



MASTERARBEIT | MASTER'S THESIS

Titel | Title

SNP distribution across different gene regions

verfasst von | submitted by
Ronit Chakraborty

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of
Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree
programme code as it appears on the
student record sheet:

UA 066 220

Studienrichtung lt. Studienblatt | Degree
programme as it appears on the student
record sheet:

Joint-Masterstudium Evolutionary Systems Biology

Betreut von | Supervisor:

Univ.-Prof. Mag. Dr. Pavel Kovarik

Acknowledgements

I extend my heartfelt appreciation to all those who have significantly contributed to my academic journey and played a crucial role in the successful completion of this MSc thesis. Their collective guidance, unwavering support, and invaluable assistance have left an indelible mark on this academic endeavour.

Foremost, I would like to express my deep gratitude to my supervisor, Pavel Kovarik, whose belief in my capabilities and the opportunity to conduct my Master's thesis in his lab have been pivotal. His consistent support, constructive feedback, and scholarly insights have been instrumental in shaping the trajectory of this research, and I am profoundly thankful for the guidance he provided.

I would also like to thank Prof. Arndt von Haeseler for his guidance throughout the Master's thesis and Fritz Sedlazeck for providing his important pointers and suggestions regarding the database.

I also want to thank all of the Kokonuts, both past and present: Annika for her challenging questions and comments during our WiP meetings (sorry for nodding off during your presentations!), Jeanne for her assistance with various thesis approaches and general advice regarding RNA-Seq analysis, Julius for being the ultimate lab buddy, and the rest of the group, which includes Mauri, Nora, Christina, Kevin, Lynn, and Liuba, for being steadfast rock stars who have greatly contributed to the lab's collaborative atmosphere.

I am deeply indebted to my family—my mother, Rinku Chakraborty; my father, Sameer Ranjan Chakraborty; and my brother, Saurabh Chakraborty—for their enduring support throughout my life. Their constant encouragement and presence have formed the bedrock of my journey, and I acknowledge with gratitude that reaching this point would not have been possible without them.

Gratitude extends to my friends who have been integral to this academic ride. Sahil Narwal, for always being there for me when I needed him; Mukesh Tiwari, for providing moral support; and my school friends, Yash Agrawal and Siddhartha Mukherjee, for infusing my school life with joy and camaraderie. Special mention goes to Mayra, whose friendship in Vienna has added a delightful dimension to my life. Lastly, I would be remiss if I did not offer my sincere thanks to Nataša Simić for being a pillar of strength during challenging times.

Abstract

Single Nucleotide Polymorphisms (SNPs) are genetic variations that occur when a single nucleotide at a specific position in the DNA sequence differs between individuals in more than 1% of the individuals. SNPs play a crucial role in disease susceptibility and evolutionary adaptation and understanding the distribution of SNPs across gene regions is essential for deciphering the genetic basis of these phenomena. In this Master's thesis, I present the generation of novel SNP count databases for the 5' Untranslated Region (5'UTR), Coding Sequence (CDS), and 3' Untranslated Region (3'UTR) regions, aiming to elucidate SNP distribution patterns and their potential implications. Leveraging publicly available genomic datasets such as the 1000 Genomes Project, I conducted a comprehensive analysis of SNP counts across different gene regions, revealing notable disparities and trends. While the CDS exhibited significantly higher SNP counts compared to the 5'UTR and 3'UTR, normalization techniques were employed to mitigate bias and facilitate comparative analysis. Despite challenges posed by normalization bias and atypical annotations, our database serves as a foundational resource for comparative SNP analysis within transcripts and gene types. Comparative analysis of housekeeping, cytokine, and chemokine genes unveiled distinct SNP patterns, that might potentially reflect their specific roles in immune response and inflammation. This database will aid in delving deeper into the effects of genetic variation on gene function and phenotype, offering a foundational platform for exploring SNP distributions and their correlations with genetic elements, thus advancing our understanding of genetic diversity and its implications for human health and evolution.

Zusammenfassung

Einzelnukleotid-Polymorphismen (SNPs) sind genetische Variationen, die auftreten, wenn sich ein einzelnes Nukleotid an einer bestimmten Position in der DNA-Sequenz bei mehr als 1 % der Individuen unterscheidet. SNPs spielen eine entscheidende Rolle bei der Krankheitsanfälligkeit und der evolutionären Anpassung, und das Verständnis der Verteilung von SNPs in verschiedenen Genregionen ist für die Entschlüsselung der genetischen Grundlagen dieser Phänomene unerlässlich. In dieser Masterarbeit präsentiere ich die Erstellung neuartiger SNP-Zählungsdatenbanken für die 5'-untranslatierte Region (5'UTR), die kodierende Sequenz (CDS) und die 3'-untranslatierte Region (3'UTR) mit dem Ziel, die SNP-Verteilungsmuster und ihre potenziellen Auswirkungen zu ergründen. Mit Hilfe öffentlich zugänglicher Genomdaten wie dem 1000 Genomes Project habe ich eine umfassende Analyse der SNP-Anzahl in den verschiedenen Genregionen durchgeführt und dabei bemerkenswerte Unterschiede und Trends aufgedeckt. Während der CDS im Vergleich zum 5'UTR und 3'UTR eine signifikant höhere SNP-Anzahl aufwies, wurden Normalisierungstechniken eingesetzt, um Verzerrungen abzuschwächen und vergleichende Analysen zu erleichtern. Trotz der Herausforderungen, die sich durch Normalisierungsfehler und atypische Annotationen ergeben, dient unsere Datenbank als grundlegende Ressource für vergleichende SNP-Analysen innerhalb von Transkripten und Gentypen. Die vergleichende Analyse von Housekeeping-, Zytokin- und Chemokin-Genen ergab unterschiedliche SNP-Muster, die ihre Rolle bei Immunreaktionen und Entzündungen widerspiegeln. Diese Datenbank wird dazu beitragen, die Auswirkungen genetischer Variationen auf die Genfunktion und den Phänotyp zu vertiefen, und bietet eine grundlegende Plattform für die Erforschung der SNP-Verteilungen und ihrer Korrelationen mit genetischen Elementen, wodurch unser Verständnis der genetischen Vielfalt und ihrer Auswirkungen auf die menschliche Gesundheit und Evolution verbessert wird.

Table of Contents

Acknowledgements.....	1
Abstract.....	3
Zusammenfassung	4
1. Introduction.....	10
1.1 Genes and Genetic Regions.....	10
1.2 Transcription.....	11
1.3 Messenger RNA (mRNA)	13
1.3.1 Structure of mRNA.....	13
1.3.2 Transcript Diversity	14
1.4 Regulation of mRNA stability	15
1.4.1 Adenine-Uridine (AU)-rich elements (AREs).....	15
1.5 Role of mRNA Half-Life in Housekeeping, Cytokine, and Chemokine Gene Expression Dynamics	16
1.5.1 Housekeeping Genes	16
1.5.2 Cytokine and Chemokine Genes.....	16
1.6 Mutations	17
1.6.1 Point Mutations.....	17
1.6.2 Insertions and Deletions.....	17
1.6.3 Duplications and Translocations.....	18
1.7 Single Nucleotide Polymorphisms (SNPs)	19
1.7.1 SNP role in diseases	19
1.8 Conserved Sequences.....	20
1.9 1000 Genomes Project.....	20

1.10	Variant Call Format (VCF).....	21
1.11	BioMart	21
1.12	Aims of the Master's Thesis	22
2.	Materials and Methods	23
2.1	Data.....	23
2.2	VCF Filtration.....	23
2.3	SNP Database Generation.....	23
2.4	SNP counts in Housekeeping, Cytokine and Chemokine genes.....	24
3.	Results	26
3.1	SNP Database for 3'UTR.....	26
3.2	Total number of normalized SNP counts is the highest in CDS as compared to the 5'UTR and 3'UTR....	28
3.3	SNP distribution in Housekeeping, Cytokine and Chemokine genes.....	30
4.	Discussions	35
4.1	Effect of Normalisation on SNP Counts	35
4.2	Effect of Conserved Sequences on Housekeeping, Cytokine and Chemokine genes	36
5.	Conclusions	37
6.	References	38
	Supplementary Materials	51

List of Tables

Table 1: Snippet of the final database for 3'UTR	26
---	----

List of Figures

Figure 1: The structure of a gene in eukaryotes.....	11
Figure 2: Transcription involves the synthesis of mRNA	12
Figure 3: Schematic diagram of eukaryotic mRNA	15
Figure 4: Types of Mutations	18
Figure 5: Example of a valid VCF	21
Figure 6: Bar Plot showing the normalised SNP counts	28
Figure 7: Bar Plot showing the number of transcripts for each gene region	29
Figure 8: A histogram showing 3'UTR length across all the transcripts within genes	30
Figure 9: Bar Plot showing the SNP counts for Housekeeping, Cytokine and Chemokine genes within the two bins within the 3'UTR	31
Figure 10: Bar Plot showing the number of genes with SNP counts for Housekeeping, Cytokine and Chemokine genes	32
Figure 11: Bar Plot showing the SNPs per Transcript ratio for Housekeeping, Cytokine and Chemokine genes within the two bins within the 3'UTR	33
Figure 12: Bar Plot showing the SNPs per Gene ratio for Housekeeping, Cytokine and Chemokine genes within the two bins within the 3'UTR	34

List of Supplementary Materials

Supplementary Table 1: Snippet of the Gene IDs for Housekeeping genes	51
Supplementary Table 2: List of Gene IDs for Cytokine genes	52
Supplementary Table 3: List of Gene IDs for Chemokine genes	53
Supplementary Table 4: Snippet of the final database for 5'UTR	54
Supplementary Table 5: Snippet of the final database for CDS	55
Supplementary Table 6: Snippet of the final database for Housekeeping Genes	56
Supplementary Table 7: Final database for Cytokine Genes	57
Supplementary Table 8: Final database for Chemokine Genes	58

1. Introduction

1.1 Genes and Genetic Regions

Genes are the fundamental units of heredity that encode the instructions that guide the development, function, and regulation of living organisms (Falk, 1986; Pearson, 2006). In the genome, genes are not uniform entities but are composed of distinct regions, that play a unique role in orchestrating cellular processes. Understanding the architecture and function of these genetic regions is paramount in unravelling the complexities of molecular biology.

At its core, a gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products (Gerstein et al., 2007). However, the linear Deoxyribonucleic Acid (DNA) sequence, constituting a gene, encompasses various segments with diverse functions which collectively contribute to the intricacies of genetic regulation. These regions include protein-coding regions (exons), non-coding regions (introns), regulatory elements (promoters, enhancers), and intergenic regions as described in **Figure 1** (Yip et al., 2012).

Promoters: Promoters are positioned upstream of a gene and act as docking sites for RNA polymerase, the enzyme responsible for initiating transcription. Promoters play a crucial role in regulating the timing and quantity of gene transcription (Smale & Kadonaga, 2003).

Enhancers: Enhancers are distinct from promoters; they are regulatory elements that can be located far from the gene they regulate. They modulate gene expression by interacting with promoters, facilitating the binding of transcription factors, and influencing the rate of transcription (Serfling et al., 1985).

Exons and Introns: The protein-coding regions of a gene are defined as exons, while non-coding regions represent introns. During transcription and subsequent RNA processing, introns are typically removed, and exons are spliced together, forming a mature mRNA transcript for translation (Berget et al., 1977; Chow et al., 1977).

Termination Signals: Termination Signals are positioned at the end of a gene. Termination signals mark the conclusion of transcription, they guide the detachment of RNA polymerase and ensure the proper termination of the mRNA molecule (Zerbini et al., 2014).

Importantly, genetic information is encoded as DNA. According to the biological dogma (CRICK, 1970), DNA is translated into RNA through transcription and then translated into proteins, converting genetic information into vital biological functions.

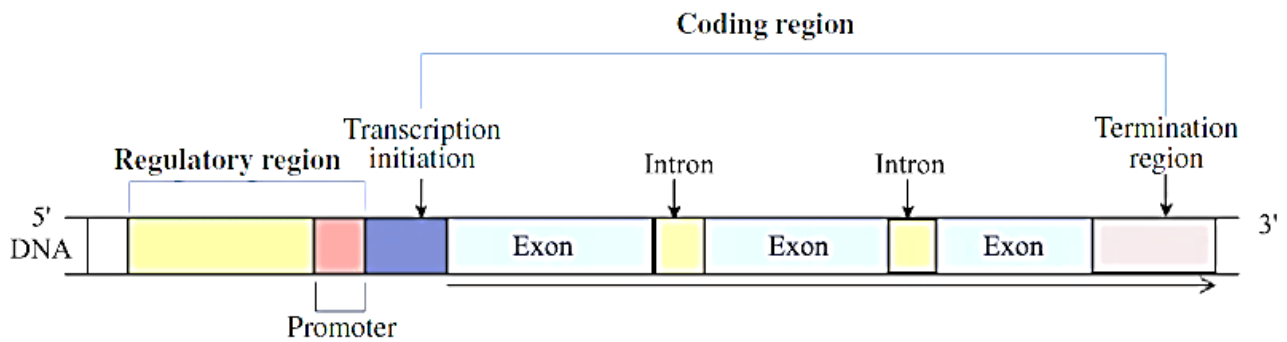


Figure 1: The structure of a gene in eukaryotes. The figure was adapted and modified from Transgenic Plants, Chapter 8 (Zerbini et al., 2014).

1.2 Transcription

As a key mechanism in molecular biology, transcription acts as a link between the functional molecules that power cellular functions and the genetic information contained in DNA. This complex process takes place in the nucleus of eukaryotic cells, directing the synthesis of RNA from a DNA blueprint (**Figure 2**). A group of molecular regulators control the accuracy and intricacy of transcription. The key enzymes of transcription are DNA-dependent RNA polymerases, which synthesise RNA strands that accurately represent the genetic code (Sentenac, 1985).

RNA polymerase is a multi-subunit enzyme with distinct functional domains that regulate transcription (Burgess, 1971). In eukaryotes, there are three major RNA polymerases—RNA polymerase I, II, and III—each dedicated to transcribing specific types of genes (ROEDER & RUTTER, 1969). RNA polymerase II (RNAPII), in particular, is responsible for transcribing protein-coding genes, making it a central player in the synthesis of messenger RNA (mRNA) (Orphanides et al., 1996; Young, 1991).

Transcription initiation is finely controlled through the interaction of regulatory elements within the DNA and RNAPII. The promoter region, a DNA sequence that serves as the RNAPII transcription start site, is at the centre of this initiation process (Fuda et al., 2009). The start of RNA synthesis is signalled by the transcription initiation complex (Plaschka et al., 2016).

Transcription is organised into defined steps. After binding to the promoter, RNA polymerase unwinds double-stranded DNA by its helicase activity, exposing a short stretch of the template strand (Kuhn et al., 1979; Lohman & Fazio, 2018). This initiation step is marked by the synthesis of a short RNA molecule, known as the nascent RNA transcript, which serves as a primer for elongation (Neugebauer, 2019; Wissink et al., 2019).

Elongation, the subsequent phase of transcription, involves the progressive movement of RNA polymerase along the DNA template, synthesizing an RNA strand complementary to the transcribed DNA strand (N. A. Hoffmann et al., 2015; Neyer et al., 2016; Sidorenkov et al., 1998). This process is regulated by factors such as nucleosome positioning and modifications of histone proteins, which impact the accessibility of the DNA template (Fuchs et al., 2009; Hartzog & Tamkun, 2007; Schones et al., 2008).

During elongation, RNA polymerase II undergoes conformational changes, allowing for the formation of the transcription bubble and the advancement of the transcription bubble along the DNA template (Gnatt et al., 2001; Nudler, 1999; Schwinghammer et al., 2013; Vassylyev et al., 2007).

Transcription termination marks the end of RNA synthesis and is governed by distinct mechanisms in eukaryotes (Porrua & Libri, 2015). In eukaryotes, termination involves polyadenylation signals and the recognition of polyadenylation sites by cleavage and polyadenylation factors. These processes ensure the accurate and regulated conclusion of transcription (BIRNSTIEL et al., 1985; Platt, 1986; Xie et al., 2023).

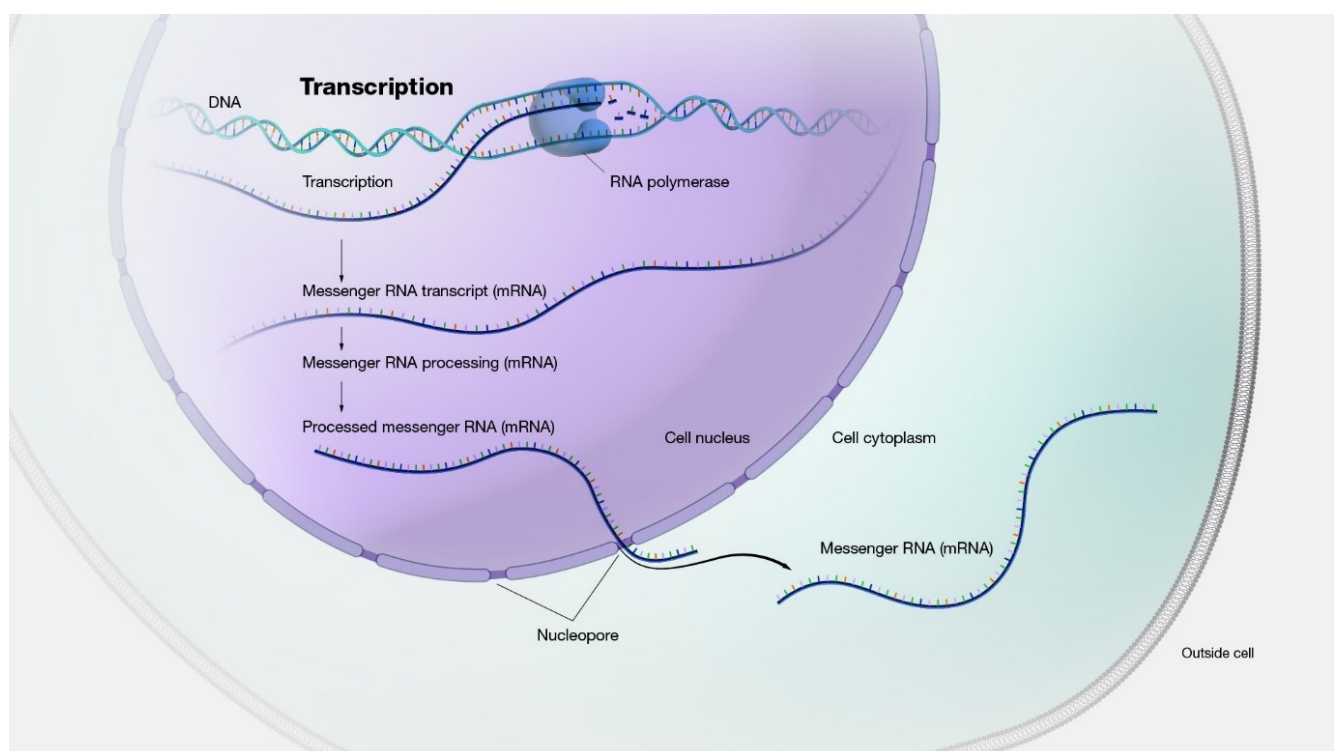


Figure 2: Transcription involves the synthesis of mRNA, essential for protein synthesis and various cellular functions (Courtesy: National Human Genome Research Institute, www.genome.gov).

