15 Alvarez-Valin, F. *et al.* (1998) Synonymous and nonsynonymous substitutions in mammalian genes: intragenic correlations. *J. Mol. Evol.* 46, 37–44

16 Smith, N.G. *et al.* (1999) The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics* 153, 1395–1402

17 Hughes, A.L. *et al.* (1997) Comparative evolutionary rates of introns and exons in murine rodents. *J. Mol. Evol.* 45, 125–130

18 Smith, N.G. *et al.* (1998) Sensitivity of patterns of molecular evolution to alterations in methodology: a critique of Hughes and Yeager. *J. Mol. Evol.* 47, 493–500

19 Bielawski, J.P. *et al.* (2000) Rates of nucleotide substitution and mammalian nuclear gene evolution. Approximate and maximum-likelihood methods lead to different conclusions. *Genetics* 156, 1299–1308

20 Bernardi, G. *et al.* (1985) The mosaic genome of warm-blooded vertebrates. *Science* 228, 953–958

21 Bulmer, M. *et al.* (1991) Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. *Proc. Natl. Acad. Sci. U. S. A.* 88, 5974–5978

22 Mouchiroud, D. *et al.* (1997) Impact of changes in GC content on the silent molecular clock in murids. *Gene* 205, 317–322

23 Francino, M.P. *et al.* (1999) Isochores result from mutation not selection. *Nature* 400, 30–31

24 Marais, G. *et al.* (2003) Neutral effect of recombination on base composition in *Drosophila*. *Genet. Res.* 81, 79–87

25 Sved, J. *et al.* (1990) The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc. Natl. Acad. Sci. U. S. A.* 87, 4692–4696

26 Wolfe, K.H. *et al.* (1989) Mutation rates differ among regions of the mammalian genome. *Nature* 337, 283–285

27 Duret, L. *et al.* (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* 17, 68–74

28 Smith, N.G. *et al.* (2003) A low rate of simultaneous double-nucleotide mutations in primates. *Mol. Biol. Evol.* 20, 47–53

29 Charlesworth, B. *et al.* (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* 134, 1289–1303

30 Smith, J.M. *et al.* (1974) The hitch-hiking effect of a favourable gene. *Genet. Res.* 23, 23–35

31 Kondrashov, A.S. (1995) Modifiers of mutation-selection balance: general approach and the evolution of mutation rates. *Genet. Res.* 66, 53–69

32 Kimura, M. (1967) On the evolutionary adjustment of spontaneous mutation rates. *Genet. Res.* 9, 23–34

33 Sniegowski, P.D. *et al.* (2000) The evolution of mutation rates: separating causes from consequences. *BioEssays* 22, 1057–1066

0168-9525/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2005.05.005

Identification of a bacterial regulatory system for ribonucleotide reductases by phylogenetic profiling

Dmitry A. Rodionov[1,2] and Mikhail S. Gelfand[1,2]

[1]Institute for Information Transmission Problems, RAS, Bolshoi Karetny per 19, Moscow, 127994, Russia
[2]State Scientific Center GosNIIGenetika. 1st Dorozhny pr. 1, Moscow 117545, Russia

Using comparative genomics approaches, we analyzed the regulation of ribonucleotide reductase genes in bacterial genomes. A highly conserved palindromic signal with consensus acaCwAtATaTwGtg, named NrdR-box, was identified upstream of most operons encoding ribonuleotide reductases from three different classes. By correlating the occurrence of NrdR-boxes with phylogenetic distribution of ortholog families, we identified a transcriptional regulator containing Zn-ribbon and ATP-cone motifs (COG1327) for the predicted ribonucleotide reductase regulon. Further characterization of the regulon and metabolic reconstruction of the regulated pathways demonstrated its functional link to replication. The method of simultaneous phylogenetic profiling of genes and conserved regulatory signals introduced in this study could be used to identify transcriptional factors regulating orphan regulons.

