Etude des variants structuraux génomiques pour comprendre les processus démographiques et adaptatifs impliqués dans la domestication des petits ruminants

Tristan Cumer

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : Biodiversité-Ecologie-Environnement

Arrêté ministériel : 25 mai 2016

Présentée par

Tristan CUMER

Thèse dirigée par **François POMPANON (CSV)**, , UGA

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

Etude des variants structuraux génomiques pour comprendre les processus démographiques et adaptatifs impliqués dans la domestication des petits ruminants

Genome structural variations to understand the adaptive and demographic processes during domestication of small ruminants.

Thèse soutenue publiquement le **13 décembre 2017**, devant le jury composé de :

Monsieur FRANÇOIS POMPANON
PROFESSEUR, UNIVERSITE GRENOBLE ALPES, Directeur de thèse
Monsieur OLIVIER HANOTTE
PROFESSEUR, UNIVERSITE DE NOTTINGHAM, Rapporteur
Monsieur THOMAS FARAUT
CHARGE DE RECHERCHE, INRA CENTRE TOULOUSE MIDI-PYRENEES, Examinateur
Monsieur MICHAEL BLUM
DIRECTEUR DE RECHERCHE, CNRS DELEGATION ALPES, Président
Madame DOMINIQUE MOUCHIROUX
PROFESSEUR, UNIVERSITE LYON 1, Rapporteur

Table des matières

INTRODUCTION GÉNÉRALE

Comment expliquer la diversité du vivant ? Pourquoi des différences existent-elles entre les individus d'une même espèce, et quelles en sont les conséquences ? Pourquoi les espèces semblent-elle adaptées à leur environnement ? Qu'advient-il lorsque les espèces sont confrontées à des environnements nouveaux ?
Voilà certaines des questions auxquelles la biologie évolutive essaye de répondre. C'est notamment la théorie de l'évolution telle qu'énoncée par Charles Darwin et complétée par la suite, qui offre un cadre robuste à l'analyse du vivant.

Contexte : la théorie de l'évolution

Au cours de son voyage aux bords du *Beagle*, le naturaliste Charles Robert Darwin (1809-1882) fit de nombreuses observations. De ces observations, il établit la théorie qui le rendit célèbre. Cette théorie est rendue publique dans son ouvrage *The origin of species*, livre dans lequel Charles Darwin pose les bases de la théorie de l'évolution des espèces. Il explique cette évolution par le jeu de la sélection naturelle, qui agit sur les différences préexistantes chez les êtres vivants. La sélection naturelle représente alors l'effet des facteurs externes (climat, compétition, prédation …) sur la capacité des organismes à survivre, se reproduire et à transmettre ces différences. De ces travaux, Charles Darwin conclut à une évolution naturelle des espèces : les individus qui ont hérité de caractères bien adaptés à leur milieu ont tendance à plus se reproduire que leurs congénères et ainsi prendre le pas sur eux (Darwin, 1872).
Indépendamment des travaux de Darwin, Johann Gregor Mendel, moine et botaniste tchéco-allemand (22 juillet 1822 - 6 janvier 1884), s'intéresse aux caractères héréditaires, transmissibles de génération en génération. Jusqu'alors, la communauté scientifique soutenait le modèle de l'hérédité par mélanges où les caractères observés chez un individu, étaient intermédiaires entre ceux de ses parents (le croisement d'un parent grand et d'un parent petit donnant par exemple un individu de taille intermédiaire). Mendel va étudier la transmission de différents caractères au fil des générations, et de son travail ressortent les lois de Mendel, décrivant la manière dont les gènes se transmettent de génération en génération. Ces travaux vont poser les bases de la génétique des populations actuelles.
La fusion de la théorie de Darwin et des lois de Mendel, complétée par la découverte de l'ADN, support matériel de l'hérédité, a permis d'interpréter l'évolution en termes de changements de proportion entre les différentes versions des gènes dans les groupements d'individus de même espèce, ou populations. C'est sur cette théorie que se basent les travaux de Julian Huxley, Ronald Fisher ou encore Ernst Mayr pour affirmer qu'une population évolue quand la fréquence des différentes versions d'un gène, les allèles, s'y modifie. On voit alors se répandre dans la population des caractères qui apportent une plus value aux individus qui en sont porteurs, caractères que l'on observe rapidement (aux échelles de temps des espèces) dans une grande partie de la population, voire de toute l'espèce. Lorsque plusieurs populations d'une même espèce sont isolées les unes des autres, chacune de ces populations peut acquérir des caractères distinctifs. Si ces populations sont dans l'impossibilité d'échanger des migrants, réduisant leurs divergences, celles-ci peuvent aller jusqu'à entraîner une inter-stérilité : elles constituent alors deux espèces différentes (Queiroz, 2005).

En complément à cette vision de l'évolution dirigée uniquement par la sélection, la théorie neutraliste de l'évolution formalisée par Motoo Kimura, pose l'hypothèse de neutralité. Selon cette hypothèse, certaines mutations n'ont aucune influence sur la valeur sélective des individus, leurs fréquences dans les populations vont alors dépendre d'évènements stochastiques (Kimura, 1983).

L'ADN : support de l'hérédité.

L'Acide DesoxyriboNucléique (ADN) a été isolé pour la première fois en 1869 par le biologiste suisse Friedrich Miescher sous la forme d'une substance riche en phosphore (Dahm, 2008). En 1927, le biologiste russe Nikolai Koltsov suggérait que l'hérédité reposait sur une « molécule héréditaire géante » constituée de « deux brins miroirs l'un de l'autre qui se reproduiraient de manière semi-conservative en utilisant chaque brin comme modèle » (Marshak, 1936). Les travaux de Frederick Griffith en 1928 (Lorenz and Wackernagel, 1994), suivi de ceux de d'Avery, MacLeod et McCarty en 1944 (Avery *et al.*, 1995) puis de ceux de Jean Brache en 1946 (Hershey and Chase, 1952) ont permis de montrer que la molécule géante dont parle Nikolai Koltsov est bel et bien l'ADN isolé par Mieischer, et confirment que cette molécule est le support de l'hérédité.
Cette longue molécule est présente dans toutes les cellules de chaque être vivant, et est formée de deux brins antiparallèles formant une double hélice (Watson and Crick, 1953). Chacun de ces brins est formé de nucléotides, eux mêmes constitués d'une base azotée (Adénine (A), Thymine (T), Cytosine (C), Guanine (G)) et liés à un désoxyribose, lui-même lié à un groupe phosphate. Les nucléotides sont liés les uns aux autres par des liaisons entre le désoxyribose d'un nucléotide et le groupe phosphate du nucléotide suivant.

Le long de cette longue séquence, plus de 2,8 milliards de paires de bases chez l'homme (Consortium and others, 2004), se dispersent des *locus* qui vont être transcrits en ARN qui seront eux même traduits en protéines.
Ces séquences font partie des gènes définis comme l'union des séquences génomiques codantes pour un ensemble de produits fonctionnels cohérents et potentiellement chevauchants (Gerstein *et al.*, 2007).

Ces gènes ne représentent pourtant qu'une part de l'ADN des espèces. Le reste de la séquence d'ADN (plus de 98 % chez l'homme (Elgar and Vavouri, 2008)) a dans un premier temps été considéré comme inutile pour les organismes, et qualifié d'ADN poubelle (Ohno, 1972). Ces séquences ne sont pas pour autant sans fonction. Nous savons maintenant que ces séquences se composent par exemple d'éléments régulant l'expression des gènes, d'ARN non codants, de gènes ayant perdu leurs fonctions suite à des mutations, les pseudogènes, ou encore des séquences répétées (transposons, ADN viraux endogènes, etc …) (ENCODE Project Consortium, 2012).

De nombreuses différences observables entre populations ont des bases génétiques (e. g. couleur de peau chez l'homme (Barsh, 2003)), et ce sont ces différences génétiques, les mutations, qui vont être le combustible de l'évolution.
Bien avant les techniques de séquençage modernes, les premières différences observées entre les génomes de différents individus étaient rares, et consistaient principalement en des changements de quantité et de structure des chromosomes.
Un exemple bien connu de différence génétique avec une forte incidence sur le phénotype (ensemble des traits observables chez un individu) est la trisomie 21, où la présence d'un chromosome 21 surnuméraire va causer le syndrome de Down (Patterson, 2009). Cependant, des différences plus petites peuvent elles aussi avoir des conséquences fortes sur les individus. Ainsi, Barbara McClintock, en étudiant les patrons de coloration du maïs, démontra que certains éléments, les transposons, étaient capables de se déplacer dans le génome. Chez le maïs, cette transposition rétablit la synthèse du pigment dans le grain de maïs initialement sans pigment, induisant le paterne de mosaïque (McClintock, 1953). Cette découverte des transposons lui valut le prix Nobel de médecine en 1983, et ouvre l'étude du polymorphisme génétique.
L'arrivée des technologies de séquençage à partir de la fin des années 1970 (Sanger and Coulson, 1975), et plus récemment avec l'essor des technologies de séquençage à haut débit, a donné accès aux génomes complets des individus, et ainsi aux mutations microscopiques et sub-microscopiques (Mukhopadhyay, 2009).

L'ADN : support de l'évolution

Parmi des différences de séquences, allant du remplacement d'une base azotée par une autre aux variation du nombre de chromosomes comme dans le cas de la trisomie, tout un panel de possibilités existe.
L'ensemble de ces mutations se regroupe généralement sous différentes catégories telles que les substitutions (changement d'une base azotée par une autre), les insertions et délétions (apparition ou disparition de certaines bases azotées), les inversions (changement dans l'ordre des bases azotées) ou les translocations (une séquence d'ADN est déplacée sur un autre chromosome, ou plus loin sur le même chromosome). Pour chacune de ces catégories, les mécanismes à l'origine sont divers. Par exemple, dans le cas d'une insertion, il peut s'agir d'un ADN viral ou retro-viral, qui s'insère après l'infection de l'individu. Il peut aussi s'agir d'une seule base azotée ajoutée par erreur lors de la réplication de l'ADN. Les conséquences de ces différences sont elles aussi variables, de la mutation neutre n'impactant pas du tout son porteur aux mutations létales ne permettant pas aux organismes de se développer.
L'idée ici n'est pas de faire l'inventaire des mécanismes génétiques conduisant au polymorphisme observable entre les individus, ni de s'attarder sur l'ensemble des conséquences possibles d'une mutation, mais d'avoir une vision de l'ADN non pas comme une molécule statique, transmissible en l'état de génération en génération, mais bien d'une entité dynamique où les prises pour l'évolution sont nombreuses.

Ainsi, les mutations composent la première des forces évolutives en générant les différences entre les individus. Ce sont pourtant les autres forces évolutives que sont la dérive, la sélection et la migration, qui font que les populations évoluent.
La dérive correspond aux fluctuations de fréquences des différents allèles d'un gène dans une population lorsque celle-ci n'est pas soumise à une autre force évolutive. Il s'agit alors d'évolution des fréquences alléliques au gré de croisements aléatoires entre les individus. La sélection, quant à elle, entre en jeu lorsque la mutation a un impact sur le phénotype et la fitness (capacité d'un individu à survivre et se reproduire) de l'individu. Dans le cas où l'impact est négatif, la mutation va être contre-sélectionnée, et sa fréquence dans la population baisser. Dans le cas où l'impact est positif pour l'individu, cette mutation va être sélectionnée favorablement et sa fréquence augmenter dans la population.
Enfin, la migration décrit le fait que des allèles vont être apportés d'une autre population via l'arrivée de nouveaux individus dans la population étudiée. Cette force peut par exemple permettre d'éviter la fixation d'allèles par dérive dans le cas de petites populations, ou réduire l'efficacité de la sélection en introduisant des allèles inadaptés dans la population.

Ce sont ces quatre forces évolutives (mutation, dérive, sélection et migration) qui vont dicter la distribution des différents allèles dans les populations. *A posteriori*, l'étude des fréquences alléliques de différentes régions génomiques peut permettre de comprendre quelles forces ont conduit au résultat observé et ainsi retracer l'histoire de cette population.

La domestication, un cas d'école dans l'étude de l'évolution

Au cours de l'histoire des espèces, de nombreux évènements peuvent entraîner des modifications de pressions des différentes forces évolutives. Ce sont d'ailleurs ces changements de pressions qui peuvent être à l'origine de l'apparition de nouvelles espèces (i.e. les radiation adaptatives) lorsque deux populations d'une même espèce sont soumises à des contraintes évolutives différentes (Gavrilets and Losos, 2009). La domestication est un bon exemple de situation où des individus d'une même espèce rencontrent des situations très différentes et pouvant aboutir à des différences fortes entre individus domestiques et sauvages.

D'un point de vue anthropologique, l'apport de la domestication aux sociétés humaines n'est plus à prouver. L'apparition de l'agriculture s'est traduite par la gestion plus ou moins marquée d'un certain nombre d'espèces par l'homme pour subvenir à ses besoins, aboutissant *in fine* à la domestication de ces espèces. Ces changements de mode de vie de l'homme marquent la transition entre le Paléolithique (-3 Million d'années à -12000 ans) et le Néolithique (-12000 à -3500 ans). Le Paléolithique est caractérisé par l'utilisation d'outils en pierre plus ou moins évolués, par des populations humaines exclusivement composées de chasseurs cueilleurs (Henke, 1944). Le Néolithique est une période marquée par l'adoption, par les groupes humains, d'une économie de production fondée sur l'agriculture et l'élevage, induisant le plus souvent une

sédentarisation (Cauwe *et al.*, 2007). Cette transition, aussi appelée révolution néolithique (Childe, 1925), marque l'un des changements majeurs dans l'histoire humaine. Cette transition a entraîné le passage de sociétés simples de chasseurs cueilleurs à des sociétés sédentaires, avec des ressources en plus grandes quantités et stables dans le temps, permettant une augmentation des densités de population.

D'un point de vue évolutif, la domestication représente un cas très particulier d'interaction interspécifique. Une définition formelle de la domestication la définit comme une relation mutualiste multi-générationnelle dans laquelle un organisme influence la reproduction et procure des soins à une autre espèce, de façon à s'assurer un apport en ressources ou en services prévisible, et au travers de laquelle l'organisme partenaire gagne un avantage sur les individus qui restent en dehors de la relation. De cette façon, la domestication augmente la fitness du domestiquant et du domestiqué (Zeder, 2015). La domestication peut, à bien des égards, être considérée comme une expérience évolutive dans laquelle l'espèce domestiquée et le domestiquant sont tout deux interdépendants pour leur survie (Zeder *et al.*, 2006).

Parmi l'ensemble des espèces domestiquées, les espèces animales représentent des cas intéressants de plusieurs points de vue. Les animaux domestiques remplissent aujourd'hui de nombreux objectifs. En plus de fournir de la nourriture (viande et lait principalement), nous procurent aussi de nombreux produits (laine, cuir) et des services (compagnie, protection) (Larson and Fuller, 2014). De plus, la différenciation est souvent forte entre les animaux domestiques et leurs relatifs sauvages les plus proches. Si chez les plantes de nombreuses différences morphologiques sont observables (c'est par exemple le cas du maïs et de la téosinte (Doebley *et al.*, 1997)), les animaux domestiques se différencient des sauvages par tout un panel de différences morphologiques, physiologiques et comportementales. La convergence phénotypique observable entre les mammifères domestiques, aussi connue comme syndrome de domestication (encadré 1) (Wilkins *et al.*, 2014), en font de bons modèles pour l'étude ce cette expérience évolutive au long court.

Encadré 1 : Le renard domestique, une domestication contrôlée

Dmitri Beliaïev (1917-1985) est un scientifique soviétique et russe qui a longuement travaillé sur la domestication. Il avait remarqué chez les chiens adultes le maintien de traits juvéniles, à la fois morphologiques (crânes plus larges que la normale par rapport à leur longueur), et comportementaux (gémissements, aboiements et attitudes de soumission). Il fit l'hypothèse que les facteurs sélectionnés lors de la domestication n'étaient ni la taille ni la reproduction, mais des traits comportementaux, en particulier la propension à la domestication via la docilité.
Il testa sa théorie sur le renard argenté, une espèce jamais domestiquée auparavant. Il a sélectionné les animaux en fonction de la faible distance de fuite, c'est-à-dire la distance minimale à laquelle l'animal pouvait être approché jusqu'à ce qu'il cherche à fuir.
Après plus de 40 années de sélection pour la docilité, les renards domestiques présentent des traits jamais sélectionnés, avec par exemple un pelage tacheté de blanc, des oreilles tombantes, la queue enroulée sur elle même, le museau raccourci et un ralentissement dans le développement.
Cette expérience illustre que des caractéristiques non sélectionnées peuvent apparaître conjointement à la docilité. Ces traits, nouvellement apparus chez le renard domestique, font partie d'un ensemble de traits reconnu comme le syndrome de domestication chez de nombreux animaux (chien, cheval, mouton chèvre …).

Avec l'essor de la biologie moléculaire, de nombreuses études se sont intéressées aux aspects génétiques de la domestication. Dans un premier temps, ces données moléculaires couplées aux données archéologiques ont permis d'identifier les espèces sauvages domestiquées, les lieux de ces domestications, ainsi que les déterminants de ces évènements (encadré 2) (Larson and Fuller, 2014). Ces données ont aussi permis d'explorer l'histoire de ces espèces domestiques, au travers de leurs domestications et de leurs migrations, et ainsi d'apporter une lumière nouvelle sur l'histoire de l'homme.
Les données génétiques à haute densité ont, par la suite, permis d'explorer ce qui a été appelé les bases génétiques de la domestication. Dans ce contexte, ces études ont permis de mettre en avant un certains nombre de mutations génétiques expliquant les traits physiques et comportementaux observables chez les animaux domestiques (Wright, 2015).

Encadré 2 : Les voies de la domestication

Pour expliquer l'origine des espèces domestiques, Francis Galton suggéra qu'à partir de la capture et le soin de louveteaux, l'homme put les domestiquer, entraînant l'apparition à terme des chiens (Galton, 1883). Cette vision simple des débuts de la domestication implique une volonté de l'homme à domestiquer des animaux. Si cette hypothèse ne semble pas un mécanisme explicatif convainquant pour Serpell (Clutton-Brock, 2014), il faut attendre des travaux récents pour entrevoir une explication convaincante. Zeder décrit la domestication comme un processus graduel au cours duquel la relation entre les humains et les animaux s'intensifie. La domestication des animaux se fait le long d'un gradient de l'interaction simple interspécifique (commensalisme, proie/prédateur) au contrôle d'individus sauvages suivi du contrôle d'individus captifs puis de l'élevage extensif et enfin de l'élevage intensif (Zeder, 2012). Trois voies de domestication semblent alors possibles pour expliquer le passage de l'état sauvage à celui de domestique : la voie commensale, la voie des proies, et la voie dirigée.

La voie commensale.
La voie commensale se base sur le fait qu'au travers de ses actions, l'homme modifie son environnement proche. Ces modifications ont pu attirer des populations animales, souhaitant profiter de ces modifications (déchets de nourriture par exemple). Les animaux les plus aptes à profiter de ces modifications sont alors les moins peureux, moins agressifs et avec la distance de fuite la plus courte. Cette anthropophilie, couplée avec l'habituation à l'homme, peut expliquer l'émergence de certaines espèces domestiques. C'est cette voie qu'auraient suivi les loups, bien avant la domestication des autres espèces domestiques (Larson *et al.*, 2012), ainsi que les chats, attirés par les rongeurs côtoyant eux-mêmes les hommes pour leurs réserves de grain (Willcox and Stordeur, 2012).

La voie des proies.
La voie des proies est portée par la volonté de l'homme, à un moment de son histoire, d'assurer la stabilité de certaines ressources. Ce faisant, l'homme a modifié ses stratégies de chasse et favorisé la survie des femelles pour la reproduction et ainsi, assurer la survie des populations de larges herbivores (Zeder, 2012).
Cette gestion des troupeaux d'animaux sauvages précède celle d'animaux captifs puis le contrôle du cycle de vie des animaux.

La voie dirigée.
La voie dirigée est la seule où l'homme entre dans une relation avec l'objectif délibéré de domestiquer une espèce (Zeder, 2012). En effet, avant de vouloir domestiquer ces espèces, l'homme avait besoin d'avoir déjà des espèces domestiques (passées par la voie commensale ou celle des proies), pour imaginer des versions domestiques d'espèces sauvages. Ces domestications arrivent en général hors des bassins de domestication initiaux et coïncident avec l'arrivée de l'agriculture (Larson and Fuller, 2014).

Qui ? Où ? Quand ? Comment ?

L'étude des animaux domestiques pose plusieurs questions auxquelles archéologues et généticiens tachent de répondre. Ces études ont permis l'identification de l'animal sauvage à l'origine des animaux domestiques. De l'aire de répartition de cet animal (ou en tout cas sont aire au moment de la domestication) ainsi que des précisions archéologiques et génétiques, il est possible de déduire la zone où a eu lieu la domestication. Les mêmes informations permettent de déduire quand cette domestication a eu lieu, et donnent les clés pour estimer quelle voie à été suivie (figure ci dessous).

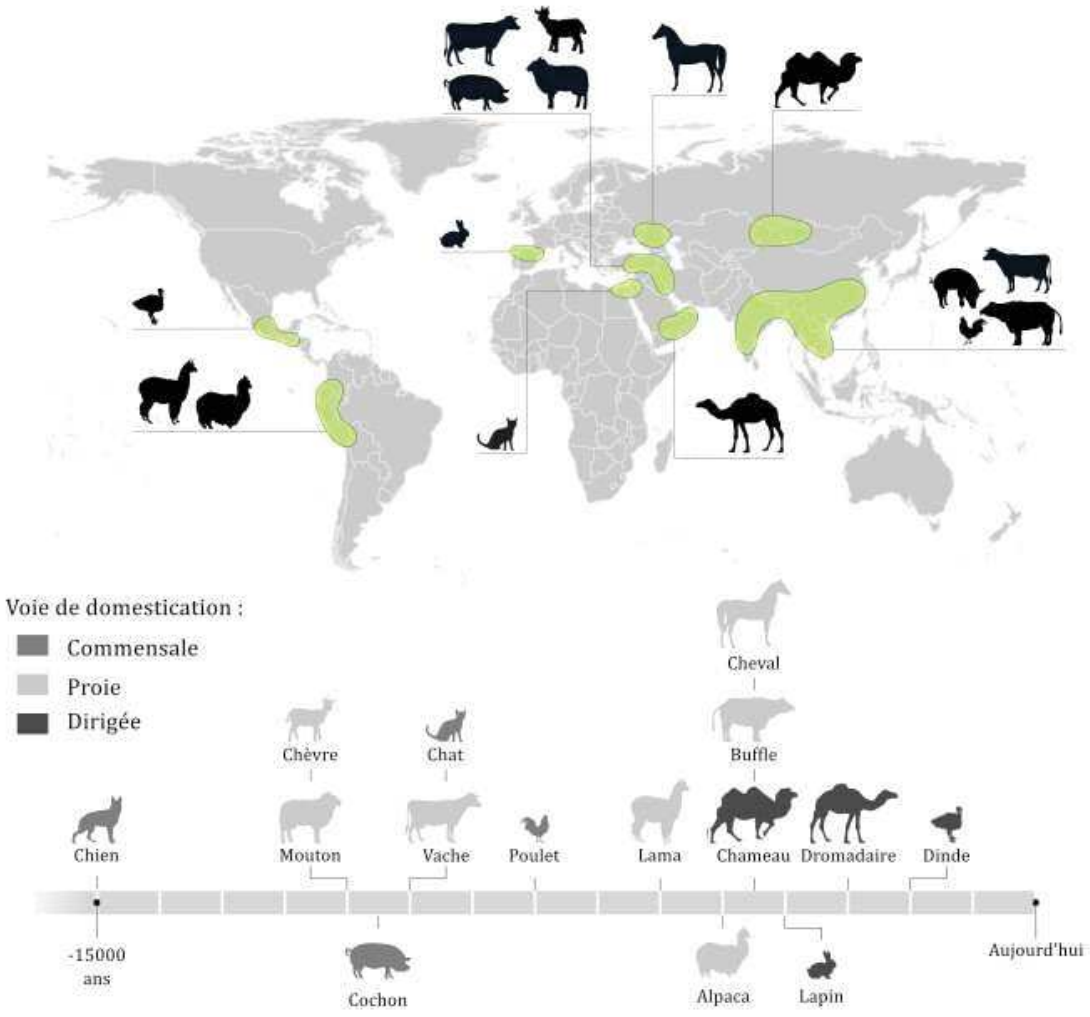


Figure 1 – Encadré 2: Lieux et dates de domestication des principaux animaux domestiques actuels. La carte représente les centres de domestication supposés des espèces dont la date de domestication est indiquée sur la frise. La couleur des animaux le long de la frise indique la voie de domestication supposée de ces espèces. Représentation inspirée de Larson and Fuller, 2014.

Problématique et plan de ce manuscrit

Dans l'aire de l'étude des données issues du reséquençage de génomes complets, de nombreuses études se basent sur les SNPs pour étudier les bases génétiques de la domestication. Cependant, en considérant l'impact des autres mutations, notamment des variants structuraux (SVs) sur les individus, il semble, en théorie, important de prendre ceux-ci en considération.
Afin de mieux mesurer l'importance de l'étude SVs au cours de l'évolution, les travaux présentés dans ce manuscrit essayent de faire émerger le rôle des SVs au cours de la domestication, notamment à travers l'exemple des petits ruminants.

Les travaux présentés ici se découpent en trois parties. La première vise à faire le point sur l'importance déjà connue des SVs chez les espèces domestiques. La seconde se concentre sur les SVs des petits ruminants décrits dans la bibliographie comme possiblement liés à la domestication. Enfin, la troisième partie se base sur une approche visant à détecter *de novo* des SVs et à les lier aux processus de domestication, d'amélioration et d'adaptation des petits ruminants.

Les petits ruminants : chèvres et moutons

Classification

La chèvre (*Capra hircus*) et le mouton (*Ovis aries*) sont des espèces domestiques de mammifères herbivores appartenant à la sous-famille des Caprinés, dans la grande famille des bovidae. La famille des bovidae (Mammalia, Ruminantia) apparue il y a environ 18,5 millions d'années (Vbra, 2000), est composée de 143 espèces actuelles (Wilson and Reeder, 2005).
Elles sont caractérisées par un estomac à quatre poches, adapté à la rumination, des sabots à deux doigts, deux cornes persistantes et creuses et leur denture est marquée par l'absence d'incisives sur le maxillaire et de canines.
Au sein des bovidae, la sous-famille des caprinae, apparue il y a 11 millions d'années, comprend 10 genres (*Ammotragus*, *Budorcas*, *Capra*, *Hemitragus*, *Naemorhedus*, *Oreamnos*, *Ovibos*, *Ovis*, *Pseudois*, *Rupicapra*) (Ropiquet and Hassanin, 2005), dont sont issus les chèvres (*Capra hircus*) et les moutons (*Ovis aries*).

De la domestication à nos jours

La chèvre (*Capra hircus*) est issue de la domestication de l'aegagre (*aegagrus*), il y a environ 10000 ans (Zeder and Hesse, 2000).
Les chèvres domestiques semblent provenir de deux centres de domestication différents, tout deux situés dans le croissant fertile. Le premier est situé au centre du plateau iranien et au sud des monts Zagros alors que le second correspond à une plus large aire, entre l'est anatolien, et le nord des monts Zagros (Figure 1) (Naderi *et al.*, 2008). Cette domestication semble avoir été précédée d'une étape de pré-domestication, consistant en une gestion durable des populations sauvages conduisant par exemple à la

protection contre les prédateurs ou la consommation des jeunes mâles. Cette étape préliminaire a conduit à une augmentation de la taille efficace de ces populations sauvages à partir desquelles certains animaux ont ensuite été domestiqués, à grande échelle et sans goulot d'étranglement (Naderi *et al.*, 2008).

Le taxon le plus proche du mouton domestique (*Ovis aries*) est le mouflon iranien (*Ovis orientalis*). La domestication de celui-ci a été localisée dans l'ouest de l'Anatolie et le nord des monts Zagros (Figure 1) (Rezaei, 2007). Cette domestication ne semble pas avoir été précédée par une phase de pré-domestication comme c'est le cas pour les chèvres, mais la diversité génétique capturée lors de la domestication tend a montrer un faible goulot d'étranglement lors de cet évènement, avec de nombreux individus impliqués (Rezaei, 2007).

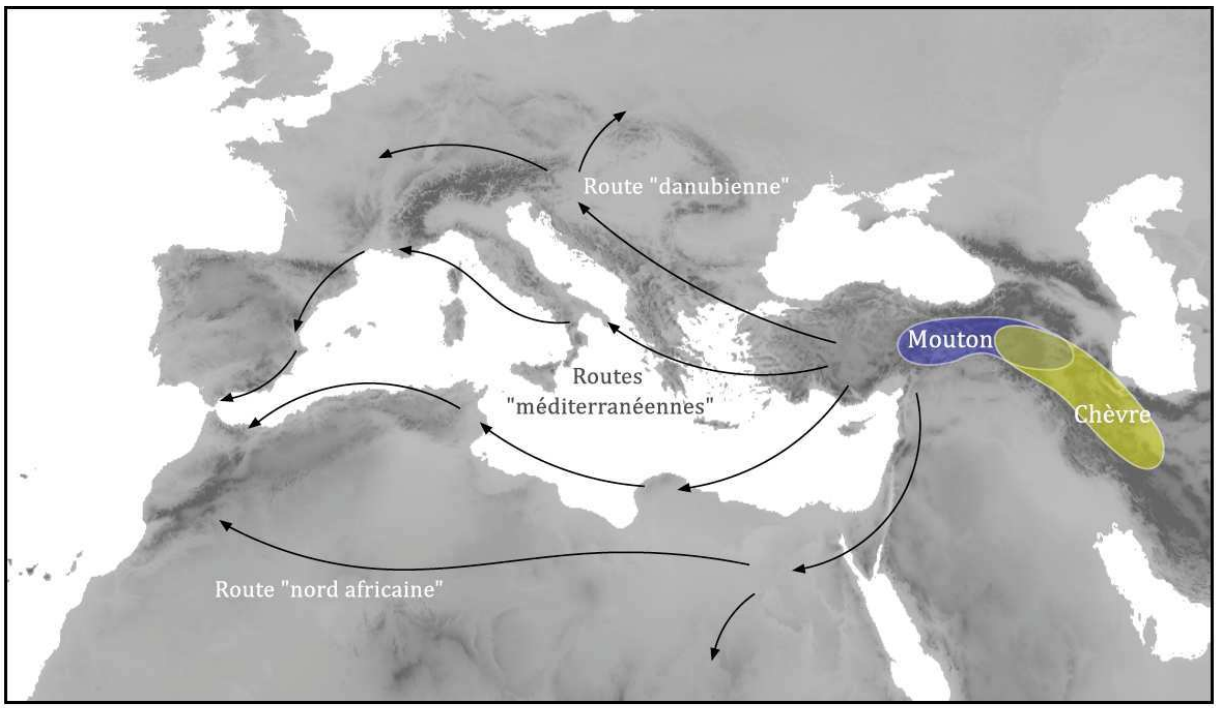


Figure 1 : Carte de l'aire de domestication et des voies de migrations Afrique et Europe inspirée de Zeder, 2008, Fernández *et al.*, 2006, Pereira *et al.*, 2009 et Muigai and Hanotte, 2013.

La dispersion des chèvres et des moutons depuis le croissant fertile vers le reste du monde est un processus complexe en plusieurs étapes. La migration vers l'Europe s'est faite durant les 4000 ans qui ont suivi la domestication, le long de deux voies principales, les voies danubienne et méditerranéenne. La voie danubienne trace la dispersion du « Neolithic package » (paquet néolithique) (poteries, animaux domestiques et sédentarisation), de la Grèce aux plaines d'Europe centrale et du nord en remontant le Danube. La seconde voie, dite méditerranéenne, suit les côtes méditerranéennes *via* des transport maritimes (Fernández *et al.*, 2006) (Bogucki, 1996).
En Afrique, la dispersion des chèvres et des moutons suit là aussi différentes routes. La voie méditerranéenne suit les côtes alors que d'autres routes, terrestres, ont permis la

dispersion des chèvres et des moutons sur l'ensemble du continent africain (Pereira *et al.*, 2009) (Muigai and Hanotte, 2013) (Figure 1).

Avec un total de 1,2 milliard de têtes de bétail de chèvres et 1,4 milliard de moutons en 2014 (http://www.fao.org/faostat/en/#data/QA) regroupés dans 665 races caprines et 1385 races ovines recensées (http://www.fao.org/documents/card/en/c/c40d538b-4765-445d-ba3c-c06eaaa49f4a), les chèvres et les moutons sont présents sur tous les continents et fournissent à l'homme de nombreux produits et services tels que du lait, de la viande, de la laine ou du cuir.

Chèvres et Moutons : de bons modèles

De part leur proximité phylogénétique (moins de 8.8 millions d'années pour l'ancêtre commun le plus proche (Ropiquet and Hassanin, 2005)), et leurs histoires de domestication, d'expansion, d'adaptation et de sélection (pour l'adaptation et la productivité) semblables, les petits ruminants sont de bons modèles pour l'étude des différents aspects de la domestication, du processus initial à l'adaptation à des climats variés, avec des objectifs de productivité divers. Ces deux évènements de domestication offrent donc deux réplicas d'une même expérience, et donc la possibilité de voir les convergences ou non et ainsi, d'approfondir nos connaissances sur les aspects évolutifs que représente l'ensemble du processus de domestication.

Contexte de recherche

Le projet NextGen

Le projet NextGen est né du constat d'érosion massive de la biodiversité au sein des animaux d'élevage. Financé dans le cadre du 7ème programme-cadre de la Commission Européenne, l'objectif du projet était 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.
Divisé en différentes composantes, il avait notamment pour objectifs :
(i) 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 petits ruminants.
(ii) L'étude des relations génome-environnement chez les chèvres et les moutons.
Pour cela, l'étude des mécanismes sous-jacents à l'adaptation locale des petits ruminants à leur environnement sur l'ensemble du gradient environnemental au Maroc.

Données : échantillonnage

Dans le cadre du premier objectif énoncé ci dessus, des échantillons de 30 *O. orientalis*, 30 *C. aegagrus*, 60 moutons (*O. aries*) et 60 chèvres (C. *hircus*) du centre de domestication, au nord ouest de l'Iran ont été collectés (Figure 2). Pour le second

objectif, un système de grille de cellules rectangulaires (0.5°x0.5°) couvrant la grande part du pays (~400.000 km2) (Figure 2) et représentatif du gradient des conditions climatiques et écologiques présentes au Maroc a été utilisé pour échantillonner au final 161 chèvres et 160 moutons. Pour l'ensemble des individus, les tissus ont été récoltés, et pour les domestiques les coordonnées spatiales et des données phénotypiques ont été collectées.

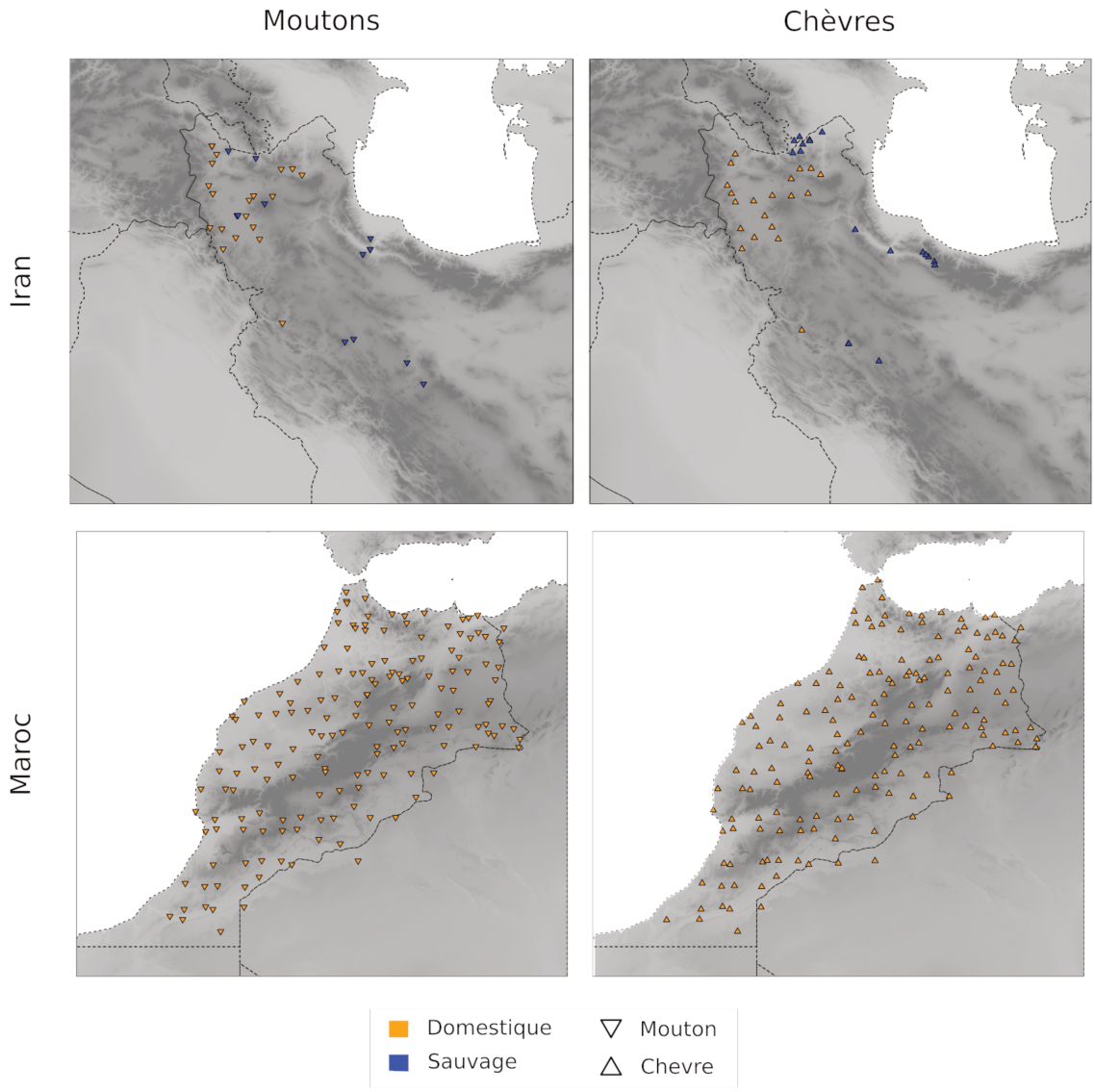


Figure 2 : Carte de l'échantillonnage des individus d séquencés en Iran et au Maroc

Données : Séquençage

Dans le cadre du projet NextGen, le génome complet de chaque individu a été séquencé via la technologies Illumina Highseq®. Pour chaque individu, après extraction de l'ADN des tissus, celui-ci a été fragmenté via sonication en fragments de taille homogène. C'est cette taille des fragments qui est appelée la taille de la librairie. Cette librairie de fragments d'ADN est ensuite séquencée à ses deux extrémités, produisant des lectures dont la taille est connue et fixée pour chaque expérience. Enfin, l'effort de séquençage a été tel que, en moyenne, la couverture du séquençage est de 12X, ce qui veut dire que chaque base de l'ADN de l'individu a été lue 12 fois en moyenne. Le séquençage d'un

individu peut donc être défini par différentes mesures que sont la taille de la librairie (en bp), la longueur des lectures (en bp) ainsi que la couverture (en X).
Les séquences de chaque individu sont ensuite alignées sur le génome de référence de l'espèce correspondante. La comparaison de ces alignements permet ensuite de détecter le polymorphisme pour chaque individu.

Projet NextGen : résultats basé sur les SNPs

Basé sur l'étude des SNPs, les données issues du projet Nextgen ont permis d'identifier certaines régions génomiques impliquées lors de la domestication des petits ruminants et l'adaptation à leurs environnements.

Concernant la domestication des petits ruminants, l'étude des SNPs a permis d'identifier 20 régions communes chez la chèvre et le mouton impactées par la domestication. Ces régions contiennent notamment des gènes associés avec le développement et le système nerveux central. Cette étude a aussi permis de montrer que pour certaines zones ciblées lors de la domestication dans une espèce, ces mêmes régions montrent un relâchement de la pression de sélection dans l'autre espèce. Ces résultats suggèrent donc que si certaines zones ont été des cibles communes lors de la domestication, d'autres non, illustrant le fait que différentes solutions ont été sélectionnées pour aboutir à des modifications phénotypiques similaires (Alberto *et al., In prep*). Pour ce qui est de l'adaptation locale des petits ruminants au Maroc, les travaux basés sur les SNPs montrent que si l'adaptation à l'altitude implique la respiration pour les deux espèces, l'adaptation au même habitat pour les deux espèces semble généralement impliquer des mécanismes différents (Benjelloun, 2015).

Problématique précise et axes de recherche

Problématique

Basés sur les données produites dans le cadre du projet NextGen, les travaux présentés dans la suite de ce manuscrit s'attachent donc à étudier l'impact des variants structuraux génomiques lors des processus de domestication et d'adaptation à leurs environnements des petits ruminants.
Ces travaux doivent donc nous permettre d'améliorer notre connaissance des variants structuraux génomiques chez les petits ruminants, ainsi que le rôle joué par ceux-ci lors de la domestication ainsi que l'adaptation des chèvres et des moutons à leur environnement.

Axes de recherche

Afin de répondre au mieux à cette problématique, les travaux présentés ici s'organisent en trois parties.
La première tente de faire une synthèse du rôle joué par les variants structuraux au travers d'exemples connus chez les animaux domestiques. Cette partie vise (i) à décrire

l'impact des SVs dans les génomes, ainsi que leurs conséquences pour les organismes, à (ii) étudier le rôle des SVs lors de la domestication, de la sélection et de l'adaptation des petits ruminant, ainsi qu'à (iii) montrer l'intérêt de l'étude des SVs chez les animaux domestiques.
Les deux parties suivantes se basent sur l'étude des données de reséquençage de génomes complets de chèvres et de moutons pour étudier l'impact des variants structuraux génomiques lors des processus de domestication et de sélection pour l'adaptation locale des petits ruminants.
La seconde partie se base sur les variants structuraux décrits dans la bibliographie, et leur recherche dans des données de reséquençage de génome complet. Cette stratégie doit nous permettre (i) de montrer qu'il est possible de retrouver ces variants dans des données de génome complet ainsi que de (ii) tester rapidement et efficacement des hypothèses sur des variants structuraux connus pour être liés aux processus de domestication et d'adaptation.

Enfin, la troisième et dernière partie se base sur une recherche sans *a priori* de l'ensemble des variants structuraux détectables dans les données de génome complet, puis tâche de lier ces variants aux processus de domestication et d'adaptation sur la base de leurs fréquences et répartitions dans les populations. Cette stratégie doit donc nous permettre de (i) détecter un grand nombre de variants structuraux, ainsi que (ii) d'étudier leur possible impact, en lien avec l'histoire de domestication et d'adaptation des petits ruminants.

Articles présentés dans ce manuscrit

Ce manuscrit prend la forme d'une thèse « sur articles », cependant tous ces articles ne sont pas égaux dans leur avancement.
Le premier est un état de l'art des SVs chez les animaux domestiques et plus de travail et de références seront nécessaires à son aboutissement.
Le second est un article près à être soumis à la revue *Molecular Biology and evolution*, en tant qu'article de recherche.
Le troisième est lui aussi près à être soumis au journal *Bioinformatics* sous la forme d'une Application Note.
Le quatrième et le cinquième quand à eux devraient être fusionné et complété par la discussion de manuscrit pour répondre plus largement à la question de l'apport, à large échelle, des variants structuraux.
Aucun de ces articles n'a, à l'heure actuelle, été soumis du fait de l'embargo posé par le papier décrivant les données et actuellement en préparation (Alberto *et al., in prep*).

PREMIÈRE PARTIE :
Variants structuraux et animaux domestiques

Introduction de la partie

Les variants structuraux (SVs) semblent impacter fortement la fitness des individus. Ce premier chapitre vise à faire un inventaire non exhaustif des SVs chez les animaux domestiques afin de mieux comprendre l'impact de ces variants sur les organismes. Cet inventaire doit aussi permettre d'entrevoir le rôle joué par ceux-ci durant la domestication. Enfin, cette étude ciblée sur les variants structuraux doit permettre de montrer l'intérêt de l'étude des SVs, notamment chez les animaux domestiques.

Article 1
Genomic Structural Variations and evolution: livestock as a study case

Introduction

Since the development of second-generation sequencing methods, the amount of genomic data produced grows continuously. Study focusing on polymorphism at whole genome scale producing those data generally focuses on single nucleotide polymorphisms (SNPs). Such markers, easy to detect are widely used to understand history and mechanism of domestication (de Simoni Gouveia *et al.*, 2014). Whole genome re-sequencing data also gives access to genomic structural variations, whose importance is possibly underestimated.

To date, multiple size-based definitions have been proposed for structural variations. On one hand, one considers SVs as any DNA sequence alteration other than a single nucleotide substitution (Feuk *et al.*, 2006). On the other hand, the traditionally used definition is more conservative and only consider as SVs genomic alteration larger than 1kb (Alkan, Coe, *et al.*, 2011). The size-based definition currently used is to consider as SVs a genomic rearrangements affecting more than 50 bp (Tattini *et al.*, 2015, Sudmant *et al.*, 2015). Considering that this 50bp threshold relies more on technical constraints than on biological considerations, the use of such size-based definition seems deprecated. Facing those limitations, structural variations could be defined based on the sequence alterations they include. Genomic structural variations can be classified whereby they are balanced or unbalanced, in other words, if they induce or not a variation in DNA quantity. Balanced structural variations are inversions and inter or intra chromosomal rearrangements, while unbalanced structural variations are insertions, deletions and Copy number variations (CNVs) (see box 1 for more information) (Tattini *et al.*, 2015).

Considering that SNPs account for only 0.1% of the human genome whereas SVs concern at least 1.5% of the genome (1.2% for indels and CNVs and 0.3% for inversions) (Pang *et al.*, 2010), the potential impact of SVs on the structure and function of genomes and their consequences at the individual, population or species level could be tremendous.

Since Darwin (Darwin, 1872), domestication is considered as the longest evolutionary experiment lead by humans (Megens and Groenen, 2012). This long-term experiment with strong selection constraints induced considerable differences between wild and domestics and between domestic breeds, to reach different productive, adaptive or ornamental objectives.

Given the potential importance of SVs during evolution, survey of non model species who have undergone strong selection pressure in their recent past, should allow us to study the role of structural variation when individuals are submitted to high selective pressures. Livestock provides thus a proper context to study how structural variations

may contribute to differentiation in an evolutionary context and how they may be important to face new selection pressures.

Focusing on livestock's genomic structural variations, the objectives of this review are to (i) understand the impact of SVs on genomes and their consequences at the organism's level, (ii) depict the role played by SVs during livestock domestication, improvement and adaptation and (iii) point the interest of studying SVs in domestics species.

SV's impact on organisms

To understand the impact of SVs on genomes and their consequences at the organism's level, the point is to focus on where SVs appends and what are the consequences. Thus, SVs may directly affect genes, impacting individual fitness directly. Individual fitness may also be impacted indirectly by SVs, through their indirect consequences on individuals.

SVs disturb genes in different ways

SVs may affect gene in multiple ways. Such gene alterations may come from a modification either of the gene product or the gene expression.

Protein modification. The fist type of alteration gathers SVs that affects the protein amino acid sequence. For example, a 57bp deletion within the ovine KAP6-1 gene modifies protein structure, with the loss of 19 amino acids in the protein, and is associated with an increase of wool fibre diameter (Zhou *et al.*, 2015).

Abnormal splicing. The second type of alteration comprises SVs that disturb gene function by disrupting mRNA splicing. For example the insertion of a transposable element at an intron-exon boundary in the *SILV* gene in dogs induces adherent splicing of the mRNA leading to the merle patterning phenotype (Clark *et al.*, 2006).

Other SVs may not affect gene product but may result in changes in expression levels.

Promoter alteration. Such deregulation may be caused by alteration of the promoter sequence. This is the case with an inversion near the KIT gene in horse, inducing an inappropriate gene expression, and the presence of the white spots of the Tobiano spotting pattern in horses (Brooks *et al.*, 2008).

Copy number variations may also deregulate gene expression. A 8kb CNV including *AMY2B* gene induce it's overexpression and this overexpression is correlated with dog's ability to digest starch (Arendt *et al.*, 2014).

All genes alterations cited above have an impact on individual phenotypes through modifying gene expression and gene product. The have *de facto* an impact on individual's fitness, and their distribution and spreading or not in populations are then ruled by selection, based on their positive or negative impact on host fitness.

"neutral SVs" affect fitness

Some SVs may not affect gene or it's expression. If they might seem to be neutral for individual carrying them (e.g. because they do not affect genes), most of them are not totally neutral in an evolutionary perspective.

This is the case, for instance, for repeated elements in genomes. Transposable elements, as any repeated sequence, play a role in evolution by their inherent ability to increase genomic instability, increasing chromosomal rearrangement rate. Such rearrangements may have deleterious effect on individuals or it's progeny. In example, the accumulation of transposable elements in the genome of Drosophila increases chromosomal rearrangements inducing early embryonic death, greatly decreasing individual fitness (Pasyukova *et al.*, 2004). On the other hand, rearrangements increase the adaptive potential of organisms. For example, rearrangements mediated by repeated sequences in the genome allow *phytoplasma* species' adaptation to diverse plant host and different insect vectors (Bai *et al.*, 2006).

