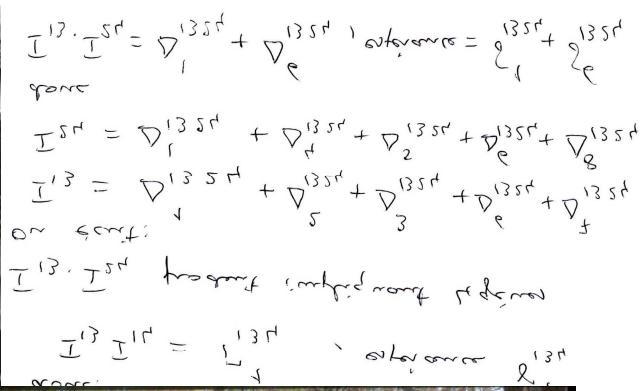
Institut National Polytechnique de Toulouse

Habilitation à Diriger des Recherches

De la Latxa à la Manech : au travers les méthodes et applications en amélioration et évaluation génétique.



```
            s=.5*(phi3(a,c,d)+phi4(ped(a,2),ped(a,3),c,d))
        else
        !abcd
            s=.5*(phi4(ped(a,2),b,c,d)+phi4(ped(a,3),b,c,d))
        endif
end function

recursive double precision function phi22(j,k,l,m) result(s)
        implicit none
        integer,intent(in):: j,k,l,m
        integer:: temp(4),a,b,c,d
        !double precision:: s
        ! I am following Karigl here because a,b,c,d is very MESSI ;-)
        a=j; b=k; c=l; d=m

        if((a*b==0).and.(c*d==0)) then
            s=0d0
        else
            ! all equal aa,aa
            if(all(a==(/b,c,d/))) then
                s=.25*(1+3*phi2(ped(a,2),ped(a,3)))
            else
                if(a<b) call swap(a,b)
```

Andrés Legarra Albizu
Chargé de Recherche, Station d'Amélioration Génétique des Animaux,
INRA Toulouse.

MEMBRES DU JURY

Mr Hervé REMIGNON	Professeur ENSAT, Toulouse – Président
Mr Vincent DUCROCQ	Directeur de Recherches INRA, Jouy en Josas - Rapporteur
Mr Alain CHARCOSSET	Directeur de Recherches INRA, Gif sur Yvette – Rapporteur
D Agustín Blasco Mateu	Professeur Universidad Politécnica de Valencia – Rapporteur
Mme Anne RICARD	Ingenieur en Chef des Ponts, des Eaux et des Forêts, Institut Français du Cheval et de l'Equitation - Examinateur
D Miguel Pérez Enciso	Professeur Universidad Autónoma de Barcelona – Examinateur

« Y hasta parece mentira, pero es cosa señalada,
que de una sangre pareja, salga la cría cambiada. » (José Larralde)

Quiero agradecer profundamente a todas las personas (demasiado numerosas para ser nombradas aquí) que me han ayudado en estos, pronto, 15 años de carrera en la investigación, dentro y fuera de ella; y a la flecha granate, porque hay una poesía de mate y de malvón.

Je tiens à remercier tous ces gens, trop nombreux pour en être cités, qui m'ont tellement aidé pendant tout ce temps de recherche.

I thank all the people (it makes for too long a list) who have been helping me throughout all these years.

Je remercie en particulier Christèle Marie pour la correction orthographique de ce manuscrit ; les erreurs sont entièrement de ma faute là où je n'ai pas suivi ses conseils.

Table de Matières

Table d’extraits des articles

1. Introduction

La génétique est une des disciplines de la science la plus fascinante ; l'amélioration génétique, fait partie de l'histoire de l'humanité depuis la domestication des animaux. Cette discipline scientifique est un mélange mystérieux de génétique, statistique, zootechnie, et maîtrise de la manipulation de grands volumes de données. Dans cette mémoire, je veux montrer mes quelques contributions.

Je suis arrivé à l'amélioration génétique grâce à l'un de mes professeurs à l'Ecole d'Agronomie de Pampelune (Javier Mendizábal), qui m'a introduit à ce monde fascinant. Bien que mon intention n'était pas de faire une thèse doctorale, un sujet de thèse proposé par Eva Ugarte me plaisait et j'ai postulé à (et eu) une bourse de thèse financé par l'Instituto Nacional de Investigaciones Agrarias (INIA, Madrid). Ainsi commençait une thèse encadré par Eva, chercheuse à ce qui s'appelait alors le CIMA (Centro de Investigación y Mejora Agraria), aujourd'hui NEIKER, et toujours situé à la « Granja Modelo de Arkaute » à coté de Vitoria. Le but de ma thèse était de travailler sur certaines améliorations dans le schéma de sélection génétique de la Latxa.

Aujourd'hui, de l'autre côté des Pyrénées, je travaille (entre autres choses) sur la Manech, « incarnation » nord-pyrénéenne de la Latxa. Quelque soit le qualificatif (ahari, béliers, carneros) ou l'espèce (des vaches, de poules, ou des souris), les animaux de rente génèrent toujours des questions de recherche fascinantes.

Dans ma carrière, j'ai travaillé sur plusieurs sujets. Aujourd'hui au moment de la réflexion j'en perçois une certaine cohérence, fruit de mes inquiétudes personnelles et celles de mes collaborateurs, cohérence aussi due aux besoins quotidiens. Mon fil conducteur est l'analyse de données et la recherche de méthodes pour l'amélioration génétique des espèces de rente. J'essaierai de présenter mes travaux dans ce mémoire, ordonnés de la manière suivante :

- Les objectifs de sélection, ou comment et pourquoi les schémas sélectionnent des différents caractères
- La modélisation des caractères non-gaussiens
- Localisation des QTLs et « phasage » des marqueurs
- L'évaluation génétique (et « génomique »)
- L'écriture des logiciels d'estimation et calcul

2. Les objectifs de sélection

Les principes classiques de l'amélioration génétique fournissent des méthodes pour la sélection multicaractère. Ils font l'hypothèse de normalité multivariée des valeurs génétiques pour les différents caractères. La sélection multicaractère idéale passe par plusieurs étapes :

- La définition d'un objectif de sélection multicaractère,
- L'estimation des composantes de la variance associées à chaque effet du modèle (génétique, résiduelle, environnementale aléatoire, etc.),
- La vérification de l'efficacité et la faisabilité technique et économique de l'objectif de sélection multicaractère,
- Et sa mise en route moyennant un schéma de contrôle de performances, d'évaluation et de sélection pertinente.

Ces étapes sont des activités classiques d'un « animal breeder ».

2.1. L'introduction des nouveaux objectifs de sélection dans le schéma de la brebis laitière de race Latxa

Classiquement, les schémas de sélection de brebis laitière commencent par une sélection pour la quantité du lait, caractère relativement facile à mesurer, pour ensuite considérer d'autres caractères d'intérêt économique et zootechnique, comme la richesse du lait, la morphologie de la mamelle, le comptage de cellules somatiques ou encore la résistance à la tremblante.

2.1.1. Estimation de paramètres génétiques

En 1998 il était temps de considérer l'introduction de certain de ces caractères dans la sélection de la Latxa, ce que j'ai fait comme part de mes travaux de thèse. J'ai estimé d'abord les composantes de variance (héritabilités, corrélations génétiques, etc.) des caractères de quantité et richesse du lait (protéine, matières grasses) [A1 ; thèse], par un REML multicaractère. Les deux estimations utilisaient deux jeux des données légèrement différents. Le premier jeu de données utilisait les performances enregistrées entre 1989-1991 et 1996-1998 lors de différents protocoles expérimentaux qui impliquaient environ d'une douzaine de troupeaux; le second jeu, des données de contrôle laitier « routinier » des campagnes 1999-2000 et 2000-2001. Les deux jeux comprenaient ≈6000 animaux avec phénotypes. Les résultats étaient légèrement differents, mais ils étaient bien classiques : corrélations génétiques fortes (>0,85) entre les quantités de lait, de matières grasses et protéiques ; corrélations génétiques négatives (≈-0,3) entre la quantité de lait et les taux ; des héritabilités moyennes pour les quantités (≈0,2), une héritabilité forte pour le taux protéique (0,5) – mais pas pour le taux butyreux (0,2), ce que nous attribuons toujours à des problèmes d'échantillonnage correct des composantes du lait, qui s'avère délicat en pratique. Voici un exemple de résultat :

Table 5 *Heritabilities (on the diagonal), genetic (above the diagonal) and phenotypic (below the diagonal) correlations between traits*†

	MY	FY	PY	F%	P%
MY	**0·20**	0·88	0·93	–0·45	–0·51
FY	0·85	**0·16**	0·88	0·02	–0·36
PY	0·96	0·84	**0·18**	–0·28	–0·15
F%	–0·21	0·29	–0·14	**0·14**	0·41
P%	–0·25	–0·12	0·02	0·25	**0·38**

Extrait 1 Paramètres génétiques de la composition du lait en Latxa [A1]

Une autre jeu de données d'une taille similaire fut utilisé pour estimer les composantes de la variance pour des caractères de morphologie mammaire, pendant et après ma thèse [thèse ; A5], les résultats étant similaires : des héritabilités modérées (0,2-0,4) et une légère opposition entre l'augmentation de la production de lait et la bonne forme de la mamelle. Néanmoins, les tendances génétiques ne montraient aucune détérioration [B3, B9], ce que nous attribuons à une sélection phénotypique dans la voie « mère à bélier ». On peut noter que les données de morphologie mammaire présentaient, à la différence de la plupart des caractères de morphologie, des mesures répétées intra-lactation et entre lactations, ce qui a demandé une modélisation spécifique et des astuces calculatoires pour ajuster le bon modèle dans les logiciels de l'époque (VCE 4.0).

En fin, la relation entre les caractères de morphologie mammaire, la production de lait et le comptage cellulaire (« somatic cell score ») fut étudiée [A5]. Les résultats montraient une corrélation génétique favorable mais faible entre la quantité de lait et le comptage cellulaire, ce que nous attribuons à un manque de linéarité dans la relation entre les deux (cet aspect sera repris plus tard), de telle manière que de faibles comptages indiquent une bonne santé de la mamelle, avec des bons niveaux de productions. Ci-dessous, un exemple de ces dernieres analyses.

Table 3. Genetic parameters [heritabilities (in bold on diagonal) and genetic correlations, ± SE of the estimates] of udder type traits, milk yield, and LSCS,[1] considering all lactations, including MYDS[2] in the model.

	Udder depth	Udder attachment	Teat placement	Teat size	Milk yield	LSCS
Udder depth	**0.26 ± 0.02**	–0.58 ± 0.05	–0.42 ± 0.04	–0.05 ± 0.05	0.43 ± 0.05	0.10 ± 0.07
Udder attachment		**0.26 ± 0.02**	0.34 ± 0.04	0.05 ± 0.05	0.10 ± 0.06	–0.27 ± 0.07
Teat placement			**0.40 ± 0.02**	0.31 ± 0.04	–0.25 ± 0.05	–0.01 ± 0.06
Teat size				**0.40 ± 0.02**	–0.10 ± 0.05	0.29 ± 0.06
Milk yield					**0.21 ± 0.02**	–0.30 ± 0.07
LSCS						**0.13 ± 0.02**

[1]LSCS = Lactational SCS.
[2]MYDS = Milk yield produced in the day of scoring.

Extrait 2 Paramètres génétiques de morphologie mammaire et comptage cellulaire en Latxa [A5]

2.1.2. Construction des objectifs de sélection

Une fois les paramètres génétiques estimés, il est possible de prédire les gains génétiques dus à l'utilisation d'un certain objectif de sélection. On peut choisir les poids de chaque caractère dans l'objectif de sélection de manière à avoir une réponse souhaitée ; ce processus est denommé comme « desired gains » ou gains souhaités. En l'absence de toute autre information économique, nous avons conçu des poids des caractères qui permettaient un gain optimale pour les caractères dont l' intérêt économique était clair (quantité de lait) sans détérioration des niveaux génétiques de morphologie mammaire et de richesse du lait [Thèse].

L'estimation des gains génétiques, tant pour les cas multicaractère et unicaractère, doit utiliser une modélisation « idéalisé » du schéma de sélection. Cette modélisation fut faite en d'accord avec les données pratiques du schéma ; plusieurs voies de progrès génétique (père de femelles, etc.) ont été décrites, et le gain génétique théorique calculé selon les formules classiques de Rendel & Robertson (1950). Ce gain génétique fut comparé à celui estimé lors des évaluations génétiques, tous les deux étant très similaires [A45].

Cuadro 4. Cálculo del progreso genético teórico en Latxa Cara Rubia
Table 4. Theoretical genetic progress in Blond-Faced Latxa

Vía	_G	L	ponderación
Hembra-Macho	1,27	4,58	1
Hembra-Hembra	0,24	4,76	1
Macho de Monta Natural-Hembra	-0.30	4,15	0,60
Macho en Testaje-Hembra	0,54	1,96	0,19
Macho mejorante-Hembra	1,16	5,14	0,21
Macho-Macho	1,16	5,14	1
Ganancia genética total anual (medida en desviaciones estándar genéticas)	0,154		

Extrait 3 Calcul du progrès génétique en Latxa Cara Rubia [A45]

L'extension au cas multivarié pour le calcul des progrès génétiques selon différents objectifs de sélection fut immédiat [Thèse]. Cette approche de gains souhaités reste néanmoins peu satisfaisante, car économiquement sous-optimale. Effectivement, on paie un prix pour le maintien des niveaux génétiques : ce que l'on ne gagne plus sur les autres caractères.

2.1.3. Poids économiques

Il a été montré dans la littérature concernant l'amélioration génétique que, en sélection (ce qui est le cas pour la plupart des espèces), les poids optimaux attribués aux caractères dans l'objectif de sélection sont déterminés par un critère simple : le gain marginal de bénéfice produit pour un gain dans le caractère d'intérêt (techniquement, la première dérivée de la fonction de bénéfice par rapport au caractère) (e.g., Goddard, 1998). La difficulté réside dans le calcul de ce gain marginal. Nous avons donc eu un projet avec financement INIA pour le calcul des poids économiques dans les schémas de sélection ovins laitiers (Manchega, Latxa). Ce travail fut le premier, à notre connaissance, dans le monde ovin laitier. Nos collègues des services d'appui à l'élevage en Pays Basque ont collecté une grande quantité de données techniques et économiques, issus d'une trentaine de fermes. En parallèle, nos collègues qui étudient la race Manchega ont entamé le même travail, notamment le thésard Manuel Ramón

Fernández (CERSYRA, Valdepeñas, Espagne ; encadré par María Dolores Pérez Guzmán) avec qui j'ai beaucoup travaillé. J'ai donc créé, à partir des données techniques et de gestion de cet ensemble de troupeaux, des équations de bénéfice comme fonction des caractères d'intérêt. Je passe sur les résultats détaillés pour deux ensembles différenciés de caractères [A7, A8].

Pour des caractères de production « classiques » (fertilité, prolificité, quantité de lait et longévité), nous avons estimé des poids économiques pour chacun des troupeaux ainsi que pour l'ensemble. Les moyennes des poids économiques étaient 138,60 € par agnelage, 40,00 € par agneau, 1,18 € par litre et 1,66 € par année de survie. Les différents gains génétiques avec des fonctions de bénéfice légèrement différentes (calculées selon les paramètres techniques de différents troupeaux) étaient très similaires, cf. l'extrait ci-dessous : on déduit que les schémas sont robustes à certaines légerès déviations dans le calcul des poids.

Table 5 *Genetic gains in the Latxa and Manchega breeds after 1 year of genetic improvement based on different economic weights*

	Criteria[†]				
	Set I, no rescaling	Set I, rescaling	Set II, no rescaling[‡]	Set II, rescaling[‡]	Set III
Latxa					
Fertility	0.004	0.003	0.004	0.004	−0.006
Prolificacy	0.010	0.010	0.010	0.010	0.002
Milk yield (kg)	2.61	2.71	2.57	2.61	3.06
Longevity (years)	0.013	0.011	0.013	0.012	0.004
Profit gain per ewe (€)	4.20	3.69	4.22	3.73	2.83
Profit gain per farm (€)	1746	1527	1757	1546	1189
Manchega					
Fertility	0.008	0.008	0.008	0.008	−0.007
Prolificacy	0.015	0.015	0.015	0.015	0.002
Milk yield (kg)	2.36	2.26	2.31	2.32	3.77
Longevity (years)	0.012	0.012	0.012	0.013	0.004
Profit gain per ewe (€)	3.39	2.89	3.51	2.94	1.97
Profit gain per farm (€)	3178.28	2707.94	3386.22	2829.38	948.45

[†] Sets: (I) using the 'median' set of economic weights; (II) using farms' individual economic weights; (III) using milk yield as the only trait in the merit index.
[‡] Average genetic gains.

Extrait 4. Gains génétiques avec différents poids économiques en Latxa et Manchega [A7]

De plus, il est bien connu que les données économiques sont souvent difficiles à décrire. Par exemple, l'amortissement des différents investissements (bâtiments, terre) est très difficile à estimer correctement. J'ai calculé des poids économiques selon un autre approche dit « rescaling » (remise à l'échelle) (Visscher et al., 1994) qui consiste à contraindre la taille du troupeau à la ressource la plus limitante – par exemple, la surface agricole –, en considérant que cette dernière a un coût infini, et de manière à ce que le bénéfice de l'entreprise ne soit pas potentiellement infini. Les résultats ont été très similaires à l'approche classique ci-dessus. Une des conclusions de ce travail fut que la fertilité reste un des caractères économiquement les plus importants, même si évaluation génétique en routine reste à faire en brebis laitière.

Un autre étude a porté sur le poids économique du « score » de cellules somatiques (SCS) dans le lait. Ce fut une nécessité car les relations de ce score avec les pertes et bénéfices de la ferme sont complexes ; en effet, le SCS est un indicateur de mammites, avec une potentielle réforme de l'individu et/ou besoin de traitements vétérinaires, et une réduction de production (qui est incluse dans la corrélation génétique avec la quantité du lait); et, en soit, le comptage cellulaire peut induire une chute du prix du lait. Les trois premiers facteurs ont été considérés *via* des fonctions relativement compliquées. Par exemple, la chute dans le prix du lait dépend de la distribution des comptages individuelles des brebis, qui est une fonction complexe qui

dépend de la taille du troupeau, de son niveau moyenne de cellules, et de la variabilité naturelle intra-troupeau. Le comptage cellulaire suit donc une distribution dite log-normale. De vraies données techniques ont été utilisées. Les poids économiques étaient donc non linéaires, comme montré dans la figure ci-dessous.

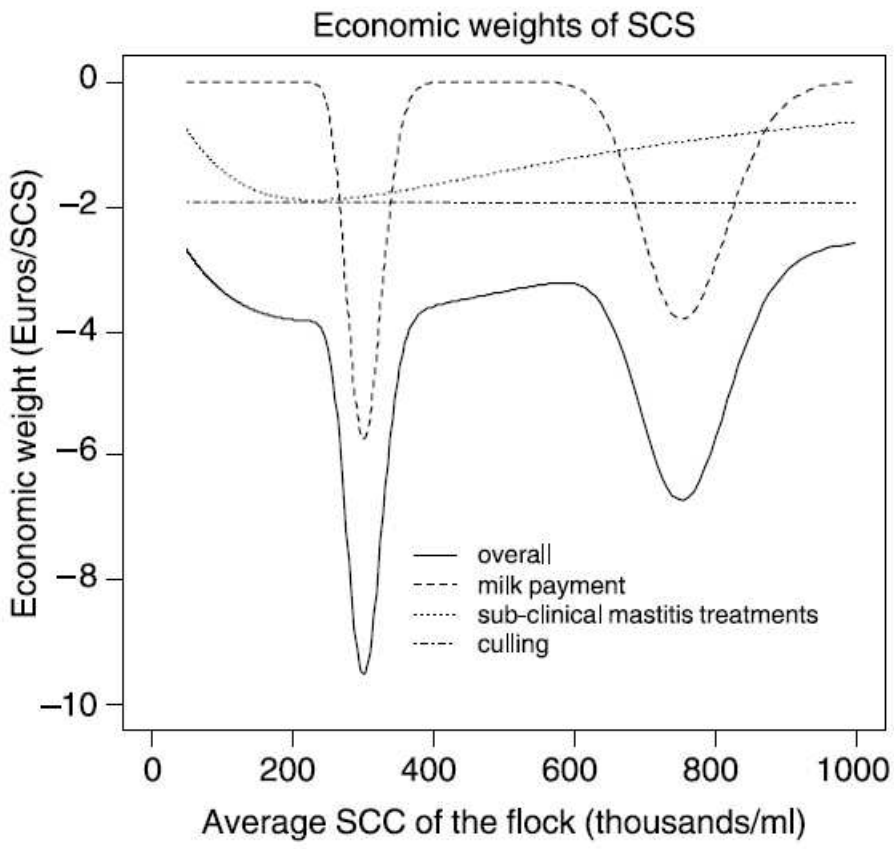


Figure 1 Overall economic weight of somatic cell score (SCS) and its components for different levels of somatic cell count (SCC) of the flock. The assumed genetic correlation between 'culling by mastitis' and SCS is 0.7.

Extrait 5 Poids économiques du score cellulaire [A7]

De plus, nous avons calculé des gains génétiques selon différents objectifs de sélection. Grossièrement, l'augmentation de bénéfice obtenu lors de l'inclusion du SCS dans la fonction de bénéfice est très légère, à cause de sa faible héritabilité : nous avons conclu que son inclusion dans l'objectif de sélection n'est pas prioritaire, sauf si le contrôle de performances pour le SCS a un cout négligeable.

En complément de ces travaux, Manuel Ramón a aussi travaillé sur la modélisation du poids économique des composants du lait [A20].

2.2. Sélection pour la résistance à la tremblante

Contexte historique : suite à la découverte du gène contrôlant la susceptibilité/résistance à la tremblante (Prnp) et aux répercussions médiatiques de l'encéphalopathie spongiforme bovine, les autorités sanitaires de l'Union Européenne ont émis une directive suggérant la sélection

des animaux porteurs d'allèle(s) de résistance ARR. Plusieurs schémas de sélection espagnols (y compris celui de la Latxa) s'y sont opposés, car la tremblante n'a jamais été un problème sévère en Espagne (à différence, par exemple, de la France avec la Manech Tête Rousse), et le surcoût de travail, la chute de progrès génétique et, surtout, la déstabilisation de tous les acteurs de la filière risquaient d'être sérieux. Finalement, le plan de sélection établi par le Ministère de l'Agriculture a été beaucoup moins ambitieux que prévu initialement.

J'ai donc été impliqué dans l'évaluation des effets potentiels d'une sélection des allèles résistants à la tremblante sur la variabilité génétique de la population Latxa [A6], dont l'auteur principal a été Leopoldo Alfonso (UPNA, Pamplona, Espagne). La perte de variabilité a été étudié avec des méthodes quantitatives (généalogies) et moléculaires (microsatellites). Les résultats étaient ambigus : selon l'indicateur ou méthodologie (infinitésimale, moléculaire) choisi la perte était acceptable ou pas. La recommandation était d'une sélection progressive avec suivi de la variabilité par méthodes de génétique moléculaire, qui sont insensibles à l'information généalogique incomplète.

