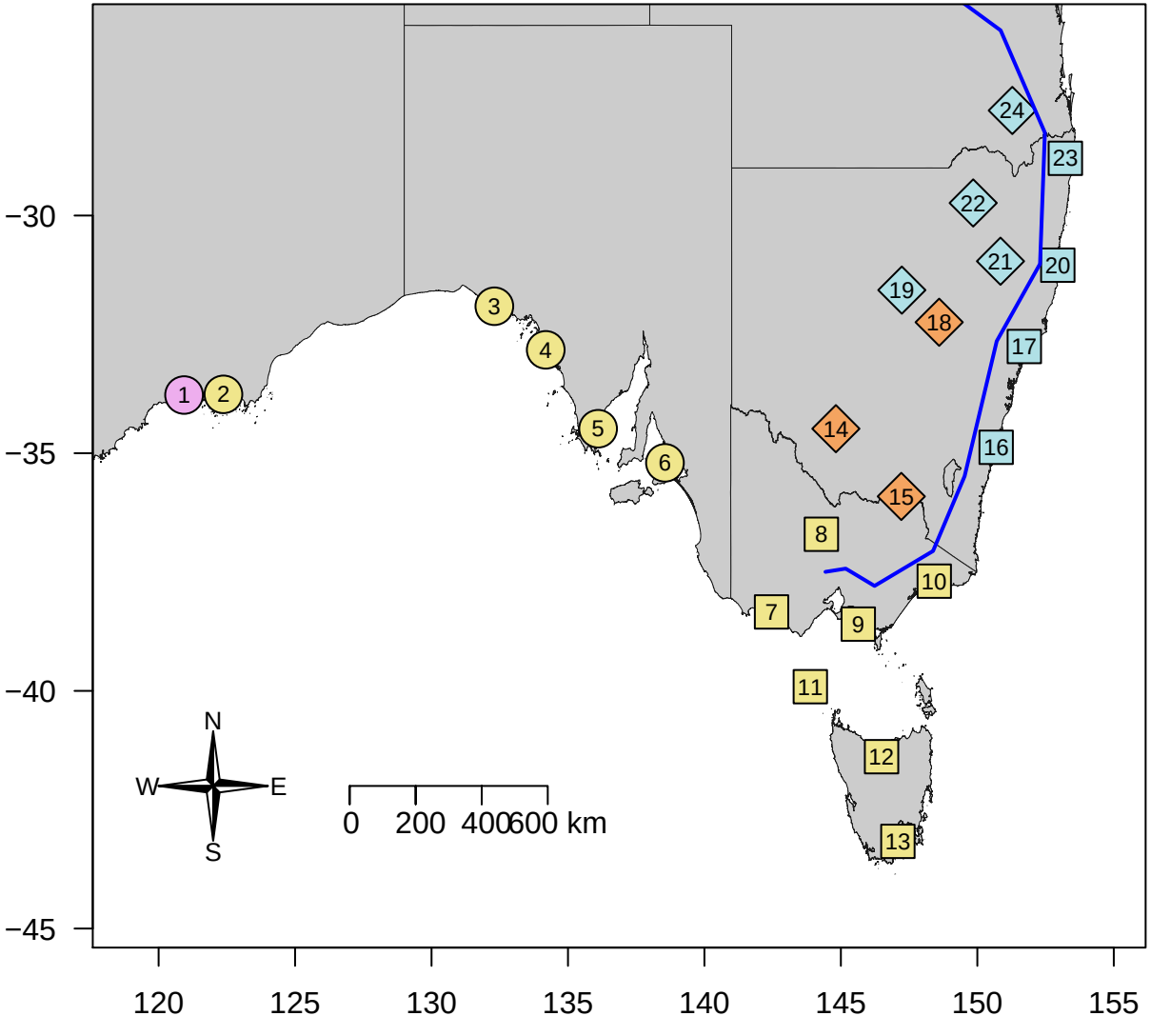
Fig. 4: The ordination plot of redundancy analysis (RDA) of 16177 SNP data for 499 *Sturnus vulgairs*, collected from 24 sample sites across 5 Australian states. Individuals are coloured by location, black arrows represent selected environmental variables. The relative placement of environmental predictor arrows to individual and location sample clustering represents the correlation between that environmental variable and the genetic differences seen in that cluster.