1.3 Messenger RNA (mRNA)

Messenger RNA (mRNA) stands as a pivotal intermediary in the transfer of genetic information from DNA to functional proteins within the cell. Structurally dynamic and functionally versatile, mRNA molecules exhibit a myriad of features essential for their role in protein synthesis and cellular regulation (**Figure 3**). Its functions include serving as a template for translation by ribosomes (Meyer, 2017; Proshkin et al., 2010) and actively participating in gene expression regulation through post-transcriptional modifications (Zhao et al., 2017), mRNA localisation (Parton et al., 2014), and interactions with regulatory molecules like RNA-binding proteins (Glisovic et al., 2008) and non-coding RNAs (Karapetyan et al., 2013).

1.3.1 Structure of mRNA

5' Cap Structure: Positioned at the 5' terminus of mRNA, the 5' cap structure consists of a dinucleoside 5',5'-triphosphate moiety where the terminal nucleoside is a 7-methylguanosine (Mikkola et al., 2005). This cap structure serves as a recognition motif for translation initiation factors, enhances mRNA stability by protecting against exonuclease degradation, and facilitates the nuclear export of mRNA (Busch, 1976; Mikkola et al., 2005).

5' Untranslated Region (UTR): Immediately following the 5' cap is the 5' UTR, a non-coding region that precedes the initiation codon (start codon) of the open reading frame (ORF) (Araujo et al., 2012). The 5' UTR harbours regulatory elements involved in translational regulation, such as internal ribosome entry sites (IRES) and upstream open reading frames (uORFs) (Leppek et al., 2018; van der Velden & Thomas, 1999).

Coding Sequences (CDS): The Coding Sequence (CDS) in mRNA comprises the region bounded by start and stop codons, directing protein synthesis during translation (Tuller & Zur, 2015). Beginning with the start codon (usually AUG) (MEIJER & THOMAS, 2002) and concluding with a stop codon (UAA, UAG, or UGA), the CDS dictates the amino acid sequence of its encoding protein (E. K. Lee & Gorospe, 2011).

Selective pressure dictates a low mutation rate within coding gene sequences, driven by the imperative to maintain protein functionality and structure (Oman et al., 2022; Wolfe et al., 1989). Even slight changes in amino acid sequences can have detrimental effects, disrupting protein folding, interactions, and enzymatic activity, thereby compromising organism fitness (Parmley & Hurst, 2007; Rodrigue et al., 2010; Sarkar et al., 2022). Natural selection acts to minimize mutations within coding regions to preserve genomic stability and ensure the presence of functional proteins crucial for organism survival and adaptation (Castle, 2011; Harris et al., 1993).

3' Untranslated Region (UTR): Following the CDS is the 3' UTR, which contains cis-acting regulatory elements involved in post-transcriptional processes such as mRNA stability, localization, and translational control (KISLAUSKIS & SINGER, 1992; Moore, 2005). The 3' UTR is often enriched in microRNA (miRNA) target sites, Iron-responsive element (IRE) and AU-rich elements (AREs), which mediate mRNA degradation and translational repression (Barreau, 2005; Bartel, 2009; Chen & Shyu, 1995; Kontoyiannis et al., 1999; La Rocca et al., 2015).

Polyadenylation Tail: At the 3' end of mRNA lies the polyadenylation tail, consisting of a string of adenine residues (poly(A) tail) (Darnell et al., 1971; Edmonds et al., 1971; S. Y. Lee et al., 1971) added post-transcriptionally by the polyadenylation machinery (Proudfoot, 2011; Tian & Graber, 2012). The poly(A) tail contributes to mRNA stability, enhances translational efficiency, and facilitates mRNA export from the nucleus (Goldstrohm & Wickens, 2008; Jackson & Standart, 1990).

1.3.2 Transcript Diversity

A hallmark feature of mRNA is its capacity for generating transcriptomic diversity through alternative splicing and other post-transcriptional modifications (Matlin et al., 2005). Alternative splicing, a prevalent mechanism in eukaryotic gene expression, enables the production of multiple mRNA isoforms from a single gene locus by differential exon inclusion or exclusion (Kelemen et al., 2013; Stamm et al., 2005). This process greatly expands the repertoire of protein isoforms and functional diversity encoded by the genome (Smith et al., 1989).

Additionally, mRNA transcripts can undergo alternative polyadenylation, alternative transcription start site usage (Tian & Manley, 2017), RNA editing (Simpson & Emeson, 1996), and other post-transcriptional modifications (Zhao et al., 2017), further enhancing transcriptomic complexity and functional specialization.

Transcript diversity is vital as it amplifies the functional possibilities of the genome, allowing for dynamic responses to environmental cues and developmental signals. This diversity underpins crucial biological processes such as tissue specialization, cellular differentiation, and disease mechanisms, emphasizing its importance in comprehending the intricacies of living systems.

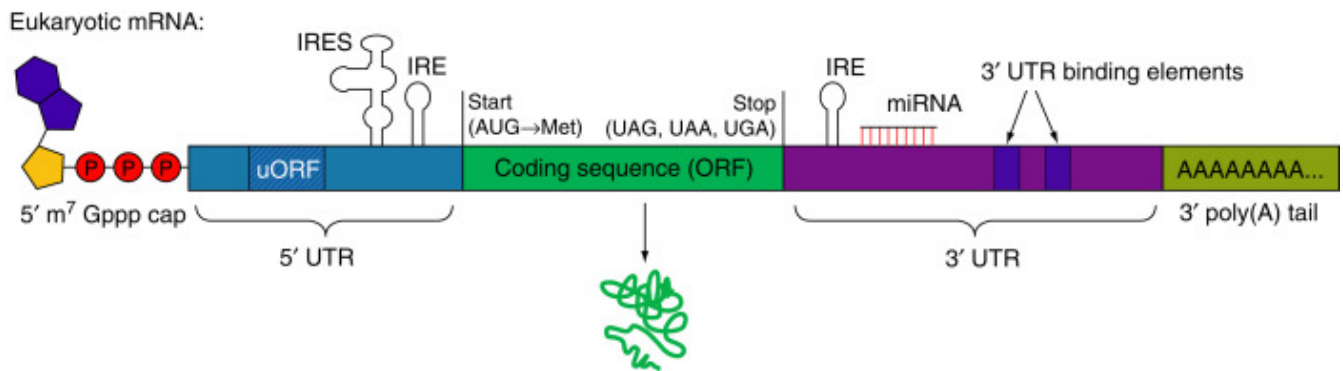


Figure 3: Schematic diagram of eukaryotic mRNA. The 5'UTR starts with a 5' cap m⁷ Gppp structure. The IRES, IRE and upstream Open Reading Frame (uORF) are located within the 5'UTR. CDS houses the codon triplets that are translated by tRNA into proteins, flanked by a start codon and a stop codon. The 3'UTR houses the AREs and miRNA that are responsible for the regulation of the gene and terminates with a Poly-A tail. The figure was adapted and modified from Messenger RNA (mRNA): The Link between DNA and Protein (Goss & Domashevskiy, 2016).

1.4 Regulation of mRNA stability

In addition to carrying genetic information from DNA to proteins, mRNA molecules host regulatory regions that control translational efficiency, stability, and localization (Guhaniyogi & Brewer, 2001; Wu & Brewer, 2012). These motifs, like AU-rich elements (AREs), internal ribosome entry sites (IRESs), and microRNA (miRNA) binding sites (Barreau, 2005; Bartel, 2009; Chen & Shyu, 1995; Kontoyiannis et al., 1999; La Rocca et al., 2015; Leppek et al., 2018; van der Velden & Thomas, 1999), orchestrate post-transcriptional gene regulation, thus shaping gene expression in response to cellular signals and environmental cues. These elements play pivotal roles in maintaining cellular balance and adapting to changing conditions (Guhaniyogi & Brewer, 2001; Mitchell & Tollervey, 2000; Ross, 1995).

1.4.1 Adenine-Uridine (AU)-rich elements (AREs)

AU-rich elements represent a prevalent class of cis-acting regulatory motifs of 50-150 nucleotides (Barreau, 2005) enriched in the 3' UTRs of mRNA transcripts (Chen & Shyu, 1995; Wilson & Brewer, 1998). Characterized by high adenine (A) and uracil (U) content elements, AREs exert a profound influence over mRNA stability and decay, playing a pivotal role in post-transcriptional gene regulation (Audic & Hartley, 2004; Crawford et al., 1997; Schiavi, 1992). ARE-mediated modulation of mRNA turnover involves interactions with RNA-binding proteins (RBPs), including members of the AU-rich element binding protein (ARE-BP) family such as AUF1/hnRNP D, tristetraprolin (TTP), butyrate-regulated factor-1 (BRF1) and KH domain-splicing regulatory protein (KSRP) among others (Wu & Brewer, 2012), which govern mRNA decay pathways and influence transcript fate.

The dynamic interplay between AREs and RBPs dictates the temporal and spatial regulation of gene expression, enabling rapid adjustments in mRNA abundance in response to cellular stimuli and environmental stressors (Kovarik et al., 2017, 2021; Sidali et al., 2021).

1.5 Role of mRNA Half-Life in Housekeeping, Cytokine, and Chemokine Gene Expression Dynamics

mRNA half-life, a critical determinant of gene expression dynamics, plays a pivotal role in shaping the temporal patterns of mRNA abundance and protein synthesis (Ross, 1995). Across different gene categories, including housekeeping, cytokine, and chemokine genes, variations in mRNA half-life contribute to the nuanced regulation of cellular processes, immune responses, and inflammatory pathways (Hargrove & Schmidt, 1989).

1.5.1 Housekeeping Genes

Housekeeping genes are defined as those genes that are “always expressed” (Watson JD et al., 1987) in every tissue to maintain cellular functions and often exhibit relatively stable mRNA levels and extended half-lives (Eisenberg & Levanon, 2013). With a median half-life of around 5 - 7 hours (Friedel et al., 2009; Sharova et al., 2009), housekeeping genes exhibit robust transcriptional and post-transcriptional regulatory mechanisms that confer stability to their transcripts.

Housekeeping genes facilitate very basic cellular functions such as transcription, translation, cellular strength and integrity, ensuring the availability of essential cellular components under steady-state conditions (Joshi et al., 2022). The sustained expression of housekeeping genes ensures the continuous production of essential proteins required for cellular respiration and energy metabolism (BUTTE et al., 2001; Eisenberg & Levanon, 2013; Krainer, 2005; WARRINGTON et al., 2000; Zhu et al., 2008).

1.5.2 Cytokine and Chemokine Genes

Cytokine and chemokine genes are fundamental components of the immune system, playing essential roles in regulating immune responses, orchestrating inflammation, and facilitating cellular communication (Ollier, 2004). These genes exhibit a wide array of functions and are subject to various regulatory mechanisms (Kovarik et al., 2017). Cytokines, a class of small signalling proteins, are secreted by diverse cell types, including immune cells, and participate in various physiological processes such as cell proliferation, differentiation, and activation of immune cells (J.-M. Zhang & An, 2007). Chemokines, which represent a subset of cytokines, predominantly regulate the migration and trafficking of various cells to specific sites of inflammation or infection, contributing significantly to the immune surveillance and defence mechanisms (Zlotnik et al., 2006).

Notably, the mRNAs of cytokine and chemokine genes often display very short half-lives due to rapid and highly regulated degradation processes (Whiteside, 1994); the half-lives may range from 18 minutes in the case of TNF to 15.5 hours in the case of IL-6 *in vivo* (Liu et al., 2021). This differential stability of mRNA transcripts can impact the temporal expression patterns of cytokines and chemokines, influencing the duration and magnitude of immune responses (Cameron & Kelvin, 2003).

1.6 Mutations

Mutations are termed as any rare change in the nucleotide sequence, usually but not always with a disease-causing attribute (Condit et al., 2002). This change in the nucleotide sequence may or may not cause phenotypic changes (Karki et al., 2015). Both inherited (germline) and acquired (somatic) mutations can occur; somatic mutations are acquired during an individual's lifetime and are the primary cause of human disorders such as cancer. They can also change amino acid sequences and in the event of a Copy Number Variation mutation, the expression of a gene is amplified (or reduced) through increased (or decreased) copy number of a locus allele (Freeman et al., 2006; F. Zhang et al., 2009).

In the gametes, germline alterations take place. Germline mutations from the parents may also exist in the progeny cells because every nucleated cell is first created by the union of an egg and a sperm. Replication mistakes, mobile genetic elements, or unrepaired DNA damage are the usual causes of mutations (Karki et al., 2015).

Mutations occur due to errors in DNA replication (Crow, 2000), environmental exposures (Kondrashov Alexey S. & Houle David, 1994), or mutagenic agents (Ling & Robinson, 1997). Despite often being associated with negative effects, mutations also drive genetic diversity and evolution. Beneficial mutations can confer advantages like enhanced adaptation, pathogen resistance, or improved physiological functions, contributing to the survival and adaptation of organisms (Foster, 2000; Galtier, 2014; Lynch, 2010).

DNA mutations fall into several main categories (**Figure 4**).

1.6.1 Point Mutations

Point mutations, the most prevalent type of genetic alteration (Copley, 2010), involve the substitution of a single nucleotide base with another, leading to sequence variations within the DNA molecule (Brown TA, 2004). Point mutations encompass transitions, where a purine base (adenine or guanine) is replaced by another purine, or a pyrimidine base (thymine or cytosine) is replaced by another pyrimidine, as well as transversions, involving the substitution of a purine base with a pyrimidine or vice versa.

1.6.2 Insertions and Deletions

Genetic changes known as insertions and deletions, which are characterised by the addition or deletion of nucleotide bases within the DNA sequence, are frequently brought about via a process known as replication slippage, sometimes referred to as slipped strand mispairing (Savino et al., 2022; Sehn, 2015). These mutational events can vary in size from single base pair alterations to extensive genomic rearrangements spanning approximately 1 kb (Sehn, 2015).

1.6.3 Duplications and Translocations

Duplications involve the replication of genomic segments, leading to the presence of multiple copies of specific DNA sequences within the genome (Mendivil Ramos & Ferrier, 2012; Reams & Roth, 2015). Translocations, on the other hand, entail the exchange of chromosomal segments between non-homologous chromosomes, resulting in the relocation of genetic material and potential disruption of gene regulation or chromosomal stability (Fraser et al., 2005; Mendivil Ramos & Ferrier, 2012).

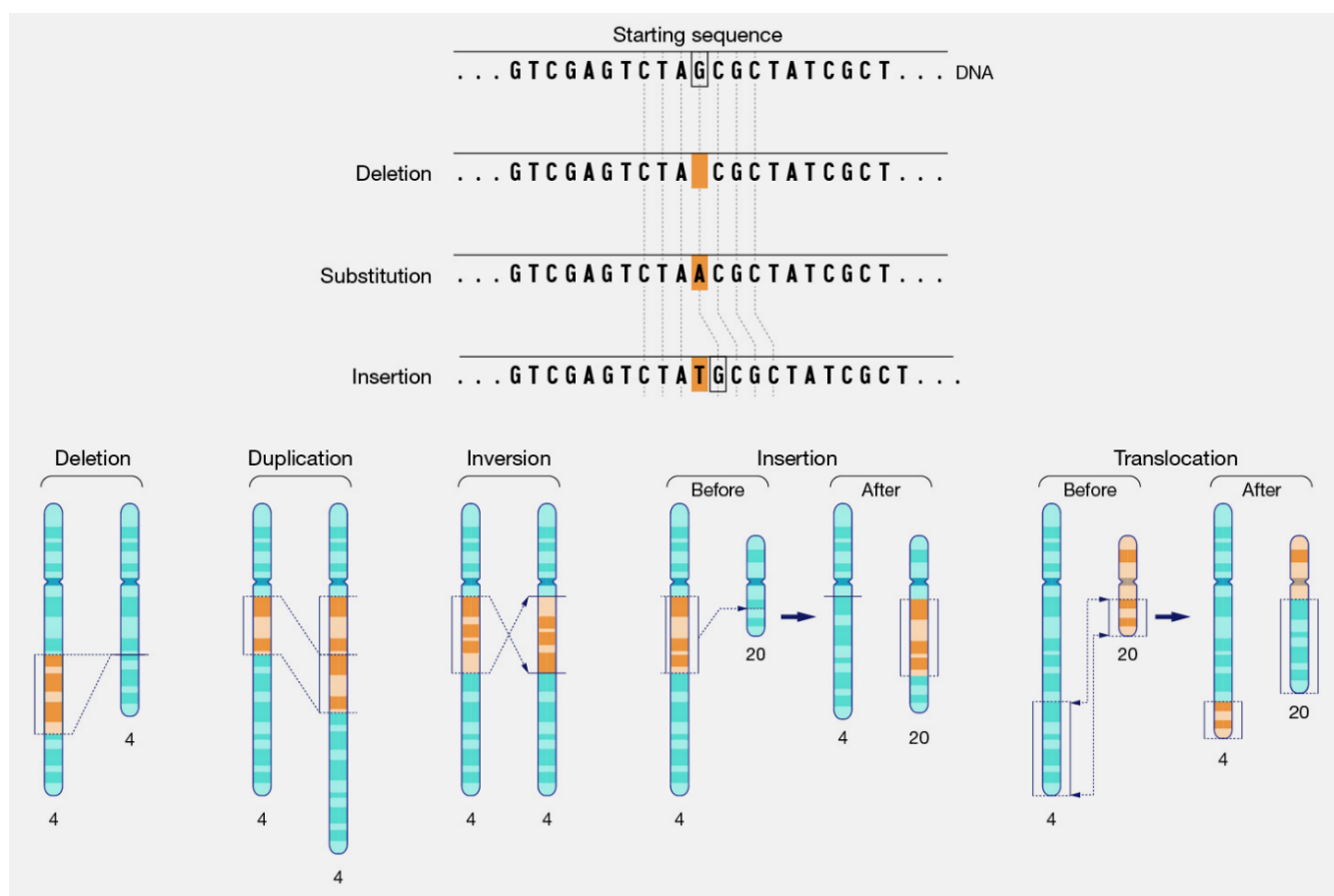


Figure 4: Types of Mutations (Courtesy: National Human Genome Research Institute, www.genome.gov).

