INRA Prod. Anim.,
2016, 29 (5), 339-350

Anticiper l'émergence d'anomalies génétiques grâce aux données génomiques

S. FRITZ[1,2], P. MICHOT[1,2], C. HOZE[1,2], C. GROHS[1], A. BARBAT-LETERRIER[1], M. BOUSSAHA[1], D. BOICHARD[1], A. CAPITAN[1,2]

[1] GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350, Jouy-en-Josas, France
[2] Allice, Maison Nationale des Éleveurs, 75595, Paris, France

Courriel : sebastien.fritz@allice.fr

En complément des approches du phénotype au génotype (ou « *bottom-up* »), les nouvelles technologies d'étude du génome, et l'analyse des données à haut débit qu'elles engendrent, permettent de mettre en évidence des anomalies avec une approche « *top-down* », c'est-à-dire avant que l'on ne détecte leur effet sur le phénotype.

Duchesne *et al* (2016 ce numéro) ont montré comment l'approche ascendante du phénotype au gène est devenue très efficace pour identifier les mutations responsables d'un phénotype d'anomalie monogénique, grâce aux nouvelles technologies de génotypage et de séquençage. Cet article a également montré les conditions d'efficacité de cette approche, en particulier la disponibilité de phénotypes précis et d'échantillons biologiques. Grohs *et al* (2016 ce numéro) ont montré l'importance et le rôle de l'Observatoire National des Anomalies Bovines (ONAB) pour satisfaire cette condition essentielle. Toutefois, lorsque les phénotypes ne sont pas facilement disponibles, cette approche ne peut pas être mise en œuvre, par exemple quand les phénotypes sont très peu spécifiques, ou peu visibles dans des conditions d'élevage, ou quand l'anomalie conduit à une mortalité précoce en gestation, se traduisant uniquement par une baisse de fertilité.

Nous présentons dans cet article deux approches complémentaires largement basées sur la fouille de données à haut débit, permettant en partie de contourner ces difficultés.

1 / Recherche de mutations responsables de mortalité embryonnaire par déficit d'homozygotes

1.1 / Recherche des régions présentant des déficits en homozygotes

Une anomalie conduisant à une mortalité précoce durant la gestation se traduit par une baisse de fertilité et passe généralement inaperçue. Ce type d'anomalie est sans doute assez fréquent car de nombreux gènes, critiques pour le développement, peuvent en être responsables. Les échantillons biologiques sont difficiles ou impossibles à collecter.

L'approche proposée a été mise en œuvre chez les bovins laitiers pour lesquels la fertilité est un problème majeur. Dans les trois principales races françaises (Holstein, Montbéliarde et Normande), une baisse du taux de conception a été observée jusqu'en 2005 (Barbat *et al* 2010). La réponse corrélée à la sélection sur la production laitière n'explique que la moitié de cette évolution, il existe donc d'autres facteurs responsables. Les races bovines laitières sont des populations récentes issues d'un faible nombre d'ancêtres fondateurs et l'augmentation continue de la consanguinité contribue à l'émergence régulière d'anomalies génétiques récessives, en ségrégation dans les populations. Certaines de ces anomalies sont visibles sur des individus nés, mais on peut supposer que d'autres conduisent à une mortalité embryonnaire à l'état homozygote. La baisse de fertilité associée est faible à l'échelle de la population et même à l'échelle d'un reproducteur, donc difficile à discerner, elle n'est très visible que dans les accouplements à risque entre parents porteurs.

Dans ces trois races, le dispositif disponible se prête très bien au type d'analyse proposé, tant en termes de génotypage que de séquençage. Profitant de l'essor de la sélection génomique et du nombre important d'animaux génotypés chaque année, VanRaden *et al* (2011) ont proposé de rechercher les régions du génome présentant un déficit d'individus homozygotes. Il est alors fait l'hypothèse que ces régions, ou tout au moins certaines d'entre elles, portent des mutations induisant une mortalité embryonnaire à l'état homozygote. La disponibilité en France de grandes populations d'individus génotypés a permis la mise en œuvre d'une approche similaire (Fritz *et al* 2013) et la détection de « *Quantitative Trait Loci* » (QTL) responsables de mortalité embryonnaire. Par ailleurs, le séquençage du génome d'un nombre croissant d'individus permet la recherche des mutations causales correspondantes, leur identification facilitant ensuite leur contre-sélection.

Le principe de l'approche est résumé sur la figure 1. Une mutation se produit sur un chromosome unique d'un ancêtre. Ce chromosome est caractérisé par les allèles de marqueurs qu'il porte et la mutation est associée à la combinaison allélique (ou haplotype) des marqueurs qui l'entourent. La mutation se répand dans la population, entourée du même haplotype qui est progressivement raboté par les recombinaisons au cours des méioses successives. Les embryons homozygotes pour la mutation et donc homozygotes pour l'haplotype meurent. La population typée montre alors un déficit, voire une absence, d'individus homozygotes.

Le choix de l'haplotype n'est pas neutre dans les résultats de l'analyse. En effet, s'il est trop court (exprimé en nombre de marqueurs successifs), il est peu informatif car le fragment a de fortes chances d'exister aussi, parfois en

forte fréquence dans la population, sans posséder la mutation. Dans ce cas, l'association haplotype – mutation n'est pas suffisante pour détecter la disparition des homozygotes à la mutation. Mais si le segment chromosomique correspondant est trop long, il n'est pas suffisamment conservé dans la population, et la mutation se retrouve associée à plusieurs haplotypes, celui d'origine mais également à d'autres haplotypes recombinants, nuisant à la précision et à la puissance de l'analyse. Le bon compromis dépend de la structure de la population et de l'âge de la mutation : plus la mutation est récente ou plus l'effet fondateur associé à la mutation est récent, plus l'haplotype conservé est long. Il dépend aussi bien sûr de la densité de marqueurs car disposer d'un haplotype assez court mais très informatif suppose une densité de marqueurs élevée. Lorsque la densité de marqueurs est insuffisante, cette approche n'est plus applicable. Les analyses bovines ont utilisé des typages réalisés avec la puce Bovine SNP50 Beadchip® d'Illumina. Avec un marqueur informatif tous les 60 kilobases (kb) environ et des haplotypes de 10 à 20 marqueurs, la taille des haplotypes a varié entre 500 kb et 2 Mégabases (Mb) environ. Avec la puce haute densité (777k, un marqueur tous les 4,5 kb environ), on pourrait détecter des haplotypes plus petits, de l'ordre de 50 à 100 kb et donc des mutations plus anciennes ; et à l'échelle de la séquence, on pourrait détecter des haplotypes de quelques kb. Cependant, les effectifs requis de populations génotypées ne nous ont pas permis de réaliser les analyses à ces densités élevées.

L'étude initiale de Fritz *et al* (2013) a utilisé les typages de 47 878 individus de race Holstein, 16 833 Montbéliards et 11 466 Normands, génotypés et issus d'un père et d'un grand-père maternels eux-mêmes typés. Aucune information phénotypique n'était requise. Seuls les 29 chromosomes autosomaux ont été considérés et 43 801 marqueurs ont été utilisés, satisfaisant différents critères de contrôle qualité (pourcentage de typages présents supérieur à 95%, fréquence de l'allèle minimal supérieur à 0,03 dans au moins une race, équilibre de Hardy-Weinberg). Ces typages ont été phasés, c'est-à-dire que les allèles reçus du père et de la mère ont été déterminés, de façon à reconstruire les chromosomes de chaque individu et disposer ainsi de l'information haplotypique sur tout le génome. À l'époque, ce phasage a été réalisé avec le logiciel « *DagPhase* » (Druet et Georges 2010), il est réalisé aujourd'hui avec « *FImpute* » (Sargolzaei *et al* 2014). Cette étape permet également d'imputer la faible proportion de typages manquants et donc de disposer d'une information complète.

L'analyse est ensuite conduite de la façon suivante. Les haplotypes de 20 marqueurs constituent une fenêtre glissante que l'on déplace d'un marqueur à la fois le long du génome. La détection est réalisée intra-race car les haplotypes de cette taille ne sont pas conservés entre races et n'auraient aucune chance de porter la même mutation entre races. À chaque position, la fréquence de chaque combinaison haplotypique (ou « allèle ») observée est calculée sur les phases d'origine maternelle. Seules les combinaisons présentant une fréquence supérieure à 1% sont conservées.

