Copyright © 2003 by the Genetics Society of America

Detecting Population Growth, Selection and Inherited Fertility From Haplotypic Data in Humans

Frédéric Austerlitz,*[,1] Luba Kalaydjieva[†,‡] and Evelyne Heyer[§]

*Laboratoire Ecologie, Systématique et Evolution, Université Paris-Sud, F-91405 Orsay, France, [†]Centre for Human Genetics, Edith Cowan University, Perth, Australia WA 6027, [‡]Western Australian Institute for Medical Research, Perth, Australia WA 6027 and [§]Centre National de la Recherche Scientifique—Laboratoire d'Anthropologie Biologique, Musée de l'Homme (MNHN), F-75116 Paris, France

Manuscript received February 13, 2003
Accepted for publication July 2, 2003

ABSTRACT

The frequency of a rare mutant allele and the level of allelic association between this allele and one or several closely linked markers are frequently measured in genetic epidemiology. Both quantities are related to the time elapsed since the appearance of the mutation in the population and the intrinsic growth rate of the mutation (which may be different from the average population growth rate). Here, we develop a method that uses these two kinds of genetic data to perform a joint estimation of the age of the mutation and the minimum growth rate that is compatible with its present frequency. In absence of demographic data, it provides a useful estimate of population growth rate. When such data are available, contrasts among estimates from several loci allow demographic processes, affecting all loci similarly, to be distinguished from selection, affecting loci differently. Testing these estimates on populations for which data are available for several disorders shows good congruence with demographic data in some cases whereas in others higher growth rates are obtained, which may be the result of selection or hidden demographic processes.

SEVERAL methods have been designed to infer past population history from molecular data (Tajima 1989; Rogers and Harpending 1992). However, events of different nature, in particular population growth and selective sweep, can leave a similar signature in a given gene. Thus, any method designed to detect selection can be used to detect population growth. For example Tajima's (1989) *D* test, originally designed to test for selection, has been widely used to detect population expansion.

The only means to discriminate between population expansion and selection is to examine several independent portions of the nuclear genome (Nielsen 2001). Demographic events leave the same signature on all genes, whereas a selective sweep will affect only the gene (and surrounding part of the genome by hitchhiking) under selective pressure. Observing the same pattern at all loci is the indication of a demographical event, whereas a single locus that stands out is likely to have been subjected to a selective event.

While this difference helps to untangle demographic from selective effects, it does nothing against the fact that different demographic processes can leave the same signature. For instance, fertility inheritance in a stationary population will, in some aspects, affect the coalescent tree in a similar way as population growth (Sibert *et al.* 2002). By fertility inheritance, we mean that a positive correlation is observed between the number of effective children of an individual and the number of effective children of his/her parents, the effective children being the children that reproduce in their own population. The availability of demographic data on the Saguenay-Lac saint-Jean (SLSJ) population has made it possible to measure fertility inheritance in the human population of SLSJ in Quebec (Austerlitz and Heyer 1998) and to assess its impact on the frequency of rare alleles and its effect on haplotypic diversity and allelic association (Austerlitz and Heyer 2000).

Most methods that aim at detecting demographic events like expansions are sensitive to the long-term history of the population, since past expansions leave a stronger signal on the molecular data, making recent demographic events difficult to detect (Lavery *et al.* 1996; Austerlitz *et al.* 1997). Thus, since the frequency of recently introduced monogenic inherited disorders in a population and the level of association between the disease allele and alleles at closely linked markers are affected only by recent history, they are very useful for inferring the recent history of the population. An independent assessment of the growth rate of several disease genes will allow the identification of any single gene that stands out with a much higher growth rate and is therefore likely to have been submitted to a selective event. If all genes show an estimated growth rate higher than what is known from demographic data or what is realistic for the population under study, then it is likely that a spe-

[1]*Corresponding author:* Laboratoire Ecologie, Systématique et Evolution, UMR CNRS 8079, Université Paris-Sud, Bâtiment 362, F-91405 Orsay Cedex, France.
E-mail: frederic.austerlitz@ese.u-psud.fr

cific demographic event, like inherited fertility, is occurring in the population.

A problem in estimating the growth rate from this kind of data is that the frequency of an inherited disorder and the level of allelic association with surrounding markers are sensitive to the assumed age of the mutation in the population. Since this age is usually unknown, it becomes a nuisance parameter for estimating the growth rate correctly. Here, we present a method that overcomes this difficulty by estimating jointly the age of the allele and the growth rate.

The principle of this new method is as follows. Two kinds of information can be used to infer the history of a given disorder: the number of copies of the mutant allele in the present population and the level of allelic association between this allele and surrounding marker loci. Concerning the number of copies, Thompson and Neel (1978) provide a simple method to evaluate the probability, for an allele introduced as a single copy, to reach a given frequency in the present population, given its growth rate. Thus they can estimate the growth rate of the population (or of the allele, if its intrinsic growth rate is different from that of the population) that is compatible with the present frequency in the population, provided that the time of introduction (through migration or mutation) of the allele in the population is known.

The most appropriate tool for estimating the time of introduction is the genetic clock (Labuda *et al.* 1997; Colombo 2000), namely the decay of allelic association through time. Using the proper Luria-Delbrück correction (Luria and Delbrück 1943), the age of the mutation in the population can be estimated from haplotypic data. This method requires knowledge of the population growth rate.

Our method combines the two methods described above. Using both the present allelic frequency of the disorder and the level of allelic association with surrounding markers, we perform a joint maximum-likelihood estimation of the age of a mutation and the population growth rate compatible with the data, assuming neutrality. To increase the performance of the genetic clock, we correct the formula used in Labuda *et al.* (1996), removing an approximation that is not valid in some cases, and develop a multipoint estimate of the age of the mutation, using all the information provided by the markers that make the haplotype. We compare our results with the coalescent-based method of Slatkin and Bertorelle (2001), which estimates population growth rate using the same kind of data, and with Reeves and Rannala's (2002) method, which estimates the mutation age. The inferred growth rates can thereafter be compared either with estimates from other mutations in the same population or with known independent demographic data, when available. We have performed this analysis on several populations (Finns, Ashkenazi Jews, French Canadians from SLSJ, and East European Gypsies) that have been widely used, due to their recent founding and subsequent isolation, to locate severe single-gene disorders. To test the applicability of the method on a larger scale, we have applied it also to the CCR5-Δ32 AIDS resistance allele in Europe.

