Human cytomegalovirus (HCMV) ORF UL73 is a polymorphic locus, encoding the viral glycoprotein gpUL73-gN, a component of the gC-II envelope complex. The previously identified gN genomic variants, denoted gN-1, gN-2, gN-3 and gN-4, were further investigated in this work by analysing a large panel of HCMV clinical isolates collected from all over the world (223 samples). Sequencing and phylogenetic analysis confirmed the existence of the four gN genotypes, but also allowed the identification of a novel subgroup belonging to the gN-3 genotype, which was designated gN-3b. The number of non-synonymous (dN) and synonymous (dS) nucleotide substitutions and their ratio (dN/dS) were estimated among the gN genotypes to evaluate the possibility of positive selection. Results showed that the four variants evolved by neutral (random) selection, but that the gN-3 and gN-4 genotypes are maintained by positive selective pressure. The 223 HCMV clinical isolates were subdivided according to their geographical origin, and four main regions of gN prevalence were identified: Europe, China, Australia and Northern America. The gN variants were found to be widespread and represented within the regions analysed without any significant difference, and no new genotype was detected. Finally, for clinical and epidemiological purposes, a rapid and low-cost method for genetic grouping of the HCMV clinical isolates was developed based on the RFLP revealed by SacI, Scal and SalI digestion of the PCR-amplified UL73 sequence. This technique enabled us to distinguish all four gN genomic variants and also their subtypes.

INTRODUCTION

Human cytomegalovirus (HCMV) is a ubiquitous beta-herpesvirus associated with a wide spectrum of disease, particularly in the newborn and in immunocompromised patients. The virus has a broad spectrum of infectivity, as documented by the various different cell types that can be infected in vitro and by the multiple organ tropism observed during infection in vivo. Differences in tissue tropism, as well as differences in the severity of clinical manifestations and in the ability of HCMV to establish persistent or latent infections, have been proposed to be related to genomic variability among strains (Chou & Dennison, 1991; Fries et al., 1994; Brown et al., 1995; Meyer-Konig et al., 1998a, b).

Viral envelope glycoproteins are usually considered important targets for neutralizing antibodies and can induce strain-specific antibodies (Klein et al., 1999). In addition, these virion components play an important role in virus
entry and cell-to-cell virus spread (Navarro et al., 1993; Rasmussen et al., 1997). Interestingly, these virion proteins have also been shown to exhibit genetic polymorphism.

Glycoprotein B (gB; gpUL55), a component of the envelope complex gC-I, is the most widely studied polymorphic viral glycoprotein. Four main gB genotypes (gB-1, gB-2, gB-3 and gB-4) have been found (Chou et al., 1994; Meyer-König et al., 1998a, b), although some rare new non-prototypic variants have been described (Shepp et al., 1998; Trincado et al., 2000). Data about the functional role of the main gB types in HCMV pathogenesis and their relationship with the clinical outcome of HCMV disease and virus cell/tissue tropism are controversial and in some cases, contradictory (Fries et al., 1994; Chern et al., 1998; Gilbert et al., 1999).

The components of the gC-III glycoprotein envelope complex, gH, gL and gO, encoded by the ORFs UL75, UL115 and UL74, respectively, have also been shown to be polymorphic. For gH and gL only two different genomic variants (gH-1 and -2, and gL-1 and -2, respectively) have been identified so far (Chou, 1992a; Rasmussen, 2001), while gO shows four different genotypes (Rasmussen, 2001; Paterson et al., 2002).

Another HCMV polymorphic surface glycoprotein has been discovered recently. This protein, named gpUL73-gN and encoded by the ORF UL73, is a component of the envelope complex gC-II in association with gM (Mach et al., 2000; Dal Monte et al., 2001; Pignatelli et al., 2002) and is able to induce neutralizing antibodies (Britt & Auger, 1985). UL73 shows four main genomic variants, denoted gN-1, gN-2, gN-3 and gN-4, and the gN-4 genotype can be further divided into three subgroups (gN-4a, gN-4b and gN-4c) (Pignatelli et al., 2001). These gN genomic variants are stable during in vitro and in vivo virus replication and show no correlation with gB genotypes. Furthermore, the UL73 gN-encoding gene seems more polymorphic than gB (Pignatelli et al., 2001) and is located just upstream of the UL74 gO gene, a new hypervariable locus in HCMV (Paterson et al., 2002).

