This dissertation is submitted for the degree of Doctor of Philosophy

Investigating the link between tRNA and mRNA abundance in mammals

Konrad Ludwig Moritz Rudolph

May 2015

St Edmund's College,
University of Cambridge

EMBL-EBI

The source code of the thesis is available at `https://github.com/klmr/thesis`.

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the relevant Degree Committee.

Contents

APPENDIX

List of abbreviations

AAB anticodon abundance bias

cDNA complementary DNA

ChIP chromatin immunoprecipitation

ChIP-seq chromatin immunoprecipitation followed by sequencing

CTCF CCCTC-binding factor

CU codon usage

CUB codon usage bias

DE differentially expressed

DNA deoxyribonucleic acid

FDR false discovery rate

FPKM fragments per kilobase of transcript per million mapped read pairs

GEO Gene Expression Omnibus

GO Gene Ontology

GRCm38 Genome Reference Consortium Mouse Build 38

HIV-1 type-1 human immunodeficiency virus

HTS high-throughput sequencing

LINE long interspersed nuclear element

lncRNA long non-coding RNA

mRNA messenger RNA

NCBIM37 NCBI mouse genome build 37

ncRNA non-coding RNA

PCA principal components analysis

PCR polymerase chain reaction

Pol I DNA-dependent RNA polymerase I

Pol II DNA-dependent RNA polymerase II

Pol III DNA-dependent RNA polymerase III

RAA relative anticodon abundance

RCU relative codon usage

RISC RNA-induced silencing complex

RNA ribonucleic acid

RNA-seq RNA sequencing

RPKM read count per kilobase of transcript per million mapped reads

rRNA ribosomal RNA

SINE short interspersed nuclear element

tAI tRNA adaptation index

TBP TATA binding protein

TF transcription factor

TFIIIB transcription factor III B

TFIIIC transcription factor III C

TPM transcripts per million

tRNA transfer RNA

TSS transcription start site

UTR untranslated region

List of Figures

List of Tables

Summary

The genetic code describes how a sequence of codons on an mRNA is translated into a sequence of amino acids, forming a protein. The genetic code manifests itself in the cell as tRNA molecules, which fall into several classes of anticodon isoacceptors, each decoding a single codon into its corresponding amino acid. In this thesis I discuss the central importance of the codon–anticodon interface to mRNA-to-protein translation, and how its stability is maintained during the life of the cell.

This thesis summarises my research into the control of the abundance of tRNAs by individual tRNA gene expression changes in mammalian organisms. I will show that tRNA gene expression is subject to tight regulation, and that the abundance of tRNA molecules is thus kept highly stable even across vastly different cellular conditions, in marked contrast with the abundance of protein-coding genes, which changes dynamically to drive cell function.

The abundance of tRNA genes defines, to a large extent, the efficiency with which mRNA can be translated into proteins. On the one hand, this serves to explain the need for the observed, stable tRNA abundance. On the other hand, this also raises questions: the change of expression of protein-coding genes means that different, specifically highly expressed protein-coding genes in different cell types will lead to a different codon demand. It could thus be beneficial for the cell to express different sets of tRNAs, trading lower overall efficiency for high efficiency in translating the most important subset of genes. To investigate this, I examine the link between mRNA expression and tRNA abundance in a variety of biological conditions across several mammalian species, establishing that changes in the pool of tRNAs are not correlated with changes in mRNA expression.

Overall, my thesis provides important insight into the interface between transcription and translation, suggesting strongly that the regulation of translation is weaker than that of transcription in mammals.

Acknowledgements

I have been supported by many people while working on this thesis. I would like to thank them for their contributions. In particular, thanks go to John Marioni who made the fatal mistake of accepting to supervise me as his PhD student. He provided keen insight and scientific support for my research and pushed me to exceed my own expectations. He also provided extensive and always useful feedback on the thesis itself. Many thanks to my collaborators of the research presented here, Bianca Schmitt and Claudia Kutter. Both performed excellent work in the generation and curation of the data, and only with their expertise was it possible to develop the research. They are also exceedingly patient for the quirks of computational researchers. The same goes for their group leader, Duncan Odom, who expertly complemented John's tutelage. I would like to extend very warm thanks to Anja Thormann who, more than anyone else, convinced me to take a risk, leave my comfort zone and come to study at the University of Cambridge. She also taught me much about the everyday life in research, and laid my foundations in statistics (which I sorely neglected during my previous studies).

A very special thanks goes to my two flat mates, Maria Xenophontos and Nils Kölling. I believe that I have discussed every single aspect of this thesis extensively with them, and they have helped me catch many embarrassing mistakes early. They also taught me the use of the ggplot2 library. Finally, they also kept kicking me, figuratively, to get to work on my thesis (sometimes more than strictly necessary, Maria). I would like to thank Myrto Kostadima and Remco Loos for teaching me to work with ChIP-seq data, and I'd like to thank Ângela Filimon Gonçalves not only for teaching me much about working with RNA-seq but also for many fruitful discussions which have impacted the analysis in many positive ways. Of my office mates I would like to thank Jean-Baptiste Pettit and Nuno Fonseca for extensive discussions about software tools and productivity. Nuno also helped me circumnavigate many software problems relating to the RNA-seq analysis. Catalina Vallejos provided excellent support for statistical questions. Antonio Scialdone, Jong Kyoung Kim, Luis Saraiva, Tim Hu and Nils Eling helped by discussing many aspects of the analysis. Anestis Touloumis in particular helped me by discussing the design of the codon compensation analysis.

Michael Schubert and Stijn van Dongen provided fruitful discussions about software. In particular, Michael's management of the cluster's software installation and Stijn's many excellent tools saved me a lot of time. Matthew Davis gave me copious feedback on many different parts of the analysis and in particular on the chapter on codon–anticodon adaptation. My thesis advisory committee, consisting of Paul Flicek, Gos Micklem and Detlev Arendt provided guidance for my research and suggested several new avenues of inquiry.

I am grateful to my parents, Gero and Vera Rudolph, as well as my sister, Sophia Rudolph, for nurturing my interest in science from early on. They are always excited to discuss every aspect of my work, and make an effort to really understand my research, which has led to much useful feedback. For moral support during the PhD, I'd like to especially thank the EMBL predoctoral community, which consists of many fine individuals and which provides, in equal measure, scientific feedback, entertainment and psychological counselling.

Finally, I would like to thank Joana Borrego Pinto for many contributions small and large to my research, in particular help with the design of figures, the presentation of data and suggestions about the interpretation of the results. But more than that, for everything else.

Introduction 1

“Nothing in biology makes sense […]”

— *Теодо́сій Григо́рович Добжа́нський (Theodosius Dobzhansky)*[*]

1.1. The central dogma

At the core of every living being is its genetic inheritance. The genetic inheritance describes information that is passed down from parents to their offspring. It contains a blueprint detailing, in essence, how to construct a new individual from a single cell. This genetic inheritance is physically present in the form of deoxyribonucleic acid (*DNA*) in almost every living cell.[†]

As a medium of information storage, DNA is complemented by two other types of molecules in the cell that, respectively, carry out the instructions encoded in the DNA by performing specific biochemical functions, and serve as an intermediary between information storage and execution. The intermediaries, which are called *ribonucleic acid (RNA)*, copy out specific parts of the complete instructions from the DNA and carry them to factories that translate the instructions into highly specialised machines. These machines are called *proteins*. The *central dogma of molecular biology* states that information is thus transmitted from DNA to RNA, and from RNA to proteins, but never from proteins back to RNA or DNA (figure 1.1) [Crick, 1958; Crick, 1970].

[*] Dobzhansky [1973]

[†] And to some extent in non-living particles called *viruses*.

When first published, the central dogma concisely summarised the available evidence at the time. Now, more than half a century later, this still largely holds true.

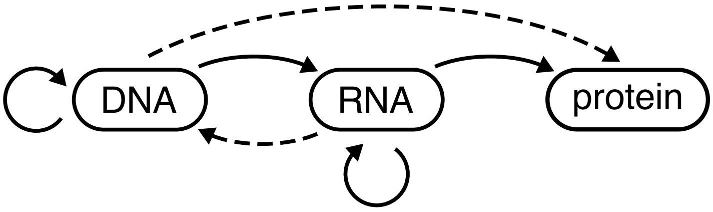
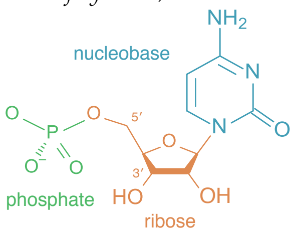


Figure 1.1: The central dogma of molecular biology. The solid arrows represent observed transfers of information; the dashed arrows represent what Crick [1970] referred to as "potential" transfers; today we know that the transfer RNA → DNA does in fact occur under certain circumstances; the transfer DNA → protein has still not been observed. Notably, the *absent* transfers in the original publication are still considered largely non-existent.

Over the years, the very high-level view of the central dogma was complemented by a detailed mechanistic description and the efforts to fill in all the details are still ongoing. In this thesis I will present the results of my exploration of one small aspect concerning the translation of RNA into proteins. To better explain how it fits into the general picture of the central dogma, we first need to understand its leading actors and their interplay. The three main roles in the central dogma are fulfilled by DNA, RNA and proteins, respectively, and we are now going to take a look at all of them in turn.

1.1.1. DNA

nucleotide *molecule consisting of a ribose, one or more phosphate groups and a nucleobase (example: cytidine monophosphate, a nucleotide of cytosine)*‡



nucleoside *nucleobase coupled to a ribose*

DNA consists of a long chain of *nucleotides*. The chemical structure of nucleotides enables them to polymerise into long, relatively stable chains. DNA is made up of nucleotide monomers, consisting of one or more phosphate groups coupled to a *nucleoside*, each of which contains any of four different types of *nucleobases*: adenine (`A`), cytosine (`C`), guanine (`G`) and thymine (`T`). Thus, DNA can be thought of as a long string of four different letters, and

‡Figure adapted from `https://commons.wikimedia.org/w/index.php?title=File:Nucleotides_1.svg&oldid=128814238`

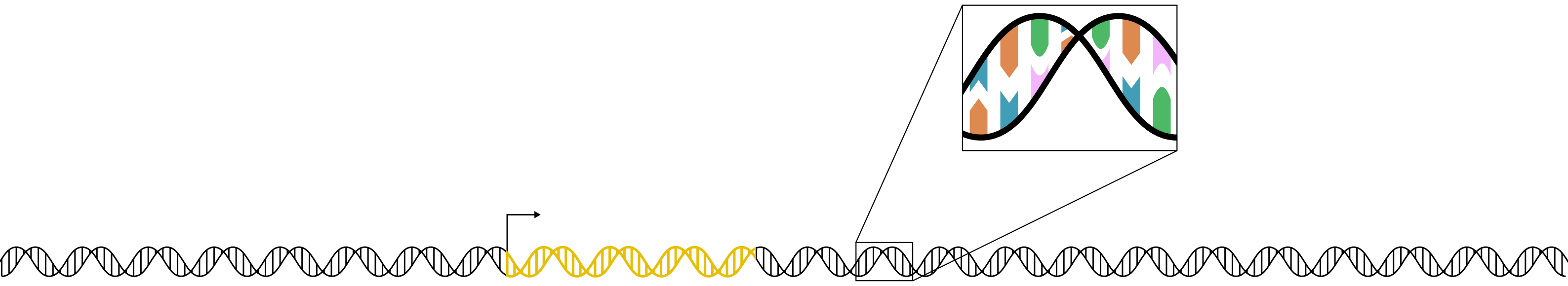


Figure 1.2: DNA double helix with a single, transcribed gene. The intertwined lines form the phosphoribose backbone of the DNA. Each vertical line connecting the backbones corresponds to a base pair. The zoomed in cartoon shows the complementary base pairing, with different colours corresponding to different nucleobases. The length of the gene is understated, the overwhelming majority of transcribed pieces of DNA are much longer (protein-coding genes being many hundred to thousands of base pairs long).

that is indeed how it is often represented. Text is written from left to right in Western cultures. By convention, DNA is written from 5′ to 3′. These numbers refer to the numbering of the carbon atoms in each nucleotide's ribose, with the 3′ carbon atom forming a covalent bond with the phosphate of the next nucleotide, which is itself attached to the 5′ carbon of its ribose. In this way, a 5′ C atom is exposed at one end of the chain, and a 3′ C is exposed at the other.

DNA is present in the cell in the form of double-stranded helices: each DNA molecule consists of two paired chains, wound tightly around each other, with the bases on each chain pairing up such that every A on one chain is paired with a T on the other, and each C is paired with a G. This striking symmetry is known as Watson–Crick base pairing, after its discoverers [Watson and Crick, 1953]. Thus, DNA is made up of two complementary strands, redundantly holding the genetic information (figure 1.2). This redundancy is used in DNA copying (which occurs at every cell division, and is the mechanism by which genetic information is passed from one cell to its offspring) to synthesise two newly formed DNA molecules, each of which contains one strand of the parent DNA molecule (*semiconservative replication*) [Meselson and Stahl, 1958].

DNA is not made up of a single polymer chain, but rather is partitioned into several long pieces, called *chromosomes*. Each chromosome forms a single molecule. However, even on a chromosome the genetic information is not stored in one consecutive piece: Rather, DNA consists of relatively short stretches encoding a specific function, separated by long stretches that do not directly encode any function. The "function" is what is transmitted, as per the central dogma, to RNA and, in many cases, on to proteins. Such self-contained, functional stretches are called *genes*. To perform its function, a gene has to be transcribed into a catalytically active form, the RNA.

gene *self-contained stretch of DNA that is transcribed to perform a function*

1.1.2. RNA

RNA is the product of transcription of a gene from DNA. RNA is chemically similar to DNA but unlike the latter, RNA is created as a single strand. This has two consequences: First, RNA is much less stable than DNA, and slowly degrades. RNA thus has a finite life-time, and the pool of RNA must be replenished by continuous transcription. Second, single-stranded ribonucleic acid spontaneously changes its spatial conformation by forming Watson–Crick base pairs between nucleotides in its own sequence, where this is *sterically* possible (i.e. where forming such a bond does not require bending the chain too much to "snap" it). The resulting structure can confer biochemical functions to the RNA. Because the structure is determined by, and exists on a higher level than the sequence identity of the RNA, it is called *secondary structure*.

steric effect *atoms occupy discrete space and cannot overlap*

Another difference between DNA and RNA is the use of slightly different nucleobases: instead of `T`, RNA uses `U` (uracil), which, like `T`, base-pairs with `A`. Despite the fact that the genetic information is encoded in virtually the same way in DNA and RNA, transcription of DNA into RNA requires a complex machinery. The core of this machinery is a complex enzyme called an *RNA polymerase*. In eukaryotes, three different, evolutionarily related RNA polymerases (*pol I*, *pol II* and *pol III*) are responsible for transcribing different types of RNA.

RNA performs numerous different functions, but one very important subcategory of RNA does not perform any function on its own; rather, it is an intermediary between the genetic information on the DNA and the final protein product, which in turn performs cellular functions. This class of RNA is called messenger RNA (*mRNA*). mRNAs are the product of the transcription of protein-coding genes by pol II. By contrast, RNAs which are not protein-coding are denoted as non-coding RNA (*ncRNA*). Transcription of mRNA requires an exquisite control, and many different transcription factors (*TF*) are known to regulate the activity of transcription of different genes in different cellular conditions.

This results in different mRNA genes being transcribed at highly different levels, leading to several orders of magnitude of difference in mRNA abundance. Even more strikingly, the same mRNA can be transcribed at different levels under different conditions. mRNA is further processed in several steps before the mature mRNA is exported from the cell's *nucleus* into the *cytoplasm*, where it is translated into proteins, and, over time, decays. Taken together, this leads to very differentiated mRNA profiles under different conditions, which imbue cells with a unique phenotype. This forms the basis of cellular differentiation into different cell types and tissues in multicellular eukaryotes.

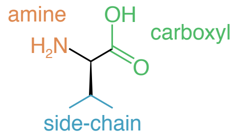
nucleus *centre of eukaryotic cells, hosting the DNA*

cytoplasm *space filling the cell, excluding all enclosed compartments, such as the nucleus*

1.1.3. Proteins

Proteins, finally, are the main effectors of cell function. Like DNA and RNA, they consist of chains of smaller molecules, so-called *amino acids*, that are strung together to form *polypeptides*. Each amino acid is a small molecule with unique properties which, jointly, shape the function of the final protein. Individual amino acids are strung together in a chemical reaction that links a carboxyl group covalently with an amino group on the next amino acid to form a *peptide bond* [Alberts & al., 2002]. Polypeptides, like many RNAs, form secondary structures via non-covalent bonds between amino acids, which are a function of the amino acid sequence. Beyond this, proteins form even higher order three-dimensional conformations called tertiary structures. When multiple proteins aggregate into a complex consisting of several subunits, we speak of quarternary structure.

amino acid *molecule consisting of an amino–carboxyl backbone and a specific side-chain (example: valine)*



peptide bond *a covalent bond formed between a carboxyl and an amino group:* $COOH + NH_2 \longrightarrow CO\text{–}NH + H_2O$

All these different levels of spatial organisation of proteins lead to the creation of highly complex structures from originally one-dimensional chains. It is their intricate structure that allows them to perform precise tasks in the cell. Because they are the work horses of the cells, proteins are highly abundant, with some proteins being present million-fold at any given moment [Milo, 2013]. This is only possible because a single gene is transcribed multiple times, and each resulting mRNA can be translated several times, and simultaneously, before being degraded. The path DNA → RNA → protein thus facilitates an amplification from a single gene copy to many orders of magnitudes more copies of the resulting protein. Despite the fact that multiple protein copies can be created from a single mRNA molecule, and that the number varies from transcript to transcript, protein abundance is predominantly determined by the abundance of mRNAs [J. J. Li & al., 2014; Jovanovic & al., 2015; Csárdi & al., 2014].

1.2. Transcription & translation

1.2.1. Transcription

motif *pattern describing a family of short sequences which, though variable, have some degree of similarity*

As mentioned previously, three different polymerases are responsible for transcribing genes encoded in the DNA into different types of RNA. The precise ways in which the different polymerases transcribe genes into their RNA products differ but the fundamental aspects of transcription are similar. In all cases, a *motif* in the DNA sequence initiates binding of a number of TF proteins to the DNA. Such motifs, called *promoters*, are found in the immediate vicinity of the transcription start site (*TSS*) of their target genes — either upstream of the TSS or following closely after it, inside the gene body. Once the TFs have bound to the DNA on top of the TSS, a polymerase attaches to the DNA and is held in place by the TFs. Subsequently, the polymerase pries the double strand apart and starts synthesising a new strand of RNA which pairs complementarily with one of the strands on the DNA (the *template strand*). The new RNA's sequence is thus identical to the other DNA strand (the *coding strand*). The RNA is produced in the direction 5′–3′, implying that the template strand is read in the direction 3′–5′ during transcription. Once the first few nucleotides of the RNA have been synthesised, the polymerase disassociates from the TF proteins, and the polymerase starts moving along the gene body, transcribing it as it goes (this may require the presence of other TFs called *activators*, which are recruited by *enhancer* motifs elsewhere on the DNA; figure 1.3).

Eukaryotic chromosomes are very long — human chromosome 1 is around 8.5 cm stretched from end to end — and, to fit into the cell, is tightly packed into a space-efficient conformation. To achieve this, DNA is coiled around *histones*, small protein complexes, to form *nucleosomes*. Too tight packing, however, has the side-effect of making the DNA inaccessible to the transcription machinery. It is thus a common feature of gene regulation to control the chromatin structure, and thus to control the accessibility of the DNA for TFs and the polymerases. Furthermore, histones can be marked in several ways — via addition or removal of acetyl or methyl groups — which are recognised by TFs and thus once again either facilitate or inhibit transcription [Alberts & al., 2002]. In addition to enhancers and promoters, chromatin structure and the modification of histones thus regulate the activity of genes.

nucleosome *complex of* DNA *wrapped around an ensemble of eight core histones*

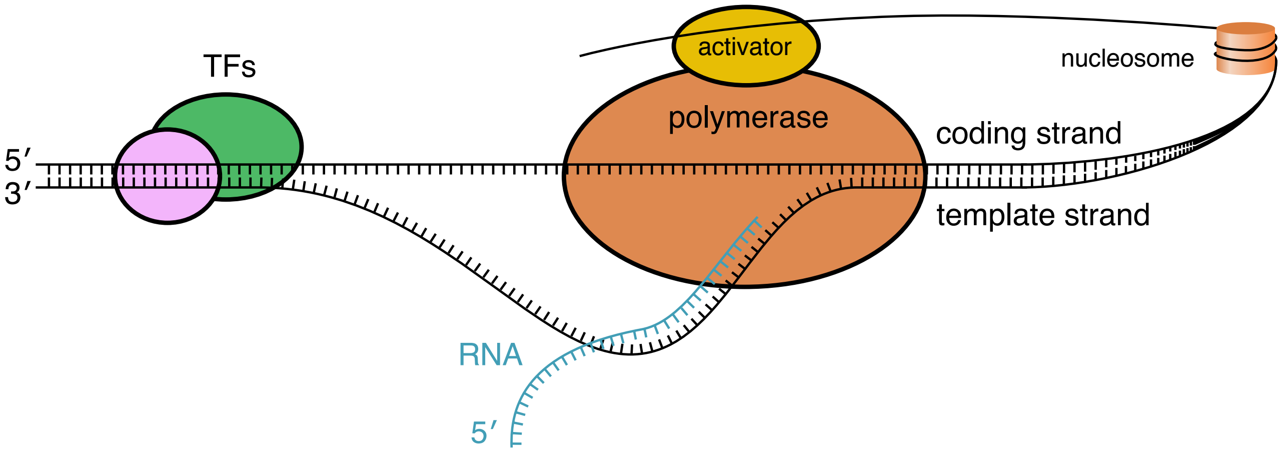


Figure 1.3: Transcription. After recruitment by TFs binding to the promoter region of the gene, a polymerase, aided by a (distally bound) activator protein, starts transcribing the RNA product from 5′ to 3′. Diagram is not to scale.

1.2.2. The genetic code

The process by which proteins are created from mRNA transcripts is more complex than the 1:1 transcription of DNA into RNA, which after all use a common alphabet to encode the information they carry. By contrast, the *translation* of mRNA transcripts into proteins requires a *code* to interpret the genetic information.

code *set of rules for interpreting a piece of information*

There are 20 different amino acids that are encoded by just 4 different nucleotides. To allow this, several nucleotides must be combined to form a larger unit coding for an amino acid. In the universal *genetic code*, shared by all known species, this is accomplished by grouping three consecutive nucleotides together to form non-overlapping, ungapped *triplet codons* along the mRNA. This results in $4^3 = 64$ possible codons, more than three times the number of amino acids. As a consequence, the genetic code is *degenerate*: most amino acids can be encoded by more than a single codon.

Codons furthermore serve as control points by defining where the translated sequence on the mRNA starts and ends. The codon `AUG`, in addition to encoding the amino acid methionine, also marks the start of the coding sequence. Three codons do not encode any amino acid, and instead signal the end of translation (`UAA`, `UAG`, `UGA`). As a consequence, every coding sequence starts with `AUG`, ends with one of the stop codons, and has a length divisible by 3. Table 1.1 contains a tabular representation of the genetic code, which is valid, with only minor variations, for all three domains of life.

During translation, individual codons on the mRNA transcript are successively paired up with their matching amino acid. However, unlike in translation, this pairing does not happen automatically. It requires an intermediate adapter molecule acting as an interface between the codon and the amino acid.

1.2.3. Transfer RNA

Individual codons are translated into their corresponding amino acid with the aid of adapter molecules carrying a specific amino acid, and which recognise the matching codon. This codon recognition is possible because the

C	A	AA	C	A	AA	C	A	AA	C	A	AA
UUU	$\overline{\text{aaa}}$	Phe	UCU	$\overline{\text{aga}}$	Ser	UAU	$\overline{\text{aua}}$	Tyr	UGU	$\overline{\text{aca}}$	Cys
UUC	$\overline{\text{gaa}}$	Phe	UCC	$\overline{\text{gga}}$	Ser	UAC	$\overline{\text{gua}}$	Tyr	UGC	$\overline{\text{gca}}$	Cys
UUA	$\overline{\text{uaa}}$	Leu	UCA	$\overline{\text{uga}}$	Ser	UAA	$\overline{\text{uua}}$	stop	UGA	$\overline{\text{uca}}$	stop
UUG	$\overline{\text{caa}}$	Leu	UCG	$\overline{\text{cga}}$	Ser	UAG	$\overline{\text{cua}}$	stop	UGG	$\overline{\text{cca}}$	Trp
CUU	$\overline{\text{aag}}$	Leu	CCU	$\overline{\text{agg}}$	Pro	CAU	$\overline{\text{aug}}$	His	CGU	$\overline{\text{acg}}$	Arg
CUC	$\overline{\text{gag}}$	Leu	CCC	$\overline{\text{ggg}}$	Pro	CAC	$\overline{\text{gug}}$	His	CGC	$\overline{\text{gcg}}$	Arg
CUA	$\overline{\text{uag}}$	Leu	CCA	$\overline{\text{ugg}}$	Pro	CAA	$\overline{\text{uug}}$	Gln	CGA	$\overline{\text{ucg}}$	Arg
CUG	$\overline{\text{cag}}$	Leu	CCG	$\overline{\text{cgg}}$	Pro	CAG	$\overline{\text{cug}}$	Gln	CGG	$\overline{\text{ccg}}$	Arg
AUU	$\overline{\text{aau}}$	Ile	ACU	$\overline{\text{agu}}$	Thr	AAU	$\overline{\text{auu}}$	Asn	AGU	$\overline{\text{acu}}$	Ser
AUC	$\overline{\text{gau}}$	Ile	ACC	$\overline{\text{ggu}}$	Thr	AAC	$\overline{\text{guu}}$	Asn	AGC	$\overline{\text{gcu}}$	Ser
AUA	$\overline{\text{uau}}$	Ile	ACA	$\overline{\text{ugu}}$	Thr	AAA	$\overline{\text{uuu}}$	Lys	AGA	$\overline{\text{ucu}}$	Arg
AUG	$\overline{\text{cau}}$	Met	ACG	$\overline{\text{cgu}}$	Thr	AAG	$\overline{\text{cuu}}$	Lys	AGG	$\overline{\text{ccu}}$	Arg
GUU	$\overline{\text{aac}}$	Val	GCU	$\overline{\text{agc}}$	Ala	GAU	$\overline{\text{auc}}$	Asp	GGU	$\overline{\text{acc}}$	Gly
GUC	$\overline{\text{gac}}$	Val	GCC	$\overline{\text{ggc}}$	Ala	GAC	$\overline{\text{guc}}$	Asp	GGC	$\overline{\text{gcc}}$	Gly
GUA	$\overline{\text{uac}}$	Val	GCA	$\overline{\text{ugc}}$	Ala	GAA	$\overline{\text{uuc}}$	Glu	GGA	$\overline{\text{ucc}}$	Gly
GUG	$\overline{\text{cac}}$	Val	GCG	$\overline{\text{cgc}}$	Ala	GAG	$\overline{\text{cuc}}$	Glu	GGG	$\overline{\text{ccc}}$	Gly

Table 1.1: The genetic code. Shown is each codon ("C"), its potential corresponding anticodon ("A") and the three-letter abbreviation of the corresponding amino acid ("AA"). AUG, in addition to encoding methionine, also signals the start of translation. Not all anticodons exist in all species. Each anticodon is the reverse complement of its corresponding codon (adapted from dos Reis, Savva, & al. [2004]).

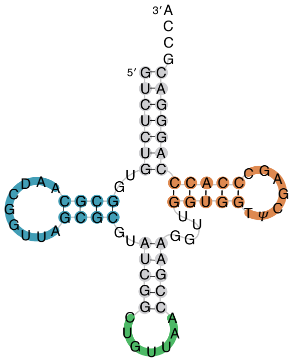


Figure 1.4: tRNA$^{\text{Asn}}$ secondary structure. The tRNA carries the anticodon $\overline{\text{guu}}$ in the middle of the anticodon loop, highlighted in green. The D loop and T loop are shown in blue and orange. Structure predicted by COVE [Eddy and Durbin, 1994], rendered by PseudoViewer 3 [Byun and Han, 2009] and manually edited.

adapter molecules are themselves RNAs, and the codon is matched via complementary base pairing of a part of the RNA sequence termed the *anticodon*. These adapter molecules are called transfer RNA (*tRNA*).

tRNAs form secondary structures with a shape that vaguely resembles a cloverleaf, consisting of a stem and three loops: the *D loop* (also known as *DHU loop* because it contains the modified nucleobase dihydrouridine), the *T loop* (or *TΨC loop*, because it contains the nucleobase pseudouridine) and the *anticodon loop*. The latter carries three nucleotides in its centre that pair with a specific codon — the anticodon [S.-H. Kim & al., 1973; Suddath & al., 1974; Robertus & al., 1974; Rich and S.-H. Kim, 1978; Schimmel and Söll, 1979].

To avoid confusion between codons and anticodons, codons in this thesis will always be typeset as 5′-CAU-3′, and anticodons (which are the reverse complements of codons) will always be typeset as 5′-$\overline{\text{aug}}$-3′. The directionality is indicated here to clarify the direction in which a codon pairs with its anticodon; it will generally be omitted in the remainder of the text. Figure 1.5 illustrates how the mRNA–tRNA interaction leads to the reverse complementarity of the codon and anticodon.

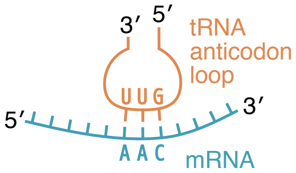


Figure 1.5: mRNA–tRNA interaction, illustrating how the 5′ end of the codon pairs with the 3′ end of the anticodon.

tRNAs are encoded by small genes, around 70 bp to 90 bp in length. Their transcription is driven by pol III. Transcription initiation for pol III can take various forms for different types of genes. In the case of tRNA genes, a so-called *class II* pol III transcription initiation is performed. Unlike protein-coding genes, whose promoters lie upstream of the actual gene body, the class II promoter is found *inside* a tRNA gene in two disjoint, strongly conserved regions called the *A box* and *B box*, respectively. The A box starts about 10 bp downstream from the TSS, whereas the B box can be found at a variable distance of about 30 bp to 60 bp downstream from the A box. tRNA transcription is initiated when the transcription factor TFIIIC binds to both motifs. This leads to the binding of the pol III recruitment factor TFIIIB immediately upstream of the tRNA gene. Subunits of TFIIIB, in particular the TATA binding protein (*TBP*), bind to upstream motifs of the tRNA, which vary strongly across evolution, but whose presence is nevertheless crucial for the initiation of transcription [Palida & al., 1993; R J White and Jackson, 1992]. Binding of TFIIIB in turn leads to the association of pol III with the gene body, and the initiation of transcription. TFIIIC is now no longer required and disassociates from the gene locus. TFIIIB remains bound and can lead to repeated transcription re-initialisation. Transcription stops when pol III encounters a short T repeat downstream of the tRNA gene (see figure 1.6)

[Robert J White, 1998; Dieci & al., 2007].

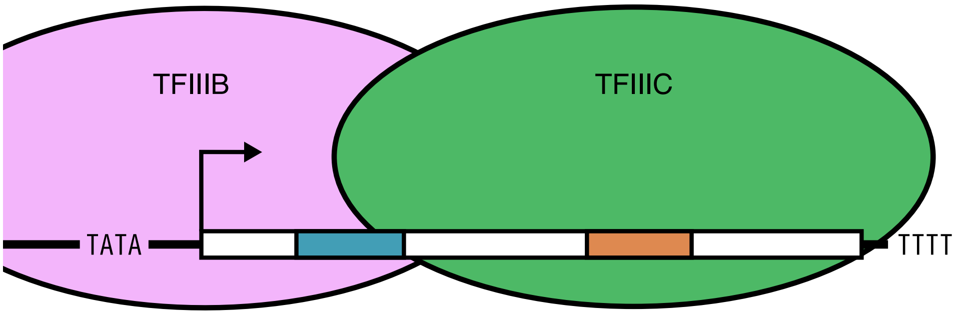


Figure 1.6: **tRNA gene transcription** immediately before pol III is recruited. The A and B box are highlighted in blue and orange, respectively. The colours show the correspondence of these regions with the loops shown in figure 1.4. The TATA motif is a non-representative example of upstream motifs which are recognised by TFIIIB. Diagram is not to scale.

tRNA genes have multiple copies across the genome. In the latest reference genome of *Mus musculus* (GRCm38 [Church & al., 2009]), 432 different tRNA genes are annotated, encoding 50 different anticodons. tRNA genes carrying the same anticodon form an *anticodon isoacceptor family*. tRNA genes for the same amino acid form an *amino acid isotype*.

The numbers of tRNA genes and anticodon isoacceptor families mentioned above exclude tRNAs for the amino acid selenocysteine, which is not part of the standard genetic code, and which was consequently excluded from the subsequent analysis. Selenocysteine is generally not included in analyses with a focus on codons in the published literature, as it requires *translational recoding*, an altogether different translation mechanism, not covered by the genetic code. Furthermore, the prevalence of proteins incorporating selenocysteine is very small [Reeves and Hoffmann, 2009]. Although this does not mean that selenocysteine is biologically irrelevant, it means that we can safely ignore its effect for whole-genome studies of codon and anticodon abundance.

After transcription, the newly formed precursor tRNA transcript undergoes processing to form a mature tRNA. As for all types of RNA, this happens while the precursor tRNA is still in the nucleus of the cell, before it is then exported into the cytoplasm where it performs its function.

The postprocessing of the tRNA is required to ensure that the tRNA folds into its correct spatial structure, can be associated with an amino acid, and recognises its target codons. It also serves as a quality control mechanism: in the case where transcription introduces errors into the tRNA sequence, no postprocessing can occur, which prevents the subsequent export of the tRNA from the nucleus. This includes splicing out introns from the tRNA gene (although this occurs rarely, as most tRNAs have no introns) and cleavage of 5′ upstream sequence elements [Alberts *&* al., 2002; Berg *&* al., 2002].

Another important processing step is the substitution of the 3′-most bases by 5′-CCA-3′, which will subsequently serve as an anchor for the attachment of an amino acid. Furthermore, several nucleotides of the tRNA are replaced by "unusual" nucleosides. Some of these are shown in figure 1.4. In particular, dihydrouridine (D) replaces several uridine[§] nucleosides in the D loop, and pseudouridine (Ψ) replaces a uridine in the T loop. In total, about 10 per cent of all nucleosides are modified in this manner, and over 70 different types of base modifications are known to occur in tRNA [Limbach *&* al., 1994; Dalluge *&* al., 1997; Alberts *&* al., 2002]. Crucially, the 5′ base of the anticodon also undergoes modification in this manner, and this plays an important part in *wobble base pairing*.

Wobble base pairing

As table 1.1 shows, there are 61 different codons. However, not all of these have corresponding anticodon tRNAs — as mentioned, there are only 50 different anticodon isoacceptors in *M. musculus*; for example, CUC codes for leucine, yet there is no matching $\overline{\texttt{gag}}$ anticodon tRNA. Instead, a CUC codon can pair with a tRNA carrying an $\overline{\texttt{aag}}$ anticodon. The mismatching 3′ base of the codon is known as the *wobble position* due to its ability to "wobble" around during codon recognition, and thus form hydrogen bonds that would not be sterically possible under normal conditions. Unlike the first two bases of the codon, the third base thus does not require a strict Watson–Crick match.

However, even under these relaxed steric constraints, A does not pair with

[§] the nucleoside of uracil

C. In order for the $\overline{\text{aag}}$ isoacceptor tRNAs to recognise the CUC codon, it therefore has to undergo a base modification of its wobble base. Indeed this is what happens, with the adenosine (the nucleoside of adenine) at the 5′ position being replaced by inosine (the nucleoside of hypoxanthine, short I). I is able to pair with A, C and U when it is in the wobble position, and the modified $\overline{\text{iag}}$ is thus able to decode CUC [Crick, 1966; Murphy and Ramakrishnan, 2004]. Despite the existence and importance of these base modifications to the anticodon, I will continue using the genomically encoded anticodon notation rather than the anticodon after base modification — in other words, I will generally write $\overline{\text{aag}}$, not $\overline{\text{iag}}$, following general convention. Table 1.2 lists the possible wobble base pairings.

5′ anticodon base	3′ codon base
C	G
G	C, U
U	A
I	A, C, U

Table 1.2: **Simplified wobble base pairing rules.** These are the rules applying to eukaryotes; prokaryotes have slightly different, more permissive pairing rules due to slight differences in the structure of the translation apparatus. In practice, more pairings are possible (though most are uncommon), and there exist several other modified bases with unique pairing properties [Murphy and Ramakrishnan, 2004]. Table based on Alberts & al. [2002].

Amino acid activation

Once mature, the tRNA is exported from the nucleus into the cytosol to aid in mRNA translation. We have seen how a tRNA recognises a specific codon via interaction with its anticodon. This still leaves the question of how the tRNA interacts with its target amino acid. In fact, so far the tRNA is "empty" — not bound to any amino acid. It needs to be "charged" with an amino acid before it can act as an adapter. Conversely, we can say that amino acids need to be *activated* by being coupled to a transfer molecule, which makes them suitable to be used in protein synthesis. This activation is handled by the protein *aminoacyl-tRNA synthetase*.

Aminoacyl-tRNA synthetases are enzymes that take a target amino acid and covalently attach it to the 3′ end of a matching tRNA. This implies that aminoacyl synthetases need to be able to recognise the correct binding partners. In most species, there are 20 different aminoacyl-tRNA synthetases — one per amino acid. Each of them is responsible for charging all tRNAs for a given amino acid isotype. The aminoacyl-tRNA synthetase recognises a matching tRNA by probing for several sequence features, including the acceptor stem, the anticodon and a specific base immediately adjacent to the 3′ end of the tRNA. Once it has associated with an amino acid and its matching tRNA, it catalyses the binding of the amino acid to the 3′ adenosine of the tRNA, to form an aminoacyl-tRNA [Schimmel and Söll, 1979; Schimmel, Giegé, & al., 1993; Ibba and Söll, 2000; Alberts & al., 2002].

Thus loaded with an amino acid, the aminoacyl-tRNA is now free to participate in the translation reaction, wherein it will bind to a matching codon and give up its attached amino acid. This will leave it once more empty but otherwise intact, so that it can be recharged immediately by another aminoacyl-tRNA synthetase. tRNA molecules are thus repeatedly reused until the molecular structure degrades, and gets recycled by the cell.

1.2.4. Translation

So far we have established that tRNAs are responsible for decoding individual codons on the mRNA transcript into individual amino acids, which are assembled into a protein. Unsurprisingly, this is not a spontaneous process in the cell. Precise coordination is required to initiate, maintain and terminate the process. All these parts are controlled by an intricate machinery, the *ribosome*. Ribosomes are large complexes of proteins and ribosomal RNA (*rRNA*), forming two subunits. In eukaryotes, these subunits are the 40S and 60S subunit, respectively. At any given moment, millions of ribosomes are present in a cell, which is necessary to to meet the demand of protein production from a limited pool of mRNA transcripts.

During translation initiation, the small ribosomal subunit binds an *initiator tRNA*^{Met} (tRNA$_\mathrm{i}^\mathrm{Met}$) in its active site. tRNA$_\mathrm{i}^\mathrm{Met}$ differs from conventional tRNA$^{\text{Met}}$ in its nucleotide composition, which enables it to bind to the small subunit of the ribosome unaided [Kolitz and Lorsch, 2010]. The complex then associates with the 5′ end of an mRNA transcript and scans along its 5′ untranslated region (*UTR*) until it finds a signal sequence surrounding a start codon, with which the bound tRNA$_{\text{i}}^{\text{Met}}$'s anticodon base-pairs [Kozak, 2002].

The 40S subunit of the ribosome is now joined by the 60S subunit. Next, the assembled ribosome starts pulling the mRNA transcript through a channel in its structure. When the next codon aligns with a particular structure within the ribosome's active site, called the *A site*, progress stalls until an aminoacyl-tRNA with a matching anticodon finds its way into the site and is able to pair with the presented codon. Next, a new peptide bond is formed between the amino end of the newly arrived amino acid and the carboxyl end of the already synthesised polypeptide. This is accompanied by a conformational change in the ribosome, which pushes the tRNA from the A site into the *P site*. Next, another conformational change moves the mRNA transcript by three bases, so that the next codon is aligned with the once again empty A site, and the process can repeat.

During the first conformational change, the tRNA that previously occupied the P site is displaced into the *E site*, from where it exits the ribosomal complex, to be either recharged by another aminoacyl-tRNA synthetase, or to be disposed of. This process continues until a stop codon is encountered. Stop codons are not recognised by tRNAs but rather by special proteins, whose binding to the stop codon triggers the termination of the protein synthesis, the release of the peptide chain, and the dissociation of the ribosomal complex.

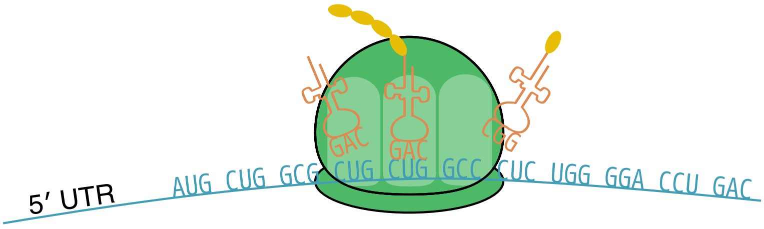


Figure 1.7: Protein translation. The different parts are: 1. the ribosome (green), the start of the mRNA (blue) with the 5′ untranslated region ahead of the start codon **AUG**, the tRNAs (orange) and the amino acids (yellow), forming a nascent polypeptide chain. The three light shaded areas are the A, P and E site, from right to left.

1.3. RNA sequencing quantifies protein-coding gene expression

1.3.1. The transcriptome reflects the state of the cell

As we have seen, the central part of the cellular machinery is abstracted by the central dogma of molecular biology, with the DNA at one end encoding the hereditary identity of the cell, and the proteins at the other end as the effectors of this information.

It is therefore proteins that determine how a cell behaves, and changes in proteins abundance ultimately determine changes in cellular function. The entirety of protein abundance in the cell at a given instance is called the *proteome*. Unfortunately, quantifying the proteome in an unbiased fashion is hard and expensive [Graumann & al., 2008]. Instead, modern biology often uses the abundance of mRNA molecules, coding for individual proteins (the *transcriptome*), as an accurate proxy of the proteome. The appropriateness of this approach has been verified in numerous studies [Nagaraj & al., 2011; Nookaew & al., 2012].

transcriptome *the entirety of the RNA molecules present in the cell at a given time*

1.3.2. Microarrays

Until recently, the abundance of mRNA has been mostly determined using specific probes, which hybridise to complementary target mRNA sequences.

Probes are tagged — for instance with a *fluorophore* — such that the presence of a probe can be detected visually. In order to determine the abundance of many different mRNAs simultaneously, large arrays of different probes can be generated and queried in parallel. Due to their miniaturisation, these arrays are known as *expression microarrays* [Schena & al., 1995].

fluorophore *small chemical that emits light after excitation by a specific wavelength*

While microarrays still play an important role in transcriptome analysis, they have recently been superseded by another technology due, mainly, to the following disadvantages [Casneuf & al., 2007; Marioni & al., 2008]: 1. Each probe is sequence specific, and only recognises its target mRNA. As a consequence, quantification is inherently biased and requires that each target is known beforehand. Microarrays thus cannot discover new target transcripts, and desiging a new microarray is technically challenging and expensive. 2. Cross-hybridisation causes low-level, non-specific binding of transcripts to non-targeted probes, skewing their reported expression strength. Since this effect is probe dependent, this means that while identical transcripts' relative abundances can be compared across arrays, abundance of different transcripts *on the same array* cannot be compared. 3. Several interesting biological questions which cannot be answered by microarray analysis at all, or only with difficulties, become tractable using new methods; these include the analysis of isoforms from alternative splicing [Katz & al., 2010] and of allele-specific expression [Pickrell & al., 2010].

