Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Scienze Farmacologiche e Tossicologiche, dello Sviluppo e del Movimento Umano

Ciclo XXVI

Settore Concorsuale di afferenza: 05/F1 - BIOLOGIA APPLICATA

Settore Scientifico disciplinare: BIO/13 - BIOLOGIA APPLICATA

New computational biology tools for the systematic analysis of the structure and expression of human genes

Presentata da: **Allison Piovesan**

Coordinatore Dottorato **Relatore**

Chiar.mo Prof. Giorgio Cantelli Forti **Chiar.mo Prof. Pierluigi Strippoli**

Esame finale anno 2014

Abstract

From the late 1980s, the automation of sequencing techniques and the computer spread gave rise to a flourishing number of new molecular structures and sequences and to proliferation of new databases in which to store them. Here are presented three computational approaches able to analyse the massive amount of publicly avalilable data in order to answer to important biological questions.

The first strategy studies the incorrect assignment of the first AUG codon in a messenger RNA (mRNA), due to the incomplete determination of its $5'$ end sequence. An extension of the mRNA $5'$ coding region was identified in 477 in human loci, out of all human known mRNAs analysed, using an automated expressed sequence tag (EST)-based approach. Proof-of-concept confirmation was obtained by in vitro cloning and sequencing for *GNB2L1*, *QARS* and *TDP2* and the consequences for the functional studies are discussed.

The second approach analyses the codon bias, the phenomenon in which distinct synonymous codons are used with different frequencies, and, following integration with a gene expression profile, estimates the total number of codons present across all the expressed mRNAs (named here "codonome value") in a given biological condition. Systematic analyses across different pathological and normal human tissues and multiple species shows a surprisingly tight correlation between the codon bias and the codonome bias.

The third approach is useful to studies the expression of human autism spectrum disorder (ASD) implicated genes. ASD implicated genes sharing microRNA response elements (MREs) for the same microRNA are co-expressed in brain samples from healthy and ASD affected individuals. The different expression of a recently identified long non coding RNA which have four MREs for the same microRNA could disrupt the equilibrium in this network, but further analyses and experiments are needed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Computational Biology

The story of Computational Biology started in the mid 1940s, when Fred Sanger published his first work on insulin [Sanger, 1945]. He used chemical and enzymatic experiments to fragment the protein and to deduce from the amino acid sequence of the fragments, the order of the amino acid in the intact protein. This led to the complete primary structure of bovine, ovine and porcine insulin ten years later [e.g. Sanger *et al*., 1955; Ryle *et al*., 1955; Brown *et al*., 1955]. It was the first time that the order of amino acids of a protein was determined. Many years later the final amino acid sequence of the first enzyme, the ribonuclease, was published [Smyth *et al*., 1963]. Consequently, the sequence of many other proteins was soon deduced. Margaret Dayhoff was the first to appreciate the importance of databases, the utility of organising biological sequences and the value of sequence comparative analysis. She began to collect all available protein sequences in order to facilitate her and others researches and pulished them in book form, the "Atlas of Protein Sequence and Structure" [Dayhoff *et al*., 1965]. Consequently, the advent of automated peptide sequencers increased the rate of sequence determination considerably. The numbers of deduced protein structures grew accordingly and it became soon necessary to develop a system in order to collect, correlate and interpret this significant information. The Cambrige Structural Database, a repository of small-molecule crystal structures, is one of the oldest dabasases being established in 1965. The common belief among scientists was that the collective use of data would lead to the discovery of new knowledge which goes beyond the results yielded by individual experiments [Kennard, 1997]. Many databases were thus established, among the others the Protein Data Bank (PDB), in 1971, in order to collect protein coordinate

data [Attwood *et al.*, 2011].

Despite the progresses in the sequence and structure determination of proteins, sequencing nucleic acids was still problematic, due to issues related to their greater size and to the difficulties during purification. In 1977, Sanger developed a technology (now known as the "Sanger method") that made possible to work with longer nucleotide fragments [Sanger *et al.*, 1977], giving origin to the field of reverse genetics. This allowed the completion of the bacteriophage phiX174, the human mitocondrial DNA and the lambda bacteriophage genomes sequencing [Sanger *et al.*, 1978; Anderson *et al.*, 1981; Sanger *et al.*, 1982] and brought cloning and sequencing into any laboratory worldwide. As well as happened many years ago for proteins, with the enormous increase in the rate of sequencing DNA fragments, a large computerised database of sequences became essential for research in molecular biology and several groups worldwide were engaged in the collection of nucleic acid sequences. This was also the time when there was the computer hardware revolution; thus tool needed to store, search and analyse new data developed alongside the tools necessary to generate the data [Smith, 1990]. In 1980, the first internationally supported resource for nucleotide sequence data was established by the European Molecular Biology Laboratory (EMBL) at Heidelberg (Germany), with its first release in April 1982. The EMBL Data Library goals were to make nucleic acid sequence data publicly available in the international molecular biology community, to encourage standardisation and free exchange of data and also to provide a European focus for computational and biological data service [Hamm and Cameron, 1986]. Meanwhile the necessity of creating an international nucleic acid sequence repository emerged also out of Europe, giving birth to GenBank, with its first release in December 1982 [Benson *et al.*, 1990]. From the beginning, GenBank and the EMBL Data Library evolved in close collaboration. They were distributed at first on magnetic tape and then on CD-ROM to anyone interested, free of charge, in order to promote scientific progress. From 1986 onwards, they started to collaborate also with the DNA Data Bank of Japan, adopting common data-entry standards and data-exange protocol in order to improve data quality and to manage the annotation of the exponential growing entries more effectively; they are currently keeping pace with the literature [Attwood *et al.*, 2011].

The late 1980s and early 1990s were fertile years thanks to the automation of sequencing techniques and to the computer spread. This period of fervent activity gave rise to a flourishing number of new molecular structures and sequences and to proliferation of new databases in which to store them. Among others, in 1991 was published a method to rapidly obtain partial RNA sequences [Adams *et al.*, 1991].

The RNA, extracted from various tissues, is converted in cDNA (complementary DNA) by the reverse transcriptase and cloned using cDNA libraries. Then each bacterial clone, containing the partial cDNA sequence complementary to the RNA expressed in the tissue from which it was extracted, is sequenced with an automated method. The obtained sequences are called expressed sequence tags (EST) and are collected in the dbEST database, which represent an important resource in diverse biological research fields [Boguski *et al.*, 1993]. This unprecedented burst of sequencing activity yielded the first complete sequenced genome: the *Haemophilus influenzae* genome [Fleischmann *et al.*, 1995], followed by a flourishing of other organism genome reports, to the human genome [Lander *et al.*, 2001; Venter *et al.*, 2001; International Human Genome Sequencing Consortium, 2004]. Together with these activities came the development of numerous databases to store and display the emerging genomic data, e.g. Ensembl [http://www.ensembl.org/; Hubbard *et al.*, 2002], the Map Viewer at the National Center for Biotechnology Information (NCBI) [http://www.ncbi.nlm.nih.gov/mapview/; Wheeler *et al.*, 2005] and the University of California at Santa Cruz (UCSC) Genome Browser [http://genome-euro.ucsc.edu/cgi-bin/hgGateway; Kent *et al.*, 2002]. They became the main sources for information about specific genes and encoded polypeptides, and the starting point for further experimentation. The term "bioinformatics" appeared in those years to indicate the research, development, or application of computational tools and approaches useful to solve problems related to management and analysis of biological data [Biomedical Information Science and Technology Consortium Definition Committee, released on July 17, 2000]. Therefore, algorithms to search these databases became a necessity and this was the time that came the Basic Local Alignment Search Tool, BLAST [Altschul *et al.*, 1997]; this offered an extended tool-set to apply any kind of sequence database search, and is still the most widely used tool in computational biology and still in continuous development [Camacho *et al.*, 2009]. The success of BLAST led to a number of more specialised sequence search methods, such as PSI-BLAST, PHI-BLAST and BLAT (BLAST-like Alignment Tool). Aside from these very popular database search tools, many other sequence, annotation and expression analysis tools were developed for a broad range of applications: e.g., for pattern recognition, for protein and RNA secondary structure prediction, for microarray data analysis and for proteome and genome annotation [Attwood *et al.*, 2011]. The public availabilty of data is of an unestimated valour, because starting from the sequence comparison, without any laboratory experiment, is possible to answer to important biological questions. The expression *in silico*, by analogy with *in vivo* and *in vitro* experiments, is usually used to

refer to this kind of approach. In this work we will see how the vast amount of public available data and the development of new computational tool to analyse it are useful to study the structure and the expression of human genes.

1.2 Human genes

A gene is the basic physical and functional unit of heredity. It is a sequence of DNA converted into a strand of so-called messenger RNA (mRNA) during the process called transcription. An mRNA could be used as the basis for building its associated molecule called protein [Pearson, 2006]. Briefly, a tipical human coding mRNA structure include a coding region (CDS, coding DNA sequence) read into the ribosome by the transfer RNA three nucleotides at time. There are 64 combinations of three nucleotides, called codons, and each codon encode for a specific amino acid (20 in all) according to the genetic code rules (Fig. 1.1). The first codon of the coding region is the translation initiation codon (AUG), which encode for a methionine (M); the last codon is one of the three stop codons (UAA, UAG, UGA), which does not specify for any amino acid and terminate the protein synthesis. The coding region is surrounded by two non coding regions: the $5'$ UTR (untranslated region) and the $3'$ UTR, surrounded in turn by the 7-methyl guanosine cap and the poly-adenine tail (poly-A), respectively (Fig. 1.2) [Alberts *et al.*, 1994].

The Human Genome Project has estimated that humans have between 20,000 and 25,000 protein coding genes [International Human Genome Sequencing Consortium, 2004]. This relatively "low" number correlated with the fact that two-thirds of the human genome are pervasively transcribe, means that the other expressed genes encode for non coding RNAs (ncRNAs). ncRNAs are generally divided into two classes based on an arbitrary length cut-off of 200 nucleotides. Those under 200 nucleotides are usually referred to as short/small ncRNAs, including the microRNAs. microRNAs are generally 21 to 25 nucleotides long and are integral components of RNA-induced silencing complex (RISC); they recognise partially complementary target mRNAs, termed microRNA response elements (MREs), to induce translational repression or mRNA degradation. ncRNAs greater than 200 nucleotides are known as long non coding RNAs (lncRNAs). The lncRNA intrinsic nucleic acid nature confers the ability to function as ligands for proteins and to bind specific RNA or DNA target sites in order to regulate gene espression [Fatica and Bozzoni, 2014]. In the following paragraphs, three aspects regarding the structure and expression of human genes studied in this thesis will be introduced.

1st position (5' end)	2nd position U	2nd position C	2nd position A	2nd position G	3rd position (3' end)
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Figure 1.1: The genetic code. Sets of three nucleotides (codons) in an mRNA molecule are translated into amino acids in the course of protein synthesis according to the rules shown.

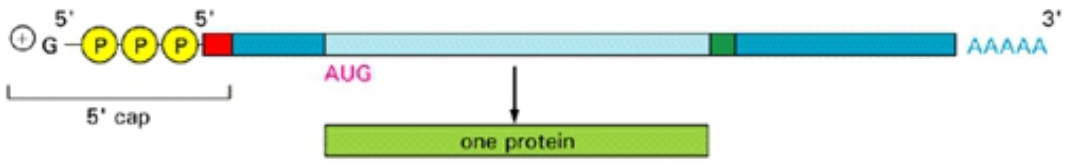


Figure 1.2: The structure of an mRNA. On the left the 7-methyl guanosine cap ($5'$ cap), which is part of the structure recognised by the small ribosomal subunit (in red). The A sequence is the poly-A tail. Light blue: the coding region with its initiation codon AUG. Dark green: the stop codon. Dark blue: the $5'$ UTR (untranslated region) and the $3'$ UTR.

1.3 mRNA 5′ coding sequence

The term "5′ end mRNA artifact" refers to the incorrect assignment of the first AUG codon in an mRNA, due to the incomplete determination of its 5′ end sequence [Casadei *et al.*, 2003]. Since the 1970s, the amino acid sequence of gene products has been routinely deduced from the nucleotide sequence of the relative cloned cDNA, according to rules for recognition of the start codon (first-AUG rule, optimal sequence context) and the genetic code [Kozak, 2002]. All standard methods for the cDNA cloning are affected by a potential inability to effectively clone the 5′ region of the mRNA [Sambrook and Russel, 2001]. This is due to the reverse transcriptase failure to extend first-strand cDNA along the full length of the mRNA template toward its 5′ end [Sambrook and Russel, 2001]. These incomplete clone sequences consequently lead to the incorrect assignment of the first AUG codon. The identification of a more complete mRNA 5′ end could reveal an additional upstream AUG, in-frame with the previously determined one, thus extending the predicted amino terminus sequence of the product and avoiding subsequent relevant errors in the experimental study of the relative cDNA [Casadei *et al.*, 2003]. An incomplete amino terminus sequence could therefore lead to errors in *in vitro* expression of proteins and in the further functional assays.

Methods to determine the full-length mRNA sequence on a large scale have been developed, such as 5′ cap trapping [Carninci *et al.*, 1996], cap analysis of gene expression (CAGE) [Kodzius *et al.*, 2006], systematic empirical annotation of a set of transcript products by 5′ rapid amplification of cDNA ends (RACE) and high-density resolution tiling arrays [Denoeud *et al.*, 2007]. However, they are experimentally labor-intensive and they have not been widely applied in comparison with the standard EST approach for fast characterisation of cDNAs [Adams *et al.*, 1991; Boguski *et al.*, 1993].

An easy and efficient computational approach to revise all the known mRNA sequences could be to compare all mRNA sequences with all EST sequences, both publicly available on the relative database (RefSeq and dbEST, respectively) and thus find EST sequences matching an mRNA 5′ end and extending it upstream. A previous manual analysis confirmed the utility of this approach, identifying sixty putative incomplete mRNAs out of the 109 human chromosome 21 protein-coding genes considered and cloning five of them [Casadei *et al.*, 2003]. The success of this approach encouraged the authors to develop a piece of software ("5'_ORF_Extender" software) in order to automate the steps that were previously performed manually. They applied this software to the *Danio rerio* (zebrafish) genome and identified a

putative extended mRNA 5′ end in the 3.3% of mRNA analysed, experimentally confirming three example cases [Frabetti *et al*., 2007]. However, it proved difficult to simply transfer the method used for *D. rerio* to *Homo sapiens*, due to the much larger size and complexity of RNA and EST sequence databases as well as the sequence analysis results file and a fully revised computational biology strategy should be adopted.

1.4 Codon bias

Codon bias is the well-known phenomenon in which distinct synonymous codons (different codons encoding the same amino acid) are used with different frequencies (reviewed in [Hershberg and Petrov, 2008]). This has been observed in species from all taxa. The codons that are used more frequently are also referred to as preferred codons or "optimal codons" [Ikemura, 1981]. Previously, optimal and non-optimal codons for each amino acid had been shown to differ between species [Grantham *et al*., 1980], in particular between distantly related species.

Codon bias can be explained by two hypotheses: the mutational (or neutral) explanation and the selectionist (or natural selection) explanation [Plotkin and Kudla, 2011]. According to the mutational explanation, codon bias originates from basal mutational processes, which cause neither advantage nor damage. The selectionist explanation asserts that synonymous mutations influence the fitness of an organism, and can thus be promoted (or repressed) throughout evolution. These two types of mechanisms are not mutually exclusive, and both are useful to understanding the phenomenon within and between species. In particular, the latter explanation is typically cited to explain variation in codon usage across a genome or across a gene [Plotkin and Kudla, 2011].

In eukaryotic genes, the most frequently used codons have a bigger content of G+C at the third codon position [Ikemura, 1985], especially in human genes, according to the mutational (or neutral) explanation of the intra-genomic heterogeneity of the human genome [Sueoka and Kawanishi, 2000]. Preferred codons also vary between genes of the same organism: expressed genes have a codon usage pattern, different from poorly expressed genes, optimised to increase translational efficiency [Ikemura, 1985] and to minimise the cost of nonsense errors during protein translation [Gilchrist *et al*., 2009]. For example, optimal codons are recognised by more abundant transfer RNA molecules in several unicellular organisms [Kanaya *et al*., 1999] and in several eukaryotes [Kanaya *et al*., 2001]. These

findings support the selectionist explanation (natural selection).

Intriguingly, Plotkin *et al*. [Plotkin *et al*., 2004] studied the role of codon usage between tissue-specific human genes. Comparing testis- to uterus-specific genes and brain- to liver-specific genes, they reported a characteristic codon usage in genes expressed in one tissue as compared to those expressed in another. Other comparisons (e.g. liver versus uterus) do not exhibit any significantly different codon usage. However, the authors suggested that codon bias might optimise translation of tissue-specific genes. Furthermore Sémon *et al*. [Sémon *et al*., 2006], analysing 2,126 human tissue-specific genes expressed in 18 different tissues, found that the difference in synonymous codon usage between tissue-specific genes expressed in different tissues is significant, but weak, as the intra-tissue variability of synonymous codon usage is much smaller than the inter-tissue variability. Additionally, these authors correlated the synonymous codon usage variability to inter-gene G+C content at the third position differences, also affecting introns and intergenic regions, due to the isochore scale variation of substitution patterns [Sémon *et al*., 2006].

At present several indexes are used to analyse codon bias, e.g. "Fop" [Ikemura, 1981], "CAI" [Sharp and Li, 1987], "E-CAI" [Puigbò *et al*., 2008], "CBI" [Bennetzen and Hall, 1982], "Nc" [Wright, 1990], "G+C content of the third codon position" [Sueoka and Kawanishi *et al*., 2000]. Several softwares for calculating these indexes are freely available on the internet (e.g. CodonW, Correspondence Analysis of Codon Usage, http://codonw.sourceforge.net/; JCat, Java Codon Adaptation Tool, http://www.jcat.de/Introduction.jsp; INCA, http://bioinfo.hr/research/inca/).

Codon bias is usually related to the genome at the level of genome sequence and noone has so far wondered if the proportion of used codons could vary during the expression of a whole transcriptome. To determine the actual pool of codons borne by all the mRNAs in the cell, the codon bias of each mRNA could be multiplied by the relative estimated number of molecules of that mRNA in the transcriptome. This could be done using once again the publicly available databases collecting the mRNA sequences and the expression experiments data, such as microarray data, in order to search for relationships between codon usage at the genome and transcriptome levels.

1.5 ASD implicated genes

Autism spectrum disorder (ASD) is a complex neurodevelopmental disorder; it is associated with impairments in social interaction and communications and with repetitive and stereotyped patterns of behaviour, interests and activities and males are affected four times more than female [Klauck, 2006]. Reviewing seven twin studies emerged that ASD has a substantial genetic component, with median values for concordance rates of 76-88% in monozygotic twin in contrast to 0-31% in dizygotic twins and an estimated heritability of 60-90% [Ronald and Hoekstra, 2011]. Extensive efforts went into identifying specific genetic causes and hundreds of ASD susceptibility loci, candidate gene mutations and chromosomal abnormalities have been studied [Betancur 2011]. An important study recently identified *MSNP1AS*, a long non coding RNA which maps within a GWAS (genome-wide association study) significant genetic marker for increased ASD risk [Kerin *et al*., 2012; Wang *et al*., 2009]. *MSNP1AS* is encoded by the antisense strand of a *MSN* pseudogene in human chromosome 5 (Fig. 1.3); *MSN*, the moesin gene, is located on the X chromosome and encodes a protein that plays a role in axon and dendrite development.

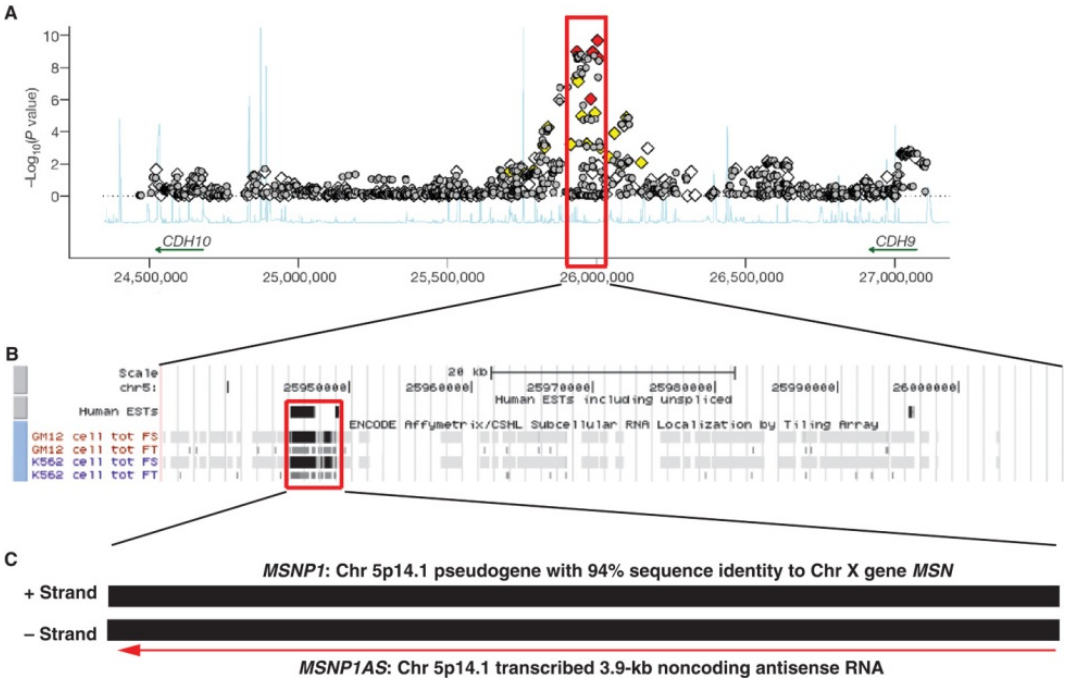


Figure 1.3: *MSNP1AS* maps within the chromosome 5p14.1 GWAS significant genetic marker for increased ASD risk. A. ASD-associated markers on chromosome 5p14.1 [adapted from Wang *et al.*, 2009]. B. A 4 kilo bases (kb) RNA transcribed from 5p14.1, as indicated by ESTs and RNA localisation (from the genome-wide ENCODE tiling aray project). C. The plus strand of the 4 kb 5p14.1 region is the *MSNP1*. The minus strand produces a non coding 3.9 kb RNA, designated *MSNP1AS* [Kerin *et al.*, 2012].

Kerin *et al*. found a strongly increased *MSNP1AS* expression in post mortem brain samples from individuals with ASD compared to those without; these higher levels correlate with the presence of the genetic marker for increased ASD risk. Furthermore the authors showed that *MSN* expression is also increased in post mortem brain samples from ASD affected individuals compared to controls and that *MSNP1AS* can bind *MSN* transcript in human neuronal cell line. Whether *MSNP1AS* higher expression can down-regulate MSN protein has not been confirmed yet, because MSN level in brain samples from ASD affected individuals does not change compared to controls [Kerin et al., 2012]. An alternative explanation about how *MSNP1AS* could contribute to disease could be found according to the mechanism of competing endogenous RNAs (ceRNAs) [Salmena *et al*., 2011]. Recent publications have shown that endogenous transcripts sharing MREs for the same microRNA can influence the expression level of each other through competitive microRNA binding: the decreased expression of one targeted transcript increases the concentration of free microRNAs that can bind the other targeted transcript and consequentely suppresses its expression; *vice versa* the overexpression of one targeted transcript leads to a decrease of available microRNAs that can bind the other targeted transcript, thereby increasing its expression [Marques *et al*., 2011]. The ceRNA network contain both coding and non conding transcripts, indeed many of protein coding genes are densely covered in MREs [Friedman *et al*., 2009] and microRNAs also regulate lncRNAs. A network involving a long non coding RNA in muscle differentiation by functioning as a ceRNA has already been described [Cesana *et al*., 2011]. As *MSNP1AS* has four MREs for a microRNA (miR-ASD, unpublished data), the overexpression of *MSNP1AS* observed in ASD affected individuals and with the risk genotype could reduce the concentration of miR-ASD, increasing the expression of other targeted transcripts that could be ASD implicated transcripts. This could be tested identifying MREs in ASD implicated genes and confirming that *MSNP1AS* and ASD implicated genes are co-expressed in the same network: a higher expression of *MSNP1AS* should lead to a higher expression of ASD implicated genes.

