4 Hackman, P. *et al.* (1997) A human compound heterozygote for two MLH1 missense mutations. *Nat. Genet.* 17, 135–136
5 Mukhopadhyay, A. *et al.* (2002) Mutations in MYOC gene of Indian primary open angle glaucoma patients. *Mol. Vis.* 8, 442–448
6 Gilbert-Dussardier, B. *et al.* (1996) Partial duplication [dup. TCAC (178)] and novel point mutations (T125M, G188R, A209V, and H302L) of the ornithine transcarbamylase gene in congenital hyperammonemia. *Hum. Mutat.* 8, 74–76
7 Pfutzer, R. *et al.* (2002) Novel cationic trypsinogen (PRSS1) N29T and R122C mutations cause autosomal dominant hereditary pancreatitis. *Gut* 50, 271–272
8 Kern, A.D. and Kondrashov, F.A. (2004) Mechanisms and convergence of compensatory evolution in mammalian mitochondrial tRNAs. *Nat. Genet.* 36, 1207–1212
9 Gao, L. and Zhang, J. (2003) Why are some human disease-associated mutations fixed in mice? *Trends Genet.* 19, 678–681
10 di Rago, J.P. *et al.* (1995) Genetic analysis of the folded structure of yeast mitochondrial Cytochrome *b* by selection of intragenic second-site revertants. *J. Mol. Biol.* 248, 804–811
11 Gagneux, P. *et al.* (1999) Mitochondrial sequences show diverse evolutionary histories of African hominoids. *Proc. Natl. Acad. Sci. U. S. A.* 96, 5077–5082
12 Shi, D. *et al.* (1998) 1.85-A resolution crystal structure of human ornithine transcarbamylase complexed with *N*-phosphonacetyl-L-ornithine. Catalytic mechanism and correlation with inherited deficiency. *J. Biol. Chem.* 273, 34247–34254
13 Shi, D. *et al.* (2001) Human ornithine transcarbamylase, crystallographic insights into substrate recognition and conformational changes. *Biochem. J.* 354, 501–509
14 Badano, J.L. *et al.* (2003) Heterozygous mutations in BBS1, BBS2 and BBS6 have a potential epistatic effect on Bardet-Biedl patients with two mutations at a second BBS locus. *Hum. Mol. Genet.* 12, 1651–1659
15 Keightley, P.D. *et al.* (2005) Evidence of widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* 3, e42
16 Whitlock, M.C. (2003) Fixation probability and time in subdivided populations. *Genetics* 164, 767–779
17 Kimura, M. (1983) *The Neutral Theory of Molecular Evolution.* Cambridge University Press
18 Moore, J.H. (2005) A global view of epistasis. *Nat. Genet.* 37, 13–14
19 Carlborg, O. and Haley, C.S. (2004) Epistasis: too often neglected in complex trait studies? *Nat. Rev. Genet.* 5, 618–625
20 Bettinelli, A. *et al.* (2005) Simultaneous mutations in the CLCNKB and SLC12A3 genes in two siblings with phenotypic heterogeneity in classic Bartter syndrome. *Pediatr. Res.* 58, 1269–1273

0168-9525/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2006.08.001

Genome Analysis

Computational characterization of multiple Gag-like human proteins

Mónica Campillos[1], Tobias Doerks[1], Parantu K. Shah[1,2] and Peer Bork[1]

[1] EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany
[2] Present address: Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA

In a genome-wide analysis, we have identified 85 human genes encoding 103 protein isoforms that resemble retroviral Gag proteins. These genes were domesticated from retrotransposons in at least five independent events during vertebrate evolution and were subsequently duplicated further in mammals. Structural insights into the mammalian proteins can be inferred by homology to Gag from viruses such as HIV; in turn, the cellular roles of the mammalian Gag homologs, such as apoptosis-related functions and binding to ubiquitin ligases, might hint at further functionality of viral Gag itself.

Introduction

Gag proteins are fast-evolving structural components of retroviruses and long terminal repeat (LTR) retrotransposons and are essential for particle formation in the viral budding process [1]. Despite the apparent importance of Gag proteins, little is known about their cellular function. Molecular functions have been assigned to three regions in the Gag sequence: (i) an N-terminal matrix domain involved in membrane binding of Gag and virion assembly; (ii) a central capsid domain that forms the core shell of virus particles, is necessary for formation of the capsid, and consists of two helical (N- and C-terminal capsid) subdomains; and (iii) a zinc-finger-containing nucleocapsid domain required for viral genome packaging and early infection process [1].