1.3.3. RNA-seq

In 2008, focus shifted from microarrays to whole-transcriptome shotgun sequencing for RNA quantification [Nagalakshmi & al., 2008; Mortazavi & al., 2008; Marioni & al., 2008]. In contrast to other RNA quantification approaches, whole-transcriptome shotgun sequencing (now typically known as RNA-seq) is entirely unbiased in that it does not rely on a pre-selected set of transcripts to assay. The approach has been shown to yield high-quality results, is highly replicable, has very low noise, and is sensitive to transcripts present at low concentration.

RNA-seq sample preparation

RNA-seq analysis starts with the enrichment of relevant transcripts from a sample's total RNA pool. This is important since the RNA fraction that we are usually interested in (mRNA constitutes 3 to 5 per cent of the total RNA [Alberts & al., 2002]) is dwarfed in abundance by the fraction of rRNA (more than 70 per cent of the cell's RNA is rRNA in eukaryotes; in prokaryotes, that number is even higher [Sittman, 1999]), and would thus dilute the signal. In practice, this enrichment can be done in one of two ways: 1. Polyadenylated RNA can be targeted due to its affinity to oligo(dT) primers, short strings of thymine; this will efficiently capture mRNA, which is post-transcriptionally 3′ tagged with poly(A) tails [Mortazavi & al., 2008]. 2. The RNA can be rRNA depleted using a RiboMinus protocol, which specifically targets rRNA molecules for removal [Cui & al., 2010]. Thus, if one wants to profile non-mRNA molecules, rRNA depletion rather than poly(A) selection is the method of choice. However, neither method is *sufficient* to reliably select only a particular type of RNA. For one thing, neither method is 100 per cent specific; in addition, mRNAS are not the only type of RNA that is polyadenylated: a large fraction of ncRNA is also polyadenylated and exported from the nucleus into the cytoplasm [Cheng & al., 2005].

The enriched RNA is subsequently fragmented to a uniform length of about 200 nt, and reverse transcribed into complementary DNA (*cDNA*). Ultimately, the result is a cDNA library of fragments which are then sequenced on a high-throughput sequencing machine (figure 1.8).

Computational processing of the sequencing information

Sequencing the RNA-seq samples yields short-read libraries, typically several tens of millions of reads in size, with reads of uniform length from 25 bp up to (currently) about 125 bp. These are then mapped to a reference — either the whole genome or the transcriptome — or assembled *de novo* (usually in absence of a suitable reference). This assigns reads to genomic locations, which can be queried and matched to features (usually genes or exons) [Cox, 2007; D. Kim & al., 2013; Anders, Pyl, & al., 2014].

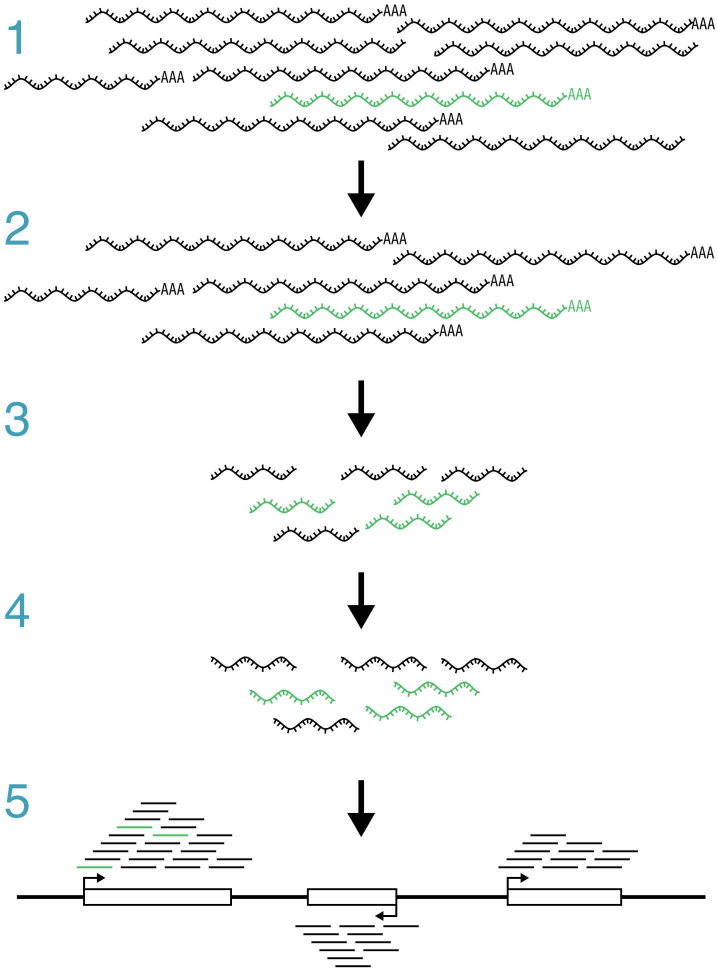


Figure 1.8: Typical RNA-seq workflow. Starting from a total RNA sample (1), the RNA of interest is enriched either via poly(A) selection or rRNA depletion (2). The enriched fraction is fragmented and size-selected (3). The fragments are reverse transcribed into cDNA (4) and sequenced. The resulting sequencing reads are aligned back to a reference (5) and counted on features of interest (adapted from Mortazavi & al. [2008]).

It is important to distinguish whether the cDNA fragments were sequenced from both ends or from one end only. In the first case, instead of a single read per fragment we end up with paired-end reads, which are separated by an approximately known distance (the fragment length minus the lengths of the sequence reads). Mapping and assembling paired-end reads creates further algorithmic challenges but increases the amount of information contained in a read (pair), which increases the amount of unambiguously mappable data (figure 1.9) [Langmead and Salzberg, 2012].

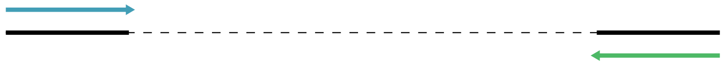


Figure 1.9: **Transcript fragment with paired-end reads.** The transcript is sequenced from both ends, resulting in a read pair whose distance to each other can be inferred from the known approximate fragment length. As long as one of the reads is "anchored" by being uniquely mapped, the paired read can be used to improve the mapping confidence in a repeat region.

Another distinction is made for fragments which are tagged and selected in such a way as to be strand-specific. This allows identification of the strand from which fragments originate, and aids in the unambiguous reassembly of the sequenced reads, as well as the identification of antisense information within (intronic regions in) a gene body, which would otherwise confound the expression estimate (figure 1.10) [Yassour & al., 2010].

1.3.4. Expression normalisation

Since the ultimate goal of RNA-seq is to quantify transcript abundance, the next step is to quantify the number of reads mapping to genomic features. In the simplest case, one can count the number of reads overlapping with a feature's genomic range. This is what e.g. HTSeq [Anders, Pyl, & al., 2014] does. However, read counts obtained in this manner vary with the length of the mappable region — longer features originate more sequenced fragments, and hence more reads, with equal coverage, compared to shorter features — and with the total size of the sequenced library.

Mortazavi & al., 2008 therefore introduced a relative measure of transcript abundance, the read count per kilobase of transcript per million mapped reads (*RPKM*), defined as

$$x_i^* = \frac{x_i}{\tilde{l}_i \cdot 10^{-3} \cdot n \cdot 10^{-6}}\,, \tag{1.1}$$

where x_i^* is the RPKM of transcript i, x_i is the raw number of reads an-

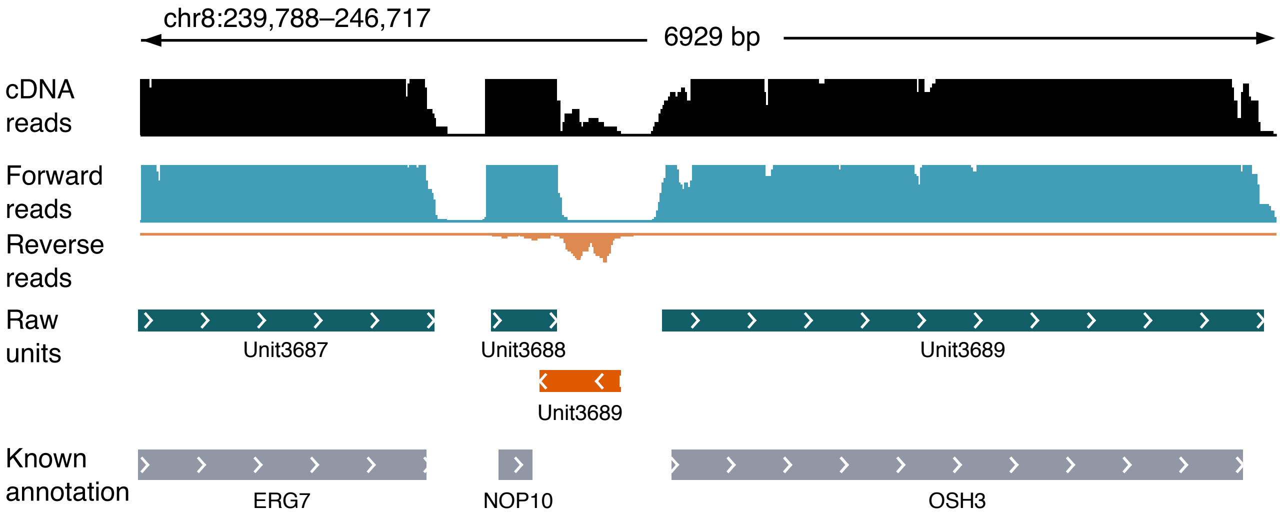


Figure 1.10: Effect of strand-specific mapping on feature identification. RNA-seq coverage data from *S. cerevisiae* without strand specificity (top) and from two libraries prepared with forward- and reverse-strand specific protocols, as well as the feature units thus found. This example shows that strand-specific mapping can lead to the discovery of new reverse-strand features, which may skew the expression quantification of forward-strand features (figure modified from Yassour & al. [2010], licensed under CC-BY).

notated with transcript i, $\tilde{l}_i$ is the effective length of the transcript (i.e. its length minus the fragment length plus 1) and n is the library size (in number of reads). The constant multiplier 1×10^9 merely serves to make otherwise very small values more manageable. With the advent of paired-end sequencing the RPKM has been supplanted by the FPKM, simply replacing the number of reads in the equation with the number of fragments (i.e. the read pairs rather than single reads in the case of paired-end data).

A related measure is transcripts per million (*TPM*), which additionally normalise by the total transcript abundance [B. Li & al., 2010]. In other words,

$$x_i^* = \frac{x_i}{\tilde{l}_i} \cdot \left(\sum_j \frac{x_j}{\tilde{l}_j} \right)^{-1} \cdot 10^6 \,. \qquad (1.2)$$

TPM succinctly answers the question, "given one million transcript in my sample, how often will I see transcript i?"

It is important to note that both approaches give a sample dependent abundance, and thus neither of these units makes measured transcript abundance comparable across different experiments: Consider two biological conditions which are identical except for a transcript X, which is highly abundant in condition A, and not present in condition B. If we sequence the same amount of mRNA from both samples, relatively fewer fragments of the non-X transcript will exists for condition A relative to condition B, even though their absolute abundance in the original sample was the same [Robinson and Oshlack, 2010].

To compare transcript abundance across biological samples, a different approach has to be taken.

1.3.5. Differential expression

One of the main uses of RNA-seq data is to compare and contrast gene expression between different conditions, such as between tissues, between different species, between healthy and tumour tissues, &c. By doing so, we hope to establish which genes characterise differences and may thus be *causal* of the phenotypic change.

To find statistically significant differences in gene expressions between two samples, we want to test the null hypothesis that the gene count in both samples comes from the same distribution with the same mean ($H_0 : \mu_{iA} = \mu_{iB}$ for a gene i between two conditions A and B).

Reads are assumed to be sampled independently from a population, thus read counts on a given feature can be approximated by a Poisson distribution [Mortazavi & al., 2008; Marioni & al., 2008]. However, actual expression data has been shown to be over-dispersed compared to this model [Robinson and Smyth, 2007]. In order to accurately model this greater dispersion, a negative binomial distribution can be used instead.

$$X_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma_{ij}^2)\,, \tag{1.3}$$

for each gene i in library j, with mean μ_{ij} and variance σ_{ij}^2.

To account for different sampling depth across libraries, it is furthermore assumed that most genes do not drastically change expression between biological samples. But the few that are highly expressed in some but not all samples have a large influence on the total count. Thus, rather than normalising by the ratios between total library sizes, each library $j = 1 \dots m$ can be normalised by a summary of the ratios between read counts for all n genes or a subset of the genes across libraries [Robinson and Oshlack, 2010; Anders and Huber, 2010]. Generalised to more than two libraries, gene expression ratios can be calculated as the ratio of each gene's read count to the geometric mean across samples:

$$s_j = \text{median}_{i=1}^{n} \frac{x_{ij}}{\sqrt[m]{\prod_{\nu=1}^{m} x_{i\nu}}} \, . \tag{1.4}$$

The parameters of the negative binomial distribution can, in principle, be estimated from the data. However, this is complicated by the typically very small number of samples. Different solutions for this problem exist. edgeR [Robinson, McCarthy, & al., 2010] fixes one parameter per sample and only estimates the other for each gene, while DESeq by Anders and Huber [2010] pools data across genes of similar expression strength, and performs local regression to find the dispersion.

1.4. Pol III ChIP-sequencing quantifies tRNA gene expression

1.4.1. ChIP-seq is a DNA binding assay

ChIP-seq *chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing*

ChIP-seq is a family of assays based on high-throughput sequencing, similar to RNA-seq, but pre-dating the latter [Johnson & al., 2007]. Unlike RNA-seq, which quantifies the abundance of RNA transcripts in the cell, ChIP-seq pinpoints loci of protein–DNA interaction for specific proteins that can be targeted with an antibody. There are several distinct applications of ChIP-seq that all rely on the identification and quantification of binding sites of specific proteins to DNA. The most common uses of ChIP-seq are: 1. the identification of novel TF binding sites by targeting specific, known TFS, and 2. the profiling of histone modifications such as methylation or acetylation, which reveal information about the transcriptional activity of the proximal sequence (see section 1.2.1) [Barski, Cuddapah, & al., 2007].

Briefly, a sample is prepared by cross-linking proteins to the DNA in solution using formaldehyde to ensure that transient interactions are captured, instead of being dissolved during the assay preparation. Next, sonication or *MNase* treatment is used to shear DNA into smaller fragments. Some of these fragments will have the protein of interest bound. Using an antibody that recognises the protein of interest as specifically and sensitively as possible, these fragments are purified. The protein is then unlinked and the remaining DNA fraction is again purified, size selected, ligated to sequencing adapters, amplified and sequenced (figure 1.11) [Park, 2009].

MNase *DNA-cleaving enzyme (nuclease) purified from specific bacteria*

1.4.2. Quantifying expression of tRNA genes

A central aspect of this thesis is the investigation of genome-wide tRNA gene expression. tRNA gene expression can be quantified via pol III ChIP-seq, using an antibody that specifically recognises the pol III subunit RPC1/155, which forms part of the active centre of tRNA gene transcription [Ablasser & al., 2009]. The reason for using this, on the first glance, indirect measure is because tRNA genes are not identifiable by their sequence alone: performing a multiple sequence alignment of tRNA genes in *M. musculus* reveals that several tRNA genes share the exact same sequence (figure 1.12).

Consequently, to identify individual tRNA genes and to quantify their expression, we cannot use conventional RNA-seq, since the RNA reads covering only the transcribed gene region are not uniquely mappable. Common strategies for counting ambiguously mapping reads, as used in ERANGE by Mortazavi & al. [2008], still require at least *some* unambiguous information, to distinguish different genes that share reads (and the problem of multi-mapping reads continues to pose challenges, as recent publications highlight [Kahles & al., 2015]).

We solve this problem by extending the tRNA gene body into the flanking regions, which are not under *purifying selection*, and therefore not conserved. As pol III ChIP-seq fragments cover the flanking regions as well as the actual gene body (figure 1.13), we can use reads uniquely mapping to the flanking

purifying selection *selective force that prevents mutations*

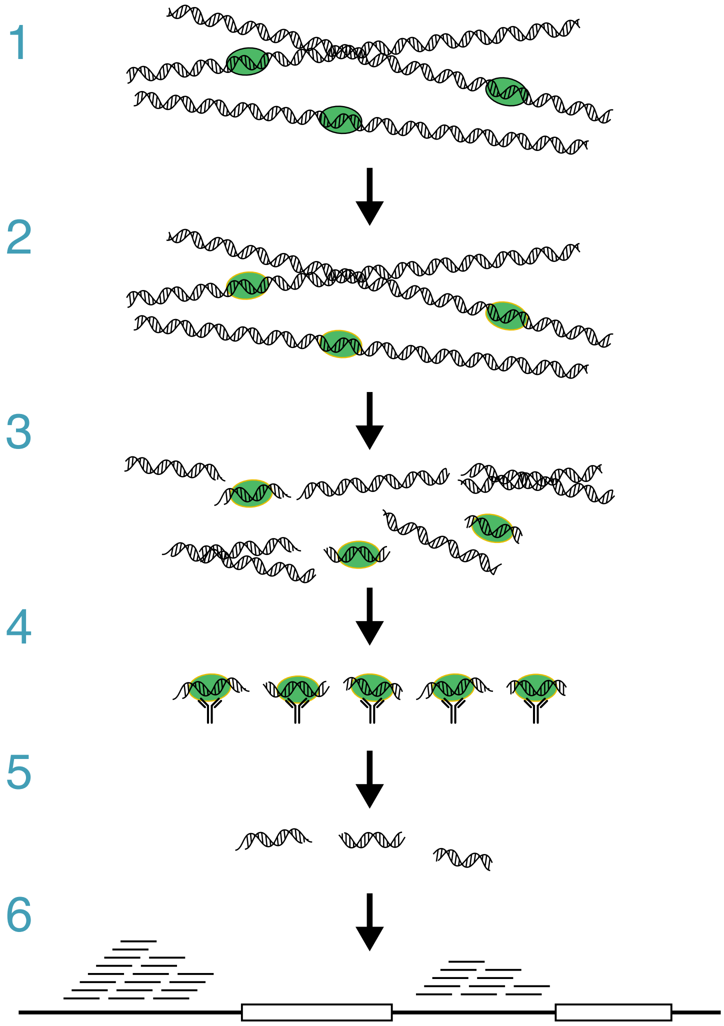


Figure 1.11: **Typical ChIP-seq workflow.** Starting from a sample with a target protein bound to DNA (1), the protein is covalently cross-linked to the DNA (mainly with formaldehyde, 2). Next, the DNA is fragmented (mainly via sonication, 3) and the protein-bound fragments are immunoprecipitated with a specific antibody targeting the protein of interest (4). The protein is unlinked and the DNA purified (5). This is followed by sequencing library preparation, sequencing and mapping of the read data back to the reference (6). This workflow follows the general outline presented in e.g. Landt & al. [2012].

regions to assign ambiguously mapped reads to the appropriate regions (figure 1.14).

More precisely, when mapping the reads, we do not discard all ambigu-

chr5.trna1044	GTCTCTGTGGCGCAATCGGTtAGCGCGTTCGGCTGTTAACCGAAAGGtTGGTGGTTCGAGCCCACCCAGGGACG
chr3.trna750	GTCTCTGTGGCGCAATCGGTtAGCGCGTTCGGCTGTTAACCGAAAGGtTGGTGGTTCGAGCCCACCCAGGGACG
chr3.trna298	GTCTCTGTGGCGCAATCGGTtAGCGCGTTCGGCTGTTAACCGAAAGGtTGGTGGTTCGAGCCCACCCAGGGACG
chr3.trna294	GTCTCTGTGGCGCAATCGGTtAGCGCGTTCGGCTGTTAACCGAAAGGtTGGTGGTTCGAGCCCACCCAGGGACG
chr3.trna289	GTCTCTGTGGCGCAATCGGTtAGCGCGTTCGGCTGTTAACCGAAAGGtTGGTGGTTCGAGCCCACCCAGGGACG
chr2.trna1947	GTCTCTGTGGCGCAATCGGTtAGCGCGTTCGGCTGTTAACCGAAAGGtTGGTGGTTCGAGCCCACCCAGGGACG
chr1.trna1014	GTCTCTGTGGCGCAATCGGTtAGCGCGTTCGGCTGTTAACCGAAAGGtTGGTGGTTCGAGCCCACCCAGGGACG
chr11.trna1446	GTCTCTGTGGCGCAATCGGTtAGCGCGTTCGGCTGTTAACCGAAAGGtTGGTGGTTCGAGCCCACCCAGGGACG
chr10.trna390	GTCTCTGTGGCGCAATCGGTtAGCGCGTTCGGCTGTTAACCGAAAGGtTGGTGGTTCGAGCCCACCCAGGGACG
chr3.trna757	GTCTCCGTGGCGCAATCGGTcAGCGCGTTCGGCTGTTAACCGAAAGGtTGGTGGTTCGAGCCCACCCGGGGACG
chr3.trna283	GTCTCTGTGGCGCAATTGGTtAGCGCGTTCGGCTGTTAACCGAAAGGtTGGTGGTTCAAGCCCACCCAGGGACG

Figure 1.12: Alignment of tRNA$^{\mathrm{Asn}}$ genes. Parts of a multiple sequence alignment of tRNA genes in *M. musculus* generated with COVE. Shown are the tRNA genes coding for asparagine. Bases which differ from the consensus sequence are highlighted in blue.

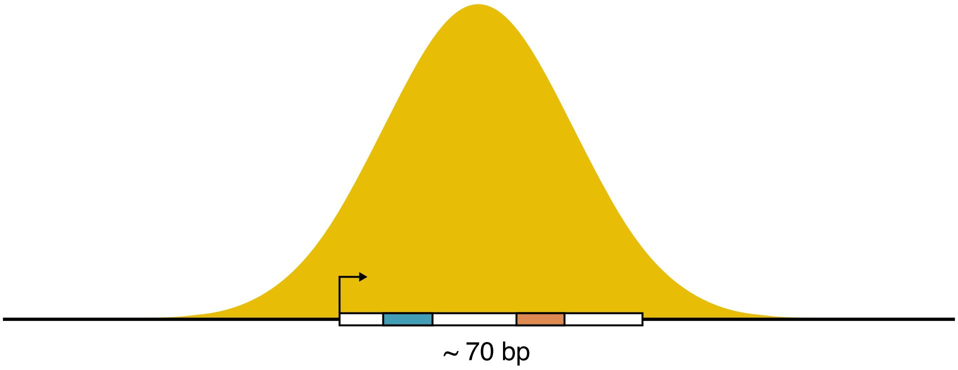
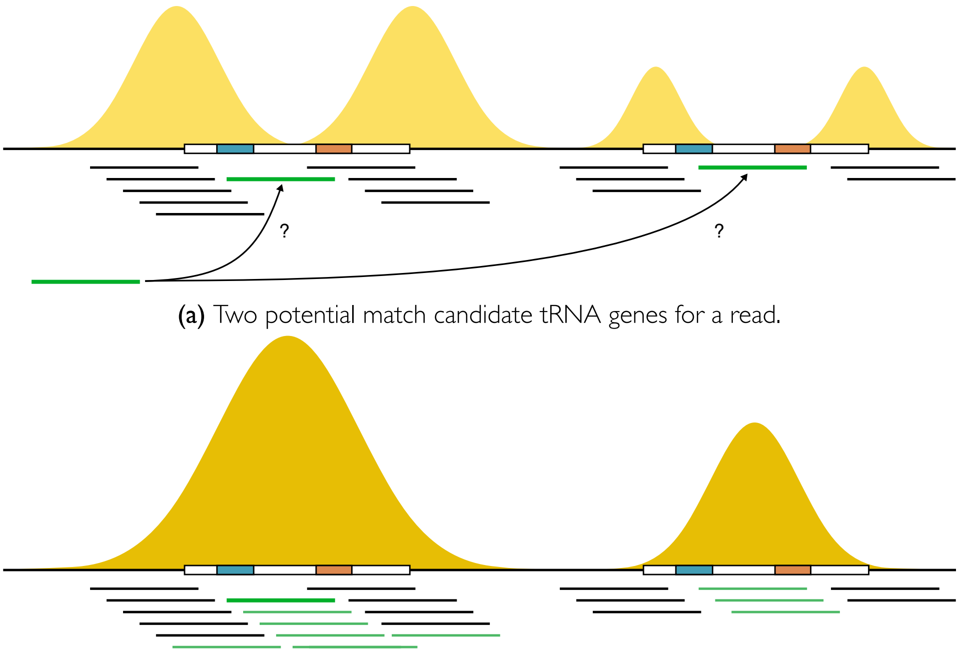


Figure 1.13: tRNA pol III ChIP binding profile. The shaded, bell-shaped area shows an idealised binding profile of ChIP-seq data spanning the tRNA gene with the A and B box highlighted, as well as its flanking regions upstream and downstream of the gene body. This overlap plays a role in identifying the individual gene.

ously mapping reads. However, we still discard reads which are likely polymerase chain reaction (PCR) duplicates, i.e. we remove all but one copy of non-unique reads in the raw input data. Despite this, we still have reads that have not been uniquely assigned to a given tRNA gene (figure 1.14a). To assign these reads, we use the number of uniquely mapping reads in tRNA genes' flanking regions to determine the most likely origin (figure 1.14b) [Kutter & al., 2011].

Formally, let i be the ith tRNA gene locus, and c_i be the count of uniquely mapped reads in its flanking region (we used ±100 bp, which has been shown to work well in practice [Kutter & al., 2011]). A multi-mapping read r, which maps to a set T of candidate tRNAs, can be allocated to a target tRNA gene i randomly with probability

$$p_i = \begin{cases} c_i \big/ \sum_{x \in T} c_x & \text{if } \sum_{x \in T} c_x \neq 0, \\ 1 \big/ |T| & \text{otherwise.} \end{cases} \tag{1.5}$$



(a) Two potential match candidate tRNA genes for a read.

(b) Using the count data from the flanking regions to extrapolate most likely mapping positions for ambiguous reads.

Figure 1.14: Mapping ambiguous ChIP reads. ChIP reads originating from tRNA genes can often not be mapped unambiguously to any given tRNA. Instead, information from the gene's flanking regions is used to determine the more likely provenance.

1.5. Mouse liver development

In chapter 2 I will present results from studies in the developing liver in *M. musculus*. Mouse liver was chosen because it is a well-studied model organ that exhibits interesting shifts in function during development. Another feature that makes liver uniquely well-suited for the study of gene expression is its homogeneity, with over 70 per cent (by volume) consisting of just one cell type, hepatocytes [Si-Tayeb & al., 2010]. The signal from whole-tissue transcriptome assays thus largely corresponds to a single cell type rather than a mixture of different signals from different cell types.

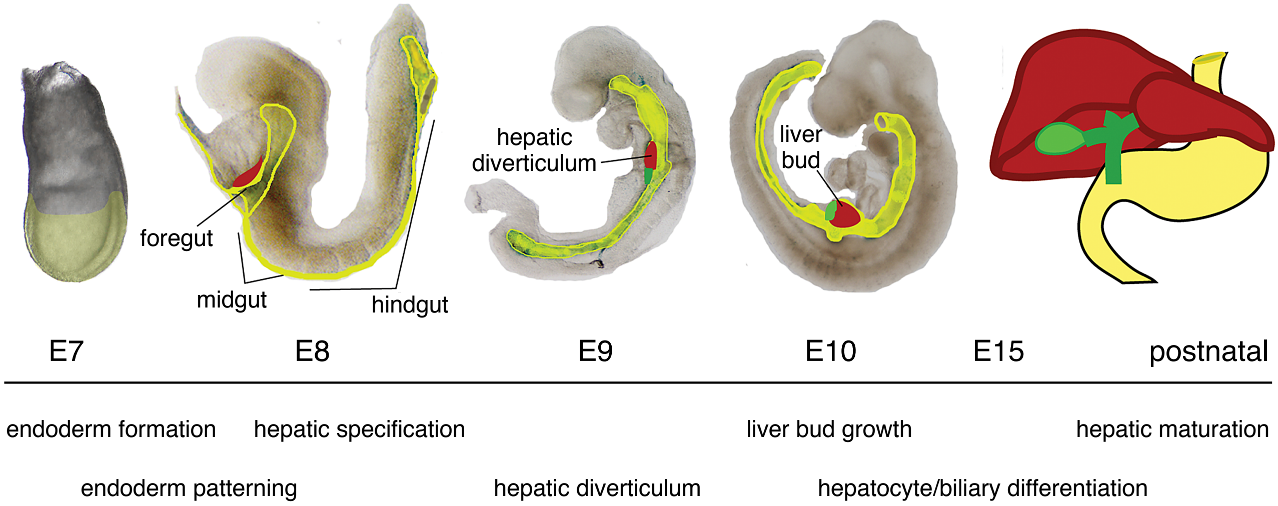


Figure 1.15: Embryonic mouse liver development. Embryo at four different stages of development, to show how the precursor of liver develops. The endoderm tissue is highlighted in yellow, the liver precursor is red, bile is green. The bottom gives a timeline of liver development (figure modified from Zorn [2008], licensed under CC-BY).

Mouse embryo development takes 19 days from fertilisation of the mouse oocyte to birth. Liver formation starts with the formation of the *hepatic diverticulum* at E9 (i.e. on day 9 after fertilisation) from a thickening of the ventral foregut endoderm (figure 1.15). The *liver bud* is subsequently formed from hepatoblasts in the anterior part of the hepatic diverticulum between E9.5 and E10, in response to growth factors from the surrounding tissue. The liver bud is also colonised by haematopoietic cells. As a consequence, the liver serves as the main haematopoietic organ in the embryo between E10 and E15 [Zorn, 2008].

Starting around E13, the liver begins to differentiate hepatoblasts into hepatocytes and biliary epithelial cells. Making up 3 per cent of the adult liver's cells, they are the second most abundant cell type in adult liver. This process leads to the formation of the characteristic tissue structure of the liver. With the maturation of the hepatocytes, this development also leads to a slow shift in organ function towards various metabolic functions and bile secretion, which are taken up immediately after birth [Si-Tayeb & al., 2010].

After birth, the liver undergoes further functional changes during the suckling–weaning transition, which happens around three weeks post birth, and which is caused by the change of diet away from fatty milk and towards solid food rich in carbohydrates [Girard & al., 1992]. Liver thus passes through three major functions during development: haematopoiesis, metabolism of fat rich diet and metabolism of carbohydrate rich diet.

1.6. Quantifying codon usage and anticodon abundance

We have seen how the genetic code defines the interface between mRNA and proteins, and how tRNAs are the physical link between the codons on the one hand and the amino acids on the other hand. As the abundance of different mRNA transcripts varies with the cell state, so does the number of different codons that are used by these transcripts. This immediately suggests that the variable demand of codons to be decoded needs to be met, in some way, by a supply of tRNA molecules carrying matching anticodons. In this thesis, I am exploring this supply–demand relationship between codons and anticodons by investigating the abundance of tRNAs in cells and its relationship to the demand of codons in the mRNA transcriptome.

The existence of a relationship between codon usage and anticodon abundance, with the codon usage adapting to the availability of matching anticodon tRNAs, was first demonstrated in *Escherichia coli* in the early 1980s [Ikemura, 1981a; Ikemura, 1981b; Ikemura, 1985]. The influence of tRNA abundance on codon choice is particularly important in organisms where no or only limited gene expression regulation exists: to modulate protein abundance, alternative, post-transcriptional regulatory mechanisms must be in charge of controlling protein abundance. Codon bias constraints imposed by tRNA anticodon isoacceptor abundance can fulfil this regulatory role [Horn, 2008].

This thesis focuses on *M. musculus* development as a mammalian model system. It has previously been established that tRNA gene transcription varies between different tissues in *Homo sapiens* using microarray expression data for a subset of tRNA genes [Dittmar & al., 2006]. Furthermore, it has been observed that tumour cells may drive overexpression of some tRNA genes [Winter & al., 2000; Pavon-Eternod, Suzanna Gomes, & al., 2009], and that, conversely, overexpressing $\text{tRNA}_\text{i}^\text{Met}$ leads to increased proliferation in human cells [Pavon-Eternod, Suzana Gomes, & al., 2013]. However, it remains unknown to what extent tRNA gene expression in mammals varies under biological conditions, and whether these changes are stochastic or coordinated.

In particular, this implies that it is unknown whether mammalian cells use the tRNA anticodon isoacceptor abundance to regulate the translation rate of protein-coding genes, and thus the flow of information from genes to proteins. As a first step in answering this question, we need to know to what extent tRNA anticodon isoacceptor abundance correlates with codon usage bias. Importantly, it is not possible to infer this from findings in non-mammalian organisms, since protein abundance is regulated differently, with a heavy focus on gene expression regulation via transcription factors in mammals. In fact, the strongly tissue-specific regulation of gene expression in mammals makes it plausible that translation-related effects have at best a limited regulatory effect on the abundance of proteins, and this is supported by recent estimates of a strong correlation between mRNA and protein abundance [J. J. Li & al., 2014; Csárdi & al., 2014; Jovanovic & al., 2015].

In the following chapters, we will take a closer look at changes in codon usage and tRNA anticodon isoacceptor abundance, and at the quantitative relationship between codon usage and anticodon abundance. Throughout the text I am going to use several related measures:

1. codon usage,
2. relative codon usage (*rcu*),
3. anticodon abundance and
4. relative anticodon abundance (*raa*).

To illustrate how they relate to each other, I am going to use the following toy example of two transcripts consisting of 5 different types of codons (plus start and stop codons):

Gene	Sequence
A	`AUG GAU UAC AAA GAC AAA GAC GAC AAA UAC AUG AAG GAC UGA`
B	`AUG AAA UAC GAU AUG AAG GAU UAC AUG AAG UAC UAC UAC AAG UAC GAU UAC UGA`

Table 1.3: Example transcripts of two genes.

The *codon usage* of a codon is the frequency with which this codon occurs in a given transcriptome. This is either the raw number of occurrences in the transcripts under consideration, or the number of occurrences, weighted by the expression of each transcript.

Using the above example, and assuming the gene expression counts *A*:10 and *B*:20, we can summarise the codon usage as follows:

Codon	A	B	$\sum(A, B)$	$\overline{A, B}$
`AAA`	3	1	4	2.0
`AAG`	1	3	4	2.0
`AUG`	2	3	5	2.5
`GAC`	4	0	4	2.0
`GAU`	1	3	4	2.0
`UAC`	2	7	9	4.5

(a) Genomic codon usage

Codon	A	B	$\sum(A, B)$	$\overline{A, B}$
`AAA`	30	20	50	25
`AAG`	10	60	70	35
`AUG`	20	60	80	40
`GAC`	40	0	40	20
`GAU`	10	60	70	35
`UAC`	20	140	160	80

(b) Codon usage weighted by gene expression

Table 1.4: Example codon usage. "Genomic" codon usage is based purely on the (genomic) sequence of codons of the transcript. Codon usage weighted by expression multiplies each codon in a transcript by the transcript's measured abundance.

Note that we have omitted the stop codon, but not the start codon, since the latter corresponds to a tRNA isoacceptor, while the former does not. The columns $\sum(A, B)$ and $\overline{A, B}$ show how the aggregated codon usage of a transcriptome consisting of these two genes would look, either as the sum of the codon usage of all genes, or as the arithmetic mean. In the following chapters, we will rarely look at the codon usage of single transcripts, and rather at either the sum codon usage or the mean codon usage of gene sets. When comparing the codon usage of gene sets of different size (to determine, for instance, which codons are used more in condition I compared to condition II), we need to use the mean codon usage rather than the sum; for sets of the same size, either metric works. However, we are generally not interested in the absolute number of codons used, but rather in the shift of use between synonymous codons. Here, using the RCU is more appropriate.

The *relative codon usage (RCU)* of a codon is that codon's contribution to the amino acid it codes for, relative to all other synonymous codons. The RCU of all synonymous codons sums to 1. Let C_a be the set of all codons coding for amino acid a, and $c \in C_a$ a codon in that set. Furthermore, let x_c be the codon usage of codon c, and m_c the RCU of c. Then, for all amino acids a and all $c \in C_a$,

$$m_c = \frac{x_c}{\sum_{i \in C_a} x_i} \,, \tag{1.6}$$

and consequently

$$\sum_{c \in C_a} m_c = 1 \,. \tag{1.7}$$

Table 1.4 illustrates two interesting things. First, when considering only a single gene, it is immaterial whether we consider expression or not: since the RCU is a *ratio* of the overall codon usage, multiplying all codon usage values with a constant changes nothing (column A is the same in figure 1.16a and figure 1.16b, and so is column B). And secondly, both in the case of the genomic RCU and in the RCU weighted by gene expression the sum RCU and the mean RCU are identical.

Codon	AA	A	B	$\sum(A, B)$	$\overline{A, B}$
AAA	Lys	0.75	0.25	0.5	0.5
AAG	Lys	0.25	0.75	0.5	0.5
AUG	Met	1.0	1.0	1.0	1.0
GAC	Asp	0.8	0.0	0.5	0.5
GAU	Asp	0.2	1.0	0.5	0.5
UAC	Tyr	1.0	1.0	1.0	1.0

(a) Genomic RCU

Codon	AA	A	B	$\sum(A, B)$	$\overline{A, B}$
AAA	Lys	0.75	0.25	0.41(6)	0.41(6)
AAG	Lys	0.25	0.75	0.58(3)	0.58(4)
AUG	Met	1.0	1.0	1.0	1.0
GAC	Asp	0.8	0.0	0.(36)	0.(36)
GAU	Asp	0.2	1.0	0.(63)	0.(63)
UAC	Tyr	1.0	1.0	1.0	1.0

(b) RCU weighted by gene expression

Table 1.4: **Example RCU** with the amino acid for each codon shown; in each column, the values for the same amino acid sum to 1. Parentheses denote periodic decimals.

The *anticodon abundance* of an anticodon is the amount of tRNA decoding a given anticodon, present in the cell at a given instance. Other publications define the anticodon abundance purely in terms of tRNA gene copy number; however, in the context of this thesis, the anticodon abundance is quantified by tRNA gene expression, and is thus an estimate of the number of tRNA molecules of each anticodon isoacceptor present in the cell.

The *relative anticodon abundance (RAA)* of an anticodon is defined equivalently to the RCU based on the anticodon abundance. That is, the contribution of an anticodon to its amino acid isotype, relative to the other anticodons in the same isotype.

Several publications use the term *codon usage bias (CUB)* to describe divergence in codon usage between different sets of genes within a genome, or differences between genomes. The CUB is then equivalent to the variation in either the codon usage as defined above, or the RCU.

codon usage bias *a divergence of the codon usage from a uniform distribution, where all alternative codons are used at the same proportions;* or *the difference in codon usage frequency or relative codon usage between sets of genes*

To study whether tRNA anticodon isoacceptor availability influences translation efficiency, we want to compare the demand in codons in transcribed genes to the supply in matching anticodon aminoacyl-tRNAs. In this thesis, we do this by calculating the correlation between the RCU and the RAA. This gives us a measure of how well the set of codons in a gene, gene set or transcriptome are adapted to the tRNA abundance. This correlation ignores wobble base pairing; in fact, some codons are not matched by any anticodon and would thus negatively impact the correlation. In the following we disregarded these codons when calculating the codon–anticodon correlation, rather than estimating the extent of matching of a given anticodon to different codons via wobbling. Despite this shortcoming, we find that this simple correlation works well in estimating codon–anticodon correlation (chapter 2, figure 2.9), and performs comparably to other measures, such as distributing a codon's usage between all its matching tRNA anticodons according to their abundance [Ikemura, 1981a].

1.7. Structure of this thesis

In the next chapters I am going to present the research I performed — in collaboration with colleagues — to study the abundance of tRNA and mRNA in mammals, and how these are linked. In chapter 2 I will present my analysis of the dynamic changes of tRNA gene expression in mouse development and how it relates to changes in protein-coding gene expression. In this chapter I will establish our model of tRNA gene expression, which postulates that changes in tRNA gene expression are nonrandom, and concerted to stabilise the abundance of tRNAs.

In chapter 3 I will focus more closely on the potential adaptation of the codon pool to the abundance of tRNAs. My interest in this topic was triggered by the publication of results closely related to those I presented in chapter 2, and which I will therefore describe and build on. I will explore the hypothesis that mammalian codon usage, like that of other organisms, is shaped by the abundance of tRNAs, and that codon bias is used to regulate the translation of cell type specific genes.

In chapter 4, finally, I will take a short glimpse at the world of the pol III transcriptome beyond tRNAs. I will quantify the association of pol III with different genomic features. Based on this, I will also introduce a future project whose aim is to specifically look at a particular set of genomic features that form part of a class of genes called *transposable elements*.

2 Developmental stability of the mRNA–tRNA interface

To study how changes in mRNA gene expression relate to changes in tRNA gene expression, we collected tissue samples from six time points in mouse (*M. musculus*) development: two before birth (E15.5 and E18.5, which stands for 15.5 and 18.5 days after fertilisation of the oocyte, respectively); two shortly after birth, happening around E20 (P0.5 and P4 — 0.5 and 4 days after birth, respectively) and two after weaning the juvenile mice (P22 and P29).

For each of these time points, tissue was collected from whole liver and whole brain (*homogenised*) and prepared for RNA-seq and pol III ChIP-seq in order to assay mRNA and tRNA gene expression as explained in sections 1.3 and 1.4. The tissues were chosen for their interesting shifts in physiology during development: the liver is a homogeneous organ predominantly made up of a single cell type — around 70 per cent hepatocytes — and liver function changes fundamentally at birth; prenatal liver serves mainly as a haematopoietic organ, whereas liver of post-natal mice is primarily a metabolic organ [Si-Tayeb & al., 2010]. Brain, by contrast, is a highly heterogeneous organ made up of many different cell types, with dynamic changes all through development [Liscovitch and Chechik, 2013].

homogenise *breaking apart and mixing the tissue in such a way that all the cell types will be evenly distributed throughout the sample*

Figure 2.1 summarises the experimental procedure.[*] Each experiment was performed in two biological replicates, which were highly correlated (figure 2.2). Table 2.1 summarises the variable names used to refer to data

[*] The wet-lab work of this project was performed by Bianca Schmitt, who was also a joint first author on the manuscript. Claudia Kutter provided guidance with the interpretation of the results and helped with the creation of the figures.

throughout this chapter.

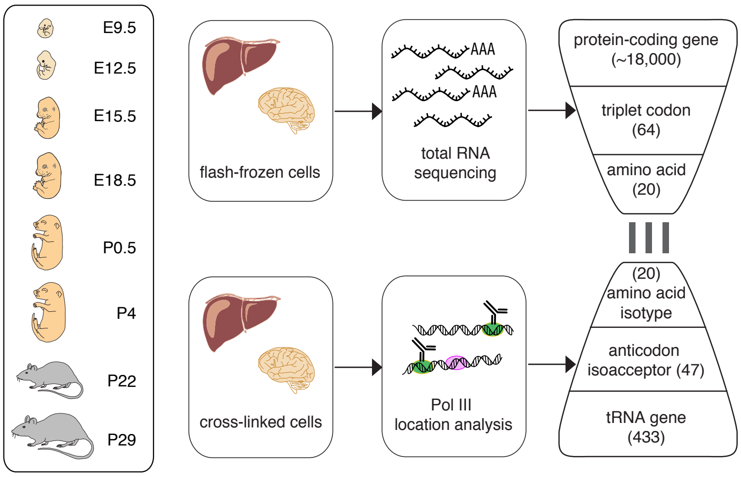


Figure 2.1: Sample analysis outline. Samples were collected in eight distinct time points. Of these, E9.5 and E12.5 were excluded from most of the subsequent analysis, except where noted. For each time point, tissue was collected from liver and brain, and on the one hand prepared for RNA-seq, and on the other hand cross-linked to the pol III antibody and prepared for ChIP-seq. The resulting data was used to quantify mRNA and tRNA gene expression, codon usage and tRNA anticodon abundance. Figure created by Claudia Kutter.

