Unbiased Mitoproteome Analyses Confirm Non-canonical RNA, Expanded Codon Translations

Herve Seligmann

Computational and Structural Biotechnology Journal 14 (2016) 391–403

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Unbiased Mitoproteome Analyses Confirm Non-canonical RNA, Expanded Codon Translations

Hervé Seligmann

Unité de Recherche sur les Maladies Infectieuses et Tropicales Émergentes, Faculté de Médecine, URMITE CNRS-IRD 198 UMER 6236, Université de la Méditerranée, Marseille, France

ARTICLE INFO

Article history:
Received 19 July 2016
Received in revised form 28 September 2016
Accepted 29 September 2016
Available online 5 October 2016

Keywords:
Frameshift
Bijective transformation
Digestive enzymes
Unbiased analyses
RNA–DNA difference

ABSTRACT

Proteomic MS/MS mass spectrometry detections are usually biased towards peptides cleaved by experimentally added digestion enzyme(s). Hence peptides resulting from spontaneous degradation and natural proteolysis usually remain undetected. Previous analyses of tryptic human proteome data (cleavage after K, R) detected non-canonical tryptic peptides translated according to tetra- and pentacodons (codons expanded by silent mono- and dinucleotides), and from transcripts systematically (a) deleting mono-, dinucleotides after trinucleotides (delRNAs), (b) exchanging nucleotides according to 23 bijective transformations. Nine symmetric and fourteen asymmetric nucleotide exchanges (X ↔ Y, e.g. A ↔ C; and X → Y → Z → X, e.g. A → C → G → A) produce swinger RNAs. Here unbiased reanalyses of these proteomic data detect preferentially non-canonical tryptic peptides despite assuming random cleavage. Unbiased analyses couldn't reconstruct experimental tryptic digestion if most detected non-canonical peptides were false positives. Detected non-tryptic non-canonical peptides map preferentially on corresponding, previously described non-canonical transcripts, as for tryptic non-canonical peptides. Hence unbiased analyses independently confirm previous trypsin-biased analyses that showed translations of del- and swinger RNA and expanded codons. Accounting for natural proteolysis completes trypsin-biased mitopeptidome analyses, independently confirms non-canonical transcriptions and translations.

1. Introduction

Protein sequences are more complex than texts written in natural human languages [1]. This implies that genes include superimposed information, overprinted on the classical protein coding gene; for example, the three frames of the shortest self-replicating circular RNA virusoid code for proteins [2]. Cryptic coding revealed by frameshifts also implies that a punctuation code regulates ribosomal frame translation. This role seems fulfilled by the natural circular code X, a set of 20 regular codons overrepresented in the coding frame of genes versus other frames. X possesses peculiar mathematical properties that enable retrieval of translational frame [3–10]. This punctuation code can also be considered as cryptic superimposed information.

The natural circular code seems to prevent unwanted ribosomal frameshifts. In addition, the structure of the genetic code implies a further mechanism against frameshifted translation. This mechanism, rather than preventing ribosomal slippage before it occurs, as assumed for X, minimizes translation after ribosomal frameshifts. This is because the genetic code's codon-amino acid assignments are such that they maximize off frame stop codons [11], avoiding metabolic waste after ribosomal slippages [12–16].

Superimposed coding is also indicated by other peculiar genetic code properties. The genetic code includes symmetries, such as Rumer's symmetry [17–20], where transformations of nucleotides into other nucleotides along specific rules reveal symmetries between codon-amino acid families.

Rumer's symmetry implies that after applying that specific transformation, all codons coding for a given amino acid are transformed into codons coding for another amino acid [21]. These theoretical observations seem related to the following empirical observations. Recent transcriptomic and proteomic findings show that genetic information is revealed by each systematic frameshifting and nucleotide transformations. Here I develop these issues and present analyses that strengthen the proteomic evidence for translation of proteins coded by overprinting associated with systematic frameshifts and systematic nucleotide transformations.

Previously detected peptides match predictions of regular translations of non-canonical mitochondrial RNAs, and non-canonical translations of codons expanded by silent mono- and dinucleotides (detailed in Fig. 1). These results seem overall robust as detected non-canonical peptides mapped on the human mitogenome with corresponding non-canonical RNAs [22,23]. Hence existences of non-canonical RNAs and peptides are validated by independent detections of non-canonical RNAs and peptides, and by convergences (associations) between detected non-canonical RNAs and peptides.

E-mail address: hselig1@gmail.com.

http://dx.doi.org/10.1016/j.csbj.2016.09.004

Systematic transformations during transcription (alternating codons underlined)

A)	Original sequence	AAACCCTTTGGG
	Translated	K P F G
B)	A<->C-swinger transformed sequence	CCCAAATTTGGG
	Translated	P K F G
C)	Expanded to tetracodons	CCCAAATTTGGG
	Translated	P K W
D)	Systematic mononucleotide deletions	CCC_AAT_TGG_
	Translated	P K W
E)	Expanded to pentacodons	CCCAAATTTGGG
	Translated	P I G
F)	Systematic dinucleotide deletions	CCC___ATT__GG
	Translated	P I G

Fig. 1. Sequence (A) and its systematic transformations and corresponding translations (B–F). B) A ↔ C systematic nucleotide exchange of sequence in A; C) assuming systematic codon expansion by silent mononucleotides; D) assuming systematic mononucleotide deletion after each trinucleotide (translation identical to that in C); E) assuming systematic codon expansion by silent dinucleotides; F) assuming systematic dinucleotide deletion after each trinucleotide (translation identical to that in E). RNAs and peptides corresponding to these alternative transcriptions and translations have been previously described for human mitochondria [22,23]. For swinger transformations, A ↔ C is only one among 23 possibilities, nine symmetric of type X ↔ Y, and 14 asymmetric, of type X → Y → Z → X. Systematic deletions of mono- and dinucleotides after each trinucleotide are annotated as $delRNA_{3-1}$ and $delRNA_{3-2}$. Systematic deletions can start at the 5′ extremity of a sequence, which is indicated by $delRNA_{3-1.0}$ and $delRNA_{3-2.0}$, deletion frames can be shifted by 0–2 and 0–3 nucleotides for $delRNA_{3-1}$ and $delRNA_{3-2}$, respectively, which can be indicated by corresponding indices.

However, MS/MS matching between observed and predicted mass spectra is biased. Hence the unconventional natures of non-canonical transcriptions and translations presumably producing these peptides require careful evaluation of proteomic analyses. Below I review the different types of non-canonical transcriptions and translations, and relevant previous results. Previous conclusions about non-canonical peptides are then re-evaluated according to analyses that account for overfitting that could have affected the previously used proteomic search algorithm [22,23]. These new analyses strengthen previous conclusions on non-canonical mitochondrial transcriptions and translations.

1.1. Non-Canonical Transcriptions: RNA–DNA Differences

Transcription is not always perfectly accurate, but in some cases, RNA–DNA differences (RDDs) are not random and are systematically detected at some specific positions, either in the form of nucleotide substitutions [24], also observed on mitochondrion-encoded RNAs [25–27]) or deletions [28]. These punctual differences between transcript and DNA occur shortly after transcripts exit the RNA polymerase, suggesting posttranscriptional RNA editing [29]. These single nucleotide modifications produce non-canonical transcripts.

In other types of non-canonical RNAs, modifications occur systematically for all nucleotides for the complete (or almost complete) RNA. Two types of systematic transformations occur: (a) systematic deletions of mono- and dinucleotides after each trinucleotide, producing delRNAs in human mitochondria [22]; and 23 types of systematic exchanges between nucleotides (nine symmetric exchanges, type X ↔ Y, e.g. A ↔ C; and fourteen asymmetric exchanges, type X → Y → Z → X, e.g. A → C → G → A, [30–33]), producing swinger RNAs.

Swinger- and delRNAs are probably not due to RNA edition, unlike the punctual RDDs. This is because transformations frequently occur systematically on sequences longer than 100 nucleotides. They seem produced by the same RNA polymerase as regular RNA, presumably after the RNA polymerase stabilizes in a hypothetical mode similar to that causing punctual nucleotide misinsertions [32–34]. This is also indicated by contiguity between regular and swinger sequences in the few detected chimeric RNAs, DNAs and peptides that consist of regular and swinger sequences [35,36].

1.2. Non-canonical Transcriptions: Systematic Deletions

Three independent lines of evidence suggest del-transcription, transcription systematically deleting/jumping mono- and dinucleotides after each trinucleotide. (a) Contiguous short RNA reads in the human transcriptome match the mitogenome transformed by systematic (mono- and dinucleotide) deletions after each trinucleotide [22]. (b) Peptides corresponding to these delRNAs were detected in proteomic data [22]. (c) The human mitogenome, after del-transformations, has more inverted palindromes potentially forming stem-loop hairpins than comparable randomly shuffled sequences [23]. Excess palindromes after del-transformations suggest biological roles for del-transformed sequences. Palindromes in del-transformed sequences apparently down-regulate del-transcriptions [37]. Convergences between these different evidences suggest actual biological roles for del-transformed DNA/RNA.

1.3. Non-canonical Transcriptions: Systematic Nucleotide Exchanges

Homology between some RNAs and the 'parent' DNA can only be detected when assuming systematic exchanges between nucleotides along the complete RNA length. This process is called 'swinger' transcription. Several independent evidences show that swinger polymerizations occasionally occur. (a) Swinger DNA has been detected [39–40], in addition to swinger RNA [30–33]. (b) Peptides corresponding to detected swinger RNAs also occur [23]. (c) The human mitogenome includes swinger repeats, meaning repeats of other parts of the mitogenome, at the condition one assumes a given swinger transformation. These swinger repeats are more numerous than for comparable randomized sequences [38,41]. (d) The swinger-transformed human mitogenome has more inverted repeats potentially forming stem-loop hairpins than randomly shuffled sequences [42]. (e) Chimeric DNA/RNA [35] and peptides [36] exist. These nucleotide and peptide sequences consist of at least two contiguous parts, where one part is 'regular' (= untransformed), the other is swinger-transformed according to one of the 23 potential bijective transformations of nucleotide sequences [19,20].

These various material evidences converge with one another: detected swinger peptides map on detected swinger RNAs [23]; mitochondrial swinger RNA abundances increase with abundances of swinger repeats in the mitogenome; and palindromes formed by swinger-transformed mitosequences associate with swinger RNA detection [42]. This association between transcripts and hairpins for swinger RNA is expected because regular mitochondrial post-transcriptional RNA processing depends on secondary structure formation for regular RNAs, a process called tRNA punctuation of mitochondrial posttranscriptional RNA processing [43]. Hence swinger RNA processing resembles regular RNA processing.

In addition, mitochondrial swinger RNA has been detected within datasets produced by classical Sanger sequencing [30–33], and by massive next generation Illumina sequencing [23]. Swinger RNA properties converge between RNAs sequenced by these different methods [23].

1.3.1. Systematic Nucleotide Exchanges and the Natural Circular Code

A specific property of the genetic code is that it includes a 'punctuation' code which enables retrieval of the protein coding frame, called the natural circular code X [44,45], putatively by interactions between mRNAs and the ribosomal decoding center [8–10]. X consists of 20 codons that are over-represented in the coding frame of genes, as compared to non-coding frames, and as a group, have several strong mathematical properties that enable detecting the coding frame.

The 23 bijective transformations (or swinger transformations), when applied to X, produce also circular codes [34,44]. The reading frame retrieval capacity (RFR) of circular codes can be quantified [5]. The RFRs of these transformations of X correlate with properties of corresponding detected swinger RNAs [34]. This means that strictly theoretical

considerations predict swinger transcription properties. Swinger RNA abundances are proportional to the invariance of circular code properties of sequences after corresponding bijective transformations.