MATERIALS AND METHODS

General presentation: Assume a population with discrete generations, with growth rate r. Assume also a rare allele at a given locus (usually a disease gene), denoted D, which appeared g generations ago in the population by mutation or migration. The carrier frequency p of this allele in the population can be estimated, for instance, from a genetic epidemiology survey. Assume also that a sample of n chromosomes carrying D have been genotyped for one or several neutral marker loci, closely linked to D. Along with D, these markers define a haplotype of size θ. Because the mutation is recent in the population, allelic association (Collins and Morton 1998) will be observed between D and the neutral loci: allelic frequencies at the marker loci among the chromosomes carrying the disease allele at D will be different from the allelic frequencies in the rest of the population. Among the n individuals, some will carry the ancestral haplotype that has not been subject to any recombination, while others will carry a recombinant haplotype, and thus share none or only a part of the alleles carried by the ancestor at the marker loci.

As we see below, both the carrier frequency (p) of the disease allele and the number of carriers of the different haplotypes depend on r, g, and the recombination rates between the different loci. Knowing the recombination rates (from the genetic maps or independently studied pedigrees), it is thus possible to jointly estimate r and g from the genetic data. The method that we present below combines the formula that gives the probability (thereafter denoted P_1) to observe the mutation at a given frequency in the population (Thompson and Neel 1978) and the Luria-Delbrück theory (Luria and Delbrück 1943; Hästbacka *et al.* 1992; Labuda *et al.* 1996) that allows us to obtain the probability (thereafter denoted P_2) to observe the proportion of the various kinds of recombinants in the sample. From these, we obtain joint maximum-likelihood estimates of r and g. We also briefly present the coalescent-based methods that aim at estimating r and g that we have used here for comparison purposes.

Frequency of the disease allele: Assume a population of growth rate r, where the number of offspring of each individual is drawn in a geometric distribution. Assume also a mutant allele introduced g generations ago in that population. Thompson and Neel (1978) provide a formula that allows a computation of the probability (P_1) for that allele to reach an exact number of copies k in the present population,

$$P_1 = (1 - R)(1 - G)G^{k-1}, \quad (1)$$

where $k = N_{\mathrm{f}}$, P is the number of copies of the allele in the final population, $R = u(1 - v)/(u + v)$, and $G = 1 - (1 - R)/M$, with $M = r^g$, $u = M - 1$, and $v = -(1 - r)^2/r$.

Allelic association (standard Luria-Delbrück): Assume a mutant allele introduced as a single copy g generations ago in the population. Assume also that, within a sample of n chromosomes carrying this mutant allele, l chromosomes carry the major haplotype, which is presumed to be ancestral. The aim of this section is to compute the probability P_2 to observe l nonrecombinant haplotypes among n sampled individuals. For this we use the classical method (Hästbacka *et al.* 1992; Labuda *et al.* 1996), in which we remove an approximation that is not valid when the growth rate is too low.

The principle is as follows: if all lineages between the ancestral gene and the present copies sampled were independent

(complete star-like genealogy, see SLATKIN and HUDSON 1991), the proportion (p_{nr}) of nonrecombinants in the sample, for a haplotype of length θ around the disease gene, would be

$$p_{nr} = (1 - \theta)^g \approx e^{-\theta g}.$$

However, this assumption of independence of the lineages is untrue, especially during the first generations after the introduction of the gene. Thus, this equation has to be corrected as proposed by HÄSTBACKA *et al.* (1992) and LABUDA *et al.* (1996), following LURIA and DELBRÜCK's (1943) method. They showed that a number g_0 of generations have to be withdrawn from g. g_0 is the expected time to the first recombination event. Denoting M_g the number of meioses that occur in g generations, g_0 is the solution of the equation

$$M_{g_0} = 1/\theta. \tag{2}$$

For a growing population with growth rate r, this number is

$$M_g = \sum_{i=1}^{g} r^i = \frac{(r^g - 1)r}{r - 1}. \tag{3}$$

LABUDA *et al.* (1996) made the simplification $r^g - 1 \approx r^g$, which is accurate only for rapidly growing populations, like the one they studied. Since several populations, including some of the populations that we study here, do not fulfill this assumption, we did not make this simplification. Thus, combining (2) and (3) and solving for g_0 yields

$$g_0 = -1 + \frac{\log(r + (r - 1)/\theta)}{\log(r)}, \tag{4}$$

and the corrected probability for an individual to carry a nonrecombinant haplotype becomes $p^c_{nr} = \exp(-\theta(g - g_0))$. The probability P_2 then becomes

$$P_2 = B(n, p^c_{nr}; l), \tag{5}$$

where $B(n, p^c_{nr}; l)$ denotes the Binomial distribution of parameters n and p^c_{nr}, evaluated at l.

