University of Zagreb
Faculty of Science
Department of Biology

Anamaria Elek

# coRdon: an R package for codon usage analysis and prediction of gene expressivity

Graduation thesis

Zagreb, 2018

This thesis is created in the bioinformatics group at the Division of Molecular Biology, under the supervision of Professor Kristian Vlahoviček. The thesis is submitted for grading to the Department of Biology at the Faculty of Science, University of Zagreb, with the aim of obtaining the Masters degree in molecular biology.

I wish to express my gratitude to people who helped me with the work presented in this thesis: my supervisor, Professor Kristian Vlahoviček, and all members of the bioinformatics group at the Division of Molecular Biology, especially Maja, who did much of the foundational work on the project and was directly involved along the way. Thanks are also due to my family, for their continuous and unconditional support, and to my friends, for making me abandon this work regularly in order to enjoy their instructive company. Last but not least, thanks to Hajduk for the five years of mostly cold and frustrating away football in Zagreb, I wouldn't have missed it for the world.

# BASIC DOCUMENTATION CARD

University of Zagreb
Faculty of Science
Division of Biology

Graduation thesis

# coRdon: an R package for codon usage analysis and prediction of gene expressivity

Anamaria Elek
Deprtment of Molecular Biology, Horvatovac 102A, 10000 Zagreb, Croatia

Not all synonymous codons are used with equal frequency in prokaryotic genomes - this selective preference is termed codon usage (CU) bias and is an independent determinant of gene expression regulation at the level of translation. The synonymous codons corresponding to the most abundant cognate tRNA are considered optimal for translation. Codon usage bias can therefore be used to predict the relative expression levels of genes, by comparing CU bias of a specific gene to the CU bias of a set of genes known to be highly expressed. This approach can be efficiently used to predict highly expressed genes within a single genome but has also been demonstrated relevant at the level of entire microbial communities – metagenomes. This is possible because CU bias is shared among the microbial species in the same environment. By analysing CU bias of a metagenome, one can identify genes with high predicted expression across the given ecological niche, and identify enriched functions, i.e. the functional fingerprint of the analysed microbial community. Software tools to efficiently analyse and visualize metagenomics data in the context of translational optimisation and CU bias are scarce and this limits the ability to generate useful knowledge. As the part of this thesis, I have developed a Bioconductor-based R package for comprehensive analysis of CU and CU-based gene expressivity. The package also includes several methods for visualization of CU, and the CU-based functional analysis of annotated DNA sequences. The package is fast and flexible enough to handle even the largest metagenomics datasets.

(49 pages, 10 figures, 3 tables, 43 references, original in English)

Thesis deposited in the Central Biological Library

Key words: translational optimization, functional analysis, metagenomics

Supervisor: Professor Kristian Vlahoviček, PhD

Reviewers: Professor Kristian Vlahoviček, PhD
Assoc. Prof Damjan Franjević, PhD
Assoc. Prof Sven Jelaska, PhD

Substitution: Asst. Prof. Rosa Karlić, PhD

Thesis accepted: September 6, 2018.

# TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu
Prirodoslovno matematički fakultet
Biološki odsjek Diplomski rad

# Razvoj paketa coRdon u programskom jeziku R za analizu korištenja sinonimnih kodona i predviđanje ekspresije gena

Anamaria Elek
Zavod za molekularnu biologiju, Horvatovac 102A, 10000 Zagreb, Hrvatska

U genomima različitih prokariota svi sinonimni kodoni ne koriste se jednakom učestalošću - njihovo selektivno korištenje zaseban je mehanizam regulacije ekspresije gena na razini translacije. Sinonimni kodoni koje prepoznaju najbrojnije tRNA molekule u stanici smatraju se optimiziranima za translaciju. Stoga se selektivno korištenje sinonimnih kodona može koristiti za predviđanje relativne ekspresivnosti gena, uspoređujući obrazac korištenja kodona u određenom genu s kodonima koji se učestalo pojavljuju u skupu visoko eksprimiranih gena. Na ovaj način moguće je identificirati potencijalno jako eksprimirane gene na razini genoma, ali još važnije, i na razini čitavih metagenoma, jer je pokazano da je uzorak selektivnog korištenja kodona zajednički organizmima unutar iste mikrobne zajednice, bez obzira na njihovu filogenetsku udaljenost. Stoga je analizirajući korištenje kodona na razini metagenoma moguće predvidjeti koji geni su relativno visoko eksprimirani u određenoj mikrobnoj zajednici, kao i koje funkcije su zastupljene među tim genima - tzv. funkcijski otisak mikrobne zajednice. U sklopu ovog rada razvila sam paket u programskom jeziku R, baziran na sintaksi i smjernicama projekta Bioconductor, za kompletnu analizu korištenja sinonimnih kodona i ekspresivnosti gena. Uz to što omogućava računanje više različitih statistika koje kvantificiraju korištenje sinonimnih kodona, paket uključuje i nekoliko metoda za vizualizaciju korištenja sinonimnih kodona, omogućava funkcionalnu analizu anotiranih slijedova DNA te je dovoljno brz i fleksibilan za analizu velikih količina metagenomskih podataka.

(49 stranica, 10 slika, 3 tablice, 43 literaturna navoda, jezik izvornika: engleski)

Rad je pohranjen u Središnjoj biološkoj knjižnici.

Ključne riječi: translacijska optimizacija, funkcijska analiza, metagenomika

Voditelj: Prof. dr. sc. Kristian Vlahoviček

Ocjenitelji: Prof. dr. sc. Kristian Vlahoviček
Izv. prof. dr. sc. Damjan Franjević
Izv. prof. dr. sc. Sven Jelaska

Zamjena: Doc. dr. sc. Rosa Karlić

Rad prihvaćen: 6. rujna 2018.

# Abbreviations

| | |
|---|---|
| B | Measure of codon bias |
| CAI | Codon Adaptation Index |
| E | Expressivity of genes |
| ENC,<br>ENC’ | Effective Number of Codon,<br>prime denotes a modified version of the same statistic |
| $F_{op}$ | Frequency of optimal codons |
| GCB | Gene Codon Usage Bias |
| MCB | Maximum Likelihood Codon Bias |
| MELP | MILC-based Expression Level Predictor |
| MILC | Measure Independent of Length and Composition |
| NGS | Next Generation Sequencing |
| OO | Object-oriented, a style of programming |
| SCUO | Synonymous Codon Usage Orderliness |

# Contents


**1. INTRODUCTION ..... 1**

**1.1 Encoding genetic information ..... 2**

**1.2 Metagenomics ..... 6**

**1.2.1 Implications for human disease ..... 7**

**1.2.2 Codon usage in metagenomic analyses ..... 7**

**1.3 Quantifying codon usage ..... 9**

**1.3.1 Codon usage statistics ..... 9**

**1.3.2 Expressivity measures ..... 15**

**2. MATERIALS AND METHODS ..... 19**

**2.1 R programming language ..... 20**

**2.2 Bioconductor and dependencies ..... 21**

**2.3 Biological data ..... 22**

**2.4 Codon usage analysis ..... 22**

**2.4.1 Loading DNA sequences ..... 22**

**2.4.2 Calculating CU statistics ..... 24**

**2.4.3 Predicting genes' expressivity ..... 24**

**2.4.4 Functional annotation ..... 24**

**3. RESULTS ..... 26**

**3.1 Implementation ..... 27**

**3.1.1 Package ..... 27**

**3.1.2 Classes and methods ..... 27**

**3.1.3 Statistics ..... 29**

**3.2 Codon usage analysis ..... 30**

**3.2.1 Codon usage bias..... 30**

**3.2.2 Relative genes' expressivity ..... 32**

**3.2.3 Functional enrichment ..... 34**

**4. DISCUSSION ..... 37**

**4.1 Software..... 38**

**4.2 Analysis ..... 39**

**5. CONCLUSION ..... 41**

**6. REFERENCES..... 43**

# 1. INTRODUCTION

## 1.1 Encoding genetic information

Deoxyribonucleic acid (DNA) contains the coded instructions for formation and functioning of all the living organisms, while proteins are the main effector molecules responsible for implementing those instructions. The transition from blocks of instructions, termed genes, contained in the polymeric DNA, to extremely diverse group of molecules such as proteins, is obviously a very complex process, requiring several steps and intermediators, and is naturally regulated at more than one level. This process is often summarised by the so-called central dogma of molecular biology, a notion first proposed by Francis Crick in 1958, that describes a flow of information in biological systems (Crick, 1958). According to this dogma, DNA sequences are always transcribed to ribonucleic acid (RNA) molecules, which are then translated to proteins. There are known exceptions from this rule – e.g. some viruses store their genetic information as RNA – which is one of the reasons why many rightly argue against labelling scientific concepts, no matter how well established, as "dogmas". However, the majority of genetic information is still stored in the DNA.

Each DNA molecule is essentially a sequence of building blocks, termed nucleotides. Each nucleotide, in turn, is composed of a sugar deoxyribose, with a nitrogenous base attached to it, connected to the adjacent nucleotides by a phosphate. There are four types of nitrogenous bases occurring in the DNA – adenine (A), guanine (G), thymine (T), and cytosine (C) – giving rise to four types of nucleotides. These four letters are the base of a genetic code. On the other hand, proteins are built using 20 different amino acids (again, there are exceptions – some archaea and bacteria use additional amino acids). A simple calculation can show that, having only four nucleotides at hand, in order to unambiguously represent the 20 amino acids, each must be encoded by at least three nucleotides ($3^1 < 20$; $3^2 < 20$; $4^3 > 20$). This was proven to be the case, and the coding triplets were termed codons. However, the number of distinct triplets ($4^3 = 64$) exceeds the number of amino acids (20). This redundancy, or codon degeneracy, is one of the important properties of the genetic code, because it provides an additional a way to regulate gene expression at the level of the translation.

Another important property of the genetic code is its universality – all organisms use the same rules for translating their DNA into proteins. Once again, there are exceptions to this rule, and there are so much as 20 variants of genetic code defined, e.g. for mitochondrial or plastid genes, and for some bacteria and archaea. Standard genetic code is summarised in the Table 1.

In the process of gene expression, coding DNAs are transcribed in messenger RNA (mRNA) molecules, and each codon on the mRNA is recognized by the specific, cognate, transfer RNA (tRNA) containing a complementary anticodon triplet. Each tRNA has the corresponding amino acid attached to it, so that by the way of sequential binding of tRNAs to mRNA, the genetic information is read, and the amino acids are concatenated in the protein sequence.

**Table 1.** Standard genetic code. Codons are grouped according to the first, second and third base (letter) in codon. Amino acids are indicated by both their three-letter and one-letter abbreviations, as well as full name. Start codon (ATG) is indicated in green, stop codons (TAA, TAG, TGA) are in red.

| 1st base | 2nd base | | | | | | | | 3rd base |
|---|---|---|---|---|---|---|---|---|---|
| | T | | C | | A | | G | | |
| T | TTT | (Phe/F) Phenyl-alanine | TCT | (Ser/S) Serine | TAT | (Tyr/Y) Tyrosine | TGT | (Cys/C) Cysteine | T |
| | TTC | | TCC | | TAC | | TGC | | C |
| | TTA | (Leu/L) Leucine | TCA | | **TAA** | Stop | **TGA** | Stop | A |
| | TTG | | TCG | | **TAG** | Stop | TGG | (Trp/W) Tryptophan | G |
| C | CTT | | CCT | (Pro/P) Proline | CAT | (His/H) Histidine | CGT | (Arg/R) Arginine | T |
| | CTC | | CCC | | CAC | | CGC | | C |
| | CTA | | CCA | | CAA | (Gln/Q) Glutamine | CGA | | A |
| | CTG | | CCG | | CAG | | CGG | | G |
| A | ATT | (Ile/I) Isoleucine | ACT | (Thr/T) Threonine | AAT | (Asn/N) Asparagine | AGT | (Ser/S) Serine | T |
| | ATC | | ACC | | AAC | | AGC | | C |
| | ATA | | ACA | | AAA | (Lys/K) Lysine | AGA | (Arg/R) Arginine | A |
| | **ATG** | (Met/M) Methionine | ACG | | AAG | | AGG | | G |
| G | GTT | (Val/V) Valine | GCT | (Ala/A) Alanine | GAT | (Asp/D) Aspartate | GGT | (Gly/G) Glycine | T |
| | GTC | | GCC | | GAC | | GGC | | C |
| | GTA | | GCA | | GAA | (Glu/E) Glutamate | GGA | | A |
| | GTG | | GCG | | GAG | | GGG | | G |

Since all amino acids except methionine and tryptophan are encoded by more than a single codon, the emerging question is whether there is any difference in the frequency with which different codons are used. It was first shown back in the 1980s that there are indeed different

patterns of codon usage, both between different genomes, as well as between genes in a single genome (Grantham *et al.*, 1981). This non-random distribution of synonymous codons was termed codon usage (CU) bias and was found to correlate with gene expressivity. Those genes which are more efficiently translated produce proteins with increased availability in cell, such as ribosomal, and have a stronger CU bias than the genes at lower expression levels, reflecting the selection pressure on the former. Authors of the early papers on the subject suggested that the choice of nucleotide in the third codon position is guided by the energy stabilization for codon-anticodon pairing, in such a way that the pairing is of intermediate strength, allowing for efficient binding and release of tRNA during translation. Beside kinetic aspect, CU bias was also found to be correlated with the abundance of the tRNA molecules in the cell (Gouy and Gautier, 1982), and those synonymous codons corresponding to the most abundant tRNA species are considered optimal for translation, because they are translated either faster or more accurately (i.e. they are less prone to cause processivity errors in translation).

Apart from different CU between genomes and genes, codons are used unequally within a single gene, allowing for the identification of nucleotide usage gradients (Hooper and Berg, 2000). Moreover, there are important codon bias variants, other than selection for single synonymous codons, namely nonsynonymous codon pairs bias and synonymous codon co-occurrence (Quax *et al.*, 2015). The former refers to the fact that some codon pairs are universally avoided or preferred, while the latter is a bias for the codons recognized by the same tRNA (including both frequent and rare ones) to appear clustered in the mRNA sequence.

For prokaryotes, translational optimization driven by codon usage bias is now recognized as an important mechanism of gene expression regulation (Supek *et al.*, 2010). While the same can be said for unicellular eukaryotes such as yeast, the story is not so straightforward in multicellular eukaryotes. In their complex genomes, the differences in synonymous codons usage can often be attributed to background nucleotide content. In mammals, for example, nucleotide concentrations vary greatly across the genome, so that several regions of fairly homogeneous GC content (i.e. percentage of guanine-cytosine pairs in the DNA) can be identified – these are termed isochores (Cozzi, Milanesi and Bernardi, 2015). However, nucleotide content alone cannot explain the degree of codon bias observed in some genes (Urrutia and Hurst, 2001), indicating that there also must be a selection acting at the level of codon usage.

Codon usage bias is therefore believed to be implicated in the regulation of gene expression in organisms at all taxonomical levels. What's more, it has been shown that CU bias is shared by

different, often phylogenetically distant, organisms within microbial communities, while being different between distinct communities (Roller *et al.*, 2013). This opens up a range of possible applications of codon usage analyses in the environmental microbial studies.

## 1.2 Metagenomics

Comprehensive analysis of environmental microbial communities has for a long time been halted by the inability to grow most of the species in the laboratory. The disproportion between microbial diversity observed under the microscope in the environmental samples and the number of species successfully cultivated on the plates was termed "the great plate-count anomaly" (Staley and Konopka, 1985). This seemingly insurmountable obstacle was finally overcome by altogether bypassing the cultivation step in the analysis. With the development of high-throughput sequencing methods it became possible to analyse vast quantities of genetic material, and to do so using limited quantities of the starting material. From there, it did not take long for the researchers to come up with an idea to directly analyse the entire genomic content of an environmental sample, without any prior culturing. This approach proved sensible, and is now termed metagenomics, because the obtained sequences are analysed in more-or-less the same way as single genomes.

In one of the first metagenome sequencing projects, DNA samples from two marine viral communities were analysed (Breitbart *et al.*, 2002), and it was found that over 65% of sequences had no significant similarity to any of the previously reported sequences, confirming the notion that most of microbial diversity goes undetected by cultivation-based analyses.

Another project aimed to analyse diversity of microbial communities from seawater samples by whole-genome shotgun sequencing (Venter *et al.*, 2004) and identified more than 1.2 million previously unknown genes, including 148 previously unknown bacterial phylotypes.

Apart from various environmental samples, metagenomes from microbial communities associated with different human body habitats were extensively analysed and characterized, primarily as a part of The National Institutes of Health Human Microbiome Project (Human Microbiome Project Consortium, 2012; Lloyd-Price *et al.*, 2017). Results of the analysis of more than two thousand metagenomic samples significantly advanced our understanding of personalized microbiome function, as well as its spatial and temporal dynamics. The second, extended phase of this huge undertaking is still ongoing.

### 1.2.1 Implications for human disease

Given the abundance of microbes inhabiting the human body, it is only reasonable to expect that they influence their host's physiology to some extent. The majority of microbes reside in the gut, their number exceeding bacterial count in all other organs by at least two orders of magnitude. The number of bacterial cell in the gut is estimated at 10 trillion ($10^{13}$), which is on the same order of magnitude as the total number of cells in human body (Sender, Fuchs and Milo, 2016). Association of the microbial genes with human phenotypes has led to the gradual improvement in our understanding of the impact of gut microbes on human health. Metagenome is now known to play a role in the onset of metabolic disorders such as obesity, metabolic syndrome and diabetes, as well as in the inflammation of the intestine, symptomatic atherosclerosis and liver cirrhosis (Devaraj, Hemarajata and Versalovic, 2013).

### 1.2.2 Codon usage in metagenomic analyses

Generally, metagenomes can be studied in one of two different ways, depending on the aim of analysis. If the goal is to estimate the phyletic distribution of microbial species in the sample, similarity searches against known microbial species' sequences are usually performed. On the other hand, if the goal is to functionally annotate the sample, identified genes (open reading frames, ORFs) are first annotated through orthology databases, such as KEGG-KO (Kyoto Encyclopaedia of Genes and Genomes—Orthology) (Kanehisa *et al.*, 2017) or COG/KOG (Clusters of Orthologous Groups of genes) (Tatusov *et al.*, 2003) and the related eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) (Huerta-Cepas *et al.*, 2016). The relative 'importance' of functions inferred from annotations is then determined according to their abundance in the sample. This approach is not without drawbacks, because only the sequences homologous to some known sequence(s) will be taken into account, while the remainder will go undetected, although it might carry an important functional information. Another problem is that just the abundance of genes encoding different functions is considered, while differential regulation is not taken into account. This is where codon usage bias and translational optimization come in handy. The relative expression level of genes can be estimated by comparing CU bias of any gene to the CU bias of a set of those genes known to be highly expressed. While this approach can be efficiently used to predict highly expressed genes in a single genome, it is especially useful for metagenome analysis, allowing for

identification of the genes with high predicted expression across the entire microbial community, and determination of the enriched functions within the community.

