Polymorphism due to multiple amino acid substitutions at a codon site within *Ciona savignyi*

Nilgun Donmez*[1], Georgii A. Bazykin†[1], Michael Brudno*‡[2], Alexey S. Kondrashov§

*Department of Computer Science, and ‡Banting and Best Department of Medical Research, University of Toronto, Toronto ON M5S 3G4 Canada

†Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow 127994, Russia

§Life Sciences Institute and Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109-2216, USA

Running head: Polymorphism due to multiple substitutions

Keywords: polymorphism, positive selection, non-synonymous substitutions, two-substitution codons, *Ciona savignyi*

[1] These two authors contributed equally to this work.
[2] *Corresponding author:*

Michael Brudno
10 King's College Rd
Pratt Bldg Rm 283
Toronto ON M5S 3G4 Canada
brudno@cs.toronto.edu
Tel.: +1 416-978-2589
Fax: +1 416 978-1455

Abstract

We compared two haploid genotypes of one *Ciona savignyi* individual and identified codons at which these genotypes differ by two non-synonymous substitutions. Using the *Ciona intestinalis* genome as an outgroup, we showed that both substitutions tend to occur in the same genotype. Only in 53 (34.4%) of 154 codons, one substitution occurred in each of the two genotypes, although 77 (50%) of such codons are to be expected if substitutions were independent. We considered two feasible evolutionary causes for the observed pattern: substitutions driven by positive selection and compensatory substitutions, as well as several potential biases. However, none of these explanations is fully compelling, and data on multiple genotypes of *C. savignyi* would help to elucidate the causes of this pattern.

Natural populations possess very different levels of nucleotide diversity, ranging from <0.001 to >0.1 (Snoke et al 2006; Lynch 2007). Among multicellular organisms, the highest nucleotide diversity has so far been observed in a marine ascidian *Ciona savignyi*, where two haploid genotypes (haplotypes) sequenced from the same individual differ from each other at 8% of nucleotide sites (Small et al 2007). Such a high diversity, which appears to be primarily due to a large effective population size of the species, makes *C. savignyi* an attractive model organism for population genetic studies. Some phenomena and patterns that can be investigated in *C. savignyi* may be almost impossible to study in less polymorphic species.

Here we consider one such phenomenon: the presence, in different haplotypes of *C. savignyi*, of allelic codons that differ from each other at two or three nucleotide sites. Such allelic codons are exceedingly rare in less diverse populations. When codons that differ from each other by multiple nonsynonymous substitutions were observed in different species, an excess of substitutions within the same lineage was interpreted as a sign of positive selection (Bazykin et al 2004, 2006). In this paper we report a similar excess in two haplotypes of *C. savignyi*, and analyze its possible causes.

MATERIALS AND METHODS

C. intestinalis annotations constituting 14,002 genes were downloaded from ftp://ftp.jgi-psf.org/pub/JGI_data/Ciona/v2.0. The *C. savignyi* genome (version 2.01) was downloaded from http://mendel.stanford.edu/SidowLab/ciona.html. We used the alignment of the two haploid genotypes, A and B, available at the site and described in (Small et al. 2007). Translated *C. intestinalis* genes were queried

against both haplotypes of *C. savignyi* using "protein2genome" functionality of the alignment software Exonerate (Slater and Birney 2005). The best hits for a gene on both haplotypes were required to have a normalized score of at least 3.0 (calculated by dividing the Exonerate raw score by the protein length involved in the alignment). In case of a tie for the best hit in either of the haplotypes, or if the normalized score of the best hits on each haplotype differed by more than 0.5, the protein was eliminated to avoid potential paralog problems. In addition, the locations of the alignments on each haplotype were checked to ensure that they correspond to the same position (+/- 5 nucleotides) in the global alignment of the two haplotypes (Small et al 2007). If the intersection of the pairwise alignments was less than 100 codons or covered less than 75% of the gene, the gene was not considered for further analysis. Finally, alignments with ambiguities in nucleotide sequences or internal stop codons were discarded. For all of the remaining genes, the two sets of pairwise Exonerate alignments were merged into 3-way alignments as follows: the intersection of "*C. intestinalis vs*. haplotype A" and "*C. intestinalis vs*. haplotype B" alignments was taken based on the *C. intestinalis* amino acid, while introducing extra gaps whenever one of the pairwise alignments had a gap in the *C. intestinalis* sequence that the other did not. The remaining dataset consisted of 5,478 triplets of orthologous genes.

