Available online at www.sciencedirect.com
SCIENCE @ DIRECT®

Biomedicine & Pharmacotherapy 57 (2003) 25–33

BIOMEDICINE & PHARMACOTHERAPY

www.elsevier.com/locate/biopha

Dossier: AIDS

Genomic studies in AIDS: problems and answers. Development of a statistical model integrating both longitudinal cohort studies and transversal observations of extreme cases

C. Huber [a], O. Pons [b], H. Hendel [c], P. Haumont [a], L. Jacquemin [a], S. Tamim [c], J.F. Zagury [c,*]

[a] *UFR Biomédicale, Université Paris V, INSERM U472, IFR69, 45, rue des Saints-Pères, 75006 Paris, France*
[b] *Unité de biométrie, INRA, 78352 Jouy-en-Josas, France*
[c] *Centre de Recherche des Cordeliers, Université Paris VI, 15, rue de l'École de Médecine, 75006 Paris, France*

Received 5 September 2002; accepted 25 September 2002

Abstract

Genomic studies developed to understand HIV-1 infection and pathogenesis have often lead to conflicting results. This is linked to various factors, including differences in cohort design and selection, the numbers of patients involved, the influence of population substructure, the ethnic origins of the participants, and phenotypic definition. These difficulties in the interpretation of results are examined through published studies on the role of polymorphisms in HLA and the chemokine receptors genes in AIDS. Our analysis suggests that the use of haplotypes will strengthen the results obtained in a given cohort, and meta-analysis including multiple cohorts to gather large-enough numbers of patients should also allow clarification of the genetic associations observed. A *P*-value of 0.001 appears to be a good compromise for significance on candidate genes in a genetic study. Due to the generally limited size of available cohorts, results will have to be validated in other cohorts.

We developed a model to fit transversal case studies (extreme case-control studies) with longitudinal cohorts (all-stages patients) for observations on two gene polymorphisms of CCR5 and NQO1. Interestingly, we observe a protective effect for the CCR5-Δ32 mutant allele in 95% of the simulations based on that model when using a population of 600 subjects; however, when using populations of 250 subjects we find a significant protection in only 59% of the simulations. Our model gives thus an explanation for the discrepancies observed in the various genomic studies published in AIDS on CCR5-Δ32 and other gene polymorphisms: they result from statistical fluctuations due to a lack of power. The sizes of most seroconverter cohorts presently available seem thus insufficient since they include less than a few hundred subjects. This result underlines the power and usefulness of the transversal studies involving extreme patients and their complementarity to longitudinal studies involving seroconverter cohorts. The transposition approach of extreme case-control data into longitudinal analysis should prove useful not only in AIDS but also in other diseases induced by chronic exposure to a foreign agent or with chronic clinical manifestations.

Keywords: Association; AIDS; Cohort; Gene; Genomic; Polymorphism; Progression

1. Introduction

Numerous studies have been performed to explain the role of the host genetic background in determining the rate of AIDS disease progression [1]. Such studies are of importance since they can lead to a better understanding of virus/host interactions and the corollary development of new tools to better fight the virus. There are two kinds of genetic studies commonly used for HIV-1 and AIDS: longitudinal cohort studies involving all-stage patients followed since their seroconversion after study enrollment and transversal studies comparing the extremes of disease progression (slow progressors (SP), fast progressors (FP)) with controls. The statistical associations found on longitudinal observations are based on Kaplan–Meier "survival" curves while the statistical associations found in transversal studies are based on Fisher's exact tests or χ^2-like tests.

Initial studies have focused on the influence of human leucocyte antigens (HLA) genes on AIDS disease progression [2], which were often inconsistent due to the large number of HLA alleles, low allele frequencies, and the limited size of the cohorts analyzed [3]. The most important studies involving HLA are recent and involved seroconverter

* Corresponding author.
E-mail address: jfz@ccr.jussieu.fr (J.F. Zagury).

patient cohorts of size 240 [4] and 470 [5]. Other genetic polymorphisms involving the cytokines TNFα [6], IFNγ [7], IL4 [8], and IL10 [9] or the myelin basic protein (MBP) [10] have shown interesting results, but the most successful genetic studies to date have dealt with the chemokines receptors [11]. In effect, while chemokines receptors were identified as the coreceptors of HIV-1 in 1996 [12], a variant of CCR5, CCR5-Δ32, was shown to have a dramatic protective effect against disease progression [13–15]. Since then, numerous studies performed on the polymorphism of chemokines and their receptors have found specific variants affecting significantly disease outcome: a polymorphism in the coreceptor CCR2 [16], a polymorphism in the promoter of the coreceptor CCR5 [17,18], a polymorphism in the promoter of the ligand of CXCR4, SDF1 [19] or in the regulatory regions of RANTES [20,21].

The protective effect of CCR5-Δ32 was not observed in all longitudinal studies [22], and conflicting results have been also observed for other chemokine receptors polymorphisms among various longitudinal cohorts, and also between longitudinal and transversal studies. In the present paper, we review the conflicting results stemming from various studies and discuss the reasons for the observed discrepancies. Finally, we try to set-up a statistical model allowing to transpose data from extreme case-control studies to predict the results one should obtain on a longitudinal cohort.

At a time, when pharmacogenomic studies are ever-developing, this work has importance not only for AIDS but for all diseases involving extreme patterns of progression: diseases induced by chronic exposure to an external agent, such as tobacco in lung cancer; or chronic clinical manifestation as in autoimmune diseases or hepatitis C; and where the genetic background of the individual is likely to influence the final outcome to therapies.

2. Material and methods

2.1. Cohorts

The review of the results is based on studies published in peer-reviewed journals. The size of the longitudinal AIDS cohorts range from 100 to 700 patients with known seroconversion dates. We voluntarily avoid seroprevalent (participants already HIV-1 infected at study enrollment) cohorts since they can induce biases in the results [23]. Transversal studies have dealt mainly with the genetics of resistance to immunodeficiency virus (GRIV) cohort developed by our group [24], which includes 250 SP and 90 rapid progressors. The GRIV cohort SP subjects correspond to 1% of the active files in hospitals which mean that the GRIV cohort corresponds to the extremes of 25&puncsp;000 patients at all stages.