Pour chaque combinaison d'haplotype k conservée, le nombre observé O(k) d'individus homozygotes est comparé au nombre attendu sous l'hypothèse de neutralité. Ce nombre attendu E(k) est estimé en tenant compte de la structure de pedigree réelle, par la formule suivante :

$$E(k) = \sum_{i=1}^{n_p} p_{ik} \sum_{j=1}^{n_{gpm}} 0{,}5(q_{jk} + f_k) \quad n_{ij}$$

avec E(k) le nombre attendu d'individus homozygotes à l'haplotype k, n_p le nombre de pères, n_{gpm} le nombre de grands-pères maternels, p_{ik} la probabilité de transmission de l'haplotype k par le père i à ses descendants, q_{jk} la probabilité de transmission de l'haplotype k par le grand-père maternel j à ses filles, f_k la fréquence de l'haplotype k dans la population, n_{ij} le nombre de descendants issus du père i et du grand-père maternel j. La probabilité de transmission de l'haplotype k (p_{ik} ou q_{jk}) est égale à 1 pour un taureau homozygote à l'haplotype k, 0,5 pour un taureau hétérozygote et 0 pour un taureau non porteur. Un QTL est détecté lorsque O(k) est significativement inférieur à E(k) par un test du χ^2. La région finale de localisation est obtenue en concaténant toutes les fenêtres consécutives présentant les mêmes estimations.

Dans le cas où la mère est typée, on gagne en puissance car la fraction homozygote augmente. La formule devient :

$$E(k) = \sum_{i=1}^{n_p} p_{ik} \sum_{j=1}^{n_m} r_{jk} \quad n_{ij}$$

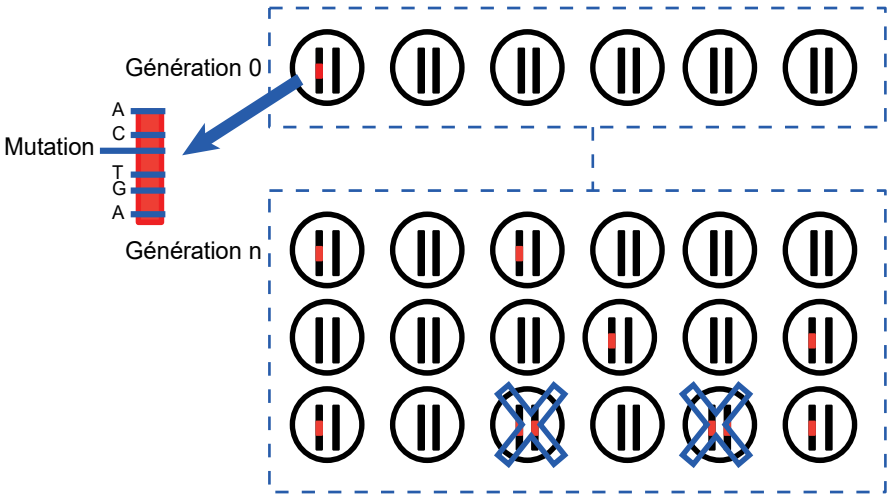
avec n_m le nombre de mères, n_{ij} le nombre de produits du père i et de la mère j, et r_{jk} la probabilité de transmission (1, ½ ou 0) de l'haplotype k par la mère. Les deux formules, portant sur des animaux différents, peuvent être combinées.

Avec cette approche, Fritz *et al* (2013) ont détecté 33 régions significatives au seuil de 10^{-5} pour l'ensemble des trois races et 2 nouvelles régions l'ont été en 2015 (tableau 1). Ce seuil tient compte du fait que 2500 intervalles non chevauchants sont testés. Pour plus de la moitié des QTL détectés, des individus homozygotes sont observés, ce qui indique que si le QTL existe, il n'est pas létal à 100% ou, plus vraisemblablement, le déséquilibre de liaison entre l'haplotype et le polymorphisme recherché est incomplet.

En race Holstein, 19 QTL sont détectés. Les QTL détectés par VanRaden *et al* (2011) sont retrouvés sauf HH2 (pour lequel la probabilité est seulement de $p = 10^{-3}$) du fait d'une fréquence faible dans la population française. L'haplotype HBY correspond à l'anomalie Brachyspina (Charlier *et al* 2012) avec laquelle l'association est quasi complète.

Les haplotypes CVM1 et CVM2 correspondent à l'anomalie CVM (« *Complex*

Figure 1. *Principe de la recherche de déficit en homozygotes.*
À la génération n, on observe un haplotype assez fréquent à l'état hétérozygote mais absent, ou plus rare qu'attendu, à l'état homozygote, ce qu'on interprète comme de la mortalité embryonnaire.



Vertebral Malformation », Thomsen *et al* 2006) située à 43,4 Mb et donc distante de plusieurs Mb des QTL détectés. Ceci s'explique par le fait que la combinaison haplotypique entourant directement CVM est très fréquente et majoritairement non porteuse de CVM. Les haplotypes CVM1 et CVM2 sont un peu plus éloignés, mais moins fréquents et davantage associés. Ce constat montre que les régions mises en évidence ne constituent pas des intervalles de confiance de la position et que la recherche des polymorphismes causaux doit se faire dans un intervalle plus large.

En race Montbéliarde, 11 QTL sont détectés. Les deux principaux QTL (MH1 et MH2) présentent un déficit d'homozygotes important avec une fréquence élevée de l'haplotype, respectivement 9 et 7%, et un nombre nul ou quasiment nul d'individus homozygotes à l'haplotype.

Seulement 7 QTL sont détectés en race Normande. Ceci s'explique par des effectifs plus faibles d'individus typés, donc des nombres attendus d'individus homozygotes à un haplotype donné plus faibles et donc une puissance de détection plus réduite.

Pour conclure, cette approche basée sur le déficit en homozygotes s'est avérée très productive dans les populations étudiées. Notons cependant qu'elle n'est efficace que sous certaines conditions assez limitantes : la fréquence de l'haplotype porteur doit être suffisamment élevée (plusieurs pour cents) et l'échantillon analysé de très grande taille (plusieurs dizaines de milliers) pour que l'effectif attendu d'homozygotes soit suffisant ; et le déséquilibre de liaison entre l'haplotype et le variant délétère suffisamment fort pour que l'absence d'homozygotes pour le variant délétère se traduise bien par un déficit significatif d'homozygotes pour l'haplotype.

1.2 / Confirmation de la létalité par analyse de la fertilité ou de la mortalité

L'absence d'homozygotes peut provenir de causes variées. Les typages proviennent essentiellement du programme de sélection génomique. Compte tenu du coût de génotypage, uniquement les candidats prometteurs sont génotypés. Ainsi, tout animal présentant un défaut non acceptable pour un reproducteur serait éliminé du génotypage et ceci peut être à l'origine de déficits en homozygotes. Ainsi, l'haplotype HH12, localisé sur le chromosome 15, est peut-être associé à la syndactylie (ou pied de mule), une anomalie induisant une fusion des onglons (Duchesne *et al* 2006). Il est donc impératif de confirmer la létalité en gestation par une analyse du taux de mise bas ou de mortinatalité.