Examples of SVs targets of human-mediated selection

SVs impact all traits selected by human

As mentioned above, SVs affect genes in a variety of ways, affecting genes directly (their expression level or their product) or indirectly (altering regulation), and such alterations are implied in the modification of a wide variety of traits selected by human during domestication, improvement and selection for local adaptation.

Productivity traits

All domestic species provide services to humans, and one of the top services is the supply of many resources, including food (milk, meat and other) and non-food products (wool, leather …). Many studies looked for the genetic bases of those differences between wilds and domestics, and highlight the role of SVs. In pigs, CNVs are known to impact genes involved in multiple traits, such as growth or meat quality (J. Jiang *et al.*, 2014). Similarly, a deletion may disrupt genetic expression or gene product, leading to strong phenotypic effects. In cattle, a 660-Kb deletion, including four different genes, is associated with strong positive effects on milk yield by QTL mapping (Kadri *et al.*, 2014).

Domestication and local adaptation

With domestication, domestics species spread out from their original geographical range. This human leaded movement of populations forced domesticated to adapt to new environments. In cattle and pigs, CNVs seems to be associated with breed-specific differences in environmental adaptation (Bickhart *et al.*, 2012, Paudel *et al.*, 2013). CNVs are also involved in major adaptive features related to domestication, such as diet changes. In dog the amylase activity correlates with AMY2B copy number, implying a better tolerance to starch rich diet induced by human feeding (Arendt *et al.*, 2014).

Reproduction and survival

The ability to reproduce and survive to changing environment is another important trait in which structural variations are implied. Therefore, intronic insertion in the pig's *SPEF2* gene increases fertility of sows and has spread in Finish population (Sironen *et al.*, 2012).

The increase of flock size induced by domestication might affect the transmission of pathogens or parasites and increase the effect of diseases on livestock populations. In

cattle, genes related to pathogens and parasites resistance, such as CATHL4 and ULBP17, are highly duplicated (Bickhart *et al.*, 2012). Genetic amplifications of disease resistance genes may be linked with dose-effect, where more copies of the genes mean more expression. This mechanism has already been described in resistance against nematode in cattle, where CNV of WC1 gene induce overexpression of the encoded eponym protein in the mesenteric lymph node of resistant animals (Hou, *et al.*, 2012).

Morphology and colour patterns

The most striking impacts of domestication are the easily viewable phenotypic differences between domestics and their wild relatives such as shape, coat colour or even size. Multiple SVs have been shown to be responsible of those traits. Coat colour is one of the most contrasting traits between wild and domestics that has been well documented in the context of domestication. In sheep and goats, the white coat colour is associated with two independent duplication including the ASIP gene and inducing it's over-expression (Norris and Whan, 2008, Fontanesi *et al.*, 2009). In cattle, two SVs, a duplication including the KIT gene and it's translocation on chromosome 29 are responsible for the white colour (Brenig *et al.*, 2013). Another structural variation, an inversion affecting the promoting sequence of the same gene, KIT, explains the Tobiano spotting pattern in horses (Brooks *et al.*, 2008). Finally, Merle patterning of the domestic dog is induced by a retrotranspon insertion in the SILV gene inducing aberrant splicing of the encoded mRNA (Clark *et al.*, 2006). This non-exhaustive list highlights a diversity of structural variations linked with a diversity of coat colours and patterns.

Another difference between many domestics and wild or between domestics is being horned or polledness. In sheep, an insertion in the 3'-UTR of the RXFP2 gene is responsible of polledness through the modification of the processing and translation of RXFP2 mRNA (Wiedemar and Drögemüller, 2015). In goats, polledness seems to be associated with a 11.7kb deletion in the same gene (Pailhoux *et al.*, 2001), even if it's causality is contested (Kijas *et al.*, 2013).

Behaviour

Differences between wild and domestics are also observable on animal behaviour. CNVs have been linked with nervous system functions, in particular nervous transmission, neuron motion and neurogenesis in Hanwoo and Holstein cattle. Those CNV may be associated with the changes in behaviour due to domestication (Shin *et al.*, 2014). In the same vein, gene annotations of pig's CNVs are enriched with genes related to sensory perception, neurological process and response to stimulus, suggesting their contribution to adaptation in the domestics and behavioural changes during domestication (Paudel *et al.*, 2013).

All these examples demonstrate that structural variations played an important role in the genetic response to domestication. They disrupt, increase or modify genes expression and products, inducing a wide range of phonotypical and behavioural modifications that may have been selected by humans during domestication, improvement and local adaptation.

A preponderant role of CNVs ?

It appears that all structural variations types are involved in modifications affecting traits selected by humans. While insertions, deletions, inversions and translocations are reported as affecting one or two of those traits, CNVs seems to be implied in all categories. However, we cannot conclude if this higher number of CNV is due to their higher impact or frequency, or if this observation is only due to our better ability to detect CNVs (Bickhart and Liu, 2014), with multiple technics available while other SVs remain harder to detect (Bickhart and Liu, 2014).

Human selection balance natural selection

Thus, SVs alter genes implied in interesting traits for human, and therefore were selected and spread among populations. However, human selection is sometimes too strong and contrary to natural selection and may lead to spread of deleterious mutations. In such cases, heterozygous individuals have better fitness than non-mutant individuals, but being homozygous mutants is lethal. This is the case with the 600Kbp deletion in the cattle, increasing milk production at heterozygous state but deleterious when homozygote (Kadri *et al.*, 2014). In pig, a LINE insertion in the SPEF2 gene increasing fertility of sows (Sironen *et al.*, 2012) but causes infertility due to « short tailed » sperm in males (Chen *et al.*, 2016), wiping individual fitness out. Such balanced selection explains the maintenance in population of deleterious alleles responsible of interesting traits at heterozygous state. Progeny or individual fitness is then balanced with selected trait.

Why study structural variations

SVs are good neutral markers

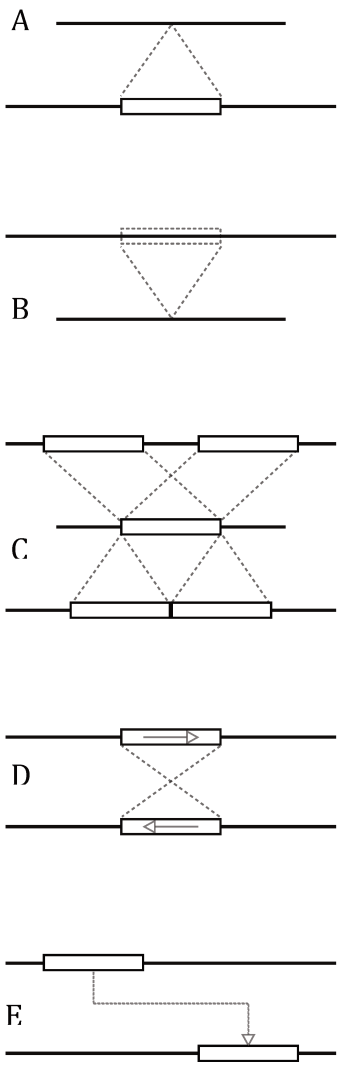
As already mentioned, impacts of SVs on individuals may be neutral or prejudicial. In the first case, if SVs don't disturb gene expression or it's regulation, SV distribution among populations is mainly ruled by drift and they may be used as neutral markers. Moreover, except for cut'n paste transposons insertions, SVs events are non-reversible. Taking this information into account allows going furtherer than with SNPs. For example, it's possible to date insertions events of LTR transposons based on the divergence between LTRs (Kijima and Innan, 2010). SVs may also be used to increase our knowledge of species / populations history. It's the case with the use of endogenous retrovirus JSRV in sheep genomes. Analysis of retro types of multiple European and Middle West population of sheep highlight two migratory episodes from domestication centre to rest of the world, providing insights into the history of sheep domestication (Chessa *et al.*, 2009).

SVs and selection; functional information

Considering non-neutral SVs, this survey highlight the fact that traits modification may be due to a wide range of mutations generally not studied during classical GWAS studies that are based only on SNPs, thus looking for SVs close to or under selective sweep may allow the identification of causal mutation. For instance, this is the case with the RXFP2 gene in sheep, associated to polled phenotype by SNPs-based GWAS (Johnston *et al.*, 2011, Kijas *et al.*, 2012), while the causative mutation seems be an insertion in the 3' UTR of the gene, probably inhibiting the proper processing and translation of the *RXFP2* gene (Wiedemar and Drögemüller, 2015). This phenomenon is also observed in dog, where a sweep is detected based on the SNPs near the AMY2B gene (Axelsson *et al.*, 2013), the CNV of the gene induce over expression and explain the better tolerance to starch rich diet of dogs than wolfs (Arendt *et al.*, 2014). In this context, SNPs are just passive witnesses of the causative changes in the DNA sequence, and studying SVs gives access to the functional information explaining the observed phenotypes.

Conclusion

As we have seen, SVs affect individual's fitness through genes modification, disrupting their expression or their product, but also indirectly, influencing genomic architecture. In the context of domestication, we have shown that all types of SVs are implied in the evolution of all traits selected by humans, illustrating their importance in improvement and local adaptation.
Thus, if the detection of SVs in WGS remains challenging (reviewed by Bickhart and Liu, 2014), structural variations have a strong impact on evolution and taking them in account seems mandatory to fully understand the genetic basis of evolution.



An **insertion** (A) is the addition of one or more nucleotide base pairs into a DNA sequence (novel or mobile-element transpositions).

A **deletion** (B) is a mutation in which a part of a chromosome or a sequence of DNA is lost, during DNA replication or after mobile-element transpositions.

CNVs (C) is a phenomenon in which sections of the genome are repeated and the number of repeats in the genome varies between individuals in populations.

An **inversion** (D) is a mutation where a segment of a chromosome is reversed end to end.

Inter or intra chromosomal rearrangements / Translocations (E) are chromosome abnormality caused by rearrangement of parts between homologous or non-homologous chromosomes.

Box 1: Definition of different genomic structural variations

SECONDE PARTIE :
Variants structuraux – Approche « variants candidats »

Partie 2 : Introduction

Dans le contexte de domestication, d'adaptation et de sélection pour différents traits d'intérêt, les variants structuraux génomiques jouent un rôle important (voir la première partie - Variants structuraux et animaux domestiques). Dans cette partie, l'objectif est de rechercher dans la bibliographie des variants structuraux connus pour être impliqués dans la domestication des chèvres et des moutons, puis de les retrouver dans des données de séquençage haut débit. Cette stratégie « variant candidat » doit nous permettre ensuite de lier ces variants à l'histoire de domestication ou d'adaptation des petits ruminants.

Partie 2 – Chapitre 1 : JSRV, enJSRV et les moutons

Contexte

Jaagsiekte sheep retrovirus (JSRV) est le beta-virus responsable de l'adénomatose pulmonaire ovine (OPA pour Ovine Pulmonary Adenocarcinoma), une tumeur chronique contagieuse des poumons des moutons. Présente dans de nombreux pays du monde, cette maladie représente un enjeu agroalimentaire important (Griffiths *et al.*, 2010).
JSRV fait partie des virus à ARN, aussi connus sous le nom de rétrovirus. Cette famille de virus se distingue notamment par la présence d'une transcriptase inverse qui rétro-transcrit leur génome d'ARN en ADN, qui est ensuite intégré dans le génome de la cellule hôte. Après l'intégration du génome rétroviral dans le génome de l'hôte, de nouvelles protéines virales sont synthétisées via la machinerie cellulaire de l'hôte ; protéines virales qui servent à la production de nouvelles particules virales qui elles-mêmes vont pouvoir contaminer de nouvelles cellules (Figure 1) (Gifford and Tristem, 2003).

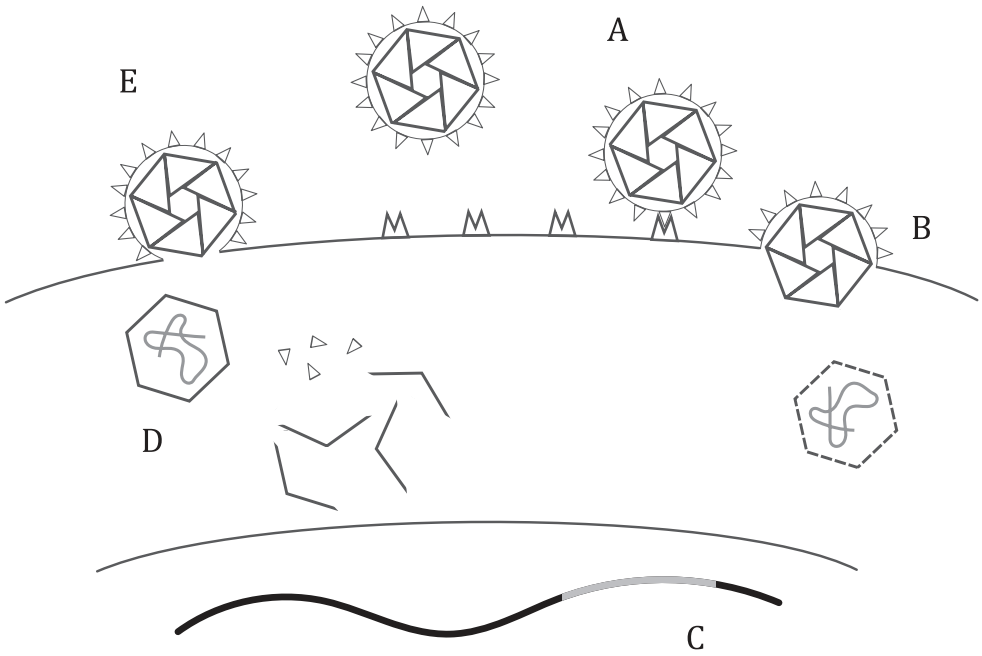


Figure 1 : cycle de vie des rétrovirus. A. Adhésion du virus à la cellule de l'hôte. B. endocytose. C. Reverse transcription de l'ARN viral en ADN et intégration dans le génome de la cellule. D. expression des gènes viraux et formation de nouvelles particules virales. E. relargage des nouvelles particules.

Le génome de JSRV fait environ 7,5 kilo-bases. Aux deux extrémités se trouvent les LTR (Long Terminal Repeat), qui encadrent les 4 gènes gag, pro, pol et env ainsi que orf-x, un ORF à la fonction inconnue (Figure 2) (Armezzani *et al.*, 2014).

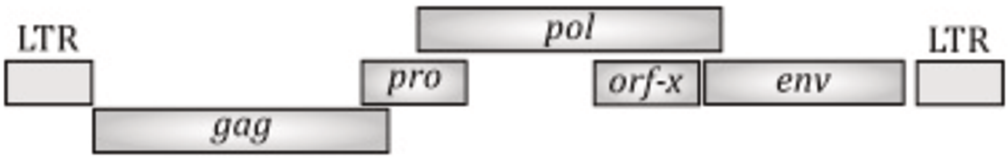


Figure 2 : organisation du génome de JSRV. gag code la polyprotéine structurale du virus, contenant les domaines de matrice (MA), capside(CA) et nucléocapside (NC). pro code pour la protease. pol code pour la transcriptase inverse, l'intégrase, une protéase, une endonucléase et une RNase H. env code pour des protéines d'enveloppe.

Si ce génome viral s'insère dans la lignée germinale de l'hôte, alors la transmission de ce génome viral se fait de façon verticale de l'hôte à son descendant et non plus de façon horizontale. Ce processus est appelé endogénisation et les séquences virales intégrées au génome de l'hôte sont dites endogènes (Figure 3) (Dupressoir *et al.*, 2012). Bien que le génome des mammifères soit constitué d'environ 8 à 10 % de séquences issues de rétrovirus endogène (Gifford and Tristem, 2003), le lien qui unit les rétrovirus et leurs hôtes reste encore mal connu.

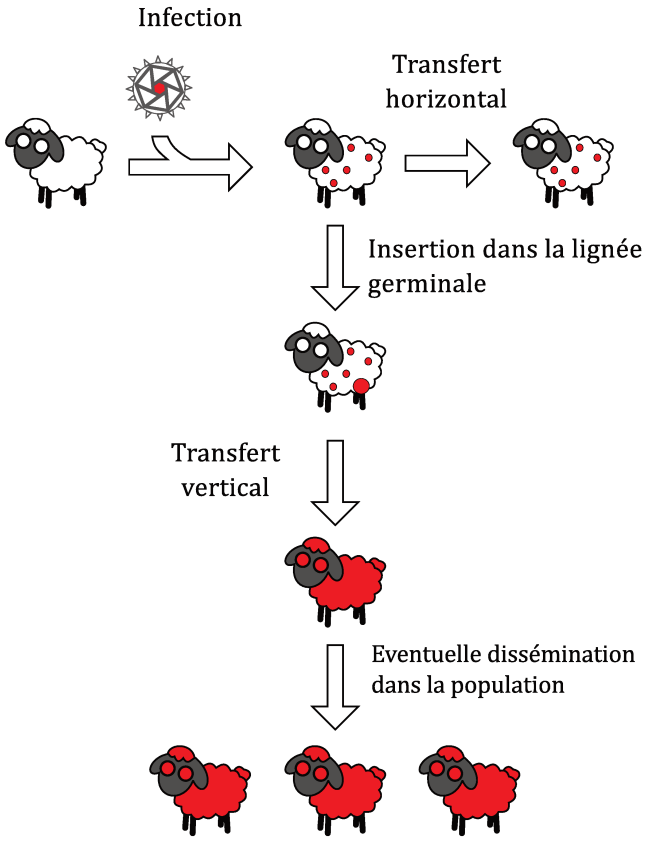


Figure 3 : Mécanisme de l'endogénisation.

Résumé de l'article

Le mouton, Jaagsiekte sheep retrovirus (JSRV), et ses copies endogènes (enJSRVs), forment un bon modèle pour étudier les relations à long terme entre les rétrovirus et leurs hôtes. En se basant sur l'étude de 76 génomes complets d'*Ovis* sauvages et domestiques, cette étude s'intéresse à cette relation à l'échelle évolutive. La méthode utilisée ici a permis de caractériser 462 enJSRVs quand seulement 27 étaient décrits précédemment. Si 14 semblent fixés au sein du genre *Ovis*, la plupart sont polymorphes et présents dans l'ensemble du genre *Ovis*, illustrant la longue coévolution entre JSRV et son hôte.
Cette étude porte ensuite un intérêt particulier sur deux copies endogènes, enJSRV20 et enJS56A1, insérée dans le locus q13 du chromosome 6. Ces copies, connues pour porter une mutation d'une arginine (R) vers un tryptophane (W) dans leur gêne gag, confèrent une résistance aux individus porteurs contre le virus exogène. Nous avons ensuite cherché le lien ente l'amplification de cette région décrite dans la bibliographie et la domestication du mouton.
Cette étude nous permet (i) de rejeter l'hypothèse d'une fixation de la mutation et d'une amplification de la région 6q13 en lien avec le processus de domestication comme aucune différence entre *Ovis* Orientalis et *Ovis* aries n'était détectable ; et (ii) de proposer un nouveau modèle pour l'histoire de la région 6q13. Dans ce modèle, le génome de JSRV a intégré le locus 6q13 après la spéciation entre *Ovis* et *Capra* et avant la radiation du genre *Ovis*. La mutation et l'amplification de la région ont eu lieu après la divergence entre les espèces « eurasiatiques » et « américaines ». Enfin, la fixation de cette mutation est due à un processus de sélection naturelle au sein de l'espèce *Ovis* orientalis avant la domestication.

Informations sur l'article

L'article 2 présenté ci-dessous a été écrit en vue d'être soumis au journal *Molecular biology and evolution.*
Les auteurs sont : T. CUMER, F. POMPANON, F. BOYER

Article 2

Old origin of a protective endogenous retrovirus (enJSRV) in the *Ovis* genus.

Abstract

Sheep, the Jaagsiekte sheep retrovirus (JSRV) and its endogenous forms (enJSRVs) are a good model to study long time relationships between retroviruses and their host. Taking advantage of 76 whole genome resequencing data of wild and domestic *Ovis*, we investigated this relation at evolutionary scale. An innovative use of re-sequencing data allowed characterising 462 differents insertions of enJSRVs, where only 27 were previously described. If only 14 insertions are fixed among genus, most of them are polymorphic and present all over the *Ovis* genus, highlighting long coevolution between JSRV and *Ovis* species.

We particularly focused on two well-known endogenous copies, enJSRV20 and enJS56A1 inserted in the q13 locus of chromosome 6 (6q13). Those two copies, known to have an arginine (R) to tryptophan (W) mutation in the *gag* gene, confer resistance against exogenous JSRV. Amplification of the 6q13 locus was described in the literature and postulated to be linked with domestication, we thus investigated, for each animal sample, the copy number of this locus to have an idea of the dynamic of copy number of the 6q13 locus at a large temporal scale

Our study allows to (i) deny previous hypothesis of fixation and amplification linked with domestication, as we were not able to detect differences between *Ovis orientalis, the actual wild relative of domesticated Ovis,* and *domestic sheep* (ii) and build a new model for the 6q13 locus history. In our model, JSRV inserted into the 6q13 locus after *Ovis-Capra* speciation and before the *Ovis* radiation. We postulate that, the protective mutation in the enJSRV 6q13 copy appeared shortly after its insertion and was followed by an amplification and that these events occurred after the divergence between "American" and "Eurasian" *Ovis* species. Lastly, fixation of the protective allele is due to natural selection before domestication in the *Ovis orientalis* species.

Introduction

Retroviruses are viruses that, during their life cycle, integrate their genome into their host's genome. This integration may happen either in somatic cells leading to an horizontal transmission of the viral DNA (exogenous retrovirus), or in the germline with then a vertical transmission of the DNA (endogenous retrovirus, ERV). Such endogenization is a quite common mechanism, since a large part of mammalian genomes is composed of retroviral sequences (8-10 %). These insertions play a major role in shaping the genome: they might alter gene structure, affect their regulation or increase genomic instability (Kaneko-Ishino and Ishino, 2012). Some genes of ERVs origins are also known to play major physiological roles, notably during mammalian

gestation, allowing fetal *in vivo* development (Lavialle *et al.*, 2013). They are transmitted across generations in a mendelian way, and their loss, spreading or fixation are ruled by evolutionary forces. Thus, neutral copies of viral DNA (e.g., no more functional due to mutations, or transcriptionally repressed by epigenetic marks…) are mainly driven by genetic drift, while the evolutionary dynamics of insertions impacting host fitness mainly depends on natural selection. Thus, studying neutral endogenous retroviruses copies allows to infer neutral demographic history of species (Sistiaga-Poveda and Jugo, 2014) or populations (Chessa *et al.*, 2009), whereas studying ERVs that alter the host fitness gives informations about adaptive histories of species.

The Jaagsiekte sheep retrovirus (JSRV) is a pathogenic agent responsible for the Ovine Pulmonary Adenocarcinoma, a transmissible lung cancer of sheep (Griffiths *et al.*, 2010). Several endogenous JSRV genome copies (enJSRV) have already been described in the genome of different species of the subfamily C*aprinae,* demonstrating an old interaction between JSRV and small ruminants (at least 5-11 million years ago (MYA)) (Hassanin *et al.*, 2012). The sheep, JSRV and enJSRV trio has long been used as a case study to address the complex relationships between retroviruses and their host, from functional and evolutionary perspectives. The sheep genome host several copies of enJSRV. Many of them remain uncharacterized, still, twenty-seven have been described up to now (Armezzani *et al.*, 2014; Sistiaga-Poveda and Jugo, 2014), among them, at least six are known to have an insertion polymorphism among sheep populations (Chessa *et al.*, 2009). EnJSRVs play a major role in sheep reproduction as the mRNA of the an *env* enJSRV gene (coding for the envelope protein in the exogenous virus, known to have a cell fusion inducing function) is required for trophectoderm formation during gestation and may play a role in immunosuppression responsive of materno-fetal tolerance (Dunlap *et al.*, 2006; Varela *et al.*, 2009).

Besides this, two loci, enJSRV-20 and enJS56A1, interfere with the exogenous JSRV during host contamination. Both endogenous copies have an arginine (R) to tryptophan (W) mutation in the *gag* gene, encoding for a transdominant protein that interfere with the exogenous protein during the late step of viral replication (Arnaud *et al.*, 2007). As a consequence, R/W mutant copies (R/W) have a protective effect against exogenous JSRV infection. These two copies are closely located at the q13 locus on chromosome 6 (6q13) (Armezzani *et al.*, 2011).
The history of the 6q13 locus seems to be closely related to the evolution of the *Ovis* genus and two hypotheses emerge form previous works. According to the first hypothesis the JSRV insertion at 6q13 would have occurred twice (Frederick Arnaud *et al.*, 2007), after the divergence between *Ovis* and *Capra (i.e.,* 5-11 MYA, (Hassanin *et al.*, 2012)), but before the divergence between the 2 *Ovis* lineages 2.4-5 MYA, i.e. the "American" (*O. nivicola*, *O. dali* and *O. canadensis*) and "Eurasian" (*O. ammon*, *O.vignei*, *O. orientalis* and *O. aries*) lineages (Rezaei *et al.*, 2010). Under this hypothesis, the R/W mutation would have appeared in both enJSRVs copies (enJSRV-20 and enJS56A1) either independently or by gene conversion. The high sequence similarity between enJSRV-20 and enJS56A1, favours the second hypothesis that enJSRV-20 is indeed the result of a recombination between enJS56A1 and another enJSRV after enJS56A1 R/W *gag* mutation (Armezzani *et al.*, 2014).

While enJS56A1 is described in all *Ovis* species, the protective allele is exclusive of the Eurasian species and is described to be fixed only in domestic's animals (Arnaud et al. 2007). The presence of enJSRV-20 copy across taxa is less clear. The insertion of enJSRV-20 is described to be fixed in *O. aries* and *O. orientalis*, polymorph in *O. ammon* and *O. cannadensis* and absent in *O. dali* and *O. nivicola* (no data available for *O. vignei*). Moreover, a protective allele exist in *O. aries*, *O. orientalis* and *O. ammon*, and is fixed in *O.aries* and *O.orientalis* (Armezzani *et al.*, 2011; Frederick Arnaud *et al.*, 2007).

The fixation of the protective alleles of both enJSRV-20 and enJS56A1 copies in domestic animals, has been hypothesised to be related to the domestication process, maybe due to management in herds and promiscuity between animals (Frederick Arnaud *et al.*, 2007). Moreover, the 6q13 region carrying the protective enJSRV allele is known to be duplicated several times, especially in domestic sheep (Armezzani *et al.*, 2011). Because of the selective advantage of the R/W enJSRV allele, this increased number of copies could lead to an enhanced resistance due to a dose-effect relationship. However, there is still no clear demonstration of the selective role of domestication on the fixation of the protective allele and on its duplication. In this context, our study aimed at testing this hypothesis by assessing the link between domestication and the presence of protective alleles, taking advantage of the whole genome sequencing of an unprecedented sampling of wild and domestic *Ovis*.

Results

global enJSRV survey.

Among the 76 individuals from 5 *Ovis* species (including 56 *O. aries*), a total of 462 JSRV insertions sites were detected in at least two individuals (Figure 1) distributed over all chromosomes (supplementary Figure 1, supplementary table 1), with a median number of 90 insertions per individuals (ranging from 69 to 109). Only 14 on the 462 insertions were found in all individuals of all species. There was a clear specific pattern of enJSRVs presence-absence (Figure 1) with a significantly higher number of ERVs insertions in *O. aries* genomes than in *O. orientalis* genomes (medians of 93.5 and 85 respectively, t. test p-value: 0.048).

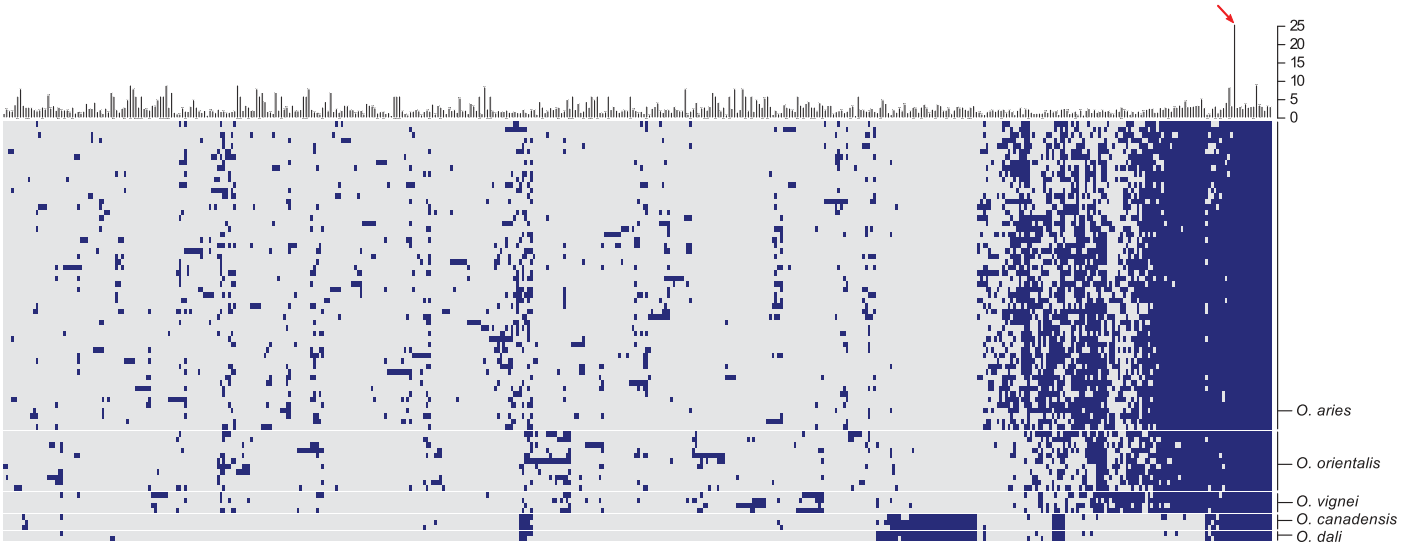


Figure 1: enJSRV presence (blue) absence (grey) heatmap. Individuals (lines) are ordered in species and enJSRV (columns) are ordered based on their relative binary distance. Only enJSRVs present in at least two individuals were kept. Layered barplot represent quantile 95 of coverage in insertion window. Red arrow is for 6q13 insertion site.

A BLASTn (Altschul *et al.*, 1990) search for the sequences flanking the 5' and 3' ends of the enJS56A1 insertions allowed positioning the 6q13 region on the sheep and goat reference genomes. In goat, the insertion site was present three times clustered on chromosome 6 (best hit between 6129894 and 6130613 bp) with no evidence for enJSRV insertion in all three duplications. In sheep, the 5' and 3' flanking sequences were present once in the contig *UnplacedScaffold_004085138.1*, between 10443 and 10819, and 1 and 342 for the 3' and 5' flanking sequences, respectively. Interestingly, this insertion site was confirmed in sheep, as it was detected in all studied individuals, with a coverage higher than for all other insertion sites (with a quantile 95 of coverage 25.56 when quantile 95 median value is 2.3 for all sites (Figure 1)).

Due to the library insert size that is too short to have direct evidence of the genome localization of gag coding sequences (see supplementary Figure S3), paired-end data showed a strong correlation between the normalized number of inserted copies at the 6q13 locus and the normalized number of putative enJSRV *gag* copies carrying the R/W mutation at the whole genome scale (Figure 2). This supports the hypothesis that enJSRV copies carrying the R/W mutation are inserted within the 6q13 locus, even if we cannot exclude the occurrence of this mutation elsewhere in another enJSRV copy. Moreover, mate-pair data obtained for one sheep showed five paired reads with one read mapped on the sequence flanking the 3' or 5' side of the insertion and the other carrying the protective mutation. This confirmed the presence of the protective mutation at the 6q13 locus (Supplementary table 2).

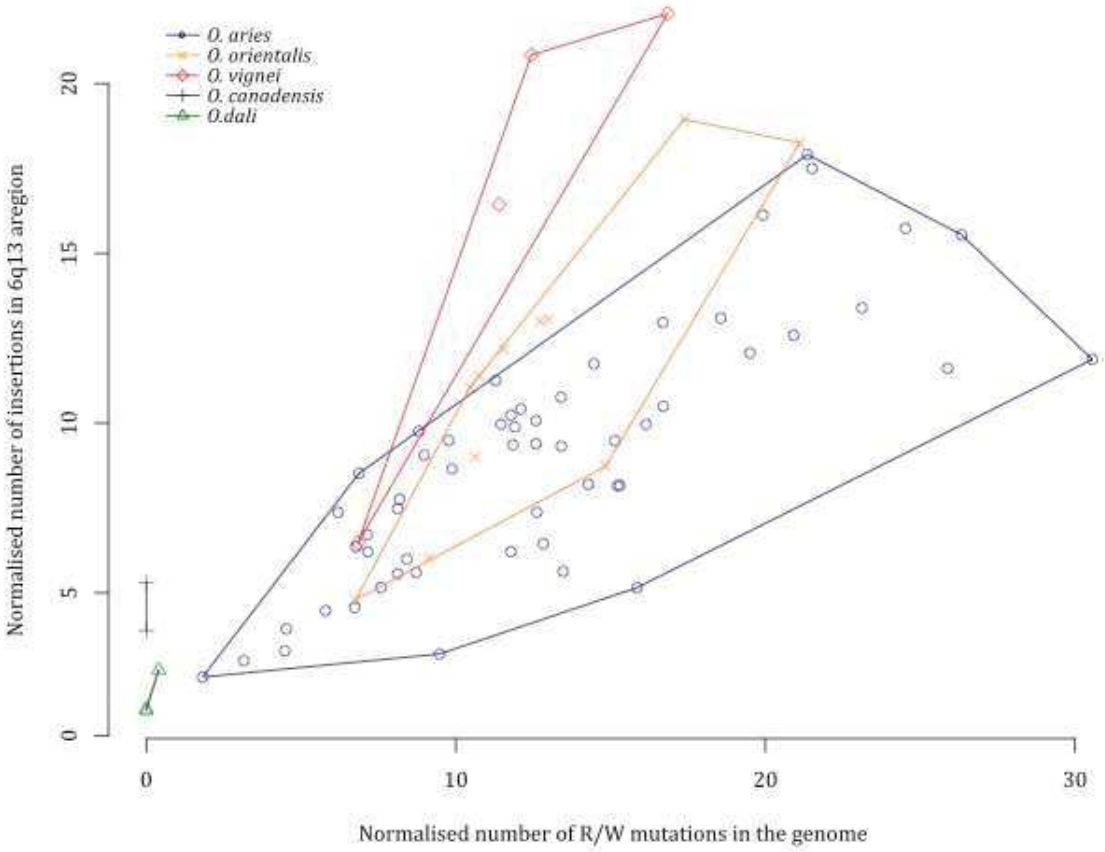


Figure 2: relationship between the number of enJSRV in 6q13 region and the number of enJSRV W mutants in the genome.

Within the 6q13 locus, the average number of enJSRV copies widely varied, from 1 to 22 among species. North American species (*O. Canadensis* and *O. dali*) hosted a low number of wild type (R) copies (ranging between 1 and 4), while domestics sheep and Eurasian wild species (*O. aries*, *O. orientalis* and *O. vignei*) all had a highly variable number of protective (W) copies (between 2 and 30) (Figure 2). There were no significant difference between domestic sheep (*O. aries*) and its closest wild relative (*O. Orientalis*) for neither the number of insertions nor the number of protective copies. Differently, for *O. vignei*, the number of R/W carrying insertions was lower than the number of inserted copies.

Evolutionary history of the 6q13 locus

The mean ratio between the number of protective copies and the number of 6q13-inserted copies highlights differences between species. As paired-end data did not allowed differentiating enJSRV-20 and enJS56A1, we denote here both types of copies as enJSRV-6q13. For *Ovis vignei*, this ratio is 0.78, indicating a higher number of enJSRV-6q13 copies than protective mutations. For *O. orientalis* and *O. aries*, ratio is respectively 1.15 and 1.45, indicating a higher number of protective mutation than the number of enJSRV inserted within 6q13 region.

Interestingly, at the 6q13 locus the number of insertion sites without enJSRV was approximately twice the number of enJSRV in all *Ovis* genomes (Supplementary Figure

2). Despite the low number of American *Ovis* individuals, we observed a clear difference between American and Eurasian species (Figure 3). American species (*O. dali* and *O. canadensis*) had a lower number of enJSRV-6q13 copies (harbouring or not the protective mutation) and a limited number of insertion sites without enJSRV, while Eurasian species (*O. vignei*, *O. orientalis* and *O. aries*) hosted a high number of mutated copies with an equivalent number of enJSRV inserted at 6q13 and twice more insertion sites without enJSRV.

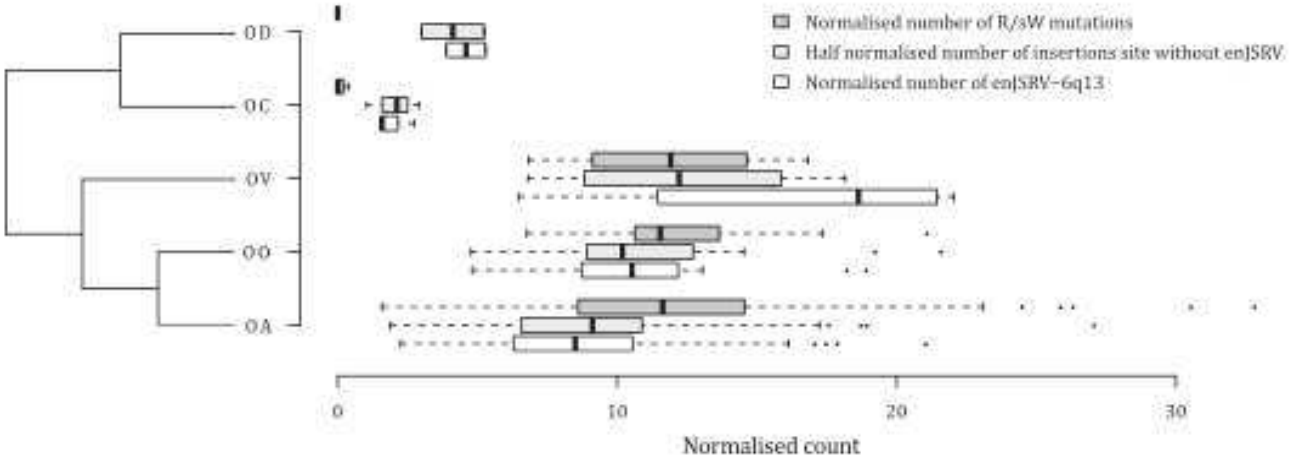


Figure 3: 6q13 region story. Phylogenetic relationship between Ovis species (Rezaei *et al.*, 2010). (b) Barplot of the number of enJSRV-6q13, the number insertion site without enJSRV divided by two and the number of W21 mutation, in function of species (OD : *Ovis dali*, OC : *Ovis canadensis*, OV : *Ovis vignei*, OO : *Ovis orientalis*, OA : *Ovis aries*).

Discussion:

Methodological issues:

The whole genome survey method described here is a reliable and simple strategy to detect mobile element insertions sites using paired-end data, inspired from the method developed by Keane et al., 2013. Nevertheless, this strategy encountered some limits, as the fragments in the library and the paired-end reads were too short to reconstruct the whole enJSRVs sequences. This was not the goal of our study but would be mandatory to further characterize the impact of each insertion, or to date insertions by comparing mutations betweens 5' and 3' LTR sequences. With such goals, producing longer reads, as available now with long read sequencing, would be necessary. However, information from our whole genome survey could be useful as a basis for future works. Moreover, our method increased the number of enJSRVs described from 27 to 462 (Arnaud et al. 2007). Our approach also allowed overcoming other limitations of the methods previously used. While the FISH method, traditionally used to detect and locate enJSRVs, only gives a rough insertion positioning at chromosomal scale (Carlson *et al.*, 2003; Armezzani *et al.*, 2011), whole genome surveys gave access to insertion site sequence with a much better resolution (100 bp compared to 10kbp). Furthermore, PCR based

detections were successfully used to detect insertion polymorphism (Sistiaga-Poveda and Jugo, 2014) but it is known that PCR dependence to hybridization site conservation among species may generate false polymorphism (e.g., false or null alleles (Pompanon *et al.*, 2005)). In our approach, the length of reads, the size of the fragments in the library and sequencing depth increased the alignment probability in insertion sites and consequently the detection power.

Endogenisation of JSRVs:
Previous works already described a long interaction between the subfamily *Caprinae* and JSRV related retroviruses, with enJSRVs shared by *Ovis* and *Capra* (Frederick Arnaud *et al.*, 2007; Sistiaga-Poveda and Jugo, 2014). Among the 462 enJSRVs described in this work, 32 were present in all *Ovis* species among which 14 were present in all individuals. This confirms that the first endogenisations of JSRV occurs before the diversification of the *Ovis* genus more than 2.4 MY ago (Rezaei *et al.*, 2010). The occurrence of differential insertions among species and populations highlight the fact that the endogenisation of JSRV accompanied the evolution of all *Ovis* species and is still in progress. The higher number (Supplementary Figure 3) of enJSRVs in domestic sheep compared to wild species may reflect a release of purifying selection and/or an increased exposure to exogenous JSRV with domestication related to management in herds and proximity between animals.

History of the 6q13 region:
The 6q13 region and its enJSRVs copies (i.e., enJSRV-20 and enJS56A1) played a special role in the evolution of the JSRV and sheep relationship, with the occurrence of a protective mutation. This mutation is known to encode for a transdominant protein that interfere with the exogenous protein during the late replication step, giving genetic resistance against exogenous JSRV (Armezzani *et al.*, 2014).

The analysis of the goat reference genome showed no evidence for provirus insertion in the 6q13 region, as the 5' and 3' flanking sequences of enJS56A1 are contiguous. Moreover, these insertion sites are part of a 15kb region repeated in tandem at least three times. In the sheep reference genome, these 5' and 3' flanking sequences are located on an unplaced contig *UnplacedScaffold_004085138.1*.

We also found that the 6q13 region hosted an endogenous JSRV in all *Ovis* individuals. This confirmed that the insertion of JSRV in the 6q13 region occurred after the *Ovis-Capra* divergence (5-11 MYA) but before the diversification of the *Ovis* genus (2.4 MYA) (Rezaei *et al.*, 2010). We detected the arginine to tryptophan (R/W) mutation previously reported in enJSRV-6q13, which gives protection against the exogenous JSRV (Armezzani *et al.*, 2014), in all individuals from the Eurasian species. It was also detected in an *O. canadensis*, but as it is supported only by a single read, it can correspond to a sequencing error. However, we cannot exclude the hypothesis of a rare tryptophan mutant in American mouflon, given the low number of individuals tested. In both case, the wide spread of protective allele in all Eurasian species and its possible presence in American species dates the 6q13 mutation during or a short time after the separation

between those two lineages (Rezaei et al. 2010). Studying more genomes from *Ovis ammon* and the American species would allow dating this mutation more precisely.

Interestingly, we also detected a genomic amplification within the 6q13 region. Moreover, the number of enJSRV-6q13 was correlated with the number of protective mutations. The most parsimonious scenario explaining those observations is that the R/W mutation occurred in the common ancestor of all Eurasian species, followed by the genomic amplification, between 1.26 and 2.42 MYA (Rezaei *et al.*, 2010).

The frequency of the protective allele was similar between domestic individuals (*O. aries*) and their closest wild relative (*O. orientalis*), but was lower in *Ovis vignei* than in the other Eurasian species (Figure 2).
This highlight the fact that not all enJSRV-6q13 in *O. vignei* carry protective mutation while all enJSRV-6q13 in *O. orientalis* and *O. aries* does, indicating a fixation after the *O. vignei* speciation, 1.26 MYA. Besides, the ratio between the number of protective mutations and the number of enJSRV-6q13, is higher than 1 in *O. orientalis* and *O. aries*. This suggests that there are more protective copies than enJSRV-6q13. This could be explained by considering that enJSRV-20 (whose exact sequence remains uncharacterised) would be a recombination of two enJS56A1 (Armezzani *et al.*, 2014), allowing two protective mutations per enJSRV-20 insertion. The resulting hypothesis is that enJSRV-20 acquired the protective mutation and spread after the speciation of *O. vignei* (1.26 MYA) but before the domestication of *O.aries* (10 KYA). A strong natural selection for an increase of the number of protective copies after the separation between *O. vignei* and the other Eurasian species would induce both the fixation of the protective allele and the spread of enJSRV-20 in the *O. orientalis* lineage.

With respect to what can be observed on the basis of the goat genome, we were also able to assess the presence of several insertions sites without enJSRV in the *Ovis* whole genome sequences. Interestingly, the normalized number of enJSRV inserted in 6q13 region and the normalized number of insertion site without enJSRV are highly correlated (Supplementary Figure 2). Moreover, we observed about twice insertion sites without enJSRV than enJSRV-6q13 inserted. These observations support a scenario involving the triplication of this region before the *Ovis*-*Capra* divergence (5-11 MYA), followed by an insertion of a JSRV provirus and then an amplification of this region in the *Ovis* lineage.

Previous work suggested that the protective allele (carrying the R/W mutation) was fixed in domestic animals in relation with the domestication process (Frederick Arnaud *et al.*, 2007), and that the genomic amplification in 6q13 region was specific to domestic sheep (Armezzani *et al.*, 2011). Our analyses showed that domestics and their closest wild relatives (*Ovis aries* and *Ovis orientalis*) host a similar number of protective enJSRV copies (R/W mutant) that this number greatly varies among individuals. No distinction between wild and domestics animals was observable, challenging the hypothesis of the impact of domestication on this locus. The polymorphism observed in domestics would more likely reflect the polymorphism present in the wild group, which was captured during domestication about 10kYA.

Bovidae and JSRV are engaged in a long-term relationship. As demonstrated by previous work (Armezzani *et al.*, 2014) and by old and shared enJSRVs among *Ovis*, this relationship was engaged before the *Caprinae* speciation (5-11 MYA). This relation took a new turn when the protective mutation occurred on the 6q13-inserted enJSRVs. Natural selection may account for the widespread amplification of this locus in all Eurasian species. Genomic amplification is well known as being a mechanism involved in resistance and adaptation, for example in resistance against nematode in cattle (Hou, Liu, *et al.*, 2012) or chemical insecticides in mosquitoes (Faucon *et al.*, 2015). If resistance against JSRV infection is linked with a dose-effect of enJSRV-6q13 *env* expression (Viginier *et al.*, 2012), then an increase in the number of copies could increase individual fitness and be selected. The recent discovery of enJSRV-26 able to escape enJSRV-6q13 resistance mechanism in the Texel breed marks a significant step forward in JSRV and sheep interaction (Armezzani *et al.*, 2011) and may be the first step of "arm race" between domestic sheep and JSRV.

Material and Methods:

Data sources:

We used as references the genome assemblies of sheep (build Oar_v4.0 - GenBank assembly accession: GCA_000298735.2), goat (build Chir_2.0 - GenBank assembly accession: GCA_000317765.2) and JSRV (RefSeq assembly accession: GCF_000850005.1), as well as the 3' and 5' flanking sequences of enJS56A1. (Arnaud, personal communication). The genomes of 76 individuals from five *Ovis* species (*O. dali*, *O. Canadensis*, *O. vignei*, *O. orientalis* and *O. aries*) and several *Ovis aries* populations and breeds were retrived from the ENA archive (accession numbers in Supplementary table 3) (Table 1). We also analyzed Mate-paire whole genome sequences for one *Ovis orientalis* sample (Accession : PRJEB3141). For a complete description of the data, refer to (Alberto *et al*, *in prep*).

Table 1: summary of individual sequences used in this study, clustered by species and populations. Data are available at: http://projects.ensembl.org/nextgen.

Species	Population	Nb of individuals	Mean coverage (sd)	Insert Size (sd)	Accession
O. aries	Iran	19	11.7 (1.1)	330 (18)	PRJEB3138
	Maroc	20	12.7 (1.0)	318 (11)	PRJEB3137
	World panel	17	12.3 (2.0)	179 (11)	PRJNA160933
O. orientalis		11	13.3 (3.1)	312 (40)	PRJEB3139
O. vignei		4	14.6 (2.4)	318 (14)	PRJEB5463
O. canadensis		3	11.8 (0.8)	170 (5)	PRJNA160933
O. dali		2	11.6 (0.1)	186 (2)	PRJNA160933

Scan for enJSRV insertions:

Reads from all animals were pooled and aligned on the JSRV genome with BWA mem algorithm (defaults parameters) (Li and Durbin, 2009). Unmapped reads whose mate aligned with high mapping quality (60) on the JSRV genome were aligned on the sheep reference genome, OAR-v4.0. Regions of the sheep genome where reads aligned with

high mapping quality (60) and a depth of coverage exceeding 5X were then considered as insertion sites. To avoid detecting multiple times the same enJSRV insertion in the reference genome, regions that are distant by less than 10kb were considered as the same insertion. Based on these regions and the aligned reads, a presence/absence table was built with information for the 76 individuals. Only sites detected independently in at least two individuals were kept for the further analysis.

