HAL
archives-ouvertes.fr

Diversité des génomes et adaptation locale des petits ruminants d'un pays méditerranéen : le Maroc

Badr Benjelloun

To cite this version:

Badr Benjelloun. Diversité des génomes et adaptation locale des petits ruminants d'un pays méditerranéen : le Maroc. Ecologie, Environnement. Université Grenoble Alpes, 2015. Français. <NNT : 2015GREAV011>. <tel-01280471>

HAL Id: tel-01280471

https://tel.archives-ouvertes.fr/tel-01280471

Submitted on 29 Feb 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ GRENOBLE ALPES

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Biodiversité Ecologie Environnement**

Arrêté ministériel : 7 août 2006

Présentée par

« Badr BENJELLOUN »

Thèse dirigée par **«Pierre TABERLET»** et
codirigée par **«François POMPANON »**

préparée au sein du **Laboratoire d'Ecologie Alpine**
dans **l'École Doctorale Chimie et Sciences du Vivant**

Diversité des génomes et adaptation locale des petits ruminants d'un pays méditerranéen : le Maroc

Thèse soutenue publiquement le **« 01 septembre 2015 »**,
devant le jury composé de :

M. Nicolas BIERNE
Directeur de Recherche, CNRS, Montpelier (Rapporteur)
M. Christophe DOUADY
Professeur, Université Lyon 1 (Rapporteur)
M. Mohamed BADRAOUI
Professeur, Directeur de l'INRA-Maroc, Rabat, Maroc (Membre)
Mme. Laurence DESPRES
Professeur, Université Grenoble Alpes (Présidente)
M. Bertrand SERVIN
Chargé de Recherche, INRA, Toulouse (Membre)
M. François POMPANON
Maitre de Conférences, Université Grenoble Alpes (Directeur de thèse)
M. Pierre TABERLET
Directeur de Recherche, CNRS, Grenoble (Directeur de thèse)

Remerciements

Cette thèse a été l'un des fruits d'une collaboration amorcée en janvier 2008 quand Pierre a effectué une visite à l'INRA Maroc en tant qu'expert dans le cadre du projet MOR 5030 financé par de l'Agence Internationale de l'Energie Atomique (AIEA). Lors de cette visite, nous avons parcouru plusieurs centaines de kilomètres et visité plusieurs élevages dans plusieurs régions du Maroc (de Tadla-Azilal à Tanger). Pierre a eu l'idée d'étudier les bases génétiques de l'adaptation des petits ruminants du Maroc à l'environnement via une approche de génomique du paysage. Ensuite, j'ai été invité à plusieurs reprises au LECA où j'ai rencontré François qui a ensuite visité l'INRA Maroc pour la mise en place et le développement de cette collaboration. Ensuite, le projet NextGen a été lancé en 2010 et m'a procuré le privilège de mener les travaux de cette thèse en étroite collaboration avec les différents partenaires du consortium NextGen.

Je commence cet exercice de reconnaissance par remercier les membres du jury. Mes remerciements chaleureux et ma profonde gratitude vont à Monsieur le Directeur de l'INRA Maroc qui a accepté de participer à l'évaluation de ce travail malgré les lourdes responsabilités associées à son poste. Mes remerciements chaleureux vont également aux autres membres du jury pour avoir accepté d'examiner mon travail de thèse : Laurence Després, Bertrand Servin, Christophe Douady et Nicolas Bierne. Un grand merci tout particulièrement à Nicolas Bierne et Christophe Douady qui se sont chargé de la lourde responsabilité de rapporteurs.

Ensuite, je tiens à souligner la fortune d'avoir rencontré mes directeurs de thèse qui m'ont épaulé, encadré et aidé pendant la mise en place et tout au long de l'aventure de cette thèse. Ainsi, je ne saurais comment remercier François pour m'avoir mis les pieds dans le monde de la génomique et pour m'avoir imprégné de son approche scientifique, son expérience et sa rigueur. Merci aussi pour tes encouragements aux moments de doute et au delà, ta sympathie, ton hospitalité ainsi que ta disponibilité malgré les lourdes responsabilités qui t'incombent.

De même, je ne saurais exprimer ma gratitude à Pierre. Je te suis redevable de plusieurs choses, notamment de la confiance que tu m'as fait pour mener de tels travaux et de mon implication dans de tels domaines de recherche. Merci également pour toutes les idées originales ainsi que pour tout le soutient et l'hospitalité tout au long des sept années de notre connaissance.

Je n'oublie pas Abdelkader Chikhi qui a su débloquer plusieurs situations délicates lors de la réalisation de l'échantillonnage au Maroc. Je te remercie pour tous les efforts administratifs et manageriels engagés pour l'aboutissement des différentes tâches. Je te remercie également pour les conseils et l'encadrement fructueux pendant les deux premières années de cette thèse avant ton départ à la retraite. De même, je n'oublie pas le rôle de Bouchaib Boulanouar dans mon initiation à la recherche scientifique en m'impliquant dans différentes activités dès mon arrivée à l'INRA Maroc en 2004. Je te remercie chaleureusement et j'espère que nous aurons l'occasion de collaborer ensemble à l'avenir.

Fréderic Boyer, c'est toi qui m'as initié à l'utilisation du schell et l'élaboration des scripts pour l'analyse des grandes masses de données et c'est grâce à ton accompagnement et ta contribution que nous avons réalisé ce travail. Merci beaucoup! De même, je n'oublie pas le rôle de Lorenzo Bomba et toute l'équipe de l'UNICAT à Plaisance, Italie sous la direction de Paolo Ajmone Marsan, notamment Marco Milanesi, Licia Colli et Ricardo Negrini dans cette initiation. Les six semaines passées au sein de cette équipe pendant le début de cette thèse étaient d'une immense utilité pour mon initiation à l'analyse des données génomiques.

Je tiens à remercier vivement Ian Streeter pour son aide et sa disponibilité pour les différents traitements, analyses, relectures ainsi que pour son sens de partage. Sa contribution dans ce travail était d'une importance capitale. Je remercie également tous les membres du consortium Nextgen pour l'ambiance de travail, la synergie et la qualité scientifique de leurs différentes contributions. Je tiens à signaler le rôle de Florian Alberto, Sylvie Stucki, Kevin Leempoel, Stéphane Joost, Pablo Orozco terWengel, Filippo Biscarini, Laura Clarke, Alessandra Stella, Adriana Alberti, Stefan Engelen, James Kijas, Mike Bruford, Paul Flicek et Eric Coissac dans la réalisation des différentes activités.

Je remercie vivement toute l'équipe qui a pris la lourde tache de l'échantillonnage au Maroc et qui ont bravé les différentes conditions difficiles de terrain. Je remercie tout particulièrement, Mohammed BenBati, Mustapha Ibnelbachyr, Abdelmajid Bechchari, Mouad Chentouf, Mouloud Laghmir, Lahbib Haounou, El Moustapha Sekkour, Sadek Mustapha et les autres. Sans vos efforts, ce travail ne pouvait pas être réalisé.

Je tiens à remercier Wahid Zamani pour le travail commun réalisé au début de cette thèse ainsi que pour les efforts engagés pour l'échantillonnage des animaux sauvages et domestiques en Iran en collaboration avec Saeid Naderi et Hamid Reza Rezaei. Je te souhaite une bonne chance et une bonne continuation.

Je remercie vivement Florian Alberto, Eric Coissac, Eric Bazin, Oscar Gaggiotti, Alexandra Vatsiou, Tristan Cumer et Eric Frichot pour les discussions et orientations fructueuses lors des mises au point de plusieurs stratégies de travail/analyse.

Mes remerciements les plus chaleureux à tous les collègues du LECA. Cela fût un véritable plaisir de vous côtoyer pendant les 31 mois passés ensemble. Merci pour la bonne ambiance, les coups de main et pour toutes les bonnes idées, je remercie plus particulièrement Florence Sagnimorte, Nancy Iacono, Kim Pla, Johan Pansu, Christian Miquel, Délphine Rioux et Carole Poillot pour toute l'aide qui m'ont consacrée tout au long de cette thèse.

Merci également à l'équipe de l'école doctorale CSV, et plus particulièrement à Magali Pourtier pour toute l'assistance dans les démarches administratives y compris en cette fin de thèse.

Mes remerciements chaleureux vont à MM. Rachid Dahan, Rachid Mrabet, Mohammed Beqqali, Yahya Baye, Mouad Chentouf, Ahmed Bellamlik, Abdellatif Ennahir, Mohammed Kadiri, Abdeljabbar Bahri et Mohamed El Asri ainsi que Mme Sanae Belhsen et Mlle. Bouchra El Amiri de l'INRA Maroc pour leur soutien et leur assistance à plusieurs niveaux pour la progression de la thèse et pour la réussite du projet.

Merci à mes chers amis Tarik Benabdelouahab et Mohammed Benbati pour leur disponibilité, leur soutien sans faille et leurs conseils précieux.

Je présente également mes vifs remerciements à Odile Pompanon et Marie-Odile Taberlet pour tout le soutient et l'hospitalité tout au long de ce parcours.

Enfin, j'adresse mes remerciements les plus chaleureux à ma chère Imane qui a été toujours présente à mes côtés, à tous les instants pour m'apporter son aide, son soutien et son amour afin de franchir les moments les plus durs. Je n'y serais pas arrivé sans toi, cette thèse c'est aussi la tienne. Merci également à Ahmed et Omar pour leur indulgence et pour avoir supporté mon éloignement pendant des moments où ils avaient besoin de ma présence. Mes remerciements chaleureux vont à ma mère, mes sœurs et frères pour le soutient inconditionnel tout au long de ce parcours.

Puisqu'il est difficile de remercier toutes les contributions à la réalisation de ce travail sans risquer d'en oublier quelques unes, je présente mes excuses ainsi que ma reconnaissance à toutes les personnes non-citées mais qui sauront se reconnaître à travers ces quelques lignes.

Tables des matières

Introduction générale **6**
1. L'étude des processus évolutifs **6**
2. L'adaptation locale **7**
2.1. Effet des autres processus évolutifs sur l'adaptation locale 8
2.2. A la recherche des bases génétiques de l'adaptation locale 9
2.3. La *landscape genetics/genomics* pour étudier l'adaptation locale 11
3. Les petits ruminants **12**
3.1. Histoire post-domestication et contexte mondial 12
3.2. Contexte marocain 15
4. Le projet NextGen **17**
5. Le travail de thèse **18**
5.1. Axes de recherche 19
5.2. Contribution personnelle 23
Références **24**

Chapitre 1: Echantillonnage représentatif des données de génomes complets **30**
Résumé et présentation de l'article **30**

Article A: What information at which cost? The reliability of variant panels and low-coverage WGS for describing genome diversity **32**
Abstract **33**
Author summary **34**
Introduction **35**
Results **37**
Estimation of population genomics statistics 37
Assessment of standard surrogates of whole genome data 42
Difference between random panels and BeadChips 42
Reliability of low-coverage re-sequencing 43
Discussion **47**
Effect of sequencing coverage on the assessment of whole genome variations 47
Effect of the density of variants 48
Ascertainment bias in non-random panels 49
Distribution of variants across the genome 50
Conclusion **51**
Material & methods **51**
Sampled individuals 51
DNA extraction and re-sequencing 52
Read mapping, SNP calling and filtering 52
Quality control of WGS data 54
Setting up datasets of variants 54
Simulating low-coverage re-sequencing data 56
Population genetics analyses 56
References **60**
Supplementary material **63**

CHAPITRE 2: Caractérisation des génomes des caprins locaux au Maroc **82**
Résumé et présentation de l'article **82**

Article B: Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (Capra hircus) using WGS data **84**
Abstract **85**
Introduction **86**

Material and Methods **88**
Sampling 88
Production of WGS Data 88
WGS Data Processing 89
Population Genomic Analyses 90
Gene Ontology Enrichment Analyses 93
Results **93**
Phylogeny of mtDNA Genomes 93
Neutral Diversity from WGS Data 94
Selection Signatures 97
Discussion **100**
Mitochondrial Variation 100
Nuclear Neutral Variation 101
Selection Signatures in Moroccan Goat Populations 103
Conclusion **107**
References **109**
Supplementary Material **114**

CHAPITRE 3: Les bases génétiques de l'adaptation locales chez les petits-ruminants domestiques **126**
Résumé et présentation de l'article **126**

Article C: Towards the genetic bases of local adaptation: a wide-scale landscape genomic approach in sheep (O. aries) and goats (C. hircus) **129**
Summary **130**
Introduction **131**
Material & Methods **133**
Sampling 133
Production of WGS datasets 134
WGS data processing 134
Genetic diversity and population structure in Moroccan sheep and goats 135
Environmental variables 136
Analyses of signatures of selection 137
Gene Ontology enrichment analyses 139
Results **140**
Population structure 140
Detection of signals of selection related to environmental variations 141
Gene Ontology enrichment analysis 145
Discussion **148**
Overall genomic variation 149
Bases of local adaptation in sheep and goats 152
Adaptive convergence 157
Differences between population-based and correlative approaches 157
Conclusion **158**
References **159**
Supplementary Material **163**

Discussion générale **176**
1. Vers des solutions alternatives adaptées aux données de génomes complets **176**
1.1. Biais dans les puces commerciales d'ADN 176
1.2. Alternatives possibles 177
1.3. Des économies sur la couverture de re-séquençage? 178
2. Les populations locales et sauvages comme ressources génétiques **178**
2.1. Rappel des principaux résultats de paramètres génétiques 179
2.2. Diversité génétique et différenciation 179

2.3. Des scénarios pour expliquer l'état actuel de diversité marocaine 181
3. Les bases génétiques de l'adaptation locale chez les chèvres et moutons 184
3.1. Aspects méthodologiques 184
3.2. Non concordance des signatures de sélection entre les méthodes corrélatives et populationnelles 185
3.3. Adaptations parallèles dans différentes populations/espèces 185
3.4. Implication de fonctions respiratoire et circulatoire dans l'adaptation à l'altitude 187
3.5. Différenciation « adaptative » le long des gradients d'altitude 189
3.6. Limites méthodologiques 190
4. Perspectives 190
4.1. Finalisation des études en cours 191
4.2. Recherches futures 191
Conclusion 194
Références 197

Annexes 202

INTRODUCTION GENERALE

Introduction générale

1. L'étude des processus évolutifs

La diversité génétique des espèces est le résultat des processus démographiques et adaptatifs qui conditionnent leur évolution en modifiant les fréquences alléliques à l'échelle des populations. Quelques processus affectent le génome entier tel que la dérive génétique (incluant les événements fondateurs et les goulots d'étranglements), la migration ou la consanguinité. D'autres processus tel que la sélection, la mutation, la recombinaison ou l'appariement assortatif ont cependant des effets spécifiques se limitant à un ou quelques locus. Depuis son émergence au début du XXème siècle, la génétique des populations s'intéresse à l'étude des variations des fréquences alléliques sous l'influence de ces processus évolutifs (Fisher 1919 ; Wright 1931 ; Fisher 1958) et fournit des outils conceptuels et statistiques pour les dissocier et inférer leur effets à partir de la variabilité génétique observée. Ceci permet de développer des approches intégrées visant à décrire simultanément les bases génétiques des variations phénotypiques et l'évolution des gènes qui sous-tendent ces variations (Black et al. 2001).

Ces études requièrent de prime abord la capacité de caractériser les variations génétiques du polymorphisme des génomes. Jusqu'à la dernière décennie, le choix des marqueurs moléculaires était limitant pour avoir une vision d'ensemble des variations qui ne sont pas réparties uniformément à l'échelle du génome. Les principaux marqueurs qui donnaient accès aux fréquences alléliques, i.e. les marqueurs co-dominants sont les microsatellites qui ne permettent pas de caractériser la variabilité de l'ensemble du génome (maximum quelques dizaines de loci caractérisables conjointement). Des marqueurs qui couvraient mieux le génome comme les AFLP sont dominants, i.e. ne permettent pas d'accéder aux fréquences alléliques. Leur utilisation pour détecter la variabilité adaptative par exemple reste limitée (Jones et al. 2013).

Plus récemment, les développements technologiques (amorcés pendant la dernière décennie du XXème siècle) permettant de typer des centaines voire des milliers de marqueurs à l'échelle d'un seul organisme et le développement des techniques permettant de détecter les *Single Nucleotide Polymorphism* (SNP) dans les régions les plus conservées ont permis une meilleure résolution lors de l'étude de ces variations. Dans ce contexte, la génomique des populations a émergé au début du XXIème siècle (Black et al. 2001) et consiste à combiner les concepts génomiques et l'utilisation des nouvelles technologies de séquençage/génotypage

pour atteindre les objectifs de la génétique des populations et comprendre les processus évolutifs (Luikart et al. 2003). A ce jour, les techniques de séquençage ne cessent de se développer (Branton et al. 2008; Clarke et al. 2009; Snyder et al. 2010; Loman and Watson 2015) et le séquençage des génomes complets de plusieurs individus est devenu possible. De même, le développement d'outils informatiques et d'algorithmes nécessaires pour traiter et analyser le raz-de-marée de données qui s'en résulte permet d'avoir un accès sans priori à la variabilité génomique et sa distribution à l'échelle du génome, e.g. (Altshuler et al. 2012; Kidd et al. 2012; Ai et al. 2015).

Ainsi, il est désormais possible de mieux répondre à des problématiques de génomique des populations en lien avec plusieurs domaines allant de la médecine et la pharmacie jusqu'à l'écologie et l'agronomie. Les modèles de génétique des populations permettent de comprendre les mécanismes qui sous-tendent la résistance ou la susceptibilité aux maladies, la réponse aux médicaments et d'en tirer profit pour améliorer la santé humaine, animale ou végétale. Dans d'autres cas, ces modèles permettent d'avoir les outils nécessaires pour la conception de schémas de sélection assistés par marqueurs pour différents objectifs. Ils permettent également de comprendre les bases génétiques de l'adaptation locale à des fins écologiques et/ou agronomiques ou de recherche fondamentale.

2. L'adaptation locale

L'origine du concept de l'adaptation est très ancienne mais Darwin au milieu du XIXème siècle proposa pour la première fois une explication scientifique solide à l'apparition des adaptations (Darwin 1859). Cette explication était basée entièrement sur le mécanisme de sélection naturelle, moteur de ce processus adaptatif par le tri des individus les plus capables de survivre et se reproduire dans un environnement. Sa motivation principale était d'éclaircir/expliquer le mystère de la spéciation et pendant longtemps, l'adaptation est restée liée aux études de la spéciation à un niveau macro-évolutif. Pourtant l'adaptation correspond bien à un processus micro-évolutif se déroulant au niveau intra-spécifique. En effet, la variation environnementale est omniprésente et les populations au sein d'une espèce donnée s'adaptent à leurs conditions locales abiotiques et biotiques, par exemple le long de gradients environnementaux ou entre des types d'habitat contrastés. Nous parlons alors d'adaptation locale qui est restreinte à un certain nombre de populations. Dans le contexte actuel de changements climatiques, c'est cette adaptation qui pourrait jouer un rôle-clé dans la survie de

la population (Franks and Hoffmann 2012), malgré que d'autres processus puissent interférer et avoir des principaux rôles également.

2.1. Effet des autres processus évolutifs sur l'adaptation locale

L'adaptation locale qui repose sur la sélection naturelle dépend également des autres processus évolutifs, notamment :

(i) La migration

Quand la population connaît des flux migratoires, l'adaptation locale ou son absence est le résultat de la balance entre la sélection et la migration. Les variations spatiales de sélection peuvent amener à l'adaptation locale et donc à la différenciation génétique, mais pas dans toutes les conditions. Une sélection variable dans le temps ou des recolonisations répétitives des populations entraveront l'adaptation locale (Kawecki and Ebert 2004). D'une façon générale, pour la mise en place de l'adaptation locale, les caractéristiques conférant une forte valeur adaptative (*fitness*) dans un environnement doivent être défavorables dans l'autre environnement et la migration ne doit pas prendre le dessus sur l'effet de cette adaptation. D'ailleurs, ce processus influence l'effet d'autres processus évolutifs sur l'adaptation locale, notamment celui de la dérive génétique.

(ii) La dérive génétique

L'influence de ce processus sur l'adaptation locale peut être de différentes natures et dépend principalement de l'intensité des flux migratoires. Des études empiriques et des simulations ont montré qu'avec un taux de migration faible à moyen, la dérive génétique affecte négativement l'adaptation locale. Cependant avec un taux de migration élevé, la dérive génétique n'a pas d'effets sur cette adaptation (Blanquart et al. 2012). En effet, la dérive peut être responsable de la fixation de mutations délétères ou de la perte de mutations bénéfiques alors qu'elles sont encore rares (Elena and Lenski 2003) ce qui suggère une action plutôt défavorable sur l'adaptation locale. Cependant, des cas de mutations neutres fixées par dérive dans une population avant de devenir adaptatives ont été démontrés (Gould and Lewontin 1979 ; Hughes 1999) suggérant ainsi une possible action avantageuse de la dérive sur l'adaptation locale.

(iii) La mutation

Les résultats montrant la rapidité de la mise en place de l'adaptation locale sont nombreux, e.g. (Huey et al. 2000; Simonson et al. 2010; Chen et al. 2012) et suggèrent que l'origine de

l'adaptation locale provient plutôt de la variabilité résidente dans le génome que de nouvelles mutations (Savolainen et al. 2013). Plusieurs adaptations liées à cette variabilité résidente du génome ont été confirmées, comme à titre d'exemple, l'implication de cette variabilité en réponse aux traitements de domestication au laboratoire de l'espèce de levure *Saccharomyces cerevisiae* chez 12 populations indépendantes (Burke et al. 2014). De même, cette variation serait à l'origine de la variation génétique fonctionnelle de l'hémoglobine au niveau des gènes *HBA* et *HBB* chez la souris sylvestre et les autres espèces du genre *Peromyscus* (Natarajan et al. 2015). Par ailleurs, la perte de diversité génétique représente une grande menace pour le potentiel adaptatif d'une population donnée (Frankham 2002). Cependant, des mutations récentes peuvent également être impliquées dans certaines adaptations locales et certains cas ont été reportés, e.g. la mutation unique qui sous-tend la forme noire de la phalène du bouleau (*Biston betularia*) qui représente une forme mieux adaptée à la pollution industrielle (van't Hof et al. 2011).

En dehors de ces processus évolutifs, le développement de l'adaptation locale est tributaire de plusieurs autres mécanismes, e.g. le développement de la plasticité phénotypique peut bien être un frein voire même une alternative à l'adaptation locale pour la survie des populations (Chevin et al. 2010; Crispo et al. 2010).

2.2. A la recherche des bases génétiques de l'adaptation locale

La multiplicité de mécanismes et processus affectant l'adaptation locale fait que ses bases génétiques sont complexes, de différentes origines et restent jusqu'à ce jour loin d'être élucidées. En effet, malgré les progrès technologiques récents qui ont permis d'identifier plusieurs variants candidats potentiellement impliqués dans les mécanismes adaptatifs, leur validation doit passer par la mise en évidence de la fonction reliant les génotypes aux phénotypes qui influenceraient la valeur adaptative (Akey 2009). Cette validation nécessite ainsi des expérimentations complexes visant à identifier les changements phénotypiques associés à l'effet du variant considéré. Subséquemment, la plupart des variants adaptatifs qui ont été bien caractérisés jusque là ont des impacts phénotypiques forts et facilement mesurables. Par exemple, les gènes impliqués dans la tolérance des métaux lourds par les plantes (Macnair 1993), ou ceux responsables de la perte de l'armure osseuse chez l'épinoche à trois épines des eaux douces (Jones et al. 2012) ou encore ceux impliqués dans la tolérance du lactose à l'âge adulte chez les humains (Bersaglieri et al. 2004). Toutefois, les caractères qui confèrent l'adaptation locale peuvent être polygéniques, quantitatifs ou alors associés à des gènes à effets inconnus ou pléiotropiques. Quelques approches ont été utilisées pour

surmonter la difficulté de caractérisation dans certains de ces cas. Ainsi, quelques variants ayant des effets pléiotropiques et jouant des rôles adaptatifs importants ont été identifiés, par exemple, la mutation *EDAR 370A* fréquente dans les populations humaines de l'Asie orientale accroit l'épaisseur des cheveux, modifie la morphologie des dents et augmente le nombre de glandes sudoripares exocrines (Fujimoto et al. 2008 ; Kimura et al. 2009 ; Park et al. 2012). Chez la souris, elle agit en outre sur les glandes mammaires et n'influence pas la morphologie dentaire (Figure 1 ; Kamberov et al. 2013). Linnen et al. (2013) de leur part ont trouvé que l'adaptation locale de la souris sylvestre *Peromyscus maniculatus* qui a colonisé récemment les plateaux des sables de Nebraska serait associée à des effets de sélections indépendantes au niveau de plusieurs mutations dans un seul gène. Chaque mutation a un effet spécifique sur un phénotype adaptatif, minimisant ainsi les actions pléiotropiques. Un autre exemple de complexité des mécanismes adaptatifs est celui de *Arabidopsis Taliana* suggéré par Fournier-Level et al. (2011) via une étude d'association pangénomique avec les valeurs adaptatives globales. Cette dernière étude a identifié une implication importante des variants inter-géniques suggérant ainsi un rôle régulateur important et complexe dans ce processus. Une autre forme adaptative est celle rapportée par Daub et al. (2013) qui ont trouvé que l'adaptation des humains aux pathogènes serait liée à de faibles effets épistatiques de plusieurs gènes sur plusieurs chromosomes dans les mêmes voies métaboliques.

D'une manière générale, la recherche des bases génétiques de l'adaptation locale représente un domaine de recherche en expansion qui se base de plus en plus sur des combinaisons de nouvelles approches génomiques, environnementales et statistiques. La compréhension de ses mécanismes génétiques et biochimiques devrait aider à améliorer la subsistance de certaines espèces/races/variétés face aux changements environnementaux, d'améliorer la productivité agricole ou simplement d'améliorer notre compréhension de la distribution de la diversité.

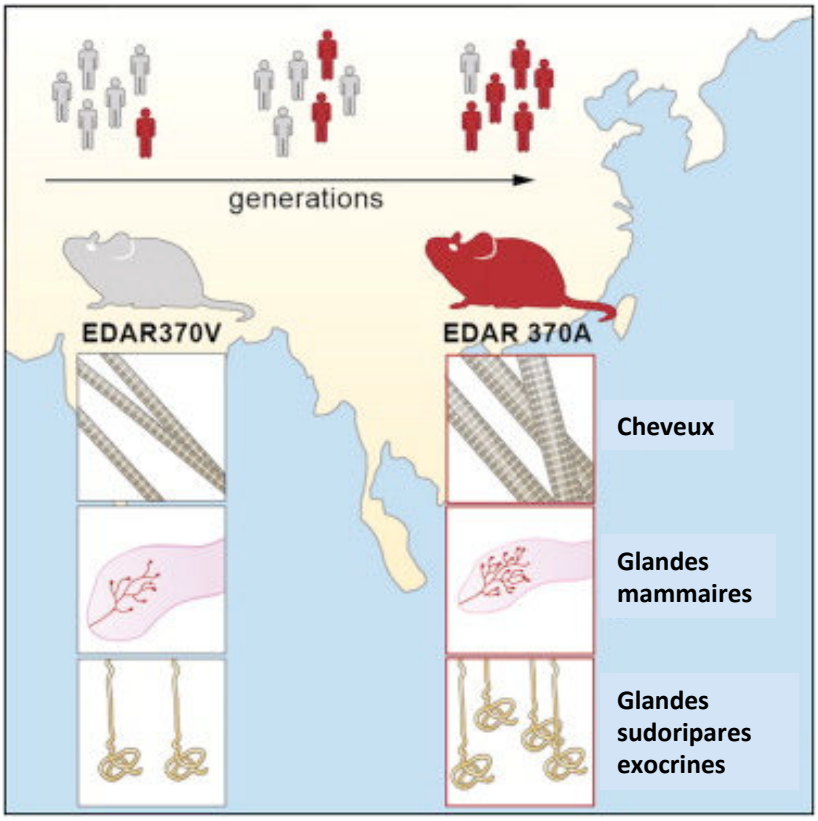

Figure 1. Schéma illustrant l'effet pléiotropique de la mutation *EDAR 370A* chez la souris (Kamberov et al. 2013).

Cette mutation très fréquente chez les humains des populations de l'Asie orientale serait apparue pour la première fois chez les chinois Hans il y a 30.000 ans selon cette étude. Elle confère également le changement de la morphologie dentaire chez les humains ce qui n'est pas mis en évidence chez la souris.

2.3. La *landscape genetics/genomics* pour étudier l'adaptation locale

La *landscape genetics* ou la « génétique du paysage » a émergé comme nouvelle approche en 2003 (Manel et al. 2003), et dès lors, elle s'intéresse à l'étude des interactions entre les caractéristiques de l'environnement et les processus évolutifs, principalement les flux de gènes et la sélection à une échelle micro-évolutive, i.e. intra-spécifique (Holderegger and Wagner 2006). Elle fournit ainsi des outils supplémentaires pour étudier l'adaptation locale et comprendre ses mécanismes. Cependant, la majorité des recherches conduites jusqu'à 2010 se sont basées sur peu de marqueurs génétiques avec une dominance de l'usage des marqueurs AFLP ou de nombres limités de SNPs (Storfer et al. 2010). Les avancées technologiques récentes dans le domaine de séquençage et l'émergence de grands jeux de données éco-climatiques ont fait émerger une nouvelle discipline qui représente une branche de la génétique du paysage. Il s'agit de la *landscape genomics* « génomique du paysage » ou « génomique environnementale » qui vise à examiner comment les facteurs éco-climatiques

influencent la variation génétique adaptative en utilisant principalement les techniques de criblage génomique et en se basant sur de grands nombres de marqueurs moléculaires. C'est un domaine florissant qui a montré son efficacité pour examiner les mécanismes adaptatifs (e.g. Jones et al. 2013; Vincent et al. 2013; De Kort et al. 2014). Jusque là, la plupart des études de « génomique du paysage » ont utilisé des approches de génétique des populations pour identifier la variation génomique adaptative (Manel and Holderegger 2013). Toutefois, des méthodes spécifiques ont été développées également. Elles sont basées sur la recherche de corrélations directes entre fréquences alléliques et variables environnementales (Joost et al. 2007; Frichot et al. 2013; Stucki et al. 2014). Cependant lors de ces études de génomique du paysage, il est nécessaire de considérer la structuration génétique et les autres effets démographiques lors des analyses (Manel and Holderegger 2013) et d'adopter des stratégies adéquates d'échantillonnage.

L'application des ces approches pour comprendre les mécanismes sous-jacents à l'adaptation locale requiert ainsi des modèles biologiques adéquats avec une grande variabilité phénotypique. A la différence de la majorité des études traitant de l'adaptation locale qui se sont focalisées sur des espèces sauvages, nous avons choisi dans le cadre de cette thèse de travailler sur des espèces d'élevage ayant un grand intérêt économique : les moutons (*Ovis aries*) et les chèvres (*Capra hircus*). Ces espèces présentent l'avantage d'avoir des génomes de référence assemblés (Dong et al. 2013 ; Jiang et al. 2014) aidant à détecter plus facilement les variations, avoir les annotations et à réaliser les criblages génomiques pour la recherche des signatures de sélection. Elles ont également des espèces apparentées respectives représentées par les ancêtres sauvages (*Ovis orientalis*) et (*Capra aegagrus*) desquelles elles ont divergé récemment (voir section « Les petits ruminants »), ce qui permettrait de mener éventuellement des investigations sur les origines des adaptations potentiellement détectées. Le caractère domestique de ces espèces ouvre également des possibilités de réalisation plus simples d'expérimentation dans des milieux contrôlés en vue de compléter éventuellement les études des mécanismes adaptatifs.

3. Les petits ruminants

3.1. Histoire post-domestication et contexte mondial

Les petits ruminants domestiques (ovins et caprins) représentent avec les bovins les principales sources de viandes rouges, de lait et de cuir à travers le monde. Il y avait 1.163

millions de moutons et 976 millions de chèvres dans le monde en 2013 et ils ont produit 8,6 et 5,4 millions de tonnes respectivement de viandes et 10,1 et 18 millions de tonnes de lait (http://faostat3.fao.org/browse/Q/QL/F). Ils représentent ainsi une composante indispensable de l'activité agricole mondiale de par leur production, leur importance socio-économique et leur contribution à l'alimentation des populations humaines.

Ces espèces ont été les premiers ongulés à être domestiqués il y a 9.900 à 10.500 ans. Les ovins ont été domestiqués dans le Croissant Fertile très probablement à partir des mouflons asiatiques *Ovis orientalis* (Peters et al. 2005, Rezaei 2007). Chez les caprins, l'ancêtre sauvage est l'aegagre *Capra aegagrus* et il y aurait deux évènements de domestication : Le premier consisterait plutôt à une pré-domestication avec une gestion extensive de troupeaux sauvages et aurait eu lieu au niveau du sud du Zagros et du Plateau Central iranien. Le second événement aurait eu lieu plutôt vers le nord des montagnes du Zagros (en Iran) et l'est de l'Anatolie (en Turquie). Les chèvres domestiques modernes seraient ainsi issues de ce deuxième événement de domestication (Naderi et al. 2008). Une très large diversité génétique a été capturée à partir des animaux sauvages (Naderi et al. 2008) et l'introgression des gènes sauvages a pu continuer après ces événements initiaux (Fernandez et al. 2006) contribuant à former la diversité génétique des animaux domestiques.

Durant les 3000 à 4000 années qui ont suivi la domestication initiale, l'agriculture a été diffusée en Europe, en Afrique et en Asie via différentes voies. Les deux voies principales vers l'Europe sont les voies méditerranéenne et danubienne (Fernandez et al. 2006). D'autres vagues de migration plus récentes ont accompagné les migrations humaines et auraient persisté jusqu'à très récemment avec des flux de gènes conséquents, e.g. (Beja-Pereira et al. 2006; Pereira et al. 2009) (Figure 2). Ces longs processus ont permis à ces animaux domestiques d'acquérir progressivement et pendant des millénaires des adaptations à leur milieu. En outre, le flux de gènes abondants via les échanges probables entre les voies migratoires aurait permis d'accumuler une diversité génétique très riche qui constituerait un potentiel adaptatif et productif fort précieux (voir section 'adaptation locale'). Tout récemment, la situation a commencé à changer dramatiquement, notamment avec la formulation du concept de race moderne au début du XIXème siècle (Porter 2002) et son application à l'élevage et à ses pratiques (Figure 2). Ceci a conduit à la formation de races bien définies sur des bases morphologiques (e.g. couleur de la robe) et la reproduction entre des phénotypes différents a été fortement freinée (Taberlet et al. 2008). Les animaux étaient ainsi exposés à une plus forte pression de sélection et ont subit des forts goulots

d'étranglement. En outre, le développement à partir du milieu du XXème siècle de nouvelles technologies appliquées à l'élevage comme l'insémination artificielle, le transfert d'embryons, les techniques d'alimentation et l'utilisation de vaccins et de traitements contre les maladies endémiques a facilité la diffusion à grande échelle de l'élevage industriel. Ceci a conduit également à une nouvelle phase dans l'histoire démographique des espèces domestiques à l'échelle internationale. Des individus de quelques races ont commencé à être diffusés à très grande échelle : à l'intérieur des pays développés et de ces derniers vers les pays en développement (FAO 2007). Les éleveurs dans les quatre coins du monde ont ainsi commencé à remplacer progressivement les races autochtones bien adaptées aux conditions locales et très diversifiées par ces races très productives dites 'industrielles' (FAO 2007 ; Taberlet et al. 2008). Ainsi, un grand nombre de races indigènes (dont la diversité génétique et les caractères adaptatifs ont été accumulée pendant les 10 millénaires de leur histoire commune avec les humains) ont disparu ou se sont trouvés en risque d'extinction (FAO 2007). Jusqu'à 2006, 13% et 3% des races inventoriées dans le monde ont disparu chez les moutons et les chèvres respectivement. Les races classées « menacées » par la FAO représentaient 13% et 14% chez les deux espèces respectivement, tandis que les races dont la situation était inconnue représentaient 30% chez les ovins et 38% chez les caprins (Figure 3). Cette situation d'érosion perdure : 62 extinctions de races ont été enregistrées chez les animaux d'élevage entre 1999 et 2006, 59 races sont passées de la classe « non menacé » à la classe « menacé » alors que 60 races qui étaient menacées en 1999 ne l'étaient plus en 2006 (FAO 2007). Les races dites 'industrielles' qui sont en très forte expansion démographique connaissent une chute énorme de diversité puisque des groupes restreints de reproducteurs ayant des performances zootechniques très élevées sont généralement utilisés à grande échelle. Ainsi, une chute de tailles efficaces et une augmentation du taux de consanguinité ont été relevées chez plusieurs populations principalement des bovins (où l'insémination artificielle est appliquée à très grande échelle), e.g. La population Holstein en Espagne aurait une taille efficace (*Ne*) entre 66 et 79 (estimations basées sur des individus nés entre 1960 et 2013) (Rodríguez-Ramilo et al. 2015). Des tailles efficaces plus faibles ont été reportées précédemment sur d'autres populations Holstein dans d'autres pays (France, USA,… ; Taberlet et al. 2008). Cependant, une réduction de taille efficace a également été reportée chez les moutons par Tapio et al. (2005). Cette réduction de diversité présente ainsi une perte énorme du potentiel adaptatif de ces races desquelles l'alimentation des populations humaines est de plus en plus dépendante (Taberlet et al. 2008).

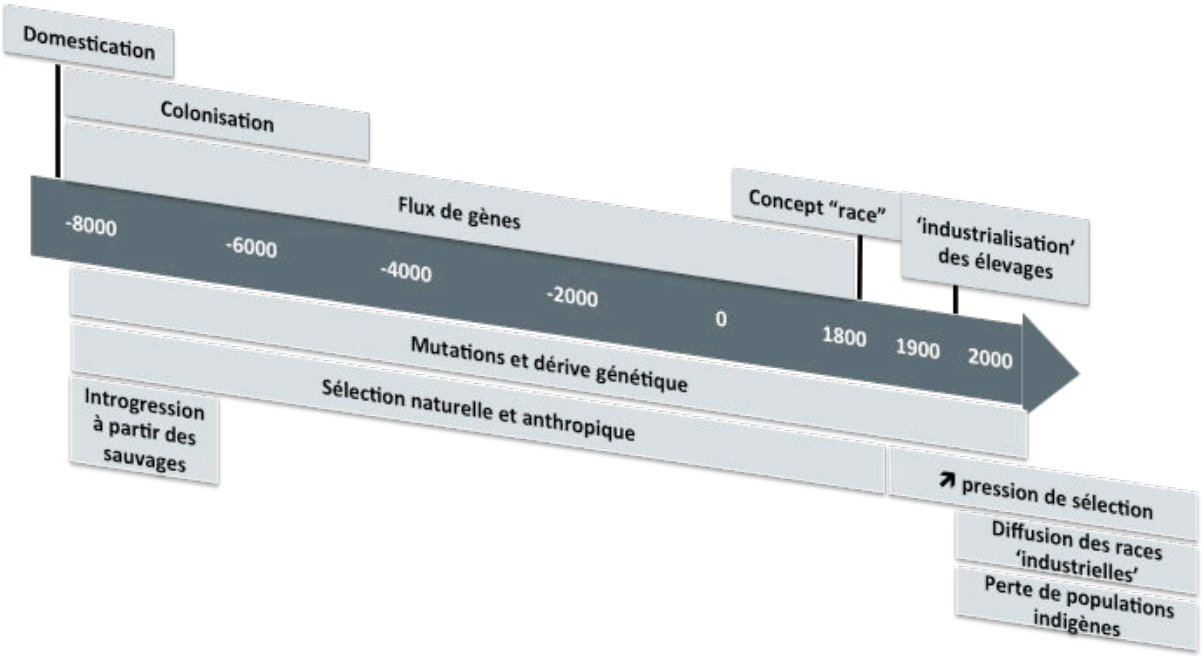

Figure 2. Illustration de l'histoire démographique post-domestication des petits ruminants (Modifiée à partir de Taberlet et al. 2008).

3.2. Contexte marocain

La situation au Maroc n'est pas si critique parce que l'élevage des petits ruminants revêt une grande importance socio-économique et intéresse plus de 65% de la population rurale. La taille du cheptel exploité est en croissance et elle était de 26,2 millions têtes en 2013 dont 20 millions d'ovins et 6,2 millions de caprins (http://faostat3.fao.org/browse/Q/QA/F), constitués à 95% de races et populations locales bien adaptées à leur environnement. L'élevage des ovins et caprins est pratiqué dans la grande part du pays sous des conditions éco-climatiques très contrastées en terme d'altitude, température, précipitations, relief, etc. et sous des pratiques d'élevage très diversifiées.

Pour les ovins, et depuis 1980, six principales races locales ont été phénotypiquement identifiées, et un programme de préservation basé sur la sélection phénotypique dans certaines zones dites «berceaux de races» a été mis en œuvre (MARA 1980). Ces six races représenteraient environ 40% de l'effectif total des ovins du Maroc (Boulanouar and Benlekhal, 2005). Les 60% restants étant constitués de races étrangères (5%) et d'individus de populations locales non encore identifiées (55%), et très peu étudiées malgré l'intérêt qu'elles peuvent représenter dans les zones d'élevage dites «difficiles» (zones de montagne). Les travaux de caractérisation génétique menés sur les ovins du Maroc qui ne concernent que les

cinq principales races identifiées, ont été basés sur l'étude des microsatellites et des variants électrophorétiques des protéines sanguines (Ouragh et al. 2002), et ont montré un important polymorphisme génétique, en même temps qu'une faible différenciation entre races.

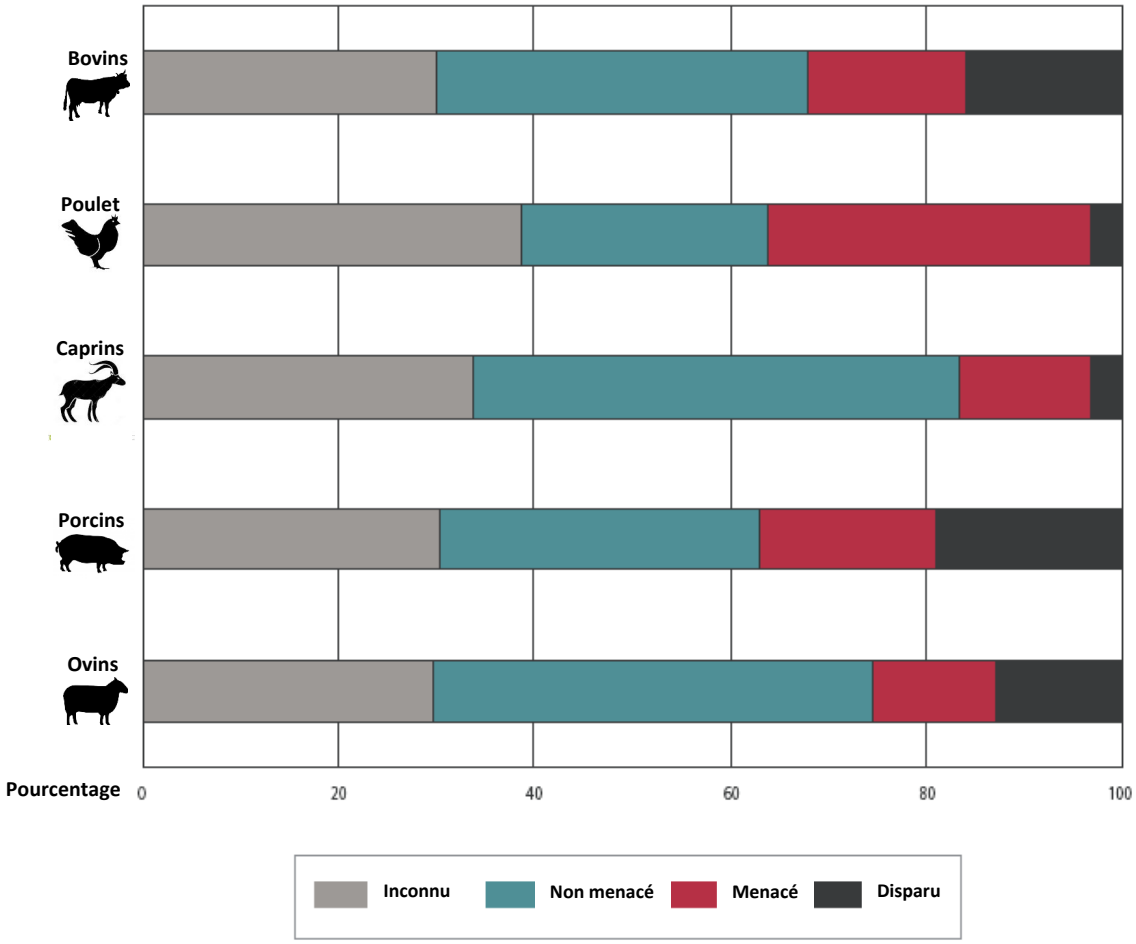

Figure 3. Situation de risque d'extinction des races des principales espèces d'élevage en 2006 (d'après FAO 2007).

Cette situation est différente de celle des caprins qui sont caractérisés par une grande diversité phénotypique et une importante hétérogénéité qu'ils doivent au brassage incontrôlé entre les différents types. Ceci a rendu difficile leur distinction en entités génétiques bien définies ou races. Toutefois, il y a eu la description de quelques populations typiques localisées dans des zones géographiques et présentant des caractéristiques assez homogènes. On a notamment différencié trois principales populations locales : la population Noire avec trois sous-populations (Atlas, Barcha et Ghazalia), la population Draa et la population du Nord. Plus récemment à partir de 2009, Atlas, Barcha et Ghazalia ont été reconnues officiellement comme races. Les élevages de ces populations sont généralement marqués par une faible

productivité qui n'assure qu'un niveau de revenu modeste aux producteurs. Les programmes d'amélioration génétique des caprins qui ont été adoptés ont visé uniquement certains troupeaux, et consistaient à effectuer des croisements avec des races étrangères ou alors à introduire ces dernières et les conduire en race pure. Par ailleurs, les études génétiques des populations caprines locales, ont été très peu nombreuses (basées sur des marqueurs microsatellites et gènes des caséines), et elles ont mis en évidence un fort polymorphisme génétique (Ouafi et al. 2002). Les études de la diversité de l'ADN mitochondrial et du chromosome Y ont mis en évidence une grande diversité très faiblement structurée selon les populations et la géographie. Ceci a été expliqué par la diversité des populations ayant colonisé le Maroc pour la première fois et des flux de gène récurrents entre les différentes populations et régions et des introgressions au nord du pays depuis la péninsule Ibérique via le détroit de Gibraltar (Pereira et al. 2009; Benjelloun et al. 2011).

4. Le projet NextGen

Dans ce contexte d'érosion massive de la biodiversité au sein des animaux d'élevage, le projet NextGen a été mis en œuvre. Il a été financé dans le cadre du 7ème programme-cadre de la Commission Européenne et avait une durée de quatre ans (avril 2010-mars 2014) pour un budget total de 3 millions €. L'objectif global du projet est de développer des méthodologies optimisées pour la préservation de la biodiversité des animaux d'élevage dans un contexte de disponibilité des données de génomes complets. Les principales composantes de recherche du projet sont:

(i) L'étude des relations génome-environnement chez les chèvres et moutons. L'objectif est de comprendre les mécanismes qui sous-tendent l'adaptation locale des petits ruminants à leur environnement d'élevage (éco-climatique) via une approche de génomique du paysage et en utilisant les données de génomes complets d'individus représentatifs de l'ensemble du gradient de variabilité environnementale au Maroc.

(ii) L'évaluation du potentiel des races locales et des ancêtres sauvages dans les centres de domestication comme ressources génétiques utiles pour la conservation à long terme de l'élevage des petit-ruminants. L'objectif est d'évaluer la diversité génétique et le potentiel adaptatif des populations locales de chèvres et de moutons en Iran et au Maroc et des populations de mouflons (*O. orientalis* et *O. vignei*) et d'aegagres (*C. aegagrus*) au niveau du centre probable de domestication dans le cadre de l'érosion génétique des animaux d'élevage.

En outre, les signatures de sélection associées au processus de domestication sont identifiées et les mécanismes probables régissant ce processus chez ces espèces sont élucidés.

(iii) La définition d'approches optimales d'échantillonnage de marqueurs représentatifs des variations de l'ensemble du génome pour évaluer correctement la biodiversité. L'objectif est de tester la précision et l'éventuel biais des techniques permettant d'échantillonner des marqueurs génomiques et d'évaluer leurs performances pour inférer la variabilité neutre et adaptative des génomes en comparaison avec les données de génomes complets.

(iv) Identifier les mécanismes sous-tendant la résistance des bovins aux maladies tropicales. L'objectif est d'identifier les gènes impliqués dans la résistance des vaches aux maladies répandues en Ouganda.

(v) Le développement de nouvelles technologies de cryoconservation et de nouvelles approches de sélection des animaux basés sur les données de génomes complets pour la cryoconservation.

Ainsi, NextGen est l'un des premiers projets ayant produit autant de génomes complets sur des espèces autres que les humains à l'échelle internationale (>400 génomes complets à un taux de couverture de 12x). En outre, la stratégie d'échantillonnage adoptée n'a jamais été implémentée à une telle échelle chez les animaux d'élevage et ouvre de nouvelles perspectives pour l'analyse des données via une approche de génomique du paysage. Enfin, l'approche intégrée permettrait d'implémenter un paquet technologique intégré pour la conservation de 3 des principaux animaux d'élevage dont les ressources génétiques subissent une érosion massive.

Cependant, ces aspects novateurs ont été naturellement associés à pas mal de défis. Le premier grand challenge était l'échantillonnage et le second était la production de données de séquences et le développement d'outils pour les traiter ainsi que la mise au point des approches de leurs analyses.

5. Le travail de thèse

Les travaux réalisés dans le cadre de cette thèse s'inscrivent globalement dans les perspectives NextGen qui sont liées à la conservation des petits ruminants dans un contexte de faisabilité des données de génomes complets. Ces travaux se placent également dans un cadre

de recherches fondamentales associées à la compréhension de la diversification des génomes et des bases génétiques de l'adaptation locale chez les mammifères.

5.1. Axes de recherche

Les travaux de cette thèse peuvent ainsi se décliner en trois principaux axes qui constituent les trois chapitres de ce document.

(i) Définition de stratégies adéquates pour l'échantillonnage des génomes afin d'étudier leur variabilité neutre et adaptative.

Malgré les avancées technologiques dans le domaine du séquençage, la production, le traitement ainsi que l'analyse des données de génomes complets de plusieurs individus restent laborieux et demandent des moyens importants. Par conséquent, les techniques permettant la réduction des coûts et de la complexité des génomes tout en gardant une représentativité correcte des génomes restent toujours sollicitées. Plusieurs approches réduisant la complexité des génomes sont très largement utilisées dans les études des processus démographiques et de l'adaptation locale. Elles vont des puces à SNPs de différentes densités qui peuvent génotyper jusqu'à 800K SNPs chez plusieurs espèces, e.g. (Kijas et al. 2012; Ramey et al. 2013) jusqu'au séquençage des génomes complets à faible taux de couverture, e.g. (Jansen et al. 2013) en passant par les techniques de RAD-seq (Miller et al. 2007; Baird et al. 2008), RNA-seq (Wilhelm et al. 2008; Mudge et al. 2011) ou la capture de l'exome, e.g. (Choi et al. 2009; Ng et al. 2009). Toutefois, peu d'études se sont penchées sur la question de la capacité de ces différentes techniques à représenter la variabilité neutre et/ou adaptative du génome. La question de fiabilité de ces données de génotypes/séquences et leur capacité à bien représenter la variabilité génomique ou au moins la partie du génome concernée par l'étude se pose avec acuité. De même, la communauté scientifique aussi bien que les professionnels dans certains domaines (e.g. l'agriculture) aurait besoin de techniques visant à réduire le coût et/ou l'effort de génotypage/séquençage tout en gardant des informations fiables sur la variabilité du génome.

Cet axe concerne ainsi un aspect méthodologique et a pour objectif d'évaluer la fiabilité (i.e. précision et biais) des principales méthodes courantes de génotypage/séquençage, et de fournir des informations permettant de développer des stratégies optimales pour échantillonner des marqueurs représentatifs des variations de l'ensemble du génome selon les objectifs des études. Nous avons abordé notamment l'étude de la diversité neutre, adaptative et du déséquilibre de liaison et nous avons comparé plusieurs stratégies d'échantillonnage des

génomes avec les données de génomes complets à 12x de couverture. Nous avons pris comme organismes modèles les petits ruminants domestiques au Maroc, en Iran et de races 'industrielles' ainsi que leurs ancêtres sauvages au centre de domestication.

(ii) Caractériser la diversité génomique neutre des populations indigènes des caprins au Maroc et identifier les gènes sous sélection au niveau de chacune de ces populations en utilisant les données de génomes complets.

La section 'Les petits ruminants' montre bien la richesse que représentent potentiellement les populations indigènes des animaux d'élevage et leurs ancêtres sauvages en terme et de qualités adaptatives et de variabilité résidente du génome (qui représente un potentiel adaptatif durable en soit). Cependant, cette richesse n'a jamais été sujette à une évaluation précise et à grande échelle. La caractérisation de la variabilité neutre et adaptative chez ces populations en utilisant les données de génomes complets permettrait de définir précisément cette richesse ainsi que leur rôle potentiel dans la préservation sur le long terme de ces espèces. Ladite section a également décrit la situation de l'élevage caprin au Maroc caractérisée globalement par la dominance des populations locales à 95% du troupeau national où les schémas de gestion actuels restent extensifs avec de faibles pressions de sélection. De même la position géographique a fait théoriquement de ce pays un point de rencontre de flux migratoires venant de l'Afrique du nord, l'Europe ou même de l'Afrique saharienne.

Dans le cadre de cet axe, nous avons utilisé les données de génomes complets (re-séquençage à 12x de couverture) de chèvres représentatives de la diversité géographique au sein des principales populations locales du Maroc afin de caractériser/évaluer la diversité neutre présente au sein de ces populations. De même, nous avons essayé d'identifier les signatures de sélection probables spécifiques à chacune de ces populations en faisant le lien avec leur variabilité phénotypique.

(iii) Etudier les bases génomiques de l'adaptation locale des moutons et chèvres aux conditions environnementales via une approche de génomique du paysage en utilisant les données massives de génomes complets.

Dans la section « L'adaptation locale », nous avons souligné l'importance des traits adaptatifs dans la persistance et la survie des populations. Cette section a passé en revue également la

complexité et la diversité des mécanismes sous-jacents à ces traits. Les études qui visent la compréhension de ces mécanismes sont de plus en plus nombreuses et confirment souvent leur multitude et leur variabilité selon l'espèce, la géographie,... (voir section « L'adaptation locale »). Les données de génomes complets permettent désormais d'avoir une bonne résolution pour identifier les mécanismes probables de l'adaptation locale. La génomique du paysage fournit des outils importants pour comprendre ces mécanismes. L'application de cette approche se limitait généralement aux populations sauvages en raison de l'effet important de la sélection anthropique sur les domestiques. Comme stipulé dans la section « Les petits ruminants », les chèvres et moutons sont bien adaptés et élevés dans des conditions environnementales très contrastées au Maroc. Les intervalles de température, précipitations, relief, saisonnalité, altitude et d'autres variables éco-climatiques sont très larges. En outre, les systèmes d'élevage sont dominés par le type « traditionnel » avec globalement une faible pression de sélection anthropique (comparés aux élevages « industriels »). Les petits ruminants de ce pays constitueraient ainsi des modèles adéquats pour examiner les bases génétiques de l'adaptation à l'environnement. De même, la présence des deux espèces (chèvres et moutons) dans les mêmes environnements présente un contexte idéal pour étudier la possible convergence adaptative chez ces deux espèces et améliorer nos connaissances sur les mécanismes régissant la distribution de la biodiversité chez les mammifères en général. La compréhension des mécanismes adaptatifs aux conditions environnementales « difficiles » nous permettrait en outre de raisonner les schémas de sélection pour l'amélioration durable des petits ruminants dans ce type de conditions.

Dans le cadre de cet axe, nous avons étudié les bases génétiques de l'adaptation locale chez les moutons et les chèvres du Maroc en se basant sur un échantillonnage à grande-échelle. Nous avons utilisé un système de grille de cellules rectangulaires (0.5°x0.5°) couvrant la grande part du pays (~400.000 km^2) (Figure 4) et représentant le gradient des conditions climatiques et écologiques présentes au Maroc. Les données de génomes complets (à 12x de couverture) de 160 moutons (et 161 chèvres respectivement) représentatifs de ce gradient ont été produites. Nous avons ensuite utilisé des méthodes différentes de criblage génomique pour rechercher les régions potentiellement sous sélection et identifier les fonctions physiologiques potentiellement impliquées dans l'adaptation locale.

Ainsi, ces trois axes ont constitué les trois différents chapitres de ce document. Le troisième axe représente une recherche qui est toujours en cours de réalisation, mais une ébauche d'article (en préparation) discutant les principaux résultats obtenus à ce jour est présentée

dans ce document. Bien entendu, nous allons continuer à travailler sur ces résultats en collaboration avec les autres membres du consortium NextGen.

Outre ces chapitres, nous avons estimé, pour traiter l'aspect méthodologique du premier chapitre, des résultats de plusieurs statistiques de génétique des populations à partir des données de génomes complets des petits ruminants sauvages et domestiques de plusieurs régions géographiques, y compris des races industrielles de moutons. Nous avons ainsi décrit ces statistiques au niveau de la « Discussion générale » pour dégager des éléments indicatifs sur l'état de la diversité génétique à l'échelle mondiale dans ces espèces.

Par ailleurs, comme pour tout le projet NextGen, la réalisation de ce travail de thèse était tributaire de la réussite de deux activités qui représentaient de véritables défis. Il s'agit de l'échantillonnage qui a commencé 2 ans avant le début du projet NextGen (en 2008) au Maroc et qui a pris fin en début 2012. La mise au point de cette opération a pris un temps considérable et l'opération elle même a été réalisée par une vingtaine de personnes réparties en quatre différentes équipes. Le deuxième défi consistait en la production et le traitement des données de génomes complets et l'adaptation de méthodologies et d'approches convenables pour leur analyse pendant la durée de la thèse. Ainsi, la production des données de séquences de tous les individus a pris un temps considérable et les données définitives de variation qui ont été utilisées dans ce travail n'ont été prêtes qu'en janvier 2014.

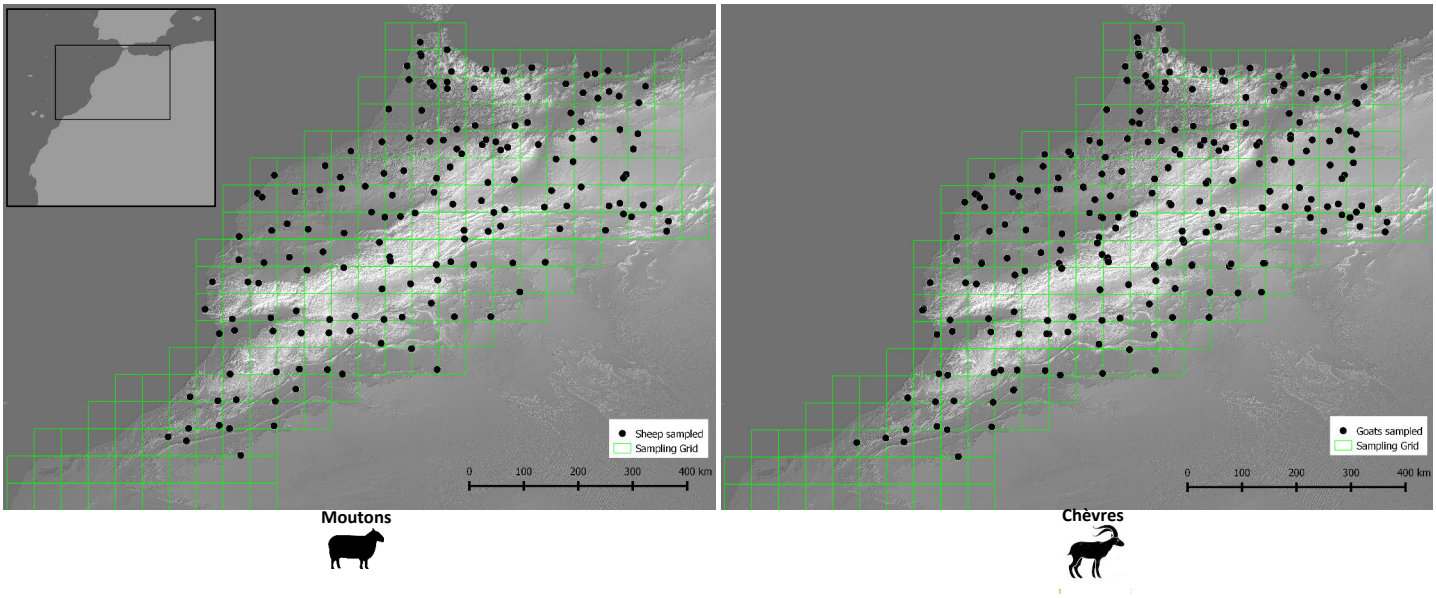

Figure 4. Répartition des ovins (gauche) et caprins (droite) inclus dans l'étude des bases génétiques de l'adaptation à l'environnement au Maroc.

Chaque point représente un individu.

5.2. Contribution personnelle

Mon travail de thèse étant inclus dans un projet collaboratif impliquant de nombreux partenaires internationaux ayant des compétences très diversifiées, il est nécessaire de préciser qu'elle a été ma contribution personnelle dans la production et l'analyse des données au sein de ce travail collectif. Mon implication a commencé avant la conception du projet NextGen en tant que chercheur à l'INRA-Maroc (partenaire n°9 du projet). Ainsi, j'ai assisté à la conception du projet et j'ai ensuite participé à la mise au point de l'opération de l'échantillonnage dont j'ai assuré la coordination pour le Maroc, tout en participant à sa réalisation. Ensuite, j'ai participé aux extractions d'ADN, et après la production des données de séquences par Adriana Alberti et Stefan Engelen au CEA-Genoscope d'Evry (Partenaire n°12 du projet), j'ai contribué à l'alignement des séquences brutes sur les génomes de référence et à la découverte des variants ainsi que le filtrage des données en collaboration avec Ian Streeter de l'EMBL-EBI (Partenaire n°5 du projet). J'ai également effectué des contrôles de qualité des données de génomes complets en les comparant aux données de génotypage par les puces à 50K SNPs dont nous avions sous-traité la production (description dans le Chapitre 1). Ensuite, j'ai réalisé les analyses des données décrites dans les différents chapitres de cette thèse, à l'exception de l'assemblage et de l'analyse des séquences d'ADN mitochondrial présenté au Chapitre 2 (réalisées par Florian Alberto et Frédéric Boyer au LECA) et de la production et traitement des données de variables environnementales et des analyses utilisant l'approche corrélative Samßada de détection de sélection décrites au niveau du Chapitre 3 (Stucki et al. 2014). Ces traitements et analyses ont été réalisés par nos collègues de l'EPFL (Partenaire n°5 du projet) Sylvie Stucki, Kevin Leempoel et Stéphane Joost.

Références

Ai H, Fang X, Yang B, Huang Z, Chen H, Mao L, Zhang F, Zhang L, Cui L, He W et al. 2015. Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nature Genetics* **47**(3): 217-+.

Akey JM. 2009. Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Research* **19**(5): 711-722.

Altshuler DM Durbin RM Abecasis GR Bentley DR Chakravarti A Clark AG Donnelly P Eichler EE Flicek P Gabriel SB et al. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422): 56-65.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *Plos One* **3**(10).

Barrett RDH, Rogers SM, Schluter D. 2009. ENVIRONMENT SPECIFIC PLEIOTROPY FACILITATES DIVERGENCE AT THE ECTODYSPLASIN LOCUS IN THREESPINE STICKLEBACK. *Evolution* **63**(11): 2831-2837.

Beja-Pereira A, Caramelli D, Lalueza-Fox C, Vernesi C, Ferrand N, Casoli A, Goyache F, Royo LJ, Conti S, Lari M et al. 2006. The origin of European cattle: Evidence from modern and ancient DNA. *Proceedings of the National Academy of Sciences of the United States of America* **103**(21): 8113-8118.

Benjelloun B, Pompanon F, Ben Bati M, Chentouf M, Ibnelbachyr M, El Amiri B, Rioux D, Boulanouar B, Taberlet P. 2011. Mitochondrial DNA polymorphism in Moroccan goats. *Small Ruminant Research* **98**(1-3): 201-205.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics* **74**(6): 1111-1120.

Black WC, Baer CF, Antolin MF, DuTeau NM. 2001. Population genomics: Genome-wide sampling of insect populations. *Annual Review of Entomology* **46**: 441-469.

Blanquart F, Gandon S, Nuismer SL. 2012. The effects of migration and drift on local adaptation to a heterogeneous environment. *Journal of Evolutionary Biology* **25**(7): 1351-1363.

Boulanouar B, et Benlekhal A. 2005. L'élevage ovin au Maroc : de la production à la consommation. In L'élevage du mouton et ses systèmes de production au Maroc. (eds. Boulanouar B. et Paquay R.), pp 3-32, INRA, Rabat, Morocco.

Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X et al. 2008. The potential and challenges of nanopore sequencing. *Nature Biotechnology* **26**(10): 1146-1153.

Chen J, Kallman T, Ma X, Gyllenstrand N, Zaina G, Morgante M, Bousquet J, Eckert A, Wegrzyn J, Neale D et al. 2012. Disentangling the Roles of History and Local Selection in Shaping Clinal Variation of Allele Frequencies and Gene Expression in Norway Spruce (Picea abies). *Genetics* **191**(3): 865-U377.

Chevin L-M, Lande R, Mace GM. 2010. Adaptation, Plasticity, and Extinction in a Changing Environment: Towards a Predictive Theory. *Plos Biology* **8**(4).

Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S, Sanjad S et al. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **106**(45): 19096-19101.

Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology* **4**(4): 265-270.

Crispo E, DiBattista JD, Correa C, Thibert-Plante X, McKellar AE, Schwartz AK, Berner D, De Leon LF, Hendry AP. 2010. The evolution of phenotypic plasticity in response to anthropogenic disturbance. *Evolutionary Ecology Research* **12**(1): 47-66.

Darwin C. 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life (eds John Murray), London, UK.

Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-Murci L, Robinson-Rechavi M, Excoffier L. 2013. Evidence for Polygenic Adaptation to Pathogens in the Human Genome. *Molecular Biology and Evolution* **30**(7): 1544-1558.

De Kort H, Vandepitte K, Bruun HH, Closset-Kopp D, Honnay O, Mergeay J. 2014. Landscape genomics and a common garden trial reveal adaptive differentiation to temperature across Europe in the tree species Alnus glutinosa. *Molecular Ecology* **23**(19): 4709-4721.

Dong Y, Xie M, Jiang Y, Xiao N, Du X, Zhang W, Tosser-Klopp G, Wang J, Yang S, Liang J et al. 2013. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). *Nature Biotechnology* **31**(2): 135-141.

Elena SF, Lenski RE. 2003. Evolution experiments with microorganisms: The dynamics and genetic bases of adaptation. *Nature Reviews Genetics* **4**(6): 457-469.

FAO. 2007. The State of the World's Animal Genetic Resources for Food and Agriculture in brief, (eds by Dafydd Pilling & Barbara Rischkowsky), Rome, Italy.

Fernandez H, Hughes S, Vigne J-D, Helmer D, Hodgins G, Miquel C, Hanni C, Luikart G, Taberlet P. 2006. Divergent mtDNA lineages of goats in an Early Neolithic site, far from the initial domestication areas. *Proceedings of the National Academy of Sciences of the United States of America* **103**(42): 15375-15379.

Fisher RA. 1919. XV. The Correlation between Relatives on the Supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh, 52, pp 399-433. doi:10.1017/S0080456800012163.

Fisher R. 1958. POLYMORPHISM AND NATURAL-SELECTION. *Bulletin of the International Statistical Institute* **36**(3): 284-289.

Fournier-Level A, Korte A, Cooper MD, Nordborg M, Schmitt J, Wilczek AM. 2011. A Map of Local Adaptation in Arabidopsis thaliana. *Science* **334**(6052): 86-89.

Frankham R. 2002. Population viability analysis. *Nature* **419**(6902): 18-19.

Frichot E, Schoville SD, Bouchard G, Francois O. 2013. Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution* **30**(7): 1687-1699.

Holderegger R, Wagner HH. 2006. A brief guide to landscape genetics. *Landscape Ecology* **21**(6): 793-796.

Huey RB, Gilchrist GW, Carlson ML, Berrigan D, Serra L. 2000. Rapid evolution of a geographic cline in size in an introduced fly. *Science* **287**(5451): 308-309.

Hughes AL. 1999. Adaptive Evolution of Genes and Genomes. pp. 270. Oxford University Press, New York, USA.

Jansen S, Aigner B, Pausch H, Wysocki M, Eck S, Benet-Pages A, Graf E, Wieland T, Strom TM, Meitinger T et al. 2013. Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. *Bmc Genomics* **14**.

Jiang Y, Xie M, Chen W, Talbot R, Maddox JF, Faraut T, Wu C, Muzny DM, Li Y, Zhang W et al. 2014. The sheep genome illuminates biology of the rumen and lipid metabolism. *Science* **344**(6188): 1168-1173.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**(7392): 55-61.