1.7 Single Nucleotide Polymorphisms (SNPs)

Single nucleotide polymorphisms, or SNPs, are the most prevalent type of genetic variation across individuals within a group (Brookes, 1999). They are believed to occur every 500–1000 base pairs throughout the human genome (Sherry et al., 1999) and account for around 90% of all sequence variants in humans (Collins et al., 1998). These genetic variations occur when a single nucleotide at a specific position in the DNA sequence differs between individuals. SNPs are defined as variations present in more than 1% (Barnes, 2002; Brookes, 1999) of the human population while variations occurring in less than 1% of the population are described as Single Nucleotide Variations (SNVs). Together, these Single Nucleotide Mutations influence the risk for the development and the outcome of disease via various mechanisms, with around 10% of all SNPs estimated to have the potential to alter some biological process (Wjst, 2004).

1.7.1 SNP role in diseases

Understanding the distribution of SNPs across different gene regions is crucial for unravelling the complex relationship between genetic variations and their functional consequences. The presence of SNPs within these regions has diverse effects on gene function (Shastry, 2009).

SNPs within coding regions are associated with septic complications in infections, inflammatory diseases, metabolic disorders, and cancers (Meng et al., 2021). In contrast, nonsynonymous SNPs may result in amino acid exchanges, potentially altering the structure and biological function of the encoded protein (Gu et al., 2015). Specifically, SNPs found in interferons and the cytokines IL1, IL4, IL6, IL7, IL10, and IL17 have been linked to the initiation of severe cytokine storms, which contribute to respiratory complications in coronavirus disease-19 (COVID-19) (Vakil et al., 2022).

In addition to their impact on protein-coding regions, SNPs occurring within the 3' untranslated region (3'UTR) of genes have garnered significant attention due to their potential functional effects such as Alternative Polyadenylation (Shulman & Elkon, 2020). SNPs within the 3'UTR can influence gene expression by altering mRNA stability, mRNA-protein interactions, microRNA (miRNA) binding sites, and RNA secondary structure (Chan et al., 2023; Haas et al., 2012; Hu & Bruno, 2011; Rykova et al., 2022). These variations can have diverse functional consequences, including modulation of gene expression levels, alternative splicing, and susceptibility to various diseases (Akdeli et al., 2014; Roy & Mallick, 2017; Shulman & Elkon, 2020). Additional research on 3'UTR polymorphisms reveals variations that interfere with TRIM4's mRNA regulation, while another SNP modifies PILRB abundance and causes transcriptional alterations associated with age-related macular degeneration (Griesemer et al., 2021). Similarly, AREs were impacted by a functional polymorphism in the 3'UTR of IFNL3 mRNA, which changed how HCV-induced microRNA bound to the AREs during infection (McFarland et al., 2014).

1.8 Conserved Sequences

Conserved sequences within genes, including cytokine and chemokine genes, represent evolutionarily preserved nucleotide sequences that are maintained across different species (Stojanovic, 1999). These sequences signify crucial functional elements essential for gene regulation, protein structure, and biological activity (Petrokovski et al., 1998). Sustained through evolutionary pressures and selective constraints, these genomic segments encompass regulatory motifs, protein-coding domains, and structural motifs vital for maintaining gene expression fidelity, ensuring protein functionality, and coordinating intercellular signalling cascades (D. Hoffmann et al., 2020; Karlin & Brocchieri, 1996; Sitbon & Petrokovski, 2007). In cytokine and chemokine genes, conserved regions often denote key structural domains such as cytokine receptor binding sites, signal peptide sequences, and conserved amino acid residues pivotal for ligand-receptor interactions (Haskill et al., 1990; Xu et al., 2023). Consequently, they play a central role in modulating precise cellular responses and immune functions across diverse biological contexts.

The conservation of these regions underscores their significance in maintaining gene function, adaptive evolution, and immune system integrity, as they are molecular indicators of both evolutionary conservation and functional indispensability within the cytokine and chemokine gene families. Mutations in these conserved sequences may result in aberrant mRNA regulation, which has been connected to several inflammatory illnesses, including inflammatory bowel disease, rheumatoid arthritis, and atherosclerosis (Alizada et al., 2021).

1.9 1000 Genomes Project

The 1000 Genomes Project is a global initiative focused on sequencing the genomes of thousands of individuals to create a comprehensive map of human genetic variation. By examining diverse populations, the project aims to uncover common and rare genetic variants, informing studies on population genetics, disease susceptibility, and evolutionary history. This collaborative effort has revolutionized genomics research, providing vital insights for biomedical studies and personalized medicine (Fairley et al., 2020; Via et al., 2010).

1.10 Variant Call Format (VCF)

VCF is a standard file format for genetic variation data, compactly storing SNPs, insertions, deletions, and structural variants. It organizes variant information, including genomic coordinates and quality scores, aiding analysis and data sharing (Danecek et al., 2011; Garrison et al., 2022).

The VCF file (**Figure 5**) comprises a header ('##' meta-info lines, '#' field definition), defining eight mandatory columns: chromosome, position, identifiers, alleles, quality, filters, and annotations. Sample data include FORMAT and ID columns. Lines are tab-delimited and match the header's count (Danecek et al., 2011).

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG00096	HG00097	HG00099
chr21	5030578	21:5030578:C:T	C	T	.	.	AC=74;AF=0.0115553;CM=0;AN=6404;AN_EAS=1170;	GT	0 1	0 0	0 0
chr21	5030596	rs50308150	A	G	.	.	AC=2;AF=0.000312305;CM=3.21821e-05;AN=6404;	GT	0 0	1 0	0 0
chr21	5030912	21:5030912:CA:C	CA	C	.	.	AC=14;AF=0.00218613;CM=0.000597157;AN=6404;	GT	0 0	0 0	0 0

Figure 5: Example of a valid VCF. ALT lists alternate alleles, referenced by genotypes, indexed from 1, with 0 for the reference haplotype. The separator indicates phased (|) multiploid data.

1.11 BioMart

BioMart, a versatile package from Ensembl, offers a user-friendly interface for accessing and querying biological data from diverse databases. It is a valuable resource for genomic research and bioinformatics analyses because it gives researchers a potent toolkit for retrieving, integrating, and analysing genomic, transcriptomic, and proteomic datasets. This allows for effective data mining and exploration across a range of species and genomic features and makes it easier to conduct thorough investigations into gene function, regulatory elements, and genetic variation. (Durinck et al., 2005; Smedley et al., 2009).

1.12 Aims of the Master's Thesis

In this master's thesis, the primary objective is to conduct a thorough investigation into the distribution patterns of SNPs across various gene regions. This analysis will utilize a publicly available genomic dataset from the 1000 Genomes Project (Fairley et al., 2020). By integrating insights from gene structure, transcriptional regulation, mRNA dynamics, and genetic repositories, this study aims to elucidate the intricate genetic landscape influencing phenotypic diversity, disease susceptibility, and evolutionary adaptation. Using variant call formats and data mining techniques, the thesis will establish a comprehensive SNP database spanning three genomic regions, 5'UTR, CDS and 3'UTR.

Furthermore, this study will aim to elucidate how SNPs affect 3'UTR which are crucial for mRNA regulation and stability. This should be initially addressed by comparing 3'UTR SNP counts of genes with stable mRNAs (i.e., Housekeeping genes) and genes with unstable mRNAs (i.e., Cytokine and Chemokine genes).

SNPs in mRNA can have significant implications for health and disease. These genetic variations can influence various aspects of mRNA processing, including splicing, polyadenylation, and stability, ultimately affecting the quantity and quality of protein production. Dysregulation of mRNA due to SNPs can lead to aberrant gene expression patterns, contributing to the development of various pathologies, including cancer, neurodegenerative disorders, and autoimmune diseases.

Whilst most studies on SNPs focus on specific genes and their effects on those genes and gene products, this study aims to generate a unique and novel database that contains the number of SNPs within each transcript for every gene. This has not been done before to the best of our knowledge and hence this will be a valuable resource for further genetic analyses, facilitating a deeper understanding of genotype-phenotype relationships and aiding in SNP prioritization for future genetic association studies.

2. Materials and Methods

2.1 Data

The variant call format (VCF) file of the most recent release for 22 autosomes and X sex chromosomes was downloaded in gzipped format from the 1000 Genomes Project (1kGP) (Fairley et al., 2020). The database encompasses a high-coverage Illumina integrated phased panel from the 1kGP, comprising filtered single nucleotide variants (SNVs), insertions and deletions (INDELs), as well as structural variants (SVs) such as large deletions (DEL), insertions (INS), duplications (DUP), and inversions (INV) across 3,202 human samples from the 1kGP dataset. (Byrska-Bishop et al., 2022).

2.2 VCF Filtration

The software package VCFtools, version: 0.1.16 (Danecek et al., 2011) was used to process the downloaded VCF files. The first step involved removing any insertions, deletions and other non-SNP variants using the command `vcftools --vcf input_file.vcf --remove-indels --recode --recode-INFO-all --stdout | gzip -c > SNPs_only.vcf.gz`. The output vcf file was then processed to remove variants that were tagged as filtered and indicated low-quality variants using this command `vcftools --gzvcf input_file.vcf.gz --remove-filtered-all --recode --stdout | gzip -c > output_PASS_only.vcf.gz`.

For additional data analysis, the generated vcf file was then examined using R, version: 4.3.2 (R Core Team, 2018).

2.3 SNP Database Generation

The resulting VCF file was then further analysed in R using a suite of Bioconductor Packages such as VariantAnnotation, version: 1.46.0 (Obenchain et al., 2014); biomaRt, version: 2.58.2 (Durinck et al., 2005, 2009); GenomicRanges, version: 1.54.1 (Lawrence et al., 2013) and GenomicFeatures, version: 1.54.3 (Lawrence et al., 2013).

The VCF file for a chromosome was first read in R using the VariantAnnotation (version: 1.46.0) package. The parameter flag was set to NA to load the files faster since the size of the VCF files ranges from 11 GB to 61 GB. The file was read with the human reference genome version: GRCh38.p14. The annotation database was generated as TxDb objects from UCSC using TxDb.Hsapiens.UCSC.hg38.knownGene (BC Team & BP Maintainer, 2019). The following step included variant location within the 3'UTR region based on the annotation database. A data frame

was generated from the resulting locations which contained the start and end position of SNPs in 3'UTR for one chromosome, while another list contained all the unique Entrez Gene IDs having SNPs in 3'UTR (Entrez Gene ID list).

Next, the biomaRt (version: 2.58.2) package was used to loop through the Entrez Gene ID list and the gene description and related HUGO Gene Nomenclature Committee (HGNC) symbol for each gene ID was obtained. Additionally, using the biomaRt package (version: 2.58.2), the 3'UTR start and end positions for each transcript of the individual gene ID were determined.

Next, an algorithm was developed to determine how many rows in the data frame, for each transcript, had SNP sites that overlapped inside the start and finish positions. Using this row count as the transcript's SNP count, a normalised SNP count for each transcript was then determined. To take into consideration the wide variation in 3'UTR lengths between the transcripts, the counts were normalised to 1000 base pairs (bp).

A data frame including the 3'UTR start, 3'UTR end, UTR length, normalised SNP count, and HGNC symbol for each transcript from each gene ID was generated. The remaining VCF files were processed in the same manner. The final SNP database (final database) for 3'UTR was created by combining the 23 data frames, each for a single VCF file.

These steps were repeated for the remaining regions: 5'UTR and CDS to generate a final database for these regions.

2.4 SNP counts in Housekeeping, Cytokine and Chemokine genes

Next, the distribution of SNP counts inside certain genes of interest related to housekeeping, cytokines, and chemokines were examined and analysed. The first step was to ascertain and define these genes of interest. Previous studies have provided lists of housekeeping genes (HSIAO et al., 2001) and cytokine and chemokine genes (Cameron & Kelvin, 2003). These lists were downloaded and converted into text files (**Supplementary Table 1**, **Supplementary Table 2** and **Supplementary Table 3**) for further usage.

The second step was to define Bins to look at SNP distribution within the 3'UTR of our genes of interest. Therefore, bins were defined as **50 bp** regions. Two bins have been generated for every gene: the first bin is located just after the stop codon, spanning base pairs **1 – 50**, while the second bin is located between positions **175 – 225 bp** downwards of the stop codon. The bins were defined in such a way as to capture and compare the SNPs within two important regions. The first bin is close to the stop codon and hence this region might be impacted by features related to termination of translation, whereas the second bin was chosen to analyse and interpret SNPs at fixed positions

downwards of the stop codon for each transcript, while also being located between the first third and the middle of the selected transcripts.

Importantly, to remove any 3'UTRs that were too long or too short, the 3'UTR length was chosen to be between the lengths **300** and **600 bp (Figure 8)**. This was performed to remove any atypical 3'UTRs that might have been overlooked due to faulty annotations (Linial, 2003; Northcutt et al., 2021).

Using an algorithm to calculate the number of rows in the data frame for each transcript that had SNP sites that overlapped inside the start and finish positions, the SNP counts were generated for the genes of interest for each VCF file. The SNP counts were restricted to transcripts whose 3'UTR was between 300 and 600 bp. For every transcript, this row count served as the SNP count. The SNP counts for the two bins were computed without any normalisation because the bins are each 50 bp in size.

As a result, a database was created including the HGNC symbol, 3'UTR length, and SNP counts for 1–50 bp and 175–225 bp for every transcript. The same procedure was applied to the other VCF files. The final SNP database, which included SNP counts within the two bins in 3'UTR, was created by combining the 23 data frames, each for a single VCF file.

SNP counts inside the various bins for genes of interest were obtained by cross-referencing the gene lists of Housekeeping, Cytokine, and Chemokine genes that were previously prepared with the final database. Plotting and comparing the SNP distribution among the bins and the three gene categories was then done using this curated database.

The work was performed on the Life Science Compute Cluster (LiSC), operated by the Division of Computational Systems Biology (CUBE) and hosted by the University of Vienna.

The scripts used to perform the database generation and the analysis, is hosted on [GitHub](#) for public access.

3. Results

3.1 SNP Database for 3'UTR

A snippet of the final SNP Database for 3'UTR is shown in **Table 1** having the normalised SNP counts for the transcripts.

3_utr_start	3_utr_end	utr_length	norm_count	hgnc_symbol
70009	71585	1576	7	OR4F5
944154	944574	420	50	SAMD11
944154	944575	421	50	SAMD11
944154	944259	105	67	SAMD11
944203	944693	490	43	NOC2L
965192	965719	527	32	KLHL17
963387	963504	117	51	KLHL17
963920	964008	88	11	KLHL17
964107	964167	60	33	KLHL17
964349	964530	181	33	KLHL17
974576	975865	1289	36	PLEKHN1
974576	975008	432	37	PLEKHN1
974576	974588	12	83	PLEKHN1
975712	975865	153	46	PLEKHN1
975198	976171	973	27	PERM1
975204	976171	967	27	PERM1
976121	976171	50	40	PERM1
998962	999058	96	42	HES4
998964	999058	94	43	HES4
998970	999058	88	34	HES4
1014479	1014540	61	49	ISG15
1054982	1056116	1134	34	AGRN
1054982	1056118	1136	34	AGRN
1070967	1071816	849	33	RNF223
1081818	1082892	1074	33	C1orf159
1081823	1082892	1069	33	C1orf159
1081823	1082987	1164	34	C1orf159
1083915	1084018	103	49	C1orf159
1087139	1087204	65	77	C1orf159
1085878	1086012	134	30	C1orf159
1084353	1084383	30	33	C1orf159
1083915	1084086	171	47	C1orf159

Table 1: Snippet of the final database for 3'UTR

The table contains the 3'UTR start site, 3'UTR end site, the length of 3'UTR, normalised count and the HGNC symbol of the genes associated with the presented transcripts. While some genes such as

OR4F5, RNF223 and NOC2L only have a single transcript, the majority of genes have multiple transcripts, ranging from two in AGRN up to fourteen transcripts in C1orf159 (**Table 1**).

This range in transcript diversity creates a potential issue with the length of the transcripts. Transcripts can vary in length due to biological reasons such as alternative splicing or alternative polyadenylation, while they can also vary in length due to errors during annotation. This difference in length makes the comparison of short 3'UTR lengths (500 bp) with long 3'UTR lengths (1500 bp) difficult, for example, since the mutation rate per base pair differs and hence the SNP counts are incomparable. To prevent this bias, the counts were all normalised to 1000 bp.

The total length of the transcripts within 3'UTR can vary anywhere from 2 bp as the smallest in one of the transcripts in the gene SH2D2A and gene TMCO3 to the longest 3'UTR at 32869 bp in one of the transcripts of the gene SMAD2.