Validation of the enJSRV insertion scan procedure :
To validate the enJSRV insertion scan procedure, we tested our ability to detect previously described polymorphic enJSRVs insertion site (Chessa *et al.*, 2009). *In silico* amplifications were performed using *isPCR* (Kuhn *et al.*, 2012) (two mismatches authorized, 10kb Max product size, primer sequences in supplementary table 4) to identify insertion sites in the OAR-v3.1 reference genome. Corresponding sequences in OAR-v4.0 were obtained by alignment (BLASTn search, defaults parameters) (Camacho *et al.*, 2009). Insertion sites for seven polymorphic enJSRVs (enJSRV-6, -7, -8, -15, -16, -18 and enJS5F16) were explored in OAR-v3.1. Only five of them could be localised in OAR-v3.1 and had precise equivalent on OAR-V4.0 (enJSRV-6, -8, -16, -18 and enJS5F16 (supplementary table 4)).
Our whole genome survey detected four of these five enJSRVs (enJSRV-6, -16, -18 and enJS5F16). The non detection of EnJSRV-8, a rare locus described in few populations from Northern Europe (Chessa *et al.*, 2009), was consistent with our sampling that did not include individuals from this geographic area.

6q13 locus identification:
Using the 5' and 3' flanking enJSRV-20 sequences (respectively 372 and 342 bp long, F. Arnaud, personal communication) as queries, BLASTn alignments (using defaults parameters) were performed against the sheep and goat reference genomes, in order to identify the 6q13 coordinates (OAR-v4.0 and Chir_2.0, respectively).

Copy number estimation:
The library (Table 1) made for paired-end sequencing had too short fragments to allow mate-pairs to anchor both on the 6q13 region and on the enJSRV mutation site (see Supplementary Fig XX). The mean fragment size of the libraries ranged from 170 to 330 bp, while the R/W mutation is at 642 bp from the 5'end of JSRV sequence. Thus it was not possible to obtain direct evidence that a copy inserted in 6q13 region carried the protective mutation. We implemented an indirect strategy relying on the estimation for each individual genome of: (i) the number of inserted enJSRV copies at the 6q13 locus and (ii) the number of enJSRV copies carrying the R to W mutation. The insertion site at the 6q13 region sequence was reconstructed by merging the enJSRV-20 5' and 3' flanking sequences and the JSRV sequence. Reads from all individuals were aligned independently on this reconstructed sequence using the BWA mem algorithm (defaults parameters) (Li and Durbin, 2009). Reads with a mapping quality lower than 60 were discarded.
(i) The number of inserted copies at the 6q13 locus (i.e., #enJSRV6q13) was estimated by counting read pairs with one read aligned on the 5' or 3' sequence and its mate aligned on the JSRV sequence. To allow comparison between individuals, and because of

possible bias due to library construction/coverage, this estimate was normalized. One hundred contiguous pairs of 350 bp windows (the size of 5' and 3' flanking sequences) were randomly selected in the genome. For each windows pair, read pairs with one read correctly aligned (quality equal to 60) in each window were counted. The normalization consisted in dividing #enJSRV6q13 by the median number of reads correctly mapped to the 100 randomly selected windows.
(ii)The number of copies carrying the R/W mutation was estimated by counting the number of reads carrying the protective mutation (reads aligning perfectly to the JSRV genome but with a A at position 642 on the JSRV reference genome) for each individual divided by the mean whole genome coverage of the individual on the OAR-v4.0 genome.
Both estimations were then compared to evaluate if the number of copies inserted in the 6q13 region was correlated to the number of copies carrying the mutation.

The number of contiguous 5' end 3' end sequences in the 6q13 region (thus showing no enJSRV insertion) was estimated by the number of pairs with one read aligned on the 5' end of the sequence and the mate aligned on 3' end corrected with the same procedure as (i).
Analyses, statistical tests and graphs were made using the R language (Ihaka and Gentleman, 1996).

Acknowledgments

We would like to thank Frederic Arnaud for providing us enJSRV-20 flanking sequences and useful suggestions.
Sequences used in this work are available at: http://projects.ensembl.org/nextgen.
This work was supported by the European Union 7th framework project NEXTGEN (Grant Agreement n°244356)
LECA is part of the Labex OSUG@2020 (ANR 10LABX56)

JSRV, enJSRV et les moutons.

Plusieurs conclusions semblent intéressantes à retenir de cette étude des JSRV endogènes au sein du genre *Ovis*.

La première se situe sur le plan méthodologique. En effet, la stratégie décrite ici est une stratégie simple permettant de rapidement détecter un grand nombre de copies d'un élément répété donné au sein de génomes d'un grand nombre d'individus.

De plus, cette étude décrit 462 copies endogènes de JSRV quand 27 étaient auparavant décrites chez le mouton. Ce résultat, en plus de confirmer l'efficacité de cette stratégie, nous permet d'étudier la distribution de ces copies au sein du genre *Ovis*. On peut ainsi voir que peu de copies sont fixées dans le genre *Ovis* (14) mais qu'il existe un fort polymorphisme au sein des espèces, avec beaucoup de copies rares. Ces résultats permettent ainsi de voir que le processus d'endogénisation de JSRV est toujours en cours et attestent de la longue coévolution entre les espèces du genre *Ovis* et JSRV.

Enfin, la troisième conclusion majeure réside dans la construction d'un modèle crédible expliquant l'histoire du locus 6q13. Dans ce modèle, JSRV a intégré le locus 6q13 après la spéciation entre *Ovis* et *Capra* et avant la diversification du genre *Ovis*. La mutation et l'amplification de la région ont eu lieu après la divergence entre les espèces « eurasiatiques » et « américaines ». Enfin, la fixation de cette mutation est due à un processus de sélection naturelle au sein de l'espèce *Ovis orientalis* avant la domestication de ceux-ci il y a 10000 ans environ.

Partie 2 - Chapitre 2 : D'autres variants

Si les copies endogènes de JSRV qui ornent les génomes des moutons sont un bon exemple de variants trouvés dans la bibliographie identifiables dans les données WGS, il existe d'autres SVs également décrits dans la bibliographie qui peuvent être eux aussi intéressants à étudier. Les deux exemples étudiés ici ont été choisis pour illustrer l'impact des SVs sur les petits ruminants.
Ces études de cas montrent le potentiel impacte des variants structuraux sur les petits ruminants mais ne constituent pas articles à proprement parler.

Domestication et convergence évolutive, le cas du gène ASIP.

Contexte

Dans la nature, les variations de la couleur du pelage des animaux sont généralement contrôlées par les pressions de sélection naturelles (survie ou reproduction) et peu variables au sein d'une espèce. Chez les espèces domestiques en revanche, la diversité des couleurs et des motifs des robes est plus importante, et cela est notamment du au fait que ces variations ont été sélectionnées dans de nombreuses espèces domestiques (Wright, 2015). Le gène du peptide signal Agouti (*ASIP*) code pour la protéine du même nom. Cette protéine est un signal paracrine agoniste inverse de l'hormone mélanotrope (α-MSH). Le récepteur de la melanocortine (MC1R) dépend de l'α-MSH pour arrêter la production de la phéomélanine (un pigment rouge-jaune) et initier la production d'eumélanine (un pigment brun noir) à la place (Hamosh *et al.*, 2005). La modification de l'activité ou l'inactivation de la protéine ASIP entraîne donc une sur-expression d'eumélanine alors qu'une augmentation de l'activité de cette protéine conduit à une forte production de phéomélanine (Hoekstra, 2006).

La mutation du gène ASIP est associée à des modifications de pigmentations chez de nombreuses espèces domestiques. Chez le cheval (Rieder *et al.*, 2001) et le cochon (Drögemüller *et al.*, 2006), des mutations ponctuelles d'ASIP sont effectivement associées à ces modifications. Chez le chien, une insertion d'un élément transposable semble lié avec des colorations marron-beiges (Dreger and Schmutz, 2011). Chez la vache, l'insertion d'un LINE entraîne là aussi la sur-expression d'un transcrit d'ASIP et induit la couleur bringée de la race normande (Girardot *et al.*, 2006).

Chez le mouton, la duplication en tandem de 190-kb incluant le gène ASIP est responsable de la couleur blanche dominante (Norris and Whan, 2008) alors que chez la chèvre, plusieurs amplifications différentes reportées semblent être liées à différents motifs de coloration de la robe (Dong *et al.*, 2015).

Basés sur l'étude de données de séquençage de génomes complets d'individus sauvages et domestiques des genres *Ovis* et *Capra*, les objectifs ont été ici (i) d'identifier les haplotypes des génomes de référence de la chèvre et du mouton, (ii) de rechercher de la

variabilité dans ces régions à partir des données de séquençage haut débit de génomes complets (WGS) de chèvres et de moutons, puis (iii) d'étudier l'impact de la domestication sur la variabilité de cette région chez les deux espèces.

Matériel et Méthodes

Haplotypage des génomes de référence :

Les génomes de référence utilisés dans cette étude pour le mouton (Oar_v4.0 / GenBank assembly accession: GCA_000298735.2) (International Sheep Genomics Consortium *et al.*, 2010) et la chèvre (Chir_ARS1 / GenBank assembly accession: GCA_001704415.1) (Bickhart *et al.*, 2017)) ont été téléchargés depuis NCBI GenBank.
L'inférence de l'haplotype des génomes de référence à été réalisée en alignant les régions codant pour le gène ASIP contre elles-mêmes pour les génomes de la chèvre et du mouton à l'aide du logiciel BLAST (blastn, paramètres par défaut) (Figure S1 et S2) (Camacho *et al.*, 2009).

Afin de clarifier l'identité de l'élément LOC102109531 situé en amont du gène ASIP chez la chèvre, la séquence allant du début de LOC102109531 à la fin du gène ASIP (des positions 63172969 à 63249542 du chromosome 13 de la chèvre) à été aligné avec la séquence du gène ASIP de l'homme (positions 34186493 à 34269344 du chromosome 11, version GRCh38.p7) avec l'aide du logiciel BLAST (blastn, en utilisant les paramètres par défaut de discontiguous megablast) (Camacho *et al.*, 2009).

Haplotypage des individus WGS :

Les données individuelles utilisées sont issues du séquençage paired-end du génome complet (whole genome sequences, WGS) de 279 individus du genre *Ovis (*151 *Ovis aries*, 19 *Ovis orientalis, 4 Ovis Vignei,* 3 *Ovis dali* et 2 *Ovis canadensis*) et 218 du genre *Capra* (197 *Capra hircus* et 21 *Capra aegagrus*) (Table S1 et S2) (toutes les données sont disponibles sur http://projects.ensembl.org/nextgen).

Le génome de référence du mouton comprenant une duplication de la zone codant pour le gène ASIP (Figure S2), n'a pu être utilisé comme référence, la présence de grandes duplications induisant une mauvaise qualité d'alignement des données WGS. De ce fait, les données WGS de tous les individus (*Ovis* et *Capra*) ont été alignées sur la région allant de la position 62000000 à 64000000 du chromosome 13 de la chèvre afin de rechercher les zones dupliquées. L'alignement a été réalisé avec BWA mem (paramètres par défaut) (Li and Durbin, 2009). Pour chaque individu, l'inférence du génotype à été déterminé en fonction de l'évolution de la couverture le long de la séquence de référence (Comparaison des médianes de la couverture des zones possiblement dupliquées avec la couverture médiane des régions flanquantes non dupliquées). Seules les lectures ayant une qualité d'alignement de 60 ont été considérées (samtools depth –Q 60 (Li *et al.*, 2009)).

Résultats et discussion

La comparaison des séquences génomiques de la chèvre et du mouton fait clairement apparaître de grosses différences entre les deux génomes pour cette zone. Par rapport au génome de référence de la chèvre, le génome de référence du mouton contient une délétion d'environ 46 kb (entre les positions 63099101 et 63147048 du chromosome 13 de la chèvre ARS1, à la position 62836494 du chromosome 13 de OAR_v4) ainsi qu'une duplication en tandem de 189,5 kbp. Cette duplication est située sur le chromosome 13 du mouton, entre les positions 62892415 et 63082093 pour la première copie, et 63082093 et 63271674 pour la seconde, et correspondant à la région comprise entre 63203054 et 63392392 du chromosome 13 du génome de référence de la chèvre (Figure 1). La zone dupliquée code pour les gènes ASIP et AHCY dans le génome de référence de la chèvre. Chez le mouton, la copie de ASIP codée entre les positions 63051795 et 63128668 est bien annotée ASIP tandis que la copie codée entre les positions 62891454 et 62940718 est annotée LOC101111988. Cette copie correspond à une version tronquée du gène, avec seulement les trois derniers exons. Pour le gène AHCY, il est annoté AHCY entre les positions 63143406 et 63159173 et LOC101112245 entre les positions 62953304 et 62969072. Dans le génome de référence de la chèvre (ARS1), aucune duplication n'est visible (Figure S2).
Cette comparaison des génomes de référence permet donc de voir que la duplication incluant ASIP décrite dans la bibliographie (Norris and Whan, 2008) est bien présente dans le génome de référence du mouton. La présence de cette duplication, possiblement liée à la couleur blanche n'est pas surprenante considérant le fait que l'individu utilisé pour faire ce génome de référence est un Texel (Y. Jiang *et al.*, 2014).

La comparaison de la séquence des gènes ASIP de l'homme avec la séquence allant du début de l'élément annoté LOC102109503 à la fin du gène ASIP de la chèvre montre une très bonne homologie des séquences codantes (Figure S3). Ce résultat tend à montrer que LOC102109503 est en réalité la première partie du Gène ASIP caprin, possiblement mal annotée dans le génome de la chèvre.

L'étude de l'évolution de la couverture le long de l'alignement des données WGS sur le génome de référence de la chèvre permet de détecter deux zones qui présentent une couverture augmentée chez certains individus du genre *Ovis* (exemples visibles sur la figure S4). La première région (A^0 sur la figure 2) couvre la séquence de 63202618 à 63392331 du chromosome 13 de la chèvre. La seconde région (B^0) couvre la séquence de 63100646 à 63187473 bp (duplications visibles sur la figure S4). De plus si l'alignement des génomes de référence montre une délétion chez le mouton dans cette zone, l'alignement des lectures issues du reséquençage de génome complet de mouton sur le génome de référence de la chèvre montre que cette séquence est bien présente dans les données de génomes complets (Figure S4). L'absence de cette séquence dans le génome de référence OAR_v4 peut venir d'une délétion chez l'individu utilisé pour construire la référence, d'une translocation de ce fragment ailleurs dans le génome, ou d'un problème lors de l'assemblage du génome.

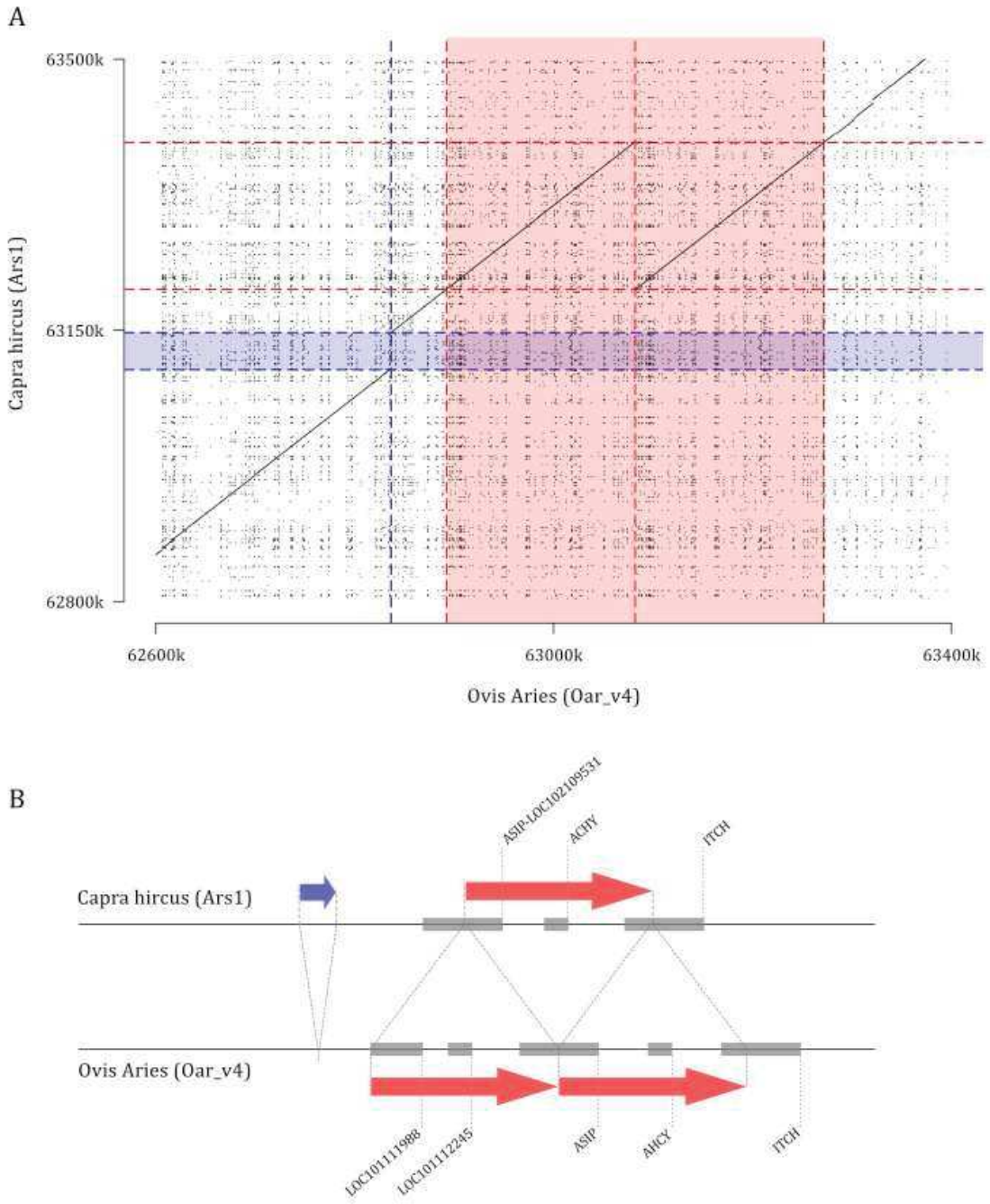


Figure 1 : Comparaison des zones entourant le gène ASIP chez la chèvre et le mouton. (A) Graphique de ressemblance (dotplot) de la séquence entre 62800k et 63500k bp du chromosome 13 du génome de référence de la chèvre (ARS1) alignée contre la séquence entre 62600k et 63400k bp du chromosome 13 du génome de référence du mouton (OAR_v4). La région en bleu est absente du génome de référence du mouton mais présente dans le génome de référence de la chèvre. Les parties en rouge illustrent la duplication au sein du génome du mouton de 189,5 kbp.
(B) Représentation des génomes de la chèvre et du mouton. Les zones absentes (en bleu) et dupliquées (en rouge) sont représentées.

L'alignement des données de génomes complets de chèvre sur la même zone montre elle aussi plusieurs duplications possibles dans cette région (exemples visibles sur la figure S3). La première région (A^C) couvre la séquence de 63228284 à 63380810 bp du chromosome 13. La seconde région (B^C) couvre la séquence de 63158172 à 63203852 bp. La troisième région (C^C) couvre la séquence de 63130270 à 63245092 bp et la quatrième (D^C) couvre la séquence de 63129216 à 63142614 bp. L'ensemble de ces duplications ainsi que les gènes impliqués sont représentés dans la figure 2. L'ensemble des génotypes des individus est recensé dans la Table S1. Aucune corrélation entre les amplifications n'est observable (données non montrées).

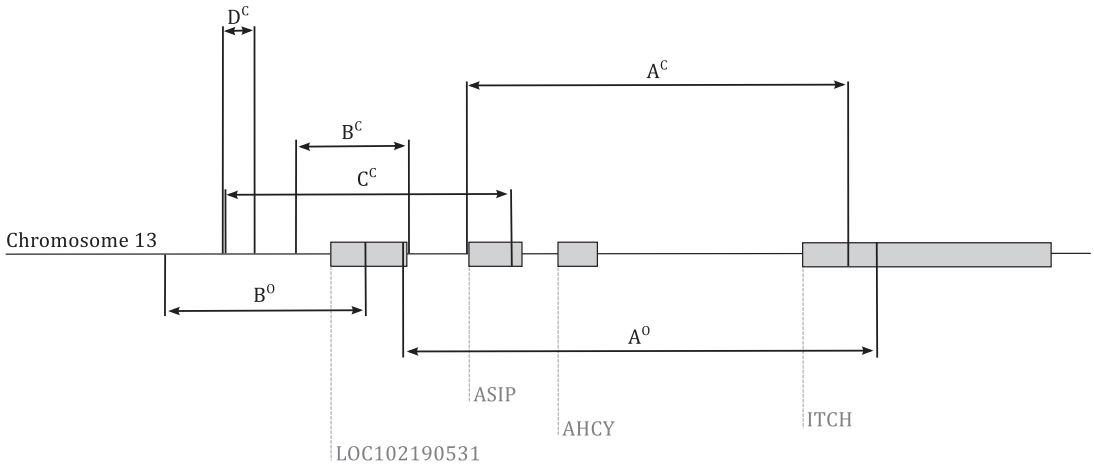


Figure 2 : Représentation schématique de la région génomique entourant le Gène ASIP, ainsi que les diverses zones soumises à variation de nombre de copies chez la chèvre (A^C,B^C, C^C, D^C) et chez le mouton (A^O, B^O).

L'étude de l'évolution de la couverture le long de cette région illustre les CNVs déjà identifiés chez le mouton pour l'amplification incluant le gène ASIP (Norris and Whan, 2008) ainsi que les différentes zones amplifiées chez la chèvre (Dong *et al.*, 2015). Nous pouvons cependant voir que chez le mouton ce sont deux zones qui sont soumises à variations de nombre de copies et non une seule (confirmée par la non corrélation des amplifications). Chez la chèvre ce sont 4 zones qui peuvent être amplifiées (figure S5).

En comparant la distribution des variations de nombre de copies de ces régions dans les différentes populations des deux espèces, nous pouvons observer que, tant pour les moutons que pour les chèvres, les individus sauvages ne présentent pas de CNV dans la région étudiée. De l'autre côté, toutes les populations domestiques présentent des CNVs dans cette zone (Figure 3). Comme les amplifications sont absentes de tous les individus sauvages étudiés, nous pouvons émettre l'hypothèse que les amplifications visibles sont apparues peu de temps après la domestication. Il est aussi possible que ces variants aient été fortement sélectionnés, si ceux-ci existent en faible présence chez les sauvages. Du fait que les amplifications n'ont pas les mêmes bornes chez la chèvre et le mouton, ces mutations semblent indépendantes. Sous l'hypothèse d'un effet sur la couleur d'une duplication du gène ASIP, il pourrait s'agir d'une convergence évolutive pour répondre à une même pression de sélection exercée par l'homme pour une variabilité de ce trait.

Etudier la répartition des allèles des SNPs en déséquilibre de liaison avec ces amplifications devrait nous permettre de tester l'hypothèse d'un effet de la domestication.

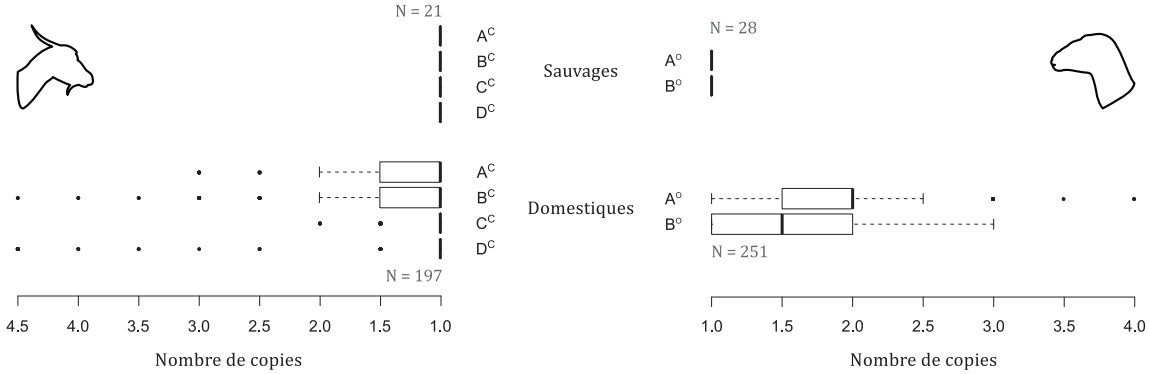


Figure 3 : Boxplot du nombre de copies des différentes régions amplifiées chez les *Capra* (à gauche) et les *Ovis* (à droite), selon que les individus sont sauvages ou domestiques.

Considérant que l'élément LOC102109531 est en réalité le début du gène ASIP, il semble donc que deux amplifications chez le mouton (A^O et B^O) et trois chez la chèvre (A^C, B^C et C^C) impactent ce gène. Au vu de l'importance connue de ce gène dans le déterminisme du pelage des individus (Rieder *et al.*, 2001, Fontanesi *et al.*, 2009, Norris and Whan, 2008), il semble intéressant de lier ces amplifications au phénotype des individus. Les données morphologiques des individus de chèvres et de moutons du Maroc ne permettant pas de faire le lien entre les amplifications et la couleur du pelage des individus, une étude plus ciblée avec une base de données morphologiques précises permettrait de tester cette hypothèse.

Conclusions :

Au vu de l'importance connue du gène ASIP sur la couleur du pelage chez les animaux domestiques, travailler sur la région entourant ce gène semble prometteur chez la chèvre et le mouton, même si ces travaux nécessitent d'être approfondis.
Ces travaux permettent tout de même de mettre en avant différents points. Le premier concerne les génomes de référence et leurs annotations. Avec ces travaux, nous pouvons montrer que le génome de référence du mouton est incomplet, avec une région présente dans le génome de référence de la chèvre ainsi que dans les données de WGS du mouton, mais absente du génome de référence. D'autre part, ces travaux ont aussi permis de monter que l'annotation du génome de référence de la chèvre est imparfaite, l'élément LOC102109531 étant probablement la première partie mal annotée du Gène ASIP.
D'autre part, ces travaux mettent en avant différentes zones amplifiées de façon indépendante chez le mouton et la chèvre. De plus, au vu de la répartition de ces amplifications uniquement visibles chez les individus domestiques, ces amplifications semblent être liées au processus de domestication ou, en tout cas, avoir été impactées lors du processus de domestication ou de sélection qui a suivi.

β-Globine ovine, un potentiel rôle adaptatif.

Contexte

L'hémoglobine a pour rôle de transporter l'oxygène dans le sang. Chez la chèvre et le mouton, les gènes codant pour la beta-globine sont issus d'une série de duplications en tandem d'un cluster de gènes ancestral suivi d'une subfonctionnalisation. Le cluster amplifié contient quatre gènes (5'-e-e-Ψb-b-3'), et ses différentes versions codent actuellement pour la globine embryonnaire (β^E), la globine juvénile (β^C) et la globine adulte (β^A), exprimés aux différents stades de développement de l'individu (Carter, 2009). La β-Globine juvénile possède une meilleure affinité à l'oxygène que la β-Globine adulte (Garner and Lingrel, 1988), et sa production, normalement observée uniquement après la naissance, peut aussi être observée durant les anémies sévères chez la chèvre et le mouton adulte.
Chez le mouton, les individus capables de synthétiser la β-Globine juvénile une fois adultes appartiennent à l'haplotype A. Un haplotype alternatif, l'haplotype B, regroupe les moutons incapables de synthétiser cette β-Globine juvénile, au stade adulte tout comme au stade juvénile. Cet haplotype B ne présente que deux copies du cluster de gènes (Garner and Lingrel, 1988).
Deux hypothèses se confrontent actuellement pour expliquer l'apparition de cet haplotype B. Cet haplotype peut être le résultat d'une délétion d'environ 37kb dans le cluster de gènes codant pour les différentes β-Globines, ou issu d'un transfert horizontal de gènes d'un autre ruminant (Jiang *et al.*, 2015).

Si, d'un côté, les individus homozygotes de l'haplotype B semblent moins tolérants à l'anémie que les individus porteurs de l'haplotype A (Huisman and Kitchens, 1968), de l'autre, l'exposition à certains pathogènes ainsi que la sélection pour un sang moins visqueux (observé chez les individus porteurs de l'haplotype B) dans les climats arides peut expliquer le maintient dans les populations de l'haplotype B (Pieragostini *et al.*, 1994).

Basés sur l'étude de données de génome complet de plus de 400 individus de chèvres et de moutons, ces travaux doivent permettre (i) de tester la variabilité des caprins en recherchant ces deux haplotypes, (ii), de rechercher un éventuel impact de la domestication des moutons dans la fréquence de l'haplotype B et (iii) d'étudier la répartition mondiale de ces deux allèles afin d'étudier leur potentiel rôle adaptatif en lien avec l'environnement.

Matériel et Méthodes

Génotype des génomes de références :

Les séquences des génomes de référence du mouton (Oar_v4.0 / GenBank assembly accession : GCA_000298735.2) (International Sheep Genomics Consortium *et al.*, 2010) et de la chèvre (ARS1 / GenBank assembly accession : GCA_001704415.1) (Bickhart *et al.*, 2017) ont été téléchargés depuis NCBI GenBank.
Le génotypage des génomes de référence a été réalisé par un alignement deux à deux des régions codantes pour la beta-globine des génomes de référence à l'aide du programme BLAST (blastn, paramètres par défauts) (Camacho *et al.*, 2009).

Analyse des données WGS :

L'ensemble des données individuelles utilisées dans cette étude est composé de séquences paired-end de génome complet (whole genome sequences, WGS) de 204 individus de 3 espèces du genre *Ovis* et de 204 individus de 2 espèces du genre *Capra* (Table S1) (toutes les données sont disponibles sur http://projects.ensembl.org/nextgen).
Les données WGS pour les individus de chaque genre (*Ovis* et *Capra*) ont été alignées sur les séquences de référence respectives (Oar_v4.0 et ARS1) selon le pipeline décrit par (Alberto *et al., In prep*).
Les zones utilisées pour l'alignement sont le chromosome 15 entre 47400000 et 47550000bp pour le mouton (Oar_v4.0), et le chromosome 15 entre 33950000 et 34200000 bp pour la chèvre (ARS1).

Basé sur un protocole similaire à celui décrit dans Jiang *et al.*, 2015, l'haplotype de chaque individu à été inféré en se basant sur l'évolution de la couverture du génome le long de la zone d'intérêt. Une couverture constante le long du génome indique un haplotype similaire à celui du génome de référence ; une couverture nulle, un individu homozygote de l'autre haplotype, et une couverture intermédiaire est signe d'un individu hétérozygote (voir Figure S1 pour un exemple de ces 3 cas chez le mouton).

Les génotypes de 70 moutons domestiques ainsi que leur origine géographique ont été récupérés de (Jiang *et al.*, 2015).

Résultats et discussion

Origine des haplotypes

L'analyse des haplotypes des génomes de référence confirme que le génome de référence du mouton (OAR-v4.0) est porteur de l'haplotype B alors que le génome de référence de la chèvre (CHIR-v2.0) est porteur de l'haplotype A (Figure 1). L'haplotypage des individus issus du projet NextGen a permis de voir que l'ensemble des individus du genre *Capra* (*C. hircus* et *C. agaegrus*) sont homozygote pour l'haplotype A (Données non montrées). D'un autre côté, au sein du genre *Ovis*, les deux haplotypes sont détectés dans des fréquences variables entre les différentes espèces et les différentes populations (Table 1). La présence de l'haplotype B chez *O. vignei*, ainsi que chez *O. canadensis* et *O.dali* (Jiang *et al.*, 2015) montre que l'apparition de cet allèle précède la diversification du genre *Ovis*, il y a 2,4 millions d'années (Rezaei *et al.*, 2010). Cette double observation, la recherche infructueuse de l'haplotype B dans les génomes de nombreux caprins d'une part, ainsi que sa présence dans l'ensemble du genre ovis de l'autre, ne nous permet pas de dater l'apparition de l'haplotype B. En effet, si l'haplotype a pu apparaître après la séparation entre *Capra* et *Ovis*, il se peut aussi que cet haplotype soit plus ancien mais que l'haplotype A ai été fixé chez la chèvre et le polymorphisme maintenu chez le mouton.

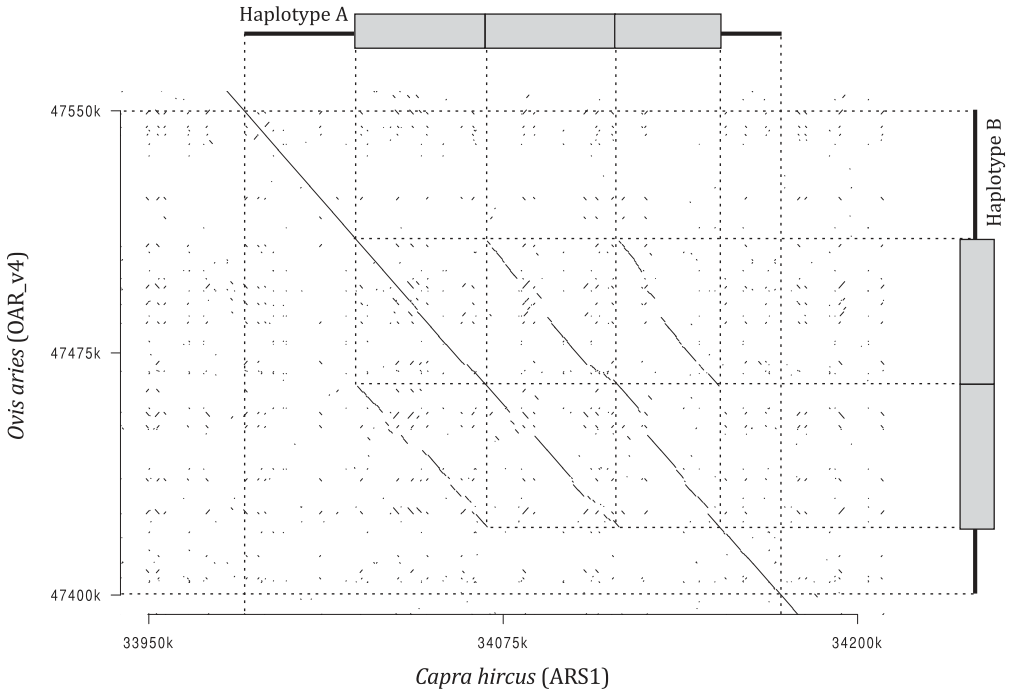


Figure 1 : Graphique de ressemblance (dotplot) de la séquence entre 33950k et 34200k bp du chromosome 15 du génome de référence de la chèvre (ARS1) alignée contre la séquence entre 47400k et 47550k bp du chromosome 15 du génome de référence du mouton (OAR_v4). Les représentations schématiques des zones dupliquées permettent de voir les deux haplotypes, A dans le génome de référence de la chèvre et B dans celui des moutons.

Table 1 : Répartition des différents haplotypes dans les génomes des individus du projet NexGen, issus de différentes espèces du genre *Ovis.*

Espèce	Population	Nombre d'individus	Nombre d'observation			Frequences haplotypiques	
			AA	AB	BB	A	B
O. aries	Marroc	161	1	8	152	0.031	0.969
	Iran	20	0	0	20	0	1
O. orientalis		19	6	8	5	0.526	0.474
O. vignei		4	0	1	3	0.125	0.875

Etude de l'impact de la domestication

La répartition mondiale des haplotypes montre que l'haplotype A est présent dans le génome des moutons de tous les continents en différentes fréquences (voir table 2, données issues de (Jiang *et al.*, 2015)).
D'un autre côté, sur la base des données WGS étudiées ici, il peut être observé, au niveau de l'Iran, bassin de domestication des moutons (Rezaei, 2007), un fort écart de fréquences entre les deux haplotypes entre les individus domestiques et les individus sauvages, avec une fréquence de 0.52 pour l'haplotype A chez les individus sauvages alors que l'haplotype B est fixé chez les individus *Ovis aries* iraniens (Table 1). Chez les moutons marocains, la fréquence de l'haplotype A (0.031) ne permet pas de tester le lien avec l'altitude (Figure 2).

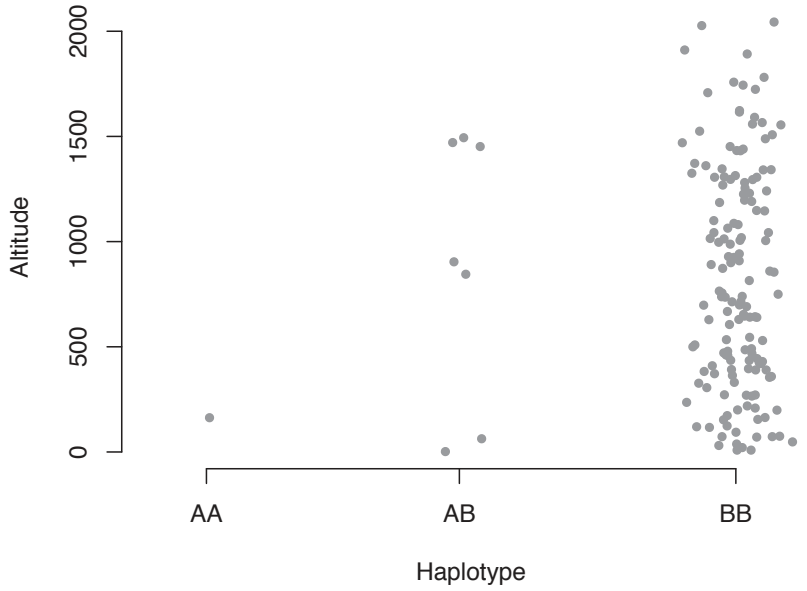


Figure 2 : Distribution des différents haplotypes chez le mouton en fonction de l'altitude au Maroc.

Plusieurs hypothèses peuvent être formulées pour expliquer ces observations.
Si la distribution dans les différentes populations des deux allèles peut être due à la dérive différentielle au sein de différentes populations, la sélection peut elle aussi expliquer cette distribution. L'hypothèse de sélection est compatible avec la fixation de l'haplotype B au niveau du bassin de domestication, couplée avec la très faible diversité nucléotidique dans les régions entourant cet haplotype (Jiang *et al.*, 2015), ce qui peut illustrer l'hypothèse de goulot d'étrangement autour de cet allèle lié au processus de domestication (Jiang *et al.*, 2015).

Table 2 : Répartition des différents haplotypes dans les génomes d'individus issus de populations de différents pays / continents. Données issues de Jiang *et al.*, 2015.

Population	Nombre d'individus	Nombre d'observation			Frequences haplotypiques	
		BB	AB	AA	B	A
Africa	6	5	1	0	0,92	0,08
Americas	7	2	4	1	0,57	0,43
Asia	13	6	5	2	0,65	0,35
Europe	23	12	8	3	0,70	0,30
Middle East	13	9	4	0	0,85	0,15
United Kingdom	8	5	1	2	0,69	0,31
Norh America	5	5	0	0	1,00	0,00

Les conséquences des différents haplotypes restant ouvertes, plusieurs hypothèses peuvent expliquer cette sélection, si sélection il y a.
D'un côté, l'haplotype B semble associé à une plus forte prolificité (Abd-Allah *et al.*, 2012), ce caractère pourrait avoir été sélectionné très tôt dans le processus de domestication et expliquerait la forte prévalence de l'haplotype B. De l'autre, un enrichissement en haplotype B semble observable parmi les populations issues de conditions arides alors que sa fréquence est variable dans d'autres populations (table 2, aussi observé par Ordás, 2004). L'hypothèse, posée par Pieragostini *et al.*, 1994 explique la forte prévalence de l'haplotype B dans ces régions par une sélection directionnelle des conditions arides auxquelles sont exposées les animaux. En effet, les individus porteurs de cet haplotype ont un sang moins visqueux. Cette faible viscosité est synonyme d'une meilleure efficacité respiratoire chez les moutons de l'haplotype B, et à été liée à une meilleur résistance au stress climatique (Pieragostini *et al.*, 2006). Dans le cadre de cette hypothèse, les conditions climatiques expliqueraient la répartition des haplotype A et B dans les cheptels, et illustrent le potentiel rôle adaptatif de l'haplotype B.

Partie 2 : Discussion

Les exemples présentés dans ce chapitre permettent de confirmer l'importance des variants structuraux dans le contexte de domestication et d'adaptation des petits ruminants.

En effet, les CNVs décrits dans la région entourant le gène ASIP montrent un fort polymorphisme au sein des domestiques, mais surtout entre les domestiques et les sauvages, puisque ceux-ci ne semblent montrer aucune amplification génomique. S'il n'est pas possible d'exclure l'hypothèse que ces CNVs sont présents en très faible fréquence chez les sauvages, leur forte prévalence chez les domestiques, contrastant avec leur absence chez les sauvages échantillonnés, tend à illustrer une sélection pour la présence de ces CNVs en lien avec la domestication. Ces observations laissent donc présager que le polymorphisme de structure observé est lié au processus de domestication de ces deux espèces. De plus, les bornes des régions amplifiées sont différentes dans les génomes de la chèvre et du mouton et indiquent que ces évènements sont indépendants (voir Partie 2 - chapitre 2, Domestication et convergence évolutive, le cas du gène ASIP).
L'évolution du cluster de gènes codants pour la beta globine illustre elle aussi l'importance de ces variants. Si les résultats présentés ici ne permettent pas de conclure sur l'origine de l'haplotype B chez le mouton, l'évolution des fréquences respectives dans les populations de moutons domestiques des différents allèles tend à montrer un potentiel rôle adaptatif de l'haplotype B lié aux climats arides (voir Partie 2 – chapitre 2, β-Globine ovine, un potentiel rôle adaptatif). Enfin, si les travaux présentés sur les copies endogènes de JSRV mettent en défaut l'hypothèse d'une sélection lors de la domestication pour la fixation des copies protectrices, ils n'en confirment pas moins l'importance des SVs dans un contexte d'évolution, à travers un exemple de mécanisme de résistance génétique à un virus lié à un CNV (voir Partie 2 - chapitre 1, partie Old origin of a protective endogenous retrovirus (enJSRV) in the *Ovis* genus).

L'ensemble de ces résultats illustre aussi qu'il est possible de retrouver efficacement des SVs dans des données de séquençage de génomes complets. Il est nécessaire pour cela de mettre en place des stratégies propres à chaque variant puisque leur détection dépend du type de variant ainsi que du génome de référence.
Ainsi, s'il est possible d'inférer le génotypes des individus pour la beta-globine basé sur l'alignement des individus sur leur génome de référence respectifs, nous pouvons voir, par l'exemple d'ASIP, qu'il est parfois nécessaire de recourir à des stratégies pour contourner les limites des techniques d'alignement sur des séquences dupliquées. De la même façon, si la recherche des copies endogènes de JSRV se révèle relativement simple et efficace, la recherche des copies enJSRV-20 et enJS56A1 montre les lacunes des génomes de référence aux assemblages imparfaits, et nécessite la mise en place de stratégies plus complexes.
Cette approche « variant candidat » a cependant des limites. La première limite réside dans le fait que l'échantillonnage des individus et la collecte de données, telles que les données morphologiques ou climatiques, n'est pas faite expressément pour

correspondre au mieux à la question posée par l'étude de ces variants. Il n'est donc pas possible de répondre à toutes les questions du fait de l'échantillonnage (soit des individus soit des données récoltées, i.e. les données morphologiques incomplètes ou pas assez précises).
La seconde limite vient du fait que si les données de séquençage de génomes complets permettent de tester *in silco* un grand nombre d'hypothèses sur ces variants issus de la bibliographie, le nombre de cas reste restreint et une approche sans *a priori* semble alors nécessaire pour contourner ces limites.

TROISIÈME PARTIE : Variants structuraux – Approche sans *a priori*

Partie 3 : Introduction

Le premier chapitre de ce manuscrit a permis de confirmer l'importance des variants structuraux génomiques en lien avec les processus liés à la domestication des petits ruminants (de la sélection pour l'adaptation locale aux variations phénotypiques). Cependant, le nombre d'exemples documentés dans la bibliographie reste limité (voir Partie 2 : Discussion). Afin de pallier ce manque, une approche sans *a priori* peut être mise en place. Contrairement à l'approche basée sur l'étude de variants candidats où l'on va chercher à détecter un variant connu, le caractériser et retracer son histoire, l'approche développée ici, que nous pourrions appeler approche « génome complet » vise à détecter un maximum de variants en utilisant des données de reséquençage de génomes complets, indépendamment de leur type, de leur fréquence ou de leur impact, puis de lier *a posteriori* ces variants à des problématiques d'intérêt, comme la domestication ou l'adaptation.

Ce chapitre se divise donc logiquement en deux partie, une première plus méthodologique, concernant la détection des variants structuraux génomiques à partir des données de reséquençage, et une seconde s'attardant sur les conséquences de ceux-ci.

Partie 3 - Chapitre 1 : Détection des SVs dans les données de reséquençage de génomes complets

Séquençage de génomes complets et variants structuraux

Comme décrit dans l'introduction de ce manuscrit, les données issues du séquençage du génome complet d'un individu (WGS) à l'aide des protocoles standards et utilisant un séquençage de type paired-end peuvent être définies par différentes mesures que sont : la taille de la librairie (en bp), la longueur des lectures (en bp) ainsi que la couverture moyenne du génome (en X).
Ces données permettent d'avoir accès au polymorphisme génétique. La détection des mutations ponctuelles (SNPs) est maintenant réalisée en routine pour les espèces modèles. Dans ce cas, la stratégie la plus utilisée comprend deux étapes : la première consiste à aligner les séquences sur un génome de référence puis, à partir de ces alignements, de déduire les polymorphismes par rapport à la référence (pour plus d'informations voir (Nielsen *et al.*, 2011)). Pour les espèces non modèles, ou en l'absence de génomes de référence, il existe des méthodes de détection se passant de la première phase et qui se basent sur une représentation compacte de l'ensemble des lectures (voir Leggett and MacLean, 2014).
Pour la détection des variants structuraux, les deux types d'approches, avec ou sans alignement sur un génome de référence, sont décrits dans la littérature. La première passe en général par l'assemblage *de novo* du génome des individus et la comparaison de ces assemblages, alors que la seconde, et la plus populaire, se base sur l'alignement des lectures des individus sur un génome de référence, puis la recherche d'anomalies dans cet alignement (Alkan *et al.*, 2011).

L'assemblage *de novo* consiste à reconstruire la séquence originale du génome à partir de la seule information des lectures associées au séquençage ainsi que des caractéristiques des librairies dont elles sont issues. Cependant, ces caractéristiques, telles que la taille de la librairie et surtout la longueur des lectures font que cette stratégie est limitée, notamment du fait des nombreux éléments répétés, présents dans les génomes, et qui posent problème lors de la reconstruction de la séquence de l'individu (Alkan *et al.*, 2011).

La stratégie basée sur l'alignement des lectures des individus sur un génome de référence permet de mettre en évidence des variations chez les individus séquencés lorsque des différences entre les lectures et la référence sont observables. Des différences ponctuelles dans la séquence permettent notamment de détecter des SNPs ou des indels.
Dans le cas de variants structuraux, ce ne sont plus seulement les différences de séquences qui importent, mais aussi les incohérences lors de l'alignement qui indiquent la présence d'un variant. Ces incohérences ont été classées en trois catégories : les incohérences dans les paires de lectures (Read-pair ou RP), dans la profondeur de séquençage (Read-depth / Read-count ou RC) ou de lectures coupées (Split-read ou SR) (Alkan *et al.*, 2011).

Les incohérences dans les paires de lectures (RP) se basent sur les caractéristiques des données de séquençage. En effet, la taille de la librairie est connue, et dans le cas où les lectures s'alignent correctement, celles-ci sont orientées l'une face à l'autre (du fait que la lecture du fragment d'ADN initial se fait par les deux extrémités). L'incohérence apparaît lorsque la taille de la librairie ne correspond pas à la taille attendue (trop petite ou trop grande) ou lorsque les deux séquences ne sont pas alignées dans la bonne direction (Korbel *et al.*, 2007).
Ainsi, deux lectures alignées trop loin l'une de l'autre seront la marque d'une délétion chez l'individu par rapport à la référence, alors que deux lectures alignées dans le même sens indiqueront une inversion.
Les incohérences de profondeur de séquençage (RC) se basent sur le fait que la couverture est en théorie la même le long du génome. Si la couverture est augmentée ou diminuée sur une portion du génome, ce sera respectivement la trace d'une duplication d'une portion du génome ou de sa délétion (Yoon *et al.*, 2009).
Les lectures coupées (SR), quant à elles, se basent sur le fait que normalement les lectures s'alignent sur toute leur longueur malgré d'éventuelles différences avec la référence. Lorsque les lectures ne s'alignent qu'en partie de part et d'autre d'une région, c'est la marque d'un point de cassure, trace d'un variant chez l'individu (Mills *et al.*, 2011, Mills *et al.*, 2006).

Chaque type de variant (insertion, délétion, variation du nombre de copies …) va, lors du réalignement des lectures d'un individu, laisser des traces bien spécifiques qu'il faut alors reconnaître et combiner pour identifier un variant (Résumé dans la figure 1).

