Genomics 101 (2013) 282–289

Contents lists available at SciVerse ScienceDirect

Genomics

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

Universal tight correlation of codon bias and pool of RNA codons (codonome): The genome is optimized to allow any distribution of gene expression values in the transcriptome from bacteria to humans

Allison Piovesan, Lorenza Vitale, Maria Chiara Pelleri, Pierluigi Strippoli *

Department of Experimental, Diagnostic and Specialty Medicine, Activity of Histology, Embryology and Applied Biology, University of Bologna, via Belmeloro 8, 40126 Bologna (BO), Italy

ARTICLE INFO

Article history:
Received 4 December 2012
Accepted 21 February 2013
Available online 1 March 2013

Keywords:
Codon bias
Codonome
Pool of RNA codons
Genome
Transcriptome

ABSTRACT

Codon bias is the phenomenon in which distinct synonymous codons are used with different frequencies. We define here the "codonome value" as the total number of codons present across all the expressed mRNAs in a given biological condition. We have developed the "CODONOME" software, which calculates the codon bias and, following integration with a gene expression profile, estimates the actual frequency of each codon at the transcriptome level (codonome bias) of a given tissue. Systematic analysis across different human tissues and multiple species shows a surprisingly tight correlation between the codon bias and the codonome bias. An aneuploidy and cancer condition such as that of Down Syndrome-related acute megakaryoblastic leukemia (DS-AMKL), does not appear to alter this relationship. The law of correlation between codon bias and codonome emerges as a property of the distribution and range of the number, sequence and expression level of the genes in a genome.

1. Introduction

Codon bias is the well-known phenomenon in which distinct synonymous codons (different codons encoding the same amino acid) are used with different frequencies (reviewed in [1]). This has been observed in species from all taxa.

The codons that are used more frequently are also referred to as preferred codons or "optimal codons" [2]. Previously, optimal and non-optimal codons for each amino acid had been shown to differ between species [3], in particular between distantly related species.

Codon bias can be explained by two hypotheses: the mutational (or neutral) explanation and the selectionist (or natural selection) explanation [4]. According to the mutational explanation, codon bias originates from basal mutational processes, which cause neither advantage nor damage. The selectionist explanation asserts that synonymous mutations influence the fitness of an organism, and can thus be promoted (or repressed) throughout evolution. These two types of mechanisms are not mutually exclusive, and both are useful to understanding the phenomenon within and between species. In particular, the latter explanation is typically cited to explain variation in codon usage across a genome or across a gene [4].

In eukaryotic genes, the most frequently used codons have a bigger content of G + C at the third codon position [5], especially in human genes, according to the mutational (or neutral) explanation of the intra-genomic heterogeneity of the human genome [6]. Preferred codons also vary between genes of the same organism: expressed genes have a codon usage pattern, different from poorly expressed genes, optimized to increase translational efficiency [5] and to minimize the cost of nonsense errors during protein translation [7]. For example, optimal codons are recognized by more abundant transfer RNA molecules in several unicellular organisms [8] and in several eukaryotes [9]. These findings support the selectionist explanation (natural selection).

Intriguingly, Plotkin et al. [10] studied the role of codon usage between tissue-specific human genes. Comparing testis- to uterus-specific genes and brain- to liver-specific genes, they reported a characteristic codon usage in genes expressed in one tissue as compared to those expressed in another. Other comparisons (e.g. liver versus uterus) do not exhibit any significantly different codon usage. However, the authors suggested that codon bias might optimize translation of tissue-specific genes.

Furthermore Sémon et al. [11], analyzing 2,126 human tissue-specific genes expressed in 18 different tissues, found that the difference in synonymous codon usage between tissue-specific genes expressed in different tissues is significant, but weak, as the intra-tissue variability of synonymous codon usage is much smaller than the inter-tissue variability. Additionally, these authors correlated the synonymous codon usage variability to inter-gene G + C content at the third position differences, also affecting introns and intergenic regions, due to the isochore scale variation of substitution patterns [11].

At present several indexes are used to analyze codon bias, e.g. "Fop" [2], "CAI" [12], "E-CAI" [13], "CBI" [14], "Nc" [15], "G + C content

* Corresponding author. Fax: +39 0512094110.
E-mail addresses: allison.piovesan2@unibo.it (A. Piovesan), lorenza.vitale@unibo.it (L. Vitale), mariachiara.pelleri2@unibo.it (M.C. Pelleri), pierluigi.strippoli@unibo.it (P. Strippoli).

0888-7543/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2013.02.009

of the third codon position" [6]. Several software programs for calculating these indexes are available free of charge on the internet (e.g. CodonW, [16]; JCat, [17]; INCA, [18]).

Codon bias is usually related to the genome at the level of genome sequence. We wondered if the proportion of used codons could vary during the expression of a whole transcriptome, introducing the new concept of determining the actual pool of codons borne by all the messenger RNAs (mRNAs) in the cell. To this end, the codon bias should be multiplied by the relative estimated number of molecules of that mRNA in the transcriptome. This would offer the possibility of searching for relationships between codon usage at the genome and transcriptome levels. Here we define the "codonome value" as the total number of codons (n) present across all the transcriptome mRNAs each expressed at a certain level (x) in a given biological condition (cv $= \Sigma(n \times x)$ for the mRNAs pool). We have developed the innovative "CODONOME" software, which is able to calculate the frequency of each codon in any reference (RefSeq) mRNA sequence and, following integration with a profile of gene expression values, to estimate the actual frequency of each codon in the mRNA pool derived from a specific tissue of a given organism (Fig. 1). In addition, to investigate a possible cell adaptation aimed to optimize the translation process, we grouped these frequencies by encoded amino acid, each being related to its specific aminoacyl-tRNA synthetase (aaRS), to determine whether some relationships exist between codon usage and aaRS mRNA expression level, a still unexplored field.

We used gene expression values obtained from independent transcriptome datasets for a certain condition available in the Gene Expression Omnibus (GEO) database [19,20] following intra- and inter-sample normalization using TRAM software [21].

We performed a systematic analysis, varying the tissue examined within human species and investigating a pool of representative species from bacteria to humans. We also tested the codonome values in a pathological condition with a general disturbance of gene expression, i.e. the aneuploid blast from Down Syndrome (DS)-related acute megakaryoblastic leukemia (AMKL), as well as in an extremely differentiated tissue with a remarkable expression preponderance of a very small number of proteins (human circulating blood erythrocytes samples). Moreover, we compared the same tissue (brain) from two different organisms (*Homo sapiens* and *Danio rerio*). In addition, we determined the codonome values in lower organisms (*Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Escherichia coli*) in order to search for general laws governing the structure of the codonome.

