Chapter 10

Replicational Mutation Gradients, Dipole Moments, Nearest Neighbour Effects and DNA Polymerase Gamma Fidelity in Human Mitochondrial Genomes

Hervé Seligmann

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/51245

1. Introduction

The large amount of available human mitochondrial genome data from Ingman and Gyllenstein and from Ruiz-Pesini et al [see http://www.mtdb.igp.uu.se/, 1 and http://www.mitomap.org/MITOMAP, 2] enables to study in some detail the spectrum of mutations observed within this species' mitochondrion. DNA mutations have two main causes: spontaneous chemical alterations of nucleotides, from one nucleotide 'species' to another, such as hydrolytic deaminations from C->T and A to hypoxanthine, which pairs with C and leads to its replacement by G (in the following summarized as A->G); and inaccuracies by the enzymatic machinery that is responsible for the polymerization of new DNA strands from the template of the existing DNA during DNA replication. Here I explore the tendency for mutations from different genes and mutation types to be explained by the first (physicochemical), or the other (more enzymatic/biological) factor, also in relation to adaptive constraints (natural selection is weaker against DNA mutations that cause no or only conservative changes at the protein level). The relative importance of these various factors affecting mutation spectra is investigated for observed human mitochondrial mutations in relation to different types of substitutions and different genes. I also explore nearest neighbour effects on the different mutation types, though the relative contribution of this factor in relation to others is not evaluated here.

The main, presumably sole DNA replicating enzyme in vertebrate mitochondria is the DNA gamma polymerase, which evolved from a bacterial tRNA synthetase [3]. Twelve types of substitutions of one nucleotide by another nucleotide occur at different frequencies. The most frequent changes occur within each of the nucleotide families, purines (adenosine, A, and

guanine, G) and pyrimidines (cytosine, C, and thymidine, T), as these involve less changes in molecular structure. These four purine to purine or pyrimidine to pyrimidine mutations are called transitions, the eight mutations from one chemical group to another are called transversions. A simplistic model predicts the frequencies of all substitutions, based on the dipole moment of the nucleotides [4], for a DNA region supposed to have no function, a pseudogene [5], so that observed substitution frequencies are believed unaffected by natural selection. The partial dipole moment of a chemical bond is proportional to the distance of an electron's mean position, in the chemical bond between atoms, from mid-distance between these atoms. The dipole moment of a molecule is the product of all partial dipole moments. G and C have high dipole moments, A and T have low dipole moments [6]. The hypothesized model assumes that a high dipole moment indicates high chemical reactivity, and hence probable alteration by chemical processes. Indeed, observed frequencies of mutations in pseudogenes of one nucleotide into another nucleotide are proportional to dipole moment changes: nucleotides with low dipole moment substitute those with high dipole moment.

Independently of the dipole moment hypothesis, some spontaneous chemical reactions, deaminations, A->G and C->T, occur preferentially while the heavy DNA strand is in the single stranded state [7, 8]. This occurs mainly during replication and transcription (DNA and RNA polymerization). Distances from replication origins and for transcription, from promoters [9, 10], determine durations that different DNA regions remain single stranded, creating gradients in deaminations in genomes with asymmetric replication, such as mitochondrial genomes (reviewed in [11, 12]). Hence gene position affects transition frequencies. Site-specific mutation rates estimated by phylogenetic reconstruction suggest that mutation gradients might also exist for some transversions [13, 14], indicating that single-strandedness might affect also substitutions that are not A->G or C->T.

Here I analyse mutation patterns observed in the 13 human mitochondrial protein coding genes, to estimate relative contributions of different processes to observed mutation patterns: replicational gradients [13, 14, 15], dipole moments [6, 16], selection against mutations that alter coding properties at the protein level [17] and gamma polymerase misincorporation [18], and potential interactions between these processes. I also explore nearest neighbour effects on mutation rates. The present analyses are also original in the sense that they are based on comparative analyses of sequence data, all the data originating from a single species, and not, as previous ones (i.e. [13, 14]), from comparisons between different, frequently evolutionarily distant species.

2. Dipole moments and the accuracy of the DNA gamma polymerase

The estimation of process-specific contributions of different mechanisms to a given phenomenon requires considering dependence among processes. Dipole moment effects are independent of the deamination gradient model which predicts a gradient in C->T, consistent with the model of dipole moment decrease, but an A->G gradient is inconsistent. This means that one deamination gradient fits into the dipole moment model, and the other does not, making both approaches approximately unrelated.

http://dx.doi.org/10.5772/51245

The issue of the accuracy of the gamma polymerase is more complex. It is indeed plausible that nucleotide misinsertion results from misrecognition of nucleotides by the polymerase, the latter due to physico-chemical similarities between nucleotides. This principle is also suggested by the high average misincorporation rates resulting in transitions (447.25±375.22) as compared to misincorporation rates causing transversions (121.64±186.6), explaining 33 percent of the variation in rates between different misincorporation types [18]. If so, the absolute value of the difference between dipole moments of nucleotides (from [16]) should be inversely proportional to misincorporation rates ('kd' in [18]), high rates occurring for nucleotides with similar dipole moments. This model differs from the model of dipole moment decrease, as it deals with the absolute value of the difference between dipole moments, and not the signed difference.

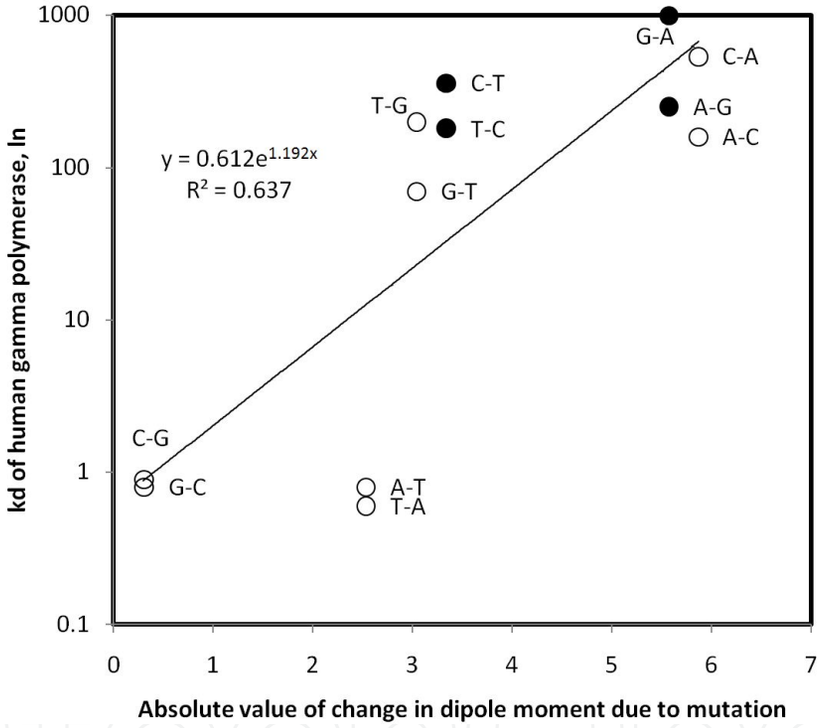

Figure 1. Misincorporation versus absolute difference between dipole moment of substituted and substituting nucleotide. Transitions (filled symbols) have high kds, but similar dipole moments decrease misincorporation kds.

The dipole similarity model for polymerase misincorporation rates can be dismissed at this point. Misincorporation rates increase, not decrease as expected, with absolute values of differences between dipole moments (r = 0.80, Figure 1). This unexplained association could reflect effects of other properties on misincorporation rates, properties that are inversely correlated with dipole moments. Note that after controlling for differences between transitions and transversions, the correlation shown in Figure 1 decreases (r = 0.54, not shown), yet the analysis confirms the principle that nucleotide substitutions with high kds tend to be substitutions between nucleotides with highly divergent dipole moments.

It is also possible that many nucleotide misincorporations result from the delay occurring between nucleotide recognition by the gamma polymerase and its incorporation in the elongating DNA polymer. One could suppose that some misincorporations are not due to misrecognitions, but to spontaneous mutations occurring after the nucleotide's accurate recognition by the polymerase, and before its incorporation. In that case, misincorporation rates should match the dipole moment model for decreased dipole moment: high rates are observed when substitutions decrease the dipole moment. This hypothesis cannot be ruled out, as misincorporation rates increase with the signed difference between nucleotide dipole moments (r = 0.50, not shown). Controlling for differences in kd between transitions and transversions, this positive association increases (r = 0.60, Figure 2).

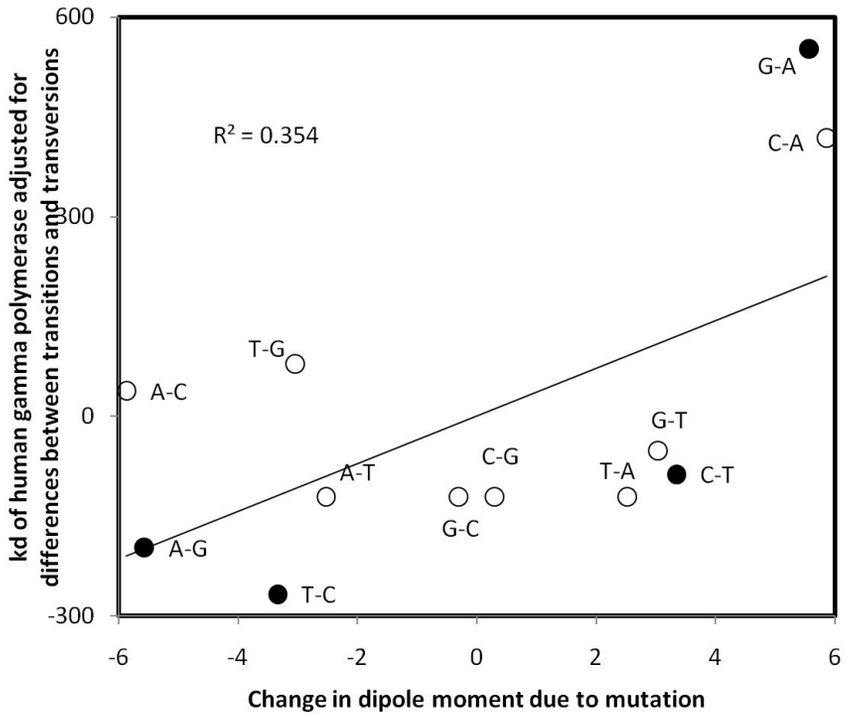

Figure 2. Adjusted misincorporation kd as a function of difference between dipole moments of substituted and substituting nucleotide. High kds imply dipole moment decrease.