De même, j'ai travaillé sur une évaluation des effets potentiels du gène Prnp sur la production laitière, effets peut-être dus à un déséquilibre de liaison au sens large [B59]. J'ai utilisé trois méthodes, avec trois jeux de données chevauchantes : un dispositif de génotypage sélectif (brebis à forte et faible production laitière), une analyse en deux étapes (valeur génétique estimé –EBV– comme fonction du génotype), et une analyse directe avec le phénotype au caractère. Plusieurs méthodes furent utilisées : maximum de vraisemblance, test de répartition (Chi-2), statistiques bayésiennes (BIC). Aucune des méthodes indiqua une relation entre le gène Prnp et la production laitière.

Globalement, les résultats de deux travaux sont conformes à la littérature existante.

2.3. Conclusion

Quel message peut-on tirer de cette partie ? La génétique quantitative a des outils conceptuels et pratiques qui servent à confronter des problèmes réels dans le déroulement quotidien des schémas de sélection. Ces questions sont d'une importance capitale et font de très bons objets de recherche. J'ai détaillé dans la partie concernant les poids économiques certaines originalités méthodologiques de nos développements. Dans la prochaine section, je détaillerai d'autres modélisations originales pour l'analyse génétique des caractères.

3. La modélisation des caractères

Les caractères quantitatifs sont, le plus souvent, considérés sous une distribution normale multivariée, sa relation étant décrite par des matrices de variances et covariances propres à chaque effet agissant sur le caractère. Ainsi, nous avons des « covariances génétiques », des « covariances résiduelles », etc.
Cette modélisation semble adéquate la plupart de temps. Néanmoins, il existe des caractères dont biologiquement on ne peut pas accepter l'hypothèse de normalité multivariée. J'ai été confronté à plusieurs de ces caractères, et j'ai contribué à leurs modélisations ainsi que à la résolution des problèmes pratiques d'estimation.

3.1. Caractères à seuils

Certains caractères ne s'observent pas ou s'enregistrent que sous la forme des données catégorielles ordonnées ; par exemple, la taille de portée, la fertilité ou la facilité de vêlage. C'était ainsi pour le caractère de facilité de vêlage chez la vache laitière ou encore sa fertilité. Evangelina López de Maturana, thésarde d'Eva Ugarte (2003-2006) à NEIKER, travaillait sur des aspects liés à l'évaluation génétique de ce caractère, et notamment la relation entre la facilité de vêlage et les caractères reproductifs. Il est bien connu que des problèmes de vêlage réduisent la fertilité des femelles ; mais cette relation restait à quantifier à l'époque.

Nous avons suivi une modélisation classique : le modèle à seuil, qui postule un phénotype « non observé » sous-jacente (« la sous-jacente ») qui a une distribution normale. La position de la sous-jacente par rapport à certaines valeurs de seuils fait exprimer un phénotype observé dans une certain catégorie (e.g. Gianola, 1982).

Evangelina a donc analysé un jeu de données complet et compliqué, car incluant des nombreux caractères à modélisation complexe. Les caractères à seuils étaient la facilité de vêlage (3 catégories), et le succès de la première insémination (2 catégories). Les autres caractères étaient le nombre d'inséminations pour une réussite, l'intervalle entre vêlage et conception (« days open »), et les jours jusqu'à la première insémination. De même, des donnés censurés et de la récursivité (sera détaillée plus tard) existaient pour ce jeu de données. Pour les données censurées (par exemple, une vache qui abandonne son troupeau à une date connue mais avant le vêlage), on fait l'hypothèse d'un phénotype « au delà » de l'enregistrement fait. C'est-à-dire, si la vache a fait trois inséminations avant de quitter le troupeaux, le nombre d'inséminations pourrait être trois, quatre ou plus car on ne connaît pas le résultat de la troisième insémination.

Nous avons choisi une méthode bayésienne, car la modélisation des données non-multinormales se fait très naturellement. Grâce aux techniques d' échantillonnage de Gibbs et « data augmentation » (Tanner, 1996), on peut échantillonner les phénotypes « non visibles », que ce soit la sous-jacente pour un caractère discrète, la donnée non observée pour un caractère censuré, ou autres. A l'aide de Luis Varona (IRTA, Lleida, Espagne) et Evangelina, j'ai conçu et programmé l'outil nécessaire (qui déviendrait le « TM », voir plus tard). Cette technique a permis à Evangelina d'estimer les relations entre tous les caractères [A9], confirmant que les problèmes de vêlage réduisent la fertilité.

3.2. Modèles récursifs

Les modèles classiques postulent une relation linéaire entre caractères décrite par une covariance. Ils ne « savent » pas dire (ou considérer) si l'un des caractères et cause et l'autre conséquence. De plus, la covariance a une interprétation linéaire, dans le sens où une augmentation de Δx du caractère A s'accompagne statistiquement d'une augmentation Δy du caractère B et cela, quelque soient les valeurs initiaux de A ou de B. Par exemple, il est concevable qu'une bonne santé mammaire (faible comptage cellulaire) soit accompagné d'une forte production laitière –covariance positive–, mais au-delà d'un seuil de production laitière la brebis est affaiblie et la santé se déteriore –covariance négative–.

Les modèles récursifs postulent que le caractère A est une fonction (eventuellement non-linéaire) du caractère B. Bien qu'étudiés en économétrie, les modèles récursifs ont été peut étudiés en génétique quantitative, mais un article de Gianola et Sorensen (2004) en a ravivé l'intérêt. Lors de la thèse d'Evangelina, l'étude de l'effet de la facilité de vêlage sur la fertilité suggérait une modélisation récursive et cela, pour deux raisons. D'abord, parce qu'il est biologiquement évident que la difficulté de vêlage va conditionner la fertilité *après* et donc, le vêlage est une cause de la fertilité et non l'inverse. En plus, il y a trois catégories de vêlage : 1 (simple), 2 (aidé), et 3 (compliqué). On ne s'attend pas à ce que passer de 1 à 2 aie autant d'effet négatif sur la fertilité que de passer de 2 à 3. C'est pour cela que nous avons postulé une relation de type « à niveaux » entre la facilité de vêlage (cause) et la fertilité (conséquence), relation du type :

$$y_i = \begin{pmatrix} x_{i,1} & x_{i,2} & x_{i,3} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

Équation 1

où y_i est le phénotype pour la fertilité et x_i est une matrice d'incidence ayant 1 dans la case correspondant au phénotype observé pour la facilité de vêlage et 0 si non (« dummy variables », ou effet croisé), et où β_j est l'effet du *j*-ème phénotype.

J'ai donc programmé ce modèle dans le logiciel TM et il a été utilisé par Evangelina dans l'étude décrite [A9]. On a pu quantifier l'effet exact d'un vêlage difficile : 0,5 insémination de plus, par exemple, comme montré ci-dessous :

Table 7. Posterior means and standard deviations (SD) for the effects of calving ease scores on fertility traits

	Calving ease score[2]					
	1		2		3	
Trait[1]	Mean	SD	Mean	SD	Mean	SD
DO, d	0.00	0.00	12.63	6.79	31.14*	14.97
DFS, d	0.00	0.00	3.34†	2.49	7.65†	5.54
NINS, no.	0.00	0.00	0.19*	0.09	0.51*	0.21
OFI-liability	0.00	0.00	–0.13*	0.06	–0.32**	0.14
OFI-observed scale, %	0.00	0.00	–5%*	2%	–12%**	5%

[1]DFS = days to first service; DO = days open; NINS = number of services per pregnancy; OFI-liability = estimate of conception rate on the underlying scale; OFI-observed scale = estimate of outcome at first insemination (expressed as percentage) on the observed scale.

[2]Calving ease scores: 1 = no problem; 2 = slight assistance; 3 = needed assistance or caesarean because of the calf's size (dystocia). Estimates for unassisted parities were set to 0. Estimates for second and third calving ease scores are contrasts to the first score.

†$P < 0.10$; *$P < 0.05$; **$P < 0.01$.

Extrait 6 Effet de la facilité de vêlage sur la fertilité [A9]

Pour effectuer la programmation d'un tel modèle, il m'a fallu comprendre le modèle en profondeur. Je me suis donc rendu compte que finalement une reparamétrisation simple permettait d'obtenir les mêmes résultats (la même vraisemblance était utilisée) et ceci, dans notre cas de modèle récursif non simultanée. On transformait l'équation décrite dans Gianola et Sorensen :

$$\begin{pmatrix} 1 & -\lambda \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1\mathbf{b}_1 + \mathbf{u}_1 + \ldots \\ \mathbf{X}_2\mathbf{b}_2 + \mathbf{u}_2 + \ldots \end{pmatrix}$$

Équation 2

Par cet autre modèle équivalent :

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1\mathbf{b}_1 + \mathbf{u}_1 + \lambda\mathbf{y}_2 + \ldots \\ \mathbf{X}_2\mathbf{b}_2 + \mathbf{u}_2 + \ldots \end{pmatrix}$$

Équation 3

Et nous avons démontré que cette équivalence conduit à une estimation simple [A9, B10] – ce dernier en collaboration avec C. Robert-Granié.

Ce type de modèle a encore eu une tournure. Varona et al. (2007) ont montré que l'Équation 2 et l'Équation 3 sont aussi équivalents à un système normal multivariée classique, dans lequel le coefficient λ est « emboité » dans les matrices de covariances. Par contre, on peut postuler des systèmes plus sophistiqués, comme notre Équation 1, ou bien des modèles non linéaires comme :

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1\mathbf{b}_1 + \mathbf{u}_1 + \lambda_1\mathbf{y}_2 + \lambda_2\mathbf{y}_2^2 + \ldots \\ \mathbf{X}_2\mathbf{b}_2 + \mathbf{u}_2 + \ldots \end{pmatrix}$$

Équation 4

dans lesquels on peut aisément modéliser des relations non linéaires entre les caractères. Nous avons fait ainsi pour [A9] et aussi dans le travail de Llibertat Tusell (IRTA, Barcelonne, Espagne ; thésarde de Miriam Piles), qui montre les relations non-linéaires (en fait presque linéaires) entre pH et fertilité [A24], cf. la figure suivante :

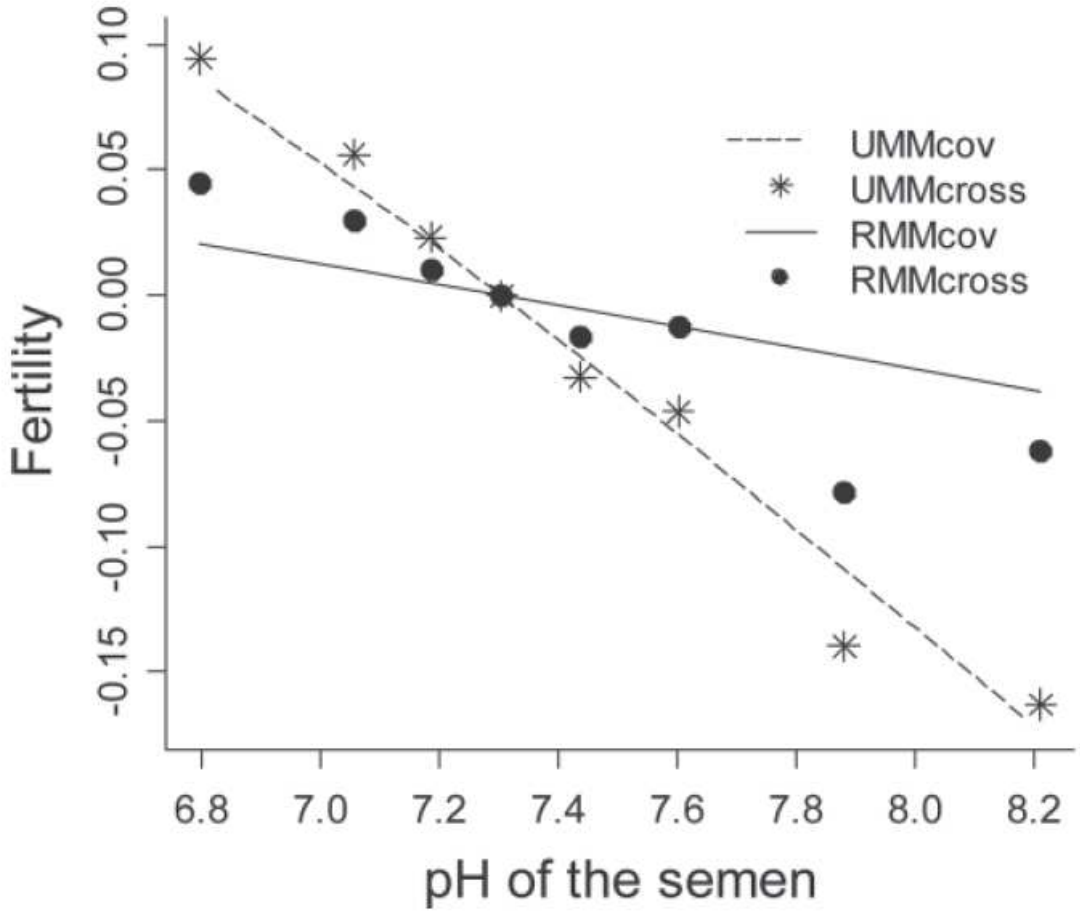


Figure 1. Effect of pH on fertility on the observed scale in the different models for pH of the semen and fertility: a mixed model without genetic and environmental correlations including pH as a covariate or as cross-classified effect in the model of fertility (models UMM_{cov} and UMM_{cross}, respectively), and recursive mixed models including pH as a covariate or as a cross-classified effect in the model of fertility (RMM_{cov} and RMM_{cross}, respectively).

Extrait 7 Effet du pH sur la fertilité avec des modèles récursifs. [A24]

3.3. Modèle produit

On peut considérer que le phénotype est une fonction de l'individu et de son environnement. Par contre, si dans l'environnement il y a d'autres individus, et ces individus ont à leur tour un déterminisme génétique pour le caractère, ce dernier mérite d'être considéré. Un exemple classique d'un tel phénomène est le modèle à effets maternels; un exemple plus à la mode sont les modèles à compétition (Bijma, 2011), dans lesquels, par exemple, un porcelet grandit comme fonction de sa capacité à grossir, mais aussi de la capacité des autres porcelets hébergés dans la même cage à lui empêcher de manger.

Un cas tout à fait particulier est la conception suite à un accouplement. Ceci était l'objet de la thèse d'Ingrid David (encadré par L. Bodin, E. Manfredi et C. Robert-Granié, de la SAGA). En effet, ce phénotype résulte de l'action jointe (et complexe) d'un ovule et un

spermatozoïde. D'abord, le phénotype observé est binaire (oui/non) ce qui suggère un traitement selon un modèle à seuil comme décrit précedemment. Traditionnellement, le modèle utilisé était un modèle additif qui sommait les effets du mâle et de la femelle dans l'échelle sous-jacente:

$$l_{i,j} = \ldots + u_{i,1} + u_{j,2}$$

Équation 5

où la sous-jacente $l_{i,j}$ suite à l'accouplement du mâle *i* avec la femelle *j* était la somme des effets des deux individus. Ceci n'était pas satisfaisant pour deux raisons ; d'une part, on sait biologiquement que les effets compensatoires sont limités (un spermatozoïde non viable ne sera jamais compensé par un ovule très viable), d'autre part cette modélisation ne permet pas de séparer (estimer séparément) les effets fixes liées à la fécondité « mâle » de ceux liés à la fécondité « femelle » quand leurs matrices d'incidence étaient confondues (par exemple, la saison). En reproduction humaine, les modèles proposent deux pseudo-phénotypes, la production d'un spermatozoïde viable (oui/non) de la part du mâle, et la production d'un ovule viable (oui/non) par la femelle. Mais le phénotype observé demeure toujours la fertilité globale (oui ou non).

J'ai donc participé à l'implémentation bayésienne de cette idée, la solution étant simple : il s'agissait d'augmenter (« data augmentation ») le problème avec les pseudo-phénotypes non observés, disons y_1 et y_2 (ovule fertile, spermatozoïde fertile). A son tour, ces deux phénotypes étaient analysés avec un modèle à seuil bi-caractère, et donc avec deux variables sous-jacentes l_1 et l_2, de la manière suivante.

$$\begin{pmatrix} \mathbf{l}_1 \\ \mathbf{l}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{u}_1 + \ldots \\ \mathbf{u}_2 + \ldots \end{pmatrix}$$

Toute la difficulté demeurait dans l'échantillonnage de y_1 et y_2 qui s'avéra finalement facile. La distribution complète conditionnelle (« full conditional ») nécessaire pour l'échantillonnage de Gibbs est une fonction en deux parties :

- D'un coté la probabilité « sachant le phénotype » :

$$Fertile \rightarrow y_1 = 1, y_2 = 1$$

$$Non\ fertile \rightarrow \begin{cases} y_1 = 0, y_2 = 0 \\ y_1 = 0, y_2 = 1 \\ y_1 = 1, y_2 = 0 \end{cases}$$

- D'autre coté la probabilité « sachant les inconnues du modèle » :

$$\Pr\left(y_1 = 0\right) = \phi\left(\hat{l}_1\right)$$

$$\Pr\left(y_2 = 0\right) = \phi\left(\hat{l}_2\right)$$

où ϕ est la fonction cumulative de probabilité normale, et $\hat{l}_1$ est la valeur de la sous-jacente pour le caractère, c'est-à-dire la somme de tous les effets (fixes, aléatoires) inclus dans le modèle, aux valeurs échantillonnées à chaque itération.

En multipliant les deux probabilités (sachant les données et les inconnues) on obtient les probabilités des quatre événements, on tire y_1 et y_2 simultanément, et on suit comme dans le déroulement d'un modèle à seuil bicaractère.

J'ai fait ce développement, qu'Ingrid a introduit dans une version *ad hoc* du logiciel TM. Ensuite, elle a montré les avantages de la méthode en interprétation de ses résultats, ainsi que la facilité d'estimation par simulation [A14]; ce modèle a été aussi utilisé pour estimer séparément les effets concernant le mâle et la femelle [A26], lors de travaux de thèse de Llibertat Tusell.

3.4. Résultats de courses hippiques

Les résultats de courses sont un caractère particulièrement compliqué, car il ne s'agit pas d'un enregistrement en unités physiques quelconques. Le résultat d'une course est, en soi, une comparaison entre les chevaux qui y ont participé. Classiquement, l'analyse de rangs se fait par un modèle dit Thurstonien (e.g. Tavernier, 1991). Ce modèle fait encore l'hypothèse d'une performance sous-jacente, dont le rang pour les différents chevaux conduit aux positions dans la course. La manipulation des distributions normales de ces performances sous-jacentes implique le calcul de certaines intégrales très complexes ; pour l'évaluation génétique des chevaux, ce calcul fut accompli par des approximations de deuxième degré. L'approche est l'œuvre d'Anne Ricard (Haras Nationaux, mais basé à ma même unité SAGA).

Récemment, Gianola et Simianer (2006) ont suggéré l'utilisation d'un algorithme de type échantillonnage de Gibbs pour l'implémentation du modèle Thurstonian. Ce algorithme est très similaire à celui pour le modèle à seuil. Anne a perçu cet algorithme comme beaucoup plus simple et bien adapté aux questions de recherche sur l'analyse de rangs. Une question immédiate : est-ce que l'algorithme d'échantillonnage de Gibbs permettait d'estimer des parts de variance ? (Ce qui restait très compliqué pour Tavernier (1991).)

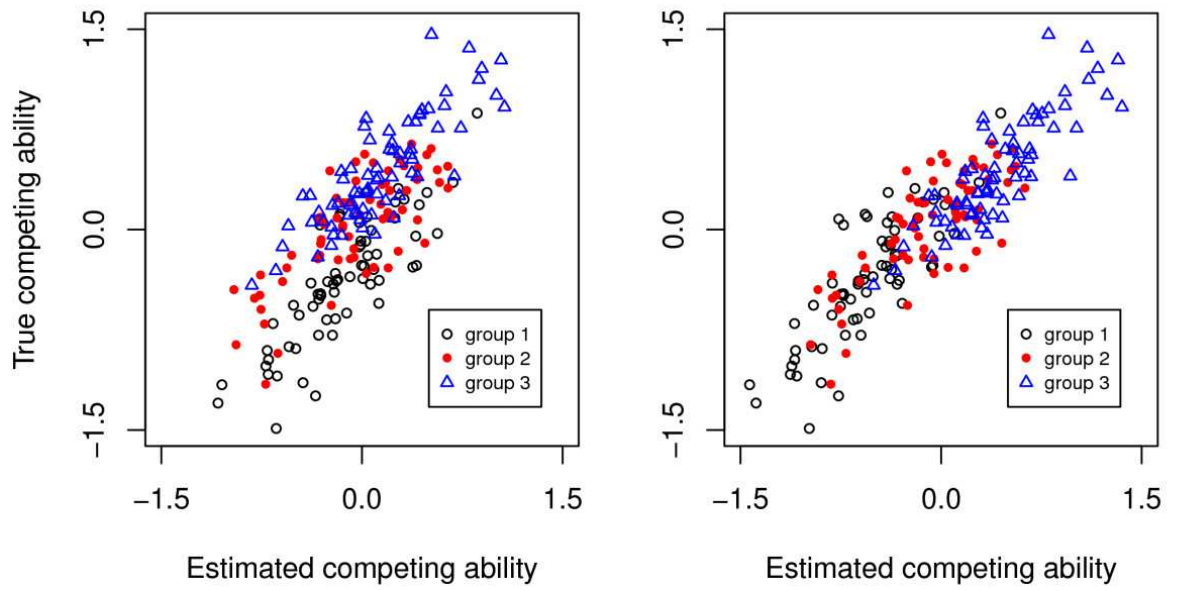
D'autre part, la répartition des chevaux dans les courses ne se passe pas au hasard ; les organisateurs de courses ont une tendance à regrouper des chevaux de « qualités » similaires. Il n'y a pas d'effet de la « course » en soi, car le résultat d'une course est un rang ; qui sera le même que la course soit « difficile » ou pas (et ceci a été démontré algébriquement dans l'article dont on parlera). Le même phénomène se passe pour de « catégories » de courses. Comment faire alors ?

Nous avons donc attaqué les deux problèmes par des développements théoriques et des simulations. J'ai surtout participé aux aspects d'implémentation de l'échantillonnage de Gibbs (notamment nous avons réfléchi ensemble au tirage des sous-jacentes, comme on verra) et à l'interprétation des résultats.

Le premier résultat fut que l'algorithme de Gianola et Simianer n'était pas correct. Le tirage de la sous-jacente d'un cheval doit se faire de la manière suivante. Si un cheval *i* est classé entre les chevaux *i-1* et *i+1*, sa performance non observée (disons l_i) doit respecter la règle $l_{i-1} < l_i < l_{i+1}$ et donc elle doit être tirée entre les sous-jacentes des chevaux précedent et suivant dans la course. Or, pour le dernier cheval, le système de Gianola et Simianer tirait sans condition. Nous avons montré ce fait et expliqué l'algorithme correct.