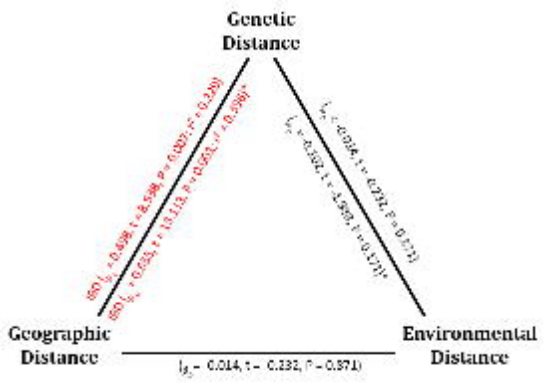
Fig. 5. The relationship between collection locality allele frequency and a locus strongly associated with a) aridity, b) Bio05 and c) Bio15. The black line in each panel shows the linear relationship between the environmental variable and allele frequency, with the grey ribbon indicating the 95% confidence interval. Each point represents a collection locality with its shape indicating the genetic cluster that it belonged to; circle = Munglinup cluster, triangle = southern Australia cluster, square = southwest New South Wales cluster, and cross = northern Australia cluster. Each panel corresponds to SNP ID 233190026, 68063014 and 196163030 respectively.

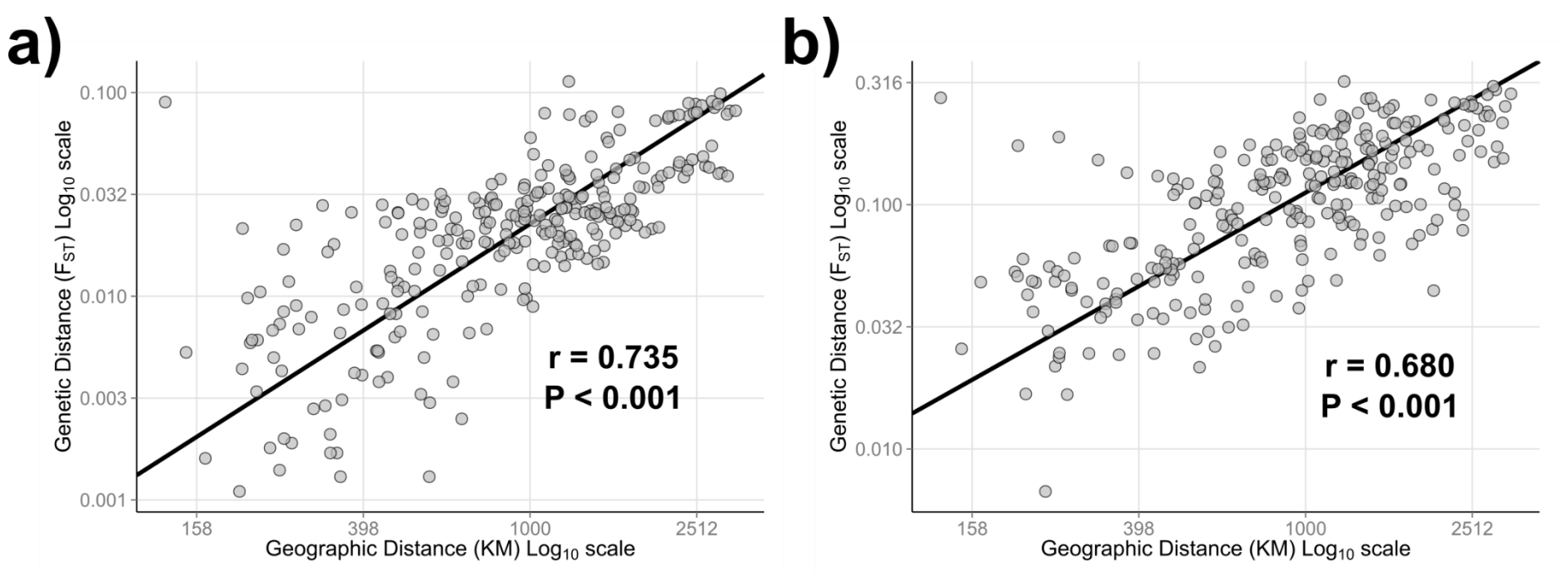
Table 1. Summary statistics for each collection locality. N = number of samples collected from each location, GC = the genetic cluster from whole-genome dataset (1 = Munglinup cluster, 2 = southern

