De novo evolution of genetic function from random sequences

Dissertation

in fulfilment of the requirements for the degree of

Doctor rerum naturalium (Dr. rer. Nat.)

of the faculty of Mathematics and Natural Sciences

at Kiel University

Submitted by

Rossy Johana Fajardo Castro

Department of Evolutionary Genetics

Max Planck Institute for Evolutionary Biology

Plön, December 2021

First Examiner: Prof. Dr. Diethard Tautz

Second Examiner: Prof. Dr. Dr. Thomas Bosch

Date of the oral examination: February 4th, 2022

To Rosa Castro Salcedo, my inspiration to be better every day.

Abstract

The study of *de novo* gene birth has opened the doors for evolutionary biologists to approach old questions about the origin of innovation in different ways and from new perspectives. The knowledge that sequences that have not previously exposed to selection may become genes with essential roles in different organisms has driven research to understand the requirements for newly expressed sequences to become genes. There are still many questions that we have just begun to answer about what makes a functional gene, how likely it is for a non-coding sequence to become one, how common is this phenomenon in nature, and how do novel genes become essential for an organism. In this thesis, I present three projects that aim to explore some of these questions.

In chapter 1 I revisit a study published in 2017 where the authors constructed a library of random sequences in *E. coli*, in order to quantify the likelihood that a random sequence expressed in a cell could provide a fitness advantage for it. I was able to successfully replicate the results of this study starting with a diluted sample of the library, and I designed a new analysis pipeline that allowed me to examine the behaviours of different sequences at much greater depth. I confirmed in these analyses that *E. coli* cells are very tolerant to the expression of random DNA sequences, and found that length—but not intrinsic disorder, GC content or aggregation probability—is a determinant factor of whether a sequence has an apparent positive effect on the fitness of the cells. Chapter 2 presents an attempt to recreate the random sequence experiment in a eukaryotic cell line. Its aim was to test whether a similar proportion of random sequences are well tolerated by the eukaryotic cells despite the different cellular complexity level. In general, the results indicate that eukaryotic cells are at least as tolerant to the expression of random sequences as the bacteria. In contrast with the results in bacteria, however, no specific feature of the sequences correlates with its tolerability. Finally, in chapter 3 I expressed three putative *de novo* genes identified in mouse in a human continuous cell line. The selected genes are young, taxonomically restricted to species of the genre *Mus*, and have transcription and proteomic evidence. Interestingly, expression of the novel genes had no major effect on the transcriptome of the cells, implying that they would also be tolerated in a human cellular background.

The combined results presented in this thesis add to the mounting evidence that cells are much more tolerant to the expression of new sequences than previously thought. This insight generates new questions about the birth of genes that should be explored in the future.

Zusammenfassung[1]

Die Untersuchung der *de novo* Evolution von Genen hat der Evolutionsbiologie neue Möglichkeiten eröffnet, um klassische Fragen über den Ursprung biologischer Innovation neu anzugehen. Sequenzen, die zuvor keiner Selektion ausgesetzt waren, können möglicherweise wesentlichen neuen Genen werden. Diese Erkenntnis erlaubt es jetzt die Voraussetzungen dafür zu verstehen, wie dieser Prozess abläuft. Es gibt dabei noch viele Fragen, deren Beantwortung jetzt erst beginnt: Was macht ein funktionelles Gen aus, wie wahrscheinlich ist es, dass eine nicht kodierende Sequenz zu einem Gen wird, wie häufig ist dieses Phänomen in der Natur, und wie werden neuartige Gene für einen Organismus essenziell. In dieser Arbeit stelle ich drei Projekte vor, die darauf abzielen, einige dieser Fragen zu untersuchen.

In Kapitel 1 greife ich eine 2017 veröffentlichte Studie auf, in der die Autoren eine Bibliothek mit zufälligen Sequenzen in *E. coli* konstruierten, um die Wahrscheinlichkeit zu bestimmen, dass eine zufällige Sequenz, die in einer Zelle exprimiert wird, dieser einen Fitnessvorteil verschaffen könnte. Ich konnte die Ergebnisse dieser Studie erfolgreich replizieren, indem ich mit einer verdünnten Probe der Bibliothek begann und eine neue Analysepipeline entwarf, die es mir ermöglichte, das Verhalten verschiedener Sequenzen in einer viel größeren Tiefe zu untersuchen. Ich bestätigte in diesen Analysen, dass *E. coli*-Zellen sehr tolerant gegenüber der Expression zufälliger DNA-Sequenzen sind, und fand heraus, dass die Länge—nicht aber die intrinsische Unordnung, der GC-Gehalt oder die Aggregationswahrscheinlichkeit—ein entscheidender Faktor dafür ist, ob eine Sequenz einen offenbar positiven Effekt auf die Fitness der Zellen hat. In Kapitel 2 wird versucht, das Experiment mit den zufälligen Sequenzen in einer eukaryotischen Zelllinie zu wiederholen. Ziel war es, zu testen, ob ein ähnlicher Anteil an Zufallssequenzen von den eukaryotischen Zellen trotz des unterschiedlichen zellulären Komplexitätsniveaus gut toleriert wird. Im Allgemeinen deuten die Ergebnisse darauf hin, dass eukaryotische Zellen mindestens genauso tolerant gegenüber der Expression von Zufallssequenzen sind wie die Bakterien. Im Gegensatz zu den Ergebnissen bei Bakterien korreliert jedoch kein spezifisches Merkmal der Sequenzen mit ihrer Verträglichkeit. In Kapitel 3 schließlich habe ich drei mutmaßliche *de novo*-Gene, die in der Maus identifiziert wurden, in einer menschlichen kontinuierlichen Zelllinie exprimiert. Bei den ausgewählten Genen handelt es sich um junge Gene, die taxonomisch auf Arten der Gattung Mus beschränkt sind und für die es transkriptions- und proteomische Nachweise gibt. Interessanterweise hatte die Expression der neuen Gene keine größeren Auswirkungen auf das Transkriptom der Zellen, was darauf hindeutet, dass sie auch in einem menschlichen zellulären Hintergrund toleriert werden würden.

Die in dieser Arbeit vorgestellten Ergebnisse ergänzen die zunehmenden Belege dafür, dass Zellen viel toleranter gegenüber der Expression neuer Sequenzen sind als bisher angenommen. Diese Erkenntnis wirft neue Fragen über die Entstehung von Genen auf, die in Zukunft erforscht werden sollten.

[1] Text kindly translated by Prof. Dr. Diethard Tautz.

Table of Contents

Section I. General introduction

1. Novel and de novo genes

The mechanism of birth and death of genes is a topic that has interested evolutionary biologists since the early 20th century (Long et al., 2013). Looking at the phenotypic diversity and complexity of different organisms, it is evident that during the course of evolution innovations have appeared, which are the force that drives evolution (Chen et al., 2013). Gene content of different organisms is different and genes are constantly gained and lost through evolution (Snel et al., 2002). Comparisons of different organisms' genomes, made possible by second-generation sequencing technologies, have shown that these innovations have a genetic basis, with new genes appearing at every taxonomic level, which are exclusive to the clade under study. These genes without homologs in other evolutionary lineages are often called in the literature novel genes, orphan genes, or taxonomically restricted genes (Daubin & Ochman, 2004a; Khalturin et al., 2009; Schmitz & Bornberg-Bauer, 2017; Tautz & Domazet-Loso, 2011).

There are different mechanisms through which novel genes may arise (Daubin & Ochman, 2004b; Kaessmann, 2010). The most common and best-studied one is the duplication and divergence of pre-existing genes. In this process, a duplication of a genomic region comprising one or several genes, or parts of a gene, reduces selective constraints on in, allowing for it to accumulate mutations and eventually diverge in function from the original one. The study of this process, ever since the late 1960's (Ohno et al., 1968), led to our current knowledge about the dynamic process of gene birth and death (Prince & Pickett, 2002). Once a gene, or part of it, has been duplicated, the accumulation of mutations could lead to different paths: homology with the parental gene could become hard to identify; the gene could become pseudogenised—the most likely outcome, according to most estimates (Lynch & Conery, 2000)—, or the copy could facilitate other events such as exon shuffling and gene fusion.

Novel genes can also arise by other mechanisms such as horizontal gene transfer, exon shuffling, retroposition, and others (Chen et al., 2013). Increasing evidence showing that the process of gene birth is highly dynamic, and that innovation at the genetic level can come from many different sources, has led to the latest addition to the list of mechanism of gene origination: *de novo* gene birth (Tautz, 2014). *De novo* genes are a subset of novel genes that originated from previously non-coding sequences (Cai et al., 2008; Snel et al., 2002). These sequences could be introns, intergenic regions, or frameshifts on coding regions (a process otherwise known as overprinting).

The possibility that genes may arise *de novo* from previously non-coding sequences was disregarded until the early 2000s. Before this, the common view was that evolution was only able to use existing "functional" pieces to create new ones—an idea famously stated in François

Jacob's influential essay "Evolution and tinkering" (Jacob, 1977). However, there is currently more and more evidence that *de novo* gene evolution is, not only possible, but also an important source of genetic innovations. Research in the past decade has identified and characterized *de novo* or putative *de novo* genes in rice (Zhang et al., 2019), *Arabidopsis* (Li et al., 2016), molluscs (Aguilera et al., 2017), *Drosophila* (Gubala et al., 2017; Reinhardt et al., 2013; Zhou et al., 2008), primates (Carelli et al., 2018; Ruiz-Orera et al., 2015), virus (Sabath et al., 2012) and bacteria (Daubin & Ochman, 2004a), among other organisms (Cai et al., 2008; Xiao et al., 2009).

Large scale searches of *de novo* gene candidates such as the ones mentioned above, were made available by the development of phylostratigraphy, a method that takes advantage of the large amount of sequencing data from many different organisms to find homologues of the protein sequence set of an organism (Domazet-Loso et al., 2007). Since the phylostratigraphy method uses sequence similarity to find potential novel gene candidates, it is not possible to be sure that these candidates are in fact *de novo*, or whether they are old genes that have diverged beyond the detection limits for homology search. The method also has the limitation of depending on the available data, which decreases its power when sequencing data for specific clades is missing. It is, however, extremely useful in directing efforts and serving as a first filter to find true *de novo* genes.

2. The process of de novo gene birth

2.1 Requirements for *de novo* gene birth

There is still much we ignore about how a non-coding sequence can become a protein-coding gene (Bornberg-Bauer & Heames, 2019). There is, however some consensus regarding the necessary pieces. The first one is, of course, the non-coding sequence itself. The most obvious sources for non-coding sequences are introns and intergenic regions, which are expected to evolve neutrally. Since mutations in such regions accumulate in a neutral way, introns and intergenic sequences provide a close to random sequence space in which, by chance, ORFs may appear and be expressed. As mentioned above, *de novo* genes have also been documented in viruses and bacteria, which lack these non-coding regions. In such cases, the new genes appear through the process of overprinting of a coding region in which a new reading frame overlapping the first one is generated through mutation (Sabath et al., 2012).

The second and third pieces necessary for the birth of a novel gene are: an ORF, and the different elements of the transcriptional/translational machinery, such as promoters and other regulatory elements. There are currently two complementary models to explain how *de novo* genes acquire these elements: ORF-first and RNA-first (McLysaght & Guerzoni, 2015). In the ORF-first model, a new gene may acquire a long enough ORF through random mutations first, and then obtain the regulatory elements from neighbouring promoters or transposable elements (ORF-first model). Intergenic regions at certain parts of the genome with high GC content are less likely to code for stop codons (which are T/A rich), and therefore have spurious ORFs that are long enough to be transcribed and translated. These ORFs could be in turn, spuriously transcribed and translated, even without the necessary regulatory elements, as it has been shown that pervasive transcription of the genome is a common phenomenon in genomes (Berretta & Morillon, 2009; Jacquier, 2009; Neme & Tautz, 2016). These spurious transcripts may confer a small advantage to the cells/organisms, and acquire afterwards the transcriptional machinery (promoters, and regulatory sequences).

Alternatively, a non-coding region of the genome might have regulatory elements present, and be partially transcribed, or already functional as a non-coding RNA (RNA-first model). Promoters and regulatory sequences of the transcriptional machinery may affect non-coding regions of the genome when they are moved around with transposable elements, or when a promoter becomes bidirectional, or through readthrough of coding genes (Carelli et al., 2018; Ruiz-Orera et al., 2015). When this happens, short sequences are transcribed, may gain longer ORFs through point mutations that remove stop codons, and can thus become new protein-coding genes. There is ample supporting evidence for both models, and it is likely that *de novo* genes arise through a combination of both paths (Schlotterer, 2015). Once a long enough ORF becomes transcribed, it can be translated. Non-coding transcripts have been repeatedly reported to be found in association with ribosomes (Bazzini et al., 2014; Durand et al., 2019; Ruiz-Orera et al., 2014; Ruiz-Orera et al., 2018; Wilson & Masel, 2011). Furthermore, there is also an abundance of reports of functional short peptides translated from ORFs in non-coding genes (for a review, see (Wang et al., 2019)), and non-coding sequences can be under purifying or positive selection (Bird et al., 2006).

Another possible mechanism for *de novo* gene birth, which provides simultaneously all necessary pieces is known as the “grow slow and moult” model (Bornberg-Bauer et al., 2015). According to this model, protein domains could originate *de novo,* not as individual genes, but as new parts of existing ones. This was show in a study of insect proteins, where the authors found evidence of “readthrough” expression at the 3’ and 5’ ends of many proteins, which could become new domains in the repertoire of the organism. It is likely that this mechanism is at work together with the other ones.

2.2 The continuum between non-coding sequences and *de novo* genes

Since different elements required for the formation of a new gene do not have to be gained simultaneously, the process of *de novo* gene birth occurs, most likely, as a gradual transformation of non-coding sequences into coding ones. This idea, known as the “protogene model”, was first proposed on 2012 (Carvunis et al., 2012). A sequence in the process of acquiring the elements required for transcription and translation, and in the process of becoming fixed as a gene in a population is known as a protogene. The origination process described above would generate a continuum of molecules ranging from short, spurious transcripts to translated sequences and finally *de novo* genes. Protogenes do not have yet a stable expression, nor have been fixed in a population, but are expressed in high enough levels to become exposed to natural selection and the corresponding evolutionary dynamics in populations. They may undergo a very rapid and dynamic gene gain/loss cycle, with only few of them becoming fixed and established as protein coding genes. The rest become pseudogenes, and go back to being non-coding sequences from which other *de novo* genes may arise.

Given that protogenes are expressed non-coding sequences that are being filtered through selection, the authors of the model suggested that they should exist in a continuum between intergenic-, and gene-like characteristics. This hypothesis is supported by their data analysis in *Saccharomyces cerevisiae*, where they identified protogenes and young genes using the phylostratigraphy method (Domazet-Loso et al., 2007), and found that properties such as length, intrinsic disorder score (IDS), hydropathicity (amino acid composition), codon adaptation index (CAI), and GC content had intermediate values in candidate protogenes compared to conserved genes and intergenic regions. The dynamics of *de novo* gene fixation in a population, and integration into metabolic and regulatory networks in the cell are not yet well understood.

Another hypothesis regarding the types of sequences that might become *de novo* genes has been proposed in contrast to the protogene model. The preadaptation hypothesis suggests that *de novo* gene birth is not a gradual process, but rather an all-or-nothing transition to expression. According to this hypothesis, young genes are born in an organism when non-coding regions have exaggerated gene-like properties. These “hopeful monsters” would have less probabilities of being deleterious than random sequences due to their exaggerated gene-like properties, which would, naturally, not be in a continuum between intergenic regions and conserved genes. Evidence to this hypothesis was presented by Wilson et al. in a publication that studied the properties of mouse novel genes and found that young, novel genes are more intrinsically disordered than both conserved (old) genes and intergenic sequences (Wilson et al., 2017).

3. Integration of novel genes into regulatory and interaction networks

In order for a new protein coding gene to have a function, it needs to integrate itself either to the regulatory or the interactome networks of an organism. It is likely that the first step is for genes to gain regulatory interactions. New regulatory interactions emerge rapidly within a few million years while protein-protein and protein-gene interactions emerge slowly. It is, of course, more likely that a novel gene does not integrate, and just becomes pseudogenized instead (Abrusan, 2013).