Chapter 2

Aim of the thesis

2.1 Systematic analysis of the human mRNA 5′ coding sequences

The aim of the first section of this work was to perform a systematic identification of coding regions at the 5′ end of all human known mRNAs. The identification of a more complete mRNA 5′ end could reveal an additional upstream AUG (in-frame with the previously determined one) thus extending the predicted amino terminus sequence of the product and avoiding subsequent relevant errors in the experimental study of the relative cDNA. The 5'_ORF_Extender software parses and makes calculations on RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/) and EST database sequences. This is done following the import of BLAT genome alignment data for human mRNAs and ESTs, in order to determine a list of genes with an incompletely described mRNA 5′ coding sequence. Due to the much larger size and complexity of RNA and EST sequence databases as well as the sequence analysis results file, the algorithm, previously described for *D. rerio* [Frabetti *et al*., 2007], has been completely revised and improved for *H. sapiens* analysis.

As a proof-of-concept, the EST-based models has been experimentally confirmed by *in vitro* cloning and sequencing of RNA 5′ coding region sequence extension for *GNB2L1*, *QARS* and *TDP2* human genes.

2.2 Relationship between codon bias and expressed RNA codons

The second part of this work was aimed to study the correlation between the codon bias (phenomenon in which distinct synonymous codons are used with different frequencies) and the number of codons present across all the expressed mRNAs, here called "codonome". Here we define the "codonome value" as the total number of codons (n) present across all the transcriptome mRNAs each expressed at a certain level (x) in a given biological condition (cv = $\sum(n \times x)$ for the mRNAs pool). The innovative "CODONOME" software has been developed to calculate the frequency of each codon in any reference (RefSeq) mRNA sequence and, following integration with a profile of gene expression values, to estimate the actual frequency of each codon in the mRNA pool derived from a specific tissue of a given organism. In addition, to investigate a possible cell adaptation aimed to optimize the translation process, these frequencies have been grouped by encoded amino acid, each being related to its specific aminoacyl-tRNA synthetase (aaRS), to determine whether some relationships exist between codon usage and aaRS mRNA expression level, a still unexplored field. Gene expression values for a certain condition were obtained from independent transcriptome datasets available in the Gene Expression Omnibus (GEO) database [Barret and Edgar, 2006; Barret *et al*., 2009] following intra- and inter-sample normalization using TRAM software [Lenzi *et al*., 2011].

A systematic analysis was performed varying the tissue examined within human species, testing a normal tissue, a pathological condition with a general disturbance of gene expression, i.e. the aneuploid blast from Down Syndrome (DS)-related acute megakaryoblastic leukemia (AMKL), and an extremely differentiated tissue with a remarkable expression preponderance of a very small number of proteins (human circulating blood erythrocytes samples) and investigating a pool of representative species from bacteria to humans. Then the codonome values was also determined in *Danio rerio*), (*Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Escherichia coli*) in order to search for general laws governing the structure of the codonome.

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

value (codonome bias) grouped by aaRS vs. the aaRS expression values.

2.3 Analysis of ASD implicated genes expression

The aim of the third part of this work was to study the role of *MSNP1AS*, a long non coding RNA, in the regulation of ASD implicated genes. As *MSNP1AS* has four MREs for miR-ASD, it could post-transcriptional regulate ASD implicated genes by competing for binding of miR-ASD. A computational approach has been adopted in order to first test the hypothesis as further experiments will be necessary to confirm the mechanism. A non redundant list of ASD implicated genes has been identified from published data available so far in the literature and their number of MREs has been predicted. Brain transcriptome data from healthy adults and ASD patients available in the GEO database have been analysed calculating the pairwise correlation of expression values between all the possible gene pairs in two groups (depending on the presence or the absence of MREs for the same microRNA), in order to study the co-expression of the ASD implicated genes by comparison with the remaining genes not implicated with ASD. Then the expression values of ASD implicated genes has then been compared in order to study their relationship with the increased ASD risk genotype.

Chapter 3

Materials and Method

3.1 Systematic analysis of the human mRNA 5′ coding sequences

3.1.1 Database construction

The 5'_ORF_Extender software parses and makes calculations on RefSeq http://www.ncbi.nlm.nih.gov/RefSeq/ and EST database sequences. This is done following the import of BLAT (BLAST-like alignment tool) genome alignment data for human mRNAs and ESTs, in order to determine a list of genes with an incompletely described mRNA 5′ coding sequence. The software has been developed using the FileMaker Pro 10 Advanced (FileMaker, Santa Clara, CA) database management system for both Windows and Macintosh operating systems. It is freely available as a stand-alone software (2.0 version) including the FileMaker runtime and a step-by-step user tutorial at http://apollo11.isto.unibo.it/software/.

Due to the very large size and high complexity of the human genome and of human EST database, together with the unavailability of a systematic assignment of mRNA and EST sequences to a defined genomic *locus* (in the form of an official gene symbol) in the UCSC data, an automated method of quality control of results has introduced. This *ex-ante* control verifies if each investigated EST has been assigned by UniGene http://www.ncbi.nlm.nih.gov/unigene system to the same locus as the mRNA sequence for which the EST is a possible candidate for 5′ end extension. This has been made possible thanks to the availability of a UniGene parser (the "UniGene Tabulator") able to produce a structured table including all UniGene updated text information [Lenzi *et al.*, 2006]. This table is imported into the "UniGene_ID" table of the 5'_ORF_Extender software as a first step, allowing

analysis to be limited to the mRNAs and corresponding ESTs that are mapped to the same defined *locus*.

Then, the human RefSeq flat file (version October 18, 2011) was downloaded from the UCSC (University of California, Santa Cruz) Genome Bioinformatics web site (http://genome.ucsc.edu/ - "Tables" section). The text file was imported into the "RefSeq_mRNA" table of the 5'_ORF_Extender software (following the software user guide) in order to obtain a local RefSeq database with all the human known reference mRNA sequences ("NM_" prefix, thus excluding RefSeq entries not supported by experimental evidence, such as "XM_" models). It is possible to only select and further analyse mRNA entries without an in-frame stop codon upstream of the described initiation codon, which are thus candidates for a possible extension at 5′ end: the presence of such a stop codon would indicate that the 5′ UTR sequence cannot be part of a longer continuous CDS. This also implies that a database of all RefSeq mRNAs that are *bona fide* complete at the 5′ end of their CDS is therefore generated.

The genome alignment data for human ESTs, assigned by UniGene to the same *locus* of the mRNAs candidates for a possible extension at 5′ end, were then downloaded from the UCSC site (version October 19, 2011) and imported into the "EST_ Data" software table. Each human mRNA (without an in-frame stop codon upstream of the described initiation codon) was then compared with all the human EST assigned to the same locus by analysing the coordinates of the pre-computed genome alignments for mRNAs and ESTs obtained by UCSC. Only those EST sequence entries presenting additional nucleotides upstream of the known 5′ mRNA end and therefore candidate to potentially extend the mRNA CDS at its 5′, were downloaded and imported into the "EST_Seq" software table.

The whole analysis for *H. sapiens*, including UniGene data, mRNA and EST data and sequences import and processing, required about 5 days for completion.

3.1.2 Computational analysis

The 5'_ORF_Extender analysis script performs the following steps (Fig. 3.1): extraction of the EST sequence stretch upstream of the matched RefSeq mRNA first base when BLAT alignment shows a 5′ extension of the EST compared with the known RefSeq sequence (following removal of introns from both EST and mRNAs genome-aligned sequences); a search in this EST stretch for the most upstream existent ATG (corresponding to AUG in RNA) in-frame with the described one in the RefSeq mRNA sequence entry; calculation of the new putative extended coding

region by merging the EST extended stretch starting from the new ATG with the previously known 5′ UTR of the RefSeq mRNA sequence; confirmation of the coding potential of this new extended sequence by excluding the presence of any in-frame stop codon within it.

It can also be estimated whether or not the determined extended CDS is complete, by searching for any in-frame stop codon that might occur in the transcript upstream of the newly determined start codon.

As a final result, the software provides a list of genes whose mRNA possesses an extended 5′ CDS on the basis of EST comparison.

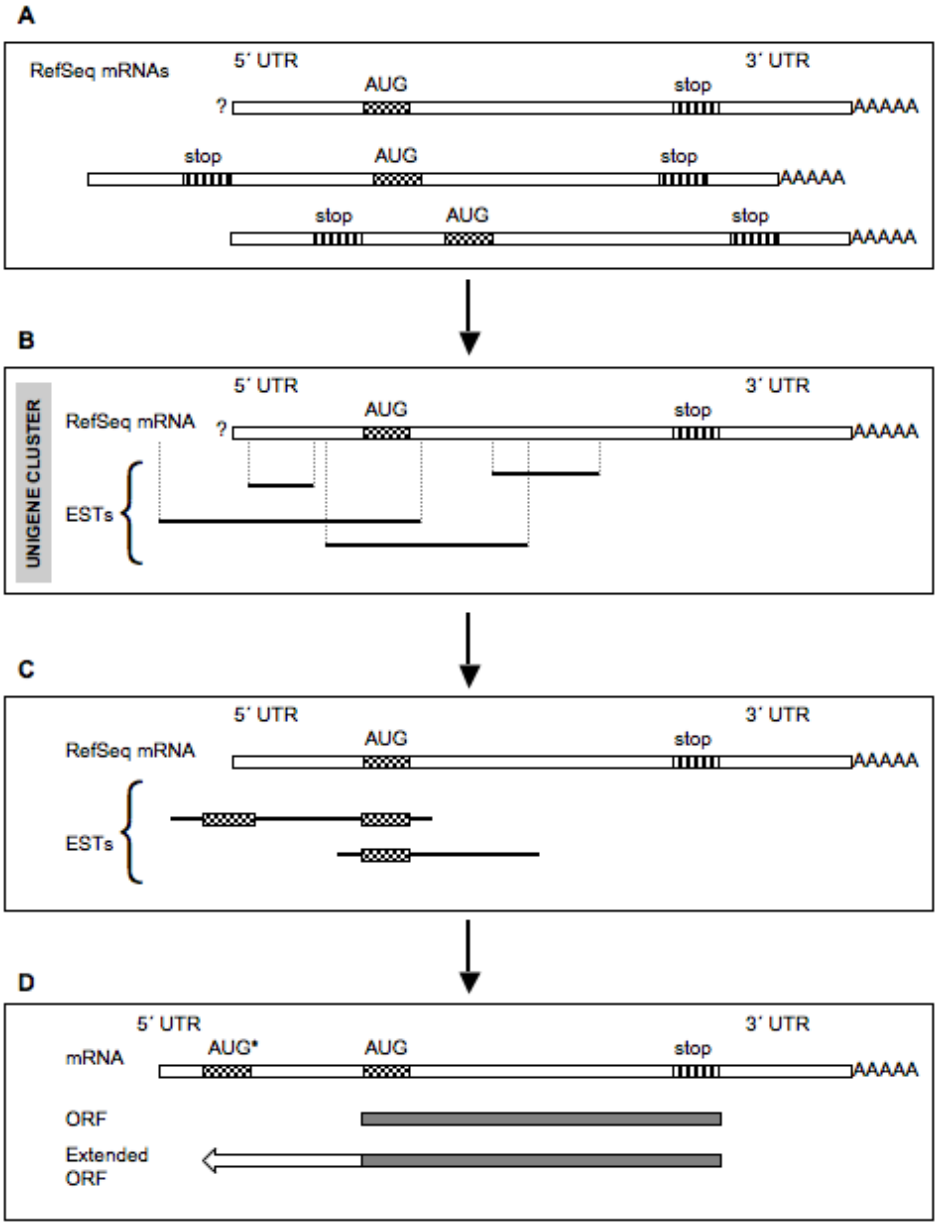


Figure 3.1: Pipeline of the 5'_ORF_Extender software version 2.0 approach. Sequence comparisons exploit BLAT-pre-computed UCSC genomic coordinates of the RefSeq and EST sequences. Detailed explanation in the text. A. Identification of RefSeq mRNA sequences without a known in-frame stop codon upstream of the described initiation codon (and thus candidates for further extension of their CDS at 5′). B. The parsed and embedded UniGene database allows the determination of those EST sequences that cluster with each RefSeq mRNA sequence and that are possible candidates for extending their 5′ coding region. C. Identification of EST sequences with an upstream in-frame AUG codon and absence of any stop codon between the previously and the newly determined AUG codons. D. Calculation of the new extended open reading frame (new AUG codon indicated with an *)

3.1.3 In vitro cloning and sequencing of the mRNA 5′ region

The sequence analysis predictions was confirmed of three example genes of the 5'_ORF_Extender results list. We utilised a reverse transcription-polymerase chain reaction (RT-PCR) approach, based on the amplification of a stretch extended from the new putatively defined 5′ UTR to at least as far as the known exon 2, in order to prove that the amplified cDNA derived from mRNA. The human RNA sources were: skeletal muscle, small intestine, ovary, brain and bone marrow total RNA purchased from Clontech (Palo Alto, CA).

Standard reverse transcription conditions were: 2 μg of total RNA, Moloney murine leukemia virus reverse-transcriptase (Promega, Madison, WI; used with the companion buffer) 400 U, oligo dT-15 2.5 μM, random nonamers 2 μM, dNTPs 500 μM each. An RNA denaturation step was performed at 95°C for 5 minutes before the addition of primers and enzyme. RT reaction was performed in a final volume of 50 μL for 60 minutes at 42°C.

PCR experiments were performed in a 25 μL final volume, containing 2 μL of cDNA, 1 U Taq polymerase (TaKaRa, Shiga, Japan) with companion reagents (0.2 mM each dNTPs, 2 mM MgCl2, 1$\times$ PCR buffer) and 0.2-0.3 μM of each primer. An initial denaturation step of 2 minutes at 94°C, followed by 40-48 cycles of 30 seconds at 94°C, 30 seconds at the indicated annealing temperature (Ta, 61-64°C), 30 seconds at 72°C, and a final extension of 7 minutes at 72°C. In one case (*TDP2* cDNA), an additional step of reamplification (20 cycles) was conducted as above, starting from 1 μL of sample obtained after the excision of the expected band from agarose gel and its subsequent syringe-squeezing [Li and Ownby, 1993].

Primers pairs were designed with "Amplify3" software [Engels, 1993] following standard criteria and are listed in Table 3.1.

All RT-PCR products obtained were gel analysed following a standard method [Davis *et al.*, 1994], purified using a GenElute kit (Sigma-Aldrich, St. Louis, MO), and then subjected to automated sequence analysis of both DNA strands for each fragment, using the same primers utilised in the respective PCR reactions. BigDye chain-terminator method (Applied Biosystems, Carlsbad, CA) was used with an automated Applied Biosystems ABI 3730 DNA automated sequencer.

Gene Symbol	**(Gene full name)**	**RefSeq mRNA GenBank Accession No.**	**Primer pairs sequence ($5' \rightarrow 3'$) (F): Forward (R): Reverse**	**RT-PCR product size (tissue sources)**	**GenBank Human EST[a]**	**Product length new/reference**	**No. of new amino acids (% of reference) length)**
GNB2L1	Guanine nucleotide binding protein (G protein), beta polypeptide 2 like 1	NM_006098	ggaattccatagttggtctc (F) cttgaatgtgcttgtttcagag (R)	470 bp (Ovary, Brain)	BU172346.1 ES313379.1 BP312588 BP244479	395/317	78 (+25%)
QARS	Glutaminyl-tRNA synthetase	NM_005051	ggatagacgaccttggagcg (F) gactccgcacatactcaagg (R)	442 bp (Skeletal Muscle, Small Intestine)	BI461626.1 BI829834.1 BI463065.1 BM560535	793/775	18 (+2%)
TDP2	Tyrosyl-DNA phosphodiesterase 2	NM_016614	cgcagctgcaccagttttccgag (F) ctcagagatggtttcaggtcg (R)	383 bp (Brain, Bone Marrow)	BM554324.1 BG719977.1 BP270589 DA431403	392/362	30 (+8%)

Table 3.1: Experimentally confirmed extended cDNA $5'$ coding region. [a]Four example EST sequences supporting an extended coding sequence at $5'$ region of the corresponding RefSeq mRNA, resulted from "5'_ORF_Extender" software analysis. *GNB2L1*, *QARS* and *TDP2* extensions were supported by a total of 5, 24 and 12 consistent ESTs, respectively.

3.1.4 Sequence analysis

In order to test whether the newly determined CDS at 5′ was conserved in different species, TBLASTN searches were performed using standard parameters, except the filter for low complexity regions was unchecked. Alignment of the protein products was made by ClustalW software (version 2.1 at:
http://www.ebi.ac.uk/Tools/msa/clustalw2/).

In order to identify novel domains which were not present in the described gene products, the predicted extended amino acid sequences for the three example genes were searched for in domain databases such as the Simple Modular Architecture Research Tool (SMART, http://www.smart.embl-heidelberg.de/) and the Conserved Domains Database (CDD, http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml).

3.2 Relationship between codon bias and expressed RNA codons

3.2.1 Database construction

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

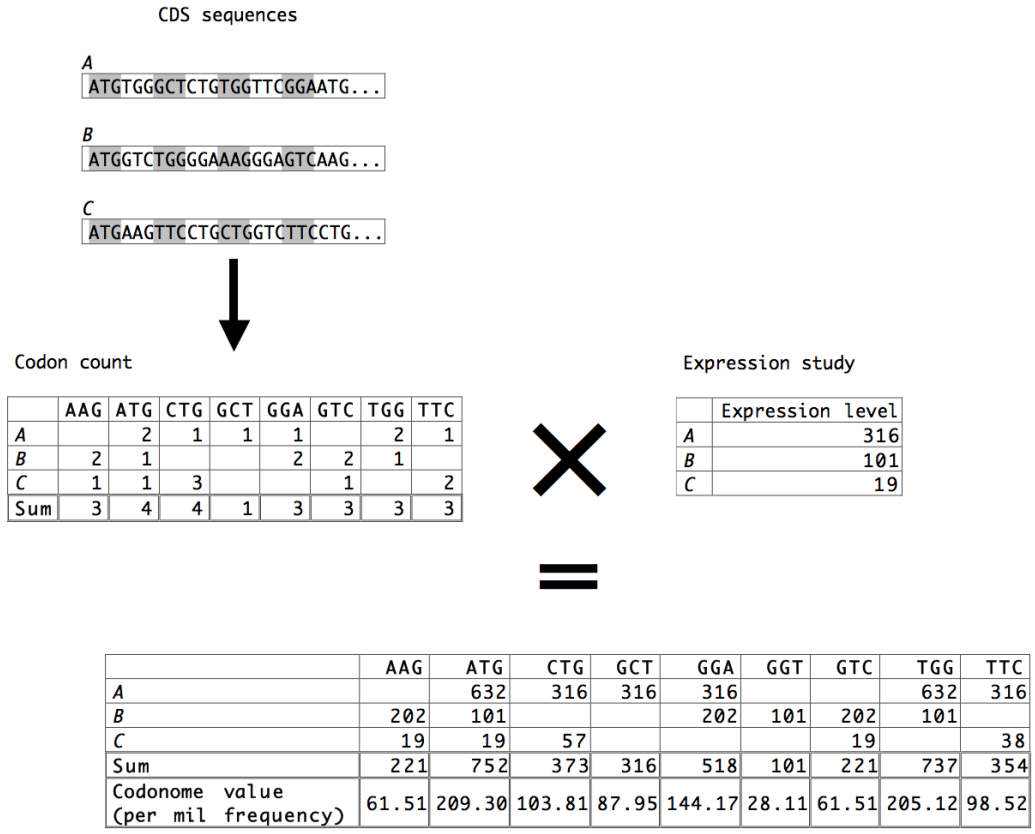
The transcriptomes from the following species were investigated in order to obtain data from higher- and lower-vertebrates as well as from invertebrates, unicellular eukaryotes, and prokaryotes: *H. sapiens*, *D. rerio*, *C. elegans*, *S. cerevisiae* and *E. coli*. First, the RefSeq mRNA flat files of the desired species were downloaded from the NCBI ftp site (*H. sapiens* version May 7, 2010; *D. rerio* version June 16, 2010; *C. elegans* and *S. cerevisiae* versions January 18, 2011; *E. coli* version March 1, 2011). Each text file was edited (in order to create a tabulator key separated file suitable for a File Maker table) and imported into the "RefSeq_Parser" table of the "CODONOME" to obtain a specific local RefSeq database.

Following the execution of the "CODONOME" command, all but the "NM_" type entries were deleted, thus excluding non-reviewed, predicted mRNA entries (*H. sapiens*: 29,538 NM entries; *D. rerio*: 14,174 NM entries; *C. elegans*: 23,894

NM entries; *S. cerevisiae*: 5,882 NM entries; *E. coli*: 4,319 NM entries). The same script also counted each codon for each mRNA individually, then summed these values to obtain the total number of each codon for the whole mRNAs pool (Fig. 3.2) and then calculated their per mil frequencies.

The expression data files for each species were downloaded from the GEO web site. The Table 3.2 and the Table 3.3 list the investigated tissues and organisms and the numbers of considered samples and experiment series. For human brain, was performed a search with the word "brain" in GEO datasets, and arbitrarily selected 24 samples from 7 different series in order to integrate representation from different platforms (Affymetrix microarrays types), different authors, and different investigated subjects, thus obtaining an integrated summarised gene expression profile that best represents the general biological transcriptome map for that tissue following both universal assignment of each probe to a specific locus via UniGene data parsing [Lenzi *et al*., 2006] and intra- and inter-sample advanced normalisation [Lenzi *et al*., 2011]. A similar process was performed in order to obtain gene expression profiles for other human tissues, including leukemic cells, as well as for other species. For *D. rerio* and *C. elegans*, for which fewer studies are available, the platform used in most experiments were chose: GPL1319 and GPL200, respectively.

Each expression data file was processed using TRAM software [Lenzi *et al*., 2011]. "Set up" and "Importing the expression data files" software sections were performed according to the software user guide. Then gene symbols with the corresponding normalised expression values were exported in a text file for each investigated species and imported it into the appropriate "Codonome" table of the "CODONOME" database.



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

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

Figure 3.2: Pipeline of the "CODONOME" software. For each RefSeq ("NM_" type) entry considered, the occurrence of each codon was counted. Then the count sum of each codon for the whole gene set was calculated (the per mil frequencies of each codon sum in relation to the sum of all codons for the whole gene set gives the codon bias). The codon count for each gene was then multiplied by the normalised expression value of that gene. Finally, the total number of each codon for the whole gene set was summed. The per mil frequencies of each codon sum in relation to the sum of all codons for the whole gene set give the transcriptome codonome bias (example simulation for a hypothetical gene set composed of three genes "*A*", "*B*" and "*C*" assuming the existence of nine codons).