Note that if causal interpretations of the associations in Figures 1 and 2 are relevant, it would be the dipole moment that affects kds. An alternative explanation to the trend in Figure 2 is that the gamma polymerase binds more readily nucleotides with low than high dipole moment, hence resulting in this biased misincorporation trend. Such a pattern could easily be caused by an overall relatively hydrophobic nature of the residues that constitute the polymerase's binding site (low dipole moment implying relative hydrophobicity). Even a very small bias for hydrophobic interactions would cause strong biases in analyses focusing on misincorporation rates. However, this hydrophobicity hypothesis does not seem to fit, at least in its simplistic form, what is known about the active site of the gamma polymerase according to the crystal structure published by Lee et al [19]. The active site consists of amino acids E895, Y951, R943 and Y955, among which one residue is positively charged (E,

http://dx.doi.org/10.5772/51245

glutamic acid), one negatively (R, arginine) and two are hydrophilic (Y, tyrosine). Note that none is classified as a hydrophobic amino acid. Hence the positive association in Figure 2 does not seem explained by active site hydrophobicity. Speculatively, electrostatic neutrality could favour misprocessing in active sites where each positive and negative charges occur, while high dipole moments would promote efficient processing.

These preliminary analyses suggest several important points on gamma polymerase fidelity: a) the causes for effects of similarity between nucleotides on misrecognition are unknown, structural similarity having effects opposite to those of dipole moment similarities; b) nucleotide properties affecting misrecognition are unknown but correlate with dipole moments; c) separating, even only conceptually, polymerase misrecognition from misincorporation, could be useful to understand polymerase accuracy; d) many misincorporations might be due to spontaneous mutations (with rates proportional to the dipole moment model) in the nucleotide occurring after accurate recognition by the polymerase, but before incorporation, resulting in misincorporation despite accurate recognition; e) alternatively, the polymerase's binding site might have in-built bias for hydrophobic misprocessing.

3. Selection on the gamma polymerase's misincorporation rates

Grantham [20] developed a matrix of dissimilarities based on major physico-chemical properties of amino acids (amino acid composition, polarity and molecular volume) that correlates best with amino acid replacement frequencies. From that matrix, Gojobori et al [17] estimated an average change in amino acid physico-chemical properties due to residue replacements for nucleotide substitutions in protein coding regions (see last line of table 4 in [17]). For example, A<->G substitutions have the lowest average impact, while G<->T have the greatest impact. One expects a negative association between impact on protein structure and the frequency of a nucleotide substitution. For pseudogenes, which do not code for proteins, the correlation between this average impact and the frequency of corresponding mutations (data from [17]) is weak ($r = -0.33$, one tailed $P = 0.15$), and even weaker after differences between transition and transversions have been accounted for ($r = -0.18$, one tailed $P = 0.29$). However, for mutation frequencies in coding sequences, natural selection against dysfunctional proteins has specifically decreased frequencies of non-conservative substitutions, and a strong negative correlation exists between impacts on protein structure and the frequency of a nucleotide substitution ($r = -0.828$, one tailed $P = 0.00044$). Accounting for differences between transitions and transversions does not alter qualitatively this result ($r = -0.749$, one tailed $P = 0.0025$).

Hence different misincorporations by the gamma polymerase [18] affect differently the coding properties of genes. The polymerase probably mainly adapted to avoid high impact nucleotide misincorporations. This can be tested by examining the correlation between misincorporation kds and the amino acid impact distances presented in [17], which will indicate to what extent these misincorporation rates resemble what is expected for pseudogenes (suggesting no selection occurs), or coding genes (suggesting the gamma polymerase is selected to minimize sub-

stitution impact on proteins). This correlation is negative, stronger than for pseudogenes, but weaker than for functional genes after selection (r = -0.434, one tailed P = 0.079). Controlling for differences between transitions and transversions does not alter much this result (r = -0.323, P = 0.15). The same holds after accounting for effects of dipole moments (Figure 2) on misincorporation rates: kds decrease with distances between replaced and replacing residues, but results are intermediate between mutation patterns observed for pseudogenes and genes that actually code for proteins (r = -0.44, one tailed P = 0.076).

This indicates that misincorporation rates include an adaptive component that minimizes the potential impact of nucleotide misincorporations on proteins. It is probable that a balance exists between minimizing different misincorporation rates, because the same active site in the polymerase is responsible for them. Hence the misincorporation pattern cannot be adapted to minimize all misincorporation rates, only to optimize misincorporation effects at protein levels. For frequencies of mutations observed in genes, selection affects each site (more or less) independently, hence impacts are minimized, resulting in much stronger correlations between mutation frequencies and impact at the protein level than observed for misincorporation rates, because the same active site produces the various types of misincorporations. The results indicate that this balancing effect due to interactions between different misincorporation types by the same active sites must be relatively strong in the gamma polymerase, otherwise the correlation with amino acid dissimilarities would resemble much more that found for coding genes. The matter of adaptively-tuned misincorporation rates by polymerases is nevertheless an interesting line of research that would gain from being developed further, including along the methods used here.

4. Gene-specific substitution matrices for human mitochondrial protein-coding genes

Misincorporation by gamma polymerases during replication is a major factor causing mutations. This factor is itself influenced by dipole moments of nucleotides, similarities between them, and greater selection pressures against specific misincorporation rates than on other rates (see previous sections). Here I examine observed mutation patterns in human mitochondrial genes.

Numbers of nucleotide substitutions for each of the 12 possible substitutions were counted from tabulations at http://www.mtdb.igp.uu.se/ [1] and http://www.mitomap.org/MITOMAP [2], separately for each gene (Table 1). Values are percentages of sites where a given mutation was observed among all sites where the substituted nucleotide mutated in that gene. The variation in that percentage within a given gene is mainly due to differences between transitions and transversions, the former dominating. Hence for further analyses, for each gene, mean percentages for transitions and transversions were calculated separately and subtracted from the observed percentages for transitions and transversions, respectively. This adjustment excludes effects due to differences between transitions and transversions in mutation percentages observed for each given gene. The two last columns in Table 1 are Pearson correlation

http://dx.doi.org/10.5772/51245

coefficients of percentages adjusted for differences between transitions and transversions and adjusted (along the same criterion) kd's of nucleotide misincorporations by the gamma polymerase (s), and after adjusting also for Grantham physico-chemical distances (s'). Correlation coefficients s are positive in 12 among 13 genes, a significant majority of cases according to a one tailed sign test ($P = 0.000854$). The correlation is significant ($P < 0.05$) at the level of a single gene for three genes, ND1, CO1 and AT8 (marked by asterisks in Table 1).

Results are only slightly altered after accounting for differences between transitions and transversions. Further analyses (s' in Table 1) using the residual misincorporation rates and the residual mutation percentages, calculated from their regressions with Grantham's amino acid dissimilarities do not change results much. These results show that variation in percentages of mutations of different types is to some extent due to misincorporation by the gamma polymerase, but a large part of the variation between substitution percentages remains unaccounted for. It is probable that natural selection against various mutations occurs, so that percentages in Table 1 are composites of misincorporation rates and other factors, such as selection against specific mutations. However, taking selection into account by using residuals from the regression of mutation frequencies with amino acid dissimilarities does not change patterns much. Hence further major factors affect observed mutation patterns, besides misincorporation rates and selection on coding impacts of mutations (and misincorporations).

5. Effects of deaminations and selection on mutation matrices

If one assumes that large parts of the variation that is not explained by the gamma polymerase's misincorporation rates in the previous analyses is due to selection, one can estimate which types of mutations are more or less prone to selection by analysing the residuals of the adjusted percentages (for each gene) from the regression with misincorporation. The line 'Res' in Table 1 indicates the number of genes for which this residual was positive, meaning that the percentage of that mutation was greater than expected from the regression with misincorporation. For two types of mutations, C->A and T->C, there were 10 such genes, which according to two tailed sign tests yields a significant tendency for observing percentages greater than expected by misincorporation ($P = 0.046$) as indicated by P in Table 1. Hence C->A and T->C are more frequent than expected by misincorporation. At least for T->C, there are two plausible explanations. T->C is a transition, and transitions cause relatively little functional effects at the level of coding properties of codons, suggesting low counterselection, hence relative over-representation (positive residuals). This explanation does not seem adequate, because the effect is not strong for other transitions (A->G, G->A and even opposite for C->T, where residuals were positive for only 2 genes ($P = 0.0095$, two tailed sign test). The latter effect on C->T is however also compatible with the second explanation for T->C. Deamination, promoted by single strandedness during replication, contributes to A->G mutations on the mitochondrial heavy strand DNA, which corresponds to T->C in Table 1 which uses the complementary light strand DNA annotation. Hence the systematic excess in T->C and systematic lack of C->T would be due to a factor that does not relate to misincorporation by the gamma polymerase, nor to selection, but presumably to the replicational

mutation gradient of A->G. Residual analysis also indicates systematic underrepresentation of a further mutation type, G->T (P = 0.0095, two tailed sign test), a transversion that might be particularly counterselected [17]. Indeed, numbers of positive residuals tend to decrease with mean physico-chemical distances between replaced and replacing amino acids associated with these nucleotide mutations (r = -0.38, not statistically significant).

