Méthodes multiblocs pour l'identification de gènes associés au vieillissement cutané

Anne Bernard, Arthur Tenenhaus, Jean-François Zagury, Christiane Guinot, Gilbert Saporta

Méthodes multiblocs pour l'identification de gènes associés au vieillissement cutané

Anne Bernard [1,2], Arthur Tenenhaus [3], Jean-François Zagury [4], Christiane Guinot [1,5], Gilbert Saporta [2]

[1] *CERIES, 20 rue Victor Noir, Neuilly sur Seine, France, anne.bernard@ceries-lab.com, christiane.guinot@ceries-lab.com*

[2] *CNAM, laboratoire CEDRIC, 292 rue Saint-Martin, Paris, France, gilbert.saporta@cnam.fr*

[3] *SUPELEC, 3, rue Joliot-Curie, Gif-sur-Yvette, France, arthur.tenenhaus@supelec.fr*

[4] *CNAM, laboratoire GBA, 292 rue Saint-Martin, Paris, France, zagury@cnam.fr*

[5] *Université Francois Rabelais, département d'informatique, 64 avenue Jean Portalis, Tours, France ;*

Résumé. Une étude a été conduite sur 502 femmes afin d'identifier des gènes ayant un impact sur le vieillissement cutané. Des données générales et phénotypiques ont été recueillies ainsi qu'un échantillon de sang pour les analyses génétiques. La sévérité du photovieillissement et de 12 signes de vieillissement cutané ont été appréciés à partir de photographies numériques du visage grâce à des échelles photographiques ordinales. Trois scores ont ensuite été calculés à partir de ces signes de vieillissement. Les SNPs (Single Nucleotide Polymorphisms) génotypés ont permis de répertorier 15198 gènes pour chacune des 502 femmes. L'analyse des liens entre ces gènes, les scores et le photovieillissement a été réalisée à l'aide d'une méthode multiblocs associée à une méthode de rééchantillonnage (bootstrap) afin de sélectionner les SNPs les plus significatifs au sein de chaque bloc. Les blocs de gènes contenant le nombre de mutations alléliques par SNPs sont les variables explicatives, et les scores et le photovieillissement sont les variables à expliquer.

Mots-clés. Bootstrap, génétique, méthodes multiblocs, méthodes sparse, photovieillissement.

Abstract. A study was conducted on 502 women to identify genes affecting skin aging. General and phenotypic data were collected and a blood sample was taken for genetic analysis. The severity of photoaging and 12 signs of skin aging were evaluated from digital photographs of the face with photographic ordinal scales. Three scores were then calculated from these signs of aging. Genotyped SNPs (Single Nucleotide Polymorphisms) allowed identifying 15.198 genes for each of the 502 women. The analysis of the relationship between these genes, photoaging and the scores was performed using a multiblock method associated with a resampling method (bootstrap) to select significative SNPs within each bloc. Blocks of genes containing the number of allelic mutations by SNPs are the independent variables, and scores and photoaging are the variables to be explained.

Keywords. Bootstrap, genetic, multiblocs methods, sparse methods, photoaging.

1 Introduction

A ce jour, aucune étude d'association génomique n'a spécifiquement recherché de liens avec le vieillissement cutané. En 2010, un projet de recherche GWAS (Genome Wide Association Study) a été mis en place pour étudier l'impact du patrimoine génétique sur les caractérisques de la peau et l'expression du vieillissement cutané (phénotypes). L'objectif de cette étude est d'analyser les possibles liens entre les génotypes et les indicateurs du vieillissement cutané (Elfakir et al., 2010). Cette étude nécessite le développement de méthodes statistiques pour le traitement de données à grande dimension. On rapporte ici les premières analyses réalisées à l'aide de la méthode multiblocs RGCCA (Regularized Generalized Canonical Correlation Analysis).