Study ID	Sample ID	Sample type	Platform	Microarray	Spots	Ref.
Pool "A" - healthy adults (n = 24)						
A1...A8 (n = 8)	GSM123271...78	Human post mortem brain tissue	GPL96	U133A	22,283	[Lockstoe *et al*., 2007]
A9 (n = 1)	GSM44690	Normal brain	GPL96	U133A	22,283	[Ge *et al*., 2005]
A10-A11 (n = 2)	GSM12688, GSM12708	Normal brain	GPL8300	U95 Version 2	12,625	[Yanai *et al*., 2005]
A12-A13 (n = 2)	GSM12689, GSM12709	Normal brain	GPL92	U95B	12,620	"
A14-A15 (n = 2)	GSM12690, GSM12710	Normal brain	GPL93	U95C	12,646	"
A16-A17 (n = 2)	GSM12691, GSM12711	Normal brain	GPL94	U95D	12,644	"
A18-A19 (n = 2)	GSM12692, GSM12712	Normal brain	GPL95	U95E	12,639	"
A20 (n = 1)	GSM52556	Normal brain	GPL96	U133A	22,283	[Detwiller *et al*., 2005; Yoon *et al*., 2006]
A21-A22 (n = 2)	GSM76949, GSM76999	Whole brain	GPL570	U133 Plus 2.0	54,675	[Nguyen and Disteche, 2006]
A23 (n = 1)	GSM136140	Human control brain tissue	GPL96	U133A	22,283	[Padden *et al*., 2007]
A24 (n = 1)	GSM112030	Brain	GPL570	U133 Plus 2.0	54,675	[Auer *et al*., 2009]
Pool "B" - healthy adult (n = 41)						
B1...B14 (n = 14)	GSM143572...85	Normal human adult red blood cells	GPL96	U133A	22,283	[Goh *et al*., 2007]
B15...B28 (n = 13)	GSM143671...76, GSM143703, GSM143706...11	Normal human adult red blood cells	GPL97	U133B	22,645	"
B29...B35 (n = 7)	GSM83897, GSM85205...10	Erythrocytes	GPL201	HG-Focus	8,793	[Kabanova *et al*., 2009]
B36...B41 (n = 6)	GSM440234...39	Reticulocytes from adult periperal blood	GPL570	U133 Plus 2.0	54,675	[Noh *et al*., 2009]
Pool "C" - DS-AMKL children (n = 31)						
C1...C3 (n = 3)	GSM491372...4	BM Sorted leukemic blasts	GPL570	U133 Plus 2.0	54,675	[Klusmann *et al*., 2010]
C4...C25 (n = 22)	GSM94245, GSM94272...92	BM or PB	GPL96	U133A	22,283	[Bourquin *et al*., 2006]
C26...C31 (n = 6)	GSM417985...90	BM or PB Sorted leukemic blasts	GPL570	U133 Plus 2.0	54,675	[Klusmann *et al*., 2010]

Table 3.2: Samples selected: *Homo sapiens* (pool "A", "B" and "C"). All Sample IDs and Platform IDs are related to GEO database. Sample type: BM, bone marrow; PB, peripheral blood. Microarray: U133A: Affymetrix Human Genome U133A Array; U95 Version 2: Affymetrix Human Genome U95 Version 2 Array; U95B: Affymetrix Human Genome U95B Array; U95C: Affymetrix Human Genome U95C Array; U95D: Affymetrix Human Genome U95D Array; U95E: Affymetrix Human Genome U95E Array; U133 Plus 2.0: Affymetrix Human Genome U133 Plus 2.0 Array; U133B: Affymetrix Human Genome U133B Array; HG-Focus: Affymetrix Human HG-Focus Target Array.

Study ID	Sample ID	Sample type	Platform	Microarray	Spots	Ref.
Pool "D" - Wildtype adults (n=23)						
D1 (n=1)	GSM74260	Brain	GPL1319	[Zebrafish] Affymetrix Zebrafish Genome Affymetrix Array	15,617	[Cameron *et al*., 2005]
D2...D4 (n=3)	GSM305891...93	"	"	"	"	[Lefebvre *et al*., 2009]
D5...D8 (n=4)	GSM280425...28	"	"	"	"	[Drew *et al*., 2008]
D9...D23 (n=15)	GSM337575...77, GSM337591...93, GSM337604...06, GSM337618...20, GSM337631...33	"	"	"	"	[Toyama *et al*., 2009]
Pool "E" - Strain N2, wildtype young adults (n=19)						
E1...E3 (n=3)	GSM214716, GSM214725, GSM214727	Whole worm	GPL200	[Celegans] Affymetrix C. elegans Genome Array	22,625	[Asikainen *et al*., 2007]
E4...E7 (n=4)	GSM419959...62	"	"	"	"	[Krajacic *et al*., 2009]
E8...E12 (n=5)	GSM250116...20	"	"	"	"	[Falk *et al*., 2008]
		"	"	"	"	[Peng *et al*., 2008]
E13...E16 (n=4)	GSM40308...11	"	"	"	"	[Falk *et al*., 2008]
		"	"	"	"	[Peng *et al*., 2008]
E17...E19 (n=3)	GSM536251...53	"	"	"	"	[Falk *et al*., 2008]
		"	"	"	"	[Peng *et al*., 2008]
Pool "F" - Strain S288C, wildtype adults (n=19)						
F1-F2 (n=2)	GSM248646-45	/	GPL5092	Bauer Center for Genomics Research Saccharomyces cerevisiae 70mer array, Hartl Lab	7,744	[Brown *et al*., 2008]
F3 (n=1)	GSM34635	/	GPL90	[YG_S98] Affymetrix Yeast Genome S98 Array	9,335	[Simons *et al*., 2006]
F4...F19 (n=16)	GSM67593, GSM67610-19, GSM67622-27, GSM67596-99, GSM67613-16, GSM67630-33, GSM67636-39, GSM67602, GSM67605-08,	/	"	"	"	[Guan *et al*., 2006]
Pool "G" - strain K-12, substr. MG1655, exponential growth, aerobic, wildtype (n=8)						
G1...G3 (n=3)	GSM247608, GSM247612, GSM247613	/	GPL199	[Ecoli_ASv2] Affymetrix E. coli Antisense Genome Array	7,312	[Dong *et al*., 2008]
G4-G5 (n=2)	GSM469137, GSM469138	/	GPL3154	[E_coli_2] Affymetrix E. coli Genome 2.0 Array	10,208	[Moon and Gottesman 2009]
G6...G8 (n=3)	GSM510322...24	/	"	"	"	[Holm *et al*., 2010]

Table 3.3: Samples selected: *Danio rerio* (pool "D"), *Caenorhabditis elegans* (pool "E"), *Saccharomyces cerevisiae* (pool "F"), *Escherichia coli* (pool "G"). All Samples IDs and Platforms IDs are related to GEO database. "/": not specified.

3.2.2 Computational analysis

For each "NM_" mRNA-type entry considered the following step were performed: was counted how many times each codon occurred; the count sum of each codon for the whole gene set and the per mil frequency of each codon sum in relation to the sum of all codon for the whole genome gene set (codon bias) were calculated; the codon count for each gene was then multiplied by the normalised expression value of that gene; the count of each codon for the whole gene set and the per mil frequency of each codon sum in relation to the sum of all codon for the whole genome gene set (codonome bias) were summed.

With these values, it is possible to search for relationships between codon usage at genome and at transcriptome level.

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

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

Lastly, a list of the twenty aaRS was created with the respective recognised codons for *H. sapiens*, *D. rerio*, *C. elegans*, *S. cerevisiae* and *E. coli*. Codon and codonome frequencies were then grouped by aaRS with the relative expression values (using the same expression data file as before).

3.2.3 Statistical analysis

Actual and simulated analyses results were exported in text files and submitted to statistical analysis using statistical software for Mac OS X ("JMP software" 5.1.2, SAS Institute Inc., Cary, USA). The correlation between paired variables was analysed through linear regression, setting the density ellipse at 0.50. In the statistical analysis results "r" is the correlation coefficient and "p" represents the p

value. The correlation was studied among the following parameters: a) codon bias, b) codonome bias, c) codon bias grouped by aaRS, d) codonome bias grouped by aaRS and e) the aaRS expression values ("a", "b", "c" and "d" are expressed as per mil frequencies).

3.3 Analysis of ASD implicated genes expression

3.3.1 ASD implicated genes

A non redundant list of ASD implicated genes was compiled from available published data [O'Roak *et al*., 2012; Neale *et al*., 2012; Sanders *et al*., 2012; Betancur, 2011; Anney *et al*., 2010; http://www.human-phenotype-ontology.org]. The software TargetScan (http://www.targetscan.org/) was used in order to predict if these ASD implicated genes have MREs for miR-ASD, in order to understand whether *MSNP1AS* (that has four MREs for the same microRNA) can post-transcriptional regulates ASD implicated genes by competing for binding of miR-ASD.

The expression data files for healty and ASD affected adults were downloaded from the GEO web site, selecting only samples with known ASD implicated risk genotype and only prefrontal or temporal cortex (Table 3.4), where expression changes associated with ASD has been found to be more pronounced [Voineagu *et al*., 2011]. An expression dataset from healthy fetal, child and adult prefrontal cortex samples was downloaded (from the GEO web site as well) in order to understand if there are changes during the normal brain development (Table 3.5).

Study ID	Sample ID	Sample type	Platform	Spots	Ref.
Pool "A" - autism spectrum disorder (n = 27)					
A1...A6 (n = 6)	GSM706412...17	Human post mortem prefrontal cortex	GPL6883	24,526	[Voineagu *et al*., 2011]
A7...A15 (n = 9)	GSM706444...56	"	"	"	"
A16...A19 (n = 4)	GSM706444...47	Human post mortem temporal cortex	"	"	"
A20...A27 (n = 8)	GSM706449...56	"	"	"	"
Pool "B" - healthy adult (n = 14)					
B1...B4 (n = 4)	GSM706429...32	Human post mortem prefrontal cortex	GPL6883	24,526	[Voineagu *et al*., 2011]
B5...B8 (n = 4)	GSM706434 GSM706436 GSM706439 GSM706441	"	"	"	"
B9...B11 (n = 3)	GSM706458...60	Human post mortem temporal cortex	"	"	"
B12...B14 (n = 3)	GSM706462 GSM706464 GSM706467	"	"	"	"

Table 3.4: Samples selected with known ASD implicated risk genotype. All Sample IDs and Platform IDs are related to GEO database. Microarray: Illumina HumanRef-8 v3.0 Expression BeadChip.

Study ID	Sample ID	Sample type	Platform	Spots	Ref.
Pool "C" - healthy fetal samples (n = 38)					
C1...C38 (n = 38)	GSM749899 GSM749900...36	Human post mortem dorsolateral prefrontal cortex	GPL4611	49,152	[Colantuoni *et al*., 2011]
Pool "D" - healthy child samples 0-10 years (n = 33)					
D1...D33 (n = 33)	GSM749937...69	Human post mortem dorsolateral prefrontal cortex	GPL4611	24,526	[Colantuoni *et al*., 2011]
Pool "E" - healthy adult samples (n = 198)					
E1...E198 (n = 198)	GSM749970...99 GSM750000...167	Human post mortem dorsolateral Prefrontal cortex	GPL4611	24,526	[Colantuoni *et al*., 2011]

Table 3.5: Samples selected for human brain development. All Sample IDs and Platform IDs are related to GEO database. Microarray: Illumina Human 49K Oligo array (HEEBO-7 set).

3.3.2 Statistical analysis

The following values were calculated developing appropriate scripts with Python (http://www.python.org/) and compared using the software R [R Development Core Team, 2008].

Expression values of each sample from ASD affected and healthy adults [Voineagu *et al*., 2011] have been normalised using the median expression value of a list of expressed housekeeping genes [Eisenberg and Levanon, 2003].

Expression values from all the healty and ASD samples considered were divided in two groups according to the following criteria: ASD implicated genes without any MREs for miR-ASD; ASD implicated genes with at least one MRE for miR-ASD. Then two groups of genes not implicated with ASD were created in order to compare the two ASD implicated genes groups with the background: genes not implicated with ASD were randomly picked in the same number of ASD implicated genes respectively without and with MREs for 1000 permutations.

In each of these groups, the Pearson's pairwise correlation coefficient of expression values between all the possible gene pairs was calculated. Then the median value of the correlation coefficients was calculated for each group and for each of the 1000 permutations. The calculated median value has been compared for ASD implicated genes without any MREs for miR-ASD and with at least one MRE for miR-ASD (separately) using as background the median values of the correlation coefficient for each of the 1000 permutations. The p values have been calculated as how many times the median value of the correlation coefficients calculated for ASD implicated genes without MREs for miR-ASD (or with at least one MREs for miR-ASD) is bigger than the median values calculated for the 1000 permutations of genes not implicated with ASD. This analysis is useful to understand whether ASD implicated genes are significantly more co-expressed than expected by chance in the normal and in the pathological situation and during normal brain development.

Expression values for ASD implicated genes with and without MREs for miR-ASD only from ASD affected adults were then compared dividing them depending on the ASD implicated risk genpotype, in order to understand whether expression values of ASD implicated genes with or without MREs for miR-ASD differ in the genotype related with a higher risk and with a higher expression of *MSNP1AS*.

Chapter 4

Results

4.1 Systematic analysis of the human mRNA 5′ coding sequences

4.1.1 Database construction and computational analysis

The processing by 5'_ORF_Extender of 30,909 human RefSeq mRNA sequences assigned by UniGene to a defined locus (out of a total of 31,903) revealed the presence of an in-frame stop codon upstream of the known start codon in 20,775 cases. 10,134 sequences had a CDS which was putatively further extendable at their 5′ end. 159,378 UCSC EST-to-genome alignments, for the ESTs candidate to potentially extend the mRNA CDS at its 5′ in these 10,134 selected human mRNAs, were then processed to identify positive final results. Following calculations executed by the software, it was possible to obtain candidate extended coding regions at 5′ end from 2,505 ESTs (Table 4.1).

Summary of analysis	
Human loci analysed	18,665
Human Reference mRNAs (RefSeq) analysed	31,903
Human RefSeq mRNA sequences assigned by UniGene to a defined locus	30,909
mRNAs with CDS not extendable at 5´ end (in-frame stop codon located upstream of the known start codon)	20,775
mRNAs with CDS possibly further extendable at 5′ end	10,134
ESTs assigned to the same locus of the 10,134 mRNAs possibly further extendable at 5′ end	7,166,113
EST-to-genome alignments for the EST candidates to potentially extend the mRNA CDS at their 5′ end	159,378
Final set of results	
ESTs with putative CDS extension	2,505
mRNAs with putative extension of their known CDS at 5′ end	615
Loci with putative extension of their known CDS at 5′ end	477
Mean number of ESTs with extended sequence per mRNA	4.1
Mean length of extended 5′ CDS	178.5
Standard Deviation of the extended 5′ CDS length	134.8
Minimum length of extension	3
Maximum length of extension	1,014
mRNAs with CDS extension supported by more than one EST	298
mRNAs with CDS extension supported by more than one EST not derived from the same library	270
Loci with CDS extension supported by more than one EST	232
Loci with CDS extension supported by more than one EST not derived from the same library	213

Table 4.1: Summary of computational analysis. CDS, coding sequence. Length is given in nucleotides.

4.1.2 Summarisation of results

The final set of 2,505 ESTs corresponded to 477 distinct human loci (2.6% of all studied genes with a RefSeq sequence) (Table 4.1). The mean number of EST sequences that allowed the extension of one mRNA sequence was 4.1, with 298 different mRNAs extended by at least two distinct EST sequences. In particular, the ESTs extending 270 out of these 298 mRNAs were not derived from the same library. The mean size of the additional open reading frames (ORF) stretch was 178.5 bases, with a standard deviation of 134.8 bases (range: 3-1,014 bases) (4.1). An example of the "Result" table of the 5'_ORF_Extender software is shown if Fig. 4.1.

For 224 genes (46.96%) it can be estimated that the determined extended CDS is complete, due to the presence of an in-frame stop codon upstream of the newly determined start codon.

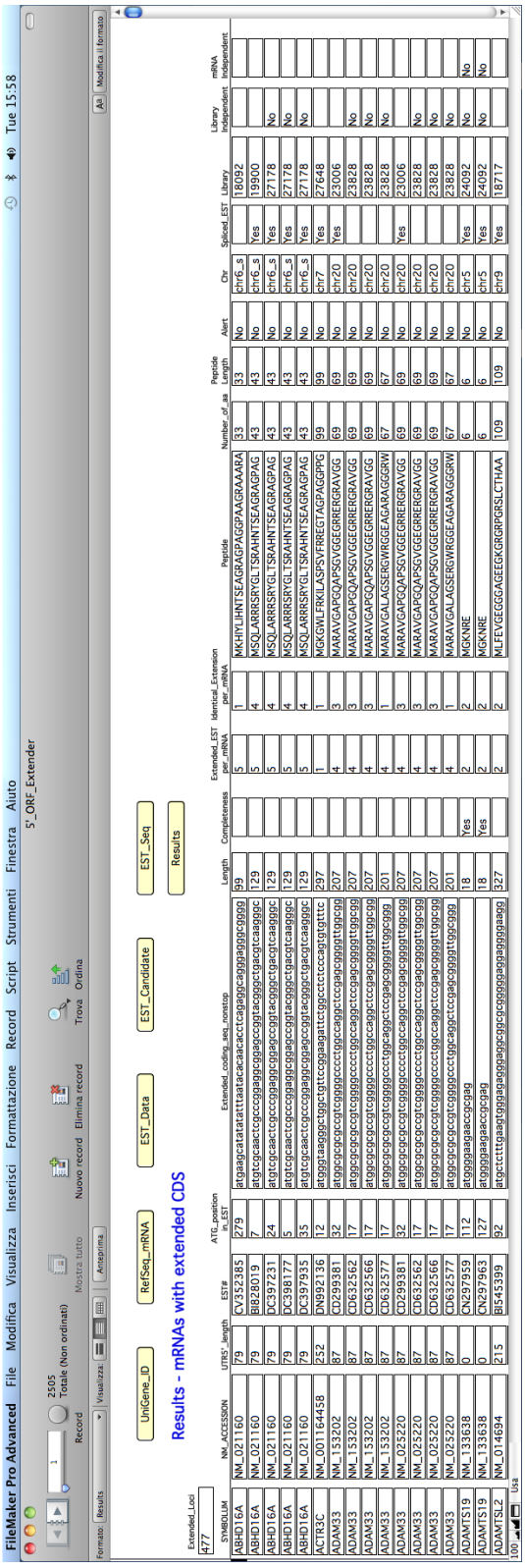


Figure 4.1: Example of the 5'_ORF_Extender software's "Results" table.

4.1.3 In vitro cloning and sequencing of the mRNA 5′ region

The predicted additional coding region was cloned for each of the three example genes: *GNB2L1*, *QARS* and *TDP2* (Table 3.1). The expected size bands corresponding to the amplified *GNB2L1*, *QARS* and *TDP2* 5′ regions are shown in Fig. 4.2 and the electropherogram obtained after *QARS* 5′ region sequencing is shown in Fig. 4.3.

The nucleotide sequences of the extended coding regions determined exactly between the 3′ end of the primer pairs for *GNB2L1*, *QARS* and *TDP2* cDNAs have been deposited in the GenBank database under accession nos. JN104586, JN104585 and JN104587, respectively.

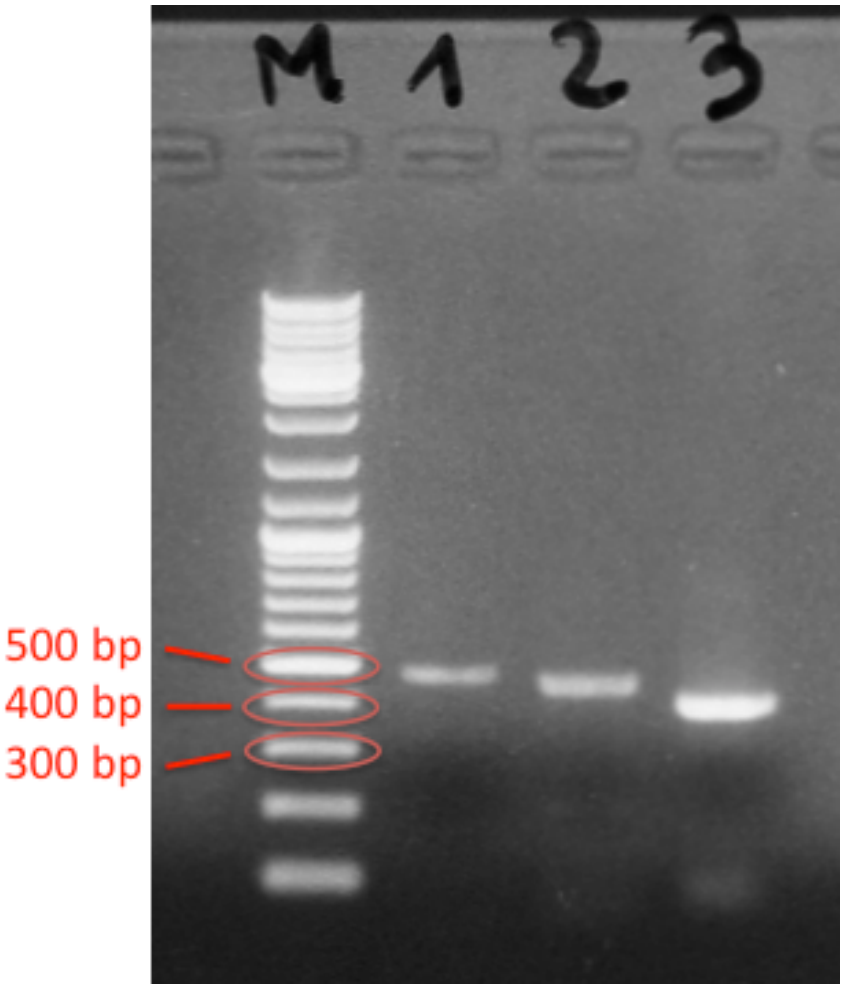


Figure 4.2: Expected band size corresponding to the amplified 5′ regions in a 1.5% agarose gel. Lane M: Marker GeneRuler Ladder. Lane 1: *GNB2L1*, 470 bp, brain. Lane 2: *QARS*, 442 bp, small intestine. Lane 3: *TDP2*, 383 bp, brain.

