COMPARATIVE ANALYSIS OF SYNONYMOUS CODON USAGE BIAS IN HUMAN MONOCYTES, B AND T LYMPHOCYTES BASED ON TRANSCRIPTOME DATA

MUHAMMAD ADIB BIN RUZMAN

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

COMPARATIVE ANALYSIS OF SYNONYMOUS CODON USAGE BIAS IN HUMAN MONOCYTES, B AND T LYMPHOCYTES BASED ON TRANSCRIPTOME DATA

MUHAMMAD ADIB BIN RUZMAN

DISSERTATION SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

INSTITUTE OF BIOLOGICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **MUHAMMAD ADIB BIN RUZMAN**

Matric No: **SGR160003**

Name of Degree: **MASTER OF BIOINFORMATICS**

Title of Thesis:

COMPARATIVE ANALYSIS OF SYNONYMOUS CODON USAGE BIAS IN HUMAN MONOCYTES, B AND T LYMPHOCYTES BASED ON TRANSCRIPTOME DATA

Field of Study: Bioinformatics

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:
Name:
Designation:

Witness's Signature Date:
Name:
Designation:

COMPARATIVE ANALYSIS OF SYNONYMOUS CODON USAGE BIAS IN HUMAN MONOCYTES, B AND T LYMPHOCYTES BASED ON TRANSCRIPTOME DATA

ABSTRACT

Human immune system comprises of many important biological components. Reduction in protein production such as hormones due to changes in codon distribution can lead to immune system disorder. In this study, the balance between mutational bias and translational selection in shaping codon usage bias in protein-coding genes in monocytes, B and T lymphocytes were examined. The protein-coding genes for monocytes, B and T lymphocytes as well as human reference protein-coding genes were obtained from RNA-Seq data from NCBI databases. This study was conducted by computing several codon usage indices and applying multivariate statistical methods. Nucleotide composition analysis showed that the protein-coding genes have GC-rich content and predicted to prefer GC-ended codons to code for the respective amino acids. Relative Synonymous Codon Usage (RSCU) analysis confirms the earlier prediction that GC-rich protein-coding genes will always prefer to use GC-ended codons except for monocytes protein-coding genes which prefer AT-ended codons. The overall codon usage bias was low in all of the cells including human reference protein-coding genes. Multivariate analysis used in this study suggested that codon usage bias is influenced by both mutational bias and translational selection. Moreover, translational selection was identified to be the dominant factor in all the immune cells studied except for monocytes in which it was heavily influenced by mutational bias. This research also provides new insights into human cells biology and contributes new information on advantages of RNA-Seq data in genomic study.

Keywords: Codon usage, codon bias, monocytes

ANALISIS PERBANDINGAN PENGGUNAAN KODON SINONIM DALAM MONOSIT, B DAN T LIMFOSIT MANUSIA BERDASARKAN DATA TRANSKRIPTOM

ABSTRAK

Sistem imun manusia terdiri daripada berbagai komponen biologi penting. Penyusutan penghasilan protin seperti hormon akibat perubahan pada pengagihan kodon akan menyebabkan gangguan sistem imun. Dalam kajian ini, imbangan antara tekanan mutasi dan translasi pilihan dalam membentuk penggunaan kodon bagi gen pengekodan protin manusia bagi monosit, B dan T limfosit telah dkenalpasti. Gen pengekodan protin bagi sel monosit, B, T limfosit dan gen rujukan manusia diperoleh dari pangkalan data NCBI. Kajian ini dijalankan dengan memasukkan beberapa indeks penggunaan kodon dan mengaplikasi kaedah analisis multivarian. Analisis komposisi nukleotida menunjukkan semua gen mempunyai kandungan GC yang tinggi maka ia dijangka akan menggunakan kodon yang diakhiri dengan GC bagi mengekod asid amino masing-masing. Analisis 'Relative Synonymous Codon Usage' (RSCU) mengesahkan jangkaan awal iaitu gen yang kaya dengan GC selalunya akan memilih untuk menggunakan kodon berakhir dengan GC kecuali gen monosit yang lebih memilih kodon diakhiri dengan AT. Keseluruhan penggunaan kodon adalah rendah bagi semua sel imun dan gen rujukan manusia. Analisis multivarian yang dijalankan mencadangkan corak penggunaan kodon adalah dipengaruhi oleh tekanan mutasi dan translasi pilihan. Tambahan pula, translasi pilihan dikenal pasti sebagai faktor utama dalam sel imun terpilih kecuali monosit di mana lebih dipengaruhi oleh tekanan mutasi. Kajian ini juga telah memberi sudut pandang baru ke dalam biologi sel manusia dan menyumbang kepada maklumat terbaru mengenai kelebihan menggunakan data RNA-Seq dalam kajian genomik.

Kata kunci: Penggunaan kodon, Kecenderungan kodon, monosit

ACKNOWLEDGEMENTS

In the name of Allah, the Most Gracious and the Most Merciful. Alhamdulillah, all praises to Allah for the strengths and His blessing in completing this thesis. First of all, I would like to express my gratitude to the University of Malaya for accepting and allowing me to pursue my dream to study here at the postgraduate level. I would also like to thank my supervisors, Prof. Dr. Amir Feisal Merican Bin Aljunid Merican and Dr. Saharuddin Bin Mohamad of the Institute of Biological Sciences, University of Malaya. They were always open whenever I ran into problems or had a question about my research or writing. They consistently guided me and steered me in the right direction.

I must express my very profound gratitude to my father, Ruzman bin Md Noor and my mother, Zalilawati binti Mohd Zain for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Finally, I would like to thank my fellow research assistants for guiding and creating a conducive working environment which allows me to feel comfortable and passionate in doing my work. Thank you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

%	:	Percent
=	:	Equals
<	:	Less than
>	:	Greater than
A	:	Adenine
C	:	Cytosine
G	:	Guanine
T	:	Thymine
Ala/ A	:	Alanine
Arg/ R	:	Arginine
Asn/ N	:	Asparagine
Asp/ D	:	Aspartate
Cys/ C	:	Cysteine
Gln/ Q	:	Gutamine
Glu/ E	:	Glutamate
Gly/ G	:	Glycine
His/ H	:	Histidine
Ile/ I	:	Isoleucine
Leu/ L	:	Leucine
Lys/ K	:	Lysine
Met/ M	:	Methionine
Phe/ F	:	Phenylalanine
Pro/ P	:	Proline
Ser/ S	:	Serine

Thr/ T	:	Threonine
Trp/ W	:	Tryptophan
Tyr/ Y	:	Tyrosine
Val/ V	:	Valine
CAI	:	Codon Adaptation Index
DNA	:	Deoxyribo Nucleic Acid
EMBOSS	:	European Molecular Biology Open Software Suite
ENC	:	Effective Number of Codon
mRNA	:	Messenger Ribo Nucleic Acid
PR2	:	Parity Rule 2
RSCU	:	Relative Synonymous Codon Usage
tRNA	:	Transfer Ribo Nucleic Acid

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Background

A codon is a set of three nucleotides that code for amino acid which is the monomeric unit of proteins (Crick et al., 1961). Protein translation is governed by the genetic code or a set of rules by which DNA or mRNA materials are translated into proteins by living cells. It is inherently redundant with 64 codons but the combinations of three nucleotide bases only able to code for 20 different amino acids. This codon characteristic is known as synonymous codons in which they encode the same amino acid and varies to each other at the third codon position (Bennetzen & Hall, 1982). Knowledge regarding the nature of codon usage bias can provide significant information on involvement of molecular evolution of genes, prediction of genomic behavior and designing cloning vectors for human (Liu et al., 2012). The information is also essential for better understanding of host-pathogen interactions in term of its association to co-evolution or adaptation of pathogens to specific hosts (Pandit & Sinha, 2011).

Codons is translated during protein synthesis as a result of the initial base pairing of cognate tRNA anticodons at the ribosomal A site by a specific tRNA complementary to the amino acid. There are 31 to 46 different tRNA anticodons found across species in which some tRNAs recognised more than one codon. This can be easily carry out through wobble base pairing due to less constraint at the first tRNA anticodon position by non-standard base pairing, thus enable the recognition of multiple third codon positions (Crick, 1966). Modification at the first anticodon position is important in order to enhance the efficiency of wobble base pairing (Holley et al., 1965; Varani & McClain, 2000).

Variation in DNA sequence composition was previously thought to be silent and would not disrupt the polypeptide chain and phenotype characteristics. However, synonymous codons have been shown to use codons in varying frequencies in different organisms (Grantham et al., 1981; Sharp et al., 1988). This phenomenon of preferring a

specific codon is known as codon usage bias and have been characterised as non-random and unique to each species. Codon usage bias is best explained as the result of mutational bias, translational selection and random genetic drift (Bulmer, 1991; Sharp & Li, 1986). Mutational bias occurs due to unequal mutational rates among the nucleotide bases by the influence of several processes and known to be the widespread feature of bacterial genomes (Lobry, 1996; Rocha et al., 2006). Influence of mutational bias in the absence of translational selection can be seen through variation in base composition (Sueoka, 1962). In various organisms, the impact of mutational bias is noticeable in GC content (Muto & Osawa, 1987) and the greatest variation was found in the nucleotide content at the third codon position (Sharp et al., 2005). The changes at the third codon position are often synonymous and accountable to less functional constraint. Thus, it is suggested that heterogeneity in base composition is the greatest source of variation in codon usage bias.

Translational selection operated by preferring codons with the most abundant corresponding tRNAs (Ikemura, 1981; Ikemura, 1985) in order to achieve effective and accurate protein translation (Andersson & Kurland, 1990; Ehrenberg & Kurland, 1984). Variations in GC content has been proven previously as the dominant factor in shaping codon usage bias but some observation suggested that base composition variation is also subjected to huge translation selection influence. The benefit of selecting optimal codons by translational selection on codon usage bias remains inconclusive. Some proposed that codon usage biases are necessary to promote the efficiency of protein synthesis for rapid growth (Andersson & Kurland, 1990; Ehrenberg & Kurland, 1984). Besides that, accurate translation can produce better yield of correctly translated protein products (Bulmer, 1991) as well as reduction in proofreading time (Ehrenberg & Kurland, 1984; Lovmar & Ehrenberg, 2006).

Further comparisons of codon usage bias across species have provided more information regarding the role of translation selection on synonymous sites. In highly

expressed genes, the rates of synonymous substitution are low and expected to be the action of translational selection (Sharp & Li, 1987). However, codon usage bias alone cannot explain for the significance of the observed reduction in divergence (Berg & Martelius 1995; Eyrewalker & Bulmer, 1995). A study investigating the role of translational selection shows a consistent perspective on highly expressed genes, suggesting selection favours codons with the most abundant tRNA in order to achieve accurate and efficient translation species (Ikemura, 1985). The preferred synonymous codons are known as optimal codons and have been identified as the most overrepresented codons in the highly expressed genes (Henry & Sharp, 2007). Identifying optimal codons was the approach used by Sharp and colleagues to investigate the evidence for natural selection on codon bias across polymorphic sites (Sharp et al., 2010). In their study, distribution of optimal codons in *Escherichia coli* and *Clostridium perfringens* tends to skewed or biased towards high frequency variants in highly expressed genes. This observation shows that it is difficult to conclude that the codon usage bias occurs due to mutational bias alone (Sharp et al., 2010). Therefore, many comprehensive studies have provided evidence that the codon usage of highly expressed genes was also subjected to translational selection.

Most of the early study analysing the codon usage bias was targeted on genetic model organism such as *Escherichia coli* and *Saccharomyces cerevisiae* (Bennetzen & Hall, 1982; Grantham et al., 1981) and information regarding tRNA abundances on each species have been characterised (Ikemura, 1981; Ikemura, 1982). From there, codon usage bias study was widely explored in other prokaryotes including *Salmonella enterica* (Sharp & Li, 1987), *Mycoplasma capricolum* (Muto et al., 1985) and *Bacillus subtilis* (Shields & Sharp, 1987), as well as in eukaryotes including *Saccharomyces* (Sharp et al., 1988), *Dictyostelium discoideum* (Sharp & Devine, 1989), *Tetrahymena* (Martindale, 1989) and *Chlamydomonas* (Campbell & Gowri, 1990).

Our current knowledge of codon usage in higher, complex organism such as mammals benefited from the availability of sequence data in large numbers of species. Codon usage has also been analysed in human but further comprehensive analyses are required in order to have a better understanding of codon usage processes and the impacting factors. The impact of codon usage bias in the human genome is less clear (Kotlar & Lavner, 2006). Isochoric structures (Bernardi, 1985) were believed to be the most influential factor shaping codon usage pattern and exhibited strong relationship with gene expression level (Vinogradov, 2003). Thus, studies on human require the balance between background nucleotide composition and expression level in order to reveal the association between gene expression level and codon preference in the human genome. Urrutia et al. (2001, 2003) reported a weak association between gene expression level and codon bias. Furthermore, in genes with high expression, Comeron (2004) showed that for the majority of amino acids with degeneracy of more than one, the codons showed increase in frequency for the most abundant tRNA in highly expressed genes compared to lowly expressed genes. Tissue specificity related to codon usage was also studied, for example Plotkin et al. (2004) proved that codon usage in genes found at specific tissue varied with other tissue specific genes, suggesting that tRNA could act differently in different type of tissues. This finding was later confirmed by Se'mon et al. (2005) in which recognized the variation in codon usage among tissues.

Thus, in this study, the patterns of synonymous codon usage bias on selected human immune cells comprise of monocytes, B and T lymphocytes will be investigated as well as comparing them to human protein-coding genes. Since the datasets were generated from RNA-Seq experiments, this study may also show the importance and benefit of using RNA-Seq data to understand more about codon usage bias on human.

1.2 Hypothesis

Codon usage bias has been identified to occur in every organism or species and have shown to also affect human's biological behavior in protein production. Due to tissue or cell specificity, each human tissue has different codon distribution that may contribute to various degree of codon usage bias. Hence, it is hypothesized that the extent of the codon usage bias in monocytes will be significantly different compared to the other human cells such as B and T lymphocytes.

1.3 Research Questions

There are two research questions based on the literature review conducted:

1) Is the codon usage pattern in protein-coding genes expressed in human monocytes differing to protein-coding genes expressed in B and T lymphocytes?
2) What are the factors that may contribute to codon usage bias in protein coding genes in human monocytes, B and T lymphocytes?

1.4 Objectives

To answer the research questions, two objectives were set up:

1) To analyse and compare the codon usage pattern of protein-coding genes in monocytes, B and T lymphocytes.

2) To investigate the involvement of mutational bias and translational selection of protein-coding genes in monocytes, B and T lymphocytes.

CHAPTER 2: LITERATURE REVIEW

2.1 Codon Degeneracy

The central dogma of molecular biology involved several main processes of protein expression from mRNA transcription to protein translation. During protein translation, information encoded in genetic materials of DNA are governed and regulated by genetic code. The main component of translation is the tRNA that enables a direct and precise communication between a triplet of nucleotides or known as codon and the corresponding amino acid. Ribosomes are the engines of translation that links the tRNA and mRNA (Figure 2.1). The role of genetic code is to determine which amino acid to be synthesised based on transcribed codons. One complicated aspect of the genetic code is the degeneracy of codon in which only 20 standard amino acids can be translated yet there are 64 different codons (61 codons encoding for amino acids and 3 stop codons). In the standard genetic code, there are three amino acids encoded by six codons, one amino acid encoded by three codons, five amino acids encoded by four codons, nine amino acids encoded by two codons and two amino acids encoded by a single codon. This has allowed more than one codon to code for a same amino acid and these different codons that code the same amino acid are called synonymous codons (Crick et al., 1961).

The fundamental molecules of translation are the set of transfer ribonucleotide acids, tRNAs where each accommodate a unique link between a triplet of nucleotides and the corresponding amino acid in the ribosome during translation process. tRNA subset recognised each of the codon with the exception of a few codons that have been reassigned in some lineages (Osawa & Jukes 1989; Osawa et al., 1990). Moreover, the genetic code is remarkably conserved, although it is still in a state of evolution (Osawa et al., 1992).

Table 2.1 shows the standard genetic code for protein synthesis in organisms (Osawa et al., 1992). The DNA codon table consist of 3 amino acids (Arg, Leu and Ser)

encoded by 6 codons, 5 amino acids (Thr, Pro, Ala, Gly and Val) by 4 codons, 1 amino acid (Ile) by 3 codons, 9 amino acids (Lys, Asn, Gln, His, Glu, Asp, Tyr, Cys and Phe) by 2 codons and 2 amino acids (Met and Trp) encoded by 1 codon. For amino acid encoded by more than 1 codon, the first two nucleotide positions are considered critical whereby any mutation or change in base composition at these positions will cause change in amino acid translated. The third codon position is however more flexible or tolerance towards any mutation and commonly known as wobble position (Osawa et al., 1992). If two out of four possible nucleotides at the third codon position can specify an amino acid, this codon is known as a 2-fold degenerate codon. These 2-fold degenerate codons are Lys, Asn, Gln, His, Glu, Asp, Tyr, Cys and Phe in which the equivalent third codon position should always be either two purines (A and G) or two pyrimidines (C and T). The 3 codons that specify Ile are 3-fold degenerate codon. The 5 amino acids (Thr, Pro, Ala, Gly and Val) encoded by 4 codons are known as 4-fold degenerate codons with any nucleotide at its third position of the codon will specify the same amino acid. Lastly, the 6 codons are known as 6-fold degenerate codons (Arg, Leu and Ser) (Osawa et al., 1992).