La base de données des inséminations artificielles utilisée pour l'évaluation génétique sur la fertilité a été mobilisée dans cet objectif. Chaque insémination réalisée entre 2000 et 2010 est préalablement qualifiée comme fécondante ou non, les inséminations au statut indéterminé étant éliminées. Puis les inséminations sont classées comme accouplement à risque ou non. Un accouplement à risque est défini par un taureau porteur

Tableau 1. *Principales régions détectées depuis 2013 présentant des déficits en homozygotes.*

Race	Nom	Chr.	Intervalle (Mb)	E(k)	O(k)	Fréquence (%)
Holstein	HBY[1]	21	20,2 - 22,3	49	0	3,6
Holstein	HH1	5	61,4 - 66,2	18	0	2,6
Holstein	HH3	8	94,0 - 96,5	21	0	2,5
Holstein	HH4	1	1,9 - 3,3	49	0	3,6
Holstein	HH5	9	92,1 - 93,2	22	0	1,4
Holstein	CVM1[2]	3	49,4 - 52,6	91	38	4,6
Holstein	CVM2[2]	3	45,8 - 47,6	68	24	3,9
Montbéliarde	MH1	19	27,6 - 29,4	131	0	9,0
Montbéliarde	MH2	29	27,9 - 29,1	80	1	7,0
Montbéliarde	MH3	6	73,3 - 74,4	122	14	7,1
Normande	NH1	24	38,1 - 39,2	12	0	1,8
Normande	NH2	11	51,1 - 54,2	31	0	3,7

Chr = Chromosome, E(k) = Espérance d'homozygotes pour l'haplotype k sous l'hypothèse neutre, O(k) = Nombre d'homozygotes effectivement observés pour l'haplotype k.
[1]HBY est un haplotype associé à l'anomalie Brachyspina.
[2]CVM1 et CVM2 sont deux haplotypes associés à l'anomalie CVM.

de l'haplotype accouplé à une fille d'un père porteur, et les trois autres types d'accouplements sont classés non à risque. Le génotype du père et du grand-père maternel de l'embryon est prédit à partir de leurs typages phasés. Si le typage du père ou du grand-père maternel est absent, l'insémination n'est pas considérée.

Théoriquement, dans le cas d'un déséquilibre complet de l'haplotype et du polymorphisme, l'effet attendu dans les accouplements à risque est de 1/4(2-f). TMB, une valeur proche de TMB/8, où TMB est le taux de mise bas moyen dans la population étudiée, f la fréquence allélique de l'anomalie et 1/4(2-f) est la fraction homozygote à l'anomalie dans la descendance. La baisse attendue dans les accouplements à risque est donc de l'ordre de 5 à 6% selon la race et le type de femelles, génisses ou vaches en lactation. Ces valeurs constituent des *maxima* car elles supposent une association complète entre QTL et haplotype et une mortalité totale à l'état embryonnaire.

La même analyse est réalisée sur les données de mortinatalité. Dans ce cas, la hausse de mortalité attendue peut atteindre 11 à 12 points environ, soit environ 1/8ème du taux de survie moyen à la naissance.

La figure 2 montre les effets des 12 haplotypes associés à une dégradation de fertilité. Les 13 autres haplotypes listés au tableau 1 ne présentent pas d'effets sur la reproduction et la survie néo-natale, suggérant une autre cause de déficit ou un artefact. Les effets des haplotypes HBY, HH1 et HH3, déjà connus, ont été confirmés et leur niveau est compatible avec une association quasi complète avec la mutation létale. Les haplotypes CVM1 et CVM2 sont également confirmés mais montrent des effets plus faibles, compatibles avec une association partielle avec CVM. Les haplotypes HH4, MH1, MH2 et NH1 sont nouveaux et présentent une perte de fertilité compatible avec une association quasi complète avec leur mutation causale. Les autres ont des effets plus faibles et restent à confirmer.

Il convient également de tester si les mutations impliquées dans les anomalies ont des effets pléiotropes, ce qui pourrait entrainer leur sélection positive à l'état hétérozygote. Dans ce cas, on peut observer des fréquences particulièrement élevées. C'est le cas par exemple de la délétion décrite par Kadri *et al* (2014), létale à l'état homozygote, mais favorisant la production laitière à l'état hétérozygote, et donc particulièrement fréquente en population Rouge nordique.

1.3 / Recherche de la mutation causale

En l'absence de matériel biologique de cas homozygotes mutés, il faut rechercher la mutation chez les porteurs hétérozygotes, en particulier chez les taureaux qui ont pu la disséminer largement. La recherche des polymorphismes sous-jacents aux QTL cartographiés consiste à exploiter les séquences de génome complet disponibles, obtenues en interne ou dans le cadre du projet « 1000 génomes bovins » (Daetwyler *et al* 2014). Pour chaque région, les polymorphismes sont étudiés de façon à identifier ceux qui respectent les conditions suivantes :

i) jamais présents à l'état homozygote (sinon ils ne seraient pas létaux) ;

ii) absents de la séquence de référence (correspondant à une vache de race Hereford) ;

iii) présents à l'état hétérozygote chez tous les taureaux considérés comme porteurs, sur la base de leur haplotype et avec effet confirmé à partir des accouplements à risque ;

iv) absents chez tous les taureaux confirmés comme non porteurs, de la même race ou d'autres.

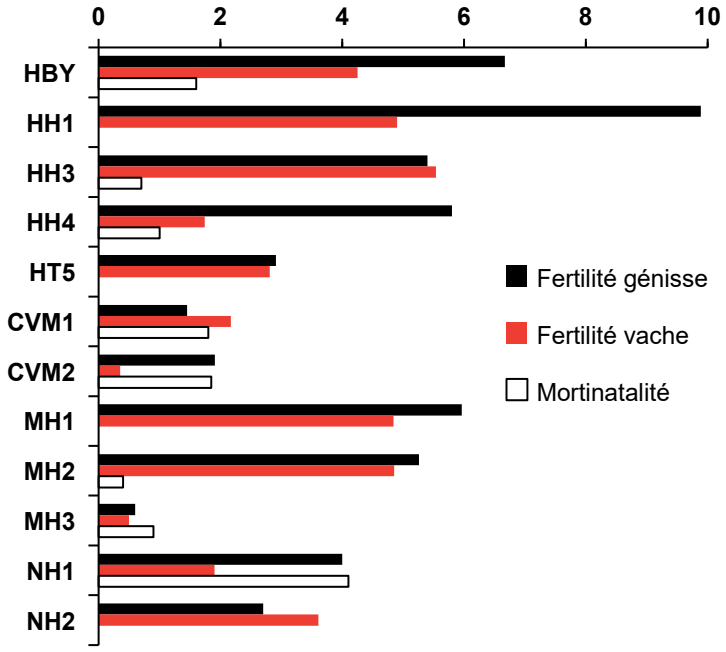
Enfin, dans un second temps, seules les séquences codantes sont conservées et les polymorphismes candidats sont sélectionnés à partir de la sévérité de l'effet prédit sur la protéine, ainsi que de la conservation de la protéine entre espèces (avec les mêmes critères que dans Duchesne *et al* 2016). L'exemple de CVM montre que la mutation causale peut être en dehors de l'haplotype diagnostic. Chaque région analysée comprend donc l'intervalle détecté étendu de 6 Mb de chaque côté, valeur suffisante pour définir un intervalle de confiance fiable.

Les polymorphismes déjà connus ont ainsi été retrouvés par cette approche : HBY (gène FANCI, Charlier *et al* 2012), CVM (SLC35A3, Thomsen *et al* 2006) et HH1 (APAF1, Adams *et al* 2016). Cinq autres polymorphismes ont été identifiés, deux en race Holstein, deux en race Montbéliarde et un en race Normande associés respectivement aux haplotypes HH3, HH4, MH1, MH2 et NH7. Ces mutations touchent des fonctions essentielles à la vie et on imagine aisément qu'elles ne soient pas tolérées à l'état homozygote. Ainsi, par exemple, le gène GART (HH4) code pour une enzyme impliquée dans la synthèse des purines ; le gène SMC2 (HH3) code pour une protéine essentielle pour la condensation des chromosomes et donc pour la mitose. Argument fort, dans ces deux cas, la conservation de l'acide aminé modifié est très large, chez tous les eucaryotes par exemple pour le gène SMC2.

Des approches analogues ont été conduites depuis par d'autres équipes (tableau 2), permettant de caractériser de nouvelles mutations létales en Holstein (HH5), Brune (BH2), Simmental (FH2

Figure 2. *Effets des haplotypes sur la baisse du taux de conception des génisses et des vaches (en point de conception) et sur la hausse de la mortinatalité (en % de mortalité) dans les accouplements à risque de type taureau porteur x fille de père porteur.*

Données de fertilité et de mortinatalité de 2000 à 2010.



et FH4), Jersey (JH1), Ayrshire (AH1) et Rouge nordique. Le développement de la sélection génomique et donc des typages dans un nombre croissant de races laisse supposer que d'autres variants délétères seront trouvés à l'avenir. Il convient de noter que si le résultat est une baisse de fertilité, aucun de ces gènes n'est impliqué dans la fonction de reproduction. Ils sont essentiels pour le développement embryonnaire ou fœtal, et les modifications induites par ces mutations entrainent la mortalité.

