bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Spatial Areas of Genotype Probability (SPAG): predicting the spatial distribution of adaptive genetic variants under future climatic conditions

Estelle Rochat[1, 2], Stéphane Joost[1, 2, 3]

[1] Laboratory of Geographic Information Systems (LASIG), School of Architecture, Civil and Environmental Engineering (ENAC), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[2] The ADAPTMAP Consortium

[3] The NEXTGEN Consortium

Abstract

In a context of rapid global change, one of the key components for the survival of species is their genetic adaptive potential. Many methods have been developed to identify adaptive genetic variants, but few tools were made available to integrate this knowledge into conservation management. We present here the SPatial Areas of Genotype probability (SPAG), using genotype-environment logistic associations to map the probability of finding beneficial variants in a study area. We define a univariate model predicting the spatial distribution of a single genotype, and three multivariate models allowing the integration of several genotypes, potentially associated with various environmental variables. We then integrate the conditions predicted by climate change scenarios to map the corresponding future spatial distribution of genotypes. The analysis of the mismatch between current and future SPAGs makes it possible to identify a) populations that are better adapted to the future climate through the presence of genetic variants able to cope with future conditions and b) vulnerable populations where genotype(s) of interest are not frequent enough for the individuals to adapt to the future climate. We use the SPAGs to study the potential adaptation of 161 Moroccan and 410 European goats to the bioclimatic conditions. In Morocco, using whole genome sequence data, we identify seven genomic regions strongly associated with the precipitation seasonality (WorldClim database). The predicted shift in SPAGs under strong climate change scenario for 2070 highlights goat's populations likely to be threatened by the expected increase in precipitation variation in the future. In Europe, we find genomic regions associated with low precipitation, the shift in SPAGs highlighting vulnerable populations not adapted to the very dry conditions expected in 2070. The SPAG methodology is successfully validated using training and test samples and provides an efficient tool to take the adaptive potential into account in general conservation frameworks.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Climate change models predict a rise in the world's mean temperature and an increase in extreme precipitation events by the end of the 21st century, which is likely to cause acidification of oceans, a rise in sea-level and more frequent extreme events like storms, drought and flooding (IPCC, 2014). When such important changes occur, many animal and plant species are confronted with a shift away from the favourable conditions necessary for their survival. In such conditions, to avoid extinction, they can either move to more favourable habitats or adapt to their new environment, plastically or genetically (Hughes, 2000). Due to the limits of the dispersal capacity, the loss of favourable habitats and the increase in landscape fragmentation, migration to new areas is often limited (Opdam and Wascher, 2004). Plastic adaptation can allow species to rapidly evolve by changing their behaviour, physiology or morphology (Reed *et al.*, 2011). However, this can potentially lead to a fitness reduction (Duputié *et al.*, 2015) and it will not necessarily ensure persistence of the next generations since it does not rely on any genetic heritable variation. In order to preserve biodiversity, it is therefore crucial to promote the conservation of the genetic adaptive potential in populations (Hoffmann and Sgrò, 2011; Sgrò *et al.*, 2011; Nicotra *et al.*, 2015; Shafer *et al.*, 2015).

The conservation of this adaptive potential is also of main importance for livestock management in order to ensure the persistence of herds, particularly in developing countries facing challenging habitat conditions (e.g. heat stress, poor food resources and the presence of parasites and diseases), where the traditional breeds have adapted to the local environment (Hoffmann, 2010). However, due to the increasing demand for food production, local breeds currently tend to be replaced by high-producing commercial breeds imported from developed countries (Rischkowsky and Pilling, 2007; Hoffmann, 2010). Such animals lack adaptive qualities and may thus be particularly vulnerable to climate change, itself likely to reinforce the already stressful conditions. It is therefore essential to highlight the adaptive potential of livestock species in order to encourage farmers to conserve local traditional breeds, and to carefully design cross-breeding, translocation or artificial selection (Scherf *et al.*, 2008; Allendorf *et al.*, 2010).

One of the essential components of the adaptive capacity of populations is genetic diversity (Allendorf and Leary, 1986). Since mutation rates are generally low, adaptation to rapid environmental changes largely depends on the amount of genetic variants already present in populations, i.e. standing genetic diversity (Orr and Unckless, 2008). With the recent increase in the availability of genetic data and the development of conservation genomics, various tools were developed to integrate the genetic diversity into conservation frameworks. For example, Bonin et al. (2007) developed an index to integrate the intraspecific genetic diversity into conservation planning, Thomassen et al. (2011)

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

proposed a method to summarize phenotypes or genotypes diversity and Vandergast et al. (2011) developed a GIS Toolbox to map the genetic diversity and divergence.

However, conserving neutral genetic diversity in populations may not be sufficient to allow rapid adaptation to increasingly stressful conditions (Reed and Frankham, 2001), and it could be more valuable to specifically preserve the standing genetic variants likely to provide a better adaptation to future conditions, i.e. adaptive genetic diversity (Hoffmann and Willi, 2008; Sgrò *et al.*, 2011; Willoughby *et al.*, 2018) or to combine both approaches (Funk *et al.*, 2012; Pauls *et al.*, 2013). An increasing attention is currently being paid to this issue in conservation discussions (Funk *et al.*, 2019; Hoelzel *et al.*, 2019; Mable, 2019).

Many methods have been developed to identify loci showing signature of local adaptation, based on various assumptions and with different limitations and advantages (Schoville *et al.*, 2012; Vitti *et al.*, 2013; Joost *et al.*, 2013; Hoban *et al.*, 2016). The results were notably used to establish prediction of future habitat range of species facing climate change (Hällfors *et al.*, 2016; Ikeda *et al.*, 2017; Garzón *et al.*, 2019; Razgour *et al.*, 2019). However, there is currently a need to integrate this knowledge in order to predict the distribution of adapted genetic variants along environmental gradients, estimate the adaptive potential of populations and identify individuals better suited to future environmental conditions. Only few studies presented methods to map the spatial distribution of adaptive alleles. Fournier et al (2011) and Banta et al. (2012) addressed this issue using the well-known species distribution modelling software Maxent (Phillips *et al.*, 2006). Fitzpatrick and Keller (2015) proposed a much powerful approach based on two community-level modelling methods (Generalised Dissimilarity Modelling and Gradient Forest) to map the current spatial distribution of adaptive variants and assess the vulnerability of populations under climate change as a function of the mismatch between current distribution of alleles and future predictions.

We propose here a new approach to predict genotype frequencies and map Spatial Areas of Genotypes Probabilities (SPAG) based on logistic genotype-environment associations (Joost *et al.*, 2007) and the theory of conditional probabilities. SPAG - whose concept had been sketched out several years ago (Joost 2006; page 138) - can be used a) to predict the probability of presence of one or many adaptive variants in non-sampled areas b) identify areas where there is a higher probability to find individuals better adapted to future climatic conditions, c) identify vulnerable populations that may be threatened by climate change and d) integrate the results into conservation framework by means of an easy combination with other georeferenced layers. We first introduce the theoretical bases of SPAGs and present an application of the approach to analyse the local adaptive potential of Moroccan and

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

European goats' populations. We apply the methodology to two independent datasets and validate the result using training and testing samples.

Material and Methods

Genetic data

Two genetic datasets characterising goat individuals (*Capra hircus*) were used for the analyses presented here. The first one was produced in the context of the NEXTGEN project (Alberto *et al.*, 2018) and the second was collected by the ADAPTMAP consortium (Stella *et al.*, 2018; http://www.goatadaptmap.org/).

The NEXTGEN consortium produced whole genome sequences data for 161 Moroccan goats from 6 different local breeds. Since goat production system in Morocco is mainly free-range, those goats are living from 8 to 12 months outdoors (Boujenane, 2005), and are confronted to contrasting environmental conditions, from the Sahara desert to the Atlas Mountains (see Figure A in Supp. File 1). The sequencing method is described by Benjelloun et al. (2015) and the final sequences contain 31.8 M of SNPS mapped to the goat's reference genome CHIR v1.0 (Dong *et al.*, 2013).

The ADAPTMAP consortium gathered genetic data for 4'563 goats from 144 breeds, sequenced worldwide with the CaprineSNP50 BeadChip and mapped on the most recent goat's reference genome ARS1 (Bickhart *et al.*, 2017). We used here a subset of these data, constituted of individuals from Switzerland, North of Italy and France. This represented 458 individuals distributed in 196 locations, with 1 to 39 individuals per site. In order to avoid overweighting of some locations, we selected a maximum of five individuals per sampling site. Those five individuals were chosen such as the subset keeping the highest Nei's genetic distances, computed with the function *dist.genpop* from the package *adegenet* (Jombart, 2008) in the R environment (R Development Core Team, 2008). The resulting dataset contains 382 individuals from 196 locations and 11 different breeds (see Figure B in Supp. File 1).

Both genetic datasets were filtered such as to keep only autosome, bi-allelic SNPs, with a maximum missingness per individuals and per site of 0.05, a minimum minor allele frequency of 0.1 and a maximum major genotype frequency of 0.9. The final datasets contain 8'497'971 SNPs for the Moroccan goats and 46'294 SNPs for the European ones.

Environmental data

The climatic conditions of the sampling locations were characterised using the 19 bioclimatic variables (Supp. File 2) from the WorldClim database, representative of the period 1960-1990 (Hijmans et al.,

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2005). Each variable was retrieved as a raster layer with a spatial resolution of 30 arc-seconds (approx. 1km2) and values were extracted for all sampling locations using the *extract* function from the R-package *raster* (Hijmans and van Etten, 2012).

In order to get similar ranges of values for all bioclimatic variables, which makes it easier to compare the subsequently derived models, all variables were standardised for each dataset, by subtracting the mean and dividing by the standard deviation. Some of the bioclimatic variables are highly correlated. However, we choose to keep all of them to be able to identify a posteriori which variable had the strongest effect. Since no models computed involved more than one environmental variable simultaneously, this collinearity will not impact the results.