## 1.3 Quantifying codon usage

### 1.3.1 Codon usage statistics

There are many statistics devised to quantify codon usage, and most can be classified into one of the two major groups. One group contains measures that essentially calculate the distance of every codon's usage to the uniform usage of synonymous codons. Two of the commonly used statistics from this group are ENC, effective number of codons (Wright, 1990), and SCUO, synonymous codon usage orderliness (Wan *et al.*, 2004). Statistics in the other group calculate the distance between the total codon usage in a given gene (or an ORF, or simply a sequence) and the codon usage in the subset of highly expressed genes. These include a modified version of the effective number of codons, ENC' (Novembre, 2002), a measure of codon bias termed B (Karlin and Mrazek, 1996), maximum likelihood codon bias, MCB (Urrutia and Hurst, 2001), and Measure Independent of Length and Composition, MILC (Supek and Vlahoviček, 2005). The statistics in both classes have their respective advantages and disadvantages, and are generally applicable in different situations. While the former statistics are at the disadvantage because they only allow for comparing CU to a uniform null distribution, the latter require a prior knowledge of the preferred codons, i.e. a subset of highly expressed genes for which CU is initially calculated.

**Effective number of codons**. One of the first devised and most commonly used measures, the effective number of codons used in a gene (ENC, also abbreviated $\widehat{N}_c$) quantifies how different the actual codon usage is from the scenario in which all codons are used equally often (Wright, 1990). This measure is based on the analogy between the usage of synonymous codons for a particular amino acid and the frequencies of alleles (alternative gene variants) at a genomic locus.

For an individual codon family of $k$ synonymous codons whose counts are $n_1, n_2, \ldots, n_k$, such that $n = \sum_i n_i$, and $p_i = n_i/n$, an estimate of the homozygosity of codon usage is defined as follows:

$$F_a = \frac{n \sum_{i=1}^{k} p_i^2 - 1}{n - 1}. \quad (1)$$

Equal codon usage would be equivalent to minimum homozygosity, whereas usage of only one codon would result in larger values of $F_a$. The effective number of codons used in a gene is then calculated like so:

$$ENC = K_1 + \frac{K_2}{\bar{F}_2} + \frac{K_3}{\bar{F}_3} + \frac{K_4}{\bar{F}_4} + \frac{K_6}{\bar{F}_6}, \quad (2)$$

where $K_i$ is the number of $i$-fold codon families, i.e. the number of amino acids in an $i$-fold redundancy class, and $\bar{F}_i$ is average of $F_a$ for each $i$-fold redundancy class: $\mathrm{F_i} = \sum_\mathrm{j} F_{a.j}/K_{i,obs}$ where $K_{i,obs}$ is the number of amino acids actually present in the $i$-fold redundancy class. Note that homozygosity is not calculated for amino acids encoded by a single codon. Also, rarely used amino acids – those for which either numerator or denominator in the equation (1) is 0 – should be excluded from calculation, i.e. $F_a$ should be averaged over only those amino acids actually present. However, if no three-fold redundant codons are observed, one should average $F_2$ and $F_4$ to obtain $F_3$. If other $i$-fold redundancy classes are unobserved, $F_i$ is assumed to equal $1/i$.

If all codons were used approximately uniformly, the ENC value wold be around 61, and if the codon usage was extremely biased, so that only one codon was used for each amino acid, ENC value would be around 20. ENC is dependent on the length of the sequence for which it is calculated, so that with increase in length, the values would approach 20 and 61, respectively. Of note, the value of ENC can be greater than 61 if the observed codon usage pattern is more uniform than expected by chance (e.g. when amino acid composition is very extreme, and the gene is very short) – in these cases, the value of ENC should be revised to 61.

When using ENC to quantify codon usage, one should bear in mind that this statistic does not take into account background nucleotide composition. This is especially important when comparing codon usage in sequences with different background nucleotide composition (e.g. from different organisms) because different background nucleotide composition can in these cases be misinterpreted as the difference in codon usage.

**Effective number of codons prime**. To allow for comparison to codon usage distribution other than uniform, a modified version of ENC, termed ENC' (Novembre, 2002), uses Pearson's $\chi^2$ statistics to measure the departure of observed from expected usage of codons for each codon family, i.e. amino acid $a$:

$$\chi_a^2 = \sum_{i=1}^{k} \frac{n\,(p_i - e_i)^2}{e_i}. \quad (3)$$

Here $p_i$ is observed usage and $e_i$ is expected usage of a codon $i$, and $n$ is the observed number of codons for that amino acid (sum of codon counts).

Homozygosity $F_a'$ for every k-fold codon family is then defined as follows:

$$F_a' = \frac{\chi_a^2 + n - k}{k\,(n-1)}, \quad (4)$$

and the rest of calculations mirror those for ENC.

It should be noted that, in the calculations of ENC' values, the author excludes amino acids that are observed fewer than five times, and this is implemented in coRdon as well.

Effectively, when the expected distribution is uniform, the expression used to calculate homozygosity given in equation (4) reduces to the homozygosity in ENC calculations, given in the equation (1). As the expected distribution departs from uniform, calculated ENC' values are decreasing, as compared to those obtained by calculating ENC.

**Codon bias**. Two equivalent statistics were devised for calculation of codon bias directed towards a set of optimal codons, one for an individual gene (Karlin and Mrazek, 1996), and the other for a set of genes (Karlin, Campbell and Mrazek, 1998).

Both start by calculating bias for a single amino acid *a* of degeneracy k:

$$B_a = \frac{1}{k} \sum_c |f_c / g_c - 1| \quad (5)$$

Here $f_c$ and $g_c$ are codon frequencies in a gene or a gene set of interest, and in the reference set of highly expressed genes, respectively, both normalised so that each sums up to 1.

If $g_c = 1/k$, the equation (5) reduces to

$$B_a = \sum_c |f_c - g_c| \quad (6)$$

Overall bias for a gene is calculated as:

$$B_g = \sum_a p_a B_a = \sum_a p_a \left( \sum_c |f_c - g_c| \right), \quad (7)$$

where $p_a$ denotes amino acid frequencies of the gene.

Equivalently, codon bias can be calculated for a gene family (F) with respect to another gene family (G):

$$B(F|G) = \sum_a p_a \left( \sum_c |f_c - g_c| \right), \quad (8)$$

and $p_a$ is now the set of amino acid frequencies of the genes of F.

The maximum possible value for B is 2, but for all practical purposes, it rarely exceeds 0.5.

Codon usage differences between two gene families generally range from 0.05 to 0.3.

**Maximum likelihood codon bias**. This measure attempts to minimise the effect of rarely occurring amino acids on the value of CU statistic (Urrutia and Hurst, 2001). Because they are more likely to be far from expected just by chance, their contribution to the value of MCB is down-scaled.

Bias for an individual amino acid *a* is:

$$B_a = \sum_a \frac{(f_c - g_c)^2}{g_c}, \quad (9)$$

where $f_c$ and $g_c$ are observed and expected frequencies of a codon *c*, respectively. Only codons that occur at least once in the sequence and also have the expected frequency $g_c \geq 0$ are used to calculate $B_a$.

Biases of individual amino acids are summed in order to obtain codon bias index for a gene:

$$B_g = \sum_a \frac{B_a \, log \, N_a}{A}, \quad (10)$$

where $N_a$ is the count of amino acid a, and *A* is the number of amino acids contributing to the index. Only amino acids with degeneracy greater than 1 are used to calculate $B_g$.

**Synonymous codon usage orderliness**. Based on the Shannon information theory, this measure is an estimate for the orderliness of synonymous codon usage and the amount of synonymous codon usage bias.

Briefly, the key measure in information theory is entropy, and it quantifies the amount of uncertainty involved in the value of a random variable. Information is then defined as an index of orderliness, or the difference between the maximum entropy and the actual entropy. The more order there is, the greater the information. In the case of codon usage, information measures the orderliness, i.e. non-randomness or bias, in synonymous codon usage, for any given amino acid.

Entropy for amino acid $a$ is defined as a sum over synonymous codons for that amino acid:

$$H_a = -\sum_i p_i \log p_i, \tag{11}$$

and $p_i$ is normalised frequency of codon $i$.

If synonymous codons were used at random, entropy would be maximum $H_a^{max} = -\log\frac{1}{k} = \log k$, where $k$ is the number of synonymous codons for amino acid a, i.e. its degeneracy. If only one of the synonymous codons is used, then $H_a^{min} = 0$.

Synonymous codon usage order for amino acid $a$ is therefore defined as:

$$O_a = \frac{H_a^{max} - H_a}{H_a^{max}} \tag{12}$$

and it follows that $0 \leq O_a \leq 1$, with $O_a = 0$ for random (uniform) CU, and $O_a = 1$ if CU is extremely biased.

The composition ratio of the amino acid $a$ in each sequence is:

$$F_a = \frac{\sum_c O_c}{\sum_a \sum_c O_c} \tag{13}$$

where numerator is summed counts of codons for amino acid a, and denominator is summed counts of codons for all amino acids with redundancy greater than 1.

Average SCUO for a gene is then the sum of amino acid orders, weighted by respective amino acid occurrences:

$$O = \sum_{a} F_a O_a \tag{14}$$

**Measure Independent of Length and Composition**. This measure aims to give an estimate of CU bias which is stable (i) over wide range of sequence lengths, starting from so little as 80 codons, and (ii) over a range of sequence compositions, from very biased to the sequences using codons uniformly (Supek and Vlahoviček, 2005).

The individual contribution of each amino acid *a* to the MILC statistic is calculated based on a log-likelihood ratio score:

$$M_a = 2 \sum_{c} O_c \, ln \frac{f_c}{g_c}, \tag{15}$$

where $f_c$ is the observed frequency of the codon *c* in a gene, and $g_c$ is the expected frequency of the same codon. MILC is then calculated as the total difference in codon usage for all amino acids, normalised by L, the length of sequence in codons, and a factor C, correction for the CU bias overestimation in shorter sequences:

$$MILC = \frac{\sum_{a} M_a}{L} - C \tag{16}$$

Correction factor is defined as follows:

$$C = \frac{\sum_{a} (r_a - 1)}{L} - 0.5, \tag{17}$$

where $r_a$ is degeneracy class for amino acid *a*. Only the amino acids that are actually present at least once in the sequence contribute to C.

### 1.3.2 Expressivity measures

Several statistics have been formulated that predict relative expression levels of genes, based on the different CU patterns. Except for GCB, all of these require a known set of highly expressed genes to be used as a reference for comparing CU.

**Codon Adaptation Index**. This measure is one of the earliest devised and one of most commonly used measures that correlates codon usage and gene expression (Sharp and Li, 1987). It is calculated based on the reference table of relative synonymous codon usage (RSCU). RSCU value for every codon is the observed codon frequency divided by the frequency expected under the assumption of equal usage of the synonymous codons:

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}} \tag{18}$$

Here $n_i$ is number of alternative codons, and $X_{ij}$ is the number of occurrences of the *j*-th codon for the *i*-th amino acid. RSCU values greater than 1 indicate that the corresponding codon is used more frequently than expected, whereas the reverse is true for RSCU values less than 1.

Relative adaptiveness of a codon, $w_{ij}$, is the frequency of that codon compared to frequency of the optimal codon for the same amino acid. It can be expressed as the ratio of either RSCU values or codon counts:

$$w_i = \frac{RSCU_{ij}}{RSCU_{imax}} = \frac{X_{ij}}{X_{imax}} \tag{19}$$

Codon adaptation index is then calculated as the geometric mean of RSCU values from the reference table corresponding to each of the codons used in that gene, divided by the maximum possible CAI for a gene with the same amino acid composition. This is equivalent to the product of ratios of the observed codon frequency and frequency of the most commonly used synonymous codon for that amino acid, with this product extending over the length of sequence in codons:

$$CAI = \frac{CAI_{obs}}{CAI_{max}} = \left(\prod_{k=1}^{L} \frac{RSCU_{obs}}{RSCU_{max}}\right)^{1/L}, \tag{20}$$

Equivalently, in terms of relative adaptiveness of codons, CAI is defined as follows:

$$CAI = \exp\left(\frac{1}{L}\sum_i \sum_j X_{ij} \ln w_{ij}\right). \quad (21)$$

If every codon corresponded to most frequently used codon in highly expressed genes, then CAI would be 1. CAI values greater than 1 correspond to codons which are used more frequently than the most frequently used codons in the set of highly expressed genes, whereas values less than 1 correspond to the less frequently used codons.

**Frequency of optimal codons**. Based on the experimentally measured relative abundance of tRNAs in E. coli and the nature of codon-anticodon interaction, it is possible to predict the order of preference among synonymous codons encoding each amino acid (Ikemura, 1981). The synonymous codon found to be the most preferred for each amino acid is designated as the optimal codon. It was shown that E. coli genes encoding abundant protein species selectively use the optimal codons, whereas other genes use optimal and "non-optimal" codons equally. The optimal codons are therefore considered to be specifically optimized for the E. coli translational system.

Frequency of optimal codons, $F_{op}$, is calculated as a ratio between the count of optimal codons and the total number of synonymous codons.

**E**. Relative expressivity of genes can be inferred based on the values of CU statistics, by comparing values calculated for each gene with the values for highly expressed genes. Codon bias statistics (B) described earlier can therefore be employed to calculate expressivity, termed E (Karlin and Mrazek, 2000). In the original paper, the authors defined a gene as predicted to be highly expressed (PHX) if it has codon frequencies similar to the codon frequencies of the ribosomal proteins (RP) genes, translation and transcription processing factors (TF) genes, and the major chaperone-degradation (CH) genes, but at the same time it deviates significantly in codon usage from the average gene of the genome (C), i.e. from the set of all protein coding genes.

Generally, for any class of reference genes S, a measure $E_S(g)$ for expression level of a gene $g$ relative to S is defined:

$$E_S(g) = B(g|C)/B(g|S) \quad (22)$$

The expressivity of a gene $g$ relative to RP, CH or TF genes can then be calculated:

$$E_{RP}(g) = \frac{B(g|C)}{B(g|RP)} \qquad E_{CH}(g) = \frac{B(g|C)}{B(g|CH)} \qquad E_{TF}(g) = \frac{B(g|C)}{B(g|TF)} \tag{23}$$

Combined, these produce the general expression measure used in the original paper:

$$E(g) = \frac{B(g|C)}{\frac{1}{2}B(g|RP) + \frac{1}{4}B(g|CH) + \frac{1}{4}B(g|TF)} \tag{24}$$

Authors note, however, that other weighted combinations can also be used. They finally defined a gene as PHX if the following two conditions are satisfied:

1) at least two among the three expression values exceed 1.05, and
2) the general expression level exceeds 1.

PHX genes in most prokaryotic genomes include, in addition to those for RP, TF, and CH proteins, the principal genes of energy metabolism and key genes involved in amino acid, nucleotide, and fatty acid biosynthesis.

**MILC-based Expression Level Predictor**. Analogously to E, the value of MELP for a gene is calculated as a ratio of MILC distance to an average CU of an entire set of genes, and a MILC distance to a reference set of highly expressed genes (Supek and Vlahoviček, 2005).

$$MELP = \frac{MILC_{set}}{MILC_{reference}} \tag{25}$$

**GCB**. In order to avoid having to rely on a set of highly expressed genes when determining relative expressivity, a new measure termed gene codon usage bias was devised (Merkl, 2003).

Codon scores for are first calculated as follows:

$$CB(cdn_j) = \log \frac{f_t(cdn_j)}{f_m(cdn_j)}, \tag{26}$$

where $f_t(cdn_j)$ is frequency of codon $j$ among synonymous codons in the target set (e.g. ribosomal genes), and $f_m(cdn_j)$ is mean frequency of the codon among synonymous codons in all genes of the genomic data set.

The extremely low CB scores obtained may be at least partially due to the small number of target codons analysed. In order not to overemphasise these codon frequencies, a lower limit of

-5 should be introduced for CB scores, i.e. for those codons with $f_t = 0$, a CB value should be set to -5.

Measure of codon usage bias in a gene is a likelihood-function computed as the normalized sum of codon scores:

$$GCB(gene) = \sum_{j=1}^{n} \frac{CB(cdn_j)}{n}, \tag{27}$$

where $gene = start, cdn_1, cdn_2, \ldots cdn_n$, and $cdn_j = cdn_1, cdn_2, \ldots cdn_n$. Note that here $n$ is the length of gene in codons, without the start codon, but including the STOP codon.

Rather than relying on a single target set of putative highly expressed genes (i.e. genes with high CU bias), an iterative approach is used for the identification of target genes and the concurrent derivation of CB scores. In the first iteration, a set of genes ("seed") known to have a biased codon composition is used as a target. CB and GCB are calculated, and genes are ranked based on the latter values. Top ranking genes are used as a target set for the next iteration. Algorithm is run until the scores are stable.

# 2. MATERIALS AND METHODS

## 2.1 R programming language

R is a free programming language and software environment for statistical computing and graphics (R Core Team, 2017). It is developed for all three major families of operating systems, Unix, Windows and Mac.

Initially, R was developed as one of three implementations of the S statistical language, but it has by now overtaken both basic S and the new S implementation, in terms of both the rate of development and the level of community support.

There are several reasons for the widespread usage of R in data analysis. For one, it is an open source scripting language in which one can easily and reproducibly perform and combine different types of data analyses (unlike the analysis performed in the equivalent point-and-click-based software).

However, probably the main reasons for the popularity of R is the fact that it is highly extensible. On top of the base R, there are many additional packages, which are collections of functions and datasets developed for wide range of tasks, by the users from various backgrounds, ranging from finance to life sciences. The packages significantly extend functionalities of base R. The official R repository, termed Comprehensive R Archive Network (CRAN), is a network of servers mirrored around the world, which store up-to-date versions of code, documentation and other R-related content. This official repository alone contains more than 10.000 published packages, and many more are available elsewhere on the Internet (e.g. on GitHub).

While the availability of packages for various purposes brings R closer to the audience with limited programming skills, its other main advantage brings it closer to other object-oriented (OO) programming languages, such as Python, Java, C++ or C#. There are two different object oriented programming systems in base R, one termed S3, and the other S4. S3 classes and methods are only loosely defined, and are essentially little more than just a naming convention. On the other hand, S4 OO system allows for classes and methods to have definitions that are more rigid. Every S4 class contains defined slots in which specific data types are stored, and can have validity method to ensure no incompatible data can be associated with it.

## 2.2 Bioconductor and dependencies

When it comes to the analysis of genomic data, the best collection of packages developed for wide range of applications, together with workflows in which they can be employed is available at Bioconductor (Huber *et al.*, 2015). The current release, Bioconductor 3.7 (May 2018), contains 1560 software packages, 342 experiment data packages, 919 annotation packages, and 21 workflows. This collection of packages provides many statistical and graphical methods for the analysis of genomic data. Bioconductor project also enables association of genomic data with biological metadata from common biological databases, processing genomic annotation data, and assembling customized annotation libraries.

Each Bioconductor package comes with associated vignettes, a task-oriented description of the package's functionality, which makes it easy for the end-users to apply the implemented methods in the analysis of their data. Providing open and accessible tools in this way also promotes reproducible research.

