PRINCÍPIOS DE GENÉTICA FORENSE

FRANCISCO CORTE-REAL
DUARTE NUNO VIEIRA

Capítulo 5

ALGUNS CONCEITOS DE GENÉTICA POPULACIONAL COM RELEVÂNCIA EM GENÉTICA FORENSE

Luís Souto

Departamento de Biologia, Universidade de Aveiro

CENCIFOR - Centro de Ciências Forenses

DOI | HTTP://DX.DOI.ORG/10.14195/978-989-26-0957-7_5

RESUMO

A Genética Forense não faz sentido sem a Genética das Populações, pois é nesse ramo da Biologia que radica a validade da aplicação das suas metodologias. A amostra forense ao ser analisada revela uma informação que tem que ser enquadrada no contexto de uma população à qual pertencerá o indivíduo ou indivíduos que produziram uma tal amostra, seja ela de referência (para comparação) ou a *amostra-problema* de um caso. Atribuir um maior ou menor peso à informação obtida analiticamente implica a compreensão de alguns princípios basilares da genética populacional entre os quais o de Hardy-Weinberg, as causas de desvios a esse princípio e as formas de o detetar.

PALAVRAS-CHAVE

Genética forense; genética populacional; Hardy-Weinberg.

SUMMARY

Forensic Genetics makes no sense without the Population Genetics, because it is the branch of biology where lies the validity of the application of its methodologies. The forensic sample analysis reveals information that has to be framed in the context of a population to which belong the individual or individuals who have produced such a sample, either reference (for comparison) or unknown samples. Assign a greater or smaller weight to information obtained analytically implies the understanding of some basic principles of population genetics including the Hardy-Weinberg, the causes of deviations from this principle and ways to their detection.

KEYWORDS

Forensic genetics; Population genetics; Hardy-Weinberg.

1. INTRODUÇÃO

1.1 DE VOLTA AOS GRUPOS SANGUÍNEOS

A noção de "grupo sanguíneo", a percepção de que podemos agrupar os humanos de acordo com a sua pertença a uma determinada classe ditada pelo seu grupo A, B, AB ou O (ou ainda Rh + ou Rh -) é algo (quase) do domínio do cidadão comum. Talvez menos popular será a noção de que essas classes não ocorrem com igual frequência, que há tipos sanguíneos mais raros e outros "vulgares".

Na verdade há muito que são reconhecidas as diferenças na frequência destas classes e o padrão de distribuição é variável de população para população. A título de exemplo, compare-se a distribuição do sistema ABO para a população portuguesa e chinesa:

Tabela 1 – Distribuição dos Grupos Sanguíneos (sistema AB0) em duas populações, portuguesa e chinesa (em percentagem)

	A	**0**	**B**	**AB**
Portugal[1]	47	42	8	3
China (Han)[2]	29	28	34	9

Este exemplo permite-nos elucidar dois pontos: primeiro, existem pesos diferentes a atribuir à informação obtida do marcador "grupo sanguíneo". Por exemplo, em Portugal, uma coincidência de resultado em que a amostra-problema tem o grupo AB e um dado suspeito tenha também esse grupo AB, tem um peso muito superior ao que teria uma conjugação de resultados em que um suspeito e amostra-problema são identificados como pertencentes ambos ao grupo A, dada a "raridade" do grupo AB na população portuguesa.

Por outro lado o exemplo mostra-nos ainda que não é indiferente considerarmos nessa ponderação, a origem étnica ou a população de referência de suspeito e amostra-problema.

Atualmente os grupos sanguíneos, que estão na base da genética forense, estão afastados da panóplia de marcadores polimórficos em uso pelos laboratórios qualificados. Hoje são os vários "polimorfismos de ADN", com relevo para os microssatélites (ou STRs, do inglês, *Short Tandem Repeats*), os marcadores-padrão. Isto não invalida que os grupos sanguíneos possam ser eles próprios "marcadores de ADN", tudo depende do nível de análise.

A determinação dos grupos sanguíneos ABO (ou Rh) feita pelas tradicionais reacções de antigénio-anticorpo, enquadra-se no que em genética chamamos de *Fenótipo*, ou seja aquilo que é expresso, que resulta de um produto celular (geralmente uma proteína). Já quando se trata dos "polimorfismos de ADN", estamos a pretender analisar informação contida no próprio ADN sem que esteja em causa a produção de algo detetável (uma proteína), situamo-nos no âmbito do *Genótipo*. Esta perspetiva é no entanto uma simplificação. O genótipo não determina por si só totalmente o fenótipo. Desde que se obteve a sequenciação do genoma humano em 2001 que as estimativas do número de genes para a espécie humana revelaram uns "humildes" 20 000 a 25 000 genes[3]

1 Distribuição para Portugal continental, arredondada à unidade para efeitos comparativos. Fonte: Duran, (2007).

2 Zhu, et al. (2010).

3 Não há um número certo para a contagem de genes, mesmo com os dados mais refinados sobre a sequenciação do genoma. O leitor mais interessado sobre este tópico pode consultar Pertea e Salzberg (2010).

(contra cerca de 30 000 da videira por exemplo). Esta constatação abriu novas e entusiasmantes avenidas de pesquisa laboratorial e até de reflexão filosófica, a ponto de alguns falarem mesmo num "eclipse do dogma genótipo-fenótipo" (Lock e Nguyen 2010).

Em genética forense, não se pesquisa a informação genética, genótipos que possam permitir inferir fenótipos (como doenças, características físicas etc). No entanto este é também um "dogma" em vias de desconstrução. Por outro lado, enquanto no caso dos grupos sanguíneos, algumas variantes dos genes (os *Alelos*) dominam sobre outras, (dominância/recessividade), nos polimorfismos do ADN como os microssatélites, a informação que recebemos por via paterna (alelo paterno) e por via materna (alelo materno) têm a mesma "força", situação já anteriormente conhecida como alelos codominantes.

Fiquemos porém no âmbito deste texto, no contexto dos marcadores polimórficos codominantes e sem (pelo menos aparente) relação com a expressão fenotípica de características, ou seja no contexto da definição da Lei nº 5/2008[4].

2. AS FREQUÊNCIAS DOS ALELOS E DOS GENÓTIPOS: O EQUILÍBRIO DE HARDY-WEINBERG

Um dos princípios mais importantes em genética de populações, o Equilíbrio de Hardy-Weinberg, foi curiosamente descoberto em simultâneo por um matemático inglês e um médico alemão, de forma independente, em 1908. Esse princípio estabelece a constância das frequências dos alelos e dos genótipos, geração após geração, sob determinados pressupostos (particularmente exigentes) e permite, no caso particular da genética forense, assumir a independência estatística usada nos cálculos com perfis genéticos, nomeadamente quando existe uma coincidência de perfis.

É também a observância desse princípio que permite que um geneticista forense utilize as tabelas de frequências (génicas e genotípicas) publicadas para a população de referência. De contrário, haveria necessidade de realizar uma amostragem a cada determinação, o que não faria sentido.

As frequências genotípicas podem ser previstas a partir das frequências alélicas, pressupondo o equilíbrio de Hardy-Weinberg.

Um caso de suposta violação na ilha de *Lupagaso*

Consideremos um caso forense de violação numa ilha em que foi analisado um marcador STR hipotético, o D007.