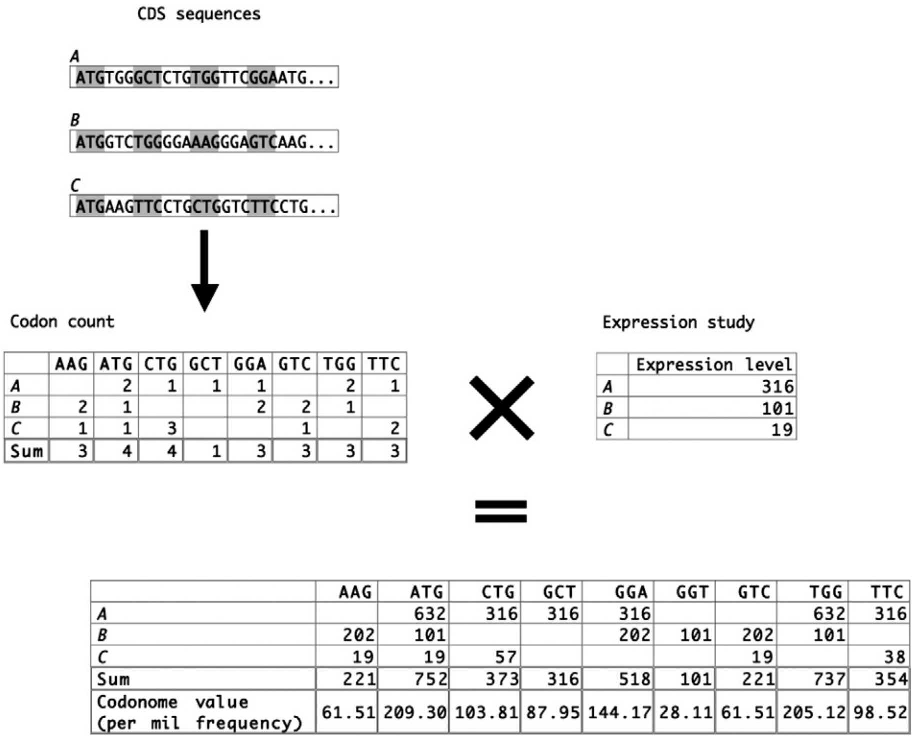
The significance of the correlation coefficients was determined for: the per mil frequencies of codons (codon bias) vs. the per mil frequencies of the codons number multiplied by expression value (codonome bias); the per mil frequencies of codons (codon bias) grouped by aaRS vs. the aaRS expression values, and the per mil frequencies of the codons number multiplied by expression value (codonome bias) grouped by aaRS vs. the aaRS expression values.

While we did not see a significant relationship between codonome values and aaRS mRNA expression level, our findings clearly show that the codon frequency in the genome is reflected proportionally in the transcriptome, irrespective of the considered tissue, species, or pathological state. This implies that transcriptome codonome values remain in excellent correlation with genome codon bias in a wide range of conditions, thus allowing the transcription of any gene subset at any level of abundance without altering the tight bond between codon bias at genome level and codonome at transcriptome level.

2. Results

2.1. Database construction and computational analysis

Following importation of the normalized expression data, we found an available expression value for: 27,850 out of 29,538 NM


CDS sequences
A
ATGTGGGCTCTGTGGTTCGGAATG...
B
ATGGTCTGGGGAAAGGGAGTCAAG...
C
ATGAAGTTCCTGCTGGTCTTCCTG...
×
=

Codon count

	AAG	ATG	CTG	GCT	GGA	GTC	TGG	TTC
A		2	1	1	1		2	1
B	2	1			2	2	1	
C	1	1	3			1		2
Sum	3	4	4	1	3	3	3	3

Expression study

	Expression level
A	316
B	101
C	19

	AAG	ATG	CTG	GCT	GGA	GGT	GTC	TGG	TTC
A		632	316	316	316			632	316
B	202	101			202	101	202	101	
C	19	19	57				19		38
Sum	221	752	373	316	518	101	221	737	354
Codonome value (per mil frequency)	61.51	209.30	103.81	87.95	144.17	28.11	61.51	205.12	98.52

Fig. 1. Pipeline of the "CODONOME" software. For each RefSeq ("NM_" type) entry considered, we counted the occurrence of each codon. We then calculated the count sum of each codon for the whole gene set (the per mil frequencies of each codon sum in relation to the sum of all codons for the whole gene set gives the codon bias). We then multiplied the codon count for each gene by the normalized expression value of that gene. Finally, we summed the total number of each codon for the whole gene set. The per mil frequencies of each codon sum in relation to the sum of all codons for the whole gene set give the transcriptome codonome bias (example simulation for a hypothetical gene set composed of three genes "A," "B" and "C," assuming the existence of nine codons).

Table 1
The ten human genes with the highest and the lowest expression values in the studied datasets. The units of expression are given, following intra- and inter-sample normalization by the TRAM software, as percentage of the mean value.

Homo sapiens					
Brain		Erythrocytes		DS-AMKL cells	
Gene symbol	Value	Gene symbol	Value	Gene symbol	Value
UBC	3088.81	*HBA2*	47816.17	*RPS18*	6592.43
TUBA1B	3044.12	*SLC25A39*	43176.51	*RPL41*	6588.66
TUBA1C	2634.12	*HBA1*	36616.34	*EEF1A1*	6583.97
UBB	2591.82	*HBB*	31649.48	*RPS10*	6233.13
CALM2	2577.65	*UBB*	25239.87	*RPS3A*	6175.32
RPL41	2549.75	*RPL21*	20966.99	*RPL23A*	6078.20
GAPDH	2316.03	*HBM*	18959.67	*RPS23*	5836.76
RPL23A	2170.38	*STRADB*	18836.62	*TPT1*	5814.83
SPARCL1	2075.33	*HBG2*	15431.72	*RPS3*	5769.17
CFL1	2019.56	*GYPC*	12556.57	*RPLP0*	5579.53
C7orf72	7.07	*RBL1*	2.19	*AWAT1*	1.15
FBXO47	7.06	*C14orf105*	2.14	*DSG4*	1.14
FABP12	7.04	*AHR*	2.12	*UBL4B*	1.14
ACER2	6.73	*ZNF165*	1.96	*MAS1L*	1.14
CXorf51	6.23	*C17orf75*	1.92	*KCTD21*	1.13
PTPRQ	5.82	*TMEM232*	1.63	*TTC16*	1.11
SLC36A2	5.75	*IFT74*	1.59	*TSSK3*	1.09
RSPH4A	5.33	*SLC16A4*	1.29	*DEFB118*	1.07
TAS2R20	4.41	*ZNF674*	1.28	*SERINC2*	1.04
C5orf52	4.36	*DMXL1*	1.13	*CCDC135*	1.04

RefSeq entries for human brain tissue; 26,589 out of 29,538 NM entries for human circulating blood erythrocyte; 27,506 out of 29,538 NM entries for human DS AMKL cells; 6,642 out of 14,174 NM entries in *D. rerio*; 19,281 out of 23,894 NM for *C. elegans*; 4,673 out of 5,882 NM for *S. cerevisiae*; 2,426 out of 4,319 NM for *E. coli*. A summary of the range in the expression data and of the main genes with the highest and lowest expression values for the considered datasets is given in Table 1 for *H. sapiens* and Supplementary Table 1 for the other investigated species available at http://apollo11.isto.unibo.it/suppl/.