Allelic association (multipoint Luria-Delbrück estimation): We have designed a new method that allows the use of the whole-haplotype information (when available). This method was initially designed to give a more accurate estimation of the age of a haplotype (HUNTER *et al.* 2002). Assume now that the mutant allele is located at a locus D surrounded by a haplotype consisting of λ markers on the left side (M_{L1}, M_{L2}, . . . , $M_{L\lambda}$) and ρ markers on the right side (M_{R1}, M_{R2}, . . . , $M_{R\rho}$). Recombination rates between D and the markers are denoted, respectively, θ_{L0}, θ_{L1}, . . . $\theta_{L\lambda}$ and θ_{R0}, θ_{R1}, . . . $\theta_{R\rho}$, with the convention that $\theta_{L0} = \theta_{R0} = 0$. The probability for a haplotype carrying D and separated by g generations from the ancestral haplotype to be of a given size θ_{Li} on the left side of the mutation (*i.e.*, to be nonrecombinant for M_{L1}, . . . , M_{Li}, but recombinant for M_{Li+1}) after g generations is given by

$$p_{Li} = \exp(-\theta_{Li}(g - g^i_0)) - \exp(-\theta_{L\,i+1}(g - g^{i+1}_0)), \tag{6}$$

where g^i_0 is the Luria-Delbrück correction, obtained from (4), replacing θ by θ_{Li}. The same calculation is applied to the right side of the mutation, yielding similar probabilities p_{Rj}, $j = 0$ . . . ρ. Then, the probability for a haplotype to be of length θ_{Li} on the left side and θ_{Rj} on the right side is $p_{i,j} = p_{Li} \times p_{Rj}$. Denote $n_{i,j}$ the numbers of carriers of each haplotype; the probability P_2 to observe these $n_{i,j}$'s in the sample of size n will be

$$P_2 = M(n, (p_{i,j}); (n_{i,j})), \tag{7}$$

where $M(n, (p_{i,j}), (n_{i,j}))$ is the multinomial distribution with parameters n and $(p_{i,j})$, taken at $(n_{i,j})$.

Joint estimation: The likelihood $L(g, r)$ of a parameter set (g, r) is the probability, for that set of parameters, to observe both the number of copies (k) in the population and the observed haplotypic variability in the sample of disease chromosomes. Thus, $L(g, r)$ is the product of the two probabilities P_1 and P_2, given by (1) and (5) or (7), respectively. $L(g, r)$ is minimized numerically using Mathematica (the notebook is available from F. Austerlitz). This method yields the maximum-likelihood estimates $\hat{g}$ and $\hat{r}$, along with their 95% confidence intervals using the standard Max − 2 rule (see, *e.g.*, KAPLAN and WEIR 1995). The parameters that are needed for the method are the final size of the population (N_f), carrier frequency of the disorders (p), frequency of the different haplotypes in the sample, and recombination rate between the different markers. These estimates assume neutrality.

If the mutant allele was generated by mutation in the population under study, $\hat{g}$ will simply be an estimate of the time of appearance of that mutation. Conversely, if the mutant allele was introduced by migration in the population as a single copy, $\hat{g}$ estimates the age of this introduction by migration in the population. However, if several migrants brought the gene into the population, $\hat{g}$ will also integrate the history of the allele in the ancestral population from which these migrants came. If the growth rate varies over time, our estimate $\hat{r}$ should be an estimate of the average growth rate over time, but the impact on g is more difficult to assess.

Coalescent-based methods: To our knowledge, as yet no coalescent-based methods allow the joint estimation of the growth rate of the population and the age of the mutation. Therefore we used two different methods. First, we used the method proposed by SLATKIN and BERTORELLE (2001) to infer the growth rate from the same kind of molecular data that we use in our method: the frequency of carriers of the disease allele in the population, the frequency of nonrecombinant haplotypes in the sample, and the size of the haplotype. This estimation was performed using the C program provided by M. Slatkin. Then the estimated growth rate was used as an input, along with the molecular data, to estimate the age of the mutation using the DMLE+ software (REEVE and RANNALA 2002). For comparison purposes, since SLATKIN and BERTORELLE's (2001) method estimates an exponential growth rate ($\hat{d}$) assuming a continuous-time model, we translated it into a discrete time growth rate ($\hat{r}$), comparable with ours, using the formula $\hat{r} = e^{\hat{d}}$.

Data used: Published data on haplotypes and carrier frequencies of different disorders in several populations were used to compare the growth rate and mutation age estimates for various diseases in the same population and check whether the method provides consistent results. For the populations for which demographic data are available, we compared the growth rate estimated from these data with our inferred growth rate. We chose four populations for which several disorders have been studied. Two of these populations are small in size (∼300,000 inhabitants) and recently founded. One is the SLSJ population, for which extensive genetic and demographic data are available. The other is the Vlax Gypsies in Bulgaria, for whom demographic data are uncertain. The other two populations are older and of larger size: the Finnish population, which numbers ∼5,000,000 inhabitants, and the Ashkenazi Jews, who are now ∼10,000,000 worldwide. Finally, we apply the method to one gene in the whole European population, to see whether the method is extendable to a larger scale.

RESULTS

Analysis of several examples: Table 1 gives the population growth rates and age of the mutations estimated

TABLE 1

Estimated growth rate and age of mutations (with their 95% confidence interval obtained with the standard Max − 2 rule) obtained with our method and the coalescent-based methods for various populations and disease genes

