2e édition des Rencontres SHS et médecine génomique
29 et 30 juin 2023 - Dijon
FRANCE MÉDECINE GÉNOMIQUE 2025 aviesan

L'apport des noms de famille dans l'élaboration d'un échantillon représentatif d'une population

Pierre Darlu[1], Pascal Chareille[2]

[1] Eco-anthropologie (EA), UMR7206, Muséum national d'Histoire naturelle, CNRS, Université de Paris Cité, France

[2] Centre Tourangeau d'Histoire et d'étude des Sources (CETHIS), Université de Tours

Les noms de famille, transmis de génération en génération en lignée masculine dans nos sociétés, constituent un substitut aux marqueurs génétiques : ils sont caractérisés par une grande diversité (plus de 400000 « allèle-noms»), une profondeur généalogique (une vingtaine de générations depuis les XIII°-XV° siècles,) une spécificité géographique et un coût modéré de collecte.

Démographes (Bienaymé, Lotka, Brouard), historiens (Bourin), anthropologues (Levi-Strauss, Ségalen), mais aussi généticiens (Cavalli-Sforza, Crow, Sokal, Lasker, Barbujani, King, etc…) se sont approprié cet objet d'investigation.

Les noms apportent une information à plusieurs niveaux (**I, II, III**) dans les stratégies d'échantillonnage :

I - Les noms permettent de décrire la structure patronymique d'une population comme proxy de sa structure génétique (avec laquelle elle est corrélée). Ils constituent ainsi une aide pour effectuer un échantillonnage sur une base géographique cohérente

II – Le choix d'un échantillon est habituellement fondé sur l'insertion géographique supposée de ses 4 grands-parents. Un critère supplémentaire est de sélectionner les personnes dont le patronyme est dominant ou exclusif (géohapax) de la région d'intérêt. Ces noms ayant souvent un enracinement historique pouvant remonter au XIII^{e} ou XIV^{e} siècle permettent d'insérer l'échantillon dans une dimension historique plus profonde. Cette attribution à une aire géographique reste probabiliste. Des exemples de cette persistance régionale sur plusieurs siècles sont ici proposés

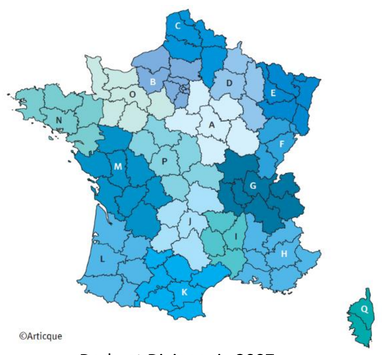
III - Lorsqu'un échantillon est préalablement constitué sans critères géographiques (échantillon de malades par exemple), les noms permettent d'attribuer une probabilité d'origine aux personnes concernées puis d'ajuster la distribution géographique de leurs données, génétiques, phénétiques ou autres, à un maillage géographique fin

I - Diversité patronymique en France établie à partir du fichier patronymique de l'INSEE.

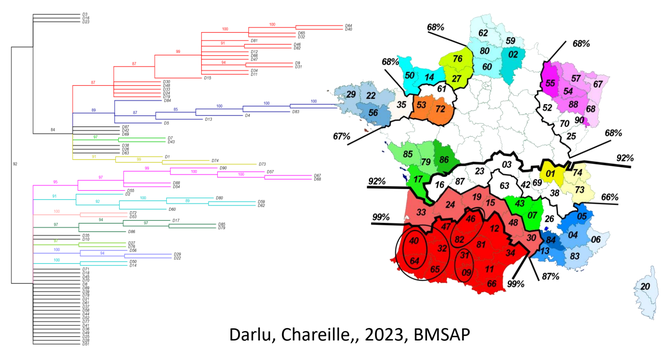
I a : Calcul d'une matrice de distances entre populations *i* et *j* à partir des fréquences de leurs *K* noms (ex: Fst ou Nei's Distance)

$$F_{ST,ij} = \frac{\sum_{k=1}^{K}\left(p_{ki} - p_{kj}\right)^{2}}{2 \times \sum_{k=1}^{K}\bar{p}_k \times (1-\bar{p}_k)} \; ; \; \phi_{ij} = \frac{\sum_{k=1}^{K} p_i p_j}{\left[\sum_{k=1}^{K} p_i^2 \times \sum_{k=1}^{K} p_j^2\right]^{1/2}} (Nei, 1973)$$

I b : Représentation spatiale à partir d'un algorithme de classification (Guénoche et al.) ou « NJ-Tree-projection » avec bootstrap



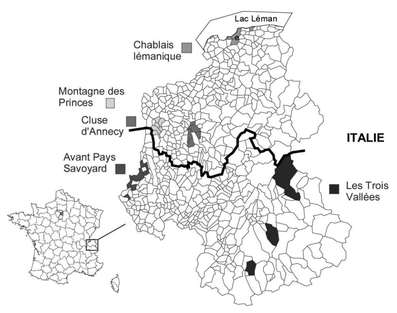
Darlu et Digioanni, 2007
Espace géographique