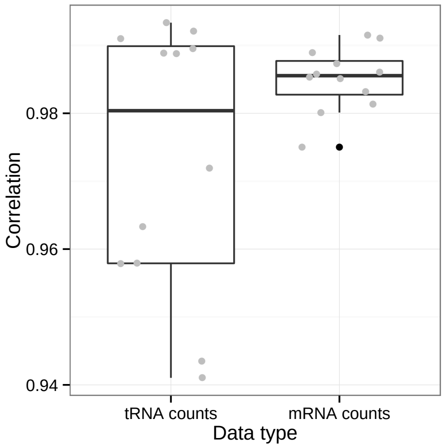


Figure 2.2: Replicate variability. Shown are the Spearman correlation coefficients between pairwise biological replicates for the tRNA count data (left) and the mRNA count data (right).

	mRNA	tRNA
Raw counts per replicate	m_{ij}	t_{ij}
Normalised counts per replicate	m^*_{ij}	t^*_{ij}
Counts of merged replicates	m'_{iy}	t'_{iy}
(Anti-)codon level counts	c_{xy}	a_{xy}

Table 2.1: **Summary of the matrix and subscript names.** i is the index of a gene; j is the index of a library replicate; x is an (anti-)codon; y is the index of a developmental stage.

In addition to the six time points described above, tissue was also collected at two earlier stages, E9.5 and E12.5. However, the embryo at such early stages of development is too small, and the tissue development has not progressed far enough, to permit collecting enough tissue-specific material. For that reason we used the whole embryo at E9.5 and separated the E12.5 embryo into torso and upper body. The subsequent analysis was performed on the six later stages in liver and brain. However, the earlier stages confirmed the general patterns found by analysing the remaining data (figure A.5).

The analysis and the results presented in this chapter are published as *"High-resolution mapping of transcriptional dynamics across tissue development reveals a stable mRNA– tRNA interface"* [Schmitt, Rudolph, & al., 2014].

2.1. Protein-coding gene expression changes dynamically during mouse development

To investigate protein-coding gene expression changes during development, we quantified the mRNA abundance from rRNA-depleted RNA-seq data (strand-specific 75 bp paired-end reads from Illumina HiSeq 2000). Reads were mapped to the *M. musculus* reference genome (NCBIM37) using iRAP [Fonseca & al., 2014] and TopHat2 [D. Kim & al., 2013]. Read counts were quantified using HTSeq [Anders, Pyl, & al., 2014], and assigned to protein-coding genes from the Ensembl release 67 [Flicek & al., 2014].

We excluded mitochondrial chromosomes from the analysis, because mitochondrial genes use a slightly distinct genetic code [Osawa & al., 1989]. Furthermore, we excluded sex chromosomes.

Changes in the expression of protein-coding genes, leading to changes in abundance of proteins, are known to drive cellular behaviour [Brawand & al., 2011]. Our data confirms that tissue development in mice is accompanied by large-scale changes in the mRNA transcriptome. This is nicely illustrated by looking at individual gene expression counts, plotted against their genomic location (figure 2.3). For example, in the liver we can confirm the functional relevance of apolipoprotein B (APOB) as the primary carrier of lipoproteins, which becomes increasingly relevant as the organ shifts to metabolism [Knott & al., 1986]. Similarly, mRNA gene expression changes highlight the role of α-fetoproteine (AFP) as the fetal version of serum albumin (figure 2.3a) [H. Chen & al., 1997]. In the brain, shifts can be seen in the activity of the neural TF FOXP2, the expression of which continuously decreases, and in calmodulin (CALM1), where transcription increases after birth [Tsui & al., 2013; Huang & al., 2011].

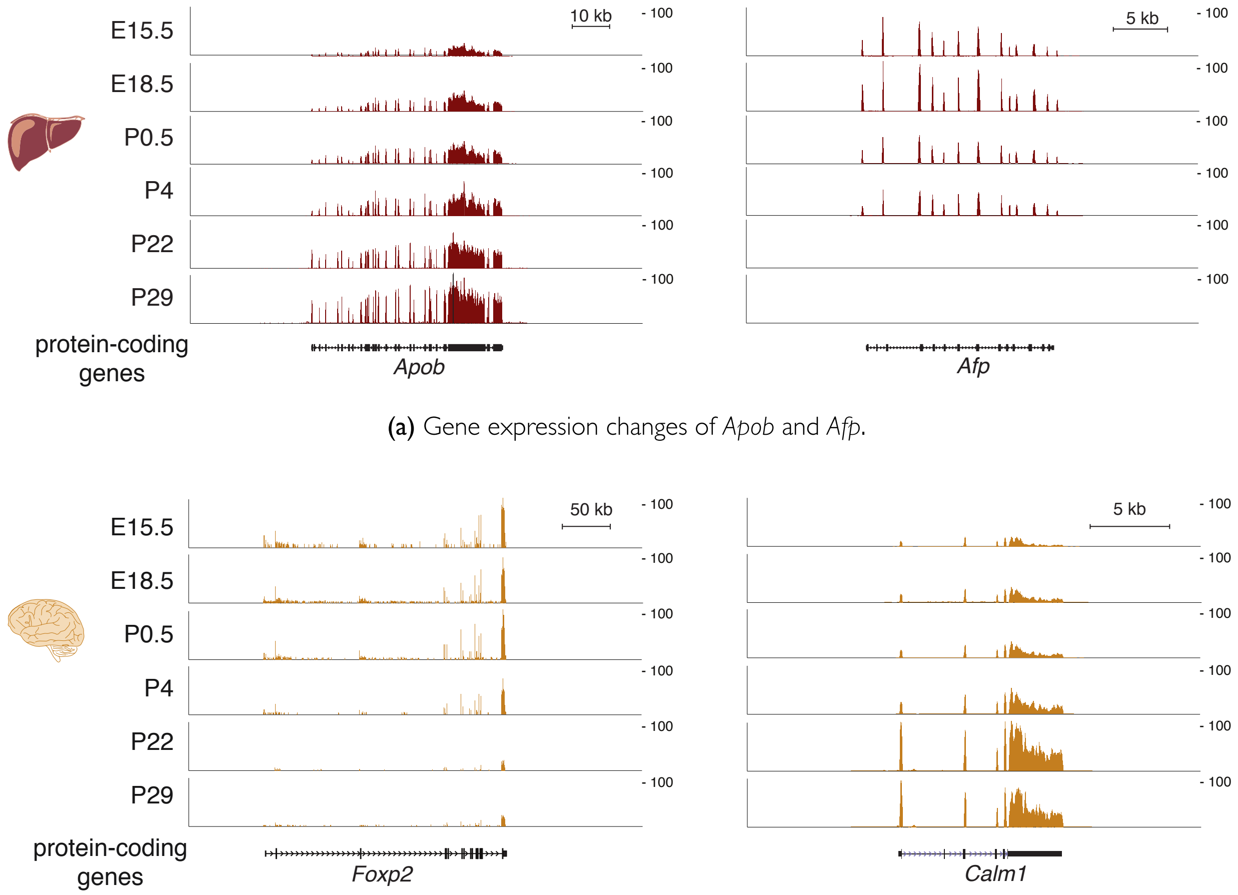
For a more systematic analysis, I took the matrix of normalised count data of all library replicates and all mRNAs, m^*_{ij}, for each mRNA gene i and each library replicate j, and calculated the pairwise Spearman rank correlation between all replicates,

$$cor_{ij} = \operatorname{cor}(m^*_{\cdot i}, m^*_{\cdot j}) \text{ for all libraries } i, j, \tag{2.1}$$

where $m^*_{\cdot i}$ denotes the ith column of the matrix M^*.

I then performed principal components analysis (*PCA*) on the correlation matrix, which allows variation in the data to be projected onto uncorrelated axes, so that the first axis represents the component that explains the most variance of the data, and the second axis represents the second component.

The resulting PCA in figure 2.4 shows that the biggest source of variance in the correlation structure of the expression data is tissue identity, which explains 97 per cent of the total variance. However, of the remaining variance, 60 per cent is explained by progression of tissue development in a way that nicely mirrors the known biology: the plot's y axis shows the linear pro-



(a) Gene expression changes of *Apob* and *Afp*.

(b) Gene expression changes of *Foxp2* and *Calm1*.

Figure 2.3: Example of gene expression changes in development. The four genes are representative for tissue- and stage-specific genes whose expression changes drive cell function. These changes can go up or down over the course of development, corresponding to either an up- or downregulation. Figure created by Bianca Schmitt.

gression of development from early stages at the bottom to late stages at the top. We observe a much stronger variation on the y axis for liver data: this could be explained by noting that the liver is more homogeneous than the brain, and changes in gene expression are therefore more coordinated; it might also reflect the change in liver function from a haematopoietic to a metabolic organ around birth.

Next, I used DESeq2 [Love & al., 2014] to identify differentially expressed genes between stages and tissues. Genes are counted as differentially expressed if their Benjamini–Hochberg FDR-corrected *p*-value is below 0.001. The number of differentially expressed genes between all pairwise developmental stages unsurprisingly shows that more distinct developmental stages have higher numbers of differentially expressed genes (figure 2.5). Furthermore, there is a clear gap between pre- and post-weaning stages, with a large jump in the number of differentially expressed genes across the weaning boundary, in both liver and brain.

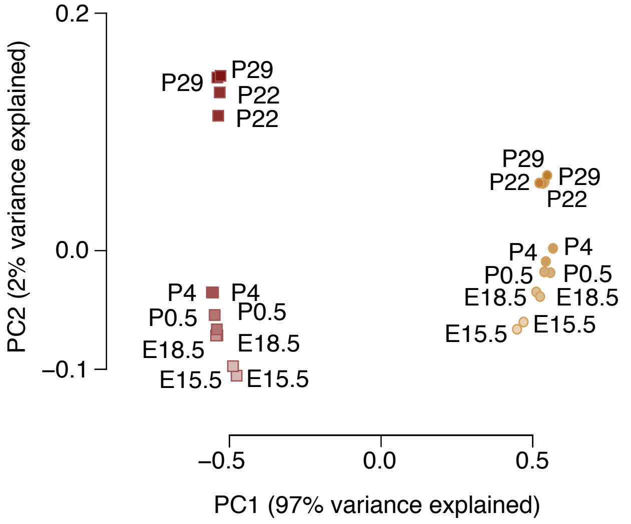
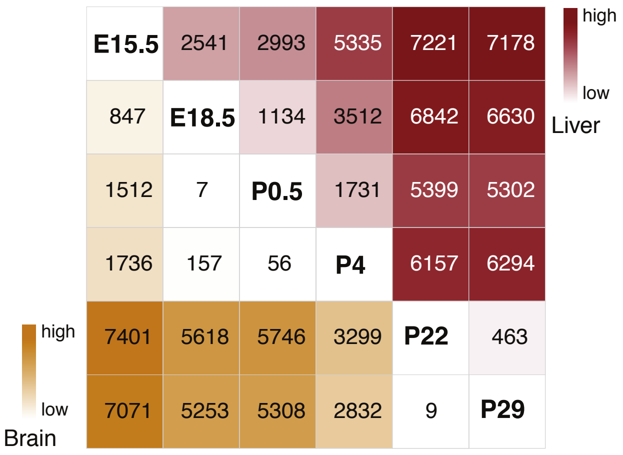


Figure 2.4: PCA of mRNA gene expression per developmental stage. Rotations 1 and 2 of the correlation matrix of protein-coding gene expression in each developmental stage. The percentage on the axes shows the amount of variance explained by each rotation. Points corresponding to liver samples are coloured in red, points correspoding to brain samples are coloured in yellow; stages of development go from light colours to dark colours.


high
low
Liver
high
low
Brain

E15.5	2541	2993	5335	7221	7178
847	**E18.5**	1134	3512	6842	6630
1512	7	**P0.5**	1731	5399	5302
1736	157	56	**P4**	6157	6294
7401	5618	5746	3299	**P22**	463
7071	5253	5308	2832	9	**P29**

Figure 2.5: Number of differentially expressed mRNA genes between stages. Each off-diagonal square shows the number of differentially expressed genes (at a significance threshold of $p < 0.01$) between the two indicated developmental stages.

These patterns are noteworthy because they recapitulate tissue identity and linear progression through the stages of tissue development. But they are not particularly surprising: cell function is dictated by the abundance of specific proteins and thus protein-coding gene transcription. The patterns of gene expression similarity shown in the PCA and in the number of differentially expressed genes hence recapitulate the expected changes in cell function between tissues and through development.

2.2. Dynamic changes of tRNA gene expression during mouse development

Quantification of tRNA genes was performed by first mapping the pol III ChIP-seq data (non-strand-specific 36 bp single-end reads sequenced by Illumina Genome Analyzer IIx or HiSeq 2000) using BWA version 0.5.9-r16 [H. Li and Durbin, 2009] using default parameters. Next, non-uniquely mapping reads were reallocated probabilistically according to the description given in the previous chapter, using the tRNA gene annotation from the Genomic tRNA Database, described in Chan and Lowe [2009]. For each tRNA gene (again excluding mitochondrial tRNA genes because the genetic code of the mitochondrial mRNA genes differs from the nuclear genetic code), reads were summed within each tRNA gene locus and in the ±100 bp flanking regions.

tRNA genes that were unexpressed in all experimental conditions were excluded from further analysis, to reduce the effect of multiple testing [Bourgon & al., 2010] and to exclude potential pseudogenes in the annotation. To be called expressed, a tRNA gene had to be present in all replicates of at least one condition with a count of at least 10, after size-factor normalisation. The threshold 10 was chosen so that small variations in either direction would have a minimal impact on the thresholding. The following analysis is thus performed using 311 expressed out of 433 total tRNA genes (72 per cent).

Unlike proteins, tRNAs do not perform a cell type specific function; instead, their continued presence is required for the maintenance of transcription in all cellular conditions. We therefore did not expect many changes in the levels of tRNA gene expression over the course of development, and we do in fact observe that many tRNA gene expression levels remain stable (figure 2.6). Nevertheless, we also observe that around 50 per cent of all tRNA genes *are* differentially expressed. Figure 2.7 shows a genomic locus containing tRNA genes that displays these different dynamics.

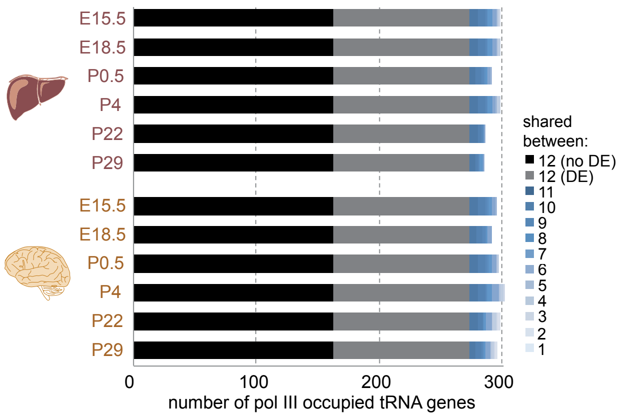


Figure 2.6: Overview over tRNA gene expression change. Bar plots show different types of tRNA gene expression dynamics: tRNA genes without change in their expression levels, tRNA genes with changes to their expression levels, which are nevertheless expressed in all stages of development across both tissues; and tRNA genes which are only expressed in a subset of all conditions. Figure created by Claudia Kutter.

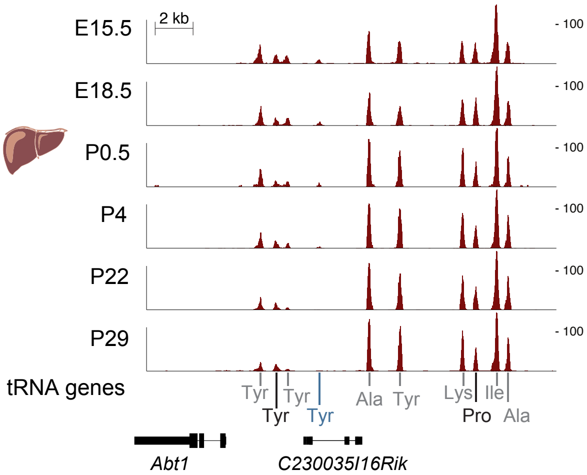


Figure 2.7: Example of dynamically changing tRNA genes. Genomic region showing different types of tRNA gene expression behaviour; the label colours on the x axis corresponds to the colours in figure 2.6. Figure created by Bianca Schmitt.

Surprisingly, the tRNA gene expression differences follow similar patterns to those observed in mRNA genes, with the first two principal components resulting from the application of PCA to the rank correlation matrix again corresponding to the tissue and developmental stage (figure 2.8). The observed patterns are incompatible with mere *random* expression changes (which would result in an unordered cloud of points). Something must account for these concerted changes in tRNA gene expression.

In the case of the protein-coding genes, we can explain the nonrandom changes in gene expression by known gene regulatory mechanisms, which control the transcriptome of each cell and developmental stage. The fact that tRNA gene expression changes across development exhibit the same patterns as mRNA gene expression changes suggests that tRNA gene expression

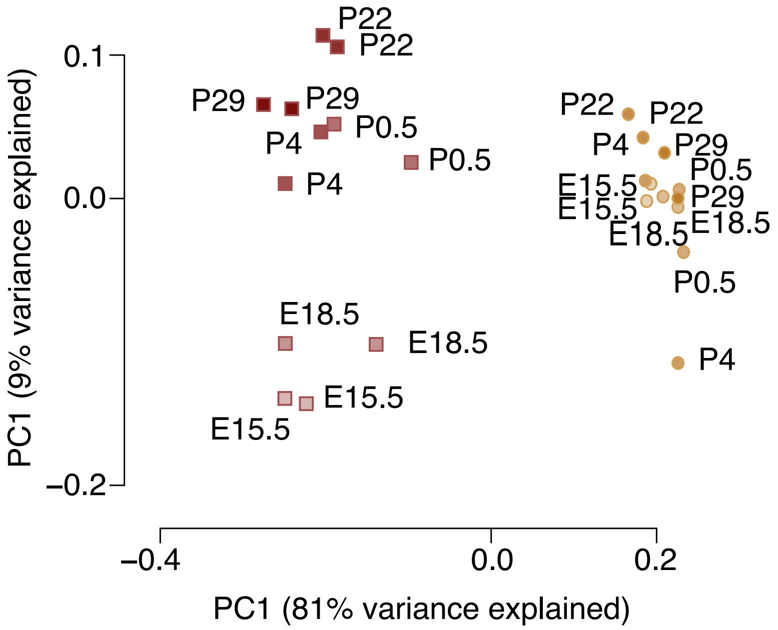


Figure 2.8: PCA of tRNA gene expression per developmental stage. Rotations 1 and 2 of the correlation matrix of tRNA gene expression in each developmental stage. The percentage on the axes shows the amount of variance explained by each rotation. Points corresponding to liver samples are coloured in red, points correspoding to brain samples are coloured in yellow; stages of development go from light colours to dark colours.

is subject to similar regulatory constraints. We therefore attempted to explain *why* tRNA gene expression requires changing in a regulated manner, and *how* this tRNA gene regulation is carried out by the cell.

Our first suspicion was that changes in mRNA gene expression might lead to changes in codon demand, since different protein-coding genes are made up from different codons. The change in codon demand in turn could lead to a change in anticodon supply in the form of differential tRNA gene expression. This would meet the need for efficient translation: mismatching codon and anticodon pools would either lead to a wasteful over-production of tRNAs of a given anticodon, or to a bottleneck in such an anticodon supply, causing efficiency loss in translation. Both scenarios present a suboptimal scenario for the fitness of the cell and should reasonably be selected against. In fact, there is some evidence that such a selection takes place [Ikemura, 1981a; Ikemura, 1985; Yang and Nielsen, 2008].

We therefore went on to quantify the codon pool corresponding to each given transcriptome, as well as the pool of available tRNA genes, grouped by their anticodon isoacceptor identity.

2.3. Every mouse mRNA transcriptome encodes the same distribution of triplet codons and amino acids

The codon pool of a given mRNA transcript is given by the distribution of triplet codons in its sequence. There are 64 possible triplet codons, of which 61 encode 20 different amino acids, and three encode the stop codon, marking the end of translation.[†] Using this information and the transcript abundance quantified by our RNA-seq data, I calculated the abundance of each triplet codon as well as each amino acid in the transcriptome of each developmental stage. I then compared these across stages to find out how they varied.

First, for every gene, the number of occurrences of each codon in the longest annotated transcript (which is often called the "canonical" transcript[‡]) was determined and this value was multiplied by the gene's expression (normalised for transcript length):

$$c_{xiy} = \text{codon}_{xi} \cdot \frac{m'_{iy}}{l_i}\ , \tag{2.2}$$

where codon_{xi} is the number of occurrences of codon x in the canonical transcript of gene i, m'_{iy} is the gene expression count of gene i in condition y (averaged across the replicates of condition y after library size normalisation), and l_i is the length of the canonical transcript of gene i.

Next, the overall usage of each codon was obtained by summing these

[†] The analysis ignores selenocysteine, which is a 21st possible amino acid, and which can be encoded by the stop codons `UGA` under rare circumstances (see section 1.2.3).

[‡] `http://www.ensembl.org/Help/Glossary?id=346`, retrieved 2014-05-12.

values across all genes:

$$c_{xy} = \sum_i c_{xiy} \,. \tag{2.3}$$

Relative codon usage (RCU) values c^*_{xy} were then calculated by dividing each codon usage value by the sum of the codon usage values of a given condition:

$$c^*_{xy} = \frac{c_{xy}}{\sum_{k \in \mathrm{syn}(x)} c_{ky}} \,, \tag{2.4}$$

where $\mathrm{syn}(c)$ is the set of synonymous codons to which c belongs.

We find that at the transcriptome level, codon abundance is highly stable across development in both tissues (Spearman's $\rho > 0.97$) — figure 2.9 top left shows this by way of example using the codons for arginine in the different stages in liver.

Given this stability, I next explored how much variation should be expected for varying transcriptomes, by simulating random transcriptomes and computing their codon usage.

I used our library-size normalised RNA-seq data to simulate background distributions in liver and brain for each specific developmental stage. I randomly rearranged the expression values across genes for the expressed ("EXPR") and all genomically annotated ("ALL") protein-coding genes. For each developmental stage, I created 100 such random background distributions. I then calculated the triplet codon usage for the rearranged protein-coding RNA expression distributions.

Figure 2.9 top middle and right shows that, even for simulated transcriptomes, the codon usage remains unchanged. In fact, both observed and simulated transcriptomes seem to simply reflect the codon abundance found in the coding part of the genome, regardless of the sometimes strong variations in gene expression between different transcriptomes.

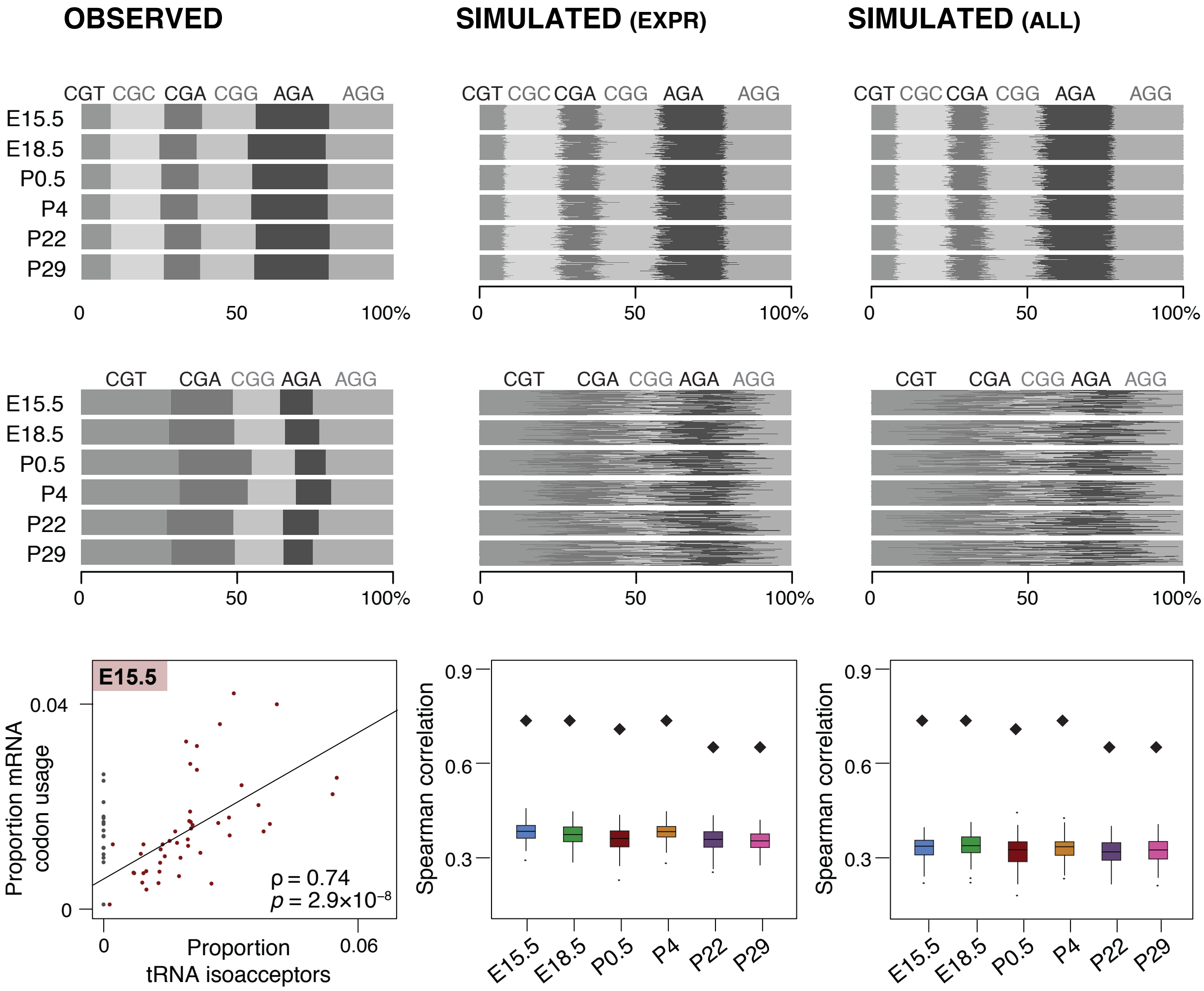


Figure 2.9: **Codon and anticodon abundance across stages of development.** The figure consists of three panels (left, middle, right) with three subfigures (top, centre, bottom) each. The left panel shows observed data for each of the developmental stages in liver (brain data is comparable). The middle and right panels show simulated data from randomised transcriptomes. The middle panel used only expressed genes of each respective stage in the simulated data, whereas the right panel uses all genes, also unexpressed ones). The top figures of each panel show relative mRNA transcript triplet codon usage, using the representative example of arginine. The centre figures show the relative tRNA anticodon abundance of the arginine isotype family. The bottom right figure shows the linear regression of relative codon usage against relative anticodon abundance in liver E15.5, along with its Spearman rank correlation. Triplet codons without directly corresponding anticodon (grey dots) were ignored in the calculation. The bottom middle and right figure shows the Spearman rank correlation coefficient of each stage's relative codon and anticodon abundance (diamond), and the range of correlation coefficients for the simulated codon and anticodon pools (box plots) for each stage. Figure created jointly with Bianca Schmitt and Claudia Kutter.

2.4. Stable isoacceptor anticodon abundance through development indicates tight regulation of tRNA gene expression

Next, I looked at the abundance of the matching tRNA isoacceptors by summing the expression of all tRNA genes belonging to the same isoacceptor family. Relative anticodon abundance (RAA) was calculated by averaging the expression values for all tRNA genes in a given anticodon isoacceptor family. Figure 2.9 centre left shows the relative anticodon isoacceptor abundance of the arginine isotype family. Again we find that the abundance stays stable across development in both tissues (Spearman's $\rho > 0.96$).

In the same way as for the mRNA transcriptome, I then simulated 100 random tRNA transcriptomes per developmental stage and calculated the relative abundance of all isoacceptor families. Unlike the observed tRNA data, we find that simulated tRNA transcriptomes create more variable pools of anticodon isoacceptors (figure 2.9 centre middle and right). The variability observed here is explained by the fact that there are only 433 tRNA genes in *M. musculus* (of which only 311 were expressed in our samples), compared to the approximately 20 000 protein-coding genes, which leads to a bigger relative influence of random sampling on the distribution. In contrast to mRNA gene expression, the stable anticodon isoacceptor abundance distribution we observe across development is therefore not compatible with random variation in the tRNA gene expression: instead, it demonstrates the necessity of a mechanism actively stabilising tRNA gene expression variation at the anticodon isoacceptor level.

2.5. mRNA triplet codon usage is highly correlated with tRNA anticodon isoacceptor abundance during development

Codons and anticodon-carrying tRNAs form the biochemical interface between the genetic code and the amino acid sequence of proteins during mRNA translation. I investigated this correspondence between mRNA-driven codon demand and tRNA anticodon supply by looking at the correlation between a codon's frequency in the mRNA transcriptome as a fraction of the overall codon count c'_{xy}, and its matching tRNA anticodon isoacceptor abundance a'_{xy} at the same developmental stage y (figure 2.9 bottom left):

$$c'_{xy} = \frac{c_{xy}}{\sum_k c_{ky}} \tag{2.5}$$

$$a'_{xy} = \frac{a_{xy}}{\sum_k a_{ky}} \tag{2.6}$$

To compare how well the anticodon supply of a given transcriptome was adapted to its codon demand, I initially calculated the Spearman rank correlation between the codon usage and the anticodon isoacceptor abundance. However, since not all codons have a corresponding anticodon-carrying tRNA, unmatched "orphan" triplet codons were discarded from the calculation. Consequently, the correlation coefficients I calculated ignore the possibility of wobble base pairing.

It is possible to account for wobble base pairing in several different ways (reviewed in Gingold and Pilpel [2011]). In particular, dos Reis, Wernisch, & al. [2003] describe the tRNA adaptation index (*tAI*), which takes into account the possible wobble base pairings when calculating the fit between codon usage and tRNA gene copy number. For my analysis, rather than accounting for all possible base pairings, I opted for a simplified version where only unmatched codons were treated differently, and all other codons were matched directly to their corresponding anticodons, as described above.

Orphan codons were matched to all anticodons that they recognise via wobble base pairing, by distributing the tRNA abundance of these tRNA genes between all matched codons; abundant codons received proportionally more of the anticodon abundance. More precisely, let x be an orphan codon and wobble(x) the anticodons that match x via wobble base pairing. We then adjust the abundance for each $x' \in \text{wobble}(x)$ according to the following rule:

$$a''_{x'y} = \left(\sum_{k \in \text{wobble}(x)} a'_{ky} \right) \cdot \frac{c'_{x'y}}{\sum_{k \in \text{wobble}(x)} c_{ky}} \,. \tag{2.7}$$

In other words, we pool the frequency of anticodons that wobble base pair with orphan codons, and let each contribute a fraction proportional to its (directly matching) codon frequency. For those anticodons x which do not wobble base pair to orphan codons, we set $a''_{xy} = a'_{xy}$.

Finally, I calculated codon–anticodon correlations between the codon usage and the adjusted anticodon abundance, $\text{cor}(c'_{\cdot y}, a''_{\cdot y})$. This yielded broadly comparable results to the simple correlations ignoring wobble base pairing and unmatched codons (figure A.12). I will briefly discuss how to improve this using a tAI adapted to tRNA gene expression in the conclusion (chapter 5).

Across both tissues and all stages of development, we find that mRNA triplet codon demand and tRNA anticodon isoacceptor abundance are highly correlated ($0.64 < \rho \leq 0.76$ Spearman's rank correlation, all $p < 0.001$), ignoring wobble base pairing. Accounting for wobble base pairing in the calculation of the adaptation of codon demand and anticodon supply does not substantially change these numbers.

We can compare these correlations between mRNA codon demand and tRNA anticodon supply with the correlations we find between our simulated mRNA and tRNA transcriptomes. To calculate correlations for the simulated transcriptomes, I first determined the means for each of the 100 shuffled triplet codon distributions and calculated their Spearman rank correlation with each of the 100 shuffled isoacceptor distributions.

In fact, correlating all 100 randomly simulated tRNA transcriptomes per tissue with the simulated mRNA transcriptomes yields a distribution of significantly lower rank correlation coefficients (figure 2.9 bottom middle and right).

This result provides further evidence that random variation of tRNA gene expression cannot account for the observed patterns of tRNA gene expression, and that tRNA gene expression must be actively regulated to stabilise the steady abundance of the of tRNA anticodon isoacceptors, matching the triplet codon demand of the corresponding mRNA transcriptome.

2.6. Variable chromatin accessibility may influence tRNA gene transcription

Having established that tRNA gene expression varies in a controlled fashion through development, we were next interested in uncovering the mechanism driving this variation. From what we know about the regulation of protein-coding genes, it seemed likely that local genomic features around each tRNA gene would be implicated in its transcriptional regulation.

Previously published results indicate that there is no clear relationship between sequence variation of the internal promoters of a tRNA gene and their expression levels [Oler & al., 2010; Canella & al., 2012]. We therefore focussed on the sequence upstream of the TSS of tRNA genes to search for *cis*-regulatory regions.

To this end, I collected the sequence on the forward and reverse strand of the 500 bp upstream regions of tRNA genes that were differentially expressed between each pair of developmental stages. These sequences were cleaned of low-complexity regions using the dust application [Bailey & al., 2009]. Motif enrichment analysis in the sequences was conducted with MEME [Bailey & al., 2009], configured to search for zero or one occurrences of one motif per sequence, up to a maximum of three distinct motifs, with a minimum motif size of 6 bp. A first-order Markov model built from the upstream regions of all nondifferentially expressed tRNAs in the appropriate stage–stage contrast was used as background.

Subsequently, TOMTOM [Gupta & al., 2007] was used to search for motifs enriched in the MEME output by exploiting databases of known TF binding sites. I used the databases `JASPAR_CORE_2009_vertebrates` and `uniprobe_mouse`. A minimum overlap of 5 bp with an E-value threshold of 10 was required. Although we found 4 significantly enriched motifs in total (at the 5 per cent significance threshold after correcting for multiple testing), these motifs were not present consistently across stages (they are enriched for differential expression between non-adjacent stages), which would be necessary to explain the differential expression we observe (table A.3). This suggests that the upstream region of tRNA genes does not contain known regulatory sequences explaining our observations.

In the absence of clear evidence for nearby TF binding sites driving differential expression, we hypothesised that the transcriptional regulation of nearby protein-coding genes might influence tRNA gene expression. I therefore went on to look for enrichment of differentially expressed (*DE*) tRNA genes in close vicinity to differentially expressed protein-coding genes: an enrichment of DE over non-DE tRNA genes near DE protein-coding genes may indicate that regulation of tRNA and protein-coding differential gene expression is driven by common factors.

A test for colocalisation of differentially expressed tRNA genes and differentially expressed mRNA genes was performed between developmental stages (E15.5–P22 in liver and P4–P29 in brain, because those were the contrasts with the largest number of differentially expressed tRNA genes). For each up-regulated tRNA gene i we counted the number of up-regulated protein-coding genes, n_i, and the total number of protein-coding genes, b_i, in a genomic region centred on the tRNA gene of interest. The analysis was performed for different window sizes (10 kb, 50 kb and 100 kb). This allowed us to compute the ratio $r_i = n_i / b_i$. We repeated this analysis for each non-differentially expressed tRNA gene j to obtain the ratio r^*_j. A Kolmogorov–Smirnov test was performed to assess whether the distribution of r, corresponding to the ratios of up-regulated protein-coding genes in the vicinity of up-regulated tRNA genes, was significantly different from the distribution r^* in the vicinity of nondifferentially expressed tRNA genes with varying significance thresholds (0.1, 0.05 and 0.01).

As for the case of TF binding sites, I was unable to demonstrate such an association unambiguously: for the contrasts mentioned above (liver E15.5–P22, brain P4–P29), I performed 9 tests each: one per combination of significance threshold and window size. The unadjusted p-values for liver are given in table 2.2. Although some of the values fall below common thresholds for significance, to properly reject the hypothesis we would require a more consistent picture; instead, modifying the parameters changes the outcome drastically. The best we can conclude is that for a window size of 50 kb there seems to be some evidence of an effect, if we choose our significance threshold for differential expression of protein-coding genes stringently. However, in brain no such effect exists (figure A.14). When controlling for multiple testing by applying Bonferroni correction, only 1 out of the 18 tests across both tissues is significant (corrected $p < 0.013$).

Window size [kb]	Threshold	p-value
10	0.1	0.285
10	0.05	0.358
10	0.01	0.409
50	0.1	0.013
50	0.05	0.005
50	0.01	0.001
100	0.1	0.118
100	0.05	0.107
100	0.01	0.10

Table 2.2: **Unadjusted p-values of colocalisation tests in liver.** The first column gives the window size (in kb) in which colocalised protein-coding genes were counted. The second column give the p-value significance threshold below which protein-coding genes were called differentially expressed.

Besides *cis*-regulatory *sequence* features, another possibility is that chromatin modifications are associated with the changes in tRNA gene expression that we observed. Previous studies by Barski, Chepelev, & al. [2010] and Oler & al. [2010] indicate that several chromatin modifications have an influence on pol III-driven transcription. Using publicly available, previously published data [Shen & al., 2012] (Gene Expression Omnibus (*GEO*) accession GSE29184), I investigated three histone modifications associated with genomic regions containing promoters and enhancers (H3K4me3, H3K4me1, H3K27ac) as well as pol II and an insulator, CCCTC-binding factor (*CTCF*). For each of these factors, I assayed their association with

1. active versus inactive tRNA genes in embryonic (E15.5) and adult (P29) tissues; and
2. differentially expressed tRNA genes between E15.5 and P29

in both liver and brain.

To test for association of the ChIP factors with tRNA gene expression, I noted whether a signal for a given ChIP target was present in the vicinity of a tRNA locus. I then compared the number of expressed versus unexpressed tRNA genes with at least one such ChIP signal in its vicinity, using Fisher's exact test, with the contingency table shown in table 2.3.

	ChIP signal	
tRNA	Present	Absent
Expressed	*a*	*b*
Unexpressed	*c*	*d*

Table 2.3: Contingency table of ChIP signal enrichment.

Occurrence of these chromatin marks was measured 0.1 kb, 0.5 kb and 1 kb upstream of and downstream from tRNA genes. Our embryonic (E15.5) and adult (P29) pol III data was complemented with embryonic (E14.5) and adult (P56) ChIP-seq data from the Shen & al. [2012] study. Although these stages do not match precisely, similar patterns of expression are present and should thus be complemented by similar patterns of histone marks. In addition, the Shen & al. [2012] data split adult brain tissue up into "cerebellum" and "cortex". I merged these two data sets for the subsequent comparison against our whole brain P29 samples by using the union of ChIP-seq binding locations.

In my test of active against inactive tRNA genes I unsurprisingly found strongly significant enrichment of histone marks for active transcription in both embryonic and adult tissues. I also found association of active tRNA genes with pol II binding and with CTCF. Enrichment of enhancer marks, by contrast, was not present — but this may simply be due to the fact that enhancers are typically more distal (table 2.4). This confirms previous findings [Barski, Chepelev, & al., 2010; Oler & al., 2010].

	Developmental stage	
Factor	Embryo	Adult
Liver		
H3K4me3	1.85×10^{-32}	1.27×10^{-29}
Enhancer	1.00	1.46×10^{-2}
H3K27ac	9.00×10^{-33}	1.99×10^{-21}
Pol II	9.38×10^{-5}	1.85×10^{-4}
CTCF	2.24×10^{-3}	2.09×10^{-4}
Brain		
H3K4me3	3.11×10^{-28}	3.18×10^{-32}
Enhancer	3.11×10^{-1}	1.00
H3K27ac	8.49×10^{-21}	1.16×10^{-14}
Pol II	3.93×10^{-13}	1.35×10^{-12}
CTCF	2.33×10^{-5}	2.76×10^{-5}

Table 2.4: Enrichment of different ChIP targets near active tRNA genes. Shown are the unajusted *p*-values for the hypothesis of no enrichment of a ChIP signal near expressed tRNA genes, compared to unexpressed tRNA genes, using a window size of ±0.5 kb; other window sizes show less evidence for enrichment.

Next, I performed the same test specifically for upregulated DE genes between E15.5 and P29. Rather than considering all expressed tRNA genes as before, I thus only consider genes that are specific to either embryonic or adult tissue. I found consistent significant enrichment of H3K27ac and, to a lesser extent, H3K4me3 and pol II in embryonic and adult liver. In contrast, brain shows no enrichment.

Though limited, this association of differentially expressed tRNA genes histone marks indicates that the accessibility of the chromatin may have an influence on tRNA gene expression, and that the observed differences may be partially influenced by changing histone modification status through the course of tissue development.

	Developmental stage	
Factor	Embryo	Adult
Liver		
H3K4me3	1.77×10^{-2}	1.44×10^{-9}
Enhancer	1.00	6.83×10^{-3}
H3K27ac	5.17×10^{-8}	1.86×10^{-9}
Pol II	8.86×10^{-3}	8.96×10^{-2}
CTCF	1.08×10^{-2}	2.42×10^{-2}
Brain		
H3K4me3	7.65×10^{-1}	5.53×10^{-1}
Enhancer	5.36×10^{-1}	1.00
H3K27ac	7.98×10^{-2}	7.37×10^{-1}
Pol II	7.16×10^{-1}	2.32×10^{-1}
CTCF	1.00	1.00

Table 2.5: Enrichment of different ChIP targets near differentially expressed tRNA genes. Shown are the unajusted *p*-values for the hypothesis of no enrichment of a ChIP signal near differentially expressed, upregulated tRNA genes, compared to non-DE tRNA genes, using a window size of ±0.5 kb; other window sizes show less evidence for enrichment.

2.7. tRNA anticodon isoacceptor families are transcriptionally compensated across development

The results thus far demonstrate that tRNA gene expression varies across development, and this variation follows clear patterns, which require active regulation. We furthermore find that variability within the transcribed tRNA pool vanishes at the isoacceptor level: the tRNA genes within each anticodon isoacceptor family vary across developmental stages, but the sum of their expression is stable (figure 2.9 centre left).

This might imply (anti-)correlation of expression across stages between the genes of an isoacceptor family. Alternatively, tRNA gene expression might vary randomly without regard to other tRNA genes in the same isoacceptor family. An example of each of these two scenarios is shown in figures 2.10a and 2.10c). To test this systematically, we compared the distribution of correlations between genes within each isoacceptor family with a background distribution. The background was generated by permuting the order of the stages before calculating the tRNA gene expression correlations. Importantly, these background distributions have a unimodal shape centred on 0 (figures 2.10b and 2.10d). This allows us to test whether the observed correlations significantly diverge from the background model:

For each isoacceptor that is encoded by more than two tRNA genes, I calculated Spearman's rank correlation (across developmental stages) between the expression values of each pair of its corresponding tRNA genes, i.e. I calculate

$$r_{ij} = \operatorname{cor}(x_{i\cdot}, x_{j\cdot}) \text{ for } i, j \in T, i < j, \tag{2.8}$$

where T is the set of tRNA genes in the isoacceptor family, and $x_{i\cdot}$ is the vector of expression values of the ith tRNA gene across all stages of development. For the same set of genes, I calculated a null set of correlations as follows:

$$b_{ijk} = \operatorname{cor}(\operatorname{perm}_k(x_{i\cdot}), x_{j\cdot}) \text{ for } i, j \in T, i < j; k \in 1 \dots |x_{i\cdot}|!\,. \tag{2.9}$$

Here, $\operatorname{perm}_k(x_{i\cdot})$ is the kth permutation of the vector $x_{i\cdot}$.