1.4 / Confirmation à grande échelle

Une validation fonctionnelle de ces variants est complexe car il est impossible de disposer d'échantillons biologiques de porteurs homozygotes. Toutefois, une confirmation statistique est possible, avec le génotypage à grande échelle de la mutation candidate. La stratégie est la suivante : la mutation candidate est ajoutée sur la puce de génotypage utilisée en sélection génomique (la puce EuroG10k en ce qui concerne nos équipes) et utilisée ensuite sur tous les candidats à la sélection. On peut alors vérifier sur un très grand échantillon qu'aucun individu n'est homozygote pour la mutation candidate. La présence d'individus homozygotes infirmerait la mutation candidate. On peut également mesurer le degré d'association entre la mutation (confirmée) et l'haplotype utilisé initialement pour la mettre en évidence et vérifier que les homozygotes à l'haplotype, s'ils existent, ne sont pas homozygotes à la mutation candidate. On mesure également la fréquence de la mutation candidate dans la population cible et on peut vérifier si elle est présente ou non dans d'autres races.

Dans la grande majorité des cas, la vérification apportée par les typages à grande échelle a confirmé et consolidé les premiers résultats (tableau 3). Un résultat cependant a été infirmé après publication, une mutation dans le gène SHBG assimilée à tort au QTL MH1. La mutation proposée est prédite pour engendrer le knock-out du gène SHBG

Tableau 2. *Liste des mutations causales découvertes par recherche de déficit en homozygotes puis séquençage.*

Race	Haplo-type	Gène	Fréquence (%)	Chromosome	Position (pb)	Références
Ayrshire	AH1	PIRM/*UBE3B*	13,0	17	65 921 497	Venhoranta *et al* (2014)
Brune	BH2	*TUBD1*	7,8	19	11 063 520	Schwarzenbacher *et al* (2016)
Holstein	HCD	*APOB*	2,5	11	77 958 995	Menzi *et al* (2016), Schütz *et al* (2016)
Holstein	HBY	*FANCI*	2,8	21	21 184 869 21 188 198	Charlier *et al* (2012)
Holstein	HH1	*APAF1*	1,3	5	63 150 400	Fritz *et al* (2013), Adams *et al* (2016)
Holstein	HH3	*SMC2*	3,7	8	95 410 507	Daetwyler *et al* (2014), McClure *et al* (2014)
Holstein	HH4	*GART*	3,8	1	1 277 227	Fritz *et al* (2013)
Holstein	HH5	*TFB1M*	2,2	9	93 223 651 93 370 998	Schütz *et al* (2016)
Jersey	JH1	*CWC15*	12,1	15	15 707 169	Sonstegard *et al* (2013)
Montbéliarde	MH1	*PFAS*	7,4%	19	28 511 199	Michot *et al* (soumis)
Montbéliarde	MH2	*SLC37A2*	5,7	29	28 879 810	Fritz *et al* (2013)
Normande	NH2	*CAD*	3,1	11	72 399 397	Michot *et al* (en préparation)
Simmental	FH2	*SLC2A2*	4,1	1	97 239 973 97 239 976	Pausch *et al* (2015)
Simmental	FH4	*SUGT1*	3,3	12	11 131 497	Pausch *et al* (2015)

Tableau 3. *Résultats de génotypage pour les mutations récessives létales étudiées en France.*

QTL	Mutation candidate	Races	Nombre d'individus typés	Fréquence de la mutation (%)	Nombre d'homozygotes mutés
HH1	APAF1	Brune	1412	0,00	0
HH1	APAF1	Holstein	136 444	1,28	0
HH1	APAF1	Montbéliarde	90 609	< 0,01	0
HH1	APAF1	Normande	30 245	< 0,01	0
HH3	SMC2	Brune	1413	0,00	0
HH3	SMC2	Holstein	136 445	3,74	0
HH3	SMC2	Montbéliarde	90 614	0,00	0
HH3	SMC2	Normande	30 246	< 0,01	0
HH4	GART	Brune	1412	0,00	0
HH4	GART	Holstein	136 443	3,83	0
HH4	GART	Montbéliarde	90 614	< 0,01	0
HH4	GART	Normande	30 246	0,00	0
MH1	SHBG	Brune	1412	0,00	0
MH1	SHBG	Holstein	136 381	< 0,01	0
MH1	SHBG	Montbéliarde	90 652	11,40	584 [1]
MH1	SHBG	Normande	30 248	0,00	0
MH1	PFAS	Brune	610	0,00	0
MH1	PFAS	Holstein	40 635	< 0,01	0
MH1	PFAS	Montbéliarde	37 484	7,40	0
MH1	PFAS	Normande	8319	0,00	0
MH2	SLC37A2	Brune	1412	0,00	0
MH2	SLC37A2	Holstein	135 681	< 0,01	0
MH2	SLC37A2	Montbéliarde	90 614	5,70	0
MH2	SLC37A2	Normande	30 245	0,00	0
NH7	CAD	Brune	610	0,00	0
NH7	CAD	Holstein	41 255	0,00	0
NH7	CAD	Montbéliarde	37 455	0,00	0
NH7	CAD	Normande	8255	3,10	0

[1] Cette mutation est invalidée par le nombre d'homozygotes observés.

codant pour un transporteur de stéroïdes. Les premiers typages n'ont pas montré d'homozygotes, confirmant cette hypothèse, et ce résultat a été publié par Fritz *et al* (2013). Toutefois, l'accumulation de typages a permis de mettre en évidence un nombre non négligeable d'homozygotes en race Montbéliarde à la suite d'une recombinaison dans l'haplotype HH1. Ces animaux sont en cours d'étude. Mais leur simple existence a remis en cause le résultat. Une autre mutation candidate dans le gène PFAS est en cours de confirmation, pour laquelle aucun homozygote n'a été observé à ce jour, malgré une fréquence élevée en race Montbéliarde et l'existence de très nombreux accouplements à risque.

2 / Recherche de mutations candidates dans les séquences de génome complet

Les mutations récessives sont déjà largement répandues dans la population depuis plusieurs générations lorsqu'apparaissent les premiers cas d'homozygotes atteints. On peut également supposer que de nombreuses anomalies passent inaperçues (mortalité embryonnaire, déficit immunitaire, problème métabolique…), faute de symptômes spécifiques. Or on dispose maintenant d'un grand nombre de séquences de génomes complets, en particulier celles des taureaux les plus importants ayant fortement contribué au patrimoine génétique de leur race. Ces séquences apportent donc une description d'une grande partie de la variabilité de l'ADN existant dans ces populations. L'idée est donc de rechercher par fouille de données, parmi les millions de polymorphismes présents, ceux qui pourraient

être responsables de futures émergences d'anomalies pour pouvoir les anticiper, voire même d'émergences actuelles non détectées.

2.1 / Recherche de variants candidats

La première étape consiste à réaliser l'inventaire de tous les variants portés par chaque individu, c'est-à-dire toutes les différences par rapport à la séquence de référence de l'espèce (pour les bovins, la séquence d'une vache Hereford). Typiquement, on en détecte de l'ordre de quelques millions par individu et quelques dizaines de millions à l'échelle de l'ensemble des animaux séquencés (Daetwyler *et al* 2014, Boussaha *et al* 2016). Un outil de référence de détection des variants de petite taille est « *Genome Analysis Tool Kit* » (GATK, McKenna *et al* 2010).

Dans un second temps, il faut donner du sens à ces variants. L'outil principal est l'annotation fonctionnelle. Cette annotation est relativement bonne dans la partie codante des gènes, moyenne dans les autres parties des gènes, et notoirement insuffisante dans les parties intergéniques. Pour les variants dans les gènes, un outil particulièrement utile est Ve!P, acronyme de « *Variant Effect Predictor* » (McLaren *et al*,2010). Ve!P fournit la liste des gènes et transcrits affectés par un variant et la position des variants dans les gènes (amont ou aval de la transcription, dans la séquence codante, dans les introns, dans les régions régulatrices…). Il prédit les conséquences des variants sur la séquence protéique. Contrairement à un variant synonyme, un variant non synonyme induit un changement d'acide aminé. Le variant peut créer un codon stop et donc induire une protéine plus courte, ou au contraire supprimer un codon stop et induire une protéine plus longue. Une insertion ou délétion d'un nombre de bases non multiple de 3 dans la séquence codante induit un décalage du cadre de lecture et conduit à une protéine totalement différente et fréquemment à l'apparition d'un codon stop prématuré. Une modification d'un site d'épissage alternatif peut conduire à un épissage différent, par exemple la perte d'exons, et peut également altérer fortement la fonction de la protéine, en particulier si ces changements touchent les domaines fonctionnels. L'outil Ve!P fournit également les scores « *SIFT* » (Kumar *et al* 2009) et « *Polyphen* » (Adzhubei, 2010) de prédiction d'impact du variant sur la fonction de la protéine et dans le cas de recherche d'anomalie, nous ciblons principalement les variants délétères.