De plus, avec l' échantillonnage de Gibbs et l'algorithme corrigé, les paramètres génétiques étaient estimés correctement, ce qui ne fut pas le cas pour d'autres méthodes « simples » (« normal scores », par exemple) qui tendait à la sous-estimation. Ceci montra l'adéquation de l'échantillonnage de Gibbs aux problèmes de rangs.

Par contre, dès que l'allocation des chevaux aux courses n'était pas au hasard, les paramètres génétiques étaient biaisés. Nous avons donc conçu un modèle de mélange : on alloue à priori les chevaux (et pas les courses) à différentes catégories (modèle de mélange), pour ensuite estimer les effets des différentes catégories par un modèle linéaire ; ceci n'est possible que si des chevaux des différentes catégories participent à la même course. Notre modèle ainsi construit n'était plus biaisé (voir figure ci-dessous). L'implémentation pratique en routine d'évaluation génétique reste à faire, car l'allocation des chevaux à des catégories reste un problème délicat.



Extrait 8 True and estimated competing ability, underlying model for ranks (left), underlying mixture model for ranks (right) [A19]

3.5. Modélisation des données longitudinales

Les données du contrôle de performances sont souvent longitudinales. Par exemple, les contrôles laitiers se suivent le cours de la vie productive d'une femelle laitière, et les enregistrements de poids le long de la croissance d'un individu. Dans un souci de précision et de généralité, et pour s'affranchir de précorrections lourdes ou complexes, de nombreux modèles ont été proposés dans la littérature, notamment les « régressions aléatoires » dans le contexte de l'évaluation génétique.

J'ai travaillé un peu sur ces aspects. D'abord, pour les contrôles laitiers élémentaires la modélisation la plus simple consiste à supposer qu'il s'agit, soit de mesures répétées du même caractère, soit de caractères corrélés. Pour les races ovines Latxa et Manchega, Malena Serrano (INIA, Madrid) a estimé les différentes composantes de variance [A2], en concluant

qu'il s'agit bien des caractères différents. Ma participation à cette étude a été la création des fichiers de base, ainsi qu'une aide à l'interprétation des résultats.

Plus tard, lors de mon post-doc avec Keith Bertrand et Ignacy Misztal (University of Georgia, Athens, Etats-Unis), j'ai travaillé sur la modélisation de courbes de croissance pour l'évaluation génétique des bovins allaitants. Effectivement, cette population dispose de pesées à la naissance, au sevrage, au bout d'un an (« birth, weaning, yearling »), théoriquement aux jours 1, 210 et 365 ; en pratique, les vraies dates se distribuent autour de celles-ci, ce qui induit la nécessité de faire des précorrections.

Un modèle pratique pour le traitement et l'évaluation génétique de ce type de données est un modèle polynomial, dans lequel la croissance est une fonction polynomial du temps (t) et de certains effets ($a_0 \ldots a_n$), à l'occasion supposés contrôlés génétiquement et donc aléatoires (d'où le nom de régression aléatoire). Le modèle pour le poids au temps t peut s'écrire ainsi :

$$y(t) = \ldots + a_0 + a_1 k_1 t + a_2 k_2 t^2 + \ldots a_n k_n t^n$$

Où les a sont des effets génétiques propres à l'individu, et les k sont des coefficients qui garantissent certains propriétés numériques optimales, (convergence, orthogonalité, etc.). Tout le problème réside dans la spécification de composants de variance pour les différents effets a, car il sont forcement corrélés. C'est-à-dire, qu'il faut trouver la matrice de covariance :

$$Var\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} G_{1,1} & G_{1,1} & \ldots & G_{1,n} \\ G_{1,2} & G_{2,2} & & \\ \vdots & & \ddots & \\ G_{n,1} & & & G_{n,n} \end{pmatrix} = \mathbf{G}_0$$

Classiquement, on estime les parts de la variance avec un REML ou échantillonnage de Gibbs (comme dans les chapitres précédents). Cependant, les fichiers de données (race Gelbvieh américaine à l'occasion) ne permettaient pas une telle estimation, car trop grands et en même temps trop incomplets (manque de poids intermédiaires).

Nous avons donc suivi une idée d'Ignacy Misztal. Il existent des équivalences entre le modèle à régression aléatoire et le modèle à caractères corrélés, plus classique. Il existe aussi des estimations de $\mathbf{G}_0$ pour d'autres races (dans notre cas, ce fut Nellore) ; ces estimations définissent des corrélations entre deux points quelconques dans le temps. Nous avons supposé des motifs de corrélations similaires pour des races différentes, et nous avons donc réintroduit les parts de variance telles que, pour des corrélations identiques, on retrouve à peu près les paramètres du modèle multicaractère Gelbvieh. Le problème était encore plus compliqué car il y avait des effets génétiques direct et maternel, et des effets aléatoires maternel et individuel. Le meilleur ajustement fut obtenu en faisant des moindres carrés. J'ai développé la méthode à partir de l'idée originale d'Ignacy et fait les calculs et programmations nécessaires, ainsi que des outils pour visualiser les résultats (voir exemple ci-dessous). Finalement, on a construit de matrices de covariances qui ajustaient correctement pour des polynômes de degré 2 à 5. Le tout a été publié [A3].

Ce travail implique de reconstruire des matrices pour ensuite faire des régressions linéaires, c'est à dire beaucoup d'algèbre matricielle, ce qui m'a bien servi pour après. La visualisation des différentes structures de covariances restait très importante, et j'ai développé des outils adaptés [B7], mais qui ne sont plus d'actualité.

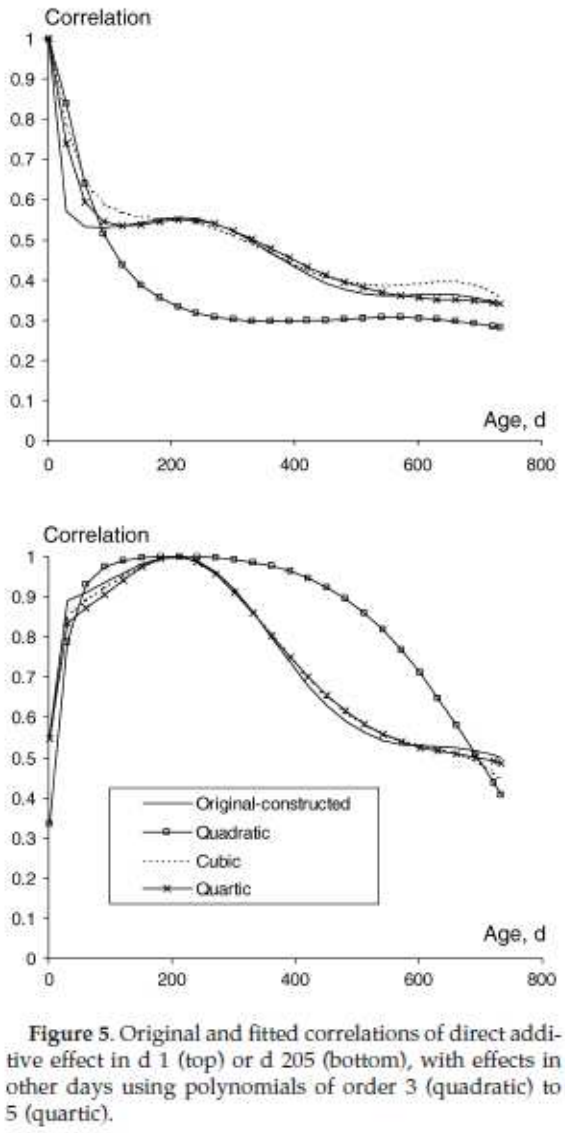


Figure 5. Original and fitted correlations of direct additive effect in d 1 (top) or d 205 (bottom), with effects in other days using polynomials of order 3 (quadratic) to 5 (quartic).

Extrait 9 Corrélations entre les effets génétiques à différentes dates [A3]

3.6. Conclusion

Je me rends compte que, dans la modélisation de tous ces caractères, il est plus difficile de bien saisir le fond biologique du caractère (un bon exemple est le modèle produit) que de l'implémenter, sauf si on travaille avec des fichiers de grande taille. En effet, les méthodes d'estimation modernes (la plus explicite est la Bayésienne avec échantillonnage de Gibbs, mais ce n'est pas la seule) permettent d'emboiter de manière naturelle une hiérarchie de paramètres, de telle manière qu'une inférence correcte passe par une étape relativement facile de « data augmentation », après laquelle on tombe sur des algorithmes classiques. D'un autre coté, la manipulation des données longitudinales m'a permis de me rendre compte des subtilités des distributions multivariées, même dans le cas normale.

Une autre conclusion est l'importance de la praticité : tous ces modèles sont conceptuellement très élégants, mais ils sont rarement utilisés en évaluation génétique routinière (les exceptions sont l'évaluation des chevaux de course, les régressions aléatoires, et le modèle à seuil),

souvent par manque d’outils ou peu d’intérêt car on a peu à gagner. Ils aident surtout à nôtre compréhension du caractère d’intérêt et, de fois, sont appliqués plus tard.

4. La localisation de QTL

Lors de mes études d'ingénieur agronome, j'ai appris que les gènes étaient des séquences d'ADN, qu'ils codifiaient des protéines, et que les changements de bases de l'ADN codant causaient un changement de la protéine, qui à son tour modifiait le phénotype de l'individu. Les gènes étaient donc des entités localisées dans le génome. L'architecture génétique admis aujourd'hui est un peu plus nuancée, car on sait qu'il y a des régions mobiles (les éléments transposables) et que les régions régulatrices de l'expression sont nombreuses, donc il n'y a pas que les régions codantes qui causent la variabilité génétique. De plus, il est admis (après un grande nombre d'études) que l'architecture génétique d'un caractère implique une myriade de gènes à petits effets en interaction, ce qui rend difficile leur identification.

De toute manière, la détection et la localisation de QTL (Quantitative Trait loci) a été et reste un des travaux majeurs de recherche en génétique, et il est une des sources principales d'inspiration pour la sélection génomique. Le principe est simple : faire l'hypothèse qu'une certaine région est porteuse d'un QTL, et vérifier si l'identité (au sens large) de la région décrit la similarité des phénotypes des individus.

4.1. Modèles linéaires de liaison et désequilibre de liaison pour la localisation de QTLs.

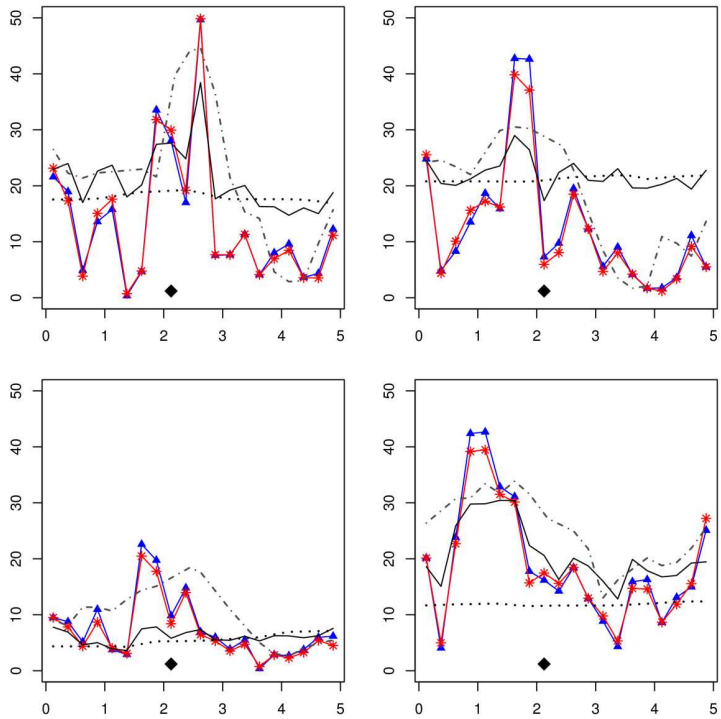
4.1.1. Modèles linéaires

Quand je suis arrivé à l'INRA (fin 2005), les méthodes de localisation de QTLs evoluaient de l'utilisation de la liaison (marqueurs peu denses de type microsatellite) à l'utilisation du déséquilibre de liaison « populationnel » (marqueurs denses de type SNP). Ce changement était dû à des raisons pratiques –disponibilité des marqueurs SNP plus denses–, et scientifiques –les analyses de liaison n'étaient pas suffisamment précises pour localiser les gènes d'intérêt. Même si de nombreuses méthodes existent pour la localisation, dans la pratique les chercheurs en génétique animale utilisent beaucoup des méthodes linéaires (Knott, 2005) : la régression linéaire (pour des données bien structurées en familles) et l'analyse en modèle mixte (pour des pedigrees complexes).

En pratique, les méthodes de déséquilibre de liaison postulent une association parfaite entre l'allèle au marqueur observé et l'allèle au QTL postulé. Les méthodes de liaison, elles, postulent que si un morceau de chromosome est transmis intégralement il est porteur du même allèle au QTL ; si le gamète transmis a été recombinant, il peut être porteur d'un allèle avec probabilité *p*, ou bien de l'autre allèle avec probabilité *1-p ;* mais rien n'est dit des allèles chez les fondateurs.

Je trouvais que les deux concepts devaient se réconcilier et conduire à une formulation simple et rigoureuse, basée sur la notion d'identité par état chez les fondateurs et par descendance dans le reste de la généalogie. Cette inquiétude était partagé par Rohan Fernando (Iowa State University, Ames, Etats-Unis), que j'avais contacté. Je suis donc allé travailler avec lui pendant trois semaines, pendant lesquelles nous avons développé les méthodes et l'algèbre en deux versions : régression et modèle mixte [A17], ainsi que fait des simulations. Cependant, au-delà de l'intérêt académique, nos simulations montraient qu'une analyse simple en

déséquilibre de liaison « pure » donnait la même précision que nos modèles plus sophistiqués, comme montré dans la figure ci-dessous. La raison est biologique : quand les marqueurs sont assez proches pour que l'identité par état ressemble à l'identité au QTL, les recombinaisons sont tellement peu fréquentes que finalement cette ressemblance marqueurs/QTL reste valable au cours des générations.



Extrait 10 Profils de détection de QTL [A17]

Interval mapping profiles in the drift and 5-cM scenario. Interval mapping profiles (minus log of the p-value) in four replicates of the drift and 5-cM scenario. LA: dotted line. LDLA: continuous line. LD decay: red, stars. Two-marker: blue, triangles. IBD: grey dot-dash line. A diamond indicates the QTL location.

Ceci dit, le modèle que nous avons publié a été introduit dans le logiciel QTLMAP (http://dga7.jouy.inra.fr/qtlmap/) et utilisé lors des nombreuses analyses de données réelles et simulées [A39, A41]. Ce dernier article fait parti des travaux de thèse de G. Sallé (encadré par Jean-Michel Elsen et Carole Moreno), qui a localisé des QTL qui agissent sur la résistance à *Haemonchus Contortus* (un parasite gastro-intestinal) chez les ovins. Cet article a montré une notable concordance entre les approches « analyse de liaison », ou nôtre approche [A17] dans deux version différentes selon que l'on prend en compte l'origine raciale ou pas, ou encore une analyse simple de type régression uni-marqueur.

4.1.2. Modèles paramétriques

Pour la réconciliation de la liaison et l'association, une formulation complètement paramétrique postule un événement qui donne lieu à l'allèle mutant au QTL, pour ensuite modéliser le déséquilibre de liaison autour du QTL chez les fondateurs, et enfin observer la transmission aux descendants (Pérez-Enciso, 2003). Non seulement cette formulation fait beaucoup d'hypothèses (et certaines approximations), mais en plus elle n'est pas plus précise,

même quand les hypothèses sont vraies ; nous avons montré ceci par simulation [A21], dans un travail en équipe, coordonné par C. Cierco et B. Mangin (MIA, INRA Toulouse).

4.1.3. Comparaison des modèles linéaires de régression et des modèles mixtes avec approximations de la coalescence

Toujours dans la localisation de QTL, Meuwissen et al. (2002) ont proposé un modèle (que l'on nommera LDLA par la suite) qui intègre liaison et déséquilibre de liaison dans le même esprit que les sections précédentes. Ce modèle utilise une théorie de coalescence (très) simplifiée pour calculer des probabilités d'identité par descendance chez les fondateurs d'une généalogie, conditionnellement aux marqueurs moléculaires ; ceci est différent de notre modèle [A17] qui postule que deux segments identiques par état le sont aussi au QTL. Le modèle de Meuwissen et al. (2002) était assez utilisé en pratique.

La thèse de Dana Roldán (encadré par Jean-Michel Elsen) visait, entre autres objectifs, à établir des protocoles optimaux pour la localisation de QTL. Dans [A39], elle a comparé notre approche dans [A17] (dans sa version régression simple, qui sera nommé RS par la suite) avec le LDLA et ceci, pour différents protocoles expérimentaux. La comparaison (article [A39]) a été faite par des simulations exhaustives et j'y ai travaillé, surtout pour la définition des différents scenarii et l'interprétation et vérification des résultats. Dana a trouvé que, généralement, le LDLA était très légèrement plus précis que le RS mais à un coût calculatoire beaucoup plus élevé : ainsi le RS est préférable pour la planification expérimentale. De même, elle a trouvé que, lorsque l'on travaille sur des dispositifs familiaux, un grand nombre de demi-frères est préférable à la même quantité de plein-frères.

4.2. Phasage des marqueurs dans une généalogie

Dans le génome d'un organisme diploïde, les deux ensembles d'allèles d'origine paternelle / maternelle sont arrivées au zygote de manière séparée, et ceux qui sont dans la même paire chromosomique sont physiquement présents dans le même chromosome. Cette notion de même origine est appelée phase. Le « phasage » ou reconstruction d'haplotypes est le processus estimant quels ensembles d'allèles ont la même origine (maternelle ou paternelle). Ce processus est de grande importance pour la localisation de QTL ou l'imputation de marqueurs manquants (en plein essor à ce jour-là).

La complexité du problème est multiple. D'abord, il y a deux sources d'information : l'apparenté connu (deux individus reliés par parenté partageront des bouts de chromosome) et l'apparenté non connu, ou ancestral, ou déséquilibre de liaison : une population non infinie est complètement apparentée, et l'on observe des motifs répétés dans les différents gamètes. Ces motifs viennent d'un même ancêtre commun. Ainsi, il existent classiquement deux familles d'algorithmes/méthodes/approximations pour « phaser » les marqueurs :

- L'un est la recherche d'un maximum de vraisemblance au sein d'une généalogie, c'est-à-dire, la configuration de méioses qui explique le plus simplement les génotypes observés. Ceci peut se traduire dans un ensemble de règles simples, en particulier pour les grandes fratries propres à la génétique animale (Wijsman, 1987). Cependant, l'optimalité globale n'est pas garantie. En plus, ces méthodes font une hypothèse de parcimonie chez les fondateurs de la généalogie: ces fondateurs sont idéalement tirés au hasard dans une population de taille infinie, ils sont non apparentés, et donc on suppose que les marqueurs aux différents loci sont en équilibre de liaison.

- L'alternative fait implicitement l'hypothèse d'une population « peu apparentée » mais de taille finie, et on utilise des méthodes semi-paramétriques qui (conceptuellement) estiment un ensemble d'haplotypes « fondateurs » à partir desquels les génotypes observés sont issus (e.g. Browning & Browning, 2011).

La thèse d'Aurélie Favier, co-encadré par Simon de Givry (MIA, INRA Toulouse) et moi, visait à utiliser des méthodes et des formalismes propres à la recherche en informatique pour attaquer ces deux problèmes. Il faut dire que la tâche était trop ambitieuse (et c'est un champ de la génétique en plein explosion) et nous nous sommes focalisés sur la première famille de méthodes, celle qui prend en compte les liens de parenté et estime l'ensemble de méioses ayant la plus grande vraisemblance.

4.2.1. Phasages dans les familles de demi-frères

L'écriture de cette vraisemblance n'est pas facile : il s'agit d'une loi discrète multivariée, avec une quantité d'inconnues égale (au sein d'une famille nucléaire) à *la taille de la fratrie x le nombre de marqueurs*, et assez lourde à calculer. Une maximisation naïve (par exemple, par dénombrement de tous les cas) n'est pas envisageable. Cependant, une collaboration avec Jean-Michel Elsen (SAGA ; qui travaillait en parallèle sur le même problème) a permis, au prix de certaines simplifications –et pour des marqueurs biallélique–, une écriture compacte de la vraisemblance, qui est détaillé dans [B31, B32]. La phase des parents est définie par un vecteur **h** ayant des valeurs possibles {-1,1} à chaque position. L'écriture de la vraisemblance revient à compter, au sein de chaque fratrie, la quantité de frères ayant hérité avec certitude des couples en « phase » ou en « opposition » (à partir d'un état initial arbitraire) pour tout couple de marqueurs, pondéré par leurs probabilités de recombinaison. Ainsi, on construit une fonction de vraisemblance qui est :

$$\log p\left(genotypes \mid \mathbf{h}\right) \propto \mathbf{h}'\mathbf{W}\mathbf{h}$$

Équation 6

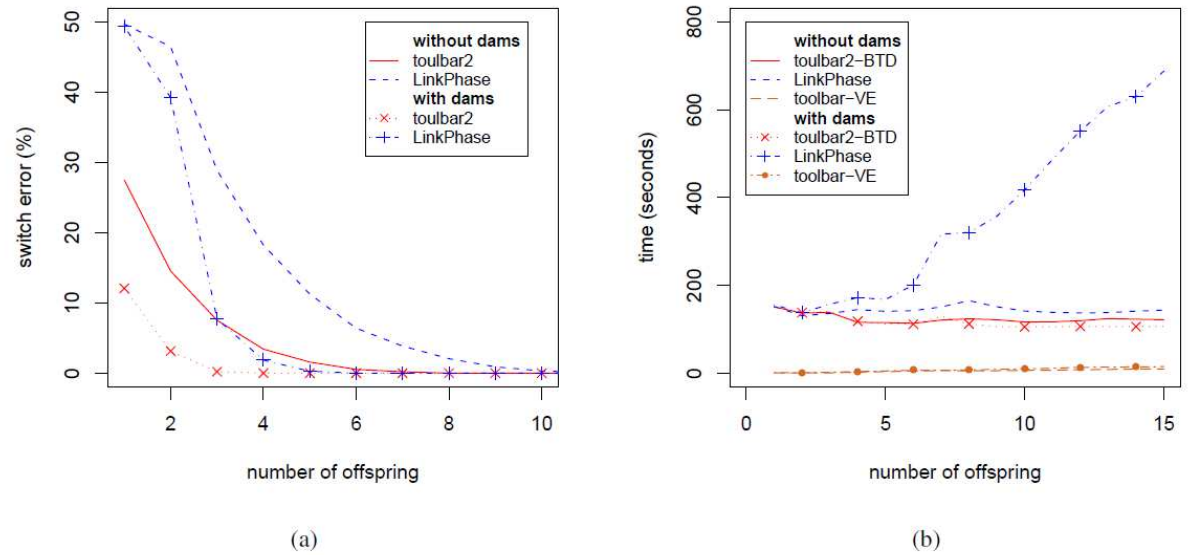
où **W** est une matrice issue du comptage ci-dessous qui est typiquement très creuse. Maximiser cette fonction n'est pas évident car le domaine de **h** est {-1,1} (et non le domaine des réels). C'est ici que l'on se sert des techniques informatiques d'optimisation combinatoire.

