ARTICLE

Population history and infrequent mutations: how old is a rare mutation? *GUCY2D* as a worked example

Sylvain Hanein¹, Isabelle Perrault¹, Sylvie Gerber¹, Nathalie Delphin¹, David Benezra², Stavit Shalev³, Rivka Carmi⁴, Josué Feingold¹, Jean-Louis Dufier⁵, Arnold Munnich¹, Josseline Kaplan¹, Jean-Michel Rozet¹ and Marc Jeanpierre^{*,6,7,8}

¹Unité de Recherches sur les Handicaps Génétiques de l'Enfant. Hôpital Necker – Enfants Malades, Paris, France; ²Hadassah University Hospital, Jerusalem, Israel; ³Technion – Israel Institute of Technology, Haifa, Israel; ⁴Genetic Institute Soroka Medical Center, Beer-Sheva, Israel; ⁵Service d'ophtalmologie, Hôpital Necker, France; ⁶Institut Cochin, Université Paris Descartes, CNRS (UMR 8104), Paris, France; ⁷Inserm, U567, Paris, France; ⁸AP-HP, Hôpital Cochin, Service de Génétique, Paris, France

The mosaic pattern of haplotypes observed around a single mutation results from one or several founder events. The difficulties involved in calculating the age of the variant are greatly reduced by assuming a single event, but this simplification may bias analysis of the genealogy of the mutation. However, if it is assumed that more than one founder event occurred, the number of genealogies is very large and the likelihood of every possible tree could not be realistically calculated. A multipoint approach is required, given the number of independent variables needed to describe a complex bifurcating genealogy. Starting from the observation that a limited number of parameters is needed for calculation of the simplest models of bifurcating genealogies, we show that the probability density of a two-ancestor model genealogy can be simply described as an algebraic function in a closed form, two coalescence times being calculated simultaneously without compromising accuracy. Implementation in a Bayesian framework is facilitated by the simplicity of the function, which describes the reciprocal relationship between the region of complete linkage disequilibrium and the branch length of the tree. We illustrate the use of haplotype information about allele-sharing decay around a mutation as a genetic clock, using data for two *GUCY2D* mutations in Mediterranean populations.

European Journal of Human Genetics (2008) 16, 115–123; doi:10.1038/sj.ejhg.5201905; published online 8 August 2007

Keywords: mutation; founder events; history; genealogy; bayesian

Introduction

The size of ancestral haplotypes around a mutation is inversely correlated to the time, in generations to the common ancestor. Once the mutation has been introduced into the population, recombination and mutation in

subsequent generations break down the ancestral haplotype at nearby markers. The rationale is apparently simple: small haplotypes indicate ancient mutations. However, the apparent simplicity of this rationale conceals real difficulties in the modeling of haplotype-sharing decay for deciphering past events.^{1,2} There are two main problems: determination of the time to the most recent common ancestor (MRCA) if the genetic locations of the variant and of the markers are known and the estimation of the location of a variant.¹ Specific methods are required to address these two problems, which represent very different challenges. For example, genetic heterogeneity – a

*Correspondence: Dr M Jeanpierre, Service de Génétique Moléculaire, Groupe Hospitalier Cochin, 123 Boulevard de Port Royal, Paris F-75014, France.

Tel: +33 1 5841 1624; Fax: +33 1 5841 1580;

E-mail: jeanpierre@cochin.inserm.fr

Received 10 November 2006; revised 11 May 2007; accepted 26 June 2007; published online 8 August 2007

well-known obstacle to fine-scale genetic mapping by linkage disequilibrium (LD) – is not generally a problem when trying to date a specific mutation.³ As LD is used to map complex diseases, large samples are analyzed and efficient algorithms do not necessarily have to be expressed algebraically.^{4,5}