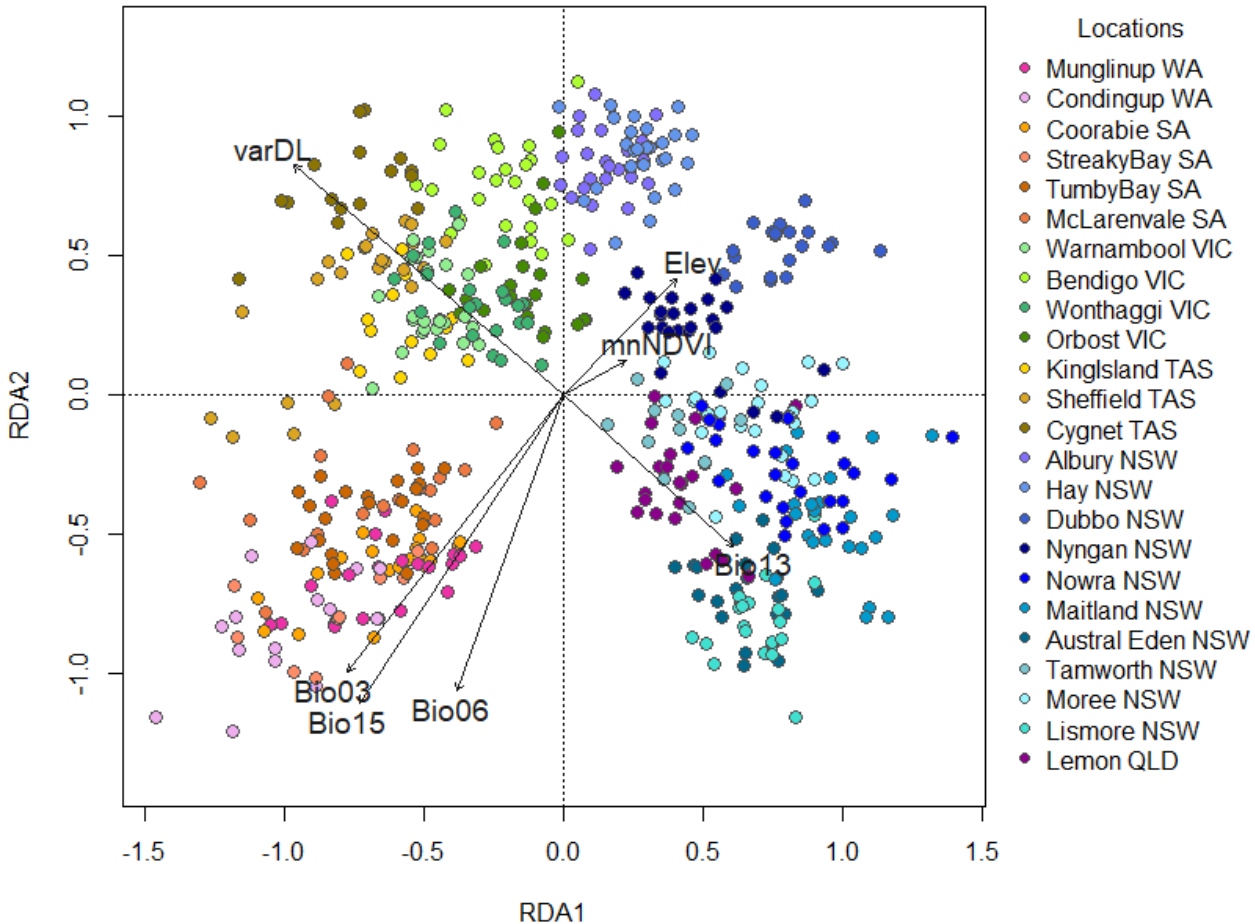
Australia cluster, 3 = southwest New South Wales cluster, 4 = northern Australia cluster), OGC = the genetic cluster from outlier dataset (1 = Western and Southern Australia, 2 = Eastern and Northern Australia), H_O = Observed Heterozygosity, H_S = the within population gene diversity , F_{ST} (vs. All) = fixation index for each population against all other individuals, % Poly SNPs = the percentage of polymorphic SNPs.

