Genetica (2008) 133:31–35
DOI 10.1007/s10709-007-9179-9

Recent duplication and positive selection of the *GAGE* gene family

Yang Liu · Qiyun Zhu · Naishuo Zhu

Received: 5 January 2007 / Accepted: 16 July 2007 / Published online: 29 July 2007

Abstract We report that the *GAGE* gene family of human Cancer/testis antigen (CTA) genes is likely to be in an early stage of its evolution. Members of this gene family are tandemly arranged on the X chromosome only in human, chimpanzee and macaque genomes and share a very high similarity. Phylogenetic trees show that the *GAGE* gene family began to duplicate after the split of human and chimpanzee. The estimated ages of the duplication events range from 4 million years ago to the present. The Ka/Ks values between the duplicates are significantly greater than 1, indicating that the mutation rate is higher in coding regions than non-coding regions of the genes, which suggests that the *GAGE* gene family is under positive selection. These findings indicate that the *GAGE* gene family may be a newly formed gene family undergoing rapid functional evolution.

Keywords *GAGE* · Primate · Duplication · Positive selection

Introduction

The *GAGE* (G Antigen) gene family is a subgroup of human cancer/testis antigen (CTA) genes, which are characterized as being restrictedly expressed in testis and some malignant tumors (Simpson et al. 2005). This gene family consists of at least eight members (*GAGE1* to 8) (de Backer et al. 1999) of high similarity in their sequences. *GAGE*s have been found to be expressed in various types of tumors, such as hepatocellular carcinoma (Kobayashi et al. 2000), stomach cancer (Kong et al. 2004), esophageal cancer (Akcakanat et al. 2006), ovarian carcinoma (Hofmann and Ruschenburg 2002), uterine cervical carcinoma Chang et al. (2005) and melanoma (Chen et al. 1998). The association between the expression of *GAGEs* and the expression of other CTA genes such as *MAGE* and *BAGE* has also been studied (Kobayashi et al. 2000; Akcakanat et al. 2006; Hofmann and Ruschenburg 2002). Most of these investigations suggest that *GAGE*s may be suitable diagnostic markers in detecting malignant diseases and targets for immunotherapy. However, little is known about the exact functions of *GAGE*s. The only exception is *GAGE*7, which is found to be an anti-apoptotic gene that confers resistance to Fas/CD95/APO-1, Interferon-γ, taxol and γ-irradiation (Cilensek et al. 2002). Much remains to be studied regarding the structure and functional domains of the expression products, the mechanisms of expression regulation and the interaction with other molecules. In this paper, we used bioinformatics methods to study the evolutionary characteristics of the *GAGE* gene family, which may provide some clues for future investigations of *GAGEs* as well as the mechanism of gametogenesis and tumorigenesis.

Materials and methods

Sequence data

All sequences were downloaded from the NCBI Genbank (http://www.ncbi.nlm.nih.gov/). The *GAGE* homologues

Yang Liu and Qiyun Zhu contributed equally to this work.

Y. Liu · Q. Zhu · N. Zhu (✉)
Laboratory of Molecular Immunology, State Key Laboratory of Genetic Engineering, School of Life Science, Fudan University, Handan Road 220, Shanghai 200433, P.R. China
e-mail: nzhu@fudan.edu.cn

were identified by running BLASTn against each available genome database available on the NCBI website (http://www.ncbi.nlm.nih.gov/blast/). Hits with an E-value greater than 0.01 were discarded. The results were then analyzed individually to obtain the precise sequences of the known and predicted *GAGE* duplicates of all species.

Sequence analysis

All *GAGE* homologues were aligned using ClustalW. The distance between each pair was calculated using MEGA3.1. In this calculation we applied the Jukes-Cantor model (Jukes and Cantor 1969) as the model of nucleotide substitution. In order to test the reliability of this calculation, a separate calculation using the Kimura 2-parameters model (Kimura 1980) was also performed and similar results were obtained (data not shown).

Phylogenies analysis

The phylogenetic trees were constructed using Maximum likelihood method (Felsenstein 1981) by DNAML of Phylip3.66.

Test for selection

The Nei & Gojobori's method (Nei and Gojobori 1986) implemented in CODEML of PAML 3.51 was used to calculate the Ka and Ks values between the human-chimpanzee and human-macaque pairs. The codon-based Fisher's exact test (Li et al. 1985) (which is suitable for sequence pairs that have poor divergence) implemented in MEGA3.1 was used over 14 human pairs. The Nielson & Yang's method (Nielsen and Yang 1998) implemented in Codeml program was then used to calculate the Ka and Ks values of every single amino acid site of human *GAGE* genes.