2 Matériel

Données collectées L'étude a été conduite sur 502 femmes caucasiennes de la cohorte SU.VI.MAX (Hercberg et al., 1998) vivant en Ile de France et âgées de 44 à 70 ans. Des informations générales, phénotypiques et médicales sur les femmes ont été recueillies au cours d'un interrogatoire médical standardisé ainsi qu'un auto-questionnaire des habitudes tabagiques et d'exposition solaire, questionnaire qui a permis de construire un score basé sur 5 items pour estimer l'intensité d'exposition au soleil sur la vie de chaque individu (Guinot et al., 2001). Les données utilisées comme covariables sont les suivantes : l'âge (en années), l'indice de masse corporelle (kg/m*m), l'intensité d'exposition au soleil sur la vie (score), le statut hormonal (non ménopausée/ ménopausée avec Traitement Hormonal de Substitution (THS)/ ménopausée sans THS) et le statut tabagique (ancien fumeur/ non fumeur/ fumeur actuel).

Phénotypes analysés Trois photographies numériques " haute-définition " du visage des femmes (une de face et deux de profil 3/4) ont été prises dans des conditions standardisées (Kodak DCS 660, 6.0 MP). Ces photographies ont ensuite été examinées par un dermatologue afin d'évaluer la sévérité du photovieillissement global (Larnier et al., 1994), et la sévérité d'une série de signes de vieillissement à l'aide d'échelles ordinales spécifiques avec illustrations photographiques (Morizot et al, 2002). Ces indicateurs ont été utilisés comme critères de jugement dans les analyses :

- photovieillissement global (grade 1-6, figure 1) ;
- score de lentigines (2 items, 0-10) ;
- score de rides (6 items, 0-10) ;
- score de relâchement (4 items, 0-10).

Ces critères se sont révélés être très liés aux 5 covariables citées précédemment. De ce fait, une régression de ces variables sur les 5 covariables a été réalisée afin de pouvoir travailler sur les résidus et non sur les critères pour la suite des analyses.

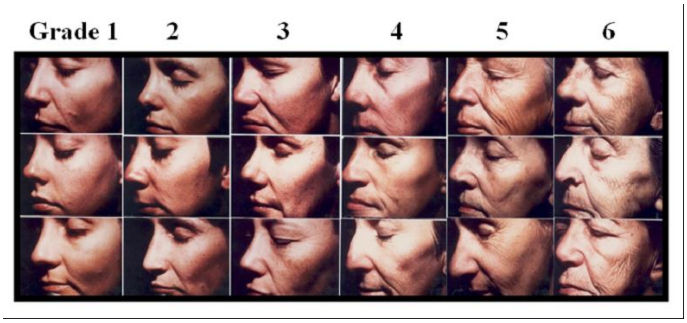

FIGURE 1 – Echelle de photovieillissement global (d'après Larnier et al., 1994)

Données génétiques Des analyses génétiques ont été réalisées à partir d'un prélèvement sanguin de chaque femme. L'ADN extrait a été analysé grâce à une puce à haut-débit Illumina Human Omni1-Quad contenant 1140000 marqueurs génétiques (SNPs : Single Nucleotid Polymorphism). Parmi les marqueurs, 91000 étaient des variants du nombre de copies, 118000 n'ont montré aucune variation, 55000 présentaient des erreurs de génotypage et 2000 étaient situés sur le chromosome Y. Après contrôle de qualité, 795063 SNPs ont été retenus dont 362223 étaient situés dans un total de 15198 gènes. Le regroupement de SNPs par gène a permis de créer 15198 blocs de SNPs (de taille différente) correspondant aux 15198 gènes répertoriés dans notre base de données. Dans les premières analyses, 13 gènes déjà étudiés en approche gène-candidat ont été réanalysés.

3 Méthode

Les liens entre les génotypes et le vieillissement cutané ont été modélisés grâce à la méthode RGCCA (Tenenhaus et Tenenhaus, 2011). Son objectif est de trouver des combinaisons linéaires entre blocs de variables telles que les composantes expliquent correctement leur propre bloc et que les composantes des blocs connectés soient fortement corrélées. Les blocs explicatifs sont les 15198 blocs de SNPs $X_1, ..., X_{15198}$, et X_{15199} (vieillissement cutané) est le bloc à expliquer (figure 2).