2.2. Genotyping

The genotyping of single nucleotide polymorphisms (SNPs) and in the case of CCR5-Δ32, a 32 bp insertion/deletion was usually performed using PCR/sequencing, PCR/RFLP or other PCR-based techniques. All GRIV participants were typed using PCR-RFLP.

2.3. Statistical methods

The statistical methods published were based on Kaplan–Meier survival analysis curves, and the Cox model for longitudinal cohorts or on χ^2 evaluations for case-control studies.

The development of the model for transposition from transversal into longitudinal observations was based on survival data analysis, using non-parametric, semi-parametric, or fully parametric models, leading respectively to Kaplan–Meier estimators of the survival functions, to regression parameters estimators for the alleles and to parametric estimators of the survival functions. The model, based on transversal observations (extreme patients and controls) was built to fit with the known longitudinal data published on seroconverter cohorts: it used a Cox regression model with a Weibul baseline hazard. Once, the model has been built, we randomly generate cohorts verifying the known cohorts behavior together with the original transversal data, and simulations were performed using the survival model established. The software used was Splus.

3. Results

3.1. Review of the literature and conflicting results

3.1.1. HLA polymorphisms

The most extensively analyzed genes have been those of the HLA with studies published as early as 1990 [2] (Table 1). HIV-1 infection is a chronic disease of the immune system and the HLA locus is of prime interest since it regulates the immune response to foreign antigens: the presentation of peptidic epitopes by class I molecules (for CTLs) and by class II molecules (for helper T-cells) may vary from one individual to the other according to his HLA genotype. Initial genetic studies included less than 100 subjects and involved patients from all origins [1,2]. Only a few studies involving HLA have used larger cohorts and they date from 1996. The necessity of using longitudinal cohorts consisting of seroincident patients and not seroprevalent patients avoids frailty bias and has been emphasized previously [23]. We will limit our analysis to the results published on seroconverter cohorts. A first study on a seroconverter cohort appeared in Kaslow et al. [4] and dealt with two cohorts representing a total of 241 men including 139 subjects from the MACS cohort. That study was further extended in Keet et al. [25] with 134 seroconverters from the Amsterdam city cohort. The work of Kaslow et al. [4] presented the B27, B57, B51, A32, A25 HLA markers associated with slower progression, while, B49 met the criteria for rapid progression (Table 1). The addition of a third cohort of 134 seroconverters [25] led to similar results, but differences were seen when TAP genes were included (part of HLA locus), showing the great impor-

Table 1
Summary of the results from the largest seroconverter cohorts and extreme patients studies

References	Cohort size	Gene allele	Nb patients	*R*	*p*	Comment
Kaslow et al. [4]	241	HLA-B27	17	0.32	NS	1
	241	HLA-B57	21	0.47	NS	1
	241	HLA-B51	25	0.52	NS	1
	241	HLA-A32	21	0.62	NS	1
	241	HLA-A25	11	0.61	NS	1
	241	HLA-B49	9	1.8	NS	1
Keet et al. [25]	375	HLA-B27	30	0.4	0.003	2
	375	HLA-B57	31	0.54	0.02	2
	375	HLA-A24	50	1.57	0.004	2
Carrington et al. [5]	330	HLA-B35	NS	2.34	$2{\times}10^{-6}$	–
	330	HLA-Cw4	NS	2.41	$2{\times}10^{-7}$	–
	330	HLA-Cw12	NS	0.61	0.03	–
	144	HLA-A29	NS	3.96	0.01	3
	144	HLA-B27	NS	6.86	0.01	3
	144	HLA-B41	NS	3.89	0.03	3
Magierowska et al. [26]	153	HLA-A3	35	0.51	0.02	4
	153	HLA-B12	40	1.73	0.05	4
	153	HLA-B17	30	0.48	0.03	4
	153	HLA-B27	17	0.2	0.001	4
Hendel et al. [24]	276	HLA-A29	27	2.91	0.008	5
	276	HLA-B14	37	0.16	0.001	5
	276	HLA-B22	12	13.07	$<10^{-4}$	5
	276	HLA-B27	35	0.34	0.02	5
	276	HLA-B35	44	1.62	0.04	5
	276	HLA-B57	33	0.26	0.01	5
	276	HLA-C8	32	0.19	0.004	5
	276	HLA-C14	18	0.16	0.03	5
	276	HLA-C16	26	2.26	0.04	5
Dean et al. [13]	309	CCR5-D32	70	0.61	0.005	–
Smith et al. [16]	648	CCR5-D32	NS	0.61	0.0003	–
	648	CCR2-64I	NS	0.64	0.001	–
Mummidi et al. [22]	174	CCR2-64I	NS	0.33	0.02	6
Kostrikis et al. [28]	117	CCR2-64I	NS	NS	0.003	–
Meyer et al. [29]	506	CCR5-D32	94	0.55	<0.01	7
Magierowska et al. [26]	153	CCR5-D32	34	0.30	0.001	4
Hendel et al. [31]	237	CCR5-D32	53	0.05	$<10^{-4}$	8
McDermott et al. [17]	417	CCR5P-1	83	1.74	0.017	9
Martin et al. [18]	694	CCR5P-1	NS	1.52	0.002	9
Faure et al. [34]	426	CX3CR1-M280	16	2.13	0.04	9

The data presented in this table are taken or computed from the publications mentioned. They are all given for the dominant model, except for CCR5P-1 [17,18] and CX3CR1-M280 [34] which correspond to a recessive model. All studies were based on longitudinal cohorts except for Magierowska et al. [26] and Hendel et al. [24,31], involving extreme patients. “Nb patients” corresponds to the total number of patients carrying the allele in the model chosen. *R* = relative hazard or relative risk, *p* = *P*-value for significance, NS = not shown.