Thanks to the fact that nearly all Bioconductor packages are under an open source license, the algorithms and data structures which they implement have been modified and improved by an active community. This continuous effort resulted in the emergence of the best packages that are now commonly used for storing and analysing data from particular types of experiments. For instance, the Biostrings package (Pages *et al.*, 2017) contains various functions for manipulation of nucleotide and protein sequences, and the `XString` and `XStringSet` classes implemented in Biostrings package are used for storing those sequences. There are many Bioconductor packages that utilize these classes and define additional methods for them. The methods and classes implemented in coRdon are devised with this compatibility in mind. Another example is `AnnotatedDataFrame` class from Bioconductor package, Biobase (Huber *et al.*, 2015).

Apart from the mentioned Bioconductor packages, there are several other R packages that are also required for proper functioning of coRdon, and they are listed as dependencies in the Imports field of DESCRIPTION file, which stores important package metadata. coRdon dependencies include Biostrings and Biobase, data.table (Dowle and Srinivasan, 2018) package for fast manipulation of tabulated data, and two packages from tidyverse (Wickham, 2017) – a collection of R packages aimed at "tidy" data analysis – namely, purr (Henry and Wickham, 2017), used for data transformations, and ggplot2 (Wickham, 2009), used for visualizations.

## 2.3 Biological data

In this thesis, codon usage is analysed in two sets of DNA sequences: gut metagenomes from healthy individuals, and from liver cirrhosis patients (Table 2). The data is randomly selected part of larger set of sequence data from the extensive gut microbiome wide association study (Qin *et al.*, 2014), available from European Nucleotide Archive (ENA), under the accession number ERP005860. Sequences in each sample were previously quality-filtered, assembled and used to predict ORFs, which were then annotated with a KO (KEGG orthology) function (Fabijanić and Vlahoviček 2016). These sequences are used as the input for codon usage analysis with coRdon.

## 2.4 Codon usage analysis

### 2.4.1 Loading DNA sequences

To calculate codon usage (CU) bias for every sequence, it is necessary to get the counts of the occurrences of each codon. This can be done by storing sequences in each sample as a `codonTable` object. The `codonTable` class is designed to be compatible with the other two structures in which DNA sequences are commonly stored, namely fasta format files and `DNAStringSet` objects from the Biostrings package. This is depicted in Figure 1. I implemented a method `readSet()` for reading fasta files and storing sequences as a `DNAStringSet` object, which can then be converted to `codonTable` using the constructor method `codonTable()`.

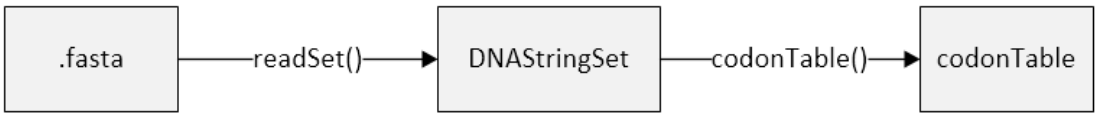


**Figure 1.** Integration of codonTable class with fasta formated files and DNAStringSet objects from Biostrings package.

**Table 2.** List of analyzed subset of sample identifiers from the original submission by Qin et al (2014). Table modified from Fabijanić and Vlahoviček (2016).

| *#* | *Individuals with cirrhosis* | *Healthy individuals* |
|---|---|---|
| *1* | LD1 | HD1 |
| *2* | LD2 | HD2 |
| *3* | LD3 | HD4 |
| *4* | LD4 | HD5 |
| *5* | LD5 | HD6 |
| *6* | LD6 | HD7 |
| *7* | LD7 | HD8 |
| *8* | LD12 | HD15 |
| *9* | LD13 | HD17 |
| *10* | LD14 | HD18 |
| *11* | LD17 | HD19 |
| *12* | LD30 | HD20 |
| *13* | LD31 | HD21 |
| *14* | LD32 | HD23 |
| *15* | LD50 | HD24 |
| *16* | LD52 | HD25 |
| *17* | LD61 | HD26 |
| *18* | LD63 | HD27 |
| *19* | LD66 | HD59 |
| *20* | LD69 | HD62 |
| *21* | LD74 | HD63 |
| *22* | LD75 | HD64 |
| *23* | LD76 | HD65 |
| *24* | LD79 | HD66 |
| *25* | LD84 | HD67 |
| *26* | LD94 | HD68 |
| *27* | LD95 | HD78 |
| *28* | LD96 | HD81 |
| *29* | LD97 | HD82 |
| *30* | LD98 | HD83 |

### 2.4.2 Calculating CU statistics

All of the statistics that measure codon usage bias, which are described in the previous chapter, can be calculated for sequences in the `codonTable` object by calling the appropriate method and giving it a `codonTable` object as the first argument.

For the data analysed in this thesis, I calculated the statistics that measure CU bias with respect to a reference set of genes (these include ENC', B, MCB, MILC), with ribosomal genes used as a reference. Statistics were calculated for those sequences longer than 80 codons, as the obtained values are generally not reliable for the short DNA sequences.

### 2.4.3 Predicting genes' expressivity

Values of codon usage statistics can be used to predict relative expression levels of genes in each sample. I calculated several different measures of CU-based gene expressivity by calling the appropriate methods, i.e. `E()`, `CAI()`, `MELP()`, and `Fop()`. Ribosomal genes were once again used as the reference, although any set of highly expressed genes can be used instead. Sequences shorter than 80 codons were again excluded.

Genes from a single sample having expressivity values in the top 10% are considered to be optimized for translation in that sample.

### 2.4.4 Functional annotation

The next analysis step was identification of those functions that are significantly enriched or depleted in the set of annotated genes predicted to have high expression level. I performed this analysis on two different levels of annotation, KEGG Orthologs (KO) and KEGG Pathways. For every level, I created a contingency table, summarising counts of genes annotated to each category among all genes in the sample, and among those predicted to be highly expressed. This can be easily achieved using `crossTab()` method implemented in coRdon, giving it as arguments both a character vector of genes' annotations, and a numeric vector of their respective expressivity values. Output of `crossTab()` is an object of `crossTab` class, containing gene annotations, respective values of the given variable, and a contingency table with counts for all genes, and for each defined subset of highly expressed genes.

Gene counts were than scaled by adding a pseudocount of 1 and transformed by MA transformation (log ratio M and mean average A). Relative enrichment was calculated as $\frac{scaled\ count\ of\ genes\ in\ the\ subset - scaled\ total\ count\ of\ genes}{scaled\ total\ count\ of\ genes} \cdot 100$. Using scaled counts, p values for enrichment were computed by binomial test with Benjamini-Hochberg correction for multiple testing.

I performed the described calculations using the `enrichment()` method implemented in coRdon. This takes a `crossTab` object and output an object of `AnnotatedDataFrame` class from Biobase package. The data stored in this object – including gene counts, M and A values, relative enrichment values, p values and adjusted p values – can be accessed using the `pData()` method from Biobase.

# 3. RESULTS

## 3.1 Implementation

### 3.1.1 Package

The coRdon package is stored in the GitHub repository BioinfoHR/coRdon (https://github.com/BioinfoHR/coRdon), and it can be downloaded directly from this repository. Additionally, coRdon can be downloaded from the development version of Bioconductor and will be included in the next release version (Bioconductor 3.8, due to be released in October).

### 3.1.2 Classes and methods

I implemented three new classes in coRdon, namely `codonTable`, `genCode`, and `crossTab`, as well as several methods for accessing and manipulating the objects of each class. All data structures are implemented in S4 OO system.

Objects of `codonTable` class are designed to store codon counts for a set of DNA sequences, along with some additional metadata used in the subsequent codon usage analysis. Every object of this class has five slots, as shown in UML class diagram in Figure 2. `ID`, `counts` and `len` are obligatory, while `KO` and `COG` can be empty vectors. ID slot contains character vector of DNA sequence identifiers. The main slot is `counts`, containing matrix of codon counts, with DNA sequences in rows and codons in columns. Sum of counts in each row is the length of each sequence (in codons), and numeric vector of these lengths is contained in `len` slot. Optional `KO` and `COG` slots can contain character vector of sequences KEGG ontology annotations or COG annotations, respectively.

Methods defined for objects of `codonTable` class are also shown in Figure 2. These include accessory methods (`show()`, `length()`, `subset()`), as well as `get` and `set` methods for accessing and modifying the slots of `codonTable` object, respectively.

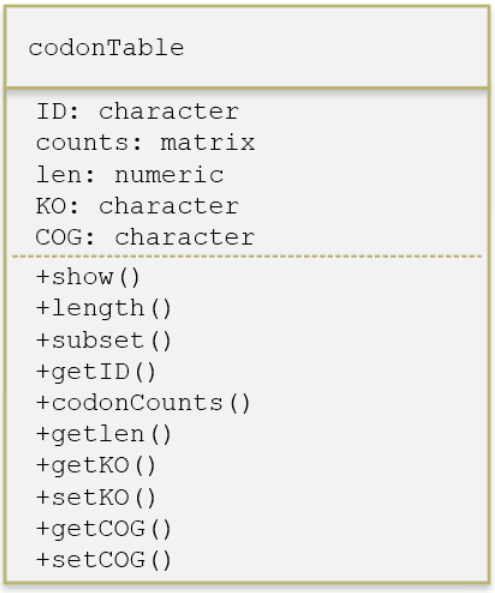


**Figure 2.** UML class diagram for `codonTable` class.

Objects of `genCode` class are not exported, and are therefore not visible to end-user, however, instances of this class are used by all the most important methods implemented in coRdon. Objects of this class contain relevant information on genetic code variant to be used in the analysis of codon usage. Figure 3 shows a UML class diagram for `genCode` class. The following slots are defined:

- `ctab`, a data.table (from data.table package) with two columns, one containing codons and the other respective amino acid they encode;
- `codons`, `stops` and `nostops`, each a character vector of codons;
- `cl`, a list, each element of which is a vector of integers indicating the positions of synonymous codons for that amino acid, when codons are alphabetically ordered;
- `deg`, a numeric vector of degeneracies for alphabetically ordered amino acids.

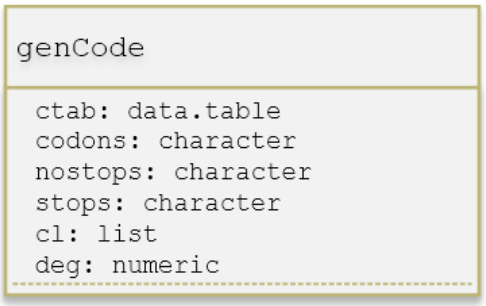


**Figure 3.** UML class diagram for `genCode` class.

The final S4 class defined in coRdon is `crossTab`. UML class diagram for this class is shown in Figure 4. Every object of `crossTab` class contains a contingency table, also referred to as a crosstab, summarizing the relationships between annotation categories of sequences in which CU is analysed, and counts of sequences in the (sub)sets defined according to the values of codon-usage-based expressivity predictors.

Slot `sequences` contains a character vector of sequences annotations; `variable` contains a numeric vector of the corresponding expressivity values; `table` contains a data.table (from data.table package) which is a contingency table with counts for all genes, and for each defined subset of genes.

Several methods are defined for this class. Apart from the usual accessor methods `show()` and `length()`, there are `getSeqAnnot()`, `getVariable()` and `contable()` methods to access the content of each slot, and a `reduceCrossTab()` method to reduce the input contingency table by mapping sequences' annotations to a broader ontology, either KEGG Pathway, KEGG Module or COG functional categories.

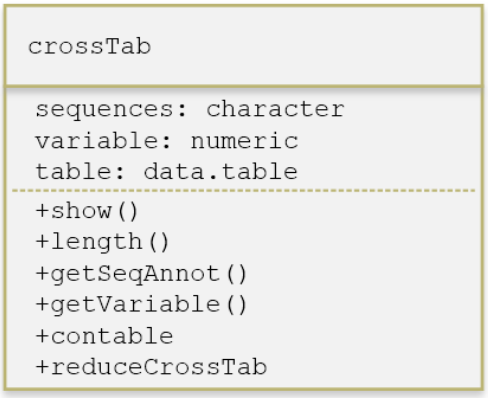


**Figure 4.** UML class diagram for `crossTab` class.

### 3.1.3 Statistics

I implemented all the statistics as described in the original publications.

For the calculation of $F_{op}$, codons are considered to be optimal if their relative adaptiveness, as defined in equation (19), is greater than 0.9.

In the analysis by (Merkl, 2003) the limit for calculating GCB was set to 6 iterations, and I included this in the coRdon implementation.

## 3.2 Codon usage analysis

A typical workflow for analysing codon usage bias includes:

- calculating one of the CU statistics
- comparing CU of every sequence (ORF, gene) to average CU of a set of highly expressed genes
- functional enrichment analysis for the subset of genes that have CU similar to a reference set of highly expressed genes and are therefore predicted to be highly expressed themselves.

### 3.2.1 Codon usage bias

Codon usage statistics calculated for all samples are generally consistent, as can be seen for the two samples on B plots in Figure 5. On a B plot, every gene is represented by a single point, its coordinates determined by the distance of genes' CU bias to overall CU bias (y axis) and to CU bias of reference genes (x axis).

A characteristic 'crescent moon' shape is visible in all the plots. ENC' plots appear flipped in comparison with the others, however, this is because larger values of ENC' correspond to CU closer to the reference, whereas for B, MCB and MILC larger values indicate departure from the reference CU.

A cut-off introduced in calculation of ENC' values (values > 61 reduced to 61) makes this statistic less reliable for sequences with CU bias close to the reference. Similarly, MCB values appear to be shrinking toward smaller values, which also limits the reliability of the statistic when measuring CU similar to the reference. On the other hand, MILC and B are better distributed over their dynamic range.

B plots for other samples display the same trends (not shown).

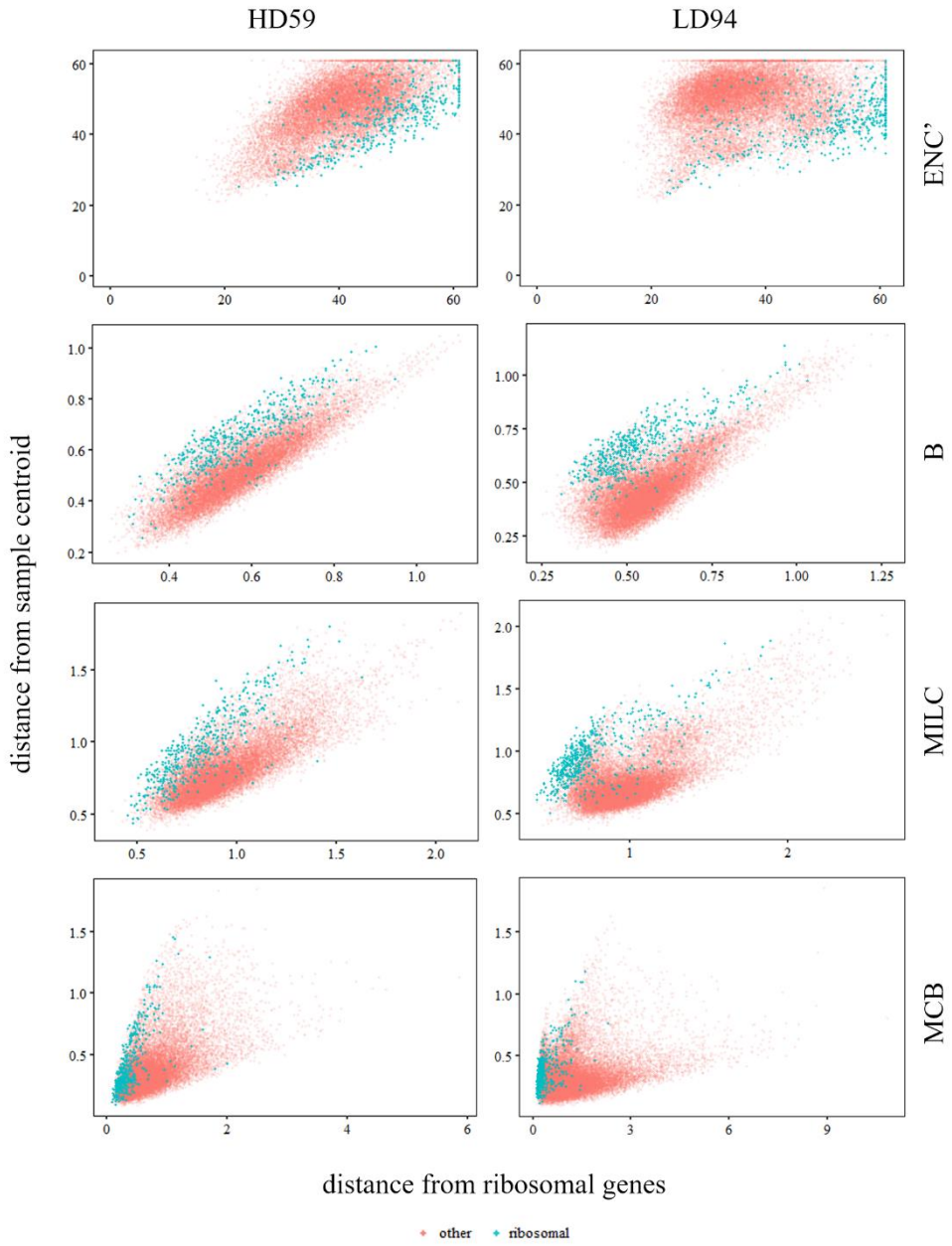


**Figure 5.** B plots for two metagenomic samples, one from healthy individual (HD59) and another from liver cirrhosis patient (LD94), showing codon usage patterns as calculated using indicated statistics (ENC, B, MILC, MCB).

The genes from any two metagenomic samples tend to cluster separately, such that genes in each sample are predominantly closer to their respective metagenome of origin. This is shown for two samples on the intra-sample B plot in Figure 6.

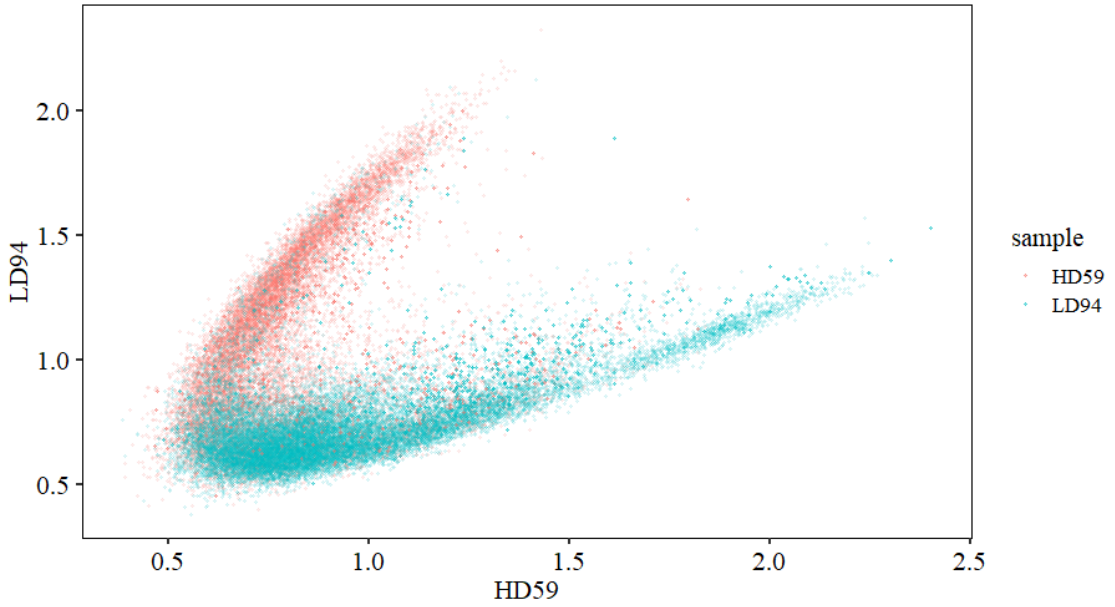


**Figure 6.** Intra-sample B plot showing MILC distance between CU bias in healthy metagenomic sample (HD59) and liver cirrhosis sample (LD94).