Ce qui a été décrit précédemment concerne les variants de petite taille, pour lesquels les outils sont assez standardisés. Les variants structuraux, c'est-à-dire les polymorphismes de grande taille (délétions, insertions, duplications, inversions) jouent un rôle tout particulier, car ils couvrent une région chromosomique, potentiellement fonctionnelle, plus grande. En effet, ils sont beaucoup moins nombreux que les variants d'une ou de quelques bases, mais ils ont une probabilité beaucoup plus forte d'avoir un impact fonctionnel, surtout lorsqu'ils affectent tout un gène ou une partie importante de ce gène. Les exemples chez les bovins ne manquent pas.

Capitan *et al* (2012) ont montré l'impact d'une grande délétion incluant le gène ZEB2 en race charolaise, avec des conséquences sévères sur différentes fonctions. Michot *et al* (2015) ont montré qu'une délétion de 4,8 kb dans le gène ITGB4 induisait une épidermolyse bulleuse jonctionnelle en race Charolaise. Kadri *et al* (2014) ont mis en évidence une grande délétion de 660 kb avec un effet très délétère sur la fertilité en race Rouge nordique. Sahana *et al* (2016) ont montré qu'une délétion de 500 kb induisait une mortalité à la naissance. Plusieurs équipes ont mis en évidence le déterminisme du syndrome CDH (déficience de cholestérol) en race Holstein, dû à l'insertion d'un retrotransposon dans le gène APOB (Menzi *et al* 2016), conduisant à sa perte de fonction. Un autre exemple est la délétion du gène FANCI, responsable du syndrome Brachyspina en race Holstein (Charlier *et al* 2012). Les variants structuraux sont donc des cibles privilégiées de ces recherches d'anomalies à partir des séquences. Dans une région spécifiée, un outil de diagnostic particulièrement efficace est « *Integrative Genome Viewer* » (IGV, Thorvaldsdóttir *et al* 2013). À l'échelle du génome entier, d'autres outils plus systématiques sont utilisés. Boussaha *et al* (2015) présentent ces approches ainsi qu'un premier inventaire des variants structuraux observés dans trois races françaises à partir de données de séquence de génome complet.

En général, l'annotation est loin d'être suffisante pour mettre en évidence de futures anomalies car elle se caractérise par un nombre élevé de faux positifs. Il avait déjà été constaté chez l'Homme (The 1000 Project Consortium, 2010) que l'annotation prédisait 100 à 150 mutations de type perte de fonction portées par individu en moyenne, soit beaucoup plus que le nombre prédit à partir des anomalies connues dans notre espèce, de l'ordre de 3 à 5 mutations portées à l'état hétérozygote. Même si le nombre d'anomalies réellement portées par individu est sous-estimé, du fait des anomalies non encore découvertes, il reste très inférieur à la valeur de 100 à 150 citée plus haut et on en déduit qu'une protéine peut voir sa fonction très altérée sans que la santé de l'organisme dans son ensemble soit fortement affectée. Ce nombre très supérieur s'explique par le fait que tous les gènes ne sont pas essentiels pour le développement et aussi en grande partie par les redondances entre voies métaboliques, permettant de maintenir la fonction même quand une des voies est atteinte.

Le constat précédent nécessite donc d'enrichir le diagnostic pour distinguer les variants réellement responsables d'anomalies des faux positifs, majoritaires. Une information très utile est disponible dans diverses bases de données décrivant les anomalies chez l'Homme (OMIM) et la souris (MGI). Ces bases sont présentées dans Grohs *et al* (2016). Si une anomalie observée chez l'Homme ou la souris est due à une mutation dans le même gène comparable à celle détectée sur la séquence étudiée, on peut supposer que les mécanismes d'échappement sont moins efficaces et que le risque d'anomalie est élevé.

Si cette recherche de variants d'intérêt dans la séquence engendre beaucoup de faux positifs, elle peut aussi manquer de puissance. Deux raisons principales expliquent ce constat : *i)* la non-détection d'une mutation, principalement par manque de couverture de séquence à ce locus, sa présence dans une zone répétée rendant son analyse trop compliquée, ou un trou dans la séquence de référence conduisant à la perte de l'information lors de la phase d'alignement des séquences produites sur la référence ; *ii)* l'absence d'annotation informative d'un variant détecté. Si les annotations sont souvent précises dans les régions codantes, elles sont encore très partielles dans le reste du génome. Des progrès sont nécessaires dans la détection des polymorphismes non codants impliqués dans les mécanismes de régulation des gènes.

2.2 / Validation des variants candidats par génotypage à grande échelle

L'analyse des données de séquence conduit à détecter plusieurs milliers de variants candidats. Ainsi, chez le bovin, Michot *et al* (2016) ont détecté quelques 2500 variants prédits comme délétères pour la fonction protéique et présents à une fréquence assez élevée (>5%) dans au moins une race. Dans une étude comparable très récente, Charlier *et al* (2016) ont détecté plus de 4000 variants non synonymes délétères et 400 variants perte de fonction. Ces nombres élevés de variants nécessitent de définir une stratégie d'analyse adaptée à cet effectif. Dans un premier temps, la stratégie, comme précédemment, est de réaliser un génotypage à grande échelle des popula-

tions à la recherche d'homozygotes. Cette stratégie est favorisée par la possibilité d'utiliser des puces à façon permettant de typer les variants candidats. Cette option est économiquement supportable si le nombre de puces utilisées est élevé, ce qui est le cas en sélection génomique bovine, en particulier avec la puce EuroG10k utilisée en France et dans différents pays européens membres d'Eurogenomics. Les variants candidats sont alors ajoutés dans un « *add on* » en complément des marqueurs utilisés en sélection génomique et sont typés en même temps. Une difficulté non négligeable est l'interprétation correcte des signaux de fluorescence de la puce, souvent source d'artefacts techniques et donc de conclusions erronées. La validation souvent manuelle de cette interprétation est une tâche lourde indispensable. Grâce aux dizaines, voire centaines de milliers de génotypages réalisés chaque année, on peut estimer la fréquence des variants candidats dans chaque race, rechercher l'absence éventuelle d'homozygotes, identifier les accouplements à risque (entre taureau porteur et vache porteuse ou fille d'un père porteur de la mutation) et suivre le devenir des veaux nés ou à naitre.

Dans le cadre du projet Bovano, plusieurs mutations létales au stade embryonnaire ont été découvertes ou confirmées. Il est également prévu de génotyper les veaux nés d'accouplements à risque pour plusieurs mutations candidates pour suivre l'évolution des animaux homozygotes mutés dans leurs premiers mois de vie et les comparer aux autres génotypes. Avec une approche similaire, Sartelet *et al* (2012) ont réalisé une étude longitudinale sur une centaine de veaux Blanc Bleu Belge issus d'accouplements à risque pour une mutation dans le gène RFN11 et ont montré un taux de mortalité élevé chez les homozygotes, dû à une réponse inflammatoire excessive, tandis que les hétérozygotes ont un avantage sélectif sur la croissance et la conformation. Cette surmortalité était impossible à déceler par l'approche classique sans connaître les génotypes des animaux. Très récemment, avec cette approche de génotypage à haut débit appliquée à 40 000 animaux Holstein néo-zélandais ou Blanc Bleu Belges, l'étude de Charlier *et al* (2016) a validé neuf nouvelles mutations responsables de mortalité embryonnaire et conclu que 15% des variants perte de fonction et 6% des variants non synonymes délétères testés induisent des pertes embryonnaires.