Recently, several independent studies have identified similarity between different human proteins and Gag domains of the Gypsy/Ty3 group of mobile elements [2,3]; in addition, a few of the human proteins have been proposed to be domesticated from the subgroup of Sushi LTR retrotransposons (see Ref. [4] and references therein). This group of mobile elements is phylogenetically related to retroviruses such as HIV that have the same organization of Gag, Pol and optional Env proteins [5].

To obtain a more global overview of Gag-like proteins in humans, we have performed a systematic and genome-wide analysis of globular Gag domains in mammals; genes related to mobile elements are often suppressed in automated genome annotations [6] and human cellular gag-like genes might hint at many functional aspects of viral Gags.

Mammalian genomes encode at least 103 domesticated Gag proteins

Genome searches with dedicated sequence profiles for capsid Gag domains (PSI-BLAST searches with a stringent cut-off *e*-value $<10^{-10}$; see Supplementary Material,

Corresponding authors: Campillos, M. (campillo@embl.de); Bork, P. (bork@embl.de)

Available online 18 September 2006.

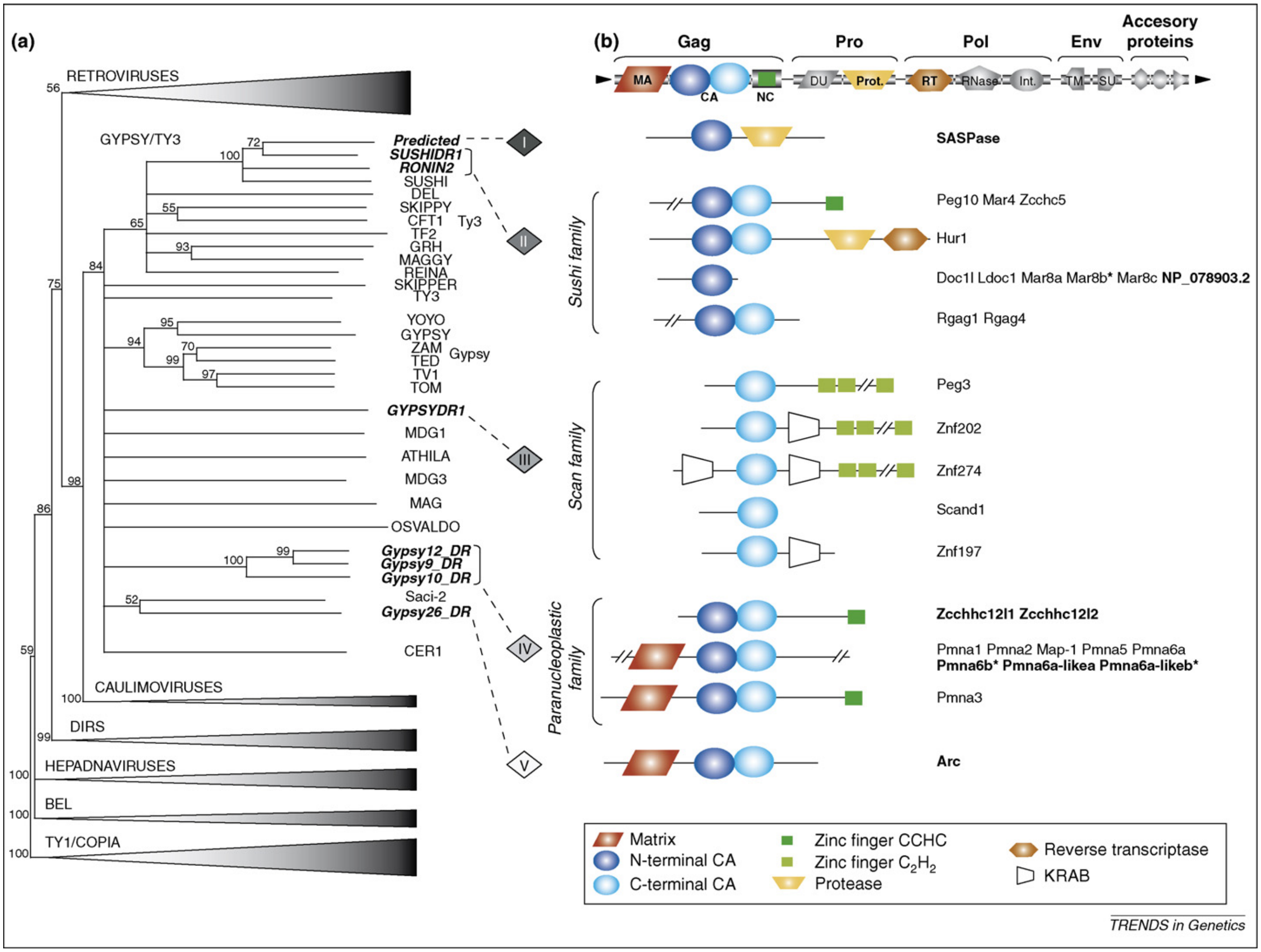

Figure 1. Domain architecture of human Gag-like proteins and their relationship to LTR retroelements. **(a)** Phylogenetic relationship of LTR retrotransposons of the Gypsy group based on reverse transcriptase sequences from which the Gag-like proteins probably originated. Bootstrap support (%) of the neighbor-joining tree for branches is shown. The most similar retrotransposons of each family of human Gag-like proteins (indicated by numbered diamonds) are marked in bold italics and connected to the families by a broken line. **(b)** The respective domain architectures of the domesticated human Gag-like proteins are compared with a simplified retrovirus domain architecture (top). Only representative sequences from the SCAN family (III) are shown. The proteins whose Gag homology was previously unknown are indicated in bold; the proteins predicted here are indicated by an asterisk. Mar8b has been reported [23] as a Gag-related protein but is not present in the current protein databases. Four previously unknown Gag homologs of the paraneoplastic family (IV) that retain only the matrix domain are not shown. Two are predicted here for the first time. Abbreviations: CA, capsid; DU, dUTPase; Env, envelope; Int, integrase; MA, matrix; NC, nucleocapsid; Pol, polyprotein; Prot, aspartic protease; RT, reverse transcriptase; SU, surface; TM, transmembrane.