We further masked codons that were not flanked, on each side, by gapless alignments of 10 amino acids or more, with at least 5 matches between the 2 haplotypes, and at least 3 matches between each haplotype and *C. intestinalis*. To remove the effect of insertion and deletion sequencing errors, we also masked frame-shifted regions: those in which the same DNA sequence of length 4 or more occurred in the aligned *C. savignyi* sequences with a shift of +/-1. The full alignments and the list of codons are available at http://compbio.cs.toronto.edu/ciona/.

The last common ancestor (LCA) codon for a pair of allelic codons in the two *C. savignyi* haplotypes was determined as follows. When the two allelic codons differed from each other at one nucleotide site, and encoded the same amino acid, we assumed that if the homologous *C. intestinalis* codon (outgroup, O) coincided either with the codon from haplotype A or with the codon from haplotype B, LCA coincided with O. In other words, we assumed parsimony, which implies that when O coincides with A (B), the single synonymous nucleotide substitution occurred in the lineage that led to B (A). If A and B differ from each other at one nucleotide site but encode different amino acids, we assumed that if O either coincides with A (B) or differs from both A and B but still encodes the same amino acid as A (B), the LCA also encodes this amino acid, and that the only nonsynonymous substitution occurred in the lineage that led to B (A). When O encodes an amino acid different from that encoded by both A and B, we assumed that the LCA could not be determined.

When the two allelic codons differ from each other at two nucleotide sites, we also assumed that the LCA coincided with O either if O coincided with A or B, or if the O codon was intermediate between codons A and B, in the sense that O differed from both A and B by a single nucleotide. In addition, when both allelic codons and the two intermediates all encoded different amino acids we assumed that the LCA encoded the same amino acid as O if O encoded the same amino acid as A, B, or one of the intermediate codons. Otherwise, we assumed that LCA could not be determined. Pairs of codons for which one of the two possible intermediate codons is a stop codon were not considered.

When the two allelic codons differ from each other at three nucleotide sites, we assumed that the LCA coincided with O either if O coincided with A or B, or if the O codon was intermediate between codons A and B, in the sense that O differed from one of them at one nucleotide site and from the other one at two nucleotide sites. In all other cases, we assumed that LCA could not be determined, as it is usually impossible to establish with confidence that all the substitutions between the two allelic codons were nonsynonymous. Pairs of codons for which one of the six possible intermediate codons is a stop codon were not considered.

Gene-specific synonymous and nonsynonymous evolutionary distances were estimated by codeml program of the PAML package (Yang et al. 1997) from pairwise nucleotide alignments for the two *C. savignyi* haplotypes and each haplotype and the *C. intestinalis* genome, taken from the triple alignments. When the distances were estimated for correlation with the occurrence of a variable codon in some region, that codon itself was excluded in distance estimation.

RESULTS

Patterns in distribution of multiple substitutions within a codon

Among the 1,251,343 homologous codons in the 5,478 analyzed genes, 93.46% are identical in the two haplotypes, and 6.40%, 0.12% and 0.005% differ at one, two and three nucleotide sites, respectively (no-, one-, two- and three-substitution codons). The mean evolutionary distance between the haplotypes is 0.086 at synonymous sites and 0.004 at non-synonymous sites, in agreement with Small *et al*. (2007). Among codons with a single synonymous substitution between the haplotype A and haplotype B, O (outgroup) coincides with either A or B in 56% of cases (Table 1). Among codons with a single nonsynonymous substitution between A and B, O encoded the same amino acid as either A or B in 60% of cases (Table 1).