The frequency of each codon at genome level corresponds to the codon bias values already known for each genome [22]. In addition, codon sums at transcriptome level (codonome value), accounting for the abundance of each mRNA bearing that codon, has been calculated as per mil frequencies of each codon, obtaining the codonome bias (see Tables 2 and 3 for *H. sapiens*, Supplementary Table 2 for the other investigated species and Supplementary Table 3 for human simulations).

We also grouped per mil frequencies for each codon (at genome level and at transcriptome level) by the corresponding aaRS and then loaded their expression values from the same normalized files as before. With the exception of *H. sapiens*, we could not find expression values for some aaRS for any of the investigated species (see Table 4 for *H. sapiens* and Supplementary Table 4 for the other investigated species).

2.2. Statistical analysis

We exported the following results in order to submit them to statistical analysis using first default and then test calculations: a) codon bias, b) codonome bias, c) codon bias grouped by aaRS, d) codonome bias grouped by aaRS, and e) the aaRS expression values ("a," "b," "c" and "d" are expressed as per mil frequencies). An example of correlation graphs for human brain is shown in Fig. 2 (see Supplementary

Table 2
The per mil frequencies of codons (codon bias) and the per mil frequencies of the codon counts multiplied by the respective expression value (codonome bias) in the human studied datasets.

	Brain		Erythrocytes		DS-AMKL cells	
Codon	Codon bias	Codonome bias	Codon bias	Codonome bias	Codon bias	Codonome bias
AAA	25.88	24.46	25.83	22.04	25.84	27.86
AAC	18.90	19.18	18.94	19.37	18.93	18.89
AAG	32.30	34.17	32.41	35.26	32.32	37.13
AAT	17.58	16.52	17.52	14.76	17.56	17.74
ACA	15.41	14.58	15.36	13.51	15.42	14.95
ACC	18.40	19.06	18.42	21.55	18.44	18.12
ACG	5.96	6.19	5.98	5.92	5.97	5.41
ACT	13.53	12.97	13.49	12.77	13.52	14.10
AGA	12.23	11.30	12.14	10.54	12.18	12.40
AGC	19.72	19.73	19.74	20.38	19.75	17.55
AGG	11.64	11.34	11.56	11.51	11.61	10.93
AGT	12.87	12.19	12.85	11.36	12.83	12.39
ATA	7.60	6.72	7.51	5.59	7.58	6.89
ATC	20.06	21.16	20.13	22.21	20.13	20.65
ATG	21.45	22.00	21.49	22.13	21.48	22.78
ATT	16.20	15.67	16.18	13.94	16.22	17.41
CAA	12.84	11.63	12.74	10.99	12.80	12.06
CAC	14.80	14.79	14.77	16.03	14.81	13.74
CAG	34.76	35.29	34.83	35.46	34.76	34.31
CAT	11.12	10.40	11.05	9.87	11.10	10.72
CCA	17.67	17.03	17.69	16.64	17.63	17.52
CCC	19.81	20.50	19.84	22.17	19.81	18.67
CCG	6.97	7.32	6.97	7.48	6.96	6.34
CCT	18.17	17.86	18.19	18.20	18.13	18.41
CGA	6.28	6.34	6.33	6.21	6.28	6.81
CGC	10.10	10.98	10.13	11.46	10.12	10.33
CGG	11.47	12.14	11.56	13.05	11.49	11.24
CGT	4.54	4.82	4.56	4.81	4.53	5.67
CTA	7.09	6.57	7.07	6.43	7.08	6.82
CTC	18.51	18.68	18.48	19.81	18.54	17.10
CTG	38.38	39.51	38.39	43.61	38.43	36.10
CTT	13.33	12.57	13.28	11.88	13.30	13.64

Table 3
The per mil frequencies of codons (codon bias) and the per mil frequencies of the codon counts multiplied by the relative expression value (codonome bias) in the human studied datasets.

	Brain		Erythrocytes		DS-AMKL cells	
Codon	Codon bias	Codonome bias	Codon bias	Codonome bias	Codon bias	Codonome bias
GAA	31.28	30.18	31.26	26.45	31.21	32.33
GAC	25.23	26.43	25.34	26.67	25.27	24.65
GAG	40.42	42.76	40.54	41.98	40.41	40.03
GAT	22.88	22.85	22.97	20.38	22.90	24.79
GCA	16.32	16.14	16.34	15.02	16.30	16.89
GCC	27.43	28.96	27.48	31.43	27.48	27.10
GCG	7.13	7.63	7.12	8.25	7.12	6.81
GCT	18.45	18.91	18.48	18.39	18.43	20.92
GGA	16.70	16.16	16.68	15.15	16.70	17.12
GGC	21.80	22.88	21.87	25.27	21.83	21.68
GGG	15.96	16.27	15.96	16.46	15.97	14.96
GGT	10.73	10.91	10.76	11.13	10.73	12.51
GTA	7.30	6.88	7.30	6.13	7.29	7.69
GTC	13.99	14.29	14.01	14.67	14.03	13.88
GTG	27.22	28.26	27.32	31.29	27.30	27.21
GTT	11.23	10.86	11.25	9.83	11.23	12.49
TAA	0.66	0.74	0.66	0.96	0.66	1.00
TAC	14.58	15.00	14.61	15.29	14.60	14.16
TAG	0.49	0.52	0.49	0.52	0.49	0.53
TAT	12.12	11.61	12.08	10.92	12.11	12.44
TCA	12.78	11.83	12.70	11.27	12.73	11.93
TCC	17.41	17.63	17.39	18.29	17.41	16.43
TCG	4.45	4.65	4.46	4.66	4.46	4.04
TCT	15.38	14.74	15.36	14.18	15.35	15.59
TGA	1.10	1.17	1.10	1.48	1.10	1.19
TGC	11.78	11.57	11.70	11.59	11.79	10.11
TGG	12.09	11.78	12.05	12.39	12.09	11.20
TGT	10.38	9.53	10.26	9.23	10.34	9.38
TTA	7.96	7.05	7.94	5.74	7.94	7.56
TTC	19.09	19.39	19.07	20.83	19.12	18.36
TTG	12.89	12.38	12.88	11.55	12.88	13.06
TTT	17.20	16.35	17.16	15.66	17.19	17.25

Table 4
The per mil frequencies of codons grouped by aminoacyl-tRNA synthetase (codon bias by aaRS), the per mil frequencies of the expressed codon grouped by aminoacyl-tRNA synthetase (codonome bias by aaRS), and the aminoacyl-tRNA synthetases expression values in human studied datasets.