Associations between empirical observations of swinger transformations and theoretical properties derived from X are strong evidence that swinger transformations increased the coding potential of short protogenomes. This is because X, shared by almost all organisms [45], is very ancient. Hence swinger transformations were embedded within the polymerization machinery since its earliest inception.

1.3.2. Swinger Transformations and tRNA-Replication Origins

A peculiar observation on palindromes formed by some human mitogenome sequences after specific swinger-transformations also suggests, among others, that swinger transformed sequences are integrated in the genome, and participate in creation of new functional sequences.

Mitochondrial light strand replication typically originates at the OL, the light strand replication origin, a stem-loop hairpin located within the largest tRNA gene group in vertebrate mitogenomes [46]. The OL loop contains the recognition and initial binding site of the mitochondrial DNA polymerase [47,48]. In several taxa, such as most birds, the OL is totally missing [49], suggesting that its function is performed by adjacent tDNAs, which form OL-like structures [50–58].

No clear homology between mitochondrial tRNAs and the OL has been observed, despite functional indications suggesting some interchangeabilities between tRNA and OL functions. These include aminoacylation of RNA corresponding to the OL [59] and similarities between tRNA (and tRNA-related enzymes) and elements of the replicational machineries of ancient viruses [60,61]. Only recent analyses searching for inverted palindromes in the swinger-transformed human mitogenome detected ten nucleotide long complementarity between the human mitochondrial OL loop and the D-arm of mitochondrial tRNA Ala [42].

Eight swinger transformations which form the group 2 bijective transformations [19], create this OL-tRNA palindrome. Hence swinger transformations reveal the previously presumed OL-tRNA homology. This suggests unsuspected evolutionary implications for swinger transformations in the context of *de novo* creation of functional structural RNAs [62]. It also confirms the above considerations that swinger polymerizations occurred since the onset of the molecular machinery of life.

1.4. Peptides Matching Translation of Codons Expanded by Silent Mono- and Dinucleotides

Several observations indicate that sequences code for many more proteins than usually assumed. For example, activity of stop-suppressor (or antitermination) tRNAs [63–66] presumably templated by the antisense sequence of regular mitochondrial tRNAs [55,65,66] might enable translation of supposed non-coding frames that include stop codons [67–71]. This is also suggested by coevolution between predicted mitochondrial suppressor tRNAs and predicted mitochondrial off-frame coding regions in several taxonomic groups (primates [67,68]; *Drosophila* [68,69]; turtles [70]; and chaetognatha [71]). These analyses assume a change in genetic code where stop codons are reassigned to code for unknown amino acid(s) [72]. This stop-codon reassignment is also suggested by comparisons between mitochondrial and other genetic codes [72–81].

Translation by another type of tRNAs, tRNAs with expanded anticodons, unleashes further coding potential. This is indicated by coevolution between predicted mitochondrial tRNAs with expanded anticodons and predicted coding sequences translated from stretches of tetracodons, codons expanded by a silent fourth nucleotide [56,57,82,83].

Presumably, regular tRNA translation of delRNAs produces the same peptides as regular RNAs translated by unusual tRNAs with expanded anticodons [84–90]. Expanded codons are compatible with symmetry and error-correcting properties of the tessera, a subset of 64 among the 264 tetracodons. Tessera are the presumed ancestors of the vertebrate mitochondrial genetic code [91]. The tessera hypothesis is compatible with the fact that regular codon–anticodon interactions are too weak for peptide elongation without ribosome, which presumably evolved after primordial translation mechanisms [92–94].

Some mitochondrial peptides match translations according to tetra- and pentacodons, including for translations of swinger-transformed versions of the human mitogenome [22,23]. This type of peptide translation is particularly peculiar. For now it is deduced from (a) coevolution between predicted tRNAs with expanded anticodons with predicted tetracoding sequences, (b) empirical matches between predicted and observed MS/MS spectrometry data, and (c) associations between detected peptides and corresponding detected non-canonical swinger RNAs. Fig. 1 shows examples of translation according to tetra- and pentacodons. Hence re-analyses are designed to avoid some biases present in previous analyses. These reanalyses confirm the validity of previously described non-canonical peptides, particularly those coded by expanded codons [22,23].

1.5. Supervised versus Unsupervised Analyses

Proteomic analyses characterize protein expression patterns from mass spectrometry data of cell proteome extracts. These typically match numerous MS/MS spectra predicted from the annotated genes in genomes with observed spectra (e.g. for the bacterium *Tropheryma whipplei*, agent of Whipple's disease [95]). This approach is biased: it optimizes the fit between observed and expected MS/MS datasets. Such supervised/biased analyses always imply some false positive detections due to overfitting between observation and prediction, particularly for large datasets [96], including microarray analyses [97–99]. Deliberate biases in analyses also improve estimations [100], but overfitting remains a problem, especially for detection of unknown phenomena.

A known bias that affects classical proteomic search algorithms is that predicted protein sequences are matched to observed data, assuming specific cleavage according to cleavage by the digestion enzyme used during protein/proteome extraction and preparation. Hence if trypsin was used during sample preparation, the amino acid at the carboxyl extremity of peptides is *a priori* supposed tryptic, specific cleavage after K or R.

Hence searches are usually biased towards peptides matching the cleavage rules of digestion enzyme(s) used during sample preparation, because this limits greatly cleavage options, saves computational machine time. Detections of non-canonical peptides would be validated if unsupervised analyses that do not predefine specific cleavage rule(s) detect mainly tryptic peptides. Unbiased analyses can only recover experimental tryptic conditions if most non-canonical peptides detections are accurate.

Proteomic search algorithms usually enable analyses assuming random cleavage when fitting observed and expected mass spectra, by options indicating 'no enzyme' or 'no specific cleavage'. This option is rarely used, because it increases search times enormously.

Here analyses assume random cleavage of actually tryptic proteomes. These should preferentially detect peptides ending by K or R, as compared to other amino acids. This biased result for unbiased analyses would validate conclusions from previous trypsin-biased analyses. The latter detected numerous peptides matching non-canonical RNAs and translations [22,23]. Here I aim at confirming these previous results using unbiased analyses assuming random cleavage.

1.6. Unsupervised Analyses and Natural Proteolysis

Numerous natural proteases are active in cells, including in mitochondria, forming the mitodegradome [101]. Natural proteolysis interferes with proteomic analyses based on artificial additions of digestion enzymes [102–104]. Analyses accounting for natural proteolysis can complete proteome descriptions [105–111]. Hence unsupervised analyses assuming random cleavage might detect some actual non-tryptic peptides produced by natural proteolysis or spontaneous protein

degradation (especially during sample preparation), potentially completing descriptions of non-canonical mitochondrial peptidomes.

Analyses examining associations between non-tryptic non-canonical peptides and previously detected corresponding non-canonical RNAs [22,23] could test whether non-tryptic peptides are false positives. Positive results would validate the existence of non-canonical transcriptions and translations, independently of the expected bias for tryptic peptides among peptide populations detected by unbiased analyses.

1.7. Hypotheses and Predictions

Here unbiased proteomic analyses search tryptic human proteome data [112] for peptides matching translations of del- and swinger-transformed versions of the human mitogenome, as done by previous biased analyses that assumed tryptic digestion [22,23]. Unbiased analyses assume random protein cleavage. They are applied to the same proteomic data as previous tryptic-biased analyses that detected non-canonical peptides that match del- and swinger-transformed versions of the human mitogenome (the latter according to three codon sizes, tri-, tetra- and pentacodons). Properties of detected non-canonical peptides are compared to those detected by classical, trypsin-biased analyses.

The working hypothesis predicts that unbiased analyses detect peptide populations biased towards trypsin-digestion. This result would mean that non-canonical peptides are not false positives. Unbiased analyses could not reconstruct tryptic experimental conditions unless a majority of detected peptides were true detection. The second aspect of the working hypothesis is that detected non-tryptic non-canonical peptides result from natural proteolysis, and hence are not false positives. In that case, these should map preferentially on detected non-canonical RNAs, as previously observed for tryptic non-canonical peptides.

The primary aim is to test whether conclusions from previous results obtained by trypsin-biased analyses can be qualitatively reproduced by unbiased (unsupervised) analyses, considering potential natural proteolysis/spontaneous protein degradation, rather than experimentally added trypsin. Confirming natural proteolysis and expanding the coverage of the non-canonical mitoproteome are secondary aims. Analyses are restricted to predictions of peptides encoded by the mitogenome and its various systematic del- and swinger transformations.

2. Materials and Methods

Materials and methods are essentially identical to the corresponding sections for peptides translated from del-transformed versions of the human mitogenome [22], and those translated from the swinger-transformed versions of the human mitogenome [23]. The only difference is in the fact that the proteomic search software Proteome Discoverer 1.3 (Thermo Fisher Scientific, Illkirch) is set to analyze proteomic data digested by 'no enzyme'. The same data as previously are analyzed [112].

As for previous analyses [22,23], associations between detected non-canonical peptides and corresponding detected non-canonical RNAs are based on human transcriptomic data [113], as previously presented (del-RNAs [22], therein Tables 1 and 2; swinger-RNAs [23], therein Table 1 and supplement). For swinger-transformed versions of the human mitogenome, predicted peptides are translated according to each tri-, tetra- and pentacodons, as previously described [22,23].

All frames of transformed sequences were translated according to the vertebrate genetic code, three, four and five frames for each positive and negative strands, for codon sizes three, four and five, respectively. For codon sizes above three, codons are translated according to the genetic code, expanding the codon by silent mono- and dinucleotides, respectively. The next codon in these cases does not include the silent nucleotide(s) (see Fig. 1). The hypothetical peptides translated from non-canonical mitogenome transformations and along non-canonical codon sizes were trypsinized *in silico*, to create the fasta file containing predicted peptides.

Stop codons are translated by the letter 'X', which the software Proteome Discoverer recognizes as Leu or Ile (not distinguishable by mass spectrometry because of equal masses). Each predicted peptide including at least one stop is represented 19 times in the input database of hypothetical predicted peptides, replacing all stops by one among the 18 remaining amino acid species, excluding Leu and Ile.

Consensus searches were handled with the Sequest (Thermo Fisher Scientific, Illkirch) algorithm with molecular mass tolerances: Parent = 1 Da and Fragment = 0.5 Da (monoisotopic masses). I activated fixed carbamidomethyl (C) and variable Oxydation (M) modifications, as well as the lysine → pyrrolysine modification.

2.1. Why Include Lysine to Pyrrolysine Modifications?

An anonymous reviewer notes that pyrrolysine is not a lysine modification, but is usually encoded by UAG stop codons [114–116]. It is presumably not encoded in eukaryotes. There are several reasons for allowing lysine → pyrrolysine modifications, despite that this probably increases search times. The first reason is methodological: results of the present analyses have to be comparable to previous searches, which allowed this modification. The second reason is that non-canonical peptides might result from mechanisms that are relicts from the mitochondrion's bacterial ancestors, which probably did translate UAG by pyrrolysine.

Hence allowing this modification might enable detecting peptides that otherwise would not be detected, when stops are assumed translated by lysine. The software does not differentiate between 'regular' lysine and lysine translated by stops. Analyses presented here do not explore issues implied by modifications, but presented data include that information for future analyses.

2.2. Unbiased Analyses Can't Include Nucleus-Encoded Proteins

This anonymous reviewer also indicates that analyses should ideally include the predicted canonical human nuclear-encoded proteome, including the mitochondrial nuclear-encoded proteins imported from the cytosol. These canonical proteins would provide valuable controls for analyses designed to detect non-canonical peptides.

First, they would prevent spurious matches between observed mass spectra and predicted non-canonical peptides resembling canonical peptides. Secondly, one expects much fewer detections of non-canonical than canonical peptides, an additional prediction that can be tested. Third, such analyses would enable to test the hypothesis that higher proportions of non-canonical than canonical peptides are non-tryptic (versus tryptic ones). This hypothesis assumes directed natural proteolysis of non-canonical, hence probably dysfunctional, peptides.