Comments: 1, two values were given for RH and we took the mean; 2, results shown are the ones, which do not take into account the TAP genes; 3, the 144 subjects are African–Americans; 4, this corresponds to the RR of 70 SPs vs. 83 standard progressors; B17 = B57 + B58; 5, this corresponds to the RR of 200 SPs vs. 76 FPs; 6, the 174 subjects are African–Americans; 7, the *P*-value is increased here because they compare the compound genotypes of CCR2, CCR5 and SDF-1; 8, this RR corresponds to the RR of 182 NPs vs. 65 RPs; however, on the 2002 cohort RR = 0.2; 9, this is the recessive model.

tance of the number of subjects to reveal robust associations. In 1999, Carrington et al. [5] dealt with 474 seroconverter patients of whom 330 were Caucasians and 144 were African–Americans. It identified in both populations clear effects for HLA-B35 and homozygosity on rapid progression, and also effects of A29, B27 and B41 among African–Americans (Table 1). The results of Kaslow et al. [4] and Carrington et al. [5] should overlap since both works rely in part on the 139 subjects from the MACS cohort: however, the results do not match. This discrepancy can be explained by the use of different statistical approaches and by the fact that the addition of 102 new subjects in the first case and of 191 new subjects in the second case modified the outcome. Interestingly, in Carrington et al. [5], the association of B27 was favoring rapid progression among African–Americans, while the other studies showed that this allele favors slow progression among Caucasians (Table 1). These data suggest again that HLA studies are very sensitive to cohort size and origin and it might need many more subjects than a few hundreds in seroconverter cohorts to yield robust results. One should note that, these two studies used different genotyping methods, serological vs. molecular.

Extreme patients case-control studies were published by Hendel et al. [24] and Magierowska et al. [26]. The work by Magierowska et al. dealt with 70 Caucasian non-progressors and 83 control infected subjects. It identified HLA A3, B17 (= B57 + B58), B27 as favoring non-progression and HLA B12 favoring rapid progression. The work by Hendel et al. dealt with 250 Caucasian non-progressors and 80 rapid progressors and identified HLA B14/C8, B27, B57, C14 as favoring non progression and HLA A29/C16, B22, B35 as favoring rapid progression as well as HLA DR11. These two latter works were based on extreme case-control studies and there are similarities for B27, B57 but there are also discrepancies. It has been shown in a more recent work on CX3CR1 [27] that these two cohorts may lead to different results probably since 70 subjects might not be a large-enough number to yield reproducible results.

Overall, the comparison of the results found for all cohorts show that B35, B27, B57, A29 have all been identified at least twice in the studies. B14 and B22, which have shown the largest odds ratios in the GRIV study, are very rare in the population (allelic frequency less than 2.5%) and this might explain why they are not seen in the other studies.

3.1.2. Chemokines receptors polymorphism

Chemokine receptors are the second set of genes, which have been most intensively analyzed for genetic associations because they have been identified as HIV-1 coreceptor together with CD4 [11,12] (Table 1). There is a dramatic effect of a polymorphism in the major HIV-1 chemokine coreceptor gene, CCR-5, a deletion of 32 bp leading to a truncated receptor associated with resistance to HIV-1 infection [13–15]. Other polymorphisms have been discovered and associations were found in CCR5 promoter [17,18] and in CCR2 [16].

The largest seroconverter cohorts described for associations of chemokine receptors in AIDS are the one used in Smith et al. [16] which included 675 Caucasian seroconverters and 154 African–American seroconverters (mostly i.v. drug users), the cohort described by Mummidi et al. [22] composed of 470 seroconverters (with about 54% Caucasians and 37% African–American), the cohort described by Kostrikis et al. [28] with 117 seroconverters mostly Caucasians, and the cohort described by Meyer et al. [29] with 506 Caucasian seroconverters. Mummidi et al. [22] could not show an effect neither of CCR5-Δ32 nor of CCR2 in the Caucasian cohort they studied (about 250 Caucasian individuals). Smith et al. [16], Kostrikis et al. [28] and Meyer et al. [29] found an effect for both CCR5-Δ32 and CCR2-64I (675, 110 and 506 individuals, respectively), however, by using compounds genotypes for CCR5, CCR2 and SDF-1 in the latter study [29]. A recent publication on the meta-analysis of CCR5-Δ32 and CCR2-64I alleles [30] confirms that both CCR5 and CCR2 polymorphisms are associated with slower progression: the total number of seroconverters studied reaches 1750. In that meta-analysis it appeared that each cohort did not necessarily exhibit an association with AIDS progression and effective associations depended on the origin of the cohort and on the route of infection [30]. In the works previously cited [16,22,28,29] the route of infection was mainly sexual.

If we compare the results on longitudinal cohorts with the ones obtained on extreme cohorts, we observe that the CCR5-Δ32 association has always been confirmed in all extreme cohorts studied [26,31,32]. The CCR2-64I effect seems weaker in case-control studies than in longitudinal cohorts [26,31], however, it becomes obvious when using compound genotypes of CCR5 and CCR2.

A SNP located in the CCR5 promoter, called CCR5P, was observed by two groups on a cohort of 417 seroconverters [17] and on a cohort of 694 seroconverters [18]. The CCR5P-1 allele is associated with faster progression to AIDS. Its effect is recessive in Caucasian seroconverter cohorts [18] while it seems dominant in African–American cohorts [33]. To analyze this association the authors have observed that the protective alleles CCR5-Δ32 and CCR2-64I are never on the same chromosome, and also never together with CCR5P-+. They distinguished among four different haplotypes (CCR5P-1/CCR5-+/CCR2-+, CCR5P-1/CCR5-+/CCR2-64I, CCR5P-1/CCR5-Δ32/CCR2-+, CCR5P-+/CCR5-+/CCR2-+) to observe the strongest effects. This effect has been confirmed in the transversal GRIV study (manuscript in preparation).

Finally, a mutation of the CX3CR1 protein was found to be associated with slower progression in a seroconverter study involving 473 patients [34]. However, the effect has not been confirmed in other longitudinal cohorts [35] and it was not observed either in the GRIV transversal observations [27].