Population Structure

The genetic population structure was estimated with a Principal Component Analysis (PCA) computed with the function *snpgdsPCA* from the *SNPRelate* R-package (Zheng *et al.*, 2012). In order to avoid a strong influence of SNP clusters on this analysis, we used here a pruned set of SNPs that are in approximate linkage equilibrium with each other. The pruning was performed with the function *snpgdsLDpruning* from the *SNPRelate* package, with a threshold D'=0.2. The resulting datasets contain 59'224 SNPs for the Moroccan goats and 14'571 SNPs for the European ones.

Logistic regressions

The SAM method (Joost *et al.*, 2007) was used to detect genotypes that are strongly associated with an environmental variable and are therefore potential adaptive variants. This method uses logistic regressions (formula 1) to assess the probability of presence of a genotype *p(G1)* as a function of an environmental variable (x1),

$$p(G1) = p(G1 = 1 \mid x_1) = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}} \qquad \text{(Formula 1)}$$

where β0 and β1 are the parameters of the regression to be fitted.

Such models were computed for each genotype with the 19 bioclimatic variables. The statistical significance of the model was assessed using Wald test and G score (log-likelihood ratio), both corrected for the false-discovery rate due to multiple comparisons using the procedure proposed by Benjamini and Hochberg (1995), under an expected false discovery rate (FDR) of 0.01 or 0.05 (i.e. 1% or 5% of the results expected to be false positives).

In order to lower the number of false positive resulting from demographic processes instead of natural selection (Li *et al.*, 2012), logistic models were computed with the addition of covariates corresponding to the coordinates of individuals on the significant components of the PCA (Stucki *et al.*, 2017). The

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

significance of the models with population covariates was assessed using a Wald test and a log likelihood ratio which compares the model with environment and covariates to the model with covariates only. An association was considered as significant if both the models without covariates and with population covariates were significant.

Finally, to identify potential functions of the SNPs involved into the significant associations, we used the NCBI Genome Browser to search for the presence of annotated genes in the genomic region of 10kbp surrounding the SNPs of interest.

All analyses were computed using a combination of the Samβada software (Stucki *et al.*, 2017) and an in-house R-script based on the *glm* function.

SPatial Areas of Genotype Probability (SPAG)

With the β parameters of the logistic regressions previously computed (formula 1), we can estimate the probability of presence of a genotype for any value of an environmental variable (x1). We consequently used them to estimate and delimit on a map the probability of presence of a genotype over the whole region of interest (Joost, 2006; Rochat *et al.*, 2016). We named such a delimited surface *Spatial Area of Genotype Probability (SPAG)*.

As more than one adaptive variant are usually identified, we also developed multivariate models to compute a single map showing the probability of presence of multiple genotypes. Three different multivariate models were developed to date: the Intersection, Union and K-Percentage.

The **Intersection** model (I-SPAG) is used to compute the probability that the variants of interest are all simultaneously present. Following the theory of conditional probability (Kolmogorov, 1956), the probability of simultaneous presence of *n* genotypes *Gi, i=1:n* can be computed using formula 2

$$p(\cap_{i=1}^{n} G_i) = p\left(\cap_{i=1}^{n-1} G_i\right) p\left(G_n \middle| \cap_{i=1}^{n-1} G_i\right) \qquad \text{(Formula 2)}$$

where $p\left(G_n \middle| \cap_{i=1}^{n-1} G_i\right)$ is a conditional probability that can be estimated using a logistic regression where $\cap_{i=1}^{n-1} G_i$ is integrated as a covariate (formula 3)

$$p\left(G_n \middle| \cap_{i=1}^{n-1} G_i\right) = \frac{e^{\beta_0+\beta_1 x_n+\beta_2 \cap_{i=1}^{n-1} G_i}}{1+e^{\beta_0+\beta_1 x_n+\beta_2 \cap_{i=1}^{n-1} G_i}} \qquad \text{(Formula 3)}$$

However, as we would like to use this model to predict the probability of presence of the genotypes for any point of the region of interest, i.e. also where *Gi* values are unknown, we suggested to estimate $\cap_{i=1}^{n-1} G_i$ by $p(\cap_{i=1}^{n-1} G_i)$, which enables the implementation of formula 3 with a recursive model using the univariate formula (see Supp. File 3 for more details).

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The **union** model (U-SPAG) is used to compute the probability that at least one of the adaptive variants is present. We implemented it with the inclusion-exclusion principle (formula 4), using an in-house R script based on the intersection model previously described.

$$p(\cup_{i=1}^{n} G_i) = \sum_{i=1}^{n} p(G_i) - \sum_{i<j} p\left(G_i \cap G_j\right) + \sum_{i<j<k} p\left(G_i \cap G_j \cap G_k\right) + \ldots + (-1)^{n-1} p(\cap_{i=1}^{n} G_i)$$

(Formula 4)

Finally, we developed a **K-percentage** model (K-SPAG) to obtain the probability that an individual carries K% of *n* adaptive variants. This probability can be computed by combining formulas from the union and intersection models (formula 5).

$$p(K\%\ G_{i=1\ldots n}) = p\left(\cup_{i=1}^{n} \cap_{1 \le i_1 < i_2 < \cdots < i_{(K\% * n+1)}} (G_{i_1} \cap G_{i_2} \cap \ldots \cap G_{i_k})\ \right) \qquad \text{(Formula 5)}$$

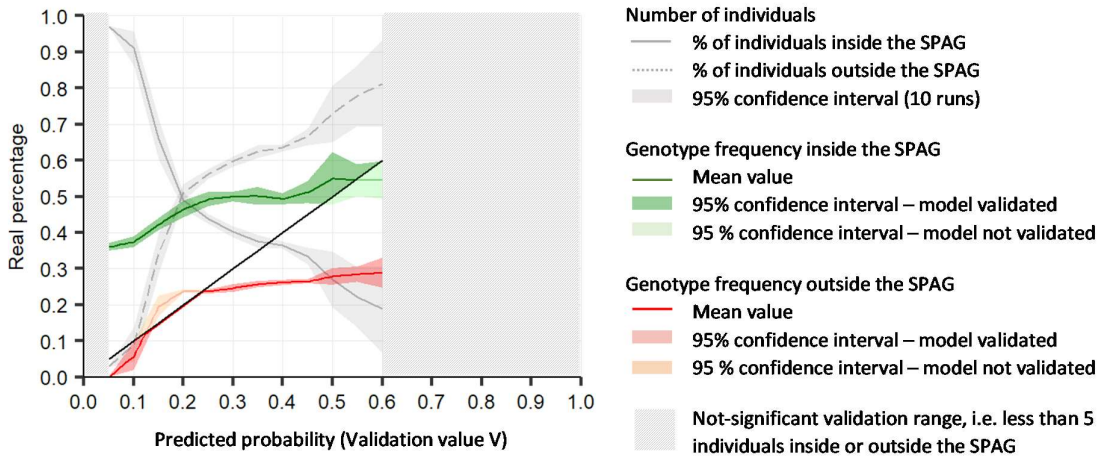
All formulas are explained in more details in Supp. File 3. Note that all multivariate models allow the integration of adaptive genotypes associated with various environmental variables since the environmental variable *xi* used to compute *p(Gi)* can be different for each *i*.

Validation of the models

To test the validity of the models proposed, we selected only 25% of the individuals to compute the SPAG and used the remaining 75% to test it. In order to capture the whole variability of the environmental variable in the training set and thus increase the predictive power of the logistic regression, the range of environmental values was divided into intervals and training individuals were randomly selected within each interval. This selection method is presented in details in Supp. File 3. The testing dataset was then used to validate the result by comparing the genotype frequency predicted by the SPAG with the observed genotype frequency among the testing individuals. We plot the results with graphs as presented in Box 1. We repeated the validation procedure 10 times, with various random selection of training samples in order to compute 95% confidence intervals.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Box 1 : Validation Procedure



For each validation value V (x-axis) between 0 and 1 with a step of 0.1 , we delimited the area where the predicted genotype frequency is higher or equal to V (SPAG≥V). We then calculated the percentage of testing individuals situated within that area (solid grey line) and the observed genotype frequency among them (green line). We computed the same values outside that area, i.e. where SPAG < V (dashed grey line and red line). If the model is valid, the observed frequency inside the SPAG≥V (green line) should effectively be higher or equal to V and it should be lower than V outside (red line). The black line indicates the limit case where the observed genotype frequency is equal to the predicted probability V. The model is therefore valid if the green line remains above the black line and the red line remains below it. In the example above, for a testing value V=0.4, the SPAG is validated since the observed genotype frequency within the SPAG≥0.4 is higher than 0.4 (green line) and is lower than 0.4 outside (red line). Inversely, the model is not validated for V=0.6, since the percentage of presence inside the SPAG≥0.6 is lower than 0.6 (the green line falls under the black line). The model is also not valid for a value V=0.2 since the genotype frequency outside the SPAG≥0.2 is higher than 0.2 (the red line goes above the black line). The green and red areas around the lines indicates the 95% confidence intervals for each line, computed on the basis of the 10 runs with different random selection of training datasets. Finally, the hatched grey areas indicate ranges of testing values where there was less than 5 individuals inside or outside the SPAG, which was therefore considered not to be significant for the validation.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Projections under climate change

In order to predict the genotype frequency optimal for future conditions, we retrieved Worldclim data for the year 2070, corresponding to a strong climate change scenario from the couple Max Planck Institute Earth System Model (MPI-ESM-LR) with a Representative Concentration Pathway equals to 8.5 (RCP 8.5). We then assume that the optimal genotype frequency for future conditions should be close to the genotype frequency currently observed in areas with climatic conditions resembling the future ones. We thus applied the current parameters of the logistic regressions on the future environmental variables in order to derive the future SPAGs for the genotypes of interest. We then study the shift between the current and future SPAG to identify vulnerable populations for which specific genotype frequencies should be much higher so that individuals can adapt to the future conditions.