There are 1610 codons separated by two substitutions, including 288 codons separated by two synonymous substitutions (such codons are rare, as they must code for either Leu or Arg), and 249 codons separated by two non-synonymous substitutions (Table 2). If the substitutions were independent, we would expect both of them to occur in the same lineage with the probability of ~50% (Bazykin *et al*., 2004). In agreement with this expectation, in codons where both substitutions were synonymous, they occur in the same lineage approximately in half of the cases.

In contrast, two non-synonymous substitutions occur in the same lineage in 66% of the cases, and in different lineages in only 34% of the cases (Table 2). Clumping of nonsynonymous substitutions is more pronounced in highly conserved genes and gene regions (Table 2). When two nonsynonymous substitutions occur in the same lineage, they tend to occur in the lineage which has a higher rate of nonsynonymous (86 of 101; chi-square, $P < 0.0001$), but not necessarily synonymous (55 of 101; chi-square, $P = 0.573$), substitutions in this gene. In contrast, when two synonymous substitutions occur in the same lineage, there is no significant difference in the rate of synonymous or nonsynonymous substitutions between the two lineages (chi-square, $P > 0.05$).

Clumping is also present in 66 three-substitution codons: while all substitutions are expected to occur in the same lineage in only 25% of cases, the observed value is 46% (chi-square, $P = 0.029$; Table 3).

Analysis

In the following subsections, we will consider three possible explanations for the observed clumping of non-synonymous substitutions: positive selection, compensatory mutations, and potential biases in the Last Common Ancestor identification, as well as other biases and phenomena that could lead to an excess of substitutions in the same lineage.

Two-substitution polymorphisms due to positive selection

Positive selection can lead to clumping of non-synonymous substitutions within a codon (Bazykin 2004, 2006). However, in contrast to the previous observations of such clumping, here we are dealing with intrapopulation polymorphisms, and transitive polymorphisms due to positive selection-driven allele replacements are short-lived (see Crow and Kimura, 1970). Thus, it is not clear whether positive selection driving both of the substitutions that convert the codon found in haplotype A into the codon found in haplotype B can provide a quantitatively feasible explanation.

We will roughly estimate the number of codons that differ by two nonsynonymous substitutions between two haplotypes that segregate within a population under positive selection. Let us assume (unrealistically) that both of these substitutions occur simultaneously. Then, in the absence of dominance, the deterministic dynamics of replacement of codon A with codon B are described by the fundamental equation

$$dp/dt = sp(1-p) \qquad (1)$$

where *p* is the frequency of codon B, 1-*p* is the frequency of codon A, and *s* is the selective advantage of codon B over codon A. The solution to this equation is given by

$$p(t) = \frac{1}{1 + (\frac{1}{p_0} - 1)e^{-st}} \qquad (2)$$

(see Crow and Kimura, 1970). If two haploid genotypes are sampled from the population every generation, the total number of heterozygous combinations over the whole course of a substitution is

$$T = \int_{-\infty}^{\infty} 2p(t)(1 - p(t))dt = 2/s \qquad (3)$$

Thus, if we observe *k* heterozygous loci under positive selection within a pair of complete haploid genotypes, the per-generation number of positive selection-driven allele replacements required to explain this fact is *ks*/2. The excess of codons where both non-synonymous substitutions occurred in the lineage of the same haplotype suggests that *k* ~50 (Table 2). Thus, even if selection is very weak (say, $s = 10^{-5}$), we still need to assume that there is a positive selection-driven replacement that results in a 2-substitution codon every 4000 generations. Probably this rate of positive selection-driven evolution is too high (Eyre-Walker 2006), because selection that simultaneously favors two non-synonymous substitutions must occur only in a minority of cases. Moreover, our calculations underestimate the required prevalence of positive selection, because in reality the two substitutions necessary to convert codon A into codon B cannot occur simultaneously. Codon A will be replaced not by codon B directly, but by an intermediate codon which has selective advantage over A, and codons A and B would coexist for a much shorter period than the two alleles where one directly replaces the other, as assumed in equation (1). Thus, simple positive selection favoring both changes is unlikely to be the leading cause of the observed pattern.