3.1.3. Analysis of the possible causes for discrepancies

We have seen with HLA the need for large numbers of subjects in the cohorts. The lack of robustness of the data is explained by the multiple alleles (more than 10 per gene) leading to a wider variability of distribution for each cohort and also statistically weaker associations since smaller number of patients are involved for each allele. The polymorphisms in the chemokine receptors seemed more robust since they were confirmed in most studies and it fits with the fact that these polymorphisms have a larger representation of patients per allele. However, some large studies [22] exhibited results on CCR5 and CCR2 different than those of most published works and it suggests that a longitudinal cohort of size 200 might not be necessary large enough to draw reliable results. In the case of CCR5P-1, it is necessary to consider haplotypes to observe an effect of the multi-site region. It is a privileged case since high linkage disequilibrium within these closely spaced SNP does not fragment the groups carrying the CCR5 and CCR2 protective alleles, thus conserving their power. The CX3CR1 case is interesting since this longitudinal study involved 426 seroconverter patients. It was not confirmed by other longitudinal and transversal studies [27,35] probably because of the rather limited *P*-value for significance ($P = 0.04$) compared to other significant polymorphisms.

3.2. A model to transpose transversal studies into longitudinal studies

3.2.1. Representation of the problem in statistical terms

The variant allele of some polymorphisms may slow down the progress of certain pathologies. There exist three major models to compute the influence of a gene polymorphism in an association studies: the dominant model which counts the individuals carrying a given allele, the recessive model which counts the homozygous individuals for a given allele, and the allelic frequency model which counts the number of the given allele. In the present study, we have set-up the parameters of our model thanks to the results obtained with two gene polymorphisms, the 32 bp deletion of CCR5 and the mutation Pro187Ser of NQO1 (NAD(P)H quinone oxidoreductase) [36]. In brief, NQO1 is a cytosolic enzyme involved in oxidative stress that catalyses the metabolic detoxication of quinones and their derivatives, and it has also been implicated in susceptibility to TNFα-mediated apoptosis [37].

We call *X*, the random time elapsed between an initial event, which may be, but is not necessarily, the onset of a disease, and some terminal stage for it, which may be, but is not necessarily, death. In general, the effect of the allele cannot be proved by a simple comparison, on a cohort of patients, of the distributions of *X* between the two kinds of patients: those who have and those who do not have the variant allele. The reason for this is that the size necessary for a cohort to give evidence for this fact should be much bigger than the usual size of the cohorts under study. One explanation is that, when the protected patients, called the SP, are relatively rare, even if the effect of the allele is very important, it is generally hidden by the overall behavior of the cohort under study as well as by the right censoring of the data.

Instead of using only the longitudinal study of a whole cohort, we propose here to consider, together with a cohort study, an independent case-control study involving two samples of the patients having extreme behaviors and for which (at least part of) the genotype is known. The first sample, called FP is a sample of the patients who have experienced a fast progression of the disease and the second one, called SP, is a sample of the patients having experienced a slow progression of the disease. The first sample is restricted to patients such that $X < t0$, and the second sample to patients such that $X > t1$, for some fixed *t*0 and *t*1 such that $t0 < t1$. The proportions of patients having such extreme behaviors are respectively, pFP and pSP. The cohort study provides us with estimates for those parameters and their standard errors through a Kaplan–Meier estimate of the survival function of *X*, taking into account possible right censoring, and its confidence intervals at *t*0 and *t*1. The case-control study gives us estimates of the proportions of the mutant allele among FP and SP, respectively denoted p1FP and p1SP and their standard errors. The problem is to find a model for the laws of *X* under $Z = 0$ and $Z = 1$ giving coherent results for the cohort and the case-control data, the difference of the behavior between FP and SP being explained by the disparity of the proportions of the allele and the disparity of the survival functions S0 and S1.

Actually, *t*0 and *t*1 are chosen wide apart in such a way that pFP + pSP is a small proportion of the entire population of patients, usually smaller than 10%. This means that when we compare two samples of respective sizes *n*FP and *n*SP, the size of the cohort necessary to achieve the same level of evidence would be more than ten times the sum *n*FP + *n*SP.

3.2.2. Building the model

Let us consider a specific disease whose evolution goes towards a defined final stage after a time *X*. We assume that the distribution of the random variable *X* depends on a mutant allele of some gene *g*, whose proportion p1 in the general population is estimated on a cohort. Here, *Z* is a variable whose value is 1 when the mutant allele is present and 0 otherwise. The survival functions of *X* conditional on $Z = 0$ and $Z = 1$ are respectively, denoted *S*0 and *S*1:

$$S0(t) = P(X \geq t | Z = 0),$$
$$S1(t) = P(X \geq t | Z = 1).$$

The problem is to test *H*0: $S1 = S0$ against *H*1: $S1 > S0$.

The survival function of *X* in the entire population is *Sg* such that:

$$Sg(t) = \mathrm{p1} \times S1(t) + (1 - \mathrm{p1}) \times S0(t).$$

The FP is defined as the patients who reach the final stage of the disease before some fixed time *t*0. The SP is defined as the patients who did not reach the final stage after a fixed period of time *t*1 ($t1 > t0$). The proportions of FP and SP in the entire population of patients are respectively, denoted by pFP and pSP. Actually, the two probabilities pFP = 1–*Sg*(*t*0) and pSP = *Sg*(*t*1) are estimated from the cohort by using Kaplan–Meier estimator for *Sg* which allows to take into account right censored data. It implies that the global survival function *Sg* and the survival functions conditional on the mutant gene must obey the following equations:

$$Sg(t0) = 1 - \mathrm{pFP} = \mathrm{p1} \times S1(t0) + (1 - \mathrm{p1}) \times S0(t0),$$
$$Sg(t1) = \mathrm{pSP} = p1 \times S1(t1) + (1 - \mathrm{p1}) \times S0(t1),$$

$$\begin{aligned}
&P(Z = 1 | X \leq t0) \\
&\quad = P(Z = 1, \quad X \leq t0)/P(X \leq t0), \\
&\quad = P(Z = 1) \times P(X \leq t0 | Z = 1)/P(X \leq t0), \quad (\mathrm{S}) \\
&\quad = \mathrm{p1} \times (1 - S1(t0))/(1 - Sg(t0)), \\
&P(Z = 1 | X \geq t1) \\
&\quad = P(Z = 1, \quad X \geq t1)/P(X \leq t1), \\
&\quad = P(Z = 1) \times P(X \geq t1 | Z = 1)/P(X \leq t1), \\
&\quad = \mathrm{p1} \times (1 - S1(t1))/(1 - Sg(t1)).
\end{aligned}$$