D007 é um sistema pouco polimórfico, uma vez que apenas tem dois alelos possíveis: 2 unidades de repetição (alelo 2) ou 4 unidades de repetição (alelo 4):

Alelo 2: GATA GATA

Alelo 4: GATA GATA GATA GATA

Consideremos então a seguinte tabela de resultados obtidos após genotipagem do marcador D007 nas amostras do caso:

4 A Lei Nº5/2008 de 12 de Fevereiro. Lei que aprova a criação de uma base de dados de perfis de ADN para fins de identificação civil e criminal, na alínea e) do artigo 2º.

Tabela 2 – Genotipagens no marcador D007 numa violação em *Lupagaso*.

Amostra	Genótipo no marcador D007
Zaragatoa poscoital	2 – 4
Vítima	4 – 4
Suspeito	2 – 2

Parece óbvia a "incriminação" do suspeito dado que, sendo a vítima homozigótica para o alelo 4, a presença do alelo 2 na zaragatoa poscoital não só é compatível com o genótipo identificado no suspeito como, sendo este homozigótico para esse alelo, essa informação se traduz num peso acrescido.

Torna-se porém imperativo avaliar até que ponto é frequente encontrar esse alelo 2 na população de referência, neste caso a população residente em *Lupagaso*.

Os cientistas recorrem a tabelas de frequências obtidas previamente através da genotipagem de amostras representativas e formadas aleatoriamente, sem qualquer grau de parentesco entre os indivíduos.

Consultada a tabela de distribuição de frequências para o sistema D007, encontra-se um valor de p = frequência do alelo 2 = 0.02 e de q= frequência do alelo 4 = 0.98.

Desde logo constatamos a raridade do alelo 2. Aliás essa variante encontra-se no limiar mínimo para que seja considerado um alelo e não uma mutação (frequência superior a 1 %).

Os valores das frequências alélicas como vemos, têm uma importância decisiva na forma como valoramos a informação (laboratorial) das genotipagens obtidas.

Importa agora esclarecer como se podem obter essas frequências alélicas inscritas nas referidas tabelas de frequências.

2.1. CONTAGEM DE GENES

Consideremos agora que se pretende conhecer a distribuição do nosso marcador hipotético D007 na população da ilha (hipotética também) de *Lupagaso*. Para tal os cientistas genotiparam um número de 100 de indivíduos cujos resultados se encontram na tabela:

Tabela 3 – Genotipagem de 100 indivíduos no sistema D007 na população da ilha de *Lupagaso*

	Genótipos			Totais
	2 -2	**2 - 4**	**4-4**	
Número de Indivíduos	30 (N_{2-2})	60 (N_{2-4})	10 (N_{4-4})	100
Frequência dos genótipos em F0	0.30	0.60	0.10	1

Podemos calcular as frequências dos alelos 2 (p) e alelo 4 (q) pela contagem de genes:

Cada indivíduo com genótipo 2 – 2 produz potencialmente dois gâmetas com alelo 2 (uma vez que quer no cromossoma de origem paterna quer no de origem materna a informação é a mesma); cada indivíduo com genótipo 2 – 4 apenas tem o potencial de transmitir um alelo 2 (e um alelo 4); cada indivíduo com genótipo 4 – 4 tem o potencial de produzir dois gâmetas com alelo 4.

Por outro lado, num sistema bialélico como D007, cada indivíduo tem o potencial de transmissão de dois alelos (quaisquer que eles sejam), logo para N indivíduos temos 2N alelos.

Ou seja, contando os genes a partir dos genótipos observados teremos:

$$p = \text{Frequência do alelo 2} = \frac{\text{Contagem genes 2}}{2N} = \frac{(2N_{2\text{-}2} + N_{2\text{-}4})}{2N}$$

$$p = \text{Frequência do alelo 4} = \frac{\text{Contagem genes 4}}{2N} = \frac{(2N_{4\text{-}4} + N_{2\text{-}4})}{2N}$$

[Expressão 1]

Podemos rearranjar matematicamente a expressão 1…

$$p = 2N_{2\text{-}2} / 2N + {}_{N2\text{-}4} / 2N$$

$$p = 2N_{4\text{-}4} / 2N + {}_{N2\text{-}4} / 2N$$

[Expressão 2]

Reconhecendo agora na expressão 2 as frequências dos genótipos f (2-2), f (2-4) e f (4-4) poderemos usar as frequências relativas de cada genótipo observado para determinar a frequência alélica:

$$p = f(2\text{-}2) + f(2\text{-}4)/2$$

$$p = f(4\text{-}4) + f(2\text{-}4)/2$$

[Expressão 3]

Aplicando ao exemplo da Tabela 3 obtém-se então a seguinte estimativa das frequências alélicas na geração F0 pela contagem génica:

$$p = \text{Frequência do alelo 2} = (30/100) + \tfrac{1}{2}(60/100)$$
$$p = 0.60$$

$$q = \text{Frequência do alelo 4} = (10/100) + \tfrac{1}{2}(60/100)$$
$$q = 0.40$$

A soma das duas frequências, como temos apenas dois alelos neste sistema hipotético, será $p + q = 1$.

A metodologia exposta é conhecida como contagem de genes e é o método corrente, sobretudo a partir do momento em que a genética forense passou a utilizar marcadores codominantes em que todos os genótipos são identificáveis tecnicamente, não havendo alelos "escondidos", ou seja recessivos, que não se expressam e outros dominantes, que mascaram a manifestação dos recessivos.

Os valores das frequências alélicas obtidos por esta contagem de genes, a partir dos genótipos observados, constituem uma *estimativa* para as frequências alélicas e genotípicas da população e não o valor real dos *parâmetros* na população. Esta é uma noção fundamental neste contexto.

Sem nos determos na complexidade do tema, importa realçar que não há um método único para estimar as frequências génicas e podemos obter por exemplo três estimativas de frequências baseando-nos numa mesma amostra da população. À medida que o tamanho da amostra aumenta e se aproxima do tamanho da população, as frequências dos genótipos observados irão ser tendencialmente mais próximas das verdadeiras frequências na população e independentemente do método utilizado para estimar, a frequência génica assim calculada será próxima da frequência génica real na população.

Estes desenvolvimentos mostram claramente a diferença entre estimativas das frequências (na prática o que é viável obter com amostragem necessariamente finita) e as (verdadeiras) frequências génicas e genotípicas na população em estudo.

É demonstrável que o método de contagem de genes pode ser considerado como uma "estimativa natural" o que em termos estatísticos será contextualizado no quadro dos métodos de estimativa por *Máxima Verosimilhança* (ML).

O método de Máxima Verosimilhança em Estatística é uma das (várias) estratégias de estimação de parâmetros, com larga aplicação, mas outras existem entre as quais as abordagens *bayesianas* que alguns autores defendem como mais apropriadas por acolherem a possibilidade de alelos não previstos.

Utilizando uma abordagem como a contagem de genes (ou máxima verosimilhança) obtemos um *estimador pontual*, ou seja um valor específico para estimativa do verdadeiro valor do parâmetro populacional. Precisamos ainda de uma ferramenta estatística que permita avaliar a qualidade da nossa estimativa e isso pode conseguir-se com recurso a intervalos de confiança, cuja metodologia de cálculo admite por si também distintas abordagens.

Cada vez que efetuarmos uma nova amostragem, um novo estudo genético-populacional em *Lupagaso* (ou…em Timor-Leste ou Portugal) iremos encontrar amostras com características diferentes e obteremos novos valores de estimativa das frequências, pelo que a estimativa, como se compreende, está associada a determinada incerteza.