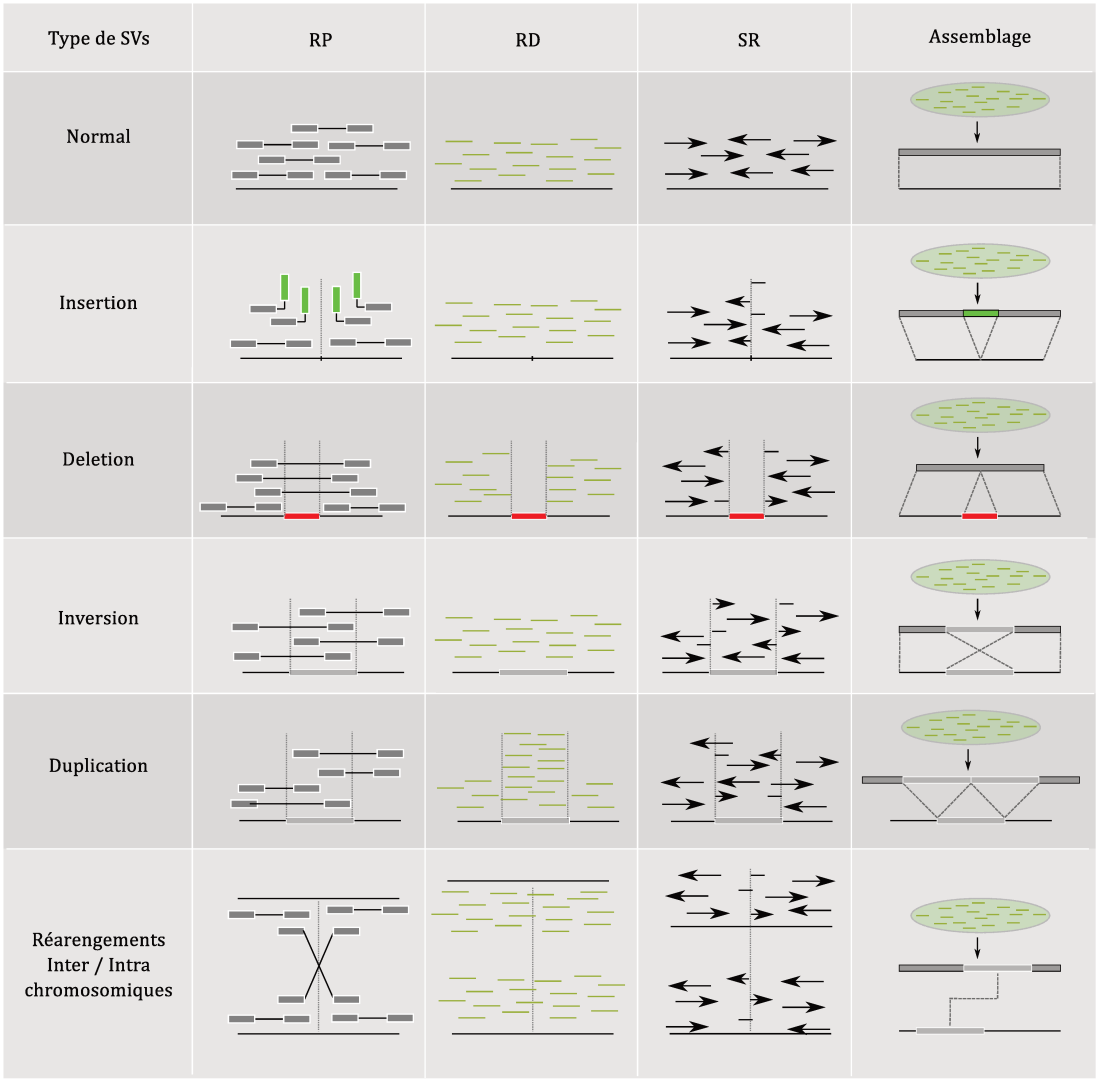


Figure 1 : Résumé des différents signaux laissés par différents types de variants structuraux.

Une multitude de programmes

À ce jour, la liste des programmes qui proposent de détecter des variants structuraux ne cesse de croitre (pour une liste non exhaustive voir Pabinger *et al.*, 2014 ou Hoogendoorn, 2012). Cette multitude s'explique par plusieurs facteurs.

Premièrement, chaque logiciel ne propose pas d'étudier l'ensemble des variants structuraux. Certains logiciels sont spécifiques d'un type de variant, c'est par exemple le cas de CNVnator (Abyzov *et al.*, 2011) pour les CNVs, TE-tracker (Gilly *et al.*, 2014) pour les insertions d'éléments transposables, ou encore Dinumt (Dayama *et al.*, 2014) pour les insertions de génome mitochondrial dans le génome nucléaire. D'autres logiciels proposent de détecter plusieurs types de variants, c'est par exemple le cas de Pindel (Ye *et al.*, 2009), capable de détecter les délétions, les duplications, les inversions ainsi que les insertions mais pas les translocations. Plusieurs en revanche semblent proposer un

panel complet (tels que BreakDancer (Chen *et al.*, 2009) ou Delly (Rausch *et al.*, 2012)) mais aucun ne semble faire consensus à ce jour.

Un second point pouvant expliquer cette multitude vient de la faible concordance entre les variants détectés par les différents logiciels (Pabinger *et al.*, 2014, Zhao *et al.*, 2013). Cette faible concordance s'explique du fait que, pour un type de variant donné, plusieurs incohérences sont détectables dans l'alignement (Figure 1), et que tous les programmes ne se basent pas sur les mêmes signaux. En effet, pour ne reprendre que les deux exemples cités ci-dessus, Delly intègre les incohérences de paires de lectures et les lectures coupées alors que Breakdancer se concentre uniquement sur les paires de lectures, pouvant entraîner des détections différentes pour les mêmes informations de base (Lin *et al.*, 2014). De plus, aucun des deux logiciels n'utilise la profondeur de séquençage alors que c'est la seule méthode utilisée par certains logiciels spécialistes des CNVs (Abyzov *et al.*, 2011).
De la même façon, si deux logiciels peuvent détecter les mêmes variants avec des signaux différents, ils peuvent aussi le faire avec des exigences différences. Des seuils sont forcément nécessaires pour détecter ou non un variant. Ainsi, pour détecter un évènement, on s'attend à ce que l'ensemble des lectures couvrant la zone indique l'évènement dans le cas d'un individu homozygote ou la moitié pour un individu hétérozygote. Mais tout comme la couverture varie, le séquençage de chaque allèle le peut également, et ces variations peuvent entraîner des biais lors de la détection. Ainsi, pour un individu séquencé avec une taille de librairie, une longueur de lecture et une profondeur données, deux logiciels intégrant les mêmes signaux mais pas avec les mêmes seuils, pourraient l'un détecter le variant mais pas l'autre.

Enfin, une autre raison à cette profusion de méthodes et de différences de prédictions vient du fait que les programmes ne définissent pas de la même façon les mêmes types de variants. Prenons l'exemple d'une insertion. Breakdancer détecte les insertions lorsque la taille de la librairie est plus petite qu'attendu (Chen *et al.*, 2009). TE-tracker, de son côté, recherche les paires de lectures avec l'une des deux lectures bien alignée sur le génome de référence, et la seconde lecture non alignée (ou alignée sur un élément répété) (Gilly *et al.*, 2014). Ces deux évènements sont bien des insertions mais les signaux détectables ne sont pas les mêmes car les insertions recherchées ne font pas la même taille. Ces deux types d'insertions peuvent être présents chez un individu mais aucun des deux logiciels n'est capable de les détecter simultanément. Il est donc nécessaire de combiner les méthodes pour prendre en compte un maximum de signaux et assurer un nombre faible de non-détection (Pabinger *et al.*, 2014).

Ainsi, si d'un côté aucun des logiciels existants ne semble satisfaisant et capturer l'ensemble des signaux présents dans les génomes (Pabinger *et al.*, 2014), de l'autre côté, ces programmes ont tous un taux de faux positifs plus ou moins élevé (Layer *et al.*, 2014) qu'il est nécessaire de contrôler.
Afin de contrôler au mieux ces limites, il semble donc nécessaire d'utiliser plusieurs logiciels pour capturer l'ensemble de la variabilité présente dans les données, tout en contrôlant le taux de faux positifs, en se concentrant sur les variants détectés en commun par plusieurs méthodes (Pabinger *et al.*, 2014).

BAdabouM : Pourquoi une méthode de plus ?

Il existe donc beaucoup de méthodes pour détecter des variants structuraux dans des données de reséquençage. Si la théorie recommande d'utiliser un maximum de ces méthodes puis de regarder la convergence (Pabinger *et al.*, 2014), la pratique montre souvent qu'il est compliqué d'installer le(s) programme(s) choisi(s), de le(s) faire fonctionner avec les données disponibles, et de comprendre comment les différents paramètres vont influencer leurs résultats.

BAdabouM est à la base un outil utilisant des critères simples et facilement interprétables nous permettant de parcourir les données de reséquençage réalignées sur un génome de référence et de détecter certains des signaux décrits ci-dessus. Au fil des améliorations, BAdabouM est devenu un outil permettant de détecter des variants structuraux en intégrant un grand nombre de signaux, de façon relativement fiable et ce, dans un temps relativement réduit.

Informations sur l'article

L'article 3 présenté ci-dessous a été écrit en vie d'être soumis au journal *Bioinformatics* sous la forme d'une application note.
Les auteurs sont : T. CUMER, F. POMPANON, F. BOYER

Article 3

BAdabouM: a structural variation discovery tool

Abstract

Summary

Genomic Structural Varations (SVs) are known to have a huge impact on individual's fitness, but remain challenging to detect in Whole Genome re-Sequencing data. Here we introduce BAdabouM, a structural variation discovery tool, able to detect SVs within a short time with high accuracy.

Availability and implementation

BAdabouM is open-source software distributed under the CeCILL license.
Source code is publicly hosted at http://github.com/cumtr/badaboum

Introduction

Rise of Whole Genome Sequencing data (WGS) gives access to huge amount of Single Nucleotide Polymorphisms (SNPs) among population; but those data also contain informations about Structural Variations (SVs). SVs are generally described as genomic rearrangements affecting more then 50 bp. They include deletions, insertions, inversions, mobile-element transpositions, translocations, tandem repeats, and copy number variants (CNVs) (Tattini *et al.*, 2015). These variations are known to have a huge impact on individual's fitness, inducing phenotypical modifications, diseases susceptibility or local adaptation and thus impacting evolution (Chain and Feulner, 2014). Despite their major role on individuals and populations, structural variations survey remains uncommon, mainly due to the lack of standardized protocol to detect them in resequenced genome.

SVs are detectable in WGS data, based on multiple signals such as reads aligned with a split (split reads), abnormally mapped pairs (read-pairs) and non-regular coverage (depth-of-coverage). Split reads are good markers for the breakpoint where the SV begin and ends. Moreover abnormally mapped pairs, with both reads aligned with the same orientation or a longer insert size than expected, and non-regular depth of coverage are good indicators of the presence of structural variations and their type (Alkan, Coe, *et al.*, 2011).

If in one hand, already developed tools seems to perform quite well to detect SVs (even if a multi tool approach is required) (Pabinger *et al.*, 2014), in the other hand, such tools might be hard to install, have non-trivial settings, implies arbitrary thresholds and runs in multiple phases lengthening calculation time.

BAdabouM

Here we introduce BAdabouM, a fast (C written), and multi signal integrating tool for discovering structural variations in diploid genomes. BAdabouM self evaluate multiple alignment parameters and use several signals to detect deletion (DEL), insertions (INS), inversions (INV), copy number variation (CNV) and inter and intra-chromosomal translocation (CTX - ITX).

Input: BAdabouM input file is an indexed bam file, with reads sorted by position.

Pre-processing: BAdabouM brows part of the file to auto evaluate experimental characteristics. Read length, library length and mean coverage are used to evaluate the minimum number of reads necessary to call each type of SVs. This threshold is defined as the ¼ of the expected number of reads under normal conditions that does not map correctly.

Discovery phase: To call SVs, BAdabouM detect specific signatures of SVs based on split reads, read-pair and depth-of coverage.
An *insertion (INS)* is a region where forward reads aligned 5' of breakpoint have dangling mate as well as reverse reads aligned 3' breakpoint. Reads overlapping breakpoint must be soft-clipped.
A *deletion (DEL)* is a region where pairs overlapping breakpoints have longer insert size than expected and where both breakpoints of the deletion are marked by soft-clipped reads.
A *copy number variation (CNV)* is a region with a higher coverage than expected and delimited by two breakpoints highlighted by soft-clipped reads.
An *inversion (INV)* is a region with two breakpoints branded by soft-clipped reads, where forward reads aligned before the first breakpoint have forward mate aligned before the second, and where reverse read aligned after the second breakpoint have reverse mate aligned after the second one.
An *intra Chromosomal Translocation (ITX)* is a region where forward reads, aligned before the first breakpoint, has reversed mates mapping on the same chromosome after the second breakpoint of the ITX. Reverse reads, aligned after the first breakpoint, has mates forward mapping before the second breakpoint. Both breakpoints are highlighted by soft-clipped reads.
An *inter Chromosomal Translocation (CTX)* is a region where forward reads, aligned before the first breakpoint, has mates mapping on an other chromosome on one side of the second breakpoint of the CTX, and where reverse reads after the first breakpoint has mates mapping on the other side of the second breakpoint. Both breakpoints are highlighted by soft-clipped reads.

Output: When BAdabouM detect a specific signature of SVs, it reports it in a table with 8 columns. The three first describes chromosome and breakpoint limits of the beginning of the SV. The three following report chromosome and breakpoint position of the end of the SV. The seventh report SV type and the height the SV size base on breakpoint limits.

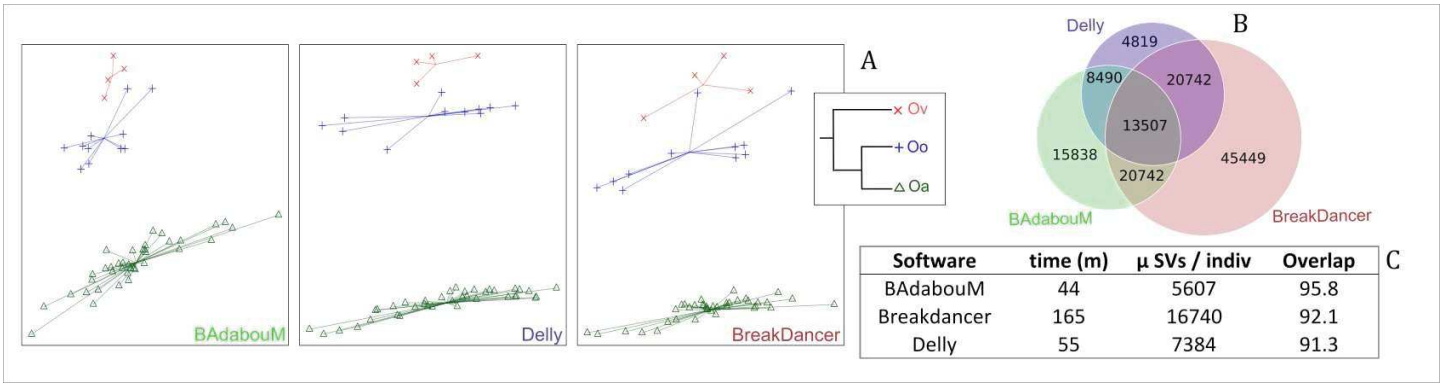
Application

To test BAdabouM's ability to detect SVs, we compared it with two commonly used methods, Delly (Rausch *et al.*, 2012) and Breakdancer (Chen *et al.*, 2009). We took advantage of a recently published dataset of medium coverage whole genome sequences (about 12X) from multiple *Ovis* species (53 individuals) (Alberto *et al., In prep*), to benchmark these different methods.
The goals were to (i) test BAdabouM's ability to discover SVs in real data, (ii) examine the concordances between BAdabouM and other methods and (iii) estimate the performances of the methods.

Based on correspondence analysis (Fig 1A), results show that BAdabouM is able, as the two other methods, to detect biological signals, fully differentiating *Ovis* species and also domestics from wild individuals.

If some variants are detected by all three methods, a greater part is detected only by two of them, and the majority by only one (Fig 1B). This previously described low concordance between methods highlights the technical difficulty to call SVs with only a part of SVs common to all methods (Pabinger *et al.*, 2014), but also differences between SVs signatures according to different methods. Indeed, BAdabouM and Breakdancer does not uses the same signatures to call insertions, inducing a low overlap for the same type of SV.

Performances analysis (Fig 1C) highlight BAdabouM's rapidity, and a lower number of SVs called for each individual. In the same time, BAdabouM's probability for an variant to be called independently in an other individual is a bit higher than the two other methods. This lower numbers of SVs per individuals, linked to the higher recall rate may be due to the BAdabouM's stringency for calling a variant, integrating multiple signals simultaneously for high confidence SVs.



C

Software	time (m)	µ SVs / indiv	Overlap
BAdabouM	44	5607	95.8
Breakdancer	165	16740	92.1
Delly	55	7384	91.3

Figure 1: Application of BAdabouM and two commonly used SV detection tool. (A) Correspondence analysis based on output of each software. (B) Venn diagram of overlap between software. (C) Table summarizing running time, mean number of structural variation per individuals and the probability for an variant to be called independently in at least two individual.

This comparison between methods shows that, as Delly and Breakdancer, BAdabouM is able to detect biological signal in real dataset. If BAdabouM is faster than the two other programs, it's also a bit more conservative; discovering less variants, with a higher recall probability.

Conclusions

Here we introduce a new tool to discover structural variations in whole genome resequencing data, BAdabouM. This tool is easy to install, handle, and is able to detect structural variations in a limited time. While our method does not claim to be flawless, we recommend using it in complement to other existing methods.

Détection de SVs dans les génomes complets : limites et perspectives

Comme listé ci-dessus (Partie 3 – Chapitre 1 : Détection des SVs dans les données de reséquençage de génomes complets), de nombreux points entraînent la mésentente entre les résultats des différentes méthodes de détection de variants structuraux à l'aide des données de reséquençage de génomes complets.

La première limite est pratique mais consommatrice en temps. Elle réside dans le fait que chaque logiciel propose son propre format de sortie sachant qu'aucun ne fait consensus, entraînant une multiplicité de ces formats et des besoins spécifiques pour fusionner ces sorties.

Une autre limite de ces méthodes réside dans le fait que les définitions des variants varient d'un logiciel à l'autre ; un même type d'évènement peut ainsi avoir plusieurs noms selon les méthodes. Si l'on prend le cas d'une région manquante dans le génome d'un individu, on peut parler d'une délétion, mais aussi d'une variation du nombre de copies (CNV) de cette région. Ces deux noms décrivent une même situation. Il en est de même avec une séquence d'ADN mitochondrial présente dans le génome nucléaire de notre individu. On peut parler de Numt (pour « nuclear mitochondrial DNA segment ») ou plus simplement d'une insertion. De l'autre côté, un même nom générique peut décrire plusieurs situations différentes. C'est par exemple le cas de l'insertion, qui peut concerner l'insertion d'un élément transposable, d'un virus endogène ou encore d'une portion d'ADN mitochondrial. Ainsi, un même évènement génomique peut avoir plusieurs noms et inversement, plusieurs noms peuvent décrire un même événement, pouvant entraîner de la confusion lors de la phase de combinaison des méthodes. Homogénéiser ces définitions permettrait sans doute d'augmenter la redondance des méthodes.
Considérant que les différents variants sont autant de boîtes possiblement imbriquées, une ontologie, telle que celle proposée par (Eilbeck *et al.*, 2005), doit permettre de classifier l'ensemble des variants selon un niveau plus ou moins évolué de l'ontologie selon les logiciels, et donc parvenir à recenser et classifier efficacement les variants.

Enfin, l'une des principales critiques que l'ont peut faire aujourd'hui aux programmes détectant les variant structuraux réside dans leur incapacité à définir de façon fiable le génotype d'un individu. Si certains programmes se penchent sur la question (Handsaker *et al.*, 2011), le génotypage des individus reste un véritable problème. La difficulté vient du fait que tout comme pour les SNPs, le génotype doit être inféré à partir de données fluctuantes (la couverture le long du génome ainsi que celle des deux allèles peux varier), avec la difficulté en plus de devoir prendre en compte la taille de la librairie ou le nombre de lectures qui prouvent ou non l'évènement. Le génotypage demande donc de définir, pour chaque type de variant, quelles caractéristiques doivent être étudiées pour évaluer le génotype et retourner celui-ci accompagné d'un score de confiance. Si cela reste possible, le temps nécessaire pour mette en place les stratégies de génotypage ainsi que celui pour le génotypage lui-même peut rapidement s'élever. La parade pourrait être dans le fait de séparer la phase de détection des variants et celle du

génotypage de ceux-ci, pour les différents individus (technique d'ailleurs utilisée pour l'appel des SNPs). Il faudrait alors développer des stratégies propres à chaque variant, permettant de récupérer efficacement les génotypes dans des régions précédemment identifiées. C'est par exemple le cas dans la seconde partie de ce manuscrit (Partie 2 : Variants structuraux – Approche « variants candidats »), où les variants sont connus *a priori* comme ceux-ci sont décrits dans la bibliographie. Dans ce cas, il est possible de le retrouver si le variant est présent dans le génome d'un individu, ainsi que d'inférer le génotype de l'individu.

L'ensemble de ces limites fait qu'il est compliqué, à l'heure actuelle, de couvrir l'ensemble des variants structuraux présents au sein d'une population, ainsi que de connaître de façon certaine le génotype de l'individu. En revanche, l'intersection de différentes méthodes intégrant différents signaux permet de cibler des régions génomiques où un évènement impactant la structure du génome a eu lieu. Lier ces régions avec les éléments génomiques proches (gènes, régions de contrôle …) et à ses caractéristiques populationnelles (fréquence, répartition …) doit tout de même permettre d'étudier le rôle potentiel de ces variants, et au besoin aller les caractériser plus finement comme il est possible de le faire au cas par cas (Partie 2 : Variants structuraux – Approche « variants candidats »).

Partie 3 - Chapitre 2 : Variants structuraux et petits ruminants.

Comme constaté à la fin du premier chapitre et comme attendu à la vue de leur importance dans le contexte de la domestication (Partie 1 : Variants structuraux et animaux domestiques), les variants structuraux semblent avoir joué un rôle important lors de la domestication des petits ruminants et dans leur adaptation à leurs environnements.
Afin d'explorer cet aspect souvent peu étudié du polymorphisme génétique, nous avons recherché les variants structuraux dans le génome de chèvres et de moutons domestiques ainsi que d'individus sauvages pour (i) obtenir un premier catalogue de SVs chez les petits ruminants et approfondir nos connaissances du rôle des SVs lors de la domestication des petits ruminants et (ii) faire le lien entre les variants structuraux et l'adaptation des petits ruminants à leurs environnements.

SVs et Domestication des petits ruminants

Résumé de l'article

Depuis leur domestication il y a dix mille ans, les chèvres et les moutons ont été sélectionnés pour atteindre différents traits (production, adaptations, comportement). Si ces traits sont connus pour avoir des bases génétiques, la plupart des études basées sur des données de génomes complets se concentrent sur les SNPs. D'autres mutations, comme les variants structuraux génomiques (SVs) restent inexplorés malgré leurs conséquences connues sur la fitness des individus. Cette étude se concentre sur l étude des SVs et leur le rôle dans le processus de domestication des petits ruminants.
Basé sur des données de reséquençage de génomes complets de chèvres et de moutons domestiques ainsi que de d'individus d'espèces sauvages apparentées, nous avons détecté 45796 SVs dans les génomes des *Ovis* et 15047 dans les génomes des *Capra*. Si la grande majorité de ces variants semble neutre et reflètent l'histoire des populations, certains variants très différenciés entre sauvages et domestiques semblent avoir été sélectionnés durant la domestication. Si aucun SVs sélectionné ne semble commun aux deux espèces, certains SVs semblent affecter des gènes déjà décrits comme lié à des traits associés à la domestication comme l'amélioration, l'immunité, la reproduction et la survie.
Un regard étroit sur les gènes possiblement impactés montre des gènes connus pour être pléiotropies comme KITLG, LYADM et ADGRG6. Ces gènes, déjà connu pour être ciblés lors de la domestication chez d'autres espèces sont de bons candidats pour des études plus approfondies.
Cette étude pointe le rôle des SVs durant la domestication des petits ruminants et souligne l'importance de leur prise en compte pour la compréhension des bases génétiques de la domestication.

Article 4

Exploring the role of genomic structural variations during small ruminant's domestication

Abstract

Since their domestication about ten thousand years ago, sheep and goats where managed by humans to reach different productive, adaptive and behavioural traits. If such treats are known to have genetic basis, most studies based on whole genome sequences remain focus on SNPs. Other mutations such as genomic structural variations (SVs) remains unexplored despite their known role on individual fitness. In this study, we focus on such variations to unravel the role of genomic SVs during domestication process.

Based on whole genome sequences of domestic's sheeps and goats and their wild relatives, we respectively called 45796 SVs within *Ovis* genomes and 15047 SVs for *Capra*'s. If a wide majority of those variants seems neutral and reflect history of theses animals, we were able to detect highly differentiated SVs possibly selected during domestication. If none of those SVs are shared by the two species, many affect genes already described as being linked with traits likely associated with domestication like improvement, immunity, reproduction and/or survival.

A closer look on theses possibly impacted genes allows identifying pleiotropic genes such as KITLG, MYADM and ADGRG6. Those genes, already described as impacted by domestication in other species, are strong candidate for further investigation.

This study highlights the role of SVs during the domestication process of small ruminant and underlines the importance of taking them in account to fully understand genetic bases of domestication.

Introduction

Plant and animal domestication represent a crucial step in human history, enabling the transition from hunting gathering to farming (Vigne, 2011). It can also be seen as a long-term evolutionary experiment (Larson and Burger, 2013). As a complement to archaeology, the use of genetic provides clues for a better understanding of domestication, specifying its starting place, date, and modality (for a complete review, see Zeder *et al.*, 2006)).

Moreover, human submitted domestics to new selective pressures, inducing phenotypical changes to reach different productive, adaptive and behavioural traits (Wright, 2015). With the rise of next generation sequencing and easy access to whole genome sequences (WGS), genetic approaches have gained much power to identify genes targeted during domestication (domestication syndrome-related characters) and

improvement (primary (e.g. meat) and secondary (e.g. milk or wool production)) process (Wiener and Wilkinson, 2011).
However, such studies mainly remain focus on Single Nucleotide Polymorphism (SNPs) data(Frantz *et al.*, 2016, Orozco-terWengel *et al.*, 2015). We know by studies focusing on both SNPs and other type of genetic variations that SNPs only account for a part of the genetic polymorphism (Pang *et al.*, 2010). Moreover, variations such as short indels or structural variations are known to have a huge impact on individual's fitness (Hoogendoorn, 2012), and may have been targeted during domestication. Structural Variations (SVs) affect more than 1.2% of human genome (including small indels) while SNPs account for 0.01% (Pang *et al.*, 2010). They include insertions, deletions and Copy number variations (CNVs), inversions and inter or intra chromosomal rearrangements (Tattini *et al.*, 2015).
Genomic structural variations have already been linked with domestic's morphological and behavioural traits, like production or reproduction. For example, in pig, CNVs are associated with multiple traits, such as growth or meat quality (J. Jiang *et al.*, 2014), while an insertion within the SPEF2 gene may influence the reproductive performance of boars (Sironen *et al.*, 2012). Changes in appearance are the most striking point when comparing wild and domestics. SVs are known to play an important role in these phenotypes. For example, an inversion near the KIT gene, explain the Tobiano spotting pattern in horses (Brooks *et al.*, 2008) and merle patterning of the domestic dog may be induced by a retrotransposon insertion in the SILV gene (Clark *et al.*, 2006).
Thus, focussing on SNPs allows catching only a part of the information and may provide an incomplete view of the genetic basis of some phenotypes. This is the case with Dog's ability to digest starch-rich diet. SNPs around the *AMY2B* gene indicate selective sweep, while structural variations analysis highlights a CNV, explaining the increase in activity of amylase (Axelsson *et al.*, 2013).

Small ruminants sheep (*Ovis aries*) and goat (*Capra hircus*) are phylogenetically close species, as their wild ancestors, the Asiatic mouflon (*Ovis orientalis*) and the Bezoar ibex (*Capra aegagrus*) diverged between 11.6 and 5.3 MYBP (Hassanin *et al.*, 2012). They both were domesticated near the Fertile Crescent about 10 KYBP where their wild relatives remain, and both spread beyond their native range following human trade routes (Zeder, 2008). Moreover, both species were managed to reach similar productive goals (i.e.: providing meat, milk or wool) (Larson and Fuller, 2014). Thus, the parallel history of those two species offers a unique context to study genetic convergences during domestication and improvement.

Tacking advantage of whole genome sequences of domestic's sheep and goats and their wild relatives, previous work based on SNPs data highlighted common genomic regions linked to domestication in both species, which were enriched, among other functions, for genes associated with neural development and the central nervous system. In this work, we focus on structural variations in small ruminants genomes. Based on the same WGS, objectives were to understand the role played by SVs during domestication and improvement of small ruminants, and study the convergence between SVs and SNPs.

Material and methods

Dataset

In total, 51 whole genome sequences for *Ovis* species and 58 for *Capra* species were retrieved from the ENA archive (Alberto *et al., In prep*) (Information available at ftp://ftp.ebi.ac.uk/pub/databases/nextgen/).
Sampling was designed specifically to detect the genetic basis of domestication. It is composed of wild individuals of both species: 11 wild *Ovis orientalis,* 18 wild *Capra aegagrus*); domestics individuals form Iran (20 sheep and 20 goats) and domestics individuals from Morocco (20 sheep and 20 goats).

SVs Calling and filtering

Read mapping

Illumina paired-end reads for *Ovis* were mapped to the sheep reference genome (build Oar_v4.0 - GenBank assembly accession: GCA_000298735.2) and for *Capra* to the goat reference genome (build Chir_2.0 - GenBank assembly accession: GCA_000317765.2) using BWA-MEM (Li and Durbin, 2009). The BAM file produced for each individual was sorted using Picard SortSam and improved using Picard MarkDuplicates (http://picard.sourceforge.net).

SVs calling

SVs where called independently for each individuals using three different methods; BAdabouM (Partie 3 – chapitre 1 : Article 3 - BAdabouM: a structural variation discovery tool), delly (Rausch *et al.*, 2012) and Breakdancer (Chen *et al.*, 2009). All three methods were run using default parameters, except for the mapping quality where a quality of 60 was required.

SVs clustering and filtering

For each individual, SVs called by two methods were considered as identical based on their reciprocal overlap. As methods do not detect breakpoints with the same accuracy, a threshold was set for congruent overlap. For inversions, deletions, duplications and intra-chromosomal translocations, a criterion of SV overlap greater than 50% was used. For insertion, as breakpoints may be narrow (only one bp), the fact of overlapping was considered as sufficient to be considered as the same event. For intra and inter-chromosomal translocation, breakpoints had to be within a 1kb window to be considered as the same event.
To consider SVs called for multiple individuals as homologous, the same strategy was applied (based on reciprocal overlap). The threshold on the reciprocal overlap was set to 70% for inversions, deletions, duplications and inter and intra-chromosomal translocations. For insertions, the fact of overlapping was considered as sufficient to be considered as the same event. Breakpoints of inter-chromosomal translocation on both chromosomes had to be within a 1kb size window to be considered as the same event.
For further analysis, only SVs called by at least two methods and presents in at least two individuals but not fixed where kept, except for insertions: insertions called only with

BAdabouM were kept as this software apply a high stringency criterions for insertion calling).
This step allowed to identify insertions, deletions, inversions, CNVs and inter or intra-chromosomal translocations. For further analysis and for each individual, each SV was considered as a dominant presence absence marker.

SVs analysis

All statistical analyses were performed using the R language (Ihaka and Gentleman, 1996).

Population characteristics of SVs

For each SV, the allele frequency was calculated as the number of time a SVs was called divided by the number of individuals.
Polymorphism within population was estimated as the number of polymorphic loci in the population divided by the total number of loci. The genetic structure of populations was inferred using sNMF algorithm implemented in the LEA package (Frichot and François, 2015). sNMF was run for a number of ancestral populations (K) ranging from 1 to 4, with 20000 iteration and 10 repetition. K=1 was considered as the best K based on the minimal cross entropy criterion (Frichot *et al.*, 2014).

SVs repartition between populations

Differentiation index (DI_{SV}) was calculated as the difference between frequency of the SV in domestics and the frequency in their wild relatives. Structural variation's Weir and Cockerham FST (FST_{SV}) (Weir and Cockerham, 1984) was calculated based on the allele frequency between wild and domestics animals.
For SVs within the 99^{th} quantile of both FST_{SV} and DI_{SV}, SNPs within 500 kb around were called using samtools mpileup on reads mapped with quality equal to 60. Weir and Cockerham FST for each SNP (FST_{SNP}) was then calculated between wild and domestics and within domestics using vcftools.

SVs and genetic features

If an SV overlapped or was close (less than 500 bp) from an annotated gene, this gene was considered to be possibly impacted by the SV. Functional annotation of those genes was then inferred based on a bibliographic review of already known functions of these genes in other livestock species.

Results

Population characteristics of SVs

In this study, a total of 58 *Capra* and 51 *Ovis* WGS where used to call genomic Structural variations. Both groups include wild animals (*Capra aegagrus* (18) and *Ovis orientalis* (11)) and two groups of domestics (*Capra hircus* from Iran (20) and Morocco (20) and *Ovis aries* from Iran (20) and Morocco (20)).

After filtering, 45796 SVs were kept with a median number of 14222 SVs per individual for *Ovis*. For *Capra* a total of 15047 SVs were called with a median number of 3639.5 SVs per individual. Composition in SVs is resumed in table 1.

Table 1: Structural variations classes and distribution within individuals.

	Ovis		Capra	
SV class	No. sites	Mean number / ind (sd)	No. sites	Mean number / ind (sd)
Total	45796	14177 (1013)	15047	3554 (362)
Deletions	30096	10076 (614)	12508	2985 (336.5)
Insertions	12816	3113 (674)	1555	303.5 (45.5)
Inversions	1806	673 (37)	743	212 (33)
CNV	240	49 (8)	73	13 (4)
ITX	99	68 (10)	4	2 (1)
CTX	739	197 (43)	164	40 (12)

The number of SVs per individuals between wild and domestics was significantly different in both sheep and goats (t-test p-values: 0.035 and 0.00498). Wild *Ovis* arbour more SVs than domestic sheep, whereas wild Capra arbour less SVs than domestic goats. There was no significant difference between domestic populations within both species (Figure 1).

The distribution of SVs within populations shows that most SVs are shared among populations (Figure 1). Most SVs are rare with 41.4% and 49.4% having an allele frequency (VAF) lower than 10% for sheep and goats respectively.
For sheep and goats, rare SVs are more likely to be specific to populations, using a threshold on the allele frequency of 0.12 and 0.15, more than 95% of SVs are shared among the different populations.
The polymorphism within population is lower in wilds animals than domestics, for both sheep and goats. Moreover, most SVs are rare with 41.4% and 49.4% with VAF < 10% for sheep and goats respectively.
For sheep, the polymorphism is 0.786 for the wild and 0.853 and 0.868 for Iranian and Moroccan *O. aries* respectively, whereas for goats, the polymorphism is 0.703 for *Capra agaegrus* and 0.805 and 0.770 for Iranian and Moroccan *C.hircus* (respectively).
Even if the structure analysis with sNMF predicts K=1 as the best number of genetic clusters for both species, when using K=2, the structure analysis, in accordance with a hierarchical clustering of individuals based on SVs show a clear differentiation between wild and domestics and within domestics (Figure 1).

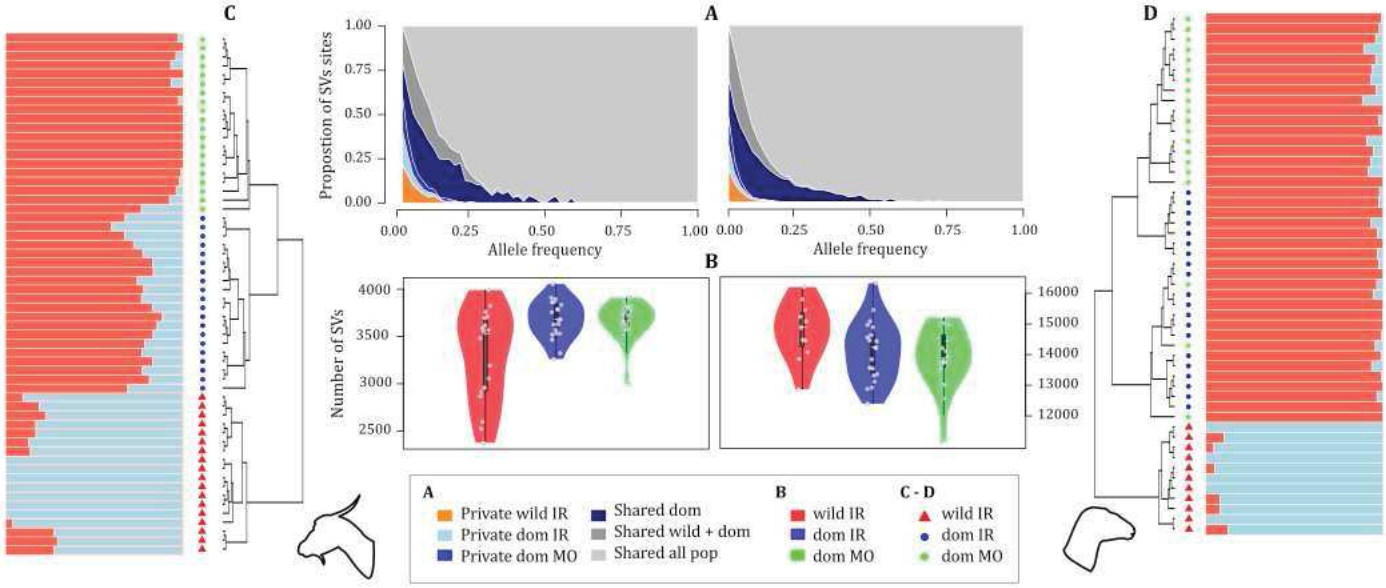


Figure 1: Population properties of SVs in both sheep and goat. **A**. SV allele sharing between populations. **B**. SVs distribution within populations. **C** and **D**. Hierarchical ascendant classification of individuals and population structure for K=2.

SVs repartition between populations

There is a clear correlation between DI_{SV} and FST_{SV} (goat - r^2: 0.6165, p-value < 2.2e-16, sheep - r^2: 0.4896, p-value < 2.2e-16).

The DI_{SV} index is based on frequency within both populations and may be biased when a locus is absent from a population and at medium frequency in the other. Thus it's value will be low while this SV contrast population. On the other hand, the FST_{SV} calculation is biased by the non-homogeneity of group size between wilds and domestics. Thus, observed heterozygosity within domestics is close form global expected heterozygosity even if the SV is absent from wilds.

Thus, to select the most differentiated SVs between wild and domestics and avoid false positives, SVs where filtered to be simultaneously within the 99th quantiles of FST_{SV} and DI_{SV}. After filtering, 135 SVs where kept for sheep and 70 for goat (Figure 2). Lowest FST_{SV} value for selected SVs is 0.25 for goat and 0.32 for sheep, and the lowest DI_{SV} is 0.58 for sheep and 0.56 for goat. 31 SVs for sheep and 26 SVs for goats were found to be overlapping with genes. Those SVs are reported in table S1. Among the selected SVs, 78.5 % are shared between wilds and domestics for goats and 81,4% for sheep.

Among the SVs detected as differentiated between wild and domestics, respectively 0.043% (3 over 70) for goats and 0.03% (4 over 135) for sheep were found within regions detected by a study based on SNPs (table S1). Among those SVs constants with SNPs based study, only two SVs overlap with a genes in Capra (and none in sheep). We detected an intronic deletion LOC10217464 and an intronic deletion KITLG in goats.

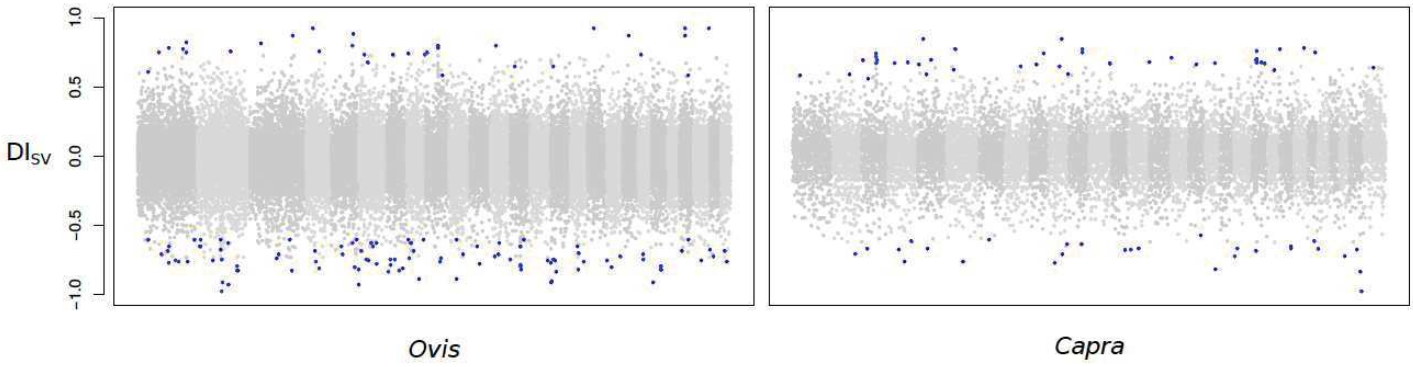


Figure 2: Manhattan plot of the DI_{SV}. SVs considered as selected with both FST_{SV} and DI_{SV} are in blue.

Table 2: list of genes possibly impacted by SVs and functions they are involved in.

Function	Ovis	Ref	Capra	Ref
Production	SLC40A1 LOC101122056 - TMED2 FMNL2 LRP1B ANKRD44 MTFR1 COPS5	Zhao *et al*, 2015 Purfield *et al,* 2017 Silva *et al*, 2016 Zhang *et al*, 2012 Gonzales *et al*, 2017 Jiang *et al*, 2009 Ramayos *et al*, 2014	KCTD8 DNMT3A LDAH VPS52	Hayes *et al*, 2008 Lui *et al*, 2015 NCBI Tanaka *et al*, 2006
Imunity	CD28 ADGRG6 CD226	Chaplin *et al*, 1999 Rigter *et al*, 2015 NCBI	LOC102174264 - GBP5 LOC102180934 - GBP6-like LOC108633177 - GBP6-like SEMA3D PPHLN1	Koltes *et al*, 2015 Kommadath *et al*, 2017 Kommadath et al, 2017 Moreti *et al*, 2006 Pant *et al*, 2010
Reproduction &survival	THSD4	Mokry *et al*, 2013	KITLG BMPR1B FOXO3 LOC102180801 - MYADM TSG101	An *et al*, 2012 Polley *et al*, 2010 Byun, 2012 Cinar *et al*, 2016 Dias *et al*, 2017

Discussion

SVs and small ruminants

This study is the first to characterize genomic structural variations on both sheep and goats in the context of domestication. Based on a combination of multiple calling tools, in this study we produced a set of high confident SVs. This set is composed in a wide majority of deletions and insertions, some CNVs and a few inversions and inter or intra chromosomal translocations, which is coherent with structural variations composition in other mammalian genomes (i.e. cattle (Chen *et al.*, 2017)) as well as in human populations (Sudmant *et al.*, 2015).

Still, a striking difference between sheep and goats in term of number of SVs is observable. This difference may be explained by ancestry effective population size. Indeed, past *Ovis* population size is larger than *Capra*'s before (Alberto *et al., In prep*). In terms of genetic variations, a larger population induces more mutations and a reduced probability for mutations to be fixed for a given period. Methodological issues may also explain this difference. Indeed, considering that calling SVs relies on assembling quality of the reference genome used to realign WGS, quality differences between sheep and goat assembly may lead to the observed differences, with false positives or true negative in one or both species (Bickhart and Liu, 2014).

Nevertheless, population properties of SVs in sheep and goats seem quite comparable. Rare SVs are mostly population specific, and SVs with a VAF>0.15 are nearly shared by all populations for both species, indicating that observed polymorphism is inherited from before domestication, which is coherent considering that our sampling is mostly composed of insertions and deletions with low mutation rate (Sudmant *et al.*, 2015).

The number of SVs between domestic and wild animals is significantly different in both species. For sheep, the number of SVs is higher in wild than domestics, reflecting either bottleneck at domestication or a higher differentiation time between wild animals and the one used as reference, than between domestics and reference, as it's had been shown for human (Li *et al.*, 2011). Such result is consistent with those obtained when looking at SNPs. Indeed, nucleotide diversity is lower in domestic sheep than in Asiatic mouflon, suggesting stable demography in the wild and/or lower effective size in domestics (Alberto *et al., In prep*). For goats, in contrast, actual populations of wild animals are fragmented and may suffer from recent bottleneck inducing observable variability within population (Alberto *et al., In prep*), which may explain the lower SVs polymorphism and the lower number of structural variations within wild population. Based on SNPs, Bezoar ibex also showed lower nucleotide diversity and higher inbreeding both populations of domestic's goats (Alberto *et al., In prep*).

However, structural variations are sufficient to reconstruct population structure and history, also visible when studying SNPs (Alberto *et al., In prep*). Thus, genomic structural variations allow a clear differentiation between wild and domestics for both species. As SNPs, SVs are good neutral markers of population history even if structural

variations does not discriminate populations as well as SNPs, witch is consistent with previous studies (Conrad and Hurles, 2007) and may be explained by the lower mutation rate of structural variation (notably insertions and deletions) (Sudmant *et al.*, 2015).

A previous work highlighted that structural variations are inequitably distributed over genomes, with higher density within heterochromatic regions (Li *et al.*, 2011). Those regions are largely composed of repeated elements, where current NGS-based methods have low sensitivity for detecting structural variations due to the short size of the reads (Medvedev *et al.*, 2009). As a consequence, our set of structural variations is incomplete, and this imprecision is probably not fairly distributed over the genome. Thus, get interest in SVs distribution over the genome seems to bias to give information's about genomic properties of SVs in small ruminants genomes.

SVs during sheep and goat domestication and improvement

Based on FST_{SV} and DI_{SV} between wild and domestic animals, some SVs are more highly differentiated than expected and have potentially been selected during domestication. Among those SVs, a more than half (respectively 77,0% (104 SVs) and 62,8% (44 SVs) for sheep and goats) is in non-coding regions. Such result is expected considering that regions highly differentiated based on SNPs are also in such regions (Alberto *et al., In prep*) and is not surprising regarding the fact that non-coding regions are important for gene regulation and still can have impact on individual fitness (Carroll *et al.*, 2008).

Close study of genes possibly impacted by SVs highlights genes and functions witch have already been linked with domestication in other livestock species. Among the 57 genes (respectively 31 in sheep and 26 in goats) impacted by highly differentiated SVs, 12 have previously been linked with improvement, 8 with immunity and 6 with reproduction and/or survival (Table 2). If the wide majority of those genes are reported as selected during domestication, at least three are known to have a pleiotropic effect. KITLG gene is known to affect prolificity and litter size in goats (An *et al.*, 2012), but is also associated with coat color in multiple domestics such as pigs (Hadjiconstantouras *et al.*, 2008) and cattle (Pausch *et al.*, 2012). The MYADM gene has been associated with high production traits (Dong *et al.*, 2015) and weight of weaned kid in goats (Gonzalez *et al.*, 2013), but also with lifetime cumulative ewe production and wool traits (U. Cinar *et al.*, 2016). ADGRG6 also known as Gpr126, is associated with weight gain in pigs(Strucken *et al.*, 2014), but also with cartilage biology (Karner *et al.*, 2015) and play a major role in neural development and myelination (Li *et al.*, 2014).

Considering regions detected by SNPs (Alberto *et al., In prep*), only a low convergence (3 SVs for goats and 4SVs for sheep within windows detected with SNPs) is observed, with only *KITLG* and *LOC102174264* gene in *Capra* detected by both studies. Low convergence between the two methods is not surprising considering the fact that SNPs based study used highly selective thresholds to avoid false positives (Alberto *et al., In prep*), then many regions under selection may have been discarded to avoid false positives. This is particularly likely when looking at the FST_{SNP} based on SNPs around some detected SVs

which highlight selective sweep (Figure 3). Thus SNPs and SVs are congruent to some point.

Moreover, both informations may be complementary. Considering for example the KITLG locus, the fact that SNPs indicates a relaxed selection in domestics supports the persistence in populations of the detected deletion, as alteration of the gene are no longer counter-selected in domestics (Alberto *et al., In prep*). Alternatively, FST_{SNP} within SLC40A1 gene form a narrow peak, witch may be explained by the presence of the structural variation (Figure 3). Thus, SVs and SNPs are complementary when looking at genetic basis of domestication and improvement.