Introduction

The rapidly increasing number of sequenced genomes provides challenges and opportunities for comparison of the whole proteomes, metabolic pathways and regulatory networks [1–3]. Functionally related genes tend to be clustered on the chromosome and to have similar patterns of occurrences in genomes [4,5]. The last assumption could be used to predict functional coupling for a pair of genes on the basis of their phylogenetic co-occurrence profiles. A modification of this approach, establishing a connection between genes and phenotypes, was used to detect potential genomic determinants of hyperthermophily [6]. In this article, we used phylogenetic profiling – correlation of genes and transcriptional regulatory elements – to identify a candidate regulator for the novel ribonucleotide reductase regulon NrdR.

Ribonucleotide reductases (RNRs) catalyze the reduction of all four ribonucleotides to the corresponding deoxyribonucleotides and are essential for the DNA synthesis [7]. There are three main types of RNRs: (i) aerobic enzymes present in prokaryotes and eukaryotes (distantly related classes Ia and Ib, represented by NrdAB and NrdEF proteins from *Escherichia coli*, respectively); (ii) bacterial and archaeal B_{12}-dependent enzymes homologous to NrdA and NrdE proteins (class II, NrdJ); and (iii) anaerobic enzymes (class III, NrdDG) [8]. In *E. coli*, the cell-cycle regulated *nrdAB* operon is activated by the

Corresponding author: Rodionov, D.A. (rodionov@iitp.ru).

DnaA, Fis and IciA transcription factors [9–10], and the anaerobically induced *nrdDG* operon is activated by Fnr [11]. Induction by hydroxyurea, an inhibitor of class I RNRs, was described for *nrd* operons in various species [12–15], suggesting upregulation of RNRs under conditions of deoxyribonucleotide starvation, although the molecular mechanism of this control was not known. Some indication of the involvement of a possible transcriptional regulator, *orfR*, was published by Torrents *et al.* [16]. Conserved consensus sequences were identified upstream of *nrd* operons in *Staphylococcus aureus* [12,17], *E. coli*, *Salmonella typhimurium* [18], and in *Streptomyces* spp. (GenBank accession nos AJ586904, AJ586905). However, the functional meaning of these sites is uncertain and no corresponding regulatory factors for the RNR genes are known. We have applied comparative genomics techniques (see the supplementary material online) to: (i) determine universal regulatory signals; (ii) identify transcription factors; (iii) describe the mode of regulation; and (iv) identify additional members of the RNR regulon.

Identification of NrdR-box

Analysis of upstream regions of *nrd* operons in various taxonomic groups enabled us to identify a highly conserved signal, named NrdR-box (for nrd Regulation), with minor taxon-specific deviations from the common consensus signal acaCwAtATaTwGtgt (Table 1). The construction of the recognition signal and our search for new regulon members are described in the supplementary material online. As result, we identified candidate NrdR-boxes upstream of all *nrd* genes in most genomes and upstream of only some *nrd* genes in a minority of genomes (for more details, see Tables S1 and S2 in the supplementary material online). In several genomes and taxonomic groups, additional members of the NrdR regulon that are involved in replication or deoxynucleotide salvage were identified (Table S1 in the supplementary material online). However, in some bacterial genomes and in all archaea and eukaryotes, no signal was observed. NrdR-boxes are highly conserved in upstream regions of RNRs from closely related species (Figures S1–S3 in supplementary material). Interestingly, NrdR-boxes occur in tandem in most cases (single NrdR-boxes were observed only in 27 of 243 operons), so that the distance between the centers of palindromes equals an integer number of DNA turns (21 bp, 31–32 bp or 41–42 bp). The presence of multiple regulatory sites at a specific distance ensures cooperative binding of NrdR molecules to DNA. All of the known promoters of NrdR-regulated genes overlapped with predicted NrdR-boxes, making it possible to predict that NrdR is a repressor (supplementary material online).

