Calcul de probabilités de parenté

Vocabulaire essentiel : partage d'allèles par ascendance commune (IBD), hypothèse nulle, Likelihood Ratio (LR), exclusion allélique, probabilité *a priori*, probabilité postérieure (ou *a posteriori*)

Les **Likelihood Ratios** (**LR**, « Rapports de vraisemblance ») sont une méthode classique de la génétique médico-légale, notamment dans le cadre de la recherche de filiations (généralement de paternités), de l'analyse de mélanges d'ADN ou de l'identification d'un individu par l'analyse de l'ADN des membres de sa famille. Cette méthode est en général associée aux kits STR d'identification, à l'origine des bases de données nationales et internationales (dont certaines peuvent fournir des fréquences alléliques de référence). La méthode dépend de l'estimation de la probabilité des génotypes observés et de la relation testée.

Chaque relation de parenté est caractérisée par la probabilité que les deux individus concernés **partagent des allèles par ascendance commune** (c'est-à-dire hérités d'un ancêtre commun récent ; on dit aussi « Identical-by-descent », « **IBD** »). Ces probabilités sont connues pour toutes les relations de parenté simples [Tableau 1] et peuvent être calculées dans toutes les situations, généralement informatiquement : on note k_0 la probabilité que les deux individus d'une paire ne partagent aucun allèle IBD ; k_1 et k_2 sont les probabilités de partager 1 ou 2 allèles IBD. La Figure 1 présente la méthode de calcul de ces coefficients IBD.

Relation de parenté	k_0	k_1	k_2
PO	0	1	0
FS	1/4	1/2	1/4
HS/AV/GC	1/2	1/2	0
CO	3/4	1/4	0
U	1	0	0

Tableau 1 - Coefficients IBD

k_0 : proportion de loci pour lesquels aucun allèle n'est partagé ou probabilité d'exclusion des loci pour une relation de parenté donnée.

k_1 : proportion de loci pour lesquels un allèle est partagé ou probabilité de partage d'un allèle IBD à un loci pour une relation de parenté donnée.

k_2 : proportion de loci pour lesquels deux allèles sont partagés ou probabilité de partage de deux allèles IBD à un loci pour une relation de parenté donnée.