						Joint estimation (our method)		Coalescent-based methods	
Population	Disorder	p	l/n	θ	N_0	$\hat{r}$	$\hat{g}$	$\hat{r}$	$\hat{g}$
Vlax	GD[b]	0.05	4/15	0.11	300,000	1.93 (1.70, 2.46)	13.7 (11.4, 17.6)	1.84 (1.42, 9.97)	16.3 (14.5, 21.0)
	CCFDN[c]	0.036	10/48	0.10		1.57 (1.43, 1.88)	18.7 (15.7, 23.0)	1.66 (1.37, 3.43)	12.3 (10.4, 17.4)
	HMSNL[d]	0.04	83/119	0.03		1.66 (1.5, 2.01)	17.0 (14.5, 20.4)	1.8 (1.45, 5.34)	9.0 (8.1, 10.6)
SLSJ	PDDR[e]	0.0385	24/37	0.07	300,000	3.09 (2.51, 4.59)	8.02 (6.74, 10.1)	4.85 (2.15, 41.9)	4.87 (4.56, 5.50)
	ARSACS[f]	0.0455	41/53	0.051		3.88 (3.06, 6.11)	6.86 (5.79, 8.61)	5.74 (2.37, 187)	4.35 (4.13, 4.69)
	ARSACS[g]		14/53	0.11		1.84 (1.63, 2.31)	14.5 (12.3, 17.7)	2.07 (1.56, 5.11)	9.37 (7.9, 12.1)
	ACCPN[h]	0.0435	38/48	0.05		4.28 (3.32, 6.96)	6.38 (5.38, 8.07)	6.5 (2.45, 369.15)	DNC[t]
Ashkenazi[a]	GD[i]	0.05	31/66	0.025	10,000,000	1.38 (1.32, 1.50)	37.2 (32.8, 44.1)	1.46 (1.26, 2.23)	25.4 (22.6, 32.6)
	BS[j]	0.0091	26/33	0.012		1.5 (1.40, 1.70)	25.9 (22.1, 32.6)	1.65 (1.18, 12.3)	16.8 (14.8,21.7)
	FXIDII[k]	0.0434	16/99	0.013		1.06 (1.05, 1.09)	165 (145, 194)	1.07 (1.00, 1.19)	112 (91.4, 167)
	FXIDIII[l]	0.04	46/73	0.013		1.28 (1.24, 1.38)	45.9 (40.2, 54.9)	1.08 (1.03, 1.23)	90 (70.1, 128)
	ITD[m]	0.002	20/36	0.023		1.29 (1.23, 1.43)	33.4 (27.8, 42.2)	1.38 (1.16, 2.58)	21.6 (17.6, 30.1)
	FD[n]	0.0225	186/435	0.036		1.47 (1.39, 1.63)	29.5 (27.0, 32.4)	1.55 (1.35, 2.31)	16.3 (14.1, 21.1)
Finland	PSOSL[o]	0.002	18/24	0.00217	5,000,000	1.03 (1.02,1.05)	199 (152, 275)	1.04 (1.01, 1.23)	104 (43.7, 179)
	PME[p]	0.015	65/87	0.00895		1.24 (1.19, 1.33)	45.9 (39.3, 55.9)	1.29 (1.16, 1.96)	25.7 (22.6, 34.8)
	CCD[q]	0.020	8/47	0.12		1.9 (1.72, 2.30)	16.9 (14.7, 20.4)	2.09 (1.59, 4.47)	12.0 (10.9, 14.7)
	DTD[r]	0.016	139/146	0.0007		1.07 (1.06, 1.1)	125 (104, 160)	1.08 (1.03, 1.47)	63.3 (49.5, 90.8)
Europe	CCR5[s]	0.18	39/46	0.006	10,000,000	1.47 (1.40, 1.62)	34.6 (30.7, 42.0)	1.57 (1.22, 4.1)	24.2 (22.6, 28.4)
					100,000,000	1.61 (1.53, 1.77)	33.2 (30.0, 39.2)	1.68 (1.28, 4.91)	21.4 (19.8, 26.1)
					500,000,000	1.72 (1.63, 1.89)	32.4 (29.6, 37.8)	1.86 (1.29, 5.72)	18.2 (17.0, 21.2)

In each case the frequency of carriers (p), the proportion of carriers of the major haplotype (l/n), the recombination fraction (θ), and the present population size (N_0) are indicated.

[a] For the Ashkenazi Jews population, we used for the frequency of the most frequent haplotype the p_{excess} value given in Colombo (2000).

[b] Galactokinase deficiency (Kalaydjieva *et al.* 1999).

[c] Congenital cataracts facial dysmorphism neuropathy (Angelicheva *et al.* 1999).

[d] Hereditary motor and sensory neuropathy-Lom (Kalaydjieva *et al.* 1996, 2000, 2001).

[e] Pseudo-vitamin D deficiency rickets (Labuda *et al.* 1996).

[f] Autosomal recessive spastic ataxia of Charlevoix-Saguenay (core haplotype; Richter *et al.* 1999; Engert *et al.* 2000).

[g] Autosomal recessive spastic ataxia of Charlevoix-Saguenay (complete haplotype; Richter *et al.* 1999; Engert *et al.* 2000).

[h] Peripheral neuropathy with or without agenesis of the corpus callosum (Casaubon *et al.* 1996).

[i] Gaucher disease (Diaz *et al.* 1999; Colombo 2000).

[j] Bloom syndrome (Ellis *et al.* 1994).

[k] Factor XI deficiency type II (Goldstein *et al.* 1999).

[l] Factor XI deficiency type III (Goldstein *et al.* 1999).

[m] Idiopathic torsion dystonia (Risch *et al.* 1995b).

[n] Familial disautomia (Blumenfeld *et al.* 1999).

[o] Polycystic lipomembranous osteodysplasia with sclerosing leukoencephalopathy (Pekkarinen *et al.* 1998).

[p] Progressive myoclonus epilepsy (Virtaneva *et al.* 1996).

[q] Congenital chloride diarrhea (Höglund *et al.* 1995).

[r] Diastrophic dysplasia (Hästbacka *et al.* 1992, 1994).

[s] CCR5-Δ32 AIDS resistance allele (Stephens *et al.* 1998).

[t] DMLE+ program did not converge probably due to the too high estimated growth rate ($\hat{r} = 6.5$).

with our method and with the coalescent-based methods. A consistent pattern for the different genes was observed in the two recently founded populations (Vlax and SLSJ). Leaving apart the case of autosomal recessive spastic ataxia of Charlevoix-Saguenay (ARSACS) in SLSJ when we considered the large 11-cM haplotype rather than the 5.1-cM core haplotype (Richter *et al.* 1999; Engert *et al.* 2000), the estimated growth rates ranged from 1.57 to 1.93 in all cases for the Vlax population and from 3.09 to 4.28 for SLSJ. The corresponding estimated ages of the mutations ranged from 13.7 to 18.7 for Vlax and from 6.38 to 8.02 for SLSJ. For ARSACS, using the core haplotype yielded results similar to those for the other genes, whereas the estimate based on the large haplotypes yielded a lower growth rate and higher age of the mutation. This discrepancy may be the consequence of other phenomena that may occur on this large haplotype, like double-recombination events.