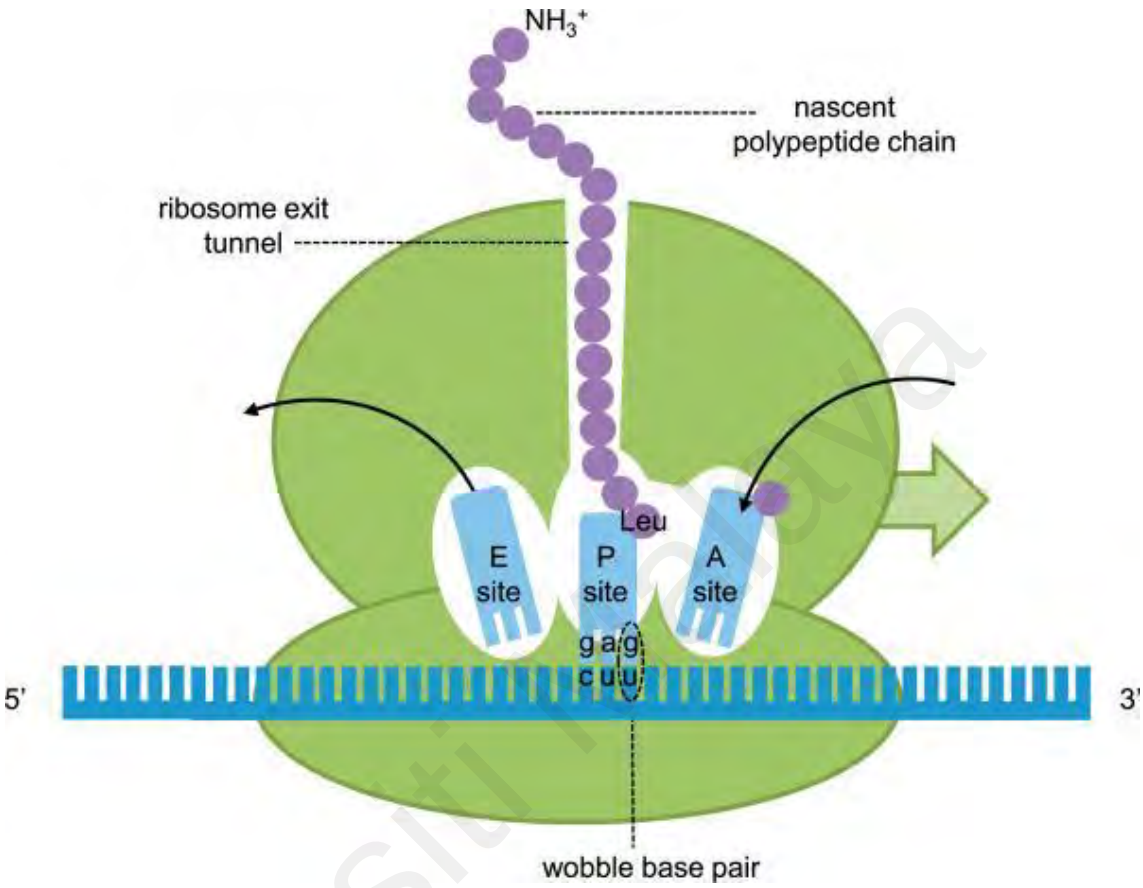

Figure 2.1: Translation in the Ribosome and tRNA Structure. Graphics of the ribosome (green) during translation of a mRNA (blue) with a wobbling codon-anticodon base pair encoding a leucine amino acid. A site, aminoacyl-tRNA site; E site, exit site; P site, peptidyl-tRNA site. Adapted from Quax et al. (2015).

Table 2.1: The standard DNA codon table. The table consist of 18 synonymous codons, 2 non-synonymous codons and 3 stop codons. Adapted from Osawa et al. (1992).

1st base	2nd base								3rd base
	T		C		A		G		
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys	T
	TTC		TCC		TAC		TGC		C
	TTA		TCA		TAA	Stop	TGA	Stop	A
	TTG		TCG		TAG		TGG	Trp	G
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	Gln	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	Lys	AGA	Arg	A
	ATG	Met	ACG		AAG		AGG		G
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	Glu	GGA		A
	GTG		GCG		GAG		GGG		G

2.2 Codon Usage Bias

Due to codon degeneracy, it was believed that all synonymous codons for any amino acid would distribute in approximately equal frequencies in the genome. However, further studies done by Grantham and colleagues (Grantham et al., 1980) suggested that the earlier prediction was incorrect whereby the codon distribution and usage were non-random and species specific (Chen et al., 2004). This occurrence is known as codon usage bias in which each codon is not used equally and some codons were used more frequently than the others (Behura & Severson, 2013). They found that codon choice of genes of the similar genomes have a specific system for selecting between codons (Grantham et al., 1980). Grantham et al. utilize mRNAs from a variety of species comprise of prokaryotes and eukaryotes and it was observed that degeneracy was detected at the third codon position (Grantham et al., 1980). Correspondence analysis from the same study on the mRNA displayed a clustering pattern arranged based on type of genome, suggesting that variation in the mRNA was genome specific (Grantham et al., 1980). The purpose of conducting genome-wide codon usage bias experiments is to investigate the specific codon preference pattern of each organism as well as their impact and consequences. This is also important to determine the factors that can construct their behavior and action in accordance to genome biology. In applied science aspects, analysis of codon usage bias can improve the knowledge of heterologous gene expression. This technique is a dynamic methodology in biotechnological area that produce recombinant products such as insulin, antibiotics, and vaccines.

2.3 Factors of Codon Usage Preference

Studies had been conducted in a wide variety of organisms in both prokaryotes and eukaryotes to identify the synonymous codon usage pattern (Gu et al., 2004; Lavner et al., 2005; Ahn et al., 2009; Liu et al., 2012), and the factors influencing the codon usage variation have been recognised since. Genomic factors such as GC content, gene expression level, gene length, recombination rate and genetic code modulation are related to codon usage bias in different organisms (Urrutia et al., 2003; Roymondal et al., 2009; Palidwor et al., 2010). From the factors identified, translational selection and mutational bias were determined to be the major factors in shaping codon usage pattern (Duret et al., 1999; Xu et al., 2011; Nair et al., 2013).

2.3.1 Mutational Bias

Codon usage bias happens because of non-randomness in the mutational patterns. Mutation occurs in some codons and thus would have lowered the equilibrium frequencies. The extent of mutational bias was previously believed to be varied between species in which may be the reason behind the differences in codon usage pattern across organisms (Behura & Severson, 2013). Mutational bias took place due to unequal mutational rates of the nucleotide bases as a result of numerous variation mechanisms (Lobry, 1996; Rocha et al., 2006). Mutational bias have been investigated in several studies and have been found to be the most important factor in shaping codon bias variation between different organisms (Marais et al., 2001; Palidwor et al., 2010). In the absence of translational selection, mutational bias can be displayed as the variation of base compositions (Sueoka, 1962) specifically the variation in GC content (Muto & Osawa, 1987) and the greatest variation in base composition between organisms can be found at the third codon position of GC content (Sharp et al., 2005). Vertebrates such as mammals are constrained strongly by nucleotide composition as the result of mutational bias (Bernardi, 1995). In agreement with other organisms, it was found that the major

compositional properties responsible for shaping codon usage bias in vertebrates was GC content at the third codon position (Urrutia et al., 2001; Rao et al., 2011).

Several multivariate analyses among organisms have identified the correlation between mutational bias towards codon usage bias (Kanaya et al., 2001; Knight et al., 2001; Chen et al., 2004; Lobry & Necsulea, 2006; Lynn et al., 2002) and have shown GC content to be the most significant parameters in shaping codon usage pattern. GC content is believed to be heavily related to genome-wide processes rather than by selective forces acting specifically on coding regions. Analysis on GC content can be a good measurement for detecting the degree of codon usage bias even in low number of genes, for example as few as 10 genes (Zavala et al., 2005). It is also found that 90% of GC content variations are observed between genomes rather than within genomes and thus make intergenic GC content suitable for codon usage bias prediction (Chen et al., 2004).

2.3.2 Translational Selection

Translational selection is believed to optimize the interaction of codons and tRNA abundances towards correspond anticodons (Figure 2.2) (Ikemura, 1985). The codon usage optimization was thought to be due to three main factors which are speed of translation elongation, accuracy of translation and cost of proofreading. All of these three factors can be minimized by the action of translational selection on the genome. By selection, gene with high levels of expression have a strong bias towards a specific set of codons while gene with low levels of expression showed equal or nearly equal usage of synonymous codon (Gouy & Gautier, 1982). It has been suggested that the optimal codons are mostly complementary to the most abundant tRNA species in the cell and perfect Crick-Watson base pairing can be established to enhance translational accuracy (Gouy & Gautier, 1982; Ikemura, 1981). The use of optimal codon has lowered the translational misincorporation rates by strongly restricted codon usage bias at the codons

site in which mistakes during translation could result in the synthesis of a costly dysfunctional peptide (Akashi, 1994).

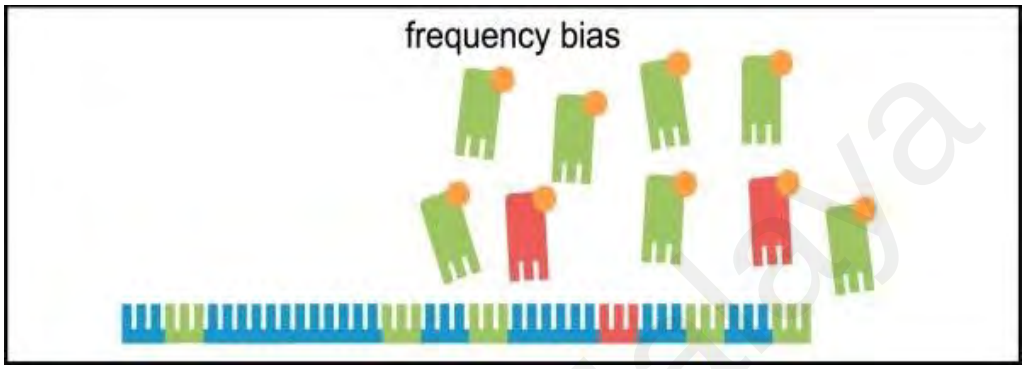

Figure 2.2: Frequency bias will result in effective protein production when the frequency of used codons matches the cellular tRNA population. Adapted from Quax et al. (2015).

Translational selection on slow growing organisms such as vertebrates is inconclusive and requires more comprehensive study to uncover the extent of translational selection on these organisms (Duret, 2002). Recent evidence indicated that only a weak translational selection acted upon the vertebrates in which the codon usage patterns were influenced strongly by compositional constraint (Qiu et al., 2011; Musto et al., 2001; Romero et al., 2003; Yang et al., 2008). In human, it is shown that there is a correlation between gene expression levels towards translational selection (Urrutia et al., 2003; Yang et al., 2008). However, there is a lack of study on the correlation of codon bias to tRNA gene copy number in vertebrates especially human whereby the availability of tRNA count can clearly show the role of translational selection on codon usage pattern.

Variation in GC content has been identified previously as the dominant factor in shaping codon usage bias but some observation suggested that base composition is also subjected to translational selection. Other isolated factors such as thermostability can be also related to translational selection. GC nucleotides are preferred in several organisms living in hot climate due to thermostability in the GC base pairing with evidence of positive correlations between GC content towards optimal growth temperature (Argos et al., 1979; Musto et al., 2004). By utilising multivariate analyses, association between GC content and growth temperature were further justified (Lobry & Chessel, 2003; Lobry & Necsulea, 2006; Lynn et al., 2002). This observation occurs due to the adaptation of nucleotide composition towards increasing temperature but this interpretation is not entirely accurate in which the correlation was suggested to be the influence of atypical codon usage of amino acids arginine in the genome (Lobry & Necsulea, 2006). However, another similar study on 764 different species was unable to identify any correlation between GC content to optimal growth temperature (Galtier & Lobry, 1997).

In *Escherichia coli*, highly expressed genes showed high codon occurrence for the 4-fold degenerate amino acids glycine and proline (Sharp & Li, 1987). Coincidently,

both amino acids showed distinct variation at the third codon position, thus this condition could not be the product of a simple mutational bias. Moreover, another study on 600 gene sequences from several species showed that codon usage pattern was varied depend on the gene expression levels (Gouy & Gautier, 1982). Since that discovery, many more intragenomic correlations of codon usage bias to gene expression level have been observed for a variety of prokaryotic species including *Escherichia coli* (Ikemura, 1985), *Streptococcus pneumoniae* (Martin-Galiano et al., 2004) and *Lactobacillus lactis* (Dressaire et al., 2009). The patterns observed in highly expressed genes are specific and unique to each species and sometimes may even oppose to the direction of existing mutation pressure. For example in AT-rich *Clostridium perfringens*, T-ended codons at the 2-fold degenerate sites are the highest across genes, however C-ended codons are the one that dominate in the highly expressed genes (Musto et al., 2003). Thus, these complex patterns have revealed the involvement of other factors rather than mutational bias alone in shaping codon usage bias.

Moreover, overrepresented codons in highly expressed genes are complementary to the tRNA species and intracellular abundances in *Escherichia coli* (Ikemura, 1981) and *Bacillus subtilis* (Kanaya et al., 1999). In *Escherichia coli*, of the four tRNAs detected for leucine, tRNA containing CAG anticodon have the highest abundances and perfectly complements to the CUG codon in which are the most frequently used among highly expressed genes (Ikemura, 1985). However in some cases, the relationship between the most abundant tRNAs and complementary codons can vary between species. For example in *Bacillus subtilis,* tRNA with the UAA anticodon was found to be the most abundant for the amino acid leucine but UAA was not the codon preferred by the highly expressed genes (Kanaya et al., 1999).

2.4 Codon Usage in Human

In the human genome, the roles of isochoric structure on codon usage bias is inconclusive and have shown to have influence on codon composition as well as expression levels (Bernardi, 1995; Vinogradov, 2003). Weak correlation between gene expression levels and codon usage bias was discovered but there was no correlation towards isoacceptor tRNA abundance (Urrutia et al., 2003; Urrutia et al., 2001). Moreover, previous study has found indication that the codon preference in mammals was related to mRNA secondary structure stability (Chamary et al., 2005). In highly expressed genes, codons with the most abundant isoacceptor tRNA gene copy numbers showed an increase in frequency compared to the lowly expressed genes (Comeron, 2004). Furthermore, codon usage bias was examined in tissue specific genes and and variation have been observed between genes from different tissues. This observation may be due to the influence of differential tRNA abundances (Plotkin et al., 2004). The variation of the codon usage was confirmed by Se'mon et al. (2005) by using internal correspondence analysis and have shown that the variation of synonymous codon usage previously discovered between tissues are non-representative and only represents 2.3% of the total variation, and that most of this is explained by variability of GC-content that affects both coding and intergenic regions (Se'mon et al., 2005).

Several evidences of splicing enhancers involvement in translation selection have also been found which eventually lead to codon usage bias (Willie et al., 2004; Chamary et al., 2005; Fairbrother et al., 2004; Parmley et al., 2005). These studies showed that codon preferred are frequently found in exonic splicing enhancers (Fairbrother et al., 2004) and further support the enhancer model (Willie et al., 2004; Chamary et al., 2005). Furthermore, a study showed that codon usage bias is at the highest levels in both highly and lowly expressed genes and the frequency of optimal codons (FOP) increase with gene expression levels (Lavner et al.,2005). From these findings, they suggested two

alternatives on how translational selection may influence codon usage bias. Firstly, translational selection may regulate the expression of lowly expressed genes by choosing codons with less abundant tRNAs and secondly, translational selection may as well improve translation accuracy by using optimal codons in genes with high concentration of amino acids.

2.5 Human Immune System and the Immune Related Genes

The immune system is a complex system essential for protection against infections from pathogenic and non-pathogenic microbes. Human immune system consists of three lines of defense. The first line of defense act as a barrier to keep out foreign substances through skin and mucus membranes. The second line of defense consists of general or alternative ways to defend against pathogens that have broken through the first line of defense via inflammatory response or fever (Janeway, 2001; Mogensen et al., 2009). Both first and second line of defense are also known as innate immunity. The innate immune response reacts quickly to any foreign agent via recognition mediated by host molecules through (PPRs) Pattern Recognition Receptors. PRRs are expressed on the surface of the cells and in intracellular compartments or secreted into the blood stream and tissue fluids (Abbas et al., 2012; Tizard, 2013) and recognized PAMPs (Pathogen-associated Molecular Patterns) in the host. A group of PRRs called TLRs (Toll-like Receptors) have an important role in recognize wide range of PAMPs, leading to activation of the immune responses (Kuby et al., 2007). TLRs binds to PAMPs and send signals to the intracellular environment via adapter proteins such as TRIF (Toll like Receptor-domain-containing adapter-inducing interferon-β) (Yamamoto, 2003) which activate both NFκB (Nuclear Factor Kappa B Subunit) signaling and MAP kinase pathway as well as pro-inflammatory cytokines secretion (Janeway & Medzhitov, 2002; Piras & Selvarajoo, 2014; Tizard, 2013) (Figure 2.1). For detection of intracellular PAMPs during virus infection, cytosolic PPRs such as NLRs (NOD-like Receptors;

Nucleotide-Binding Oligomerization Domain-like Receptors) and RLRs (RIG-I-like Receptors; Retinoic Acid-Inducible Gene-I-like Receptors) would be activated (Kawai & Akira, 2009). The activation of cytosolic PPRs induces the production of IFNs (Interferons) and proinflammatory cytokines (Gack, 2014; Weber et al., 2013) (Figure 2.1).

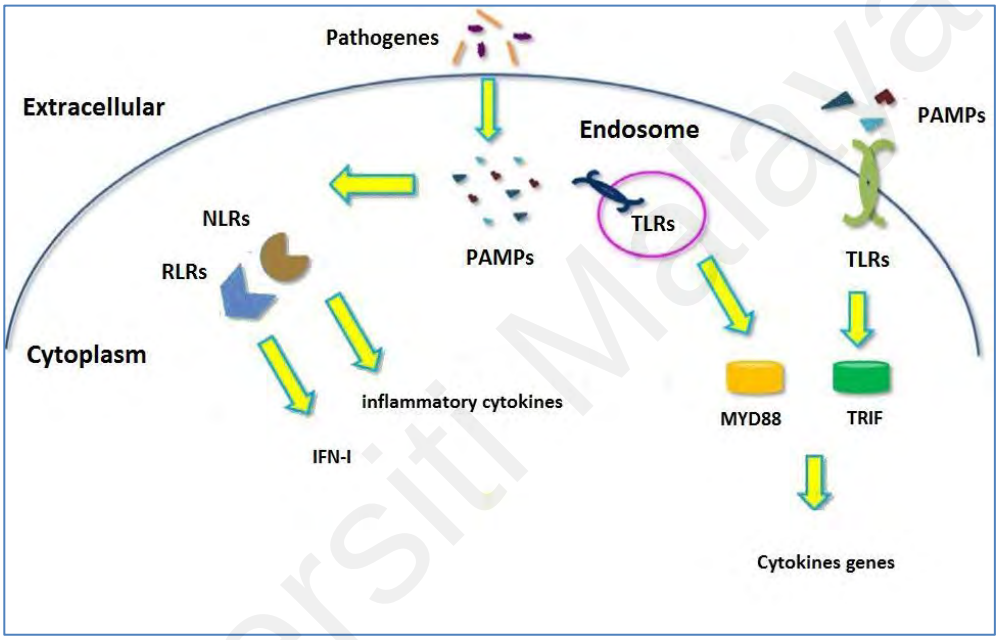

Figure 2.3: Three classes of pattern recognition receptors: (TLRs, RLRs, and NLRs) with their roles in inducing host antiviral responses. Adapted from Quax et al. (2015).