The first point is handled by a different, less time-consuming analysis described in the Results section, which shows that such spurious results are unlikely. Unfortunately, the nucleus-enocoded canonical proteome can't be included in unbiased analyses. This is not only because results from unbiased analyses have to be compared to previous biased analyses that did not include the canonical nucleus-encoded proteome (reanalyses including them are planned).

Unbiased analyses including the much larger canonical nucleus-encoded proteome are technically impossible with available computing capacities. Including these canonical proteins would manifold increase numbers of predicted peptides to be matched with observed mass spectra. This would render analyses impractical, to unknown extents, as searches excluding canonical nucleus-encoded proteins last 10 days. Hence inclusion of these controls will have to wait for commercial availability of computers and software with parallel processing capacities greater than those used now (I use a machine that has 32 parallel processors, regular PCs have 2 processors). It is adequate to remind here that analyses reported here are already control analyses for previous

Table 1
Abundances of residues at carboxyl extremities of non-canonical peptides detected by unbiased analyses. Analyses assume random cleavage of tryptic human mitoproteome. Peptides are translated from the del-, swinger-transformed human mitogenome, for codons expanded by 0–2 silent nucleotides. Column 1 indicates the residue. Columns 2, 6, 10 and 14 are numbers of detected peptides with residue indicated in 1, for each analysis assuming different transcription/translation (del-, swinger-, tetra- and pentacodon); 3, 7, 11 and 15 indicate total number of that residue in corresponding translations of the mitogenome; 4, 9, 12 and 16 indicate the bias of detecting peptides with that residue in carboxyl terminus position considering the total frequency of the residue in the corresponding translation of the mitogenome; 5, 9, 13 and 17 indicate numbers of peptides mapping on corresponding detected non-canonical RNAs. The two last lines compare results when merging tryptic vs other peptides, numbers of non-canonical peptides mapping on non-canonical RNAs are followed by expected numbers assuming random mapping.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
AA	Del				Swinger				Swinger				Swinger			
	Tri	Genome	Bias	RNA	Tri	Genome	Bias	RNA	Tetra	Genome	Bias	RNA	Penta	Genome	Bias	RNA
A	10	8328	1.49	0	2	46,838	0.30	0	6	47,048	0.75	2	9	46,178	1.33	0
C	4	5680	0.87	0	0	22,697	0.00		5	22,836	1.29	0	0	22,108	0.00	0
D	1	4662	0.27	0	3	21,545	0.97	0	0	21,752	0	1	3	20,795	0.98	0
E	6	6281	1.18	0	4	24,954	1.12	0	7	25,290	1.63	0	4	24,200	1.13	0
F	5	7868	0.79	0	4	30,878	0.91	0	4	30,982	0.76	0	4	29,626	0.92	1
G	12	12,857	1.16	0	9	57,120	1.10	1	12	57,648	1.22	0	9	55,452	1.11	0
H	0	6578	0.00		4	22,775	1.23	0	7	22,836	1.80	1	5	21,803	1.56	0
IL	26	27,890	1.15	3	11	97,382	0.79	0	15	97,914	0.90	1	6	93,464	0.44	0
K	16	8142	2.43	2	10	30,603	2.29	0	14	30,982	2.65	0	14	29,608	3.22	2
M	6	7581	0.98	0	2	22,553	0.62	0	4	22,836	1.03	0	4	21,660	1.26	0
N	5	7723	0.80	0	7	26,449	1.85	0	6	26,660	1.32	0	5	25,427	1.34	1
P	9	12,857	0.87	3	6	57,489	0.73	0	7	57,648	0.71	0	7	55,452	0.86	0
Q	7	5647	1.53	2	6	24,126	1.74	0	13	24,206	3.15	0	4	23,422	1.16	0
R	9	5876	1.90	0	11	46,786	1.64	0	18	47,048	2.25	0	10	45,626	1.49	0
S	4	16,856	0.29	1	5	68,457	0.51	0	2	68,800	0.17	3	5	65,983	0.52	0
T	3	11,954	0.31	0	9	46,822	1.34	0	1	47,047	0.13	0	4	44,813	0.61	0
V	11	11,954	1.14	2	10	46,591	1.50	1	4	47,047	0.50	4	7	44,813	1.06	1
W	9	6838	1.63	1	3	23,960	0.88	0	4	24,206	0.97	0	1	23,118	0.30	0
Y	5	7630	0.81	0	0	23,238	0.00	0	2	22,836	0.51	0	4	21,364	1.28	0
Tot	148	183,202		14/6.31	106	741,263		2/3.02	127	745,622		12/3.38	105	714,912		5/2.94
Tryps	25	14,018	2.21	2/1.07	21	77,389	1.90	0/0.68	32	78,030	2.41	0/0.19	24	75,234	2.24	2/0.70
Others	123	169,184	0.90	12/5.24	85	663,874	0.90	2/2.34	95	667,592	0.84	12/3.19	81	639,678	0.85	3/2.24

results. Hence inclusion of further controls, though valuable, has always an arbitrary component, besides the above noted technical problems.

2.3. Peptide Detection Criteria

False discovery rates FDR [117–119] were estimated against a reverse decoy database using the Percolator algorithm. No protein grouping was allowed since the database only contained non-redundant entries. Peptides are considered detected with FDR $q < 0.05$ and Xcorr > 1.99. FDR is calculated by comparing Xcorr obtained from expected and observed MS/MS mass spectra with those obtained for a decoy database of false negative predicted peptides.

Xcorr is a cross-correlation statistic that compares observed and predicted MS/MS data. It sums the products between observed (y) and expected (x) values for series of data. In this case, these are the observed and expected mass spectrometry data [120]:

$$\mathrm{Xcorr} = \sum_{i=0}^{n-1} x(i) * y(i+\tau),$$

where τ is a displacement (lag) between observed and expected data for position i, with n positions in the data.

Peptide posterior error probabilities (PEP) are also indicated. PEP estimates confidence in detections of specific individual peptides. This approach differs from q, designed to estimate confidence for groups of detected peptides. The latter optimizes between false positive and false negative rates. PEP should be used with caution because it inflates false negatives [117]. Analyses focus on peptide populations and hence do not integrate PEP, but PEP is indicated because it could be useful in the context of future analyses focusing on specific peptides.

Table 2
Pearson correlation coefficient r between abundances of non-canonical peptides detected by unsupervised proteomic analyses of trypsin-digested human mitochondrial proteomic MS/MS data and abundances of corresponding, previously detected non-canonical RNAs [22,23]. Correlations are calculated separately for tryptic peptides (carboxyl extremity K or R) and other peptides. Non-canonical transcripts are del-and swinger-transformations of the human mitogenome, the latter translated along codons expanded by 0, 1 and 2 silent nucleotides. P values are one tailed, expecting positive correlations. Fisher's method for combining P values sums the $-2 \times \log$ Pi, where i runs from 1 to k. This sum follows a chi-square statistic distribution with $2 \times k$ degrees of freedoms, where k is the number of Ps combined (here $k = 4$). Bold indicates statistical significance at $P < 0.05$.

Pearson r	Unbias Tryps		Other		
Transformation	r	P	r	P	All
Del	0.358	0.172	0.270	0.241	0.401
Swinger	**0.446**	0.016	0.171	0.217	0.253
Swinger tetra	0.109	0.310	0.099	0.327	0.143
Swinger penta	0.192	0.190	0.306	0.078	0.186
Combined chi	**17.41**	0.026	13.24	0.104	

3. Results

3.1. Unsupervised Analyses Assuming Random Cleavage

Proteomic analyses of the 96 human proteome extracts [112], when no specific cleavage enzyme is specified, lasted about 10 days for each unsupervised analysis. Four such unbiased 10-day long searches were performed, for peptides matching translations of the nine del-transformations of the human mitogenome (four del-transformations deleting a mononucleotide after each trinucleotide, and five del-transformations, deleting a dinucleotide after each trinucleotide), and of the 23 swinger-transformed versions of the human mitogenome. The latter are translated according to three codon sizes: regular tricodons, tetra- and pentacodons, which are regular codons expanded by silent mono- and dinucleotides.

Comparable analyses of these data and predicted peptides, but trypsin-biased, last 8–9 h for each analysis. The longer times required

for analyses reflect much greater potential cleavage combinations when comparing observed and expected peptide mass spectra for unbiased analyses.

3.1.1. Unsupervised Analyses: Bias for Tryptic Peptides

All four unsupervised searches matching observed MS/MS mass spectra with predicted ones detected preferentially tryptic peptides (carboxyl-extremity K or R, Table 1). Hence without *a priori* biasing searches, populations of detected non-canonical peptides are for these analyses biased towards tryptic peptides.

Biases for residue identity at the carboxyl extremity of detected peptides are calculated as the frequency of observing a given amino acid at that position in detected peptides, divided by that amino acid's frequency in hypothetical peptides translated from the corresponding complete transformed human mitogenome. The highest bias favors lysine (K) for translation of delRNAs, swinger RNAs, and swinger RNAs according to pentacodons, and second highest (after bias for Q) for swinger RNAs according to tetracodons. Lysine is one among 19 possibilities, so the probability to obtain the strongest bias for K is $1/19 = 0.053$, for tetracodons the result has $P = 0.11$.

According to Fisher's method for combining independent P values [121], the overall result for bias in favor of K across all four analyses is $P = 0.0046$. Hence results show a strong bias favoring K at the end of detected peptides, despite assuming random cleavage.

Bias for arginine (R) at the carboxyl extremity of detected peptides is the second highest for peptides matching del-transformed versions of the human mitogenome (excluding K, this has $P = 1/18 = 0.056$), the fourth highest for regular translations of the 23 swinger-transformed versions of the human mitogenome ($P = 0.17$), and the third highest for their translation according to tetra- and pentacodons ($P = 0.11$, each). According to Fisher's method for combining independent P values [121], the overall result for bias in favor of K across all four analyses is $P = 0.02$.

Bias for combined K and R ending peptides is highly significant according to chi-square tests for each of the four independent analyses: delRNAs, $P = 0.00023$; swinger RNAs according to tricodons, $P = 0.0016$; tetracodons, $P = 0.00000006$; and pentacodons, $P = 0.000038$.

This means that unsupervised searches for non-canonical peptides detect preferentially non-canonical peptides that match the known tryptic sample preparation. This result could not be obtained if the majority of detections were false positives. Hence these biased results obtained from unbiased analyses confirm that the various populations of non-canonical peptides (peptides translated from delRNAs, and from swinger RNAs, these translated according to regular tri-, tetra and pentacodons) exist. This result is not trivial, and independently confirms conclusions from previous trypsin-biased analyses [22,23].

3.2. Search Bias and Absolute Versus Relative Majority of Tryptic Peptides

An anonymous reviewer notes that tryptic peptides are only a relative majority among detected peptides, rather than an absolute majority. This might to some extent contradict the above conclusion of bias towards tryptic peptides. This issue can be understood by comparing the tryptic bias obtained from unbiased analyses, to that obtained for chymotrypsin-biased analyses.

These chymotrypsin-biased analyses were for the same proteomic data and the same predicted non-canonical peptides ([122], therein supplementary data). They differ from previous analyses assuming tryptic digestion, and the current unbiased ones, because analyses assumed chymotryptic digestion. Analyses assuming chymotryptic digestion detected 479 non-canonical peptides, among which 131 (27.35%) had carboxyl terminal residues matching chymotryptic digestion (W, Y and F). All other peptides (72.65%) were tryptic. Hence the absolute majority of peptides detected by these analyses biased towards three possible non-tryptic carboxyl terminal residues are tryptic peptides, as expected.