As in the 3'UTR database, the final 5'UTR (**Supplementary Table 4**) and CDS (**Supplementary Table 5**) database showed similar differences in transcript length. The longest transcript in the 5'UTR is 6565 bp for the gene ZC3H11A, while the smallest transcript in the region is 2 bp long and belongs to a transcript of the gene ZNF511. Similarly, the longest transcript in CDS is 21692 bp for the gene MUC16, while the smallest transcript in CDS is 2 and found in several genes, including API5, GATM, and MYO1C, among others.

Therefore, the variance in transcript length prompted the normalisation of the counts to a 1000 bp baseline level, enabling thorough analysis of the SNP counts among the transcripts and insightful comparisons within genes.

3.2 Total number of normalized SNP counts is the highest in CDS as compared to the 5'UTR and 3'UTR

The completion of the final database files prompted an analysis of the normalized SNP counts across the three defined gene regions (**Figure 6**).

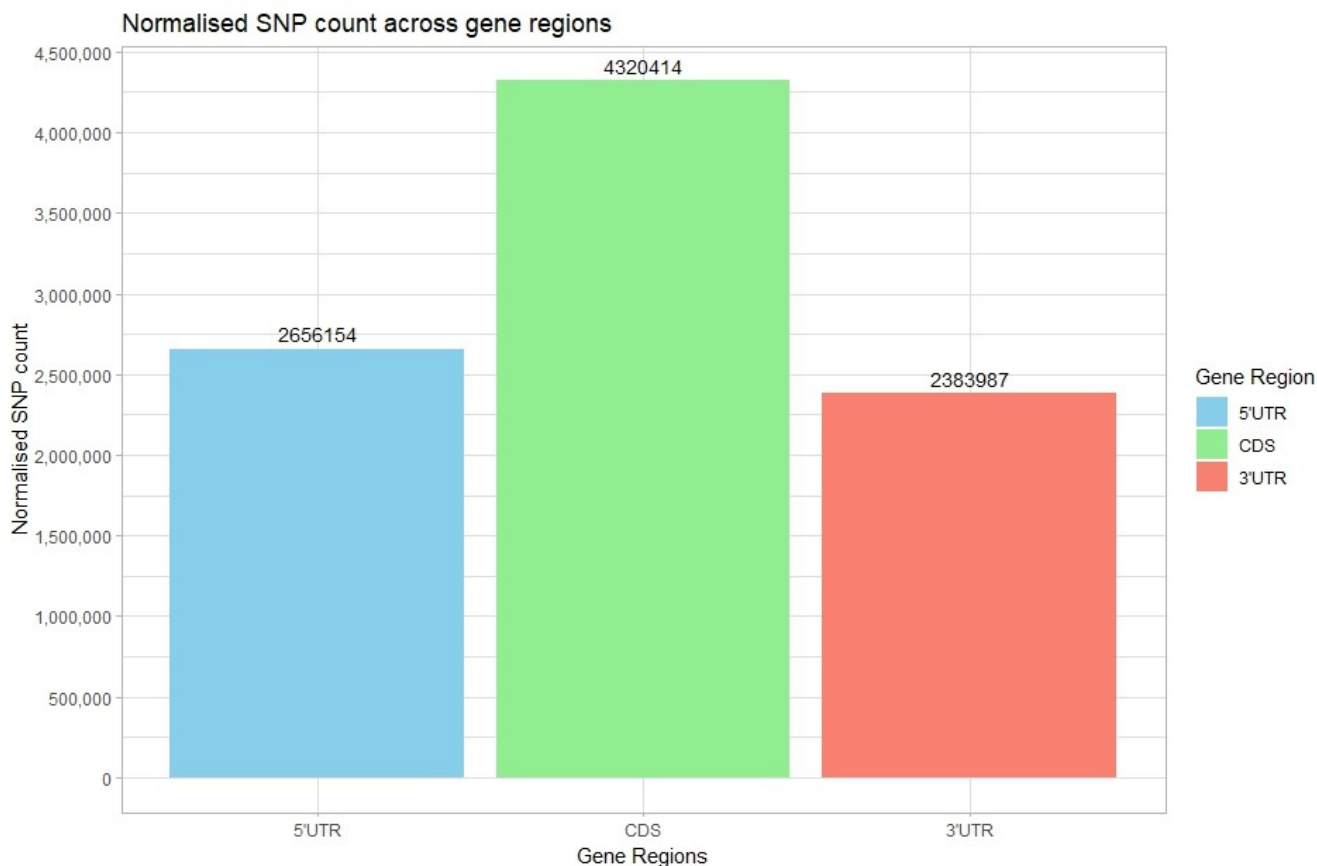


Figure 6: Bar Plot showing the normalised SNP counts. The total count of normalised SNPs is shown above the bars for each region.

Surprisingly, the SNP count was highest within CDS, followed by the 5'UTR and the 3'UTR. However, this can be explained, given the highest number of transcripts (189,984) within the CDS (**Figure 7**). The high transcript count can be explained by differential splicing frequently occurring within CDS and/or by erroneous annotations, both of which could increase the transcript count for CDS. Therefore, the high total SNP count in the CDS relative to the 5' and 3' UTRs is probably the result of the high transcript count in the CDS and the normalization of SNP count based on transcript length rather than transcript count.

Further examination revealed that the ratio of SNP per transcript, for CDS is around 23 SNPs per transcript, while for the 3'UTR it is around 22 SNPs per transcript, and for 5'UTR it is around 32 SNPs per transcript, uncovering the highest SNP count to be within the 5'UTR. The high SNP counts per transcript in 5'UTR might be attributed to a potential bias in the SNP normalization (i.e. SNP count

per 1000 bp): If a 5' UTR is erroneously short (e.g. 5 bp) owing to an incorrect annotation, then a SNP found in this very short 5' UTR would result in a very high normalized SNP count, thereby artificially increasing the overall SNP counts in 5' UTR. Given the technical difficulties in precisely mapping the transcription start site, it can be assumed that the 5' UTR annotations are less accurate than CDS and 3' UTR annotations. Indeed, the average 5' UTR transcript length is shorter (176 bp) than that of CDS (190 bp) and 3' UTR (747 bp).

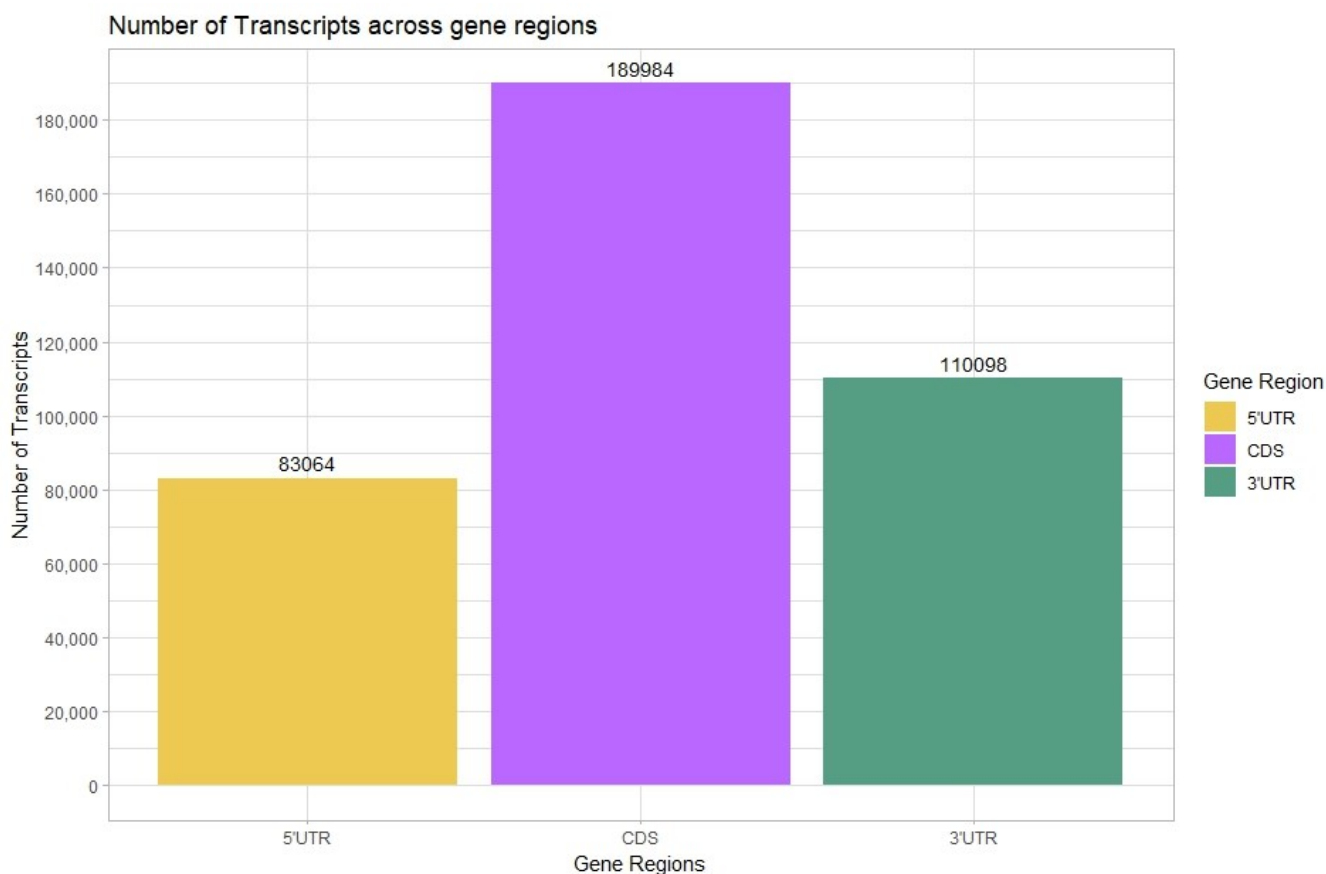


Figure 7: Bar Plot showing the number of transcripts for each gene region. The total of transcripts is shown above the bars for each region.

3.3 SNP distribution in Housekeeping, Cytokine and Chemokine genes

After the analysis of SNP counts in gene regions, the SNP counts and their distribution within the 3'UTR of Housekeeping, Cytokine, and Chemokine genes of interest were examined. SNPs were categorized into two bins of 50 bp each, with the first bin spanning the initial 50 bp following the stop codon, and the second bin covering positions 175 to 225 bp downstream of the stop codon. Normalization was not conducted as the bins were of equal length. Additionally, filtering was applied to retain only those transcripts falling within the length range of 300 to 600 bp (**Figure 8**).

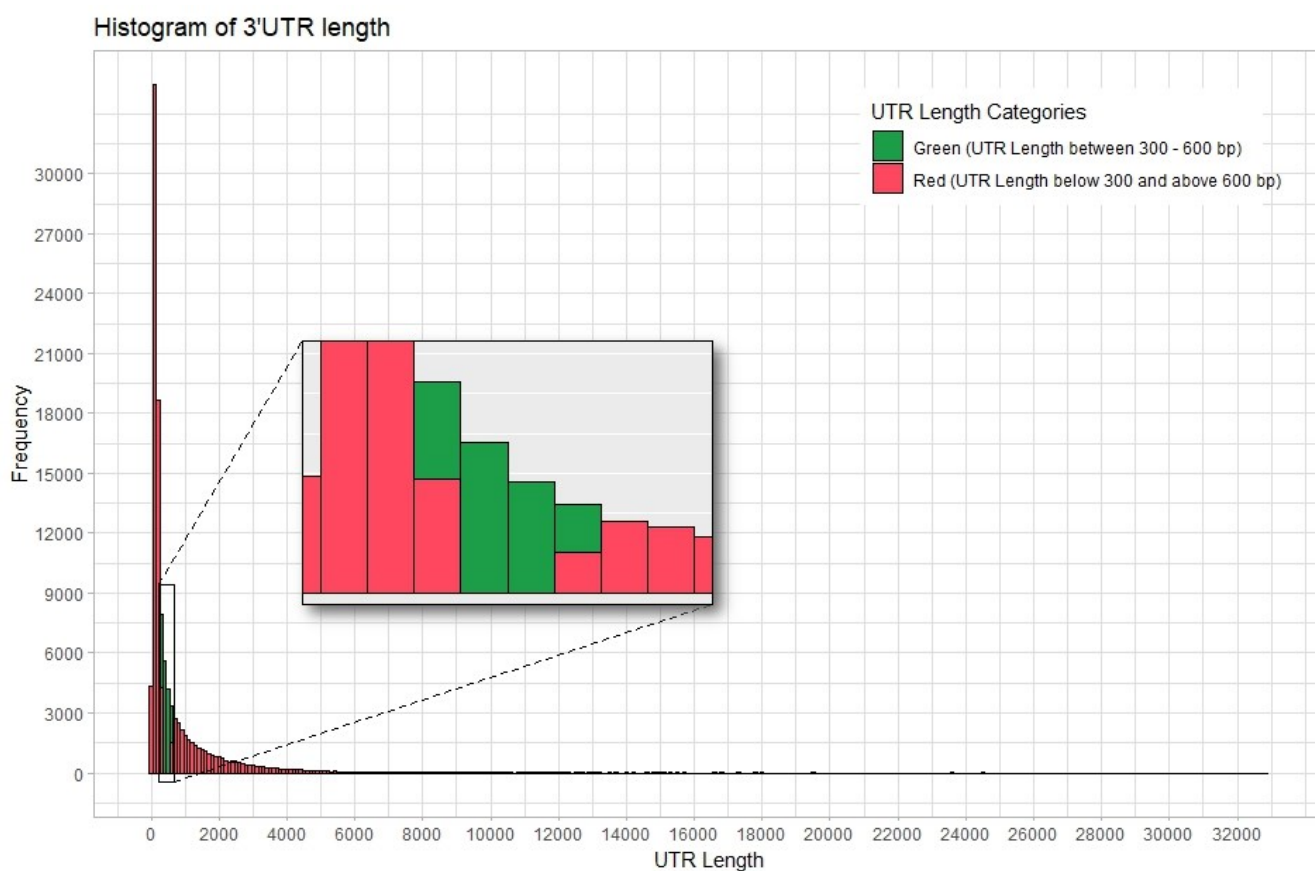


Figure 8: A histogram showing 3'UTR length across all the transcripts within genes. The inset shows the 3'UTR region between 250 – 650 bp. The green bars signify the subset of 3'UTR lengths (between 300 - 600 bp) that were analysed for SNP counts in Housekeeping, Cytokine and Chemokine genes of interest, while the red bars show the dataset that was excluded from the analysis. The Y axis shows the frequency of the 3'UTR lengths.

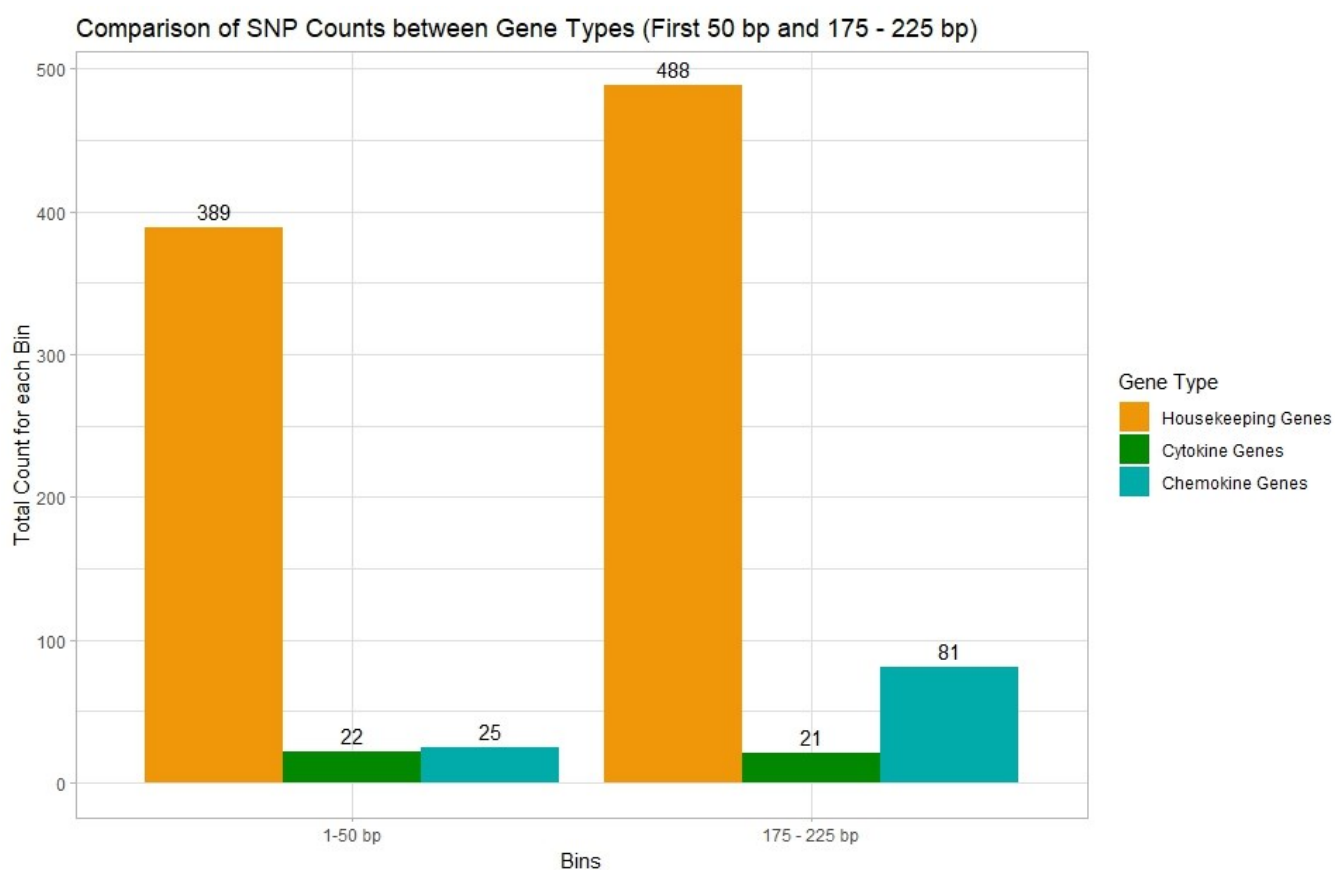


Figure 9: Bar Plot showing the SNP counts for Housekeeping, Cytokine and Chemokine genes within the two bins within the 3'UTR. The total SNP count for each region is shown above the bars. Gene regions are colour-coded in Yellow (Housekeeping genes), Green (Cytokine genes) and Blue (Chemokine genes).