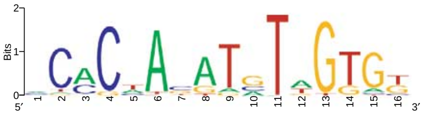
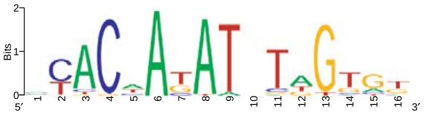
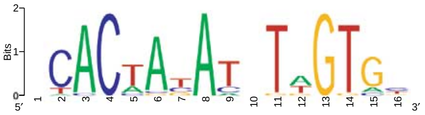
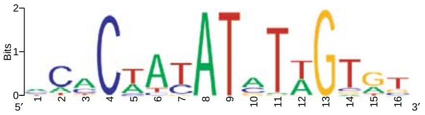
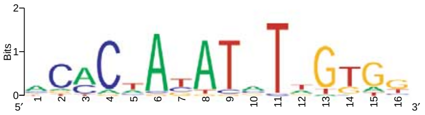
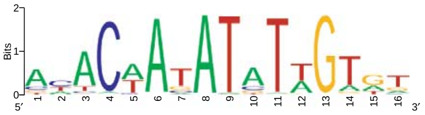
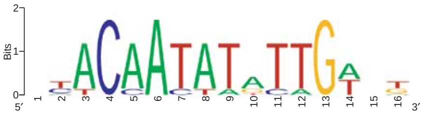
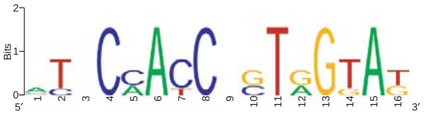
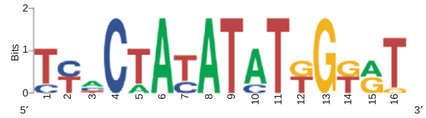
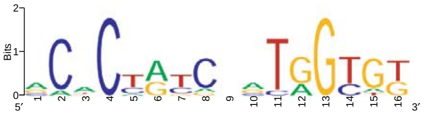
Conservation of the signal suggested the existence of a universal regulatory mechanism. The palindromic structure of the NrdR-box and its size are characteristic of many prokaryotic transcription factors. In an attempt to identify the transcription factor, we analyzed the clusters of orthologous groups of proteins (COGs) [19] using as a query a compiled phylogenetic profile with two categories of bacterial genomes, those with and those without the predited NrdR-boxes. We used an extended phylogenetic pattern search tool that enabled identification of COGs that contain or exclude selected organisms (http://web.dmz.uni-wh.de/projects/protein_chemistry/epps/) [20]. The exact pattern search among 63 genomes identified only one COG, COG1327, that was present in most bacteria but absent in archaea and eukaryotes. COG1327 was not found in the available genomes of the *Bacteroidetes*/*Chlorobi* group or in ε-proteobacteria, nor in Aquifex aeolicus, which constitutes a single-genome lineage. Among most other taxonomic groups, COG1327 showed a mosaic distribution (Table 1, supplementary material online).

COG1327, represented by hypothetical proteins *ybaD* from *E. coli* and *ytcG* from *Bacullis subtilis*, is annotated as ‘predicted transcriptional regulator consisting of Zn-ribbon and ATP-cone domains’ [19]. At most one member of this COG is present in any genome. Metal-binding Zn-ribbons consist of four conserved cysteines and participate in DNA or RNA binding in many different proteins including transcriptional factors [21]. The ATP-cone is an ATP-binding regulatory domain [22] that could be involved in sensing of deoxyribonucleotides to induce the DNA-binding activity of the Zn-ribbon domain of COG1327. We have tentatively re-named this protein NrdR.

We supplemented this analysis by considering the NrdR occurrence in taxonomic groups with mosaic distribution of NrdR-boxes (Table 1). Among α- and γ-proteobacteria, both NrdR-boxes and the *nrdR* genes were absent only in obligate intracellular parasites and endosymbionts *Rickettsia*, *Wolbachia*, *Buchnera* and *Wigglesworthia*. Both NrdR-boxes and *nrdR* were not found in two δ-proteobacteria (*Desulfovibrio* spp.), one cyanobacteria (*Nostoc* sp.) or two lactobacilli (*Oenococcus oeni* and *Leuconostoc mesenteroides*) in the *Bacillus*/*Clostridium* group. By contrast, among four sequenced spirochetes, only *Treponema denticola* has the *nrdR* gene and a candidate NrdR-box, which are both located upstream of the ribonucleotide reductase *nrdD*. For a control, we applied all of the constructed NrdR-box profiles with low thresholds to the genomes without *nrdR*, and found no candidate sites upstream of *nrd* genes. Finally, we analyzed upstream regions of the *nrd* genes in taxonomic groups without *nrdR*, and found no conserved signals that could serve as variant NrdR-boxes.

