UNIVERSITY OF CALIFORNIA, SAN DIEGO

A Diverse Set of Evolutionary Questions that have been Answered Using Completely Sequenced Genomes

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in

Biology

by

Fyodor Alexeevich Kondrashov

Committee in charge:

Professor Doris Bachtrog, Chair
Professor Peter Andolfatto
Professor John Huelsenbeck
Professor Pavel Pevzner
Professor Chris Wills

2008

The dissertation of Fyodor Alexeevich Kondrashov is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2008

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This thesis has been made possible by many people supporting my endeavors by scientific and administrative input. My advisor, Doris Bachtrog has deserves the most credit for enduring my eccentric flustering in the course of my time in San Diego. Peter Andolfatto went out of his way more than once to show his support in both science and life. Without the kind and timely action of John Huelsenbeck, Pavel Pevzner, Chris Wills and Lin Chao I would have stalled in my progression of acquiring academic regalia. The entirety of Andolfatto and Bachtrog lab deserves special mention for creating a working environment. Finally, the input of all of my collaborators has continuously improved my scientific inquiries. In particular, I would like to single out Eugene Koonin and Alexey Kondrashov who deserve the status of my first and most important mentors in science.

Special thanks go out to all of my coauthors on several papers that are a part of this dissertation.

Chapter 1, in full, is a reprint of the material as it appears in Popadin K, Mamirova L and Kondrashov FA (2007). A manually curated database of tetrapod mitochondrial tRNAs. BMC Bioinformatics **8**, 441. Biomed Central 2007. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material at it appears in Kondrashov FA. (2005) Prediction of pathogenic mutations in mitochondrially encoded

human tRNAs. *Human Molecular Genetics* **14,** 2415-2419. Oxford University Press 2005. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in Plotnikova OV, Kondrashov FA, Vlasov PK, Ginter EK and Rogaev EI (2007). Conversion and compensatory evolution of the γ-crystallin genes and identification of a cataractogenic mutation that reverses the sequence of the human CRYGD gene to an ancestral state. Am. J. Hum. Genet. **81**, 32-43. The American Society of Human Genetics 2007. The dissertation author was one of the two primary investigators and authors of this paper.

Chapter 4, in full, is a reprint of the material as it appears in Kondrashov FA, Ogurtsov AY, Kondrashov AS. (2006) Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. Journal of Theoretical Biology **240**, 616-626. Elsevier Ltd. 2007. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is a reprint of the material as it appears in Kondrashov FA, Koonin EV, Morgunov IG, Finogenova TV, Kondrashova MN. (2006) Evolution of glyoxylate cycle enzymes in Metazoa: evidence of multiple horizontal transfer events and pseudogene formation. Biol Direct **1**, 31. Biomed Central 2007. The dissertation author was the primary investigator and author of this paper.

Chapter 6, in full, is a manuscript of the material as it appears in Kondrashov FA , Gurbich TA and Vlasov PK (2007). Selection for functional uniformity of tuf duplicates in gamma-proteobacteria. Trends in Genetics **23**, 215-218. Elsevier Ltd. 2007. The dissertation author was the primary investigator and author of this paper.

Chapter 7, in full, is a manuscript intended for publication as Assis R, Kondrashov AS, Koonin, EV and Kondrashov FA (2008) Nested genes and increase of organizational complexity in metazoan genomes.The dissertation author was the primary investigator and author of this paper.

VITA

2000 Bachelor of Arts
Simon's Rock College of Bard

2004 Master of Science
University of California, Davis

2008 Doctor of Philosophy
University of California, San Diego

PUBLICATIONS

Kondrashov FA. (2005) Prediction of pathogenic mutations in mitochondrially encoded human tRNAs. *Human Molecular Genetics* **14,** 2415-2419.

Kondrashov FA, Ogurtsov AY, Kondrashov AS. (2006) Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. *Journal of Theoretical Biology* **240**, 616-626.

Kondrashov FA, Koonin EV, Morgunov IG, Finogenova TV, Kondrashova MN. (2006) Evolution of glyoxylate cycle enzymes in Metazoa: evidence of multiple horizontal transfer events and pseudogene formation. *Biol Direct* **1**, 31.

Kondrashov FA , Gurbich TA and Vlasov PK (2007). Selection for functional uniformity of tuf duplicates in gamma-proteobacteria. *Trends in Genetics* **23**, 215-218

Plotnikova OV, Kondrashov FA, Vlasov PK, Ginter EK and Rogaev EI (2007). Conversion and compensatory evolution of the γ-crystallin genes and identification of a cataractogenic mutation that reverses the sequence of the human CRYGD gene to an ancestral state. *Am. J. Hum. Genet.* **81**, 32-43

Popadin K, Mamirova L and Kondrashov FA (2007). A manually curated database of tetrapod mitochondrial tRNAs. *BMC Bioinformatics* **8**, 441

Assis R, Kondrashov AS, Koonin, EV and Kondrashov FA (2008) Nested genes and increase of organizational complexity in metazoan genomes. *Trends in Genetics* (submitted).

ABSTRACT OF THE DISSERTATION

A Diverse Set of Evolutionary Questions that have been Answered Using Completely Sequenced Genomes

by

Fyodor Alexeevich Kondrashov

Doctor of Philosophy in Biology

University of California, San Diego, 2008

Professor Doris Bachtrog, Chair

In the past decade or so, the availability of completely sequenced genomes and their annotations opened up previously unthinkable opportunities to explore evolutionary and functional aspects of organic life forms. The rate of deciphering new genomes shows no signs of slowing down. While a few years ago every self-respecting genomicist could name every single available genome, currently I will be hard pressed to name those that have been completed in the past year alone. Whether or not the race to sequence more, faster and cheaper will continue to revolutionize our understanding of biology is an open question largely irrelevant to this thesis. However, it is undeniably evident that in the past decade a new set of approaches, tools and knowhow have emerged in the fienld of computational genomics, which will continue to be used for many years to come.

It so happened that in the past three years I have been interested in several different questions that forced me to utilize almost every aspect of comparative computational genomics. In the course of answering questions that tickled my curiosity I have created a database, performed an evolutionary analysis of a newly sequenced genome, worked with secondary and crystal protein and RNA structures, measured the rate and mode of selection in coding and noncoding sequences, made functional annotations of proteins, and in the course of doing so have added to our understanding of several important evolutionary questions. Thus, this thesis is more of a demonstration of my capabilities as a computational comparative genomicist rather than a comprehensive attempt to resolve some long-standing dispute in biology. The first part of this thesis deals with some aspects and examples of compensatory evolution in a framework of Compensatory Pathogenic Deviations. The second part is a collection of works where the primary concept is the use of negative selection to reveal functional and evolutionary novel aspects of genes and genomes.

In Chapter 1, I describe a database of mitochondrial tRNA sequences and secondary structures from completely sequenced metazoan mitochondrial genomes. This database has been compiled mostly by hand, such that secondary structure predictions were matched with evolutionary conserved regions while eliminating annotation errors, resulting in an impressive 6060 curated tRNAs structure predictions. After its completion, but before publication,

this database has been used to describe patterns of compensatory evolution in mt tRNAs, which is the topic of Chapter 2. Previously to my work, it has been thought that it is impossible to create an exclusively computational method for prediction of pathogenic mutations in human mt tRNAs. I have been able to show the contrary using two simple improvements. Firstly, I have shown that sequences of more closely related species are much better predictors of fitness impacts in orthologous human sites than distant species. Indeed, this is an intuitive concept but it has not been utilized in a predictive matter previously. Secondly, I have used patterns of compensatory evolution in tRNA stem structures, which also greatly increased the predictive power of pathogenicity in orthologous sites. It is not enough to look at sequence conservation of a site to claim functional conservation, since sites may evolve in quickly even while being under functional and selective constraint. Such rapid evolution is most easily reconciled with functional conservation under the framework of structural compensatory evolution. For example, in a tRNA the nucleotides forming a Watson-Crick pair in a stem structure may rapidly change between G-C pair and an A-T pair. The destruction of each pair may be deleterious; however, each site may be rapidly evolving. By keeping track of sites potentially evolving in a compensatory manner, I have been able to further improve my prediction of pathogenic mutations in mt tRNAs.

In Chapter 3 I use the sequence of the γ-crystallins from several mammalian species to study compensatory evolution in this gene. A disease-

causing variant in one of the γ-crystallins was found in other, healthy mammals. Such events are called a Compensatory Pathogenic Deviation (CPD), and are thought to be caused by structural compensations in the homologous proteins. In this case, using a correlation sequence analysis it was possible to identify a probable compensatory site in the γ-crystallin. Curiously, on the crystal structure this site was in a 180 degrees symmetrical position to the site with the pathogenic substitution. This allowed us to conclude that crystallins are likely to be packed together such that individual proteins are assembled in strings with alternating 180 degree rotations. In addition, these genes showed interesting patterns of gene conversion. Two γ-crystallin pseudogenes showed clear signs of negative selection despite clearly being pseudogenes. This observation, coupled with signs of gene conversion in this gene family, led to the conclusion that gene conversion can lead to apparent selection in cases where the rate of conversion is rapid.

Chapter 4 is entirely devoted to the issue of selection on synonymous sites in human protein coding genes. Contrary to general belief, negative selection does not always lead to a decrease in the rate of evolution. If a preferred nucleotide is highly mutable, then the rate of evolution may be increased in comparison to a completely neutral site. This will occur due to a preferred to un-preferred nucleotide substitution achieving fixation through drift and driven by a high rate of mutation, while selection will drive the reverse process of un-preferred to preferred substitution. Mammalian synonymous sites

appear to be a mixture of sites with different rates of mutation spanning almost two orders of magnitude due to the highly mutable CpG context. The analysis reported in this chapter has shown that highly mutable synonymous sites evolve faster than intron sites with the same CpG context, while synonymous sites outside CpGs, those with a low rate of mutation, evolve slower than intron sites outside the CpG context. Assuming weak selection preferring GC nucleotides in synonymous sites leads to a perfect fit between several independent observations and theoretical predictions. This work remains the most comprehensive study of negative selection on synonymous sites that utilizes both empirical observations and theory.

Chapter 5 reports a genome-wide search for two glyoxylate cycle-specific enzymes, isocitrate lyase and malate synthase, in vertebrate genomes. The presence of glyoxylate cycle in metazoans has always been controversial, with all textbooks in biochemistry claiming that the glyoxylate cycle is not present in higher animals. I utilized sequence pattern searches in completely sequenced genomes, and found both glyoxylate cycle-specific enzymes in non-mammalian vertebrates. In addition, malate synthase appears to be still functional in non-placental mammals while being present as a pseudogene in placentals. Interestingly, both of these enzymes show a high rate of horizontal gene transfer throughout eukaryote evolution. In Chapter 6 a new method to study selection in duplicated genes is described. The method assumes that substitution in two gene copies with a high rate of paralogous gene conversion are under selection

simultaneously in both genes and, therefore, the strength of selection in such gene copies should resemble selection in a single copy gene. Thus, by comparing substitutions that have, and have not, been subject to gene conversion, we were able to evaluate selection on two diverging gene copies. Interestingly, selection against nonsynonymous substitutions was stronger in two independently evolving gene copies than in gene copies that were evolving in concert. This may be possible due to strong selection for maintaining functional uniformity of two gene copies.

Chapter 7 deals with organizational complexity of several metazoan genomes. The “beans on a string” model of gene arrangement has been abandoned since the discovery of a large fraction of nested genes. I was curious to analyze the evolution of such complex, nested gene arrangement. Through many genome comparisons it became clear that in recent evolutionary history complex, nested, gene arrangements have been much more commonly created than destroyed, implying a constant increase in genome organizational complexity. By looking at expression patterns of nested gene pairs no evidence has been found in support of selection playing a role in this independent increase of complexity. This study is the first study that looked at the evolution of complexity by analyzing evolutionary events directly, in this case gains of nested gene structures, rather than analyzing correlations of different genomic parameters. Equilibrium of genome organizational complexity appears to be very far away, on the order of hundreds of millions of years away.

Part I.

Compensatory Pathogenic Deviations

Chapter 1.

A database of metazoan mitochondrial tRNA genes

BMC Bioinformatics

BioMed Central

Database

Open Access

A manually curated database of tetrapod mitochondrially encoded tRNA sequences and secondary structures

Konstantin Yu Popadin[1], Leila A Mamirova[1] and Fyodor A Kondrashov*[2]

Address: [1]Institute for Information Transmission Problems RAS, Bolshoi Karetny pereulok 19, Moscow 127994, Russia and [2]Section on Ecology, Behavior and Evolution, Division of Biological Sciences, University of California at San Diego, 2218 Muir Biology Building, La Jolla, CA 92093, USA

Email: Konstantin Yu Popadin - konstantinpopadin@gmail.com; Leila A Mamirova - leilamamirova@gmail.com; Fyodor A Kondrashov* - fkondrashov@ucsd.edu

* Corresponding author

Published: 14 November 2007

BMC Bioinformatics 2007, **8**:441 doi:10.1186/1471-2105-8-441

Received: 14 March 2007
Accepted: 14 November 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/441

Abstract

Background: Mitochondrial tRNAs have been the subject of study for structural biologists interested in their secondary structure characteristics, evolutionary biologists have researched patterns of compensatory and structural evolution and medical studies have been directed towards understanding the basis of human disease. However, an up to date, manually curated database of mitochondrially encoded tRNAs from higher animals is currently not available.

Description: We obtained the complete mitochondrial sequence for 277 tetrapod species from GenBank and re-annotated all of the tRNAs based on a multiple alignment of each tRNA gene and secondary structure prediction made independently for each tRNA. The mitochondrial (mt) tRNA sequences and the secondary structure based multiple alignments are freely available as Supplemental Information online.

Conclusion: We compiled a manually curated database of mitochondrially encoded tRNAs from tetrapods with completely sequenced genomes. In the course of our work, we reannotated more than 10% of all tetrapod mt-tRNAs and subsequently predicted the secondary structures of 6060 mitochondrial tRNAs. This carefully constructed database can be utilized to enhance our knowledge in several different fields including the evolution of mt-tRNA secondary structure and prediction of pathogenic mt-tRNA mutations. In addition, researchers reporting novel mitochondrial genome sequences should check their tRNA gene annotations against our database to ensure a higher level of fidelity of their annotation.

Background

Mitochondrially encoded tRNAs (mt-tRNAs) are an excellent object of study for researchers in several fields for a variety of reasons. The primary reason is the wide variety of available completely sequenced mitochondrial genomes, which provides a large data sample from a broad phylogenetic background. Besides the obvious availability factor, mt-tRNAs show several unusual properties. mt-tRNAs are of particular interest to structural biologists, since the secondary structure of the mt-tRNAs is not as conserved as that of their nuclear encoded counterparts [1], and some mt-tRNAs in several lineages show accelerated rates of secondary structure evolution [2]. Although some changes of the secondary structure may be

the D-stem structure, which is a structural evolutionary change particularly common in mt-tRNAs [2].

Utility and Discussion

Most tetrapod mitochondrial genomes code for 22 different tRNAs with the exception of Metatherians that have lost the mt-tRNA^{Lys} [5]. In addition, some tetrapod mitochondrial genomes that were labeled as complete were only partially finished, such that seven mammalian genomes did not have sequences for tRNA^{Phe} (*Dromiciops gliroides*, *Metachirus nudicaudatus*, *Macrotis lagotis*, *Notoryctes typhlops*, *Perameles gunnii*, *Pseudocheirus peregrinus* and *Thylamys elegans*) and five mammals did not have the sequence for tRNA^{Pro} (*Arctocephalus forsteri*, *Dromiciops gliroides*, *Macrotis lagotis*, *Perameles gunnii* and *Thylamys elegans*). Thus, our database contains complete manually curated sequence and secondary structure information for 6060 mitochondrially encoded tRNA molecules.

Our database is available in 22 text files, one for each tRNA, with sequences of the 277 different species presented in the same order in each file. The order of the species in the alignment is the same for each mt-tRNA gene and roughly recapitulates the tetrapod phylogeny. Each file in the database includes the species common and scientific names, basic phylogenetic information and a multiple alignment of the tRNA with unaligned flanking sequence and annotated secondary structure (Figure 1 and 2). The "|" characters in the alignment delineate the conserved secondary structure prediction that was made using the alignment of all tRNA genes. The capital and lowercase letters in the files represent paired nucleotides according to the secondary structure prediction that was made with mfold. The two methods of secondary structure prediction generally showed similar results but small differences were common. For example, according to the mfold prediction many species in the tRNA^{Asn} gene form 3 WC pairs in the D-stem, while the classical tRNA structure supported by the alignment predicts 4 interacting nucleotides in this stem (Figure 1b). The value of showing separate predictions made by the alignment and the secondary structure is more evident in complicated cases, such as the case of the anticodon stem in the tRNA^{Asn} of the common iguana. In this case the alignment delineates the overall area where the anticodon stem should be formed, while mfold predicts which nucleotides form WC pairs in the structure (Figure 1b). Our database has a simple tab-delimited format with a set number of species in exactly the same order in each file making it especially useful for those researchers that wish to use our database in batch by parsing information on the secondary structure from our files.

The first database of mammalian mt-tRNAs which we used as a kernel in our alignment reports only mammalian species, it does not report any secondary structure that is independent of a multiple alignment and excludes complicated cases, such as the loss of D-stems [1]. Another, more current database that includes nuclear and mitochondrial tRNAs from the entire diversity of life forms has been, unfortunately, derived automatically [20] and is unlikely to be useful to researchers requiring a high level of sequence and structure annotation fidelity. In addition, both of these databases are difficult to use in batch mode, as they do not represent their results in a parsing-friendly format. Thus, our database is likely to be more useful for researchers that require a low level of annotation error, a phylogenetically diverse sample or prefer to work with many tRNA genes in simple text files. However, our database is not tailored to the needs of researchers that require a graphical interface for their work.

In the course of re-annotation and the compilation of a secondary-structure based multiple alignment, we have modified the annotation of the mt-tRNA gene location for 13% of all mt-tRNAs presented in our database. Such a high error rate in the annotation of such seemingly simple molecules as mt-tRNAs underscores the importance of availability of manually annotated databases such as the one reported here. In particular, we suggest for researchers reporting novel mitochondrial genome sequences to check their tRNA gene annotations against our database to ensure a higher level of fidelity of their annotation. Manually curated databases have an inherent advantage of a lower error rate than automatically created ones. However, a manual assembly of such an extensive database as the one reported here is a resource-intensive enterprise, and it is unlikely that the current database will be considerably expanded using the same manual approach. Rather the aim for the further development of this resource is to use the alignments reported here as a basis for further automatic enlargement.

Conclusion

We report a secondary structure based multiple alignment of 6060 mt-tRNAs from 277 tetrapod species. In the course of our work, we have re-annotated a large fraction of mt-tRNA genes, and manually checked all secondary structure predictions. We expect that our database will facilitate further research of mitochondrially encoded tRNAs from a structural, evolutionary and medical perspectives. Currently, mammalian mitochondrial tRNAs are thought to have a high level of similarity to the canonical tRNA secondary structure [1]. However, an analysis of exceptions to the canonical tRNA structures among the vertebrate mt-tRNAs, which is made possible with the database reported here, has not been undertaken. The evolutionary implications of compensations on a molecular level have been investigated previously [4], however, the study of CPDs in mt-tRNAs has been performed only

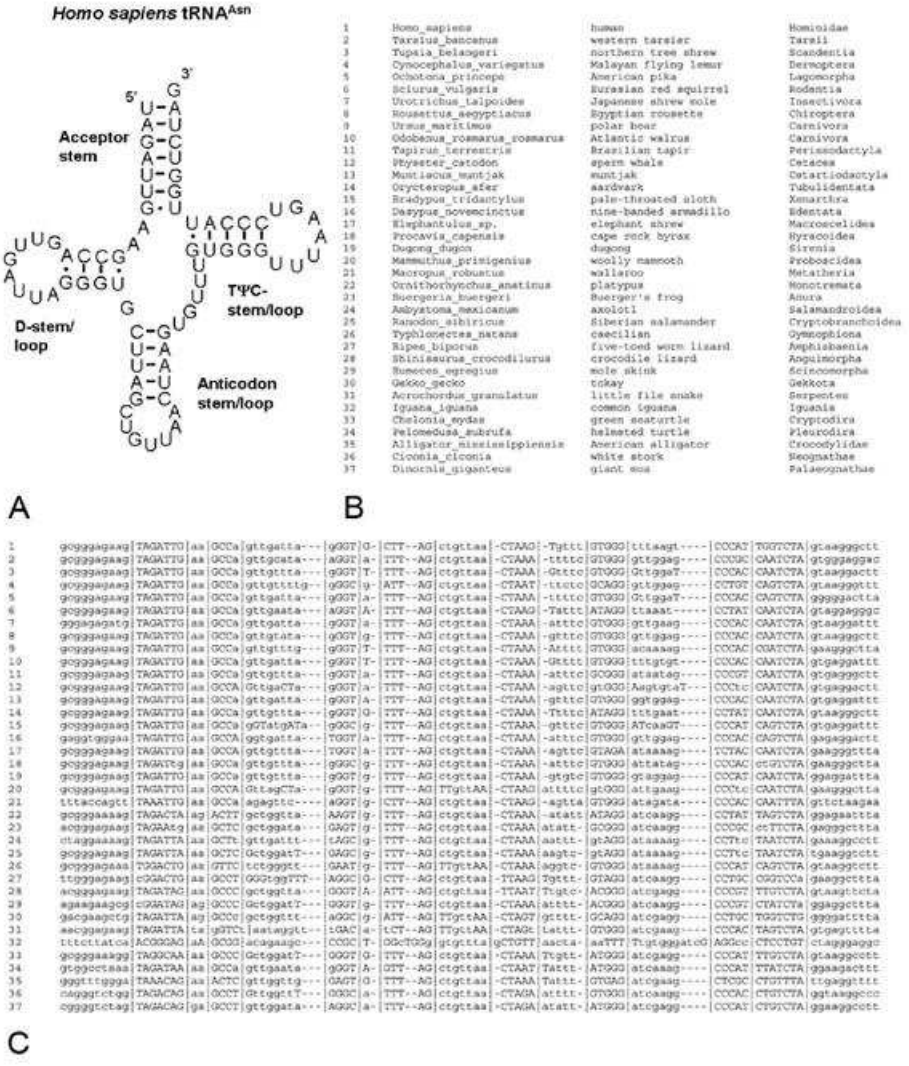


Figure 1
The secondary structure of the human mt-tRNA^{Asn} (a) and the multiple alignment with annotated secondary structure for selected species of mt-tRNA^{Asn} (b, c). The "|" characters separate the loops and stems based on the accepted basic secondary structure of mt-tRNAs form Helm *et al.* (2000) while capital letters denote those nucleotides that are predicted by mfold to participate in WC or GU pairing in stem structures.

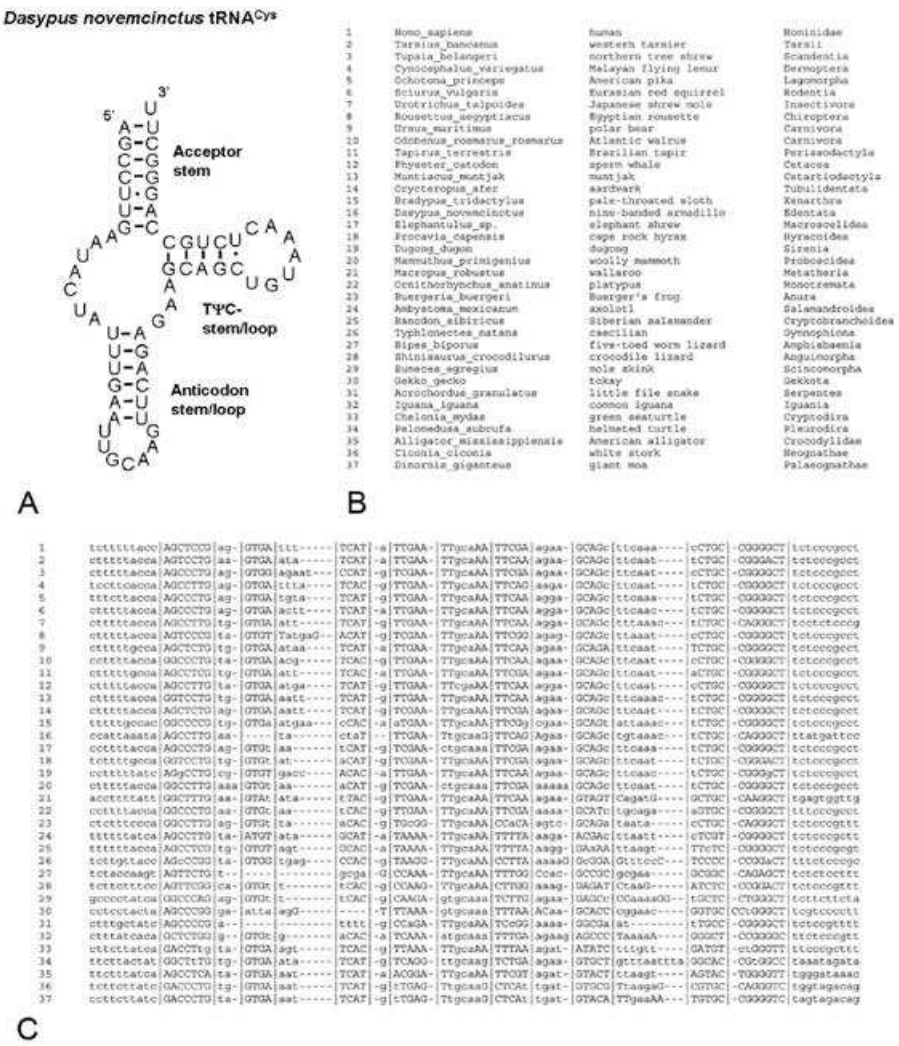


Figure 2
The secondary structure of the the nine-banded armadillo mt-tRNA^{Cys} (a), and the multiple alignment with annotated secondary structure for selected species of mt-tRNA^{Cys} (b, c). The "|" characters separate the loops and stems based on the accepted basic secondary structure of mt-tRNAs form Helm *et al.* (2000) while capital letters denote those nucleotides that are predicted by mfold to participate in WC or GU pairing in stem structures. The secondary structure of mt-tRNA^{SerAGY} in our database resembles the one of the nine-banded armadillo mt-tRNA^{Cys} (c).

on mammalian mt-tRNAs. Finally, prediction of pathogenic mutations in mt-tRNAs relies heavily on evolutionary conservation [13,14] and the availability of a secondary structure-based alignment of an expanded set of species may contribute to a more accurate prediction of the phenotypic consequences of mt-tRNA mutations.

Availability and requirements

Project name: mt tRNA tetrapod database;

Project home page: http://www.umich.edu/~kondrash/Database/;

Operating system(s): Platform independent;

Programming language: none

License: no restriction;

Any restrictions to use by non-academics: no restriction.

Authors' contributions

KYuP, LAM and FAK conceived the construction of the database, and participated in the construction of the initial and final alignments, corrected erroneously annotated tRNAs and were involved in secondary structure prediction. FAK drafted the paper, and all authors read and approved the final manuscript.

Acknowledgements

KYuP and LAM were supported by the Molecular and Cellular Biology Program of the Russian Academy of Science. KYuP was supported by the Russian Fund of Basic Research (grant 04-04-49623). LAM was partially supported by grants from the Howard Hughes Medical Institute (55005610), INTAS (05-1000008-8028). FAK is a National Science Foundation Graduate Research Fellow.

References

1. Helm M, Brule H, Friede D, Giege R, Putz D, Florentz C: **Search for characteristic structural features of mammalian mitochondrial tRNAs.** *RNA* 2000, **6**:1356-1379.
2. Macey JR, Larson A, Ananjeva NB, Papenfuss TJ: **Replication slippage may cause parallel evolution in the secondary structures of mitochondrial transfer RNAs.** *Mol Biol Evol* 1997, **14**:30-39.
3. Kondrashov FA: **Convergent evolution of secondary structure of mitochondrial cysteine tRNA in the nine-banded armadillo *Dasypus novemcinctus*.** *Biofizika* 2005, **50**:396-403.
4. Kern AD, Kondrashov FA: **Mechanisms and convergence of compensatory evolution in mammalian mitochondrial tRNAs.** *Nat Genet* 2004, **36**:1207-1212.
5. Janke A, Feldmaier-Fuchs G, Thomas WK, von Haeseler A, Paabo S: **The marsupial mitochondrial genome and the evolution of placental mammals.** *Genetics* 1994, **137**:243-256.
6. Borner GV, Yokobori S, Morl M, Dorner M, Paabo S: **RNA editing in metazoan mitochondria: staying fit without sex.** *FEBS Lett* 1997, **409**:320-324.
7. Helm M, Brule H, Degoul F, Cepanec C, Leroux JP, Giege R, Florentz C: **The presence of modified nucleotides is required for cloverleaf folding of a human mitochondrial tRNA.** *Nucleic Acids Res* 1998, **26**:1636-1643.
8. Brandon MC, Lott MT, Nguyen KC, Spolim S, Navathe SB, Baldi P, Wallace DC: **MITOMAP: a human mitochondrial genome database – 2004 update.** *Nucleic Acids Res* 2005, **33**:D611-D613.
9. Wittenhagen LM, Kelley SO: **Impact of disease-related mitochondrial mutations on tRNA structure and function.** *Trends Biochem Sci* 2003, **28**:605-611.
10. Mahata B, Mukherjee S, Mishra S, Bandyopadhyay A, Adhya S: **Functional delivery of a cytosolic tRNA into mutant mitochondria of human cells.** *Science* 2006, **314**:471-474.
11. Moreno-Loshuertos R, Acin-Perez R, Fernandez-Silva P, Movilla N, Perez-Martos A, Rodriguez de Cordoba S, Gallardo ME, Enriquez JA: **Differences in reactive oxygen species production explain the phenotypes associated with common mouse mitochondrial DNA variants.** *Nat Genet* 2006, **38**:1261-1268.
12. Florentz C, Sissler M: **Disease-related versus polymorphic mutations in human mitochondrial tRNAs. Where is the difference?** *EMBO Rep* 2001, **2**:481-486.
13. McFarland R, Elson JL, Taylor RW, Howell N, Turnbull DM: **Assigning pathogenicity to mitochondrial tRNA mutations: when "definitely maybe" is not good enough.** *Trends Genet* 2004, **20**:591-596.
14. Kondrashov FA: **Prediction of pathogenic mutations in mitochondrially encoded human tRNAs.** *Hum Mol Genet* 2005, **14**:2415-2419.
15. Smith PM, Ross GF, Taylor RW, Turnbull DM, Lightowlers RN: **Strategies for treating disorders of the mitochondrial genome.** *Biochim Biophys Acta* 2004, **1659**:232-239.
16. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2006, **34**:D16-D20.
17. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2007, **35**:D5-D12.
18. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**:1792-1797.
19. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction.** *Nucleic Acids Res* 2003, **31**:3406-15.
20. Sprinzl M, Vassilenko KS: **Compilation of tRNA sequences and sequences of tRNA genes.** *Nucleic Acids Res* 2005, **33**:D139-D140.

Chapter 1, in full, is a reprint of the material as it appears in Popadin K, Mamirova L and Kondrashov FA (2007). A manually curated database of tetrapod mitochondrial tRNAs. BMC Bioinformatics **8**, 441. Biomed Central 2007. The dissertation author was the primary investigator and author of this paper.

Chapter 2.

Prediction of pathogenic mutations in mitochondrially encoded human tRNAs

Human Molecular Genetics, 2005, Vol. 14, No. 16
doi:10.1093/hmg/ddi243
Advance Access published on July 13, 2005

Prediction of pathogenic mutations in mitochondrially encoded human tRNAs

Fyodor A. Kondrashov[1,2,*]

[1]Section on Ecology, Behavior and Evolution, Division of Biological Sciences, University of California at San Diego, 2218 Muir Biology Building, La Jolla, CA 92093, USA and [2]Engelhardt Institute of Molecular Biology, Vavilova 32, 119991 Moscow, Russia

Received May 30, 2005; Revised June 24, 2005; Accepted July 1, 2005

Some mutations in human mitochondrial tRNAs are severely pathogenic. The available computational methods have a poor record of predicting the impact of a tRNA mutation on the phenotype and fitness. Here patterns of evolution at tRNA sites that harbor pathogenic mutations and at sites that harbor phenotypically cryptic polymorphisms were compared. Mutations that are pathogenic to humans occupy more conservative sites, are only rarely fixed in closely related species, and, when located in stem structures, often disrupt Watson–Crick pairing and display signs of compensatory evolution. These observations make it possible to classify ~90% of all known pathogenic mutations as deleterious together with only ~30% of polymorphisms. These polymorphisms segregate at frequencies that are more than two times lower than frequencies of polymorphisms classified as benign, indicating that at least ~30% of known polymorphisms in mitochondrial tRNAs affect fitness negatively.

INTRODUCTION

A variety of often severe genetic disorders, mostly neuromuscular and neurodegenerative, are caused by mutations in mitochondrial (mt) DNA (1–4). These mutations may be located in mitochondrially encoded proteins, rRNAs, tRNAs and even in regulatory regions (5). However, mutations of the 22 mt tRNAs are of particular interest because these tRNAs span only 10% of the human mitochondrial genome yet they harbor more than half of all known mitochondrial pathogenic mutations (5) and, as tRNAs have a specific, cloverleaf secondary structure, such mutations may be studied from the perspective of the secondary structure of the tRNAs (6).

It is thought that the accurate identification of pathogenic mutations would enable researchers to describe their molecular and biochemical characteristics (7,8), which may lead to more successful treatment of the resulting pathologies (8,9). Currently, a mutation can be identified as pathogenic by a variety of different criteria (10), one of which is the preference for the mutation to be found at an evolutionarily conserved site, which by itself is a poor predictor of pathogenic nature of a variant (11–13). The other criteria, such as the requirement for the pathogenic mitochondria to be heteroplasmic, are not computational and often require extensive work in the laboratory (10). Thus, the availability of an accurate computational approach should aid rapid and inexpensive identification of novel pathogenic mutations.

Previous comparisons of pathogenic mutations and phenotypically cryptic polymorphisms in human mt tRNAs showed that pathogenic mutations are more often located at conservative sites (11–14) and stem structures (14) and tend to disrupt Watson–Crick (WC) nucleotide pairing in stems (11,14). However, these observations turned out to be insufficient to predict the impact of a nucleotide substitution in an mt tRNA on the phenotype and fitness (11,14). Here, an evolutionarily based computational analysis of the differences between pathogenic and non-pathogenic substitutions is described.

RESULTS

Pathogenic mutations and cryptic polymorphisms were mapped on multiple alignments of the 22 mt tRNAs from 138 different mammals. Both nucleotides which act as mutations and as polymorphisms in humans may be found at orthologous sites of closely related species (15). However, compensatory substitutions, usually restoring WC pairing (not including GU pairs), often accompany fixations of pathogenic mutations (11), but not of cryptic polymorphisms, in a non-human mammal. For example, a human pathogenic

*To whom correspondence should be addressed. Tel: +1 7916 2566774; Fax: +1 858 534718; Email: kondrashov@ucdavis.edu

mutation in $tRNA^{Gly}$ that is a part of the normal sequence of its non-human orthologs appears to be subject to WC compensation (Fig. 1A). In contrast, the WC correspondence between sites that harbor human polymorphisms and the interacting sites is less than perfect (Fig. 1B). The tendency of mt tRNA stem sites harboring pathogenic mutations to co-evolve with their complementary stems sites was used to distinguish pathogenic mutations from polymorphisms. On the basis of the criteria of conservation and compensatory co-evolution, a human variant can be classified as either benign or deleterious.

There are 94 known variants (47 pathogenic mutations and 47 polymorphisms) in mt tRNA stems that disrupt WC pairing. For each variant, the number of species was counted in which a variant was found without the compensatory substitution, and the number of species in which a potentially WC pairing-restoring compensatory substitution was found without the variant. For example, in Figure 1A, both of these numbers are equal to 0 (variant A is always found opposite a T) and in Figure 1B, these values are 1 and 2 for variant G, and 4 and 0 for variant A. The sum of these two numbers, obtained separately for primates only and for all mammals, was used as a gauge of evolutionary independence of the interacting sites. A variant was classified as benign if this sum was greater than two for primates (four pathogenic mutations; 15 polymorphisms) or if this sum was greater than nine for all mammals, and the site was conserved in less than 100 of the 138 (80 out of 119 for $tRNA^{Lys}$ variants, see Materials and Methods) available mammals (one pathogenic mutation; five polymorphisms). All other variants were classified as deleterious (42 pathogenic mutations; 27 polymorphisms).

If a variant within an mt tRNA stem that does not disrupt a WC pair (two pathogenic mutation and 38 polymorphisms total) affects a site that is conserved in more than 128 of the 138 available mammals (108 out of 119 for $tRNA^{Lys}$ variants, see Materials and Methods), and the variable nucleotide was found in fewer than five mammals, it was classified as deleterious (two pathogenic mutation; four polymorphisms) and other variants (34 polymorphisms) were classified as benign. Together, these criteria place 94% pathogenic mutations localized in the stems, and 36% of such cryptic polymorphisms, into the deleterious category.

Without tertiary structure information, the only available data on tRNA loops are sequence conservation. Previous attempts to distinguish pathogenic mutations from polymorphisms relied on sequence conservation in all the available mammalian orthologs (11). However, this dataset is too diverse, because mt tRNA loops are often not conserved even within primates (Fig. 1), and the use of non-primate mammals may obscure the pattern of conservation. In general, the use of closely related species is preferred, because of the possibility of functional or compensatory changes in more distant orthologs.