Gene	A		C		G		T		A-C	A-G	A-T	C-A	C-G	C-T	G-A	G-C	G-T	T-A	T-C	T-G	s	s'
ND1	272	116	344	124	112	45	228	48	2.8	88.9	8.3	13.3	4.7	82.0	87.8	6.1	6.1	7.7	78.2	14.1	55*	53*
ND2	326	114	349	109	99	33	268	77	3.8	91.5	4.7	15.3	5.9	78.8	80.5	9.8	9.8	3.3	89.1	7.6	-5	-11
CO1	419	121	462	121	250	59	410	97	5.7	89.5	4.8	12.7	4.2	83.1	91.8	3.3	4.9	6.7	90.4	2.9	55*	47
CO2	196	65	214	59	102	39	172	55	7.7	83.1	9.2	4.3	10.0	84.7	90.2	7.3	2.4	9.4	85.9	4.7	13	6
AT8	80	42	69	31	13	9	45	26	0.0	95.2	4.8	11.1	0.0	88.9	100	0	0	0	92.9	7.1	74*	71*
AT6	206	115	230	81	71	47	174	95	4.2	90.8	5.0	8.1	5.8	86.2	90.2	7.8	2.0	3.0	87.0	10.0	40	43
CO3	210	87	249	70	116	44	209	69	11.6	84.9	3.5	6.3	3.8	90.0	93.2	4.6	2.3	9.5	87.8	2.7	45	37
ND3	102	41	102	27	37	13	105	29	18.4	73.7	7.9	13.8	6.9	79.3	84.6	15.4	0	5.1	92.3	2.6	17	3
ND4l	84	27	92	19	36	12	85	23	7.7	84.6	7.7	7.7	0	92.3	84.6	0	15.4	8.7	78.3	13.0	20	47*
ND4	416	144	473	133	137	32	352	92	10.8	84.2	5.0	10.2	5.7	84.1	87.9	12.1	0	5.6	87.9	6.5	31	15
ND5	518	207	580	183	190	49	416	117	7.8	85.5	5.7	18.7	18.7	62.6	85.6	8.1	6.3	4.9	90.9	4.2	19	3
ND6	198	53	187	72	37	21	103	32	4.8	82.5	12.7	13.4	10.5	76.1	90.9	9.1	0	3.6	92.7	3.6	27	-25
Cytb	326	142	391	141	137	67	287	95	4.3	88.7	7.0	8.6	6.2	85.2	89.7	10.3	0	1.9	92.5	5.7	12	-1
Res									4	7	7	10	5	2	6	7	2	5	10	5		
Dssh									-29	16	3	38	23	-39	-13	30	-17	-63*	35	-9		
Pos 1									-12	17	-11	15	4	-20	-6	12	-7	23	-18	2		
Pos 2									-3	-1	11	10	-15	4	-1	1	-1	-27	13	4		
Pos 3									-11	-23	53*	29	33	-35	-56*	52*	20	-39	28	-1		
Dloop									22	-39	27	-9	42	-26	20	30	-41	-31	53*	-42		
Pos 1									-13	19	-11	5	2	-7	49*	-18	-54	30	-3	-19		
Pos 2									9	-20	36	41	-37	-4	-19	13	13	-21	33	-34		
Pos 3									39	-50	38	-9	43	-20	-40	60*	-26	-46	53*	-26		

Table 1. Percentage of mutations observed in each human mitochondrial protein coding gene. A, C, G, T indicate the number of that nucleotide in that gene, followed by the number of sites with that nucleotide that are polymorphic. 's' is the Pearson correlation coefficient of percentages adjusted for differences between transitions and transversions and adjusted nucleotide misincorporations by the gamma polymerase (* indicates P < 0.05). The last lines (from Res on) and s' are explained in the text.

http://dx.doi.org/10.5772/51245

6. Mutation gradients across mitochondrial genomes

The previous section indicates that some mutations might be systematically more frequent than expected by misincorporations by the gamma polymerase, and suggests that mutations due to replicational deamination gradients could cause this effect. The study of mutational gradients has used different methods to compare mutation rates at different locations in the genome. Some studies infer mutation rates from phylogenetic comparisons among species of nucleotide contents at given sites (i.e. [13, 14]). Phylogeny-inferred kinetics for A->G and C->T gradients match the properties of the underlying chemical processes: the chemically faster C->T deamination saturates faster in computational analyses with duration spent single stranded than the slower A->G reaction [13, 14, 21]. Other studies infer mutation rates from gene nucleotide contents: for the C->T deamination, one expects relatively high C and low T contents in regions close to replication origin(s), and the opposite for genes with high durations spent single stranded [11, 12, 15, 21, 22, 23].

The method used here is closer to direct observation of mutations, because it compares only between genomes from the same species (*Homo sapiens* in this case). This means that one is closer to an 'instantaneous' observation of mutations. This procedure decreases numbers of undetected multiple changes. I did not use a full phylogenetic model of all human mitochondria to infer mutation rates. Data in Table 1 are for a simplified procedure that counts numbers of sites within a gene where a given type of mutation was observed and calculates the percentage of sites with that nucleotide where that mutation occurred, assuming that the most common nucleotide at any given site is the ancestral nucleotide.

Durations spent single stranded are calculated as previously [11, 12, 21, 22, 24]. I explored for replicational and transcriptional gradients (Dssh and Dloop in Table 1) for each of the 12 mutations, not only for A->G and C->T. This is because time spent single stranded might also affect other mutations, notably transversions [13]. Correlational analyses for gradients (analysis across rows, one per column in Table 1) used the residuals of mutation rates from their regression with misincorporation rates (residual analysis is across columns, one regression calculated per gene/row in Table 1), in order to exclude effects of polymerase inaccuracy on mutational gradients. However note that using the raw mutation percentage data as in Table 1, gradient analyses do not change much.

Two potential gradients in duration of singlestrandedness are considered, singlestrandedness during replication and during transcription (indicated in Table 1 by Dssh and Dloop, respectively). The last rows in Table 1 show Pearson correlation coefficients between residual mutation percentages and times spent single stranded during replication (Dssh) and transcription (Dloop). The hypothesis of singlestrandedness expects positive correlations, but this was observed only for half the cases, for each replication and transcription. There was a significant drop in T->A mutations along the replicational gradient, and a significant increase in T->C mutations along the transcription gradient (Figure 3). The latter effect is predicted by deamination gradients. Deamination gradients are also expected for G->A, but were not observed. Data in Table 1 only support the hypothesis of a deamination gradient for T->C. They cannot differentiate between replicational and transcriptional gradients. It is

notable that in this case the predicted G->A gradient is not stronger than gradients observed for other mutations. Apparently, another mutation, T->A, reacts to single strandedness, but in the direction opposite to that expected (singlestrandedness is predicted to increase mutations, not decrease them). Other, less direct methods based on phylogenetic reconstructions, perhaps fail to detect this gradient because selection, at larger evolutionary scale, might have weeded out many mutations such as the transversion T->A (this type of mutation implies non-conservative changes at the amino acid coding level), leaving mainly neutral and close to neutral ones. Indeed, transitions affect less coding properties than transversions (transitions cause on average more conservative amino acid changes than transversions). This would explain why phylogenetic comparisons detected weaker signals for transversion than transition gradients, while analyses in Table 1 for almost instantaneous mutations are apparently less affected by natural selection occurring after a mutation happened and do not show differences in gradients between transitions and transversions. These comparative data restricted to *Homo sapiens* confirm only the (heavy strand) deamination of A->G (corresponding to T->C in the annotation used here) at the level of a transcriptional gradient.

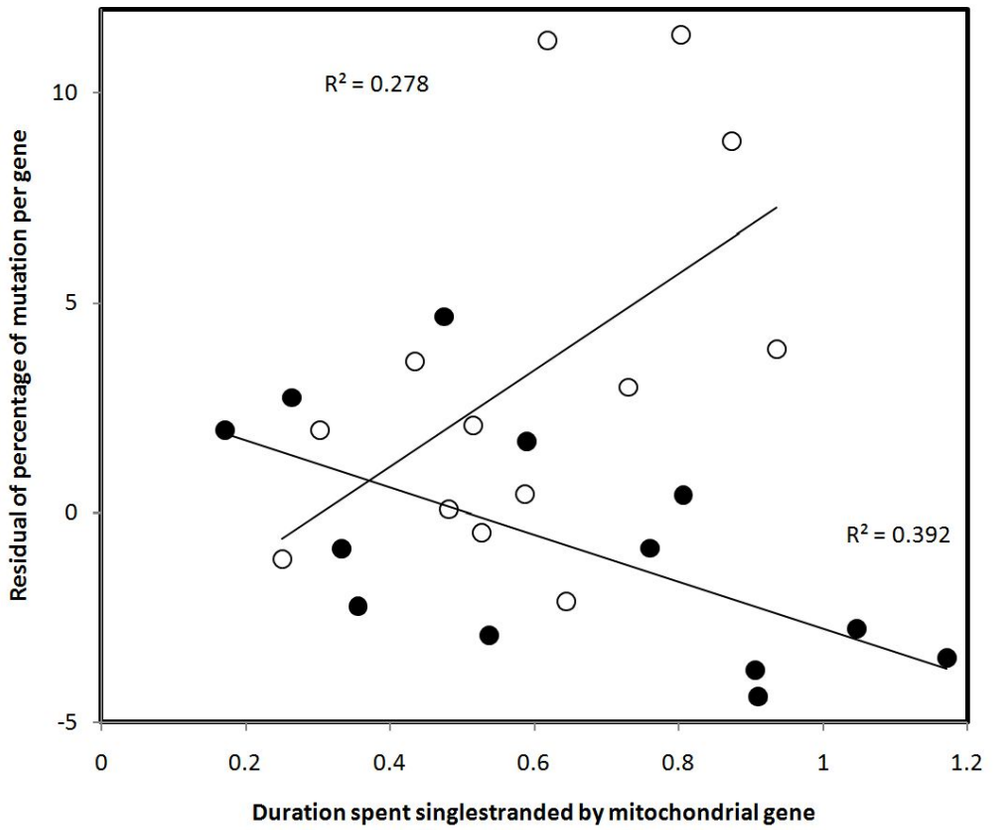

Figure 3. Mutations versus singlestranded during replication (T->A, filled symbols) and transcription (T->C, circles). Mutation percentages are residuals from regressions with misincorporation by the gamma polymerase, calculated based on data from Table 1.

The results suggest that mutation rates estimated from sequence comparisons within a single species reflect misincorporation rates, but barely confirm well established observations

http://dx.doi.org/10.5772/51245

of deamination gradients, which were based on comparisons between evolutionary more distant sequences, and on nucleotide contents of single sequences. Apparently, instantaneous mutation rates reflect misincorporation by gamma polymerase, while the effects of deamination gradients, which result from a biased cumulation of mutations, might result from long term processes and are therefore more detectable at a wider evolutionary scale.

7. Mutation gradients and selection at different codon positions

The issue of effects of selection on mutational gradients can also be investigated by analysing separately codon positions, as indicated for replicational and transcriptional gradients in Table 1. In terms of replicational gradients, there were no gradients detectable for any mutation at first and second codon positions, but there were three gradients, one negative (G->A) and two positive (A->T and G->C) at third codon positions. Hence these analyses confirm that replication gradients are more detectable where the mutation is synonymous or has little impact because causing a conservative amino acid change, as occurs at third codon positions, but not or much less at first and second codon positions. However, the specifically predicted deamination gradients are not detected. The opposite is observed for G->A (corresponding to C->T mutations on the heavy DNA strand), this mutation unexpectedly decreases along the singlestrandedness gradient, while an increase was expected.

Assuming a transcriptional gradient in singlestrandedness, the expected positive G->A gradient is detected for first codon positions. This is the only statistically significant gradient detected that is not at third codon position. The transcriptional gradient analyses at third codon position confirm the gradient observed for pooled codon positions for T->C, which fits the deamination gradient, and detects a gradient for G->C mutations.