Calculating the age of a mutation is greatly simplified by assuming that all present day haplotypes descended from a single common ancestor, with all branches of equal length and evolving independently. Such ‘star-shaped’ or ‘star-like’ genealogies are representative of samples taken from exponentially growing populations and are by far the easiest to analyze, as this model bypasses all the difficulties of translating a set of mosaic haplotypes into a specific genealogy for a genetic variant. Branch length is the only parameter required to describe a star-like genealogy. The assumption that the entire genealogy arose at a single time point in the past leads to a simple formula relating the proportion of identical-by-descent (IBD) alleles and the age of the mutation.^{6–9} Star-shaped genealogies underestimate the number of generations from the common ancestor, and correction factors have been proposed to give more realistic figures for actual age.^{10,11} Assuming a single ancestor being an unsatisfactory approximation, testing of the fit to the single ancestor model have been proposed.¹² The question of the genealogy of a mutation is of particular importance if the mutation concerned is thought to have been subjected to recent positive selection, because unexpectedly long haplotypes for a frequent mutation are thought to be a genetic signature of recent positive selection.^{13–15} The resolution of time intervals for possible population expansions requires the demonstration of star genealogies rather than their prior assumption.

We need to describe the genealogy of a mutation appropriately, according to the level of accuracy desired. Backwards inference from present haplotypes to past processes is difficult owing to the large number of possible genealogies and the stochastic nature of haplotype decay. As an illustration of the complexity of the problem, 10 haplotypes would give rise to 104 topologically different trees, and for each shape of tree, nine branch lengths must be evaluated ($n-1$ nodes representing root and intermediate founders). The number of parameters needed to describe all the possible genealogies of a mutation is generally much larger than the number of independent parameters obtained from the analysis of genetic markers.

We consider here the simplest bifurcating model – a two-ancestor model. The probability density of haplotype block size is shown to be defined in a closed form, facilitating its implementation in a multipoint Bayesian algorithm. In populations with a long tradition of marriages between relatives, secondary founder effects appear as large regions of IBD alleles. We used *GUCY2D* – the gene most frequently implicated in our series of patients with Leber congenital amaurosis (LCA, MIM 204000) – to illustrate our methods,

focusing on two recessive mutations present in Mediterranean populations. This scenario may correspond to an approach midway between oversimplistic and overcomplicated genealogical models.

Methods

Families

Thirty-six unrelated families originating from North Africa and one family from Portugal were screened for mutations in all known LCA genes. All patients fulfilled the minimal criteria for the diagnosis of LCA described elsewhere.¹⁶ Twenty-two of the 36 North African families (16 consanguineous) were from Algeria, seven (all consanguineous) were from Tunisia, five (all consanguineous) were from Morocco, and two (both consanguineous) were from Egypt.

Genotyping

DNA samples available from individuals belonging to the seven families harboring the 387delC mutation were subjected to PCR amplification, using primers specific for exon 2 of the *GUCY2D* gene, as described elsewhere.¹⁷ Intronic primers were used to amplify the 8th exon of the *GUCY2D* gene of all members of the four unrelated families carrying the Phe565Ser mutation for whom DNA samples were available. The purified PCR fragments were directly sequenced, using the Big Dye Terminator Cycle Sequencing Kit. Single nucleotide polymorphism (SNP) analysis was carried out and the segregation of two *GUCY2D* SNPs (SNP; 531G>T, 12717T>C; Genbank accession no. AJ222657) was studied in families harboring either the 387delC microdeletion ($n=7$) or the Phe565Ser missense mutation ($n=4$). PCR amplifications were carried out with the following primers.

ForwardG531T: (5′–3′) CATGGGTTACTCGGGCTTGA GAAA; reverse531G>T: (5′–3′) GAGAGAAGATGGGGTCG CAAGCCCA and forward12717T>C: (5′–3′): TGCTCCCTG TCCCATCTG; reverseT12717C: (5′–3′): AGACAGTATGCC TTTATTTCAC.

An annealing temperature of 60°C was used for both 531G>T and 12717T>C. The 315 bp (for 531G>T) and 200 bp (for 12717T>C) fragments amplified were directly sequenced.