These chymotrypsin-biased analyses were completed by three separate analyses biased to detect only peptides with a specific, chymotryptic ending, hence separately W, Y, or F. In these analyses set to detect peptides with only one possible chymotryptic residue at its carboxyl terminus, tryptic peptides represent on average across all analyses 87.3% of all detected peptides.

In order to compare these results with those from unbiased analyses presented here, I calculated the bias between tryptic peptides and peptides matching each other possible (non-tryptic) carboxyl terminus for data in Table 1. Tryptic peptides are on average 82.76% of the peptides in Table 1 when considering only one alternative residue at the carboxyl terminus. This average value is very comparable to the above mentioned 87.3% tryptic peptide majority obtained for chymotrypsin-biased analyses searching for only one of the three chymotryptic carboxyl termini.

The meaning of this is mathematically trivial. Tryptic bias decreases the more other options are allowed. It is highest when analyses are biased towards tryptic peptides: all detected peptides were tryptic. This bias decreases when analyses are biased towards a single different possibility (separating W, Y and F). Tryptic bias further decreases when all three chymotryptic carboxyl termini are considered (W, Y and F). Tryptic bias is lowest, yet still statistically significant, when analyses are unbiased, as when considering all results in Table 1. When averaging abundances of non-tryptic carboxyl termini in Table 1, the tryptic bias is comparable to that obtained for analyses biased towards only one non-tryptic carboxyl terminus, as obtained for analyses biased separately towards W, Y or F.

The issue of relative versus absolute tryptic majority is only a matter of doing adequate comparisons. I agree that adding canonical nucleus-encoded proteins in the analyses would probably yield valuable further insights in this context, regarding potential biases for tryptic canonical peptides, versus more non-tryptic natural digestions for non-canonical peptides. Currently such analyses are technically impossible in the context of unbiased analyses.

3.3. Associations between Non-canonical RNA and Peptide Abundances

Peptides with non-tryptic carboxyl extremity could represent false positive detections. They might alternatively result from natural proteolysis/spontaneous degradation of proteins. The bias for tryptic peptides (previous section) corresponds to experimental trypsin-preparation of extracts. Remaining non-tryptic peptides are not necessarily false positives.

This can be tested by exploring associations between non-canonical peptides and corresponding non-canonical RNAs, as previously described for trypsin-biased analyses [22,23]. Two independent methods are used in this respect: (a) Pearson correlation analyses between abundances of detected non-canonical peptides and corresponding RNAs, expecting positive correlations between abundances; (b) precise mapping of individual peptides and RNAs, which expects that more detected peptides map on detected RNAs than expected by chance. These positive associations between non-canonical peptides and RNAs would show the regular causal link between RNA and peptides, for the various non-canonical transcriptions and translations.

Numbers of non-canonical peptides are counted from lists of peptides in the supplementary data, mitogenome coverages by non-canonical RNA for delRNAs and for swinger RNAs are from previous publications (delRNAs, [22], therein Tables 1 and 2, for systematic mono- and dinucleotide deletions, respectively; swinger RNAs [23], therein Table 1). These non-canonical RNAs had been detected from human transcriptome data [113], using blastn with Megablast default alignment parameters [123].

For delRNAs, correlation analyses of associations between abundances depend on nine observations, based on coverages of the del-transformed human mitogenome. There are four and five observations, according to the four frames of systematic mononucleotide deletions,

and the five frames of systematic dinucleotide deletions, respectively (data in [22], therein Tables 1 and 2). Correlation analyses for swinger transformations are based on 23 observations, for each swinger transformation of the human mitogenome (RNA coverage data from [23], therein Table 1).

For tryptic peptides detected by unbiased analyses, abundances of detected swinger peptides coded by regular codons are proportional to corresponding swinger RNA coverage of the human mitogenome ($r = 0.45$, $P = 0.016$, one tailed test). Correlations are positive but not statistically significant at $P < 0.05$ for swinger analyses of other codon sizes, and del-transformations. Combining the four P values using Fisher's method for combining P values [121] yields an overall significant positive association ($P = 0.026$). These results from unbiased analyses confirm previous trypsin-biased analyses [22,23].

Correlation analyses for non-tryptic peptides detected by unbiased analyses are also positive, though never statistically significant at $P < 0.05$, also not after combining P values ($P = 0.104$). These weaker associations between peptide abundances and RNA coverage for non-tryptic peptides suggest that a greater proportion of these peptides could be false positive detections, though the overall positive trends are rather compatible with them resulting from natural proteolysis. Local mappings on RNAs below test this point.

3.4. Local Mapping of Non-canonical RNA and Peptides

Some peptides detected by unsupervised proteomic analyses map on previously detected, corresponding non-canonical RNAs (Table 1). Previous similar analyses for non-canonical peptides detected by trypsin-biased searches showed that detected peptides map more frequently than expected by chance on corresponding detected non-canonical RNAs, for del- and swinger-transformations [22,23].

Analyses across unsupervised analyses for del- and swinger-transformations find that 4 among 102 detected non-canonical tryptic peptides (3.9%) map on previously detected non-canonical RNAs. Only 2.64 among these 102 detected tryptic peptides should map on detected RNAs if mapping is random. Small sample size does not enable statistical testing, but suggests a non-significant difference corresponding to the expected association between non-canonical RNAs and peptides. Though this result is not statistically significant, it should be considered as confirmative as it is in line with previous, statistically significant results for tryptic peptides detected by analyses biased towards tryptic peptides.

For non-tryptic peptides detected by unsupervised analyses, 29 among 388 (7.5%) map on previously detected RNAs. This is statistically significantly more than the 13.01 expected according to random mapping (chi-square test, $P = 0.000009$). Rates of mapping on RNAs do not differ between tryptic and non-tryptic peptides according to a chi-square test. Hence overall, there are not more false positive detections of non-tryptic than tryptic peptides.

In total, 33 among the 490 non-canonical peptides detected by unbiased analyses map on non-canonical RNAs (6.7%), which is statistically significantly more than the 15.65 expected by chance (chi-square test, $P = 0.000012$).

These results for non-tryptic peptides detected by unbiased analyses confirm conclusions about non-canonical transcriptions and translations, independently of previous results for tryptic peptides detected by trypsin-biased proteomic analyses. Notably, results from independent analyses strengthen conclusions that swinger RNAs are also translated according to tetra- and pentacodons.

3.5. Unique Versus Multiple Detections of Tryptic Peptides

Unbiased analyses confirm previous trypsin-biased analyses in two ways. First, they detect preferentially tryptic peptides. This corresponds to the tryptic experimental design. Second, detected peptides associate with previously detected RNAs, for tryptic and other peptides, as found in previous publications for tryptic peptides detected by trypsin-biased analyses [22,23].

These results independently confirm trypsin-biased analyses because most tryptic peptides detected by trypsin-biased analyses differ from those detected by the unbiased analyses presented here. Only some tryptic peptides detected by trypsin-biased analyses are also detected by unbiased analyses (one or two peptides). Hence analyses confirm independently of previous analyses, the previous results for trypsin-biased analyses.

Note that analyses of the same data, testing the same hypotheses, and assuming chymotryptic digestion, yield similar conclusions. These analyses detect majorities of tryptic peptides. Both tryptic and chymotryptic peptides associate with detected RNAs [122]. Hence unbiased analyses yield a third independent confirmation of results obtained by tryptic- and chymotryptic-biased analyses.

3.6. Negative Control: Residues after the Carboxyl-Terminal of Detected Peptides

A further analysis shows a peculiar unknown fact about natural mitochondrial proteolysis. Unbiased analyses yield peptide populations with diverse residues at their carboxyl extremity, which might mainly reflect proteolysis by naturally occurring digestion enzymes in human mitochondria. The alternative hypothesis (or rather the null hypothesis) is that peptides were actually randomly cleaved, which would also be compatible with random peptide detections, possibly due to majorities of false detections.

Biases for tryptic peptides overall falsify the random cleavage hypothesis, but remaining peptide populations, after excluding tryptic peptides, might nevertheless fit random cleavage. Here control analyses test for this by recording the first amino acid expected according to non-canonical translations, after the carboxyl extremity of the detected peptides. According to the compilation by ExPaSy PeptideCutter (http://web.expasy.org/peptide_cutter/peptidecutter_enzymes.html, accessed 6VI2016), the majority of listed specific cleavage rules relate to the carboxyl extremity of peptides, rather than the N-terminal of the next peptide. Nevertheless, the possibility that populations of detected peptides include biases for N-terminal cleavage in relation to the 'downstream'-encoded amino acid is plausible.

Table 3 presents biases, calculated as for Table 1 by using total abundances of amino acids, for amino acids at the N-terminal of the peptide located after the detected peptides, for the various unsupervised analyses (peptides translated from del- and swinger-transformed human mitogenome, and translated according to tetra- and pentacodons for the swinger-transformed versions).

These biases do not resemble biases detected for the carboxyl extremity of the detected peptides, when considering the same amino acid species. Bias distributions are systematically less extreme for N-terminals of the next undetected peptide than for the carboxyl extremity of detected peptides. For the N-terminal of the next peptide, the lowest bias is 0.44, 0.40, 0.33 and 0.42 for detected peptides translated from del-, swinger-transformed versions of the human mitogenome, and for swinger-transformed mitogenome translations according to tetra- and pentacodons. Such biases below "1" indicate cleavage avoidance. For the carboxyl terminal of detected peptides, corresponding minimal biases are 0 for each non-canonical translation, the strongest possible negative bias.

Maximal biases for N-terminals of undetected peptides next to detected peptides are 1.59, 1.98, 1.64 and 2.31. For carboxyl extremities of detected peptides, corresponding maximal biases are 2.43, 2.29, 3.15 and 3.22. Overall, distributions for biases of amino acid identities at the N-terminal of next peptides are much closer to the value '1', indicating no bias, and seem random around this value. This suggests that there is no evidence for N-terminal specific cleavage in these data for the human mitochondrial proteome.

Table 3
Bias in amino acid identity at the N-terminal (column 1) of the peptide after detected peptides, for unbiased analyses assuming random cleavage. Analysis search for peptides matching translations of the del- (columns 2–4) and swinger-transformed human mitogenome (columns 5–7), and translations of the swinger mitogenomes according to tetra- (columns 8–10) and pentacodons (columns 11–13). Columns 2, 5 , 8 and 11 incicate numbers of detections. 'Genome' (columns 3, 6, 9, 12) indicates abundances of that residue in the corresponding hypothetical translations of the complete mitogenome after transformations and non-canonical translations. Biases (columns 4, 7, 10, 13) do not resemble those for carboxyl-extremities of detected peptides (Table 1) and are less extreme. Overall they match random distributions around '1', indicating lack of bias. This suggests that there is no or very little natural proteolysis with cleavage specificity related to the N-terminal of peptides after detected peptides.