The third line of defense or adaptive immunity activated through specific cells or molecules as listed in Table 2.2 in order to eliminate the disease-causing pathogens. This cell such as B lymphocytes acts by producing antibodies in the extracellular fluid, while T lymphocytes recognize and kill cells that have become infected. The immune system is closely associated to the lymphatic system, with B and T lymphocytes being found primarily within the lymph nodes (Chaplin, 2010; Danilova, 2012).

Monocytes are the precursor cells of differentiated macrophages. Monocytes make up only 1-6% of the white blood cells. Once out of the blood stream, monocytes enlarged and differentiated into macrophages. Some macrophages protect the tissues by moving along via amoeboid motion while others remain settled in one place. Macrophages are the primary scavenger cells in removing larger particles such as dead erythrocytes and neutrophils. Phagocytic macrophages play an important role in the development of adaptive immunity in which they engulf molecular or cellular antigens while the fragments of processed antigen are inserted into its own membrane as part of surface protein complexes (Ginhoux et al., 2014). Lymphocytes are the main cells that regulate and control the adaptive immune response of the body even though only 5% of lymphocytes are found in circulation or 20-30% of the white blood cells. Most lymphocytes are found in lymphoid tissues, where they are more likely to encounter foreign invaders. By one estimate, the adult body contains a trillion lymphocytes at one time (Alberts et al., 2002).

Physiological studies on immune cells have been conducted extensively and at the molecular level, reduction in protein production such as hormones can promote the development of immune system disorder such as Systemic Lupus Erythematosus (Fimmel et al., 2005). Furthermore, codon usage bias has displayed their influence on Toll-like receptor genes of monocytes in which change in codon preference can alter the normal expression level (Zhong et al., 2005; Newman et al., 2016). Unfortunately, until now there

is no report on codon usage patterns in human immune cells. With the knowledge of codon preference of the cells, further studies can be conducted on altering codon distribution and protein production for normal human immunological function. Identification of codon preference in any particular cell can later be used for techniques such as codon-optimization of mRNAs in pharmaceuticals and nucleic acid therapies (Mauro et al., 2014).

Table 2.2: Components of human immune system. Adapted from Benito-Martin et al. (2015).

Immunity	**Cells**	**Molecules**
Innate	Monocytes Macrophages Dendritic cells Natural killer (NK) cells Neutrophils Mast cells Basophils Eosinophils	Cytokines Chemokines Complement
Adaptive	T cells: Helper T (CD4+) Killer T (CD8+) Memory T Suppressor T B cells	Cytokines Antibodies

2.6 Codon Optimization

The amino acids can be encoded by more than one codon due to the degenerate nature of the genetic code as discussed before. Therefore, change in codon preference may alter drastically the amount of protein expressed (Welch, 2009; Kudla, 2009). Codon-optimized mRNA sequences can be performed via various methods based on codon usage profile. The idea of altering codon distribution has accelerated the use of codon-optimized mRNAs for protein production in pharmaceuticals therapies. However, substantial evidence shows that changing synonymous codon preferences in natural mRNAs could have unanticipated results. This changes ranges from change in protein stability and conformation or change in protein function and post-translational modifications (Ward, 2011; Shabalina, 2013). The worst-case scenario is that synonymous mutations may lead to the development of various diseases (Shabalina, 2013; Hunt, 2009; Chen, 2010). Besides that, several risks have been identified as a result of codon-optimization includes production of anti-drug antibodies which can cause allergic reactions and also may lowered the drug efficacy (Katsnelson, 2011; Sauna, 2011; Kimchi, 2013; USFDA, 2003).

2.7 Use of RNA-Seq Technology for Codon Usage Bias Study

Advancement in next generation sequencing technology have shifted sequencing techniques from microarrays to whole-transcriptome shotgun sequencing or RNA sequencing (RNA-Seq) for RNA quantification (Nagalakshmi et al., 2008; Mortazavi et al., 2008; Marioni et al., 2008). In contrast to other RNA quantification methods, RNA-Seq does not rely on a preselected set of transcripts to assay and have proven to yield high quality results. This sequencing method is also highly replicable with very low noise and is sensitive to transcripts present at low concentration. Even though with the understanding of variation of codon usage bias among different species, the study only focuses on lower, less complex organisms such as *Escherichia coli* and *Saccharomyces*

cerevisiae. RNA-Seq has made it possible to gather a large amount of sequence data which can provide a new direction of codon usage study by using non-model organisms (Plotkin et al., 2011; Novoa et al., 2012; Shabalina, 2012; Hershberg et al., 2008). Through RNA sequencing, faster and robust identification of mRNA and tRNA abundance, proteomics and ribosome density profiling could finally be realised (Quax et al., 2015). Therefore, the opportunity to discover and better understand the human codon usage bias can be accomplished using RNA-Seq data. Codon usage bias study in human is important in order to learn the adaptation and evolutionary process occur in human genome.

2.8 Codon Usage Indices

Ever since the establishment of correlation between codon usage and tRNA abundance (Ikemura, 1981), comprehensive studies on codon–anticodon adaptation have been conducted (Bulmer, 1987; Bulmer, 1991; Xia, 1998; Xia, 2008; Higgs & Ran, 2008; Jia & Higgs, 2008; Palidwor, 2010). This has led to the development of several analytical tests using alternative theoretical predictions (Xia, 1996; Xia, 2005; Carullo & Xia, 2008; van Weringh, 2011), as well as the establishment of widely used codon usage indices (Sharp & Li, 1987; Wright, 1990; Xia, 2007). The most accepted index is Codon Adaptation Index, CAI (Sharp & Li, 1987; Xia, 2007) which has shown to give a better insight of codon usage bias. By using CAI, new findings such as positive correlation between codon usage bias towards splicing strength of yeast in intronic region and translation elongation efficiency have been identified (Ma & Xia 2011).

Among the codon usage indices proposed, they are often identified as codon specific or gene specific. A representative of the codon specific type of codon usage index is the Relative Synonymous Codon Usage (Sharp, 1986), and representatives of the gene specific type are the Codon Adaptation Index, CAI (Sharp & Li, 1987; Xia, 2007), Effective Number of Codons, ENC (Wright, 1990), the Frequency of Optimal Codons,

FOP (Ikemura, 1981) and the Codon Bias Index, CBI (Bennetzen & Hall, 1982). CAI have been suggested as the most effective and reliable measurement in predicting gene expression levels based on several comprehensive studies (Comeron & Aguade, 1998; Duret & Mouchiroud, 1999; Coghlan & Wolfe, 2000), but ENC has an advantage over CAI and other codon usage indices in which ENC only require codon frequencies data of the genes studied. Meanwhile, utilizing other index require external information which is often unavailable. For example, CAI requires a reference set of highly expressed genes for particular species, FOP and CBI needs information on relative tRNA abundance and gene expression levels. Extra information on tRNA have not significantly helpful in identifying codon usage bias in genes according to several studies in which have suggested that there is a low or no influence of tRNA abundance on the level of gene expressed (Rocha, 2004). Hence, ENC has been frequently used in biological research particularly in codon usage study. However, several problems arise when using ENC especially due to computer specifications whereby it can slow down performance and limiting its usage.

2.9 Online Bioinformatics Tools for Analysis

Codon usage analysis requires tools that have the ability to examined codon usage factors such as expression level. Highly expressed genes have been identified to have strong codon usage bias for example in *Escherichia coli, Saccharomyces cerevisiae* and *Bacillus subtilis* (Sharp & Cowe, 1991; Sharp & Li, 1987; Shields & Sharp, 1987). However, the opposite observations suggesting lowly expressed genes tend to have low codon usage bias are still unclear and inconclusive. (Fitch & Strausbaugh, 1993; Kliman & Hey, 1993; Kliman & Hey, 1994).

In many cases, the limiting factor preventing a comprehensive analysis of codon usage is often lack of sequence information. A fundamental part of any analysis conducted is to ensure that the sample size is large enough to represent the target

population. The sample size also needs to fulfill the magnitude of the selective coefficient and the effective population size. The tedious part of selecting the sample size is that the factors involve is often unpredictable and unknown and must be empirically done.

Online tools to measure codon usage bias are often in early stage of development and usually have various utilities and limitations (Peden, 2000). Therefore, selecting established software is important in order to yield a good result. Online tools such as CodonW and EMBOSS cusp have been used in many codon usage bias study and several algorithms and formulas have been created to describe the codon bias indices. There is only a few software available to assist the analysis of codon usage bias. Some simple program can tabulate codon usage for particular species directly from the available public nucleic acid databases such as GenBank (Nakamura, 1996). Moreover, some programs could also measure specific codon usage indices such ENC (Wright, 1990), CBI (Bennetzen & Hall, 1982) and several other general codon usage indices (Goldman et al., 1995; Krishnaswamy & Shanmugasundaram, 1995; Rodriguezbelmonte et al., 1996). Each program has their own unique way to calculate codon usage using multiple input formats and allows user to choose any of the index required.

2.9.1 EMBOSS cusp

An example of basic program available online is EMBOSS cusp. This program was designed ultimately to create codon usage table based on input nucleotide sequences. The function "cusp" creates and tabulate a codon usage table based on one or more nucleotide coding sequences and writes the table to the designated file. Information generated on the codon usage table includes sequence of each codon and respective amino acids. From there, the proportion of usage among synonymous codons can be quantified as well as the calculation of expected number of codons per 1000 bases of the input sequences.

2.9.2 CodonW

Another example of free online tool software for measuring codon usage indices is the CodonW which have been used widely in codon usage studies. The purpose of using this software is to simplify the analysis of codon usage by incorporate codon usage indices with multivariate statistical analysis in a single program. This software designed to be simple to use and portable in terms of both operating systems and machine architecture (Peden, 2000). Since the advancement of sequencing technology and the rapid increase of available sequence databases, CodonW was also created to accommodate unlimited number of input sequences or sequence length that can be included in the statistical analysis. Universal genetic codes as well as other alternative genetic codes are also integrated in this program allowing much more studies on variety of species can be established. CodonW generates most of the output in a tabulate form using a specific command line. One particular disadvantage of this program is that there is no built-in graphics to view results in the form of plots and graphics (Peden, 2000). Therefore, other programs designed to work specifically with numerical data such as Excel, R Studio and Minitab are required.

CHAPTER 3: MATERIALS AND METHODOLOGY

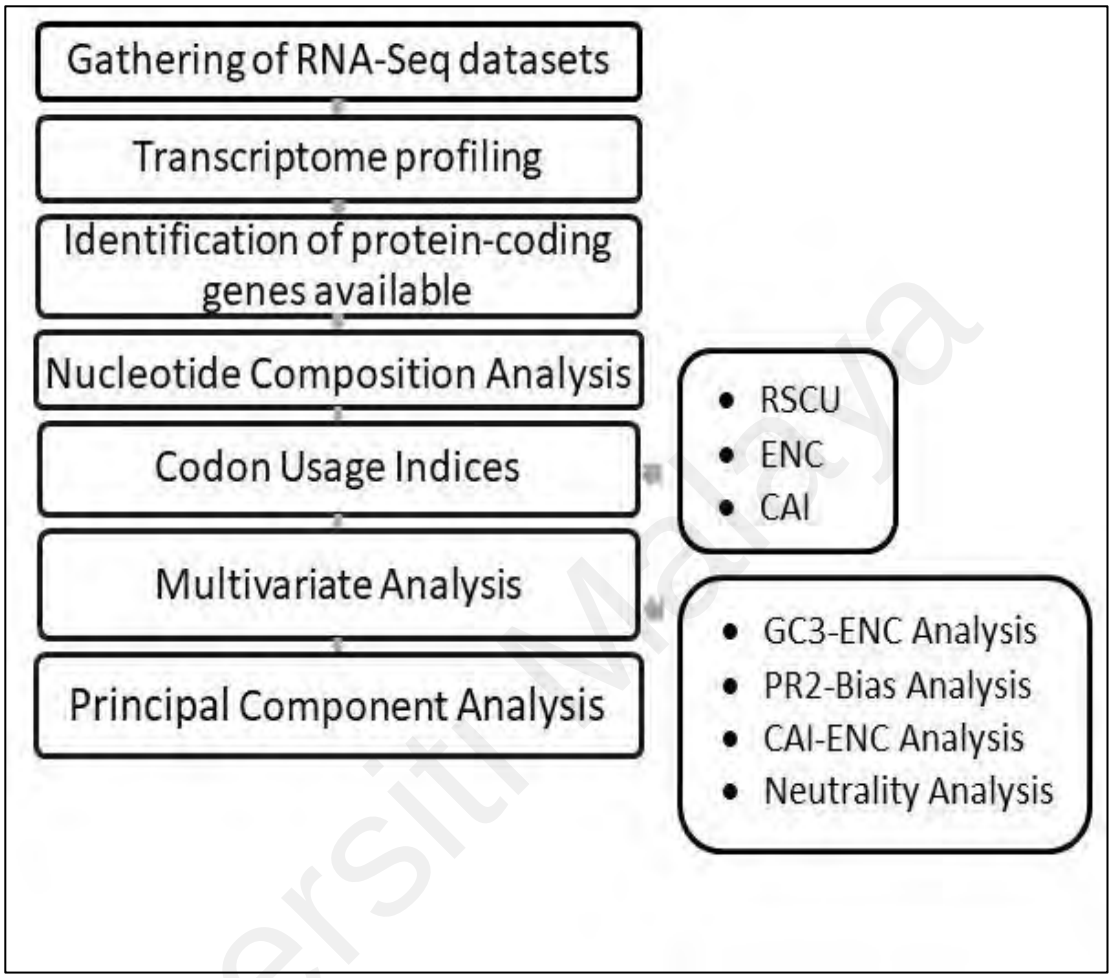

Figure 3.1: Schematic representation of workflow from gathering datasets to analyses involved in the study.

3.1 RNA-Seq datasets

Protein-coding genes sequences analysed in this study are extracted from transcriptomic datasets of monocytes, B and T lymphocytes cells. All of the transcriptome data have been sequenced using next generation sequencing technologies and the datasets are obtained from public sequence database (www.ncbi.nlm.nih.gov/). Monocytes, B lymphocytes and T lymphocytes profiling data are downloaded from Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) under accession number GSE80095, GSE29158 and GSE85527 respectively. Human protein-coding sequences data from GENCODE Release 22 (www.gencodegenes.org) are also examined and used as a reference. The source of the data used, respective accession number are listed in Table 3.1. From the RNA-Seq data obtained, genome alignment, transcript assembly and gene expression profiling are performed to identify the protein-coding genes present in monocytes, B and T lymphocytes as well as human protein-coding genes. The genes selected in this research are only selected from high-throughput RNA-Seq data. Only protein-coding genes are selected from the list of genes identified with the exclusion of mitochondrial genes, pseudogenes and novel genes.

Table 3.1: Source of RNA-Seq data

Data	Source	Accession Number/ Version
Monocytes	www.ncbi.nlm.nih.gov/geo/	GSE80095
B lymphocytes	www.ncbi.nlm.nih.gov/geo/	GSE29158
T lymphocytes	www.ncbi.nlm.nih.gov/geo/	GSE85527
Human Protein-Coding Genes	https://www.gencodegenes.org	GENCODE Release 22

3.2 Transcriptome profiling of the datasets

3.2.1 Genome alignment and transcript assembly

The quality of all sequences reads from monocytes, B and T lymphocytes and human reference are reviewed using FastQC (https://www.bioinformatics.babraham.ac.uk/) (Andrews, 2010). The adaptors and low-quality bases are trimmed from the sequences using Trimmomatic (http://www.usadellab.org/cms/) (Bolger et al., 2014). For each data, the standard score at base across reads is set at Q > 20. Trimmed raw reads are then aligned to the human reference genome sequence (GRCh38.79) using HISAT version 0.1.4 (https://ccb.jhu.edu/software/hisat/) (Kim et al., 2015) with GENCODE version 22 is used as a guided reference annotation. The aligned reads for each data are assembled into transcripts by StringTie version 1.3.3 (https://ccb.jhu.edu/software/stringtie) (Pertea et al., 2015) using a GENCODE reference annotation GTF file version 22 and separate GTF files are generated for each of the samples. The transcripts abundance is estimated as Fragments Per Kilobase of exon per Million fragments mapped (FPKM) (Trapnell et al., 2010).

3.2.2 Gene expression profiling

For detecting gene expression pattern in monocytes, B and T lymphocytes and human genome, transcript assemblies (GTF files) are merged together to form a single set of non-redundant transcripts using Cuffmerge (a part of cufflinks, version 2.2.1) (http://cole-trapnell-lab.github.io/cufflinks/) (Trapnell et al., 2010). Cuffquant (a part of cufflinks, version 2.2.1) (Trapnell et al., 2010) is used to quantify the expression levels of transcripts and to create individual binary files (CXB format). The Cuffnorm (a part of cufflinks, version 2.2.1) (Trapnell et al., 2010) is used to normalize FPKM between the samples. The FPKM > 0.1 threshold is set to determine expressed transcripts. The merged assembly is then compared with a GENCODE reference annotation GTF file version 22,

which contains protein-coding genes, non-coding genes, pseudogenes and alternative transcribed variants. From the comparative analysis results, intergenic transcripts are considered as putatively novel transcripts. These transcripts are filtered against the non-redundant database from NCBI using Basic Local Alignment Search Tool for Nucleotide (BLASTN version 2.4.0) (https://www.ncbi.nlm.nih.gov/BLAST/) with threshold of e-value < 1e-10.

3.3 Nucleotide composition properties analysis

The following nucleotide composition properties are calculated for the protein-coding genes: i) Overall nucleotide compositions (A%, T%, G% and C%), ii) Nucleotide composition at the third codon position (A3%, T3%, C3%, and G3%), iii) Frequencies of nucleotide combination (GC%, AT%, AT3% and GC3%) for each set of genes (Singer, 2000). These calculations are determined using an in-house Perl scripts. GC% content is the percentage of G+C frequency in a coding gene while GC3% contents are the percentage frequency of G+C at the third positions of codons. A3%, T3%, G3% and C3% contents are the percentage frequencies of A, T, G and C at the synonymous third position of codons, respectively (Singer, 2000).