Among the gene types, housekeeping genes exhibited the highest number of SNPs, compared to Cytokine and Chemokine genes. Out of 406 selected housekeeping genes, 166 genes displayed SNPs within the 3'UTR for both bins. Similarly, SNPs were observed in 11 Cytokine genes out of 84 genes of interest, while 12 Chemokine genes exhibited SNPs from a list of 150 genes (**Figure 10**).

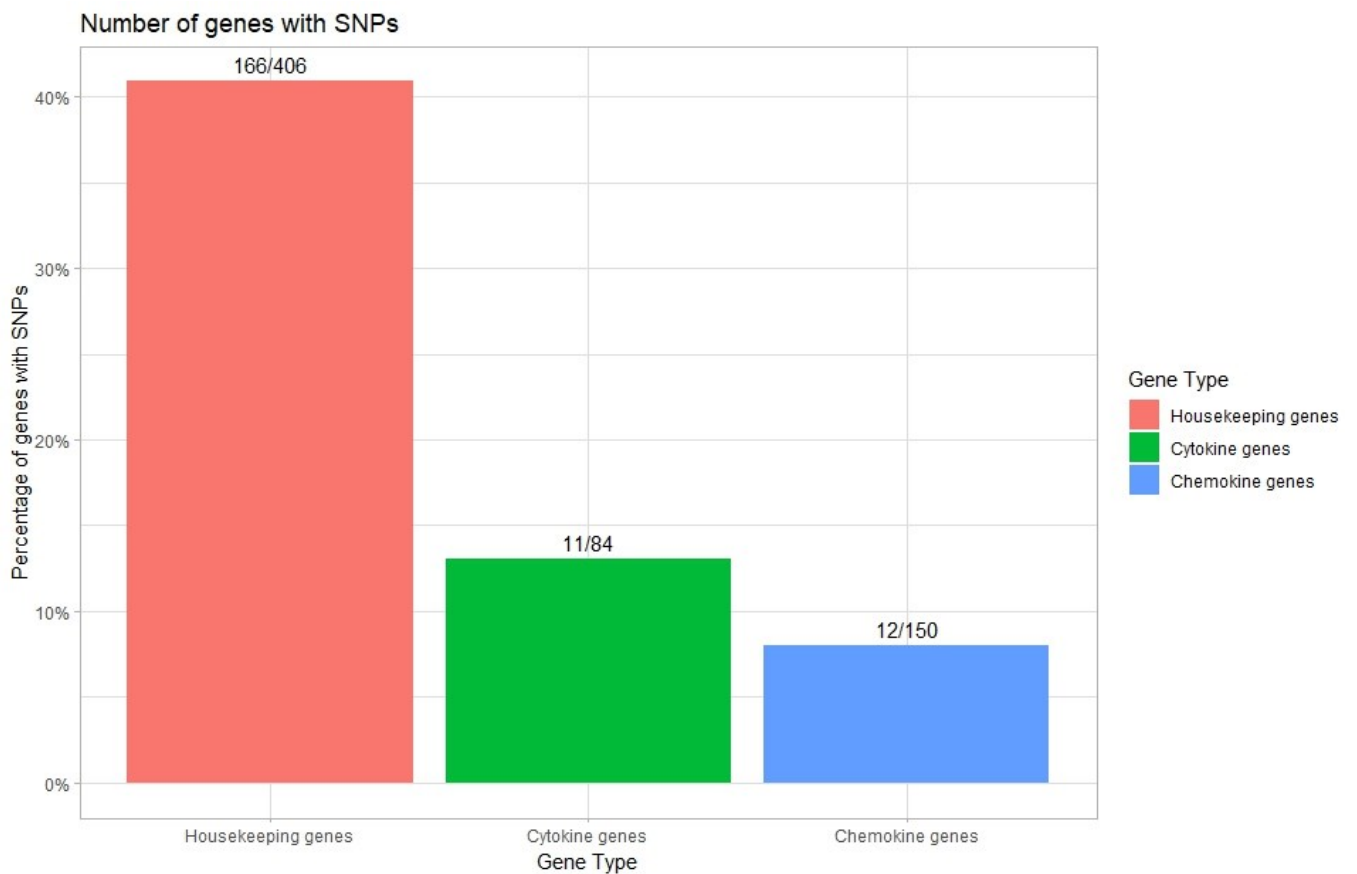


Figure 10: Bar Plot showing the number of genes with SNP counts for Housekeeping, Cytokine and Chemokine genes. The Y axis represents the fraction of genes from the total genes that have SNPs for the three regions. The number of genes with SNPs is represented above the bars.

The housekeeping genes have the highest SNP count across both bins as it has the highest number of genes having SNPs and also a high number of transcripts. A closer look at the ratio of SNP per transcript for the first bin reveals that this increase in the number of transcripts does not increase the SNP ratio. The ratio stands at around 1 SNP per transcript for Housekeeping, Cytokine and Chemokine genes. This trend is also followed second bin, where both housekeeping and cytokine genes have the SNP per transcript ratio of 1 compared to a ratio of 2 for chemokine genes (**Figure 11**).

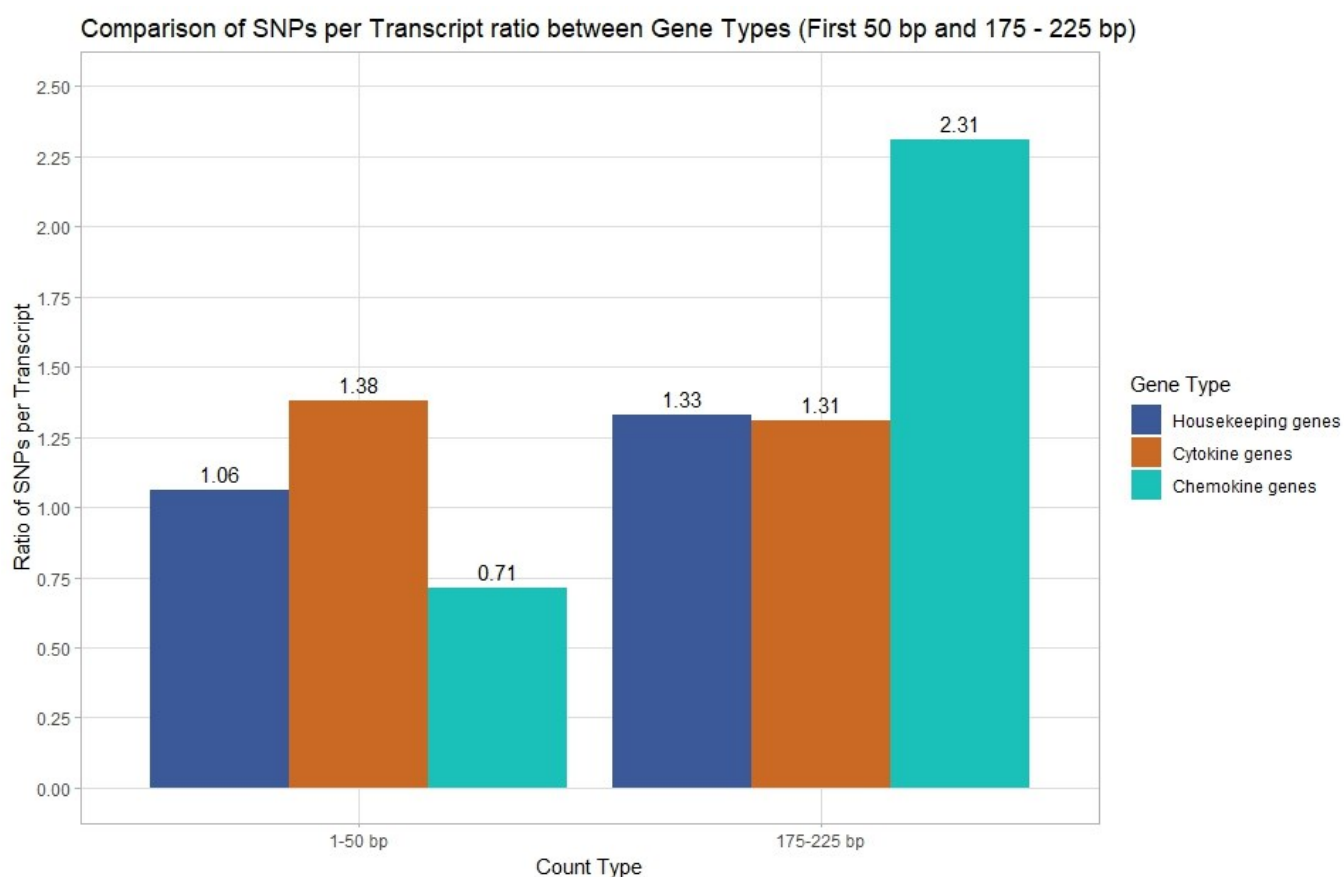


Figure 11: Bar Plot showing the SNPs per Transcript ratio for Housekeeping, Cytokine and Chemokine genes within the two bins within the 3'UTR. The actual ratio for each region is shown above the bars. Gene regions are colour-coded in Blue (Housekeeping genes), Brown (Cytokine genes) and Teal (Chemokine genes).

For detailed SNP counts in the two bins, a section of the final database is available in **Supplementary Table 6** for housekeeping genes, **Supplementary Table 7** for Cytokine genes, and **Supplementary Table 8** for Chemokine genes.

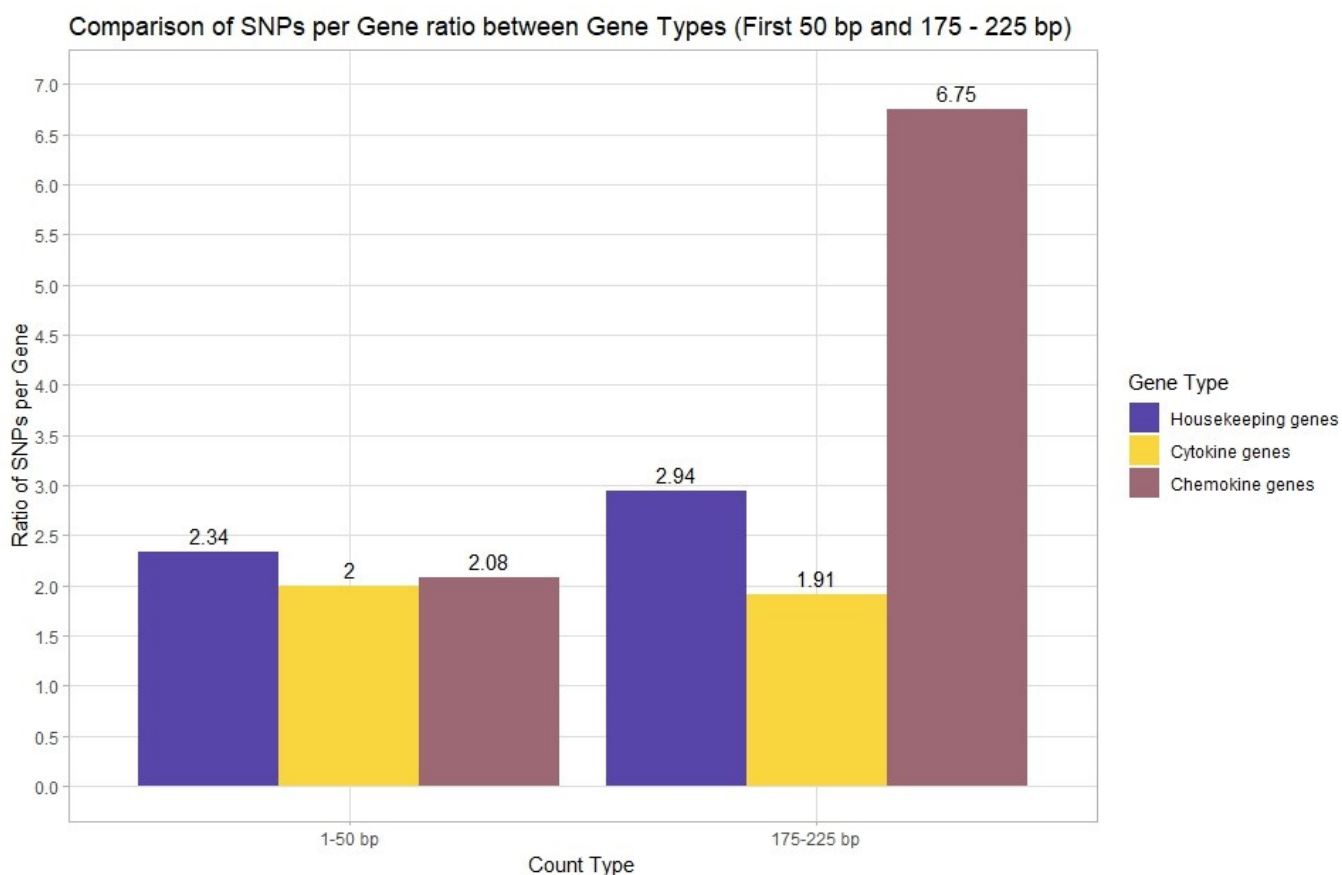


Figure 12: Bar Plot showing the SNPs per Gene ratio for Housekeeping, Cytokine and Chemokine genes within the two bins within the 3'UTR. The actual ratio for each region is shown above the bars. Gene regions are colour-coded in Purple (Housekeeping genes), Yellow (Cytokine genes) and Brown (Chemokine genes).

A closer examination of the SNPs per gene ratio (**Figure 12**) shows that the Housekeeping, Cytokine, and Chemokine genes have ratios of about 2 SNPs per gene over the first 50 bp. In contrast, for the second bin, the ratio increases from two to three SNPs per gene for Housekeeping genes and from two to about seven SNPs per gene for Chemokine genes. The ratio for Cytokine genes, on the other hand, remains stable at around two SNPs per gene.

4. Discussions

4.1 Effect of Normalisation on SNP Counts

Since the lengths of transcripts among genes were not similar in length for the 5'UTR, CDS and 3'UTR, normalisation was implemented. Counts were normalised to 1000 bp to correct the bias for longer transcripts to accumulate more SNPs.

This transcript bias can be illustrated by the transcript diversity within genes such as PLEKHN1, where the 3'UTR length can vary from 12 to 1289 bp. This huge range in 3'UTR length necessitates the implementation of normalisation since without normalisation, the counts cannot be compared even for two transcripts of the same gene.

Although normalisation offers a useful framework for comparing SNPs across transcripts, it also introduces a normalisation bias. The SNP counts of transcripts containing very short gene regions, such as 2 – 20 bp, will be greatly exaggerated when normalised to 1000 bp. A transcript having a 3'UTR of 2 bp and one SNP, for instance, would have 50 SNPs after normalisation. Similarly, if the transcript has a very long gene region, such as 5000 – 32000 bp, the SNP count from these regions is reduced by a lot to account for normalisation. For example, a transcript having a 3'UTR of 32000 bp having an SNP count of 100 SNPs would only have 3 SNPs after normalisation.

This normalisation bias is a major issue since, although normalisation makes it possible to compare SNP counts, it also obscures any potential natural variation resulting from mutation rates within gene regions. Additionally, as seen in **Figure 6** and **Figure 7**, a high concentration of longer transcripts within a gene area can significantly reduce the number of SNPs. For instance, the average length of the CDS is 190 bp, compared to 747 bp for the 3'UTR. Additionally, the normalisation pushes the overall SNP counts for 3'UTR to be significantly lower than 3'UTR because it contains fewer transcripts (83604) than 189984 in CDS.

Normalization can lead to inflated or lowered SNP counts, but it provides a critical framework for standardising transcript-to-transcript comparisons, especially when evaluating SNPs within gene regions with significant length disparities.

4.2 Effect of Conserved Sequences on Housekeeping, Cytokine and Chemokine genes

Investigating the SNP counts for Housekeeping, Cytokine and Chemokine genes of interest uncovered a trend where the first 50 bp after the stop codon had fewer SNPs on average than the second bin. This is in line with previous work describing a high conservation of the 3'UTR just after the stop codon (Greener et al., 2002) for most genes since any mutations can be detrimental to the functioning of the stop codon. Our data supports this since the ratio of SNPs per gene for the three gene types, where Housekeeping, Cytokine and Chemokine genes all exhibit around 2 SNPs per gene within the first bin.

The SNP-to-gene ratio of the second bin reveals more SNPs in the housekeeping and chemokine genes than in the cytokine genes. This is unexpected considering the strong selective pressure that evolution has placed on housekeeping genes to maintain the steady synthesis of essential proteins needed for energy metabolism and cellular respiration.

Conserved sites might be attributed to the decrease in SNPs observed within cytokine genes. We know that conserved regions often denote key structural domains such as AREs (Asson-Batres et al., 1994) which are important for post-transcriptional gene regulation that needs to be maintained and thus have a high evolutionary pressure to keep SNP counts low since any polymorphism can contribute to diseases such as cancer, autoimmune and inflammatory diseases (Hitti, 2012). The substantial increase in the SNP count of chemokine genes, however, is not due to conserved sites and we suspect that there might be a region of high polymorphism present around the 175 – 225 bp instead of a conserved region that might be causing this increase in the SNPs.

Since cytokine and chemokine genes have a shorter lifespan (Whiteside, 1994) and play a crucial role in regulating immune responses and inflammation, the 3'UTR is important for preserving the half-life of mRNAs. As a result, as **Figure 9** illustrates, there is an observed decrease in SNP frequency within the 3'UTR of these genetic regions.

This preliminary study sheds light on SNP distributions across housekeeping, cytokine, and chemokine genes, revealing intriguing trends and patterns. Additional analysis is necessary to completely comprehend the fundamental processes that underlie these findings and explore conserved sites. This study represents an initial step towards a deeper understanding of genetic variation in these gene types, paving the way for future investigations.