Results

Population structure

For the Moroccan dataset, the cumulated variance explained by the 10 first PCA components represents only 8.1% of the total variance and the increase in variance explained is almost proportional to the number of components, which highlights that there is no clear sub-structure. We therefore do not include any population structure on the subsequent analysis and computed only logistic regressions without any covariates.

For the European goats, the first component of the PCA explains 6.2% of the total variance when the second, third and fourth components explain 2.0%, 1.7% and 1.6% respectively. The low variance explained by each PCA component indicates no clear population structure. However, since the variance explained by the first component is much higher than what explained by the next ones, it is possible that the first component is partially related to population structure. We therefore computed logistic regressions without covariates and then logistic regressions with a covariate corresponding to the coordinates of goat individuals on the first component of the PCA.

Logistic Regressions

For the Moroccan dataset, more than 483 million of association models were computed. After correction for false discovery rate with a significant threshold of 5%, none models are significant according to the Wald score but seven models are significant according to the G score (Table 1). Three of them indicated an association of SNPs on chromosome 1 with the mean diurnal range (bio2), three others relate SNPs on chromosome 6 to the coefficient of precipitation (bio15) and one to the

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

precipitation of the driest quarter (bio17). None of the SNPs identified correspond to any annotated gene on the goat assembly CHIR 1.0.

Bioclim	Chr	BP	Genotype	GF	β0	β1	G	W	qG
bio2	1	10'309'616	GG	5.59	-6.98	-3.73	41.68	13.63	0.017
bio2	1	10'311'246	CC	5.59	-6.98	-3.73	41.68	13.63	0.017
bio2	1	10'311'252	CC	5.59	-6.98	-3.73	41.68	13.63	0.017
bio17	6	12'244'775	TT	8.07	-6.82	-5.15	40.17	11.59	0.028
bio15	6	12'276'168	AA	21.74	-1.80	1.51	38.74	24.77	0.034
bio15	6	12'285'545	CC	21.74	-1.80	1.51	38.74	24.77	0.034
bio15	6	12'285'617	TT	21.74	-1.80	1.51	38.74	24.77	0.034

Table 1 *:* Significant models obtained for the analysis of Moroccan datasets, considering all bioclimatic variables, after FDR correction with a threshold of 5%. Chr=Chromosome, BP=Position in base pairs, GF=Genotype Frequency, β0 and β1 = parameters of the logistic regression, G=G score (Log Likelihood ratio) of the model, W=Wald score of the model, qG= p-value of G score corrected for FDR

We then investigated in more details the adaptation to the precipitation seasonality (bio15), which is a measure of the variation of monthly precipitation over the year. This variable shows a high variability over the sampling area, with values from 25% to 50% in the Atlas Mountains and the northeast of Morocco (i.e. the standard deviation of monthly precipitation is equal to 25%-50% of the mean of monthly precipitation), values higher than 70% in the coastal areas, and a maximum of more than 100% close to the Sahara. When considering only the associations involving this bioclimatic variable (25’447’348 models), 78 models are significant after FDR-correction of G score, with a significant threshold of 5% (Supp. File 4). Those SNPs are located on seven different genomic regions (Table 2), corresponding to four annotated genes on the reference genome CHIR 1.0: DSG4, CDH2, KCTD1 and WRN. The peak on chromosome 6, containing 33 significant models, does not correspond to any annotated gene. However, two SNPs from the CaprineSNP50 BeadChip fall within this genomic region when mapped on CHIR1.0 and are situated within a long intergenic noncoding RNA (LincRNA) when mapped on the new genome assembly ARS1.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ID	Chr	Start (BP)	End (BP)	Peak (BP)	Geno	GF	G	qG	β0	β1	Genes
M1	6	12'174'332	12'298'321	12'276'168	AA	21.74	38.74	0.004	-1.80	1.51	*(LincRNA)*
M2	13	43'436'394	43'438'732	43'436'394	GG	10.56	29.02	0.042	-3.14	1.80	-
M3	24	19'436'980	19'436'980	19'436'980	CC	76.40	34.75	0.008	1.55	1.29	-
M4	24	25'852'900	25'860'754	25'860'754	AG	38.51	34.75	0.008	-0.29	-1.07	DSG4
M5	24	28'799'029	28'833'762	28'833'253	TT	12.42	27.87	0.046	-2.72	1.58	CDH2
M6	24	30'566'869	30'584'692	30'566'869	TT	2.48	27.99	0.046	-25.69	-15.44	KCTD1
M7	27	25'930'079	25'933'133	25'930'079	GG	78.88	32.88	0.012	1.76	-1.35	WRN

Table 2: Significant models obtained for the analysis of Moroccan datasets with precipitation seasonality (bio15) after FDR correction. Chr=Chromosome, Start=Start in base pairs of the region identified as under selection, End=End in base pairs of the region, Peak SNP = SNP of the most significant model on that region, Geno = corresponding Genotype, GF=corresponding Genotype Frequency, β0 and β1 = parameters of the logistic regression, G=G score (Log Likelihood ratio) of the model, qG=corresponding p-value corrected for FDR, Genes = Annotated genes on the genomic region.

For the European dataset, 2'638'758 associations were computed, among which 44'191 models (1.67%) were significant both without covariate and with the first PCA-component as covariate, according to both G score and Wald score corrected for a false positive rate of 1% (Supp. File 5). The fifteen models with the strongest G scores when computed without covariate correspond to 10 genomic regions (Table 3), associated with three bioclimatic variables related to precipitation (bio13 = precipitation of wettest month, bio16 = precipitation of wettest quarter, bio18 = precipitation of warmest quarter) and two bioclimatic variables related to temperature (bio3 = isothermality, bio8 = mean temperature of wettest quarter). Seven annotated genes correspond exactly to one of the SNPs identified: KRT12, CSN1S2, CACNB, PRDM5, LOC102174324, PALM and NAV3.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ID	ENV	CHR	BP	GENO	GF	qGpop	qWpop	qG0	qW0	β0	β1	Genes
E1a	bio18	19	40696776	GG	33.8	3.3E-09	3.9E-07	3.8E-17	1.2E-11	-1.41	1.30	KRT12
E1b	bio13	19	40696776	GG	33.8	1E-08	7.5E-07	4.1E-17	1.2E-11	-1.40	1.30	KRT12
E1c	bio16	19	40696776	GG	33.8	2.9E-08	1.3E-06	6E-16	2.8E-11	-1.36	1.23	KRT12
E1d	bio13	19	40696776	AA	40.3	4.3E-09	5.4E-07	1.4E-13	6.9E-10	-0.52	-1.07	KRT12
E2	bio18	1	38183832	AA	31.7	3E-09	1.3E-07	1.6E-15	1.2E-11	-0.97	1.13	-
E3a	bio3	6	86081075	CC	31.2	3.5E-11	1.2E-08	3.1E-14	4.2E-11	0.65	-1.06	CSN1S2
E3b	bio18	6	86081075	CC	31.2	9.6E-11	6.9E-08	5.7E-14	5E-10	0.71	1.12	CSN1S2
E4	bio8	13	32300758	GG	31.2	6.1E-10	9E-08	5.1E-14	5.6E-11	0.11	-1.02	CACNB2
E5a	bio18	5	23213822	GG	39.0	6E-10	4.1E-08	7.3E-14	5.6E-11	-0.52	1.02	-
E5b	bio16	5	23213822	GG	39.0	1.3E-09	9.9E-08	1.4E-13	2.3E-10	-0.51	1.02	-
E6	bio8	6	4945809	AA	19.4	9.3E-07	7.9E-05	1E-13	9.2E-08	-2.06	1.55	PRDM5
E7	bio18	16	76397454	GG	22.8	1.1E-10	5.3E-07	1E-13	8E-09	1.30	1.29	LOC102174324
E8	bio18	7	67159272	CC	35.1	6.9E-10	1.3E-07	1E-13	2.3E-10	0.27	1.03	PALM
E9	bio3	5	7093719	GG	39.8	7E-11	1.8E-08	1E-13	9E-11	-0.63	1.02	NAV3
E10	bio18	14	85434737	AA	47.4	2.6E-09	1.6E-07	1.3E-13	2.2E-10	-0.16	-1.01	-

Table 3 Fifteen most significant models (based on G score of the model without covariate) obtained for the analysis of the European dataset, considering all bioclimatic variables, after FDR correction with a threshold of 1% for both G and Wald Score. Chr=Chromosome, BP=Position in base pairs, GENO=Genotype of interest, GF=Genotype Frequency, qGpop (resp. qWpop) = FDR-corrected p-values of Gscore (resp. Wald score) of the model with the first PCA-component as covariate, qG0 (resp. qW0) = FDR-corrected p-values of Gscore (resp. Wald score) of the models without any covariate, β0 and β1 = parameters of the logistic regression without covariate, Genes = Annotated genes on the genomic region.

SPatial Areas of Genotype Probability

Spatial Areas of Genotypes probability were computed for the models presented on tables 2 and 3. Figure 1 show examples of results obtained for the Moroccan goats and Figure 2 for the European ones. Univariate results for all the other models from tables 2 and 3 are available in Supp. File 6.

For the Moroccan dataset, Figure 1A corresponds to the univariate SPAG for the genotype of model M1 (Table 2) that is mainly carried by goats living close to the coast and the Sahara, where high variations of precipitation can be observed. The predicted genotype frequency is the highest in the extreme southwest of the country where the variations of precipitation are the most important and all goats carry this marker. This is not the case in coastal areas, where goats without the variant are living close to goats having it, the predicted genotype frequency in that area being coherently close to 0.5. In the Atlas Mountains and northeast of the country, the predicted variant frequency is much lower (<0.2 in most areas) which corresponds to the observation of an absence of the genotype in

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

most goats sampled in this region. Two other markers positively correlated with bio15 were highlighted by models M3 and M5 (Table 2). However, the simultaneous presence of those two genotypes is very unlikely, as indicated by the I-SPAG (Figure 1B) showing a predicted frequency lower than 0.1 for most part of the territory. Indeed, although those two markers are similarly associated with the environment, only 11% of the sampled goats carry them both. Nevertheless, both variants are absent in the Atlas Mountains and the northeast of the country where variations of precipitation are the lowest, what explains their similar positive correlation with bio15. When considering the probability to find at least one of the genotypes from the three models M1, M3 and M5 all positively associated with the coefficient of precipitation, the U-SPAG (Figure 1C) predicts a very high probability in many parts of the territory. Indeed, 78% of all goats carry at least one of those three genotypes and only some goats living in the northeast of the country have none of them. Finally the K-SPAG presented on Figure 1D shows the probability that goats carry at least 50% of the four variants positively associated with the coefficient of precipitation (M1, M2, M3, M5), i.e. the probability to find at least two of those genotypes. This map is the most contrasted one, with a very high probability of presence close to the coast and Sahara (> 0.9) and a very low frequency (<0.2) in the centre and northeast of the country.