Comparing the absolute values of the correlation coefficients in Table 1 for replicational and transcriptional gradients, correlations are stronger with transcriptional singlestrandedness, however this analysis does not account for the expected positive direction of the correlations of mutations with singlestrandedness. If one assumes that correlations should be positive (singlestrandedness should increase mutations), one does not detect any systematic difference between replication and transcription. The human mutation data might be better explained by transcriptional singlestrandedness, but the matter remains unclear. Deamination gradients are more detectable assuming transcriptional than replicational singlestrandedness, suggesting that deaminations observed in human sequences occurred mainly during transcription. The fact that more gradients are detected at third codon positions than at other positions indicates that selection against mutations affecting protein structure occurs and prevents detecting mutational gradients due to singlestrandedness.

8. Mutation gradients and misincorporations

Analyses in the previous section suggest that mutational gradients exist in mitochondria, but are less detectable at the evolutionary scale reflected by sequence variation within *Homo*

sapiens populations than when comparing between evolutionary more divergent sequences belonging to different species. Nevertheless, additional analyses show that replicational gradients confound effects of misincorporation by the gamma polymerase. Indeed, the column 's' in Table 1 shows that while mutation patterns in most genes overall fit the pattern predicted by misincorporation, this extent varies widely among genes (from -5 for ND2 to 73 for AT8). My first guess was that gene size (from 69 to over 600 codons, for AT8 and ND5, respectively) differences cause this. My assumption was that estimations of mutation patterns are less accurate in short genes, causing low correlations (low s) between observed mutation patterns and misincorppration rates. However, if this was true, one would expect a better match with misincorporation patterns in long genes, but surprisingly, patterns fit best in AT8: sampling inaccuracy does not explain variation in 's'.

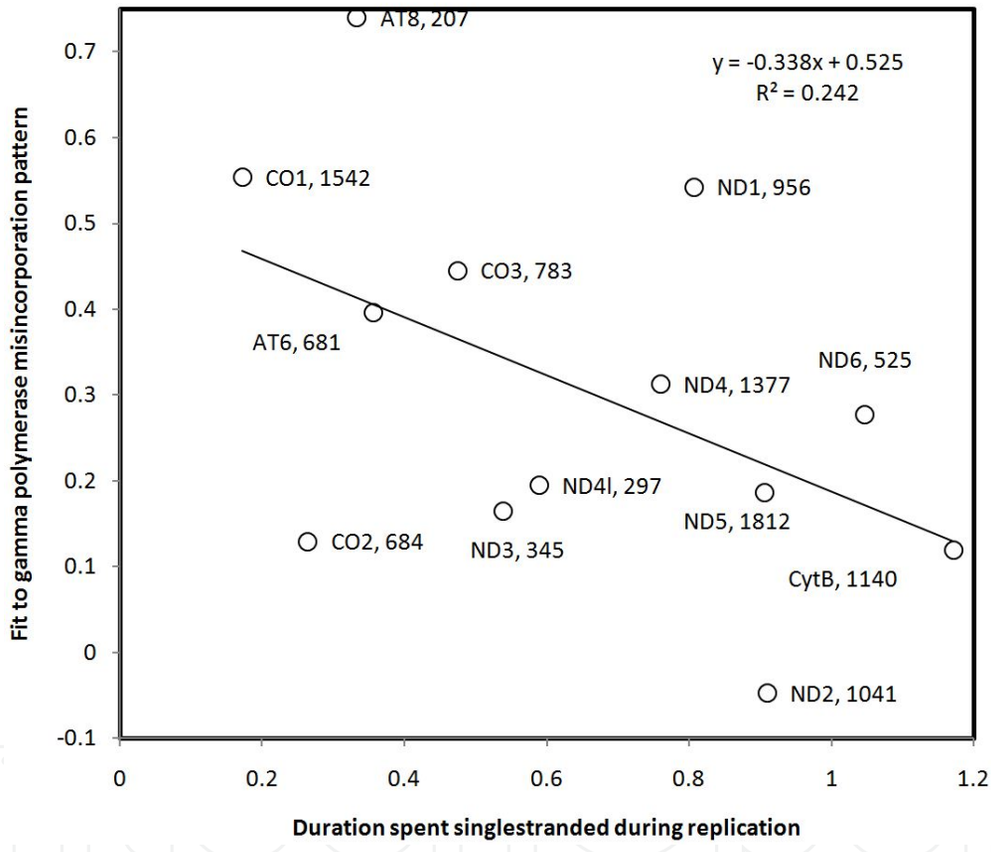

Figure 4. s from Table 1 as a function of singlestrandedness during replication. Mutation patterns resemble those predicted by misincorporation by the gamma polymerase in genes that remain singlestranded for a short time during replication. Values indicate gene lengths.

Replicational mutation gradients might explain variation in s between genes: mutation patterns in genes that endure short periods of singlestrandedness during replication should be least affected by replication gradients, and fit best the pattern predicted by gamma polymerase misincorporation, and vice versa (Figure 4). Indeed, s decreases with singlestrandedness during replication ($r = -0.49$, $P = 0.045$, one sided test; but there was no correlation of s with singlestrandedness during transcription, $r = -0.27$, $P < 0.10$).

http://dx.doi.org/10.5772/51245

Inaccurate 's' estimation due to short genes affects results in Figure 4. Short genes fit less the trend in Figure 4 than large genes (gene size is indicated in Figure 4): absolute values of residuals calculated from the regression in Figure 4 decrease with gene size (r = -0.45). Hence 21 percent of variation in s unexplained by singlestrandedness is from sampling effects. Accounting for them, the correlation in Figure 4 is r = -0.63. This means that sampling effects affect less 's' (and estimates of observed mutation rates) from Table 1 than singlestrandedness. This stresses the importance of mutational gradients despite weak results in Table 1.

Singlestrandedness during replication is an even better predictor of the fit between observed mutation patterns and gamma polymerase misincorporation when residual analyses account for each Grantham distances between replaced and replacing amino acids (s' in Table 1). This s' decreases more than s with replicational singlestrandedness (r = -0.6277, one tailed P = 0.011). Interestingly, using transcriptional singlestrandedness yields r = -0.468 (one tailed P = 0.053). Accounting for total singlestrandedness during both replication and transcription by summing both up and analysing the correlation of s' with this sum of replicational and transcriptional singlestrandedness yields r = -0.649 (one tailed P = 0.0083). In each of these analyses using s', gene size had a significant impact on residuals. Accounting for that effect systematically increased correlations between s' and replicational, transcriptional, and the combination of both singlestrandedness (r = -0.811, r = -0.68 and r = -0.89).

9. Mutation patterns: effects of dipole moments or gamma polymerase misincorporations?

Figure 2 shows that even after accounting for differences between transitions and transversions on misincorporation rates, differences between dipole moments of the substituted and the substituting nucleotides explain part of the variation in misincorporation rates. Hence both factors (dipole moment or misincorporation by the gamma polymerase) are confounded, and one cannot be sure which affects mutation patterns, or whether they affect each independently observed mutation patterns. For that reason I calculated residuals of adjusted kds from the regression with signed differences in dipole moments (data from Figure 2) and calculated correlations between these residuals and the adjusted mutation percentages (calculated from Table 1) for each gene. This version of s is adjusted for effects of dipole moments on misincorporation rates, and is positive for all genes. This adjusted s increased as compared to s from Table 1 in 8 (and decreased in 5) genes. Hence adjusting misincorporation for effects of dipole moments only slightly increases its fit with observed mutation patterns.

However, when examining the increase in s after adjusting for dipole moment effects in relation to replicational singlestrandedness, this increase is proportional to singlestrandedness during replication (not shown). This suggests that effects of the component of misincorporation that is independent of dipole moments increase with singlestrandedness. Hence singlestrandedness interacts with gamma polymerase fidelity. In these analyses, this fidelity is separated into a component associated with dipole moments, and a different component.

According to the analyses, it is the latter, unknown factor that increases its effects on observed mutation patterns with singlestrandedness during replication.

Similar residual analyses for dipole moments show that observed mutation patterns do not fit well with differences in dipole moments after calculating residuals from their regression with misincorporation rates (these analyses inverse between dependent and independent in Figure 2). These correlations were negative in 11 among 13 genes, suggesting a weak effect that is opposite to that expected by the hypothesis that mutations decrease dipole moments [4].

The latter analyses indicate that dipole moments affect mutation rates through their effects on misincorporation by polymerases, but not directly on spontaneous alterations of single stranded DNA. Misincorporation by gamma polymerase has at least two components, one related to dipole moments, and another one, unrelated to dipole moments. Effects of the latter on mutation patterns increase with singlestrandedness. Analyses in a previous section suggested that distinguishing between misincorporation due to nucleotide misrecognition versus misincorporation due to nucleotide alteration after accurate recognition could prove valuable. It is not clear whether the effect independent of dipole moments that increases with singlestrandedness relates to misrecognition, alteration after recognition, or a subcomponent of any of these. Hydrophobic bias (for low dipole moment) in relation to misincorporations by the gamma polymerase binding site for nucleotides is not explained by a simplistic analysis of the residues composing the active site of gamma polymerase. Nevertheless, these results indicate that the 'age' of the replication fork has some effect on its fidelity.

A similar comparison can be done between s and s' in Table 1. Here one sees that s', as compared to s, is lower than s in 11 among 13 genes. Hence gene-specific mutation patterns match misincorporation rates after accounting for differences between transitions and transversions better than after accounting, in addition, for selection against non-conservative amino acid replacements resulting from nucleotide substitutions: s' as compared to s decreases with singlestrandedness. Hence in this case, accounting for selection against non-conservative amino acid replacements improves slightly the match of observed mutations with misincorporations for genes with short singlestranded exposure, but mainly decreases that match for those with long singlestrandedness. This effect is opposite to the one reported in a previous paragraph for accounting for dipole moment effects. Accounting for the latter improves the match between mutation and misincorporation patterns with singlestrandedness, while accounting for selection decreases that match.

10. More evidence for complex indirect effects of mutation gradients

Sampling inaccuracy might affect estimates of s and s' in Table 1. A further potential indirect factor with opposite effect might exist. Duration spent single stranded used is for a gene's midpoint, a good approximation for short genes, but increasingly inaccurate the longer the gene. In order to evaluate this, absolute residuals of mutation percentages (Table 1) from their regressions with misincorporation rates (both adjusted for differences between transitions and transversions) are plotted versus numbers of potential sites that could mutate for that mutation type

http://dx.doi.org/10.5772/51245

in that gene (Figure 5). Residuals tend to decrease with sample sizes up to approximately 250 nucleotides, which corresponds to an average of a sequence of 1000 base pairs (for each mutation type, there is only one substituted nucleotide, so on average, these mutations occurred over a total sequence that is about four times longer). The absolute value of residual mutation percentages increases with sample size up from about 250 nucleotides.