1	2	3	4	5	6	7	8	9	10	11	12	13
AA	Del			Swinger			Swinger			Swinger		
	Tri	Genome	Bias	Tri	Genome	Bias	Tetra	Genome	Bias	Penta	Genome	Bias
A	8	8328	1.19	6	46,838	0.79	10	47,048	1.09	3	46,178	0.42
C	4	5680	0.87	3	22,697	0.82	6	22,836	1.34	8	22,108	2.31
D	6	4662	1.59	3	21,545	0.86	7	21,752	1.64	5	20,795	1.54
E	6	6281	1.18	8	24,954	1.98	7	25,290	1.41	2	24,200	0.53
F	5	7868	0.79	2	30,878	0.40	5	30,982	0.82	2	29,626	0.43
G	13	12,857	1.25	8	57,120	0.87	8	57,648	0.71	4	55,452	0.46
H	6	6578	1.13	4	22,775	1.09	4	22,836	0.90	2	21,803	0.59
IL	20	27,890	0.89	20	97,382	1.27	25	97,914	1.30	18	93,464	1.23
K	5	8142	0.76	4	30,603	0.81	5	30,982	0.82	4	29,608	0.86
M	5	7581	0.81	6	22,553	1.64	5	22,836	1.12	4	21,660	1.18
N	9	7723	1.44	2	26,449	0.47	6	26,660	1.15	3	25,427	0.75
P	11	12,857	1.06	5	57,489	0.54	12	57,648	1.06	11	55,452	1.27
Q	2	5647	0.44	4	24,126	1.24	4	24,206	0.84	4	23,422	1.09
R	3	5876	0.61	8	46,786	1.06	3	47,048	0.33	13	45,626	1.82
S	12	16,856	0.88	8	68,457	0.72	19	68,800	1.41	11	65,983	1.06
T	11	11,954	1.14	12	46,822	1.58	9	47,047	0.98	8	44,813	1.14
V	9	11,954	0.93	13	46,591	1.72	5	47,047	0.54	6	44,813	0.96
W	7	6838	1.27	2	23,960	0.52	2	24,206	0.42	2	23,118	0.55
Y	6	7630	0.97	2	23,238	0.53	4	22,836	0.90	2	21,364	0.60
Tot	148	183,202		120	741,263		146	745,622		112	714,912	

These analyses show non-random patterns in cleavages for detected non-canonical peptide populations, for carboxyl termini. In this respect, results for N-termini function as negative controls and strengthen confidence in results.

3.7. Few Nuclear Contaminations: Peptides Follow the Mitochondrial Vertebrate Code

Eukaryotic nuclear genomes include numerous inserts of the mitogenome. Hence detected non-canonical peptides could originate from non-canonical transcriptions and translations of such nuclear mitogenome inserts, or from translations of nuclear sequences that by chance resemble the transformed mitogenome. This possibility is tested by translating the transformed mitogenome using the nuclear genetic code, and by checking whether detected non-canonical peptides are compatible with translation according to the nuclear genetic code.

Considering that coding assignments of 60 among 64 codons (93.75%) are identical for the nuclear and the vertebrate mitochondrial genetic codes, I calculated numbers of peptides, considering lengths of detected peptides, expected to match also nuclear genetic code translation. I used equation $N \times 0.9375^{-k}$, where N is the number of detected peptides with k residues. This equation expresses the fact that all codons coding for the peptide must be among those invariant between the two genetic codes, when one or more codons belong to the four codons differing between these two genetic codes, the detected peptide is incompatible with the nuclear genetic code.

There are 177.86 peptides expected compatible with both codes across all analyses. This is far more than the 91 detected non-canonical peptides with translations identical according to both genetic codes. Comparisons between expected and observed peptides compatible with translation according to the nuclear genetic code, separately according for the four different non-canonical transcriptions and translations, follow the same principle: observed numbers of peptides compatible also with the nuclear genetic code are far fewer than expected (Table 4).

This bias means that observed non-canonical peptides match specifically more than expected by chance translation according to the mitochondrial vertebrate genetic code. This result also excludes that detections of non-canonical peptides are incorrect, that these mass spectra actually correspond to similar, nucleus-encoded canonical peptides. This is because the analysis reported in this section accounts for the extreme and plausible situation where sequences identical to the mitogenome were translated. The fact that analyses differentiate between nuclear versus mitochondrial translations of the mitogenome is incompatible with nuclear contaminations.

4. General Discussion

Analyses presented here are mainly designed to test conclusions from previous analyses of the human mitochondrial peptidome (data from [112]), where non-canonical peptides matching translations of del- and swinger-transformed versions of the human mitogenome were detected, including translations of expanded codons. Del-transformations assume transcription that systematically deletes mono- and dinucleotides after every third transcribed nucleotide [22]. Swinger-transformed RNAs result presumably from systematic nucleotide exchanges, during transcription along 23 exchange rules, also called bijective transformations [34]. The human proteome includes peptides matching detected swinger RNA, translated according to tri-, tetra- and pentacodons (expanded by silent mono- and dinucleotides) [23].

Table 4
Observed (column 4) and expected (column 5) numbers of detected non-canonical peptides compatible with translations according to each nuclear and mitochondrial vertebrate genetic codes. Predictions account for peptide length (mean length and standard deviation in columns 2 and 3), considering that translation of 60/64 (0.9375) codons is identical between these genetic codes. Results indicate strong biases against detection of peptides compatible with both genetic codes, showing that detected populations of peptides are specifically translated according to the mitochondrial vertebrate genetic code. This systematic bias excludes that detected non-canonical peptides have cytosolic origins.

1	2	3	4	5
Transformation	AAs	Sd	Obs	Exp
Del	18.28	5.96	23	48.75
Swinger tri	21.23	9.75	27	36.26
Swinger tetra	17.37	7.38	19	52.45
Swinger penta	17.37	7.79	23	40.40

These previous analyses assumed tryptic proteome preparation [112]. Hence the first set of analyses was biased by information corresponding to sample preparation. Here analyses of the same data were repeated without using that information on tryptic-digestion, but assuming random cleavage. Results indicate a positive bias towards detection of tryptic non-canonical peptides by unsupervised analyses. This result is a strong confirmation that overall, populations of detected non-canonical peptides are not false positives: otherwise, unbiased analyses would not detect positive bias for tryptic peptides. This implies that these non-canonical transcriptions and translations are a biological reality.

Results of unbiased analyses also suggest the possibility that the proteome underwent other specific cleavages, presumably resulting from natural proteolytic activity in the biological sample, such as described for chymotrypsin [116]. Overall, tryptic and non-tryptic non-canonical peptides associate with previously detected corresponding non-canonical RNAs [22,23,116]. Convergences between peptide and RNA detections are further evidence that overall, tryptic and other detected peptides are not false positives.

In addition, detected non-canonical peptides preferentially match translation according to the vertebrate mitochondrial genetic code: fewer than expected by chance are compatible with translation according to the nuclear genetic code, considering that 93.75% of codons follow the same translation rules according to both genetic codes. This result is incompatible with detection of peptides originating from the cytosol, even for nuclear DNA sequences identical to the mitogenome.

4.1. Statistical Considerations and Peptide Detection

One can argue that non-canonical peptides were detected by chance and are false positives, due to a very large number of comparisons between predicted peptides and observed mass spectra. If it was so, (a) peptide detections would not be biased towards independently detected RNAs, (b) towards translation specific to the vertebrate mitochondrial genetic code, and (c) peptide populations detected by unsupervised analyses would not be biased towards experimental tryptic cleavage. In addition, peptide detections are confirmed by false detection rates q, based on decoy peptides that function as negative controls. FDR takes into account sample sizes (as do usual P values), but also the number of statistical tests done.

The last point in this argument is because analyses account that at stops, every possible amino acid could be inserted. Hence matching observed and expected peptides based on their molecular weight is not sufficient to ascertain the sequence of the peptide: the program can adjust any MS/MS spectrum with a close weight to one of the 19 peptides produced by sequences including at least one stop.

This point does not consider that MS/MS spectrometry accounts not only for total mass, but also for masses of secondary fragments. The simple example of peptide EFG can be helpful here. EFG has the same molecular weight as peptides EGF, FEG, FGE, GEF and GFE and can't be differentiated from these five other peptides by its total mass. However, the estimate of that mass is typically combined with estimates of secondary fragments. Only two peptides, EFG and GEF are compatible with detection of the mass of EF. The same point is valid for observing a mass corresponding to FG, which is also compatible with two peptides, EFG and FGE. The combined observation of masses corresponding to these two fragments characterizes the entire peptide sequence.

In addition, mass spectrometry analyses consider separately b and y ions. Hence the same sequence characterization may occur independently according to both ion types. The score Xcorr integrates these pieces of information, and is the statistic on base of which FDR is calculated to minimize false positives. In fact, numerous tryptic peptides were not detected by original analyses assuming that tryptic digestion was detected by analyses assuming non-tryptic digestion. In addition, previous analyses showed that tryptic peptides detected twice, by analyses assuming tryptic and chymotryptic digestions do not differ in detection accuracy from those detected only by one of these analyses [116]. This suggests that the methodology used for peptide detections is rather prone to false negatives, rather than false positives. False positives are probably a small minority reduced to few individual cases that would not qualitatively alter conclusions.

4.2. Potential Confounding Factors: Nuclear Contaminations

The first detected swinger-transformed sequences are RNA and DNA sequences in Genbank's databases (EST data for RNA) longer than 100 nucleotides. These were detected by blast using default megablast alignment search parameters for input sequences consisting of *in silico* swinger transformed mitogenome versions [30–33]. The detected GenBank sequences aligning with high identity levels with *in silico* produced swinger mitogenome versions (>90% identity) were sequenced by the classical Sanger technology. Similar searches in Genbank's human transcriptome SRA (sequence read archives) data produced by RNA seq (Illumina) next generation sequencing technology using blastn (also with default search parameters) confirmed the relative abundances of swinger RNAs [23].

Further megablast analyses could not detect alignments between any human nuclear chromosome sequence and del-, swinger-transformed mitogenome versions. However, blastn analyses detected such alignments that could potentially confound several alignments between the transformed mitogenome and RNA seq data. Nevertheless, the majority of RNA seq alignments are due to RNAs originating from the mitochondrion, and are not nuclear, because identities between the transformed mitogenome versions and RNA seq sequences are greater than with corresponding nuclear chromosome sequences, and this for each del- and swinger-transformed mitogenomes [22,23].

Note that nuclear chromosome sequence alignments with the del- and swinger-transformed mitogenome imply that besides regular mitochondrial mitogenome inserts in nuclear chromosomes (numts, [124–141]), transformed versions of the mitogenome (or part of) occur in the nuclear genome. Alternatively, regular numts are transcribed according to del and swinger non-canonical systematic transformations. At this point, the main issue is the existence of polymerizations producing systematic transformations, independently of cell compartment where these occur, or whether produced by replication, reverse transcription or transcription. Hence answering with certitude to these questions beyond explained above, though important, is secondary at this point.

In addition, nuclear contaminations are at most minor for peptides presented here, because detected non-canonical peptides are less frequently compatible with both nuclear and vertebrate mitochondrial genetic codes than expected by chance. This bias suggests high specificity for mitochondrial origin of detected peptides.

4.3. Potential Confounding Factors: Heteroplasmy

Heteroplasmy [142–144] is a further known phenomenon that could explain results. However, single nucleotide substitutions can't explain observations of long, non-canonical peptides. Hence only length heteroplasmies, especially those resulting from insertions, could by chance explain non-canonical peptides predicted from systematic mitogenome transformations.

However, the most common length heteroplasmies are relatively few and mainly located in the mitochondrial control region [145], while the peptides detected for the various non-canonical transcriptions and translations are distributed all around the mitogenome. This excludes length heteroplasmies as a major confounding factor for detections of non-canonical peptides.

4.4. Potential Confounding Factors: Fused Transcripts

Some transcripts result from fusion of RNA transcribed from DNA regions that are not contiguous [146,147]. This can result from reverse-transcription artifacts during cDNA production [148]. Fused swinger RNAs also exist [35]. Fusions of regular RNAs are unlikely to produce RNAs that would mimick products of systematically transforming transcriptions. Hence only few single detected non-canonical peptides could by chance correspond to RNA fusions. Artificial transcript fusions during cDNA production could not have produced detected peptides.

4.5. Natural Proteolysis of Canonical Versus Non-canonical Peptides

An anonymous reviewer suggested that proteomic analyses should include classical, canonical proteins. This would enable comparing results between canonical and non-canonical peptides, expecting fewer non-canonical peptides than canonical ones. In addition, the reviewer expected that non-canonical peptides would more frequently match non-tryptic, hence natural proteolysis, than canonical peptides. The rationale behind this prediction is that one could expect that non-canonical products are preferentially digested as waste than products of canonical genes.

