babilités p_1, p_2, etc. dans la population, et évaluons le pouvoir de discrimination de ce marqueur génétique sous l'angle de vue de la capacité à disculper un innocent lorsque la police dispose d'une trace laissée par l'auteur d'un crime. Il y a une probabilité p_1 que la trace soit de génotype G_1 et dans ce cas tous les individus de génotype autres que G_1 sont disculpés. Les chances que l'individu à innocenter soit de génotype autre que G_1 sont de $1 - p_1$. Les chances d'avoir cette configuration (trace de génotype G_1 et individu que l'on réussit à disculper parce qu'il est de génotype autre que G_1) sont de $p_1 (1 - p_1)$. La trace pourrait être de génotype p_2 et seuls les individus de génotype autre que G_2 seraient disculpés. Les chances d'avoir cette configuration sont de $p_2 (1 - p_2)$. On peut poursuivre la réflexion avec tous les génotypes possibles. Au final, s'il y a n génotypes possibles, nos chances de disculper un innocent sont de

$$p_1 (1 - p_1) + p_2 (1 - p_2) + \ldots. p_n (1 - p_n)$$

C'est cette expression que l'on appelle le pouvoir de discrimination (P_d). Elle peut être développée:

$$\begin{aligned} P_d &= p_1 - p_1^2 + p_2 - p_2^2 + \ldots + p_n - p_n^2 \\ &= (p_1 + p_2 + \ldots + p_n) - (p_1^2 + p_2^2 + \ldots + p_n^2) \end{aligned} \tag{8.2}$$

Etant donné que $p_1 + p_2 + \ldots. p_n = 1$,

$$\begin{aligned} P_d &= 1 - (p_1^2 + p_2^2 + \ldots + p_n^2) \\ P_d &= 1 - \sum_{i=1}^{n} p_i^2 \end{aligned} \tag{8.3}$$

L'expression entre parenthèses dans l'équation (8.3) est en quelque sorte l'inverse du pouvoir de discrimination. Elle exprime la probabilité d'avoir deux génotypes identiques en analysant deux individus pris au hasard dans la population considérée. C'est ce que l'on appelle bien maladroitement la ***probabilité d'identité*** (P_{id}), et que nous préférons désigner sous l'appellation *probabilité de correspondance* (P_c). Les anglophones parlent volontiers de *probability of match*

$$\begin{aligned} P_{id} &= \sum_{i=1}^{n} p_i^2 \\ \Rightarrow P_d &= 1 - P_{id} \end{aligned} \tag{8.4}$$

Le théorème de Bayes et les principes de l'interprétation

La formule mathématique utilisée dans l'approche bayésienne découle du théorème de Bayes, du nom du Révérend Thomas Bayes. Elle constitue la clef mathématique d'une théorie du raisonnement, permettant d'ajuster rationnellement une conviction ou un point de vue à la lumière de nouvelles informations et offre une solution au problème de l'induction (Hacking, 2001). Elle peut être présentée de la manière suivante sous forme de «chances» (dans le sens du terme anglosaxon ***odds***):

$$\underbrace{\frac{P(H_1 \mid E,I)}{P(H_2 \mid E,I)}}_{\text{chances a posteriori}} = \underbrace{\frac{P(E \mid H_1,I)}{P(E \mid H_2,I)}}_{\text{rapport de vraisemblance}} \times \underbrace{\frac{P(H_1 \mid I)}{P(H_1 \mid I)}}_{\text{chances a priori}} \qquad (8.12)$$

où H_1 et H_2 sont les hypothèses en concurrence. $P(H_1|I)$ et $P(H_2|I)$ expriment la probabilité que chacune des hypothèses H_1 et H_2 soient vraies connaissant les informations (I) de l'affaire. E correspond à une observation qui a été faite, les profils ADN par exemple. Cette observation E est plus ou moins vraisemblable selon les hypothèses qui sont envisagées. $P(E|H_1,I)$ et $P(E|H_2,I)$ décrivent la probabilité que l'on fasse l'observation E si l'une respectivement l'autre hypothèse est vraie, en tenant compte toujours du contexte I. $P(H_1|E,I)$ et $P(H_2|E,I)$ décrivent la probabilité des hypothèses H_1 et H_2 une fois que l'observation E a été faite.