As for the older populations, the Ashkenazi Jews showed much older mutations ($\hat{g}$ ranged from 25.9 to 45.9) and smaller population growth ($\hat{r}$ ranged from 1.28 to 1.5) except for factor XI deficiency of type II, where $\hat{r} = 1.06$ and $\hat{g} = 165$. The Finnish population showed contrasting patterns depending on the disease, with a high growth rate ($\hat{g} = 16.9$, $\hat{r} = 1.9$) estimated with recent mutations and a low growth rate ($\hat{g} = 199$, $\hat{r} = 1.03$) with old ones.

Finally, we treated the case of the CCR5-Δ32 AIDS resistance gene in Europe. Because Europe cannot be considered as a single, homogenous population, we tried different values for its assumed final size, ranging from 10,000,000 to 500,000,000, this latter value being approximately the present census size of Europe. The inferred growth rate ranged from 1.47 to 1.72 with an age of the mutation from 32.4 to 34.6.

Comparison with coalescent-based methods: Both methods yielded similar results in terms of the estimated growth rates. The estimates obtained using the Slatkin and Bertorelle (2001) method were almost always slightly higher than ours and the upper range of their confidence interval was always much higher than ours and clearly unrealistic in some cases. The ages estimated with DMLE+ were in almost all cases lower than the ages estimated with our method but they were in the same order of magnitude. To test whether this discrepancy in the estimate of the age came from the difference in the estimate of population growth rate, we performed the DMLE+ analysis using the estimate of r obtained with our method. This yielded higher estimates of the age of the mutation, but still lower than our estimate (result not shown). For instance, in the case of hereditary motor and sensory neuropathy-Lom (HMSNL) in the Vlax population, $\hat{g}$ increased from 9.0 to 9.9, still lower than the 17.0 obtained with our method.

Multipoint estimates: We performed this procedure for three cases (see Table 2), for which we had the necessary data (position of all markers and frequency of carrier of each haplotype). In two cases out of three [ARSACS in SLSJ and polycystic lipomembranous osteodysplasia with sclerosing leukoencephalopathy (PLOSL) in Finland], we found similar estimates for minimum growth rate and age of the mutation, compared with the case when we counted only recombinant and nonrecombinant haplotypes (compare with Table 1). The confidence interval was similar for growth rate but reduced for the age of the mutation: the difference between the upper and lower limits of the confidence interval decreased from 123 to 113 generations for PLOSL and from 5.4 to 4.4 for ARSACS. In the last case (galactokinase deficiency in the Vlax population), the estimate of growth rate was lower (1.61 *vs.* 1.91) and conversely the age of the mutation was higher (123 *vs.* 113).

DISCUSSION

An important result is that our estimates, which are based solely on genetic data, are consistent with the general history of the populations, as described in the literature. The recently founded populations (Vlax and SLSJ) presented a constant pattern of "young" disorders associated with a high growth rate, whereas the populations established for a longer time (Ashkenazi and Finnish) showed a general trend of older diseases associated with a lower estimated growth rate.

In addition to this global consistency between our estimates and the demographic data, we were able to detect some specific phenomena. For the SLSJ data, the $\hat{r}$ values are much higher than the known growth rate of the population (1.4; Austerlitz and Heyer 1998) for all loci in the present study. It is quite unlikely that heterozygous advantage could be the explanation for these high estimated growth rates. Indeed, if these genes showed heterozygote advantage, they would be found in nonnegligible frequencies in other populations, like the French population from which the founders of SLSJ came. Moreover, if heterozygote advantage was general for disease genes, we would estimate an excessive growth rate in all populations.

As we indicated above, we have demonstrated in a previous study that the high carrier frequencies of these disorders are explained mainly by fertility inheritance: a correlation in effective reproduction from one generation to the next (Austerlitz and Heyer 1998). In other words, the individuals that come from large sibships that mostly remained in the community tend to have also a lot of children that settle themselves in the community; therefore, this fertility inheritance is mainly cultural. A disease gene carried by such individuals will have an intrinsic growth rate that is much higher than the population growth rate, hence the very high values estimated here.

SLSJ is a case study to check whether fertility inheritance can be detected from molecular data. Indeed, our

TABLE 2

Joint estimate of growth ($\hat{r}$) and age of the mutation ($\hat{g}$) with their 95% confidence interval obtained with the standard Max − 2 rule using multipoint analysis for three cases where it was possible

Population	Disorder	$\hat{r}$	$\hat{g}$	Reference
Vlax	Galactokinase deficiency	1.61 (1.47, 1.93)	18.2 (15.3, 22.8)	Kalaydjieva *et al.* (1999)
SLSJ	ARSACS	1.91 (1.68, 2.40)	13.9 (12.0, 16.4)	Richter *et al.* (1999)
Finland	PSOSL	1.03 (1.02, 1.05)	186 (143, 256)	Pekkarinen *et al.* (1998)

estimates of growth rate are similar for all loci and much higher than the known population growth rate. As a side effect, it yields a slight underestimate of the age of the mutation: for all disorders in SLSJ, we estimated an age between 6 and 8 generations. However, we know from demographic data that the mutations were present in the population when it was founded 12 generations ago (Bouchard and De Braekeleer 1991; Labuda *et al.* 1996), a value that is above the upper limit of the confidence interval in all cases. Since we have similar results for several genes, this discrepancy is indeed an indication of a real bias. The known increase in allelic association caused by fertility inheritance (Austerlitz and Heyer 2000) may explain in part this slight downward bias. The demographic estimate of 12 generations might also be an overestimate, since it is based on an assumption of a generation length of 25 years, whereas the true value might be closer to 30 years according to a study based precisely on the French Canadian population (Tremblay and Vezina 2000).