### 3.2.2 Relative genes' expressivity

Of the calculated gene expressivity statistics, E and MELP are highly correlated, as well as CAI and Fop (Figure 7 and Table 3**Table 3**). This is not unexpected, given the theoretical similarities in the calculation of these statistics.

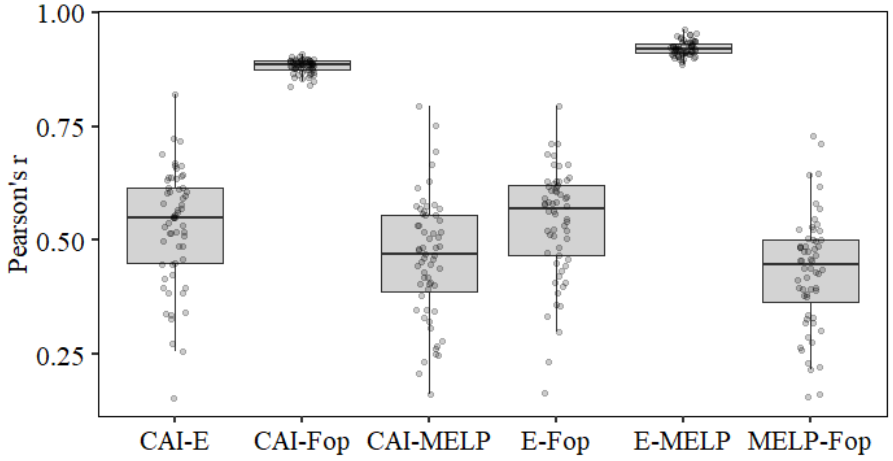


**Figure 7.** Correlation of codon usage based expressivity measures, calculated for all samples (n = 60). Box plots show median and interquartile range, with whiskers extending from 10$^{th}$ to 90$^{th}$ quantile. Overlaid dot plots show values for individual samples.

**Table 3.** Correlation of codon usage based expressivity measures, calculated for all samples (n = 60).

| *statistics* | *Pearson's r (mean ± sd)* |
|---|---|
| CAI – E | 0.53 ± 0.13 |
| CAI – Fop | 0.88 ± 0.02 |
| CAI – MELP | 0.46 ± 0.13 |
| E – Fop | 0.54 ± 0.12 |
| E – MELP | 0.92 ± 0.02 |
| MELP – Fop | 0.43 ± 0.12 |

For each sample, those genes with the expressivity in the top 10% values of the calculated statistic are predicted to be highly expressed. When comparing what genes are predicted to be highly expressed using different statistics, averaged over all samples, it turns out that about 5% are common for all the statistics (i.e. 0.5% of the total number of genes are predicted to be highly expressed, regardless of the statistic used). Another 20% of the highly expressed genes are predicted using E and MELP, but not with any other statistic, as well as another 20% predicted with CAI and Fop only. Additional ~20% of the genes predicted to be highly expressed are common for any three statistics. This is shown in the Venn diagram in Figure 8. Treating healthy and diseased samples separately did not improve the results (not shown).

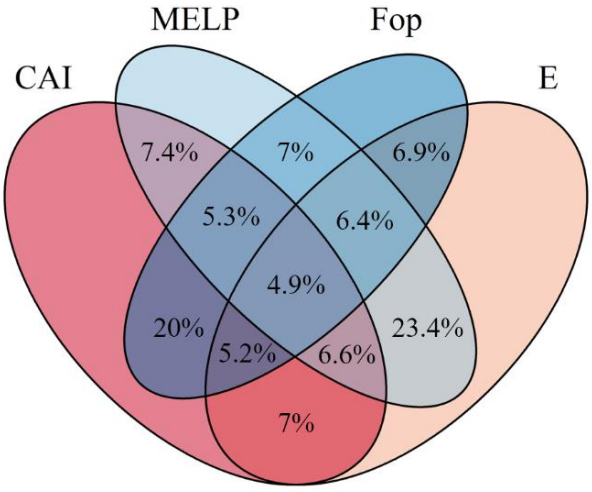


**Figure 8.** Venn diagram showing the proportions of genes with top 10% expressivity values, averaged over all samples (n = 60).

### 3.2.3 Functional enrichment

Enriched and depleted KEGG orthologues among highly expressed genes are visualized on the MA plots, which are commonly used for plotting the results of differential gene expression analysis. Examples of this plot for both diseased and healthy metagenomic sample are shown in Figure 9. Significantly enriched or depleted categories are defined as those with p value < 0.05.

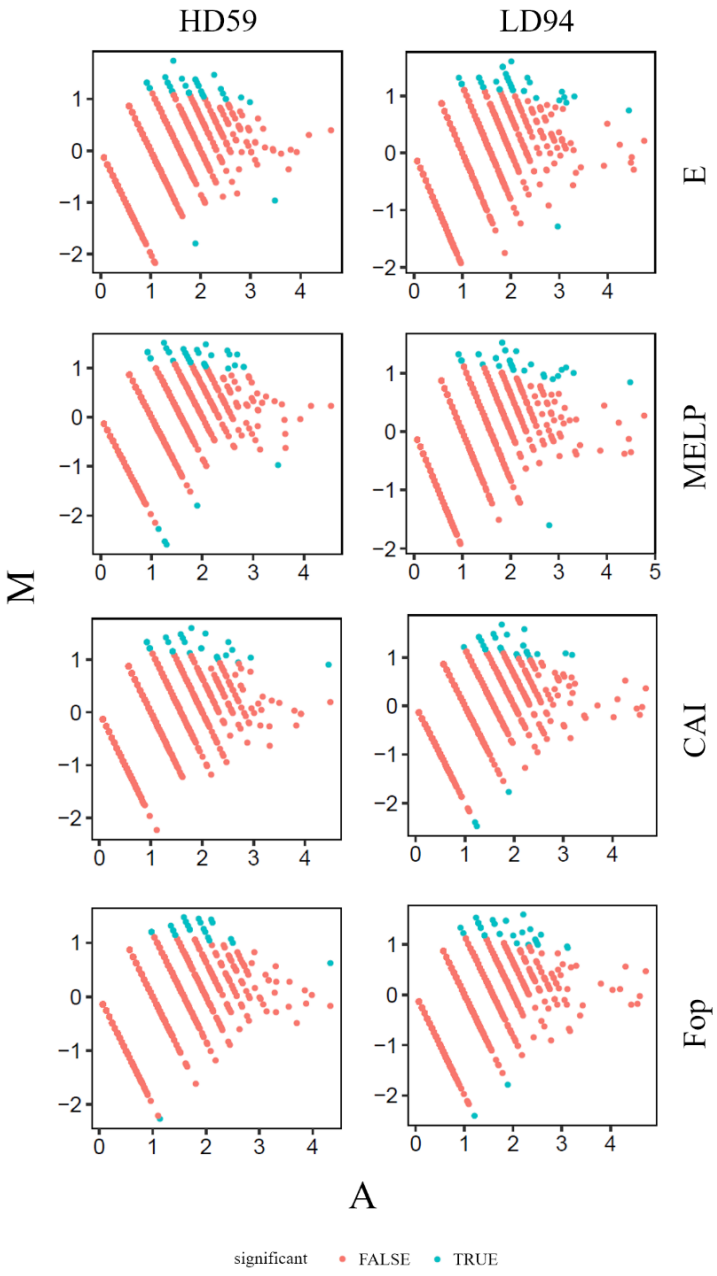


**Figure 9.** MA plots for two metagenomic gut samples, one healthy (HD59) and the other from the liver cirrhosis patient (LD94). Plotted enrichment values were calulated using indicated CU-bases expressivity measures (E, MELP, CAI or Fop). Significance is determined by p value < 0.05, calculated by binomial test.

Apart from the KO categories, enrichment was analysed at the level of KEGG Pathways. The resulting enriched pathways for one sample are shown on the bar plots in the Figure 10. Consistent with previously observed trends, using CAI and Fop to calculate enrichment results in more common categories between the two sets of results, than there are between the results of the analyses using the other statistics. Fop appears to be the most restrictive statistic, resulting in the fewest number of enriched categories in all samples. Generally, only a small number of categories is consistently found to be enriched when using any expressivity statistic (e.g. terpenoide backbone synthesis pathway is consistently found to be enriched in LD94 sample). This is also true for the other samples (data not shown).

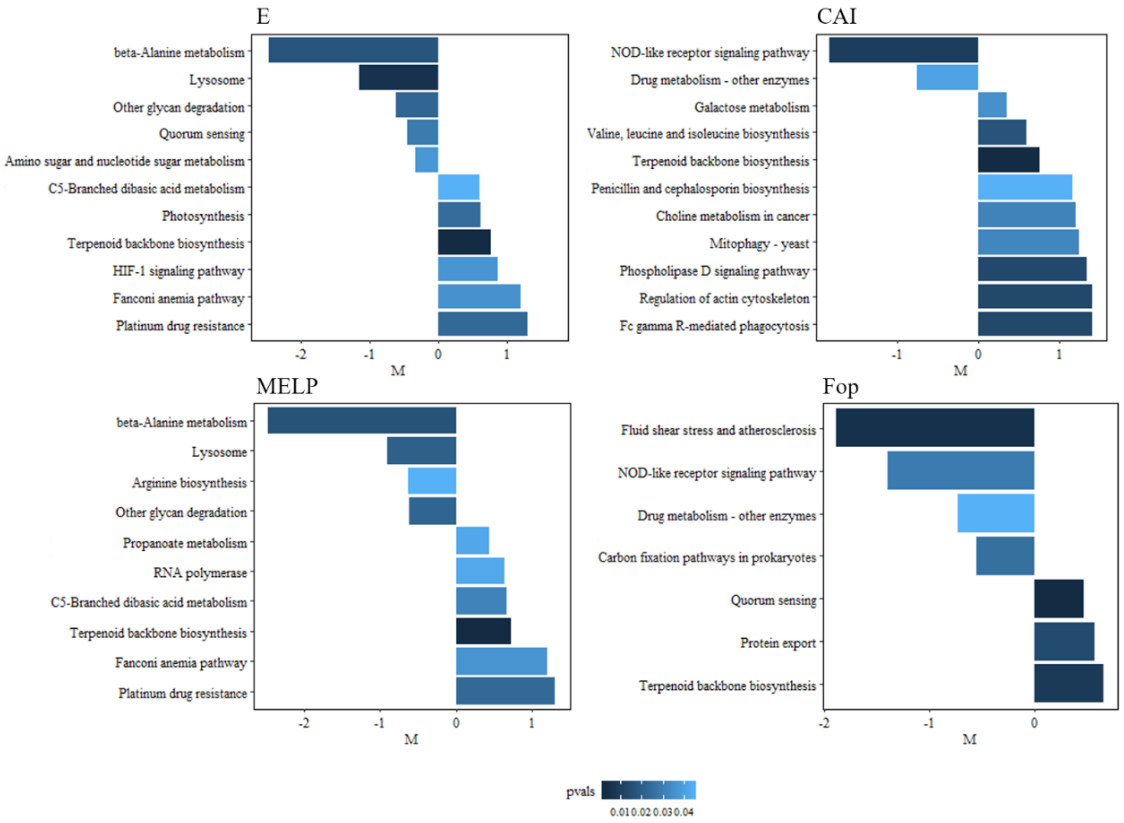


**Figure 10**. Enriched KEGG Pathways in the gut metagenome of liver cirrhosis patient (sample LD94). Enrichment is calculated using the codon usage based expressivity statistics indicated above each bar plot, and expressed as log ratio M (see Materials and Methods for details); p value < 0.5, calculated by binomial test.

From all pathways found to be enriched in any healthy sample, flagellar assembly pathway was enriched in the highest number of them. ABC transporters pathways, fatty acids, sugars and glycans metabolism, bacterial chemotaxis and ferroptosis pathways were also enriched in several healthy metagenomic samples. Somewhat unexpectedly, the other glycan degradation pathway (KEGG: map00511) is depleted in healthy samples, along with lipopolysaccharide biosynthesis and pyrimidine metabolism pathways.

In diseased samples, on the other hand, enriched pathways include phagocytosis and mitophagy, carbon fixation and amino acids metabolism, as well as fatty acids metabolism. Depleted are amino acids' biosynthetic pathways, thiamine metabolism pathway, bacterial secretion system and fluid shear stress pathway. Full list of enriched and depleted KEGG pathways can be found in Supplementary Tables 1 to 4.

# 4. DISCUSSION

## 4.1 Software

Over the years, numerous researchers have devised their own statistics to measure codon usage in genes, and some of them accompanied their publications with the software to facilitate calculation of the new statistics. Consequently, apart from the theoretical background and evaluation results giving support to different measures, the efficiency with which they could be computed also influenced their more or less frequent use by the scientific community.

To the best of my knowledge, the existing software for codon usage analysis is mostly limited by one or more of the following issues. The available tools are either specified for calculation of single statistics (e.g. ENCprime (Novembre, 2002), CodonO (Angellotti *et al.*, 2007)), or the calculations are limited to the level of a single genomes, most often that of model organisms, because they use internally implemented reference sets of highly expressed genes (e.g. CodonW (Peden, 1999)).

Although there are some tools that can be used to calculate various CU statistics, these generally have a point-and-click interface (e.g. INCA (Supek and Vlahovicek, 2004), ACUA (Vetrivel, Arunkumar and Dorairaj, 2007), DnaSP (Rozas *et al.*, 2017)). Despite many available options, this still limits the flexibility of analysis, while at the same time often not contributing to the ease of use because of the many available options that may not necessarily be relevant for the specific analysis.

Additionally, most tools only produce tabulated output, and lack any visualization, although there are exceptions (e.g. INCA). This necessitates the use of another tool for visualization and complicates the discovery process.

Finally, the performance of the existing tools often does not scale well with very large inputs. This is especially problematic for the software which is only available as a web server, because of the need to transfer over the Internet very large amounts of data, which is highly relevant in case of Next Generation Sequencing technologies.

All these reasons prompted me to develop a comprehensive tool for analysing codon usage in DNA sequences, regardless of their originating genome, and not limited to a single genome. As R programming language offers scalability and flexibility not available with user-interface-based tools, and Bioconductor project has in the recent years become the standard for genomic analyses (Huber *et al.*, 2015), I choose to develop an R package that complies with the Bioconductor guidelines. Contributing to such repository of tools for different purposes that

play nicely with each other opens a way for comprehensive in-depth analysis of genomic data sets that includes translational optimization.

Moreover, as a part of Bioconductor, with the next official release (due to be in October 2018) the package will be automatically introduced to a wide audience of biologists, and this will positively affect the maintenance process as it will hopefully result in further improvements based on the user feedback.

## 4.2 Analysis

Because translational optimization acts to regulate gene expression in prokaryotic organisms, codon usage analysis is potentially a good approach towards estimating relative expression levels of genes. The statistics used to quantify codon usage bias, however, are not always consistent, owing to different mathematical approaches employed in calculations. Nonetheless, by comparing results obtained for the data at hand using different statistics it is possible, for one, validate the results to some extent, and two, identify the highly specific set of genes expected to encode most abundant proteins in the cell.

A comparison like this can be easily performed using coRdon, as the syntax for calculation of any of the implemented codon usage statistics is essentially the same.

Going further, given that the amount of translational selection acting on a gene is related to its functional category (Supek *et al.*, 2010), it follows that functional analysis based on codon usage statistics should reveal important functions in the input set of sequences, even without the need for these sequences to have any prior functional annotation.

For gut metagenome samples, analysis of genes predicted to be highly expressed identified as enriched those functions known to be important to the interactions of microbes with host (Qin *et al.*, 2010), e.g. pathways for degradation of polysaccharides and fatty acids metabolism, as well as ABC transporters pathways, which enable gut microbiomes to partake in the metabolic reactions of the intestine, and chemotaxis and flagellar assembly pathway, responsible for the retention of microbiomes in the intestine. In the samples from patients with liver cirrhosis, phagocytosis and mitophagy pathways are enriched, possibly indicating activation of the non-specific immune system. Furthermore, some of the important functions were found to be depleted in the diseased samples, namely photosynthesis and oxidative phosphorylation, amino acid biosynthesis and vitamin metabolism, as well as the fluid shear stress pathway. This

observed functional depletion can be hypothesised to result in impaired metabolic contribution of gut microbiota and its less effective retention in the gut, respectively.

Combining the results obtained by analysing codon usage with the results of other omics analyses could lend accuracy to what we know so far about the molecular functions of gut metagenome.

# 5. CONCLUSION

I developed a complete R package, with accompanying documentation, tests and vignettes, for comprehensive analysis of codon usage in DNA sequences. The package is deposited in the GitHub repository BioinfoHR/coRdon (URL: https://github.com/BioinfoHR/coRdon) and can also be directly downloaded from the Bioconductor repository (URL: https://bioconductor.org/packages/devel/bioc/html/coRdon.html).

coRdon can be used to analyse codon usage in various unannotated or KEGG/COG annotated DNA sequences. It calculates different measures of CU bias and CU-based predictors of gene expression and performs gene set enrichment analysis for annotated sequences. Several methods for visualization of CU and enrichment analysis results are also implemented, including B plots for visualizing CU bias in a set of sequences, and between two sets of sequences, as well as MA plots and bar plots for visualizing enriched and depleted functions in a set of sequences. Lastly, coRdon implements a method that allows seamless integration of the obtained results with other analyses and visualisation using other methods available in R and Bioconductor.

As a working example of codon usage, in this thesis I analysed codon usage in the gut metagenome samples from healthy individuals and from liver cirrhosis patients. I was able to determine levels of genes' translational optimization, and predict which genes are optimized for high levels of expression, as well as determine enriched functions in intestinal microbiomes of cirrhotic patients and healthy individuals.

# 6. REFERENCES

Angellotti, M. C., Bhuiyan, S. B., Chen, G. and Wan, X.-F. (2007) ‘CodonO: codon usage bias analysis within and across genomes’, *Nucleic Acids Research*. Oxford University Press, 35 (Web Server).

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., Azam, F. and Rohwer, F. (2002) ‘Genomic analysis of uncultured marine viral communities’, *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 99(22), pp. 14250–5.