Dans chacun des blocs, le nombre de SNPs varie entre 1 et 10000 SNPs. Les composantes à optimiser sont les $y_j = X_j a_j$ (a_j poids externe). Le problème d'optimisation est le suivant :

$$\max_{a_1,...,a_j} \sum_{j,k=1,j\neq k}^{J} c_{jk} \mid cov(X_j a_j, X_k a_k) \mid$$
$$\text{s.c. } \tau_j \|a_j\|^2 + (1-\tau_j) Var(X_j a_j) = 1, j = 1, ..., J$$

avec $y_j = X_j a_j$, $C = \{c_{jk}\}$ matrice telle que $c_{jk} = 1$ si X_j et X_k sont liés, 0 sinon et τ, paramètre de régularisation fixé à 1 dans notre exemple (mode A).

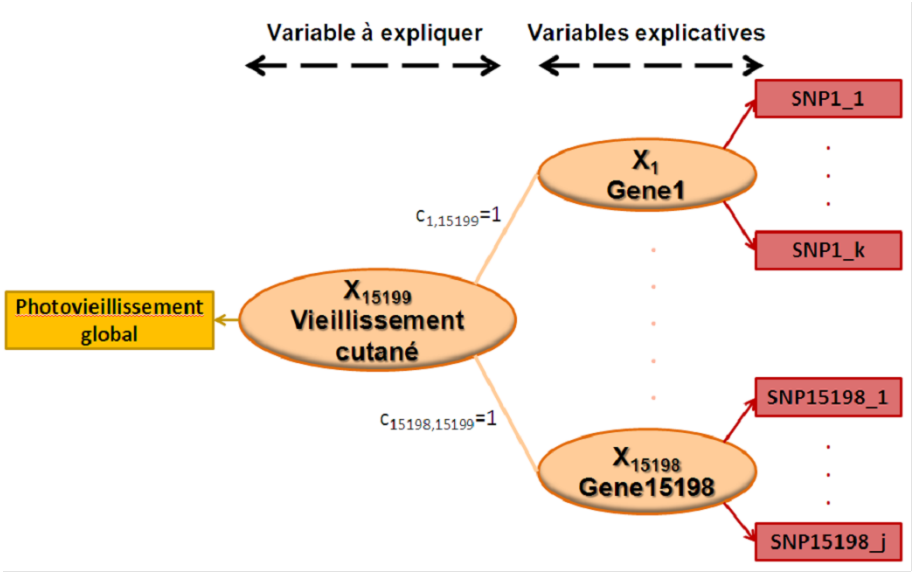

FIGURE 2 – Modèle des liens possibles entre gènes et vieillissement cutané avec remplacement successif du photovieillissement par les différents scores (jaune et rouge : modèle externe liant variables observées à variables latentes, beige : modèle interne liant les variables latentes entre elles)

Algorithme avec l'hypothèse $\tau = 1$:
1. Initialisation des a_j
2. Calcul des estimations externes $y_j = X_j a_j$ sous la contrainte $||a_j||^2 = 1$
3. Calcul des estimations internes $z_j = \sum_{j \neq k} e_{jk} y_k$ (expliquant la relation entre les blocs)
4. Calcul des nouveaux $a_j = \frac{(X_j^t z_j)}{\|X_j^t z_j\|}$
5. Itérer jusqu'à convergence

Sélection de SNPs par bloc Afin de sélectionner dans chaque bloc les SNPs ayant un lien significatif avec la variable réponse, la méthode de rééchantillonnage bootstrap est utilisée (Efron et Tibshirani, 1993). Le principe du rééchantillonnage bootstrap est de tirer au hasard avec remise des observations dans l'échantillon dont on dispose. La distribution de la population est inconnue ou non-normale et le bootstrap permet de construire une distribution proche de la distribution inconnue en calculant variance et intervalle de confiance. Les différentes étapes à réaliser pour la sélection de SNPs par bloc sont les suivantes. On considère le premier bloc de SNPs correspondant au gène 1. Soit $X = (SNP1_1, ..., SNP1_k)$ les variables explicatives du bloc 1 et Y la variable à expliquer (vieillissement cutané). On considère B réplications.

Étape 1 Réaliser la RGCCA sur l'échantillon initial. Conserver les poids externes $a_1, ..., a_k$ associés à chacune des variables explicatives observées ($SNP1_1$, ..., $SNP1_k$).

Étape 2 Répéter pour $b = 1, ..., B$.
1. Tirer un échantillon aléatoire de taille N avec remise, noté $Z^{*b} = (X^{*b}, Y^{*b})$.
2. Réaliser la RGCCA sur le nouvel échantillon et calculer le nouveau poids externe a_j^{*b} associé à chaque variable.