As for the regulatory interactions, the evidence suggests that it is likely that *de novo* genes are first regulated by enhancers. Novel genes in mammals have been found to be preferentially closer to enhancers than to promoters (Majic & Payne, 2020). Enhancers can evolve from many different sources including transposons, promoter switching, co-option, and even *de novo* (although this last mechanism has not been well studied) (Rebeiz & Tsiantis, 2017). Furthermore, enhancers evolve rapidly, and they are usually derived from ancient sequences. Novel promoters, on the other hand, evolve from novel sequences. (Villar et al., 2015). Later in the process, it is possible for enhancers to become promoters, for stronger regulatory interactions, as has been documented in mammals (Carelli et al., 2018).

Young genes evolve rapidly and are exposed to less selective constraints than older genes, similar to intergenic sequences (Heames et al., 2020). In the initial stages of evolution of any gene, rates of evolution are initially fast, and then slow down (Elena et al., 1996; Goodman, 1981). Several authors agree that future studies should focus more on the study of how new genes emerge, using functional characterization studies and experimental evolution approaches (Light et al., 2014; Schlotterer, 2015; Schmitz & Bornberg-Bauer, 2017).

3.1 Features of novel and *de novo* genes

In order for a *de novo* gene to be unequivocally identified as such, the syntenic non-coding region in closely related species must be found (McLysaght & Hurst, 2016). However, this usually proves to be a difficult task due to the fast divergence of non-coding regions and detection limits of current tools (Light et al., 2014). This difficulty in having a big sample of unequivocal genes to analyse has also interfered with estimations of the frequency in which *de novo* genes appear, the probability that protogenes become genes integrated into cellular pathways, and the evolutionary dynamics in play in this process.

One more aspect of *de novo* genes that has received considerable attention in the literature, is the characteristics that they have in common. In general, since it is difficult to unequivocally identify the novo genes, researchers have focused mainly in the differences between novel and ancient (conserved, core, essential) genes. There is agreement in that young genes are shorter than older ones, have lower transcription levels, are expressed in a tissue-specific manner, and evolve rapidly (Basile et al., 2017; Carvunis et al., 2012; Heames et al., 2020; Luis Villanueva-Canas et al., 2017; Schmitz et al., 2018). There are also some features for which no consensus exists and seem to be dependent on the lineage or species of study (Van Oss & Carvunis, 2019). These features are in particular GC content, intrinsic disorder and aggregation propensity.

A 2017 study found that GC content is highly correlated with intrinsic disorder since amino acids encoded by codons with high GC are disorder promoting. In most organisms, orphans are more disordered than older genes. (except for yeast, which has low GC) (Basile et al., 2017)**.** A 2016 study, reported that younger genes in *E. coli* are less random than older ones, as calculated by randomness tests, they have lower GC content, and they are shorter (Wang et al., 2016). *De novo* and novel domains also have high intrinsic disorder and evolve rapidly in insects (Klasberg et al., 2018).

4. Random sequences as a tool to study gene evolution

Since non-coding sequences are expected to accumulate mutations in a neutral—i.e., a nearly random—way, random sequences could be used as a proxy to study *de novo* gene evolution. The hypothesis that random sequences could be the origin of proteins was already postulated and explored mathematically in the early 1990s in the context of the ancestral set of peptides at the origin of life (White & Jacobs, 1990, 1993). Random sequences have proven repeatedly to be raw material from which novel function can arise in experimental evolution experiments. Some examples of this include the evolution of functional ATP-binding sequences from libraries of random sequences and phage display (Keefe & Szostak, 2001); increasing phage infectivity from a library of random sequences (Hayashi et al., 2003); an experimental evolution experiment to find peptides of 140 residues with DNA binding affinity of a specific sequence (Nakashima et al., 2007); evolution of resistance to nickel in bacteria from a 20-mer peptide with 12 random positions (Stepanov & Fox, 2007); and the evolution of antibiotic resistance genes from random sequences (Knopp, 2019; Knopp et al., 2021).

Some studies that could be helpful to guide our efforts in understanding this complexity have been published in the past years. These studies take advantage of next generation sequencing, and the potential of synthetic random sequences to study the possibility that something that is inherently non-coding could be tolerated and even used by living cells. One study, for example, replaced the promoter sequence of a LAC operon in *Escherichia coli* with random sequences of the same length in order to determine which sequences could be used by the bacteria as promoters, or how fast new, functional promoters could evolve from them (Yona et al., 2018). The surprising results showed that 11% of the sequences used could promote the transcription of a reporter gene from the beginning, and over 50% of the remaining sequences became functional promoters with just one mutation. A similar study, done in *Saccharomyces cerevisiae*, showed that a large percentage of random sequences can also act as promoters in eukaryotic cells (de Boer et al., 2020).

A 2017 study using random sequences in *E. coli* showed interesting features of the expression of random sequences (Neme et al., 2017). In their paper, the authors tried to experimentally address the question of what is the likelihood of a coding sequence to have a biological activity that might impact the fitness of a cell. The aim of the work was to test whether random sequences expressed in a bacterial cell could have an effect that gave an advantage or disadvantage to the cell in terms of growth speed. To do this, the authors ligated a pool of random 150bp sequences generated by adding equimolar amounts of each nucleotide at every synthesis step into a commercial inducible expression vector (pFLAG-CTC, Sigma). This vector has start and stop codons in frame of the restriction site used, which means that the random sequences were flanked by a constant sequence of 12 bp on the 5'-end and 36 bp on the 3'-end. The downstream constant sequence encodes a FLAG-tag before the stop codon. The pool of vectors was used to transfect *E. coli* (DH10B) to generate a library of bacterial cells. Expression of the cloned peptides could be induced by adding the commonly used compound isopropyl β-d-1-thiogalactopyranoside (IPTG) to the culture media. The total number of sequences successfully cloned into the cells was not clearly reported in the publication, but the authors reported results for at least 1000 sequences coding for the full-length 65 amino acid-long peptides in their analyses.

Comparisons between random sequences and naturally occurring ones have also shed some light on what we could expect to be the characteristics of *de novo* genes. For example, in a study comparing random sequences with proteins from public databases, Angyan et al. showed that peptides translated from random sequences with GC content ranging 40 to 60 % show similar aggregation propensity, intrinsic disorder and transmembrane domain predictions to human (Angyan et al., 2012). On the other hand, a more recent study found that natural sequences have longer disordered regions than random sequences (Yu et al., 2016). A study on random peptide resistance to protease degradation also showed that at least 20 % of random peptides obtained by phage display show stable three-dimensional folding (Chiarabelli, 2006).

5. About this thesis

This thesis is the compilation of three projects developed in order to address some of the unresolved questions about the birth of *de novo* genes. In Chapter 1 I take a look at the results of the 2017 work in *E. coli* by reproducing the experiment with the library of random sequences, and combining it with the sequencing data of the 2017 publication. The aim was to address in depth some of the questions still unanswered about the likelihood that certain sequences might be favoured over others to be retained in a population, and about the molecular features that make it possible. This work has resulted in a recent publication.

In Chapter 2, I extended the question about what we can quantify of the effect of random sequences to eukaryotic cells. Using a human cell line, I generated a new library of random sequences in a heterologous expression system, which—combined with an amplicon sequencing approach—allowed me to track the changes in frequency of single clones in the library through 20 cell divisions.

Finally, in Chapter 3 I chose several candidates from a list of putative mouse *de novo* genes to express them in a human cell line in order to study their effect on the transcriptome. These candidates, which can be considered in effect *de novo* mouse sequences expressed in the human cells, could provide further indication about the tolerance of the cells to random sequences and how they could interact with the regulatory networks of the cells.

Section II. Using random sequences and a eukaryotic expression system to study de novo gene birth

Chapter 1. The effects of sequence length and composition of random sequence peptides on the growth of E. coli cells

The work done for this chapter has been accepted for publication in the journal Genes for the special issue "How do New Genes Originate and Evolve?" (Editors Manyuan Long and Esther Bethran) and it is also available as a pre-print article on bioRxiv under the DOI: https://doi.org/10.1101/2021.11.22.469569.

Johana Fajardo Castro [1], and Diethard Tautz [2],*

[1] Max Planck Institute for Evolutionary Biology, August-Thienemann Strasse 2, 24306 Plön, Germany; jfajardo@evolbio.mpg.de

[2] Max Planck Institute for Evolutionary Biology, August-Thienemann Strasse 2, 24306 Plön, Germany; tautz@evolbio.mpg.de

* Correspondence: tautz@evolbio.mpg.de

Abstract: We study the potential for the *de novo* evolution of genes from random nucleotide sequences using libraries of *E. coli* expressing random sequence peptides. We assess the effects of such peptides on cell growth by monitoring frequency changes of individual clones in a complex library through four serial passages. Using a new analysis pipeline that allows to trace peptides of all lengths, we find that over half of the peptides have consistent effects on cell growth. Across nine different experiments, around 16 % of clones increase in frequency and 36 % decrease, with some variation between individual experiments. Shorter peptides (8–20 residues), are more likely to increase in frequency, longer ones are more likely to decrease. GC content, amino acid composition, intrinsic disorder and aggregation propensity show slightly different patterns between peptide groups. Sequences that increase in frequency tend to be more disordered with lower aggregation propensity. This coincides with the observation that young genes with more disordered structures are better tolerated in genomes. Our data indicate that random sequences can be a source of evolutionary innovation, since a large fraction of them are well tolerated by the cells or can provide a growth advantage.

Keywords: *de novo* gene evolution, random peptide sequences, fitness, *E. coli*, protein structure genes MDPI

Article

The Effects of Sequence Length and Composition of Random Sequence Peptides on the Growth of *E. coli* Cells

Johana F. Castro and Diethard Tautz *

Max Planck Institute for Evolutionary Biology, August-Thienemann Strasse 2, 24306 Plön, Germany; jfcastro@evolbio.mpg.de
* Correspondence: tautz@evolbio.mpg.de

Abstract: We study the potential for the *de novo* evolution of genes from random nucleotide sequences using libraries of *E. coli* expressing random sequence peptides. We assess the effects of such peptides on cell growth by monitoring frequency changes in individual clones in a complex library through four serial passages. Using a new analysis pipeline that allows the tracing of peptides of all lengths, we find that over half of the peptides have consistent effects on cell growth. Across nine different experiments, around 16% of clones increase in frequency and 36% decrease, with some variation between individual experiments. Shorter peptides (8–20 residues), are more likely to increase in frequency, longer ones are more likely to decrease. GC content, amino acid composition, intrinsic disorder, and aggregation propensity show slightly different patterns between peptide groups. Sequences that increase in frequency tend to be more disordered with lower aggregation propensity. This coincides with the observation that young genes with more disordered structures are better tolerated in genomes. Our data indicate that random sequences can be a source of evolutionary innovation, since a large fraction of them are well tolerated by the cells or can provide a growth advantage.

Keywords: de novo gene evolution; random peptide sequences; fitness; *E. coli*; protein structure

check for updates

Citation: Castro, J.F.; Tautz, D. The Effects of Sequence Length and Composition of Random Sequence Peptides on the Growth of *E. coli* Cells. *Genes* **2021**, *12*, 0. https://doi.org/

Academic Editors: Manyuan Long and Esther Betran

Received: 17 October 2021
Accepted: 26 November 2021
Published: 28 November 2021

1. Introduction

New genes can arise by two alternative mechanisms [1–6]. The first is through duplication and/or recombination of existing genes or gene fragments, which later accumulate mutations that render them different from their parental genes. The second is de novo evolution from previously non-coding sequences. While this was long thought to be unlikely, there is now plenty of evidence that the process has probably been active throughout evolution [7–13]. However, since it is difficult to distinguish de novo evolution from duplication followed by divergence beyond sequence recognition [14], one can prove true de novo evolution only for relatively recent events, where evolutionary time has not been enough for accumulation of too many mutations [1]. Several dedicated studies on individual genes, including functional analyses, have been published [15–19]. In addition to this, there are well-documented cases of peptides with biological function derived from randomly synthesized sequences [20–24]. Overall genome comparisons between recently separated species have suggested that de novo evolved genes arise continuously with a high rate, but can also get lost at high rates [25–28]. This dynamic transformation of non-coding sequences into coding ones is very clear, especially in eukaryotes, where large parts of the non-coding genome are transcribed. Comparisons between closely related mouse populations and species revealed the transcription of these non-coding regions is subject to fast evolutionary change, such that within a time span of 10 million years the whole genome can become transcribed and thus subjected to evolutionary testing [8]. Hence, the raw material for de novo evolution, namely transcripts from initially non-coding DNA regions, is abundantly present.

Based on these insights, we previously developed an experimental approach to ask which fraction of random sequences has a potential biological function that could become subject to further adaptive evolution [29]. We expressed a library of sequences with random sequence composition in bacterial cells and monitored which sequences could provide a growth advantage or disadvantage to the cell in the context of four growth cycles of the whole library. The general experimental design for this experiment is shown in Figure 1.

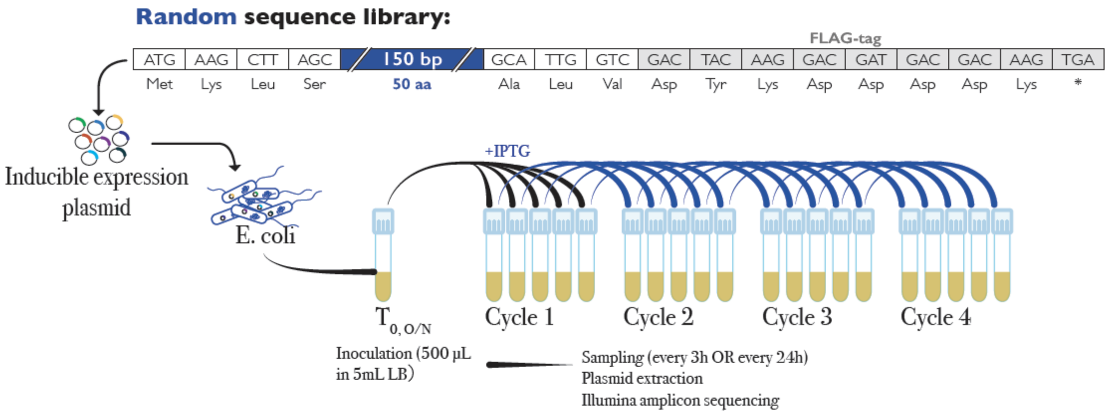

Random sequence library:
FLAG-tag
ATG AAG CTT AGC 150 bp GCA TTG GTC GAC TAC AAG GAC GAT GAC GAC AAG TGA
Met Lys Leu Ser 50 aa Ala Leu Val Asp Tyr Lys Asp Asp Asp Asp Lys *
Inducible expression plasmid
E. coli
+IPTG
$T_{0,O/N}$
Cycle 1
Cycle 2
Cycle 3
Cycle 4
Inoculation (500 µL in 5mL LB)
Sampling (every 3h OR every 24h)
Plasmid extraction
Illumina amplicon sequencing

Figure 1. Experimental design to evaluate the fraction of bioactive sequences in a library of random sequences. A pool of random 150 bp sequences generated by adding equimolar amounts of each nucleotide at every synthesis step was ligated into a commercial inducible expression vector (pFLAG-CTC, Sigma). This vector has start and stop codons in frame of the restriction site used for cloning, which means that the random sequences were flanked by a common sequence of 12 bp on the 5′-end and 36 bp on the 3′-end with a FLAG-tag (grey boxes). The resulting 195 nucleotide and 65 amino acid full sequences are shown. The pool of clones was used to transfect *E. coli* (DH10B) to generate a library of bacterial cells. Expression of the cloned peptides was induced by adding isopropyl β-d-1-thiogalactopyranoside (IPTG) to the culture media. Replicates were sampled every three h for a total of 12 h (3-h cycles, 12-h experiments) whereby one tenth of the culture volume was used for seeding the culture at each passage. The overall experiment replicated the one described in [29], where the analysis focused on full-length peptides only and also included experiments with 24-h growth cycles (5-day experiments). Here we use a newly designed pipeline to analyze all experiments and all peptide lengths.

The experiments showed that a surprisingly large fraction of random sequences affected cell growth, either by enhancing it, or by slowing it down. In the initial analysis, between 11 and 25% of the sequences increased in frequency in all replicates of each experiment, whereas 18 to 53%, decreased [29]. However, the study focused exclusively on the full-length peptides in the library, although the design strategy with random synthesis of the insert produces also a large number of truncated peptides with premature stop codons.

In the present study, we first reproduced the experiment, but with a lower concentration of starting library in an attempt to reduce the possible impact of very many low-frequency clones on the overall mean fitness of the complex library. Plus, we designed a new pipeline to analyze the new experiment, as well as all of the previous experiments. This new pipeline allowed us to include the clones expressing truncated peptides and to assess whether the expressed vector without insert could have a growth effect on the cells harboring it.