For the European dataset, the univariate SPAG presented (Figure 2A) corresponds to a genotype negatively associated with the precipitation of the warmest quarter (bio18). This variable ranges from less than 100 mm in the Mediterranean border to more than 500 mm in the Swiss Alps, with values between 250 and 300 mm in the Jura, the French Alps and the Swiss Plateau, and between 150 and 200 mm in most of other parts of France and of Northern Italy. Consequently, the predicted genotype frequency is the lowest in the Swiss Alps (<0.1), slightly higher in Jura, French Alps and Swiss Plateau (between 0.2 and 0.4) and higher than 0.5 everywhere else, with a maximum around 0.8 in the Mediterranean border. We then computed the I-SPAG (Figure 2B) of this same variant (model E10) with the model E1d corresponding to a genotype AA associated with low precipitation in the wettest month (bio13). For this last model, the genotype GG of the same SNP shows the strongest positive association with bio18 (model E1a). The two models E1d and E10 may therefore correspond to a similar adaptation to low values of precipitation during the warmest quarter. However, the I-SPAG indicates that their simultaneous presence is not very likely (predicted frequency < 0.6 everywhere). Indeed, 23% of the total goats carry those two genotypes, but they are essentially localised in areas where other goats carry only one of the two markers. The predicted probability is the highest in the centre-north of France (regions Centre, Iles de France, East of Pays de la Loire, Normandie and South of Hauts de France, see Sup. File 1 for regions' map) and in the southern part (Occitanie and West of Provence), whereas in the Alps, Jura and most part of Switzerland, no goats carry the two genotypes and the

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

predicted frequency is close to 0. The probability to find at least one of those two variants, presented on the U-SPAG in Figure 2C, indicates a trend similar to the probability of presence of E10 alone (Figure 2A), but with even stronger contrast between the Alps-Jura-Switzerland area (frequencies < 0.3) and the rest of the territory (frequencies > 0.7). Finally the K-SPAG (Figure 2D) shows the probability to find at least 50% of five genotypes negatively associated with the precipitation of the warmest quarter (bio18), i.e. the probability to find at least three of them. Note that for the models positively associated with bio18 in Table 3, we used the alternative genotype that was the most significantly negatively correlated with bio18 (indicated after the model ID). The resulting SPAG is very close to the I-SPAG of E1d and E10 (Figure 2B).

For all cases presented, the validation graphs indicate that the SPAGs computed with 25% of the individuals generally correctly predict the genotype frequency of the 75% remaining individuals. However, for the Moroccan dataset, the U-SPAG (Figure 1C) tends to slightly overestimate the probability of presence since the observed genotype frequency outside the SPAG (red line) tends to be slightly higher than the predicted value (black line).

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

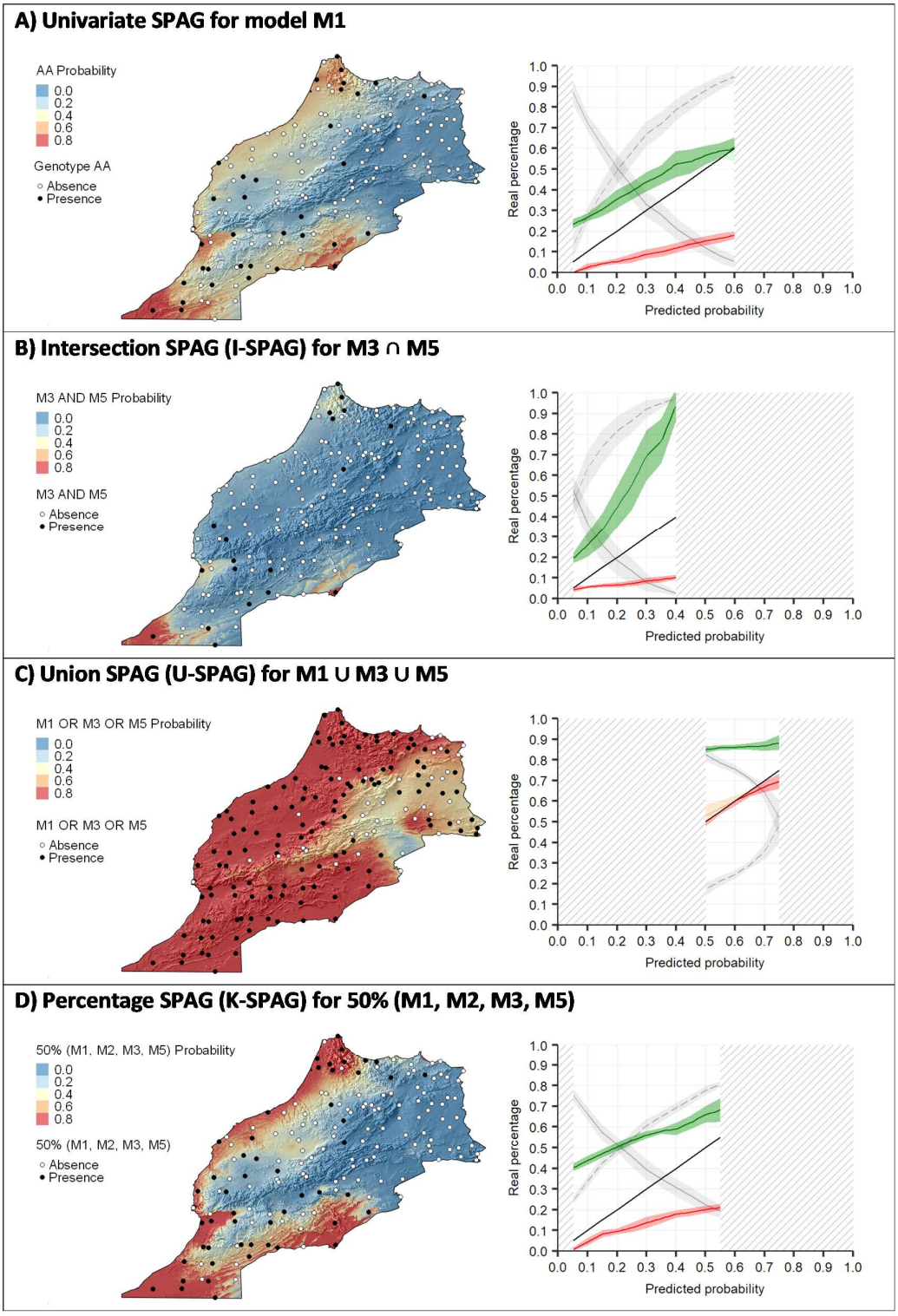


Figure 1 : Univariate and Multivariate Spatial Areas of Genotypes Probability for the Moroccan dataset. The identifiers of the presented models (M1, M2, M3, M5) refers to Table 2. The maps show the average genotype(s) frequency(ies) based on the 10 runs computed with different random selection of training sets containing 25% of the total number of individuals. Please refer to Box 1 to interpret the validation graphs shown on the right of each map.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

