LES ERREURS DANS LES GENEALOGIES ET LEURS INCIDENCES SUR L'ESTIMATION DE LA CONSANGUINITE

Pierre DARLU *

Différents types d'erreur peuvent se glisser dans la reconstruction de l'histoire généalogique d'une population : attributions erronées d'un père ou d'une mère à un enfant, erreurs ou confusions dans les registres d'état-civil ou de paroisse, mauvaises transcriptions lors des différentes étapes de la saisie... Dans ces conditions il n'est pas sans risque de considérer l'histoire généalogique ainsi reconstruite recouvre avec exactitude la véritable histoire biologique de la population.

Depuis Dahlberg (1928), Bernstein (1930), Sutter et Tabah (1951), et d'autres à leur suite, il est habituel de caractériser l'endogamie et l'isolement d'une population par un coefficient, le coefficient moyen de consanguinité. Lorsque ce coefficient est défini à partir des fréquences de mariages entre divers apparentés (cousins de degrés divers, oncle-nièce, etc.), on obtient le coefficient de consanguinité moyen "apparent". En revanche, lorsque ce coefficient est défini à partir

* Unité de recherches de Génétique épidémiologique, I.N.S.E.R.M. U 155 et R.C.P. "Différenciation anthropologique, démographique et génétique des populations humaines" (C.N.R.S.)

Ecologie Humaine, Vol 4, n°1
1986, pp.6372

de l'ensemble de l'information généalogique disponible sur cette population, on obtient le coefficient moyen de consanguinité "élargi". Ces coefficients ne sont en fait que des résumés de l'information plus ou moins importante dont on dispose sur les liens de parenté entre tous les individus de la population : information réduite dans le cas du coefficient "apparent", information plus complète dans le cas de la consanguinité élargie. En ce sens, la valeur de ce coefficient "apparent" ne peut qu'être inférieure au coefficient "élargi".

Il est en général difficile de vérifier la véracité des liens biologiques dans toute une généalogie. Lorsque des marqueurs génétiques sont analysés, il est possible de vérifier la validité biologique des relations familiales sur deux ou trois générations (Langaney, 1974, Lathrop et al., 1983). Pour des généalogies plus profondes, il faut utiliser des gènes ou des haplotypes rares et vérifier que leur ségrégation est compatible avec les relations familiales déclarées (Chaventré, 1976). Cependant, ces gènes étant rares, ils ne permettent en général pas de valider toute une généalogie, mais seulement quelques-uns de ses segments.

Il est donc important de pouvoir évaluer le biais introduit par l'hypothèse d'absence d'erreur, dans le calcul d'un coefficient moyen de consanguinité.

METHODE

Dans ce but, je propose une formulation modifiée du coefficient moyen de consanguinité. Elle s'inspire de la formulation donnée par Sutter et Tabah (1951) qui prenait en compte la possibilité de mutation et de migration.

$$\alpha_g = \frac{1}{\sum_c n_c} \cdot \sum_c n_c \sum_i \left(\frac{\delta i}{2}\right)^{n_1+n_2} \left(\frac{1+F_i}{2}\right)$$

avec $\delta i = 0$ lorsqu'une erreur survient sur la chaîne i
et $\delta i = 1$ autrement, nc étant le nombre d'enfants du couple c à la génération g-1, les sommations se faisant sur l'ensemble des couples de la génération g-1 et sur l'ensemble des différentes chaînes reliant entre eux les deux individus de chaque couple c, avec n1 commun dont la consanguinité est fj etn2 chaînes pour revenir de cet ancêtre à l'autre individu du couple.

Soient p la probabilité de non paternité,
m la probabilité de non maternité,
et r la probabilité d'autres erreurs.

La probabilité pour que $\delta i = 1$ peut s'écrire alors :
$P(\delta i=1) = (1-p)(1-r)$, pour une chaîne allant d'un homme à son enfant et
$P(\delta i=1) = (1-m)(1-r)$, pour une chaîne allant d'une femme à son enfant.

EXEMPLES

1) L'exemple du tableau 1 est tiré de la généalogie des habitants d'Arthez d'Asson en Béarn. Il montre l'ampleur du biais dans l'estimation du coefficient de parenté ϕij, entre deux individus I et J, sous différentes hypothèses concernant p, m et r.