This work has focused on the gN genotyping of a large population of HCMV clinical isolates (223 samples), collected from all over the world, in order to investigate: (i) the phylogeny of gN genomic variants in a large sample of HCMV strains; (ii) the worldwide geographical distribution of the identified variants; and (iii) the ratio of non-synonymous/synonymous nucleotide substitutions \( d_{N}/d_{S} \) to evaluate the action of selective pressure on gN genotypes. Moreover a new, rapid and inexpensive assay to detect gN genomic variants by PCR-RFLP analysis has been developed for epidemiological and clinical purposes.

METHODS

Study population and clinical features. HCMV isolates (223) from different patient populations were analysed. Strains were obtained from 20 children infected post-natally, 97 congenitally infected infants, 29 AIDS patients, 50 solid organ transplant recipients, 10 bone marrow transplant recipients and 17 isolates from patients whose underlying disease was not available.

Viral DNA extraction. Clinical specimens (urine, saliva, biopsy, blood and amniotic fluid) were collected in our diagnostic laboratory and in other centres. The geographical origin of the HCMV clinical strains analysed in this study (available for 212 of the isolates) was as follows: Europe, 116; Northern America, 38; Australia, 38; China, 20.

Urine and saliva were inoculated in human embryonic lung fibroblasts (HELFs) to allow virus replication before DNA extraction. As noted above, gN genomic variants are stable during in vitro virus replication (Pignatelli et al., 2001). In contrast, the other clinical specimens were used for direct DNA extraction and PCR amplification.

Viral DNA extraction was optimized in relation to the type of specimen. Specifically, four different methods were used depending on the source of virus: (i) viral DNA was extracted from infected HELFs according to the method of Pignatelli et al. (2001); (ii) amniotic fluid was treated with IsoQuick Nucleic Acid Extraction Kit (ORCA Research); (iii) the Instagene Whole Blood Kit (Bio-Rad) was used to collect DNA directly from blood samples or PMNLs, isolated using dextran, as described by Van der Bij et al. (1988); and (iv) DNA from biopsies was extracted with the DNeasy Tissue Kit (Qiagen). The DNA obtained from clinical specimens was used for subsequent PCR amplification and restriction analysis.

PCR amplification, sequencing and gN genotyping of clinical strains. DNA was diluted in water to optimal concentration for use as a template for PCR amplification of the entire ORF UL73 from the HCMV genome. PCR was performed on ~300 ng total DNA using a pair of primers that had been selected in the conserved region surrounding UL73, shown by previous sequencing analysis (Pignatelli et al., 2001). This method allowed amplification of the ORF from all the samples tested, without the need for a nested step, giving usable amounts of PCR products. The chosen primers were: gN-up (5′-TGGTGTGATGGAGTGGAAC-3′; nt 105730–105748, with reference to the AD169 genome) and gN-lw (5′-TGGTGTGATGGAGTGGAAC-3′; nt 106130–106149). After a hot-start step for AmpliTaq-Gold (Perkin-Elmer) activation, samples underwent 35 cycles of denaturation at 96°C for 1 min, annealing at 55°C for 1 min and extension at 72°C for 1 min, followed by a final extension at 72°C for 10 min. Primers for PCR amplification were synthesized by Primm Srl. Amplification was carried out with a PTC-200 Thermal Cycler (MJ-Research).

UL73 sequences were determined for purified PCR-amplified fragments (QIquick PCR Purification Kit; Qiagen) by automated DNA sequencing (MWG-Biotech). Genotyping was performed according to Pignatelli et al. (2001) and allowed the assignment of all HCMV isolates analysed to one of the four gN genotypes.

Evaluation of \( d_{S} \) and \( d_{N} \) nucleotide substitutions and search for positive selection. Seven consensus sequences representative of each genotype were constructed by alignment of the entire set of sequences and used to search for positive selection on UL73 among genotypes. The estimated number of \( d_{S} \) and \( d_{N} \) nucleotide substitutions for all pairwise sequence comparisons, both within and between samples, was calculated by the method of Nei & Gojobori (1986).

In order to identify particular regions where positive selection might be operating, the window analysis was performed according to Endo
et al. (1996) using the WINA program. Briefly, synonymous distance and non-synonymous distance (Nei & Gojobori, 1986) were estimated along the codon site for a sequence of 20 codons in length (window region) on an alignment of homologous genes. In the window analysis, a region showing $d_N > d_S$ was considered the within-gene region on which positive selection may operate.