Le théorème de Bayes correspond parfaitement au processus judiciaire [Champod et Taroni, 1993]. Au départ, il y a deux hypothèses en concurrence, celle de l'accusation (par exemple, «l'accusé est à l'origine du cheveu trouvé sur la scène de crime») et celle de la défense («un autre individu inconnu et non pas l'accusé est à l'origine du cheveu trouvé sur la scène de crime»). Chacune de ces hypothèses a une certaine probabilité d'être vraie et il est possible d'exprimer ces probabilités par un rapport $P(H_1|I)/P(H_2|I)$. Les examens amènent ensuite des observations, qui risquent fort de modifier notre conviction quant à la véracité respective des deux hypothèses. Le rapport des probabilités initiales (on parle souvent de ***probabilités a priori***) s'en trouve modifié pour donner un nouveau rapport des probabilités tenant compte des observations $P(H_1|E,I)/P(H_2|E,I)$. Ces nouvelles probabilités sont souvent appelées les ***probabilités a posteriori***. Le processus peut d'ailleurs se poursuivre avec des

$$P(a\,/\,b) = 2p_a(1-F_{ST})p_b = 2(1-F_{ST})p_a p_b \qquad (8.21)$$

On procède de façon similaire pour calculer le dénominateur $P(a/b,a/b)$:

$$P(a\,/\,b,a\,/\,b) = 2(1-F_{ST})p_a p_b \cdot \frac{F_{ST}+(1-F_{ST})p_a}{(1+F_{ST})} \cdot \frac{F_{ST}+(1-F_{ST})p_b}{(1+2F_{ST})} \qquad (8.22)$$

Ces formules peuvent être intégrées dans l'équation (8.20), ce qui, après simplification, aboutit à la formule générale ci-après [Balding et Nichols, 1994]. Le cas de l'homozygote se construit de façon analogue.

$$P(a\,/\,b \mid \text{suspect} = a\,/\,b) = \frac{2[F_{ST}+(1-F_{ST})p_a][F_{ST}+(1-F_{ST})p_b]}{(1+F_{ST})(1+2F_{ST})} \qquad (8.23)$$

$$P(a\,/\,a \mid \text{suspect} = a\,/\,a) = \frac{[2F_{ST}+(1-F_{ST})p_a][3F_{ST}+(1-F_{ST})p_a]}{(1+F_{ST})(1+2F_{ST})} \qquad (8.24)$$

Comme leur formulation mathématique l'indique, ces formules sont à comprendre comme étant les probabilités d'observer un individu de génotype donné dans la population considérée, sachant qu'un individu possédant ce génotype (celui de la personne d'intérêt) a déjà été observé.

Les valeurs de facteur de consanguinité mesurées dans les populations européennes sont en général inférieures à 0,002 [Gill, 2003]. Conformément à ce qui est connu de ces populations, c'est en Finlande que la valeur la plus élevée a été mesurée: 0,005. Pour des estimations du coefficient de consanguinité dans différentes populations, on renvoie à la lecture de Foreman *et al.* [1998] et Foreman et Lambert [2000]. La prise en compte du facteur de consanguinité affaiblit évidemment la valeur probante d'un résultat. Si l'on prend par exemple les deux cas de la figure 8.3, le calcul avec un paramètre F_{ST} de 0,01 donne pour le cas n° 1 une valeur de $2{,}58 \times 10^{-10}$ au lieu de $1{,}75 \times 10^{-10}$; pour le cas n° 2, on obtient de $1{,}78 \times 10^{-24}$ au lieu de $3{,}74 \times 10^{-27}$. L'effet de la prise en compte du facteur de consanguinité est bien sûr plus marqué lorsque les allèles sont rares. Comme on peut le voir dans ces exemples chiffrés et dans des évaluations plus sophistiquées et étendues [Buckleton, 2005b], la prise en compte du facteur de consanguinité a un impact somme toute modeste. Dans des situations où le coefficient de consanguinité n'est pas connu, par exemple parce que la population est très hétérogène, l'utilisation d'un F_{ST} de 0,03 présente des garanties largement suffisantes pour éviter toute sous-estimation de l'impact de ce paramètre [Buckleton, 2005a].