The new goal of this project was to explore the possible effects of shorter peptides in relation to the full-length peptides studied before. All peptides in the original analysis have common C-terminal residues (FLAG-tag—see Figure 1), which may have contributed to their stability and/or biological effects. Since naturally de novo evolved peptides would not have such a common C-terminus, it is important to verify whether the same spectrum of effects is also seen with peptides that have random C-termini. Furthermore, we wanted to explore sequence features of the peptides that could make them more or less likely to be tolerated by the cells and to be maintained in the population through several cycles of growth. Finally, we wanted to address the critical points that were raised against our original experiment, where [30] and [31] suggested a vector effect driving the patterns of peptides that rise in frequency. In this view, the vector itself would have a negative effect due to expressing a 38 amino acid peptide (or a secondary RNA structure) under induction conditions, which would be relieved when a "neutral" random sequence was replacing it, giving the impression that the "neutral" sequence acts positively. While we had argued that this effect could not fully explain the data that we had at that time [32], further analysis of this question is certainly warranted.

2. Materials and Methods

2.1. Library and Replication Experiment

We used the original library described in [29] from a stock frozen in 20% glycerol. The general design of the library and the experiment are depicted in Figure 1. In order to assess whether there could be a complexity effect, we repeated the original experiment using a 100-fold dilution of the original library and a 1-day sampling schedule, with samplings every 3 h for a total of 4 samplings in 12 h. This was done by seeding 5 µL from the stock on 25 mL LB liquid medium with 500 µg/mL ampicillin, and allowed it to grow overnight at 37 °C with constant shaking (250 rpm). After 16 h, 500 µL of the liquid culture were transferred into five 5 mL tubes containing 4.5 mL of LB medium with 10^{-3} mol/L IPTG to induce expression of the random sequences. For each cycle, 500 µL of culture from each tube were used to seed a new replicate after 3 h of growth (37 °C, 250 rpm). From the remaining bacterial culture for each replicate, 3 mL were collected and used for plasmid extraction using a QIAprep Spin Miniprep kit (QIAGEN, Hilden, Germany). Extracted plasmids were eluted in 30 µL of elution buffer and stored at -20 °C until use.

Amplicon sequencing of the library was performed using specific barcoded primers to amplify a 356-nucleotide fragment including the random sequences in a one-step PCR using PHUSION HF master mix (Invitrogen) (all primers used are listed in Supplementary Table S1). The cycling program consisted of an initial denaturation at 98 °C for 30 s, followed by 25 cycles of 98 °C for 10 s, 65 °C for 20 s, and 72 °C for 1 minute. After a final elongation step of 72 °C for 10 minutes, samples were purified using a Qiagen MinElute Gel Extraction kit. Concentration of samples was calculated through relative quantification in an agarose gel, using a Molecular Imager(R) Gel Doc(TM) XR+ System with the Image Lab(TM) Software (Bio-Rad). Barcoded samples were pooled together in equal concentrations to obtain the sequencing library. Sequencing was done using Illumina's MiSeq Reagent Kit v3 with 300 cycles to get overlapping 300-nucleotide paired-end reads.

Available data

In addition to sequencing data from the diluted library experiment described above, we used the original fastq files for eight experiments described in [29]. The original experiments were done following two different sampling schedules: either a 1-day course with samplings every 3 h, or a 4-day course with samplings every 24 h. In either case, four timepoints were sampled. The number of replicates, cycle duration, and experiment length for each of the experiments are summarized in Table 1. In addition to three experiments with 10 replicates of each type of sampling schedule, we used two 4-day experiments with 5 replicates. One of them (experiment 7) was done with a treatment control without induction with IPTG, while the other one (experiment 8) was sequenced more deeply (5x more reads than the other experiments) to capture even rare clones present at low frequencies in the population.

Table 1. Clone performance in different experiments.

Exp [1]	Cycle Length/Experiment Length (Replicates)	N [2]	POS [3]	NEG [3]	NS [3]	Range Log2-Fold Change (Average)	Empty Vector Log2-Fold Change [4]
1	3 h/1 day (n = 10)	5625	0.11	0.36	0.53	−8.0 to 2.7 (−1.1)	1.1
2	3 h/1 day (n = 10)	5606	0.17	0.43	0.41	−7.7 to 2.2 (−1.2)	0.5
3	3 h/1 day (n = 10)	5638	0.18	0.40	0.42	−7.8 to 2.7 (−1.1)	0.4
4	24 h/4 days (n = 8)	5623	0.14	0.30	0.56	−5.2 to 5.2 (−0.5)	0.1
5	24 h/4 days (n = 10)	5596	0.10	0.26	0.64	−5.4 to 5.0 (−0.6)	−1.7
6	24 h/4 days (n = 10)	5632	0.26	0.41	0.32	−5.9 to 2.2 (−0.7)	−1.1
7	24 h/4 days (n = 5)	5623	0.07	0.28	0.65	−7.2 to 4.0 (−0.9)	−0.2
8	24 h/4 days (n = 5)	5689	0.27	0.46	0.27	−11.2 to 1.4 (−1.6)	−0.4
9	3 h/1 day (n = 5)/diluted library	5651	0.16	0.32	0.51	−8.4 to 5.6 (−0.7)	−0.7
	All experiments averages [5]:						
	All clones	5621	0.16	0.36	0.48		
	Clones with 4aa ORF	200	0.04	0.73	0.18		
	Clones with 5aa ORF	221	0.02	0.52	0.44		
	Clones with 6aa ORF	209	0.06	0.37	0.56		
	Clones with FLAG sequence	638	0.03	0.77	0.17		
	Clones with FLAG + 1 sequence	129	0.03	0.68	0.20		
	Clones with FLAG + 2 sequence	126	0.05	0.64	0.23		
	Clones 48+ aa without FLAG	237	0.06	0.55	0.34		

[1] For experiments 1–8 we reanalyzed the original fastq data from [29], experiment 9 with a diluted starting library was done within the framework of this study. [2] Number of clones detected among the 5701 unique sequence clones in the database for which at least 5 reads were mapped in each experiment. [3] Fraction of clones in each category. POS and NEG were assigned when $p_{adj} < 0.05$; otherwise, the clone was categorized as non-significant (NS). [4] All vector clone frequency changes were highly significant ($p_{adj} < 0.01$) in their respective experiment, except for Exp 4 ($p_{adj} > 0.05$). [5] The distribution of values from the 9 experiments is not significantly different from a normal distribution (Shapiro–Wilk test, $p > 0.5$).

2.2. Analysis Pipeline

First, the paired end reads for each experiment were trimmed using Trimmomatic (v. 0.36), and merged using the software USEARCH10 (-fastq_mergepairs, -fastq_maxdiffs 30, -fastq_minmergelen 100) [33]. Since each read in a pair covers the entire random sequence, up to 30 mismatches were allowed between the paired forward and reverse reads. The fastq_mergepairs algorithm resolves discrepancies between the forward and reverse reads by comparing the quality score for the conflicting position in each read. It keeps the residue with the best quality score in the merged read. Merging the reads with this algorithm reduces the percentage of sequencing errors kept in each read. Note that it is not possible to account for PCR errors that have occurred during the library preparation.

To remove reads that do not belong to a PCR product from the plasmids in the library, a custom Perl script was used to find and save all merged reads containing pre-defined sequences up- and downstream of the random sequences on the pFLAG-CTC plasmid. The pre-defined sequences were a 18 bp sequence around the start codon, and the FLAG-tag, including the stop codon. The reads thus selected are considered clean amplicon reads, trimmed around the pre-defined sequences, and used for all subsequent analyses.

2.3. Database Generation

To generate a database of all unique sequences in the library that could be detected by the amplicon sequencing approach, all clean reads from all available experiments and replicates were first dereplicated using USEARCH10. Dereplication was done in 3 rounds. In the first round, the nucleotide sequences were sorted alphabetically, and the -fastx_uniques option was used to remove duplicate sequences, keeping only one sequence of each type in the database while keeping track of the number of total sequences of each type with the -sizeout option. In this way repeated identical sequences were removed and a "size" annotation was added to the read name indicating how many identical matches were present in the clean read files. In the second round, all files with singletons removed were merged into a single file of all amplicon sequences available, sorted, and de-replicated again using the same exact-match method. This exact matching approach is prone to enrichment of PCR or sequencing errors, since any two reads with even a single nucleotide difference are kept as individual sequences in the database. Singleton reads—more likely to be PCR or sequencing errors—were removed and a third dereplication round using a clustering approach was implemented.

The third round of dereplication aimed to remove reads generated by PCR or sequencing artefacts. The clustering approach used is based on the one used for OTU validation in microbiome analyses. Reads were sorted in decreasing order of size annotation, and the -cluster_smallmem option of USEARCH10 was used with an identity cut-off of 0.97. The clustering algorithm used by USEARCH is a greedy clustering approach. Here, sorting by the size annotation means that high-frequency reads are used as centroids or seeds for clusters first. This strategy relies on the assumption that reads found in high frequencies are more likely to be real, and less-common, highly-similar reads are probably generated through PCR or sequencing errors. The identity threshold of 0.97 allows less frequent reads with, for example, up to 5 mismatches in the expected 195-nucleotide sequence to join the high-frequency centroids forming the clusters. Using an additional filter of minimum cluster size of 8 reads, commonly used in microbiome amplicon sequencing analyses, removes other artefacts from the database. The resulting library of unique clusters (Supplementary Figure S1, full database: SuppData_BACT_tableinfo.tsv and SuppData_BACT.tsv) was used as the final database.

This database served also basis for the simulation of a 100.000 sequence library in R by sampling A, T, G, and C using the calculated probabilities for each nucleotide at each position (see below).

2.4. Sequence Features

Several parameters were used to characterize the sequences in the complete database, as well as in the sequence groups generated after mapping of the reads to find changes in frequency. ORFs were predicted using the program getorf form the EMBOSS suite [34] using the full database as input (-minsize 12, -find 3). Only the first ORF was kept for each sequence. Predicted ORFs were translated in the first frame using transeq from the EMBOSS suite, and the first predicted peptide was kept for each ORF. Sequence length was calculated for each read, as well as the predicted ORF and peptide using bash programs.

The number of peptides of each length depends on the probability of getting a stop codon at each consecutive position, and not before. This is best described by the probability function of a geometric distribution:

$$(1\text{-}p)^{(k\text{-}1)}{*}p, \quad (1)$$

where k is the number of trials, in this case, the number of positions or the length of the sequence; and p is the probability of "success" or getting a stop codon. Multiplying this probability distribution by the number of synthesized sequences, one gets the expected count of peptides of each length. The resulting expected distribution of peptide lengths was used to confirm library quality (Supplementary Figure S2).

GC content was calculated as the percentage of guanine (G) and cytosine (C) in a sequence relative to its length using custom Perl scripts. This was done for the complete read, the random part of the sequence (obtained by trimming 12 nucleotides on the 5′-end and 33 nucleotides from the 3′- end of the clean reads), and the predicted ORF. Amino acid composition of the database and different sequence groups were calculated using the Biostrings package (V 2.58.0) from Bioconductor in R. Lists of sequences from each database formatted as AAStringSets were used as input for the letterfrequency function and amino acid frequencies were plotted for each sequence correcting for length. For the complete database, full-length predicted peptides were used, and frequencies were calculated for each sequence independently in order to obtain frequency distributions. For the group analysis, the flanking sequences were trimmed from the peptides and amino acid frequencies were calculated for the complete set of random amino acids as a single sequence.

Intrinsic disorder was calculated using the command line version of IUPred (IUPred2A) [35] with the -short option. Intrinsic disorder scores were averaged for each peptide to obtain single average disorder values. In addition to this, the fraction of residues with a predicted disorder score equal to or larger than 0.5 was calculated, producing comparable results (data not shown).

Protein aggregation propensity was calculated for all sequences in the database using the program PASTA 2.0 on the web server of The BioComputing POS lab of the University of Padua (Italy) (http://old.protein.bio.unipd.it/pasta2/ last accessed November 2021) [36]. For each sequence, free energy for the single-best pairing was obtained using the default settings for peptides. The best-energy pairing for self-aggregation was obtained for each sequence from the output files, and energies of -5 or less were considered indicative of a high probability of aggregation.

2.5. Mapping of Reads to Full Database

Clean reads for all replicates and timepoints in each experiment were mapped to the database using a global alignment-based method from the program USEARCH10 (option -usearch_global). For consistency with the clustering analysis, alignments had a minimum required identity of 0.97, minimum query coverage of 0.9, and maximum one hit, and 5 gaps. Hits were extracted from the search results and counted using custom bash scripts to generate count tables for each replicate in each experiment.

2.6. Frequency Change Determination and Group Assignment

Raw count tables for each experiment were used as input for statistical analyses using the package DESeq2 in R [37]. Count data of each experiment were analyzed independently using cycle number as explanatory variable, and only sequences that had at least 5 reads mapped in the whole experiment were kept.

DESeq2 was designed mostly for the analysis of RNASeq data, but is broadly applicable to a large range of data types that require to control for large dynamic range and dispersion effects [37]. This makes different experiments better comparable between each other. Based on the log2-fold changes provided by DESeq2 (full data in: SuppData_DESeq2_ALLexp_Cycle4vs1.tsv) we classify the clones into NEG for negative changes and POS for positive changes. In addition, we chose the multiple-testing corrected p_{adj} value (provided by the program) as a cut-off to create a category of NS ("non-significant") clones. For category assignment, a flag was added to each sequence on the database table depending on whether its fold-change was positive (1) or negative (1), and significant (p_{adj} 0.05) or non-significant (p_{adj} 0.05) for each experiment. For the overall assignments of sequences to one of the three categories, category flags were compared across all experiments, and a general flag (sign.most, in the database table) was assigned when at least the strict majority of experiments had the same flag (5 or more).

While p-values should normally not be used for a ranking between experiments, we believe that errors created in this way are small, or at least smaller than the variances that we see between the experiments anyway. A possible alternative for ranking the clones would be to calculate their individual fitness effects in the background of the mean fitness of the whole library, as suggested by [38]. However, these authors advise against using their procedure under conditions where fitness distribution are broadly spread, as is the case in our experiments (compare Figures 1 and 2 in [29]).

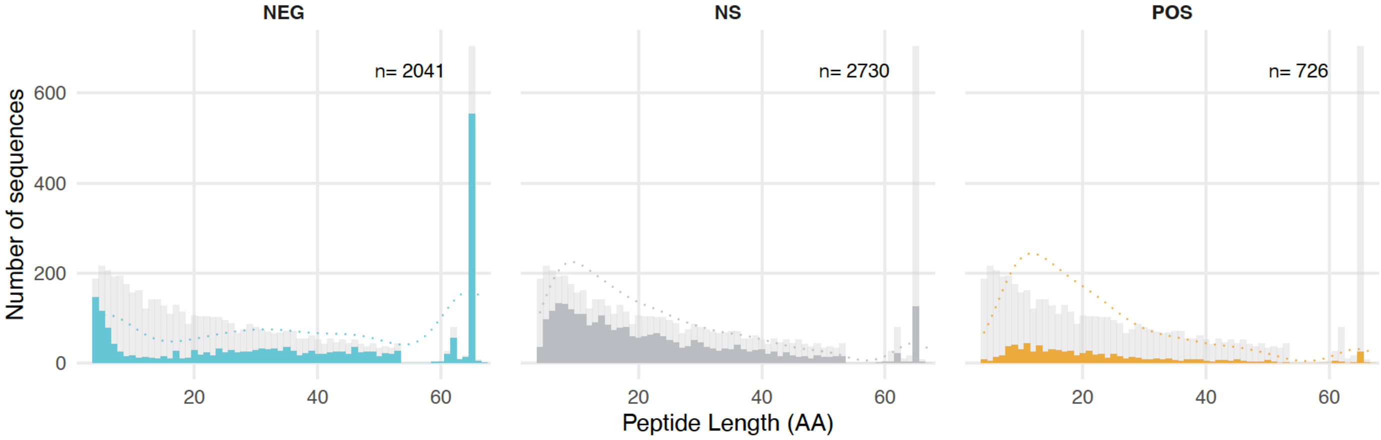


Figure 2. Length distribution of all predicted peptides in the random sequence database and assignment to response groups. Histogram of sequence lengths for each group of sequences. The colored bars represent the number of peptides of each length assigned to each group in the experiments. Light blue: peptides showing a decrease in frequency (NEG); dark grey: peptides showing no significant change in frequency (NS); orange: peptides showing an increase in frequency (POS). The light grey bars in each panel represent the predicted peptide lengths of the complete database (compare suppl. Figure 2). Dashed lines represent the kernel density estimates for each category.