Two-substitution polymorphisms due to compensatory mutations

A second potential explanation for the observed pattern is compensatory evolution. Let us assume that codons A and B have the same fitness, but the intermediate codons have a reduced fitness, $1\text{-}s$. In this case, in each pair of the substitutions, only the second (from an intermediate codon to either A or B) is driven by positive selection, and there is no long-term increase in fitness. Then, the deterministic equilibrium frequency of the intermediate codons will be $2m/s$, where m is the per-nucleotide mutation rate, and codons A and B appear from these intermediate codons, due to mutation, at the rate $m*(2m/s) = 2m^2/s$. If we treat coexistence of A and B as a selectively neutral polymorphism with the "effective mutation rate" $m_{\text{eff}} = 2m^2/s$, the expected nucleotide diversity is $\pi = 4N_e m_{\text{eff}} = 8N_e m^2/s$. Assuming $m = 10^{-8}$, $N_e = 10^6$ (Small et al. 2007), and $s = 10^{-5}$, we obtain $\pi \sim 10^{-4}$. Thus, in order to explain the 50 extra codons where the variants found in haplotypes A and B differ by two non-synonymous substitutions (Table 2), we need to assume that such compensatory selection operates on $5*10^5$ codons, *i. e.* on 40% of all codons. The assumption that such a large fraction of protein sites are under this kind of selection, with at least two amino acids conferring the same high fitness, does not seem to be very likely. Of course, if the selection against the intermediate codons is weaker, a smaller fraction of codons under compensatory selection will suffice: if $s = 10^{-6}$ only 4% of the codons could be under this selection. As s declines past 10^{-6}, however, selection becomes inefficient given $N_e = 10^6$, and the intermediate codon(s) will become effectively neutral.

Biased misidentification of the ancestral codon

The observed clumping could also be caused by biased misidentification of the LCA codon for the two *C. savignyi* haplotypes. Because *C. intestinalis* is not a close outgroup for within-species polymorphism in *C. savignyi*, in a substantial fraction of cases, LCA cannot be identified under the assumption of parsimony (Table 2), and in some cases, LCA is likely to be misidentified. Unbiased mistakes in identification of the LCA codon will not produce the observed pattern: if, in the case of a two-substitution pair of *C. savignyi* codons A and B, the LCA codon was drawn randomly from the four possible codons (A, B, and the two intermediates), the two substitution that distinguish A from B will be attributed to the same and the two different haplotypes with equal probabilities. However, there may also be a systematic bias in misidentification of LCA, due to two reasons.

First, it may be possible that only one of the two intermediate codons confers a high fitness, and the other one confers a low fitness. For example, if the codons A and B are AAT (encoding Asn) and GGT (encoding Gly), the intermediate codon AGT (encoding Ser) may be fit, and the intermediate codon GAT (encoding Asp) may be unfit. Then, if the outgroup is very distant from the LCA, it will carry the fit intermediate codon in only 1/3 of cases (assuming that at equilibrium codons AAT, GGT, and AGT are equally common). As a result, one could conclude that for 2/3 of codon pairs, both substitutions occurred in one *C. savignyi* haplotype, because the LCA (as revealed by the outgroup) coincides with the other haplotype.

Unfortunately, *C. intestinalis* is the closest known outgroup that can be used to polarize the *C. savignyi* polymorphism. We investigated the potential effect of biased misidentification of LCA by testing the robustness of the excess of multiple substitutions in the same lineage in interspecific divergence (Bazykin et al 2004) to the choice of the outgroup. We identified the codons with 2 non-synonymous substitutions between human and mouse, and determined LCA in 3 ways: i) using only sites where dog and opossum carry the same codon (most confident), ii) using dog as an outgroup, and iii) using opossum as an outgroup (least confident). The corresponding fractions of codons where both substitutions in two-substitution human-mouse codons occurred in the same lineage (amino acid-level pattern) were 76%, 69%, and 67%, respectively. Thus, the excess of codons where both substitutions were attributed to the same lineage diminishes when a more distant outgroup is used. These data argue against biased misidentification of the LCA as an explanation of the observed pattern in divergence between independent evolutionary lineages, although this conclusion does not necessarily apply to within-species polymorphism of *C. savignyi*.