**Sequence and phylogenetic analysis.** The nucleotide and amino acid sequences described in this study were aligned using the MultiAlign software (http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html), option MAP (Huang, 1994) and displayed as a printable output by the BOXSHADE Server (http://www.ch.embnet.org/software/BOX_form.html). Phylogenetic analysis was performed by the Hein method (Hein & Stovlbaek, 1996) and the Lasergene ‘99 System package (DNASTAR), choosing the MegAlign option. Phylogenetic analysis was confirmed by ClustalW software and MEGA software alignment and tree display methods (Eddy, 1995).

**RFLP analysis for gN variant detection.** UL73 restriction pattern prediction was performed with Rebase software (http://www.rebase.neb.com/NEBcutter/) on the previously sequenced HCMV isolates. According to the predicted digestions, a set of three enzymes was selected to distinguish the four gN genomic variants and their subgroups. The enzymes chosen were SacI, ScaI and SalI (Mbi Fermentas); a combination of the RFLP patterns created by restriction using these endonucleases identified each genogroup (see Table 2 for details). RFLP analysis was carried out by digesting 15 μl of PCR products at 37°C for 2 h; the digested fragments, supplemented with tracking dye (2 μl of 0.25% bromophenol blue, 0.25% xylene cyanol and 30% glycerol), were electrophoresed in TAE buffer (40 mM Tris/acetate, 10 mM EDTA, pH 8) on 3.5% Nusieve agarose gels (FMC Bioproducts). They were sized by comparison with the migration distances of 50 bp ladder standard size DNA marker (Mbi Fermentas) and visualized by ethidium bromide staining. Gels and fragment size analysis were performed with the EagleEye II Imaging Acquiring System (Stratagene).

**Statistical analysis.** Distributions of the gN genotypes and the relationship to geographical origin of HCMV strains were analysed using contingency tables and chi-square test.

**RESULTS**

**Phylogenetic analysis of gN genotypes from 223 HCMV clinical isolates: identification of a novel subgroup**

Sequencing analysis of ORF UL73 obtained from this large population of HCMV clinical isolates (223 samples) confirmed the presence of four main gN genomic variants, as previously described (Pignatelli et al., 2001). The branch identifying the gN-3 group (Fig. 1) could be divided into two phylogenetically linked subgroups, similarly to those previously described for the gN-4 genotype, which shows three distinct subtypes (gN-4a, gN-4b and gN-4c). The subgroup gN-3a was previously described as the gN-3 genotype, while the new branching analysis of this larger panel of HCMV isolates identified the novel subgroup gN-3b.

When compared with group 1, the arbitrary reference group, which includes the AD169 laboratory-adapted strain, the subgroup gN-3b showed 66–68 nucleotide substitutions, of which 57% were non-synonymous (15·7% of divergence from AD169; 99–100% of identity intra-group). This led to 30–31 amino acid substitutions with respect to group 1 isolates (Fig. 2).

The genotypic frequencies of the variants analysed were as follows: gN-1 27·3%, gN-2 2·2%, gN-3 15% and gN-4 55·5%. These results confirmed previous findings and indicated that the most common variant is gN-4, while the gN-2 genotype is the rarest.

**Estimation of the $d_N/d_S$ ratio**

The comparison of synonymous (silent) and non-synonymous (amino acid-altering) substitutions among HCMV gN genotypes was performed according to Nei & Gojobori (1986) and a distance matrix was constructed. Considering the entire gene UL73, data were consistent with the absence of positive selection, since the $d_N$ value was never greater than the $d_S$ value (see Table 1 for details).

![Fig. 1. Results of phylogenetic analysis of 189 UL73 DNA sequences. The tree was generated using MegAlign software. Scale measures distance among sequences.](http://vir.sgmjournals.org)
Using the WINA program (Endo et al., 1996), the average number of dS and dN and the dN/dS ratio at each window were calculated. The graph shown in Fig. 3, where the dN/dS ratio is plotted along nucleotide sites on the ORF UL73, shows the results of the window analysis for each gN genotype. The data showed that genotypes gN-3 and gN-4 contain regions on which positive selection is operating (characterized by a dN value larger than dS and thus a dN/dS ratio >1), whereas positive selection did not seem to be acting on gN-1 and gN-2. Furthermore, positive selective pressure acted in different domains depending on the genotype considered; the regions subjected to positive selection comprised aa 60–73 of gN-3 and aa 8–23 of gN-4.