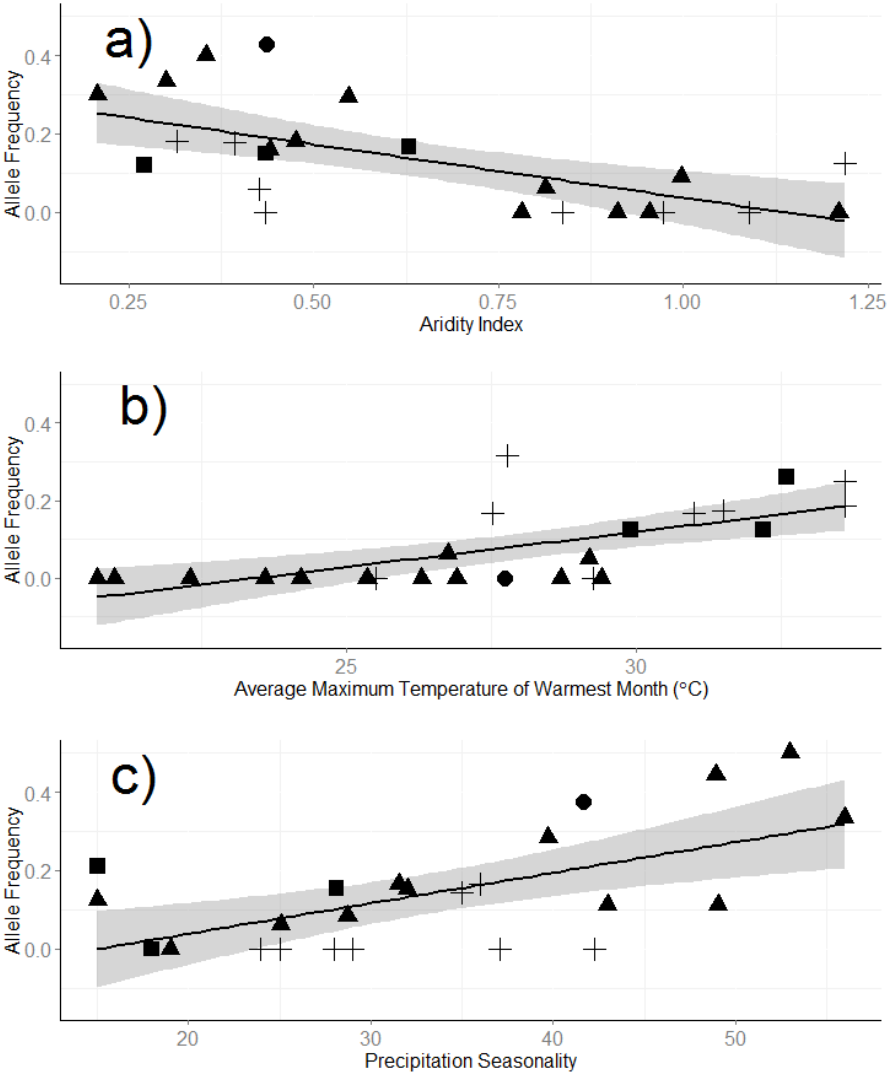
Table 2. Summary of candidate SNPs under putative selection, as detected by Fst outlier (Bayescan) and environmental association (Bayenv2 and RDA) approaches. Bayenv2 analysis is broken down into the three tested environmental variables Aridity, Bio5 (maximum temperature of the warmest month), and Bio 15 (precipitation seasonality). RDA is broken down into three key RDA axis. Overlapping and unique SNPs associated for each method are displayed, as well as common SNPs across the approaches, displayed in a matrix format.