3. Results

3.1. Replication with Diluted Library

In experiments with a complex library, all clones compete against each other, but rare clones generate only few reads that cannot be reliably analyzed. Hence, these unaccounted background clones can influence the behavior of the more frequent clones. In an attempt to test this possibility, we repeated the experiment of [29], but with a starting stock that was diluted by 100-fold compared to the previous ones and used a sampling schedule with samplings every 3 h for a total of 4 samplings in 12 h. The further experimental steps were conducted as described in [29]. The overall results showed that there was no major difference compared to the previous results (see Table 1 below). The majority of clones identified in the previous experiments could again be detected even with a 100-fold dilution. Hence, we decided to do the in-depth analysis described below across all available data.

3.2. Characterization of the Sequences in the Random Clone Library

To analyze all experiments done with the given clone library, we first produced a reference sequence database including all different sequences reliably detected in any of the sequencing experiments. This required the establishment of a pipeline for filtering of PCR and sequencing errors, which we conducted based on a common approach that is also used in microbiome studies. We required that each sequence was represented by at least eight reads, biasing against rare variants that can be generated in the PCR amplification steps before sequencing.

The median number of paired-end reads per replicate was 284,875. On average, 79.3% of them could be successfully merged, and both known plasmid-derived sequence regions could be found in 96.44% (±1.46%) of those merged. The resulting database consisted of 5701 unique sequences with minimum cluster size of eight. This included 647 peptides with the FLAG-tag sequence, of which 25 were not full-length due to internal deletions. There were 253 peptides that end with frameshift versions of the FLAG-tag sequence. Furthermore, since for the random part of peptides of lengths 4, 5 and 6, there were only 1, 21, and 441 possible different amino acid sequences, respectively, different clones could code for the same peptide. For example, the library includes 200 clones coding for the shortest possible peptide (MKLS—derived from the vector, see Figure 1), where the first triplet in the random sequence is a stop codon. Overall, the 5701 unique sequence clones coded for 5234 different peptides.

The dereplication algorithms used to generate the database provided information about the frequency of the different sequences in the library. The cluster size distribution is shown in Supplementary Figure S1. It has a right skewed distribution (mean: 20,274, median: 9870 sequences per cluster) with one extreme outlier with 4.2×10^7 sequences, which corresponds to the vector plasmid without insert ("empty" vector).

The ORF length distribution in the database has the expected composition and features of a random database of sequences, i.e., it follows largely the expected distribution of predicted peptide lengths (Supplementary Figure S2). Deviations concern mostly the longest sequence classes, due to the constant sequences derived from the vector. Note that some of the longest classes were also partly derived from frameshift versions.

With respect to GC content, we found that the sequences in the databases did not fully reflect a completely random synthesis. The mean and median GC content of the full reads was 53.8%, and median GC content of the predicted ORFs was slightly higher (mean 53.04%, median 54.6%) with larger variance due to the shorter sequences (Supplementary Figure S3A). A closer look to the GC content at every position in the database for reads with exactly the designed sequence length revealed a generalized bias towards lower A and higher G content at every position, remarkably larger on the 3′ end of the sequences starting at position 36 (Supplementary Figure S3B). This is probably due to a bias during library synthesis, with a presumptive new supply of chemicals in between. Still, given that the length distribution of resulting peptides conformed mostly to the random expectation (compare Supplementary Figure S2), we considered the library as being primarily made up of random nucleotide sequences.

A relevant descriptor of the structural properties of an amino acid sequence is its intrinsic disorder level. Intrinsically disordered proteins lack defined secondary and tertiary structures, and naturally occurring genes have a higher intrinsic disorder than random sequences [10,39–41]. In addition to intrinsic disorder scores, GC content [40] and amino acid content are used as indicators of the disorder levels of proteins in a database of sequences. Since large, hydrophobic amino acids are more likely to promote aggregation or formation of secondary structures, they are called order-inducing amino acids. The propensity of amino acids to induce order or disorder is one of the factors used for the calculation of intrinsic disorder scores [42].

Intrinsic disorder for the proteins in the database was calculated as the average intrinsic disorder score (IDS) of all residues in the peptide, using the -short setting of IUPred2A (see Methods). Average IDS values have a right-skewed bimodal distribution with the majority of sequences having an average IDS of 1.00 (Supplementary Figure S4A). This was due to the large number of short sequences in the database that were very unlikely to be able to make any secondary structures and were also under the limit of detection of the software used. Grouping the sequences into length classes shows this effect clearly. The mean of the distribution of average IDS shifts to smaller values for longer peptide lengths, ranging from 0.947 for the shortest peptides with less than 10 residues, to 0.281 for those with 48 or more residues (Supplementary Figure S4B). There is also a general correlation of IDS with length (Supplementary Figure S4C), as well as with GC content (Supplementary Figure S4D).

3.3. Frequency Changes in Clones during the Growth Experiments

For all sequencing files from the experiments, 80–90% of clean reads were successfully mapped to the database, allowing us to calculate frequency changes during the experiments. Raw count tables were used to do enrichment analyses using DESeq2. Although this algorithm was originally designed for the analysis of RNAseq sequencing data, it is also frequently used for the analysis of amplicon sequencing data. The assumption behind this is that the distribution of data in amplicon sequencing should follow a near-log normal distribution, with many low frequency counts and few high-frequency ones. The overall results of the DESeq2 analyses with respect to categorizing clones with positive (POS), negative (NEG), or non-significant (NS) changes are summarized in Table 1.

There is some variation between single experiments, especially with respect to the number of clones in the POS group. This is not directly related to the experiment type, i.e., the two experiments with the lowest fraction of POS clones (Exp 1 and Exp 7) have different cycle times (3 h vs. 24 h). Similarly, the range of log2-fold changes for individual clones varies considerably (Table 1). This suggests that even small variations in experimental conditions can lead to somewhat different outcomes. However, for all experiments there were always more NEG clones than POS clones. The average across all experiments shows 16% POS clones, 36% NEG clones, and 48% NS clones.

With the new pipeline, we could also trace the overall performance of the empty vector in the different experiments using the log2-fold change values. In experiments 1 to 3 it went slightly up, in experiments 5 to 9 it went slightly down, and in experiment 4 there was no significant change (Table 1). Only in experiment 5 the down trend was stronger than the average in this experiment. Note, however, that the DESeq2 normalization procedure penalizes against large count numbers in a way that could make negative trends stronger. Overall, we concluded from these data that the peptide and RNA expressed from the vector itself has no strong influence on growth.

We assessed also whether translation of the flanking sequences has a specific effect. The first four amino acids (MKLS) of the peptides were coded by the vector (see Figure 1). Of the clones that expressed only these first four amino acids due to a direct stop codon in the random sequence only 4% were POS while 78% are NEG, indicating a negative effect of this peptide compared to the overall clone performance (Table 1). Interestingly, this overall negative effect is relieved when only one or two additional amino acids were translated, with the percentage of NEG clones falling to 53% and 38%, respectively (Table 1). From this analysis we concluded that the vector-derived, constant N-terminal amino acids of the peptides have an overall negative effect on growth, which can be overcome by additionally coded amino acids or the RNA sequence components in the clones.

The C-terminus of the full-length peptides was formed by three constant amino acids plus the eight amino-acid FLAG tag sequence (see Figure 1). Of the different clones with this translated FLAG tag sequence, only 3% were POS while 75% were NEG across most experiments (Table 1), which would indicate a negative effect of this sequence. However, we find also the two frameshift translation versions of this sequence among the clones and both showed a similar excess of NEG versus POS effects (Table 1). This suggests that it is not the FLAG tag sequence that acted negatively, but that longer peptides have a generally higher likelihood of being in the NEG group. This is also supported by the fact that peptides with a length of 48+, but without including any of the FLAG tag versions, showed a similar bias towards NEG (Table 1) (see also the further analysis of the length effects below).

3.4. Length, GC Content, and Amino Acid Composition Dependence

For the further analysis, we assigned each sequence in the database into the categories POS, NEG, or NS, based on having consistent category assignments in the majority of experiments (see Methods). Since most sequences (95%) fell consistently within one of these three categories (with the remainder being inconsistent and therefore not further analyzed), one can compare whether peptides of different length are equally represented in each of these groups. Figure 2 shows that this is not the case. The fraction of NEG clones is particularly high for the shortest and the longest peptides. This is most likely caused by the negative effects of the vector derived parts of the sequence, as discussed above. The relative fraction of POS and NS clones is particularly high in the length classes between 8 and 20 amino acids.

The GC content distribution of ORFs in each of the groups is depicted in Figure 3. The peaks are similar for all three classes at about 57% GC, slightly higher than the average for the whole library, which is at 53% GC. The POS and NS peptides show broader distributions than the NEG peptides, with a stronger shoulder towards lower GC contents. The NEG peptides show a second peak at 42% GC, mostly driven by the negative effect of the shortest clones with only the vector-derived peptide (see above).

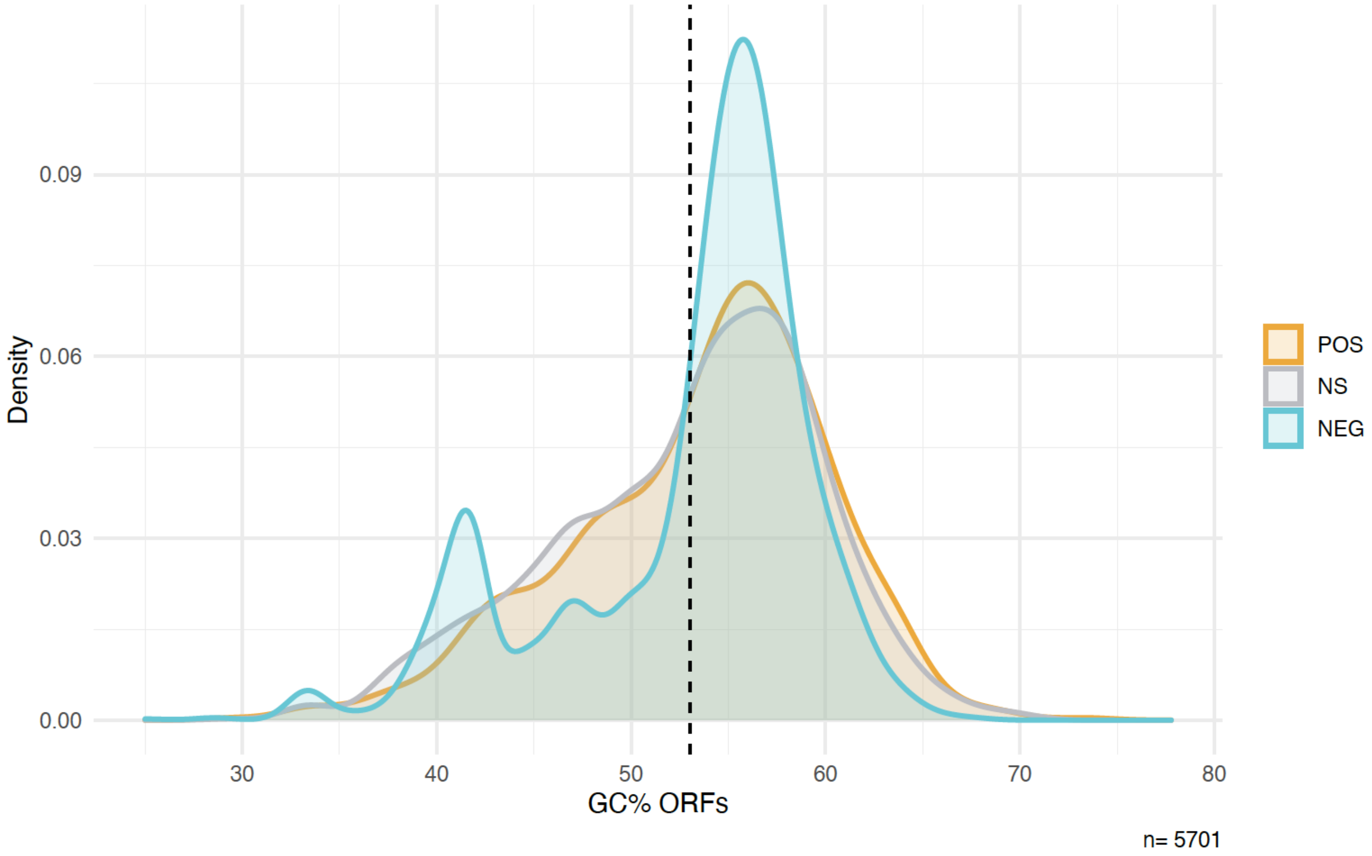


Figure 3. Density plots for the GC content of sequences in each of the three clone groups. Dashed line: average GC content of all ORFs in the library (53%). The colors represent the assignment to the three groups of clones in the experiments: Light blue: sequences showing a decrease in frequency (NEG); dark grey: sequences showing no significant change in frequency (NS); orange: sequences showing an increase in frequency (POS).

We also compared amino acid compositions of the peptides from the three clone groups. For this analysis we excluded the vector derived parts of the sequences. The overall frequencies for the whole database and the three groups of peptides are presented in Supplementary Table S2. Figure 4 shows the differences for each group compared to the database. The largest differences are found for A, G, and S in the comparison between the POS and NEG groups. It is also notable that the frequency of 7 out of the 10 amino acids considered to be more disorder-inducing is lower in the NEG group than in the database, while 9 out of 10 of the order-inducing amino acids are depleted in the NS group. The POS group shows in general the largest deviations from the database.

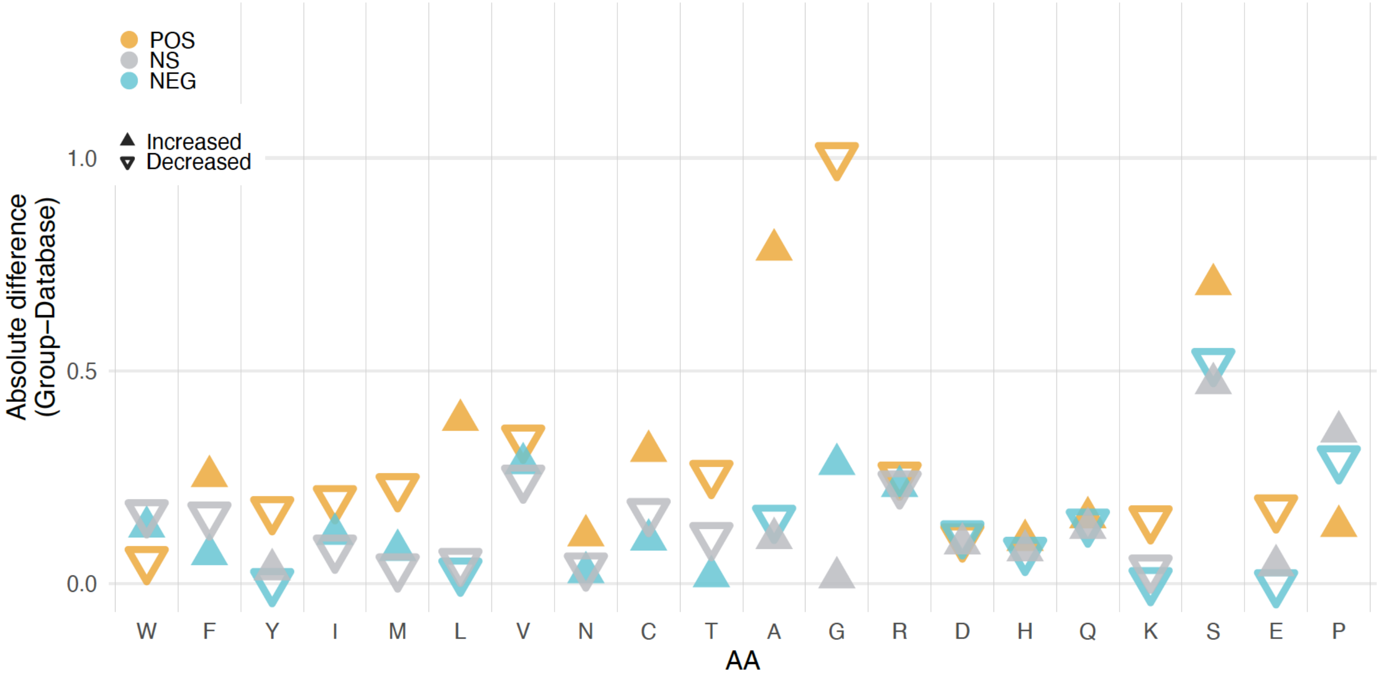


Figure 4. Differences in amino acid frequencies for the three groups of peptides. Frequencies were calculated as the percentage of each amino acid in all sequences in the groups and the complete database. Differences are shown as absolute values and the direction of the change is represented as +, if positive, or – is negative for each comparison. Light blue: NEG minus the database; grey: NS minus the database; orange: POS minus the database. Amino acids are ordered from left to right according to the TOP-IDP scale that reflects propensity for disorder induction from order promoting (left) to disorder promoting (right) [42].