Can we detect fertility inheritance in other populations? In the case of the Vlax community in Bulgaria, the estimated $\hat{r}$ are rather high (from 1.57 to 1.93) for the three disorders under study. If we consider the population size of 17,000 Roma in the 14th century [a reasonable approximation given the available information on the historical demography of the Roma (Marushiakova and Popov 1997)] and the current population of 8 million Romani in Europe (Liegeois 1994), the overall growth rate, for a generation time of 25 years, is 1.32. As for SLSJ, this discrepancy could be explained by fertility inheritance. This type of correlation in effective reproduction could be the consequence of the social and cultural subdivision of this community. The studies of disease genes in the Vlax Gypsies have involved three groups: Rudari, Lom, and Kalderas. If for any reason these three groups had a different mean effective number of children, this could lead to an overall correlation in effective children, the children of the people from the group with the highest number of effective children being likely to remain also in this group. This hypothesis remains to be tested. Nevertheless, since we did not detect as much fertility inheritance as in the SLSJ, we expect a smaller bias on the age of the mutations. Indeed, the estimated age of the various disorders coincides well with the founding of the population in the 14th century, when Vlax groups arrived in Romania and were confined there until the 19th century (Fraser 1992; Liegeois 1994).

In the case of the Ashkenazi Jews, growth rates are estimated at ~1.4 (except for factor XI deficiency type II), compatible with the value of 1.5 [exp(0.4)] that has been estimated from demographical data (Risch *et al.* 1995b; Labuda *et al.* 1997). Therefore our results are in agreement with previous conclusions (Labuda *et al.* 1997; Colombo 2000) that the frequencies of inherited disorders in the Ashkenazi Jews could be explained simply by demographic growth, without the need of invoking heterozygous advantage or specific demographic behavior, like the social selection process proposed by several authors (Motulsky 1995; Risch *et al.* 1995a,b). So even if this population was subdivided into small communities until the early 19th century, differential growth rates among communities would not be sufficient to create a fertility inheritance effect.

Regarding the age of the mutation, our estimate of the age of the idiopathic torsion dystonia mutation, namely 33.4 generations, is consistent with the 32 generations estimated previously by Labuda *et al.* (1997). Considering factor XI deficiency, we come to the same conclusion as Goldstein *et al.* (1999) that type II is much older than type III: 165 *vs.* 46 generations. Our estimates are higher than their estimates (120 for type II and 31 for type III) but, as they point out, their estimates are based on the coalescent time of the sample and are thus an underestimate of the age of the mutation, which predates the coalescent time of all carriers of the disease gene.

Whereas estimates obtained for disorders in the recently founded populations appear consistent, a more variable pattern is observed in the case of an older population like Finland, where situations range from recent disorders associated with a rapid growth rate to old disorders with a much lower growth rate. This result is rather logical since, in a recent population, it is likely that the disorders observed at present were introduced simultaneously by the migrants that founded the population. In older populations, however, disease mutations could have been introduced, by mutation or by migration, at various points in time.

Geographical structure, if any, is also more likely to have an impact on these older populations. Thus a vari-

ant can arise in a given subpopulation and increase rapidly in frequency. This is consistent with the patterns observed in Finland, where some disorders are older and have a wide geographical distribution, whereas others are younger with a more localized distribution (de la Chapelle and Wright 1998; Peltonen *et al.* 1999). When disorders have a different age, it is difficult to compare the growth rate estimate since populations do not have a steady growth. We have examined only one recent gene (CDD), which has a local distribution and a very high growth rate (~1.9). The estimated rapid growth could be due to a high local growth rate or a fertility correlation in the subpopulation where this gene is found or to a selective effect. We would need data on other similar genes to distinguish between these different explanations.

Finally, for the CCR5-Δ32 AIDS-resistance allele in Europe, we estimated a growth rate between 1.47 and 1.72, clearly higher than what we know from past European demography: the European population (excluding the countries of the former USSR) increased from ~32 million inhabitants in 1500 to ~492 million at present (Biraben 1979), *i.e.*, a growth rate of ~1.1. The difference between the two estimates is consistent with the hypothesis that selective advantage of heterozygotes is responsible for the high frequency of CCR5-Δ32 in Europe (Stephens *et al.* 1998). Alternatively this difference could be a consequence of geographic structure that would have allowed the gene to increase more rapidly in some areas, as was shown in Finland. More data on the same geographical and historical scale are needed to evaluate the relative impact of demographic and selective factors.

Comparing our method with those based on coalescent simulations suggests that, while the estimates are generally in agreement, our values are usually slightly smaller for $\hat{r}$ and higher for $\hat{g}$. Our confidence intervals are smaller for $\hat{r}$ but larger for $\hat{g}$. Moreover, the upper value for the confidence interval of the growth rate is much smaller in our cases, coalescent methods yielding an exaggerated value in several cases. More theoretical work is needed to understand these discrepancies.

Similarly we have an indication that the multipoint method that takes into account the whole distribution of recombinants and the distance at which the recombination occurred in each case yields more accurate results, at least in terms of the width of the confidence interval. This aspect is in need of confirmation with data on other diseases and by theoretical work (simulations).

Our method like the coalescent-based methods assumes that the frequency of these genes changes as if they were neutral. This assumption might appear contradictory with the fact that most of the genes studied are recessive lethal disorders. However, since these genes are in low frequency, the occurrence of homozygotes is very rare and thus negative selection acts only very moderately. Thus, this assumption of neutrality, which is made in several methods that use allelic association (Kaplan *et al.* 1995; Collins and Morton 1998), is unlikely to yield a bias in our estimates.