Les considérations ci-dessus permettent de tenir compte de la consanguinité pour le calcul de la fréquence d'un génotype (éq. (8.15) et (8.16)) et pour le calcul de

la probabilité $P(E|H_2)$ à l'échelle d'un marqueur, qui envisage qu'il y ait une personne autre que la personne connue (ex. le suspect) qui possède le même génotype pour le marqueur considéré (éq. (8.23) et (8.24)). Il faudrait pouvoir ajouter un dernier échelon de prise en compte de la consanguinité dans la phase de multiplication des résultats obtenus pour chaque marqueur. En effet, en utilisant les équations susmentionnées, la consanguinité est prise en compte pour les calculs pour chaque marqueur mais pas pour la combinaison des marqueurs. Dans les équations (8.18) et (8.19) et pour la combinaison de 2 marqueurs, l'expression $P(G_T|G_S,H_2,I)$ devrait s'écrire $P(G_{1T},G_{2T}|G_{1S},G_{2S},H_2,I) = P(G_{1T}|G_{1S},G_{2S},H_2,I) \times P(G_{2T}|G_{1T},G_{1S},G_{2S},H_2,I)$, G_{1T}, G_{2T}, G_{1S}, G_{2S} étant les génotypes de la trace et du suspect pour les marqueurs 1 et 2. On devine la complexité des développements pour des génotypes à 10 ou 15 marqueurs, ainsi que les nombreux coefficients de dépendance à connaître. Malheureusement, la mesure de ces éventuelles dépendances entre marqueurs est une difficulté considérable, qui incite certains auteurs à suggérer un plancher arbitraire pour la valeur de $P(E|H_2,I)$, pour les profils avec 10 marqueurs ou plus que l'on obtient aujourd'hui [Foreman et Evett, 2001]. Pour les 10 marqueurs du kit SGM+ par exemple, leur proposition est de limiter $P(E|H_2,I)$ à une valeur de 1 sur 1 milliard. Des recherches récentes [Hopwood, 2012] sur les kits avec 15 marqueurs montrent que même si théoriquement les rapports de vraisemblance obtenus avec des correspondances sur 15 marqueurs sont plus grands que ceux obtenus avec 10 marqueurs, il n'y a pas de réel bénéfice à reporter des probabilités de profils ADN qui soient plus petites qu'un sur un milliard. Ils proposent les valeurs seuils suivantes: pour des personnes non apparentées 1 sur un milliard, pour des frères ou des sœurs 1 sur 10^5, pour des parents/enfant 1 sur 10^7, pour des demi-frères/sœurs ou oncle/neveu 1 sur 10^9 également.
Curran *et al.* (2003) et par la suite Buckleton *et al.* (2006c) ont étudié quelle serait l'ampleur de l'erreur sur l'estimation de la probabilité d'un profil en utilisant les équations (8.6) et (8.7), c'est-à-dire en ne considérant pas le facteur de consanguinité et en admettant donc que la population est en équilibre de Hardy-Weinberg [les Anglo-Saxons appellent ce mode de calcul simpliste la *product rule*]. Dans les scénarios de modélisation qu'ils ont envisagés, la population considérée est divisée en n sous-populations et le facteur de consanguinité est supposé être de 0,01 ou 0,03. L'étude montre que l'application des formules (8.23) et (8.24) donne des estimations conformes aux attentes dans de telles populations structurées, alors que l'utilisation de la *product rule* a une nette tendance à surestimer la rareté du profil.

La dérive génétique (*Genetic drift*)

La dérive génétique est une fluctuation de la fréquence des allèles qui survient au sein d'une petite population par les hasards de l'échantillonnage des gamètes.