Sendo p a verdadeira frequência de um alelo A, a variância da frequência alélica estimada pelo método de contagem génica é dada pela expressão **p (1-p) / (2n)**, sendo n o tamanho da amostra. O desvio padrão é por sua vez a raiz quadrada da variância.

O intervalo de confiança de 95% para o verdadeiro valor da frequência alélica p é cerca de duas vezes o valor do desvio padrão em cada lado do intervalo. Quanto maior a amostragem (n) menor serão variância e desvio padrão e mais precisa será a estimativa obtida para uma dada frequência alélica, algo que conhecemos em múltiplas aplicações da Estatística.

Sendo o cálculo das frequências génicas e genotípicas relevante para a genética forense, persiste discussão científica em torno da pertinência da inclusão dos intervalos de confiança no reporte da informação ao poder judicial[5].

2.2. EQUILÍBRIO DE HARDY-WEINBERG

Podemos então considerar que os dois alelos 2 e 4 do sistema D007 se juntarão aleatoriamente

5 A este propósito consulte-se Charles Brenner em http://dna-vie.com/noconfid.htm#challenge).

e com as proporções definidas pelas frequências p e q respetivamente.

Na situação de um organismo diplóide (ou seja com dois conjuntos de cromossomas homólogos, um de origem paterna e outro de origem materna) como é o caso da espécie humana, num sistema bi-alélico, com dois alelos com frequências p e q, podemos recorrer a um tradicional "quadrado de Punnet", um diagrama onde se equacionam todas as possibilidades de combinações entre os gâmetas (paterno e materno) e os respetivos resultados probabilísticos.

Note-se que o entendimento do quadrado de Punnet se baseia num princípio da estatística segundo o qual a probabilidade de dois acontecimentos simultâneos e não mutuamente exclusivos se obtém pela multiplicação das probabilidades de cada um desses acontecimentos independentes.

Tabela 4 -Quadrado de Punnet para um sistema bialélico com dois alelos 2 e 4 com frequências p e q respetivamente na geração inicial F0. A sombreados os genótipos esperados pelas diferentes combinações alélicas na geração F1

	Gâmetas masculinos (espermatozóides)		
Gâmetas femininos (ovócitos)		Alelo 2 (p)	Alelo 4 (q)
	Alelo 2 (p)	2-2 (p^2)	2-4 (pq)
	Alelo 4 (q)	2-4 (pq)	2-4 (q^2)

Seguindo a Tabela 4, espermatozóides com o alelo 2, têm probabilidade p, igual à frequência desse alelo na população, de fecundar ovócitos com probabilidades p e q respetivamente, consoante estes ovócitos tenham o alelo 2 ou 4.

Teremos para as frequências dos vários genótipos possíveis (frequências genotípicas esperadas na geração F1) respetivamente:

$$f\ (\text{Alelo 2- Alelo 2}) = p^2$$

$$f\ (\text{Alelo 2- Alelo 4}) = 2pq$$

$$f\ (\text{Alelo 4- Alelo 4}) = q^2$$

A matemática reconhece estas quantidades como a expansão binomial $(p+q)^2$, cuja soma é a unidade.

Ou seja: $\mathbf{(p+q)^{2=}\ p^2+2pq+q^2=1}$.

A expressão binomial permite prever as frequências dos vários genótipos (frequências genotípicas) numa dada geração a partir das frequências p e q dos gâmetas produzidos pela geração anterior. A distribuição binomial, em termos da estatística, é um caso particular da "distribuição multinomial", que descreve a probabilidade de obter uma amostra com um número determinado de objetos em cada uma de diversas classes, o que na binomial se traduz em duas classes.

Voltando ao exemplo considerado e substituindo p e q, pelos respetivos valores teríamos as frequências genotípicas esperadas na geração seguinte (geração F1):

$$f\ (\text{Alelo 2 /Alelo2}) = p^2 = (0.6)^2 = 0.36$$

$$f\ (\text{Alelo 2 /Alelo 4}) = 2pq = 2x\ 0.6x0.4 = 0.48$$

$$f\ (\text{Alelo 4 / Alelo 4}) = q^2 = (0.4)^2 = 0.16$$

Aplicando estas frequências a uma nova amostra de 100 indivíduos da nossa ilha de *Lupagaso*, teríamos agora os seguintes valores esperados para os três genótipos:

Tabela 5 – Frequência esperada dos vários genótipos na geração F1.

Genótipos do sistema D007	Frequência Esperada
2-2	36
2-4	48
4-4	16

Segundo o equilíbrio de Hardy-Weinberg estas frequências genotípicas obtidas após uma geração de panmixia (F1), isto é, em que os cruzamentos são feitos ao acaso, irão permanecer constantes em subsequentes gerações.

Com os valores na geração F1 retomávamos o mesmo valor inicial das **frequências alélicas** calculado para a geração parental inicial F0, ou seja 0.6 e 0.4 respetivamente. No entanto como se vê na Tabela 5, a distribuição genotípica esperada em F1 difere da inicial. Pode-se provar porém que a partir desta primeira geração panmítica F1, a distribuição genotípica iria permanecer constante.

Este princípio exemplificado para um sistema de dois alelos pode ser matematicamente demonstrado também para um sistema de n vários alelos como na realidade acontece com os marcadores microssatélites.

O equilíbrio de Hardy-Weinberg diz-nos *que quer as frequências dos genes (frequências alélicas) quer as frequências genotípicas se irão manter com certo valor indefinidamente se se mantiver um conjunto de condições e tal é passível de demonstração matemática*.

Outra forma de enunciar o equilíbrio de Hardy-Weinberg é dizer que a expressão $p^2+2pq+q^2=1$ prevê de forma precisa as frequências genotípicas a partir dos valores p e q das frequências alélicas.

2.3.1. CONDIÇÕES E UTILIDADE DO EQUILÍBRIO DE HARDY-WEINBERG

Para que a previsibilidade e constância do equilíbrio de Hardy-Weinberg se concretizem há um conjunto de pressupostos. A população tem que ser panmítica como referido, ou seja os cruzamentos devem ocorrer de forma aleatória. A fertilidade deve ser idêntica. Não se admite sobreposição de gerações, mutações, nem seleção natural, nem fatores que alterem o fundo genético inicial (não há movimentos migratórios).

Estas condições como é óbvio, não se observam nas populações humanas e nem tão pouco noutros seres vivos. No entanto mantém-se a sua utilidade e observa-se até um surpreendente grau de concordância.

A utilidade do formalismo de Hardy-Weinberg consiste no fato de ele servir como modelo, como *hipótese nula* (um termo caro à estatística), que nos permite prever as frequências dos genótipos na ausência de fatores perturbadores da simples e aleatória combinação dos gâmetas portadores dos vários alelos. A comparação com aquela situação ideal, em que apenas a hereditariedade contribui para os valores das frequências génicas, permite-nos em cada situação real, evidenciar e avaliar as forças evolutivas (e até eventuais causas de erro técnico ou amostral) ao comparar a distribuição genotípica e as frequências génicas obtidas com o esperado segundo o modelo (teórico, não existente nas populações que estudamos) do equilíbrio de Hardy-Weinberg.