	Brain			Erythrocytes			DS-AMKL cells		
Gene symbol	Codon bias	Codonome bias	Expression value	Codon bias	Codonome bias	Expression value	Codon bias	Codonome bias	Expression value
AARS	69.33	71.62	340.77	69.43	73.21	111.55	69.25	71.43	182.57
CARS	22.19	21.13	171.68	22.01	20.82	65.49	22.24	19.61	112.81
DARS	48.08	49.27	90.51	48.29	47.04	27.13	48.00	49.33	265.46
EPRS	134.34	135.65	97.28	134.55	132.93	45.11	134.03	133.57	139.06
FARSA	36.30	35.73	116.50	36.24	36.50	45.93	36.31	35.49	88.23
FARSB	36.30	35.73	83.59	36.24	36.50	96.01	36.31	35.49	72.70
GARS	65.26	66.21	240.98	65.33	68.30	78.81	65.04	65.97	350.40
HARS	25.92	25.20	192.71	25.83	25.88	8.31	25.97	24.50	116.31
IARS	49.10	49.88	246.07	49.10	49.92	21.82	49.08	50.17	357.47
KARS	58.19	58.65	202.52	58.24	57.26	227.30	58.16	64.99	629.51
LARS	98.21	96.84	141.49	98.08	98.94	51.21	98.19	94.30	169.68
MARS	16.17	15.64	122.54	16.14	13.92	48.41	16.22	17.36	97.31
NARS	36.43	35.66	425.68	36.43	34.20	20.20	36.47	36.56	291.97
QARS	47.63	46.94	162.10	47.58	46.40	23.15	47.59	46.46	863.36
RARS	56.30	56.97	78.05	56.32	57.59	11.25	56.38	57.85	87.05
SARS	82.57	80.85	255.34	82.48	79.96	60.12	82.71	78.21	187.42
TARS	53.20	52.67	76.46	53.15	53.56	79.01	53.30	52.36	144.87
VARS	59.71	60.30	75.19	59.81	61.83	72.68	59.72	61.00	64.98
WARS	12.08	11.75	137.52	12.03	12.40	29.93	12.16	11.23	135.00
YARS	26.70	26.58	172.73	26.68	26.34	87.21	26.65	26.67	267.99

Fig. 1 for human circulating blood erythrocyte graphs, Supplementary Fig. 2 for human DS-AMKL cells graphs and Supplementary Figs. 3–6 for the other investigated species, *D. rerio*, *C. elegans*, *S. cerevisiae* and *E. coli,* respectively). Correlation coefficients and *p* values for each comparison are listed in Table 5.

The comparisons between the codon bias and the codonome bias, as well as these values grouped by aaRS, show correlation coefficients very close to 1, with a *p* value always <0.0001, for all the investigated tissues and species. When random and permuted numbers are used instead of human real expression values, the pattern does not change; rather, the correlation coefficient is often even closer to 1.

When grouped by aaRS, codon bias and codonome bias, when compared to aaRS mRNA expression values, show no correlation, with really low coefficients (sometimes even negative ones), and *p* values of at

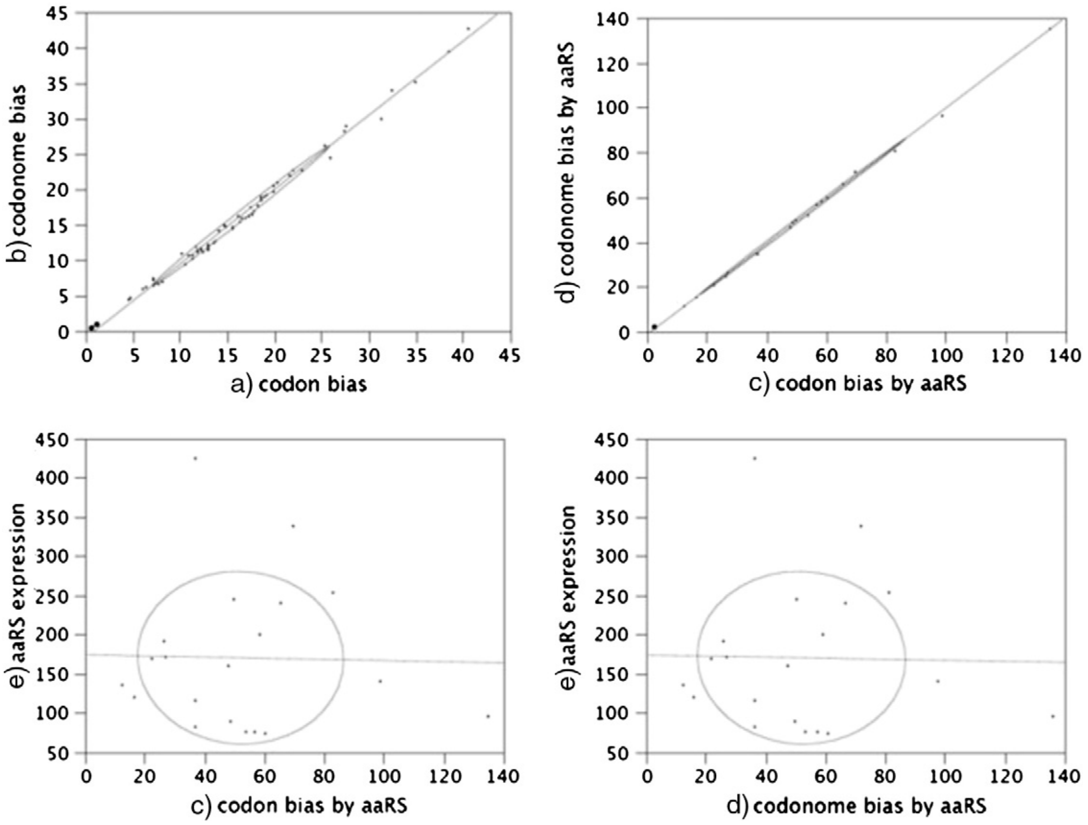


Fig. 2. Correlation graphs in human brain "a": the per mil frequencies of each codon at genome level (codon bias); "b": the per mil frequencies of each codon multiplied by expression value (codonome bias); "c": the per mil frequencies of codons grouped by aminoacyl-tRNA synthetase (codon bias by aaRS); "d": the per mil frequencies of real expressed codons grouped by aaRS (codonome bias by aaRS); "e": the aaRS expression values. See Table 5 for correlation coefficients and *p* values. The elliptic line represents the density ellipse at 0.50.