Une branche de l'informatique qui traite de ce problème s'appelle les problèmes de satisfaction pondérées (« WCSP : weighted constrained satisfaction problems ») qui, à leur tour, sont formellement équivalents (à certaines transformations près) à des modèles probabilistes graphiques bayésiens, les réseaux bayésiens (« Bayesian networks »). Cette transformation implique convertir les lois de probabilité en fonctions de coût : $f\left(x_i; parent\left(x_i\right)\right) = -c \log \Pr\left(x_i \middle| parent\left(x_i\right)\right)$.

Aurélie a transformé le problème précédant (qui peut être vu comme un réseau bayésien) dans un problème WCSP et, avec la boîte à outils des WCSP, Aurélie a rapidement codé et testé extensivement par simulation le problème du calcul de phase et cela, avec plusieurs méthodes d'optimisation (DFBB,BTD,VE) du solver `toulbar2` [B31, B32, A45, A46]. Il faut remarquer que cette méthode est exacte (elle trouve l'ensemble des phases les plus probables), tandis que la plupart des méthodes itératives existantes peuvent trouver des maxima locaux.

Ses résultats principaux sont les suivants :

- Les résultats obtenus sont égaux ou meilleurs que ceux obtenus par d'autres méthodes
- L'utilisation de l'équation 6 rend le problème très simple tant en espace mémoire qu'en temps de calcul. La matrice **W** a des propriétés grossièrement prévisibles (le nombre d'éléments non nuls grandit linéairement avec le nombre de marqueurs).
- Des tests en situation réaliste (déséquilibre de liaison chez les fondateurs) donnent aussi d'excellents résultats, comme montré dans la figure ci-dessous (36000 SNP du chromosome X humain)



(a) (b)

Figure 4: Human chr. X dataset w/wout genotyped dams ($L = 36000, n \in [1, 15]$).

Extrait 11 Qualité de la reconstruction d'haplotypes [B32]

Ainsi, cette méthode a été implémenté dans le logiciel QTLMAP (utilisé par exemple dans [A39, A41]). Nous avons aussi proposé une mesure [Thèse A. Favier] pour informer l'utilisateur de la qualité d'un phasage. Cette mesure est une approximation de la probabilité *a posteriori* qu'un marqueur *k* ait une certaine phase. L'approximation vient du fait que l'on suppose le reste des marqueurs correctement phasés ; de cette manière le calcul est abordable, de la manière suivante :

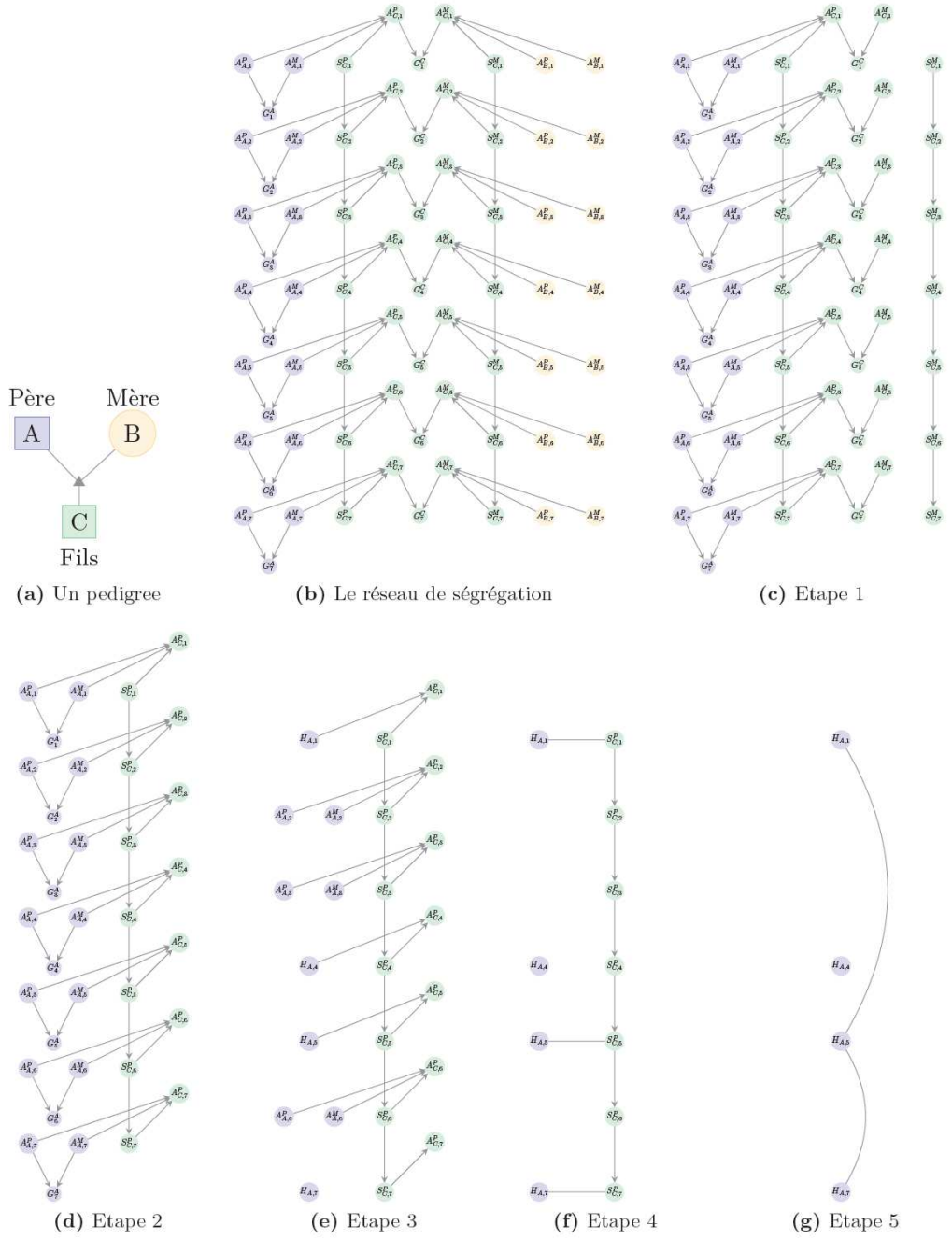
$$p\left(h_k = 1 \middle| genotypes\right) \approx p\left(h_k = 1 \middle| genotypes, \mathbf{h}_{-k}\right) = \frac{\exp\left\{\begin{pmatrix}\mathbf{h}_{-k}\\1\end{pmatrix}'\begin{pmatrix}\mathbf{W}_{-k,-k} & \mathbf{w}_{-k,k}\\ \mathbf{w}_{k,-k} & w_{k,k}\end{pmatrix}\begin{pmatrix}\mathbf{h}_{-k}\\1\end{pmatrix}\right\}}{\exp\left\{\begin{pmatrix}\mathbf{h}_{-k}\\1\end{pmatrix}'\begin{pmatrix}\mathbf{W}_{-k,-k} & \mathbf{w}_{-k,k}\\ \mathbf{w}_{k,-k} & w_{k,k}\end{pmatrix}\begin{pmatrix}\mathbf{h}_{-k}\\1\end{pmatrix}+\begin{pmatrix}\mathbf{h}_{-k}\\-1\end{pmatrix}'\begin{pmatrix}\mathbf{W}_{-k,-k} & \mathbf{w}_{-k,k}\\ \mathbf{w}_{k,-k} & w_{k,k}\end{pmatrix}\begin{pmatrix}\mathbf{h}_{-k}\\-1\end{pmatrix}\right\}}$$

Où « *-k* » indique l'ensemble de loci autres que *k*. Cette estimateur est non biaisé, c'est à dire, ni sous-estime ni surestime l'erreur.

4.2.2. Décompositions fonctionnelles et exactes : vers les phasages dans des familles complexes

Malgré les bons résultats de l'approche précédente, cette écriture simplifié ne pouvait se faire que pour des familles de demi-frères. La généralisation à des généalogies complexes, de grande importance en génétique animale, n'était pas envisageable en l'état.

Aurélie a refait, avec ses outils et formalismes informatiques, la décomposition proposé par Jean-Michel (voir ci-dessous).



(a) Un pedigree (b) Le réseau de ségrégation (c) Etape 1

(d) Etape 2 (e) Etape 3 (f) Etape 4 (g) Etape 5

Figure 8.2 – Soit une famille nucléaire 8.2a. Supposons que le père et le fils soient génotypés sur 5 marqueurs tels que :
Père : $(ab\ bb\ aa\ ab\ ab\ aa\ ab)$. Fils : $(aa\ bb\ aa\ ab\ aa\ ab\ aa)$.
Le réseau d'origine de ségrégation est représenté en 8.2b.
Les réseaux 8.2c à 8.2g sont le résultat graphique des étapes de la réduction du calcul $\sum_{\mathbf{V} \setminus \mathcal{P}} \mathbb{P}(\mathbf{V}, \mathbf{G})$

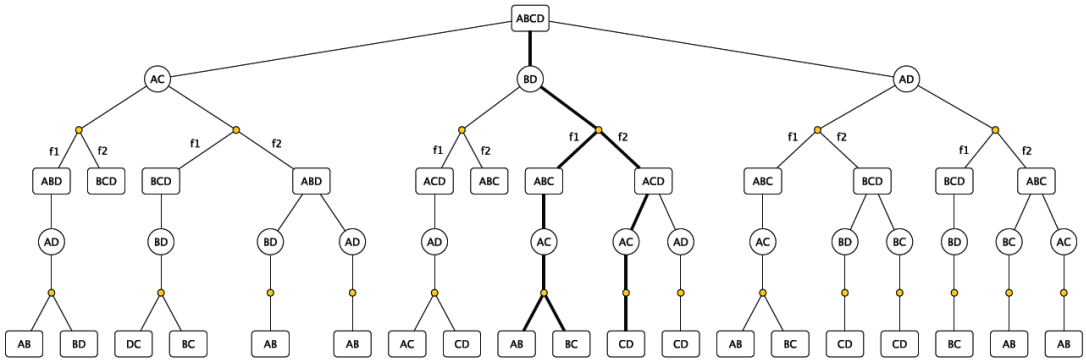
Extrait 12 Simplification de la vraisemblance pour le phasage de demi-frères [Thèse A Favier]

Cette décomposition était un cas particulier d'un type de décomposition dans le formalisme WCSP. Si l'on pouvait étendre cette décomposition à des généalogies quelconques, alors on aurait avancé d'un grand pas vers la résolution du problème des généalogies complexes, et en même temps en méthodes informatiques. Donc, Aurélie s'est plongée dans ce sujet et elle a

aboutie à des résultats très prometteurs [A47, A48, thèse A. Favier]. D'abord, elle a montré les connexions profondes entre les formalismes de WCSP et les réseaux bayésiens (je perçois personnellement les réseaux bayésiens comme étant un type de WCSP). Ensuite, elle a montré que, quand les réseaux peuvent se décomposer de manière efficace (on réduit un problème complexe à plusieurs problèmes simples : $f(A,B,C) = f(A,B) + f(B,C)$), savoir si la décomposition existe ou pas est vérifiable avec une procédure de test efficace. Ce travail a impliqué un algèbre sur les coûts infinis (=probabilités nulles). De plus, elle montre comment simplifier encore le problème en faisant des projections et des soustractions (voir exemple ci-dessous).

A	B	C	D	$f(A,B,C,D)$
0	0	0	0	0
0	0	0	1	0
0	0	1	0	5
0	0	1	1	4
0	1	0	0	6
0	1	0	1	6
0	1	1	0	4
0	1	1	1	3
1	0	0	0	0
1	0	0	1	0
1	0	1	0	5
1	0	1	1	4
1	1	0	0	9
1	1	0	1	9
1	1	1	0	7
1	1	1	1	6

(a) Une fonction de coûts



(b) Représentation de toutes les décompositions possibles de la fonction

Figure 4.3 – Trois types de nœuds sont à identifier dans la figure (b) : les nœuds rectangulaires pour identifier les fonctions, les nœuds ronds pour identifier par rapport à quelques variables la décomposition par paire est réalisée et enfin les nœuds (ronds) sans label pour symboliser la construction des deux fonctions résultantes de la décomposition. Le chemin en gras représente les décompositions par paire réalisées avec l'ordre DAC : $A < B < C < D$.

Extrait 13 Simplification d'un problème WCSP [Thèse A Favier]

$f_1 = g(A,B,D)$		D	
		0	1
A	B		
0	0	0	0
	1	4	3
1	0	0	0
	1	7	6

(a)

$f_2 = h(B,C,D)$		D	
		0	1
B	C		
0	0	0	0
	1	5	4
1	0	2	3
	1	0	0

(b)

Table 4.1 – Application du théorème 4.26 sur la fonction $f(A,B,C,D)$ de la figure 4.3a avec $f_1 = g(A,B,D)$ et $f_2 = h(B,C,D)$. (La fonction $h(B,C,D)$ n'est pas décomposable.)

$f_2 = g(A,B,D)$		D	
		0	1
A	B		
0	0	0	0
	1	0	0
1	0	0	0
	1	3	3

(a)

$f_1 = h(B,C,D)$		D	
		0	1
B	C		
0	0	0	0
	1	5	4
1	0	6	6
	1	4	3

(b)

Table 4.2 – Application du théorème 4.26 sur la fonction $f(A,B,C,D)$ de la figure 4.3a avec $f_1 = g(A,B,D)$ et $f_2 = h(B,C,D)$. (Les deux fonctions ternaires sont encore décomposables.)

Extrait 14 Simplification d'un problème WCSP [Thèse A Favier]

Deux problèmes existent pour l'applicabilité immédiate de la procédure. Le premier est l'absence d'un algorithme pour faire les décompositions de manière optimale ; le deuxième est le cas des problèmes non-décomposables. Dans ce dernier cas, Aurélie propose une décomposition approchée qui ignorerait certains coûts ; on peut regarder cette dernière idée comme une simplification probabiliste qui fait l'hypothèse que certains événements sont indépendants quand en fait ils ne le sont pas. Cette décomposition approchée n'a pas pu être approfondie, par manque de temps.

4.3. Conclusion

Je ne me suis pas trop impliqué dans la recherche de QTL en soi. Cependant, le développement des modèles de détection de QTL sert à affûter beaucoup des compétences, et aussi à réaliser que les méthodes les plus sophistiquées ne sont pas forcément les meilleures.
D'un autre coté, j'ai beaucoup appris en encadrant la thèse d'Aurélie. Scientifiquement, j'ai appris énormément de choses et j'en suis sorti plus modeste : d'autres disciplines ont des outils formidables. Humainement, la direction de thèse est une aventure quotidienne très riche, et qui conduit à se poser des vraies questions.

5. L'évaluation génétique (et génomique)

La sélection des animaux de rente se fait par le choix comme reproducteurs de ceux qui donneront la descendance plus profitable (ou utile dans un sens). Il est donc primordiale de bien choisir ces animaux, ce qui se fait aujourd'hui à partir d'une évaluation. Quand je suis arrivé à la recherche, l'évaluation génétique était très figée, voire immuable : on obtenait les phénotypes, l'information environnementale et la généalogie *via* le contrôle de performances et on utilisait un BLUP. Le BLUP et le modèle mixte sont aussi une vaste source d'idées pour les développements méthodologiques. Tout ceci a évolué dès que l'information génomique a commencé à être utilisée de manière massive, avec les puces à ADN.

5.1. Travaux en évaluation classique

5.1.1. Comparaison de modèles d'évaluation génétique Latxa

Le modèle d'amélioration génétique de la Latxa fut conçu par Dunixi Gabiña (à l'époque au CIMA, Vitoria, Espagne) dans les années 1980, en s'inspirant des modèles utilisés en France. Dans un souci d'éviter les biais, il avait construit un modèle avec *deux* groupes de contemporaines : « troupeau-année » et « troupeau-mois de mise bas-âge de la femelle ». Effectivement, les travaux d'Eduardo Urarte (CIMA) ont montré que la conduite des troupeaux différenciait la conduite des différentes classes d'âge-nombre de mise bas. Néanmoins, et après une douzaine d'années, on se rendait compte que ce modèle était compliqué à expliquer, à comprendre, et que l'hypothèse d'un effet « troupeau-mois de mise bas-âge de la femelle » constant au cours des années n'était plus tenable. En plus, certains groupes de contemporaines avaient une taille trop petite.

Dans ma thèse, j'ai entrepris la mise à jour du modèle d'évaluation. Il est très difficile de comparer des modèles d'évaluation. Boichard et al. (1995) ont proposé deux types de test qui consistaient à comparer les évaluations génétiques avec des sous-ensembles des données : le test qui considère les évaluations au fil des années, et celui qui considère la précision additionnelle apportée par les générations successives de filles. Les deux tests donnaient des résultats décevants en Latxa : en effet, une grand partie de la généalogie n'étant pas connue, les tests étaient une fonction du modèle choisi pour les groupes génétiques. Nous avons alors choisi une autre approche.

Nous sommes rentrés dans un cadre bayésien qui permet de comparer des modèles non emboités. J'ai travaillé avec Pedro López Romero (thésard à l'INIA à la même époque) ; nous avons utilisé son outil de comparaison de modèles, basé sur l' échantillonnage de Gibbs. Cet outil permettait de calculer certains critères d'évaluation, notamment le Facteur de Bayes, qui est un rapport entre les probabilités des modèles, et le « Deviance Information Criteria ». De plus, j'ai introduit dans son logiciel une forme de « leave-one-out cross-validation », c'est-à-dire de validation croisée « une donnée à chaque fois ». Cette forme bayésienne permettait de ne pas faire tourner les analyses des centaines de milliers de fois, et produisait des diagnostics de biais ou précision. Tout ce travail, qui fut publié dans [A4], montra la préférence pour un modèle « troupeau-année-âge-mise bas », plus simple que le précédent ; un exemple de

diagnostic est montré ci-dessous. Il faut mentionner que ce travail me prépara pour les tests empiriques d'efficacité de la sélection génomique que l'on verra plus tard.

Table 6
Estimates of the checking functions for different strains and models

Model	Strain and checking function							
	Blond-Faced Latxa				Black-Faced Latxa			
	$E(\|d_1\|)^{a,b}$	$E(d_1^2)$	$E(\|d_2\|)^c$	$E(d_2^2)$	$E(\|d_1\|)$	$E(d_1^2)$	$E(\|d_2\|)$	$E(d_2^2)$
1	25.86	1503.05	0.3389	0.1396	24.12	1272.95	0.3399	0.1402
2	25.52	1411.76	0.3412	0.1408	24.08	1241.52	0.3411	0.1408
3	26.00	1446.60	0.3454	0.1436	24.23	1243.50	0.3445	0.1429

[a] E=expectation.
[b] d_1=Observation−prediction.
[c] $d_2=p(d_1>0)-0.5$.

Extrait 15 Comparaison de modèles d'évaluation en Latxa [A4]

5.1.2. Evaluation génétique multi-raciale en bovin allaitant

Mon post-doctorat aux USA avait pour objectif de mettre à jour l'évaluation génétique de certaines races de bovins allaitants qui étaient évaluées par l'Université de Georgia, et dont le responsable scientifique était Keith Bernard. Les fermes (« ranches ») américaines ont une race prédominante, mais cela ne veut pas dire que la race se reproduit en pur. En pratique, on utilise des taureaux d'autres races au sein d'une race : par exemple un taureau Angus peut être utilisé pour augmenter la qualité de la viande d'un troupeau Gelbvieh. Cela ne veut pas dire pour autant que tout le troupeau deviendra Angus ; plusieurs mélanges coexistent en même temps. Il est donc important pour l'évaluation génétique d'une race (à l'occasion, la race Gelbvieh) de considérer :

- L'hétérosis générée lors d'un croisement de races
- L'utilisation des évaluations génétiques « externes » des taureaux d'autres races (comme l'Angus mentionné) puisque l'on utilise des taureaux évalués de manière précise dans leurs races respectives.

Le travail se complique vue la multiplicité des races utilisées aux Etats Unis ; voir par exemple le tableau de composition raciale ci-dessous. L'objectif était d'avoir une indexation plus précise mais aussi plus robuste.

Table 1 Breed composition of animals in pedigree and data files

Breed composition[a]	No. animals in pedigree file	No. animals in data file
8/8 Angus	32 150	97
8/8 Brahman	206	0
8/8 Charolais	715	1
8/8 Hereford	4058	0
8/8 Limousin	546	0
8/8 Simmental	1229	0
8/8 Unknown[b]	42 416	132
8/8 Gelbvieh	99 258	86 699
7/8 Gelbvieh	374 120	360 685
6/8 Gelbvieh	120 732	114 526
5/8 Gelbvieh	25 337	23 919
4/8 Gelbvieh	110 318	96 085
3/8 to 1/8 Gelbvieh	32 923	22 855
Other	22 617	85

[a]Animals with 8/8 breed composition are considered purebreds; Angus and Red Angus were grouped together and Hereford and Polled Hereford were grouped together.
[b]Animals with unknown breed designation were grouped into British Breed category.

Extrait 16 Composition raciale de l'évaluation multi-raciale Gelbvieh

Le croisement implique deux aspects. D'abord, il y a le niveau de base d'une race par rapport à l'autre, qui se modélisa avec des groupes de parents inconnus par race et date de naissance. Ces groupes génétiques furent reliés *a priori* par une structure d'auto-corrélation pour éviter de larges déviations.

Pour l'hétérosis (donc la dominance), il existe des modèles théoriques très complets (Lo et al., 1993). Ces modèles sont inapplicables ici car on n'a pas accès aux données des autres races. De plus, les animaux croisés sont relativement peu nombreux et les croisements non symétriques. Pour l'ajustement de l'hétérosis, ce fut donc une approche empirique, mais très répandue, qui considère un effet de la covariable « composition race A *x* composition race B ». Nous recueillions des informations *a priori* des effets d'hétérosis (disponibles lors des décades de protocoles de croisement) qui furent (et sont) l'information a priori pour les évaluations. Tout ce travail a été fait avec mes collègues de UGA, K. Bertrand, R. Sapp, I. Misztal, T. Strabel et JP Sanchez (qui a complété mon travail après mon départ).