−30
−35
−40
−45
120
125
130
135
140
145
150
155
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
N
W
E
S
0
200
400
600 km


Genetic Distance
Geographic Distance
Environmental Distance


a)
0.100
0.032
0.010
0.003
0.001
Genetic Distance (F_{ST}) Log_{10} scale
158
398
1000
2512
Geographic Distance (KM) Log_{10} scale
r = 0.735
P < 0.001
b)
0.316
0.100
0.032
0.010
Genetic Distance (F_{ST}) Log_{10} scale
158
398
1000
2512
Geographic Distance (KM) Log_{10} scale
r = 0.680
P < 0.001


Locations
Munglinup WA
Condingup WA
Coorabie SA
StreakyBay SA
TumbyBay SA
McLarenvale SA
Warnambool VIC
Bendigo VIC
Wonthaggi VIC
Orbost VIC
KinglsIand TAS
Sheffield TAS
Cygnet TAS
Albury NSW
Hay NSW
Dubbo NSW
Nyngan NSW
Nowra NSW
Maitland NSW
Austral Eden NSW
Tamworth NSW
Moree NSW
Lismore NSW
Lemon QLD
varDL
Elev
mnNDVI
Bio13
Bio03
Bio15
Bio06
RDA1
RDA2
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5


a)
Allele Frequency
0.4
0.2
0.0
0.25
0.50
0.75
1.00
1.25
Aridity Index
b)
Allele Frequency
0.4
0.2
0.0
25
30
Average Maximum Temperature of Warmest Month (°C)
c)
Allele Frequency
0.4
0.2
0.0
20
30
40
50
Precipitation Seasonality

Collection Locality	State	N	GC	OGC	Env. Cluster	H_O	H_S	F_{ST} (vs. All)	% Poly. SNPs	Tajima's D
Munglinup	WA	18	1	1	1	0.092	0.225	0.094	55.2	0.137
Condingup	WA	16	2	1	1	0.092	0.257	0.051	66.1	-0.129
Coorabie	SA	19	2	1	1	0.137	0.265	0.028	80.9	-0.395
Streaky Bay	SA	9	2	1	1	0.069	0.267	0.137	35.9	NA
Tumby Bay	SA	22	2	1	1	0.114	0.269	0.030	83.3	-0.134
McLarenvale	SA	20	2	1	1	0.141	0.266	0.032	86.0	-0.423
Warrnambool	VIC	24	2	1	3	0.117	0.271	0.026	88.6	-0.286
Bendigo	VIC	25	2	1	2	0.159	0.273	0.027	93.4	-0.449
Wonthaggi	VIC	22	2	1	3	0.119	0.272	0.030	81.6	-0.292
Orbost	VIC	22	2	1	3	0.148	0.270	0.028	89.0	-0.468
King Island	TAS	14	2	1	3	0.135	0.272	0.027	80.9	-0.395
Sheffield	TAS	21	2	1	3	0.098	0.265	0.037	79.2	-0.106
Cygnet	TAS	17	2	1	3	0.095	0.262	0.037	72.5	-0.190
Albury	NSW	24	3	1	2	0.200	0.259	0.038	89.5	-0.899
Hay	NSW	26	3	1	2	0.158	0.259	0.025	86.2	-0.581
Dubbo	NSW	20	3	2	2	0.165	0.257	0.030	82.9	-0.671
Nyngan	NSW	25	4	2	2	0.126	0.273	0.023	84.7	-0.391
Nowra	NSW	25	4	2	3	0.133	0.268	0.030	89.2	-0.278
Maitland	NSW	25	4	2	3	0.149	0.266	0.034	91.7	-0.391
Austral Eden	NSW	21	4	2	3	0.142	0.273	0.024	87.0	-0.349
Tamworth	NSW	11	4	2	2	0.129	0.271	0.029	64.6	-1.409
Moree	NSW	24	4	2	2	0.166	0.271	0.028	91.6	-0.546
Lismore	NSW	22	4	2	3	0.123	0.270	0.024	86.9	-0.393
Lemon Tree	QLD	24	4	2	2	0.141	0.272	0.024	89.0	-0.418

Approach		Total SNPs	Unique SNPs	Bayescan	Aridity	Bio05	Bio15	Axis 1	Axis 2	Axis 3
F_{st} Outlier	Bayescan	89	69	-						
Environmental association: Bayenv2	Aridity	28	23	2	-					
	Bio05	174	156	8	0	-				
	Bio15	43	40	1	0	0	-			
Environmental association: RDA	Axis 1	32	23	8	0	1	1	-		
	Axis 2	27	24	1	0	0	2	0	-	
	Axis 3	52	40	0	3	9	0	0	0	-
All		**445**	**375**							