Genetic markers flanking the LCA1 locus were studied for LD in families carrying one of these two mutations (Hanein *et al*, 2004¹⁷). The position of the markers (with physical distances in megabases, shown in parentheses) were estimated from human genome working draft data available from the University of California, Santa Cruz (UCSC). We also studied *GUCY2D* SNPs, selected based on their position within the gene. The (c.531G>T) SNP was located 229 bp upstream from the 387delC mutation and 11957 bp downstream from the Phe565Ser mutation, whereas the (c.12717T>C) SNP was located 1163 bp

upstream from the 387delC mutation and 11023 bp downstream from the Phe565Ser mutation (Table 1).

Statistical analysis

Two-ancestor model problem

Comparing a trio of haplotypes side by side, the sizes of the complete LD region and the partial disequilibrium region can be determined on both sides of the mutation (Figure 1a and b). We first assume that a high density of markers clarifies the question of the intervals, and that allele sharing implies IBD. Identity-by-state (IBS) will later be taken into consideration for the calculation of haplotype density. If the same pair of haplotypes shares the longest track of identical alleles on both sides of the mutation, the LD pattern is said to be concordant (Figure 1c).

Haplotype core boundary

The existence of a region of identical alleles surrounding the mutations indicates a common origin of present-day mutation-carrying haplotypes. The size of this region in complete LD is represented by the distribution of the most proximal recombination event. The probability that no recombination occurred in an interval θ after m meioses is $(1-\theta)^m$, whereas the probability that a single recombination occurred exactly at a distance θ is proportional to $(1-\theta)^{m-1}$. The distribution normalized to unity over the interval $[0,1]$ is

$$\int_0^1 m(1-\theta)^{m-1} d\theta = 1 \quad (1)$$

The number of meioses m is $2n_1 + n_2$ in our two-ancestor model problem. As any given value of m may result from several combinations of n_1 and n_2 , the size of the haplotype core (region in complete LD) cannot resolve the two branch parameters unambiguously. Distal and proximal boundaries must be considered together to obtain insight into the history of the mutation. The power function,

$(1-\theta)^n$, describing the probability of there being no recombination at a genetic distance θ from a mutation is usually approximated by the exponential $e^{-n\theta}$, with no notable consequences, when θ is small and n is large. As normalization constants, which are required to ensure that the probabilities sum to 1, are easier to calculate for power functions than for exponential functions, the expression of the base case has not been simplified unnecessarily. The relationship between the number of meioses and distance from the mutation is best visualized as a 3-dimensional plot (Figure 2).

Distal haplotype boundary

We will first consider the construction of a three-branch tree from a simple two-branch model (Figure 3). If we start from a pair of haplotypes and add a third chromosome, two possibilities must be considered; the third sequence may be smaller or larger than the homologous region of

Figure 2 Representation of the relationship between the number of meioses and the size of the region in complete linkage disequilibrium. Each orthogonal point of view represents either the probability of haplotype block edge expressed for the interval (1, 5 cM) or the probability of ancestry expressed for the interval (0, 100 meioses). In a Bayesian framework, the recombination interval is known and the number of generations is calculated.

Figure 1 Schematic representation of the two-ancestor problem. (a) The genealogy of the mutation is defined by two coalescence times (n_1 , n_2) and the position of the node. The left to right arrangement of haplotypes is taken into account. (b) On each side of the mutation, the parameters studied are the size of the linkage disequilibrium regions, either complete (θ_{1a} and θ_{1b}) or partial (θ_{2a} and θ_{2b}). Dark gray bars: alleles inherited from ancestor n_1 . Light gray bars: alleles inherited from ancestor n_2 . One of the several possible combinations is shown. (c) The likelihood that the distal LD boundaries concern the same pair of haplotypes on both sides of the mutation is a function of the ratio n_2/n_1 . Concordance of distal haplotype blocks suggests a low n_2/n_1 ratio. Left, concordant pattern; right, discordant pattern.

Figure 3 A graphical approach to the construction of probability density functions. (a) A branch is added to a pair of haplotypes, corresponding to the simplest possible tree. (b) The dotted line represents the power (approximately exponential) function for the common region of two haplotypes. The solid line represents the common region of three haplotypes. The dashed line represents the distal boundary function, expressed as the difference between two power functions. The X-axis is the distance from the mutation, and the Y-axis is a dimensionless quantity.

the initial pair. As these probabilities are alternative events, they are complementary. Since two of the three probability density functions are described by equation (1), the last function corresponds to the difference between the other two. When the three branches are of identical length, the distribution of the resulting tree is

$$6n((1-\theta)^{2n-1} - (1-\theta)^{3n-1}) \quad (2)$$

The first factor ($6n$) is the normalizing constant making the integral of the density function equal to 1.