sections 1 and 2, for details) identified 85 human gag-like genes (see Supplementary Material, Figure S2.3), which can be classified into five distinct families, termed I–V, on the basis of their sequence similarity to Gag proteins from different retrotransposon families (Figure 1).

We detected the retrovirus-related protease SASPase (family I) [7], for which similarity to and acquisition from Gag has not been reported before. We also identified three large gene families known to be associated with Gag: the retroviral-capsid-related Sushi family [4] (family II; 11 members described, 1 additional member identified here); the SCAN family of transcription factors [8] (family III; 56 genes annotated in the Ensembl database; structural similarity of the SCAN domain to parts of the Gag capsid domain has previously been noted [9]), and 15 human members of the recently reported Gag-related Paraneoplastic family (family IV; only six members have been previously reported [6,10]), 3 of which have been implicated in neuronal paraneoplastic diseases [11] and 1 of which, Map-1, is involved in apoptosis [12]. Finally, we found that activity-regulated cytoskeleton-associated protein (Arc) [13], currently the only member of family V, is related to Gag (see Figure 1 for an overview of all proteins in the five families and their domain architectures).

Taken together, we have identified the three known Gag-like families and two new ones (families I and V), we have predicted five novel genes, and we have shown that 12 known human proteins have homology to Gag. Although only a few of the 103 Gag-like human proteins have been characterized experimentally, the presence of orthologs of almost all of them in mouse and rat (exceptions are genes that emerged by duplication after divergence from rodents), coupled with expressed sequence tag support for most of them (see Supplementary Material, Figure S2.3), confirms that Gag-like proteins have a large functional repertoire in mammals.

Origin of Gag-like proteins in mammals

Phylogenetic analysis of Gag genes in vertebrate genomes (including those from retrotransposons) suggests that at least five independent Gag domestication events happened before the split of Eutherians (see Supplementary Material, section 5); three of the respective genes then duplicated, giving rise to the large families II–IV (Figure 1a and Supplementary Material, section 5), which seem to have expanded further through recent duplications. The exact timing of the domestication events is difficult to determine owing to limited data; at least, the Arc protein seems to have been acquired before the divergence of amphibians because the *arc* gene in *Xenopus tropicalis* is in synteny with its mammalian orthologs (data not shown).