Two criteria were applied to 18 pathogenic mutations and 98 polymorphisms located outside of stems: (i) whether a variant is found in one of the five non-human Greater Ape species and (ii) the level of conservation in all of the 17 primate species of the site of a variant. A variant was classified as benign if it was present in at least one Greater Ape species (one pathogenic mutation; 42 polymorphisms) or if its site was not conservative, in the sense that its most common nucleotide was found in less than 15 out of 17 primate species (one pathogenic mutation; 35 polymorphisms). Otherwise, a variant was classified as deleterious (16 pathogenic mutations; 21 polymorphisms). This simple approach classifies as deleterious 89% of pathogenic mutations and only 21% of polymorphisms located outside of stem structures.

Among all the known variants within all human mt tRNAs, ~70% of polymorphisms were classified as benign, whereas the rate of false negative predictions was low: only ~10% of all known pathogenic mutations were classified as benign (Fig. 2). The classification of ~90% of all pathogenic mutation as deleterious is a substantial improvement over previous results (11,14). Still, classification of ~30% of phenotypically cryptic polymorphisms as deleterious may appear as a deficiency of the proposed analysis, should it represent the rate of false positive prediction. Previous analyses assumed that all segregating polymorphisms are selectively neutral (11,14). However, owing to a very high mutation rate of the mitochondrial DNA (16), many segregating variants may be deleterious (17). In addition, some polymorphisms may be sequencing errors or heteroplasmic variants.

Indeed, the average frequency of polymorphisms classified as deleterious (0.0014) is more than two times lower than that of the polymorphisms classified as benign (0.0032), which is highly significant ($n_1 = 131$, $n_2 = 52$, $U = 4401.0$, $P < 0.0018$, two-tailed Mann Whitney U-test). Most of the polymorphisms that were classified as deleterious are singletons, i.e. they are present in only one of the 2064 individuals from which the data on segregating variants have been obtained (18) and few were found in more than two individuals (Fig. 3). The 52 different polymorphisms that were classified as deleterious were found 340 times in the sample of 2064 individuals from the human population, although all the 183 polymorphisms were found 1415 times. Thus, ~30% of all polymorphisms in human tRNAs classified as deleterious (52/183), indeed, reduce fitness, despite the lack of obvious phenotypic manifestation. The probability that a polymorphism drawn at random from the human population is deleterious is ~25% (340/1415). The high estimate is not surprising because some of the polymorphisms found in the human population used in this study are known to contribute to the progression of mitochondrial disease (17). However, these estimates were made using a highly non-random sample of the human population (18) and more accurate measurements are needed to make a more reliable qualitative estimate.

Taken together, these observations suggest that the described method has a high rate of accuracy for distinguishing benign variants from severely and slightly pathogenic ones. To aid the identification of new pathogenic variants, this method was applied to all possible mutations of the 22 mt tRNAs (Supplementary Material). As expected, the mutations disrupting WC pairing in stems were predicted to have the highest probability of being deleterious (2218 deleterious mutations out of 2346 mutations total), whereas mutations in stems that to do not disrupt WC pairs have the lowest (104 out of 354). Mutations that are not located in stems have an intermediate probability of being deleterious (1095 out of 1767), most likely due to the inclusion of the highly conserved anticodon loop. Obviously, most lethal mutations should also be classified as deleterious by this approach.

Figure 1. Multiple alignment of (**A**) mt $tRNA^{Gly}$ and (**B**) mt $tRNA^{Thr}$. Pathogenic mutations are labeled in red and their compensatory substitutions in blue. Polymorphic states are in green and their compensatory states in yellow. Sequences are, top to bottom, from *Homo sapiens, Pan troglodytes, Pan paniscus, Gorilla gorilla, Pongo pygmaeus, Pongo pygmaeus abelii, Papio hamadryas, Macaca sylvanus, Macaca mulatto, Colobus guereza, Trachypithecus obscurus, Hylobates lar, Cebus albifrons, Lemur catta, Nycticebus coucang, Tarsius bancanus, Tupaia belangeri.*

DISCUSSION

Polymorphisms classified here as deleterious are not necessarily overtly pathogenic, because even a slight decrease in fitness is enough to reduce the frequency of a variant (19). However, the identification of a relatively high proportion of slightly deleterious polymorphisms is not necessarily a disadvantage of the proposed method even from a medical perspective. It is likely that many of the identified deleterious polymorphisms either contribute epistatically to the progression of many mitochondrial disorders (17) or slightly reduce life expectancy (20,21). Investigation of the impact of such polymorphisms on fitness may lead to more accurate description of the impact of these polymorphisms on morbidity and mortality.

Similar attempts to discriminate pathogenic mutations from polymorphisms were made for amino acid replacements in nuclear encoded proteins (22,23). However, the described analysis of pathogenic mutations in mt tRNAs is slightly more successful than that in proteins, possibly because of the difficulty of predicting patterns of compensatory evolution in proteins (24), relative to tRNAs (15). Many polymorphisms

Human Molecular Genetics, 2005, Vol. 14, No. 16

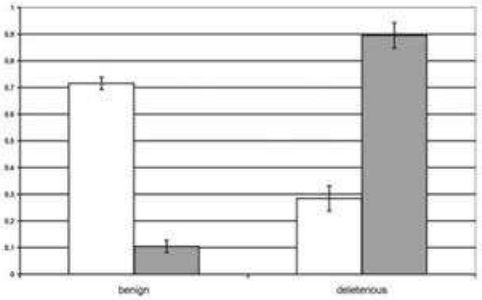


Figure 2. Frequency of 184 polymorphisms (white) and 68 pathogenic mutations (grey) classified as benign (131 polymorphisms; seven pathogenic mutations) and deleterious (52 polymorphisms; 60 pathogenic mutations). Error bars represent the standard deviation.

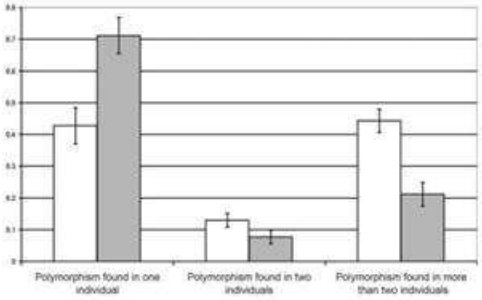


Figure 3. Frequency distribution of polymorphisms classified as benign (white) and deleterious (grey). Error bars represent the standard deviation.

found in nuclear encoded proteins are deleterious (22,23); future application of the method described here for the identification of deleterious variants in the two rRNA genes encoded in the mitochondria and the application of protein-based methods (25) for the identification of deleterious variants in 13 mitochondrially encoded proteins should lead to an estimate of the genetic load (26) of the mitochondrial genome.

There are four reasons why the method described here appears to have a higher accuracy than those that were published previously (11,14). First, sites in mt tRNAs were segregated into three, rather than two, structural categories. Secondly, this method allows for the possibility of having a variable site that excludes a particular nucleotide, i.e. it allows for the possibility that A, T and G are neutral equivalents and C is deleterious. Thirdly, use patterns of compensatory evolution were used in stem sites that disrupt WC pairs. Finally, sequence conservation information is used only in closely related species. Further improvement of this method may be based fine-tuning the cutoff values that were used, or on the addition of other ways to distinguish pathogenic and polymorphic variants. For example, the use of tertiary structure information improves the prediction of pathogenic mutations in proteins (23) and should help to improve the method of pathogenic mutation prediction for mt tRNAs as well. However, given that sequence similarity of closely related species is a good predictor of pathogenicity, the present method can be improved by the availability of mitochondrial sequences of other primate species, especially from *Platyrrhini* and *Strepsirhini*, of which only three species are currently available.

MATERIALS AND METHODS

A list of pathogenic mutations was taken from MitoMap (5) (http://www.mitomap.org/) and a list of polymorphisms from mtDB (18) (http://www.genpat.uu.se/mtDB/). Those pathogenic mutations whose status was listed as 'unclear' in MitoMap were excluded from the analysis. Data from the recent re-evaluation of the pathogenicity of mutations reported in MitoMap (14) were used with a slight modification: data on sequence conservation was excluded from the score of pathogenicity [see Supplemental Material from McFarland *et al.* (14)]. Those mutations that were listed in MitoMap and scored less than six on the scale of pathogenicity (minus the score from the conservation column from McFarland *et al.* (14)] were removed from the list of pathogenic mutations. In addition, the mutation 606(A → G) was excluded from the list of pathogenic mutations because of the raised doubts of its pathogenic nature (27). Finally, variants that were listed as pathogenic mutations and as polymorphisms were not included in either category. The final dataset included 67 pathogenic mutations and 183 polymorphisms. The cutoff values that were used to classify variants as pathogenic or benign were chosen according to the expected fraction of non-WC compensations among all observed compensatory events in primates and mammals, which were published previously (15). The fraction of non-WC compensations was ~5% for pathogenic compensations in WC pairs (11), and as there are two interacting nucleotides in a WC pair, it implies that 10% of the nucleotides involved in WC pairing may be subject to non-WC compensations. Thus, the cutoff values for WC compensations in WC pairs were two out of 17 (~12%) in primate species and 15 out of 138 (~11%) in all mammals. Sequence conservation cutoff values were taken as ~5% for non-WC pairs, as was done previously (11), and a more relaxed threshold of ~20% was chosen for sites located outside of stem structures because of a substantially higher fraction of compensatory evolution in such sites (15). Complete mitochondrial sequences from 138 different mammalian species were obtained from GenBank by using 'mammal AND complete AND genome AND mitochondria' as a keyword in the Entrez retrieval system (28). Only one mitochondrial sequence was used from each mammal, and the possibility that a polymorphism in humans randomly occurs at the same site as a polymorphism in another species was ignored. Alignments of tRNA genes were made with CLUSTALW (29) and manually corrected using secondary structure information published previously (6). Annotations of tRNA genes were manually corrected for several

species using data on sequence similarity from closely related species. Different cutoff values were used for variants located in mt $tRNA^{Lys}$ when estimating the level of conservation in stem sites because this gene is missing in marsupial species (30). In calculating average frequencies of benign and deleterious polymorphisms, but not in the application of the Mann Whitney *U*-test, three outliers (polymorphisms with a frequency > 0.05) were removed.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

The author thanks P. Andolfatto, D. Bachtrog, N. Esipova, S. Makeev, A. Kondrashov, V. Ramensky, V. Tumanyan and P. Vlasov for a critical reading of the manuscript. The author is an NSF Graduate Research Fellow. This work was supported by a Contract of the Russian Ministry of Science and Education (02.434.11.1008) and a grant on Molecular and Cellular Biology from RAS.

Conflict of Interest statement. None declared.

REFERENCES

1. De Vivo, D.C. (1993) The expanding clinical spectrum of mitochondrial diseases. *Brain Dev.*, **15**, 1–22.
2. Taylor, R.W. and Turnbull, D.M. (2005) Mitochondrial DNA mutations in human disease. *Nat. Rev. Genet.*, **6**, 389–402.
3. Zeviani, M. and Di Donato, S. (2004) Mitochondrial disorders. *Brain*, **127**, 2153–2172.
4. Jacobs, H.T. (2003) Disorders of mitochondrial protein synthesis. *Hum. Mol. Genet.*, **12**, R293–R301.
5. Brandon, M.C., Lott, M.T., Cuong Nguyen, K., Spolim, S., Navethe, S.B., Baldi, P. and Wallace, D.C. (2005) MITOMAP: a human mitochondrial genome database—2004 update. *Nucleic Acids Res.*, **33**, D611–D613.
6. Helm, M., Brule, H., Friede, D., Giege, R., Putz, D. and Florentz, C. (2000) Search for characteristic structural features of mammalian mitochondrial tRNAs. *RNA*, **6**, 1356–1379.
7. Florentz, C., Sohm, B., Tryoen-Toth, P., Putz, J. and Sissler, M. (2003) Human mitochondrial tRNAs in health and disease. *Cell. Mol. Life Sci.*, **60**, 1356–1375.
8. Wittenhagen, L.M. and Kelley, S.O. (2003) Impact of disease-related mitochondrial mutations on tRNA structure and function. *Trends Biochem. Sci.*, **28**, 605–611.
9. Smith, P.M., Ross, G.F., Taylor, R.W., Turnbull, D.M. and Lightowlers, R.N. (2004) Strategies for treating disorders of the mitochondrial genome. *Biochim. Biophys. Acta*, **1659**, 232–239.
10. DiMauro, S. and Schon, E.A. (2001) Mitochondrial DNA mutations in human disease. *Am. J. Med. Genet.*, **106**, 18–26.
11. Florentz, C. and Sissler, M. (2001) Disease-related versus polymorphic mutations in human mitochondrial tRNAs. Where is the difference? *EMBO Rep.*, **2**, 481–486.
12. Schon, E.A., Bonilla, E. and DiMauro, S. (1997) Mitochondrial DNA mutations and pathogenesis. *J. Bioenerg. Biomembr.*, **29**, 131–149.
13. Sternberg, D., Chatzoglou, E., Laforêt, P., Fayet, G., Jardel, C., Bloudy, P., Fardeau, M., Amselem, S., Eymard, B. and Lombès, A. (2001) Mitochondrial DNA transfer *RNA* gene sequence variations in patients with mitochondrial disorders. *Brain*, **124**, 984–994.
14. McFarland, R., Elson, J.L., Taylor, R.W., Howell, N. and Turnbull, D.M. (2004) Assigning pathogenicity to mitochondrial tRNA mutations: when 'definitely maybe' is not good enough. *Trends Genet.*, **20**, 591–596.
15. Kern, A.D. and Kondrashov, F.A. (2004) Mechanisms and convergence of compensatory evolution in mammalian mitochondrial tRNAs. *Nat. Genet.*, **36**, 1207–1212.
16. Ballard, J.W.O. and Whitlock, M.C. (2004) The incomplete natural history of the mitochondria. *Mol. Ecol.*, **13**, 729–744.
17. Wallace, D.C. (1994) Mitochondrial DNA sequence variation in human evolution and disease. *Proc. Natl Acad. Sci. USA*, **91**, 8739–8746.
18. Ingman, M., Kaessmann, H., Paabo, S. and Gyllensten, U. (2000) Mitochondrial genome variation and the origin of modern humans. *Nature*, **408**, 708–713.
19. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.
20. Niemi, A.K. *et al.* (2003) Mitochondrial DNA polymorphisms associated with longevity in a Finnish population. *Hum. Genet.*, **112**, 29–33.
21. Tanaka, M., Takeyasu, T., Fuku, N., Li-Jun, G. and Kurata, M. (2004) Mitochondrial genome single nucleotide polymorphisms and their phenotypes in the Japanese. *Ann. NY Acad. Sci.*, **1011**, 7–20.
22. Sunyaev, S.R., Lathe, W.C., III, Ramensky, V.E. and Bork, P. (2000) SNP frequencies in human genes an excess of rare alleles and differing modes of selection. *Trends Genet.*, **16**, 335–337.
23. Sunyaev, S., Ramensky, V., Koch, I., Lathe III, W., Kondrashov, A.S. and Bork, P. (2001) Prediction of deleterious human alleles. *Hum. Mol. Genet.*, **10**, 591–597.
24. Kondrashov, A.S., Sunyaev, S. and Kondrashov, F.A. (2002) Dobzhansky–Muller incompatibilities in protein evolution. *Proc. Natl Acad. Sci. USA*, **99**, 14878–14883.
25. Ramensky, V., Bork, P. and Sunyaev, S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.
26. Eyre-Walker, A. and Keightley, P.D. (1999) High genomic deleterious mutation rates in hominids. *Nature*, **397**, 344–347.
27. McFarland, R., Taylor, R.W., Chinnery, P.F., Howell, N. and Turnbull, D.M. (2004) A novel sporadic mutation in cytochrome *c* oxidase subunit II as a cause of rhabdomyolysis. *Neuromuscul. Disord.*, **14**, 162–166.
28. Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.
29. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
30. Janke, A., Feldmaier-Fuchs, G., Thomas, W.K., von Haeseler, A. and Paabo, S. (1994) The marsupial mitochondrial genome and the evolution of placental mammals. *Genetics*, **137**, 243–256.

Chapter 2, in full, is a reprint of the material at it appears in Kondrashov FA. (2005) Prediction of pathogenic mutations in mitochondrially encoded human tRNAs. *Human Molecular Genetics* **14,** 2415-2419. Oxford University Press 2005. The dissertation author was the primary investigator and author of this paper.

Chapter 3.

Conversion and compensatory evolution of the human γ-crystallin genes

ARTICLE

Conversion and Compensatory Evolution of the γ-Crystallin Genes and Identification of a Cataractogenic Mutation That Reverses the Sequence of the Human *CRYGD* Gene to an Ancestral State

Olga V. Plotnikova,* Fyodor A. Kondrashov,* Peter K. Vlasov, Anastasia P. Grigorenko, Evgeny K. Ginter, and Evgeny I. Rogaev

We identified a mutation in the *CRYGD* gene (P23S) of the γ-crystallin gene cluster that is associated with a polymorphic congenital cataract that occurs with frequency of ~0.3% in a human population. To gain insight into the molecular mechanism of the pathogenesis of γ-crystallin isoforms, we undertook an evolutionary analysis of the available mammalian and newly obtained primate sequences of the γ-crystallin genes. The cataract-associated serine at site 23 corresponds to the ancestral state, since it was found in *CRYGD* of a lower primate and all the surveyed nonprimate mammals. Crystallin proteins include two structurally similar domains, and substitutions in mammalian CRYGD protein at site 23 of the first domain were always associated with substitutions in the structurally reciprocal sites 109 and 136 of the second domain. These data suggest that the cataractogenic effect of serine at site 23 in the N-terminal domain of CRYGD may be compensated indirectly by amino acid changes in a distal domain. We also found that gene conversion was a factor in the evolution of the γ-crystallin gene cluster throughout different mammalian clades. The high rate of gene conversion observed between the functional *CRYGD* gene and two primate γ-crystallin pseudogenes (*CRYGEP1* and *CRYGFP1*) coupled with a surprising finding of apparent negative selection in primate pseudogenes suggest a deleterious impact of recently derived pseudogenes involved in gene conversion in the γ-crystallin gene cluster.

Cataracts are characterized by opaqueness of all or part of the eye crystallin lens[1] and are the most common cause of blindness in the world, with congenital cataracts frequently resulting in blindness or visual impairment in children.[2] The estimated prevalence is 2.2–2.49 cases per 10,000 live births,[3] and ~50% of all infantile cataract cases are genetic.[2] Most cases occur as isolated pathologies, but some forms are associated with other abnormalities.[4] Although congenital cataracts can be transmitted as a recessive or an X-linked trait, autosomal dominant inheritance occurs most frequently and exhibits both clinical variability and genetic heterogeneity.[2]

Clinical and molecular genetics studies have led to the identification of multiple candidate disease loci for congenital cataracts. Mutations in genes encoding four specific types of proteins have been described in association with the phenotype of nonsyndromic inherited cataracts. These include members of the α-, β-, and γ-crystallin families[5,6] (MIM +123580, +123590, *123610, *123620, *123630, *123631, +600929, +123680, +123690, and *123730), three transcription factors (MAF[7] [MIM *177075], PITX3[8] [MIM +602669], and HSF4[9] [MIM *602438]), cytoskeletal protein BFSP2[10] (MIM *603212), and membrane-transport proteins MIP[1] (MIM +154050), GJA3 (CX46)[11] (MIM *121015), and GJA8 (CX50)[12] (MIM *600897). Approximately half of all mutations associated with congenital cataracts are located in crystallin genes.[13]

Crystallins are the major water-soluble structural proteins expressed in the mammalian eye lens and consist of three major families—the α-, β-, and γ- crystallins[14]—with the γ-crystallin composing up to 40% of the soluble proteins expressed in the lens.[15] In humans, the γ-crystallin gene cluster is located on chromosome 2q33-q35 and consists of genes *CRYGA* (MIM *123660; GenBank accession numbers M17315 and M17316), *CRYGB* (MIM *123670; GenBank accession number M19364), *CRYGC* (MIM +123680; GenBank accession numbers K03003 and K03004), and *CRYGD* (MIM +123690; GenBank accession numbers K03005 and K03006)[16] (encoding γ*A*-, γ*B*-, γ*C*-, and γ*D*-crystallins, respectively), with cataract-associated mutations in two of these genes (*CRYGD* and *CRYGC*) that code for the most abundant γ-crystallin proteins in the lens.[17] Two other γ-crystallin genes—*CRYGEP1* (GenBank accession numbers K03007 and K03008) (encoding γ*E*-crystallin) and *CRYGFP1* (GenBank accession numbers K03009 and K03010) (encoding γ*F*-crystallin) (both MIM

From the Laboratory of Molecular Brain Genetics, Research Center of Mental Health (O.V.P.; A.P.G.; E.I.R.), Engelhardt Institute of Molecular Biology (P.K.V.), Research Center of Medical Genetics (E.K.G.), Vavilov Institute of General Genetics (E.I.R.), Russian Academy of Sciences, and Lomonosov Moscow State University (E.I.R.), Moscow; Department of Psychiatry, Brudnick Neuropsychiatric Research Institute, University of Massachusetts Medical School, Worcester (O.V.P.; A.P.G.; E.I.R.); and Section on Ecology, Behavior and Evolution, Division of Biological Sciences, University of California at San Diego, La Jolla (F.A.K.)

Received January 3, 2007; accepted for publication March 30, 2007; electronically published May 16, 2007.

Address for correspondence and reprints: Dr. Evgeny I. Rogaev, Department of Psychiatry, Brudnick Neuropsychiatric Research Institute, University of Massachusetts Medical School, 303 Belmont Street, Worcester, MA 01604. E-mail: Evgeny.Rogaev@umassmed.edu

* These two authors contributed equally to this work.

Am. J. Hum. Genet. 2007;81:32–43. 0002-9297/2007/8101-0004$15.00
DOI: 10.1086/518616

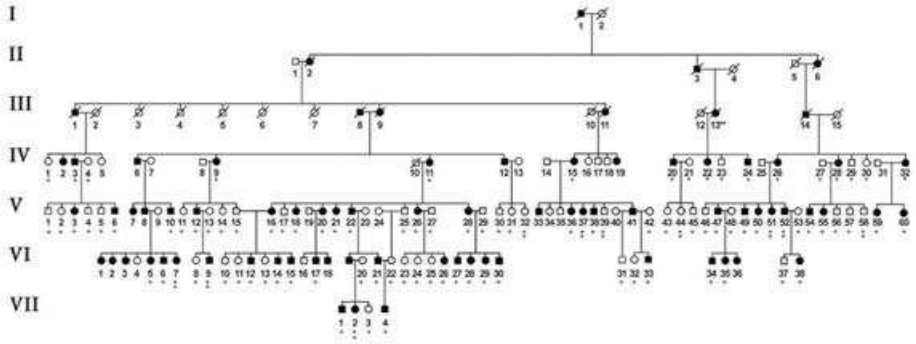


Figure 1. Abridged PCC-affected pedigree selected from the genetic isolate. The mutation is transmitted as an autosomal dominant trait. The affected individuals are represented by blackened squares (males) and circles (females), and the unaffected individuals are represented by unblackened symbols. Family members participating in this study are indicated by an asterisk (*). One asterisk indicates subjects genotyped by restriction enzyme–digestion analysis, and two asterisks indicate individuals genotyped by direct-sequencing analysis.

*123660)—are also located in the same cluster. However, in humans, they harbor a stop codon and are considered pseudogenes, whereas, in other nonprimate mammals, these genes appear functional.[18] Cataractogenesis is anticipated to be a strong factor in selection processes of genes for lens proteins; however, very little is yet known about the evolution of different members of γ-crystallin genes, especially in human and primate lineages. We have previously linked nonnuclear polymorphic congenital cataract (PCC [MIM %601286]) to the γ-crystallin gene cluster (*CRYG*) on the human chromosome 3q33-35 in a large pedigree from a Central Asian population.[19] Here, we screened the PCC-affected pedigree for mutations in the *CRYGA–CRYGD* genes and performed an evolutionary and structural analysis of the mutation and the γ-crystallin gene family.

Material and Methods

Mutation Analysis

The collection of DNA samples from subjects with PCC was described elsewhere.[19] The genomic sequence of human *CRYGA–CRYGD* genes was obtained from the GenBank database.[20] To search for mutations, the protein-coding regions of these genes were amplified by PCR, by use of genomic DNA from probands of the PCC-affected pedigree. Pairs of oligonucleotide primers flanking the exons of human *CRYGA–CRYGD* genes were designed manually or by Primer3 and were used for PCR amplification and sequencing of the PCR products (primer oligonucleotide sequences are available from the authors on request). PCR was performed for 32 cycles at 94°C for 3 min, with an annealing temperature of 56°C–58°C for 30 s, and at 72°C for 4 min. Each PCR was performed in a volume of 25 μl that contained 10–20 pmol of each primer, 1× reaction buffer, 50 ng DNA, 200 μM dNTP, 2.5–3 mM $MgCl_2$, and 0.2 U *Taq* polymerase. The PCR products were purified with electrophoresis in a 1% agarose gel, 1× TBE buffer, and the QIAEX II Kit gel extraction kit (QIAGEN). The purified PCR products were sequenced directly with use of an ABI Prism 310 Automated Sequencer with the ABI Prism Big-Dye Terminator cycle sequencing kits (Applied Biosystems).

The C70T mutation in the *CRYGD* gene was initially found in selected probands by direct sequencing. The presence or absence of the mutation was elucidated further by restriction enzyme–digestion assay in genomic DNA samples from all affected and unaffected family members of the pedigree. To distinguish the genotypes of unaffected and heterozygous individuals for this particular mutation, we designed nucleotide substitutions in one of the primers (reverse int) to create a new site for *Bpm*I-restriction endonuclease in the mutant C70T allele. Exon 2 of the *CRYGD* gene was amplified by two rounds of PCR with the primers direct ext (5′-GCAGCCCCACCCGCTCA-3′) and reverse ext (5′-GGGTA-ATACTTTGCTTATGTGGGG-3′) and then with internal primers direct int (5′-AGCCATGGGGAAGGTGAG-3′) and reverse int (5′-AGTAGGGCTGCAGGCTGG-3′). The PCR products were digested for 3–4 h at 37°C with *Bpm*I, and resulting DNA restriction fragments were analyzed on a 7% polyacrylamide gel and were visualized using ethidium bromide staining. In total, we analyzed 54 individuals with cataract and 46 unaffected individuals from the Middle Asian PCC pedigree. In addition, families with obesity from the same genetic isolate (22 individuals) were genotyped. We also tested 512 control chromosomes from white (206 chromosomes from Russians) and mixed white and Mongolian (224 chromosomes from Tatars and 82 chromosomes from Bashkirs) populations. The cataract-associated mutation (C70T) was detected in affected individuals from the PCC-affected pedigree only.

Sequencing of Primate Genes

To determine nucleotide sequences for ORFs of functional γ-crystallin genes (*CRYGA–CRYGD*) in primates, we used the PCR oligonucleotide primers based on human sequences or redundant oligonucleotide primers based on macaque, chimpanzee, and hu-

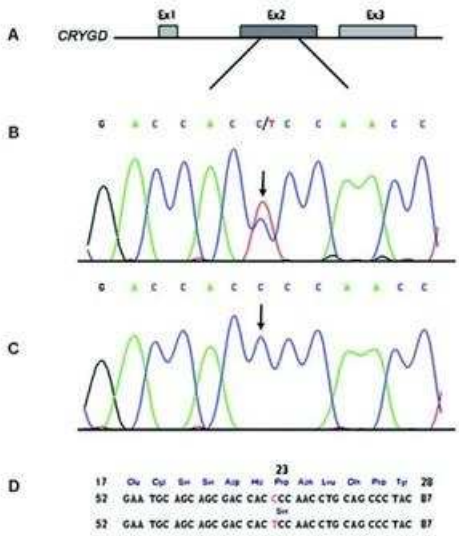


Figure 2. Identification of the mutation in individuals with PCC. *A*, The three exons (Ex) that comprise *CRYGD*. *B* and *C*, Sequence chromatogram from a heterozygous patient carrying the P23S mutation (*B*) and a homozygous wild-type sequence from an unaffected individual (*C*). *D*, Exon sequence of *CRYGD* bearing the cataract-associated PS23S mutation.

man intronic sequences, such that these primers flanked exons 1, 2, and 3. The PCRs and sequencing were performed as described above. The γ-crystallin gene sequences were determined in species of the following primate families: *Hominidae* (*Pan paniscus* [pygmy chimpanzee, GenBank accession numbers EF467187, EF467196, EF467205, and EF467214]; *Pan troglodytes* [chimpanzee, GenBank accession numbers EF467190, EF467199, EF467208, and EF467217]; *Gorilla gorilla* [gorilla, GenBank accession numbers EF467183, EF467192, EF467201, and EF467210]; *Pongo pigmaeus* [orangutan, GenBank accession numbers EF467188, EF467197, EF467206, and EF467215]); *Hylobatidae* (*Hylobates lar* [gibbon, GenBank accession numbers EF467184, EF467193, EF467202, and EF467211); *Cecropithecidae* (*Macaca mulatta* [rhesus monkey, GenBank accession numbers EF467186, EF467195, EF467204, and EF467213]); *Cebidae* (*Lagothrix lagotricha* [common woolly monkey, GenBank accession numbers EF467185, EF467194, EF467203, and EF467212]); *Ateles geoffroyi* [black-handed spider monkey, GenBank accession numbers EF467182, EF467191, EF467200, and EF467209]); and *Callitrichidae* (*Saguinus labiatus* [red-chested mustached tamarin, GenBank accession numbers EF467189, EF467198, EF467207, and EF467216]). In addition, we determined gene sequence for the putative pseudogene *CRYGFP1* in *Pan paniscus* (GenBank accession number EF492219), *G. gorilla* (GenBank accession number EF492217), *Pongo pigmaeus* (GenBank accession number EF492220), and *H. lar* (GenBank accession number EF492218). PCR primer oligonucleotide sequences are available from the authors on request. Genomic DNA samples were obtained from Coriell Cell Repositories and from our own collection of primate DNAs.

Bioinformatic and Structural Analysis

In addition to the sequenced primate genes, we used the sequence information of the γ-crystallin genes from GenBank.[20] We also used the information on gene order from completely sequenced mammalian genomes of *Monodelphis domestica*, *Canis familiaris*, *Mus musculus*, and *Bos taurus*, using the UCSC Genome Browser.[21] The rat genome (*Rattus norvegicus*) was excluded from the synteny analysis because of a likely error of assembly of the γ-crystallin cluster. A multiple alignment of all sequences was made using the MUSCLE alignment program.[22] We reconstructed the phylogeny of the γ-crystallin genes, using a Bayesean approach as implemented in MrBayes, with 1 million iterations (mcmc ngen = 1,000,000 in MrBayes) with the General Time Reversible model.[23] Sequence divergence between genes was estimated using the codeml program in the PAML package.[24] Exon 1, which encodes 3 aa, was omitted from the phylogenetic analysis.

Results

A Novel Pathogenic Mutation in the GRYGD *Gene*

Elsewhere, we established a link between PCC and the cluster of γ-crystallin genes (*CRYG*) at chromosome 2q33-35 in a large, unique pedigree of a family from a Central Asian population.[19] This population of mixed white and Mongolian origin is characterized by tribe ancestry, high endogamy, and complex ethnic genesis. Two inherited monogenic diseases were accumulated in these populations: an autosomal dominant cataract (PCC), with frequency of the mutant gene of ~0.26%, and autosomal recessive obesity, with frequency of the mutant gene of ~2.47%.[19,25,26] The large 7-generation PCC-affected pedigree characterized by high endogamy and a high coefficient of inbreeding (>3%) was selected for the molecular-genetic study. Of the 157 pedigree individuals, 105 had PCC, and DNA samples from 100 members, including 54 affected individuals, were available for mutation analysis of the *CRYG* cluster (fig. 1).

We amplified and sequenced the coding regions of the *CRYGA–CRYGD* genes and identified a novel nonsynonymous C70T (P23S) mutation in the coding region of exon 2 of the *CRYGD* gene, which was found to cosegregate with the disease (fig. 2). The presence or absence of the mutation was confirmed by sequence analysis of four affected and four unaffected individuals from the PCC-affected pedigree and then by restriction enzyme–digestion assay (as described in the "Material and Methods" section) of DNA samples from all available members from the PCC-

The figure is available in its entirety in the online edition of *The American Journal of Human Genetics*.

Figure 3. Sequence alignment of γ-crystallins. The legend is available in its entirety in the online edition of *The American Journal of Human Genetics*.

Species	Sites and residues												
Primates	22	23	24	…	48	49	50	…	108	109	110	…	136	137	138
γD-*H. sap*	H	P	N		P	N	Y		C	S	C		S	N	Y
γD-*P.trog*	H	P	N		P	N	Y		C	S	C		S	N	Y
γD-*P.panis*	H	P	N		P	N	Y		C	S	C		S	N	Y
γD-*P.pyg*	H	P	N		P	N	Y		C	S	C		S	N	Y
γD-*G.gor*	H	P	N		P	N	Y		C	S	C		S	N	Y
γD-*H.lar*	H	P	N		P	N	Y		C	S	C		S	N	Y
γD-*M.mul*	H	P	N		P	N	Y		C	S	S		P	N	Y
γD-*A.geof*	C	P	N		P	N	Y		C	T	S		P	N	Y
γD-*L.lagot*	C	P	N		P	N	Y		C	T	S		P	N	Y
γD-*S.lab*	H	S	N		P	N	Y		C	T	S		P	N	Y
Others															
γD-*B.taur*	H	S	N		P	N	Y		C	S	S		P	N	Y
γD-*C.fam*	H	S	N		P	N	Y		C	S	S		P	N	Y
γD-*R.nor*	H	S	N		P	N	F		C	P	S		T	N	Y
γD-*M.mus*	H	S	N		P	N	F		C	P	S		T	N	Y
Opossum	H	S	N		P	N	Y		C	S	S		P	N	Y

Figure 4. Sequence patterns in the interacting fragments of two γ-crystallin domains. Correlated serine and proline residues in CRYGD are shown in red and green, respectively. *H. sap* = *H. sapiens*; *P. trog* = *Pan troglodytes*; *P. panis* = *Pan paniscus*; *P. pyg* = *Pongo pygmaeus*; *G. gor* = *G. gorilla*; *M. mul* = *Macaca mulatta*; *A. geof* = *A. geoffroyi*; *L. lagot* = *L. lagotricha*; *S. lab* = *S. labiatus*; *B. taur* = *B. taurus*; *C. fam* = *C. familiaris*; *R. nor* = *R. norvegicus*; *M. mus* = *Mus musculus*.

affected pedigree. All affected individuals—but none of the related unaffected individuals from the PCC-affected pedigree or other unaffected, unrelated control individuals—were found to be heterozygous for this mutation. Three common synonymous SNPs in *CRYGB* (C192T) and *CRYGD* (T51C and T392C) were also detected. These polymorphic changes, however, showed no cosegregation with PCC in this pedigree. We found no *CRYGD* C70T cataract-associated mutation in unaffected individuals from the same genetic isolate or in 512 control chromosomes from populations of white or mixed white and Mongolian origin (see the "Material and Methods" section). The data strongly demonstrated that the nonsynonymous C70T (P23S) substitution in *CRYGD* is the only mutation in the γ-crystallin gene cluster that segregates with PCC.

Compensatory Evolution Reveals Structural Characteristics of the γ-Crystallins

Mutations that have a pathogenic effect when they are harbored in a human gene can be benign in other, sometimes closely related, organisms.[27,28] Such cases are known as "compensated pathogenic deviations" (CPDs), since it is thought that the deleterious effect of such mutations is neutralized by another, compensatory mutation. Unlike mutations in γ-crystallins described elsewhere, the mutation we describe here also appears to be a CPD, such that the disease-causing variant is found in the normal *CRYGD* sequence of several wild-type organisms, including one primate (fig. 3).

The basis of compensations of CPDs is usually the maintenance of structural stability within a single molecule,[27–30] although, in a few cases, the compensatory mutation and the CPD may be located on two different interacting proteins.[27] To investigate the molecular nature of the compensation of the P23S substitution, we assembled sequences of the γ-crystallin genes from the available mammalian genomes (human, chimpanzee, macaque, dog, mouse, rat, cow, and opossum) and the genew sequences for several primates that we determined in this study (*A. geoffroyi*, *S. labiatus*, *L. lagotricha*, *Macaca mulatta*, *H. lar*, *Pongo pygmaeus*, *G. gorilla*, *Pan paniscus*, and *Pan troglodytes*). The resulting multiple alignment was analyzed on the basis of the available crystal structure of the human CRYGD protein[31] (Protein Data Bank ID 1h4a).

The β- and γ-crystallin polypeptides fold into Greek key motifs that form two structurally similar domains.[31–33] On the γD-crystallin structure, site 23 interacts with position 49, and, since many compensatory substitutions for CPDs have been found in interacting sites, we surveyed the amino acid at site 49 in mammalian γD-crystallin genes. However, we found that site 49 and the neighboring sites are generally conserved throughout evolution and show no evidence of compensatory evolution with site 23 (fig. 4).

Thus, we undertook a correlation analysis in search of compensatory substitutions in the entire γD-crystallin protein. We searched for the compensatory site on the basis of the expected pattern of compensatory evolution[27]; that is, all species harboring serine at site 23 must have a single predicted compensatory amino acid at another site. No single site conformed to this prediction. However, sites 109 and 136 conformed in conjunction, such that the pathogenic state S23 was not observed together with the human state in either site 109 or 136 (fig. 4). Thus, in the course of evolution at site 23, in the common ancestor of

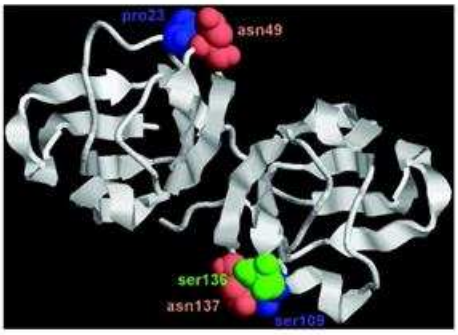


Figure 5. Contact regions in two symmetrical crystallin domains and four similar Greek key motifs form β-strands in two joined protein domains. The sample is based on the structure of the human γD-crystallin protein.