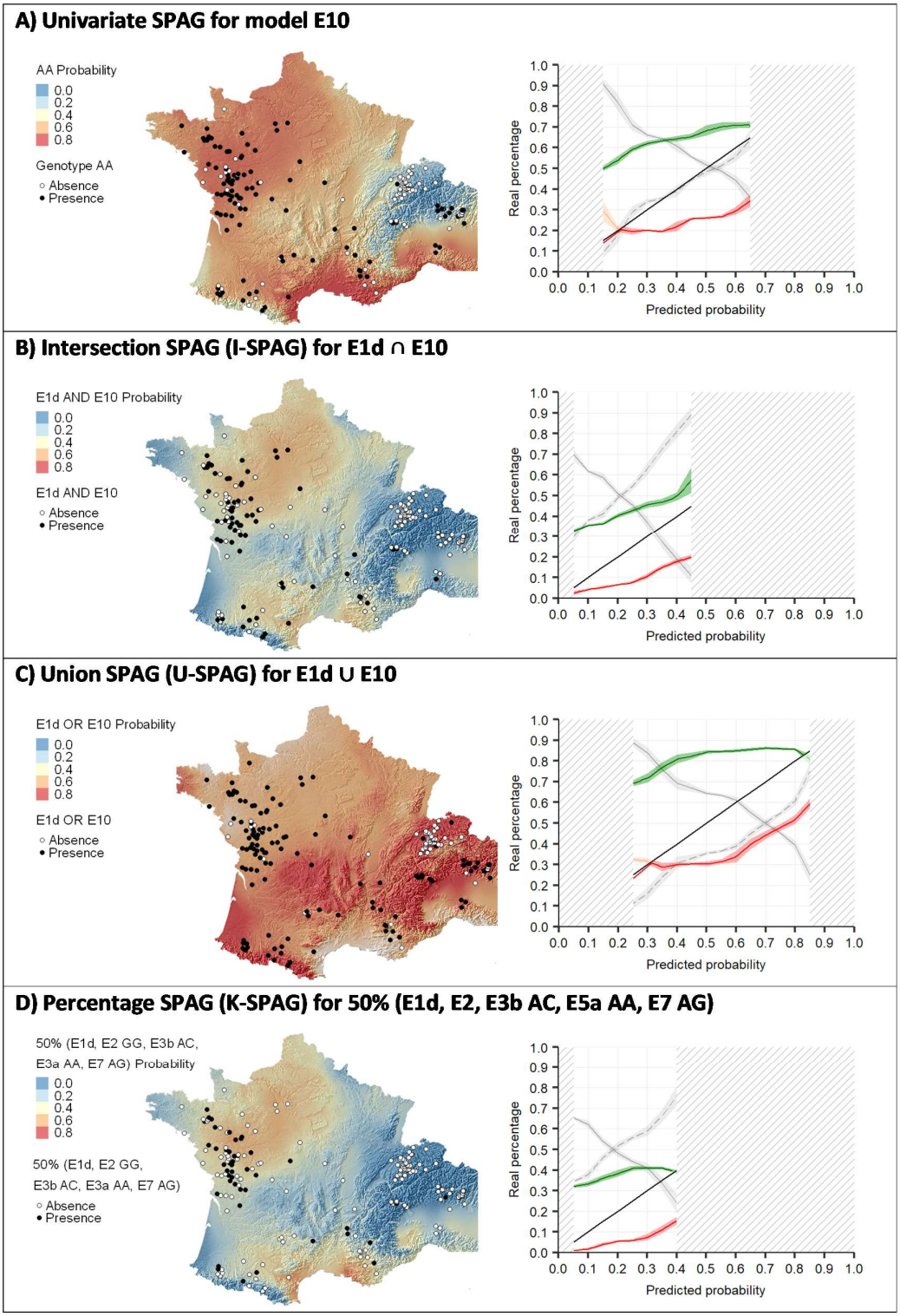


Figure 2 : Univariate and Multivariate Spatial Areas of Genotypes Probability for the European dataset. The identifiers of the presented models (E10, E1d, E2, E3b, E5a, E7) refers to Table 3. The maps show the average genotype(s) frequency(ies) based on the 10 runs computed with different random selection of training sets containing 25% of the total number of individuals. Note that since up to five individuals can be localised on the same site, a black dot indicates a presence if at least 50% of the individuals of the site carry the marker(s). Please refer to Box 1 to interpret the validation graphs shown on the right of each map.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

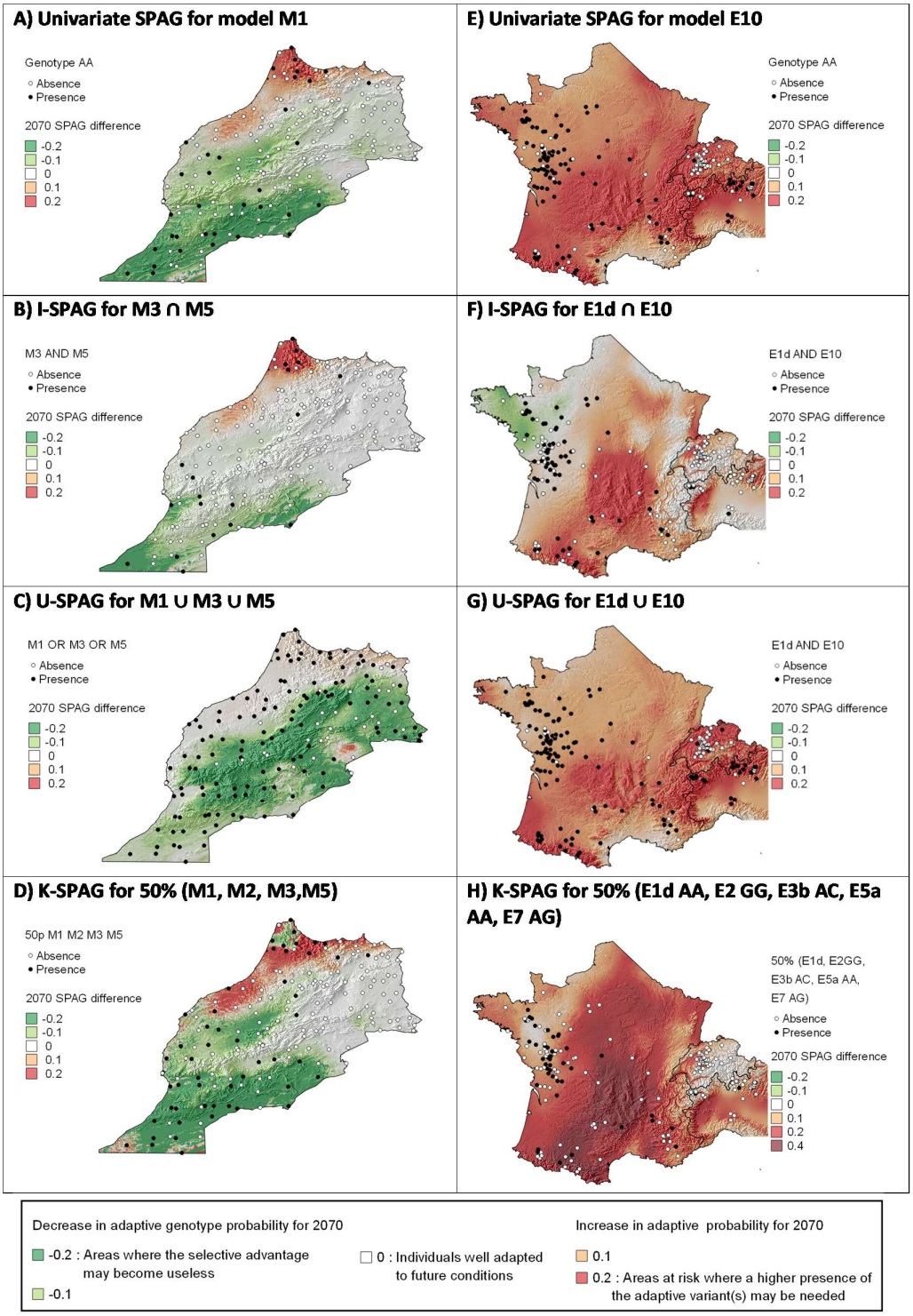


Figure 3: Predicted SPAG difference for 2070 considering the MPI-ESM-LR climate change scenario with RCP 8.5, for the Moroccan (left) and European (right) datasets. The identifiers of the presented models (M1, M2, M3, M5, E10, E1d, E2, E3b, E5a, E7) refer to Table 2 (Moroccan dataset) and Table 3 (European dataset). The maps show the average difference in genotype(s) frequency(ies) based on the 10 runs computed with different random selection of training sets containing 25% of the total number of individuals.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Projections under climate change

Figure 3 shows the differences between the current SPAGs presented in Figures 1 and 2 and their corresponding projections for 2070. In Morocco, the precipitation seasonality (bio15) is predicted to increase in the northwest of the country, with a maximum increase of 5 to 10 in the extreme northwestern region (Tangier-Tetouan, see region's map in Supp. File 1) and to decrease in other areas, especially in the Atlas Mountains and close to the Sahara (from -10 to -20). The evolution of the univariate SPAG for model M1 (Figure 3A) consequently indicates the highest risk in the Tangier-Tetouan area, where the genotype frequency should be approximately 20% higher to be well adapted to future conditions. However, many individuals in this area already carry the favourable genotype and the risk for the population may therefore be reduced thanks to natural gene flow. Nevertheless, this is not the case in the southwest of that area (Rabat, Casablanca) where the genotype frequency should be 10-20% higher and none of the goats sampled there currently carry the adaptive variant. Similar observations can be made as regards the I-SPAG of M3 and M5 (Figure 3B), two other markers that potentially confer an adaptation to high variations of precipitation. However, the U-SPAG (Figure 3C) highlights no vulnerable areas, which indicates that if the presence of at least one of the adaptive variants is sufficient to enable the adaptation to high variations of precipitation, no populations are threatened. Finally the K-SPAG (Figure 3D) shows once again a risk area in the northwest of the country, where the probability to carry at least 50% of the adaptive variants should be approximately 20% higher. Again, individuals in the most northern part of this risk area may be less threatened thanks to the close presence of goats that already carry the favourable genotypes, whereas the population from the Rabat-Casablanca area may be more threatened due to the current much lower presence of the adapted variants.

For the European dataset, the model E10 presented in Figure 2A was associated with low precipitation of the warmest quarter (bio 18), which is predicted to decrease from 20 to 180 mm over the whole study area until 2070. The predicted precipitation loss is maximum in the Alps (-120 to -180), in the Mediterranean coast (-100 to -130) and in the centre-south of France (West of Auvergne and East of Nouvelle Acquitaine, -90 to -120 mm, see region's map in Supp. File 1). As a consequence, the evolution of the univariate SPAG for model E10 (Figure 3E) indicates that all populations may be vulnerable, since the genotype frequency should be 10 to 20% higher everywhere. Even though the potentially adaptive genotype is already present in most French goats (75% carry it), this is not the case for much of the goats populations from the Alps and Switzerland that currently lack the adaptive variant and may be particularly vulnerable. When considering the I-SPAG of E10 and E1d (Figure 3F), a higher risk is highlighted in the centre-south of France (West of Auvergne and its surrounding), northeast of France (Alsace in East of Grand-Est), Swiss Plateau, and northwest of Italy. Our dataset does not contain any

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

goat sampled in Alsace and northwest of Italy. However, in centre-south of France and Swiss-Plateau, the genotypes are almost never simultaneously present in the goats sampled and the populations may therefore be threatened. The U-SPAG of the same genotypes (Figure 3G) shows results very similar to the univariate SPAG for E10 alone (Figure 3E). If the presence of at least one of those two genotypes may be sufficient to allow an adaptation to low precipitation, the region the most threatened will then once again be the Alps and Switzerland, where the two markers are currently absent. Finally, the evolution of the K-SPAG (Figure 3H) shows a high risk in a large part of France and north of Italy, where the genotype frequency should be more than 20% higher. The risk is particularly high in the centre and south of France (Auvergne, Occitanie) and populations from Auvergne may be particularly threatened due to the current low genotype frequencies on the goats sampled there.