Ainsi, l'erreur relative sur ϕij, Δ, peut atteindre 26 %, dans une situation pourtant courante où les exclusions de paternité surviendraient dans 4 % des naissances, les erreurs de transcriptions et de codages également dans 4 % et les erreurs de maternité, pouvant correspondre à des adoptions non déclarées, dans 1 % des cas...

2) Comme il est logique de supposer que les deux probabilités p et m sont différentes, l'erreur relative Δ sera évidemment dépendante du sexe des ancêtres de I et J. Par exemple, on peut imaginer deux types de cousins issus de germains, l'un où les ancêtres intermédiaires seraient tous des hommes, l'autre tous des femmes. Dans l'hypothèse où p=.04 et m=r=.01, l'erreur relative sur le coefficient de parenté entre les deux cousins $(\Delta=[\phi o-\phi ij)\phi o)$, peut atteindre, en moyenne, jusqu'à 24 % dans le premier type de cousins, au lieu de 14 % dans le deuxième type...

3) Un niveau d'un coefficient moyen de consanguinité apparente, on s'attend à des biais de même ampleur. Reprenant les observations de Barraï et al. (1962) sur les fréquences des diverses catégories de mariages consanguins dans les diocèses de Parma et de Piacenza entre 1851

et 1957, on peut évaluer l'erreur relative sur le coefficient moyen de consanguinité apparente, sous différentes hypothèses concernant m, p et r. Comme le montre le tableau 2, cette erreur est importante et d'autant plus élevée que l'on s'intéresse à des apparentés plus lointains. Ainsi, en supposant l'absence d'erreurs dans les relations de parenté, on est conduit à surestimer le coefficient moyen de consanguinité de 14,75 %, dans l'hypothèse port plausible où p=.04 et m=r=.01...

4) Ces résultats remettent en question l'intérêt des estimations des coefficients moyens de consanguinité, en particulier dans les petites populations isolées. En effet, dans de telles populations, on s'attend à ce que le coefficient moyenb de consanguinité soit élevé : ce coefficient tend vers 1 comme 1-exp(-g/2N) (Jacquard, 1975) et il s'en rapproche d'ailleurs d'autant plus que l'isolement est ancien, la population réduite mais aussi que les données généalogiques sont plus profondes et complètes.

Malheureusement, dans le même temps, la qualité informative de ce coefficient moyen de consanguinité diminue, en raison de l'augmentation rapide du biais de son estimation sous l'effet cumulatif des divers types d'erreur que nous avons évoqué plus haut. Paradoxalement, plus la quantité d'information généalogique que résume ce coefficient est importante, moins bonne en sera la qualité...

5) L'effectif efficace d'une population supposée panmictique est très souvent estimé à partir de la fréquence des mariages entre cousins, particulièrement les cousins issus de germains

(Dahlberg, 1948), ou à partir du coefficient moyen de consanguinité (Malecot, 1966). Quelque soit la méthode utilisée, si l'hypothèse d'absence d'erreur dans la reconstitution des pedigrees n'est pas fondée, on obtiendra une sous-estimation de la taille efficace de la population. Les isolats ne sont certainement pas aussi réduits que le laissent pense de telles méthodes d'estimation, et le biais de cette estimation sera d'autant plus grand que les diverses erreurs dans les attributions des enfants à leurs parents seront importantes...

CONCLUSIONS

Ces diverses observations soulignent la nécessité d'une évaluation précise des valeurs des probabilités p, m et r. Cela conditionne en effet la validité des études basées sur des comparaisons, dans le temps ou dans l'espace, de divers coefficients de consanguinité. La tâche est rendue d'autant plus difficile que de nombreux facteurs exercent des influences certaines sur ces probabilités : elles fluctuent en effet d'une génération à une autre, par exemple à cause de l'évolution des moeurs, d'une inégale attention portée à la tenue des registres d'une années sur l'autre, d'un changement dans les coutumes d'attribution des noms de famille. Elles peuvent également être différentes selon le degré d'apparentement que l'on considère : ainsi on peut imaginer que les exclusions de paternité sont plus rares - ou plus fréquentes ? - pour les couples de cousins que pour les couples non apparentés, ou que les erreurs de registres sont moins nombreuses pour les mariages entre apparentés qui font l'objet de plus d'attention... Enfin, ces probabilitésa varient certainement d'une population à l'autre.