3.4 Measurement Indices of Codon Usage Bias

Codon usage indices are a group of index that is used to measure the degree of codon usage bias in organism (Peden, 2000). RSCU value for all datasets are determined using EMBOSS cusp software version 6.6.0.0. Meanwhile, CAI and ENC values are determined and analysed using CodonW version 1.4.4 software (http://codonw.sourceforge.net/). The CodonW is designed to work with genetic code and it is useful in determining the codon usage pattern, the number of codon present in the coding sequences and calculating the amount of individual nucleotide (Peden, 2000).

3.4.1 Relative Synonymous Codon Usage (RSCU) analysis

The RSCU values for all of the coding sequences are calculated to determine the characteristics of synonymous codon usage without the influence of amino acid composition and genome size by the different gene samples following a previously established method (Sharp, 1986). RSCU values are the ratio between the observed usage frequency of one codon in a gene sample and the expected usage frequency in the synonymous codon family, given that all codons for the particular amino acid are used equally. The synonymous codons with RSCU values > 1.0 have positive codon usage bias and are defined as abundant codons, whereas those with RSCU values < 1.0 have negative codon usage bias and are defined as less abundant codons. When the RSCU value equal to 1.0, it means there is no codon usage bias for that amino acid and the codons are chosen equally or randomly (Sharp et al., 1986). Moreover, the synonymous codons with RSCU values > 1.6 and < 0.6 are treated as over-represented and under-represented codons, respectively (Wong et al., 2010). RSCU value is calculated as follows:

$$RSCU = \frac{g_{ab}}{\sum_{b}^{n_a} g_{ab}} n_a$$

where, g_{ab} is the observed number of the a^{th} codon for the b^{th} amino acid which has n_a kinds of synonymous codons.

3.4.2 Codon Adaptation Index (CAI) analysis

Codon Adaptation Index (CAI) analysis is a quantitative method that predicts the expression level of a gene and estimate the degree of codon biasness based on coding sequences (Sharp, 1987). CAI analysis is the most widely used codon usage index due to its reliability in measuring expression level of genes. CAI values range from 0 represent low expression level to 1 represent high expression level. A higher CAI value suggests and reflects a stronger codon usage bias in a gene and suggested to be preferred over the genes with lower CAI value. The most frequent codons have the highest relative adaptiveness and sequences with higher CAI values are suggested to be preferred over those with lower CAI (Sharp, 1987). The synonymous codon usage patterns of *Homo sapiens* are used as references. Non-synonymous codons and termination codons are excluded from the calculation. The reference datasets for *Homo sapiens* was obtained from the Codon Usage Database (Nakamura, 2000). The CAI is estimated using the equation given by Sharp and Li (1986) as follows:

$$CAI = \exp \frac{1}{L} \sum_{k=1}^{L} \ln \omega_{c(k)}$$

where, L is the number of codons in the gene and $\omega_{c(k)}$ is the ω (relative adaptiveness) value for the k^{th} codon in the gene.

3.4.3 Effective Number of Codon (ENC) analysis

Effective Number of Codon (ENC) analysis is used to quantify the absolute codon usage bias by evaluating the degree of codon usage bias exhibited by the coding sequences, regardless of gene length and the number of amino acids (Wright, 1990). ENC values range from 20, indicating extreme codon usage bias using only one of the possible synonymous codons for the corresponding amino acid, to 61, indicating no bias using all possible synonymous codons equally for the corresponding amino acid. The larger the extents of codon usage bias in a gene, the smaller the ENC value. It is also generally accepted that genes have a significant codon usage bias when the ENC value is less than or equal to 35 (Wright, 1990; Comeron, 1998). The ENC values are computed using the formula given by Wright (1990) as follows:

$$ENC = \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6}$$

where, F_k (k = 2, 3, 4 or 6) is the average of the Fk values for k-fold degenerate amino acids. The F value denotes the probability that two randomly chosen codons for an amino acid with two codons are identical.

3.5 ENC Plot Analysis

An ENC-GC3 plot (ENC-plot) is generally used to determine the dominant factors between mutational bias and translational selection that may influence the codon usage bias (Wright, 1990). The ENC values are plotted against the GC content at the third codon position and a standard curve is drawn to represent the maximum influence of GC3 on codon usage bias. It is suggested that if the codon usage variation is only constrained by mutational bias, the ENC values will lie on or around the standard curve. However, if ENC values lie far lower than the standard curve, other factors such as translational selection play a major role in shaping the codon usage bias. In general, if genes distribution pattern approaches the expected standard curve, then the codon usage bias of genes is mainly influenced by mutational bias related to compositional constraints. In contrast if genes distribution pattern went further away from the expected standard curve, then the codon usage bias of genes is also influenced by mutational bias but other factors such as translational selection involved as well. The expected standard curve is plotted using calculation formulated by Wright (1990):

$$ENC^{expected} = 2 + s + \frac{29}{s^2 + (1 - s^2)}$$

where *s* represent the GC3 value for each protein-coding genes.

3.6 Parity Rule 2 (PR2) Bias Analysis

The Parity Rule 2 (PR2) bias analysis plot is used to determine the influence of mutational bias and translational selection on codon usage bias. PR2 bias is a plot of AT-bias [A3/(A3 + T3)] as the ordinate and GC-bias [G3/(G3+C3)] as the abscissa at the third codon position of the entire genes. The centre of the plot, where both coordinates are 0.5, is the place where A = T and G = C (PR2), with total equilibrium between influence of mutational bias and translational selection. A vector from the centre represents the extent and direction of biases from PR2. PR2 bias plots are highly informative when PR2 biases at the third position of codons in the amino acids of individual genes are plotted (Sueoka, 1995; Sueoka, 1999).

3.7 CAI-ENC Plot Analysis

CAI-ENC plot is used to investigate the balance between mutational bias and translational selection in shaping codon usage bias (Nasrullah et al., 2015; Vicario et al., 2007). This plot is drawn with ENC value as ordinate and CAI value as abscissa, and each point represents an individual gene. The strength of relationship or the degree of codon usage bias between both CAI and ENC is measured by r value of the plot. If the correlation, r value of the two indices approaches –1, this indicates that the translational selection is preferred over mutational bias, meanwhile if the r value approaches 0 (no correlation), mutational bias may be more influential than translational selection.

3.8 Neutrality Plot Analysis

Neutrality plot is used to investigate and compare the balance of between mutational bias and translational selection influences in shaping codon usage bias (Sueoka, 1988). Neutrality plot is drawn with average GC content at the first and second codon position (GC12) as ordinate and GC content at the third codon position (GC3) as abscissa, and each point represents an individual gene. Regression analysis on GC3 is

regarded as mutational bias-translational selection equilibrium coefficient and the evolutionary extent of mutational bias and translational selection is represented as the slope of the regression line. A regression slope closer to 0 or tend to sloped to the horizontal axis, indicating translational selection as the dominant factor, while a slope closer to 1 or the points lie along the diagonal distribution, indicating mutational bias as the dominant factor (Sueoka, 1988).

3.9 Principal Component Analysis (PCA)

Principal Component Analysis is a multivariate statistical method to transform a set of observations of correlated variables into a set of linearly uncorrelated variables spanning a space of lower dimensionality (Jolliffe, 2002). In the present study, PCA is used to analyse the major trends in codon usage patterns among coding sequences. PCA involves a mathematical procedure that transforms correlated variables which are RSCU values into a smaller number of uncorrelated variables called principal components. The transformation is defined so that the first principal component accounts for the largest possible variance of the data, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.

3.10 Software used to draw figures

PCA was performed using the XLSTAT 2015 software (https://www.xlstat.com/). Neutrality analysis plot, ENC-GC3 plot, CAI-ENC plot and PR2 bias plot were generated using R Studio software version 0.99.902 (https://www.rstudio.com/products/rstudio/).

CHAPTER 4: RESULTS

4.1 Transcriptome profiling

The transcripts were aligned and then merged together to form a single non-redundant set of transcripts. The abundance of assembled transcripts was estimated using Fragments Per Kilobase of exon per Million fragments mapped (FPKM) value. By setting the FPKM > 0.1 threshold and only selecting protein-coding genes, a total of 11,644 monocytes protein-coding transcripts were identified. Meanwhile, for B and T lymphocytes and human reference, a total of 13,574, 11,480 and 19,814 protein-coding transcripts were identified respectively. The distribution of identified genes and transcripts were listed in Table 4.1.

4.2 Nucleotide composition analysis

Compositional constraints of the genome have been reported to influence the codon usage preference (Jenkins et al., 2003). Therefore, it is important to analyse the overall nucleotide composition of protein-coding genes expressed in monocytes and compare it to the nucleotide composition of protein-coding genes expressed in B and T lymphocytes and human protein-coding genes (Table 4.2). Nucleotide composition at the third codon position was also calculated due to of its great influence on codon usage preference (Table 4.3). The GC content in protein-coding genes expressed in monocytes (50.2 ± 0.05) was lower compared to the one expressed in B lymphocytes (53.1 ± 0.05), T lymphocytes (51.8 ± 0.05) and human protein-coding genes (53.0 ± 0.05). Interestingly, for B and T lymphocytes and human protein-coding genes, the GC at the third position of codon (GC3) content were much higher compared to the GC content. Not many changes were identified in the GC content at the third position of the codon in protein-coding genes expressed in monocytes. These observations indicated that protein-coding genes

expressed in monocytes are enriched with A and T nucleotide especially at the third position of the codon.

Table 4.1: Number of protein-coding genes identified in each dataset.

Data	Number of Protein-Coding Genes
Monocytes	11,644
T Lymphocytes	13,574
B Lymphocytes	11,480
Human Protein-Coding Genes	19,814

Table 4.2: Overall composition of nucleotides. The percentage displayed represents the average value of each nucleotide in each set of protein-coding genes.

Data	A%	T%	G%	C%	AT%	GC%
Monocytes	24.9 ± 0.05	24.9 ± 0.05	25.0 ± 0.04	25.2 ± 0.05	49.8 ± 0.05	50.2 ± 0.05
B lymphocytes	25.4 ± 0.06	21.5 ± 0.04	27.0 ± 0.04	26.1 ± 0.06	46.9 ± 0.05	53.1 ± 0.05
T lymphocytes	26.2 ± 0.09	22.0 ± 0.04	26.5 ± 0.04	25.3 ± 0.05	48.2 ± 0.07	51.8 ± 0.05
Human Protein-Coding Genes	25.2 ± 0.06	21.8 ± 0.05	26.7 ± 0.04	26.4 ± 0.06	47.0 ± 0.06	53.0 ± 0.05

Table 4.3: Composition of nucleotide at the third codon position. The percentage displayed represents the average value of each nucleotide at the third codon position.

Data	A3%	T3%	G3%	C3%	AT3%	GC3%
Monocytes	23.4 ± 0.05	26.3 ± 0.06	25.0 ± 0.05	25.3 ± 0.06	49.7 ± 0.05	50.3 ± 0.05
B lymphocytes	18.7 ± 0.09	21.5 ± 0.09	29.8 ± 0.08	30.0 ± 0.10	40.2 ± 0.09	59.8 ± 0.09
T lymphocytes	20.0 ± 0.09	22.7 ± 0.09	28.8 ± 0.08	28.5 ± 0.10	42.7 ± 0.09	57.3 ± 0.09
Human Protein-Coding Genes	18.6 ± 0.09	21.4 ± 0.08	29.3 ± 0.08	30.7 ± 0.11	40.0 ± 0.09	60.0 ± 0.09

4.3 Codon usage pattern and preferences

In order to investigate the extent of codon usage bias in protein-coding genes expressed, the RSCU values of each codon were calculated. The results of RSCU analysis is presented in Table 4.4 and summarized in Table 4.5. Among the frequently used codons for each amino acids, monocytes displays fourteen AT-ended codons (AGA, TCT, GGA, CCT, ACA, ATT, TGT, GAT, GAA, TTT, CAT, AAA, AAT and TAT) with (A-ended: 5; T-ended: 9) and the remaining four (CTG, GCC, GTG, CAG) were GC-ended codons. It is interesting to note that B, T lymphocytes and human protein-coding genes preferred to used significantly different codons in which B lymphocytes preferred only GC-ended codons (CTG, CGG, AGC, GCC, GGC, CCC, ACC, GTG, ATC, TGC, GAC, GAG, TTC, CAC, AAG, AAC, CAG and TAC) with (C-ended: 12; G-ended: 6). T lymphocytes only prefer to use two AT-ended codons (AGA, TGT and AAT) while human protein-coding genes only prefer one AT-ended codon (AGA). There was a significant difference between monocytes protein-coding genes to the rest of protein-coding genes studied in term of the preference towards AT ended codons ($p < 0.05$). It is proven from RSCU analysis that monocytes protein-coding genes showed higher codon usage bias towards AT compared to GC-ended codons.

More observation can be done based on RSCU whereby the values can be divided into three categories: (A) codons with RSCU values less than 0.6 represent under-represented codon, (B) codons with RSCU values between 0.6 and 1.6 represent codon with a little or no bias, or (C) codons with RSCU values more than 1.6 represent over-represented codon (Wong, 2010). Analysis of RSCU value based on over or under-represented codons showed that preferred and non-preferred codons ranges between 0.29 and 2.37 with majority fell between the 0.6 to 1.6 categories. It is remarkable to note that majority of over-represented codons are G-ended codons while majority of under-represented codons are A or G-ended codons (Table 4.4). We have not found any similar

over-represented codon across all set of protein-coding genes but we found many similarity in under-represented codons which are CTA, CGT, TCG, GCG, CCG and ACG. Regardless of the RSCU value obtained, similar codon preference across all set of protein-coding genes can be seen in four amino acids which are Leucine, Alanine, Valine and Glutamine. For 6 folded degeneracy amino acid (Arg, Leu and Ser), the preferred codon for Leu in all datasets is CTG. Moreover, CTG codon is over represented (RSCU > 1.6) in B and T lymphocytes and human protein-coding genes, while high RSCU value was identified (RSCU = 1.47) for monocytes. The preferred codon for Ser in monocytes was TCT, meanwhile for human protein-coding genes and both B and T lymphocytes were TCC and AGC respectively. For Arg, CGG was selected for B lymphocytes while others including monocytes preferred AGA. In 4-fold degeneracy amino acids (Ala, Gly, Pro, Thr and Val), except for Ala and Val, all 4 degeneracy amino acids showed different preference pattern in monocytes compared to the other two immune cells and human protein-coding genes studied. Different preference pattern was also identified for 3-fold degeneracy amino acid (Ile) in monocytes compared to the other cell types. Monocytes preference towards codon encoded for 2-fold degeneracy amino acids (Asn, Asp, Cys, Glu, Gln, His, Lys, Phe, and Tyr) was identified to be differ compared to the other cells except for the Gln that was encoded by CAG in all cells. In general, we identified that except for 4 amino acids (Ala, Gln, Leu and Val) the rest of the degeneracy amino acids in monocytes are having preference towards codons with either A or T at the third position.

Table 4.4: Synonymous codon usage of protein-coding genes. The codon usage displayed as RSCU[a] value.

Amino Acid	Codon	Monocytes	B lymphocytes	T lymphocytes	Human Protein-Coding Genes
Leu, L	CTA	0.59	0.44	0.46	0.43
	CTC	1.01	1.11	1.06	1.14
	CTG	1.47*	2.37*	2.22*	2.35*
	CTT	1.07	0.82	0.88	0.81
	TTA	0.86	0.48	0.55	0.48
	TTG	1.01	0.79	0.84	0.78
Arg, R	AGA	2.02*	1.25	1.38*	1.29*
	AGG	1.79	1.18	1.22	1.26
	CGA	0.44	0.70	0.70	0.65
	CGC	0.62	1.09	1.00	1.10
	CGG	0.72	1.28*	1.20	1.22
	CGT	0.41	0.50	0.50	0.48
Ser, S	AGC	1.19	1.42*	1.36*	1.43
	AGT	1.00	0.95	0.99	0.91
	TCA	1.16	0.93	0.98	0.93
	TCC	1.12	1.25	1.18	1.29*
	TCG	0.29	0.32	0.30	0.33
	TCT	1.24*	1.14	1.18	1.12
Ala, A	GCA	1.12	0.93	0.98	0.92
	GCC	1.25*	1.57*	1.51*	1.60*
	GCG	0.47	0.42	0.40	0.43
	GCT	1.17	1.07	1.11	1.05

Table 4.4, continued

Gly, G	GGC	1.04	1.34*	1.28*	1.35*
	GGA	1.18*	1.01	1.07	1.00
	GGG	1.05	1.00	0.96	0.99
	GGT	0.72	0.66	0.69	0.65
Pro, P	CCA	1.20	1.12	1.16	1.10
	CCC	1.12	1.27*	1.21*	1.28*
	CCG	0.46	0.45	0.43	0.46
	CCT	1.22*	1.16	1.20	1.15
Thr, T	ACA	1.36*	1.16	1.21	1.15
	ACC	1.10	1.36*	1.29*	1.40*
	ACG	0.37	0.45	0.42	0.44
	ACT	1.16	1.02	1.08	1.01
Val, V	GTA	0.74	0.50	0.54	0.48
	GTC	0.83	0.91	0.88	0.94
	GTG	1.38*	1.84*	1.76*	1.83*
	GTT	1.05	0.75	0.82	0.74
Ile, I	ATA	0.9	0.52	0.56	0.53
	ATC	0.84	1.36*	1.27*	1.37*
	ATT	1.26*	1.13	1.17	1.10
Cys, C	TGC	0.95	1.05*	0.99	1.07*
	TGT	1.05*	0.95	1.01*	0.93
Asp, D	GAC	0.96	1.04*	1.00*	1.06*
	GAT	1.04*	0.96	1.00*	0.94

Table 4.4, continued

Glu, E	GAA	1.01*	0.86	0.92	0.86
	GAG	0.99	1.14*	1.08*	1.14*
Phe, F	TTC	0.76	1.05*	1.00*	1.06*
	TTT	1.24*	0.95	1.00*	0.94
His, H	CAC	0.98	1.13*	1.08*	1.15*
	CAT	1.02*	0.87	0.92	0.85
Lys, K	AAA	1.15*	0.88	0.93	0.89
	AAG	0.85	1.12*	1.07*	1.11*
Asn, N	AAC	0.86	1.03*	0.98	1.04*
	AAT	1.14*	0.97	1.02*	0.96
Gln, Q	CAA	0.80	0.52	0.56	0.54
	CAG	1.2*	1.48*	1.44*	1.46*
Tyr, Y	TAC	0.84	1.09*	1.04*	1.10*
	TAT	1.16*	0.91	0.96	0.90

a RSCU, Relative Synonymous Codon Usage
* codon preferred or highest RSCU value

Table 4.5: The summary of codon preference of each amino acid. Preferred codons are known as codons which are used more among the synonymous codons.