Discussion

Signature of adaptive potential

In Morocco, very few models are significant when considering all bioclimatic variables, which may be due to the huge number of associations computed, resulting in a strong FDR-correction. When considering the precipitation seasonality only, we identified seven genomic regions potentially under selection and corresponding to four annotated genes. Among them, DSG4 and KCTD1 may be related to the development of hair properties. **DSG4** is involved in the creation of a desmosomal cadherin found in epidermal and hair follicles cells (Kljuic *et al.*, 2003) and has been reported to play a role in the development of wool properties in sheep (Ling *et al.*, 2014) and in the regulation of keratin in goats (Wang *et al.*, 2017; Zhang *et al.*, 2019). **KCTD1** codes for a protein interacting with the β-catenin involved in the regulation of the WNT signalling pathway (Li *et al.*, 2014), which acts in the development of hair follicle in animals (Galbraith, 2010). Another gene identified, the **N-cadherin** (CDH2), may be related to skin properties since it encodes for a protein which mediates cell-cell adhesion and may interact with melanocytes (Hayashi *et al.*, 2007). Those results suggest that goats confronted with high variations of precipitation may have developed an adaptation related to hair or skin properties, which could for example ensure a better water repulsion. In addition, the genomic region containing the highest number of significant associations (Supp. File 3) is located on chromosome 6, in a genomic area corresponding to a long intergenic noncoding RNAs (LincRNA) on the new assembly ARS1. Although we do not have any indication regarding the function of this LincRNA, intergenic non-coding RNA have also been reported to be implicated in hair follicle growth or skin pigmentation in goats (Ren *et al.*, 2016; Wang *et al.*, 2017; Ge *et al.*, 2018). The potential adaptive function of the last gene identified, **WRN**, is less evident since this gene codes for a protein associated with the Werner syndrome responsible for premature and accelerated aging (Goto, 1997).

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Since this disease notably includes an early greying of the hair, the WRN gene could interact with some actors of the hair development, but it may also likely be a false positive.

For the European dataset, the number of significant models was much higher, which may be explained by 1) a less strong FDR-correction due to a lower number of associations computed, 2) the pre-selection of SNPs with a chip purposely chosen to represent selected traits and 3) a larger study area with a higher number of different breeds analysed. From the fifteen most significant models, seven candidate genes for natural selection were identified, among which one LOC gene of unknown function, associated with the precipitation of warmest quarter and thus potentially conferring an adaptation to drought. Two of the other genes are related to properties of the cornea: **KRT12** is involved in the creation of a protein forming molecules of the corneal epithelium (Kao *et al.*, 1996) and **PRDM5** is implicated in the regulation of the fibrillary collagen needed for the maintain of the cornea resistance (Burkitt Wright *et al.*, 2011, p. 5). This last gene may also confer an adaptive advantage to pigs living in high altitudes (Ai *et al.*, 2014). The SNPs corresponding to those genes were significantly associated with the precipitation of the warmest quarter and the mean temperature of the wettest quarter respectively. They may therefore highlight an adaptation to drought conditions that can be correlated with higher UV-radiation potentially influencing cornea properties. Among the other genes identified, **CSN1S2** is related to the casein content of the milk (Ramunno *et al.*, 2001) and has already been reported as a gene potentially under selection in goats (Pariset *et al.*, 2009; Bertolini *et al.*, 2018). The SNP associated with this gene was correlated with the isothermality (bio3) and the precipitation of the warmest month (bio18) and may therefore relate to a specific property of milk resulting either from an adaptation to drought conditions or to an artificial selection. Another identified gene **CACNB2** was also previously identified as a potential candidate for selection in goats (Brito *et al.*, 2017; Mdladla *et al.*, 2018). This gene is involved in the calcium channel and energy pathway and may confer an advantage for adaptation to cold temperature in Siberian human populations (Cardona *et al.*, 2014). It corresponds here to a variant negatively correlated with the mean temperature of the wettest quarter and may therefore also confer an adaptive potential for goats confronted with cold habitat. Finally, the two last genes identified may be related to skin properties. The **PALM** gene codes for a protein involved in the cell formation and potentially related to pigmentation (Kutzleb *et al.*, 1998), whereas, **NAV3** may be involved in the development of cutaneous T-cells lymphoma (Karenko *et al.*, 2005, p. 3).

Many of the genes highlighted on the two case studies may therefore be associated with a function that can be influenced by climate, which reinforces the potential that they are true signatures of local adaptation. However, although previous studies show the power of genotype-environment associations to predict phenotype (Lasky *et al.*, 2015; Vangestel *et al.*, 2018) or fitness (Fournier-Level *et al.*, 2011; Hancock *et al.*, 2011), more investigations are needed to verify that the variants identified

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

are really conferring an adaptive advantage and that they are heritable (Funk *et al.*, 2019). Moreover, adaptation process is complex and often involves polygenic traits (Pritchard and Rienzo, 2010), for which the detection power of the genotype-environment associations may be much reduced (Villemereuil *et al.*, 2014; Harrisson *et al.*, 2014) . In that case, it may be advisable to use multivariate genotype-environment association models (Forester *et al.*, 2017) or to integrate other methods to identify SNPs related to polygenic adaptation (Zhou *et al.*, 2013; Lasky *et al.*, 2015).

From SPAG to conservation

The first utility of the SPAGs is to quantify the current probability to find beneficial alleles or the expression of favourable traits in plant and animal populations, even in regions where no individuals have been sampled. Our results show that with few training individuals (i.e. 40 goats in Morocco and 120 in Europe), a good estimation of the genotype frequency is possible, for one or many genotypes of interest. The univariate models presented here were already applied to map the genotype frequencies of adaptive variants of the Scandinavian brown bears (Joost, 2006), Moroccan sheep (Rochat et al., 2016) and coral reefs from the Ryukyu Archipelago (Selmoni et al., 2019). The multivariate models are presented here for the first time and according to the validation procedure applied, they appear to be powerful in estimating the combined probability to find many genotypes potentially correlated with different environmental variables. With the I-SPAG, the resulting probabilities may rapidly become very low, but this model could be used when we suspect that the simultaneous presence of some adaptive genotypes is needed to ensure the adaptation or when we would like to highlight the probability to find simultaneously variants that may confer an adaptive response to different environmental variables (for example low precipitation and high temperature). At the other extreme, the U-SPAG may rapidly indicate high probabilities of presence in most parts of the territory, but it can be used when suspecting that the presence of at least one of the variants may be sufficient to confer the adaptive potential. Since it is usually difficult to know if the simultaneous presence of variants is needed or if an union is sufficient, the K-SPAG offers an interesting compromise, allowing the identification of populations that retain a given percentage of variants, which may allow the delimitation of areas where there is the highest probability to find goats with a high adaptive potential.

On the other hand, the study of the shift in SPAG under climate change conditions can help identify 1) well-adapted populations, where individuals currently show adaptive variants that seem to be optimal under future conditions, 2) populations at risk where the current genotype frequency is not optimal, but where the favourable variants are already present in the population, thus potentially allowing a natural increase of the genotype frequency through gene flow and 3) threatened populations where optimal variants are currently missing but would be needed in order to ensure an adaptation to future

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

climate. Those identifications may be of high utility in conservation planning. Indeed, when priorizing areas for conservation, the success may be increased by choosing to preserve preadapted individuals that already carry functional variants conferring a good adaptation to future climate (Orr and Unckless, 2008). Moreover SPAG can also be used to prevent translocation of individuals that currently do not carry the variants favourable for the future conditions of the target site, which would result in a reduction or loss of the adaptive potential of target populations (Weeks *et al.*, 2011). In addition, conservation plans may be developed to increase the survival capacity of threatened populations. This can imply assisted gene flow to import the adaptive variants in a population where they are missing (Aitken and Whitlock, 2013; Kelly and Phillips, 2016) or artificial selection of individuals already pre-adapted to future conditions (Hoffmann, 2010). However, this has to be undertaken carefully since the selection of locally adapted individuals can result in a loss of genetic diversity (Savage *et al.*, 2018), which may decrease the potential of populations to adapt to new environmental changes. Kardos and Shafer (2018) therefore proposed that gene-targeted conservation measures should be done only with traits affecting vital processes of the species and when phenotypic variations are large enough to ensure a high probability of success. Moreover, since the results of the shift under climate change may be highly dependent on the climate change scenario considered, computations should be made with various scenarios and less weight should be given to the conclusions not consistent within scenarios (Reside *et al.*, 2018). Finally, an analysis of connectivity should be done to highlight the potential of natural gene flow to increase the genotype frequency.

The Moroccan case study highlighted that the goats' populations from the surroundings of Rabat and Casablanca may lack adaptive variants potentially conferring an advantage to face high variation of precipitation. If the adaptive role of those genotypes is validated, the goat populations of this area may be threatened. Due to the high economic and social importance of goats in Morocco, it is crucial to preserve viable populations. Indeed, in this country, agriculture contributes to 12 to 24% of the national GDP and employs 40% of the total active population (Boujenane, 2005). Livestock, especially small ruminants, is the most important sector in agriculture and goat farms represent 20% of the total number of agricultural farms (Boujenane, 2005). It is therefore important to consider preserving or introducing the adaptive variant on each vulnerable population. This could be done for example by favouring crossbreeding with individuals from the southern or north-eastern part of the country, where the adaptive genotypes are currently well present and avoiding breeding or translocation with exotic goats or goats from the Atlas or Oriental areas. In the Northern part of Morocco (Tanger-Tetouan regions), goats population represent 12% of the national goats (Chentouf, 2014) and play an important role to preserve food security (Godber *et al.*, 2016). In that area, crossbreeding with exotic breeds have been introduced to improve milk production (Boujenane, 2005; Godber *et al.*, 2016).

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

However, our result show that the genotype frequency of adaptive variants should increase in the goat populations from this region and that it is therefore of main importance to maintain there local individuals that show the necessary adaptive variants.