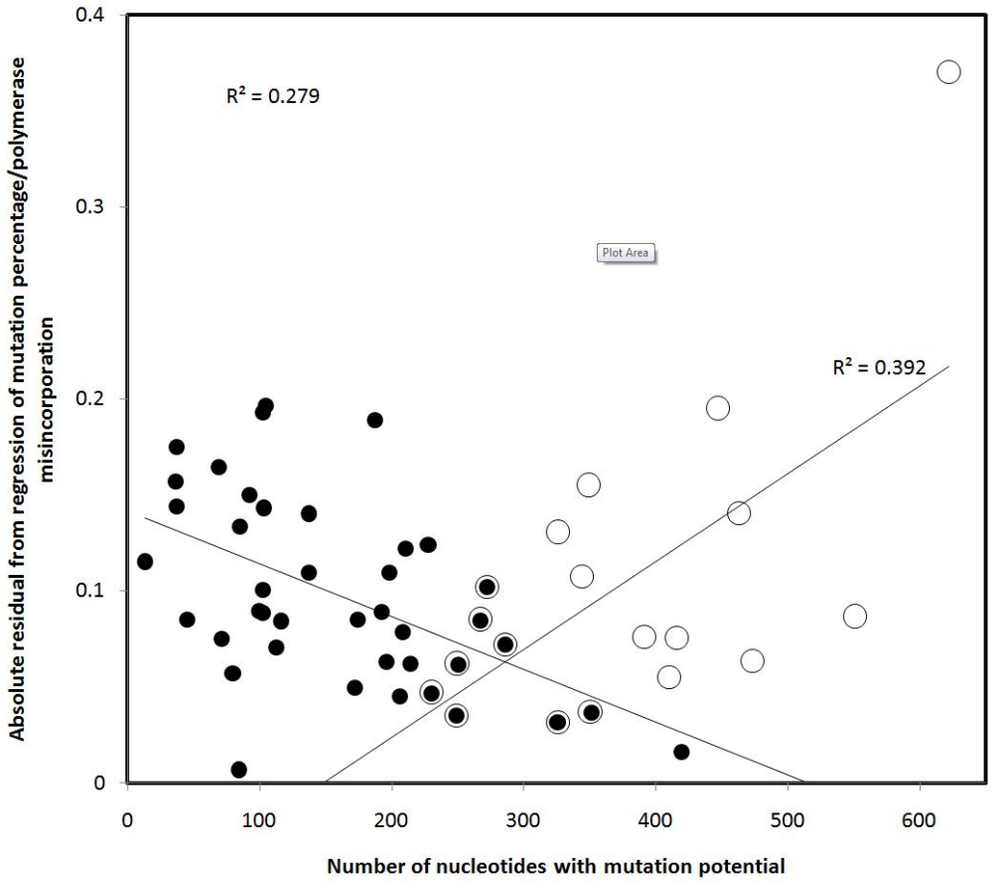

Figure 5. Residual mutation percentage (absolute value) from regression with gamma polymerase kds for each nucleotide in each gene, versus numbers of potentially mutating nucleotides. The decrease indicates a sampling effect: samples up to 200-300 nucleotides fit better misincorporation patterns because of sampling effects. Beyond 250, inaccuracy increases, perhaps because different gene regions have different mutation regimes.

The decreasing pattern is what one expects from sampling effects: up to about 1000 base pairs, longer genes enable to estimate better mutation patterns (the absolute residual is small). But for genes longer than that threshold, absolute residuals increase, hence mutation patterns tend to fit less well misincorporation as predicted by the gamma polymerase. This could be due to the mixing of regions with different singlestrandedness, which perhaps alters non-linearly mutation patterns. A similar effect where estimation inaccuracy of mutation rates decreases, then increases with sequence length exists for the correlation between rates of morphological and molecular evolution [25]. The threshold was for sequence lengths around 1200 base pairs, indicating that estimates of mutation rates (mainly from vertebrate mitochondrial protein coding sequences, as those analysed for *Homo sapiens* here) de-

creased beyond that sequence length. It was suggested, as for Figure 5 here, that mutation patterns change with the relative position of a gene, and that for long regions, more than one mutation regime might be mixed, decreasing the accuracy of analyses. Figure 5 follows that principle, and indicates a similar threshold.

11. Mutational gradients after accounting for amino acid replacement impacts on proteins

Previous sections show that indirect effects of gradients in singlestrandedness on mutation patterns exist (i.e. Figures 4 and 5). Yet analyses of mutation percentages, or mutation percentages after accounting for differences between transitions and transversions, and after accounting for effects of misincorporation by gamma polymerase, do only marginally enable to detect mutation gradients with singlestrandedness, and this for any codon position. The analyses of gradients that separate codon positions indicate that natural selection might affect mutation patterns (Table 1), and could mask mutational gradients according to singlestrandedness. Selection against non-conservative amino acid replacements also affects mutation percentages. Analyses for singlestrandedness gradients did not yet account for that latter factor, in addition to misincorporation by the gamma polymerase and differences between transitions and transversions.

I calculated residuals of mutation percentages (adjusted for differences between transitions and transversions) from their regression with mean physico-chemical (Grantham's) distances between replaced and replacing amino acids resulting from that nucleotide substitution in coding sequences (for mutation percentages across all codon positions), separately for each of the protein coding genes. Hence this analysis is across columns, for each row in Table 1. Then, for each substitution type, I calculated correlations with singlestrandedness during replication, transcription, and their sum (these analyses are across rows, for each column, on residuals produced by the latter 'row' analysis across columns). This yields correlations between residual mutation rates and singlestrandedness for each mutation type. The majority of these are positive correlations (Table 2): mutation percentages (after accounting by residual analyses for differences between transitions and trasversions, misincorporation rates and Grantham distances (assumed to reflect selection against dysfunctional proteins)) increase with singlestrandedness during replication (11 among 12 cases, exception A->C mutations), transcription (11 among 12 cases, exception A->G mutations) and their sum (all cases). Hence overall, singlestrandedness promotes all types of nucleotide substitutions, not only deaminations A->G and C->T, for both replicational and transcriptional singlestrandedness. Their sum improves correlations in half the cases. Correlations were statistically significant (one tailed $P < 0.05$) for one correlation with replicational singlestrandedness (T->C), two with transcriptional singlestrandedness (C->G and T->C) and three with the sum of both (A->T, T->C and T->G).

Correlations were stronger with transcriptional singlestrandedness than replicational singlestrandedness in 7 among 12 cases, which does not indicate which among the two is the most

http://dx.doi.org/10.5772/51245

important factor. Possibly, singlestrandedness during replication and during transcription affect differently different substitution types, or differences are random. These analyses clearly show that after accounting for mean effects of substitutions on proteins, percentages of all types of nucleotide substitutions increase with singlestrandedness during each replication and transcription. These clear patterns were not detectable without accounting for mean nucleotide substitution impact on physico-chemical properties of coded amino acids. It seems these effects prevented detecting mutation gradients for substitutions that were not deaminations. Singlestrandedness increases at least slightly probabilities of all types of substitutions.

Substitution	Rep	Trans	Both
A->C	-0.159	0.401	0.066
A->G	0.428	-0.242	0.241
A->T	0.400	0.447	0.481*
C->A	0.378	0.063	0.291
C->G	0.319	0.468*	0.433
C->T	0.243	0.315	0.312
G->A	0.091	0.339	0.216
G->C	0.395	0.322	0.421
G->T	0.425	0.340	0.448
T->A	0.301	0.381	0.382
T->C	0.478*	0.531*	0.573*
T->G	0.452	0.350	0.469*

Table 2. Pearson correlation coefficients of time spent singlestranded during replication, transcription, and their sum versus substitution percentages in the 13 human mitochondrial protein coding genes adjusted for differences between transitions and transversions, misincorporation rates and for mean effect of the substitution on Grantham's physico-chemical distances between replaced and replacing amino acids.

Causes for differences in gradient strengths for different substitution types are not known. Gradients are strongest for substitutions involving a small absolute change in nucleotide dipole moment, and weakest for those where the absolute change in dipole moment is large. Speculatively, large dipole differences may affect even when singlestrandedness is short, so that no strong gradient is detectable, because the main effect is the dipole moment, independently of singlestrandedness. For small dipole moment differences, the dipole moment effect woult hence be enhanced by singlestrandedness, resulting in a gradient.

12. Nearest neighbour effects on mutation rates

Previous analyses of mutation patterns in human mitochondrial protein coding genes fit expectations according to several factors: misincorporation by the gamma polymerase, selec-

tion against mutations that alter amino acid properties, and dipole moments of nucleotides. A hierarchy between these factors exists. In addition, they interact: misincorporation rates are also affected by selection against non-conservative mutations; and gradients in single-strandedness affect extents by which the various factors affect mutation patterns. Only after adequate accounting for misincorporation and selection (and differences between transitions and transversions), mutation gradients along durations of singlestrandedness are cleary observed for all types of nulceotide substitutions.

Flank		**5'**						**3'**					
A		**Tot**	**Mut**	**A**	**C**	**G**	**T**	**Tot**	**Mut**	**A**	**C**	**G**	**T**
	A->	968	319		15	290	14	928	238		19	193	26
	C->	1037	289	22		10	257	1069	363	21		12	330
	G->	461	166	150	11		5	371	89	79	6		4
	T->	897	278	8	244	26		1275	470	18	395	57	
C													
	A->	1097	326		36	260	30	1063	447		22	401	24
	C->	1285	327	41			270	1293	390	35		34	321
	G->	311	114	100	8		6	505	203	175	20		8
	T->	1110	245	23	203	19		981	286	17	240	29	
G													
	A->	378	158		6	145	7	447	229		12	204	13
	C->	503	156	15		28	113	322	154	21		8	125
	G->	254	72	60	8		4	256	72	68	2		2
	T->	227	83	6	67	10		323	115	6	102	7	
T													
	A->	890	425		18	373	34	888	303		20	271	12
	C->	823	252	19		21	212	1102	252	57		14	181
	G->	328	137	123	8		3	222	124	111	10		3
	T->	676	295	13	264	18		668	198	13	171	14	
	A				40	-54	1				72*	40	48*
	C					-37	36					43	60*
	G						51*						91*

Table 3. Dinucleotide sites and mutating sites in human mitochondrial protein coding sequences, separating 5′ and 3′ nucleotide identity. Last 3 lines are correlations, see text.