Results

Structure and arrangement of human *GAGE*s

All Human *GAGE* gene family members are located at Xp11.4-p11.2. Fifteen full-length duplicates can be obtained (the 1st to 7th at 1,6328–8,3004 nt of NT_079573, the 8th to 14th at 1,589–77,715 nt of NT_086939). Each duplicate is approximately 9.5 kb in length, including a 7.3 kb transcribed region (gene or pseudogene) and a 2.2 kb intergenic region located upstream of it. All duplicates rank closely in the same direction with clear repeat boundaries and no gaps between them. There is a 50 kb region between the 7th and 8th duplicates remaining unsequenced in the human genome database, suggesting that the total number of *GAGE* duplicates may be greater. The last of the 15 duplicates contains a 1 kb region at its 3′ terminal that is quite different from the other 14, so this duplicate was excluded in the following analysis.

We labeled the 14 duplicates sequentially as Human01–Human14. Within each duplicate we predicted the transcribed sequence (7.3 kb), intergenic sequence (2.2 kb) and coding sequence (351 bp or 354 bp). Four of the duplicates have been annotated as members of *GAGE* family in Genbank. Preliminary analyses of the other 10 duplicates showed that none of their coding sequences is disrupted by the stop codon and all of their introns have a GT-AG terminal structure, indicating that none of them can be asserted as a pseudogene.

GAGE homologues in other species

The BLAST results show that only the chimpanzee (Pan troglodytes) and macaque (Macaca mulatta) genomes contain homologues of the human *GAGE* gene (3 hits and 4 hits, respectively), which also clustered on the short arm of X chromosome. No hits were detected in any other available genome databases, indicating that *GAGE* family is unique to primates. By analyzing BLAST results we estimated that there are at least 8 duplicates in the chimpanzee's genome. However, due to the incompleteness and low statistical coverage of the chimpanzee genome sequence, only one full-length duplicate sequence was obtained (116,198–125,727nt of NW_001251829). The rest of the hits were fractions of full-length sequences. As for the macaque genome, at least 4 hits of *GAGE* homologues were detected. Since the data of the macaque genome were even less complete, we were unable to obtain a single full-length duplicate sequence. Thus, we resorted to joining up adjacent segments manually and got one full-length duplicate sequence (joined by 5,736–11,792nt and 2,762–5,735nt of AANU01113531).

Divergence between the *GAGE*s

We aligned the 14 human duplicates, 1 chimpanzee duplicate and 1 macaque duplicate sequences in terms of the transcribed sequence, intergenic sequence and coding sequence. The distances between each pair were calculated (Table 1), showing that members of the human *GAGE* family are highly similar. The distance between the human and chimpanzee coding sequences is greater than that for the transcribed sequences, which is opposite to most cases in which coding sequences are less likely to accumulate mutation under evolutionary restraints (Fig. 1).

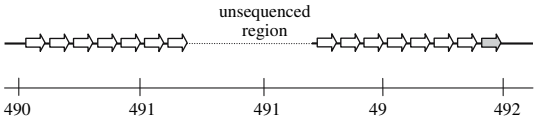

Fig. 1 Ranking of the *GAGE* duplicates on X chromosome. All 15 duplicates rank in tandem at Xp11.2–11.4. The arrows stand for the transcribed sequences and their transcribing direction. The line sections upstream of the arrows stand for the untranscribed sequences. The grey one was excluded in this study because the corresponding sequence is disrupted

Recent duplication of the *GAGE*s

We constructed the phylogenetic trees of the *GAGE* homologues (Fig. 2). The trees show that the chimpanzee duplicate is at a separate branch from the human duplicates, from which we can infer that the duplication event of human *GAGE*s occurred after the split of human and chimpanzee. The fact that the chimpanzee, macaque and human genomes each contain multiple duplicates indicates that the *GAGE* family duplicated separately in primate species.