Pour les informations externes, je fis une dérivation approchée qui permettait de les introduire de manière naturelle, en modifiant les équations du modèle mixte :

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{0} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{*-1} & \mathbf{G}^{*-1}\mathbf{Q}' \\ \mathbf{0} & \mathbf{Q}'\mathbf{G}^{*-1} & \mathbf{Q}'\mathbf{G}^{*-1}\mathbf{Q} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \\ \mathbf{g} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y}+\mathbf{G}^{*-1}\boldsymbol{\mu}_0 \\ \mathbf{Q}'\mathbf{G}^{*-1}\boldsymbol{\mu}_0 \end{bmatrix}$$

Avec $$\mathbf{G}^{*-1} = \begin{bmatrix} \mathbf{G}^{EE}+\mathbf{D}^{-1}-\mathbf{G}_{EE}^{-1} & \mathbf{G}^{EI} \\ \mathbf{G}^{IE} & \mathbf{G}^{II} \end{bmatrix}$$

C'est-à-dire, en incluant les informations externes dans le coté droit ($\boldsymbol{\mu}_0$) et en modifiant la matrice de parenté en incluant les précisions externes (matrice **D**). Cette méthode ressemble énormément au Single Step que l'on verra plus tard.

J'ai programmé un logiciel d'itérations sur des données qui permettait de faire les évaluations génétiques avec ce modèle pour des fichiers de grande taille (~800 000 animaux, effets directs et maternels et 3 caractères, donc 5 000 000 inconnues). La description de la procédure et des résultats se trouvent dans [A10].

5.2. Evaluation génomique

5.2.1. Evaluation empirique de la précision de l'évaluation génomique en souris

En 2006, l'évaluation génomique qui s'approchait était une complète inconnue. Il faut préciser que, si la dérivation du BLUP a précédé son utilisation massive, dans le cas de l'évaluation génomique l'application et le développement théorique se déroulent toujours en parallèle. De plus, la faisabilité de la sélection génomique était controversée. Eduardo Manfredi (SAGA) a eu l'idée de faire un test « grandeur nature » de la sélection génomique en lapin ou souris; par manque de financement, ce projet n'a jamais vu le jour, mais il nous a inspiré pour la suite.

Effectivement, les simulations étaient insatisfaisantes et les données en bovin laitier tardaient encore deux ou trois ans à venir. Nous avons eu l'idée (avec Eduardo, Jean Michel Elsen et Christèle Robert-Granié) d'analyser un jeu de données publiques de souris de laboratoire, disponible à http://gscan.well.ox.ac.uk/, qui était destiné à la détection de QTLs. Et, puisque un test de l'efficacité de la sélection n'était pas possible, nous avons analysé la qualité de l' évaluation génomique par validation croisée –comme dans les approches ci-dessus. Cette idée devint le standard dans la comparaison de modèles génomiques.

L'analyse du jeu de données souris [A12] fut très enrichissant. Nous avons utilisé des modèles linéaires mixtes, i.e. BLUP, mais avec des estimations des parts de variance associé (à

l'époque le résultat décrit dans Gianola et al. (2009) qui met en relation variance génétique et variance de l'effet SNP n'était pas connu).

D'abord nous avons montré que la performance de l'évaluation génomique dépendait du caractère : le gain sur l'évaluation généalogique a été nul pour le poids adulte, mais correct pour la taille, la vitesse de croissance ou encore l'index corporel. Ensuite, et par différents schémas de validation croisée, nous avons montré que la performance des évaluations (génomique ou généalogique) dépendait fortement des liens de parenté. C'est à dire, le phénotype (et donc la valeur génétique) d'un individu était prédit plus précisément si l'on utilisait des apparentés proches pour la prédiction, comme montré dans le tableau ci-dessous :

TABLE 2

Predictive ability of different models for genomic selection

	Model		
Trait	1	2	3
	Across families[a]		
Weight	0.07	0.25	0.20
Growth slope	0.04	0.26	0.19
Body length	0.05	0.16	0.12
Body mass index	0.06	0.17	0.12
	Within families[b]		
Weight	0.67	0.67	0.63
Growth slope	0.54	0.55	0.51
Body length	0.24	0.27	0.25
Body mass index	0.32	0.35	0.33

[a] Standard errors ~0.03.
[b] Standard errors ~0.02.

Extrait 17 Efficacité de la sélection génomique chez la souris

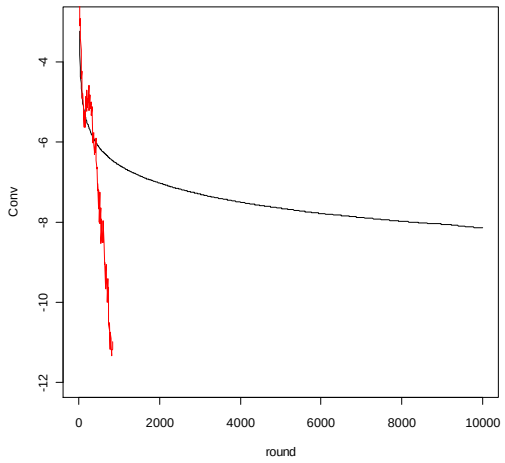
Ce résultat, vérifié depuis par de nombreuses études, remettait en cause les promesses de la sélection génomique : la précision des évaluations, et donc les plans de sélection, dépendaient toujours de la distribution des individus dans des familles. Or, la promesse initiale était que tous les individus génotypés, même les plus exotiques, pouvaient être évalués avec la même précision. La cause ? D'abord, seuls les SNPs ne suffisent pas à identifier un ensemble de QTL ayant d'effets transposables d'une famille à l'autre. Ce qui veut (peut-être) dire que les effets des gènes sont locaux et interagissent avec le fond génétique, qui lui est très similaire entre apparentés. De plus, ce résultat a servi à ce que les gens comprennent que les marqueurs ainsi que le déséquilibre de liaison sont des indicateurs de la parenté (Goddard, 2009).

Ce même jeu de données souris fut utilisé par Gustavo de los Campos [A13] pour réaliser la première application du Lasso Bayésien à la sélection génomique.

5.2.2. Algorithme d'estimation des effets SNP

Pour le travail précédant, allouer assez de mémoire pour un BLUP (un « BLUP_SNP » en fait) avec 10,000 inconnues (SNP) et une matrice d'incidence dense n'était pas facile : cela demanda un algorithme de résolution original, que j'ai conçu à partir des idées des algorithmes d'itération sur les données (Misztal & Gianola, 1987). L'algorithme consista à écrire les mises à jour (« update ») d'une inconnue dans l'algorithme de Gauss Seidel, non pas à partir des équations du modèle mixte, mais à partir des données (**y**) corrigées « par le reste des inconnues ». Ceci permet d'itérer les méthodes avec seulement la matrice d'incidence des effets **Z** et non pas leur produit croisé **Z'Z**. Cet algorithme n'est pas le seul à fonctionner ainsi: la méthode des gradients preconditionnés conjugués (PCG) le permet aussi. A l'aide d'Ignacy Misztal, j'ai fini de peaufiner cet algorithme et je l'ai comparé avec d'autres, y compris PCG ; nous l'avons baptisé GSRU (« Gauss Seidel with Residual Update »), e voici un pseudo-code et un aperçu de sa performance:

```
Double precision::
xpx(neq),y(ndata),e(ndata),X(ndata,neq), &
sol(neq),lambda,lhs,rhs,val
do i=1,neq
       xpx(i)=dot_product(X(:,i),X(:,i))
!form diagonal of X?X
enddo
lambda=vare/vara
e=y
do until convergence
       do i=1,neq
              !form lhs
              lhs=xpx(i)+lambda
              ! form rhs with y corrected by
other effects (formula 1)
              rhs=dot_product(X(:,i),e) &
                  +xpx(i) *sol(i)
              ! do Gauss Seidel
              val=rhs/lhs
              ! MCMC sample solution from
its conditional (commented out here)
              ! val=normal(rhs/lhs,1d0/lhs)
              ! update e with current
estimate (formula 2)
              e=e - X(:,i)*(val-sol(i))
              !update sol
              sol(i)=val
       enddo
enddo
```



Convergence de PCG (rouge) GSRU (noir)

Extrait 18 Pseudocode et performance du GSRU [A11]

Cet algorithme (qui a été découvert en parallèle et indépendamment par G. de los Campos, R. Fernando, et sans doute d'autres) a connu un certain succès car, bien que moins efficace que le PCG, il permet, avec des modifications très mineures, d'implémenter de manière efficace des méthodes Bayésiens comme le Lasso Bayésien, le BayesB, etc. Le code que j'ai développé a evolué pour devenir GS3 –dont je parlerai plus tard.

5.2.3. Tests d'évaluation génomique – autres espèces

En 2009 nous avons commencé un projet (Amasgen ; coordonné par Vicent Ducrocq, GABI, INRA Jouy-en-Josas) qui visait à comparer des méthodes pour évaluation génomique des bovins laitiers ; en parallèle, un autre projet (Roquefort'In) en ovins laitiers de race Lacaune (coordonné par Francis Barillet, SAGA, INRA Toulouse), et un troisième (Genomia) en race Manech et Latxa (coordonné par moi-même) a démarré en 2010. Je me suis fortement impliqué dans ces trois projets.

Tous ces projets ont testé les méthodes d'évaluation génomique. En bovin laitier, nous avons testé plusieurs méthodes (BLUP_QTL, PLS/Sparse PLS, BayesCPi, Lasso Bayésien, Elastic Net, GBLUP), sans que l'une soit nettement meilleure que l'autre, sauf pour les caractères ayant un gène majeur (taux butyreux, taux protéique) où les méthodes non linéaires telles que l'Elastic Net ou le BayesCPi donnent de meilleurs résultats [A23, A32, A36, A43]. Ceci dit, n'importe quelle méthode génomique est plus précise qu'une évaluation classique. Voici un des ces tableaux de résultats comme illustration:

Table 4. Correlations between observed daughter yield deviations (DYD) and predicted DYD provided by partial least squares regression (PLS), sparse PLS (sPLS), pedigree-based BLUP (BLUP), and genomic BLUP (GBLUP)

Item	Milk yield	Fat yield	Protein yield	Fat content	Protein content	Conception rate
BLUP	0.38	0.40	0.44	0.44	0.47	0.28
PLS	0.52	0.58	0.55	0.71	0.71	0.25
sPLS	0.50	0.59	0.53	0.72	0.72	0.21
GBLUP	0.56	0.59	0.55	0.72	0.73	0.35

Extrait 19 Comparaison des précisions de quelques méthodes d'évaluation génomique [A36]

Pour le projet Genomia, les données sont en cours d'analyse. Quand aux données de Lacaune, si bien une évaluation « expérimentale » est en cours [B52], Sandrine Duchemin a déjà comparé trois méthodes : GBLUP, PLS et BayesCPi [A37]; aucune de ces trois méthodes n'était supérieure à l'autre, ce qui suggère la non existence de gènes majeurs dans cette population.

Chez la poule, j'ai (avec Fanny Calenge et Catherine Beaumont, URA, INRA Tours) testé la performance d'une puce SNP à faible densité (1500 SNPs) pour évaluer une sélection génomique pour la résistance au portage de *Salmonella enterica*. Puisque le jeu de données était petit, nous avons entamé (1) l'estimation des parts de variance associées, soit aux SNPs, soit à la généalogie et (2) calculé la précision théorique à partir des modèles avec ou sans informations génomiques. En fait, la variance génétique était divisé pour moitié entre les deux parentés, et la précision n'augmentait pas ; il fut conclu que le jeu de données n'était pas suffisant ni en taille (<500 individus) ni en densité (<1600 marqueurs) [A22].

5.2.4. Paramétrisations et qualité du Lasso Bayésien

Lors du projet Amasgen, j'ai pris le Lasso Bayésien qui avait été montré par Gustavo et je l'ai testé avec nos jeux des données. Or, la paramétrisation du Lasso Bayésien ne me semblait pas intuitive : la distribution *a priori* des effets des SNPs dépendait de la variance résiduelle (ce qui pour moi n'a pas trop de sens biologique). De plus, Gianola et al. (2009) avaient montré la manière de « passer » conceptuellement de la distribution des effets des SNPs à la distribution

des valeurs génétiques, c'est-à-dire à la variance génétique, *via* la formule $Variance\,génétique \approx 2\sum p_i q_i Var(a)$ où *Var(a)* est la variance des effets des SNPs. Cela conduisait à ce que la variance résiduelle rentre dans la variance génétique.

Nous avons d'abord montré qu'une reparamétrisation simple du Lasso Bayésien (qui en fait reprenait le Lasso original de Tibshirani, 1996) éliminait cette nuance et, de plus, avec des précisions accrues et des estimations de la variance génétique semblables à celles obtenues avec des modèles classiques. Quant à la prédiction des valeurs génétiques, le Lasso Bayésien était au moins aussi efficace que d'autres méthodes comme le GBLUP mais, pour certains caractères contrôlés par des gènes majeurs, il fut capable d'incorporer cette information avec une augmentation de la précision, comme montré dans le tableau ci-dessous.

Table 4. *Accuracies: correlations between GEBVs and 2DYDs in the validation data set, in Holstein*

Trait	BL1Var	BL2Var	GBLUP	MCMC-GBLUP	HetVar-GBLUP
MY	0·28	0·41	0·42	0·40	0·41
FY	0·35	0·37	0·34	0·37	0·36
PY	0·27	0·30	0·31	0·30	0·30
FP	0·53	0·73	0·59	0·61	0·71
PP	0·36	0·48	0·44	0·46	0·47

MY, milk yield; FY, fat yield; PY, protein yield; FP, fat percentage; PP, protein percentage.

Extrait 20 Performance du Lasso Bayésien [A23]

Nous proposions aussi un GBLUP dont le poids relatif de chaque marqueur est extrait des résultats du Lasso Bayésien ; ce HetVar-GBLUP avait de très bonnes précisions. Tout a été publié [A23].

5.2.5. Relation entre les parentés génomiques et généalogiques

VanRaden (2008) a été le premier à réaliser la connexion entre la parenté généalogique et la parenté génomique. C'est-à-dire, il montra que les marqueurs permettaient de construire une matrice de parenté dite « génomique », **G** ($\mathbf{G} = \mathbf{ZDZ}'$, où **Z** sont les génotypes aux SNPs et **D** une matrice diagonale avec hétérozygoties). Cette matrice montrait une dualité : d'une part, elle permettait d'écrire d'une manière compacte les modèles mixtes pour l'effet du SNP (du « BLUP_SNP » vers le « GBLUP ») avec les mêmes résultats (au sens strict : les deux modèles sont équivalents), d'autre part, la matrice résultante était un estimateur de la parenté réalisée. La parenté généalogique fait l'hypothèse d'un nombre infini de gènes non liés. La liaison et la taille physique du génome produisent des déviations de la parenté moyenne, qui peuvent être observés dans **G**.

Or, ce résultat, plus montré que démontré, était mal compris. Quand Miguel Toro (Universidad Politécnica de Madrid, Espagne) m'a posé la question : « mais quelle relation entre cette matrice **G** et la parenté moléculaire des généticiens de la conservation? » je n'ai pas su répondre et ceci a déclenchée une recherche avec lui et Luis Alberto García-Cortés (INIA, Madrid). En génétique de conservation et des espèces sauvages, des estimateurs de parenté existent depuis des lustres, le plus simple étant la parenté moléculaire ($F_{Mi,j}$), c'est-à-

dire la probabilité pour deux individus *i,j* que deux allèles tirés au hasard de chacun soient identiques (par état). Des relations avec la parenté classique étaient connues.

Nous avons d'abord proposé un autre estimateur de la parenté moléculaire : la covariance entre les doses géniques ({0,1,2} pour {AA,Aa,aa}) des individus, c'est-à-dire $Cov_M(i,j) = \mathbf{z}_i\mathbf{z}'_j$. Nous avons ensuite ré-découvert Cockerham (1969) qui montre que F_M et Cov_M ont des relations simples avec la parenté généalogique (« coancestry », ϕ) : $F_{Mi,j} = \phi_{i,j} + p^2 + q^2$, $Cov_{Mi,j} = pq\phi_{i,j}$. L'estimateur de VanRaden (2008) suit naturellement, mais il y en a d'autres. Nous avons aussi montré que tous les estimateurs sont sensibles à l'estimation des fréquences alléliques, surtout dans le cas de dérive, et qu'en prenant des espérances sur la distribution des fréquences les estimateurs étaient plus précis (voir ci-dessous).

Table 1 Features of the regression of genealogical coancestry *f* on molecular coancestry (f_M) and molecular covariance (Cov_M)

Nb SNP	Nb replicates	Regression on coancestry			Regression on covariance		
		a	b	R^2	a	b	R^2
$p = 0.50$							
100	1000	-0.66 (0.03)	1.38 (0.06)	0.69 (0.03)	0.03 (0.00)	2.77 (0.12)	0.69 (0.03)
10000	50	-0.99 (0.00)	1.99 (0.01)	1.00 (0.01)	0.00 (0.00)	3.98 (0.03)	1.00 (0.01)
Expected values		$-\frac{p^2+q^2}{2pq} = -1$	$\frac{1}{2pq} = 2$		0	$\frac{1}{pq} = 4$	
$p_i \sim Beta(1, 1)$							
100	1000	-1.01 (0.08)	1.58 (0.10)	0.52 (0.06)	-0.22 (0.04)	3.17 (0.21)	0.52 (0.06)
10000	50	-1.98 (0.02)	2.97 (0.03)	0.99 (0.02)	-0.50 (0.00)	5.95 (0.06)	0.99 (0.00)
Expected values		$-\frac{\bar{p}^2+\bar{q}^2+2Var(p)}{2\bar{p}\bar{q}-2Var(p)}$ $= -2$	$\frac{1}{2\bar{p}\bar{q}-2Var(p)}$ $= 3$		$-\frac{Var(p)}{\bar{p}\bar{q}-Var(p)}$ $= -0.5$	$\frac{1}{\bar{p}\bar{q}-Var(p)}$ $= 6$	

Intercept (a), slope (b) and coefficient of determination (R^2), with standard deviations across replicates, of the regression equation of genealogical coancestry *f* on molecular coancestry (f_M) and molecular covariance (Cov_M), based on simulated data, when the distribution of allele frequencies in the founders (p) is known and fixed ($p = 0.5$) or variable ($p_i \sim$ Beta(1,1)).

Extrait 21 Caractéristiques de quelques estimateurs de la parenté avec des marqueurs [A30]

Finalement, ce travail avait une fonction plus pédagogique (connexion des estimateurs dans deux écoles de génétique différentes) qu'innovatrice, mais la question de l'estimation des fréquences alléliques pour le calcul de **G** reste ouverte, comme on verra par la suite.

5.2.6. Modèle multicaractère pour l'évaluation génomique multi-raciale

Une des promesses de l'évaluation génomique est d'évaluer plus précisément les races de petite taille grâce au partage de QTL avec d'autres races plus grandes. Ce discours est un peu contradictoire avec le constat que les parentés les plus proches apportent la majeure partie de l'information ; il faudrait donc une grande quantité d'apparentés lointaines (d'autres races) pour compenser le faible nombre d'apparentés « proches » (intra-race). Des études empiriques (e.g. Pryce et al., 2011) montrent une certaine valeur de l'utilisation d'une population, par exemple, Holstein+Brune pour la prédiction de la Brune.

Ces modèles font l'hypothèse d'équivalence des effets des SNPs dans toutes les races. Or, l'ensemble des gènes en ségrégation n'est pas le même dans les deux races (par exemple,

DGAT1 est très polymorphe en Holstein mais peu en Montbéliarde), ni les motifs de déséquilibre de liaison. Varona et al. (2010) ont eu l'idée de postuler les effets des SNP dans les différentes races, **g**, comme corrélés mais pas identiques :

$\mathbf{g} = \begin{Bmatrix} \mathbf{g}_{race1} \\ \mathbf{g}_{race2} \end{Bmatrix}; \mathbf{g} \sim \mathbf{N}\left(\begin{matrix} \mathbf{0} \\ \mathbf{0} \end{matrix}, \mathbf{I} \otimes \begin{bmatrix} \sigma^2_{race1} & \sigma^2_{1,2} \\ \sigma^2_{2,1} & \sigma^2_{race1} \end{bmatrix} \right)$. Ce modèle conduit, après transformation en modèle équivalent, à un modèle de type GBLUP dans lequel on utilise une parenté génomique qui relie les individus de toutes les races :

$\mathbf{G} = \begin{pmatrix} \mathbf{G}_{race1} & \mathbf{G}_{race1,race2} \\ \mathbf{G}_{race2,race1} & \mathbf{G}_{race2} \end{pmatrix}$ mais en multi-caractère, c'est-à-dire seuls les individus de la race *A* ont une phénotype dans le caractère *A*. Ce modèle ressemble MACE (Schaeffer, 1994), mais dans notre cas aucun individu a de phénotypes dans plus d'une race.

Sofiene Karoui (thésard de María Jesús Carabaño et Clara Díaz à l'INIA, Madrid) a ensuite estimé les corrélations génétiques entre les différentes races françaises Holstein, Normande et Montbéliarde (c'est à dire entre les effets des SNPs dans ces différentes races). Seules les corrélations génétiques entre la Montbéliarde et la Holstein étaient appréciables (0,66 pour le taux butyreux, 0,79 pour le lait). Cela veut dire qu'un taureau bon en Holstein pour le taux butyreux n'est pas forcement bon en Normande, puisque la corrélation n'est que de 0.35.

Ensuite, Sofiene a montré que le gain de précision de l'évaluation génomique de plusieurs populations ensemble était en fait une fonction de cette corrélation génétique : dès qu'elle était faible (cas de la Normande avec les autres races), on n'obtient pas de gain en précision.

	Corrélation entre races	Montbéliarde	Normande	Holstein
Milk	1	0.19	0.13	0.30
	0	0.17	0.12	0.30
Fat content	1	0.33	0.40	0.47
	0	0.27	0.39	0.49
Fertility	1	0.20	0.07	0.10
	0	0.19	0.07	0.10

Extrait 22 Précision (R^2) de l'évaluation génomique multi-raciale (corrélation=1) ou uni-raciale (corrélation=0) [A42]

Ce travail est en cours de publication [A42].

5.2.7. Procédure à une étape (Single Step)

Les puces à SNP sont chères. Dans les populations d'animaux de rente, seuls certains animaux sont génotypés –principalement les mâles, même si ceci est en train de changer avec les puces bovines à faible coût. La situation se détériore dans les espèces laitières, où le mâle n'a pas de phénotype propre. Pour faire une évaluation génomique, soit on créé des pseudo-caractères en condensant l'information des apparentés des animaux génotypés, soit on construit un modèle plus général. Ce modèle plus général pouvait passer, selon moi, par la construction d'une matrice de parenté « unifiée » , **H**, qui combinerait les matrices de parentés généalogiques et génomiques :

$$\mathbf{A}=\begin{bmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{bmatrix};\mathbf{H}=\begin{bmatrix}? & ?\\ ? & \mathbf{G}\end{bmatrix}$$

où **A** et **H** sont divisés en blocs d'individus génotypés et non génotypés, et **G** est une matrice de parentés génomiques (cette matrice étant « très bonne » et donc fixée).