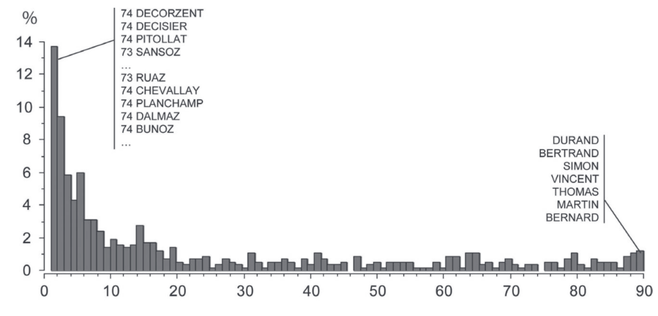


Darlu, Chareille,, 2023, BMSAP

***Conclusion** : Les clusters patronymiques fournissent un maillage géographique raisonné pour constituer des échantillons avec des assises historiques fiables. La taille des échantillons doivent être modulés en fonction de la robustesse des clusters (en termes de %BP par ex.)*

II -

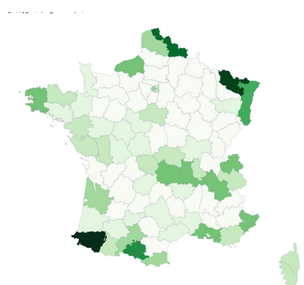




Pourcentage de patronymes, parmi 609 attestés dans le corpus de 5 régions de Savoie entre 1710 et 1940, qui sont présents dans un seul département (géohapax comme Deplanchin, Bunoz…= ou simultanément dans 2, 3, …90 départements. Il y a 14% de géohapax dans ce corpus

	INSEE 1891-1940	XVIII°	XV°
Touraine (dep. 37)	176	39	
Dijonnais (dep. 21)	171		40-50*
Calvados (dep. 14)	274		50-70*

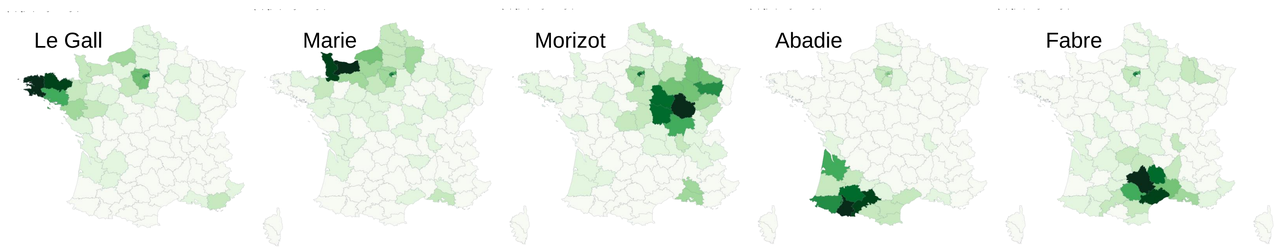
Nombre de géohapax (INSEE, N≥5) déjà attestés au XVIII° ou XV°
** selon le degré de lemmatisation*



Distribution départementale des géohapax représentés par au moins 10 naissances (1891-1915)

***Conclusion** : pour avoir une bonne représentativité à la fois géographique et avec profondeur généalogique, la stratégie consiste à échantillonner les personnes portant un géohapax local ou régional, ou un patronyme attesté par une longue présence historique*.

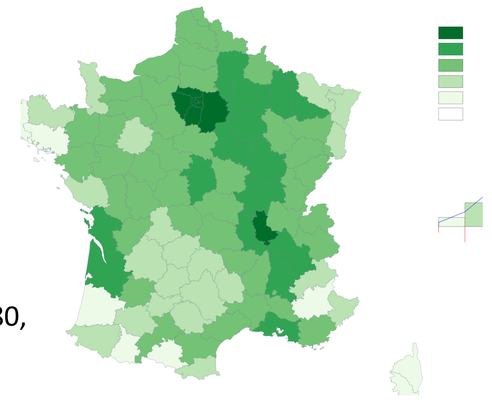
III - Un vecteur de fréquence géographique peut être attaché à chaque nom : exemple ;



Exemple de la représentation de la diversité génétique du HLA des donneurs de moelle établit par département à partir de leurs noms de famille.

$$p_{jk} = \frac{\sum_{i=1}^{s} \omega_{ijk} f_{ik}}{\sum_{j=1}^{r}\sum_{i=1}^{s} \omega_{ijk} f_{ik}} \qquad \omega_{ijk} = \frac{\delta_{ijk}\pi_{jk}}{\sum_{j=1}^{r} \delta_{ijk}\pi_{jk}}$$

EJHG (2003) 11, 794–80,



***Conclusion** : Les noms permettent de décrire, en probabilité, la distribution spatiale de phénotypes, génotypes, maladies, etc… et de les mettre en relation avec de possibles données environnementale, historique, épidémiologique liées à la géographie…*