Jones MR, Forester BR, Teufel AI, Adams RV, Anstett DN, Goodrich BA, Landguth EL, Joost S, Manel S. 2013. INTEGRATING LANDSCAPE GENOMICS AND SPATIALLY EXPLICIT APPROACHES TO DETECT LOCI UNDER SELECTION IN CLINAL POPULATIONS. *Evolution* **67**(12): 3455-3468.

Joost S, Bonin A, Bruford MW, Despres L, Conord C, Erhardt G, Taberlet P. 2007. A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology* **16**(18): 3955-3969.

Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, Yang Y, Li S, Tang K, Chen H et al. 2013. Modeling Recent Human Evolution in Mice by Expression of a Selected EDAR Variant. *Cell* **152**(4): 691-702.

Kawecki TJ, Ebert D. 2004. Conceptual issues in local adaptation. *Ecology Letters* **7**(12): 1225-1241.

Kidd JM, Gravel S, Byrnes J, Moreno-Estrada A, Musharoff S, Bryc K, Degenhardt JD, Brisbin A, Sheth V, Chen R et al. 2012. Population Genetic Inference from Personal Genome Data: Impact of Ancestry and Admixture on Human Genomic Variation. *American Journal of Human Genetics* **91**(4): 660-671.

Kijas JW, Lenstra JA, Hayes B, Boitard S, Neto LRP, San Cristobal M, Servin B, McCulloch R, Whan V, Gietzen K et al. 2012. Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. *Plos Biology* **10**(2).

Kruglyak L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* **22**(2): 139-144.

Linnen CR, Poh Y-P, Peterson BK, Barrett RDH, Larson JG, Jensen JD, Hoekstra HE. 2013. Adaptive Evolution of Multiple Traits Through Multiple Mutations at a Single Gene. *Science* **339**(6125): 1312-1316.

Loman NJ, Watson M. 2015. Successful test launch for nanopore sequencing. *Nature Methods* **12**(4): 303-304.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. 2003. The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics* **4**(12): 981-994.

Macnair MR. 1993. THE GENETICS OF METAL TOLERANCE IN VASCULAR PLANTS. *New Phytologist* **124**(4): 541-559.

Manel S, Holderegger R. 2013. Ten years of landscape genetics. *Trends in Ecology & Evolution* **28**(10): 614-621.

Manel S, Schwartz MK, Luikart G, Taberlet P. 2003. Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution* **18**(4): 189-197.

MARA. 1980. Plan moutonnier. Direction de l'élevage. Division de la Production Animale, Rabat, Morocco.

Marsden CD, Lee Y, Kreppel K, Weakley A, Cornel A, Ferguson HM, Eskin E, Lanzaro GC. 2014. Diversity, Differentiation, and Linkage Disequilibrium: Prospects for Association Mapping in the Malaria Vector Anopheles arabiensis. *G3-Genes Genomes Genetics* **4**(1): 121-131.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* **17**(2): 240-248.

Mudge JM, Frankish A, Fernandez-Banet J, Alioto T, Derrien T, Howald C, Reymond A, Guigo R, Hubbard T, Harrow J. 2011. The Origins, Evolution, and Functional Potential of Alternative Splicing in Vertebrates. *Molecular Biology and Evolution* **28**(10): 2949-2959.

Naderi S, Rezaei H-R, Pompanon F, Blum MGB, Negrini R, Naghash H-R, Balkiz O, Mashkour M, Gaggiotti OE, Ajmone-Marsan P et al. 2008. The goat domestication process inferred from large-scale mitochondrial DNA analysis of wild and domestic individuals. *Proceedings of the National Academy of Sciences of the United States of America* **105**(46): 17659-17664.

Natarajan C, Hoffmann FG, Lanier HC, Wolf CJ, Cheviron ZA, Spangler ML, Weber RE, Fago A, Storz JF. 2015. Intraspecific Polymorphism, Interspecific Divergence, and the Origins of Function-Altering Mutations in Deer Mouse Hemoglobin. *Molecular Biology and Evolution* **32**(4): 978-997.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE et al. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**(7261): 272-U153.

Ouafi AT, Babilliot JM, Leroux C, Martin P. 2002. Genetic diversity of the two main Moroccan goat breeds: phylogenetic relationships with four breeds reared in France. *Small Ruminant Research* **45**(3): 225-233.

Ouragh L, Amigues Y, Nguyen TC, Boscher MY. 2002. Analyses génétique des races ovines marocaines - Genetic analysis of Moroccan sheep breeds. 9èmes Rencontres Recherches Ruminants, 4 et 5 Décembre 2002, pp 99, Paris, France.

Pereira F, Queiros S, Gusmao L, Nijman IJ, Cuppen E, Lenstra JA, Davis SJM, Nejmeddine F, Amorim A, Econogene C. 2009. Tracing the History of Goat Pastoralism: New Clues from Mitochondrial and Y Chromosome DNA in North Africa. *Molecular Biology and Evolution* **26**(12): 2765-2773.

Peters J, von den Driesch A, Helmer D. 2005. The upper Eurphrates-Tigris basin: cradle of agro-pastoralism?. In: *The First Steps of Animal Domestication. New Archaeological Approaches* (eds Vigne JD, Peters J, Helmer D), pp. 96–124. Oxbow Books, Oxford, UK.

Porter V. 2002. *Mason's world dictionary of livestock breeds, types and varieties*, Wallingford, UK.

Prasad KVSK, Song B-H, Olson-Manning C, Anderson JT, Lee C-R, Schranz ME, Windsor AJ, Clauss MJ, Manzaneda AJ, Naqvi I et al. 2012. A Gain-of-Function Polymorphism Controlling Complex Traits and Fitness in Nature. *Science* **337**(6098): 1081-1084.

Ramey HR, Decker JE, McKay SD, Rolf MM, Schnabel RD, Taylor JF. 2013. Detection of selective sweeps in cattle using genome-wide SNP data. *Bmc Genomics* **14**.

Rezaei S. 2007. Phylogénie moléculaire du Genre *Ovis* (Mouton et Mouflons), Implications pour la Conservation du Genre et pour l'Origine de l'Espèce Domestique. Thèse de Doctorat. Université Joseph Fourier. pp 162 p, Grenoble, France.

Savolainen O, Lascoux M, Merila J. 2013. Ecological genomics of local adaptation. *Nature Reviews Genetics* **14**(11): 807-820.

Simonson TS, Yang Y, Huff CD, Yun H, Qin G, Witherspoon DJ, Bai Z, Lorenzo FR, Xing J, Jorde LB et al. 2010. Genetic Evidence for High-Altitude Adaptation in Tibet. *Science* **329**(5987): 72-75.

Snyder M, Du J, Gerstein M. 2010. Personal genome sequencing: current approaches and challenges. *Genes & Development* **24**(5): 423-431.

Storfer A, Murphy MA, Spear SF, Holderegger R, Waits LP. 2010. Landscape genetics: where are we now? *Molecular Ecology* **19**(17): 3496-3514.

Stucki S, Orozco-terWengel P, Bruford MW, Colli L, Masembe C, Negrini R, Taberlet P, Joost S. 2014. High performance computation of landscape genomic models integrating local indices of spatial association. *arxiv* **1405.7658v2**.

Taberlet P, Valentini A, Rezaei HR, Naderi S, Pompanon F, Negrini R, Ajmone-Marsan P. 2008. Are cattle, sheep, and goats endangered species? *Molecular Ecology* **17**(1): 275-284.

Tapio M, Tapio I, Grislis Z, Holm LE, Jeppsson S, Kantanen J, Miceikiene I, Olsaker I, Viinalass H, Eythorsdottir E. 2005. Native breeds demonstrate high contributions to the molecular variation in northern European sheep. *Molecular Ecology* **14**(13): 3951-3963.

van't Hof AE, Edmonds N, Dalikova M, Marec F, Saccheri IJ. 2011. Industrial Melanism in British Peppered Moths Has a Singular and Recent Mutational Origin. *Science* **332**(6032): 958-960.

Vincent B, Dionne M, Kent MP, Lien S, Bernatchez L. 2013. LANDSCAPE GENOMICS IN ATLANTIC SALMON (SALMO SALAR): SEARCHING FOR GENE-ENVIRONMENT INTERACTIONS DRIVING LOCAL ADAPTATION. *Evolution* **67**(12): 3469-3487.

Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**(7199): 1239-U1239.

Wright S. 1931. Evolution in Mendelian populations. *Genetics* **16**(2): 0097-0159.

CHAPITRE 1: Echantillonnage représentatif des données de génomes complets

Chapitre 1: Echantillonnage représentatif des données de génomes complets

Résumé et présentation de l'article

Le développement récent des nouvelles technologies de séquençage a rendu possible l'utilisation des données de génomes complets (WGS) dans les études de génétique des populations afin d'étudier les processus démographiques et rechercher les gènes sous sélection. Ceci a ouvert de nouveaux horizons pour comprendre la répartition de la diversité ainsi que la variabilité le long des génomes. Cependant la production et le traitement de ces données nécessitent des moyens et des efforts importants ce qui limite leur utilisation. Par conséquent, de nouvelles techniques de génotypage sont de plus en plus développées et sont largement utilisées (e.g. puces à ADN, Rad-seq, RNA-seq, capture de l'exome, etc.). Elles consistent généralement à génotyper un échantillon de marqueurs le long ou sur une partie des génomes. De même, il y a moyen de réduire le coût de séquençage en réduisant le taux de couverture (nombre de lecture moyen de chaque base du génome). Cependant, l'utilisation de ces stratégies est confrontée à la question clé de leur représentativité de la variation génomique et de sa répartition.

Dans ce chapitre qui traite ce volet méthodologique, nous essayons de répondre à cette question en testant le biais et/ou la précision de plusieurs approches d'échantillonnage de génomes qui sont largement utilisées en les comparant aux données de génomes complets produites à 12X de couverture. Nous avons produit 173 génomes de chèvres et moutons au Maroc et en Iran et de races « industrielles » ainsi que des populations sauvages. A partir de ces données, nous avons extrait 946 différents jeux de données qui correspondent à différentes stratégies d'échantillonnage. Nous avons extrait les génotypes qui correspondent à deux puces à ADN ovines à moyenne (≈50K SNP) et à haute (≈600K SNPs) densité de marqueurs et à une puce caprine à 54K SNPs. Nous avons également extrait les génotypes correspondant à la capture de l'exome ovin. En outre, nous avons extrait des jeux de données qui correspondent à des panels de 1K à 5M de variants aléatoires répartis sur les génomes. Nous avons également testé la variation de la taille d'échantillons entre 5 et 48 individus. Outre les panels de variants aléatoires et non aléatoires, nous avons extrait à partir des données brutes à 12X des séquences pour simuler les données de génomes complets à des taux de couverture de 1X, 2X et 5X.

Cette étude montre que des panels de 5K à 10K variants aléatoires avec une densité d'un variant tous les 500 à 250 kb sont suffisants pour avoir une estimation assez précise de la diversité génomique neutre globale (i.e. la diversité nucléotidique, l'hétérozygotie, le coefficient de consanguinité,…). En outre, les panels à 10K variants extraits d'une population donnée (moutons ou chèvres marocains) permettent une estimation non biaisée de la diversité dans d'autres populations ainsi que de la différenciation entre ces populations (domestiques iraniens, « industriels » et sauvages). Par contre, les puces à ADN et la capture de l'exome, bien qu'elles présentent des densités de marqueurs plus élevés engendrent un biais considérable et renversent l'ordre de diversité des populations dans certains cas. Par ailleurs, notre étude suggère que la détection de signatures de sélection ou le calcul du déséquilibre de liaison requièrent des densités plus élevées de variants d'au moins 1M (1 variant tous les ≈3kb). L'un des résultats importants de cette étude est que le taux de couverture de 5X peut être suffisant pour avoir une estimation assez précise de la variation génomique contrairement aux taux de couverture plus faibles.

Cette étude nous permet de conclure que le biais de recrutement des puces à ADN est conséquent et que celui-ci doit être pris en compte lors des études de génétique des populations utilisant ces techniques. De même, ce chapitre ouvre la voie vers d'autres approches pour définir des marqueurs génétiques (i.e. choix d'un panel aléatoire) lors de la conception de nouvelles puces. Notre étude définit des nombres de variants minimums selon l'objectif de l'étude, et qui peuvent être caractérisés via des techniques de génotypage par séquencage. Par ailleurs, cette étude permet de conclure que le taux de couverture peut représenter un paramètre important sur lequel on peut agir pour réduire le coût de séquençage sans perdre énormément de précision dans l'estimation de la variation génomique. Ces résultats peuvent bien être généralisés à d'autres espèces de mammifères où l'ordre de grandeur et la variabilité du déséquilibre de liaison ne sont pas très différents de nos modèles.

Nous présentons ce chapitre sous forme d'article finalisé qui sera soumis très prochainement.

Article A: What information at which cost? The reliability of variant panels and low-coverage WGS for describing genome diversity

Running head: Surrogates for whole genome sequences

Badr Benjelloun[1,2,3], Frédéric Boyer[1,2], Ian Streeter[4], Wahid Zamani[1,2], Stefan Engelen[5], Adriana Alberti[5], Mohamed BenBati[3], Mustapha Ibnelbachyr[6], Mouad Chentouf[7], Abdelmajid Bechchari[8], Florian J. Alberto[1,2], Hamid R. Rezaei[9], Saeid Naderi[10], Alessandra Stella[11], Abdelkader Chikhi[6], Laura Clarke[4], James Kijas[12], Paul Flicek[4], Pierre Taberlet[1,2], François Pompanon[1,2]

To be submitted to Plos Genetics

[1] Université Grenoble-Alpes, Laboratoire d'Ecologie Alpine, F-38000 Grenoble, France

[2] CNRS, Laboratoire d'Ecologie Alpine, F-38000 Grenoble, France

[3] National Institute of Agronomic Research (INRA Maroc), Regional Centre of Agronomic Research, Beni-Mellal, Morocco

[4] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

[5] CEA-Institut de Génomique, Genoscope, Centre National de Séquençage, France.

[6] National Institute of Agronomic Research (INRA_Mor), CRRA Errachidia, Morocco

[7] National Institute of Agronomic Research (INRA_Mor), CRRA Tangier, Morocco

[8] National Institute of Agronomic Research (INRA_Mor), CRRA Oujda, Morocco

[9] Environmental Sciences Department, Gorgan University of Agriculture and Natural Resources, Gorgan, Iran

[10] Natural Resources Faculty, University of Guilan, Guilan, Iran

[11] Parco Technologico Padano, Lodi, Italy

[12] Commonwealth Scientific and Industrial Research Organisation Animal Food and Health Sciences, St Lucia, QLD 4067, Australia.

Abstract

Recent technological advances made possible the development of several genotyping methods for a powerful characterization of genome diversity. However, the reduction of genome complexity through these methods can lead to biased and/or imprecise estimations of genome variations, and it is critical to choose an appropriate genotyping strategy to get an accurate sample of genome variations. In that context, we produced 173 Whole Genome Sequences (WGS) at 12X coverage for ovine and caprine domestic animals and their wild relatives. More than 43.4 M variants were discovered in sheep (*Ovis aries*), 31.7 M in goats (*Capra hircus*), 29.2 M in *O. orientalis* and 17.4 M in *C. aegagrus*. From these data we generated 946 datasets corresponding to random panels of 1K to 5M variants, commercial 50K and HD (600K) SNP chips and exome capture, for sample sizes of 5 to 48 individuals. We also simulated low-coverage genome re-sequencing of 1X, 2X and 5X by randomly sub-sampling reads from the 12X re-sequencing data. Our main goals were to assess the influence of the variant panel and sample size on the estimation of population genomic parameters and on the ability of detecting a known signature of selection. Globally, 5K to 10K random variants (representing one variant each 500 Kb to 260 Kb) were enough for an accurate estimation of genome diversity. Moreover, 10K variants randomly chosen within one population gave unbiased estimates for worldwide panels of other populations within a species. Conversely, commercial panels (i.e., 50K – 600K SNP chips) and exome capture displayed strong ascertainment biases and even sometimes modified the ranking of populations based on diversity estimates. Besides the characterization of the neutral diversity, the detection of the signature of selection and the accurate estimation of linkage disequilibrium required panels of at least 1 M variants. Finally, whole genome re-sequencing coverage of at least 5X appeared to be necessary for accurate assessment of genomic variations.

Author summary

Although whole genome sequences (WGS) greatly increase the ability to infer demographic processes using population genomics approaches, they are still not affordable for a representative number of individuals/populations and their handling remains time consuming and requires high performance computing. Here we produce 173 WGS at 12X coverage in a worldwide panel of populations of sheep and goats and their wild relatives Asiatic mouflon and bezoars. We attempt then to define samples of random and non-random variants allowing accurate inferences of whole genome variation. In addition, we tested the accuracy of low-coverage genome sequencing to study the individual genomic variability. We find that commonly used techniques designed to reduce genome complexity (i.e. medium and high density commercial SNP chips and exome capture) display substantial ascertainment bias to estimate most population genomics indices. Conversely, panels of 5,000 to 10,000 random variants extracted from the WGS are effective to study the genomes. However, studying linkage disequilibrium or selection signatures requires high density of markers (i.e. 1 million or more). Moreover, a sequencing coverage of 5X appears to be sufficient to obtain reliable information, but lower-coverage (i.e. <=2X) should be used cautiously with the commonly used variant calling approaches.

Introduction

Demographic and adaptive processes such as migration, genetic bottleneck and selection are evolutionary forces that have imprinted genomes. Combined with genetic processes such as recombination they result in a non-uniform distribution of genetic variations across the genomes. Since the middle of the last century [1,2], population genetics has been providing theoretical models to infer how these processes have shaped evolution by studying genetic variations among individuals, populations or species. This set up a conceptual framework for inferring the role of these processes from the study of current genetic variations. Thus, a mandatory prerequisite of evolutionary studies has been the design of panels of molecular markers representative of genome variations. This step has always been challenging. Until the last decade, the efficiency of molecular markers was mainly limited by technical issues. Co-dominant markers such as microsatellites give access to allelic frequencies and are informative for inferring demographic processes (e.g. [3,4]) but a maximum of a few dozen markers was usually genotyped. Other markers such as Amplified Fragment Length Polymorphism (AFLP) are more representative of whole genome variations, as a few hundred can be genotyped simultaneously, but they are dominant and do not give access to allelic frequencies. Even co-dominant markers that can be genotyped across the whole genome such as Single Nucleotide Polymorphisms (SNPs) were first revealed in limited numbers, e.g. [5]. Due to these limitations, there was a strong risk for misestimating whole genome variations and linkage disequilibrium [6], and to detect inaccurately both past demographic and selection signatures. Recent technological developments made possible the typing of very large numbers of co-dominant markers, mostly SNPs, which considerably increased the representativeness of the genomes and lead to the development of population genomics approaches [7-10]. So far, such an approach has essentially been limited to model organisms because of the need for whole genome reference data. Today, methods that overcome this challenge have been developed [11,12], and it is possible to manage whole genome data on nearly any studied species. Moreover, references genomes now have been produced for more and more species [13-15] [16]. Nevertheless, producing reference genomes and setting up population genomic studies by re-sequencing whole genomes of several individuals at sufficient coverage remains both costly and computationally time consuming. Thus, several strategies have been developed in order to reduce these costs while aiming to keep reliable and representative information of whole genome variations. One strategy is to reduce the depth of coverage of the WGS data to obtain information on the whole genome. A few studies

recently promoted the use of low to medium coverage WGS [17-19]. Nevertheless there is a strong risk of losing accuracy in variant calling and individual genotyping. Simulation studies (e.g. [17,20-23]) have proposed to increase either the number of samples, the number of sequenced individuals or to sequence key individuals to overcome these problems, but the reliability of low to medium coverage in WGS has not been experimentally/empirically evaluated yet. A second strategy to reduce the costs is to avoid whole genome sequencing and genotype a panel of a limited number of variants. For instance, commercial DNA chips or arrays for SNP typing are already available for several species (e.g. human, cattle, sheep, chicken) and can be designed for the purpose of any species. The Restriction-site Associated DNA sequencing (RAD-seq) method reduces genome complexity by re-sequencing stretches of genomic DNA adjacent to restriction endonuclease sites [24,25]. The RNA-seq method gives access to the transcriptome by sequencing the complementary DNA (cDNA) [26,27]. Genome enrichment methods allow the extraction of targeted regions of the genome, and one main application is the exome capture used for sequencing protein-coding regions [28-31], which can be used in population genomics studies [32,33]. However, when using these approaches we face the key question of their ability to produce accurate genotypes. The panel of genotyped variants should reliably represent genome variations for all studied individuals to avoid the ascertainment bias that results in the misestimating of genetic parameters. Only very few studies evaluated the accuracy of such genotyping approaches, and showed the impact of ascertainment bias on measures of population divergence [34]. Moreover, until now, no study evaluated the impact of subsampling panels of variants compared to WGS data, when studying genome diversity, population genetic structure and genes under selection.

In this context, our study aimed at assessing the accuracy of different variant subsampling methods for describing whole genome diversity. We produced 173 WGS at 12X coverage for four mammal species: sheep (*Ovis aries*), goat (*Capra hircus*) and their closely related wild species, the Asiatic mouflon (*Ovis orientalis*) and the bezoar (*Capra aegagrus*). From these WGS data, we extracted panels of genomic variants corresponding to different genome sampling strategies (i.e., exome capture, commercial SNP chips or random panels) in order to evaluate the impact of variants subsampling on the estimation of genome diversity and on the detection of a selection signature. This allowed defining appropriate variant densities for population genomic studies. We also simulated lower re-sequencing coverage to evaluate the impact of sequencing depth on the assessment of whole genome diversity.

Results

The assessment of random and non-random variant panels was made through two successive rounds. The first round was the identification of possible surrogates for the WGS by subsampling and testing random and non-random variant panels among a wide range of densities in Moroccan and wild animals. Variant calling was done using 12X coverage whole genome sequencing data for each species. It allowed the discovery of 29.96, 29.04, 21.71 and 17.32 million polymorphic variants for Moroccan sheep, Asiatic mouflon, Moroccan goats and bezoars, respectively (see Table 1). In the second round, the putative surrogates for the WGS (variant panels) that were identified in Morrocan sheep and goats were tested in other populations and their performances were compared to the WGS and non-random panels. Variant calling was done in Iranian domestics and genotypes corresponding to the discovered bi-allelic SNPs in the Ovis genus (i.e. 52,201,792) were inferred for a worldwide Industrial sheep group. Lastly, low-coverage re-sequencing data were generated by randomly sampling various fixed percentages of reads from the raw 12X re-sequencing data (see 'Material and Methods' section) and genotypes were compared to the 12X WGS variants.

Estimation of population genomics statistics

- Genetic diversity within groups

We assessed the effect of variants subsampling on describing genetic diversity by comparing the observed heterozygosity (*Ho*), inbreeding coefficient (*F*), nucleotide diversity (π) and Linkage disequilibrium ($r^2_{0.15}$) estimated with subsamples of variants to that calculated with the WGS dataset (see Table 1). Even at low-densities, random panels of variants gave generally information similar to that from WGS. Accurate estimates were obtained with all random panels of 5K or more markers for inbreeding (*F*) and nucleotide diversity (π), (Figures 1, S1, S2, S3) and with random panels of 10K or more markers for observed heterozygosity (*Ho*) (Figure S4). On the contrary, non-random panels of variants strongly biased estimations in general. The *ovine* 50K SNP and HD BeadChips and the *caprine* 50K SNP BeadChip from Illumina® highly overestimated π and *Ho*. For example, in the 30 Moroccan individuals there was an overestimation of 129%, 108% and 194% for π and 61%, 47% and 102% for *Ho* for these three panels, respectively. The dataset simulating exome capture underestimated π and *Ho* (e.g., underestimation of 20% and 6% for π and 8% and 5%

for *Ho* in Moroccan sheep and Asiatic mouflon, respectively). The underestimation of the inbreeding coefficient (F) was higher in domestic but not in wild animals.

Table 1. Population genomics statistics from WGS data for wild and Moroccan domestic small ruminants.

Species/Populations	*Ovis*		*Capra*	
	Moroccan sheep	Asiatic mouflon	Moroccan goats	Bezoars
Number of individuals (*n*)	30	14	30	18
Number of variants	43,478,084	29,274,713	31,775,474	17,449,771
Number of polymorphic variants	29,958,788	29,039,121	21,709,831	17,321,976
Short indels	2,805,416	2,713,334	2,139,714	1,344,653
Variants with > 2 alleles	817,859	265,998	219,706	109,520
Heterozygosity (*Ho*)	0.222 ± 0.026	0.223 ± 0.032	0.189 ± 0.018	0.194 ± 0.025
Inbreeding coefficient (*F*)	0.061 ± 0.108	0.186 ± 0.118	0.056 ± 0.092	0.182 ± 0.106
Linkage disequilibrium $r^2_{0.15}$ (Kb)	5	2.65	3.53	3.94
Nucleotide diversity (π)	0.165	0.273	0.137	0.237

At least 1M random markers in Moroccan sheep and 500K random markers in the other groups were necessary to have an estimation of LD ($r^2_{0.15}$) similar to that obtained with the WGS dataset. Smaller random panels and non-random panels biased this estimation (Figures 2, S5, S6). Exome capture especially biased the *LD* estimation in Moroccan sheep (but not in Asiatic mouflon with 10 individuals and more). Moreover, in all groups, decreasing the number of individuals highly increased $r^2_{0.15}$. In particular, Asiatic mouflon had an $r^2_{0.15}$ of 2.65Kb for 14 individuals and of 92.9Kb for 5 individuals (Figure S5).

- Genetic structure and differentiation

We assessed the influence of the variant panels on two methods describing the genetic structure and differentiation of wild versus Moroccan populations.

First, we estimated the Weir & Cockerham [35] differentiation index (*Fst*), which was rather high between wild and domestic animals (*Fst* = 0.105 in *Ovis* and *Fst* = 0.087 in *Capra* from WGS data; Figures 3, S7). Independently of the number of variants used, there was a strong sampling effect due to the individuals selected for estimating *Fst*. For a given set of individuals the number of random variants did not influence greatly the mean *Fst* values compared to that obtained with WGS data. The smallest random panels (from1K to 50K) increased the variance in *Fst* estimates among marker-set replicates for a given set of

individuals (Figures 3, S7). The caprine 50K SNP BeadChipIllumina® overestimated *Fst* values by 28% on average (Figure 3) and the ovine 50K and HD SNP BeadChips, and the exome capture slightly underestimated the *Fst* (2 to 13%). However, all non-random panels kept the ranking found with the WGS datasets for *Fst* estimated with different sets of individuals (r always > 0.98). Except for the caprine 50K BeadChip, the effect of the subsampling strategy on the *Fst* estimation was lower than that of the sample size.

Second, we used the clustering method implemented in sNMF [36] to estimate individual ancestry coefficients. The estimations depended neither on the number of markers used nor on the number of individuals in the sample. For the most likely number of clusters (K=2 for *Ovis* and *Capra* from the sNMF cross-validation values [36]), all variant panels led to similar results (Figure S8).

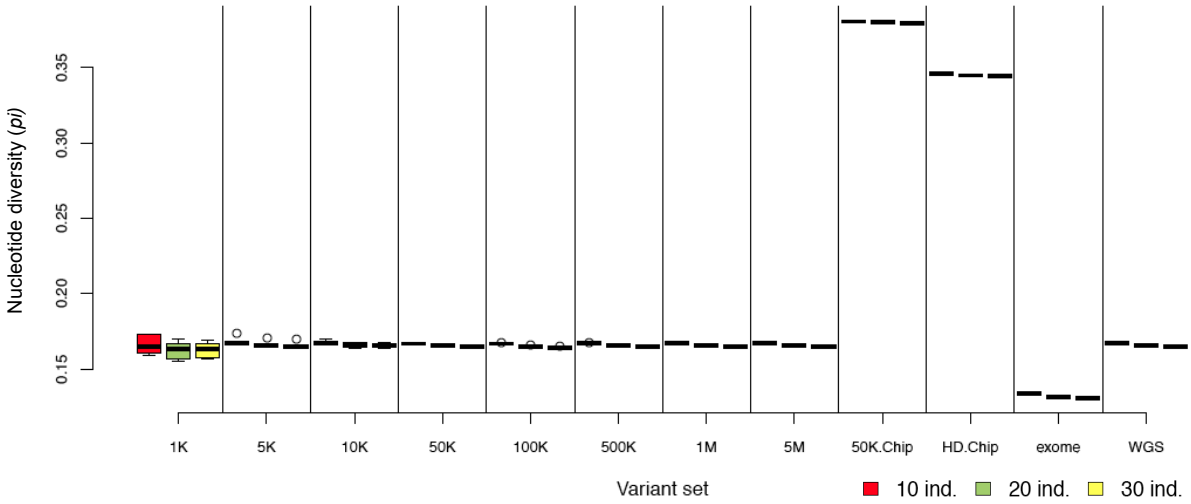

Figure 1. Nucleotide diversity (π) in Moroccan sheep calculated from WGS data and from random and non-random panels of variants.

Nucleotide diversity (π) was estimated for each replicate of the different numbers of variants of the random panels and for each non-random panel. Sample sizes varied for each estimate from 10 to 30 individuals.

Random panels are denoted by their number of variants (from 1K to 5M) and non-random panels by: 50K.Chip (Illumina® ovine 50K SNP Beadchip), HD.Chip (Illumina® ovine HD Beadchip) exome (exome capture simulation), WGS (all variants extracted from whole genome sequences). For each panel of variants the sample sizes are from left to right: 10 (red), 20 (green) and 30 (yellow) individuals.

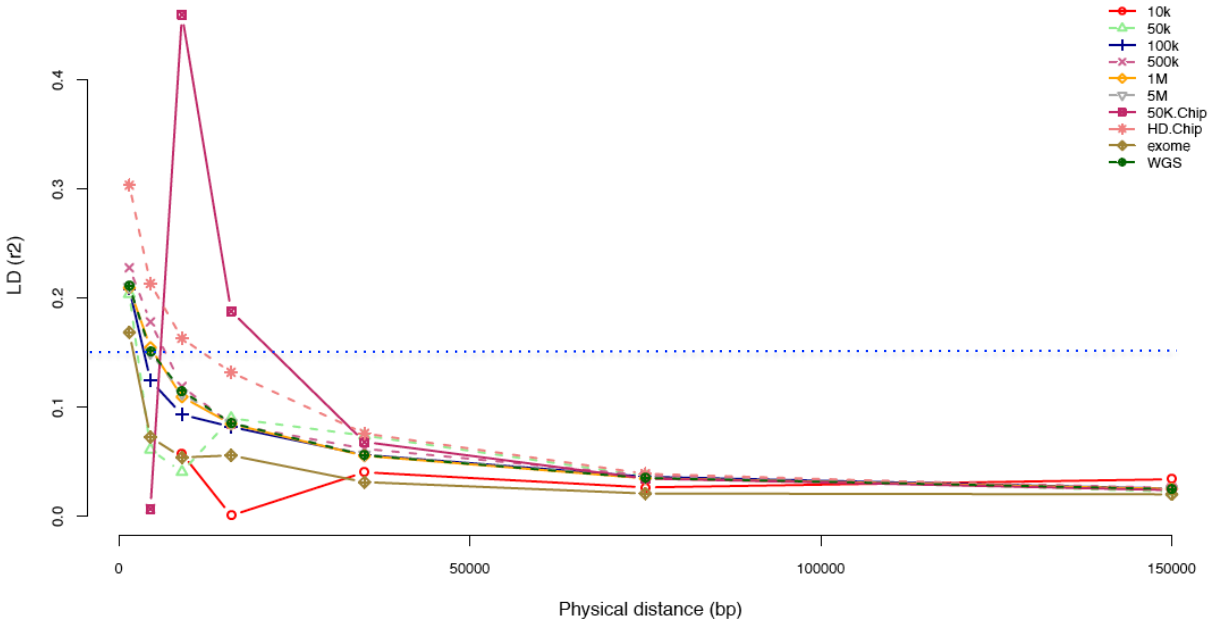

Figure 2. Decay of Linkage disequilibrium as a function of physical distance for different panels of variant in Moroccan sheep.

The Linkage Disequilibrium (r^2) was calculated on 30 Moroccan sheep. Inter-SNP distances (bp) were binned into the following classes: 1K-3K; 3K-6K; 6K-12K; 12K-20K; 20K-50K; 50K-100K; 100K-200K. Random and non-random panels are denoted according to Figure 1.

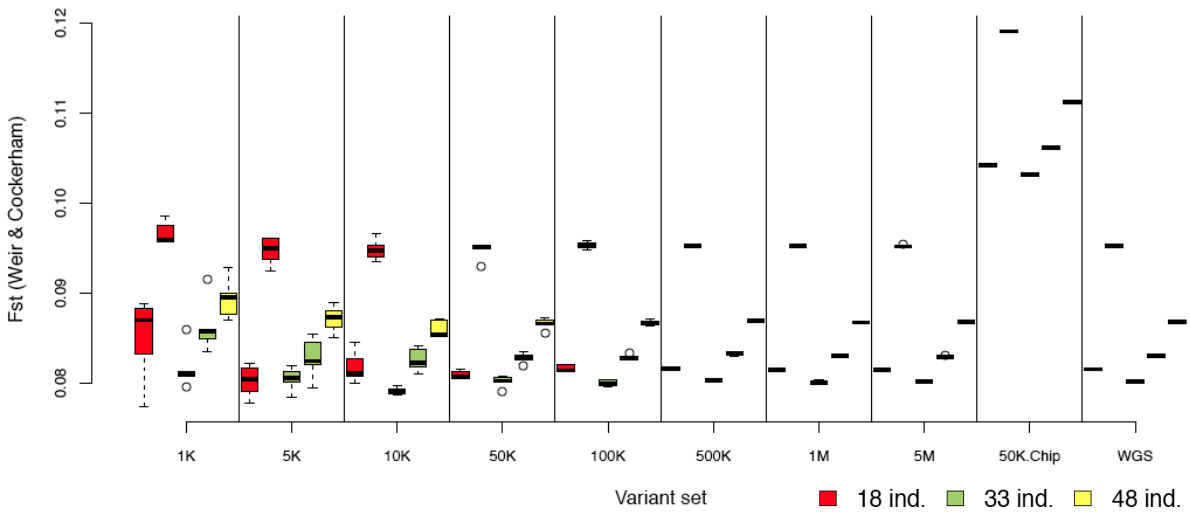

Figure 3. Fixation index (*Fst*) between Moroccan goats and bezoars for different panels of variants and different samples of individuals.

The fixation index *Fst* [35] was estimated for each random panel for the 5 independent replicates , and for each non-random dataset for each sample size. Random and non-random panels are denoted according to Figure 1.

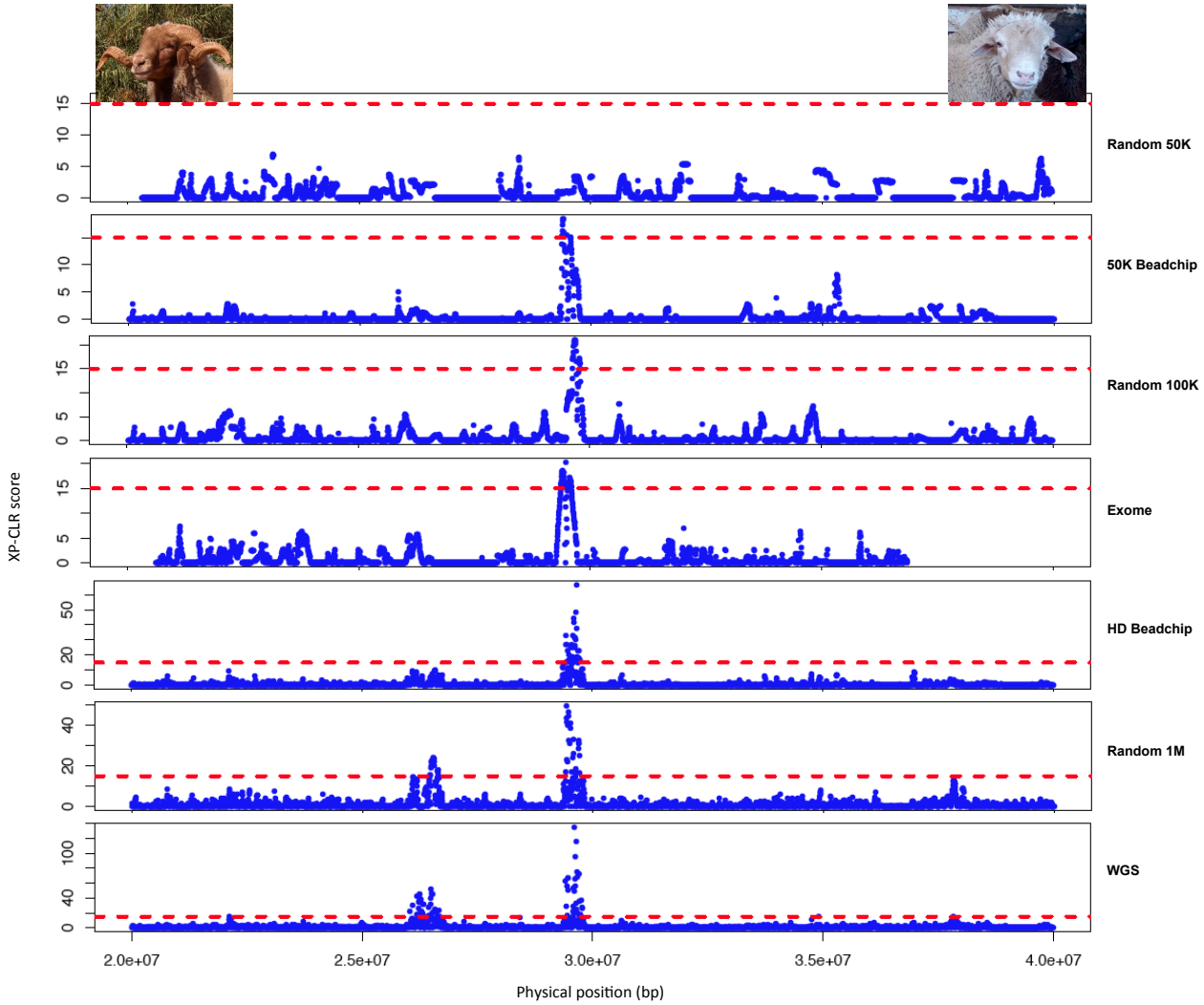

Figure 4. XP-CLR scores calculated along the 20M-40M bp segment on chromosome 10 in a horned-polled Moroccan sheep comparison for different sets of variants.

The two peaks of XP-CLR scores showed in the WGS data plot are located respectively in the two genes *NBEA* (chr 10: 26,007,917 - 26,592,574) and *MAB21L1* (chr 10: 26,231,353 - 26,232,432) and in the *RXFP2* gene (chr 10: 29,454,677 - 29,502,617 bp). The horizontal dashed line represents a XP-CLR score of 15 to represent a scale among the different plots.

- Detection of a signature of selection

By contrasting 15 horned and 15 polled Moroccan sheep, the XP-CLR method [37] applied on the WGS dataset allowed detection of the signal of selection previously reported on the Relaxin/Insulin-Like Family Peptide Receptor 2 gene *RXFP2* [38]. This signal was also detected with random panels of 100K markers and more, with the *ovine* 50K and HD BeadChips and with the exome capture (Figure 4). However, the intensity of the signal decreased progressively with the density of markers. Therewith, another non-previously reported signal of selection was detected only with the WGS dataset and with random panels

of 5M and 1M variants. This signal was located in the region of the Neurobeachin *NBEA* and Mab21-like 1 *MAB21L1* genes on chromosome 10 (positions 26,007,917-26,592,574 and 26,231,353-26,232,432 on OAR v3.1, respectively).

Assessment of standard surrogates of whole genome data

To assess the ability of a panel of variants defined on a specific set of individuals to act as an accurate standard surrogate of whole genome data on worldwide populations, we genotyped wilds and Iranian and industrial domestics using 3 panels of 10 K variants set up with Moroccan individuals. Among the 10k variants, on average 4959 and 4679 were actually polymorphic in mouflon and bezoars respectively, 6125 and 5826 in the Moroccan sheep (n=20) and goats (n=20) respectively, 5609 and 5688 in the Iranian sheep and goats and 5851 in the industrial sheep. Whatever the group, the three panels of 10K random variants gave estimates of nucleotide diversity (π), observed heterozygosity (*Ho*), inbreeding coefficient (*F*) and genetic differentiation (*Fst*) similar to that obtained from the WGS data. On the contrary, as in previous analyses, the BeadChips showed considerable ascertainment biases by overestimating *Fst* between domestic groups (52% for the 50K BeadChip and 49% for the HD BeadChip in sheep and 90 for the 50K BeadChip in goats) and the diversity in all groups (Figures 5, 6, S9-S13). This ascertainment bias did not affect the estimation of the inbreeding coefficient (F). Whatever the bias, the ranking of individual *Ho* and *F* were not affected by the panel of variants used, but for π, the wilds even appeared less diverse than the domestics while WGS data showed the opposite (Figure 5).

Difference between random panels and BeadChips

One major difference between setting-up of the random panels of variants and the BeadChips relies on the distribution of variants across the genome. Figure 7 illustrates this in showing the distributions of the physical distances between adjacent variants in various panels for Moroccan sheep and goats. The random 50K variants as well as the random 500K variants and the HD ovine BeadChip showed a similar L-shaped curve indicating that variants were evenly distributed across the genomes. On the other hand, the caprine 50K BeadChip displayed an almost complete lack of SNPs separated by less than around 30Kb, while for the ovine 50K BeadChip the lack of SNPs in these categories is less drastic, at most around a half of the expected distribution for the shorter distances. The exome capture simulation displayed a very high occurrence of distances lower than 200 bp and a quasi absence of distance larger then 10kb (Figures 7, S14).

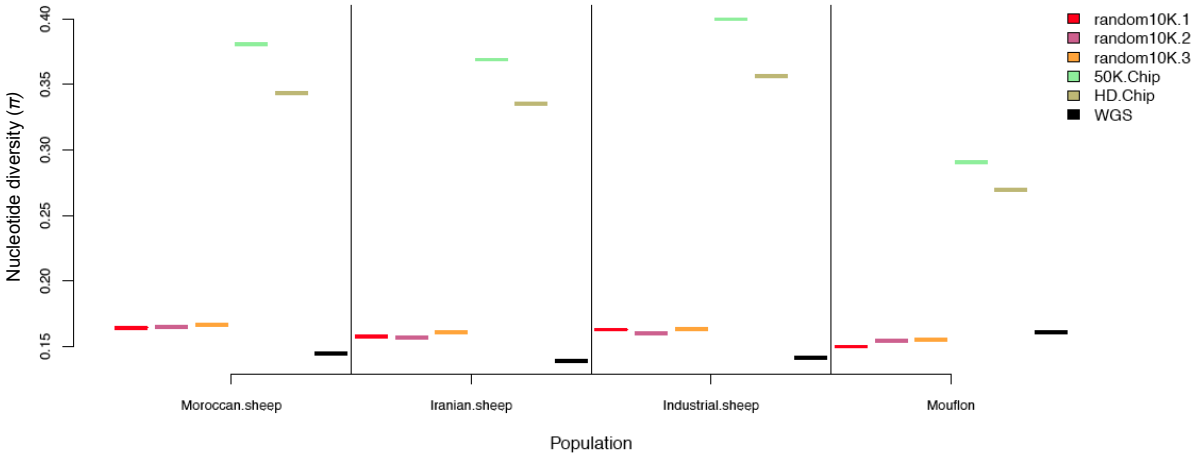

Figure 5. Nucleotide diversity (π) estimated in four *Ovis* groups using the surrogate 10K panels of variants.

Plot of Nucleotide diversity (π) estimated with 3 independent sets of 10K variants defined in Moroccan sheep (10K), and with Illumina® ovine 50K SNP Beadchip (50K.Chip), Illumina® ovine HD Beadchip (HD.Chip), and variants extracted from whole genome sequences (WGS).

Reliability of low-coverage re-sequencing

1X, 2X and 5X whole genome sequencing coverage were simulated by randomly sampling reads in the 12X WGS data, and used for calling genotypes in 30 goats and 30 sheep. The 12X WGS allowed genotyping at 31,775,474 variant sites (31,735,229 at which more than 95% of individuals had genotypes called) for goats and 43,478,084 for sheep (43,105,056 at which more than 95% of individuals had genotypes called), and decreasing the coverage strongly reduced the number of variants that could be genotyped (missing genotypes, Table 2), while the number of variants wrongly genotyped remained rather low (mis-matching genotypes, Table 2). Heterozygous genotypes were more affected than homozygous ones. Moreover, the decreasing coverage resulted in an increasing underestimation of *Ho* (around 1.2, 3 and 6 times for 5X, 2X and 1X, respectively), and in a decreasing preservation of the relative ranking of *Ho* values among individuals (Table 2). This ranking was better preserved in sheep than in goats.

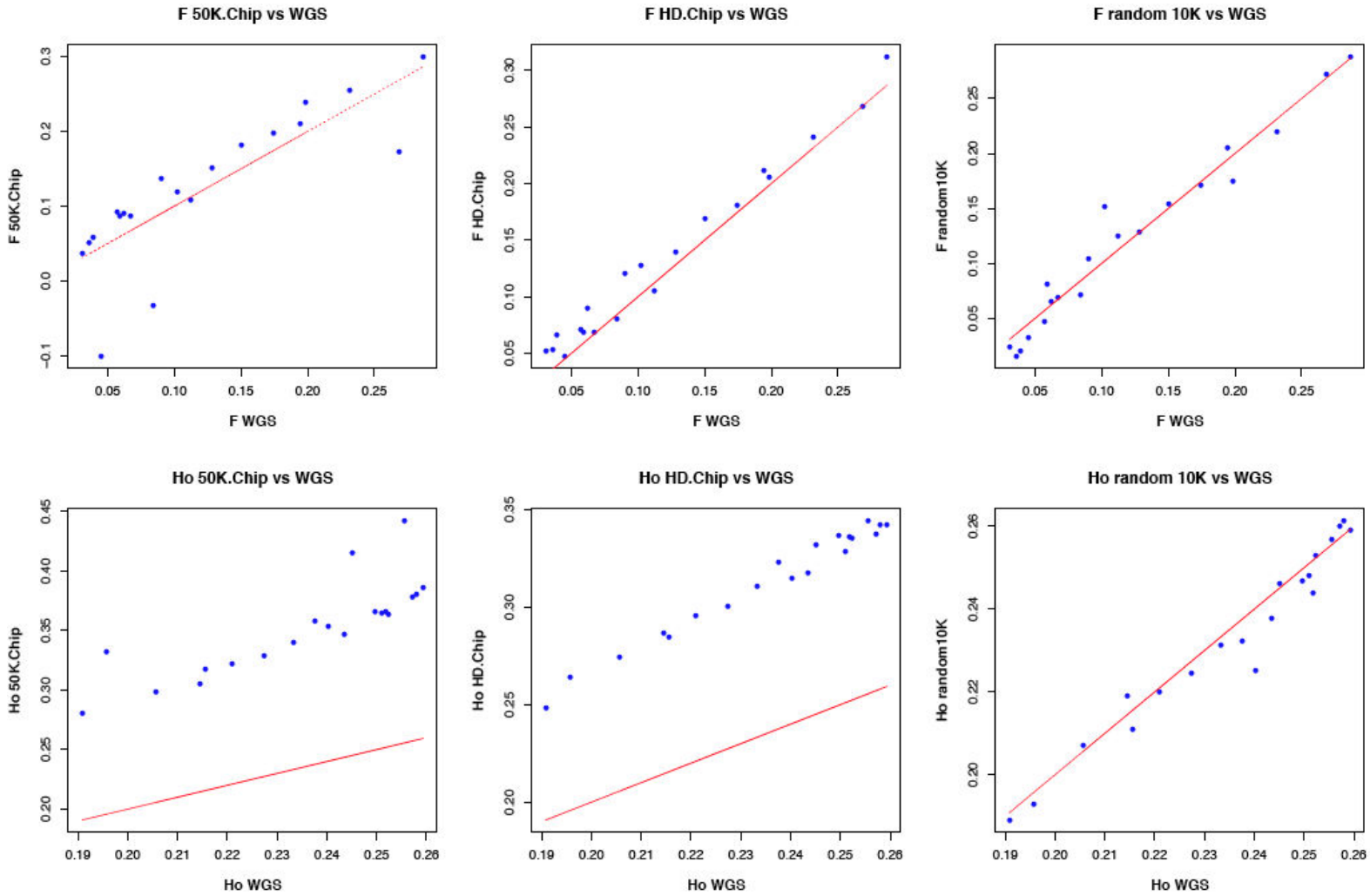

Figure 6. Estimates of individual inbreeding coefficient (*F*) and observed heterozygosity (*Ho*) from different panels of variants compared to WGS data estimates in industrial sheep breeds.

Plot of individual inbreeding coefficient (*F*; top) and observed Heterozygosity (*Ho*; bottom) estimated with variants extracted from whole genome sequences (WGS) versus inferences with Illumina® ovine 50K SNP Beadchip (50K.Chip), Illumina® ovine HD Beadchip (HD.Chip), and 1 set of 10K variants defined in Moroccan sheep (random 10K). The red lines represent the relationship for which the estimates of the different panels are identical to the ones of WGS inferences.

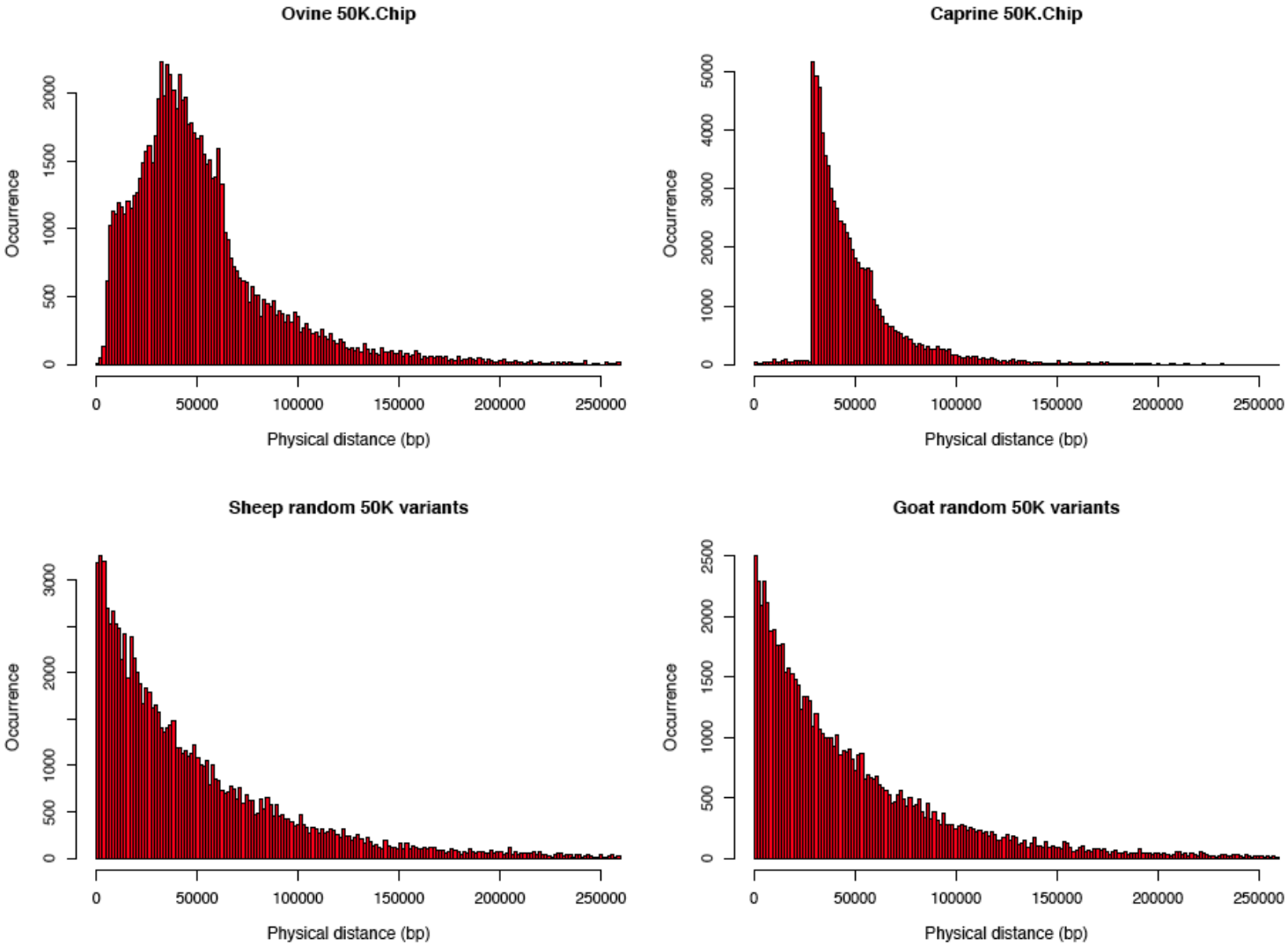

Figure 7. Distribution of physical distances between adjacent variants in 50K BeadChips and random panels of 50K variants.

Ovine 50K.Chip: Illumina® ovine 50K SNP Beadchip; Caprine 50K.Chip: Illumina® caprine 50K SNP Beadchip; Sheep random 50K variants: random panel of 50K variants defined in Moroccan sheep; Goat random 50K variants: random panel of 50K variants defined in Moroccan goats.

Table 2. Concordance between low-coverage re-sequencing and 12X coverage for homozygous and heterozygous genotypes.

Coverage		**1x**		**2x**		**5x**	
Species		**Sheep**	**Goats**	**Sheep**	**Goats**	**Sheep**	**Goats**
Genotypes for > 95% of individuals	Nb. of sites	12,603,362	10,701,885	18,615,123	14,906,837	37,473,708	27,472,390
	Nb. of polymorphic variants	268,491	254,313	4,327,012	3,324,740	24,074,435	17,108,812
Genotypes for 100% of individuals	Nb. of sites	12,550,038	10,662,633	15,617,291	12,930,734	28,419,192	20,974,409
	Nb. of polymorphic variants	259,177	249,056	1,783,255	1,737,328	15,847,527	11,296,131
Heterozygote genotypes	Matching 12x (%)	0.12 ± 0.02	0.19 ± 0.03	6.20 ± 0.79	5.68 ± 0.72	74.1 ± 8.6	71.5 ± 7.1
	Mis-matching 12x (%)	3.14 ± 0.33	4.38 ± 0.65	2.97 ± 0.30	3.66 ± 0.53	2.27 ± 0.24	2.16 ± 0.23
	Missing (%)	96.7 ± 11.1	95.4 ± 9.4	90.8 ± 10.3	90.7 ± 9.0	23.6 ± 2.6	26.4 ± 2.6
Homozygote genotypes	Matching 12x (%)	34.1 ± 0.1	38.1 ± 0.1	49.2 ± 0.2	52.3 ± 0.2	87.8 ± 1.7	87.5 ± 1.1
	Mis-matching 12x (%)	0.02 ± 0.02	0.02 ± 0.02	0.03 ± 0.01	0.02 ± 0.01	0.09 ± 0.10	0.07 ± 0.02
	Missing (%)	65.9 ± 2.1	61.9 ± 1.4	50.8 ± 1.9	47.7 ± 1.4	12.2 ± 0.4	12.5 ± 0.4
Correlations of *Ho* with 12X estimates	r (Pearson)	0.642	0.173	0.989	0.507	0.999	0.989
	Slope (Pearson)	0.149	0.176	0.329	0.281	0.864	0.818
	r (Spearman)	0.586	0.203	0.802	0.522	0.942	0.900

Number of sites and polymorphic variants were defined using two different percentages of genotyped individuals thresholds: > 95% and 100%. Other estimates were inferred from 95% filtering. *Ho* correlations were estimated according to Pearson and Spearman to compare rankings of individuals. Slopes were estimated by forcing the intercept of the linear regression to be 0.

Discussion

The recent development of sequencing technologies makes now possible the whole sequencing of individual genomes, which may greatly extend the information inferred through population genetics approaches [39,40]. However, re-sequencing large numbers of individuals is still not affordable in most of the studies and WGS analyses remain time-consuming and require high performance computing. In this context, we assessed the efficiency and drawbacks of different genome-wide sampling strategies to give accurate characterization of genomes diversity.

Effect of sequencing coverage on the assessment of whole genome variations

We identified many high confidence variants in each species using WGS 12X coverage re-sequencing data (from 17.5M in bezoars to 43.5M in Moroccan sheep), corresponding mostly to SNPs but also to small indels, which represented 6% to 10% of the variants. Overall, the approach used to call and filter variants was efficient according to the high concordance between 12X re-sequencing data and the 50K SNP BeadChips genotyping. The lowest concordance was obtained for bezoars, and would result from a higher number of indels that correspond to SNPs in the caprine 50K BeadChip. Besides, the higher number of variants discovered in the domestics compared to the wild animals could partially result from the high number of domestics used for the variant discovery. The slightly higher percentage of mapped reads in *Ovis* compared to *Capra* (99.4% vs 98.9%) might result from the higher quality of the sheep genome assembly and would explain at least partially the lower number of variants called in *Capra* species.

The simulation of 1X, 2X and 5x WGS datasets from the 12X WGS confirmed the sensitivity of population genetics inferences to the sequencing coverage previously found (e.g. [17,23]), and helped to depict the effect of reducing the coverage. As might be expected, we found that homozygote genotypes were more correctly called than heterozygote ones whatever the coverage. This is due to the fact that more reads should be mapped at a position for calling the two alleles of an heterozygote than for calling the unique allele of an homozygote. Additionally, the filtering process for variant calling induced a higher percentage of missing data for heterozygotes because it discarded any heterozygous genotype for which one allele was under or over-represented.

Thus, the decrease in WGS coverage first resulted in a decrease in variant density (increasing proportion of missing data). The density of reliable variants obtained when decreasing the

coverage (> 250k for 1X and > 3M from 2X, see Table 2) would still have been sufficient to allow accurate estimation of population genetics parameters and detection of selection signatures (see below 'effect of variant density'). However, the trend is combined to a bias that strongly affected the estimations. This bias concerned both missing and erroneous genotyping, which affected mostly heterozygotes (even more when the coverage decreases) where the erroneous genotyping mostly produces homozygotes. This resulted in an underestimation of heterozygosity (*Ho*). However, the values obtained for the 5X coverage appeared to be just as accurate as those inferred from the 12X WGS (highly correlated values of *Ho*, and thus of *F*), for both studied species. This result is coherent with the findings of [20] who showed that in association studies, genotyping 3,000 individuals at 4X depth provided similar power to 30X sequencing of about 2,000 individuals.

Effect of the density of variants

Population genetics parameters were estimated for different sample sizes. The objective was not to characterize in itself a sample size effect that we obviously expected, but to assess the effect of the density of variants under various conditions. It should be noted that we generally observed a sample size effect on the estimation of summary statistics, whatever the panel of variants, and that this effect was especially strong when measuring population differentiation, even greater than the effect of variant density. For any of the chosen sample sizes, the density of variants was determinant to get a representative view of genome variations.

Many population genetics studies that infer demographic processes still rely on just a few dozens to a few hundreds of genetic markers aiming to be representative of the whole genome variations [41-44]. We could, in fact, get a representative view of the whole genome variations by using a relatively small set of variants, provided they are randomly distributed across the genome. Low-density random panels of variants (i.e. 5K or 10K corresponding to 1 variant every ~300 or ~600Kb) gave estimates of summary statistics similar to those calculated from 12X WGS data. The assessment of population structure through calculation of coefficients of ancestry was reliable whatever the panel density, while the estimations of *Fst* required at least 100K variants. Furthermore, the estimation of linkage disequilibrium and the detection of signatures of selection required higher variant densities: around one variant every 3 to 6Kb, which gave similar estimates to 12X WGS data with roughly one variant every 100 to 200bp.

The adequate densities of variants required for a reliable description of genomic variations depend on the pattern of *LD* decay across the genome. In the four studied species, those patterns were globally the same, with r^2 dropping bellow 0.15 within at most 5Kb. Consequently, we needed 500k to 1M variants to accurately estimate *LD* decay. All panels of fewer than 100K variants (~1 variant per 30kb) produced incorrect estimations of r^2 for small distances (until 50Kb depending on the panel). We could expect that the same orders of magnitude of variant densities would be required in species characterized by similar patterns of *LD* decay (e.g. *Anopheles arabiensis* with r^2 dropping bellow 0.2 within 200 bp [45]). However, genomic patterns of *LD* decay depend on the demographic histories of populations, and reflect the changes in effective population sizes. It is likely that populations with smaller effective population sizes, which have experienced for example strong bottlenecks such as industrial breeds, could require smaller variant densities.

Selective sweeps, when they occur, increase LD in regions of several Kb surrounding the selected allele. This signature is then reduced by recombination, and the older the selective sweep the smaller will be the region still influenced around the selected allele [46,47]. In the case of the selective sweep that has occurred in the *RXFP2* gene, the signal is still extending ~350Kb and required at least a random panel of 100K variants in order to be detected. Therefore, higher density random panels would be needed to detect any weaker selective sweep (i.e. associated to lower *LD*).

Ascertainment bias in non-random panels

The estimation of almost all population genetics parameters was biased when using variants from commercial SNP BeadChips or exome. Measurements both of genome diversity and of population differentiation were affected. SNPs included in the design of the commercial panels were chosen according to their high level of polymorphism in several breeds (mainly European industrials [41]), and this ascertainment bias lead to an overestimation of the genomic diversity. The *ovine* HD BeadChip suffered less from this bias compared to the 50K *ovine* BeadChip due to the inclusion of high, medium and low frequency SNPs [48]

The exome capture data, while representing highly conserved regions, logically underestimated genetic diversity. If scientists are widely aware of the possible consequences of such ascertainment biases, as already reported for microsatellites markers [49-52], only one study has addressed this question for SNP chips until now to our knowledge [34].

The biased estimation of genetic diversity and genetic differentiation would be less problematic as long as the ranking of estimated values is preserved (e.g., the most variable individuals are actually those with the highest measured diversity). For example, when estimating animal genetic resources, this will allow finding the more diverse populations/breeds. However, we showed that this ranking was inverted when comparing the diversity of wilds and domestics with the ovine and caprine SNP Beadchips, which should be used with caution when comparing well-differentiated populations. We showed that random panels of 10K variants, even designed from a unique population, were more efficient in describing genomic variations by producing unbiased estimates. Thus, our WGS data, as with other datasets available in public databases, can be used to set up panels of variants representing accurate surrogates for WGS data.

Besides the biases they induce, non-random panels of variants such as SNP chips also impact the estimation of genomic diversity according to the density of variants and their distribution across the genome.

Distribution of variants across the genome

Besides the effects of variant density and ascertainment bias, the distribution of variants across the genome also impacts the reliability of the characterization of the genome. For similar numbers of variants, the ovine and caprine50K BeadChips were less accurate than random panels for estimating the *LD* decay over short distances. This is not surprising given the underrepresentation of close adjacent SNPs (<6Kb in ovine and <30kb in caprine BeadChips, Figure 7) in these Beadchips. Moreover, the local density of Beadchip SNPs varied across the genome with some regions being far well covered than others. This explains why, like [38], we were able to detect the signal of selection associated to the *RXFP2* gene with the *ovine* 50K SNP BeadChip but not with 50K variants random panels. The commercial BeadChip has four SNPs in a 148 Kb window centred on the *RXFP2* gene, which appeared to be enough for detecting selection, while the random 50K panel used had no variant in that window.

The distribution of variants across the genome obviously determines the ability to detect selection signatures**,** and high-density variant panels are required to detect selected regions. One needs variants from regions under selection to find the associated signature, which is not necessarily assumed by low and medium-density panels of variants. This is more limiting

when studying populations characterised by low overall linkage disequilibrium, which is generally the case of indigenous domestic and wild populations.

Conclusion

The accuracy of panels of variants to describe genome variations depends on the distribution of these variants across the genome, according to the level of LD and its proper variability. While high to medium coverage genome sequencing produces reliable genotyping, it remains costly both in terms of money and in data management, and thus surrogates of WGS data are still needed.

For model species, commercial standardized panels are generally already available and one should know their potential biases and use them cautiously. This is particularly true if the studied populations or breeds are genetically divergent from the individuals used for designing the set of variants. Our results showed that a few thousands of markers randomly chosen across the genome provide unbiased information. Therefore, it could be valuable to include such sets of variants when designing new high density SNP chips. In non-model species, the genotyping of individuals by SNP chips could be replaced by genotyping by sequencing approaches (RAD-seq), which would theoretically approximate a random distribution of markers across the genome, and could thus provide convenient surrogates for WGS data. As shown by our results, a suitable variant density should be targeted according to the aim of the study and the resources allocated. Finally, when considering Whole Genome Sequencing approaches, low-coverage (1X and 2X) sequencing is not appropriate for setting up population genomics studies due to the important underestimation of heterozygote genotypes. A medium coverage of 5X could provide summary statistics with a reasonable underestimation.

Material & methods

Sampled individuals

Tissue samples were collected for 48 sheep (*Ovis aries*) and 30 goats (*Capra hircus*) widely spread across the Northern half of Morocco (North of latitude 28°) between January 2008 and March 2012 (Table S1). In North-western Iran, 20 sheep and 20 goats were collected between August 2011 and July 2012. Tissues from the distal part of the ear were collected and placed

in alcohol for one day, before transfer into silica-gel tubes until DNA extraction. Tissues from 15 Asiatic mouflon (*Ovis orientalis*) and 20 bezoars (*Capra aegagrus*) were collected in Iran, either from captive or recently hunted animals and conserved in silica-gel after one day in alcohol, or from frozen corpses or tissues archived in alcohol by the Iranian local Department of Environment and transferred in silica-gel until extraction. Additionally, the International Sheep Genome Consortium (http://www.sheephapmap.org/) provided whole genome re-sequencing data at 12X coverage for 20 sheep representing a worldwide panel of 20 industrial breeds (Table S1).

DNA extraction and re-sequencing

DNA extraction was done at *Parco Tecnologico Padano* (Lodi, Italy) using the Puregene Tissue Kit from Qiagen® following the manufacturer's instructions. Then, 500ng of genomic DNA were sheared to a 150-700 bp range using the Covaris® E210 instrument. Sheared DNA was used for Illumina® library preparation by a semi-automatized protocol. Briefly, end repair, A tailing and Illumina® compatible adaptors (BiooScientific) ligation were performed using the SPRIWorks Library Preparation System and SPRI TE instrument (Beckmann Coulter), according to the manufacturer protocol. A 300-600 bp size selection was applied in order to recover most of the fragments. DNA fragments were amplified by 12 cycles PCR using Platinum Pfx Taq Polymerase Kit (Life® Technologies) and Illumina® adapter-specific primers. Libraries were purified with 0.8x AMPure XP beads (Beckmann Coulter). After library profile analysis by Agilent 2100 Bioanalyzer (Agilent® Technologies) and qPCR quantification, the libraries were sequenced using 100 bp length read chemistry in paired-end flow cell on the Illumina® HiSeq2000.

Read mapping, SNP calling and filtering

Illumina paired-end reads of *Ovis* were mapped on the sheep reference genome (OAR v3.1, GenBank assembly GCA_000317765.1 [53]) and those of *Capra* on the goat reference genome (CHIR v1.0, GenBank assembly GCA_000317765.1 [14]) using BWA mem [54]. 99.4% (± 0.1%), 99.3% (± 0.2%), 98.9% (± 0.1%) and 98.8% (± 0.4%) of the reads were mapped on the reference assembly for sheep, mouflon, goats and bezoar, respectively. The BAM files produced were then sorted using Picard SortSam and improved using Picard MarkDuplicates (http://picard.sourceforge.net), GATK RealignerTargetCreator, GATK IndelRealigner [55] and Samtools calmd [56].

Variant sites were initially called using three different algorithms: Samtools mpileup [56], GATK UnifiedGenotyper [57] and Freebayes [58]. Variants were called for each group independently: Moroccan sheep, Iranian sheep, Iranian mouflon, Moroccan goat, Iranian goat and Iranian bezoar. Note that a larger dataset than that used in this study was used for variant discovery in Moroccan groups (160 sheep and 161 goats; for European Nucleotide Archive ID, see Table 3). Then we ran two successive rounds of filtering variant sites. Filtering stage 1 merged together calls from the three algorithms, whilst filtering out the lowest-confidence calls. A variant site passed if it was called by at least two different calling algorithms with variant phred-scaled quality > 30. An alternate allele at a site passed if it was called by any one of the calling algorithms, and the genotype count > 0. Filtering stage 2 used Variant Quality Score Recalibration by GATK. First, we generated a training set of the highest-confidence variant sites where (i) the site is called by all three variant callers with variant phred-scaled quality > 100; (ii) the site is biallelic; (iii) the minor allele count is at least 3, counting only samples with genotype phred-scaled quality > 30. The training set was used to build a Gaussian model using the tool GATK VariantRecalibrator using the following variant annotations from UnifiedGenotyper: QD, HaplotypeScore, MQRankSum, ReadPosRankSum, FS, DP, InbreedingCoefficient. The Gaussian model was applied to the full data set, generating a VQSLOD (log odds ratio of being a true variant). Sites were filtered out if VQSLOD < cutoff value. The cutoff value was set for each population by the following: Minimum VQSLOD = {the median value of VQSLOD for training set variants} - 3 * {the median absolute deviation VQSLOD of training set variants}. Measures of the transition / transversion ratio of SNPs suggest that this chosen cut-off criterion gives the best balance between selectivity and sensitivity.