Table 5

Correlation coefficients (r) and p values of comparisons. a) The per mil frequencies of codons (codon bias), b) the per mil frequencies of codons number multiplied by expression value (codonome bias), c) the per mil frequencies of codons grouped by aminoacyl-tRNA synthetase (codon bias by aaRS), d) the per mil frequencies of real expressed codon grouped by aminoacyl-tRNA synthetase (codonome bias by aaRS), e) the aminoacyl-tRNA synthetases mRNA expression values (aaRS expression). NS, not significant.

Subset	*X* variable	*Y* variable	(r)	*p* Value
Human brain	a) Codon bias	b) Codonome bias	0.996517	<0.0001
	c) Codon bias by aaRS	d) Codonome bias by aaRS	0.999457	<0.0001
	c) Codon bias by aaRS	e) aaRS expression	−0.022970	NS
	d) Codonome bias by aaRS	e) aaRS expression	−0.021290	NS
Human brain with absolute numbers instead of per mil frequencies	a) Codon bias	b) Codonome bias	0.996546	<0.0001
	c) Codon bias by aaRS	d) Codonome bias by aaRS	0.999457	<0.0001
	c) Codon bias by aaRS	e) aaRS expression	−0.051980	NS
	d) Codonome bias by aaRS	e) aaRS expression	−0.050950	NS
Human brain with a first permutation of the expression values	a) Codon bias	b) Codonome bias	0.999838	<0.0001
	c) Codon bias by aaRS	d) Codonome bias by aaRS	0.999937	<0.0001
	c) Codon bias by aaRS	e) aaRS expression	−0.024030	NS
	d) Codonome bias by aaRS	e) aaRS expression	−0.022320	NS
Second human brain with a second permutation of the expression values	a) Codon bias	b) Codonome bias	0.999943	<0.0001
	c) Codon bias by aaRS	d) Codonome bias by aaRS	0.999982	<0.0001
	c) Codon bias by aaRS	e) aaRS expression	−0.023000	NS
	d) Codonome bias by aaRS	e) aaRS expression	−0.023420	NS
Human brain with non-normalized expression values	a) Codon bias	b) Codonome bias	0.998791	<0.0001
	c) Codon bias by aaRS	d) Codonome bias by aaRS	0.999925	<0.0001
	c) Codon bias by aaRS	e) aaRS expression	−0.052180	NS
	d) Codonome bias by aaRS	e) aaRS expression	−0.050720	NS
Human brain with random expression values from 1 to 10^4	a) Codon bias	b) Codonome bias	0.999990	<0.0001
	c) Codon bias by aaRS	d) Codonome bias by aaRS	0.999996	<0.0001
	c) Codon bias by aaRS	e) aaRS expression	−0.052180	NS
	d) Codonome bias by aaRS	e) aaRS expression	−0.051980	NS
Human circulating blood erythrocytes	a) Codon bias	b) Codonome bias	0.979111	<0.0001
	c) Codon bias by aaRS	d) Codonome bias by aaRS	0.998358	<0.0001
	c) Codon bias by aaRS	e) aaRS expression	0.119425	NS
	d) Codonome bias by aaRS	e) aaRS expression	0.126501	NS
Human circulating blood erythrocytes with random expression values from 1 to 10^5	a) Codon bias	b) Codonome bias	0.998791	<0.0001
	c) Codon bias by aaRS	d) Codonome bias by aaRS	0.937790	<0.0001
	c) Codon bias by aaRS	e) aaRS expression	0.074207	NS
	d) Codonome bias by aaRS	e) aaRS expression	0.027150	NS
Human brain from patients affected by Trisomy 21 and Acute Megakaryoblastic Leukemia	a) Codon bias	b) Codonome bias	0.990428	<0.0001
	c) Codon bias by aaRS	d) Codonome bias by aaRS	0.996594	<0.0001
	c) Codon bias by aaRS	e) aaRS expression	0.015140	NS
	d) Codonome bias by aaRS	e) aaRS expression	0.041824	NS
Danio rerio brain	a) Codon bias	b) Codonome bias	0.986128	<0.0001
	c) Codon bias by aaRS	d) Codonome bias by aaRS	0.992635	<0.0001
	c) Codon bias by aaRS	e) aaRS expression	0.384812	>0.1743
	d) Codonome bias by aaRS	e) aaRS expression	0.431892	>0.1230
Caenorhabditis elegans	a) Codon bias	b) Codonome bias	0.979831	<0.0001
	c) Codon bias by aaRS	d) Codonome bias by aaRS	0.991042	<0.0001
	c) Codon bias by aaRS	e) aaRS expression	0.026048	NS
	d) Codonome bias by aaRS	e) aaRS expression	0.026811	NS
Saccharomyces cerevisiae	a) Codon bias	b) Codonome bias	0.991204	<0.0001
	c) Codon bias by aaRS	d) Codonome bias by aaRS	0.996790	<0.0001
	c) Codon bias by aaRS	e) aaRS expression	0.224813	>0.3698
	d) Codonome bias by aaRS	e) aaRS expression	0.217939	>0.3850
Escherichia coli	a) Codon bias	b) Codonome bias	0.988796	<0.0001
	c) Codon bias by aaRS	d) Codonome bias by aaRS	0.996650	<0.0001
	c) Codon bias by aaRS	e) aaRS expression	0.510390	>0.0519
	d) Codonome bias by aaRS	e) aaRS expression	0.502108	>0.0565

least >0.1 (p values at least >0.05 only in the case of *E. coli* dataset), even when using random and permuted expression values.

3. Discussion

Codon bias is a well-known phenomenon, observed in species from bacteria to mammals. Preferred codons can differ dramatically between species and also within a genome. The direct application of this phenomenon is usually the optimization of the heterolog expression of a protein exploiting the codon bias of the guest. It has been demonstrated that the use of particular codons can increase the expression of a transgene by over 1,000-fold [23].

However, codon bias is often studied among few genes, and always at the genome level. Here we have presented a computational system capable of studying codon bias in a new way. We developed software useful for studying codon bias at mRNA level, which counts each time that a given codon is represented in the transcriptome, thus accounting for the abundance of each mRNA bearing that codon (Fig. 1).