Elle contribue de façon déterminante à générer des différences de fréquences alléliques entre les populations humaines. Elle fait partie des phénomènes que l'on peut mettre sous l'étiquette «erreurs d'échantillonnage». Dans un échantillon de N enfants, nés dans une population où la fréquence de l'allèle a est de pa, la probabilité que l'on rencontre exactement i allèles a est de $[2N!/i!(2N\text{-}i)!]p_i(1\text{-}p)^{2N-i}$. Si l'on fait des simulations numériques de l'évolution de la fréquence au fil des générations, on obtient des courbes du type de la figure 8.4, où l'on voit que plus la taille N de la population est faible plus la dérive potentielle est rapide et importante.

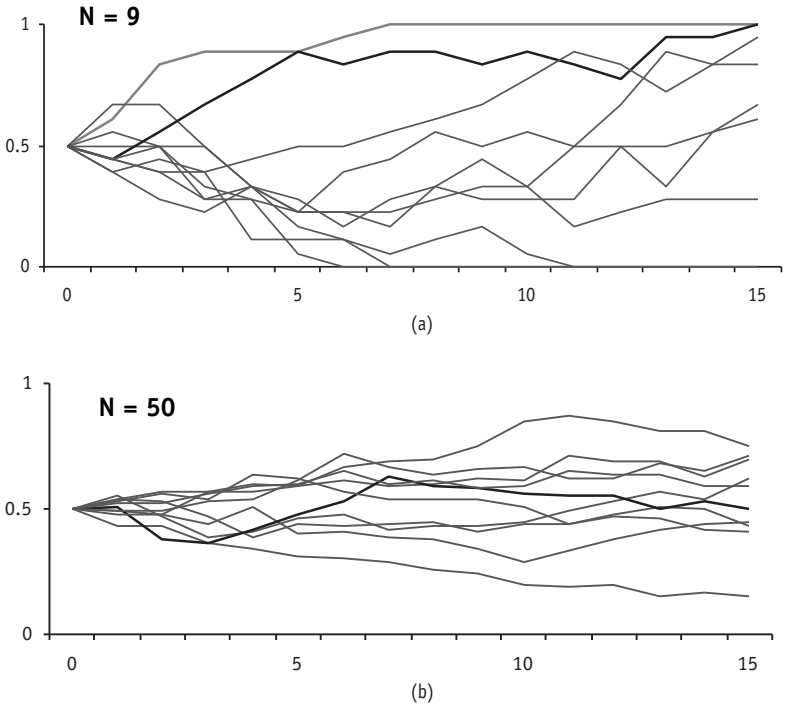

Fig. 8.4 Illustration du phénomène de la dérive génétique. Evolution de la fréquence d'un allèle (axe vertical) ayant une fréquence initiale de 50% au cours de 15 générations (axe horizontal). (a) suivi de l'évolution dans 10 petites populations de 9 individus chacune. (b) suivi de l'évolution dans 10 populations de 50 individus chacune. On voit bien que la fréquence initiale de 50% dérive très vite, d'autant plus vite que la population est petite.

secondaire. Les bases de données permettant d'estimer les fréquences d'un haplotype commencent à être considérables et les statistiques disposent d'outils fournissant des chiffres pour les fréquences qu'on peut considérer comme fiables.

Par contre, le fait que le même ADNmt soit commun à tous les membres d'une descendance très large est nettement plus difficile à apprécier statistiquement. La deuxième hypothèse examinée dans le rapport de vraisemblance envisage que la trace provienne d'une personne inconnue autre que le suspect. Il est essentiel de bien définir qui est cette personne inconnue. Elle peut être un individu quelconque non apparenté mais pourrait aussi être un lointain cousin ayant hérité du même ADNmt que le suspect par un ancêtre commun. Dans ce cas, on peut utiliser des hypothèses multiples comme nous l'avons vu précédemment. La justice française a été confrontée à des difficultés de ce type dans une affaire d'assassinat en Corse, où les suspects, accusés par la correspondance du profil de leur ADNmt avec celui d'une trace, pouvaient à juste titre faire valoir que cet ADNmt était certainement commun à de nombreuses personnes de leur région de la Corse [*Le Monde*, 15 novembre 2002]. Ce cas souligne l'importance de bien définir la population d'intérêt et d'utiliser des fréquences alléliques locales.