Finally, the SPAG maps could be integrated into decision frameworks considering the adaptive potential when defining the vulnerability of species (Bonin *et al.*, 2007; Williams *et al.*, 2008; Sgrò *et al.*, 2011; Dawson *et al.*, 2011; Razgour *et al.*, 2018) or in more global decision frameworks considering other factors of vulnerability such as predation level or habitat loss. Moreover, SPAG could also be used to predict the presence of genotype(s) associated with other pressures showing a spatial distribution, such as the presence of a parasite (Vajana *et al.*, 2018) or a predator (Cousyn *et al.*, 2001) or the urbanization level (Harris and Munshi-South, 2016) . Very similar models can also be derived to predict allele frequencies instead of genotypes frequencies or to integrate other covariates (for example to account for autocorrelation). SPAG therefore constitute a valuable tool to support conservation decisions, especially under current changing climatic conditions.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

Ai H, Yang B, Li J, Xie X, Chen H, Ren J (2014). Population history and genomic signatures for high-altitude adaptation in Tibetan pigs. *BMC Genomics* **15**: 834.

Aitken SN, Whitlock MC (2013). Assisted Gene Flow to Facilitate Local Adaptation to Climate Change. *Annual Review of Ecology, Evolution, and Systematics* **44**: 367–388.

Alberto FJ, Boyer F, Orozco-terWengel P, Streeter I, Servin B, Villemereuil P de, *et al.* (2018). Convergent genomic signatures of domestication in sheep and goats. *Nat Commun* **9**: 1–9.

Allendorf FW, Hohenlohe PA, Luikart G (2010). Genomics and the future of conservation genetics. *Nat Rev Genet* **11**: 697–709.

Allendorf FW, Leary RF (1986). Heterozygosity and fitness in natural populations of animals. *Conservation biology: the science of scarcity and diversity*: 57–76.

Banta JA, Ehrenreich IM, Gerard S, Chou L, Wilczek A, Schmitt J, *et al.* (2012). Climate envelope modelling reveals intraspecific relationships among flowering phenology, niche breadth and potential range size in Arabidopsis thaliana. *Ecology Letters* **15**: 769–777.

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 (Methodological)* **57**: 289–300.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Benjelloun B, Alberto FJ, Streeter I, Boyer F, Coissac E, Stucki S, *et al.* (2015). Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (Capra hircus) using WGS data. *Front Genet* **6**.

Bertolini F, Servin B, Talenti A, Rochat E, Kim ES, Oget C, *et al.* (2018). Signatures of selection and environmental adaptation across the goat genome post-domestication. *Genetics Selection Evolution* **50**: 57.

Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, *et al.* (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat Genet* **49**: 643–650.

Bonin A, Nicole F, Pompanon F, Miaud C, Taberlet P (2007). Population Adaptive Index: a New Method to Help Measure Intraspecific Genetic Diversity and Prioritize Populations for Conservation. *Conservation Biology* **21**: 697–708.

Boujenane I (2005). *Small Ruminant Breeds of Morocco*.

Brito LF, Kijas JW, Ventura RV, Sargolzaei M, Porto-Neto LR, Cánovas A, *et al.* (2017). Genetic diversity and signatures of selection in various goat breeds revealed by genome-wide SNP markers. *BMC Genomics* **18**: 229.

Burkitt Wright EMM, Spencer HL, Daly SB, Manson FDC, Zeef LAH, Urquhart J, *et al.* (2011). Mutations in PRDM5 in Brittle Cornea Syndrome Identify a Pathway Regulating Extracellular Matrix Development and Maintenance. *The American Journal of Human Genetics* **88**: 767–777.

Cardona A, Pagani L, Antao T, Lawson DJ, Eichstaedt CA, Yngvadottir B, *et al.* (2014). Genome-Wide Analysis of Cold Adaptation in Indigenous Siberian Populations. *PLOS ONE* **9**: e98076.

Chentouf M (2014). Systèmes de production caprine au Nord du Maroc: Contraintes et propositions d'amélioration. *Opt Médit A: Médit Semin* **108**: 25–32.

Cousyn C, Meester LD, Colbourne JK, Brendonck L, Verschuren D, Volckaert F (2001). Rapid, local adaptation of zooplankton behavior to changes in predation pressure in the absence of neutral genetic changes. *PNAS* **98**: 6256–6260.

Dawson TP, Jackson ST, House JI, Prentice IC, Mace GM (2011). Beyond Predictions: Biodiversity Conservation in a Changing Climate. *Science* **332**: 53–58.

Dong Y, Xie M, Jiang Y, Xiao N, Du X, Zhang W, *et al.* (2013). Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nature Biotechnology* **31**: 135–141.

Duputié A, Rutschmann A, Ronce O, Chuine I (2015). Phenological plasticity will not help all species adapt to climate change. *Global Change Biology* **21**: 3062–3073.

Fitzpatrick MC, Keller SR (2015). Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters* **18**: 1–16.

Forester BR, Lasky JR, Wagner HH, Urban DL (2017). Using genotype-environment associations to identify multilocus local adaptation. *bioRxiv*: 129460.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fournier-Level A, Korte A, Cooper MD, Nordborg M, Schmitt J, Wilczek AM (2011). A Map of Local Adaptation in Arabidopsis thaliana. *Science* **334**: 86–89.

Funk WC, Forester BR, Converse SJ, Darst C, Morey S (2019). Improving conservation policy with genomics: a guide to integrating adaptive potential into U.S. Endangered Species Act decisions for conservation practitioners and geneticists. *Conserv Genet* **20**: 115–134.

Funk WC, McKay JK, Hohenlohe PA, Allendorf FW (2012). Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution* **27**: 489–496.

Galbraith H (2010). Fundamental hair follicle biology and fine fibre production in animals. *animal* **4**: 1490–1509.

Garzón MB, Robson TM, Hampe A (2019). ΔTraitSDMs: species distribution models that account for local adaptation and phenotypic plasticity. *New Phytologist* **222**: 1757–1765.

Ge W, Wang S-H, Sun B, Zhang Y-L, Shen W, Khatib H, *et al.* (2018). Melatonin promotes Cashmere goat (Capra hircus) secondary hair follicle growth: a view from integrated analysis of long non-coding and coding RNAs. *Cell Cycle* **17**: 1255–1267.

Godber OF, Laroussi BF, Chentouf M, Wall R (2016). Intensification of Mediterranean Goat Production Systems: A Case Study in Northern Morocco. *Agriculture* **6**: 16.

Goto M (1997). Hierarchical deterioration of body systems in Werner's syndrome: Implications for normal ageing. *Mechanisms of Ageing and Development* **98**: 239–254.

Hällfors MH, Liao J, Dzurisin J, Grundel R, Hyvärinen M, Towle K, *et al.* (2016). Addressing potential local adaptation in species distribution models: implications for conservation under climate change. *Ecological Applications* **26**: 1154–1169.

Hancock AM, Brachi B, Faure N, Horton MW, Jarymowycz LB, Sperone FG, *et al.* (2011). Adaptation to Climate Across the Arabidopsis thaliana Genome. *Science* **334**: 83–86.

Harris SE, Munshi-South J (2016). Scans for positive selection reveal candidate genes and local adaptation of Peromyscus leucopus populations to urbanization. *bioRxiv*: 038141.

Harrisson KA, Pavlova A, Telonis-Scott M, Sunnucks P (2014). Using genomics to characterize evolutionary potential for conservation of wild populations. *Evol Appl* **7**: 1008–1025.

Hayashi R, Yamato M, Sugiyama H, Sumide T, Yang J, Okano T, *et al.* (2007). N-Cadherin Is Expressed by Putative Stem/Progenitor Cells and Melanocytes in the Human Limbal Epithelial Stem Cell Niche. *STEM CELLS* **25**: 289–296.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25**: 1965–1978.

Hijmans RJ, van Etten J (2012). raster: Geographic analysis and modeling with raster data. R package version 2.0-12. http://CRAN.R-project.org/package=raster.

Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, *et al.* (2016). Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. *The American Naturalist* **188**: 379–397.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Hoelzel AR, Bruford MW, Fleischer RC (2019). Conservation of adaptive potential and functional diversity. *Conserv Genet* **20**: 1–5.

Hoffmann I (2010). Climate change and the characterization, breeding and conservation of animal genetic resources. *Animal Genetics* **41**: 32–46.

Hoffmann AA, Sgrò CM (2011). Climate change and evolutionary adaptation. *Nature* **470**: 479–485.

Hoffmann AA, Willi Y (2008). Detecting genetic responses to environmental change. *Nat Rev Genet* **9**: 421–432.

Hughes L (2000). Biological consequences of global warming: is the signal already apparent? *Trends in Ecology & Evolution* **15**: 56–61.

Ikeda DH, Max TL, Allan GJ, Lau MK, Shuster SM, Whitham TG (2017). Genetically informed ecological niche models improve climate change predictions. *Global Change Biology* **23**: 164–176.

IPCC (2014). Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.

Jombart T (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**: 1403–1405.

Joost S (2006). The geographical dimension of genetic diversity : a GIScience contribution for the conservation of animal genetic resources.

Joost S, Bonin A, Bruford MW, Després L, Conord C, Erhardt G, *et al.* (2007). A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology* **16**: 3955–3969.

Joost S, Vuilleumier S, Jensen JD, Schoville S, Leempoel K, Stucki S, *et al.* (2013). Uncovering the genetic basis of adaptive change: on the intersection of landscape genomics and theoretical population genetics. *Mol Ecol* **22**: 3659–3665.

Kao WW, Liu CY, Converse RL, Shiraishi A, Kao CW, Ishizaki M, *et al.* (1996). Keratin 12-deficient mice have fragile corneal epithelia. *Invest Ophthalmol Vis Sci* **37**: 2572–2584.

Kardos M, Shafer ABA (2018). The Peril of Gene-Targeted Conservation. *Trends in Ecology & Evolution* **33**: 827–839.

Karenko L, Hahtola S, Päivinen S, Karhu R, Syrjä S, Kähkönen M, *et al.* (2005). Primary Cutaneous T-Cell Lymphomas Show a Deletion or Translocation Affecting NAV3, the Human UNC-53 Homologue. *Cancer Res* **65**: 8101–8110.