2.3.2. TESTANDO O EQUILÍBRIO DE HARDY-WEINBERG

A comparação entre os resultados das genotipagens obtidas com uma dada amostra da população e a hipótese nula (a do equilíbrio de Hardy-Weinberg) faz-se recorrendo a testes estatísticos como o de χ^2 de Pearson:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

No teste χ^2 obtém-se o valor χ^2 pelo somatório dos quadrados das diferenças entre os genótipos observados e esperados divididos pelos genótipos esperados. Note-se que o numerador da expressão nos fornece a diferença entre valores observados e esperados em cada classe genotípica mas esta diferença é computada em termos absolutos ou seja não nos interessa em que sentido ocorre essa diferença, se a mais ou a menos, e daí se elevar ao quadrado. Essa diferença absoluta é dividida pelos valores esperados pois temos que relativizar em relação a um valor esperado (uma diferença de 4 em 100 tem um peso muito diferente de uma diferença de 4 em 8, respetivamente 4% e 50%). Cada uma destas diferenças ponderadas calculadas em cada classe genotípica é somada às restantes previstas de acordo com as combinações possíveis em cada marcador genético atendendo ao seu polimorfismo.

O teste χ^2 irá fornecer a probabilidade de obtermos o valor encontrado (ou maior ainda) para a diferença entre os valores observados e os esperados segundo o modelo teórico (hipótese nula do equilíbrio de Hardy-Weinberg) apenas pelo acaso (e não por qualquer fator evolutivo ou experimental).

Se as diferenças entre os valores de distribuição dos nossos genótipos e os valores que seriam de esperar na situação de equilíbrio de Hardy-Weinberg forem muito grandes, então haverá razões para suspeitar que essas diferenças não se devem ao acaso apenas.

A analogia com o lançamento de uma moeda ao ar é evidente: se as diferenças entre o número de caras e coroas esperado pelo mero fator sorte e o número de caras e coroas efetivamente obtido se traduzir numa diferença grande, *significativa*, então suspeitamos de alguma viciação da moeda uma vez que partimos do modelo (hipótese nula) de que as duas faces da moeda teriam à partida a mesma probabilidade de sair.

Mas até que ponto esse valor de probabilidade é suficiente para tomarmos uma decisão sobre se a nossa distribuição segue ou não o equilíbrio de Hardy-Weinberg?

Torna-se necessário comparar o valor de χ^2 obtido com uma distribuição de χ^2. À medida que o valor de χ^2 cresce, a probabilidade de a diferença entre os valores observados e esperados ser devida apenas ao acaso diminui (quanto maiores as diferenças entre valores observados e esperados, menos provável é a nossa hipótese nula, isto é, o equilíbrio de Hardy-Weinberg). Por convenção em geral adota-se o critério de um patamar de probabilidade 0.05 ou menos para as diferenças obtidas em χ^2 para rejeitar a hipótese nula. Significa que se o acaso apenas explica as diferenças obtidas em 5% ou menos,

então é de rejeitar que a população se encontre em equilíbrio de Hardy-Weinberg.

Note-se ainda que ao estimar a estatística χ^2 temos que ter em conta os "graus de liberdade", a quantidade de informação usada no respetivo cálculo. No exemplo do marcador DS007 para prever os três genótipos esperados (3 classes) utilizamos duas informações: o número de indivíduos e uma das frequências alélicas (com dois alelos, uma vez determinada uma das frequências, p a outra é função da primeira, q=1-p). O número de graus de liberdade é dado pelo número de classes menos o número de informações usadas para o cálculo das várias proporções esperadas. No exemplo teríamos 3-2 = 1 grau de liberdade.

Dado que, com os marcadores de ADN atualmente utilizados, muitas classes de genótipos esperados (segundo o equilíbrio de Hardy-Weinberg) resultariam em valores baixos, recorre-se antes a uma adaptação do teste exato de Fisher, sendo usual a metodologia desenvolvida por Guo e Thompson (Guo e Thompson 1992) de tipo Monte Carlo ou em alternativa de tipo Cadeias de Markov.

O método de Fisher assenta no uso da probabilidade condicionada que resulta de agregação de todas as possíveis combinações genotípicas que seriam tão ou menos prováveis que os genótipos observados (seriam pelo menos tão raras quanto a distribuição genotípica encontrada) empregando as mesmas frequências alélicas.

Como verificámos o enunciado do equilíbrio de Hardy-Weinberg implica conhecimento de uma das frequências num sistema bialélico (no exemplo hipotético a frequência de um dos alelos, 2 ou 4, p ou q).

A dedução dessa frequência assenta em amostragens de uma dada população de referência (os referidos estudos populacionais). Ora sabendo que usaremos essas frequências para vários cálculos, na verdade praticamente todos, aplicáveis no contexto da genética forense e designadamente o cálculo da "raridade de um dado perfil genético", então percebemos como o grau de maior ou menor incerteza na determinação dessas frequências afeta imediatamente o grau de incerteza sobre os nossos cálculos de âmbito forense.

O erro é minimizável (não eliminável) pelo alargamento da amostragem. Desde que assegurado que esta amostragem é constituída por indivíduos representativos da população de referência, não relacionados entre si, obtida de forma aleatória, não há um número ideal ou sequer um número N de indivíduos na amostra que se possa recomendar com base em critérios objetivos. No entanto alguns autores referem as N=100 como um valor aceitável (Butler 2012).

Este alargamento da amostra é controlável pelo investigador, ou seja temos ao nosso alcance minimizar o erro associado a esta "amostragem estatística". A estatística fornece-nos as técnicas de avaliação desse erro de amostra (da "amostragem estatística"). Note-se que esse erro é meramente estatístico e nada tem a ver com possíveis erros laboratoriais ou outras causas de âmbito genético ou demográfico.

Porém, na "amostragem genética", a sorte de gâmetas transportando as suas combinações alélicas, está fora do controle do investigador, sendo determinada pelo processo evolutivo subjacente à diferenciação das populações referido anteriormente.

3. SUBESTRUTURAÇÃO POPULACIONAL

3.1. A POPULAÇÃO

A população, como bem lembrado por Lock e Nguyen (Lock e Nguyen 2010) é um conceito abstrato que "não existe na natureza antes definido formalmente por entidades nacionais e estatais". Nos Estados Unidos, o FBI utiliza como suporte ao seu sistema de base de dados CODIS, estudos genético-populacionais em que as populações de referência são grandes grupos étnico-populacionais no mínimo discutíveis como afro-americanos; caucasianos; hispânicos. Eis-nos pois com a possibilidade de "inventar" as populações humanas consoante o nosso critério, os nossos objetivos.

É atualmente praticamente consensual que as atuais populações humanas resultarão de um processo de expansão e diversificação a partir de África há cerca de 100 000 anos, um período de tempo muito curto se comparado com os milhões de anos que foram consumidos na evolução que conduziu à nossa espécie, o *Homo sapiens*.

Nesta caminhada a partir de África ("out of África"), os grupos humanos encontraram (terão eventualmente até se cruzado embora tal seja matéria de controvérsia científica), com outros hominídeos com menor capacidade cognitiva, como os Neandertais existentes em zonas como a Europa ocidental. Estes novos *Homo sapiens* terão vencido pela sua inteligência, pelo seu valor "cultural" e foram-se fixando pelos distintos continentes e em função de barreiras geográficas e por força destas, barreiras socioculturais e linguísticas, foram-se também diferenciando em "populações", mas não em subespécies a tal ponto que a unidade se mantém, ou seja a inter-reprodutibilidade entre os humanos independentemente do grupo e localização geográfica a que pertencem. Como seria de esperar a diversidade proporcionada pela informação genética, a diversidade genética, é consideravelmente superior no continente africano e decresce com o afastamento deste. Na Europa temos uma grande "monotonia genética".