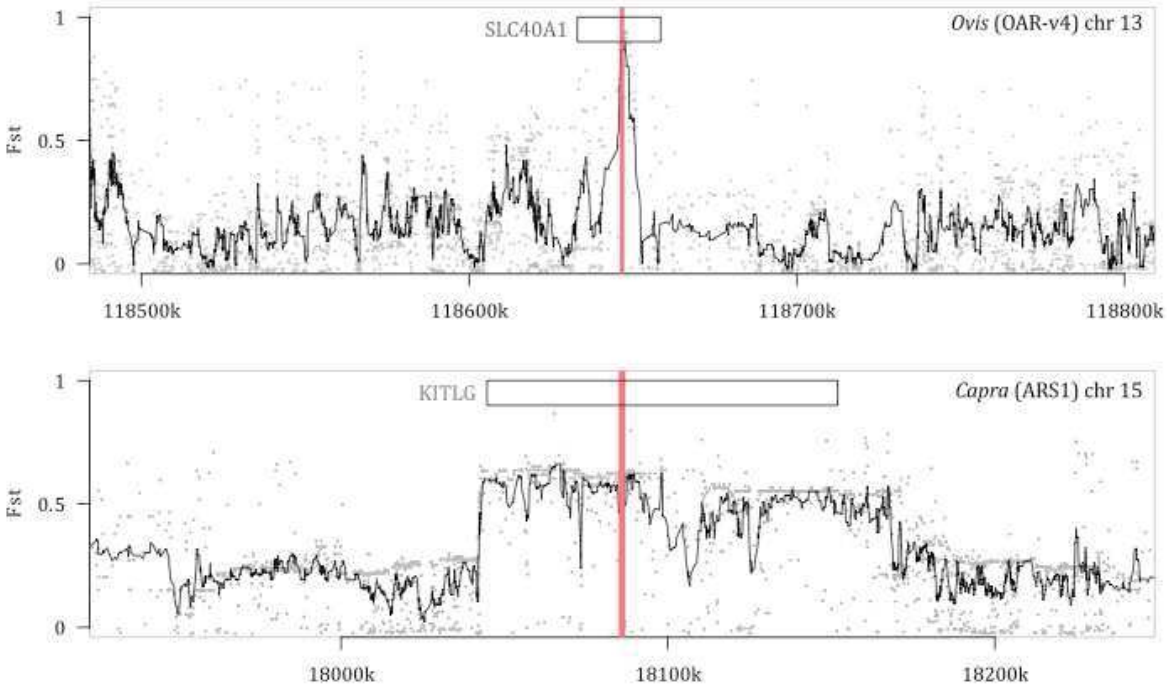


Figure 3: Fst between wild and domestics animals along regions surrounding putatively selected SVs within genes. Top panel is in the 13th chromosome of sheep. Down panel is in the 15th chromosome of goat.

Structural variation and domestication

Our study, only based on SVs, does not highlight converging selection between *Ovis* and *Capra*. Such result is in opposition with those obtained with SNPs (Alberto *et al., In prep*). Considering SVs possibly selected, most of them are probably inherited from the wild ancestors as they are present in few copies in the wild relative population. Thus, SVs selected during domestication and improvement are non-novel, which is compatible with the low mutation rate of SVs (Sudmant *et al.*, 2015) and may explain this non-convergence.

It's hard to speculate about the functional impact of SVs described in this study. As none of the selected SVs are in exon, we can not say if SVs detected as selected during domestication (or contribute to the selected phenotype), if they are detected as selected due to some hitchhiking by a nearby causal mutation (SNP, short indel …) or if observed differentiation is only due to drift. Anyway, the fact that all those SVs are in intronic

regions does not mean they are neutral.(i.e. the intronic insertion in the SPEF2 gene in pig is influencing the reproductive performance of boars (Sironen *et al.*, 2012)). In order to validate the biological consequences of the SVs detected as associated with domestication in this study, further functional investigations such as expression levels measurements or protein conformation and addressing testing would be mandatory.

This study highlights the importance of considering structural variations when studying domestication process. Indeed, our survey allows to identify Genes possibly impacted by SVs and selected during domestication process in both sheep and goats genomes.

SVs et Adaptation des petits ruminants

Résumé de l'article

Le processus au travers duquel les individus d'une population évoluent pour être plus adaptées à leur environnement que les autres individus de la même espèce est appelé adaptation locale. Depuis leur domestication, les petits ruminants ont été propagés en dehors de leur aire de répartition naturelle, et sont élevés dans des climats variés. De ce fait, ils ont accumulé des traits adaptatifs à ces environnements, offrant un contexte idéal pour l'étude des bases génétiques de l'adaptation local.

Si l'étude des bases génétiques de l'adaptation sont maintenant bien étudiées, le rôle des variants structuraux génomiques (SVs) reste inexploré malgré leur conséquences connues sur la fitness des individus.

Basé sur des données de reséquençage de génomes complets de chèvres et de moutons échantillonnés pour couvrir l'ensemble du territoire du Maroc, l'objectif à été d'étudier les bases génétiques de l'adaptation le long de gradients environnementaux continus.

Après avoir détecté 57775 SVs chez les moutons et 18172 SVs chez les chèvres, nous avons été capables de détecter 130 SVs pour les moutons et 35 pour les chèvres dont la distribution covarie avec des variables environnementales. Cette étude permet ainsi de voir que les SVs jouent possiblement un rôle dans l'adaptation locale des petits ruminants en affectant des gènes impliqués dans la morphologie, l'immunité et le métabolisme.

Ainsi, les variants structuraux génomiques affectent une large part des génomes des petits ruminants et semblent jouer un rôle important dans l'adaptation locale des petits ruminants. La prise en compte des SVs complète notre compréhension des bases génétiques de l'adaptation.

Article 5
Small ruminants adaptation: deciphering the role of structural variations

Abstract

The process through which a population has evolved to be more well-suited to its environment than other members of the same species is called local adaptation. After their domestication, small ruminant spread out their natural range, were raised under various geo-climatic conditions, and accumulated adaptive traits to their environments, offering ideal context for studying genomic basis of adaptation.
If genetic basis of local adaptation are now well studied, the role of genomic structural variations remains unexplored despite their documented impact on individual's fitness.
Based on whole genome sequences from sheep and goats covering the Moroccan territory, we aimed to study genetic basis of local adaptation along continuous climatic gradients.
After calling 57775 SVs for sheep and 18172 SVs for goats, we were able to detect 130 SVs for sheep and 35 SVs for goats that covariate with environmental variables. Our results highlight that structural variations may have played a role during local adaptation by affecting genes implied in morphology, immunity, and metabolism.
Thus, genomic structural variations affect a large part of small ruminant genomes and may play an important role for local adaptation. Tacking SVs into account complete our knowledge of genetic bases of local adaptation.

Introduction

By selecting more adapted individuals, the environment imposes natural selection and average fitness of individuals increase in their local environment over generations. As a consequence, individuals have a better fitness in their local environment than individuals from other populations (Kawecki and Ebert, 2004).
After initial domestication, livestock spread out their natural range and were raised under various environments representing a wide range of geo-climatic conditions (Taberlet *et al.*, 2008). These habitats include conditions that are very different from their natural one and are very contrasted such as arid deserts or Himalayan Mountains. Thus, domestics accumulated adaptive traits to their new environments and offer ideal context for studying genomic basis of adaptation.

The Moroccan territory represents a good case study for studying local adaptation to various environments. It's geography spans from the Atlantic Ocean on the west side, to mountainous areas in the centre and east side, with the Sahara desert on the south and the Mediterranean sea on the north. Hence, the northern part of the country is under Mediterranean climate, with moderately hot and dry summers and mild and wet

winters. The south-eastern portions of Morocco are very hot, and include portions of the Sahara Desert. The west coast is under Ocean influences, with cool summers and cool winters, and irregular precipitation in winter and intense drought in the summer months. Finally, high atlas and high plateaus are characterised by two types of alpine climates (Zereini and Hötzl, 2008).

Domestics sheep and goats are found throughout Morocco and mostly belong to indigenous landraces locally adapted to their environment (Benjelloun, 2015). Thus, Moroccan small ruminants provide ideal context for studying genetic basis of local adaptation effects.

To date, the detection of adaptive genomic basis in landscape-genomic studies was mainly based on population genomic approaches by contrasting populations submitted to different conditions (Manel and Holderegger, 2013). However, correlation between allele frequencies and environment factors, implemented in new methods (Frichot *et al.*, 2013) (Stucki *et al.*, 2017) allows to clearly link genetic and environmental variations.

The rise of Whole Genome Shotgun Sequencing (WGS) offered the possibility to finely map genomic regions involved in adaptation process. The vast majority of studies focus on Single Nucleotide Polymorphism (SNPs), thus accounting only for a part of polymorphism. Other variations such as short indels or structural variations (SVs), are known to have a huge impact on individual's fitness through regulation effect or changes in genomic architecture (Feuk *et al.*, 2006), and may play a role in adaptation (Bickhart *et al.*, 2012; Paudel *et al.*, 2013). Structural variations include insertions, deletions, inversions, inter and intra chromosomal translocations and copy number variations. They are known to alter more than 1.2% of human genome while SNPs account for only 0.01% (Pang *et al.*, 2010). A previous work was able to associate SNPs to signatures of selection along climatic gradient, highlighting the genetic adaptation of Moroccan small ruminants to their local environment (Benjelloun, 2015).

Tacking advantage of the same dataset of whole genome sequences of more than 300 sheep and goats from Morocco, our objectives were (i) to detect genomic structural variations differentially distributed across environmental parameters, and (ii) link those structural variations with local adaptation of small ruminants in Morocco.

Material and Methods

Dataset

A total of 159 low coverage (12x mean coverage) 101bp paired-end shotgun sequencing for *Ovis aries* individuals and 160 for *Capra hircus* from Morocco were retrieved from the ENA archive (Alberto *et al., In prep*) (Information available at ftp://ftp.ebi.ac.uk/pub/databases/nextgen/). Those individuals were previously selected to cover the whole Moroccan territory together with a wide range of climatic conditions (Benjelloun, 2015). For a complete description of the sequencing protocol, see (Alberto *et al., In prep*).

SVs Calling and filtering

Read mapping

Paired-end reads were mapped to the sheep reference genome (build Oar_v4.0 - GenBank assembly accession: GCA_000298735.2) and to the goat reference genome (build ARS_1 - GenBank assembly accession: GCA_000317765.2) for sheep and goats respectively using BWA-MEM algorithm (Li and Durbin, 2009). The BAM file produced for each individual was sorted using Picard SortSam and improved using Picard MarkDuplicates (http://picard.sourceforge.net).

SVs calling

Structural variations (SVs) where called independently for each individuals using three different methods: BAdabouM (Partie 3 – chapitre 1 : Article 3 - BAdabouM: a structural variation discovery tool), DELLY (Rausch *et al.*, 2012) and Breakdancer (Chen *et al.*, 2009). All tree methods were run using default parameter, except for the reads mapping quality, where a quality of 60 was required.

SVs clustering and filtering

For each individual, SVs called by two methods were considered as identical based on their reciprocal overlap. As methods do not detect breakpoints with the same accuracy, a threshold was set for congruent overlap. For inversions, deletions, duplications and intra-chromosomal translocations, a criterion of SV overlap greater than 50% was used. For insertion, as breakpoints may be narrow (only one bp), the fact of overlapping was considered as sufficient to be considered as the same event. For intra and inter-chromosomal translocation, breakpoints had to be within a 1kb window to be considered as the same event.

To consider SVs called for multiple individuals as homologous, the same strategy was applied (based on reciprocal overlap). The threshold on the reciprocal overlap was set to 70% for inversions, deletions, duplications and inter and intra-chromosomal translocations. For insertions, the fact of overlapping was considered as sufficient to be considered as the same event. Breakpoints of inter-chromosomal translocation on both chromosomes had to be within a 1kb size window to be considered as the same event.

For further analysis, only SVs called by at least two methods and presents in at least two individuals but not fixed where kept, except for insertions: insertions called only with

BAdabouM were kept as this software apply a high stringency criterions for insertion calling).
This step allowed to identify insertions, deletions, inversions, CNVs and inter or intra-chromosomal translocations. For further analysis and for each individual, each SV was considered as a dominant presence absence marker.

Environmental data & Biome reconstruction

6 climatic variables from 67 (previously extracted from the WorldClim dataset (Hijmans et al. 2005 from the individuals sampling locations) were considered, as it was done in previous study (Benjelloun, 2015). Those climatic variables are: April average monthly precipitation (Prec_4), July average monthly mean temperature (Tmean_7), isothermality (Bio_3), temperature annual range (Bio_7), mean temperature of wettest quarter (Bio_8) and precipitation Seasonality (Bio_15).
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 2 DEM-derived variables were computed with SAGA GIS (Conrad *et al.*, 2015): altitude (Alti), and solar radiation in June (TI2112).

Using all individuals' data for sheep and goats, a PCA was run using ade4 package (Dray *et al.*, 2017) to resume environmental data variability into uncorrelated principal components (PC). Those PC were then used as synthetic environmental variables for the rest of the analysis.

Correlation between SVs and environmental variables

To detect SVs linked with environment variables, correlative approaches was use to check, for each variant, if allelic distribution fits over the environmental gradient. Two models were selected; a linear model using latent factor mixed models (implemented in LFMM (Frichot *et al.*, 2013)) and a logistic model (implemented in Sambada (Stucki *et al.*, 2017)).

LFMM
Neutral genetic structure based on SVs was inferred using sNMF (Frichot *et al.*, 2014). sNMF was run for ancestral populations (K) from 1 to 10, with 20000 iteration and 10 repetition. K=1 was considered as the best K based on minimal cross entropy criterion (figure S1) for both sheep and goats.
For each of the two first PC, LFMM software was run 5 times with 100000 iterations and 10000 burn-in steps for K= 1. P-values from z-scores were re-calibrate using inflation factor (see (Frichot and François, 2015) for details).

SamBada
We performed logistic regressions between SVs distribution and environmental variables to estimate the probability that an individual carries a specific genetic marker given each of the two first PC of the PCA using SamBada (Stucki *et al.*, 2017) with

defaults parameters. Wald score for each SV was then converted to p-values using chi-squared distribution as a prior (Stucki *et al.*, 2017).

Output comparisons and Functional annotation of SV

To compare output from LFMM and SamBada, and detect candidate loci common to the two methods, we used a FDR approach by converting all p-values into q-values. This step was done using the qvalue package of R (Ihaka and Gentleman, 1996). Correlation between q-values outputted by the two programs was tested using Kendall test (implemented in Kendall package for R (McLeod, 2005)).

A SV was considerate as a candidate locus if the q-values of the two methods were lower than 10^{-2}. Each SV was then relocated on the annotated reference genome. If an SV overlapped or was close (less than 500 bp) from an annotated gene, we considered this gene to be impacted by the SV. Functional annotation of those genes was then set based on bibliographic research of already known function of the gene in sheep and goats or other livestock species.

Results

In this study, a total of 160 goats and 159 sheep WGS where used to call genomic structural variations. After filtering, 57775 polymorphic SVs were kept with a mean number of 12927 SVs per sheep and 18172 SVs for goats with a mean number of 3444 SVs per individuals.
For Sheep, deletions represent 61% of the total of SVs, insertions 33% and inversions 10%. For goats, deletions represent 81% of the total, insertions 12% and inversions 4% (Composition in SVs is resumed in table 1).

Table 1: Structural variation types and their distributions for both sheep and goats.

	Sheep		Goat	
SV class	No. sites	Mean number / ind (sd)	No. sites	Mean number / ind (sd)
Total	57775	12927.4 (829.8)	18172	3444.1 (201.8)
Deletions	35232	8878.2 (579.4)	14849	2911.6 (176.5)
Insertions	19266	3234.5 (440.1)	2229	298.0 (36.6)
Inversions	2085	533.7 (36.2)	839	173.1 (16.5)
CNV	325	49.5 (5.4)	78	15.3 (2.9)
ITX	90	55.2 (3.8)	6	1.7 (1.0)
CTX	777	176.2 (31.3)	171	44.3 (9.9)

Biome reconstruction

The three first axis of the PCA summarises respectively 35.6%, 31.0% and 14.7% of the initial environmental variable set. The first PC opposes the precipitation seasonality (Bio_15) and isothermality (Bio_3) to the altitude (Alti) and temperature annual range (Bio_7). The second PC mainly summarise variability of April average monthly precipitation (Prec_4) with July average monthly mean temperature (Tmean_7) and mean temperature of wettest quarter (Bio_8) (Figure 1). Third axis mainly resume variability of solar radiation in June (TI2112), and was not analysed further. The first axis opposes the west side with oceanic climate characterised by mild winter and cool summers, with high altitude climate of the Eastern and Atlas part of the country characterised by harsh winter and warm summers. The second axis opposes the Rif and Mediterranean coast characterised by high precipitation seasonality and steady temperature, with arid climates of the southern desert characterised by low precipitations and cold winters and warm summers (Figure 1) (Zereini and Hötzl, 2008).

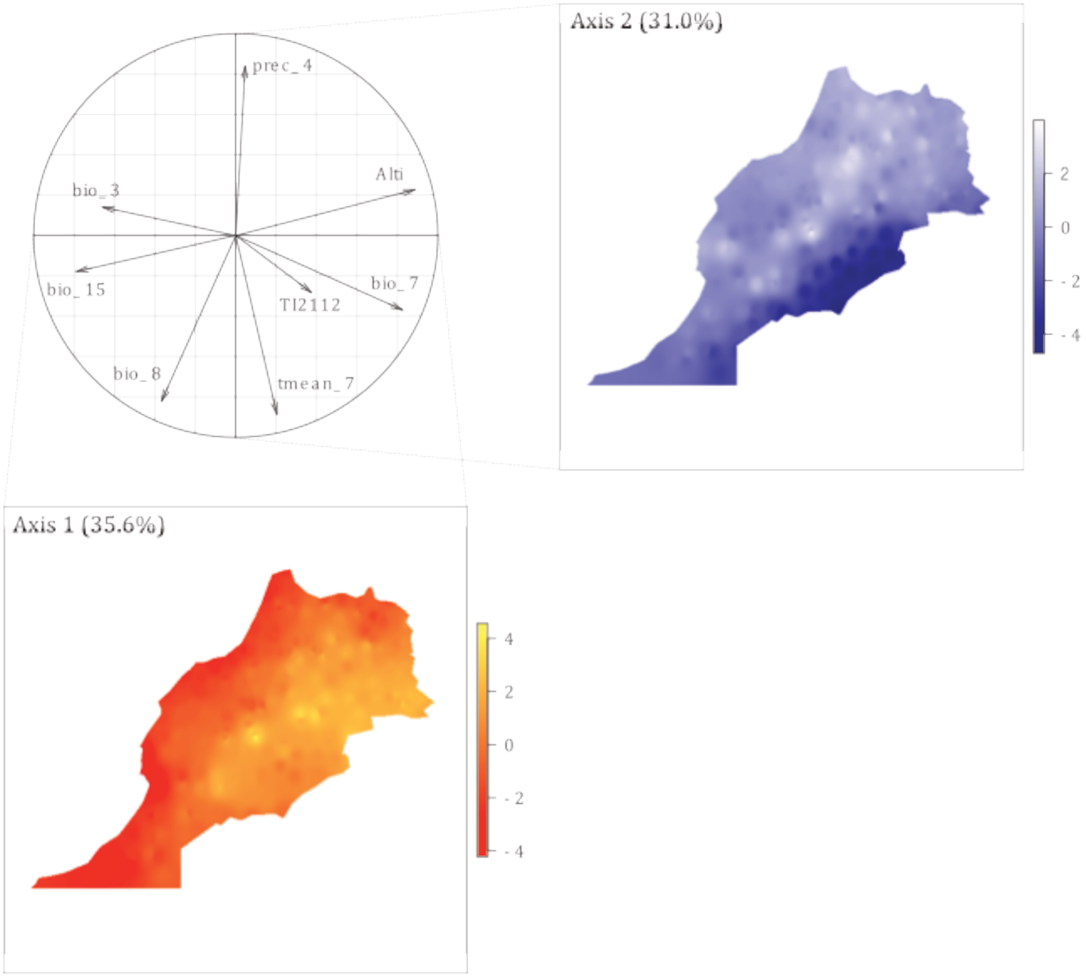


Figure 1: Environmental PCA results. Top-left figure: contribution of different environmental variables to the two first axis of the PCA. Bottom-left and top-right figures: Geographical extrapolation of the two first axis of the PCA.

Selection signatures

No population structure was detected for both sheep and goat, i.e. using sNMF, the lowest Cross-Entropy value was for K=1 (figure S1).
To test association between between LFMM and Sambada output, Kendall tests was run. P-values were all highly significant (p-values =< 2.22e-16 for all comparisons).
Using 10^{-2} as false discovery rate threshold, multiple SVs were found to be correlated with the environment. Hence, when considering the SVs predicted by the two methods, 45 SVs for sheep (32 deletions, 11 insertions, 1 inter chromosomal translocation (CTX) and 1 inversion) and 5 SVs for goats (3 deletions and 1 CTX) were correlated with the first PC and 85 (44 insertions, 38 deletions, 2 inversions and 1 CNV) and 30 SVs (25 deletions, 2 insertions, 2 CNV and 1 inversion) with the second PC respectively. Among those SVs, 52 (42%) were found associated with a gene in sheep and 18 (51%) for goats (Figure 2, Table 2, full list in Table S1).

		Nb outiliers LFMM	Nb outliers Sambada	Common	Associated gene
Ovis	PC1	890	45	45	16
	PC2	1275	105	85	36
Capra	PC1	357	5	5	2
	PC2	405	34	30	27

Table 2: Significant structural variations according to climatic principal component for both sheep and goats.

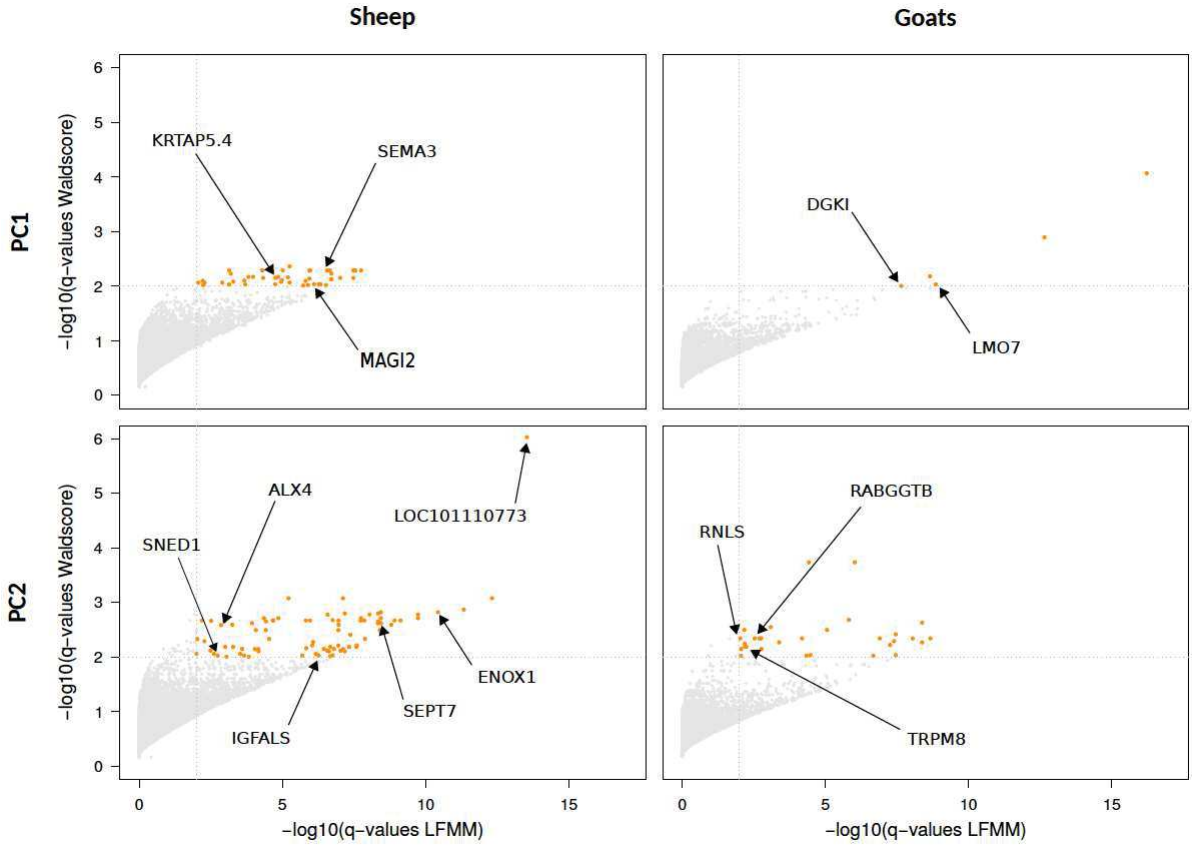


Figure 2: Correlation between sheep and goats genomic structural variations and the two first axis of the PCA, estimated by two methods, Sambada (Waldscore) and LFMM. Orange dots are structural variations with q-value higher than 10^{-2} with both methods.

Discussion

Despite the known importance of structural variations in local adaptation of domestic animals (Paudel *et al.*, 2013; Bickhart *et al.*, 2012), large-scale study remains rare.

Biome reconstruction.

Principal Component Analysis (PCA) is a multivariate statistical technique that converts a set of correlated variables into a set of orthogonal and uncorrelated axes called principal components. Applied on environmental data, PCA are often used to summarize bioclimatic variables and to reconstruct bioclimatic niches (James and McCulloch, 1990). On Moroccan data, environmental PCA allows to have a synthetic view of the different biomes small ruminants are confronted to. Thus, the two first axis of the PCA represent the wide majority of climatic condition that can be found in Morocco.
The first axis opposes the west side with oceanic climate and alpine climate of the Atlas part of the country, while the second axis opposes the Rif and Mediterranean coast with arid climates of the southern desert. These biomes reconstruction cover the wide range of climatic condition across the Moroccan territory and confirms it's potential to be a good proxy of bioclimatic variation over the whole territory for studying local adaptation.

Genomic structural variations overview

Large scale prediction of structural variations based on WGS of Moroccan small ruminants allowed to identify a large number of non documented variations, with 57775 polymorphic SVs for sheep and 18172 SVs for goats. The fact that less SVs were detectable in goat than in sheep is coherent with SNPs calling data based on the same WGS (as sheep showed 7 million more SNPs than goats and a higher nucleotide diversity) (Benjelloun, 2015), and with the fact that goats harboured already less SVs than sheep at the time of domestication (see Part 3 – Chapter 2 : Article 4 - Exploring the role of genomic structural variations during small ruminant's domestication).

Structural variation compositions differ slightly between both species, with proportionally, less insertion and more deletion on sheep genomes than in goats' genomes (respectively 61% of insertions and 33% of deletion in sheep compared to 81% and 12% in goat). Given that those proportions are the same when studying SVs in the context of domestication for both species (see see Part 3 – Chapter 2 : Article 4 - Exploring the role of genomic structural variations during small ruminant's domestication), we can exclude a post domestication process to explain those differences. This difference is probably rather due to technical limitations to call SVs. Indeed, it's important to keep in mind that calling SVs relies greatly on the quality of the reference genome used to realign WGS. Thus, quality differences between sheep and goat assembly may lead to observed differences, with false positives or true negative in one or both species (Bickhart and Liu, 2014).

Beside those technical limitations, no genetic structure was observable based on those structural variations, as previously reported based on SNPs data (Benjelloun, 2015). Thus, local breeds of the same species are not highly divergent from each other. This may be explained by moderate intensity of selection and gene flow during breeds' formation.

Genomic structural variations and selection signatures

When considering sampling covering environmental gradients, population genomic approaches contrasting two sets of individuals taken from opposed environmental conditions seem deprecated. As a consequence, we favoured, given the data we used, correlative methods to detect adaptive loci. Indeed, these methods seem well more suited to catch signature of selection along a gradient.

Such methods correlate the distribution of a variant and an environmental variable. Generally, they have a huge false positives rate, mainly due to demographic history or complexe selective scenarii (de Villemereuil *et al.*, 2014). We address those issues by different ways. First of all, we controlled for a possible population structure that could lead to artifactual predictions due to confounding effects. Based on both SNP data and our SVs calling, no population structure could be inferred. We then applied two different correlative methods based on two different models and kept only SVs predicted by both methods thus leading to a very conservative procedure. Moreover, we used restrictive thresholds (10^{-2}) to increase the confidence in the results, involving a decrease of power and a high level of false negatives.

As explained above, the absence of remarkable genetic structure within goat or sheep populations is a good point for limiting confounding effect (Novembre and Di Rienzo, 2009) Several genomic structural variations are significantly linked with environment despite conservative thresholds (Figure 2). Among the 130 SVs for sheep and the 35 SVs for goats detected as correlated with one of the two principal components, a large part (60% for sheep and 48.6% for goats) of SVs does not seem to impact genes directly. This is not surprising regarding the fact that these kind of mutations could disrupt the structure of the genes and then being possibly too disturbing for the function of the gene-product and also by the fact that non-coding regions can be important for gene regulation (Carroll *et al.*, 2008).

In addition, we detected less SVs correlated with environments for goats than for sheep that is consistent with the fact that less SVs were called in goats. Moreover, this also reflects the lower number of SVs that were captured for goats than for sheep at the time of domestication (see Part 3 – Chapter 2 : Article 4 - Exploring the role of genomic structural variations during small ruminant's domestication). Proportionally, with 0.23% for sheep and 0.19% for goats, only a small part of whole set of SVs seems implied in the genetic basis of local adaptions.

However, several structural variations associated with environmental variables are localised within genes. 14 SVs are associated with the first PC and possibly impact genes

in sheep. A first example is a large deletion (including two exons) within the *KRPAT5.4* (*Keratin associated protein 5.4*) gene is correlated with this variable. This gene is expressed in the hair cuticle of wool sheep (Jenkins and Powell, 1994), and may play a role during cuticule differentiation in sheep possibly affecting wool quantity or quality (Zhao *et al.*, 2017). Another example is an intronic deletion within the MAGI2 (Membrane Associated Guanylate Kinase) gene is also correlated with the first environmental variable. This gene have already been linked with feeding efficiency in cattle (Hou, Bickhart, *et al.*, 2012) and altitude in Tibetan pigs (Ai *et al.*, 2014). Considering the climatic conditions resumed within the first axis, those genetic information seems to highlight possible morphological and physiological adaptation to rude climatic condition in Atlas regions, with modification of wool quality or quantity along the gradient and selection for a better feeding efficiency in scanty resources area.

For goats, two SVs, linked with this first variable impact possibly genes. The *DGKI* (diacylglycerol kinase iota) gene may be impacted by an intronic deletion linked with the first variable. *DGKI* have already been linked with *Trypanosoma congolense* resistance in cattle (Noyes *et al.*, 2011) a parasite absent from Morocco, where other trypanosomes such as *Trypanosoma evansi* are presents (Desquesnes *et al.*, 2013). Moreover, prevalence of trypanosomosis was negatively correlated with altitude in Ethiopia (Duguma *et al.*, 2015). Hence, it's possible that exposition to trypanosome in low altitude regions of Morocco selected resistant allele in goats.

Thirty-six SVs associated with genes were found to be correlated with the second PC in sheep. Among them, the *SNED1* (sushi, nidogen and EGF-like domains 1) gene affected by an intronic deletion and the *ENOX1* (Ecto NOX Disulfide-thiol Exchanger 1) gene affected by an intronic deletion, have been associated with feed efficiency respectively in cattle and pig (N. V. Serão *et al.*, 2013; Reyer *et al.*, 2017), while several variants seems to impact genes previously linked with morphology (an insertion in *SEPT7* gene (McClure *et al.*, 2010)), milk (an intronic deletion in *AGBL4* gene (Flori *et al.*, 2009), and an intronic deletion in *GNA14* gene (Saowaphak *et al.*, 2017)) and beef (a deletion in *AGBL4* gene (Flori *et al.*, 2009) and an exonic insertion in *IGFALS* gene (Guo *et al.*, 2012)) production. Interestingly, this axis contract climate but also oppose the Timahdite breed in Mediteranean region with the D'man breed in Saharan region. Productivity and morphology of those two breeds are highly different, with a more productive breed in the northern part of the country (Timahdite) and a more rustic breed (D'man) in the southern part (Boujenane, 2005). If those breeds are not enough differentiated to be captured by structures analysis, structural variations highlighted here suggest that those differences may be genetic basis, with selection impacting production genes in one hand, and adaptation to rude climatic condition on the other.

Similarly, for goats, the second PC does not only contrast environments but also breeds, with the Northern breed on the north part of the country and the Draa from southern Morocco (Benjelloun *et al.*, 2015). Among the genes impacted by SVs that are correlated with the second variable in goat, we detect genes linked with feed efficiency (an intronic CNV in *RNLS* gene (Serão *et al.*, 2013)**),** which is coherent with the Draa's ability to maintain an unchanged food intake in rude climatic conditions (Hossaini-Hilali and

Mouslih, 2002). We also detected an intronic deletion within the *TRPM8* (transient receptor potential cation channel subfamily M member 8) gene, involved in regulation of body temperature at low temperatures. This gene, previously detected as selected in worldwide sheep populations (Fariello *et al.*, 2014), may play a role in goat's adaptation to daily and annual climatic variations of Saharan desert.
RABGGTB (Rab geranylgeranyltransferase beta subunit) gene, linked with milk productivity in cattle (Li *et al.*, 2010), seems to be impacted by an intronic deletion contrasting goats along the gradient. As it has been postulated that the gene is linked with milk productivity, of one of the two alleles, may have been selected in the productive Northern breed (Benjelloun *et al.*, 2015).

The survey of SVs possibly linked with the two synthetic variables highlight the potential role of structural variations in local adaptation. Our results tend to show that environmental conditions encountered by small ruminants in Morocco selected specific alleles including SV, and the comparisons with selection signatures based on SNPs should allow us to evaluate the convergence and complementarity between both types of markers.

Structural Variations and Single Nucleotide Polymorphism comparison

Considering that SVs are probably in linkage disequilibrium with SNPs around them, we should be able to detect the same signal with both markers.
Our work is based on synthetic environmental variables while previous work based on SNPs considered independently each environmental variable (Benjelloun, 2015), as a consequence, the comparison between both analyses needs to be carefully conducted. Thus, considering that our environmental variable resume information contained in initial environmental data, comparison between SVs correlated with principal component and SNPs detected with environment variables that compose the PC should be possible.

The first PC oppose oceanic coast with Atlas Mountains. Genes possibly impacted by structural variations differentially distributed along this gradient suggest a possible morphological and physiological adaptation to rude climatic condition in Atlas regions for sheep, with modification of wool quality or quantity, and selection for a better feeding efficiency in scanty resources area. Such adaptation to altitude is also visible in signature of selection based on SNPs, which highlight genes possibly impacting resistance to hypoxia at high altitude variable(Benjelloun, 2015).
Based on the same environmental gradient, SVs analysis spotted a possible allele selected for the better resistance to trypanosomes at lower altitude in goats. SNPs tends to point out selection for a higher resistance against pathogens in goats raised under hot and wet conditions, corresponding to the climatic condition of low altitude regions on the gradient (Benjelloun, 2015).

The second PC oppose Saharan desert with Mediterranean coast. This gradient clearly differentiates locally adapted breeds for both sheep and goats. For both species,

northern breeds, raised under favourable climate are more productive than southern breeds that seem more rustics, raided under rude climate.
Genes affected by SVs highlighted here suggest that differences observed between breeds may have genetic basis, with selection impacting genes linked with productivity in one hand and genes related with rusticity, liked with feed efficiency or temperature regulations in the other. Beside, it's interesting to note the convergent selection for better feeding efficiency in sheep and goat exposed to rude environment in Saharan desert. On one hand, SNPs signals pinpoint selection for better thermoregulation in goats raised in highly fluctuating temperature environment. On the other, a SVs may affect regulation of body temperature at low temperatures. Thus, both markers tends to illustrate goat's adaptation to variation of climatic condition of Saharan desert.

It is however important to notice that if adaptive mechanism pinpointed by SNPs and those highlighted by SVs are similar to some extent, the targeted genes are not the same, except for one region (described bellow). This non-convergence between signals from the two types of markers may have diverse origins. First, both studies does not uses the same methods to link markers with environmental data, witch may introduce confusion. Indeed, SNP based study uses both population genomic approaches and correlation between allele frequencies and environmental factors (Benjelloun, 2015) while we only use the second approach here. Moreover, the use of synthetic variables instead of environmental variables surely induces some differences. Another source of non-convergence between both methods may be the low linkage disequilibrium of sheep and goats genomes (Benjelloun, 2015). Thus, selective sweeps may be narrow around a SV, and not detected by the previous analysis. Moreover, SVs detected in this study may also be false positives, particularly due to the fact that SVs are encoded as dominant markers, with may skew our detection method. They also may be in linkage disequilibrium with other variations such as SNPs affecting nearby gene, not recorded by previous analysis.

As mentioned above, both methods (using SVs and SNPs) point a region on sheep's chromosome 10 encoding for two annotated features: On one hand, SNPs around and within the RXFP2 gene seems associated with Prec_4 variable, on the other, the deletion of the nearby LOC101110773 element is correlated with the second synthetic variable (Figure 3). The presence of the deleted element was previously reported as causal for polled sheep (Pailhoux *et al.*, 2001; Wiedemar and Drögemüller, 2015). Thus, both variations highlight the same genetic regions, and in this case SNPs are passives witness of the causal variant. In addition, this example illustrates the importance of taking in account all genomic variations to understand genetic basis of adaptation, particularly while attending to identify causal mutation.

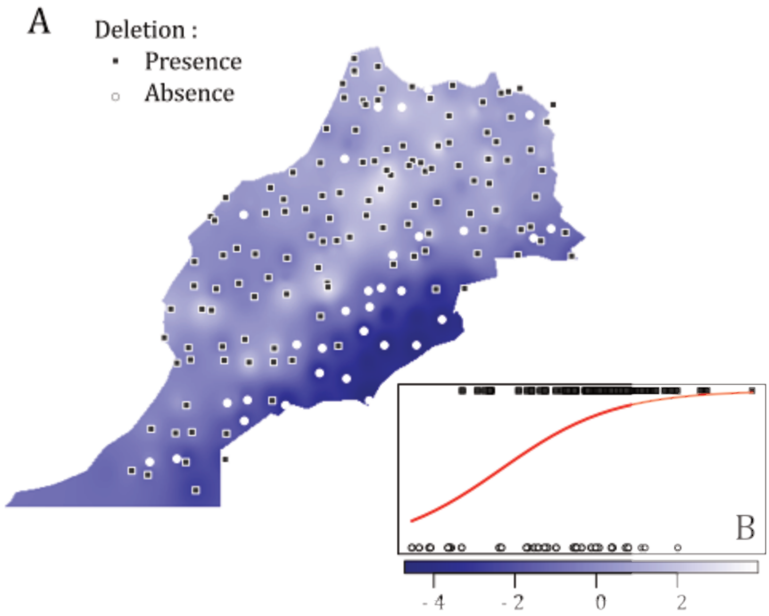


Figure 3: Geographical distribution of the deletion located near the RXFP2 gene (A) and it's repartition along the environmental gradient (B).

Besides illustrating convergence between both analysis and both markers, the case of the RXFP2 gene is interesting and illustrate the possible impact of confounding effects. As previously noticed, the second principal component, in addition to contrast Mediterranean and arid environments, also opposes D'man breed distribution in the desert with other Moroccan sheep breeds. This breed is the only Moroccan breed with both polled male and female (Boujenane, 2005). Thus, it may be wrong to conclude that the deletion only linked with the environment gradient as it can be also due to breeds' distributions. Such results illustrate the importance of tacking precautions while analysing correlation between genotype and environmental variables or phenotypes, to minimize co-variables interferences as much as possible. In this case, it's not the locus that is a false positive, but the correlation itself (Rellstab *et al.*, 2015).

Anyway, we cannot exclude a pleiotropic effect of RXFP2 gene, where the polled character would be one of the consequences on the mutation. The superposition of environmental gradients and breeds repartition may add confounding effect to the study and bring false positives. Nevertheless, the differentiation between breeds is weak and breeds are locally adapted to their environment (Benjelloun, 2015), so trying to correct for breed effect may be counterproductive, by filtering adaptive loci.

It's still hard to speculate about real impact of SVs without functional validation. It would be mandatory to study impact of SVs on the expression of proteins and their activity to validate or invalidate such hypothesis. Moreover, it seems important to notice that all types of genetics variations may have impact on individual fitness and integrating all variations should allow a better understanding of the genetic basis of adaptation.

Conclusion

SVs affect a large part of small ruminant genomes and may play an important role for local adaptation. Even if further work is needed to validate the role of SVs detected here, tacking SVs into account seem mandatory to have a global vision of the genetic bases of local adaptation.

Partie 3 : Discussion

Ce chapitre s'intéresse à la détection des SVs dans des données de génomes complets de petits ruminants. Sans informations préalables sur les variants à rechercher, cet axe se décompose en deux parties. La première est méthodologique et concerne les moyens à mettre en œuvre pour les détecter, alors que la seconde s'intéresse directement aux variants et à leurs conséquences sur les organismes.

Au niveau méthodologique, ce travail permet de voir qu'il est possible de détecter des variations de structure dans l'ADN, mais que de nombreuses difficultés ne permettent pas de détecter l'ensemble des SVs dans les données de reséquençage telles qu'elles nous étaient fournies.De nombreuses méthodes existent pour détecter les SVs (Lin *et al.*, 2014), mais aucune de ces méthodes ne permet de détecter tous les SVs (Zhao *et al.*, 2013). En outre, toutes ces méthodes présentent un plus ou moins fort taux de faux positifs qu'il est nécessaire de prendre en compte (Layer *et al.*, 2014). Il faut alors combiner ces méthodes pour capter un maximum de signaux et ainsi minimiser le taux de faux positifs (Pabinger *et al.*, 2014).

S'il est encore compliqué de détecter des variants, inférer le génotype l'est encore plus (Handsaker *et al.*, 2011). Pour résoudre ce problème, une solution pourrait être de séparer les phases de découverte et de génotypage des individus comme c'est le cas pour l'appel des SNPs. La découverte peut se faire rapidement, en intégrant indépendamment les différents signaux pour essayer d'identifier avec un maximum de précision un maximum de zones suspectes ainsi que d'essayer de préciser quels types de variants il s'agit. Dans un second temps, le génotypage des individus pourrait se faire *via* des stratégies adaptées à chaque type de variant, de façon à génotyper efficacement les individus (comme développé dans la partie 2 : Variants structuraux – Approche « variants candidats »).
Malgré ces limites, il est possible de détecter des variants structuraux dans des génomes complets et de s'intéresser à l'histoire des espèces sans *a priori*.

Ainsi, les méthodes utilisées dans les travaux présentés ci dessus permettent de détecter un grand nombre de variants structuraux dans les génomes des petits ruminants.
Ces variants témoignent de l'histoire des espèces. Ils permettent de discriminer les animaux domestiques des animaux sauvages pour les deux espèces étudiées ici (Partie 3 - Chapitre 2 : Article 4 - Exploring the role of genomic structural variations during small ruminant's domestication), montrant bien la séparation génétique entre ces deux populations. Ces résultats viennent confirmer la capacité des SVs à être des marqueurs neutres de l'histoire évolutive (Conrad and Hurles, 2007).
Certains variants, discriminant très nettement les sauvages des domestiques, sont de bons candidats dans l'étude des variants possiblement sélectionnés lors de la domestication. La localisation de ces variants montre que si certains sont dans des gènes, une grande partie est localisée dans des régions non géniques. Si ces régions sont non codantes, elles ne sont pas pour autant non fonctionnelles (Spielmann and Mundlos, 2016), et les améliorations des annotations des génomes de référence permettront de

mieux comprendre le rôle de ces variants. Certains variants possiblement impliqués lors de la domestication sont localisés dans des gènes dont certains déjà identifiés comme impactés dans le processus de domestication, chez la chèvre, le mouton ou d'autres espèces domestiques.

Parmi les variants identifiés, certains semblent aussi jouer un rôle dans l'adaptation des petits ruminants à leurs environnements. L'approche développée dans les travaux présentés ici se base sur l'utilisation de variables climatiques synthétiques correspondant aux premiers axes d'une ACP, permettant ainsi de résumer efficacement la variabilité climatique en grands biomes climatiques, qui ont plus de sens qu'une simple variable parfois difficile à interpréter. Les axes de l'ACP représentent donc des variables climatiques synthétiques, qui opposent les milieux favorables sur les côtes, avec des climats océaniques et méditerranéens, opposés aux climats plus rudes de l'Atlas ou du désert saharien. La distribution de certains variants semble en corrélation avec ces variables climatiques. Là aussi, un grand nombre de variants ne semble pas impacter de gènes. En se focalisant sur les gènes impactés, certains ont déjà été décrits chez des espèces domestiques comme importants pour des fonctions telles que la productivité, ou la robustesse. Ces résultats peuvent s'expliquer à l'aune des biomes représentés. Ces résultats sont cohérents avec la théorie de l'allocation des ressources, qui veut que les organismes répondent à la sélection jusqu'à ce que leur fitness ne puisse plus s'améliorer. C'est alors le moment où les organismes utilisent toutes les ressources disponibles de la façon la plus efficace (Beilharz *et al.*, 1993). En effet, si l'on considère plus faciles les conditions de vie dans les climats favorables, alors plus de ressources sont disponibles pour être allouées à la production de lait, de viande ou de laine. En revanche, quand le climat est moins favorable, ce sont les caractéristiques plus rustiques qui sont sélectionnées et une part plus grande des ressources est allouée à la survie et à l'efficacité de cette survie.
Dans le cas des petits ruminants au Maroc, les individus issus de conditions difficiles allouent une plus grande part des ressources limites à leur survie au travers d'une rusticité plus importante, alors que les individus en conditions plus faciles peuvent allouer plus de ressources à la production.

Que ce soit dans l'étude des SVs dans le contexte de domestication ou d'adaptation, il n'est pas facile de dire si les mutations identifiées sont sélectionnées ou si elles sont emportées par déséquilibre de liaison (effet auto-stop ou 'hitchhicking') avec la mutation sélectionnée. Ces résultats illustrent tout de même que les SVs ont pu jouer un rôle lors de la domestication. Ce rôle est d'autant plus facile à imaginer que si la fréquence de mutation reste plus rare, le potentiel de perturbation d'un évènement donné est plus important pour les SVs que pour les SNPs (Pang *et al.*, 2010). En effet, les mutations ponctuelles touchent une seule base. Si certaines de ces mutations modifient de façon certaine le fonctionnement des organismes, en affectant la séquence codante (et en modifiant la séquence protéique) ou en touchant une zone régulatrice, elles peuvent aussi être sans impact même en affectant ces zones, du fait de redondance du code génétique et les mutations synonymes. D'un autre côté, l'introduction de ce manuscrit montre bien le potentiel perturbateur des SVs et il semble que cette importance soit retrouvée chez les petits ruminants.

DISCUSSION GÉNÉRALE

Les travaux présentés dans ce manuscrit s'attachent à approfondir notre connaissance du rôle des variants structuraux génomiques dans l'histoire évolutive, en ciblant les mécanismes sélectifs liés à la domestication et l'adaptation des petits ruminants à leurs environnements. Ces travaux se sont donc organisés en deux grands axes. Le premier, basé sur l'étude des variants présents dans la bibliographie, permet de cibler précisément un variant, puis d'étoffer nos connaissances sur ce cas particulier. Le second axe vise, à partir d'un ensemble de donnée issues du séquençage de génomes complets, à rechercher les variants, sans connaissances *a priori*, puis de lier ceux-ci à la domestication et à l'adaptation des petits ruminants.

Au travers de ce travail, deux aspects imbriqués l'un dans l'autre, et imbriqués au sein des axes de recherche semblent se dégager. Le premier aspect est méthodologique. Il inclut l'ensemble des points pratiques et techniques concernant la détection des variants structuraux génomiques dans des données de séquençage de génomes complets. Le second aspect est quant à lui biologique et traite de l'impact des variants structuraux génomiques chez les petits ruminants, en s'intéressant notamment au rôle joué par ceux-ci au cours des processus de domestication et d'adaptation.

Cette discussion va donc essayer de faire la synthèse de ces deux aspects du travail présenté dans ce manuscrit. Cette synthèse devra, dans un second temps, permettre de répondre à la question posée initialement, à savoir quel est l'apport des variants structuraux dans la compréhension des mécanismes sous-jacents à la domestication des petits ruminants.

Variants structuraux génomiques : Aspects méthodologiques

SVs et reséquençage de génomes complets.

Au travers des travaux présentés dans ce manuscrit, nous avons pu détecter des variants structuraux génomiques dans des données de reséquençage de génomes complets. Ce travail a aussi permis de montrer que la difficulté de cette détection est dépendante des variants. Elle peut être simple, dans le cas où il suffit d'étudier l'évolution de la couverture le long du génome (cas de la globine, Partie 2 – Chapitre 2 : β-Globine ovine, un potentiel rôle adaptatif). Elle peut aussi être plus compliquée, notamment dans le cas où la zone est dupliquée dans le génome de référence, biaisant de fait l'alignement des lecture et impliquant de réaligner ces lectures sur la région non dupliquée pour évaluer correctement les génotypes des individus (cas du gène ASIP, Partie 2 – Chapitre 2 : Domestication et convergence évolutive, le cas du gène ASIP). Plus compliquée aussi lorsqu'elle nécessite la mise en place des stratégies en plusieurs étapes pour identifier les SVs (inventaire des enJSRV, Partie 2 – Chapitre 1 : Article 2 - Old origin of a protective endogenous retrovirus (enJSRV) in the Ovis genus). Enfin, cette détection peut être complexe, quand les régions qui hébergent les variants structuraux ne sont pas assemblées dans le génome de référence et que la détection du variant nécessite l'usage de stratégies indirectes (comme l'estimation du nombre de copies mutées d'enJSRV-

6q13, Partie 2 – Chapitre 1 : Article 2 - Old origin of a protective endogenous retrovirus (enJSRV) in the Ovis genus).