Second, biased misidentification of LCA may appear if the amino acid composition of proteins is out of equilibrium (Jordan et al. 2005). If, for a double-substitution codon, the amino acid substitution that creates the terminal amino acid from the intermediate amino acid is more likely than the reciprocal substitution, some clumping could be an artefact of systematic errors in inferring the LCA from the outgroup codon (Bazykin et al. 2004). To test this hypothesis, we used the data on codons identical between the two *C. savignyi* haplotypes to infer the rates of each nonsynonymous substitution between *C. intestinalis* and *C. savignyi*. Next, for each 2-substitution pair of codons, we compared the rate of substitutions from each of the two intermediate codons into each of the two terminal codons (a total of four substitutions) with the rate of the reciprocal substitutions. A substitution from intermediate to terminal codon can lead to false excess of clumping for the given 2-substitution codon if its rate is higher than the rate of the reciprocal substitution. We compared the clumping between 2-substitution codons with 0 or 1 false excess amino acid substitutions ("false deficit" codons) with that in 2-substitution codons with 2-4 false excess amino acid substitutions ("false excess" codons). The difference between the two values was not large (Table 2; chi-square, $P > 0.1$), arguing against non-equilibrium composition of proteins as a leading cause for the observed clumping.

Other explanations

The observed clumping might also be due to errors in alignment, or closely correlated nearby single nucleotide sequencing errors. We believe, however, that our filtering criteria are effective in eliminating such cases: we make sure that the two orthologous *C. savignyi* exons are also aligned to each other in the haplotype alignments, and that the level of conservation between the two exons in *C. savignyi* is above a threshold. The observed elevated prevalence of nonsynonymous substitutions in the vicinity of codons where haplotypes A and B differ by two nonsynonymous substitutions is unlikely to be due to sequencing errors, because this effect is absent when A and B differ by two synonymous substitutions. Moreover, the clumping of nonsynonymous substitutions is the strongest when the nearby codons are completely conserved between the A and B haplotypes, further reducing the chance that sequencing or alignment errors play a dominant role.

This clumping cannot be explained by hypermutable CpG dinucleotides, because substitutions are also clumped in codon pairs where no intermediate codons contain CpGs (Table 2). Due to the small number of 2-substitution amino acids with mutations at 1st and 3rd nucleotides (Table 2), we cannot immediately dismiss that some of the clumping is due to mutation events spanning two adjacent nucleotides. In the interspecific case, however, Bazykin et al (2004) rejected this explanation in the presence of a larger dataset.

DISCUSSION

Our analysis of differences between the two haploid genotypes from one *C. savignyi* individual follows closely that for differences between mouse and rat genomes (Bazykin *et al*. 2004). Surprisingly, the results of these analyses are also rather similar to each other. Our data reveal clumping of within-population non-synonymous polymorphisms at the same codon that is similar in magnitude to clumping of non-synonymous substitutions that distinguish different genomes (Tables 2 and 3). Clumping of non-synonymous substitutions in different mammalian species (Bazykin *et al*. 2004) and HIV-1 strains (Bazykin *et al*. 2006) was interpreted as the signature of positive selection which occurred at some time during the divergence of the analyzed lineages. Although the exact fraction of positive selection-driven amino acid substitutions remains controversial, it is almost certainly substantial (Eyre-Walker 2006).

In contrast, the role of positive selection in maintaining polymorphism at the molecular level is believed to be small (Kimura 1983). Indeed, our rough estimates suggest that the observed clumping of non-synonymous substitutions cannot be easily explained by positive selection. We considered two different scenarios, assuming that either both, or only one, of the two non-synonymous differences between the two *C. savignyi* codons are favored by positive selection, and in both cases it takes very high prevalence of positive selection to explain the observed clumping.

Another feasible explanation for the observed clumping of non-synonymous substitutions is the biased misidentification of the LCA. Qualitatively, this is feasible if only one intermediate codon has a high enough fitness to be present. Quantitatively, however, this mechanism appears to be unlikely to explain what we see. We also considered the effect of sequencing errors and correlated mutations at adjacent nucleotides, and do not believe these to be the leading causes of the observed clumping.