We estimated the ages of the duplication events in human *GAGE* family from the substitution rate of the transcribed sequences and the divergence time between the human and chimpanzee lineage (approximately 6–7 million years ago) (Brunet et al. 2002). The mean age of the duplication should be $6 \times (0.71/1.88) \approx 2$ million years ago or nearer, in which 1.88(%) is the mean distance between human and chimpanzee and 0.71(%) is the mean distance within human. Using the same method, we estimated that the earliest duplication (which is between Human01 and Human13 with a maximum distance 1.26%) was no earlier than 4 million years ago and the latest duplication (which is between Human09 and Human13 with a minimum distance 0.04%) was just within the past tens of thousands of years, which implies that the *GAGE* family has duplicated continuously and may still be duplicating.

Positive selection of *GAGE*s

The commonly used method to test if positive selection is taking place on a gene is to calculate the Ka/Ks value. Ka is the number of nonsynonymous substitutions per nonsynonymous site and Ks is the number of synonymous substitutions per synonymous site. A Ka/Ks value greater than 1 is a strong evidence of positive selection (Nei 2005b).

We calculated the Ka/Ks values of each pair of the coding sequences of the *GAGE* homologues. The mean value between human and chimpanzee is 2.445 and the standard error 0.291. One-tailed *T*-test shows that the probability to accept null-hypothesis Ka/Ks = 1 is 2.55E-10, indicating that positive selection does act on the *GAGE* gene family.

We applied another method to calculate the possibilities to accept Ka/Ks > 1 of each pair of the 14 human duplicates. The results are shown in Table 2. The mean probability is 0.713 and the standard error 0.368, which reinforces the conclusion that *GAGE*s are under positive selection.

We then calculated the Ka and Ks values of every amino acid site of human *GAGE* amino acid sequence (Fig. 3). Among all 117 residues, eight were found to be positively selected (Table 3), three of which are significant.

Discussion

Our study shows that the human *GAGE* family contains 15 or more duplicates that are highly similar, the *GAGE*

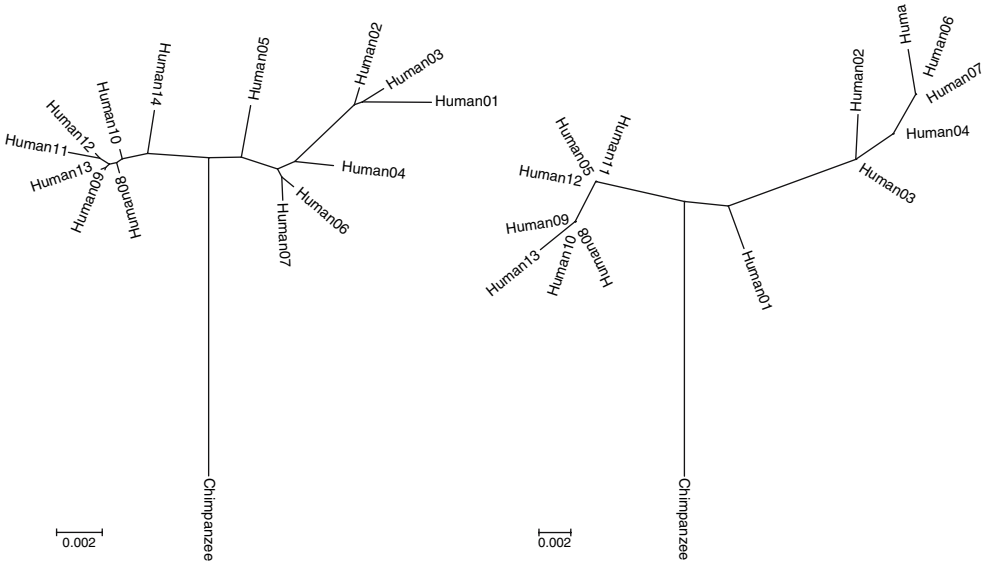

Fig. 2 Phylogenetic trees of human and chimpanzee *GAGE* homologues. In terms of full-length sequences (left) and coding sequences (right). The macaque homologue is also at a separated branch which is relatively far from human and chimpanzee. It is not shown due to limited space

Table 1 Mean Jukes-Cantor distances (%) and standard errors among duplicates

	Full-length sequence	Transcribed sequence	Intergenic sequence	Coding sequence
Within human	0.65 ± 0.06	0.71 ± 0.06	0.42 ± 0.08	1.10 ± 0.37
Human-chimpanzee	1.78 ± 0.13	1.88 ± 0.16	1.53 ± 0.24	2.68 ± 0.79
Human-macaque	11.26 ± 0.38	11.01 ± 0.36	11.89 ± 0.74	10.12 ± 1.62
Chimpanzee-macaque	11.12 ± 0.37	10.86 ± 0.35	11.82 ± 0.73	11.69 ± 1.84