Next, we used a χ^2-test to investigate whether there was a significant difference between the background b and the observed correlation distributions c. We only performed the test for the 27 isoacceptor families with six or more genes, since isoacceptor families with less than six genes did not contain enough points for meaningful interpretation.

The distribution of observed correlations in some cases has a bimodal shape, which can be clearly distinguished from the unimodal background (figure 2.10b). In total, 16 out of 27 isoacceptor families (59 per cent) with more than five genes show significantly different foreground and background distributions (all FDR-corrected $p < 0.0199$, see table 2.6).

The bimodal shape of the correlation distribution can be interpreted as the existence of two distinct clusters of tRNA genes within the isoacceptor families, which compensate for each others' expression changes. However, these clusters of genes do not form genomic clusters, i.e. the tRNA genes within each cluster are not closer to one another than to other clusters.

To establish this, I defined 69 clusters of all genomically annotated tRNA genes that lie within 7.5 kb of each other. I counted how many active tRNA genes of an isoacceptor family colocalised in a genomic cluster with tRNA genes of the same isoacceptor family, before calculating the fraction of tRNA genes for each isoacceptor family belonging to a genomic cluster. To test whether genes in isoacceptor families tend to genomically colocalise more than expected by chance, we randomly assigned tRNA genes to isoacceptor families (preserving the actual isoacceptor family gene numbers) 1000 times. I then tested whether the mean percentage of clustering tRNA genes per isoacceptor family differed from the mean percentage expected by chance, by using a binomial test. Finally, I tested whether there was a difference in these percentages between isoacceptor families that show evidence for compensation, and isoacceptor families that show no such evidence by applying a χ^2-test.

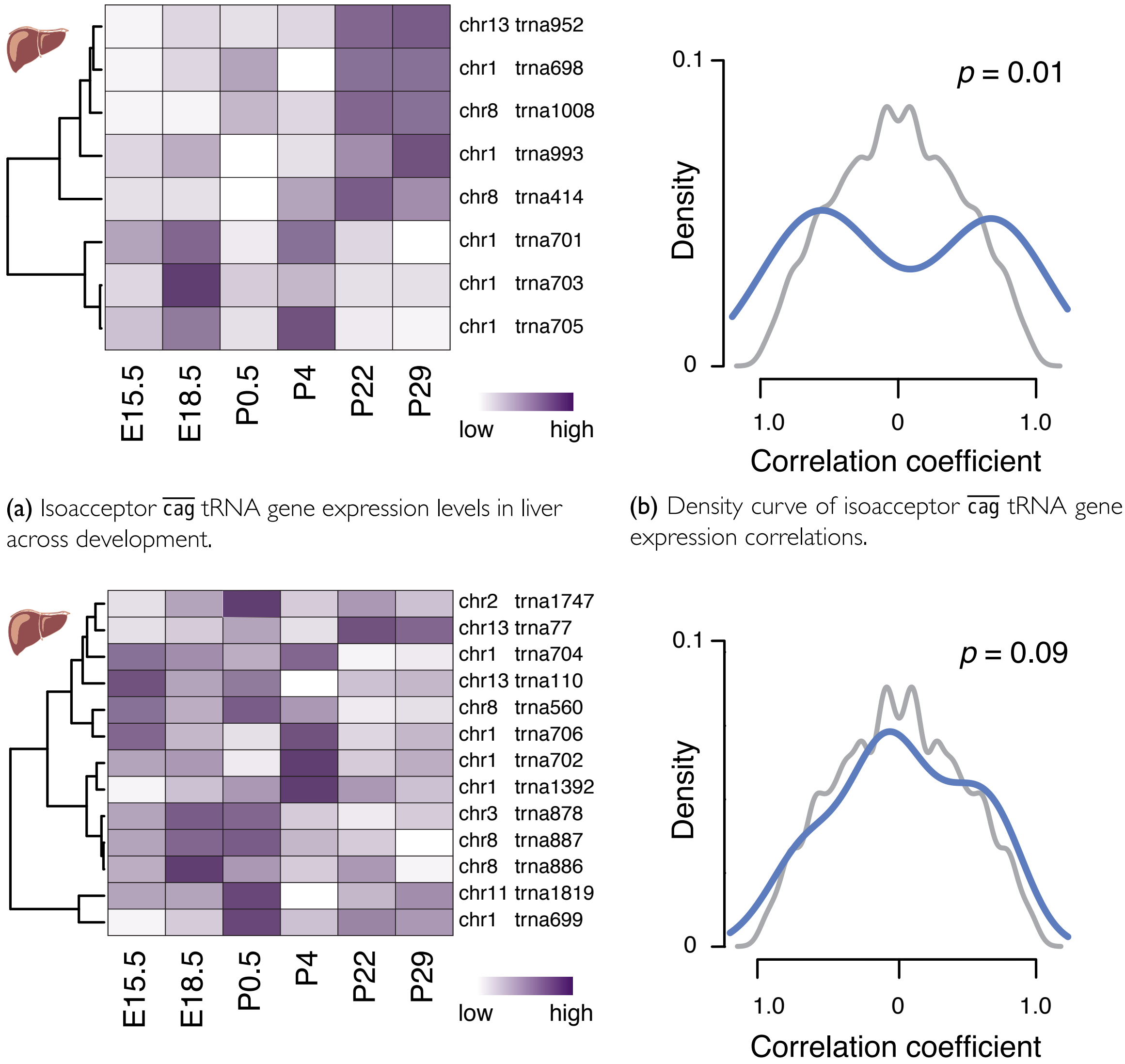
❖

In summary, we have shown that tRNA gene expression varies pervasively across mouse development in different tissues. This variation follows concerted patterns that provide evidence of specific regulation of tRNA gene transcription. Although we have been unable to pinpoint a mechanism for the specific gene expression patterns we observed, there is a broad correlation between tRNA gene activity and the existence of histone marks for active gene expression. The precise purpose of the tRNA gene regulation remains similarly unclear, since we found that both the codon usage and the anticodon isoacceptor abundance are stable across development, and thus do not need to be adjusted using specific tRNA gene expression changes.

Indeed, when looking at individual anticodons, we found that tRNA genes within many isoacceptor families are acting in concert to compensate for changes in each others' expression, with the net result of producing a stable abundance of tRNA molecules of the anticodon isoacceptor. The regulatory mechanism enacting this compensation has thus far not been described, and its identification poses a new challenge.

Isoacceptor	p-value
$\overline{\text{gug}}$	1.01×10^{-25}
$\overline{\text{agc}}$	2.26×10^{-12}
$\overline{\text{gca}}$	9.17×10^{-11}
$\overline{\text{cca}}$	1.15×10^{-10}
$\overline{\text{cug}}$	3.60×10^{-9}
$\overline{\text{ugg}}$	3.68×10^{-7}
$\overline{\text{aac}}$	3.68×10^{-7}
$\overline{\text{gua}}$	5.69×10^{-6}
$\overline{\text{cau}}$	8.53×10^{-6}
$\overline{\text{uuc}}$	2.42×10^{-5}
$\overline{\text{ugc}}$	2.56×10^{-5}
$\overline{\text{aga}}$	6.57×10^{-5}
$\overline{\text{cuc}}$	6.81×10^{-5}
$\overline{\text{cac}}$	2.06×10^{-4}
$\overline{\text{guc}}$	8.76×10^{-4}
$\overline{\text{cag}}$	1.99×10^{-2}
$\overline{\text{guu}}$	6.01×10^{-2}
$\overline{\text{agu}}$	1.26×10^{-1}
$\overline{\text{gcc}}$	1.33×10^{-1}
$\overline{\text{uuu}}$	3.35×10^{-1}
$\overline{\text{gcu}}$	3.45×10^{-1}
$\overline{\text{aau}}$	3.45×10^{-1}
$\overline{\text{gaa}}$	6.50×10^{-1}
$\overline{\text{acg}}$	8.01×10^{-1}
$\overline{\text{ucc}}$	8.01×10^{-1}
$\overline{\text{cuu}}$	8.12×10^{-1}
$\overline{\text{agg}}$	8.44×10^{-1}

Table 2.6: Evidence against absence of compensation. The first column contains the tRNA anticodon isoacceptor families. The second column contains the FDR-adjusted p-values of H_0: there is no effect of the order of the stages on coordinated gene expression of tRNA genes within an isoacceptor family.



(a) Isoacceptor $\overline{\texttt{cag}}$ tRNA gene expression levels in liver across development.

(b) Density curve of isoacceptor $\overline{\texttt{cag}}$ tRNA gene expression correlations.

(c) Isoacceptor $\overline{\texttt{gcc}}$ tRNA gene expression levels in liver across development.

(d) Density curve of isoacceptor $\overline{\texttt{gcc}}$ tRNA gene expression correlations.

Figure 2.10: tRNA gene expression is compensated at the anticodon isoacceptor level during mouse development. Panels (a) and (c) show two examples of tRNA gene isoacceptor families and their gene expression across development (row-normalised). In panel (a) we can see two clusters of coordinated expression: the top 5 genes are lowly expressed at first, and start being highly expressed at P22. The bottom 3 genes show a roughly opposite trend. In panel (c), no such clusters are obviously present. Panels (b) and (d) show the corresponding pairwise gene–gene correlation coefficients, plotted as a density curve (blue), as well as the density curve of the background distribution.

3 Implications of codon–anticodon interaction on the regulation of translation

In the previous chapter I have shown that codon usage and its interaction with tRNA anticodons remains remarkably stable despite substantial variability of the transcriptome during mammalian development.

Shortly after the publication of our research on tRNA gene regulation during mouse development, Gingold, Tehler, & al. [2014] published *"A Dual Program for Translation Regulation in Cellular Proliferation and Differentiation"*. They report that cells with different functional characteristics preferentially use distinct sets of codons, and that the pool of active tRNAs adapts dynamically to decode this set of codons with high efficiency.

Because its conclusions are highly relevant to our own research we evaluated the paper carefully. In the following, I will first describe the paper's main findings to the extent that they are relevant to our own research. I will then discuss new experimental data and computational analyses we performed to explore how our previous results relate to this paper.

3.1. "A dual program for translation regulation in cellular proliferation and differentiation"

Gingold, Tehler, & al. [2014] investigated the abundance of tRNA anticodons and the codon usage of different groups of protein-coding genes in patient tissue samples and human-derived cell lines in different cellular conditions — five primary cancers, induced differentiation, release from serum starvation, senescence, and *MYC* and *RAS* overexpression — with the aim of characterising the differences in tRNA gene expression and tRNA anticodon abundance. They hypothesise that the balance between tRNA anticodon supply and codon demand might influence the rate of production of proteins from mRNA [Gingold and Pilpel, 2011].

3.1.1. tRNA anticodons change in abundance in tumour tissues

Gingold, Tehler, & al. [2014] designed custom microarrays to probe the expression levels of the tRNAs of most anticodons present in the human genome, excluding those prone to cross-hybridisation, or where a low in silico screening score predicts that they may be pseudogenes. Using these microarrays, they assayed the abundance of anticodon isoacceptors in healthy B cells and in B-cell derived lymphomas. They showed that there are specific anticodon isoacceptors whose abundance changes reproducibly in tumours compared to healthy tissue, while other anticodon isoacceptors do not change their abundance. This mirrors results reported previously [Pavon-Eternod, Suzanna Gomes, & al., 2009].

In addition, they used their array-based assays to determine tRNA anticodon abundance levels for all samples and applied hierarchical clustering to the resulting matrix of tRNA expression profiles. They observed that samples clustered into two groups, corresponding to tumour cell types and differentiating cell types. When performing an equivalent clustering on mRNA gene expression values for the same samples, they observe that samples cluster not by whether they are tumours, but rather by their tissue of origin, i.e. by the tissue from which the tumour or cell line was derived.

3.1.2. Codon usage differs between genes involved in cell proliferation and genes involved in differentiation

Next, Gingold, Tehler, & al. [2014] looked at protein-coding genes within Gene Ontology (GO) terms that they functionally associated with healthy, adult tissue (“pattern specification process”) on the one hand, and tumour tissue (“M phase of mitotic cell cycle”) on the other hand. They show that the codon usage bias in these two GO terms (averaged over all containing genes) clearly differs (figure 3.1): the genes in both GO terms preferentially use different synonymous codons.

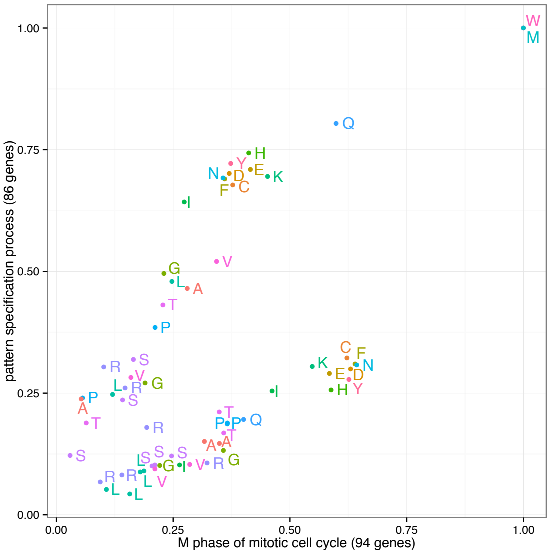


Figure 3.1: **Codon usage bias in two GO terms.** Each point represents one codon, whose corresponding amino acid is given by the label. The position of the point is given by the codon usage bias in each GO term, respectively. Codons for the same amino acid share the same colour (reproduction of figure 2A in Gingold, Tehler, & al. [2014]).

They expand their analysis by calculating the mean codon usage across many GO categories and perform PCA on the resulting matrix (figure 3.2). This reveals that the largest contributor to the variation stems from the split of the GO terms into two distinct sets, one encompassing multi-cellular GO terms and the other cell autonomous GO terms. They argue that these two sets of GO terms correspond to GO terms functionally responsible for maintaining cell homoeostasis on the one hand, and rapid cellular division, such as that found in tumours, on the other hand. In other words, gene families specific to rapidly dividing as opposed to healthy, mature cells have distinct codon composition.

It is however worth noting that the choice of GO terms highlighted in the PCA is somewhat arbitrary: It is for instance not clear why "translation" should be more active during cellular division than in stable cells. The authors rather describe the relevant set as "cell autonomous" GO terms, but the paper's argument implicitly assumes that this is associated with cell division. Furthermore, the allocation of genes to GO terms was performed via simple textual matching, such that GO sub-categories whose description contains the text "differentiation and proliferation" would be counted as belonging to the GO super-category "differentiation".

A better strategy for the selection of GO terms would ensure that allocated uniquely to a single GO term without overlap between the sets, to ensure pairwise independence between the data points: because the GO hierarchy allows individual GO terms to have multiple parents, a naive GO term selection can lead to child terms (and hence genes) present in several categories. This can be corrected by applying a GO trimming algorithm (e.g. Jantzen & al. [2011]).

Furthermore, GO terms generally do not distinguish between positive and negative influence of a gene on a phenotype. As a consequence, the GO term "M phase of mitotic cell cycle" contains both activators and inhibitors of M phase, and the containing genes' expression may be anticorrelated.

3.1.3. Differential codon usage in cell-condition specific gene sets matches tRNA anticodon abundance in corresponding cells

Finally, Gingold, Tehler, & al. [2014] compared the PCA of the per-GO codon usage to the actual gene expression of mRNAS and tRNAS in different cellular conditions. The analysis was done in two parts. Firstly, the fold change of gene expression between normal tissue and a given condition was calculated for each protein-coding gene. The mean expression fold change of each GO term, mapped to a colour map, was plotted on top of the GO terms in the PCA (figure 3.3 bottom). This was done separately for various conditions, including several tumour samples and cell lines with induced differentiation. In all examined conditions, there seems to be a marked gradi-

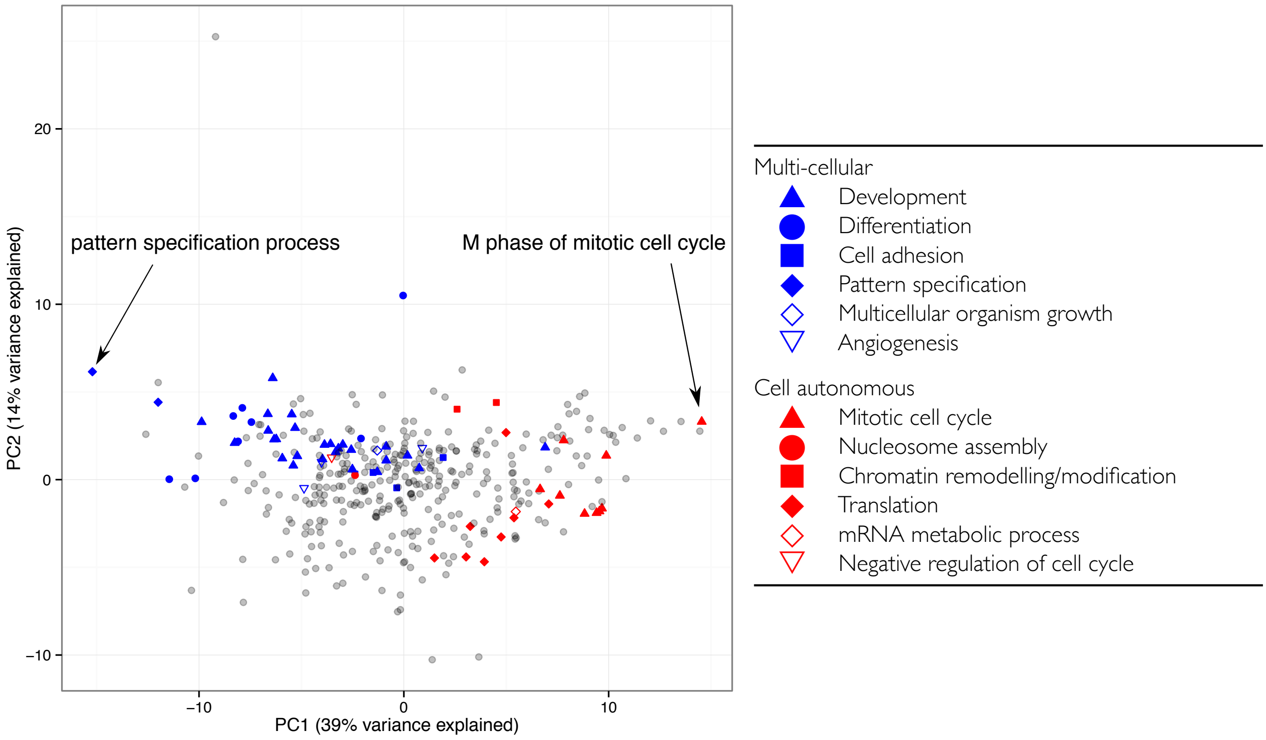


Figure 3.2: PCA of mean GO term codon usage. Each dot corresponds to one GO term. The position of the dots is derived by rotating the matrix of the mean codon usage of the genes belonging to each GO term. Created using methods of Gingold, Tehler, & al. [2014] (the precise numbers differ slightly due to the use of different implementations to perform the analysis but this does not impact the interpretation).

ent of mRNA expression enrichment from left to right. The authors used this to suggest that the first principal component of the PCA separates GO terms by their specificity to either proliferating or differentiating cellular programmes.

Secondly, to estimate translation efficiency they calculated the fold change of tRNA gene expression for isoacceptor tRNAs that decode the codons of the genes in each GO term, and averaged across them, weighted by codon usage. Adopting the notation for anticodon abundance from chapter 2, we can define the fold change in translation efficiency e_{AB} between two conditions A, B for a multiset of codons C as:

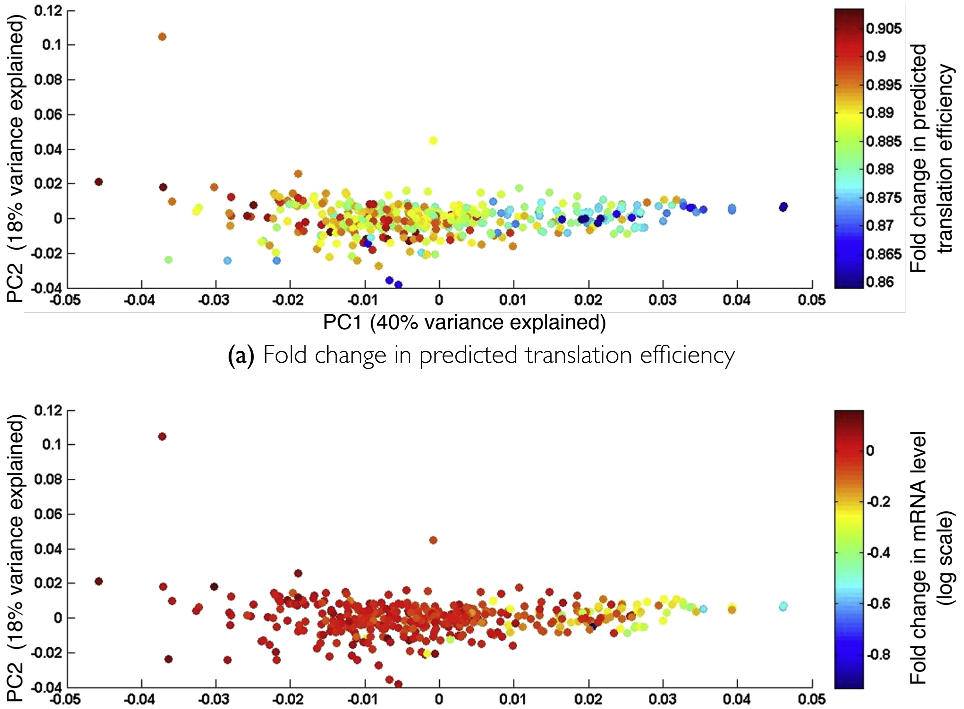
$$F = \frac{a_{\cdot A}}{a_{\cdot B}} \tag{3.1}$$

$$e_{AB} = \sum_{c \in C} \frac{1}{|d(c)|} \sum_{f \in F_{d(c)}} f \ , \tag{3.2}$$

where $d(C)$ is the set of anticodons that can decode the codon c.

The fold change in translation efficiency between normal tissue and the sample under consideration was again mapped to a colour gradient, which was overlaid on the PCA (figure 3.3 top). Again there appears to be a gradient of how well different GO terms are adapted to the tRNA anticodon pool of a given condition, following the first principal component of the PCA.

Although interesting, the authors did not calculate correlations between the first principal component and either the codon usage bias or the translation efficiency fold change. Instead, they visualised the presumed correspondence using a potentially misleading rainbow colour map [Borland and Taylor, 2007]. Furthermore, the absolute range of changes is extremely small compared to the overall fold change (range of fold change 0.86 to 0.905 in one case, corresponding to less than 5 per cent of the maximum fold change), and the lack of statistical analysis makes it impossible to tell whether these changes are in fact significant. It is also important to note that the entire fold change range in predicted translation efficiency is < 1: Consequently, GO categories that are specific for a given condition according to figure 3.2, and should thus be better adapted to the tRNA anticodon abundance, are in fact less well adapted according to this analysis (figure 3.3). This is a recurring feature for all conditions tested in this way, and is potentially inconsistent with the claim that the tRNA anticodon pool is specially adapted to cell condition specific gene sets.



(a) Fold change in predicted translation efficiency

(b) Fold change in mRNA level

Figure 3.3: Projection of the tRNA and mRNA expression changes onto the codon usage map for cells after induced differentiation via retinoic acid. Both panels show the same PCA reproduced in figure 3.2. In the top panel, the colours correspond to the fold change in predicted translation efficiency of the genes constituting each GO term, given the change in cellular abundance of tRNA anticodons compared to normal cells. The bottom panel shows the mean fold change in mRNA levels per GO term compared to normal cells. Figure modified from Gingold, Tehler, & al. [2014].

3.2. Are tRNA anticodon abundance and codon usage highly adapted to different cellular conditions?

To explore how these two findings — the stability of the anticodon pool through development on the one hand [Schmitt, Rudolph, & al., 2014], and the malleability of the anticodon pool to match the codon demand of highly expressed protein-coding genes on the other hand [Gingold, Tehler, & al., 2014] — we examined differences between healthy tissue and tumour cell lines in *M. musculus* and *H. sapiens*.

3.2.1. The effect of gene set size on codon usage bias

Since we were already in possession of some relevant tRNA data we decided to use our own data to recapitulate these findings. A first observation was that we had looked at the whole transcriptome, while Gingold, Tehler, & al. [2014] had looked at specific subsets. Although both approaches are valid, some caution is necessary when comparing the results: the smaller the set of genes, the larger the effects of random sampling of the genes become. Subsequently, when analysing a particular feature, such as codon usage bias, random sampling will contribute proportionally more to the variation between two small sets than to differences between two larger sets. Consequently, if one wants to assert that deviations are nonrandom, one thus has to account for this effect.

To assess how much of the variation in GO term codon usage is explainable by such effects, I generated random sets of different numbers of genes, corresponding to the whole range of GO term gene set sizes. For each simulated gene set, I calculated the total codon usage of the genes in the resulting set. To ensure that the codon usage of different gene sets was comparable, all codon frequencies within a sample were normalised by the total codon frequency in the sample, to obtain codon proportions:

$$x_i^* = \left(\sum_{g \in G} x_{ig}\right) \Big/ \left(\sum_j \sum_{g \in G} x_{jg}\right) \tag{3.3}$$

Where $i, j \in \texttt{AAA}, \ldots, \texttt{TTT}$ are the codons (excluding stop codons), $g \in G$ are the genes in the sample, x_{ig} is the codon frequency of codon i in gene g, and x_i^* is the relative codon frequency of codon i in the sample. We have $\sum_i x_i^* = 1$.

This was done 10 000 times for each GO term gene set size. Next, I calculated the Pearson correlation between the codon proportions of each sample with the genome-wide codon proportions (across all annotated coding sequences). Figure 3.4 plots the distribution of the correlations against the

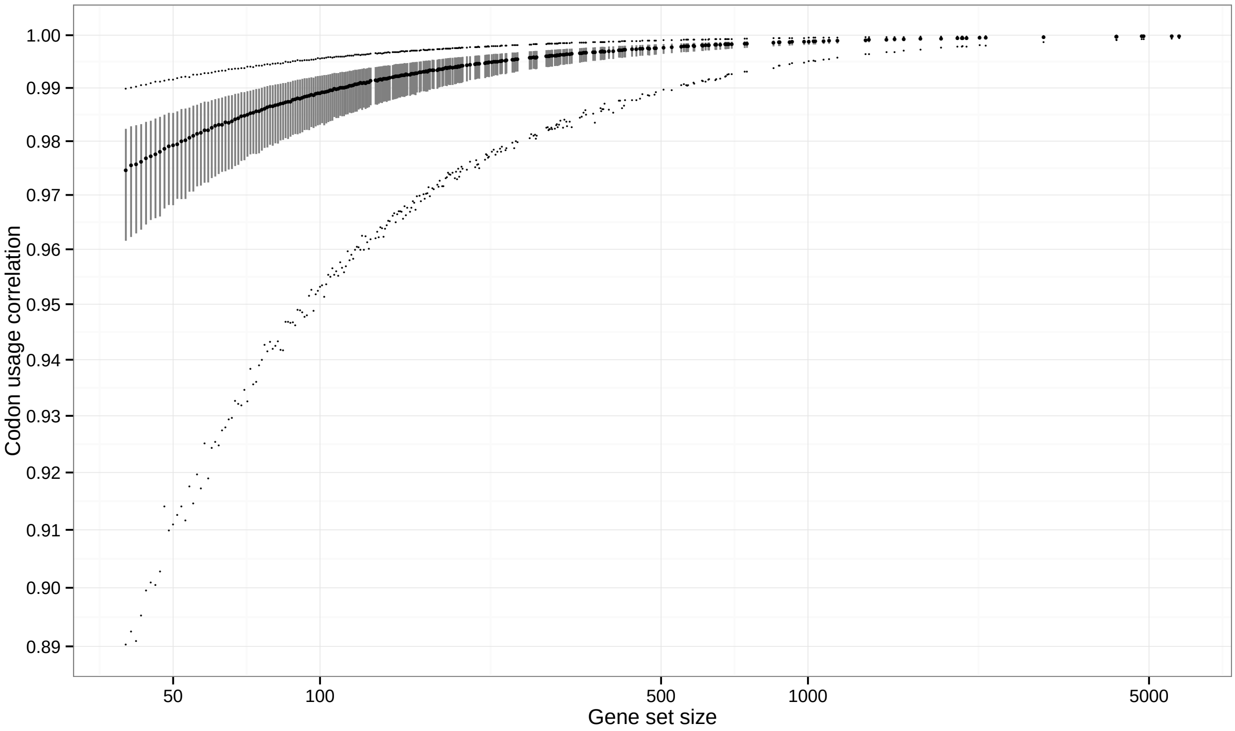


Figure 3.4: Dependence of codon usage variability on sample size. Genes were randomly sample from the human genome to create sets of sizes given by x (in grey). The mean codon usage of the sets was calculated, and their Pearson correlation to the genomic background is shown on the y axis. For each size, the distribution of 10 000 repeated samples is summarised in the figure by a vertical segment connecting the first with the third quartile; the mean is indicated as a black dot. The second and ninety eighth percentile are hinted at with smaller dots.

gene set sizes.

I used these distributions to calculate empirical p-values for each GO term in order to test the null hypothesis H_0: the divergence in codon usage from a background distribution is indistinguishable from random variation for this GO term. Figure 3.5 plots the codon usage variation of each GO term, along with an indication of whether the observed variation is explained by random variation alone.

The plot illustrates that the variation of many GO terms is expected by

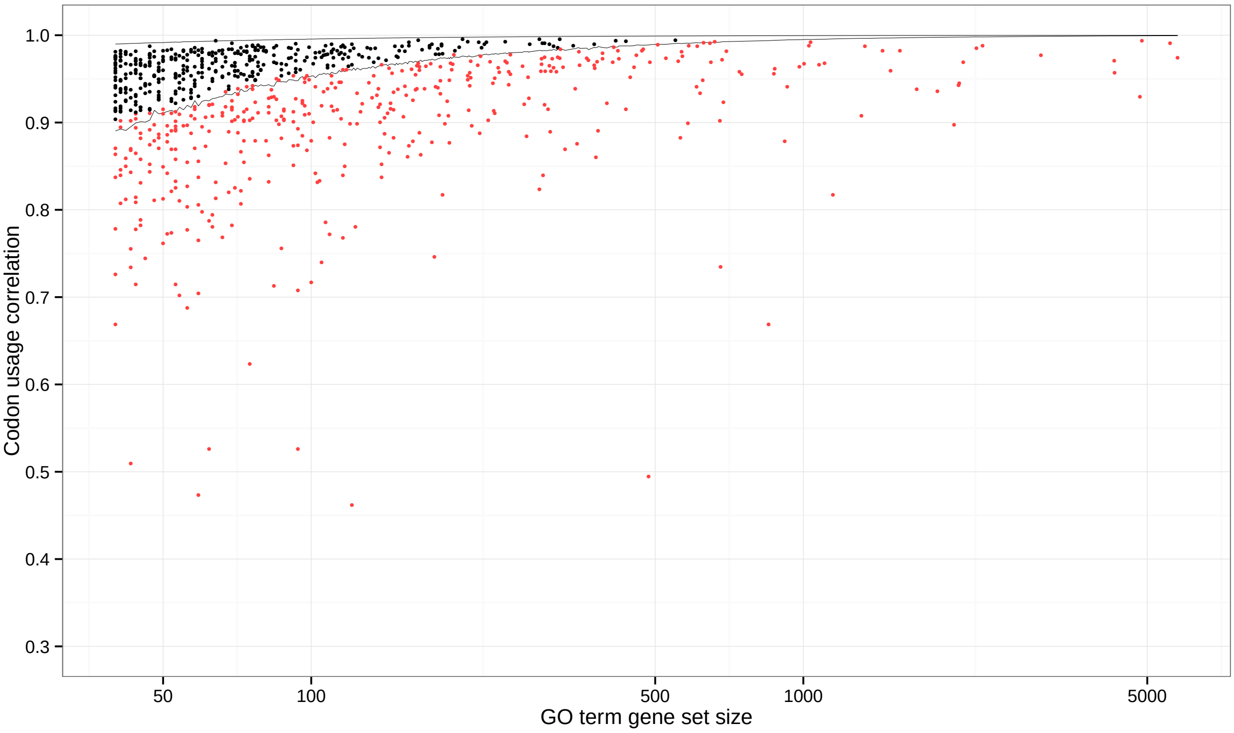


Figure 3.5: GO term codon usage variation. Each point corresponds to a GO term whose Pearson correlation of codon usage with the genomic background is plotted against its gene set size. GO terms with an FDR-adjusted empirical $p < 0.05$ are coloured in red. The continuous lines denote the second and ninety eighth percentile of the simulated data from figure 3.4.

chance; however, more than half (421 out of 791, 53 per cent) show significantly higher variation. Thus, despite an influence of random sampling especially for small gene set sizes, this observation is unlikely to explain the variation seen in figure 3.2.

3.2.2. The extent of anticodon adaptation to distinct cellular conditions

Next, in order to test whether the anticodon pool adapts to specific cellular conditions as suggested by figure 3.3, I examined whether the codon–anticodon adaptation is better between matching than between mismatching mRNA and tRNA transcriptomes. For this purpose, I calculated codon–anticodon adaptation as the correlation between matching codons and anticodons, each as a proportion of the contribution to their corresponding amino acid — in other words, the correlation between RCU and RAA. I similarly tested whether the codon–anticodon adaptation was better for genes that are specific to a cellular condition, compared to the overall transcriptome. In other words, I looked at the mRNA pool of each cellular condition, and calculated the codon–anticodon correlations for

1. the whole transcriptome and its matching tRNA pool ("matching"),
2. the whole transcriptome and all mismatching tRNA pools ("mismatching"),
3. the set of differentially, highly expressed mRNA genes and the tRNA pool of the same cellular condition ("DE"), and
4. a condition specific gene set (for the GO terms "M phase of mitotic cell cycle" & "pattern specification process") and the tRNA pool of the same cellular condition ("GO").

I performed the analysis using data generated from *M. musculus* and *H. sapiens*. For both species, I compared healthy tissue (adult whole-liver tissue, homogenised) to two different hepatocyte derived cancer cell lines: HepG2 and Huh7 in *H. sapiens*, and Hepa1-6 and Hepa1c1c7 in *M. musculus*. At least two biological replicates were present for each sample.[*] Distinct tumours were used because we wanted to test the hypothesis that there exists a tRNA abundance adaptation specific to the cellular programme in proliferating cells, which would be present across different tumour types (figure 3.6).

The analysis compares different scenarios that should, according to the hypothesis proposed by Gingold, Tehler, & al. [2014], show markedly different codon–anticodon adaptation distributions: codon and anticodon pools

[*] The tissue collection and sequence library preparation was performed by Bianca Schmitt and Diego Villar.

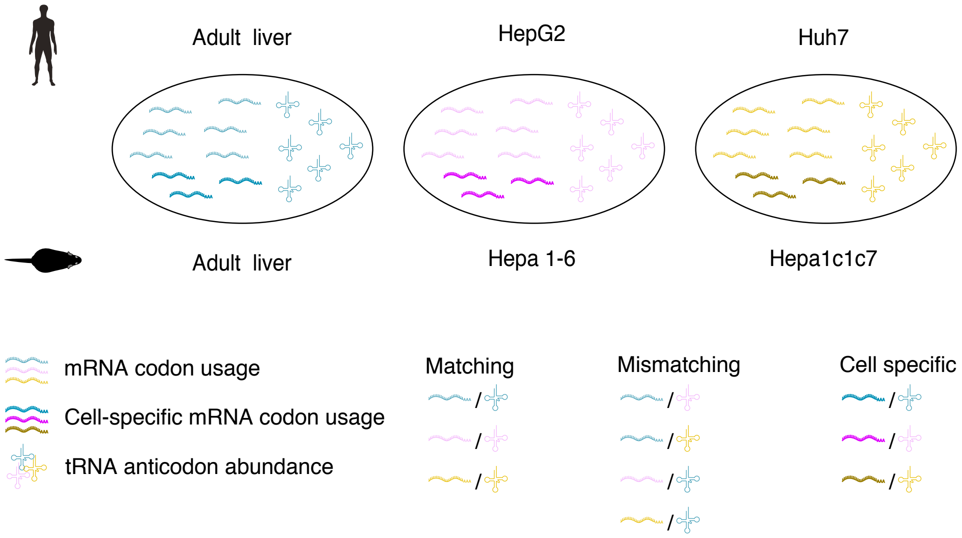


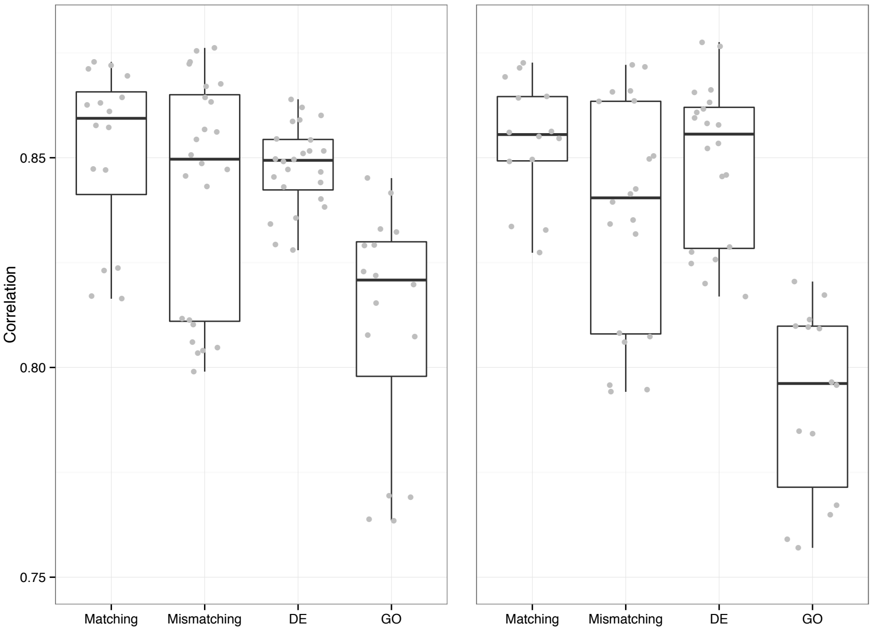
Figure 3.6: Experimental setup. mRNA and tRNA expression was quantified for three different cellular conditions in both human (healthy adult liver, HepG2 & Huh7) and mouse (healthy adult liver, Hepa1-6 & Hepa1c1c7). Each experiment was performed in multiple replicates. Each category ("matching", "mismatching" and "cell specific") corresponds to a comparison between the codons of a gene set and the anticodons of a sample. The cartoon shows how these were combined to obtain the distributions. Each sample condition has a different colour. mRNA transcripts in bolder colours correspond to cell type specific transcripts.

of the same cellular conditions should be highly coordinated (correlation type "matching"). Codons and anticodons of *mismatching* conditions (i.e. the codon pool of a healthy cell and the anticodon pool of a tumour cell, or vice versa) should be less well coordinated (correlation type "mismatching"). Furthermore, the codon usage of just a subset of the overall transcriptome corresponding to genes specific to the cellular condition should be coordinated with their anticodon pool to an even greater extent. Conversely, if we did not expect a tRNA anticodon pool adapted to specific subsets of the transcriptome, we might expect that the correlation between the codon usage of such subsets and the anticodon abundance was *less* than that of the overall transcriptome, due to higher stochastic variability in smaller gene sets (figure 3.6).

Cell specific subsets of the transcriptome can be defined in several ways. My analysis used two possible definitions: Firstly, I looked at differentially expressed genes between healthy and tumour cells, and called genes condition-specific if they were significantly differentially expressed (adjusted $p < 0.001$), their expression was high (mean expression across conditions in the upper quartile of all mean expression levels), and their fold change between the conditions was also high (taking the 200 genes with the highest fold change; correlation type "DE"). Secondly, I took the gene sets of the two most extreme GO terms found by Gingold, Tehler, & al. [2014] according to the variance in codon usage, which they describe as highly specific for proliferating cells (GO term "M phase of mitotic cell cycle") and differentiating/differentiated cells (GO term "pattern specification process"; correlation type "GO").

Unfortunately, GO term annotations differ vastly between *H. sapiens* and *M. musculus* and, in particular, one of the two condition specific GO terms (according to the PCA analysis, figure 3.2) is badly annotated in *M. musculus* (with only 3 genes annotated for "M phase of mitotic cell cycle"). As a consequence, I did not use the existing GO term annotations for mouse but rather used the orthologous gene sets from the human GO term annotation (determined from the gene name, which is identical in *M. musculus* and *H. sapiens* [Wain & al., 2002]).

We want to test whether the "mismatching" correlations are lower than the "matching" ones and whether the "DE" and "GO" correlations are higher, respectively, than the "matching" ones. I therefore used one-tailed tests for significant deviations from the null. In all but one cases, I failed to reject the null hypothesis of no difference. The only case where we find some support for rejecting H_0 is the comparison of "matching"–"mismatching" in *H. sapiens* ($p = 0.048$, one-tailed Mann–Whitney–Wilcoxon test). On the other hand, the "GO" correlations are significantly *lower* than the "matching" correlations in both species ($p < 4.4 \times 10^{-5}$ in *M. musculus*; $p < 2.5 \times 10^{-8}$ in *H. sapiens*). This offers evidence against the hypothesis put forward by Gingold, Tehler, & al. [2014] (figure 3.7). This is consistent with the alternative hypothesis that the observed variation in codon usage is caused by random variation due to the small sample size of the GO term gene sets, but it may also be due to the issues related to GO term selection that I have described in section 3.1.2.



(a) *M. musculus* (b) *H. sapiens*

	mRNA		tRNA
Matching			
M. musculus	healthy liver	–	healthy liver
	Hepa1-6	–	Hepa1-6
	Hepa1c1c7	–	Hepa1c1c7
H. sapiens	healthy liver	–	healthy liver
	HepG2	–	HepG2
	Huh7	–	Huh7
Mismatching			
M. musculus	healthy liver	–	Hepa1-6
	healthy liver	–	Hepa1c1c7
H. sapiens	healthy liver	–	HepG2
	healthy liver	–	Huh7

Figure 3.7: Codon–anticodon correlations. Each box shows a distribution of codon–anticodon Spearman correlations. A codon–anticodon correlation is computed from relative contributions to the respective amino acid (i.e. RCU versus RAA), ignoring wobble base pairing. "Matching" shows correlations of the codon and anticodon pool of the same condition. "Mismatching" shows correlations of the codon and anticodon pool of mismatching conditions (e.g. Hepa1-6 codons versus healthy liver anticodons); the two cancer replicates are *not* counted as mismatching conditions. "DE" shows correlations of the codon pool of highly, differentially expressed mRNA genes, and the anticodon pool of the same condition. "GO" shows correlations of the codon pool of condition-specific GO term gene sets, and the anticodon pool of the same condition. For all distributions, correlations were calculated between all pairwise combinations of replicates.

It is worth nothing that the "DE" codon–anticodon correlations vary non-negligibly, depending on how exactly the top 200 condition-specific genes are chosen; alternative strategies to the one outlined were not found to change the overall result, however: none of the strategies yielded correlations that were significantly higher than the transcriptome-wide "matching” correlations.

3.2.3. Other genomic features may drive the perceived codon usage bias

The trend seen in figure 3.2 exists, albeit to a lesser degree, when plotting amino acid usage rather than codon usage (figure 3.8). The first principal component separates the GO term “keratinisation” from the rest. This is due to the fact that the proteins in this GO term are uniquely highly enriched in the amino acid proline, which forms more rigid peptide bonds than all other amino acids, with implications for the robustness of, amongst other things, the cellular skeleton. The proteins in this GO term are thus implicated in the structural formation of the epithelium [Hohl *&* al., 1995]. The second principal component, by contrast, shows the same separation of GO categories as the first principal component of figure 3.2. Moreover, these two principal components are highly correlated (Pearson’s $\rho = 0.95$). Clearly, this correlation cannot be explained by *codon* identity, since amino acids, not codons, conventionally determine protein function. However, amino acid usage alone does not explain the pattern either: After removing amino acid identity as a confounding factor (by computing RCU instead of codon usage), the GO term PCA still displays the previously observed separation of GO categories (figure 3.9).