System of equations (S) gives linear equations for the four unknowns $S0(t0)$, $S0(t1)$, $S1(t0)$, $S1(t1)$ as functions of p1, pFP, pSP and the proportions of allele $Z = 1$ among FP and SP, respectively, denoted p1FP and p1SP. It follows that:

$$S0(t1) = (\mathrm{pSP} \times (1 - \mathrm{p1SP}))/(1 - \mathrm{p1}),$$
$$S1(t1) = S0(t1) \times (\mathrm{p1SP} \times (1 - \mathrm{p1}))/(\mathrm{p1} \times (1 - \mathrm{p1SP})), \quad (\mathrm{S}')$$
$$S1(t0) = (\mathrm{p1} - \mathrm{p1FP} \times \mathrm{pFP} \times \mathrm{pFP})/\mathrm{p1},$$
$$S0(t0) = (1 - \mathrm{pFP} - \mathrm{p1} \times S1(t0))/(1 - \mathrm{p1}).$$

In order to find a model that can fulfill those equations, we propose to use Weibull distributions for the conditional survival functions $S0$ and $S1$, with respective scale parameters $1/\lambda0$ and $1/\lambda1$ and shape parameters $\alpha0$ and $\alpha1$:

$$S0(t) = \exp(-(\lambda0 \times t)^{\alpha0}), \quad (\mathrm{M})$$
$$S1(t) = \exp(-(\lambda1 \times t)^{\alpha1}).$$

Let $a0 = S0(t0)$, $b0 = S0(t1)$, $a1 = S1(t0)$, $b1 = S1(t1)$. Then function $S0$ meets the two points ($t0$, $a0$) and ($t1$, $b0$) and function $S1$ meets the two points ($t0$, $a1$) and ($t1$, $b1$). From Eqs. (S') and (M) we deduce the values of the four parameters $\lambda0$, $\lambda1$, $\alpha0$, and $\alpha1$ from the values of the initial parameters p1, pFP, pSP, p1FP, and p1SP. Denoting ll the function $u \longrightarrow \log(\log(u))$, we finally get:

$$\log(\lambda0) = (\mathrm{ll}(a0)\log(t1) - \mathrm{ll}(b0)\log(t0))/(\mathrm{ll}(b0) - \mathrm{ll}(a0)),$$
$$\log(\lambda1) = (\mathrm{ll}(a1)\log(t1) - \mathrm{ll}(b1)\log(t0))/(\mathrm{ll}(b1) - \mathrm{ll}(a1)), \quad (S'')$$
$$\alpha0 = -ll(b0)/(\log(\lambda0) + \log(t1)),$$
$$\alpha1 = -ll(b1)/(\log(\lambda1) + \log(t1)).$$

Replacing in Eq. (S"), the true values of the initial parameters p1, pFP, pSP, p1FP, and p1SP by their estimates from the cohort data, we obtain estimates of $\lambda0$, $\lambda1$, $\alpha0$ and $\alpha1$, and their standard errors may be deduced from the standard errors of the initial parameters. We have thus confidence bands for the distributions $S0$ and $S1$, which depend on the size of the all-stage patients' cohort and the extreme patients samples. Those confidence bands for the distributions $S0$ and $S1$, may overlap or not, giving evidence, in the latter case, of an effect of the allele: we can then reject $H0$ for $H1$ if $S1 > S0$. At this stage, one may define the smallest difference between $S0$ and $S1$, which makes sense in order to declare that the allele is protective.

3.2.3. *Proving and quantifying the effect of the allele: simulations*

Once the parametric model based both on the longitudinal cohorts and the case-control transversal data has been established (see Fig. 1A, C), one can make simulations according to that model, by generating an exact survival time (between infection and death) for SPs and FPs, still respecting the pFP, pSP, p1, p1FP, and p1SP proportions. For this, we use $S0$ and $S1$ as survival functions respectively for people carrying or not the allele. A number *ns* of simulations of samples of size *n* of the time to onset of the terminal event, including right censoring, is performed, based on the model built on the joint analysis of the cohort and the extreme cases transversal data. The data thus generated are shown to be coherent with the initial cohort and the initial transversal data. A non-parametric test of the slowing down effect of the mutant allele is performed, using Kaplan–Meier survival confidence bands. Then, if the efficiency of the mutant allele is thus proven, in order to quantify the effect of the presence of the allele Z on the progression of the disease through a single real parameter β, one assumes a Cox proportional hazard model. It means that the survival S, conditional on Z is modeled as Sc such that:

$$Sc(t|Z = z) = S0(t)^{\exp(\beta.z)}\ \beta\,\text{real}, \quad z = 0 \quad \text{or} \quad 1. \quad (\mathrm{M}')$$

Then testing $H0$: $\beta = 0$ against $H1$: $\beta < 0$ one gets the relative risk ρ for people carrying the allele versus the ones without the allele, of the terminal event to happen, which is equal to $\rho = \exp(\beta)$.

3.2.4. *Application to AIDS*

The random variable X is the time elapsed between the initial event, which is the infection time and the terminal event, which is the achievement of a CD4 level below 500. The SPs are defined as those people who still have a CD4 level over 500 after $t1 = 14$ years, and the FPs as those people who have less than 300 CD4 at a time smaller than 3 years. Thus, in that case, $t0 = 3$ and $t1 = 14$. The proportions of SPs and FPs have been estimated respectively, as pSP = 0.01 and pFP = 0.05. From the transversal data on SP and FP, the two genes alleles CCR5-Δ32 and NQO1-187Ser are suspected to exert a protective effect against disease.