Dans la seconde partie de ce manuscrit, le passage en revue des différents types de variants structuraux, des signaux laissés par ceux-ci et que l'on peut détecter lors du réalignement des génomes complets, ainsi que l'évaluation d'un certain nombre de logiciels se proposant de les détecter, ont permis de voir que la détection des SVs à l'échelle du génome complet est un problème encore d'actualité. Ce problème est notamment du au fait que les méthodes disponibles sont toutes soumises à de forts taux de faux positifs (avec de nombreux SVs détectés par une seule méthode, voir Partie 3 – Chapitre 1 : Détection des SVs dans les données de reséquençage de génomes complets) (Layer *et al.*, 2014) et de faux négatifs (SVs non détectés) (Pabinger *et al.*, 2014). Il semble donc nécessaire de prendre en compte les variants détectés par plusieurs de ces méthodes pour intégrer une large gamme de signaux et ainsi réduire le taux de faux négatifs.
En outre, cette utilisation conjointe de plusieurs logiciels doit permettre de cibler l'intersection des SVs trouvés par ces méthodes, et ainsi contrôler le risque de faux positifs.

Limites et complémentarité des approches "candidat" et "sans a priori"

Les travaux présentés ici sont donc basés sur deux approches aux granulométries différentes, tant par le nombre de variants ciblés que par la précision avec laquelle ces SVs étaient étudiés. La première approche, axée sur l'étude de variants candidats permet de préciser les informations connues sur ces variants (leurs bornes, le génotype des individus, ect …) avec une grande précision et ainsi de tester efficacement une hypothèse sur un variant donné. La seconde approche se base quant à elle sur une recherche sans *a priori* des SVs, puis cherche à les lier aux questions d'intérêt. Cette approche permet d'avoir une vision globale mais relativement grossière (types de SVs pas toujours bien définis, points de cassure imprécis suite à la fusion des résultats) bien que porteuse d'informations biologiques, puisqu'elle permet de différencier les individus sauvages des individus domestiques (voir la partie concernant le logiciel BadabouM, voir Partie 3 – Chapitre 1 : Article 3 - BAdabouM: a structural variation discovery tool) et d'identifier des SVs possiblement ciblés lors de la domestication ou par la sélection pour l'adaptation.

Cette double granulométrie permet donc de couvrir un large spectre de SVs mais ces deux méthodes présentent cependant chacune des limites que l'autre approche ne peut compenser. Dans le cas des variants candidats, leur nombre est limité et l'approche se révèle potentiellement très chronophage puisqu'elle demande de développer une stratégie propre à chaque variant. Dans le cas de l'approche sans *a priori*, la granulométrie actuelle est trop grossière (inférence de caractères présence/absence et non de génotypes, breakpoints plus ou moins précis), et ne permet pas de rentrer dans les détails pour l'ensemble des variants identifiés.

Limites de la détection des SVs

L'inventaire des SVs présentés dans ce manuscrit n'est cependant pas exhaustif. En effet, la seconde partie de ce manuscrit montre que la détection des variants est dépendante de plusieurs paramètres tels que la qualité de l'assemblage des génomes de référence ainsi que des logiciels utilisés.
Dès lors, faire des hypothèses sur les caractéristiques des SVs (telles que leurs répartitions dans les génomes, l'enrichissement en certains types, ou toute autre caractéristique de ces variants) semble biaisé et il serait difficile de dissocier ce qui est lié aux variants eux-mêmes ou à la stratégie employée pour les détecter. Le choix a donc été fait de ne pas s'intéresser à ces caractéristiques. Cependant, et malgré cet inventaire incomplet, les variants retenus pour les analyses de génomes complets l'ont été sur la base de critères exigeants, réduisant la liste des SVs possibles à une liste restreinte de SVs de confiance, permettant l'analyse de leurs conséquences biologiques.

Variants structuraux génomiques : Aspects biologiques

SVs et évolution des petits ruminants.

Les résultats présentés dans la première partie de ce manuscrit ont permis de montrer, au travers de trois exemples, que des SVs ont joué un rôle lors de l'évolution du genre *Ovis* (JSRV), de leur domestication (ASIP), et potentiellement lors de leur adaptation aux climats arides (β-globine). Ces exemples fouillés mettent en avant, pour chacun d'entre eux, les mécanismes impactés par le variant (surexpression du gène, modification du transcrit …) ainsi que leurs conséquences sur les individus (résistance au virus exogène, impact sur la pigmentation ….).
Dans la seconde partie, l'étude de ces variants structuraux sur l'ensemble du génome a permis de montrer qu'il est possible de lier certains SVs détectés sans *a priori* avec des événements de l'histoire des petits ruminants comme la domestication ou leur adaptation à leurs environnements.
Ainsi, l'introduction de ce manuscrit pointe l'importance des variants structuraux dans le contexte des espèces domestiques, et les petits ruminants ne semblent pas échapper à cette règle. De par leur nombre dans les deux espèces ainsi que les fonctions des gènes possiblement impactés, les SVs semblent avoir joué un rôle lors de la domestication puis de la sélection pour l'adaptation des petits ruminants à leurs environnements.

L'étude parallèle de ces deux espèces, de par leur proximité phylogénétique, ainsi que leur histoire de domestication similaire (temporalité de la domestication, lieu de cette domestication, histoire d'expansion, objectifs de sélection …), rend possible l'analyse des similitudes ou différences entre ces deux taxons et permet d'étudier des convergences évolutives entre ces taxons. Il est intéressant de noter que, dans les résultats présentés dans ce manuscrit, les SVs détectés comme sélectionnés chez la chèvre et le mouton ne sont jamais les mêmes, et ne sont donc pas issus d'un même polymorphisme ancestral

hérité et sélectionné de façon convergente. Nous pouvons cependant voir des similitudes entre ces deux taxons, au niveau des gènes et des fonctions impactées :
- Au niveau des gènes avec l'exemple du gène ASIP et des duplications qui l'englobent. En effet, si des duplications semblent inclure ce gène (ou des parties de ce gène) dans les deux taxons, ces duplications ne font pas la même taille et n'ont pas les mêmes bornes chez les deux espèces, preuve que ces duplications ne sont pas héritées d'un ancêtre commun. Dans l'hypothèse où ces SVs affecteraient la même voie métabolique (pigmentation), il pourrait alors s'agir d'une convergence entre chèvres et moutons dans le contexte de la domestication (voir Partie 2 – Chapitre 2 : Domestication et convergence évolutive, le cas du gène ASIP).
- Au niveau fonctionnel dans l'étude des variants à l'échelle des génomes complets. Dans le cas de la domestication, des variants sont identifiés dans des gènes connus pour modifier la physiologie des animaux, leur production, leur immunité ou leur reproduction (Voir Partie 3 – Chapitre 2 : Variants structuraux et petits ruminants).

Dès lors, il est possible de dire que les SVs semblent jouer un rôle important au sein des processus évolutifs et de l'histoire des espèces, en réponse aux pressions de sélection auxquelles elles sont soumises.

Il est cependant difficile de dire si les variants identifiés sont causaux (i. e. qui modifient réellement la fonction de la séquence), ou en déséquilibre de liaison avec d'autres mutations, qui seraient elles causales (au moins dans un premier temps). Il est donc nécessaire d'inclure les SVs dans leur contexte génomique, et d'intégrer l'ensemble des informations contenues dans ce contexte pour avoir une vision globale des différences et de leurs conséquences sur les individus.

Apports de l'étude des variants structuraux

SVs et SNPs : complémentarité des marqueurs.

Ce manuscrit pose une première pierre dans l'étude du rôle des SVs dans la domestication des petits ruminants. Pour répondre à la question posée au début de ce travail, à savoir l'apport de l'étude de ces variants, il semble intéressant de lier ces SVs aux SNPs, utilisés plus couramment dans les études de génomes complets.

- De la convergence des informations.

Cette étude des SVs, menée en parallèle de celle basée sur les SNPs (Alberto *et al., In prep*) permet d'étudier la complémentarité entre ces deux types de marqueurs. Si, d'une part, les deux approches convergent et pointent des zones identiques, tant pour la domestication (KITLG) que pour l'adaptation (RXFP2), de l'autre, les deux études pointent des régions différentes. Ainsi, Les SNPs indiquent des zones non détectées avec les SVs. Ceci n'est pas surprenant si l'on considère la densité de ces deux types de marqueurs le long du génome. Les SNPs étant beaucoup plus nombreux s'ils couvrent de façon plus fiable l'ensemble du génome et permettent de détecter un plus grand nombre de régions sous sélection.

Plus étonnamment, certains SVs sont détectés sous sélection alors que ceux-ci se trouvent dans des régions non identifiées avec les SNPs. Plusieurs éléments peuvent tout de même expliquer cette observation. Il peut avant tout s'agir de faux positifs. En effet, en plus des limites techniques qui entourent la détection des SVs (Voir Partie 3 – Chapitre 1 : Détection des SVs dans les données de reséquençage de génomes complets), les SVs utilisés sont des marqueurs dominants, la détection de signaux de sélection peut être biaisée par cette information incomplète. En effet, ces données ne donnent pas accès à l'hétérozygotie des individus. Il se peut alors que certaines mutations détectées comme liées au processus de domestication ou d'adaptation à l'environnement, le soient de façon artéfactuelle, en raison de la non exactitude des marqueurs.

D'un autre côté, les études menées avec les SNPs visaient à détecter des signaux convergents entre chèvres et moutons (Alberto *et al., In prep*). Pour ce faire, l'objectif était de limiter les faux positifs avec des seuils exigeants et d'utiliser des méthodes différentes de celles utilisées sur les SVs. Ces seuils exigeants et ces différences méthodologiques ont pu entraîner l'exclusion de régions sous sélection contenant les SVs détectés ici.

- SVs et SNPs forment des haplotypes sous sélection.

L'étude des SNPs en déséquilibre de liaison avec les SVs (par exemple dans la Partie 3 – Chapitre 2 : Article 4 - Exploring the role of genomic structural variations during small ruminant's domestication) permet tout de même de mettre en avant le fait que SVs et SNPs composent ensemble des haplotypes qui peuvent être sélectionnés (c'est par exemple le cas du gène KITLG chez la chèvre, voir Partie 3 – Chapitre 2 : Article 4 - Exploring the role of genomic structural variations during small ruminant's domestication). Bien qu'il soit nécessaire de faire des analyses plus approfondies pour identifier la mutation causale, il semble important d'avoir accès à l'ensemble du polymorphisme génétique pour interpréter au mieux le signal observé et identifier les mutation possiblement impliquées.

- SVs et SNPs ne contiennent pas les mêmes informations.

Il semble important de garder à l'esprit que ces deux marqueurs ne permettent pas d'accéder aux mêmes informations. Ainsi, les SNPs sont nombreux dans les génomes ; ils permettent donc de couvrir l'ensemble du génome. Ces informations permettent alors de retracer l'histoire neutre des espèces, mais aussi d'identifier les zones possiblement sous sélection. D'un autre côté, l'étude des SVs permet de compléter ces informations. En effet, du fait de leur non réversibilité ainsi que de leur faible fréquence, la probabilité qu'un même événement se soit produit au même endroit de deux façons indépendantes est faible, la présence (ou l'absence selon le génotype de l'individu utilisé comme référence) d'un SVs chez plusieurs individus indique donc leur proximité. De plus, il est possible de dater certains SVs. C'est par exemple le cas des éléments transposables à LTR où la divergence entre les séquences des deux LTRs, combinée avec l'hypothèse de l'horloge moléculaire, renseigne sur la date de l'insertion (Kijima and Innan, 2010). Il semble donc important de regarder ces deux types de marqueurs comme complémentaires.

L'ensemble de ces conclusions permet de mettre en avant le fait que l'étude des SVs et des SNPs est complémentaire pour intégrer l'ensemble du signal présent dans des données issues de reséquençage de génomes complets et être capable ensuite de les lier aux mécanismes biologiques et évolutifs.

Perspectives

Perspectives méthodologiques, approche hiérarchisée et pan-génomique.

Si d'un côté l'étude des SVs permet de mettre en lumière certains mécanismes possiblement impactés dans le contexte de la domestication et de l'adaptation, montrant ainsi l'intérêt d'inclure les SVs dans les études basées sur des données WGS, de l'autre, les limites qui entourent la détection et l'usage des SVs posent la question de la méthodologie à employer.

Au vu des problèmes que pose la détection des SVs, une approche hiérarchisée pourrait être mise en place pour permettre une détection rapide des zones sous sélection et identifier si des variants sont impliqués. Une première étape pourrait être de détecter les zones sous sélection avec des méthodes fiables et utilisées en routine basées sur les SNPs, et une deuxième, de regarder ces zones plus en profondeur et essayer d'identifier les haplotypes sélectionnés en intégrant un maximum de variations possibles. Ainsi, cette stratégie de scan rapide, puis d'approfondissement, permettrait de capter l'ensemble du signal et d'identifier au mieux la (ou les) variation(s) causale(s) qui est (sont) sélectionnée(s).

Cette approche ne résout cependant pas les problèmes méthodologiques qui s'opposent à la découverte des variants structuraux. Les deux approches déployées ici, ainsi que dans les travaux précédents (Bickhart and Liu, 2014), pointent la qualité de reconstruction des génomes de référence actuels comme limitant dans la détection des variants structuraux génomiques.

Une autre façon de voir le problème posé par la qualité des génomes de référence actuels est qu'ils se composent d'une séquence consensus unique et linéaire sensée représenter la diversité des génomes présents au sein d'une espèce, alors qu'elle ne représente en réalité qu'un seul individu. En prenant ce problème à l'envers, il semble que le génome de référence d'une espèce devrait rassembler toute l'information contenue au sein d'une espèce. Ce génome de référence global, aussi appelé pan-génome (Computational Pan-Genomics Consortium, 2016), peut être représenté sous la forme d'un graphe. Cette représentation en graphe du génome de référence d'une espèce permettrait de résumer l'ensemble des allèles connus à chaque locus pour une espèce donnée, et le génome d'un individu n'est alors qu'un chemin dans ce graphe (ou deux dans le cas d'un individu d'une espèce diploïde). Ainsi, l'usage pan-génome de référence permettrait de capturer l'ensemble de la variabilité au sein de ces individus, et ainsi, potentiellement, de contenir l'ensemble des informations génétiques d'une espèce. Considérant un graphe de référence, il est alors possible de lier le chemin emprunté par

un individu dans ce graphe avec des informations fonctionnelles et phénotypiques attachées aux nœuds parcourus par cet individu.
Un autre avantage d'un tel graphe est qu'il permettrait aussi de faire la synthèse des différents marqueurs actuels. Ainsi, indépendamment de sa taille et de sa fréquence, tout variant (SNP, indel, micro/macro-satellites, variants structuraux …) serait un locus alternatif, résolvant ainsi les problèmes de définition relevés tout au long de ce manuscrit.

Si nous avons vu que les techniques actuelles sont limitées pour la détection des SVs, l'essor de la pan-génomique devrait permettre d'intégrer toutes les informations de polymorphisme, indépendamment de leur taille ou de la définition qui leur est donnée.
Ainsi, en utilisant les données du projet Nextgen, couplées avec des techniques de construction et d'alignement de données de reséquençage sur un tel génome en graphe (Limasset *et al.*, 2016) (Liu *et al.*, 2016), il semble possible d'outrepasser les limites identifiées dans ce manuscrit. Cette approche pourrait notamment permettre d'explorer les caractéristiques génomiques des SVs chez les petits ruminants ainsi que d'intégrer les informations sous la forme d'haplotypes et non plus sous la forme de mutations indépendantes.

Perspectives biologiques, vers l'intégration de tous les signaux

Les travaux présentés dans ce manuscrit pointent donc l'importance d'intégrer tous les marqueurs génomiques pour interpréter au mieux les phénomènes évolutifs qui régissent la vie des individus. Ainsi, s'il est nécessaire d'inclure les SVs dans leur contexte génétique, avec les autres variations (telles que les SNPs les indels de petite taille, etc …), il est aussi nécessaire de prendre en compte d'autres informations telles que les informations épigénétiques (méthylation de l'ADN ou des histones) pour avoir une vision plus globale de ces processus évolutifs.
Cette approche globale doit permettre, dans le contexte de domestication et d'adaptation, d'obtenir la vision la plus complète des variations qui existent entre les individus et de comprendre les mécanismes qui permettent aux individus de répondre aux conditions environnementales.
Dans une optique évolutive, intégrer l'ensemble de ces signaux doit permettre d'étudier comment les individus et les populations évoluent, donnant ainsi des clés pour la compréhension des mécanismes micro-évolutifs.

BIBLIOGRAPHIE

Abd-Allah,M. *et al.* (2012) Relationships between haemoglobin (Hb) type and productive and reproductive performance of Rahmani ewes and lambs. *Online J. Anim. Feed Res. OJAFR*, **2**, 40–44.
Abyzov,A. *et al.* (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, **21**, 974–984.
Ai,H. *et al.* (2014) Population history and genomic signatures for high-altitude adaptation in Tibetan pigs. *BMC Genomics*, **15**.
Alberto,F.J. *et al.* Convergent genomic signatures of domestication in sheep and goats. *Prep*.
Alkan,C., Coe,P., *et al.* (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**, 363–376.
Alkan,C., Sajjadian,S., *et al.* (2011) Limitations of next-generation genome sequence assembly. *Nat. Methods*, **8**, 61–65.
Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
An,X.P. *et al.* (2012) Polymorphism identification in the goat KITLG gene and association analysis with litter size. *Anim. Genet.*, **43**, 104–107.
Arendt,M. *et al.* (2014) Amylase activity is associated with *AMY2B* copy numbers in dog: implications for dog domestication, diet and diabetes. *Anim. Genet.*, **45**, 716–722.
Armezzani,A. *et al.* (2014) "Ménage à Trois": The Evolutionary Interplay between JSRV, enJSRVs and Domestic Sheep. *Viruses*, **6**, 4926–4945.
Armezzani,A. *et al.* (2011) The Signal Peptide of a Recently Integrated Endogenous Sheep Betaretrovirus Envelope Plays a Major Role in Eluding Gag-Mediated Late Restriction ▽. *J. Virol.*, **85**, 7118–7128.
Arnaud,F. *et al.* (2007) A Paradigm for Virus–Host Coevolution: Sequential Counter-Adaptations between Endogenous and Exogenous Retroviruses. *PLoS Pathog*, **3**, e170.
Arnaud,F. *et al.* (2007) Mechanisms of Late Restriction Induced by an Endogenous Retrovirus. *J. Virol.*, **81**, 11441–11451.
Avery,O.T. *et al.* (1995) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. 1944. *Mol. Med.*, **1**, 344–365.
Axelsson,E. *et al.* (2013) The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature*, **495**, 360–364.
Bai,X. *et al.* (2006) Living with Genome Instability: the Adaptation of Phytoplasmas to Diverse Environments of Their Insect and Plant Hosts. *J. Bacteriol.*, **188**, 3682–3696.
Barsh,G.S. (2003) What Controls Variation in Human Skin Color? *PLoS Biol.*, **1**.
Beilharz,R.G. *et al.* (1993) Quantitative genetics and evolution: Is our understanding of genetics sufficient to explain evolution? *J. Anim. Breed. Genet. Z. Tierzuchtung Zuchtungsbiologie*, **110**, 161–170.
Benjelloun,B. *et al.* (2015) Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (Capra hircus) using WGS data. *Front. Genet.*, **6**.
Benjelloun,B. (2015) Diversité des génomes et adaptation locale des petits ruminants d'un pays méditerranéen: le Maroc.
Bickhart,D.M. *et al.* (2012) Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res.*, **22**, 778–790.
Bickhart,D.M. *et al.* (2017) Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.*, **49**, 643–650.
Bickhart,D.M. and Liu,G.E. (2014) The challenges and importance of structural variation

detection in livestock. *Evol. Popul. Genet.*, **5**, 37.
Bogucki,P. (1996) The spread of early farming in Europe. *Am. Sci.*, **84**, 242–253.
Boujenane,I. (2005) Small Ruminant Breeds of Morocco. In, *Characterisation of Small Ruminant Breeds in West Asia and North Africa North Africa*., pp. 5–54.
Brenig,B. *et al.* (2013) Molecular genetics of coat colour variations in White Galloway and White Park cattle. *Anim. Genet.*, **44**, 450–453.
Brooks,S.A. *et al.* (2008) A chromosome inversion near the KIT gene and the Tobiano spotting pattern in horses. *Cytogenet. Genome Res.*, **119**, 225–230.
Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
Carlson,J. *et al.* (2003) Chromosomal Distribution of Endogenous Jaagsiekte Sheep Retrovirus Proviral Sequences in the Sheep Genome. *J. Virol.*, **77**, 9662–9668.
Carroll,S.B. *et al.* (2008) Regulating evolution. *Sci. Am.*, **298**, 60–67.
Carter,A.M. (2009) Evolution of factors affecting placental oxygen transfer. *Placenta*, **30 Suppl A**, S19-25.
Cauwe,N. *et al.* (2007) Le Néolithique en Europe Armand Colin.
Chain,F.J.J. and Feulner,P.G.D. (2014) Ecological and evolutionary implications of genomic structural variations. *Evol. Popul. Genet.*, **5**, 326.
Chen,K. *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.
Chen,L. *et al.* (2017) Detection and validation of structural variations in bovine whole-genome sequence data. *Genet. Sel. Evol. GSE*, **49**.
Chen,R. *et al.* (2016) Detection of one large insertion/deletion (indel) and two novel SNPs within the <i>SPEF2</i> gene and their associations with male piglet reproduction traits. *Arch. Anim. Breed.*, **59**, 275–283.
Chessa,B. *et al.* (2009) Revealing the history of sheep domestication using retrovirus integrations. *Science*, **324**, 532–536.
Childe,V.G. (1925) The dawn of European civilization, K. Paul, Trench, Trubner & Co.; A.A. Knopf, London; New York.
Clark,L.A. *et al.* (2006) Retrotransposon insertion in SILV is responsible for merle patterning of the domestic dog. *Proc. Natl. Acad. Sci.*, **103**, 1376–1381.
Clutton-Brock,J. (2014) The Walking Larder: Patterns of Domestication, Pastoralism, and Predation Routledge.
Computational Pan-Genomics Consortium (2016) Computational pan-genomics: status, promises and challenges. *Brief. Bioinform.*
Conrad,D.F. and Hurles,M.E. (2007) The population genetics of structural variation. *Nat. Genet.*, **39**, S30–S36.
Conrad,O. *et al.* (2015) System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geosci. Model Dev.*, **8**, 1991–2007.
Consortium,I.H.G.S. and others (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
Dahm,R. (2008) Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Hum. Genet.*, **122**, 565–581.
Darwin,C. (1872) The Origin Of Species.
Dayama,G. *et al.* (2014) The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic Acids Res.*, **42**, 12640–12649.
Desquesnes,M. *et al.* (2013) Trypanosoma evansi and surra: a review and perspectives on origin, history, distribution, taxonomy, morphology, hosts, and pathogenic effects. *BioMed Res. Int.*, **2013**, 194176.
Doebley,J. *et al.* (1997) The evolution of apical dominance in maize. *Nature*, **386**, 485–488.
Dong,Y. *et al.* (2015) Reference genome of wild goat (capra aegagrus) and sequencing of

goat breeds provide insight into genic basis of goat domestication. *BMC Genomics*, **16**, 431.
Dray,S. *et al.* (2017) ade4: Analysis of Ecological Data : Exploratory and Euclidean Methods in Environmental Sciences.
Dreger,D.L. and Schmutz,S.M. (2011) A SINE insertion causes the black-and-tan and saddle tan phenotypes in domestic dogs. *J. Hered.*, **102 Suppl 1**, S11-18.
Drögemüller,C. *et al.* (2006) The mutation causing the black-and-tan pigmentation phenotype of Mangalitza pigs maps to the porcine ASIP locus but does not affect its coding sequence. *Mamm. Genome Off. J. Int. Mamm. Genome Soc.*, **17**, 58–66.
Duguma,R. *et al.* (2015) Spatial distribution of Glossina sp. and Trypanosoma sp. in south-western Ethiopia. *Parasit. Vectors*, **8**, 430.
Dunlap,K.A. *et al.* (2006) Endogenous retroviruses regulate periimplantation placental growth and differentiation. *Proc. Natl. Acad. Sci.*, **103**, 14390–14395.
Dupressoir,A. *et al.* (2012) From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. *Placenta*, **33**, 663–671.
Eilbeck,K. *et al.* (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.*, **6**, R44.
Elgar,G. and Vavouri,T. (2008) Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet. TIG*, **24**, 344–352.
ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
Fariello,M.-I. *et al.* (2014) Selection Signatures in Worldwide Sheep Populations. *PLOS ONE*, **9**, e103813.
Faucon,F. *et al.* (2015) Identifying genomic changes associated with insecticide resistance in the dengue mosquito *Aedes aegypti* by deep targeted sequencing. *Genome Res.*, **25**, 1347–1359.
Fernández,H. *et al.* (2006) Divergent mtDNA lineages of goats in an Early Neolithic site, far from the initial domestication areas. *Proc. Natl. Acad. Sci.*, **103**, 15375–15379.
Feuk,L. *et al.* (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 85–97.
Flori,L. *et al.* (2009) The Genome Response to Artificial Selection: A Case Study in Dairy Cattle. *PLOS ONE*, **4**, e6595.
Fontanesi,L. *et al.* (2009) 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.
Frantz,L.A.F. *et al.* (2016) Genomic and archaeological evidence suggest a dual origin of domestic dogs. *Science*, **352**, 1228–1231.
Frichot,E. *et al.* (2014) Fast and Efficient Estimation of Individual Ancestry Coefficients. *Genetics*, **196**, 973–983.
Frichot,E. *et al.* (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol. Biol. Evol.*, **30**, 1687–1699.
Frichot,E. and François,O. (2015) LEA : An R package for landscape and ecological association studies. *Methods Ecol. Evol.*, **6**, 925–929.
Galton,F. (1883) Inquiries Into Human Faculty and Its Development.
Garner,K.J. and Lingrel,J.B. (1988) Structural organization of the beta-globin locus of B-haplotype sheep. *Mol. Biol. Evol.*, **5**, 134–140.
Gavrilets,S. and Losos,J.B. (2009) Adaptive Radiation: Contrasting Theory with Data. *Science*, **323**, 732–737.
Gerstein,M.B. *et al.* (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Res.*, **17**, 669–681.
Gifford,R. and Tristem,M. (2003) The Evolution, Distribution and Diversity of Endogenous Retroviruses. *Virus Genes*, **26**, 291–315.

Gilly,A. *et al.* (2014) TE-Tracker: systematic identification of transposition events through whole-genome resequencing. *BMC Bioinformatics*, **15**.
Girardot,M. *et al.* (2006) The insertion of a full-length Bos taurus LINE element is responsible for a transcriptional deregulation of the Normande Agouti gene. *Pigment Cell Res.*, **19**, 346–355.
Gonzalez,M.V. *et al.* (2013) A divergent Artiodactyl MYADM-like repeat is associated with erythrocyte traits and weight of lamb weaned in domestic sheep. *PloS One*, **8**, e74700.
Griffiths,D.J. *et al.* (2010) Pathology and Pathogenesis of Ovine Pulmonary Adenocarcinoma. *J. Comp. Pathol.*, **142**, 260–283.
Guo,J. *et al.* (2012) A genome-wide association study using international breeding-evaluation data identifies major loci affecting production traits and stature in the Brown Swiss cattle breed. *BMC Genet.*, **13**, 82.
Hadjiconstantouras,C. *et al.* (2008) Characterization of the porcine KIT ligand gene: expression analysis, genomic structure, polymorphism detection and association with coat colour traits. *Anim. Genet.*, **39**, 217–224.
Hamosh,A. *et al.* (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514-517.
Handsaker,R.E. *et al.* (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.*, **43**, 269–276.
Hassanin,A. *et al.* (2012) Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. *C. R. Biol.*, **335**, 32–50.
Henke,W. (1944) Handbook of Paleoanthropology Springer.
Hershey,A.D. and Chase,M. (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.*, **36**, 39–56.
Hoekstra,H.E. (2006) Genetics, development and evolution of adaptive pigmentation in vertebrates. *Heredity*, **97**, 222–234.
Hoogendoorn,E. (2012) Computational methods for the detection of structural variation in the human genome.
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. Resour. Génétiques Anim. Genéticos Anim.*, **32**, 49–56.
Hou,Y., Bickhart,D.M., *et al.* (2012) Analysis of copy number variations in Holstein cows identify potential mechanisms contributing to differences in residual feed intake. *Funct. Integr. Genomics*, **12**, 717–723.
Hou,Y., Liu,G.E., *et al.* (2012) Genomic regions showing copy number variations associate with resistance or susceptibility to gastrointestinal nematodes in Angus cattle. *Funct. Integr. Genomics*, **12**, 81–92.
Huisman,T.H. and Kitchens,J. (1968) Oxygen equilibria studies of the hemoglobins from normal and anemic sheep and goats. *Am. J. Physiol.*, **215**, 140–146.
Ihaka,R. and Gentleman,R. (1996) R: A Language for Data Analysis and Graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.
International Sheep Genomics Consortium *et al.* (2010) The sheep genome reference sequence: a work in progress. *Anim. Genet.*, **41**, 449–453.
James,F.C. and McCulloch,C.E. (1990) Multivariate Analysis in Ecology and Systematics: Panacea or Pandora's Box? *Annu. Rev. Ecol. Syst.*, **21**, 129–166.
Jenkins,B.J. and Powell,B.C. (1994) Differential expression of genes encoding a cysteine-rich keratin family in the hair cuticle. *J. Invest. Dermatol.*, **103**, 310–317.
Jiang,J. *et al.* (2014) Global copy number analyses by next generation sequencing provide insight into pig genome variation. *BMC Genomics*, **15**, 593.
Jiang,Y. *et al.* (2015) Beta-globin gene evolution in the ruminants: evidence for an

ancient origin of sheep haplotype B. *Anim. Genet.*, **46**, 506–514.
Jiang,Y. *et al.* (2014) The sheep genome illuminates biology of the rumen and lipid metabolism. *Science*, **344**, 1168–1173.
Johnston,S.E. *et al.* (2011) Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population: GWAS OF SEXUAL WEAPONRY IN A WILD POPULATION. *Mol. Ecol.*, **20**, 2555–2566.
Kadri,N.K. *et al.* (2014) A 660-Kb Deletion with Antagonistic Effects on Fertility and Milk Production Segregates at High Frequency in Nordic Red Cattle: Additional Evidence for the Common Occurrence of Balancing Selection in Livestock. *PLoS Genet.*, **10**, e1004049.
Kaneko-Ishino,T. and Ishino,F. (2012) The role of genes domesticated from LTR retrotransposons and retroviruses in mammals. *Front. Microbiol.*, **3**.
Karner,C.M. *et al.* (2015) Gpr126/Adgrg6 deletion in cartilage models idiopathic scoliosis and pectus excavatum in mice. *Hum. Mol. Genet.*, **24**, 4365–4373.
Kawecki,T.J. and Ebert,D. (2004) Conceptual issues in local adaptation. *Ecol. Lett.*, **7**, 1225–1241.
Kijas,J.W. *et al.* (2013) Genetic diversity and investigation of polledness in divergent goat populations using 52 088 SNPs. *Anim. Genet.*, **44**, 325–335.
Kijas,J.W. *et al.* (2012) Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. *PLoS Biol*, **10**, e1001258.
Kijima,T.E. and Innan,H. (2010) On the estimation of the insertion time of LTR retrotransposable elements. *Mol. Biol. Evol.*, **27**, 896–904.
Kimura,M. (1983) The Neutral Theory of Molecular Evolution Cambridge University Press.
Korbel,J.O. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.
Kuhn,R.M. *et al.* (2012) The UCSC genome browser and associated tools. *Brief. Bioinform.*, bbs038.
Larson,G. *et al.* (2012) Rethinking dog domestication by integrating genetics, archeology, and biogeography. *Proc. Natl. Acad. Sci.*, **109**, 8878–8883.
Larson,G. and Burger,J. (2013) A population genetics view of animal domestication. *Trends Genet.*, **29**, 197–205.
Larson,G. and Fuller,D.Q. (2014) The Evolution of Animal Domestication. *Annu. Rev. Ecol. Evol. Syst.*, **45**, 115–136.
Lavialle,C. *et al.* (2013) Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **368**, 20120507.
Layer,R.M. *et al.* (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, **15**, R84.
Leggett,R.M. and MacLean,D. (2014) Reference-free SNP detection: dealing with the data deluge. *BMC Genomics*, **15 Suppl 4**, S10.
Li,H. *et al.* (2010) Genome-wide scan for positional and functional candidate genes affecting milk production traits in Canadian Holstein Cattle. *Proc 9th WCGALP Leipz. Ger.*, **26**.
Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.*, **25**, 2078–2079.
Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
Li,M. *et al.* (2014) Whole-genome sequencing of Berkshire (European native pig) provides insights into its origin and domestication. *Sci. Rep.*, **4**, 4678.
Li,Y. *et al.* (2011) Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nat. Biotechnol.*, **29**, 723–

730.
Limasset,A. *et al.* (2016) Read mapping on de Bruijn graphs. *BMC Bioinformatics*, **17**, 237.
Lin,K. *et al.* (2014) Making the difference: integrating structural variation detection tools. *Brief. Bioinform.*, bbu047.
Liu,B. *et al.* (2016) deBGA: read alignment with de Bruijn graph-based seed and extension. *Bioinforma. Oxf. Engl.*, **32**, 3224–3232.
Lorenz,M.G. and Wackernagel,W. (1994) Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol. Rev.*, **58**, 563–602.
Manel,S. and Holderegger,R. (2013) Ten years of landscape genetics. *Trends Ecol. Evol.*, **28**, 614–621.
Marshak,A. (1936) The Structure of the Chromosomes of the Salivary Gland of Drosophila melanogaster. *Am. Nat.*, **70**, 181–184.
McClintock,B. (1953) Induction of Instability at Selected Loci in Maize. *Genetics*, **38**, 579–599.
McClure,M.C. *et al.* (2010) A genome scan for quantitative trait loci influencing carcass, post-natal growth and reproductive traits in commercial Angus cattle. *Anim. Genet.*, **41**, 597–607.
McLeod,A.I. (2005) Kendall rank correlation and Mann-Kendall trend test. *R Package Kendall*.
Medvedev,P. *et al.* (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods*, **6**, S13–S20.
Megens,H.-J. and Groenen,M. a. M. (2012) Domesticated species form a treasure-trove for molecular characterization of Mendelian traits by exploiting the specific genetic structure of these species in across-breed genome wide association studies. *Heredity*, **109**, 1–3.
Mills,R.E. *et al.* (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.*, **16**, 1182–1190.
Mills,R.E. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.
Muigai,A.W.T. and Hanotte,O. (2013) The Origin of African Sheep: Archaeological and Genetic Perspectives. *Afr. Archaeol. Rev.*, **30**, 39–50.
Mukhopadhyay,R. (2009) DNA sequencers: the next generation. *Anal. Chem.*, **81**, 1736–1740.
Naderi,S. *et al.* (2008) The goat domestication process inferred from large-scale mitochondrial DNA analysis of wild and domestic individuals. *Proc. Natl. Acad. Sci.*, **105**, 17659–17664.
Nielsen,R. *et al.* (2011) Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.*, **12**, 443–451.
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.
Novembre,J. and Di Rienzo,A. (2009) Spatial patterns of variation due to natural selection in humans. *Nat. Rev. Genet.*, **10**, 745–755.
Noyes,H. *et al.* (2011) Genetic and expression analysis of cattle identifies candidate genes in pathways responding to Trypanosoma congolense infection. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 9304–9309.
Ohno,S. (1972) So much 'junk' DNA in our genome. *Brookhaven Symp. Biol.*, **23**, 366–370.
Ordás,J.G. (2004) Structure of European ovine populations from directional autocorrelations between proteins. *J. Anim. Breed. Genet.*, **121**, 229–241.
Orozco-terWengel,P. *et al.* (2015) Revisiting demographic processes in cattle with genome-wide population genetic analysis. *Front. Genet.*, **6**, 191.
Pabinger,S. *et al.* (2014) A survey of tools for variant analysis of next-generation genome

sequencing data. *Brief. Bioinform.*, **15**, 256–278.
Pailhoux,E. *et al.* (2001) A 11.7-kb deletion triggers intersexuality and polledness in goats. *Nat. Genet.*, **29**, 453–458.
Pang,A.W. *et al.* (2010) Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.*, **11**, R52.
Pasyukova,E.G. *et al.* (2004) Accumulation of transposable elements in the genome of Drosophila melanogaster is associated with a decrease in fitness. *J. Hered.*, **95**, 284–290.
Patterson,D. (2009) Molecular genetic analysis of Down syndrome. *Hum. Genet.*, **126**, 195–214.
Paudel,Y. *et al.* (2013) Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC Genomics*, **14**, 449.
Pausch,H. *et al.* (2012) Identification of QTL for UV-Protective Eye Area Pigmentation in Cattle by Progeny Phenotyping and Genome-Wide Association Analysis. *PLOS ONE*, **7**, e36346.
Pereira,F. *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.
Pieragostini,E. *et al.* (2006) Functional effect of haemoglobin polymorphism on the haematological pattern of Gentile di Puglia sheep. *J. Anim. Breed. Genet.*, **123**, 122–130.
Pieragostini,E. *et al.* (1994) Hemoglobin phenotypes and hematological factors in Leccese sheep. *Small Rumin. Res.*, **13**, 177–185.
Pompanon,F. *et al.* (2005) Genotyping errors: causes, consequences and solutions. *Nat. Rev. Genet.*, **6**, 847–846.
Queiroz,K. de (2005) Ernst Mayr and the modern concept of species. *Proc. Natl. Acad. Sci.*, **102**, 6600–6607.
Rausch,T. *et al.* (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, i333–i339.
Rellstab,C. *et al.* (2015) A practical guide to environmental association analysis in landscape genomics. *Mol. Ecol.*, **24**, 4348–4370.
Reyer,H. *et al.* (2017) Exploring the genetics of feed efficiency and feeding behaviour traits in a pig line highly selected for performance characteristics. *Mol. Genet. Genomics MGG*.
Rezaei,H. (2007) Phylogénie moléculaire du Genre Ovis (Mouton et Mouflons), Implications pour la Conservation du Genre et pour l'Origine de l'Espèce Domestique.
Rezaei,H.R. *et al.* (2010) Evolution and taxonomy of the wild species of the genus Ovis (Mammalia, Artiodactyla, Bovidae). *Mol. Phylogenet. Evol.*, **54**, 315–326.
Rieder,S. *et al.* (2001) Mutations in the agouti (ASIP), the extension (MC1R), and the brown (TYRP1) loci and their association to coat color phenotypes in horses (Equus caballus). *Mamm. Genome Off. J. Int. Mamm. Genome Soc.*, **12**, 450–455.
Ropiquet,A. and Hassanin,A. (2005) Molecular phylogeny of caprines (Bovidae, Antilopinae): the question of their origin and diversification during the Miocene. *J. Zool. Syst. Evol. Res.*, **43**, 49–60.
Sanger,F. and Coulson,A.R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, **94**, 441–448.
Saowaphak,P. *et al.* (2017) Genetic correlation and genome-wide association study (GWAS) of the length of productive life, days open, and 305-days milk yield in crossbred Holstein dairy cattle. *Genet. Mol. Res. GMR*, **16**.
Serão,N.V. *et al.* (2013) Single nucleotide polymorphisms and haplotypes associated with feed efficiency in beef cattle. *BMC Genet.*, **14**, 94.

Serão,N.V.L. *et al.* (2013) Bivariate genome-wide association analysis of the growth and intake components of feed efficiency. *PloS One*, **8**, e78530.
Shin,D.-H. *et al.* (2014) Deleted copy number variation of Hanwoo and Holstein using next generation sequencing at the population level. *BMC Genomics*, **15**, 240.
de Simoni Gouveia,J.J. *et al.* (2014) Identification of selection signatures in livestock species. *Genet. Mol. Biol.*, **37**, 330–342.
Sironen,A. *et al.* (2012) L1 insertion within SPEF2 gene is associated with increased litter size in the Finnish Yorkshire population. *J. Anim. Breed. Genet. Z. Tierzuchtung Zuchtungsbiologie*, **129**, 92–97.
Sistiaga-Poveda,M. and Jugo,B.M. (2014) Evolutionary dynamics of endogenous Jaagsiekte sheep retroviruses proliferation in the domestic sheep, mouflon and Pyrenean chamois. *Heredity*, **112**, 571–578.
Spielmann,M. and Mundlos,S. (2016) Looking beyond the genes: the role of non-coding variants in human disease. *Hum. Mol. Genet.*, **25**, R157–R165.
Strucken,E.M. *et al.* (2014) Genomewide study and validation of markers associated with production traits in German Landrace boars. *J. Anim. Sci.*, **92**, 1939–1944.
Stucki,S. *et al.* (2017) High performance computation of landscape genomic models including local indicators of spatial association. *Mol. Ecol. Resour.*, **17**, 1072–1089.
Sudmant,P.H. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.
Taberlet,P. *et al.* (2008) Are cattle, sheep, and goats endangered species? *Mol. Ecol.*, **17**, 275–284.
Tattini,L. *et al.* (2015) Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Front. Bioeng. Biotechnol.*, **3**.
U. Cinar,M. *et al.* (2016) P5059 Ovine MYADM-like repeat gene association with lifetime cumulative ewe production and wool traits. *J. Anim. Sci.*, **94**, 144.
Varela,M. *et al.* (2009) Friendly Viruses: The Special Relationship between Endogenous Retroviruses and Their Host. *Ann. N. Y. Acad. Sci.*, **1178**, 157–172.
Vbra,E. (2000) Antelopes, Deer, and Relatives: Fossil Record, Behavioral Ecology, Systematics, and Conservation.
Viginier,B. *et al.* (2012) Copy Number Variation and Differential Expression of a Protective Endogenous Retrovirus in Sheep. *PLoS ONE*, **7**, e41965.
Vigne,J.-D. (2011) The origins of animal domestication and husbandry: a major change in the history of humanity and the biosphere. *C. R. Biol.*, **334**, 171–181.
de Villemereuil,P. *et al.* (2014) Genome scan methods against more complex models: when and how much should we trust them? *Mol. Ecol.*, **23**, 2006–2019.
Watson,J.D. and Crick,F.H.C. (1953) Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, **171**, 737–738.
Weir,B.S. and Cockerham,C.C. (1984) Estimating F-Statistics for the analysis of population structure. *Evol. Int. J. Org. Evol.*, **38**, 1358–1370.
Wiedemar,N. and Drögemüller,C. (2015) A 1.8-kb insertion in the 3′-UTR of *RXFP2* is associated with polledness in sheep. *Anim. Genet.*, **46**, 457–461.
Wiener,P. and Wilkinson,S. (2011) Deciphering the genetic basis of animal domestication. *Proc. Biol. Sci.*, **278**, 3161–3170.
Wilkins,A.S. *et al.* (2014) The 'domestication syndrome' in mammals: a unified explanation based on neural crest cell behavior and genetics. *Genetics*, **197**, 795–808.
Willcox,G. and Stordeur,D. (2012) Large-scale cereal processing before domestication during the tenth millennium cal BC in northern Syria. *Antiquity*, **86**, 99–114.
Wilson,D.E. and Reeder,D.M. (2005) Mammal Species of the World: A Taxonomic and Geographic Reference JHU Press.
Wright,D. (2015) The Genetic Architecture of Domestication in Animals. *Bioinforma. Biol.*

Insights, **9**, 11–20.
Ye,K. *et al.* (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.
Yoon,S. *et al.* (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592.
Zeder,M.A. (2015) Core questions in domestication research. *Proc. Natl. Acad. Sci.*, **112**, 3191–3198.
Zeder,M.A. *et al.* (2006) Documenting domestication: the intersection of genetics and archaeology. *Trends Genet.*, **22**, 139–155.
Zeder,M.A. (2008) Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 11597–11604.
Zeder,M.A. (2012) The domestication of animals. *J. Anthropol. Res.*, **68**, 161.
Zeder,M.A. and Hesse,B. (2000) The initial domestication of goats (Capra hircus) in the Zagros mountains 10,000 years ago. *Science*, **287**, 2254–2257.
Zereini,F. and Hötzl,H. eds. (2008) Climatic Changes and Water Resources in the Middle East and North Africa Springer Berlin Heidelberg, Berlin, Heidelberg.
Zhao,J. *et al.* (2017) Identification of genes and proteins associated with anagen wool growth. *Anim. Genet.*, **48**, 67–79.
Zhao,M. *et al.* (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, **14**, S1.
Zhou,H. *et al.* (2015) A 57-bp deletion in the ovine KAP6-1 gene affects wool fibre diameter. *J. Anim. Breed. Genet.*, **132**, 301–307.

MATÉRIEL SUPPLÉMENTAIRE

Matériel supplémentaire - SECONDE PARTIE : Variants structuraux – Approche « variants candidats »

Matériel supplémentaire - Article 2 : Old origin of a protective endogenous retrovirus (enJSRV) in the Ovis genus.

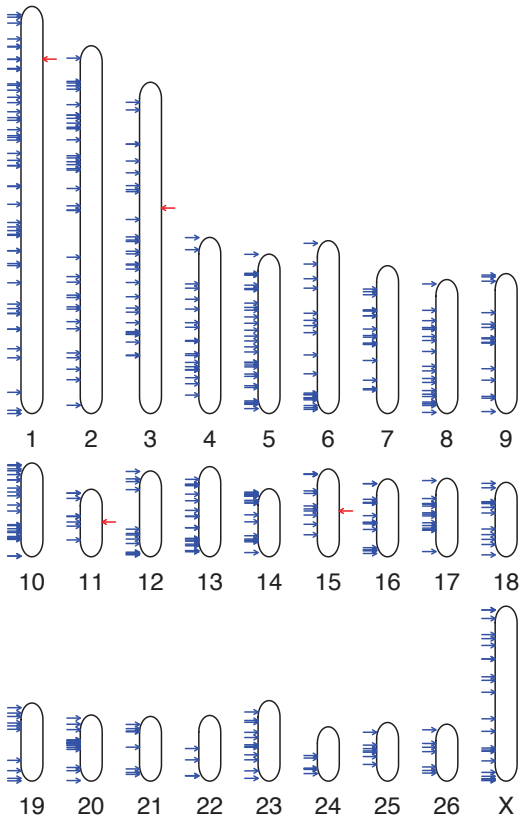


Figure S1: EnJSRVs chromosomal distribution. Blue arrows indicate insertion site detected by whole genome survey. Red arrows indicate insertion site of previously described (Chessa et al. 2009). On chromosome 3, enJSRV-8 is absent from our dataset.

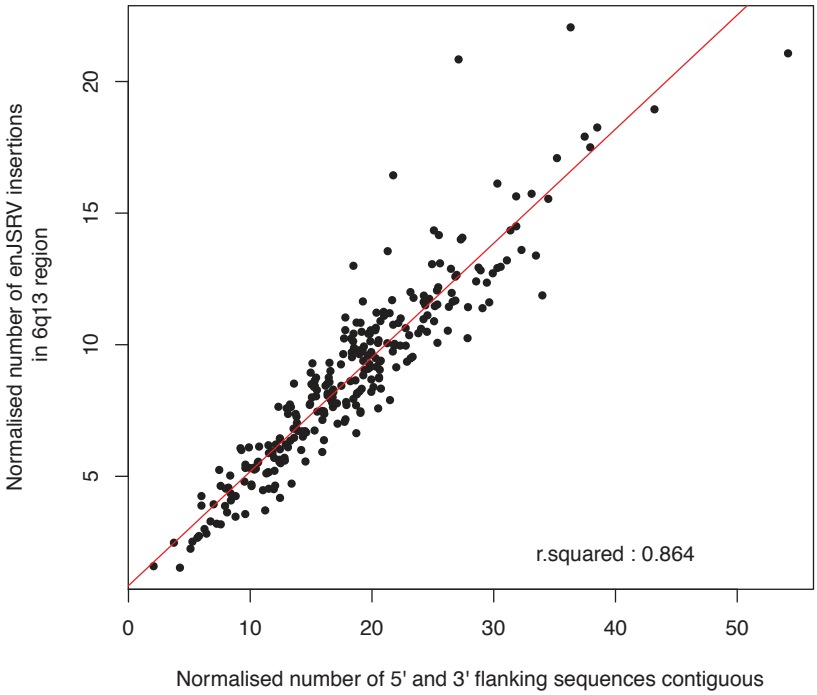


Figure S2: Normalised number of enJSRV insertions in the 6q13 region in function of the normalised number of 5’ and 3’ flanking sequences contiguous in the 6q13 region.