Genotypes were improved and phased by Beagle 4 [59], and then filtered out where the genotype probability calculated by Beagle is less than 0.95. The genotype call sets generated at this stage constituted the WGS datasets used for within-population analyses. For cross-populations comparisons and validation of the identified WGS surrogates that were performed in each genus (i.e. *Capra* and *Ovis*), we generated a set of filtered variant sites per genus by merging the positions of filtered bi-allelic SNPs called in the different groups. For each sample, genotypes were called at each SNP position using GATK UnifiedGenotyper using the option GENOTYPE_GIVEN_ALLELES. Genotypes were improved and phased by Beagle 4 [59], and then filtered out where the genotype probability calculated by Beagle is less than 0.95.

Quality control of WGS data

To further assess the quality of the filtered WGS datasets, a subset of the sequenced individuals were genotyped using commercial SNP Chips by *Laboratorio Genetica e Servizi* (Cremona, Italy). 29 Moroccan sheep and 8 Asiatic mouflon were genotyped with the Illumina® ovine 50K SNPs BeadChip, and 27 Moroccan goats and 8 bezoars with the Illumina® caprine 50K SNPs BeadChip. In order to establish the concordance between WGS and chip data the coordinates of the SNPs on the chips were obtained by mapping the probes used for chip design onto the corresponding reference genome (OAR v3.1 or CHIR v1.0) using BWA aln and sample [54]. The raw data in Plink format [60] were updated for SNP coordinates and were filtered for each group by applying the following inclusion criteria: SNPs in a known chromosome (from our mapping); minor allele frequency (MAF) > 0.02, genotype call rate (SNPs) > 0.95, genotype call rate (Animals) > 0.95 and identity-by-state (Animals) < 0.95. The filtered datasets were converted to harmonize the reference alleles with the reference genomes using a script based on the programs PlinkSeq v 0.08 (http://atgu.mgh.harvard.edu/plinkseq/index.shtml) and Plink v 1.07 [60] which was necessary for the quality control of the re-sequencing data. After removing the positions corresponding to short indels and tri-allelic variants, which are incorrectly genotyped by the BeadChips, the number of SNPs both genotyped with the Chip and by whole genome sequencing was 47,122 for sheep 49,467 for goats, 37,779 for Asiatic mouflon and 41,751 SNPs for bezoars. The comparison of the *ovine* and *caprine* 50K BeadChips genotyping data with the WGS data was performed. The average (± s.d.) genotype concordance between the *ovine*/*caprine* 50K BeadChips and the WGS was 99.9% (± 0.1%) in sheep, 99.7% (± 0.0%) in goats, 99.7% (± 0.1%) in Asiatic mouflon and 98.5% (± 0.3%) in bezoars.

Setting up datasets of variants

From the 173 individuals, we defined different groups depending on the question addressed. First, to evaluate the impact of sampling panels of variants and reducing the WGS coverage on the estimation of genetic parameters we designed four groups corresponding to 30 Moroccan sheep, 30 goats, 14 Asiatic mouflon and 18 bezoars. In order to assess the effect of individual sample size, each analysis was performed for the whole groups and for two random subsets corresponding approximately to one third and two thirds of the total (i.e. respectively 10 and 20 Moroccan sheep and goats, 5 and 10 Asiatic mouflon and 8 and 13 bezoars). Second, for detecting a signal of selection associated to the *RXFP2* locus [38] and related to the presence/absence of horns, we had to consider additional Moroccan sheep to constitute 2

contrasted groups of 15 horned and 15 polled individuals (Figure S15; Table S1). Third, in order to assess the ability of random panels of variants defined in Moroccan sheep and goat groups to accurately represent the genome variations in other groups of the same genus, we used 4 groups of *Ovis* (20 Moroccan, 20 Iranian and 20 industrial sheep, 15 Asiatic mouflon), and 3 groups of *Capra* (20 Moroccan and 20 Iranian goats and 20 bezoars).

For each group of individuals, a 12X WGS dataset was composed of all the SNPs called (see 'Read mapping, SNP calling and filtering' section) and used for within population analyses. Note that for cross-populations comparisons we only kept the variants found polymorphic in both groups considered, in order to prevent from any possible biased calling or filtering error. Then, variant panels were extracted from the 12X WGS datasets. Random panels were extracted using GATK SelectVariants [57] consisting in 5 independent replicates for each of the 8 following numbers of variants: 1K, 5K, 10K, 50K, 100K, 500K, 1M and 5M. We also created non-random panels simulating the data obtained with commercial BeadChips or through exome capture. BeadChip data were obtained by calling variants at the Illumina® 50K Ovine or Caprine BeadChip SNP coordinates. We successfully extracted 42,117 variants for Moroccan sheep, 47,245 variants for Moroccan goats, 26,141 variants for Asiatic mouflon and 33,951 variants for bezoars. The combined datasets used for cross-population analyses included 30,870 variants for *Ovis* and 38,641 variants for *Capra*. In sheep, the High Density BeadChip genotyping was simulated by calling WGS variants at the coordinates of the Illumina® ovine HD BeadChip. This gave 601,456 variants for Moroccan sheep and of 444,169 variants for Asiatic mouflon. The combined dataset had 419,041 variants.

We simulated an exome capture only for *Ovis* because of the annotation of the goat genome was insufficiently advanced. The exome annotation was obtained from the sheep genome annotation that was available in ENSEMBL database by the time of analysis (25th September 2013) (ftp://ftp.ensembl.org/pub/pre/) and corresponded to 224,871 exons in 45,972 genes. The number of variants from these regions extracted from 12X WGS was 278,568 for Moroccan sheep and 155,236 for Asiatic mouflon. The 93,409 variants polymorphic in both groups constituted the combined dataset. Thus, for the identification of potential surrogates for the WGS, the genotypes produced for the different variant panels and the different groups of individuals constituted a total of 946 datasets of which 516 were used for estimating within-group genetic diversity and 430 for cross-populations comparisons.

The second round of the analysis consisted on the assessment of the ability of panels of random 10K variants defined in one population (i.e., Moroccan domestics) to characterize

genome diversity in other populations. We selected 3 panels of 10K variants from the five replicates previously set-up. For these 10K variants, we extracted from the 12X WGS data the genotypes of 15-20 individuals per group analysed (i.e., Moroccan and Iranian populations, industrial sheep and wilds).

Simulating low-coverage re-sequencing data

The 12X WGS data were subsampled to simulate the output of a sequencing experiment with fewer reads were generated. For each of the 30 Moroccan sheep and the 30 Moroccan goats groups, three sub-sampled WGS datasets were generated comprising (i) 15 million, (ii) 30 million, (iii) 75 million paired reads, corresponding approximately to a 1X, 2X, and 5X sequencing coverage of the genome, respectively. Paired reads were randomly chosen from the full sequencing data using Picard Downsample, in such a way that all reads had an equal probability of being chosen, including duplicate or unaligned ones. Next, Picard MarkDuplicates was used to tag reads that appeared as duplicates in the sub-sampled datasets.

For each variant of the list generated for the 12X WGS, genotypes were called using GATK UnifiedGenotyper with the option GENOTYPE_GIVEN_ALLELES. Genotypes were improved and phased with Beagle4, and filtered at the individual level when the genotype probability was less than 0.95. Variants were kept when the genotype was called for more than 95% of the individuals. For each of the simulated coverages, the genotype at each variant position was compared to that obtained for the 12X coverage and classified as matching, un-matching or missing, for homozygotes and heterozygotes separately. Additionally, the individual observed heterozygosity was inferred for each coverage and used to estimate (i) Pearson correlation, (ii) Spearman correlation with 12X inferences. Slope values (*b)* were estimated for each depth by setting the intercept to 0.

Population genetics analyses

- Genetic diversity within groups

Using Vcftools [61] we estimated the observed heterozygosity (*Ho*) and inbreeding coefficient (*F*) with polymorphic diploid bi-allelic SNPs, and the nucleotide diversity (π) by taking the averaged nucleotide diversity over all fully diploid variants. Pairwise SNPs linkage disequilibrium (r^2) was also estimated with Vcftools on all bi-allelic variants (SNPs and indels) for 5 segments of 2Mbp selected on 5 chromosomes (physical positions between 1

Mbp and 3 Mbp on chromosomes 5, 10, 15, 20 and 25). The extent of the linkage disequilibrium was assessed by the physical distance corresponding to $r^2 = 0.15$ ($r^2_{0.15}$).

- Genetic differentiation and structure

The genetic structure and differentiation was measured between Moroccan domestics and Iranian wilds, as representative of population having diverged about 10,000 years ago (i.e., at the time of domestication). The averaged Fst [35] was estimated for bi-allelic variants with Vcftools. Additionally, genetic structure was investigated through the Bayesian clustering approach sNMF [36] using bi-allelic variants. This method was specifically developed to estimate individual admixture coefficients on large genomic datasets.

- Detection of a selection signature

We targeted the genomic region surrounding the *Relaxin/insulin-like family peptide receptor 2* gene (*RXFP2*; Chr 10: 29,454,677 – 29,502,617bp), which already showed a signature of selection related to polledness in sheep [38,62]. We extracted variants between positions 20 Mb and 40 Mb on chromosome 10 for 15 horned and 15 polled sheep and searched for selected sweeps in this region using XP-CLR [37] to infer genetic distances, we estimated a constant recombination rate for this region based on the random 1M variants dataset, using the PAIRWISE program of LDhat v2.2 [63] with recommended parameters. XP-CLR scores were calculated for each grid point placed along the segment considered with a spacing of 5Kb. A maximum of 300 bi-allelic variants was considered in a sliding window of 0.5cM around the grid point and we down-weighted contributions of highly correlated SNPs (r^2>0.99).

Accession numbers

The variant call sets are archived in the European Nucleotide Archive with accession numbers provided in Table 3. The accession of the sample in the Biosamples archive, and of the corresponding aligned bam file in the ENA archive are listed in Table S1.

Acknowledgments

This work was funded by the UE FP7 project *NEXTGEN* 'Next generation methods to preserve farm animal biodiversity by optimizing present and future breeding options'; grant

agreement no. *244356*. We would like to thank the International Sheep Genome Consortium for providing re-sequencing data for 20 industrial sheep breeds. We thank Eric Coissac who helped in setting-up the overall approach and Bertrand Servin for the useful discussions. We are grateful to R. Hadria, M. Laghmir, L. Haounou, E. Hafiani, E. Sekkour, M. ElOuatiq, A Dadouch, A. Lberji, C. Errouidi and M. Bouali for helping in sampling in Morocco.

Table 3. References of the files containing the WGS variant data and the Chip genotyping data.

File	ENA accession	Species	Taxonomy ID	Country	Description
MOOA.population_sites.OARv3_1.20140328.vcf.gz	ERZ019290	Ovis aries	9940	Morocco	WGS filtered population variant calls
MOOA.genus_snps.OARv3_1.20140328.vcf.gz	ERZ019292	Ovis aries	9940	Morocco	WGS genotypes called at all known ovis SNPs
MOOA.ovineSNP50.OARv3_1.20140307.vcf.gz	{pending}	Ovis aries	9940	Morocco	Ovine SNP50 bead chip genotypes
IROO.population_sites.OARv3_1.20140307.vcf.gz	ERZ020211	Ovis orientalis	469796	Iran	WGS filtered population variant calls
IROO.genus_snps.OARv3_1.20140528.vcf.gz	{pending}	Ovis orientalis	469796	Iran	WGS genotypes called at all known ovis SNPs
IROO.ovineSNP50.OARv3_1.20140307.vcf.gz	{pending}	Ovis orientalis	469796	Iran	Ovine SNP50 bead chip genotypes
IROA.genus_snps.OARv3_1.20140307.vcf.gz	ERZ019270	Ovis aries	9940	Iran	WGS genotypes called at all known ovis SNPs
IROA.ovineSNP50.OARv3_1.20140307.vcf.gz	{pending}	Ovis aries	9940	Iran	Ovine SNP50 bead chip genotypes
ISGC.genus_snps.OARv3_1.20140131.vcf.gz	{pending}	Ovis aries	9940	Various	WGS genotypes called at all known ovis SNPs
MOCH.population_sites.CHIR1_0.20140307.map.gz	ERZ020631	Capra hircus	9925	Morocco	WGS filtered population variant calls
MOCH.genus_snps.CHIR1_0.20140307.vcf.gz	ERZ019391	Capra hircus	9925	Morocco	WGS genotypes called at all known capra SNPs
MOCH.goatSNP50.CHIR1_0.20140307.vcf.gz	{pending}	Capra hircus	9925	Morocco	Goat SNP50 bead chip genotypes
IRCA.population_sites.CHIR1_0.20140307.vcf.gz	ERZ020627	Capra aegagrus	9923	Iran	WGS filtered population variant calls
IRCA.genus_snps.CHIR1_0.20140409.vcf.gz	{pending}	Capra aegagrus	9923	Iran	WGS genotypes called at all known capra SNPs
IRCA.goatSNP50.CHIR1_0.20140307.vcf.gz	{pending}	Capra aegagrus	9923	Iran	Goat SNP50 bead chip genotypes
IRCH.genus_snps.CHIR1_0.20140307.vcf.gz	ERZ019386	Capra hircus	9925	Iran	WGS genotypes called at all known capra SNPs
IRCH.goatSNP50.CHIR1_0.20140307.vcf.gz	{pending}	Capra hircus	9925	Iran	Goat SNP50 bead chip genotypes

References

1. Wright S (1931) Evolution in Mendelian populations. Genetics 16: 0097-0159.

2. Fisher R (1958) POLYMORPHISM AND NATURAL-SELECTION. Bulletin of the International Statistical Institute 36: 284-289.

3. Di Rienzo A, Donnelly P, Toomajian C, Sisk B, Hill A, et al. (1998) Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. Genetics 148: 1269-1284.

4. Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: A study of Y chromosome microsatellites. Molecular Biology and Evolution 16: 1791-1798.

5. Holloway JW, Beghe B, Turner S, Hinks LJ, Day INM, et al. (1999) Comparison of three methods for single nucleotide polymorphism typing for DNA bank studies: Sequence-specific oligonucleotide probe hybridisation, TaqMan liquid phase hybridisation, and microplate array diagonal gel electrophoresis (MADGE). Human Mutation 14: 340-347.

6. Jones MR, Forester BR, Teufel AI, Adams RV, Anstett DN, et al. (2013) INTEGRATING LANDSCAPE GENOMICS AND SPATIALLY EXPLICIT APPROACHES TO DETECT LOCI UNDER SELECTION IN CLINAL POPULATIONS. Evolution 67: 3455-3468.

7. Black WC, Baer CF, Antolin MF, DuTeau NM (2001) Population genomics: Genome-wide sampling of insect populations. Annual Review of Entomology 46: 441-469.

8. Goldstein DB, Weale ME (2001) Population genomics: Linkage disequilibrium holds the key. Current Biology 11: R576-R579.

9. Jorde LB, Watkins WS, Bamshad MJ (2001) Population genomics: a bridge from evolutionary history to genetic medicine. Human Molecular Genetics 10: 2199-2207.

10. Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: From genotyping to genome typing. Nature Reviews Genetics 4: 981-994.

11. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, et al. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 12: 499-510.

12. Everett MV, Grau ED, Seeb JE (2011) Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. Molecular Ecology Resources 11: 93-108.

13. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, et al. (2013) The zebrafish reference genome sequence and its relationship to the human genome. Nature 496: 498-503.

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

15. Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, et al. (2012) Reference genome sequence of the model plant Setaria. Nature Biotechnology 30: 555-+.

16. Carneiro M, Rubin C-J, Di Palma F, Albert FW, Alfoeldi J, et al. (2014) Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. Science 345: 1074-1079.

17. Jansen S, Aigner B, Pausch H, Wysocki M, Eck S, et al. (2013) Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. Bmc Genomics 14.

18. Dastjerdi A, Robert C, Watson M (2014) Low coverage sequencing of two Asian elephant (Elephas maximus) genomes. GigaScience 3: 12-12.

19. Bizon C, Spiegel M, Chasse SA, Gizer IR, Li Y, et al. (2014) Variant calling in low-coverage whole genome sequencing of a Native American population sample. Bmc Genomics 15.

20. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR (2011) Low-coverage sequencing: Implications for design of complex trait association studies. Genome Research 21: 940-951.

21. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, et al. (2012) Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nature Genetics 44: 631-U641.

22. Alex Buerkle C, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? Molecular ecology 22: 3028-3035.

23. Han E, Sinsheimer JS, Novembre J (2014) Characterizing Bias in Population Genetic Inferences from Low-Coverage Sequencing Data. Molecular Biology and Evolution 31: 723-735.

24. Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Research 17: 240-248.

25. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, et al. (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. Plos One 3.

26. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, et al. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453: 1239-U1239.

27. Mudge JM, Frankish A, Fernandez-Banet J, Alioto T, Derrien T, et al. (2011) The Origins, Evolution, and Functional Potential of Alternative Splicing in Vertebrates. Molecular Biology and Evolution 28: 2949-2959.

28. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272-U153.

29. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, et al. (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proceedings of the National Academy of Sciences of the United States of America 106: 19096-19101.

30. Teer JK, Mullikin JC (2010) Exome sequencing: the sweet spot before whole genomes. Human Molecular Genetics 19: R145-R151.

31. Cosart T, Beja-Pereira A, Chen S, Ng SB, Shendure J, et al. (2011) Exome-wide DNA capture and next generation sequencing in domestic and wild species. Bmc Genomics 12.

32. Mascher M, Richmond TA, Gerhardt DJ, Himmelbach A, Clissold L, et al. (2013) Barley whole exome capture: a tool for genomic research in the genus Hordeum and beyond. Plant Journal 76: 494-505.

33. Campbell N, Sinagra G, Jones KL, Slavov D, Gowan K, et al. (2013) Whole Exome Sequencing Identifies a Troponin T Mutation Hot Spot in Familial Dilated Cardiomyopathy. Plos One 8.

34. Albrechtsen A, Nielsen FC, Nielsen R (2010) Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. Molecular Biology and Evolution 27: 2534-2547.

35. Weir BS, Cockerham CC (1984) ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION-STRUCTURE. Evolution 38: 1358-1370.

36. Frichot E, Mathieu F, Trouillon T, Bouchard G, Francois O (2014) Fast and Efficient Estimation of Individual Ancestry Coefficients. Genetics 196: 973-+.

37. Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. Genome Research 20: 393-402.

38. Kijas JW, Lenstra JA, Hayes B, Boitard S, Neto LRP, et al. (2012) Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. Plos Biology 10.

39. Kidd JM, Gravel S, Byrnes J, Moreno-Estrada A, Musharoff S, et al. (2012) Population Genetic Inference from Personal Genome Data: Impact of Ancestry and Admixture on Human Genomic Variation. American Journal of Human Genetics 91: 660-671.

40. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.

41. Alhaddad H, Khan R, Grahn RA, Gandolfi B, Mullikin JC, et al. (2013) Extent of Linkage Disequilibrium in the Domestic Cat, Felis silvestris catus, and Its Breeds. Plos One 8.

42. Olson ZH, Whittaker DG, Rhodes OE (2013) Translocation History and Genetic Diversity in Reintroduced Bighorn Sheep. Journal of Wildlife Management 77: 1553-1563.

43. Garza JC, Gilbert-Horvath EA, Spence BC, Williams TH, Fish H, et al. (2014) Population Structure of Steelhead in Coastal California. Transactions of the American Fisheries Society 143: 134-152.

44. Huang H, Wang H, Li L, Wu Z, Chen J (2014) Genetic Diversity and Population Demography of the Chinese Crocodile Lizard (Shinisaurus crocodilurus) in China. Plos One 9.

45. Marsden CD, Lee Y, Kreppel K, Weakley A, Cornel A, et al. (2014) Diversity, Differentiation, and Linkage Disequilibrium: Prospects for Association Mapping in the Malaria Vector Anopheles arabiensis. G3-Genes Genomes Genetics 4: 121-131.

46. Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, et al. (1998) Dating the origin of the CCR5-Delta 32 AIDS-resistance allele by the coalescence of haplotypes. American Journal of Human Genetics 62: 1507-1515.

47. Kim Y, Nielsen R (2004) Linkage disequilibrium as a signature of selective sweeps. Genetics 167: 1513-1524.

48. Kijas JW, Porto-Neto L, Dominik S, Reverter A, Bunch R, et al. (2014) Linkage disequilibrium over short physical distances measured in sheep using a high-density SNP chip. Animal Genetics 45: 754-757.

49. Wan QH, Wu H, Fujihara T, Fang SG (2004) Which genetic marker for which conservation genetics issue? Electrophoresis 25: 2165-2176.

50. Vowles EJ, Amos W (2006) Quantifying ascertainment bias and species-specific length differences in human and chimpanzee microsatellites using genome sequences. Molecular Biology and Evolution 23: 598-607.

51. Curtis D, Vine AE, Knight J (2008) Investigation into the ability of SNP chipsets and microsatellites to detect association with a disease locus. Annals of Human Genetics 72: 547-556.

52. Miller JM, Malenfant RM, David P, Davis CS, Poissant J, et al. (2014) Estimating genome-wide heterozygosity: effects of demographic history and marker type. Heredity 112: 240-247.

53. Jiang Y, Xie M, Chen W, Talbot R, Maddox JF, et al. (2014) The sheep genome illuminates biology of the rumen and lipid metabolism. Science 344: 1168-1173.

54. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.

55. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43: 491-+.

56. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

57. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20: 1297-1303.

58. Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. arXiv.

59. Browning BL, Browning SR (2013) Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. Genetics 194: 459-+.

60. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81: 559-575.

61. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, et al. (2011) The variant call format and VCFtools. Bioinformatics 27: 2156-2158.

62. Dominik S, Henshall JM, Hayes BJ (2012) A single nucleotide polymorphism on chromosome 10 is highly predictive for the polled phenotype in Australian Merino sheep. Animal Genetics 43: 468-470.

63. Auton A, McVean G (2007) Recombination rate estimation in the presence of hotspots. Genome Research 17: 1219-1227.

Supplementary material

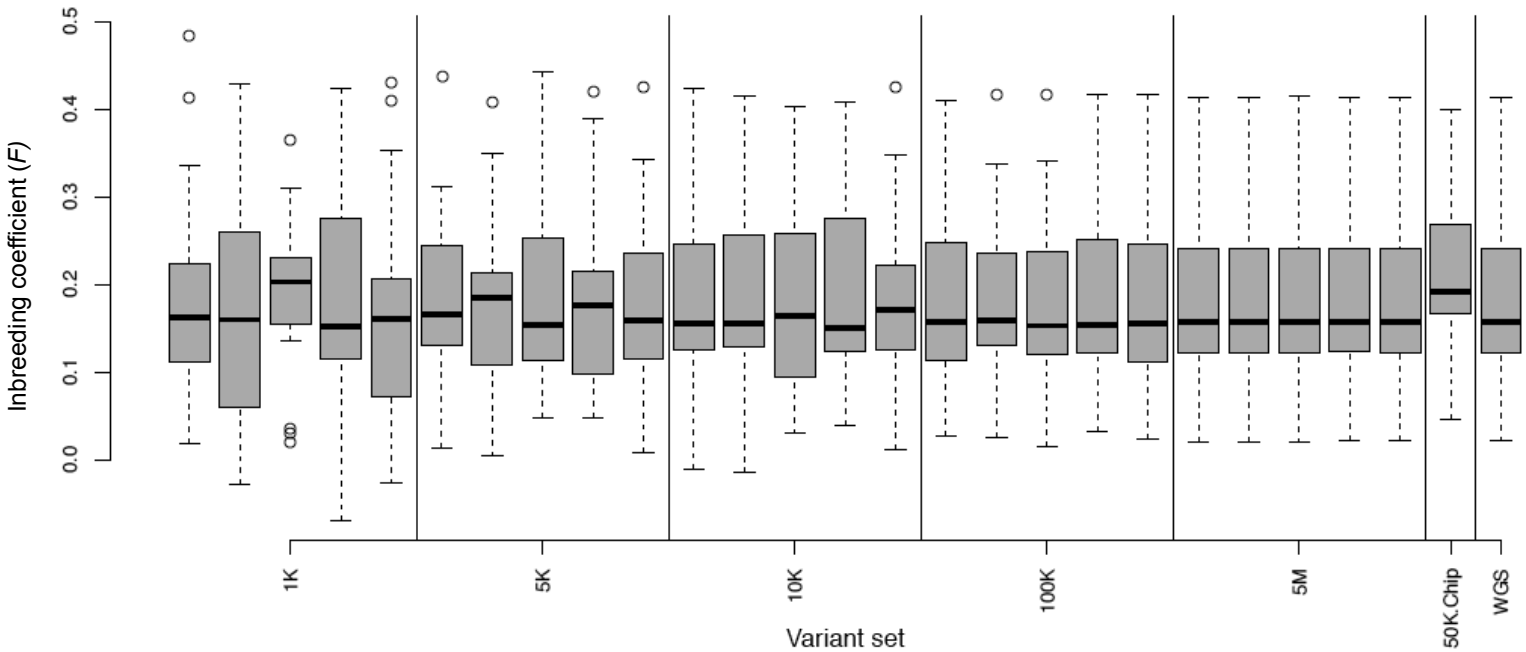

Figure S1. Inbreeding coefficient (*F*) in bezoars calculated from Whole Genome Sequence data and random and non-random subsamples of variants.

The figure presents for each replicate of random panel and for each non-random panel the boxplot for 18 individual estimates of *F*.

Random panels are denoted by their size (i.e. 1k to 5M), and non-random panels by: 50K.Chip (Illumina® ovine 50K SNP Beadchip), WGS (all variants extracted from whole genome sequences).

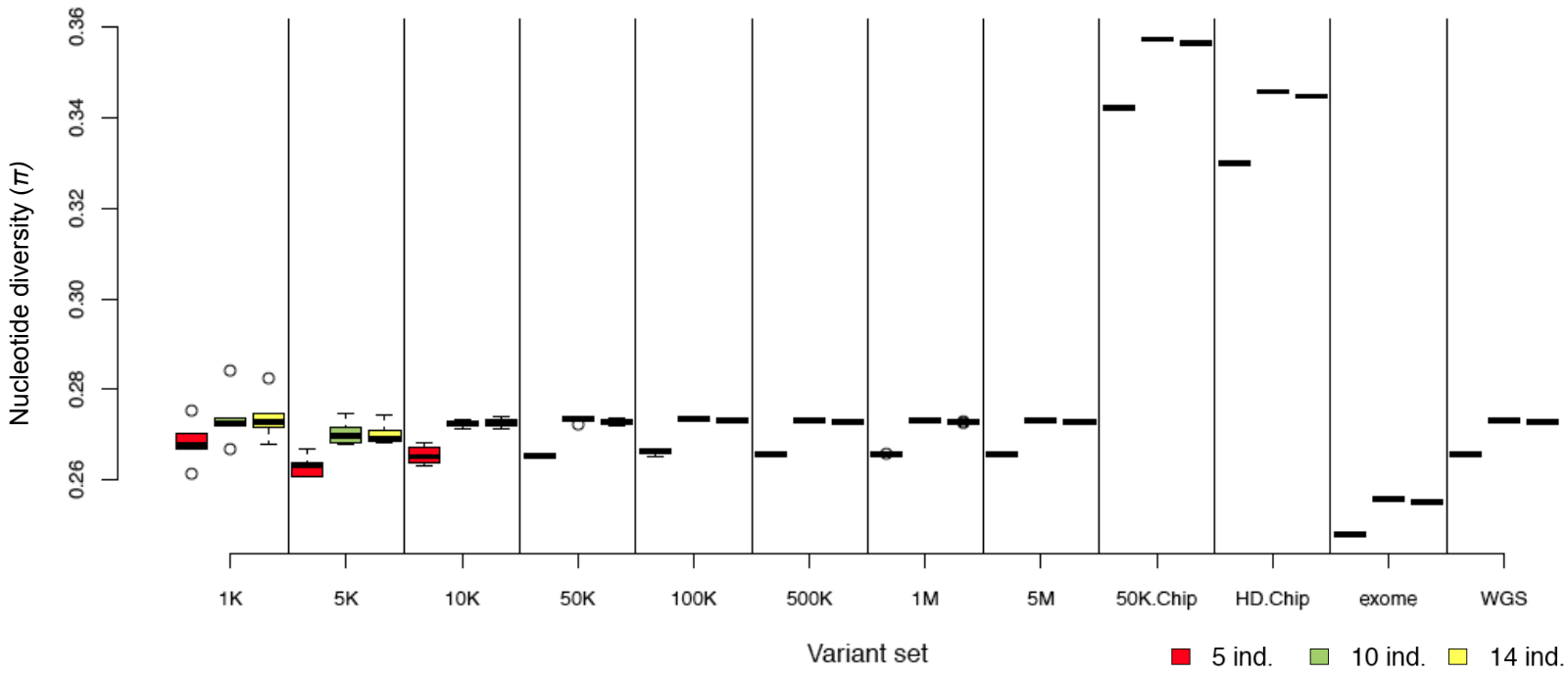

Figure S2. Nucleotide diversity (π) in Asiatic mouflon calculated from Whole Genome Sequence data and random and non-random subsamples of variants

Nucleotide diversity (π) estimated for 132 datasets in Asiatic mouflon. The figure presents for each size of random panel the boxplot for 5 independent replicates for each sample size, and for each non-random dataset the pi value for each sample size.

Random panels are denoted by their size (i.e. 1k to 5M), and non-random panels by: 50K.Chip (Illumina® ovine 50K SNP Beadchip), HD.Chip (Illumina® ovine HD Beadchip) exome (exome capture simulation), WGS (all variants extracted from whole genome sequences). For each set of variants the sample sizes are from left to right: 5 (red), 10 (green) and 14 (yellow) individuals.

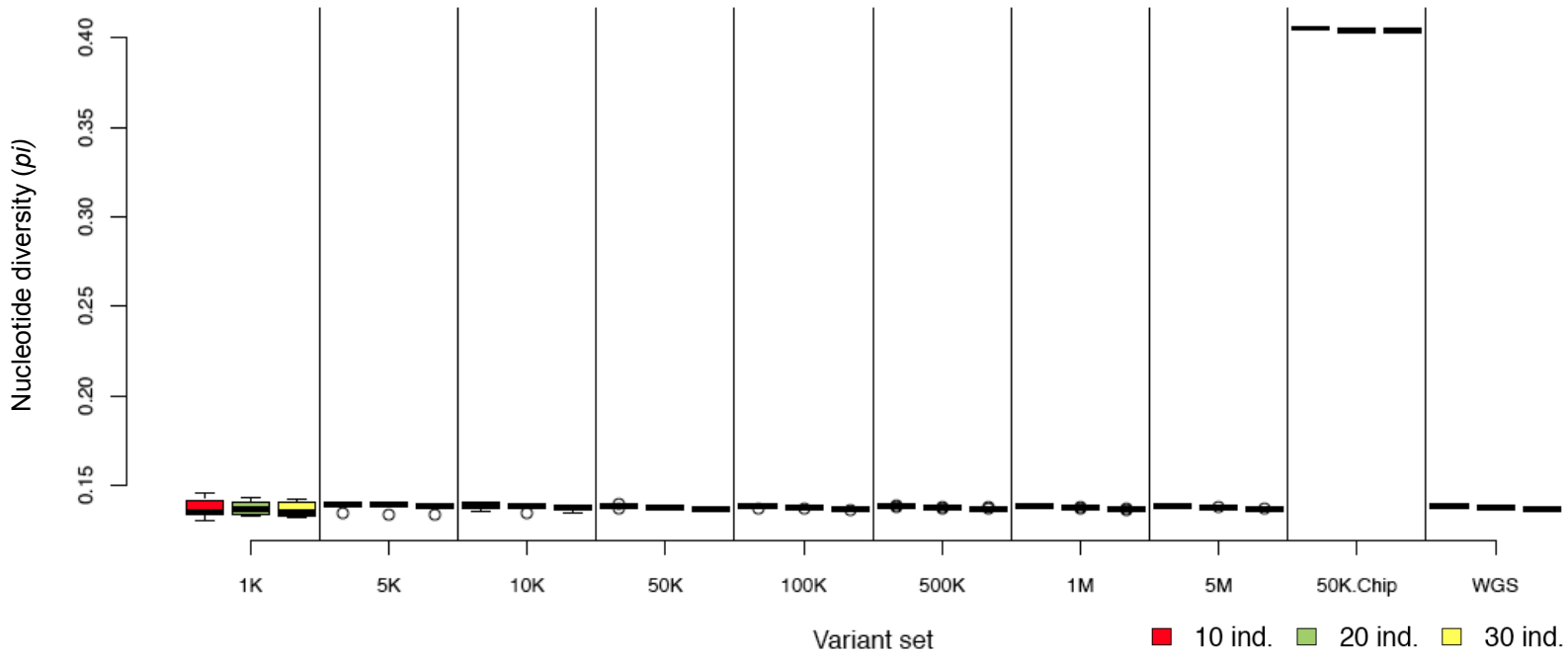

Figure S3. Nucleotide diversity (π) in Moroccan goats calculated from Whole Genome Sequence data and random and non-random subsamples of variants.

Nucleotide diversity (π) estimated for 126 datasets in goats. The figure presents for each size of random panel the boxplot for 5 independent replicates for each sample size, and for each non-random dataset the π value for each sample size.

Random panels are denoted by their size (i.e. 1k to 5M), and non-random panels by: 50K.Chip (Illumina® caprine 50K SNP Beadchip), WGS (all variants extracted from whole genome sequences). For each set of variants the sample sizes are from left to right: 10 (red), 20 (green) and 30 (yellow) individuals.

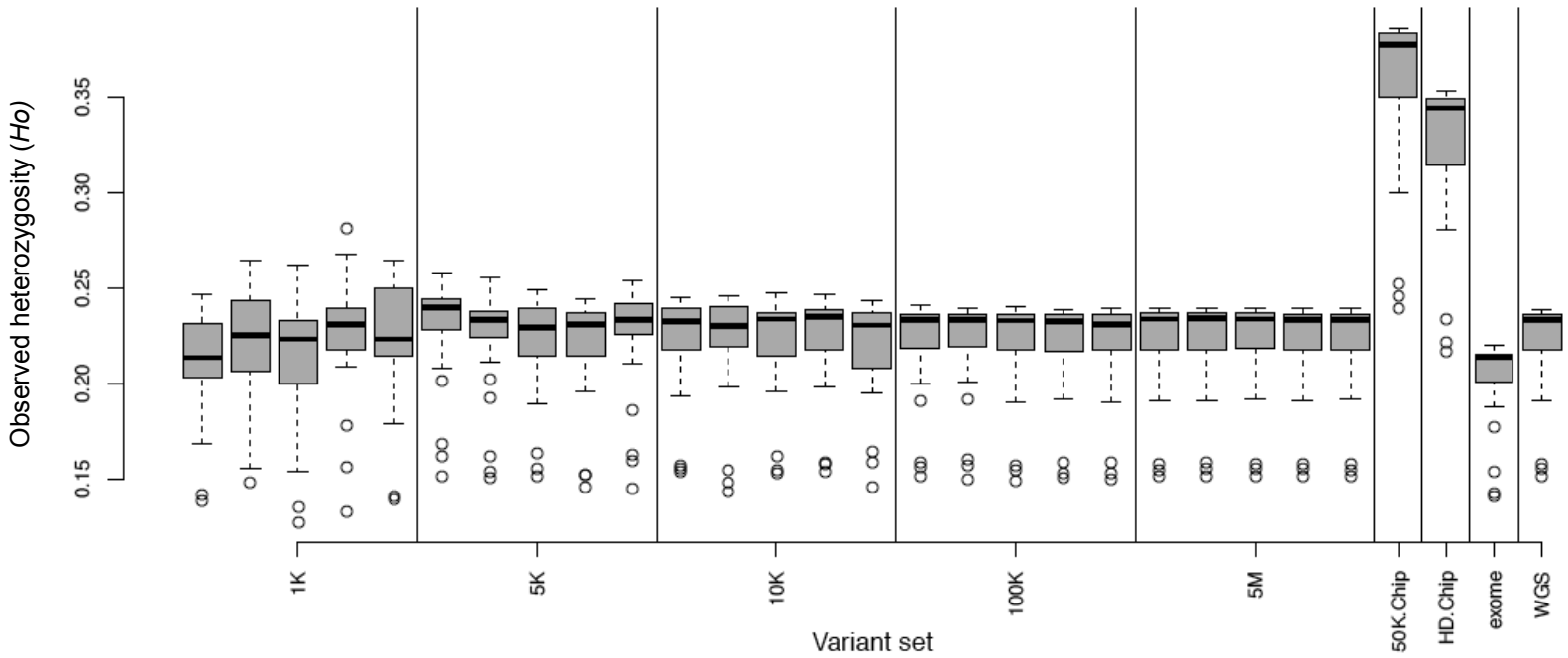

Figure S4. Observed heterozygosity (*Ho*) in Moroccan sheep calculated from Whole Genome Sequence data and random and non-random subsamples of variants.

The figure presents for each replicate of random panel and for each non-random panel the boxplot for 30 individual estimates of *Ho*.

Random panels are denoted by their size (i.e. 1k to 5M), and non-random panels by: 50K.Chip (Illumina® ovine 50K SNP Beadchip), HD.Chip (Illumina® ovine HD Beadchip) exome (exome capture simulation), WGS (all variants extracted from whole genome sequences).

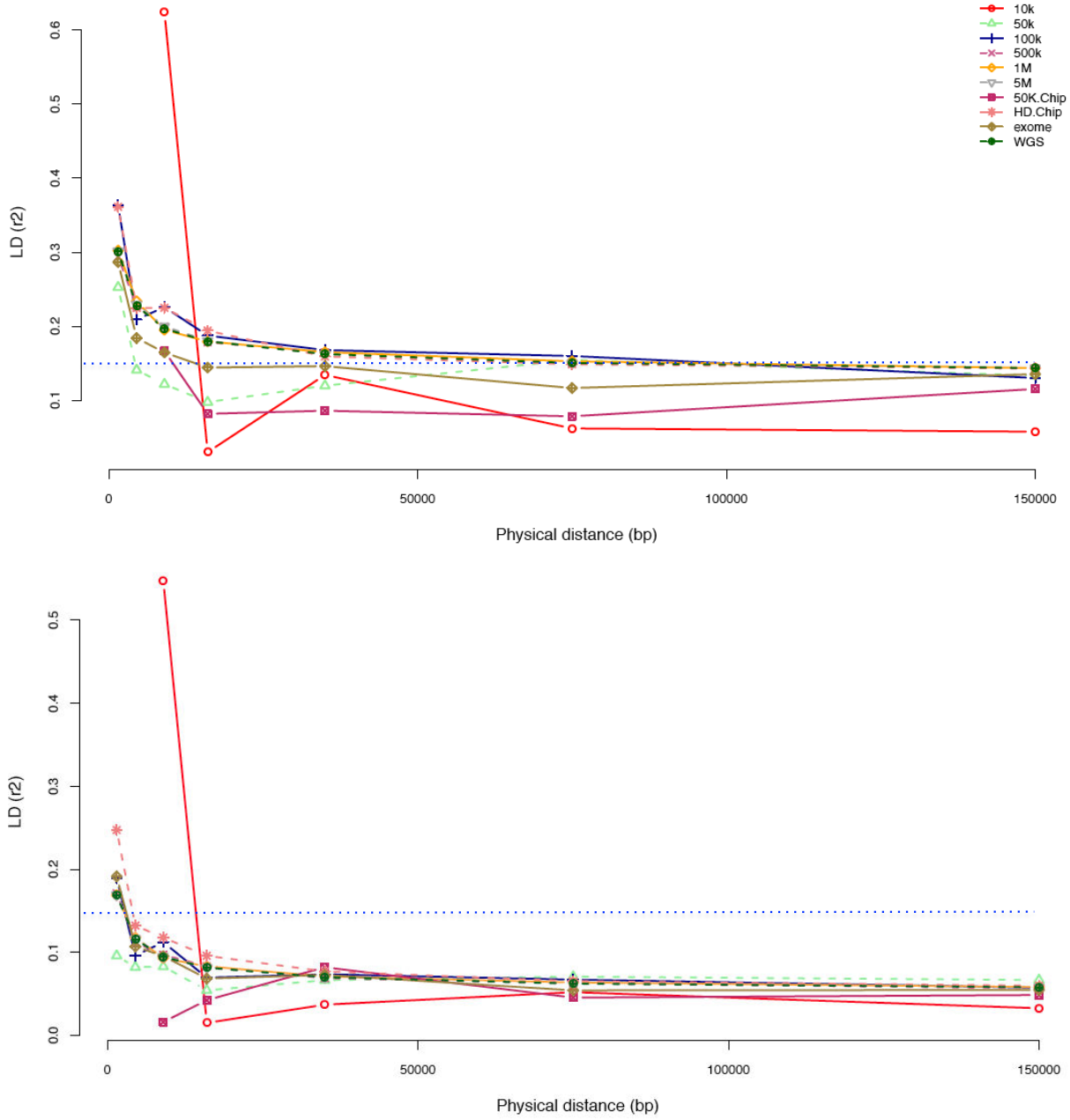

Figure S5. Decay of Linkage disequilibrium (r^2) as a function of physical distance for different panels of variant in Asiatic mouflon.

The Linkage Disequilibrium was calculated on 5 (top) and on 14 individuals (bottom). Inter-SNP distances (bp) were binned into the classes: 1-3K; 3K-6K; 6K-12K; 12K-20K; 20K-50K; 50K-100K; 100K-200K. Random panels are denoted by their size (i.e. 1k to 5M), and non-random panels by: 50K.Chip (Illumina® ovine 50K SNP Beadchip), HD.Chip (Illumina® ovine HD Beadchip) exome (exome capture simulation), WGS (variants extracted from whole genome sequences).

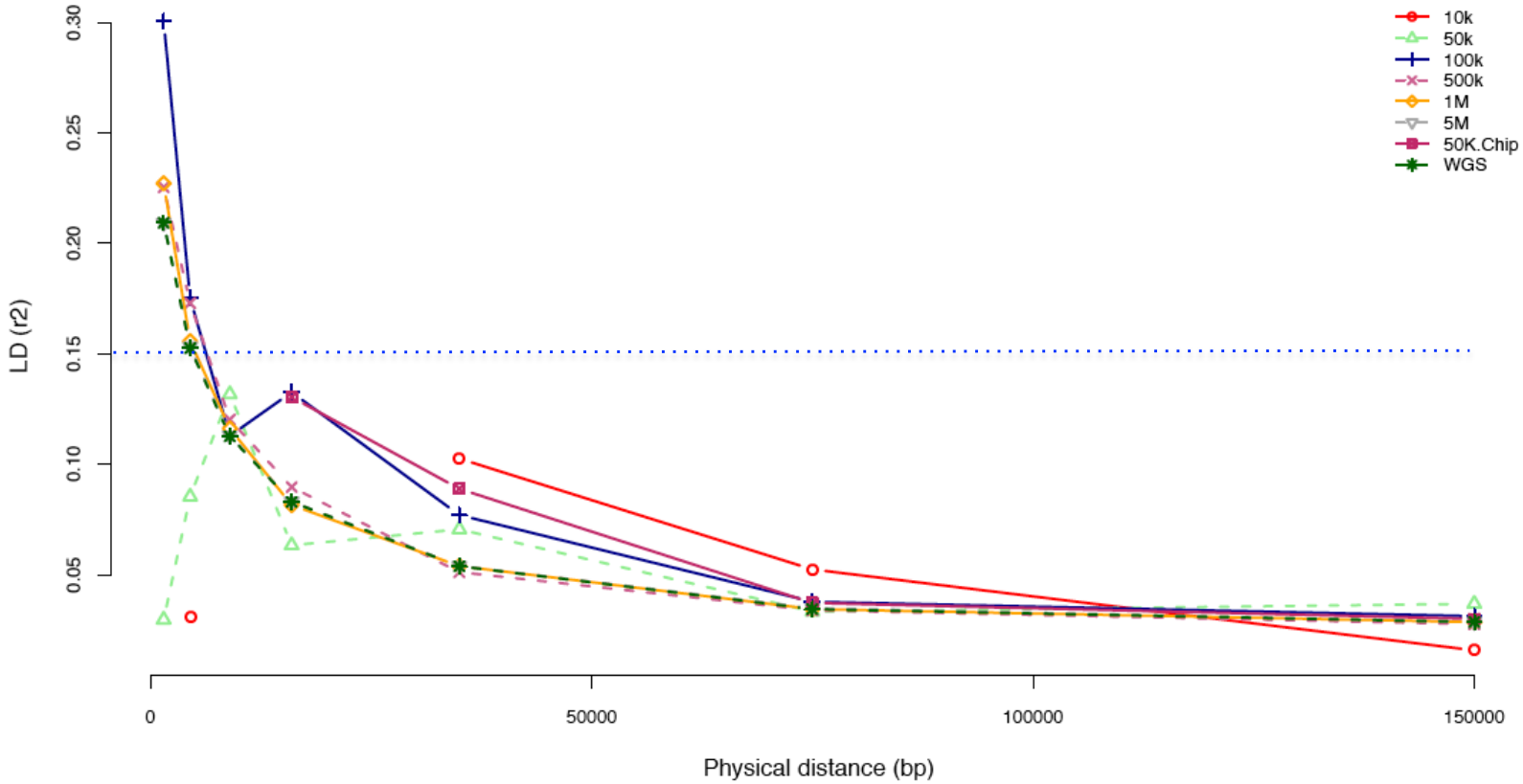

Figure S6. Decay of Linkage disequilibrium (r^2) as a function of physical distance for different panels of variant in Moroccan goats.

The Linkage Disequilibrium was calculated on 30 goats. Inter-SNP distances (bp) were binned into the classes: 1-3K; 3K-6K; 6K-12K; 12K-20K; 20K-50K; 50K-100K; 100K-200K. Random panels are denoted by their size (i.e. 1k to 5M), and non-random panels by: 50K.Chip (Illumina® ovine 50K SNP Beadchip), WGS (variants extracted from whole genome sequences).

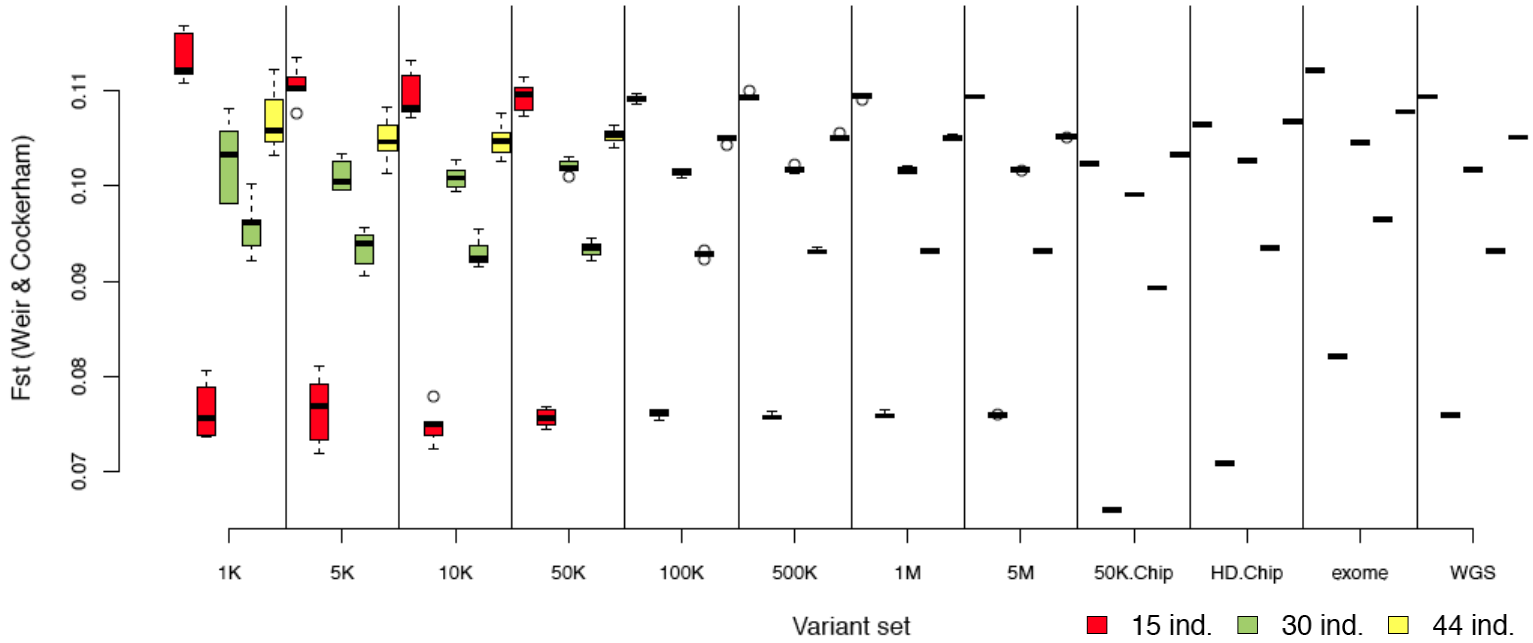

Figure S7. Fixation index (Fst) between Moroccan sheep (*O. aries*) and Asiatic mouflon (*O. orientalis*) for different panels of variants and different samples of individuals.

The fixation index Fst (Weir and Cockerham, 1984) estimated from 220 datasets of *Ovis* combining different panels of variants and different sample sizes. The figure presents for each size of random panel the boxplot for 5 independent replicates for each sample size, and for each non-random dataset the Fst value for each sample size. Random panels are denoted by their size (i.e. 1k to 5M), and non-random panels by: 50K.Chip (Illumina® ovine 50K SNP Beadchip), HD.Chip (Illumina® ovine HD Beadchip) exome (exome capture simulation), WGS (all variants extracted from whole genome sequences). For each set of variants the sample sizes are from left to right: 15 (red, 2 replicates), 30 (green, 2 replicates) and 44 (yellow) individuals.

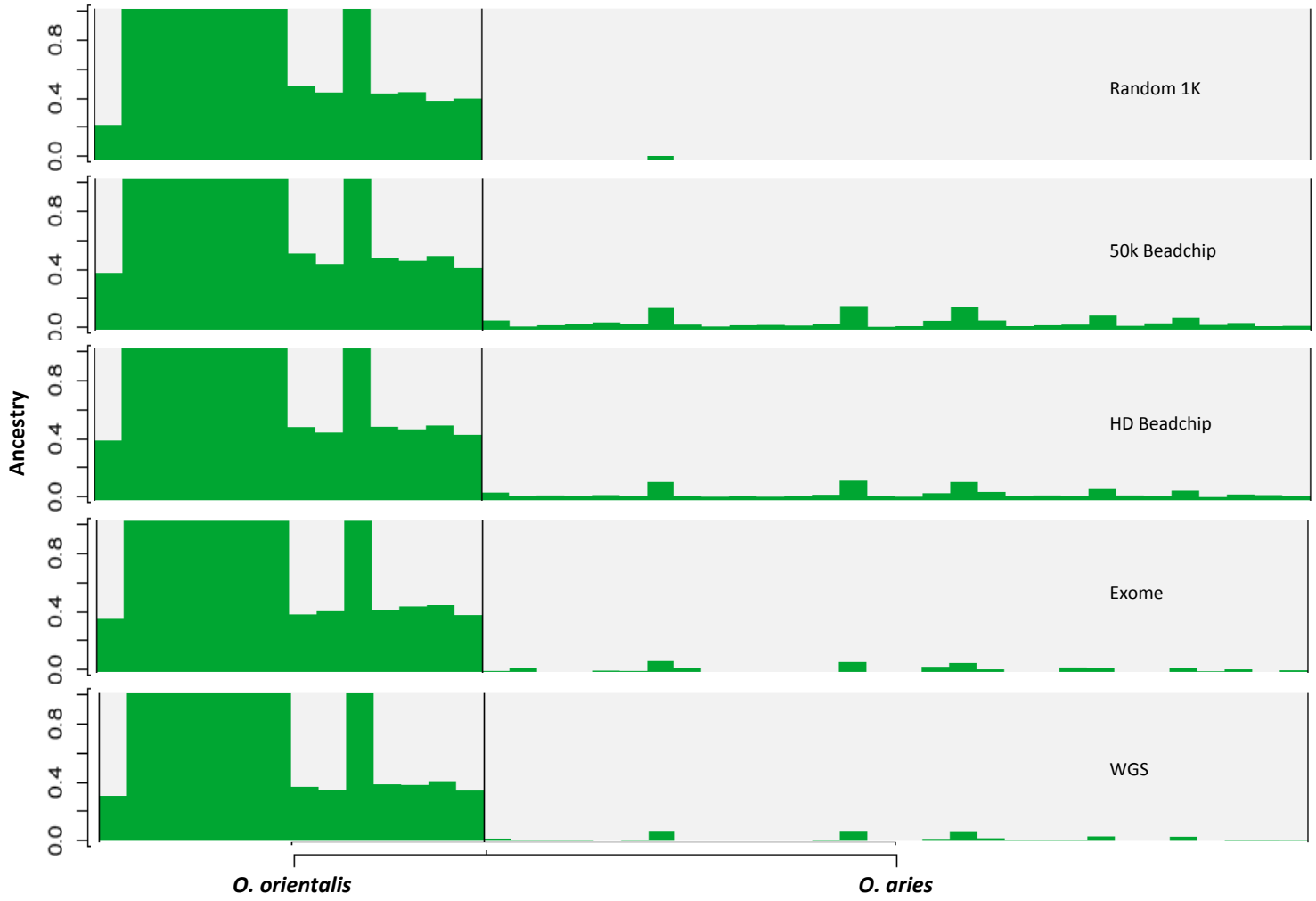

Figure S8. Population structure in 44 sheep and Asiatic mouflon for different panels of variants.

Plot of sNMF Ancestry estimates for $k = 2$. Each bar represents the estimated membership coefficients for each accession in each of the 2 clusters.

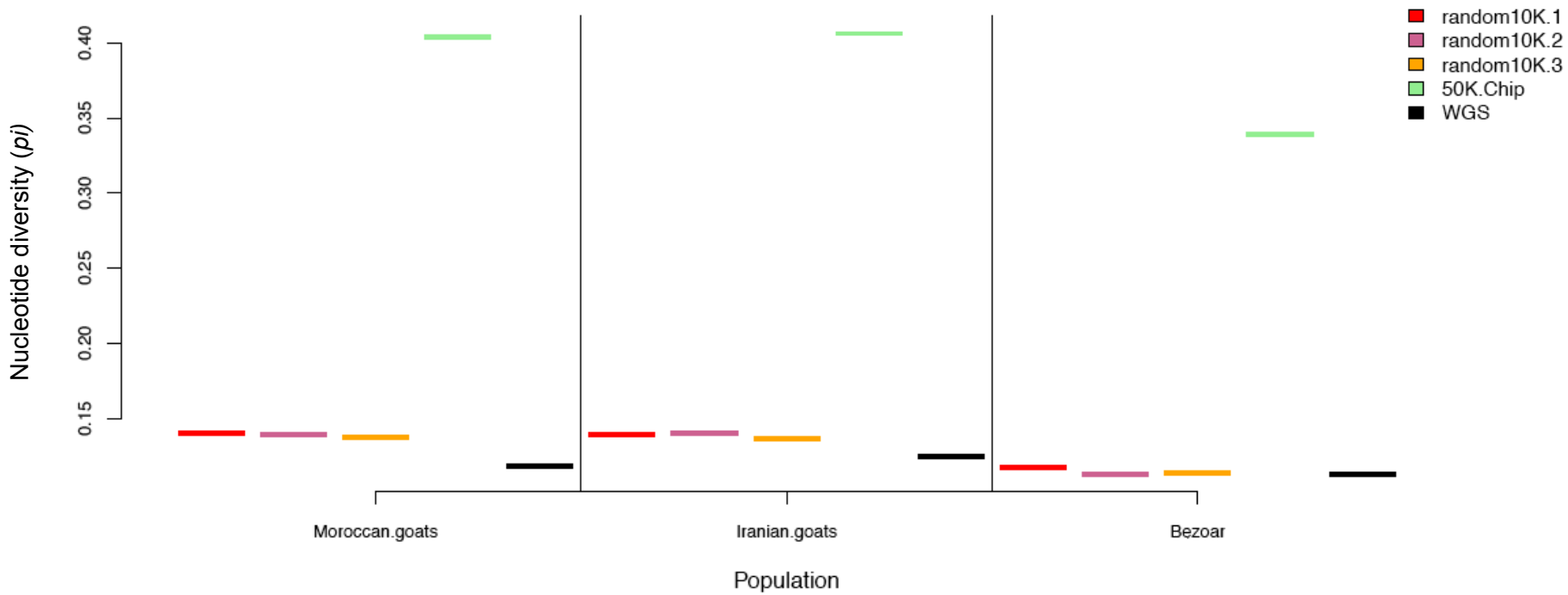

Figure S9. Nucleotide diversity (π) estimated in 2 domestic and wild *Capra* populations with random and commonly used panels of variants.

Plot of Nucleotide diversity (π) estimated with 3 independent sets of 10K variants defined in Moroccan goats (10K), and with Illumina® caprine 50K SNP Beadchip (50K.Chip), and variants extracted from whole genome sequences (WGS).

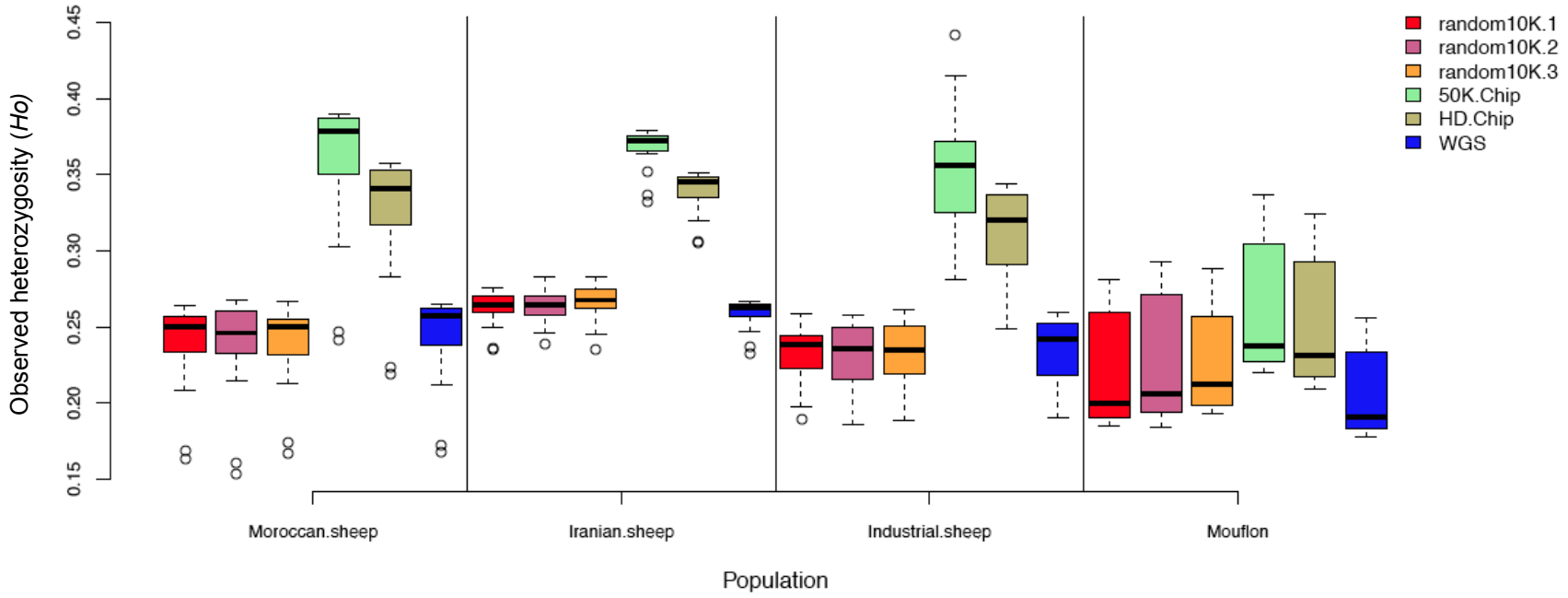

Figure S10. Observed heterozygosity (*Ho*) estimated in 3 domestic and wild *Ovis* populations with random and commonly used panels of variants.

Plot of individual heterozygosity (*Ho*) estimated with 3 independent sets of 10K variants defined in Moroccan sheep (10K), and with Illumina® ovine 50K SNP Beadchip (50K.Chip), Illumina® ovine HD Beadchip (HD.Chip), and variants extracted from whole genome sequences (WGS).

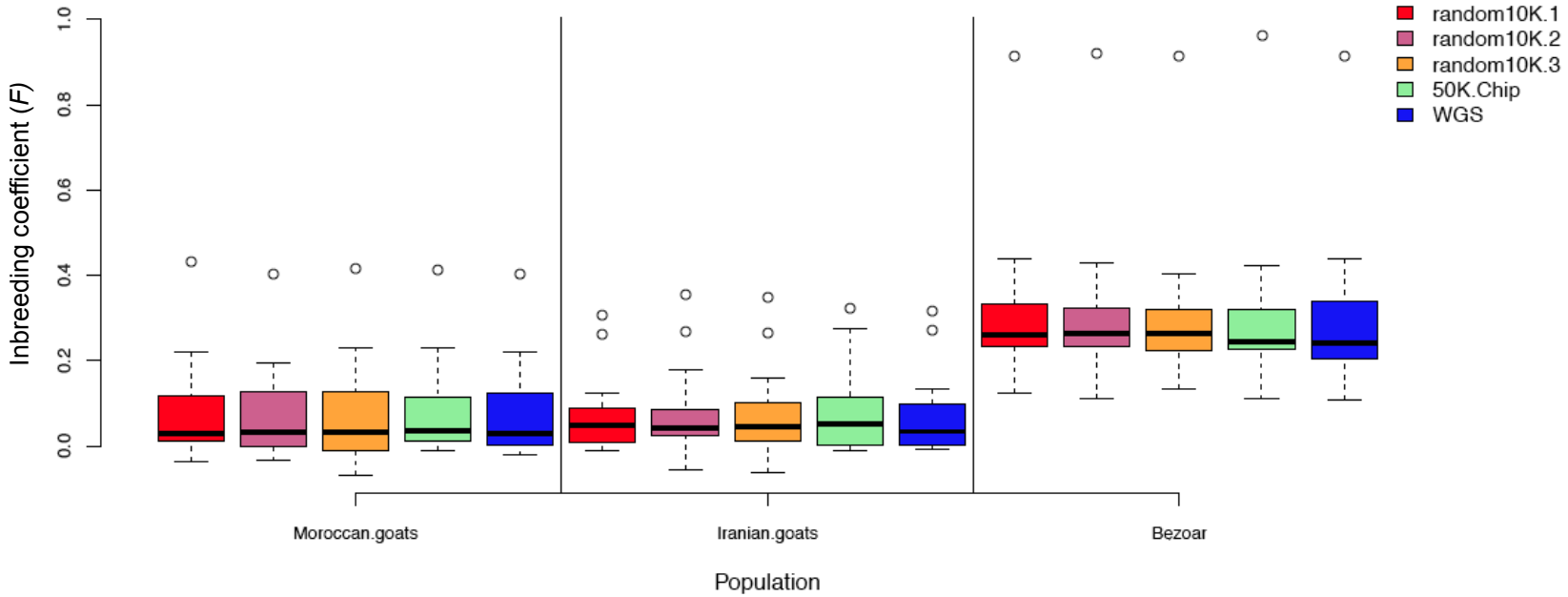

Figure S11. Inbreeding coefficient (*F*) estimated in 2 domestic and wild *Capra* populations with random and commonly used panels of variants.

Plot of individual Inbreeding coefficient (*F*) estimated with 3 independent sets of 10K variants defined on Moroccan goats (10K), and with Illumina® caprine 50K SNP beadchip (50K.Chip), and variants extracted from whole genome sequences (WGS).

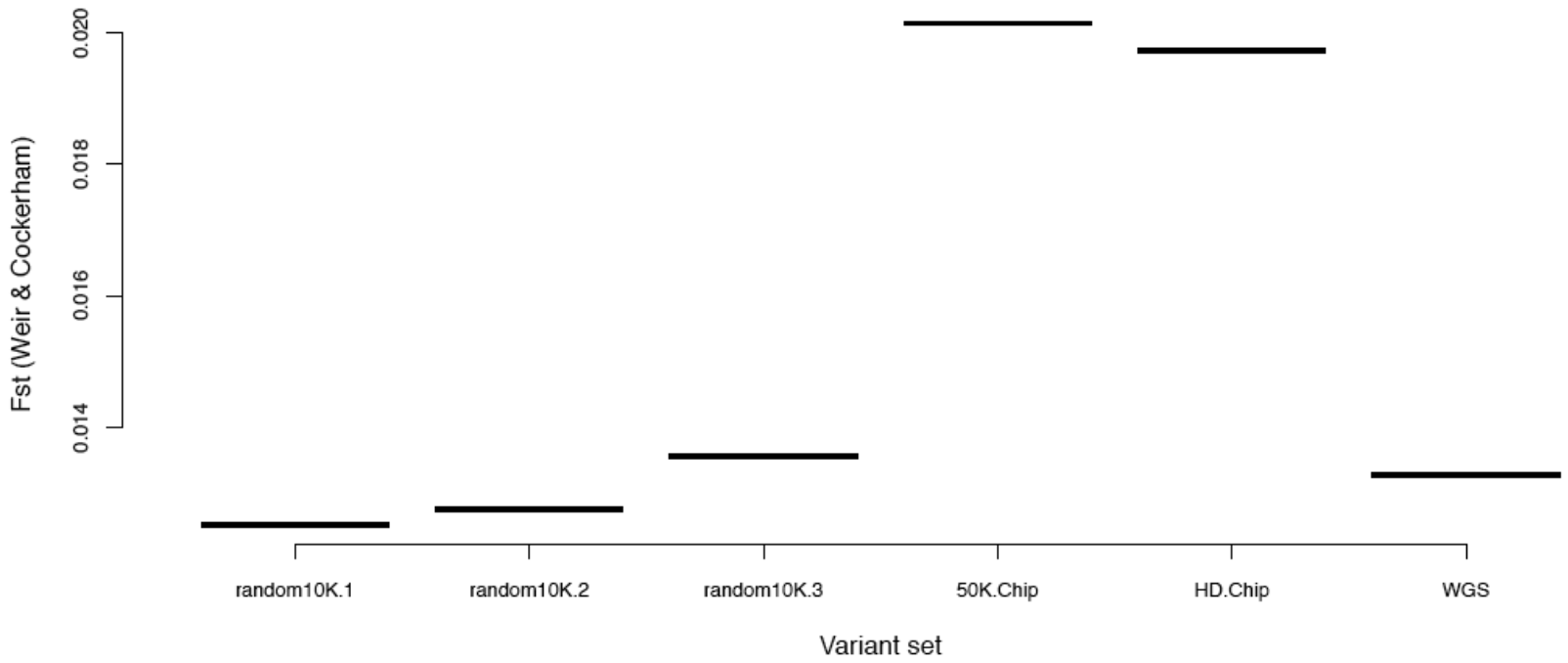

Figure S12. The fixation index (Fst) between Moroccan, Iranian and industrial sheep goats estimated with random and commonly used panels of variants.

Plot of Fixation index Fst (Weir and Cockerham, 1984) estimated with 3 independent sets of 10K variants defined in Moroccan sheep (10K), and with Illumina® ovine 50K SNP Beadchip (50K.Chip), Illumina® ovine HD Beadchip (HD.Chip), and variants extracted from whole genome sequences (WGS).

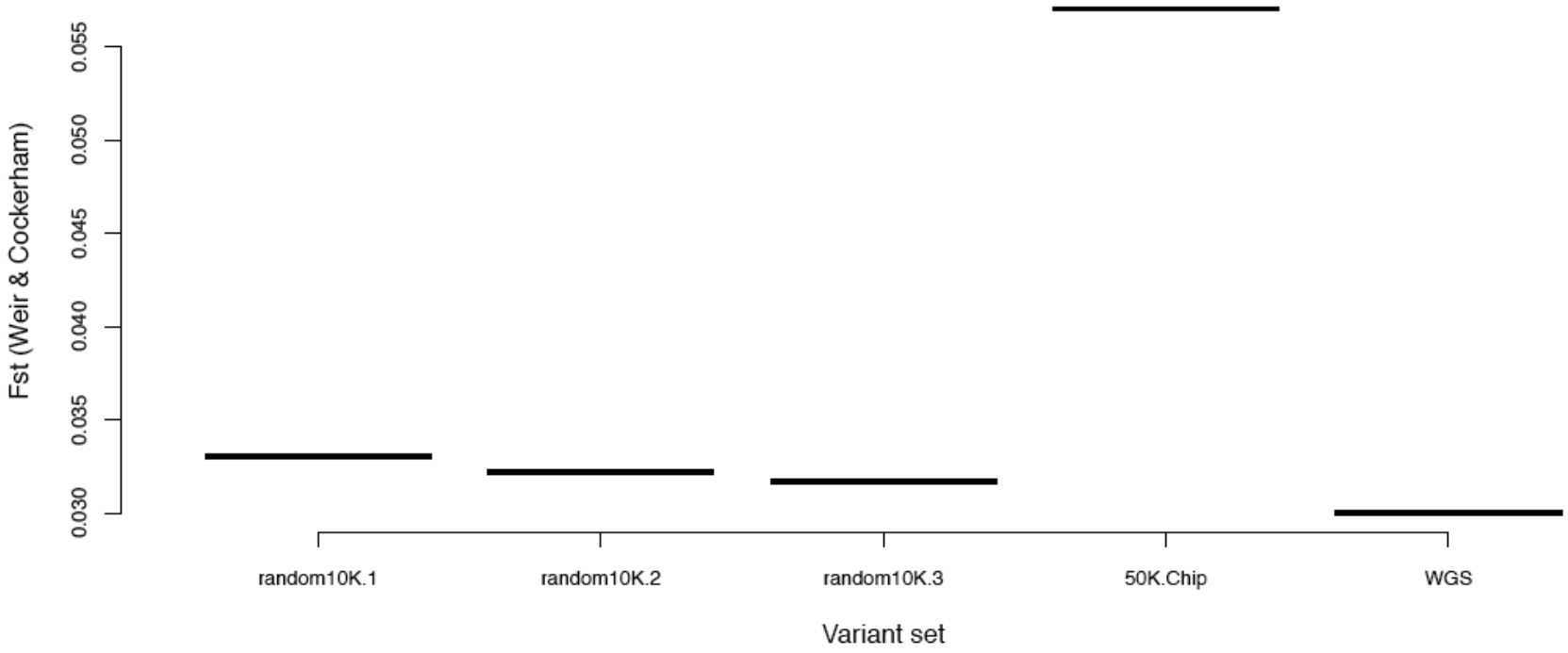

Figure S13. The fixation index (Fst) between Moroccan and Iranian goats estimated with random and commonly used panels of variants.