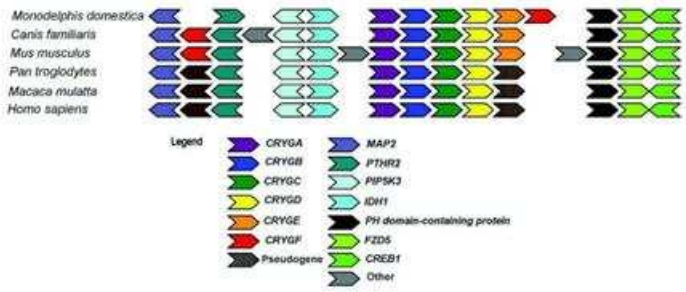


Figure 6. Gene order of paralogous γ-crystallin genes in mammalian genomes

primates, the ancestral serine changed to proline. In *S. labiatus*, the proline reversed to the ancestral serine, apparently without deleterious effects, whereas the same reversal in humans results in cataract formation. The deleterious effect of P23S substitution in humans is most likely related to the P136S substitution that also occurred in the primate, apparently the *Hominidae-Hylobatidae* common ancestor. Interestingly, the amino acid is also reverted to the ancestral state (S136P) in *S. labiatus* that have 23S in the wild-type allele. In general, the presence of serine at site 23 appears to be tolerated under the condition that either site 109 or site 136 is occupied by a proline. Remarkably, site 109 corresponds to the same position in the second domain as site 23 in the first γD-crystallin domain, whereas the corresponding site of the interaction site 49 in the second domain is site 137 (fig. 5).

Apparent Selection and Gene Conversion in the γ-Crystallin Gene Family

There is a large diversity of crystallin genes in higher animals; the γ-crystallin family is mammalian specific.[34] The γ-crystallin family is located in tandem with one of the genes (*CRYGFP1*) slightly removed from the rest of the cluster (fig. 6). The high sequence similarity of some of the γ-crystallin genes in humans[14] and a phylogenetic analysis of the gene family in rats[35] indicated that the γ-crystallin genes undergo gene conversion, which is apparently restricted to exon 2.[35] In addition, it is thought that two of the six γ-crystallin genes (*CRYGEP1* and *CRYGFP1*) have turned into a pseudogene in humans and chimpanzees, as evidenced by the presence of stop codons in the beginning of the second codon.[14,18] Several diseases are caused by gene conversion from a degenerate pseudogene into a functional gene,[36] including cataractogenesis by gene conversion of the β-crystallin pseudogene to one of the β-crystallin functional genes.[37] The presence of closely related pseudogenes in the γ-crystallins coupled with the reported gene-conversion events opens up the possibility that the same mechanism is the cause of some fraction of familial cataracts.

Gene conversion can make it particularly difficult to establish orthology[38]; therefore, we relied on synteny to resolve orthologous relationships within the γ-crystallin gene family. Indeed, the syntenic structure is well preserved within the mammalian clade, with only one of the genes (*CRYGFP1*) separated from an otherwise tandem arrangement of the γ-crystallin genes in the common ancestor of placental mammals (fig. 6).

In rats, gene conversion appears to preferentially affect the second exon of the γ-crystallins.[35] Thus, we constructed separate phylogenies for exons 2 and 3 for all six of the γ-crystallin genes for all available mammals, which, for the first time, included primates (fig. 7). On a phylogeny, gene-conversion events appear at a point of common ancestry of paralagous sequences.[38] In exon 2, gene conversion was found across all of the genes in the γ-crystallin family and in all surveyed taxa, but it was particularly common in nonprimate mammals and in *CRYGD*, *CRYGEP1*, and *CRYGFP1* (fig. 7*A*).

Since the divergence of the macaque and human lineages, there have been at least two fixed gene-conversion events in the macaque genome (one between *CRYGEP1* and *CRYGFP1* and one between *CRYGD* and the preconverted *CRYGE* or *CRYGFP1* sequence). In the human-chimpanzee lineage, there have been a gene-conversion event between *CRYGE* and *CRYGFP1* after the macaque split and a probable gene-conversion event around the time of divergence from macaques (fig. 7*A*). The rate of gene conversion in primates appears to be higher in *CRYGEP1*, *CRYGFP1*, and *CRYGD*, since only one gene-conversion event involving other genes (between *CRYGB* and *CRYGC*) was fixed around the time of divergence of the surveyed primate species (fig. 7*A*). The rate of gene conversion in exon 3 was not as high as in exon 2, but some conversion events were still observed (fig. 7*A*). Exon 3 of *CRYGEP1* and *CRYGFP1* was involved in at least five gene-conversion

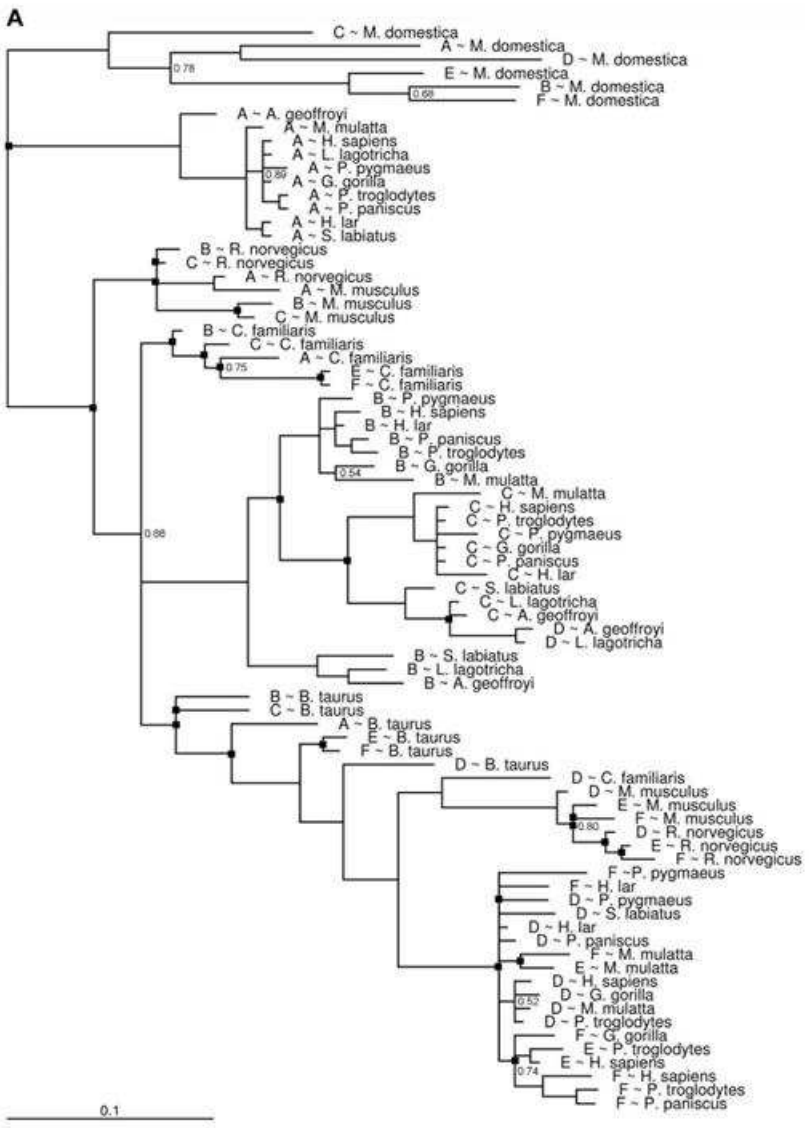


Figure 7. Bayesian phylogenies of exon 2 (*A*) and exon 3 (*B*) of the γ-crystallins. Gene-conversion events are shown on the phylogeny with a blackened square, and only posterior probabilities <0.90 are indicated. A, B, C, D, E, and F correspond to the *CRYGA*, *CRYGB*, *CRYGC*, *CRYGD*, *CRYGE*, and *CRYGF* genes, respectively.

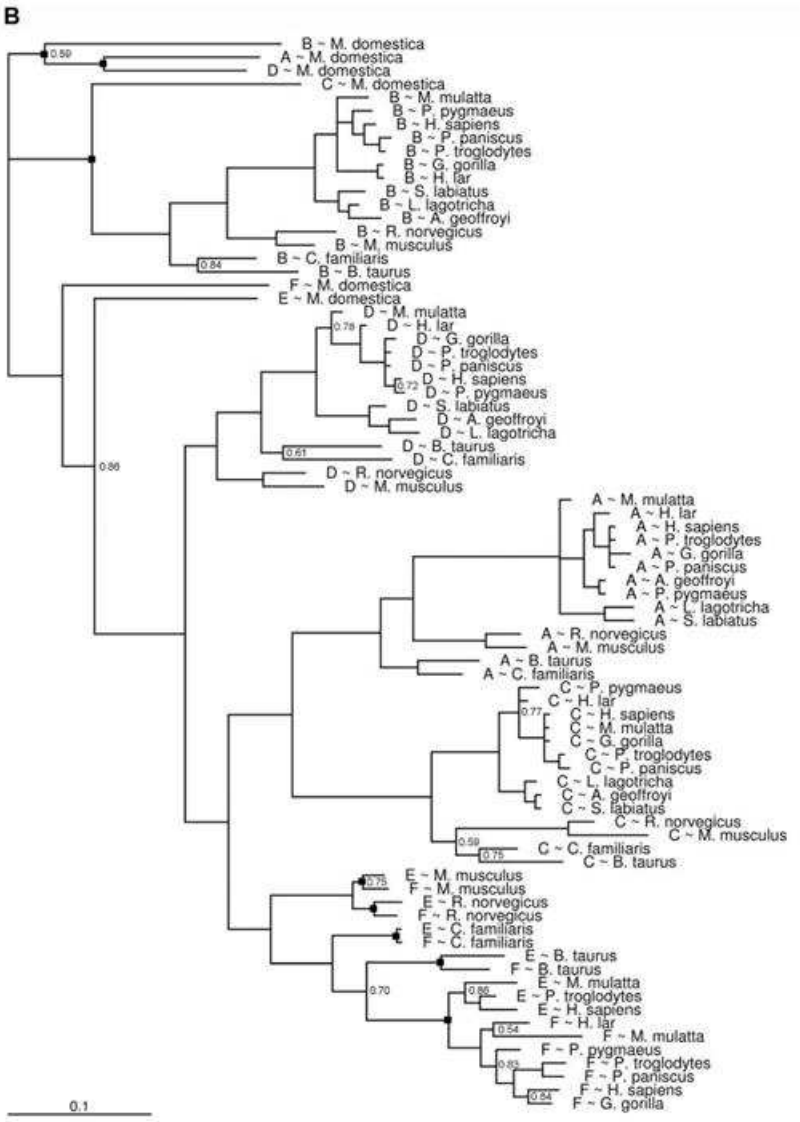

B
B ~ M. domestica
A ~ M. domestica
D ~ M. domestica
C ~ M. domestica
B ~ M. mulatta
B ~ P. pygmaeus
B ~ H. sapiens
B ~ P. paniscus
B ~ P. troglodytes
B ~ G. gorilla
B ~ H. lar
B ~ S. labiatus
B ~ L. lagotricha
B ~ A. geoffroyi
B ~ R. norvegicus
B ~ M. musculus
B ~ C. familiaris
B ~ B. taurus
F ~ M. domestica
E ~ M. domestica
D ~ M. mulatta
D ~ H. lar
D ~ G. gorilla
D ~ P. troglodytes
D ~ P. paniscus
D ~ H. sapiens
D ~ P. pygmaeus
D ~ S. labiatus
D ~ A. geoffroyi
D ~ L. lagotricha
D ~ B. taurus
D ~ C. familiaris
D ~ R. norvegicus
D ~ M. musculus
A ~ M. mulatta
A ~ H. lar
A ~ H. sapiens
A ~ P. troglodytes
A ~ G. gorilla
A ~ P. paniscus
A ~ A. geoffroyi
A ~ P. pygmaeus
A ~ L. lagotricha
A ~ S. labiatus
A ~ R. norvegicus
A ~ M. musculus
A ~ B. taurus
A ~ C. familiaris
C ~ P. pygmaeus
C ~ H. lar
C ~ H. sapiens
C ~ M. mulatta
C ~ G. gorilla
C ~ P. troglodytes
C ~ P. paniscus
C ~ L. lagotricha
C ~ A. geoffroyi
C ~ S. labiatus
C ~ R. norvegicus
C ~ M. musculus
C ~ C. familiaris
C ~ B. taurus
E ~ M. musculus
F ~ M. musculus
E ~ R. norvegicus
F ~ R. norvegicus
E ~ C. familiaris
F ~ C. familiaris
E ~ B. taurus
F ~ B. taurus
E ~ M. mulatta
E ~ P. troglodytes
E ~ H. sapiens
F ~ H. lar
F ~ M. mulatta
F ~ P. pygmaeus
F ~ P. troglodytes
F ~ P. paniscus
F ~ H. sapiens
F ~ G. gorilla
0.59
0.84
0.79
0.72
0.61
0.86
0.77
0.59
0.75
0.75
0.70
0.80
0.54
0.83
0.84
0.1

Table 1. Sequence Divergence between Human and Macaque γ-Crystallin Genes

Gene	No. of Substitutions		dn:ds[a]	P[b]
	Nonsynonymous	Synonymous		
CRYGA	4.12	6.25	.1338	<.05
CRYGB	7.22	14.93	.1001	<.001
CRYGC	5.16	5.08	.1953	NS
CRYGD	9.32	4.31	.3322	NS
CRYGEP1	11.46	15.12	.1041	<.001
CRYGFP1	21.22	17.06	.2236	<.01

[a] The rate of synonymous (ds) and nonsynonymous (dn) evolution.
[b] The *P* value of a χ^2 test of the number of synonymous and nonsynonymous substitutions observed compared with the number expected under neutrality. NS = not significant.

events since the radiation of placental mammals: two in rodents, one each in dog and cow, and one in primates before the divergence of humans and macaques.

The relatively high rate of gene conversion in primates that involves two of the genes that have been pseudogenized in humans and a functional gene reveals the possibility that some of the degenerate mutations that have accumulated in the pseudogene may be transferred to the functional gene, with deleterious or pathogenic consequences. Such frequent gene conversion between a pseudogene and a functional gene should lead to a pattern of negative selection in the pseudogene, because selection would act against degenerate mutations in the pseudogene if gene conversion events frequently transfer such mutations to a functional gene.[39,40]

Although *CRYGEP1* and *CRYGFP1* are thought to be pseudogenes in humans and chimpanzees,[18] it is not known whether they are functional in other primates. We retrieved the macaque *CRYGEP1* and *CRYGFP1* sequences from orthologous (determined by synteny) sections of the macaque genome, and we compared the obtained sequences with those of other γ-crystallin genes from different species. The macaque *CRYGFP1* sequence contained three frameshift mutations, strongly suggesting that it is a pseudogene. The functional status of *CRYGEP1* was not as apparent. Although the macaque *CRYGEP1* gene showed seven amino acid changes in evolutionarily conserved sites in all γ-crystallins, no inframe stop codons or frameshift mutations were observed.

A common way to test the strength of selection acting on protein-coding genes is to compare the rate of synonymous (ds) and nonsynonymous (dn) evolution.[38] A neutrally evolving sequence, such as a pseudogene, is expected to show the same rate of evolution in these two types of sites (dn:ds = 1). We compared the rates of evolution between human and macaque sequences for all six genes in the γ-crystalline cluster (table 1). We found no evidence to support a neutral level of evolution in *CRYGEP1* (dn:ds = 0.1041) that would suggest that *CRYGEP1* is an active gene in the macaque; however, we also found that the dn:ds ratio significantly differs from 1 even for the *CRYGFP1* pseudogene (dn:ds = 0.2236). Such apparent selection is commonly observed on the phenotypic level[41]; however, to our knowledge, it has been observed only once on a molecular level, in a mouse c-ubiquitin cluster by an uneven crossing-over event.[42] We determined the *CRYGFP* sequence in *Pan paniscus, Pongo pygmaeous,* and *H. lar* and found signs of pseudogenization of the gene in different primate species (table 2) (see the "Discussion" section). To our knowledge, this is the first study that observes apparent selection in nonfunctional sequence due to gene conversion with a functional gene.

Discussion

To date, six different mutations in the *CRYGD* gene have been found in patients with cataracts: R14C,[42] R36S,[44,45] R58H,[46] W156X,[47] P23T,[47] and E107A[48] (MIM +123690.0001–123690.0006). The seventh mutation identified here (P23S) appears to be the cause of the nonnuclear PCC. This mutation is accumulated in an isolated population (with frequency of ~0.26%) along with autosomal recessive obesity (with frequency of the mutant-gene allele of ~2.5%). We found no connection or cosegregation of these pathologies, which were probably inherited from different ancestral founders. The PCC type of cataract is characterized by a nonprogressive phenotype and partial opacity of the lens, which has a variable location on the periphery between the fetal nucleus of the lens and the equator. The opacities are irregular and look similar to a bunch of grapes or a lump of cotton balls and may be present simultaneously in different lens layers.[19] Another amino acid substitution at the same site has been described elsewhere, a proline→threonine substitution (P23T) in a family with congenital cataracts.[47,49–53] Generally, the clinical manifestation of the PCC-affected family was different from that of patients with other forms of familial cataracts caused by the P23T mutation of the *CRYGD* gene. The P23T mutation has been shown to cosegregate with variable phenotypes, such as the lamellar cataract in an Indian family and the fasciculiform, coralliform, and cerulean cataracts (MIM 608983) in a Moroccan family.[47,49–53] The segregation frequency of PCC strongly corresponds to an autosomal dominant form of inheritance (frequency [±SD] of 0.56 ± 0.04) with >90%–97% penetrance.[19,25] Thus, the variability of clinical phenotype may be caused by epigenetic factors during embryonic development or by gene modifiers.

In vitro experimental data about the nature of the P23T substitution suggested that this change does not significantly alter the stability structure of the protein but, rather, affects the protein solubility,[52,53] resulting in clusters of the P23T-mutant protein. Because of a high similarity of the side chains of thrionine and serine, Evans et al.[52] also considered the impact of a nonnatural P23S substitution and found that P23S also affects the solubility of γD-crystallin, although not as profoundly. This effect may be caused by a change in the hydrogen-binding characteristics of the protein-water interface. A substitution of a

Table 2. Signs of Pseudogenization in the Primate *CRYGFP1* Pseudogene

Species	Sign(s) of Pseudogenization
H. sapiens	56 Y→STOP and 1-nt deletion at site 422
Pan troglodytes	56 Y→STOP, 1-nt deletion at site 422 and deletion of the 5′ end of exon 3 sites 252–267
Pan paniscus	56 Y→STOP, 1-nt deletion at site 422, deletion of the 5′ end of exon 3 sites 252–267, and A insertion at site 455
G. gorilla	56 Y→STOP
Pongo pygmaeus	42 C→STOP
H. lar	Second intron splice sites AT…GG
Macaca mulatta	1 M→K, G deletion at site 30, 28-nt insertion at site 425, and 4-nt insertion at site 511

proline, since it is an imino acid that does not have a hydrogen bond–forming NH group, is particularly capable of affecting solubility in water. Thus, it is likely that the P23S substitution is compensated for in another part of the γD-crystallin protein.

This observation is in agreement with our observation that the probable compensatory substitution is a reversal to a proline in structurally reciprocal sites in a different domain. The distribution patterns of the proline/serine residues that we describe here may play an important role in the protein-water–system stabilization. In particular, these sites can be the part of the crystallin interaction interface with other lens proteins where, possibly, γD-crystallins are connected with each other. Substitutions in these sites may lead to protein aggregation in solution that dramatically changes the lens crystal transparency. The observation that two sites, 109 and 136, may compensate for the same CPD (P23S) is not completely unexpected, since examples of compensatory interactions involving more than two interacting sites have been described.[28,29]

We find evidence to support active gene conversion between the pseudogene *CRYGFP1* and the functional *CRYGD* gene copy in recent evolution (since the divergence of human and macaque lineages). In addition, we find that a conversion event between these two genes also occurred sometime in the primate lineage before the divergence of humans and macaques (fig. 7). The phylogeny can reveal only fixed gene-conversion events, such that, if gene conversion events still occur but are not fixed, they will not be observed. In particular, the probability of fixation of a gene-conversion event will be much lower if it brings a deleterious substitution into a functional gene.[39,40] Thus, gene conversion between the pseudogenes and the functional gene copies may have occurred with a much higher frequency on a mutational level than is apparent from the phylogeny that reveals only fixed events on an evolutionary scale.

The observation of apparent negative selection in pseudogenes depends on the assumption that these genes became pseudogenes before the divergence of humans and macaques. There are no shared stop codons or frameshift mutations between the human and macaque pseudogenes; however, frequent gene conversion, which is seen between these pseudogenes and their functional copies, will erase such shared stop codons. To demonstrate this point, we sequenced the *CRYGFP1* pseudogene from *Pan paniscus*, *G. gorilla*, *Pongo pygmaeus*, and *H. lar*, in addition to an already available sequence from *H. sapiens*, *P. troglodytes*, and *Macaca mulatta*. We found that the *CRYGFP1* sequence is a pseudogene in each of these species; however, whereas the higher ape species share a stop codon, other species show different signs of pseudogenization (table 2). The possibility of very recent and independent pseudogenizations of the same gene in four separate lineages is extremely remote and is unlikely to explain our observations.

There are two nonexclusive ways in which gene conversion between a functional gene and a pseudogene can lead to apparent negative selection. The first is gene conversion of the functional sequence over that of the pseudogene. A comparison of pseudogene sequences from two species would reveal apparent negative selection ($dn < ds$) if such gene-conversion events were fixed after the divergence of the two species. Alternatively, selection may act on mutations in the pseudogene if such mutations are deleterious when they are converted to the functional gene. Whereas the first model is simple and requires only one fixed gene conversion of a functional copy to the pseudogene, the latter model encompasses many parameters, such as mutation rate and gene-conversion rate, on a population level. To delineate the exact conditions under which gene conversion from a pseudogene to a functional gene can lead to apparent selection in a pseudogene, we would require an extensive population genetics model, which is beyond the scope of our work here. However, it is clear that, for apparent selection to show such a strong pattern of selection (dn:ds for the pseudogene γF-crystallin was very similar to dn:ds for other functional *CRYG* genes and substantially deviates from a neutral expectation of $dn:ds = 1$; see table 1), the rate of gene conversion should be at least on the order of the rate of emergence of the potentially deleterious substitutions in the pseudogene.

Since we observe apparent selection in the pseudogene, it is reasonable to hypothesize that some mutations that may cause cataracts in humans may originate in functional genes from gene-conversion events from either the *CRYGEP1* or the *CRYGFP1* pseudogene. Indeed, the activation of the *CRYGEP1* pseudogene may lead to Coppock-like cataracts,[16] and gene conversion leading to genetic disorders has been observed for several diseases,[36] including cataract formation through gene conversion in the β-crystallins.[37] The observation that the *CRYGD* gene has the most cataract-causing mutations described in humans (including P23T) of all functional *CRYG* genes[3] may be

explained either by a higher level of expression of the *CRYGD* gene compared with the other *CRYG* genes[16] or by a higher rate of gene conversion of *CRYGD* with pseudogenes, which may harbor such mutations.

We checked the sequence of mutations known to cause cataracts in humans against those of the two pseudogenes (fig. 3). Of the seven surveyed mutations, we found one that corresponded to a state found in a pseudogene—P23T in *CRYGD*, which is the most common and likely independently derived mutation underlying clinical heterogeneity for different forms of cataracts. It is unclear whether this or other mutations also originate through gene conversion. Nevertheless, our evolutionary analysis suggested that negative selection of γ-crystallin pseudogenes is likely driven by gene conversion of the pseudogenes with functional genes that may result in cataractogenic γ-crystallin alleles.

We describe a human polymorphic congenital cataract caused by a mutation that reversed an amino acid in the *CRYGD* gene to an ancestral state found in nonprimate mammals. This cataract-associated mutation may be compensated for by indirect mechanisms related to the overall protein solubility, through substitutions in a symmetric protein domain. In addition, we found gene-conversion events in the γ-crystallin gene cluster in several mammalian species that involve the interaction of pseudogenes and functional genes in the primate lineage. The observed negative selection in the pseudogene in the course of human-macaque divergence is likely to be the result of apparent selection due to frequent gene conversions between the pseudogenes and the functional genes. The data suggest that some cataractogenic mutations might appear in functional γ-crystallin genes from pseudogenes through gene-conversion events contributing to conservation of the pseudogene sequence.

Acknowledgments

This study was supported by the Biodiversity and Dynamics of Gene Pools program of the Presidium of the Russian Academy of Sciences (support to E.I.R.). E.I.R. is also supported in part by the National Institute of Diabetes and Digestive and Kidney Diseases and National Institute of Neurological Disorders and Stroke (National Institutes of Health), and F.A.K. is supported by a National Science Foundation graduate research fellowship.

Web Resources

Accession numbers and URLs for data presented herein are as follows:

GenBank, http://www.ncbi.nlm.nih.gov/Genbank/ (for sequence information on *CRYGA* [accession numbers M17315 and M17316], *CRYGB* [accession number M19364], *CRYGC* [accession numbers K03003 and K03004], *CRYGD* [accession numbers K03005 and K03006], *CRYGEP1* [accession numbers K03007 and K03008], and *CRYGFP1* [accession numbers K03009 and K03010]; *Pan paniscus CRYGA, CRYGB, CRYGC, CRYGD,* and *CRYGFP1* [accession numbers EF467187, EF467196, EF467205, EF467214, and EF492219, respectively]; *Pan troglodytes CRYGA, CRYGB, CRYGC,* and *CRYGD* [accession numbers EF467190, EF467199, EF467208, and EF467217, respectively], *G. gorilla CRYGA, CRYGB, CRYGC, CRYGD,* and *CRYGFP1* [accession numbers EF467183, EF467192, EF467201, EF467210, and EF492217, respectively]; *Pongo pygmaeus CRYGA, CRYGB, CRYGC, CRYGD,* and *CRYGFP1* [accession numbers EF467188, EF467197, EF467206, EF467215, and EF492220, respectively]; *H. lar CRYGA, CRYGB, CRYGC, CRYGD,* and *CRYGFP1* [accession numbers EF467184, EF467193, EF467202, EF467211, and EF492218, respectively]; *Macaca mulatta CRYGA, CRYGB, CRYGC,* and *CRYGD* [accession numbers EF467186, EF467195, EF467204, and EF467213, respectively]; *L. lagotricha CRYGA, CRYGB, CRYGC,* and *CRYGD* [accession numbers EF467185, EF467194, EF467203, and EF467212, respectively]; *A. geoffroyi CRYGA, CRYGB, CRYGC,* and *CRYGD* [accession numbers EF467182, EF467191, EF467200, and EF467209, respectively]; and *S. labiatus CRYGA, CRYGB, CRYGC,* and *CRYGD* [accession numbers EF467189, EF467198, EF467207, and EF467216, respectively])

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for PCC and other genetic forms of cataracts)

Primer3, http://frodo.wi.mit.edu/

Protein Data Bank, http://www.pdb.org/ (for the human CRYGD protein [ID 1h4a])

University of California Santa Cruz (UCSC) Genome Browser, http://genome.ucsc.edu/

References

1. Francis PJ, Berry V, Moore AT, Bhattacharya S (1999) Lens biology: development and human cataractogenesis. Trends Genet 15:191–196
2. Francis PJ, Berry V, Bhattacharya SS, Moore AT (2000) The genetics of childhood cataract. J Med Genet 37:481–488
3. Reddy MA, Francis PJ, Berry V, Bhattacharya SS, Moore AT (2004) Molecular genetic basis of inherited cataract and associated phenotypes. Surv Ophthalmol 49:300–315
4. SanGiovanni JP, Chew EY, Reed GF, Remaley NA, Bateman JB, Sugimoto TA, Klebanoff MA (2002) Infantile cataract in the collaborative perinatal project: prevalence and risk factors. Arch Ophthalmol 120:1559–1565
5. Blundell T, Lindley P, Miller L, Moss D, Slingsby C, Tickle I, Turnell B, Wistow G (1981) The molecular structure and stability of the eye lens: x-ray analysis of γ-crystallin II. Nature 289:771–777
6. Graw J (1997) The crystallins: genes, proteins and diseases. Biol Chem 378:1331–1348
7. Jamieson RV, Perveen R, Kerr B, Carette M, Yardley J, Heon E, Wirth MG, van Heyningen V, Donnai D, Munier F, et al (2002) Domain disruption and mutation of the bZIP transcription factor, MAF, associated with cataract, ocular anterior segment dysgenesis and coloboma. Hum Mol Genet 11:33–42
8. Semina EV, Ferrell RE, Mintz-Hittner HA, Bitoun P, Alward WL, Reiter RS, Funkhauser C, ack-Hirsch S, Murray JC (1998) A novel homeobox gene *PITX3* is mutated in families with autosomal-dominant cataracts and ASMD. Nat Genet 19:167–170
9. Bu L, Jin Y, Shi Y, Chu R, Ban A, Eiberg H, Andres L, Jiang H,

Zheng G, Qian M, et al (2002) Mutant DNA-binding domain of HSF4 is associated with autosomal dominant lamellar and Marner cataract. Nat Genet 31:276–278
10. Conley YP, Erturk D, Keverline A, Mah TS, Keravala A, Barnes LR, Bruchis A, Hess JF, FitzGerald PG, Weeks DE, et al (2000) A juvenile-onset, progressive cataract locus on chromosome 3q21-q22 is associated with a missense mutation in the beaded filament structural protein–2. Am J Hum Genet 66: 1426–1431
11. Bassnett S, Missey H, Vucemilo I (1999) Molecular architecture of the lens fiber cell basal membrane complex. J Cell Sci 112:2155–2165
12. Berry V, Francis P, Kaushal S, Moore A, Bhattacharya S (2000) Missense mutations in MIP underlie autosomal dominant "polymorphic" and lamellar cataracts linked to 12q. Nat Genet 25:15–17
13. Hejtmancik JF, Smaoui N (2003) Molecular genetics of cataract. Dev Ophthalmol 37:67–82
14. Meakin SO, Breitman ML, Tsui LC (1985) Structural and evolutionary relationships among five members of the human γ-crystallin gene family. Mol Cell Biol 5:1408–1414
15. Meakin SO, Du RP, Tsui LC, Breitman ML (1987) γ-Crystallins of the human eye lens: expression analysis of five members of the gene family. Mol Cell Biol 7:2671–2679
16. Brakenhoff RH, Henskens HA, van Rossum MW, Lubsen NH, Schoenmakers JG (1994) Activation of the γ E-crystallin pseudogene in the human hereditary Coppock-like cataract. Hum Mol Genet 3:279–283
17. Pande A, Pande J, Asherie N, Lomakin A, Ogun O, King JA, Lubsen NH, Walton D, Benedek GB (2000) Molecular basis of a progressive juvenile-onset hereditary cataract. Proc Natl Acad Sci USA 97:1993–1998
18. Brakenhoff RH, Aarts HJ, Reek FH, Lubsen NH, Schoenmakers JG (1990) Human γ-crystallin genes: a gene family on its way to extinction. J Mol Biol 216:519–532
19. Rogaev EI, Rogaeva EA, Korovaitseva GI, Farrer LA, Petrin AN, Keryanov SA, Turaeva S, Chumakov I, St George-Hyslop P, Ginter EK (1996) Linkage of polymorphic congenital cataract to the γ-crystallin gene locus on human chromosome 2q33–35. Hum Mol Genet 5:699–703
20. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2006) GenBank. Nucleic Acids Res 34:D16–D20
21. Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A, et al (2006) The UCSC Genome Browser database: update 2007. Nucleic Acids Res 34:D590–D598
22. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797
23. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574
24. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555–556
25. Ginter EK, Turaeva S, Revazov AA, Panteleeva OA, Artykov OA (1983) Medical genetic study of the population of Turkmenia. III. Hereditary pathology in Turkmen Nokhurlis. Genetika 19:1344–1352
26. Ginter EK, Petrin AN, Spitsyn VA, Rogaev EI (1991) An attempt to locate the gene for congenital cataracts using linkage analysis. Genetika 27:1840–1849
27. Kondrashov AS, Sunyaev S, Kondrashov FA (2002) Dobzhansky-Muller incompatibilities in protein evolution. Proc Natl Acad Sci USA 99:14878–14883
28. Kern AD, Kondrashov FA (2004) Mechanisms and convergence of compensatory evolution in mammalian mitochondrial tRNAs. Nat Genet 36:1207–1212
29. Fukami-Kobayashi K, Schreiber DR, Benner SA (2002) Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. J Mol Biol 319: 729–743
30. Poon A, Chao L (2005) The rate of compensatory mutation in the DNA bacteriophage phiX174. Genetics 170:989–999
31. Basak A, Bateman O, Slingsby C, Pande A, Asherie N, Ogun O, Benedek GB, Pande J (2003) High-resolution X-ray crystal structures of human γD crystallin (1.25 A) and the R58H mutant (1.15 A) associated with aculeiform cataract. J Mol Biol 328:1137–1147
32. Evans P, Wyatt K, Wistow GJ, Bateman OA, Wallace BA, Slingsby C (2004) The P23T cataract mutation causes loss of solubility of folded γD-crystallin. J Mol Biol 343:435–444
33. Pande A, Annunziata O, Asherie N, Ogun O, Benedek GB, Pande J (2005) Decrease in protein solubility and cataract formation caused by the Pro23 to Thr mutation in human γ D-crystallin. Biochemistry 44:2491–2500
34. Aarts HJ, den Dunnen JT, Leunissen J, Lubsen NH, Schoenmakers JG (1988) The γ-crystallin gene families: sequence and evolutionary patterns. J Mol Evol 27:163–172
35. den Dunnen JT, Moormann RJ, Lubsen NH, Schoenmakers JG (1986) Concerted and divergent evolution within the rat γ-crystallin gene family. J Mol Biol 189:37–46
36. Bischof JM, Chiang AP, Scheetz TE, Stone EM, Casavant TL, Sheffield VC, Braun TA (2006) Genome-wide identification of pseudogenes capable of disease-causing gene conversion. Hum Mutat 27:545–552
37. Vanita, Sarhadi V, Reis A, Jung M, Singh D, Sperling K, Singh IR, Burger I (2001) A unique form of autosomal dominant cataract explained by gene conversion between β-crystallin B2 and its pseudogene. J Med Genet 38:392–396
38. Li WH (1997) Molecular evolution. Sinauer, Sunderland, MA
39. Teshima KM, Innan H (2004) The effect of gene conversion on the divergence between duplicated genes. Genetics 166: 1553–1560
40. Kondrashov FA, Gurbich TA, Vlasov PV (2007) Selection for functional uniformity of *tuf* duplicates in γ-proteobacteria. Trends Genet 23:215–218
41. Barton NH (1990) Pleiotropic models of quantitative variation. Genetics 124:773–782
42. Perelygin AA, Kondrashov FA, Rogozin IB, Brinton MA (2002) Evolution of the mouse polyubiquitin-C gene. J Mol Evol 55: 202–210
43. Stephan DA, Gillanders E, Vanderveen D, Freas-Lutz D, Wistow G, Baxevanis AD, Robbins CM, VanAuken A, Quesenberry MI, Bailey-Wilson J, et al (1999) Progressive juvenile-onset punctate cataracts caused by mutation of the γD-crystallin gene. Proc Natl Acad Sci USA 96:1008–1012
44. Kmoch S, Brynda J, Asfaw B, Bezouska K, Novak P, Rezacova P, Ondrova L, Filipec M, Sedlacek J, Elleder M (2000) Link between a novel human γD-crystallin allele and a unique cataract phenotype explained by protein crystallography. Hum Mol Genet 9:1779–1786
45. Gu J, Qi Y, Wang L, Wang J, Shi L, Lin H, Li X, Su H, Huang

S (2005) A new congenital nuclear cataract caused by a missense mutation in the γD-crystallin gene (*CRYGD*) in a Chinese family. Mol Vis 11:971–976

46. Héon E, Priston M, Schorderet DF, Billingsley GD, Girard PO, Lubsen N, Munier FL (1999) The γ-crystallins and human cataracts: a puzzle made clearer. Am J Hum Genet 65:1261–1267
47. Santhiya ST, Shyam MM, Rawlley D, Vijayalakshmi P, Namperumalsamy P, Gopinath PM, Loster J, Graw J (2002) Novel mutations in the γ-crystallin genes cause autosomal dominant congenital cataracts. J Med Genet 39:352–358
48. Messina-Baas OM, Gonzalez-Huerta LM, Cuevas-Covarrubias SA (2006) Two affected siblings with nuclear cataract associated with a novel missense mutation in the *CRYGD* gene. Mol Vis 12:995–1000
49. Mackay DS, Andley UP, Shiels A (2004) A missense mutation in the γD crystallin gene (*CRYGD*) associated with autosomal dominant "coral-like" cataract linked to chromosome 2q. Mol Vis 10:155–162
50. Xu WZ, Zheng S, Xu SJ, Huang W, Yao K, Zhang SZ (2004) Autosomal dominant coralliform cataract related to a missense mutation of the γD-crystallin gene. Chin Med J 117: 727–732
51. Nandrot E, Slingsby C, Basak A, Cherif-Chefchaouni M, Benazzouz B, Hajaji Y, Boutayeb S, Gribouval O, Arbogast L, Berraho A, et al (2003) Gamma-D crystallin gene (CRYGD) mutation causes autosomal dominant congenital cerulean cataracts. J Med Genet 40:262–267
52. Shentu X, Yao K, Xu W, Zheng S, Hu S, Gong X (2004) Special fasciculiform cataract caused by a mutation in the γD-crystallin gene. Mol Vis 10:233–239
53. Burdon KP, Wirth MG, Mackey DA, Russell-Eggitt IM, Craig JE, Elder JE, Dickinson JL, Sale MM (2004) Investigation of crystallin genes in familial cataract, and report of two disease associated mutations. Br J Ophthalmol 88:79–83

Chapter 3, in full, is a reprint of the material as it appears in Plotnikova OV, Kondrashov FA, Vlasov PK, Ginter EK and Rogaev EI (2007). Conversion and compensatory evolution of the γ-crystallin genes and identification of a cataractogenic mutation that reverses the sequence of the human CRYGD gene to an ancestral state. Am. J. Hum. Genet. **81**, 32-43. The American Society of Human Genetics 2007. The dissertation author was one of the two primary investigators and authors of this paper.