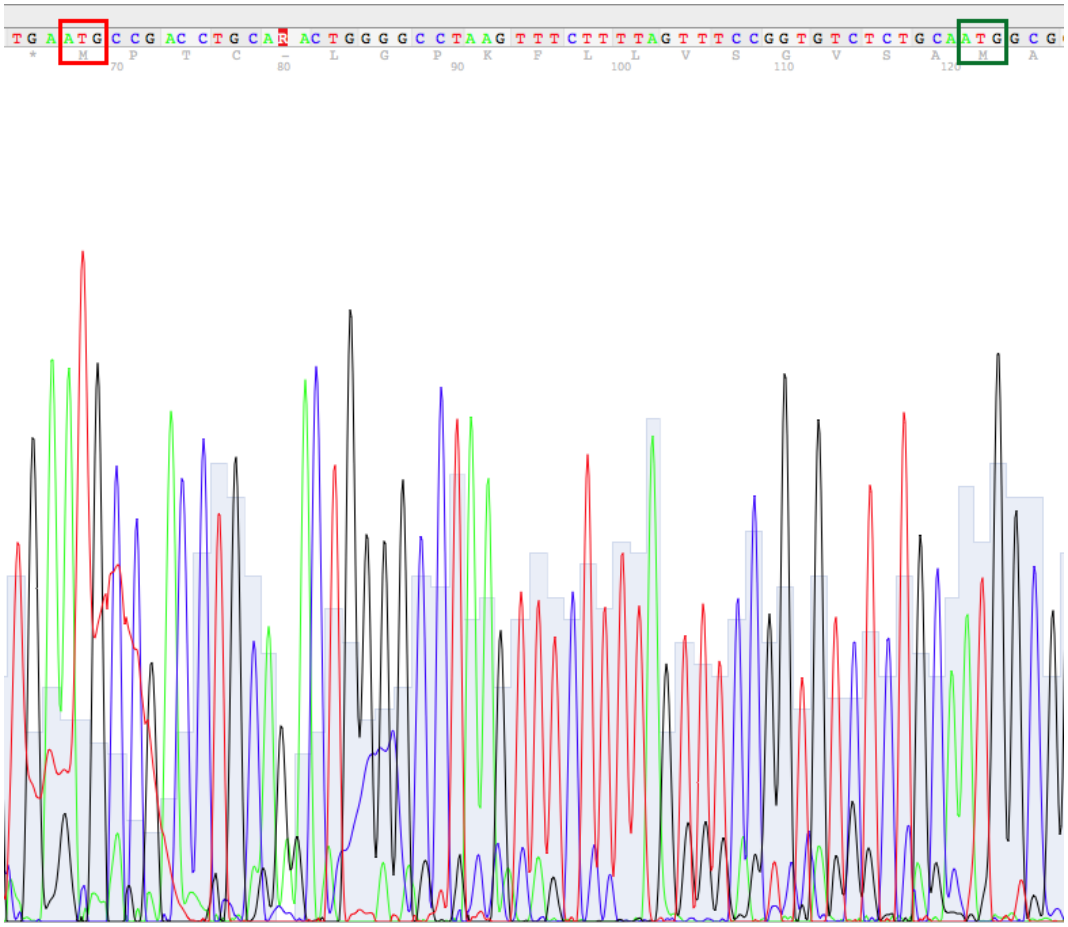


Figure 4.3: Electropherogram of *QARS* 5′ region sequencing. The green and the red rectangles underline the previously and newly determined AUG codons, respectively. Asterisk: in-frame stop codon upstream of the newly determined start codon.

4.1.4 Sequence analysis

The extended coding sequences for *GNB2L1*, *QARS* and *TDP2* were analysed using the TBLASTN program in order to compare them with known nucleotide sequences deposited in the NCBI databases. This confirmed that no human matching sequence had been previously deposited in the "mRNA" (molecular type) division of GenBank, except two sequences (#AK302867 and #AK298699) relating to *QARS* and *TDP2*, respectively. Although these sequences are not present in the GenBank EST division, they were generated in the context of the NEDO large-scale cDNA sequencing project [Yudate *et al*., 2001] and the relative entries were not tagged with the corresponding gene symbol as well as their predicted proteins (classified as "unnamed protein product"). They were not used by the genome browsers NCBI Map Viewer [Sayers *et al*., 2011] and University of California at Santa Cruz (UCSC) Genome Browser [Sanborn *et al*., 2011] to build mRNA models with the extended CDS. mRNA models including the extended CDS reported here for *QARS* (Ensembl Entry ENST00000420147) and *TDP2* (Ensembl Entry ENST00000545995), but not for *GNB2L1*, were available at the European Bioinformatics Institute (EBI) Ensembl genome browser [Flicek *et al*., 2011]. These CDSs were not however included in the entries containing coding sequences (Ensembl CCDS) available for the two genes, respectively, and the mRNA models were mainly based on mRNA sequences. These include the aforementioned "mRNA" sequences relating to *QARS* and *TDP2*, with limited support from available ESTs (2 ESTs out of the 24 identified by 5'_ORF_Extender in the case of *QARS* and 2 out of the 12 in the case of *TDP2*). In addition, as stated in the Ensembl genome annotation documentation, EST alignments are displayed on the website but are not usually used as supporting evidence in the gene-building process. The nucleotide and amino acid analysis data are summarised in the Table 3.1.

Sequence comparison also showed the presence of high conservation of the extended stretch with predicted proteins in non human primates, a finding consistent with the coding nature of these regions (Fig. 4.4).

The amino acid sequences predicted at the amino terminus of these three genes did not show new known functional domains through database searches.

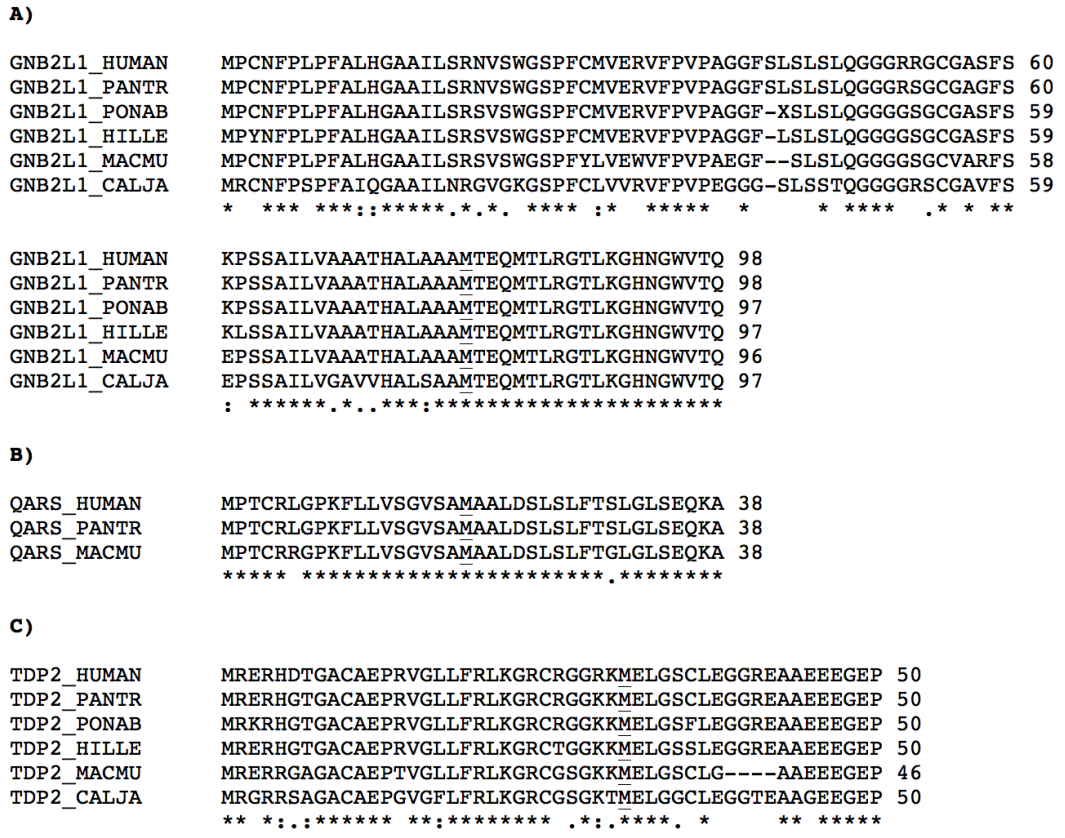


Figure 4.4: ClustalW alignment of GNB21L (A), QARS (B) and TDP2 (C) protein sequences from different species. Human sequences are derived from the original cDNA sequencing data presented here. The methionine corresponding to the previously determined start codon in the human mRNA reference sequence is underlined, followed by the first 20 amino acids of the reference protein sequence. HUMAN: *Homo sapiens*, PANTR: *Pan troglodytes*, PONAB: *Pongo abelii*, HILLE: *Nomascus leucogenys*, MACMU: *Macaca mulatta*, CALJA: *Callithrix jacchus*. Asterisk: residue conserved in all sequences; colon: conservative substitution; dot: less conservative substitution.

4.2 Relationship between codon bias and expressed RNA codons

4.2.1 Database construction and computational analysis

Following importation of the normalised expression data, available expression values were found for: 27,850 out of 29,538 NM RefSeq entries for human brain tissue; 26,589 out of 29,538 NM entries for human circulating blood erythrocyte; 27,506 out of 29,538 NM entries for human DS AMKL cells; 6,642 out of 14,174 NM entries in *D. rerio*; 19,281 out of 23,894 NM for *C. elegans*; 4,673 out of 5,882 NM for *S. cerevisiae*; 2,426 out of 4,319 NM for *E. coli*. A summary of the range in the expression data and of the main genes with the highest and lowest expression values for the considered datasets is given in Table 4.2 for *H. sapiens* and Table 4.3 for the other investigated species available.

The frequency of each codon at genome level corresponds to the codon bias values already known for each genome [Nakamura Y, Codon Usage Database http://www.kazusa.or.jp/codon/]. In addition, codon sums at transcriptome level (codonome value), accounting for the abundance of each mRNA bearing that codon, has been calculated as per mil frequencies of each codon, obtaining the codonome bias (see Table 4.4 and Table 4.5 for *H. sapiens*, Table 4.6 for the other investigated species and Table 4.7 for human simulations).

Per mil frequencies for each codon (at genome level and at transcriptome level) were also grouped by the corresponding aaRS and then their expression values were loaded from the same normalised files as before. With the exception of *H. sapiens*, we could not find expression values for some aaRS for any of the investigated species (see Table 4.8 for *H. sapiens* and Table 4.9 for the other investigated species).

Homo sapiens

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

Table 4.2: The ten human genes with the highest and the lowest expression values in the studied datasets. The units of expression are given, following intra- and inter-sample normalisation by the TRAM software, as percentage of the mean value.

Danio rerio		***Caenorhabditis elegans***		***Saccharomyces cerevisiae***		***Escherichia coli***	
Brain		**Whole worm**					
Gene symbol	**Value**	**Gene symbol**	**Value**	**Gene symbol**	**Value**	**Gene symbol**	**Value**
mrps7	7051.00	*col-93*	1501.63	*CCW12*	561.08	*rplA*	989.00
epd	3800.07	*rpl-3*	1394.01	*RPL42B*	488.58	*rplE*	358.99
rpl12	2004.90	*rpl-2*	1388.10	*SSA1*	471.06	*rpsN*	860.42
rpl35a	1902.82	*col-92*	1383.15	*CWP2*	457.85	*rplV*	844.00
rps25	1815.57	*rps-25*	1378.99	*TDH3*	456.72	*rplP*	841.89
rpl6	1776.76	*asp-1*	1363.07	*ENO1*	448.49	*cspC*	835.24
uba52	1708.42	*rps-4*	1353.35	*HOR7*	445.87	*rplC*	803.48
ef1a	1662.62	*rps-1*	1350.21	*RPS21B*	431.01	*rpsA*	784.36
rps14	1589.81	*eft-3*	1333.51	*CDC19*	428.76	*rplD*	766.37
rpl35	1581.87	*act-2*	1331.33	*HHF2*	423.98	*atpF*	766.21
zgc:136367	1.66	*srg-45*	35.20	*UAF30*	33.23	*ydi*S	19.94
crygm2c	1.65	*srx-116*	35.00	*MND1*	32.22	*mfd*	15.58
ahr1a	1.61	*srsx-22*	34.92	*OSW1*	32.21	*ydiM*	15.43
otx5	1.60	*K05C4.3*	34.90	MEI4	31.89	*yegD*	6.30
pth1rb	1.54	*C16C4.1*	34.83	*YPT53*	31.18	*lolE*	5.44
lhx3	1.51	*C33E10.6*	34.55	*SPO74*	30.85	*yeiT*	4.04
nr5a5	1.51	*fbxa-159*	34.51	*ERR2*	29.28	*mobB*	3.90
mpll	1.49	*srh-88*	34.03	*RPL34B*	28.73	*yihU*	3.59
hoxa9b	1.44	*sre-20*	33.75	*YRF1-7*	27.51	*fadL*	2.98
insb	1.44	*srv-30*	33.15	*RPL40B*	27.09	*HHF2*	0.24

Table 4.3: The ten genes with the highest and the lowest expression values in the studied datasets. The units of expression are given, following intra- and inter-sample normalisation by the TRAM software, as percentage of the mean value.

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

Table 4.4: The per mil frequencies of codons (codon bias) and the per mil frequencies of the codon counts multiplied by the respective expression value (codonome bias) in the human studied datasets. First part.

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

Table 4.5: The per mil frequencies of codons (codon bias) and the per mil frequencies of the codon counts multiplied by the respective expression value (codonome bias) in the human studied datasets. Second part.

	D. rerio		*C. elegans*		*S. cerevisiae*		*E. coli*	
Codon	**Codon bias**	**Codonome bias**	**Codon bias**	**Codonome bias**	**Codon bias**	**Codonome bias**	**Codon bias**	**Codonome bias**
AAA	29.31	28.93	37.00	33.64	42.56	39.61	35.04	39.83
AAC	23.69	25.12	18.33	19.70	24.67	25.19	21.87	23.73
AAG	31.60	39.32	25.82	30.59	31.22	33.21	10.26	11.96
AAT	15.63	14.22	30.11	27.63	36.41	33.35	17.53	15.07
ACA	16.50	14.68	20.40	18.37	17.69	16.51	7.13	5.99
ACC	16.20	18.02	10.34	12.45	12.35	13.54	23.55	24.59
ACG	7.32	6.66	8.84	7.99	8.05	7.47	14.29	12.59
ACT	14.05	13.15	19.34	19.22	20.11	21.14	9.03	10.23
AGA	14.79	13.55	15.32	14.99	21.45	22.43	2.03	1.77
AGC	18.50	17.29	8.28	7.94	9.79	9.08	16.09	15.32
AGG	10.46	10.83	3.73	3.09	9.53	8.67	1.23	0.96
AGT	13.20	11.24	12.25	10.85	14.38	13.28	8.56	7.07
ATA	6.81	5.17	9.20	7.04	17.81	15.55	4.54	3.42
ATC	23.29	26.72	18.73	20.62	16.90	17.98	25.00	27.98
ATG	15.56	15.52	32.17	29.52	30.40	30.53	29.82	27.78
ATT	24.92	26.47	26.14	25.33	20.65	20.57	27.23	27.55
CAA	12.31	10.54	27.72	28.64	27.25	27.55	15.14	13.50
CAC	14.90	14.10	9.00	9.65	7.50	7.88	9.78	10.61
CAG	34.91	34.52	14.47	14.67	12.36	11.53	29.17	29.46
CAT	11.27	10.20	14.08	13.36	13.52	12.96	12.82	11.35
CCA	15.89	14.61	26.74	30.36	17.77	19.17	8.39	7.75
CCC	12.89	13.59	4.32	3.81	6.80	6.44	5.38	4.38
CCG	8.17	7.11	9.61	8.83	5.18	4.72	23.60	24.20
CCT	16.92	16.34	8.97	8.17	13.31	13.13	6.82	6.34
CGA	7.08	6.29	12.08	10.95	2.88	2.39	3.51	2.91
CGC	9.85	11.07	5.02	6.37	2.49	2.32	22.27	22.01
CGG	6.87	6.27	4.58	3.95	1.70	1.47	5.50	4.44
CGT	7.15	9.26	11.31	13.49	6.26	6.67	20.83	24.58
CTA	6.17	5.20	7.75	6.43	13.44	12.85	3.96	3.16
CTC	17.01	16.97	14.61	16.51	5.23	4.74	11.15	10.14
CTG	37.44	37.67	11.95	10.79	10.66	9.97	53.11	55.63
CTT	12.64	11.69	21.39	22.30	11.90	11.04	10.81	9.46
GAA	25.12	24.12	41.64	40.65	46.79	46.98	39.50	42.83
GAC	28.21	29.38	17.22	18.42	20.36	21.05	19.46	21.89
GAG	44.54	47.92	24.68	27.25	19.81	18.69	17.99	18.41
GAT	25.15	24.97	36.67	36.68	38.74	37.36	32.25	30.78
GCA	16.81	16.26	20.38	19.04	16.24	15.72	19.98	20.72
GCC	19.82	23.22	12.57	15.75	12.47	14.16	25.74	23.79
GCG	8.52	7.86	8.20	7.55	6.21	5.94	34.01	33.19
GCT	21.25	23.49	22.90	25.88	20.71	24.31	15.12	18.20
GGA	21.81	22.65	32.11	37.70	10.90	10.09	7.83	6.61
GGC	17.58	19.59	6.55	6.14	9.72	9.68	30.31	31.33
GGG	9.97	9.40	4.31	3.71	6.09	5.81	11.11	9.44
GGT	13.90	16.06	11.03	10.92	23.48	27.76	24.82	28.10
GTA	6.37	5.75	9.90	8.79	11.83	10.95	10.81	12.03
GTC	14.63	15.60	13.63	15.92	11.39	13.08	15.10	14.28
GTG	27.63	27.82	14.37	13.34	10.87	10.62	26.50	25.35
GTT	13.68	13.18	24.55	25.29	21.98	23.86	17.97	20.91
TAA	1.22	1.69	1.04	1.21	0.94	1.07	1.98	2.30
TAC	16.39	17.17	13.71	14.53	14.33	15.28	12.20	13.04
TAG	0.53	0.70	0.37	0.38	0.43	0.43	0.24	0.20
TAT	11.97	11.38	17.43	15.64	18.60	17.67	15.82	14.21
TCA	13.17	10.92	20.88	19.09	18.64	17.29	7.05	5.89
TCC	15.19	15.12	10.50	11.40	14.10	15.05	8.70	9.42
TCG	5.49	4.82	12.22	11.86	8.57	8.04	8.98	7.73
TCT	16.84	15.51	16.97	16.97	23.52	24.59	8.39	9.55
TGA	2.17	2.18	0.69	0.59	0.55	0.55	1.01	0.92
TGC	11.02	10.15	8.81	8.84	4.53	4.21	6.34	5.77
TGG	11.52	10.86	10.87	10.42	10.28	10.36	15.28	13.50
TGT	11.06	8.85	10.89	9.64	7.78	7.80	4.90	4.31
TTA	6.25	4.67	9.57	7.72	26.55	25.46	13.56	10.81
TTC	19.99	20.69	23.53	24.43	18.05	19.15	16.39	18.27
TTG	11.81	10.58	19.88	18.60	27.34	29.41	13.37	11.56
TTT	17.07	15.11	22.32	18.39	25.95	24.68	21.85	18.88

Table 4.6: The per mil frequencies of codons (codon bias) and the per mil frequencies of codon counts multiplied by the relative expression value (codonome bias) in: *Danio rerio*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Escherichia coli*.

Homo sapiens							
		Brain				**Erythrocytes**	
		Permutation 1	**Permutation** 2	**Raw Data**	**Random Numbers (from 1 to 10^4)**	**Random Numbers (from 1 to 10^5)**	
Codon	**Codon bias**	**Codonome bias**	**Codonome bias**	**Codonome bias**	**Codonome bias**	**Codon bias**	**Codonome bias**
AAA	25.88	26.24	25.78	24.90	25.88	25.83	22.88
AAC	18.90	18.83	18.89	19.35	18.94	18.94	15.45
AAG	32.30	32.65	32.19	33.40	32.32	32.41	27.83
AAT	17.58	17.53	17.47	16.99	17.51	17.52	15.12
ACA	15.41	15.53	15.31	15.01	15.42	15.36	14.73
ACC	18.40	18.39	18.30	18.79	18.47	18.42	15.68
ACG	5.96	5.94	5.97	6.16	5.98	5.98	6.76
ACT	13.53	13.64	13.40	13.06	13.53	13.49	13.90
AGA	12.23	12.21	12.26	11.48	12.16	12.14	13.40
AGC	19.72	19.66	19.94	19.84	19.81	19.74	19.47
AGG	11.64	11.55	11.82	11.26	11.59	11.56	8.74
AGT	12.87	12.93	12.81	12.48	12.83	12.85	15.20
ATA	7.60	7.71	7.54	6.97	7.56	7.51	7.47
ATC	20.06	20.03	19.97	20.97	20.01	20.13	15.34
ATG	21.45	16.39	16.12	15.86	16.15	21.49	19.38
ATT	16.20	21.57	21.41	21.88	21.42	16.18	10.71
CAA	12.84	12.87	12.85	12.03	12.78	12.74	8.12
CAC	14.80	14.74	14.75	14.74	14.84	14.77	16.71
CAG	34.76	34.34	34.95	35.46	34.78	34.83	38.22
CAT	11.12	11.16	11.13	10.55	11.10	11.05	9.06
CCA	17.67	17.86	17.67	17.42	17.70	17.69	23.64
CCC	19.81	19.68	19.99	20.05	19.81	19.84	27.68
CCG	6.97	6.92	7.07	6.89	6.98	6.97	8.57
CCT	18.17	18.31	18.24	18.03	18.18	18.19	24.51
CGA	6.28	6.22	6.36	6.45	6.30	6.33	7.67
CGC	10.10	9.85	10.10	10.48	10.14	10.13	11.18
CGG	11.47	11.15	11.52	12.02	11.50	11.56	16.70
CGT	4.54	4.48	4.53	4.70	4.56	4.56	4.52
CTA	7.09	7.14	7.10	6.80	7.07	7.07	6.33
CTC	18.51	18.42	18.58	18.62	18.51	18.48	22.69
CTG	38.38	37.84	38.52	39.13	38.43	38.39	38.97
CTT	13.33	13.44	13.29	12.81	13.30	13.28	12.20
GAA	31.28	31.40	31.35	30.61	31.24	31.26	28.01
GAC	25.23	25.07	25.31	26.13	25.24	25.34	25.77
GAG	40.42	40.26	40.69	42.02	40.41	40.54	39.27
GAT	22.88	23.00	22.73	23.16	22.89	22.97	30.11
GCA	16.32	16.46	16.27	16.23	16.31	16.34	15.45
GCC	27.43	27.33	27.51	28.20	27.51	27.48	30.64
GCG	7.13	7.10	7.15	7.11	7.17	7.12	8.71
GCT	18.45	18.46	18.39	18.65	18.46	18.48	18.71
GGA	16.70	16.61	16.79	16.38	16.67	16.68	12.79
GGC	21.80	21.62	21.99	22.33	21.91	21.87	20.63
GGG	15.96	15.96	16.00	16.09	15.97	15.96	20.74
GGT	10.73	10.75	10.75	10.84	10.78	10.76	10.78
GTA	7.30	7.37	7.23	7.07	7.28	7.30	5.82
GTC	13.99	14.04	14.01	14.29	14.06	14.01	12.99
GTG	27.22	27.21	27.17	28.10	27.29	27.32	28.64
GTT	11.23	11.39	11.08	10.97	11.20	11.25	10.47
TAA	0.66	0.72	0.68	0.61	0.66	0.66	0.38
TAC	14.58	14.44	14.52	15.01	14.61	14.61	18.41
TAG	0.49	0.51	0.50	0.48	0.50	0.49	0.37
TAT	12.12	12.05	12.13	11.82	12.08	12.08	7.45
TCA	12.78	12.92	12.61	12.13	12.69	12.70	11.52
TCC	17.41	17.59	17.39	17.54	17.40	17.39	19.89
TCG	4.45	4.43	4.52	4.52	4.47	4.46	4.74
TCT	15.38	15.52	15.32	14.99	15.35	15.36	17.44
TGA	1.10	1.16	1.12	1.06	1.10	1.10	0.77
TGC	11.78	11.57	11.68	11.50	11.76	11.70	8.84
TGG	12.09	12.02	12.13	11.86	12.05	12.05	8.90
TGT	10.38	10.32	10.23	9.68	10.36	10.26	5.68
TTA	7.96	8.01	7.94	7.37	7.91	7.94	8.60
TTC	19.09	19.22	18.95	19.38	19.12	19.07	21.06
TTG	12.89	12.95	12.87	12.56	12.86	12.88	11.67
TTT	17.20	17.30	17.16	16.69	17.12	17.16	15.88

Table 4.7: The per mil frequencies of codons (codon bias) and the per mil frequencies of expressed codons (codonome bias). In these analyses the real genes expression values has been substituted with: permuted values (two examples), raw data (values non-normalised by TRAM), random numbers from 1 to 10^4 and random numbers from 1 to 10^5.