Kelly E, Phillips BL (2016). Targeted gene flow for conservation. *Conservation Biology* **30**: 259–267.

Kljuic A, Bazzi H, Sundberg JP, Martinez-Mir A, O'Shaughnessy R, Mahoney MG, *et al.* (2003). Desmoglein 4 in Hair Follicle Differentiation and Epidermal Adhesion: Evidence from Inherited Hypotrichosis and Acquired Pemphigus Vulgaris. *Cell* **113**: 249–260.

Kolmogorov A (1956). Foundations of the Theory of Probability. *Chelsea Pub Co*.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Kutzleb C, Sanders G, Yamamoto R, Wang X, Lichte B, Petrasch-Parwez E, *et al.* (1998). Paralemmin, a Prenyl-Palmitoyl–anchored Phosphoprotein Abundant in Neurons and Implicated in Plasma Membrane Dynamics and Cell Process Formation. *The Journal of Cell Biology* **143**: 795–813.

Lasky JR, Upadhyaya HD, Ramu P, Deshpande S, Hash CT, Bonnette J, *et al.* (2015). Genome-environment associations in sorghum landraces predict adaptive traits. *Science Advances* **1**: e1400218.

Li X, Chen C, Wang F, Huang W, Liang Z, Xiao Y, *et al.* (2014). KCTD1 Suppresses Canonical Wnt Signaling Pathway by Enhancing β-catenin Degradation. *PLOS ONE* **9**: e94343.

Li J, Li H, Jakobsson M, Li S, Sjödin P, Lascoux M (2012). Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Molecular Ecology* **21**: 28–44.

Ling YH, Xiang, H, Zhang G, Ding JP, Zhang ZJ, Zhang YH, *et al.* (2014). Identification of complete linkage disequilibrium in the DSG4 gene and its association with wool length and crimp in Chinese indigenous sheep. *GMR | Genetics and Molecular Research | The Original by FUNPEC-RP* **13**: 5617–5625.

Mable BK (2019). Conservation of adaptive potential and functional diversity: integrating old and new approaches. *Conserv Genet* **20**: 89–100.

Mdladla K, Dzomba EF, Muchadeyi FC (2018). Landscape genomics and pathway analysis to understand genetic adaptation of South African indigenous goat populations. *Heredity* **120**: 369–378.

Nicotra AB, Beever EA, Robertson AL, Hofmann GE, O'Leary J (2015). Assessing the components of adaptive capacity to improve conservation and management efforts under global change. *Conservation Biology* **29**: 1268–1278.

Opdam P, Wascher D (2004). Climate change meets habitat fragmentation: linking landscape and biogeographical scale levels in research and conservation. *Biological Conservation* **117**: 285–297.

Orr HA, Unckless RL (2008). Population Extinction and the Genetics of Adaptation. *The American Naturalist* **172**: 160–169.

Pariset L, Joost S, Marsan PA, Valentini A, Econogene Consortium (EC) (2009). Landscape genomics and biased FST approaches reveal single nucleotide polymorphisms under selection in goat breeds of North-East Mediterranean. *BMC Genet* **10**: 7.

Pauls SU, Nowak C, Bálint M, Pfenninger M (2013). The impact of global climate change on genetic diversity within populations and species. *Mol Ecol* **22**: 925–946.

Phillips SJ, Anderson RP, Schapire RE (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling* **190**: 231–259.

Pritchard JK, Rienzo AD (2010). Adaptation – not by sweeps alone. *Nat Rev Genet* **11**: 665–667.

R Development Core Team (2008). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Ramunno L, Cosenza G, Pappalardo M, Longobardi E, Gallo D, Pastore N, *et al.* (2001). Characterization of two new alleles at the goat CSN1S2 locus. *Animal Genetics* **32**: 264–268.

Razgour O, Forester B, Taggart JB, Bekaert M, Juste J, Ibáñez C, *et al.* (2019). Considering adaptive genetic variation in climate change vulnerability assessment reduces species range loss projections. *PNAS* **116**: 10418–10423.

Razgour O, Taggart JB, Manel S, Juste J, Ibáñez C, Rebelo H, *et al.* (2018). An integrated framework to identify wildlife populations under threat from climate change. *Molecular Ecology Resources* **18**: 18–31.

Reed DH, Frankham R (2001). How Closely Correlated Are Molecular and Quantitative Measures of Genetic Variation? A Meta-Analysis. *Evolution* **55**: 1095–1103.

Reed TE, Schindler DE, Waples RS (2011). Interacting Effects of Phenotypic Plasticity and Evolution on Population Persistence in a Changing Climate. *Conservation Biology* **25**: 56–63.

Ren H, Wang G, Chen L, Jiang J, Liu L, Li N, *et al.* (2016). Genome-wide analysis of long non-coding RNAs at early stage of skin pigmentation in goats (Capra hircus). *BMC Genomics* **17**: 67.

Reside AE, Butt N, Adams VM (2018). Adapting systematic conservation planning for climate change. *Biodivers Conserv* **27**: 1–29.

Rischkowsky B, Pilling D (2007). *The state of the world's animal genetic resources for food and agriculture*. Food & Agriculture Org.

Rochat E, Leempoel K, Vajana E, Colli L, Ajmone-Marsan P, Joost S, *et al.* (2016). Map of genotype frequency change in autochtonous Moroccan sheep breeds due to global warming.

Savage AE, Mulder KP, Torres T, Wells S (2018). Lost but not forgotten: MHC genotypes predict overwinter survival despite depauperate MHC diversity in a declining frog. *Conserv Genet* **19**: 309–322.

Scherf B, Rischkowsky B, Hoffmann I, Wieczorek M, Montironi A, Cardellino R (2008). Livestock Genetic Diversity in Dry Rangelands. In: Lee C, Schaaf T (eds) *The Future of Drylands*, Springer Netherlands, pp 89–100.

Schoville SD, Bonin A, François O, Lobreaux S, Melodelima C, Manel S (2012). Adaptive Genetic Variation on the Landscape: Methods and Cases. *Annual Review of Ecology, Evolution, and Systematics* **43**: 23–43.

Selmoni O, Rochat E, Lecellier G, Berteaux-Lecellier V, Joost S (2019). Seascape genomics as a new tool to empower coral reef conservation strategies: an example on north-western Pacific Acropora digitifera. *bioRxiv*: 588228.

Sgrò CM, Lowe AJ, Hoffmann AA (2011). Building evolutionary resilience for conserving biodiversity under climate change. *Evolutionary Applications* **4**: 326–337.

Shafer ABA, Wolf JBW, Alves PC, Bergström L, Bruford MW, Brännström I, *et al.* (2015). Genomics and the challenging translation into conservation practice. *Trends in Ecology & Evolution* **30**: 78–87.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Stella A, Nicolazzi EL, Van Tassell CP, Rothschild MF, Colli L, Rosen BD, *et al.* (2018). AdaptMap: exploring goat diversity and adaptation. *Genet Sel Evol* **50**: 61.

Stucki S, Orozco-terWengel P, Forester BR, Duruz S, Colli L, Masembe C, *et al.* (2017). High performance computation of landscape genomic models including local indicators of spatial association. *Molecular Ecology Resources* **17**: 1072–1089.

Thomassen HA, Fuller T, Buermann W, Milá B, Kieswetter CM, Jarrín-V. P, *et al.* (2011). Mapping evolutionary process: a multi-taxa approach to conservation prioritization. *Evolutionary Applications* **4**: 397–413.

Vajana E, Barbato M, Colli L, Milanesi M, Rochat E, Fabrizi E, *et al.* (2018). Combining Landscape Genomics and Ecological Modelling to Investigate Local Adaptation of Indigenous Ugandan Cattle to East Coast Fever. *Front Genet* **9**.

Vandergast AG, Perry WM, Lugo RV, Hathaway SA (2011). Genetic landscapes GIS Toolbox: tools to map patterns of genetic divergence and diversity. *Molecular Ecology Resources* **11**: 158–161.

Vangestel C, Eckert AJ, Wegrzyn JL, St. Clair JB, Neale DB (2018). Linking phenotype, genotype and environment to unravel genetic components underlying cold hardiness in coastal Douglas-fir (Pseudotsuga menziesii var. menziesii). *Tree Genetics & Genomes* **14**: 10.

Villemereuil P de, Frichot É, Bazin É, François O, Gaggiotti OE (2014). Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology* **23**: 2006–2019.

Vitti JJ, Grossman SR, Sabeti PC (2013). Detecting Natural Selection in Genomic Data. *Annual Review of Genetics* **47**: 97–120.

Wang S, Ge W, Luo Z, Guo Y, Jiao B, Qu L, *et al.* (2017). Integrated analysis of coding genes and non-coding RNAs during hair follicle cycle of cashmere goat (Capra hircus). *BMC Genomics* **18**: 767.

Weeks AR, Sgro CM, Young AG, Frankham R, Mitchell NJ, Miller KA, *et al.* (2011). Assessing the benefits and risks of translocations in changing environments: a genetic perspective. *Evolutionary Applications* **4**: 709–725.

Williams SE, Shoo LP, Isaac JL, Hoffmann AA, Langham G (2008). Towards an Integrated Framework for Assessing the Vulnerability of Species to Climate Change. *PLOS Biol* **6**: e325.

Willoughby JR, Harder AM, Tennessen JA, Scribner KT, Christie MR (2018). Rapid genetic adaptation to a novel environment despite a genome-wide reduction in genetic diversity. *Molecular Ecology* **27**: 4041–4051.

Zhang Y, Wu K, Wang L, Wang Z, Han W, Chen D, *et al.* (2019). Comparative study on seasonal hair follicle cycling by analysis of the transcriptomes from cashmere and milk goats. *Genomics*.

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**: 3326–3328.

Zhou X, Carbonetto P, Stephens M (2013). Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLOS Genetics* **9**: e1003264.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.20.884114; this version posted December 20, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.