2.3 / Anomalies non létales : l'exemple du gène RP1

Parmi les quelques 2500 mutations détectées dans 1900 gènes et prédites comme délétères, Michot *et al* (2016) ont observé un enrichissement important en gènes impliqués dans le système nerveux, la vision et le système auditif, ainsi que dans le développement musculaire et squelettique. Il est fait l'hypothèse que la domestication apporte une sécurité (alimentation, logement, vie de groupe, sécurité vis-à-vis des prédateurs….) rendant possible l'apparition de défauts sensoriels qui seraient fortement contre-sélectionnés à l'état sauvage. Parmi les quelques 1000 variants candidats génotypés, l'un d'entre eux a fait l'objet d'une étude toute particulière, du fait de sa forte fréquence en race Normande et de son annotation dans OMIM. Il s'agit d'une mutation dans le gène RP1 prédite, sur la base d'arguments chez l'Homme, pour induire une dégénérescence rétinienne. Une analyse approfondie de bovins homozygotes pour cette mutation a montré une dégénérescence rétinienne à partir de l'âge de 4 ans, avec perte progressive de la vision nocturne puis diurne. Cette mutation semble expliquer une part importante des troubles de vision connus pour être fréquents dans cette race. Elle est ancienne puisqu'on la retrouve, à une fréquence moindre, dans d'autres races et la taille de l'haplotype conservé l'entourant permet de la dater à 560 générations, soit environ 3000 ans, bien après la domestication et bien avant la constitution des races. Cette situation en race Normande n'est pas récente puisqu'une estimation à partir d'animaux des années 1970 montre une fréquence encore plus élevée. Les raisons de cette situation atypique dans cette race ne sont pas connues et relèvent sans doute d'un effet fondateur massif. Contrairement à des situations similaires où l'anomalie a vu sa fréquence augmenter du fait de son association à un caractère fortement sélectionné (Sartelet *et al* 2012, Kadri *et al* 2014), une seule association de RP1 avec les caractères sélectionnés a été mise en évidence et expliquerait plutôt la baisse de fréquence dans les années récentes puisque l'anomalie est associée à une mauvaise conformation de mamelle.

Cet exemple illustre parfaitement le fait qu'une anomalie avec des conséquences compatibles avec l'élevage, bien qu'invalidante, peut se maintenir à une fréquence élevée à la faveur de dérives génétiques dans les populations animales.

2.4 / Cas des variants rares

Michot *et al* (2016) ont ciblé des mutations relativement fréquentes. Mais les mutations rares sont beaucoup plus nombreuses. Il peut s'agir de mutations récessives, mais dont la diffusion pourrait s'accroitre, ou de mutations dominantes qui pourraient s'exprimer directement chez les produits de première génération.

Si ce sont des mutations dominantes, elles sont généralement très récentes et peuvent être apparues *de novo*. Elles s'expriment chez une fraction des produits. Comme indiqué dans Boichard *et al* (2016 ce numéro), le fait que l'individu muté n'exprime pas le phénotype peut être dû à un mosaïcisme et n'est pas incompatible avec une mutation dominante. Pour la même raison, moins de 50% des produits peuvent être atteints. Plusieurs exemples sont bien connus : « *Bulldog* » chez le taureau « *Igale Masc* » (Daetwyler *et al* 2014), avec un atteint pour 200 veaux nés ; une pousse excessive du poil et une absence de thermorégulation chez un taureau néo-zélandais, présentant une mutation dans le gène du récepteur à la prolactine (Littlejohn *et al* 2014). Avec la disparition du testage sur descendance, le nombre de produits d'un taureau diffusé sur la base de son évaluation génomique est assez élevé avant que l'on ne constate l'anomalie et qu'on puisse interrompre sa carrière et il aurait été hautement préférable d'identifier la mutation de novo délétère avant la diffusion du taureau. L'objectif est donc de rechercher des mutations privées, c'est-à-dire présentes chez un seul individu. L'exemple du taureau Montbéliard « *Etsar* », responsable du syndrome de CHARGE, suggère qu'une analyse de la séquence d'ADN aurait permis de prédire ce statut avant l'utilisation du taureau. À noter cependant que le mosaïcisme peut être un handicap pour cette approche. Dans le cas d'« *Igale Masc* » et du syndrome « *Bulldog* », il aurait fallu une couverture de séquence très élevée pour identifier l'anomalie chez le taureau mosaïque à environ 1%.

Les mutations récessives et à faible fréquence ne constituent pas un problème, mais portées par des reproducteurs élites, elles peuvent voir cette fréquence augmenter rapidement, et il convient de les identifier le plus précocement possible. Une première étude (Bourneuf *et al*, soumis) a porté sur la séquence de 43 taureaux (24 Holstein, 11 Montbéliards, 5 Normands et 3 Charolais) nés au cours des quatre dernières générations. Ils sont donc suffisamment jeunes pour qu'une nouvelle mutation ne se traduise pas encore par des cas, faute d'accouplements consanguins. Nous avons ainsi identifié 18 mutations délétères privées (c'est-à-dire portées par un seul individu) dont sept ont été confirmées par séquençage Sanger. Pour cinq de ces mutations dans les gènes COL6A3, EDAR, ITGA3, SLC35A3 et SOWAHB, les conséquences prédites sont sévères, conduisant à une mortalité embryonnaire ou périnatale, ou nécessité d'euthanasie, tandis que les deux autres gènes, FAM189A1 et CSNK1G2, sont encore mal connus. La validation de ces prédictions n'est pas aisée et repose sur la détection, voire la

réalisation, d'accouplements à risque entre porteurs.

Toutefois, en race allaitante Charolaise, l'intervalle de génération est plus court qu'en race laitière et il a été possible de rechercher des individus consanguins descendants du taureau porteur d'une mutation « décalage du cadre de lecture » dans le gène EDAR. L'analyse du pedigree charolais complet nous prédit qu'en espérance 14 veaux parmi les 344 564 veaux nés en 2013 et d'état civil connu seraient homozygotes pour cette mutation. En parallèle, une enquête nous a permis de détecter 10 de ces veaux, tous atteints de dysplasie anhidrotique ectodermale, c'est-à-dire d'absence de poils et de dents. Ce phénotype est similaire chez les Hommes portant des mutations dans les gènes de la voie de l'ectodysplasine. Le génotypage de sept de ces veaux atteints a montré qu'ils portaient tous l'haplotype présent chez le taureau ancêtre séquencé. Cette mutation est une mutation *de novo* chez le taureau ancêtre, et donc absente de chez ses parents. Nous avons pu analyser trois autres veaux homozygotes pour le même haplotype mais descendants du père de l'ancêtre et ne portant donc pas la mutation, et ces trois veaux sont non atteints. Cette observation démontre donc que la mutation ciblée, qui est la seule différence d'ADN dans la région chromosomique entre les veaux atteints et non atteints, est causale, et aucune expérimentation supplémentaire n'est nécessaire. Cette stratégie peut être adaptée pour caractériser les conséquences phénotypiques de gènes mal connus comme FAM189A1 et CSNK1G2 et ainsi améliorer l'annotation du génome de mammifères.

La diminution du coût du séquençage rend envisageable une généralisation de cette approche de séquençage de tous les reproducteurs promis à une forte diffusion, de façon à anticiper les futures émergences d'anomalies. On peut donc supposer que cette pratique va devenir la norme dans un futur très proche. Une difficulté reste à surmonter cependant, il convient d'améliorer sensiblement notre capacité prédictive de l'effet réel de nombreux variants supposés délétères. Si la prédiction est possible quand des syndromes équivalents existent chez la souris ou l'Homme, elle reste très aléatoire dans la grande majorité des cas. Nul doute cependant que des progrès importants seront réalisés dans les prochaines années dans la prédiction des effets biologiques des variants.