En fait, Ignacy Misztal (UGA) et son thésard Ignacio Aguilar (INIA, Uruguay) étaient sur la même piste et nous avons travaillé ensemble. Notre première idée fut :

$\mathbf{H}=\begin{bmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{G}\end{bmatrix}$ [A15], ce qui conduisait à un algorithme simple d'évaluation génétique, par des équations non symétriques (Henderson, 1984) qui utilisaient la matrice **A** et non $\mathbf{A}^{-1}$ (comme détaillé dans [A16]. Néanmoins, très rapidement, nous nous sommes aperçu que la matrice **H** ainsi définie n'était pas positive définie et, pire, n'avait pas de sens biologique : deux animaux pouvaient être apparentés et leur descendance non apparentée. L'information de **G** devait remonter aux ascendants et descendre aux descendants. Transmettre l'information de **G** aux descendants était simple *via* des matrices de transmission mendélienne [A15]. Pour l'ensemble de la généalogie, rien de plus naturel qu'utiliser des index de sélection, et nous l'avons formalisé de cette manière :

$$p(\mathbf{u}_2)=N(\mathbf{0},\mathbf{G});\ p(\mathbf{u}_1|\mathbf{u}_2)=N\left(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2,\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)$$

L'écriture de la matrice de covariance ainsi construite est :

$$\mathbf{H}=\begin{bmatrix}\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}+\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G}\end{bmatrix}$$

qui est une expression compliquée mais utilisable dans la méthode non-symétrique. Des inspections des matrices résultantes avec de petits exemples étaient satisfaisantes biologiquement, et deux articles [A15, A16] furent ainsi publiés.

Cependant, les équations non symétriques ne convergeaient pas pour de grosses masses de données (plus de 2 000 000 animaux), ce qui était troublant. Dave Johnson (LIC, Nouvelle Zelande) a déverrouillé la situation en montrant que

$\mathbf{H}^{-1}=\mathbf{A}^{-1}+\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{-1}+\mathbf{A}_{22}^{-1}\end{bmatrix}$, dont $\mathbf{G}^{-1}$ et $\mathbf{A}_{22}^{-1}$ sont des matrices denses mais de petites tailles (au moins en 2009). A ce moment, nous (principalement Ignacio Aguilar) avons développé des stratégies de calcul efficaces [A31] ; par exemple, des algorithmes de parallélisation ou l'algorithme de Colleau (2002) pour la construction de $\mathbf{A}_{22}$. Ceci permettra la première évaluation génétique avec toutes les données : phénotypes, généalogie et marqueurs [A18], sur des fichiers de grandes tailles (8 million d'équations, 6000 individus génotypés), avec des résultats comparables à l'évaluation en plusieurs étapes (très soignée et longue), mais beaucoup plus simple du point de vue calculatoire et conceptuel. Une autre dérivation du Single Step fut fait par Christensen et Lund (2010), sous une autre perspective d'imputation de génotypes manquants. Cette coïncidence fut très satisfaisante car elle nous rassura. Un exemple des résultats dans [A18] suit.

Table 1. Coefficients of determination (R^2) and coefficients (δ) for regression of 2009 daughter deviations (DD) or corresponding estimated breeding values (EBV_{09}) for bulls progeny tested from 2005 through 2009 on 2004 predictions obtained by different algorithms

	DD		EBV_{09}	
Prediction method	R^2 (%)	δ	R^2 (%)	δ
Parent average	24	0.76	36	0.79
Multiple-step	40	0.86	50	0.82
Single-step[1]				
G5	41	0.76	49	0.70
GB	38	0.68	45	0.63
GC	37	0.71	45	0.66
GG – G5	41	0.79	50	0.73
GG – GB	38	0.77	46	0.71
GG – GC	39	0.79	46	0.73

[1]Assumed allele frequency of 0.5 (G5), base population (GB), current population (GC), or calculated as in [30] of Gianola et al. (2009) (GG).

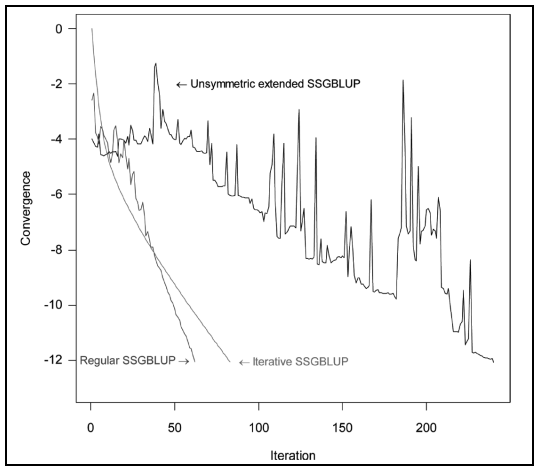
Extrait 23 Performance du Single Step [A18]

5.2.8. Stratégies calculatoires pour le Single Step

Le Single Step commença ainsi et est en train d'être utilisé par beaucoup de pays (Nouvelle Zélande, Finlande, etc.), même si le calcul de $\mathbf{G}^{-1}$ et $\mathbf{A}_{22}^{-1}$ reste long, voir impossible dans certains cas (bovins laitiers américains ou français avec plus de 100 000 individus génotypés). Ceci motiva Vincent Ducrocq à écrire un modèle similaire, dans lequel la valeur génétique d'un individu est la somme d'un effet polygénique « classique » et d'un effet « déviation » dû à l'écart entre l'information des SNPs et l'information généalogique: $\mathbf{u} = \mathbf{u}^* + \mathbf{d}$, et $Var(\mathbf{d}_2) = \mathbf{G} - \mathbf{A}_{22}$. En parallèle, j'étais conscient de la faiblesse du Single Step pour des fichiers de grandes tailles et j'avais développé des systèmes équivalents (ayant les mêmes solutions) pour simplifier le calcul. Ces systèmes n'étaient pas forcement faciles à programmer. Nous avons donc rejoint nos idées.

Il s'avéra que l'idée de Vincent était une dérivation (une troisième) du Single Step. Cette idée conduit à un système itératif (ci-dessous) pour résoudre le Single Step comme une succession d'évaluations génétiques « classiques » et « génomiques », ces dernières n'impliquant pas forcément le calcul explicite, ni de **G**, ni de $\mathbf{A}_{22}$. Nous avons démontré que la convergence du système est garantie, même quand $\mathbf{G} - \mathbf{A}_{22}$ n'est pas positive définie, à condition que les mises à jours entre évaluations suivent un schéma dit de « Succesive underrelaxation », avec un poids compris entre 0 et 1.

```
# Pseudocode for iterative SSGBLUP
# pick w between 0 and 1
# Xpy = (X'y ; Z'y)
phi_old = 0; gamma_old=0; sol_old=0
while ( ! convergence){
  RHS=Xpy
  #update of RHS of BLUP
  RHS(first_genotyped:last_genotyped)=
          RHS(first_genotyped:last_genotyped)
          +alpha*(phi-gamma)
  # solving regular BLUP
  sol=solve(LHS,RHS)
  # SUR updates
  sol=w*sol+(1-w)*sol_old
  u2=sol(first_genotyped:last_genotyped)
  # genomic solving
  phi=solve(G,u2)
  gamma=solve(A22,u2)
  # SUR updates
  phi=w*phi+(1-w)*phi_old
  gamma=w*gamma+(1-w)*gamma_old
  #storage of old solutions
  sol_old=sol
  gamma_old=gamma
  phi_old=phi
}
```



Convergence

Pseudocode : résolution itérative du Single Step.
Extrait 24 Algorithmes pour Single Step de grande taille [A38]

De plus, nous avons montré un système équivalent ayant les mêmes solutions que le Single Step et ne demandant pas la formation explicite ni de **G**, ni de $\mathbf{A}_{22}$:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}_1'\mathbf{W}_1 & \mathbf{X}_2'\mathbf{W}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{W}_1'\mathbf{X}_1 & \mathbf{W}_1'\mathbf{W}_1+\alpha_u\mathbf{A}^{11} & \alpha_u\mathbf{A}^{12} & \mathbf{0} & \mathbf{0} \\ \mathbf{W}_1'\mathbf{X}_2 & \alpha_u\mathbf{A}^{12} & \mathbf{W}_2'\mathbf{W}_2+\alpha_u\mathbf{A}^{22} & \alpha_u\mathbf{I} & -\alpha_u\mathbf{I} \\ \mathbf{0} & \mathbf{0} & \alpha_u\mathbf{I} & \alpha_u\mathbf{A}_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \alpha_u\mathbf{I} & \mathbf{0} & \alpha_u\mathbf{G} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \\ -\hat{\boldsymbol{\varphi}} \\ -\hat{\boldsymbol{\gamma}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}_1'\mathbf{y}_1 \\ \mathbf{W}_2'\mathbf{y}_2 \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

Les deux solveurs ont été montrés comme compétitifs dans une simulation, et nous avons montré que la quantité de mémoire à utiliser croît linéairement avec la taille du problème. Il reste à vérifier sa performance à échelle réelle. Ces idées ont été publiées dans [A38].

5.2.9. Effets de la sélection dans le Single Step

L'évaluation génétique classique par BLUP généalogique était résistante à la sélection, car l'information de la sélection était (grossièrement) contenue dans le BLUP même. Par contre, Ducrocq et Patry (2011) ont montré que, dès que les jeunes taureaux sont sélectionnés par évaluations génomiques, un biais est produit car le BLUP ne sait pas lire cette information. Le Single Step est en principe robuste à ce type de sélection car il utilise l'information génomique. Par contre, un autre type de sélection existe : le génotypage sélectif des animaux. Typiquement, soit seules les dernières générations sont génotypés, soit un ensemble d'individus sélectionnés sont génotypés (par exemple, tous les mâles élite de la population). En tous cas, cela induit qu'*a priori*, la moyenne des individus génotypés (sa base génétique

dans la parenté décrite par **G**) est différente de celle des individus non génotypés (la base génétique de **A**), contrairement aux hypothèses du Single Step. De plus, la variance génétique est réduite, du fait de la sélection. Comment pallier cela ? En fait, si l'on savait les fréquences alléliques dans la population de base de **A**, **G** serait construite en fonction de cette population de base ; or, ces fréquences sont difficiles à estimer.

Ce problème s'est révélé en même temps de plusieurs manières : dans nos simulations préliminaires pour Vitezica et al. [A28] et dans l'analyse de vraies données dans [A25], on constatait une détérioration de la précision dans le Single Step par rapport aux modèles en deux étapes.

Nous avons réalisé qu'en fait, ce décalage (μ) entre les moyennes de la population de base en **G** et en **A** pouvait être inclus dans le modèle, et qu'il était aléatoire car fonction des effets aléatoires. Sa variance était une fonction de la dérive possible. C'est-à-dire :

$$p(\mathbf{u}_2) = N(\mathbf{1}\mu, \mathbf{G}); p(\mathbf{u}_1|\mathbf{u}_2) = N\left(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2, \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right); p(\mu) = N(0, \alpha)$$

dont la valeur de α pouvait se déduire analytiquement comme fonction de **G** et $\mathbf{A}_{22}$. L'inclusion dans le Single Step était très facile. De manière très intéressante, l'article de Powell et al. (2010) nous aida à mettre en relation notre α avec la théorie des statistiques *F* de Wright, qui modélise la dérive dans une population panmictique, ce qui donna une formule qui corrige pour la dérive moyenne de la population mais aussi pour la réduction de variance génétique.

Finalement, la correction (BLUP$_{FST}$) devint $\mathbf{G}^* = (1 - 0.5\alpha)\mathbf{G} + \mathbf{11}'\alpha$, avec $\alpha = \bar{\mathbf{A}}_{22} - \bar{\mathbf{G}}$, la différence entre les parentés moyennes des deux populations ; et en fait $F_{st} = \alpha/2$. Cette dérivation est égale aux moyennes de la diagonale et « off-diagonal » de $\mathbf{A}_{22}$ et **G** en population panmictique, et elle s'est avérée très juste en population non panmictique (Christensen et al., 2012). Les données simulées de [A28] montraient un très bonne performance de BLUP$_{FST}$:

Table 4. *Squared correlations between true and EBVs (SDs) for different heritabilities and prediction methods under* P_Y *and* P_{EBV}

Heritability	Prediction method	P_Y	P_{EBV}
0·05	$BLUP_{PED}$	7 (4)	10 (4)
	$BLUP_{DYD}$	23 (5)	28 (7)
	$BLUP_{1STEP}$	29 (5)	25 (7)
	$BLUP_{\alpha}$	29 (5)	27 (8)
	$BLUP_{F_{ST}}$	29 (5)	27 (7)
0·30	$BLUP_{PED}$	20 (4)	23 (6)
	$BLUP_{DYD}$	49 (5)	56 (6)
	$BLUP_{1STEP}$	54 (4)	47 (6)
	$BLUP_{\alpha}$	55 (5)	60 (5)
	$BLUP_{F_{ST}}$	55 (5)	60 (5)
0·50	$BLUP_{PED}$	21 (4)	30 (6)
	$BLUP_{DYD}$	56 (5)	64 (6)
	$BLUP_{1STEP}$	61 (5)	49 (7)
	$BLUP_{\alpha}$	61 (5)	67 (5)
	$BLUP_{F_{ST}}$	61 (5)	67 (5)

Extrait 25 Précision d'un Single Step avec correction pour la sélection [A28]

De plus, on peut apprécier dans [A25] la manque de biais:

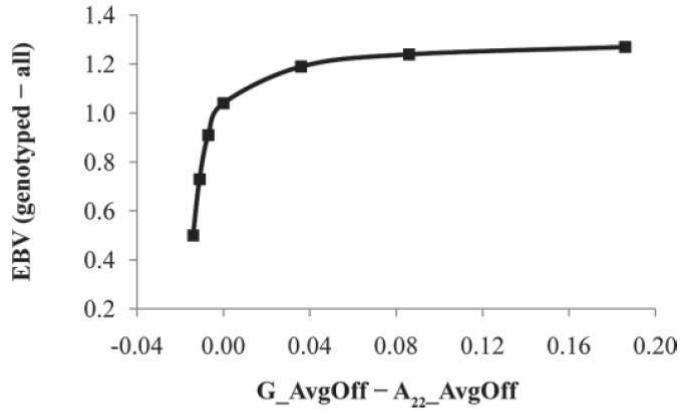


Figure 1. The average difference between the EBV of genotyped birds and all birds as a function of the difference between mean off-diagonal (AvgOff) elements of the genomic (**G**) and relationship ($\mathbf{A}_{22}$) matrices. The average difference between the EBV of genotyped birds and all birds in BLUP was 1.10.

Extrait 26 Biais empirique dans une population de poulet [A25]

Cette correction, qui reste empirique, revient en quelque sorte à postuler une matrice de covariance dont l'*a priori* serait **A** et l'observation serait **G**. Plusieurs idées peuvent (doivent) être développées encore, par exemple, pour utiliser l'imputation de génotypes ou considérer la profondeur de la généalogie. Ces travaux ont été fait en collaboration, entre autres, avec Zulma Vitezica (ENSAT, Toulouse), Ignacy Misztal (UGA), Ignacio Aguilar (INIA, Uruguay), Ching Yi Chen (UGA).

5.2.10. Localisation de QTL par GWAS avec le Single Step

Dans la détection de QTL, nous trouvons le même problème que dans l'évaluation génomique : seul un sous-ensemble d'individus est génotypé. L'approche traditionnelle consiste à condenser l'information des apparentés sur les individus en question, comme par exemple performances de ses descendantes pour le bovin laitier. Cette procédure peut s'étendre, mais elle fait l'hypothèse d'individus (fondateurs) non apparentés et de plus, a besoin d'une structure de données homogène (c'est-à-dire, toutes les familles sont du même type). Notre groupe « Single Step » (Ignacy Misztal, Ignacio Aguilar, Huiyu Wang et autres) a soupçonné que l'on pouvait utiliser le Single Step pour ce type de situations. Nous avons fait les développements suivants.

Les modèles de type GBLUP font une hypothèse de normalité multivariée des effets des SNP et des valeurs génétiques des individus. On peut montrer que, de la même manière que la valeur génétique d'un individu est la somme des effets des SNPs, les effets des SNPs se calculent à partir des solutions des valeurs génétiques des individus : $\widehat{SNPs} = Cov(SNPs, \mathbf{u}) Var(\mathbf{u})^{-1} \hat{\mathbf{u}}$ (Henderson, 1973); dans une notation plus formelle :

$$\hat{\mathbf{a}} = \mathbf{DZ'G}^{-1}\hat{\mathbf{u}} = \mathbf{DZ'}(\mathbf{ZDZ'})^{-1}\hat{\mathbf{u}}.$$

Cette expression est valable tant en GBLUP qu'en Single Step et on peut donc obtenir des solutions des SNP à partir du Single Step. De plus, on peut pondérer chaque SNP dans la matrice **D** par son effet (dans ce travail-ci nous avons fait une approximation de ce que nous avons fait dans [A23] pour le HetVarGBLUP). En itérant, on obtient des approximations plus précises des effets des SNP. En regardant individuellement leurs effets, on peut localiser les QTL soupçonnés.

Joy (Huiyu Wang), thésarde d'Ignacy Misztal a donc entamé des simulations pour vérifier nos idées. Le Single Step a donné des résultats comparables à BayesB en précision, tout en étant plus simple et rapide [A35]. Un exemple est montré dans la figure ci-dessous :

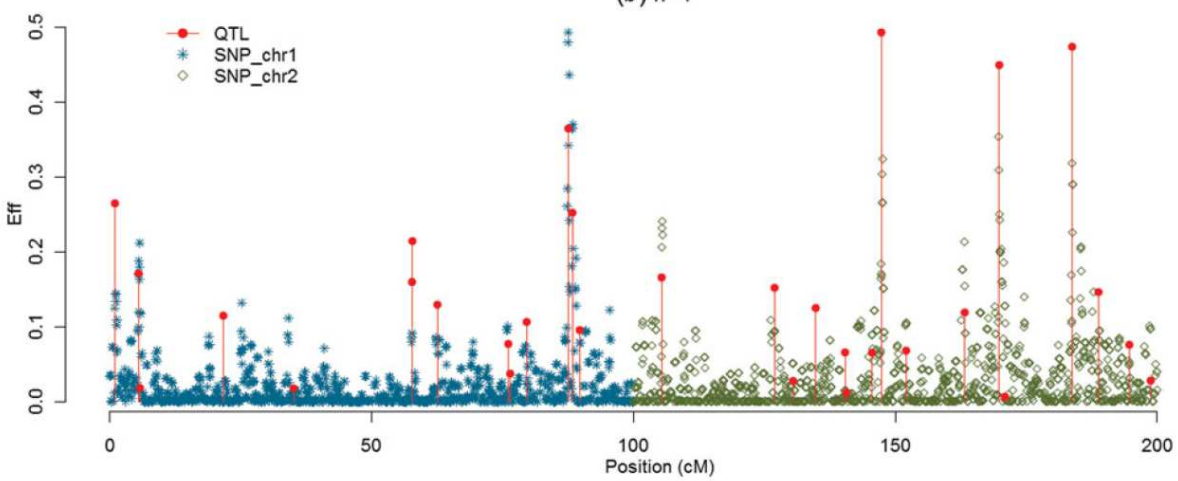


Fig. 2. SNP solutions and their four-point moving averages from ssGBLUP/S1 in the third iteration: (*a*) SNP solutions and (*b*) four-point moving average.

Extrait 27 GWAS avec le Single Step [A35]

5.3. Conclusion

Mes nombreux travaux en évaluation génomique sont le fruit d'un goût personnel mais aussi du hasard, qui a fait que des nombreux problèmes liés à l'évaluation génomique apparaissent au cours de ces dernièrrs annéess. Il faut tirer la conclusion de qu'il est nécessaire d'entretenir le savoir-faire car on ne sait pas quand il sera nécessaire.
Je trouve personnellement que l'évaluation génomique permet de réconcilier trois mondes qui ne se parlaient (plus) trop : les évaluateurs BLUP avec leurs algorithmes dédiés, mais dont la théorie n'évoluait guère depuis les années 80, les généticiens des populations, qui manipulent les notions de gènes au sein d'une population, et les chasseurs de QTLs, qui fournissent les outils de manipulation statistique des morceaux du génome. Cette réconciliation est longue mais fascinante.

6. L'implémentation des méthodes

Je voudrais dédier un petit chapitre aux outils que j'ai fait ou mis en place. Notre rôle n'est pas seulement de générer une connaissance obscure pour qu'elle soit gardé précieusement dans les journaux, mais aussi faire vivre les méthodes et l'amélioration génétique en générale avec des outils conviviaux.

6.1. TM

Le TM (de « Threshold Model ») est un logiciel que nous avions construit pour la thèse d'Evangelina López de Maturana et avec la participation de Luis Varona (IRTA, Lleida), car le peu de logiciels existants ne s'adaptaient pas bien à nos besoins. C'est donc un logiciel capable d'estimer les parts de variance par des méthodes Bayésiennes d' échantillonnage de Gibbs, pour des modèles à multiples effets fixes et aléatoires (y compris génétiques), et pour des phénotypes continus, catégoriels, et censurés. Il est disponible sur le site http://snp.toulouse.inra.fr/~alegarra . Nous avons voulu le faire convivial et il dispose d'un interface de fichiers de paramètres ainsi que de sorties facilement utilisables. Nous nous sommes servi du TM, tantôt en tant que tel, tantôt comme base de développement [A5, A9, A14, A19, A24, A33]. Une liste d'articles considérable cite TM, même s'il n'a pas été publié formellement.

6.2. GS3

GS3 (contraction de « Genomic Selection, Gauss Seidel, Gibbs Sampling ») est un logiciel que j'ai développé lors des travaux publiés dans [A12] pour l'estimation d'effets des SNP en sélection génomique. A l'origine, il estimait des effets des SNPs selon l'algorithme de GSRU déjà mentionné, ainsi que les composantes de la variance. Après, et avec la participation d'Anne Ricard et Olivier Filangi (GAREN, INRA Rennes), nous avons introduit le BayesCPi (algorithme de mélange qui permet en principe de fixer certains SNP à 0) ainsi que le Lasso Bayésien. Il a la capacité d'introduire plusieurs effets fixes et aléatoires, y compris un effet génétique additif modélisé selon la généalogie. Il s'utilise avec un fichier de paramètres. Nous l'avons utilisé de manière intense pour la sélection génomique [A11, A12, A23, A36, A42] et il commence à être utilisé à l'extérieur de l'INRA.

7. Directions futures

Lors de mes travaux de recherche, j'ai été confronté à plusieurs problèmes que j'ai résolu avec plus ou moins de succès. Je suis à une étape où je me rends compte de l'importance de la transmission des connaissances, non seulement *via* les publications mais aussi à travers la diffusion d'outils, l'encadrement de thésards, la direction de projets, et l'enseignement. J'ai encadré une thésarde (Aurélie Favier) et j'ai collaboré étroitement avec beaucoup d'autres thésards (Evangelina López de Maturana, Manuel Ramón, L. Tusell, Sofiene Karoui, C. Colombani, I. David, S. Duchemin, I. Aguilar, P. López-Romero, H. Wang, F. Ytournel, G. Sallé). Je voudrais continuer à explorer l'amélioration génétique, avec un œil sur les applications, et en même temps mieux diffuser les connaissances.