Table 2 Pairwise probability of Ka/Ks > 1 of 14 human *GAGE* genes

	01	02	03	04	05	06	07	08	09	10	11	12	13
01													
02	0.692												
03	1.000	0.777											
04	0.690	0.603	0.777										
05	1.000	1.000	1.000	1.000									
06	0.599	0.468	0.603	0.777	1.000								
07	0.599	0.468	0.603	0.777	1.000	1.000							
08	0.692	1.000	1.000	1.000	0.778	1.000	1.000						
09	0.692	1.000	1.000	1.000	0.778	1.000	1.000	1.000					
10	0.692	1.000	1.000	1.000	0.778	1.000	1.000	1.000	1.000				
11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.778	0.778	0.778			
12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.778	0.778	0.778	1.000		
13	0.601	1.000	1.000	1.000	0.605	1.000	1.000	0.781	0.781	0.781	0.605	0.605	
14	0.513	0.363	0.468	0.603	1.000	0.777	0.777	1.000	1.000	1.000	1.000	1.000	1.000

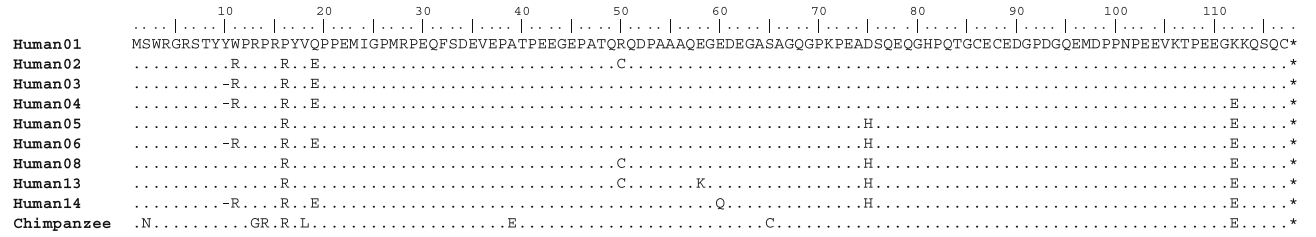

Fig. 3 Alignment of amino acid sequences of *GAGE* homologues. Human06 and Human07; Human05, Human11 and Human12; Human08, Human09 and Human10 are identical and thus only one of each group is shown

Table 3 Possible positively selected sites in human *GAGE* amino acid sequence

Position of site	Ka/Ks value and standard error	Posterior probability of Ka/Ks > 1
11	6.808 ± 3.783	0.742
16	7.030 ± 3.679	0.769
19	6.685 ± 3.834	0.728
50	8.893 ± 1.652	0.999*
58	6.515 ± 3.897	0.708
60	6.528 ± 3.892	0.709
75	8.558 ± 2.283	0.956*
112	8.547 ± 2.300	0.955*

Significant results are denoted by asterisks

homologues can only be found in the primate lineage, the divergence of the coding sequence is greater than that of the non-coding sequence, the ages of the duplication events were estimated to range from 4 million years ago to today, the calculation result met the criterion of positive selection Ka/Ks > 1. Based on the above results, we can infer that the *GAGE* family genes duplicated recently and are under positive selection.

Gene duplication is thought to be the dominating mechanism of functional evolution (Prince and Pickett 2002). The fact that the *GAGE* duplicates are located next to each other in the genome infers that unequal crossover (Smith 1976) may be the mechanism of *GAGE*s' duplication. According to the classical theory (Haldane 1933), there are two possible outcomes of gene duplication after

the duplication event. Usually one of the duplicates keeps its original function and remains steady under purifying selection while the other loses selection constraints and gradually becomes a pseudogene. Alternatively, when environmental changes require new phenotypes to emerge, mutations tend to accumulate in the gene sequence under positive selection, leading to entirely new functions or modifications of existing ones. Such conditions can only be seen in a small portion of cases (Prince and Pickett 2002).

We further inferred that *GAGE* may relate to a certain evolving characteristic of the primate species. Previous studies indicate that this gene family may play a role in gametogenesis and tumorigenesis. We have identified possible positively selected amino acid sites in this study, which may provide useful information for further investigations of the functional sites of *GAGE* gene as well as the evolution of primates.