In the general situation in which the third branch is not identical to the first two, three haplotype patterns can be recognized, according to the position of the critical recombination events in the mutation tree (Figure 4a and b). The density probabilities for these three possibilities are

$$\text{X1} \quad 2n_2((1-\theta)^{2n_2-1} - (1-\theta)^{n_1+n_2-1}) \quad (3)$$

$$\text{X2} \quad 2n_2((1-\theta)^{n_1+n_2-1} - (1-\theta)^{m-1}) \quad (4)$$

$$\text{YZ} \quad 4n_1((1-\theta)^{2n_1-1} - (1-\theta)^{m-1}) \quad (5)$$

Patterns X1 and X2 are distinguished here only for the calculation of the proportion n_1/n_2 in the mosaic haplotype. By simplifying Equations (3) and (4), we obtain the density function (normalized to unity) of the distal edge:

$$4n_1((1-\theta)^{2n_1-1} - (1-\theta)^{m-1}) + 2n_2((1-\theta)^{2n_2-1} - (1-\theta)^{m-1}) \quad (6)$$

The complete LD region and distal edges result from shared stochastic events and cannot not be considered to be independent variables. The probability of observing both quantities θ_1 (complete LD region size) and θ_2 (partial LD region size) must be calculated from transitional probability expressions. The probability generating functions as a function of relative branch length are $(1-\theta_1)^{2n_1-n_2-1}$ and $(1-\theta_1)^{n_2-1}$ for the X and YZ conditions, respectively (Figure 4c). When all branches

have the same length, the density probability of n is

$$2n^2(1-\theta_2)^{2n-1}(1-\theta_1)^{n-1} \quad (7)$$

The base case functions being expressed in a closed form, the modal figure (maximum likelihood of number of generations) is simply calculated from the derivative, as the expression

$$(1-\theta)^{2n-1} - (1-\theta)^{3n-1}$$

gives rise to the derivative

$$\partial_n = 2(1-\theta)^{2n-1} \log[1-\theta] - 3(1-\theta)^{3n-1} \log[1-\theta]$$

The value of n that passes through zero at the peak's inflection points is

$$n = \frac{\log[2/3]}{\log[1-\theta]}$$

This is Risch's formula, which relates the proportion of identical alleles and the number of generations from the MRCA.⁶ This derivation also holds for determination of the proximal edge in larger samples, as for $n-1$ out of n haplotypes.

Allele sharing: IBS or IBD?

Coalescence times may be severely underestimated if allele sharing is assumed to be accounted for by IBD alone.¹⁰ For each marker, the probability of observing, by chance, x identical alleles of a total of z alleles if the allele frequency is f , follows a truncated binomial distribution with parameter f . These probabilities, calculated from present-day allele frequencies, are taken as prior probabilities of IBS. A synthetic picture of coalescence time density functions is obtained as a compound probability for every proportion of identical alleles actually corresponding to IBD.

Results

The results of the haplotype analysis of the markers flanking the *GUCY2D* gene are shown in Table 1. The

Figure 4 Haplotype patterns and their probabilities. (a) On each side of the mutation, three possible patterns are recognized. X1 and X2 patterns are the 'natural' patterns, where the shortest haplotype corresponds to the longest branch (X1 ancestor 2 determines the common region; X2 ancestor 1 determines the common region). YZ is the pattern and probability when the shortest segment is either one of the two haplotypes at the tip of the forked branch. (b) Density probability of the three possible patterns (X-axis: recombination fraction, Y-axis: arbitrary units). (c) Graphical representation of the expressions of the conditional probabilities. Solid lines with no symbols indicate the theoretical expectations for the proximal edge (θ_1), given the distal edge (θ_2). Circles and squares indicate simulations for $\theta_2 = 0.1$ and $\theta_2 = 0.15$ respectively simulations ($n_1 = 6$; $n_2 = 4$). Unblackened and blackened symbols indicate the position of the recombinations for the X and YZ patterns respectively.