We refer to the total number of codons (n) present across all the mRNAs pool, each expressed at a certain level (x) in a given biological condition as its codonome value ($\mathrm{cv} = \Sigma(n \times x)$ for the mRNAs pool). This is an entirely new concept in genomics, which allows us to determine the consistence of the actual pool of codons physically existent in the mRNA space of a cell, rather than the codon frequency at the level of the gene sequence. The innovative "CODONOME" software is able to calculate these parameters, offering the possibility to test whether there are limits that constrain representation of a codon in the whole transcriptome given its frequency at the genome level (codon bias). We used as reference data input gene expression profiles calculated by integration and normalization of different datasets for a given tissue, following the demonstration that this approach gives a more accurate representation of a reference transcriptome as compared to the use of platform- or experimental-skewed datasets [21]. As expected, the normalized expression profiles show that the genes with the highest expression values are housekeeping genes (see Table 1 and Supplementary Table 1). In addition, the human circulating blood erythrocytes expression profile highlights the preponderance of the most frequently expressed hemoglobin subunits. These findings emphasize the consistence of the reference gene expression profiles we have calculated with the known biology of the considered tissues.

In addition, the "CODONOME" software may show the codons grouped in relation to the aaRS that recognize each group, to explore whether cells organized in such a way optimize the translation process, expressing preferentially aaRS that recognize most frequent codons.

Our findings highlight some new concepts of general relevance about the relationship between the codon bias at genome level and the transcriptome output in term of pool of codons.

First, we demonstrate a surprisingly tight correlation ($r > 0.97$, with the exception of a single case with $r > 0.93$) between the frequency of each codon at genome level (codon bias) and the proportion of that codon in the transcriptome (codonome bias) in different human tissues. This is not trivial because, due to the highly skewed representation of particular gene subsets in various differentiated tissues and to codon bias alteration in singular gene sequence, a more or less relevant loss of correlation could be expected. It seems that a global compensation may exist between codon bias of highly and of poorly expressed genes, even in extremely differentiated tissues with a remarkable expression preponderance of a small number of proteins, as we found in human circulating blood erythrocytes analysis.

Moreover, this high correlation level is maintained across multiple species, from bacteria to humans. This finding clearly implies that the proportional representation of each codon in the DNA and mRNA pool is a general law of nature. It is reasonable to hypothesize that this correlation, resulting from the interaction of the gene number, the skewing of genome codon bias for each gene, and the allowed gene expression value range, allows for a maximal optimization of the transcription and translation processes. Indeed, replacement of actual expression values by random numbers in different ranges shows that the universal law of correlation between codon bias and codonome at a genome scale is not limited to the real gene subsets expressed in nature, but emerges as a general property of the distribution and range of the number, sequence, and expression level of the genes included in a genome. This also implies the important conclusion that there is no constraint, in terms of codon bias, for the global distribution of gene expression values during transcription of a genome.

An additional key finding of this study is the demonstration that the codon bias/codonome correlation is not disrupted by a profound alteration of normal gene expression profile such as may be found in aneuploidy or cancer. We tested the transcriptome of DS-AMKL cells, a condition grouping an aneuplody state with a cancer state, and confirmed the universal value of this correlation.

On the other hand, we found no correlation between aaRS mRNA level expression and their respective recognized codons in the codonome, so it would seem that cells do not use this process to optimize the translation. The explanation may be that aaRS, essential enzymes, are usually in molar excess in the translation machinery and that fine-tuning of their expression in relation to the codonome to be translated is not needed. An alternative explanation could be a tuning of an aaRS expression at translation level of their mRNAs rather than at the transcription level investigated here.

4. Conclusion

In this study we have presented a novel biological concept in genomics, the codonome, indicating the codon pool in the mRNA molecules of a cell. We have also developed a freely available software program, "CODONOME," which is able to calculate the parameters connected to codon bias and codonome concepts. Systematic analysis across multiple tissues, species, and conditions shows that representation of codon bias in the transcriptome (codonome) is tightly linked to the genome bias at codon level, and that codon bias/codonome correlation is a general property of natural genomes.

5. Materials and methods

5.1. Database construction

We developed the "CODONOME" software to parse and integrate RefSeq entries and expression values data and to then calculate how many codons are actually represented in the transcriptome of a given tissue of an organism. We based our software on the FileMaker Pro 10 Advanced (FileMaker, Santa Clara, CA) database management system for both Windows and Macintosh. We made stand-alone software, including the FileMaker runtime with a user guide included, freely available to basic users at http://apollo11.isto.unibo.it/software/.

We investigated the transcriptomes from the following species: *H. sapiens*, *D. rerio*, *C. elegans*, *S. cerevisiae* and *E. coli* in order to obtain data from higher- and lower-vertebrates as well as from invertebrates, unicellular eukaryotes, and prokaryotes.

First, we downloaded the RefSeq mRNA flat files of the desired species from the NCBI ftp site (*H. sapiens* version May 7, 2010; *D. rerio* version June 16, 2010; *C. elegans* and *S. cerevisiae* versions January 18, 2011; *E. coli* version March 1, 2011). Each text file was edited and imported into the appropriate "CODONOME" database table (see the software user guide) to obtain a specific local RefSeq database.

Following the execution of the "CODONOME" command, all but the "NM_" type entries were deleted, thus excluding non-reviewed, predicted mRNA entries (*H. sapiens*: 29,538 NM entries; *D. rerio*: 14,174 NM entries; *C. elegans*: 23,894 NM entries; *S. cerevisiae*: 5,882 NM entries; *E. coli*: 4,319 NM entries). The same script also counted each codon for each mRNA individually, then summed these values to obtain the total number of each codon for the whole mRNAs pool (Fig. 1) and then calculated their per mil frequencies.

We downloaded the expression data files for each species from the GEO web site. The Table 6 and the Supplementary Table 5 list the investigated tissues and organisms and the numbers of considered samples and experiment series. For human brain, we searched for "brain" in GEO datasets, and arbitrarily selected 24 samples from 7 different series in order to integrate representation from different platforms (Affymetrix microarrays types), different authors, and different investigated subjects, thus obtaining an integrated summarized gene expression profile that best represents the general biological transcriptome map for that tissue following both universal assignment of each probe to a specific locus via UniGene data parsing [38] and intra- and inter-sample advanced normalization [21]. We performed a similar process to obtain gene expression profiles for other human tissues, including leukemic cells, as well as for other species (Table 6 and Supplementary Table 5). For *D. rerio* and *C. elegans*, for which fewer studies are available, we chose the platform used in most experiments: GPL1319 and GPL200, respectively.

We processed each expression data file using TRAM software [21]. We performed "Set up" and "Importing the expression data files" software sections according to the software user guide. Then we exported gene symbols with the corresponding normalized expression values in a text file for each investigated species and imported it into the appropriate "CODONOME" database table. "CODONOME" may also accept as input data any data file in text (tab-delimited) format containing two columns separated by a "TAB" key (ASCII9): the official gene symbol and the corresponding numerical (linear) gene expression value, respectively.