In conclusion, our method provides an efficient way for tracing back the recent history of populations or of disorders in these populations. Thus, it will be especially helpful for populations for which no demographic data are available. It is consistent across disorders in several populations and enables us to detect factors like selection or cultural events that allow a gene to reach a high frequency within a few generations. Distinguishing the effects of these factors needs the study of several loci within the same population. It would be inappropriate to reject neutrality at a locus if studied alone and not in contrast with other loci, because it would be impossible to determine if the high intrinsic growth rate of an allele is really the result of selection specifically at this locus or of a demographic process that affects all loci. This need of contrasting several loci for testing neutrality is also pointed out by Nielsen (2001). Finally, even if the present design was applied here only on disease genes (and one AIDS resistance gene) in human populations, it could be extended to any haplotypic data when such data become available.

The availability of demographic data in some cases has allowed us to detect culturally inherited fertility, as in the documented case of the SLSJ. We have an indication that such a phenomenon could exist in the Vlax population. Further theoretical work on this subject is needed to develop more accurate methods to detect and gauge fertility correlation. The fine study of coalescent trees is a promising avenue since fertility correlation changes not only the scale of the tree but also its symmetry (Sibert *et al.* 2002). This issue is important since fertility inheritance can bias estimated population growth, age of mutation, and also recombination rate (Austerlitz and Heyer 2000). Furthermore, it has a tremendous impact on effective population size, reducing it by a factor of >10 in the case of SLSJ (Austerlitz and Heyer 1998), and could lead to an erroneous detection of population growth in stationary populations (Sibert *et al.* 2002).

We thank Montgomery Slatkin for sending us his program for estimating growth rate, Jeff Reeve for a corrected version of the DMLE+ software and his help on its use, and two anonymous reviewers for helpful comments and suggestions. L.K. acknowledges support from the Australian Research Council and the Wellcome Trust.

LITERATURE CITED

Angelicheva, D., I. Turnev, D. Dye, D. Chandler, P. K. Thomas *et al.*, 1999 Congenital cataracts facial dysmorphism neuropathy (CCFDN) syndrome: a novel developmental disorder in Gypsies maps to 18qter. Eur. J. Hum. Genet. **7:** 560–566.

Austerlitz, F., and E. Heyer, 1998 Social transmission of reproductive behavior increases frequency of inherited disorders in a young-expanding population. Proc. Natl. Acad. Sci. USA **95:** 15140–15144.

Austerlitz, F., and E. Heyer, 2000 Allelic association is increased

by correlation of effective family size. Eur. J. Hum. Genet. **8:** 980–985.

Austerlitz, F., B. Jung-Muller, B. Godelle and P.-H. Gouyon, 1997 Evolution of coalescence times, genetic diversity and structure during colonization. Theor. Popul. Biol. **51:** 148–164.

Biraben, J.-N., 1979 Essai sur l'évolution du nombre des hommes. Population **1:** 13–25.

Blumenfeld, A., S. A. Slaugenhaupt, C. B. Liebert, V. Temper, C. Maayan *et al.*, 1999 Precise genetic mapping and haplotype analysis of the familial dysautonomia gene on human chromosome 9q31. Am. J. Hum. Genet. **64:** 1110–1118.

Bouchard, G., and M. De Braekeleer, (Editors), 1991 *Histoire d'un Génome. Population et Génétique dans l'est du Québec*. Presses de l'Université du Québec, Sillery, Quebec, Canada.

Casaubon, L. K., M. Melanson, I. Lopes-Cendes, C. Marineau, E. Andermann *et al.*, 1996 The gene responsible for a severe form of peripheral neuropathy and agenesis of the corpus callosum maps to chromosome 15q. Am. J. Hum. Genet. **58:** 28–34.

Collins, A., and N. E. Morton, 1998 Mapping a disease locus by allelic association. Proc. Natl. Acad. Sci. USA **95:** 1741–1745.

Colombo, R., 2000 Age estimate of the N370S mutation causing Gaucher disease in Ashkenazi Jews and European populations: a reappraisal of haplotype data. Am. J. Hum. Genet. **66:** 692–697.

de la Chapelle, A., and F. A. Wright, 1998 Linkage disequilibrium mapping in isolated populations: the example of Finland. Proc. Natl. Acad. Sci. USA **95:** 12416–12423.

Diaz, A., M. Montfort, B. Cormand, B. Zeng, G. M. Pastores *et al.*, 1999 Gaucher disease: the N370S mutation in Ashkenazi Jewish and Spanish patients has a common origin and arose several thousand years ago. Am. J. Hum. Genet. **64:** 1233–1238.

Ellis, N. A., A. M. Roe, J. Kozloski, M. Proytcheva, C. Falk *et al.*, 1994 Linkage disequilibrium between the FES, D15S127, and BLM loci in Ashkenazi Jews with Bloom syndrome. Am. J. Hum. Genet. **55:** 453–460.

Engert, J. C., P. Berube, J. Mercier, C. Dore, P. Lepage *et al.*, 2000 ARSACS, a spastic ataxia common in northeastern Quebec, is caused by mutations in a new gene encoding an 11.5-kb ORF. Nat. Genet. **24:** 120–125.

Fraser, A. M., 1992 *The Gypsies*. Blackwell, Oxford.

Goldstein, D. B., D. E. Reich, N. Bradman, S. Usher, U. Seligsohn *et al.*, 1999 Age estimates of two common mutations causing factor XI deficiency: recent genetic drift is not necessary for elevated disease incidence among Ashkenazi Jews. Am. J. Hum. Genet. **64:** 1071–1075.

Hästbacka, J., A. de la Chapelle, I. Kaitila, P. Sistonen, A. Weaver *et al.*, 1992 Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat. Genet. **2:** 204–211.

Hästbacka, J., A. de la Chapelle, M. M. Mahtani, G. Clines, M. P. Reeve-Daly *et al.*, 1994 The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. Cell **78:** 1073–1087.

Höglund, P., P. Sistonen, R. Norio, C. Holmberg, A. Dimberg *et al.*, 1995 Fine mapping of the congenital chloride diarrhea gene by linkage disequilibrium. Am. J. Hum. Genet. **57:** 95–102.