So far, no active retrotransposons corresponding to those from which the five families of domesticated Gag proteins originated have been reported in the mammalian genomes containing the domesticated proteins, and our domain searches did not reveal their presence. For example, although the human Gag-like SASPase protein has an ortholog in rat, the respective Sushi-related retrotransposon family cannot be detected in either the rat or the human genome. This observation suggests that the Gag-like proteins might have caused the 'death' of their parental retrotransposons and might also protect the genome against infection by related viruses and retrotransposons [2]. In agreement with this hypothesis, Fv1, a Gag-like protein in mouse that is similar to a murine endogenous retrovirus, prevents infection by Friend leukemia virus [14,15].

Transfer of structural information to mammalian Gag-like proteins

The similarity of the Gag-like human genes to each other and to retroviruses with known 3D structures such as HIV (see Supplementary Material, Figure S3) enables us to transfer structural information to all 85 Gag-like human proteins, including those involved in neuronal paraneoplastic diseases (only the SCAN domain structure has been determined [9] and might have been transferred to closely related family III members). Most of the human proteins identified contain several Gag-like domains and, in fact, their domain architectures often resemble those of viral Gags for which 3D structures are available.

During the analysis, we noted some surprising sequence similarity between the N- and C-terminal capsid subdomains. A comparison of the 3D structures confirmed the homology between these two subdomains [$P(m) = 1.11 \times 10^{-3}$ (probability that sequence identity found after structure-based alignment occurred by chance); see Supplementary Material, section 4], indicating that they might have molecular functions in the mammalian Gag-like proteins similar to their functions in retroviruses.

Functional similarities among Mammalian Gag-like proteins

Although little functional information is known for most of the individual human Gag-like proteins, their common