families are grouped according to mutation, delC or F565S. All patients carry two copies of the same mutation, and none is a compound heterozygote. As would be predicted for patients born to first or second cousins, F565S carriers presented long blocks (up to 20 cM) of identical alleles. Although all F565S-bearing families originate from the same region, we found no genealogical evidence for a common ancestor of two patients. Haplotype analysis of the markers flanking the *GUCY2D* gene suggested a common founder for each mutation.

Pairwise analysis of allele-sharing decay, using Risch's formula, provided different coalescence numbers, according to the marker considered. For delC, the age of the common ancestor with two-point methods varied from 25 to 146 generations. For F565S, estimation of the age of the common ancestor with a two-point formula, using the

proportion of identical alleles as a parameter, gave even more variable results for the number of generations elapsed. The long region of apparent allele identity observed for each patient and the familial context of marriage between cousins suggest that a grand parent or a great grand parent was probably an intermediate ancestor. Based on this hypothesis, the number of haplotypes to be analyzed can be reduced from eight to four, and the calculation with the pairwise method gives variable coalescence times extremely sensitive to the choice of the marker analyzed (20–110 generations).

Analyzing every possible haplotype trio confirmed the hypothesis of intermediate ancestors (Figure 5a). Averaging over all possible trios is not appropriate, because not all trios of a sample can have the same weight for calculation of the root of the tree. A branch-and-bound approach,

Table 1 Haplotype analysis of the markers flanking the *GUCY2D* gene

Families		Individual	Tel																			Cen																							
			AFMc024aa5-D17S1866 (82514 pb)	494469	AFM207aa1-D17S926 (576983)	419127	AFM022xb6 - D17S1529 (996110)	861040	AFM038x4 - D17S831 (1837150)	2434620	AFMa061za9 - D17S1876 (4291770)	1898222	AFM263wb5 - D17S938 (6189992)	2254	AFM177hb6 - D17S796 (6192246)	81397	AIPL1 (6273643)	1454245	AFMA183YE1-D17S1796 (7727888)	119360	c.531 G>T (7847248)	229	c.387delC (7847477)	1234	p.Phe56Ser (7848711)	10723	c.12717 T>C (7859434)	151948	RETGC147KB (8011382)	191184	AFMA298WA9-D17S1812 (8202566)	894584	AFMa134x5 - D17S1791 (9097150)	708951	AFM225ZC1-D17S804 (9806101)	650125	AFMb336zg9 - D17S1852 (10456226)	2655455	AFM192yH - D17S799 (13111681)	1089643	AFM191xh12 - D17S921 (14201324)				
1	1												178	144	288		183	G	del	T		306	156	286	162	216																			
	2												178	144	288		183	G	del	T		306	160	270	158	200																			
2	1												178	143	288		183	G	del	T		306	156	286	162	202																			
	2												178	143	290		183	G	del	T		306	160	270	158	214																			
3	1												174	164	304		183	G	del	T		306	148	276	156	210																			
	2												180	170	290		183	G	del	T		306	148	276	158	210																			
	3												174	164	304		183	G	del	T		306	148	276	166	202																			
	4												180	170	290		183	G	del	T		306	148	276	158	210																			
4	1												180	144	288		183	G	del	T		306	156	280	158	206																			
	2												172	164	292		183	G	del	T		306	160	270	162	192																			
5	1												178	170	290		183	G	del	T		306	160	270	162	200																			
	2												178	170	288		183	G	del	T		306	156	280	160	202																			
6	1												164	168	288		183	G	del	T		306	160	272	158	210																			
	2												180	166	288		183	G	del	T		306	158	270	162	202																			
7	1												166	156	280		183	G	del	T		306	158	274	158	194																			
	2												172	164	288		183	G	del	T		308	158	278	162	206																			
A	1			165	238	106	167	164	286									G		mt	T		314			272		200																	
	2			171	238	106	167	164	286									G		mt	T		314			272		206																	
	3			165	238	106	167	164	286									G		mt	T		314			272		200																	
	4			171	238	106	167	164	286									G		mt	T		314			272		206																	
B	1			165	238	106	167	164	286									G		mt	T		314			272		200																	
	2			171	238	106	167	164	286									G		mt	T		314			272		206																	
	3			165	238	106	167	164	286									G		mt	T		314			272		200																	
	4			171	238	106	167	164	286									G		mt	T		314			272		206																	
C	1												166	164	306			G		mt	T		314			282		200																	
	2												164	166	290			G		mt	T		314			282		200																	
	3												164	166	290			G		mt	T		314			274		198																	
	4												164	166	290			G		mt	T		314			282		200																	
D	1												164	166	290			G		mt	T		314			282		200																	
	2												166	166	290			G		mt	T		314			282		200																	
	3												234	108	176	170	294			G		mt	T	308		278		207		188		177													
	4												228	108	176	170	294			G		mt	T	308		278		207		192		183													
E	1												234	108	176	170	294			G		mt	T	308		278		207		188		177													
	2												228	108	176	170	294			G		mt	T	308		278		207		192		183													
	3												228	108	176	170	294			G		mt	T	308		278		207		188		177													
	4												228	108	176	170	294			G		mt	T	308		278		207		192		183													
F	1			166	242	154	228	106	179	162	288							G		mt	T		314			282		200																	
	2			166	242	154	228	106	179	162	288							G		mt	T		314			282		200																	