We thus conclude that both changes in codon usage and changes in amino acid usage independently correlate with the separation of GO terms into two functional categories. Furthermore, given the results presented in the previous chapter, we note that we have no evidence for a functional adaptation of the tRNA pool to specific gene sets. I therefore looked for other factors that could explain the observed patterns in codon usage. In some prokaryotic species, gene codon usage can be predicted by the intergenic sequence GC bias [S. L. Chen *&* al., 2004]. Mammalian genomic GC content varies drastically across large regions of the genome. However, within regions called *isochores* GC content is stable, and variation in GC bias between isochores is highly correlated with variation in codon usage of protein-coding genes between isochores [Sharp and Matassi, 1994].

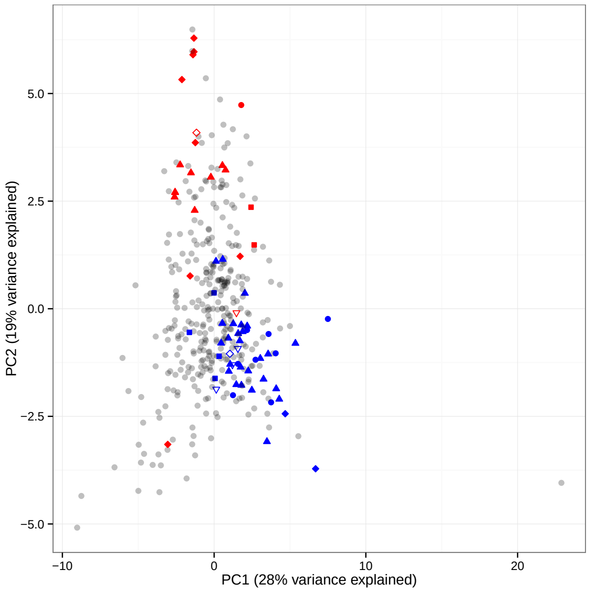


Figure 3.8: **PCA of mean GO term amino acid usage.** Each dot corresponds to one GO term. The position of the dots is derived by rotating the matrix of the mean amino acid usage of the genes belonging to each GO term. Colours and symbols as in figure 3.2.

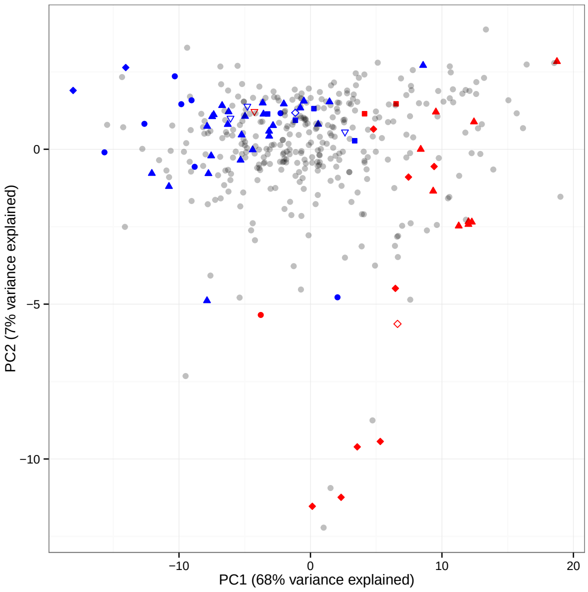
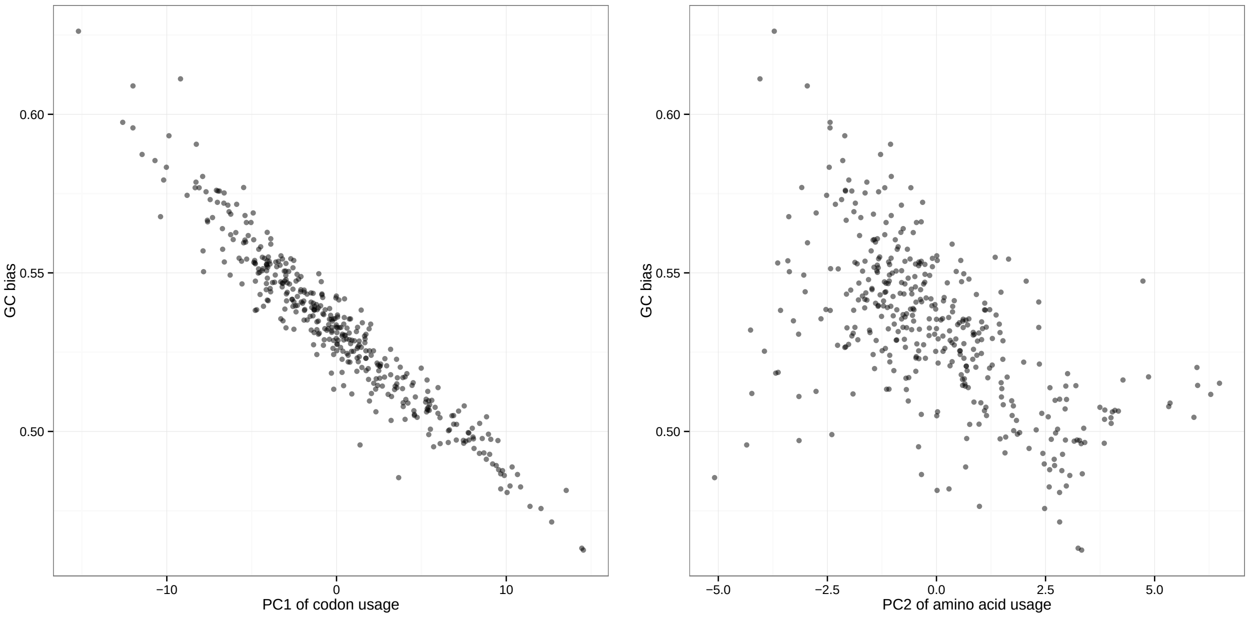


Figure 3.9: **PCA of mean GO term RCU.** Each dot corresponds to one GO term. The position of the dots is derived by rotating the matrix of the mean RCU of the genes belonging to each GO term. Colours and symbols as in figure 3.2.

In fact, we find that the first principal component of the codon usage spread (figure 3.2) as well as the second principal component of amino acid spread (figure 3.8) correlate highly with the GC bias in the GO term gene sets (Pearson's $\rho = -0.96$, $\rho = -0.58$; figures 3.10a and 3.10b). This suggests that genomic features other than gene function might influence gene codon usage. However, it is entirely conceivable that the differences in GC bias of the GO terms are caused by the changes in either amino acid or codon usage, rather than the other way round: any genetic point mutation from `A` or `T` to `C` or `G` (or vice versa) by definition changes the GC content of that region. We thus do not expect changing codon composition to be GC neutral.

To establish whether the driver of this correlation is large-scale GC bias adaptation, changes in protein function and thus amino acid usage or subtle changes in codon usage to regulate translation, we therefore need to examine the GC bias of the intergenic regions surrounding each GO term's genes: not being part of any mRNA transcript, the GC bias of the intergenic regions cannot be driven by codon-level selection. It can thus serve as a background for the expected variation in GC to compare against the GC variation that we see between GO terms. An outline of the necessary investigation to answer this question can be found in the conclusion in chapter 5.



(a) GC bias against codon usage (Pearson's $\rho = -0.96$).

(b) GC bias against amino acid usage (Pearson's $\rho = -0.58$).

Figure 3.10: GC bias plotted against a principal component of the GO term codon & amino acid usage PCA. Each point represents one GO term. The x coordinate corresponds to that of the named principal component of the PCA in figures 3.2 and 3.8. The y coordinate corresponds to the mean GC bias of the GO gene set.

3.3. Discussion

In this chapter I extended the analysis from chapter 2, and contrasted it with the findings by Gingold, Tehler, & al. [2014]. We have moved beyond the variation of tRNA gene expression and have questioned the apparent stability of the codon usage and anticodon isoacceptor abundance in the transcriptome in the face of major changes in cell function. However, contrary to previous reports I was unable to find evidence for a selective adaptation of the tRNA pool to cell-specific subsets of genes. The work presented in this chapter constitutes the beginning of our investigation into the codon usage variability of subsets of the transcriptome. Many questions still remain open.

3.3.1. Codon–anticodon adaptation to distinct cellular conditions

One caveat of the analysis of the codon–anticodon correlation between subsets of the transcriptome is the treatment of replicates. In figure 3.7, biological replicates of each sample were treated independently, rather than pooled: correlations were calculated between all pairwise combinations of RNA-seq and pol III ChIP-seq data of the samples under consideration. Each data point thus corresponds to a pairwise combination of two replicates; this implies that data points within one distribution are *not* independent of each other. The alternative would have been to average over biological replicates, and correlate the averaged data. However, this has two important drawbacks:

1. It does not account for variance in the expression data and thus underestimates the variance in the distributions of the correlations.
2. If mRNA and tRNA gene expression values are dispersed around the same mean, averaging across replicates would yield a higher correlation than actually present in vivo.

The current analysis measured tRNA anticodon adaptation using the codon–anticodon correlation, ignoring wobble base pairing. A potentially better measure of tRNA adaptation is the tAI, which accounts for wobble base pairing in a species-specific manner. Extending the analysis to use the tAI is ongoing work.

Furthermore, our selection of GO terms for proliferation and differentiation may be an inadequate description of cell-type specific gene expression. In addition to concerns already outlined in section 3.1.2, we need to question the assumption that the go term "pattern specification process" is specific to cells from adult liver. In fact, although Gingold, Tehler, & al. [2014] contrast tumour samples with healthy, differentiated cells, they test codon–anticodon adaptation on cells after induced differentiation rather than differentiated cells, which may be better matched by the "pattern specification process" go term. To improve our analysis, I could use go terms resulting from a gene set enrichment analysis of the adult liver samples [Subramanian & al., 2005].

3.3.2. Evolutionary conservation of codon–anticodon adaptation

If a mechanism leading to the adaptation of the cellular trna pool to different cellular conditions existed, we would expect this to be well conserved across mammalian evolution. Indeed, if the trna pool is differentially adapted to the codon usage of subsets of genes, then the codon usage of these genes would be under negative selection, and should thus show conservation. Furthermore, this conservation should therefore be *stronger* than genomic features under neutral selection, such as gc content. Thus, if we can show that codon usage bias of go term gene sets is no more conserved across mammalian evolution than gc bias, this may provide additional evidence against the hypothesis of a functional relevance for gene set specific codon bias.

Using genome data from different mammalian species and homology information for the annotated go categories in humans, we are able to infer how strongly conserved codon usage is compared to the conservation of other genomic features such as gc bias for each orthologous gene. Under the assumption that codon usage is selected for to drive differential translation rates in different cellular conditions, we would expect a comparatively high correlation of codon usage compared to other genomic features.

4 The pol III transcriptome consists of more than just tRNA

4.1. A profile of pol III binding across different features

Less than 5 per cent of the human genome is protein-coding. Much less than 1 per cent of the genome codes for tRNAs [Lander & al., 2001]. Nevertheless, pol III can transcribe a variety of RNA species, with tRNA genes forming only a part of the overall pol III transcriptome. Notable other targets of pol III transcription are the 5S rRNA (a part of the ribosome) as well as the U6 snRNA which forms part of the spliceosome, a large riboprotein that is implicated in the post-transcriptional processing of RNAs [Robert J White, 1998]. In this chapter I briefly interrogate other members of the pol III transcriptome.

For this analysis, I re-examined the ChIP-seq data in the developing liver of *M. musculus* from chapter 2. I was interested in generating a profile of how much binding of pol III occurs at different functionally annotated regions in the genome. Chapter 2 examined just one such region, the tRNA genes. To compare this with the amount of binding in other genomic features, I quantified the ChIP-seq reads that mapped to annotated features from the GRCm38 mouse genome annotation curated by Ensembl (release 75) [Flicek & al., 2014]. Furthermore, I used data from annotated repeats because it is known that pol III binds to, and potentially drives transcription of, several types of retrotransposons [Carrière & al., 2012], which are screened and annotated by RepeatMasker [Smit and Hubley, 2014]. Pol III ChIP-seq reads were mapped to the *M. musculus* reference genome GRCm38 using Bowtie [Langmead, Trapnell, & al., 2009].

As explained in section 1.4, many tRNA genes arise through gene duplication events and we thus expect many reads to map to multiple locations. This problem also exists for the other annotation types we are interested in. However, the strategy also explained in section 1.4 cannot be applied to non-tRNA annotations since we cannot make the same assumptions about the binding profiles in the flanking regions of the genomic features. In particular, while tRNA transcription uses a type II transcription initiation, other pol III targets use different types of initiation, due to their different promoter and enhancer structure. These differences could have strong effects on the binding profile of active pol III on the target loci.

I overcame this challenge by only reporting a single match per read, even if multiple matches were possible. This assigns the read to an arbitrary locus amongst its possible matches. As long as all potential match positions for a read fall into the same type of annotation, this should not pose a problem for the analysis: all we are interested in is which annotation type a read falls into, not where on the genome.

The annotations used in this analysis required some manual curation. Excluded from the subsequent analysis are repeat types "simple repeats", which stands for small repeats of a few bases, and "tandem repeats", where the variability of the copy number between the reference and the sample makes it impossible to quantify the ChIP signal. Furthermore, I merged other types of repeats, except those corresponding to retrotransposons and those corresponding to other gene copies (such as tRNAs), which I counted separately.

I then counted reads overlapping with each of the annotations mentioned above and calculated TPMs. Figure 4.1 compares the abundance of pol III binding on different features. As expected, tRNA accounts for a majority of the total binding. Unfortunately, this makes it hard to assess the remaining variability visually. The remainder of the data is therefore summarised again in tabular format, for just one stage (table 4.1).

Interestingly, the proportions with the extreme skew towards tRNA genes shown in figure 4.1 are similar to those reported in Raha & al. [2010] and Canella & al. [2012] for pol II association with pol III genes and repeats. As

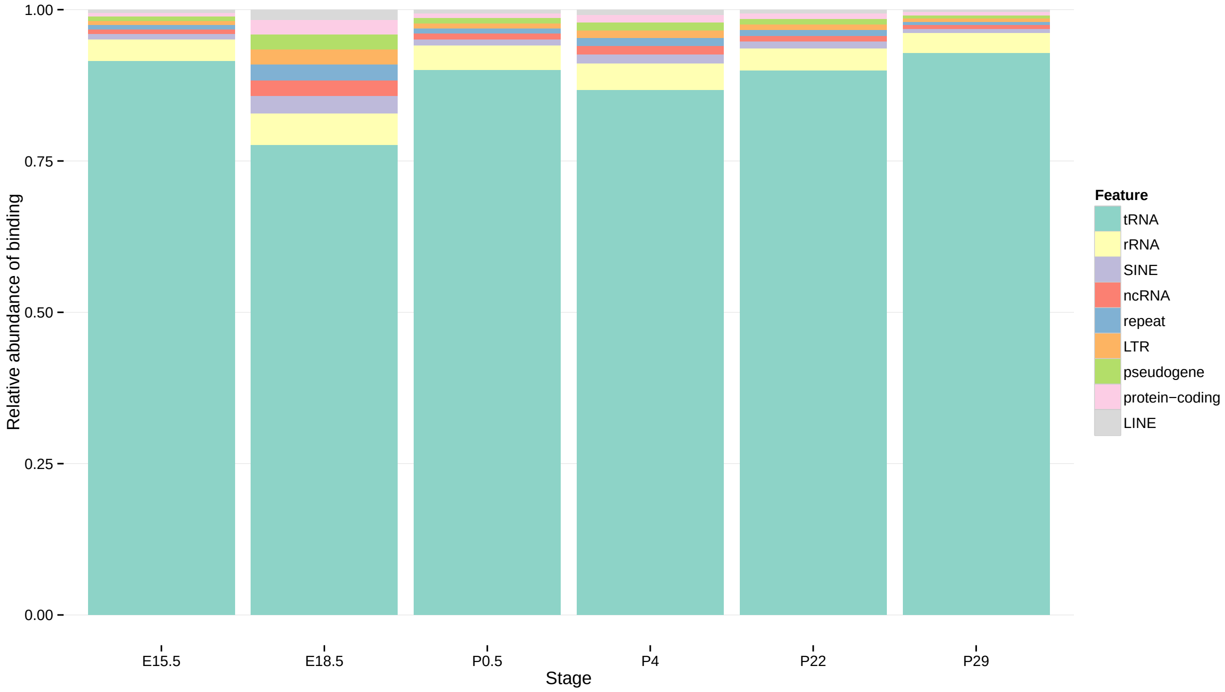


Figure 4.1: Polymerase III coverage, compared across different feature types in six stages of development in liver. Strikingly, the E18.5 stage shows strongly reduced overall tRNA activity. This is due to a single, divergent replicate, which pulls down the average.

Feature	Prop (%)
rRNA	31.1
SINE	11.7
ncRNA	10.6
repeat	10.3
LTR	10.2
pseudogene	9.7
protein-coding	9.3
LINE	7.2

Table 4.1: Polymerase III coverage excluding tRNA genes for liver E15.5.

can be seen in table 4.1, with the exception of rRNA and long interspersed nuclear elements (*LINE*), all features have very similar coverage proportions.

These numbers are probably skewed by the pol III ChIP-seq background signal, which has not been removed from the data, on the assumption that, as long as all features are susceptible to the same amount of spurious signal, the influence on the proportions would be negligible. Comparing the signal strength from the input libraries and the pol III ChIP reveals that this may not be the case (see figure C.1). Consequently, in the future it will be important to verify that the results described below are robust after properly modelling background noise levels.

4.1.1. SINEs are transcribed by pol III

Taken together with the observation reported by Carrière & al. [2012], the results compelled me to take a closer look at the binding of short interspersed nuclear elements (*SINE*) by pol III. SINES are a type of retrotransposon that are highly abundant in the genome: as much as 13 per cent of the genome is a SINE in mammals, and there are of the order of 1 500 000 gene copies [Lander & al., 2001]. SINES are are anywhere from 100 bp to 700 bp in length and are typically composed of three elements: the *head*, the *body* and the *tail*. The head of SINES is derived from pol III-transcribed RNAS. The body is an unrelated piece of sequence often containing fragments from LINES. The tail, finally, usually consists of a simple repeat.

In mammals, most SINES, which form the class of *Alu* elements, are derived from the (pol III-transcribed) 7SL ncRNA. Many others, forming different classes of SINES, are derived from different tRNA genes [Vassetzky and Kramerov, 2013].

The fact that they possess a pol III promoter means that they can, in some cases, be transcribed in vitro [Robert J White, 1998]. It is generally assumed that they do not perform a function in the cell and are usually not actively transcribed. However, as the results by Carrière & al. [2012] indicate that there is indeed limited transcription of SINE in mammals, we should be able to observe this in our pol III ChIP data. It is my hope that in examining SINE transcription, we may be able to learn more about the transcription regulation of tRNA genes: the similar promoter structure of tRNA-derived SINES suggests that the mechanism of gene regulation may be similar here. Furthermore, since SINES are not evolutionarily conserved and perform no vital function in the cell, their active transcription will always happen incidentally. Comparing the upstream sequence composition of active and inactive SINES and of related tRNA genes may enable us to pinpoint common sequence motifs necessary for active transcription.

Testing this is not trivial, since SINES are repeat elements and we thus cannot directly map reads to unique locations, similar to the case of tRNA genes. In the following, rather than looking at the flanking sequences to disambiguate non-uniquely mapping reads as I did for tRNA, I pooled classes of SINE gene copies into 676 classes (downloaded from RepBase [Jurka & al., 2005]). I then generated a reference transcriptome from the consensus sequences of the SINE classes. Reads from pol III ChIP-seq were then mapped to this reference using Bowtie.

Across two tissues (liver and brain) and six stages of development I find that on average 63 per cent of the SINE classes (431 out of 676) show nonzero coverage. In these, there is low but significant enrichment of pol III binding compared to the input libraries (ChIP-seq experiment performed without antibody to quantify the noise introduced by the experiment). The amount of pol III binding to the various SINE classes varies across 4 orders of magnitude. An overview over the magnitude of binding is shown in figure 4.2.

Despite the challenges of working with repeat regions, the quantification of changes in SINE binding by pol III during mouse development seems thus feasible. The next problem to solve is the handling of experimental noise, shown in the input libraries. Conventional ChIP-seq protocols for the identification of protein binding sites or histone modifications use these libraries during peak-calling to quantify and filter out the background signal [Zhang & al., 2008]. The naive way of handling this, simply subtracting the input count from the pol III count, will yield negative values in many cases where the signal is low and noise is high (about 40 per cent in this dataset).

After accounting for noise, the remaining expressed SINE classes can be classified based on the RNA they are derived from. Different classes of SINE have a highly variable number of gene copies, from 100 to 1 000 000. When

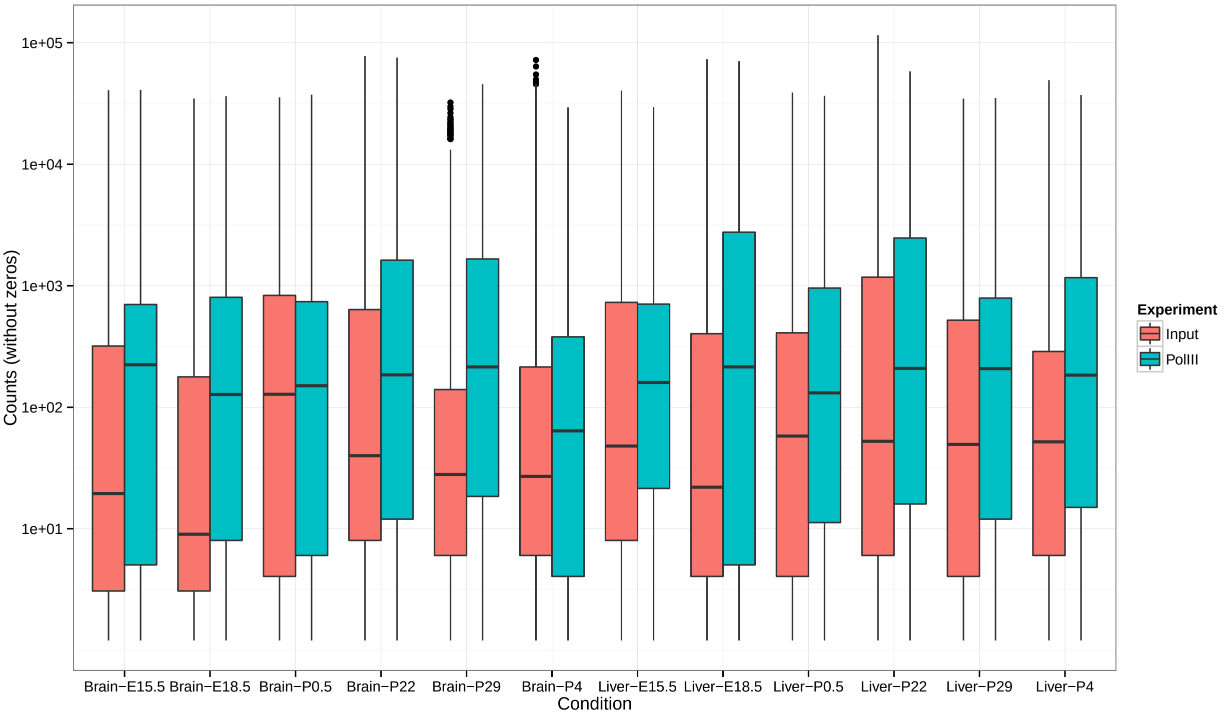


Figure 4.2: SINE binding by pol III across development in liver and brain. Raw counts of pol III binding to different SINE classes. Classes with count equal to 0 have been filtered out.

comparing pol III binding between these classes, it is important to account for these differences, since higher gene copies will have proportionally higher binding.

❖

In summary, I have shown that tRNA genes account for the vast majority of the overall pol III binding. After this, rRNA genes show the most binding, and following this, SINEs. I have mapped the pol III ChIP data against consensus sequences for 676 SINE classes to show that many of these classes have nonzero pol III binding. In the future, I plan to investigate how pol III binding varies between different classes of SINEs, in particular with regards to their origin. Furthermore, I plan to investigate how this binding changes across development, and whether this variation in binding correlates between any of the SINE classes and related tRNA genes. Finally, if such a correlation is found, I will look for common upstream sequence elements that are absent in other SINE classes. However, this last step may require looking at the upstream sequence of individual SINEs rather than that of SINE classes unless it turns out that the upstream sequences of individual SINEs within a class is conserved.

I have focused on SINEs due to their interesting similarity to tRNA genes. A similar analysis as for SINEs can of course also be performed for other repeat classes that showed similar levels of pol III binding in the previous analysis. In fact, there is potential for the SINE analysis to yield a framework to facilitate this kind of sequencing analysis for repeat features, which are notoriously hard to work with in the context of high-throughput sequencing due to their inherent non-mappability.

Conclusion 5

… except in the light of genetics.

Evolution by natural selection is the implicit assumption underlying all of modern biology. Dobzhansky [1973] famously argues against ill-informed criticism of evolution, stating that

> Nothing in biology makes sense except in the light of evolution.

This statement is as relevant today as it was then — both as an admonishment about the still prevalent ignorance of basic biological facts, and serving as a succinct summary of our understanding of biology. In fact, modern evolutionary synthesis, which is the prevailing explanatory model employed today, and which is itself an evolution of the Darwinian theory of descent with modification by means of natural selection, is unchallenged in this status, and is corroborated by every new piece of evidence.

In *"The making of the fittest"*, Sean Carroll argues that the best evidence for evolution we have is the genetic record that we can now read directly in the DNA of all living species [Carroll, 2006]. Carroll was mainly talking about the similarity between homologous genes in different species. But one of the most striking examples of strong conservation is the near-universal genetic code: Throughout billions of years of evolution, the instruction set used to encode the building blocks of proteins has remained virtually unchanged.

In contrast with the conservation of the genetic code itself, the implementation of this code shows some degree of variation, most notably in the variation of the selection of preferential synonymous codons, i.e. the codon bias (reviewed in Ermolaeva [2001]) on the one hand, and in the divergence of the tRNA genes [Kutter & al., 2011] on the other hand.

5.1. Discussion

5.1.1. Regulation of tRNA gene expression

In chapter 2, I have summarised our research into the variability of the tRNA genes, not across evolution, but across development. Our findings mirror a common theme: despite pervasive variation of tRNA gene expression, the anticodon isoacceptor tRNA abundance remains very stable across different stages of development, and matches the codon demand of the transcriptome which, itself, also shows very little variation.

As there is no effect on the abundance of the tRNA anticodon isoacceptor pools, it remains unclear why such variation of the tRNA gene expression exists; however, I have shown that the variation is not stochastic but rather the product of coordinated regulation, precisely to ensure the stability of the anticodon pool. One might thus suspect that the observed variation in tRNA gene expression is simply a consequence of a stabilising mechanism to provide a buffer against extrinsically caused changes in the expression of individual tRNA genes.

On the other hand, the variability in tRNA gene expression may be due to a regulatory role, which is a common theme of ncRNAs: Many different types of ncRNA are known to be implicated in the regulation of gene expression through a variety of mechanisms [Mattick and Makunin, 2006]. For example, long non-coding RNAs (*lncRNA*) have been shown to bind to chromatin and help shape its conformation [Rinn and Chang, 2012], and small RNA fragments from different sources are involved in RNA silencing via a machinery known as RNA-induced silencing complex (*RISC*) [Hamilton and Baulcombe, 1999; Hammond & al., 2000].

Although tRNAs have a "canonical" role — to act as adapters in the process of translation — this of course does not preclude other roles. In fact, there is evidence that ties tRNA-derived fragments to different regulatory roles. As mentioned in the introduction, about ten per cent of the bases in a typical tRNA's transcript are post-transcriptionally modified. Some of these modifications impact the stability of the transcript. For example, it is known that cytosine-5 methylation in the anticodon loop of tRNAs, which is prevalent in actively transcribed tRNAs, inhibits endonucleolytic cleavage. Absence of this methylation leads to cleavage and the accumulation of 3′ and 5′ fragments [Thompson & al., 2008]. Furthermore, there is evidence that overabundance of 5′ tRNA fragments lead to cellular stress [Blanco & al., 2014].

But not only the 5′ fragment of tRNAs is catalytically active: the 3′ ends of specific tRNAs have been shown to act as primers for the transcription of endogenous retroviruses such as type-1 human immunodeficiency virus (*HIV-1*) in a highly sequence-specific manner [Litvak & al., 1994]. In general, the expression of such retroviruses is detrimental for the cell and, by implication, excess abundance of specific tRNA-derived fragments affects the cell's fitness negatively. This exerts a selective pressure to evolve a mechanism for suppressing the expression of such fragments. One way of depleting their abundance is to downregulate the expression of the tRNA genes from which they are derived in response to the detection of excess tRNA fragments.

In sum, the regulatory role of tRNA transcripts adds another dimension to the need for the regulation of their abundance. In fact, the dependence on specific enzymes (such as NSUN2 in mouse) to methylate tRNAs, and thus to ensure their stability hints at the fundamental importance of preventing the formation of excess tRNA-derived fragments [Blanco & al., 2014].

The precise mechanisms that regulate the expression of tRNA genes remain unclear. Corroborating previous reports [Oler & al., 2010], I have found some evidence that tRNA gene activity correlates with specific histone marks. However, it is unclear whether this is a cause or a consequence of differential regulation, and it is insufficient to account for differences in the expression of tRNA genes in close vicinity. Furthermore, there is so far no mechanism for the dynamic feedback necessary for effecting the compensatory effect between genes in an isoacceptor family.

5.1.2. Absence of evidence for codon bias-dependent translation efficiency in mammals

Despite the existence of large variations in codon usage between subsets of genes, some of which are cell type specific, I was unable to find evidence for a regulatory effect of this codon bias on translation rates via higher adaptation to a cell type specific tRNA anticodon isoacceptor pool in mammals. On the contrary, the variation in the tRNA anticodon abundance does not seem to correlate with cell type specific codon demand. This finding, presented in chapter 3, lends support to a view that has recently been challenged [Gingold, Tehler, & al., 2014; Wilusz, 2015]: that translational selection via codon bias, if present at all, plays a negligible role in mammalian systems. It will be interesting to see how this controversy will unfold.

If true, this implies that, in mammals, codon bias has not conserved the regulatory role it plays in unicellular organisms, where it is well established that codon bias influences translation efficiency to control gene-specific expression levels (reviewed in Plotkin and Kudla [2010]). Why would this central role of codon bias be present in unicellular organisms but not in complex multicellular animals? The following is an attempt at an explanation.

In contrast to unicellular organisms, multicellular organisms need to encode the fundamentally different functionality of distinct cell types in a single, static genome. This is achieved through a highly sophisticated regulatory machinery whose dynamic evolution has been revealed in recent years, in particular with a focus on mammalian systems [Villar & al., 2015]. While organisms consisting of a single cell type also employ gene regulatory networks, the complexity of the regulatory machinery in organisms with many cell types is much higher in comparison, and in particular relies heavily on distal regulatory elements and epigenetic modification.

However, despite only consisting of a single cell type, unicellular life still needs to dynamically adapt to different environments. Besides using simple transcriptional control, they also rely heavily on translational control by means of gene specific codon bias variation. In multicellular organisms, the existence of a more direct control of gene expression via transcription regulation obviates the need for this less direct mechanism.

The effect this has is further heightened by the simpler genomic tRNA landscape in unicellular organisms: while mammals have several hundred tRNA genes, with up to dozens of gene copies per anticodon, unicellular organisms possess fewer anticodons, and usually just a single tRNA gene copy per anticodon [Chan and Lowe, 2009]. As a consequence, the downregulation of even just a single tRNA gene copy has profound consequences on the cytosolic tRNA concentration, and thus the ability of the cell to translate individual codons. This acts as a powerful pressure on the selection of suitable codons.

I suggest that these two factors — the relatively higher complexity of transcriptional regulation in mammals, and the higher impact of variation in tRNA gene transcription on variation in tRNA availability and thus on translation efficiency in unicellular organisms — are sufficient to explain the results we observe here as well as established results reported in the literature.

At the end of the project outlined in chapter 3, I have started exploring other potential sources of the cell type-specific codon bias observed in mammals, unrelated to the regulation of translation rate. My first intuition was that the codon bias might be a stochastic artefact caused by the small size of the gene sets under consideration. However, whilst stochastic variation does have an effect on codon bias, it is insufficient to explain all the observed codon bias in most gene sets. I will continue exploring genomic GC bias as another potential cause of this effect.

5.1.3. The extended pol III transcriptome

The pol III ChIP-seq data generated for the projects presented in this thesis provides a wealth of information beyond just tRNA gene activity. Chapter 4 takes a brief glimpse at genome-wide pol III binding and confirms that pol III binding can be used to assess the activity of genes with known pol III-driven transcription.

In particular, I was able to assess binding of pol III to the promoter region of SINE loci. The problem of multi-mapping reads and the high number of SINE gene copies makes it hard to assess the activity of individual SINE genes. However, by collapsing SINE gene families, I could corroborate previous reports of SINE transcription in vivo [Carrière & al., 2012].

5.2. Future directions

5.2.1. Regulation of tRNA transcription

While providing unprecedented insight into the controlled variability of tRNA gene transcription, chapter 2 has failed to establish a mechanism for the differential regulation of tRNA gene transcription. Known features of pol III recruitment, such as transcription factor binding and specific histone marks, could not conclusively be shown to cause the differences I observed in the tRNA gene transcription between different stages of development. My analysis deliberately excluded the internal promoters of tRNA genes from the search for specific motifs since it has previously been reported that variation of the internal promoter of tRNA genes is unrelated to variation in gene expression [Oler & al., 2010; Canella & al., 2012].

However, results in Gingold, Tehler, & al. [2014] indicate that this may have been premature, as they find significant differences in the B box of tRNA genes which they reported as differentially expressed between conditions. This suggests that internal promoter variation may contribute to the observed variability of the tRNA transcriptome after all. I intend to run the methods they used on our data to test this hypothesis.

5.2.2. Codon usage adaptation

The question of what causes codon usage bias variability across functional subsets of the transcriptome remains wide open. GC bias, in particular, is worth exploring further. On the one hand, I observed a robust correlation between GC bias and codon usage, and we know that codon usage can sometimes be predicted from intergenic GC content [S. L. Chen & al., 2004]. On the other hand, Duret [2002] show that, at least in *Drosophila melanogaster* and *Caenorhabditis elegans*, GC bias is uncorrelated with codon usage bias.

To explore this further, two avenues present themselves:

1. It is known that intergenic isochore GC content in mammals predicts codon usage. Since intergenic regions are non-coding, this suggests that differential codon usage between genes has, at best, a minor functional relevance. So far, I have only looked at GC bias in coding sequences. To make similar conclusions, I will have to instead compare codon usage to the GC bias in the flanking regions of protein-coding genes.

2. GC bias can vary between synonymous and non-synonymous codons. If GC bias is indeed causal for codon usage, we would expect that GC bias correlates highly between the first two nucleotide positions of the codon and its wobble position. However, if the wobble position's GC bias is uncorrelated to the GC bias of the other codon positions in a gene set, this would require a different explanation.

Another feature known to constrain codon deployment is the presence of other sequence features in the coding region of genes. This includes binding sites for enhancers and splicing factors [Hyder & al., 1995; Blencowe, 2000]. The extent of this constraint has recently been shown to be much more widespread than previously assumed [Stergachis & al., 2013]. It is conceivable that condition specific gene sets carry enhancers for their own transcription in their gene bodies, which would contribute to a codon usage bias. It would be worthwhile to investigate the enrichment of such binding sites in gene sets with strong codon bias, in particular those reported by [Gingold, Tehler, & al., 2014].

The usage of a simple correlation between matching codons and anticodons, disregarding wobble base pairing, has proved adequate to demonstrate an overall high correlation between the codon demand and matching tRNA anticodon isoacceptor pool. However, arguing about the relative adaptiveness of different gene sets or transcriptomes may make it necessary to consider wobble base pairing to model the codon–anticodon interaction more precisely. The TAI [dos Reis, Wernisch, & al., 2003] offers a way of quantifying codon–anticodon adaptation by considering (simplified, see table 1.2) wobble base pairing rules. However, it is conventionally based on the TRNA isoacceptor gene copy number as a measure of anticodon abundance. I plan to improve this using our TRNA gene expression data, which offer a more accurate anticodon isoacceptor abundance measure, instead. In addition, it may be possible to extend the TAI metric by considering extended wobble base pairing rules [Murphy and Ramakrishnan, 2004] — although it is likely that this will have a very limited effect on the adaptation value.

5.2.3. Pol III transcription of SINEs

The next step in the analysis of SINE expression requires a framework for the robust quantification of pol III binding signal over background noise. I intend to use alignment-free quantification methods [Patro & al., 2014; Bray & al., 2015] to quantify pol III binding on individual gene loci, extending the approach used for TRNA genes. This approach also allows a robust estimate of background noise from ChIP input libraries. To verify the validity of this approach, lacking independent suitable SINE expression data, I will computationally generate high-throughput sequencing read libraries from simulated SINE gene expression profiles, and test for concordance between the simulation and the expression estimated by the pipeline.

Using this framework for estimating SINE gene expression, I can then investigate whether TRNA gene derived SINES possess elements of transcription regulation in comon with their TRNA gene of origin. To do this, individual SINE classes will to be grouped by their origin, and their expression across mouse development will be correlated with the expression of their source TRNA gene to look for common factors.

APPENDIX

Supplementary material for chapter 2

A

The material in this section has been taken from the supplementary figures & methods of Schmitt, Rudolph, & al. [2014] with minimal changes to the figure legends. The figures and their captions have been created jointly by Bianca Schmitt, Claudia Kutter and me.

A.1. Code

The code used in the analysis of the data for this chapter can be found at `https://github.com/klmr/trna` and `https://github.com/klmr/trna-chip-pipeline`.

A.2. Supplementary figures and tables

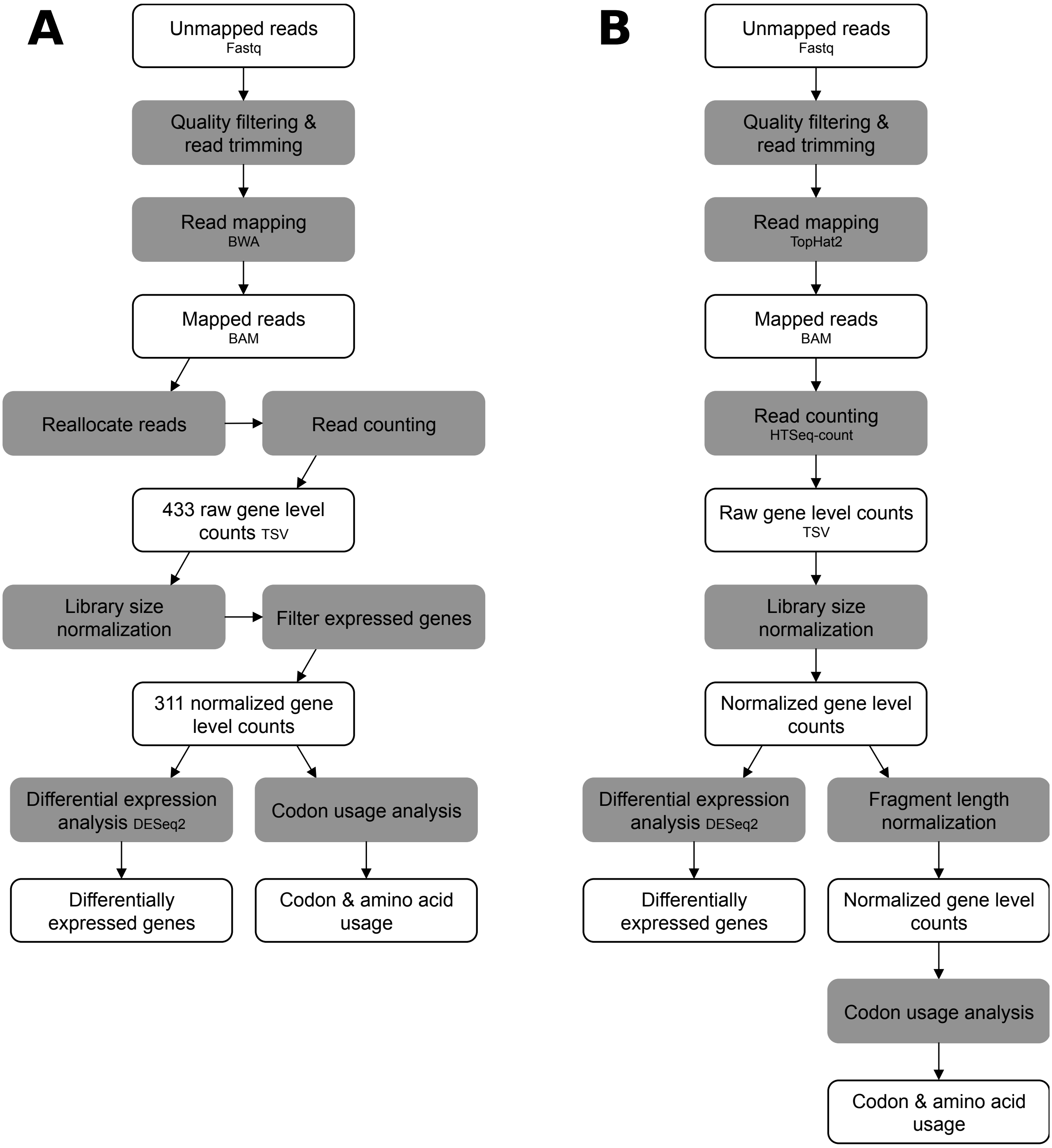

A
Unmapped reads
Fastq
Quality filtering & read trimming
Read mapping
BWA
Mapped reads
BAM
Reallocate reads
Read counting
433 raw gene level counts TSV
Library size normalization
Filter expressed genes
311 normalized gene level counts
Differential expression analysis DESeq2
Codon usage analysis
Differentially expressed genes
Codon & amino acid usage
B
Unmapped reads
Fastq
Quality filtering & read trimming
Read mapping
TopHat2
Mapped reads
BAM
Read counting
HTSeq-count
Raw gene level counts
TSV
Library size normalization
Normalized gene level counts
Differential expression analysis DESeq2
Fragment length normalization
Differentially expressed genes
Normalized gene level counts
Codon usage analysis
Codon & amino acid usage

Figure A.1: Workflow of the genome-wide identification and analysis of protein-coding and tRNA genes. (A) RNA-seq analysis of protein-coding gene expression, differential expression analysis and codon usage analysis. (B) ChIP-seq analysis of pol III occupancy at tRNA gene loci, differential expression analysis of tRNA genes, and anticodon isoacceptor abundance analysis.

Table A.1: The 311 tRNA genes found expressed across mouse development.