Despite the relative complexity of factors described and affecting mutation patterns, this is not an exhaustive list of effects on mutation rates. Notably, nearest neighbour effects exist [26], where identities of nucleotide(s) flanking the mutating site affect mutation rates, as in-

http://dx.doi.org/10.5772/51245

dicated by the editor of this volume after reviewing a former version of this chapter. G and C, the nucleotides with the highest dipole moments, seem to increase mutation rates in various organisms along similar patterns [26-30]. This suggests a physico-chemical basis for nearest neighbour effects, possibly along the lines of dipole moment effects and the stability of DNA duplexes surrounding the mutating nucleotide [26]. These biases are strong enough to justify the need of incoprorating at least the strongest nearest neighbour effect in models designed to detect natural selection on mutations [31], which is not surprising as CpG dinucleotides are disproportionately represented among sites with pathogenic polymorphisms [32,33]. Moreover, nearest neighbour effects interact with gene location and the frequency of transcription, suggesting interactions with singlestrandedness [34, 35]. Nearest neighbour analysis of mutation patterns requires large sample sizes, and therefore is unfortunately incompatible with a gene by gene analysis as a function of singlestrandedness in the context of this mitochondrial dataset.

However, even after pooling mutation data from all genes, one would ideally examine the twelve substitutions in relation to each of the 16 combinations of nucleotides at the 5′ and 3′ positions. Such detailed analyses are also not possible with this dataset. Nevertheless, as known to this author, nearest neighbor effects have not yet been examined in the context of mitochondrial genomes, hence even simplified analyses pooling mutations from all genes and codon positions together may still be valuable. In addition, most nearest neighbour analyses examined do not analyse substitutions in relation to their direction (they pool X->Y with X<-Y), but this can be done on this dataset. Mutation data from all genes and codon positions were pooled, and analysed each time separately in relation to the identity of their 5′, and their 3′ flanking nucleotide. This yields reasonable samples, and the mutation patterns can be compared according to the different flanking nucleotide identities (Table 3).

The data in Table 3 enable a number of different analyses, only one is presented here, though many others are of interest. For example, biases exist in terms of dinucleotide frequencies, between 5′ or 3′ flanking by the same nucleotide. I focus here on the analysis of mutation patterns. Numbers in each row in Table 3 were divided by the number of mutating sites among all possible dinucleotide sites for that category (Mut). The column Tot in Table 3, which indicates the total number of dinucleotide sites found independently of the occurrence of a mutation at that site, is indicated but not used in further analyses. The substitution matrices that result are very similar, comparing 5′ and 3′, and different nucleotide contexts. This is because the overwhelming majority of the variation in mutation rates is due to the difference between transitions and transversions. For that reason, effects of transitions versus transversions were accounted for by subtracting observed mutations rates from the average for transitions and transversions, respectvely, as done in previous analyses. Then these data adjusted for differences between transitions and transversions are compared between different substitution matrices, so that effects of the difference between transitions and transversions is accounted for before comparing the matrices with different neighbours.

The three last lines in Table 3 show Pearson correlation coefficients (x100) between these mutation patterns (adjusted for differences between transitions and transversions). Even after accounting for differences between transversions and transitions, substitution patterns

across different 3′ neighbouring nucleotides resemble each other: all six correlations are positive, 4 among these are statistically significant ($P < 0.05$, one tailed tests because positive associations are expected, see asterisks in Table 3). 3′ G and T had most similar patterns. Hence grouping of mutation patterns according to 3′ nucleotides does not follow purine/pyrimidine nor dipole moment differences. 3′ G seems to affect most mutation patterns, hence results are probably not random also for 3′ nearest neighbour effects.

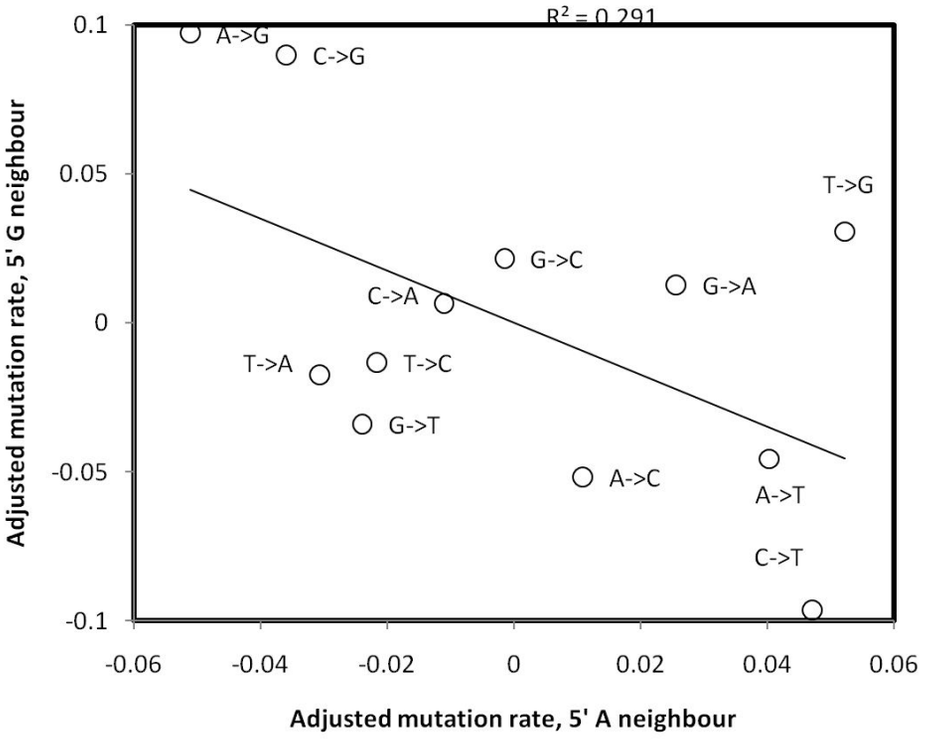

Figure 6. Mutation rates adjusted for differences between transition and transversions for 5′ A and G neighbours in human mitochondrial protein coding genes.

The same analysis for 5′ flanking nucleotides reveals a similar, more enhanced situation. Here, the only statistically significant association is between mutation patterns with 5′ G and T as nearest 5′ neighbour. The weak positive correlations in the 3′ context are negative in the 5′ context, one being close to statistically significant (the comparison between 5′ A and G, see Figure 6): 5′ G affects mutation rates in a way that tends to be systematically opposite to what is observed in other contexts, so that relatively high mutation rates become relative low, and vice versa. Effects of 5′ G on mutation rates are expected, considering previous reports. However, these have mainly shown effects on C->T mutations. The results here show that 5′ G has a systematic effect on all mutation types, some increasing, as expected, but others decreasing in the 5′ G context.

It is notable that the correlation matrices for 5′ and 3′ contexts (in the 3 last lines of Table 3) are very similar, if not in their values, but in their pattern: the ranks, from least to most positive correlation coefficients, are identical (Figure 7). This means that the same effects are at

http://dx.doi.org/10.5772/51245

work for 5′ and 3′ flanks, but that effects are stronger for 5′ flanking nucleotides. In this context, it is important to remember that the annotation used here is that of the light strand DNA in the mitochondrion, which bears the coding sequence of most genes. In the elongating light DNA strand, the 3′ nucleotide is already present before the mutating nucleotide is added, while the 5′ nucleotide is not yet there, and could not possibly have any effect. This is not compatible with a 5′ effect during replication, unless one considers that the effect is from the neighbouring nucleotide on the template heavy strand DNA. In that case, the inverse complement would have the major flanking effects, with the strongest effect by the nucleotide that is not yet complemented by the nascent strand (the 5′ of the light strand becomes the 3′ in the heavy strand), and a weaker but similar effect by the neighbouring nucleotide that is already complemented by the replication process. Along that scenario, neighbouring nucleotides would affect misincorporation rates. This scenario would be very compatible with electrostatic effects, due to dipole moments.

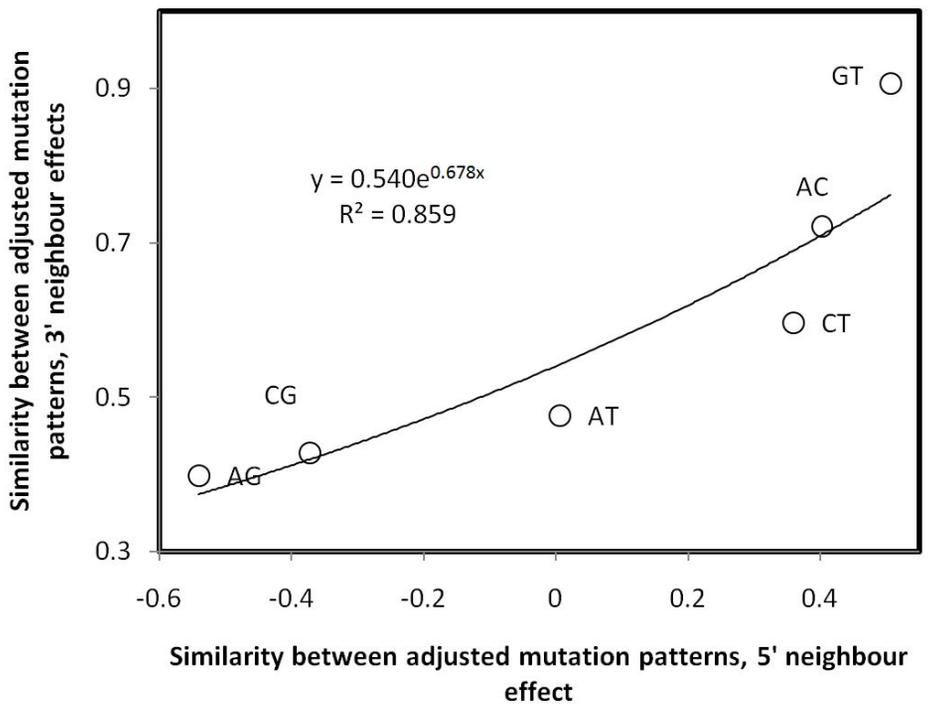

Figure 7. Similarities (Pearson correlation coefficients in the three last lines of Table 3) between transition versus transversion adjusted mutation patterns for 3′ neighbouring nucleotides as a function of similarities for mutation patterns found for 5′ neighbouring nucleotides (see Table 3). Letters near datapoints indicate the neighbouring nucleotides whose mutation patterns are compared.

It is notable that the 5′ G mutation pattern is very similar to the 3′ C mutation pattern as these are observed for the light strand (r = 0.87). These are the most similar mutation patterns found when comparing 5′ and 3′ mutation patterns. Because 3′ C on the light strand is 5′ G on the heavy strand, this similarity indicates that the factor at work involves both strands, always involving the 5′ G nucleotide.