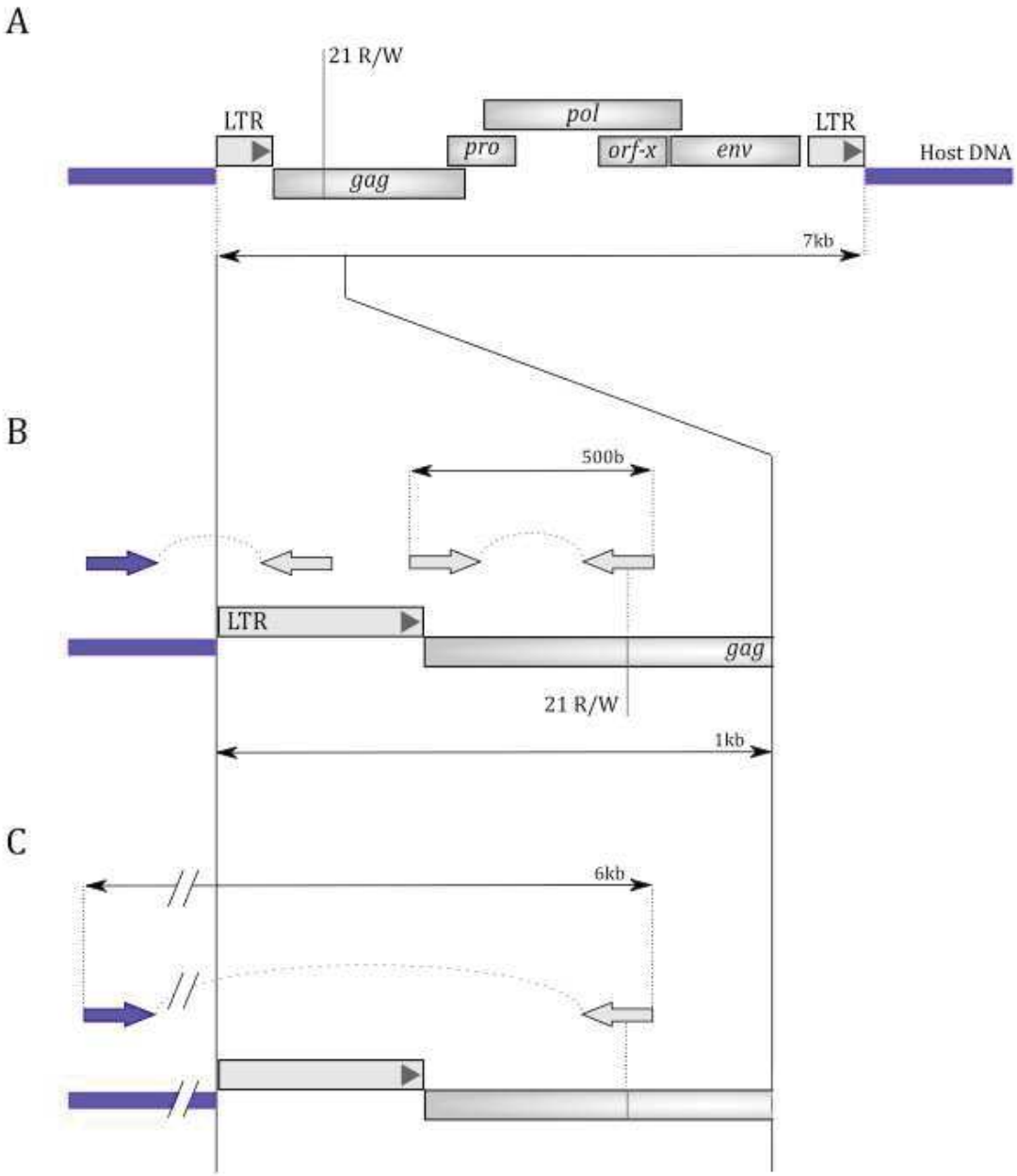


Figure S3: Schematic representation of the 6q13 locus. **A.** Global representation of an enJSRV6q13. **B.** Representation of the limits of the library size to detect in the same time the number of insertions and the number of mutations. **C.** Mate-pair library allows confirming insertion and mutation in the 6q13 locus.

TABLE S1. putative positions of enJSRVs distribution along chromosomes

Chromosome	Start	Stop
1	342	639
1	2205879	2207196
1	14384701	14384836
1	37717873	37718219
1	43856484	43857177
1	57059277	57060170
1	57193894	57194075
1	67591577	67591695
1	67917957	67918857
1	88525704	88525910
1	100177541	100178031
1	101531283	101532174
1	110124268	110124408
1	110263276	110263474
1	120684708	120685538
1	121300002	121300139
1	123844156	123844338
1	125907498	125907620
1	126354935	126355308
1	129345396	129346155
1	141667485	141667680
1	153680801	153681018
1	167653959	167654268
1	176116007	176117307
1	185111597	185112429
1	186979733	186980689
1	188119890	188120306
1	192148370	192149100
1	198528740	198528956
1	200333820	200334198
1	210472404	210472641
1	214110346	214111402
1	218916724	218916906
1	222227096	222227380
1	223161909	223162608
1	233159705	233161286

	1	233623777	233624453
	1	239700322	239701283
	1	239924941	239930218
	1	248163897	248164833
	1	253585865	253587353
	1	264331689	264332073
	1	268322867	268323216
	1	270038023	270038249
	2	5598443	5598576
	2	22893847	22894449
	2	30029537	30029780
	2	37783110	37783587
	2	41010075	41010427
	2	62201200	62202028
	2	62637353	62637461
	2	64185165	64185284
	2	70700842	70701233
	2	72243833	72244031
	2	78555763	78555917
	2	79424743	79424924
	2	80143849	80144423
	2	88969215	88969522
	2	92714889	92716243
	2	93868997	93869107
	2	105835597	105835951
	2	137474462	137474704
	2	137769423	137769567
	2	140646171	140646904
	2	164826311	164826436
	2	165751902	165752550
	2	170493291	170493730
	2	173047129	173047554
	2	174526102	174526762
	2	185298227	185298488
	2	192813156	192813337
	2	196599818	196600413
	2	199868504	199869102
	2	202012987	202013360
	2	202026037	202026537
	2	209102491	209103442
	2	221859647	221860026

	2	223807837	223808755
	2	224857243	224857983
	2	240619718	240620642
	3	9329933	9330100
	3	39186126	39195498
	3	39574854	39575953
	3	48394403	48395059
	3	53562177	53563626
	3	53672831	53673582
	3	53720670	53720839
	3	54152338	54152464
	3	54681942	54682743
	3	55447772	55448277
	3	61509401	61509781
	3	71396208	71396504
	3	79081810	79082366
	3	88725088	88725419
	3	97794805	97795019
	3	100574634	100575018
	3	116512522	116512982
	3	117447687	117448152
	3	119600616	119601166
	3	123271753	123271880
	3	131375845	131376194
	3	131971918	131972455
	3	150242656	150242853
	3	151579466	151581898
	3	154289872	154290161
	3	182331929	182332744
	3	182488558	182489731
	3	192140869	192141026
	3	208008872	208009059
	3	210708657	210709487
	4	494978	495078
	4	12499965	12500560
	4	20187309	20187763
	4	24268081	24268316
	4	29570695	29570870
	4	29581486	29581833
	4	29795041	29795282
	4	31710120	31710731

4	34617114	34617910
4	39333289	39334026
4	40461352	40462280
4	48933475	48934365
4	49466400	49466557
4	58425562	58426328
4	61512008	61512190
4	62950672	62950871
4	70715644	70715869
4	77382933	77383791
4	84885218	84885920
4	87757497	87757613
4	110929300	110929939
4	119106962	119107190
5	2245220	2245329
5	3337947	3339051
5	6254426	6254769
5	7712478	7717661
5	8556834	8557511
5	17652410	17652755
5	25191266	25191724
5	31588136	31588423
5	32859808	32860389
5	34453325	34453478
5	34532466	34532758
5	41689666	41690047
5	45136894	45138484
5	49158918	49159101
5	56429867	56430303
5	61146356	61147585
5	61546003	61546166
5	65403799	65404400
5	71849837	71850193
5	74760148	74760777
5	83919859	83920710
5	86914702	86915038
5	93777799	93778038
5	93943311	93943958
5	94060618	94061471
5	94954928	94955431
5	99403130	99403273

	5	107785561	107785761
	6	3909611	3909877
	6	5407826	5408627
	6	9284715	9285367
	6	10564342	10564581
	6	10752461	10753041
	6	12716680	12717339
	6	12968927	12970271
	6	13860124	13860527
	6	15864048	15864153
	6	26903030	26903347
	6	44111200	44111348
	6	54758386	54759077
	6	59522738	59523456
	6	63100958	63102002
	6	67900058	67901549
	6	91264533	91265051
	6	115251702	115253046
	7	16069121	16069287
	7	16800216	16801160
	7	22564532	22565074
	7	23552547	23552709
	7	31323220	31323369
	7	35925425	35926538
	7	42270045	42270223
	7	46770191	46770875
	7	47180532	47180829
	7	47717322	47717815
	7	51680986	51681311
	7	52791858	52792682
	7	65962454	65962654
	7	66193144	66193393
	7	69512154	69512978
	7	70225681	70225881
	7	82525096	82525384
	7	84282139	84282420
	8	762798	763312
	8	4266341	4266467
	8	5477298	5477671
	8	6601534	6601715
	8	6791797	6792086

8	7805026	7805638
8	11545439	11546171
8	13114725	13116194
8	13144610	13144806
8	15554617	15555445
8	21067792	21068150
8	26395548	26395679
8	31964735	31966191
8	42025182	42025521
8	42833095	42833246
8	49447089	49452035
8	52141755	52142010
8	58040696	58042435
8	58230403	58230661
8	62883886	62884086
8	69871446	69871546
9	1335132	1335646
9	10396888	10397060
9	11582255	11582477
9	22596749	22597569
9	30339085	30339764
9	47826153	47826260
9	47986582	47992598
9	48069202	48069435
9	48085756	48086199
9	50560132	50560493
9	51282785	51283023
9	58298367	58298995
9	60543405	60543645
9	68283882	68284577
9	74377327	74377509
9	89616074	89616425
9	93749753	93751200
10	647010	647635
10	16003412	16004140
10	18084299	18085221
10	18685131	18686087
10	21617729	21617890
10	25832743	25833387
10	29588207	29589131
10	42138555	42139995

10	47691830	47692284
10	56177117	56177482
10	70818749	70818892
10	70875412	70875672
10	74044429	74044539
10	75754428	75754614
10	79300547	79300775
10	83982083	83982720
10	84860626	84861247
11	28240905	28242188
11	32040573	32049553
11	37169037	37169726
11	37275426	37275626
11	37288586	37288848
11	39691484	39691648
11	53828077	53828457
11	58645670	58645801
11	58760228	58760944
12	1817176	1818635
12	3051196	3051882
12	3379651	3380810
12	4212217	4213191
12	12223631	12223761
12	20120926	20121787
12	21411994	21412760
12	25451474	25451935
12	54127796	54127968
12	55909805	55910005
12	61869488	61874567
12	69295656	69295840
12	71813569	71814164
13	4613535	4614205
13	5660528	5660822
13	10166106	10166457
13	11624541	11624679
13	14374019	14375234
13	15369608	15370479
13	16657853	16658356
13	16923769	16925070
13	26668709	26669011
13	26802177	26804085

13	37028781	37028902
13	37473254	37474460
13	42751972	42752630
13	51646054	51646354
13	65538088	65538800
13	65912334	65913023
13	66757785	66757987
13	66805801	66806368
13	71396838	71397780
13	73488152	73488299
14	3856048	3856908
14	13773683	13780959
14	14971659	14980135
14	15223892	15224121
14	16361403	16361881
14	38663513	38664335
14	44115742	44115868
14	49408238	49409612
14	50398708	50398990
14	52017299	52017499
14	55968772	55969680
14	56091086	56091274
14	56130210	56130428
14	57477208	57478067
14	57501294	57509925
14	57938885	57939519
14	59600134	59600248
15	38451168	38451300
15	42210989	42218388
15	44088138	44089566
15	46872217	46873045
15	53278089	53278204
15	60727882	60728229
15	75925987	75926108
15	76228047	76228225
15	76378592	76378792
16	3014416	3015130
16	5469289	5470177
16	6306642	6307224
16	25168923	25169204
16	31490866	31491325

16	31606724	31607033
16	31637575	31638584
16	37880919	37881875
16	49444446	49445003
16	49827306	49827643
16	50083698	50084419
16	67340837	67343461
16	67429063	67429214
17	4799548	4799935
17	24180310	24180483
17	24476595	24476909
17	24952057	24952255
17	26386142	26386299
17	31340912	31341665
17	37652428	37653197
17	39306448	39306959
17	40515405	40515647
17	49320735	49320874
17	70266242	70266410
18	1745731	1746577
18	11313776	11314646
18	16031940	16032335
18	20421369	20421533
18	45540229	45540450
18	49039604	49040011
18	50745977	50746816
18	51648143	51648305
18	59484167	59484271
18	60858720	60858855
18	63585684	63586276
18	67684215	67684835
19	216298	217044
19	16088646	16089816
19	40146810	40147510
19	42778494	42778746
19	52693205	52698033
19	56858629	56859268
20	439875	440708
20	8078714	8079957
20	10560277	10560448
20	24548420	24548603

20	25894311	25895216
20	27503439	27512324
20	27644831	27645035
20	27681519	27682215
20	29086233	29086861
20	29134959	29135562
20	29373765	29374538
20	29803622	29805080
20	30009950	30010084
20	30939896	30940223
20	31625205	31625393
20	39882607	39882894
20	44049314	44049592
20	48822223	48824333
21	6670613	6673917
21	9453275	9453971
21	26305784	26306380
21	38172977	38173101
21	38826302	38827077
21	40635749	40635938
22	4210324	4210623
22	16472241	16472356
22	25223297	25223959
22	33186290	33186404
23	2167537	2167990
23	7944860	7945077
23	16500828	16502291
23	20365671	20365981
23	26993479	26994305
23	33759588	33759808
23	37782226	37782808
23	46951814	46953083
23	62173878	62173998
24	5838772	5839243
24	9079889	9081007
24	9696720	9696925
24	10000606	10001303
24	17847402	17847597
24	19785012	19785833
24	26936023	26936191
25	12919655	12919981

25	19012317	19012959
25	22399435	22399745
25	23373665	23373832
25	25140637	25141145
25	27882202	27882550
25	37835974	37837470
26	6217951	6218700
26	8240712	8241365
26	22834694	22835611
26	26203038	26204564
26	29143308	29143872
26	39733535	39734340
X	410747	412151
X	854922	855027
X	1839426	1840180
X	6767081	6767711
X	15662387	15662647
X	23137844	23138163
X	24058816	24059051
X	25613659	25613933
X	38550273	38552731
X	48250470	48251866
X	68808481	68809297
X	78053684	78054013
X	80796664	80798020
X	94485352	94485669
X	94701774	94702431
X	105056320	105057753
X	110305261	110306549
X	113448410	113449223
X	125917159	125925275
X	132518091	132518197
X	132532853	132533798
UnplacedScaffold_004080257.1	11568	11784
Unplaced_Contig646_fixed	88	329
UnplacedScaffold_004080949.1	5887	6361
UnplacedScaffold_004080951.1	871	1616
UnplacedScaffold_004081385.1	13936	14136
Unplaced_Contig2508_fixed	1848	2062
Unplaced_Contig2542_fixed	1632	2082
UnplacedScaffold_004081482.1	52	3365

UnplacedScaffold_004081513.1	36197	36587
UnplacedScaffold_004081611.1	5105	5750
UnplacedScaffold_004081704.1	38852	39202
UnplacedScaffold_004082124.1	90	127
UnplacedScaffold_004082173.1	896	6590
UnplacedScaffold_004082437.1	4298	4735
UnplacedScaffold_004082547.1	26498	27063
UnplacedScaffold_004082613.1	385	680
UnplacedScaffold_004082672.1	1821	3366
UnplacedScaffold_004082774.1	6891	7023
UnplacedScaffold_004082890.1	5819	6716
UnplacedScaffold_004083582.1	13635	14355
UnplacedScaffold_004083595.1	913	1083
UnplacedScaffold_004085138.1	99	10866
UnplacedScaffold_004085449.1	9790	9997
Unplaced_Contig10233_fixed	798	1364

Table S2. Informations about mate-pairs reads with one mapped on the sequence flanking the 3' or 5' side of the insertion and the other carrying the protective mutation (in red).

Read Name	Library size	Mapping Quality	Mate Position	Sequence
ERR169319.20584717	2900	60	5' flanking sequence	AAATATGGGACAGACGCATAGTCGTCAGTTGTTTGTGCATATGTTATCTGTAATGTTAAAACAT**T**GGGGAATTACTGTTTCTAAACCTAAATTAATCAATT
ERR169327.1128786	8020	60	3' flanking sequence	TTGTTTGTGCATATGTTATCTGTAATGTTAAAACAT**T**GGGGAATTACTGTTTCTAAACCTAAATTAATCAATTTTCTTTCATTCATCGAGGAAGTTTGCCC
ERR169327.1422824	8020	60	3' flanking sequence	TTGTTTGTGCATATGTTATCTGTAATGTTAAAACAT**T**GGGGAATTACTGTTTCTAAACCTAAATTAATCAATTTTCTTTCATTCATCGAGGAAGTTTGCCC
ERR169327.10969254	8020	60	3' flanking sequence	TTGTTTGTGCATATGTTATCTGTAATGTTAAAACAT**T**GGGGAATTACTGTTTCTAAACCTAAATTAATCAATTTTCTTTCATTCATCGAGGAAGTTTGCCC
ERR169327.13131248	8020	60	3' flanking sequence	TTGTTTGTGCATATGTTATCTGTAATGTTAAAACAT**T**GGGGAATTACTGTTTCTAAACCTAAATTAATCAATTTTCTTTCATTCATCGAGGAAGTTTGCCC
ERR169317.4373880	8020	60	3' flanking sequence	GTTTGTGCATATGTTATCTGTAATGTTAAAACAT**T**GGGGAATTACTGTTTCTAAACCTAAATTAATCAATTTTCTTTCATTCATCGAGGAAGTTTGCCC

TABLE S3. Summary of sheep used in this study.

sample_name	sample_accession	biosamples_id	sample_provide	species	taxonomy_id	breed	country
IROA-B2-5296	ERS239046	SAMEA2065588	NEXTGEN	Ovis aries	9940	.	Iran
IROA-B3-5134	ERS239047	SAMEA2065589	NEXTGEN	Ovis aries	9940	.	Iran
IROA-B4-5190	ERS154865	SAMEA2012929	NEXTGEN	Ovis aries	9940	.	Iran
IROA-B5-5295	ERS154863	SAMEA2012927	NEXTGEN	Ovis aries	9940	.	Iran
IROA-B6-5139	ERS239048	SAMEA2065590	NEXTGEN	Ovis aries	9940	.	Iran
IROA-C3-5212	ERS154869	SAMEA2012933	NEXTGEN	Ovis aries	9940	.	Iran
IROA-C6-5187	ERS239049	SAMEA2065591	NEXTGEN	Ovis aries	9940	.	Iran
IROA-C7-5042	ERS239050	SAMEA2065592	NEXTGEN	Ovis aries	9940	.	Iran
IROA-D5-5081	ERS239051	SAMEA2065593	NEXTGEN	Ovis aries	9940	.	Iran
IROA-D6-5152	ERS154866	SAMEA2012930	NEXTGEN	Ovis aries	9940	.	Iran
IROA-D7-5033	ERS239052	SAMEA2065594	NEXTGEN	Ovis aries	9940	.	Iran
IROA-E5-5157	ERS239053	SAMEA2065595	NEXTGEN	Ovis aries	9940	.	Iran
IROA-E6-5351	ERS239054	SAMEA2065596	NEXTGEN	Ovis aries	9940	.	Iran
IROA-E7-5036	ERS239055	SAMEA2065597	NEXTGEN	Ovis aries	9940	.	Iran
IROA-F10-5068	ERS239056	SAMEA2065598	NEXTGEN	Ovis aries	9940	.	Iran
IROA-F3-5142	ERS154867	SAMEA2012931	NEXTGEN	Ovis aries	9940	.	Iran
IROA-F5-5051	ERS154868	SAMEA2012932	NEXTGEN	Ovis aries	9940	.	Iran
IROA-G3-5095	ERS239057	SAMEA2065599	NEXTGEN	Ovis aries	9940	.	Iran
IROA-G4-5205	ERS154862	SAMEA2012926	NEXTGEN	Ovis aries	9940	.	Iran
IROO-C3-0001	ERS154526	SAMEA2012637	NEXTGEN	Ovis orientalis	469796	.	Iran
IROO-D6-0002	ERS154527	SAMEA2012638	NEXTGEN	Ovis orientalis	469796	.	Iran
IROO-D6-0005	ERS154530	SAMEA2012641	NEXTGEN	Ovis orientalis	469796	.	Iran
IROO-E3-5492	ERS154533	SAMEA2012643	NEXTGEN	Ovis orientalis	469796	.	Iran
IROO-E5-5146	ERS154531	SAMEA2012642	NEXTGEN	Ovis orientalis	469796	.	Iran
IROO-J11-0602	ERS239060	SAMEA2065602	NEXTGEN	Ovis orientalis	469796	.	Iran
IROO-J11-0905	ERS239061	SAMEA2065603	NEXTGEN	Ovis orientalis	469796	.	Iran
IROO-K7-2301	ERS419581	SAMEA2395411	NEXTGEN	Ovis orientalis	469796	.	Iran
IROO-K7-2303	ERS419580	SAMEA2395410	NEXTGEN	Ovis orientalis	469796	.	Iran
IROO-M12-9997	ERS239022	SAMEA1964491	NEXTGEN	Ovis orientalis	469796	.	Iran
IROO-N13-5061	ERS239063	SAMEA2065604	NEXTGEN	Ovis orientalis	469796	.	Iran
IROV-AB8-1004	ERS403315	SAMEA2358289	NEXTGEN	Ovis vignei	59896	.	Iran
IROV-AB8-1005	ERS403316	SAMEA2358290	NEXTGEN	Ovis vignei	59896	.	Iran
IROV-AB8-1006	ERS403317	SAMEA2358291	NEXTGEN	Ovis vignei	59896	.	Iran
IROV-AD7-1002	ERS403313	SAMEA2358287	NEXTGEN	Ovis vignei	59896	.	Iran
MOOA-J17-1384	ERS154709	SAMEA2012231	NEXTGEN	Ovis aries	9940	local populations	Morocco
MOOA-L15-0414	ERS154724	SAMEA2012246	NEXTGEN	Ovis aries	9940	local populations	Morocco
MOOA-L19-1322	ERS154726	SAMEA2012248	NEXTGEN	Ovis aries	9940	local populations	Morocco
MOOA-M17-1293	ERS154734	SAMEA2012324	NEXTGEN	Ovis aries	9940	D'man	Morocco
MOOA-O11-0271	ERS154746	SAMEA2012336	NEXTGEN	Ovis aries	9940	Sardi	Morocco
MOOA-O9-3220	ERS154751	SAMEA2012340	NEXTGEN	Ovis aries	9940	local populations	Morocco
MOOA-Q12-0163	ERS154764	SAMEA2012431	NEXTGEN	Ovis aries	9940	Sardi	Morocco
MOOA-R10-0005	ERS154762	SAMEA1967786	NEXTGEN	Ovis aries	9940	Boujaad	Morocco
MOOA-R5-0027	ERS154776	SAMEA2012442	NEXTGEN	Ovis aries	9940	local populations	Morocco
MOOA-S11-0138	ERS154782	SAMEA2012524	NEXTGEN	Ovis aries	9940	Timahdite	Morocco
MOOA-T10-0238	ERS154797	SAMEA2012537	NEXTGEN	Ovis aries	9940	Timahdite	Morocco
MOOA-T5-0041	ERS154802	SAMEA2012589	NEXTGEN	Ovis aries	9940	local populations	Morocco
MOOA-T7-3100	ERS154805	SAMEA2012592	NEXTGEN	Ovis aries	9940	local populations	Morocco
MOOA-U10-0242	ERS154810	SAMEA2012596	NEXTGEN	Ovis aries	9940	local populations	Morocco
MOOA-U11-1027	ERS154811	SAMEA2012597	NEXTGEN	Ovis aries	9940	local populations	Morocco
MOOA-U14-1068	ERS154814	SAMEA2012600	NEXTGEN	Ovis aries	9940	D'man	Morocco
MOOA-W8-2287	ERS154838	SAMEA2012853	NEXTGEN	Ovis aries	9940	Beni Guil	Morocco
MOOA-X11-1023	ERS154834	SAMEA2012849	NEXTGEN	Ovis aries	9940	local populations	Morocco
MOOA-Z11-2196	ERS154856	SAMEA2012920	NEXTGEN	Ovis aries	9940	Ouled Djellal	Morocco
MOOA-Z9-2154	ERS154861	SAMEA2012925	NEXTGEN	Ovis aries	9940	Ouled Djellal	Morocco
OARI_BCS3	SRS335706	SAMN01000774	ISGC	Ovis aries	9940	Brazilian Creole	Americas
OARI_BMN4	SRS335671	SAMN01000739	ISGC	Ovis aries	9940	Morada Nova	Americas
OARI_BSI4	SRS335670	SAMN01000738	ISGC	Ovis aries	9940	Santa Inì»s	Americas
OARI_CAS3	SRS335688	SAMN01000756	ISGC	Ovis aries	9940	Castellana	SW Europe
OARI_FIN1	SRS335716	SAMN01000784	ISGC	Ovis aries	9940	Finnsheep	Northern Europe
OARI_GAR4	SRS335736	SAMN01000804	ISGC	Ovis aries	9940	Indian Garole	Asia
OARI_GCN5	SRS335738	SAMN01000806	ISGC	Ovis aries	9940	Gulf Coast native	Americas
OARI_KRS5	SRS335681	SAMN01000749	ISGC	Ovis aries	9940	Karakas	SW Asia
OARI_LAC1	SRS335682	SAMN01000750	ISGC	Ovis aries	9940	Meat Lacaune	SW Europe
OARI_LAC84	SRS335683	SAMN01000751	ISGC	Ovis aries	9940	Mmilk Lacaune	SW Europe
OARI_MER454	SRS335684	SAMN01000752	ISGC	Ovis aries	9940	Merino	SW Europe
OARI_MERC1	SRS335700	SAMN01000768	ISGC	Ovis aries	9940	Merino	SW Europe
OARI_NDZ1	SRS335721	SAMN01000789	ISGC	Ovis aries	9940	Norduz	SW Asia
OARI_SALA2	SRS335674	SAMN01000742	ISGC	Ovis aries	9940	Salz	.
OARI_SBF454	SRS335676	SAMN01000744	ISGC	Ovis aries	9940	Scottish Blackface	Northern Europe
OARI_SUM2	SRS335713	SAMN01000781	ISGC	Ovis aries	9940	Sumatra	Asia
OARI_VBS2	SRS335694	SAMN01000762	ISGC	Ovis aries	9940	Valais Blacknose	Central Europe
OCAN_OCAN1	SRS335678	SAMN01000746	ISGC	Ovis canadensis	37174	Ovis canadensis	.
OCAN_OCAN2	SRS335680	SAMN01000748	ISGC	Ovis canadensis	37174	Ovis canadensis	.
OCAN_OCAN3	SRS335679	SAMN01000747	ISGC	Ovis canadensis	37174	Ovis canadensis	.
ODAL_ODAL1	SRS335696	SAMN01000764	ISGC	Ovis dalli	9943	Ovis dalli	.
ODAL_ODAL2	SRS335717	SAMN01000785	ISGC	Ovis dalli	9943	Ovis dalli	.

TABLE S4. Summary of information unsed to validate the enJSRV insertion scan procedure.

enJSRV	amorces		Postion OAR v4.0	Detected by Whole Genome Survey
	Forward	Reverse		
enJSRV-6	ccagttccagaagggaagggag	caggggaataactggtgctacct	1:239929928-239926832	Y
enJSRV-7	tgtgcacacgtggtgggagtc	actcgagaggaagcacgagggtc	(multiple locus amplified)	N
enJSRV-8	tcagtggatcaatggtctgcga	tgggtgagcatgcacacacg	3:139105810-139106319	Y
enJSRV-15	gtgggaagaaattgcttggtacac	ccataagggtggtatcgtctgca	(Sequence not detected)	N
enJSRV-16	tgctcagtttcccaggtgccca	gtgcaagagctagagctggaagg	UnplacedScaffold_004083582.1:13920-14138	Y
enJSRV-18	gggaagattcgtttcttaggcgctc	atgaaccggactccatggcag	11:32040949-32041591	Y
enJS5F16	ggataagctacactataaaaccaaag	ccatatgtagggattgggggtg	15:42218266-42211217	Y

Matériel supplémentaire - Partie 2 - Chapitre 2 : D'autres variants.

Domestication et convergence évolutive, le cas du gène ASIP

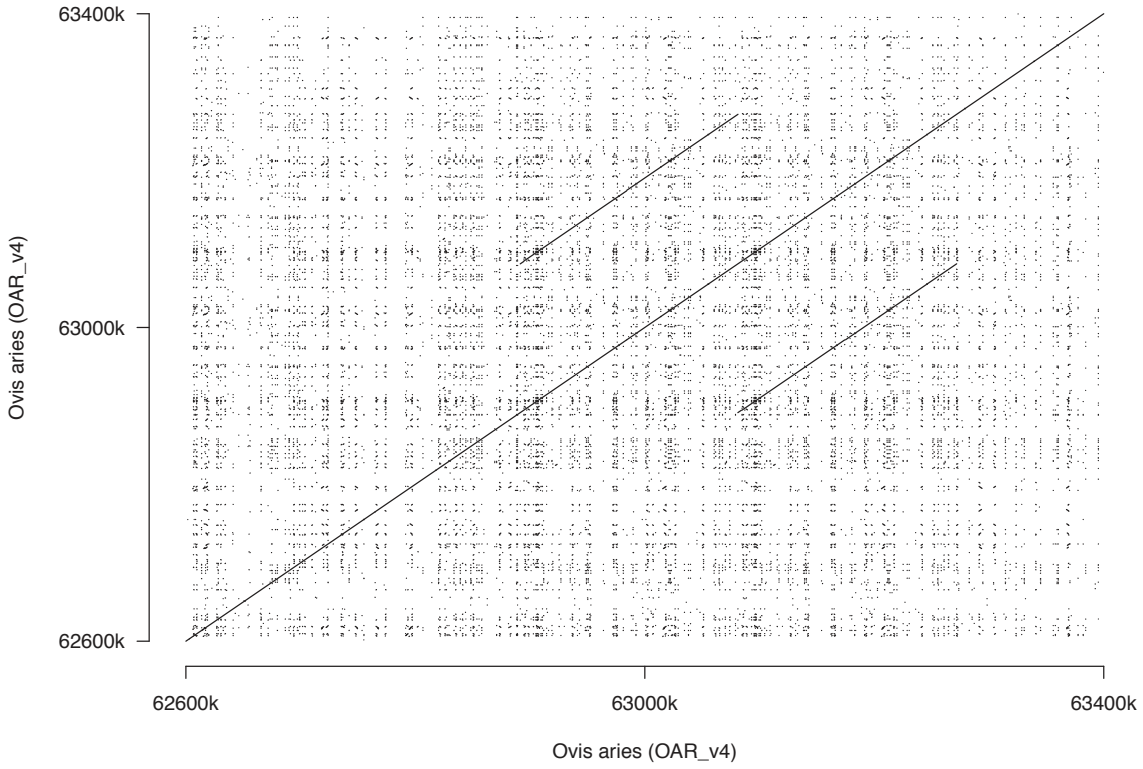


Figure S1 : Graphique de ressemblance (dotplot) de la séquence entre 62800k et 63500k bp du chromosome 13 du génome de référence de la chèvre (ARS1) alignée contre elle même.

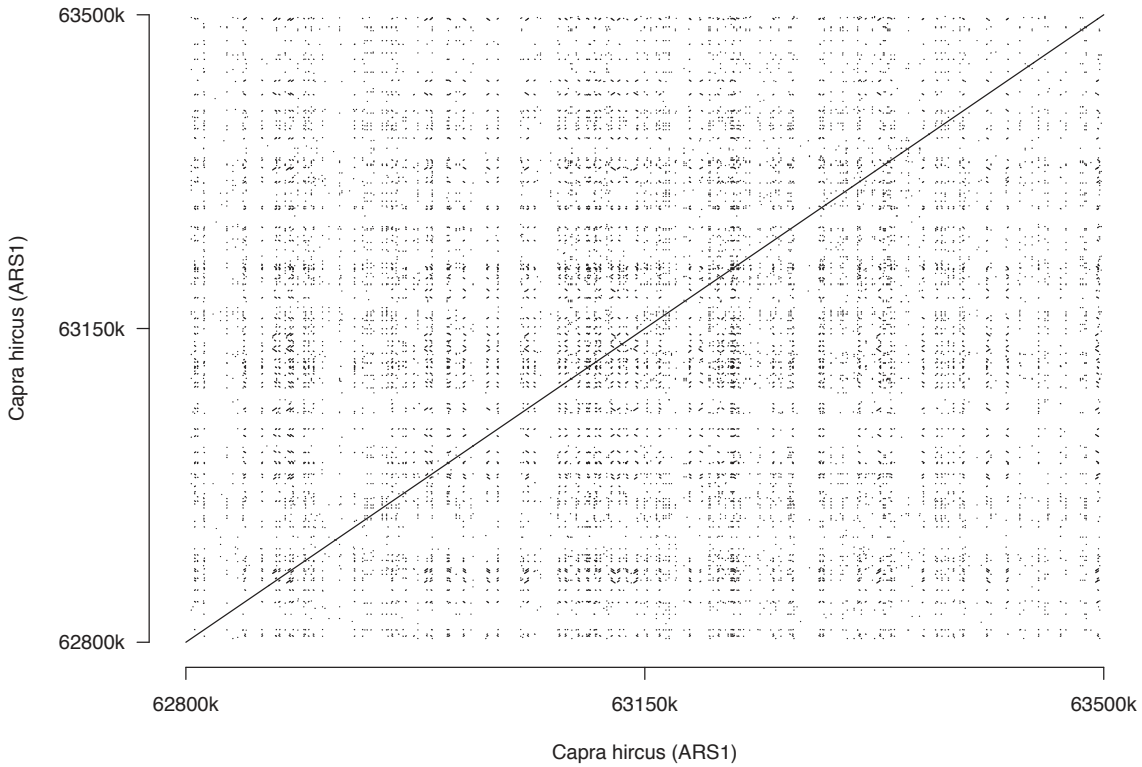


Figure S2 : Graphique de ressemblance (dotplot) de la séquence entre 62600k et 63400k bp du chromosome 13 du génome de référence du mouton (OAR_v4) alignée contre elle même.

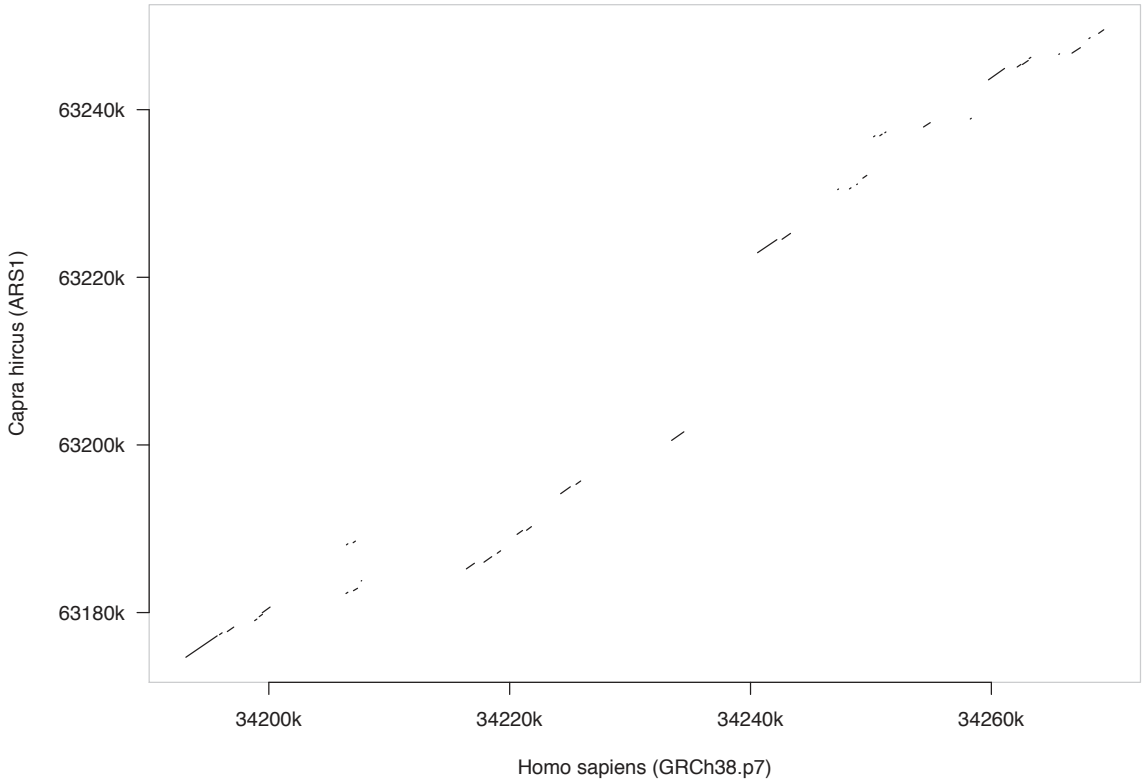


Figure S3 : Graphique de ressemblance (dotplot) de la séquence entre les positions 63172969 et 63249542 du chromosome 13 du génome de référence de la chèvre (ARS1) et la séquence allant des positions 34186493 à 34269344 du chromosome 11 du génome de référence de l'homme (GRCh38.p7).

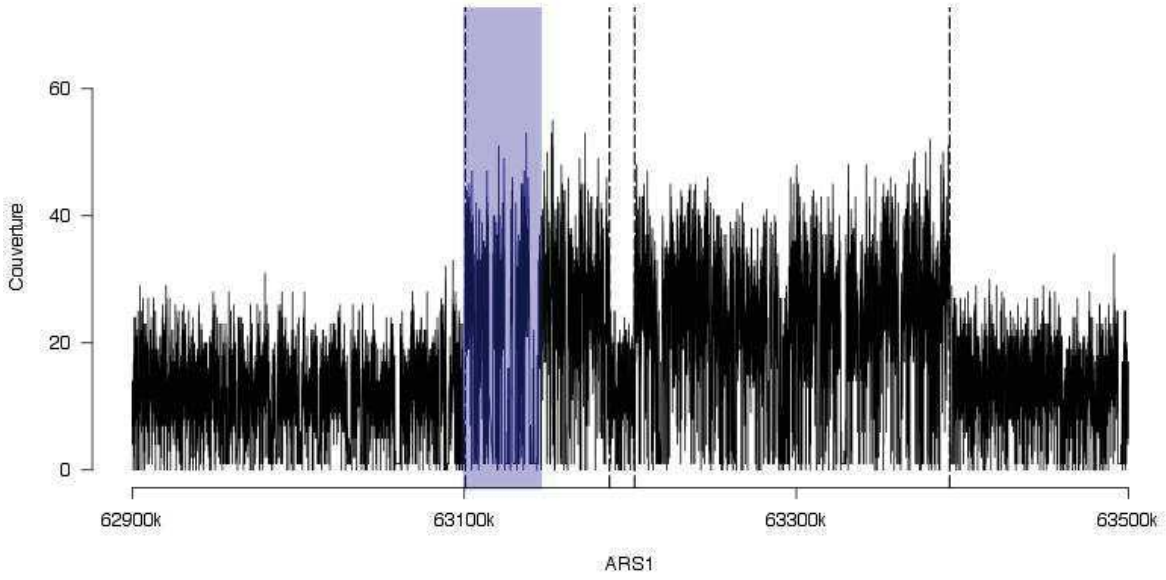


Figure S4 : Évolution de la couverture lors du réalignement d'un génome complet de mouton sur le génome de référence de la chèvre dans la région entourant le gène ASIP du chromosome 13 de la chèvre (ARS1). La région bleue correspond à la zone absente du génome de référence du mouton mais présente dans l'ensemble des individus étudiés ici.

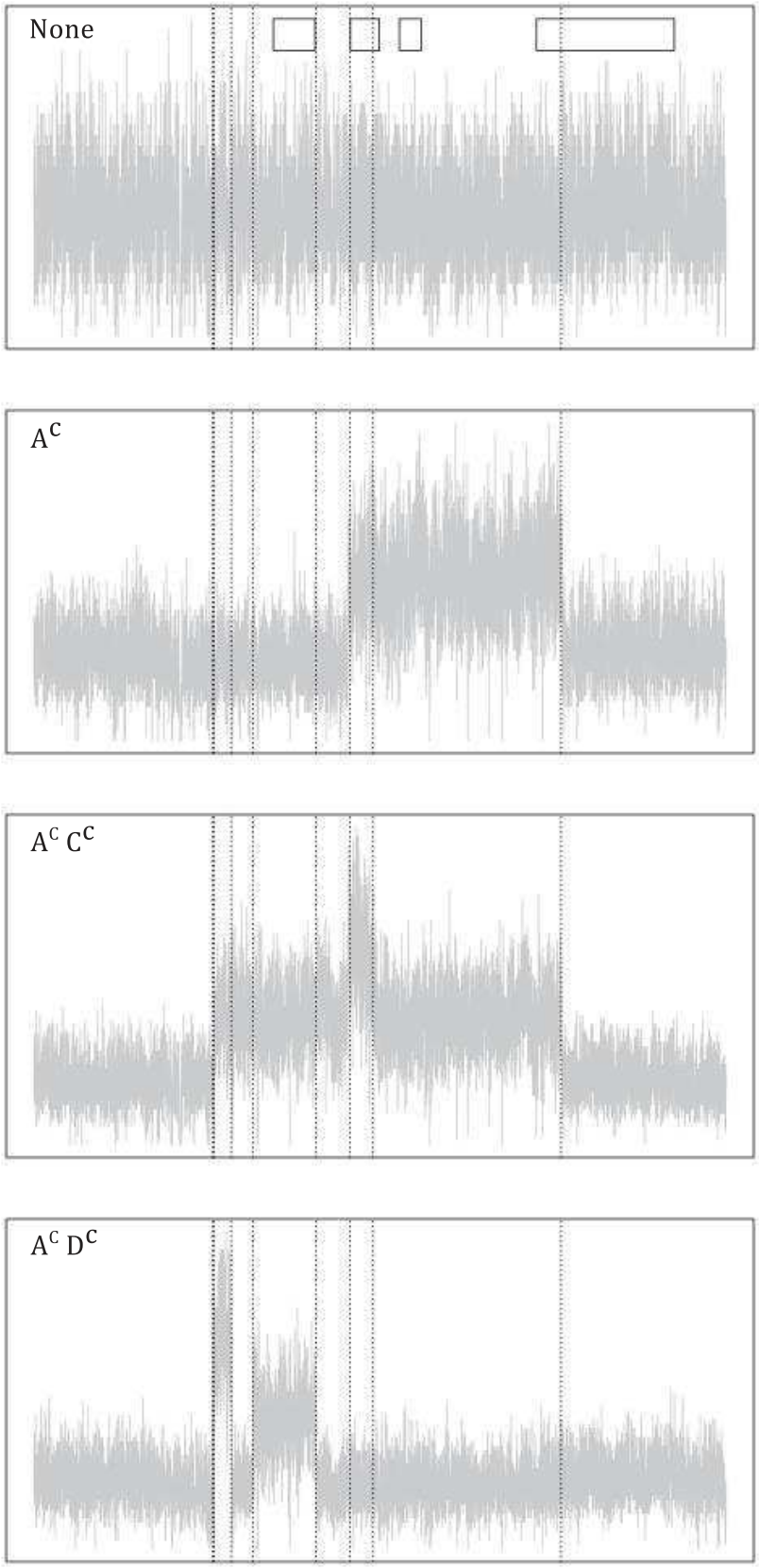


Figure S4 : Évolution de la couverture lors du réalignement de données de séquençage de génomes complets de la chèvre dans la région entourant le gène ASIP du chromosome 13 de la chèvre (ARS1). Les traits verticaux indiquent les bornes de toutes les amplifications possibles. Les notes à gauche de chaque alignement indiquent le code des zones amplifiées.