5. Conclusions

The generation of novel SNP count databases for 5'UTR, CDS and 3'UTR provides us with important insights into the distribution of SNPs across the different gene regions and its influence on disease susceptibility and evolutionary adaptation.

Comparing SNP count across the different gene regions informs us that while CDS does show significantly more SNP counts than 5'UTR and 3'UTR, a closer look indicates a higher number of transcripts within CDS that drives up the SNP count. Alongside this, atypical annotations lead to a further increase in the counts. While normalisation helps in analysing the counts, normalisation bias can cause further complications.

This database can serve as a starting point for comparing the SNPs within various transcripts, despite its limitations. This can also be utilised to examine the effects of SNPs on various gene types. For instance, comparing the Housekeeping, Cytokine and Chemokine genes of interest, we observe high SNP counts in 3'UTR for housekeeping genes, while Cytokine and Chemokine genes have fewer SNPs. Short mRNA half-life, coupled with their important role in inflammation and immune response could explain a lack of SNPs within these gene types.

Moving forward, this database will help explore the impact of genetic variation on gene function and phenotype with greater precision and depth. By providing a comprehensive framework for investigating SNP distributions and their associations with various genetic elements, this database lays the groundwork for future studies aimed at unravelling the complexities of genetic variation and its consequences for human health and evolution.

6. References

- Akdeli, N., Riemann, K., Westphal, J., Hess, J., Siffert, W., & Bachmann, H. S. (2014). A 3'UTR polymorphism modulates mRNA stability of the oncogene and drug target Polo-like Kinase 1. *Molecular Cancer*, 13(1), 87. <https://doi.org/10.1186/1476-4598-13-87>
- Alizada, A., Khyzha, N., Wang, L., Antounians, L., Chen, X., Khor, M., Liang, M., Rathnakumar, K., Weirauch, M. T., Medina-Rivera, A., Fish, J. E., & Wilson, M. D. (2021). Conserved regulatory logic at accessible and inaccessible chromatin during the acute inflammatory response in mammals. *Nature Communications*, 12(1), 567. <https://doi.org/10.1038/s41467-020-20765-1>
- Araujo, P. R., Yoon, K., Ko, D., Smith, A. D., Qiao, M., Suresh, U., Burns, S. C., & Penalva, L. O. F. (2012). Before It Gets Started: Regulating Translation at the 5' UTR. *Comparative and Functional Genomics*, 2012, 1–8. <https://doi.org/10.1155/2012/475731>
- Asson-Batres, M. A., Spurgeon, S. L., Diaz, J., DeLoughery, T. G., & Bagby, G. C. (1994). Evolutionary conservation of the AU-rich 3'untranslated region of messenger RNA. *Proceedings of the National Academy of Sciences*, 91(4), 1318–1322. <https://doi.org/10.1073/pnas.91.4.1318>
- Audic, Y., & Hartley, R. S. (2004). Post-transcriptional regulation in cancer. *Biology of the Cell*, 96(7), 479–498. <https://doi.org/10.1016/j.biolcel.2004.05.002>
- Barnes, M. R. (2002). SNP and Mutation Data on the Web – Hidden Treasures for Uncovering. *Comparative and Functional Genomics*, 3, 590124. <https://doi.org/10.1002/cfg.131>
- Barreau, C. (2005). AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Research*, 33(22), 7138–7150. <https://doi.org/10.1093/nar/gki1012>
- Bartel, D. P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, 136(2), 215–233. <https://doi.org/10.1016/j.cell.2009.01.002>
- BC Team, & BP Maintainer. (2019). *TxDb.Hsapiens.UCSC.hg38.knownGene: Annotation package for TxDb object(s). R package version 3.4.6.*
- Berget, S. M., Moore, C., & Sharp, P. A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences*, 74(8), 3171–3175. <https://doi.org/10.1073/pnas.74.8.3171>
- BIRNSTIEL, M., BUSSLINGER, M., & STRUB, K. (1985). Transcription termination and 3' processing: the end is in site! *Cell*, 41(2), 349–359. [https://doi.org/10.1016/S0092-8674\(85\)80007-6](https://doi.org/10.1016/S0092-8674(85)80007-6)
- Brookes, A. J. (1999). The essence of SNPs. *Gene*, 234(2), 177–186. [https://doi.org/10.1016/S0378-1119\(99\)00219-X](https://doi.org/10.1016/S0378-1119(99)00219-X)

- Brown TA. (2004). *Genomes, Chapter 14, Mutation, Repair and Recombination* (2nd ed.). Oxford: Wiley-Liss.
- Burgess, R. R. (1971). RNA POLYMERASE. *Annual Review of Biochemistry*, 40(1), 711–740. <https://doi.org/10.1146/annurev.bi.40.070171.003431>
- Busch, H. (1976). The Function of the 5' Cap of mRNA and Nuclear RNA Species. *Perspectives in Biology and Medicine*, 19(4), 549–567. <https://doi.org/10.1353/pbm.1976.0064>
- BUTTE, A. J., DZAU, V. J., & GLUECK, S. B. (2001). Further defining housekeeping, or “maintenance,” genes Focus on “A compendium of gene expression in normal human tissues.” *Physiological Genomics*, 7(2), 95–96. <https://doi.org/10.1152/physiolgenomics.2001.7.2.95>
- Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Paul Flicek, Germer, S., Brand, H., Hall, I. M., Talkowski, M. E., ... Xiao, C. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, 185(18), 3426–3440.e19. <https://doi.org/10.1016/j.cell.2022.08.004>
- Cameron, M. J., & Kelvin, D. J. (2003). *Cytokines and Chemokines—Their Receptors and Their Genes: An Overview* (pp. 8–32). https://doi.org/10.1007/978-1-4615-0171-8_2
- Castle, J. C. (2011). SNPs Occur in Regions with Less Genomic Sequence Conservation. *PLoS ONE*, 6(6), e20660. <https://doi.org/10.1371/journal.pone.0020660>
- Chan, J. J., Tabatabaeian, H., & Tay, Y. (2023). 3'UTR heterogeneity and cancer progression. *Trends in Cell Biology*, 33(7), 568–582. <https://doi.org/10.1016/j.tcb.2022.10.001>
- Chen, C.-Y. A., & Shyu, A.-B. (1995). AU-rich elements: characterization and importance in mRNA degradation. *Trends in Biochemical Sciences*, 20(11), 465–470. [https://doi.org/10.1016/S0968-0004\(00\)89102-1](https://doi.org/10.1016/S0968-0004(00)89102-1)
- Chow, L. T., Gelinas, R. E., Broker, T. R., & Roberts, R. J. (1977). An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, 12(1), 1–8. [https://doi.org/10.1016/0092-8674\(77\)90180-5](https://doi.org/10.1016/0092-8674(77)90180-5)
- Collins, F. S., Brooks, L. D., & Chakravarti, A. (1998). A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation: Table 1. *Genome Research*, 8(12), 1229–1231. <https://doi.org/10.1101/gr.8.12.1229>
- Condit, C. M., Achter, P. J., Lauer, I., & Sefcovic, E. (2002). The changing meanings of mutation: A contextualized study of public discourse. *Human Mutation*, 19(1), 69–75. <https://doi.org/10.1002/humu.10023>
- Copley, S. D. (2010). Evolution and the Enzyme. In *Comprehensive Natural Products II* (pp. 9–46). Elsevier. <https://doi.org/10.1016/B978-008045382-8.00670-5>

- Crawford, E. K., Ensor, J. E., Kalvakolanu, I., & Hasday, J. D. (1997). The Role of 3' Poly(A) Tail Metabolism in Tumor Necrosis Factor- α Regulation. *Journal of Biological Chemistry*, 272(34), 21120–21127. <https://doi.org/10.1074/jbc.272.34.21120>
- CRICK, F. (1970). Central Dogma of Molecular Biology. *Nature*, 227(5258), 561–563. <https://doi.org/10.1038/227561a0>
- Crow, J. F. (2000). The origins, patterns and implications of human spontaneous mutation. *Nature Reviews Genetics*, 1(1), 40–47. <https://doi.org/10.1038/35049558>
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- Darnell, J. E., Wall, R., & Tushinski, R. J. (1971). An Adenylic Acid-Rich Sequence in Messenger RNA of HeLa Cells and Its Possible Relationship to Reiterated Sites in DNA. *Proceedings of the National Academy of Sciences*, 68(6), 1321–1325. <https://doi.org/10.1073/pnas.68.6.1321>
- Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16), 3439–3440. <https://doi.org/10.1093/bioinformatics/bti525>
- Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, 4(8), 1184–1191. <https://doi.org/10.1038/nprot.2009.97>
- Edmonds, M., Vaughan, M. H., & Nakazato, H. (1971). Polyadenylic Acid Sequences in the Heterogeneous Nuclear RNA and Rapidly-Labeled Polyribosomal RNA of HeLa Cells: Possible Evidence for a Precursor Relationship. *Proceedings of the National Academy of Sciences*, 68(6), 1336–1340. <https://doi.org/10.1073/pnas.68.6.1336>
- Eisenberg, E., & Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends in Genetics*, 29(10), 569–574. <https://doi.org/10.1016/j.tig.2013.05.010>
- Fairley, S., Lowy-Gallego, E., Perry, E., & Flicek, P. (2020). The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research*, 48(D1), D941–D947. <https://doi.org/10.1093/nar/gkz836>
- Falk, R. (1986). What is a gene? *Studies in History and Philosophy of Science Part A*, 17(2), 133–173. [https://doi.org/https://doi.org/10.1016/0039-3681\(86\)90024-5](https://doi.org/https://doi.org/10.1016/0039-3681(86)90024-5)
- Foster, P. L. (2000). Adaptive mutation: implications for evolution. *BioEssays*, 22(12), 1067–1074. [https://doi.org/10.1002/1521-1878\(200012\)22:12<1067::AID-BIES4>3.0.CO;2-Q](https://doi.org/10.1002/1521-1878(200012)22:12<1067::AID-BIES4>3.0.CO;2-Q)

- Fraser, J. A., Huang, J. C., Pukkila-Worley, R., Alspaugh, J. A., Mitchell, T. G., & Heitman, J. (2005). Chromosomal Translocation and Segmental Duplication in *Cryptococcus neoformans*. *Eukaryotic Cell*, 4(2), 401–406. <https://doi.org/10.1128/EC.4.2.401-406.2005>
- Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler-Smith, C., Hurles, M. E., Carter, N. P., Scherer, S. W., & Lee, C. (2006). Copy number variation: New insights in genome diversity. *Genome Research*, 16(8), 949–961. <https://doi.org/10.1101/gr.3677206>
- Friedel, C. C., Dölken, L., Ruzsics, Z., Koszinowski, U. H., & Zimmer, R. (2009). Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. *Nucleic Acids Research*, 37(17), e115–e115. <https://doi.org/10.1093/nar/gkp542>
- Fuchs, S. M., Laribee, R. N., & Strahl, B. D. (2009). Protein modifications in transcription elongation. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1789(1), 26–36. <https://doi.org/10.1016/j.bbagr.2008.07.008>
- Fuda, N. J., Ardehali, M. B., & Lis, J. T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, 461(7261), 186–192. <https://doi.org/10.1038/nature08449>
- Galtier, N. (2014). Mutation-Driven Evolution. *Systematic Biology*, 63(1), 113–114. <https://doi.org/10.1093/sysbio/syt055>
- Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S., & Prins, P. (2022). A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLOS Computational Biology*, 18(5), e1009123. <https://doi.org/10.1371/journal.pcbi.1009123>
- Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korb, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., & Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Research*, 17(6), 669–681. <https://doi.org/10.1101/gr.6339607>
- Glisovic, T., Bachorik, J. L., Yong, J., & Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters*, 582(14), 1977–1986. <https://doi.org/10.1016/j.febslet.2008.03.004>
- Gnatt, A. L., Cramer, P., Fu, J., Bushnell, D. A., & Kornberg, R. D. (2001). Structural Basis of Transcription: An RNA Polymerase II Elongation Complex at 3.3 Å Resolution. *Science*, 292(5523), 1876–1882. <https://doi.org/10.1126/science.1059495>
- Goldstrohm, A. C., & Wickens, M. (2008). Multifunctional deadenylase complexes diversify mRNA control. *Nature Reviews Molecular Cell Biology*, 9(4), 337–344. <https://doi.org/10.1038/nrm2370>
- Goss, D. J., & Domashevskiy, A. V. (2016). Messenger RNA (mRNA): The Link between DNA and Protein. In *Encyclopedia of Cell Biology* (pp. 341–345). Elsevier. <https://doi.org/10.1016/B978-0-12-394447-4.10040-9>

- Greener, M. J., Sewry, C. A., Muntoni, F., & Roberts, R. G. (2002). The 3'-untranslated region of the dystrophin gene – conservation and consequences of loss. *European Journal of Human Genetics*, 10(7), 413–420. <https://doi.org/10.1038/sj.ejhg.5200822>
- Griesemer, D., Xue, J. R., Reilly, S. K., Ulirsch, J. C., Kukreja, K., Davis, J. R., Kanai, M., Yang, D. K., Butts, J. C., Guney, M. H., Luban, J., Montgomery, S. B., Finucane, H. K., Novina, C. D., Tewhey, R., & Sabeti, P. C. (2021). Genome-wide functional screen of 3'UTR variants uncovers causal variants for human disease and evolution. *Cell*, 184(20), 5247–5260.e19. <https://doi.org/10.1016/j.cell.2021.08.025>
- Gu, W., Gurguis, C. I., Zhou, J. J., Zhu, Y., Ko, E.-A., Ko, J.-H., Wang, T., & Zhou, T. (2015). Functional and Structural Consequence of Rare Exonic Single Nucleotide Polymorphisms: One Story, Two Tales. *Genome Biology and Evolution*, 7(10), 2929–2940. <https://doi.org/10.1093/gbe/evv191>
- Guhaniyogi, J., & Brewer, G. (2001). Regulation of mRNA stability in mammalian cells. *Gene*, 265(1–2), 11–23. [https://doi.org/10.1016/S0378-1119\(01\)00350-X](https://doi.org/10.1016/S0378-1119(01)00350-X)
- Haas, U., Sczakiel, G., & Laufer, S. (2012). MicroRNA-mediated regulation of gene expression is affected by disease-associated SNPs within the 3'-UTR via altered RNA structure. *RNA Biology*, 9(6), 924–937. <https://doi.org/10.4161/rna.20497>
- Hargrove, J. L., & Schmidt, F. H. (1989). The role of mRNA and protein stability in gene expression. *The FASEB Journal*, 3(12), 2360–2370. <https://doi.org/10.1096/fasebj.3.12.2676679>
- Harris, L. F., Sullivan, M. R., & Hickok, D. F. (1993). Conservation of genetic information: a code for site-specific DNA recognition. *Proceedings of the National Academy of Sciences*, 90(12), 5534–5538. <https://doi.org/10.1073/pnas.90.12.5534>
- Hartzog, G. A., & Tamkun, J. W. (2007). A new role for histone tail modifications in transcription elongation. *Genes & Development*, 21(24), 3209–3213. <https://doi.org/10.1101/gad.1628707>
- Haskill, S., Peace, A., Morris, J., Sporn, S. A., Anisowicz, A., Lee, S. W., Smith, T., Martin, G., Ralph, P., & Sager, R. (1990). Identification of three related human GRO genes encoding cytokine functions. *Proceedings of the National Academy of Sciences*, 87(19), 7732–7736. <https://doi.org/10.1073/pnas.87.19.7732>
- Hitti, E. (2012). Sequence variations affecting AU-rich element function and disease. *Frontiers in Bioscience*, 17(1), 1846. <https://doi.org/10.2741/4023>
- Hoffmann, D., Diderrich, R., Reithofer, V., Friederichs, S., Kock, M., Essen, L.-O., & Mösch, H.-U. (2020). Functional reprogramming of *Candida glabrata* epithelial adhesins: the role of conserved and variable structural motifs in ligand binding. *Journal of Biological Chemistry*, 295(35), 12512–12524. <https://doi.org/10.1074/jbc.RA120.013968>

- Hoffmann, N. A., Jakobi, A. J., Moreno-Morcillo, M., Glatt, S., Kosinski, J., Hagen, W. J. H., Sachse, C., & Müller, C. W. (2015). Molecular structures of unbound and transcribing RNA polymerase III. *Nature*, 528(7581), 231–236. <https://doi.org/10.1038/nature16143>
- HSIAO, L.-L., DANGOND, F., YOSHIDA, T., HONG, R., JENSEN, R. V., MISRA, J., DILLON, W., LEE, K. F., CLARK, K. E., HAVERTY, P., WENG, Z., MUTTER, G. L., FROSCH, M. P., MacDONALD, M. E., MILFORD, E. L., CRUM, C. P., BUENO, R., PRATT, R. E., MAHADEVAPPA, M., ... GULLANS, S. R. (2001). A compendium of gene expression in normal human tissues. *Physiological Genomics*, 7(2), 97–104. <https://doi.org/10.1152/physiolgenomics.00040.2001>
- Hu, Z., & Bruno, A. E. (2011). The Influence of 3'UTRs on MicroRNA Function Inferred from Human SNP Data. *Comparative and Functional Genomics*, 2011, 1–9. <https://doi.org/10.1155/2011/910769>
- Jackson, R. J., & Standart, N. (1990). Do the poly(A) tail and 3' untranslated region control mRNA translation? *Cell*, 62(1), 15–24. [https://doi.org/https://doi.org/10.1016/0092-8674\(90\)90235-7](https://doi.org/https://doi.org/10.1016/0092-8674(90)90235-7)
- Joshi, C. J., Ke, W., Drangowska-Way, A., O'Rourke, E. J., & Lewis, N. E. (2022). What are housekeeping genes? *PLOS Computational Biology*, 18(7), e1010295. <https://doi.org/10.1371/journal.pcbi.1010295>
- Karapetyan, A., Buiting, C., Kuiper, R., & Coolen, M. (2013). Regulatory Roles for Long ncRNA and mRNA. *Cancers*, 5(4), 462–490. <https://doi.org/10.3390/cancers5020462>
- Karki, R., Pandya, D., Elston, R. C., & Ferlini, C. (2015). Defining “mutation” and “polymorphism” in the era of personal genomics. *BMC Medical Genomics*, 8(1), 37. <https://doi.org/10.1186/s12920-015-0115-z>
- Karlin, S., & Brocchieri, L. (1996). Evolutionary conservation of RecA genes in relation to protein structure and function. *Journal of Bacteriology*, 178(7), 1881–1894. <https://doi.org/10.1128/jb.178.7.1881-1894.1996>
- Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., & Stamm, S. (2013). Function of alternative splicing. *Gene*, 514(1), 1–30. <https://doi.org/10.1016/j.gene.2012.07.083>
- KISLAUSKIS, E., & SINGER, R. (1992). Determinants of mRNA localization. *Current Opinion in Cell Biology*, 4(6), 975–978. [https://doi.org/10.1016/0955-0674\(92\)90128-Y](https://doi.org/10.1016/0955-0674(92)90128-Y)
- Kondrashov Alexey S., & Houle David. (1994). Genotype—environment interactions and the estimation of the genomic mutation rate in *Drosophila melanogaster*. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 258(1353), 221–227. <https://doi.org/10.1098/rspb.1994.0166>
- Kontoyiannis, D., Pasparakis, M., Pizarro, T. T., Cominelli, F., & Kollias, G. (1999). Impaired On/Off Regulation of TNF Biosynthesis in Mice Lacking TNF AU-Rich Elements. *Immunity*, 10(3), 387–398. [https://doi.org/10.1016/S1074-7613\(00\)80038-2](https://doi.org/10.1016/S1074-7613(00)80038-2)
- Kovarik, P., Bestehorn, A., & Fesselet, J. (2021). Conceptual Advances in Control of Inflammation by the RNA-Binding Protein Tristetraprolin. *Frontiers in Immunology*, 12. <https://doi.org/10.3389/fimmu.2021.751313>