Homo sapiens

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

Table 4.8: The per mil frequencies of codons grouped by aminoacyl-tRNA synthetase (codon bias by aaRS), the per mil frequencies of the expressed codon grouped by aminoacyl-tRNA synthetase (codonome bias by aaRS), and the aminoacyl-tRNA synthetases expression values in human studied datasets.

Species	Gene symbol	Codon bias	Codonome bias	Expression value	Species	Gene symbol	Codon bias	Codonome bias	Expression value
D. rerio	*aars*	66.40	70.84	139.36	*S. cerevisiae*	*ALA1*	55.62	60.13	162.49
	cars	22.07	19.00	23.19		*CDC60*	95.12	93.46	143.93
	eprs	123.52	123.69	/		*DED81*	61.08	58.54	192.22
	farsa	37.06	35.80	/		*DPS1*	59.10	58.40	260.83
	farsb	37.06	35.80	43.88		*FRS1*	44.00	43.83	260.76
	hars	26.16	24.31	103.73		*FRS2*	44.00	43.83	156.46
	iars	55.03	58.36	12.94		*GLN4*	39.62	39.08	118.51
	kars	60.91	68.25	115.32		*GRS1*	50.19	53.35	180.08
	lars	91.33	86.78	/		*GUS1*	66.61	65.67	219.64
	mars	15.56	15.52	9.68		*HTS1*	21.02	20.84	162.48
	qars	47.22	45.05	13.39		*ILS1*	55.36	54.10	154.64
	rars	56.20	57.27	16.17		*KRS1*	73.79	72.82	218.47
	sars	82.39	74.91	/		*MES1*	30.40	30.53	164.98
	si:dkey-274m14.2	39.33	39.34	/		*SES1*	88.99	87.33	188.41
	si:dkey-276i5.1	63.27	67.70	105.94		*THS1*	58.21	58.65	224.02
	tars	54.08	52.51	43.84		*TYS1*	32.94	32.94	147.53
	vars	62.30	62.35	10.03		*VAS1*	56.08	58.51	168.58
	wars	11.52	10.86	27.10		*WRS1*	10.28	10.36	155.73
	wu:fc17a11	53.36	54.35	/		*YDR341C*	44.31	43.93	/
	yars	28.37	28.56	23.35		*YHR020W*	43.06	43.47	/
						YNL247W	12.31	12.00	/
C. elegans	*ars-2*	64.05	68.23	149.15	*E. coli*	*alaS*	94.84	95.90	364.21
	crs-1	19.70	18.48	123.51		*argS*	55.38	56.67	174.99
	drs-1	53.89	55.10	249.92		*asnS*	39.40	38.80	/
	ers-1	42.19	43.31	161.41		*aspS*	51.70	52.68	171.33
	ers-3	66.31	67.90	82.72		*cysS*	11.25	10.08	100.07
	frs-1	45.84	42.82	202.79		*glnS*	15.14	13.50	/
	frs-2	45.84	42.82	192.17		*glsS*	29.17	29.46	/
	grs-1	54.00	58.47	152.32		*gltX*	57.49	61.24	247.61
	hrs-1	23.07	23.02	171.03		*glyQS*	74.08	75.48	/
	irs-1	54.07	52.99	220.43		*hisS*	22.60	21.96	194.27
	krs-1	62.82	64.24	323.80		*ileS*	56.77	58.95	129.04
	lrs-1	85.16	82.34	143.51		*leuS*	105.96	100.77	322.29
	mrs-1	32.17	29.52	185.49		*lysU*	45.30	51.79	/
	nrs-1	48.44	47.32	231.11		*metG*	29.82	27.78	122.34
	prs-1	49.64	51.17	221.29		*pheS*	38.24	37.15	288.63
	rrt-1	52.05	52.84	199.94		*pheT*	38.24	37.15	299.49
	srs-1	81.11	78.12	135.54		*proS*	44.19	42.67	29.79
	trs-1	58.92	58.03	175.71		*serS*	57.76	54.97	/
	vrs-2	62.44	63.35	146.32		*thrS*	54.00	53.40	457.90
	wrs-1	10.87	10.42	/		*trpS*	15.28	13.50	92.78
	yrs-1	31.14	30.16	90.52		*tyrS*	28.02	27.25	/
						valS	70.38	72.58	162.41

Table 4.9: The per mil frequencies of codons grouped by aminoacyl-tRNA synthetase (codon bias by aaRS), the per mil frequencies of the expressed codons grouped by aminoacyl-tRNA synthetase (codonome bias by aaRS), and the aminoacyl-tRNA synthetases expression values in *Danio rerio*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Escherichia coli*. "/": value not found.

4.2.2 Statistical analysis

The following results were exported in order to submit them to statistical analysis using first default and then test calculations: a) codon bias, b) codonome bias, c) codon bias grouped by aaRS, d) codonome bias grouped by aaRS, and e) the aaRS expression values ("a", "b", "c" and "d" are expressed as per mil frequencies). Correlation coefficients and p values for each comparison are listed in Table 4.10. See Fig. 4.5 for human brain correlation graphs, Fig. 4.6 for human circulating blood erythrocyte graphs, Fig. 4.7 for human DS-AMKL cells graphs and Fig. 4.8-4.11 for the other investigated species, *D. rerio*, *C. elegans*, *S. cerevisiae* and *E. coli*, respectively.

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

When grouped by aaRS, codon bias and codonome bias, when compared to aaRS mRNA expression values, show no correlation, with really low coefficients (sometimes even negative ones), and p values of at least > 0.1 (p values at least > 0.05 only in the case of *E. coli* dataset), even when using random and permuted expression values.

Subset	X variable	Y variable	(r)	p value
Human brain	a) codon bias	b) codonome bias	0.996517	< 0.0001
	c) codon bias by aaRS	d) codonome bias by aaRS	0.999457	< 0.0001
	c) codon bias by aaRS	e) aaRS expression	-0.022970	NS
	d) codonome bias by aaRS	e) aaRS expression	-0.021290	NS
Human brain with absolute numbers instead of per mil frequencies	a) codon bias	b) codonome bias	0.996546	< 0.0001
	c) codon bias by aaRS	d) codonome bias by aaRS	0.999457	< 0.0001
	c) codon bias by aaRS	e) aaRS expression	-0.051980	NS
	d) codonome bias by aaRS	e) aaRS expression	-0.050950	NS
Human brain with a first permutation of the expression values	a) codon bias	b) codonome bias	0.999838	< 0.0001
	c) codon bias by aaRS	d) codonome bias by aaRS	0.999937	< 0.0001
	c) codon bias by aaRS	e) aaRS expression	-0.024030	NS
	d) codonome bias by aaRS	e) aaRS expression	-0.022320	NS
Second human brain with a second permutation of the expression values	a) codon bias	b) codonome bias	0.999943	< 0.0001
	c) codon bias by aaRS	d) codonome bias by aaRS	0.999982	< 0.0001
	c) codon bias by aaRS	e) aaRS expression	-0.023000	NS
	d) codonome bias by aaRS	e) aaRS expression	-0.023420	NS
Human brain with non-normalised expression values	a) codon bias	b) codonome bias	0.998791	< 0.0001
	c) codon bias by aaRS	d) codonome bias by aaRS	0.999925	< 0.0001
	c) codon bias by aaRS	e) aaRS expression	-0.052180	NS
	d) codonome bias by aaRS	e) aaRS expression	-0.050720	NS
Human brain with random expression values from 1 to 10^4	a) codon bias	b) codonome bias	0.999990	< 0.0001
	c) codon bias by aaRS	d) codonome bias by aaRS	0.999996	< 0.0001
	c) codon bias by aaRS	e) aaRS expression	-0.052180	NS
	d) codonome bias by aaRS	e) aaRS expression	-0.051980	NS
Human circulating blood erythrocytes	a) codon bias	b) codonome bias	0.979111	< 0.0001
	c) codon bias by aaRS	d) codonome bias by aaRS	0.998358	< 0.0001
	c) codon bias by aaRS	e) aaRS expression	0.119425	NS
	d) codonome bias by aaRS	e) aaRS expression	0.126501	NS
Human circulating blood erythrocytes with random expression values from 1 to 10^5	a) codon bias	b) codonome bias	0.998791	< 0.0001
	c) codon bias by aaRS	d) codonome bias by aaRS	0.937790	< 0.0001
	c) codon bias by aaRS	e) aaRS expression	0.074207	NS
	d) codonome bias by aaRS	e) aaRS expression	0.027150	NS
Human brain from patients affected by Trisomy 21 and Acute Megakaryoblastic Leukemia	a) codon bias	b) codonome bias	0.990428	< 0.0001
	c) codon bias by aaRS	d) codonome bias by aaRS	0.996594	< 0.0001
	c) codon bias by aaRS	e) aaRS expression	0.015140	NS
	d) codonome bias by aaRS	e) aaRS expression	0.041824	NS
Danio rerio brain	a) codon bias	b) codonome bias	0.986128	< 0.0001
	c) codon bias by aaRS	d) codonome bias by aaRS	0.992635	< 0.0001
	c) codon bias by aaRS	e) aaRS expression	0.384812	> 0.1743
	d) codonome bias by aaRS	e) aaRS expression	0.431892	> 0.1230
Caenorhabditis elegans	a) codon bias	b) codonome bias	0.979831	< 0.0001
	c) codon bias by aaRS	d) codonome bias by aaRS	0.991042	< 0.0001
	c) codon bias by aaRS	e) aaRS expression	0.026048	NS
	d) codonome bias by aaRS	e) aaRS expression	0.026811	NS
Saccharomyces cerevisiae	a) codon bias	b) codonome bias	0.991204	< 0.0001
	c) codon bias by aaRS	d) codonome bias by aaRS	0.996790	< 0.0001
	c) codon bias by aaRS	e) aaRS expression	0.224813	> 0.3698
	d) codonome bias by aaRS	e) aaRS expression	0.217939	> 0.3850
Escherichia coli	a) codon bias	b) codonome bias	0.988796	< 0.0001
	c) codon bias by aaRS	d) codonome bias by aaRS	0.996650	< 0.0001
	c) codon bias by aaRS	e) aaRS expression	0.510390	> 0,0519
	d) codonome bias by aaRS	e) aaRS expression	0.502108	> 0,0565

Table 4.10: Correlation coefficients (r) and p values of comparisons. a) The per mil frequencies of codons (codon bias), b) the per mil frequencies of codons number multiplied by expression value (codonome bias), c) the per mil frequencies of codons grouped by aminoacyl-tRNA synthetase (codon bias by aaRS), d) the per mil frequencies of real expressed codon grouped by aminoacyl-tRNA synthetase (codonome bias by aaRS), e) the aminoacyl-tRNA synthetases mRNA expression values (aaRS expression). NS, not significant.

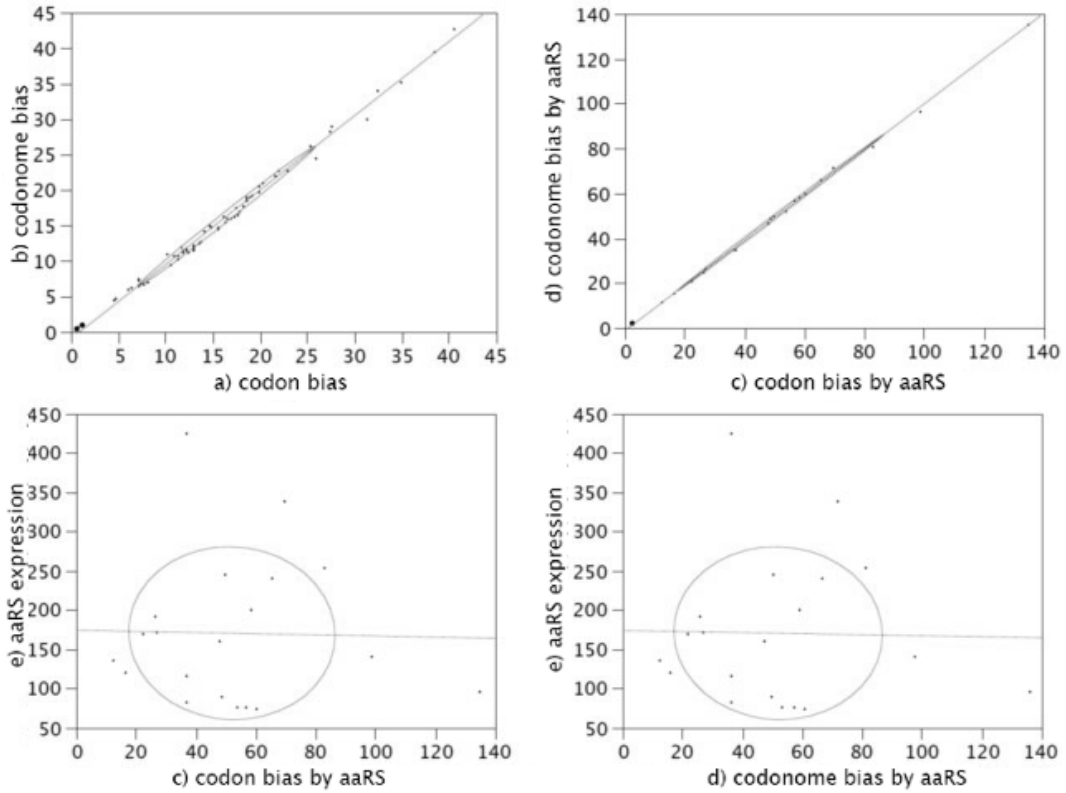


Figure 4.5: Correlation graphs in human brain. "a": the per mil frequencies of each codon at genome level (codon bias); "b": the per mil frequencies of each codon multiplied by expression value (codonome bias); "c": the per mil frequencies of codons grouped by aminoacyl-tRNA synthetase (codon bias by aaRS); "d": the per mil frequencies of real expressed codons grouped by aaRS (codonome bias by aaRS); "e": the aaRS expression values. See Table 4.10 for correlation coefficients and p values. The elliptic line represents the density ellipse at 0.50.

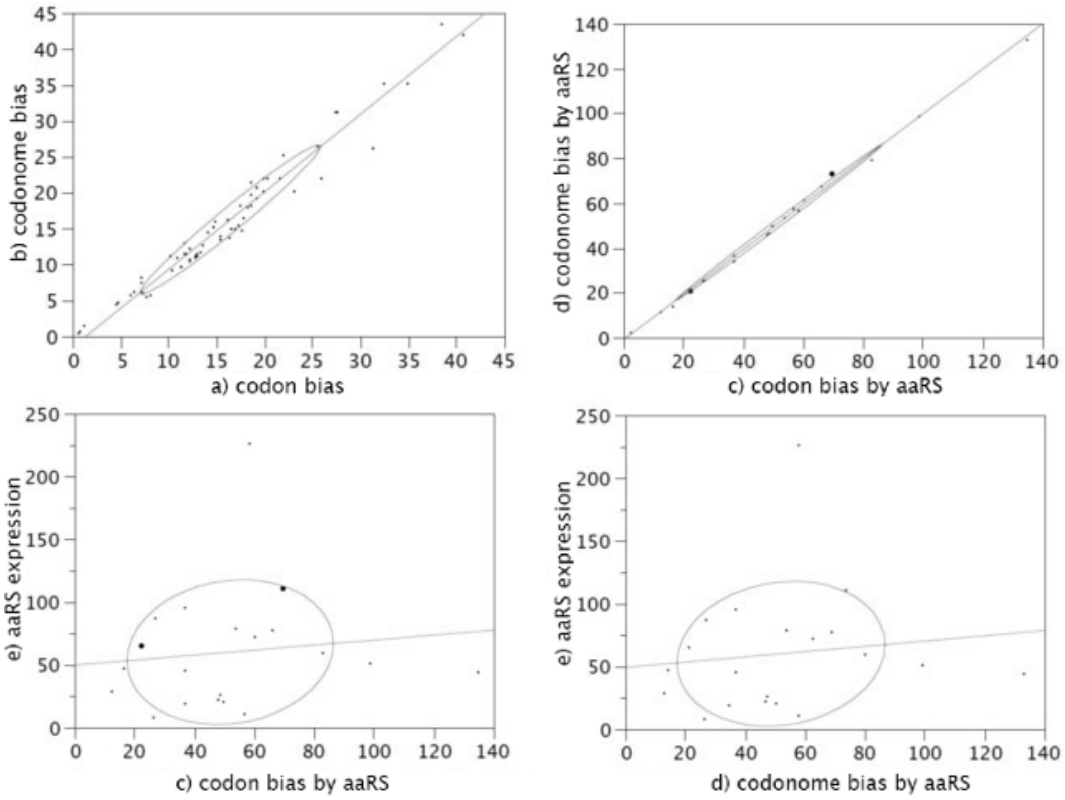


Figure 4.6: Correlation graphs in human circulating blood erythrocytes. "a": the per mil frequencies of each codon at genome level (codon bias); "b": the per mil frequencies of each codon multiplied by expression value (codonome bias); "c": the per mil frequencies of codons grouped by aminoacyl-tRNA synthetase (codon bias by aaRS); "d": the per mil frequencies of real expressed codons grouped by aaRS (codonome bias by aaRS); "e": the aaRS expression values. See Table 4.10 for correlation coefficients and p values. The elliptic line represents the density ellipse at 0.50.

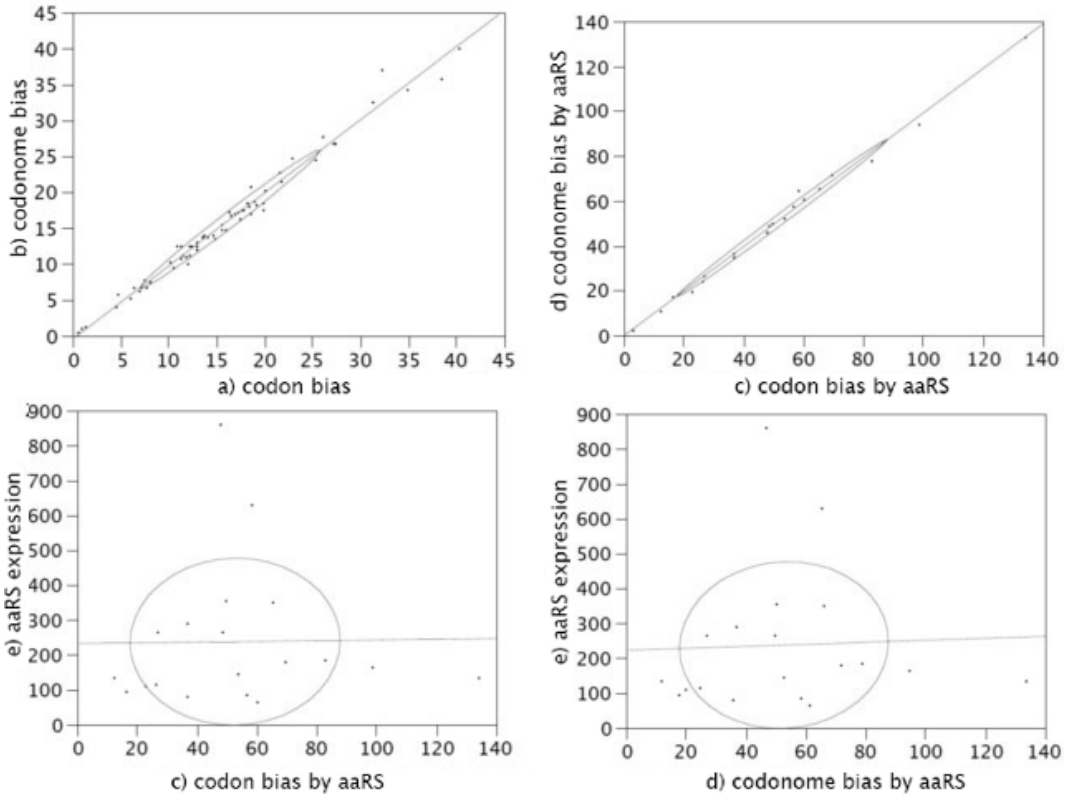


Figure 4.7: Correlation graphs in human Down Syndrome-related acute megakaryoblastic leukemia cells. "a": the per mil frequencies of each codon at genome level (codon bias); "b": the per mil frequencies of each codon multiplied by expression value (codonome bias); "c": the per mil frequencies of codons grouped by aminoacyl-tRNA synthetase (codon bias by aaRS); "d": the per mil frequencies of real expressed codons grouped by aaRS (codonome bias by aaRS); "e": the aaRS expression values. See Table 4.10 for correlation coefficients and p values. The elliptic line represents the density ellipse at 0.50.

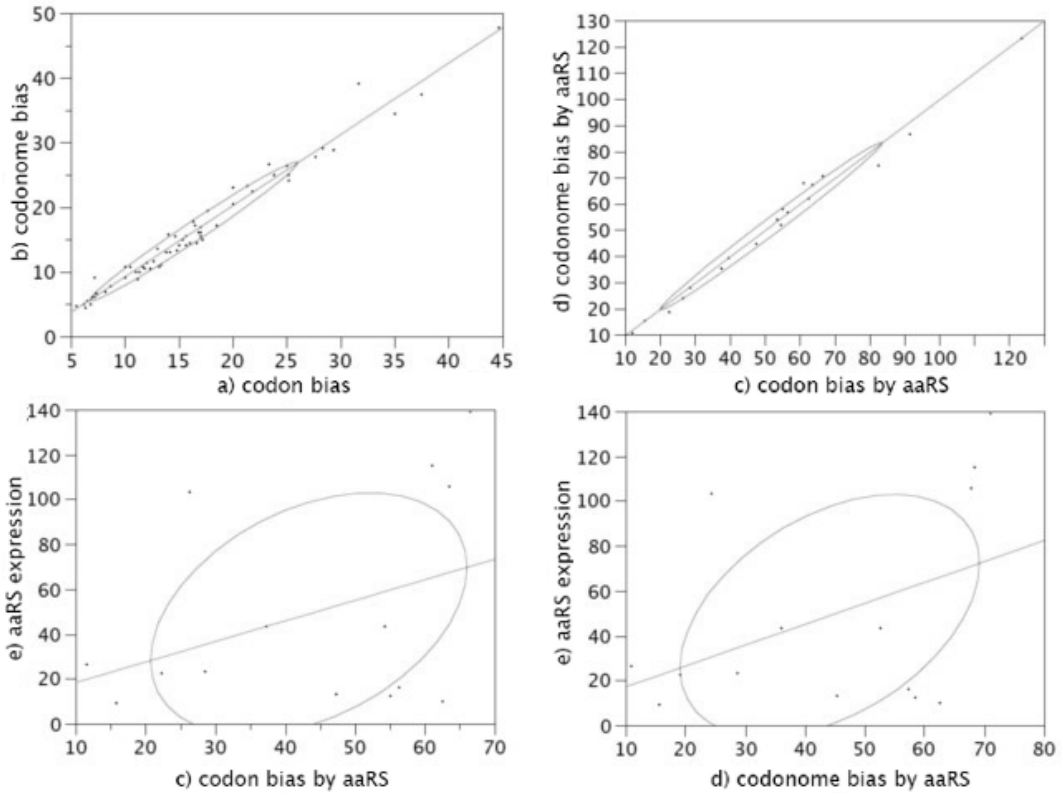


Figure 4.8: Correlation graphs in *Danio rerio* brain. "a": the per mil frequencies of each codon at genome level (codon bias); "b": the per mil frequencies of each codon multiplied by expression value (codonome bias); "c": the per mil frequencies of codons grouped by aminoacyl-tRNA synthetase (codon bias by aaRS); "d": the per mil frequencies of real expressed codons grouped by aaRS (codonome bias by aaRS); "e": the aaRS expression values. See Table 4.10 for correlation coefficients and p values. The elliptic line represents the density ellipse at 0.50.