Gene	Isoacceptor	Isotype
chr1.trna1000	$\overline{\mathtt{gtc}}$	Asp
chr1.trna1001	$\overline{\mathtt{tcc}}$	Gly
chr1.trna1002	$\overline{\mathtt{ctc}}$	Glu
chr1.trna1004	$\overline{\mathtt{tcc}}$	Gly
chr1.trna1005	$\overline{\mathtt{gtc}}$	Asp
chr1.trna1006	$\overline{\mathtt{gtt}}$	Asn
chr1.trna1022	$\overline{\mathtt{cgg}}$	Pro
chr1.trna1184	$\overline{\mathtt{ttt}}$	Lys
chr1.trna1392	$\overline{\mathtt{gcc}}$	Gly
chr1.trna1547	$\overline{\mathtt{ttc}}$	Glu
chr1.trna485	$\overline{\mathtt{ttt}}$	Lys
chr1.trna672	$\overline{\mathtt{agg}}$	Pro
chr1.trna698	$\overline{\mathtt{cag}}$	Leu
chr1.trna699	$\overline{\mathtt{gcc}}$	Gly
chr1.trna701	$\overline{\mathtt{cag}}$	Leu
chr1.trna702	$\overline{\mathtt{gcc}}$	Gly
chr1.trna703	$\overline{\mathtt{cag}}$	Leu
chr1.trna704	$\overline{\mathtt{gcc}}$	Gly
chr1.trna705	$\overline{\mathtt{cag}}$	Leu
chr1.trna706	$\overline{\mathtt{gcc}}$	Gly
chr1.trna707	$\overline{\mathtt{gtc}}$	Asp
chr1.trna708	$\overline{\mathtt{tcc}}$	Gly
chr1.trna709	$\overline{\mathtt{ctc}}$	Glu
chr1.trna710	$\overline{\mathtt{cac}}$	Val
chr1.trna730	$\overline{\mathtt{tct}}$	Arg
chr1.trna993	$\overline{\mathtt{cag}}$	Leu
chr1.trna994	$\overline{\mathtt{gtc}}$	Asp
chr1.trna995	$\overline{\mathtt{tcc}}$	Gly
chr1.trna996	$\overline{\mathtt{ctc}}$	Glu

chr1.trna997	$\overline{\texttt{gtc}}$	Asp
chr1.trna998	$\overline{\texttt{tcc}}$	Gly
chr1.trna999	$\overline{\texttt{ctc}}$	Glu
chr10.trna1095	$\overline{\texttt{tga}}$	Ser
chr10.trna1316	$\overline{\texttt{taa}}$	Leu
chr10.trna381	$\overline{\texttt{gtt}}$	Asn
chr10.trna688	$\overline{\texttt{cga}}$	Ser
chr10.trna81	$\overline{\texttt{ctc}}$	Glu
chr10.trna851	$\overline{\texttt{gtc}}$	Asp
chr10.trna856	$\overline{\texttt{gtc}}$	Asp
chr10.trna857	$\overline{\texttt{cca}}$	Trp
chr10.trna961	$\overline{\texttt{gaa}}$	Phe
chr11.trna1022	$\overline{\texttt{cct}}$	Arg
chr11.trna1023	$\overline{\texttt{tcg}}$	Arg
chr11.trna1234	$\overline{\texttt{cct}}$	Arg
chr11.trna1432	$\overline{\texttt{gca}}$	Cys
chr11.trna1433	$\overline{\texttt{gca}}$	Cys
chr11.trna1442	$\overline{\texttt{gca}}$	Cys
chr11.trna1444	$\overline{\texttt{gca}}$	Cys
chr11.trna1446	$\overline{\texttt{gtt}}$	Asn
chr11.trna1493	$\overline{\texttt{ttg}}$	Gln
chr11.trna1816	$\overline{\texttt{ttt}}$	Lys
chr11.trna1817	$\overline{\texttt{ctg}}$	Gln
chr11.trna1818	$\overline{\texttt{tct}}$	Arg
chr11.trna1819	$\overline{\texttt{gcc}}$	Gly
chr11.trna1820	$\overline{\texttt{cca}}$	Trp
chr11.trna1821	$\overline{\texttt{gct}}$	Ser
chr11.trna1822	$\overline{\texttt{agt}}$	Thr
chr11.trna1823	$\overline{\texttt{aat}}$	Ile
chr11.trna1824	$\overline{\texttt{tcc}}$	Gly
chr11.trna1849	$\overline{\texttt{cca}}$	Trp
chr11.trna1880	$\overline{\texttt{cca}}$	Trp
chr11.trna1911	$\overline{\texttt{caa}}$	Leu

chr11.trna1912	$\overline{\texttt{ctc}}$	Glu
chr11.trna2021	$\overline{\texttt{aag}}$	Leu
chr11.trna2022	$\overline{\texttt{tgc}}$	Ala
chr11.trna2023	$\overline{\texttt{ctt}}$	Lys
chr11.trna204	$\overline{\texttt{cac}}$	Val
chr11.trna205	$\overline{\texttt{acc}}$	Gly
chr11.trna206	$\overline{\texttt{tgt}}$	Thr
chr11.trna207	$\overline{\texttt{tgg}}$	Pro
chr11.trna208	$\overline{\texttt{aac}}$	Val
chr11.trna393	$\overline{\texttt{aat}}$	Ile
chr11.trna394	$\overline{\texttt{aga}}$	Ser
chr11.trna395	$\overline{\texttt{agt}}$	Thr
chr11.trna396	$\overline{\texttt{cgg}}$	Pro
chr11.trna397	$\overline{\texttt{gtc}}$	Asp
chr11.trna398	$\overline{\texttt{cca}}$	Trp
chr11.trna399	$\overline{\texttt{agt}}$	Thr
chr11.trna400	$\overline{\texttt{cga}}$	Ser
chr11.trna401	$\overline{\texttt{tag}}$	Leu
chr11.trna550	$\overline{\texttt{cgt}}$	Thr
chr11.trna791	$\overline{\texttt{gca}}$	Cys
chr11.trna945	$\overline{\texttt{ccg}}$	Arg
chr12.trna470	$\overline{\texttt{aat}}$	Ile
chr12.trna790	$\overline{\texttt{ctt}}$	Lys
chr13.trna100	$\overline{\texttt{aat}}$	Ile
chr13.trna1000	$\overline{\texttt{ttt}}$	Lys
chr13.trna1001	$\overline{\texttt{ctc}}$	Glu
chr13.trna101	$\overline{\texttt{gta}}$	Tyr
chr13.trna102	$\overline{\texttt{agc}}$	Ala
chr13.trna103	$\overline{\texttt{ctt}}$	Lys
chr13.trna104	$\overline{\texttt{agt}}$	Thr
chr13.trna105	$\overline{\texttt{ttc}}$	Glu
chr13.trna106	$\overline{\texttt{gta}}$	Tyr
chr13.trna107	$\overline{\texttt{cca}}$	Trp

chr13.trna108	$\overline{\texttt{cat}}$	Met
chr13.trna109	$\overline{\texttt{cca}}$	Trp
chr13.trna110	$\overline{\texttt{gcc}}$	Gly
chr13.trna111	$\overline{\texttt{cat}}$	Met
chr13.trna112	$\overline{\texttt{tga}}$	Ser
chr13.trna113	$\overline{\texttt{ttg}}$	Gln
chr13.trna114	$\overline{\texttt{ttg}}$	Gln
chr13.trna115	$\overline{\texttt{gct}}$	Ser
chr13.trna60	$\overline{\texttt{gaa}}$	Phe
chr13.trna61	$\overline{\texttt{cat}}$	Met
chr13.trna62	$\overline{\texttt{aag}}$	Leu
chr13.trna63	$\overline{\texttt{aag}}$	Leu
chr13.trna64	$\overline{\texttt{ctg}}$	Gln
chr13.trna65	$\overline{\texttt{caa}}$	Leu
chr13.trna66	$\overline{\texttt{agc}}$	Ala
chr13.trna67	$\overline{\texttt{agc}}$	Ala
chr13.trna68	$\overline{\texttt{tgc}}$	Ala
chr13.trna69	$\overline{\texttt{agc}}$	Ala
chr13.trna70	$\overline{\texttt{cgc}}$	Ala
chr13.trna71	$\overline{\texttt{agc}}$	Ala
chr13.trna72	$\overline{\texttt{cgt}}$	Thr
chr13.trna73	$\overline{\texttt{tgt}}$	Thr
chr13.trna74	$\overline{\texttt{tcg}}$	Arg
chr13.trna75	$\overline{\texttt{cgt}}$	Thr
chr13.trna77	$\overline{\texttt{gcc}}$	Gly
chr13.trna78	$\overline{\texttt{cat}}$	Met
chr13.trna81	$\overline{\texttt{agt}}$	Thr
chr13.trna82	$\overline{\texttt{cat}}$	Met
chr13.trna83	$\overline{\texttt{ttt}}$	Lys
chr13.trna84	$\overline{\texttt{gtc}}$	Asp
chr13.trna85	$\overline{\texttt{caa}}$	Leu
chr13.trna86	$\overline{\texttt{aga}}$	Ser
chr13.trna87	$\overline{\texttt{ctg}}$	Gln

chr13.trna88	$\overline{\texttt{aga}}$	Ser
chr13.trna89	$\overline{\texttt{ttt}}$	Lys
chr13.trna90	$\overline{\texttt{cat}}$	Met
chr13.trna91	$\overline{\texttt{cac}}$	Val
chr13.trna92	$\overline{\texttt{aat}}$	Ile
chr13.trna93	$\overline{\texttt{aac}}$	Val
chr13.trna94	$\overline{\texttt{agc}}$	Ala
chr13.trna947	$\overline{\texttt{cat}}$	Met
chr13.trna948	$\overline{\texttt{tcg}}$	Arg
chr13.trna949	$\overline{\texttt{tcg}}$	Arg
chr13.trna95	$\overline{\texttt{aac}}$	Val
chr13.trna950	$\overline{\texttt{aga}}$	Ser
chr13.trna951	$\overline{\texttt{acg}}$	Arg
chr13.trna952	$\overline{\texttt{cag}}$	Leu
chr13.trna953	$\overline{\texttt{acg}}$	Arg
chr13.trna954	$\overline{\texttt{cac}}$	Val
chr13.trna955	$\overline{\texttt{cgc}}$	Ala
chr13.trna956	$\overline{\texttt{aat}}$	Ile
chr13.trna957	$\overline{\texttt{agg}}$	Pro
chr13.trna958	$\overline{\texttt{gta}}$	Tyr
chr13.trna959	$\overline{\texttt{gta}}$	Tyr
chr13.trna96	$\overline{\texttt{agc}}$	Ala
chr13.trna960	$\overline{\texttt{gta}}$	Tyr
chr13.trna961	$\overline{\texttt{gta}}$	Tyr
chr13.trna962	$\overline{\texttt{aac}}$	Val
chr13.trna963	$\overline{\texttt{agc}}$	Ala
chr13.trna964	$\overline{\texttt{cac}}$	Val
chr13.trna965	$\overline{\texttt{aac}}$	Val
chr13.trna966	$\overline{\texttt{agc}}$	Ala
chr13.trna968	$\overline{\texttt{aat}}$	Ile
chr13.trna969	$\overline{\texttt{tat}}$	Ile
chr13.trna97	$\overline{\texttt{cac}}$	Val
chr13.trna970	$\overline{\texttt{gct}}$	Ser

chr13.trna971	$\overline{\texttt{agt}}$	Thr
chr13.trna972	$\overline{\texttt{cga}}$	Ser
chr13.trna973	$\overline{\texttt{acg}}$	Arg
chr13.trna974	$\overline{\texttt{aac}}$	Val
chr13.trna975	$\overline{\texttt{ctg}}$	Gln
chr13.trna976	$\overline{\texttt{gct}}$	Ser
chr13.trna978	$\overline{\texttt{aga}}$	Ser
chr13.trna979	$\overline{\texttt{gtc}}$	Asp
chr13.trna98	$\overline{\texttt{cac}}$	Val
chr13.trna980	$\overline{\texttt{aga}}$	Ser
chr13.trna981	$\overline{\texttt{gtc}}$	Asp
chr13.trna982	$\overline{\texttt{ctg}}$	Gln
chr13.trna983	$\overline{\texttt{cat}}$	Met
chr13.trna984	$\overline{\texttt{tga}}$	Ser
chr13.trna985	$\overline{\texttt{tct}}$	Arg
chr13.trna987	$\overline{\texttt{tat}}$	Ile
chr13.trna988	$\overline{\texttt{gaa}}$	Phe
chr13.trna989	$\overline{\texttt{aat}}$	Ile
chr13.trna99	$\overline{\texttt{cat}}$	Met
chr13.trna990	$\overline{\texttt{aat}}$	Ile
chr13.trna991	$\overline{\texttt{taa}}$	Leu
chr13.trna994	$\overline{\texttt{cat}}$	Met
chr13.trna995	$\overline{\texttt{ttg}}$	Gln
chr13.trna996	$\overline{\texttt{agt}}$	Thr
chr13.trna997	$\overline{\texttt{ccg}}$	Arg
chr13.trna998	$\overline{\texttt{caa}}$	Leu
chr13.trna999	$\overline{\texttt{cat}}$	Met
chr14.trna188	$\overline{\texttt{aag}}$	Leu
chr14.trna190	$\overline{\texttt{tgt}}$	Thr
chr14.trna191	$\overline{\texttt{gta}}$	Tyr
chr14.trna192	$\overline{\texttt{tgg}}$	Pro
chr14.trna209	$\overline{\texttt{acg}}$	Arg
chr14.trna347	$\overline{\texttt{ttc}}$	Glu

chr14.trna359	$\overline{\mathtt{ttc}}$	Glu
chr14.trna457	$\overline{\mathtt{gaa}}$	Phe
chr14.trna703	$\overline{\mathtt{tag}}$	Leu
chr14.trna704	$\overline{\mathtt{tgt}}$	Thr
chr14.trna705	$\overline{\mathtt{agg}}$	Pro
chr15.trna913	$\overline{\mathtt{cat}}$	Met
chr16.trna50	$\overline{\mathtt{cgt}}$	Thr
chr17.trna1000	$\overline{\mathtt{ctt}}$	Lys
chr17.trna113	$\overline{\mathtt{ccc}}$	Gly
chr17.trna458	$\overline{\mathtt{gca}}$	Cys
chr17.trna516	$\overline{\mathtt{tat}}$	Ile
chr17.trna82	$\overline{\mathtt{ctt}}$	Lys
chr17.trna83	$\overline{\mathtt{cgg}}$	Pro
chr17.trna84	$\overline{\mathtt{ctt}}$	Lys
chr17.trna994	$\overline{\mathtt{ccg}}$	Arg
chr17.trna995	$\overline{\mathtt{cct}}$	Arg
chr17.trna996	$\overline{\mathtt{tgg}}$	Pro
chr17.trna998	$\overline{\mathtt{tgg}}$	Pro
chr19.trna106	$\overline{\mathtt{gaa}}$	Phe
chr19.trna107	$\overline{\mathtt{ttt}}$	Lys
chr19.trna108	$\overline{\mathtt{gaa}}$	Phe
chr19.trna109	$\overline{\mathtt{tac}}$	Val
chr19.trna110	$\overline{\mathtt{tac}}$	Val
chr19.trna637	$\overline{\mathtt{tct}}$	Arg
chr19.trna638	$\overline{\mathtt{taa}}$	Leu
chr19.trna639	$\overline{\mathtt{ttt}}$	Lys
chr19.trna711	$\overline{\mathtt{gct}}$	Ser
chr19.trna8	$\overline{\mathtt{agc}}$	Ala
chr2.trna1431	$\overline{\mathtt{gtg}}$	His
chr2.trna1432	$\overline{\mathtt{gtg}}$	His
chr2.trna1509	$\overline{\mathtt{gct}}$	Ser
chr2.trna1747	$\overline{\mathtt{gcc}}$	Gly
chr2.trna1947	$\overline{\mathtt{gtt}}$	Asn

chr2.trna263	$\overline{\texttt{cgc}}$	Ala
chr2.trna586	$\overline{\texttt{gtg}}$	His
chr3.trna1	$\overline{\texttt{ctt}}$	Lys
chr3.trna1040	$\overline{\texttt{acg}}$	Arg
chr3.trna27	$\overline{\texttt{gta}}$	Tyr
chr3.trna28	$\overline{\texttt{gta}}$	Tyr
chr3.trna283	$\overline{\texttt{gtt}}$	Asn
chr3.trna284	$\overline{\texttt{cac}}$	Val
chr3.trna286	$\overline{\texttt{ttc}}$	Glu
chr3.trna287	$\overline{\texttt{ccc}}$	Gly
chr3.trna289	$\overline{\texttt{gtt}}$	Asn
chr3.trna29	$\overline{\texttt{agc}}$	Ala
chr3.trna291	$\overline{\texttt{gtg}}$	His
chr3.trna292	$\overline{\texttt{ctt}}$	Lys
chr3.trna293	$\overline{\texttt{gtg}}$	His
chr3.trna294	$\overline{\texttt{gtt}}$	Asn
chr3.trna295	$\overline{\texttt{gtg}}$	His
chr3.trna297	$\overline{\texttt{ctg}}$	Gln
chr3.trna298	$\overline{\texttt{gtt}}$	Asn
chr3.trna303	$\overline{\texttt{ctc}}$	Glu
chr3.trna309	$\overline{\texttt{ctg}}$	Gln
chr3.trna48	$\overline{\texttt{aac}}$	Val
chr3.trna628	$\overline{\texttt{tct}}$	Arg
chr3.trna745	$\overline{\texttt{ctc}}$	Glu
chr3.trna746	$\overline{\texttt{tcc}}$	Gly
chr3.trna747	$\overline{\texttt{gtg}}$	His
chr3.trna748	$\overline{\texttt{ctt}}$	Lys
chr3.trna749	$\overline{\texttt{gtg}}$	His
chr3.trna750	$\overline{\texttt{gtt}}$	Asn
chr3.trna751	$\overline{\texttt{gtg}}$	His
chr3.trna752	$\overline{\texttt{ccc}}$	Gly
chr3.trna753	$\overline{\texttt{ctg}}$	Gln
chr3.trna754	$\overline{\texttt{ttc}}$	Glu

chr3.trna755	$\overline{\texttt{ccc}}$	Gly
chr3.trna756	$\overline{\texttt{ctg}}$	Gln
chr3.trna757	$\overline{\texttt{gtt}}$	Asn
chr3.trna792	$\overline{\texttt{cat}}$	Met
chr3.trna878	$\overline{\texttt{gcc}}$	Gly
chr3.trna92	$\overline{\texttt{tgg}}$	Pro
chr3.trna93	$\overline{\texttt{agg}}$	Pro
chr4.trna16	$\overline{\texttt{aga}}$	Ser
chr4.trna1697	$\overline{\texttt{gtg}}$	His
chr5.trna1043	$\overline{\texttt{gtt}}$	Asn
chr5.trna109	$\overline{\texttt{gta}}$	Tyr
chr5.trna110	$\overline{\texttt{agc}}$	Ala
chr5.trna1314	$\overline{\texttt{gtc}}$	Asp
chr5.trna1315	$\overline{\texttt{gaa}}$	Phe
chr5.trna1316	$\overline{\texttt{gtc}}$	Asp
chr5.trna1317	$\overline{\texttt{tgc}}$	Ala
chr5.trna702	$\overline{\texttt{tgc}}$	Ala
chr6.trna1021	$\overline{\texttt{gca}}$	Cys
chr6.trna1022	$\overline{\texttt{gca}}$	Cys
chr6.trna1025	$\overline{\texttt{gca}}$	Cys
chr6.trna1029	$\overline{\texttt{gca}}$	Cys
chr6.trna107	$\overline{\texttt{cct}}$	Arg
chr6.trna157	$\overline{\texttt{gca}}$	Cys
chr6.trna317	$\overline{\texttt{ccc}}$	Gly
chr6.trna46	$\overline{\texttt{agg}}$	Pro
chr7.trna1213	$\overline{\texttt{agt}}$	Thr
chr7.trna1276	$\overline{\texttt{ttt}}$	Lys
chr7.trna156	$\overline{\texttt{tat}}$	Ile
chr7.trna337	$\overline{\texttt{ttc}}$	Glu
chr7.trna387	$\overline{\texttt{tcg}}$	Arg
chr7.trna441	$\overline{\texttt{tgg}}$	Pro
chr7.trna559	$\overline{\texttt{aag}}$	Leu
chr7.trna86	$\overline{\texttt{tca}}$	SeC

chr7.trna861	$\overline{\mathtt{tag}}$	Leu
chr7.trna977	$\overline{\mathtt{agg}}$	Pro
chr8.trna1008	$\overline{\mathtt{cag}}$	Leu
chr8.trna414	$\overline{\mathtt{cag}}$	Leu
chr8.trna560	$\overline{\mathtt{gcc}}$	Gly
chr8.trna783	$\overline{\mathtt{cat}}$	Met
chr8.trna886	$\overline{\mathtt{gcc}}$	Gly
chr8.trna887	$\overline{\mathtt{gcc}}$	Gly
chr9.trna1035	$\overline{\mathtt{ctt}}$	Lys
chr9.trna342	$\overline{\mathtt{ctg}}$	Gln
chr9.trna592	$\overline{\mathtt{gca}}$	Cys
chr9.trna593	$\overline{\mathtt{gca}}$	Cys
chr9.trna783	$\overline{\mathtt{acg}}$	Arg
chr9.trna961	$\overline{\mathtt{ttc}}$	Glu
chrX.trna371	$\overline{\mathtt{tgc}}$	Ala
chrX.trna375	$\overline{\mathtt{tgc}}$	Ala
chrX.trna459	$\overline{\mathtt{tac}}$	Val
chrX.trna637	$\overline{\mathtt{tgc}}$	Ala

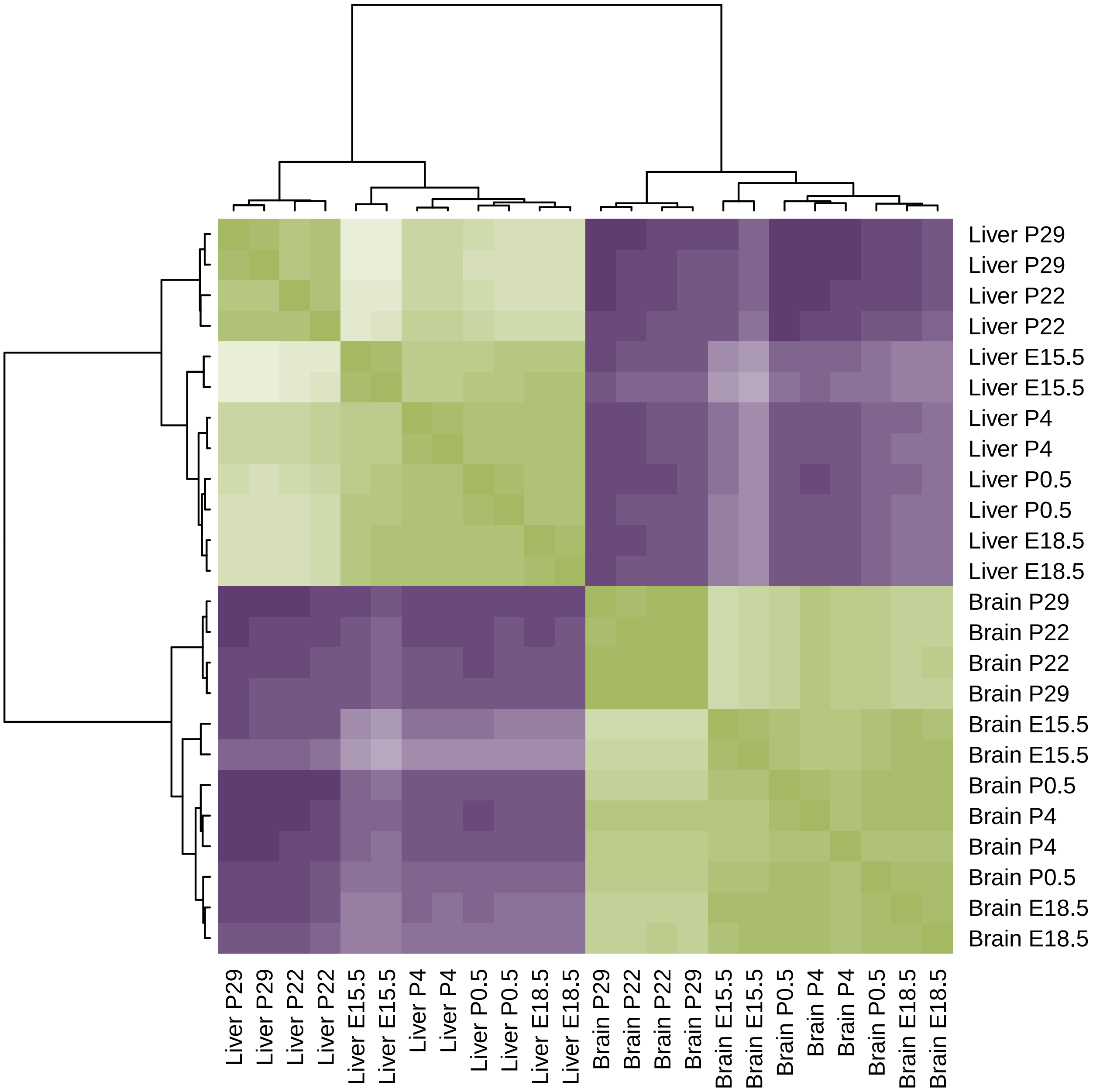


Figure A.2: Hierarchical clustering of mRNA gene expression correlations. The heatmap shows the Spearman correlations of mRNA gene expression values, representing the same data as figure 2.4. The samples cluster hierarchically by tissue, followed by developmental stage.

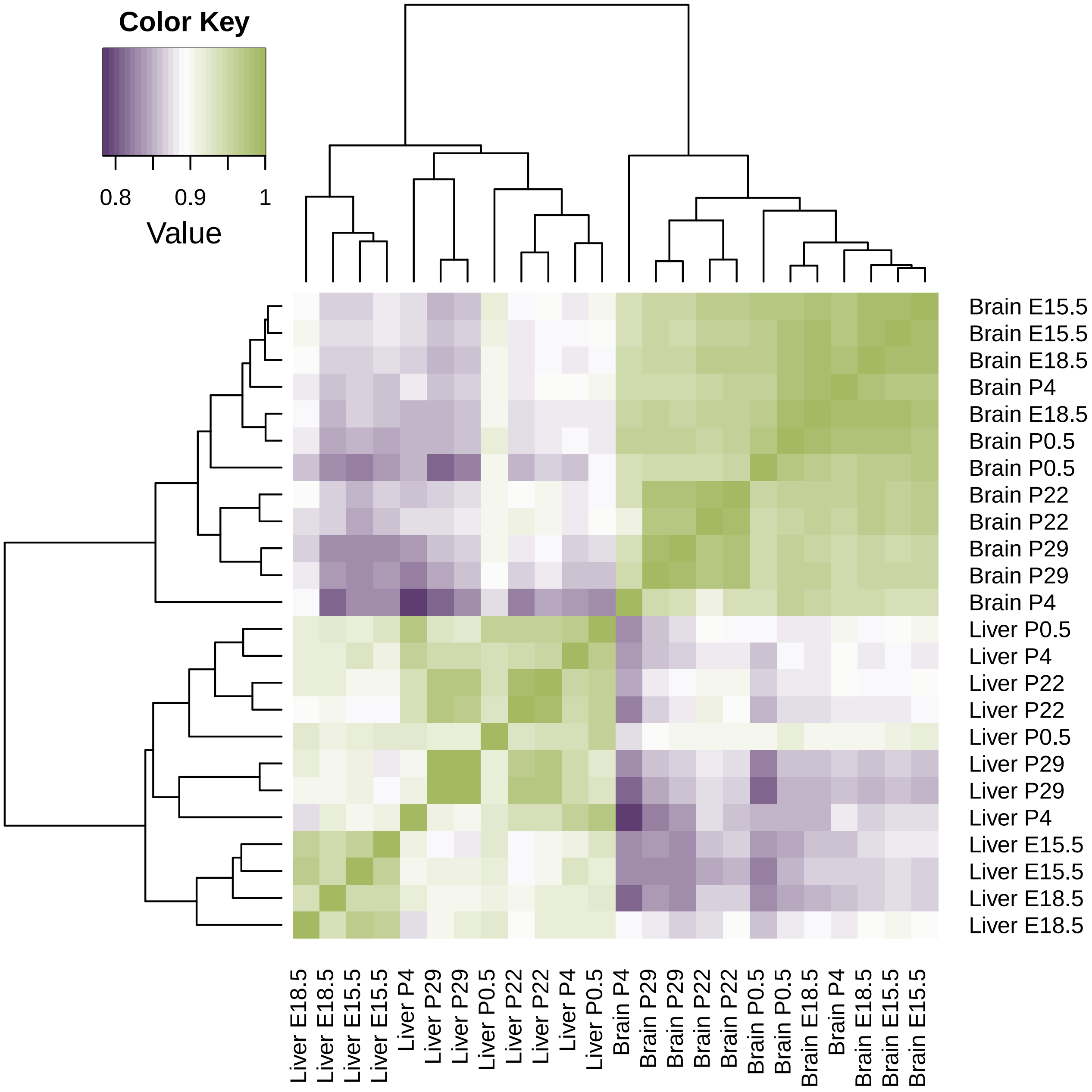
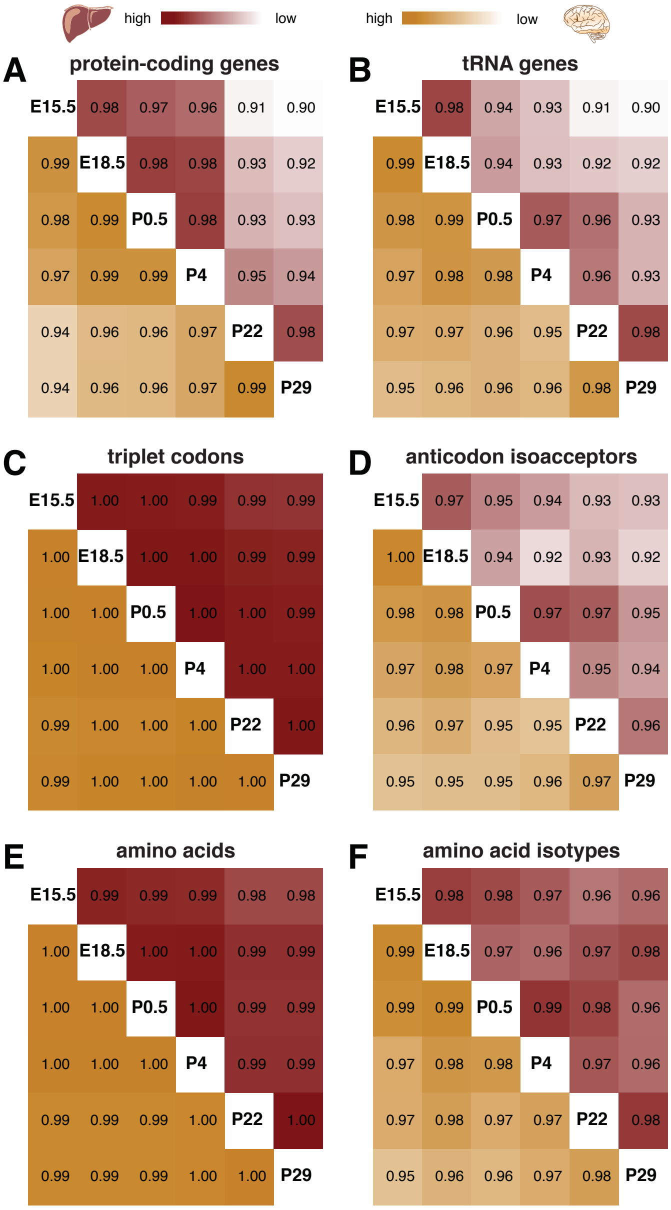


Figure A.3: Hierarchical clustering of tRNA gene expression correlations. The heatmap shows the Spearman correlations of tRNA gene expression values, representing the same data as figure 2.8. The samples cluster hierarchically by tissue, followed by developmental stage, with few exceptions.



A protein-coding genes

E15.5	0.98	0.97	0.96	0.91	0.90
0.99	E18.5	0.98	0.98	0.93	0.92
0.98	0.99	P0.5	0.98	0.93	0.93
0.97	0.99	0.99	P4	0.95	0.94
0.94	0.96	0.96	0.97	P22	0.98
0.94	0.96	0.96	0.97	0.99	P29

B tRNA genes

E15.5	0.98	0.94	0.93	0.91	0.90
0.99	E18.5	0.94	0.93	0.92	0.92
0.98	0.99	P0.5	0.97	0.96	0.93
0.97	0.98	0.98	P4	0.96	0.93
0.97	0.97	0.96	0.95	P22	0.98
0.95	0.96	0.96	0.96	0.98	P29

C triplet codons

E15.5	1.00	1.00	0.99	0.99	0.99
1.00	E18.5	1.00	1.00	0.99	0.99
1.00	1.00	P0.5	1.00	1.00	0.99
1.00	1.00	1.00	P4	1.00	1.00
0.99	1.00	1.00	1.00	P22	1.00
0.99	1.00	1.00	1.00	1.00	P29

D anticodon isoacceptors

E15.5	0.97	0.95	0.94	0.93	0.93
1.00	E18.5	0.94	0.92	0.93	0.92
0.98	0.98	P0.5	0.97	0.97	0.95
0.97	0.98	0.97	P4	0.95	0.94
0.96	0.97	0.95	0.95	P22	0.96
0.95	0.95	0.95	0.96	0.97	P29

E amino acids

E15.5	0.99	0.99	0.99	0.98	0.98
1.00	E18.5	1.00	1.00	0.99	0.99
1.00	1.00	P0.5	1.00	0.99	0.99
1.00	1.00	1.00	P4	0.99	0.99
0.99	0.99	0.99	1.00	P22	1.00
0.99	0.99	0.99	1.00	1.00	P29

F amino acid isotypes

E15.5	0.98	0.98	0.97	0.96	0.96
0.99	E18.5	0.97	0.96	0.97	0.98
0.99	0.99	P0.5	0.99	0.98	0.96
0.97	0.98	0.98	P4	0.97	0.96
0.97	0.98	0.97	0.97	P22	0.98
0.95	0.96	0.96	0.97	0.98	P29

Figure A.4: Correlation of RNA-seq and pol III ChIP-seq data during mouse liver and brain development. Correlation of (A) protein-coding gene expression across developmental stages, (B) tRNA gene expression as measured by pol III occupancy, (C) triplet codon usage in protein-coding genes, (D) tRNA anticodon isoacceptor, (E) amino acid usage of protein-coding genes and (F) tRNA amino acid isotype.

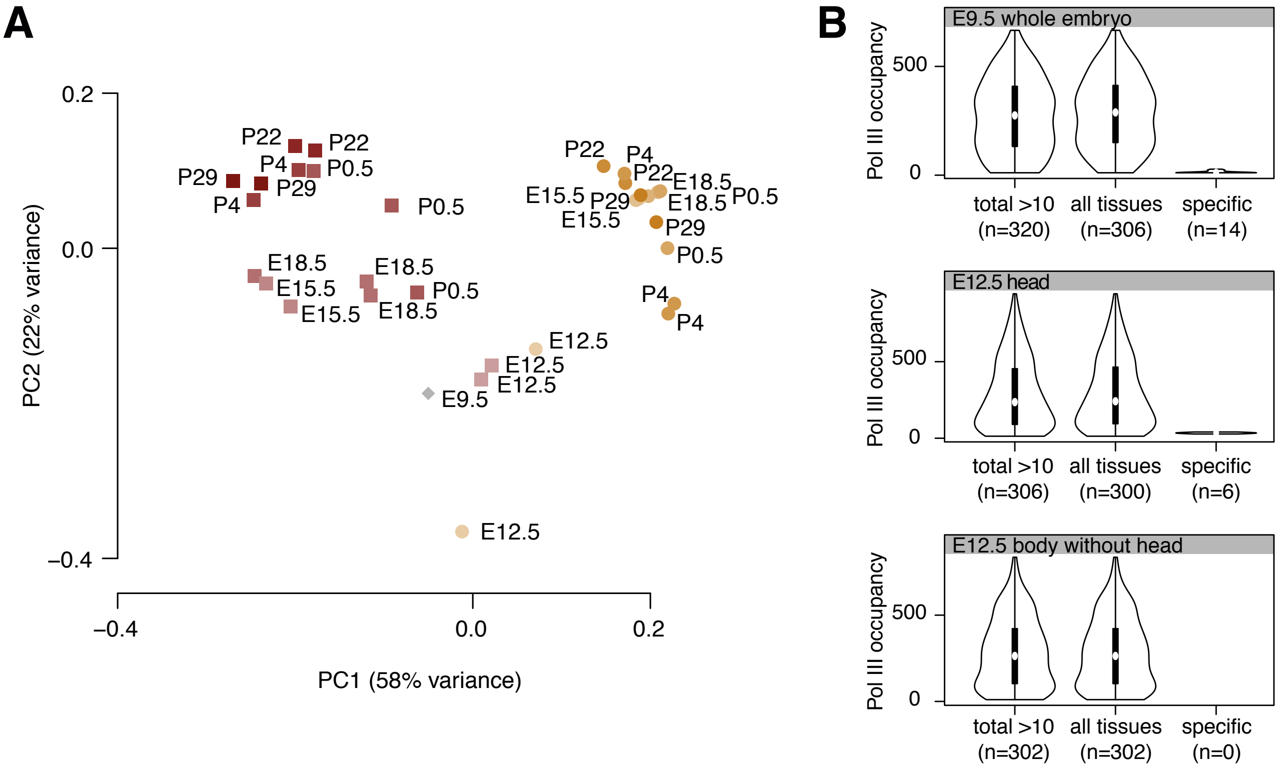


Figure A.5: Early developmental stage-specific tRNA genes are lowly expressed. (A) Factorial map of the PCA of pol III occupied tRNA gene expression levels in liver (red), brain (yellow), embryonic body without head (light red) and head (light yellow) of stage E12.5, as well as whole E9.5 embryo (grey). The proportion of variance explained by the PC is indicated in parenthesis.
(B) Violin plots represent normalized enrichment of pol III at tRNA genes identified in E9.5 whole embryo (top), E12.5 head (middle) and E12.5 body without head (bottom) tissue. In parentheses are the numbers of tRNA genes transcribed in the particular embryonic stage ("total $>$ 10"), which are subdivided into tRNA genes that can be found in the 12 developmental stages according to figure 2.6 ("all tissues") and those that are specific for the embryonic stage ("specific").

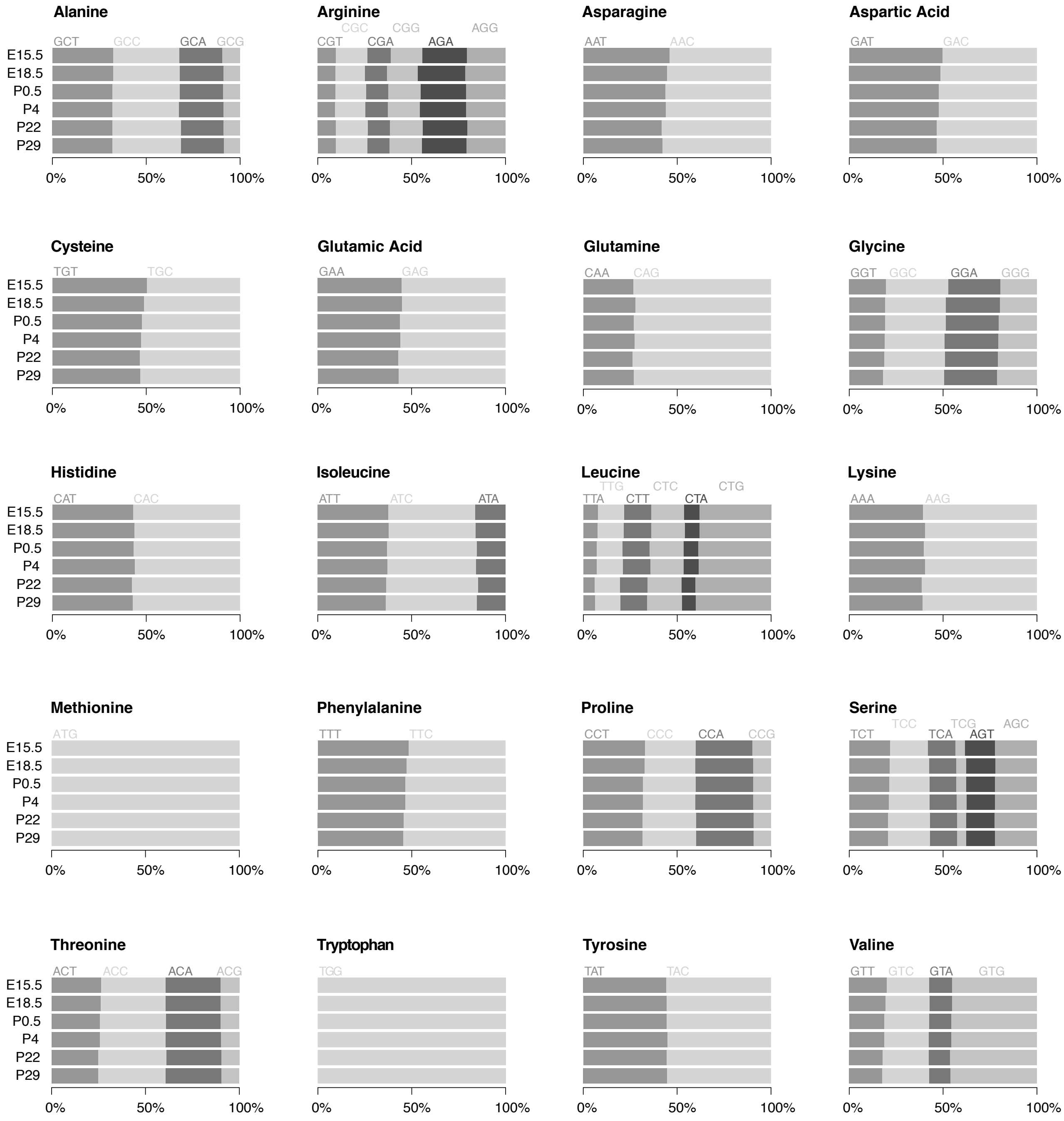


Figure A.6: Observed codon usage in mRNA transcriptomes of developing mouse liver. Proportional frequencies (RCU) weighted by transcript expression are shown for triplet codons ordered by amino acid as a bar plot, where grey shading is by triplet codon. Data is obtained from liver RNA-seq data of all 6 developmental stages.