The Fixation index Fst (Weir and Cockerham, 1984) estimated with 3 independent sets of 10K variants defined on Moroccan goats (10K), and with Illumina® caprine 50K SNP beadchip (50K.Chip), and variants extracted from whole genome sequences (WGS).

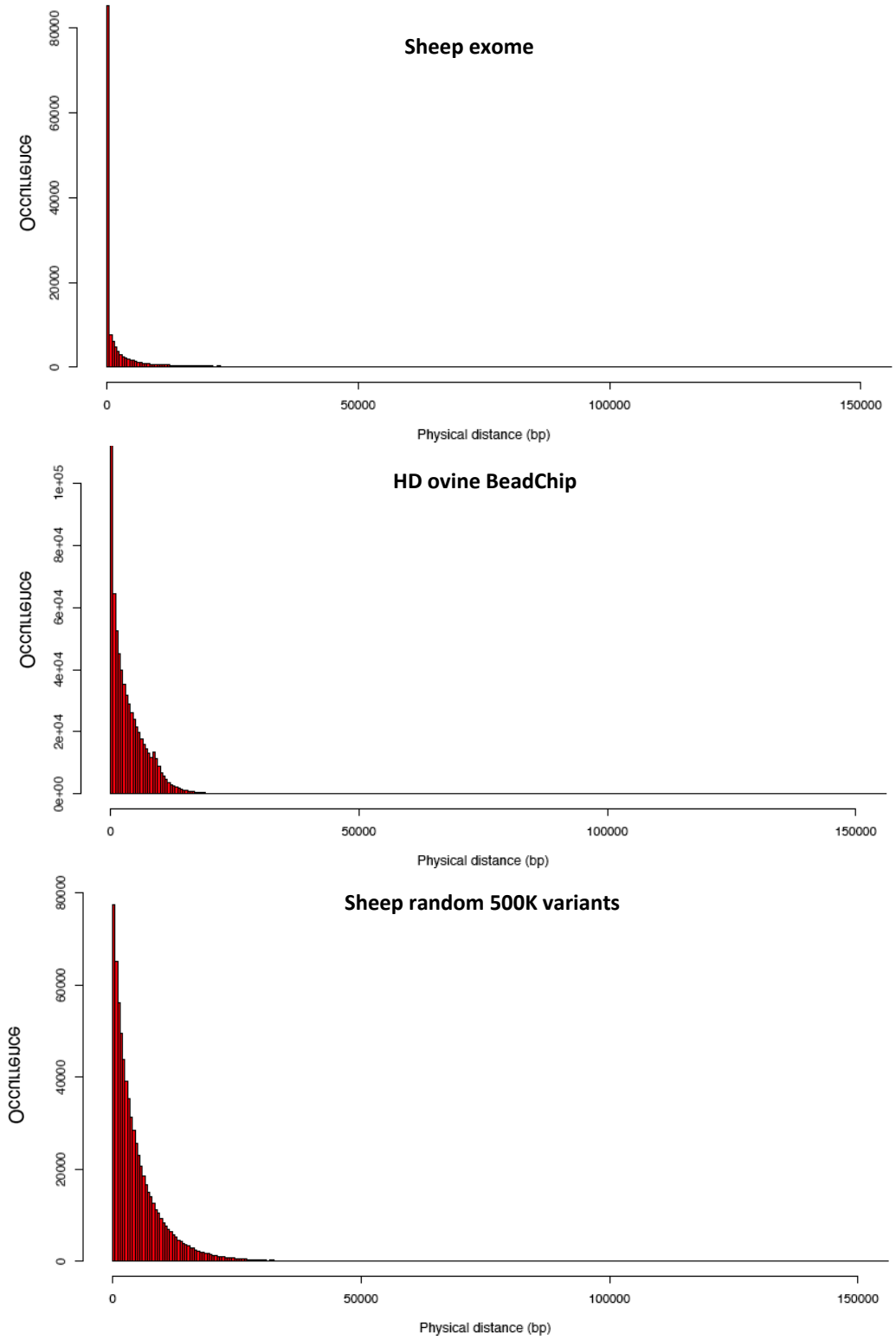

Figure S14. Distribution of physical distances between adjacent variants in sheep exome, HD ovine BeadChip and a random panel of 500K variants.

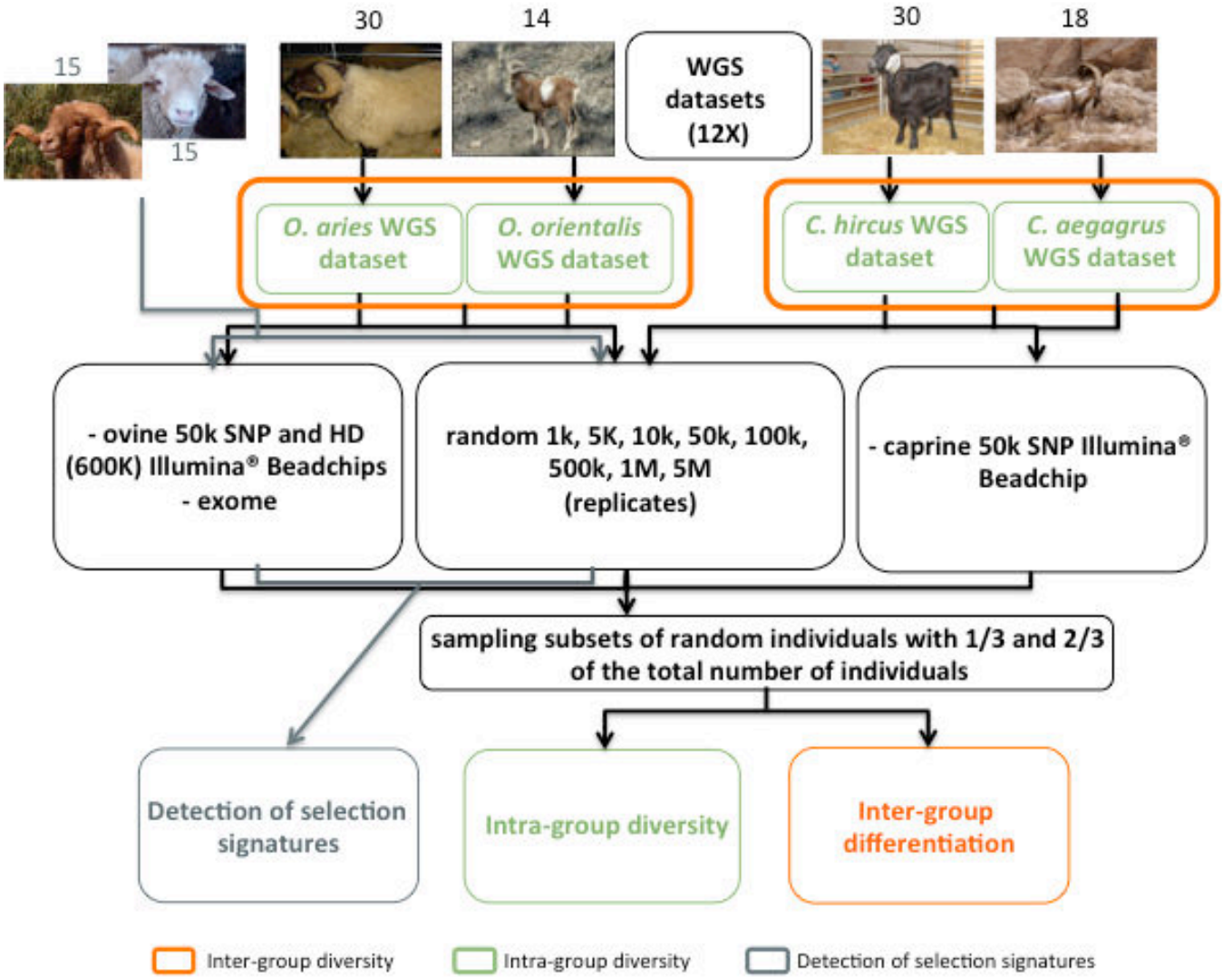

Figure S15. Flow-chart describing sampling random and non-random panels of variants and individuals.

Whole genome sequences are denoted by WGS.

Table S1. Table listing the samples used for the analyses described in this paper, the accession ID of the sample in the Biosamples archive, the accession ID of the aligned bam file in the ENA archive.

Sample name	Biosample accession	ENA aligned bam file	Species	Location
IRCA-C3-1001	SAMEA2065212	ERZ018663	Capra aegagrus	Iran
IRCA-F2-5026	SAMEA2065213	ERZ018662	Capra aegagrus	Iran
IRCA-F2-5064	SAMEA2065214	ERZ018650	Capra aegagrus	Iran
IRCA-F2-5066	SAMEA2065215	ERZ018658	Capra aegagrus	Iran
IRCA-F3-0597	SAMEA2065216	ERZ018659	Capra aegagrus	Iran
IRCA-F3-0600	SAMEA2065217	ERZ018651	Capra aegagrus	Iran
IRCA-G2-0568	SAMEA2065218	ERZ018656	Capra aegagrus	Iran
IRCA-G2-5063	SAMEA2065220	ERZ018660	Capra aegagrus	Iran
IRCA-G2-5065	SAMEA2188056	ERZ018649	Capra aegagrus	Iran
IRCA-I11-0001	SAMEA2065221	ERZ018657	Capra aegagrus	Iran
IRCA-I11-0002	SAMEA2065222	ERZ018653	Capra aegagrus	Iran
IRCA-I11-0003	SAMEA2065223	ERZ018655	Capra aegagrus	Iran
IRCA-I6-5237	SAMEA2065224	ERZ018665	Capra aegagrus	Iran
IRCA-K12-0005	SAMEA2065225	ERZ018664	Capra aegagrus	Iran
IRCA-K7-0009	SAMEA2065226	ERZ018654	Capra aegagrus	Iran
IRCA-M12-0008	SAMEA2065227	ERZ018666	Capra aegagrus	Iran
IRCA-M7-0652	SAMEA2065421	ERZ018652	Capra aegagrus	Iran
IRCA-N8-0006	SAMEA1966535	ERZ018661	Capra aegagrus	Iran
IRCH-B3-5031	SAMEA1966659	ERZ018647	Capra hircus	Iran
IRCH-B4-5209	SAMEA1968884	ERZ018631	Capra hircus	Iran
IRCH-B5-5032	SAMEA2065422	ERZ018645	Capra hircus	Iran
IRCH-C3-5039	SAMEA2065423	ERZ018638	Capra hircus	Iran
IRCH-C5-5206	SAMEA2065424	ERZ018646	Capra hircus	Iran
IRCH-C6-5204	SAMEA2065425	ERZ018633	Capra hircus	Iran
IRCH-C7-5144	SAMEA2065426	ERZ018634	Capra hircus	Iran
IRCH-D5-5240	SAMEA2065427	ERZ018639	Capra hircus	Iran
IRCH-D6-5189	SAMEA2065428	ERZ018641	Capra hircus	Iran
IRCH-D7-5132	SAMEA2065429	ERZ018628	Capra hircus	Iran
IRCH-E5-5053	SAMEA2065430	ERZ018643	Capra hircus	Iran
IRCH-E6-5087	SAMEA2065431	ERZ018636	Capra hircus	Iran
IRCH-E7-5193	SAMEA2065432	ERZ018635	Capra hircus	Iran
IRCH-F11-5140	SAMEA2065433	ERZ018637	Capra hircus	Iran
IRCH-F3-5044	SAMEA2065434	ERZ018644	Capra hircus	Iran
IRCH-F4-5093	SAMEA2065435	ERZ018629	Capra hircus	Iran
IRCH-F5-5133	SAMEA2065436	ERZ018642	Capra hircus	Iran
IRCH-G3-5210	SAMEA2065437	ERZ018630	Capra hircus	Iran
IRCH-G4-5194	SAMEA2065438	ERZ018640	Capra hircus	Iran
IRCH-G5-5185	SAMEA2065587	ERZ018632	Capra hircus	Iran
MOCH-AB11-2160	SAMEA2012652	ERZ018686	Capra hircus	Morocco
MOCH-H19-1343	SAMEA2012697	ERZ018783	Capra hircus	Morocco
MOCH-K14-0425	SAMEA2012702	ERZ018809	Capra hircus	Morocco
MOCH-L12-0379	SAMEA2012709	ERZ018708	Capra hircus	Morocco
MOCH-L14-0418	SAMEA2012711	ERZ018773	Capra hircus	Morocco
MOCH-L18-1280	SAMEA2012707	ERZ018765	Capra hircus	Morocco
MOCH-M15-1213	SAMEA2012749	ERZ018736	Capra hircus	Morocco
MOCH-N10-3078	SAMEA2012752	ERZ018729	Capra hircus	Morocco
MOCH-O11-0304	SAMEA2012761	ERZ018720	Capra hircus	Morocco
MOCH-O14-1203	SAMEA2012763	ERZ018741	Capra hircus	Morocco
MOCH-P12-0217	SAMEA2012762	ERZ018739	Capra hircus	Morocco
MOCH-P16-1251	SAMEA2012823	ERZ018779	Capra hircus	Morocco
MOCH-Q10-0090	SAMEA2012826	ERZ018694	Capra hircus	Morocco
MOCH-Q12-0031	SAMEA2012828	ERZ018690	Capra hircus	Morocco
MOCH-Q13-0153	SAMEA2012829	ERZ018703	Capra hircus	Morocco
MOCH-R14-1105	SAMEA2012839	ERZ018758	Capra hircus	Morocco
MOCH-S12-1071	SAMEA2012896	ERZ018693	Capra hircus	Morocco
MOCH-S16-1135	SAMEA2012900	ERZ018755	Capra hircus	Morocco

MOCH-S5-0045	SAMEA2012902	ERZ018801	Capra hircus	Morocco
MOCH-S8-2252	SAMEA2012906	ERZ018671	Capra hircus	Morocco
MOCH-T11-1036	SAMEA2037793	ERZ018810	Capra hircus	Morocco
MOCH-T4-3026	SAMEA2012911	ERZ018817	Capra hircus	Morocco
MOCH-T5-0057	SAMEA2012963	ERZ018689	Capra hircus	Morocco
MOCH-T6-0074	SAMEA2012964	ERZ018688	Capra hircus	Morocco
MOCH-U10-0279	SAMEA2012969	ERZ018713	Capra hircus	Morocco
MOCH-U14-1058	SAMEA2012973	ERZ018794	Capra hircus	Morocco
MOCH-V12-1042	SAMEA2012983	ERZ018712	Capra hircus	Morocco
MOCH-V8-2274	SAMEA2013048	ERZ018730	Capra hircus	Morocco
MOCH-X5-2098	SAMEA2013062	ERZ018811	Capra hircus	Morocco
IROO-C3-0001	SAMEA2012637	ERZ017220	Ovis orientalis	Iran
IROO-D6-0002	SAMEA2012638	ERZ017221	Ovis orientalis	Iran
IROO-D6-0003	SAMEA2012639	ERZ017222	Ovis orientalis	Iran
IROO-D6-0004	SAMEA2012640	ERZ017223	Ovis orientalis	Iran
IROO-D6-0005	SAMEA2012641	ERZ017224	Ovis orientalis	Iran
IROO-D6-0006	SAMEA2065600	ERZ017225	Ovis orientalis	Iran
IROO-D6-5104	SAMEA2065601	ERZ017226	Ovis orientalis	Iran
IROO-E3-5492	SAMEA2012643	ERZ017227	Ovis orientalis	Iran
IROO-E5-5146	SAMEA2012642	ERZ017228	Ovis orientalis	Iran
IROO-F5-5079	SAMEA1967031	ERZ017229	Ovis orientalis	Iran
IROO-J11-0602	SAMEA2065602	ERZ017230	Ovis orientalis	Iran
IROO-J11-0905	SAMEA2065603	ERZ017231	Ovis orientalis	Iran
IROO-K7-0642	SAMEA1972234	ERZ017232	Ovis orientalis	Iran
IROO-N13-5061	SAMEA2065604	ERZ017233	Ovis orientalis	Iran
IROA-B2-5037	SAMEA2012928	ERZ017155	Ovis aries	Iran
IROA-B2-5296	SAMEA2065588	ERZ017157	Ovis aries	Iran
IROA-B3-5134	SAMEA2065589	ERZ017161	Ovis aries	Iran
IROA-B4-5190	SAMEA2012929	ERZ017160	Ovis aries	Iran
IROA-B5-5295	SAMEA2012927	ERZ017162	Ovis aries	Iran
IROA-B6-5139	SAMEA2065590	ERZ017156	Ovis aries	Iran
IROA-C3-5212	SAMEA2012933	ERZ017154	Ovis aries	Iran
IROA-C6-5187	SAMEA2065591	ERZ017148	Ovis aries	Iran
IROA-C7-5042	SAMEA2065592	ERZ017159	Ovis aries	Iran
IROA-D5-5081	SAMEA2065593	ERZ017152	Ovis aries	Iran
IROA-D6-5152	SAMEA2012930	ERZ017146	Ovis aries	Iran
IROA-D7-5033	SAMEA2065594	ERZ017163	Ovis aries	Iran
IROA-E5-5157	SAMEA2065595	ERZ017164	Ovis aries	Iran
IROA-E6-5351	SAMEA2065596	ERZ017158	Ovis aries	Iran
IROA-E7-5036	SAMEA2065597	ERZ017151	Ovis aries	Iran
IROA-F10-5068	SAMEA2065598	ERZ017149	Ovis aries	Iran
IROA-F3-5142	SAMEA2012931	ERZ017234	Ovis aries	Iran
IROA-F5-5051	SAMEA2012932	ERZ017150	Ovis aries	Iran
IROA-G3-5095	SAMEA2065599	ERZ017153	Ovis aries	Iran
IROA-G4-5205	SAMEA2012926	ERZ017147	Ovis aries	Iran
MOOA-NN-9999	SAMEA2012112	ERZ019218	Ovis aries	Morocco
MOOA-J17-1384	SAMEA2012231	ERZ017297	Ovis aries	Morocco
MOOA-L12-0351	SAMEA2012243	ERZ017306	Ovis aries (H)	Morocco
MOOA-L15-0414	SAMEA2012246	ERZ017309	Ovis aries	Morocco
MOOA-L19-1322	SAMEA2012248	ERZ017313	Ovis aries	Morocco
MOOA-M10-3204	SAMEA2012249	ERZ017315	Ovis aries	Morocco
MOOA-M12-0323	SAMEA1965439	ERZ017317	Ovis aries (H)	Morocco
MOOA-M13-0303	SAMEA2012250	ERZ017318	Ovis aries (H)	Morocco
MOOA-M14-0421	SAMEA2012251	ERZ017319	Ovis aries (H)	Morocco
MOOA-M17-1293	SAMEA2012324	ERZ017322	Ovis aries	Morocco
MOOA-M18-1317	SAMEA2012325	ERZ017323	Ovis aries (P)	Morocco
MOOA-N16-1259	SAMEA2012333	ERZ017330	Ovis aries (P)	Morocco
MOOA-N17-1272	SAMEA2012334	ERZ017331	Ovis aries (P)	Morocco
MOOA-O11-0271	SAMEA2012336	ERZ017333	Ovis aries (H)	Morocco
MOOA-O9-3220	SAMEA2012340	ERZ017339	Ovis aries	Morocco
MOOA-Q12-0163	SAMEA2012431	ERZ017351	Ovis aries (H)	Morocco
MOOA-Q15-1173	SAMEA2012434	ERZ017354	Ovis aries (P)	Morocco
MOOA-Q9-0169	SAMEA2012437	ERZ017356	Ovis aries (H)	Morocco
MOOA-R10-0005	SAMEA1967786	ERZ017357	Ovis aries (H)	Morocco

MOOA-R13-1131	SAMEA2012440	ERZ017361	Ovis aries (P)	Morocco
MOOA-R14-1133	SAMEA2012441	ERZ017362	Ovis aries	Morocco
MOOA-R5-0027	SAMEA2012442	ERZ017364	Ovis aries	Morocco
MOOA-S10-0227	SAMEA2012523	ERZ017369	Ovis aries (H)	Morocco
MOOA-S11-0138	SAMEA2012524	ERZ017370	Ovis aries (H)	Morocco
MOOA-S12-1086	SAMEA2012525	ERZ017371	Ovis aries (P)	Morocco
MOOA-S14-0098	SAMEA2012527	ERZ017373	Ovis aries (P)	Morocco
MOOA-T10-0238	SAMEA2012537	ERZ017383	Ovis aries (H)	Morocco
MOOA-T12-1088	SAMEA2012540	ERZ017386	Ovis aries (P)	Morocco
MOOA-T14-1115	SAMEA2012528	ERZ017387	Ovis aries (P)	Morocco
MOOA-T5-0041	SAMEA2012589	ERZ017389	Ovis aries	Morocco
MOOA-T7-3100	SAMEA2012592	ERZ017392	Ovis aries	Morocco
MOOA-T9-0213	SAMEA2012594	ERZ017396	Ovis aries (H)	Morocco
MOOA-T9-0218	SAMEA2012595	ERZ017397	Ovis aries (H)	Morocco
MOOA-U10-0242	SAMEA2012596	ERZ017398	Ovis aries (H)	Morocco
MOOA-U11-1027	SAMEA2012597	ERZ017399	Ovis aries	Morocco
MOOA-U12-0070	SAMEA2012598	ERZ017400	Ovis aries (P)	Morocco
MOOA-U14-1068	SAMEA2012600	ERZ017401	Ovis aries (P)	Morocco
MOOA-V11-0123	SAMEA2012607	ERZ017410	Ovis aries (P)	Morocco
MOOA-V12-1049	SAMEA2012768	ERZ017411	Ovis aries (P)	Morocco
MOOA-V13-0128	SAMEA2012599	ERZ017412	Ovis aries (P)	Morocco
MOOA-W12-1057	SAMEA2012850	ERZ017421	Ovis aries (P)	Morocco
MOOA-W6-3133	SAMEA2012852	ERZ017423	Ovis aries	Morocco
MOOA-W8-2287	SAMEA2012853	ERZ017425	Ovis aries	Morocco
MOOA-X11-1023	SAMEA2012849	ERZ017428	Ovis aries	Morocco
MOOA-X8-2116	SAMEA2012859	ERZ017431	Ovis aries	Morocco
MOOA-Y5-2077	SAMEA2012915	ERZ017436	Ovis aries	Morocco
MOOA-Z11-2196	SAMEA2012920	ERZ017442	Ovis aries	Morocco
MOOA-Z9-2154	SAMEA2012925	ERZ017447	Ovis aries	Morocco
OARI_AFS33	SAMN01000771		Ovis aries	SW Asia
OARI_AW454	SAMN01000791		Ovis aries	.
OARI_BCS1	SAMN01000755		Ovis aries	Americas
OARI_BMN4	SAMN01000739		Ovis aries	Americas
OARI_BSI4	SAMN01000738		Ovis aries	Americas
OARI_CAS3	SAMN01000756		Ovis aries	SW Europe
OARI_FIN1	SAMN01000784		Ovis aries	N Europe
OARI_GAR4	SAMN01000804		Ovis aries	Asia
OARI_GCN5	SAMN01000806		Ovis aries	Americas
OARI_KRS5	SAMN01000749		Ovis aries	SW Asia
OARI_LAC1	SAMN01000750		Ovis aries	SW Europe
OARI_LAC84	SAMN01000751		Ovis aries	SW Europe
OARI_MER454	SAMN01000752		Ovis aries	SW Europe
OARI_MERC1	SAMN01000768		Ovis aries	SW Europe
OARI_NDZ1	SAMN01000789		Ovis aries	SW Asia
OARI_OJA4	SAMN01000809		Ovis aries	SW Europe
OARI_SALA2	SAMN01000742		Ovis aries	.
OARI_SBF454	SAMN01000744		Ovis aries	N Europe
OARI_SMS2	SAMN01000760		Ovis aries	Central Euro
OARI_SUM2	SAMN01000781		Ovis aries	Asia

The fourth column indicates the species and sheep included in the analysis for *RXFP2* locus signal were denoted by (H) for horned animals and (P) for polled individuals.

CHAPITRE 2: Caractérisation des génomes des caprins locaux au Maroc

CHAPITRE 2: Caractérisation des génomes des caprins locaux au Maroc

Résumé et présentation de l'article

Nous avons discuté dans l'Introduction générale le contexte mondial des ressources génétiques au sein des animaux d'élevage. Nous vivons une expansion des animaux « industriels » qui sont caractérisés par une chute de diversité génétique et qui sont en train de remplacer les populations locales à travers le monde. Cette situation conduit à l'érosion massive de la biodiversité, incluant la perte de traits adaptatifs présents dans les populations indigènes et qui auraient été sélectionnés pendant les 10.000 ans de leur histoire commune avec l'Homme. Cependant, cette biodiversité n'a jamais été évaluée via les données de génomes complets et à grande échelle. Malgré leur substitution par les races « industrielles » dans certaines régions, les caprins du Maroc sont nombreux et ont une diversité morphologique et adaptative très importante. En outre le Maroc, de par sa position géographique, représenterait un point de rencontre de plusieurs flux migratoires. Ses populations caprines représentent ainsi un modèle intéressant pour étudier les populations locales.

Nous avons caractérisé dans ce chapitre la diversité neutre et les signatures de sélection au sein des principales populations locales des chèvres au Maroc (i.e. Noire, Nord et Draa) en partant d'un échantillon de 44 individus issus de localités géographiquement très distantes. A partir des données de génomes complets à une couverture de 12X nous avons étudié le polymorphisme de l'ADN mitochondrial, le niveau de diversité nucléaire global et la structuration génétique, et nous avons caractérisé des signatures de sélection liées à des caractères propres à chaque race/population.

Cette étude montre la forte diversité génétique dans ces populations avec la présence de plus de 24 millions de variants polymorphes dont 1,6 millions de courtes insertion/délétions. Cette forte variation est associée à un très faible déséquilibre de liaison avec une distance qui correspond à r^2=0,2 de 5,4kb sans considérer les variants rares (i.e. fréquence de l'allèle mineur<0,05). Cette diversité est faiblement structurée entre régions et populations (Fst très faibles de 0,001 à 0,004). La population Noire a plus de variants exclusifs (3,7 millions versus 1,9 millions dans la Draa et 1,3 millions dans la population du Nord), mais ceci semble être lié à la taille d'échantillon plus élevée pour cette population. Cette étude a mis en évidence dans chaque population plusieurs signatures de sélection, et pour une grande partie d'entre

elles nous avons pu identifier des gènes candidats. Ces derniers ont permis de caractériser plusieurs processus métaboliques potentiellement impliqués dans les traits spécifiques à chaque population. L'un des processus identifiés nous permet de proposer l'hypothèse d'une adaptation à la chaleur via deux différents mécanismes dans les populations Noire et Draa. La première favoriserait la transpiration et la seconde l'halètement.

Finalement, ce travail qui a été publié dans *Frontiers in Genetics* (Benjelloun et al. 2015) montre la diversité très riche présente au sein des populations locales qui devrait être préservée et gérée d'une façon durable, et ouvre la voie à plusieurs études fonctionnelles en vue de la validation des fonctions potentiellement impliquées dans la différenciation morphologique et adaptative entre les populations de chèvres au Maroc.

Article B: Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (*Capra hircus*) using WGS data

Badr Benjelloun[1,2,3*], Florian J. Alberto[1,2], Ian Streeter[4], Frédéric Boyer[1,2], Eric Coissac[1,2], Sylvie Stucki[5], Mohammed BenBati[3], Mustapha Ibnelbachyr[6], Mouad Chentouf[7], Abdelmajid Bechchari[8], Kevin Leempoel[5], Adriana Alberti[9], Stefan Engelen[9], Abdelkader Chikhi[6], Laura Clarke[4], Paul Flicek[4], Stéphane Joost[5], Pierre Taberlet[1,2], François Pompanon[1,2] and Nextgen Consortium[10]

Published in Frontiers in Genetics 6:107. doi: 10.3389/fgene.2015.00107

[1] Laboratoire d'Ecologie Alpine, Université Grenoble-Alpes, Grenoble, France

[2] Laboratoire d'Ecologie Alpine, Centre National de la Recherche Scientifique, Grenoble, France

[3] National Institute of Agronomic Research (INRA Maroc), Regional Centre of Agronomic Research, Beni-Mellal, Morocco

[4] European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK

[5] Laboratory of Geographic Information Systems (LASIG), School of Civil and Environmental Engineering (ENAC), École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[6] Regional Centre of Agronomic Research Errachidia, National Institute of Agronomic Research (INRA Maroc), Errachidia, Morocco

[7] Regional Centre of Agronomic Research Tangier, National Institute of Agronomic Research (INRA Maroc), Tangier, Morocco

[8] Regional Centre of Agronomic Research Oujda, National Institute of Agronomic Research (INRA Maroc), Oujda, Morocco

[9] Centre National de Séquençage, CEA-Institut de Génomique, Genoscope, Évry, France

[10] NextGen Consortium, http://nextgen.epfl.ch/

* **Correspondence**: Badr Benjelloun (badr.benjelloun@gmail.com; badr.benjelloun@ujf-grenoble.fr)

Abstract

Since the time of their domestication, goats (*Capra hircus*) have evolved in a large variety of locally adapted populations in response to different human and environmental pressures. In the present era, many indigenous populations are threatened with extinction due to their substitution by cosmopolitan breeds, while they might represent highly valuable genomic resources. It is thus crucial to characterize the neutral and adaptive genetic diversity of indigenous populations. A fine characterization of whole genome variation in farm animals is now possible by using new sequencing technologies. We sequenced the complete genome at 12× coverage of 44 goats geographically representative of the three phenotypically distinct indigenous populations in Morocco. The study of mitochondrial genomes showed a high diversity exclusively restricted to the haplogroup A. The 44 nuclear genomes showed a very high diversity (24 million variants) associated with low linkage disequilibrium. The overall genetic diversity was weakly structured according to geography and phenotypes. When looking for signals of positive selection in each population we identified many candidate genes, several of which gave insights into the metabolic pathways or biological processes involved in the adaptation to local conditions (e.g., panting in warm/desert conditions). This study highlights the interest of WGS data to characterize livestock genomic diversity. It illustrates the valuable genetic richness present in indigenous populations that have to be sustainably managed and may represent valuable genetic resources for the long-term preservation of the species.

Key words: *Capra hircus*, WGS, genomic diversity, population genomics, selection signatures, indigenous populations, Morocco

Introduction

Livestock species play a major socio-economic role in the world since they provide many goods and services to human populations. Goats (*Capra hircus*) in particular are one of the more important livestock species, because of their high potential of adaptation to harsh environments. They had a worldwide population of about 1006 million in 2013 (http://faostat3.fao.org/browse/Q/QA/E) and, together with cattle and sheep, they represent the most important source of meat, milk, and skin.

Goats are considered to be the first ungulate to be domesticated, about 10,500 to 9900 years ago near the Fertile Crescent (Zeder, 2005; Naderi et al., 2008). Following human migrations and trade routes, goats rapidly spread over the rest of the world, mainly in Eurasia and Africa (Taberlet et al., 2008; Tresset and Vigne, 2011). During this expansion, they became adapted to different climatic conditions and husbandry practices. In response to these environmental and anthropic selection pressures, a large variety of locally-adapted populations emerged. These populations were managed in a traditional way, i.e., with moderate selection for traits of interest and reproduction allowing important gene flows among them, thus maintaining high levels of phenotypic diversity (Taberlet et al., 2008). However, the rise of the breed concept during mid-1800s (Porter, 2002), and its application to husbandry practices, led to the creation of well-defined breeds. This process aimed at standardizing phenotypic traits mainly associated with morphological aspects (e.g., coat color). Selection of animals for these traits was generally moderated, while crossing among different phenotypes was reduced (Taberlet et al., 2008). More recently, since mid-1900s, industrial breeding has become more widespread, backed by the progress of husbandry practices including the introduction of artificial insemination, embryo transfer, the improvements in feed technology and the use of vaccines and therapeutics against endemic diseases. This has led breeders to progressively substitute the many locally-adapted indigenous breeds for very few highly productive cosmopolitan ones for short-term economic reasons (Taberlet et al., 2008). Thus, FAO in 2013 estimated that 18% of local goat breeds over the world were threatened or already extinct (http://faostat.fao.org/). Consequently, a part of the highly valuable genetic resources captured from the wilds and gradually accumulated over 98% of their common history with humans is now threatened (Taberlet et al., 2008).

Thus, it appears crucial to assess the genetic resources of indigenous populations in order to manage them sustainably and to propose zootechnical approaches that take into account the

preservation of genetic resources. This might be critical in the current context of global environmental changes. To accurately characterize genetic resources, it is necessary to access variation data across the whole genome. This would allow the identification of alleles related to contrasted environmental conditions and those potentially playing an adaptive role. Recent progress in sequencing technologies has opened new perspectives toward the magnitude of genetic analysis that is possible. Sequencing cost and time have dramatically decreased (Snyder et al., 2010) and it is now possible to obtain Whole Genome Sequencing (WGS) data for several dozen individuals, which allows access to variation data sets of the whole genome (Altshuler et al., 2012; Kidd et al., 2012). It is thus possible to combine WGS data and population genomic approaches to characterize neutral and adaptive variation in an unprecedented way. This allows an accurate characterisation of genetic resources and their geographic distribution. The Moroccan territory represents an ideal case-study for evaluating the potential of indigenous breeds for constituting neutral and adaptive genomic resources. Despite the massive introduction of "cosmopolitan" breeds to improve goat milk production in some areas, indigenous populations still represent about 95% of Moroccan goats. This proportion has been continually decreasing and this could lead in a mid-long term to the complete absorption of some indigenous populations by cosmopolitan breeds. In Morocco there are more than 6.2 Million goats (http://faostat3.fao.org/browse/Q/QA/E). Direct anthropic selection was relatively modest and until recently it was difficult to distinguish well-defined breeds. However, several phenotypic groups displaying specific morphological and adaptive characteristics have been identified. They will be referred hereafter here as populations. The three major groups are: (i) the Black goats with three sub-populations that have been recently officially recognized (Atlas, Barcha and Ghazalia), (ii) the Draa population, (iii) and the Northern population. Besides these three main populations/breeds, the major proportion of Moroccan goats presents intermediate phenotypes and non-recognized local populations. The Black population is characterized by its dark color, long hair, a low water turnover and thus good resistance to water stress (Hossainihilali et al., 1993). It presents a good acclimation to various environmental conditions in Morocco (from the Eastern plateaus to Atlas Mountains and the Souss valley more in the South). The Northern population displays some phenotypic similarities with Spanish breeds such as the Murciana-Granadina, Malaguena or Andalusia breeds (Benlekhal and Tazi, 1996). It is bred for milk and meat production although it presents a lower level of production than cosmopolitan industrial dairy breeds (Analla and Serradilla, 1997). It shows a substantial reproductive seasonality related to photoperiod variation (Chentouf et al., 2011). Following an extensive

breeding system, it is the preferred breed to be raised in the harsh mountains of the extreme North of Morocco with oceanic influence and a milder climate. The Draa population is bred in the oasis in Southern Morocco, which is characterized by arid/desert climate conditions. Its water turnover is low compared to European goat breeds studied in similar environments. The Draa goat also has the ability to maintain an unchanged food intake during periods of water deprivation (Hossaini-Hilali and Mouslih, 2002). It displays relatively higher performances of reproduction (i.e., prolificacy, earliness; Ibnelbachyr et al., 2014) and hornless individuals represent about 54.1% of the total (Ibnelbachyr et al., in preparation). In this study, we applied a population genomic framework using WGS data to (i) describe neutral genomic diversity and population structure in the main Moroccan indigenous goat populations (ii) identify potential genomic regions differentially selected among the main populations according to their specific traits. To address these issues, we sequenced at 12× coverage 44 goats representing the Moroccan-wide geographic diversity of the three main goat indigenous populations in the country.

Material and Methods

Sampling

Sample collection was performed in a wide part of Morocco [~400,000 km^2; Northern part of Morocco in latitude range (28°–36°)]. A total of 44 individuals unambiguously assigned to one of the three main indigenous populations (i.e., Black, Draa and Northern) were sampled (Table S1) in a way that maximized individuals' spread over the sampling area. This resulted in sampling spatially distant unrelated individuals, ensuring a spatial representativeness of all regions (Figure 1). For each individual, tissue samples were collected from the distal part of the ear and placed in alcohol for 1 day, and then transferred to a silica-gel tube until DNA extraction.

Production of WGS Data

DNA extractions were done using the Puregene Tissue Kit from Qiagen® following the manufacturer's instructions. Then, 500 ng of DNA were sheared to a 150–700 bp range using the Covaris® E210 instrument (Covaris, Inc., USA). Sheared DNA was used for Illumina® library preparation by a semi-automatized protocol. Briefly, end repair, A tailing and Illumina® compatible adaptors (BiooScientific) ligation were performed using the SPRIWorks

Library Preparation System and SPRI TE instrument (Beckmann Coulter), according to the manufacturer protocol. A 300–600 bp size selection was applied in order to recover the most of fragments. DNA fragments were amplified by 12 cycles PCR using Platinum Pfx Taq Polymerase Kit (Life® Technologies) and Illumina® adapter-specific primers. Libraries were purified with 0.8× AMPure XP beads (Beckmann Coulter). After library profile analysis by Agilent 2100 Bioanalyzer (Agilent® Technologies, USA) and qPCR quantification, the libraries were sequenced using 100 base-length read chemistry in paired-end flow cell on the Illumina HiSeq2000 (Illumina®, USA).

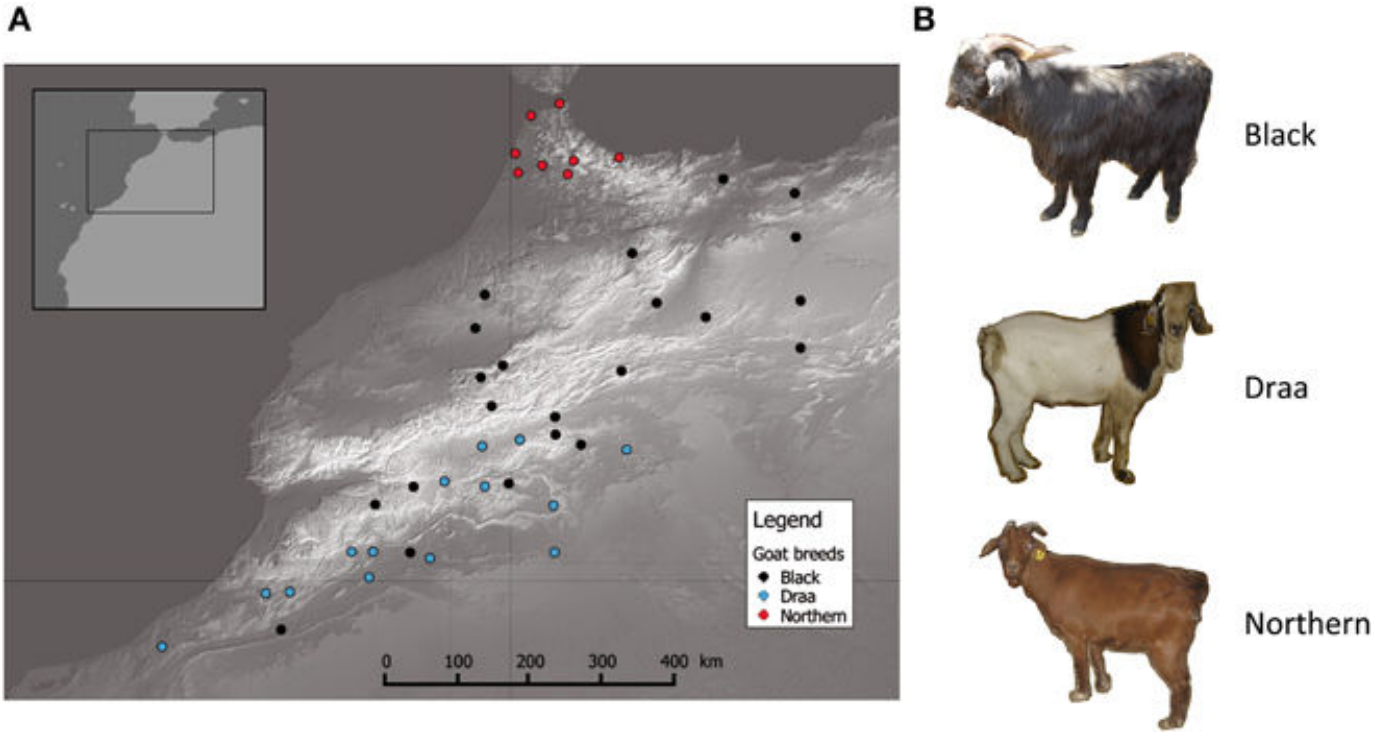

Figure 1. Distribution of goats sampled.

(A) Geographic map showing the distribution of the 44 goats sampled in this study. Each point represents one individual and different colors illustrate different populations. **(B)** Striking phenotypic differences between the 3 main goat populations in Morocco.

WGS Data Processing

Paired-end reads were mapped to the goat reference genome (CHIR v1.0, GenBank assembly GCA_000317765.1) (Dong et al., 2013) using BWA mem (Li and Durbin, 2009). The BAM files produced were then sorted using Picard SortSam and improved using Picard MarkDuplicates (http://picard.sourceforge.net), GATK RealignerTargetCreator, GATK IndelRealigner (Depristo et al., 2011), and Samtools calmd (Li et al., 2009). Variant calling

was done using three different algorithms: Samtools mpileup (Li et al., 2009), GATK UnifiedGenotyper (McKenna et al., 2010), and Freebayes (Garrison and Marth, 2012).

There were two successive rounds of filtering variant sites. Filtering stage 1 merged together calls from the three algorithms, whilst filtering out the lowest-confidence calls. A variant site passed if it was called by at least two different calling algorithms with variant phred-scaled quality >30. An alternate allele at a site passed if it was called by any one of the calling algorithms, and the genotype count >0. Filtering stage 2 used Variant Quality Score Recalibration by GATK. First, we generated a training set of the highest-confidence variant sites where (i) the site is called by all three variant callers with variant phred-scaled quality >100, (ii) the site is biallelic (iii) the minor allele count is at least 3 while counting only samples with genotype phred-scaled quality >30. The training set was used to build a Gaussian model using the tool GATK VariantRecalibrator using the following variant annotations from UnifiedGenotyper: QD, HaplotypeScore, MQRankSum, ReadPosRankSum, FS, DP, InbreedingCoefficient. The Gaussian model was applied to the full data set, generating a VQSLOD (log odds ratio of being a true variant). Sites were filtered out if VQSLOD < cutoff value. The cutoff value was set for each population by the following: Minimum VQSLOD = {the median value of VQSLOD for training set variants} – 3* {the median absolute deviation VQSLOD of training set variants}. Measures of the transition / transversion ratio of SNPs suggest that this chosen cutoff criterion gives the best balance between selectivity and sensitivity. Genotypes were improved and phased by Beagle 4 (Browning and Browning, 2013), and then filtered out where the genotype probability calculated by Beagle is less than 0.95.

The whole mitochondrial genome (mtDNA) was assembled from a subset of random 20,000,000 reads using the ORGASM tool (Coissac, unpublished). We then extracted the sequence of the HVI segment of the control region for each individual in order to compare with the haplogroup references discovered worldwide (see below).

Population Genomic Analyses

Characterisation of mtDNA Diversity

The number of polymorphic sites and the number of haplotypes were calculated from the whole mitochondrial sequences using DNAsp (Librado and Rozas, 2009). We also calculated these parameters for the hyper variable segment (HVI) of the control region, for which 22 reference sequences representing the diversity of the 6 haplogroups found over the world

were available (Naderi et al., 2007). We were interested in the level of resolution of the HVI segment to discriminate the different haplotypes compared to the whole mitochondrion.

Then, using the sequences corresponding to the HVI segment for our dataset and the reference sequences, we drew a network of the haplotypes to identify the different haplogroups present in our dataset. The best evolutionary model was determined using jModelTest v 2.1.4 (Darriba et al., 2012). A median joining network representing the relationships between haplotypes was drawn using SplitsTree4 (Huson and Bryant, 2006).

Characterisation of Neutral Nuclear Diversity

Neutral nuclear genomic variations were characterized to evaluate the level of genetic diversity present in Morocco and within populations. The total number of variants and the number of variants within each population were calculated. Allele frequencies and the percentage of exclusive variants (i.e., variants polymorphic in only one population) were estimated at the population scale using the Perl module vcf-compare of Vcftools (Danecek et al., 2011). The level of nucleotide diversity (π) was calculated within each population and averaged over all of the biallelic and fully diploid variants for which all individuals had a called genotype. The observed percentage of heterozygote genotypes per individual (*Ho*) was calculated considering only the biallelic SNPs with no missing genotype calls. From *Ho*, the inbreeding coefficients (*F*) were calculated for each individual using population allelic frequencies over all 44 individuals. The relatedness among individuals was assessed using the pairwise identity-by-state (*IBS*) distances calculated as the average proportion of alleles shared using Vcftools.

Pairwise linkage disequilibrium (LD) was assessed through the correlation coefficient (r^2). It was estimated in 5 segments of 2 Mb on different chromosomes (physical positions between 5 and 7 Mb on chromosomes 6, 11, 16, 21, and 26). LD was estimated either by using the whole set of reliable variants or after discarding rare variants with a minor allele frequency (MAF) less than 0.05. For both estimations, we calculated r^2 values between all pairs of bi-allelic variants (SNPs and indels) on the same segment using Vcftools. Inter-SNP distances (kb) were binned into the following 7 classes: 0–0.2, 0.2–1, 1–2, 2–10, 10–30, 30–60, and 60–120 kb and observed pairwise *LD* was averaged for each inter-SNP distance class and used to draw *LD* decay. Due to the insufficient number of individuals per population we made these estimations for the whole set of individuals without considering each population individually.

Genetic structure was assessed using three different methods: (i) a principal component analysis (PCA) was done using an *LD* pruned subset of bi-allelic SNPs. *LD* between SNPs in windows containing 50 markers was calculated before removing one SNP from each pair where *LD* exceeded 0.95. Subsequently, only 12,543,534 SNPs among a total of 22,304,702 bi-allelic SNPs were kept for this analysis. The R package adegenet v1.3-1 (Jombart and Ahmed, 2011) was used to run PCA and Plink v1.90a (https://www.cog-genomics.org/plink2) was used for *LD* pruning. (ii) An analysis with the clustering method sNMF (Frichot et al., 2014) was carried-out. This method was specifically developed to analyse large genomic datasets in a fast, efficient and reliable way. It is based on sparse non-negative matrix factorization to estimate admixture coefficients of individuals. All biallelic variants were used and five runs for each *K* value from 1 to 10 were performed using a value of *alpha* parameter of 8. For each run, the cross-entropy criterion was calculated with 5% missing data to identify the most likely number of clusters. The run showing the lowest cross-entropy value for a given K was considered. (iii) Finally, the *Fst* index was estimated according to Weir and Cockerham (1984) for each polymorphic site and then weighted to obtain one value over the whole genome. The overall *Fst* between the three groups and the population pairwise values were calculated using Vcftools.

Detection of Selection Signatures

A genome scan approach was performed using the XP-CLR method (Chen et al., 2010) to identify potential regions differentially selected among the three populations. It is a likelihood method for detecting selective sweeps that involves jointly modeling the multi-locus allele frequency differentiation between two populations. This method is robust to detect selective sweeps and especially with regards to the uncertainty in the estimation of local recombination rate (Chen et al., 2010). Due to the absence of genomic position, the physical position (1 Mb ≈ 1 cM) was used. An in-house script based on overlapped segments of a maximum of 27 cM was designed to estimate and assemble XP-CLR scores using the whole set of bi-allelic variants. Overlapping regions of 2 cM were applied and the scores related to the extreme 1 cM were discarded, except at the starting and the end of chromosomes on the CHIR v1.0 assembly. XP-CLR scores were calculated using grid points spaced by 2500 bp with a maximum of 250 variants in a window of 0.5 cM and by down-weighting contributions of highly correlated variants ($r^2 > 0.95$) in the reference group.

To equilibrate the number of individuals per population, only 14 Black goats were randomly sampled among the 22. They were included with the 14 Draa and the 8 Northern individuals.

Each population was tested using a reference group including individuals from the two other populations. The 0.1% genomic regions with highest XP-CLR scores revealed by the analysis were identified and lists of genes partially or fully covered by these regions were then established. To ensure the coverage of short genes (i.e., genes shorter than the distance between adjacent grid points), two segments of 1500 bp each surrounding both sides of genes were also considered. NCBI databases were used to identify coordinates of the 20700 annotated autosomal genes on the CHIR v1.0 genome assembly (http://www.ncbi.nlm.nih.gov/genome?term=capra%20hircus).

Gene Ontology Enrichment Analyses

To explore the biological processes in which the top candidate genes are involved, Gene Ontology (GO) enrichment analyses were performed using the application GOrilla (Eden et al., 2009). The 12,669 goat genes associated with a GO term were used as background reference. Significance for each individual GO-identifier was assessed with *P*-values that were corrected using FDR *q*-value according to the Benjamini and Hochberg (1995) method. GO terms identified in each population were clustered into homogenous groups using REVIGO (Supek et al., 2011). Medium similarity among GO terms in a group was applied and the weight of each GO term was assessed by its *p*-value.

Results

Phylogeny of mtDNA Genomes

The whole mitochondrial genome was assembled successfully for 41 individuals and represented 16,651 bp length sequences. A total of 239 polymorphic sites were detected, which allowed discriminating 41 haplotypes. In an alternative complementary approach, the 481 bp length sequenced of the HVI segment of the control region was extracted, and this revealed 64 polymorphic sites identifying 40 single haplotypes. We constructed a network using the GTK + G + I model, which showed the best likelihood. The network (Figure 2) including the 22 reference haplotypes (i.e., haplogroups A, B, C, D, F, and G; Naderi et al., 2007) showed that the 40 haplotypes all belonged to the haplogroup A. We did not detect any coherent pattern of geographic structure among the haplotypes. There was also no clear differentiation of the haplotypes according to the three considered populations.

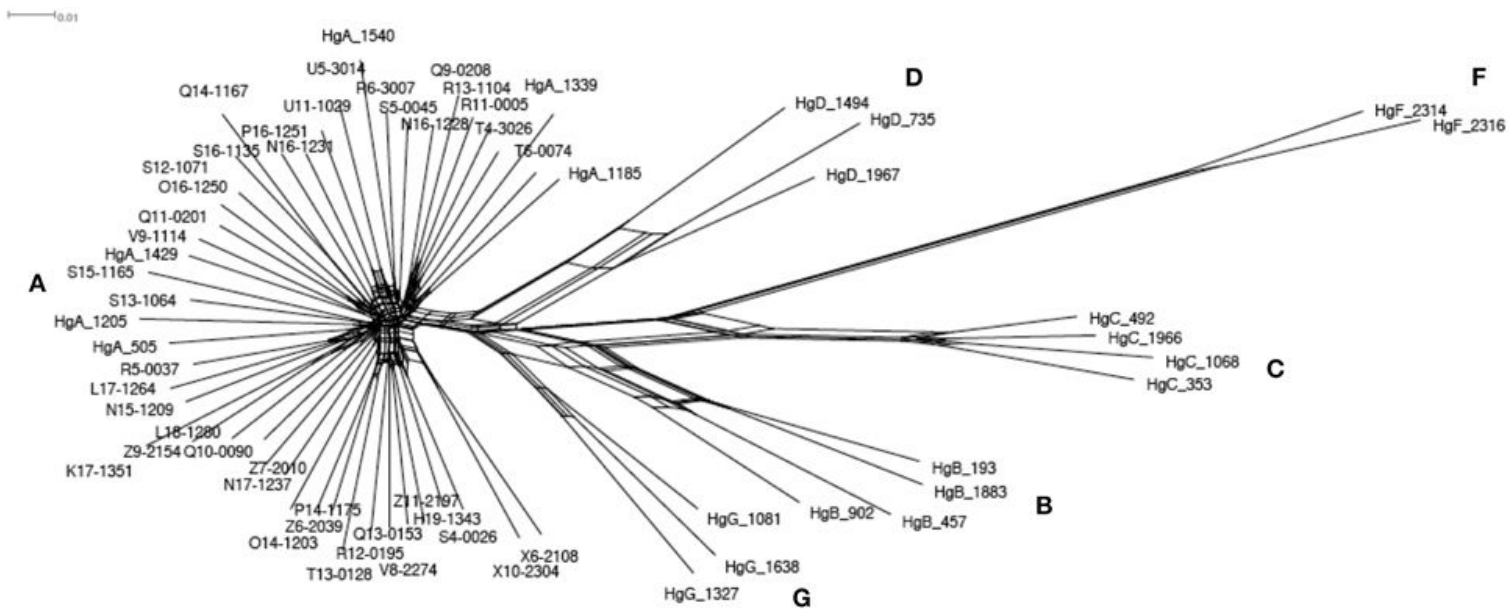

Figure 2. Phylogenetic network based on the mitochondrial HVI segment of the control region.

Sequences of 41 Moroccan goats and the 22 references representing the worldwide diversity (Naderi et al., 2007) were used. The 22 reference identifiers start with « Hg » and the following letter indicate which haplogroup each belongs to. The other identifiers correspond to the Moroccan goats. The red letters give the names of the 6 haplogroups.

Neutral Diversity from WGS Data

The whole nuclear genomes were assembled on the goat reference genome CHIR1.0 along the 30 chromosomes. We mapped unambiguously 99.0% (±0.1%) of reads to the CHIR v1.0 assembly. However, the mapped reads properly paired constituted 90.3% (±0.1%) of reads in average. After the filtering processes, a total of 24,022,850 variants were found to be polymorphic in the total dataset among which 22,396,750 were SNPs and 1,626,100 were small indels. There were a total of 15,948,529 transitions and 6,540,478 transversions leading to a ts/tv ratio of 2.44. Due to differences in quality among individuals, the number of variants called per individual was at least 23,273,239 and 24,003,837 on average. As a consequence, a total of 23,059,968 variants showed no missing genotype over the 44 samples, among which 22,963,257 were biallelic.

Among the 24,022,850 polymorphic variants, only 12,024,778 variants were polymorphic within each of the three populations. The remaining variants were either polymorphic in only one or in two populations. When considering variants exclusive to each population, 3,704,299 were found polymorphic only in the Black population (n = 22), 1,887,724 only in the Draa

population (n = 14) and 1,305,561 only in the Northern population (n = 8) (Figure 3). Rare variants (MAF < 0.05) represented a total of 10,892,203 (45.3%).

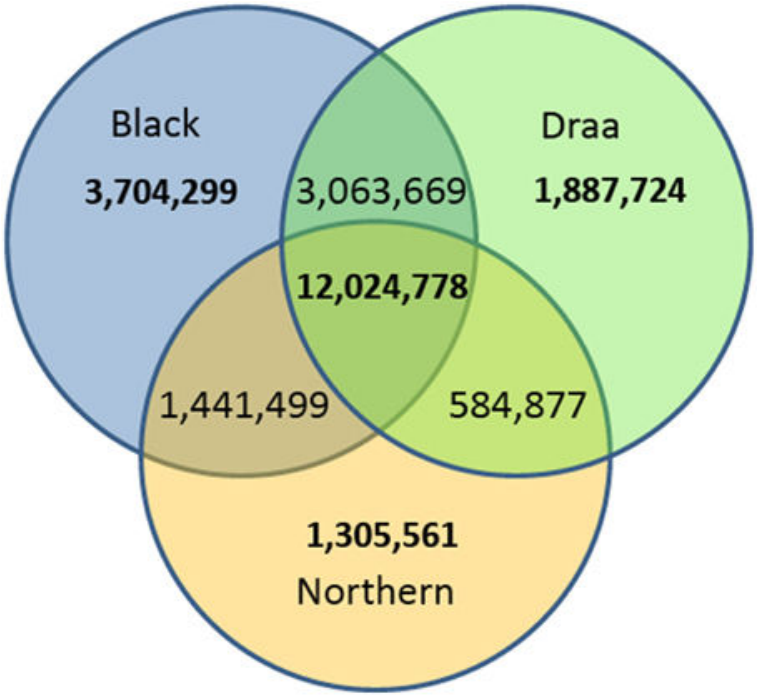

Figure 3. Venn diagram of the number of polymorphic variants in the three Moroccan goat populations.

Considering the 44 goats together, the average nucleotide diversity (π) calculated from 22,963,257 biallelic variants without missing genotype calls was 0.180. The Draa and the Black populations displayed similar π values amounting to 0.180 and 0.181 respectively. Among the 8 individuals representing the Northern population, π was slightly higher, amounting to 0.189. The observed percentage of heterozygote genotypes per individual (*Ho*) was 17.2% on average, ranging from 12.1% to 18.4%. The average inbreeding coefficient (F) was globally rather low (0.05 ± 0.07) and values were evenly distributed among populations. Similar average values were obtained for the Northern and Black populations (respectively 0.04 ± 0.07 and 0.04 ± 0.05). The Draa goats were slightly more inbred (average F = 0.07 ± 0.09), particularly due to one individual showing F = 0.32.

We assessed LD by calculating the pairwise r^2 values between polymorphic sites for five chromosome regions. When withdrawing rare variants (i.e., MAF < 0.05), the average r^2 value was 0.40 for the first bin (0–0.2 kb) and decayed to less than 0.20 in 5.4 kb (Figure 4). Using the whole set of reliable variants, the average r^2 was 0.21 for the first bin and decreased

rapidly to less than 0.20 in 239 bp of distance. Moreover, it decayed to less than 0.15 in about 1.33 kb distance (Figure S2).

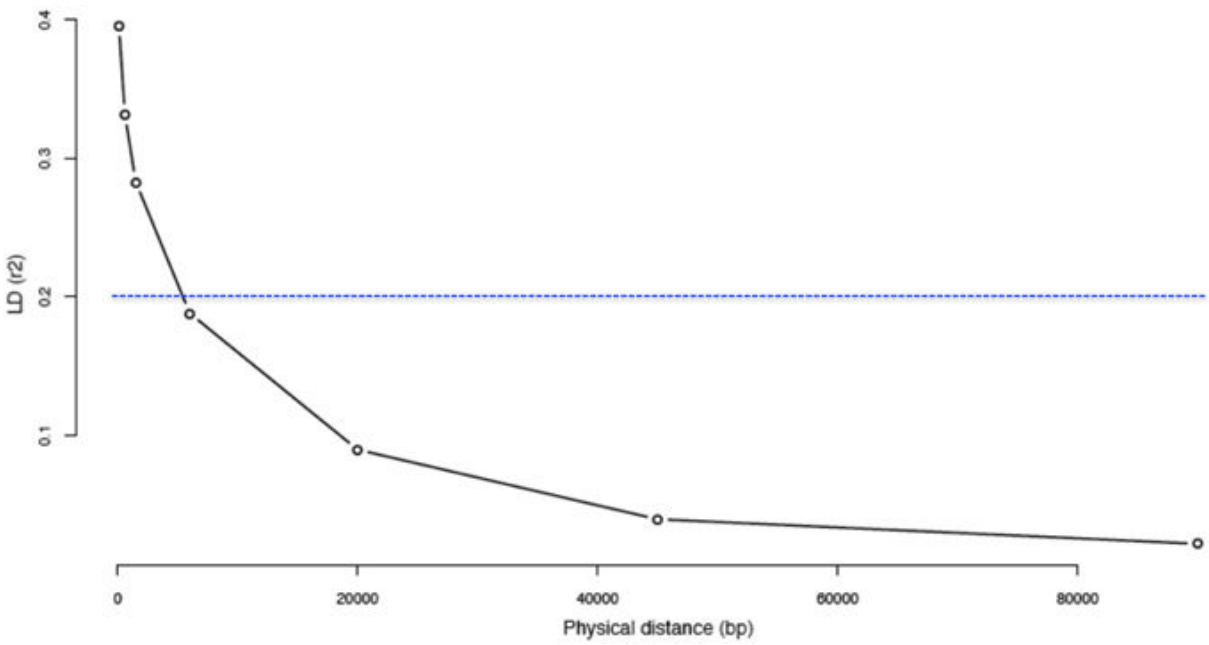

Figure 4. Decay of linkage disequilibrium (r^2) as a function of physical distance by excluding "rare" variants.

The Linkage Disequilibrium (*LD*) was calculated for the 44 Moroccan goats on 5 different segments of 2 Mb each on 5 different chromosomes. Inter-variant distances (bp) were binned and averaged into the classes: 0–0.2, 0.2–1, 1–2, 2–10, 10–30, 30–60, and 60–120 kb.

Among the three populations, the level of genetic differentiation over the whole nuclear genome was extremely low (*Fst* = 0.0024). The pairwise *Fst* values varied from 0.001 for the Black-Draa comparison to 0.004 for the Northern-Draa comparison. Between the Black and Northern populations the pairwise *Fst* was 0.003.

The PCA analysis showed a very low population structure in the 44 Moroccan goats. The 3 main principal components (PCs) explained 5.8% of variance. The first PC tended to distinguish the Northern and Draa populations while the Black populations formed an in-between group. The second PC acted predominantly to distinguish individuals within the Draa and the Northern populations (Figure S1).

The clustering analysis of the genetic structure using sNMF (Frichot et al., 2014) showed that the 44 Moroccan goats belonging to the three populations were more likely represented by only one cluster according to the "crossentropy" criterion (lower values for $K = 1$). However,

this criterion is not straightforward and when increasing until $K = 3$ we observed a weak pattern of genetic structure (Figure 5). At $K = 2$, the Northern goats were all strongly assigned to one distinct cluster. The second cluster was characterized by high assignment from the Draa population, except for two individuals that belong to the same cluster as the Northern goats. Finally, the Black goats showed variable levels of admixture between the two clusters (Figure 5A). When mapping the assignment results on a map we observed a geographic pattern with one cluster represented mainly in the north of Morocco (red component; Figure 5B) and the second cluster more present in the south (Figure 5B). At $K = 3$, the additional cluster was mostly represented in the Black goats which are located in the center of the sampling area (Figure 5A). The two other clusters still mostly represented the separation of Northern and Draa populations but the pattern was less evident. It was difficult to disentangle the relationship of genetic structure with populations and geography because the two factors were confounding.

Selection Signatures

We applied the XP-CLR genome scan method (Chen et al., 2010) on the whole genomes of 36 goats from the three phenotypic populations (14 Black, 14 Draa, and 8 Northern). We identified selective sweep genes in each population considering the top 0.1% genome-wide scores. Our approach highlighted respectively 142, 167, and 176 candidate genes in the Black, Draa, and Northern populations. The region showing the strongest XP-CLR score was located on chromosome 6 for the Black goats (Figure S3) and on chromosome 22 for the Northern goats (Figure S4), but they did not match any annotated gene. The annotated genes showing the strongest selective sweeps were *HTT, MSANTD1*, and *LOC102170765* in the Black goats, and *FOXP2, TRAP1* and *DNASE1* in the Northern goats (Table 1). In the Draa population, the highest XP-CLR scores corresponded to *LOC102190531, ADD3*, and *ASIP* genes (Figure 6). The enrichment categories of the identified candidate genes in the Black goats were associated with 15 GO terms (Table S2). They clustered into the following four differentiated categories by REVIGO (Supek et al., 2011): tube development, calcium ion transmembrane import into mitochondrion, negative regulation of transcription from RNA polymerase II promoter during mitosis and response to fatty acid. The enrichment of the identified candidate genes in Draa goats highlighted the significance of 25 GO terms (Table S3) clustering into five differentiated categories: regulation of respiratory gaseous exchange, behavior, postsynaptic membrane organization, protein localization to synapse, and neuron cell-cell

adhesion. In the Northern goats, we did not find significant enrichment categories for the candidate genes identified.

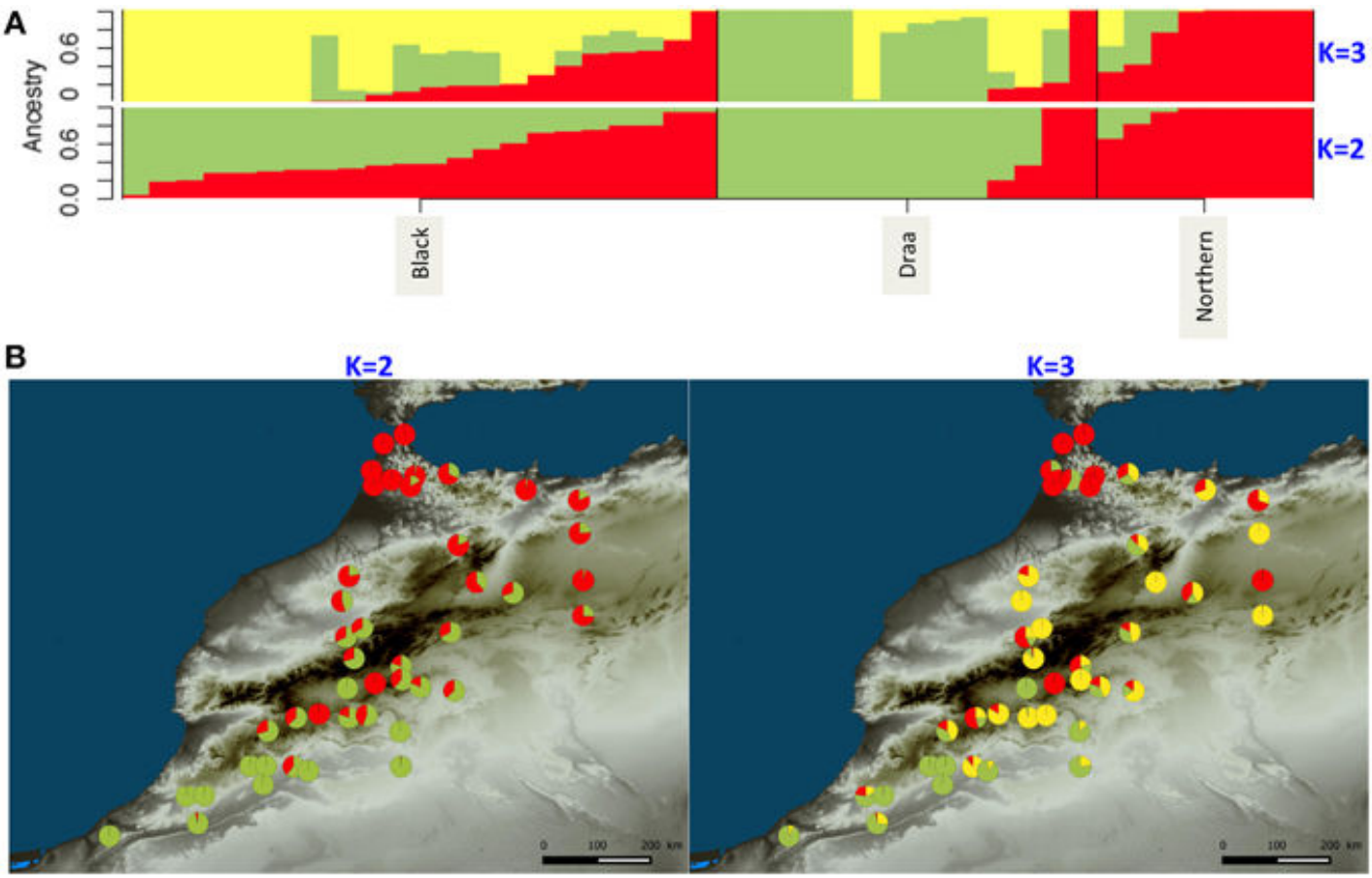

Figure 5. WGS ancestry estimates for Moroccan goats for $K = 2$ and $K = 3$ clusters.

(A) Each bar represents one individual. Different colors illustrate the assignment proportion (Q score) to each one of the assumed clusters. **(B)** Geographical distribution of individual Q-score values.

Table 1. Top-20 candidate genes under positive selection in each Moroccan goat population using the top-0.1% XP-CLR scores autosomal-wide cut-off level.