3.2. DERIVA GENÉTICA E FLUXO GENÉTICO

As populações humanas apresentam graus variáveis de diferenciação, consequência de processos como migração, mutação e deriva.

A deriva genética traduz o fato de as frequências génicas apresentarem uma flutuação ao longo do tempo apenas por força do acaso. É algo que não podemos controlar por melhor que seja a nossa amostragem aleatoriamente retirada da população. No limite as frequências podem chegar a extremos, ou seja termos um único alelo presente na população (fixação) e a eliminação de alelos (um alelo que ocorra com baixa frequência numa população de tamanho reduzido a probabilidade de esse alelo se perder em cada geração é considerável). Num sistema bi-alélico isso equivaleria termos frequência p=1, q=0.

Em cada geração se tivermos dois alelos com frequências p e q as mesmas iriam se manter constantes segundo o equilíbrio de Hardy-Weinberg se a totalidade dos gâmetas presentes numa geração estivesse presente na seguinte nas respetivas proporções. No entanto em cada geração, nem todos os indivíduos vão contribuir para a reprodução e assim apenas uma sub-amostra (aleatória) da população real é usada nas combinações dos gâmetas

masculino e feminino para a geração seguinte. Este "erro de amostra genético" é tanto maior quanto menor for o tamanho do efetivo populacional (as populações reais também não são infinitas como exigem os pressupostos do equilíbrio de Hardy-Weinberg). O mecanismo da deriva genética é um fator que tende a homogeneizar geneticamente uma população, reduz a diversidade genética.

Quanto maior o desvio padrão das frequências alélicas (provocado pelo erro de amostra genético), maior a diversidade genética e menor a deriva genética.

Se duas populações forem geneticamente distantes isso reflete-se numa grande diferença entre as respetivas frequências génicas.

Segundo o equilíbrio de Hardy-Weinberg, teríamos uma população panmítica mas, sobretudo quando consideramos populações grandes, verifica-se que os "cruzamentos não são ao acaso", havendo uma maior facilidade em, por exemplo, um homem português de Trás-os-Montes vir a ter filhos com uma transmontana do que com uma algarvia. Isto significa que uma dada população (a população portuguesa no exemplo) pode em certas circunstâncias, compreender subunidades atuando já de forma independente, significa que uma tal população se encontra *subestruturada*. Essas subunidades serão tanto mais distintas quanto menor for a taxa de mistura genética, quanto menor for o fluxo genético entre ambas. Um acidente geográfico como uma cadeia montanhosa pode ser o elemento causal de uma diminuição de fluxo genético e consequente subestruturação populacional de uma população.

Se as duas subunidades ou duas populações tiverem características genéticas muito distintas, um aumento do fluxo genético irá contribuir para o aumento da diversidade genética da população e será mais evidente o impato desse fluxo genético. Por exemplo as linhagens femininas (mitocondriais) provenientes da África subsaariana atualmente integrantes do "património genético" português, são claramente reconhecíveis e vice-versa dado o ponto de partida tão geneticamente divergente (Pereira e Ribeiro, 2009).

As frequências de alguns genes mostram uma distribuição "clinal" ou seja, apresentam uma variação contínua desde um ponto de frequência mais elevada até zonas de frequência mais rara, à medida que nos distanciamos. Isto acontece como consequência de ser mais provável um cruzamento com indivíduos geograficamente próximos do que com pessoas localizadas mais longinquamente e portanto menos acessíveis. Este padrão clinal tem sido observado para alguns polimorfismos humanos, sendo um exemplo clássico o grupo sanguíneo B (Cavalli-Sforza 1996). Apesar da recente tendência para a facilitação da mobilidade populacional, continua a ser possível identificar esse gradiente na distribuição de algumas das frequências génicas humanas. Uma boa parte das pessoas continua a realizar uniões num certo (embora cada vez maior) raio de proximidade.

Recentemente os estudos de grande magnitude envolvendo um conjunto de até 300000 SNPs (algo inimaginável há poucos anos) e recorrendo à estratégia de WGP (*Whole Genome Polymorhisms*) vieram a corroborar os trabalhos já clássicos de Cavalli-Sforza (uma das grandes referências na genética das populações humanas), evidenciando por técnicas de tratamento de dados PCA (*Principal Component Analysis*) uma notável correlação entre a posição de um

dado indivíduo no "espaço genético" com a sua posição no espaço geográfico (McCoy, 2009).

Por outro lado, uma diminuição do fluxo genético deixa, a prazo, maior potencialidade para a atuação do mecanismo aleatório da deriva genética, diferenciando cada vez mais as (agora) subpopulações a ponto de elas serem já geneticamente distinguíveis.

A situação com consequências do ponto de vista forense é a de termos duas "subpopulações" sob uma população *a priori* considerada, o que obriga então a utilizar uma diferente base de dados de distribuição de frequências na hora de aplicar ao cálculo da raridade/vulgaridade de uma dado perfil genético coincidente.

O cromossoma Y, com a sua especificidade (na porção não recombinante com o cromossoma X) aliado à prática frequente da patrilocalidade (i.e. a mobilidade para a constituição do casal é assegurada pela mulher permanecendo o homem no seu local original) tem sido objeto da maior atenção pelos cientistas forenses numa perspetiva da subestruturação populacional, inclusive a um nível "microgeográfico" (Brion 2004).

Quando a diminuição do fluxo genético é resultante da decrescente probabilidade de cruzamentos em função do aumento da distância entre indivíduos ou populações, estamos no conceito do *isolamento por distância*, formulado por Sewall Wright.

3.2.1. Avaliação do grau de substruturação: da heterozigotia aos índices Fst.

Wright (Wright 1951) desenvolveu o primeiro de uma longa série de índices que permitem avaliar o grau de subestruturação populacional de uma população. Genericamente designados por *Fst*, estes índices assumiram posteriores desenvolvimentos a que correspondem sucessivas nomenclaturas: GST (Nei 1973); θ (Weir e Cockerham 1984); RST (Sltakin 1995).

Quando os membros de uma população são aparentados, então a probabilidade de obterem descendentes homozigóticos é maior do que na situação de cruzamentos ao acaso (a probabilidade de parentes partilharem um mesmo alelo é obviamente superior à da população em geral). Na natureza temos casos extremos, de *autogamia*, em que a reprodução se faz com hermafroditismo (ambos os sexos no mesmo indivíduo) como certas espécies de plantas.

Uma medida da diversidade genética numa população é a heterozigotia (em cada marcador ou locus).

A heterozigotia pode ser obtida pela contagem do número de heterozigotos numa população (HO, heterozigotia observada) ou a partir das frequências génicas (HE, Heterozigotia esperada[6], dada pela expressão $He = 1 - \sum p_i^2$ (pi é a frequência de um alelo i num dado locus). Quando maior o número de heterozigotos maior a diversidade genética. A noção de heterozigotia num dado locus pode ser aplicada a múltiplos loci[7].

6 Esta expressão corresponde ao termo "diversidade genética de Nei". A este respeito consulte-se por exemplo Hamilton, M. (2009).