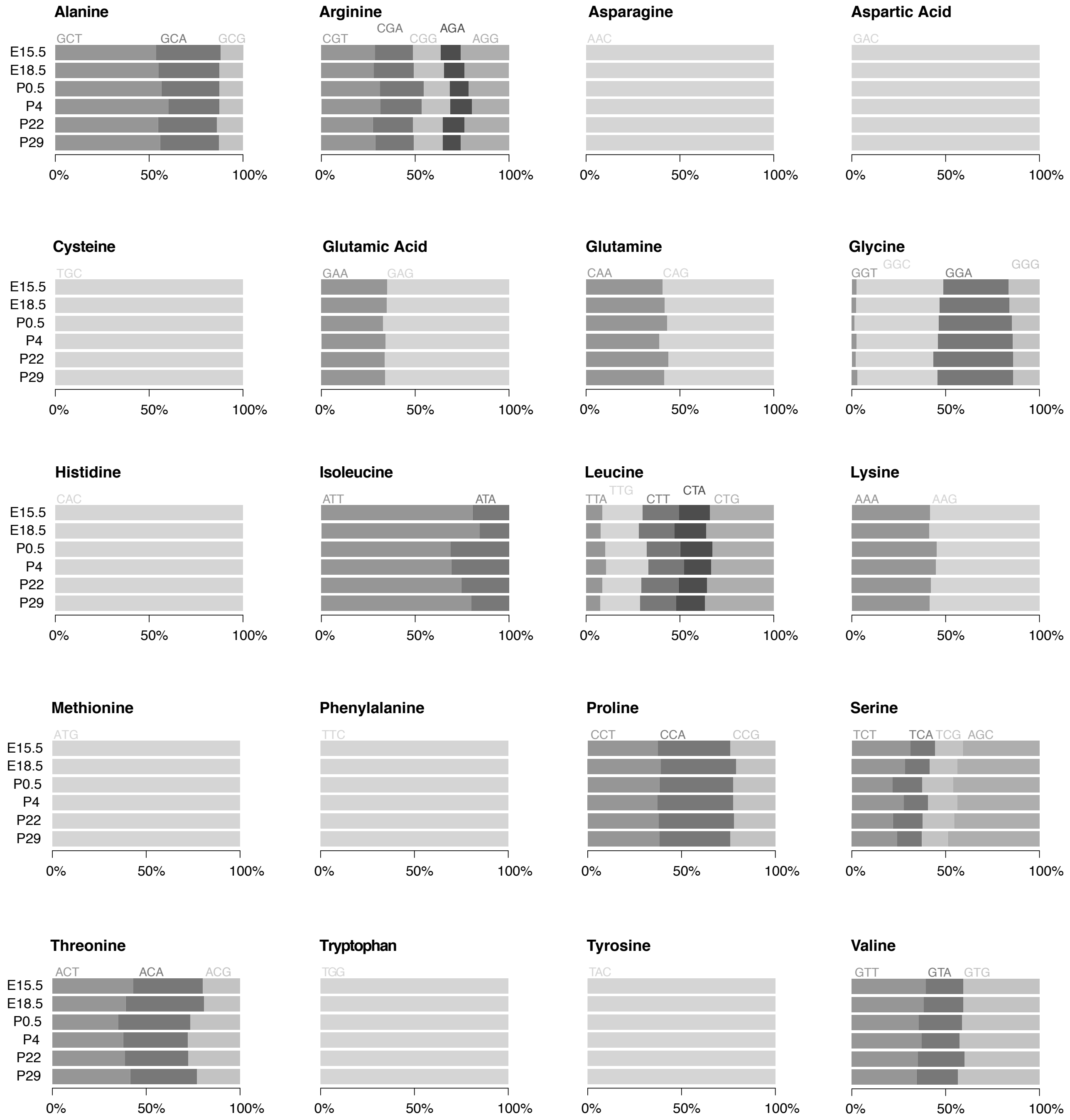


Figure A.7: Observed anticodon abundance of tRNA isoacceptors of developing mouse liver. Proportional frequencies weighted by tRNA gene expression (RAA) are shown for anticodon isoacceptors ordered by amino acid isotype as a bar plot, where grey shading is by anticodon. Data is obtained from liver pol III ChIP-seq data of all 6 developmental stages.

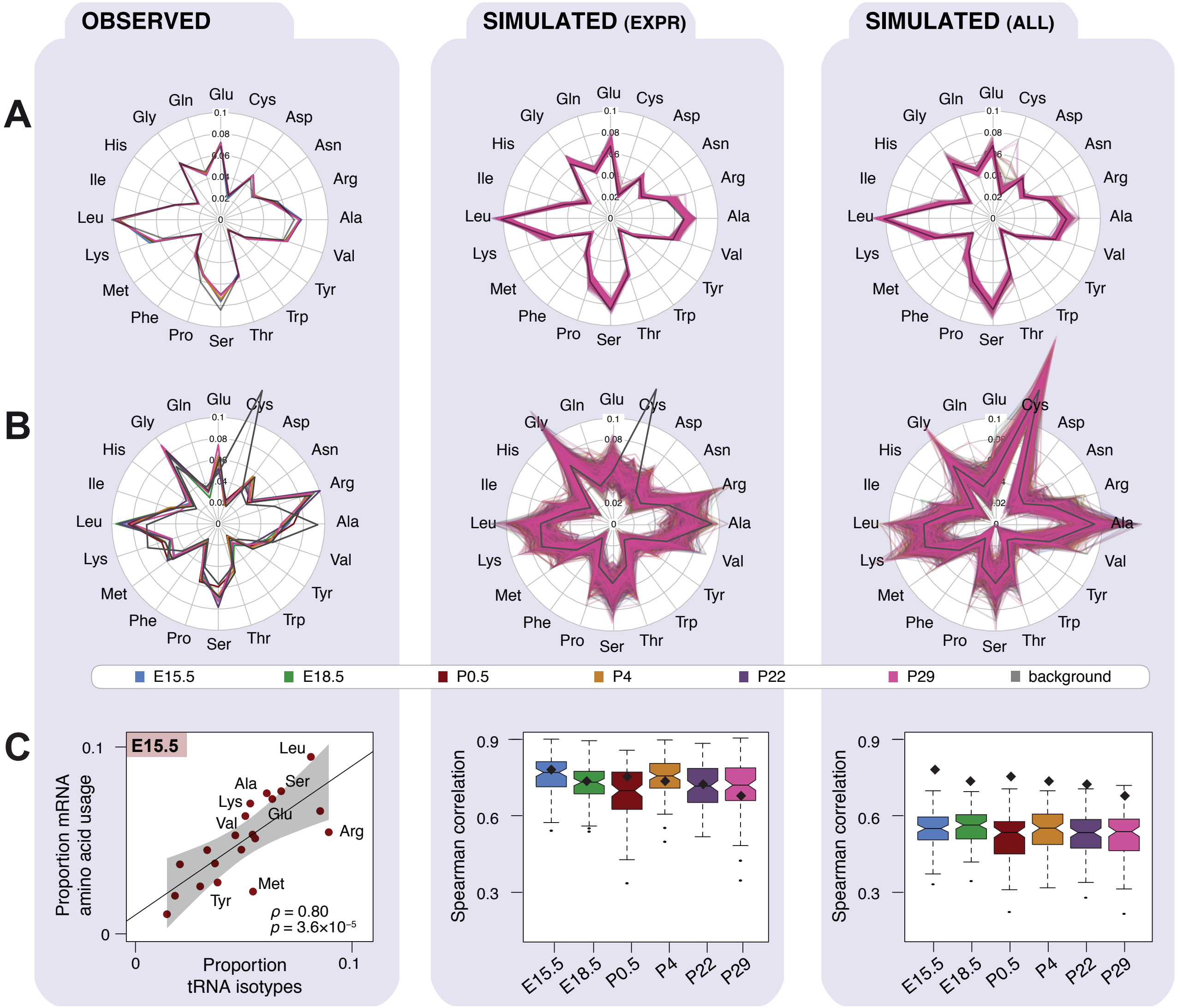


Figure A.8: Observed and simulated amino acid and isotype usage in transcriptomes across mouse liver development. Each panel (A–C) consists of three columns: experimentally observed data (left), simulated patterns of transcription randomized among either the expressed genes (middle) or all genomically encoded genes (right). Transcriptomes of each developmental stage were simulated 100 times. Proportional frequencies weighted by transcript expression are shown for (A) 20 amino acids as a radial plot, where data lines are coloured by developmental stage and the background of all genomically annotated mRNA genes is in grey. Labels within grid of radial plot describe ratios. Proportional frequencies weighted by pol III binding are shown for (B) 20 isotypes as a radial plot, both coloured as above (grey: background of all genomically annotated tRNA genes). (C) Plot right panel shows Spearman's rank correlation coefficients (ρ) and p-values (p) of pol III binding to tRNA isotypes (x-axis) and transcriptomic amino acid frequencies weighted by expression obtained from RNA-seq data (y-axis) in E15.5 liver (experimentally observed data) and all six developmental stages (simulated data). Amino acid isotypes outside the 99 per cent confidence interval (grey area within plot in C right) are named. Observed Spearman's rank correlation coefficients across all stages (coloured as above) are indicated by black diamonds in plot C middle and left panels.

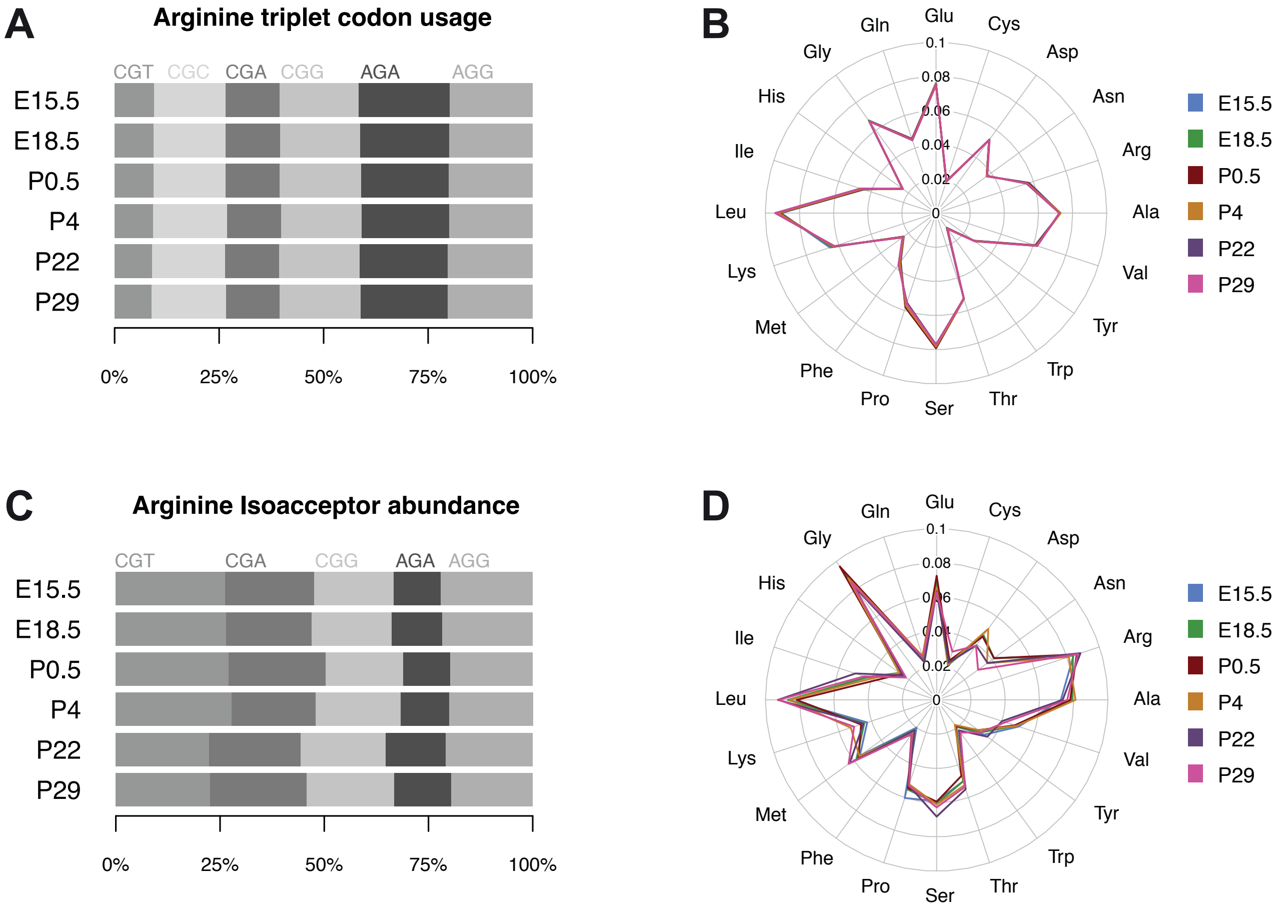


Figure A.9: mRNA codon usage and pol III occupancy of tRNA isotypes in developing mouse brain tissue. Proportional frequency weighted by transcript expression of (A) arginine triplet codons, (B) amino acids, (C) pol III binding of arginine isoacceptors and (D) pol III binding of amino acid isotypes. Grey shading is by triplet codon (A) or tRNA anticodon (C). Labels within grid of radial plot describe proportions.

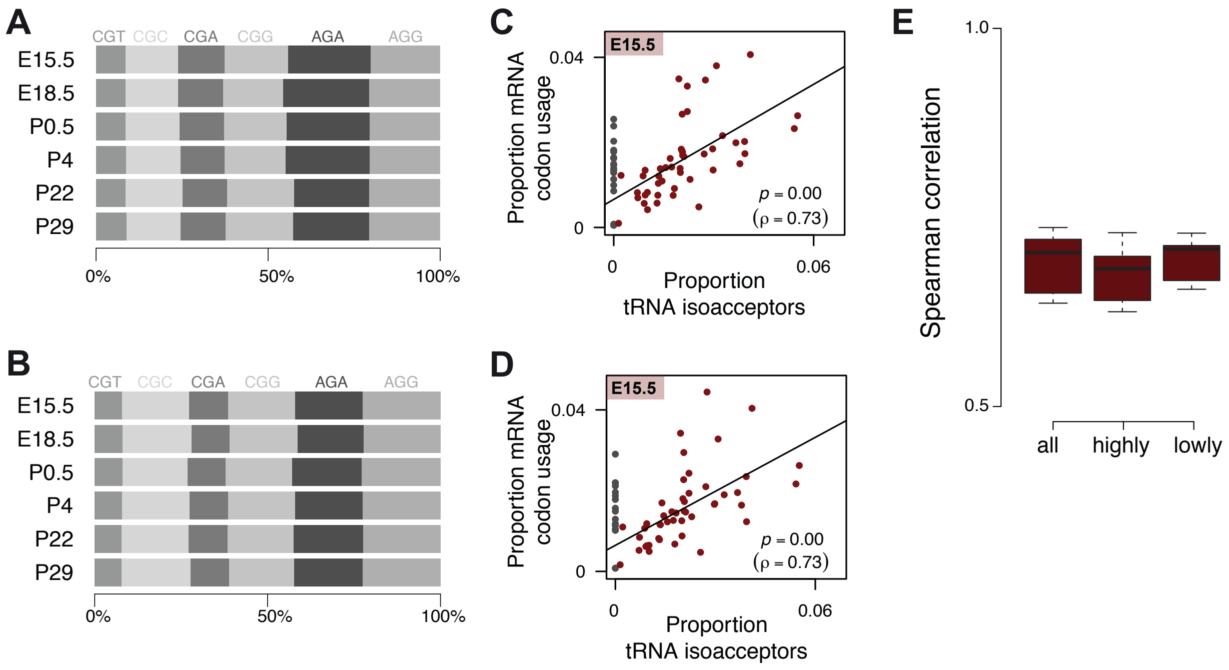


Figure A.10: Highly versus lowly expressed protein-coding genes show no differential codon usage. Proportional frequencies weighted by transcript expression are shown for arginine triplet codons as a bar plot of (A) highly (90th–95th percentile) and (B) lowly expressed (25th–50th percentile) protein-coding genes during liver development, where grey shading is by triplet codon. Plots show Spearman's rank correlation coefficients (ρ) and p-values (p) of pol III binding to tRNA isoacceptors (x-axis) and transcriptomic codon frequencies weighted by expression obtained from RNA-seq data (y-axis) in E15.5 liver of (C) highly and (D) lowly expressed protein-coding genes. Anticodon isoacceptors (grey dots in plots) are not encoded in the mouse genome and were excluded from calculating the correlation coefficients. (E) Variances of correlation values over all stages in liver (i) all expressed protein-coding genes, (ii) highly and (iii) lowly expressed protein-coding gene sets.

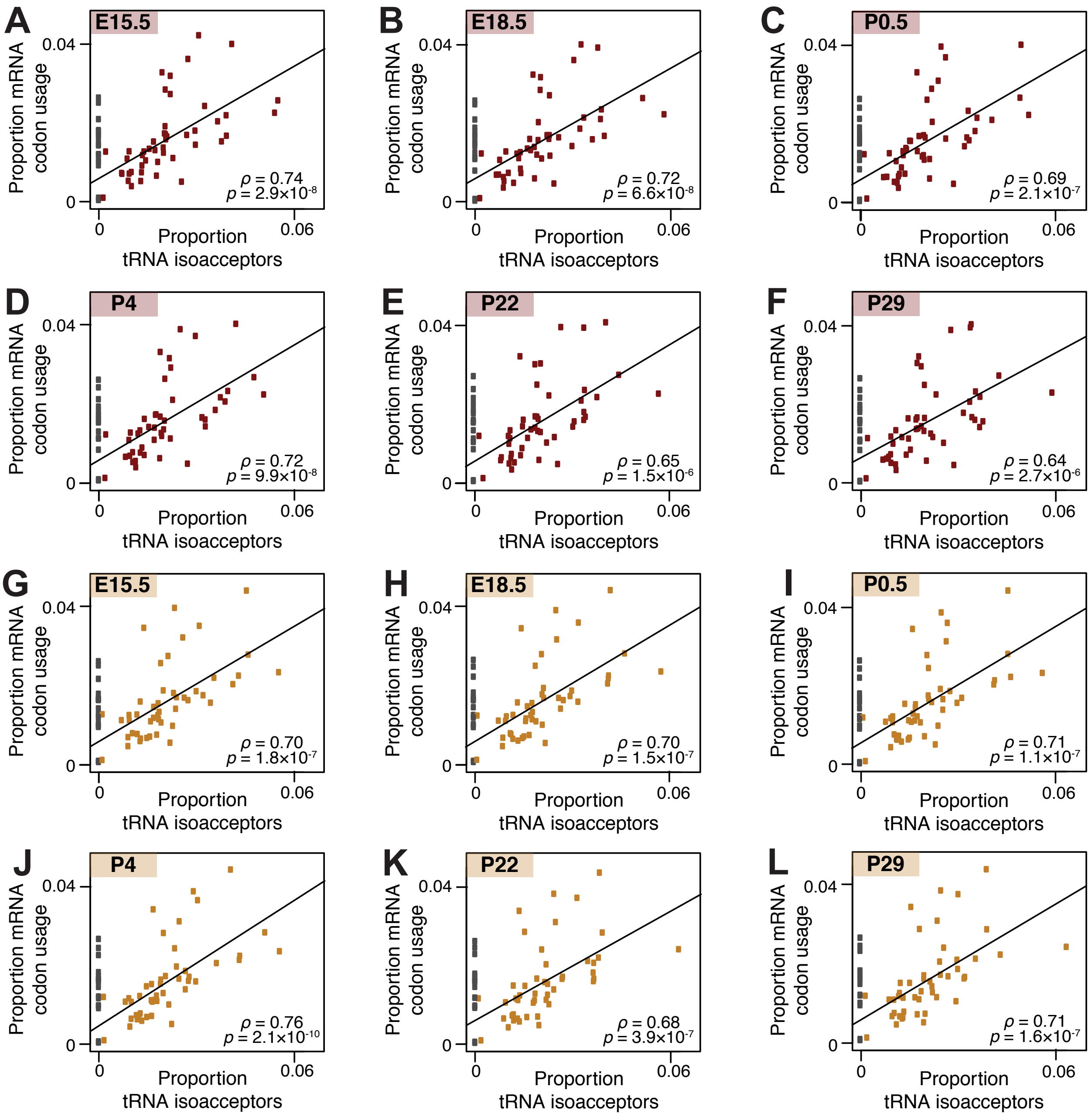


Figure A.11: Transcriptomic mRNA codon usage and pol III binding to tRNA isoacceptors correlate in developing mouse liver and brain. Plots show correlation of proportional pol III binding to tRNA isoacceptors (x-axis) and transcriptomic codon frequencies weighted by expression obtained from RNA-seq data (y-axis). Correlation plots for developing liver (A–F) and brain (G–L) are shown. Indexed box in top left indicates developmental stage. Grey dots represent degenerated codons. Spearman's rank correlation coefficients (ρ) are reported along with their p-values (p) in bottom right of each panel.

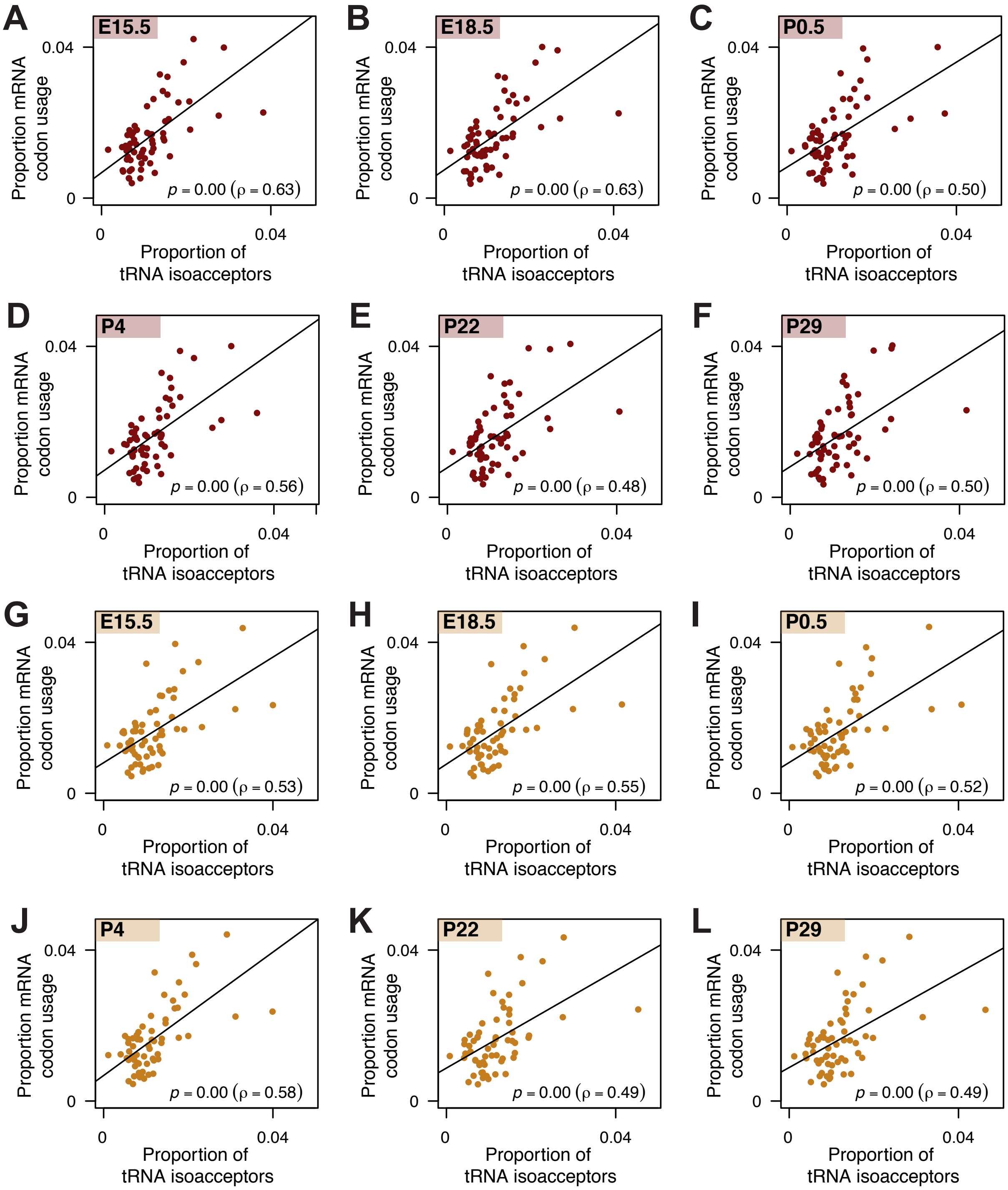


Figure A.12: Transcriptomic mRNA codon usage and wobble corrected pol III binding to tRNA isoacceptors correlate in developing mouse liver and brain. Plots show correlation of proportional pol III binding to tRNA isoacceptors corrected according to wobble pairing (x-axis) and transcriptomic codon frequencies weighted by expression obtained from RNA-seq data (y-axis). Correlation plots for developing liver (A–F) and brain (G–L) are shown. Indexed box in top left indicates developmental stage. Spearman's rank correlation coefficients (ρ) are reported along with their p-values (p) in bottom right of each panel.

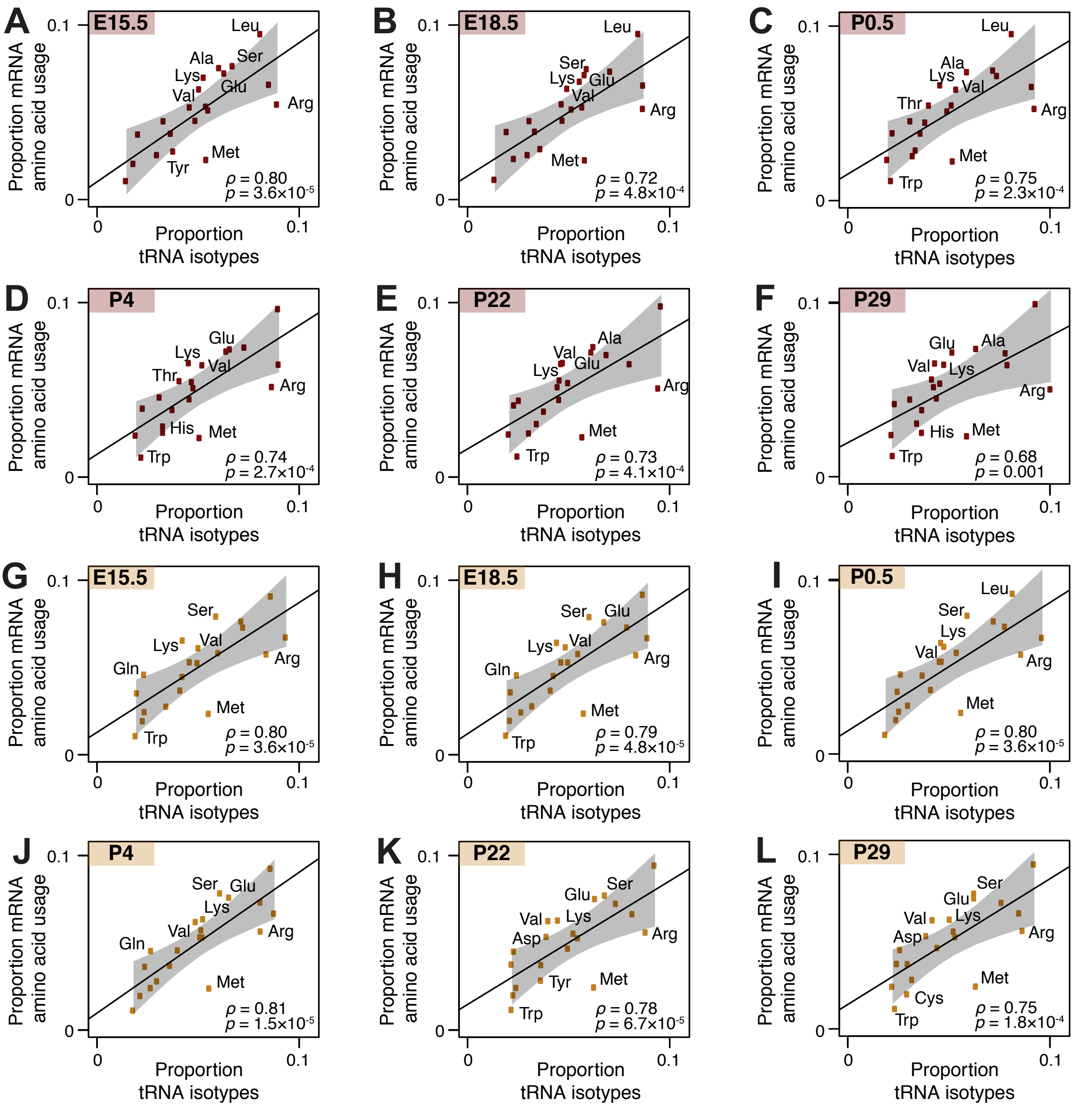


Figure A.13: Transcriptomic mRNA amino acid usage and pol III binding to tRNA isotypes correlate in developing mouse liver and brain. Plots show correlation of pol III binding to tRNA isotypes (x-axis) and transcriptomic amino acid frequencies weighted by expression obtained from RNA-seq data (y-axis). Correlation plots for developing liver (A–F) and brain (G–L) are shown. Indexed box in top left indicates developmental stage. Grey area represents 99 per cent confidence interval. Spearman's rank correlation coefficients (ρ) and the corresponding p-values (p) are reported in top left and bottom right, respectively of each panel. Amino acid isotypes outside the 99 per cent confidence interval (grey area) are named.

Tissue	Contrast	Query ID	Target ID	*p*-value	Target consensus
Brain	E18.5–P22	TTAGCTTTGTTTCTTTGTTTT	MA0041.1	0.01	GAATGTTTGTTT
Brain	E18.5–P22	TTAGCTTTGTTTCTTTGTTTT	MA0042.1	0.03	GGATGTTTGTTT
Brain	P4–P22	GTCAACTCCCTCCCCAGATCCCACCCGCC	MA0068.1	0.04	GAAAAATTTCCCATACTCCACTCCCCCCCC
Brain	P4–P22	GTCAACTCCCTCCCCAGATCCCACCCGCC	MA0079.2	0.04	CCCCGCCCCC

Table A.3: Significantly enriched MEME hits. Shown are hits that are significantly enriched at the 5 per cent threshold after correcting for multiple testing, given with their corrected *p*-values.

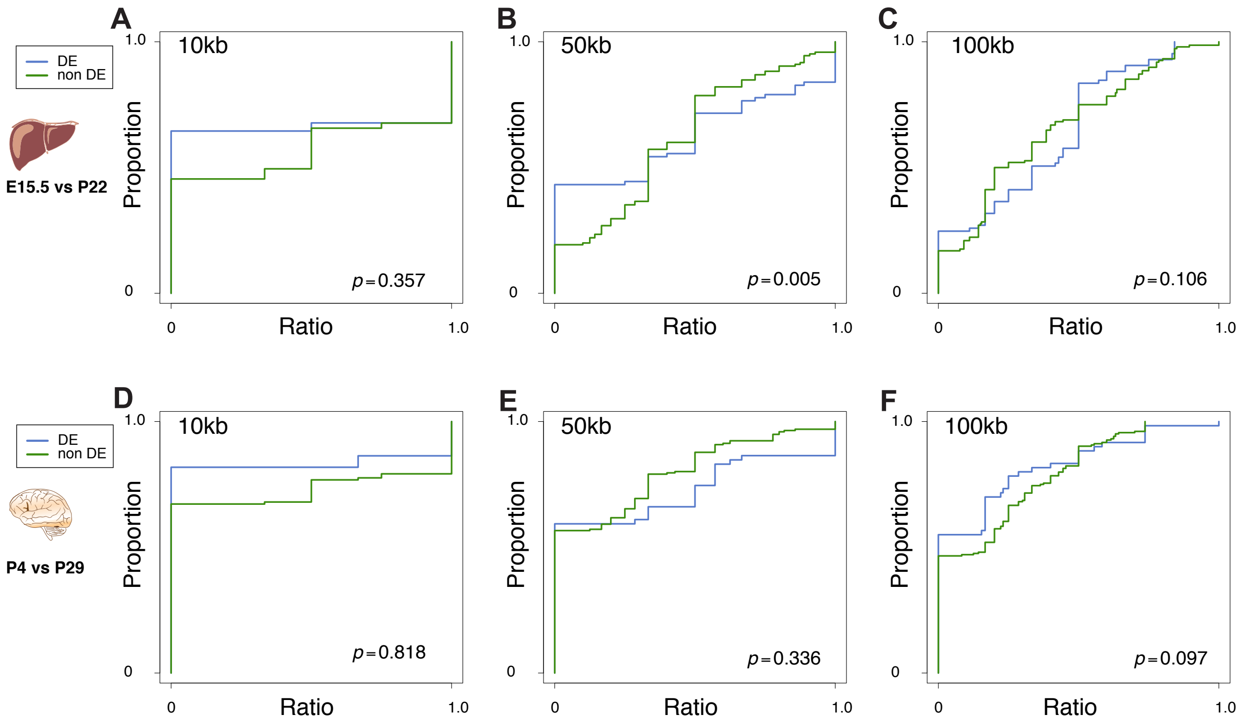


Figure A.14: Differentially expressed tRNA genes show no colocalisation with differentially expressed protein-coding genes. In each plot, the blue line is the cumulative distribution of the ratio of the number of upregulated mRNA genes to the number of all mRNA genes in the neighbourhood of each upregulated tRNA gene. The green line is the cumulative distribution of the ratios of the number of upregulated mRNA genes (FDR cutoff 0.01) to the number of all mRNA genes, in the neighbourhood of each tRNA gene that is not differentially expressed. Significant differences between these two distributions reveal situations where upregulated tRNA genes are significantly (by Kolmogorov–Smirnov test) associated with upregulated protein-coding genes. Different window sizes were used, ranging from 10 kb, 50 kb and 100 kb around tRNA genes. Pairwise comparison of (A–C) E15.5 and P22 in liver as well as (D–F) P4 and P29 in brain are shown. This analysis was repeated using two additional FDR cutoffs (0.05 and 0.0, data for liver in table 2.2, not shown for brain). Under the assumption that there was an observable colocalisation effect, we would expect there to be a robust signal, i.e. consistent significance across different tested parameters. However, of the 18 tests, only one was significant (corrected $p < 0.013$), after correcting for multiple testing (Bonferroni), indicating the absence of any strong localisation effect.

Supplementary material for chapter 3

B

B.1. Code

The code used in the analysis of the data for this chapter can be found at `https://github.com/klmr/codons`.

B.2. Supplementary tables

Table B.1: Gene identifiers for GO term "M phase of mitotic cell cycle".

Gene	Human ID	Mouse ID
MAD1L1	ENSG00000002822	ENSMUSG00000029554
PAFAH1B1	ENSG00000007168	ENSMUSG00000020745
NUP160	ENSG00000030066	ENSMUSG00000051329
CENPQ	ENSG00000031691	ENSMUSG00000023919
MPHOSPH9	ENSG00000051825	ENSMUSG00000038126
KIF2A	ENSG00000068796	ENSMUSG00000021693
NUP133	ENSG00000069248	ENSMUSG00000039509
SMC1A	ENSG00000072501	ENSMUSG00000041133
NDE1	ENSG00000072864	ENSMUSG00000022678
CLASP1	ENSG00000074054	ENSMUSG00000064302
NUP37	ENSG00000075188	ENSMUSG00000035351
NDC80	ENSG00000080986	ENSMUSG00000024056
XPO1	ENSG00000082898	ENSMUSG00000020290
SEH1L	ENSG00000085415	ENSMUSG00000079614
ZW10	ENSG00000086827	ENSMUSG00000032264
BIRC5	ENSG00000089685	ENSMUSG00000017716
NUDC	ENSG00000090273	ENSMUSG00000028851
CENPM	ENSG00000100162	ENSMUSG00000068101
RANGAP1	ENSG00000100401	ENSMUSG00000022391
CDC25B	ENSG00000101224	ENSMUSG00000027330
MAPRE1	ENSG00000101367	ENSMUSG00000027479
STAG2	ENSG00000101972	ENSMUSG00000025862
CENPI	ENSG00000102384	ENSMUSG00000031262
CENPT	ENSG00000102901	ENSMUSG00000036672
SMC3	ENSG00000108055	ENSMUSG00000024974
NUP98	ENSG00000110713	ENSMUSG00000063550
NUP107	ENSG00000111581	ENSMUSG00000052798
FBXO5	ENSG00000112029	ENSMUSG00000019773
KIF20A	ENSG00000112984	ENSMUSG00000003779

GORASP1	ENSG00000114745	ENSMUSG00000032513
CENPA	ENSG00000115163	ENSMUSG00000029177
CDC20	ENSG00000117399	ENSMUSG00000006398
NSL1	ENSG00000117697	ENSMUSG00000062510
CENPF	ENSG00000117724	ENSMUSG00000026605
STAG1	ENSG00000118007	ENSMUSG00000037286
NUP43	ENSG00000120253	ENSMUSG00000040034
CENPL	ENSG00000120334	ENSMUSG00000026708
KIF18A	ENSG00000121621	ENSMUSG00000027115
ZWINT	ENSG00000122952	ENSMUSG00000019923
CENPK	ENSG00000123219	ENSMUSG00000021714
B9D2	ENSG00000123810	ENSMUSG00000063439
NUP85	ENSG00000125450	ENSMUSG00000020739
DLGAP5	ENSG00000126787	ENSMUSG00000037544
SGOL1	ENSG00000129810	ENSMUSG00000023940
CLIP1	ENSG00000130779	ENSMUSG00000049550
CDCA8	ENSG00000134690	ENSMUSG00000028873
MPHOSPH6	ENSG00000135698	ENSMUSG00000031843
KIF23	ENSG00000137807	ENSMUSG00000032254
CASC5	ENSG00000137812	ENSMUSG00000027326
CENPO	ENSG00000138092	ENSMUSG00000020652
KIF20B	ENSG00000138182	ENSMUSG00000024795
CENPE	ENSG00000138778	ENSMUSG00000045328
KIF2B	ENSG00000141200	ENSMUSG00000046755
ITGB3BP	ENSG00000142856	ENSMUSG00000028549
KIF2C	ENSG00000142945	ENSMUSG00000028678
NUF2	ENSG00000143228	ENSMUSG00000026683
INCENP	ENSG00000149503	ENSMUSG00000024660
DSN1	ENSG00000149636	ENSMUSG00000027635
SPC25	ENSG00000152253	ENSMUSG00000005233
CENPH	ENSG00000153044	ENSMUSG00000045273
RANBP2	ENSG00000153201	ENSMUSG00000003226
AHCTF1	ENSG00000153207	ENSMUSG00000026491

BUB3	ENSG00000154473	ENSMUSG00000066979
SKA1	ENSG00000154839	ENSMUSG00000036223
BUB1B	ENSG00000156970	ENSMUSG00000040084
SEC13	ENSG00000157020	ENSMUSG00000030298
CDC25C	ENSG00000158402	ENSMUSG00000044201
TAOK1	ENSG00000160551	ENSMUSG00000017291
PMF1	ENSG00000160783	ENSMUSG00000028066
SPC24	ENSG00000161888	ENSMUSG00000074476
SGOL2	ENSG00000163535	ENSMUSG00000026039
CLASP2	ENSG00000163539	ENSMUSG00000033392
CDC25A	ENSG00000164045	ENSMUSG00000032477
MAD2L1	ENSG00000164109	ENSMUSG00000029910
RAD21	ENSG00000164754	ENSMUSG00000022314
CENPN	ENSG00000166451	ENSMUSG00000031756
NDEL1	ENSG00000166579	ENSMUSG00000018736
PLK1	ENSG00000166851	ENSMUSG00000030867
MIS12	ENSG00000167842	ENSMUSG00000040599
BUB1	ENSG00000169679	ENSMUSG00000027379
ZWILCH	ENSG00000174442	ENSMUSG00000032400
CKAP5	ENSG00000175216	ENSMUSG00000040549
APITD1	ENSG00000175279	ENSMUSG00000073705
RPS27	ENSG00000177954	ENSMUSG00000090733
AURKB	ENSG00000178999	ENSMUSG00000020897
RCC2	ENSG00000179051	ENSMUSG00000040945
SKA2	ENSG00000182628	ENSMUSG00000020492
KNTC1	ENSG00000184445	ENSMUSG00000029414
PPP1CC	ENSG00000186298	ENSMUSG00000004455
ERCC6L	ENSG00000186871	ENSMUSG00000051220
CENPP	ENSG00000188312	ENSMUSG00000021391
NDE1	ENSG00000275911	
CENPC1		ENSMUSG00000029253

Table B.3: Gene identifiers for GO term “M phase of mitotic cell cycle”.

Gene	Human ID	Mouse ID
FOXP3	ENSG00000049768	ENSMUSG00000039521
ALX4	ENSG00000052850	ENSMUSG00000040310
FOXN3	ENSG00000053254	ENSMUSG00000033713
FOXJ2	ENSG00000065970	ENSMUSG00000003154
TP63	ENSG00000073282	
GLI2	ENSG00000074047	ENSMUSG00000048402
TBX5	ENSG00000089225	ENSMUSG00000018263
CHRD	ENSG00000090539	ENSMUSG00000006958
CRKL	ENSG00000099942	ENSMUSG00000006134
MFNG	ENSG00000100060	ENSMUSG00000018169
MID1	ENSG00000101871	ENSMUSG00000035299
FOXF1	ENSG00000103241	ENSMUSG00000042812
LFNG	ENSG00000106003	ENSMUSG00000029570
HOXA5	ENSG00000106004	ENSMUSG00000038253
GLI3	ENSG00000106571	ENSMUSG00000021318
FOXN1	ENSG00000109101	ENSMUSG00000002057
FOXM1	ENSG00000111206	ENSMUSG00000001517
BMP5	ENSG00000112175	ENSMUSG00000032179
CDX1	ENSG00000113722	ENSMUSG00000024619
HES1	ENSG00000114315	ENSMUSG00000022528
ACVR2B	ENSG00000114739	ENSMUSG00000061393
FOXP1	ENSG00000114861	ENSMUSG00000030067
FOXO3	ENSG00000118689	ENSMUSG00000048756
HOXB1	ENSG00000120094	ENSMUSG00000018973
FOXA2	ENSG00000125798	ENSMUSG00000037025
PAX1	ENSG00000125813	ENSMUSG00000037034
SIX1	ENSG00000126778	ENSMUSG00000051367
FOXP2	ENSG00000128573	ENSMUSG00000029563
SMO	ENSG00000128602	ENSMUSG00000001761
HOXD11	ENSG00000128713	ENSMUSG00000042499

HOXD13	ENSG00000128714	ENSMUSG00000001819
FOXA1	ENSG00000129514	ENSMUSG00000035451
FOXJ1	ENSG00000129654	ENSMUSG00000034227
RAX	ENSG00000134438	ENSMUSG00000024518
HEY2	ENSG00000135547	ENSMUSG00000019789
NKX2-1	ENSG00000136352	ENSMUSG00000001496
FOXP4	ENSG00000137166	ENSMUSG00000023991
FOXF2	ENSG00000137273	ENSMUSG00000038402
SHROOM3	ENSG00000138771	ENSMUSG00000029381
ASCL1	ENSG00000139352	ENSMUSG00000020052
FOXN4	ENSG00000139445	ENSMUSG00000042002
FOXK2	ENSG00000141568	ENSMUSG00000039275
ZEB1	ENSG00000148516	ENSMUSG00000024238
FOXO1	ENSG00000150907	ENSMUSG00000044167
ZIC1	ENSG00000152977	ENSMUSG00000032368
ZIC3	ENSG00000156925	ENSMUSG00000067860
GRHL3	ENSG00000158055	ENSMUSG00000037188
FOXQ1	ENSG00000164379	ENSMUSG00000038415
SHH	ENSG00000164690	ENSMUSG00000002633
UNCX	ENSG00000164853	ENSMUSG00000029546
FOXK1	ENSG00000164916	ENSMUSG00000056493
CDX2	ENSG00000165556	ENSMUSG00000029646
FOXI1	ENSG00000168269	ENSMUSG00000047861
RFNG	ENSG00000169733	ENSMUSG00000025158
FOXD4	ENSG00000170122	ENSMUSG00000051490
HOXD12	ENSG00000170178	ENSMUSG00000001823
FOXN2	ENSG00000170802	ENSMUSG00000034998
SOSTDC1	ENSG00000171243	ENSMUSG00000036169
FOXB1	ENSG00000171956	ENSMUSG00000059246
FOXR1	ENSG00000176302	ENSMUSG00000074397
FOXL1	ENSG00000176678	ENSMUSG00000097084
FOXE1	ENSG00000178919	ENSMUSG00000070990
NOG	ENSG00000183691	ENSMUSG00000048616

FOXL2	ENSG00000183770	ENSMUSG00000050397
TBX1	ENSG00000184058	ENSMUSG00000009097
FOXO4	ENSG00000184481	ENSMUSG00000042903
FOXD4L1	ENSG00000184492	
FOXD4L4	ENSG00000184659	
NRG3	ENSG00000185737	ENSMUSG00000041014
PTCH1	ENSG00000185920	ENSMUSG00000021466
FOXD2	ENSG00000186564	ENSMUSG00000055210
FOXI2	ENSG00000186766	ENSMUSG00000048377
FOXE3	ENSG00000186790	ENSMUSG00000044518
FOXD3	ENSG00000187140	ENSMUSG00000067261
FOXR2	ENSG00000189299	ENSMUSG00000071665
SYNGAP1	ENSG00000197283	ENSMUSG00000067629
ASPH	ENSG00000198363	ENSMUSG00000028207
FOXJ3	ENSG00000198815	ENSMUSG00000032998
FOXO6	ENSG00000204060	ENSMUSG00000052135
FOXB2	ENSG00000204612	ENSMUSG00000056829
FOXD4L5	ENSG00000204779	
FOXI3	ENSG00000214336	ENSMUSG00000055874
FOXD1	ENSG00000251493	ENSMUSG00000078302
FOXD4L6	ENSG00000273514	
LHX1	ENSG00000273706	ENSMUSG00000018698
LHX1	ENSG00000274577	

C Supplementary material for chapter 4

C.1. Code

The code used in the analysis of the data for this chapter can be found at `https://github.com/klmr/pol3-seq`.