Part II.

Revealing function and selection in genes and genomes

Chapter 4.

Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites

Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Theoretical Biology 240 (2006) 616–626
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites

Fyodor A. Kondrashov[a], Aleksey Y. Ogurtsov[b], Alexey S. Kondrashov[b,*]

[a]Section of Ecology, Behavior and Evolution, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0346, USA
[b]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20894, USA

Received 20 June 2005; received in revised form 26 October 2005; accepted 27 October 2005
Available online 15 December 2005

Abstract

The impact of synonymous nucleotide substitutions on fitness in mammals remains controversial. Despite some indications of selective constraint, synonymous sites are often assumed to be neutral, and the rate of their evolution is used as a proxy for mutation rate. We subdivide all sites into four classes in terms of the mutable CpG context, nonCpG, postC, preG, and postCpreG, and compare four-fold synonymous sites and intron sites residing outside transposable elements. The distribution of the rate of evolution across all synonymous sites is trimodal. Rate of evolution at nonCpG synonymous sites, not preceded by C and not followed by G, is ~10% below that at such intron sites. In contrast, rate of evolution at postCpreG synonymous sites is ~30% above that at such intron sites. Finally, synonymous and intron postC and preG sites evolve at similar rates. The relationship between the levels of polymorphism at the corresponding synonymous and intron sites is very similar to that between their rates of evolution. Within every class, synonymous sites are occupied by G or C much more often than intron sites, whose nucleotide composition is consistent with neutral mutation–drift equilibrium. These patterns suggest that synonymous sites are under weak selection in favor of G and C, with the average coefficient $s \sim 0.25/N_e \sim 10^{-5}$, where N_e is the effective population size. Such selection decelerates evolution and reduces variability at sites with symmetric mutation, but has the opposite effects at sites where the favored nucleotides are more mutable. The amino-acid composition of proteins dictates that many synonymous sites are CpGprone, which causes them, on average, to evolve faster and to be more polymorphic than intron sites. An average genotype carries ~10^7 suboptimal nucleotides at synonymous sites, implying synergistic epistasis in selection against them.
Published by Elsevier Ltd.

Keywords: Mutation; Selection; Synonymous site; Evolution; Genetic drift

1. Introduction

Throughout all life, synonymous codons are used non-randomly (Grantham et al., 1980; see Li, 1997, Chapter 7, for review). There is a general agreement that selection plays a major role in this phenomenon (Andersson and Kurland, 1990; McVean and Vieira, 2001; Duret, 2002; Carlini and Stephan, 2003; Nielsen and Akashi, 2003). Synonymous substitutions affect mRNA translation (Ikemura, 1985; Sorensen et al., 1989; Sharp et al., 1995; Akashi, 1995, 1999a, b, 2003) and thus can cause translational selection which influences codon usage in many, although perhaps not in all (Kanaya et al., 1999), organisms. Synonymous substitutions also affect important properties of mRNAs which are "not directly related to the codon–anticodon interaction" (Duan and Antezana, 2003), in particular, their secondary structures (Hartl et al., 1994; Innan and Stephan, 2001; Duan et al., 2003; Katz and Burge, 2003; Chamary and Hurst, 2005a).

However, the importance of selection at synonymous sites in mammals remains unclear. Although their codon usage is obviously non-random, due to elevated frequencies of G and C at synonymous sites (Debry and Marzluff, 1994; Eyre-Walker, 1999), the causes of this pattern are controversial. Some authors argue for an important role of selection (Debry and Marzluff, 1994; Eyre-Walker, 1999; Keightley and Gaffney, 2003; Urrutia and Hurst, 2003;

*Corresponding author. Tel.: 1 301 435 8944; fax: 1 301 480 2288.
E-mail addresses: fkondras@ncbi.nlm.nih.gov (F.A. Kondrashov), ogurtsov@ncbi.nlm.nih.gov (A.Y. Ogurtsov), kondrashov@ncbi.nlm.nih.gov (A.S. Kondrashov).

0022-5193/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.jtbi.2005.10.020

Nielsen and Akashi, 2003; Chamary and Hurst, 2004; Comeron, 2004; Lu and Wu, 2005; Chamary and Hurst, 2005a), but others disagree (e. g., Sharp et al., 1995; Smith and Hurst, 1999; Duret and Hurst, 2001; Urrutia and Hurst, 2001; Duret, 2002; Subramanian and Kumar, 2003) and favor alternative explanations, such as biased mutation (Wolfe et al., 1989) or biased gene conversion (Duret, 2002).

The arguments for or against selection at synonymous sites in mammals are undermined by conflicting data on whether evolution at four-fold synonymous sites is slower than at presumably neutral intron or pseudogene sites, which is often thought to be the obligatory signature of any selection at synonymous sites. Hughes and Yager (1997) and Chamary and Hurst (2004) reported similar levels of rat–mouse divergence at synonymous and intron sites, Bustamante et al. (2002) found that synonymous sites evolve more slowly than homologous pseudogene sites, Subramanian and Kumar (2003) reported that in primates synonymous sites evolve faster than intron sites, and Hellman et al. (2003) reached the opposite conclusion.

Because effective neutrality of synonymous substitutions in mammals is widely accepted, mutation rates are routinely estimated through rates of synonymous substitution in mammalian evolution (Smith and Hurst, 1999; Keightley and Eyre-Walker, 2000; Kumar and Subramanian, 2002) and patterns in synonymous substitutions are generalized to the whole genome (Duret et al., 2002). Similarly, levels of intrapopulation variability and rates of interspecies divergence at synonymous sites have been accepted as the neutral point of reference in tests for positive selection (e.g. Fay et al., 2001).

We study selection at synonymous sites through patterns in human–chimpanzee divergence and in intrahuman polymorphism. Using such a close pair of species guarantees against errors caused by multiple substitutions at the same site (Li, 1997) and by ambiguous alignments of introns. Similar to several previous analyses, ours takes into account elevated mutability in mammals of the CpG context, i.e. of nucleotides within 5′CG3′ segments on the DNA sequence (see Li, 1997; Nachman and Crowell, 2000). However, the commonly used classification of sites into those residing and not residing within a CpG context (e.g. Hellman et al., 2003) may obscure the patterns in divergence, since substitutions at a site can affect its placement within this classification (Keightley and Gaffney, 2003).

Thus, we subdivide all sites into four non-overlapping classes: those not preceded by C and not followed by G (nonCpG, Keightley and Gaffney 2003), preceded by C but not followed by G (postC), followed by G but not preceded by C (preG), and preceded by C and followed by G (postCpreG). Sites from the last three classes are CpGprone, as they can reside within CpG context. This approach makes it possible to disentangle the impacts of mutation and selection and to show that weak selection in favor of G and C is a major factor of evolution of synonymous sites in mammals.

2. Materials and methods

2.1. Data

We obtained the human–chimpanzee (hg17-panTro1) alignments and annotation from the Genome Center at U.C. Santa Cruz (Karolchik et al., 2003). Transposable element (TE)-derived intron sites are those masked by RepeatMasker in these alignments. The first 40 and the last 40 nucleotides of an intron, as well as all sites preceded and/or followed by a human–chimpanzee mismatch were excluded from the analysis. Expression level was assayed by the number of ESTs. Mapped polymorphisms were taken from the U.C. Santa Cruz Genome Center annotation of dbSNP release 123 to assembly hg17 of the human genome. For our analyses we used polymorphisms that were obtained in genome-wide, non-exon targeted assays. Only SNPs with the following Submitter Handles in dbSNP flatfiles were used: CSHL-HAPMAP, BCM_SSAHASNP, SC_JCM, SSAHASNP, WI_SSAHASNP, TSC-CSHL, WUGSC_SSAHASNP, SC_SNP, SC. The data used are located as follows.

Human–chimpanzee alignments:

http://hgdownload.cse.ucsc.edu/goldenPath/hg17/vsPanTro1/axtNet/

Human genome annotation:
http://hgdownload.cse.ucsc.edu/goldenPath/hg17/database/knownGene.txt.gz

Human polymorphisms mapping to the human genome:
http://hgdownload.cse.ucsc.edu/goldenPath/hg17/database/snp.txt.gz

Human polymorphism annotation in dbSNP flatfiles:
ftp://ftp.ncbi.nlm.nih.gov/snp/human/ASN1_flat/

2.2. Review of theory

Consider stochastic mutation–selection–drift equilibrium at a locus (site) with four alleles: A, T, G, and C. Assuming that all mutation rates are low (well below N_e^{-1}, where N_e is the effective population size), a population (approximately) is fixed with one of the alleles most of the time, and occasionally undergoes switches between fixations of different alleles. The frequency of the ith allele, p_i, is the fraction of time when it is fixed. When the population is fixed for the ith allele, the flux of switches to fixation of the jth allele (the per generation probability of a switch), $f_{i>j}$, is the corresponding mutation rate $\mu_{i>j}$, times the population size N, times the probability $g_{i>j}$ that a mutant carrying the jth allele which appeared in a population fixed with the ith allele will reach fixation. The formula for $g_{i>j}$ can be found in Bulmer (1991, Eq. (7)). Equilibrium allele frequencies p_i^{EQ} can be obtained by solving the system of

linear equations which describes the equality of the total rates of switches from and to fixations of each allele (Bulmer, 1991, Eq. (10)):

$$\begin{aligned}
p_A(f_{A>T}+f_{A>G}+f_{A>C}) &= p_T f_{T>A}+p_G f_{G>A}+p_C f_{C>A},\\
p_T(f_{T>A}+f_{T>G}+f_{T>C}) &= p_A f_{A>T}+p_G f_{G>T}+p_C f_{C>T},\\
p_G(f_{G>A}+f_{G>T}+f_{G>C}) &= p_A f_{T>G}+p_T f_{T>G}+p_C f_{C>G},\\
p_C(f_{C>A}+f_{C>T}+f_{C>G}) &= p_A f_{A>C}+p_T f_{T>C}+p_G f_{G>C}.
\end{aligned} \tag{1}$$

The total rate of evolution (per generation probability of a switch between some allele fixations) at equilibrium is

$$R = p_A^{EQ}(f_{A>T}+f_{A>G}+f_{A>C})+\cdots+p_C^{EQ}(f_{C>A}+f_{C>T}+f_{C>G}) \tag{2}$$

and the total heterozygosity at equilibrium is

$$P = p_A^{EQ}N(\mu_{A>T}H_{A>T}+\mu_{A>G}H_{A>G}+\mu_{A>C}H_{A>C})+\cdots, \tag{3}$$

where $H_{i>j}$ is the expected contribution to heterozygosity by a mutant carrying the jth allele which appeared in a population where the ith allele is fixed (see McVean and Charlesworth, 1999, Eq. (10)).

3. Results

3.1. Intron sites: data

Table 1 presents data on frequencies of the four nucleotides, rate of evolution R (assayed through human–chimpanzee divergence, i.e. the fraction of mismatches in the alignments), and the level of intrahuman polymorphism P (assayed through the density of SNPs) at four-fold synonymous sites and intron sites within 13 533 loci that contain 53 792 introns. First, let us consider introns.

At nonCpG sites, frequencies of G and C are only slightly below 25%. In contrast, postC sites are strongly depleted of G, preG sites are strongly depleted of C, and postCpreG sites are depleted of both G and C (this difference is highly statistically significant, as well as all the differences mentioned below). Of course, this is just another way of saying that introns are depleted of CpG contexts (Bird, 1980). R and P are the lowest at nonCpG sites, and the highest at postCpreG sites. At intron sites of TE origin, frequencies of G and C, as well as P and R, are higher than at nonTE intron sites from the corresponding classes.

Fig. 1 presents data on polarized polymorphisms, those where the ancestral allele is G or C and the derived allele is

Table 1
Properties of sites classified according to their possible CpG context

	All	nonCpG	postC	preG	postCpreG
(a) *Intron sites outside transposable elements*					
Number	52954891	33887011 (64%)	7978950 (15%)	8673858 (16%)	2415072 (5%)
Frequencies					
A	0.2827	0.2606	0.3007	0.3104	0.4331
T	0.3111	0.2883	0.3373	0.3347	0.4588
G	0.2097	0.2330	0.0423	0.3161	0.0538
C	0.1965	0.2180	0.3197	0.0387	0.0543
Divergence	0.01064	0.00932	0.01319	0.01178	0.01663
Polymorphism	0.001294	0.001165	0.001573	0.001416	0.001767
(b) *Intron sites inside transposable elements*					
Number	32287369	19894875 (62%)	5275940 (16%)	5401718 (17%)	1714836 (5%)
Frequencies					
A	0.2702	0.2414	0.3051	0.2924	0.4272
T	0.2923	0.2682	0.3039	0.3236	0.4367
G	0.2207	0.2479	0.0570	0.3290	0.0670
C	0.2169	0.2425	0.3340	0.0551	0.0690
Divergence	0.01274	0.01056	0.01568	0.01468	0.02280
Polymorphism	0.001574	0.001409	0.001815	0.001699	0.002351
(c) *Four-fold synonymous sites*					
Number	1949372	682032 (35%)	654300 (34%)	293573 (15%)	319467 (16%)
Frequencies					
A	0.2165	0.1478	0.2262	0.2039	0.3550
T	0.2320	0.1597	0.2399	0.2169	0.3840
G	0.2425	0.3479	0.0859	0.4751	0.1245
C	0.3090	0.3446	0.4480	0.1042	0.1365
Divergence	0.01282	0.00831	0.01351	0.01182	0.02195
Polymorphism	0.001441	0.001051	0.001529	0.001251	0.002267

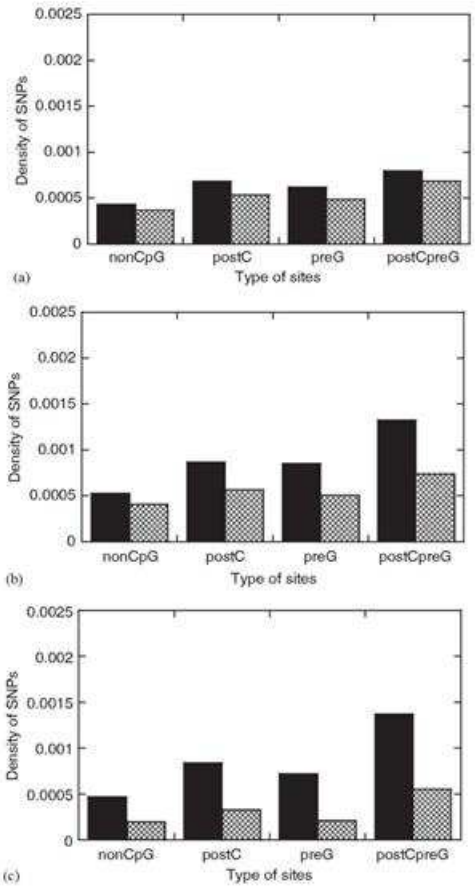


Fig. 1. Densities of reciprocal polymorphisms, GC > AT (black bars) vs. AT > GC (gray bars), at intron nonTE sites (a), intron TE sites (b), and four-fold synonymous sites (c).

A or T (GC > AT, Lercher et al., 2002b), and the reciprocal (AT > GC) (currently, the lack of a close enough outgroup for *Homo* and *Pan* genomes make it impossible to obtain the analogous data on polarized substitutions). At nonTE intron sites, there is only a small excess of GC > AT polymorphisms over AT > GC polymorphisms. In contrast, at TE intron sites this excess is much larger, especially at CpGprone sites.

Fig. 2 presents data on human–chimpanzee divergence at orthologous intron sites located within TEs from different families. On average, TEs which were inserted more recently evolve much faster.

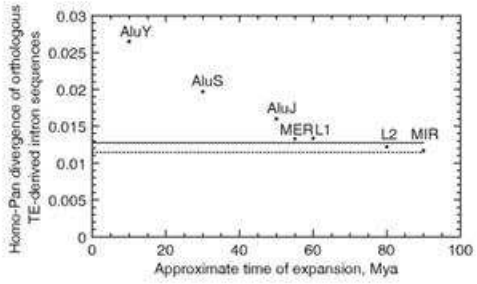


Fig. 2. The dependence of human–chimpanzee divergence between orthologous within-intron copies of transposable elements from different families and subfamilies on the approximate ages of their expansions (Kapitonov and Jurka, 1996; International Human Genome Sequencing Consortium, 2001). The solid line shows the average human–chimpanzee divergence of all intron sequences identified as TEs by RepeatMasker, and the dotted line shows the divergence of intron sequences which remain unmasked.

3.2. *Intron sites: equilibrium with asymmetric mutation at nonTE sites*

The patterns observed at intron sites are consistent with their selective neutrality. NonTE sites appear to be close to mutation–drift equilibrium, while sites of TE origin are losing G's and C's at CpGprone sites.

The ratio of $\mu_{trv\text{-}nonCpG}$, $\mu_{tri\text{-}nonCpG}$, $\mu_{trv\text{-}CpG}$, and $\mu_{tri\text{-}CpG}$, the rates of transversions (of each of the two possible ones) outside CpG, transitions outside CpG, transversions within CpG, and transitions within CpG, is ~1:3:5:30 among mammalian nucleotide substitutions (Nachman and Crowell, 2000; Ebersberger et al., 2002; Kondrashov, 2003; our data; $\mu_{trv\text{-}nonCpG} \sim 0.4 \times 10^{-8}$). Thus, at mutation–drift equilibrium, postC (preG) sites must be depleted of G (C), and postCpreG sites must be depleted of both G and C. CpGprone sites also must evolve faster and be more polymorphic than nonCpG sites.

Indeed, at CpGprone sites some mutation rates are the same as at nonCpG cites but other mutation rates are higher. At a selectively neutral site (locus) with only two alleles, B_1 and B_2, the equilibrium frequency of B_1 (i.e. the probability that B_1 is fixed at a random moment) is $v/(u+v)$ (e.g. Sueoka, 1962), and R, defined as the per generation frequency of switches between fixations of B_1 and of B_2, is $2uv/(u+v)$, where u and v are rates of mutation from B_1 to B_2 and back (e.g. Bulmer, 1991, Eqs. (6) and (7)). Thus, R doubles when v increases from u to infinity. P, defined as heterozygosity, is $4N_e uv/(u+v)$ (McVean and Charlesworth, 1999, Eq. (10)). Without selection, P always (with any number of alleles) changes with the mutation rates in exactly the same way as does R.

This analysis can be extended to mutation–drift equilibrium at a site with four nucleotides (alleles), A, T, G, and C (Bulmer, 1991, Eq. (10), see Methods). Fig. 3 shows how

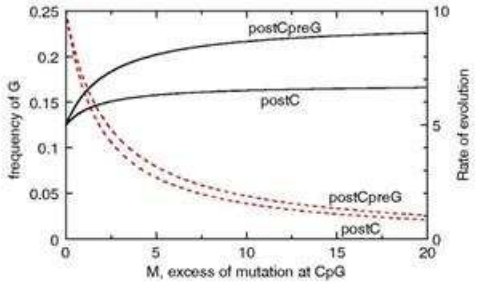


Fig. 3. Frequencies of G (broken red lines) and rates of evolution R in the units of $\mu_{trv\text{-}nonCpG}$ (solid black lines) at the mutation–drift equilibrium without selection as functions of M, where $\mu_{trv\text{-}nonCpG} = 1$, $\mu_{tsi\text{-}nonCpG} = 3$, $\mu_{trv\text{-}CpG} = (1 + 4M/9)\,\mu_{trv\text{-}nonCpG}$, and $\mu_{tsi\text{-}CpG} = (1 + M)\,\mu_{tsi\text{-}nonCpG}$. PostC sites (shown) and preG (not shown) sites evolve at identical rates. Frequency of C at preG sites is the same as frequency of G at postC sites, and at postCpreG sites frequencies of G and C are identical. Properties of nonCpG sites correspond to $M = 0$.

the predicted parameters of such sites depend on M, the relative excess of transitions within the CpG context. For intron nonTE sites, the frequencies of G observed at postCpreG and postC sites, 5.4% and 4.2% (Table 1), imply $M = 8.4$ and $M = 9.0$, respectively (Fig. 3), in excellent agreement with $M \sim 9.0$ which follows from direct data on mutation rates (Kondrashov, 2003). Excesses of R (of P) at postCpreG or at postC sites over the corresponding parameters of nonCpG sites are (Table 1) 1.784 (1.517) or 1.415 (1.350) and imply $M = 14.5$, $M = 3.2$, $M \to \infty$, or $M = 20$, respectively (Fig. 3). However, since under selective neutrality R and P are essentially independent of M when $M > 5$ (Fig. 3), frequencies of mutable nucleotides are more suitable for indirect estimates of high values of M.

Similarity of the levels of the reciprocal polymorphisms GC > AT and AT > GC suggest that nonTE intron sites are close to mutation–drift equilibrium without selection (Eyre-Walker, 1997; Smith and Eyre-Walker, 2001), although, on average, these sites are slowly losing G and C (Fig. 1a; Lercher and Hurst, 2002). This conclusion is supported by direct data on their evolution, obtained for a small fraction of *Homo-Pan* genome alignments for which a suitable outgroup is available (Webster et al., 2003).

3.3. Intron sites: loss of CpG context in transposable elements

Intron sites of TE origin deviate substantially from mutation–drift equilibrium and rapidly lose G and C at CpGprone sites (Fig. 1b). At the moment of insertion, many TEs have a higher (and not a lower, Duret and Hurst, 2001) GC-content and a higher proportion of mutable CpG contexts than nonTE intron sites (Chen et al., 2001). It takes almost 100 Myr for a TE-derived intron segment to reach mutation–drift equilibrium (Fig. 2) and, before this happens, the segment remains more GC-rich and CpG-rich and, thus, evolves faster and is more polymorphic than at equilibrium.

However, even within the nonCpG class, R and P at TE intron sites are ~15% higher than at nonTE intron sites, although for CpGprone classes these excesses are higher (Table 1). This pattern can be caused by slow dissolution of other (different from CpG) mutable contexts within TE-derived intron segments (Hwang and Green, 2004) and/or by negative selection affecting ~10% of nonTE intron sites (Shabalina et al., 2001). Still, we will use nonTE intron sites as a neutral mutation–drift equilibrium point of reference, since they appear to be much closer to this equilibrium than any other sites.

3.4. Four-fold synonymous sites: data

The average rate of evolution at four-fold synonymous sites is similar to that at intron sites (Hughes and Yager, 1997; but see Smith and Hurst, 1998; Chamary and Hurst, 2004) apparently suggesting the lack of selective constraint. However, this overall similarity is misleading (Chamary and Hurst, 2004) and hides a complex pattern.

Synonymous sites from all the four classes are strongly enriched by G and C, relative to the corresponding intron sites (Table 1). In particular, frequencies of G and C at postCpreG synonymous sites are 2.5 times above those expected at mutation–drift equilibrium with $M = 9$ and observed at postCpreG nonTE sites within introns.

In contrast, there is no uniform relationship between R or P at synonymous and the corresponding nonTE intron sites. NonCpG synonymous sites evolve 10% slower, postC and preG synonymous sites evolve at approximately the same rate, and postCpreG synonymous sites evolve ~30% faster, with the levels of polymorphism displaying a very similar pattern (Table 1). As the result of this diversification of P and R at synonymous sites, the ratios of their values at nonCpG, postC or preG, and PostCpreG synonymous sites are ~1:1.5:2.5.

In contrast to nonTE intron sites, GC > AT polymorphisms at synonymous sites are ~2.5 times more common than AT > GC polymorphisms (Fig. 1c; Smith and Eyre-Walker, 2001).

3.5. Four-fold synonymous sites: selection for G and C

There is no reason to assume that context-dependent mutation rates are different between exons and introns. However, the observed contrasts between nonTE intron sites and four-fold synonymous sites can be readily explained by weak selection. Let us make an oversimplified assumption that uniform, constant selection with the coefficient $s \sim 0.25 N_e^{-1}$ favors G or C over A or T at all

synonymous sites. Intermediate dominance will be assumed, with selective advantage $2s$ to homozygotes for the favored allele. Naturally, such selection will always increase frequencies of G and C. In contrast, R and P will be affected differently at sites from different classes, being reduced at nonCpG sites, but elevated at postCpreG sites.

Indeed, constant selection always reduces R and P if all mutation rates are equal or if less mutable alleles are favored (e.g. Akashi, 1999b, c). However, constant selection favoring more mutable allele(s) may increase R (Eyre-Walker, 1992; Eyre-Walker and Bulmer, 1995; McVean and Charlesworth, Figs. 5c and 6c) and P (McVean and Charlesworth, 1999, Fig. 2). At a site with two alleles, B_1 and B_2, and selection for B_2 with coefficient s, at equilibrium $R = 2Suv/[(1-e^{-S})(u+ve^{S})]$ (Bulmer, 1991, Eq. (6) and (7)) and

$$P = \frac{4N_e uv}{u+ve^{-S}}\left(\frac{e^{-S}}{1+e^{-S}}+\frac{1-e^{-S}}{2S}\right) \quad (4)$$

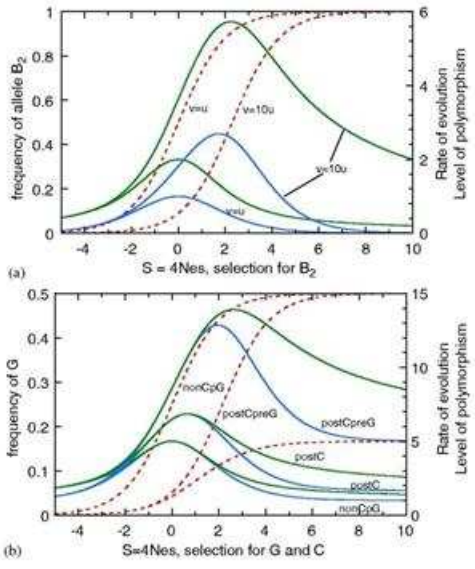


Fig. 4. Equilibrium allele frequencies (broken red lines), rates of evolution in the units of $\mu_{TV\text{-}nonCpG}$ (solid blue lines), and levels of polymorphism in the units of $2N_e\mu_{TV\text{-}nonCpG}$ (solid green lines) as functions of $S = 4N_e s$. (a) Two alleles, B_1 and B_2, under selection with coefficient s in favor of B_2. (b) Four alleles, A, T, G, and C, under selection with coefficient s in favor of G and C, and $M = 9$. PostC sites (shown) and preG sites (not shown) evolve at identical rates. Frequency of C at preG sites is the same as frequency of G at postC sites, and frequencies of G and C are identical at nonCpG and postCpreG sites.

(McVean and Charlesworth 1999, Eq. (15)), where $S = 4N_e s$. If B_2 is more mutable than B_1 ($v > u$), R and P are maximal at some $S > 0$ (Fig. 4a).

The analogous patterns persist in the case of four alleles (Bulmer, 1991, Eq. (10)); McVean and Charlesworth, 1999, Eq. (10)), see Methods). Thus, at nonCpG sites, where mutation is symmetric, R and P are maximal at $S = 0$. In contrast, at CpGprone sites R and P are maximal at a positive S, i.e. under selection favoring more mutable allele G and/or C (Fig. 4b).

These patterns can be used to estimate S roughly. Frequencies of G at nonCpG, postC or preG, and postCpreG synonymous sites imply $S \sim 0.8$, ~ 1.25, and 1.1, respectively (Table 1 and Fig. 4b). The values of R and P at synonymous sites deviate from their values at the corresponding nonTE intron sites by -10%, 0%, and $+30\%$ at nonCpG, postC or preG, and postCpreG sites, respectively, which implies $S \sim 0.9$, $S \sim 1.3$ (or, alternatively, $S = 0$), and $S \sim 0.9$ (or $S > 3.0$), respectively (Fig. 4b). Thus, it appears that $S \sim 1$, so that a typical value of s at a synonymous site is $\sim 0.25 N_e^{-1}$.

The 2.5-fold difference between the levels of reciprocal GC > AT and AT > GC polymorphisms at synonymous sites (Maside et al., 2004) of all classes (Fig. 1c) suggests a higher $s \sim 0.55 N_e^{-1}$ (data not reported). However, the different levels of reciprocal polymorphisms may be to some extent caused by factors other than selective advantage of G and C (Smith and Eyre-Walker, 2001; Lercher et al., 2002a, b), which work even at nonTE intron sites (Fig. 1a). Thus, $s \sim 0.55 N_e^{-1}$ is probably an overestimation.

3.6. Heterogeneity of the observed patterns across genes

The position of a mammalian gene within isochores, genome regions with different GC-contents (see Eyre-Walker and Hurst, 2001) affects the patterns described above. Not surprisingly, genes located within GC-rich genome regions (as assayed by GC-content of their introns) have proportionally more G (Fig. 5a) and C (data not reported) at their synonymous sites. Still, the relationships between rates of evolution at nonTE intron sites and fourfold synonymous sites from the corresponding classes remain the same for genes with all GC-contents, except for those which are very GC-poor, where synonymous sites do not evolve faster than intron sites (Fig. 5b). Perhaps, a factor which favors G and C at synonymous sites is counterbalanced, in genes residing within the most GC-poor genome regions, by another factor responsible for the low regional GC-content. The patterns in P depend on the regional GC-content similarly (data not reported).

In contrast, the nucleotide frequencies (Fig. 6a), R (Fig. 6b), and P (data not reported) depend very little on the expression level of a gene (Duret and Mouchiroud, 2000). A slight increase of the rate of evolution at CpGprone sites with the expression may by due to positive correlation of expression with the GC-content of the gene

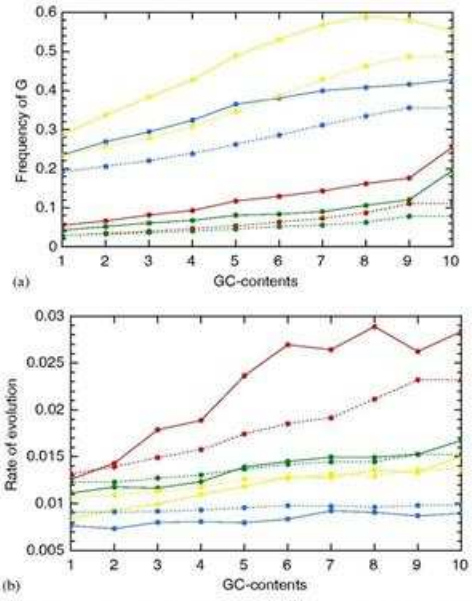


Fig. 5. Frequencies of nucleotide G (a) and the rates of evolution (b) at four-fold synonymous (solid lines) and nonTE intron sites (broken lines) from the four classes (nonCpG—blue, postC—green, preG—yellow, postCpreG—red) in genes split into ten bins of equal sizes according to the GC-content of their introns.

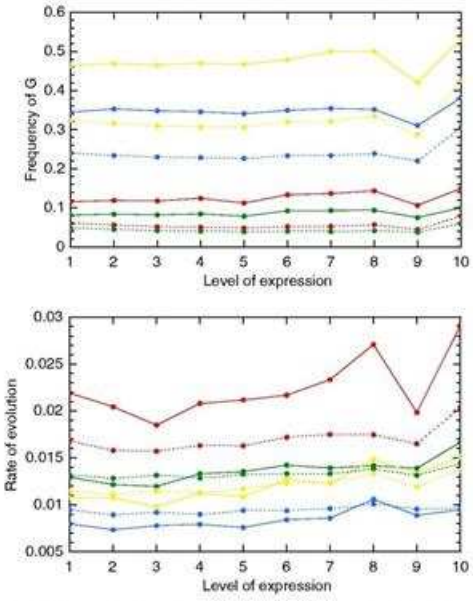


Fig. 6. Frequencies of nucleotide G (a) and the rates of evolution (b) at four-fold synonymous (solid lines) and nonTE intron sites (broken lines) sites from the four classes (nonCpG—blue, postC—green, preG—yellow, postCpreG—red) in genes split into ten bins of equal sizes according to their expression levels.

in mammals (Lercher et al., 2002a; Urrutia and Hurst, 2003).

3.7. Heterogeneity of the observed patterns within genes

The observed patterns are not exactly uniform within genes. Synonymous sites are more GC-enriched within first exons than within last exons of genes. The difference is particularly substantial, 40% vs. 24%, for postCpreG sites. Indeed, first exons are often covered by CpG islands, located in the 5′ ends of genes (Takai and Jones, 2003). However, the rate of evolution of postCpreG sites within first exons is not higher, and even slightly lower than within last exons (data not reported). Perhaps, coefficients of selection in favor of G and C are higher within first exons, and exceed, at some sites, the values which leads to the maximal R. Alternatively, CpG contexts may be less mutable within CpG islands. Different patterns in codon bias at the beginnings vs. the ends of genes have also been observed in bacteria (Hartl et al., 1994). In contrast, there is no difference between GC-contents of first and last introns of genes, although postCpreG sites located close to edges of all introns are more GC-rich and evolve faster than such sites deep inside introns (data not reported).

3.8. More detailed classification of sites

Table 2 presents data on nonTE intron sites and four-fold synonymous sites subdivided into classes according to all 4×4 immediate contexts (the genetic code does not admit postA four-fold synonymous sites). Not surprisingly (e.g. Hess et al., 1994; Hwang and Green, 2004), there is some heterogeneity within sites lumped into nonCpG, postC, or preG inclusive classes of our 2×2 classification. In particular, there is a strong tendency for postApreA sites to be occupied by A more often than by T, and postTpreT sites are occupied by T more often than by A, implying the lack of strand asymmetry in this pattern. Apparently, at postApreA (postTpreT) sites A>T (T>A) substitutions are rarer than T>A (A>T) substitutions.

Still, CpG is by far the most important context, which is not surprising since its impact on the mutation rate is an order of magnitude higher than that of all other contexts (Hwang and Green, 2004). Also, predictions based on the full 4×4 classification of sites are currently impossible, due to lack of data on the impacts of contexts, other than CpG, at the mutation rate in primates.

Table 2
Properties of sites classified according to all 4 × 4 immediate contexts

	postA		postT		postG		postC	
	Intron	4-fold	Intron	4-fold	Intron	4-fold	Intron	4-fold
preA								
A	0.365		0.259	0.118	0.327	0.265	0.332	0.239
T	0.213		0.281	0.130	0.199	0.090	0.274	0.164
G	0.235		0.239	0.405	0.256	0.283	0.039	0.060
C	0.186		0.220	0.346	0.219	0.362	0.354	0.537
Divergence	0.0090		0.0084	0.0073	0.0076	0.0064	0.0111	0.0108
preT								
A	0.280		0.184	0.095	0.234	0.186	0.288	0.207
T	0.307		0.400	0.213	0.283	0.171	0.349	0.275
G	0.217		0.206	0.346	0.220	0.164	0.046	0.073
C	0.196		0.210	0.346	0.263	0.480	0.316	0.446
Divergence	0.0106		0.0104	0.0108	0.097	0.0091	0.0154	0.0156
preG								
A	0.325		0.245	0.125	0.344	0.311	0.431	0.342
T	0.306		0.364	0.208	0.323	0.230	0.450	0.384
G	0.326		0.349	0.600	0.284	0.306	0.059	0.124
C	0.043		0.042	0.068	0.050	0.153	0.060	0.150
Divergence	0.0132		0.0107	0.0108	0.0113	0.0132	0.0166	0.0220
PreC								
A	0.251		0.180	0.093	0.202	0.179	0.298	0.218
T	0.242		0.321	0.206	0.220	0.143	0.363	0.274
G	0.283		0.246	0.446	0.289	0.309	0.052	0.133
C	0.225		0.253	0.255	0.289	0.368	0.288	0.376
Divergence	0.0095		0.0085	0.0085	0.0091	0.0085	0.0122	0.0143

4. Discussion

Elevated frequencies of nucleotides G and C at synonymous sites, as well as complex relationships between the rates of divergence and levels of polymorphism at synonymous sites vs. intron sites suggests that the majority of synonymous sites of human and chimpanzee genes are under weak selection that favors nucleotides G and C. Comparison of the properties of such sites with those of intron sites of nonTE origin (Table 1), which appear to be close to selectively neutral mutation–drift equilibrium (Fig. 1), and to theoretical predictions (Fig. 4b) implies that the average coefficient of selection s in favor of nucleotides G and C at a human synonymous site is $\sim 0.25N_e^{-1}$, with not too much variation across individual sites. The data are clearly inconsistent with strong selection for G and/or C at some synonymous sites and selective neutrality at other sites: elevated frequencies of G and C can be generated in this way, but elevated rates of evolution (Eyre-Walker, 1992) and levels of polymorphism (McVean and Charlesworth, 1999) at CpGprone sites cannot (Fig. 4b).

A variety of methods produced the following estimates for s: $\sim 1.3N_e^{-1}$ in *Escherichia coli* (Hartl et al., 1994), $\sim 2.2N_e^{-1}$ in *Drosophila simulans* (Akashi, 1995), $\sim 4.6N_e^{-1}$ in *D. pseudoobscura* (Akashi and Schaeffer, 1997), and $\sim 0.65N_e^{-1}$ in *D. americana* (Maside et al., 2004) and *D. miranda* (Bartolomé et al., 2005). Thus, in the units of the corresponding $1/N_e$ values, selection at synonymous sites is apparently weaker in hominids than in *Drosophila*. In *D. simulans*, $N_e \sim 5 \times 10^6$ (Ayala and Hartl, 1993). Estimates of N_e in hominids are to some extent controversial: in modern humans and chimpanzees $N_e \sim (1-2) \times 10^4$ (Yu et al., 2003); however, in the human–chimpanzee common ancestor it was either the same (Rannala and Yang, 2003) or 2–5 times higher (Satta et al., 2004). Thus, the absolute strength of selection at synonymous sites in hominids, $s \sim 10^{-5}$, is close to or even higher than in *Drosophila*, where $s \sim 5 \times 10^{-6}$.