Practical reasons prevented me from performing these tests. These additional analyses require including among predicted peptides the complete human proteome (corresponding to more than 20,000 genes). This increases numbers of predicted peptides to extents that, for unbiased analyses, are incompatible with current computing capacities. For these reasons, previous and current analyses have been restricted to peptides encoded by the human mitogenome, excluding nucleus-encoded mitochondrial proteins, which are imported from the cytosol into the mitochondrion [149–152].

A possible solution to this technical problem is to sample the canonical proteome. Analyses searching for peptides matching the swinger-transformed versions of the human mitogenome, translated according to regular tricodons, included such a control. These analyses included peptides predicted according to the regular translation of the untransformed human mitogenome, with the canonical mitochondrion-encoded genes. Fifteen among the detected peptides correspond to translation of the untransformed human mitogenome, among which a single tryptic peptide (6.7% of detected peptides encoded by the untransformed mitogenome). However, 20.9% of the remaining non-canonical peptides are tryptic.

This difference is not compatible with the hypothesis that natural proteolysis digests preferentially non-canonical peptides. However, this qualitative result is not statistically significant, due to small sample size. In addition, the fifteen peptides translated from the regular mitogenome are not restricted to canonical translation of the 13 proteins encoded by the human mitogenome. They include translations of other frames of these genes, and of other sequences (e.g. rRNAs etc). This hypothesis requires analyses specifically designed to test its predictions, which are beyond the frame of present analyses.

4.6. Amino Acids Inserted at Stops

A further useful comment by a reviewer suggested to investigate which amino acids are detected inserted at stops. Table 5 shows the distribution of amino acids inserted at stops for the various types of investigated non-canonical peptides, those translated from delRNAs, swinger RNAs, and from the latter, translated according to tetra- and pentacodons. These distributions overall resemble each other, hence biases for insertion of specific amino acid species at stops are explored for the sum of amino acids across all types of non-canonical peptides.

This distribution is compared to the distribution of amino acids in the 13 canonical, mitogenome-encoded proteins. Chi-square tests detect statistically significant positive biases for five amino acids, in decreasing order of bias: K, Q, C, D and E. The two first amino acids are identical to regular amino acids found most frequently inserted at stops by Aerni et al. [153]. This is a further indication that results presented here are not due to random false detections. In addition, this suggests that the mitochondrial system for translating stops resembles that found in bacteria, at least that from *Escherichia coli*.

Table 5
Distributions of amino acids inserted at stops in detected non-canonical peptides (columns Del, Swinger tri, Swinger tetra and Swinger penta), compared to the distribution of amino acids in canonical proteins encoded by the human mitogenome (Mito). Bias is the ratio between the frequency of the amino acid across all non-canonical peptides (column All) and its frequency in canonical proteins. P values are calculated using a chi-square test. Statistically significant results at $P < 0.05$ are underlined, and in bold when these are positive biases indicating greater than expected insertions at stop codons.

AA	Mito	Del	Swinger tri	Swinger tetra	Swinger penta	All peptides	Bias	P
A	225	1	10	7	1	19	0.65	0.062
C	22	3	3	3	0	9	3.15	**0.002**
D	66	2	5	10	0	17	1.98	**0.010**
E	88	5	9	4	2	20	1.75	**0.020**
F	216	3	2	4	1	10	0.36	0.0006
G	212	16	9	12	1	38	1.38	0.058
H	97	0	4	4	0	8	0.64	0.208
I,L	963	19	6	19	46	90	0.72	0.0006
K	95	18	22	22	11	73	5.92	**4×10^{-40}**
M	208	8	9	6	5	28	1.04	0.853
N	164	8	7	5	3	23	1.08	0.723
P	219	5	3	9	5	22	0.77	0.237
Q	90	18	7	8	10	43	3.68	**2×10^{-14}**
R	63	4	2	2	3	11	1.35	0.359
S	274	9	5	10	10	34	0.96	0.796
T	351	10	8	6	5	29	0.64	0.013
V	167	5	5	5	2	17	0.78	0.328
W	104	2	2	3	3	10	0.74	0.356
Y	135	2	1	7	4	14	0.80	0.414

4.7. Associations Between Independent Transcriptomic and Proteomic Data

A further important point raised by an anonymous reviewer relates to the origins of transcriptomic data, which are from patients with myeloid leukemia, versus the origins of proteomic data, which are from healthy patients. I previously discussed this issue for analyses of these data [26] along the following lines.

It is clear that if RNA and peptide data were obtained from the same cells, associations between RNA and peptide data would be strongest. The strength of the association would decrease if RNA and peptide were from the same tissues of the same individual(s), but not the same cells. Along that rationale, they would further decrease if RNA and peptide data were obtained from different individuals with similar backgrounds (e.g. all healthy).

Current analyses were done on data that were available to this author, in formats readily analyzable by available software, and for adequate quantities of data. The RNA and peptide data differ in cells, tissues, individuals and backgrounds. This means that statistically significant associations were repeatedly detected between RNA and peptide data despite a number of confounding factors that could mask RNA-peptide associations. The fact that associations between non-canonical RNAs and peptides were nevertheless repeatedly detected implies that the actual phenomenon is much stronger than evaluated in these suboptimal conditions.

A noisy background is more likely to mask than create statistically significant signals. In addition, noise would only occasionally create spurious associations, but associations were repeatedly detected. In fact, discrepancies between RNA and peptide origins explain why relatively few detected peptides map on detected RNAs. Nevertheless, these discrepancies could not prevent detecting associations between non-canonical RNAs and corresponding peptides.

5. Conclusions

- Unbiased analyses assuming random cleavage for tryptic data yield results biased towards tryptic peptides for peptides translated from non-canonical RNAs and along non-canonical translations. Results confirm previous trypsin-biased analyses that detected non-canonical peptides.
- Detected non-canonical RNAs associate with tryptic and non-tryptic peptides.
- Detected non-canonical peptides are overwhelmingly incompatible with translation according to the nuclear genetic code, and specifically match the mitochondrial vertebrate genetic code.
- Overall, results confirm translation of non-canonical RNAs (del- and swinger RNAs), and along expanded codons, in addition to detections of other types of non-canonical peptides, such as peptides translated from contiguous regular and swinger-transformed RNA [36].
- Proteomic analyses assuming random cleavage detect non-canonical peptides digested by natural proteolysis, expand proteomic coverage.

Acknowledgments

This work has been carried out thanks to the support of the A*MIDEX project (no ANR-11-IDEX-0001-02) funded by the "Investissements d'Avenir" French Government program, managed by the French National Research Agency (ANR).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.csbj.2016.09.004.

References

[1] Popov O, Segal DM, Trifonov EN. Linguistic complexity of protein sequences as compared to texts of human languages. Biosystems 1996;38:65–74.
[2] AbouHaidar MG, Venkataraman S, Golshani A, Liu B, Ahmad T. Novel coding, translation, and gene expression of a replicating covalently closed circular RNA of 220 nt. Proc Natl Acad Sci U S A 2014;111:14542–7.
[3] Arquès DG, Michel CJ. A complementary circular code in the protein coding genes. J Theor Biol 1996;182:45–58.
[4] Ahmed A, Frey G, Michel CJ. Frameshift signals in genes associated with the circular code. In Silico Biol 2007;7:155–68.
[5] Ahmed A, Frey G, Michel CJ. Essential molecular functions associated with the circular code evolution. J Theor Biol 2010;264:613–22.
[6] Michel CJ. Circular code motifs in transfer and 16S ribosomal RNAs: a possible translation code in genes. Comput Biol Chem 2012;37:24–37.
[7] Michel CJ. Circular code motifs in transfer RNAs. Comput Biol Chem 2013;45:17–29.
[8] El Soufi K, Michel CJ. Circular code motifs in the ribosome decoding center. Comput Biol Chem 2014;52:9–17.
[9] El Soufi K, Michel CJ. Circular code motifs near the ribosome decoding center. Comput Biol Chem 2015;59a:158–76.
[10] El Soufi K, Michel CJ. Circular code motifs in genomes of eukaryotes. J Theor Biol 2016;408:198–212.
[11] Itzovitz S, Alon U. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. Genome Res 2007;17:405–12.
[12] Seligmann H, Pollock DD. The ambush hypothesis: hidden stop codons prevent off-frame gene reading. Midsouth Computational Biology and Bioinformatics Society; 2003[Abstract 36].
[13] Seligmann H, Pollock DD. The ambush hypothesis: hidden stops prevent off-frame gene reading. DNA Cell Biol 2004;23:701–5.
[14] Seligmann H. Cost minimization of ribosomal frameshifts. J Theor Biol 2007;249:162–7.
[15] Seligmann H. The ambush hypothesis at the whole organism level: off frame, 'hidden' stops in vertebrate mitochondrial genes increase developmental stability. Comput Biol Chem 2010;34:80–5.
[16] Seligmann H. Coding constraints modulate chemically spontaneous mutational replication gradients in mitochondrial genomes. Curr Genomics 2012;13:37–54.
[17] Jestin J-L, Soulé C. Symmetries by base substitutions in the genetic code predict 2(′) or 3(′) aminoacylation of tRNAs. J Theor Biol 2007;247:391–4.
[18] Jestin J-L. A rationale for the symmetries by base substitutions of degeneracy in the genetic code. Biosystems 2010;99:1–5.
[19] Fimmel E, Giannerini S, Gonzalez DL, Strüngmann L. Dinucleotide circular codes and bijective transformations. J Theor Biol 2015;386:159–65.
[20] Gumbel M, Fimmel E, Danielli A, Strüngmann L. On models of the genetic code generated by binary dichotomic algorithms. Biosystems 2015;128:9–18.
[21] Kozyrev SV, Khrnnikov AY. 2-Adic numbers in genetics and Rumer's symmetry. Dokl Math 2010;81:128–30.
[22] Seligmann H. Codon expansion and systematic transcriptional deletions produce tetra-, pentacoded mitochondrial peptides. J Theor Biol 2015;387:154–65.
[23] Seligmann H. Translation of mitochondrial swinger RNAs according to tri-, tetra- and pentacodons. Biosystems 2016;140:38–48.
[24] Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG. Widespread RNA and DNA sequence differences in the human transcriptome. Science 2011;333:53–8.
[25] Bar-Yaacov D, Levin AG, Richards AL, Hachen N, Rebolledo Jaramillo B, Nekrutenko A, Zarivach R, Mishmar D. RNA–DNA differences in human mitochondria restore ancestral form of 16S ribosomal RNA. Genome Res 2013;23:1789–96.
[26] Hodgkinson A, Idaghdour Y, Gbeha E, Grenier JC, Hip-Ki E, Bruat V, Goulet JP, de Malliard T, Awadalla P. High resolution genomic analysis of human mitochondrial RNA sequence variation. Science 2014;344:413–5.
[27] Moreira S, Valachi M, Aoulad-Aissa M, Otto C, Burger G. Novel modes of RNA editing in mitochondria. Nucleic Acids Res 2016;44:4907–19.
[28] Chen C, Bundschuh R. Systematic investigation of insertional and deletional RNA–DNA differences in the human transcriptome. BMC Genomics 2012;13:616.
[29] Wang IX, Core LJ, Kwak H, Brady L, Bruzel A, McDaniel L, Richards AL, Wu M, Grunseich C, Lis JT, Cheung VG. RNA–DNA differences are generated in human cells within seconds after RNA exits polymerase II. Cell Rep 2014;6:906–15.
[30] Seligmann H. Overlapping genes coded in the 3′-to-5′-direction in mitochondrial genes and 3′-to-5′ polymerization of non-complementary RNA by an 'invertase'. J Theor Biol 2012;318:38–52.
[31] Seligmann H. Polymerization of non-complementary RNA: systematic symmetric nucleotide exchanges mainly involving uracil produce mitochondrial RNA transcripts coding for cryptic overlapping genes. Biosystems 2013;111:156–74.
[32] Seligmann H. Systematic asymmetric nucleotide exchanges produce human mitochondrial RNAs cryptically encoding for overlapping protein coding genes. J Theor Biol 2013;324:1–20.
[33] Seligmann H. Triplex DNA:RNA, 3′-to-5′ inverted RNA and protein coding in mitochondrial genomes. J Comput Biol 2013;20:1–12.
[34] Michel CJ, Seligmann H. Bijective transformation circular codes and nucleotide exchanging RNA transcription. Biosystems 2014;118:39–50.
[35] Seligmann H. Swinger RNAs with sharp switches between regular transcription and transcription systematically exchanging ribonucleotides: case studies. Biosystems 2015;135:1–8.
[36] Seligmann H. Chimeric peptides from contiguous regular and swinger RNA. Comput Struct Biotechnol J 2016;14:283–97.
[37] Seligmann H. Systematically frameshifting by deletion of every 4th or 4th and 5th nucleotides during mitochondrial transcription: RNA self-hybridization regulates delRNA expression. Biosystems 2016;142:43–51.
[38] Seligmann H. Mitochondrial swinger replication: DNA replication systematically exchanging nucleotides and short 16S ribosomal DNA swinger inserts. Biosystems 2014;125:22–31.
[39] Seligmann H. Species radiation by DNA replication that systematically exchanges nucleotides? J Theor Biol 2014;363:216–22.
[40] Seligmann H. Sharp switches between regular and swinger mitochondrial replication: 16S rDNA systematically exchanging nucleotides A ↔ T + C ↔ G in the mitogenome of *Kamimuria wangi*. Mitochondrial DNA a DNA Mapp. Seq Anal 2015;27:2440–6.
[41] Seligmann H. Systematic exchanges between nucleotides: genomic swinger repeats and swinger transcription in human mitochondria. J Theor Biol 2015;384:70–7.
[42] Seligmann H. Swinger RNA self-hybridization and mitochondrial non-canonical swinger transcription, transcription systematically exchanging nucleotides. J Theor Biol 2016;399:84–91.
[43] Ojala D, Montoya J, Attardi G. tRNA punctuation model of RNA processing in human mitochondria. Nature 1981;290:470–4.
[44] Michel CJ, Pirillo G. Identification of all trinucleotide circular codes. Comput Biol Chem 2010;34:122–5.
[45] Michel CJ. The maximal C(3) self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses. J Theor Biol 2015;380:156–77.
[46] Clayton DA. Replication of animal mitochondrial DNA. Cell 1982;28:693–705.
[47] Hixson JE, Wong TW, Clayton DA. Both the conserved stem-loop and divergent 5′-flanking sequences are required for initiation at the human mitochondrial origin of light-strand DNA replication. J Biol Chem 1986;261:2384–90.
[48] Wanrooij S, Falkenberg M. The human mitochondrial replication fork in health and disease. Biochim Biophys Acta 1797;2010:1378–88.
[49] Desjardins P, Morais R. Nucleotide sequence and evolution of coding and noncoding regions of a quail mitochondrial genome. J Mol Evol 1991;32:153–61.
[50] Seligmann H, Krishnan NM. Mitochondrial replication origin stability and propensity of adjacent tRNA genes to form putative replication origins increase developmental stability in lizards. J Exp Zool B Mol Dev Evol 2006;306:433–49.
[51] Seligmann H, Krishnan NM, Rao BJ. Possible multiple origins of replication in primate mitochondria: alternative role of tRNA sequences. J Theor Biol 2006;241:321–32.
[52] Seligmann H, Krishnan NM, Rao BJ. Mitochondrial tRNA sequences as unusual replication origins: pathogenic implications for *Homo sapiens*. J Theor Biol 2006;243:375–85.
[53] Seligmann H. 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 2008;379:188–99.
[54] Seligmann H. 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 2010;99:85–93.
[55] Seligmann H. Pathogenic mutations in antisense mitochondrial tRNAs. J Theor Biol 2011;269:287–96.
[56] Seligmann H. Pocketknife tRNA hypothesis: anticodons in mammal mitochondrial tRNA side-arm loops translate proteins? Biosystems 2013;113:165–76.