The origin of new genes is a hot research topic. Our understanding of the molecular mechanisms and dynamics involved in the creation of new genes, however, remains unclear. Critical to this literature is the discovery of actual cases of new genes, which tend to have considerably original characteristics (Long et al. 2003). In this study we discovered that GAGE is a young gene. Therefore, further investigation of this gene may help to understand the mechanism of gene origination.

Further investigations require more data, such as more complete and reliable sequence data of the genomes of human and other species, with which we can describe the evolution of *GAGE* gene in more detail and provide new evidence to shed light on current controversies in the evolution model of the duplicated genes (Nei and Rooney 2005a) and the outcome of gene duplicates (Prince and Pickett 2002). Moreover, determining if the *GAGE* gene has the copy number polymorphism (CNP) warrants a comprehensive examination of individuals, as this phenomenon is considered to be more important than SNP in evolution (Sebat et al. 2004). Gathering additional experimental data promises to reveal more about this nascent gene and the mechanisms involved in positive selection.

Acknowledgments This work was supported by the National Sciences Foundation of China (NSFC, No. 30471906 and 30571650). We thank Fuqu Yu, Richard Callahan and Gary Potikyan for critical reading of the manuscript, and Tonghai Dou for helpful discussion.

References

Akcakanat A, Kanda T, Tanabe T et al (2006) Heterogeneous expression of GAGE, NY-ESO-1, MAGE-A and SSX proteins in esophageal cancer: implications for immunotherapy. Int J Cancer 118(1):123–128

Brunet M, Guy F, Pilbeam D et al (2002) A new hominid from the upper miocene of Chad, Central Africa. Nature 418(6894):145–151

Chang HK, Park J, Kim W et al (2005) The expression of *MAGE* and *GAGE* genes in uterine cervical carcinoma of Korea by RT-PCR with common primers. Gynecol Oncol 97(2):342–347

Chen ME, Lin SH, Chung LW et al (1998) Isolation and characterization of PAGE-1 and GAGE-7. New genes expressed in the LNCaP prostate cancer progression model that share homology with melanoma-associated antigens. J Biol Chem 273(28):17618–17625

Cilensek ZM, Yehiely F, Kular RK et al (2002). A member of the GAGE family of tumor antigens is an anti-apoptotic gene that confers resistance to Fas/CD95/APO-1, interferon-γ, taxol and γ-irradiation. Cancer Biol Ther 1(4):380–387

de Backer O, Arden KC, Boretti M et al (1999) Characterization of the *GAGE* genes that are expressed in various human cancers and in normal testis. Cancer Res 59(13):3157–3165

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17(6):368–376

Haldane JBS (1933) The part played by recurrent mutation in evolution. Am Nat 67(708):5–19

Hofmann M, Ruschenburg I (2002) mRNA detection of tumor-rejection genes *BAGE*, *GAGE*, and *MAGE* in peritoneal fluid from patients with ovarian carcinoma as a potential diagnostic tool. Cancer 96(3):187–193

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian Protein Metabolis. New York: Academic Press, pp 21–132

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16(2):111–120

Kobayashi Y, Higashi T, Nouso K et al (2000) Expression of *MAGE*, *GAGE* and *BAGE* genes in human liver diseases: utility as molecular markers for *Hepatocellular carcinoma*. J Hepatol 32(4):612–617

Kong U, Koo J, Choi K et al (2004) The expression of *GAGE* gene can predict aggressive biologic behavior of intestinal type of stomach cancer. Hepatogastroenterology 51(59):1519–1523

Li WH, Wu CL, Luo CC (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2(2):150–174

Long M, Betran E, Thornton K et al (2003) The origin of new genes: glimpses from the young and old. Nat Rev Gen 4(11):865–875

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3(5):418–426

Nei M, Rooney AP (2005a) Concerted and birth-and-death evolution of multigene families. Ann R Genet 39:121–152

Nei M (2005b) Selectionism and neutralism in molecular evolution. Mol Biol Evol 22(12):2318–2342

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148(3):929–936

Prince VE, Pickett FB (2002) Splitting pairs: the diverging fates of duplicated genes. Nat Rev Gen 3(11):827–837

Sebat J, Lakshmi B, Troge J et al (2004) Large-scale copy number polymorphism in the human genome. Science 305(5683):525–528

Simpson AJ, Caballero OL, Jungbluth A et al (2005) Cancer/testis antigens, gametogenesis and cancer. Nat Cancer 5(8):615–625

Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. Science 191(4227):528–535