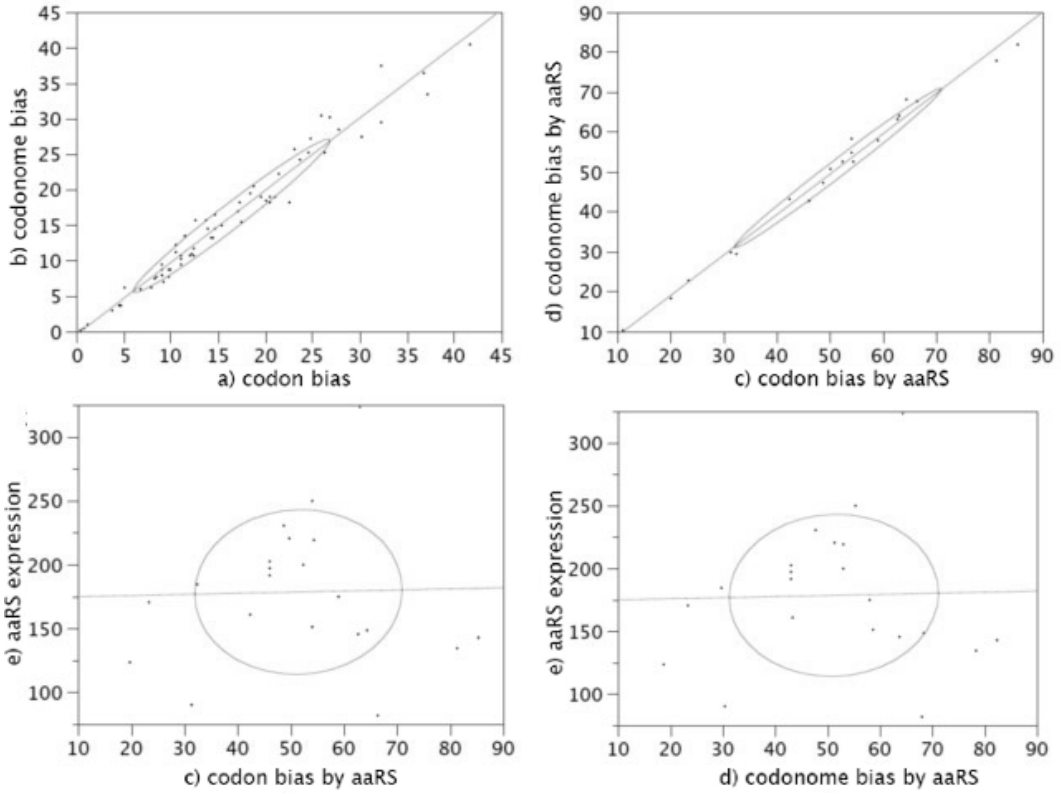


Figure 4.9: Correlation graphs in *Caenorhabditis elegans*. "a": the per mil frequencies of each codon at genome level (codon bias); "b": the per mil frequencies of each codon multiplied by expression value (codonome bias); "c": the per mil frequencies of codons grouped by aminoacyl-tRNA synthetase (codon bias by aaRS); "d": the per mil frequencies of real expressed codons grouped by aaRS (codonome bias by aaRS); "e": the aaRS expression values. See Table 4.10 for correlation coefficients and p values. The elliptic line represents the density ellipse at 0.50.

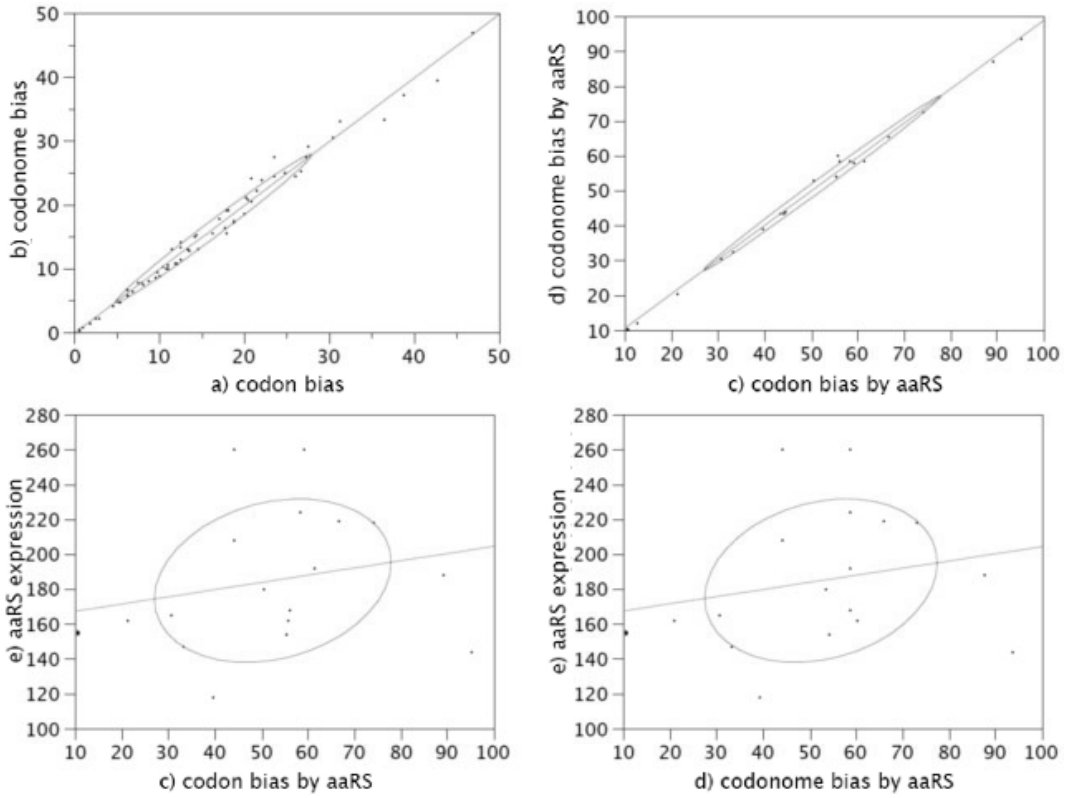


Figure 4.10: Correlation graphs in *Saccharomices cerevisiae*. "a": the per mil frequencies of each codon at genome level (codon bias); "b": the per mil frequencies of each codon multiplied by expression value (codonome bias); "c": the per mil frequencies of codons grouped by aminoacyl-tRNA synthetase (codon bias by aaRS); "d": the per mil frequencies of real expressed codons grouped by aaRS (codonome bias by aaRS); "e": the aaRS expression values. See Table 4.10 for correlation coefficients and p values. The elliptic line represents the density ellipse at 0.50.

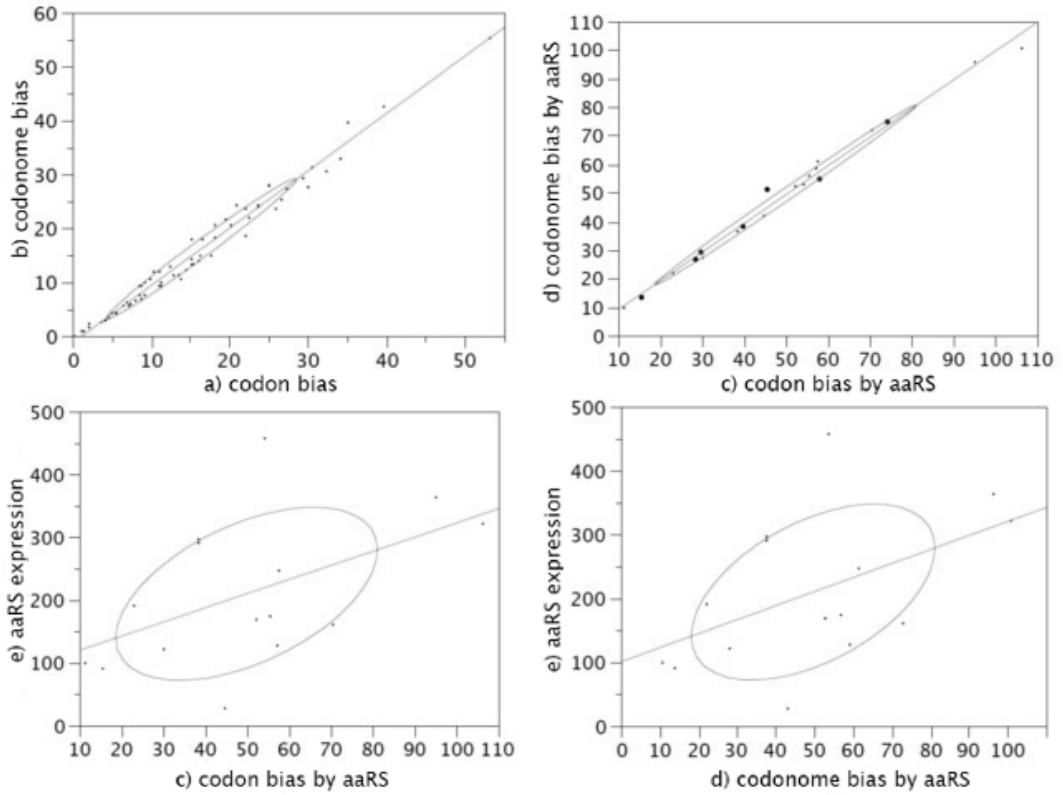


Figure 4.11: Correlation graphs in *Escherichia coli*. "a": the per mil frequencies of each codon at genome level (codon bias); "b": the per mil frequencies of each codon multiplied by expression value (codonome bias); "c": the per mil frequencies of codons grouped by aminoacyl-tRNA synthetase (codon bias by aaRS); "d": the per mil frequencies of real expressed codons grouped by aaRS (codonome bias by aaRS); "e": the aaRS expression values. See Table 4.10 for correlation coefficients and p values. The elliptic line represents the density ellipse at 0.50.

4.3 Analysis of ASD implicated genes expression

4.3.1 ASD implicated genes

TargetScan prediction found, out of a total of 447 ASD implicated genes identified from the available literature, 130 genes not having MREs for miR-ASD and 317 genes having at least one predicted MREs for miR-ASD. Among them, in healthy and ASD considered samples from the first dataset [Voineagu *et al.*, 2011], an expression value was available for 66 ASD-associated genes that do not have MREs and 160 ASD-associated genes that have MREs. In fetal, child and adult samples from the second dataset [Colantuoni *et al.*, 2011], an expression value was available for 114 ASD implicated genes that do not have MREs and 281 ASD implicated genes that have MREs.

4.3.2 Statistical analysis

Median values of Pearson's pairwise correlation coefficients and p values calculated for ASD implicated genes without MREs for miR-ASD and for ASD implicated genes with at least one MRE for miR-ASD for all the analysed data sets are listed in Table 4.11.

Comparing these values using as a background the median values of Pearson's pairwise correlation coefficients calculated for all the 1000 permutations of randomly picked genes not implicated with ASD, significant co-expression was found for ASD implicated genes that share MREs for miR-ASD with *MSNP1AS* in healthy and ASD affected adults (Fig. 4.12 and Fig. 4.13, respectively). During the normal cortex development, a significant co-expression of ASD implicated genes was found in child and adult samples (Fig. 4.14, Fig. 4.15, respectively) but not in fetal samples (Fig. 4.16).

Considering the ASD implicated risk genotype in ASD affected adults, not significant differences were found comparing the expression values of the two groups of ASD implicated genes (Fig. 4.17).

	ASD implicated genes			
	without MREs		with MREs	
	Median value	**p value**	**Median value**	**p value**
Healthy adults	0.029	0.147	0.073	0.000
ASD affected adults	0.053	0.076	0.066	0.002
Healthy fetal samples	0.006	0.115	0.005	0.068
Healthy child samples	0.010	0.073	0.019	0.001
Healthy adults samples	0.006	0.080	0.018	0.000

Table 4.11: Median values of Pearson's pairwise correlation coefficients. Pearson's pairwise correlation coefficients were calculated for expression values between all the possible gene pairs in ASD implicated genes without and with MREs for miR-ASD, respectively.

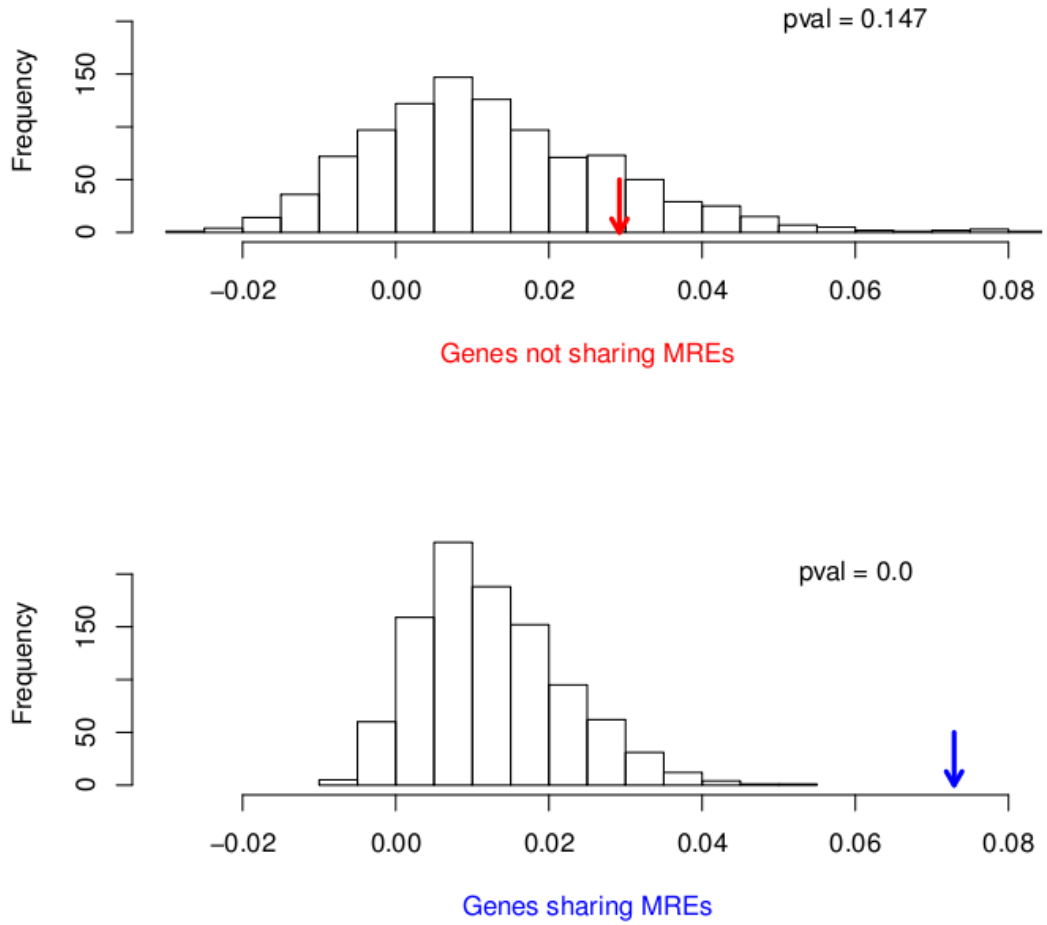


Figure 4.12: Pearson's pairwise correlation coefficients in healty adults. The arrows represent the median value of coefficients for each group of autism associated genes: ASD implicated genes without MREs for miR-ASD in red and ASD implicated genes with at least one MRE for miR-ASD in blue. See Table 4.11 for median and p values. The background is the median value of coefficients calculated for 1000 permutations of genes not implicated with ASD. pval: p value.

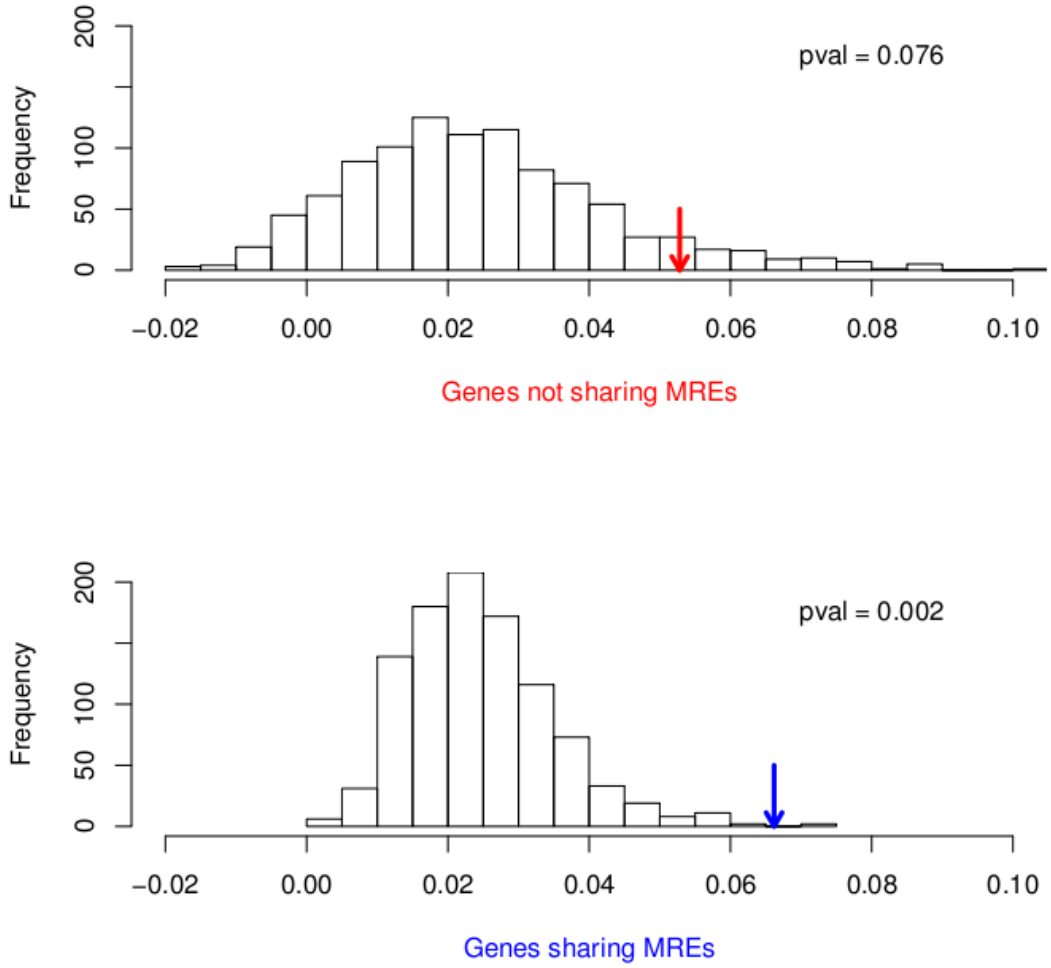


Figure 4.13: Pearson's pairwise correlation coefficients in ASD affected adults. The arrows represent the median value of coefficients for each group of autism associated genes: ASD implicated genes without MREs for miR-ASD in red and ASD implicated genes with at least one MRE for miR-ASD in blue. See Table 4.11 for median and p values. The background is the median value of coefficients calculated for 1000 permutations of genes not implicated with ASD. pval: p value.

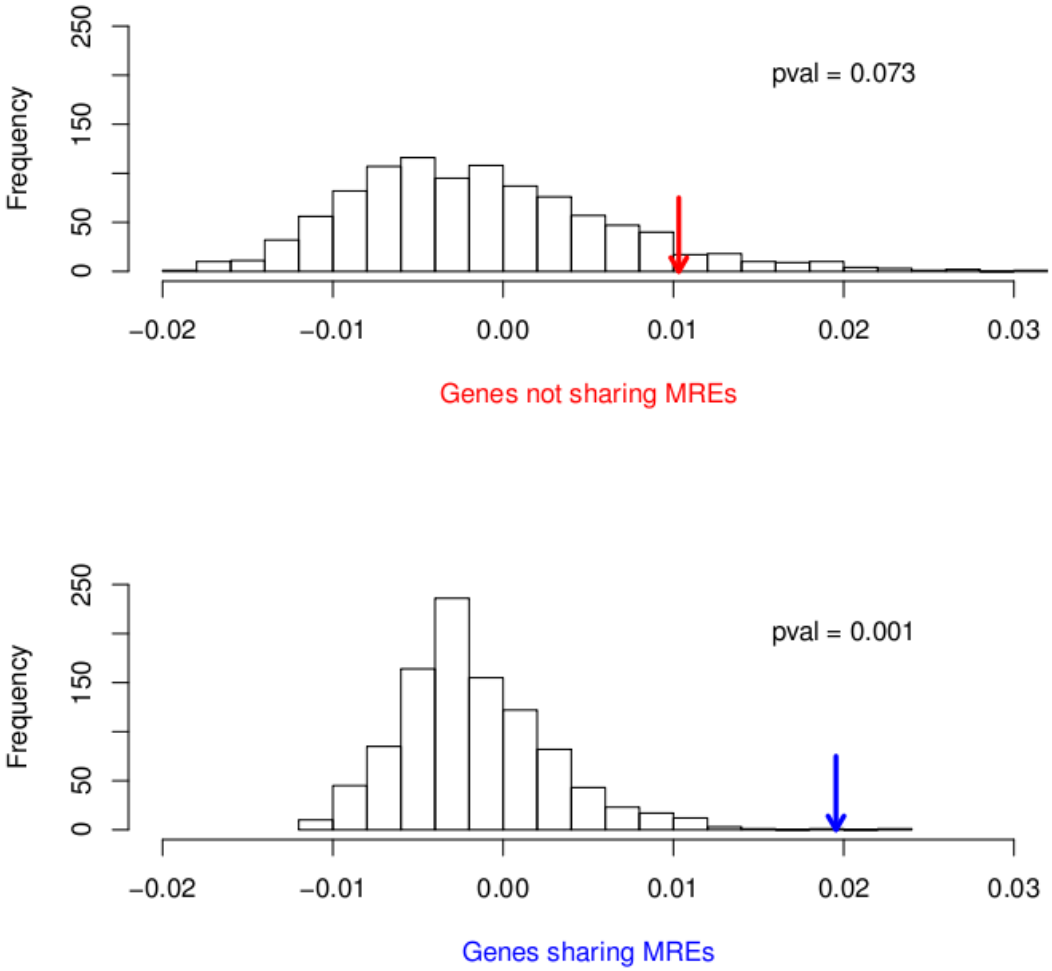


Figure 4.14: Pearson's pairwise correlation coefficients in healthy child samples. TThe arrows represent the median value of coefficients for each group of autism associated genes: ASD implicated genes without MREs for miR-ASD in red and ASD implicated genes with at least one MRE for miR-ASD in blue. See Table 4.11 for median and p values. The background is the median value of coefficients calculated for 1000 permutations of genes not implicated with ASD. pval: p value.