The claim that multiple non-synonymous substitutions at a codon tend to occur in clumps (Bazykin et al. 2004) was disputed by Friedman and Hughes (2005). They argued that codons with two nonsynonymous substitutions occur less frequently than codons with one synonymous and one non-synonymous substitution. Because in most genes non-synonymous substitutions are much rarer than synonymous substitutions, their result is certainly true. This, however, in no way affects the argument that codons with two non-synonymous substitutions in one lineage are overrepresented compared to codons with two non-synonymous mutations in different lineages. The "model-free approach" of Friedman and Hughes (2005) was also criticized by Yang (2006).

In summary, we observe a pattern – the clumping of non-synonymous substitutions that distinguish the two known haplotypes of *C. savignyi* – that appears to be real but lacks a definite explanation. The situation is even more intriguing because the analogous pattern in divergence of different lineages can be naturally explained by positive selection. We believe that the sequencing and analysis of additional haplotypes of *C. savignyi* will resolve this puzzle. In particular, if positive selection does play a role in the observed clumping, nucleotide diversity in the vicinity of a young, derived allele created by several advantageous substitutions must be reduced due to hitch-hiking (*e. g*., Evans et al., 2005, Helgason et al. 2007). Simultaneously the allele frequencies of intermediate codons will make it possible to ascertain adaptive landscapes associated with individual polymorphic codons and to elucidate the factors responsible for our observations.

ACKNOWLEDGEMENTS

This study was supported by National Science and Engineering Research Council of Canada (NSERC) Discovery grant to M.B., and grants from the Russian Academy of Sciences (programs “Biodiversity” and “Molecular and Cellular Biology”) and Russian Foundation of Basic Research (08-04-01394 and 08-04-09680). G.A.B. was partially supported by the Russian Science Support Foundation.

Literature Cited

Bazykin, G. A., F. A. Kondrashov, A. Y. Ogurtsov, S. Sunyaev, and A. S. Kondrashov, 2004 Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. Nature **429:** 558-62.

Bazykin, G. A., J. Dushoff, S. A. Levin, and A. S. Kondrashov, 2006 Bursts of nonsynonymous substitutions in HIV-1 evolution reveal instances of positive selection at conservative protein sites. Proc. Natl. Acad. Sci. USA **103:** 19396-401.

Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory*. Harper & Row: New York.

Evans, P. D., S. L. Gilbert, N. Mekel-Bobrov, E. J. Vallender, J. R. Anderson *et al.,* 2005 Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. Science **309:** 1717-20.

Eyre-Walker, A. 2006 The genomic rate of adaptive evolution. Trends Ecol. Evol. **21:** 569-75.

Friedman, R., and A. L. Hughes, 2005 The pattern of nucleotide difference at individual codons among mouse, rat, and human. Mol. Biol. Evol. **22:** 1285–1289.

Huelsenbeck, J. P., S. Jain, S. W. Frost, and S. L. Pond, 2006 A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. Proc. Natl. Acad. Sci. USA. **103:** 6263-8.

Helgason, A, S. Palsson, G. Thorleifsson, S. F. A. Grant, V. Emilsson *et al*., 2007 Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. Nat. Genet. **39:** 218-25.

Jordan, I. K., F. A. Kondrashov, I. A. Adzhubei, Yu. I. Wolf, E. V. Koonin *et al.*, 2005 A universal trend of amino acid gain and loss in protein evolution. Nature **433:** 633-638.

Kimura M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge Univ. Press.

Lynch M., 2007. *The Origins of Genome Architecture*. Sinauer: Sunderland.

Slater, G. S., and E. Birney, 2005 Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. **6:** 31.

Small, K. S., M. Brudno, M. M. Hill, and A. Sidow, 2007 Extreme genomic variation in a natural population. Proc. Natl. Acad. Sci. USA. **104:** 5698-703.

Snoke, M. S., T. U. Berendonk, D. Barth, and M. Lynch, 2006 Large global effective population sizes in *Paramecium*. Mol. Biol. Evol. **23:** 2474-2479.