Cozzi, P., Milanesi, L. and Bernardi, G. (2015) ‘Segmenting the Human Genome into Isochores’, *Evolutionary bioinformatics online*. SAGE Publications, 11, pp. 253–61.

Crick, F. (1958) ‘On Protein Synthesis’, *Symposia of the Society for Experimental Biology*. Cambridge University Press, 12, pp. 138–163.

Devaraj, S., Hemarajata, P. and Versalovic, J. (2013) ‘The Human Gut Microbiome and Body Metabolism: Implications for Obesity and Diabetes’, *Clinical Chemistry*, 59(4), pp. 617–628.

Dowle, M. and Srinivasan, A. (2018) ‘data.table: Extension of `data.frame`’. R package version 1.11.5. Available at: https://cran.r-project.org/package=data.table.

Gouy, M. and Gautier, C. (1982) ‘Codon usage in bacteria: correlation with gene expressivity’, *Nucleic Acids Research*. Oxford University Press, 10(22), pp. 7055–74.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) ‘Codon catalog usage is a genome strategy modulated for gene expressivity’, *Nucleic Acids Research*. Oxford University Press, 9(1), pp. r43-74.

Henry, L. and Wickham, H. (2017) ‘purrr: Functional Programming Tools’. R package vesion 0.2.5. Available at: https://cran.r-project.org/package=purrr.

Hooper, S. D. and Berg, O. G. (2000) ‘Gradients in nucleotide and codon usage along Escherichia coli genes’, *Nucleic Acids Research*. Oxford University Press, 28(18), pp. 3517–23.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., Pagès, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L. and Morgan, M. (2015) ‘Orchestrating high-throughput genomic analysis with Bioconductor’, *Nature Methods*, 12(2),

pp. 115–121.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., von Mering, C. and Bork, P. (2016) 'eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences', *Nucleic Acids Research*. Oxford University Press, 44(D1), pp. D286–D293.

Human Microbiome Project Consortium (2012) 'Structure, function and diversity of the healthy human microbiome', *Nature*, 486(7402), pp. 207–214.

Ikemura, T. (1981) 'Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system', *Journal of Molecular Biology*. Academic Press, 151(3), pp. 389–409.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K. (2017) 'KEGG: new perspectives on genomes, pathways, diseases and drugs', *Nucleic Acids Research*, 45(D1), pp. D353–D361.

Karlin, S., Campbell, A. M. and Mrazek, J. (1998) 'Comparative DNA analysis across diverse genomes', *Annual Review of Genetics*, 32(1), pp. 185–225.

Karlin, S. and Mrazek, J. (1996) 'What Drives Codon Choices in Human Genes?', *Journal of Molecular Biology*. Academic Press, 262(4), pp. 459–472.

Karlin, S. and Mrazek, J. (2000) 'Predicted highly expressed genes of diverse prokaryotic genomes', *Journal of Bacteriology*. American Society for Microbiology, 182(18), pp. 5238–50.

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., Brady, A., Creasy, H. H., McCracken, C., Giglio, M. G., McDonald, D., Franzosa, E. A., Knight, R., White, O. and Huttenhower, C. (2017) 'Strains, functions and dynamics in the expanded Human Microbiome Project', *Nature*, 550(7674), pp. 61–66.

Merkl, R. (2003) 'A Survey of Codon and Amino Acid Frequency Bias in Microbial Genomes Focusing on Translational Efficiency', *Journal of Molecular Evolution*, 57(4), pp. 453–466.

Novembre, J. A. (2002) ‘Accounting for Background Nucleotide Composition When Measuring Codon Usage Bias’, *Molecular Biology and Evolution*. Oxford University Press, 19(8), pp. 1390–1394.

Pages, H., Aboyoun, P., Gentleman, R. and DebRoy, S. (2017) ‘Biostrings: Efficient manipulation of biological strings’. R package version 2.48.0. Available at: https://bioconductor.org/packages/release/bioc/html/Biostrings.html/.

Peden, J. (1999) ‘CodonW’. Available at: http://codonw.sourceforge.net/.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S. D., Wang, J. and Wang, J. (2010) ‘A human gut microbial gene catalogue established by metagenomic sequencing’, *Nature*, 464(7285), pp. 59–65.

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., Guo, J., Le Chatelier, E., Yao, J., Wu, L., Zhou, J., Ni, S., Liu, L., Pons, N., Batto, J. M., Kennedy, S. P., Leonard, P., Yuan, C., Ding, W., Chen, Y., Hu, X., Zheng, B., Qian, G., Xu, W., Ehrlich, S. D., Zheng, S. and Li, L. (2014) ‘Alterations of the human gut microbiome in liver cirrhosis’, *Nature*. Nature Publishing Group, 513(7516), pp. 59–64.

Quax, T. E. F., Claassens, N. J., Söll, D. and van der Oost, J. (2015) ‘Codon Bias as a Means to Fine-Tune Gene Expression’, *Molecular Cell*, 59(2), pp. 149–161.

R Core Team (2017) ‘R: A language and environment for statistical computing’. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.r-project.org/.

Roller, M., Lucić, V., Nagy, I., Perica, T. and Vlahoviček, K. (2013) ‘Environmental shaping of codon usage and functional adaptation across microbial communities’, *Nucleic Acids Research*, 41(19), pp. 8842–8852.

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E. and Sánchez-Gracia, A. (2017) ‘DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets’, *Molecular Biology and Evolution*, 34(12), pp. 3299–3302.

Sender, R., Fuchs, S. and Milo, R. (2016) ‘Revised Estimates for the Number of Human and Bacteria Cells in the Body’, *PLOS Biology*. Public Library of Science, 14(8), p. e1002533.

Sharp, P. M. and Li, W. H. (1987) ‘The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications’, *Nucleic Acids Research*. Oxford University Press, 15(3), pp. 1281–95.

Staley, J. T. and Konopka, A. (1985) ‘Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats’, *Annual Review of Microbiology*, 39(1), pp. 321–346.

Supek, F., Škunca, N., Repar, J., Vlahoviček, K. and Šmuc, T. (2010) ‘Translational Selection Is Ubiquitous in Prokaryotes’, *PLoS Genetics*. Edited by N. A. Moran. Public Library of Science, 6(6), p. e1001004.

Supek, F. and Vlahovicek, K. (2004) ‘INCA: synonymous codon usage analysis and clustering by means of self-organizing map’, *Bioinformatics*, 20(14), pp. 2329–2330.

Supek, F. and Vlahoviček, K. (2005) ‘Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity’, *BMC Bioinformatics*, 6(1), p. 182.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V, Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V, Vasudevan, S., Wolf, Y. I., Yin, J. J. and Natale, D. A. (2003) ‘The COG database: an updated version includes eukaryotes’, *BMC Bioinformatics*, 4(1), p. 41.

Urrutia, A. O. and Hurst, L. D. (2001) ‘Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection’, *Genetics*, 159(3), pp. 1191–9.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H. and Smith, H. O. (2004) ‘Environmental Genome Shotgun Sequencing of the Sargasso Sea’, *Science*, 304(5667), pp. 66–74.

Vetrivel, U., Arunkumar, V. and Dorairaj, S. (2007) ‘ACUA: a software tool for automated codon usage analysis’, *Bioinformation*. Biomedical Informatics Publishing Group, 2(2), pp. 62–3.

Wan, X.-F., Xu, D., Kleinhofs, A. and Zhou, J. (2004) ‘Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes’, *BMC Evolutionary Biology*, 4(1), p. 19.

Wickham, H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer. R package version 3.0.0. Available at: http://ggplot2.org.

Wickham, H. (2017) ‘tidyverse: Easily Install and Load the “Tidyverse”’. R package version 1.2.1. Available at: https://cran.r-project.org/package=tidyverse.

Wright, F. (1990) ‘The “effective number of codons” used in a gene’, *Gene*, 87(1), pp. 23–9.

# CURRICULUM VITAE

Anamaria Elek

## EDUCATION

October 2016 - September 2018
University Graduate Programme in Molecular Biology
Faculty of Science, University of Zagreb (Croatia)

October 2013 - September 2016
Bachelor of Science in Molecular Biology
Magna Cum Laude
Faculty of Science, University of Zagreb (Croatia)

## WORK EXPERIENCE

July - September 2018
Internship
Exaltum, Zagreb (Croatia)

April 2018
Internship
OmicX, Rouen (France)

July - August 2017
Summer Undergraduate Internship
Max F. Perutz Laboratories, Vienna (Austria)
Alwin Köhler Group

May - September 2016
Student Internship
Ruđer Bošković Institute, Zagreb (Croatia)
Iva Tolić Group

## SKILLS

Languages
Croatian (native proficiency)
English (full professional proficiency)
German (elementary)
Italian (elementary)

Software
R, Git, Linux

# SUPPLEMENT

**Supplementary Table 1.** KEGG Pathways found to be significantly enriched and depleted in gut metagenomes of healthy individuals and liver cirrhosis patients. Functional analysis is based on codon usage in the subset of genes predicted to be highly expressed, as quantified by CAI (Sharp and Li, 1987). Details in the text.

| enriched in healthy samples | number of samples |
|---|---|
| Flagellar assembly | 6 |
| Antigen processing and presentation | 4 |
| Estrogen signaling pathway | 4 |
| IL-17 signaling pathway | 4 |
| Progesterone-mediated oocyte maturation | 4 |
| Protein processing in endoplasmic reticulum | 4 |
| Th17 cell differentiation | 4 |
| ABC transporters | 3 |
| Arginine biosynthesis | 3 |
| Fatty acid degradation | 3 |
| Ferroptosis | 3 |
| Glycerophospholipid metabolism | 3 |
| Photosynthesis | 3 |
| PI3K-Akt signaling pathway | 3 |
| Prostate cancer | 3 |
| Valine, leucine and isoleucine degradation | 3 |
| Adipocytokine signaling pathway | 2 |
| Adrenergic signaling in cardiomyocytes | 2 |
| Ascorbate and aldarate metabolism | 2 |
| Butanoate metabolism | 2 |
| cAMP signaling pathway | 2 |
| Carbapenem biosynthesis | 2 |
| Cell cycle | 2 |
| Cell cycle - Caulobacter | 2 |
| Choline metabolism in cancer | 2 |
| D-Alanine metabolism | 2 |
| Glucagon signaling pathway | 2 |
| Glucosinolate biosynthesis | 2 |
| Glycerolipid metabolism | 2 |
| Human papillomavirus infection | 2 |
| Insect hormone biosynthesis | 2 |
| Lysine degradation | 2 |
| Novobiocin biosynthesis | 2 |
| Oxidative phosphorylation | 2 |
| Pathways in cancer | 2 |
| Peroxisome | 2 |
| Plant-pathogen interaction | 2 |
| Polyketide sugar unit biosynthesis | 2 |
| PPAR signaling pathway | 2 |
| Primary bile acid biosynthesis | 2 |
| Propanoate metabolism | 2 |
| Proteoglycans in cancer | 2 |
| Pyrimidine metabolism | 2 |
| Regulation of actin cytoskeleton | 2 |

| enriched in healthy samples | number of samples |
|---|---|
| Ribosome | 2 |
| RNA degradation | 2 |
| Salivary secretion | 2 |
| Staphylococcus aureus infection | 2 |
| Sulfur relay system | 2 |
| Two-component system | 2 |
| Type II diabetes mellitus | 2 |
| Ubiquitin mediated proteolysis | 2 |
| Viral carcinogenesis | 2 |
| Vitamin B6 metabolism | 2 |
| 2-Oxocarboxylic acid metabolism | 1 |
| Acarbose and validamycin biosynthesis | 1 |
| African trypanosomiasis | 1 |
| Alanine, aspartate and glutamate metabolism | 1 |
| Aldosterone synthesis and secretion | 1 |
| alpha-Linolenic acid metabolism | 1 |
| Aminoacyl-tRNA biosynthesis | 1 |
| Antifolate resistance | 1 |
| Apelin signaling pathway | 1 |
| Arginine and proline metabolism | 1 |
| Atrazine degradation | 1 |
| Autophagy - yeast | 1 |
| Bacterial chemotaxis | 1 |
| Bacterial secretion system | 1 |
| Basal transcription factors | 1 |
| Benzoate degradation | 1 |
| beta-Lactam resistance | 1 |
| Biofilm formation - Escherichia coli | 1 |
| Biofilm formation - Pseudomonas aeruginosa | 1 |
| Biosynthesis of ansamycins | 1 |
| Biosynthesis of secondary metabolites | 1 |
| Biosynthesis of siderophore group nonribosomal peptides | 1 |
| Biosynthesis of vancomycin group antibiotics | 1 |
| C5-Branched dibasic acid metabolism | 1 |
| Calcium signaling pathway | 1 |
| Carbon fixation in photosynthetic organisms | 1 |
| Carbon metabolism | 1 |
| Cationic antimicrobial peptide (CAMP) resistance | 1 |
| Cell cycle - yeast | 1 |
| Central carbon metabolism in cancer | 1 |
| cGMP-PKG signaling pathway | 1 |
| Chagas disease (American trypanosomiasis) | 1 |
| Chlorocyclohexane and chlorobenzene degradation | 1 |
| Cytosolic DNA-sensing pathway | 1 |
| D-Glutamine and D-glutamate metabolism | 1 |
| Degradation of aromatic compounds | 1 |
| Dioxin degradation | 1 |
| DNA replication | 1 |
| Drug metabolism - other enzymes | 1 |
| Endocytosis | 1 |

| enriched in healthy samples | number of samples |
|---|---|
| Epithelial cell signaling in Helicobacter pylori infection | 1 |
| Epstein-Barr virus infection | 1 |
| Ether lipid metabolism | 1 |
| Fatty acid biosynthesis | 1 |
| Flavone and flavonol biosynthesis | 1 |
| Fluid shear stress and atherosclerosis | 1 |
| Fluorobenzoate degradation | 1 |
| Fructose and mannose metabolism | 1 |
| Glycine, serine and threonine metabolism | 1 |
| Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate | 1 |
| Glycosphingolipid biosynthesis - globo and isoglobo series | 1 |
| Glyoxylate and dicarboxylate metabolism | 1 |
| Histidine metabolism | 1 |
| Insulin resistance | 1 |
| Limonene and pinene degradation | 1 |
| Linoleic acid metabolism | 1 |
| Lipopolysaccharide biosynthesis | 1 |
| Longevity regulating pathway - multiple species | 1 |
| Lysine biosynthesis | 1 |
| MAPK signaling pathway - yeast | 1 |
| Metabolic pathways | 1 |
| Mismatch repair | 1 |
| Mitophagy - yeast | 1 |
| Necroptosis | 1 |
| Neomycin, kanamycin and gentamicin biosynthesis | 1 |
| Non-alcoholic fatty liver disease (NAFLD) | 1 |
| Osteoclast differentiation | 1 |
| p53 signaling pathway | 1 |
| Pancreatic secretion | 1 |
| Pantothenate and CoA biosynthesis | 1 |
| Parkinson disease | 1 |
| Phenazine biosynthesis | 1 |
| Phenylalanine metabolism | 1 |
| Phenylalanine, tyrosine and tryptophan biosynthesis | 1 |
| Phosphonate and phosphinate metabolism | 1 |
| Porphyrin and chlorophyll metabolism | 1 |
| Primary immunodeficiency | 1 |
| Prolactin signaling pathway | 1 |
| Protein digestion and absorption | 1 |
| Protein export | 1 |
| Purine metabolism | 1 |
| Quorum sensing | 1 |
| Retrograde endocannabinoid signaling | 1 |
| Rheumatoid arthritis | 1 |
| Ribosome biogenesis in eukaryotes | 1 |
| RNA transport | 1 |
| Secondary bile acid biosynthesis | 1 |
| Selenocompound metabolism | 1 |
| Sphingolipid metabolism | 1 |
| Spliceosome | 1 |

| enriched in healthy samples | number of samples |
|---|---|
| Thermogenesis | 1 |
| Thyroid hormone synthesis | 1 |
| Toll-like receptor signaling pathway | 1 |
| Toll and Imd signaling pathway | 1 |
| Toluene degradation | 1 |
| Tropane, piperidine and pyridine alkaloid biosynthesis | 1 |
| Valine, leucine and isoleucine biosynthesis | 1 |
| Vancomycin resistance | 1 |
| Xylene degradation | 1 |
| **depleted in healthy samples** | **number of samples** |
| Other glycan degradation | 4 |
| Amino sugar and nucleotide sugar metabolism | 2 |
| Drug metabolism - other enzymes | 2 |
| Fatty acid degradation | 2 |
| Glucagon signaling pathway | 2 |
| Glycerophospholipid metabolism | 2 |
| Insulin resistance | 2 |
| Monobactam biosynthesis | 2 |
| Pentose and glucuronate interconversions | 2 |
| Pyrimidine metabolism | 2 |
| Ribosome | 2 |
| RNA polymerase | 2 |
| Aminoacyl-tRNA biosynthesis | 1 |
| Antifolate resistance | 1 |
| Bacterial chemotaxis | 1 |
| Bacterial secretion system | 1 |
| Base excision repair | 1 |
| Biofilm formation - Escherichia coli | 1 |
| Biotin metabolism | 1 |
| Carbon fixation in photosynthetic organisms | 1 |
| Carbon fixation pathways in prokaryotes | 1 |
| Cell cycle - Caulobacter | 1 |
| Citrate cycle (TCA cycle) | 1 |
| D-Glutamine and D-glutamate metabolism | 1 |
| Degradation of aromatic compounds | 1 |
| Fatty acid metabolism | 1 |
| Ferroptosis | 1 |
| Flagellar assembly | 1 |
| Folate biosynthesis | 1 |
| Fructose and mannose metabolism | 1 |
| Glycosaminoglycan degradation | 1 |
| Glyoxylate and dicarboxylate metabolism | 1 |
| Insulin signaling pathway | 1 |
| Lipopolysaccharide biosynthesis | 1 |
| Lysosome | 1 |
| Metabolic pathways | 1 |
| MicroRNAs in cancer | 1 |
| Nicotinate and nicotinamide metabolism | 1 |
| NOD-like receptor signaling pathway | 1 |
| One carbon pool by folate | 1 |
| Oxidative phosphorylation | 1 |
| Peptidoglycan biosynthesis | 1 |