Conclusion

Les évolutions technologiques de séquençage et le génotypage de masse dans le cadre de la sélection génomique permettent de développer de nouvelles approches encore inenvisageables il y a quelques années. Ces méthodes sont complémentaires de l'approche classique du phénotype au génotype. Elles permettent de rechercher des anomalies peu visibles au niveau phénotypique comme les mortalités embryonnaires ou des anomalies aux effets peu spécifiques. Les résultats sont nombreux, particulièrement à partir de l'analyse des séquences seules, et la principale difficulté réside dans le risque de détection de faux positifs et dans le manque de validation des résultats. S'il ne fournit aucune preuve formelle, le génotypage de masse apporte une information statistique qui permet de renforcer les convictions et il aide à l'utilisation des résultats en établissant le génotype des reproducteurs. Une autre difficulté est la multiplication des résultats obtenus au cours de ces dernières années. L'approche du déficit en homozygotes a permis de caractériser cinq anomalies létales à l'état embryonnaire en deux ans en race Holstein, et trois en race Montbéliarde. Selon Charlier *et al* (2016), 6 à 15% des variants délétères, selon leur nature, correspondraient à des anomalies (mais sont heureusement à des fréquences très faibles). Ces résultats deviennent trop nombreux pour envisager une éradication immédiate comme c'est souvent le cas pour les anomalies, et de nouvelles approches sont nécessaires pour prendre en compte ces résultats en sélection.

Remerciements

Ces travaux ont été réalisés dans le cadre du projet Bovano financé par l'ANR et Apisgene (ANR-14-CE19-0011). Les typages utilisés proviennent en majorité de l'activité de sélection génomique, ainsi que de différents projets de recherche. Les séquences ont été obtenues dans le cadre d'une dizaine de projets de recherche, dont Cartoseq (ANR-10-GENM-0018) financé par l'ANR et ApisGene.

Références

Adams H.A., Sonstegard T.S., VanRaden P.M., Null D.J., Van Tassell C.P., Larkin D.M., Lewin H.A., 2016. Identification of a nonsense mutation in APAF1 that is likely causal for a decrease in reproductive efficiency in Holstein dairy cattle. J. Dairy Sci., 99, 6693-6701.

Adzhubei I.A., Schmidt S., Peshkin L., Ramensky V.E., Gerasimova A., Bork P., Kondrashov A.S., Sunyaev S.R., 2010. A method and server for predicting damaging missense mutations. Nat. Methods, 7, 248-249.

Barbat A., Le Mezec P., Ducrocq V., Mattalia S., Fritz S., Boichard D., Ponsart C., Humblot P., 2010. Female fertility in French dairy breeds: current situation and strategies for improvement. J. Reprod. Dev., 56, S15-S21.

Boichard D., Grohs C., Danchin-Burge C., Capitan A., 2016. Les anomalies génétiques : définition, origine, transmission et évolution, mode d'action. In : Anomalies génétiques. Boichard D. (Ed). Dossier, INRA Prod. Anim., 29, 297-306.

Bourneuf E., Otz P., Pausch H., Jagannathan V., Michot P., Grohs C., Piton G., Ammermüller S., Deloche M.C., Fritz S., Leclerc H., Péchoux C., Boukadiri A., Saintilan R., Créchet F., Mosca M., Segelke D., Guillaume F., Bouet S., Baur A., Vasilescu A., Genestout L., Thomas A., Allais-Bonnet A., Rocha D., Colle M.A., Klopp C., Esquerré D., Wurmser C., Flisikowski K., Schwarzenbacher H., Burgstaller J., Brügmann M., Dietschi E., Huth N., Freick M., Barbey S., Fayolle G., Danchin-Burge C., Schibler L., Bed'hom B., Hayes B.J., Daetwyler H.D., Fries R., Boichard D., Pin D., Drögemüller C., Capitan A., 2016. Rapid Discovery of De Novo Deleterious Mutations in Cattle Using Genome Sequence Data: Enhancing the Value of Farm Animals as Model Species. Soumis.

Boussaha M., Esquerre D., Barbieri J., Djari A., Pinton A., Letaief R., Salin G., Escudie F., Roulet A., Fritz S., Samson F., Grohs C., Bernard M., Klopp C., Boichard D., Rocha D., 2015. Genome-wide study of structural variants in bovine Holstein, Montbeliarde and Normande dairy breeds. Plos One, 10: e0135931.

Boussaha M., Michot P., Letaief R., Hoze C., Fritz S., Grohs C., Esquerre D., Duchesne A., Philippe R., Blanquet V., Phocas F., Floriot S., Rocha D., Klopp C., Capitan A., Boichard D., 2016. Construction of a large collection of small genome variations in French dairy and beef breeds using whole genome sequences. Genet. Select. Evol., 48, 87.

Capitan A., Bonnet A., Pinton A., Marquant-Le Guienne B., Le Bourhis D., Grohs C., Bouet S., Clement L., Salas-Cortes L., Venot E., Chaffaux S., Weiss B., Delpeuch A., Noe G., Rossignol M.N., Barbey S., Dozias D., Cobo E., Barasc H., Auguste A., Pannetier M., Deloche M.C., Lhuilier E., Bouchez O., Esquerre D., Salin G., Klopp C., Donnadieu C., Chantry-Darmon C., Hayes H., Gallard Y, Ponsart C., Boichard D., Pailhoux E., 2012. A 3.7 Mb deletion encompassing ZEB2 causes a novel Polled and Multisystemic Syndrome in the progeny of a somatic mosaic bull. Plos One, 7, e49084.

Charlier C., Agerholm J.S., Coppieters W., Karlskov-Mortensen P., Li W., de Jong G., Fasquelle C., Karim L., Cirera S., Cambisano N., Ahariz N., Mullaart E., Georges M., Fredholm M., 2012. A Deletion in the Bovine FANCI Gene Compromises Fertility by Causing Fetal Death and Brachyspina. Plos One, 7, e43085.

Charlier C., Li W., Harland C., Littlejohn M., Coppieters W., Creagh F., Davis S., Druet T.,

Faux P., Guillaume F., Karim L., Keehan M., Kadri N.K., Tamma N., Spelman R., Georges M., 2016. NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. Genome Res., 26, 1333-1341.

Daetwyler H.D., Capitan A., Pausch H., Stothard P., Van Binsbergen R., Brøndum R.F., Liao X., Djari A., Rodriguez S.C., Grohs C., Esquerré D., Bouchez O., Rossignol M.N., Klopp C., Rocha D., Fritz S., Eggen A., Bowman P., Coote D., Chamberlain A.J., Anderson C., Van Tassell C.P., Hulsegge I., Goddard M.E., Guldbrandtsen B., Lund M.S., Veerkamp R.F., Boichard D., Fries R., Hayes B.J., 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet., 46, 858-867.

Druet T., Georges M. 2010. A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics, 184, 789-798.

Duchesne A., Gautier M., Chadi S., Grohs C., Floriot S., Gallard Y., Caste G., Ducos A., Eggen A., 2006. Identification of a doublet missense substitution in the bovine LRP4 gene as a candidate causal mutation for syndactyly in Holstein cattle. Genomics, 88, 610-621.

Duchesne A., Grohs C., Michot P., Boichard D., Floriot S., Fritz S., Capitan A., 2016. Du phénotype à la mutation causale d'anomalies récessives bovines. In : Anomalies génétiques. Boichard D. (Ed). Dossier, INRA Prod. Anim., 29, 319-328.

Fritz S., Capitan A., Djari A., Rodriguez S.C., Barbat A., Baur A., Grohs C., Weiss B., Boussaha M., Esquerre D., Klopp C., Rocha D., Boichard D., 2013. Detection of haplotypes associated with prenatal death in dairy cattle and identification of deleterious mutations in GART, SHBG, and SLC37A2. Plos One, 8, e65550.

Grohs C., Duchesne A., Floriot S., Deloche M.C., Boichard D., Ducos A., Danchin-Burge C. 2016. L'Observatoire National des Anomalies Bovines, son action et ses résultats pour une aide efficace à la gestion des anomalies génétiques. In : Anomalies génétiques. Boichard D. (Ed). Dossier, INRA Prod. Anim., 29, 307-318.

Kadri N.K., Sahana G., Charlier C., Iso-Touru T., Guldbrandtsen B., Karim L., Nielsen U.S., Panitz F., Aamand G.P., Schulman N., Georges M., Vilkki J., Lund M.S., Druet T., 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 Genetics, 10, e1004049.

Kumar P., Henikoff S., Ng P.C., 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc., 4, 1073-1081.