[57] Seligmann H. Putative anticodons in mitochondrial tRNA sidearm loops: pocketknife tRNAs? J Theor Biol 2014;340:155–63.
[58] Seligmann H, Labra A. The relation between hairpin formation by mitochondrial WANCY tRNAs and the occurrence of the light strand replication origin in Lepidosauria. Gene 2014;542:248–57.
[59] Yu CH, Liao JY, Zhou H, Qu LH. The rat mitochondrial Ori L encodes a novel small RNA resembling an ancestral tRNA. Biochem Biophys Res Commun 2008;372:634–8.
[60] Maizels N, Weiner AM. Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation. Proc Natl Acad Sci U S A 1994;91:6729–34.
[61] Maizels N, Weiner AM. Phylogeny from function: the origin of tRNA is in replication, not translation. Chapter 2. In: Fitch WM, Ayala FJ, editors. Tempo and mode in evolution: genetics and paleontology 50 Years after Simpson. Washington: National Academies Press; 1995. p. 25–40 [http://www.ncbi.nlm.nih.gov/books/NBK232211/].
[62] Seligmann H, Raoult D. Unifying view of stem-loop hairpin RNA as origin of current and ancient parasitic and non-parasitic RNAs, including in giant viruses. Curr Opin Microbiol 2016;31:1–8.
[63] Capone JP, Sharp PA, RajBhandary UL. Amber, ochre and opal suppressor tRNA genes derived from a human serine tRNA gene. EMBO J 1985;4:213–21.
[64] Beier H, Grimm M. Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. Nucleic Acids Res 2001;29:4767–82.
[65] Seligmann H. Avoidance of antisense antiterminator tRNA anticodons in vertebrate mitochondria. Biosystems 2010;101:42–50.
[66] Seligmann H. Undetected antisense tRNAs in mitochondria? Biol Direct 2010;5:39.
[67] Seligmann H. Two genetic codes, one genome: frameshifted primate mitochondrial genes code for additional proteins in presence of antisense antitermination tRNAs. Biosystems 2011;105:271–85.
[68] Seligmann H. Putative protein-encoding genes within mitochondrial rDNA and the D-loop region. In: Lin Z, Liu W, editors. Ribosomes: molecular structure, role in biological functions and implications for genetic diseases. Nova Science Publishers; 2013. p. 67–86.
[69] Seligmann H. An overlapping genetic code for frameshifted overlapping genes in *Drosophila* mitochondria: antisense antitermination tRNAs UAR insert serine. J Theor Biol 2012;298:51–76.
[70] Seligmann H. Overlapping genetic codes for overlapping frameshifted genes in Testudines, and *Lepidochelys olivacea* as a special case. Comput Biol Chem 2012;41:18–34.
[71] Barthélémy R-M, Seligmann H. Cryptic tRNAs in chaetognath mitochondrial genomes. Cryptic tRNAs in chaetognath mitochondrial genomes. Comput Biol Chem 2016;62:119–32.
[72] Massey SE, Garey JR. A comparative genomics analysis of codon reassignments reveals a link with mitochondrial proteome size and a mechanism of genetic code change via suppressor tRNAs. J Mol Evol 2007;64:399–410.
[73] Seligmann H. Phylogeny of genetic codes and punctuation codes within genetic codes. Biosystems 2015;129:36–43.
[74] Knight RD, Landweber LF, Yarus M. How mitochondria redefine the code. J Mol Evol 2001;53:299–313.
[75] Sengupta S, Yang X, Higgs PG. The mechanisms of codon reassignments in mitochondrial genetic codes. J Mol Evol 2007;64:662–88.
[76] Ring KL, Cavalcanti AR. Consequences of stop codon reassignment on protein evolution in ciliates with alternative genetic codes. Mol Biol Evol 2008;25:179–86.
[77] Vallabhaneni H, Fan-Minogue H, Bedwell DM, Farabaugh PJ. Connection between stop codon reassignment and frequent use of shifty stop frameshifting. RNA 2009;15:889–97.
[78] Johnson LJ. Pseudogene rescue: an adaptive mechanism of codon reassignment. J Evol Biol 2010;23:1623–30.
[79] Johnson LJ, Cotton JA, Lichtenstein CP, Elgar GS, Nichols RA, Polly PD, Le Comber SC. Stops making sense: translational trade-offs and stop codon reassignment. BMC Evol Biol 2011;11:227.
[80] Campbell JH, O'Donoghue P, Campbell AG, Schwientek P, Sczyrba A, Woyke T, Söll D, Podar M. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. Proc Natl Acad Sci U S A 2013;110:5540–5.
[81] Ivanova NN, Schwientek P, Tripp HJ, Rinke C, Pati A, Huntemann M, Visel A, Woyke T, Kyrpides NC, Rubin EM. Stop codon reassignment in the wild. Science 2014;344:909–13.
[82] Seligmann H. Putative mitochondrial polypeptides coded by expanded quadruplet codons, decoded by antisense tRNAs with unusual anticodons. Biosystems 2012;110:84–106.
[83] Seligmann H, Labra A. Tetracoding increases with body temperature in Lepidosauria. Biosystems 2013;114:155–63.
[84] Riddle DL, Roth JR. Frameshift suppressors. 3. Effects of suppressor mutations on transfer RNA. J Mol Biol 1972;66:495–506.
[85] O'Connor M, Gesteland RF, Atkins JF. tRNA hopping: enhancement by an expanded anticodon. EMBO J 1989;8:4315–23.
[86] Tuohy TM, Thompson S, Gesteland RF, Atkins JF. Seven, eight and nine-membered anticodon loop mutants of tRNA(2Arg) which cause +1 frameshifting. Tolerance of DHU arm and other secondary mutations. J Mol Biol 1992;228:1042–54.
[87] Walker SE, Frederick K. Recognition and positioning of mRNA in the ribosome by tRNAs with expanded anticodons. J Mol Biol 2006;360:599–609.
[88] Dunham CM, Selmer M, Phelps SS, Kelley AC, Suzuki T, Joseph S, Ramakrishnan V. Structures of tRNAs with an expanded anticodon loop in the decoding center of the 30S ribosomal subunit. RNA 2007;13:817–23.
[89] Maehigashi T, Dunkle JA, Miles SJ, Dunham CM. Strucural insights into +1 frameshifting promoted by expanded or modification-deficient anticodon stem loops. Proc Natl Acad Sci U S A 2014;111:12740–5.
[90] Beznosková P, Gunišová S, Valášek LS. Rules of UGA-N decoding by near-cognate tRNAs and analysis of readthrough on short uORFs in yeast. RNA 2016;22:456–66.
[91] Gonzalez DL, Giannerini S, Rosa R. On the origin of the mitochondrial genetic code: towards a unified mathematical framework for the management of genetic information. Nature Precedings; 2012. http://dx.doi.org/10.1038/npre.2012.7136.1.
[92] Baranov PV, Venin M, Provan G. Codon size reduction as the origin of the triplet genetic code. PLoS One 2009;4:e5708.
[93] Root-Bernstein M, Root-Bernstein R. The ribosome as a missing link in the evolution of life. J Theor Biol 2015;367:130–58.
[94] Root-Bernstein R, Root-Bernstein M. The ribosome as a missing link in prebiotic evolution II: ribosomes encode ribosomal proteins that bind to common regions of their own mRNAs and rRNAs. J Theor Biol 2016;397:115–27.
[95] Kowalczewska M, Villard C, Lafitte D, Fenollar F, Raoult D. Global proteomic pattern of *Tropheryma whipplei*: a Whipple's disease bacterium. Proteomics 2009;9:1593–616.
[96] Guerra L, McGarry LM, Robles V, Bielza C, Larrañaga P, Yuste R. Comparison between supervised and unsupervised classifications of neuronal cell types: a case study. Dev Neurobiol 2011;71:71–82.
[97] Brazma A, Vilo J. Gene expression data analysis. FEBS Lett 2000;480:17–24.
[98] Causton H, Quackenbusch J, Brazma A. Analysis of gene expression data matrices. Chapter 4. In: Causton H, Quackenbusch J, Brazma A, editors. Microarray Gene expression data analysis: a beginner's guide. ISBN 1-40510-682-4; 2003. p. 72–133.
[99] Paik S, Kim C. Evolving role of pathology in modern oncology. Chapter 2. In: Bonadonna G, Hortobagyi GN, Valagussa P, editors. Textbook of breast cancer: a clinical guide to therapy3rd ed. ; 2006. p. 17–31 [ISBN 13:978-1-4822-0287-8].
[100] Efron B. Biased versus unbiased estimation. Adv Math 1975;16:259–77.
[101] Quirós PM, Langer T, López-Otin C. New roles for mitochondrial proteases in health, ageing and disease. Nat Rev Mol Cell Biol 2015;16:345–59.
[102] Timerbaev AR, Buchberger W. Inorganic analysis and speciation. Chapter 22. In: Deyl Z, Miksik I, Tagliaro F, Tesatova E, editors. Advanced chromatographic and Electromigration methods in biosciences. Elseviers Science; 1998. p. 963–1012.
[103] Abonnenc M, Mayr M. Proteomics of atherosclerosis. Chapter 13. In: Wick G, Grundtman C, editors. Inflammation and atherosclerosis. Wien: Springer; 2012. p. 249–66.
[104] Piatkov KI, Oh J-H, Liu Y, Varshavsky A. Calpain-generated natural protein fragments as short-lived substrates of the N-end rule pathway. Proc Natl Acad Sci U S A 2014;111:E817–26.
[105] Schmidt W, Egbring R, Havemann K. Effect of elastase-like and chymotrypsin-like natural proteases from human granulocytes on isolated clotting factor XIII. Thromb Res 1975;6:315–29.
[106] Andrews AT, Alichanidis E. Proteolysis of caseins and the proteose-peptone fraction of bovine milk. J Dairy Res 1983;50:275–90.
[107] Rietschel B, Arrey TN, Meyer B, Bornemann B, Schuerken M, Karas M, Poetsch A. Elastase digests. New ammunition for shotgun membrane proteomics. Mol Cell Proteomics 2009;8:1029–43.
[108] Wildes D, Wells JA. Sampling the N-terminal proteome of human blood. Proc Natl Acad Sci U S A 2010;107:4561–6.
[109] Leonelli L, Pelton J, Schoeffler A, Dahlbeck D, Berger J, Wemmer DE, Staskawicz B. Structural elucidation and functional characterization of the Hyaloperonospora arabidopsidis effector protein ATR13. PLoS Pathog 2011;7:e1002428.
[110] Venter E, Smith RD, Payne SH. Proteogenomic analysis of bacteria and Archaea: a 46 organism case study. PLoS One 2011;6:e27587.
[111] Volkmann G, Volkmann V, Liu XQ. Site-specific protein cleavage in vivo by an intein-derived protease. FEBS Lett 2012;586:79–84.
[112] Gueugneau M, Coudy-Gandilhon C, Gourbeyre O, Chambon C, Combaret L, Polge C, Taillandier D, Attaix D, Friguet B, Maier AB, Butler-Browne G, Béchet D. Proteomics of muscle chronological ageing in post-menopausal women. BMC Genomics 2014;15:1165.
[113] Garzon R, Volinia S, Papaioannou D, Nicolet D, Kohlschmidt J, Yan PS, Mrozek K, Bucci D, Carroll AJ, Baer MR, Wetzler M, Carter TH, Powell BL, Kolitz JE, Moore JO, Eisfeld AK, Blachly JS, Blum W, Caliguiri MA, Stone RM, Marcucci G, Croce CM, Byrd JC, Bloomfield CD. Expression and prognostic impact of lncRNAs in acute myeloid leukemia. Proc Natl Acad Sci U S A 2014;111:18679–84.
[114] Lobanov AV, Turanov AA, Hatfield DL, Gladyshev VN. Dual functions of codons in the genetic code. Crit Rev Biochem Mol Biol 2010;45:257–65.
[115] O'Donoghue P, Prat L, Heinemann IU, Ling J, Odoi K, Liu WR, Söll D. Near-cognate suppression of amber, opal and quadruplet codons competes with aminoacyl-tRNAPyl for genetic code expansion. FEBS Lett 2012;583:3931–7.
[116] Odoi KA, Huang Y, Rezenom YH, Liu WR. Nonsense and sense suppression abilities of original and derivative *Methanosarcina mazei* pyrrolysyl-tRNA synthetase-tRNA(Pyl) pairs in the *Escherichia coli* BL21(DE3) cell strain. PLoS One 2013;8:e57035.
[117] Käll L, Storey JD, MacCoss MJ, Noble WS. Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Res 2008;7:40–4.
[118] Brosch M, Yu L, Hubbard T, Choudhary J. Accurate and sensitive peptide identification with MascotPercolator. J Proteome Res 2009;8:3176–81.
[119] Spivak M, Weston J, Bottou L, Käll L, Noble WS. Improvements to the percolator algorithm for peptide identification from shotgun proteomics data sets. J Proteome Res 2009;8:3737–45.
[120] Eng JK, McCormack AL, Yates III JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 1994;5:976–89.
[121] Fisher RA. Combining independent tests of significance. Am Stat 1648;2:30–1.
[122] Seligmann H. Natural chymotrypsin-like-cleaved human mitochondrial peptides confirm tetra-, pentacodon, non-canonical RNA translations. Biosystems 2016;147:78–93.
[123] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
[124] Lopez JV, Yukhi N, Masuda R, Modi W, O'Brien SJ. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J Mol Evol 1994;39:174–90.
[125] Lopez JV, Cumver M, Stephens JC, Johnoson WE, O'Brien SJ. Rates of nuclear and cytoplasmic mitochondrial DNA sequence divergence in mammals. Mol Biol Evol 1997;14:277–86.
[126] Lopez JV, Stephens JC, O'Brien SJ. The long and short nuclear mitochondrial DNA (Numt) lineages. Trends Ecol Evol 1997;12:114.