3.5. *Structural Features*

The intrinsic disorder score (IDS) differs between the different clone groups, with the NEG group showing a stronger bimodal distribution than the two other groups (Figure 5). When breaking up the IDS in peptide length classes, it becomes clear that the highest IDS were due to the shortest classes (1–17 amino acids), for which the IDS calculation is anyway not very meaningful (compare also Supplementary Figure S4). The lowest IDS scores were seen for the longest peptides (48+ amino acids), but otherwise there was no clear difference, especially between the POS and NS groups of peptides (Figure 5).

There are generally few highly ordered sequences in the library (i.e., sequences with an average IUPred2 score of less than 0.25). This could be due to the fact that highly ordered sequences tend to aggregate, and are expected to be highly insoluble and detrimental to the cells. In order to assess aggregation propensity, we used the software PASTA 2.0. It calculates the free energy of predicted ß-strand intermolecular pairings for each sequence and reports the lowest value for each peptide as the best pairing [36]. Lower aggregation energies mean that it is easier for the peptides to form amyloids or to aggregate. In general, aggregation energies lower than -5 pasta energy units (PEU) are considered evidence for possible amyloid formation. Sequences in the NEG group show generally lower PEU values than the two other groups with a peak at - 4 PEU and a distribution shifted towards even lower values (Figure 6A). There is also a secondary peak at aggregation energies higher than the other two groups (Figure 6A).

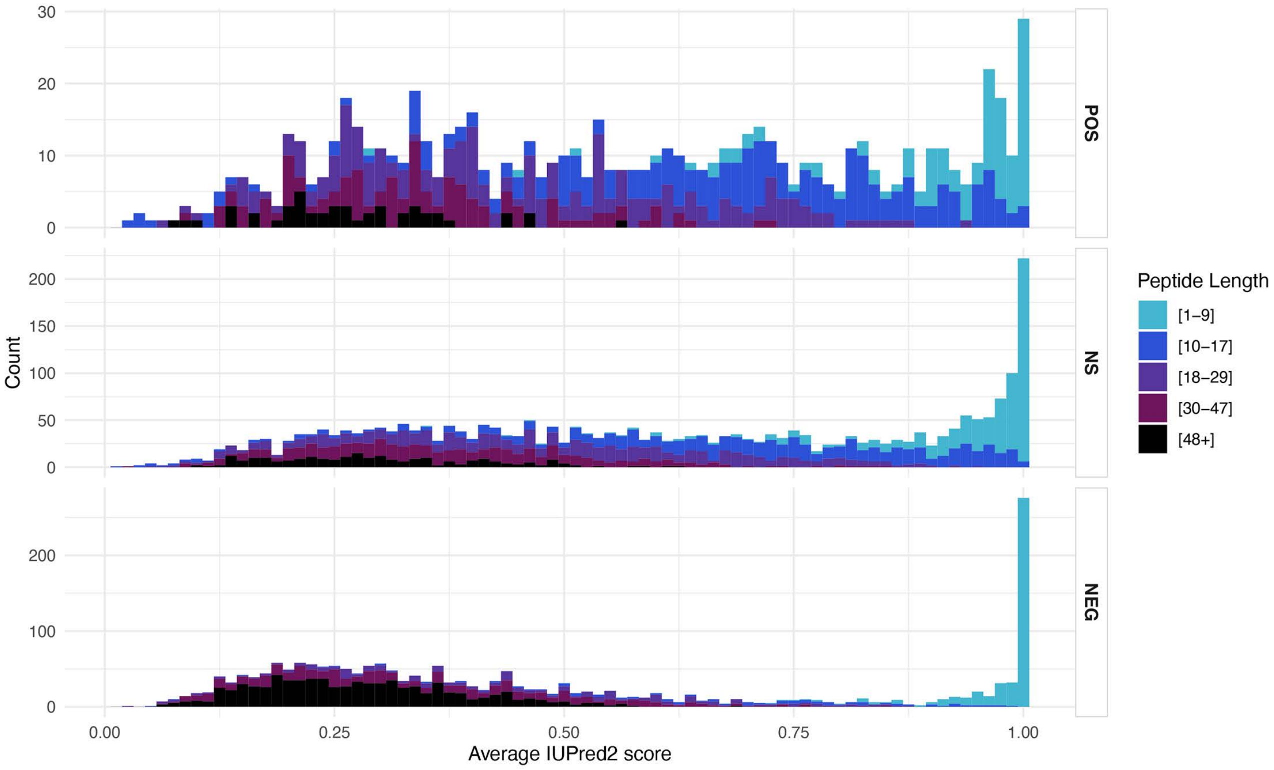


Figure 5. Intrinsic disorder scores (IDS) for the three groups of peptides. Histograms of average IUPred2 scores (IDS) colored by the length categories depicted to the right. OR Empirical cumulative distribution of average IUPred2 scores (IDS) colored by the length categories.

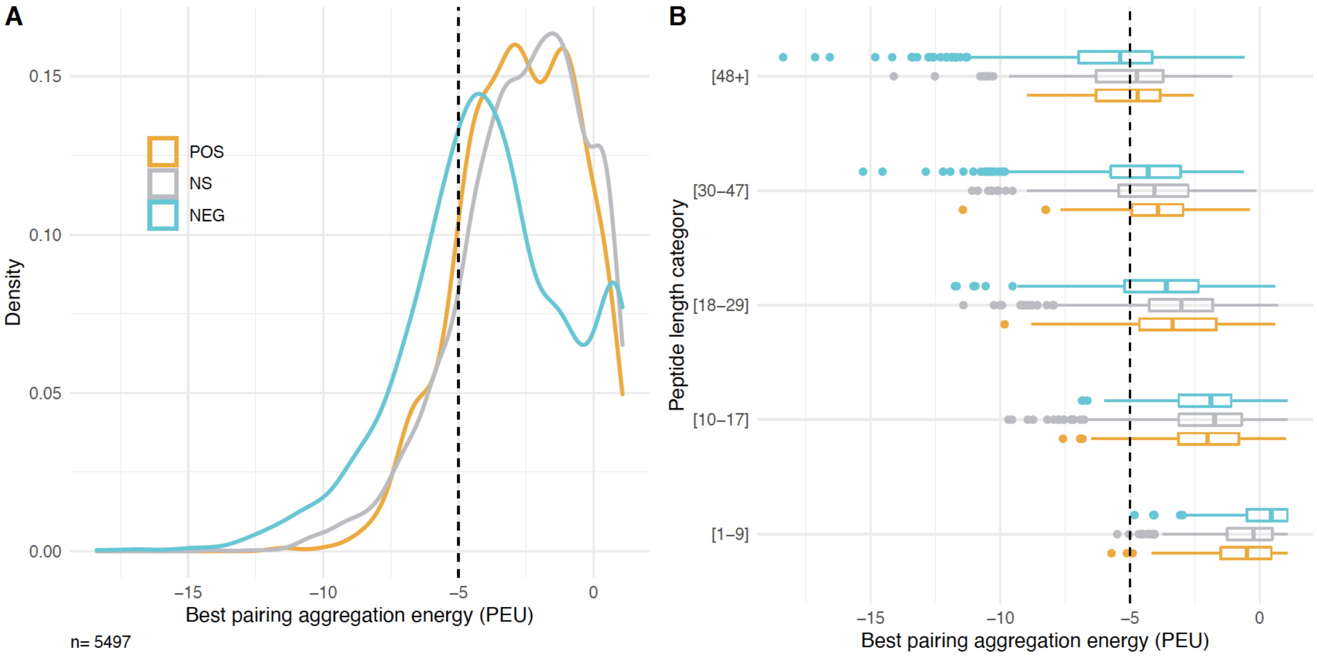


Figure 6. Aggregation energy analysis for the three clone groups and different length classes of peptides: (**A**) Density plots for the best aggregation energy of sequences in each group. (**B**) Best aggregation energy of sequences in each group and each length class. The lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The whiskers are the minimum and maximum data points up to 1.5 times the closest IQR.

In order to see whether sequences of a particular length are driving the observed pattern, we compared the data distribution for the different peptide length classes (Figure 6B). Interestingly, the distributions showed different patterns between the groups of clones at the highest and lowest length values, but more similar ones at the intermediate ones. The secondary peak found in the NEG group at very high aggregation energies seems to be generated mostly by the shortest peptides, and the shift towards negative values, by the longest ones.

4. Discussion

Here we have performed an in-depth analysis of all available data from amplicon sequencing experiments of a library of *E. coli* cells expressing different randomly synthesized sequences and grown for four expansion cycles to allow competition between clones. The experiments, first described by [29], were set up to assess which fraction of random DNA sequences expressed in a living organism has the potential of producing molecules that have an effect on cell growth or fitness. This question is relevant for the study of the origin of innovation in biological systems and, in particular, of de novo genes derived from more or less random non-coding sequences.

The main goal of the present study was to broaden the analysis to all peptides in the data, irrespective of their length. The authors in [29] had originally focused on the full-length peptides only, with FLAG-tags derived from the vector sequences. The broadening of the focus to all expressed peptides allowed us to investigate whether the constant sequences flanking the random inserts in the library influenced the growth effects. In addition to this, we wanted to test whether there are particular molecular or structural features of the sequences driving their effect on the growth of the cells in this population. Studies looking at young and novel genes in diverse species report that novel genes and proto-genes can have distinct features such as ORF length or intrinsic disorder levels that differentiate them from older genes and intergenic regions [10,12,39,41,43,44]. A possible explanation for these observations is that certain features make sequences more likely to be positively selected—or at least not selected against.

4.1. New Analysis Pipeline

The first step in our analysis was to ask whether amino acid sequences of different lengths present in the population and not analyzed in the original publication show similar behavior to the full-length (65 amino-acid-long) sequences that were the focus of the [29] study. This required us to generate a new analysis pipeline that addressed the three limitations of the original one. The first two—the incomplete removal of PCR and sequencing errors from the database, and the resulting artificial redundancy—are the result of using the predicted translation of the ORFs to generate the reference database of clones, instead of the nucleotide sequences. This was done to take advantage of the genetic code redundancy and to, at least partially, compensate for PCR and sequencing errors. The strategy, however, was insufficient as can be seen from the finding in the original publication of several very similar clones coding for peptides with only one or a few substitutions—an extremely improbable event in a library composed of random sequences of that length.

To compensate for such PCR and sequencing errors in this new analysis, we used a dereplication approach that removes all singleton reads from each sequencing file before joining them together and clustering them to a 97% identity. We used the full-length merged and trimmed reads for database generation and mapping, which allowed us to keep track of all clones that code for the same (shorter) peptides independently, which was important to detect protein vs. RNA effects (see further discussion below). The database generated was composed of more than 5000 reliably identifiable sequences predicted to code for peptides of all expected lengths. Furthermore, from its composition, it is possible to confirm that the library was indeed generated from random sequences, albeit with a slight bias towards a higher guanidine content during the synthesis process.

The final improvement to the pipeline was to change the algorithm used for mapping the reads back to the database from a local to a global alignment strategy. This greatly improved the speed and accuracy of the pipeline. Over 90% of all reads containing the flanking sequences mapped back to our database of unique sequences in all experiment replicates. This represents a 30–50% increase in mapped reads when compared with the pipeline used for analyses in the [29] study. As a result of this, we found that the change in frequencies can actually be much larger than what was initially reported. Some sequences, for example, have a decrease in frequency between the first and the last cycle of the experiments of up to 1000-fold.

4.2. Clone Effects

Having identified how the frequency of sequences changes in the available experiments, we were able to classify the sequences in groups according to the direction of the change. With the improved mapping pipeline, we found that over 80% of the sequences in the database had consistent behavior in at least five of the nine experiments, suggesting that the observed results were indeed an effect of the sequences and not due to chance or drift. This is noteworthy, considering that the experiments were performed independently, by different researchers, at different times, and have variations between sampling schedules, seed size, sequencing depths, and number of replicates.

Over half of the sequences in the database were consistently assigned to be either neutral (48%) or to go up in frequency (16%), suggesting that they are at least not very deleterious to the cells expressing them. This large proportion of sequences tolerated in a population of *E. coli* suggests that random sequences could also be expressed and maintained in large numbers in natural populations, making them an abundant source for possible evolutionary innovations.

We also specifically evaluated the vector effects that were suggested to have indirectly caused the observed positive effect [30,31]. In dedicated experiments, [31] found that just expressing the 38 amino acid peptides from the empty vector (i.e., without a cloned insert), had a slightly negative effect on the exponential growth of the *E. coli* cells. By disrupting this vector peptide with a potentially neutral peptide, one could generate an apparent positive effect. However, we found in our analysis that this peptide behaves mostly like a neutral peptide in the context of the full experiment, i.e., when not only focusing on the exponential growth phase as done in [31], but taking all competition cycles into account. While there is, on average, a small negative effect across experiments, it is not strong enough to explain the growth of most POS clones as merely its relief. Hence, we conclude that, in principle, the justified reservations about positive effects in our experiments [30,31] are not warranted in the face of the full data shown here, as well as the arguments provided previously [32].

4.3. Negative Effects of Vector Coded Amino Acids

Our data showed that the first four amino acids expressed by the vector had by themselves a negative fitness effect on the cells, with 73% of clones encoding only the first four residues, consistently decreasing in frequency in five or more experiments. A reason for this might be that the second and third codons in the sequence—lysine (AAG) and leucine (CTT), respectively—are not the most commonly used by *E. coli* for these amino acids. Interestingly, this negative effect diminishes quickly when one or two additional amino acids are translated. Hence, it is not of much concern for the overall experiment with mostly longer peptides; although, it contributes to the observed bimodal distributions of peptide length, intrinsic disorder, and aggregation propensity for the NEG peptides.

However, the same observation demonstrates that not only the coding part of a random sequence is important, for its maintenance in a population. Over 20% of the clones coding for the same peptide, but with different RNA sequences, show different growth trends in at least five experiments. In other words, clones with the same coding peptide had different effects on the growth trajectory of cells, due to the non-coding parts of their sequence. The authors in [29] had already shown that the RNA can have a different effect on growth than the protein by introducing a stop codon in single clones, disrupting the reading frame but keeping the rest of the sequence intact.

Systematic studies on replacing non-coding positions in an artificially expressed GFP RNA in *E. coli* have also shown that even small differences in RNA sequence can have differential fitness consequences for the cells [45], although this might be mostly caused by perturbing co-translational protein folding [46]. On the other hand, transcription has also been shown to contribute strongly to the metabolic burden that is caused by overexpressing genes in *E. coli* [47]. It is thus expected that the clone effects that we find are a combination of effects from the expressed RNA and protein together.

4.4. Protein Structure Correlations

Notwithstanding the possible fitness contribution of the RNA of the clones, we have analyzed protein structural properties in the three different groups of peptides. The most compelling difference between clones with POS or NEG responses is their length. Shorter peptides in the length range of 8–20 aa are prevalent in the POS and NS groups, while longer ones are prevalent in the NEG group. While it is generally known that newly evolved genes are shorter than older genes [7,43,44], the differences we observe here are at a much smaller scale than what is usually studied, since the ORF lengths of 4–65aa in our database are often not even annotated. Interestingly, in a study on the phenotypic impact of random sequences in Arabidopsis, [23] used also very short peptides (with cores of 6 or 12 random amino acids) and found a substantial fraction having an effect on the phenotype, including possibly beneficial ones.

The NEG group of peptides showed on average lower intrinsic disorder and higher aggregation propensity compared to the POS group. This is in line with the observation that naturally occurring young genes are more likely to have higher intrinsic disorder [10], which could be the reason why they are better tolerated by the cells [48].

We find no major differences in the three groups of peptides with respect to GC-content. However, there are some differences with respect to overall amino acid composition. The largest contrasts occur between POS and NEG peptides, whereby POS peptides have more alanine and serine but less glycine. With its six codons, serine is a frequent amino acid in the random sequences and it has a strong disorder promoting effect [42]. This could explain the higher disorder tendency in the POS peptides. Alanine and glycine, on the other hand, have both four codons and are therefore expected to occur equally frequently in random sequences, and they have similar disorder promoting effects. It is therefore unclear why alanine is more prevalent in POS and glycine is more prevalent in NEG clones.