(a) *Testing for a potential longitudinal effect of CCR5*. The proportion of carriers of the mutant CCR5-Δ32 protective allele is known experimentally to be as p1 = 0.18 in the entire population [13–15], the proportion of this allele among SPs is $P(Z = 1 \mid X > t1) = 0.25$ and among FPs is $P(Z = 1 \mid X \leq t0) = 0.05$ (from the GRIV data). This allows for a computation of the respective survival functions $S0$ and $S1$ as well as the relative risk $\exp(\beta)$.

We obtain the estimated Weibull parameters of $S0$, $\alpha_0 = 2.835$, $\lambda_0 = 0.1233$, and of $S1$, $\alpha_1 = 3.706$ and $\lambda_1 = 0.1057$. This leads to the survival curves $S0$ and $S1$ shown in Fig. 1A.

We now generated randomly precise survival times for FPs and SPs still respecting the proportions p1, p1FP, p1SP, pFP, and pSP and performed simulations according to the previous law. For $ns = 1000$ such simulations of a sample of size $n = 500$, we show evidence of an effect of the CCR5-Δ32 protective allele at a level of 90%. We also obtain a relative risk equal to 0.67, with a mean 95% confidence interval equal to [0.5493–0.8432], confirming the protective effect of this allele in the longitudinal analysis, with a median excellent P-value equal to 0.00172. Fig. 1B presents an example of the estimated survival curve for 500 subjects with the corresponding intervals for 95% confidence: the two confidence

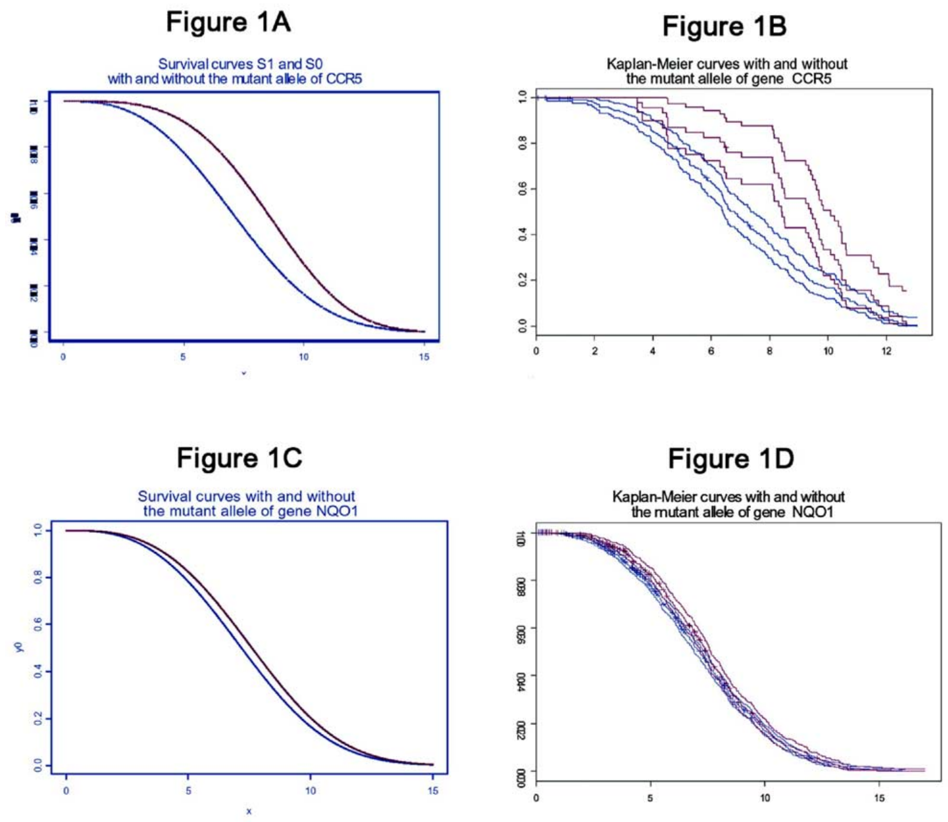

Fig. 1. (A) Survival curves *S*0 and *S*1 obtained from Eq. (M) from the raw data pFP, pSP, p1, p1FP, and p1SP. In red is the curve for the mutant CCR5-Δ32 carriers, in blue is the curve for the WT individuals. (B) Example of a survival curve obtained from a simulation for a cohort of size 500 respecting the data pFP, pSP, p1, p1FP, and p1SP proportions. The middle red curve for the CCR5-Δ32 carriers is surrounded by its 95% confidence interval curves, and the middle blue curve for the WT subjects is surrounded by its 95% confidence interval curves. The confidence intervals are clearly separated. (C) Survival curves *S*0 and *S*1 obtained from Eq. (M) from the raw data pFP, pSP, p1, p1FP, and p1SP. In red is the curve for the mutant NQO1-187Ser carriers, in blue is the curve for the WT individuals. (D) Example of a survival curve obtained from a simulation for a cohort of size 500 respecting the data pFP, pSP, p1, p1FP, and p1SP proportions. The middle red curve for the NQO1-187Ser carriers is surrounded by its 95% confidence interval curves, and the middle blue curve for the WT subjects is surrounded by its 95% confidence interval curves. In that case, the confidence intervals are totally overlapping.

intervals for the carriers of CCR5-Δ32 vs. non-carriers are clearly separated.

However, we computed the minimal size of a simulated cohort according to the survival functions *S*0 and *S*1 able to show evidence, at the more usual 95% level, of the effect of this mutant allele: it is equal to n = 600. Even if the sample size is as small as 250, and for ns = 1000 simulations, the protective effect of the mutant allele of CCR5-Δ32 will be observed in 59% of the cases. The median *P*-value we obtained is equal to 0.0287 and the mean 95% confidence interval for the relative risk is equal to [0.5071–0.9353].

(b) *Testing for a potential longitudinal effect of NQO1*. The proportion of the potentially protective NQO1-187Ser mutant is p1 = 0.346 in the entire population. Its proportion among SPs is $P(Z = 1 \mid X > t1) = 0.41$ and among FPs is $P(Z = 1 \mid X \leq t0) = 0.28$. This allows the computation of the respective survival functions *S*1 as well as the relative risk exp(β). We obtain the estimated Weibull parameters of *S*0, α_0 = 2.640, λ_0 = 0.1124, and of *S*1, α_1 = 2.796 and λ_1 = 0.1064. The representation of the survival curves for NQO1 is given in Fig. 1C.