using the density functions described in the methods section as a skeleton, reduces the number of haplotypes to be considered as a first step and then calculates the time, in generations, to the MRCA, using haplotype sizes as parameters (Figure 5b).

Discussion

This paper presents a new approach for constructing mutation genealogies from present-day haplotypes. Using the simplest branched model as a starting point, exact calculations can be made, using the size of the preserved regions around the mutation.

When the ancestral haplotype is known, the expected genetic distance from either edge of the ancestral haplotype is simply the reciprocal of the age of the variant.¹ In practice, the ancestral haplotype is not known, and

determining whether any block of alleles actually represents the ancestral sequence of the root ancestor or an intermediate ancestor is subject to the same problems as the direct calculation of coalescence times. The core region around the mutation therefore seems to be the best region to start from, as this region in complete LD almost certainly correspond to the sequence of the root ancestor. The size of the region in complete LD reflects only the sum of the number of meioses in each branch, and is an imperfect estimator of time to the common ancestor. The number of meioses accounting for the distribution of the size of the complete LD block is sensitive to sample size (number of haplotypes studied): large samples may have very small core haplotypes that nonetheless correspond to a recent mutation; the mean ancestor haplotype length is however insensitive to sample size.¹⁸ Furthermore, the uncertainty associated with the size of the region in complete LD is always large. This uncertainty is explained

Figure 5 Ancestry of the *GUCY2D* mutations. (a) Calculation of the n_2/n_1 ratio for a trio of haplotypes (delC, haplotypes 1, 3, and 14, Table 1). (b) Compound probability of time to common ancestor. The dashed line indicates the number of generations for the mutation F565S and the solid line the number of generations for the mutation delC.

by the dimension of the density function, which is proportional to $ne^{-n\theta}$, whereas the distribution of haplotypes approximates to a gamma function in star-like genealogies,¹ and is therefore proportional to $n^xe^{-xn\theta}$ and represented by a much narrower density function in large samples.

Risch's formula describes the relationship between the proportion of shared alleles for a single marker of known distance from the mutation and the number of generations from the common ancestor. As Risch's formula is a moment estimator method, we would expect to find a difference between the mode and the mean if the distribution has an asymmetric shape. Analysis of the relationship between the modal and mean values of the unnormalized distribution of the distal edge provided an unexpected link between our Bayesian framework and Risch's approach.