This functional assignment of COG1327 is corroborated by another comparative genomic technique: analysis of gene neighborhoods. It is well known that transcriptional factors often directly regulate adjacent genes on the chromosome [23]. Indeed, in many microbial genomes, the *nrdR* genes are clustered with ribonucleotide reductase genes or with those that are involved in the chromosome replication, for example, *dnaB, dnaI, polA* (Figure 1). In γ-proteobacteria, experimental and predicted binding sites for the main regulator of replication, DnaA, precede RNR genes. In some studies, *ybaD* (COG1327) was predicted to be the regulator of riboflavin biosynthesis [24,25] and named *ribX*. Indeed, in proteobacteria, *ybaD* is often clustered with the riboflavin biosynthesis *rib* genes, the glycine metabolism gene *glyA* and the transcription antitemination factor *nusB* (Figure 1). However, no candidate binding sites were found upstream of these operons and, therefore, the link

Table 1. The NrdR regulons in bacteria

Taxonomic group of bacteria	Sequence logo of NrdR-boxes	Distribution of NrdR[a]	Distribution of *nrdR* genes and NrdR-boxes in bacterial genomes	Other candidate regulon members
Actinobacteria		+	Present in all actinobacteria	None
α-proteobacteria		±	Present in all α-proteobacteria, except *Wolbachia* and *Rickettsia* spp.	None
β-proteobacteria		+	Present in all β-proteobacteria	None
γ-proteobacteria		±	Present in all γ-proteobacteria, except *Buchnera* and *Wigglesworthia spp.*	*topA* (DNA topoisomerase I) in *Pseudomonas* spp.; *dnaA* (replication initiator) in *Shewanella* spp.
δ-proteobacteria		±	Present in all δ-proteobacteria, except *Desulfovibrio* spp.	*dnaA* (replication initiator) in *Myxococcus xanthus, Desulfotalea psychrophila*; *COG1192* (chromosome partitioning) in *Desulfuromonas* spp.
Bacillus/ *Clostridium* group		±	Present in all *Bacillus*/ *Clostridium* group members, except *Leuconostoc mesenteroides* and *Oenococcus oeni*	*dgk-pnuC* (dNTP salvage) in lactobacilli; *nucA* (nucleotidase) in *Lactococcus lactis*; *yvdD-yvdC* (unknown function) in *Bacillus* spp.; *ligA* (DNA ligase) in *Clostridium acetobutylicum*
Thermotogales		+	Present in all thermotogales	None
Thermus/ *Deinococcus* group		+	Present in all *Thermus*/ *Deinococcus* group	*DR1775* (DNA helicase II) in *Deinococcus radiodurans*
Chlamydiales		+	Present in all chlamydiales	None
Cyanobacteria		±	Present in all cyanobacteria, except *Nostoc* sp.	None
ε-proteobacteria	n/a	−	Absent in ε-proteobacteria	None
Bacteroidetes/ *Chlorobi* group	n/a	−	Absent in all members of the *Bacteroidetes*/*Chlorobi* group	None
Mycoplasmatales	n/a	−	Absent in mycoplasmatales	None
Spirochaetes	n/a	±	Present only in *Treponema denticola*	None
Other	Diverse	±	Present in *Pirellula* sp., *Chloroflexus aurantiacus* and *Fusobacterium nucleaticum*, but absent in *Aquifex aeolicus*	None

[a] The presence, absence or mosaic distribution of NrdR boxes is indicated by +, − or ±, respectively.