Since coefficients of selection at synonymous sites can vary over many orders of magnitude, their concentration within a narrow range may appear unlikely (Gillespie, 1994). A plausible cause for this concentration is synergistic epistasis (Li, 1987; Akashi, 1995, p. 1074; Akashi, 1996, p. 1305), expected, for example, if synonymous sites are involved in maintaining the structure of mRNA (Innan and Stephan, 2001; Katz and Burge, 2003; Chamary and Hurst, 2005a). With synergistic epistasis, selection against a deleterious nucleotide is negligible when most of the sites of the molecule are occupied by beneficial nucleotides; however, selection gradually gets stronger when the number of deleterious nucleotides increases and eventually becomes sufficient to arrest their further accumulation (Kondrashov, 1994; Piganeau et al., 2001; Berg et al., 2004). This happens when s grows past $\sim 0.1N_e^{-1}$ (Akashi, 1996; Ohta, 2002), and the further growth of s (past $\sim 5.0N_e^{-1}$, Maside et al., 2004; Fig. 2b) can eventually eliminate almost all deleterious nucleotides, making selection negligible again. Thus, at mutation–selection–drift equilibrium, coefficients of selection against deleterious nucleotides at the majority of sites must be confined between $\sim 0.1N_e^{-1}$ and $\sim 5.0N_e^{-1}$ (Akashi, 1996). High values of s in hominids probably suggest that their mRNAs are far from optimal.

Two factors differentially affect the rates of evolution at mammalian nonTE intron vs. four-fold synonymous sites. First, synonymous sites are CpGprone much more often, which is dictated by amino-acid composition of proteins and the genetic code. In particular, highly mutable postCpreG sites are 3 times more common among synonymous sites than among intron sites (Table 1). Second, the interplay of mutation biases and constant selection for G and C reduces R and P at synonymous nonCpG sites, but increases them at such postCpreG sites.

Together, these two factors cause the average values of R and P across all four-fold synonymous sites to be $\sim 20\%$ and $\sim 10\%$, respectively, above the R and P values for intron sites of nonTE origin. An elevated rate of evolution of synonymous sites, where selection favors more mutable G:C pairs, has been reported for *Drosophila* (McVean and Vieira, 2001). However, in hominids, R and P are also elevated at intron sites of TE origin, due to their deviation from mutation–drift equilibrium, and not to selection (Table 1, Figs. 1 and 2). It is a mere coincidence

(Chamary and Hurst, 2004) that the average rate of evolution at all four-fold synonymous sites is very close to that at all intron sites (Hughes and Yager, 1997; Subramanian and Kumar, 2003). Similarly, all four nucleotide frequencies at four-fold synonymous sites are close to 25% (Table 1) due to selection in favor of G and C being counterbalanced by their elevated mutability, and not to selective neutrality.

With $s \sim 0.25 N_e^{-1}$, selection is weak enough to allow fixations of many slightly deleterious nucleotides. If suboptimal nucleotides (mostly, A and T) with $s \sim 10^{-5}$ occupy ~30% from $\sim 3 \times 10^7$ synonymous sites in the diploid mammalian genome, an organism carries $\sim 10^7$ deleterious nucleotides at such sites, which constitute ~100 lethal equivalents. The survival of a population of such organisms requires synergistic epistasis among loci (Kondrashov, 1995). A substantial fraction of new mutations replaces a suboptimal nucleotide with the optimal one and, thus, are slightly beneficial.

Our analysis makes it possible to explain several observations. At synonymous sites, the frequency of C is higher than the frequency of G (Chamary and Hurst, 2004) because four-fold postC sites, where G is rare, are >2 times more common than preG sites, where C is rare, (Table 1). This fact, dictated by the genetic code, where all codon families with C at the second position are four-fold degenerate, could be responsible for strand asymmetry in evolution at synonymous sites (Webster and Smith, 2004). Synonymous sites of constitutive exons have higher frequencies of G and C and evolve more rapidly than such sites of alternatively spliced exons (Iida and Akashi, 2000) because selection in favor of G and C at synonymous sites is stronger in constitutive exons. If synonymous sites are released from selective constraint, for example after a gene turns into a pseudogene, this leads to a large, temporary increase in the rate of their evolution (Bustamante et al., 2002), due to the above mutation–drift equilibrium frequencies of mutable C and G at CpGprone synonymous sites. The same mechanism leads to temporarily elevated rates of evolution of newly inserted transposons (Fig. 2).

Since biased gene conversion can lead to the same dynamics as selection (e.g. Lercher et al., 2002a), we cannot formally discriminate between the two. However, an important role of biased gene conversion in creating the patterns described above is unlikely (Eyre-Walker, 1999), because of the contrasts between exons and introns of the same genes. While selection can obviously be very different at synonymous exon sites vs. intron sites, it is unclear how the rate of biased gene conversion could change drastically at exon/intron boundaries. The increased probability of fixation of AT>GC mutations (Webster and Smith, 2004), as well as the excess of GC>AT over AT>GC polymorphisms (Fig. 1c) can be due to selection for G and C (Smith and Eyre-Walker, 2001; Lercher et al., 2002a, b; Webster et al., 2003). The exon–intron contrasts also argue against transcription-coupled repair bias (Green et al., 2003; Majewski, 2003) as a cause of patterns reported here.

Widespread, weak advantage of nucleotides G and C at synonymous sites supports selection on mRNA stability as an important factor in the dynamics of such sites (Chamary and Hurst, 2005a). In contrast, this advantage appears to be inconsistent with the possible involvement of such sites in splice regulation (Eskesen et al., 2004; Fairbrother et al., 2004; Willie and Majewski, 2004), since GC-content at such sites diminishes near exon–intron junctions (Chamary and Hurst, 2005b).

The available methods of estimating K_s, the evolutionary distance at synonymous sites (Yang, 1997), do not accommodate trimodal distributions of the rates of evolution and nucleotide frequencies at individual sites (Table 1). Since the context of a synonymous site is mostly determined by the surrounding non-synonymous sites, a synonymous site remains within the same class for a long time. Even for a relatively close mouse–rat pair, divergences at CpGprone four-fold synonymous sites, estimated using the Tamura–Nei formula (Tamura and Nei, 1993), are well below, relative to the divergence at nonCpG sites, of what is expected from the 1:1.5:2.5 ratios observed in the human–chimpanzee pair (data not reported). This indicates that multiple substitutions occurred at CpGprone synonymous sites since rat–mouse divergence, and that the Tamura–Nei formula, which assumes equal rates of reciprocal substitutions, underestimates divergences at such sites. In the case of mouse–human divergence, saturation at CpGprone sites is much more pronounced (data not reported). Thus, estimates of K_s between distant mammals are unreliable.

Even the correct values of K_s should not be used to estimate mutation rates in mammals, due to lack of neutrality (Kondrashov, 2001). Probably, neutral divergence between a pair of mammals (and, thus, the mutation rates outside CpG context multiplied by the number of generations of their independent evolution) can be approximated as ~1.1 times the (correctly estimated) K_s at nonCpG four-fold synonymous sites. Whether synonymous sites are a suitable neutral point of reference (Fay et al., 2001, 2002; Anisimova et al., 2002; Smith and Eyre-Walker, 2002; Eyre-Walker, 2002) in tests for positive selection, is not clear (Akashi, 1995), although the answer may be affirmative for hominids, since selection with $s \sim 0.25 N_e^{-1}$ at synonymous sites affects R and P in almost the same way (Fig. 4b).

Supporting information

C codes for calculating the properties of mutation–drift–selection equilibrium are available from ftp://ftp.ncbi.nih.gov/pub/kondrashov/k4.

Acknowledgment

This research was supported in part by the Intramural Research Program of the NIH, National Library of Medicine.

References

Akashi, H., 1995. Inferring weak selection from patterns of polymorphism and divergence at silent sites in Drosophila DNA. Genetics 139, 1067–1076.

Akashi, H., 1996. Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. Genetics 144, 1297–1307.

Akashi, H., 1999a. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics 151, 221–238.

Akashi, H., 1999b. Within- and between-species DNA sequence variation and the 'footprint' of natural selection. Gene 238, 39–51.

Akashi, H., 1999c. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics 151, 221–238.

Akashi, H., 2003. Translational selection and yeast proteome evolution. Genetics 164, 1291–1303.

Akashi, H., Schaeffer, S.W., 1997. Natural selection and the frequency distributions of "silent" DNA polymorphism in Drosophila. Genetics 146, 295–307.

Andersson, S.G., Kurland, C.G., 1990. Codon preferences in free-living microorganisms. Microbiol. Rev. 54, 198–210.

Anisimova, M., Bielawski, J.P., Yang, Z., 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol. Biol. Evol. 19, 950–958.

Ayala, F.J., Hartl, D.L., 1993. Molecular drift of the bride of sevenless (boss) gene in Drosophila. Mol. Biol. Evol. 10, 1030–1040.

Bartolomé, C., Maside, X., Yi, S., Grant, A.L., Charlesworth, B., 2005. Patterns of selection on synonymous and nonsynonymous variants in Drosophila miranda. Genetics 169, 1495–1507.

Berg, J., Willmann, S., Lässig, M., 2004. Adaptive evolution of transcription factor binding sites. BMC Evol. Biol. 4: Research42.

Bird, A.P., 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. 8, 1499–1504.

Bulmer, M., 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129, 897–907.

Bustamante, C.D., Nielsen, R., Hartl, D.L., 2002. A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. Mol. Biol. Evol. 19, 110–117.

Carlini, D.B., Stephan, W., 2003. In vivo introduction of unpreferred synonymous codons into the Drosophila Adh gene results in reduced levels of ADH protein. Genetics 163, 239–243.

Chamary, J.-V., Hurst, L.D., 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selective-driven codon usage. Mol. Biol. Evol. 21, 1014–1023.

Chamary, J.V., Hurst, L.D., 2005a. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol. 6, R75.

Chamary, J.V., Hurst, L.D., 2005b. Biased codon usage near intron–exon junctions: selection on splicing enhancers, splice-site recognition or something else? Trends Genet. 21, 256–259.

Chen, F.-C., Vallender, E.J., Wang, H., Tzeng, C.-S., Li, W.-H., 2001. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. J. Hered. 92, 481–489.

Comeron, J.M., 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. Genetics 167, 1293–1304.

Debry, R.W., Marzluff, W.F., 1994. Selection on silent sites in the rodent H3 histone gene family. Genetics 138, 191–202.

Duan, J., Antezana, M.A., 2003. Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. J. Mol. Evol. 57, 694–701.

Duan, J.B., Wainwright, M.S., Comeron, J.M., Saitou, N., Sanders, A.R., Gelernter, J., Gejman, P.V., 2003. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum. Mol. Genet. 12, 205–216.

Duret, L., 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. 12, 640–649.

Duret, L., Hurst, L.D., 2001. The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. Mol. Biol. Evol. 18, 757–762.

Duret, L., Mouchiroud, D., 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. 17, 68–74.

Duret, L., Semon, M., Piganeau, G., Mouchiroud, D., Galtier, N., 2002. Vanishing GC-rich isochores in mammalian genomes. Genetics 162, 1837–1847.

Ebersberger, I., Metzler, D., Schwarz, C., Paabo, S., 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. Am. J. Hum. Genet. 70, 1490–1497.

Eskesen, S.T., Eskesen, F.N., Ruvinsky, A., 2004. Natural selection affects frequencies of AG and GT dinucleotides at the 5′ and 3′ ends of exons. Genetics 167, 543–550.

Eyre-Walker, A., 1992. The effect of constraint on the rate of evolution in neutral models with biased mutation. Genetics 131, 233–234.

Eyre-Walker, A., 1997. Differentiating selection and mutation bias. Genetics 147, 1983–1987.

Eyre-Walker, A., 1999. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics 152, 675–683.

Eyre-Walker, A., 2002. Changing effective population size and the McDonald-Kreitman test. Genetics 162, 2017–2024.

Eyre-Walker, A., Bulmer, M., 1995. Synonymous substitution rates in enterobacteria. Genetics 140, 1407–1412.

Eyre-Walker, A., Hurst, L.D., 2001. The evolution of isochores. Nat. Rev. Genet. 2, 549–555.

Fairbrother, W.G., Holste, D., Burge, C.B., Sharp, P.A., 2004. Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol. 2, E268.

Fay, J.C., Wyckoff, G.J., Wu, C.I., 2001. Positive and negative selection on the human genome. Genetics 158, 1227–1234.

Fay, J.C., Wyckoff, G.J., Wu, C.-I., 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature 415, 1024–1026.

Gillespie, J.H., 1994. Substitutional processes in molecular evolution. III. Deleterious alleles. Genetics 138, 943–952.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., Pave, A., 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8, r49–r62.

Green, P., Ewing, B., Miller, W., Thomas, P.J., Green, E.D., NISC Comparative Sequencing Program, 2003. Transcription-associated mutational asymmetry in mammalian evolution. Nat. Genet. 33, 514–517.

Hartl, D.L., Moriyama, E.N., Sawyer, S.A., 1994. Selection intensity for codon bias. Genetics 138, 227–234.

Hellman, I., Zollner, S., Enard, W., Ebersberger, I., Nickel, B., Paabo, S., 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res. 13, 831–837.

Hess, S.T., Blake, J.D., Blake, R.D., 1994. Wide variations in neighbor-dependent substitution rates. J. Mol. Biol. 236, 1022–1033.

Hughes, A.L., Yager, M., 1997. Comparative evolutionary rates of introns and exons in murine rodents. J. Mol. Evol. 45, 125–130.

Hwang, D.G., Green, P., 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc. Natl Acad. Sci. USA 101, 13994–14001.

Iida, K., Akashi, H., 2000. A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. Gene 261, 93–105.

Ikemura, T., 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2, 13–34.

Innan, H., Stephan, W., 2001. Selection intensity against deleterious mutations in RNA secondary structure and rate of compensatory nucleotide substitutions. Genetics 159, 389–399.
International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.
Kanaya, S., Yamada, Y., Kudo, Y., Ikemura, T., 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238, 143–155.
Kapitonov, V., Jurka, J., 1996. The age of Alu subfamilies. J. Mol. Evol. 42, 59–65.
Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D., Kent, W.J., 2003. The UCSC genome browser database. Nucleic Acids Res. 31, 51–54.
Katz, L., Burge, C.B., 2003. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res. 13, 2042–2051.
Keightley, P.D., Eyre-Walker, A., 2000. Deleterious mutations and the evolution of sex. Science 290, 331–333.
Keightley, P.D., Gaffney, D.J., 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. Proc. Natl Acad. Sci. USA 100, 13402–13406.
Kondrashov, A.S., 1994. Muller's ratchet under epistatic selection. Genetics 136, 1469–1473.
Kondrashov, A.S., 1995. Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? J. Theor. Biol. 175, 583–594.
Kondrashov, A.S., 2001. Sex and U. Trends Genet. 17, 75–77.
Kondrashov, A.S., 2003. A direct estimate of human per nucleotide spontaneous mutation rate. Hum. Mutat. 21, 12–27.
Kumar, S., Subramanian, S., 2002. Mutation rates in mammalian genomes. Proc. Natl Acad. Sci. USA 99, 803–808.
Lercher, M.J., Hurst, L.D., 2002. Can mutation or fixation biases explain the allele frequency distribution of human single nucleotide polymorphisms (SNPs)? Gene 300, 53–58.
Lercher, M.J., Smith, N.G., Eyre-Walker, A., Hurst, L.D., 2002a. The evolution of isochores: evidence from SNP frequency distributions. Genetics 162, 1805–1810.
Lercher, M.J., Urrutia, A.O., Hurst, L.D., 2002b. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat. Genet. 31, 180–183.
Li, W.H., 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J. Mol. Evol. 24, 337–345.
Li, W.-H., 1997. Molecular Evolution. Sinauer, Sunderland.
Lu, J., Wu, C.I., 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. Proc. Natl Acad. Sci. USA 102, 4063–4067.
Majewski, J., 2003. Dependence of mutational asymmetry on gene-expression levels in the human genome. Am. J. Hum. Genet. 73, 688–692.
Maside, X., Lee, A.W., Charlesworth, B., 2004. Selection on codon usage in *Drosophila americana*. Curr. Biol. 14, 150–154.
McVean, G., Charlesworth, B., 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. Genet. Res. 74, 145–158.
McVean, G.A., Vieira, J., 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. Genetics 157, 245–257.
Nachman, M.W., Crowell, S.L., 2000. Estimate of the mutation rate per nucleotide in humans. Genetics 156, 297–304.
Nielsen, R., Akashi, H., 2003. Action of Purifying Selection at Silent Sites. Encyclopedia of the Human Genome. Nature Publishing Group, London.
Ohta, T., 2002. Near-neutrality in evolution of genes and gene regulation. Proc. Natl Acad. Sci. USA 99, 16134–16137.
Piganeau, G., Westrelin, R., Tourancheau, B., Gautier, C., 2001. Multiplicative versus additive selection in relation to genome evolution: a simulation study. Genet. Res. 78, 171–175.
Rannala, B., Yang, Z., 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164, 1645–1656.
Satta, Y., Hickerson, M., Watanabe, H., O'hUigin, C., Klein, J., 2004. Ancestral population sizes and species divergence times in the primate lineage on the basis of intron and BAC end sequences. J. Mol. Evol. 59, 478–487.
Shabalina, S.A., Ogurtsov, A.Y., Kondrashov, V.A., Kondrashov, A.S., 2001. Selective constraint in intergenic regions of human and mouse genomes. Trends Genet. 17, 373–376.
Sharp, P.M., Averof, M., Lloyd, A.T., Matassi, G., Peden, J.F., 1995. DNA sequence evolution: the sounds of silence. Phil. Trans. R. Soc. London B 349, 241–247.
Smith, N.G.C., Eyre-Walker, A., 2001. Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. Mol. Biol. Evol. 18, 982–986.
Smith, N.G.C., Eyre-Walker, A., 2002. Adaptive protein evolution in *Drosophila*. Nature 415, 1022–1024.
Smith, N.G.C., Hurst, L.D., 1998. Sensitivity of patterns of molecular evolution to alterations in methodology: a critique of Hughes and Yeager. J. Mol. Evol. 47, 493–500.
Smith, N.G.C., Hurst, L.D., 1999. The causes of synonymous rate variation in the rodent genome: can substitution rates be used to estimate sex bias in mutation rate? Genetics 152, 661–673.
Sorensen, M.A., Kurland, C.G., Pedersen, S., 1989. Codon usage determines translation rate in *E. coli*. J. Mol. Biol. 207, 365–377.
Subramanian, S., Kumar, S., 2003. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. Genome Res. 13, 838–844.
Sueoka, N., 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl Acad. Sci. USA 48, 582–592.
Takai, D., Jones, P.A., 2003. The CpG island searcher: a new WWW resource. Silico Biol. 3, 235–240.
Tamura, K., Nei, M., 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10, 512–526.
Urrutia, A.O., Hurst, L.D., 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics 159, 1191–1199.
Urrutia, A.O., Hurst, L.D., 2003. The signature of selection mediated by expression on human genes. Genome Res. 13, 2260–2264.
Webster, M.T., Smith, N.G.C., 2004. Fixation biases affecting human SNPs. Trends Genet. 20, 122–126.
Webster, M.T., Smith, N.G.C., Ellegren, H., 2003. Compositional evolution of noncoding DNA in the human and chimpanzee genomes. Mol. Biol. Evol. 20, 278–286.
Willie, E., Majewski, J., 2004. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. Trends Genet. 20, 534–538.
Wolfe, K.H., Sharp, P.M., Li, W.-H., 1989. Mutation rates differ among regions of the mammalian genome. Nature 337, 283–285.
Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS 13, 555–556.
Yu, N., Jensen-Seaman, M.I., Chemnick, L., Kidd, J.R., Deinard, A.S., Ryder, O., Kidd, K.K., Li, W.-H., 2003. Low nucleotide diversity in chimpanzees and bonobos. Genetics 164, 1511–1518.

Chapter 4, in full, is a reprint of the material as it appears in Kondrashov FA, Ogurtsov AY, Kondrashov AS. (2006) Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. Journal of Theoretical Biology **240**, 616-626. Elsevier Ltd. 2007. The dissertation author was the primary investigator and author of this paper.

Chapter 5.

Evolution of glyoxylate cycle enzymes in Metazoa: evidence of multiple horizontal transfer events and pseudogene formation

Biology Direct

BioMed Central

Research

Open Access

Evolution of glyoxylate cycle enzymes in Metazoa: evidence of multiple horizontal transfer events and pseudogene formation

Fyodor A Kondrashov*[1], Eugene V Koonin[2], Igor G Morgunov[3], Tatiana V Finogenova[3] and Marie N Kondrashova[4]

Address: [1]Section on Ecology, Behavior and Evolution, Division of Biological Sciences, University of California at San Diego, 2218 Muir Biology Building, La Jolla, CA 92093, USA, [2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA, [3]Skryabin Institute of Biochemistry and Physiology of Microorganisms, Russian Academy of Sciences, Pushchino, Russian Federation and [4]Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, Pushchino, Russian Federation

Email: Fyodor A Kondrashov* - fkondrashov@ucsd.edu; Eugene V Koonin - koonin@ncbi.nlm.nih.gov; Igor G Morgunov - morgunovs@rambler.ru; Tatiana V Finogenova - finog@ibpm.puschino.ru; Marie N Kondrashova - kondrashova@iteb.ru

* Corresponding author

Published: 23 October 2006

Biology Direct 2006, **1**:31 doi:10.1186/1745-6150-1-31

Received: 16 October 2006
Accepted: 23 October 2006

This article is available from: http://www.biology-direct.com/content/1/1/31

Abstract

Background: The glyoxylate cycle is thought to be present in bacteria, protists, plants, fungi, and nematodes, but not in other Metazoa. However, activity of the glyoxylate cycle enzymes, malate synthase (MS) and isocitrate lyase (ICL), in animal tissues has been reported. In order to clarify the status of the MS and ICL genes in animals and get an insight into their evolution, we undertook a comparative-genomic study.

Results: Using sequence similarity searches, we identified MS genes in arthropods, echinoderms, and vertebrates, including platypus and opossum, but not in the numerous sequenced genomes of placental mammals. The regions of the placental mammals' genomes expected to code for malate synthase, as determined by comparison of the gene orders in vertebrate genomes, show clear similarity to the opossum MS sequence but contain stop codons, indicating that the MS gene became a pseudogene in placental mammals. By contrast, the ICL gene is undetectable in animals other than the nematodes that possess a bifunctional, fused ICL-MS gene. Examination of phylogenetic trees of MS and ICL suggests multiple horizontal gene transfer events that probably went in both directions between several bacterial and eukaryotic lineages. The strongest evidence was obtained for the acquisition of the bifunctional ICL-MS gene from an as yet unknown bacterial source with the corresponding operonic organization by the common ancestor of the nematodes.

Conclusion: The distribution of the MS and ICL genes in animals suggests that either they encode alternative enzymes of the glyoxylate cycle that are not orthologous to the known MS and ICL or the animal MS acquired a new function that remains to be characterized. Regardless of the ultimate solution to this conundrum, the genes for the glyoxylate cycle enzymes present a remarkable variety of evolutionary events including unusual horizontal gene transfer from bacteria to animals.

Reviewers: Arcady Mushegian (Stowers Institute for Medical Research), Andrey Osterman (Burnham Institute for Medical Research), Chris Ponting (Oxford University).

Open peer review

This article was reviewed by Arcady Mushegian (Stowers Institute for Medical Research), Andrei Osterman (Burnham Institute for Medical Research), Chris Ponting (Oxford University). For the full reviews, please go to the Reviewers' comments section.

Background

Glyoxylate cycle is a distinct, anaplerotic variant of the tricarboxylic acid (TCA) cycle the net effect of which is the conversion of two molecules of acetyl-CoA to succinate gluconeogenesis. The glyoxylate cycle shares three of the five involved enzymes with the TCA cycle. skips the two rate-limiting decarboxylation steps of the latter, which are catalyzed by isocitrate dehydrogenase and α-ketoglutarate dehydrogenase (Figure 1; [1,2]). The glyoxylate cycle deviates from the TCA cycle when isocitrate, instead of being decarboxylated into α-ketoglutarate by isocitrate dehydrogenase, is converted into glyoxylate and succinate by isocitrate lyase (ICL). Malate synthase (MS) completes the shortcut by producing malate from glyoxylate and acetyl-CoA (Figure 1). Succinate produced by the glyoxylate cycle is utilized, primarily, for carbohydrate synthesis. Both ICL and MS are essential for the function of this pathway and are thought to be dedicated, glyoxylate cycle-specific enzymes such that their activities are often considered to be signatures of this pathway [2].

It is widely accepted that the glyoxylate cycle operates in bacteria [2], fungi [3], some protists [4,5], and plants [6]; in addition, recent reports [7,8] identified a bifunctional enzyme in nematodes with both ICL and MS activities that apparently evolved by the fusion of the respective genes (see also [2]). Although several authors reported ICL and/or MS activity in other Metazoa, including birds [9], reptiles [10,11], and placental mammals [12-20], the claim that the glyoxylate cycle functions in animals other than nematodes remains controversial [21,22]. One of the major problems with regard to the existence of the glyoxylate cycle in Metazoa is the failure to identify the ICL and MS genes in metazoan genomes, except for those of the nematodes. Here, we undertake a bioinformatic analysis aimed at detection of orthologous genes for the two glyoxylate cycle-specific enzymes in the available complete and draft genomes of various animals and reveal the existence of a MS pseudogene in placental mammals. We further examine the phylogenies of these enzymes and derive evolutionary scenarios that include multiple horizontal gene transfer (HGT) events.

Results

Isocitrate lyase and malate synthase genes and pseudogenes in animals

In addition to the previously identified ICL homologs in the nematodes *Caenorhabditis* and *Strongyloides*, a putative ICL gene has been annotated in the mosquito *Anopheles gambiae* and the sea anemone *Nematostella vectensis*, and we also found an incomplete homolog in the mosquito *Aedes aegypti*. However, the extremely high similarity of the protein sequence of the mosquito and bacterial genes (only 20% divergence between *A. gambiae* and *E. coli*) and the lack of introns in the mosquito sequence strongly suggested a bacterial contamination. Indeed, such contamination appears to be common at least in the *A. gambiae* genome (S. L. Mekhedov and EVK, unpublished observations). In contrast, the predicted sea anemone ICL sequence contained introns and was identical to several EST sequences. In addition, we identified ICL homolgos among EST sequences for two other Cnidarians (*Acropora millepora, Hydractinia echinata*) and several nematodes (*Ancylostoma ceylanicum, Globodera rostochiensis, Heterodera glycines, Parastrongyloides trichosuri, Pristionchus pacificus, Meloidogyne hapla, Meloidogyne javanica, Xiphinema index*). The high sequence conservation of ICL (Table 1) implies that, if intact copies of this gene were present in other completely sequenced metazoan genomes, we would have been able to detect them easily. Thus, it appears that, of all Metazoa with sequenced genomes, only nematodes and Cnidaria encode ICL.

In contrast, apparent MS orthologs are readily identifiable in several animal genomes including nematodes (*C. elegans, C. briggsae, C. remanei*), cnidarians (*N. vectensis*), insects (*A. gambiae, A. aegypti, Bombyx mori*), echinoderms (*Strongylocentrotus purpuratus*), and vertebrates (*Danio rerio, Tetraodon nigroviridis, Fugu rubripes, Xenopus tropicalis, Monodelphis domestica*). The *Ornithorhynchus anatinus* (platypus) and *Oryzias latipes* (fish) genomes also appear to possess the MS gene. In addition, for an insect (*Spodoptera frugiperda*), two cnidarians (*Hydractinia echinata, Acropora millepora*), a variety of nematodes (*Ancylostoma caninum, Ancylostoma ceylanicum, Globodera rostochiensis, Heterorhabditis bacteriophora, Heterodera glycines, Heterodera schachtii, Meloidogyne arenaria, Meloidogyne incognita, Meloidogyne javanica, Parastrongyloides trichosuri, Pristionchus pacificus, Trichostrongylus vitrinus, Trichuris vulpis, Xiphinema index*), a primitive chordate (*Branchiostoma floridae*, lancelet), and several vertebrates (*Trichosurus vulpecula, Hippoglossus hippoglossus, Oryzias latipes, Salmo salar, Pimephales promelas, Xenopus laevis, Gasterosteus aculeatus, Fundulus heteroclitus*), we detected at least one EST corresponding to the MS gene. However, these sequences were excluded from further analysis because they did not cover the entire coding sequence. None of the detected MS homologs from animals have been characterized experimentally although some of them are annotated in Genbank as proteins similar to the nematode malate synthase/isocitrate lyase bifunctional protein. In the genome sequence of the sea anemone (a Cnidarian) *N. vectensis*, we detected two distinct MS genes; however, we

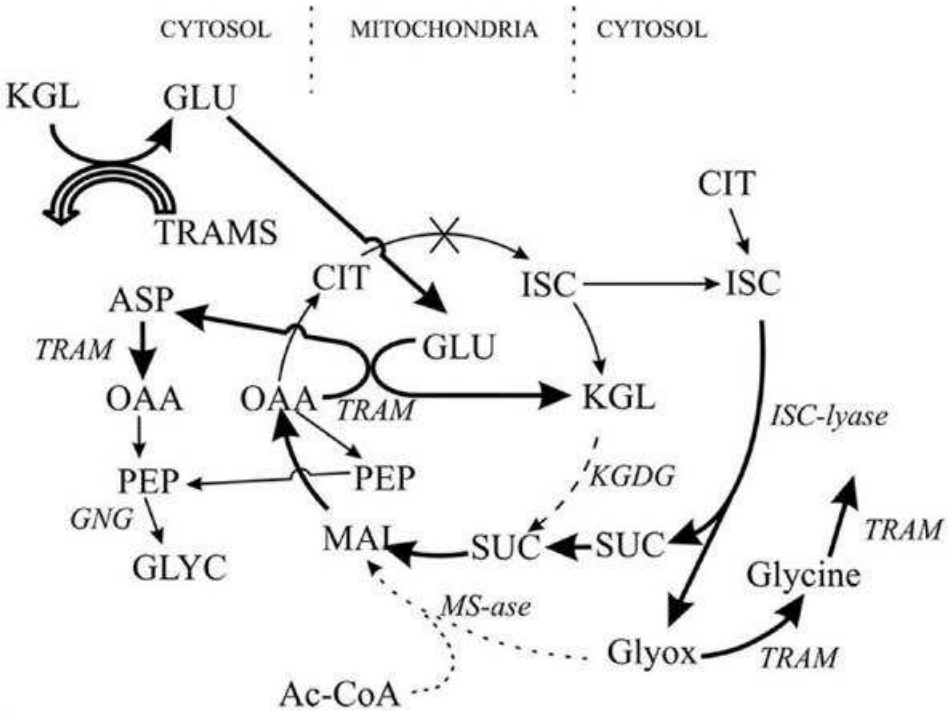


Figure 1

The complete Krebs, truncated Krebs and glyoxylate cycles. Abbreviations: α-ketoglytarate -KGL, α-ketoglytarate dehydrogenase – KGDH, acetyl-CoA – Ac-CoA, aspartate – ASP, citrate-CIT, gluconeogenesis – GNG, glutamate – GLU, glycogene – GLYC, glyoxylate-Glyox, isocitrate – ISC, isocitrate lyase – ICL, malate-MAL, malate synthase – MS, oxalacetate – OAA, phosphoenolpyruvate – PEP, succinate-SUC, transaminases (aminotransferases) – TRAM. The truncated Krebs cycle includes the OAA + GLU -> ASP + KGL reaction that is catalyzed by glutamate-glyoxylate-aminotransferase [58, 59]. TRAM reactions in mitochondria and cytosol are connected by common amino and keto acids. Thick lines represent rapid reaction steps, dashed lines – slow and easily inhibited steps, crossed out like is the blocked aconitase reaction, dotted like – malate synthase pathway that may have been recently lost in placental mammal common ancestor.

strongly suspect that one of these is a contamination from an animal source because ESTs corresponding to this genes were not detectable, and it clustered with vertebrates and sea urchin in a phylogenetic tree that included insect MS genes as an outgroup (data not shown).

Despite the lack of experimental evidence, there are several indications that animal MS homologs (in addition to those from nematodes) are functional enzymes. Firstly, the coding sequences of these genes do not contain nonsense or frameshift mutations or large insertions or deletions, and the protein sequences retain the conserved motifs characteristic of bacterial MS (data not shown). Secondly, the gene structure is preserved between closely related species, and all introns have the canonical splicing sites (GT...AG), suggesting that the transcripts of these genes are properly spliced. Thirdly, most of these sequences contain regions that are identical or nearly identical to EST sequences. Finally, the rate of non-synonymous substitutions in these genes is substantially lower than the rate of synonymous substitutions, which indicates that these genes are subject to purifying selection at the level of the protein function (Table 2).

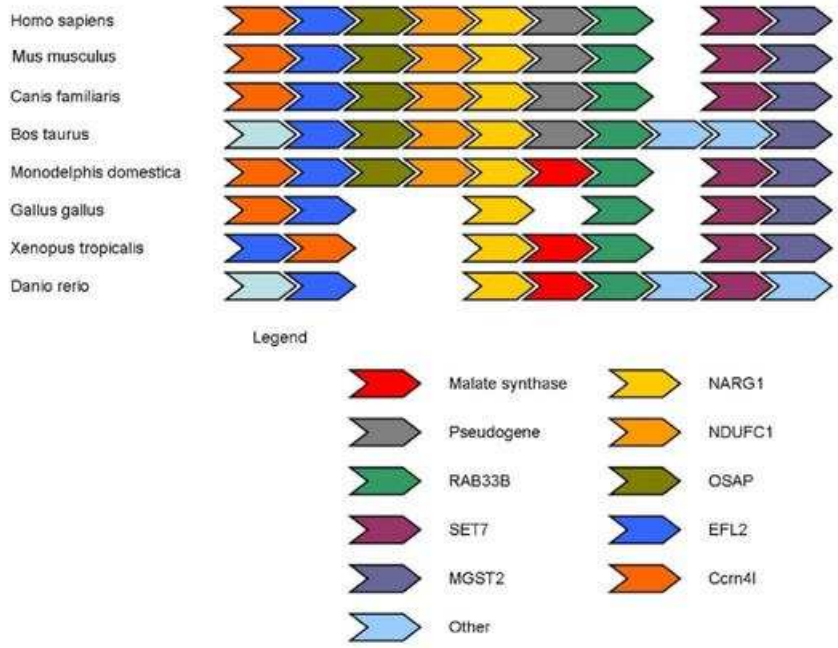


Figure 2
The syntenic region around malate synthase orthologs in completely sequenced Coelomate genomes.

We have not found the MS gene in more than 15 sequenced genomes of placental mammals. Given the substantial number of genomes searched, the high conservation of the MS sequence in other animals (Table 2), and the fact that the closely related opossum sequence was used as the query to search other mammalian genomes, it seems most unlikely that we have missed this gene. A TBLASTN search of the placental mammal genome sequences using the opossum MS sequence as the query showed several marginally significant hits to the same genomic region where the MS is found in the opossum genome. However, these searches detected only a small portion of the MS sequence in placental mammal genomes, and some of the identifiable sequences contained stop codons (Table 3). A comparison of the gene orders in the corresponding genomic regions shows that the sequences similar to MS were located in the exact position occupied by the MS gene in other animals (Figure 2). Thus, it appears that these searches detect the true orthologs of MS but the gene was inactivated and became a pseudogene in the placental mammal lineage. Although the general synteny conservation in this genomic region extends to the chicken genome (Figure 2), we found no evidence of a functional gene or a pseudogene in that region in chicken. Thus, it appears likely that the MS gene has been independently disrupted beyond recognition in the chicken genome. Similarly, the gene order is conserved between *A. gambiae* and the numerous sequenced genomes of *Drosophila* species (data not shown), however, we have not been able to find any traces of a pseudogene in *Drosophila*.

The presence of the detectable MS pseudogene in several mammalian genomes seemed unexpected because pseudogenes are usually not recognizable after ~100 million years that separate the mammalian orders from their com-

Table 1: Protein divergence (p-distance) of isocitrate lyase from selected genomes.

	Euglena gracilis	*Arabidopsis thaliana*	*Saccharomyces cerevisiae*	*Caenorhabditis elegans*	*Dictyostelium discoideum*	*Chlamydomonas reinhardtii*	*Escherichia coli*	*Brucella melitensis*
Arabidopsis thaliana	0.743							
Saccharomyces cerevisiae	0.758	0.479						
Caenorhabditis elegans	0.714	0.606	0.612					
Dictyostelium discoideum	0.708	0.600	0.608	0.369				
Chlamydomonas reinhardtii	0.716	0.601	0.614	0.380	0.357			
Escherichi coli	0.703	0.580	0.606	0.403	0.383	0.360		
Brucella melitensis	0.710	0.607	0.613	0.275	0.344	0.348	0.363	
Sulfolobus solfataricus	0.717	0.566	0.598	0.397	0.394	0.370	0.379	0.369

mon ancestor, as indicated by several studies of human and mouse genome divergence [23,24]. However, the rodent lineage appears to evolve substantially faster than other mammalian orders [25-27], and indeed, the MS pseudogene was not detected by genome-wide TBLASTN searches of rodent genomes (Table 3). Thus, some ancestral pseudogenes might have evolved beyond recognition only in the fastest evolving mammalian orders but remain recognizable in others.