Alternative explanations not involving effects on misincorporation rates, such as dipole moments and 'spontaneous' (non-enzymatic) mutations are also very plausible. The latter are more compatible with the similarities in patterns between 5′ and 3′ effects and effects on both strands, but less with the strong directional effect detected (less similar mutation patterns between 5′ than 3′ substitution patterns).

Hence strong neighbouring effects are detected on mutation patterns observed in human mitochondrial genomes, yet their cause remain unknown, and might, as for other effects on mutation patterns, have different physico-chemical causes, combined with some biological factors.

13. Dipole moments and retrotranscription rates by the gamma polymerase

The various analyses described show complex effects, most of them are confirmative of phenomena that have already been described. Indeed, it is quite trivial that misincorporation rates affect mutation frequencies, and that these frequencies are decreased by selection against dysfunctional proteins. Gradients in singlestrandedness as affecting mutation rates are also known, though the fact that they affect all or most mutation types is relatively original to the analyses presented here. A similar rationale relates to the original component of the results from the nearest neighbour analyses. However, the fact that so many different factors are jointly considered in the analysis of a single dateset of mutations is not the sole major originality in terms of potential mechanisms explored in this chapter.

The hypothesis that dipole moments affect misincorporation, mutation rates, and mutation gradients, is a major potential novelty. Unfortunately, when its effects are detected, these are not well understood: the main effect on misincorporation is that of absolute dipole moment change, and the bias favouring low dipole moments remains unexplained.

The suggestion by David Stuart, the editor of this volume, to examine associations between dipole moments and elongation rates in relation to the inserted nucleotide [36] yields interesting results in this respect, confirming that dipole moments affect the incorporation rate of nucleotides into nascent DNA. Results below indicate complex mechanisms, and should be considered as preliminary and with extreme caution. First, it seems that kms of incorporations of nucleotides increase with dipole moments, as one would expect if high dipole moments enable quick processing by the (charged and hydrophilic) active site of the gamma polymerase, though this effect is not statistically significant at $P < 0.05$.

However, electron singlets and or triplets of molecules can be in an 'excited' state, which modifies the dipole moment of the molecule, as calculated by Bergmann and Weiler-Feilchenfeld [6] (therein table VII) for nucleotides. Dipole moments for the excited triplet state correlate positively with nucleotide insertion rates ($r = 0.9865$, $P = 0.007$, one tailed test). Considering that more than one correlation test was done (for the regular and the two excited dipole moments), this result is not statistically very strong (especially that only 4 datapoints are involved in the analysis).

http://dx.doi.org/10.5772/51245

I assume that implicitly, the hypothesis developed by the editor, following my initial interest in effects of dipole moments on polymerase activity, is that if dipole moments affect nucleotide incorporation rates, discrimination against incorporation of the much more common ribonucleotides should associate negatively with (deoxyribo)nucleotide dipole moment. Indeed, the activity of the gamma polymerase as measured by Kasiviswanathan and Copeland [36] includes also the (mis?) incorporation rates of ribonucleotides on the template of DNA, and these associate negatively with nucleotide triplet excited dipole moment ($r = -0.963$, one tailed $P = 0.019$). In addition, the rate of reverse transcription by the gamma polymerase, where deoxyribonucleotides are inserted on the template of (mis)inserted ribonucleotides correlates positively with the mean of the singlet and triplet excited dipole moments ($r = 0.99984$, one tailed $P = 0.00008$, see Figure 8).These analyses yield notable results, though they are not necessarily as statistically robust as they seem, due to the low number of degrees of freedom (only four datapoints). In addition, correlations between each of the kms and each the dipole moment, the singlet and excited, and their average was calculated, in total 12 correlations. In these cases, according to a strict Bonferroni criterion to correct for multiple testing, to keep $P < 0.05$ while testing 12 times a hypothesis, one should use the threshold of $P = 0.05/12 = 0.0042$. According to that often overconservative criterion, only the result in Figure 8 remains statistically significant.

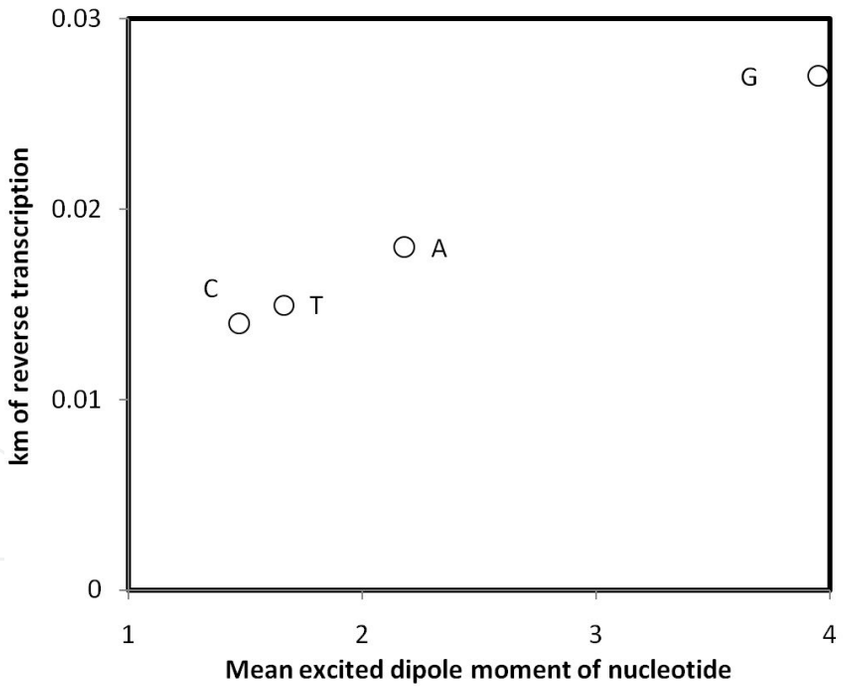

Figure 8. Rate of deoxyribonucleotide 'reverse' incorporation as a function of its mean excited dipole moment on the template of ribonucleotide.

The data nevertheless confirm the hypothesis that nucleotides are processed on the basis of their dipole moment, where nucleotides with high dipole moments are more rapidly correctly processed. This result might actually explain also the results obtained in earlier sec-

tions on associations between dipole moment changes and nucleotide misrecognitions. The rate of a process, and its accuracy are frequently negatively associated. Hence if correct incorporation is proportional to dipole moments, misincorporation might be (as observed) inversely proportional, explaining that patterns in Figures 1 and 2 are opposite to predictions: the hydrophilic active site will handle correctly more rapidly a nucleotide with high dipole moment, and more probably mishandle a nucleotide with low dipole moment.

14. General discussion

The analyses presented above show that mutation patterns estimated from the simple comparison between sequences from a species confirm the patterns expected from experimentally determined misincorporation rates for the gamma polymerase. This is an important confirmation that comparative analyses yield trustable estimates of mutation patterns and rates. Analyses support, to lesser extents, that mutation patterns across genes are determined by durations spent singlestranded, and suggest that in order to detect such effects, comparisons involving longer evolutionary time spans than those implied by separations between different individuals from a single species are required to detect the cumulation of mutations due to singlestrandedness.

Grantham's physico-chemical distances between replaced and replacing amino acids affect misincorporation rates by the gamme polymerase, and percentages of mutations observed in protein coding genes. This suggests that natural selection to conserve protein function affects each of these two different patterns. Gradients of mutations with singlestrandedness are barely detectable without controlling for effects of Grantham distances on mutation percentages, several indirect effects are observable that indicate interactions between singlestrandedness and misincorporation patterns by the gamma polymerase. After the effects of Grantham distances on mutation percentages are accounted for by residual analyses, the expected increase in mutation percentages with singlestrandeness becomes detectable in all types of substitutions. This is a notable result, because singlestrandedness was believed until now to affect only or mainly substitutions due to deaminations (A->G and C->T). Analyses do not succeed to indicate which of replicational and transcriptional singlestrandedness is most relevant to predict mutations. Further chemical processes accounting for effects of singlestrandedness on mutation types besides deaminations have to be investigated and suggested.

It seems that gamma polymerase misincorporation patterns change with single strandedness, which may reflect the duration of activity by the replication fork's molecular 'machinery'. Only molecular experiments much more developed than those used until now could yield such results. This shows that combined analyses of bioinformatic and experimental data enable to suggest the existence of previously unknown biochemical phenomena.

Further points raised are the involvement of nucleotide dipole moments in the interactions between the nucleotide and the gamma polymerase. Analyses at this point do not yield much information beyond the fact that such effects occur. More functional hypotheses could help in this respect.

http://dx.doi.org/10.5772/51245

The data on mutation patterns from *Homo sapiens* do not enable to establish whether mutations cumulate during singlestrandedness due to transcription or replication. It is probable that the ratio between these two types of events that open double stranded DNA, changes with the longevity of an individual/species, where greater lifespan increases the transcription component [11]. It is probable that analyses similar to those done here based on ample mitochondrial sequence data available for other mammal species with shorter lifespans could help in this respect. Comparing results from different species would probably be fruitful. In addition, these analyses could preliminarily reveal whether misincorporation patterns by the gamma polymerases of the different species differ. This could be an exciting line of research, that could potentially link differences in mutation patterns with differences in the gamma polymerases from these species. Such analyses could yield a workable model for the efficiency and fidelity of gamma polymerase in relation to its detailed structure. It is notable that much information necessary for such analyses is already available online and only awaits the interest of enthousiastic students of molecular biology.

Nearest neighbour effects as detected for mitochondrial mutation patterns confirm what is known from previous studies on nuclear chromosomes. They also show that the 5′ G effect on mutation rates is more complex, as it affects differently different types of mutations. Unfortunately, nearest neighbour analyses require samples that are not compatible with the data at hand, so that its analysis in combination with other factors could not be done. The fact that nearest neighbour effects tend to increase (though marginally so), with the thermodynamic stability of the DNA duplex where these neighbouring effects occur, is in itself compatible with dipole moment effects as the causes for the nearest neighbour effects on mutations because nucleotide dipole moments predict duplex thermodynamic stabilities [37]. It is possible that the direct cause for this is thermodynamic stability, through the fact that regions forming stable duplexes are more able to tolerate a misinserted nucleotide. But the association between nearest neighbour effects and stability is weak, indicating that another, associated factor is at work. Possibly, it is the electrostatic effect of nucleotide dipole moments of neighbouring nucleotides on the fidelity of the gamma polymerase that causes these effects. Such effects are particularly probable, considering that the gamma polymerase active site includes two charged residues, and that nucleotide processing seems to depend to some extent on the nucleotide's dipole moments. Hence nearest neighbour effects could be due to interferences between the electrostatic fields of the active site, the incorporated nucleotide, and the nearest neighbours, especially when these nearest neighbours have high dipole moment.