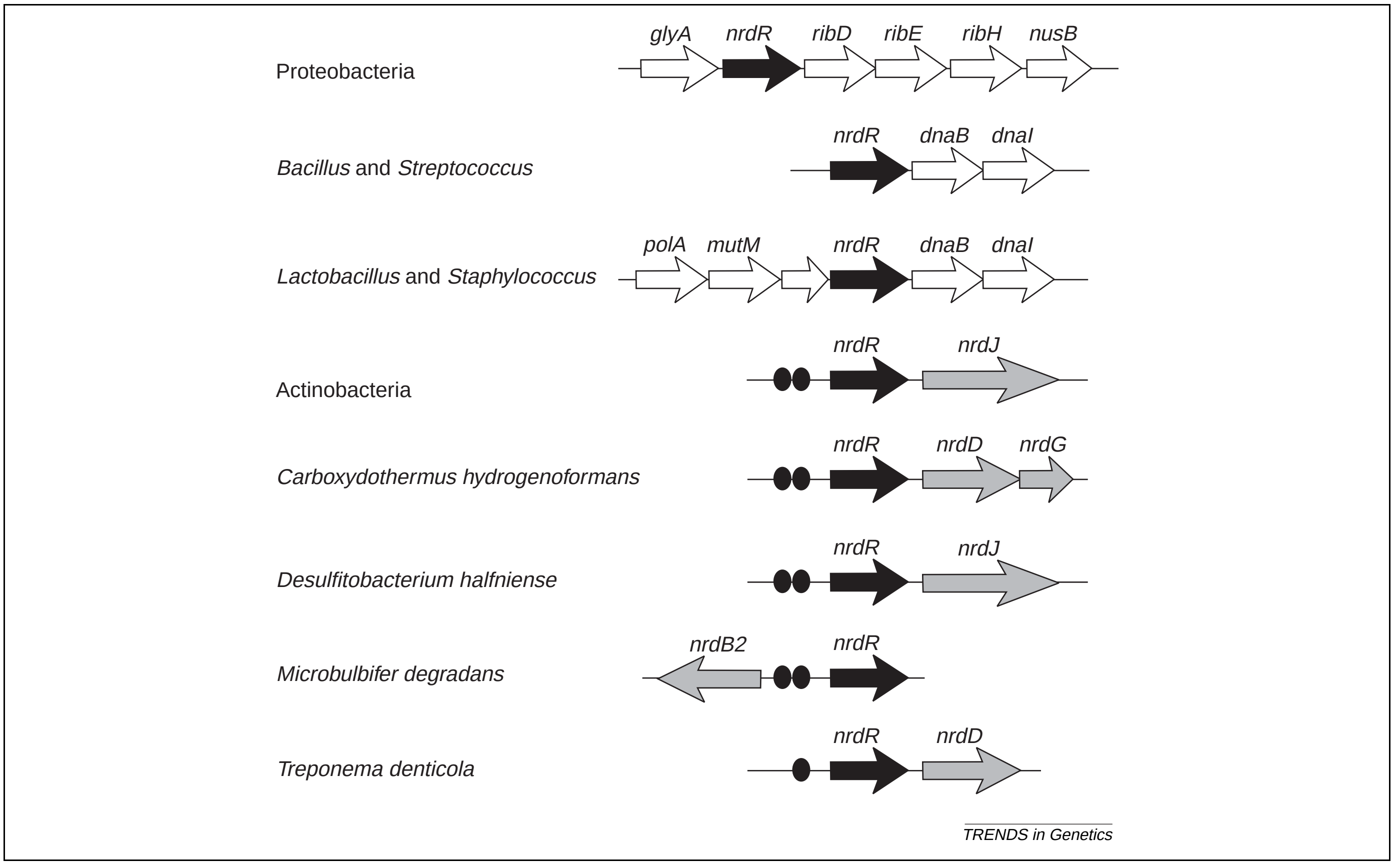


Figure 1. Genomic organization of the *nrdR*-containing loci in some bacterial genomes. Genes encoding the predicted ribonucleotide reductase regulator NrdR and the ribonucleotide reductase components are shown in black and grey, respectively. The black circles indicate the predicted NrdR-sites. The direction of transcription is indicated by the arrows.

between NrdR and riboflavin biosynthesis remains unexplained.