- Kovarik, P., Ebner, F., & Sedlyarov, V. (2017). Posttranscriptional regulation of cytokine expression. *Cytokine*, 89, 21–26. <https://doi.org/10.1016/j.cyto.2015.11.007>
- Kuhn, B., Abdel-Monem, M., & Hoffmann-Berling, H. (1979). DNA Helicases. *Cold Spring Harbor Symposia on Quantitative Biology*, 43(0), 63–67. <https://doi.org/10.1101/SQB.1979.043.01.011>
- La Rocca, G., Olejniczak, S. H., González, A. J., Briskin, D., Vidigal, J. A., Spraggon, L., DeMatteo, R. G., Radler, M. R., Lindsten, T., Ventura, A., Tuschl, T., Leslie, C. S., & Thompson, C. B. (2015). In vivo, Argonaute-bound microRNAs exist predominantly in a reservoir of low molecular weight complexes not associated with mRNA. *Proceedings of the National Academy of Sciences*, 112(3), 767–772. <https://doi.org/10.1073/pnas.1424217112>
- Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, 9(8), e1003118. <https://doi.org/10.1371/journal.pcbi.1003118>
- Lee, E. K., & Gorospe, M. (2011). Coding region. *RNA Biology*, 8(1), 44–48. <https://doi.org/10.4161/rna.8.1.13863>
- Lee, S. Y., Mendecki, J., & Brawerman, G. (1971). A Polynucleotide Segment Rich in Adenylic Acid in the Rapidly-Labeled Polyribosomal RNA Component of Mouse Sarcoma 180 Ascites Cells. *Proceedings of the National Academy of Sciences*, 68(6), 1331–1335. <https://doi.org/10.1073/pnas.68.6.1331>
- Leppek, K., Das, R., & Barna, M. (2018). Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nature Reviews Molecular Cell Biology*, 19(3), 158–174. <https://doi.org/10.1038/nrm.2017.103>
- Ling, M. M., & Robinson, B. H. (1997). Approaches to DNA Mutagenesis: An Overview. *Analytical Biochemistry*, 254(2), 157–178. <https://doi.org/10.1006/abio.1997.2428>
- Linial, M. (2003). How incorrect annotations evolve – the case of short ORFs. *Trends in Biotechnology*, 21(7), 298–300. [https://doi.org/10.1016/S0167-7799\(03\)00139-2](https://doi.org/10.1016/S0167-7799(03)00139-2)
- Liu, C., Chu, D., Kalantar-Zadeh, K., George, J., Young, H. A., & Liu, G. (2021). Cytokines: From Clinical Significance to Quantification. *Advanced Science*, 8(15). <https://doi.org/10.1002/advs.202004433>
- Lohman, T. M., & Fazio, N. T. (2018). How Does a Helicase Unwind DNA? Insights from RecBCD Helicase. *BioEssays*, 40(6). <https://doi.org/10.1002/bies.201800009>
- Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*, 26(8), 345–352. <https://doi.org/10.1016/j.tig.2010.05.003>
- Matlin, A. J., Clark, F., & Smith, C. W. J. (2005). Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, 6(5), 386–398. <https://doi.org/10.1038/nrm1645>

- McFarland, A. P., Horner, S. M., Jarret, A., Joslyn, R. C., Bindewald, E., Shapiro, B. A., Delker, D. A., Hagedorn, C. H., Carrington, M., Gale, M., & Savan, R. (2014). The favorable IFNL3 genotype escapes mRNA decay mediated by AU-rich elements and hepatitis C virus-induced microRNAs. *Nature Immunology*, 15(1), 72–79. <https://doi.org/10.1038/ni.2758>
- MEIJER, H. A., & THOMAS, A. A. M. (2002). Control of eukaryotic protein synthesis by upstream open reading frames in the 5'-untranslated region of an mRNA. *Biochemical Journal*, 367(1), 1–11. <https://doi.org/10.1042/bj20011706>
- Mendivil Ramos, O., & Ferrier, D. E. K. (2012). Mechanisms of Gene Duplication and Translocation and Progress towards Understanding Their Relative Contributions to Animal Genome Evolution. *International Journal of Evolutionary Biology*, 2012, 1–10. <https://doi.org/10.1155/2012/846421>
- Meng, L., Song, Z., Liu, A., Dahmen, U., Yang, X., & Fang, H. (2021). Effects of Lipopolysaccharide-Binding Protein (LBP) Single Nucleotide Polymorphism (SNP) in Infections, Inflammatory Diseases, Metabolic Disorders and Cancers. *Frontiers in Immunology*, 12. <https://www.frontiersin.org/articles/10.3389/fimmu.2021.681810>
- Meyer, M. M. (2017). The role of mRNA structure in bacterial translational regulation. *WIREs RNA*, 8(1). <https://doi.org/10.1002/wrna.1370>
- Mikkola, S., Salomaki, S., Zhang, Z., Maki, E., & Lonnberg, H. (2005). Preparation and Properties of mRNA 5-cap Structure. *Current Organic Chemistry*, 9(10), 999–1022. <https://doi.org/10.2174/1385272054368402>
- Mitchell, P., & Tollervey, D. (2000). mRNA stability in eukaryotes. *Current Opinion in Genetics & Development*, 10(2), 193–198. [https://doi.org/10.1016/S0959-437X\(00\)00063-0](https://doi.org/10.1016/S0959-437X(00)00063-0)
- Moore, M. J. (2005). From Birth to Death: The Complex Lives of Eukaryotic mRNAs. *Science*, 309(5740), 1514–1518. <https://doi.org/10.1126/science.1111443>
- Neugebauer, K. M. (2019). Nascent RNA and the Coordination of Splicing with Transcription. *Cold Spring Harbor Perspectives in Biology*, 11(8), a032227. <https://doi.org/10.1101/cshperspect.a032227>
- Neyer, S., Kunz, M., Geiss, C., Hantsche, M., Hodorau, V.-V., Seybert, A., Engel, C., Scheffer, M. P., Cramer, P., & Frangakis, A. S. (2016). Structure of RNA polymerase I transcribing ribosomal DNA genes. *Nature*, 540(7634), 607–610. <https://doi.org/10.1038/nature20561>
- Northcutt, C. G., Athalye, A., & Mueller, J. (2021). Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *ArXiv Preprint ArXiv*, 2103(14749). <https://doi.org/https://doi.org/10.48550/arXiv.2103.14749>
- Nudler, E. (1999). Transcription elongation: structural basis and mechanisms. *Journal of Molecular Biology*, 288(1), 1–12. <https://doi.org/10.1006/jmbi.1999.2641>

- Obenchain, V., Lawrence, M., Carey, V., Gogarten, S., Shannon, P., & Morgan, M. (2014). VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*, 30(14), 2076–2078. <https://doi.org/10.1093/bioinformatics/btu168>
- Ollier, W. E. R. (2004). Cytokine genes and disease susceptibility. *Cytokine*, 28(4–5), 174–178. <https://doi.org/10.1016/j.cyto.2004.07.014>
- Oman, M., Alam, A., & Ness, R. W. (2022). How Sequence Context-Dependent Mutability Drives Mutation Rate Variation in the Genome. *Genome Biology and Evolution*, 14(3). <https://doi.org/10.1093/gbe/evac032>
- Orphanides, G., Lagrange, T., & Reinberg, D. (1996). The general transcription factors of RNA polymerase II. *Genes & Development*, 10(21), 2657–2683. <https://doi.org/10.1101/gad.10.21.2657>
- Parmley, J. L., & Hurst, L. D. (2007). How do synonymous mutations affect fitness? *BioEssays*, 29(6), 515–519. <https://doi.org/10.1002/bies.20592>
- Parton, R. M., Davidson, A., Davis, I., & Weil, T. T. (2014). Subcellular mRNA localisation at a glance. *Journal of Cell Science*, 127(10), 2127–2133. <https://doi.org/10.1242/jcs.114272>
- Pearson, H. (2006). What is a gene? *Nature*, 441, 398+. <https://link.gale.com/apps/doc/A632098680/HRCA?u=anon~acb74edf&sid=googleScholar&xid=0d4e5127>
- Petrokovski, S., Henikoff, J. G., & Henikoff, S. (1998). Exploring protein homology with the Blocks server. *Trends in Genetics*, 14(4), 162–163. [https://doi.org/10.1016/S0168-9525\(98\)01421-8](https://doi.org/10.1016/S0168-9525(98)01421-8)
- Plaschka, C., Hantsche, M., Dienemann, C., Burzinski, C., Plitzko, J., & Cramer, P. (2016). Transcription initiation complex structures elucidate DNA opening. *Nature*, 533(7603), 353–358. <https://doi.org/10.1038/nature17990>
- Platt, T. (1986). TRANSCRIPTION TERMINATION AND THE REGULATION OF GENE EXPRESSION. *Annual Review of Biochemistry*, 55(1), 339–372. <https://doi.org/10.1146/annurev.bi.55.070186.002011>
- Porrua, O., & Libri, D. (2015). Transcription termination and the control of the transcriptome: why, where and how to stop. *Nature Reviews Molecular Cell Biology*, 16(3), 190–202. <https://doi.org/10.1038/nrm3943>
- Proshkin, S., Rahmouni, A. R., Mironov, A., & Nudler, E. (2010). Cooperation Between Translating Ribosomes and RNA Polymerase in Transcription Elongation. *Science*, 328(5977), 504–508. <https://doi.org/10.1126/science.1184939>
- Proudfoot, N. J. (2011). Ending the message: poly(A) signals then and now. *Genes & Development*, 25(17), 1770–1782. <https://doi.org/10.1101/gad.17268411>
- R Core Team. (2018). *R: A language and environment for statistical computing* (4.3.2). R Foundation for Statistical Computing.

- Reams, A. B., & Roth, J. R. (2015). Mechanisms of Gene Duplication and Amplification. *Cold Spring Harbor Perspectives in Biology*, 7(2), a016592. <https://doi.org/10.1101/cshperspect.a016592>
- Rodrigue, N., Philippe, H., & Lartillot, N. (2010). Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proceedings of the National Academy of Sciences*, 107(10), 4629–4634. <https://doi.org/10.1073/pnas.0910915107>
- ROEDER, R. G., & RUTTER, W. J. (1969). Multiple Forms of DNA-dependent RNA Polymerase in Eukaryotic Organisms. *Nature*, 224(5216), 234–237. <https://doi.org/10.1038/224234a0>
- Ross, J. (1995). mRNA stability in mammalian cells. *Microbiological Reviews*, 59(3), 423–450. <https://doi.org/10.1128/mr.59.3.423-450.1995>
- Roy, J., & Mallick, B. (2017). Altered gene expression in late-onset Alzheimer's disease due to SNPs within 3'UTR microRNA response elements. *Genomics*, 109(3), 177–185. <https://doi.org/https://doi.org/10.1016/j.ygeno.2017.02.006>
- Rykova, E., Ershov, N., Damarov, I., & Merkulova, T. (2022). SNPs in 3'UTR miRNA Target Sequences Associated with Individual Drug Susceptibility. *International Journal of Molecular Sciences*, 23(22), 13725. <https://doi.org/10.3390/ijms232213725>
- Sarkar, A., Panati, K., & Narala, V. R. (2022). Code inside the codon: The role of synonymous mutations in regulating splicing machinery and its impact on disease. *Mutation Research/Reviews in Mutation Research*, 790, 108444. <https://doi.org/10.1016/j.mrrev.2022.108444>
- Savino, S., Desmet, T., & Franceus, J. (2022). Insertions and deletions in protein evolution and engineering. *Biotechnology Advances*, 60, 108010. <https://doi.org/10.1016/j.biotechadv.2022.108010>
- Schiavi, S. (1992). Regulation of proto-oncogene mRNA stability. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1114(2–3), 95–106. [https://doi.org/10.1016/0304-419X\(92\)90009-N](https://doi.org/10.1016/0304-419X(92)90009-N)
- Schones, D. E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., & Zhao, K. (2008). Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell*, 132(5), 887–898. <https://doi.org/10.1016/j.cell.2008.02.022>
- Schwinghammer, K., Cheung, A. C. M., Morozov, Y. I., Agaronyan, K., Temiakov, D., & Cramer, P. (2013). Structure of human mitochondrial RNA polymerase elongation complex. *Nature Structural & Molecular Biology*, 20(11), 1298–1303. <https://doi.org/10.1038/nsmb.2683>
- Sehn, J. K. (2015). Insertions and Deletions (Indels). In *Clinical Genomics* (pp. 129–150). Elsevier. <https://doi.org/10.1016/B978-0-12-404748-8.00009-5>
- Sentenac, A. (1985). Eukaryotic RNA Polymerase. *Critical Reviews in Biochemistry*, 18(1), 31–90. <https://doi.org/10.3109/10409238509082539>

- Serfling, E., Jasin, M., & Schaffner, W. (1985). Enhancers and eukaryotic gene transcription. *Trends in Genetics*, 1, 224–230. [https://doi.org/10.1016/0168-9525\(85\)90088-5](https://doi.org/10.1016/0168-9525(85)90088-5)
- Sharova, L. V., Sharov, A. A., Nedorezov, T., Piao, Y., Shaik, N., & Ko, M. S. H. (2009). Database for mRNA Half-Life of 19 977 Genes Obtained by DNA Microarray Analysis of Pluripotent and Differentiating Mouse Embryonic Stem Cells. *DNA Research*, 16(1), 45–58. <https://doi.org/10.1093/dnares/dsn030>
- Shastri, B. S. (2009). SNPs: Impact on Gene Function and Phenotype. In A. A. Komar (Ed.), *Single Nucleotide Polymorphisms: Methods and Protocols* (pp. 3–22). Humana Press. https://doi.org/10.1007/978-1-60327-411-1_1
- Sherry, S. T., Ward, M., & Sirotkin, K. (1999). dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. *Genome Research*, 9(8), 677–679. <https://doi.org/10.1101/gr.9.8.677>
- Shulman, E. D., & Elkon, R. (2020). Systematic identification of functional SNPs interrupting 3'UTR polyadenylation signals. *PLOS Genetics*, 16(8), e1008977-. <https://doi.org/10.1371/journal.pgen.1008977>
- Sidali, A., Teotia, V., Solaiman, N. S., Bashir, N., Kanagaraj, R., Murphy, J. J., & Surendranath, K. (2021). AU-Rich Element RNA Binding Proteins: At the Crossroads of Post-Transcriptional Regulation and Genome Integrity. *International Journal of Molecular Sciences*, 23(1), 96. <https://doi.org/10.3390/ijms23010096>
- Sidorenkov, I., Komissarova, N., & Kashlev, M. (1998). Crucial Role of the RNA:DNA Hybrid in the Processivity of Transcription. *Molecular Cell*, 2(1), 55–64. [https://doi.org/10.1016/S1097-2765\(00\)80113-6](https://doi.org/10.1016/S1097-2765(00)80113-6)
- Simpson, L., & Emeson, R. B. (1996). RNA Editing. *Annual Review of Neuroscience*, 19(1), 27–52. <https://doi.org/10.1146/annurev.ne.19.030196.000331>
- Sitbon, E., & Pietrokovski, S. (2007). Occurrence of protein structure elements in conserved sequence regions. *BMC Structural Biology*, 7(1), 3. <https://doi.org/10.1186/1472-6807-7-3>
- Smale, S. T., & Kadonaga, J. T. (2003). The RNA Polymerase II Core Promoter. *Annual Review of Biochemistry*, 72(1), 449–479. <https://doi.org/10.1146/annurev.biochem.72.121801.161520>
- Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., & Kasprzyk, A. (2009). BioMart – biological queries made easy. *BMC Genomics*, 10(1), 22. <https://doi.org/10.1186/1471-2164-10-22>
- Smith, C. W. J., Patton, J. G., & Nadal-Ginard, B. (1989). ALTERNATIVE SPLICING IN THE CONTROL OF GENE EXPRESSION. *Annual Review of Genetics*, 23(1), 527–577. <https://doi.org/10.1146/annurev.ge.23.120189.002523>
- Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T. A., & Soreq, H. (2005). Function of alternative splicing. *Gene*, 344, 1–20. <https://doi.org/10.1016/j.gene.2004.10.022>