Dans ces conditions, il est inconsistant de vouloir comparer entre eux les coefficients moyens de consanguinité apparente et de consanguinité élargie pour une même populationi, aussi bien que de vouloir comparer les coefficients de consanguinité apparente de deux populations différentes ou de deux générations successives, si l'on ne s'est pas assuré au préalable que les probabilités p, m et r y sont bien comparables...

BIBLIOGRAPHIE

Barraï I, Cavalli-Sforza LL, Moroni A, 1962. Frequencies of pedigree of consanguineous marriages and mating structure or the population. Ann. Hum. Génet. 25: 347-377.

Bernstein F, 1930. Fortgesetste Untersuchungen aus der theorie der blutgruppen Zeit. ind. Abst. Vererb. 56: 233-273.

Chaventré A, 1976. Les généalogies. Techniques de présentation et de vérifications. In : L'étude des Isolats, A. Jacquard Ed. I.N.E.D (Paris), pp 233-244.

Dahlberg G, 1928. Inbreeding in man. Genetics 14: 421-454.

Jacquard A, 1975. Inbreeding : one Word, Several Meanings Theor. Pop. Biol, 7: 338-363.

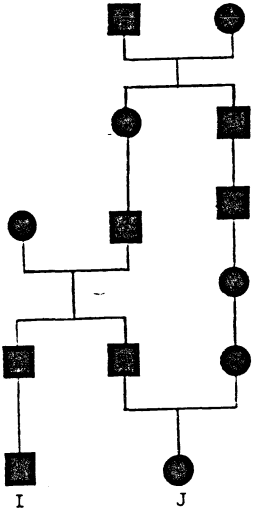
Langaney A, 1974. Computation of abnormal genetic inheritance estimation of parentage exclusion probabilities, legitimacy and adoption rates. In : Genealogical Mathematics. Ballanoff PA Ed (Paris). Maison des Sciences de l'Homme, pp. 127-138.

Lathrop MG, Hooper AB, Huntsman JW, Ward RH, 1983. Evaluating Pedigree Data. I. The estimation of Pedigree Error in the Presence of Marker Mistyping. Am. J. Hum. Genet. 35: 241-262.

Malécot G, 1966. Probabilité et hérédité, INED/PUF (paris)

Sutter J, Tabah L, 1951. La mesure de l'endogamie et ses applications démographiques. J. Soc. Stat de Paris, 92: 243-267.

TABLEAU 1



$$\phi_{ij} = (1-p)^2(1-m)^2(1-r)^4(\tfrac{1}{2})^5 + (1-p)^4(1-r)^4(\tfrac{1}{2})^5 + (1-p)^5(1-m)^4(1-r)^5(\tfrac{1}{2})^{10} + (1-p)^3(1-m)^6(1-r)^9(\tfrac{1}{2})^{10}$$

HYPOTHESES	ϕ_{ij}	$\Delta = (\phi_0-\phi_{ij})/\phi_0$
p=m=r=0	ϕ_0 = .0644	-
p=.01; m=r=0	.0625	3.0 %
p=.04; m=r=0	.0569	11.6 %
p=.04; m=r=.01	.0540	16.1 %
p=r=.04; m=.01	.0476	26.1 %

TABLEAU 2

LIENS DE PARENTE	HYPOTHESES p = 0.01 m = 0.00 r = 0.00	p = 0.04 m = 0.00 r = 0.00	p = 0.04 m = 0.01 r = 0.01	p = 0.04 m = 0.01 r = 0.04
Oncle-nièce Tante-neveu	1.33 %	5.22 %	9.60 %	17.57 %
Cousins-germains	1.93 %	7.52 %	13.03 %	23.10 %
Cousins inégaux	2.54 %	9.83 %	16.37 %	28.30 %
Cousins issus de germains	3.28 %	12.54 %	19.90 %	32.98 %
TOTAL	2.26 %	8.77 %	14.75 %	25.61 %

Erreurs relatives sur quelques coefficients de consanguinité et sur le coefficient moyen de consanguinité apparente (dernière ligne) en fonction de diverses hypothèses sur les erreurs dans les pedigrees. Ces résultats ont été obtenus à partir des données de Barraï et al. (1962).