Beyond effects of nucleotide dipole moments on incorporation rates, results suggest natural selection decreasing nucleotide misincorporations with high impact on protein structure. These are encouraging results that could yield further insights if similar analyses are applied to different types of polymerases.

Variation in mutation patterns for genes with different locations along singlestrandedness gradients might have explanations that differ from the ones suggested. The genome structure might be designed so that genes that cannot afford, from a functional point of view, large mutation rates, are located so as to endure little singlestrandedness. It is important to remember that this factor might interact with the results presented. It is also important to

remember that natural selection probably affects observed mutation frequencies in ways not accounted for by presented analyses. This effect might be weaker in genes located far from replication origins, and hence probably more able to tolerate mutations. Hence observed mutations would in these cases much more reflect the original processes, not confounded by effects of natural selection due to gene function.

A further point relates to the patterns observed in Figure 5, where gene length seems to affect the mutation pattern. A possible factor here is the capacity of longer sequences to form more secondary structures by self-hybridization. Considering that secondary structure protects against mutations due to singlestrandedness, this factor could hence indirectly affect mutation patterns, especially in longer genes, assuming that in some ways, genes are replicated as functional units, a possibility that cannot be ruled out *a priori*, especially if secondary structure formation is designed to involve a gene as a unit, for example in the mRNA [38]. It is also possible that secondary structures affect the function of the gamma polymerase, causing differences in misincorporation patterns between regions forming more or less secondary structures, as previous analyses possibly indicated [13, 14].

An important point to stress here is that the data that are available at this point do not limit our capacities to analyses, along multiple dimensions, the various factors that cause mutation patterns, and understand their details in relation to these factors. The computational power and statistical tools are also not limiting and close to adequate. The limiting factor is the time invested by the adequately skilled manpower, or more correctly, the financial investment to support such activity based mainly on analysing valuable molecular data of different types.

15. Conclusions

Combined analyses of comparative sequence data and experimentally determined gamma polymerase misincorporation data, together with models for substitutions based on nucleotide dipole moments and models for substitution impacts on protein structure reveal that observed human mitochondrial protein coding gene mutation patterns are affected in decreasing order of importance by gamma polymerase misincorporation rates, selection against non-conservative amino acid replacements, and gradients in singlestrandedness during replication and transcription. Gamma polymerase misincorporation rates are selected to optimize effects of substitutions on non-conservative amino acid replacements, and favour nucleotides with low dipole moments, suggesting hydrophobic bias in nucleotide misbinding. Further analyses confirm this: the hydrophilic active site of the gamma polymerase handles faster nucleotides with high dipole moment and mishandles more often those with low dipole moment, suggesting that process accuracy limits its rate. The wealth of results confirms known and expected patterns, and expands beyond them, revealing selection on polymerase fidelity, and spontaneous tendencies during single stranded DNA states for all substitutions, not only those previously known to react to singlestrandedness.

Author details

Hervé Seligmann*

Address all correspondence to: podarcissicula@gmail.com

The National Collections of Natural History at the Hebrew University of Jerusalem, Israel

References

[1] Ingman, M., & Gyllenstein, U. (2006). mtDB: Human mitochondrial genome database, a resource for population genetics and medical sciences. *Nuc Acids Res*, 34, D749-D751.

[2] Ruiz-Pesini, E., Lott, M. T., Procaccio, V., Poole, J., Brandon, M. C., Mishmar, D., Yi, C., Kreuziger, J., Baldi, P., & Wallace, D. C. (2007). An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nuc Acids Res*, 35(Database issue: D823-D828).

[3] Kaguni, L. S. (2004). DNA polymerase gamma, the mitochondrial replicase. *Ann Rev Biochem*, 73, 293-320.

[4] Seligmann, H. (2006). Error propagation across levels of organization: From chemical stability of ribosomal RNA to developmental stability. *J Theor Biol*, 242, 69-80.

[5] Li, W. H., Wu, C. I., & Luo, C. C. (1984). Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J Mol Evol*, 21, 58-71.

[6] Bergmann, E. D., & Weiler-Feilchenfeld, H. (1973). The dipole moments of purines and pyrimidines. Chapter 1. *In: Duchesne J (ed) Physico-chemical properties of nucleic acids*, I., Academic Press, London.

[7] Frederico, L. A., Kunkel, T. A., & Shaw, B. R. (1990). A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochem*, 29, 2532-2537.

[8] Frederico, L. A., Kunkel, T. A., & Shaw, B. R. (1993). Cytosine deamination in mismatched base-pairs. *Biochem*, 32, 6523-6530.

[9] Francino, M. P., Chao, L., Riley, M. A., & Ochman, H. (1996). Asymmetries generated by transcription-coupled repair in enterobacterial genes. *Science*, 272, 107-109.

[10] Francino, M. P., & Ochman, H. (1993). Deamination as the basis of strandasymmetric evolution in transcribed Escherichia coli sequences. *Mol Biol Evol*, 18, 1147-1150.

[11] Seligmann, H. (2011). Mutation patterns due to converging mitochondrial replication and transcription increase lifespan, and cause growth rate-longevity tradeoffs. *In: Seligmann H. (ed.) DNA Replication-Current Advances*, Rijeka, InTech, Chapter 6, 151-180.

[12] Seligmann, H. (2012). Coding constraints modulate chemically spontaneous mutational replication gradients in mitochondrial genomes. *Curr Genomic*, 13, 37-54.

[13] Krishnan, N. M., Seligmann, H., Raina, S. Z., & Pollock, D. D. (2004). Detecting gradients of asymmetry in site-specific substitutions in mitochondrial genomes. *DNA & Cell Biol*, 23, 707-714.

[14] Krishnan, N. M., Seligmann, H., Raina, S. Z., & Pollock, D. D. Phylogenetic analysis of site-specific perturbations in asymmetric mutation gradients. *In: A. Gramada, and P.E. Bourne (eds.) Currents in Computational Molecular Biology*, ACM Press, San Diego, CA, 266-267.

[15] Seligmann, H., Krishnan, N. M., & Rao, B. J. (2006). Possible multiple origins of replication in primate mitochondria: alternative role of tRNA sequences. *J Theor Biol*, 241, 321-332.

[16] Madariaga, S. T., Contreras, J. G., & Seguel, C. G. (2005). Interaction energies in non Watson-Crick pairs: an ab initio study of G U and U U pairs. *J Chil Chem Soc*, 50, 435-438.

[17] Gojobori, T., Li, W. H., & Graur, D. (1982). Patterns of nucleotide substitution in pseudogenes and functional genes. *J Mol Evol*, 18, 360-369.

[18] Lee, H. R., & Johnson, K. A. (2006). Fidelity of the human mitochondrial DNA polymerase. *J Biol Chem*, 281, 36236-36240.

[19] Lee, Y. S., Kennedy, W. D., & Yin, Y. W. (2009). Structural insight into processive human mitochondrial DNA synthesis and disease-related polymerase mutations. *Cell*, 139, 312-324.

[20] Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science*, 185, 862-864.

[21] Seligmann, H. (2008). Hybridization between mitochondrial heavy strand tDNA and expressed light strand tRNA modulates the function of heavy strand tDNA as light strand replication origin. *J Mol Biol*, 379, 188-199.

[22] Seligmann, H., Krishnan, N. M., & Rao, B. J. (2006). Mitochondrial tRNA sequences as unusual replication origins: pathogenic implications for Homo sapiens. *J Theor Biol*, 243, 375-385.

[23] Seligmann, H. (2010). Mitochondrial tRNAs as light strand replication origins: similarity between anticodon loops and the loop of the light strand replication origin predicts initiation of DNA replication. *Biosystems*, 99, 85-93.

[24] Tanaka, M., & Ozawa, T. (1994). Strand asymmetry in human mitochondrial mutations. *Genomics*, 22, 327-335.

[25] Seligmann, H. (2010). Positive correlations between molecular and morphological rates of evolution. *J Theor Biol*, 264, 799-807.

[26] Krawczak, M., Ball, E. V., & Cooper, D. N. (1998). Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *Am. J. Hum. Genet.*, 63, 474-488.

[27] Zhao, Z., & Boerwinkle, E. (2002). Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res*, 12, 1679-1686.

[28] Zhanga, F., & Zhao, Z. (2004). The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs. *Genomics*, 84, 786-796.

[29] Zhao, H., Li, Q.-Z., Zeng, C.-Q., Yang, H.-M., & Yu, J. (2005). Neighboring-nucleotide effects on the mutation patterns of the rice genome. *Geno. Prot. Bioinfo.*, 3(3).

[30] Zhang, W., Bouffard, G. G., Wallace, S. S., & Bond, J. P. (2007). NISC Comparative Sequencing Program. Estimation of DNA sequence context-dependent mutation rates using primate genomic gequences. *J Mol Evol*, 65, 207-214.

[31] Suzuki, Y., Gojobori, T., & Kumar, S. (2009). Methods for incorporating the hypermutability of CpG dinucleotides in detecting natural selection operating at the amino acid sequence level. *Mol Biol Evol*, 26, 2275-2284.

[32] Cheunga, L. W. T., Leeb, Y. F., Ngb, T. W., Chingb, W. K., Khooc, U. S., Ngd, M. K. P., & Wong, A. S. T. (2007). CpG/CpNpG motifs in the coding region are preferred sites for mutagenesis in the breast cancer susceptibility genes. *FEBS Letters*, 681, 4668-4674.

[33] Antonarakis, S. E. (2006). *CpG Dinucleotides and Human Disorders. eLS*.

[34] Misawa, K. (2011). A codon substitution model that incorporates the effect of the GC contents, the gene density and the density of CpG islands of human chromosomes. *BMC Genomics*, 12, 397.

[35] Chen-L, C., Rappailles, A., Duquenne, L., Huvet, M., Guilbaud, G., Farinelli, L., Audit, B., d'Aubenton-Carafa, Y., Arneodo, A., Hyrien, O., & Thermes, C. (2010). Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res*, 20, 447-457.

[36] Kasiviswanathan, R., & Copeland, W. C. (2011). Ribonucleotid discrimination and reverse transcription by the human mitochondrial DNA polymerase. *J Biol Chem*, in press.

[37] Seligmann, H., & Amzallag, G. N. (2002). Chemical interactions between amino acid and RNA: multiplicity of the levels of specificity explains origin of the genetic code. *Naturwissenschaften*, 89, 542-551.

[38] Krishnan, N. M., Seligmann, H., & Rao, B. J. (2008). Relationship between mRNA secondary structure and sequence variability in chloroplast genes: possible life history implications. *BMC Genomics*, 9, 48.

IntechOpen