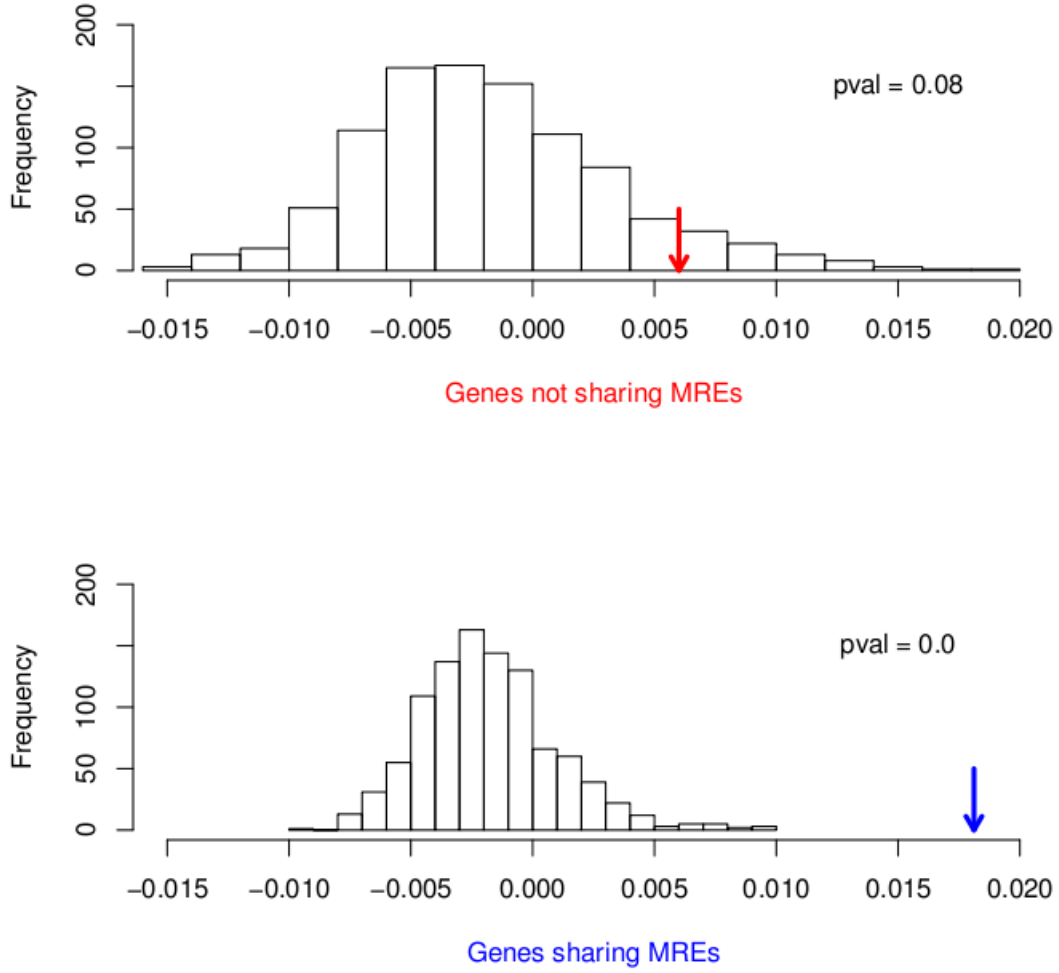


Figure 4.15: Pearson's pairwise correlation coefficients in healthy adult samples. The arrows represent the median value of coefficients for each group of autism associated genes: ASD implicated genes without MREs for miR-ASD in red and ASD implicated genes with at least one MRE for miR-ASD in blue. See Table 4.11 for median and p values. The background is the median value of coefficients calculated for 1000 permutations of genes not implicated with ASD. pval: p value.

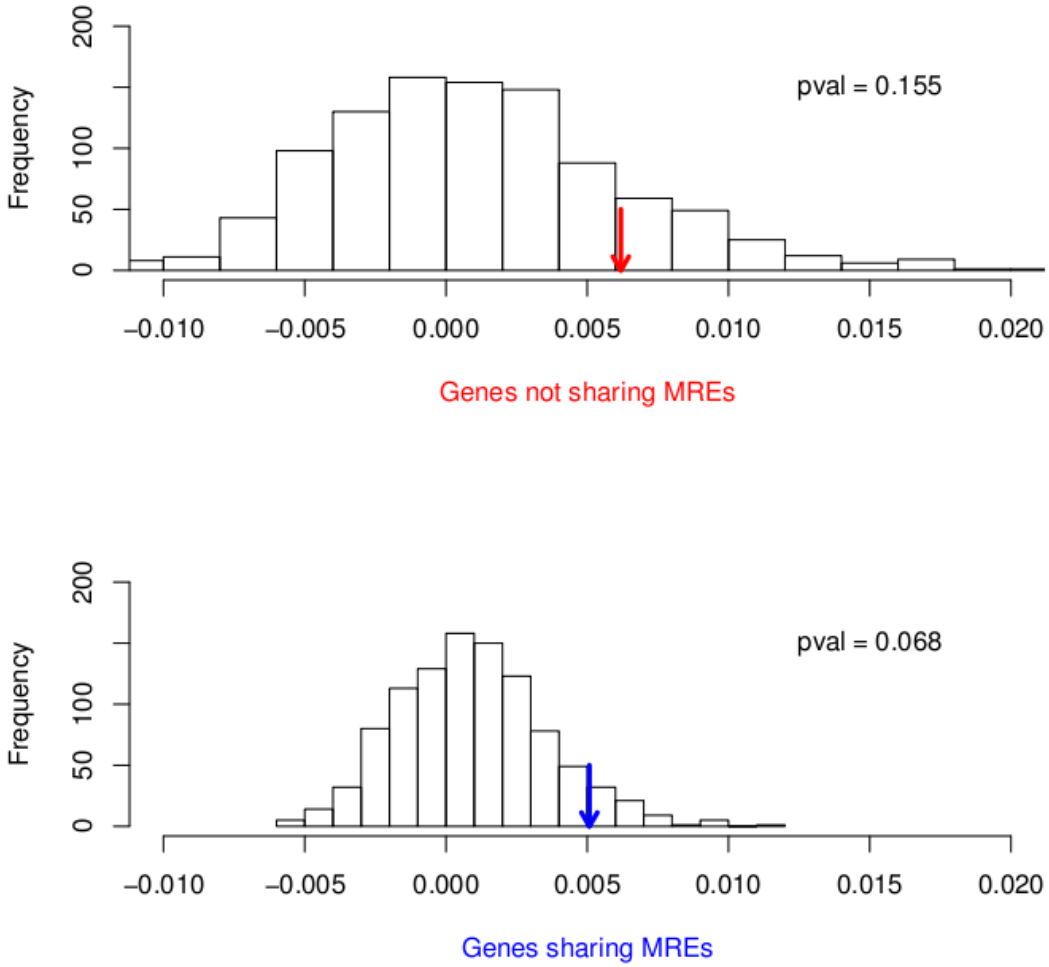


Figure 4.16: Pearson's pairwise correlation coefficients in healthy fetal samples. The arrows represent the median value of coefficients for each group of autism associated genes: ASD implicated genes without MREs for miR-ASD in red and ASD implicated genes with at least one MRE for miR-ASD in blue. See Table 4.11 for median and p values. The background is the median value of coefficients calculated for 1000 permutations of genes not implicated with ASD. pval: p value.

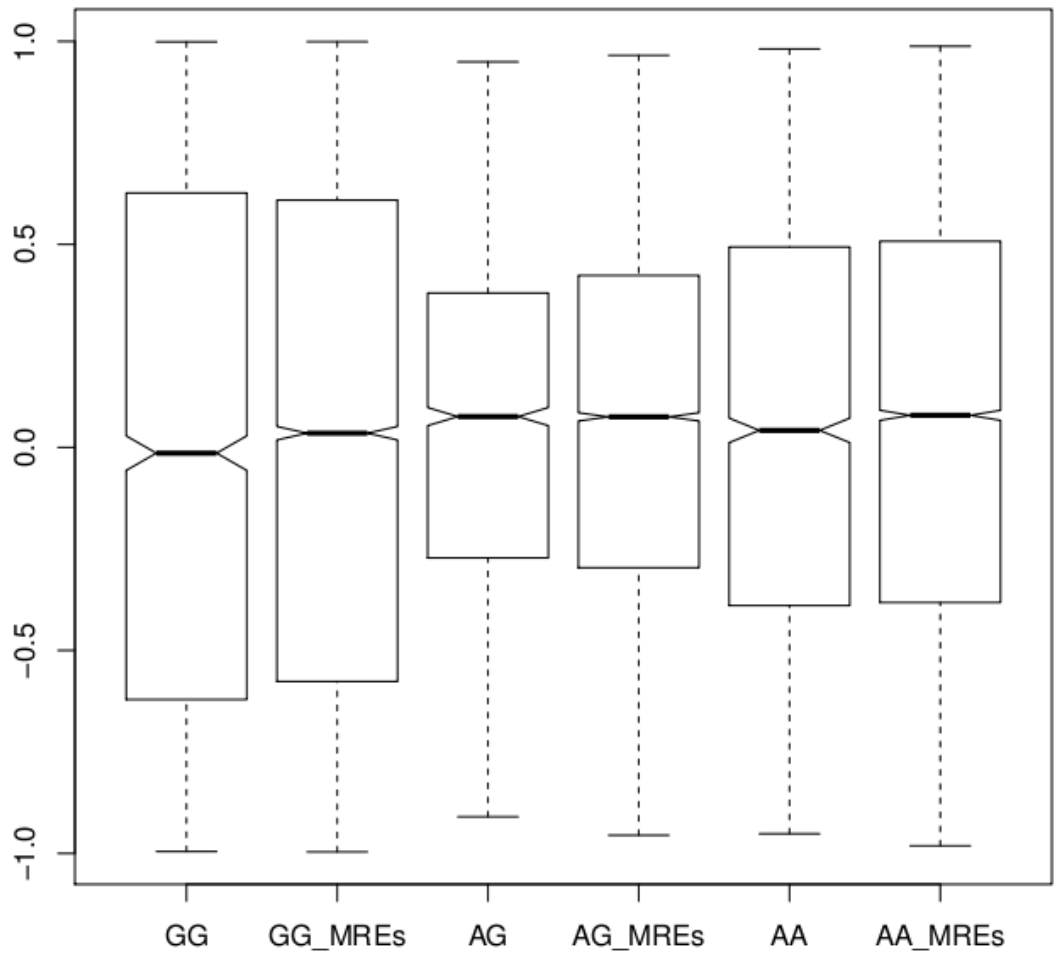


Figure 4.17: Expression values of ASD implicated genes expressed in ASD samples divided by genotype: AA is the genotype associated with ASD and with a higher expression of *MSNP1AS*. For each genotype, ASD implicated genes without MREs for miR-ASD are on the left and ASD implicated genes with at least one MRE for miR-ASD on the right.

Chapter 5

Discussion

The late 1980s and early 1990s were fertile years thanks to the automation of sequencing techniques and to the computer spread. This period of fervent activity gave rise to a flourishing number of new molecular structures and sequences and to proliferation of new databases in which to store them. The public availabilty of data is of an unestimated valour, because the collective use of data leads to the discovery of new knowledge which goes beyond the results yielded by individual experiments. In this work were presented three examples about the development of new computational tools able to analyse the massive amount of publicly avalilable data, which is often not completely exploited, in order to answer to important biological questions. Each subject will be discuss in a separate section.

5.1 mRNA 5′ coding sequence

The continuous incorporation of information derived from individual and large-scale cDNA sequencing projects (including those specifically designed to characterise mRNA 5′ end [Carninci et al., 1996; Suzuki et al., 2000; Porcel et al., 2004]) in the last few years led to continuous improvement of completeness of mRNA reference sequences (e.g., RefSeq), and also to the corresponding protein coding sequences. However, genome browsers do not appear to systematically extract useful information from the vast quantity of EST data. To date, EST data remain invaluable due to significantly longer continuous RNA sequences they may provide in comparison with the very short fragments typically deposited in current high-throughput nucleotide sequencing databases. We first showed in zebrafish that EST analysis by 5'_ORF_Extender software could extend the currently known mRNA CDS [Frabetti et al., 2007], thereby differing from other methods, which do not

incorporate prediction of the putative CDS extension (e.g., [Kitagawa et al., 2005]).

In this work, we have presented a modified strategy that was able to analyse the much more numerous human sequences. Firstly, we fully revised the software algorithm by using pre-computed coordinates of the UCSC-downloaded RefSeqs and ESTs genome alignment data (rather than the results of a large scale BLAST comparison), and specific UCSC-downloaded EST sequence entries. Rather than GenBank EST raw entries, these are EST sequence entries in which nucleotides which are unaligned to the genome are removed, and undetermined ("N") or mismatched nucleotides are replaced by the corresponding nucleotides present on the genome. This key change significantly improved a number of areas: the software speed of analysis, sensitivity (due to the implementation of management of sequence in 'complement' orientation with respect to the genome recorded DNA strand, with consequent identification of previously undetected mRNA extensions thank to ESTs in opposite orientation to the corresponding mRNA), specificity (due to the use of EST sequence entries processed by UCSC as described above, thereby avoiding false positive identification of start codons in the EST sequence, and possibly false negatives too, thus further improving sensitivity), and usability (due to the removal of all steps previously requiring Unix functions, such as local running of BLAST and manipulations of large text files). Furthermore, we adopted an original quality filter which was able to test if each single EST candidate with sequence information of possible use for extending a known mRNA, was attributed to the same locus of that mRNA by an updated, complete and embedded version of UniGene. Lastly, we automated data summarisation for an analysed genome.

Following these improvements, 5'_ORF_Extender recognised a total of 477 loci, out of the 18,665 human loci represented in the mRNA reference set, as bona fide candidates for extension. The percentage of genes with an estimated incomplete mRNA 5′coding sequence (2.6%) is in the lower range compared with previous estimates (in the range of 2-5%), which were based on more limited samples of sequences [Carninci et al., 1996; Suzuki et al., 2000; Porcel et al., 2004]. The sensitivity of the method depends on the size of the EST repertoire available. Although EST single-pass sequencing itself is prone to experimental errors, we strongly suggest that the mRNAs for which more than one EST was found, deriving from two independent cDNA libraries and leading to the same prediction, possess a longer CDS than the one described so far.

The identification of the most upstream currently definable AUG start codon in an mRNA sequence cannot itself formally exclude that in some cases a downstream AUG codon may also be used by the ribosome, due to the phenomenon of alternative

translation [Bazykin and Kochetov, 2011]. In addition, due to the availability of a large number of tissue- or stage-specific EST data, the EST-based extended CDS and/or the mRNA with the incomplete ORF could possibly derive from alternative transcription starting sites and/or splicing at the investigated locus. Nevertheless, the protein-coding nature of additional nucleotides at the 5′ of the locus is highlighted, and in the results each distinct alternative RefSeq mRNA isoform mapping to the same locus is associated only with the EST-based extended CDS with which it is compatible.

As a proof-of-concept, we have experimentally confirmed the EST-based models showing an extended coding region at 5′ end for three randomly chosen mRNAs: *GNB2L1* (guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1), *QARS* (glutaminyl-tRNA synthetase) and *TDP2* (tyrosyl-DNA phosphodiesterase 2) (Table 3.1). In these three cases, cross-species comparison at amino acid level indicated a very high grade of conservation of the extended sequence among primates (Fig. 4.4). Therefore, the predicted product for these three human genes should be redefined for functional studies.

GNB2L1 (located on 5q35.3), also known as *RACK1* (Receptor for activated C-kinase 1), is an ubiquitously expressed gene encoding a protein, homologous to the G protein ß subunit, which can coordinate the interaction of a variety of key signaling molecules. It is believed to play a central role in many biological processes (cell growth, translation, apoptosis, migration, cell cycle, cell division) [McCahill et al., 2002]. Interestingly, while in interaction studies this protein was retrieved as a prey starting from many other proteins as a bait (interactions listed in the Entrez Gene entry at: http://www.ncbi.nlm.nih.gov/gene/10399), to date no study appears to have been designed using GNB2L1 as a bait. Such a study could have the advantage of expressing a more complete product for GNB2L1, including 78 amino acids at its amino terminus (Fig. 4.4A), which could reveal additional interactions compared with the product encoded by the currently known cDNA. For example, when Fomenkov et al. [Fomenkov et al., 2004] reported the extension of the interacting *GNB2L1* cDNA CDS, it appeared not to include the region described here, and the interaction was localised to the C-terminal region of GNB2L1 protein. Notably, our analysis also identifies an in-frame stop codon upstream of the newly determined *GNB2L1* start codon. This would suggest that the extended coding region at 5´ for this mRNA is now complete, since the use of possibly existent further start codons would hesitate in translation stopping upstream of the translation start in the correct frame.

The *QARS* gene maps on 3p21.31 and encodes an aminoacyl-tRNA synthetase.

Functional data are summarised at the Entrez Gene entry:
http://www.ncbi.nlm.nih.gov/gene/5859. The *QARS* cDNA was used in at least one study of human protein-protein pair wise interaction [Rual et al., 2005] and it was derived from human ORFeome v1.1 database [Rual et al., 2004], which, in its current 5.1 version (http://horfdb.dfci.harvard.edu/), still lacks codons we have determined here by cDNA sequencing. In fact, in the ORFeome, the cDNA clones from which the *QARS* ORF was deduced reported only few bases upstream of the putative start codon, thus hampering the individuation of further in-frame upstream codons, which were instead identified by our EST-based analysis. Although, in this case, the extra amino acids are few (Fig. 4.4B), making unlikely significant changes in the interaction study, this finding stresses the need for gene annotation refinements. As in the case of *GNB2L1*, the presence of an in-frame stop codon upstream of the newly determined *QARS* start codon suggests that the extended coding region at $5'$ for this mRNA is now complete (Fig. 4.3.

The *TDP2* gene, located on 6p22.3-p22.1, encodes a member of a superfamily of divalent cation-dependent phosphodiesterases and also has several interactions described summarised at the Entrez Gene entry:
http://www.ncbi.nlm.nih.gov/gene/51567, where its cDNA was not used as a bait. Its known CDS appears to lack 30 conserved amino acids corresponding to the protein amino terminus (Fig. 4.4C).

5.2 Codon bias

Codon bias is a well-known phenomenon, observed in species from bacteria to mammals. Preferred codons can differ dramatically between species and also within a genome. The direct application of this phenomenon is usually the optimisation of the heterolog expression of a protein exploiting the codon bias of the guest. It has been demonstrated that the use of particular codons can increase the expression of a transgene by over 1,000-fold [Gustafsson et al., 2004].

However, codon bias is often studied among few genes, and always at the genome level. Here we have presented a computational system capable of studying codon bias in a new way. We developed software useful for studying codon bias at mRNA level, which counts each time that a given codon is represented in the transcriptome, thus accounting for the abundance of each mRNA bearing that codon (Fig. 3.2).

We refer to the total number of codons (n) present across all the mRNAs pool, each expressed at a certain level (x) in a given biological condition as its codonome

value (cv = $\sum(n \times x)$ for the mRNAs pool). This is an entirely new concept in genomics, which allows us to determine the consistence of the actual pool of codons physically existent in the mRNA space of a cell, rather than the codon frequency at the level of the gene sequence. The innovative "CODONOME" software is able to calculate these parameters, offering the possibility to test whether there are limits that constrain representation of a codon in the whole transcriptome given its frequency at the genome level (codon bias). We used as reference data input gene expression profiles calculated by integration and normalisation of different datasets for a given tissue, following the demonstration that this approach gives a more accurate representation of a reference transcriptome as compared to the use of platform- or experimental-skewed datasets [Lenzi et al., 2011]. As expected, the normalised expression profiles show that the genes with the highest expression values are housekeeping genes (see Table 4.2 and Table 4.3). In addition, the human circulating blood erythrocytes expression profile highlights the preponderance of the most frequently expressed hemoglobin subunits. These findings emphasise the consistence of the reference gene expression profiles we have calculated with the known biology of the considered tissues.

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

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

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

Moreover, this high correlation level is maintained across multiple species, from bacteria to humans. This finding clearly implies that the proportional representation of each codon in the DNA and mRNA pool is a general law of nature. It

is reasonable to hypothesise that this correlation, resulting from the interaction of the gene number, the skewing of genome codon bias for each gene, and the allowed gene expression value range, allows for a maximal optimisation of the transcription and translation processes. Indeed, replacement of actual expression values by random numbers in different ranges shows that the universal law of correlation between codon bias and codonome at a genome scale is not limited to the real gene subsets expressed in nature, but emerges as a general property of the distribution and range of the number, sequence, and expression level of the genes included in a genome. This also implies the important conclusion that there is no constraint, in terms of codon bias, for the global distribution of gene expression values during transcription of a genome.

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

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

5.3 ASD implicated genes

Autism spectrum disorder (ASD) is a highly heritable and complex neurodevelopmental condition that results in behavioral, social and communication impairments. The genetic complexity of this disorder is underlined by the fact that despite extensive efforts to elucidate the causes of ASD uncovering hundreds of susceptibility loci and candidate genes, only a few of these markers represent clear targets for further analyses.

Our analyses were aimed to find a common network in which ASD candidate genes act in tune leading to the disorder development. Recent publications have shown that endogenous transcripts, both coding and non coding, sharing microRNA

response elements (MREs) for the same microRNA can influence the expression level of each other through competitive microRNA binding. As a long non coding RNA, *MSNP1AS*, been recently identified associated with ASD and the increased ASD risk genotype, has four MREs for a microRNA, we think that it could regulate the expression of ASD implicated genes by competing with them for the microRNA binding. Our prediction shows indeed that 70.9% of ASD candidate genes, out of all genes identified to be implicated with ASD from the literature, share at least one MRE for the same microRNA with *MSNP1AS*. Our computational analyses confirmed that only ASD implicated genes sharing at least one MREs with *MSNP1AS* for the same microRNA are more co-expressed in the prefrontal cortex than expected by chance, in both healty and ASD affected adults. ASD implicated genes are as well co-expressed in both healthy children and healthy adults, but this association was not found in analysed fetal samples, confirming gene expression changes during the normal brain development [Colantuoni *et al*., 2011] and suggesting a possible role of ASD implicated genes regulated by *MSNP1AS* during the pathology development. ASD implicated genes without any MRE for the microRNA never show significant p values, which indicates that these genes are not involved in the network with *MSNP1AS*.

Considering the ASD risk genotype, if *MSNP1AS* is highly expressed, competing for the microRNA binding with the ASD implicated genes, it would lead to a over expression of the ASD implicated genes because less microRNA molecules are available to bind (and thus to regulate) them. This is not what we observed, because we are not able to find differences in ASD implicated genes' expression in the ASD increased risk genotype.