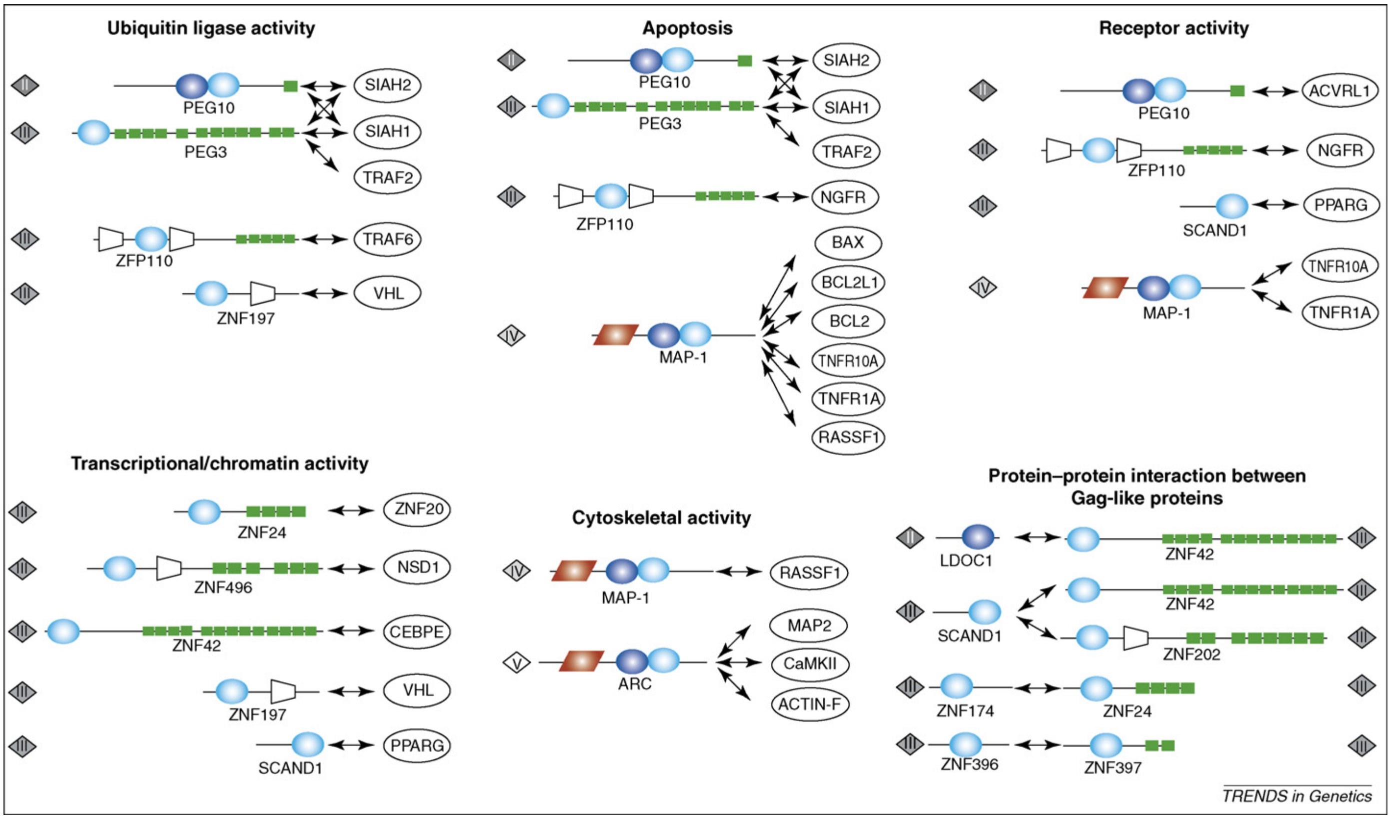

Figure 2. Functional classification of the interaction partners of mammalian Gag-like proteins. The interaction partners are organized by biological activity. Although the list is unlikely to be complete, there is a statistically significant enrichment of specific functional categories (see Supplementary Material, section 6) that might be relevant for retroviral Gag proteins. For domain symbols see Figure 1. Diamonds represent the different Gag-like families.

origin, as identified here, should provide a unified basis for the study of viral function.

The well-characterized human Gag-like proteins (Arc, Map-1, Peg-10 and Peg3) regulate apoptosis, interact with the cytoskeleton, or are involved in transcription regulation [13,16–18]; in other words, they have roles that are compatible with those ascribed to viral Gag proteins. Interaction partners have been identified for several human Gag-like proteins (Table 6.2) and the functions of these partners also reflect the roles of Gags; for example, protein–protein interactions between human Gag-like proteins and E3 ubiquitin ligases, nuclear receptors and transmembrane receptors seem to be a common theme (see Figure 2 for an overview and Supplementary Material, section 6, for the statistically significant enrichment of certain functions in human Gag-like proteins and their interaction partners). In this regard, the protein responsible for protecting rhesus macaque against HIV-1 infection, TRIM5α, interacts specifically with the Gag capsid domain and has been proposed to have ubiquitin ligase activity [19].

The functions known for human Gag-like proteins are compatible with those inferred from the analysis of their interacting partners, and both together suggest that retroviral Gag proteins interfere with (anti)apoptotic cellular signaling in the same way as human Gag-like proteins such as Map-1, which triggers apoptosis by interacting with cytoplasmic receptor regions and apoptotic mitochondrial proteins [16]. Similar interactions could also be responsible for the membrane reshaping in which viral Gag has been implicated during the budding process of the retrovirus [1]. This process resembles the synaptic remodeling process in neurons, in which the human Gag-like Arc protein is thought to participate and which requires membrane and cytoskeleton reorganization [13].

A comparison of the domain architectures of Gag-like proteins involved in similar biological processes (inferred from an analysis of interactions partners; Figure 2) suggests that the matrix domain confers a cytoskeletal binding property, in agreement with actin binding of viral Gags (reviewed in Ref. [20]). The C-terminal capsid subdomain seems to be involved in the interaction with receptors and ubiquitin ligases, although the involvement of the faster-evolving, homologous N-terminal capsid subdomain cannot be discarded.

Concluding remarks

Taken together, the unified identification and characterization of the domain architecture of 85 human Gag-like proteins not only enables their structural annotation but also facilitates the transfer of functional information from viral Gag proteins to the human proteins and vice versa. The domestication, and likely integration into apoptosis regulation, of many of the human Gag-like proteins illustrates how flexible mammalian genomes are in the evolution of gene function. The continuous expansions of these gene families indicate the fast diversification among mammals of the processes in which they are involved.