Fréquences des profils ADNmt et des profils STR Y

Aux paragraphes 5.3.3 et 5.4.1 ont été présentées les sources d'information sur les fréquences des profils ADNmt et STR Y. Les experts sont fréquemment confrontés à des cas où le profil d'une trace n'a encore jamais été observé ou qu'un nombre très limité d'individus présentent ce profil dans les bases de données utilisées pour l'estimation des fréquences. Dans la ligne des considérations du paragraphe 8.3.1 sur la précision des chiffres, il faut tenir compte des incertitudes sur la fréquence des profils. Avec l'ADNmt et les STR Y, nous sommes confrontés à des haplotypes et non à des génotypes combinant deux allèles.

Imaginons la situation suivante: Y représente l'haplotype observé dans la trace. Le suffixes C et S représentent la trace et le suspect (ou la personne connue analysée), respectivement. Les hypothèses d'intérêt sont l'échantillon retrouvé sur la scène du crime provient du suspect (ou d'une personne de la même ligné paternelle (H_p) et l'échantillon retrouvé sur la scène du crime provient d'une personne non apparentée au suspect (H_d). L'haplotype Y est évalué en quantifiant le rapport de vraisemblance suivant:

$$RV = \frac{P(Y_C, Y_S \mid H_p, I)}{P(Y_C, Y_S \mid H_d, I)} = \frac{P(Y_C \mid Y_S, H_p, I)}{P(Y_C \mid Y_S, H_d, I)}$$

suspect. Ceci correspond à une formule différente, mais qui donne des ordres de grandeur similaires [Selvin, 1983]:

$$P = [Np(1-p)^{N-1}] / [1-(1-p)^{N}] \tag{8.33}$$

Bien qu'elle paraisse raisonnable à plusieurs titres, l'idée d'affirmer une identification avec un profil ADN, comme on le fait (trop souvent) avec une empreinte digitale, peine à s'imposer, et ceci pour plusieurs raisons:

- Tout d'abord, l'identification est une décision et les décideurs utilisent de fait une approche normative et prescriptive, qui ne fait pas uniquement intervenir les probabilités mais également la notion forcément plus subjective de «gain» ou «perte» liée aux conséquences.
- Ensuite, la réalité génétique fait que les profils ADN ne sont pas uniques à une seule personne, et ceci dans un nombre de cas non négligeables qui sont les vrais jumeaux [Scientific Sleuthing Review, 1996]. Et c'est là une différence d'importance avec les empreintes digitales, les jumeaux ayant des empreintes digitales distinctes.
- Les profils ADN actuels paraissent certes avoir des fréquences extraordinairement basses (fig. 8.3). Néanmoins, si l'on envisage de comparer deux à deux tous les individus de la planète, on s'aperçoit que même avec les profils très rares obtenus de nos jours, il y a une quasi-certitude qu'il existe des paires d'individus sur la planète possédant le même profil ADN[11]. C'est seulement lorsque la fréquence des profils ADN descend en-dessous de 1×10^{-22}, que ce risque descend en dessous de 1% [Aitken & Taroni, 2004].
- Les fréquences des profils ADN ne sont pas des fréquences mesurées, mais sont des fréquences calculées à partir des fréquences des allèles, et en utilisant un modèle génétique choisi. Les chiffres utilisés sont grevés d'incertitudes et les calculs effectués s'appuient sur de nombreux présupposés (§ 8.4.1), notamment des présupposés d'indépendance. Ces aléas sont d'autant plus importants que les fréquences sont rares et que les paramètres combinés sont nombreux. Dès lors, plus les fréquences des profils ADN s'abaissent pour s'approcher des valeurs qui permettraient peut-être d'affirmer une identification,