The proportion of the population with ASD that has been sampled to date is limited and not accurate. Further analyses should be conducted with a greater number of samples with known genotype and of a better RNA quality, grouped by sex and age and from different brain regions. Furthermore *in vitro* and *in vivo* experiments are needed to study the over and the under expression of *MSNP1AS* and their consequences in the abundance of the microRNA and of the ASD implicated genes, in order to test the hypothesis of a network in wich *MSNP1AS* regulates ASD implicated genes.

Conclusions

In conclusion, in this study we have presented three different computational biology approaches useful in order to study the structure, as in the case of mRNA $5'$ coding region sequence, and expression, as in case of the codonome bias and the ASD implicated genes, of human genes.

Our first approach has been able to generate, on a genome scale, 477 EST-driven original extended CDSs of human mRNAs, which are now available to researchers interested in these loci. In addition, software users can access a list of 20,775 human mRNAs in which the presence of an in-frame stop codon upstream of the known start codon indicates completeness of the coding sequence at $5'$ in the current form.

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

Lastly, the third approach is useful to study the expression of human genes. We have found the existence of a network in which autism spectrum disorder implicated genes sharing microRNA response elements for the same microRNA are co-expressed. We think that changes in the expression of a recently identified long non coding RNA which have four microRNA response elements for the same microRNA can regulate the expression of these genes disrupting the equilibrium in this network, but further analyses and experiments are needed.

These three examples showed how the massive amount of publicly available data are of an unestimated valour being still useful to study biological processes, sometimes even different from the purpose for which they have been created.

References

[1] Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, and Moreno RF. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252:1651–1656, 1991.

[2] Alberts B, Bray D, Lewis J, Raff M, Roberts K, and Watson JD. *Molecular Biology of the Cell*. New York: Garland Science, third edition, 1994.

[3] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Research*, 25(17):3389–3402, 1997.

[4] Anderson SF, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, and Young IG. Sequence and organization of the human mitochondrial genome. *Nature*, 290(5806):457–465, 1981.

[5] Anney R, Klei L, Pinto D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, Sykes N, Pagnamenta AT, Almeida J, Bacchelli E, Bailey AJ, Baird G, Battaglia A, and *et al*. A genome-wide scan for common alleles affecting risk for autism. *Hum Mol Genet*, 19(20):4072–4082, 2010.

[6] Asikainen S, Storvik M, Lakso M, and Wong G. Whole genome microarray analysis of *C. elegans* rrf-3 and eri-1 mutants. *FEBS Letters*, 581(26):5050–5054, 2007.

[7] Attwood TK, Gisel A, Eriksson N-E, and Bongcam-Rudloff E. *Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective*. Bioinformatics - Trends and Methodologies, Dr. Mahmood A. Mahdavi (Ed.), InTech, 2011.

[8] Auer H, NewsomDL, and Kornacker K. Expression profiling using Affymetrix GeneChip microarrays. *Methods Molecular Biology*, 509:35–46, 2009.

[9] Barrett T and Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods in Enzymology*, 411:352–369, 2006.

[10] Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, and Edgar R. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acid Research*, 37(Database issue):D885–D890, 2009.

[11] Bazykin GA and Kochetov AV. Alternative translation start sites are conserved in eukaryotic genomes. *Nucleic Acids Research*, 39(2):567–577, 2011.

[12] Bennetzen JL and Hall BD. Codon selection in yeast. *The Journal of Biological Chemistry*, 257(6):3026–3031, 1982.

[13] Benson D, Boguski M, Lipman D, and Ostell J. The National Center for Biotechnology Information. *Genomics*, 6(2):389–391, 1990.

[14] Betancur C. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain Reseach*, 1380:42–77, 2011.

[15] Boguski MS, Lowe TM, and Tolstoshev CM. dbEST–database for "expressed sequence tags". *Nature Genetics*, 4(4):332–333, 1993.

[16] Bourquin JP, Subramanian A, Langebrake C, Reinhardt D, Bernard O, Ballerini P, Baruchel A, Cavé H, Dastugue N, Hasle H, Kaspers GL, Lessard M, Michaux L, Vyas P, and van Wering E *et al*. Identification of distinct molecular phenotypes in acute megakaryoblastic leukemia by gene expression profiling. *Proceedings of the National Academy of Science of the United States of America*, 103(9):3339–3344, 2006.

[17] Brown H, Sanger F, and Kitai R. The structure of pig and sheep insulins. *The Biochemical Journal*, 60(4):556–565, 1955.

[18] Brown KM, Landry CR, Hartl DL, and Cavalieri D. Cascading transcriptional effects of a naturally occurring frameshift mutation in *Saccharomyces cerevisiae*. *Mol Ecology*, 17(12):2985–2997, 2008.

[19] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, and Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics*, 10:421, 2009.

[20] Cameron DA, Gentile KL, Middleton FA, and Yurco P. Gene expression profiles of intact and regenerating zebrafish retina. *Molecular Vision*, 11:775–791, 2005.

[21] Carninci P, Kvam C, Kitamura A, Ohsumi T, Okazaki Y, Itoh M, Kamiya M, Shibata K, Sasaki N, Izawa M, Muramatsu M, Hayashizaki Y, and Schneider C. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics*, 37(3):327–336, 1996.

[22] Casadei R, Strippoli P, D'Addabbo P, Canaider S, Lenzi L, Vitale L, Giannone S, Frabetti F, Facchin F, Carinci P, and Zannotti M. mRNA $5'$ region sequence incompleteness: a potential source of systematic errors in translation initiation codon assignment in human mRNAs. *Gene*, 321:185–193, 2003.

[23] Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, Chinappi M, Tramontano A, and Bozzoni I. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell*, 147(2):358–369, 2011.

[24] Colantuoni C, Lipska BK, Ye T, Hyde TM, Tao R, Leek JT, Colantuoni EA, Elkahloun AG, Herman MM, Weinberger DR, and Kleinman JE. Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature*, 478(7370):519–523, 2011.

[25] Davis LG, Kuehl WM, and Battey JF. *Basic methods in molecular Biology*. Appleton & Lange, Norwalk, second edition, 1994.

[26] Dayhoff, MO, Eck RV, Chang MA, and Sochard MR. *Atlas of protein sequence and structure*. National Biomedical Research Foundation, Silver Spring, Maryland, USA, 1965.

[27] Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J, Dike S, Wyss C, Henrichsen CN, Holroyd N, Dickson MC, and et al. Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Research*, 17(6):746–759, 2007.

[28] Detwiller KY, Fernando NT, Segal NH, Ryeom SW, D'Amore PA, and Yoon SS. Analysis of hypoxia-related gene expression in sarcomas and effect of hypoxia on RNA interference of vascular endothelial cell growth factor A. *Cancer Research*, 65(13):5881–5889, 2005.

[29] Dong T, Kirchhof MG, and Schellhorn HE. RpoS regulation of gene expression during exponential growth of *Escherichia coli* K12. *Molecular Genetics Genomics*, 279(3):267–277, 2008.

[30] Drew RE, Rodnick KJ, Settles M, Wacyk J, Churchill E, Powell MS, Hardy RW, Murdoch GK, Hill RA, and Robison BD. Effect of starvation on transcriptomes of brain and liver in adult female zebrafish (*Danio rerio*). *Physiological Genomics*, 35(3):283–295, 2008.

[31] Eisenberg E and Levanon EY. Human housekeeping genes are compact. *Trends in Genetics*, 19(7):362–365, 2003.

[32] Engels WR. Contributing software to the internet: the Amplify program. *Trends in Biochemical Sciences*, 18(11):448–450, 1993.

[33] Falk MJ, Zhang Z, Rosenjack JR, Nissim I, Daikhin E, Nissim I, Sedensky MM, Yudkoff M, and Morgan PG. Metabolic pathway profiling of mitochondrial respiratory chain mutants in *C. elegans*. *Molecular Genetics and Metabulism*, 93(4):388–397, 2008.

[34] Fatica A and Bozzoni I. Long non-coding RNAs: new players in cell differentiation and development. *Nature Reviews Genetics*, 15(1):7–21, 2014.

[35] Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, and et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* rd. *Science*, 269(5223):496–512, 1995.

[36] Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, and Kähäri A *et al*. Ensembl 2011. *Nucleic Acids Research*, 39(Database issue):D800–D806, 2011.

[37] Fomenkov A, Zangen R, Huang YP, Osada M, Guo Z, Fomenkov T, Trink B, Sidransky D, and Ratovitski EA. RACK1 and stratifin target DeltaNp63alpha for a proteasome degradation in head and neck squamous cell carcinoma cells upon DNA damage. *Cell Cycle*, 3(10):1285–1295, 2004.

[38] Frabetti F, Casadei R, Lenzi L, Canaider S, Vitale L, Facchin F, Carinci P, Zannotti M, and Strippoli P. Systematic analysis of mRNA $5'$ coding sequence incompleteness in danio rerio: an automated EST-based approach. *Biology Direct*, 2:34–54, 2007.

[39] Friedman RC, Farh KK, Burge CB, and Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105, 2009.

[40] Ge X, Yamamoto S, Tsutsumi S, Midorikawa Y, Ihara S, Wang SM, and Aburatani H. Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics*, 86(2):127–141, 2005.

[41] Gilchrist MA, Shah P, and Zaretzki R. Measuring and detecting molecular adaptation in codon usage against nonsense errors during protein translation. *Genetics*, 183(4):1493–1505, 2009.

[42] Goh SH, Josleyn M, Lee YT, Danner RL, Gherman RB, Cam MC, and Miller JL. The human reticulocyte transcriptome. *Physiological Genomics*, 30(2):172–178, 2007.

[43] Grantham R, Gautier C, Gouy M, Mercier R, and Pavé A. Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*, 8(1):r49–r62, 1980.

[44] Guan Q, Zheng W, Tang S, Liu X, Zinkel RA, Tsui KW, Yandell BS, and Culbertson MR. Impact of nonsense-mediated mRNA decay on the global expression profile of budding yeast. *PLoS Genetics*, 2(11):e203, 2006.

[45] Gustafsson C, Govindarajan S, and Minshull J. Codon bias and heterologous protein expression. *Trends in Biotechnology*, 22(7):346–353, 2004.

[46] Hamm GH and Cameron GN. The EMBL data library. *Biochemical Society transaction*, 14(1):5–9, 1986.

[47] Hershberg R and Petrov DA. Selection on codon bias. *Annual Review of Genetics*, 42:287–299, 2008.

[48] Holm AK, Blank LM, Oldiges M, Schmid A, Solem C, Jensen PR, and Vemuri GN. Metabolic and transcriptional response to cofactor perturbations in *Escherichia coli*. *Journal of Biological Chemistry*, 85(23):17498–17506, 2010.

[49] Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, and *et al*. The Ensembl genome database project. *Nucleic Acid Research*, 30(1):38–41, 2002.

[50] Ikemura T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *Journal of Molecular Biology*, 151(3):389–409, 1981.

[51] Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution*, 2(1):13–34, 1985.

[52] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[53] Kabanova S, Kleinbongard P, Volkmer J, Andrée B, Kelm M, and Jax TW. Gene expression analysis of human red blood cells. *International Journal of Medical Sciences*, 6(4):156–159, 2009.

[54] Kanaya S, Yamada Y, Kudo Y, and Ikemura T. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, 238(1):143–155, 1999.

[55] Kanaya S, Yamada Y, Kinouchi M, Kudo Y, and Ikemura T. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *Journal of Molecular Evolution*, 53(4-5):290–298, 2001.

[56] Kennard O. *From private data to public knowledge*. In *The Impact of Electronic Publishing on the Academic Community*, an International Workshop organised by the Academia Europaea and the Wenner-Gren Foundation, Wenner-Gren Center, Stockholm, 16-20 April, 1997. Ian Butterworth, Ed. Published by Portland Press Ltd., London, UK, 1997.

[57] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D. The human genome browser at UCSC. *Genome Research*, 12(6):996–1006, 2002.

[58] Kerin T, Ramanathan A, Rivas K, Grepo N, Coetzee GA, and Campbell DB. A noncoding RNA antisense to moesin at 5p14.1 in autism. *Science Translational Medicine*, 4(128):128ra40, 2012.

[59] Kitagawa N, Washio T, Kosugi S, Yamashita T, Higashi K, Yanagawa H, Higo K, Satoh K, Ohtomo Y, Sunako T, Murakami K, Matsubara K, Kawai J, Carninci P, Hayashizaki Y, and *et al*. Computational analysis suggests that alternative first exons are involved in tissue-specific transcription in rice (*Oryza sativa*). *Bioinformatics*, 21(9):1758–1763, 2005.

[60] Klauck SM. Genetics of autism spectrum disorder. *European Journal of Human Genetics*, 14(6):714–720, 2006.

[61] Klusmann JH, Godinho FJ, Heitmann K, Maroz A, Koch ML, Reinhardt D, Orkin SH, and Li Z. Developmental stage-specific interplay of GATA1 and IGF signaling in fetal megakaryopoiesis and leukemogenesis. *Genes and Development*, 24(15):1659–1672, 2010.

[62] Klusmann JH, Li Z, Böhmer K, Maroz A, Koch ML, Emmrich S, Godinho FJ, Orkin SH, and Reinhardt D. miR-125b-2 is a potential oncomiR on human chromosome 21 in megakaryoblastic leukemia. *Genes and Development*, 24(5):478–490, 2010.

[63] Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, Hayashizaki Y, and Carninci P.

CAGE: cap analysis of gene expression. *Nature Methods*, 3(3):211–222, 2006.

[64] Kozak M. Pushing the limits of the scanning mechanism for initiation of translation. *Gene*, 299:1–34, 2002.

[65] Krajacic P, Hermanowski J, Lozynska O, Khurana TS, and Lamitina T. *C. elegans* dysferlin homolog fer-1 is expressed in muscle, and fer-1 mutations initiate altered gene expression of muscle enriched genes. *Physiological Genomics*, 40(1):8–14, 2009.

[66] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, and *et al*. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[67] Lefebvre KA, Tilton SC, Bammler TK, Beyer RP, Srinouanprachan S, Stapleton PL, Farin FM, and Gallagher EP. Gene expression profiles in zebrafish brain after acute exposure to domoic acid at symptomatic and asymptomatic doses. *Toxicological Sciences*, 107(1):65–77, 2009.

[68] Lenzi L, Frabetti F, Facchin F, Casadei R Vitale L, Canaider S, Carinci P, Zannotti M, and Strippoli P. Unigene Tabulator: a full parser for the UniGene format. *Bioinformatics*, 22(20):2570–2571, 2006.

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

[70] Li Q and Ownby CL. A rapid method for extraction of DNA from agarose gels using a syringe. *BioTechniques*, 15(6):976–978, 1993.

[71] Lockstone HE, Harris LW, Swatton JE, Wayland MT, Holland AJ, and Bahn S. Gene expression profiling in the adult Down syndrome brain. *Genomics*, 90(6):647–660, 2007.

[72] Marques AC, Tan J, and Ponting CP. Wrangling for microRNAs provokes much crosstalk. *Genome Biology*, 12(11):132, 2011.

[73] McCahill A, Warwicker J, Bolger GB, Houslay MD, and Yarwood SJ. The RACK1 scaffold protein: a dynamic cog in cell response mechanisms. *Molecular Pharmacology*, 62(6):1261–1273, 2002.

[74] Moon K and Gottesman S. A PhoQ/P-regulated small RNA regulates sensitivity of escherichia coli to antimicrobial peptides. *Molecular Microbiology*, 74(6):1314–1330, 2009.

[75] Neale BM1, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V, Polak P, Yoon S, Maguire J, Crawford EL, Campbell NG, and *et al*. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, 485(7397):242–245, 2012.

[76] Noh SJ, Miller SH, Lee YT, Goh SH, Marincola FM, Stroncek DF, Reed C, Wang E, and Miller JL. Let-7 microRNAs are developmentally regulated in circulating human erythroid cells. *Journal of Translational Medicine*, 7:98, 2009.

[77] Nguyen DK and Disteche CM. Dosage compensation of the active X chromosome in mammals. *Nature Genetics*, 38(1):47–53, 2006.

[78] O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD, Turner EH, Stanaway IB, Vernot B, Malig M, Baker C, and *et al*. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 485(7397):246–250, 2012.

[79] Padden M, Leech S, Craig B, Kirk J, Brankin B, and McQuaid S. Differences in expression of junctional adhesion molecule-a and beta-catenin in multiple sclerosis brain tissue: increasing evidence for the role of tight junction pathology. *Acta Neuropathologica*, 113(2):177–186, 2007.

[80] Pearson H. Genetics: what is a gene? *Nature*, 441(7092):398–401, 2006.

[81] Peng M, Falk MJ, Haase VH, King R, Polyak E, Selak M, Yudkoff M, Hancock WW, Meade R, Saiki R, Lunceford AL, Clarke CF, and Gasser DL. Primary coenzyme Q deficiency in Pdss2 mutant mice causes isolated renal disease. *PLoS Genetics*, 4(4):e1000061, 2008.

[82] Plotkin JB and Kudla G. Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics*, 12(1):32–42, 2011.

[83] Plotkin JB, Robins H, and Levine AJ. Tissue-specific codon usage and the expression of human genes. *Proceedings of the National Academy of Science of the United States of America*, 101(34):12588–12591, 2004.

[84] Porcel BM, Delfour O, Castelli V, De Berardinis V, Friedlander L, Cruaud C, Ureta-Vidal A, Scarpelli C, Wincker P, Schachter V, Saurin W, Gyapay G, Salanoubat M, and Weissenbach J. Numerous novel annotations of the human genome sequence supported by a 5'-end-enriched cdna collection. *Genome*, 14:463–471, 2004.

[85] Puigbò P, Bravo IG, and Garcia-Vallvé S. E-CAI: a novel server to estimate an expected value of Codon Adaptation Index (eCAI). *BMC Bioinformatics*, 9:65, 2008.

[86] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.

[87] Ronald A and Hoekstra RA. Autism Spectrum Disorders and Autistic Traits: A Decade of New Twin Studies. *American Journal of Medical Genetics Part B Neuropsychiatric genetics*, 156:255–274, 2011.

[88] Rual JF, Hirozane-Kishikawa T, Hao T, Bertin N, Li S, Dricot A, Li N, Rosenberg J, Lamesch P, Vidalain PO, Clingingsmith TR, Hartley JL, Esposito D, Cheo D, Moore T, and *et al*. Human ORFeome version 1.1: a platform for reverse proteomics. *Genome Research*, 14(10B):2128–2135, 2004.

[89] Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, and *et al*. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

[90] Ryle AP, Sanger F, Smith LF, and Kitai R. The disulphide bonds of insulin. *The Biochemical Journal*, 60(4):541–556, 1955.

[91] Salmena L, Poliseno L, Tay Y, Kats L, and Pandolfi PP. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*, 146(3):353–358, 2011.

[92] Sambrook J and Russell DW. *Rapid amplification of 5′ cDNA Ends*. Molecular Cloning - A Laboratory Manual, Cold Spring Harbor Laboratory Press, New York, 2001.

[93] Sanborn JZ, Benz SC, Craft B, Szeto C, Kober KM, Meyer L, Vaske CJ, Goldman M, Smith KE, Kuhn RM, Karolchik D, Kent WJ, Stuart JM, Haussler D, and Zhu J. The UCSC cancer genomics browser: update 2011. *Nucleic Acids Research*, 39(Database issue):D951– D959, 2011.

[94] Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, Walker MF, Ober GT, Teran NA, Song Y, El-Fishawy P, and *et al*. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, 485(7397):237–241, 2012.

[95] Sanger F. The free amino groups of insulin. *The Biochemical Journal*, 39(5):507–515, 1945.

[96] Sanger F, Nicklen S, and Coulson AR. DNA sequencing with chain-terminating inhibitors. *Biochemistry*, 74:5463–5467, 1977.

[97] Sanger F, Thompson EOP, and Kitai R. The amide groups of insulin. *The Biochemical Journal*, 59(3):509–518, 1955.

[98] Sanger F, Coulson AR, Friedmann T, Air GM, Barrell BG, Brown NL, Fiddes JC, Hutchison CA 3rd, Slocombe PM, and Smith M. The nucleotide sequence of bacteriophage phiX174. *Journal of Molecular Biology*, 125(2):225–246, 1978.

[99] Sanger F, Coulson AR, Hong GF, Hill DF, and Petersen GB. Nucleotide sequence of bacteriophage lambda DNA. *Journal of Molecular Biology*, 162(4):729–773, 1982.

[100] Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, and *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 39(Database issue):D38–D51, 2011.

[101] Sharp PM and Li WH. The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, 15(3):1281–1295, 1987.

[102] Sémon M, Lobry JR, and Duret L. No evidence for tissue-specific adaptation of synonymous codon usage in humans. *Molecular Biology and Evolution*, 23(3):523–529, 2006.

[103] Simons V, Morrissey JP, Latijnhouwers M, Csukai M, Cleaver A, Yarrow C, and Osbourn A. Dual effects of plant steroidal alkaloids on *Saccharomyces cerevisiae*. *Antimicrobial Agents Chemotherapy*, 50(8):2732–2740, 2006.

[104] Smyth DG, Stein WH, and Moore S. The sequence of amino acid residues in bovine pancreatic ribonuclease: Revisions and confirmations. *The Journal of Biological Chemistry*, 238:227–234, 1963.

[105] Smith TF. The history of the genetic sequence databases. *Genomics*, 6(4):701–707, 1990.

[106] Sueoka N and Kawanishi Y. DNA G+C content of the third codon position and codon usage biases of human genes. *Gene*, 261(1):53–62, 2000.

[107] Suzuki Y, Ishihara D, Sasaki M, Nakagawa H, Hata H, Tsunoda T, Watanabe M, Komatsu T, Ota T, Isogai T, Suyama A, and Sugano S. Statistical analysis of the 5' untranslated region of human mrna using "oligo-capped" cdna libraries. *Genomics*, 64(3):286–297, 2000.

[108] Toyama R, Chen X, Jhawar N, Aamar E, Epstein J, Reany N, Alon S, Gothilf Y, Klein DC, and Dawid IB. Transcriptome analysis of the zebrafish pineal gland. *Developmental Dynamics*, 238(7):1813–1826, 2009.

[109] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, and *et al*. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[110] Wang K, Zhang H, Ma D, Bucan M, Glessner JT, Abrahams BS, Salyakina D, Imielinski M, Bradfield JP, Sleiman PM, Kim CE, Hou C, Frackelton E, Chiavacci R, Takahashi N, and *et al*. Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature*, 459(7246):528–533, 2009.

[111] Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, and *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 33(Database issue):D39–D45, 2005.

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

[113] Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, Lancet D, and Shmueli O. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, 21(5):650–659, 2005.

[114] Yoon SS, Segal NH, Park PJ, Detwiller KY, Fernando NT, Ryeom SW, Brennan MF, and Singer S. Angiogenic profile of soft tissue sarcomas based on analysis of circulating factors and microarray gene expression. *Journal Surgical Research*, 135(2):282–290, 2006.

[115] Yudate HT, Suwa M, Irie R, Matsui H, Nishikawa T, Nakamura Y, Yamaguchi D, Peng ZZ, Yamamoto T, Nagai K, Hayashi K, Otsuki T, Sugiyama T, Ota T, Suzuki Y, Sugano S, Isogai T, and Masuho Y. HUNT: launch of a full-length cDNA database from the helix research institute. *Nucleic Acids Research*, 29(1):185–188, 2001.