PO, FS, HS, AV, GC, CO : voir fiche 9 – La parenté en génétique médico-légale

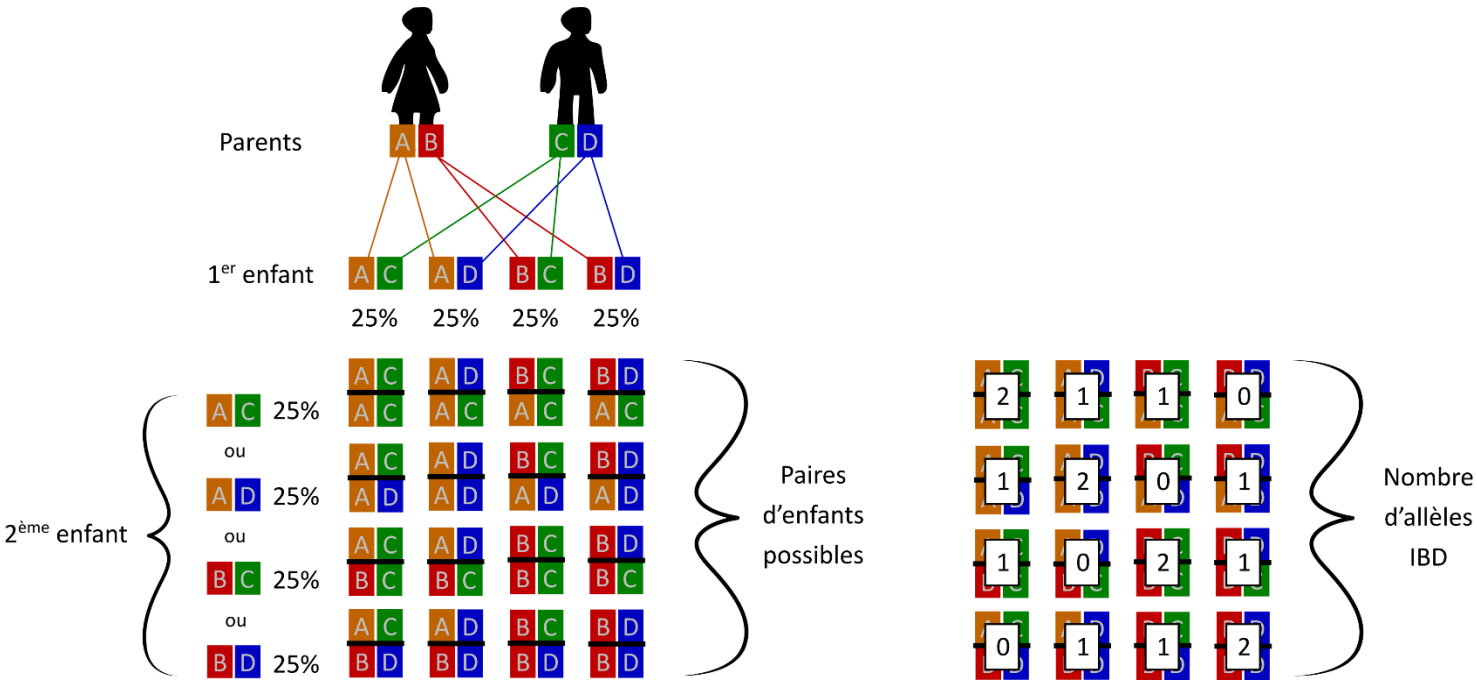

Figure 1 - Calcul des coefficients IBD pour une paire Frère/Sœur

Deux parents de génotype A/B et C/D sont à l'origine de 4 génotypes possibles dans la génération suivante : A/C, A/D, B/C et B/D.

Parmi les 16 combinaisons de deux enfants possibles on observera 25% de paires ne partageant aucun allèle IBD, 50% un seul allèle IBD et 25% deux allèles IBD.

LIKELIHOOD RATIOS (LR)

<u>Principe de la méthode : la probabilité du génotype observé</u>

La méthode des Likelihood Ratios (« rapports de vraisemblance ») dépend en premier lieu du calcul de la probabilité qu'un génotype soit observé en considérant une relation de parenté donnée (ou l'absence d'une telle relation). Pour tout génotype isolé, on peut d'abord calculer une probabilité d'apparition qui dépend des fréquences alléliques utilisées (et donc de la population de référence qui a servi à déterminer ces fréquences). Dans une population qu'on suppose être à l'équilibre de Hardy-Weinberg, on peut déterminer la probabilité d'un génotype a/b (hétérozygote pour le marqueur) ou a/a (homozygote pour le marqueur) avec l'une des deux équations suivantes :

$$Pr(G = a/b) = 2p_a p_b$$

$$Pr(G = a/a) = p_a^2$$

Où p_a est la fréquence de l'allèle « a » et p_b la fréquence de l'allèle « b ». Par exemple, la probabilité qu'un individu porte les allèles 11 et 14 au locus D8S1179 est égale à $2p_{11}p_{14}$. La probabilité du profil STR ou SNP complet (en supposant que tous les marqueurs soient indépendants) est le produit des probabilités de tous les loci [Figure 2]. Revoyez la fiche « 7 – Principes de l'identification génétique ».

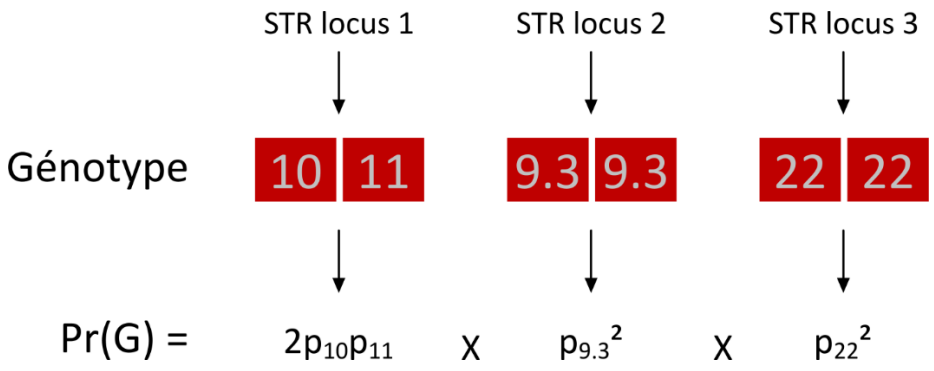

Figure 2 - Probabilité d'un génotype

Soit Pr(G) la probabilité d'occurrence d'un génotype G dans une population en fonction des fréquences alléliques STR, avec p_x, la fréquence d'un allèle x.

Vraisemblance des génotypes en fonction du lien de parenté

Pr(H|G) désigne la **probabilité** de l'hypothèse H étant donné l'observation des génotypes. On appelle « vraisemblance » la probabilité inverse Pr(G|H) qui désigne la probabilité des génotypes sous l'hypothèse H. Pr(G|PO) est donc la **vraisemblance** de G sous l'hypothèse PO. Le Théorème de Bayes permet d'exprimer Pr(H|G) en fonction de Pr(G|H) :

$$\Pr(H|G) = \frac{\Pr(H) \times \Pr(G|H)}{\Pr(G)}$$

Où Pr(H) et Pr(G) sont les probabilités *a priori* de H et G, qui seront discutées plus bas. Pour chaque catégorie de parenté, il est possible d'établir la vraisemblance des deux génotypes d'une paire. Cette probabilité [Tableau 2] dépend des coefficients IBD [Tableau 1] et des fréquences alléliques des marqueurs utilisés. Chaque paire de génotypes a donc une vraisemblance définie par les fréquences alléliques, qui est pondérée par les coefficients IBD, eux-mêmes dépendants de la relation de parenté étudiée.