Folding Degeneracy	Amino Acid	Monocytes	B lymphocytes	T lymphocytes	Human Protein-Coding Genes
6	Leu, L	CTG	CTG	CTG	CTG
	Arg, R	AGA	CGG	AGA	AGA
	Ser, S	TCT	AGC	AGC	TCC
4	Ala, A	GCC	GCC	GCC	GCC
	Gly, G	GGA	GGC	GGC	GGC
	Pro, P	CCT	CCC	CCC	CCC
	Thr, T	ACA	ACC	ACC	ACC
	Val, V	GTG	GTG	GTG	GTG
3	Ile, I	ATT	ATC	ATC	ATC
2	Cys, C	TGT	TGC	TGT	TGC
	Asp, D	GAT	GAC	-	GAC
	Glu, E	GAA	GAG	GAG	GAG
	Phe, F	TTT	TTC	-	TTC
	His, H	CAT	CAC	CAC	CAC
	Lys, K	AAA	AAG	AAG	AAG
	Asn, N	AAT	AAC	AAT	AAC
	Gln, Q	CAG	CAG	CAG	CAG
	Tyr, Y	TAT	TAC	TAC	TAC

- represent no codon preference

4.4 Strength of codon usage bias

To quantify the extent of variation and degree of codon usage bias among protein-coding gene expressed in monocytes, B and T lymphocytes as well as human protein-coding genes, the Effective Number of Codon (ENC) were calculated. The mean ENC values among the protein-coding genes studied ranged from 48.33 to 52.92. With the ENC value of 52.92 ± 3.35, protein-coding genes expressed in monocytes were identified to be less bias compare to the protein-coding genes expressed in B and T lymphocytes (Table 4.6). ENC values also showed that the degree of codon usage bias in B and T lymphocytes are quite similar to the one from human protein-coding genes. Considering protein-coding genes with ENC value of less than 35 as highly bias gene, only 0.08% of monocytes protein-coding genes were identified as highly bias in our analysis (Table 4.7). While analyzing the ENC value in monocytes in Figure 4.1, we detected statistically significant different between ENC value in monocytes compared to B and T lymphocytes as well as human protein-coding genes ($p < 0.05$). Besides ENC analysis, we also determine the codon usage preferences of monocytes with B and T lymphocytes as well as human protein-coding genes using Codon Adaptation Index (CAI). CAI analysis revealed that not much different for the mean of the CAI value for all datasets analyzed. However, there is a slight difference in the distribution of the CAI value in monocytes compare to the other cells as shown in the Figure 4.1.

Table 4.6: Statistical data of ENC[a] and CAI[b] values for each set of protein-coding genes. The mean is the average value of the data, median is the middle number of data and standard deviation (STD) is the measure of spreading of numbers. The minimum and maximum values were also recoded for each protein-coding gene in the table.

Data	Index	Mean	Median	STD	Min	Max
Monocytes	ENC	52.92	53.32	3.35	26.93	61.00
	CAI	0.22	0.216	0.02	0.091	0.44
B lymphocytes	ENC	48.47	49.63	6.59	24.56	61.00
	CAI	0.23	0.228	0.04	0.058	0.50
T lymphocytes	ENC	49.21	50.47	6.39	25.45	61.00
	CAI	0.23	0.224	0.04	0.058	0.76
Human Protein-Coding Genes	ENC	48.33	49.66	6.79	20.00	61.00
	CAI	0.23	0.227	0.04	0.033	0.78

a ENC: Effective Number of Codons
b CAI: Codon Adaptation Index

Table 4.7: Number of protein-coding genes according to ENC values. ENC is the effective number of codons used to quantify the codon usage bias.

Data	ENC < 35[a]	ENC > 35[b]	< 35 (%)
Monocytes	9	11635	0.08
B Lymphocytes	427	13147	3.15
T Lymphocytes	281	11199	2.45
Human Protein-Coding Genes	785	19029	3.96

[a] Number of protein-coding genes with ENC value less than 35
[b] Number of protein-coding genes with ENC value more than 35

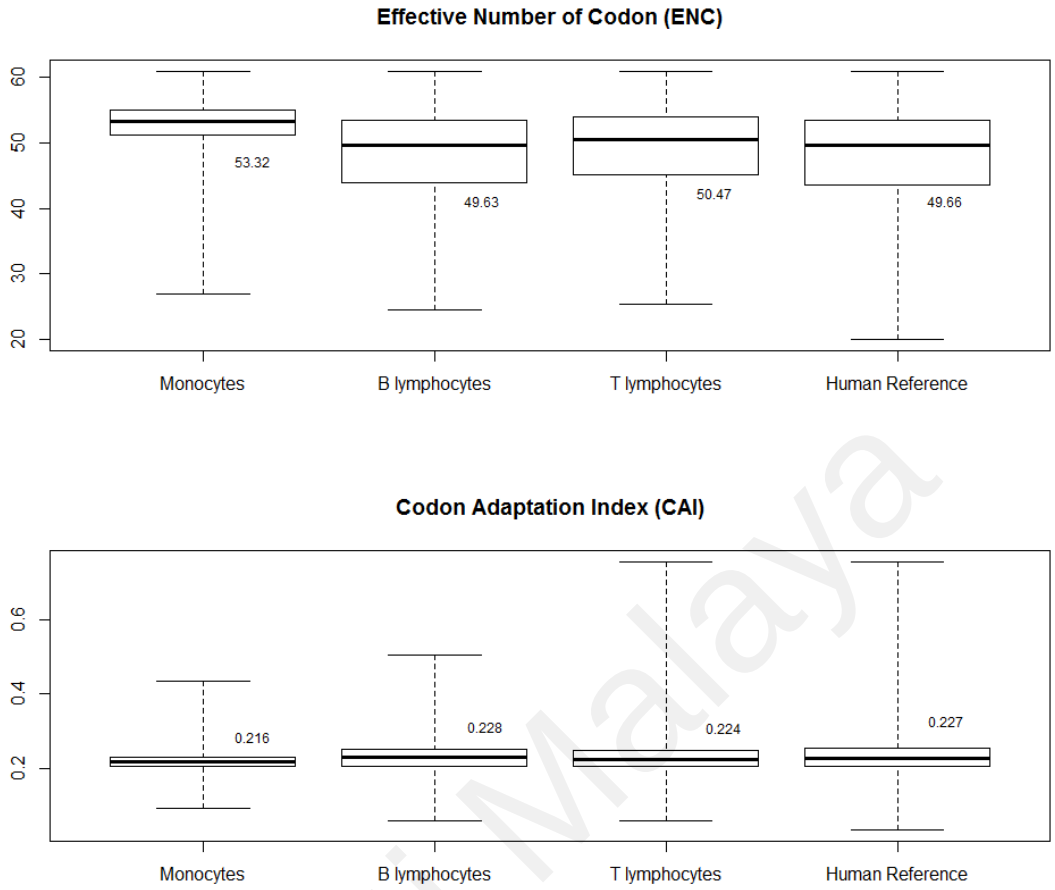

Figure 4.1: Box plot of Effective Number of Codon (ENC) and Codon Adaptation Index (CAI). Distribution of data displayed based on the five-number summary: minimum, first quartile, median, third quartile, and maximum value for each dataset.

4.5 The relationship between ENC and CAI

CAI analysis was performed to predict the level of gene expression (Naya, 2001; Gupta, 2004) in which higher CAI value represent elevated levels of gene expression. ENC analysis was used to quantify the general codon usage bias by evaluating the degree of codon usage bias exhibited by the coding sequences, regardless of gene length and the number of amino acids. The relationship between codon usage bias and gene expression levels was analysed using Pearson's correlation analysis using both ENC and CAI. Pearson's correlation analysis often used to measure the linear correlation between two variables by providing the value between +1 to -1 where +1 represent positive correlation and -1 represent negative correlation. Thus, if the result appears 0, that means there is no correlation between both variables. Table 4.8 shows the Pearson's correlation analysis result between ENC and CAI. It can be seen that all of the protein-coding genes showed negative and weak correlation ($p < 0.05$). Negative correlation indicated that higher gene expression levels have higher degree of codon usage bias.

Table 4.8: Pearson's correlation analysis. Pearson's correlation analysis for each set of protein-coding genes. The variables are Effective Number of Codons (ENC) and Codon Adaptation Index (CAI).

Data	N[a]	Pearson's Correlation[b] ENC[c]/CAI[d]
Monocytes	11644	-0.385
B lymphocytes	13574	-0.470
T lymphocytes	11480	-0.420
Human Protein-Coding Genes	19814	-0.460

[a] represented the number of valid
[b] correlation is significant at the 0.05 level
[c] Effective Number of Codon
[d] Codon Adaptation Index

4.6 Role of mutational bias and translational selection in shaping codon usage bias

The heterogeneity of codon usage pattern in monocytes was analysed by plotting the ENC values of each protein-coding genes expressed in monocytes against the third position of each codons in the corresponding genes (GC3). This analysis was performed to determine the role of nucleotide compositional constraint or mutational bias on shaping the synonymous codon usage pattern. A standard curve of expected ENC value against GC3 was calculated based on the Wrights' methods (Wright, 1990). If the codon usage in a gene is affected only by mutational bias, the corresponding point for the gene would lie on or close to the expected curve. However, if other factors such as translational selection were involved in the codon usage pattern of a gene, the corresponding point would depart away below the expected curve (Wright, 1990). In ENC-GC3 plot (Figure 4.2) for all the set of protein-coding genes studied, the point representing each gene clustered together below the expected ENC curve with only a few of the point fell on the expected curve. This result indicated the presence of mutational bias in shaping codon usage bias in monocytes, B and T lymphocytes. This observation also suggested that mutational bias was not the sole factor determining the codon usage bias in them but other factors such as translational selection involved as well.

To further confirm the influence of both mutational bias and translational selection in the datasets, PR2-bias plot was performed. PR2 bias plot is an analysis to determine the relations between purines (A/G) and pyrimidines (C/T) in genes. It is suggested that if only mutation bias influence codon usage bias, nucleotide G and C, A and T should be used equally among the genes (G=C, A=T) (Zhang et al., 2007), in which the average position of genes would be positioned exactly at the centre of PR2 plot, where both coordinates are 0.5. The result showed that the nucleotides are not used proportionally and also that C and T were seen to appear more frequent than A and G in

all the gene sets. The result also showed that the centre of data distribution were shifted to the left, indicating that mutational bias was not the only factor affecting codon usage bias (Figure 4.3). This differences between nucleotide content suggested that not only mutational bias influenced the codon usage bias but also other factors such as translation selection could also involve.

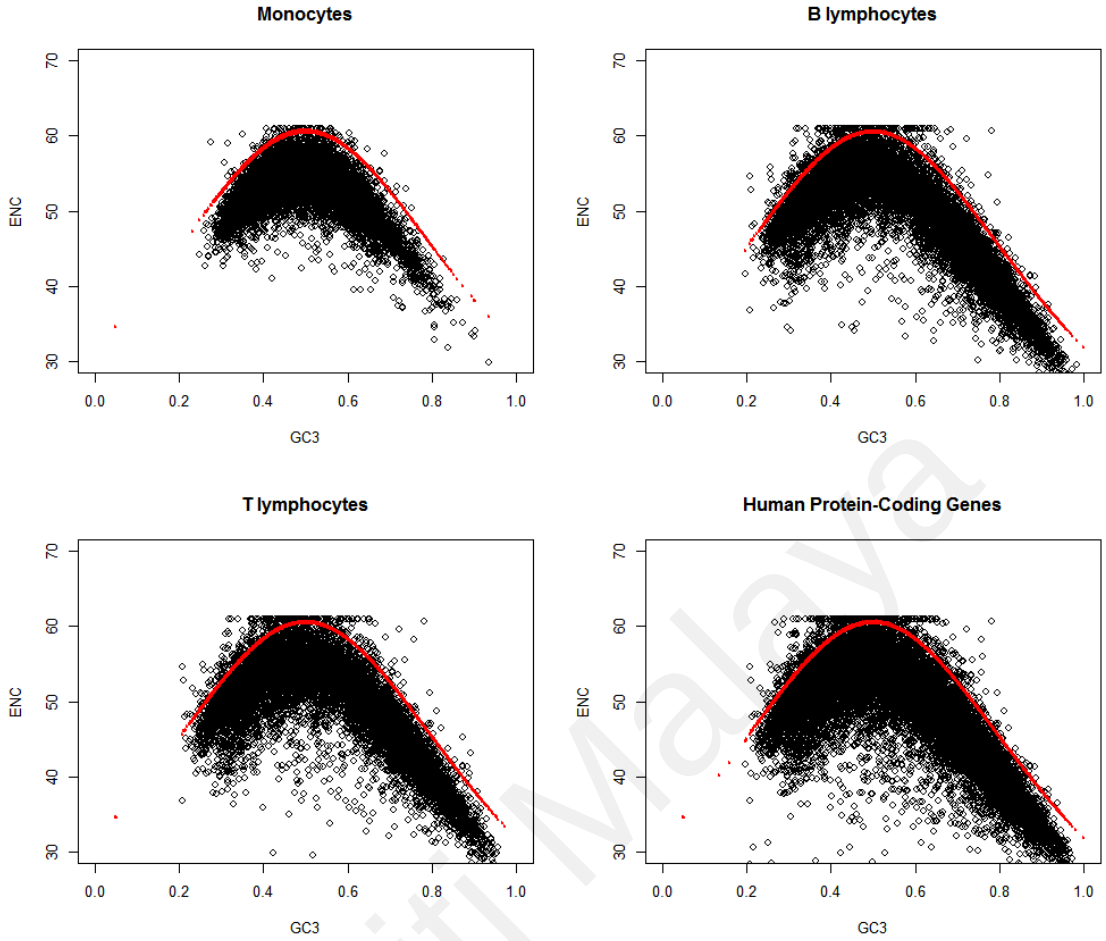

Figure 4.2: GC3 vs. ENC plot. Scatter plot of GC3 (X axis) vs. ENC (Y axis) for monocytes, B and T lymphocytes and human protein-coding genes. ENC represent the effective number of each genes and GC3 represent G+C content at the third codon position of each genes. The expected curve represents the maximum influence of GC3 or mutational bias on codon usage bias.

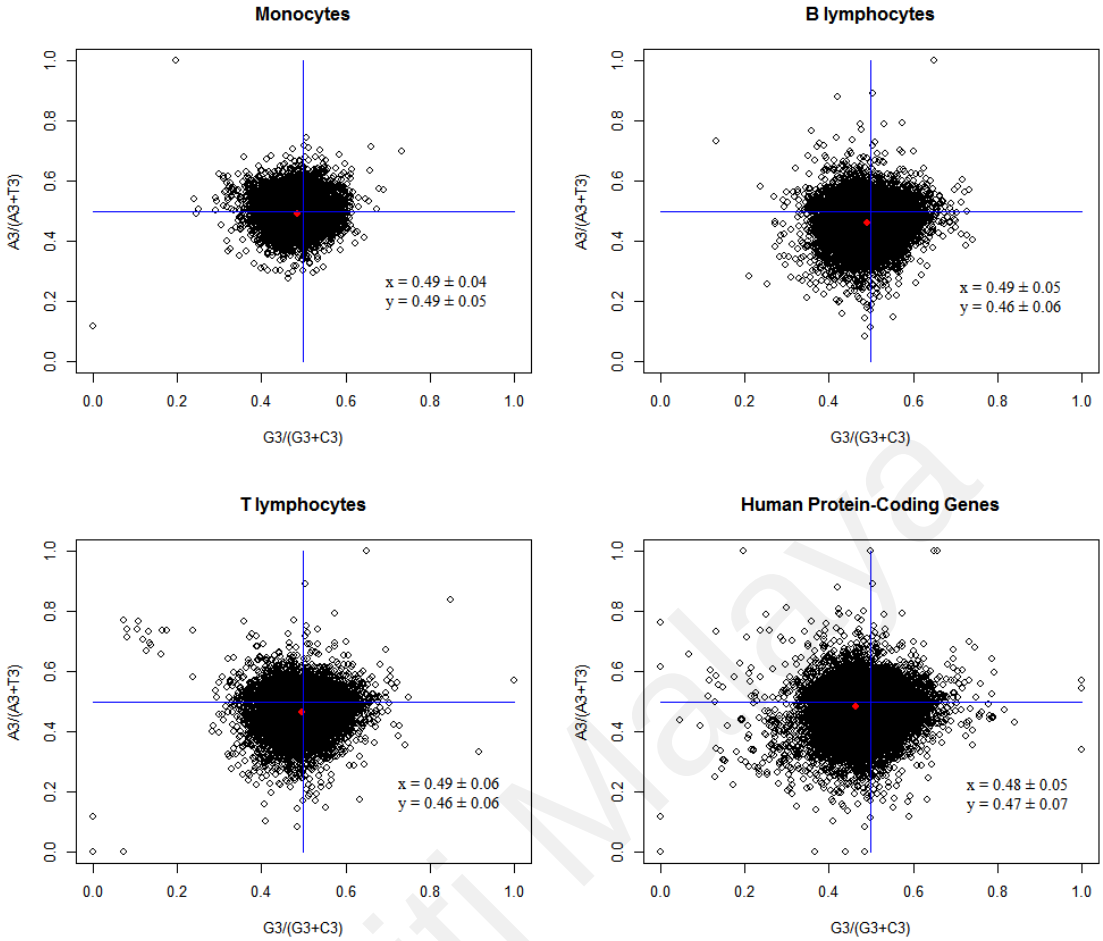

Figure 4.3: PR2-bias plot. Protein-coding genes are plotted based on their GC bias [G3/(G3+C3)] and AT bias [A3/(A3+T3)] in the third codon position. Center point was made by two intersecting lines in the middle of the plot represent the state of no codon usage bias.