### Worldwide geographical distribution of gN genomic variants

The population examined in this work was divided into four groups according to the geographical provenance of HCMV clinical isolates. This information was available for 212 strains, as summarized in Methods.

Fig. 4 shows a map of the world, with the four main regions of origin of the clinical strains (Europe, Australia, Northern America and China) analysed in this study shown in different colours and the genotypic frequencies of each gN
variant represented by the histograms; for each region the number of cases analysed is indicated.

The variants gN-1, gN-3 and gN-4 and their subgroups were detectable in all four geographical areas, although all had different frequencies. The rarest genotype, gN-2, was well represented in Northern America and detectable in Europe, but was not identified in Chinese or Australian samples. No new variants were detected in the geographical regions examined in this study.

Statistical analysis on the genotypic frequency distribution showed no statistically significant association \( (P > 0.05) \) of gN genomic variants and geographical source of the isolates. Interestingly, Europe, Australia and China showed a superimposable distribution of genotypes.

**RFLP analysis for detection of gN variants**

As reported in Methods, UL73 restriction pattern prediction, performed by Rebase software on the previously sequenced HCMV isolates, allowed the selection of a set of three enzymes to distinguish the four gN genotypes and their subgroups. The enzymes chosen were *SacI*, *ScaI* and *SalI*; a combination of the RFLP patterns created by these restriction endonucleases identified each genotype. Table 2 shows the predicted digestions for each gN variant and its subtypes, listing the number of cut sites and the expected
size of the fragments of each UL73 genotype following digestion. Each gN genomic variant can be distinguished by these predicted restriction patterns. For the most represented genotype, gN-4, five distinct restriction patterns should be detectable.

The HCMV DNAs whose gN genotype had previously been characterized by sequencing analysis were used as a control of digestion predictions. Fig. 5 shows the restriction patterns of the four gN variant prototype strains; by comparing these patterns with the expected fragment sizes (Table 2), we were readily able to deduce the gN genotype of each strain.

In order to evaluate further the reliability of the PCR-RFLP-based assay, 60 HCMV strains, previously sequenced in the UL73 genomic region, were RFLP-genotyped as described above, in a blind assay (six cases of gN-1 type, four of gN-2, seven of gN-3a, five of gN-3b, 11 of gN-4a, 12 of gN-4b and 15 of gN-4c). There was no discrepancy between subtyping based on UL73 sequences and subtyping based on UL73 PCR-RFLP.

To evaluate the accuracy of the method, 30 newly collected samples were sequenced and found to corroborate the genotype assigned on the basis of a previously performed RFLP-typing assay. Thus, this simple method was considered useful for genetic subtyping.

**DISCUSSION**

This work has focused on the analysis of gpUL73-gN polymorphisms in a large panel of HCMV clinical isolates collected from four different regions of the world: Europe, Northern America, China and Australia. This study
Analysis of HCMV gpUL73-gN genomic variants

is the most comprehensive CMV comparison thus far reported.

Since human cytomegalovirus is considered to be widespread and has long been established in humans, the relatively limited gN variation encountered in the population of HCMV isolates examined in our previous work (Pignatelli et al., 2001) raised the possibility that our strains reflected only the local distribution of gN genotypes and did not represent the gN genotypes present in other locations. By analysing a larger population, we confirmed the existence of four main gN genomic variants and no new genotype was detected, although the previously identified gN-3 group showed an internal subdivision into two phylogenetically linked subgroups, denoted as gN-3a and gN-3b, which was analogous to the previously described gN-4 genotype subdivision.