| depleted in healthy samples | number of samples |
|---|---|
| Phenylalanine, tyrosine and tryptophan biosynthesis | 1 |
| Phenylpropanoid biosynthesis | 1 |
| Phosphotransferase system (PTS) | 1 |
| Photosynthesis | 1 |
| Polyketide sugar unit biosynthesis | 1 |
| Porphyrin and chlorophyll metabolism | 1 |
| Terpenoid backbone biosynthesis | 1 |
| Tropane, piperidine and pyridine alkaloid biosynthesis | 1 |
| Tyrosine metabolism | 1 |
| Vancomycin resistance | 1 |
| **enriched in diseased samples** | **number of samples** |
| Phosphotransferase system (PTS) | 5 |
| Propanoate metabolism | 4 |
| Ribosome | 4 |
| Benzoate degradation | 3 |
| Carbon fixation pathways in prokaryotes | 3 |
| Choline metabolism in cancer | 3 |
| Fc gamma R-mediated phagocytosis | 3 |
| Limonene and pinene degradation | 3 |
| Mitophagy - yeast | 3 |
| Novobiocin biosynthesis | 3 |
| Regulation of actin cytoskeleton | 3 |
| 2-Oxocarboxylic acid metabolism | 2 |
| Aminobenzoate degradation | 2 |
| Biosynthesis of siderophore group nonribosomal peptides | 2 |
| Butanoate metabolism | 2 |
| Carbon metabolism | 2 |
| Cell cycle | 2 |
| Cell cycle - yeast | 2 |
| Complement and coagulation cascades | 2 |
| D-Arginine and D-ornithine metabolism | 2 |
| D-Glutamine and D-glutamate metabolism | 2 |
| Dioxin degradation | 2 |
| Drug metabolism - cytochrome P450 | 2 |
| Endocytosis | 2 |
| Glycerolipid metabolism | 2 |
| Hippo signaling pathway | 2 |
| Histidine metabolism | 2 |
| Insect hormone biosynthesis | 2 |
| Legionellosis | 2 |
| Lysine biosynthesis | 2 |
| Lysine degradation | 2 |
| MAPK signaling pathway - yeast | 2 |
| Mismatch repair | 2 |
| Monobactam biosynthesis | 2 |
| Nucleotide excision repair | 2 |
| Pantothenate and CoA biosynthesis | 2 |
| Phenylalanine metabolism | 2 |
| Phosphatidylinositol signaling system | 2 |
| Phospholipase D signaling pathway | 2 |
| Porphyrin and chlorophyll metabolism | 2 |
| Proximal tubule bicarbonate reclamation | 2 |

| enriched in diseased samples | number of samples |
|---|---|
| Pyruvate metabolism | 2 |
| Sphingolipid metabolism | 2 |
| Synthesis and degradation of ketone bodies | 2 |
| Terpenoid backbone biosynthesis | 2 |
| Two-component system | 2 |
| Type I diabetes mellitus | 2 |
| Ubiquitin mediated proteolysis | 2 |
| Valine, leucine and isoleucine biosynthesis | 2 |
| Valine, leucine and isoleucine degradation | 2 |
| Xylene degradation | 2 |
| Adrenergic signaling in cardiomyocytes | 1 |
| AGE-RAGE signaling pathway in diabetic complications | 1 |
| Alcoholism | 1 |
| Amphetamine addiction | 1 |
| Antifolate resistance | 1 |
| Antigen processing and presentation | 1 |
| Apelin signaling pathway | 1 |
| Basal transcription factors | 1 |
| beta-Alanine metabolism | 1 |
| Betalain biosynthesis | 1 |
| Bile secretion | 1 |
| Biosynthesis of secondary metabolites | 1 |
| Biosynthesis of unsaturated fatty acids | 1 |
| C5-Branched dibasic acid metabolism | 1 |
| cAMP signaling pathway | 1 |
| Caprolactam degradation | 1 |
| Carbapenem biosynthesis | 1 |
| Carbohydrate digestion and absorption | 1 |
| Cardiac muscle contraction | 1 |
| Cellular senescence | 1 |
| Central carbon metabolism in cancer | 1 |
| Citrate cycle (TCA cycle) | 1 |
| Cocaine addiction | 1 |
| D-Alanine metabolism | 1 |
| DNA replication | 1 |
| Dopaminergic synapse | 1 |
| Estrogen signaling pathway | 1 |
| Flagellar assembly | 1 |
| Galactose metabolism | 1 |
| Gastric acid secretion | 1 |
| Geraniol degradation | 1 |
| Glutathione metabolism | 1 |
| Glycerophospholipid metabolism | 1 |
| Glycosylphosphatidylinositol (GPI)-anchor biosynthesis | 1 |
| Glyoxylate and dicarboxylate metabolism | 1 |
| Homologous recombination | 1 |
| Human papillomavirus infection | 1 |
| IL-17 signaling pathway | 1 |
| Indole alkaloid biosynthesis | 1 |
| Longevity regulating pathway - worm | 1 |
| Meiosis - yeast | 1 |
| Metabolism of xenobiotics by cytochrome P450 | 1 |

| enriched in diseased samples | number of samples |
|---|---|
| MicroRNAs in cancer | 1 |
| Necroptosis | 1 |
| Nitrogen metabolism | 1 |
| Nonribosomal peptide structures | 1 |
| One carbon pool by folate | 1 |
| p53 signaling pathway | 1 |
| Penicillin and cephalosporin biosynthesis | 1 |
| Pentose phosphate pathway | 1 |
| Phenazine biosynthesis | 1 |
| Photosynthesis | 1 |
| PI3K-Akt signaling pathway | 1 |
| Platinum drug resistance | 1 |
| Polyketide sugar unit biosynthesis | 1 |
| Primary immunodeficiency | 1 |
| Prostate cancer | 1 |
| Protein export | 1 |
| Protein processing in endoplasmic reticulum | 1 |
| Purine metabolism | 1 |
| Pyrimidine metabolism | 1 |
| Quorum sensing | 1 |
| Retinol metabolism | 1 |
| RIG-I-like receptor signaling pathway | 1 |
| Salivary secretion | 1 |
| Salmonella infection | 1 |
| Selenocompound metabolism | 1 |
| Serotonergic synapse | 1 |
| Staphylococcus aureus infection | 1 |
| Starch and sucrose metabolism | 1 |
| Streptomycin biosynthesis | 1 |
| Sulfur metabolism | 1 |
| Taurine and hypotaurine metabolism | 1 |
| Th17 cell differentiation | 1 |
| Thyroid hormone signaling pathway | 1 |
| Tryptophan metabolism | 1 |
| Tuberculosis | 1 |
| Type II diabetes mellitus | 1 |
| Ubiquinone and other terpenoid-quinone biosynthesis | 1 |
| Viral carcinogenesis | 1 |
| **depleted in diseased samples** | **number of samples** |
| Fluid shear stress and atherosclerosis | 3 |
| Photosynthesis | 3 |
| Tuberculosis | 3 |
| Degradation of aromatic compounds | 2 |
| Glucagon signaling pathway | 2 |
| Peroxisome | 2 |
| Quorum sensing | 2 |
| Sphingolipid metabolism | 2 |
| Sulfur metabolism | 2 |
| Acarbose and validamycin biosynthesis | 1 |
| Antifolate resistance | 1 |
| Benzoate degradation | 1 |
| Biofilm formation - Vibrio cholerae | 1 |

| depleted in diseased samples | number of samples |
|---|---|
| Biosynthesis of ansamycins | 1 |
| Carbon fixation in photosynthetic organisms | 1 |
| Carbon fixation pathways in prokaryotes | 1 |
| Cell cycle - Caulobacter | 1 |
| Central carbon metabolism in cancer | 1 |
| D-Alanine metabolism | 1 |
| D-Glutamine and D-glutamate metabolism | 1 |
| DNA replication | 1 |
| Drug metabolism - other enzymes | 1 |
| Epithelial cell signaling in Helicobacter pylori infection | 1 |
| Fatty acid biosynthesis | 1 |
| Fatty acid degradation | 1 |
| Fatty acid metabolism | 1 |
| GABAergic synapse | 1 |
| Galactose metabolism | 1 |
| Glycerolipid metabolism | 1 |
| Glycine, serine and threonine metabolism | 1 |
| Glycosphingolipid biosynthesis - globo and isoglobo series | 1 |
| Histidine metabolism | 1 |
| Inositol phosphate metabolism | 1 |
| Longevity regulating pathway - worm | 1 |
| Lysosome | 1 |
| Monobactam biosynthesis | 1 |
| Nitrogen metabolism | 1 |
| NOD-like receptor signaling pathway | 1 |
| Oxidative phosphorylation | 1 |
| Pentose and glucuronate interconversions | 1 |
| Phenylalanine, tyrosine and tryptophan biosynthesis | 1 |
| Phosphotransferase system (PTS) | 1 |
| Porphyrin and chlorophyll metabolism | 1 |
| Protein digestion and absorption | 1 |
| Purine metabolism | 1 |
| Pyrimidine metabolism | 1 |
| Ribosome | 1 |
| Steroid hormone biosynthesis | 1 |
| Terpenoid backbone biosynthesis | 1 |
| Tropane, piperidine and pyridine alkaloid biosynthesis | 1 |

**Supplementary Table 2.** KEGG Pathways found to be significantly enriched and depleted in gut metagenomes of healthy individuals and liver cirrhosis patients. Functional analysis is based on codon usage in the subset of genes predicted to be highly expressed, as quantified by E (Karlin and Mrazek, 2000). Details in the text.

| enriched in healthy samples | number of samples |
|---|---|
| Fatty acid biosynthesis | 4 |
| Flagellar assembly | 4 |
| Galactose metabolism | 4 |
| p53 signaling pathway | 4 |
| Ribosome | 4 |
| ABC transporters | 3 |
| Adipocytokine signaling pathway | 3 |
| Arginine biosynthesis | 3 |
| Bacterial chemotaxis | 3 |
| Base excision repair | 3 |
| Chloroalkane and chloroalkene degradation | 3 |
| Fatty acid metabolism | 3 |
| Ferroptosis | 3 |
| HIF-1 signaling pathway | 3 |
| Neomycin, kanamycin and gentamicin biosynthesis | 3 |
| Ubiquitin mediated proteolysis | 3 |
| African trypanosomiasis | 2 |
| Aldosterone synthesis and secretion | 2 |
| Aminoacyl-tRNA biosynthesis | 2 |
| Autophagy - yeast | 2 |
| Biofilm formation - Vibrio cholerae | 2 |
| Biosynthesis of siderophore group nonribosomal peptides | 2 |
| C5-Branched dibasic acid metabolism | 2 |
| Caprolactam degradation | 2 |
| Cell cycle | 2 |
| Cell cycle - yeast | 2 |
| Chagas disease (American trypanosomiasis) | 2 |
| D-Alanine metabolism | 2 |
| Dioxin degradation | 2 |
| Fatty acid degradation | 2 |
| Glutathione metabolism | 2 |
| Glycosylphosphatidylinositol (GPI)-anchor biosynthesis | 2 |
| Insulin signaling pathway | 2 |
| Legionellosis | 2 |
| Lipoic acid metabolism | 2 |
| N-Glycan biosynthesis | 2 |
| Pentose and glucuronate interconversions | 2 |
| Phenazine biosynthesis | 2 |
| Phenylalanine metabolism | 2 |
| Phosphonate and phosphinate metabolism | 2 |
| Photosynthesis | 2 |
| Platinum drug resistance | 2 |
| Porphyrin and chlorophyll metabolism | 2 |
| PPAR signaling pathway | 2 |

| enriched in healthy samples | number of samples |
|---|---|
| Primary bile acid biosynthesis | 2 |
| Protein digestion and absorption | 2 |
| Secondary bile acid biosynthesis | 2 |
| Sphingolipid metabolism | 2 |
| Valine, leucine and isoleucine biosynthesis | 2 |
| Alanine, aspartate and glutamate metabolism | 1 |
| Aminobenzoate degradation | 1 |
| AMPK signaling pathway | 1 |
| Antifolate resistance | 1 |
| B cell receptor signaling pathway | 1 |
| Biofilm formation - Escherichia coli | 1 |
| Carbapenem biosynthesis | 1 |
| Carbon metabolism | 1 |
| Cationic antimicrobial peptide (CAMP) resistance | 1 |
| Cell cycle - Caulobacter | 1 |
| Central carbon metabolism in cancer | 1 |
| D-Glutamine and D-glutamate metabolism | 1 |
| Drug metabolism - other enzymes | 1 |
| Endocytosis | 1 |
| Fanconi anemia pathway | 1 |
| Flavone and flavonol biosynthesis | 1 |
| FoxO signaling pathway | 1 |
| Fructose and mannose metabolism | 1 |
| Glucosinolate biosynthesis | 1 |
| Glycerophospholipid metabolism | 1 |
| Glycolysis / Gluconeogenesis | 1 |
| Homologous recombination | 1 |
| Huntington disease | 1 |
| Longevity regulating pathway - multiple species | 1 |
| Meiosis - yeast | 1 |
| Mitophagy - yeast | 1 |
| Naphthalene degradation | 1 |
| NF-kappa B signaling pathway | 1 |
| Nitrogen metabolism | 1 |
| Novobiocin biosynthesis | 1 |
| One carbon pool by folate | 1 |
| Other glycan degradation | 1 |
| Oxidative phosphorylation | 1 |
| Phenylalanine, tyrosine and tryptophan biosynthesis | 1 |
| Phosphotransferase system (PTS) | 1 |
| Plant-pathogen interaction | 1 |
| Prolactin signaling pathway | 1 |
| Proximal tubule bicarbonate reclamation | 1 |
| Pyrimidine metabolism | 1 |
| Pyruvate metabolism | 1 |
| Quorum sensing | 1 |
| Salmonella infection | 1 |
| Starch and sucrose metabolism | 1 |
| Steroid hormone biosynthesis | 1 |
| Sulfur metabolism | 1 |

| enriched in healthy samples | number of samples |
|---|---|
| Synthesis and degradation of ketone bodies | 1 |
| T cell receptor signaling pathway | 1 |
| Terpenoid backbone biosynthesis | 1 |
| Thermogenesis | 1 |
| Tryptophan metabolism | 1 |
| Valine, leucine and isoleucine degradation | 1 |
| **depleted in healthy samples** | **number of samples** |
| Cyanoamino acid metabolism | 3 |
| Galactose metabolism | 3 |
| Phenylpropanoid biosynthesis | 3 |
| Photosynthesis | 3 |
| Quorum sensing | 3 |
| ABC transporters | 2 |
| Aminoacyl-tRNA biosynthesis | 2 |
| beta-Alanine metabolism | 2 |
| Fluid shear stress and atherosclerosis | 2 |
| Lipopolysaccharide biosynthesis | 2 |
| Lysine biosynthesis | 2 |
| Nicotinate and nicotinamide metabolism | 2 |
| Pyrimidine metabolism | 2 |
| Ribosome | 2 |
| Selenocompound metabolism | 2 |
| Sulfur relay system | 2 |
| Terpenoid backbone biosynthesis | 2 |
| Valine, leucine and isoleucine biosynthesis | 2 |
| Amino sugar and nucleotide sugar metabolism | 1 |
| Bacterial secretion system | 1 |
| Benzoate degradation | 1 |
| Biofilm formation - Vibrio cholerae | 1 |
| Central carbon metabolism in cancer | 1 |
| Glucagon signaling pathway | 1 |
| Glutathione metabolism | 1 |
| Glycerolipid metabolism | 1 |
| Glycine, serine and threonine metabolism | 1 |
| Glyoxylate and dicarboxylate metabolism | 1 |
| Neomycin, kanamycin and gentamicin biosynthesis | 1 |
| Nitrotoluene degradation | 1 |
| Oxidative phosphorylation | 1 |
| Pantothenate and CoA biosynthesis | 1 |
| Pentose phosphate pathway | 1 |
| Peroxisome | 1 |
| Pertussis | 1 |
| Plant-pathogen interaction | 1 |
| Porphyrin and chlorophyll metabolism | 1 |
| Propanoate metabolism | 1 |
| Protein export | 1 |
| Sphingolipid metabolism | 1 |
| Valine, leucine and isoleucine degradation | 1 |

| enriched in diseased samples | number of samples |
|---|---|
| Cationic antimicrobial peptide (CAMP) resistance | 4 |
| Homologous recombination | 4 |
| Naphthalene degradation | 4 |
| Ribosome | 4 |
| Staphylococcus aureus infection | 4 |
| ABC transporters | 3 |
| Ascorbate and aldarate metabolism | 3 |
| DNA replication | 3 |
| Flagellar assembly | 3 |
| Pentose and glucuronate interconversions | 3 |
| Riboflavin metabolism | 3 |
| RNA polymerase | 3 |
| Vitamin B6 metabolism | 3 |
| Acarbose and validamycin biosynthesis | 2 |
| Arginine and proline metabolism | 2 |
| Bacterial secretion system | 2 |
| beta-Alanine metabolism | 2 |
| Carbon fixation pathways in prokaryotes | 2 |
| Carbon metabolism | 2 |
| Citrate cycle (TCA cycle) | 2 |
| D-Glutamine and D-glutamate metabolism | 2 |
| Epithelial cell signaling in Helicobacter pylori infection | 2 |
| Fanconi anemia pathway | 2 |
| Fatty acid degradation | 2 |
| Flavone and flavonol biosynthesis | 2 |
| Fructose and mannose metabolism | 2 |
| Glucagon signaling pathway | 2 |
| Glycine, serine and threonine metabolism | 2 |
| Isoquinoline alkaloid biosynthesis | 2 |
| Lysine degradation | 2 |
| Nitrotoluene degradation | 2 |
| Pertussis | 2 |
| Phosphotransferase system (PTS) | 2 |
| Propanoate metabolism | 2 |
| Proteoglycans in cancer | 2 |
| Retinol metabolism | 2 |
| Ribosome biogenesis in eukaryotes | 2 |
| Spliceosome | 2 |
| Sulfur metabolism | 2 |
| Taste transduction | 2 |
| Taurine and hypotaurine metabolism | 2 |
| Thiamine metabolism | 2 |
| Toll and Imd signaling pathway | 2 |
| Adipocytokine signaling pathway | 1 |

| enriched in diseased samples | number of samples |
|---|---|
| Alcoholism | 1 |
| Aminoacyl-tRNA biosynthesis | 1 |
| Aminobenzoate degradation | 1 |
| Apelin signaling pathway | 1 |
| Biosynthesis of ansamycins | 1 |
| Biosynthesis of antibiotics | 1 |
| Biosynthesis of secondary metabolites | 1 |
| Biosynthesis of unsaturated fatty acids | 1 |
| Butanoate metabolism | 1 |
| Calcium signaling pathway | 1 |
| Cell cycle - Caulobacter | 1 |
| cGMP-PKG signaling pathway | 1 |
| Chagas disease (American trypanosomiasis) | 1 |
| Chloroalkane and chloroalkene degradation | 1 |
| Chlorocyclohexane and chlorobenzene degradation | 1 |
| D-Arginine and D-ornithine metabolism | 1 |
| Fluid shear stress and atherosclerosis | 1 |
| Folate biosynthesis | 1 |
| FoxO signaling pathway | 1 |
| Glycosaminoglycan degradation | 1 |
| Human papillomavirus infection | 1 |
| Insect hormone biosynthesis | 1 |
| Lipoic acid metabolism | 1 |
| Lipopolysaccharide biosynthesis | 1 |
| Longevity regulating pathway | 1 |
| Longevity regulating pathway - multiple species | 1 |
| Lysine biosynthesis | 1 |
| MAPK signaling pathway - fly | 1 |
| MAPK signaling pathway - plant | 1 |
| Metabolism of xenobiotics by cytochrome P450 | 1 |
| Microbial metabolism in diverse environments | 1 |
| Mismatch repair | 1 |
| Nitrogen metabolism | 1 |
| Nucleotide excision repair | 1 |
| Pancreatic secretion | 1 |
| Phosphonate and phosphinate metabolism | 1 |
| Primary bile acid biosynthesis | 1 |
| Purine metabolism | 1 |
| Pyrimidine metabolism | 1 |
| Pyruvate metabolism | 1 |
| Ras signaling pathway | 1 |
| Salmonella infection | 1 |
| Secondary bile acid biosynthesis | 1 |
| Starch and sucrose metabolism | 1 |
| Styrene degradation | 1 |
| Terpenoid backbone biosynthesis | 1 |
| Zeatin biosynthesis | 1 |

| depleted in diseased samples | number of samples |
|---|---|
| Thiamine metabolism | 4 |
| Valine, leucine and isoleucine biosynthesis | 4 |
| Alanine, aspartate and glutamate metabolism | 2 |
| Arginine and proline metabolism | 2 |
| Arginine biosynthesis | 2 |
| Butanoate metabolism | 2 |
| Degradation of aromatic compounds | 2 |
| Fructose and mannose metabolism | 2 |
| Lipopolysaccharide biosynthesis | 2 |
| Oxidative phosphorylation | 2 |
| Propanoate metabolism | 2 |
| Protein digestion and absorption | 2 |
| Two-component system | 2 |
| Ubiquinone and other terpenoid-quinone biosynthesis | 2 |
| 2-Oxocarboxylic acid metabolism | 1 |
| Amino sugar and nucleotide sugar metabolism | 1 |
| Ascorbate and aldarate metabolism | 1 |
| Base excision repair | 1 |
| Biofilm formation - Pseudomonas aeruginosa | 1 |
| Biosynthesis of amino acids | 1 |
| Biosynthesis of ansamycins | 1 |
| Biotin metabolism | 1 |
| C5-Branched dibasic acid metabolism | 1 |
| Central carbon metabolism in cancer | 1 |
| Folate biosynthesis | 1 |
| Glutamatergic synapse | 1 |
| Glycine, serine and threonine metabolism | 1 |
| Longevity regulating pathway - worm | 1 |
| Lysine degradation | 1 |
| Microbial metabolism in diverse environments | 1 |
| Necroptosis | 1 |
| Peroxisome | 1 |
| Phosphatidylinositol signaling system | 1 |
| RNA polymerase | 1 |