Horizontal gene transfer in the evolution of glyoxylate cycle enzymes in eukaryotes

Several lines of evidence suggest that there was extensive HGT of bacterial MS and ICL genes into several eukaryotic lineages [28]. The two genes are fused to form a bifunctional gene in the nematodes and *Euglena*, but in the nematodes, the ICL domain precedes the MS domain, whereas *Euglena* has the reversed domain order [5]. Since ICL and MS are encoded in the same *ace* operon in many bacteria, and the gene order in the operon also varies, it has been suggested that nematodes and *Euglena* acquired these genes via HGT from bacteria with the respective gene orders in the *ace* operon [5]. This hypothesis predicts that, in phylogenetic trees, the domains from the bifunctional eukaryotic genes should cluster with homologs from bacteria that have the same gene order in the *ace* operon. In practice, testing this prediction was not a straightforward task. The ICL domain of the bifunctional enzymes of the nematodes showed very high (>70% identity) sequence similarity to the ICL of α-proteobacteria, in particular, those of the genus *Brucella*, and clustered with these bacterial proteins in the phylogenetic tree (Figure 3). However, in the sequenced α-proteobacterial genomes, the MS gene is located in a region distant from the ICL gene such that there is no *ace* operon. Interestingly, the MS domain sequence of the bifunctional nematode enzyme showed by far greater similarity to the MS from a different assemblage of bacteria, in particular, several species of Gram-positive bacteria of the genus *Bacillus* (~57% identity), in contrast to the much lower similarity to the MS of *Brucella* (~25% identity). In the phylogenetic tree of MS, the nematode sequences did not cluster with any specific bacterial clade but rather was positioned at the root of the bacterial subtree (Figure 3). This might result from acceleration of evolution of the MS domain in the nematode lineage and/or the absence of the actual bacterial source of the nematode gene in the current databases. Taken together, the evidence seems to be compelling for the horizontal transfer of the ICL-MS gene from bacteria into the nematode lineage. The most likely scenario would involve HGT into the nematode lineage of a "hybrid" *ace* operon containing a "proteobacterial-type" ICL and a "Gram-positive-type" MS; a bacterial genome with such a "hybrid" *ace* operon (or the actual fusion of the two genes) remains to be discovered.

Interestingly, the ICL and MS genes in the cnidarian genome appear to be encoded in tandem but on the opposite strands, in the convergent, 5'-5' orientation. However, without further sequencing of this genomic region from other cnidarians, it is unclear if the two genes originate from an ancestral *ace* operon but one of them was inverted in the sea anemone or the current gene organization is an assembly artifact. In the reconstructed phylogenies (Figure 3), the sea anemone ICL and MS genes cluster within different sets of bacteria which, as in the case of nematodes, might reflect acquisition of a "hybrid" ace operon from an unknown bacterial source.

With regard to the bifunctional gene of *Euglena*, specific phylogenetic inferences were not feasible because of the extremely high rate of evolution in the *Euglena* lineage

Table 2: Pairwise comparisons of malate synthase genes in Coelomata genomes.

	Anopheles gambiae	*Aedes aegypti*	*Bombyx mori*	*Strongylocentrotus purpuratus*	*Danio rerio*	*Takifugu rubripes*	*Tetraodon nigroviridis*	*Xenopus tropicalis*	*Monodelphis domestica*
Anopheles gambiae		65.2561	66.8261	66.3264	65.3852	21.7031	59.0776	62.9706	65.6821
Aedes aegypti	0.1874		64.6634	66.2250	66.4475	63.7981	65.8231	12.0822	66.8287
Bombyx mori	0.4153	0.4429		64.8519	7.2214	66.0626	66.5992	66.4243	65.9305
Strongylocentrotus purpuratus	0.4654	0.4957	0.5145		7.0848	27.4334	18.0603	56.6819	64.4132
Danio rerio	0.4748	0.4742	0.5009	0.3354		3.6424	6.9404	12.1555	12.9374
Takifugu rubripes	0.4522	0.4805	0.4817	0.3317	0.1844		0.3413	12.8377	37.2786
Tetraodon nigroviridis	0.4414	0.4585	0.4728	0.3298	0.1919	0.0441		36.8792	19.8208
Xenopus tropicalis	0.4877	0.5024	0.4543	0.3374	0.2440	0.2583	0.2530		3.5119
Monodelphis domestica	0.4395	0.4348	0.4682	0.3013	0.2222	0.2211	0.2189	0.1533	

The lower half of the table shows the rates of evolution in nonsynonymous sites, and the upper half shows the rates of evolution in synonymous sites. Most of the synonymous evolution rates were at the saturation levels. However, in each case, the estimated nonsynonymous substitution rate was significantly lower than the corresponding synonymous rate, which is indicative of purifying selection at the amino acid sequence level.

(Figure 3). Nevertheless, acquisition of a MS-ICL operon via HGT remains a distinct possibility. In addition to the apparent HGT of the bifunctional gene into the nematode lineage, the phylogenetic tree of ICL suggests at least three other independent instances of bacteria-to-eukaryotes HGT – into the *Nematostella*,*Dictyostelium* and Chlamydomonadaceaelineages (Figure 3a). The rest of the eukaryotic ICLs, i.e., those from plants, fungi, and the ciliate *Tetrahymena*, form a well-defined clade with one of the two copies of the ICL gene from *Mycobacteria* and *Anaeromyxobacter dehalogenans* (Figure 3a). The monophyly of this clade is additionally supported by the presence of a distinctive inserted domain which seems to be a derived shared character (Figure 4). Thus, considering all the evidence, the most likely evolutionary scenario for ICL seems to include the following events (Figure 5): i) early acquisition of the ICL gene by an ancestral eukaryote from bacteria, most likely, the mitochondrial endosymbiont, ii) evolution of the insertion domain, possibly, by internal duplication with subsequent radical divergence, iii) secondary, reverse HGT of the ICL gene from an early eukaryote to a bacterium (possibly, an ancestral *Mycobacterium*), iii) loss of the ICL gene at the outset of animal evolution, iv) at least five additional HGTs from bacteria

Table 3: Malate synthase pseudogenes in placental mammals

Species	Number of exons found by TBLASTN	Percent of opossum MS gene covered by TBLASTN hits	Number of stop codons in hits
Bos Taurus	2	8.6%	0
Canis familiaris	2	10.6%	0
Cavia porcellus	Not Found		
Dasypus novemcinctus	3	26.5%	3
Echinops telfairi	2	21.4%	3
Felis catus	4	32.5%	0
Homo sapiens	2	27.0%	0
Loxodonta africana	2	14.8%	1
Macaca mulatta	4	18.2%	0
Mus musculus	Not Found		
Oryctolagus cuniculus	1	13.5%	2
Rattus norvegicus	Not Found		
Sorex araneus	2	17.5%	1
Myotis lucifugus	2	18.0%	1
Otolemur garnettii	2	22.2%	1
Spermophilus tridecemlineatus	2	25.2%	1

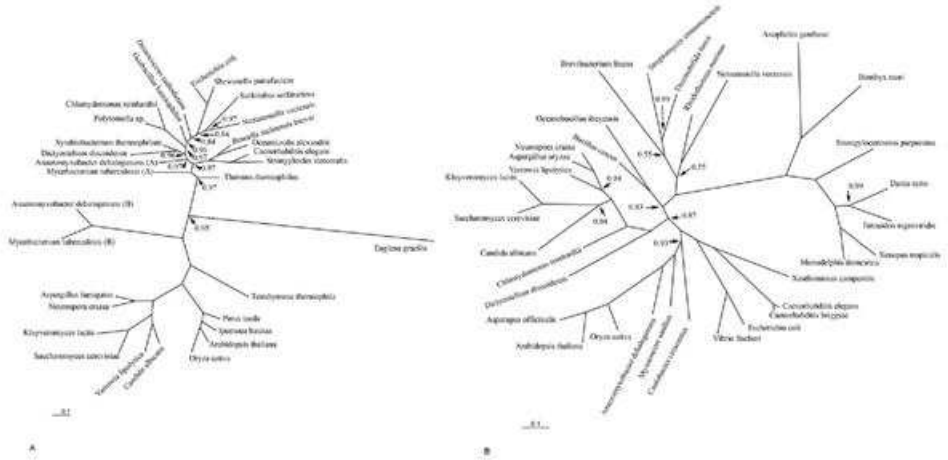


Figure 3
The phylogenies of isocitrate lyase (a) and malate synthase (b). The tree was constructed using the Bayesian approach with the posterior probabilities shown on the tree. Posterior probabilities of 1.0 are not shown.

to eukaryotes, resulting in displacement of the ancestral eukaryotic form of ICL by various bacterial forms in chlamydomonads, *Dictyostelium*, *Euglena*, cnidaria, and nematodes. In the case of nematodes and Euglena, and possibly, cnidaria as well, HGT was accompanied by fusion of ICL and MS genes, probably, facilitated by the juxtaposition of these genes in the respective bacterial *ace* operons (it is also conceivable that the fusion occurred within a bacterial genome prior to the HGT). An alternative scenario would involve the origin of the eukaryotic-type ICL in a distinct bacterial lineage (possibly, an ancestral *Mycobacterium*) with subsequent HGT into an early prokaryote. Given that all bacteria that have the eukaryotic-type ICL also possess a second, typical bacterial ICL, this scenario seems less likely. Regardless of the exact evolutionary scenario of ICL, the unusual, for animals, acquisition of the bifunctional ICL-MS enzyme by nematodes via HGT from a bacterial source appears undeniable.

The MS phylogenetic tree is less well-resolved than the ICL tree (Figure 3b), and the multiple alignment of MS has not revealed any plausible derived shared characters, such as lineage-specific large inserts (Figure 4), complicating the inference of the evolutionary scenario. In order to assess the monophyly of eukaryotic MS, we compared intron positions in eukaryotic genes. Many introns are conserved in orthologous genes from plants and animals, whereas independent gain of introns in the same position in different lineages is unlikely [29,30]. Therefore, the presence of even one shared intron strongly suggests monophyly of the respective genes as opposed to origin via independent HGT events. Indeed, although plants and coelomate animals did not form a clade in the MS tree (Figure 3b) and instead appeared to cluster with different bacterial species, plant and coelomate MS genes shared one intron in the same position (Figure 6), which is best compatible with their origin from a common eukaryotic ancestor. In contrast, the nematode and the cnidarian gene do not share introns with other animal, fungi or plant genes (or with each other), in either the MS and ICL sequences, which is consistent with a history of HGT (see above).

Discussion and conclusion

The evolution of glyoxylate cycle enzymes, ICL and MS, seems to have involved a remarkable array of events. These include at least three independent gene fusions, in nematodes, *Euglena*, and *Anaeromyxobacter dehalogenans*, multiple HGTs, and gene loss, in particular, in animals. The probable acquisition of the bifunctional ICL-MS gene

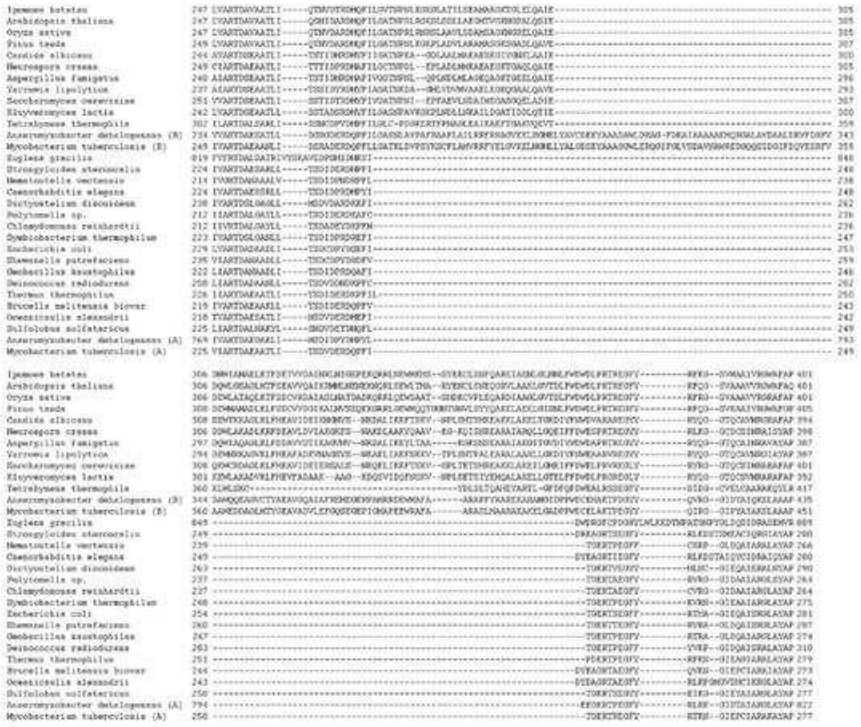


Figure 4
Multiple alignment of isocitrate lyase in the vicinity of the plant- and fungal-specific insertion.

via HGT from bacteria in nematodes and cnidarians is of special note because the very reality of acquisition of new genes by animals via HGT from bacteria is a highly controversial topic, and there are very few strongly supported cases [31-35]. The HGT of the bifunctional ICL-MS is supported by multiple lines of evidence, namely: i) unusually high similarity between the respective animal and bacterial genes, at least, in the case of ICL, ii) confident placement of the ICL and MS domains of the animal bifunctional enzymes within specific bacterial branches in phylogenetic trees, iii) juxtaposition of ICL and MS that is not seen in other eukaryotes but is common in bacteria (*ace* operons), iv) absence of shared intron positions between the bifunctional enzymes and the stand-alone homologs from other eukaryotes. Collectively, these observations seem to make the nematode and cnidarian ICL-MS true "smoking guns" of HGT from bacteria to specific lineages of animals. We believe that this is an important proof of principle that justifies a systematic search for other such cases.

Given that most archaea lack the glyoxylate cycle enzymes (with a few exceptions that, in all likelihood, can be attributed to HGT from bacteria ([36], Figure 3)), it appears most likely that eukaryotes originally acquired these genes from the mitochondrial endosymbiont. The ICL gene was lost early in metazoan evolution but was reacquired in the nematode and cnidarian lineages. In contrast, the MS gene was generally retained throughout the evolution of the eukaryotes but became a pseudogene in placental

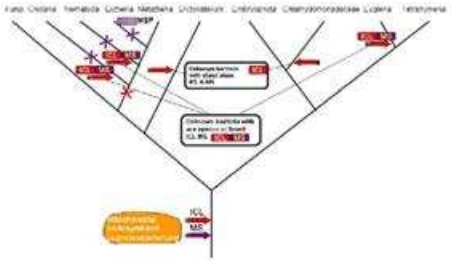


Figure 5
The inferred scenario for ICL and MS during eukaryotic evolution. The schematic shows only selected branches of the phylogenetic tree of eukaryotes, those that are relevant to inferred events in the evolution of the glyoxylate cycle. Block arrows show horizontal gene transfer, and crosses show gene loss; MSP stands for malate synthase pseudogene.

mammals. Combined with conflicting experimental data, these observations stress the conundrum around the function of the glyoxylate cycle-specific enzymes in coelomate animals. One possibility is that these enzymes, ICL and MS, have been lost in Coelomates, but MS was recruited to perform a new function. However, there is currently no experimental evidence of any function of MS other than its involvement in the glyoxylate cycle, and no indication of acceleration of evolution of the MS gene in the Coelomate lineage, which would be expected in the case of a substantial change of function. Alternatively, the ICL gene might have been lost after a different, perhaps, distantly related or unrelated gene evolved the isocitrate lyase function in the Coelomate lineage – a potential case of non-orthologous gene displacement, a fairly common evolutionary phenomenon [37]. This explanation is compatible with several experimental reports that demonstrate the presence of the ICL and MS activities in Coelomates [9-20]. However, since the validity of these experimental results has been challenged [21,22], determination of the function(s) of the MS in Coelomates would be a major step towards the resolution of the conundrum.

The functional significance of the pseudogenization of MS in placental mammals and possible independent loss of MS in birds is another enigma. One possibility is that the generally higher transaminase activity in warm-blooded mammals [38] enhanced the removal of the toxic glyoxylate through transamination by several glyoxylate-animotransferases ([39-43], Figure 1), rendering the MS activity non-essential. The alternatives are that, even if other Coelomates possess the glyoxylate cycle, placental mammals and birds have lost it entirely, or yet another gene evolved the malate synthase function in an additional case of non-orthologous gene displacement.

The extreme evolutionary mobility of the glyoxylate cycle enzymes might seem puzzling although, as far as prokary-

Monodelphis domestica	97	TERFLRALGSTAQGI Q^VDFD_DGNCPTFFNQIKGIYNIFQAV	137
Xenopus tropicalis	100	TERFHRALLSSAQGI Q^VDFD_DGNCPTFFNQIKGIFNIYQVV	140
Danio rerio	103	TQRLIMGLKSTAQGLQ^VDFD_DGNCPTYRNQIKGIYNVYQAV	143
Tetraodon nigroviridis	97	TERFIKALQTPAQGI Q^VDFD_DGNCPTYHNQIKGIHNVMKAV	137
Fugu rubripes	97	TERFIKALQTPAEGI Q^VDFD_DGNCPTYHNQIKGIHNVLKAV	137
Strongylocentrotus purpuratus	93	IQHFTRSLQSSAQGI Q^TDFD_DGHCPTRRTQIEGLYNVYRAV	133
Anopheles gambeae	98	TIHFTDCLYAEVQGI Q^VDFD_DGHCPTWRNTVLGLFNVTRAV	138
Caenorhabditis elegans	534	RKMMINAMNSGANVFM_ADFE_DSNSPTWRNQLEGQINLYDAV	574
Caenorhabditis briggsae	527	RKMMINAMNSGANVFM_ADFE_DSNSPTWRNQLEGQINLYDAV	567
Yarrowia lipolytica	118	RKMMINALNSDVWTYM_ADFE_DSSAPTWSNMDGQVNLYDGV	158
Candida albicans	105	RKMMINALNSNVATYM_ADFE_DSLTPAWKNLVEGQVNLYDGV	145
Saccharomyces cerevisiae	109	RNMLINALNAPVNTYM_TDFE_DSASPTWRNMVYGQVNLYDAI	149
Cryptococcus neoformans	97	RKMMINALNSGAKTFM_ADFE^DSNSPTWSNMLGQVNLYDAI	137
Aspergillus fumigatus	101	RKMMVNALNADVWTYM_ADFE^DSSAPTWANMNGQVNLYDAI	141
Kluyveromyces lactis	109	RNMLVNALNSDVKTYM_TDFE_DSASPTWNVIYGQVNLYDAI	149
Arabidopsis thaliana	110	RKMIINALNSGAKVFM^ADFE_DALSPSWENLMRGHVNLKDAV	150
Oryza sativa	110	RKMMINALNSGAKVFM^ADFE_DALSPTWENLMRGQVNLRDAV	150
Dictyostelium discoideum	96	RKMMINALNSGAKVFM_ADFE_DANCPNWENSIHGQQNMIDAN	136

Figure 6
Multiple alignment of malate synthase in the vicinity of the intron common to higher animals and plants. The presence of an intron is shown with the carrot symbol (^) while the absence of one is shown with an underscore (_).

otic metabolic pathways are concerned, it is not entirely unprecedented [44]. The key biological consideration appears to be that the two enzymes of the glyoxylate cycle comprise a compact, readily transferable functional unit, especially, when the two genes are juxtaposed or fused. Acquisition of this unit immediately endows the recipient with new metabolic capabilities – to produce succinate and to eliminate the toxic glyoxylate – which could be a selective advantage, at least, under some metabolic regimes.

Methods

We employed a series of similarity searches for isocitrate lyases and malate synthases in GenBank, and complete and draft genomes of all Metazoans available at NCBI and EMBL. The genomes searched and the parameters of the searches where identical for the two proteins.

The *Saccharomyces cerevisiae* sequences (isocitrate lyase – NC_001137; malate synthase – NC_001146) to the non-redundant protein sequence database at NCBI [45] using the BLASTP program [46] in order to identify all Metazoan homologues that have already been annotated in protein sequence. This approach identified isotrate lyase genes that were annotated in *A. gambiae* (XP_561347), *C. elegans* (NP_503306), *C. briggsae* (CAE62276), *Strongyloides ratti* (BAD89436) and *S. stercoralis* (AAF00535), and malate synthase genes in *A. gambiae* (XP_315354), *C. elegans* (NP_503306), *C. briggsae* (CAE62276), *S. purpuratus* (XP_782946), *D. rerio* (XP_685378) and *T. nigroviridis* (CAF91513). These homologues were identified unambiguously, with low expectation values ($E < 1 \times 10^{-30}$) and with at least 40% identity.

All genes that were predicted from complete genomes rather than obtained by direct sequencing of mRNAs (genes from *A. gambiae*, *D. rerio* and *T. nigroviridis* in this case) were checked for consistency of the annotation. To do this, the predicted protein sequence were mapped to the complete genomes available at the UC Santa Cruz Genome Browser [47] using the BLAT program [48] and checked for correct splice sites in introns (GT...AG), for start and stop codons in the first and last exons, and for the absence of nonsense or frameshift mutations in the retrieved sequence. Where appropriate, the annotation was modified to fit these criteria, and the resulting protein sequence checked by alignment to closest homologues that have been sequenced directly from an mRNA.

The next step of the sequence query was a recursive BLAST search of the available draft and complete Metazoan genomes. First, all isocitrate lyase and malate synthase protein sequences, which were identified in the step described above, were compared with the nucleotide sequences of these genomes using the TBLASTN program [46]. When a homolog was found in one of the genomes, it was annotated according to the sequence similarity with the respective protein sequence, and then checked for correct splice sites, start and stop codons, and the lack of frameshift and nonsense mutations. To complete the search cycle, the newly identified genes was then used as a query in a new TBLASTN search of the Metazoan genomes. The following genomes were queried: *Homo sapiens*, *Pan troglodytes*, *Macaca mulatta*, *Mus musculus*, *Rattus norvegicus*, *Cavia porcellus*, *Canis familiaris*, *Felis catus*, *Bos taurus*, *Dasypus novemcinctus*, *Echinops telfairi*, *Loxodonta africana*, *Oryctolagus cuniculus*, *Sorex araneus*, *Myotis lucifugus*, *Otolemur garnettii*, *Spermophilus tridecemlineatus*, *Monodelphis domestica*, *Ornithorhynchus anatinus*, *Gallus gallus*, *Xenopus tropicalis*, *Takifugu rubripes*, *Tetraodon negroviridis*, *Danio rerio*, *Ciona savignyi*, *Ciona intestinalis*, *Strongylocentrotus purpuratus*, *Bombyx mori*, *Aedes aegypti*, *Anopheles gambiae*, *Tribolium castaneum*, *Nematostella vectensis*, *Caenorhabditis elegans*, *Cenorhabditis briggsae* and twelve *Drosophila* species. To marginalize the possibility that the ICL and MS gene sequences are the result of bacterial contamination, we checked for the presence of an introns by BLAT [48] and ESTs by a TBLASTN search [46] in dbEST [49].

Finally, position-specific search implemented in PSI-BLAST [50] was used to search for possible missed homologues among the annotated genes from human, mouse, rat and *Drosophila* genomes and a ScanProsite [51] search of all genes in the UniProt (Swiss-Prot and TrEMBL) [52] and PDB [53] databases. This procedure has not revealed any Metazoan isocitrate lyase or malate synthase sequences that were not picked up with the BLASTP or TBLASTN searches.

Syntenic regions of genomes revealed by BLAST and BLAT searches of genes adjacent to the ICL and MS genes; only assembled genomes were considered. Multiple protein alignments were constructed using the MUSCLE program [54] with default parameters and manually checked for errors and for consistency of the alignment with the ScanProsite [51] ICL and MS amino acid patterns. Rates of synonymous and nonsynonymous evolution were calculated with the PAML package [55].

Phylogenetic trees were constructed by two methods, the neighbor joining procedure with 10,000 bootstrap replicates using with the MEGA program [56] and the Bayesian inference approach implemented in the MrBayes program [57] run with a GTR model assuming a gamma-distribution of substitution rates across sites for 1 million iterations (mcmc ngen = 1000000 in MrBayes). The two methods revealed, largely, congruent phylogenies.

Abbreviations

α-ketoglutarate -KGL, α-ketoglutarate dehydrogenase – KGDH, acetyl-CoA – Ac-CoA, aspartate – ASP, citrate-CIT, gluconeogenesis – GNG, glutamate – GLU, glycogene – GLYC, glyoxylate-Glyox, isocitrate – ISC, isocitrate lyase – ICL, malate-MAL, malate synthase – MS, oxalacetate – OAA, phosphoenolpyruvate – PEP, succinate-SUC, transaminases (aminotransferases) – TRAM.

Reviewers' comments

Reviewer's report 1

Arcady Mushegian, Stowers Institute for Medical Research (with additional contribution from Manisha Goel).

This is an interesting work, starting to trace the unusual path of evolution of malate synthase and isocitrate lyase in animal kingdom, with additional discussion of what might have been going on with these genes in bacteria.

I suggest that the authors do the following:

1. Due diligence with the databases of unfinished genomes: I did a quick tblastn against the environmental sequence genomes at NCBI and saw at least one entry that codes for the same domain tandem as the two-domain nematode protein: is it one ORF or two, from a nematode or perhaps from a bacterium? The unfinished bacterial genomes – perhaps the donor of two genes to the nematode lineage can be identified among them?

Author response: *We significantly expanded the scope of searches in the revised version. Indeed, there are some very similar sequences of the ICL-MS fused nematode gene in the environmental sequence database. We find sequences that are highly similar to the nematode gene in two different configurations, with ICL and MS or not fused. Unfortunately, however, it is impossible to tell whether these sequences are from bacteria or eukaryotes, and therefore, we cannot use this information to resolve any of the issues regarding the potential donor of the ICL-MS fused gene in the nematodes.*

When the authors say 'gene transfer into the nematode lineage', how do they know it is not an earlier event (search Schmidtea genome traces perhaps, also Coelenterata)? The same databases, plus ESTs, are needed to account for additional ICLs (I think I can see some in corals).

Author response: *Since the submission of the first draft of this manuscript, the cnidarian Nematostella vectensis genome draft has been completed, and now we have included the ICL and MS gene sequences found in this genome into our analysis. Interestingly, the sea anemone genes appear to cluster with bacterial genes as well, albeit with different lineages of bacteria than the nematode genes. We believe that HGT of the fused gene (or operon, with subsequent fusion) in the nematode lineage is the most parsimonious solution. An earlier HGT, e.g., to the common ancestor of Metazoa, would require genes losses in addition.*

Reviewer's report 2

Andrei Osterman, Burnham Institute for Medical Research

The strength of the manuscript by F. Kondrashov et al. on Evolution of glyoxylate cycle enzymes in Metazoa is in the detailed analysis of possible evolutionary scenarios that included multiple horizontal gene transfer (HGT) events from bacteria to eukaryotes beyond a symbiotic ancestor of extant mitochondria. Such a case-study is a best possible contribution to the heated debates on this exciting albeit highly controversial topic. Based on a solid comparative analysis of genomic sequences of multiple bacterial and eukaryotic species, which included the delineation of intron/exon structures and "pseudogenized" regions in genomes of Metazoa, the authors presented several plausible scenarios. While differing in details, all of them inevitably include several independent cross-kingdom HGT and gene fusion events. In that regard this paper is a highly recommended reading and thinking material. A weaker aspect of this study is a relatively low impact on our understanding of a metabolic driving force behind these amazing events. Despite a heroic attempt to build on the existing fragmental and highly controversial biochemical data, an emerging picture remains largely obscure. The above notion hardly argues against the authors of this study, but rather provides another illustration of a profound disregard of the basic metabolic biochemistry by the overwhelming majority of the experimental research community in the post-genomic era. A juxtaposition of a monumental effort (and quite a stunning progress) on elucidating minute details of signaling cascades, transcription machinery and other complex systems versus an apparent lack of any drive to finally straighten out basic questions such as: (i) presence or absence of malate synthase activity or (ii) actual function of a malate synthase homologs in placental animals, can hardly be reconciled other than by a popular misconception of the actual depth of our knowledge of basic metabolism. Contrasting this problem and putting it in a fundamental evolutionary context is another (likely unintended) impact of this article.

Overall, I firmly support the publication of the submitted article in "Biology Direct", and I believe that it is a perfect fit for the mission of this distinguished Journal.

Author response: *Actually, the original motivation behind this article was to apply computational approaches in an attempt to resolve the paradox of the glyoxylate cycle in mammals and birds: several laboratories have reported that this*

pathway was functional but the participant enzymes could not be identified. As it happens, the paradox only deepened as we ascertained the presence of "orphan" MS in many animals and pseudogenes in mammals. So it was very much our intent, indeed, our primary goal, to attract attention to the mysterious function of the animal MS, and hopefully, to stimulate relevant biochemical experimentation. The discovery of interesting cases of HGT was, in a sense, a by-product of our research, even if it might have the greatest general impact of the observations reported here.

Reviewer's report 3
Chris Ponting, Oxford University

This is an interesting study aiming at resolving the long-standing issue of whether malate synthase and isocitrate lyase genes are functional in many animal genomes. It is argued that the malate synthase gene is functional in non-eutherian mammals, other vertebrates, echinoderms and arthropods, but that eutherians have lost this gene through pseudogenization. Meanwhile, the isocitrate lyase gene appears to be absent from all animals with the notable exception of nematode worms (where it is fused with malate synthase). It is argued that these two genes were transferred horizontally into several eukaryotic lineages.

One of the issues with which the authors had to contend was of contamination. I agree that the homologues purported to be in mosquito genome sequences appear to result instead from contamination from bacterial sources. Certainly there is evidence that the mosquito genomes are contaminated with these bacterial genes, particularly because they appear to be single exon genes present between clone gaps in unplaced sequence. Contamination might also be an explanation for the postulated horizontal gene transfer of ICL into *Dictyostelium* and *Chlamydomonas* lineages, but this wasn't (but should have been) considered by the authors.

Author response:*Contamination can be a serious problem in studies such as this one. We have done our best to exclude the possibility of bacterial or other contaminations of Metazoan sequences by examining gene structure, EST sequences, and level of sequence similarity. We concluded that the ICL sequence from the mosquito Anopheles gambeae and one of the MS sequences from the cnidarian Nematostella vectensis are likely contaminants. By contrast, other Metazoan ICL and MS genes, including those of Dictyostelium and Chlamydomonas, contained introns, and most have several independent EST sequences in GenBank which effectively rules out bacterial contamination.*

A main finding of this report is that the isocitrate lyase gene is absent from "completely sequenced metazoan genomes". Whilst this appears true, there are numerous isocitrate lyase ESTs from cnidarians apparent in public databases. The authors will need to determine whether these represent independent horizontal or else vertical acquisitions, or else consider whether they are contaminants of the type seen in mosquito genomes.

Author response:*Indeed, these EST sequences are identical to the ICL and one of the MS genes from the genome. Thus, we have included the sequences from the recently completed genome draft of Nematostella vectensis; however, we have not included the EST sequences of other species with ESTs for which we could not obtain the cognate genomic sequences.*

Table 2 shows PAML non-synonymous and synonymous substitution rates between diverse metazoans. The vast majority of these estimates are not meaningful since saturation of substitution will have occurred, and so Table 2 should not be kept. If the authors believe me to be in error here, they should demonstrate their point using classic tests for saturation.

Author response: *Indeed many of the values reported in that table are beyond saturation. However, what is clear from this table is that the nonsynonymous divergence (which is nowhere near saturation levels) is much lower than the rate of synonymous divergence, thus demonstrating functional constraint. Therefore, we opted to keep the table after adding a note of caution to the reader regarding the interpretation of the estimates of the synonymous divergence.*

I do not think the authors have presented sufficient evidence that there "was extensive HGT of bacterial MS and ICL genes into several eukaryotic lineages". They imply that these genes have been acquired independently from separate bacterial sources before being fused in nematodes. They discount the possibility that the evolutionary rate of nematode malate synthase might be particularly high without explanation and instead choose a scenario that the bacterial source is as yet unknown. I do not consider that this provides "compelling" evidence for horizontal transfer of both ICL and MS genes from bacteria, which has implications for the title of the manuscript.

Author response: *Suppose the genes in the nematode, as well as Chlamydomonas, Dictyostelium and Nematostella (but not other eukaryotic species) have experienced a substantial acceleration of the evolution rate. In order for such acceleration to result in the nesting of these sequences within bacterial clades (which is what we observe) this acceleration must have been coupled with extensive* ***convergent*** *evolution as well. We believe that such convergent mode of evolution is highly unlikely, to say the least, and a much more parsimonious explanation for the phylogenetic tree reported here is a series of HGT events. Thus, we stand by our statement that the evidence for several cases of*

HGT from bacteria to specific eukaryotic lineages, including nematodes and Cnidaria, is compelling, and there was no reason to modify the title of the paper. Partly in consideration of these comments, the discussion of HGT was expanded in the revised version of the paper.

Moreover, if one considers that the horizontal acquisition of ICL by slime mold and *Chlamydomonas* lineages might instead be accounted for by contamination of sequences by bacterial sources, there is even less evidence for the "four additional HGTs from bacteria to eukaryotes" proposed.

Author response: *As discussed above in our response to Mushegian's comments, it is highly unlikely that the source of ICL and MS sequences from Chlamydomonas, Dictyostelium and Nematostella is bacterial contamination since the genes in these species have introns and ESTs corresponding to these genes are available for all three o fthese species. We mention these observations in the revised text.*

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

FAK, EVK, IGM, MNK designed the study, FAK and EVK carried out the bioinformatic analysis, and all authors participated in drafting of the manuscript, and approved the final version.

Acknowledgements

The authors thank Alexey Kondrashov for suggesting the possibility of non-orthologous gene displacement in glyoxylate cycle specific enzymes and for critical reading of this manuscript. FAK is a National Science Foundation Graduate Fellow.

References

1. Kornberg H, Krebs HA: **Synthesis of cell constitutions from C2-units by a modified tricarboxylic cycle.** *Nature* 1957, **179**:988-991.
2. Nelson DL, Cox MM: *Principles of Biochemistry* Fourth edition. New York, Freeman Publishers; 2005.
3. Lowenstein JM: **The tricarboxylic acid cycle (VI) Modifi ed tricarboxylic acid cycle.** In *Metabolic Pathways* 3rd edition. Edited by: Grenberg DM. New York, Academic Press; 1967:209-213.
4. Levy MR, Scherbaum OH: **Induction of the glyoxylate cycle in Tetrahymena.** *Arch Biochem Biophys* 1965, **109**:116-121.
5. Nakazawa M, Minami T, Teramura K, Kumamoto S, Hanato S, Takenaka S, Ueda M, Inui H, Nakano Y, Miyatake K: **Molecular characterization of a bifunctional glyoxylate cycle enzyme, malate synthase/isocitrate lyase, in Euglena gracilis.** *Comp Biochem Physiol B Biochem Mol Biol* 2005, **141**:445-452.
6. Kornberg HL, Beevers H: **A mechanism of conversion of fat to carbohydrate in castor beans.** *Nature* 1957, **180**:35-36.
7. Liu F, Thatcher JD, Barral JM, Epstein HF: **Bifunctional glyoxylate cycle protein of Caenorhabditis elegans: a developmentally regulated protein of intestine and muscle.** *Dev Biol* 1995, **169**:399-414.
8. Siddiqui AA, Stanley CS, Berk SL: **Cloning and expression of isocitrate lyase from human round worm Strongyloides stercoralis.** *Parasite* 2000, **7**:233-236.
9. Davis WL, Jones RG, Farmer GR, Dickerson T, Cortinas E, Cooper OJ, Crawford L, Goodman DB: **Identification of glyoxylate cycle enzymes in chick liver – the effect of vitamin D3: cytochemistry and biochemistry.** *Anat Rec* 1990, **227**:271-284.
10. Goodman DBP, Davis WL, Jones RG: **Glyoxylate cycle in toad urinary bladder: Possible stimulation by aldosterone.** *Proc Natl Acad Sci USA* 1980, **77**:1521-1525.
11. Davis WL, Jones RG, Goodman DBP: **Cytochemical localization of malate synthase in amphibian fat body adipocytes: possible glyoxylate cycle in a vertebrate.** *J Histochem Cytochem* 1986, **34**:689-692.
12. Kondrashova MN, Rodionova MA: **Realization of glyoxylate cycle in mitochondria of animal tissues.** *Dokl Akad Nauk SSSR* 1971, **196**:1225-1227. in Russian
13. Davis WL, Jones RG, Farmer GR, Cortinas E, Matthews JL, Goodman DB: **The glyoxylate cycle in rat epiphyseal cartilage: the effect of vitamin-D3 on the activity of the enzymes isocitrate lyase and malate synthase.** *Bone* 1989, **10**:201-206.
14. Davis WL, Matthews JL, Goodman DB: **Glyoxylate cycle in the rat liver: effect of vitamin D3 treatment.** *FASEB J* 1989, **3**:1651-1655.
15. Davis WL, Goodman DB, Crawford LA, Cooper OJ, Matthews JL: **Hibernation activates glyoxylate cycle and gluconeogenesis in black bear brown adipose tissue.** *Biochim Biophys Acta* 1990, **1051**:276-278.
16. Popov VN, Igamberdiev AU, Schnarrenberger C, Volvenkin SV: **Induction of glyoxylate cycle enzymes in rat liver upon food starvation.** *FEBS Lett* 1996, **390**:258-260.
17. Popov VN, Volvenkin SV, Eprintsev AT, Igamberdiev AU: **Glyoxylate cycle enzymes are present in liver peroxisomes of alloxan-treated rats.** *FEBS Lett* 1998, **440**:55-58.
18. Popov VN, Volvenkin SV, Kosmatykh TA, Susd A, Schnarrenberger C, Eprintcev AT: **Induction of a peroxisomal malate dehydrogenase isoform in liver of starved rats.** *Biochemistry (Mosc)* 2001, **66**:496-501.
19. Kokavec A, Crowe SF: **Alcohol consumption in the absence of adequate nutrition may lead to activation of the glyoxylate cycle in man.** *Med Hypotheses* 2002, **58**:411-415.
20. Morgunov IG, Kondrashova MN, Kamzolova SV, Sokolov AP, Fedotcheva NI, Finogenova TV: **Evidence of the glyoxylate cycle in the liver of newborn rats.** *Med Sci Monit* 2005, **11**:BR57-60.
21. Holmes RP: **The absence of glyoxylate cycle enzymes in rodent and embryonic chick liver.** *Biochim Biophys Acta* 1993, **1158**:47-51.
22. Jones JD, Burnett P, Zollman P: **The glyoxylate cycle: does it function in the dormant or active bear?** *Comp Biochem Physiol B Biochem Mol Biol* 1999, **124**:177-179.
23. Shabalina SA, Ogurtsov AY, Kondrashov VA, Kondrashov AS: **Selective constraint in intergenic regions of human and mouse genomes.** *Trends Genet* 2001, **17**:373-376.
24. Mouse Genome Sequencing Consortium: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420**:520-562.
25. Cooper GM, Brudno M, Green ED, Batzoglou S, Sidow A: **NISC Comparative Sequencing Program. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes.** *Genome Res* 2003, **13**:813-820.
26. Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C, Antonarakis SE: **Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs).** *Science* 2003, **302**:1033-1035.
27. Rat Genome Sequencing Project Consortium: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution.** *Nature* 2004, **428**:493-521.
28. Schnarrenberger C, Martin W: **Evolution of the enzymes of the citric acid cycle and the glyoxylate cycle of higher plants. A case study of endosymbiotic gene transfer.** *Eur J Biochem* 2002, **269**:868-883.
29. Fedorov A, Merican AF, Gilbert W: **Large-scale comparison of intron positions among animal, plant, and fungal genes.** *Proc Natl Acad Sci USA* 2002, **99**:16128-16133.
30. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV: **Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution.** *Curr Biol* 2003, **13**:1512-1517.

31. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2000, **409:**860-921.
32. Salzberg SL, White O, Peterson J, Eisen JA: **Microbial genes in the human genome: lateral transfer or gene loss?** *Science* 2001, **292:**1903-1906.
33. Stanhope MJ, Lupas A, Italia MJ, Koretke KK, Volker C, Brown JR: **Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates.** *Nature* 2001, **411:**940-944.
34. Genereux DP, Logsdon JM Jr: **Much ado about bacteria-to-vertebrate lateral gene transfer.** *Trends Genet* 2003, **19:**191-195.
35. Mitreva M, Blaxter ML, Bird DM, McCarter JP: **Comparative genomics of nematodes.** *Trends Genet* 2005, **21:**573-581.
36. Serrano JA, Bonete MJ: **Sequencing, phylogenetic and transcriptional analysis of the glyoxylate bypass operon (ace) in the halophilic archaeon Haloferax volcanii.** *Biochim Biophys Acta* 2001, **1520:**154-162.
37. Koonin EV, Galperin MY: *Sequence – Evolution – Function: Computational Approaches in Comparative Genomics* Norwell MA, Kluwer Academic Publishers; 2003.
38. Akhmerov RN, Sultanov S, Allamuratov SI, Almatov KT: **The effects of transaminase inhibitors on oxygen consumption in mitochondria and live frogs and mice.** In *Succinic acid in Medicine, Food Industry and Agriculture* Edited by: Kondrashova MN, Kaminsky YG, Maevsky EI. Pushchino; 1997:8-13. in Russian
39. Meister A, Sober HA, Tice SV, Fraser PE: **Transamination and associated deamidation of asparagine and glutamine.** *J Biol Chem* 1952, **197:**319-330.
40. Tolbert N: **Mammalian peroxisomes (microbodies).** *J Histochem Cytochem* 1973, **21:**941-948.
41. Noguchi T, Minatogawa Y, Takada Y, Okuno E, Kido R: **Subcellular distribution of pyruvate (glyoxylate) aminotransferases in rat liver.** *Biochem J* 1978, **170:**173-175.
42. Noguchi T, Okuno E, Takada Y, Minatogawa Y, Okai K, Kido R: **Characteristics of hepatic alanine-glyoxylate aminotransferase in different mammalian species.** *Biochem J* 1978, **169:**113-122.
43. Noguchi T, Fujiwara S: **Identification of mammalian aminotransferases utilizing glyoxylate or pyruvate as amino acceptor. Peroxisomal and mitochondrial asparagine aminotransferase.** *J Biol Chem* 1988, **263:**182-186.
44. Koonin EV, Galperin MY: *Sequence – Evolution – Function: Computational Approaches in Comparative Genomics* Norwell (MA), Kluwer Academic Publishers; 2003.
45. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2006, **34:**D16-20.
46. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
47. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, Hillman-Jackson J, Kuhn RM, Pedersen JS, Pohl A, Raney BJ, Rosenbloom KR, Siepel A, Smith KE, Sugnet CW, Sultan-Qurraie A, Thomas DJ, Trumbower H, Weber RJ, Weirauch M, Zweig AS, Haussler D, Kent WJ: **The UCSC Genome Browser Database: update 2006.** *Nucl Acids Res* 2006, **34:**D590-598.
48. Kent WJ: **BLAT – the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656-664.
49. Boguski MS, Lowe TM, Tolstoshev CM: **dbEST – database for "expressed sequence tags".** *Nat Genet* 1993, **4:**332-333.
50. Altschul SF, Koonin EV: **Iterated profile searches with PSI-BLAST – a tool for discovery in protein databases.** *Trends Biochem Sci* 1998, **23:**444-447.
51. Gattiker A, Gasteiger E, Bairoch A: **ScanProsite: a reference implementation of a PROSITE scanning tool.** *Applied Bioinformatics* 2002, **1:**107-108.
52. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS: **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2005, **33:**D154-159.
53. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Research* 2000, **28:**235-242.
54. Edgar RC: **MUSCLE: Multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32:**1792-1797.
55. Yang Z: **PAML: A program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13:**555-556.
56. Kumar S, Tamura K, Nei M: **MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment.** *Brief Bioinform* 2004, **5:**150-163.
57. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19:**1572-1574.
58. Parlo RA, Coleman PS: **Enhanced rate of citrate export from cholesterol-rich hepatoma mitochondria. The truncated Krebs cycle and other metabolic ramifications of mitochondrial membrane cholesterol.** *J Biol Chem* 1984, **259:**9997-10003.
59. Kondrashova MN: **Interaction of transamination and oxidation processes.** *Biochemistry (Mosc)* 1991, **56:**388-405.

Chapter 5, in full, is a reprint of the material as it appears in Kondrashov FA, Koonin EV, Morgunov IG, Finogenova TV, Kondrashova MN. (2006) Evolution of glyoxylate cycle enzymes in Metazoa: evidence of multiple horizontal transfer

events and pseudogene formation. Biol Direct **1**, 31. Biomed Central 2007. The dissertation author was the primary investigator and author of this paper.

Chapter 6.

Selection for functional uniformity of tuf duplicates in gamma-proteobacteria

Selection for functional uniformity of *tuf* elongation factor duplicates in gamma proteobacteria

Fyodor A. Kondrashov[1*], Tatiana A. Gurbich[2], Peter K. Vlasov[3],

[1]Section on Ecology, Behavior and Evolution, Division of Biological Sciences, University of California at San Diego, 2218 Muir Biology Building, La Jolla, CA 92093, USA

[2]Bard College, Annandale-on-Hudson, NY 12504, USA

[3]Engelhardt Institute of Molecular Biology, Vavilova 32, 119991 Moscow, Russian Federation

Teaser: Selection prevents functional divergence of elongation factor *tuf* gene copies.

Keywords: gene conversion, duplication, elongation factor, evolution, protein-RNA interaction

*Corresponding author: Kondrashov, F.A. (fkondrashov@ucsd.edu)

ABSTRACT

An extra copy of a gene is though to provide some functional redundancy, leading to a higher rate of evolution in duplicated genes. We estimate the impact of gene duplication on selection in *tuf* elongation factor paralogs and find that without gene conversion, they evolved significantly slower compared with when gene conversion has been a factor in their evolution. Thus, *tuf* gene copies evolve under selection that ensures their functional uniformity, and gene conversion reduces selection against amino acid substitutions that effect the function of the EF-Tu protein.

Introduction

Gene duplications are routinely assumed to evolve under weaker selection than nonduplicated genes because extra gene copies are thought to be, at least partially, redundant [1-3]. Such functional redundancy, which may take place regardless of a fitness increase associated with a gene duplication [4], leads to a relaxation of selection and a subsequent acceleration of evolution in recently duplicated genes [3-7]. However, commonly used methods of investigating the acceleration of the rate of evolution in duplicated genes are not applicable to individual gene families. Here, we employ a new method of investigating the strength and mode of selection acting on duplicated genes by analyzing the impact of paralogous gene conversion on the rate of evolution of gene copies. We performed this comparison on the paralogs of the elongation factor *tuf* gene (EF-Tu protein) in gamma proteobacteria. EF-Tu plays a crucial role in protein synthesis by binding aminoacylated tRNAs to the ribosome, which is essential in all living organisms [8-10].

Gene conversion and selection

The impact of gene conversion, a form of recombination, on the evolution of the involved gene copies is well known [11-15]. Conversion between gene copies prevents their divergence, leading to concerted evolution [3,11], and will occur until gene copies become too diverged, such that recombination cannot

occur between them [8,14]. This process can either increase the strength of selection acting on gene copies, or weaken it. If a gene duplication does not bring enough fitness advantage (is partially redundant), the strength of selection acting on two copies of the gene is expected to be relaxed compared to that of the same single copy gene [4]. Gene conversion can reduce such relaxation of selection in a duplicated gene by spreading “degenerate” mutations (mutations that are benign in one copy but deleterious if present in both gene copies, which is a direct result of genetic redundancy [2,5]), thereby increasing the efficiency of selection against such mutations. Alternatively, gene conversion can eliminate such degenerate mutations by gene converting the original copy over the copy that carries the degenerate mutation.

Gene conversion can also reduce the strength of selection acting on the gene copies. This may occur if there is selection against functional divergence of gene copies, which may occur, for example, with evolution of expression rate of interacting protein subunits [16]. Thus, if a mutation occurring in one of the gene copies is deleterious, but is benign as long as it occurs at the same time in both copies, gene conversion can relax selection acting against such mutations. If fitness increases with the number of gene copies in a non-epistatic fashion there will be no genetic redundancy, such that the strength of selection against deleterious mutations will be the same regardless of the number of gene copies [4]. Therefore, gene conversion will have no impact on the selection pressure in gene duplicates as long as the duplication event increased fitness in a non-

epistatic fashion [4]. In essence, frequent gene conversion ensures that gene copies evolve as a single unit, such that selection and the rate of evolution in the two converting copies would resemble that of a single copy.

Thus, a comparison in the rate of evolution in lineages with and without gene conversion can reveal the change in the mode and strength of selection on gene copies after a gene duplication event. Two highly similar copies of the *tuf* gene, *tufA* and *tufB* were found in *Escherichia coli* and *Salmonella typhimurium* [17,18], suggesting that this gene undergoes gene conversion. A more recent phylogenetic analysis of these genes in several complete bacterial genomes confirmed that the *tuf* gene undergoes gene conversion in proteobacteria [19]. Since the publication of this study [19], the number of available complete bacterial genomes substantially increased, enabling a quantitative analysis of this process.

Faster rate of evolution in nonindependently evolving gene copies

We constructed a phylogeny of the *tuf* gene in gamma proteobacteria (see supplementary material online), which revealed that *tuf* gene copies from the same genome cluster together, indicating a high rate of gene conversion (Supplementary Figure 1, [1,19]). Furthermore, the *tuf* gene copies were found in the same syntenic region (Supplementary Figure 2, [2,19]) implying that the phylogenetic pattern observed here (Figure 1, Supplementary Figure 1) is

unlikely to be the product of multiple independent gene duplications. Although the deep branches of the phylogeny showed weak bootstrap support, terminal branches of the tree were reconstructed reliably, making it possible to map the most recent gene conversion events on the tree (Figure 1, Supplementary Figure 1). All branches in the phylogeny could then be partitioned into two sets: a) branches that represent evolution that occurred since the last gene conversion event (red branches in Figure 1), and b) branches that represent the evolutionary history before the most recent gene conversion event (black branches in Figure 1).

We then measured the rates of synonymous (ds) and nonsynonymous (dn) evolution on these two types of branches of the tree using the PAML package ([20], see supplementary methods online). First, we used the rate of evolution along the branches that have not been affected by gene conversion to estimate the rate of gene conversion. The average number of synonymous substitutions per site between gene copies along these branches was 0.030; assuming that both synonymous substitutions and gene conversion events are (nearly) neutral, the rate of gene conversion (c) is approximately 30 times faster than the point mutation rate (μ) ($ds = \mu/c$, see [15]), which is comparable to the rate seen in other species [15, 21]. The average number of nonsynonymous substitutions along the same branches evolution was ~1.5, such that two gene copies diverge by only 1.5 amino acid substitutions before they undergo gene conversion. Such a rapid rate of gene conversion in *tuf* genes suggests that most

amino acid substitutions that occurred along the branch segments corresponding to evolution before the last gene conversion event have been subject to selection in both gene copies.

Next we compared the number of nonsynonymous and synonymous substitutions to estimate the strength of selection acting in branches before and after the last gene conversion event. Instead of averaging the dn/ds values from each branch, we estimated the total number of nonsynonymous and synonymous substitutions in the two types of phylogenetic branch segments ([22], see supplementary methods online). We found that the number of amino acid substitutions in the branches after the last gene conversion event was significantly lower (Table 1), and the dn/ds ratios were equivalent to 0.0577 in branch segments without gene conversion and to 0.0758 in branch segments with gene conversion ($p < 0.010$, Fisher's exact test), indicating stronger selection against amino acid substitutions when gene copies evolve independently. This result is contrary to what is expected under the usual assumption of the redundancy of gene duplications [2,5]. We observe that when the gene copies are accumulating amino acid substitutions independently they are more conserved. However, this result is expected if there is selection to maintain similar functional characteristics in both gene copies. In that case, the selection pressure against functionally important amino acid substitutions occurring in independently evolving genes can be alleviated by gene conversion maintaining sequence and functional homogeneity of the converting gene copies.

Because the EF-Tu protein binds to the aminoacylated tRNAs [8-10], the amino acid residues on the protein-tRNA interface may be particularly important to the functional properties of EF-Tu. Thus, we compared the rate of protein evolution at the protein-tRNA interface (using three available crystal structures of protein-tRNA complexes, Supplementary Figure 3, see supplementary methods online [23-25]) on the branches representing the rate of evolution after the last gene conversion versus those branches representing the rate of evolution before the last gene conversion event. We estimated the number of amino acid substitutions inside and outside the protein-tRNA interface in the two types of branches, confirming that amino acid substitutions in the protein-tRNA interface occurred less frequently when the two gene copies were evolving independently, than if their evolution was affected by gene conversion (Table 1).

A novel mode of selection?

The acceleration of evolution in non-independently evolved gene copies, preferentially affecting the protein-RNA interface, suggests that the *tuf* gene copies are subject to a different type of selection than is generally expected in duplicated genes. It appears that the *tuf* elongation factor genes are evolving under selection that is aimed at functional homogenization of its copies, preventing their sequence divergence. Thus, some substitutions that are deleterious if present in one copy, would be (nearly) neutral when present in both copies simultaneously. This selection may be based on the functional interaction

of proteins coded by diverging gene copies; to our knowledge, such selection in duplicated genes has yet to have been described. There may be several reasons why functional heterogeneity of EF-Tu proteins is deleterious. Different tRNA-binding coefficients of EF-Tu copies may lead to inefficient regulation of their expression, or cause translation rate heterogeneity. The subfunctionalization of EF-Tu gene copies that may lead the two copies to preferentially function with different tRNAs may be particularly deleterious if the different copies interfere with the function of the other copy through the formation of nonfunctional protein complexes with the non-specific RNA. Thus, selection for functional uniformity may also be thought of as selection against subfunctionalization (functional specialization). We expect for such selection for functional uniformity to be particularly strong in gene copies where independent functional or expressional regulation is not straightforward, such as when gene copies are located in different operons, which is the case with the Tu elongation factor (17-19).

Methods

Rationale for our comparison

Nonorthologous gene conversion is a form of recombination that leads to the replacement of one DNA sequence with another [26]. This process appears to occur rapidly [27], and leads to concerted evolution, which is the observation of non-independent evolution of paralagous sequences [26]. On a phylogenetic

tree, gene conversion events look identical to independent duplications (see Supplementary Figure 1), however, further support for gene conversion can be obtained through a comparison of synteny (Supplementary Figure 2). It is expected for the two sequences to diverge slightly before a conversion event reverts the diverging sequences to an identical state [28,29] involving either the entire or just a fraction of the paralogous copies [26,28]. Thus, novel mutations that arise in one of the copies can either be removed by conversion by reverting to the original state or may spread to other copies by the reciprocal gene conversion event.

Gene conversion can change the nature of selection acting on mutations arising in gene copies [29,30]. Consider a haploid organism with two gene copies with an extremely fast rate of gene conversion. Since a novel mutation would either rapidly disappear or spread to the other copy through gene conversion, only two genotypes (both copies mutant or wild-type) would be visible to selection. Thus, in a system with an infinite rate of gene conversion, selection is equivalent to that acting on a single gene copy since novel mutants would be screened by selection in both copies simultaneously [31,32]. In a haploid system with a relatively fast but not an instantaneous rate of gene conversion there will be a time period between the emergence of a novel mutant and a gene conversion event, such that the intermediate genotype (one copy mutant and one copy wild type) may persist or even fix in the population. Thus, substitutions fixed in the population without the influence of gene conversion will be subject to

selection pressure that considers the state of the two gene copies independently, while substitutions that have been converted and then fixed must have been subject to selection pressure that considered their impact on both copies simultaneously.

Phylogeny and genome location of the tuf gene

We assembled all *tuf* genes from all the complete genomes of gamma proteobacteria available in GenBank [33] Majority (47 out of 56) of the gamma proteobacteria genomes contained two copies of the gene, while no genomes with three copies of the gene were found. To make sure that no copies of the *tuf* gene were omitted, we performed TBLASTN [34] searches of the *tuf* protein sequence against the complete genomes. We found no evidence to suggest that any *tuf* gene copies were missed in our analysis.

Gene conversion can maintain a high sequence similarity between gene copies. A phylogenetic comparison of gene copies from several genomes is often used to reveal gene conversion, with a clustering of gene copies on the branches of the phylogenetic tree in multiple closely related species [27]. We reconstructed the phylogeny of *tuf* genes from the 56 complete genomes of gamma subdivision of proteobacteria using a neighbor joining approach as implemented in MEGA3 [35] with 10 000 bootstrap replicates, and a Bayesean approach as implemented in MrBayes [36] with 1 million iterations (mcmc ngen = 1000000 in MrBayes)

using the General Time Reversible model [36]. Both methods yielded congruent phylogenies. All gene copies of the *tuf* gene tended to cluster together on the branch tips of the phylogenetic tree (Supplementary Figure 1) indicating that they undergo rapid concerted evolution in gamma proteobacteria. While the accuracy of the phylogenetic reconstruction is questionable for deep branches of the tree, the reconstruction of the terminal branches is robust, as shown by the high bootstrap values on the phylogeny (Supplementary Figure 1).

Genomic location

One of the ways to rule out the possibility of multiple independent recent gene duplications, which in theory can produce the same phylogenetic pattern, is to analyze the genomic location (synteny) of the gene copies [37]. Genomic location of gene duplicates is not expected to change in the course of gene conversion, while independent gene duplications should place the new non-tandem copies in a random location. While it is unlikely that so many independent duplications occurred in the case of the *tuf* gene, we nevertheless looked at the genomic location of *tuf* genes in an attempt to find evolutionary changes of genomic location in the course of gene conversion.

Both of the *tuf* gene genomic locations are conserved within genomic regions that code for proteins with translation-related function (Supplementary Figure 2, [37]), with the *tufA* gene being close to the *fusA* gene coding for the

elongation factor EF-G and the *tufB* gene being closer to the *CoA* gene. In two cases, we have observed a change in the genomic location of one of the *tuf* genes, which is most likely related to the mechanism of conversion [38].

Estimating the rate of evolution

We then estimated the rates of nonsynounymous (dn) and synonymous (ds) substitution rates on all of the branches of the phylogenetic tree using the PAML package [39]. We only used branches with ds < 1.0 when reporting the rate of evolution in branches representing evolution before and after gene conversion. An accumulation of slightly deleterious polymorphisms may skew the evolutionary dn/ds estimates in terminal branches [40,41], which were overwhelmingly branches representing evolution occurring before a gene conversion event. However, the direction of this effect is opposite to the difference between terminal and deeper branches that we report here [40,41], and cannot explain our results. Instead of the average rate of evolution (average(dn/ds) = average(N/Ns / S/Ss)), where N and S are the nonsynonymous and synonymous substitutions per gene, respectively, and Ns and Ss are the number of nonsynonymous and synonymous sites, we looked at dn/ds = sum(N)/sum(Ns) / sum(S)/sum(Ss). This approach is appropriate when for many of the averaged ratios the numerator or the denominator is close to zero [42]. To estimate the number of substitutions occurring on the branch segments not affected by gene conversion we estimated each clade separately (species

connected by red or blue branches in Supplementary Figure 1). To estimate the number of substitutions on the deeper branches effected by gene conversion (black branches in Supplementary Figure 1) we allowed different dn/ds ratios (model=2 parameter in PAML control file) for the branches effected and not effected by gene conversion.

We used three crystal structures of the tRNA- protein complex (PDB ids: 1b23, 1ob2 and 1ob5; [43-45]) to compare the rate of evolution between the tRNA-protein interface versus the rest of the protein globula. We defined the protein-tRNA interface by selecting residues that were within 4Å of any tRNA nucleotide in any of the three available structures and nearest neighbors of contact residues in the protein chain (Supplementary Figure 3). We used PAML [39] to estimate the number of substitutions occurring along different types of branches separately for the tRNA-protein interface and the rest of the protein. A phylogeny-independent method yielded similar results (data not shown).

Acknowledgements

The authors thank Peter Andolfatto, Doris Bachtrog, Robert Cutler, Hideki Innan, Eugene Koonin, Alexey Kondrashov and Martin Lercher for comments on the manuscript and discussions of the interplay between gene conversion and selection.

Table 1. The influence of gene conversion on the number of substitutions in *tuf* gene copies

	Branch segments not affected by gene conversion	Branch segments affected by gene conversion	Significance by Fisher's exact test
Number of synonymous substitutions in the entire protein	615.6	6101.3	$P < 0.010$
Number of nonsynonymous substitutions in the entire protein	92.8	1205.3	$P < 0.010$
Number of nonsynonymous substitutions in the protein–tRNA interface	0	91	$P < 0.0022$
Number of nonsynonymous substitutions outside the protein–tRNA interface	80.7	1113.4	$P < 0.0022$

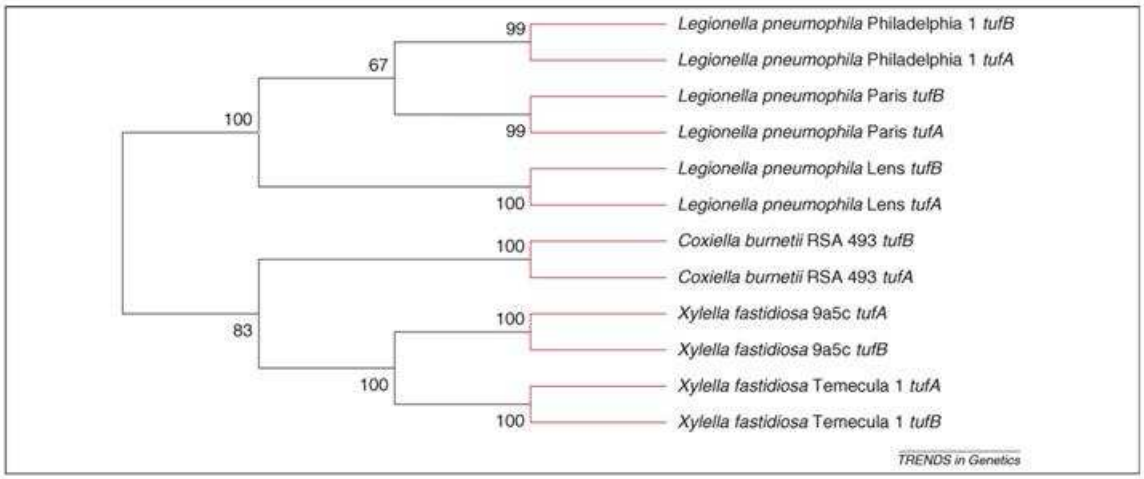


Figure 1. Neighbor-joining *tuf* gene phylogeny with bootstrap values, for selected species of γ-proteobacteria. Red branch segments denote evolution that occurred in the absence of gene conversion, whereas black branch segments denote evolutionary history before the most recent gene conversion event.

Figure S1. *tuf* gene phylogeny in 56 gamma proteobacteria species. Red and blue branches represent evolution that occurred without the action of gene conversion, while black branches represent the evolutionary history before the most recent gene conversion event. Red branches signify cases where gene conversion occurred since the last speciation event, while blue branches signify cases where at least one speciation event occurred after the most recent gene conversion event.

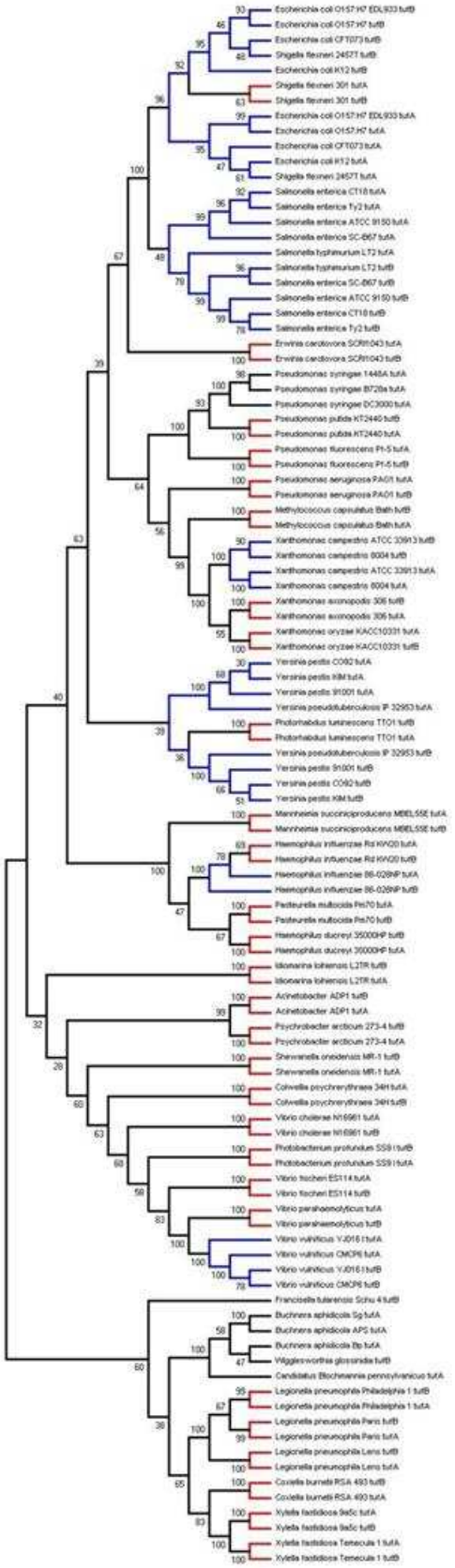
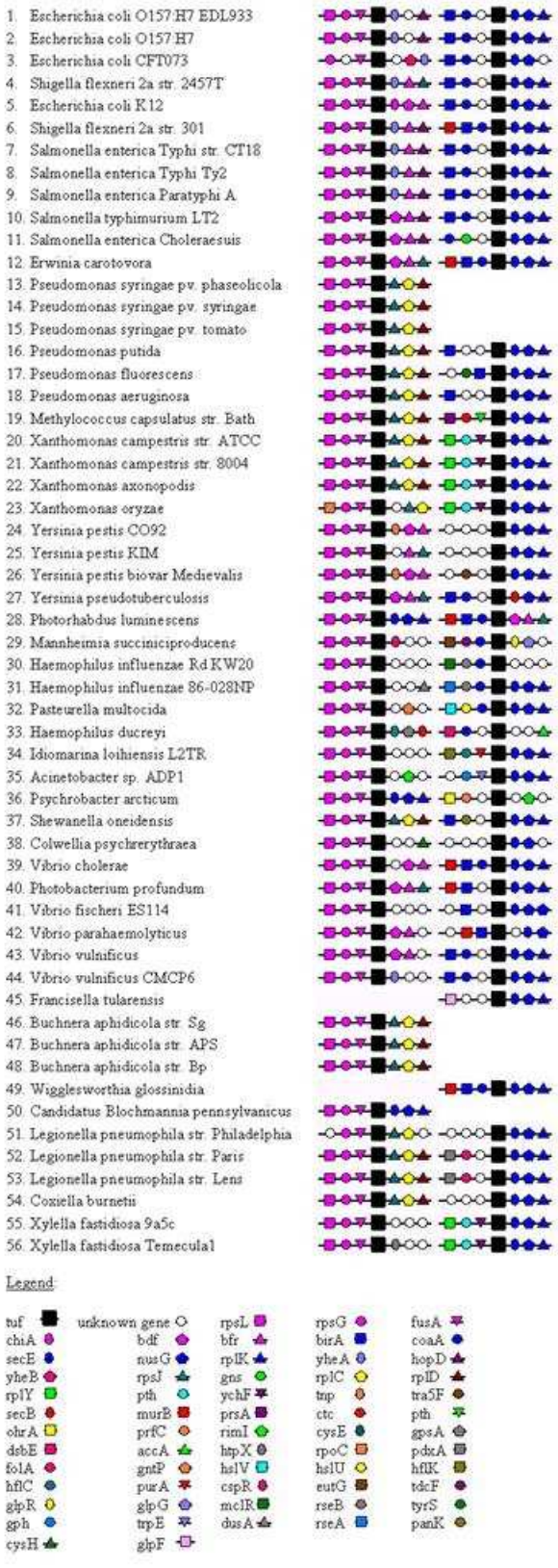

Escherichia coli O157:H7 EDL933 tufB
Escherichia coli O157:H7 tufB
Escherichia coli CFT073 tufB
Shigella flexneri 2457T tufB
Escherichia coli K12 tufB
Shigella flexneri 301 tufA
Shigella flexneri 301 tufB
Escherichia coli O157:H7 EDL933 tufA
Escherichia coli O157:H7 tufA
Escherichia coli CFT073 tufA
Escherichia coli K12 tufA
Shigella flexneri 2457T tufA
Salmonella enterica CT18 tufA
Salmonella enterica Ty2 tufA
Salmonella enterica ATCC 9150 tufA
Salmonella enterica SC-B67 tufA
Salmonella typhimurium LT2 tufA
Salmonella typhimurium LT2 tufB
Salmonella enterica SC-B67 tufB
Salmonella enterica ATCC 9150 tufB
Salmonella enterica CT18 tufB
Salmonella enterica Ty2 tufB
Erwinia carotovora SCRI1043 tufA
Erwinia carotovora SCRI1043 tufB
Pseudomonas syringae 1448A tufA
Pseudomonas syringae B728a tufA
Pseudomonas syringae DC3000 tufA
Pseudomonas putida KT2440 tufB
Pseudomonas putida KT2440 tufA
Pseudomonas fluorescens Pf-5 tufA
Pseudomonas fluorescens Pf-5 tufB
Pseudomonas aeruginosa PAO1 tufA
Pseudomonas aeruginosa PAO1 tufB
Methylococcus capsulatus Bath tufB
Methylococcus capsulatus Bath tufA
Xanthomonas campestris ATCC 33913 tufB
Xanthomonas campestris 8004 tufB
Xanthomonas campestris ATCC 33913 tufA
Xanthomonas campestris 8004 tufA
Xanthomonas axonopodis 306 tufB
Xanthomonas axonopodis 306 tufA
Xanthomonas oryzae KACC10331 tufA
Xanthomonas oryzae KACC10331 tufB
Yersinia pestis CO92 tufA
Yersinia pestis KIM tufA
Yersinia pestis 91001 tufA
Yersinia pseudotuberculosis IP 32953 tufA
Photorhabdus luminescens TTO1 tufB
Photorhabdus luminescens TTO1 tufA
Yersinia pseudotuberculosis IP 32953 tufB
Yersinia pestis 91001 tufB
Yersinia pestis CO92 tufB
Yersinia pestis KIM tufB
Mannheimia succiniciproducens MBEL55E tufA
Mannheimia succiniciproducens MBEL55E tufB
Haemophilus influenzae Rd KW20 tufA
Haemophilus influenzae Rd KW20 tufB
Haemophilus influenzae 86-028NP tufA
Haemophilus influenzae 86-028NP tufB
Pasteurella multocida Pm70 tufA
Pasteurella multocida Pm70 tufB
Haemophilus ducreyi 35000HP tufB
Haemophilus ducreyi 35000HP tufA
Idiomarina loihiensis L2TR tufB
Idiomarina loihiensis L2TR tufA
Acinetobacter ADP1 tufB
Acinetobacter ADP1 tufA
Psychrobacter arcticum 273-4 tufB
Psychrobacter arcticum 273-4 tufA
Shewanella oneidensis MR-1 tufB
Shewanella oneidensis MR-1 tufA
Colwellia psychrerythraea 34H tufA
Colwellia psychrerythraea 34H tufB
Vibrio cholerae N16961 tufA
Vibrio cholerae N16961 tufB
Photobacterium profundum SS9 tufB
Photobacterium profundum SS9 tufA
Vibrio fischeri ES114 tufA
Vibrio fischeri ES114 tufB
Vibrio parahaemolyticus tufA
Vibrio parahaemolyticus tufB
Vibrio vulnificus YJ016 tufA
Vibrio vulnificus CMCP6 tufA
Vibrio vulnificus YJ016 tufB
Vibrio vulnificus CMCP6 tufB
Francisella tularensis Schu 4 tufB
Buchnera aphidicola Sg tufA
Buchnera aphidicola APS tufA
Buchnera aphidicola Bp tufA
Wigglesworthia glossinidia tufB
Candidatus Blochmannia pennsylvanicus tufA
Legionella pneumophila Philadelphia 1 tufB
Legionella pneumophila Philadelphia 1 tufA
Legionella pneumophila Paris tufB
Legionella pneumophila Paris tufA
Legionella pneumophila Lens tufB
Legionella pneumophila Lens tufA
Coxiella burnetii RSA 493 tufB
Coxiella burnetii RSA 493 tufA
Xylella fastidiosa 9a5c tufA
Xylella fastidiosa 9a5c tufB
Xylella fastidiosa Temecula 1 tufA
Xylella fastidiosa Temecula 1 tufB

Figure S2. Synthenic region around the *tuf* gene copies.