In a mutation history tree, branch length represents time of successive meioses. In our model with only a small number of nodes, a star-like or star-shaped genealogy is not assumed, but may result from a calculation in which the two coalescence times are found to be very similar. As our two-ancestor equations could be transposed to the broader situation of the two most proximal edges of haplotype-sharing regions for larger samples, it is tempting to extrapolate the equations to the case of a proportion of shared alleles of any size with more than two ancestors. However, such model-free formulae are unfortunately unattainable because the expected distribution of haplotypes always depends on an underlying model of genealogy and model-free formulae must therefore represent every possible tree for a given number of chromosomes. An absolute calculation is at odds with the inevitable uncertainty of experimental parameters (recombination

rates, allele, and mutation frequencies) and is mostly not required and difficult to represent.

The probability of a mutation genealogy is easier to understand in terms of a limited number of parameters *ad hoc* algebraic equations based on simplified genealogies could be constructed for any specific shape of tree because we can derive an algebraic correspondence between tree shapes and IBD blocks. Inference about mutation history is likely to concern only a few ancestors near the root of the tree, suggesting that the 'small number of ancestors' model may provide the framework of a heuristic trade-off between accuracy and simplicity. Analyses based on haplotype size naturally leads to a multipoint method, the distribution of the proximal and distal edges being sufficient for the calculation of two branch lengths at a time. Considering the entire haplotype leads to more robust estimates than pairwise disequilibrium.¹⁹

Bayesian procedures require specification of prior distributions for all the parameters.²⁰ We implicitly assume in our analysis that the prior age is drawn from a uniform distribution. This assumption of a uniform prior distribution is appropriate for a rare mutation because the large frequency variations associated with the small number of mutation carriers usually generate an approximately flat distribution. A well-known relationship links the frequency of a DNA variant and effective population size.²¹ For rare variants, the expected figure is often blurred by stochastic variations of variant frequencies. Known distributions from the same population could be used to define the prior distribution required for a fully Bayesian approach.

As most recessive disease-causing mutations are recent, such mutations are a potentially valuable tool for probing the recent molecular history of populations. Branching

points are more likely to occur when the population is expanding, and are surrogate markers for such periods. When the frequency of a DNA variant is large and accurately known, the estimated age of this variant may be used to calculate the population expansion coefficients because we have both the size of the population and the time needed to reach the observed frequency.

In this paper, we have analyzed genotypic data from a series of families with *GUCY2D* mutations. The 387delC and F565S mutations were observed in North African and Southern European (Portuguese) patients recruited from five different countries. Founder effects may account for this high prevalence of the two mutations analyzed here in North African patients. No compound heterozygotes were observed. This observation suggests that mutation frequency was variable in the regions considered, possibly due to secondary founder effects. All carriers of the 387delC mutation had a conserved haplotype over 1-cM interval containing the *GUCY2D* locus. As mean ancestor haplotype length is the inverse of time to the common ancestor, the apparent midpoint of haplotypes upon visual inspection provides a first approximation of ancestry (1 cM = 100 generations). Unfortunately, haplotype midpoint is only a proxy for the mean residual ancestral haplotype, one of the most remarkable figures in the molecular history of a mutation because this mean haplotype is sensitive neither to sample size nor to a specific genealogical tree.¹⁸ The key challenge is not obtaining a crude estimate of ancestry based on visual inspection, but estimating the precision of the calculation of branching points as proxies for periods of population expansion. The branch-and-bound algorithm groups together similar haplotypes and reduces the number of branches. The final step calculates the expected number of generations for these independent branches. This final heuristic step is justified by the simplicity of the algebraic description of unconnected branches. The precision of the estimation is limited by the uncertainty about the local variations of the sex-average genetic map: more recombinations are expected from female ancestors than male ancestors, these differences of the male-specific and female-specific map are expected to be limited on the long run. As recombination hotspots occur at intervals of 200 kb or less on average,²² fine-scale variations in recombination rates limit the time span of methods based on haplotype-decay models.

Selection against homozygotes has a limited impact on the calculation, even in populations with high inbreeding coefficient, because the fraction of mutations removed at each generation is approximately proportional to a usually low gene frequency. After 100 generations, a mutation of frequency 0.01 in a population of inbreeding coefficient 0.02 is affected by less than 2%.