Littlejohn M.D., Henty K.M., Tiplady K., Johnson T., Harland C., Lopdell T., Sherlock R.G., Li W.B., Li W., Lukefahr S.D., Shanks B.C., Garrick D.J., Snell R.G., Spelman R.J., Davis S.R., 2014. Functionally reciprocal mutations of the prolactin signalling pathway define hairy and slick cattle. Nat. Comm., 5, 5861.

McClure M.C., Bickhart D., Null D., VanRaden P., Xu L.Y., Wiggans G., Liu G., Schroeder S., Glasscock J., Armstrong J., Cole J.B., Van Tassell C.P., Sonstegard T.S., 2014. Bovine exome sequence analysis and targeted SNP genotyping of recessive fertility defects BH1, HH2, and HH3 reveal a putative causative mutation in SMC2 for HH3. Plos One, 9, e92769.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M., DePristo M.A., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297-1303.

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

Menzi F., Besuchet-Schmutz N., Fragniere M., Hofstetter S., Jagannathan V., Mock T., Raemy A., Studer E., Mehinagic K., Regenscheit N., Meylan M., Schmitz-Hsu F., Drogemuller C., 2016. A transposable element insertion in APOB causes cholesterol deficiency in Holstein cattle. Anim. Genet., 47, 253-257.

Michot P., Fantini O., Braques R., Allais-Bonnet A., Saintilan R., Grohs C., Barbieri J., Genestout L., Danchin C., Gourreau J.M., Boichard D., Pin D., Capitan A., 2015. Whole-genome sequencing identifies a homozygous deletion encompassing exons 17 to 22 of the Integrin Beta 4 Gene in a Charolais calf with Junctional Epidermolysis Bullosa. Genet. Select. Evol., 47, 37.

Michot P., Chahory S., Marete A., Grohs C., Dagios D., Donzel E., Aboukadiri A., Deloche M.C., Allais-Bonnet A., Chambrial M., Barbey S., Boussaha M., Danchin-Burge C., Fritz S., Boichard D., Capitan A. 2016. A reverse genetic approach identifies an ancient frameshift mutation in RP1 causing recessive progressive retinal degeneration in European cattle breeds. Genet. Select. Evol., 48, 56.

Michot P., Fritz S., Barbat A., Boussaha M., Deloche M.C., Grohs C., Hoze C., Le Berre L., Le Bourhis D., Desnoes O., Salvetti P., Schibler L., Boichard D., Capitan A. A missense mutation in PFAS is likely causal for embryonic lethality associated with the MH1 haplotype in Montbeliarde dairy cattle. J. Dairy Sci., soumis.

Pausch H., Schwarzenbacher H., Burgstaller J., Flisikowski K., Wurmser C., Jansen S., Jung S., Schnieke A., Wittek T., Fries R., 2015. Homozygous haplotype deficiency reveals deleterious mutations compromising reproductive and rearing success in cattle. BMC Genomics, 16, 312.

Sahana G., Iso-Touru T., Wu X.P., Nielsen U.S., de Koning D.J., Lund M.S., Vilkki J., Guldbrandtsen B., 2016. A 0.5-Mbp deletion on bovine chromosome 23 is a strong candidate for stillbirth in Nordic Red cattle. Genet. Select. Evol., 48, 35.

Sargolzaei M., Chesnais J.P., Schenkel F.S., 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics, 15, 478.

Sartelet A., Druet T., Michaux C., Fasquelle C., Geron S., Tamma N., Zhang Z.Y., Coppieters W., Georges M., Charlier C., 2012. A splice site variant in the bovine RNF11 gene compromises growth and regulation of the inflammatory response. Plos Genet., 8, e1002581.

Schutz E., Wehrhahn C., Wanjek M., Bortfeld R., Wemheuer W.E., Beck J., Brenig B., 2016. The Holstein Friesian Lethal Haplotype 5 (HH5) Results from a Complete Deletion of TBF1M and Cholesterol Deficiency (CDH) from an ERV-(LTR) Insertion into the Coding Region of APOB. Plos One, 11, e0154602.

Schwarzenbacher H., Burgstaller J., Seefried F.R., Wurmser C., Hilbe M., Jung S., Fuerst C., Dinhopl N., Weissenbock H., Fuerst-Waltl B., Dolezal M., Winkler R., Grueter O., Bleul U., Wittek T., Fries R., Pausch H., 2016. A missense mutation in TUBD1 is associated with high juvenile mortality in Braunvieh and Fleckvieh cattle. BMC Genomics, 17, 400.

Sonstegard T.S., Cole J.B., VanRaden P.M., Van Tassell C.P., Null D.J., Schroeder S.G., Bickhart D., McClure M.C., 2013. Identification of a nonsense mutation in CWC15 associated with decreased reproductive efficiency in Jersey cattle. PLoS ONE, 8, e54872.

The 1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. Nature, 467, 1061-1073.

Thomsen B., Horn P., Panitz F., Bendixen E., Petersen A.H., Holm L.E., Nielsen V.H., Agerholm J.S., Arnbjerg J., Bendixen C., 2006. A missense mutation in the bovine SLC35A3 gene, encoding a UDP-N-acetylglucosamine transporter, causes complex vertebral malformation". Genome Res., 16, 97-105.

Thorvaldsdóttir H., Robinson J.T., Mesirov J.P., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics, 14, 178-192.

VanRaden P.M., Olson K.M., Null D.J., Hutchison J.L., 2011. Harmful recessive effects on fertility detected by absence of homozygous haplotypes. J. Dairy Sci., 94, 6153-6161.

Venhoranta H., Pausch H., Flisikowski K., Wurmser C., Taponen J., Rautala H., Kind A., Schnieke A., Fries R., Lohi H., Andersson M., 2014. In frame exon skipping in UBE3B is associated with developmental disorders and increased mortality in cattle. BMC Genomics, 15, 890.

Résumé

Avec le récent développement des approches reposant sur le génotypage et séquençage à haut débit, identifier la mutation responsable d'une anomalie à partir de quelques cas est beaucoup plus simple qu'auparavant. Toutefois, il n'est pas toujours possible d'observer des cas et de disposer du matériel biologique correspondant. Cet article présente deux approches reposant sur l'analyse de données

à haut débit, permettant d'identifier des mutations récessives létales ou d'orienter plus efficacement leur recherche à partir des données de séquence du génome. La première approche utilise les données de génotypage à haut débit pour rechercher des régions du génome présentant un déficit en homozygotes. Cette méthode a déjà permis de caractériser plusieurs mutations dans chacune des races bovines analysées. Dans la seconde approche, on recherche dans les données de séquence disponibles des variants dont l'annotation suggère qu'ils ne sont pas tolérés à l'état homozygote. Le nombre de faux positifs est élevé, mais ces données permettent d'orienter la phase d'observation et de diagnostic plus efficacement et, dans les cas les plus favorables, d'anticiper l'émergence de l'anomalie. Des exemples sont fournis avec le gène RP1 responsable de dégénérescence rétinienne ou le gène EDAR responsable de l'absence de poils et de dents.

Abstract

Anticipate the emergence of genetic defects from genomics data

With the recent development of approaches using high throughput genotyping and sequencing, identifying the causal mutation underlying a genetic defect from several cases has become much simpler. However, cases, or their corresponding biological material, are not always available. This article presents two approaches relying on high throughput data analysis to identify recessive lethal mutations or to efficiently orient their research from genome sequence. The first approach uses genotyping data to look for genomic regions presenting a deficit in homozygotes. This method has proven to be efficient with several mutations already characterized in each analyzed bovine breed. In the second approach, DNA variants are searched in the available whole genome sequences, with strong annotations suggesting that they are not tolerated in the homozygous state. The number of false positives is relatively high but these variants can orient the observation step toward mating at risk or homozygous animals and, in the most favorable cases, to anticipate the outbreak of the defect. Examples are provided with the RP1 gene responsible for retina degeneration or the EDAR gene responsible for the ectodermal anhidrotic syndrome.

FRITZ S., MICHOT P., HOZE C., GROHS C., BARBAT-LETERRIER A., BOUSSAHA M., BOICHARD D., CAPITAN A., 2016. Anticiper l'émergence d'anomalies génétiques grâce aux données génomiques. In : Anomalies génétiques. Boichard D. (Ed). Dossier, INRA Prod. Anim., 339-350.