Quelles lignes de recherche ? Je prévois (de manière peut-être très naïve) certaines directions de recherche possibles. Elles tournent autour de la conciliation de la génétique quantitative « moléculaire » et « infinitésimale » pour l'évaluation génétique, ainsi que de la définition des objectifs et stratégies de sélection.

7.1. Dans l'immédiat

Le déséquilibre de liaison et l'identité aux marqueurs sont à la base des méthodes de détection de QTLs et de la sélection génomique. Ils sont une fonction de la parenté réalisée. Or, cette fonction a été très peu caractérisée dans la littérature, même si nous y avons un peu contribué [A30]. Une vraie conciliation des apparentés génomique et généalogique passe par l'utilisation de la parenté généalogique comme un *a priori* et des génotypes comme une « observation ». L'« observation » et l'*a priori* doivent se combiner pour créer une notion de parenté *a posteriori* (cette parenté peut être locale –au locus– ou à l'échelle du génome). Nous sommes en train de développer (avec Miguel Toro, Luis Alberto García Cortés, et Claude Chevalet – LGC, INRA Toulouse) les expressions qui permettent de calculer la distribution à un locus de la parenté « vraie » (réalisée) en fonction de la parenté généalogique ; par exemple, quelle est la variance de la parenté réalisé entre deux frères (qui peuvent avoir en réalité 0, 1 ou 2 allèles au locus identiques par descendance). Ce développement devrait permettre de pondérer l'information apporté par la parenté généalogique par rapport aux données génomiques.

Avec l'arrivée des données de séquence, je prévois aussi de travailler avec mes collègues sur des méthodes d'estimation efficaces pour l'évaluation génomique, comme dans [A11, A31, A38]. Et ceci, pour des évaluations « pures » (tous les individus sont génotypés) ou de type Single Step (un sous-ensemble des animaux est génotypé), avec une possible généralisation à des méthodes non-linéaires mais efficaces telles que le Lasso. De fait, une extension non-linéaire du Single Step, qui est esquissé dans [A38] est en cours.

Un vrai défi est l'inclusion des effets non additifs (dominants) dans les évaluations génomiques. D'abord, il faut préciser le concept : l'effet additif ou de substitution dépend de l'effet de chaque *génotype* et des fréquences alléliques dans la population. L'effet de dominance est un écart qui dépend des fréquences supposées. Les équivalences par rapport aux modèles infinitésimaux doivent être établies, ainsi que la connexion entre la corrélation génétique entre populations [A42] et le modèle classique avec dominance (Lo et al., 1993). Par contre, une vraie différence par rapport aux approches infinitésimales est que l'on *voit*

dans le génotype l'incidence de la dominance au SNP, ce qui donnera des précisions d'estimation beaucoup plus grandes que dans les approches qui utilisent la généalogie. Des algorithmes dédiés et une connexion avec le Single Step devront être développés, ainsi que des vérifications de son utilité pratique. Ensuite, une stratégie d'utilisation optimale en sélection devrait être développée (mais ceci est loin d'être ma spécialité).

7.2. Dans le futur

Pour la compréhension et la réconciliation de la génétique quantitative « moléculaire » et « généalogique », une ligne de recherche encore plus ambitieuse passe par l'inclusion de la théorie de la coalescence, qui permet de gérer des évènements rarement (voir jamais) considérés en évaluation génétique, comme la mutation ou la divergence de races et sous-espèces.

Les données de séquence ont une autre particularité : la présence de variants que l'on ne peut pas classifier comme « allèles » (CNV, insertions, délétions, transpositions…). Comment peut-on inclure cette information dans la théorie quantitative et les modèles statistiques ? Peut-on se servir des annotations, des comparaisons entre espèces… ?

L'aspect « interactions » peut s'étendre au-delà de l'interaction au sein d'un locus (dominance) pour inclure l'interaction entre SNPs (épistasie) mais aussi l'interaction avec l'environnement, qui peut être macroscopique (interaction avec une variable mesurée comme, par exemple, la pluviométrie) ou microscopique (control génétique de la variabilité). Pour ce faire, il en existe des outils classiques en quantitative infinitésimale, mais ils sont peu utilisés. Tous ces aspects n'ont pas toujours été considérés dans la construction de programmes de sélection : vaut-il mieux une variété robuste ou performante ? Comment quantifier la robustesse? Comment quantifier économiquement son intérêt ?

8. Références additionnelles

Bijma, P. 2011. A General Definition of the Heritable Variation That Determines the Potential of a Population to Respond to Selection. Genetics 189:1347-1359.

Boichard, D., B. Bonaiti, A. Barbat, and S. Mattalia. 1995. Three methods to validate the estimation of genetic trend for dairy cattle. Journal of Dairy Science 78:431-437.

Browning, S. R. and B. L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. The American Journal of Human Genetics 81:1084-1097.

Christensen, O., P. Madsen, B. Nielsen, T. Ostersen, and G. Su. 2012. Single-step methods for genomic evaluation in pigs. Animal: first view doi:10.1017/S1751731112000742.

Christensen, O. F. and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet Sel Evol 42:2.

Cockerham, C. C. 1969. Variance of gene frequencies. Evolution 23:72-84.

Colleau, J. J. 2002. An indirect approach to the extensive calculation of relationship coefficients. Genetics Selection Evolution 34:409-422.

Gianola, D. 1982. Theory and analysis of threshold characters. Journal of Animal Science 54:1079.

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando. 2009. Additive genetic variability and the Bayesian alphabet. Genetics 183:347-363.

Gianola, D. and H. Simianer. 2006. A Thurstonian model for quantitative genetic analysis of ranks: A Bayesian approach. Genetics 174:1613-1624.

Gianola, D. and D. Sorensen. 2004. Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. Genetics 167:1407-1424.

Goddard, M. 1998. Consensus and debate in the definition of breeding objectives. Journal of Dairy Science 81:6-18.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245-257.

Henderson, C.R. 1973. Sire evaluation and genetic trends. Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr D. L. Lush. Champaign, Ill., Pages 10-41.

Henderson, C. R. 1984. Applications of Linear Models in Animal Breeding. University of Guelph, Guelph.

Lo, L. L., R. L. Fernando, and M. Grossman. 1993. Covariance between relatives in multibreed populations – Additive Model. Theoretical and Applied Genetics 87:423-430.

Meuwissen, T. H. E., A. Karlsen, S. Lien, I. Olsaker, and M. E. Goddard. 2002. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics 161:373-379.

Misztal, I. and D. Gianola. 1987. Indirect solution of mixed model equations. Journal of Dairy Science 70:716-723.

Patry, C. and V. Ducrocq. 2011. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. Journal of Dairy Science 94:1011-1020.

Pérez-Enciso, M. 2003. Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: a Bayesian unified framework. Genetics 163:1497-1510.

Powell, J. E., P. M. Visscher, and M. E. Goddard. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nature Reviews Genetics 11:800-805.

Pryce, J., B. Gredler, S. Bolormaa, P. Bowman, C. Egger-Danner, C. Fuerst, R. Emmerling, J. Sölkner, M. Goddard, and B. Hayes. 2011. Short communication: Genomic selection using a multi-breed, across-country reference population. Journal of Dairy Science 94:2625-2630.

Rendel, J. and A. Robertson. 1950. Estimation of genetic gain in milk yield by selection in a closed herd of dairy cattle. Journal of Genetics 50:1-8.

Tanner, M. A. 1996. Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions. Springer Verlag.

Tavernier, A. 1991. Genetic evaluation of horses based on ranks in competitions. Genetics Selection Evolution 23:59-173.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B 58:267-288.

VanRaden, P. M. 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91(11):4414-4423.

Varona, L., C. Moreno, N. Ibanez-Escriche, and J. Altarriba. 2010. Whole genome evaluation for related populations. Page 460 in Proc. 9th World Congress on Genetics Applied to Livestock Production, Leipzig.

Varona, L., D. Sorensen, and R. Thompson. 2007. Analysis of litter size and average litter weight in pigs using a recursive model. Genetics 177:1791-1799.

Visscher, P., P. Bowman, and M. Goddard. 1994. Breeding objectives for pasture based dairy production systems. Livestock Production Science 40:123-137.

Wijsman, E. M. 1987. A deductive method of haplotype analysis in pedigrees. American Journal of Human Genetics 41:356.

9. Publications

9.1. Articles dans journaux à comité de lecture

[A1] A. Legarra and E. Ugarte, 2001. Genetic parameters of milk traits in Latxa dairy sheep. *Animal Science,* 73, 407-412.

[A2] M. Serrano, E. Ugarte, J.J. Jurado, M.D. Pérez-Guzmán and A. Legarra, 2001. Test-day models and genetic parameters in Latxa and Manchega dairy ewes. *Livestock production science,* 67:253-264.

[A3] A. Legarra, I. Misztal, J.K. Bertrand, 2004. Constructing Covariance Functions for Random Regression Models for Growth in Gelbvieh Beef Cattle. *Journal of Animal Science,* 82: 1564-1571.

[A4] A. Legarra, P. López-Romero, and E. Ugarte, 2005. Bayesian model selection of contemporary groups for BLUP genetic evaluation in Latxa dairy sheep. *Livestock Production Science,* 93: 205-212.

[A5] A. Legarra, E. Ugarte. 2005. Genetic Parameters of Udder Traits, Somatic Cell Score and Milk Yield in Latxa Sheep. *Journal of Dairy Science*, 88: 2238-2245.

[A6] L. Alfonso, A. Parada, A. Legarra, E. Ugarte, A. Arana. 2006. Effects on genetic variability of selection against scrapie sensitivity in the Latxa Black-Faced sheep. *Genetics, Selection, Evolution,* 38:495-512.

[A7] A. Legarra, M. Ramón, E. Ugarte, M.D. Pérez-Guzmán, 2007. Economic weights of fertility, prolificacy, milk yield and longevity in dairy sheep. *Animal*, Volume 1, Issue 02, pp 193-203.

[A8] A. Legarra, M. Ramón, E. Ugarte, M.D. Pérez-Guzmán, J.Arranz, 2007 Economic weights of somatic cell score in dairy sheep. *Animal*, Volume 1, Issue 02, pp 205-212.

[A9] E. López de Maturana, A. Legarra, E. Ugarte, 2007. Analysis of Fertility and Dystocia Using Recursive Multivariate Models, Handling Censored and Categorical Data in Holsteins. *Journal of Dairy Science,* 90: 2012-2024.

[A10] A. Legarra, J.K. Bertrand, T. Strabel, R. Sapp, J.P. Sanchez and I. Misztal, 2007. Multi-breed genetic evaluation in a Gelbvieh population. *Journal of Animal Breeding and Genetics*, 124: 286-295.

[A11] A. Legarra, I. Misztal, 2008. Computing strategies in genome-wide selection. *Journal of Dairy Science,* J. Dairy Sci. 91:360–366.

[A12] A. Legarra, C. Robert-Granié, E. Manfredi and J.M. Elsen. 2008. Performance of genomic selection in mice. *Genetics,* 180:611-618.

[A13] G. de los Campos, H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, J.M. Cotes. 2009. Predicting Quantitative Traits with Regression Models for Dense Molecular Markers and Pedigrees. *Genetics,* 182: 375–385.

[A14] I. David, L. Bodin, A. L.egarra, C. Robert-Granié. 2009. Product versus additive threshold models for analysis of reproduction outcomes in animal genetics. *Journal of Animal Science* 87:2510-2518.

[A15] A. Legarra, I. Aguilar, I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science*, 92:4656-4663.

[A16] I. Misztal, A. Legarra, I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree and genomic information. *Journal of Dairy Science*, 92:4648-4655.

[A17] A. Legarra and R. Fernando. 2009. Linear models for joint association and linkage QTL Mapping. *Genetics Selection Evolution* 41:43.

[A18] I. Aguilar, I. Misztal, D. Johnson, A. Legarra, S. Tsuruta, T. Lawlor, 2010. A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science*, 93 :743–752.

[A19] A. Ricard, A. Legarra, 2010. Validation of models for analysis of ranks in horse breeding evaluation. *Genetics Selection Evolution* ,42:3.

[A20] M. Ramón, A. Legarra, E. Ugarte, J. J. Garde, and M. D. Pérez-Guzmán. 2010. Economic weights for major milk constituents of Manchega dairy ewes. *J. Dairy Sci.* 93 :3303–3309

[A21] C Cierco-Ayrolles, S Dejean, A Legarra, H Gilbert, T Druet, F Ytournel , D Estival , N Oumouhou and B Mangin. 2010. Does probabilistic modelling of linkage disequilibrium evolution improve the accuracy of QTL location in animal pedigrees? Genetics Selection Evolution, 42:38.

[A22] A. Legarra, F. Calenge, P. Mariani, P. Velge, C. Beaumont. Use of a reduced set of SNP for genetic evaluation of resistance to Salmonella carrier state in laying hens. Poultry Science, 2011, 90:731-736.

[A23] A. Legarra, C. Robert-Granié, P. Croiseau, F. Guillaume and S. Fritz. 2011. Improved Lasso for Genomic Selection. Genetics Research, 93, pp. 77–87

[A24] Ll. Tusell, A. Legarra, M. García-Tomás, O. Rafel, J. Ramon, and M. Piles. 2011. Different ways to model biological relationships between fertility and the pH of the semen in rabbits. Journal of Animal Science, 89:1294-1303.

[A25] C. Y. Chen, I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. Effect of different genomic relationship matrices on accuracy and scale. 2011. Journal of Animal Science, 89:2673-2679.

[A26] L Tusell, I David, L Bodin, A Legarra, O Rafel, M López-Bejar, M Piles. 2011. Using the product threshold model for estimating separately the effect of temperature on male and female fertility. Journal of Animal Science, 89:3983-3995.

[A27] R. Simeone, I. Misztal, I. Aguilar, and A. Legarra. 2011. Evaluation of the utility of diagonal elements of the genomic relationship matrix as a diagnostic tool to detect mislabelled genotyped animals in a broiler chicken population. Journal of Animal Breeding and Genetics, 128:386-393.

[A28] Z. G. Vitezica, I. Aguilar, I. Misztal and A. Legarra. 2011. Bias in Genomic Predictions for Populations Under Selection. Genetics Research, 93:357-366.

[A30] M.A. Toro, L.A. García-Cortés, A. Legarra. 2011. A note on the rationale for estimating genealogical coancestry from molecular markers. Genetics, Selection, Evolution, 43:27.

[A31] Aguilar, I., Misztal, I., Legarra, S. Tsuruta. 2011. Efficient computations of genomic relationship matrix and other matrices used in the single-step evaluation. Journal of Animal Breeding and Genetics, 128:422-428.

[A32] P. Croiseau, A. Legarra, F. Guillaume, S. Fritz, A. Baur, C. Colombani, C. Robert-Granié, D Boichard and V. Ducrocq. Fine tuning genomic evaluations in dairy cattle through SNP pre-selection with the Elastic-Net algorithm. Genetics Research, in press.

[A33] L. Tusell, A. Legarra, M. García-Tomás, O. Rafel, J. Ramon and M. Piles. Genetic basis of semen traits and their relationship with growth rate in rabbits. Journal of Animal Science, J ANIM SCI November 18, 2011 jas.2011-4165.

[A34] C. Marie-Etancelin, B. Basso, S. Davail, K. Gontier, X. Fernandez, Z. G. Vitezica, D. Bastianelli, E. Baéza, M.-D. Bernadet, G. Guy, J.-M. Brun, A. Legarra. 2011. Genetic parameters of product quality and hepatic metabolism in fattened mule ducks.

[A35] H. Wang, I. Misztal, I. Aguilar, A. Legarra, W.M. Muir, 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res., Camb. (2012), 94, pp. 73–83

[A36] C. Colombani, P. Croiseau, S. Fritz, F. Guillaume, A. Legarra, V. Ducrocq, and C. Robert-Granié. A comparison of PLS and sparse PLS regressions in genomic selection in French dairy cattle. Journal of Dairy Science, 95:2120-2131.

[A37] S. I. Duchemin, C. Colombani, A. Legarra, G. Baloche, H. Larroque, J-M. Astruc, F. Barillet, C. Robert-Granié, and E. Manfredi. Genomic selection in French Lacaune dairy sheep breed. Journal of Dairy Science, 95:2723-2733.

[A38] A. Legarra, and V. Ducrocq. Computational strategies for national integration of phenotypic, genomic and pedigree data in a single-step BLUP. Journal of Dairy Science, 95:4629–4645.

[A39] D. L. Roldan, H. Gilbert, J. Henshall., A. Legarra, J.M. Elsen. Fine-mapping quantitative trait loci with a medium-high marker density panel: efficiency of population structures and comparison of linkage disequilibrium linkage analysis models. Genetics Research, in press.

[A40] Ytournel F., Teyssèdre S., Roldan D., Erbe M., Simianer H., Boichard D., Gilbert H., Druet T., Legarra A. LDSO: A program to simulate pedigrees and molecular information under various evolutionary forces. Journal of Animal Breeding and Genetics, in press, doi: 10.1111/j.1439-0388.2011.00986.x.

[A41] G. Sallé, P. Jacquiet, L. Gruner, J. Cortet, C. Sauvé, F. Prévot, C. Grisez, J. P. Bergeaud, L. Schibler, A. Tircazes, D. François, C. Pery, F. Bouvier, J. C. Thouly, J. C. Brunel, A. Legarra, J. M. Elsen, J. Bouix, R. Rupp and C. R. Moreno. A genome scan for QTL affecting resistance to Haemonchus contortus in sheep. Journal of Animal Science, in press, doi: 10.2527/jas.2012-5121.

Soumis

[A42] S. Karoui, M. J. Carabaño, C. Diaz, A. Legarra. Genomic evaluation combining different French dairy cattle breeds. Submitted to Genetics, Selection Evolution.

[A43] C. Colombani, A. Legarra, S. Fritz, F. Guillaume, P. Croiseau, V. Ducrocq, and C. Robert-Granié. Application of Bayesian Lasso and Bayes Cπ CPi for genomic selection in French Holstein and Montbéliarde breeds. Submitted to Journal of Dairy Science.

En Français

[A44] Robert-Granié C., Legarra A., Ducrocq V., 2011. Principes de base de la sélection génomique. In : Numéro spécial, Amélioration génétique. Mulsant P., Bodin L., Coudurier B., Deretz S., Le Roy P., Quillet E., Perez J.M. (Eds). INRA Prod.Anim., 24, 331-340. (Invited)

En Espagnol

[A45] A. Legarra and E. Ugarte, 2001. Resultados de la aplicación de metodologías de extensión de lactaciones a 120 días en ovejas de la raza Latxa. *ITEA* 97A:104-116.

[A46] A. Legarra, E. Ugarte y F. Arrese, 2003. Análisis del progreso genético en el esquema de mejora de la raza Latxa. *ITEA* 99A: 192-202.

En congrès à comité de lecture

[A47] A. Favier, S. de Givry, A. Legarra, T. Schiex, Pairwise decomposition for combinatorial optimization in graphical models, Proceeding of 22th international joint conference on artificial intelligence (IJCAI'11), Barcelona, 2011.

En congrès à comité de lecture, en Français

[A48] A. Favier, S. de Givry, A. Legarra, T. Schiex, Décomposition par paire pour l'optimisation combinatoire dans les modèles graphiques, 7èmes Journées Francophone de Programmation par Contraintes, Lyon, 2011

9.2. Communications

[B1] A. Legarra and F. J. Mendizábal: Sheep milk quality in the Basque Country Community and Navarra from 1996 to 1998. Poster. Seminar on Production systems and product quality, Murcia, Spain, 23-25 September 1999, FAO-CIHEAM Network on sheeps and goats. *Options Méditerranéennes*, serie A, 46, 145-149.

[B2] E. Ugarte, A. Legarra, I. Beltrán de Heredia and J. Arranz. Udder morphology: a new trait to introduce in the Latxa breeding programme. 52nd Annual Meeting of the EAAP. Budapest, August 26-29 2001.

[B3] E. Ugarte, A. Legarra. Scientific background of the selection program in the Latxa breed. Proceedings of the meeting of the FAO-CIHEAM os Genetic resources of shhep and goats. Sassari, Italia, 9-11 de mayo de 2002. *Options Méditerranéennes*, serie A, 55, 91-98.

[B4] E. Ugarte, A. Legarra. Organisational changes in the Latxa breeding programme to introduce selection for milk quality. Proceedings of the meeting of the FAO-CIHEAM os Genetic resources of sheep and goats. Sassari, Italia, 9-11 de mayo de 2002. *Options Méditerranéennes*, serie A, 55, 141-146.

[B5] Legarra, A., López-Romero, P., y Ugarte, E., 2002. Bayesian model selection : an application to genetic evaluation of the Latxa dairy sheep. Proc 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France. CD-ROM communication nº 20-07.

[B6] Legarra, A, Strabel, T., Bertrand, J.K., Misztal, I., 2003. Setting up the Gelbvieh multiple breed evaluation. ADSA and ASAS joint meeting, Phoenix, EEUU, 2003. J. Dairy Sci., Vol. 86, Suppl.1, p.198.

[B7] Legarra, A, Misztal, I., Jamrozik, J. 2003. Plotting covariance functions from random regression models. ADSA and ASAS joint meeting, Phoenix, EEUU, 2003. J. Dairy Sci., Vol. 86, Suppl.1, p.113.

[B8] K.R. Robbins, I. Misztal, J. K. Bertrand, A. Legarra, and S. Tsuruta, 2004. A practical longitudinal model for evaluating growth in Gelbvieh cattle. ADSA and ASAS joint meeting, St Louis, USA, 2004. J. Anim Sci., Vol. 82, Suppl.I, p.243.

[B9] A. Legarra, E. Ugarte, 2004. A rationale to introduce more traits in the Latxa breeding program. 55th annual meeting of the European Association for Animal Production. Book of abstracts, p. 237

[B10] A. Legarra, C. Robert-Granié. Computation of recursive models and an example on fertility traits. 8th World Congress on Genetics applied to Livestock Production, Belo Horizonte, Brasil, August 2006, CD-ROM communication 24-07.

[B11] Z.G. Vitezica, A. Legarra. Accuracy of genotype estimation using loop breakers. 8th World Congress on Genetics applied to Livestock Production, Belo Horizonte, Brasil, August 2006, CD-ROM communication 20-13.

[B12] E. López de Maturana, A. Legarra, E. Ugarte. Effects of calving ease on fertility in the basque Holstein population using recursive methodology. 8th World Congress on Genetics applied to Livestock Production, Belo Horizonte, Brasil, August 2006, CD-ROM communication 01-23.

[B13] Legarra, A., C. Robert-Granié, E. Manfredi, and J.M. Elsen. 2007. Does genomic selection work in a mice population? Pages 66-74 in XI QTLMAS 2007. Papers and abstracts from the Workshop on QTL and Marker Assisted Selection, 22-23 March 2007, Toulouse, France. http://germinal.toulouse.inra.fr/qtlmas/

[B14] Legarra, A., E. Manfredi, C. Robert-Granié, and J.M. Elsen. 2007. Validation of genomic selection in an outbred mouse. 58th annual meeting of the EAAP, Dublin, 2007. Book of abstracts page 162.