Our in-depth study, targeted at a single protein (Gag) from a subgroup of mobile elements, suggests that many more mammalian genes might have originated from mobile elements than is currently thought [21]. Only around 40 domesticated genes had been previously identified in the human genome by large-scale studies aimed at all proteins encoded by mobile elements [6,22]; it is now likely that several hundreds exist.

Acknowledgements

We thank members of the Bork group and Aidan Budd for helpful discussions. M.C. is recipient of a FEBS long-term fellowship. This work was supported in part by the Bundesministerium fuer Bildung und Forschung (Nationales Genomforschungsnetz grant 01GR0454 to P.B.).

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2006.09.006.

References

1 Morita, E. and Sundquist, W.I. (2004) Retrovirus budding. *Annu. Rev. Cell Dev. Biol.* 20, 395–425

2 Lynch, C. and Tristem, M. (2003) A co-opted gypsy-type LTR-retrotransposon is conserved in the genomes of humans, sheep, mice, and rats. *Curr. Biol.* 13, 1518–1523

3 Voltz, R. *et al.* (1999) A serologic marker of paraneoplastic limbic and brain-stem encephalitis in patients with testicular cancer. *N. Engl. J. Med.* 340, 1788–1795

4 Ono, R. *et al.* (2006) Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. *Nat. Genet.* 38, 101–106

5 Song, S.U. *et al.* (1994) An env-like protein encoded by a *Drosophila* retroelement: evidence that gypsy is an infectious retrovirus. *Genes Dev.* 8, 2046–2057

6 Zdobnov, E.M. *et al.* (2005) Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. *Nucleic Acids Res.* 33, 946–954

7 Bernard, D. *et al.* (2005) Identification and characterization of a novel retroviral-like aspartic protease specifically expressed in human epidermis. *J. Invest. Dermatol.* 125, 278–287

8 Edelstein, L.C. and Collins, T. (2005) The SCAN domain family of zinc finger transcription factors. *Gene* 359, 1–17

9 Ivanov, D. *et al.* (2005) Mammalian SCAN domain dimer is a domain-swapped homolog of the HIV capsid C-terminal domain. *Mol. Cell* 17, 137–143

10 Wills, N.M. *et al.* (2006) A functional −1 ribosomal frameshift signal in the human paraneoplastic MA3 gene. *J. Biol. Chem.* 281, 7082–7088

11 Rosenfeld, M.R. *et al.* (2001) Molecular and clinical diversity in paraneoplastic immunity to Ma proteins. *Ann. Neurol.* 50, 339–348

12 Tan, K.O. *et al.* (2001) MAP-1, a novel proapoptotic protein containing a BH3-like motif that associates with Bax through its Bcl-2 homology domains. *J. Biol. Chem.* 276, 2802–2807

13 Lyford, G.L. *et al.* (1995) Arc, a growth factor and activity-regulated gene, encodes a novel cytoskeleton-associated protein that is enriched in neuronal dendrites. *Neuron* 14, 433–445

14 Lilly, F. (1967) Susceptibility to two strains of Friend leukemia virus in mice. *Science* 155, 461–462

15 Best, S. *et al.* (1996) Positional cloning of the mouse retrovirus restriction gene Fv1. *Nature* 382, 826–829

16 Baksh, S. *et al.* (2005) The tumor suppressor RASSF1A and MAP-1 link death receptor signaling to Bax conformational change and cell death. *Mol. Cell* 18, 637–650

17 Okabe, H. *et al.* (2003) Involvement of PEG10 in human hepatocellular carcinogenesis through interaction with SIAH1. *Cancer Res.* 63, 3043–3048

18 Relaix, F. *et al.* (2000) Pw1/Peg3 is a potential cell death mediator and cooperates with Siah1a in p53-mediated apoptosis. *Proc. Natl. Acad. Sci. U. S. A.* 97, 2105–2110

19 Xu, L. *et al.* (2003) BTBD1 and BTBD2 colocalize to cytoplasmic bodies with the RBCC/tripartite motif protein, TRIM5δ. *Exp. Cell Res.* 288, 84–93