Black population					Draa					Northern population				
Gene	Chr	Number of top-scores	Distance/ grid point	Higher score	Gene	Chr	Number of top-scores	Distance/ grid point	Higher score	Gene	Chr	Number of top-scores	Distance/ grid point	Higher score
HTT	6	29	4739	82.6	*LOC102190531*	13	9	2493	94.0	*FOXP2*	4	14	33163	48.7
MSANTD1	6	3	5501	61.4	*ADD3*	26	24	5409	74.4	*TRAP1*	25	5	3977	42.8
LOC102170765	6	1	699	54.0	*ASIP*	13	2	995	71.3	*DNASE1*	25	4	2485	41.8
FAM160B1	26	2	42409	45.6	*VPS13B*	14	36	21697	70.1	*FAM227B*	10	9	25094	39.8
STRIP1	3	5	3069	43.6	*RALY*	13	9	5294	66.1	*CREBBP*	25	14	9497	38.6
NDUFA6	5	4	2786	41.8	*ICAM3*	7	5	1696	62.4	*PAPSS2*	26	1	43841	35.7
HNRNPA3	2	3	1183	40.3	*HIVEP2*	9	15	6353	61.4	*SLX4*	25	2	9472	32.4
KITLG	5	7	15223	39.7	*GGH*	14	10	2984	59.3	*PGM5*	8	9	23504	32.0
ALX3	3	1	7499	39.6	*PLSCR3*	19	2	1770	58.2	*BCAS3*	19	11	53928	31.4
IFT88	12	13	4139	39.6	*SOX6*	15	17	28387	54.8	*GALK2*	10	3	51372	31.2
XPO4	12	25	3141	39.5	*JARID2*	23	27	8506	52.9	*MAB21L1*	12	2	1213	31.0
VPS13B	14	16	48818	38.0	*NOL4*	24	6	78026	49.7	*NBEA*	12	31	21437	30.9
LOC102183160	14	1	298	37.5	*TIMP3*	5	3	4339	49.4	*LOC102182654*	25	1	2127	30.8
FLI1	29	6	10103	36.9	*EIF2S2*	13	2	7340	48.8	*LCOR*	26	7	8873	30.6
C4H7orf10	4	11	70095	35.7	*TTC39C*	24	10	10009	48.2	*RANBP10*	18	10	5817	29.5
TTC21A	22	3	11417	35.6	*PCBP3*	1	2	103547	46.0	*SLC12A4*	18	2	10049	28.0
LATS2	12	4	6149	34.3	*TTPA*	14	7	8967	45.0	*ROR1*	3	5	39379	27.8
NSMCE2	14	5	46323	33.7	*ASTN2*	8	11	79003	43.1	*MRPL54*	7	1	2005	27.7
ATG2B	21	3	25060	33.4	*GALNT7*	8	10	10949	42.2	*PDE1A*	2	2	146650	27.4
FAT2	7	2	42599	33.4	*MUC13*	1	7	3341	41.7	*KRT8*	5	2	3717	27.3

Coordinates of 20700 autosomal genes on the CHIR v1.0 goat assembly were used to identify candidate genes matching XP-CLR top scores. Genes were ranked according to the higher XP-CLR score. Chr, Chromosome. Number of top-scores, Number of grid points among the top-0.1% XP-CLR scores matching the gene. Distance/grid point: gene length/number of top-scores. Grid points in XP-CLR analysis were separated by 2.5 Kb.

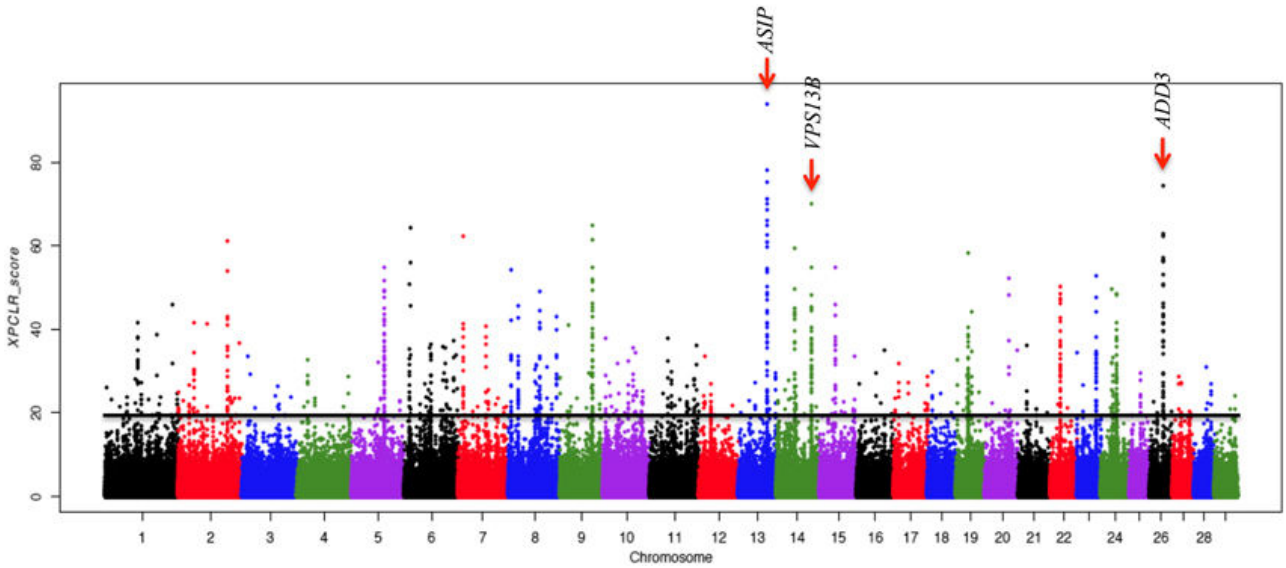

Figure 6. Plot of XP-CLR scores along autosomes in selective sweep analysis for the Draa goat population.

The horizontal line indicates a 0.1% autosomal-wide cut-off level. Red arrows and names indicate the top three candidate genes.

Discussion

Indigenous/traditional goats have been raised for a long time for various purposes and they have gradually accumulated several traits making them well adapted to their environments. The mechanisms underlying these adaptive traits have been poorly studied until now. The recent development of sequencing technologies has now made possible the sequencing of individuals' whole genomes and this may greatly expand our understanding of genomic diversity. Except for a few studies based on medium density SNP panels (about 50,000 SNPs) (Kijas et al., 2013; Tosser-Klopp et al., 2014), previous population genetic studies on goats have been limited to just a few dozens of markers (i.e., microsatellites). In this study we used variants spanning the whole genome to characterize indigenous goat populations of Morocco.

Mitochondrial Variation

Complete mitochondrial sequences were successfully assembled from a low portion of reads for 41 individuals. In terms of its ability to discriminate between the different haplotypes, the 481 bp length of the HVI segment of the control region was almost as accurate as the whole mitochondrion sequence of 16,651 bp length from which it was extracted. Only a small difference in the total number of haplotypes defined was found (41 against 40 haplotypes

respectively). This result shows that despite a low number of variable sites, the dense variability found in the control region (26.8% of the total number of variants for only 2.8% of the sequence length) concentrated most of the phylogenetic information. Thus, the HVI segment of the control region seems to be a good surrogate of the whole mitochondrial polymorphism. This study confirmed previous results based on the HVI segment of the control region (Pereira et al., 2009; Benjelloun et al., 2011) where Moroccan domestic goats showed only haplotypes from the A haplogroup (HgA). In a larger study using 2430 samples with a worldwide distribution, Naderi et al. (2007) found that most of the domestic goats displayed HgA (about 94%). Thus, it seems that the mitochondrial categorization in Morocco is rather representative of the rest of the world, even if the remaining haplogroups were not identified in our sampling. Besides this, the mtDNA diversity was weakly structured according to geography, as already reported by (Benjelloun et al., 2011) on the HVI region.

We did not find any clear structure of the mitochondrial haplotypes among the three populations. The high mitochondrial diversity characterizing these three populations probably indicates the diversity present in the first domesticated goats that arrived in Morocco and/or recurrent gene flows from diverse origins. According to (Pereira et al., 2009), Moroccan goat populations would have been established via two main colonization routes, one a North African land route and the other a Mediterranean maritime route across the Strait of Gibraltar. The high gene flows between populations, mediated by humans, would be ultimately responsible for the absence of structure across Morocco.

Nuclear Neutral Variation

Although the low percentage of the properly paired mapped reads (about 10%) in comparison with the percentage of mapped reads (about 99%) would illustrate a possible fragmentation of the genome assembly used, we identified many high confidence variants (approximately 24 million among which 6.8% were small indels) over the whole nuclear genomes of the 44 Moroccan goats studied. This is much higher than was found in all previous studies detecting variants in large sample cohorts from whole genome sequencing. For example, the human 1000 Genomes Project (Altshuler et al., 2012) detected approximately 15 million SNPs and 1 million short indels, while in the 1001 Genomes Project of Arabidopsis thaliana about 5 million SNPs and 81,000 small indels were found (Cao et al., 2011). The polymorphism detected in the Moroccan goats remains huge even when considered in proportion to the genome size of the species.

This huge number of variants did not show a strong genetic structure either among populations or over geographic space. The globally weak genetic structure suggests that extensive gene flows along with low level of selection have produced this pattern. Our findings contrast with most previous studies, which generally show a clear structure among goat breeds or populations (Cañon et al., 2006; Agha et al., 2008; Serrano et al., 2009; Di et al., 2011; Hassen et al., 2012; Kijas et al., 2013). Several reasons could explain this difference. First, most of the previous studies used microsatellite markers exhibiting high mutation rate. Thus, compared to SNP markers, microsatellites could more likely show imprints of recent demographic events such as differentiation between recently isolated populations. Moreover, the microsatellite markers generally used (Serrano et al., 2009; Di et al., 2011) were recommended by FAO and designed to exacerbate genetic differentiation among breeds, which was thus artificially inflated. In a more recent study, (Kijas et al., 2013) used a panel of SNP markers from a chip designed with animals representing industrial breeds for the SNP discovery (Tosser-Klopp et al., 2014). In that case the results were certainly inflated by the ascertainment bias due to the chip design. However, it is also likely that in our case the demographic history of Moroccan goats differs from that of the breeds previously studied, and in particular from the ones compared at larger geographic scales such as Europe and Middle East (Cañon et al., 2006), or China, Iran and Africa (Di et al., 2011). The structured diversity found in these latter two studies would result from the strong isolation between countries. However, even at smaller scales the selection pressures exerted by breeding processes and husbandry practices may have increased isolation among breeds, and thus reinforced population differentiation compared to Morocco. The situation found in Morocco is close to the one described by Hassen et al. (2012) for six Ethiopian goat ecotypes, where even with microsatellite markers most of the diversity was found within populations, showing low levels of genetic differentiation. This result was explained by the existence of uncontrolled breeding strategies and agricultural extensive systems. In Morocco, it seems that goat populations have experienced moderate levels of selection and that most of the genetic diversity has been preserved during the breeding process which led to the three phenotypic populations. However, a weak genetic pattern was revealed by sNMF, which seems to be partially related to populations as well as geography. When mapping the clustering results (for $K = 3$, Figure 5B), a pattern appeared across Morocco, with Northern goats displaying a higher assignment probability to one distinct cluster. The Northern population is observably slightly more diverse than the others for which higher numbers of individuals were studied. This higher diversity and the slightly higher genetic differentiation of the Northern goats

support the hypothesis of an influence of Iberian gene flows through the strait of Gibraltar in the North of Morocco (Analla and Serradilla, 1997).

The goal of our study was not to visualize the *LD* variations along chromosomes by covering all regions including centromeres and chromosomal inversions that are reportedly characterized by an elevated *LD* (Weetman et al., 2010; Marsden et al., 2014). Rather, we aimed to generate a global representation of *LD* across the genome by covering segments of 2 Mb in 5 different chromosomes taking all the reliable variants found from WGS data. Furthermore, knowing the effect of rare variants on *LD* estimation (Andolfatto and Przeworski, 2001) and to compare our findings with previous studies, we also estimated *LD* after discarding rare variants (MAF < 0.05). The extent of LD reported without rare variants ($r^2 < 0.20$ after 5.4 kb on average) is clearly shorter compared to all previous studies on farm animals, where it largely exceeds 10 kb for $r^2 = 0.20$ (Meadows et al., 2008; Villa-Angulo et al., 2009; Wade et al., 2009; McCue et al., 2012; Ai et al., 2013; Veroneze et al., 2013). In these studies, whole genome variants were not available and potential biases due to the use of SNP chips may partially explain the results. However, we consider that our finding would mainly result from the extensive breeding system favoring high gene flows among Moroccan goat populations/herds and low inbreeding and from the absence until now of strong selection during the breeding processes. Results on *LD* and genetic variability illustrate the important diversity present in indigenous populations in comparison with industrial breeds on which previous studies mainly focussed (e.g., Meadows et al., 2008; Villa-Angulo et al., 2009). This should be considered in the establishment of future programs aimed at improving these populations to preserve this highly valuable genetic diversity.

Beside this, when using the whole set of reliable variants we found a much lower *LD* ($r^{20}_{0.20}$ = 239 bp). We do believe that this value should be considered in genome wide association and genome scan studies. Indeed most of studies remove rare variants for genotyping quality issues. In our case, the quality filtering produced reliable rare variants (about 45%) that would give a more realistic estimation of LD. To our knowledge, very few studies included rare variants to estimate *LD* (e.g., Mackay et al., 2012).

Selection Signatures in Moroccan Goat Populations

The weakly structured genetic diversity in Moroccan goats was suitable to detect selection signatures, avoiding possible false positives potentially generated by genetic structure. Despite a common genomic background and this weak population structure in Moroccan

goats, the three main populations have been bred in various conditions and thereby have been subject to different anthropic and environmental selections in their recent history. As a result, they differ in their physiology, behavior and morphology. The observation of rapid phenotypic changes raises the question of the underlying genetic changes that would be shaped by selection. We identified numerous signatures of selection corresponding to genomic regions potentially under selection in each population.

A difficulty in identifying the genes or metabolic pathways under selection resides in the currently incomplete annotation of the goat genome. The stronger selective sweeps corresponded to regions in the Black population (chromosome 6) and in the Northern population (chromosome 22) matching un-annotated genes on the CHIR v1.0 assembly. This is probably due to either the incomplete annotation of the caprine genome or the fact that the selected functional mutations within each of these regions are not located within or close to a protein-coding gene. The incomplete genome annotation prevented us from identifying several known selected genes among Moroccan goat populations. For example, the *melanocortin-1 receptor* (*MC1R*) gene that is reported to be involved in coat color differentiation in goats (e.g., Fontanesi et al., 2009a) is not associated to any chromosome on the CHIR v1.0 assembly. Therefore, we were not able to detect its possible associated signal of selection in populations where the coat color is fixed knowing that we looked for selection signatures on autosomes only. Another problem consisted in the presence of several annotated genes that were not identified (i.e., no known orthologs, gene identifier starting with "LOC"). Thus, many genes potentially under selection could not be used in our GO enrichment analyses (e.g., the higher-score candidate gene in Draa population on Chromosome 13; Table 1). Despite these restrictions, we identified several sets of strong candidate genes in the three studied populations.

In the Black population the top-ranked candidate gene identified was *huntingtin* (*HTT;* Table 1). It has been comprehensively studied in humans where it is associated with Huntington's disease, an inherited autosomal dominant neurodegenerative disorder (Mende-Mueller et al., 2001; Sathasivam et al., 2013). The *HTT* protein directly binds the endoplasmic reticulum (ER) and may play a role in autophagy triggered by ER stress (Atwal and Truant, 2008). Thus, we could speculate a possible involvement of this gene in the adaptation to physiological or pathological conditions leading to ER stress. This gene, among other candidates, was involved in the enrichment of GO terms *pattern specification process* (GO:0007389) and *organ development* (GO:0048513). These two categories were clustered

together with the enriched *neuron maturation* term (GO:0042551) (Table S2). Hence, we could hypothesize a possible role of genes involved in these categories in some morphological traits specific to the Black goat population. Besides this, we noticed the enrichment of genes associated with the response to fatty acids GO terms (GO:0070542; GO:0071398). Candidate genes in these categories include *CPT1A* that encodes for a mitochondrial enzyme responsible for the formation of acyl carnitines that enables activated fatty acids to enter the mitochondria (van der Leij et al., 2000; Vaz and Wanders, 2002). The *SREBF1* gene encodes for a family of transcription factors (*SREBPs*) that regulate lipid homeostasis (Yokoyama et al., 1993; Eberle et al., 2004). The *GNPAT* gene encodes an essential enzyme to the synthesis of ether phospholipids. The last gene in these categories is *CPS1* and it encodes for a mitochondrial enzyme that catalyzes synthesis of carbamoyl phosphate (Aoshima et al., 2001). This suggests that selection acted upon the metabolism of fatty acids and lipids in the Black population, reflecting the possible development of an effective metabolism that could be linked to a higher amount of volatile fatty acids generated by the rumen microbial flora (Bergman, 1990).

In the Draa population, which is raised in oasis/desert areas and well adapted to high temperatures (Hossaini-Hilali and Mouslih, 2002), the enrichment of GO terms associated with the regulation of respiratory system and gaseous exchange categories (GO:0002087; GO:0043576; GO:0044065) would reflect the likely use of panting in evaporative heat loss. Goats could use panting as well as sweating for body thermo regulation according to the level of hydration and solar radiation (Dmiel and Robertshaw, 1983; Baker, 1989), and the type of regulatory system also depends on the breed/population (e.g., The Black Bedouin goats of Sinai Peninsula that use sweating in preference to panting) (Dmiel et al., 1979). Panting compared to sweating helps animals to better preserve their blood plasma volume (no losses of salt) and involves cooling of the blood passing the nasal area, which makes it possible to keep brain temperature lower than body temperature (Baker, 1989). Differences between Draa and Black populations in coat color, hair length and head size (larger in Black, Ibnelbachyr et al., in preparation) would support the hypothesis of different mechanisms of adaptation. Black goats would favor sweating and Draa panting as the more beneficial adaptation to warm environments. Mechanisms underlying dissipation should be further studied in these populations to elucidate the adaptive processes involved.

The enrichment of GO terms associated with lactate transport (GO:0015727; GO:0035873) (Table S3) in the Draa population could be linked to the stronger specific energetic demand associated with pregnancy and lactation in this population. The prolificacy in this population

is much higher than in the rest of Moroccan goats (about 1.51 kids/birth vs. about 1 kid/birth; Ibnelbachyr et al., 2014). Thereby lactate transport may play a crucial role to meet this higher energetic requirement by shuttling lactate to a variety of sites where it could be oxidized directly, re-converted back to pyruvate or glucose and oxidized again, allowing the process of glycolysis to restart and ATP provision maintained (Brooks, 2000; Philp et al., 2005). This corroborates the higher concentration of lactate in cells during lactation than during dry-off period 5 weeks before parturition in cattle reported by Schwarm et al. (2013). Besides this, a top candidate gene in the Draa population was the *agouti signaling protein* (*ASIP*) (Table 1), which plays a key role in the modulation of hair and skin pigmentation in mammals (Lu et al., 1994; Furumura et al., 1996; Kanetsky et al., 2002) by antagonizing the effect of the *melanocortin-1 receptor gene* (*MC1R*) and promoting the synthesis of phaeomelanin, a yellow–red pigment (Hida et al., 2009). *ASIP* was associated with different coat colors in cattle and sheep (Seo et al., 2007; Norris and Whan, 2008). The strong selective sweep related to this gene could be linked to the higher variation in Draa's coat color when compared to other populations (Ibnelbachyr et al., in preparation). This variation in coat color was highly represented in the 14 Draa samples used in this study (Table S4). However, previous studies focussing on this gene identified an important polymorphism in worldwide goat breeds without any clear association with differences in coat color (Badaoui et al., 2011; Adefenwa et al., 2013). Fontanesi et al. (2009b) reported the presence of a copy number variation (CNV) affecting *ASIP* and *AHCY* genes, and might be associated to the white color in Girgentana and Saneen breeds. Nevertheless, the design of our study was not adapted to identify CNV and we cannot link the selection signature detected here in this gene to the findings of this study.

In the Northern population, no GO term was enriched but the second ranked candidate gene identified was *TRAP1*, which encodes a mitochondrial chaperone protein (Felts et al., 2000). Under stress conditions this gene was shown to protect cells from reactive oxygen species, (ROS)-induced apoptosis and senescence (Im et al., 2007; Pridgeon et al., 2007). Such regulation of the cellular stress response would play a role in the adaptation of this population to harsh environments (e.g., mountainous areas in the North of Morocco).

Finally, several strong signals of selection pointed to genes or pathways for which possible functions remained ambiguous. For example in the Northern population, the strong signal of selection associated with *FOXP2*, which encodes for a regulatory protein, is required for proper development of language in Humans (Lai et al., 2001), song learning in songbirds (Haesler et al., 2004), and learning of rapid movement sequences in mice (Groszer et al.,

2008). This gene could be involved in learning but its possible functions in goats cannot be hypothesized easily. A similar case was found in the Draa population for which GO categories linked to behavior and vocalization behavior (GO:0071625; GO:0030534; GO:0007610) were enriched. We were not able to predict the possible functions of these genes. Furthermore, the *NR6A1* gene that was identified potentially under selection in Draa (within the top 0.1% XP-CLR scores) was previously associated with the number of vertebras in pigs (Mikawa et al., 2007; Rubin et al., 2012). Considering the larger body length and size in this population in comparison with the Black population (Ibnelbachyr et al., in preparation), we could hypothesize a similar role of this gene in the body elongation in goats. A future characterization of this morphologic trait in Draa goats would confirm or refute this hypothesis.

Conclusion

Our study characterized whole genome variation in the main goat indigenous populations at a countrywide scale in an unprecedented way. The whole genome data and the wide geographic spread of animals allowed for a precise characterization of the distribution of genomic diversity in various populations. The position of Morocco has made it subject to various colonization waves for domestic animals. Additionally, previous and present management schemes have favored gene flow between goat populations. This created and maintained a very high level of total genetic diversity that is weakly structured according to geography and populations. A part of the overall diversity corresponded to potentially adaptive variation, as several genes appeared to be under selection. The different populations studied appeared to bear specific adaptations, even when submitted to similar conditions such as those related to a warm/desert context. This would demonstrate the potential of different indigenous livestock populations to constitute complementary reservoirs of possibly adaptive diversity that would be highly valuable in the context of global environmental changes. However, these populations are threatened due to their substitution by more productive cosmopolitan breeds that should not have the potential to become locally adapted to harsh environments. It is thus extremely important to promote the sustainable management of these genetic resources with emphasis on both overall neutral and adaptive diversity. This study has also identified several genes as potentially under selection and further studies are needed to depict the underlying mechanisms.

Accession Numbers

The accession numbers of the 44 samples in the BioSamples archive, the accession numbers of the sequencing data and aligned bam files in the ENA archive are reported in the Table S1. The variant calls and genotype calls used in this paper are archived in the European Variation Archive with accession ID ERZ020631.

Author Contributions

PT, FP, SJ, PF designed the study. PT and FP supervised the study. BB, MB, MI, MC, AB, AC sampled individuals. AA, SE produced whole genome sequences. BB, FJA, IS, FB, EC, SS, KL, MI, LC analyzed the data and interpreted the results. BB, FJA, FP, KL, SJ, IS, AA wrote the Manuscript. All authors revised and accepted the final version of the manuscript.

Funding

This work was funded by the UE FP7 project *NEXTGEN* "Next generation methods to preserve farm animal biodiversity by optimizing present and future breeding options"; grant agreement no. *244356*.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

We are greatful to R. Hadria, M. Laghmir, L. Haounou, E. Hafiani, E. Sekkour, M. ElOuatiq, A Dadouch, A. Lberji, C. Errouidi and M. Bouali for their great efforts in sampling goats in Morocco. We thank T. Benabdelouahab for his contribution in the production of some maps. We also thank the two reviewers for valuable suggestions to improve this paper.

References

Adefenwa, M. A., Peters, S. O., Agaviezor, B. O., Wheto, M., Adekoya, K. O., Okpeku, M., et al. (2013). Identification of single nucleotide polymorphisms in the agouti signaling protein (ASIP) gene in some goat breeds in tropical and temperate climates. *Mol. Biol. Rep.* 40, 4447–4457. doi: 10.1007/s11033-013-2535-1

Agha, S. H., Pilla, F., Galal, S., Shaat, I., D'andrea, M., Reale, S., et al. (2008). Genetic diversity in Egyptian and Italian goat breeds measured with microsatellite polymorphism. *J. Anim. Breed. Genet.* 125, 194–200. doi: 10.1111/j.1439-0388.2008.00730.x

Ai, H., Huang, L., and Ren, J. (2013). Genetic diversity, linkage disequilibrium and selection signatures in chinese and Western pigs revealed by genome-wide SNP markers. *PLoS ONE* 8:e56001. doi: 10.1371/journal.pone.0056001

Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65. doi: 10.1038/nature11632

Analla, M., and Serradilla, J. M. (1997). "Problems of selection criteria and genetic evaluations of the goat population in the north of Morocco," in *Data Collection and Definition of Objectives in Sheep and Goat Breeding Programmes: New Prospects. Zaragoza: CIHEAM. Options Méditerranéennes: Série A. Séminaires Méditerranéens; n. 33*, eds D. Gabiña and L. Bodin (Toulouse: Options Méditerranéennes CIHEAM), 153–156

Andolfatto, P., and Przeworski, M. (2001). Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. *Genetics* 158, 657–665.

Aoshima, T., Kajita, M., Sekido, Y., Kikuchi, S., Yasuda, I., Saheki, T., et al. (2001). Novel mutations (H337R and 238-362del) in the CPS1 gene cause carbamoyl phosphate synthetase I deficiency. *Hum. Hered.* 52, 99–101. doi: 10.1159/000053360

Atwal, R. S., and Truant, R. (2008). A stress sensitive ER membrane-association domain in Huntingtin protein defines a potential role for Huntingtin in the regulation of autophagy. *Autophagy* 4, 91–93. doi: 10.4161/auto.5201

Badaoui, B., D'Andrea, M., Pilla, F., Capote, J., Zidi, A., Jordana, J., et al. (2011). Polymorphism of the goat agouti signaling protein gene and its relationship with coat color in Italian and Spanish breeds. *Biochem. Genet.* 49, 523–532. doi: 10.1007/s10528-011-9427-7

Baker, M. A. (1989). Effects of dehydration and rehydration on thermoregulatory sweating in goats. *J. Physiol. Lond.* 417, 421–435. doi: 10.1113/jphysiol.1989.sp017810

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* 57, 289–300.

Benjelloun, B., Pompanon, F., Ben Bati, M., Chentouf, M., Ibnelbachyr, M., El Amiri, B., et al. (2011). Mitochondrial DNA polymorphism in Moroccan goats. *Small Ruminant Res.* 98, 201–205. doi: 10.1016/j.smallrumres.2011.03.041

Benlekhal, A., and Tazi, M. S. (1996). "Situation du secteur caprin au Maroc," in *Les Perspectives de Développement de la Filière Lait de Chèvre Dans le Bassin Méditerranéen. Une Réflexion Collective Appliquée au cas Marocain. (Goat Sector in Morocco. in: Perspectives for Goat Milk Production in the Mediterranean Area. A Collective Reflexion Applied to the Moroccan Situation.)*, Rome: FAO Animal Production and Health paper, FAO.

Bergman, E. N. (1990). Energy contributions of volatile fatty-acids from the gastrointestinal-tract in various species. *Physiol. Rev.* 70, 567–590.

Brooks, G. A. (2000). Intra- and extra-cellular lactate shuttles. *Med. Sci. Sports Exerc.* 32, 790–799. doi: 10.1097/00005768-200004000-00011

Browning, B. L., and Browning, S. R. (2013). Detecting identity by descent and estimating genotype error rates in sequence data. *Am. J. Hum. Genet.* 93, 840–851. doi: 10.1016/j.ajhg.2013.09.014

Cañon, J., Garcia, D., Garcia-Atance, M. A., Obexer-Ruff, G., Lenstra, J. A., Ajmone-Marsan, P., et al. (2006). Geographical partitioning of goat diversity in Europe and the Middle East. *Anim. Genet.* 37, 327–334. doi: 10.1111/j.1365-2052.2006.01461.x

Cao, J., Schneeberger, K., Ossowski, S., Guenther, T., Bender, S., Fitz, J., et al. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* 43, 956–U960. doi: 10.1038/ng.911

Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Res.* 20, 393–402. doi: 10.1101/gr.100545.109

Chentouf, M., Bister, J. L., and Boulanouar, B. (2011). Reproduction characteristics of North Moroccan indigenous goats. *Small Ruminant Res.* 98, 185–188 doi: 10.1016/j.smallrumres.2011.03.037

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., Depristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9, 772–772. doi: 10.1038/nmeth.2109

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491. doi: 10.1038/ng.806

Di, R., Vahidi, S. M. F., Ma, Y. H., He, X. H., Zhao, Q. J., Han, J. L., et al. (2011). Microsatellite analysis revealed genetic diversity and population structure among Chinese cashmere goats. *Anim. Genet.* 42, 428–431. doi: 10.1111/j.1365-2052.2010.02072.x

Dmiel, R., and Robertshaw, D. (1983). The control of panting and sweating in the Black Bedouin goat—a comparison of 2 modes of imposing a heat load. *Physiol. Zool.* 56, 404–411.

Dmiel, R., Robertshaw, D., and Choshniak, I. (1979). Sweat gland secretion in the black bedouin goat. *Physiol. Zool.* 52, 558–564.

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). *Nat. Biotechnol.* 31, 135–141. doi: 10.1038/nbt.2478

Eberle, D., Hegarty, B., Bossard, P., Ferre, P., and Foufelle, F. (2004). SREBP transcription factors: master regulators of lipid homeostasis. *Biochimie* 86, 839–848. doi: 10.1016/j.biochi.2004.09.018

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10:48. doi: 10.1186/1471-2105-10-48

Felts, S. J., Owen, B. A. L., Nguyen, P., Trepel, J., Donner, D. B., and Toft, D. O. (2000). The hsp90-related protein TRAP1 is a mitochondrial protein with distinct functional properties. *J. Biol. Chem.* 275, 3305–3312. doi: 10.1074/jbc.275.5.3305

Fontanesi, L., Beretti, F., Riggio, V., Dall'Olio, S., Gonzalez, E. G., Finocchiaro, R., et al. (2009a). Missense and nonsense mutations in melanocortin 1 receptor (MC1R) gene of different goat breeds: association with red and black coat colour phenotypes but with unexpected evidences. *BMC Genet.* 10:47. doi: 10.1186/1471-2156-10-47

Fontanesi, L., Beretti, F., Riggio, V., Gonzalez, E. G., Dall'Olio, S., Davoli, R., et al. (2009b). Copy number variation and missense mutations of the agouti signaling protein (ASIP) gene in goat breeds with different coat colors. *Cytogenet. Genome Res.* 126, 333–347. doi: 10.1159/000268089

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., and Francois, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics* 196, 973. doi: 10.1534/genetics.113.160572

Furumura, M., Sakai, C., Abdelmalek, Z., Barsh, G. S., and Hearing, V. J. (1996). The interaction of agouti signal protein and melanocyte stimulating hormone to regulate melanin formation in mammals. *Pigment Cell Res.* 9, 191–203. doi: 10.1111/j.1600-0749.1996.tb00109.x

Garrison, E., and Marth, G. (2012). *Haplotype-Based Variant Detection from Short-Read Sequencing*. arXiv.

Groszer, M., Keays, D. A., Deacon, R. M. J., De Bono, J. P., Prasad-Mulcare, S., Gaub, S., et al. (2008). Impaired synaptic plasticity and motor learning in mice with a point mutation implicated in human speech deficits. *Curr. Biol.* 18, 354–362. doi: 10.1016/j.cub.2008.01.060

Haesler, S., Wada, K., Nshdejan, A., Morrisey, E. E., Lints, T., Jarvis, E. D., et al. (2004). FoxP2 expression in avian vocal learners and non-learners. *J. Neurosci.* 24, 3164–3175. doi: 10.1523/JNEUROSCI.4369-03.2004

Hassen, H., Lababidi, S., Rischkowsky, B., Baum, M., and Tibbo, M. (2012). Molecular characterization of Ethiopian indigenous goat populations. *Trop. Anim. Health Prod.* 44, 1239–1246. doi: 10.1007/s11250-011-0064-2

Hida, T., Wakamatsu, K., Sviderskaya, E. V., Donkin, A. J., Montoliu, L., Lamoreux, M. L., et al. (2009). Agouti protein, mahogunin, and attractin in pheomelanogenesis and melanoblast-like alteration of melanocytes: a cAMP-independent pathway. *Pigment Cell Melanoma Res.* 22, 623–634. doi: 10.1111/j.1755-148X.2009.00582.x

Hossaini-Hilali, J., and Mouslih, Y. (2002). La chèvre Draa. Potentiel de production et caractéristiques d'adaptation aux contraintes de l'environnement aride. *Anim. Genet. Res. Inf.* 32, 49–56. doi: 10.1017/S1014233900001553

Hossainihilali, J., Benlamlih, S., and Dahlborn, K. (1993). Fluid balance and milk secretion in the fed and feed-deprived black moroccan goat. *Small Ruminant Res.* 12, 271–285. doi: 10.1016/0921-4488(93)90063-N

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267. doi: 10.1093/molbev/msj030

Ibnelbachyr, M., Boujenane, I., Chikhi, A., and et Er-rouidi, C. (2014). "Le système de conduite de 3 chevrotages en 2 ans: Outil de gestion moderne de la conduite technique de la race caprine locale Draa," in *Technology Creation and Transfer in Small Ruminants: Roles of Research, Developement Services and Farmer Associations. Options Méditerranéennes: Série A. Séminaires Méditerranéens; n. 108*, eds M. Chentouf, A. López-Francos, M. Bengoumi, and D. Gabiña (Tangier: Options Méditerranéennes CIHEAM), 199–207.

Im, C. N., Lee, J. S., Zheng, Y., and Seo, J. S. (2007). Iron chelation study in a normal human hepatocyte cell line suggests that tumor necrosis factor receptor-associated protein 1 (TRAP1) regulates production of reactive oxygen species. *J. Cell. Biochem.* 100, 474–486. doi: 10.1002/jcb.21064

Jombart, T., and Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27, 3070–3071. doi: 10.1093/bioinformatics/btr521

Kanetsky, P. A., Swoyer, J., Panossian, S., Holmes, R., Guerry, D., and Rebbeck, T. R. (2002). A polymorphism in the agouti signaling protein gene is associated with human pigmentation. *Am. J. Hum. Genet.* 70, 770–775. doi: 10.1086/339076

Kidd, J. M., Gravel, S., Byrnes, J., Moreno-Estrada, A., Musharoff, S., Bryc, K., et al. (2012). Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. *Am. J. Hum. Genet.* 91, 660–671. doi: 10.1016/j.ajhg.2012.08.025

Kijas, J. W., Ortiz, J. S., McCulloch, R., James, A., Brice, B., Swain, B., et al. (2013). Genetic diversity and investigation of polledness in divergent goat populations using 52088 SNPs. *Anim. Genet.* 44, 325–335. doi: 10.1111/age.12011

Lai, C. S. L., Fisher, S. E., Hurst, J. A., Vargha-Khadem, F., and Monaco, A. P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* 413, 519–523. doi: 10.1038/35097076

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452. doi: 10.1093/bioinformatics/btp187

Lu, D. S., Willard, D., Patel, I. R., Kadwell, S., Overton, L., Kost, T., et al. (1994). Agouti protein is an antagonist of the melanocyte-stimulating-hormone receptor. *Nature* 371, 799–802. doi: 10.1038/371799a0

Mackay, T. F. C., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D. H., et al. (2012). The Drosophila melanogaster genetic reference panel. *Nature* 482, 173–178. doi: 10.1038/nature10811

Marsden, C. D., Lee, Y., Kreppel, K., Weakley, A., Cornel, A., Ferguson, H. M., et al. (2014). Diversity, differentiation, and linkage disequilibrium: prospects for association mapping in the malaria vector Anopheles arabiensis. *G3* 4, 121–131. doi: 10.1534/g3.113.008326

McCue, M. E., Bannasch, D. L., Petersen, J. L., Gurr, J., Bailey, E., Binns, M. M., et al. (2012). A high density SNP array for the domestic horse and extant perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. *PLoS Genet*. 8:e1002451. doi: 10.1371/journal.pgen.1002451

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 20, 1297–1303. doi: 10.1101/gr.107524.110

Meadows, J. R. S., Chan, E. K. F., and Kijas, J. W. (2008). Linkage disequilibrium compared between five populations of domestic sheep. *BMC Genet*. 9:61. doi: 10.1186/1471-2156-9-61

Mende-Mueller, L. M., Toneff, T., Hwang, S. R., Chesselet, M. F., and Hook, V. Y. H. (2001). Tissue-specific proteolysis of huntingtin (htt) in human brain: evidence of enhanced levels of N- and C-terminal htt fragments in Huntington's disease striatum. *J. Neurosci*. 21, 1830–1837.

Mikawa, S., Morozumi, T., Shimanuki, S. I., Hayashi, T., Uenishi, H., Domukai, M., et al. (2007). Fine mapping of a swine quantitative trait locus for number of vertebrae and analysis of an orphan nuclear receptor, germ cell nuclear factor (NR6A1). *Genome Res*. 17, 586–593. doi: 10.1101/gr.6085507

Naderi, S., Rezaei, H.-R., Pompanon, F., Blum, M. G. B., Negrini, R., Naghash, H.-R., et al. (2008). The goat domestication process inferred from large-scale mitochondrial DNA analysis of wild and domestic individuals. *Proc. Natl. Acad. Sci. U.S.A*. 105, 17659–17664. doi: 10.1073/pnas.0804782105

Naderi, S., Rezaei, H. R., Taberlet, P., Zundel, S., Rafat, S. A., Naghash, H. R., et al. (2007). Large-scale mitochondrial DNA analysis of the domestic goat reveals six haplogroups with high diversity. *PLoS ONE* 2:e1012. doi: 10.1371/journal.pone.0001012

Norris, B. J., and Whan, V. A. (2008). A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep. *Genome Res*. 18, 1282–1293. doi: 10.1101/gr.072090.107

Pereira, F., Queiros, S., Gusmao, L., Nijman, I. J., Cuppen, E., Lenstra, J. A., et al. (2009). Tracing the history of goat pastoralism: new clues from mitochondrial and y chromosome DNA in North Africa. *Mol. Biol. Evol*. 26, 2765–2773. doi: 10.1093/molbev/msp200

Philp, A., Macdonald, A. L., and Watt, P. W. (2005). Lactate - a signal coordinating cell and systemic function. *J. Exp. Biol*. 208, 4561–4575. doi: 10.1242/jeb.01961

Porter, V. (2002). *Mason's World Dictionary of Livestock Breeds, Types and Varieties, 5th Edn*. Wallingford: CABI Publishing.

Pridgeon, J. W., Olzmann, J. A., Chin, L. S., and Li, L. (2007). PINK1 protects against oxidative stress by phosphorylating mitochondrial chaperone TRAP1. *PLoS Biol*. 5, 1494–1503. doi: 10.1371/journal.pbio.0050172

Rubin, C. J., Megens, H. J., Barrio, A. M., Maqbool, K., Sayyab, S., Schwochow, D., et al. (2012). Strong signatures of selection in the domestic pig genome. *Proc. Natl. Acad. Sci. U.S.A*. 109, 19529–19536. doi: 10.1073/pnas.1217149109

Sathasivam, K., Neueder, A., Gipson, T. A., Landles, C., Benjamin, A. C., Bondulich, M. K., et al. (2013). Aberrant splicing of HTT generates the pathogenic exon 1 protein in Huntington disease. *Proc. Natl. Acad. Sci. U.S.A*. 110, 2366–2370. doi: 10.1073/pnas.1221891110

Schwarm, A., Viergutz, T., Kuhla, B., Hammon, H. M., and Schweigel-Rontgen, M. (2013). Fuel feeds function: energy balance and bovine peripheral blood mononuclear cell activation. *Comp. Biochem. Physiol. A Mol. Integr. Physiol*. 164:101–110. doi: 10.1016/j.cbpa.2012.10.009

Seo, K. S., Mohanty, T. R., Choi, T. J., and Hwang, I. (2007). Biology of epidermal and hair pigmentation in cattle: a mini-review. *Vet. Dermatol*. 18, 392–400. doi: 10.1111/j.1365-3164.2007.00634.x

Serrano, M., Calvo, J. H., Martinez, M., Marcos-Carcavilla, A., Cuevas, J., Gonzalez, C., et al. (2009). Microsatellite based genetic diversity and population structure of the endangered Spanish Guadarrama goat breed. *BMC Genet*. 10:61. doi: 10.1186/1471-2156-10-61

Snyder, M., Du, J., and Gerstein, M. (2010). Personal genome sequencing: current approaches and challenges. *Genes Dev*. 24, 423–431. doi: 10.1101/gad.1864110

Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* 6:e21800. doi: 10.1371/journal.pone.0021800

Taberlet, P., Valentini, A., Rezaei, H. R., Naderi, S., Pompanon, F., Negrini, R., et al. (2008). Are cattle, sheep, and goats endangered species? *Mol. Ecol.* 17, 275–284. doi: 10.1111/j.1365-294X.2007.03475.x

Tosser-Klopp, G., Bardou, P., Bouchez, O., Cabau, C., Crooijmans, R., Dong, Y., et al. (2014). Design and characterization of a 52K SNP chip for goats. *PLoS ONE* 9:e86227. doi: 10.1371/journal.pone.0086227

Tresset, A., and Vigne, J. D. (2011). Last hunter-gatherers and first farmers of Europe. *C. R. Biol.* 334, 182–189. doi: 10.1016/j.crvi.2010.12.010

van der Leij, F. R., Huijkman, N. C. A., Boomsma, C., Kuipers, J. R. G., and Bartelds, B. (2000). Genomics of the human carnitine acyltransferase genes. *Mol. Genet. Metab.* 71, 139–153. doi: 10.1006/mgme.2000.3055

Vaz, F. M., and Wanders, R. J. A. (2002). Carnitine biosynthesis in mammals. *Biochem. J.* 361, 417–429. doi: 10.1042/0264-6021:3610417

Veroneze, R., Lopes, P. S., Guimaraes, S. E. F., Silva, F. F., Lopes, M. S., Harlizius, B., et al. (2013). Linkage disequilibrium and haplotype block structure in six commercial pig lines. *J. Anim. Sci.* 91, 3493–3501. doi: 10.2527/jas.2012-6052

Villa-Angulo, R., Matukumalli, L. K., Gill, C. A., Choi, J., Van Tassell, C. P., et al. (2009). High-resolution haplotype block structure in the cattle genome. *BMC Genet.* 10:19. doi: 10.1186/1471-2156-10-19

Wade, C. M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., et al. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* 326, 865–867. doi: 10.1126/science.1178158

Weetman, D., Wilding, C. S., Steen, K., Morgan, J. C., Simard, F., and Donnelly, M. J. (2010). Association mapping of insecticide resistance in wild *Anopheles gambiae* populations: major variants identified in a low-linkage disequilbrium genome. *PLoS ONE* 5:e13140. doi: 10.1371/journal.pone.0013140

Weir, B. S., and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population-structure. *Evolution* 38, 1358–1370. doi: 10.2307/2408641

Yokoyama, C., Wang, X. D., Briggs, M. R., Admon, A., Wu, J., Hua, X. X., et al. (1993). SREBP-1, a basic-helix-loop-helix-leucine zipper protein that controls transcription of the low-density-lipoprotein receptor gene. *Cell* 75, 187–197. doi: 10.1016/S0092-8674(05)80095-9

Zeder, M. A. (2005). "A view from the Zagros: new perspectives on livestock domestication in the Fertile Crescent," in *The First Steps of Animal Domestication. New Archaeological Approaches*, eds J. D. Vigne, J. Peters, and D. Helmer (Oxford: Oxbow Books), 125–146.

Supplementary Material

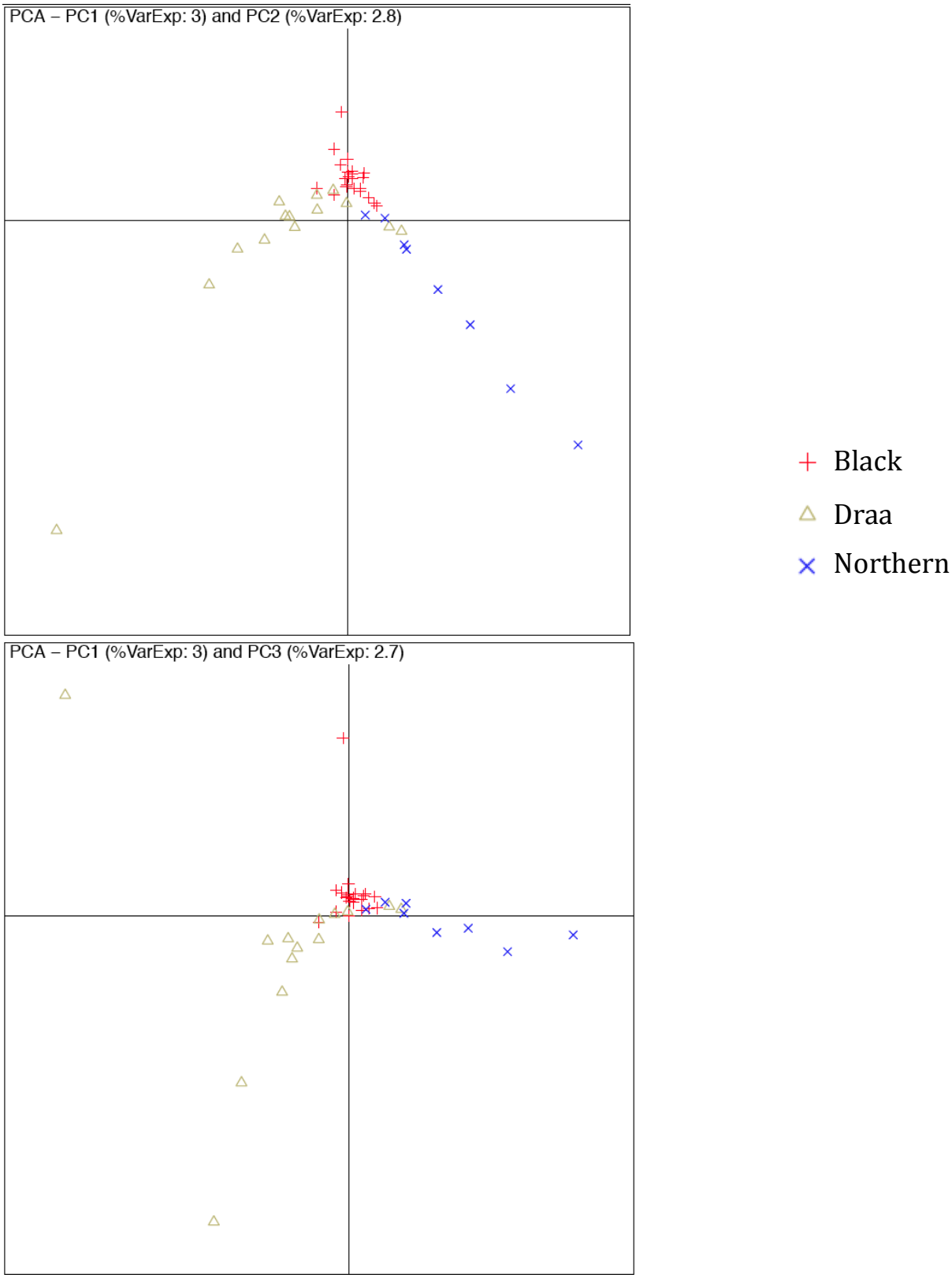

Figure S1. Principal Component Analysis based on the whole genome SNPs for the 44 Moroccan goats.

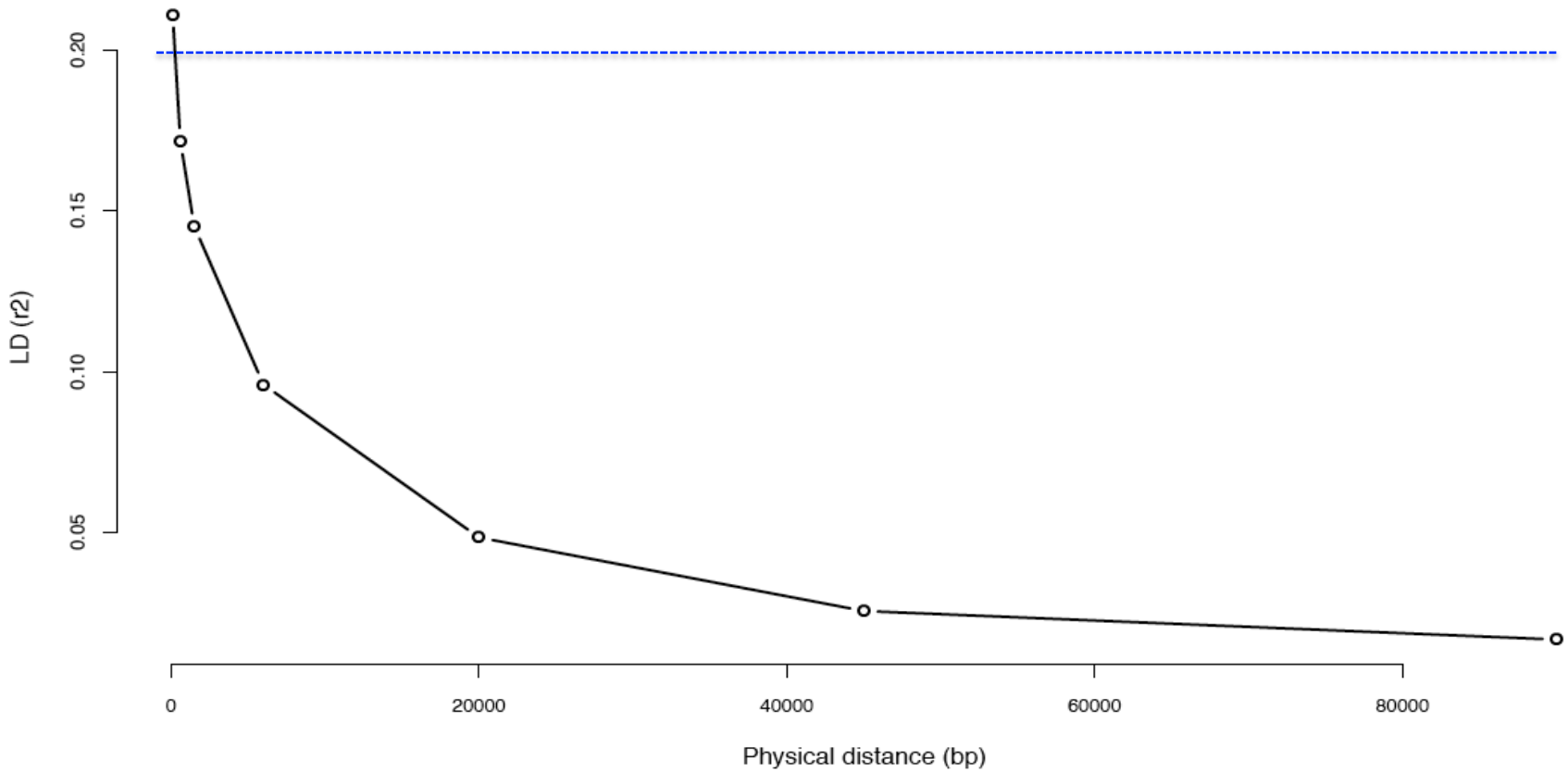

Figure S2. Decay of linkage disequilibrium (r^2) as a function of physical distance including "rare" variants.
The Linkage Disequilibrium (*LD*) was calculated for the 44 Moroccan goats on 5 different segments of 2Mb each on 5 different chromosomes. Inter-variant distances (bp) were binned and averaged into the classes: 0–0.2, 0.2–1, 1–2, 2–10, 10-30, 30-60 and 60-120 kb.

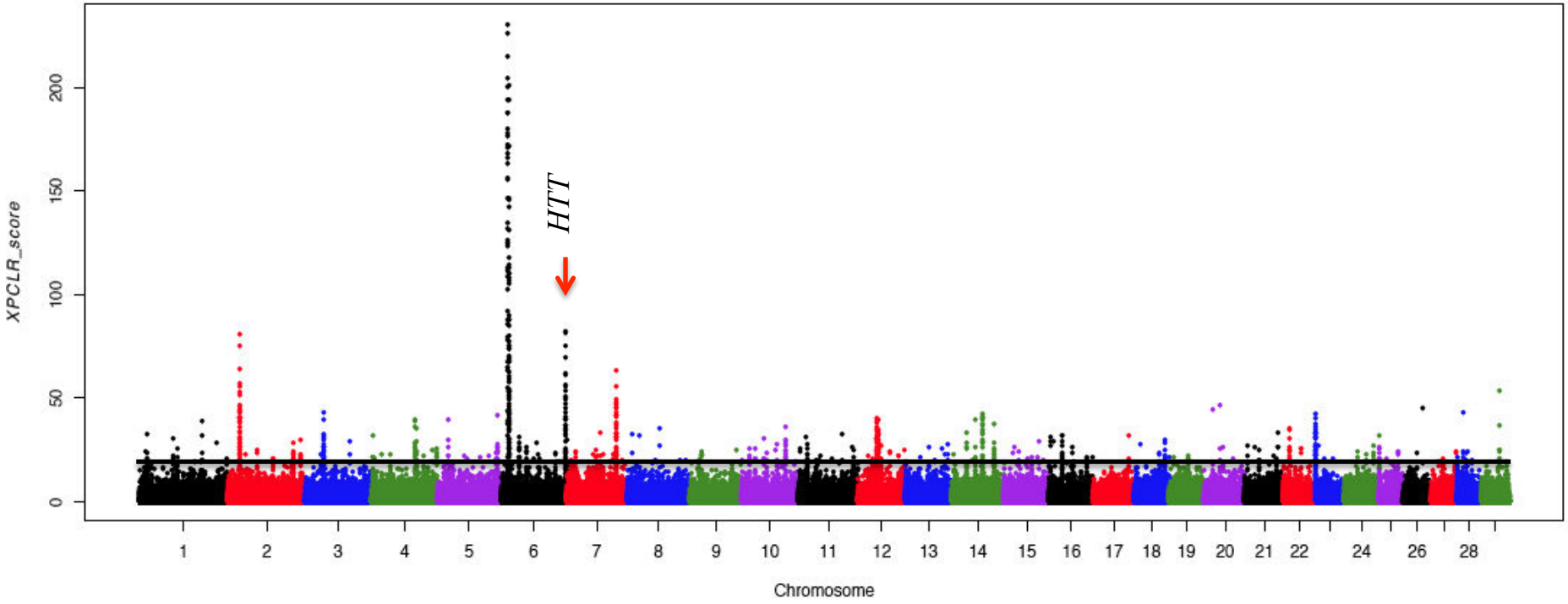

Figure S3. Plot of XP-CLR scores along autosomes in selective sweep analysis for the Black goat population.

The horizontal line indicates a 0.1% autosomal-wide cut-off level. The red arrow and name indicates the top candidate gene. The higher scores linked to the stronger signal on chromosome 6 were not associated to any annotated gene on the goat assembly (CHIR v1.0).

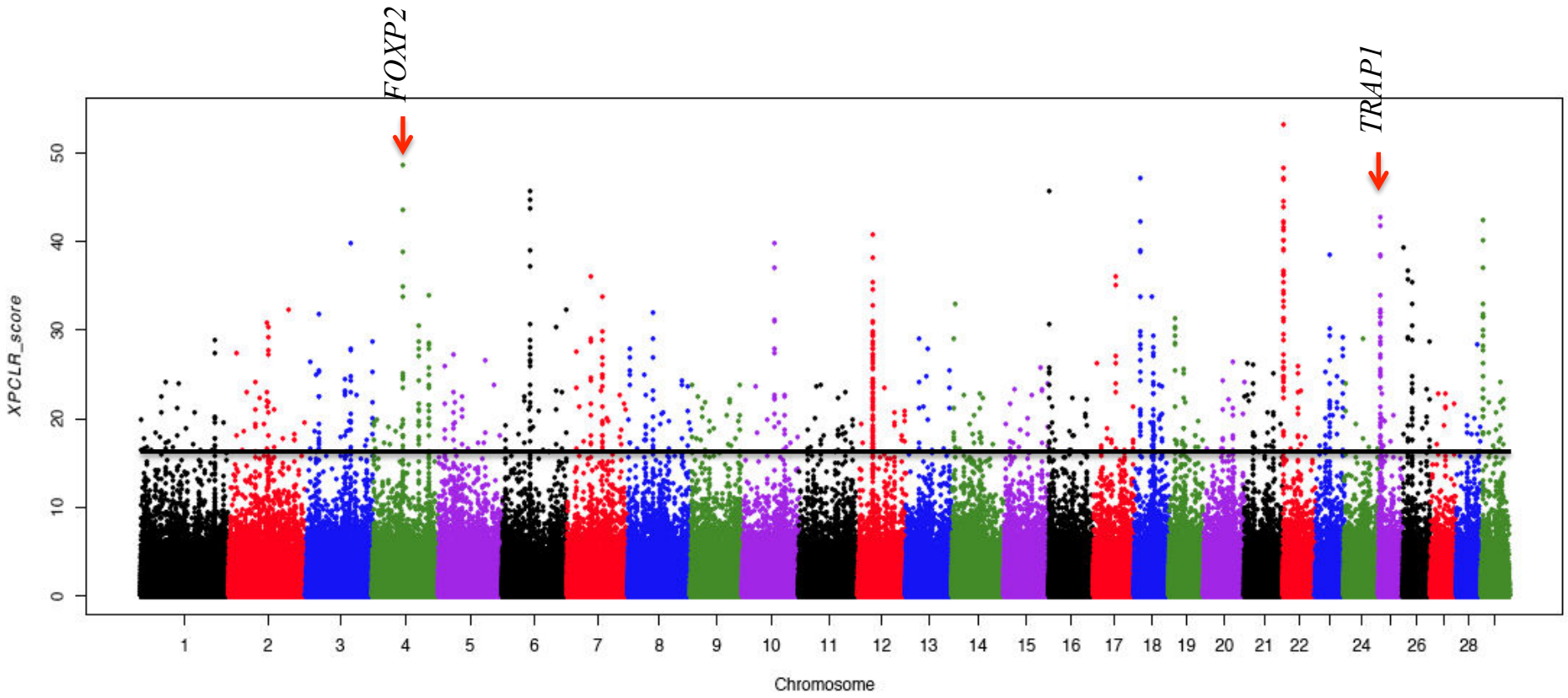

Figure S4. Plot of XP-CLR scores along autosomes in selective sweep analysis for the Northern goat population.

The horizontal line indicates a 0.1% autosomal-wide cut-off level. Red arrows and names indicate the two top candidate genes. The higher scores linked to the stronger signal on chromosome 22 were not associated to any annotated gene on the goat assembly (CHIR v1.0).

Table S1. Characteristics of the 44 samples used for the analyses, their accession numbers in the Biosamples archive and the accession numbers of the sequencing data and aligned bam files in the ENA archive.

Sample name	Biosample accession	ENA sequencing run	ENA aligned bam file	Estimated age (months)	Sex	Longitude (degrees)	Latitude (degrees)	Population
MOCH-H19-1343	SAMEA2012697	ERR248929	ERZ018783	12	female	-11.05555	+28.26907	Draa
MOCH-K17-1351	SAMEA2012705	ERR315500	ERZ018802	18	male	-9.5984	+29.02136	Draa
MOCH-L17-1264	SAMEA2012714	ERR315508	ERZ018745	36	female	-9.25929	+29.0379	Draa
MOCH-L18-1280	SAMEA2012707	ERR234315	ERZ018765	24	male	-9.38272	+28.50969	Black
MOCH-N15-1209	SAMEA2012756	ERR234304	ERZ018699	36	female	-8.05893	+30.26711	Black
MOCH-N16-1228	SAMEA2012757	ERR234305	ERZ018677	12	female	-8.38965	+29.60393	Draa
MOCH-N16-1231	SAMEA2012758	ERR315503	ERZ018727	36	female	-8.08965	+29.60393	Draa
MOCH-N17-1237	SAMEA2012759	ERR315516	ERZ018668	12	male	-8.14666	+29.2436	Draa
MOCH-O14-1203	SAMEA2012763	ERR246143	ERZ018741	36	female	-7.52117	+30.51695	Black
MOCH-O16-1250	SAMEA2012765	ERR315510	ERZ018743	36	female	-7.56962	+29.59379	Black
MOCH-P14-1175	SAMEA2012822	ERR315498	ERZ018763	18	female	-7.08635	+30.59224	Draa
MOCH-P16-1251	SAMEA2012823	ERR246153	ERZ018779	18	female	-7.29038	+29.51272	Draa
MOCH-Q10-0090	SAMEA2012826	ERR229484	ERZ018694	>50	female	-6.6516	+32.7464	Black
MOCH-Q11-0201	SAMEA2012827	ERR248933	ERZ018806	24	female	-6.5748	+32.0533	Black
MOCH-Q13-0153	SAMEA2012829	ERR229476	ERZ018703	.	male	-6.5553	+31.0857	Draa
MOCH-Q14-1167	SAMEA2012830	ERR315512	ERZ018805	>50	female	-6.51941	+30.52099	Draa
MOCH-Q9-0208	SAMEA2012834	ERR248926	ERZ018790	30	female	-6.5208	+33.2172	Black
MOCH-R11-0005	SAMEA2012835	ERR229478	ERZ018721	.	female	-6.26669	+32.22269	Black
MOCH-R12-0195	SAMEA2012836	ERR229479	ERZ018782	48	female	-6.4226	+31.649	Black
MOCH-R13-1104	SAMEA2012838	ERR313264	ERZ018770	>50	female	-6.02545	+31.17796	Draa
MOCH-R14-1105	SAMEA2012839	ERR248928	ERZ018758	24	female	-6.18389	+30.56283	Black
MOCH-R5-0037	SAMEA2012891	ERR340429	ERZ018807	.	female	-6.0898	+35.201	Northern
MOCH-R6-3007	SAMEA2012892	ERR315796	ERZ018704	96	female	-6.05	+34.93	Northern
MOCH-S12-1071	SAMEA2012896	ERR229471	ERZ018693	12	male	-5.5314	+31.50069	Black

MOCH-S13-1064	SAMEA2012897	ERR313261	ERZ018675	>50	female	-5.52492	+31.2501	Black
MOCH-S15-1165	SAMEA2012899	ERR313254	ERZ018757	>50	female	-5.55291	+30.25376	Draa
MOCH-S16-1135	SAMEA2012900	ERR234318	ERZ018755	24	female	-5.53758	+29.59546	Draa
MOCH-S4-0026	SAMEA2012901	ERR229473	ERZ018710	60	female	-5.87	+35.73	Northern
MOCH-S5-0045	SAMEA2012902	ERR219547	ERZ018801	>50	female	-5.713	+35.0314	Northern
MOCH-T13-0128	SAMEA2012898	ERR340425	ERZ018737	>50	female	-5.1657	+31.1077	Black
MOCH-T4-3026	SAMEA2012911	ERR345980	ERZ018817	96	female	-5.469	+35.902	Northern
MOCH-T5-0057	SAMEA2012963	ERR219544	ERZ018689	>50	female	-5.2712	+35.0991	Northern
MOCH-T6-0074	SAMEA2012964	ERR219543	ERZ018688	6	male	-5.3537	+34.9089	Northern
MOCH-U11-1029	SAMEA2012970	ERR313258	ERZ018774	12	male	-4.59874	+32.14459	Black
MOCH-U13-1059	SAMEA2012972	ERR313268	ERZ018723	>50	female	-4.52598	+31.03844	Draa
MOCH-U5-3014	SAMEA2012974	ERR332584	ERZ018819	24	female	-4.633	+35.145	Northern
MOCH-V8-2274	SAMEA2013048	ERR246139	ERZ018730	96	female	-4.45	+33.8	Black
MOCH-V9-1114	SAMEA2013052	ERR248923	ERZ019216	24	female	-4.10499	+33.10143	Black
MOCH-X10-2304	SAMEA2013060	ERR315795	ERZ018822	4	male	-3.416	+32.9	Black
MOCH-X6-2108	SAMEA2013063	ERR315782	ERZ018754	72	female	-3.16944	+34.84444	Black
MOCH-Z11-2197	SAMEA2012106	ERR332588	ERZ018676	60	female	-2.0833	+32.4667	Black
MOCH-Z6-2039	SAMEA2012109	ERR315513	ERZ018669	48	female	-2.1639	+34.6431	Black
MOCH-Z7-2010	SAMEA2012110	ERR248934	ERZ018682	60	female	-2.14917	+34.02583	Black
MOCH-Z9-2154	SAMEA2037794	ERR332576	ERZ018687	12	female	-2.0833	+33.1333	Black

Table S2. Summary of results from enrichment analysis for putative genes under selection in the Moroccan Black goat population.

GO term	Biological process	Number of genes associated	Number of candidate genes associated	P-value	Enrichment
GO:0007070	Negative regulation of transcription from RNA polymerase II promoter during mitosis	2	2	8.42E-5	108.52
GO:0007068	Negative regulation of transcription during mitosis	2	2	8.42E-5	108.52
GO:0035295	Tube development	119	7	1.13E-4	6.38
GO:0097091	Synaptic vesicle clustering	3	2	2.51E-4	72.35
GO:0036444	Calcium ion transmembrane import into mitochondrion	3	2	2.51E-4	72.35
GO:0097479	Synaptic vesicle localization	3	2	2.51E-4	72.35
GO:0007389	Pattern specification process	231	9	2.81E-4	4.23
GO:0042551	Neuron maturation	15	3	3.2E-4	21.70
GO:0070542	Response to fatty acid	39	4	4.39E-4	11.13
GO:0046021	Regulation of transcription from RNA polymerase II promoter, mitotic	4	2	4.99E-4	54.26
GO:0048513	Organ development	829	18	5.58E-4	2.36
GO:0009653	Anatomical structure morphogenesis	908	19	6.08E-4	2.27
GO:0009887	Organ morphogenesis	263	9	7.21E-4	3.71
GO:0045896	Regulation of transcription during mitosis	5	2	8.27E-4	43.41
GO:0071398	Cellular response to fatty acid	21	3	8.99E-4	15.50

Table S3. Summary of results from enrichment analysis for putative genes under selection in Draa goat population.

GO term	Biological process	Number of genes associated	Number of candidate genes associated	P-value	Enrichment
GO:0001941	Postsynaptic membrane organization	5	3	1.26E-5	54.79
GO:0071625	Vocalization behaviour	6	3	2.51E-5	45.66
GO:0002087	Regulation of respiratory gaseous exchange by neurological system process	7	3	4.35E-5	39.14
GO:0044065	Regulation of respiratory system process	8	3	6.91E-5	34.25
GO:0060078	Regulation of postsynaptic membrane potential	41	5	8.00E-5	11.14
GO:0060080	Regulation of inhibitory postsynaptic membrane potential	9	3	1.03E-4	30.44
GO:0007158	Neuron cell-cell adhesion	10	3	1.46E-4	27.40
GO:0030534	Adult behavior	108	7	1.8E-4	5.92
GO:0044708	Single-organism behavior	279	11	2.46E-4	3.60
GO:0043576	Regulation of respiratory gaseous exchange	12	3	2.63E-4	22.83
GO:2000463	Positive regulation of excitatory postsynaptic membrane potential	12	3	2.63E-4	22.83
GO:0035418	Protein localization to synapse	12	3	2.63E-4	22.83
GO:0051899	Membrane depolarization	84	6	3.12E-4	6.52
GO:0097113	Alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionate selective glutamate receptor clustering	3	2	3.55E-4	60.88
GO:0097104	Postsynaptic membrane assembly	3	2	3.55E-4	60.88
GO:0072553	Terminal button organization	3	2	3.55E-4	60.88
GO:0060079	Regulation of excitatory postsynaptic membrane potential	35	4	5.54E-4	10.44
GO:0042391	Regulation of membrane potential	218	9	6.56E-4	3.77
GO:0097119	Postsynaptic density protein 95 clustering	4	2	7.04E-4	45.66
GO:0015727	Lactate transport	4	2	7.04E-4	45.66
GO:0035873	Lactate transmembrane transport	4	2	7.04E-4	45.66

GO:0051965	Positive regulation of synapse assembly	17	3	7.81E-4	16.12
GO:0016337	Single organismal cell-cell adhesion	182	8	8.73E-4	4.01
GO:0045838	Positive regulation of membrane potential	18	3	9.29E-4	15.22
GO:0007610	Behavior	380	12	9.5E-4	2.88

Table S4. Coat colors for the 14 Draa goats used in the analyses. Colors were ordered according to their proportion in the individual coat.

Sample name	Coat color
MOCH-U13-1059	Dark brown, Black, White
MOCH-R13-1104	Light brown
MOCH-S16-1135	Black, White, Light brown
MOCH-S15-1165	White, Black, Dark brown
MOCH-Q14-1167	Dark brown, White, Black
MOCH-P14-1175	Dark brown
MOCH-N16-1228	White
MOCH-N16-1231	White, Dark brown, Black
MOCH-N17-1237	White, Light brown
MOCH-P16-1251	Light brown
MOCH-L17-1264	Light brown, White, Black
MOCH-H19-1343	White, Light brown
MOCH-K17-1351	Black
MOCH-Q13-0153	.

CHAPITRE 3: Les bases génétiques de l'adaptation locales chez les petits-ruminants domestiques

CHAPITRE 3: Les bases génétiques de l'adaptation locales chez les petits-ruminants domestiques

Résumé et présentation de l'article

L'adaptation locale représente l'un des mécanismes les plus importants qui permettent la survie des populations. Il repose principalement sur la sélection des individus les mieux adaptés pour survivre et se reproduire, mais interagit avec plusieurs autres processus évolutifs. Ses mécanismes sont ainsi complexes et loin d'être complètement élucidés. La génomique du paysage représente une discipline émergente qui fournit des outils importants pour étudier l'adaptation locale. Depuis leur domestication il y a environ 10.000 ans, les chèvres et moutons ont été élevés de façon traditionnelle sous une grande diversité de conditions et ont été sujets à des pressions de sélection variables dans le temps et l'espace. Ils auraient ainsi acquis graduellement pendant des millénaires des traits adaptatifs spécifiques à leur environnement. Les ovins et caprins au Maroc représentent un cas très intéressant pour étudier ces traits adaptatifs parce que ces animaux sont nombreux et bien répartis sur tout le territoire qui est caractérisé par des conditions écologiques et climatiques très contrastées.