An additional possible implication of the enrichment of serine in both the POS and NS groups is its potential for evolution. Creixell et al. [49] found that serine is the fastest-evolving amino acid and attribute this to fact that its six codons can be divided into two, very different, groups (AGY and TCN). The fact that the codons are so different facilitates non-synonymous substitutions, which allows evolution to explore a large sequence space in a shorter period of time. If, as our data seem to show, sequences containing larger fractions of serine are better tolerated by the cells, such sequences would be excellent starting material for the evolution of new functional peptides.

5. Conclusions

Although no single determining feature of a sequence could be identified that would earmark individual peptides as having potentially positive or negative effects on the cells, some differences exist with respect to structural properties. In particular, we found that shorter and more disordered peptides have a greater potential for being retained in a population as a primary source for novel genes, supporting the conclusions by James et al. on the general patterns of protein domain evolution [12]. Most importantly, our data confirm that random sequences have the potential of being beneficial for the cell, especially in the context of the complex competition between clones that we study in these experiments. However, we show in the accompanying paper [50] that individual candidate POS clones can also provide a growth advantage in pairwise competition experiments, although not necessarily with the same strength as seen in the bulk experiments. We conclude that our experiments support the notion that random sequences are an abundant source for generating evolutionary novelty.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/genes12120000/s1, Supplementary files: Figure S1: Cluster size distribution of clones, Figure S2: Predicted peptide length distribution of all clones in the library, Figure S3: GC content analysis of clones in the library, Figure S4: Intrinsic disorder scores for peptides, Table S1: Amino acid frequencies, Table S2: List of primers used, SuppData_BACT.tsv: results table for library analysis; SuppData_DESeq2_ALLexp_Cycle4vs1.tsv: DESeq2 results table for individual experiments; SuppData_BACT_tableInfo.tsv: Table legends.

Author Contributions: Conceptualization, J.F.C., and D.T.; data curation, J.F.C.; formal analysis, J.F.C.; investigation, J.F.C.; supervision, D.T.; validation, J.F.C. and D.T.; visualization, J.F.C.; writing—original draft, J.F.C.; writing—review and editing, D.T. All authors have read and agreed to the published version of the manuscript.

Funding: The work was funded by institutional funds of the MPG to DT. JF is a member of the IMPRS for Evolutionary Biology.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: Dryad http://dx.doi.org/10.5061/dryad.6f356 and at the European Nucleotide Archive (ENA) under the project number PRJEB19640. Sequencing data files for the replication experiments will are available at ENA, project number: PRJEB48999. All codes for data organization and analysis are available as GitLab project under https://github.com/johanafc/random-seqs-ecoli.

Acknowledgments: We thank Ellen McConnell for her help in getting the replication experiment established and Rafik Neme for his comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

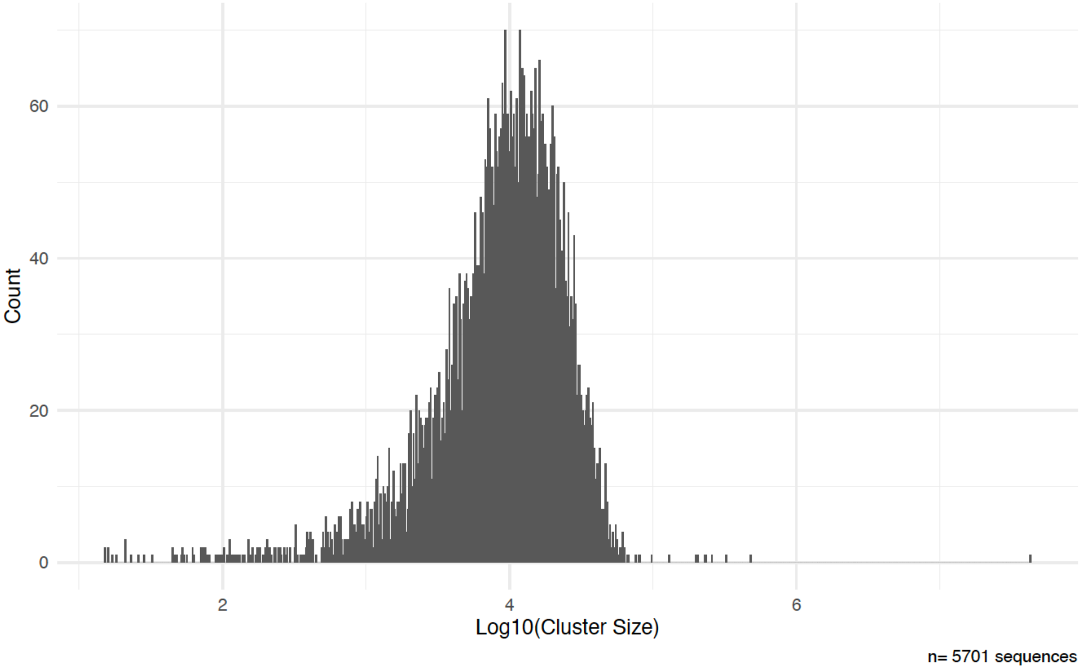
References

1. Tautz, D.; Domazet-Loso, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **2011**, *12*, 692–702. [CrossRef]
2. Chen, S.; Krinsky, B.H.; Long, M. New genes as drivers of phenotypic evolution. *Nat. Rev. Genet.* **2013**, *14*, 645–660. [CrossRef] [PubMed]
3. Schlötterer, C. Genes from scratch—The evolutionary fate of de novo genes. *Trends Genet.* **2015**, *31*, 215–219. [CrossRef] [PubMed]
4. McLysaght, A.; Guerzoni, D. New genes from non-coding sequence: The role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philos. Trans. R. Soc. B Biol. Sci.* **2015**, *370*, 20140332. [CrossRef] [PubMed]
5. Van Oss, S.B.; Carvunis, A.-R. De novo gene birth. *PLoS Genet.* **2019**, *15*, e1008160. [CrossRef]
6. Andersson, D.I.; Jerlström-Hultqvist, J.; Näsvall, J. Evolution of New Functions De Novo and from Preexisting Genes. *Cold Spring Harb. Perspect. Biol.* **2015**, *7*, a017996. [CrossRef]
7. Neme, R.; Tautz, D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genom.* **2013**, *14*, 1–13. [CrossRef] [PubMed]
8. Neme, R.; Tautz, D. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *eLife* **2016**, *5*, e09977. [CrossRef] [PubMed]
9. Ruiz-Orera, J.; Messeguer, X.; Subirana, J.; Alba, M.M. Long non-coding RNAs as a source of new peptides. *eLife* **2014**, *3*, e03523. [CrossRef] [PubMed]
10. Wilson, B.A.; Foy, S.G.; Neme, R.; Masel, J. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat. Ecol. Evol.* **2017**, *1*, 1–6. [CrossRef] [PubMed]
11. Pavesi, A.; Magiorkinis, G.; Karlin, D.G. Viral Proteins Originated De Novo by Overprinting Can Be Identified by Codon Usage: Application to the "Gene Nursery" of Deltaretroviruses. *PLoS Comput. Biol.* **2013**, *9*, e1003162. [CrossRef] [PubMed]
12. James, J.E.; Willis, S.M.; Nelson, P.G.; Weibel, C.; Kosinski, L.J.; Masel, J. Universal and taxon-specific trends in protein sequences as a function of age. *eLife* **2021**, *10*, e57347. [CrossRef]
13. Zhang, L.; Ren, Y.; Yang, T.; Li, G.; Chen, J.; Gschwend, A.R.; Yu, Y.; Hou, G.; Zi, J.; Zhou, R. Rapid evolution of protein diversity by de novo origination in Oryza. *Nat. Ecol. Evol.* **2019**, *3*, 679–690. [CrossRef]

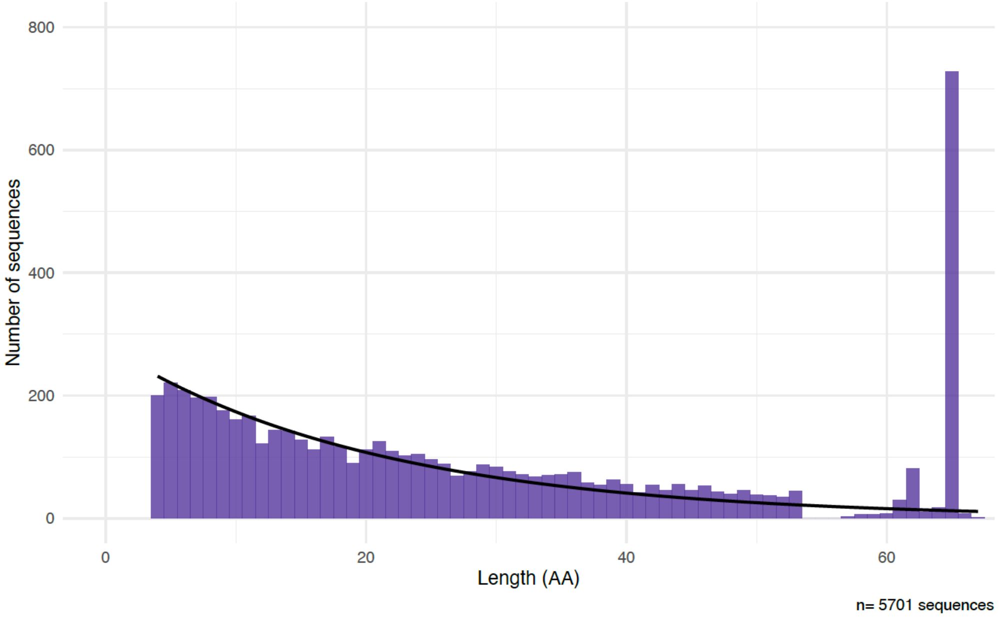
14. Weisman, C.M.; Murray, A.W.; Eddy, S.R. Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biol.* **2020**, *18*, e3000862. [CrossRef]
15. Heinen, T.J.; Staubach, F.; Häming, D.; Tautz, D. Emergence of a New Gene from an Intergenic Region. *Curr. Biol.* **2009**, *19*, 1527–1531. [CrossRef]
16. Xie, C.; Bekpen, C.; Künzel, S.; Keshavarz, M.; Krebs-Wheaton, R.; Skrabar, N.; Ullrich, K.K.; Tautz, D. A de novo evolved gene in the house mouse regulates female pregnancy cycles. *eLife* **2019**, *8*, e44392. [CrossRef]
17. Cai, J.; Zhao, R.; Jiang, H.; Wang, W. De Novo Origination of a New Protein-Coding Gene in Saccharomyces cerevisiae. *Genetics* **2008**, *179*, 487–496. [CrossRef] [PubMed]
18. Li, D.; Dong, Y.; Jiang, Y.; Jiang, H.; Cai, J.; Wang, W. A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Res.* **2010**, *20*, 408–420. [CrossRef] [PubMed]
19. Reinhardt, J.; Wanjiru, B.M.; Brant, A.T.; Saelao, P.; Begun, D.J.; Jones, C.D. De Novo ORFs in Drosophila Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-coding Sequences. *PLoS Genet.* **2013**, *9*, e1003860. [CrossRef] [PubMed]
20. Stepanov, V.G.; Fox, G.E. Stress-Driven In Vivo Selection of a Functional Mini-Gene from a Randomized DNA Library Expressing Combinatorial Peptides in Escherichia coli. *Mol. Biol. Evol.* **2007**, *24*, 1480–1491. [CrossRef] [PubMed]
21. Knopp, M.; Gudmundsdottir, J.S.; Nilsson, T.; König, F.; Warsi, O.; Rajer, F.; Ädelroth, P.; Andersson, D.I. De Novo Emergence of Peptides That Confer Antibiotic Resistance. *mBio* **2019**, *10*, e00837-19. [CrossRef] [PubMed]
22. Knopp, M.; Babina, A.M.; Gudmundsdóttir, J.S.; Douglass, M.V.; Trent, M.S.; Andersson, D.I. A novel type of colistin resistance genes selected from random sequence space. *PLoS Genet.* **2021**, *17*, e1009227. [CrossRef] [PubMed]
23. Bao, Z.; Clancy, M.A.; Carvalho, R.F.; Elliott, K.; Folta, K.M. Identification of Novel Growth Regulators in Plant Populations Expressing Random Peptides. *Plant Physiol.* **2017**, *175*, 619–627. [CrossRef] [PubMed]
24. Keefe, A.D.; Szostak, J.W. Functional proteins from a random-sequence library. *Nat. Cell Biol.* **2001**, *410*, 715–718. [CrossRef] [PubMed]
25. Zhao, L.; Saelao, P.; Jones, C.D.; Begun, D.J. Origin and Spread of de Novo Genes in Drosophila melanogaster Populations. *Science* **2014**, *343*, 769–772. [CrossRef] [PubMed]
26. Palmieri, N.; Kosiol, C.; Schlötterer, C. The life cycle of Drosophila orphan genes. *eLife* **2014**, *3*, e01311. [CrossRef]
27. Neme, R.; Tautz, D. Evolution: Dynamics of De Novo Gene Emergence. *Curr. Biol.* **2014**, *24*, R238–R240. [CrossRef]
28. Durand, É.; Gagnon-Arsenault, I.; Hallin, J.; Hatin, I.; Dubé, A.K.; Nielly-Thibault, L.; Namy, O.; Landry, C.R. Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. *Genome Res.* **2019**, *29*, 932–943. [CrossRef]
29. Neme, R.; Amador, C.; Yildirim, B.; McConnell, E.; Tautz, D. Random sequences are an abundant source of bioactive RNAs or peptides. *Nat. Ecol. Evol.* **2017**, *1*, 1–7. [CrossRef]
30. Weisman, C.; Eddy, S.R. Gene Evolution: Getting Something from Nothing. *Curr. Biol.* **2017**, *27*, R661–R663. [CrossRef]
31. Knopp, M.; Andersson, D.I. No beneficial fitness effects of random peptides. *Nat. Ecol. Evol.* **2018**, *2*, 1046–1047. [CrossRef] [PubMed]
32. Tautz, D.; Neme, R. Reply to 'No beneficial fitness effects of random peptides'. *Nat. Ecol. Evol.* **2018**, *2*, 1048. [CrossRef]
33. Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **2010**, *26*, 2460–2461. [CrossRef]
34. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277. [CrossRef]
35. Mészáros, B.; Erdős, G.; Dosztányi, Z. IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **2018**, *46*, W329–W337. [CrossRef]
36. Walsh, I.; Seno, F.; Tosatto, S.C.; Trovato, A. PASTA 2.0: An improved server for protein aggregation prediction. *Nucleic Acids Res.* **2014**, *42*, W301–W307. [CrossRef] [PubMed]
37. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 1–21. [CrossRef]
38. Li, F.; Salit, M.L.; Levy, S.F. Unbiased Fitness Estimation of Pooled Barcode or Amplicon Sequencing Studies. *Cell Syst.* **2018**, *7*, 521–525. [CrossRef] [PubMed]
39. Heames, B.; Schmitz, J.; Bornberg-Bauer, E. A Continuum of Evolving De Novo Genes Drives Protein-Coding Novelty in Drosophila. *J. Mol. Evol.* **2020**, *88*, 382–398. [CrossRef] [PubMed]
40. Basile, W.; Sachenkova, O.; Light, S.; Elofsson, A. High GC content causes orphan proteins to be intrinsically disordered. *PLoS Comput. Biol.* **2017**, *13*, e1005375. [CrossRef]
41. Yuedong, Y.; Cao, Z.; Yang, Y.; Wang, C.-L.; Su, Z.-D.; Zhao, Y.-W.; Wang, J.-H.; Zhou, Y. Natural protein sequences are more intrinsically disordered than random sequences. *Cell. Mol. Life Sci.* **2016**, *73*, 2949–2957. [CrossRef]
42. Campen, A.; Williams, R.; Brown, C.; Meng, J.; Uversky, V.; Dunker, A. TOP-IDP-Scale: A New Amino Acid Scale Measuring Propensity for Intrinsic Disorder. *Protein Pept. Lett.* **2008**, *15*, 956–963. [CrossRef] [PubMed]
43. Carvunis, A.-R.; Rolland, T.; Wapinski, I.; Calderwood, M.; Yildirim, M.; Simonis, N.; Charloteaux, B.; Hidalgo, C.; Barbette, J.; Santhanam, B.; et al. Proto-genes and de novo gene birth. *Nat. Cell Biol.* **2012**, *487*, 370–374. [CrossRef]
44. Schmitz, J.F.; Ullrich, K.K.; Bornberg-Bauer, E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat. Ecol. Evol.* **2018**, *2*, 1626–1632. [CrossRef]