7 A extensão a múltiplos (k) alelos da heterozigotia esperada obtém-se somando todos os possíveis homozigóticos e subtraindo da unidade: $He= 1 \sum_{i=1}^{k} p_i^2$ enquanto a heterozigotia observada será o somatório de todos os heterozigóticos observados na amostra populacional para k alelos: $Ho = \sum_{i=1}^{h} Hi$ em que Hi é a frequência observada em cada genótipo e h o número de heterozigotos possíveis num sistema de k alelos (sendo que esse número possível é dado pela expressão h= k (k-1) /2).

O índice de fixação F[8] (ou F_{IS} na notação original de Wright) compara ambas as heterozigotias (a observada e a esperada caso houvesse cruzamentos ao acaso) considerando uma única população: *F = He-Ho / He*, em que He é a frequência esperada de heterozigotos a partir das frequências alélicas da população de acordo com o equilíbrio de Hardy-Weinberg e Ho é a frequência de heterozigotos observados na população.

O índice de fixação para dada população (até que ponto os seus membros cruzam ao acaso ou pelo contrário há diminuição de heterozigotia face ao esperado) pode ser explorado no contexto de várias subpopulações.

Na realidade as populações humanas não são panmíticas e assim a estimativa das frequências génicas e genotípicas deve acautelar a estratificação das populações. Nem sempre é viável dispor de bases de dados de distribuição de frequências para todas as populações. No contexto forense, alguns indivíduos podem não estar representados nas bases de dados de frequências disponíveis para uma dada população pela sua especificidade étnico-racial ou geográfica. Por exemplo imagine-se o cenário de um crime perpetrado em Portugal por um indivíduo pertencente a uma tribo índia do interior da Amazónia.

No âmbito da subestruturação populacional (da população em subpopulações) os *Fsts* surgem na perspetiva da comparação da diversidade genética de cada subpopulação com a diversidade genética total da população, pois, como exposto acima, a subestruturação acarreta incremento da diversidade genética.

A alteração na heterozigotia (face à frequência esperada de heterozigotos segundo o equilíbrio de Hardy-Weinberg) situa-se agora por um lado dentro de cada subpopulação (devida ao fato de os cruzamentos não serem ao acaso) e por outro entre as subpopulações de uma população, devido à subestruturação desta última.

Entre os vários coeficientes que visam avaliar o grau de subdivisão de uma população consideremos θ, F e f (F_{ST}, F_{IT} e F_{IS}, respectivamente na nomenclatura de Wright).

F_{IS} (I, Indivíduos; S, Subpopulações) representa a média da diferença entre a heterozigotia observada e a esperada (segundo HW) **dentro de cada subpopulação** como consequência de os cruzamentos não serem ao acaso. Este coeficiente corresponde a correlação entre dois alelos num genótipo ao acaso em qualquer das subpopulações, ou seja trata-se do índice acima descrito para uma única população, F, mas agora como média de várias subpopulações. F_{IS} ou f traduz o "coeficiente de inbreeding" ou "consanguinidade" dentro de cada subpopulação. De acordo com Holsinger e Weir (Holsinger e Weir 2009), nas populações humanas os desvios intra-populacionais em relação às proporções do equilíbrio de Hardy-Weinberg serão pequenos, algo avaliado pelo índice F_{IS} que compara as frequências

8 Este índice pode ser encarado como uma generalização à escala populacional do "coeficiente de inbreeding" (notação ***f***) ou de consanguinidade calculado família a família e significa determinar a probabilidade de dois alelos num dado indivíduo serem "idênticos por descendência" (IBD, *Identical By Descendt*), o que implica que haverá um ancestral comum aos dois alelos presentes. Quando os membros de um casal têm parentes em comum (por exemplo são primos em segundo grau) a probabilidade de transportarem alelos idênticos por descendência no seu genótipo é aumentada o que origina uma tendência para excesso de homozigotos (e diminuição da heteroziogotia).

genotípicas observadas em relação às esperadas (segundo HW) dentro de uma população[9].

Por sua vez, o Fst, $\mathbf{F_{ST} = (H_T - H_S) / H_T}$ traduz a redução na heterozigotia devido a diferença nas frequências alélicas **entre** subpopulações. Corresponde à diferença entre a heterozigotia média esperada das subpopulações e a heterozigotia esperada da população total. Este coeficiente mede pois o **grau de diferenciação entre subpopulações** e interessa sobremaneira aquando da decisão de considerar ou não uma dada base de dados populacionais para fins forenses como adequada ao caso em apreço. O valor de FST varia entre 0.0 (não diferenciação) e 1.0 (subpopulações totalmente diferenciadas com alelos fixados).

Finalmente o terceiro índice, F_{IT}, $\mathbf{F_{IT} = H_T\text{-}H_I / H_T}$ representa a comparação entre a heterozigotia média observada nas subpopulações com a heterozigotia esperada para a população total.

Os índices de fixação conheceram sucessivamente desenvolvimentos desde a definição inicial de Wright que fora formulada no contexto de um locus com dois alelos. Assim, por exemplo, o termo G_{ST} já referido resulta da aplicação a alelos múltiplos e loci múltiplos (Wright 1978, Nei 1973).

Os índices de fixação podem alternativamente ser encarados numa perspetiva de análise de variância, comparando a variância das frequências alélicas face à variância da população total.

Se considerarmos uma dada população e várias sub-amostras dessa população então podemos usar a *variância* [σ] para avaliar até que ponto os valores observados se afastam dos valores médios, da *média*.

Dado que se considera que as subpopulações humanas partilham um mesmo passado evolutivo, por mais longínquo que seja, então as respetivas frequências génicas partilham a mesma média variando por força do efeito de "amostragem genética" mencionado.

De facto podemos ter duas distribuições de frequências alélicas com a mesma média mas distintas variâncias, ou seja o espectro de observações ser menos ou mais disperso.

Nesta perspetiva F_{ST} passa a ser a proporção da variância genética dentro de uma subpopulação (S) relativa à variância genética total. O equivalente do coeficiente de *inbreeding* F_{IS}, será agora a proporção da variância de uma subpopulação contida num indivíduo. Surgem assim novos índices de subestruturação como sejam os coeficientes θ ou θ_{ST} (Weir 1996), Φ_{ST} (Excoffier 1992), este para dados haplotípicos (análise AMOVA, análise de variância molecular).

Com o advento dos marcadores microssatélites foram introduzidos índices (como R_{ST}) baseados nas diferenças no número de repetições entre alelos em cada locus microssatélite[10].

9 Representando H_I a heterozigotia média observada por indivíduo dentro das subpopulações; HS a heterozigotia média esperada dentro de subpopulações nas condições de cruzamento aleatório; HT a heterozigotia esperada nas condições de cruzamentos aleatórios para a totalidade da população (sem considerar subpopulações) temos (por analogia com a expressão de F acima): $\mathbf{F_{IS} = (H_S\text{-}H_I) / HS.}$

10 Este método assenta num modelo mutacional, designado SMM, "stepwise mutation" em que cada mutação cria um novo alelo por adição/deleção de uma unidade de repetição, com igual probabilidade, pelo que os alelos com uma grande diferença de tamanho entre dois alelos traduz uma menor proximidade de relação). Assim compreende-se que estes métodos de avaliação da subestruturação sejam adequados quando a mutação tenha sido o mecanismo decisivo de diferenciação entre populações, enquanto os métodos "tradicionais" de avaliação do grau de estruturação (mas não totalmente postos de parte) se baseavam nas diferenças entre as frequências génicas de subpopulações.