In this new evaluation on a larger sampling of HCMV isolates, we found that gN variants are widespread but do not necessarily exhibit the same frequency distribution. Although there was no geographical grouping of isolates that seemed to have evolved completely independently from strains collected in other regions of the world, we cannot exclude the possibility that rare or new non-prototypic variants exist elsewhere, probably in regions geographically and migrationally isolated from the rest of the world.

gpUL73-gN has been demonstrated to have an N-terminal polymorphic domain (Pignatelli et al., 2001). Since gN, complexed with gM, appears to be a target of human antibody responses during natural infection (Kari & Gehrz, 1990; Mach et al., 2000), gN could also be subjected to selective immune pressure. To evaluate whether the gN polymorphisms had been subjected to positive selection, the \( d_{NS}/d_{S} \) nucleotide substitutions ratio was analysed. An excess of non-synonymous substitutions (\( d_{NS}/d_{S} > 1 \)) is considered an unequivocal index of positive selective pressure at the molecular level (Endo et al., 1996; Zhang et al., 1997). The data presented in this study indicated that, overall, the gene UL73 had not evolved genomic variants as a response to positive selective pressure, but possibly secondarily to neutral selection (Kimura, 1983).

Interestingly, after the divergence of the four gN genotypes, we showed by window analysis that the gN-3 and gN-4 types are maintained and supported by positive selection, which operates in two different domains of the protein, depending on the genotype examined. However, the significance of these results is unclear and requires additional study.

In general, a preponderance of synonymous substitutions (silent, indicating neutral or negative selection) have been reported in the majority of genes from various organisms, including viruses (Endo et al., 1996; Gojobori et al., 1990, 1994), but the existence of positive selection at the molecular level, represented by an excess of non-synonymous (amino acid-altering) substitutions, has been detected for surface molecules that interface with the immune system (Ina & Gojobori, 1994; Endo et al., 1996; Yamaguchi-Kabata & Gojobori, 2000). This kind of analysis has never been performed for other polymorphic HCMV genes, such as UL55 (gB), UL75 (gH), or UL144.

To maintain viral infectivity, amino acids that interact with the receptor of the target cell or that are responsible for correct protein folding are those most conserved among surface proteins. These regions are subjected to negative selection to maintain the proper folding, structure and function of the protein. In contrast, amino acid changes may produce antigenic variations and may enable the virus to escape recognition by the host immune system (Sasaki, 1994; Endo et al., 1996). Furthermore, other selective forces, such as viral cell tropism, virulence, stability and complex formation may induce the accumulation of specific viral variants with a higher fitness in a given selective environment (Zhang et al., 1997). Thus, it is usually accepted that both positive and negative selections are taking place during the evolution of surface proteins.

Recent studies have shown that gN genomic variants should be related to CMV-induced immunopathogenesis in the immunocompromised host and in congenital infected infants. In particular, HIV patients exhibit a prevalence of HCMV strains with the gN-1 genotype, which seem to replicate favourably in these immunocompromised hosts (Pignatelli et al., 2003a), and the gN-1 variant seems to be associated with a favourable chronic outcome in HCMV congenital infections (Pignatelli et al., 2003b). Thus, the importance of gN genotyping for epidemiological and clinical studies could require a faster and cheaper alternative to sequencing technology.

Here, we have presented a simple PCR-RFLP detection method, based on ethidium bromide-stained agarose gels, a technique readily available to all diagnostic laboratories. This method is accurate and inexpensive making the large-scale screening of gN genomic variants feasible. In addition, it eliminates the need for special equipment or facilities for sequencing of HCMV strains from different clinical specimens. This method will be extremely useful for the rapid diagnosis of HCMV infection with potentially more virulent variants and to clarify the epidemiological and clinical impact of HCMV genomic variants.

HCMV genetic diversity is relevant to vaccine design and diagnosis and may also have implications in the transmission and natural history of HCMV disease. Most candidate vaccines are produced using the gB envelope glycoprotein, but thus far these preparations have not provided the anticipated protection. We do not yet know which epitopes from gN will give protective immunity. However, gN is an exposed structural protein (Pignatelli et al., 2002), able to induce a neutralizing antibody response in the host (Britt & Auger, 1985) and whose variants show clinically relevant differences between strains (see above). These features suggest that gN may be useful in a cocktail of
immunogenic peptides, especially if proteins representative of different subtypes or typical of certain HCMV strains are used to design a prophylactic vaccine, as reviewed by Plotkin (2001).

Furthermore, UL73 appears suitable for the differentiation of strains, as it seems to be more variable than gB in the HCMV genome (Pignatelli et al., 2001), and could thus be used as an efficient marker for epidemiological studies.

ACKNOWLEDGEMENTS

This work was partially supported by the University of Bologna and the Italian Ministry of Education (60 and 40%) and by the AIDS Project of the Italian Ministry of Public Health.

REFERENCES