Concluding remarks

Thus, we have tentatively characterized the regulation of ribonucleotide reductases in bacteria using comparative genomic analysis. A combination of various techniques, such as phylogenetic profiling of genes and regulatory signals, phylogenetic footprinting of regulatory sites and positional gene clustering, enabled us to produce a detailed description of a regulatory system that is almost completely uncharacterized experimentally. We assigned the role of regulator of RNR genes in most bacterial genomes to NrdR (COG1327), and identified its universal DNA-binding signal, which occurs mostly in tandems suggesting co-operative binding, and we predicted that NrdR acts as a repressor by phylogenetic footprinting of NrdR-sites and of known promoters. We identified new members of the NrdR regulon involved in deoxynucleotide metabolism and replication, and thus characterized the functional role of NrdR in bacteria.

Our analysis shows that a combination of the diverse techniques used in comparative genomics analysis, such as phylogenetic profiling, positional clustering and phylogenetic footprinting, enables a detailed description of a system that is little studied experimently, whereas relying on any single type of evidence might be somewhat misleading. Indeed, while this study was being completed, Borovok and colleagues showed that NrdR was a transcriptional regulator of class Ia and class II RNR genes in *Streptomyces* [26]. Finally, the suggested modification of phylogenetic profiling based on the co-occurrence of regulatory motifs and genes seems to be useful for the analysis of unknown regulatory proteins, although it has obvious limitations, because it requires some conservation (or at least tractable evolution) of the signal and is not immediately applicable to the analysis of large regulator families where it is not possible to resolve orthology relationships.

Acknowledgements

This study was supported in part by grants from the Howard Hughes Medical Institute (55000309 to M.G.), the Russian Fund of Basic Research (04–04–49361 to D.R.), the Russian Science Support Fund (M.G.) and the Russian Academy of Sciences (Programs 'Molecular and Cellular Biology' and 'Origin and Evolution of the Biosphere').

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2005.05.011

References

1 Huynen, M. *et al*. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* 10, 1204–1210

2 Osterman, A. and Overbeek, R. (2003) Missing genes in metabolic pathways: a comparative genomics approach. *Curr. Opin. Chem. Biol.* 7, 238–251

3 Gelfand, M.S. (1999) Recognition of regulatory sites by genomic comparison. *Res. Microbiol.* 150, 755–771

4 Pellegrini, M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4285–4288

5 Glazko, G.V. and Mushegian, A.R. (2004) Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. *Genome Biol.* 5, R32

6 Makarova, K.S. *et al.* (2003) Potential genomic determinants of hyperthermophily. *Trends Genet.* 19, 172–176

7 Reichard, P. (1993) From RNA to DNA, why so many ribonucleotide reductases? *Science* 260, 1773–1777

8 Torrents, E. *et al.* (2002) Ribonucleotide reductases: divergent evolution of an ancient enzyme. *J. Mol. Evol.* 55, 138–152

9 Jacobson, B.A. and Fuchs, J.A. (1998) Multiple cis-acting sites positively regulate *Escherichia coli nrd* expression. *Mol. Microbiol.* 28, 1315–1322

10 Han, J.S. *et al.* (1998) Effect of IciA protein on the expression of the *nrd* gene encoding ribonucleoside diphosphate reductase in *E. coli*. *Mol. Gen. Genet.* 259, 610–614

11 Boston, T. and Atlung, T. (2003) FNR-mediated oxygen-responsive regulation of the *nrdDG* operon of *Escherichia coli*. *J. Bacteriol.* 185, 5310–5313

12 Masalha, M. *et al.* (2001) Analysis of transcription of the *Staphylococcus aureus* aerobic class Ib and anaerobic class III ribonucleotide reductase genes in response to oxygen. *J. Bacteriol.* 183, 7260–7272