45. Mittal, P.; Brindle, J.; Stephen, J.; Plotkin, J.B.; Kudla, G. Codon usage influences fitness through RNA toxicity. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 8639–8644. [CrossRef]
46. Walsh, I.M.; Bowman, M.A.; Santarriaga, I.F.S.; Rodriguez, A.; Clark, P.L. Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 3528–3534. [CrossRef] [PubMed]
47. Li, Z.; Rinas, U. Recombinant protein production associated growth inhibition results mainly from transcription and not from translation. *Microb. Cell Factories* **2020**, *19*, 1–11. [CrossRef] [PubMed]
48. Tretyachenko, V.; Vymětal, J.; Bednárová, L.; Kopecký, V.; Hofbauerova, K.; Jindrová, H.; Hubálek, M.; Souček, R.; Konvalinka, J.; Vondrášek, J.; et al. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Sci. Rep.* **2017**, *7*, 1–9. [CrossRef]
49. Creixell, P.; Schoof, E.M.; Tan, C.S.H.; Linding, R. Mutational properties of amino acid residues: Implications for evolvability of phosphorylatable residues. *Philos. Trans. R. Soc. B Biol. Sci.* **2012**, *367*, 2584–2593. [CrossRef]
50. Bhave, D.; Tautz, D. Effects of the expression of random sequence clones on growth and transcriptome regulation in Escherichia coli. *bioRxiv* **2021**. [CrossRef]

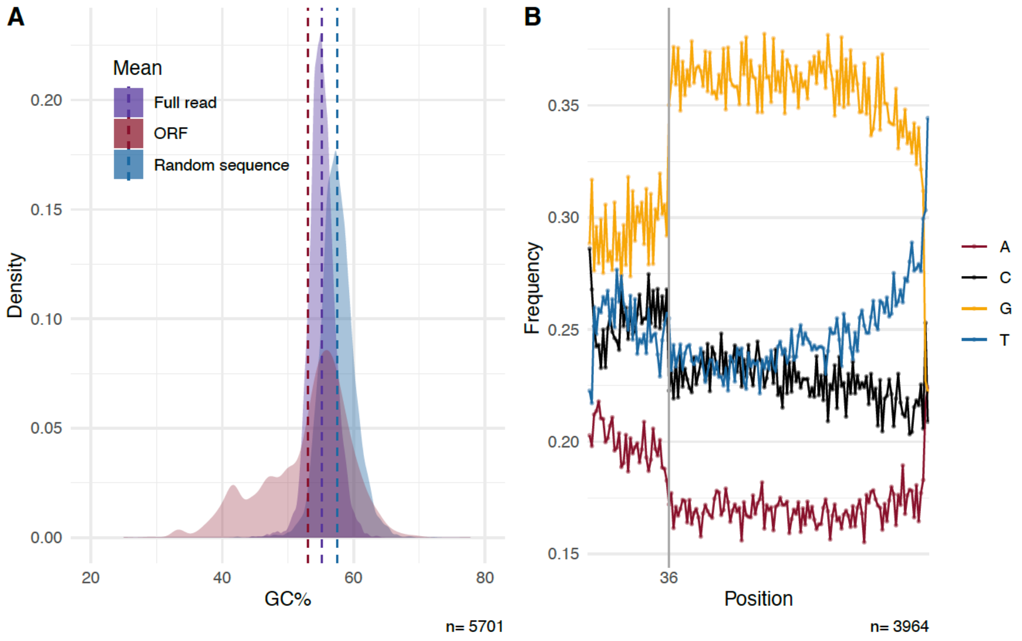
1.1 Supplementary Results



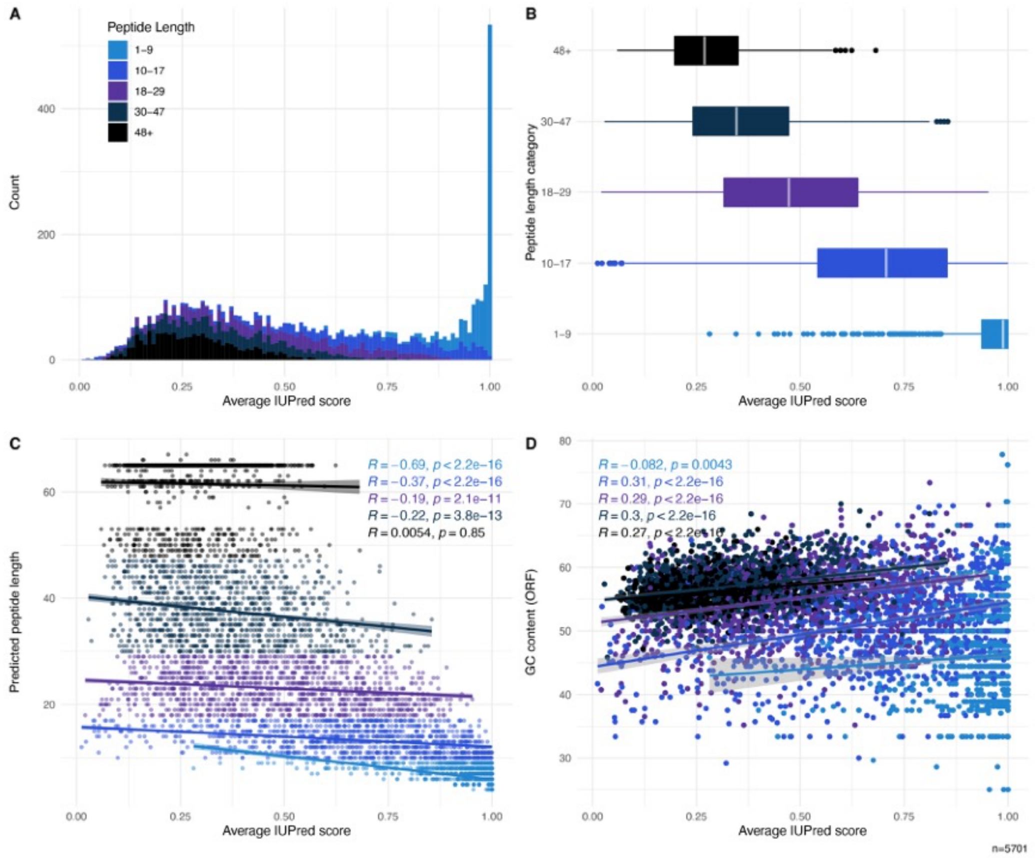
Supplementary Figure 1-1. Cluster size distribution of clones.
Cluster sizes correspond to the number of reads across all experiments assigned to each of the 5701 clones in the database. Most clones had between 103 and 105 reads assigned to them (mean 2.03x104 reads, median 9.9x103 reads). There is a single clone with 4.23x107 reads, which corresponds to the pFLAG-CTC plasmid without an insert.



Supplementary Figure 1-2. Predicted peptide length distribution of all clones in the library.
The library follows the expected distribution of stop codons in 5701 random sequences, defined as the probability mass function of a geometric distribution with p = 3/64 stop codons (black line). Only the first ORF in each sequence was considered. Minimum peptide length corresponds to the 4 residues encoded by the plasmid and insert design (MKLS). Residues 55 to 65 are constant in the sequence design (ALVDYKDDDDK), which should result in a single peak for sequences of length 65. However, the process of library synthesis or cloning generated 1735 clones with unexpected sequence lengths (30.43% of the clones in the library). This resulted in 171 clones with predicted peptide lengths between 55 and 64 residues or longer than 65 residues.*



Supplementary Figure 1-3. *GC content analysis of clones in the library.*
A) GC content distributions calculated over the complete reads, between start and FLAG-tag sequences (this includes the flanking sequences of the plasmid, with a GC content of 46.2%the full sequenced reads (purple), the ORFs only (red) and simulated random sequences (blue). The ORF GC distribution is much more broadly spread, due to increased variance caused by very short ORFs. B) Fraction of each nucleotide along the positions of the randomly synthesized sequence stretch.



Supplementary Figure 1-4. Intrinsic disorder scores for peptides in the whole library in four different representations (A-D; see main text).

Average intrinsic disorder scores become smaller for longer peptides (length 1-9 = 0.947, length 10-17 = 0.677, length 18-29 = 0.479, length 30-47 = 0.362, length 48+ = 0.281.

Supplementary Table 1-1. Amino acid frequencies for the database, as well as the three groups of peptides

AA	Database	UP	DOWN	NS
W	0.0284	0.0279	0.0298	0.0268
F	0.0320	0.0345	0.0326	0.0304
Y	0.0221	0.0204	0.0221	0.0224
I	0.0314	0.0294	0.0325	0.0306
M	0.0156	0.0134	0.0164	0.0153
L	0.0918	0.0957	0.0915	0.0913
V	0.0850	0.0816	0.0878	0.0825
N	0.0189	0.0200	0.0191	0.0185
C	0.0431	0.0462	0.0441	0.0414
T	0.0448	0.0422	0.0450	0.0437
A	0.0884	0.0963	0.0869	0.0895
G	0.1134	0.1034	0.1162	0.1136
R	0.1141	0.1116	0.1164	0.1118
D	0.0268	0.0257	0.0257	0.0277
H	0.0243	0.0253	0.0235	0.0251
Q	0.0264	0.0280	0.0250	0.0277
K	0.0205	0.0190	0.0204	0.0201
S	0.0904	0.0974	0.0852	0.0951
E	0.0282	0.0264	0.0282	0.0286
P	0.0545	0.0558	0.0516	0.0580

Supplementary Table 1-2. List of primers used

Primer Name	Primer Sequence
pFLAG-CTC FWD-1	AATGATACGGCGACCACCGAGATCTACAC AACCGCAT ACACTCTTTCCCTACACGACGCTCTTCCGATCT CATCATAACGGTTCTGGCAAATATTC
pFLAG-CTC FWD-2	AATGATACGGCGACCACCGAGATCTACAC AAGGCCTT ACACTCTTTCCCTACACGACGCTCTTCCGATCT T CATCATAACGGTTCTGGCAAATATTC
pFLAG-CTC FWD-3	AATGATACGGCGACCACCGAGATCTACAC AGAGTGTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT GT CATCATAACGGTTCTGGCAAATATTC
pFLAG-CTC FWD-4	AATGATACGGCGACCACCGAGATCTACAC CACAAGTC ACACTCTTTCCCTACACGACGCTCTTCCGATCT CGA CATCATAACGGTTCTGGCAAATATTC
pFLAG-CTC FWD-5	AATGATACGGCGACCACCGAGATCTACAC CGTTCCTA ACACTCTTTCCCTACACGACGCTCTTCCGATCT ATGA CATCATAACGGTTCTGGCAAATATTC
pFLAG-CTC FWD-6	AATGATACGGCGACCACCGAGATCTACAC GCTTGGAT ACACTCTTTCCCTACACGACGCTCTTCCGATCT TGCGA CATCATAACGGTTCTGGCAAATATTC
pFLAG-CTC FWD-7	AATGATACGGCGACCACCGAGATCTACAC GTCAACAC ACACTCTTTCCCTACACGACGCTCTTCCGATCT GAGTGG CATCATAACGGTTCTGGCAAATATTC

pFLAG-CTC RWD-A	CAAGCAGAAGACGGCATACGAGAT AACCGGAA GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT A CTGTATCAGGCTGAAAATCTTCT
pFLAG-CTC RWD-B	CAAGCAGAAGACGGCATACGAGAT AGAGTGAC GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT TC CTGTATCAGGCTGAAAATCTTCT
pFLAG-CTC RWD-C	CAAGCAGAAGACGGCATACGAGAT CAACTGGT GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT CTA CTGTATCAGGCTGAAAATCTTCT
pFLAG-CTC RWD-D	CAAGCAGAAGACGGCATACGAGAT CGTTCGTT GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT GATA CTGTATCAGGCTGAAAATCTTCT
pFLAG-CTC RWD-E	CAAGCAGAAGACGGCATACGAGAT CTGTTCAC GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT ACTCA CTGTATCAGGCTGAAAATCTTCT
pFLAG-CTC RWD-F	CAAGCAGAAGACGGCATACGAGAT GCTTGCAA GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT TTCTCT CTGTATCAGGCTGAAAATCTTCT
pFLAG-CTC RWD-G	CAAGCAGAAGACGGCATACGAGAT GTCAACTG GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT CACTTCT CTGTATCAGGCTGAAAATCTTCT

Chapter 2. Effects of random sequences expressed in a population of eukaryotic cells

2.1 INTRODUCTION

The time when non-coding parts of eukaryotic genomes were considered to be junk DNA is long gone. We now know that non-coding sequences have functions beyond being a buffer for the effects of mutation and transposable elements, and are also key for the regulation of expression of genes through various mechanisms and for chromatin formation and stability (Bernardi, 2019). In the context of evolution, there are different levels of sequence conservation between different non-coding parts of the genome, which supports the idea that they can have different functions (Bird et al., 2006). An even more impressive function of non-coding regions of the genome became evident with the discovery of *de novo* genes. We now know that it is possible that non-coding sequences provide raw material for more than regulatory sequences, and that they may also be the source of new coding genes.

Even though it can be expected that ancestral mechanisms for the evolution of innovation are conserved across all organisms, it is also to be expected that different mechanisms could be specific to different domains of life. There are several examples of phenomena that depend heavily on unique features of an organism, as it is the case of horizontal gene transfer, being much more frequent in prokaryotes than eukaryotes (Keeling & Palmer, 2008). The fact that there are so many more non-coding sequences in eukaryotes than in prokaryotes suggests that *de novo* gene evolution might be one of these phenomena.

That *de novo* genes can be found in eukaryotic organisms, and that they are functional and even essential for the correct function of some physiological processes in some cases has been well documented (For a review, see (Van Oss & Carvunis, 2019)). Given the abundance of non-coding sequences in eukaryotes compared to prokaryotes, it is not surprising that most studies of *de novo* genes have been done in eukaryotic organisms. One could even argue that this mechanism of gene birth should be studied in eukaryote models where it is more likely to occur, just due to the abundance of raw material. In this way, *de novo* genes have been identified and characterised in different eukaryote model organisms including yeast (Li et al., 2014; Vakirlis et al., 2018), plants (Arendsee et al., 2014; Xiao et al., 2009), fruit flies (Heames et al., 2020), nematodes (Prabh & Rodelsperger, 2019), mice (Xie et al., 2019) and primates (Guerzoni & McLysaght, 2016; Knowles & McLysaght, 2009). Many more studies have identified young, taxonomically restricted genes as *de novo* gene candidates in an even wider range of organisms (Schmitz et al., 2020; Wang et al., 2020; Wissler et al., 2013). The protogene model proposed to explain the process of *de novo* gene birth was based on a study of the yeast Saccharomyces cerevisiae (Carvunis et al., 2012).

The likelihood that a sequence that has accumulated mutations neutrally, and could therefore be considered almost a random sequence, can be tolerated by a eukaryotic cell when it is expressed or becomes otherwise functional is unknown. As described in the first chapter for bacteria, random sequences have the potentials of being a useful tool to understand how *de novo* genes are born in eukaryotes. An interesting study of random sequences in yeast promoters showed that even short (80 nucleotides) random sequences have the potential of serving as promoters in yeast. The study presented in Chapter 1 of this thesis, connected to the previous study by Neme et al. 2017 (Neme et al., 2017), and similar studies using random sequences in bacteria (Chiarabelli, 2006; Keefe & Szostak, 2001; Knopp et al., 2021; Tretyachenko et al., 2017) have shown that a significant fraction of random sequences could be not only well tolerated, but even beneficial for bacterial cells. Beyond the study exploring the likelihood that a random sequence can function as a promoter in yeast (de Boer et al., 2020), there are no studies yet that have systematically looked at the possible functionality of random sequences in other eukaryotes. The complexity of eukaryote systems has probably been a deterrent to performing large scale studies of their tolerance to the expression of random sequences. However, from studies such as the mouse (Neme & Tautz, 2016), we know that most of the genome is being transcribed at some point, this being one argument for the existence of *de novo* genes in the first place. So, tolerance of eukaryotic cells to random sequences is expected to be at least as high as in the bacteria.