- Stojanovic, N. (1999). Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Research*, 27(19), 3899–3910. <https://doi.org/10.1093/nar/27.19.3899>
- Tian, B., & Graber, J. H. (2012). Signals for pre-mRNA cleavage and polyadenylation. *WIREs RNA*, 3(3), 385–396. <https://doi.org/10.1002/wrna.116>
- Tian, B., & Manley, J. L. (2017). Alternative polyadenylation of mRNA precursors. *Nature Reviews Molecular Cell Biology*, 18(1), 18–30. <https://doi.org/10.1038/nrm.2016.116>
- Tuller, T., & Zur, H. (2015). Multiple roles of the coding sequence 5' end in gene expression regulation. *Nucleic Acids Research*, 43(1), 13–28. <https://doi.org/10.1093/nar/gku1313>
- Vakil, M. K., Mansoori, Y., Al-Awsi, G. R. L., Hosseini-pour, A., Ahsant, S., Ahmadi, S., Ekrahi, M., Montaseri, Z., Pezeshki, B., Mohaghegh, P., Sohrabpour, M., Bahmanyar, M., Daraei, A., Dadkhah Jouybari, T., Tavassoli, A., & Ghasemian, A. (2022). Individual genetic variability mainly of Proinflammatory cytokines, cytokine receptors, and toll-like receptors dictates pathophysiology of COVID-19 disease. *Journal of Medical Virology*, 94(9), 4088–4096. <https://doi.org/https://doi.org/10.1002/jmv.27849>
- van der Velden, A. W., & Thomas, A. A. M. (1999). The role of the 5' untranslated region of an mRNA in translation regulation during development. *The International Journal of Biochemistry & Cell Biology*, 31(1), 87–106. [https://doi.org/10.1016/S1357-2725\(98\)00134-4](https://doi.org/10.1016/S1357-2725(98)00134-4)
- Vassilyev, D. G., Vassilyeva, M. N., Perederina, A., Tahirov, T. H., & Artsimovitch, I. (2007). Structural basis for transcription elongation by bacterial RNA polymerase. *Nature*, 448(7150), 157–162. <https://doi.org/10.1038/nature05932>
- Via, M., Gignoux, C., & Burchard, E. (2010). The 1000 Genomes Project: new opportunities for research and social challenges. *Genome Medicine*, 2(1), 3. <https://doi.org/10.1186/gm124>
- WARRINGTON, J. A., NAIR, A., MAHADEVAPPA, M., & TSYGANSKAYA, M. (2000). Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiological Genomics*, 2(3), 143–147. <https://doi.org/10.1152/physiolgenomics.2000.2.3.143>
- Watson JD, Hopkins NH, Roberts JW, Steitz JA, & Weiner AM. (1987). *Molecular Biology of the Gene* (4th ed.) (4th ed., Vol. 1). Benjamin/Cummings.
- Whiteside, T. L. (1994). Cytokines and cytokine measurements in a clinical laboratory. *Clinical Diagnostic Laboratory Immunology*, 1(3), 257–260. <https://doi.org/10.1128/cdli.1.3.257-260.1994>
- Wilson, G. M., & Brewer, G. (1998). *The Search for Trans-Acting Factors Controlling Messenger RNA Decay* (pp. 257–291). [https://doi.org/10.1016/S0079-6603\(08\)60510-3](https://doi.org/10.1016/S0079-6603(08)60510-3)
- Wissink, E. M., Vihervaara, A., Tippens, N. D., & Lis, J. T. (2019). Nascent RNA analyses: tracking transcription and its regulation. *Nature Reviews Genetics*, 20(12), 705–723. <https://doi.org/10.1038/s41576-019-0159-6>

- Wjst, M. (2004). Target SNP selection in complex disease association studies. *BMC Bioinformatics*, 5(1), 92. <https://doi.org/10.1186/1471-2105-5-92>
- Wolfe, K. H., Sharp, P. M., & Li, W.-H. (1989). Mutation rates differ among regions of the mammalian genome. *Nature*, 337(6204), 283–285. <https://doi.org/10.1038/337283a0>
- Wu, X., & Brewer, G. (2012). The regulation of mRNA stability in mammalian cells: 2.0. *Gene*, 500(1), 10–21. <https://doi.org/10.1016/j.gene.2012.03.021>
- Xie, J., Libri, D., & Porrua, O. (2023). Mechanisms of eukaryotic transcription termination at a glance. *Journal of Cell Science*, 136(1). <https://doi.org/10.1242/jcs.259873>
- Xu, H., Lin, S., Zhou, Z., Li, D., Zhang, X., Yu, M., Zhao, R., Wang, Y., Qian, J., Li, X., Li, B., Wei, C., Chen, K., Yoshimura, T., Wang, J. M., & Huang, J. (2023). New genetic and epigenetic insights into the chemokine system: the latest discoveries aiding progression toward precision medicine. *Cellular & Molecular Immunology*, 20(7), 739–776. <https://doi.org/10.1038/s41423-023-01032-x>
- Yip, K. Y., Cheng, C., Bhardwaj, N., Brown, J. B., Leng, J., Kundaje, A., Rozowsky, J., Birney, E., Bickel, P., Snyder, M., & Gerstein, M. (2012). Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biology*, 13(9), R48. <https://doi.org/10.1186/gb-2012-13-9-r48>
- Young, R. A. (1991). RNA POLYMERASE II. *Annual Review of Biochemistry*, 60(1), 689–715. <https://doi.org/10.1146/annurev.bi.60.070191.003353>
- Zerbini, F. M., Silva, F. N. da, Urquiza, G. P. C., & Basso, M. F. (2014). Transgenic Plants. In *Biotechnology and Plant Breeding* (pp. 179–199). Elsevier. <https://doi.org/10.1016/B978-0-12-418672-9.00008-8>
- Zhang, F., Gu, W., Hurler, M. E., & Lupski, J. R. (2009). Copy Number Variation in Human Health, Disease, and Evolution. *Annual Review of Genomics and Human Genetics*, 10(1), 451–481. <https://doi.org/10.1146/annurev.genom.9.081307.164217>
- Zhang, J.-M., & An, J. (2007). Cytokines, Inflammation, and Pain. *International Anesthesiology Clinics*, 45(2), 27–37. <https://doi.org/10.1097/AIA.0b013e318034194e>
- Zhao, B. S., Roundtree, I. A., & He, C. (2017). Post-transcriptional gene regulation by mRNA modifications. *Nature Reviews Molecular Cell Biology*, 18(1), 31–42. <https://doi.org/10.1038/nrm.2016.132>
- Zhu, J., He, F., Hu, S., & Yu, J. (2008). On the nature of human housekeeping genes. *Trends in Genetics*, 24(10), 481–484. <https://doi.org/10.1016/j.tig.2008.08.004>
- Zlotnik, A., Yoshie, O., & Nomiya, H. (2006). The chemokine and chemokine receptor superfamilies and their molecular evolution. *Genome Biology*, 7(12), 243. <https://doi.org/10.1186/gb-2006-7-12-243>

Supplementary Materials

Supplementary Table 1: Snippet of the Gene IDs for Housekeeping genes

FAM193A	ARL6IP1	IFITM1	IRAK1	MYH9	CD34
RPLP2	KARS1	PGAM1	PSEN1	STMN1	TMSB10
COX6B1	AARS1	LAMP1	ACTB	RPS27	DSTN
HYAL2	HLA-A	LGALS1	LINC02967	HLA-DOA	SERPINB6
RER1	PSMD8	SNRPN	FTL	XBP1	TXNIP
ANXA2	RFTN1	EEF1A1	H3-3A	RPS15	BRD2
MARCKS	VPS72	COX8A	ESD	CD81	RPS27A
ATP5MC1	ATP6V1F	HSP90AB1	CD74	FKBP1A	B2M
APEX1	HLA-B	RPA2	RPS14	HLA-DQA1	HNRNPA1
RPL24	PSMD7	HLA-DQB1	RPSA	RPL7A	ARAF
CCND2	CCNI	PSMC1	ENO1	ARF4	NONO
EIF4A1	MDH1	RPS21	PTMA	H2AZ1	ARHGAP1
ECHS1	HMGB1	RPL37A	FYN	S100A10	TLE5
GDI2	SEPTIN2	RPL12	GUSB	PIM1	FMOD
CD164	CIRBP	PSME1	RPL36AL	EEF1G	RPL23A
RPS23	PSMD2	VDAC2	HNRNPC	LGALS3	MLH1
SPCS2	OAZ1	APLP2	TMSB4X	UBA1	SRP14
PRDX6	YWHAH	PLP2	RPLP0	RPS16	ATP5MC3
PTDSS1	RPL39	CANX	RPLP1	NCL	RPL9P7
ATP5F1A	BCLAF1	HPN	RPS17	CLU	BUD31
MORF4L2	ATIC	FNTA	ACTG1	RPS25	FCGRT
CAST	RBPMS	RPL18	CALM2	C1QBP	RPL10A
ATP5F1C	ERH	ZNF91	ATP5F1B	NFKBIA	RPL35
CLTC	DVL3	CAP1	CLTA	MSN	PTP4A2
TAGLN2	NPIP3	RPL27	GNAS	ARF3	RPL27A
EIF4A3	HTRA1	ANXA11	HSPD1	RPS6	RPL28
RPL4	NCSTN	H2AC18	NDUFV2	TCEA1	RPS5
NEDD8	RPL14	ARHGDIB	CAPN2	RPS19	RPS9
MPRIIP	NDRG1	FTH1	CD3E	ARF1	RPS10
RAC1	PFDN5	RHOA	NPM1	RPS3A	RPS29
EIF4H	ATF4	SEC61B	RACK1	YWHAZ	SNRPD2
PSMB3	B2M	EIF1	GSTP1	NAP1L1	GPS1
PSMB2	SLC25A5	HNRNPF	PTMA	FKBP4	COPA
PSMB4	PSAP	STAT3	UQCRB	ZFP36	RPL21

Supplementary Table 2: List of Gene IDs for Cytokine genes

IL1A	CD70	AGIF
IL1B	CD153	CSF3
IL1RA	CD178	NK CELL STIMULATORY FACTOR
IL18	TRAIL	LEUKEMIA INHIBITORY FACTOR
IL2	TWEAK	ONCOSTATIN M
IL4	TRANCE	CSIF
IL9	TGFB1	HMWBCGF
IL13	EPO	LCF
IL3	TPO	CTLA8
IL3	SCF	CD40L
IL5	SCF	TRAP
GMCSF	MCSF	CACHECTIN
IL6	MSP	LTA
IL6	MSP	TALL2
IL11	HEMATOPOIETIN1	CD27L
GCSF	CATABOLIN	CD30L
IL12	IL1 RECEPTOR ANTAGONIST	FASL
LIF	INTERFERON γ INDUCING FACTOR	APO2L
OSM	T CELL GROWTH FACTOR	APO3L
IL10	BSF1	OPGL
IL14	T CELL GROWTH FACTOR P40	TGFB
IL16	P600	ERYTHROPOIETIN
IL17	MULTIPOTENTIAL CSF	MGDF
CD154	MCGF	STEM CELL FACTOR
CD154	BCDF1	CKIT LIGAND
TNFA	CSF2	CSF1
TNFB	IFNB2	MACROPHAGE STIMULATING FACTOR
APRIL	BSF2	MST1

Supplementary Table 3: List of Gene IDs for Chemokine genes

XCL1	CCL22	SCM1A	STCP1
XCL2	CCL23	SCM1B	CKB8
CCL1	CCL24	MCP1	CKB6
CCL3	CCL26	MIP1B	ESKINE
CCL4	CCL27	LAG1	CTACK
CCL5	CXCL1	MCP3	MGSAA
CCL13	CXCL2	HCC1	GROG
CCL14	CXCL3	HCC2	MGSAG
CCL16	CXCL7	LEC	PPBP
CCL17	CXCL8	MIP3B	MDNCF
CCL20	CXCL13	6CKINE	BCA1

Supplementary Table 4: Snippet of the final database for 5'UTR

5_utr_start	5_utr_end	utr_length	norm_count	hgnc_symbol
923923	924431	508	43	SAMD11
925150	925189	39	26	SAMD11
925922	925941	19	53	SAMD11
925731	925800	69	14	SAMD11
960584	960693	109	28	KLHL17
966482	966531	49	61	PLEKHN1
966502	966531	29	69	PLEKHN1
982065	982093	28	36	PERM1
981030	981173	143	42	PERM1
981167	981173	6	167	PERM1
982065	982117	52	58	PERM1
999974	1000172	198	45	HES4
999974	1000097	123	49	HES4
1001138	1001281	143	28	ISG15
1008194	1008279	85	12	ISG15
1013984	1014004	20	50	ISG15
1001145	1001263	118	25	ISG15
1013497	1013573	76	39	ISG15
1020120	1020172	52	19	AGRN
1020123	1020373	250	28	AGRN
1022201	1022413	212	14	AGRN
1034106	1034555	449	27	AGRN
1074016	1074306	290	41	RNF223
1116060	1116081	21	95	C1orf159
1091991	1092103	112	27	C1orf159
1091544	1091565	21	48	C1orf159
1116060	1116089	29	103	C1orf159
1116060	1116095	35	86	C1orf159
1091544	1092174	630	43	C1orf159
1105956	1106127	171	29	C1orf159
1116060	1116356	296	30	C1orf159
1105956	1106049	93	54	C1orf159
1116060	1116361	301	30	C1orf159
1091544	1091609	65	77	C1orf159
1092566	1092813	247	28	C1orf159
1106650	1106871	221	45	C1orf159
1173880	1173926	46	22	TTLL10
1174285	1174321	36	28	TTLL10
1174424	1174489	65	31	TTLL10
1179189	1179215	26	38	TTLL10

Supplementary Table 5: Snippet of the final database for CDS

genomic_coding_start	genomic_coding_end	cds_length	norm_count	hgnc_symbol
69037	70008	971	2	OR4F5
924432	924948	516	23	SAMD11
925922	926013	91	11	SAMD11
930155	930336	181	44	SAMD11
931039	931089	50	20	SAMD11
935772	935896	124	73	SAMD11
939040	939129	89	56	SAMD11
939275	939412	137	51	SAMD11
941144	941306	162	19	SAMD11
942136	942251	115	43	SAMD11
942410	942488	78	26	SAMD11
942559	943058	499	32	SAMD11
943253	943377	124	73	SAMD11
943698	943808	110	36	SAMD11
943908	944153	245	69	SAMD11
939272	939412	140	50	SAMD11
935772	935793	21	48	SAMD11
939275	939460	185	43	SAMD11
939272	939460	188	43	SAMD11
943971	944153	182	66	SAMD11
930312	930336	24	125	SAMD11
943698	944153	455	57	SAMD11
959215	959240	25	40	NOC2L
958929	959081	152	13	NOC2L
957099	957273	174	40	NOC2L
956894	957025	131	31	NOC2L
956095	956215	120	42	NOC2L
955923	956013	90	33	NOC2L
954004	954082	78	77	NOC2L
953782	953892	110	36	NOC2L
953175	953288	113	53	NOC2L
952412	952600	188	27	NOC2L
952000	952139	139	58	NOC2L
951127	951238	111	18	NOC2L
948490	948603	113	35	NOC2L
948131	948232	101	20	NOC2L
946402	946545	143	28	NOC2L
946173	946286	113	35	NOC2L
945518	945653	135	37	NOC2L
945057	945146	89	34	NOC2L

Supplementary Table 6: Snippet of the final database for Housekeeping Genes

hgnc_symbol	utr_length	count_50bp	count_175_225bp
AARS1	532	1	0
AARS1	417	2	2
AARS1	415	2	2
AARS1	596	1	2
AARS1	413	1	2
AARS1	422	2	2
AARS1	450	2	2
ABLIM1	305	1	2
ACTB	595	0	3
ACTB	460	0	3
ACTB	344	1	2
ACTB	438	1	2
ACTB	599	0	3
ACTB	600	0	3
ACTB	595	0	3
ACTB	460	0	3
ACTB	344	1	2
ACTB	438	1	2
ACTB	599	0	3
ACTB	600	0	3
ACTG1	352	3	3
ACTG1	316	3	4
ACTG1	438	2	0
ACTG1	354	4	2
ACTG1	416	0	3
ANP32B	511	0	1
ANXA2	545	1	0
ANXA2	485	1	0
ANXA2	422	1	5
ANXA2	478	2	0
ANXA2	352	0	1
AP3S1	590	0	3
APLP2	357	4	0
ARAF	450	2	1
ARF1	365	2	0
ARF3	392	1	1
ARHGDIB	476	0	2
ARHGDIB	490	0	2
ARL6IP1	304	1	0
ARL6IP1	491	0	1

Supplementary Table 7: Final database for Cytokine Genes

hgnc_symbol	utr_length	count_50bp	count_175_225bp
CSF2	317	1	0
EPO	566	3	2
IL10	525	2	2
IL11	447	1	4
IL16	493	1	0
IL18	342	0	1
IL3	412	1	1
IL3	412	1	1
IL5	365	0	1
IL6	424	1	2
IL6	428	1	2
IL6	424	1	2
IL6	428	1	2
OSM	360	2	1
TPO	543	3	0
TPO	540	3	0

Supplementary Table 8: Final database for Chemokine Genes

hgnc_symbol	utr_length	count_50bp	count_175_225bp
CCL13	478	0	3
CCL14	509	2	2
CCL14	512	1	2
CCL15	486	0	3
CCL15	486	0	3
CCL15	486	0	3
CCL15	486	0	3
CCL2	375	1	2
CCL2	375	1	2
CCL21	388	1	0
CCL21	388	1	0
CCL21	388	1	0
CCL25	424	0	4
CCL25	432	0	3
CCL25	434	0	4
CCL25	396	0	4
CCL25	420	0	4
CCL25	424	0	4
CCL25	432	0	3
CCL25	434	0	4
CCL25	396	0	4
CCL25	420	0	4
CCL3	415	2	1
CCL3	415	2	1
CCL4	301	2	2
CCL4	313	2	2
CCL4	301	2	2
CCL4	313	2	2
CCL4	301	2	2
CCL4	313	2	2
CCL5	300	0	1
CCL7	381	0	1
CCL7	439	0	1
CCL8	494	1	2
CX3CL1	350	0	1