For ns = 1000 simulations on a sample of size n = 500, we obtain a relative risk equal to 0.89, with a mean 95% confidence interval equal to [0.7405–1.076] for the mutant allele of NQO1-187Ser, thus showing no evidence for the protective effect of this allele. The median *P*-value was found to be equal to 0.2. This does not allow us to reject the null hypothesis (*P*-value equal to 0.20) that the mutant allele of NQO1 has no effect on disease progression. Fig. 1D presents an example of the estimated survival curve for a simulation on 500 subjects with the intervals for 95% confidence: the confidence intervals for the carriers of NQO1-187Ser vs. WT individuals are clearly overlapping.

If we increase the size of the samples to n = 1000 patients, still no significant effect of the allele is proven (median *P*-value is 0.0586) with a mean 95% confidence interval for the relative risk equal to [0.7405–1.076] covering 1. With ns = 1000 simulations of a sample size of n = 4000, the mean 95% confidence interval for the relative risk is [0.8213–0.939] which does not cover 1. This size of n = 4000 is the smallest size of a simulated cohort according to the survival functions *S*0 and *S*1 able to show, at a level of 95%, an effect of the NQO1-187Ser mutant allele. However, a parametric model for an observed cohort of this size may be unrealistic and we cannot conclude that the size of $n = 4000$ would allow us to prove an effect of the NQO1-187Ser mutant allele in 95% of actual cohorts.

4. Discussion and conclusion

The review of the literature has shown that major associations such as CCR5-Δ32 were robust in most cohorts; however, discrepancies arose as soon as small numbers of patients compared to the number of alleles were at stake such as in HLA, or when insufficient *P*-values were found such as in CX3CR1.

From Table 1, it appears difficult to define a cut-off of *P*-values warranting the reliability (i.e. high reproducibility) of an association identified. For instance, HLA-B57 has been regularly found associated with slow progression in various cohorts [4,24–26] but with a *P*-value higher than 0.01, while HLA-A24 was found only once associated with rapid progression with the low *P*-value 0.004. A threshold of 0.001 for *P*-values would certainly bring much reliability underlining the associations of HLA-B14, HLA-B22, HLA-B27, HLA-B35, CCR5-Δ32, CCR2-64I (Table 1). Since genomic studies aim at identifying the pathways used by the virus for its survival and its pathogenicity, presenting data with *P*-values comprised within 0.001 and 0.05 might also be relevant: in such cases, one should emphasize that the data need to be confirmed by other studies.

In most of the published studies, Bonferroni corrections were rarely applied. Published results correspond to a small percentage of the genotypes effectively analyzed in research laboratories and very few studies would remain significant if one would apply strictly the Bonferroni principle. However, this is defendable when one knows that the candidate genes studied are the most relevant ones biologically. Until now, the only clear biological explanation for an association between a genetic polymorphism and AIDS progression has been obtained for the deletion CCR5-Δ32, which inactivates the receptor.

The case of the multiple polymorphisms in the CCR5/CCR2 locus (chromosome 5) shows how the use of haplotypes might help unravel new associations or increase their significance [17,18]. On the one hand, haplotypes will diminish the strength of the studies by dissecting the population in smaller groups; on the other hand it can reveal concealed effects of some polymorphisms as for CCR5P. The study of single SNPs remains, however, indispensable since a biological effect can be linked with a single nucleotide mutation (for instance fixation of a transcription factor on a promoter).

Overall our review has shown that strong associations found in longitudinal studies are also found in transversal studies. To assess the reciprocity of this statement, we have developed a protocol for integrating transversal with longitudinal data. In order to operate such a transposition, various models are available: we used the Weibull model, which provides a flexible family of survival functions. Knowing the allelic distributions from GRIV patients and controls for the two genes CCR5 and NQO1, we set the parameters of the model to fit with the longitudinal data available on these two genes from all-stage patients' cohorts.

When there is evidence of the effect of an allele through the identification of a model integrating both the longitudinal cohort data and the extreme cases transversal observations, then the disparity of the survival functions with and without the given allele gives credit to the protective effect of the allele. The relative risk of the people carrying the allele vs. the ones without it, to develop the terminal event, may then be estimated through a Cox proportional hazard model used on a simulation of the obtained coherent model. Actually, though Cox model assumes that the effect of the allele on the risk is constant in time, which is not always true, this relative risk is an easy indicator of the amount of the protective effect of the allele. We will try to take into account a possible time-dependency of the risk in future models.

The simulations based on our model suggest that the smallest sample size allowing for a cohort to show a 95% significant difference between the survival curves *S*1 and *S*0 with and without the mutant allele CCR5-Δ32 is about $n = 600$ while for NQO1-187Ser it is $n = 4000$. As CCR5-Δ32 is rarer than NQO1-187Ser, it underlines the better protection given by CCR5-Δ32 compared to NQO1-187Ser. NQO1 gives an example how an effect can be observed in a transversal study but not validated in longitudinal studies because of its weakness. Interestingly, the size of 600 obtained for a 95% confidence interval in the CCR5-Δ32 allele association is fully compatible with the meta-analysis gathering all CCR5-Δ32 data where four studies have no effect while five studies exhibit a protective effect [30]. In particular, there was a study on a cohort of 254 patients with no effect of CCR5-Δ32 [22].

There may be another difficulty for proving the effect of an allele using only longitudinal studies. Real cohorts surely suffer from truncation in two respects: FPs may be too fast to enter the cohort and SPs may be skipped also from the cohort as they show poor signs of being ill. Our simulated cohorts are more "perfect" ones in the sense that they do not suffer from any truncation even if we have assumed to suffer right censoring which is very natural when SP are expected.

Acknowledgements

The authors are grateful to all the patients and medical staff who have kindly collaborated with the project. This work was supported by grants of ANRS (Agence Nationale de Recherche sur le SIDA) and ACV Development Foundation.