[B15] Garreau H, Eady S.J., Hurtaud J., Legarra A., 2008. Genetic parameters of production traits and resistance to digestive disorders in a commercial rabbit population. Proc. 9th World Rabbit Congress, Verona, Italy, June 10-13.

[B16] A. Legarra, R. Fernando, 2009. Joint association and linkage QTL mapping on half-sib families by regression. 13th QTLMAS workshop, Wageningen, the Netherlands.

[B17] A. Legarra, I. Aguilar, I. Misztal. Whole-population relationship matrix including pedigree and markers for genomic selection. 2009. Poster, Symposium SGLPGE, Madison, Wisconsin.

[B18] A. Legarra, I. Aguilar, I. Misztal. Whole-population relationship matrix including pedigree and markers for genomic selection. 2009. Poster, EAAP meeting, Barcelona, Spain.

[B19] I. Misztal, A. Legarra, I. Aguilar. Computing procedures for genetic evaluation including phenotypic, full pedigree and genomic information. 2009. Oral communication, EAAP meeting, Barcelona, Spain.

[B20] Elsen J. M., Filangi O., Gilbert H., Legarra A., Le Roy P., Moreno-Romieu C., 2009. QTLMAP a software for the detection of QTL in full and half sib families. 60th Annual Meeting EAAP, Barcelone, Espagne, 24-27 Aout 2009, Session 53, Theatre 13, p.603.

[B21] Ricard A., Legarra A., 2009. Ranking in competition : an efficient tool to measure aptitude. 60th Annual Meeting EAAP, Barcelone, Espagne, 24-27 Aout 2009, Session 19, Theatre 7, p. 216.

[B22] I. Misztal, I. Aguilar, D. Johnson, A. Legarra, S. Tsuruta, T.J. Lawlor. A unified approach to utilize phenotypic, full pedigree, and genomic information for a genetic evaluation of Holstein final score. 2009. Interbull Bulletin 40.

[B23] C. Marie-Etancelin, X Fernandez, S. Davail, J.M. André, D. Bastianelli, E. Baéza, M.D. Bernadet, G. Guy, A. Legarra, J.M. Brun. 2009. Genetic parameters of mule ducks' meat and fatty liver performances simultaneously estimated in both parental lines. IX European Symposium on the Quality of Poultry Meat, Turku, Finland, 2009.

[B24] I. Aguilar, I. Misztal, and A. Legarra. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree and genomic information. ADSA and ASAS joint meeting, Montreal, Canada, 2009.

[B25] I. Misztal, A. Legarra, and I. Aguilar. 2009. Genetic evaluation including phenotypic, full pedigree and genomic information. ADSA and ASAS joint meeting, Montreal, Canada, 2009.

[B26] C. Colombani, A. Legarra, P. Croiseau, F. Guillaume, S. Fritz, V. Ducrocq, C. Robert-Granie. 2010. Application of PLS and Sparse PLS regression in Genomic Selection. 9th World Congress on Genetics applied to Livestock Production, Leipzig, Germany, August 2010 CD-ROM Comm. 0439.

[B27] L. Tusell Palomero, A. Legarra Albizu, M. García-Tomás, O. Rafel Guarro, J. Ramón Riba, M. Piles Rovira. 2010. Different ways to model the biological relationship between fertility and pH of the semen in rabbits. 9th World Congress on Genetics applied to Livestock Production, Leipzig, Germany, August 2010 CD-ROM Comm. 0569.

[B28] I. Misztal, I. Aguilar, A. Legarra, D. Johnson, S. Tsuruta, T. Lawlor. A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation. 2010. 9th World Congress on Genetics applied to Livestock Production, Leipzig, Germany, August 2010 CD-ROM Comm. 0050.

[B29] A. Legarra, C. Robert-Granié, P. Croiseau, F. Guillaume, S. Fritz, V. Ducrocq. Aptitude of Bayesian Lasso for genomic selection. 2010. 9th World Congress on Genetics applied to Livestock Production, Leipzig, Germany, August 2010 CD-ROM Comm. 0118.

[B30] Z. Vitezica, I. Aguilar, A. Legarra. 2010. One-step vs. multi-step methods for genomic prediction in presence of selection. 9th World Congress on Genetics applied to Livestock Production, Leipzig, Germany, August 2010 CD-ROM Comm. 0131.

[B31] A. Favier, J.-M. Elsen, A. Legarra, S. de Givry. 2010. Exact haplotype reconstruction in half-sib families with dense marker maps. 9th World Congress on Genetics applied to Livestock Production, Leipzig, Germany, August 2010 CD-ROM Comm. 0260.

[B32] A. Favier, J-M. Elsen, S. de Givry, and A. Legarra. 2010. Optimal haplotype reconstruction in half-sib families. In: ICLP-10 workshop on Constraint Based Methods for Bioinformatics, Edinburgh, UK, 2010

[B33] P. Croiseau, C. Colombani, A. Legarra, F. Guillaume, S. Fritz, A. Baur, R. Dassonneville, C. Patry, C. Robert-Granié, V. Ducrocq. 2010. Improving genomic evaluation strategies in dairy cattle through SNP pre-selection. 9th World Congress on Genetics applied to Livestock Production, Leipzig, Germany, August 2010 CD-ROM Comm. 0360.

[B34] G. Sallé, P. Jacquiet, L. Gruner, J. Cortet, C. Sauvé, F. Prevot, D. François, F. Bouvier, C. Pery, J. Bouix, A. Legarra, R. Rupp, C. Moreno. 2010. Preliminary results of a QTL detection study for resistance to Haemonchus contortus in sheep using the ovineSNP50 Beadchip. 9th World Congress on Genetics applied to Livestock Production, Leipzig, Germany, August 2010 CD-ROM Comm. 0463.

[B35] G.Salle, P.Jacquiet, L.Gruner, J.Cortet, C.Sauve, F.Prevot, F.Bouvier, D.Francois, C.Pery, J.Bouix, A.Legarra, R.Rupp, and C.R.Moreno. 2010. Preliminary results of a QTL detection study for resistance to Haemonchus contortus in sheep using the ovineSNP50 Beadchip. AGAH-2010. Animal Genomics for Animal Health International Symposium. 31 May - 2 June 2010 - "Maison de la Chimie", Paris (France).

[B36] S. Tsuruta, I. Aguilar, I. Misztal, A. Legarra, T. Lawlor. 2010. Multiple trait genetic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. 9th World Congress on Genetics applied to Livestock Production, Leipzig, Germany, August 2010 CD-ROM Comm. 0489_PP2-150.

[B37] I. Aguilar, I. Misztal, A. Legarra, S. Tsuruta. 2010. Efficient computations of genomic relationship matrix and other matrices used in the single-step evaluation. 9th World Congress on Genetics applied to Livestock Production, Leipzig, Germany, August 2010 CD-ROM Comm. 0768.

[B38] S. Tsuruta, I. Aguilar, I. Misztal, A. Legarra, and T. J. Lawlor. 2010. Multiple trait genetic evaluation of linear type traits using genomic and phenotypic information in US Holsteins. 2010 ASAS meeting, July11-15, Colorado.

[B39] I. Misztal, I. Aguilar, A. Legarra, and T. J. Lawlor. 2010. Choice of parameters for single-step genomic evaluation for type. 2010 ASAS meeting, July11-15, Colorado.

[B40] C. Moreno, J.M. Elsen, A. Legarra, R. Rupp, J. Bouix, F. Barillet, I. Palhière, H. Larroque, D. Allain, D. François, C. Robert, G. Tosser-Klopp, L. Bodin, P. Mulsant. 2010. Searching for genes of interest in sheep. Book of abstracts, p. 116. EAAP 2010, Heraklion, Greece.

[B41] I. Misztal, I. Aguilar, A. Legarra, and T. J. Lawlor. 2010. Choice of parameters for single-step genomic evaluation for type. . Book of abstracts, p. 357. EAAP 2010, Heraklion, Greece.

[B42] I. Misztal, C. Y. Chen, I. Aguilar, Z. G. Vitezica, A. Legarra, and W. M. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale.Book of abstracts joint ADSA-ASAS meeting, J. Anim. Sci. Vol. 89, E-Suppl. 1/J. Dairy Sci. Vol. 94, E-Suppl. 1, p.163.

[B43] G.Sallé, C.R.Moreno, L. Gruner, J. Cortet, C. Sauvé, F. Prévot, F. Bouvier, D.François, C.Pery, J.Bouix, A.Legarra, R.Rupp and P.Jacquiet. Genomic Mapping of Resistance to Haemonchus contortus Using the High-density OvineSNP50 Beadchip. 23rd. International Conference of the World Association for the Advancement of Veterinary Parasitology in Buenos Aires, 21-25 August, 2011

[B44] Misztal I., Tsuruta S., Aguilar I., Legarra A., Lawlor T. J., 2011. Approximation of genomic accuracies in single-step genomic evaluation. Interbull Meeting, Stavanger, Norvège, 27-28 Aout 2011

[B45] Legarra, A., Misztal, I. and Aguilar, I. 2011. The single step: genomic evaluation for all. Invited talk. Book of abstracts of the EAAP 2011, p. 1.

[B46] Croiseau, P., Hozé, C., Fritz, S., Guillaume, F., Colombani, C., Legarra, A., Baur, A., Robert-Granié, C., Boichard, D. and Ducrocq, V. 2011. Description of the French genomic evaluation approach. Book of abstracts of the EAAP 2011, p. 3.

[B47] Colombani, C., Legarra, A., Croiseau, P., Fritz, S., Guillaume, F., Ducrocq, V. and Robert-Granié, C. 2011. Bayes Cpi versus GBLUP, PLS regression, sparse PLS and elastic net methods for genomic selection in french dairy cattle. Book of abstracts of the EAAP 2011, p. 7.

[B48] Casellas, J., Esquivelzeta, C. and Legarra, A. 2011. Genomic BLUP with additive mutational effects. Book of abstracts of the EAAP 2011, p. 8.

[B49] Vitezica, Z.G. and Legarra, A. 2011. How to remove bias in genomic predictions? Book of abstracts of the EAAP 2011, p. 31.

[B50] Toro, M.A., Garcia-Cortes, L.A. and Legarra, A. 2011. Rationale for estimating genealogical coancestry from molecular markers. Book of abstracts of the EAAP 2011, p. 175.

[B51] Beltran De Heredia, I., Ugarte, E., Aguerre, X., Soulas, C., Arrese, F., Mintegi, L., Astruc, J.M., Maeztu, F., Lasarte, M., Legarra, A. and Barillet, F. 2011. Genomia: across-Pyrenees genomic selection for dairy sheep Book of abstracts of the EAAP 2011, p. 181.

[B52] Baloche, G., Larroque, H., Astruc, J.M., Babilliot, J.M., Boscher, M.Y., Boulenc, P., Chantry-Darmon, C., De Boissieu, C., Frégeat, G., Giral-Viala, B., Guibert, P., Lagriffoul, G., Moreno, C., Panis, P., Robert-Granié, C., Salle, G., Legarra, A. and Barillet, F. 2011. Work in progress on genomic evaluation using GBLUP in French Lacaune dairy sheep breed. Book of abstracts of the EAAP 2011, p.345.

[B53] Robert-Granie, C., Duchemin, S., Larroque, H., Baloche, G., Barillet, F., Moreno, C., Legarra, A. and Manfredi, E. 2011 A comparison of various methods for the computation of genomic breeding values in French Lacaune dairy sheep breed. Book of abstracts of the EAAP 2011, p. 346.

[B54] Misztal, I., Wang, H., Aguilar, I., Legarra, A., Muir, W. Tools for Genomic Analyses Using Single-Step Methodology. XX Plant and Animal Genome (PAG), San Diego, CA, 15/1/2012.

En Espagnol

[B55] A. Legarra , J. Arranz, I. Beltrán de Heredia and E. Ugarte. Sistema de calificación de la morfología mamaria en ovejas de Raza Latxa: resultados preliminares. VIII Jornadas de Producción Animal, Zaragoza, Spain, May 11-13, 1999. *ITEA* Vol. Extra 20, 345-347.

[B56] A. Legarra, E. Ugarte and E. Ruiz de Zárate: Adecuación de la estructura de datos del caballo de carne en Alava. X Reunión de mejora genética animal, Caldes de Montbui, Barcelona, Spain, June 8-9 2000. *ITEA* 96A:311-317.

[B57] A. Legarra, E. Ugarte, I. Beltrán de Heredia and J. Arranz. Parámetros genéticos y respuestas a diferentes índices de selección de caracteres de morfología mamaria en la raza Latxa. IX Jornadas de Producción Animal, Zaragoza, Spain, April 25-27, 2001. *ITEA* Vol. Extra 22, 24-26.

[B58] J. Arranz, A. Legarra. Importancia de la metodología de la toma de muestra en la composición de la leche de oveja. IX Jornadas de Producción Animal, Zaragoza, Spain, April 25-27, 2001. *ITEA* Vol. Extra 22, 649-651.

[B59] A. Legarra, E. Ugarte, I. Beltrán de Heredia, 2004. Análisis de asociación entre genotipos de PrP y producción de leche en la raza Latxa. XII Reunión nacional de mejora genética animal. Arucas, Las Palmas, Canarias. ITEA, 100A, 127-133

[B60] E. López de Maturana, A. Legarra, E. Ugarte. Estudio de la relación genética entre los caracteres facilidad de parto materna y fertilidad de la hembra en el ganado vacuno lechero de la CAPV. XII Reunión nacional de mejora genética animal. Arucas, Las Palmas, Canarias. ITEA 100A, 156-166

[B61] M. Ramón, A. Legarra, M.D. Pérez-Guzmán, E. Ugarte. Análisis técnico-económico de ganaderías de raza Latxa y Manchega como paso previo al cálculo de pesos económicos. (Poster). XI Jornadas sobre Producción Animal, Zaragoza, Spain, May 11-12, 2005. *ITEA* Vol. Extra 26, 90-92.

[B62] E. López de Maturana, A. Legarra, E. Ugarte. Estudio integral del carácter facilidad de parto en ganado vacuno Frisón de la CAPV. XI Jornadas sobre Producción Animal, Zaragoza, Spain, May 11-12, 2005. *ITEA* Vol. Extra 26, 123-125.

[B63] M. Ramón, A. Legarra, M.D. Pérez-Guzmán, E. Ugarte. Obtención de pesos económicos para selección por rentabilidad. XI Jornadas sobre Producción Animal, Zaragoza, Spain, May 11-12, 2005. *ITEA* Vol. Extra 26, 129-131.

[B64] A. Legarra, M. Ramón, E. Ugarte, M.D. Pérez-Guzmán. Pesos económicos en ovino lechero en razas Latxa y Manchega. XI Jornadas sobre Producción Animal, Zaragoza, Spain, May 11-12, 2005. *ITEA* Vol. Extra 26, 132-134.

[B65] A. Legarra, C. Robert-Granié, P. Croiseau, F. Guillaume y S. Fritz. 2010. Lasso Bayesiano Mejorado para selección genómica. XV Reunión Nacional de Mejora Genética, Vigo, Espagne, Junio 2010.

[B66] A Legarra. 2010. Mejora animal mediante selección genómica: nuevos aires para el parecido entre parientes. XVIII Seminario de Genética de poblaciones y evolución. Guitiriz, Espagne, Mai 2010.

[B67] Toro, M.A., García-Cortés, L.A., Legarra, A., Estimación del parentesco molecular mediante marcadores moleculares. XXXVIII Congreso de la Sociedad Española de Genética, Murcia, 2011.

En Français

[B68] Marie-Etancelin C., André J.M., Baéza E., Basso B., Bastianelli D., Bernadet M.D., Brun J.M, Davail S., Dubos F., Fernandez X., Guémené D., Gontier K., Guy G., Legarra A. Parametres génétiques d'indicateurs du métabolisme hépatique durant le gavage, de la qualité des produits et du taux de corticostérone chez le canard, estimes dans le cadre du programme « GENECAN ». To be presented at the « Journées de la recherche avicole », Octobre 2008.

[B69] Guillaume F., Fritz S., Croiseau P., Legarra A., Robert-Granié C., Colombani C., Patry C., Boichard D., Ducrocq V., 2009. Modèles d'évaluation génomique : Application aux populations bovines laitières françaises. Rencontres Recherches Ruminants, Paris, France, 2-3 Décembre 2009, 1-8.

[B70] Moreno C., Kopp C., Mulsant P., Robert-Granié C., Rupp R., Barillet F., Delmas C., Bouix J., Allain D., Manfredi E., Bodin L., Elsen J. M., Robelin D., Mangin B., Faraut T., Servin B., Jacquiet P., Foucras G., Legarra A., 2009. Utilisation d'une puce 60 000 SNP pour cartographier finement des QTL affectant des caractères de production, de résistance aux

maladies et de comportement chez les ovins. 16ème Rencontres Recherches Ruminants, Paris, France, 2-3 décembre 2009, Abstract, 1 p.

[B71] A. Ricard, A. Legarra, S. Danvy, C. Guyon, J.C. Meriaux, G. Guèrin. Peut-on prédire la qualité d'un reproducteur équin pour le CSO à partir de la génomique ? 37ème Journée de la Recherche Equine, 24 février 2011.

[B72] A. Ricard, A. Legarra, J.C. Meriaux, S. Danvy, G. Guerin. Résultats mitigés de l'évaluation génomique chez les chevaux de concours hippique. 38ème Journée de la Recherche Equine, 2012.

CURRICULUM VITAE

31 Juillet 2012

Andrés LEGARRA ALBIZU

Né à Pamplona (Navarra, Espagne) le 10 Mai, 1972.

Domicile : 11 Jardins Occitans, Ramonville St Agne (Haute Garonne, France), 31520.

Travail : INRA UR631(SAGA), BP 52627, 31326 Castanet Tolosan CEDEX, France. e-mail: andres.legarra@toulouse.inra.fr. Téléphone : 05 61 28 51 82. Fax: 05 61 28 53 53.

Titres

Ingenieur Agronome, Universidad Pública de Navarra, Pamplona, Espagne, 1997.

Docteur, Universidad Publica de Navarra, Pamplona, Espagne, 2002. Travail de thèse supervisé par Eva Ugarte: “Optimización del esquema de mejora de la raza Latxa: análisis del modelo de valoración e introducción de nuevos caracteres en el objetivo de selección”.

Situation professionnelle

Postdoc, University of Georgia, 2002-2003

Chercheur, Neiker, Vitoria, Espagne, 2003-2005

Chercheur, INRA, 2005-

Encadrement

Aurélie Favier, thésarde co-encadré avec Simon De Givry, 2011-2009. Thèse : « Décompositions fonctionnelles et structurelles dans les modèles graphiques probabilistes appliquées à la reconstruction d’haplotypes », soutenue le 3 décembre 2011.

Enseignement

Cours “Genome-wide association mapping and genomic selection”, Universidad de Buenos Aires, 3-7 Mars 2009.

Cours d’une journée, “Genetic improvement in Animal Science”, Master EURAMA, ESAP, Toulouse, 17/9/2008; 24/9/2009. 15/9/2010.

Cours de 3 jours “Genomic selection”, Universidad Autonoma de Barcelona, 29-30 et 1 Octobre 2010; University of Georgia, Athens, 28-1 Juin 2012. ; INIA, Madrid, 11-13 Juin 2012

6 h de cours “Génétique quantitative ”, Master MABS : « Microbiologie, Agrobiosciences, Bioinformatique et biologie des Systèmes, UE : Génomique et Génétique Statistique », Université Paul Sabatier (Toulouse, France). Saison 2011-2012.

Projets financés

Très impliqué:

INIA (Espagne) : RTA02-002-C2. “Economic weights in Dairy sheep genetic improvement programs”. Budget total: 20000 euros

ANR (France) : SheepSNPQTL. “ Utilisation d'une puce 60 000 SNP pour cartographier finement des QTL affectant des caractères de production, de resistance aux maladies de comportement chez les ovins » Budget total: 300000 euros.

ANR (France) : Rules&Tools. “Statistical methods for the dissection of trait variability with SNP chips”. Budget total: 400000 euros.

ANR (France) : Amasgen. “Methodological approaches and applications of geneomic selection in dairy cattle” Budget total: 300000 euros.

Correspondant INRA:

POCTEFA (European FEDER Funds): GENOMIA. “Renforcement des schémas de sélection des races ovines laitières locales d’intérêt économique, écologique et social. » Budget total: 650000 euros.

Autres

Membre du comité editorial de: Journal of Animal Science 2005-2007 ; Genetics, Selection, Evolution 2012-. Reviewer: Plos ONE; Journal of Animal Science; Journal of Dairy Science; Genetics, Selection, Evolution; Animal; Heredity; Genetics; World Rabbit Science; Canadian Journal of Animal Science; Crop Science; Información Técnico-Económica Agraria (ITEA), Spanish Journal of Agricultural Research. Reviewer for the 8th World Congress on Genetics Applied to Livestock Production.

Organiseur du XI QTLMAS 2007. Workshop on QTL and Marker Assisted Selection. 22-23 March 2007. Toulouse, France.

Professeur “libre” (*docente libre*) Universidad de Buenos Aires, Argentina.

Rapporteur et/ou membre du jury de plusieurs thèses en Espagne: Evangelina López de Maturana, Manuel Ramón Fernández, Yang Bin, Llibertat Tusell.

Communications invitées :

-Sélection génomique: Séminaire des départements GA et GAP, INRA, 18/9/2007

- Sélection génomique: Séminaire du “réseau avicole” de l’INRA, 6/12/2007

- Sélection génomique: Universidad de Buenos Aires, 13/2/2007

-Performance of genomic selection in mice: Iowa State University, 3/4/2008

-Practical application of Genomic selection: September 7-8 2008 in Salzburg, Austria, University of Natural Resources and Applied Life Sciences.

-One-step genetic evaluation including pedigree, genomic, and phenotypic data. Universidad de Zaragoza, Espagne, 10/11/2009.

-Genomic selection (‘Mejora animal mediante selección genómica: nuevos aires para el parecido entre parientes’). XVIII Seminario de Genética de Poblaciones y Evolución. Guitiriz, Espagne, 5-7 Mai 2010.

-Improved Bayesian Lasso for genomic selection (‘Lasso Bayesiano Mejorado para selección genómica’). XIV Reunión nacional de mejora genética animal. Vigo, Espagne, 16-18 Juin 2010.

-The Single Step: genomic selection for all. Meeting of the European Association of Animal Production, 28 Août 2011, Stavanger, Norway.

-Genomic relationships (“El parentesco genómico: ese desconocido”). Universidad de Buenos Aires, 2 Mars 2011

- Improved Bayesian Lasso for genomic selection (‘Lasso Bayesiano Mejorado para selección genómica’). Universidad de Buenos Aires, 2 Mars 2011