Étape 3 Répéter pour $j = 1, ..., k$.
1. $E_j = (a_j^{*1}, ..., a_j^{*B})$ est l'échantillon bootstrap de taille B du poids externe a_j (associé à X_j).
2. Calculer l'intervalle de confiance I_j^* pour a_j.
3. Si $0 \in I_j^*$, éliminer la variable X_j.

Les bornes inférieure et supérieure $a_j^{*(inf)}$ et $a_j^{*(sup)}$ de l'intervalle I_j^* sont calculées, pour un seuil de confiance α fixé, de la façon suivante :
$a_j^{*(inf)} = 100 \cdot \alpha^{ième}$ percentile et $a_j^{*(sup)} = 100 \cdot (1-\alpha)^{ième}$ percentile obtenus à partir de E_j.

	Mode A
AVE (Gene1)	0,83
AVE (Gene2)	0,47
AVE (Gene3)	0,28
AVE (Gene4)	0,29
AVE (Gene5)	0,55
AVE (Gene6)	0,34
AVE (Gene7)	0,28
AVE (Gene8)	0,10
AVE (Gene9)	0,20
AVE (Gene10)	0,14
AVE (Gene11)	0,03
AVE (Gene12)	0,13
AVE (Gene13)	0,10
AVE (Rides)	1,00
AVE (modèle externe)	0,10
AVE (modèle interne)	0,02

TABLE 1 – Indices de qualité du modèle

4 Résultats - Conclusion

Les analyses ont été réalisées sur 13 gènes ayant déjà été étudiés avec une approche gène-candidat. Les composantes expliquent correctement leur propre bloc (ex : table 1 AVE(Gene1) = 0,83), cependant la qualité du modèle interne est basse (AVE(inner model) = 0,02). L'aspect unidimensionnel de la RGCCA ne permet pas d'obtenir une bonne qualité de modèle, il faut donc l'étendre au multidimensionnel. La solution est donc de chercher à construire pour chaque bloc plus d'une composante externe (déflation). Le nombre de composantes nécessaires a été calculé par validation croisée (régression PLS). Le rééchantillonnage par bootstrap est en cours et permettra de sélectionner les SNPs les plus significativement liés à la variable à expliquer. Les interactions éventuelles entre gènes et entre SNPs n'ont pas encore été prises en compte, il faudrait donc par la suite ajouter des liens entre blocs (gènes) et généraliser la méthode aux 15198 gènes. Par ailleurs, des méthodes sparse telle que le lasso (Tibshirani, 1996) pourront être associées à la méthode multibloc afin de sélectionner les SNPs les plus significatifs par bloc et les comparer à ceux sélectionnés avec la méthode bootstrap.

Bibliographie

[1] Elfakir, A. et al. (2010), Functional MC1R gene variants are associated with an increased risk for severe photoaging of facial skin, *J Invest Dermatol*, 130, 1107-15.
[2] Hercberg, S. et al. (2004), The SU.VI.MAX Study : a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals, *Arch Intern Med*, 164, 2335-42.
[3] Guinot, C. et al. (2001), Sun exposure behaviour of a general adult population in France, *Skin and Environment – Perception and Protection*, 10th European Academy of Dermatology and Venerology Congress, Munich, 2001. Bologne : Monduzzi editore, 2001, p. 1099-106.
[4] Larnier, C. et al. (1994), Evaluation of cutaneous photodamage using a photographic scale,*Br J Dermatol*, 130, 167-173.
[5] Morizot, F. et al. (2002), Development of photographic scales documenting features of skin ageing based on digital images, *Ann Dermatol Venereol*, XXème Congrès Mondial de Dermatologie, Paris, 1–5 juillet 2002, 129 :1s402 [Abstr.].
[6] Tenenhaus, A. et Tenenhaus, M. (2011), Regularized Generalized Canonical Correlation Analysis, *Psychometrika*, 76, 257-284.
[7] Efron, B. et Tibshirani, R. (1993), An introduction to the Bootstrap, *Chapman and Hall*, London.
[8] Tibshirani, R. (1996), Regression shrinkage and selection via the lasso,*J Roy Statist Soc*, 58, 267–288.