Dans ce chapitre qui constitue aussi une partie importante du projet NextGen, nous avons adopté une approche de génomique du paysage qui a été appréhendée par un large échantillonnage basé sur un système de grille de cellules rectangulaires (0.5°x0.5°) couvrant la grande part du Maroc (≈400.000 km^2). C'est une zone caractérisée par l'élevage de toutes les races et populations indigènes du pays sous des conditions climatiques et écologiques très contrastées. Une banque de 1412 et 1283 échantillons non apparentés respectivement d'ovins et de caprins issus de 164 cellules a été constituée, et les données environnementales caractéristiques des lieux d'échantillonnage ont été collectées. 160 moutons et 161 chèvres représentatifs de l'ensemble du gradient de variation du climat ont été sélectionnés, et leurs génomes complets ont été séquencés à un taux de couverture de 12X. Nous avons caractérisé la structuration génétique des groupes étudiés et nous avons adopté deux approches de détection des signatures de sélection. Une première approche spécifique à la génomique du paysage est basée sur l'identification des sites polymorphes dont la variation est corrélée à une variation environnementale donnée. La seconde approche est populationnelle et consiste à contraster deux groupes d'individus qui se placent aux deux extrémités d'un gradient environnemental pour identifier les portions de génome qui les distinguent. Cette approche a été appliquée pour étudier l'adaptation à sept variables (cinq par espèce) représentatives des

grandes catégories environnementales (altitude, pente, température, précipitations), tandis que l'approche corrélative a été testée sur 81 différentes variables éco-climatiques en éliminant les plus corrélées d'entre elles (|r|>0,9).

Cette étude montre une forte diversité qui est très faiblement structurée selon les régions ou les populations dans les deux espèces. Elle identifie via l'approche populationnelle plusieurs signatures de sélection localisée en grande partie dans les portions non-codantes du génome, suggérant ainsi l'importance probable de la sélection des régions régulatrices dans les mécanismes adaptatifs. Une autre partie de ces signatures de sélection est associée aux gènes (dont une partie de variation non-sens) qui permettent d'identifier plusieurs voies métaboliques qui seraient sous-jacentes aux traits adaptatifs. Les voies majeures identifiées impliquent des mécanismes respiratoires et des processus cardiaques pour l'adaptation à l'altitude, et la biosynthèse de l'ATP pour l'adaptation à la pente.

Ce chapitre montre que les mécanismes impliqués dans l'adaptation à un même facteur environnemental seraient généralement différents chez les chèvres et moutons. Moins de 1% des gènes identifiés sont communs aux deux espèces pour les mêmes variables environnementales. Cependant, certains gènes sont identifiés chez les deux espèces. C'est le cas du locus *NFIB* qui est impliqué dans la maturation des poumons et la différenciation des cellules de Clara (cellules progénitrices dans les petites voies respiratoires) et qui est identifié pour l'adaptation à l'altitude chez les deux espèces. Cette étude montre également l'implication possible du locus *MCM3* dans l'adaptation des moutons à l'altitude. Ce gène est connu pour avoir une action régulatrice sur la famille de gènes *HIF,* dont le gène *EPAS1* qui a été identifié dans les populations tibétaines comme impliqué dans l'adaptation à l'altitude chez l'Homme (Yi et al. 2010 ; Simonson et al. 2010). Ceci suggère une certaine forme de convergence adaptative chez différentes espèces. De plus, cette étude caractérise l'évolution de la différenciation des zones sous sélection le long du gradient d'altitude. Cette différenciation présente, selon les gènes, différents patrons de variation, permettant de visualiser les altitudes clés auxquelles les modifications génétiques seraient sélectionnées.

Par ailleurs, l'approche corrélative ne permet d'identifier que 25 variants candidats chez les moutons et 54 chez les chèvres qui seraient associés à au moins l'une des 81 variables éco-climatiques étudiées. La moitié de ces variants n'a pas été détectée par l'approche populationnelle.

Compte tenu de la nature et de la masse des données en jeu, la mise en œuvre de ce travail requiert la modification de plusieurs outils d'analyse existants et un temps considérable. Ainsi, l'analyse est toujours en cours de réalisation. Comme stipulé dans l'Introduction générale, nous présentons dans ce chapitre les résultats obtenus jusqu'à maintenant sous forme d'ébauche d'article (en préparation). Bien entendu, en collaboration avec les autres partenaires impliqués dans cette étude, nous sommes en train de prospecter l'utilisation d'autres approches corrélatives (e.g., LFMM ; Frichot et al. 2013), de finaliser l'identification des gènes et voies métaboliques liés aux signatures de sélection (incluant l'identification des effets spécifiques des variations non-sens identifiées). Nous finaliserons également la comparaison des mécanismes identifiés pour nos deux modèles d'études, la chèvre et le mouton.

Article C: Towards the genetic bases of local adaptation: a wide-scale landscape genomic approach in sheep (*O. aries*) and goats (*C. hircus*)

Badr Benjelloun[1,2,3*], Kevin Leempoel[4*], Sylvie Stucki[4*], Ian Streeter[5], Pablo Orozco Ter-Wengel[6], Frédéric Boyer[1,2], Florian J. Alberto[1,2], Filippo Biscarini[7], Mustapha Ibnelbachyr[8], Mohamed BenBati[3], Mouad Chentouf[9], Abdelmajid Bechchari[10], Stefan Engelen[11], Adriana Alberti[11], Abdelkader Chikhi[9], Laura Clarke[5], Michael W. Bruford[6], Alessandra Stella[7], Paul Flicek[5], Pierre Taberlet[1,2], Stéphane Joost[4], François Pompanon[1,2]

[1] Laboratoire d'Ecologie Alpine, Université Grenoble-Alpes, Grenoble, France

[2] Laboratoire d'Ecologie Alpine, Centre National de la Recherche Scientifique, Grenoble, France

[3] National Institute of Agronomic Research (INRA Maroc), Regional Centre of Agronomic Research, Beni-Mellal, Morocco

[4] Laboratory of Geographic Information Systems (LASIG), School of Civil and Environmental Engineering (ENAC), École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[5] European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK

[6] School of Biosciences, Cardiff University, Cardiff, UK

[7] Parco Technologico Padano, Lodi, Italy

[8] Regional Centre of Agronomic Research Errachidia, National Institute of Agronomic Research (INRA Maroc), Errachidia, Morocco

[9] Regional Centre of Agronomic Research Tangier, National Institute of Agronomic Research (INRA Maroc), Tangier, Morocco

[10] Regional Centre of Agronomic Research Oujda, National Institute of Agronomic Research (INRA Maroc), Oujda, Morocco

[11] Centre National de Séquençage, CEA-Institut de Génomique, Genoscope, Évry, France

Paper under preparation

Summary

Since their domestication 10 kyears ago followed by a worldwide spread, sheep and goats have accumulated highly valuable adaptive traits allowing them to be raised within highly diversified environments. Besides the current rapid development and wide spread of just a few productive cosmopolitan breeds marked by limited genetic diversity, indigenous populations may keep adaptive traits that would constitute crucial genomic resources in the context of environmental changes. We sequenced the genomes of 160 indigenous sheep and 161 goats representative of the Moroccan-wide diversity in ecology, climate and geographic origin. We detected 39 million variants in sheep and 32 million in goats showing a very weak geographic structure over the country in both species. We used population-based and correlative approaches to identify several sets of loci and genes that likely have a role in local adaptation globally to altitude, slope, rainfall, temperature and their variation. The main adaptive pathways were associated with respiration and circulation for the adaptation to altitude as well as ATP biosynthesis for slope. The major genes identified to be related to altitude showed different patterns of variation of genetic differentiation along the altitudinal gradient. Candidate genes for adaptation to the same environmental variable were generally different between the two species, suggesting different adaptive mechanisms in sheep and goats. However, similar or functionally linked genes responding to the same environmental variable were also found such as *NFIB* loci that is associated to lung maturation and was putatively associated to the adaptation to altitude in sheep and goats.

Key words

Local adaptation, whole genome sequences, sheep, goats, landscape genomics, selection

Introduction

Local adaptation is the adjustment or changes in behaviour, physiology and structure of an organism to become more suited to its environment. It relies on the increase in frequency in a population of traits that are advantageous under its local environmental context. The strict criterion is that a population must have higher fitness at its native site than any other population introduced to that site (Kawecki and Ebert 2004). In a context of climate change, remaining locally adapted would permit an efficient population survival (Franks and Hoffmann 2012), although other mechanisms would have also an important role in this persistence (Loarie et al. 2009; Chevin et al. 2010).

Thus, understanding the genetic mechanisms of local adaptation requires depicting the genetic basis of fitness variation within and across natural environments (Bergelson and Roux 2010). Several approaches have been developed from the characterization of association mapping by correlating phenotypes with genotypes to the study of genetic differentiation using population genetics approaches (Fournier-Level et al. 2011; Savolainen et al. 2013).

The genetic bases of several adaptive traits with relatively simple modes of inheritance have already been characterized, such as heavy metal tolerance in plants (Macnair 1993) or marine–freshwater adaptation in threespine sticklebacks (Jones et al. 2012). Such adaptations typically involve one or a few major loci. However, most of adaptive traits are affected by many segregating loci, show a large non-genetic variability (Ward and Kellis 2012) and so far have been less well addressed, which has led to a poor understanding of most of adaptative mechanisms. Only a few studies elucidated any complex adaptation issues. For example, (Fournier-Level et al. 2011) identified an important role of regulatory variation in Arabidopsis Thaliana local adaptation at the European scale by integrating loci fitness in natural environments using a GWAS approach. Similarly, Daub et al. (2013) found evidence for association between small polygenic epistatic effects and human adaptation to pathogens. Otherwise, since 2003, landscape genetics has emerged as a new research area that enables the spatial mapping of allele frequencies from one or more species (or populations) and, subsequently, the correlation of such patterns with the landscape variations (Manel et al. 2003). It integrates population genetics as well as landscape ecology and spatial statistics and it gives more options to understand adaptation as well as gene flow and their interaction, e.g. (Dionne et al. 2008). More recently, technical progress in sequencing and the emergence of large environmental datasets have opened a new branch of landscape genetics: landscape

genomics, which aims to identify environmental or landscape factors that influence adaptive genetic diversity using genome scans with a large numbers of molecular markers genotyped (Manel and Holderegger 2013). So far, most landscape-genomic studies have used population genomic approaches to detect adaptive genomic variation (Manel and Holderegger 2013). However, specific landscape genomic approaches have also been developed and they directly correlate allele frequencies with environment factors (Joost et al. 2007; Frichot et al. 2013; Stucki et al. 2014). One important issue in such studies remains the need to account for genetic structure and/or demographic effects during analysis (Manel and Holderegger 2013).

Sheep and goats play a crucial role in feeding human populations throughout the world. In 2013, they had a global population of 1.2 and 1 billions respectively [http://faostat.fao.org/site/573/] and they represent together with cattle the main source of meat and milk at the worldwide scale. These species were among the first ungulates to be domesticated, between 10.5 and 9.9 kyears ago near the Fertile-Crescent (Peters et al. 2005; Zeder, 2005). During this process, a large part of genetic diversity present in wild animals was captured (Naderi et al. 2008; Taberlet et al. 2008). Then, domesticated animals rapidly spread over the rest of the old-world where they were raised for a long time under various environments representing a wide range of geo-climatic conditions and husbandry practices (Taberlet et al. 2008). These populations were managed with only moderate selection for traits of interest, and reproduction practices allowed important gene flows among them. They accumulated gradually highly valuable adaptive traits to their environments (i.e. climate, ecology and husbandry) and maintained high levels of phenotypic diversity (Taberlet et al. 2008). Therefore, indigenous populations would constitute a suitable model to study mechanisms underlying their local adaptation. Furthermore, these mechanisms would be of great interest for long-term conservation of these species in a combined context of biodiversity loss in farm animals and environmental changes.

Sheep and goats in Morocco are very diverse; they are mostly indigenous and their breeding systems are characterized by moderate anthropic selection pressures (for goats, see (Benjelloun et al. 2015)). Furthermore, indigenous small-ruminants are raised over almost the whole country under a very wide geo-climatic and ecologic diversity (e.g. 15°C<Temperature annual range<42°C; Climatic Research Unit (New et al. 2002)).

Here we sequenced at 12X coverage 160 sheep and 161 goats representing the Moroccan-wide geo-climatic diversity and we used a landscape genomics framework to identify

selection signatures across genomes and elucidate mechanisms underlying local adaptation in these species to a wide range of eco-climatic variables.

Material & Methods

Sampling

Sample collection was performed in a wide part of Morocco covering a range of highly contrasted environments (~400,000 km^2; Northern part of Morocco in latitude range [28°-36°]; Figure 1). A sampling grid consisting of 162 cells of 0.5° of longitude and latitude was established and a maximum of 3 unrelated animals have been sampled by flock in 3 different flocks per cell for each species. For each individual, tissue samples were collected from the distal part of the ear and placed in alcohol for one day, and then transferred to a silica-gel tube until DNA extraction. A total of 412 flocks were covered from which we had to select 164 individuals. The most important criterion was to optimize the selection in order to represent a wide range of environmental conditions among our samples. The second point was to take geographic space into account in order to maximise individuals' spread over the covered area and to assure a spatial representativeness of all regions. Traditional random sampling cannot take these criteria into account.

In order to choose samples to be as different as possible, we first performed a principal component analysis (PCA) on the 117 variables extracted from the Climatic Research Unit (CRU) dataset (New et al. 2002). The PCA allows us to maximise the ecological distance between the farms (separately for sheep and goats). Afterwards, we performed an ascending hierarchical classification on the first 7 PCA-axis (96% of the variance) to regroup sampled farms in function of their ecological distances. Using the Ward criteria, we reduced the number of classes to 164 (Escoffier & Pages 2008).

After regrouping, we selected one individual per class. In order to assure spatial representativeness, we performed 50 random samplings and chose the one with the maximal index of repartition (i.e. the maximal sum of distances between each farm and its nearest neighbour). After sequencing, we had to remove individuals due to low sequence quality, and 161 goats and 160 sheep were kept.

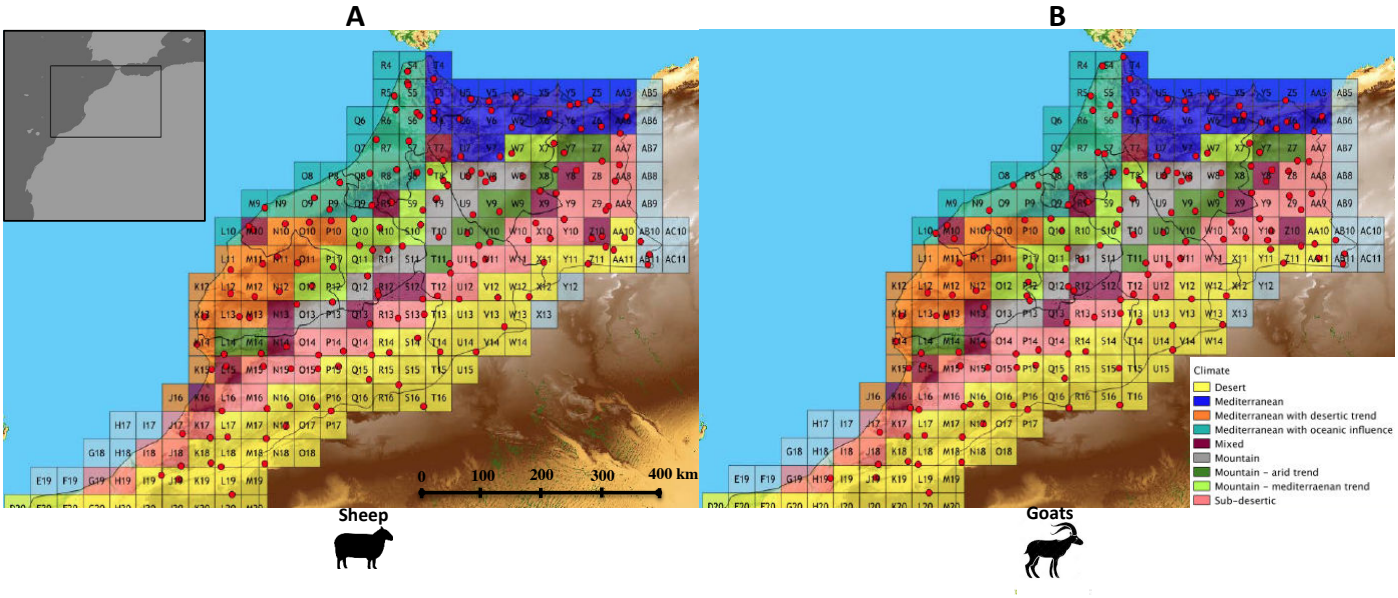

Figure 1. Distribution of sheep and goats sampled.

Maps showing the distribution of the 160 sheep (A) and 161 goats (B) sampled over the main climate categories. Each point represents one individual.

Production of WGS datasets

DNA extractions were done using the Puregene Tissue Kit from Qiagen® following the manufacturer's instructions. Then, 500ng of DNA were sheared to a 150-700 bp range using the Covaris® E210 instrument (Covaris, Inc., USA). Sheared DNA was used for Illumina® library preparation by a semi-automatized protocol. Briefly, end repair, A tailing and Illumina® compatible adaptors (BiooScientific) ligation were performed using the SPRIWorks Library Preparation System and SPRI TE instrument (Beckmann Coulter), according to the manufacturer protocol. A 300-600 bp size selection was applied in order to recover the most of fragments. DNA fragments were amplified by 12 cycles PCR using Platinum Pfx Taq Polymerase Kit (Life® Technologies) and Illumina® adapter-specific primers. Libraries were purified with 0.8x AMPure XP beads (Beckmann Coulter). After library profile analysis by Agilent 2100 Bioanalyzer (Agilent® Technologies, USA) and qPCR quantification, the libraries were sequenced using 100 base-length read chemistry in paired-end flow cell on the Illumina HiSeq2000 (Illumina®, USA).

WGS data processing

Illumina paired-end reads for sheep were mapped to the sheep reference genome (OAR v3.1, GenBank assembly GCA_000317765.1 (Jiang et al. 2014)) and those for goats were mapped

to the goat reference genome (CHIR v1.0, GenBank assembly GCA_000317765.1 (Dong et al. 2012)) using BWA mem (Li and Durbin 2009). The BAM files produced were then sorted using Picard SortSam and improved using Picard Markduplicates (http://picard.sourceforge.net), GATK RealignerTargetCreator, GATK IndelRealigner (DePristo et al. 2011) and Samtools calmd (Li et al. 2009). Variant calling was done using three different algorithms: Samtools mpileup (Li et al. 2009), GATK UnifiedGenotyper (McKenna et al. 2010) and Freebayes (Garrison and Marth 2012).

There were two successive rounds of filtering variant sites. Filtering stage 1 merged together calls from the three algorithms, whilst filtering out the lowest-confidence calls. A variant site passed if it was called by at least two different calling algorithms with variant quality > 30. An alternate allele at a site passed if it was called by any one of the calling algorithms, and the genotype count > 0. Filtering stage 2 used Variant Quality Score Recalibration by GATK. First, we generated a training set of the highest-confidence variant sites where (i) the site is called by all three variant callers with variant quality > 100, (ii) the site is biallelic (Palti et al. 2015) the minor allele count is at least 3 while counting only samples with genotype quality > 30. The training set was used to build a Gaussian model using the tool GATK VariantRecalibrator using the following variant annotations from UnifiedGenotyper: QD, HaplotypeScore, MQRankSum, ReadPosRankSum, FS, DP, InbreedingCoefficient. The Gaussian model was applied to the full data set, generating a VQSLOD (log odds ratio of being a true variant). Sites were filtered out if VQSLOD < cutoff value. The cutoff value was set for each population by the following: Minimum VQSLOD = {the median value of VQSLOD for training set variants} - 3 * {the median absolute deviation VQSLOD of training set variants}. Measures of the transition / transversion ratio of SNPs suggest that this chosen cutoff criterion gives the best balance between selectivity and sensitivity. Genotypes were improved and phased by Beagle 4 (Browning and Browning 2013), and then filtered out where the genotype probability calculated by Beagle is less than 0.95.

Genetic diversity and population structure in Moroccan sheep and goats

Neutral genomic variation was characterized to evaluate the level of genetic diversity present in Moroccan sheep and goats. The total number of variants and the number of variants within each population were calculated. The level of nucleotide diversity (π) was calculated in each species and averaged over all of the biallelic and fully diploid variants for which all individuals had a called genotype using Vcftools (Danecek et al. 2011). The observed percentage of heterozygote genotypes per individual (*Ho*) was calculated considering only the

biallelic SNPs with no missing genotype calls. From *Ho,* the inbreeding coefficients (*F*) were calculated for each individual using population allelic frequencies over all individuals.

Pairwise linkage disequilibrium (LD) was assessed through the correlation coefficient (r^2). It was estimated in 5 segments of 2Mb on different chromosomes (physical positions between 5 Mb and 7 Mb on chromosomes 6, 11, 16, 21 and 26). LD was estimated either by using the whole set of reliable variants or after discarding rare variants with a minor allele frequency (MAF) less than 0.05. For both estimations, r^2 values between all pairs of bi-allelic variants (SNPs and indels) on the same segment were calculated using Vcftools. Inter-SNP distances (kb) were binned into the following 7 classes: 0–0.2, 0.2–1, 1–2, 2–10, 10-30, 30-60 and 60-120 kb and observed pairwise *LD* was averaged for each inter-SNP distance class and used to draw *LD* decay.

Genetic structure among individuals was assessed using two different methods: (i) a principal component analysis (PCA) was done using an *LD* pruned subset of bi-allelic SNPs. *LD* between SNPs in windows containing 50 markers was calculated before removing one SNP from each pair where *LD* exceeded 0.95. Subsequently, only 12,543,534 SNPs among a total of 29,427,980 bi-allelic SNPs were kept for this analysis in goats and 14,056,772 out of 30,069,299 of bi-allelic SNPs for sheep. The R package adegenet v1.3-1 (Jombart and Ahmed 2011) was used to run PCA and Plink v1.90a (https://www.cog-genomics.org/plink2) was used for *LD* pruning. (ii) An analysis with the clustering method sNMF (Frichot et al. 2014) was carried-out. This method was specifically developed for fast analysis of large genomic datasets. It is based on sparse non-negative matrix factorization to estimate admixture coefficients of individuals. All bi-allelic variants were used and five runs for each K value from 1 to 10 were performed using a value of *alpha* parameter of 16. For each run, the cross-entropy criterion was calculated with 5 % missing data to identify the most likely number of clusters. The run showing the lowest cross-entropy (CE) value for a given K was considered, and similarly, the number of clusters associated with the lower CE was considered as the most likely representative of our data structure.

Environmental variables

67 climatic variables were extracted from the WorldClim dataset (Hijmans *et al*. 2005; http://www.worldclim.org/current) for the Moroccan sheep and goats sampling locations. These variables are based on data collected over 30 years and provide temperature and precipitation measurements as well as bioclimatic indices (i.e. derived from the monthly

temperature and rainfall values in order to generate more biologically meaningful variables) with an initial resolution of 1x1km. Additionally, we used a Digital Elevation Model (DEM) with a resolution of 90m (SRTM; http://earthexplorer.usgs.gov; courtesy of the U.S. Geological Survey) to obtain topography related variables. A total of 14 DEM-derived variables were computed from SAGA GIS (Böhner *et al*. 2006) and include for instance altitude, slope, solar radiation, etc. A multi-scale analysis framework was used to evaluate the sensitivity of associations to a change of resolution. For DEM-derived variables, we used a Gaussian pyramid to generalize the DEM at resolutions of 180, 360, 720, 1440, 2880m. Each DEM-derived variables was then computed on each of these DEMs. For climatic variables, we increased the window size (from a 3x3 window of pixels to 33x33km) in order to consider a larger habitat for each individual. Afterwards, we conducted pairwise correlation analysis between all 81 variables to remove highly correlated ones ($|r|>=0.9$). Because the sampling locations of sheep and goats are slightly different, these analyses were performed separately for both datasets (Figure 1). We found high correlations between WorldClim temperature variables and precipitation. Most of the bioclimatic variables were highly correlated with temperature and precipitations. However, DEM-derived variables were not highly correlated and most were kept (e.g. Solar radiation variables were not highly correlated to slope or to eastness/northness). Thus, 31 out of 81 variables were retained for sheep and 27 for goats (Table S1). These selected variables were all included in our correlative approach to find selection signatures (see section "Analyses of signatures of selection"). Due to the difficulty to run our population-based method for each variable, we selected five representative ones of the main categories for each species. Therefore, we included altitude, slope, mean temperature in July (temp7), precipitation in April (prec4) and temperature annual range (bio7) in sheep analyses. For goats, we included altitude, slope, temp7 as well as rainfall in March (prec3) and precipitation seasonality (bio15).

Analyses of signatures of selection

Our landscape genetics framework was based on two different approaches to identify selection signatures associated with the environmental variation.

Correlative approach

A correlative approach was applied using Samβada (Stucki et al. 2014). It is an improved version of the spatial analysis method SAM (Joost et al. 2007), which increases its computational power over large datasets on one hand, and provides multivariate models on

the other hand. This individual-based method performs logistic regressions in which the binary genetic marker is either present or absent and correlates with quantitative environmental variables. Therefore, it provides the probability of occurrence of a genotype for each individual in relation with environmental parameters as well as spatial statistics that are helpful for the interpretation of significant results with regards to spatial autocorrelation.

Due to the size of our datasets, we first performed uni-variate models using initial resolution of each variablesin order to compute associations within a reasonable time and facilitate the handling of output files. Those analyses were performed using a pruned subset of biallelic SNPs, *LD* between SNPs in windows containing 50 markers was calculated before removing one SNP from each pair where *LD* exceeded 0.5. These subsets were constituted of 5.1 and 5.3 million SNPs for sheep and goats respectively. Multi-scale analysis was then performed on a subset of SNPs associated with a Q-value of 0.4 in the first step with the identified variables, except latitude and longitude. Therefore, 243 SNPs and 24 variables were used in sheep and 1341 SNPs and 24 variables were used in goats A False discovery threshold of 0.2 was then applied on Samβada's results to identify candidate SNPs.

Population-based approach

A genome scan method based on population genetics models was applied on our datasets. We worked on 7 variables representing various environmental categories, i.e. climatic variables temperature, precipitations and DEM-derived altitude and slope with their respective initial resolutions (see "Environmental variables" section). For each variable, two pools of 20 individuals were constituted, each representing one extreme of the gradient of variation of the variable. The XP-CLR method (Chen et al. 2010) was then run to identify potential regions differentially selected in each extreme pool. It is a likelihood method for detecting selective sweeps that involves jointly modelling the multi-locus allele frequency differentiation between two populations. It is based on a reference population and an object one. Theoretically, this method is designed to identify genomic regions under positive selection in the object population. Therefore, for each variable, we did the analysis twice to consider each extreme group as an object one in one analysis knowing that our groups were not homogenous and could not be considered as populations. This method is robust to detect selective sweeps and especially with regards to the uncertainty in the estimation of local recombination rate (Chen et al. 2010). Due to the absence of genomic position, the physical position (1 Mb ≈ 1 cM) was used. We used overlapped segments of a maximum of 27 cM to estimate and assemble XP-CLR scores using the whole set of bi-allelic variants as described

in Benjelloun et al. (2015). Overlapping regions of 2cM were applied and the scores related to the extreme 1cM were discarded, except at the starting and the end of chromosomes on the OAR v3.1 and the CHIR v1.0 genome assemblies. XP-CLR scores were calculated using grid points spaced by 2500 bp with a maximum of 250 variants in a window of 0.1 cM and by down-weighting contributions of highly correlated variants (r^2>0.95) in the reference group. The 0.1% genomic regions with highest XP-CLR scores revealed by the analysis were identified and the top differentiated variants between the two pools and located within those top XP-CLR windows (0.1 cM each) were defined using a 0.1% genome-wide cut-off level of *Fst* (Weir and Cockerham 1984). In addition, the top-XP-CLR windows that were overlapped were grouped into pools representing top-peaks that were ranked across autosomes. Lastly, the identified variants were classified in various categories (i.e. intron, exon, synonymous, missense, inter-genic…) using the Variant Effect Predictor (VEP) tools (McLaren et al. 2010) for both species. Lists of genes that include or less than 5 kb away from the identified candidate variants (Downstream 5'-end and upstream 3'-end) were established and used for the Gene Ontology enrichment analyses.

For each species we aimed at depicting the pattern of differentiation of the top candidate genes under selection across the environmental gradients. For that, for each environmental variable we ranked the 160 sheep (and 161 goats, respectively) according to the ranking of their geographic position on the environmental gradient considered. A sliding limit moving by steps of 10 individuals was applied to define 2 groups among which the *Fst* value (Weir and Cockerham 1984) was estimated based on the candidate variants associated to those genes. The minimum number of individuals per group was 20 and the maximum 140. Then, this allowed plotting the variation of the *Fst* value along the environmental gradient.

Gene Ontology enrichment analyses

To explore the biological processes in which the candidate genes identified are involved, Gene Ontology (GO) enrichment analyses were performed using the application GOrilla (Eden et al. 2009). The 12,669 goat and 14620 sheep genes associated with a GO term were used as background references. Significance for each individual GO-identifier was assessed with P-values that were corrected using FDR q-value according to the Benjamini and Hochberg (Benjamini and Hochberg 1995) method. GO terms identified for each variable were clustered into homogenous groups using REVIGO and allowing medium similarity (0.7) (Supek et al. 2011). Low similarity among GO terms in a group was applied and the weight of each GO term was assessed by its p-value. Due to the insufficient numbers of genes identified

using our correlative approach, these analyses were restricted to the genes identified by our population-based method (i.e. genes associated to variants meeting both criteria: 0.1% top XP-CLR scores and 0.1% top *Fst*).

Results

Population structure

We mapped unambiguously 99.4% (± 0.1%) of sheep reads on the OAR v3.1 assembly and 98.9% (± 0.1%) of goat reads on the CHIR v1.0 assembly. 38,599,873 variants were successfully called in sheep, among which 38,278,356 were polymorphic, 2,607,680 (6.8%) were small insertions/deletions (indels) and 808,753 (2.1%) variants displayed more than two alleles (mainly tri-allelic). For goats, 31,743,850 variants were discovered in the total dataset among which 31,650,083 were polymorphic, 2,137,479 (6.7%) were small indels and 219,236 (0.7%) variants displayed more than two alleles. Rare variants characterized by a minor allele frequency (MAF) less than 5% represented 17,022,878 (44.1%) in sheep and 18,513,669 (58.3%) in goats. The whole genome nucleotide diversity was 0.174 in sheep and 0.126 in goats. The average (± s.d.) heterozygosity (*Ho*) and inbreeding coefficient (*F*) were 0.166 (± 0.014) and 0.045 (± 0.081) respectively in sheep and 0.119 (± 0.012) and 0.056 (± 0.096) respectively in goats. Linkage disequilibrium was assessed by the pairwise r^2 value between polymorphic sites in the studied genomic regions. Using the whole set of reliable variants, the genomic distance at which it decayed to less than 0.15 was 655 bp in sheep and 166 bp in goats. Moreover, r^2 decayed to less than 0.1 in 3.12 kb and 2.1 kb in sheep and goats respectively (Figure S1). When withdrawing rare variants (MAF<0.05), the average r^2 decayed to less than 0.2 in 3.6 kb in sheep and 5.8 kb in goats. It decayed to less than 0.15 in 4.4 kb in sheep and 8.1 kb in goats (Figure S1).

PCA analysis showed that the first and second principal components explained together less than 2% of variation in both species, and reflected no obvious pattern of population structure (Figure S2). Consistently, sNMF suggested no significant effect of population structure as the data was better explained by a single cluster in each species (Figure S3). However, a weak pattern of geographic structure when considering the existence of three clusters (Figures 2 and 3), especially in goats where the red component was more prominent in the North and the yellow one in the Southwestern of the sampling grid (Figure 3B).

Detection of signals of selection related to environmental variations

We developed two different genome scan approaches to look at selection signatures related to environmental variations. On one hand, a population-based approach based on XP-CLR method and *Fst* was applied. Candidate variants in the windows associated to the top XP-CLR-scores were identified using a cut-off of genome wide *Fst* estimates. On the other hand, we used the correlative algorithm of Samßada to look at the influence of a wide set of environmental variables on the allelic frequencies along environmental gradients.

Population-based approach

Combining the XP-CLR and *Fst* methods, we highlighted 5981 (± 746) different candidate variants and 141 (± 20) different candidate genes on average in each one of the 5 studied variables in sheep (Table 1) and 4930 (± 564) candidate variants and 214 (± 25) candidate genes in each extreme group for the 5 studied variables in goats (Table 2). Most of the identified variants were inter-genic (65.9% ± 5.18% in sheep and 61.1 ± 4.02 in goats). Missense candidate variants represented 0.19% (± 0.06%) and 0.04% (± 0.02%) in sheep and goats respectively and synonymous variants represented 0.39 (± 0.12%) in sheep and 0.02 (± 0.04%) in goats. Intron variants represented 26.0 (± 4.63) in sheep and 32.5 (± 3.57) in goats (Tables 1, 2, S2 and S3).

In sheep, we identified 136 candidate genes related to altitude and 112 candidate genes were identified for rainfall in April. Similarly, 165 genes were identified for temperature annual range (bio7), 150 genes for mean temperature of July (temp7) and 144 genes were identified for slope (Table 1). Candidate genes in goats were 252 for altitude, 209 for rainfall in March (prec3), 201 for rainfall seasonality (bio15), 221 for mean temperature of July (temp7) and 185 genes for slope (Table 2). The 8 annotated genes showing the strongest XP-CLR scores in both analyses for each of the main environmental variables are presented in Table 3.

The differentiation of candidate variants and genes (i.e. associated to high peaks of XP-CLR scores; e.g. Figure 4) along environmental gradients showed different clear patterns, generally with a highest differentiation close to one or both extremes of the gradient forming “U” or “S” shapes (Figures 5 and 6).

Lastly, 3 of the genes associated to each of the three variables altitude, “temp7” and slope were common to sheep and goats. They represent less than 1% of the total number of genes identified for these three environmental variables.

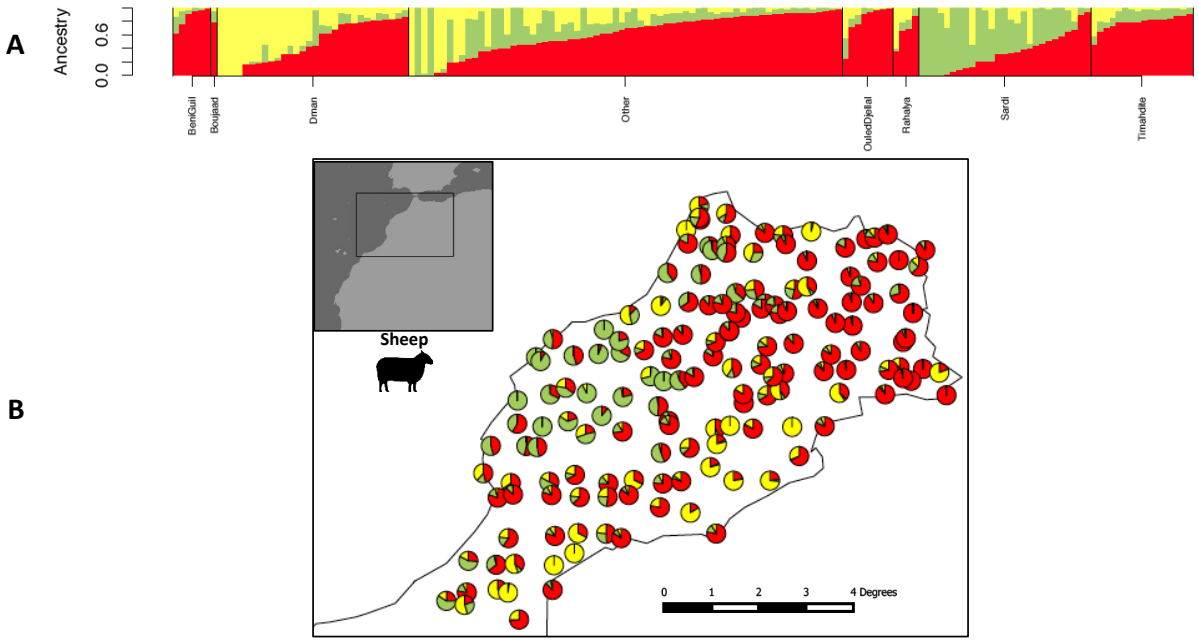

Figure 2. Admixture coefficient estimates for Moroccan sheep for K=3 clusters.

(A) Each bar represents one individual. Different colours illustrate the assignment proportion (Q score) to each one of the assumed clusters. Individuals were grouped according to various breeds or populations. (B) Geographical distribution of individual Q-score values.

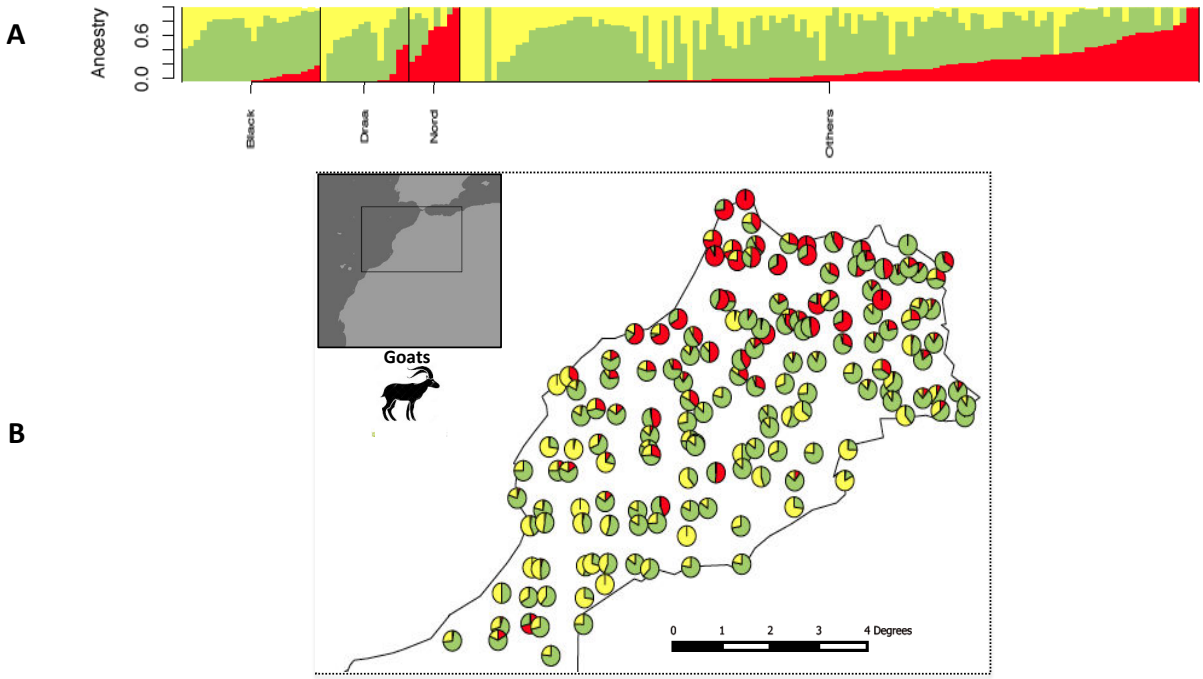

Figure 3. Admixture coefficient estimates for Moroccan goats for K=3 clusters.

(A) Each bar represents one individual. Different colours illustrate the assignment proportion (Q score) to each one of the assumed clusters. (B) Geographical distribution of individual Q-score values.

Table1. Number of candidate genes and variants under positive selection detected by the population-based approach for the five environmental variables studied in Moroccan sheep.

		Mean ± SD.	Altitude		Prec4		Bio7		Temp7		Slope	
			High	Low	Low	High	High	Low	High	Low	High	Low
Number of genes		141 ± 20	136		112		165		150		144	
Number of variants		5981 ± 746	5436		6250		5811		7133		5275	
Proportions (%)	**Missense**	0.19 ± 0.06	0.16	0.21	0.26	0.24	0.14	0.26	0.24	0.19	0.10	0.13
	Synonymous	0.39 ± 0.12	0.28	0.21	0.31	0.54	0.34	0.29	0.41	0.53	0.56	0.45
	Splice-region/Synonymous	0.01 ± 0.01	0.03	0	0	0	0	0	0	0.02	0	0
	Non-coding exon	0.02 ± 0.05	0	0	0	0	0	0	0.14	0.07	0	0
	Intron	26.0 ± 4.63	21.8	24.7	24.0	18.7	26.5	27.0	34.2	22.7	31.4	28.7
	Splice-region/Intron	0.08 ± 0.04	0.12	0.06	0.07	0.06	0.03	0.03	0.07	0.05	0.13	0.13
	5' UTR	0.04 ± 0.03	0.03	0.03	0	0.03	0.03	0	0.02	0.07	0.07	0.10
	3' UTR	0.30 ± 0.25	0.03	0	0.65	0.57	0.03	0.03	0.43	0.43	0.13	0.41
	Downstream	3.68 ± 1.30	1.62	2.15	3.21	5.19	3.55	2.93	4.39	5.83	4.22	3.76
	Upstream	3.74 ± 0.90	3.57	4.60	3.69	3.01	2.65	2.90	5.04	5.09	3.01	3.79
	Inter-genic	65.9 ± 5.18	72.4	68.1	68.5	72.2	66.8	66.6	55.4	65.4	60.6	62.9

Genes and variants were displayed for each of the two XP-CLR/*Fst* analyses done by environmental variable and by merging the results of the two analyses to show complete lists per variable. Percentages were estimated by analysis.

Table 2. Candidate genes and variants under positive selection detected by the population-based approach for the five environmental variables studied in Moroccan goats.

		Mean ± SD.	Altitude		Prec3		Bio15		Temp7		Slope	
			High	Low	Low	High	High	Low	High	Low	High	Low
Number of genes		214 ± 25	252		209		201		221		185	
Number of variants		4930 ± 564	5408		4963		5470		4719		4090	
Proportions (%)	**Missense**	0.04 ± 0.03	0.08	0.05	0	0.06	0.02	0.03	0.04	0.03	0	0.04
	Synonymous	0.02 ± 0.04	0.08	0.11	0.04	0	0.02	0	0	0	0	0
	Intron	32.5 ± 3.57	39.1	30.5	30.6	27.9	34.9	31.2	28.7	35.1	30.8	32.3
	Splice-region/Intron	0.05 ± 0.04	0.11	0.08	0.04	0	0	0.07	0.04	0	0.04	0.13
	5' UTR	0.22 ± 0.09	0.34	0.32	0.25	0.29	0.27	0.10	0.17	0.25	0.29	0.04
	Splice-region/5' UTR	0.01 ± 0.02	0.04	0.03	0.04	0	0	0	0	0	0	0
	3' UTR	0.73 ± 0.25	1.03	1.00	0.50	0.97	0.98	0.93	0.50	0.62	0.81	0.55
	Splice-region/3' UTR	0.00 ± 0.01	0	0.03	0	0	0	0	0	0	0	0
	Downstream	2.96 ± 1.00	4.46	4.04	2.37	4.33	3.18	2.03	2.84	3.93	2.32	1.68
	Upstream	3.08 ± 1.52	3.16	3.81	2.29	2.95	2.39	1.51	3.39	2.93	2.85	1.68
	Inter-genic	61.1 ± 4.02	52.7	61.1	64.3	64.4	59.2	65.1	64.8	57.8	63.7	64.2

Genes and variants were displayed for each of the two XP-CLR/*Fst* analyses done by environmental variable and by merging the results of the two analyses to show complete lists per variable. Percentages were estimated by analysis.

Table 3. The eight top annotated candidate genes associated with the higher XP-CLR scores for each environmental parameter

	Sheep					Goats				
Environmental variable	**Altitude**	**Temp7**	**Slope**	**Prec4**	**Bio7**	**Altitude**	**Temp7**	**Slope**	**Prec3**	**Bio15**
Top candidate genes	*GMDS*	*ENSOARG00000011616*	*ENSOARG00000003995*	*LDLRAD4*	*GMDS*	*KHDRBS2*	*SPAG17*	*C14H8orf37*	*ARHGAP15*	*CNTN5*
	MCM3	*RXFP2*	*NEUROG3*	*ENSOARG00000016955*	*ENSOARG00000023809*	*LOC102180412*	*MTDH*	*HTR2A*	*PRKCE*	*CDH2*
	ASRGL1	*STK19*	*SV2B*	*SLC35F2*	*FOXP4*	*PRAMEF12*	*DEGS2*	*SNTG2*	*GRSF1*	*LOC102175357*
	OXR1	*CHMP4C*	*EDN3*	*BCAS1*	*RYBP*	*GATAD2A*	*EVL*	*DUSP2*	*RUFY3*	*UBE2D1*
	NFIB	*C4orf17*	*CYP19A1*	*ENSOARG00000011119*	*LAMB1*	*LOC100861296*	*AIMP1*	*PHTF1*	*C16H1orf53*	*LOC102175300*
	U6	*ENSOARG00000015390*	*ENSOARG00000000908*	*ENSOARG00000017401*	*ENSOARG00000008550*	*LOC102180242*	*GIMD1*	*FAM105B*	*LHX9*	*RNASE4*
	IRAK4	*ENSOARG00000015647*	*KNTC1*	*ENSOARG00000017428*	*EDAR*	*PMPCA*	*TBCK*	*JRK*	*HTR2A*	*INPP4A*
	PUS7L	*ENSOARG00000011616*	*FAM193A*	*MGME1*	*PTPN4*	*SDCCAG3*	*PET112*	*EXD2*	*PLCG1*	*SAPS3*

The four top candidate genes identified in each one of the two XP-CLR/*Fst* analyses for each environmental parameter were considered.

Temp7 is the mean temperature in July; Prec4 is rainfall in April; Bio7 is Temperature annual range; Bio15 is rainfall variation (variation coefficient).

Correlative approach

Samβada identified 25 candidate variants in sheep among which nine were associated to eight genes (in the intron or downstream genes; Table S4). Samßada detected 8 variants for the four variables studied also by the population based approach. Six of them were also identified by this last approach (Table 4).

In goats, our approach with Samßada identified 56 variants among which 15 were associated with 15 various genes (Table S5). Samßada identified 20 SNPs associated with three variables studied also by XP-CLR/*Fst* approach. Only eight from them were identified also by the last approach (Table 5).

Gene Ontology enrichment analysis

The genes that were identified by our population genetic approach were used for Gene Ontology (GO) enrichment analyses (See "Material and methods"). Seven and eight GO categories were enriched for adaptation to altitude respectively in sheep and goats including as main categories Muscle contraction, positive regulation of leukocyte proliferation and Regulation of DNA recombination in sheep (Table 6) and Heart contraction and process, Clara cell differentiation and regulation of ion transmembrane transport in goats (Table7). The enrichment of genes associated with slope from our analysis highlighted the significance of four GO categories in sheep (Table S6) and six different enrichment categories in goats including mainly Regulation of ATP biosynthetic (Table S7). Genes identified for rainfall in April did not allow the identification of any significant GO category in sheep but those identified for rainfall in March in goats were associated with 30 GO categories that clustered into five highly differentiated categories (if small REVIGO similarity=0.5 is required) (Supek et al. 2011) including Neutrophil chemotaxis, immune response-regulating cell surface receptor signalling pathway involved in phagocytosis, positive regulation of multicellular organismal process and regulation of ion transport (Table S8). Seven GO terms were enriched for temperature annual range (bio7) in sheep (Table S9) and 16 GO categories were enriched for rainfall seasonality (bio15) in goats.

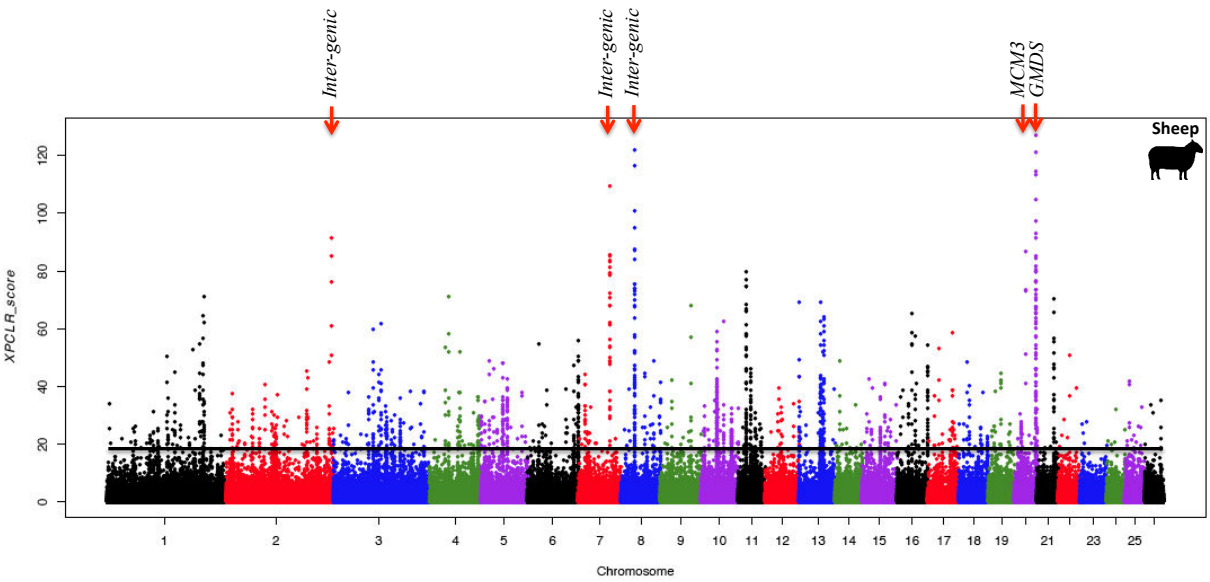

Figure 4. Plot of XP-CLR scores along autosomes in selective sweep analysis for the high altitude in Moroccan sheep.

The horizontal lines indicate a 0.1% autosomal-wide cut-off level. Red arrows and names indicate the nature or names of genes associated with the four top signals candidate genes.

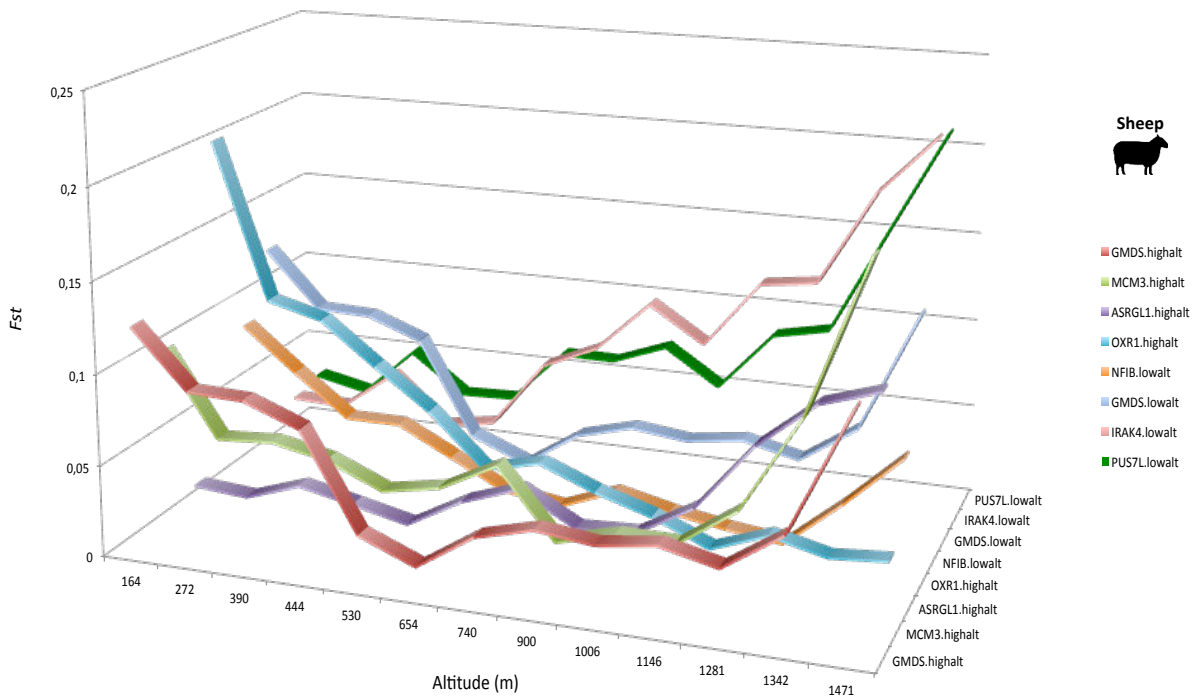

Figure 5. Evolution of differentiation index (*Fst*) for a sliding limit along an altitudinal gradient in the eight top-score candidate genes identified in sheep.

The sliding limit was applied from the higher altitude recorded in the 20 sheep low altitude extreme group (164 m) to the lower altitude in the high altitude 20 individuals extreme group (1471 m). The sliding limit moved each time using steps of 10 individuals.

Table 4. Candidate variants/genes identified by multi-resolution analysis with Samßada in Moroccan sheep for the environmental variables studied with the population-based approach.

Environmental variable	Chr	Position	Best Resolution	SNP type	Gene	Detection by XP-CLR/Fst
Prec_4	23	43794976	Initial	Intron	*LDLRAD4*	Yes
	23	43812782	Initial	Intron	*FAM210A*	Yes
	23	43847594	3x3km	Intron	*RNMT*	Yes
	23	43861704	9x9km	Inter-genic	-	Yes
	23	43874160	17x17km	Downstream	*MC1R*	Yes
	23	44038684	3x3km	Inter-genic	-	Yes
	23	44084253	3x3km	Inter-genic	-	No
Bio7	14	875672	17x17km	Intron	*VAC14*	No

Altitude was not correlated to any variant with Smßada in sheep.

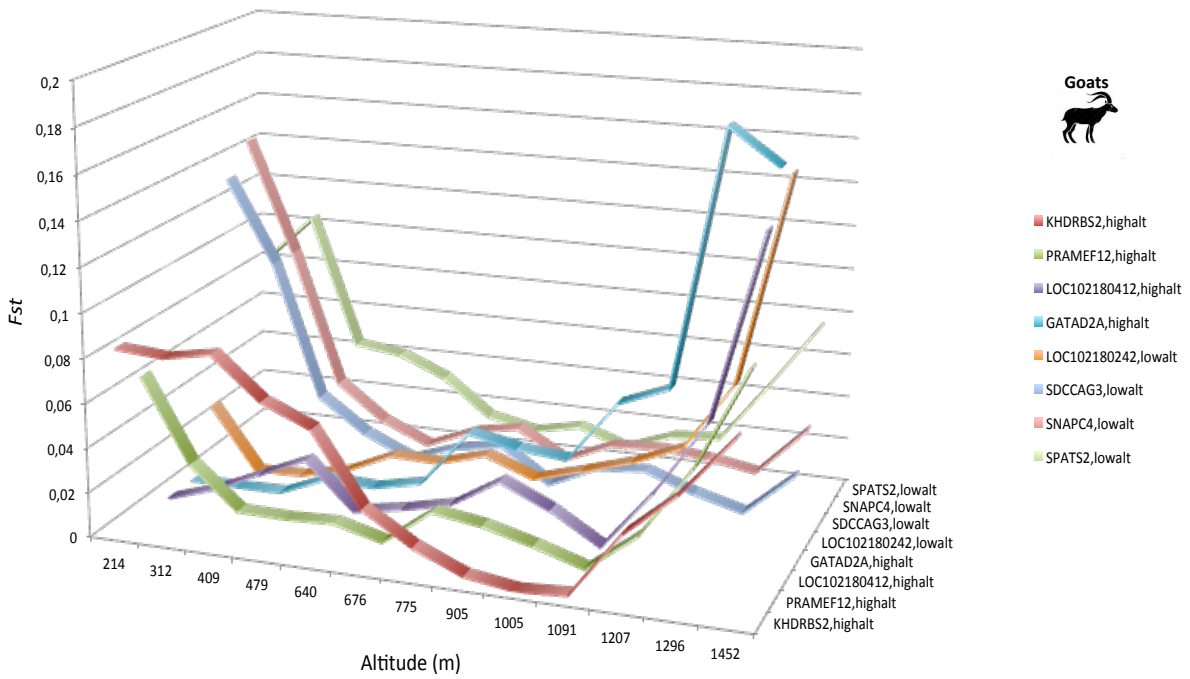

Figure 6. Evolution of differentiation index (*Fst*) for a sliding limit along an altitudinal gradient in the eight top-score candidate genes identified in goats.

The sliding limit was applied from the higher altitude recorded in the 20 individuals low altitude extreme group (234 m) to the lower altitude in the high altitude 20 individuals extreme group (1452 m). The sliding limit moved each time using steps of 10 individuals.

Table 5. Candidate variants/genes identified by muli-resolution analysis with Samßada in Moroccan goats for the environmental variables studied with the population-based approach.

Environmental variable	Chr	Position	Best resolution	SNP type	Gene	Detection by XP-CLR/Fst
Altitude	1	23002164	1439x1439m	Inter-genic	-	No
	22	42794456	1439x1439m	Intron	PXK	Yes
	28	40372626	1439x1439m	Intron	ARHGAP22	No
	6	12207826	2879x2879m	Inter-genic	-	Yes
	6	12218302	2879x2879m	Inter-genic	-	No
	6	12254244	2879x2879m	Inter-genic	-	Yes
	6	12259667	2879x2879m	Inter-genic	-	Yes
	6	25849772	Initial	Inter-genic	-	No
	6	26455416	720x720m	Inter-genic	-	No
Slope	22	41805444	180x180m	Inter-genic	-	No
	8	46850695	Initial	Inter-genic	-	No
Bio15	24	19436980	17x17km	Inter-genic	-	No
	24	28807953	Initial	Inter-genic	-	No
	4	95035251	5x5km	Intron	EXOC4	No
	6	12187316	17x17km	Inter-genic	-	Yes
	6	12218302	17x17km	Inter-genic	-	No
	6	12242353	17x17km	Inter-genic	-	No
	6	12254244	9x9km	Inter-genic	-	Yes
	6	12259667	Initial	Inter-genic	-	Yes
	6	12276649	Initial	Inter-genic	-	Yes

Discussion

Since their domestication, indigenous sheep and goats have been raised for a long time under highly diversified conditions and have gradually accumulated several characteristics making them well adapted to their environments. The mechanisms underlying these adaptions have been poorly studied until now. The Moroccan small ruminants constitute an interesting case study for investigating the genetic bases of local adaptation. Morocco exhibits a very large geo-climatic and ecologic diversity, and its geographic position has made it subject to numerous colonization waves for domestic animals in general and small-ruminants especially resulting in a low geographic structure of genetic variations (Pereira et al. 2009; Benjelloun et al. 2015). Furthermore, small-ruminants are numerous, typically indigenous, well distributed in the whole territory and raised under a wide range of husbandry practices. In this study, we used whole genome sequences for 321 sheep and goats representing the Moroccan-wide environmental variety.

The sampling strategy allowed a wide coverage of the environmental and genetic diversity spanning in Morocco. It allowed also more options for data analysis and a higher resolution for scientific investigation to find association between genomes and numerous environmental variables.

Overall genomic variation

The huge number of WGS data produced allowed an unprecedented resolution when describing genomic variation in small ruminants. The high proportions of mapped reads in both species would illustrate the high quality of sequence data produced and the relative completeness of the genome assemblies used. The slightly higher percentage of mapped reads in sheep (0.5% more) would result from the possible higher completeness of sheep genome assembly in comparison with CHIR1.0. However, sheep and goats displayed very large counts of genomic variants (38.6 M and 31.7 M respectively) enlarging substantially the worldwide catalogue of ovine and caprine variants. Sheep showed 6.9 million more variants than goats, with a higher nucleotide diversity that could be linked to a higher percentage of rare variants in goats (58% of variants showing MAF<0.05 versus 44% in sheep). The number of variants discovered in sheep is much higher than most of the previous studies discovering variants using whole genome sequences from large sample sets. For example, the human 1000 Genomes Project (Altshuler et al. 2012) detected approximately 15 million SNPs and 1 million short indels. However, in a recent study describing the worldwide human variation identified 81 million SNPs and 3.4 million short indels from more than 2500 individuals (1000 GenomesConsortium, submitted). The polymorphism shown here was comparable to that reported by Ai et al. (2015) who discovered about 41 million variants over 69 Chinese pig sequences including wild boars. Otherwise, we report here approximately 7 million additional caprine variants than Benjelloun et al. (2015) who used a quarter subset of the Moroccan goats from the present study (n=44).

Sheep were more heterozygotes than goats and less inbred but linkage disequilibrium was slightly lower in goats. However, *LD* value is highly influenced by the percentage of rare variants and when we removed them sheep displayed even lower *LD*. This fact could also partly explain differences in heterozygosity and inbreeding coefficients between the two species. Generally, *LD* extents found here complete on one hand the findings of Benjelloun et al. (2015) who found a longer *LD* extent ($r^2_{0.15}$=1.33kb using the whole set of variants and $r^2_{0.15}$≈12 kb when excluding rare variants) using a subset of our goat dataset. The difference in reported LD results from the fact that we used here many more animals for that estimate. On the other hand, *LD* values reported here are shorter than all those reported on other domestic animals (i.e. horses, cattle, pigs) where it largely exceeds 10 kb for r^2=0.20 (Villa-Angulo et al. 2009; Wade et al. 2009; McCue et al. 2012; Ai et al. 2013; Veroneze et al. 2013). As described by Benjelloun et al. (2015), in some of these studies, whole genome variants were

not available and potential biases due to the use of SNP chips would partially explain our findings. However, our results would mainly illustrate a high effective population size and the effect of the very common extensive breeding systems favouring high gene flows among Moroccan sheep and goats and the absence until now of very strong selection pressure.

The very high polymorphism in Moroccan indigenous sheep and goats was weakly structured over geographic regions and among phenotypic groups. Only a weak pattern of geographic structure was shown in goats by sNMF (with k=3), with Northern individuals displaying a higher assignment probability to one distinct cluster. As advanced by Benjelloun et al. (2015) this may be explained by a possible influence of Iberian gene flows through the strait of Gibraltar in the North of Morocco. However, such patterns were not displayed in sheep. Typically, the weak population structure observed in Moroccan small ruminants would demonstrate that there have been no strong bottlenecks experienced by those populations. This could be linked to moderate intensity of selection associated with abundant gene flows and/or a high genetic diversity that was preserved even during the processes leading to the formation of various breeds and populations.

Table 6. Enrichment analysis for putative genes under selection in relation with altitude in Moroccan sheep.

GO Term	Biological process	Number of genes associated	Number of candidate genes associated	P-value	Enrichment
GO:0070665	Positive regulation of leukocyte proliferation (a)	106	6	1.32E-4	7.66
GO:0010569	Regulation of double-strand break repair via homologous recombination	14	3	1.34E-4	29.01
GO:0006936	Muscle contraction	178	7	3.5E-4	5.32
GO:0000018	Regulation of DNA recombination	50	4	4.99E-4	10.83
GO:0050671	Positive regulation of lymphocyte proliferation (a)	101	5	9.05E-4	6.70
GO:0032946	Positive regulation of mononuclear cell proliferation (a)	102	5	9.46E-4	6.64
GO:0070663	Regulation of leukocyte proliferation (a)	153	6	9.48E-4	5.31

Biological processes marked by the same letter in parenthesis (a) were clustered together using REVIGO with medium similarity (Supek et al. 2011).

Table 7. Enrichment analysis for putative genes under selection in relation with altitude in Moroccan goats.

GO term	Biological process	Number of genes associated	Number of candidate genes associated	P-value	Enrichment
GO:0060306	Regulation of membrane repolarization (a)	25	4	4.15E-4	11.12
GO:0034765	Regulation of ion transmembrane transport (a)	309	13	5.44E-4	2.92
GO:0086067	AV node cell to bundle of His cell communication (b)	3	2	6.12E-4	46.34
GO:0060486	Clara cell differentiation	3	2	6.12E-4	46.34
GO:0003015	Heart process (b)	13	3	7.55E-4	16.04
GO:0086005	Ventricular cardiac muscle cell action potential (b)	13	3	7.55E-4	16.04
GO:0060047	Heart contraction (b)	13	3	7.55E-4	16.04
GO:0034762	Regulation of transmembrane transport (a)	323	13	8.21E-4	2.80

Biological processes marked by the same letter in parenthesis (a) or (b) were clustered together using REVIGO with medium similarity (Supek et al. 2011).

Bases of local adaptation in sheep and goats

This very weak population structure was particularly suitable for identifying selective sweeps likely associated to local adaptation, avoiding possible confusions associated with demography. We used a population-based approach to identify selective sweeps linked to environmental conditions (i.e. altitude, temperature, humidity and slope). We used a stringent approach based on a haplotype-based method (i.e. XP-CLR) and a single variant differentiation *Fst* to identify selective sweeps. It allowed us to identify first candidate variants and then the associated candidate genes. Furthermore, we ran a correlative method using all environmental parameters available (after having discard the highly correlated ones).

At the inverse of the correlative approach, the population-based one identified, for each environmental parameter, several sets of candidate variants and genes (29.905 variants inter-genic or associated to 707 genes for five environmental variables in sheep and 14.279 variants inter-genic or associated to 607 genes for five variables in goats). Generally, inter-genic and intronic variants represented the largest part of these candidate variants (about 60% and 30% respectively). Similar findings have been reported by several previous studies looking for selective sweeps, e.g. (Ai et al. 2015). Non-coding variants could highly influence and regulate gene transcription and thus phenotypes via diverse known and unknown mechanisms (Ward and Kellis 2012). They could be a part of sequences regulating translation, stability and localization (i.e. un-translated regions), or of promoter regions or enhancers that could be very far from the genes they influence (Noonan and McCallion 2010; Dunham et al. 2012). There are a few large-scale projects such as the Encyclopaedia of DNA Elements (Dunham et al. 2012), which released comprehensive maps of chromatin states, transcription factor binding and transcription for a selection of cell lines and DNase maps for many primary cells in humans (Dunham et al. 2012). However, our understanding of the functional non-coding variation is still far from being complete and defining a complete regulatory annotation on a genome-wide scale is still unattainable. Therefore, our findings, which show that candidate variants are mostly noncoding and that some highest selective sweeps cover only inter-genic regions (e.g. Table S2 and S3), suggest that adaptation to the environmental parameters we are studying would be partly controlled by several regulatory elements.

However, several of the identified candidate variants were within protein-coding genes and some of them were missense, which would have an understandable biochemical effect.

As described before, several hundreds of candidate genes were identified under positive selection in each species using the population-based approach. We used those sets of genes to investigate biological processes enriched for each environmental variable considered. Then, for each one of those variables, we investigated roles of the eight top-candidate genes based on XP-CLR scores (i.e. presented in Table 3) and the whole sets of enriched biological terms. Possible adaptive roles of several genes and enriched biological terms were not easily hypothesised for the corresponding environmental variable. However, many genes and biological processes displayed a likely direct role in the corresponding adaptation. Therefore, we limited our discussion to those genes and terms, although we recognize that our approach could miss several adaptive mechanisms that could be of high interest.

Adaptation to altitude in goats

The enrichment of the GO term associated with Clara cell differentiation for altitude in goats is consistent with the nature of these cells, which are epithelial on the luminal surface of airways of the mammalian lung (Massaro et al. 1994). In addition to their secretory and xenobiotic roles (Serabjitsingh et al. 1980), they are the progenitor cells in small pulmonary airways (Giangreco et al. 2002). They were shown to be numerous and prominent with big apical caps in llama living at high altitude (Heath et al. 1976). They presented also signs of pathological alteration and marks of their compensatory proliferation after exposure to hypoxia in rabbits (Uhlik et al. 2005). Genes that were involved in the enrichment of this GO category included *NFIB* and *GATA6*. The first gene was also identified as a top candidate gene for adaptation of sheep to altitude (Table S2) and it is essential for lung maturation in mice (Steele-Perkins et al. 2005). This would support an important role of these genes in the protection of the highlander goats (and *NFIB* gene in highlander sheep) against possible damages caused by hypoxia conditions in the epithelium of bronchioles. Gene Ontology analysis identified also an over-representation of genes involved in ventricular cardiac muscle cell action potential (GO:0086005), heart process and contraction and AV node cell to bundle of His cell communication (GO:0003015; GO:0060047; GO:0086067) and regulation of trans-membrane transport (GO:0034762; GO:0060306; GO:0034765). Differentiation of the action potentials allows the different electrical characteristics of the different portions of the heart and it was previously demonstrated that chronic high-altitude exposure induces an increase in the size of the right ventricular cells in rats. Hypertrophied cells showed prolongation of action potential (AP) (Chouabe et al. 1997). The enrichment of this GO term for altitude in goats is consistent with a possible role of the candidate genes identified

(*NEDD4L*, *SCN5A* and *GJA5*) in the prolongation of ventricular AP during ischemia or lack of oxygen in high altitude hypoxia as described by Zhou et al. (2015) who reported this effect in experimental conditions in rats. Similarly, GO term related to heart contraction and AV node cell to bundle of His cell communication are consistent with a likely response to hypoxia. The AV node is a part of the electrical conduction system of the heart located at the centre, in the floor of the right atrium, between the atria and ventricles. It takes the signal from the Sinus Node, slows and regulates it, and then sends the electrical impulses from the atria to the ventricles (bundle of His) (James and Spence 1966). Hypoxia generally decreases the amplitude of action potentials of the AV node as shown in rabbits (Senges et al. 1979); (Kohlhardt and Haap 1980); (Hirata 1990). Our findings therefore support a better regulation response of the AV node of highlander goats to oxygen deprivation in comparison with low altitude goats. Besides, the over-representation of genes associated with heart process would also be related to a possible role of goat heart metabolism to respond to oxygen deprivation and to limit damage induced by hypoxia. Such a case was reported by (Calmettes et al. 2010) who demonstrated a high elasticity of ATP production in rat hearts adapted to Chronic Hypoxia, compared to controls measured under low oxygen perfusion. This elasticity induced an improved response of energy supply to cellular energy demand. Finally, the significant enrichment of GO terms associated with regulation of trans-membrane transport is consistent with the role of the inward Na-Ca in increasing duration of the low plateau of rat ventricular AP in altitude cardiac hypertrophy described by (Espinosa et al. 2000). We hypothesize a similar mechanism in goats occurring in high altitude.