References

[1] Roger M. Influence of host genes on HIV-1 disease progression. FASEB 1998;12:625–32.

[2] Al Jabri AA. HLA and in vitro susceptibility to HIV infection. Mol Immunol 2002;38:959–67.

[3] Hill AV. HIV and HLA: confusion or complexity. Nature M 1996;2: 395.

[4] Kaslow RA, et al. Influence of combinations of human major histocompatibility complex genes on the source of HIV-1 infection. Nature M 1996;2:405–11.

[5] Carrington M, et al. HLA and HIV-1: heterozygote advantage and B35-Cw4 disadvantage. Science 1999;283:1748–52.

[6] Khoo SH, et al. Tumor necrosis factor c2 microsatellite allele is associated with rate of HIV disease progression. AIDS 1997;11: 423–8.

[7] Bream JH, et al. Polymorphisms of the human IFNG noncoding regions. Immunogenetics 2000;51:50–8.

[8] Nakayama EE, et al. Protective effect of IL-T polymorphism on HIV-1 disease progression: relationship with viral load. J Infect Dis 2002; 185:1183–6.

[9] Shin HD, et al. Genetic restriction of HIV-1 pathogenesis to AIDS by promoter alleles of IL10. Proc Natl Acad Sci, USA 2000;97: 14467–72.

[10] Garred P, et al. Susceptibility to HIV infection and progression of AIDS in relation to variant alleles of mannose-binding lectin. Lancet 1997;349:236–40.

[11] O'Brien SJ, Moore JP. The effect of genetic variation in chemokines and their receptors on HIV transmission and progression to AIDS. Immunol Rev 2000;177:99–111.

[12] Feng Y, Broder CC, Kennedy PE, Berger EA. HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. Science 1996;272:872–7.

[13] Dean M, et al. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CCR-5 structural gene. Science 1996;273:1856–62.

[14] Samson M, et al. Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature 1996;382:722–5.

[15] Rappaport J, Cho Y- Y, Hendel H, Schwartz E, Schächter F, Zagury J- F. The CCR5 32 bp deletion confers resistance to fast progression among HIV-1 infected heterozygous individuals. Lancet 1997;349: 922–3.

[16] Smith MW, et al. Contrasting influence of CCR2 and CCR5 variants on HIV-1 infection and disease progression. Science 1997;277: 959–65.

[17] McDermott DH, Zimmerman PA, Guignard F, Kleeberger CA, Leitman SF, Murphy PM. CCR5 promoter polymorphism and HIV-1 disease progression. Lancet 1998;352:866–70.

[18] Martin MP, et al. Genetic acceleration of AIDS progression by a promoter variant of CCR5. Science 1998;282:1907–11.

[19] Winkler C, et al. Genetic restriction of AIDS pathogenesis by an SDF1 chemokine gene variant. Science 1998;279:389–93.

[20] McDermott DH, et al. Chemokine RANTES promoter polymorphism affects risk of both HIV infection and disease progression in the multicenter AIDS cohort study. AIDS 2000;14:2671–8.

[21] An P, et al. Modulating influence on HIV/AIDS by interacting RANTES gene variants. Proc Natl Acad Sci, USA 2002;99:10002–7.

[22] Mummidi S, et al. Genealogy of the CCR5 locus and chemokine system gene variants associated with altered rates of HIV-1 disease progression. Nature M 1998;4:786–93.

[23] Smith MW, et al. CCR2 chemokine receptor and AIDS progression. Nature M 1998;3:1052–3.

[24] Hendel H, et al. New class I, and II HLA alleles strongly associated with opposite patterns of progression to AIDS. J Immunol 1999;162: 6942–6.

[25] Keet IP, et al. Consistent associations of HLA class I, and II and transporter gene products with progression of HIV-1 infection in homosexual men. J Infect Dis 1999;180:299–309.

[26] Magierowska M, et al. Combined genotypes of CCR5, CCR2, SDF1, and HLA genes can predict the long-term nonprogressor status in HIV-1 infected individuals. Blood 1999;93:936–41.

[27] Hendel H, et al. Validation of genetic case-control studies in AIDS and application to the CX3CR1 polymorphism. J Acquir Immune Defic Syndr 2001;26:507–11.

[28] Kostrikis LG, et al. A chemokine receptor CCR2 allele delays HIV-1 disease progression and is associated with a CCR5 promoter mutation. Nature M 1998;4:350–3.

[29] Meyer L, et al. CC-chemokine receptor variants, SDF-1 polymorphism, and disease progression in 720 HIV-infected patients. AIDS 1999;13:624–5.

[30] Ioannidis JP, et al. Effects of CCR5-D32, CCR2-64I, and SDF-1 3'A alleles on HIV-1 disease progression: an international meta-analysis of individual-patient data. Ann Intern M 2001;135:782–95.

[31] Hendel H, et al. Distinctive effects of CCR5, CCR2, and SDF1 genetic polymorphisms in AIDS progression. J Acquir Immune Defic Syndr Hum Retrovirol 1998;19:381–6.

[32] Stewart GJ, et al. Increased frequency of CCR-5 delta 32 heterozygotes among long term non progressors with HIV-1 infection. AIDS 1997;11:1833–8.

[33] An P, et al. Influence of CCR5 promoter haplotypes on AIDS progression in African–Americans. AIDS 2000;14:2217–22.

[34] Faure S, et al. Rapid progression to AIDS in HIV+ individuals with a structural variant of the chemokine receptor CX3CR1. Science 2000; 287:2274–7.

[35] Mc Dermott, et al. Genetic polymorphism in CX3CR1 and risk of HIV disease. Science 2000;290:2031.

[36] Traver RD, et al. NAD(P)H:quinone–oxidoreductase gene expression in human colon carcinoma cells: characterization of a mutation which modulates DT-diaphorase activity and mitomycin sensitivity. Cancer Res 1992;52:797–802.

[37] Siemankowski LM, Morreale J, Butts BD, Briehl MM. Increased TNF-α sensitivity of MCF-7 cells transfected with NAD(P)H:quinone reductase. Cancer Res 2000;60:3638–44.