11 Le problème est similaire au classique calcul des chances que 2 personnes dans un groupe aient leur anniversaire le même jour, qui se calcule par la formule $J!/[(J-N)! \times J^N]$, avec J = nombre de jours dans l'année, et N = nombre de personnes. Ce calcul, qui a vite fait de dépasser les capacités de calcul de l'ordinateur, peut être facilité en utilisant une formulation approximativement équivalente: $e^{(-N(N+1)/2J)}$
Avec les profils ADN, J peut être considéré comme égal à $1/p$, où p est la fréquence d'un profil, N étant le nombre de personnes dans la population.

de I_1 par héritage commun. On peut ensuite dresser le tableau des valeurs $P(E|H_1,I)$, de $P(E|H_2,I)$ et du rapport de vraisemblance pour les différentes combinaisons de génotypes possibles (tab. 8.14). Le calcul du rapport de vraisemblance est alors aisé à réaliser en utilisant les probabilités k_2, k_1 et k_0 adéquates. Les valeurs pour quelques liens de parenté L_P sont présentées (tab. 8.15). Si l'on utilise les valeurs adéquates pour le lien de parenté père - enfant, on retrouve les données du tableau 8.13.

Tableau 8.14 Tableau des valeurs $P(E|H_1)$, de $P(E|H_2)$ et du rapport de vraisemblance *RV* pour différentes combinaisons de profils ADN possibles pour les individus R_1 et R_2.

I_1	I_2	$P(E\|H_1)$	$P(E\|H_2)$	RV
a/a	a/a	$1 \times k_2 + p_a \times k_1 + p_a \times k_1 + p_a^2 \times k_0$	p_a^2	$(k_2+2k_1p_a+k_0pa^2)/p_a^2$
a/a	a/b	$0 \times k_2 + p_b \times k_1 + p_b \times k_1 + 2p_ap_b \times k_0$	$2p_ap_b$	$(k_1+k_0p_a)/p_a$
a/a	b/b	$0 \times k_2 + 0 \times k_1 + p_b^2 \times k_0$	p_b^2	k_0
a/a	b/c	$0 \times k_2 + 0 \times k_1 + 2p_bp_c \times k_0$	$2p_bp_c$	k_0
a/b	a/b	$1 \times k_2 + p_b \times k_1 + p_a \times k_1 + 2p_ap_b \times k_0$	$2p_ap_b$	$[k_2+k_1\ (p_a+p_b)+k_02p_ap_b]/2p_ap_b$
a/b	a/a	$0 \times k_2 + p_a \times k_1 + p_a^2 \times k_0$	p_a^2	$(k_1+k_0p_a)/p_a$
a/b	a/c	$0 \times k_2 + p_c \times k_1 + 2p_ap_c \times k_0$	$2p_ap_c$	$(k_1+2k_0p_a)/2p_a$
a/b	c/d	$0 \times k_2 + 0 \times k_1 + 2p_cp_d \times k_0$	$2p_cp_d$	k_0
a/b	c/c	$0 \times k_2 + 0 \times k_1 + p_c^2 \times k_0$	p_c^2	k_0

Tableau 8.15 Valeur des probabilités k_2, k_1 et k_0 pour quelques types de lien de parenté entre les individus R_1 et R_2. Les valeurs pour d'autres types de liens peuvent être déduites assez simplement.

Lien de parenté	k_2	k_1	k_0
père - enfant	0	1/2	0
frère - frère (sœur)	1/4	1/4	1/4
frère - demi-frère (demi-sœur)	0	1/4	1/2
grand-parent - enfant	0	1/4	1/2
oncle (tante) - neveu (nièce)	0	1/4	1/2
cousin - cousin (cousine)	0	1/8	3/4

Si l'on envisage que les deux individus soient des frères, on peut assez vite se rendre compte que le rapport de vraisemblance pour un marqueur ne peut pas être inférieur à 1/4. C'est la valeur que l'on obtient lorsqu'ils n'ont aucun allèle en commun. Dans le cas de demi-frères, il ne peut pas être inférieur à 1/2. Il n'y a par contre pas de limite supérieure, ce rapport devenant d'autant plus élevé que la fréquence des allèles communs entre les deux individus est basse.