Therefore, GO term enrichment analysis for altitude in goats showed the existence of adaptive paths involving the functioning of heart and lung, which represent the main organs helping to face oxygen shortage.

Adaptation to altitude in sheep

In sheep, enriched GO terms for altitude concern mainly the regulation of leucocyte, lymphocyte and mononuclear proliferation (GO:0070665; GO:0050671; GO:0032946; GO:0070663). Indeed, leukocyte invasion into hypoxic tissues is well-known and circulating monocytes and/or mononuclear fibrocytes are recruited to the pulmonary circulation of chronically hypoxic animals. These cells play an important role to face the pulmonary hypertensive process in response to low-input oxygen conditions (Stenmark et al. 2005). This suggests that regulation of leukocyte, lymphocyte and mononuclear

proliferation would be implied in sheep adaptation to high-altitude and genes enriched in these categories (*CLCF1*, *TMIGD2*, *ZP4*, *TLR4*, *KITLG* and *EBI3*) may play a certain role in this adaptation through the mechanism cited above. However, these categories may also display a possible adaptation of sheep in low altitude to face possible pathogens that could be dominant in lowland environment. Further investigations on prevalence of pathogens in Moroccan lowlands would depict this possible involvement.

Another enriched GO term in altitudinal sheep variation was associated with muscle contraction (GO:0006936). The effects of acute or prolonged exposure to hypoxia on human skeletal muscle performance and contractile properties were previously reported (for a review, see (Perrey and Rupp 2009)). This review reported also that the adaptation to chronic hypoxia minimizes the effects on skeletal muscle dysfunction (i.e. impairment during fatigue resistance exercise and in muscle contractile properties). Thus we could predict a possible role of genes enriched in this category (*RYR3*, *ITGB5*, *ARHGEF11*, *TPM1*, *VIPR1*, *ADRBK1* and *P2RX3*) in helping highlander sheep to reduce the impact of chronic hypoxia on skeletal muscle or possibly on other muscle-types (i.e. cardiac and smooth) disturbance.

A top-candidate gene identified for altitude in sheep was *MCM3*, which is one of the mini-chromosome maintenance proteins (*MCM* 2-7). Results showed a higher differentiation in the highlander sheep group (Figure 5). *MCM* proteins are components of a DNA helicase that plays an essential role in DNA replication and cell proliferation (Maiorano et al. 2006). Recently, it was demonstrated that they inhibit *HIF-1* (hypoxia-inducible factor 1) transcriptional activity and thus decrease proliferation in response to hypoxia in many cell types (Hubbi et al. 2011). *HIF-1* was identified under positive selection for adaptation to high-altitude in Tibetans with its paralog *HIF-2* (*EPAS1*) (Beall et al. 2010; Simonson et al. 2010; Yi et al. 2010). Mutations in these genes were associated with haemoglobin concentration. Their role in maintaining oxygen in tissues in hypoxic conditions was thus suggested (Yi et al. 2010). Our findings support a likely implication of *MCM3* in the regulation of one or multiple *HIF* genes in response to hypoxia conditions linked to high-altitude in sheep, and suggest a possible form of adaptive convergence in sheep and humans. Such form hypothesises the involvement of two different genes acting on the same function in these two species.

Other top candidate genes identified in the high-low altitude comparison in sheep include *OXR1* gene. It is a conserved eukaryotic gene that is known to protect yeast and human cells from oxidative damage induced by reactive oxygen species (ROS). It was also identified as a vital protein that controls the sensitivity of neuronal cells to oxidative stress and protects against oxidative stress-induced neurodegeneration (Oliver et al. 2011). We could thus hypothesise a possible implication of that gene linked to the oxygen level at various altitudes. Our analysis identified a missense candidate variation in *PUS7L* gene in sheep for altitude. GO annotations related to this gene include pseudouridine Ψ (an anti-mutagenic and invariant region of tRNA) synthase activity and RNA binding. A study suggested a possible involvement of Ψ in the reduction of chromosome aberrations (dicentric) caused by radiation linked to X-rays and carbon ions (Monobe et al. 2003). Additionally, *PUS7L* was identified under positive selection in both Deedu Mangolians and Tibetan humans who both live in high-altitude environment (Xing et al. 2013). A large number of candidate variants identified for altitude variation in sheep were associated with *GMDS* "GDP-mannose 4,6 dehydratase" (71 intronic variants and 8 downstream). This gene is implied in the de novo biosynthesis of GDP-(L)-fucose. The latter forms part of a number of glycoconjugates, and defects in its metabolism have been associated with leukocyte adhesion deficiency type II in humans (Karsan et al. 1998). However, we could not predict its possible role in such differentiation. We could only speculate a possible involvement in the immune system possibly linked to the environmental context either in low or high altitude. Further investigations would be needed to depict such a role.

Further adaptations: slope and rainfall in goats

The enrichment of GO categories associated with ATP biosynthetic processes (GO:2001171; GO:2001169; GO:1903580) in adaptation of goats to slope is consistent with a higher need for synthesised energy in animals raised in steep slopes (mountainous areas) in comparison with moderate-slope goats. This would make sense knowing that Moroccan mountains goats are generally raised following an extensive system where they move a lot for grazing depending on forage availability. Only *PINK1* and *PID1* were involved in the enrichment of the three categories.

The enrichment of GO categories associated with neutrophil, Granulocyte, leukocyte and cell chemotaxis in goats for rainfall in March supports an important role of the immune system in the protection of goats following chemical variations linked to humidity. The latter is generally associated with a higher prevalence of some pathogens in ruminants, e.g.

Salmonella that cause diarrheic adult goats (Mahmood et al. 2014), Fascioliasis in Buffaloes (Bhutto et al. 2012) or bluetongue over ruminant species (Trebas et al. 2004). Candidate genes associated with these GO categories were *EDN3*, *SYK*, *PDE4D*, *IL1B*, *PDE4B* and *S100A9*.

As mentioned above, other GO terms were significantly enriched in goats and sheep in relation with the other variables but we could not predict the biochemical mechanism underlying the possible adaptation, e.g. positive regulation of filopodium assembly (GO:0051491) associated with slope in sheep or positive regulation of oocyte development (GO:0060282) in goats for the same variable.

Adaptive convergence

The dissimilar GO categories and the very low percentage of candidate genes identified simultaneously in sheep and goats for the same environmental variable (<1% for altitude, temp7 and slope) would support generally different adaptive mechanisms in sheep and goats for these three variables. Similarly, our identified candidate genes in the two species for adaptation to altitude are different from those found in Chinese pigs (Ai et al. 2015). However, this could also be associated to the use of different methods/approaches to detect selective sweeps. Interestingly, the regulatory relation between genes *MCM3* we associated here with sheep adaptation to altitude and *NIF* widely reported to be under selection in Tibetans (Simonson et al. 2010; Yi et al. 2010; Daub et al. 2013) would illustrate an interesting adaptive convergence case in humans and sheep based not necessarily on the same genes but on genes likely associated to the same biochemical action.

Differences between population-based and correlative approaches

Our findings showed that population-based and correlative approaches did not detect similar selection signatures for the same environmental variable and species except for some cases (6 SNPs in sheep and 10 in goats; Tables 4 and 5). The limited number of candidate variants detected by our correlative approach favoured this as well as the way in which both approaches work. We hypothesize that the population structure, even weak could have an impact on the correlative method robustness.

Finally, one limitation for identifying genes and metabolic pathways under selection was due to the fact that several annotated genes in the sheep and goat genomes were not identified and do not have known orthologs in other species (e.g. gene names starting with 'LOC' in goats).

Conclusion

Our study used a landscape genomic framework to depict the genetic bases of local adaptation in farm animals. The 321 sheep and goat whole genome sequences and from a wide range of biotic and abiotic conditions represent a unique resource for studying evolutionary processes. We identified several sets of candidate variants, genes and biological processes that are likely involved in local adaptation to various eco-climatic conditions. We could show the variation of genetic differentiation over environmental gradients according to several different patterns. Therefore, this study showed the likely effect of local adaptation on genomes not only in contrasted environments but also over a continuous environmental gradient in two livestock species. This contributes to our understanding on how local adaptation could act and opened new horizons for better understanding how genetic diversity is distributed and how it can be a valuable resource for conservation purposes.

Accession numbers

The variant calls and genotype calls used in this paper are archived in the European Variation Archive with accession ID ERZ019290 for sheep and ID ERZ020631 For goats. The data are accessible at ftp://ftp.ebi.ac.uk/pub/databases/nextgen/

References

Ai H, Fang X, Yang B, Huang Z, Chen H, Mao L, Zhang F, Zhang L, Cui L, He W et al. 2015. Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nature Genetics* **47**(3): 217-+.

Ai H, Huang L, Ren J. 2013. Genetic Diversity, Linkage Disequilibrium and Selection Signatures in Chinese and Western Pigs Revealed by Genome-Wide SNP Markers. *Plos One* **8**(2).

Altshuler DM Durbin RM Abecasis GR Bentley DR Chakravarti A Clark AG Donnelly P Eichler EE Flicek P Gabriel SB et al. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422): 56-65.

Beall CM, Cavalleri GL, Deng L, Elston RC, Gao Y, Knight J, Li C, Li JC, Liang Y, McCormack M et al. 2010. Natural selection on EPAS1 (HIF2 alpha) associated with low hemoglobin concentration in Tibetan highlanders. *Proceedings of the National Academy of Sciences of the United States of America* **107**(25): 11459-11464.

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**(1): 289-300.

Benjelloun B, Alberto FJ, Streeter I, Boyer F, Coissac E, Stucki S, BenBati M, Ibnelbachyr M, Chentouf M, Bechchari A et al. 2015. Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (Capra hircus) using WGS data. *Frontiers in genetics* **6**: 107-107.

Bergelson J, Roux F. 2010. Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. *Nature Reviews Genetics* **11**(12): 867-879.

Bhutto B, Arijo A, Phullan MS, Rind R. 2012. Prevalence of Fascioliasis in Buffaloes under Different Agro-climatic Areas of Sindh Province of Pakistan. *International Journal of Agriculture and Biology* **14**(2): 241-245.

Böhner J, McCloy KR, Strobl J. 2006. SAGA – Analysis and Modelling Applications. *Göttinger Geographische Abhandlungen*, **115**, 130.

Browning BL, Browning SR. 2013. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. *Genetics* **194**(2): 459-+.

Calmettes G, Deschodt-Arsac V, Gouspillou G, Miraux S, Muller B, Franconi J-M, Thiaudiere E, Diolez P. 2010. Improved Energy Supply Regulation in Chronic Hypoxic Mouse Counteracts Hypoxia-Induced Altered Cardiac Energetics. *Plos One* **5**(2).

Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. *Genome Research* **20**(3): 393-402.

Chevin L-M, Lande R, Mace GM. 2010. Adaptation, Plasticity, and Extinction in a Changing Environment: Towards a Predictive Theory. *Plos Biology* **8**(4).

Chouabe C, Espinosa L, Megas P, Chakir A, Rougier O, Freminet A, Bonvallet R. 1997. Reduction of I-Ca,I-L and I(to1)density in hypertrophied right ventricular cells by simulated high altitude in adult rats. *Journal of Molecular and Cellular Cardiology* **29**(1): 193-206.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**(15): 2156-2158.

Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-Murci L, Robinson-Rechavi M, Excoffier L. 2013. Evidence for Polygenic Adaptation to Pathogens in the Human Genome. *Molecular Biology and Evolution* **30**(7): 1544-1558.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**(5): 491-+.

Dionne M, Caron F, Dodson JJ, Bernatchez L. 2008. Landscape genetics and hierarchical genetic structure in Atlantic salmon: the interaction of gene flow and local adaptation. *Molecular Ecology* **17**(10): 2382-2396.

Dong Y, Xie M, Jiang Y, Xiao N, Du X, Zhang W, Tosser-Klopp G, Wang J, Yang S, Liang J et al. 2013. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). *Nature Biotechnology* **31**(2): 135-141.

Dunham I Kundaje A Aldred SF Collins PJ Davis C Doyle F Epstein CB Frietze S Harrow J Kaul R et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414): 57-74.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *Bmc Bioinformatics* **10**.

Espinosa L, Chouabe C, Morales A, Lachuer J, Georges B, Fatemi M, Terrenoire C, Tourneur Y, Bonvallet R. 2000. Increased sodium-calcium exchange current in right ventricular cell hypertrophy induced by simulated high altitude in adult rats. *Journal of Molecular and Cellular Cardiology* **32**(4): 639-653.

Fournier-Level A, Korte A, Cooper MD, Nordborg M, Schmitt J, Wilczek AM. 2011. A Map of Local Adaptation in Arabidopsis thaliana. *Science* **334**(6052): 86-89.

Franks SJ, Hoffmann AA. 2012. Genetics of Climate Change Adaptation. *Annual Review of Genetics, Vol 46* **46**: 185-208.

Frichot E, Mathieu F, Trouillon T, Bouchard G, Francois O. 2014. Fast and Efficient Estimation of Individual Ancestry Coefficients. *Genetics* **196**(4): 973-+.

Frichot E, Schoville SD, Bouchard G, Francois O. 2013. Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution* **30**(7): 1687-1699.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. Vol 1207.3907v2. arXiv.

Giangreco A, Reynolds SD, Stripp BR. 2002. Terminal bronchioles harbor a unique airway stem cell population that localizes to the bronchoalveolar duct junction. *American Journal of Pathology* **161**(1): 173-182.

Heath D, Smith P, Harris P. 1976. CLARA CELLS IN LLAMA. *Experimental Cell Biology* **44**(2): 73-82.

Hirata A. 1990. EFFECT OF HYPOXIA ON ELECTRICAL-ACTIVITY OF ATRIOVENTRICULAR NODAL CELLS AND ATRIAL CELLS OF THE RABBITS HEART. *Journal of Electrocardiology* **23**(1): 69-76.

Hubbi ME, Luo W, Baek JH, Semenza GL. 2011. MCM Proteins Are Negative Regulators of Hypoxia-Inducible Factor 1. *Molecular Cell* **42**(5): 700-712.

James TN, Spence CA. 1966. DISTRIBUTION OF CHOLINESTERASE WITHIN SINUS NODE AND AV NODE OF HUMAN HEART. *Anatomical Record* **155**(2): 151-&.

Jiang Y, Xie M, Chen W, Talbot R, Maddox JF, Faraut T, Wu C, Muzny DM, Li Y, Zhang W et al. 2014. The sheep genome illuminates biology of the rumen and lipid metabolism. *Science* **344**(6188): 1168-1173.

Jombart T, Ahmed I. 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**(21): 3070-3071.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**(7392): 55-61.

Joost S, Bonin A, Bruford MW, Despres L, Conord C, Erhardt G, Taberlet P. 2007. A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology* **16**(18): 3955-3969.

Karsan A, Cornejo CJ, Winn RK, Schwartz BR, Way W, Lannir N, Gershoni-Baruch R, Etzioni A, Ochs HD, Harlan JM. 1998. Leukocyte Adhesion Deficiency Type II is a generalized defect of de novo GDP-fucose biosynthesis - Endothelial cell fucosylation is not required for neutrophil rolling on human nonlymphoid endothelium. *Journal of Clinical Investigation* **101**(11): 2438-2445.

Kawecki TJ, Ebert D. 2004. Conceptual issues in local adaptation. *Ecology Letters* **7**(12): 1225-1241.

Kohlhardt M, Haap K. 1980. THE INFLUENCE OF HYPOXIA AND METABOLIC-INHIBITORS ON THE EXCITATION PROCESS IN ATRIOVENTRICULAR NODE. *Basic Research in Cardiology* **75**(6): 712-727.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data P. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Loarie SR, Duffy PB, Hamilton H, Asner GP, Field CB, Ackerly DD. 2009. The velocity of climate change. *Nature* **462**(7276): 1052-U1111.

Macnair MR. 1993. THE GENETICS OF METAL TOLERANCE IN VASCULAR PLANTS. *New Phytologist* **124**(4): 541-559.

Mahmood AK, Khan MS, Khan MA, Bilal M. 2014. PREVALENCE OF SALMONELLA IN DIARRHEIC ADULT GOATS IN FIELD CONDITIONS. *Journal of Animal and Plant Sciences* **24**(1): 98-102.

Maiorano D, Lutzmann M, Mechali M. 2006. MCM proteins and DNA replication. *Current Opinion in Cell Biology* **18**(2): 130-136.

Manel S, Holderegger R. 2013. Ten years of landscape genetics. *Trends in Ecology & Evolution* **28**(10): 614-621.

Manel S, Schwartz MK, Luikart G, Taberlet P. 2003. Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution* **18**(4): 189-197.

Massaro GD, Singh G, Mason R, Plopper CG, Malkinson AM, Gail DB. 1994. BIOLOGY OF THE CLARA CELL. *American Journal of Physiology* **266**(1): L101-L106.

McCue ME, Bannasch DL, Petersen JL, Gurr J, Bailey E, Binns MM, Distl O, Guerin G, Hasegawa T, Hill EW et al. 2012. A High Density SNP Array for the Domestic Horse and Extant Perissodactyla: Utility for Association Mapping, Genetic Diversity, and Phylogeny Studies. *Plos Genetics* **8**(1).

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**(9): 1297-1303.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**(16): 2069-2070.

Monobe M, Arimoto-Kobayashi S, Ando K. 2003. beta-pseudouridine, a beer component, reduces radiation-induced chromosome aberrations in human lymphocytes. *Mutation Research-Genetic Toxicology and Environmental Mutagenesis* **538**(1-2): 93-99.

Naderi S, Rezaei H-R, Pompanon F, Blum MGB, Negrini R, Naghash H-R, Balkiz O, Mashkour M, Gaggiotti OE, Ajmone-Marsan P et al. 2008. The goat domestication process inferred from large-scale mitochondrial DNA analysis of wild and domestic individuals. *Proceedings of the National Academy of Sciences of the United States of America* **105**(46): 17659-17664.

New M, Lister D, Hulme M, Makin I. 2002. A high-resolution data set of surface climate over global land areas. *Climate Research* **21**(1): 1-25.

Noonan JP, McCallion AS. 2010. Genomics of Long-Range Regulatory Elements. *Annual Review of Genomics and Human Genetics, Vol 11* **11**: 1-24.

Oliver PL, Finelli MJ, Edwards B, Bitoun E, Butts DL, Becker EBE, Cheeseman MT, Davies B, Davies KE. 2011. Oxr1 Is Essential for Protection against Oxidative Stress-Induced Neurodegeneration. *Plos Genetics* **7**(10).

Palti Y, Gao G, Liu S, Kent MP, Lien S, Miller MR, Rexroad CE, III, Moen T. 2015. The development and characterization of a 57K single nucleotide polymorphism array for rainbow trout. *Molecular Ecology Resources* **15**(3): 662-672.

Pereira F, Queiros S, Gusmao L, Nijman IJ, Cuppen E, Lenstra JA, Davis SJM, Nejmeddine F, Amorim A, Econogene C. 2009. Tracing the History of Goat Pastoralism: New Clues from Mitochondrial and Y Chromosome DNA in North Africa. *Molecular Biology and Evolution* **26**(12): 2765-2773.

Perrey S, Rupp T. 2009. Altitude-Induced Changes in Muscle Contractile Properties. *High Altitude Medicine & Biology* **10**(2): 175-182.

Savolainen O, Lascoux M, Merila J. 2013. Ecological genomics of local adaptation. *Nature Reviews Genetics* **14**(11): 807-820.

Senges J, Mizutani T, Pelzer D, Brachmann J, Sonnhof U, Kubler W. 1979. EFFECT OF HYPOXIA ON THE SINOATRIAL NODE, ATRIUM, AND ATRIOVENTRICULAR NODE IN THE RABBIT HEART. *Circulation Research* **44**(6): 856-863.

Serabjitsingh CJ, Wolf CR, Philpot RM, Plopper CG. 1980. CYTOCHROME-P-450 - LOCALIZATION IN RABBIT LUNG. *Science* **207**(4438): 1469-1470.

Simonson TS, Yang Y, Huff CD, Yun H, Qin G, Witherspoon DJ, Bai Z, Lorenzo FR, Xing J, Jorde LB et al. 2010. Genetic Evidence for High-Altitude Adaptation in Tibet. *Science* **329**(5987): 72-75.

Steele-Perkins G, Plachez C, Butz KG, Yang GH, Bachurski CJ, Kinsman SL, Litwack ED, Richards LJ, Gronostajski RM. 2005. The transcription factor gene Nfib is essential for both lung maturation and brain development. *Molecular and Cellular Biology* **25**(2): 685-698.

Stenmark KR, Davie NJ, Reeves JT, Frid MG. 2005. Hypoxia, leukocytes, and the pulmonary circulation. *Journal of Applied Physiology* **98**(2): 715-721.

Stucki S, Orozco-terWengel P, Bruford MW, Colli L, Masembe C, Negrini R, Taberlet P, Joost S. 2014. High performance computation of landscape genomic models integrating local indices of spatial association. *arxiv* **1405.7658v2**.

Supek F, Bosnjak M, Skunca N, Smuc T. 2011. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *Plos One* **6**(7).

Taberlet P, Valentini A, Rezaei HR, Naderi S, Pompanon F, Negrini R, Ajmone-Marsan P. 2008. Are cattle, sheep, and goats endangered species? *Molecular Ecology* **17**(1): 275-284.

Uhlik J, Konradova V, Vajner L, Adaskova J. 2005. Normobaric hypoxia induces mild damage to epithelium of terminal bronchioles in rabbits (ultrastructural study). *Veterinarni Medicina* **50**(10): 432-438.

Veroneze R, Lopes PS, Guimaraes SEF, Silva FF, Lopes MS, Harlizius B, Knol EF. 2013. Linkage disequilibrium and haplotype block structure in six commercial pig lines. *Journal of Animal Science* **91**(8): 3493-3501.

Villa-Angulo R, Matukumalli LK, Gill CA, Choi J, Van Tassell CP, Grefenstette JJ. 2009. High-resolution haplotype block structure in the cattle genome. *Bmc Genetics* **10**.

Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR et al. 2009. Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse. *Science* **326**(5954): 865-867.

Ward LD, Kellis M. 2012. Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology* **30**(11): 1095-1106.

Weir BS, Cockerham CC. 1984. ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION-STRUCTURE. *Evolution* **38**(6): 1358-1370.

Xing J, Wuren T, Simonson TS, Watkins WS, Witherspoon DJ, Wu W, Qin G, Huff CD, Jorde LB, Ge R-L. 2013. Genomic Analysis of Natural Selection and Phenotypic Variation in High-Altitude Mongolians. *Plos Genetics* **9**(7).

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS et al. 2010. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science* **329**(5987): 75-78.

Zeder MA. 2005. "A view from the Zagros: new perspectives on livestock domestication in the Fertile Crescent," in *The First Steps of Animal Domestication. New Archaeological Approaches*, eds J. D. Vigne, J. Peters, and D. Helmer (Oxford: Oxbow Books), 125–146.

Supplementary Material

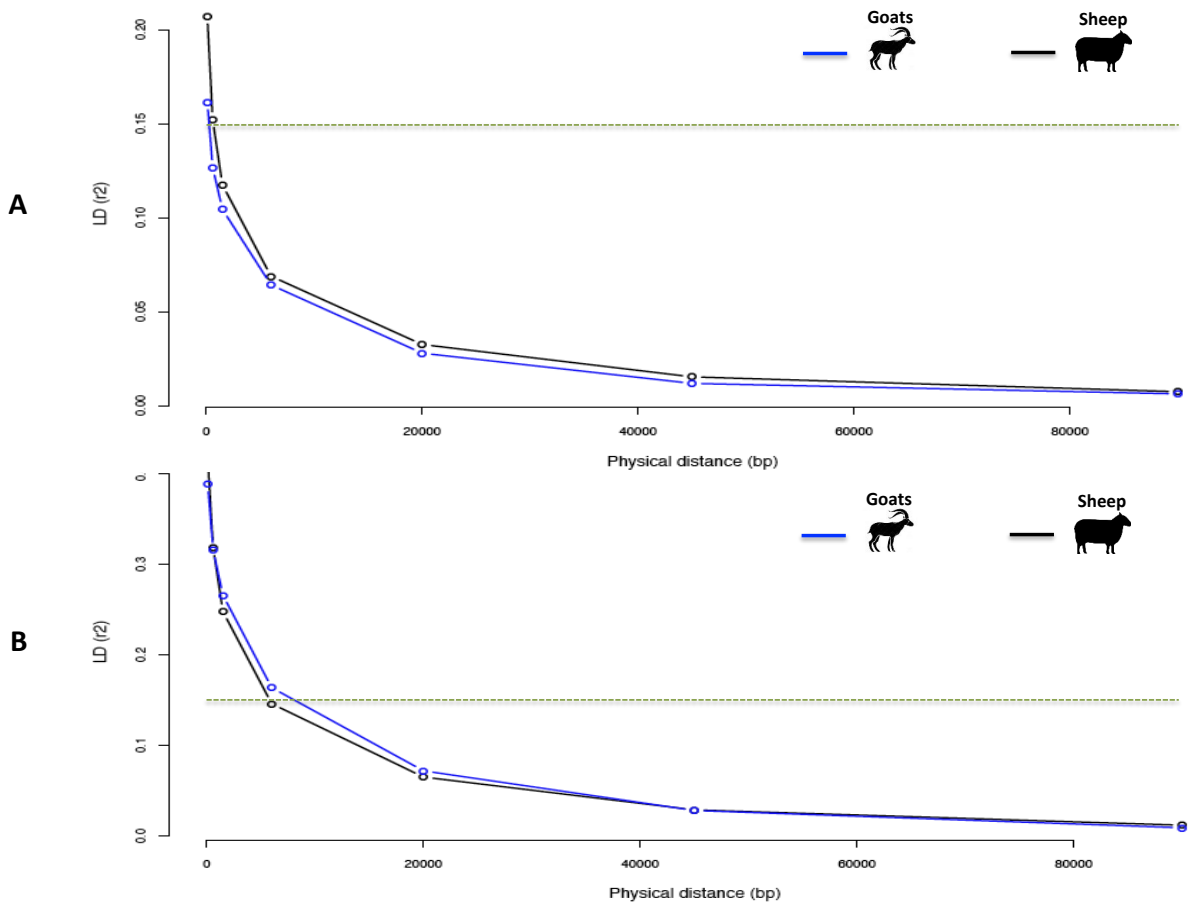

Figure S1. Decay of linkage disequilibrium (r^2) as a function of physical distance in sheep and goats.

The Linkage Disequilibrium (*LD*) was estimated for the 160 sheep and 161 goats on 5 different segments of 2Mb each on 5 different chromosomes either by considering the whole set of reliable variants (A) or by excluding rare variants with MAF<0.05 (B). Inter-variant distances (bp) were binned and averaged into the classes: 0–0.2, 0.2–1, 1–2, 2–10, 10-30, 30-60 and 60-120 kb.

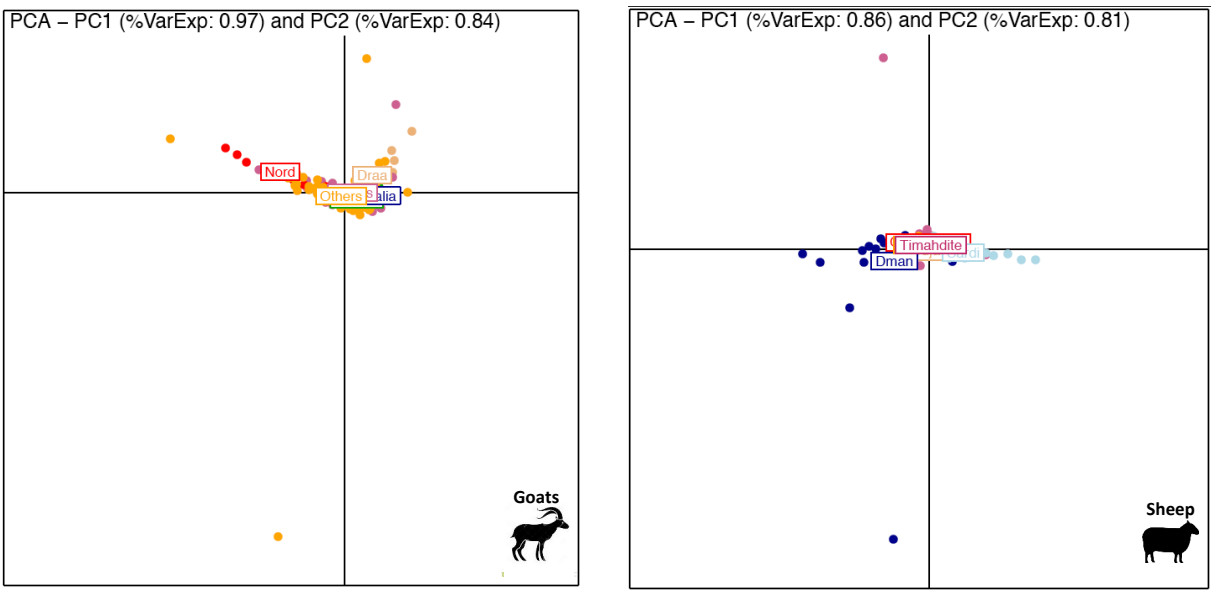

Figure S2: Distribution of the 160 sheep (right) and 161 goats (left) according to the main two principal components (PC).

The variance explained by each PC is mentioned in the top of the graphs. Colours distinguish the phenotypic groups of individuals (i.e. breeds, populations or sub-populations).

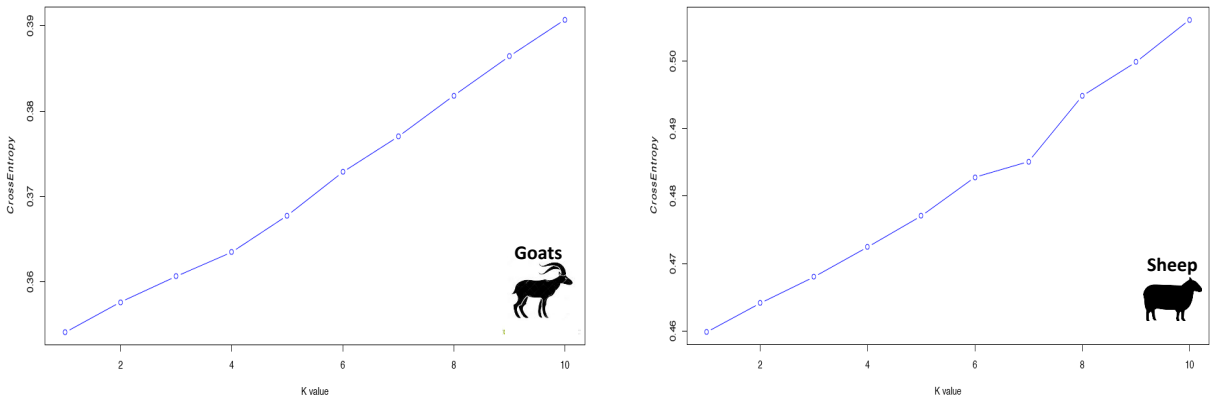

Figure S3. Cross validation coefficients for the different numbers of clusters tested between one and 10.

Each plots shows the CrossValidation coefficient (CV; y-axis) in relation to the different numbers of clusters tested (x-axis). The most likely number of clusters in the dataset is defined for the value of K that shows the smallest CV.

Table S1. Environmental variables studied.

Sheep		Goats	
Selected variable	**Correlated variables |r|>=0.9**	**Selected variable**	**Correlated variables |r|>=0.9**
Altitude	tmin_12, tmin_11, tmin_3, tmin_2, tmin_1, tmean_12, tmean_1, bio_11, bio_6	Altitude	tmin_12, tmin_11, tmin_3, tmin_2, tmin_1, tmean_12, tmean_1, bio_11, bio_6
Northness		Northness	
Easthness		Easthness	
Curvature		Slope	VRM, TWI
Plan Curvature		Curvature	
Profile Curvature		Plan Curvature	
Ruggdness		Profile Curvature	
Slope	Sky View Factor (SVF), Topographic witness index (TWI), Energy200	Sky View Factor (SVF)	Catching_Slope, Energy200
Total Insolation (TI) 21/06	Direct solar radiation (Di) 21/06, Difuse solar radiation (Df) 21/6	TI 21/06	Di 21/06, Df 21/06
TI 21/12	Di 21/12, Df 21/12	TI 21/12	Di 21/12, Df 21/12
Orientation200		Orientation200	
Coherency200		Coherency200	
tmin_5	tmin_9, tmin_8, tmin_7, tmin_6	tmin_8	tmin_9, tmin_7, tmin_6, tmin_5, tmean_5, bio_1
tmin_4	tmin_11, tmin_10, tmin_9, tmin_3, tmin_2, tmean_12, tmean_11, tmean_10, tmean_4	tmean_6	tmean_9, tmean_8, tmean_7, tmean_5, tmax_5, bio_10
tmean_3	tmean_2, tmean_1, tmax_12, tmax_11, bio_11, bio_1	tmean_2	tmin_11, tmin_10, tmin_9, tmin_4, tmin_3, tmin_2, tmean_12, tmean_11
tmean_9	tmean_6, tmax_10, bio_1	tmean_10	tmean_4, tmean_3, tmean_1, tmax_12, tmax_11, tmax_2, tmax_1
tmean_5	tmin_8, tmin_7, tmean_6	prec_9	bio_17, bio_11, bio_1
tmax_9	tmean_8, tmean_7, tmax_6, tmax_5, bio_10	tmax_10	tmean_10, tmean_9, tmean_4, tmax_11, tmax_4, tmax_3, tmax_2, tmax_1, bio_1
tmax_8	tmean_7, tmax_7, tmax_6, bio_5	tmax_7	tmax_8, tmax_6, bio_5
tmax_4	tmax_10, tmax_5, tmax_3	prec_8	
prec_10	prec_11, prec_3, bio_16, bio_13, bio_12	prec_6	prec_5, bio_17
prec_9	tmax_2, tmax_1, bio_18, bio_17	prec_1	prec_12, prec_11, prec_4, prec_3, prec_2, bio_19, bio_16, bio_13, bio_12
prec_8		bio_15	
prec_6	prec_5, bio_17	bio_14	prec_7, bio_17
prec_4	prec_5, prec_3, prec_2, bio_12	bio_9	
bio_15		bio_8	
bio_14	prec_7, bio_17	bio_7	bio_4, bio_2
bio_9		bio_3	
bio_8			
bio_7	bio_4, bio_2		
bio_3			

Selected variables are those included in the correlative approach and correlated variables are those withdrawn because of their high correlation with the selected ones. Prec, tmin, tmax and tmean are rainfall, lower temperature, higher temperature and mean temperature respectively for each month specified from 1 to 12. Variables starting with "Bio" are various bioclimatic variables derived from temperature and rainfall and presented in http://worldclim.org/bioclim. Other names represent various DEM-derived variables.

Table S2. Candidate variants and genes associated with the 20 top-XP-CLR signals identified in each analysis associated with local adaptation of Moroccan sheep to altitude.

High altitude						
Gene	**Variant type**	**Number of variants**	**Rank**	**Score**	**Chr.**	**Overlapping Top scores**
GMDS	Downstream	8	1	127	20	69
	Intron	71				
-	Inter-genic	130				
-	Inter-genic	4	2	122	8	50
-	Inter-genic	4	4	91.5	2	5
MCM3	Downstream	2	5	86.9	20	5
	Intron	22				
	Splice region/intron	1				
	Synonymous	4				
	Upstream	2				
-	Inter-genic	16				
-	Inter-genic	1	7	71.4	4	6
-	Inter-genic	14	8	71.3	1	10
ASRGL1	Intron	3	9	70.6	21	17
-	Inter-genic	1				
OXR1	Intron	81	12	68.1	9	6
-	Inter-genic	12	14	64.5	1	10
ZNF831	Intron	25	15	64.1	13	17
-	Inter-genic	42	16	62.5	10	6
ENSOARG00000021651	Downstream	1	17	62.0	3	10
	Upstream	2				
-	Inter-genic	17				
BIRC6	Intron	43	18	60.0	3	5
	Splice region/intron	2				
	Synonymous	2				
IFT88	Intron	1	19	59.0	10	50
ENSOARG00000004678	Downstream	1	20	58.8	17	7
RNFT2	Downstream	2				
-	Inter-genic	24				

Low altitude						
Gene	**Variant type**	**Number of variants**	**Rank**	**Score**	**Chr.**	**Overlapping scores**
NFIB	Downstream	1	1	92.8	2	33
	Intron	8				
-	Inter-genic	12	2	69.7	25	10
-	Inter-genic	18	3	69.6	1	19
GMDS	Downstream	8	4	68.1	20	40
	Intron	71				
-	Inter-genic	120				
-	Inter-genic	12	5	67.8	2	7
U6	Downstream	5	7	66.6	16	7
	Upstream	4				
-	Inter-genic	5				
U6	Downstream	3	9	63.1	5	6
	Upstream	10				
-	Inter-genic	22				
-	Inter-genic	10	12	59.8	7	6
IRAK4	Downstream	1	13	58.7	3	12
	Intron	6				
	Upstream	4				
PUS7L	Intron	2				
	Missense	2				
	Upstream	8				
TMEM117	Intron	1				
TWF1	Downstream	3				
-	Inter-genic	16				
-	Inter-genic	1	14	57,3	7	21
-	Inter-genic	71	16	56,1	3	3
-	Inter-genic	18	18	52,6	3	5
-	Inter-genic	27	20	51,5	20	12

For each analysis, in each window associated with the 20 higher XP-CLR signals, variants marked by an *Fst* higher than the 0.1% genome-wide threshold were classified in different inter-genic and genic categories. Rank represents the autosomal-wide rank of the corresponding XP-CLR signal based on its higher score. "Score" represent the higher XP-CLR score identified for that XP-CLR signal. "Overlapping top-scores" represents the number of overlapping windows marked by an XP-CLR score higher than the 0.1% autosomal-wide threshold.

Table S3. Candidate variants and genes associated with the 20 top-XP-CLR signals identified in each analysis associated with local adaptation of Moroccan goats to rainfall seasonality "bio15".

High "bio15"						
Gene	**Variant type**	**Number of variants**	**Rank**	**Score**	**Chr.**	**Overlapping signals**
-	Inter-genic	56	1	127	8	21
CNTN5	5' UTR	1	2	112	15	21
	Intron	35				
CDH2	Intron	28	3	105	24	14
	Upstream gene	1				
-	Inter-genic	11				
-	Inter-genic	17	4	99	9	25
LOC102175357	Missense	1	5	69.4	26	30
	Upstream gene	1				
UBE2D1	Intron	2				
-	Inter-genic	10				
LYRM9	3' UTR	18	6	64.4	19	2
	5' UTR	1				
	Downstream gene	10				
	Intron	27				
	Upstream gene	4				
-	Inter-genic	17				
RRP36	5' UTR	1	7	62.2	23	11
	Intron	1				
	Upstream gene	6				
PPP2R5D	Downstream gene	3				
MRPL2	Downstream gene	1				
	Intron	1				
CUL7	Intron	1				
KLC4	Downstream gene	1				
-	Inter-genic	2				
-	Inter-genic	52	8	61.2	21	5
-	Inter-genic	25	9	61.1	24	10
-	Inter-genic	103	10	60.7	6	14
MS4A13	Downstream gene	4	11	57.3	15	7
	Intron	12				
	Upstream gene	1				
-	Inter-genic	5				
LOC102190926	Intron	72	13	50.5	3	3
ARHGEF38	Intron	60	14	50.2	6	1
-	Inter-genic	17	15	49.9	11	7
-	Inter-genic	3	16	48.2	9	2
-	Inter-genic	5	18	47.3	9	7
-	Inter-genic	15	19	46.0	20	11
LOC102186025	Downstream gene	2	20	45.9	11	10

Low "bio15"						
Gene	**Variant type**	**Number of variants**	**Rank**	**Score**	**Chr.**	**Overlap signa**
-	Inter-genic	140	1	318	6	60
-	Inter-genic	2	2	126	17	38
-	Inter-genic	123	3	114	6	25
RNASE4	Intron	3	5	96	10	22
LOC102175300	Upstream gene	1				
INPP4A	Intron	5	6	89.2	11	26
	Upstream gene	10				
-	Inter-genic	21				
-	Inter-genic	25	7	78.3	14	25
SAPS3	3' UTR	1	8	78.1	29	18
	Downstream gene	1				
	Intron	6				
	Splice region/Intron	1				
PRR12	Intron	3	9	76.0	18	16
-	Inter-genic	10	11	71.4	7	8
ARHGEF38	Intron	60	13	70.4	6	3
DDX50	Intron	2	14	69.0	28	16
KIAA1279	Intron	1				
-	Inter-genic	30	15	68.0	13	2
WDR48	3' UTR	1	16	64.8	22	4
	Downstream gene	6				
	Intron	14				
LOC102177939	Intron	1				
TTC21A	Intron	1				
XIRP1	Intron	1				
-	Inter-genic	4				
DPY19L4	Intron	44	18	60.0	14	19
ESRP1	Intron	1				
AK3	Upstream gene	1	19	55.2	8	12
RCL1	Intron	10				
-	Inter-genic	9				
DSG2	5' UTR	1	20	55.1	24	9
	Intron	34				

	Intron	1		
	Upstream gene	3		
POMT1	3' UTR	1		
	Intron	7		
RAPGEF1	Downstream gene	1		
	Intron	4		
UCK1	5' UTR	1		
	Downstream gene	6		
	Synonymous/3' UTR	1		
-	Inter-genic	3		

For each analysis, in each window associated with the 20 higher XP-CLR signals, variants marked by an *Fst* higher than the 0.1% genome-wide threshold were classified in different inter-genic and genic categories. Rank represents the autosomal-wide rank of the corresponding XP-CLR signal based on its higher score. "Score" represent the higher XP-CLR score identified for that XP-CLR signal. "Overlapping top-scores" represents the number of overlapping windows marked by an XP-CLR score higher than the 0.1% autosomal-wide threshold.

Table S4. Candidate variants/genes identified by multi-resolution analysis with Samßada in Moroccan sheep.

Environmental variable	Chr	Position	SNP type	Gene	Detection by XP-CLR
prec_4	23	43794976	Intron	*LDLRAD4*	Yes
	23	43812782	Intron	*FAM210A*	Yes
	23	43847594	Intron	*RNMT*	Yes
	23	43861704	Inter-genic	-	Yes
	23	43874160	Downstream	*MC1R*	Yes
	23	44038684	Inter-genic	-	Yes
	23	44084253	Inter-genic	-	No
catch_slope	10	13537408	Inter-genic	-	No
	20	50510912	Inter-genic	-	No
TI2112	5	70648057	Inter-genic	-	-
TI216	5	60766984	Inter-genic	-	-
bio_14	19	2170224	Inter-genic	-	-
	7	48256781	Intron	*RNF111*	-
	7	48262822	Intron	*RNF111*	-
bio_15	1	38304177	Inter-genic	-	-
bio_3	18	11745070	Downstream	*MCTP2*	-
	2	66041147	Inter-genic	-	-
bio_7	14	875672	Intron	*VAC14*	No
bio_8	18	60885624	Inter-genic	-	-
prec_10	23	43794976	Intron	*LDLRAD4*	-
	23	43812782	Intron	*FAM210A*	-
prec_8	19	2167574	Inter-genic	-	-
	19	2170224	Inter-genic	-	-
	7	48256781	Intron	*RNF111*	-
	7	48262822	Intron	*RNF111*	-
prec_9	19	2162818	Inter-genic	-	-
	19	2167574	Inter-genic	-	-
	19	2170224	Inter-genic	-	-
	7	48256781	Intron	*RNF111*	-
	7	48262822	Intron	*RNF111*	-
tmax_4	1	190582	Intron	*DTYMK*	-
	23	43794976	Intron	*LDLRAD4*	-
	23	43812782	Intron	*FAM210A*	-
	23	43874160	Downstream	*MC1R*	-
tmax_8	1	190582	Intron	*DTYMK*	-
	14	875672	Intron	*VAC14*	-
	23	43794976	Intron	*LDLRAD4*	-
	23	43812782	Intron	*FAM210A*	-
	23	43874160	Downstream	*MC1R*	-
tmax_9	1	190582	Intron	*DTYMK*	-
	3	211734411	Inter-genic	-	-

Table S5. Candidate variants/genes identified by muli-resolution analysis with Samßada in Moroccan goats.

Environmental variable	Chr	Position	SNP type	Gene	Detection by XP-CLR
Altitude	1	23002164	Inter-genic	-	No
	22	42794456	Intron	PXK	Yes
	28	40372626	Intron	ARHGAP22	No
	6	12207826	Inter-genic	-	Yes
	6	12218302	Inter-genic	-	No
	6	12254244	Inter-genic	-	Yes
	6	12259667	Inter-genic	-	Yes
	6	25849772	Inter-genic	-	No
	6	26455416	Inter-genic	-	No
Slope	22	41805444	Inter-genic	-	No
	8	46850695	Inter-genic	-	No
bio_15	24	19436980	Inter-genic	-	No
	24	28807953	Inter-genic	-	No
	4	95035251	Intron	EXOC4	No
	6	12187316	Inter-genic	-	Yes
	6	12218302	Inter-genic	-	No
	6	12242353	Inter-genic	-	No
	6	12254244	Inter-genic	-	Yes
	6	12259667	Inter-genic	-	Yes
	6	12276649	Inter-genic	-	Yes
TI216	1	124570528	Intron	TFDP2	-
	1	147737581	Inter-genic	-	-
	6	48842226	Inter-genic	-	-
	7	79661490	Intron	HAPLN1	-
	9	38668915	Inter-genic	-	-
	9	38675107	Inter-genic	-	-
TI2112	3	104936887	Inter-genic	-	-
bio_14	4	95035251	Intron	EXOC4	-
	6	12254244	Inter-genic	-	-
	6	26455416	Inter-genic	-	-
	6	48842226	Inter-genic	-	-
	9	55810891	Intron	EPB41L2	-
bio_7	1	10309616	Inter-genic	-	-
	11	15823825	Inter-genic	-	-
	13	74456761	Inter-genic	-	-
	20	4481114	Inter-genic	-	-
	9	14309947	Intron	NKAIN2	-
bio_8	1	77790195	Inter-genic	-	-
	26	3818155	Inter-genic	-	-
bio_9	12	17515212	Inter-genic	-	-
	14	26405742	Intron	FER1L6	-
northness	15	53255703	Intron	GDPD4	-
	20	25965154	Intron	ITGA2	-
prec_1	1	56842826	Inter-genic	-	-
	14	1335043	Intron	RIMS1	-
	14	45215445	Inter-genic	-	-
	4	25093566	Intron	HDAC9	-
	4	51418120	Intron	FOXP2	-
	6	47914533	Inter-genic	-	-
prec_6	1	74038394	Inter-genic	-	-
prec_8	4	95035251	Intron	EXOC4	-
	6	12254244	Inter-genic	-	-
	6	12259667	Inter-genic	-	-
	6	48842226	Inter-genic	-	-
	7	54532252	Inter-genic	-	-
tmax_10	17	4322848	Intron	TRIM2	-
	20	28161553	Inter-genic	-	-
tmax_7	1	10309616	Inter-genic	-	-
	13	74456761	Inter-genic	-	-
	2	133961081	Inter-genic	-	-
	20	4481114	Inter-genic	-	-
	9	38857907	Inter-genic	-	-
tmean_2	1	59516318	Intron	LSAMP	-
	25	6830009	Inter-genic	-	-
	6	26455416	Inter-genic	-	-
	6	48842226	Inter-genic	-	-
tmean_6	2	133961081	Inter-genic	-	-
tmin_8	11	25021331	Inter-genic	-	-

	11	25030523	Inter-genic	-	-

Table S6. Enrichment analysis for putative genes under selection in relation with slope in Moroccan sheep.

GO term	Biological process	Number of genes associated	Number of candidate genes associated	P-value	Enrichment
GO:0019374	Galactolipid metabolic process	4	2	3.27E-4	67.06
GO:0097264	Self proteolysis	5	2	5.43E-4	53.65
GO:0042537	Benzene-containing compound metabolic process	23	3	6.4E-4	17.50
GO:0051491	Positive regulation of filopodium assembly	23	3	6.4E-4	17.50

Table S7. Enrichment analysis for putative genes under selection in relation with slope in Moroccan goats.

GO term	Biological process	Number of genes associated	Number of candidate genes associated	P-value	Enrichment
GO:2001171	Positive regulation of ATP biosynthetic process (a)	3	2	3.27E-4	63.41
GO:0060282	Positive regulation of oocyte development	3	2	3.27E-4	63.41
GO:0045979	Positive regulation of nucleoside metabolic process	16	3	5.77E-4	17.83
GO:2001169	Regulation of ATP biosynthetic process (a)	4	2	6.5E-4	47.56
GO:1903580	Positive regulation of ATP metabolic process (a)	4	2	6.5E-4	47.56
GO:0009214	Cyclic nucleotide catabolic process	17	3	6.95E-4	16.78

Biological processes marked by the same letter in parenthesis were clustered together using REVIGO with medium similarity (Supek et al. 2011).

Table S8. Enrichment analysis for putative genes under selection in relation with rainfall in March in Moroccan goats.

GO term	Biological process	Number of genes associated	Number of candidate genes associated	P-value	Enrichment
GO:0030593	Neutrophil chemotaxis (a)	36	6	5.73E-6	13.02
GO:1990266	Neutrophil migration (a)	39	6	9.3E-6	12.02
GO:0071621	Granulocyte chemotaxis (a)	42	6	1.45E-5	11.16
GO:0051240	Positive regulation of multicellular organismal process	1083	31	1.81E-5	2.24
GO:0002433	Immune response-regulating cell surface receptor signaling pathway involved in phagocytosis	65	7	1.87E-5	8.42
GO:0038096	Fc-gamma receptor signaling pathway involved in phagocytosis (b)	65	7	1.87E-5	8.42
GO:0002431	Fc receptor mediated stimulatory signaling pathway (b)	66	7	2.07E-5	8.29
GO:0038094	Fc-gamma receptor signaling pathway (b)	66	7	2.07E-5	8.29
GO:0097530	Granulocyte migration (a)	46	6	2.48E-5	10.19
GO:1901897	Regulation of relaxation of cardiac muscle (c)	6	3	4.01E-5	39.07
GO:0043269	Regulation of ion transport (d)	483	18	4.86E-5	2.91
GO:1901077	Regulation of relaxation of muscle (c)	8	3	1.1E-4	29.30
GO:0010959	Regulation of metal ion transport (d)	260	12	1.32E-4	3.61
GO:0097529	Myeloid leukocyte migration (a)	63	6	1.5E-4	7.44
GO:1901078	Negative regulation of relaxation of muscle (c)	2	2	1.63E-4	78.15
GO:1901898	Negative regulation of relaxation of cardiac muscle (c)	2	2	1.63E-4	78.15
GO:0060326	Cell chemotaxis (a)	123	8	1.82E-4	5.08
GO:0030534	Adult behaviour (e)	125	8	2.03E-4	5.00
GO:0002429	Immune response-activating cell surface receptor signaling pathway	235	11	2.23E-4	3.66
GO:0050851	Antigen receptor-mediated signaling pathway (b)	99	7	2.77E-4	5.53
GO:0051094	Positive regulation of developmental process	879	24	3.58E-4	2.13
GO:0008344	Adult locomotory behaviour (e)	75	6	3.92E-4	6.25
GO:0030595	Leukocyte chemotaxis (a)	80	6	5.55E-4	5.86
GO:0032743	Positive regulation of interleukin-2 production	31	4	6.23E-4	10.08
GO:0007628	Adult walking behaviour (e)	31	4	6.23E-4	10.08
GO:0051239	Regulation of multicellular organismal process	1929	41	6.35E-4	1.66
GO:0034765	Regulation of ion transmembrane transport (d)	309	12	6.36E-4	3.03
GO:1903522	Regulation of blood circulation	186	9	6.56E-4	3.78
GO:0044708	Single-organism behaviour (e)	321	12	8.87E-4	2.92
GO:0097485	Neuron projection guidance (f)	368	13	9.19E-4	2.76
GO:0007411	Axon guidance (f)	368	13	9.19E-4	2.76
GO:0034762	Regulation of transmembrane transport (d)	323	12	9.36E-4	2.90

Biological processes marked by the same letter in parenthesis were clustered together using REVIGO with medium similarity (Supek et al. 2011).

Table S9. Enrichment analysis for putative genes under selection in relation with temperature annual range "bio7" in Moroccan sheep.

GO term	Biological process	Number of genes associated	Number of candidate genes associated	P-value	Enrichment
GO:0097089	Methyl-branched fatty acid metabolic process	2	2	5.92E-5	129.38
GO:0090158	Endoplasmic reticulum membrane organization	2	2	5.92E-5	129.38
GO:0007566	Embryo implantation	35	4	1.47E-4	14.79
GO:0044027	Hypermethylation of CpG island (a)	3	2	1.77E-4	86.25
GO:0044026	DNA hypermethylation (a)	3	2	1.77E-4	86.25
GO:0015908	Fatty acid transport	46	4	4.3E-4	11.25
GO:0001561	Fatty acid alpha-oxidation	6	2	8.7E-4	43.13

Biological processes marked by the same letter in parenthesis were clustered together using REVIGO with medium similarity (Supek et al. 2011).

DISCUSSION GENERALE

Discussion générale

Ce travail de thèse a abordé des aspects de conservation de deux espèces d'élevage qui contribuent amplement à l'alimentation des populations et qui ont un très grand intérêt agronomique et socio-économique à travers le monde. De ces aspects, il a ouvert des questions de recherches fondamentales liées à la compréhension de diversification des génomes chez les mammifères. Il s'est penché sur ces questions sous différents angles. Il y a notamment un aspect méthodologique d'actualité qui est lié aux performances des techniques d'échantillonnage des génomes, incluant l'utilisation de faibles taux de couverture de re-séquençage de génomes complets, pour inférer correctement la variabilité des génomes (diversité neutre et adaptative). Il y a également l'aspect de l'étude de la diversité neutre et des signatures de sélection présentes dans les populations indigènes pour comprendre les bases génétiques de la diversification et de l'adaptation locale chez ces espèces, par l'étude de génomique du paysage au Maroc. Nous avons également estimé via les données de génomes complets la richesse et la diversité génomique globale au sein d'un plus large échantillon de populations/races de chèvres et de moutons à travers le globe, y compris les ancêtres sauvages et les races industrielles. Ceci nous a permis d'émettre une ébauche de conclusion sur la distribution de la biodiversité au sein de ces espèces à plus grande échelle.

1. Vers des solutions alternatives adaptées aux données de génomes complets

1.1. Biais dans les puces commerciales à ADN

De nos jours, le développement de puces à ADN est en plein essor, e.g. (Houston et al. 2015; Malenfant et al. 2015; Palti et al. 2015; Silva-Junior et al. 2015). Il est basé principalement sur l'utilisation de SNPs à fort polymorphisme dans quelques populations de référence (généralement races industrielles) et ne représentant pas de façon similaire toutes les portions du génome e.g. (Tosser-Klopp et al. 2014). Ces puces sont très sollicitées/utilisées et par la communauté scientifique dans des recherches génomiques et par les professionnels de l'élevage. Plusieurs recherches se sont basées sur ces outils de génotypage dans des études de la diversité génétique neutre et adaptative, e.g. (Ramey et al. 2013). Nos résultats dans le Chapitre 1 montrent que les puces classiques comportent un biais de recrutement conséquent. Nous avons bien relevé qu'elles inversent l'ordre des individus dans certains cas et même des

populations lorsqu'on les compare sur la base de leur diversité. La faible densité des SNPs génotypés par les puces à 50K SNPs chez la chèvre et le mouton (1 SNP en moyenne tous les 60kb), et leur distribution le long du génome biaisent également les estimations du déséquilibre de liaison (*LD*). De même, l'absence de variants rares qui représentent en réalité une bonne partie de variation génomique (45 à 58% des variants ayant une fréquence de l'allèle mineure<0,05 ; Chapitres 2 et 3) impactent l'estimation du *LD* même par les puces à hautes densités comme le montrent nos résultats au niveau des Chapitre 2 et 3. De même, nous estimons que la densité insuffisante de marqueurs dans les puces à 50K SNPS empêcherait la détection de certaines signatures de sélection même si la puce ovine a permis de détecter le signal lié au locus *RXFP2* chez les moutons sans cornes, dans notre étude comme (Chapitre 1) dans celle préalable de Kijas et al. (2012). Nous pensons que cette puce aurait de faibles chances à détecter d'autres signatures de sélection présentes dans des populations différentes de celles ayant servi pour sa conception (i.e. absence de polymorphisme à ces locus chez les races ayant servi à sa conception). Ca devrait être probablement le cas de signatures de sélection caractérisées par un faible déséquilibre de liaison. Par exemple le signal lié aux locus *NBEA* et *MAB21L1* (Chapitre 1) qui n'est détecté que par nos jeux de données avec 1M variants aléatoires et plus et qui n'a pas été identifié avec les puces même à haute densité comportant plus de 600K SNPs. Enfin, on peut dire que les utilisateurs de ces techniques d'échantillonnage de génomes pour étudier la variation génomique devraient être conscients de leurs limites et des différents types de biais qu'ils présentent, y compris le biais de recrutement. Ce biais dans les puces à ADN a été soulevé par certaines études, e.g. Albrechtsen et al. (2010) dans une puce humaine à 500K SNPs et tout récemment par McTavish et al. (2015) dans une puce à 50K SNPs bovine qui l'ont inféré via des simulations. Mais, c'est la première fois à notre connaissance qu'on le met en évidence via l'étude de données réelles basées sur la caractérisation de génomes complets.

1.2. Alternatives possibles

Contrairement à la recherche de régions sous sélection, les panels à faible densité de variants aléatoires échantillonnés sur tout le génome (e.g. 10K variants) permettent d'approcher beaucoup mieux la réalité pour décrire la diversité neutre et caractériser la différenciation entre populations. En outre, ces panels de variants aléatoires sont efficaces pour estimer les paramètres génétiques même au sein des populations très différenciées de celles à partir desquelles ils ont été échantillonnés. Nous pensons ainsi que cette approche d'échantillonnage devrait compléter l'ancienne approche utilisée pour la conception de puces à ADN destinées

aux études de génétique des populations. En dehors des études de la diversité génétique « neutre », nous avons défini des nombres minimaux de variants aléatoires permettant d'inférer correctement différents paramètres étudiés. Ces nombres pourraient bien servir à adapter les moyens de génotypages en fonction de l'objectif de l'étude. Différentes techniques de génotypage par séquençage pourraient être choisies selon les densités requises de marqueurs. Ces techniques ne permettent pas d'avoir un échantillon complètement aléatoire mais elles peuvent assurer une couverture plus efficace du génome sans prendre en considération le niveau de polymorphisme d'un SNP pour le considérer. Cependant, nos résultats ont montré que l'intensité des signatures de sélection ressorties par notre comparaison des moutons sans cornes à ceux cornus est beaucoup plus forte en utilisant les données de génomes complets. Nous estimons ainsi que ces données de génomes complets, malgré l'effort nécessaire pour leur mise en œuvre ont toujours une grande valeur ajoutée dans certaines études et permettent d'avoir la meilleure résolution possible et la détermination de la variabilité génomique à différentes échelles, y compris l'échelle des variants. En outre, elles permettent d'accéder à d'autres sources de variation génomiques notamment la variation structurale (inversions, duplications, large insertions/délétions, etc.). Ceci représente certes un avantage inégalé pour l'étude de la variabilité adaptative notamment.

1.3. Des économies sur la couverture de re-séquençage?

Nous avons montré dans le cadre de ce travail qu'un taux de couverture de 5X permet d'avoir les génotypes de plus de 37M de variants sur 43M chez les moutons et 27M sur 32M chez les chèvres comparé à la couverture de 12X. Ce taux de 5X n'a généré que 2% de faux génotypes des variants hétérozygotes et moins de 0.1% de faux génotypes de variants homozygotes comparés toujours aux données de la couverture 12X. Le taux de couverture peut représenter ainsi un paramètre sur lequel on peut agir pour réduire le coût/effort de séquençage à environ la moitié (par rapport à 12X) sans affecter amplement les résultats.

2. Les populations locales et sauvages comme ressources génétiques

En cherchant les alternatives potentielles des données de génomes complets dans la première partie de cette thèse (Chapitre 1), nous avons estimé des paramètres de génétique des populations chez les groupes étudiés via les données de génomes complets, ce qui nous permet de soulever des questions biologiques très intéressantes. Nous ne les avons pas discutées dans l'article qui était centré sur les aspects méthodologiques, mais du point de vue

de la conservation des espèces d'élevage, ces résultats permettent d'émettre des conclusions de grande importance dans le contexte actuel d'érosion massive des ressources génétiques, et nous allons les discuter ici, en lien avec les résultats obtenus dans les autres parties de cette thèse.

2.1. Rappel des principaux résultats de paramètres génétiques

Un rappel des résultats de diversité génétique fournis dans le Chapitre 1 est présenté dans le Tableau 1. Ils concernent les quatre groupes des moutons marocains, chèvres marocaines, mouflons asiatiques et aegagres en partant des variants découverts à l'échelle intra-groupe. Ces résultats montrent que les sauvages sont plus consanguins que les domestiques avec des hétérozygoties comparables à l'échelle de chaque genre. Le déséquilibre de liaison évalué par la distance qui correspond à r^2 =0,15 ($r^2_{0,15}$) est plus long chez les moutons que chez les mouflons alors qu'il est plus court chez les chèvres marocaines que chez les aegagres. En comparant tous les groupes, les mouflons ont le $r^2_{0,15}$ le plus faible et les moutons en ont le plus long. L'indice de différenciation *Fst* (Weir and Cockerham 1984) est plus élevé entre les moutons et les mouflons (0,105) qu'entre les chèvres et les aegagres (et 0,087). La diversité génétique dans les quatre groupes et les domestiques iraniens et les moutons industriels en considérant tous les variants intergroupes montre que les moutons iraniens ont la valeur de π la plus faible et les mouflons en ont la valeur la plus forte chez les *Ovis* (Tableau 2). Chez Capra, les sauvages ont la diversité la plus faible comparés aux domestiques iraniens et marocains (Tableau 2). Le nombre de variants exclusifs aux groupes étudiés est très grand chez les mouflons asiatiques (8,5 millions) comparés aux domestiques. Chez les caprins, les iraniens sauvages et domestiques ont des nombres de variants exclusifs plus grands que les marocains (Figure 1).

2.2. Diversité génétique et différenciation

Le *Fst* de Weir and Cockerham (1984) entre les chèvres marocaines et les aegagres (0,087) est comparable à ceux obtenus chez l'Homme entre les populations ibériques et les populations de l'Asie orientale pour lesquelles ce paramètre varie entre 0,075 to 0,079 en utilisant des données de génomes complets également (Altshuler et al. 2012). Le *Fst* obtenu entre les moutons marocains et les mouflons asiatiques est un peu plus élevé (0,105) mais reste généralement plus faible que ceux rapportés par Ai et al. (2015) entre les sangliers du sud de la Chine et les populations locales porcines dans le même pays et qui varient entre 0,096 et 0,2 en utilisant l'estimateur de Akey et al. (2002) via des données de génomes

complets également. Contrairement aux petits-ruminants, la domestication des cochons aurait été effectuée via des événements parallèles en Asie et en Europe approximativement à la même époque (Rubin et al. 2012) pour un intervalle de génération proche à celui des petits-ruminants. Ceci était inattendu parce que nous nous attendions à ce qu'entre les petits ruminants domestiques au Maroc et les ancêtres sauvages dans les montagnes du Zagros, l'isolement aurait été plus accentué qu'entre les porcins sauvages et domestiques proches géographiquement.

Les taux de consanguinité qui sont plus élevés chez les sauvages comparés à leurs domestiques apparentés seraient le résultat de différents mécanismes. En effet, chez les mouflons, une sous-structure générée par la différenciation d'un groupe de sept individus (Figure S8 du Chapitre 1) issus d'une population récemment introduite par les humains dans une péninsule pour des objectifs de chasse. Ces individus sont marqués par des taux élevés de consanguinité par rapport aux autres mouflons. Chez les aegagres, les individus séquencés venaient de différentes localités et des taux forts de consanguinité sont relevés chez tous les individus. Ceci serait très probablement dû à leur persistance sous forme de petites populations isolées géographiquement dans des habitats constitués principalement de montagnes rudes empêchant ainsi de forts flux de gènes. Par contre, les taux de consanguinité plus élevés chez les domestiques marocains seraient dus aux modes extensifs de conduite basés sur des flux importants de gènes et des pressions modestes de sélection (voir la section « Des scénarios pour expliquer l'état actuel de diversité marocaine »).

Par ailleurs, l'ampleur du déséquilibre de liaison (*LD*) dans le génome représente un élément clé qui conditionne les études d'associations pangénomiques et les recherches des signatures de sélection via des échantillons de marqueurs génomiques (Lewis and Knight 2012) . Il dépend aussi de la taille efficace de la population (*Ne*) (Tenesa et al. 2007). Ainsi, ces populations d'aegagres ayant un *LD* plus élevé que les domestiques marocains et une diversité nucléotidique π plus faible comparée aux domestiques iraniens, auraient probablement subit un fort goulot d'étranglement comme ça a été relevé par Naderi et al. (2008), même s'ils enregistrent un nombre très élevé de variants qui ne sont pas partagés avec les autres échantillons (4,5 millions). Chez *Ovis*, les mouflons sont les plus diversifiés génétiquement, malgré leur plus faible taille d'échantillon comparée aux autres groupes (n=15 versus n=20) ce qui pourrait en principe sous-estimer la diversité. Les résultats du déséquilibre de liaison confirment ce constat qui pourrait être lié à une taille efficace plus élevée que celle des

moutons marocains et des caprins étudiés. Les estimations de tailles efficaces pourraient confirmer cela.

Malgré leur répartition géographique très large (présence sur trois continents) et la diversité génétique qu'ils présentent (π = 0,141), les moutons industriels ont peu de variants spécifiques par rapport aux autres groupes pour lesquels la répartition géographique se limitait à l'échelle d'un pays. Ceci peut être en partie lié au fait que dans notre étude nous n'avons pas fait de découverte de variants dans ce groupe. Cependant le génome de référence utilisé (OAR v3.1) étant assemblé à partir de deux individus Texel (Jiang et al. 2014), les variants exclusifs à ces derniers individus sont présents dans notre jeu de données. Mais ce n'est probablement qu'une faible partie de la variabilité exclusive aux races industrielles. Quoi qu'il en soit, notre échantillon de races « industrielles » qui représente une grande diversité géographique a une diversité génétique plus faible que les populations de moutons du Maroc. Aussi, les variants exclusifs à ces derniers sont plus nombreux que ceux des moutons iraniens. En effet, la forte diversité des populations iraniennes locales n'est pas surprenante sachant qu'une très grande diversité a été capturée durant le processus de domestication (Fernandez et al. 2006; Naderi et al. 2008). Nous attendons aussi une baisse de la diversité depuis le centre de domestication le long des voies de diffusion des animaux domestiques (Bruford et al. 2003). Ainsi, la forte diversité chez les chèvres et moutons marocains en comparaison avec les sauvages et domestiques iraniens était beaucoup moins attendue et corrobore les résultats des chapitres 2 et 3 suggérant une grande diversité dans ces populations.

Ainsi, cette forte diversité génétique devrait bien être prise en compte en terme de ressources génétiques animales. Le niveau élevé de diversité ainsi que le nombre très large de variants exclusifs identifiés dans les différentes populations locales (marocaines et iraniennes) et les animaux sauvages en comparaison avec un échantillon cosmopolite de moutons « industriels » montre leur potentiel de constituer un réservoir précieux d'allèles pouvant être d'une extrême importance adaptative (voir Introduction générale).