[127] Zhang DX, Hewitt GM. The long and short of nuclear mitochondrial DNA (Numt) lineages reply from D-X. Trends Ecol Evol 1997;12:114.
[128] Bensasson D, Zhang D, Hartl DL, Hewitt GM. Mitochondrial pseudogenes: evolution's misplaces witnesses. Trends Ecol Evol 2001;16:314–21.
[129] Tourmen Y, Baris O, Dessen P, Jacques C, Malthièry Y, Reynier P. Structure and chromosomal distribution of human mitochondrial pseudogenes. Genomics 2002;80:71–7.
[130] Bensasson D, Feldman MW, Petrov DA. Rates of DNA duplication and mitochondrial DNA insertion in the human genome. J Mol Evol 2003;57:343–54.
[131] Ricchetti M, Tekala F, Dujon B. Continued colonization of the human genome by mitochondrial DNA. PLoS Biol 2004;2:E273.
[132] Thalman O, Hebler J, Poinar HN, Päabo S, Vigilant L. Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes. Mol Ecol 2004;13:321–35.
[133] Richly E, Leister D. NUMTs in sequenced eukaryotic genomes. Mol Biol Evol 2004;21:1081–4.
[134] Schmitz J, Piskurek O, Zischler H. Forty million years of independent evolution: a mitochondrial gene and its corresponding nuclear pseudogene in primates. J Mol Evol 2005;61:1–11.
[135] Thalman O, Serre D, Hofreiter M, Lukas D, Eriksson J, Vigilant L. Nuclear insertions help and hinder inference of the evolutionary history of gorilla mtDNA. Mol Ecol 2005;14:179–88.
[136] Yao YG, Kong QP, Salas A, Bandelt HJ. Pseudomitochondrial genome haunts disease studies. J Med Genet 2008;45:769–72.
[137] Hazkani-Covo E, Zeller RM, Martin W. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet 2010;6:e1000834.
[138] Ramos A, Barbena E, Mateiu L, del Mar González M, Mairal Q, Lima M, Montiel R, Aluja MP, Santos C. Nuclear insertions of mitochondrial origin: database updating and usefulness in cancer studies. Mitochondrion 2011;11:946–53.
[139] Tsuji J, Frith MC, Tomii K, Horton P. Mammalian NUMT insertion is non-random. Nucleic Acids Res 2012;40:9073–88.
[140] Soto-Calderón ID, Lee EJ, Jensen-Seaman MI, Anthony NM. Factors affecting the relative abundance of nuclear copies of mitochondrial DNA (numts) in hominoids. J Mol Evol 2012;75:102–11.
[141] Soto-Calderón ID, Clark NJ, Wildschutte JV, DiMattio K, Jensen-Seaman MI, Anthony NM. Identification of species-specific nuclear insertions of mitochondrial DNA (numts) in gorillas and their potential as population genetic markers. Mol Phylogenet Evol 2014;81:61–70.
[142] Smigrodzki RM, Khan SM. Mitochondrial microheteroplasmy and a theory of aging and age-related disease. Rejuvenation Res 2005;8:172–98.
[143] Rose G, Passarino G, Scornaienchi V, Romeo G, Dato S, Bellizzi D, Mari V, Feraco E, Maletta R, Bruni A, Franceschi C, De Benedictis G. The mitochondrial DNA control region shows genetically correlated levels of heteroplasmy in leukocytes of centenarians and their offspring. BMC Genomics 2007;8:293.
[144] Stefano GB, Kream RM. Mitochondrial DNA heteroplasmy in human health and disease. Biomed Rep 2016;4:259–62.
[145] Ramos A, Santos C, Mateiu L, del Mar Gonzalez M, Alvarez L, Azevedo L, Amorim A, Pilar Aluja M. Frequency and pattern of heteroplasmy in the complete human mitochondrial genome. PLoS One 2013;8:e74636.
[146] Frenkel-Morgenstern M, Gorohovski A, Lacroix V, Rogers M, Ibanez K, Boullosa C, Andres Leon E, Ben-Hur A, Valencia A. ChiTaRS: a database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data. Nucleic Acids Res 2013;41:D142–51.
[147] Yang W, Wu J-M, Bi A-D, Ou-yang Y, Shen H-H, Chim G-W, Zhou J-H, Weiss E, Holman EP, Liao DJ. Possible formation of mitochondrial-RNA containing chimeric or trimeric RNA implies a post-transcriptional and post-splicing mechanism for RNA fusion. PLoS One 2013;8:e77016.
[148] Xie B, Yang W, Chen L, Jiang H, Liao Y, Liao DJ. Two RNAs or DNAs may artificially fuse together at a short homologous sequence (SHS) during reverse transcription or polymerase chain reactions, and thus reporting an SHS-containing chimeric RNA requires extra caution. PLoS One 2016;11:e0154855.
[149] Allen JF. Why chloroplasts and mitochondria retain their own genomes and genetic systems: colocation for redox regulation of gene expression. Proc Natl Acad Sci U S A 2015;112:10231–8.
[150] Bauer NC, Doetsch PW, Corbett AH. Mechanisms regulating protein localization. Traffic 2015;16:1039–61.
[151] Horvath SE, Rampelt H, Oeljeklaus S, Warscheid B, van der Laan M, Pfanner N. Role of membrane contact sites in protein import into mitochondria. Protein Sci 2015;24:277–97.
[152] Kunze M, Berger J. The similarity between N-terminal targeting signals for protein import into different organelles and its evolutionary relevance. Front Physiol 2015;6:259.
[153] Aerni HR, Shifman MA, Rogulina S, O'Donoghue P, Rinehart J. Revealing the amino acid composition of proteins within an expanded genetic code. Nucleic Acids Res 2015;43:e8.