Génotype A	**Génotype B**	**Probabilité de la paire de génotypes pour n'importe quelle relation de parenté testée**
a/a	a/a	$k_2p_a^2 + k_1p_a^3 + k_0p_a^4$
a/a	b/b	$k_0p_a^2p_b^2$
a/a	a/b	$k_1p_a^2p_b + 2k_0p_a^3p_b$
a/a	b/c	$2k_0p_a^2p_bp_c$
a/b	a/b	$2k_2p_ap_b + k_1p_ap_b(p_a+p_b) + 4k_0p_a^2p_b^2$
a/b	a/c	$k_1p_ap_bp_c + 4k_0p_a^2p_bp_c$
a/b	c/d	$4k_0p_ap_bp_cp_d$

Tableau 2 - Probabilités des paires de génotypes

La probabilité d'une paire de génotypes peut être exprimée en fonction des coefficients IBD et des fréquences des allèles a, b, c et d, respectivement p_a, p_b, p_c et p_d.

Le calcul de la vraisemblance d'une paire de génotypes en fonction des fréquences et de la relation de parenté peut ainsi être effectué [Tableau 2]. Soit G une paire de génotypes G1=a/b et G2=a/c et H toute relation testée :

$$Pr(G|H) = k_1 p_a p_b p_c + 4k_0 p_a^2 p_b p_c$$

On teste par exemple les relations FS et HS. Comme vu plus haut [Tableau 1], pour une relation FS on a $k_1 = ½$ et $k_0 = ¼$ et pour une relation HS on a $k_1 = ½$ et $k_0 = ½$. D'où :

$$Pr(G|FS) = ½\, p_a p_b p_c + ¼ \times 4k_0 p_a^2 p_b p_c = ½\, p_a p_b p_c + p_a^2 p_b p_c$$

$$Pr(G|HS) = ½\, p_a p_b p_c + ½ \times 4k_0 p_a^2 p_b p_c = ½\, p_a p_b p_c + 2p_a^2 p_b p_c$$

Il existe deux catégories de parenté particulières. La vraisemblance des génotypes de deux individus non-apparentés (**U** ou « **Unrelated** ») est simplement égale au produit de la probabilité de chaque génotype (l'intersection de deux probabilités indépendantes), car elle ne dépend pas du partage d'allèles par ascendance commune. En ce sens, il ne s'agit pas d'une catégorie de parenté, mais de la valeur seuil généralement utilisée par les tests statistiques pour déterminer la probabilité des autres catégories de parenté.

La catégorie de parenté **Parent-Enfant**, notée **PO**, est également particulière, parce qu'elle suppose la transmission d'un allèle pour chaque marqueur. Cela signifie que **chaque locus présente au moins un allèle partagé** et que la probabilité de transmission de cet allèle partagé est de 50% si le parent est hétérozygote pour le marqueur et de 100% s'il est homozygote. Si au moins un locus ne présente pas d'allèle partagé (constituant une **exclusion allélique**), la vraisemblance des génotypes sous l'hypothèse d'une parenté PO sera égale à zéro. On peut par exemple calculer, pour un parent supposé P=a/a et un enfant supposé E=b/b :

$$k_0 = 0$$

$$Pr(G|PO) = k_0 p_a^2 p_b^2 = 0 \times p_a^2 p_b^2 = 0$$

Cela étant vérifié pour toutes les fréquences alléliques, **une exclusion allélique invalidera automatiquement toute relation PO supposée**, à moins de prendre en compte la possibilité d'une ou plusieurs mutations. Dans l'analyse de nombres importants de marqueurs (milliers ou millions de SNP), il est nécessaire d'inclure un taux de mutation dans les calculs, afin de ne pas éliminer des parentés PO réelles pour lesquelles il y aurait justement eu mutation ou erreur de détermination des allèles.

Dans tous les cas, il convient d'établir l'**hypothèse nulle** qu'il n'y a pas de relation de parenté et l'**hypothèse alternative** que la relation recherchée est vérifiée. Le calcul permettra d'établir le rapport des vraisemblances des deux hypothèses (le **Likelihood Ratio** lui-même) et la probabilité que l'hypothèse nulle soit rejetée (la **probabilité *a posteriori***).