Um exemplo da importância da subestruturação no contexto forense pode ser o da população australiana onde "escondidas" no seio de importantes aglomerados populacionais cosmopolitas coexistem minorias de populações indígenas aborígenes, com pequenos núcleos dispersos com baixos efetivos e forte tendência endogâmica e até neste caso, poligâmica (Taylor 2012; Walsh 2007). Na Austrália a quantificação dos perfis de DNA é feita com recurso a três bases de dados de frequências, Aborígenes, Europeus e Asiáticos, realizando-se frequentemente cálculos em diferentes bases alternativas incluindo a base de dados associada à origem étnica do réu caso este declare a sua afiliação étnica (Taylor2012), em substituição do sistema inicial em que se recorria a uma única base de dados representativa de cada estado federado australiano dado que por exemplo os aborígenes poderiam estar sub-representados nessa amostragem (recorde-se o que atrás dissemos sobre a amostragem em estudos populacionais).

Para os marcadores autossómicos está previsto desde há muito um fator de correção θ para os efeitos da subestruturarão populacional. A SWGDAM[11], seguindo o *National Research Council* (NRCII) recomenda as seguintes correções:

Para homozigotos em lugar de p^2 deve-se entrar com a expressão: $p^2+p(1-p)\theta$, mantendo a expressão *2pq* para os heterozigotos.

Em populações grandes há um menor número de homozigóticos do que nas subpopulações que a constituem ("efeito Whalund"). Assim introduz-se um fator de correção para prevenir a subestimativa de homozigóticos devida a subestruturação. As entidades internacionais citadas propõem o uso de valores de θ de 0.01 para populações em geral e de 0.03 para pequenas populações (mais endogâmicas) (Council 1996).

A consideração do fator θ aquando da estimativa da probabilidade de ocorrência de um perfil genético coincidente (entre suspeito e amostra-evidência) traduz-se numa redução do valor em favor do suspeito. Quanto maior o valor de θ, maior será a probabilidade de obter uma coincidência por mero acaso (Holsinger 2009). Estudos com o conjunto de marcadores STRs recomendados atualmente pelas entidades europeias (Budowle 2011) mostram por exemplo valores de FST para 15 STRs de Fst=0.0161, pelo que a aplicação dos valores dos coeficientes θ recomendados é "conservativa".

Nos marcadores de linhagem como cromossoma Y e mitocondrial, são esperados efeitos de subestruturação ainda maiores. No entanto apesar de haver recomendações por instâncias internacionais como a *DNA Commission of the International Society of Forensic Genetics* (ISFG) no sentido de os efeitos de subestruturação populacional serem tidos em consideração neste tipo de marcadores de ADN, a metodologia para tal não está ainda definida e deve-se ter em conta a pertinência ou não do uso de tal fator de correcção sobre as frequências, pois se as circunstâncias do caso não apontarem para uma especificidade étnica ou geográfica ou se as populações não estiverem subestruturadas tal não fará sentido (Cockerton 2012).

Com os marcadores polimórficos SNPs, que justificam um crescente interesse entre a comunidade científica forense, acontece que as diferenças entre as frequências de distintas populações humanas podem ser muito grandes (a frequência de

11 *Scientific Working Group on X Analysis Methods*, grupo que representa laboratórios forenses ao nível federal, estatal e local nos EUA e Canadá.

um dado alelo pode ir desde 0 numa população a 1 noutra, ou seja a fixação de um alelo e perda do outro é verosímil). Este dado torna pertinente o conhecimento da distribuição de frequências em múltiplas bases de dados populacionais quando se pretende utilizar tais tipos de marcadores na rotina forense embora haja estudos orientados no sentido da obtenção de um painel "universal" de marcadores SNPs (Pakstis 2007).

Por outro lado essa diferença de frequências tão marcada entre distintas populações (um elevado Fst) pode ser utilizada com vantagens: o marcador torna-se então um possível AIM (*Ancestral Informative Marker*) um marcador de ancestralidade e/ou origem biogeográfica o que pode ser muito útil nas aplicações forenses. Muitas vezes não há acesso a dados que informem a origem biogeográfica nomeadamente quando lidamos com amostras de vestígios (sangue, saliva) e esses marcadores AIM podem fornecer algum grau de indicação quanto à origem biogeográfica do indivíduo que produziu tais vestígios (teria sido um europeu? um africano?) o que permite estreitar o âmbito das investigações policiais quando não há uma coincidência de perfis de ADN com os suspeitos considerados ou ainda auxiliando na procura de desaparecidos.

4. PARA ALÉM DA DERIVA GENÉTICA: OUTROS FATORES DA DINÂMICA POPULACIONAL

A aplicação de um certo valor de θ é discutível: não só os distintos marcadores apresentam uma considerável variação de Fsts (pela sua história genealógica) como não são de excluir efeitos de outro fator evolutivo determinante, a *seleção natural*.

Os marcadores aprovados para genética forense são-no com base num modelo neutral ou seja admite-se que não estejam sob efeito da *seleção natural*, a qual se iria traduzir em marcadas diferenças nas frequências em certas subpopulações sujeitas a esse mecanismo. Este é um tema ainda pouco explorado mas a possibilidade aberta por trabalhos com conjuntos de altas densidades de SNPs junto a marcadores do sistema CODIS vem alertar para essa possibilidade (Anderson e Weir 2006).

Também a independência entre os vários loci é outra questão relevante no contexto da genética forense. Alguns marcadores polimórficos podem residir em loci que se apresentam em "desequilíbrio de ligação" e como tal tendem a ser transmitidos em conjunto e não de forma independente como consequência da sua proximidade num mesmo cromossoma o que se vai refletir numa menor frequência de recombinação[12].

Na situação de desequilíbrio de ligação, dois alelos em diferentes loci (note-se bem que agora estamos a considerar distintos loci e não dois alelos num dado locus como acima) tendem a ser co-herdados com maior frequência do que a esperada (se fossem verdadeiramente independentes um do outro). Daqui pode resultar uma

12 A recombinação genética ocorre de forma aleatória quando dois cromossomas homólogos trocam partes do seu ADN constitutivo como acontece no processo de formação dos gâmetas (na meiose). Quanto maior a distância entre dois segmentos de ADN maior a probabilidade de a recombinação ocorrer, pelo que a medida da recombinação entre dois *loci* é também uma medida da distância genética. Uma percentagem de 1% de recombinação corresponde a 1 centimorgan (cM) que por sua vez se pode relacionar com a distância física entre dois loci uma vez que 1cM se traduz em cerca de 1Mb (um milhão de pares de bases de ADN).

sobre-estimativa da força da evidência aquando dos respetivos cálculos forenses.

Esta possibilidade tem sido muito menos estudada pois exige uma abordagem populacional complementada com estudos familiares. Nos estudos populacionais (i.e. com base em indivíduos não relacionados) o desequilíbrio de ligação não é facilmente detetável, pelo que há necessidade de obtenção de famílias para esse efeito.