Table 6
Samples selected: *Homo sapiens* (pool “A,” “B” and “C”). All Sample IDs and Platform IDs are related to GEO database. Sample type: BM, bone marrow; PB, peripheral blood. Microarray: U133A: Affymetrix Human Genome U133A Array; U95 Version 2: Affymetrix Human Genome U95 Version 2 Array; U95B: Affymetrix Human Genome U95B Array; U95C: Affymetrix Human Genome U95C Array; U95D: Affymetrix Human Genome U95D Array; U95E: Affymetrix Human Genome U95E Array; U133 Plus 2.0: Affymetrix Human Genome U133 Plus 2.0 Array; U133B: Affymetrix Human Genome U133B Array; HG-Focus: Affymetrix Human HG-Focus Target Array.

Study ID	Sample ID	Sample type	Platform	Microarray	Spots	Ref.
Pool “A” — healthy adults ($n = 24$)						
A1…A8 ($n = 8$)	GSM123271…78	Human post-mortem brain tissue	GPL96	U133A	22,283	[24]
A9 ($n = 1$)	GSM44690	Normal brain	GPL96	U133A	22,283	[25]
A10–A11 ($n = 2$)	GSM12688, GSM12708	Normal brain	GPL8300	U95 Version 2	12,625	[26]
A12–A13 ($n = 2$)	GSM12689, GSM12709	Normal brain	GPL92	U95B	12,620	[26]
A14–A15 ($n = 2$)	GSM12690, GSM12710	Normal brain	GPL93	U95C	12,646	[26]
A16–A17 ($n = 2$)	GSM12691, GSM12711	Normal brain	GPL94	U95D	12,644	[26]
A18–A19 ($n = 2$)	GSM12692, GSM12712	Normal brain	GPL95	U95E	12,639	[26]
A20 ($n = 1$)	GSM52556	Normal brain	GPL96	U133A	22,283	[27,28]
A21–A22 ($n = 2$)	GSM76949, GSM76999	Whole brain	GPL570	U133 Plus 2.0	54,675	[29]
A23 ($n = 1$)	GSM136140	Human control brain tissue	GPL96	U133A	22,283	[30]
A24 ($n = 1$)	GSM112030	Brain	GPL570	U133 Plus 2.0	54,675	[31]
Pool “B” — healthy adult ($n = 41$)						
B1…B14 ($n = 14$)	GSM143572…85	Normal human adult red blood cells	GPL96	U133A	22,283	[32]
B15…B28 ($n = 13$)	GSM143671…76, GSM143703, GSM143706…11	Normal human adult red blood cells	GPL97	U133B	22,645	[32]
B29…B35 ($n = 7$)	GSM83897, GSM85205…10	Erythrocytes	GPL201	HG-Focus	8,793	[33]
B36…B41 ($n = 6$)	GSM440234…39	Reticulocytes from adult periperal blood	GPL570	U133 Plus 2.0	54,675	[34]
Pool “C” — DS-AMKL children ($n = 31$)						
C1…C3 ($n = 3$)	GSM491372…4	BM Sorted leukemic blasts	GPL570	U133 Plus 2.0	54,675	[35]
C4…C25 ($n = 22$)	GSM94245, GSM94272…92	BM or PB	GPL96	U133A	22,283	[36]
C26…C31 ($n = 6$)	GSM417985…90	BM or PB Sorted leukemic blasts	GPL570	U133 Plus 2.0	54,675	[37]

5.2. Computational analysis

For each “NM_” mRNA-type entry considered, we counted how many times each codon occurred; we then calculated the count sum of each codon for the whole gene set and the per mil frequency of each codon sum in relation to the sum of all codon for the whole genome gene set (codon bias). We then multiplied the codon count for each gene by the normalized expression value of that gene. Finally, we summed the count of each codon for the whole gene set and the per mil frequency of each codon sum in relation to the sum of all codon for the whole genome gene set (codonome bias). With these values, it is possible to search for relationships between codon usage at genome and at transcriptome level.

To test the requirements for maintaining these relationships, we simulated casual changes in the expression values of real genes in several tests. For the human brain subset we twice permuted the real genes' expression values. We performed another test importing non-normalized expression values exported from TRAM. In the last test, we substituted the actual gene expression values with random numbers from 1 to 10^4, reflecting the order of magnitude of the original dataset, thereby executing a script.

For the human circulating blood erythrocytes subset, we performed another test with random numbers (from 1 to 10^5, bigger than the actual maximum genes expression value) using the random numbers generator at www.randomizer.org, with these parameters: 1 set of 26,589 unique and unsorted numbers per set, from 1 to 10^5. We then exported the created numbers that we manually imported in place of the real expression values in a text file.

Lastly, we created a list of the twenty aaRS with the respective recognized codons for *H. sapiens*, *D. rerio*, *C. elegans*, *S. cerevisiae* and *E. coli*. We then grouped codon and codonome frequencies by aaRS with the relative expression values (using the same expression data file as before; see details in the software documentation).

5.3. Statistical analysis

We exported actual and simulated analyses results in text files and submitted them to statistical analysis using statistical software for Mac OS X (“JMP software” 5.1.2, SAS Institute Inc., Cary, USA). We then analyzed the correlation between paired variables through linear regression. We set the density ellipse at 0.50. In the statistical analysis results “*r*” is the correlation coefficient and “*p*” represents the *p* value. We studied the correlation among the following parameters: a) codon bias, b) codonome bias, c) codon bias grouped by aaRS, d) codonome bias grouped by aaRS and e) the aaRS expression values (“a,” “b,” “c” and “d” are expressed as per mil frequencies).

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2013.02.009.

Acknowledgments

We are profoundly grateful to the Costa family, their friends (in particular the Dal Monte newlyweds), and the community of Dozza (Bologna, Italy), for their generous support to our laboratory's genomics and trisomy 21 research.

References

[1] R. Hershberg, D.A. Petrov, Selection on codon bias, Annu. Rev. Genet. 42 (2008) 287–299.

[2] T. Ikemura, Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system, J. Mol. Biol. 151 (1981) 389–409.

[3] R. Grantham, C. Gautier, M. Gouy, R. Mercier, A. Pavé, Codon catalog usage and the genome hypothesis, Nucleic Acids Res. 8 (1980) r49–r62.

[4] J.B. Plotkin, G. Kudla, Synonymous but not the same: the causes and consequences of codon bias, Nat. Rev. Genet. 12 (2011) 32–42.

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

[6] N. Sueoka, Y. Kawanishi, DNA G + C content of the third codon position and codon usage biases of human genes, Gene 261 (2000) 53–62.

[7] M.A. Gilchrist, P. Shah, R. Zaretzki, Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation, Genetics 183 (2009) 1493–1505.

[8] S. Kanaya, Y. Yamada, Y. Kudo, T. Ikemura, Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis, Gene 238 (1999) 143–155.