4.7 Mutational bias versus translational selection in shaping codon usage bias

CAI-ENC plot was constructed to analyse the influence of mutational bias and translational selection on the codon usage bias. A linear correlation analysis between the two variables was measured. Both CAI and ENC are considered because both signify the extent of codon usage bias exhibited and relationship between them have been shown in Figure 4.4 in form of r value. If the correlation, between both parameters is close to -1, this suggests that the translational selection is preferred over mutational bias. Otherwise, if the r value approaches 0, mutational bias may be more influential than translational selection. As shown in Figure 4.4, the correlation between CAI and ENC in monocytes is closer to zero compared to B and T lymphocytes. Interestingly the correlations between CAI and ENC in B and T lymphocytes are found to be almost similar to the one in human protein-coding genes. This result indicated that protein-coding genes expressed in monocytes were heavily influenced by mutational bias in shaping the codon usage bias compared to translational selection. However, as shown in the CAI-ENC plot, for B and T lymphocytes, the correlation analysis revealed the dominant role of translational selection in determining the codon usage bias of the expressed genes.

Further analysis was performed using neutrality plot analysis to identify the role of key determinant factors which are translational selection and mutational bias in structuring codon usage pattern. The neutrality plot is a regression analysis of average GC content at the first and second positions (GC12) on GC content at the third codon position (GC3). A significant positive correlation was observed between GC3 and GC12 of in all the plots shown in the result (Figure 4.5). The positive regression slope in neutrality plots indicated that intragenic GC mutational bias affects the GC content at all codon positions in a uniform pattern. Monocytes protein-coding genes shows regression line with a slope value of 0.595, indicating relative neutrality of 59.5 % with relative constraint of 40.5 %, indicating the minor influence of translational selection on the codon

usage patterns. In contrast with monocytes, B and T lymphocytes as well as human protein-coding genes exhibited regression slope value closer to 0, indicating the dominant role of translational selection in shaping codon usage bias. Both analyses indicated the major factor that affecting codon usage bias in monocytes was mutational bias, while for B and T lymphocytes the major factor affecting codon usage bias was translational selection.

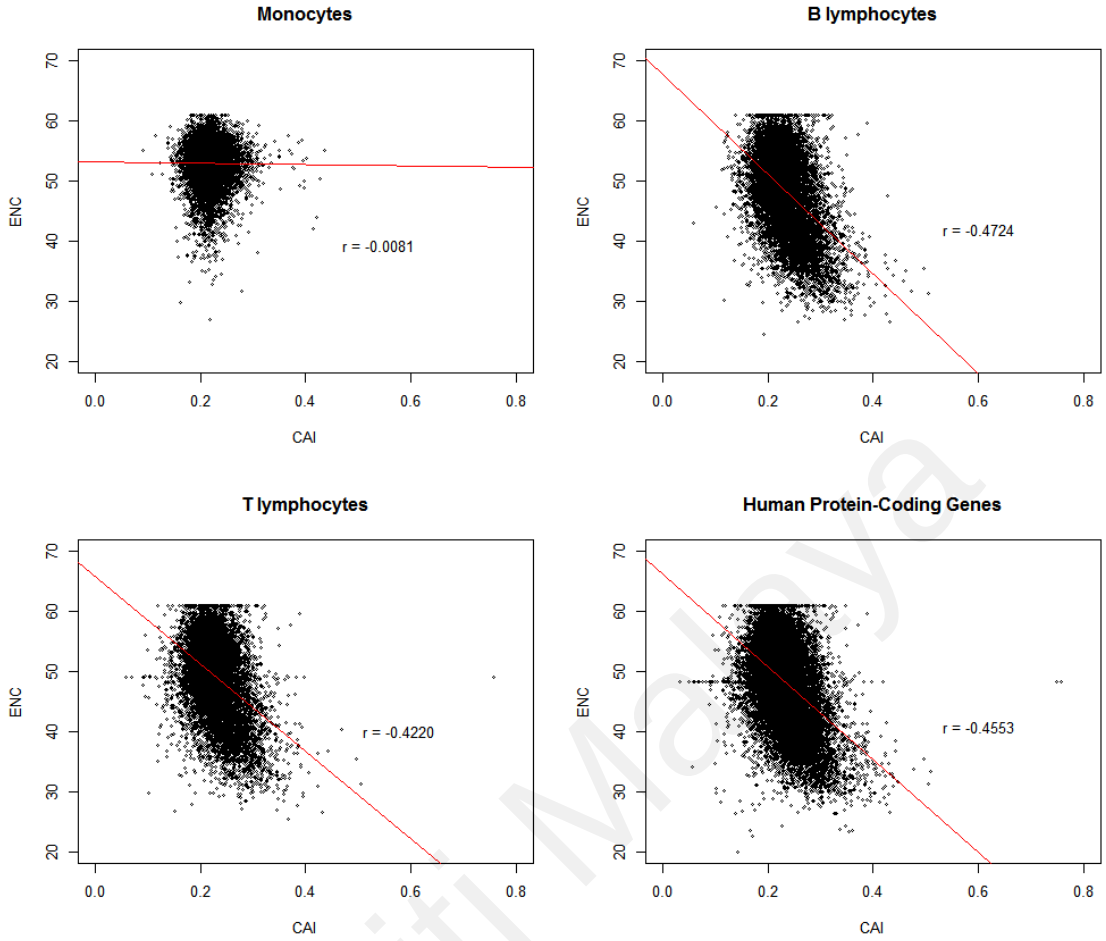

Figure 4.4: CAI vs. ENC plot. Scatter plot of CAI (X axis) vs. ENC (Y axis) for monocytes, B and T lymphocytes and human protein-coding genes. ENC represent effective number of each genes and CAI represent index to measure level of gene expression. Individual genes are plotted based on the CAI value versus the ENC value.

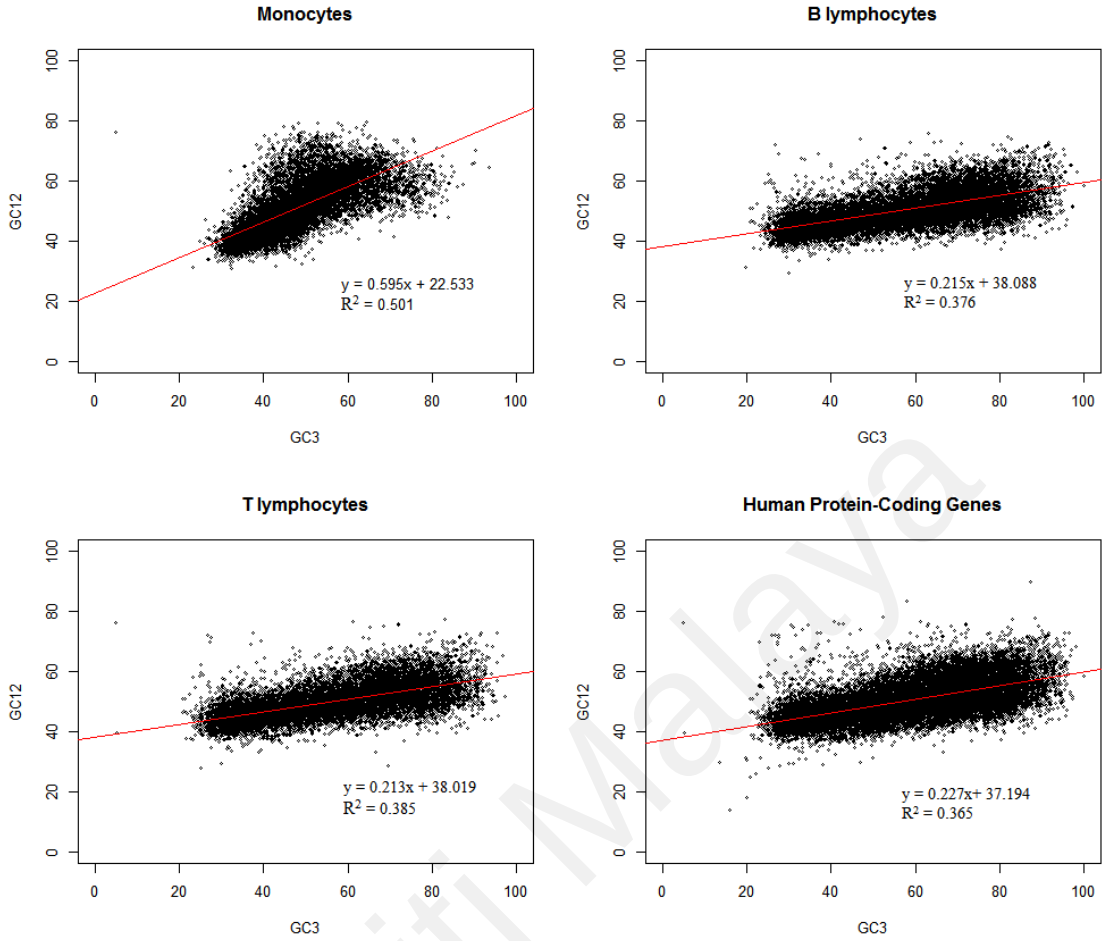

Figure 4.5: Neutrality Plot (GC3 vs. GC12). Scatter plot of GC3 (X axis) vs. GC12 (Y axis) for monocytes, B and T lymphocytes and human protein-coding genes. Individual protein-coding genes are plotted based on the mean GC content in the first and second codon position versus the GC content of the third codon position. Regression lines and coefficient of determination, R2 are shown in the plot.

4.8 Principal Component Analysis (PCA)

PCA was done to identify the similarities and differences of codon usage patterns in different human cells. From the analysis in Figure 4.6, the major trend was identified, in which the first principal axis (F1) accounted for 92.01 % of the total variation indicating substantial similarity in amino acid usage between genes, and the second principal axis (F2) accounted for 7.55 % of the total variation in synonymous codon usage. Based on the point located on the plot, monocytes have significantly different codon usage pattern compare to B and T lymphocytes as well as human protein-coding genes. Pearson's correlation analysis was also performed to identify the strength of the first and second principal axis towards codon usage indices which are ENC and CAI. Based on Figure 4.7, both indices have high correlation towards the first and second principal axis with ENC have the highest correlation towards the first principal axis, followed by CAI (r value = -0.99, 0. 065, respectively). For second principal axis, ENC again show the highest correlation along with GC3 (r value = 0.99, -0.99, respectively) while CAI again exhibit the lowest correlation with r value of -0.88.

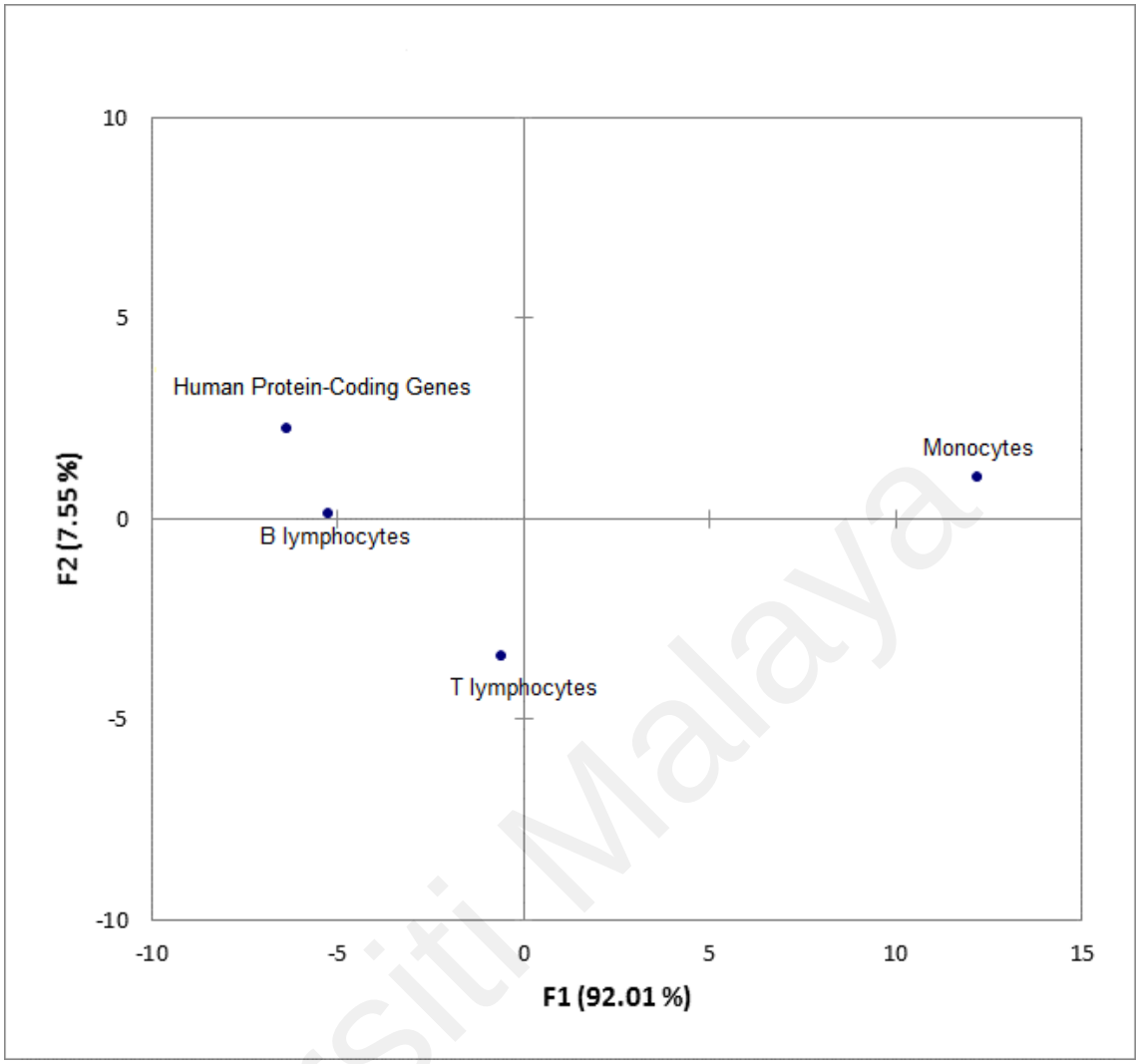

Figure 4.6: Principal Component Analysis (PCA). The points represent the average trend in codon usage of each datasets. This analysis depicts the variation among the RSCU values of codons of the protein-coding genes studied.

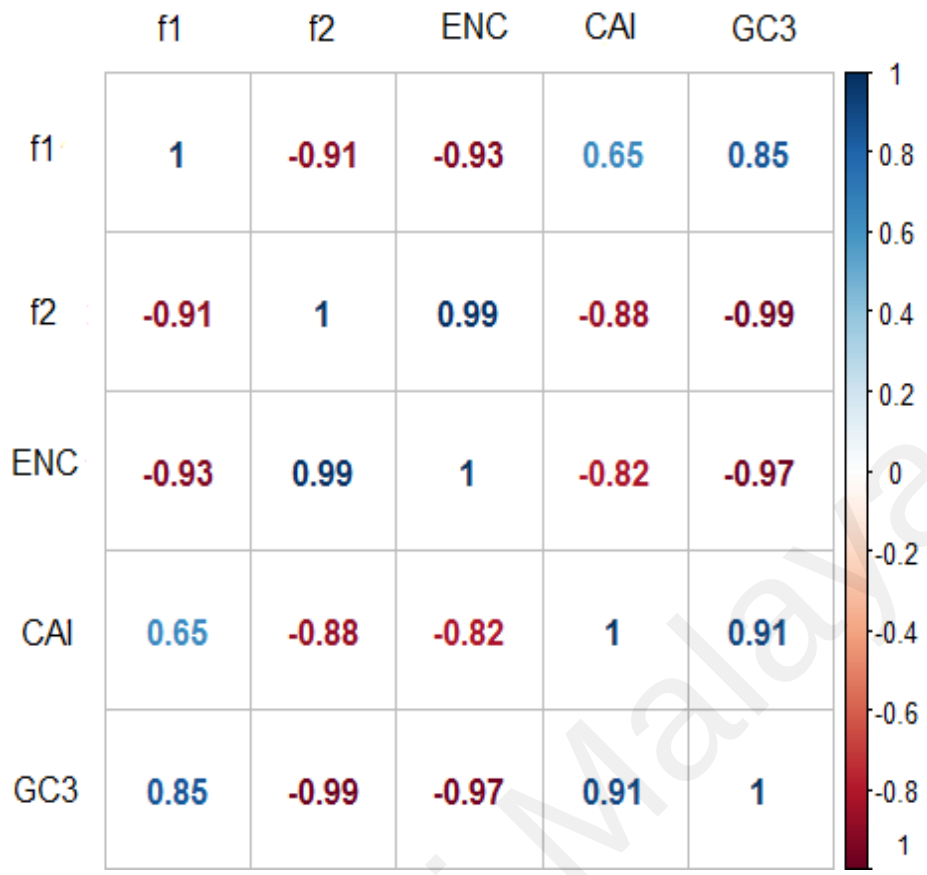

Figure 4.7: Pearson's correlation analysis. Correlation analysis between first and second principal axis towards ENC, CAI and GC content at the third codon position.

CHAPTER 5: DISCUSSION

This study has highlighted the codon usage patterns of protein-coding genes in a comparative manner between selected human immune cells. As synonymous codon usage pattern between organisms are non-random and species specific, we are curious to know if it is also applied to different type of cells in a same organism. Selection-mutation drift model suggested that codon usage bias is mainly influenced by a balance between translation selection and mutational bias. This theory can also be applied to human genome as several findings suggest translational selection was responsible for most part of the codon preference activity with the influence of mutational bias as well (Kotlar, 2006; Plotkin, 2004). Tissue specificity in human has been shown to have significant difference in codon usage across different tissues (Plotkin, 2004). This variation may be due to different tRNA abundant for each tissues that lead to different tRNA activity (Se'mon, 2005). Therefore, it is important to know the balance between mutational bias and translational selection in shaping codon usage bias in the respective immune cells as both factors might contribute greatly in shaping codon usage bias.

Nucleotide composition could be one of the important factors in codon usage bias in human protein-coding genes. Here we found that average GC content and GC content at the third codon position are higher than AT content in all immune cells studied including human protein-coding genes. In a study conducted by Bernardi (1995), mammalian genomes including human exhibited large-scale variation in GC content in both coding and noncoding regions. This variation may suggest the possibility of nucleotide composition in influencing codon preference.

The codon usage pattern was found to be significantly similar between human, B and T lymphocytes protein-coding genes. However, monocytes protein-coding genes showed significantly different codon usage pattern based on the RSCU value recorded.