C.2. Supplementary figures

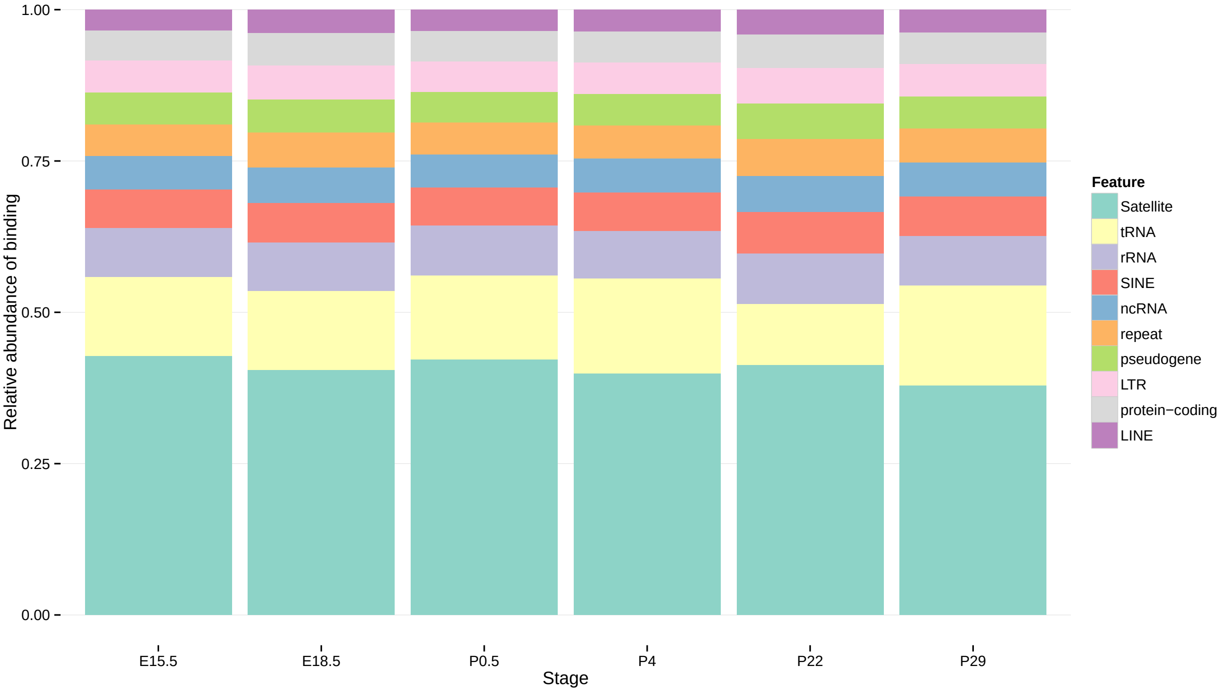


Figure C.1: Input library coverage of different features in six stages of development in liver. The analysis was performed under the assumption that different features have similar amount of input binding (normalised for feature length). As we can see here, this is not quite the case.

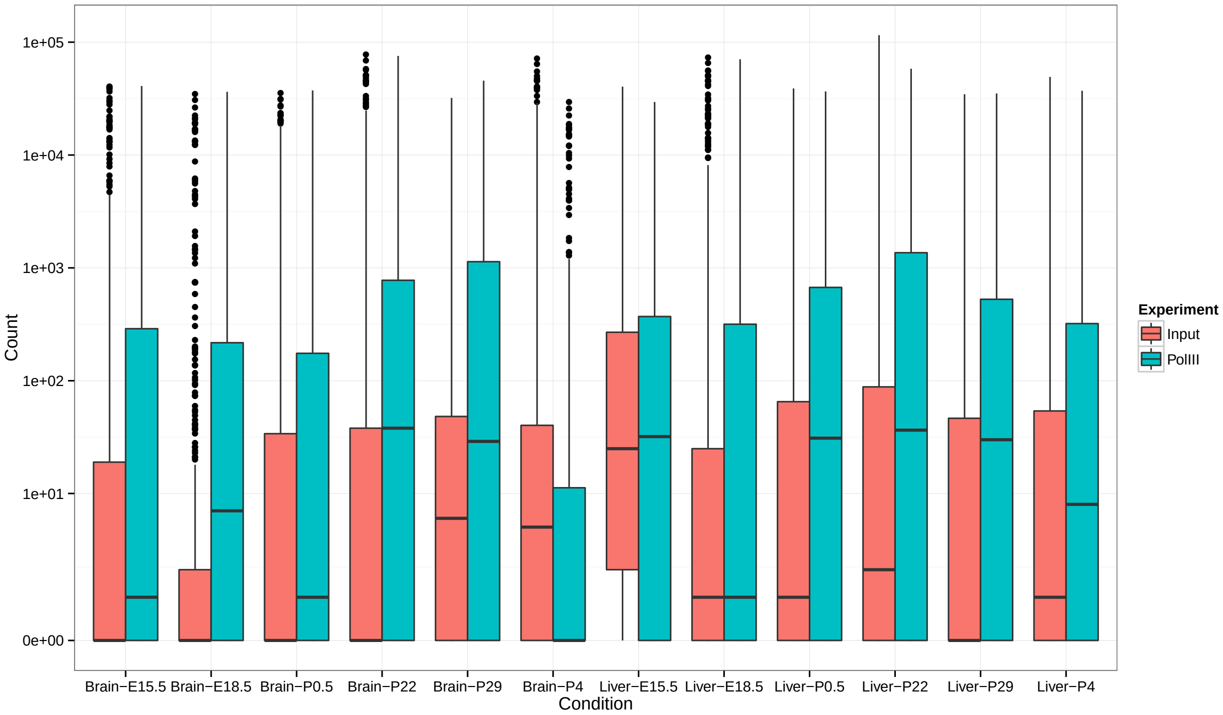


Figure C.2: SINE binding by pol III across development in liver and brain. Raw counts of pol III binding to different SINE classes, including those classes where no binding occurs.

Bibliography

Ablasser, Andrea, Franz Bauernfeind, Gunther Hartmann, Eicke Latz, Katherine A Fitzgerald, and Veit Hornung (2009). "RIG-I-dependent sensing of poly(dA:dT) through the induction of an RNA polymerase III-transcribed RNA intermediate". In: *Nature immunology* 10.10, pp. 1065–1072. ISSN: 1529-2908, 1529-2916. DOI: 10.1038/ni.1779.

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter (2002). *Molecular Biology of the Cell*. Garland Science.

Anders, Simon and Wolfgang Huber (2010). "Differential expression analysis for sequence count data". In: *Genome Biology* 11.10, R106. ISSN: 1465-6906. DOI: 10.1186/gb-2010-11-10-r106.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber (2014). "HTSeq – A Python framework to work with high-throughput sequencing data".

Bailey, Timothy L, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble (2009). "MEME SUITE: tools for motif discovery and searching". In: *Nucleic acids research* 37.Web Server issue, W202–8. ISSN: 0305-1048. DOI: 10.1093/nar/gkp335.

Barski, Artem, Iouri Chepelev, Dritan Liko, Suresh Cuddapah, Alastair B Fleming, Joanna Birch, Kairong Cui, Robert J White, and Keji Zhao (2010). "Pol II and its associated epigenetic marks are present at pol III-transcribed noncoding RNA genes". In: *Nature structural & molecular biology* 17.5, pp. 629–634. ISSN: 1545-9993, 1545-9985. DOI: 10.1038/nsmb.1806.

Barski, Artem, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao (2007). "High-resolution profiling of histone methylations in the human genome". In: *Cell* 129.4, pp. 823–837. ISSN: 0092-8674. DOI: 10.1016/j.cell.2007.05.009.

Berg, Jeremy M, John L Tymoczko, and Lubert Stryer (2002). *Biochemistry*. W H Freeman.

Blanco, Sandra, Sabine Dietmann, Joana V Flores, Shobbir Hussain, Claudia Kutter, Peter Humphreys, Margus Lukk, Patrick Lombard, Lucas Treps, Martyna Popis, Stefanie Kellner, Sabine M Hölter, Lillian Garrett, Wolfgang Wurst, Lore Becker, Thomas Klopstock, Helmut Fuchs, Valerie Gailus-Durner, Martin Hrabě de Angelis, Ragnhildur T Káradóttir, Mark Helm, Jernej Ule, Joseph G Gleeson, Duncan T Odom, and Michaela Frye (2014). "Aberrant methylation of tRNAs links cellular stress to neuro-developmental disorders". In: *The EMBO journal* 33.18, pp. 2020–2039. issn: 0261-4189, 1460-2075. doi: 10.15252/embj.201489282.

Blencowe, B J (2000). "Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases". In: *Trends in biochemical sciences* 25.3, pp. 106–110. issn: 0968-0004.

Borland, David and M Russell Taylor 2nd (2007). "Rainbow color map (still) considered harmful". In: *IEEE computer graphics and applications* 27.2, pp. 14–17. issn: 0272-1716.

Bourgon, Richard, Robert Gentleman, and Wolfgang Huber (2010). "Independent filtering increases detection power for high-throughput experiments". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.21, pp. 9546–51. issn: 0027-8424, 1091-6490. doi: 10.1073/pnas.0914005107.

Brawand, David, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, Frank W Albert, Ulrich Zeller, Philipp Khaitovich, Frank Grützner, Sven Bergmann, Rasmus Nielsen, Svante Pääbo, and Henrik Kaessmann (2011). "The evolution of gene expression levels in mammalian organs". In: *Nature* 478.7369, pp. 343–348. issn: 0028-0836. doi: 10.1038/nature10532.

Bray, Nicolas, Harold Pimentel, Páll Melsted, and Lior Pachter (2015). "Near-optimal RNA-Seq quantification". In: arXiv: 1505.02710 [q-bio.QM].

Byun, Yanga and Kyungsook Han (2009). "PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots". In: *Bioinformatics* 25.11, pp. 1435–1437. issn: 1367-4803.

Canella, Donatella, David Bernasconi, Federica Gilardi, Gwendal LeMartelot, Eugenia Migliavacca, Viviane Praz, Pascal Cousin, Mauro Delorenzi, Nouria Hernandez, and CycliX Consortium (2012). "A multiplicity of factors contributes to selective RNA polymerase III occupancy of a subset of RNA polymerase III genes in mouse liver". In: *Genome research* 22.4, pp. 666–680. issn: 1088-9051. doi: 10.1101/gr.130286.111.

Carrière, Lucie, Sébastien Graziani, Olivier Alibert, Yad Ghavi-Helm, Fayçal Boussouar, Hélène Humbertclaude, Sylvie Jounier, Jean-Christophe Aude, Céline Keime, Janos Murvai, Mario Foglio, Marta Gut, Ivo Gut, Mark Lathrop, Julie Soutourina, Matthieu Gérard, and Michel Werner (2012). "Genomic binding of Pol III transcription machinery and relationship with TFIIS transcription factor distribution in mouse embryonic stem cells". In: *Nucleic acids research* 40.1, pp. 270–283. issn: 0305-1048, 1362-4962. doi: 10.1093/nar/gkr737.

Carroll, Sean B (2006). *The making of the fittest : DNA and the ultimate forensic record of evolution*. New York, N.Y.: W.W. Norton & Co.

Casneuf, Tineke, Yves Van de Peer, and Wolfgang Huber (2007). "In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation". In: *BMC bioinformatics* 8, p. 461. issn: 1471-2105. doi: 10.1186/1471-2105-8-461.

Chan, Patricia P and Todd M Lowe (2009). "GtRNAdb: a database of transfer RNA genes detected in genomic sequence". In: *Nucleic acids research* 37.Database issue, pp. D93–7. issn: 0305-1048, 1362-4962. doi: 10.1093/nar/gkn787.

Chen, Hongmin, Jack O Egan, and Jen-Fu Chiu (1997). "Regulation and Activities of α-Fetoprotein". In: *Critical Reviews in Eukaryotic Gene Expression* 7.1-2. issn: 1045-4403, 2162-6502.

Chen, Swaine L, William Lee, Alison K Hottes, Lucy Shapiro, and Harley H McAdams (2004). "Codon usage between genomes is constrained by genome-wide mutational processes". In: *Proceedings of the National Academy* *of Sciences of the United States of America* 101.10, pp. 3480–3485. issn: 0027-8424. doi: 10.1073/pnas.0307827100.

Cheng, Jill, Philipp Kapranov, Jorg Drenkow, Sujit Dike, Shane Brubaker, Sandeep Patel, Jeffrey Long, David Stern, Hari Tammana, Gregg Helt, Victor Sementchenko, Antonio Piccolboni, Stefan Bekiranov, Dione K Bailey, Madhavan Ganesh, Srinka Ghosh, Ian Bell, Daniela S Gerhard, and Thomas R Gingeras (2005). "Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution". In: *Science* 308.5725, pp. 1149–1154. issn: 0036-8075, 1095-9203. doi: 10.1126/science.1108625.

Church, Deanna M, Leo Goodstadt, Ladeana W Hillier, Michael C Zody, Steve Goldstein, Xinwe She, Carol J Bult, Richa Agarwala, Joshua L Cherry, Michael DiCuccio, Wratko Hlavina, Yuri Kapustin, Peter Meric, Donna Maglott, Zoë Birtle, Ana C Marques, Tina Graves, Shiguo Zhou, Brian Teague, Konstantinos Potamousis, Christopher Churas, Michael Place, Jill Herschleb, Ron Runnheim, Daniel Forrest, James Amos-Landgraf, David C Schwartz, Ze Cheng, Kerstin Lindblad-Toh, Evan E Eichler, Chris P Ponting, and Mouse Genome Sequencing Consortium (2009). "Lineage-specific biology revealed by a finished genome assembly of the mouse". In: *PLoS biology* 7.5, e1000112. issn: 1544-9173, 1545-7885. doi: 10.1371/journal.pbio.1000112.

Cox, Anthony J (2007). "ELAND – Efficient Large-Scale Alignment of Nucleotide Databases".

Crick, Francis H C (1958). "On Protein Synthesis". In: *Symp Soc Exp Biol* XII, pp. 138–163.

– (1966). "Codon–anticodon pairing: the wobble hypothesis". In: *Journal of molecular biology* 19.2, pp. 548–555. issn: 0022-2836.

– (1970). "Central dogma of molecular biology". In: *Nature* 227.5258, pp. 561–563. issn: 0028-0836.

Csárdi, Gábor, Alexander Franks, David S Choi, Eduardo M Airoldi, and D Allan Drummond (2014). "Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast". In: *bioRxiv*.

Cui, Peng, Qiang Lin, Feng Ding, Chengqi Xin, Wei Gong, Lingfang Zhang, Jianing Geng, Bing Zhang, Xiaomin Yu, Jin Yang, Songnian Hu, and Jun Yu (2010). "A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing". In: *Genomics* 96.5, pp. 259–265. issn: 0888-7543, 1089-8646. doi: 10.1016/j.ygeno.2010.07.010.

Dalluge, J J, T Hamamoto, K Horikoshi, R Y Morita, K O Stetter, and J A McCloskey (1997). "Posttranscriptional modification of trna in psychrophilic bacteria". In: *Journal of bacteriology* 179.6, pp. 1918–1923. issn: 0021-9193.

Dieci, Giorgio, Gloria Fiorino, Manuele Castelnuovo, Martin Teichmann, and Aldo Pagano (2007). "The expanding RNA polymerase III transcriptome". In: *Trends in Genetics* 23.12, pp. 614–22. issn: 0168-9525. doi: 10.1016/j.tig.2007.09.001.

Dittmar, Kimberly A, Jeffrey M Goodenbour, and Tao Pan (2006). "Tissue-specific differences in human transfer RNA expression". In: *PLoS genetics* 2.12, e221. issn: 1553-7390. doi: 10.1371/journal.pgen.0020221.

Dobzhansky, Theodosius (1973). "Nothing in Biology Makes Sense except in the Light of Evolution". In: *The American biology teacher* 35.3, pp. 125–129. issn: 0002-7685.

Dos Reis, Mario, Renos Savva, and Lorenz Wernisch (2004). "Solving the riddle of codon usage preferences: a test for translational selection". In: *Nucleic acids research* 32.17, pp. 5036–5044. issn: 0305-1048. doi: 10.1093/nar/gkh834.

Dos Reis, Mario, Lorenz Wernisch, and Renos Savva (2003). "Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome". In: *Nucleic acids research* 31.23, pp. 6976–6985. issn: 0305-1048, 1362-4962. doi: 10.1093/nar/gkg897.

Duret, Laurent (2002). "Evolution of synonymous codon usage in metazoans". In: *Current opinion in genetics & development* 12.6, pp. 640–649. issn: 0959-437X.

Eddy, Sean R and Richard Durbin (1994). "RNA sequence analysis using covariance models". In: *Nucleic acids research* 22.11, pp. 2079–2088. issn: 0305-1048. doi: 10.1093/nar/22.11.2079.

Ermolaeva, M D (2001). “Synonymous codon usage in bacteria”. In: *Current issues in molecular biology* 3.4, pp. 91–97. issn: 1467-3037.

Flicek, Paul, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah Hunt, Nathan Johnson, Thomas Juettemann, Andreas K Kähäri, Stephen Keenan, Eugene Kulesha, Fergal J Martin, Thomas Maurel, William M McLaren, Daniel N Murphy, Rishi Nag, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Emily Pritchard, Harpreet S Riat, Magali Ruffier, Daniel Sheppard, Kieron Taylor, Anja Thormann, Stephen J Trevanion, Alessandro Vullo, Steven P Wilder, Mark Wilson, Amonida Zadissa, Bronwen L Aken, Ewan Birney, Fiona Cunningham, Jennifer Harrow, Javier Herrero, Tim J P Hubbard, Rhoda Kinsella, Matthieu Muffato, Anne Parker, Giulietta Spudich, Andy Yates, Daniel R Zerbino, and Stephen M J Searle (2014). “Ensembl 2014”. In: *Nucleic acids research* 42.Database issue, pp. D749–55. issn: 0305-1048, 1362-4962. doi: 10.1093/nar/gkt1196.

Fonseca, Nuno A., Robert Petryszak, John Marioni, and Alvis Brazma (2014). “iRAP - a meta-pipeline for RNA-seq data analysis”. In: *Medical mission to Austria, July 1-August 8, 1947. American Unitarian Service Committee in cooperation with World Health Organization Interim Commission. [Abridged report submitted by Erwin Kohn]. American Unitarian Association. Unitari...*

Gingold, Hila and Yitzhak Pilpel (2011). “Determinants of translation efficiency and accuracy”. In: *Molecular systems biology* 7, p. 481. issn: 1744-4292. doi: 10.1038/msb.2011.14.

Gingold, Hila, Disa Tehler, Nanna R Christoffersen, Morten M Nielsen, Fazila Asmar, Susanne M Kooistra, Nicolaj S Christophersen, Lise Lotte Christensen, Michael Borre, Karina D Sørensen, Lars D Andersen, Claus L Andersen, Esther Hulleman, Tom Wurdinger, Elisabeth Ralfkiær, Kristian Helin, Kirsten Grønbæk, Torben Ørntoft, Sebastian M Waszak, Orna Dahan, Jakob Skou Pedersen, Anders H Lund, and Yitzhak Pilpel (2014). “A Dual Program for Translation Regulation in Cellular Proliferation and Differentiation”. In: *Cell* 158.6, pp. 1281–1292. issn: 0092-8674. doi: 10.1016/j.cell.2014.08.011.

Girard, J, P Ferré, J P Pégorier, and P H Duée (1992). "Adaptations of glucose and fatty acid metabolism during perinatal period and suckling-weaning transition". In: *Physiological reviews* 72.2, pp. 507–562. issn: 0031-9333.

Graumann, Johannes, Nina C Hubner, Jeong Beom Kim, Kinarm Ko, Markus Moser, Chanchal Kumar, Jürgen Cox, Hans Schöler, and Matthias Mann (2008). "Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins". In: *Molecular & cellular proteomics: MCP* 7.4, pp. 672–683. issn: 1535-9476, 1535-9484. doi: 10.1074/mcp.M700460-MCP200.

Gupta, Shobhit, John A Stamatoyannopoulos, Timothy L Bailey, and William Stafford Noble (2007). "Quantifying similarity between motifs". In: *Genome biology* 8.2, R24. issn: 1465-6906. doi: 10.1186/gb-2007-8-2-r24.

Hamilton, A J and D C Baulcombe (1999). "A species of small antisense RNA in posttranscriptional gene silencing in plants". In: *Science* 286.5441, pp. 950–952. issn: 0036-8075.

Hammond, S M, E Bernstein, D Beach, and G J Hannon (2000). "An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells". In: *Nature* 404.6775, pp. 293–296. issn: 0028-0836. doi: 10.1038/35005107.

Hohl, D, P A de Viragh, F Amiguet-Barras, S Gibbs, C Backendorf, and M Huber (1995). "The small proline-rich proteins constitute a multigene family of differentially regulated cornified cell envelope precursor proteins". In: *The Journal of investigative dermatology* 104.6, pp. 902–909. issn: 0022-202X.

Horn, David (2008). "Codon usage suggests that translational selection has a major impact on protein expression in trypanosomatids". In: *BMC genomics* 9, p. 2. issn: 1471-2164. doi: 10.1186/1471-2164-9-2.

Huang, K-P, F L Huang, and P K Shetty (2011). "Stimulation-mediated translocation of calmodulin and neurogranin from soma to dendrites of mouse hippocampal CA1 pyramidal neurons". In: *Neuroscience* 178, pp. 1–12. issn: 0306-4522, 1873-7544. doi: 10.1016/j.neuroscience.2011.01.027.

Hyder, Salman M, Zafar Nawaz, Constance Chiappetta, Koshinaga Yokoyama, and George M Stancel (1995). "The Protooncogene c- jun Contains an Unusual Estrogen-inducible Enhancer within the Coding Sequence". In: *The Journal of biological chemistry* 270.15, pp. 8506–8513. issn: 0021-9258.

Ibba, Michael and Dieter Söll (2000). "Aminoacyl-trna synthesis". In: *Annual review of biochemistry* 69, pp. 617–650. issn: 0066-4154. doi: 10.1146/annurev.biochem.69.1.617.

Ikemura, T (1981a). "Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes". In: *Journal of molecular biology* 146.1, pp. 1–21. issn: 0022-2836.

– (1981b). "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". In: *Journal of molecular biology* 151.3, pp. 389–409. issn: 0022-2836.

– (1985). "Codon usage and trna content in unicellular and multicellular organisms". In: *Molecular biology and evolution* 2.1, pp. 13–34. issn: 0737-4038.

Jantzen, Stuart G, Ben Jg Sutherland, David R Minkley, and Ben F Koop (2011). "GO Trimming: Systematically reducing redundancy in large Gene Ontology datasets". In: *BMC research notes* 4, p. 267. issn: 1756-0500. doi: 10.1186/1756-0500-4-267.

Johnson, David S, Ali Mortazavi, Richard M Myers, and Barbara Wold (2007). "Genome-wide mapping of in vivo protein-DNA interactions". In: *Science* 316.5830, pp. 1497–1502. issn: 0036-8075, 1095-9203. doi: 10.1126/science.1141319.

Jovanovic, Marko, Michael S Rooney, Philipp Mertins, Dariusz Przybylski, Nicolas Chevrier, Rahul Satija, Edwin H Rodriguez, Alexander P Fields, Schraga Schwartz, Raktima Raychowdhury, Maxwell R Mumbach, Thomas Eisenhaure, Michal Rabani, Dave Gennert, Diana Lu, Toni Delorey, Jonathan S Weissman, Steven A Carr, Nir Hacohen, and Aviv Regev (2015). "Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens". In: *Science* 347.6226, p. 1259038. issn: 0036-8075, 1095-9203. doi: 10.1126/science.1259038.

Jurka, J, V V Kapitonov, A Pavlicek, P Klonowski, O Kohany, and J Walichiewicz (2005). "Repbase Update, a database of eukaryotic repetitive elements". In: *Cytogenetic and genome research* 110.1-4, pp. 462–467. issn: 1424-8581, 1424-859X. doi: 10.1159/000084979.

Kahles, Andre, Jonas Behr, and Gunnar Rätsch (2015). "MMR: A Tool for Read Multi-Mapper Resolution". In: *bioRxiv*.

Katz, Yarden, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge (2010). "Analysis and design of RNA sequencing experiments for identifying isoform regulation". In: *Nature methods* 7.12, pp. 1009–1015. issn: 1548-7091, 1548-7105. doi: 10.1038/nmeth.1528.

Kim, Daehwan, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". In: *Genome Biol.* 14.4 (4), R36. issn: 1465-6906. doi: 10.1186/gb-2013-14-4-r36.

Kim, Sung-Hou, G J Quigley, F L Suddath, A McPherson, D Sneden, J J Kim, J Weinzierl, and Alexander Rich (1973). "Three-dimensional structure of yeast phenylalanine transfer RNA: folding of the polynucleotide chain". In: *Science* 179.4070, pp. 285–288. issn: 0036-8075.

Knott, T J, R J Pease, L M Powell, S C Wallis, S C Rall Jr, T L Innerarity, B Blackhart, W H Taylor, Y Marcel, and R Milne (1986). "Complete protein sequence and identification of structural domains of human apolipoprotein B". In: *Nature* 323.6090, pp. 734–738. issn: 0028-0836. doi: 10.1038/323734a0.

Kolitz, Sarah E and Jon R Lorsch (2010). "Eukaryotic initiator trna: finely tuned and ready for action". In: *FEBS letters* 584.2, pp. 396–404. issn: 0014-5793, 1873-3468. doi: 10.1016/j.febslet.2009.11.047.

Kozak, Marilyn (2002). "Pushing the limits of the scanning mechanism for initiation of translation". In: *Gene* 299.1-2, pp. 1–34. issn: 0378-1119.

Kutter, Claudia, Gordon D Brown, Ângela Gonçalves, Michael D Wilson, Stephen Watt, Alvis Brazma, Robert J White, and Duncan T Odom (2011). "Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among trna genes". In: *Nature Genetics* 43.10, pp. 948–55. issn: 1061-4036. doi: 10.1038/ng.906.

Lander, E S & al. (2001). “Initial sequencing and analysis of the human genome”. In: *Nature* 409.6822, pp. 860–921. issn: 0028-0836. doi: 10.1038/35057062.

Landt, Stephen G, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, Peter Bickel, James B Brown, Philip Cayting, Yiwen Chen, Gilberto DeSalvo, Charles Epstein, Katherine I Fisher-Aylor, Ghia Euskirchen, Mark Gerstein, Jason Gertz, Alexander J Hartemink, Michael M Hoffman, Vishwanath R Iyer, Youngsook L Jung, Subhradip Karmakar, Manolis Kellis, Peter V Kharchenko, Qunhua Li, Tao Liu, X Shirley Liu, Lijia Ma, Aleksandar Milosavljevic, Richard M Myers, Peter J Park, Michael J Pazin, Marc D Perry, Debasish Raha, Timothy E Reddy, Joel Rozowsky, Noam Shoresh, Arend Sidow, Matthew Slattery, John A Stamatoyannopoulos, Michael Y Tolstorukov, Kevin P White, Simon Xi, Peggy J Farnham, Jason D Lieb, Barbara J Wold, and Michael Snyder (2012). “ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia”. In: *Genome research* 22.9, pp. 1813–1831. issn: 1088-9051. doi: 10.1101/gr.136184.111.

Langmead, Ben and Steven L Salzberg (2012). “Fast gapped-read alignment with Bowtie 2”. In: *Nature methods* 9.4, pp. 357–359. issn: 1548-7091, 1548-7105. doi: 10.1038/nmeth.1923.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg (2009). “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome”. In: *Genome biology* 10.3, R25. issn: 1465-6906, 1465-6914. doi: 10.1186/gb-2009-10-3-r25.

Li, Bo, Victor Ruotti, Ron M Stewart, James A Thomson, and Colin N Dewey (2010). “RNA-Seq gene expression estimation with read mapping uncertainty”. In: *Bioinformatics* 26.4, pp. 493–500. issn: 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btp692.

Li, Heng and Richard Durbin (2009). “Fast and accurate short read alignment with Burrows-Wheeler transform”. In: *Bioinformatics* 25.14 (14), pp. 1754–1760. issn: 1367-4803. doi: 10.1093/bioinformatics/btp324.

Li, Jingyi Jessica, Peter J Bickel, and Mark D Biggin (2014). “System wide analyses have underestimated protein abundances and the importance of transcription in mammals". In: *PeerJ* 2, e270. issn: 2167-8359. doi: 10.7717/peerj.270.

Limbach, P A, P F Crain, and J A McCloskey (1994). "Summary: the modified nucleosides of RNA". In: *Nucleic acids research* 22.12, pp. 2183–2196. issn: 0305-1048. doi: 10.1093/nar/22.12.2183.

Liscovitch, Noa and Gal Chechik (2013). "Specialization of gene expression during mouse brain development". In: *PLoS computational biology* 9.9, e1003185. issn: 1553-734X, 1553-7358. doi: 10.1371/journal.pcbi.1003185.

Litvak, S, L Sarih-Cottin, M Fournier, M Andreola, and L Tarrago-Litvak (1994). "Priming of HIV replication by trna(Lys3): role of reverse transcriptase". In: *Trends in biochemical sciences* 19.3, pp. 114–118. issn: 0968-0004.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2".

Marioni, John C, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays". In: *Genome research* 18.9, pp. 1509–1517. issn: 1088-9051. doi: 10.1101/gr.079558.108.

Mattick, John S and Igor V Makunin (2006). "Non-coding RNA". In: *Human molecular genetics* 15 Spec No 1, R17–29. issn: 0964-6906. doi: 10.1093/hmg/ddl046.

Meselson, M and F W Stahl (1958). "The replication of DNA in Escherichia coli". In: *Proceedings of the National Academy of Sciences of the United States of America* 44.7, pp. 671–682. issn: 0027-8424. doi: 10.1073/pnas.44.7.671.

Milo, Ron (2013). "What is the total number of protein molecules per cell volume? A call to rethink some published values". In: *BioEssays: news and reviews in molecular, cellular and developmental biology* 35.12, pp. 1050–1055. issn: 0265-9247, 1521-1878. doi: 10.1002/bies.201300066.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature methods* 5.7, pp. 621–8. issn: 1548-7091. doi: 10.1038/nmeth.1226.

Murphy 4th, Frank V and V Ramakrishnan (2004). “Structure of a purine–purine wobble base pair in the decoding center of the ribosome”. In: *Nature structural & molecular biology* 11.12, pp. 1251–1252. issn: 1545-9993, 1545-9985. doi: 10.1038/nsmb866.

Nagalakshmi, Ugrappa, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder (2008). “The transcriptional landscape of the yeast genome defined by RNA sequencing”. In: *Science* 320.5881, pp. 1344–1349. issn: 0036-8075, 1095-9203. doi: 10.1126/science.1158441.

Nagaraj, Nagarjuna, Jacek R Wisniewski, Tamar Geiger, Juergen Cox, Martin Kircher, Janet Kelso, Svante Pääbo, and Matthias Mann (2011). “Deep proteome and transcriptome mapping of a human cancer cell line”. In: *Molecular systems biology* 7, p. 548. issn: 1744-4292. doi: 10.1038/msb.2011.81.

Nookaew, Intawat, Marta Papini, Natapol Pornputtapong, Gionata Scalcinati, Linn Fagerberg, Matthias Uhlén, and Jens Nielsen (2012). “A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae”. In: *Nucleic acids research* 40.20, pp. 10084–10097. issn: 0305-1048, 1362-4962. doi: 10.1093/nar/gks804.

Oler, Andrew J, Ravi K Alla, Douglas N Roberts, Alexander Wong, Peter C Hollenhorst, Katherine J Chandler, Patrick A Cassiday, Cassie A Nelson, Curt H Hagedorn, Barbara J Graves, and Bradley R Cairns (2010). “Human RNA polymerase III transcriptomes and relationships to pol II promoter chromatin and enhancer-binding factors”. In: *Nature structural & molecular biology* 17.5, pp. 620–628. issn: 1545-9993. doi: 10.1038/nsmb.1801.

Osawa, S, T Ohama, T H Jukes, and K Watanabe (1989). “Evolution of the mitochondrial genetic code. I. Origin of AGR serine and stop codons in metazoan mitochondria”. In: *Journal of molecular evolution* 29.3, pp. 202–207. issn: 0022-2844.

Palida, Fakhruddin A, Charles Hale, and Karen U Sprague (1993). “Transcription of a silkworm $\text{tRNA}^{\text{Ala}}_{\text{C}}$ gene is directed by two AT-rich upstream sequence elements”. In: *Nucleic acids research* 21.25, pp. 5875–5881. issn: 0305-1048. doi: 10.1093/nar/21.25.5875.

Park, Peter J (2009). "ChIP–seq: advantages and challenges of a maturing technology". In: *Nature reviews. Genetics* 10.10, pp. 669–680. issn: 1471-0056, 1471-0064. doi: 10.1038/nrg2641.

Patro, Rob, Stephen M Mount, and Carl Kingsford (2014). "Sailfish enables alignment-free isoform quantification from RNA-seq reads using light-weight algorithms". In: *Nature biotechnology* 32.5, pp. 462–464. issn: 1087-0156, 1546-1696. doi: 10.1038/nbt.2862.

Pavon-Eternod, Mariana, Suzana Gomes, Marsha R Rosner, and Tao Pan (2013). "Overexpression of initiator methionine trna leads to global reprogramming of trna expression and increased proliferation in human epithelial cells". In: *RNA* 19.4, pp. 461–466. issn: 1355-8382, 1355-8382. doi: 10.1261/rna.037507.112.

Pavon-Eternod, Mariana, Suzanna Gomes, Renaud Geslain, Qing Dai, Marsha Rich Rosner, and Tao Pan (2009). "trna over-expression in breast cancer and functional consequences". In: *Nucleic Acids Res* 37.21, pp. 7268–80. issn: 0305-1048, 1362-4962. doi: 10.1093/nar/gkp787.

Pickrell, Joseph K, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard (2010). "Understanding mechanisms underlying human gene expression variation with RNA sequencing". In: *Nature* 464.7289, pp. 768–772. issn: 0028-0836, 1476-4687. doi: 10.1038/nature08872.

Plotkin, Joshua B and Grzegorz Kudla (2010). "Synonymous but not the same: the causes and consequences of codon bias". In: *Nature reviews. Genetics* 12.1, pp. 32–42. issn: 1471-0056. doi: 10.1038/nrg2899.

Raha, Debasish, Zhong Wang, Zarmik Moqtaderi, Linfeng Wu, Guoneng Zhong, Mark Gerstein, Kevin Struhl, and Michael Snyder (2010). "Close association of RNA polymerase II and many transcription factors with Pol III genes". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.8, pp. 3639–3644. issn: 0027-8424, 1091-6490. doi: 10.1073/pnas.0911315106.

Reeves, M A and P R Hoffmann (2009). "The human selenoproteome: recent insights into functions and regulation". In: *Cellular and molecular life* *sciences: CMLS* 66.15, pp. 2457–2478. issn: 1420-682X, 1420-9071. doi: 10.1007/s00018-009-0032-4.

Rich, Alexander and Sung-Hou Kim (1978). "The three-dimensional structure of transfer RNA". In: *Scientific American* 238.1, pp. 52–62. issn: 0036-8733.

Rinn, John L and Howard Y Chang (2012). "Genome regulation by long noncoding RNAs". In: *Annual review of biochemistry* 81, pp. 145–166. issn: 0066-4154, 1545-4509. doi: 10.1146/annurev-biochem-051410-092902.

Robertus, J D, J E Ladner, J T Finch, D Rhodes, R S Brown, B F Clark, and Aaron Klug (1974). "Structure of yeast phenylalanine trna at 3 A resolution". In: *Nature* 250.467, pp. 546–551. issn: 0028-0836.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1, pp. 139–140. issn: 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btp616.

Robinson, Mark D and Alicia Oshlack (2010). "A scaling normalization method for differential expression analysis of RNA-seq data". In: *Genome Biology* 11.R25, R25. issn: 1465-6906, 1465-6914. doi: 10.1186/gb-2010-11-3-r25.

Robinson, Mark D and Gordon K Smyth (2007). "Moderated statistical tests for assessing differences in tag abundance". In: *Bioinformatics* 23.21, pp. 2881–2887. issn: 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btm453.

Schena, M, D Shalon, R W Davis, and P O Brown (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray". In: *Science* 270.5235, pp. 467–470. issn: 0036-8075.

Schimmel, Paul R, R Giegé, D Moras, and S Yokoyama (1993). "An operational RNA code for amino acids and possible relationship to genetic code". In: *Proceedings of the National Academy of Sciences of the United States of America* 90.19, pp. 8763–8768. issn: 0027-8424.

Schimmel, Paul R and Dieter Söll (1979). "Aminoacyl-trna synthetases: general features and recognition of transfer RNAs". In: *Annual review of biochemistry* 48, pp. 601–648. issn: 0066-4154. doi: 10.1146/annurev.bi.48.070179.003125.

Schmitt, Bianca M, Konrad L M Rudolph, Panagiota Karagianni, Nuno A Fonseca, Robert J White, Iannis Talianidis, Duncan T Odom, John C Marioni, and Claudia Kutter (2014). "High-resolution mapping of transcriptional dynamics across tissue development reveals a stable mRNA–tRNA interface". In: *Genome Res* 24.11, pp. 1797–807. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.176784.114.

Sharp, P M and G Matassi (1994). "Codon usage and genome evolution". In: *Current opinion in genetics & development* 4.6, pp. 851–860. ISSN: 0959-437X.

Shen, Yin, Feng Yue, David F McCleary, Zhen Ye, Lee Edsall, Samantha Kuan, Ulrich Wagner, Jesse Dixon, Leonard Lee, Victor V Lobanenkov, and Bing Ren (2012). "A map of the cis-regulatory sequences in the mouse genome". In: *Nature* 488.7409, pp. 116–120. ISSN: 0028-0836. DOI: 10.1038/nature11243.

Sittman, Donald B (1999). "RNA Synthesis and Processing". In: *Biochemistry*. Ed. by Victor L Davidson and Donald B Sittman. The National medical series for independent study. Lippincott, William & Wilkins, pp. 145–164. ISBN: 9780683305036.

Smit, A F A and R Hubley (2014). *RepeatMasker*.

Stergachis, Andrew B, Eric Haugen, Anthony Shafer, Wenqing Fu, Benjamin Vernot, Alex Reynolds, Anthony Raubitschek, Steven Ziegler, Emily M LeProust, Joshua M Akey, and John A Stamatoyannopoulos (2013). "Exonic transcription factor binding directs codon choice and affects protein evolution". In: *Science* 342.6164, pp. 1367–1372. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1243490.

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.43, pp. 15545–15550. ISSN: 0027-8424. DOI: 10.1073/pnas.0506580102.

Suddath, F L, G J Quigley, A McPherson, D Sneden, J J Kim, S H Kim, and A Rich (1974). "Three-dimensional structure of yeast phenylalanine transfer

RNA at 3.0angstroms resolution". In: *Nature* 248.5443, pp. 20–24. issn: 0028-0836.

Si-Tayeb, Karim, Frédéric P Lemaigre, and Stephen A Duncan (2010). "Organogenesis and development of the liver". In: *Developmental cell* 18.2, pp. 175–189. issn: 1534-5807, 1878-1551. doi: 10.1016/j.devcel.2010.01.011.

Thompson, Debrah M, Cheng Lu, Pamela J Green, and Roy Parker (2008). "tRNA cleavage is a conserved response to oxidative stress in eukaryotes". In: *RNA* 14.10, pp. 2095–2103. issn: 1355-8382, 1469-9001. doi: 10.1261/rna.1232808.

Tsui, David, John P Vessey, Hideaki Tomita, David R Kaplan, and Freda D Miller (2013). "FoxP2 regulates neurogenesis during embryonic cortical development". In: *The Journal of neuroscience: the official journal of the Society for Neuroscience* 33.1, pp. 244–258. issn: 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.1665-12.2013.

Vassetzky, Nikita S and Dmitri A Kramerov (2013). "SINEBase: a database and tool for SINE analysis". In: *Nucleic acids research* 41.Database issue, pp. D83–9. issn: 0305-1048, 1362-4962. doi: 10.1093/nar/gks1263.

Villar, Diego, Camille Berthelot, Sarah Aldridge, Tim F Rayner, Margus Lukk, Miguel Pignatelli, Thomas J Park, Robert Deaville, Jonathan T Erichsen, Anna J Jasinska, James M A Turner, Mads F Bertelsen, Elizabeth P Murchison, Paul Flicek, and Duncan T Odom (2015). "Enhancer Evolution across 20 Mammalian Species". In: *Cell* 160.3, pp. 554–566. issn: 0092-8674, 1097-4172. doi: 10.1016/j.cell.2015.01.006.

Wain, Hester M, Elspeth A Bruford, Ruth C Lovering, Michael J Lush, Mathew W Wright, and Sue Povey (2002). "Guidelines for human gene nomenclature". In: *Genomics* 79.4, pp. 464–470. issn: 0888-7543. doi: 10.1006/geno.2002.6748.

Watson, James D and Francis H C Crick (1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid". In: *Nature* 171.4356, pp. 737–738. issn: 0028-0836.

White, R J and S P Jackson (1992). "Mechanism of TATA-binding protein recruitment to a TATA-less class III promoter". In: *Cell* 71.6, pp. 1041–1053. issn: 0092-8674.

White, Robert J (1998). *RNA Polymerase III Transcription*. 2nd ed. Berlin: Springer-Verlag.

Wilusz, Jeremy E (2015). "Controlling translation via modulation of tRNA levels". In: *Wiley interdisciplinary reviews. RNA*. ISSN: 1757-7004, 1757-7012. DOI: 10.1002/wrna.1287.

Winter, A G, G Sourvinos, S J Allison, K Tosh, P H Scott, D A Spandidos, and R J White (2000). "RNA polymerase III transcription factor TFIIIC2 is overexpressed in ovarian tumors". In: *Proc Natl Acad Sci U S A* 97.23, pp. 12619–24. ISSN: 0027-8424. DOI: 10.1073/pnas.230224097.

Yang, Ziheng and Rasmus Nielsen (2008). "Mutation-Selection Models of Codon Substitution and Their Use to Estimate Selective Strengths on Codon Usage". In: *Molecular biology and evolution* 25.3, pp. 568–579. ISSN: 0737-4038.

Yassour, Moran, Jenna Pfiffner, Joshua Z Levin, Xian Adiconis, Andreas Gnirke, Chad Nusbaum, Dawn-Anne Thompson, Nir Friedman, and Aviv Regev (2010). "Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species". In: *Genome biology* 11.8, R87. ISSN: 1465-6906, 1465-6914. DOI: 10.1186/gb-2010-11-8-r87.

Zhang, Yong, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu (2008). "Model-based analysis of ChIP-Seq (MACS)". In: *Genome biology* 9.9, R137. ISSN: 1465-6906, 1465-6914. DOI: 10.1186/gb-2008-9-9-r137.

Zorn, Aaron M (2008). "Liver development". In: *StemBook*. Ed. by Alexander F Schier. Cambridge (MA): Harvard Stem Cell Institute.