**Supplementary Table 3.** KEGG Pathways found to be significantly enriched and depleted in gut metagenomes of healthy individuals and liver cirrhosis patients. Functional analysis is based on codon usage in the subset of genes predicted to be highly expressed, as quantified by MELP (Supekand Vlahovicek, 2005). Details in the text.

| enriched in healthy samples | number of samples |
|---|---|
| ABC transporters | 4 |
| Pentose and glucuronate interconversions | 4 |
| Ribosome | 4 |
| Cationic antimicrobial peptide (CAMP) resistance | 3 |
| Flagellar assembly | 3 |
| Adipocytokine signaling pathway | 2 |
| Aminoacyl-tRNA biosynthesis | 2 |
| Arginine biosynthesis | 2 |
| Ascorbate and aldarate metabolism | 2 |
| Base excision repair | 2 |
| Chloroalkane and chloroalkene degradation | 2 |
| Fatty acid biosynthesis | 2 |
| Fatty acid degradation | 2 |
| Fatty acid metabolism | 2 |
| Ferroptosis | 2 |
| Flavone and flavonol biosynthesis | 2 |
| Fructose and mannose metabolism | 2 |
| Galactose metabolism | 2 |
| Glutathione metabolism | 2 |
| Legionellosis | 2 |
| Lipoic acid metabolism | 2 |
| Longevity regulating pathway - multiple species | 2 |
| Naphthalene degradation | 2 |
| Neomycin, kanamycin and gentamicin biosynthesis | 2 |
| p53 signaling pathway | 2 |
| Pyrimidine metabolism | 2 |
| Pyruvate metabolism | 2 |
| RNA polymerase | 2 |
| Staphylococcus aureus infection | 2 |
| Ubiquitin mediated proteolysis | 2 |
| Vitamin B6 metabolism | 2 |
| Acarbose and validamycin biosynthesis | 1 |
| African trypanosomiasis | 1 |
| Alanine, aspartate and glutamate metabolism | 1 |
| Aldosterone synthesis and secretion | 1 |
| Antifolate resistance | 1 |
| Arginine and proline metabolism | 1 |
| Autophagy - yeast | 1 |
| Bacterial chemotaxis | 1 |
| Bacterial secretion system | 1 |
| beta-Alanine metabolism | 1 |
| Biofilm formation - Vibrio cholerae | 1 |
| Biosynthesis of antibiotics | 1 |
| Biosynthesis of secondary metabolites | 1 |

| enriched in healthy samples | number of samples |
|---|---|
| Biosynthesis of siderophore group nonribosomal peptides | 1 |
| Biosynthesis of unsaturated fatty acids | 1 |
| Butanoate metabolism | 1 |
| C5-Branched dibasic acid metabolism | 1 |
| Calcium signaling pathway | 1 |
| Caprolactam degradation | 1 |
| Carbapenem biosynthesis | 1 |
| Carbon fixation pathways in prokaryotes | 1 |
| Carbon metabolism | 1 |
| Cell cycle | 1 |
| Cell cycle - Caulobacter | 1 |
| Cell cycle - yeast | 1 |
| cGMP-PKG signaling pathway | 1 |
| Chagas disease (American trypanosomiasis) | 1 |
| Chlorocyclohexane and chlorobenzene degradation | 1 |
| Citrate cycle (TCA cycle) | 1 |
| D-Alanine metabolism | 1 |
| D-Arginine and D-ornithine metabolism | 1 |
| D-Glutamine and D-glutamate metabolism | 1 |
| Dioxin degradation | 1 |
| DNA replication | 1 |
| Drug metabolism - other enzymes | 1 |
| Epithelial cell signaling in Helicobacter pylori infection | 1 |
| Fanconi anemia pathway | 1 |
| FoxO signaling pathway | 1 |
| Glucagon signaling pathway | 1 |
| Glucosinolate biosynthesis | 1 |
| Glycerophospholipid metabolism | 1 |
| Glycine, serine and threonine metabolism | 1 |
| Glycosylphosphatidylinositol (GPI)-anchor biosynthesis | 1 |
| HIF-1 signaling pathway | 1 |
| Homologous recombination | 1 |
| Huntington disease | 1 |
| Insect hormone biosynthesis | 1 |
| Isoquinoline alkaloid biosynthesis | 1 |
| Lysine degradation | 1 |
| Meiosis - yeast | 1 |
| Metabolism of xenobiotics by cytochrome P450 | 1 |
| N-Glycan biosynthesis | 1 |
| Nitrogen metabolism | 1 |
| Nitrotoluene degradation | 1 |
| Novobiocin biosynthesis | 1 |
| Nucleotide excision repair | 1 |
| Other glycan degradation | 1 |
| Oxidative phosphorylation | 1 |
| Pancreatic secretion | 1 |
| Pertussis | 1 |
| Phenazine biosynthesis | 1 |
| Phenylalanine metabolism | 1 |
| Phenylalanine, tyrosine and tryptophan biosynthesis | 1 |

| enriched in healthy samples | number of samples |
|---|---|
| Phosphonate and phosphinate metabolism | 1 |
| Phosphotransferase system (PTS) | 1 |
| Photosynthesis | 1 |
| Plant-pathogen interaction | 1 |
| Platinum drug resistance | 1 |
| Porphyrin and chlorophyll metabolism | 1 |
| PPAR signaling pathway | 1 |
| Primary bile acid biosynthesis | 1 |
| Prolactin signaling pathway | 1 |
| Propanoate metabolism | 1 |
| Protein digestion and absorption | 1 |
| Proximal tubule bicarbonate reclamation | 1 |
| Purine metabolism | 1 |
| Retinol metabolism | 1 |
| Riboflavin metabolism | 1 |
| Ribosome biogenesis in eukaryotes | 1 |
| Secondary bile acid biosynthesis | 1 |
| Sphingolipid metabolism | 1 |
| Spliceosome | 1 |
| Starch and sucrose metabolism | 1 |
| Steroid hormone biosynthesis | 1 |
| Sulfur metabolism | 1 |
| Synthesis and degradation of ketone bodies | 1 |
| Taste transduction | 1 |
| Taurine and hypotaurine metabolism | 1 |
| Terpenoid backbone biosynthesis | 1 |
| Thermogenesis | 1 |
| Thiamine metabolism | 1 |
| Valine, leucine and isoleucine biosynthesis | 1 |
| Valine, leucine and isoleucine degradation | 1 |
| **depleted in healthy samples** | **number of samples** |
| Central carbon metabolism in cancer | 2 |
| Galactose metabolism | 2 |
| Lipopolysaccharide biosynthesis | 2 |
| Thiamine metabolism | 2 |
| Valine, leucine and isoleucine biosynthesis | 2 |
| ABC transporters | 1 |
| Alanine, aspartate and glutamate metabolism | 1 |
| Amino sugar and nucleotide sugar metabolism | 1 |
| Aminoacyl-tRNA biosynthesis | 1 |
| Arginine and proline metabolism | 1 |
| Arginine biosynthesis | 1 |
| Bacterial secretion system | 1 |
| beta-Alanine metabolism | 1 |
| Biofilm formation - Vibrio cholerae | 1 |
| Biosynthesis of ansamycins | 1 |
| Butanoate metabolism | 1 |
| C5-Branched dibasic acid metabolism | 1 |
| Cyanoamino acid metabolism | 1 |
| Degradation of aromatic compounds | 1 |
| Fluid shear stress and atherosclerosis | 1 |
| Folate biosynthesis | 1 |

| depleted in healthy samples | number of samples |
|---|---|
| Fructose and mannose metabolism | 1 |
| Glutathione metabolism | 1 |
| Glycerolipid metabolism | 1 |
| Glyoxylate and dicarboxylate metabolism | 1 |
| Lysine biosynthesis | 1 |
| Lysine degradation | 1 |
| Necroptosis | 1 |
| Neomycin, kanamycin and gentamicin biosynthesis | 1 |
| Nicotinate and nicotinamide metabolism | 1 |
| Nitrotoluene degradation | 1 |
| Oxidative phosphorylation | 1 |
| Pantothenate and CoA biosynthesis | 1 |
| Peroxisome | 1 |
| Pertussis | 1 |
| Phenylpropanoid biosynthesis | 1 |
| Phosphatidylinositol signaling system | 1 |
| Photosynthesis | 1 |
| Plant-pathogen interaction | 1 |
| Propanoate metabolism | 1 |
| Protein digestion and absorption | 1 |
| Protein export | 1 |
| Pyrimidine metabolism | 1 |
| Quorum sensing | 1 |
| Ribosome | 1 |
| RNA polymerase | 1 |
| Selenocompound metabolism | 1 |
| Sphingolipid metabolism | 1 |
| Sulfur relay system | 1 |
| Terpenoid backbone biosynthesis | 1 |
| Two-component system | 1 |
| Ubiquinone and other terpenoid-quinone biosynthesis | 1 |
| **enriched in diseased samples** | **number of samples** |
| Fatty acid metabolism | 4 |
| Isoquinoline alkaloid biosynthesis | 4 |
| Biotin metabolism | 3 |
| Chloroalkane and chloroalkene degradation | 3 |
| Fatty acid biosynthesis | 3 |
| Prodigiosin biosynthesis | 3 |
| Ribosome | 3 |
| Tryptophan metabolism | 3 |
| Valine, leucine and isoleucine degradation | 3 |
| 2-Oxocarboxylic acid metabolism | 2 |
| Aminoacyl-tRNA biosynthesis | 2 |
| AMPK signaling pathway | 2 |
| Arginine biosynthesis | 2 |
| Atrazine degradation | 2 |
| Biosynthesis of unsaturated fatty acids | 2 |
| C5-Branched dibasic acid metabolism | 2 |
| Carbapenem biosynthesis | 2 |
| Cationic antimicrobial peptide (CAMP) resistance | 2 |
| Degradation of aromatic compounds | 2 |
| Fatty acid degradation | 2 |

| enriched in diseased samples | number of samples |
|---|---|
| Glycerolipid metabolism | 2 |
| Glycosphingolipid biosynthesis - globo and isoglobo series | 2 |
| Glycosylphosphatidylinositol (GPI)-anchor biosynthesis | 2 |
| Homologous recombination | 2 |
| Insulin resistance | 2 |
| Longevity regulating pathway - multiple species | 2 |
| Lysine biosynthesis | 2 |
| Meiosis - yeast | 2 |
| Naphthalene degradation | 2 |
| Necroptosis | 2 |
| Nitrotoluene degradation | 2 |
| Nucleotide excision repair | 2 |
| Phenylpropanoid biosynthesis | 2 |
| Phosphotransferase system (PTS) | 2 |
| Platinum drug resistance | 2 |
| Porphyrin and chlorophyll metabolism | 2 |
| Proximal tubule bicarbonate reclamation | 2 |
| Pyrimidine metabolism | 2 |
| Pyruvate metabolism | 2 |
| Renin-angiotensin system | 2 |
| RNA degradation | 2 |
| RNA polymerase | 2 |
| Styrene degradation | 2 |
| Tropane, piperidine and pyridine alkaloid biosynthesis | 2 |
| Tyrosine metabolism | 2 |
| Alanine, aspartate and glutamate metabolism | 1 |
| Amyotrophic lateral sclerosis (ALS) | 1 |
| Arachidonic acid metabolism | 1 |
| Ascorbate and aldarate metabolism | 1 |
| Autophagy - yeast | 1 |
| Bacterial secretion system | 1 |
| Basal transcription factors | 1 |
| Benzoate degradation | 1 |
| Biofilm formation - Vibrio cholerae | 1 |
| Biosynthesis of ansamycins | 1 |
| Biosynthesis of vancomycin group antibiotics | 1 |
| Butanoate metabolism | 1 |
| Calcium signaling pathway | 1 |
| cAMP signaling pathway | 1 |
| Carbohydrate digestion and absorption | 1 |
| Carbon fixation in photosynthetic organisms | 1 |
| Carbon fixation pathways in prokaryotes | 1 |
| Cell cycle - yeast | 1 |
| Chlorocyclohexane and chlorobenzene degradation | 1 |
| Cyanoamino acid metabolism | 1 |
| Cytosolic DNA-sensing pathway | 1 |
| D-Alanine metabolism | 1 |
| D-Glutamine and D-glutamate metabolism | 1 |
| Dioxin degradation | 1 |
| DNA replication | 1 |
| Drug metabolism - other enzymes | 1 |
| Ether lipid metabolism | 1 |

| enriched in diseased samples | number of samples |
|---|---|
| Fanconi anemia pathway | 1 |
| Fc gamma R-mediated phagocytosis | 1 |
| Ferroptosis | 1 |
| Fluorobenzoate degradation | 1 |
| Fructose and mannose metabolism | 1 |
| Galactose metabolism | 1 |
| Gastric acid secretion | 1 |
| Glucagon signaling pathway | 1 |
| Glycerophospholipid metabolism | 1 |
| Glycosaminoglycan degradation | 1 |
| Histidine metabolism | 1 |
| Huntington disease | 1 |
| Insect hormone biosynthesis | 1 |
| Insulin signaling pathway | 1 |
| Legionellosis | 1 |
| Limonene and pinene degradation | 1 |
| Lipopolysaccharide biosynthesis | 1 |
| MAPK signaling pathway - plant | 1 |
| MAPK signaling pathway - yeast | 1 |
| Metabolic pathways | 1 |
| Metabolism of xenobiotics by cytochrome P450 | 1 |
| MicroRNAs in cancer | 1 |
| Mismatch repair | 1 |
| N-Glycan biosynthesis | 1 |
| Nitrogen metabolism | 1 |
| Nonribosomal peptide structures | 1 |
| Novobiocin biosynthesis | 1 |
| Oxytocin signaling pathway | 1 |
| Pantothenate and CoA biosynthesis | 1 |
| Peptidoglycan biosynthesis | 1 |
| Pertussis | 1 |
| Phenylalanine metabolism | 1 |
| Photosynthesis | 1 |
| Primary bile acid biosynthesis | 1 |
| Primary immunodeficiency | 1 |
| Propanoate metabolism | 1 |
| Protein digestion and absorption | 1 |
| Protein processing in endoplasmic reticulum | 1 |
| Purine metabolism | 1 |
| Quorum sensing | 1 |
| Riboflavin metabolism | 1 |
| Steroid hormone biosynthesis | 1 |
| Sulfur metabolism | 1 |
| Synthesis and degradation of ketone bodies | 1 |
| Terpenoid backbone biosynthesis | 1 |
| Thermogenesis | 1 |
| Thiamine metabolism | 1 |
| Thyroid hormone synthesis | 1 |
| Toluene degradation | 1 |
| Tuberculosis | 1 |
| Two-component system | 1 |
| Type II diabetes mellitus | 1 |

| enriched in diseased samples | number of samples |
|---|---|
| Ubiquitin mediated proteolysis | 1 |
| Viral carcinogenesis | 1 |
| Xylene degradation | 1 |
| **depleted in diseased samples** | **number of samples** |
| Bacterial secretion system | 3 |
| Citrate cycle (TCA cycle) | 3 |
| Phosphotransferase system (PTS) | 3 |
| Biofilm formation - Escherichia coli | 2 |
| Biofilm formation - Vibrio cholerae | 2 |
| C5-Branched dibasic acid metabolism | 2 |
| Carbon fixation pathways in prokaryotes | 2 |
| Cell cycle - Caulobacter | 2 |
| Glycine, serine and threonine metabolism | 2 |
| Methane metabolism | 2 |
| Oxidative phosphorylation | 2 |
| Peptidoglycan biosynthesis | 2 |
| Protein export | 2 |
| Sulfur metabolism | 2 |
| Terpenoid backbone biosynthesis | 2 |
| 2-Oxocarboxylic acid metabolism | 1 |
| Acarbose and validamycin biosynthesis | 1 |
| Alanine, aspartate and glutamate metabolism | 1 |
| Arginine biosynthesis | 1 |
| beta-Alanine metabolism | 1 |
| beta-Lactam resistance | 1 |
| Biosynthesis of amino acids | 1 |
| Biosynthesis of antibiotics | 1 |
| Butanoate metabolism | 1 |
| Carbon metabolism | 1 |
| Cyanoamino acid metabolism | 1 |
| Degradation of aromatic compounds | 1 |
| Fatty acid degradation | 1 |
| Fructose and mannose metabolism | 1 |
| Glutathione metabolism | 1 |
| Glycerolipid metabolism | 1 |
| Glycolysis / Gluconeogenesis | 1 |
| Insulin resistance | 1 |
| Lipopolysaccharide biosynthesis | 1 |
| Longevity regulating pathway - worm | 1 |
| Lysine biosynthesis | 1 |
| Lysosome | 1 |
| Microbial metabolism in diverse environments | 1 |
| Monobactam biosynthesis | 1 |
| Novobiocin biosynthesis | 1 |
| Other glycan degradation | 1 |
| Pantothenate and CoA biosynthesis | 1 |
| Phenylalanine metabolism | 1 |
| Phenylalanine, tyrosine and tryptophan biosynthesis | 1 |
| henylpropanoid biosynthesis | 1 |
| Photosynthesis | 1 |
| Polyketide sugar unit biosynthesis | 1 |
| Prodigiosin biosynthesis | 1 |

| depleted in diseased samples | number of samples |
|---|---|
| Propanoate metabolism | 1 |
| Quorum sensing | 1 |
| Riboflavin metabolism | 1 |
| Ribosome | 1 |
| RNA degradation | 1 |
| Starch and sucrose metabolism | 1 |
| Streptomycin biosynthesis | 1 |
| Thiamine metabolism | 1 |
| Vancomycin resistance | 1 |