2.3. Des scénarios pour expliquer l'état actuel de diversité marocaine

Il y a un manque de littérature relative à l'arrivée des premières populations au Maroc. Les quelques références existantes concernent d'autres localités proches géographiquement, notamment au sud de la Lybie où la première pratique de production laitière mise en évidence remonte au 5ème millénaire avant J.C. (Dunne et al. 2012). De même, les chèvres seraient arrivé en Ibérie il y a environ 7.400 ans (Fernandez et al. 2006). Une présence au

Maroc/Maghreb occidental des animaux d'élevage à cette époque semble ainsi être probable. Par ailleurs, outre les résultats sur les comparaisons entre populations locales et sauvages, les Chapitres 2 et 3 ont montré une très faible structuration actuelle de la diversité génétique chez les petits ruminants du Maroc selon les régions ou les races/populations. Cet ensemble de constats favorise l'hypothèse d'une multiplicité de vagues de colonisation de ce pays via différentes voies potentielles (i.e. Nord-africaine, Ibérienne, Saharienne) comme rapporté pour les chèvres par Pereira et al. (2009) et Benjelloun et al. (2011) (Annexe 1) et avec une forte hétérogénéité de ces populations « colonisatrices ». Ensuite un flux de gènes assez important aurait pu être maintenu avec de faibles pressions de sélection, ne générant pas de forts goulots d'étranglement même lors de la création des races dans ce pays, et ce pour les deux espèces. Des hypothèses ont été émises sur les origines des principales races ovines (Boujenane 1999), mais nous ne maitrisons pas l'histoire exacte ni sur leurs origines ni sur le processus de leur formation. Il semblerait ainsi que celui-ci aurait concerné des troupeaux/populations très larges. Dans ce cas, la standardisation des caractères morphologiques/adaptatifs ne devait pas être le seul objectif imminent de cette formation de races, et celle-ci aurait également été accompagnée du maintient de la diversité dans les populations. Ce scénario semble être aussi soutenu que celui préconisant un flux de gènes récent généralisé et maintenu dans tout le pays après la formation de races. Les résultats du Chapitre 2 sur les signatures de sélection chez les populations de chèvres ont montré que la formation des races distinctes pourrait être liée au développement des mécanismes physiologiques différents pour s'adapter à l'environnement climatique, e.g. halètement/transpiration associés à l'environnement désertique/chaud.

Tableau 1. Rappel des paramètres de génétique des populations intragroupes chez les domestiques marocains et les ancêtres sauvages.

Populations	***Ovis***		***Capra***	
	Domestiques marocains	**Mouflons asiatiques**	**Domestiques marocains**	**Aegagres**
Nombre d'individus (*n*)	30	14	30	18
Heterozygotie (*Ho*)	0,222 ± 0,026	0,223 ± 0,032	0,189 ± 0,018	0,194 ± 0,025
Coefficient de consanguinité (*F*)	0,061 ± 0,108	0,186 ± 0,118	0,056 ± 0,092	0,182 ± 0,106
Déséquilibre de liaison ($r^2_{0.15}$) en Kb	5	2,65	3,53	3,94
Coefficient de différenciation (*Fst*)	0,105		0,087	

Le coefficient de différenciation *Fst* (Weir and Cockerham 1984) est estimé entre les sauvages et domestiques.

Tableau 2. Distribution de la variation génomique dans différents groupes d'*Ovis* et de *Capra* (variants intergroupes).

Populations	*Ovis*				*Capra*		
	Domestiques marocains	Domestiques iraniens	Races Industrielles	Mouflons asiatiques	Domestiques marocains	Domestiques Iraniens	Aegagres
Nombre d'individus (*n*)	20	20	20	15	20	20	20
Nombre de variants polymorphiques	28.426.390	27.244.995	27.432.087	31.193.888	19.807.929	22.874.401	19.120.668
Diversité nucléotidique (π)	0,145	0,139	0,141	0,161	0,118	0,125	0,113

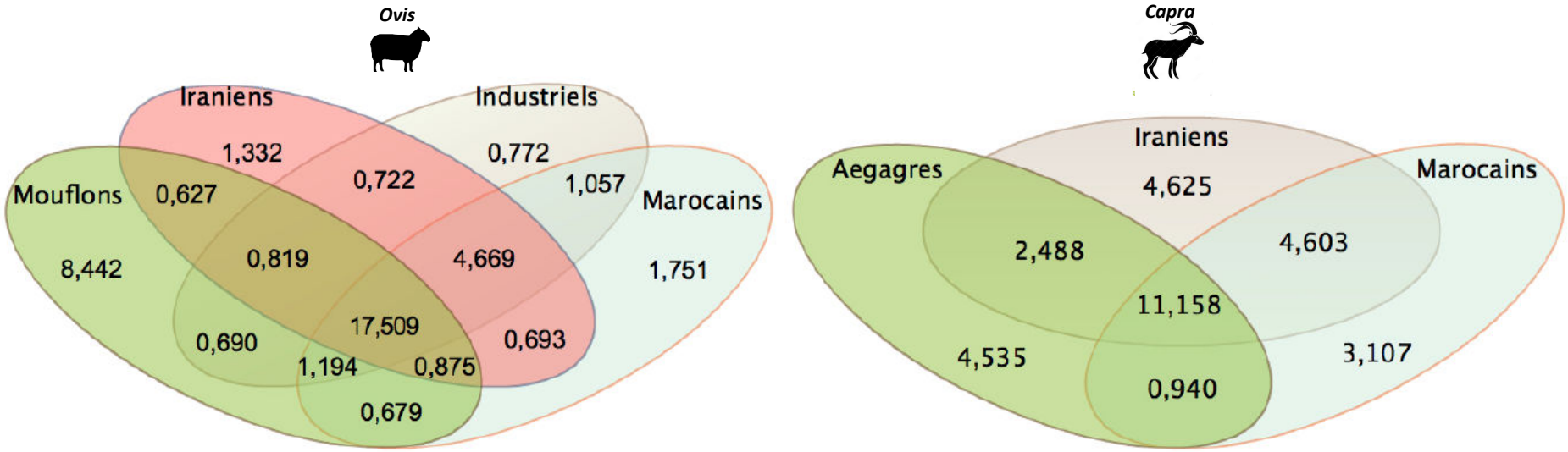

Figure 1. Diagrammes de Venn illustrant les nombres de variants exclusifs et partagés (en millions) entre différentes combinaisons de populations chez les genres *Ovis* et *Capra*.

Les populations domestiques locales sont désignées par leurs origines et les races industrielles sont désignées par « Industriels ».

3. Les bases génétiques de l'adaptation locale chez les chèvres et moutons

3.1. Aspects méthodologiques

Au delà de la diversité génétique globale, nous avons essayé d'identifier les bases génétiques de la différenciation phénotypique, y compris celles de l'adaptation à l'environnement dans les Chapitres 2 et 3 de cette thèse. D'abord, les approches de détection de signatures de sélection n'ont pas été les mêmes dans les deux Chapitres. Le premier a été basé complètement sur une approche populationnelle parce que nous cherchions à identifier les balayages sélectifs au niveau des principales populations caprines individualisées. Le second s'est basé sur une approche de « génomique du paysage » où l'échantillonnage individu-centré a permis d'avoir des échantillons le long de gradients environnementaux. Une approche corrélative spécifique à la génomique du paysage a été appliquée (i.e. Samßada ; Stucki et al. 2014) pour détecter les allèles corrélés aux variations de variables environnementales. Une approche populationnelle a également été mise en œuvre dans ce contexte. Nous avons ainsi constitué des groupes d'individus selon leur positionnement sur chaque gradient environnemental, abstraction faite de leur position géographique. En effet, la quasi-absence de la structure de population a été favorable aux deux approches en minimisant les effets confondants. A la différence du Chapitre 2, l'approche populationnelle appliquée était basée sur une combinaison de XP-CLR (Chen et al. 2010) et du *Fst* par variant permettant ainsi d'identifier les variants candidats faisant partie des 0,1% scores XP-CLR et des 0,1% *Fst* les plus élevés, puis d'identifier les gènes qui en sont associés. Cette approche est conservative et avait pour objectif de réduire la part des faux-positifs, en prenant le risque d'éliminer des variants/gènes faiblement différenciés et qui pourraient être impliqués dans des mécanismes adaptatifs via des faibles effets épistatiques, tel a été le cas de l'adaptation possible des populations humaines aux pathogènes (Daub et al. 2013). Un autre problème lié à l'utilisation de l'approche populationelle est relatif à l'absence de carte génétique chez les espèces étudiées. XP-CLR n'étant pas sensible au taux de recombinaison tel que ça a été rapporté par Chen et al. (2010), mais nous estimons que la définition de ces cartes génétiques à l'échelle des génomes devraient permettre d'appliquer en outre des méthodes basées sur le déséquilibre de liaison pour détecter les balayages sélectifs.

3.2. Non concordance des signatures de sélection entre les méthodes corrélatives et populationnelles

Un des points importants révélé par notre étude est la non concordance entre les signatures de sélection identifiées par la méthode corrélative (Samßada) et l'approche populationnelle (XP-CLR/*Fst*). Un point essentiel semble être la détection de peu de signatures de sélection par la première approche. L'une des raisons possibles à cela serait la prise en compte d'une partie des individus pour détecter les signatures de sélection par l'approche populationnelle (i.e., groupes extrêmes) ce qui n'est pas le cas dans l'approche corrélative. L'autre explication serait liée à la présence d'une certaine structure de population (même si elle est très faible) ce qui peut être une source potentielle de baisse de performance des méthodes corrélatives en général (Frichot et al. 2013; Stucki et al. 2014). Cependant, la différenciation observée des variants candidats le long de gradients altitudinaux montre bien une certaine corrélation entre les variations alléliques (illustrées par la différenciation) au niveau de ces groupes de variants et les variables environnementales. Nous sommes toujours en train d'examiner les raisons de cette faible détection des signatures de sélection par l'approche corrélative.

3.3. Adaptations parallèles dans différentes populations/espèces

L'un des résultats les plus probants de l'étude des signatures de sélection au sein des 3 principales populations caprines marocaines est lié aux indications sur différents mécanismes physiologiques possibles aidant à la thermorégulation chez deux populations différentes. En effet, nos résultats aussi bien que les caractéristiques morphologiques des animaux (e.g. taille de la tête) permettent de suggérer que la chèvre Draa favoriserait une dissipation de la chaleur via l'halètement. La population Noire devrait favoriser la transpiration pour s'adapter au même milieu d'élevage. L'existence de ces deux mécanismes aidant à la dissipation de la chaleur chez les chèvres a été rapportée (Dmiel and Robertshaw 1983; Baker 1989), mais le fait de relever que deux populations faiblement différenciées sur le plan génétique et présentes dans la même zone géographique aient des mécanismes adaptatifs complètement différents était surprenant. Cette hypothèse suggère que la même adaptation (au climat chaud) peut se produire via 2 mécanismes physiologiques différents (l'halètement et la transpiration) au sein de la même espèce et dans le même environnement (Figure 2A). Quelques facteurs peuvent expliquer cela, notamment, le coût adaptatif qui amènerait probablement à sélectionner un trait plutôt que l'autre dans une population. Dans ce cas spécifique, le pouvoir de la ventilation assurée par l'halètement à refroidir le sang passant par la région nasale et à garder le cerveau de l'animal à une température convenable pourrait être déterminant (Baker 1989).

En effet, comme stipulé dans le Chapitre 2, l'halètement par rapport à la transpiration présente l'avantage de refroidir le cerveau en préservant le volume plasmatique sanguin (pas de perte de sels). Mais son pouvoir à assurer ce refroidissement pourrait être tributaire de la taille de la tête qui est plus petite chez la Draa. Cette observation au sein de la même espèce nous amène à soulever la non implication des mêmes gènes ni des mêmes fonctions dans l'adaptation des chèvres et moutons aux mêmes environnements. En effet, dans le Chapitre 3, nous avons rapporté que moins de 1% des gènes candidats sont communs aux deux espèces pour la même variable environnementale considérée. De même, les termes de *Gene Ontology* (GO) qui ont été enrichis sont différents. Ceci suggère a priori des mécanismes adaptatifs différents. Cependant, un cas semble être très important : le gène *MCM3* est identifié sous sélection pour l'adaptation à l'altitude chez les moutons. Ce gène appartient à une famille qui inhibe l'activité transcriptionnelle du gène *HIF-1* en réponse aux conditions d'hypoxie (Hubbi et al. 2011). Ce dernier gène avec *HIF-2* (*EPAS1*) représentent les principaux gènes qui seraient impliqués chez l'Homme dans l'adaptation des Tibétains à l'altitude (Beall et al. 2010; Simonson et al. 2010; Yi et al. 2010). Il s'agit là d'un cas où la même adaptation chez deux espèces différentes (humains et moutons) serait contrôlée par le même mécanisme physiologique via un gène dans la première espèce et à travers un autre gène qui régule le premier dans la seconde espèce (Figure 2B). Par ailleurs, les effets régulateurs jouent probablement un rôle primordial dans les processus adaptatifs étudiés parce que 60% des variants identifiés dans nos analyses au niveau du Chapitre 3 sont inter-génique et 30% sont dans des introns. Nous ignorons la part des faux-positifs dans nos détections, mais ces pourcentages corroborent les résultats sur l'adaptation des cochons à l'altitude (Ai et al. 2015) et sur d'autres adaptations. Les mécanismes par lesquels ces variants agissent ne sont pas entièrement élucidés (Ward and Kellis 2012). Différentes types d'action ont été déjà décrits, e.g. régulation de la transcription, promoteurs, amplificateurs. Chez les humains, le projet ENCODE a permis de générer une large carte de ces éléments dans plusieurs cellules souches à l'échelle du génome (Dunham et al. 2012). De telles cartes chez d'autres espèces ne sont pas établies jusqu'à aujourd'hui.

Un autre type d'adaptation révélée via nos analyses et qui est aussi important est celui du gène *NFIB* qui est identifié pour l'adaptation à l'altitude chez les chèvres et les moutons à la fois. C'est un gène qui est impliqué dans la maturation des poumons chez la souris (Steele-Perkins et al. 2005) et dans la différenciation des cellules de Clara qui sont des cellules progénitrices dans les petites voies respiratoires (Giangreco et al. 2002). Il est très probable qu'il s'agit là

d'un même gène impliqué dans le même mécanisme physiologique (protection contre les dommages lié à l'hypoxie au niveau des bronchioles ; Chapitre 3) et impliqué dans la même adaptation (Figure 2C). Plusieurs cas de ce types d'adaptation parallèle ont été rapportés, e.g. l'implication des gènes *MC1R* et son antagoniste *ASIP* dans la couleur de la peau de plusieurs mammifères les aidant ainsi à se dissimuler dans leurs environnements respectifs (Manceau et al. 2010 ; Chapitre 2).

De toute façon, les résultats relatifs aux différences de fonctions et de gènes identifiés pour l'adaptation au même environnement dans deux espèces relativement proches (*O. aries* et *C. hircus*), voire même dans deux populations de la même espèce, nous permettent de suggérer que l'adaptation aux mêmes conditions environnementales peut se mettre en place via une multitude de traits adaptatifs qui sont régis par des fonctions et des gènes différents. Ceci peut bien expliquer la multiplicité des mécanismes génétiques régissant l'adaptation locale comme ça a été décrit dans l'Introduction générale de ce document.

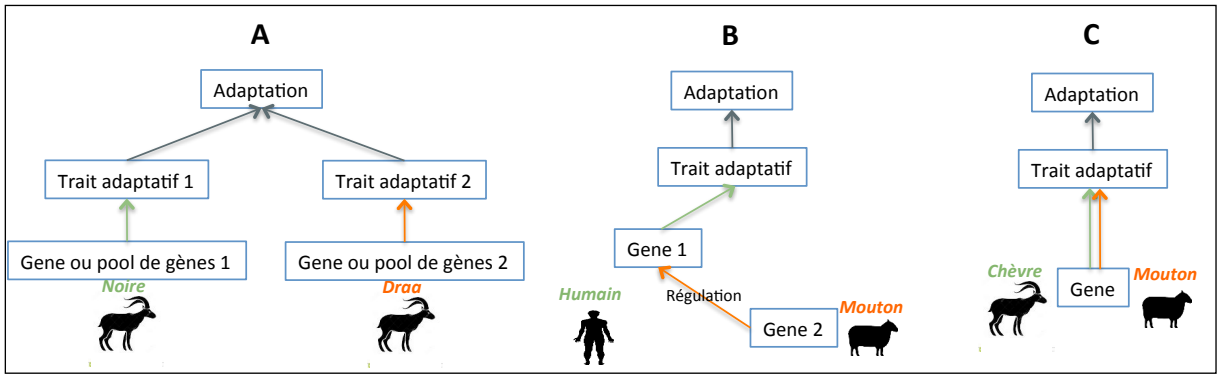

Figure 2. Cas de convergences génétiques probables régissant les adaptations de plusieurs espèces/populations au même environnement relevées dans notre étude.

L'adaptation au même environnement peut être induite par : (A) différents gènes probablement impliqués dans différentes fonctions et différents traits dans deux populations au sein de la même espèce, mais ce cas a aussi été noté dans différentes espèces (chèvres et moutons) ; (B) un gène sélectionné dans une espèce et dans l'autre le gène impliqué est probablement celui qui régule le premier ; (c) dans deux espèces, le même gène est probablement impliqué dans le même trait adaptatif. Les couleurs distinguent les actions probables des gènes candidats au niveau de chaque espèce/population.

3.4. Implication de fonctions respiratoire et circulatoire dans l'adaptation à l'altitude

Le Chapitre 3 a permis d'associer plusieurs voies métaboliques à l'adaptation à plusieurs variables environnementales. L'une des principales voies métaboliques identifiées pour l'adaptation des chèvres à l'altitude est la « Différenciation des cellules de Clara ». Ces

cellules qui sont des cellules mères des cellules épithéliales des petites voies pulmonaires sont nombreuses chez les lamas de haute altitude (Heath et al. 1976) et présentent des signes de prolifération compensatrices dans des conditions d'hypoxie chez les lapins en conditions contrôlées (Uhlik et al. 2005). Les changements qu'elles subissent chez les caprins de haute altitude devront être déterminés mais il semblerait qu'elles ont un rôle au niveau des poumons. D'ailleurs, le gène *NFIB* impliqué dans l'enrichissement du terme GO associé à cette fonction est lié à la signature de sélection la plus élevée lors d'une comparaison entre les moutons de basse altitude à ceux de haute altitude. C'est un gène qui joue également un rôle important dans la maturation des poumons (Steele-Perkins et al. 2005). Des études physiologiques des poumons des moutons et chèvres de haute altitude devraient permettre d'élucider les rôles de la fonction respiratoire dans cette adaptation. Les autres principales fonctions métaboliques qui sont associées à l'altitude chez les chèvres étaient liées aux fonctions cardiaques, sa contraction et son système de conduction électrique. Comme stipulé dans le Chapitre 3, des études distinctes ont mis en évidence des implications directes et indirectes de ces fonctions pour compenser le manque d'oxygène chez des animaux de laboratoire principalement (Calmettes et al. 2010; Zhou et al. 2015), e.g. la prolongation du potentiel d'action ventriculaire chez les rats durant l'ischémie (Zhou et al. 2015). L'implication de ces fonctions respiratoire et circulatoire dans l'adaptation à l'altitude chez les chèvres suggère que cette adaptation se produit via des mécanismes compensant le manque d'oxygène ou limitant ses dommages. Des mécanismes ayant des effets identiques représentent les principales cibles des gènes sous sélection détectées jusqu'à aujourd'hui chez les populations humaines du Tibet avec l'implication des gènes *HIF-1α* et *HIF-2α/EPAS1* (Simonson et al. 2010; Yi et al. 2010). Ces gènes sont notamment exprimés dans les poumons et ils sont associés avec le nombre d'érythrocytes et la concentration de l'hémoglobine (Yi et al. 2010). Leur implication dans le maintien de suffisamment d'oxygène tissulaire a été suggérée par ces derniers auteurs. Chez le mouton, le gène *MCM3* qui régule le *HIF-1* a été fortement associé à l'altitude (voir section « Adaptations parallèles dans différentes populations/espèces »). En outre, le gène *NFIB* est impliqué dans la maturation des poumons comme cité dans la même section. La contraction musculaire est également associée à la haute altitude chez le mouton. Si on suppose qu'il s'agit des muscles squelettiques, nous pouvons suggérer que cette adaptation consiste en la réduction des effets du manque d'oxygène sur les propriétés contractiles des muscles comme ça a été mis en évidence par Perrey and Rupp (2009). Ceci supporte l'hypothèse que ces mécanismes compensatoires du

manque d'oxygène représentent l'un des principaux traits sélectionnés dans les conditions de haute altitude chez plusieurs espèces de mammifères.

Notre approche basée sur les populations met en évidence plusieurs autres fonctions métaboliques qui sont associées à d'autres variables environnementales. Par exemple, les processus de biosynthèse des ATP qui sont associés à la pente chez les chèvres qui peuvent être expliqués par le mode de conduite souvent extensif chez les chèvres locales marocaines, principalement celles des montagnes. Ainsi, les chèvres du relief accidenté auraient besoin de plus d'énergie pour chercher du pâturage selon sa disponibilité. Dans d'autres cas, les gènes identifiés ne permettent pas d'inférer clairement les fonctions correspondantes et leur implication dans l'adaptation. Ceci n'enlève rien de leur importance et des investigations sur ces fonctions devront être entreprises par la suite. D'une façon générale, des études moléculaires et fonctionnelles détaillées seraient requises pour examiner l'ampleur des changements d'expression de gènes identifiés dans cette étude.

3.5. Différenciation « adaptative » le long des gradients d'altitude

Dans le Chapitre 3, nous avons caractérisé l'évolution de la différenciation (*Fst*) entre les chèvres (respectivement moutons) de basse et haute altitude en utilisant une limite glissante le long du gradient altitudinal. Nous avons mené cette approche sur les variants candidats inclus dans les huit gènes associés aux scores XP-CLR les plus élevés. Plusieurs types de variation clinale sont relevés. Des courbes en « U », indiquant un maximum de différenciation en bas et en haut du gradient, sont observées pour les gènes *MCM3* et *GMDS* chez les ovins et *KHDRBS2* chez les caprins. Ce type de forme indiquerait la présence d'allèles différents à fortes fréquences (voire fixés) aux deux extrémités du gradient et présentant des fréquences équivalentes en milieu du gradient. De même nous avons observé des courbes en « S » comme pour les gènes *OXR1* et *PUS7L* chez les moutons et *GATAD2A* et *SNAPC4* chez les chèvres. Ces formes indiqueraient la prédominance croissante d'un allèle lorsqu'on se rapproche de l'extrémité d'un gradient. L'origine géographique très diversifiée des individus (i.e. une même altitude est représentée par des individus de populations et localités différentes) confirme que ces variations de différenciations liées aux fréquences alléliques, qui sont de plus locus spécifiques, ne sont pas liées à des flux géniques en lien avec des phénomènes d'isolement par la distance. De nombreuses études ont mis en évidence des variations clinales liées à la sélection, e.g. les clines latitudinaux de la taille de *Drosophila subobscura* (Huey et al. 2000) ou de temps de floraison chez *Arabidopsis thaliana* qui a été attribué à la sélection sur un gène « *FRIGIDA* » (Stinchcombe et al. 2004). Cependant, il est

toujours difficile de distinguer l'effet de l'histoire démographique de celui de la sélection (Savolainen et al. 2013). Notre approche, par la nature de l'échantillonnage permet de s'affranchir de cette confusion.

3.6. Limites méthodologiques

Dans la caractérisation des mécanismes adaptatifs, notre approche d'identification des fonctions biologiques comporte des sources de biais : (i) L'absence de modèles biologiques proches des moutons et des chèvres dans les outils d'enrichissement des termes GO limite le choix du modèle à considérer. Nous avons opté pour *Homo sapiens* pour lequel l'annotation fonctionnelle est la plus complète parmi toutes les espèces. Cette stratégie pourrait être une source de sous ou sur-estimation de la représentativité de certains catégories GO. Ceci serait lié au fait que les fonctions des gènes peuvent varier d'une espèce à l'autre comme a été stipulé dans l'Introduction générale et rapporté par Manceau et al. (2010) et Kamberov et al. (2013). (ii) L'outil d'enrichissement des termes GO que nous avons utilisé dans les Chapitres 2 et 3 ne prend pas en compte la longueur des gènes sur le génome pour déterminer la probabilité de leur identification par erreur. Il y a d'autres méthodes qui prennent en considération cette source potentielle d'erreur et nous prévoyons les appliquer prochainement.

L'une des questions sur laquelle nous avons passé beaucoup de temps était associée au seuil du score XP-CLR et *Fst* à appliquer pour identifier les variants/gènes candidats. Nous avons appliqué des seuils-pourcentages (0,1% des scores XP-CLR et des *Fst*), mais nous sommes convaincus qu'ils peuvent bien ne pas représenter la réalité. Nous pensons que le développement d'approches permettant de fixer des seuils variables selon l'étude peut aider à surmonter ceci. Cependant, nos approches restent conservatives. C'est bien le cas du Chapitre 3 où l'approche basée sur les populations a utilisé une combinaison de deux méthodes de détection de signatures de sélection en conservant les variants identifiés par les deux méthodes (i.e. XP-CLR et *Fst*).

4. Perspectives

Comme stipulé au début de cette discussion, ce travail de thèse a permis d'appréhender plusieurs aspects de conservation génétique et d'autres aspects généraux liés à la compréhension de bases génétiques possibles de l'adaptation locale et de la distribution de la biodiversité. Il a permis naturellement d'ouvrir plusieurs horizons, d'abord pour finaliser le

travail qui est en cours et qui est inscrit dans un cadre plus global du projet NextGen, et pour orienter des recherches futures.

4.1. Finalisation des études en cours

L'étude des mécanismes adaptatifs représente un travail qui est toujours en cours de finalisation. Nous prévoyons de compléter les analyses en appliquant d'autres méthodes corrélatives de détection de sélection (i.e. LFMM; Frichot et al. 2013) et d'identifier l'effet des variants non-synonymes détectés sur les protéines et identifier les effets possibles sur les fonctions. Nous pensons également à étudier cet effet en regardant les haplotypes sélectionnés pour certains gènes différenciés au niveau des pools extrêmes d'individus (e.g. faible et haute altitude).

Une autre étape consistera à examiner si dans les mêmes gènes identifiés dans les deux espèces pour la même adaptation, ce sont les mêmes variants qui sont impliqués ou pas, tel que cela a été déjà rapporté dans la littérature dans la coloration pâle chez deux espèces de lézards (Manceau et al. 2010). Cela nous permettra d'étudier les différentes formes possibles de convergence génétique liée à l'adaptation locale à l'échelle de plus d'une espèce en partant des similitudes entre variants.

Enfin, un travail qui est en cours de réalisation par les membres du consortium NextGen et auquel je contribue en partie est relatif à l'étude des bases génétiques des mécanismes qui auraient été sélectionnés lors des processus de domestication des chèvres et moutons. La compréhension de ces bases génétiques et des fonctions des traits ayant été sélectionnés lors de ces processus pourrait aider à comprendre les bases génétiques d'une transition clé dans l'histoire humaine. Au delà de leur valeur scientifique, ces bases pourraient bien être utiles pour l'amélioration et la conservation des populations domestiques.

4.2. Recherches futures

(i) Echantillonnage des génomes

L'étude des stratégies d'échantillonnage des variants pour étudier la variation génomique a permis d'identifier le biais dans les puces commerciales à ADN et de proposer de nouvelles approches basées sur l'échantillonnage de nombres variables de variants aléatoires selon l'objectif de l'étude. Cependant, les aspects techniques sur la faisabilité de cet échantillonnage en absence des données de re-séquençage n'ont pas été abordés. Des techniques de génotypage par séquençage (e.g. RAD-seq) ont été proposées pour approximer cet

échantillonnage mais, il serait profitable d'évaluer la capacité de ces techniques à représenter la variation génomique. De même d'autres méthodes pourraient être testées, en les simulant par extraction à partir des données de génomes complets, pour déterminer leurs limites et leurs avantages (e.g. RNA-seq ; Pool-seq).

(ii) Conservation des animaux domestiques

L'état actuel de la diversité génétique au sein des populations locales et sauvages évalué dans le cadre de cette thèse devrait déboucher sur des recherches visant à désigner des schémas efficients de suivi et de gestion de cette biodiversité en considérant les variables prédictives et les modèles de changements environnementaux. En outre, au vu de la forte érosion génétique que subissent les espèces domestiques (FAO 2007 ; Taberlet et al. 2008), une tâche qui semble primordiale à l'heure actuelle est la sensibilisation des décideurs à la richesse que présentent les populations indigènes d'animaux d'élevage. Ensuite des activités de recherche ayant pour objectif de concevoir des méthodes efficaces d'amélioration durable de la productivité des élevages en assurant la préservation des ressources génétiques locales devraient être mises en oeuvre.

(iii) Etudes fonctionnelles

Les résultats obtenus sur les bases génétiques de l'adaptation locale ont ressorti un nombre très élevé de gènes candidats et de fonctions métaboliques. L'élucidation complète de ces fonctions nécessite des recherches moléculaires et fonctionnelles détaillés spécifiques à chaque mécanisme identifié tel que nous l'avons suggéré d'une façon ponctuelle dans les sections précédentes.

(iv) Cartes génétiques

La réalisation d'une carte génétique des chèvres (et respectivement des moutons). Cette carte permettrait d'illustrer la variation de la recombinaison et du déséquilibre de liaison à l'échelle des génomes. Avec les 400 génomes de petits ruminants qui ont été produits dans le cadre du projet NextGen, l'établissement d'une telle carte génétique serait réalisable. Nous l'avons abordé pour la recherche des signatures de sélection liées au locus *RXFP2* dans le Chapitre 1 mais ça n'a concerné que 20Mb sur le chromosome 10 du génome. Une telle carte permettrait d'inclure les variations du déséquilibre de liaison le long du génome pour chercher les signatures de sélection (Sabeti et al. 2007).

(v) Eléments régulateurs du génome

L'implication de la variation non-codante dans la variation phénotypique est connue et elle se produit via plusieurs mécanismes connus en partie (Ward and Kellis 2012). Chez les humains une carte de variation régulatrice a été établie dans le cadre du projet ENCODE (Dunham et al. 2012). Ainsi, la compréhension de l'adaptation locale chez les petits ruminants ne peut être complète sans l'annotation des éléments régulateurs sur leurs génomes. Nous sommes conscient que l'établissement d'une telle carte dans le court terme est difficilement envisageable, mais elle représenterait certes une grande valeur ajoutée dans les études de la variation adaptative.

(vi) Adaptation aux pathogènes

La résistance aux pathogènes constitue un aspect adaptatif très important dans le contexte actuel de changements climatiques. L'étude de ses bases génétique représente plusieurs intérêts et applications. Outre la compréhension de ces résistances, il y a notamment l'intégration de ces aspects dans les schémas de gestion de la biodiversité ainsi que de l'amélioration de la productivité des animaux d'élevage. Les données de génomes complets produits dans le cadre de NextGen concernent une large zone géographique au Maroc qui est caractérisée par une prévalence plus ou moins variée de plusieurs pathogènes spécifiques aux petits-ruminants. Il serait très important de compléter l'échantillonnage NextGen en fonction de cette prévalence pour étudier les bases génétiques de la résistance aux pathogènes (chez les humains, voir (Daub et al. 2013)). Dans ce cadre, un projet a été soumis en réponse à l'appel d'offres ARIMNet2 en intégrant 9 partenaires dont l'INRA-Maroc. L'objectif est d'étudier ces aspects à une échelle méditerranéenne via une approche intégrée.

(vii) Plasticité et épi-génétique

Il s'agit d'un des mécanismes aidant à la survie des populations dans le contexte de changements environnementaux (Chevin et al. 2010). La plasticité peut affecter la dynamique évolutive en interagissant notamment avec l'adaptation locale de différentes manières (Jablonka 2013). L'étude des aspects épigénétiques et notamment des patrons de méthylation dans des conditions environnementales contrastées chez les chèvres et moutons du Maroc permettrait de comprendre ces interactions. Dans ce cadre, un projet est en cours de réalisation traitant cet aspect à une échelle plus grande en impliquant les petits ruminants du Maroc « ClimGen ».

Conclusion

Les travaux réalisés dans le cadre de cette thèse ont permis de clarifier plusieurs aspects appliqués et d'autres fondamentaux. Le premier aspect traité est méthodologique et a permis de déterminer les biais présents dans les approches classiques d'échantillonnage des génomes dans les études de génomique des populations. Il a permis également de proposer des stratégies adaptées pour obtenir une image représentative de la variation génomique en fonction des objectifs assignés à l'étude. Ces propositions pourraient aider à concevoir des méthodes de génotypage de nouvelle génération, plus robustes dans certains cas d'études en suivant d'autres approches pour choisir les variants/régions génomiques à génotyper. En outre, ce travail de thèse a utilisé les données de génomes complets pour mettre l'accent sur la richesse précieuse que représentent les populations indigènes et les populations sauvages issues directement des ancêtres des animaux d'élevage, et ce dans un contexte combiné d'érosion massive de la biodiversité et de changements environnementaux. Les résultats devraient aider à mieux concevoir les plans de monitoring de la biodiversité au sein des animaux de ferme et à élaborer des programmes d'amélioration durable de l'élevage à l'échelle planétaire en prenant en compte la valeur de la biodiversité encore présente en dehors des élevages « industriels ». Les résultats obtenus dans le cadre de ce volet devraient supporter les efforts de préservation entrepris à l'échelle mondiale pour freiner l'extinction des populations indigènes en fournissant des éléments quantifiés sur la valeur qu'elles représentent. Cette thèse a également permis de développer un volet adaptatif via une approche de génomique du paysage sans précédent. Outre les bases génétiques rapportées et qui ont permis de poser les bases de certains mécanismes liés à l'adaptation à leur environnement contraignant, ce travail a ouvert la voie à l'étude de convergences adaptatives entre espèces proches partageant un même environnement, voire même entre populations d'une même espèce. Ce travail a également permis de caractériser les variations clinales de différenciation de fragments chromosomiques liés à des gènes candidats le long de gradient environnementaux. De telles variations se limitaient jusque là à des aspects théoriques ou alors à des études ne pouvant pas avoir une telle résolution. Ce travail a permis ainsi d'approfondir nos connaissances sur les mécanismes d'adaptation locale en utilisant des modèles biologiques peu utilisés dans ce genre d'études (i.e. animaux domestiques), mais qui présentent un grand nombre d'avantages qui incluent : (i) des génomes de références de mieux en mieux caractérisés, (ii) la possibilité d'élevages en conditions contrôlées et d'expérimentations pour une éventuelle validation fonctionnelle, (iii) la présence d'espèces

apparentées représentées par les ancêtres sauvages avec une divergence récente d'environ 10.000 ans qui permettront de prospecter les origines des traits adaptatifs, (iv) la possibilité d'intégration des traits adaptatifs validés dans les schémas de sélection pour un développement durable de l'élevage dans des environnements contraignants.

Références

Ai H, Fang X, Yang B, Huang Z, Chen H, Mao L, Zhang F, Zhang L, Cui L, He W et al. 2015. Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nature Genetics* **47**(3): 217-+.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Research* **12**(12): 1805-1814.

Albrechtsen A, Nielsen FC, Nielsen R. 2010. Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. *Molecular Biology and Evolution* **27**(11): 2534-2547.

Altshuler DM Durbin RM Abecasis GR Bentley DR Chakravarti A Clark AG Donnelly P Eichler EE Flicek P Gabriel SB et al. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422): 56-65.

Baker MA. 1989. EFFECTS OF DEHYDRATION AND REHYDRATION ON THERMOREGULATORY SWEATING IN GOATS. *Journal of Physiology-London* **417**: 421-435.

Beall CM, Cavalleri GL, Deng L, Elston RC, Gao Y, Knight J, Li C, Li JC, Liang Y, McCormack M et al. 2010. Natural selection on EPAS1 (HIF2 alpha) associated with low hemoglobin concentration in Tibetan highlanders. *Proceedings of the National Academy of Sciences of the United States of America* **107**(25): 11459-11464.

Benjelloun B, Pompanon F, Ben Bati M, Chentouf M, Ibnelbachyr M, El Amiri B, Rioux D, Boulanouar B, Taberlet P. 2011. Mitochondrial DNA polymorphism in Moroccan goats. *Small Ruminant Research* **98**(1-3): 201-205.

Boujenane I. 1999. Les ressources génétiques ovines au Maroc. Actes Editions, pp 136, Rabat, Maroc.

Bruford MW, Bradley DG, Luikart G. 2003. DNA markers reveal the complexity of livestock domestication. *Nature Reviews Genetics* **4**(11): 900-910.

Calmettes G, Deschodt-Arsac V, Gouspillou G, Miraux S, Muller B, Franconi J-M, Thiaudiere E, Diolez P. 2010. Improved Energy Supply Regulation in Chronic Hypoxic Mouse Counteracts Hypoxia-Induced Altered Cardiac Energetics. *Plos One* **5**(2).

Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. *Genome Research* **20**(3): 393-402.

Chevin L-M, Lande R, Mace GM. 2010. Adaptation, Plasticity, and Extinction in a Changing Environment: Towards a Predictive Theory. *Plos Biology* **8**(4).

Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-Murci L, Robinson-Rechavi M, Excoffier L. 2013. Evidence for Polygenic Adaptation to Pathogens in the Human Genome. *Molecular Biology and Evolution* **30**(7): 1544-1558.

Dmiel R, Robertshaw D. 1983. THE CONTROL OF PANTING AND SWEATING IN THE BLACK BEDOUIN GOAT - A COMPARISON OF 2 MODES OF IMPOSING A HEAT LOAD. *Physiological Zoology* **56**(3): 404-411.

Dunham I Kundaje A Aldred SF Collins PJ Davis C Doyle F Epstein CB Frietze S Harrow J Kaul R et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414): 57-74.

Dunne J, Evershed RP, Salque M, Cramp L, Bruni S, Ryan K, Biagetti S, di Lernia S. 2012. First dairying in green Saharan Africa in the fifth millennium BC. *Nature* **486**(7403): 390-394.

FAO. 2007. The State of the World's Animal Genetic Resources for Food and Agriculture in brief, (eds by Dafydd Pilling & Barbara Rischkowsky), Rome, Italy.

Fernandez H, Hughes S, Vigne J-D, Helmer D, Hodgins G, Miquel C, Hanni C, Luikart G, Taberlet P. 2006. Divergent mtDNA lineages of goats in an Early Neolithic site, far from the initial domestication areas. *Proceedings of the National Academy of Sciences of the United States of America* **103**(42): 15375-15379.

Frichot E, Schoville SD, Bouchard G, Francois O. 2013. Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution* **30**(7): 1687-1699.

Giangreco A, Reynolds SD, Stripp BR. 2002. Terminal bronchioles harbor a unique airway stem cell population that localizes to the bronchoalveolar duct junction. *American Journal of Pathology* **161**(1): 173-182.

Heath D, Smith P, Harris P. 1976. CLARA CELLS IN LLAMA. *Experimental Cell Biology* **44**(2): 73-82.

Houston DD, Mitchell KS, Clouse JW, Maughan PJ, Creighton JC, Smith AN, Bybee SM, Belk MC. 2015. SNP development in North American burying beetles (Coleoptera: Silphidae): a tool to inform conservation decisions. *Conservation Genetics Resources* **7**(2): 349-352.

Hubbi ME, Luo W, Baek JH, Semenza GL. 2011. MCM Proteins Are Negative Regulators of Hypoxia-Inducible Factor 1. *Molecular Cell* **42**(5): 700-712.

Huey RB, Gilchrist GW, Carlson ML, Berrigan D, Serra L. 2000. Rapid evolution of a geographic cline in size in an introduced fly. *Science* **287**(5451): 308-309.

Jablonka E. 2013. Epigenetic inheritance and plasticity: The responsive germline. *Progress in Biophysics & Molecular Biology* **111**(2-3): 99-107.

Jiang Y, Xie M, Chen W, Talbot R, Maddox JF, Faraut T, Wu C, Muzny DM, Li Y, Zhang W et al. 2014. The sheep genome illuminates biology of the rumen and lipid metabolism. *Science* **344**(6188): 1168-1173.

Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, Yang Y, Li S, Tang K, Chen H et al. 2013. Modeling Recent Human Evolution in Mice by Expression of a Selected EDAR Variant. *Cell* **152**(4): 691-702.

Kijas JW, Lenstra JA, Hayes B, Boitard S, Neto LRP, San Cristobal M, Servin B, McCulloch R, Whan V, Gietzen K et al. 2012. Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. *Plos Biology* **10**(2).

Lewis CM, Knight J. 2012. Introduction to genetic association studies. *Cold Spring Harbor protocols* **2012**(3): 297-306.

Malenfant RM, Coltman DW, Davis CS. 2015. Design of a 9K illumina BeadChip for polar bears (Ursus maritimus) from RAD and transcriptome sequencing. *Molecular Ecology Resources* **15**(3): 587-600.

Manceau M, Domingues VS, Linnen CR, Rosenblum EB, Hoekstra HE. 2010. Convergence in pigmentation at multiple levels: mutations, genes and function. *Philosophical Transactions of the Royal Society B-Biological Sciences* **365**(1552): 2439-2450.

McTavish EJ, Hillis DM. 2015. How do SNP ascertainment schemes and population demographics affect inferences about population history? *Bmc Genomics* **16**.

Naderi S, Rezaei H-R, Pompanon F, Blum MGB, Negrini R, Naghash H-R, Balkiz O, Mashkour M, Gaggiotti OE, Ajmone-Marsan P et al. 2008. The goat domestication process inferred from large-scale mitochondrial DNA analysis of wild and domestic individuals. *Proceedings of the National Academy of Sciences of the United States of America* **105**(46): 17659-17664.

Palti Y, Gao G, Liu S, Kent MP, Lien S, Miller MR, Rexroad CE, III, Moen T. 2015. The development and characterization of a 57K single nucleotide polymorphism array for rainbow trout. *Molecular Ecology Resources* **15**(3): 662-672.

Pereira F, Queiros S, Gusmao L, Nijman IJ, Cuppen E, Lenstra JA, Davis SJM, Nejmeddine F, Amorim A, Econogene C. 2009. Tracing the History of Goat Pastoralism: New Clues from Mitochondrial and Y Chromosome DNA in North Africa. *Molecular Biology and Evolution* **26**(12): 2765-2773.

Perrey S, Rupp T. 2009. Altitude-Induced Changes in Muscle Contractile Properties. *High Altitude Medicine & Biology* **10**(2): 175-182.

Ramey HR, Decker JE, McKay SD, Rolf MM, Schnabel RD, Taylor JF. 2013. Detection of selective sweeps in cattle using genome-wide SNP data. *Bmc Genomics* **14**.

Rubin CJ, Megens HJ, Barrio AM, Maqbool K, Sayyab S, Schwochow D, Wang C, Carlborg O, Jern P, Jorgensen CB et al. 2012. Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences of the United States of America* **109**(48): 19529-19536.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**(7164): 913-U912.

Savolainen O, Lascoux M, Merila J. 2013. Ecological genomics of local adaptation. *Nature Reviews Genetics* **14**(11): 807-820.

Silva-Junior OB, Faria DA, Grattapaglia D. 2015. A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 Eucalyptus tree genomes across 12 species. *New Phytologist* **206**(4): 1527-1540.

Simonson TS, Yang Y, Huff CD, Yun H, Qin G, Witherspoon DJ, Bai Z, Lorenzo FR, Xing J, Jorde LB et al. 2010. Genetic Evidence for High-Altitude Adaptation in Tibet. *Science* **329**(5987): 72-75.

Steele-Perkins G, Plachez C, Butz KG, Yang GH, Bachurski CJ, Kinsman SL, Litwack ED, Richards LJ, Gronostajski RM. 2005. The transcription factor gene Nfib is essential for both lung maturation and brain development. *Molecular and Cellular Biology* **25**(2): 685-698.

Stinchcombe JR, Weinig C, Ungerer M, Olsen KM, Mays C, Halldorsdottir SS, Purugganan MD, Schmitt J. 2004. A latitudinal cline in flowering time in Arabidopsis thaliana modulated by the flowering time gene FRIGIDA. *Proceedings of the National Academy of Sciences of the United States of America* **101**(13): 4712-4717.

Stucki S, Orozco-terWengel P, Bruford MW, Colli L, Masembe C, Negrini R, Taberlet P, Joost S. 2014. High performance computation of landscape genomic models integrating local indices of spatial association. *arxiv* **1405.7658v2**.

Taberlet P, Valentini A, Rezaei HR, Naderi S, Pompanon F, Negrini R, Ajmone-Marsan P. 2008. Are cattle, sheep, and goats endangered species? *Molecular Ecology* **17**(1): 275-284.

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Research* **17**(4): 520-526.

Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, Dong Y, Donnadieu-Tonon C, Eggen A, Heuven HCM, Jamli S et al. 2014. Design and Characterization of a 52K SNP Chip for Goats. *Plos One* **9**(1).

Uhlik J, Konradova V, Vajner L, Adaskova J. 2005. Normobaric hypoxia induces mild damage to epithelium of terminal bronchioles in rabbits (ultrastructural study). *Veterinarni Medicina* **50**(10): 432-438.

Ward LD, Kellis M. 2012. Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology* **30**(11): 1095-1106.

Weir BS, Cockerham CC. 1984. ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION-STRUCTURE. *Evolution* **38**(6): 1358-1370.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS et al. 2010. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science* **329**(5987): 75-78.

Zhou J-J, Ma H-J, Liu Y, Guan Y, Maslov LN, Li D-P, Zhang Y. 2015. The anti-arrhythmic effect of chronic intermittent hypobaric hypoxia in rats with metabolic syndrome induced with fructose. *Canadian Journal of Physiology and Pharmacology* **93**(4): 227-232.

ANNEXES

Annexes

Annexe 1. Benjelloun B, Pompanon F, Ben Bati M, Chentouf M, Ibnelbachyr M, El Amiri B, Rioux D, Boulanouar B, Taberlet P. 2011. Mitochondrial DNA polymorphism in Moroccan goats. *Small Ruminant Research* **98**(1-3): 201-205.

Small Ruminant Research 98 (2011) 201–205

Contents lists available at ScienceDirect

Small Ruminant Research

journal homepage: www.elsevier.com/locate/smallrumres

Mitochondrial DNA polymorphism in Moroccan goats☆

B. Benjelloun[a,*], F. Pompanon[b], M. Ben Bati[a], M. Chentouf[c], M. Ibnelbachyr[d], B. El Amiri[e], D. Rioux[b], B. Boulanouar[f], P. Taberlet[b]

[a] Centre Régional de la Recherche Agronomique de Tadla, INRA, BP 567, Beni Mellal 23000, Morocco
[b] Laboratoire d'Ecologie Alpine, CNRS UMR 5553, Université Joseph Fourier, BP 53, 38041 Grenoble cedex 9, France
[c] Centre Régional de la Recherche Agronomique de Tanger, 78 Avenue Sidi Mohamed Ben Abdellah, Tanger, Morocco
[d] Centre Régional de la Recherche Agronomique de Errachidia, Avenue Moulay Ali Chérif Errachidia B.P. 2 Errachidia principale, Errachidia, Morocco
[e] Centre Régional de la Recherche Agronomique de Settat, Route tertiaire 1406, A 5 Km de Settat, Morocco
[f] Institut National de la Recherche Agronomique, Avenue Ennasr Rabat, Maroc, BP 415 RP Rabat, Morocco

ARTICLE INFO

Article history:
Available online 29 March 2011

Keywords:
Goats
Morocco
Genetic diversity
Haplotype
HVI control region

ABSTRACT

The present study characterizes the mitochondrial DNA diversity of Moroccan goats. 150 goats of different phenotypic entities were sampled over four geographic regions covering most of the Moroccan territory and the HVI segment of their mitochondrial DNA (mtDNA) control region was sequenced. The 150 Moroccan goats represented 97 haplotypes for this mtDNA fragment. Most of this diversity was present within phenotypic entities and within geographic regions. This weak genetic structure may result from the fact that all haplotypes were already mixed in the populations that colonized Morocco and/or the existence of recurrent gene flows from Mediterranean routes. Comparing the Moroccan haplotype diversity to that of 21 haplotypes representative of the worldwide diversity showed that all the Moroccan goats studied belonged to the A haplogroup that is preponderant in the world. The haplotypes of the Northern region appeared to be less diverse, what would probably reflect a stronger founder effect in this region.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Small ruminants have a major socio-economic and ecological role throughout the world and allow the production of 11.1 million of tons of meat per year. This is especially true in Morocco, where the breeding of small ruminants interests more than 65% of the rural population (MADRPM, 2004) and plays an essential role in the rural economic activity. However, only a few studies have described the genetic structure of these local populations until now. For goats, it is difficult to distinguish well-defined Moroccan breeds. However, several investigations have described typical populations with specific localization and phenotypic characteristics. The three major local populations are: (i) the Black population of the Atlas with two subpopulations; (ii) the Northern population; (iii) and the population Draa.

The study of microsatellites and casein genes highlighted the strong genetic polymorphism of such Moroccan populations (Tadlaoui Ouafi et al., 2002). They are characterized by a high diversity and a significant heterogeneity due to an uncontrolled mixing between various populations. A comprehensive study of mitochondrial and Y chromosome DNA diversity in Northern African goats (Pereira et al., 2009) confirmed the high level of variability and suggested that their colonization was influenced by recurrent gene flows through maritime routes (instead of a unidirectional terrestrial route) as

☆ This paper is part of the special issue entitled: Technological development and associative attempts to a sustainable goat production – A selection of Plenary and Oral presentations from the 10th International Conference on Goats, Guest edited by Dr. Marta Madruga.

* Corresponding author. Tel.: +212 5 23 44 00 06, fax: +212 5 23 44 00 83.
E-mail address: badr.benjelloun@gmail.com (B. Benjelloun).

0921-4488/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.smallrumres.2011.03.041

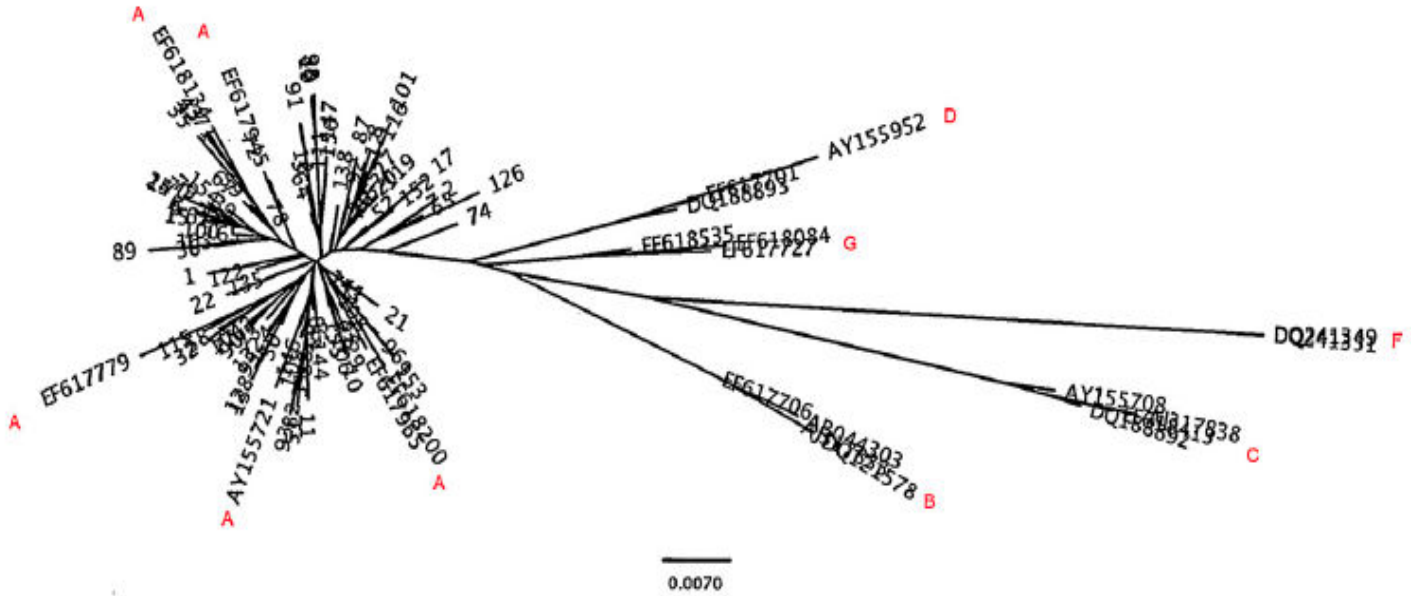

Fig. 1. Neighbor-joining tree of the 97 Moroccan haplotypes and 21 reference haplotypes representing the worldwide diversity. Codes starting with two letters represent the GenBank Accession numbers of the 21 reference sequences, the other numbers correspond to the Moroccan haplotypes. The red letters give the name of the 6 haplogroups identified in the world. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Table 1
Distribution of genetic variance in populations and geographic regions revealed by AMOVA.

Source of variation	Within populations	Among populations	Within regions	Among regions
d.f.	3	146	3	146
% Variation	7,5	92,5	6,8	93,2
P value	$P < 0.001$		$P < 0.001$	

well as by gene flows between Morocco and the Iberian Peninsula.

The present study aimed to characterize the mitochondrial diversity of Moroccan goats (HVI segment of the control region) in relation with the haplotype diversity already described at the worldwide level (Naderi et al., 2007).

2. Materials and methods

Samples were collected from 150 goats from four main geographic regions: (i) the plains of the central region (provinces of Beni Mellal and Khouribga) (ii) the mountains of the central region (province of Azilal) (iii) the Northern Region (provinces Tangier, Larache and Chefchaouen), and (iv) the South-Eastern area (provinces of Errachidia, Ouarzazate and Zagora). These goats belonged to four entities: (i) Northern population, (ii) Black population of the Atlas, (iii) Draa breed and (iv) other phenotypes.

Tissue samples from the distal part of the ear were collected and placed in alcohol for one day, and then transferred to a tube filled with silica gel until extraction. DNA was extracted using the Qiagen DNeasy tissue kit following the manufacturer's instructions. The HVI segment of the control region was sequenced using the primers CAP-F (5′-CGTGTATGCAAGTACATTAC-3′) and CAP-R (3′-CTGATTAGTCATTAGTCCATC-5′) that amplified a fragment of 598 bp (without primers) corresponding to the positions 15,653 to 16,250 on the complete goat mitochondrial sequence of reference (Parma et al., 2003; GenBank accession number AF533441). PCR amplifications were conducted in a 25 μl volume with 2 mM $MgCl_2$, 200 mM of each dNTP, 1 mM of each primer and 1 unit of AmpliTaq Gold Polymerase (Applied Biosystems). After a 10 min period at 95 °C for polymerase activation, 35 cycles were run with the following steps: 95 °C: 30 s, 55 °C: 30 s, 72 °C: 1 min. PCR products were purified using the Qiaquick PCR purification kit (Qiagen). 35 ng of purified DNA from this PCR product was used for sequencing with the CAP-F or CAP-R primers. Sequence reactions were performed for both DNA strands with the CAP-F or CAP-R primers by using the ABI PRISM Dye Terminator Cycle Sequencing Reaction Kit (Applied Biosystems) in a 20 ml volume with 2 mM of each primer. 25 cycles were run with the following steps 96 °C: 30 s, 55 °C: 30 s, 60 °C: 4 min. Excess dye terminators were removed by spin-column purification and the products were electrophorized on an ABI 3130 PRISM DNA sequencer (Applied Biosystems) using the POP 7 polymer.

The sequences obtained were edited for correction with SeqScape v2.5 (Applied Biosystems). They were aligned together with 21 reference sequences (Naderi et al., 2007) using Mega v3.1 (Kumar et al., 2004), and then adjusted by eye. For analyses, we kept the 481 bp long region (GenBank Accession Numbers HQ455369–HQ455518) usually used for characterizing the goat mitochondrial diversity (e.g., Luikart et al., 2001; Naderi et al., 2007). AMOVA were performed on the Moroccan dataset using ARLEQUIN v3.0 (Excoffier et al., 2005) in order to test the partition of the genetic variance among and within populations and geographic areas. A median joining network representing the relationships between haplotypes was drawn using NETWORK v4.5.1.6. A Neighbor-joining tree was constructed using the Moroccan haplotypes and the 21 haplotypes representing 6 domestic goat haplogroups identified in the world (Naderi et al., 2007).

3. Results

The HV1 fragment of the control region showed a high polymorphism in Moroccan goat populations, with a haplotype diversity (Hd) of 0.9925. The 150 sequences had 98 variable sites over 481, and corresponded to 97 haplotypes. The Neighbor-joining tree (Fig. 1) made with these 97 haplotypes and 21 haplotypes representing the diversity of the 6 haplogroups found over the world (i.e., A, B, C, D, F and G, Naderi et al., 2007) showed that the 150 goats studied were from the haplogroup A. Even if the AMOVA showed a significant population effect on the mt variation ($P < 0.001$), 92% of this variation was distributed within populations (Table 1). Also, the differentiation between geographic regions was low but significant ($P < 0.001$), and about 93% of the genetic variability was distributed within regions (Table 1).

4. Discussion

4.1.1. Moroccan goats' diversity in the worldwide context

The very high haplotype diversity obtained for the Moroccan goats is similar to that measured by Naderi et al. (2007) on 1440 haplotypes from the A haplogroup across the world. It also confirms the high diversity already described in Moroccan and other Northern African goats (Pereira et al., 2009). This high diversity would result from the heterogeneity of domestic goat populations that first colonized Morocco. Because the goat domestication process is recent at the evolutionary timescale (about 10,000 years ago), only a few mutations should have occurred since the initial steps of domestication despite the high mutation rate of the control region. Thus, the high diversity observed now would result from the capture of a large part of the wild diversity during domestication. This has been confirmed by the comparison of the genetic diversity of domestic goats to that of their wild ancestor, showing among others large initial effective population size for domestic goats (see Naderi et al., 2007).

4.2. Structure of Moroccan diversity

The high level of variability observed within populations and geographic regions is consistent with the high diversity found within geographical regions at the worldwide scale (Sultana et al., 2003; Joshi et al., 2004; Chen et al., 2005; Pereira et al., 2005; Naderi et al., 2007). The median joining network of the Moroccan haplotypes (Fig. 2) confirms the high diversity of haplotypes within geographic regions. However, it points out a tendency for a higher similarity between haplotypes from the Northern region. The results are in accordance with two non-exclusive hypotheses that are (i) the high heterogeneity of founder populations in Morocco due to a mix of haplotypes in the first domesticated populations (Naderi et al., 2007), and (ii) the recurrent influx of genetic diversity via the Mediterranean Sea (Pereira et al., 2009). The higher similarity between Northern haplogroups would probably result

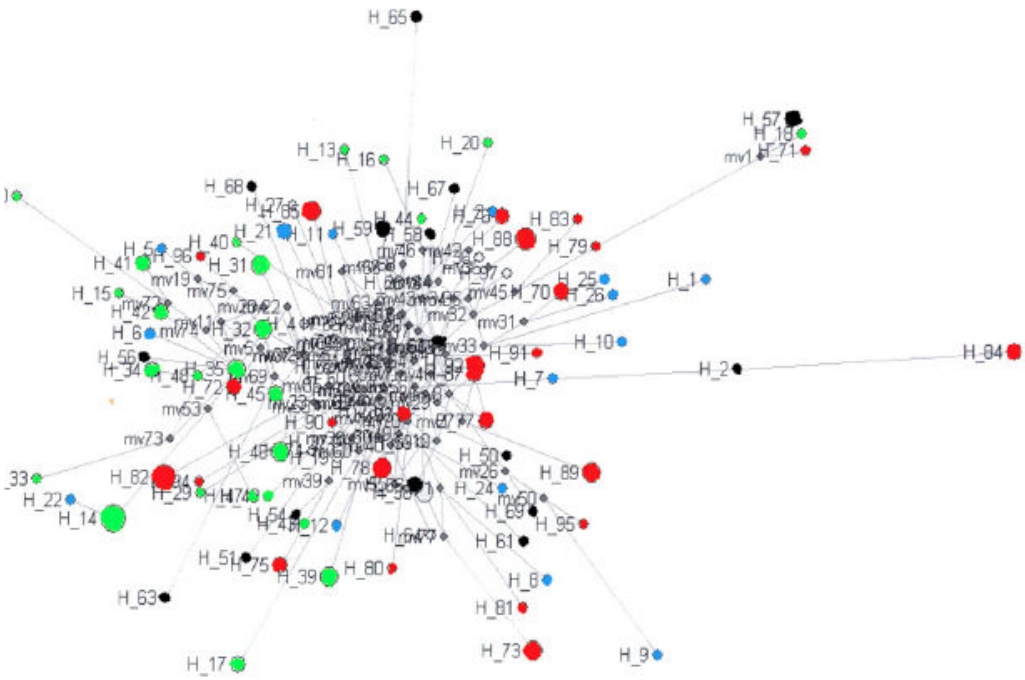

Fig. 2. Phylogenetic network based on the HVI control region haplotypes. The different colours are related to the geographical origin. Green: Northern region; Red: South-Eastern region; Blue: mountains of the center; Black: plains of the center. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

from a strong founder effect, probably from populations originating in the Iberian Peninsula, given the past occurrence of admixture across the strait of Gibraltar (Anderung et al., 2005; Pereira et al., 2009).

5. Conclusion

The 150 Moroccan goats studied are all from the A haplogroup that is predominant in the world and especially in Western Europe. It appears that, as elsewhere in the world, these populations are characterized by a high mitochondrial diversity that is very weakly structured according to regions and breeds/populations. This variability would reflect the diversity already present in the first domesticated individuals that arrived in Morocco and/or the existence of recurrent gene flows from Mediterranean routes. Despite the high diversity within geographic regions, the Northern population seems more homogeneous, probably because of a stronger founder at the origin of these populations.

Complementary studies on selected genes would allow understanding the adaptive history of these populations in relation with local conditions (i.e., climate, pathogenic context, breeding system, etc.).

Conflict of interest

None.

Acknowledgements

Laboratory work of this study was conducted in the Laboratoire d'Ecologie Alpine, Université Joseph Fourrier, Grenoble, France, and has been supported by the International Atomic Energy Agency (IAEA) into the Technical Project no. MOR 5030.

References

Anderung, C., Bouwman, A., Persson, P., Carretero, J.M., Ortega, A.I., Elburg, R., Smith, C., Arsuaga, J.L., Ellegren, H., Gotherstrom, A., 2005. Prehistoric contacts over the straits of Gibraltar indicated by genetic analysis of Iberian Bronze age cattle. Proc. Natl. Acad. Sci. U.S.A. 102, 8431–8435.

Chen, S.Y., Su, Y.H., Wu, S.F., Sha, T., Zhang, Y.P., 2005. Mitochondrial diversity and phylogeographic structure of Chinese domestic goats. Mol. Phylogenet. Evol. 37 (3), 804–814.

Excoffier, L., Laval, G., Schneider, S., 2005. Arlequin ver3. 0: an integrated software package for population genetics data analysis. Evol. Bioinform. Online 1, 47–50.

Joshi, M.B., Rout, P.K., Mandal, A.K., Tyler-Smith, C., Singh, L., et al., 2004. Phylogeography and origin of Indian domestic goats. Mol. Biol. Evol. 21 (3), 454–462.

Kumar, S., Tamura, K., Nei, M., 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignement. Brief Bioinform. 5, 150–163.

Luikart, G., Gielly, L., Excoffier, L., Vigne, J.D., Bouvet, J., et al., 2001. Multiple maternal origins and weak phylogeographic structure in domestic goats. Proc. Natl. Acad. Sci. U.S.A. 98, 5927–5932.

MADRPM, 2004. L'élevage en Chiffres 2003. Direction de l'Elevage, Rabat, Maroc.

Naderi, S., Rezaei, H.-R., Taberlet, P., Zundel, S., Rafat, S.-A., et al., 2007. Large-scale mitochondrial DNA analysis of the domestic goat reveals six haplogroups with high diversity. PLoS ONE 2 (10), e1012.

Parma, P., Feligini, M., Greeppi, G., Enne, G., 2003. The complete nucleotide sequence of goat (*Capra hircus*) mitochondrial genome. Goat mitochondrial genome. DNA Seq. 14 (3), 199–203.

Pereira, F., Pereira, L., Van Asch, B., Bradley, D., Amorim, A., 2005. The mtDNA catalogue of all Portuguese autochthonous goat (*Capra hircus*) breeds: high diversity of female lineages at the western fringe of European distribution. Mol. Ecol. 14, 2313–2318.

Pereira, F., Queirós, S., Gusmão, L., Nijman, I.J., Cuppen, E., Lenstra, J.A., Econogene, C., Davis, S.J.M., Nejmeddine, F., Amorim, A, 2009. Tracing the history of goat pastoralism: new clues from mitochondrial

and Y chromosome DNA in North Africa. Mol. Biol. Evol. 26, 2765–2773.
Sultana, S., Mannen, H., Tsuji, S., 2003. Mitochondrial DNA diversity of Pakistani goats. Anim. Genet. 34 (6), 417–421.
Tadlaoui Ouafi, A., Babilliot, J.-M., Leroux, C., Martin, P., 2002. Genetic diversity of the two main Moroccan goat breeds: phylogenetic relationships with four breeds reared in France. Small Rumin. Res. 45, 225–233.

Résumé

Les progrès technologiques récents nous permettent d'accéder à la variation des génomes complets ce qui nous ouvre la porte d'une meilleure compréhension de leur diversification via des approches de génomique des populations et de génomique du paysage. Ce travail de thèse se base sur l'analyse des données de génomes complets (WGS) pour caractériser la diversité génétique des petits ruminants (chèvre et moutons) et rechercher les bases génétiques de l'adaptation locale.

Dans un premier temps, ce travail appréhende un aspect méthodologique et examine la précision et le biais de différentes approches d'échantillonnage des génomes pour caractériser la variabilité génétique, en les comparant aux données WGS. Nous mettons en évidence un fort biais des approches classiques (i.e. puces à ADN, capture de l'exome) ainsi que des séquençages de génomes à faibles taux de couverture (1X et 2X), et nous suggérons des alternatives basées sur un échantillonnage aléatoire de marqueurs dont la densité est variable selon les objectifs de l'étude (évaluation de la diversité neutre, déséquilibre de liaison, signatures de sélection). Le jeu de données produit a permis d'évaluer l'état des ressources génétiques de différentes populations domestiques (races locales marocaines, iraniennes, races industrielles) et sauvages (aegagres, mouflons asiatiques). Nous relevons une très forte diversité génétique dans les populations indigènes et sauvages qui constituent des réservoirs d'allèles et peuvent jouer un rôle important pour préserver le potentiel adaptatif des petits ruminants domestiques dans un contexte de changement climatique. L'étude plus approfondie des populations de chèvres du Maroc montre une forte diversité génétique faiblement structurée géographiquement, et met en évidence des portions de génome présentant des signaux de sélection. Leur étude montre l'existence de mécanismes adaptatifs potentiellement différents selon les populations (e.g. transpiration/halètement dans l'adaptation probable à la chaleur).

Enfin, nous explorons les bases génétiques de l'adaptation locale à l'environnement chez les moutons et les chèvres via une approche de génomique du paysage. En scannant les génomes de 160 moutons et 161 chèvres représentant la diversité éco-climatique du Maroc, nous identifions de nombreux variants et gènes candidats qui permettent d'identifier les voies physiologiques potentiellement sous-jacentes à l'adaptation locale. En particulier, il apparait que les mécanismes respiratoires et les processus cardiaques joueraient un rôle clé dans l'adaptation à l'altitude. Les résultats suggèrent que les chèvres et moutons ont probablement développé différents mécanismes adaptatifs pour répondre aux mêmes variations environnementales. Cependant, nous identifions plusieurs cas probables de voies adaptatives communes à plusieurs espèces. Par ailleurs, nous avons caractérisé les patrons de variations du niveau de différenciation de régions chromosomiques sous sélection en fonction de l'altitude. Cela nous permet de visualiser la diversité des réponses adaptatives selon les gènes (par exemple, sélection de variants à faible et/ou haute altitude). Ainsi, ce travail pose les bases pour une meilleure compréhension de certains mécanismes génomiques de l'adaptation locale.

Mots clés : génomes complets, adaptation locale, génomique des populations, génomique du paysage, petits ruminants, ressources génétiques

Abstract

Recent technological developments allow an unprecedented access to the whole genome variation and would increase our knowledge on genome diversification using population and landscape genomics. This work is based on the analysis of Whole Genome Sequence data (WGS) with the purpose of characterising genetic diversity in small ruminants (sheep and goats) and exploring genetic bases of local adaptation.

First, we addressed a methodological aspect by investigating the accuracy and possible bias in the widely used genotyping approaches to characterize genetic variation in comparison with WGS data. We highlighted strong bias in conventional approaches (SNP chips and exome capture) and also in low-coverage whole genome re-sequencing (1X and 2X), and we suggested effective solutions based on sampling panels of random markers over the genome depending the purpose of the study (assessing neutral diversity, linkage disequilibrium, selection signatures). The various datasets produced allowed assessing genetic resources in various domestic (Moroccan and Iranian indigenous breeds and industrials) and wild populations (bezoars and Asiatic mouflons). We identified a very high diversity in indigenous and wild populations. They constitute a reservoir of alleles allowing them to play a possible key role in the preservation of these species in the context of global changes. The deep study of Moroccan goats showed a high diversity that is weakly structured in geography and populations, and highlighted numerous genomic regions showing signatures of selection. These regions identified different putative adaptive mechanisms according to the population (e.g. panting/sweating to adapt to warm/desert environment).

Then, we explored genetic bases of local adaptation to the environment in sheep and goats using a landscape genomics framework. We scanned genomes of 160 sheep and 161 goats representing the eco-climatic Moroccan-wide diversity. We identified numerous candidate variants and genes, which allowed for identifying physiological pathways possibly underlying local adaptation. Especially, it seems that respiration and cardiac process have key roles in the adaptation to altitude. Our results suggest dissimilar adaptive mechanisms for the same environment in sheep and goats. However, we highlighted several cases of common metabolic pathways in different species. Moreover, we characterized some patterns for the variation of genetic differentiation in some candidate genomic regions over environmental gradients. This allowed us to visualise different adaptive reaction depending genes. This work points the way towards a better understanding of some mechanisms underlying local adaptation.

Key words: whole genomes, local adaptation, population genomics, landscape genomics, small-ruminants, genetic resources