In this study, we have described the density function of haplotype block edges in a simple bifurcating model. Because it is difficult to translate haplotype

sequences into molecular history, genealogies of intermediate complexity, with a small number of ancestors, represent a heuristic middle-ground between simple but unrealistic star-like trees and general models of untractable complexity.

Acknowledgements

We especially thank the families for their participation in this study. We thank Julie Sappa for useful suggestions.

References

- McPeck MS, Strahs A: Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am J Hum Genet* 1999; **65**: 858–875.
- Slatkin M, Rannala B: Estimating allele age. *Annu Rev Genomics Hum Genet* 2000; **1**: 225–249.
- Rannala B, Bertorelle G: Using linked markers to infer the age of a mutation. *Hum Mut* 2001; **18**: 87–100.
- Morris AP, Whittaker JC, Balding DJ: Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. *Am J Hum Genet* 2004; **74**: 945–953.
- Morris AP, Whittaker JC, Balding DJ: Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am J Hum Genet* 2002; **70**: 686–707.
- Risch N, de Leon D, Ozelius L *et al*: Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. *Nat Genet* 1995; **9**: 152–159.
- Guo SW, Xiong M: Estimating the age of mutant disease alleles based on linkage disequilibrium. *Hum Hered* 1997; **47**: 315–337.
- Colombo R: Age estimate of the N370S mutation causing Gaucher disease in Ashkenazi Jews and European populations: A reappraisal of haplotype data. *Am J Hum Genet* 2000; **66**: 692–697.
- Serre JL, Simon-Bouy B, Mornet E *et al*: Studies of RFLP closely linked to the cystic fibrosis locus throughout Europe lead to new considerations in populations genetics. *Hum Genet* 1990; **84**: 449–454.
- Labuda D, Zietkiewicz E, Labuda M: The genetic clock and the age of the founder effect in growing populations: a lesson from French Canadians and Ashkenazim. *Am J Hum Genet* 1997; **61**: 768–771.
- Labuda M, Labuda D, Korab-Laskowska M *et al*: Linkage disequilibrium analysis in young populations: pseudo-vitamin D-deficiency rickets and the founder effect in French Canadians. *Am J Hum Genet* 1996; **59**: 633–643.
- Rosenberg NA, Hirsh AE: On the use of star-shaped genealogies in inference of coalescence times. *Genetics* 2003; **164**: 1677–1682.
- Vallender EJ, Lahn BT: Positive selection on the human genome. *Hum Mol Genet* 2004; **13** (Spec No. 2): R245–R254.
- Sabeti PC, Reich DE, Higgins JM *et al*: Detecting recent positive selection in the human genome from haplotype structure. *Nature* 2002; **419**: 832–837.
- Bersaglieri T, Sabeti PC, Patterson N *et al*: Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* 2004; **74**: 1111–1120.
- Perrault I, Rozet JM, Gerber S *et al*: Leber congenital amaurosis. *Mol Genet Metab* 1999; **68**: 200–208.
- Hanein S, Perrault I, Gerber S *et al*: Leber congenital amaurosis: comprehensive survey of the genetic heterogeneity, refinement of the clinical definition, and genotype–phenotype correlations as a strategy for molecular diagnosis. *Hum Mutat* 2004; **23**: 306–317.

- 18 Piccolo F, Jeanpierre M, Leturcq F *et al*: A founder mutation in the gamma-sarcoglycan gene of gypsies possibly predating their migration out of India. *Hum Mol Genet* 1996; **5**: 2019–2022.
- 19 Liu JS, Sabatti C, Teng J, Keats BJ, Risch N: Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res* 2001; **11**: 1716–1724.
- 20 Beaumont MA, Rannala B: The Bayesian revolution in genetics. *Nat Rev Genet* 2004; **5**: 251–261.
- 21 Kimura M, Ohta T: The age of a neutral mutant persisting in a finite population. *Genetics* 1973; **75**: 199–212.
- 22 McVean GA, Myers SR, Hunt S *et al*: The fine-scale structure of recombination rate variation in the human genome. *Science* 2004; **304**: 581–584.