[9] S. Kanaya, Y. Yamada, M. Kinouchi, Y. Kudo, T. Ikemura, Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis, J. Mol. Evol. 53 (2001) 290–298.

[10] J.B. Plotkin, H. Robins, A.J. Levine, Tissue-specific codon usage and the expression of human genes, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 12588–12591.

[11] M. Sémon, J.R. Lobry, L. Duret, No evidence for tissue-specific adaptation of synonymous codon usage in humans, Mol. Biol. Evol. 23 (2006) 523–529.

[12] P.M. Sharp, W.H. Li, The codon Adaptation Index — a measure of directional synonymous codon usage bias, and its potential applications, Nucleic Acids Res. 15 (1987) 1281–1295.

[13] P. Puigbò, I.G. Bravo, S. Garcia-Vallvé, E-CAI: a novel server to estimate an expected value of Codon Adaptation Index (eCAI), BMC Bioinformatics 9 (2008) 65–71.

[14] J.L. Bennetzen, B.D. Hall, Codon selection in yeast, J. Biol. Chem. 257 (1982) 3026–3031.

[15] F. Wright, The effective number of codons used in a gene, Gene 87 (1990) 23–29.

[16] J. Peden, Correspondence Analysis of Codon Usage, http://codonw.sourceforge.net/.

[17] A. Grote, Java Codon Adaptation Tool, http://www.jcat.de/Introduction.jsp.

[18] INCA, http://bioinfo.hr/research/inca/.

[19] T. Barrett, R. Edgar, Gene expression omnibus: microarray data storage, submission, retrieval, and analysis, Methods Enzymol. 411 (2006) 352–369.

[20] T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, R.N. Muertter, R. Edgar, NCBI GEO: archive for high-throughput functional genomic data, Nucleic Acids Res. 37 (2009) D885–D890.

[21] L. Lenzi, F. Facchin, F. Piva, M. Giulietti, M.C. Pelleri, F. Frabetti, L. Vitale, R. Casadei, S. Canaider, S. Bortoluzzi, A. Coppe, G.A. Danieli, G. Principato, S. Ferrari, P. Strippoli, TRAM (Transcriptome Mapper): database-driven creation and analysis of transcriptome maps from multiple sources, BMC Genomics 12 (2011) 121–138.

[22] Y. Nakamura, Codon Usage Database, http://www.kazusa.or.jp/codon/.

[23] C. Gustafsson, S. Govindarajan, J. Minshull, Codon bias and heterologous protein expression, Trends Biotechnol. 22 (2004) 346–353.

[24] H.E. Lockstone, L.W. Harris, J.E. Swatton, M.T. Wayland, A.J. Holland, S. Bahn, Gene expression profiling in the adult Down syndrome brain, Genomics 90 (2007) 647–660.

[25] X. Ge, S. Yamamoto, S. Tsutsumi, Y. Midorikawa, S. Ihara, S.M. Wang, H. Aburatani, Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues, Genomics 86 (2005) 127–141.

[26] I. Yanai, H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar, R. Ophir, A. Bar-Even, S. Horn-Saban, M. Safran, E. Domany, D. Lancet, O. Shmueli, Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification, Bioinformatics 21 (2005) 650–659.

[27] K.Y. Detwiller, N.T. Fernando, N.H. Segal, S.W. Ryeom, P.A. D'Amore, S.S. Yoon, Analysis of hypoxia-related gene expression in sarcomas and effect of hypoxia on RNA interference of vascular endothelial cell growth factor A, Cancer Res. 65 (2005) 5881–5889.

[28] S.S. Yoon, N.H. Segal, P.J. Park, K.Y. Detwiller, N.T. Fernando, S.W. Ryeom, M.F. Brennan, S. Singer, Angiogenic profile of soft tissue sarcomas based on analysis of circulating factors and microarray gene expression, J. Surg. Res. 135 (2006) 282–290.

[29] D.K. Nguyen, C.M. Disteche, Dosage compensation of the active X chromosome in mammals, Nat. Genet. 38 (2006) 47–53.

[30] M. Padden, S. Leech, B. Craig, J. Kirk, B. Brankin, S. McQuaid, Differences in expression of junctional adhesion molecule-A and beta-catenin in multiple sclerosis brain tissue: increasing evidence for the role of tight junction pathology, Acta Neuropathol. 113 (2007) 177–186.

[31] H. Auer, D.L. Newsom, K. Kornacker, Expression profiling using Affymetrix GeneChip microarrays, Methods Mol. Biol. 509 (2009) 35–46.

[32] S.H. Goh, M. Josleyn, Y.T. Lee, R.L. Danner, R.B. Gherman, M.C. Cam, J.L. Miller, The human reticulocyte transcriptome, Physiol. Genomics 30 (2007) 172–178.

[33] S. Kabanova, P. Kleinbongard, J. Volkmer, B. Andrée, M. Kelm, T.W. Jax, Gene expression analysis of human red blood cells, Int. J. Med. Sci. 6 (2009) 156–159.

[34] S.J. Noh, S.H. Miller, Y.T. Lee, S.H. Goh, F.M. Marincola, D.F. Stroncek, C. Reed, E. Wang, J.L. Miller, Let-7 microRNAs are developmentally regulated in circulating human erythroid cells, J. Transl. Med. 7 (2009) 98.

[35] J.H. Klusmann, F.J. Godinho, K. Heitmann, A. Maroz, M.L. Koch, D. Reinhardt, S.H. Orkin, Z. Li, Developmental stage-specific interplay of GATA1 and IGF signaling in fetal megakaryopoiesis and leukemogenesis, Genes Dev. 24 (2010) 1659–1672.

[36] J.P. Bourquin, A. Subramanian, C. Langebrake, D. Reinhardt, O. Bernard, P. Ballerini, A. Baruchel, H. Cavé, N. Dastugue, H. Hasle, G.L. Kaspers, M. Lessard, L. Michaux, P. Vyas, E. van Wering, C.M. Zwaan, T.R. Golub, S.H. Orkin, Identification of distinct molecular phenotypes in acute megakaryoblastic leukemia by gene expression profiling, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 3339–3344.

[37] J.H. Klusmann, Z. Li, K. Böhmer, A. Maroz, M.L. Koch, S. Emmrich, F.J. Godinho, S.H. Orkin, D. Reinhardt, miR-125b-2 is a potential oncomiR on human chromosome 21 in megakaryoblastic leukemia, Genes Dev. 24 (2010) 478–490.

[38] L. Lenzi, F. Frabetti, F. Facchin, R. Casadei, L. Vitale, S. Canaider, P. Carinci, M. Zannotti, P. Strippoli, UniGene Tabulator: a full parser for the UniGene format, Bioinformatics 22 (2006) 2570–2571.