**Supplementary Table 4.** KEGG Pathways found to be significantly enriched and depleted in gut metagenomes of healthy individuals and liver cirrhosis patients. Functional analysis is based on codon usage in the subset of genes predicted to be highly expressed, as quantified by FOP (Ikemura, 1981). Details in the the text.

| enriched in healthy samples | number of samples |
|---|---|
| Flagellar assembly | 6 |
| ABC transporters | 4 |
| Fatty acid metabolism | 4 |
| Ferroptosis | 4 |
| Ribosome | 4 |
| Adipocytokine signaling pathway | 3 |
| Bacterial chemotaxis | 3 |
| Butanoate metabolism | 3 |
| Cell cycle - Caulobacter | 3 |
| Fatty acid biosynthesis | 3 |
| Peroxisome | 3 |
| Photosynthesis | 3 |
| PPAR signaling pathway | 3 |
| Adrenergic signaling in cardiomyocytes | 2 |
| Alanine, aspartate and glutamate metabolism | 2 |
| beta-Lactam resistance | 2 |
| Biosynthesis of unsaturated fatty acids | 2 |
| C5-Branched dibasic acid metabolism | 2 |
| cAMP signaling pathway | 2 |
| Carbapenem biosynthesis | 2 |
| Cell cycle | 2 |
| Chagas disease (American trypanosomiasis) | 2 |
| Cytosolic DNA-sensing pathway | 2 |
| Glutamatergic synapse | 2 |
| Epstein-Barr virus infection | 2 |
| Fatty acid degradation | 2 |
| Fluorobenzoate degradation | 2 |
| Fructose and mannose metabolism | 2 |
| Glutamatergic synapse | 2 |
| Glycerophospholipid metabolism | 2 |
| Insulin signaling pathway | 2 |
| MAPK signaling pathway - yeast | 2 |
| Oxidative phosphorylation | 2 |
| Porphyrin and chlorophyll metabolism | 2 |
| Proteoglycans in cancer | 2 |
| Purine metabolism | 2 |
| Pyrimidine metabolism | 2 |
| Salivary secretion | 2 |
| Thermogenesis | 2 |
| Toluene degradation | 2 |
| Valine, leucine and isoleucine biosynthesis | 2 |
| African trypanosomiasis | 1 |
| AGE-RAGE signaling pathway in diabetic complications | 1 |
| Alcoholism | 1 |

| enriched in healthy samples | number of samples |
|---|---|
| Aldosterone synthesis and secretion | 1 |
| AMPK signaling pathway | 1 |
| Antifolate resistance | 1 |
| Antigen processing and presentation | 1 |
| Apelin signaling pathway | 1 |
| Arginine biosynthesis | 1 |
| Atrazine degradation | 1 |
| Bacterial secretion system | 1 |
| Basal transcription factors | 1 |
| beta-Alanine metabolism | 1 |
| Biofilm formation - Pseudomonas aeruginosa | 1 |
| Biosynthesis of ansamycins | 1 |
| Biotin metabolism | 1 |
| Calcium signaling pathway | 1 |
| Carbon fixation in photosynthetic organisms | 1 |
| Cell cycle - yeast | 1 |
| Cellular senescence | 1 |
| cGMP-PKG signaling pathway | 1 |
| Chloroalkane and chloroalkene degradation | 1 |
| Chlorocyclohexane and chlorobenzene degradation | 1 |
| Citrate cycle (TCA cycle) | 1 |
| D-Alanine metabolism | 1 |
| Epithelial cell signaling in Helicobacter pylori infection | 1 |
| Estrogen signaling pathway | 1 |
| Flavone and flavonol biosynthesis | 1 |
| Glucosinolate biosynthesis | 1 |
| Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate | 1 |
| Glycosaminoglycan degradation | 1 |
| Glycosphingolipid biosynthesis - ganglio series | 1 |
| Glycosphingolipid biosynthesis - globo and isoglobo series | 1 |
| Glyoxylate and dicarboxylate metabolism | 1 |
| IL-17 signaling pathway | 1 |
| Insect hormone biosynthesis | 1 |
| Insulin resistance | 1 |
| Legionellosis | 1 |
| Limonene and pinene degradation | 1 |
| Longevity regulating pathway - multiple species | 1 |
| Lysine biosynthesis | 1 |
| Lysine degradation | 1 |
| Methane metabolism | 1 |
| MicroRNAs in cancer | 1 |
| Mitophagy - yeast | 1 |
| Necroptosis | 1 |
| Nitrogen metabolism | 1 |
| Novobiocin biosynthesis | 1 |
| Nucleotide excision repair | 1 |
| One carbon pool by folate | 1 |
| p53 signaling pathway | 1 |
| Pancreatic secretion | 1 |
| Pathogenic Escherichia coli infection | 1 |

| enriched in healthy samples | number of samples |
|---|---|
| Pathways in cancer | 1 |
| Pentose and glucuronate interconversions | 1 |
| Pentose phosphate pathway | 1 |
| Phenazine biosynthesis | 1 |
| Phenylalanine, tyrosine and tryptophan biosynthesis | 1 |
| Phosphonate and phosphinate metabolism | 1 |
| Polyketide sugar unit biosynthesis | 1 |
| Primary bile acid biosynthesis | 1 |
| Prodigiosin biosynthesis | 1 |
| Prolactin signaling pathway | 1 |
| Propanoate metabolism | 1 |
| Prostate cancer | 1 |
| Protein digestion and absorption | 1 |
| Protein export | 1 |
| Proximal tubule bicarbonate reclamation | 1 |
| Regulation of actin cytoskeleton | 1 |
| Ribosome biogenesis in eukaryotes | 1 |
| RNA degradation | 1 |
| RNA polymerase | 1 |
| Salmonella infection | 1 |
| Secondary bile acid biosynthesis | 1 |
| Spliceosome | 1 |
| Staphylococcus aureus infection | 1 |
| Streptomycin biosynthesis | 1 |
| Sulfur metabolism | 1 |
| Sulfur relay system | 1 |
| Terpenoid backbone biosynthesis | 1 |
| Th17 cell differentiation | 1 |
| Thiamine metabolism | 1 |
| Toll and Imd signaling pathway | 1 |
| Tropane, piperidine and pyridine alkaloid biosynthesis | 1 |
| Tyrosine metabolism | 1 |
| Ubiquitin mediated proteolysis | 1 |
| Valine, leucine and isoleucine degradation | 1 |
| Vancomycin resistance | 1 |
| Various types of N-glycan biosynthesis | 1 |
| **depleted in healthy samples** | **number of samples** |
| 2-Oxocarboxylic acid metabolism | 2 |
| ABC transporters | 2 |
| Amino sugar and nucleotide sugar metabolism | 2 |
| Antifolate resistance | 2 |
| Arginine and proline metabolism | 2 |
| Base excision repair | 2 |
| Carbon fixation pathways in prokaryotes | 2 |
| Histidine metabolism | 2 |
| Lipopolysaccharide biosynthesis | 2 |
| Other glycan degradation | 2 |
| Pentose and glucuronate interconversions | 2 |
| Pertussis | 2 |
| Propanoate metabolism | 2 |
| Pyrimidine metabolism | 2 |
| Ribosome | 2 |

| depleted in healthy samples | number of samples |
|---|---|
| RNA degradation | 2 |
| Starch and sucrose metabolism | 2 |
| Acarbose and validamycin biosynthesis | 1 |
| Biofilm formation - Escherichia coli | 1 |
| Biofilm formation - Pseudomonas aeruginosa | 1 |
| Biotin metabolism | 1 |
| Butanoate metabolism | 1 |
| Carbon fixation in photosynthetic organisms | 1 |
| D-Glutamine and D-glutamate metabolism | 1 |
| Drug metabolism - other enzymes | 1 |
| Fatty acid degradation | 1 |
| Fatty acid metabolism | 1 |
| Flagellar assembly | 1 |
| Fructose and mannose metabolism | 1 |
| Glycerophospholipid metabolism | 1 |
| Glycosaminoglycan degradation | 1 |
| Glycosphingolipid biosynthesis - ganglio series | 1 |
| MicroRNAs in cancer | 1 |
| Monobactam biosynthesis | 1 |
| Neomycin, kanamycin and gentamicin biosynthesis | 1 |
| One carbon pool by folate | 1 |
| Pantothenate and CoA biosynthesis | 1 |
| Pentose phosphate pathway | 1 |
| Phenylalanine, tyrosine and tryptophan biosynthesis | 1 |
| Phosphonate and phosphinate metabolism | 1 |
| Polyketide sugar unit biosynthesis | 1 |
| Porphyrin and chlorophyll metabolism | 1 |
| Protein export | 1 |
| Pyruvate metabolism | 1 |
| RNA polymerase | 1 |
| Streptomycin biosynthesis | 1 |
| Sulfur metabolism | 1 |
| Tropane, piperidine and pyridine alkaloid biosynthesis | 1 |
| Tyrosine metabolism | 1 |
| Ubiquinone and other terpenoid-quinone biosynthesis | 1 |
| Valine, leucine and isoleucine degradation | 1 |
| Various types of N-glycan biosynthesis | 1 |
| **enriched in diseased samples** | **number of samples** |
| Fructose and mannose metabolism | 3 |
| Glycerolipid metabolism | 3 |
| Limonene and pinene degradation | 3 |
| Mismatch repair | 3 |
| Phosphatidylinositol signaling system | 3 |
| Phosphotransferase system (PTS) | 3 |
| Porphyrin and chlorophyll metabolism | 3 |
| Primary immunodeficiency | 3 |
| Propanoate metabolism | 3 |
| Valine, leucine and isoleucine degradation | 3 |
| AGE-RAGE signaling pathway in diabetic complications | 2 |
| Alcoholism | 2 |
| Benzoate degradation | 2 |
| Bile secretion | 2 |

| enriched in diseased samples | number of samples |
|---|---|
| Biosynthesis of siderophore group nonribosomal peptides | 2 |
| Carbapenem biosynthesis | 2 |
| Carbohydrate digestion and absorption | 2 |
| Carbon fixation pathways in prokaryotes | 2 |
| D-Alanine metabolism | 2 |
| Degradation of aromatic compounds | 2 |
| Flagellar assembly | 2 |
| Glyoxylate and dicarboxylate metabolism | 2 |
| Insect hormone biosynthesis | 2 |
| Legionellosis | 2 |
| MAPK signaling pathway - plant | 2 |
| Methane metabolism | 2 |
| MicroRNAs in cancer | 2 |
| Monobactam biosynthesis | 2 |
| Naphthalene degradation | 2 |
| Oxidative phosphorylation | 2 |
| Phenazine biosynthesis | 2 |
| Phenylalanine metabolism | 2 |
| Primary bile acid biosynthesis | 2 |
| Progesterone-mediated oocyte maturation | 2 |
| Protein export | 2 |
| Quorum sensing | 2 |
| Ribosome | 2 |
| Tryptophan metabolism | 2 |
| Type I diabetes mellitus | 2 |
| Valine, leucine and isoleucine biosynthesis | 2 |
| 2-Oxocarboxylic acid metabolism | 1 |
| Adrenergic signaling in cardiomyocytes | 1 |
| Amphetamine addiction | 1 |
| Amyotrophic lateral sclerosis (ALS) | 1 |
| Antifolate resistance | 1 |
| Apoptosis | 1 |
| Arginine biosynthesis | 1 |
| Ascorbate and aldarate metabolism | 1 |
| Basal transcription factors | 1 |
| Base excision repair | 1 |
| beta-Alanine metabolism | 1 |
| Betalain biosynthesis | 1 |
| Biosynthesis of vancomycin group antibiotics | 1 |
| Butanoate metabolism | 1 |
| C5-Branched dibasic acid metabolism | 1 |
| cAMP signaling pathway | 1 |
| Caprolactam degradation | 1 |
| Carbon metabolism | 1 |
| Cardiac muscle contraction | 1 |
| Cell cycle | 1 |
| Cell cycle - yeast | 1 |
| Cellular senescence | 1 |
| Chloroalkane and chloroalkene degradation | 1 |
| Chlorocyclohexane and chlorobenzene degradation | 1 |
| Choline metabolism in cancer | 1 |
| Cocaine addiction | 1 |

| enriched in diseased samples | number of samples |
|---|---|
| Complement and coagulation cascades | 1 |
| Cytosolic DNA-sensing pathway | 1 |
| D-Glutamine and D-glutamate metabolism | 1 |
| Dioxin degradation | 1 |
| Dopaminergic synapse | 1 |
| Drug metabolism - cytochrome P450 | 1 |
| ECM-receptor interaction | 1 |
| Endocytosis | 1 |
| Epstein-Barr virus infection | 1 |
| Ether lipid metabolism | 1 |
| Fc gamma R-mediated phagocytosis | 1 |
| Ferroptosis | 1 |
| Flavone and flavonol biosynthesis | 1 |
| Fluid shear stress and atherosclerosis | 1 |
| Gastric acid secretion | 1 |
| Geraniol degradation | 1 |
| Glucosinolate biosynthesis | 1 |
| Glycosphingolipid biosynthesis - globo and isoglobo series | 1 |
| HIF-1 signaling pathway | 1 |
| Hippo signaling pathway | 1 |
| Huntington disease | 1 |
| Indole alkaloid biosynthesis | 1 |
| Inositol phosphate metabolism | 1 |
| Lipopolysaccharide biosynthesis | 1 |
| Longevity regulating pathway - worm | 1 |
| Lysine degradation | 1 |
| Lysosome | 1 |
| MAPK signaling pathway - yeast | 1 |
| Metabolism of xenobiotics by cytochrome P450 | 1 |
| Mitophagy - yeast | 1 |
| Neomycin, kanamycin and gentamicin biosynthesis | 1 |
| Nonribosomal peptide structures | 1 |
| One carbon pool by folate | 1 |
| p53 signaling pathway | 1 |
| Pantothenate and CoA biosynthesis | 1 |
| Penicillin and cephalosporin biosynthesis | 1 |
| Pentose and glucuronate interconversions | 1 |
| Phenylalanine, tyrosine and tryptophan biosynthesis | 1 |
| Phospholipase D signaling pathway | 1 |
| Platelet activation | 1 |
| Platinum drug resistance | 1 |
| Polyketide sugar unit biosynthesis | 1 |
| Proteoglycans in cancer | 1 |
| Proximal tubule bicarbonate reclamation | 1 |
| Purine metabolism | 1 |
| Pyruvate metabolism | 1 |
| Regulation of actin cytoskeleton | 1 |
| Relaxin signaling pathway | 1 |
| Retinol metabolism | 1 |
| Ribosome biogenesis in eukaryotes | 1 |
| RIG-I-like receptor signaling pathway | 1 |
| RNA polymerase | 1 |

| enriched in diseased samples | number of samples |
|---|---|
| RNA transport | 1 |
| Salivary secretion | 1 |
| Salmonella infection | 1 |
| Secondary bile acid biosynthesis | 1 |
| Selenocompound metabolism | 1 |
| Serotonergic synapse | 1 |
| Spliceosome | 1 |
| Staphylococcus aureus infection | 1 |
| Steroid hormone biosynthesis | 1 |
| Streptomycin biosynthesis | 1 |
| Sulfur metabolism | 1 |
| Taurine and hypotaurine metabolism | 1 |
| Terpenoid backbone biosynthesis | 1 |
| Thiamine metabolism | 1 |
| Thyroid hormone signaling pathway | 1 |
| Toluene degradation | 1 |
| Tuberculosis | 1 |
| Two-component system | 1 |
| Tyrosine metabolism | 1 |
| Ubiquinone and other terpenoid-quinone biosynthesis | 1 |
| Ubiquitin mediated proteolysis | 1 |
| Zeatin biosynthesis | 1 |
| **depleted in diseased samples** | **number of samples** |
| Photosynthesis | 3 |
| Ribosome | 3 |
| Citrate cycle (TCA cycle) | 2 |
| Fluid shear stress and atherosclerosis | 2 |
| Galactose metabolism | 2 |
| Lysosome | 2 |
| Nitrogen metabolism | 2 |
| Polyketide sugar unit biosynthesis | 2 |
| Pyrimidine metabolism | 2 |
| Selenocompound metabolism | 2 |
| Sphingolipid metabolism | 2 |
| Streptomycin biosynthesis | 2 |
| 2-Oxocarboxylic acid metabolism | 1 |
| Antifolate resistance | 1 |
| Arginine and proline metabolism | 1 |
| Arginine biosynthesis | 1 |
| Bacterial chemotaxis | 1 |
| Benzoate degradation | 1 |
| Biofilm formation - Escherichia coli | 1 |
| Biofilm formation - Vibrio cholerae | 1 |
| Biosynthesis of amino acids | 1 |
| Biosynthesis of ansamycins | 1 |
| Carbon fixation pathways in prokaryotes | 1 |
| Cyanoamino acid metabolism | 1 |
| Degradation of aromatic compounds | 1 |
| DNA replication | 1 |
| Drug metabolism - cytochrome P450 | 1 |
| Drug metabolism - other enzymes | 1 |
| Epithelial cell signaling in Helicobacter pylori infection | 1 |

| depleted in diseased samples | number of samples |
|---|---|
| Fatty acid biosynthesis | 1 |
| Fatty acid degradation | 1 |
| Fatty acid metabolism | 1 |
| Flagellar assembly | 1 |
| GABAergic synapse | 1 |
| Glycerolipid metabolism | 1 |
| Glyoxylate and dicarboxylate metabolism | 1 |
| Histidine metabolism | 1 |
| Homologous recombination | 1 |
| Inositol phosphate metabolism | 1 |
| Insulin resistance | 1 |
| Lipopolysaccharide biosynthesis | 1 |
| Methane metabolism | 1 |
| Mismatch repair | 1 |
| Neomycin, kanamycin and gentamicin biosynthesis | 1 |
| NOD-like receptor signaling pathway | 1 |
| One carbon pool by folate | 1 |
| Other glycan degradation | 1 |
| Oxidative phosphorylation | 1 |
| Pentose phosphate pathway | 1 |
| Peptidoglycan biosynthesis | 1 |
| Phosphotransferase system (PTS) | 1 |
| Protein digestion and absorption | 1 |
| Purine metabolism | 1 |
| Pyruvate metabolism | 1 |
| Quorum sensing | 1 |
| RNA polymerase | 1 |
| Sulfur metabolism | 1 |
| Tropane, piperidine and pyridine alkaloid biosynthesis | 1 |
| Tryptophan metabolism | 1 |