Table S1 : Liste des individus du genre Ovis utilisés dans cette étude, ainsi que leur génotype inféré. Le nombre pour chaque allèle indique le nombre de copies présentes chez l'individu. Les séquences de tous les individus sont disponibles sur http://projects.ensembl.org/nextgen

Individu	A	B
IROA-B2-5037	2	1
IROA-B2-5296	1.5	1
IROA-B3-5134	2	1
IROA-B4-5190	2	1
IROA-B5-5295	1	1
IROA-B6-5139	1	1
IROA-C3-5212	1	1
IROA-C6-5187	2	1.5
IROA-C7-5042	2	1
IROA-D5-5081	1.5	1.5
IROA-D6-5152	2	1
IROA-D7-5033	2	1
IROA-E5-5157	2.5	1
IROA-E6-5351	2.5	1.5
IROA-E7-5036	2.5	1
IROA-F10-5068	1.5	2
IROA-F3-5142	1.5	1
IROA-F5-5051	2	1.5
IROA-G3-5095	2	1
IROA-G4-5205	2.5	1
IROO-C3-0001	1	1
IROO-C3-2743	1	1
IROO-D6-0002	1	1
IROO-D6-0003	1	1
IROO-D6-0004	1	1
IROO-D6-0005	1	1
IROO-D6-0006	1	1
IROO-D6-5104	1	1
IROO-E3-5492	1	1
IROO-E5-5146	1	1
IROO-F5-5079	1	1
IROO-J11-0602	1	1
IROO-J11-0905	1	1
IROO-K7-0639	1	1
IROO-K7-0642	1	1
IROO-K7-2301	1	1
IROO-K7-2303	1	1
IROO-M12-9997	1	1
IROO-N13-5061	1	1
IROV-AB8-1004	1	1
IROV-AB8-1005	1	1
IROV-AB8-1006	1	1
IROV-AD7-1002	1	1
MOOA-AA10-2191	2	1
MOOA-AA11-2176	3	1
MOOA-AA6-2030	2	1
MOOA-AA6-2036	2	2
MOOA-AA8-2131	2.5	1.5
MOOA-AB10-2186	2	2.5
MOOA-AB11-2161	2	1.5
MOOA-I19-1360	1	1
MOOA-J17-1384	2.5	3
MOOA-J18-1352	1.5	1
MOOA-J19-1336	1.5	1.5
MOOA-K13-0333	1.5	2
MOOA-K14-0398	1	1.5
MOOA-K15-0407	1	1
MOOA-K17-1375	1.5	2.5
MOOA-K18-1345	1	2
MOOA-L11-0291	2	2.5
MOOA-L12-0351	2	2
MOOA-L13-0315	1.5	2
MOOA-L14-0391	1.5	1.5
MOOA-L15-0414	2.5	1
MOOA-L16-1405	1.5	1.5
MOOA-L17-1300	1	2
MOOA-L18-1311	1.5	1
MOOA-L19-1322	1.5	1
MOOA-M10-3200	2	1.5
MOOA-M10-3204	1.5	1.5
MOOA-M11-0276	1.5	1
MOOA-M12-0323	2.5	2
MOOA-M13-0303	2	1.5
MOOA-M14-0421	2.5	1
MOOA-M15-1243	1.5	2
MOOA-M16-1255	1.5	2.5
MOOA-M17-1293	2.5	2
MOOA-M18-1317	1	1
MOOA-M9-3196	1.5	1
MOOA-N10-3188	1.5	1.5
MOOA-N11-0301	3	2
MOOA-N12-0265	1	2
MOOA-N14-0436	2	1.5

MOOA-N15-1238	2	1
MOOA-N16-1259	1.5	1
MOOA-N17-1272	1.5	1
MOOA-NN-9999	1.5	1
MOOA-O10-3264	2.5	2
MOOA-O11-0271	2	2.5
MOOA-O12-0254	2	2
MOOA-O13-0367	2	1
MOOA-O14-1228	2	1
MOOA-O15-1215	2	1
MOOA-O16-1276	1.5	1
MOOA-O9-3220	2	2.5
MOOA-P10-3252	2	2.5
MOOA-P11-0191	2.5	2
MOOA-P13-0362	1.5	1
MOOA-P14-1198	2	1.5
MOOA-P15-1211	1.5	2.5
MOOA-P16-1282	1.5	1
MOOA-P8-3229	2	2
MOOA-P9-3170	2.5	2.5
MOOA-Q10-0058	2.5	2.5
MOOA-Q10-0118	2	1.5
MOOA-Q11-0159	3.5	2
MOOA-Q12-0163	2	2.5
MOOA-Q13-0107	1	3
MOOA-Q14-1195	1	2
MOOA-Q15-1173	1.5	1.5
MOOA-Q8-3248	1.5	1
MOOA-Q9-0169	1.5	2
MOOA-R10-0005	2	2
MOOA-R11-0146	1	1.5
MOOA-R12-0010	2.5	1
MOOA-R12-0013	1.5	1
MOOA-R13-1131	2	2
MOOA-R14-1133	2.5	2
MOOA-R16-1166	1	1
MOOA-R5-0027	2	2.5
MOOA-R6-3022	3	1.5
MOOA-R7-3064	3	2
MOOA-R8-3242	2.5	2.5
MOOA-R9-0196	2.5	2
MOOA-S10-0227	2	1
MOOA-S11-0138	2	2.5
MOOA-S12-1086	3	2
MOOA-S13-1079	1	1
MOOA-S14-0098	2	2
MOOA-S16-1160	2	1.5
MOOA-S4-3009	1.5	2.5
MOOA-S5-3046	2	1.5
MOOA-S5-3051	2.5	1.5
MOOA-S6-3015	1.5	2
MOOA-S6-3018	1	3
MOOA-S7-3083	2.5	2
MOOA-S8-2250	2	1
MOOA-S9-0210	2	2.5
MOOA-T10-0238	4	2
MOOA-T11-0130	1.5	2
MOOA-T11-1040	1.5	1.5
MOOA-T12-1088	1	1
MOOA-T14-1115	2	2
MOOA-T4-3056	1	2
MOOA-T5-0041	2	3
MOOA-T6-0044	2	2
MOOA-T6-3044	1.5	2
MOOA-T7-3100	2	1.5
MOOA-T8-2257	2.5	2
MOOA-T8-2261	2.5	1.5
MOOA-T8-2262	2	1.5
MOOA-T9-0213	2.5	1
MOOA-T9-0218	2	2.5
MOOA-U10-0242	2	2
MOOA-U11-1027	2.5	2
MOOA-U12-0070	1	1
MOOA-U14-1068	1	1.5
MOOA-U5-3029	3	1
MOOA-U6-3273	2	2
MOOA-U7-3116	1.5	2.5
MOOA-U8-2266	2	1
MOOA-U8-2270	1.5	1.5
MOOA-U9-1108	2.5	2
MOOA-V10-1097	2.5	1.5
MOOA-V11-0062	1.5	2
MOOA-V11-0123	2	3
MOOA-V12-1049	4	2.5
MOOA-V13-0128	1	2
MOOA-V5-3137	1	1.5
MOOA-V6-3153	2	1
MOOA-V7-3125	1.5	1.5
MOOA-V8-2274	2	2
MOOA-V8-2279	2.5	1
MOOA-V8-2281	2	2.5
MOOA-V9-1147	2	1.5
MOOA-W10-1157	1	2
MOOA-W12-1057	1.5	1.5
MOOA-W5-3146	1.5	1.5
MOOA-W6-3133	2	1.5
MOOA-W7-3127	2.5	1.5
MOOA-W8-2287	2.5	1

MOOA-X10-2244	1.5	1
MOOA-X10-2305	1.5	1.5
MOOA-X11-1023	2	2
MOOA-X6-2102	1.5	2.5
MOOA-X7-2066	2	2
MOOA-X8-2116	2.5	2
MOOA-X9-2237	2	2
MOOA-X9-2243	2	1
MOOA-Y10-2233	1.5	2
MOOA-Y5-2073	2	2
MOOA-Y5-2077	2	2
MOOA-Y6-2048	1.5	2.5
MOOA-Y7-2063	2.5	2
MOOA-Y8-2142	1.5	1.5
MOOA-Z10-2212	2.5	2
MOOA-Z10-2215	1.5	1.5
MOOA-Z11-2196	2	1.5
MOOA-Z11-2204	2	2
MOOA-Z5-2082	2.5	1.5
MOOA-Z6-2039	1.5	2
MOOA-Z7-2012	2	2
MOOA-Z9-2154	1.5	1.5
MOOA-Z9-2158	2	1
OARI_AFS32	2.5	1.5
OARI_AFS33	2	1.5
OARI_AW454	2.5	1
OARI_AWD1	2.5	2.5
OARI_AWD3	2	2.5
OARI_AWT1	2	1.5
OARI_AWT2	2	1.5
OARI_BCS1	1.5	1.5
OARI_BCS3	2	1
OARI_BGE2	1.5	1
OARI_BGE4	1	1
OARI_BMN3	2	1
OARI_BMN4	3	1
OARI_BSI3	2	1
OARI_BSI4	2	1.5
OARI_CAS1	2	1
OARI_CAS3	2	1
OARI_CC50	2.5	1.5
OARI_CHA02	1.5	1
OARI_CHA05	1	1.5
OARI_CHU1	2	1
OARI_CHU2	2	1
OARI_CHVA1	1.5	2.5
OARI_CHVC1	2	2.5
OARI_DWM1	1.5	2.5
OARI_EMZ1	1.5	1.5
OARI_FIN1	1.5	1
OARI_FIN4	2	2
OARI_GAR14	1	1
OARI_GAR4	1.5	1
OARI_GCN4	1	2.5
OARI_GCN5	2	1.5
OARI_GUR4	2	2
OARI_GUR5	1	1.5
OARI_KR4	1.5	2.5
OARI_KRS3	2	1
OARI_KRS5	1.5	2
OARI_LAC1	2.5	2.5
OARI_LAC84	2	3
OARI_MER454	2	1.5
OARI_MERA1	2.5	3
OARI_MERC1	3	1.5
OARI_NDZ1	1	1
OARI_NDZ4	1.5	2
OARI_NQA11	1.5	1
OARI_OJA4	2	1
OARI_OJA5	2.5	2
OARI_PD454	2.5	3
OARI_RDA2	2.5	1
OARI_RDA4	2	1
OARI_ROM454	2.5	2
OARI_SALA1	2.5	2
OARI_SALA2	2	1
OARI_SALC1	2	1
OARI_SBF454	2.5	2
OARI_SKZ1	1	3
OARI_SKZ4	1	2.5
OARI_SMS2	2.5	2
OARI_SUM2	1.5	1
OARI_SUM7	1.5	1
OARI_SWAA27	1.5	2
OARI_SWAA29	2	2.5
OARI_SWAN3	2	2
OARI_SWAN4	3	1.5
OARI_TEX454	2	2.5
OARI_TWM1	2.5	2
OARI_VBS2	1.5	1
OARI_WHSF1	2.5	2.5
OARI_ZB08	1	1.5
OARI_ZD11	2	2
OCAN_OCAN1	1	1
OCAN_OCAN2	1	1
OCAN_OCAN3	1	1
ODAL_ODAL1	1	1
ODAL_ODAL2	1	1

Table S2 : Liste des individus du genre Capra utilisés dans cette étude, ainsi que leur génotype inféré. Le nombre pour chaque allèle indique le nombre de copies présentes chez l'individu. Les séquences de tous les individus sont disponibles sur http://projects.ensembl.org/nextgen

Individu	A	B	C	D
AUST_Boer_942a	3	2	1	1
AUST_Boer_P439	2	2	1	1
AUST_Cash_E040	1	2.5	1	1
AUST_RL_100a	2	1.5	1	1
AUST_RL_200a	1	1	1	1
FRCH-AL-0001	3.5	1	1.5	1
FRCH-AL-0002	4.5	1	1	1
FRCH-SA-0001	1	3	1	1
FRCH-SA-0002	1	3	1	1
IRCA-C3-1001	1	1	1	1
IRCA-F2-5026	1	1	1	1
IRCA-F2-5064	1	1	1	1
IRCA-F2-5066	1	1	1	1
IRCA-F3-0597	1	1	1	1
IRCA-F3-0600	1	1	1	1
IRCA-G2-0568	1	1	1	1
IRCA-G2-5063	1	1	1	1
IRCA-G2-5065	1	1	1	1
IRCA-I11-0001	1	1	1	1
IRCA-I11-0002	1	1	1	1
IRCA-I11-0003	1	1	1	1
IRCA-I6-5237	1	1	1	1
IRCA-K12-0005	1	1	1	1
IRCA-K7-0009	1	1	1	1
IRCA-M12-0008	1	1	1	1
IRCA-M7-0652	1	1	1	1
IRCA-M7-5041	1	1	1	1
IRCA-M7-5147	1	1	1	1
IRCA-N8-0006	1	1	1	1
IRCA-N8-5141	1	1	1	1
IRCH-B3-5031	1	1.5	1	1
IRCH-B4-5209	1	1	1	1
IRCH-B5-5032	1	1	1	1
IRCH-C3-5039	3	2	1	1
IRCH-C5-5206	3	2	1	1
IRCH-C6-5204	1	1.5	1	1
IRCH-C7-5144	1	1	1	2.5
IRCH-D5-5240	1	1	1	5
IRCH-D6-5189	2	1	1	4
IRCH-D7-5132	1.5	1	1.5	1
IRCH-E5-5053	1	1	1	4.5

IRCH-E6-5087	1	1	1	1
IRCH-E7-5193	3	2	1	1
IRCH-F11-5140	4	2	1	1.5
IRCH-F3-5044	3	1.5	1	1
IRCH-F4-5093	1	1.5	1	1
IRCH-F5-5133	1	1	1	1
IRCH-G3-5210	1.5	1	1	1
IRCH-G4-5194	1	1.5	1	1
IRCH-G5-5185	1	1	1	4
ITCH-SA-0001	1	3	1	1
ITCH-SA-0002	1	3	1	1
ITCH-SA-0003	1	3	1	1
ITCH-SA-0004	1	3	1	1
ITCH-SA-0005	1	2.5	1	1
MOCH-AA10-2195	1	1	1	5.5
MOCH-AA11-2174	1	1	1	7.5
MOCH-AA6-2031	1.5	1	1	1
MOCH-AA6-2034	1.5	1.5	1	1
MOCH-AA7-2026	1	1	1	5
MOCH-AA9-2152	1	1.5	1.5	2.5
MOCH-AB10-2181	2	1	1	3.5
MOCH-AB11-2160	1	1	1	1
MOCH-AB11-2167	1	1	1	8
MOCH-H19-1343	1.5	1.5	1	1
MOCH-J17-1355	1	1	1	1.5
MOCH-J18-1324	1	1	1	1
MOCH-J19-1309	1	1	1	1
MOCH-K13-0366	1	1	1	1
MOCH-K14-0425	1	1	1	1.5
MOCH-K15-0440	1	1.5	1	1
MOCH-K16-1367	1	1	1	1
MOCH-K17-1351	1	1.5	1	1
MOCH-K18-1315	1	1.5	1	1
MOCH-L10-3100	1	1	1	1
MOCH-L12-0379	1	1	1	1
MOCH-L13-0350	1	1	1	1
MOCH-L14-0418	1	1	1	1
MOCH-L15-0443	1	1	1	1
MOCH-L16-1370	1	1	1	1
MOCH-L17-1264	3	1	1	1.5
MOCH-L18-1280	1	1	1	1
MOCH-L19-1290	1	1	1	1
MOCH-M10-3089	1	2	2	1
MOCH-M10-3107	2.5	1.5	1	1
MOCH-M11-0314	1	1	1	1
MOCH-M12-0351	1	1.5	1	1
MOCH-M13-0333	1	1	1	1
MOCH-M14-0455	1	1	1	4.5
MOCH-M15-1213	1	1	1	1

MOCH-M16-1227	1	1	1.5	3.5
MOCH-M17-1262	1.5	1.5	2	1
MOCH-M18-1285	1	1	1	1
MOCH-N10-3078	1	1.5	1	1
MOCH-N11-0330	3	2.5	1	1
MOCH-N12-0292	1	2	1	1
MOCH-N13-0411	1	1.5	1	1
MOCH-N14-0459	2.5	1	2	1.5
MOCH-N15-1209	1.5	1	2	1
MOCH-N16-1228	1	1.5	1	1
MOCH-N16-1231	1	1	1	1
MOCH-N17-1237	2.5	1.5	1	1
MOCH-N9-3126	1	1	1	1
MOCH-NN-9998	1	1	1	4.5
MOCH-NN-9999	1	1	1	4.5
MOCH-O11-0304	4	2	1	1
MOCH-O14-1203	1.5	1	1	4.5
MOCH-O15-1195	2	1	1	1
MOCH-O16-1250	1	1	1	1
MOCH-O8-3153	1	1.5	1	1
MOCH-P10-3074	1	1	1	1
MOCH-P11-0222	1	1.5	1	1
MOCH-P12-0217	3.5	1	1	1.5
MOCH-P12-0387	1	1	1	1
MOCH-P13-0389	2.5	1.5	1	1
MOCH-P14-1175	1	2	1	1
MOCH-P15-1186	1	1	1	1
MOCH-P16-1251	2.5	1.5	1	1
MOCH-P8-3156	1	1.5	1	1
MOCH-Q10-0090	1.5	1	1.5	1
MOCH-Q10-0096	2	1	1	1
MOCH-Q11-0193	2	1.5	1	1
MOCH-Q11-0201	2	1	1	1
MOCH-Q12-0031	1.5	1	1.5	1
MOCH-Q13-0153	1	1.5	1	1
MOCH-Q14-1167	1	1	1	1
MOCH-Q15-1143	1	1	1	1
MOCH-Q16-1147	2	1	1	1
MOCH-Q8-3165	1	2	1	1
MOCH-Q9-0208	1	1	1.5	1
MOCH-R11-0005	1	1	1	1
MOCH-R12-0030	1	1	1	1
MOCH-R12-0195	1	1	1	1
MOCH-R13-1104	1	2	1	1
MOCH-R14-1105	1	1	1	6.5
MOCH-R5-0037	1.5	1.5	1	1
MOCH-R6-3007	2	1.5	1	1
MOCH-R8-3167	2.5	2	1	1
MOCH-R9-0231	1	1	1	1

MOCH-S10-0262	1	1	1	1
MOCH-S11-0188	1	1	1	1
MOCH-S12-1071	1	1	1	1
MOCH-S13-1064	1	1	1	1
MOCH-S15-1165	1.5	1	1	1
MOCH-S16-1135	1.5	1	1	1
MOCH-S4-0026	2	1.5	1.5	1.5
MOCH-S5-0045	1	1.5	1	1
MOCH-S6-0065	1	2.5	1	1
MOCH-S7-3176	1	1.5	1.5	1
MOCH-S7-3179	2	1.5	1	1
MOCH-S8-2252	1	1	1	1
MOCH-S9-0238	2.5	2	1	1
MOCH-T10-0266	1	1.5	1	1
MOCH-T11-1036	2	1	1	1
MOCH-T12-0183	1	1.5	1	1
MOCH-T12-1078	2	1	1	1
MOCH-T13-0128	1	1	1	3
MOCH-T4-3021	1	1.5	1	1
MOCH-T4-3026	1.5	1.5	1.5	1.5
MOCH-T5-0057	1	1	1	1
MOCH-T6-0074	1	1	1	1
MOCH-T8-2258	1	1.5	1	1
MOCH-T8-2261	1	1	1	1
MOCH-T8-2262	1	1.5	1	1
MOCH-T9-0252	1	2	1	1
MOCH-U10-0279	1	1	1	1
MOCH-U11-1029	1	1	1	4.5
MOCH-U12-0113	1	1	1	1
MOCH-U13-1059	4.5	1	1	1.5
MOCH-U14-1058	1	1	1	1
MOCH-U5-3014	1	1	1	1
MOCH-U6-3081	1.5	2	1	1
MOCH-U7-3038	1	1	1	1
MOCH-U8-2266	1	1.5	1	1
MOCH-U8-2270	1	1.5	1	1
MOCH-U9-1088	2.5	2	1	1
MOCH-V10-1083	2.5	1.5	1	1
MOCH-V11-0103	1	1	1	4.5
MOCH-V12-1042	2.5	1.5	1	1
MOCH-V5-3059	1	2	2	1
MOCH-V6-3070	1	1.5	1.5	1
MOCH-V7-3047	2.5	1	1	1
MOCH-V8-2274	1	1	1	1.5
MOCH-V8-2278	1.5	1	1.5	1
MOCH-V8-2280	2.5	2	1	1
MOCH-V9-1114	1	1	1	1
MOCH-W12-1051	1	1	1	4.5
MOCH-W13-1009	1	1	1	5

MOCH-W5-3064	1	1	1	1
MOCH-W6-3053	1	1.5	1	1
MOCH-W7-3048	1	1	1	1
MOCH-W8-2290	1	1.5	1	1
MOCH-W9-2295	2.5	2	1	1
MOCH-X10-2245	1.5	1	1	7
MOCH-X10-2304	1	1	1	1.5
MOCH-X5-2098	1	1	1	4.5
MOCH-X6-2103	1	2.5	1	1
MOCH-X6-2108	1	1	1	1
MOCH-X7-2064	1	1.5	1	1
MOCH-X8-2111	1.5	1	1	1
MOCH-X9-2235	1	1	1	1.5
MOCH-Y10-2228	1	1	1	1
MOCH-Y10-2231	1	1	1	1
MOCH-Y10-2234	1	1.5	1	1
MOCH-Y11-2217	1	1	1	4.5
MOCH-Y6-2048	1	1.5	1.5	1
MOCH-Y6-2051	1.5	1	1	1
MOCH-Y7-2063	1	1	1	1
MOCH-Y8-2141	3	1	1	1
MOCH-Z11-2197	2	1	1	1
MOCH-Z11-2203	1	1	1	1
MOCH-Z5-2083	1.5	1.5	1.5	1
MOCH-Z6-2039	1.5	1	1	1
MOCH-Z6-2044	1	1	1	1
MOCH-Z7-2010	1	1	1	1
MOCH-Z8-2009	1	1	1	4.5
MOCH-Z9-2154	1	1	1	1
MOCH-Z9-2206	1	1	1	4.5

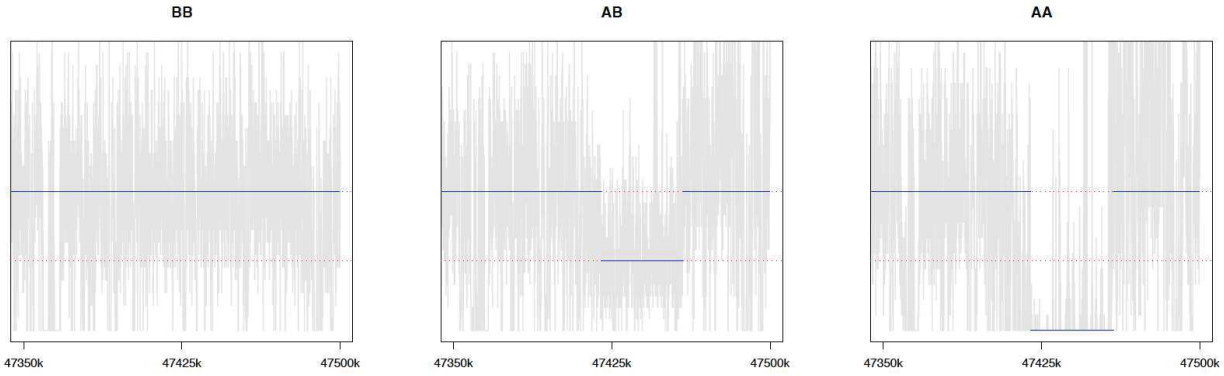


Figure S1 : Évolution de la couverture chez trois individus de moutons et inférence de son haplotype. Les lignes pointillées représentent la couverture normale (la plus haute) et la moitié de la couverture normale (la plus basse) de l'individu. La ligne bleue représente l'évolution de la couverture arrondie aux valeurs possibles selon les trois haplotypes (0 = AA, ½ couverture normal = AB, couverture normale =AA).

Table S1 : Liste des individus utilisés dans cette étude.

Genre	Individu
Capra	IRCA-I11-0002
	IRCA-F2-5064
	IRCA-G2-5065
	IRCA-G2-5063
	IRCA-F2-5026
	IRCA-I6-5237
	IRCA-C3-1001
	IRCA-I11-0003
	IRCA-M7-0652
	IRCA-F3-0600
	IRCA-F3-0597
	IRCA-K12-0005
	IRCA-K7-0009
	IRCA-M12-0008
	IRCA-G2-0568
	IRCA-N8-5141
	IRCA-N8-0006
	IRCA-M7-5147
	IRCA-I11-0001
	IRCA-F2-5066
	IRCA-M7-5041
	IRCH-C6-5204
	IRCH-B3-5031
	IRCH-B4-5209
	IRCH-C5-5206
	IRCH-D6-5189
	IRCH-D5-5240
	IRCH-B5-5032
	IRCH-C7-5144
	IRCH-E5-5053
	IRCH-C3-5039
	IRCH-E6-5087
	IRCH-G3-5210
	IRCH-G4-5194
	IRCH-F5-5133
	IRCH-F4-5093
	IRCH-F3-5044
	IRCH-F11-5140
	IRCH-E7-5193
	IRCH-G5-5185
	IRCH-D7-5132
	MOCH-AA9-2152
	MOCH-AA6-2031
	MOCH-AB11-2160
	MOCH-L19-1290

	MOCH-K18-1315
	MOCH-L10-3100
	MOCH-L12-0379
	MOCH-L14-0418
	MOCH-AA7-2026
	MOCH-L18-1280
	MOCH-M10-3089
	MOCH-L17-1264
	MOCH-L16-1370
	MOCH-N11-0330
	MOCH-L15-0443
	MOCH-M13-0333
	MOCH-M17-1262
	MOCH-M10-3107
	MOCH-M14-0455
	MOCH-N13-0411
	MOCH-N15-1209
	MOCH-N17-1237
	MOCH-N14-0459
	MOCH-N16-1228
	MOCH-N12-0292
	MOCH-M18-1285
	MOCH-N16-1231
	MOCH-N9-3126
	MOCH-NN-9998
	MOCH-M16-1227
	MOCH-M11-0314
	MOCH-M12-0351
	MOCH-P12-0217
	MOCH-NN-9999
	MOCH-Q11-0193
	MOCH-P13-0389
	MOCH-O14-1203
	MOCH-P8-3156
	MOCH-P14-1175
	MOCH-P15-1186
	MOCH-Q10-0096
	MOCH-O11-0304
	MOCH-P16-1251
	MOCH-P10-3074
	MOCH-Q11-0201
	MOCH-Q12-0031
	MOCH-Q13-0153
	MOCH-Q14-1167
	MOCH-O15-1195
	MOCH-O16-1250
	MOCH-P11-0222

	MOCH-Q16-1147
	MOCH-O8-3153
	MOCH-Q8-3165
	MOCH-Q9-0208
	MOCH-R11-0005
	MOCH-S16-1135
	MOCH-R8-3167
	MOCH-S6-0065
	MOCH-S7-3176
	MOCH-R5-0037
	MOCH-T11-1036
	MOCH-T10-0266
	MOCH-S11-0188
	MOCH-S12-1071
	MOCH-T12-0183
	MOCH-S7-3179
	MOCH-P12-0387
	MOCH-S9-0238
	MOCH-S15-1165
	MOCH-R6-3007
	MOCH-S10-0262
	MOCH-R14-1105
	MOCH-S5-0045
	MOCH-R9-0231
	MOCH-R13-1104
	MOCH-R12-0195
	MOCH-S13-1064
	MOCH-S8-2252
	MOCH-T4-3021
	MOCH-S4-0026
	MOCH-R12-0030
	MOCH-T13-0128
	MOCH-T12-1078
	MOCH-T5-0057
	MOCH-T4-3026
	MOCH-U7-3038
	MOCH-U5-3014
	MOCH-U12-0113
	MOCH-T8-2262
	MOCH-T6-0074
	MOCH-V5-3059
	MOCH-T8-2258
	MOCH-U8-2270
	MOCH-U13-1059
	MOCH-U10-0279
	MOCH-V11-0103
	MOCH-U9-1088

	MOCH-T8-2261
	MOCH-T9-0252
	MOCH-U8-2266
	MOCH-V12-1042
	MOCH-V10-1083
	MOCH-V8-2274
	MOCH-U14-1058
	MOCH-V7-3047
	MOCH-V6-3070
	MOCH-U11-1029
	MOCH-U6-3081
	MOCH-V8-2280
	MOCH-V8-2278
	MOCH-V9-1114
	MOCH-X10-2304
	MOCH-W5-3064
	MOCH-Y8-2141
	MOCH-X8-2111
	MOCH-W13-1009
	MOCH-W7-3048
	MOCH-Y10-2228
	MOCH-Y10-2234
	MOCH-W9-2295
	MOCH-Z11-2197
	MOCH-W12-1051
	MOCH-W8-2290
	MOCH-X7-2064
	MOCH-W6-3053
	MOCH-Y7-2063
	MOCH-Y6-2051
	MOCH-X5-2098
	MOCH-Z9-2154
	MOCH-Z5-2083
	MOCH-Y6-2048
	MOCH-Z11-2203
	MOCH-Z8-2009
	MOCH-Z9-2206
	MOCH-Z7-2010
	MOCH-Z6-2039
	MOCH-Z6-2044
	MOCH-AA10-2195
	MOCH-AB11-2167
	MOCH-K15-0440
	MOCH-K17-1351
	MOCH-J18-1324
	MOCH-K14-0425
	MOCH-J17-1355
	MOCH-J19-1309
	MOCH-H19-1343
	MOCH-K16-1367
	MOCH-X9-2235
	MOCH-K13-0366
	MOCH-X6-2108
	MOCH-Y11-2217
	MOCH-X6-2103
	MOCH-Y10-2231
	MOCH-Q10-0090
	MOCH-X10-2245
	MOCH-L13-0350
	MOCH-Q15-1143
	MOCH-AB10-2181
	MOCH-AA11-2174
	MOCH-AA6-2034
	MOCH-M15-1213
	MOCH-N10-3078
Ovis	MOOA-AA6-2030
	MOOA-AA8-2131
	MOOA-AA11-2176
	MOOA-AA10-2191
	MOOA-AB11-2161
	MOOA-AA6-2036
	MOOA-AB10-2186
	MOOA-I19-1360
	MOOA-J17-1384
	MOOA-K15-0407
	MOOA-J18-1352
	MOOA-K18-1345
	MOOA-L13-0315
	MOOA-L12-0351
	MOOA-L11-0291
	MOOA-L15-0414
	MOOA-M10-3200
	MOOA-L16-1405
	MOOA-L19-1322
	MOOA-K17-1375
	MOOA-M11-0276
	MOOA-M10-3204
	MOOA-L17-1300
	MOOA-L18-1311
	MOOA-K13-0333
	MOOA-N11-0301
	MOOA-O11-0271
	MOOA-M13-0303

	MOOA-M18-1317
	MOOA-M12-0323
	MOOA-NN-9999
	MOOA-M9-3196
	MOOA-N17-1272
	MOOA-N14-0436
	MOOA-N16-1259
	MOOA-N12-0265
	MOOA-N10-3188
	MOOA-O10-3264
	MOOA-M15-1243
	MOOA-O13-0367
	MOOA-M16-1255
	MOOA-M17-1293
	MOOA-K14-0398
	MOOA-N15-1238
	MOOA-O14-1228
	MOOA-O9-3220
	MOOA-P10-3252
	MOOA-O16-1276
	MOOA-P11-0191
	MOOA-P8-3229
	MOOA-P15-1211
	MOOA-Q10-0058
	MOOA-P14-1198
	MOOA-P16-1282
	MOOA-Q13-0107
	MOOA-P13-0362
	MOOA-Q14-1195
	MOOA-Q11-0159
	MOOA-Q12-0163
	MOOA-R11-0146
	MOOA-Q10-0118
	MOOA-R10-0005
	MOOA-P9-3170
	MOOA-R12-0010
	MOOA-R14-1133
	MOOA-R16-1166
	MOOA-R5-0027
	MOOA-R7-3064
	MOOA-Q9-0169
	MOOA-R8-3242
	MOOA-S5-3046
	MOOA-S14-0098
	MOOA-S6-3018
	MOOA-S5-3051
	MOOA-Q8-3248

	MOOA-S6-3015
	MOOA-S4-3009
	MOOA-R6-3022
	MOOA-S13-1079
	MOOA-S16-1160
	MOOA-S10-0227
	MOOA-S8-2250
	MOOA-S12-1086
	MOOA-S9-0210
	MOOA-T14-1115
	MOOA-T11-0130
	MOOA-S11-0138
	MOOA-T6-0044
	MOOA-T8-2261
	MOOA-T10-0238
	MOOA-T5-0041
	MOOA-T4-3056
	MOOA-T12-1088
	MOOA-T7-3100
	MOOA-T6-3044
	MOOA-U12-0070
	MOOA-T9-0218
	MOOA-U5-3029
	MOOA-U14-1068
	MOOA-T9-0213
	MOOA-Y10-2233
	MOOA-Y5-2073
	MOOA-Y5-2077
	MOOA-Y8-2142
	MOOA-Z9-2154
	MOOA-Z7-2012
	MOOA-Z9-2158
	MOOA-Z6-2039
	MOOA-Y6-2048
	MOOA-Y7-2063
	MOOA-U11-1027
	MOOA-Z11-2196
	MOOA-V7-3125
	MOOA-V8-2279
	MOOA-Z10-2215
	MOOA-Z11-2204
	MOOA-W10-1157
	MOOA-W5-3146
	MOOA-X10-2244
	MOOA-W8-2287
	MOOA-X9-2237
	MOOA-X6-2102

	MOOA-X8-2116
	MOOA-T8-2262
	MOOA-W7-3127
	MOOA-Z5-2082
	MOOA-W12-1057
	MOOA-W6-3133
	MOOA-X11-1023
	MOOA-V8-2274
	MOOA-X9-2243
	MOOA-V8-2281
	MOOA-Z10-2212
	MOOA-U8-2270
	MOOA-U8-2266
	MOOA-U7-3116
	MOOA-V5-3137
	MOOA-U9-1108
	MOOA-V10-1097
	MOOA-V12-1049
	MOOA-V11-0123
	MOOA-V11-0062
	MOOA-V13-0128
	MOOA-V6-3153
	MOOA-O12-0254
	MOOA-O15-1215
	MOOA-M14-0421
	MOOA-T8-2257
	MOOA-X10-2305
	MOOA-X7-2066
	MOOA-J19-1336
	MOOA-R12-0013
	MOOA-U10-0242
	MOOA-Q15-1173
	MOOA-S7-3083
	MOOA-R13-1131
	MOOA-U6-3273
	MOOA-T11-1040
	MOOA-L14-0391
	MOOA-V9-1147
	MOOA-R9-0196
	IROA-B2-5296
	IROA-F10-5068
	IROA-B2-5037
	IROA-F5-5051

	IROA-B3-5134
	IROA-G4-5205
	IROA-G3-5095
	IROA-F3-5142
	IROA-B4-5190
	IROA-D6-5152
	IROA-C6-5187
	IROA-C7-5042
	IROA-C3-5212
	IROA-E6-5351
	IROA-E7-5036
	IROA-D5-5081
	IROA-D7-5033
	IROA-E5-5157
	IROA-B5-5295
	IROA-B6-5139
	IROO-C3-2743
	IROO-C3-0001
	IROO-D6-0003
	IROO-E3-5492
	IROO-K7-0639
	IROO-D6-5104
	IROO-D6-0004
	IROO-E5-5146
	IROO-D6-0006
	IROO-F5-5079
	IROO-K7-2303
	IROO-K7-0642
	IROO-D6-0005
	IROO-J11-0602
	IROO-N13-5061
	IROO-M12-9997
	IROO-K7-2301
	IROO-J11-0905
	IROO-D6-0002
	IROV-AB8-1006
	IROV-AB8-1005
	IROV-AD7-1002
	IROV-AB8-1004

Matériel supplémentaire - TROISIÈME PARTIE : Variants structuraux – Approche sans *a priori*

Matériel supplémentaire - Article 3 : BAdabouM: a structural variation discovery tool

1.1 Structural Variations detection

BAdabouM implements a sliding window running along the genome looking for specific signatures. This sliding window is divided in three parts, allowing detecting, on both sides parts abnormally mapped pairs and aberrant coverage, and split reads in the middle one (Figure S1).

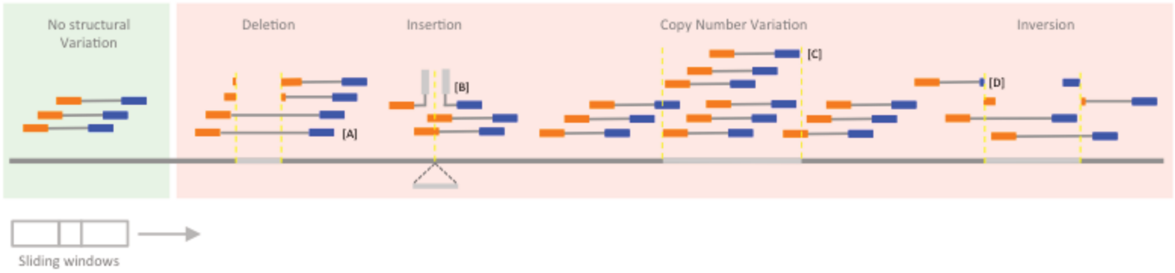


Figure S1 : schematic representation of the sliding windows and specific signature of different SVs. To call SVs, BAdabouM detects incoherent alignment. Space between pairs of reads is more important than normal fluctuation [A]. Pairs of read whose one of the read is not mapped close to his partner (also called dangling) [B]. Too many or too few reads are aligned to sequence [C]. Split-reads confirm event and allow precise placement of breakpoint [D].

1.2 Workflow

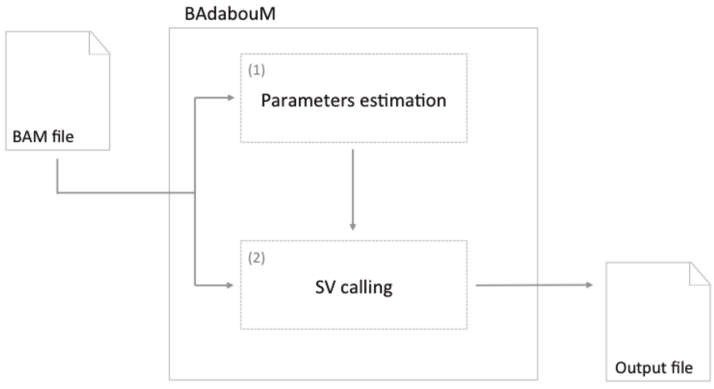


Fig S2 : schematic representation of the BAdabouM workflow.

1.3 Implementation and distribution

BAdabouM is a Cpp software distributed under CeCILL.
BAdabouM is freely available at http://github.com/cumtr/badaboum

Table S1: List of structural variations highly differentiated between domestics and wilds sheep and goats.

	Chr	Start	Stop	SV Type	DI	Fst	SNP	Genes
Capra	1	16976363	16979242	DEL	0,586	0,390		
	2	86202222	86203704	DEL	0,592	0,297		
	2	120008999	120010940	DEL	-0,703	0,296		
	3	2911050	2913151	DEL	0,694	0,286		
	3	28705021	28706244	DEL	-0,667	0,268		CDCP2
	3	32086569	32087116	DEL	0,561	0,312		DAB1
	3	66416764	66417253	DEL	0,744	0,334		
	3	66440617	66442167	DEL	0,719	0,309	X	LOC102174264
	3	66496268	66501875	DEL	0,675	0,274		LOC102180934
	3	66750305	66752505	DEL	0,694	0,286		LOC108633177
	4	20750930	20752684	DEL	0,675	0,274		
	4	44024708	44025316	DEL	-0,672	0,268		
	4	70559139	70560351	DEL	-0,758	0,355		
	4	84803016	84810829	DEL	0,678	0,333		SEMA3D
	4	99827760	99828733	DEL	-0,614	0,278		
	5	5082186	5082798	DEL	0,664	0,266		
	5	18085186	18086810	DEL	0,850	0,442	X	KITLG
	5	34378974	34380021	DEL	0,592	0,297		
	5	37745694	37747062	DEL	-0,667	0,268		PPHLN1
	5	51938751	51940456	DEL	0,700	0,292		
	6	21867632	21869268	DEL	0,628	0,264		
	6	30214611	30238961	DEL	0,775	0,357		BMPR1B
	6	63632820	63636363	DEL	-0,758	0,355		KCTD8
	7	41045092	41048727	DEL	-0,600	0,632		
	8	65027857	65130061	INV	0,653	0,296		LOC102188314
	9	28845462	28846015	DEL	0,664	0,266		FOXO3
	9	71769577	71770746	DEL	0,744	0,334		
	10	30675872	30678546	DEL	-0,769	0,462		PPM1A
	10	64257602	64258225	DEL	0,653	0,296		
	10	76159923	76160790	DEL	0,850	0,442		

	10	77040729	77041733	DEL	-0,708	0,315		
	10	96077579	96078046	DEL	-0,633	0,260		
	10	98779403	98781307	DEL	0,597	0,257		
	11	73723878	73724297	DEL	-0,633	0,260		DNMT3A
	11	77823513	77824250	DUP	0,750	0,333		LDAH
	11	77860097	77861697	DEL	0,775	0,357		LDAH
	12	80860412	80863062	DEL	0,669	0,266		
	12	80872046	80873212	DEL	0,669	0,266		
	13	68879514	68880091	DEL	-0,678	0,277		
	14	5761599	5762230	DEL	-0,672	0,268		
	14	47817802	47817912	INS	-0,664	0,314		
	15	888894	889413	DEL	0,683	0,304		
	16	4750235	4750816	DEL	0,714	0,316		
	17	36677696	36678245	DEL	0,664	0,266		SPATA5
	17	50990468	50991021	DEL	-0,569	0,308		
	18	51655040	51655585	DEL	0,675	0,274		
	18	55987467	55990998	DEL	-0,814	0,433		
	20	3801785	3802183	DEL	-0,722	0,310		FBXW11
	20	15190282	15192537	DEL	-0,664	0,314		RNF180
	21	19348176	19349072	DEL	-0,683	0,296		LOC102180801
	21	19424294	19499375	INV	0,694	0,286		LOC102181632 LOC102181900 LOC102182171 LOC102182732
	21	19464041	19465415	DEL	0,764	0,381		
	21	19535350	19536098	DEL	0,703	0,379		
	21	19540035	19540992	DEL	0,678	0,333		
	21	19562825	19563664	DEL	0,678	0,333	X	
	21	45613963	45614657	DEL	0,678	0,333		
	21	55788851	55789224	DEL	0,669	0,266		CATSPERB
	22	26315819	26317493	DEL	0,622	0,293		
	23	904927	905387	DEL	0,775	0,357		GMDS
	23	41160675	41161209	DEL	-0,650	0,675		
	23	41249400	41249895	DEL	-0,664	0,314		VPS52
	24	43889095	43889764	DUP	0,783	0,484		
	25	33220182	33220292	INS	-0,614	0,278		

	25	37037938	37170870	INV	0,750	0,333		PVRIG STAG3 GPC2 GAL3ST4 C25H7orf43 LAMTOR4 LOC108633863
	26	603005	603456	DEL	-0,669	0,382		
	28	25503410	25504668	DEL	-0,719	0,421		
	29	25608705	25609798	DEL	-0,672	0,268		TSG101
	29	41130020	41130130	INS	-0,833	0,420		
	29	46285276	46285771	DEL	-0,975	0,975		
	30	33128961	33359953	INV	0,642	0,434		
Ovis	1	41438288	41438838	DEL	0,611	0,410		
	1	43284348	43284891	DEL	-0,600	0,632		
	1	93886322	93888694	DEL	0,750	0,333		
	1	102404087	102404188	INS	-0,709	0,366		
	1	127942057	127942538	DEL	-0,684	0,347		CYYR1
	1	130341176	130341978	DEL	-0,768	0,358		
	1	131164003	131165519	DEL	0,784	0,397		
	1	134214858	134215647	DEL	-0,650	0,675		
	1	164575774	164576759	DEL	-0,750	0,764		
	1	181373999	181374655	DEL	-0,759	0,405		
	1	210628960	210629378	DEL	0,775	0,357		
	1	229039060	229047720	DEL	0,750	0,333		
	1	229047753	229048848	DEL	0,825	0,410		
	1	237108325	237109780	DEL	-0,759	0,405		
	1	264662049	264663112	DEL	-0,600	0,632		
	2	5898879	5899856	DEL	-0,650	0,675		
	2	9053457	9055190	DEL	-0,650	0,675		
	2	9715163	9715656	DEL	-0,600	0,632		ZNF618
	2	78782822	78786698	DEL	-0,750	0,764		
	2	114943019	114944648	DEL	-0,600	0,632		

	2	115288630	115296732	DEL	-0,675	0,697			
	2	118646293	118647186	DEL	-0,975	0,975		SLC40A1	
	2	121200441	121200953	DEL	-0,743	0,337			
	2	122382060	122384088	DEL	-0,909	0,537		LOC101122056	
	2	152320794	152321973	DEL	-0,625	0,654			
	2	155873231	155873830	DEL	-0,925	0,926		FMNL2 LOC101117395	
	2	166999200	167000476	DEL	0,759	0,363		LRP1B	
	2	198386109	198386592	DEL	-0,825	0,832		ANKRD44	
	2	199692868	199693533	DEL	-0,793	0,380			
	2	204654442	204654847	DEL	-0,825	0,832		CD28	
	3	28371757	28373019	DEL	0,818	0,826			
	3	110244754	110245430	DEL	-0,734	0,385			
	3	119500827	119500928	INS	-0,709	0,366			
	3	173397833	173398800	DEL	-0,600	0,632			
	3	181834549	181835883	DEL	-0,825	0,832		FGD4	
	3	185630850	185631877	DEL	0,875	0,478			
	4	35806392	35807879	DEL	0,925	0,570			
	4	47459828	47460391	DEL	-0,684	0,347			
	4	50724447	50725317	DEL	-0,759	0,405		CFTR	
	4	69624592	69625143	DEL	-0,809	0,447			
	4	70626208	70626809	DEL	0,759	0,363		C4H7orf31	
	5	22382550	22385494	DEL	-0,650	0,675			
	5	30508126	30508830	DEL	-0,709	0,366			
	5	74299611	74301463	DEL	0,800	0,382			
	5	75541015	75541584	DEL	-0,625	0,654			
	5	78545659	78546395	DEL	0,884	0,641			
	5	89679555	89681796	DEL	-0,800	0,809			
	5	97757044	97757640	DEL	-0,600	0,632			
	5	106924559	106925517	DEL	-0,925	0,926			

	5	107109446	107109884	DEL	-0,818	0,402	X	
	6	9684925	9689428	DEL	-0,684	0,347		
	6	25182141	25189870	DEL	0,734	0,332		
	6	34174696	34175129	DEL	-0,743	0,337		
	6	46016093	46016998	DEL	0,677	0,398		
	6	48273467	48274097	DEL	-0,750	0,764		
	6	61420957	61421417	DEL	-0,625	0,654		
	6	69942598	69944240	DEL	-0,650	0,675		
	6	86540005	86540530	DEL	-0,625	0,654		GC
	6	88318439	88319553	DEL	-0,775	0,786		
	7	6158343	6159426	DEL	-0,834	0,468		
	7	16641473	16642093	DEL	-0,743	0,337		
	7	18155512	18156585	DEL	-0,784	0,426		THSD4
	7	24531817	24533049	DEL	0,734	0,332		
	7	34487885	34489735	DEL	-0,743	0,337		
	7	60687624	60689715	DEL	-0,825	0,832		
	7	83541841	83542648	DEL	-0,809	0,447		
	8	8201539	8203314	DEL	-0,600	0,632		
	8	8752059	8753213	DEL	0,743	0,411		
	8	12141555	12144154	DEL	-0,734	0,385		
	8	22323217	22325142	DEL	-0,625	0,654		
	8	23660948	23661447	DEL	-0,625	0,654		
	8	38460515	38461467	DEL	-0,684	0,347		
	8	66738489	66739529	DEL	-0,884	0,513		ADGRG6
	8	88092865	88093357	DEL	-0,600	0,632		RPS6KA2
	8	89945094	89945681	DEL	0,734	0,332		WDR27
	9	6752386	6752886	CTX	0,750	0,333		
	9	42801398	42802170	DEL	0,800	0,382		MTFR1
	9	43874328	43883923	DEL	0,800	0,382		COPS5
	9	43885207	43886335	DEL	0,784	0,397		COPS5

	9	54568234	54568675	DEL	-0,759	0,405		
	9	60206759	60208285	DEL	-0,734	0,385		LOC106991339
	9	68926344	68927109	DEL	0,586	0,330		RSPO2
	10	37008030	37009035	DEL	-0,600	0,632		
	10	37795740	37796447	DEL	-0,884	0,513		
	10	69089053	69092359	DEL	-0,709	0,366		GPC6
	11	23219139	23219833	DEL	-0,650	0,675		
	11	32041138	32049055	DEL	-0,775	0,786		LOC106990188
	12	11633236	11634339	DEL	-0,725	0,741		
	12	17503146	17504887	DEL	0,800	0,382		USH2A
	12	32681287	32683647	DEL	-0,709	0,366		PLD5
	12	35996788	35998069	DEL	-0,625	0,654		
	13	765824	766209	DEL	-0,743	0,337		
	13	24011415	24011897	DEL	0,652	0,341		
	13	49039438	49041911	DEL	-0,650	0,675		
	13	49052920	49054518	DEL	-0,650	0,675		
	13	49708903	49711796	DEL	-0,784	0,426		
	13	49814583	49814960	DEL	-0,793	0,380	X	
	13	50418909	50419668	DEL	-0,818	0,402	X	
	13	62586071	62587200	DEL	-0,600	0,632		
	14	58254631	58255129	DEL	-0,750	0,764		
	15	383111	384673	DEL	-0,768	0,358		
	15	3289576	3291521	DEL	-0,909	0,537		
	15	3842434	3843373	DEL	-0,900	0,902		
	15	11015022	11015698	DEL	0,652	0,341		
	15	11947034	11948747	DEL	-0,743	0,337		
	15	16223812	16223913	INS	-0,750	0,764		
	15	21909851	21910694	DEL	-	0,468	X	

	16	34728997	34730533	INV	-0,834	0,675		
	16	38320241	38320883	DEL	-0,650	0,402		
	16	43289371	43289959	DEL	-0,818	0,719		
	16	64839501	64847073	DEL	-0,700	0,405		
	17	33049151	33051799	DEL	-0,759	0,570		
	18	2950342	2951996	DEL	0,925	0,764		
	18	23071399	23072905	DEL	-0,750	0,809		
	18	65928541	65930551	DEL	-0,800	0,741		
	19	29525711	29527320	DEL	-0,725	0,478		
	20	10858507	10859246	DEL	0,875	0,332		C20H6orf89
	20	18044321	18045690	DEL	0,734	0,366		
	20	21474599	21475013	DEL	-0,709	0,337		
	21	6843033	6848718	DEL	-0,743	0,537		
	21	41086261	41095046	DEL	-0,909	0,402		
	21	41096167	41096988	DEL	-0,818	0,402		
	21	41106439	41106983	DEL	-0,818	0,380		
	21	41111738	41120889	DEL	-0,793	0,402		
	23	7346515	7348166	DEL	-0,818	0,347		CD226
	23	21385982	21387844	DEL	-0,684	0,570		C23H18orf21
	23	21388034	21389329	DEL	0,925	0,478		C23H18orf21
	23	37779630	37792772	DEL	0,875	0,330		
	23	40569983	40572058	DEL	0,586	0,632		
	24	24278103	24279511	DEL	-0,600	0,764		
	24	41202990	41204124	DEL	-0,750	0,570		
	24	41943554	41943959	DEL	0,925	0,697		
	25	13190796	13191990	DEL	-0,675	0,741		
	26	18341714	18342302	DEL	-0,725	0,347		PDGFRL
	26	31296611	31297383	DEL	-0,684	0,405		LOC105605231
					-0,759			

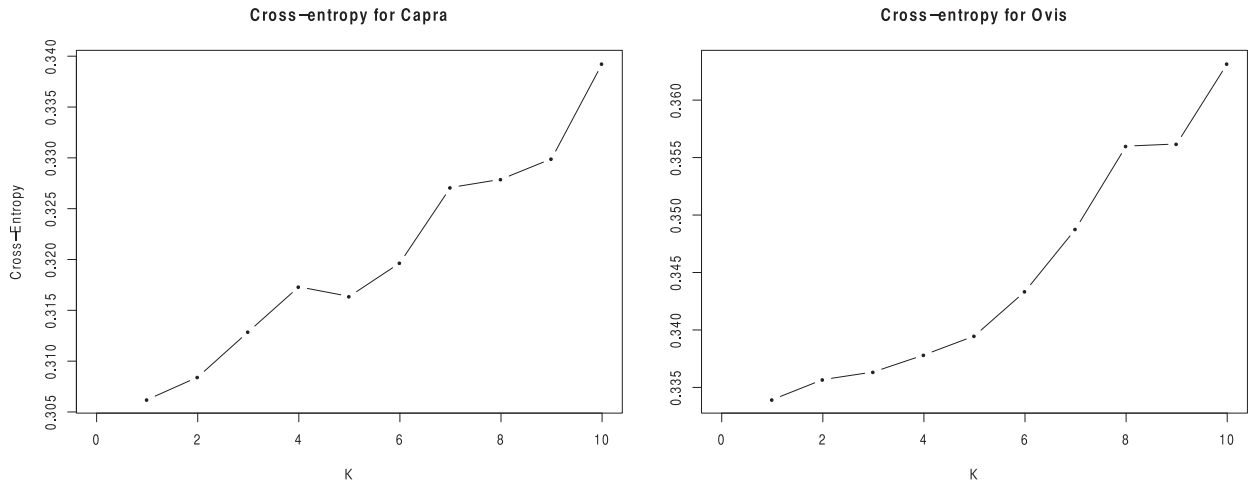


Figure S1 : sNMF Cross-Entropy score according to different values of K. K=1 is the lowest for both sheep and goat.