Monocytes protein-coding genes showed a tendency to translate codon ended with AT nucleotides even though the average coding sequences are rich in GC content. The codon usage preference is in contrast to the findings in several studies that showed GC-rich genome usually have the tendency to encode amino acids with GC-ended codons while AT-rich genome tend to prefer AT-ended codons (Singer, 2000; Li, 2015). However, this occurrence is not surprising due to similar findings in previous studies conducted on other species (Anderson, 1993; Rodriguez-trellez, 2000). The best explanation that could describe this phenomenon is that the codon usage pattern seen in monocytes protein-coding genes may be due to intra-genomic variation as proposed by Sharp et al. (2005). It was suggested that even though monocytes protein-coding genes are GC-rich, more than half of the protein-coding genes identified may located in a single large cluster composed of unusual base composition. This cluster is AT-rich in composition differ with the average nucleotide composition of the monocytes protein-coding genes and tends to prefer AT-ended codons. Besides that, another possible reason for the monocytes protein-coding genes displaying such codon usage pattern is due to evolutionary process in part of the genes which may change the codon preference. Anderson et al. (1993) have also observed similar codon usage pattern in their studies on multiple *Drosophila* species whereby one of the species, *Drosophila willistoni* exhibited had GC-rich genome but preferred to encode AT-ended codons compared to other *Drosophila* species that were examined. The shifted codon preference may be due to the impact of the deletion on part of intron in the *Adh* genes of the *willistoni* group (Anderson, 1993; Rodriguez-trellez, 2000). However, the possibility of deletion or insertion occurring in the monocytes protein-coding genes remain unclear, thus further examinations are required. Therefore, based on both the nucleotide composition and RSCU analyses, it is suggested that the codon preference might be influenced mostly by compositional constraints, which is also related to the presence of mutational bias.

Moreover, the mean ENC values ranged from 47.17 ± 6.56 to 52.92 ± 3.35, indicating that the codon usage bias is considered low in all set of protein-coding genes studied. It is also observed that the ENC values were highly correlated with levels of GC in the third codon position. Based on the ENC value recorded, all of the protein-coding genes displayed value of more than 35, suggesting low codon usage bias. The high mean ENC value recorded indicated that all set of protein-coding genes include human protein-coding genes display relatively stable and conserved genomic composition. Our analysis suggested that codon usage bias in monocytes is slightly lower and might be affected by nucleotide compositions. A previous study on 15 different vertebrates had recorded quite similar result of which the ENC values were also relatively high with value more than 35 ranging from 42.41 in *Petromyzon marinus* to 57.00 in *Anolis carolinensis* genes (Qiu, 2011). In human protein-coding genes, the mean ENC value showed quite similar result to the study conducted by Wright (1990) whereby ENC value reported were close to 45 with the distribution of values for individual genes ranges from 30 to 61.

Meanwhile, the CAI values ranged from 0.22 ± 0.02 to 0.24 ± 0.04, indicating that all set of protein-coding genes studied has low mean expression levels. Genes with a higher CAI value are associated with high codon bias but this parameter alone does not distinguish the bias from GC related mutational bias to translational selection. In other words, CAI could not specifically identify the extent of each contribution factors in shaping codon usage bias (Carbone, 2005). In human, it is believed that CAI is a less effective index in which CAI is suitable for organisms with higher rate of replication as seen in prokaryotes and lower eukaryotes (He, 2016). This is because the proteins involved in transcription and translation of bacteria and virus are often highly expressed and leads to higher codon bias (Carbone, 2005; Willenbrock, 2006). Besides that, CAI is also expected to not perform well in GC-rich organism with low mutational bias (Grocock

& Sharp, 2002) but can be very well used as rough estimation of expression levels of genes studied.

Mutational bias and translational selection are considered the two major factors that shape codon usage pattern (Tatarinova, 2010). To identify the influence of both factors in shaping codon usage bias, ENC-GC3 plots were generated and significant positive correlations were observed between ENC and GC3 in all datasets studied including human protein-coding genes. If codon choice is constrained only by mutational bias, the points on the plot will lie on the expected curve (Wright, 1990). The clustering pattern in Figure 4.2 which majority of the points situated below the expected curve of the plots suggested that mutational bias is not the only factor contributed to the codon usage bias but at the same time other factors such as translational selection may also involve. This pattern is in agreement with previous study on other vertebrate such as sea lamprey in which there is only weak translational selection acting upon the codon usage pattern (Qiu, 2011).

To further confirm the influence of both translational selection and mutational bias in codon usage, PR2-bias plots were generated. In this analysis, if synonymous codon usage bias is caused by mutational bias alone, GC or AT should be used proportionally among the degenerate codon groups in a gene (Zhang, 2013). The association between purines (A, G) and pyrimidines (C, T) was analysed by Parity Rule 2 (PR2) bias plot (Figure 4.3) and revealed that the GC and AT are not use equally whereby C and T were observed more frequently than A and G nucleotides in all datasets. The unequal usage of GC and AT in this analysis further reflects the fact that translational selection has played an important role in driving degenerate codon positions in the analysis. From these findings, we can conclude that both mutational bias and translational selection have contributed to the codon bias in all datasets. Similar results have been reported by Sueoka (1988, 1992) in which high frequency of C and T nucleotides observed in the coding

sequences. This was explained to be the result of mutational effect than the selective effect at the DNA base composition level (Sueoka, 2002).

In this study, the presence of translational selection and mutational bias in codon usage bias of monocytes, B and T lymphocytes were identified and therefore further analysis is needed to identify the major influencer between those factors. From CAI-ENC plot analysis and neutrality plot analysis (Figures 4.4 & 4.5), translational selection is believed to play a minor part in monocytes protein-coding genes in contrast to other set of protein-coding genes in which translational selection was found to be the major contributors in shaping the codon usage pattern. These observations clearly showed that codon usage in monocytes are heavily dependent on compositional constraint as they have the lowest GC content among the protein-coding genes studied, suggesting strong mutational bias influence (Urrutia, 2003). Translational selection in human may be due to the need to minimize the mistakes in incorporation of the amino acids and maximize the speed of elongation. It is also essential to increase the cellular concentration of free ribosomes for protein translation (Hershberg, 2008).

Considering the multivariate nature of codon usage, PCA analysis was performed on RSCU values to determine the trends of codon usage variations in coding sequences (Figure 4.6). The result showed that first principal axis, F1 accounted for the major portion of codon usage variation followed by second principal axis, F2. The difference in clustering pattern between monocytes protein-coding genes to other protein-coding genes can be seen clearly due to different in codon preference as shown in Table 4.4. Moreover, Pearson's correlation was performed to evaluate the relationship between the first two axes towards ENC, CAI and GC3 (Figure 4.7). ENC showed the highest correlation value ($r = -0.93$, $p < 0.001$) towards first principal axis followed by GC3 ($r = 0.85$, $p < 0.001$). This correlation study showed strong relationship towards low ENC value as displayed in Table 4.8, suggesting that the codon bias in protein-coding genes studied were low.

CHAPTER 6: CONCLUSION

In this thesis, a comprehensive study on codon usage bias in monocytes, B and T lymphocytes were discussed, including human protein-coding genes as the reference. With the availability of whole transcriptome datasets provided by RNA-Seq technology, comparative study on the global codon usage pattern in human monocytes, B and T lymphocytes as well as human protein-coding genes have been realised. This is the first study that codon usage bias is systematically investigated in human immune cells and it is also the first time RNA-Seq technology was used in analysing codon usage bias in human.

The results showed that monocytes, B and T lymphocytes exhibited distinctive codon usage pattern. Codon usage bias in all of protein-coding genes studied was low according to the codon usage indices performed. Observation on individual cells showed that monocytes have distinct codon usage pattern compared to other cells whereby monocytes prefer to use AT-ended codons while other cells prefer to use GC-ended codons. Each cell have their own unique codon preference to code for each amino acid and only 4 amino acids out of 18 showed similar codon preference across different set of protein-coding genes.

Involvement of both mutational bias and translational selection were detected for all set of protein-coding genes and the degree of involvement differed between cells. In monocytes, the mutational bias was identified to be the dominant factor in shaping its codon usage bias while in contrast with B and T lymphocytes, translation selection was determined to be the major factor influencing the codon preference. Using Principal Component Analysis (PCA), monocytes showed a very distinct codon usage pattern compared to B and T lymphocytes.

In summary, the findings have provided sufficient evidence in order to fulfill all the objectives of the study. It has been shown that monocytes have significantly different codon usage preference compared to B and T lymphocytes. Besides that, the major factor that contribute in shaping codon usage bias in monocytes also differ to B and T lymphocytes. This observation has suggested the nature of codon usage bias in human which are each tissue or cell have their own unique codon usage pattern. This information enables us to identify the evolutionary events that occur between the cells involved and importance of the evolution. With increasing evidences of the role of codon usage bias influence in cell physiology, this report provides new insight in understanding the impact of codon usage bias in human immune system.

REFERENCES

Abbas, A. K., Lichtman, A. H., & Pillai, S. (2012). *Cellular and molecular immunology* (7th ed.). Philadelphia: Elsevier/Saunders.

Agashe, D., Martinez-Gomez, N. C., Drummond, D. A., & Marx, C. J. (2013). Good codons, bad transcript: Large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. *Molecular Biology and Evolution*, *30*(3), 549–560.

Ahn, I., Jeong, B. J., & Son, H. S. (2009). Comparative study of synonymous codon usage variations between the nucleocapsid and spike genes of coronavirus, and C-type lectin domain genes of human and mouse. *Experimental and Molecular Medicine*, *41*(10), 746-756.

Akashi, H. (1994). Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. *Genetics*, *136*, 27–35.

Alberts, B., Johnson, A., & Lewis, J. (2002). *Lymphocytes and the cellular basis of adaptive immunity: Molecular biology of the cell* (4th ed.). New York: Garland Science.

Anderson, C. L., Carew, E. A., & Powell, J. R. (1993). Evolution of the Adh locus in the *Drosophila willistoni* group: The loss of an intron, and shift in codon usage. *Molecular Biology and Evolution*, *10*(3), 605–618.

Andersson, S. G., & Kurland, C. G. (1990). Codon preferences in free-living microorganisms. *Microbiological Reviews*, *54*(2), 198–210.

Argos, P., Rossmann, M. G., Grau, U. M., Zuber, H., Frank, G., & Tratschin, J. D. (1979). Thermal stability and protein structure. *UCLA Forum in Medical Sciences*, *21*, 159-169.

Behura, S. K., & Severson, D. W. (2013). Codon usage bias: Causative factors, quantification methods and genome-wide patterns with emphasis on insect genomes. *Biological Reviews of the Cambridge Philosophical Society*, *88*(1), 49-61.

Benito-Martin, A., Di Giannatale, A., Ceder, S., & Peinado, H. (2015). The new deal: A potential role for secreted vesicles in innate immunity and tumor progression. *Frontiers in Immunology*, *6*.

Bennetzen, J. L., & Hall, B. D. (1982). Codon selection in yeast. *The Journal of Biological Chemistry*, *257*, 3026–3031.

Berg, O. G., & Martelius, M. (1995) Synonymous substitution-rate constants in *Escherichia coli* and *Salmonella typhimurium* and their relationship to gene expression and selection pressure. *Journal of Molecular Evolution*, *41*, 449–456.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., & Rodier, F. (1985). The mosaic genome of warm-blooded vertebrates. *Science*, *228*, 953–958.

Bernardi, G., (1995) The human genome: Organization and evolutionary history. *Annual Review of Genetics*, *29*, 445–476.

Bulmer, M. (1987). Coevolution of codon usage and transfer RNA abundance. *Nature*, *325*(6106), 728-730.

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

Campbell, W. H., & Gowri G. (1990). Codon usage in higher plants, green algae, and *cyanobacteria*. *Plant Physiology*, *92*, 1-11.

Carbone, A., Képès, F., & Zinovyev, A. (2005). Codon bias signatures, organization of microorganisms in codon space, and lifestyle. *Molecular Biology and Evolution*, *22*(3), 547–561.

Carullo, M., & Xia, X. (2008). An extensive study of mutation and selection on the wobble nucleotide in tRNA anticodons in fungal mitochondrial genomes. *Journal of Molecular Evolution*, *66*, 484–493.

Chamary, J.V., & Hurst, L.D. (2005). Biased codon usage near intron-exon junctions: Selection on splicing enhancers, splice-site recognition or something else? *Trends in Genetics*, *21*, 256–259.

Chamary, J.V., & Hurst, L.D. (2005). Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biology*, *6*, 75.71–75.12.

Chamary, J. V., Parmley, J. L., & Hurst, L. D. (2006). Hearing silence: Nonneutral evolution at synonymous sites in mammals. *Nature Reviews Genetics*, *7*, 98–108.

Chaplin, D. D. (2010). Overview of the immune response. *The Journal of Allergy and Clinical Immunology*, *125*(2), 3–23.

Chen, R. (2010). Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PLOS ONE*, *5*, 13574.

Chen, S. L., Lee, W., Hottes, A. K., Shapiro, L., & McAdams, H. H. (2004). Codon usage between genomes is constrained by genome-wide mutational processes. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(10), 3480-3485.

Coghlan, A., & Wolfe, K. H. (2000). Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast*, *16*, 1131–1145.

Comeron, J. M., & Aguade, M. (1998). An evaluation of measures of synonymous codon usage bias. *Journal of Molecular Evolution*, *47*(3), 268–274.

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

Crick, F. H., Barnett, L., Brenner, S. & Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins. *Nature*, *192*, 1227–1232.

Crick, F. H. C. (1966). Codon–anticodon pairing: The wobble hypothesis. *Journal of Molecular Biology*, *19*, 548–555.

Danilova, N. (2012). *The evolution of adaptive immunity.* New York, NY: Springer US, 218-235.

Dressaire, C., Cristophe, G., & Pascal L. (2009). Transcriptome and proteome exploration to model translation efficiency and protein stability in *Lactococcus lactis*. *PLOS Computational Biology*, *5*, 606.

Duret, L., & Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proceedings of the National Academy of Sciences*, *96*(8), 4482-4487.

Duret, L. (2002). Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetics & Development*, *12*(6), 640-649.

Ehrenberg, M., & Kurland, C. (1984). Costs of accuracy determined by a maximal growth rate constraint. *Quarterly Reviews of Biophysics*, *17*(1), 45-82.

Eyre-Walker, A. & Bulmer, M. (1995). Synonymous substitution rates in *Enterobacteria*. *Genetics*, *140*, 1407–1412.

Fairbrother, W. G., Holste, D., Burge, C. B., & Sharp, P. A. (2004). Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLOS Biology*, *2*, 268.

Fimmel, S., & Zouboulis, C. C. (2005). Influence of physiological androgen levels on wound healing and immune status in men. *Aging Male*, *8*(3-4), 166-174.

Fitch, D. H. A., & Strausbaugh, L. D. (1993). Low codon bias and high-rates of synonymous substitution in *Drosophila hydei* and *Drosophila melanogaster* histone genes. *Molecular Biology and Evolution*, *10*, 397-413.

Gack, M. U. (2014). Mechanisms of RIG-I-like receptor activation and manipulation by viral pathogens. *Journal of Virology*, *88*(10), 5213–5216.

Galtier, N., & Lobry. J. R. (1997). Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of Molecular Evolution*, *44*(6), 632-636.

Ginhoux, F., & Jung, S. (2014). Monocytes and macrophages: Developmental pathways and tissue homeostasis. *Nature Reviews Immunology*, *14*(6), 392-404.

Goldman, E., Alan, H. R., & William, F. S. (1995). Consecutive low usage leucine codons block translation only when near the 5' end of a message in *Escherichia coli*. *Journal of Molecular Biology*, *245*, 467-473.

Gouy, M., & Gautier, C. (1982). Codon usage in bacteria: Correlation with gene expressivity. *Nucleic Acids Research*, *10*(22), 7055-7074.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., & Pave, A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*, *8*, 49-62.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., & Mercier, R. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Research*, *9*, 43–74.

Grocock, R. J. & Sharp, P. M. (2002). Synonymous codon usage in *Pseudomonas aeruginosa* PA-O1. *Gene*, *289*, 131-139.

Gu, W., Zhou, T., Ma, J., Sun, X., & Lu, Z. (2004). Analysis of synonymous codon usage in *SARS Coronavirus* and other viruses in the *Nidovirales*. *Virus Research*, *101*(2), 155-161.

Gupta, S. K., Bhattacharyya, T. K., & Ghosh, T. C. (2004). Synonymous codon usage in *Lactococcus lactis*: Mutational bias versus translational selection. *Journal of Biomolecular Structure and Dynamics*, *21*(4), 527-536.

He, B., Dong, H., Jiang, C., Cao, F., Tao, S., & Xu, L. (2016). Analysis of codon usage patterns in *Ginkgo biloba* reveals codon usage tendency from A/U-ending to G/C-ending. *Scientific Reports*, *6*, 35927.

Henry, I. & Sharp, P. M. (2007). Predicting gene expression level from codon usage bias. *Molecular Biology and Evolution*, *24*, 10-12.

Hershberg, R., & Petrov, D.A. (2008). Selection on codon bias. *Annual Review of Genetics*, *42*, 287-299.

Higgs, P.G., & Ran, W. (2008). Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Molecular Biology and Evolution*, *25*, 2279-2291.

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., … Zamir, A. (1965). Structure of ribonucleic acid. *Science*, *147*, 1462–1465.

Hunt, R. (2009). Silent (synonymous) SNPs: Should we care about them? *Methods Molecular Biology*, *578*, 23-39.

Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *Journal of Molecular Biology*, *151*, 389–409.

Ikemura, T. (1982). Correlation between the abundance of yeast tRNAs and the occurrence of the respective codons in protein genes. *Journal of Molecular Biology*, *158*, 573–597.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution*, 2, 13–34.

Janeway, C. A., Jr. (2001). How the immune system works to protect the host from infection: a personal view. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(13), 7461-7468.

Janeway, C. A., & Medzhitov, R. (2002). Innate immune recognition. *Annual Review of Immunology*, *20*, 197–216.

Jenkins, G. M., & Holmes, E. C. (2003). The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Research*, *92*, 1-7.

Jolliffe, I. T. (2002). Principal component analysis. *Journal of the American Statistical Association*, *98*, 487.

Kanaya, S., Yamada, Y., Kudo, Y., & Ikemura, T. (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs. *Gene*, *238*, 143-155.

Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y., & Ikemura, T. (2001). Codon usage and tRNA genes in eukaryotes: Correlation of codon usage diversity with translation efficiency. *Journal of Molecular Evolution, 53*, 90–98.

Katsnelson, A. (2011). Breaking the silence. *Nature Medicine*, *17*, 1536–1538.