No painel de marcadores STRs adotados para fins forenses os marcadores *loci* vWA e D12S391 estão situados no mesmo cromossoma 12 e com proximidade física. Estes marcadores têm sido alvo de estudos sobre desequilíbrio de ligação (Budowle 2011; Oconnor 2011; Gill 2012), que apontam para que não devam ser considerados como loci independentes e como tal tratados de forma diferente dos restantes marcadores aquando dos cálculos de probabilidade de coincidência de perfis genéticos ou em testes de parentesco.

O conceito de desequilíbrio de ligação (LD, *Linkage Desequilibrium*) não se confunde com *Ligação* (*linkage*), uma vez que dois *loci* podem apresentar um LD sem que estejam em *linkage*. Além do *desequilíbrio de ligação* resultante da proximidade física de marcadores, vários fatores da dinâmica populacional como a substruturação populacional e em particular a deriva genética, podem concorrer para a observação de *desequilíbrio de ligação* entre *loci* não (fisicamente) ligados. O fluxo genético entre populações geneticamente distintas tende a criar desequilíbrio de ligação por mistura populacional (ALD, *Admixture Linkage Desequilibrium*) entre *loci* (ligados e não ligados) com frequências alélicas distintas nas populações originais (Pfaff 2001).

Enquanto a deriva genética constitui fator de redução da diversidade genética, outros processos contribuem para a dinâmica populacional, como as mutações ou as migrações ou a *mistura genética* os quais atuam em sentido contrário.

A *mistura genética* resulta na obtenção de uma população com características híbridas a partir da contribuição de duas ou mais populações ancestrais as quais se encontravam inicialmente isoladas (o que distingue do termo fluxo genético, em que há um contato prolongado e troca de genes entre duas populações).

A maior ou menor diversidade genética e como tal a compreensão da atual configuração das populações humanas, abrange o âmbito da mistura genética populacional (um exemplo claro foram as movimentações de grandes efetivos populacionais africanos para a América) se bem que não possam ser ignorados os efeitos dos processos de deriva genética, mutação, seleção ou migração ou ainda o efeito das próprias alterações nos efetivos populacionais[13] como os "efeitos de gargalo" (*bottleneck*) em que por força de um acontecimento sociopolítico (uma grande catástrofe, uma guerra) ou geodinâmico (dependendo da escala de tempo em que nos situamos para análise), há uma redução drástica de forma significativa do efetivo populacional o que se traduz em diminuição da diversidade genética.

13 Importa fazer a distinção entre o tamanho da população e o tamanho efetivo da população, compreendendo este último a porção da população que de fato participa da reprodução contribuindo com gâmetas para a geração seguinte. Trata-se de algo difícil de determinar pelo que se considera como o tamanho de uma população idealizada com as mesmas características (quanto a deriva genética) que a população real em estudo e independentemente do censo populacional ou seja da contagem de indivíduos na população.

BIBLIOGRAFIA

Anderson, A.D., Weir, B.S. (2006). An assessment of the behavior of the population structure parameter, θ, at the CODIS loci. *International Congress Series,* 1288, 495– 497.

Brion, M., Quintans, B., Zarrabeitia, M., Gonzalez-Neira, A., Salas, A., Lareu, V.,Tyler-Smith, C., Carracedo, A. (2004). Micro-geographical differentiation in Northern Iberia revealed by Y-chromosomal DNA analysis. *Gene,* 329, 17-25.

Budowle, B., Ge, J. Y., Chakraborty, R., Eisenberg, A. J., Green, R., Mulero, J., Lagace, R., Hennessy, L. (2011). Population genetic analyses of the NGM STR loci. *International Journal of Legal Medicine*, 125 (1), 101-109.

Cavalli-Sforza, L.L. (1996). *Genes Povos e Línguas*. Lisboa, Instituto Piaget. ISBN: 9789727712632

Cockerton, S., McManus, K., Buckleton, J. (2012). Interpreting lineage markers in view of subpopulation effects. *Forensic Science International-Genetics*, 6 (3), 393-397.

Duran, J.A., Chabert, T., Rodrigues, F., Pestana, D. (2007). Distribuição dos Grupos Sanguíneos na População Portuguesa. *AB0* Número 29 Jan/Mar

Excoffier, L., Smouse, P.E., Quattro, J.M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131, 479-91.

Gill, P., Phillips, C., McGovern, C., Bright, J., Buckleton, J. (2012). An evaluation of potential allelic association between the STRs vWA and D12S391: Implications in criminal casework and applications to short pedigrees. *Forensic Science International: Genetics,* 6, 477–486.

Guo, S.G., Thompson E.A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics,* 48, 361-372.

Hamilton, M.B. (2009). Population Genetics. Wiley-Blackwell.

Holsinger, K.E., Weir, B.S. (2009). Fundamental Concepts in Genetics. Genetics in geographically structured populations: defining, estimating and interpreting F-ST. *Nature Reviews Genetics*, 10 (9), 639-650.

Lock, M., Nguyen, V. (2010). An Anthropology of Biomedicine. Oxford, Wiley-Blackwell.

McEvoy, B.P., Montgomery, G.W., McRae, A.F., Samuli, R., Perola, M., Spector, T.D. ...Visscher, P. (2009). Geographical structure and differential natural selection among North European populations. *Genome Research,* 19 (5), 804-814.

Nei M. (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences,* 70, 3321–3323.

O'Connor, K. L., Tillmar, A.O. (2012). Effect of linkage between vWA and D12S391 in kinship analysis. *Forensic Science International-Genetics,* 6 (6), 840-844.

Pakstis, A.J.; Speed, W.C.; Kidd, J.R.; Kidd, K.K. (2007). Candidate SNPs for a universal individual identification panel. *Human Genetics*, 121, 304–317.

Pereira, L., Ribeiro, F.M. (2009). O Património Genético Português, a história humana preservada nos genes. Lisboa, Gradiva.

Pertea, M., Salzberg, S.L. (2010). Between a chicken and a grape: estimating the number of human genes. *Genome Biology*: 11:206.

Pfaff, C.L., Parra, E.J., Bonilla, C., Hiester, K., McKeigue, P.M., Kamboh, M.I. ...Shriver, M.D. (2001). Population Structure in Admixed Populations: Effect of Admixture Dynamics on the Pattern of Linkage Disequilibrium. *American. Journal of Human Genetics*, 68:198–207.

Slatkin M. (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, 139, 457–462.

Taylor, D., Nagle, N., Ballantyne, K.N., Van Oorschot, R.A.H., Wilcox, S., Henry, J., Turakulov, R., Mitchell, R.J. (2012). An investigation of admixture in an Australian Aboriginal Y-chromosome STR database. *Forensic Science International-Genetics*, 6 (5), 532-538.

Walsh, S.J., Mitchell, R.J., Torpy, F., Buckleton, J.S. (2007). Use of subpopulation data in Australian forensic DNA casework. *Forensic Science International-Genetics*, 1(3-4), 238-246.

Weir, B. S. (1996). Genetic Data Analysis II. Sinauer Associates.

Weir B.S., Cockerham C.C. (1984) Estimating F-Statistics for the analysis of population structure. *Evolution*, 38, 1358–1370.

Wright, S. (1951). The genetical structure of populations. *Ann. Eugenics*, 15:159–171

Zhu, F., Tao, S., Xu, X., Ying, Y., Hong, X., Zhu, H., Yan, L. (2010). Distribution of AB0 blood group allele and identification of three novel alleles in the Chinese Han population. *Vox Sanguinis,* 98, 554-559.