13 Scotti, C. *et al.* (1996) The *Bacillus subtilis* genes for ribonucleotide reductase are similar to the genes for the second class I NrdE/NrdF enzymes of Enterobacteriaceae. *Microbiology* 142, 2995–3004

14 Smalley, D. *et al.* (2002) Aerobic-type ribonucleotide reductase in the anaerobe *Bacteroides fragilis*. *J. Bacteriol.* 184, 895–903

15 Borovok, I. *et al.* (2002) *Streptomyces* spp. contain class Ia and class II ribonucleotide reductases: expression analysis of the genes in vegetative growth. *Microbiology* 148, 391–404

16 Torrents, E. *et al.* (2003) *Corynebacterium ammoniagenes* class Ib ribonucleotide reductase: transcriptional regulation of an atypical genomic organization in the *nrd* cluster. *Microbiology* 149, 1011–1020

17 Alkema, W.B. *et al.* (2004) Regulog analysis: detection of conserved regulatory networks across bacteria: application to *Staphylococcus aureus*. *Genome Res.* 14, 1362–1373

18 Jordan, A. *et al.* (1995) Two different operons for the same function: comparison of the *Salmonella typhimurium nrdAB* and *nrdEF* genes. *Gene* 167, 75–79

19 Tatusov, R.L. *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29, 22–28

20 Reichard, K. and Kaufmann, M. (2003) EPPS: mining the COG database by an extended phylogenetic patterns search. *Bioinformatics* 19, 784–785

21 Aravind, L. and Koonin, E.V. (1999) DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res.* 27, 4658–4670

22 Aravind, L. *et al.* (2000) The ATP-cone: an evolutionarily mobile, ATP-binding regulatory domain. *J. Mol. Microbiol. Biotechnol.* 2, 191–194

23 Korbel, J.O. *et al.* (2004) Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat. Biotechnol.* 22, 911–917

24 Wolf, Y.I. *et al.* (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.* 11, 356–372

25 Doerks, T. *et al.* (2004) Global analysis of bacterial transcription factors to predict cellular target processes. *Trends Genet.* 20, 126–131

26 Borovok, I. *et al.* (2004) Alternative oxygen-dependent and oxygen-independent ribonucleotide reductases in *Streptomyces*: cross-regulation and physiological role in response to oxygen limitation. *Mol. Microbiol.* 54, 1022–1035

0168-9525/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2005.05.011

Letter

Assessing the signatures of selection in *PRNP* from polymorphism data: results support Kreitman and Di Rienzo's opinion

Marta Soldevila[1], Francesc Calafell[1], Agnar Helgason[2], Kári Stefánsson[2] and Jaume Bertranpetit[1]

[1]Unitat de Biologia Evolutiva, Facultat de Ciències de la Salut i de la Vida, Universitat Pompeu Fabra, Dr. Aiguader 80, 08003 Barcelona, Catalonia, Spain
[2]deCode Genetics, Sturlugata 8, IS-101 Reykjaviík, Iceland

Kreitman and Di Rienzo [1] highlighted an important issue in the analysis of polymorphism data and in the detection of the footprint of natural selection. In their article, they discussed the ascertainment bias that can be introduced in neutrality tests when genotyping large samples after an initial partial ascertainment of variation, by the sequencing of a limited number of chromosomes, implying that low-frequency variants are excluded from the analysis (see Ref. [2] and references therein). This could have a strong impact in the field of evolutionary genetics because numerous studies based on single nucleotide polymorphism (SNP) data are being produced with previously known SNPs, and detecting selection in the human genome is usually performed with neutrality tests based on the allele-frequency spectrum.

The balancing-selection hypothesis

Ascertainment bias was suggested in a study by Mead *et al.* [3], who proposed that balancing selection was the mechanism that maintained a polymorphism in the gene encoding the human prion protein (*PRNP*; GenBank accession no. M13899) in codon 129 (M129V). Mead *et al.* [3] argued that repeated episodes of endocannibalism in ancient human populations could have produced the

Corresponding author: Bertranpetit, J. (jaume.bertranpetit@upf.edu).

Available online 23 May 2005