Kawai, T., & Akira, S. (2009). The roles of TLRs, RLRs and NLRs in pathogen recognition. *International Immunology, 21*(4), 317–337.

Kimchi-Sarfaty, C. (2013). Building better drugs: Developing and regulating engineered therapeutic proteins. *Trends in Pharmacological Science, 34,* 534–548.

Kliman, R. M., & J. Hey. (1993). Reduced natural-selection associated with low recombination in *Drosophila melanogaster*. *Molecular Biology and Evolution*, *10*, 1239-1258.

Kliman, R. M., & J. Hey. (1994). The effects of mutation and natural-selection on codon bias in the genes of *Drosophila*. *Genetics*, *137*, 1049-1056.

Knight, R. D., Freeland, S. J., & Landweber, L. F. (2001). A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology, 2*, 10.

Kotlar, D., & Lavner, Y. (2006). The action of selection on codon bias in the human genome is related to frequency, complexity, and chronology of amino acids. *BMC Genomics, 7*, 67.

Krishnaswamy, S., & Shanmugasundaram, S. (1995). Codon analysis of *cyanobacterial* genes. *Current Science, 69*, 182-185.

Kuby, J., Kindt, T. J., Goldsby, R. A., & Osborne, B. A. (2007). *Immunology* (6th ed.). New York: Freeman.

Kudla, G. (2009). Coding-sequence determinants of gene expression in *Escherichia coli*. *Science, 324*, 255–258.

Lavner, Y., & Kotlar, D. (2005). Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene, 345*(1): 127-138.

Li, J., Zhou, J., Wu, Y., Yang, S., & Tian, D. (2015). GC content of synonymous codons profoundly influences amino acid usage. *G3: Genes|Genomes|Genetics, 5*(10), 2027–2036.

Li, W. H., (1987). Models of nearly neutral mutations with particular implications for non-random usage of synonymous codons. *Journal of Molecular Evolution, 24*, 337-345.

Liu, X., Zhang, Y., Fang, Y., & Wang, Y. (2012). Patterns and influencing factor of synonymous codon usage in porcine *circovirus*. *Virology Journal, 9*, 68.

Lobry, J. R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular Biology and Evolution, 13*, 660–665.

Lobry, J. R., & D. Chessel. (2003). Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. *Journal of Applied Genetics, 44*, 235-261.

Lobry, J. R., & Necsulea, A. (2006). Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. *Gene, 385*, 128-136.

Lovmar, M., & Ehrenberg, M. (2006). Rate, accuracy and cost of ribosomes in bacterial cells. *Biochimie, 88*, 951–61.

Lu, H., Zhao, W. M., Zheng, Y., Wang, H., Qi, M., & Yu, X. P. (2005). Analysis of synonymous codon usage bias in *Chlamydi*a. *Acta Biochimica and Biophysica Sinica, 37*(1), 1-10.

Lynn, D. J., Singer, G. A., & Hickey, D. A. (2002). Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Research*, *30*(19), 4272-4277.

Ma, P., & Xia, X. (2011). Factors affecting splicing strength of yeast genes. *Comparative and Functional Genomics, 21*-146.

Marais, G., & Duret, L. (2001). Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. *Journal of Molecular Evolution, 52*(3), 275-280.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research, 18*(9), 1509-1517.

Martindale, D. W. (1989). Codon usage in *Tetrahymena* and other ciliates. *The Journal of Protozoology, 36*, 2934.

Martin-Galiano, A. J., Wells, J. M., & de la Campa, A. G. (2004). Relationship between codon biased genes, microarray expression values and physiological characteristics of *Streptococcus pneumoniae*. *Microbiology, 150*, 2313-2325.

Mauro, V. P., & Chappell, S. A. (2014). A critical analysis of codon optimization in human therapeutics. *Trends in Molecular Medicine, 20*(11), 604–613.

Mogensen, T. H. (2009). Pathogen recognition and inflammatory signaling in innate immune defenses. *Clinical Microbiology Reviews, 22*(2), 240–273.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods, 5*(7), 621-628.

Musto, H., Cruveiller, S., D'Onofrio, G., Romero, H., & Bernardi, G. (2001). Translational selection on codon usage in *Xenopus laevis*. *Molecular Biology and Evolution, 18*(9), 1703-1707.

Musto, H., Romero, H., & Zavala, A. (2003). Translational selection is operative for synonymous codon usage in *Clostridium perfringens* and *Clostridium acetobutylicum*, *Microbiology, 149*, 855.

Musto, H., Naya, H., Zavala, A., Romero, H., & Alvarez-Valin, F. (2004). Correlations between genomic GC levels and optimal growth temperatures in prokaryotes, *Federation of European Biochemical Societies, 573*, 73-77.

Muto, A., Yamao, F., & Kawauchi, Y. (1985). Codon usage in *Mycoplasma capricolum*. *Proceedings of the Japan Academy, 61*, 12-15.

Muto,A., & Osawa,S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. *Proceedings of the National Academy of Sciences of the United States of America, 84*, 166–169.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science, 320*(5881), 1344-1349.

Nair, R. R., Nandhini, M. B., Sethuraman, T., & Doss, G. (2013). Mutational pressure dictates synonymous codon usage in freshwater unicellular alpha -

cyanobacterial descendant *Paulinella chromatophora* and beta - *cyanobacterium Synechococcus elongatus* PCC6301. *Springerplus, 2*, 492.

Nakamura, Y., Wada, K., Wada, Y., Doi, H., & Kanaya, S. (1996). Codon usage tabulated from the international DNA sequence databases. *Nucleic Acids Research*, *24*, 214-215.

Nakamura, Y., Gojobori, T., & Ikemura, T. (2000). Codon usage tabulated from international DNA sequence databases. *Nucleic Acids Research*, 28(1), 292.

Nasrullah, I., Butt, A.M., Tahir, S., Idrees, M., & Tong, Y. (2015). Genomic analysis of codon usage shows influence of mutation pressure, natural selection, and host features on Marburg virus evolution. *BMC Evolutionary Biology, 15*, 174.

Naya, H., Romero, H., Carels, N., Zavala, A., & Musto, H. (2001). Translational selection shapes codon usage in the GC-rich genomes of *Chlamydomonas reinhardtii*. *Federation of European Biochemical Societies Letters, 501*(2-3), 127–130.

Naya, H., Romero, H., Zavala, A., Alvarez, B., & Musto, H. (2002). Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. *Journal of Molecular and Evolution, 55*, 260-264.

Newman, L. A., & Gold, P. E. (2016). Attenuation in rats of impairments of memory by scopolamine, a muscarinic receptor antagonist, by mecamylamine, a nicotinic receptor antagonist. *Psychopharmacology*, *233*(5), 925-932.

Novoa, E. M., Pavon-Eternod, M., Pan, T., & Ribas de Pouplana, L. (2012). A role for tRNA modifications in genome structure and codon usage. *Cell*, *149*(1), 202-213.

Osawa, S., & T. H. Jukes, (1989). Codon reassignment (codon capture) in evolution. *Journal of Molecular Evolution, 29*, 271-278.

Osawa, S., Muto, A., Jukes, T., & Ohama, T. (1990). Evolutionary changes in the genetic code. *Proceedings of the Royal Society of London Series B, 241,* 19-28.

Osawa, S., Jukes, T. H., Watanabe, K., & Muto, A. (1992). Recent evidence for evolution of the genetic code. *Microbiology Reviews, 56*, 229-264.

Palidwor, G. A., Perkins, T. J., & Xia, X. (2010). A general model of codon bias due to GC mutational bias. *PLOS ONE, 5*(10), 13431.

Pandit, A., & Sinha, S. (2011). Differential trends in the codon usage patterns in HIV-1 genes. *PLOS ONE, 6*(12), 28889.

Peden, J. F. (2000). *Analysis of codon usage.* Doctoral dissertation, University of Nottingham.

Piras, V., & Selvarajoo, K. (2014). Beyond MyD88 and TRIF pathways in toll-like receptor signaling. *Frontiers in Immunology, 5*.

Plotkin, J. B., Robins, H., & Levine, A. J. (2004). Tissue-specific codon usage and the expression of human genes. *Proceedings of the National Academy of Sciences of the United States of America, 101*(34), 12588-12591.

Plotkin, J. B., & Kudla, G. (2011). Synonymous but not the same: The causes and consequences of codon bias. *Nature Reviews Genetics, 12*(1), 32-42.

Qiu, H., Hildebrand, F., Kuraku, S., & Meyer, A. (2011). Unresolved orthology and peculiar coding sequence properties of lamprey genes: The KCNA gene family as test case. *BMC Genomics, 12*(1), 325.

Quax, T. E. F., Claassens, N. J., Söll, D., & van der Oost, J. (2015). Codon bias as a means to fine-tune gene expression. *Molecular Cell, 59*(2), 149–161.

Rao, Y., Wu, G., Wang, Z., Chai, X., Nie, Q., & Zhang, X. (2011). Mutation bias is the driving force of codon usage in the *Gallus gallus* genome. *DNA Research, 18*(6), 499-512.

Rocha, E. P. C. (2004). Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Research. 14*(11), 2279–2286.

Rocha, E. P. C., Maynard, S. J., Hurst, L. D., Holden, M. T., Cooper, J. E., Smith N. H., & Feil, E. (2006). Comparisons of dN/dS are time-dependent for closely related bacterial genomes. *Journal of Theoretical Biology, 239*, 226–235.

Rodriguez-Belmonte, E., Freire Picos, M. A., Rodriguez-Torres A.M., Gonzalez-Siso, M. I., & Cerdan, M. E. (1996). PICDI, a simple program for codon bias calculation. *Molecular Biotechnology, 5*, 191-195.

Rodríguez-Trelles, F., Tarrío, R., & Ayala, F. J. (2000). Fluctuating mutation bias and the evolution of base composition in Drosophila. *Journal of Molecular Evolution, 50*(1), 1-10.

Romero, H., Zavala, A., Musto, H., & Bernardi, G. (2003). The influence of translational selection on codon usage in fishes from the family *Cyprinidae*. *Gene, 317*(1-2), 141-147.

Roymondal, U., Das, S., & Sahoo, S. (2009). Predicting gene expression level from relative codon usage bias: An application to *Escherichia coli* genome. *DNA Research, 16*(1), 13-30.

Sauna, Z. E., & Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics, 12*, 683–691.

Se'mon, M., Lobry, J. R., & Duret, L. (2005). No evidence for tissue-specific adaptation of synonymous codon usage in humans. *Molecular Biology and Evolution, 23*, 1–7.

Shabalina, S.A. (2013). Sounds of silence: Synonymous nucleotides as a key to biological regulation and complexity. *Nucleic Acids Research, 41*, 2073–2094.

Sharp, P. M., Tuohy, T. M., & Mosurski, K. R. (1986). Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research, 14*(13), 5125–5143.

Sharp, P. M., & Li, W. H. (1987). The rate of synonymous substitution in *enterobacterial* genes is inversely related to codon usage bias. *Molecular Biology and Evolution, 4*, 222-230.

Sharp, P.M., & Li, W.H. (1987). The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research, 15*, 1281–1295.

Sharp, P. M., Cowe, E., Higgins, D. G., Shields, D. C., Wolfe, K. H., & Wright, F. (1988). Codon usage patterns in *E. coli, B. subtilis, S. cerevisiae, S. pombe, D. melanogaster* and *Homo sapiens*. *Nucleic Acids Research, 16*, 8207–8211.

Sharp, P. M., & Devine, K. M. (1989). Codon usage and gene-expression level in *Dictyostelium discoideum* - highly expressed genes do prefer optimal codons. *Nucleic Acids Research, 17*, 5029-5039.

Sharp, P. M. (1991). Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium* codon usage, map position, and concerted evolution. *Journal of Molecular Evolution, 33*, 23-33.

Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F., & Sockett, R. E. (2005). Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Research, 33*(4), 1141-53.

Sharp, P. M., Emery, L. R., & Zeng, K. (2010). Forces that influence the evolution of codon bias. *Philosophical Transactions of the Royal Society of London, 365*, 1203-1212.

Shields, D. C., & Sharp, P. M. (1987). Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Research, 15*, 8023-8040.

Singer, G. A. C., & Hickey, D.A. (2000). Nucleotide bias causes a genome wide bias in the amino acid composition of proteins. *Molecular Biology and Evolution, 17*(11), 1581–8.

Spencer, P. S. (2012). Silent substitutions predictably alter translation elongation rates and protein folding efficiencies. *Journal of Molecular Biology, 422*, 328–335.

Sueoka, N. (1962). On the genetic basis of variation and heterogeneity of DNA base composition. *Proceedings of the National Academy of Sciences of the United States of America, 48*, 582–591.

Sueoka, N. (1988). Directional mutation pressure and neutral molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America, 85*(8), 2653-7.

Sueoka, N. (1992). Directional mutation pressure, selective constraints, and genetic equilibria. *Journal of Molecular Evolution, 34*(2), 95–114.

Sueoka, N. (1995). Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *Journal of Molecular Evolution, 40*(3), 318–325.

Sueoka, N. (1999). Translation-coupled violation of Parity Rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position. *Gene, 238*(1), 53–58.

Sueoka, N. (2002). Wide intra-genomic G+C heterogeneity in human and chicken is mainly due to strand-symmetric directional mutation pressures. *Gene, 300*, 141–154.

Tatarinova, T. V., Alexandrov, N. N., Bouck, J. B., & Feldmann, K. A. (2010). GC3 biology in corn, rice, sorghum and other grasses. *BMC Genomics, 11*, 308.

Tizard, I. R. (2013). *Veterinary immunology* (9th ed.). St. Louis, Missouri: Elsevier/Saunders.

Tsai, C. J. (2008). Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. *Journal of Molecular Biology, 383*, 281–291.

U. S. Food & Drug Administration (2003). *Paving the way for personalized medicine: FDA's role in a new era of medical product development.* U.S. Department of Health and Human Services.

Osawa, S., & Jukes, T.H. (1989). Codon reassignment (codon capture) in evolution. *Journal of Molecular Evolution, 29*, 271-278.

Urrutia, A.O., & Hurst, L.D. (2001). Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics, 159*, 1191–1199.

Urrutia, A.O., & Hurst, L.D. (2003). The signature of selection mediated by expression on human genes. *Genome Research, 13*, 2260–2264.

Van Weringh, A., Ragonnet-Cronin, M., Pranckeviciene, E., Pavon-Eternod, M., Kleiman, L., & Xia, X. (2011). HIV-1 modulates the tRNA pool to improve translation efficiency. *Molecular Biology and Evolution, 28*, 1827–1834.

Varani, G., & McClain, W. H. (2000). The G·U wobble base pair: A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Reports, 1*(1), 18–23.

Vicario, S., Moriyama, E.N., & Powell, J.R. (2007). Codon usage in twelve species of *Drosophila*. *BMC Evolutionary Biology, 7*, 226.

Vinogradov, A. E. (2003). Isochores and tissue-specificity. *Nucleic Acids Research, 31*, 5212–5220.

Ward, N. J. (2011). Codon optimization of human factor VIII cDNAs leads to high-level expression. *Blood, 117*, 798–807.

Weber, M., Gawanbacht, A., Habjan, M., Rang, A., Borner, C., Schmidt, A. M., & Weber, F. (2013). Incoming RNA virus nucleocapsids containing a 5′ triphosphorylated genome activate RIG-I and antiviral signaling. *Cell Host and Microbe, 13*(3), 336–346.

Welch, M. (2009). Optimizing genes for protein expression. *Journal of the Royal Society, Interface the Royal Society, 6*(4), 467–476.

Wilke, C. O., & D. A. Drummond. (2006). Population genetics of translational robustness. *Genetics, 173*, 473–81.

Willenbrock, H., Friis, C., Juncker, A. S., & Ussery, D. W. (2006). An environmental signature for 323 microbial genomes based on codon adaptation indices. *Genome Biology, 7*(12), 114.

Willie, E., & Majewski, J. (2004). Evidence for codon bias selection at the pre-mRNA level in eukaryotes. *Trends in Genetics*, 20, 534–538

Wong, E. H., Smith, D. K., Rabadan, R., Peiris, M., & Poon, L. L. (2010). Codon usage bias and the evolution of influenza A viruses. Codon usage biases of influenza virus. *BMC Evolutionary Biology, 10*(1), 253.

Wright, F. (1990). The effective number of codons used in a gene. *Gene, 87*(1), 23–29.

Xia, X. (1996). Maximizing transcription efficiency causes codon usage bias. *Genetics, 144*, 1309–1320.

Xia, X. (1998). How optimized is the translational machinery in *Escherichia coli, Salmonella typhimurium* and *Saccharomyces cerevisiae*? *Genetics, 149*, 37–44.

Xia, X. (2007). An improved implementation of Codon Adaptation Index. *Evolutionary Bioinformatics, 3*, 53–58.

Xia, X. (2008). The cost of wobble translation in fungal mitochondrial genomes: Integration of two traditional hypotheses. *BMC Evolutionary Biology, 8*, 211.

Xu, C., Cai, X., Chen, Q., Zhou, H., Cai, Y., & Ben, A. (2011). Factors affecting synonymous codon usage bias in chloroplast genome of *Oncidium Gower Ramsey*. *Evolutionary Bioinformatics, 7*, 271-278.

Yamamoto, M. (2003). Role of adaptor TRIF in the MyD88-independent toll-like receptor signaling pathway. *Science, 301*(5633), 640–643.

Yang, Z., & Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular Biology and Evolution, 25*(3), 568-579.

Zavala, A., Naya, H., Romero, H., Sabbia, V., Piovani, R., & Musto, H. (2005). Genomic GC content prediction in prokaryotes from a sample of genes. *Gene, 357*(2), 137-143.

Zhang, F. (2010). Differential arginylation of actin isoforms is regulated by coding sequence-dependent degradation. *Science, 329*, 1534–1537.

Zhang, Y. M., Shao, Z. Q., Yang, L. T., Sun, X. Q., Mao, Y. F., & Chen, J. Q. (2013). Non-random arrangement of synonymous codons in archaea coding sequences. *Genomics, 101*(6), 362–367.

Zhong, F., Cao, W., Chan, E., Tay, P. N., Cahya, F. F., Zhang, H., & Lu, J. (2005). Deviation from major codons in the toll-like receptor genes is associated with low toll-like receptor expression. *Immunology, 114*(1), 83-93.

Zhou, J.H, (2013). The effects of the synonymous codon usage and tRNA abundance on protein folding of the 3C protease of foot and mouth disease virus. *Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases, 16*, 270–274.