20 Cudmore, S. *et al.* (1997) Viral manipulations of the actin cytoskeleton. *Trends Microbiol.* 5, 142–148

21 Marques, A.C. *et al.* (2005) Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol.* 3, e357

22 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

23 Brandt, J. *et al.* (2005) Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. *Gene* 345, 101–111

0168-9525/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2006.09.006

Chromosomal clustering of nuclear genes encoding mitochondrial and chloroplast proteins in *Arabidopsis*

Andrey Alexeyenko[1], A. Harvey Millar[2], James Whelan[2] and Erik L.L. Sonnhammer[1]

[1] Center for Genomics and Bioinformatics, Karolinska Institutet, S-17177, Stockholm, Sweden
[2] ARC Centre of Excellence in Plant Energy Biology, M316, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia

We present a statistical analysis of chromosomal clustering among nuclear genes encoding mitochondrial or chloroplast proteins in *Arabidopsis*. For both organelles, the clustering was significantly increased above the expectation, but the clustering effect was weak, and most clusters were small and dispersed. Clustered genes showed coexpression but not more than expected, and no substantial synteny was detected in other eukaryotic genomes. We propose that the unexpected clustering results from continuous selection favoring chromosomal proximity of genes acting in the same organelle.

Introduction

Mitochondria and chloroplasts both originated from endosymbiotic events. During the course of evolutionary history, most of the essential genes required for mitochondrial and chloroplast function were transferred to the nuclear genome [1]. The organelle protein sets encoded in the nucleus today are not simply the genes originally transferred from the ancient endosymbiont but have a much more complicated genetic history. Evidence for proteins derived from the original endosymbiont, from the host genome, and even originating from other endosymbionts is found in both mitochondrial and chloroplast protein sets [2,3,4]. Due to the discrete metabolic roles of the two plant energy organelles, their origins, the mechanisms of gene transfer and the need for coordination of nuclear and organelle function in plants, it is possible that chromosomal organization of nuclear genes according to function could be an important aspect of regulation. Genes with various kinds of functional relationships are known to be in clusters on the chromosomes of a number of organisms [5–10]. Combining the experimental organelle sets [11–13] with clustering of their location and their expression in *Arabidopsis*, we have attempted to determine if any of the complex factors noted above link physical location with function or origin in this model plant.

Results

Chromosomal clustering of genes encoding targeted organelle proteins

Clusters of neighboring genes were built by finding stretches of organelle genes closer than 10 kb to another organelle gene along the five chromosomes of *Arabidopsis*. To resolve the problem of tandemly duplicated genes, clustered homologs (those with a unidirectional BLAST similarity score >100 bits) were counted as one gene. To calculate the statistical significance of the number of clustered genes in the experimental sets of 473 mitochondrial and 664 chloroplast genes, we picked the same number of genes randomly from the genome 5000 times to generate a probability distribution (Figure 1 and Table 1). The *P*-values for the observed clustered genes in the mitochondrial and chloroplast sets were 0.0034 and 0.0004, which are greatly significant.

No large clusters were found using a 10-kbp cutoff. Most chloroplast and mitochondrial clusters contained two genes. Five clusters had three organelle genes, and one contained four genes. Variable numbers of other genes were found in these clusters – the largest total cluster size was observed with three mitochondrial genes and six other genes. The importance of the cutoff distance was assessed by using a number of cutoffs between 5 and 80 kbp, as summarized in Figure 2. Approximately 50% of all *Arabidopsis* genes have a neighbor within 80 kbp, hence it was not meaningful to investigate higher cutoffs. For the chloroplast genes, the observed clustering was significant ($P < 0.05$) at all cutoffs except 80 kbp, whereas the mitochondrial genes are generally less significantly clustered, particularly at the 5-kbp cutoff.

The clustering analysis was also made with a gene-based distance cutoff, that is, counting the number of intermediate genes rather than the number of base pairs. The cutoffs 0, 1, 2, 4 and 9 in-between genes were used. In this case, we found the clustering tendency much weaker, and, again, less significant in the mitochondrial set (only one *P*-value was <0.05). The optimum was seen at two genes in between for both groups. This indicates that the genes clustered by absolute distance tend to reside

Corresponding author: Sonnhammer, E.L.L. (Erik.Sonnhammer@sbc.su.se)
Available online 18 September 2006.