Comparaison des probabilités : rapports de vraisemblance et probabilité postérieure

La vraisemblance des génotypes étant déterminée, l'objectif des tests de parenté est d'établir Pr(PO|G), la **probabilité** de l'hypothèse PO pour un ensemble de génotypes observés. Le théorème de Bayes décrit plus haut peut être exprimé sous forme de « ratio » (le rapport de deux probabilités), en fonction des **probabilités *a priori*** et des vraisemblances des génotypes :

$$\frac{\Pr(\mathrm{PO}|\mathrm{G})}{\Pr(\mathrm{U}|\mathrm{G})} = \frac{\Pr(\mathrm{PO}) \times \Pr(\mathrm{G}|\mathrm{PO})}{\Pr(\mathrm{G})} \times \frac{\Pr(\mathrm{G})}{\Pr(\mathrm{U}) \times \Pr(\mathrm{G}|\mathrm{U})} = \frac{\Pr(\mathrm{G}|\mathrm{PO})}{\Pr(\mathrm{G}|\mathrm{U})} \times \frac{\Pr(\mathrm{PO})}{\Pr(\mathrm{U})}$$

Etant donné qu'il n'est en général pas possible d'exprimer numériquement les hypothèses de parenté présentées par la justice ou l'archéologie, on présuppose que les deux probabilités a priori sont égales, c'est-à-dire que le rapport Pr(PO)/Pr(U) est égal à 1 et peut être négligé. On considère donc que le rapport des probabilités est égal au LR correspondant :

$$\frac{\Pr(\mathrm{PO}|\mathrm{G})}{\Pr(\mathrm{U}|\mathrm{G})} = \frac{\Pr(\mathrm{G}|\mathrm{PO})}{\Pr(\mathrm{G}|\mathrm{U})} = \mathrm{LR}_{\mathrm{PO/U}}$$

Bien que les LR soient suffisants pour identifier quelle catégorie de parenté est favorisée par les tests, il est parfois nécessaire de pouvoir présenter une probabilité entre 0 et 1, en particulier dans le domaine juridique. Pour ce faire, on doit pouvoir montrer (ou supposer) que les deux catégories de parenté testées sont les seules envisageables ou, en d'autres termes pour l'exemple présenté ici, que PO et U sont la liste exhaustive des événements possibles et que la somme de leurs probabilités vaut donc 1. La **probabilité postérieure** est alors donnée par la formule suivante :

$$\Pr(\mathrm{PO})_{\mathrm{posterior}} = 1 - \frac{1}{\mathrm{LR}_{\mathrm{PO/U}}}$$

Il faut noter qu'il est possible de calculer des LR entre n'importe quelles catégories de parenté, en comparant par exemple PO à FS ou FS à HS. Il n'est cependant pas possible de calculer le LR de U, car la probabilité d'une paire de génotypes est égale à leur vraisemblance sous l'hypothèse U. D'où :

$$\Pr(\mathrm{U}|\mathrm{G}) = \frac{\Pr(\mathrm{U}) \times \Pr(\mathrm{G}|\mathrm{U})}{\Pr(\mathrm{G})}$$

Pr(G|U) = Pr(G)

D'où :

Pr(U|G) = Pr(U)

La probabilité de U n'est pas affectée par la vraisemblance des génotypes observés. Il convient donc de calculer tous les LR par rapport à U et d'interpréter chaque test pour chaque catégorie de parenté. On n'établira pas la probabilité que deux individus ne soient pas apparentés, au contraire on rejettera les probabilités de parenté trop faibles.

Limites théoriques

Les LR sont, par définition, dépendants des hypothèses de parenté formulées *a priori*. Il est possible de proposer des hypothèses complexes (si les coefficients de partage IBD correspondants sont identifiables), mais les résultats des tests ne permettent pas de suggérer une relation de parenté non-testée. Par exemple, le rejet des hypothèses PO, FS, HS et CO ne constitue pas une indication que CO2 est plus probable, **chaque hypothèse est indépendante**. Par ailleurs, comme noté plus haut, il n'est pas possible d'établir une probabilité pour U avec la méthode des LR.

Limites d'application : petites populations mal définies et ADN dégradé

Même dans les conditions techniques idéales (profils génétiques de qualité et sans ambigüités), l'étude des populations anciennes ou mal étudiées implique deux limites spécifiques : l'hétérogénéité des échantillons et un nombre limité d'individus. Parce que l'application des LR requiert la constitution de fréquences alléliques de référence, elles doivent habituellement être construites à partir de petits nombres d'individus, ce qui signifie qu'elles ne sont parfois pas fiables. Ce cas de figure concerne le matériel archéologique mais aussi les populations modernes très minoritaires ou celles chez qui seuls quelques centaines d'individus ont pu être étudiés.

Il faut ajouter à ces deux limites les conséquences de la dégradation de la molécule d'ADN au cours du temps. En effet, certains prélèvements ne permettent pas d'obtenir de profils STR complets ou même presque complets.