1. Escherichia coli O157:H7 EDL933
2. Escherichia coli O157:H7
3. Escherichia coli CFT073
4. Shigella flexneri 2a str. 2457T
5. Escherichia coli K12
6. Shigella flexneri 2a str. 301
7. Salmonella enterica Typhi str. CT18
8. Salmonella enterica Typhi Ty2
9. Salmonella enterica Paratyphi A
10. Salmonella typhimurium LT2
11. Salmonella enterica Choleraesuis
12. Erwinia carotovora
13. Pseudomonas syringae pv. phaseolicola
14. Pseudomonas syringae pv. syringae
15. Pseudomonas syringae pv. tomato
16. Pseudomonas putida
17. Pseudomonas fluorescens
18. Pseudomonas aeruginosa
19. Methylococcus capsulatus str. Bath
20. Xanthomonas campestris str. ATCC
21. Xanthomonas campestris str. 8004
22. Xanthomonas axonopodis
23. Xanthomonas oryzae
24. Yersinia pestis CO92
25. Yersinia pestis KIM
26. Yersinia pestis biovar Medievalis
27. Yersinia pseudotuberculosis
28. Photorhabdus luminescens
29. Mannheimia succiniciproducens
30. Haemophilus influenzae Rd KW20
31. Haemophilus influenzae 86-028NP
32. Pasteurella multocida
33. Haemophilus ducreyi
34. Idiomarina loihiensis L2TR
35. Acinetobacter sp. ADP1
36. Psychrobacter arcticum
37. Shewanella oneidensis
38. Colwellia psychrerythraea
39. Vibrio cholerae
40. Photobacterium profundum
41. Vibrio fischeri ES114
42. Vibrio parahaemolyticus
43. Vibrio vulnificus
44. Vibrio vulnificus CMCP6
45. Francisella tularensis
46. Buchnera aphidicola str. Sg
47. Buchnera aphidicola str. APS
48. Buchnera aphidicola str. Bp
49. Wigglesworthia glossinidia
50. Candidatus Blochmannia pennsylvanicus
51. Legionella pneumophila str. Philadelphia
52. Legionella pneumophila str. Paris
53. Legionella pneumophila str. Lens
54. Coxiella burnetii
55. Xylella fastidiosa 9a5c
56. Xylella fastidiosa Temecula1
Legend
tuf
chiA
secE
yheB
rplY
secB
ohrA
dsbE
folA
hflC
glpR
gph
cysH
unknown gene
bdf
nusG
rpsJ
pth
murB
prfC
accA
gntP
purA
glpG
trpE
glpF
rpsL
bfr
rplK
gns
ychF
prsA
rimI
htpX
hslV
cspR
mclR
dusA
rpsG
birA
yheA
rplC
tnp
ctc
cysE
rpoC
hslU
eutG
rseB
rseA
fusA
coaA
hopD
rplD
tra5F
pth
gpsA
pdxA
hflK
tdcF
tyrS
panK

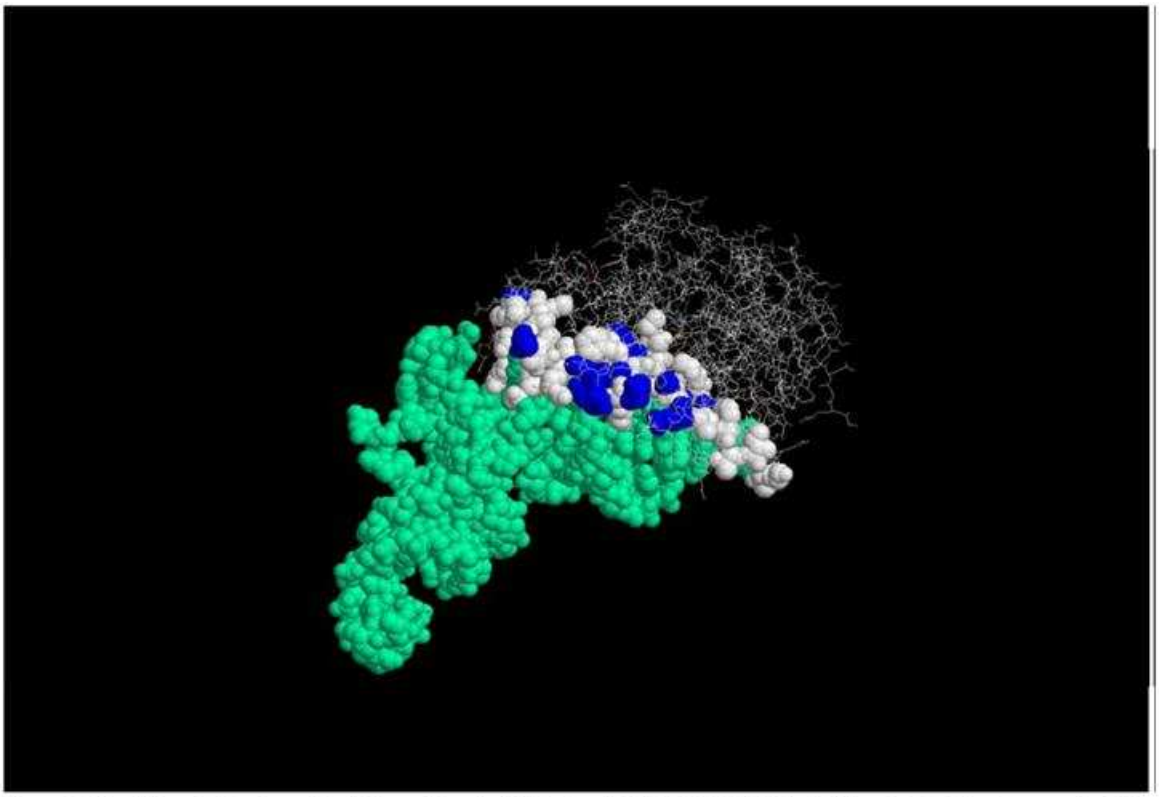


Figure S3. The RNA-protein interface of the EF-Tu protein and a tRNA molecule (PDB id: 1b23) [18]. The tRNA molecule is shown in green, the RNA-protein interface of the EF-Tu protein is shown in blue and white, such that the residues where substitutions that occurred in the course of evolution prior to the most recent gene conversion event are shown in blue.

References

1. Ohno, S. (1970) Evolution by Gene Duplication. Berlin-Heidelberg-New York: Springer-Verlag.

2. Lynch, M. and Force, A. (2000) The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459-473.

3. Li, W. H. (1997) Molecular Evolution. Sunderland, MA: Sinauer.

4. Kondrashov, F. A. *et al.* (2002) Selection in the evolution of gene duplications. Genome Biol. 3: RESEARCH0008.

5. Lynch, M. and Conery, J. S. (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151-1155.

6. Davis, J. C. and Petrov, D. A. (2004) Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol. 2: E55.

7. Jordan, I. K. *et al.* (2004) Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evol Biol. 4:22.

8. Stark, H. *et al.* (1997) Visualization of elongation factor Tu on the Escherichia coli ribosome. Nature. 389: 403-406.

9. Stark, H. *et al*. (2002) Ribosome interactions of aminoacyl-tRNA and elongation factor Tu in the codon-recognition complex. Nat Struct Biol. 9: 849-854.

10. Nilsson, J. and Nissen, P. (2005) Elongation factors on the ribosome. Curr Opin Struct Biol. 15:349-54

11. Walsh JB. (1987) Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion? Genetics. 117: 543-557.

12. Ohta, T. (1990) How gene families evolve. Theor Popul Biol 37:213- 219.

13. Innan, H. (2003) A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. Proc Natl Acad Sci U S A. 100: 8793-8798.

14. Teshima, K. M. and Innan, H. (2004) The effect of gene conversion on the divergence between duplicated genes. Genetics. 166: 1553-1560.

15. Gao, L. Z. and Innan, H. (2004) Very low gene duplication rate in the yeast genome. Science. 306: 1367-1370.

16. Papp, B. *et al.* (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature. 424: 194-197.

17. Abdulkarim, F. and Hughes, D. (1996) Homologous recombination between the tuf genes of Salmonella typhimurium. J Mol Biol. 260: 506-22.

18. Zuurmond, A. M. *et al.* (1999) Either of the chromosomal tuf genes of E. coli K-12 can be deleted without loss of cell viability. Mol Gen Genet. 260: 603-607.

19. Lathe, W. C. 3rd. and Bork, P. (2001) Evolution of tuf genes: ancient duplication, differential loss and gene conversion. FEBS Lett. 502: 113-116.

20. Yang, Z. (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13: 555-556.

21. Innan, H. (2003) The coalescent and infinite-site model of a small multigene family. Genetics. 163: 803-810.

22. Smith, N. G. and Eyre-Walker, A. (2002) Adaptive protein evolution in Drosophila. Nature. 415: 1022-1024.

23. Nissen, P. *et al.* (1995) Crystal structure of the ternary complex of Phe-tRNAPhe, EF-Tu, and a GTP analog. Science. 270: 1464-1472.

24. Kristensen, O. *et al.* (1996) Isolation, crystallization and X-ray analysis of the quaternary complex of Phe-tRNA(Phe), EF-Tu, a GTP analog and kirromycin. FEBS Lett. 399: 59-62.

25. Parmeggiani, A. *et al.* (2006) Enacyloxin IIa pinpoints a binding pocket of elongation factor Tu for development of novel antibiotics. J Biol Chem. 281: 2893-900.

26. Li, W. H. (1997) Molecular Evolution. Sunderland, MA: Sinauer.

27. Gao LZ, Innan H. (2004) Very low gene duplication rate in the yeast genome. Science. 306: 1367-70.

28. Walsh JB. (1987) Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion? Genetics. 117: 543-557.

29. Teshima, K. M. and Innan, H. (2004) The effect of gene conversion on the divergence between duplicated genes. Genetics. 166: 1553-1560.

30. Hurst, L. D. and Smith, N. G. C. (1998) The evolution of concerted evolution. Proc. Royal Soc. London B Biol. Sci. 265: 121-127

31. Lynch, M. and Force, A. (2000) The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459-473.

32. Kondrashov, F. A. *et al.* (2002) Selection in the evolution of gene duplications. Genome Biol. 3: RESEARCH0008.

33. Benson, D. A. *et al.* (2006) GenBank. *Nucleic Acids Res.* 34: D16-20.

34. Altschul, S. F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.

35. Kumar, S. *et al.* (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief. Bioinform.* 5: 150-163.

36. Ronquist, F., and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572-1574.

37. Lathe, W.C. 3rd, Bork, P. (2001) Evolution of tuf genes: ancient duplication, differential loss and gene conversion. FEBS Lett. 502: 113-116.

38. Hughes, D. (2000) Co-evolution of the tuf genes links gene conversion with the generation of chromosomal inversions. J Mol Biol. 297: 355-364.

39. Yang, Z. (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13: 555-556.

40. Golding, G.B. (1987) The detection of deleterious selection using ancestors inferred from a phylogenetic history. Genet Res. 49: 71-82.

41. Golding, B. and Felsenstein, J. (1990) A maximum likelihood approach to the detection of selection from a phylogeny. J Mol Evol. 31: 511-23.

42. Smith, N.G. and Eyre-Walker, A. (2002) Adaptive protein evolution in Drosophila. Nature. 415:1022-1024.

43. Stark, H. *et al.* (1997) Visualization of elongation factor Tu on the Escherichia coli ribosome. Nature. 389: 403-406.

44. Kristensen, O. *et al.* (1996) Isolation, crystallization and X-ray analysis of the quaternary complex of Phe-tRNA(Phe), EF-Tu, a GTP analog and kirromycin. FEBS Lett. 399: 59-62.

45. Parmeggiani, A. *et al.* (2006) Enacyloxin IIa pinpoints a binding pocket of elongation factor Tu for development of novel antibiotics. J Biol Chem. 281: 2893-900.

Chapter 6, in full, is a manuscript of the material as it appears in Kondrashov FA , Gurbich TA and Vlasov PK (2007). Selection for functional uniformity of tuf duplicates in gamma-proteobacteria. Trends in Genetics **23**, 215-218. Elsevier Ltd. 2007. The dissertation author was the primary investigator and author of this paper.

Chapter 7.

Nested genes and increasing organizational complexity of metazoan genomes

Nested genes and increasing organizational complexity of metazoan genomes

Raquel Assis[1], Alexey S. Kondrashov[1], Eugene V. Koonin[2] and Fyodor A. Kondrashov[3]

[1]Center for Computational Medicine and Biology and the Life Sciences Institute, University of Michigan, Ann Arbor MI, 48109 USA.

[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda MD, 20894 USA.

[3]Section on Ecology, Behavior and Evolution, Division of Biological Sciences, University of California at San Diego, 2218 Muir Biology Building, La Jolla CA, 92093 USA.

Corresponding author: Kondrashov, F.A. (fkondrashov@ucsd.edu)

The most common form of protein-coding gene overlap in eukaryotes is a simple nested structure, whereby one gene is embedded in an intron of another. Analysis of these structures in vertebrates, fruit flies, and nematodes revealed substantially higher rates of evolutionary gains than losses that could not be attributed to any obvious functional relationships between nested gene members. Thus, accumulation of

nested gene structures seems to be increasing genome organizational complexity via a neutral process.

Eukaryotes are typically more complex than prokaryotes on the molecular, systems, and phenotypic scales of biological organization. In particular, genomes of multicellular eukaryotes possess a complex architecture that involves substantial overlapping of their transcribed regions [1-3] and protein-coding genes [4-6], forming an interleaving mosaic of exon and intron sequences. Although it is clear that such complex genome organization is made possible by the presence of introns, the rates and mechanisms of evolutionary events leading to gains and losses of overlapping gene arrangements have not been studied previously.

Evolutionary dynamics of nested gene structures

The most common form of overlap between protein-coding genes in eukaryotes is a nested gene structure, and in a majority of such structures, the internal gene lies entirely within one intron of the external gene [5, 6]. Thus, we investigated the evolution of this class of nested gene structures in vertebrates, *Drosophila*, and *Caenorhabditis*. A search of NCBI annotation records yielded 428, 815, 440, and 608 nested gene pairs in *H. sapiens*, *D. melanogaster*, *C. elegans*, and *C. briggsae* genomes, respectively. After eliminating gene pairs that

might have been misannotated (see Supplementary Material), we arrived at sets of 128, 792, 429, and 233 nested gene pairs, respectively. Only a small minority of the protein sequences encoded by internal genes from each of these three major taxa show significant sequence similarity to internal genes products in the other two taxa (data not shown), suggesting that either these structures emerged independently and relatively late during evolution, or that they were extensively and repeatedly lost.

By examining gene annotations and constructing sequence alignments, we identified the closest species with a completely sequenced genome in which each nested gene structure was absent. Absence of the nested structure in an appropriate outgroup species indicates its emergence (gain) in the respective lineage, whereas presence of the nested structure in the outgroup indicates its loss (Figure 1). Gains were found in all three taxa, with the emergence of 55 internal genes in at least 40 independent events in vertebrates, 52 internal genes in at least 48 events in *Drosophila*, and 22 internal genes in as many events in *Caenorhabditis*. The rate of these acquisitions was approximately uniform throughout the course of evolution (Figure 2). In contrast, losses of nested gene structures were much rarer, with no losses in vertebrates, 17 in *Drosophila*, and 2 in *Caenorhabditis*.

Acquisition of nested gene structures

At least four scenarios are plausible for the formation of a nested gene structure: a) An internal gene can evolve by insertion of a DNA sequence into an intron of a pre-existing gene, b) an internal gene can evolve *de novo* from an intronic sequence of a pre-existing gene, c) a gene can become internal after an adjacent gene acquires an additional exon(s), or d) a gene can become internal after fusion of two genes that flank it from the opposite sides (Figure 3).

By comparing the gene structures and encoded protein sequences of internal and external genes to complete gene sets from the respective species, we deduced the mechanisms of formation of vertebrate nested gene structures (Table 1). Nearly all nested gene structures appear to have emerged by insertion of a DNA sequence, which arose by gene duplication or retrotransposition, into an intron of a pre-existing gene. The origin of an internal gene was classified as a retrotransposition when it was intronless in a given species, whereas its non-nested orthologs in a sister species contained introns. A duplication at the DNA level was inferred when both the internal gene and its non-nested ortholog in a sister species had introns. In cases where the internal gene and an non-nested ortholog in a sister species were both intronless, retrotransposition and duplication at the DNA level could not be discriminated. Five internal genes in humans are candidates for *de novo* origin from intron sequences (see supplementary data), including one case with no sequence similarity beyond apes (PLAC4) and another with no similarity beyond old world monkeys (STH) (Table S1). Analysis of the 12 recently sequenced *Drosophila* genomes showed

that the majority of *de novo* genes originate in introns [7]. Consistent with this observation, we found 11 internal genes in *D. melanogaster* with no sequence similarity to any genes in the genome of the closely related *D. yakuba*. We did not identify any nested gene structures that evolved via the remaining two scenarios.

No functional significance of nested gene structures

At least three hypotheses could explain the parallel accumulation of nested gene structures in different taxa. First, a nested structure might confer a selective advantage due to a functional or co-regulatory relationship between its members [8–12]. Second, according to the transcriptional collision model, members of a nested gene structure could interfere with each other's transcription [13, 14], resulting in alternative expression of these genes in different tissues or during different times in development. Finally, acquisition of a nested gene structure could be a neutral process [15–20], driven by the presence of numerous long introns that provide niches for insertion of genes. Each of these hypotheses leads to a distinct prediction about the relationship between the expression of internal and external genes in a nested pair. The functional co-regulation hypothesis predicts a positive correlation between levels of their expression in similar tissues, the transcriptional collision hypothesis predicts a negative correlation, and the neutral hypothesis predicts no correlation.

We compared correlations of gene expression levels in 73 tissues between 109 nested gene pairs and 1000 random sets of 109 adjacent genes in the human genome (see Supplementary Material). Although weak positive correlations were detected in both cases, there was no significant difference between the sets of nested and non-nested genes (mean correlation coefficients were 0.33 ± 0.03 for nested gene pairs and 0.33 ± 0.0008 for non-nested pairs), which is consistent with the neutral hypothesis. The observation that external genes have substantially more and longer introns than average in the respective species (Ref. 6 and Supplementary Material) is also compatible with the neutral hypothesis. Furthermore, examination of the available functional information for nested gene pairs (Table S1) did not reveal any obvious connections [6]. Fixation of originally neutral or even slightly deleterious sequence segments, such as introns and transposable elements, through genetic drift acting in relatively small populations is a common phenomenon in eukaryotic evolution that may be partially responsible for the evolution of complex phenotypes [16–20]. The increase in organizational complexity of intron-rich genomes via emergence of nested gene structures appears to be another facet of this process.

Predicting the course of genome structure evolution

The neutral hypothesis implies that the preferential evolutionary gain of nested gene structures is due to metazoan genomes being far from neutral equilibrium with respect to birth and death of intron-contained genes [16]. We

estimated the rate of acquisition of nested gene structures as approximately 0.4, 0.9, and 0.2 events per million years in the *H. sapiens*, *D. melanogaster*, and *C. elegans* lineages, respectively (see Supplementary Material). Since animal genomes currently contain ~500-800 nested gene pairs, these rates indicate that nested gene structures began to emerge ~1 billion years ago, perhaps concurrent with the substantial intron gain that apparently occurred at the onset of metazoan evolution [21]. The present results suggest that metazoan introns are still far from saturation by internal genes and that the organizational complexity of metazoan genomes will continue to increase for many millions of years via the emergence of new nested gene structures. By the time metazoan genomes reach organizational complexity equilibrium, the overlap of functional elements is expected to be much greater than what we observe in extant taxa and will likely include numerous Russian doll-like nested structures.

Conclusions and perspective

We have shown that the evolution of metazoan genomes is accompanied by a steady rise in the prevalence of nested arrangements of protein-coding genes, leading to increasingly complex genome architectures. In addition to overlaps between protein-coding genes, animal genomes contain numerous complex arrangements involving genes that encode small RNAs. In particular, a substantial fraction of microRNA (miRNA) and small nucleolar RNA genes are either fully contained within introns of protein-coding genes or overlap with

protein-coding exons [22]. MiRNA genes are highly dynamic components of animal genomes, and it will be of major interest to determine whether the trend of increasingly complex genome organization applies to these genes as well.

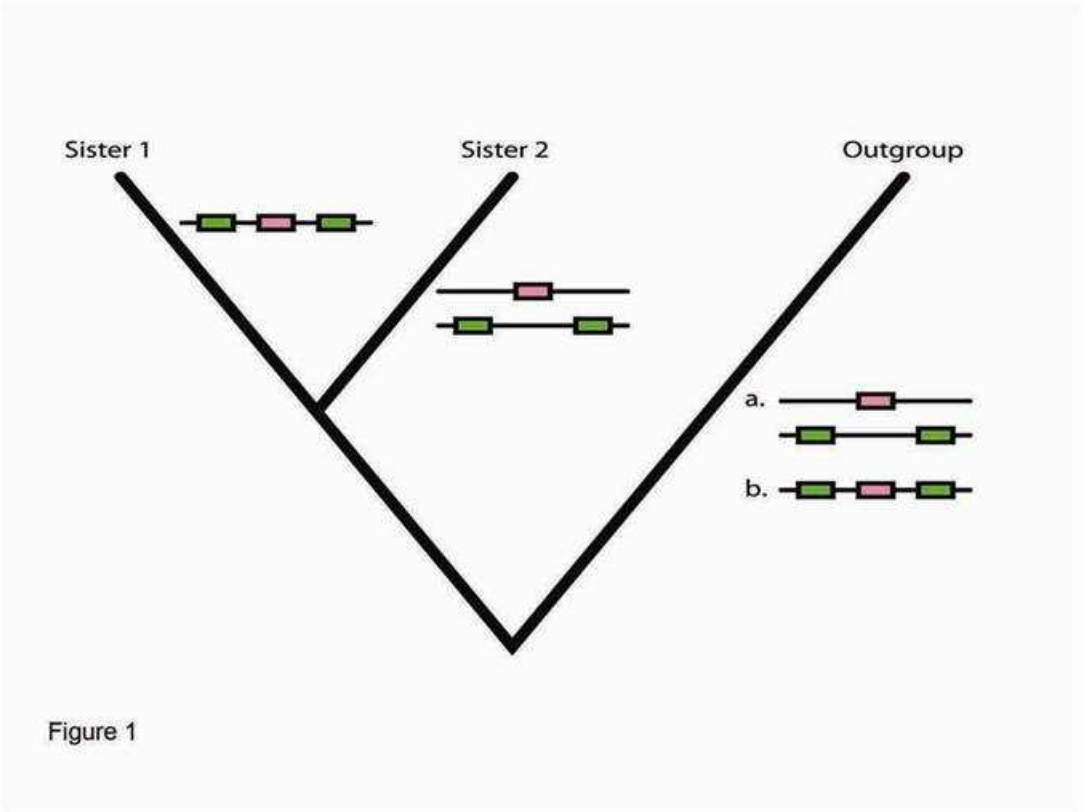
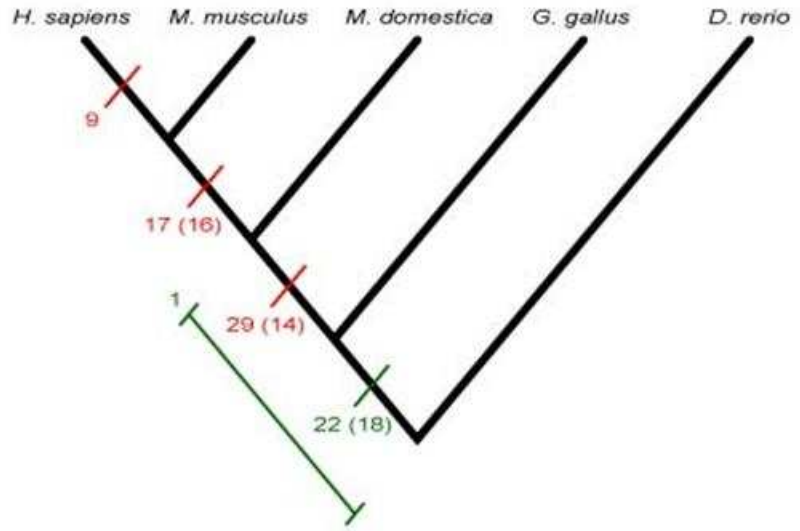


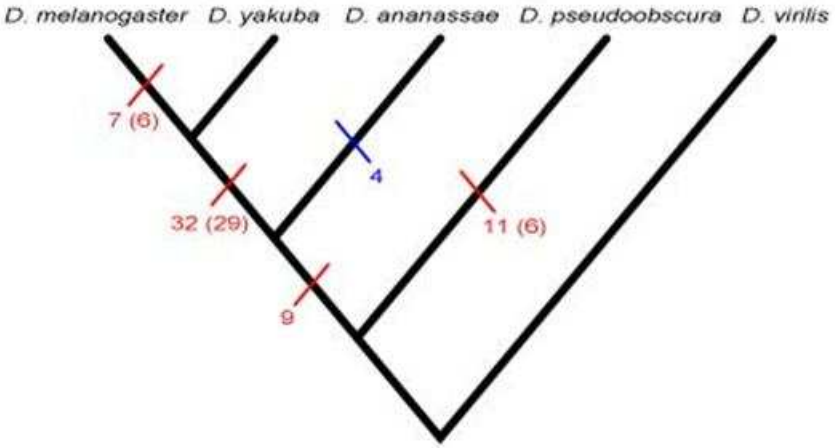
Figure 1. Phylogenetic analysis of gains and losses of nested gene structures. Gain or loss of a nested gene structure must have occurred if, within a pair of sister species, the structure is present in one but absent in the other. a, Absence of the nested structure in the outgroup indicates its gain in sister 1. b, Presence in the outgroup indicates its loss in sister 2.

Figure 2. Dynamics of gain and loss of nested gene structures. Gains and losses of internal genes are labelled on the a) vertebrate, b) *Drosophila* and c) nematode phylogenies in red and blue, respectively. Nested gene structures that have a different nested state in the most distant outgroup, and therefore cannot be resolved between gains or losses, are shown in green. Independent events, or those that occur in different introns, are shown in parentheses. Events that could not be timed with a high enough resolution are shown on the side of each phylogeny.

a.


H. sapiens
M. musculus
M. domestica
G. gallus
D. rerio
9
17 (16)
29 (14)
1
22 (18)

b.


D. melanogaster
D. yakuba
D. ananassae
D. pseudoobscura
D. virilis
7 (6)
32 (29)
4
11 (6)
9

c.

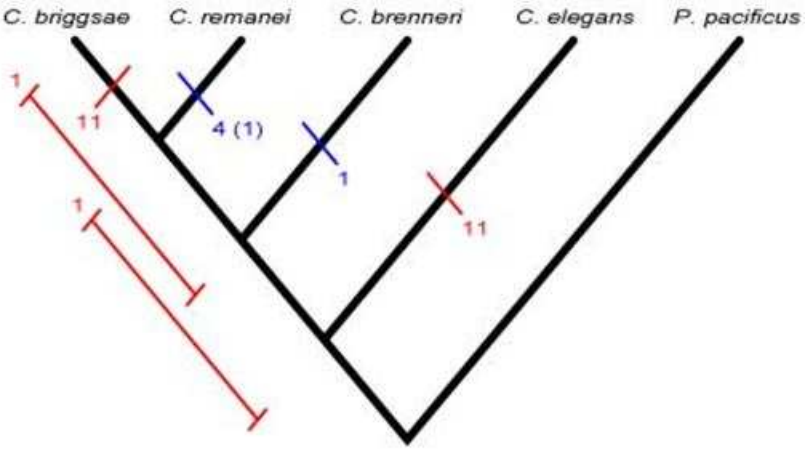
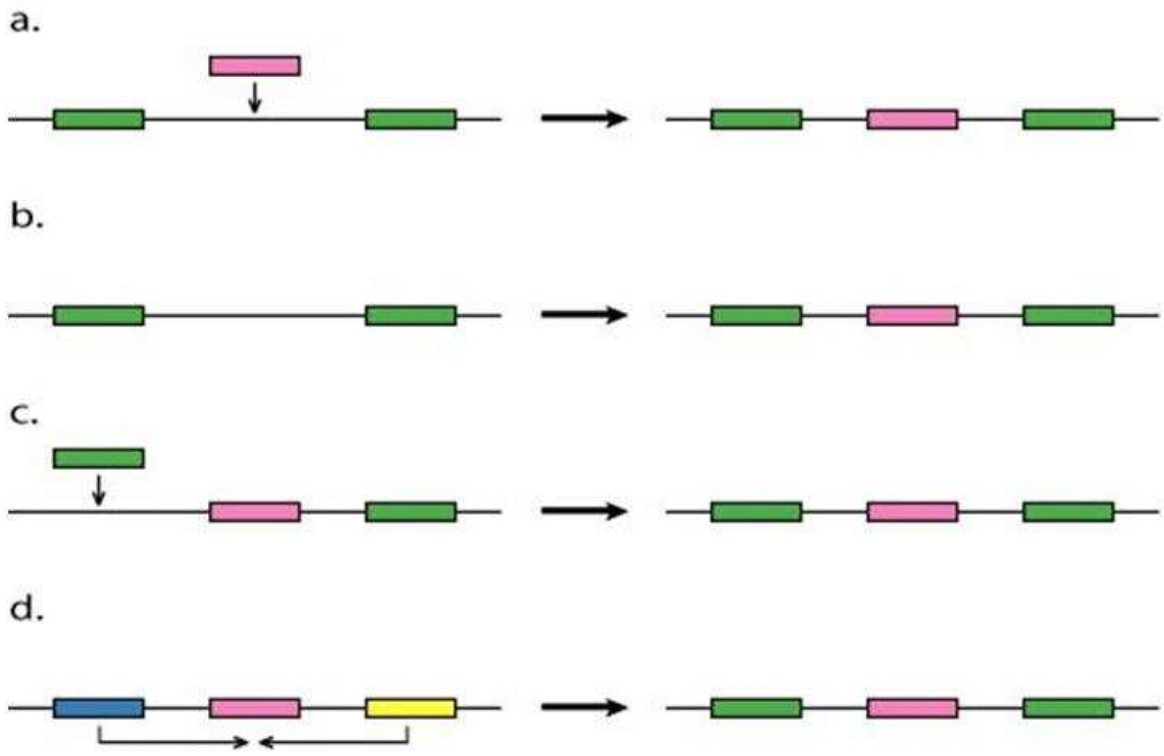

C. briggsae
C. remanei
C. brenneri
C. elegans
P. pacificus
1
11
4 (1)
1
1
11

Figure 3. Scenarios for the origin of a nested gene structure.



a, Evolution of an internal gene by insertion of a DNA sequence into an intron of a pre-existing gene. b, *de novo* evolution of a gene from an intronic sequence of a pre-existing gene. c, Internalization of a gene after exon(s) acquisition of an adjacent gene. d, Internalization of a gene via fusion of two flanking genes. Color key: pink – internal gene, green – exons of the external gene, blue and yellow – flanking genes.

Table 1. Mechanisms of origin of human internal genes.

	After human - mouse	After human - opossum	After human - chicken	Total
Duplication	2	2	4	8
Retrotransposition	3	11	6 (5)	21 (20)*
Duplication or retrotransposition	2	3 (2)	16 (2)	21 (6)
De novo candidates	2	1	2	5

Independent events are shown in parentheses.
*One retrotransposition event was not dated to the degree of accuracy as other cases.

Supplementary Methods

Identification and quality control of nested gene pairs

Sequences and annotations for *H. sapiens*, *D. melanogaster*, *C. elegans*, and *C. briggsae* genomes were downloaded from the NCBI GenBank [23] database at ftp://ftp.ncbi.nih.gov/genomes/. After selecting the longest isoform of each gene, we identified 428, 815, 440, and 608 nested gene pairs in each genome, respectively. Several measures were taken to exclude erroneously annotated nested genes. For the *H. sapiens*, *D. melanogaster*, and *C. elegans* genomes, we retained only RefSeq genes [24]. We also excluded all human genes with the

labels “hypothetical” or “predicted” in the defline of the GenBank-derived fasta file. For the *D. melanogaster* and *C. elegans* genomes, we kept only those genes that showed >95% sequence identity over >90% of the length of the best nucleotide BLAST [25] hit with complete mRNAs sequenced from the same species. The mRNAs were obtained from GenBank with the Entrez retrieval system [23], using the species names and “complete” as key words and setting the limits option to mRNA molecules. Because annotation of the *C. briggsae* genome was the least reliable, we required that all *C. briggsae* genes have significant BLAST hits to protein sequences from the final set of *C. elegans* genes. In addition, all cases of nested gene evolution involving *C. briggsae* gene annotations were checked manually against *C. elegans* annotations using the BLAT program [26] on the UCSC genome browser [27].

Comparative genomic analysis of nested gene structures

Genes that passed the above inclusion criteria were compared to the genomes of sister species and outgroups. We used the protein BLAT alignment tool on the UCSC genome browser, as well as the TBLASTN program [25], to compare protein sequences of internal and external genes to complete genomes. If an ortholog for an internal gene was not identified using either of these two methods, a TBLASN search was performed against the orthologous intron from the external gene. Thus, in order to classify a nested gene structure as having been gained or lost in evolution, we required that both the internal and external genes be found in the sister species and an outgroup. It is easier to find an

internal gene within the orthologous intron of an external gene in an outgroup, which was our expectation for an evolutionary loss, than it is to find it in the entire genome of the outgroup, which was the requirement for an evolutionary gain (Figure 1). Thus, our approach was conservative and could have slightly biased the results in favor of discovery of evolutionary losses. Also, the requirement of finding both genes in both genomes prevented us from misidentifying as evolutionary events genes that are absent due to incomplete genome sequences. For vertebrates, an additional method was employed to analyze the evolution of nested gene structures. Alignments of regions in the sister and outgroup species orthologous to the nested gene pair were constructed using OWEN [28]. We began all alignments with a strict requirement of 16 successive matches and $p < 10^{-8}$ and progressively relaxed these parameters to 8 successive matches and $p < 0.01$, using the greedy algorithm to resolve any conflicts. Presence or absence of an internal gene in the orthologous external gene was judged based on the quality of the alignment. A gap in the alignment opposite the entire span of an internal gene in human indicated the absence of the internal gene in that genome. Both methods yielded the same results, with the exception of 5 cases, which are candidates for *de novo* gene creation. Candidate *de novo* genes were identified when both TBLASTN and BLAT revealed no sequence similarity of an internal gene in a sister species. We did not apply the latter method to invertebrate genomes due to the higher degree of their divergence, which also prevented us from performing a systematic analysis of the modes of internal gene evolution in invertebrates.

Analysis of gene expression

Gene expression data were obtained from [29], which included 73 healthy human tissues measured on the HG-U133A Affymetrix array. We computed the correlation of mean levels of expression of internal and external genes for 109 nested genes in humans. We next identified all adjacent pairs of RefSeq annotated genes in the human genome and randomly selected 109 such pairs 1000 times. We than compared the correlation coefficient of the 109 nested genes to the average correlation coefficient of the 1000 trials of 109 adjacent pairs.

Estimating the rate of nested gene evolution

Of the 128 definite nested gene structures in the human lineage, we identified 55 that emerged after the divergence of human and zebrafish lineages ~450 million years ago [30]. Assuming that these 128 nested gene structures are representative of the overall 428 annotations in the human genome, the observed number of internal gene gains give an estimate of ~ 0.4 gains per million years for all nested genes in the human genome (55/128 * 428/450). In the *D. melanogaster* lineage, 48 internal genes were gained since the divergence of *D. melanogaster* and *D. pseudoobscura* ~55 million years ago [31], indicating a rate of ~0.9 gains per million years. Our analysis of the *C. elegans* genome was more restricted due to large distances between the *C. elegans*, *C. briggsae*, and *Pristionchus pacificus* genomes. Because we never considered cases where

sequence similarity was not high enough to determine orthology, we described only a handful of cases of nested gene evolution. Nevertheless, an approximation was still possible due to the total number of nested genes showing a high enough sequence similarity between *C. elegans*, *C. briggsae*, and *P. pacificus* genomes. Of the 440 total *C. elegans* internal genes, exactly one half (220) were found in *C. briggsae* and *P. pacificus*, 11 of which were gains. Thus, the overall rate of nested gene gain was 22 per ~100 million years of evolution separating *C. elegans* and *C. briggsae* [32], or ~0.2 per million years.

References

1. Mironov, A.A. *et al.* (1999) Frequent alternative splicing of human genes. *Genome Res.* 9, 1288–1293.

2. Willingham, A. T. *et al.* (2006) Transcriptional landscape of the human and fly genomes: nonlinear and multifunctional modular model of transcriptomes. *Cold Spring Harb. Symp. Quant. Biol.* 71, 101–110.

3. Kapranov, P. *et al.* (2007) Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.* 8, 413–423.

4. Veeramachaneni, V. *et al.* (2004) Mammalian overlapping genes: the comparative perspective. *Genome Res.* 14, 280–286.

5. Misra, S. *et al.* (2002) Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. *Genome Biol.* 3, research0083.1–0083.22.

6. Yu, P. *et al.* (2005) Nested genes in the human genome. *Genomics* 86, 414–422.

7. Drosophila 12 Genomes Consortium. (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450, 203–218.

8. Henikoff, S. and Eghtedarzadeh, M.K. (1987) Conserved arrangement of nested genes at the Drosophila Gart locus. *Genetics* 117, 711–725.

9. Habib, A. A. *et al.* (1998) The OMgp gene, a second growth suppressor within the NF1 gene. *Oncogene* 16, 1525–1531.

10. Jaworski, D. M. *et al.* (2007) Potential regulatory relationship between the nested gene DDC8 and its host gene tissue inhibitor of metalloproteinase-2. Physiol. *Genomics* 28, 168–178.

11. Furia, M. *et al.* (1993) Dense cluster of genes is located at the ecdysone-regulated 3C puff of *Drosophila melanogaster*. *J. Mol. Biol.* 231, 531–538.

12. Davies, W. *et al.* (2004) Expression patterns of the novel imprinted genes Nap1l5 and Peg13 and their non-imprinted host genes in the adult mouse brain. *Gene Exp. Pat.* 4, 741–747.

13. Crampton, N. *et al*. (2006) Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy *Nucleic Acids Res.* 34, 5416–5425.

14. Osato, N. *et al.* (2007) Transcriptional interferences in cis natural antisense transcripts of humans and mice. *Genetics* 2, 1299–1306.

15. Da Lage, JL. *et al*. (2003) A nested alpha-amylase gene in Drosophila ananassae. *J. Mol. Evol*. 57, 355–362.

16. Lynch, M. and Conery, J.S. (2003) The origins of genome complexity. *Science* 302, 1401–1404.

17. Lynch, M. (2006) The origins of eukaryotic gene structure. *Mol. Biol. Evol.* 23, 450–468.

18. Yi, S.V. (2006) Non-adaptive evolution of genome complexity. *Bioessays* 28, 979–982.

19. Lynch, M. (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl. Acad. Sci. USA* 104, 8597–8604.

20. Lynch, M. (2002) Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci. USA* 99, 6118–6123.

21. Carmel, L. *et al.* (2007) Patterns of intron gain and conservation in eukaryotic genes. *BMC Evol. Biol.* 7, 192.

22. Mattick, J.S. and Makunin, I.V. (2005) Small regulatory RNAs in mammals. *Hum. Mol. Genet.* 14, R121-R132.

23. Benson D.A. *et al.* (2008) GenBank. *Nucleic Acids Res.* 36, D25–D30.

24. Pruitt, K. D. *et al.* (2007) Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins *Nucleic Acids Res.* 35, D61–D65.

25. Altschul, S. F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

26. Kent, W. J. (2002) BLAT - the BLAST-like alignment tool. *Genome Res.* 12, 656–664.

27. Karolchik, D. *et al.* (2008) The UCSC Genome Browser Database: update. *Nucleic Acids Res.* 36, D773–D779.

28. Ogurtsov, A. Y. *et al.* OWEN: Aligning long collinear regions of genomes. *Bioinformatics* 18, 1703–1704 (2002).

29. Su, A. I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* 101, 6062–6067.

30. Kumar, S. and Hedges, S. B. (1998) A molecular timescale for vertebrate evolution. *Nature* 392, 917–920.

31. Tamura, K. *et al.* (2003) Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol. Biol. Evol.* 21, 36–44.

32. Stein, L. D. *et al.* (2003) The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol.* 1, E45

Chapter 7, in full, is a manuscript intended for publication as Assis R, Kondrashov AS, Koonin, EV and Kondrashov FA (2008) Nested genes and increase of organizational complexity in metazoan genomes.The dissertation author was the primary investigator and author of this paper.