Yang Z., 2006 On the Varied Pattern of Evolution of 2 Fungal Genomes: A Critique of Hughes and Friedman. Mol. Biol. Evol. **23:** 2279–2282

Yang Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13:** 555-6.

Table 1. Divergence at codons where haplotypes A and B differ at 1 nucleotide site[a]

	Substitution in lineage of A [a]	Substitution in lineage of B [a]	LCA unknown
Synonymous substitution	20,023 (50.0%)	20,052 (50.0%)	31,901 (44.3%)
Non-synonymous substitution	2,452 (49.7%)	2,481 (50.3%)	3,236 (39.6%)

[a] Frequencies in the first two columns are only within codons where the LCA is known (see Methods for details).

Table 2. Divergence at codons where haplotypes A and B differ at 2 nucleotide sites

	Both substitutions in same haplotype	Substitutions in different haplotypes	LCA unknown
Two synonymous substitutions	82 (43.9%)	105 (56.2%)	101 (35.1%)
One synonymous and one non-synonymous substitution	117 (43.0%)	155 (57.0%)	404 (59.8%)
Two synonymous or two non-synonymous substitutions	52 (69.3%)	23 (30.7%)	69 (47.9%)
None or one synonymous substitution, two or one non-synonymous substitutions	68 (68.7%)	31 (31.3%)	126 (56.0%)
Two non-synonymous substitutions	101 (65.6%)	53 (34.4%)	95 (38.2%)
Codons:			
Possible false excess	63 (68.5%)	29 (31.5%)	74 (44.6%)
Possible false deficit	38 (61.3%)	24 (38.7%)	21 (25.3%)
1,3-substitution[a]	13 (56.5%)	10 (43.5%)	10 (30.3%)
CpG-free[b]	69 (66.3%)	35 (33.7%)	62 (37.3%)
Regions:[c]			
Very strong conservation	12 (92.3%)	1 (07.7%)	4 (23.5%)
Strong conservation	15 (78.9%)	4 (21.1%)	11 (36.7%)
Moderate conservation	19 (55.9%)	15 (44.1%)	18 (34.6%)
All others	55 (62.5%)	33 (37.5%)	62 (41.3%)
Genes:[d]			
Low D_n	18 (94.7%)	1 (05.3%)	4 (17.4%)
Medium D_n	28 (58.3%)	20 (41.7%)	32 (40.0%)

High D_n	55 (63.2%)	32 (36.8%)	59 (40.4%)

[a] 1,3-substitution codons are those where the two *C. savignyi* haplotypes differ from each other at the first and third nucleotide sites.

[b] CpG-free codons are those in which neither of the two possible intermediate states between rat and mouse codons includes CpG context, neither inside the codon nor on its boundary.

[c] Regions with very strong, strong and moderate conservation are those in which the codon under consideration is flanked from each side by gapless alignments of two *C. savignyi* genomes and *C. intestinalis* of length 10 or more each with 9 or 10, 8, and 7 invariant amino acids, respectively.

[d] Genes were split into 3 bins of equal size (low, medium and high D_n) according to the average of D_n values between *C. intestinalis* and each of the haplotypes of *C. savignyi*.

Table 3. Divergence at codons where genomes A and B differ at 3 nucleotide sites [a]

Number of paths involving synonymous substitutions	All substitution in same lineage	Two substitutions in same lineage, one substitution in the other lineage	LCA unknown
Six	9 (52.9%)	8 (47.1%)	10 (37.0%)
Five	0 (00.0%)	1 (100.0%)	5 (83.3%)
Four	2 (100.0%)	0 (00.0%)	1 (33.3%)
Three	1 (50.0%)	1 (50.0%)	6 (75.0%)
Two	0 (00.0%)	4 (100.0%)	6 (60.0%)
One	0 (00.0%)	0 (00.0%)	3 (100.0%)
None	0 (00.0%)	0 (00.0%)	0 (00.0%)
Total	**12 (46.2%)**	**14 (53.8%)**	**31 (54.4%)**

[a] There are six evolutionary paths between the two codons that differ from each other at all three sites, depending on the order in which the substitutions occur.