Hunter, M., E. Heyer, F. Austerlitz, D. Angelicheva, V. Nedkova *et al.*, 2002 The P28T mutation in the GALK1 gene accounts for galactokinase deficiency in Roma (Gypsy) patients across Europe. Pediatr. Res. **51:** 602–606.

Kalaydjieva, L., J. Hallmayer, D. Chandler, A. Savov, A. Nikolova *et al.*, 1996 Gene mapping in Gypsies identifies a novel demyelinating neuropathy on chromosome 8q24. Nat. Genet. **14:** 214–217.

Kalaydjieva, L., A. Perez-Lezaun, D. Angelicheva, S. Onengut, D. Dye *et al.*, 1999 A founder mutation in the GK1 gene is responsible for galactokinase deficiency in Roma (Gypsies). Am. J. Hum. Genet. **65:** 1299–1307.

Kalaydjieva, L., D. Gresham, R. Gooding, L. Heather, F. Baas *et al.*, 2000 N-myc downstream-regulated gene 1 is mutated in hereditary motor and sensory neuropathy-Lom. Am. J. Hum. Genet. **67:** 47–58.

Kalaydjieva, L., D. Gresham and F. Calafell, 2001 Genetic studies of the Roma (Gypsies): a review. BMC Med. Genet. **2:** 5.

Kaplan, N. L., and B. S. Weir, 1995 Are moment bounds on the recombination fraction between a marker and a disease locus too good to be true? Allelic association mapping revisited for simple genetic diseases in the Finnish population. Am. J. Hum. Genet. **57:** 1486–1498.

Kaplan, N. L., W. G. Hill and B. S. Weir, 1995 Likelihood methods for locating disease genes in nonequilibrium populations. Am. J. Hum. Genet. **56:** 18–32.

Labuda, M., D. Labuda, M. Korab-Laskowska, D. E. Cole, E. Zietkiewicz *et al.*, 1996 Linkage disequilibrium analysis in young populations: pseudo-vitamin D-deficiency rickets and the founder effect in French Canadians. Am. J. Hum. Genet. **59:** 633–643.

Labuda, D., E. Zietkiewicz and M. Labuda, 1997 The genetic clock and the age of the founder effect in growing populations: a lesson from French Canadians and Ashkenazim. Am. J. Hum. Genet. **61:** 768–771.

Lavery, S., C. Moritz and D. R. Fielder, 1996 Genetic patterns suggest exponential growth in a declining species. Mol. Biol. Evol. **13:** 1106–1113.

Liegeois, J. P., 1994 *Roma, Gypsies, Travellers*. Council of Europe Press, Strasbourg, France.

Luria, S. E., and M. Delbrück, 1943 Mutations of bacteria from virus sensitivity to virus resistance. Genetics **28:** 491–511.

Marushiakova, E., and V. Popov, 1997 *Gypsies (Roma) in Bulgaria*. Peter Lang, Frankfurt am Main, Germany.

Motulsky, A. G., 1995 Jewish diseases and origins. Nat. Genet. **9:** 99–101.

Nielsen, R., 2001 Statistical tests of selective neutrality in the age of genomics. Heredity **86:** 641–647.

Pekkarinen, P., M. Kestila, J. Paloneva, J. Terwilliger, T. Varilo *et al.*, 1998 Fine-scale mapping of a novel dementia gene, PLOSL, by linkage disequilibrium. Genomics **54:** 307–315.

Peltonen, L., A. Jalanko and T. Varilo, 1999 Molecular genetics of the Finnish disease heritage. Hum. Mol. Genet. **8:** 1913–1923.

Reeve, J. P., and B. Rannala, 2002 DMLE+: Bayesian linkage disequilibrium gene mapping. Bioinformatics **18:** 894–895.

Richter, A., J. D. Rioux, J. P. Bouchard, J. Mercier, J. Mathieu *et al.*, 1999 Location score and haplotype analyses of the locus for autosomal recessive spastic ataxia of Charlevoix-Saguenay, in chromosome region 13q11. Am. J. Hum. Genet. **64:** 768–775.

Risch, N., D. de Leon, S. Fahn, S. Bressman, L. Ozelius *et al.*, 1995a ITD in Ashkenazi Jews—genetic drift or selection? Nat. Genet. **11:** 14–15.

Risch, N., D. de Leon, L. Ozelius, P. Kramer, L. Almasy *et al.*, 1995b Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. Nat. Genet. **9:** 152–159.

Rogers, A. R., and H. Harpending, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9:** 552–569.

Sibert, A., F. Austerlitz and E. Heyer, 2002 Wright-fisher revisited: the case of fertility correlation. Theor. Popul. Biol. **62:** 181–197.

Slatkin, M., and G. Bertorelle, 2001 The use of intraallelic variability for testing neutrality and estimating population growth rate. Genetics **158:** 865–874.

Slatkin, M., and R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.

Stephens, J. C., D. E. Reich, D. B. Goldstein, H. D. Shin, M. W. Smith *et al.*, 1998 Dating the origin of the *CCR5*-Δ32 AIDS-resistance allele by the coalescence of haplotypes. Am. J. Hum. Genet. **62:** 1507–1515.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–596.

Thompson, E. A., and J. V. Neel, 1978 Probability of founder effect in a tribal population. Proc. Natl. Acad. Sci. USA **75:** 1442–1445.

Tremblay, M., and H. Vezina, 2000 New estimates of intergenerational time intervals for the calculation of age and origins of mutations. Am. J. Hum. Genet. **66:** 651–658.

Virtaneva, K., J. Miao, A. L. Traskelin, N. Stone, J. A. Warrington *et al.*, 1996 Progressive myoclonus epilepsy EPM1 locus maps to a 175-kb interval in distal 21q. Am. J. Hum. Genet. **58:** 1247–1253.

Communicating editor: M. A. Asmussen