With this hypothesis in mind, in this chapter, I describe the effects of expressing artificial random sequences in a eukaryotic cell line. Specifically, I quantify the proportion of cells expressing specific clones in a library of random sequences, and how it changes over time. This study has three main goals: First, it aims to test whether eukaryotic cells respond to the expression of random sequences like bacteria. Second, to identify which molecular features of the sequences—if any—drive different effects of the random sequences on the cells. By answering these questions, it aims to provide a better understanding of different factors that could allow an ORF from a non-coding sequence to remain through generations as standing variation in a population of eukaryotic cells.

The experimental system used for this study is a library of random sequences expressed in a human cell line: HEK293. Each cell in the library is modified to express a single 174 nucleotide-long sequence with 150 random nucleotides flanked by two constant codons at the 5' end, and a 6-histidine tag at the 3' end. Each random sequence acts as a barcode, and this makes it possible to quantify the relative number of cells in the population expressing it using an amplicon sequencing approach. In this way, one can monitor how the proportion of individual sequences changes over time. The library of random sequences was cultured with doxycycline to induce expression of the peptides over a period of 20 days, sampled every 48 hours and sequenced to determine the proportion of cells containing each sequence. The experimental design is similar to that used before in *E. coli* (Neme et al., 2017), but using a eukaryotic model allowed me to incorporate the complexity and specific genomic features of eukaryotes. Other important improvements of this study include the targeted genome integration of the expression construct, a Kozak sequence to facilitate translation, shorter leading sequences on the 5'-end of the random sequence, and codon-optimised flanking sequences on both sides of the random sequence using frequently used codons in the human genome (Figure 2-1).

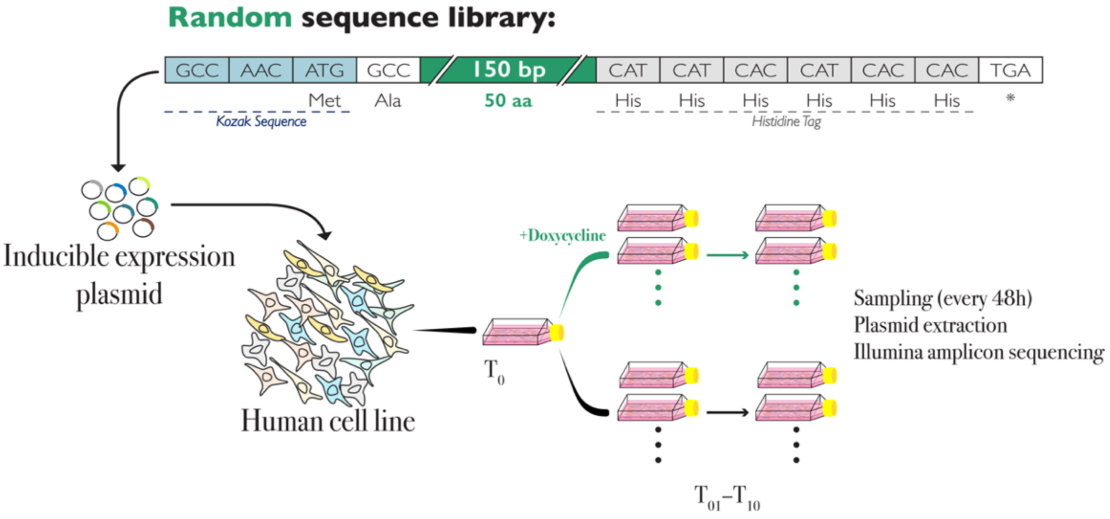


Figure 2-1. Experimental design.
The constant sequence flanking the random part of the oligonucleotides are depicted, including the Kozak sequence before the start codon and the 6-x histidine tag before the stop codon.

I used the Flp-In™ T-REx™ 293 cell line (FITR293, ThermoFisher Scientific). Flp-In™ T-REx™ is a protein expression system, selected for this study because of three key features: it permits the generation of stable expression cell lines; the expression construct is integrated into a specific target site on the genome; and expression is inducible with a Tet-On system (Figure 2-2). The FITR293 is generated by inserting two plasmids into the genome of the commonly used HEK293 cell line: the first one—pcDNA™6/TR—stably expresses the tetracycline repressor gene (tetR) and a Blasticidin resistance gene; the second one—pFRT/lacZeo—contains a Zeocin resistance gene with an FRT recombination site inside. The provider of the cell line provides no information about the location of either plasmid in the genome. The recombination of these plasmids into the genome can happen in any region, and requires testing to make sure that it is a single insertion without detrimental effects to the cells, and in a transcriptionally active site. The commercial cell line has already been tested by the provider so that it fulfils these conditions.

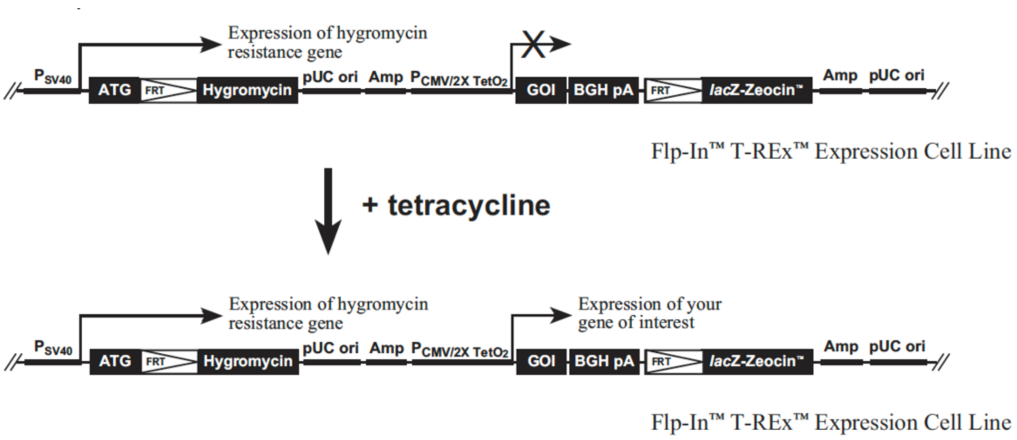

Expression of hygromycin resistance gene
P SV40
ATG
FRT
Hygromycin
pUC ori
Amp
P CMV/2X TetO2
GOI
BGH pA
FRT
lacZ-Zeocin™
Amp
pUC ori
Flp-In™ T-REx™ Expression Cell Line
+ tetracycline
Expression of hygromycin resistance gene
Expression of your gene of interest
P SV40
ATG
FRT
Hygromycin
pUC ori
Amp
P CMV/2X TetO2
GOI
BGH pA
FRT
lacZ-Zeocin™
Amp
pUC ori
Flp-In™ T-REx™ Expression Cell Line

Figure 2-2. Flp-In™ T-REx™ system. From the provider's Core Kit manual.

For the generation of the library, the oligonucleotides with the random sequences were inserted in the multiple-cloning site of a third plasmid—pcDNA5/FRT/TO. It contains a strong CMV promoter with two tetracycline operator ($TetO_2$) sequences, an FRT site for targeted recombination with the one already on the genome, and a Hygromycin B resistance gene (*hygB*) without a start codon. This plasmid was co-transfected with a pOG44 plasmid, which contains a FLP integrase gene. The integrase, expressed transiently in the cells after transfection, mediates DNA recombination between the FRT site already on the genome, and the one in the plasmid with the insert to be expressed. Successful recombination resulted in the *hygB* gene gaining a start codon from the pFRT/lacZeo plasmid, which gave the cells Hygromycin resistance for selection of successful clones.

Amplicon sequencing of the library at each timepoint was done using specific primers on the pcDNA5/FRT/TO plasmid. Changes in frequency were calculated, and sequences were assigned to three groups according to the direction of the frequency change in the population: UP, if they increased significantly ($p<0.05$), DOWN, if they decreased, or NS, if they had no significant change. In general, the percentages of sequences in each group are within the range of those observed in *E. coli* (See Chapter 1). However, unlike bacteria, it seems that length is not an important factor for whether a sequence is better tolerated by the cells.

A puzzling result of this study is that it appears that some feature of this library prevents the tetracycline induction system from functioning correctly. In other words, the library is permanently expressed regardless of the presence of tetracycline/doxycycline in the medium. These observations represent a first empirical approach at understanding how eukaryotic cells react to the expression of novel protein sequences and, therefore, attempt to shed some light over the way in which *de novo* genes can be kept in populations of cells as standing variation that might become useful for eukaryotic organisms.

2.2 RESULTS

The main challenge of this study's experimental design was to achieve stable and inducible expression of the random sequences in eukaryotic cells by integrating the expression construct into their genome. HEK293 cell lines are very well known and widely used in research, which means that there is plenty of information available about most of them, including genomic and transcriptomic resources (Lin et al., 2014). However, Flp-In™ T-REx™ 293 cells have been genetically modified by the provider, and there is little information made public about how they differ from the parental cell line. I characterised the cell line before generating the library with random sequences using the following assays: First, I generated a cell line with inducible expression of GFP to use as expression control; second, I titrated the tolerance of the cells to the selection antibiotic hygromycin B with an antibiotic kill curve. Third, I generated growth curves for the non-transfected Flp-In™ T-REx™ 293 cells and the GFP control. Next, I determined the minimum effective dose of doxycycline needed to induce expression; and, finally, I sequenced their genome in order to identify the location of the FRT site for recombination.

2.2.1 GFP CONTROL CELL LINE

A GFP control plasmid was successfully generated by introducing a GFP sequence into the pcDNA5/FRT/TO plasmid. Efficiency of transfection was 80 % to 90 % with all methods used. However, efficiency of integration into the genome was very low every time from 0.000045 % to 0.000079 %. Cells transfected with the GFP control plasmid are morphologically indistinguishable from non-transfected cells. Expression of GFP is detectable in less than 1% of cells even without an induction reagent (Figure 2-3).

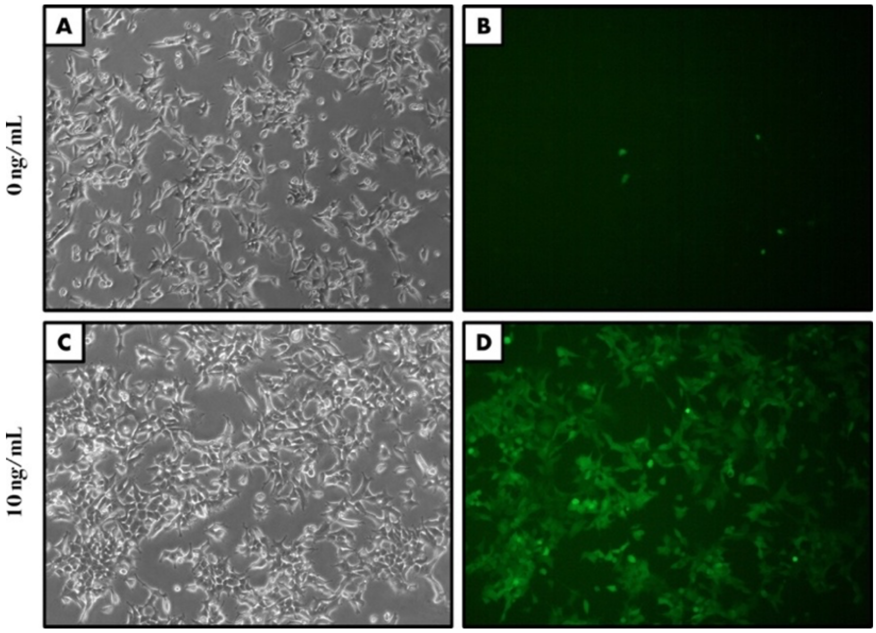


Figure 2-3. Inducible expression of GFP in FITR293 cells.
A–B. Cells with no addition of doxycycline to the medium. C–D. Induction of expression with 10 ng/mL doxycycline.

2.2.2 Hygromycin B kill curve

The hygromycin B kill curve was done as described in the methods section for six concentrations between 100 and 800 µg/mL of the antibiotic. Cells were killed completely within one week even at the lowest concentration used. Given the sensitivity of the cells to the antibiotic, and following the recommendations of the provider in the protocol, I decided to use the lowest concentration for the selection of transfectants in all experiments: 100 µg/mL.

2.2.3 Growth curves

The growth curve shown in Figure 2-4A shows some characteristic features of the cell line. Doubling times vary with cell density from 36 to 16 hours in the lowest and highest densities, respectively, with an average of 26 hours in normal culture conditions. Furthermore, as long as the media is being refreshed daily, the cells do not show contact inhibition as it would be expected, even when they have reached over 100 % confluence.

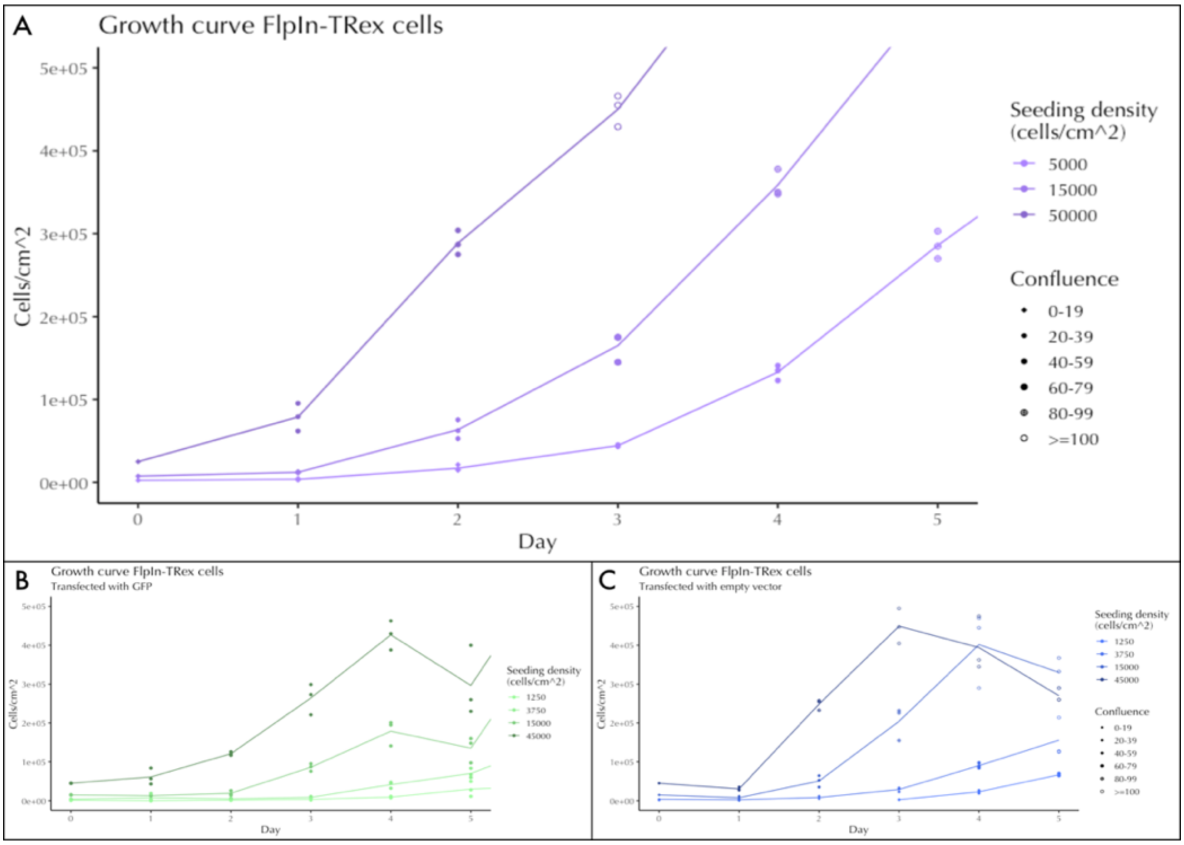


Figure 2-4. Growth curves FITR293 cells.
A. Cells growing in medium with Blasticidin and Zeocin, before transfection. B. GFP-control cells. C. Cells transfected with an empty pcDNA5/FRT/TO plasmid. Transfected cells are grown in medium with Blasticidin and Hygromycin. Growth curves B and C were generated with data partially generated by Jun Ishigohoka under my supervision.

Additional growth curves were generated for two controls, one transfected with the empty plasmid (pcDNA5/FRT/TO) and one transfected with a plasmid for GFP expression (pcDNA5/FRT/TO/GFP). These growth curves were obtained following normal culture conditions without refreshing the medium for cells after they have reached 100 % confluence. Cells transfected with the empty vector, which has the FRT insertion site but does not have an insert to be expressed in the cells, show the same growth pattern as the original cell line, with similar doubling times at similar concentrations. However, without refreshing of the medium, the cells quickly reach confluence, and a stationary phase followed by cell death (Figure 2-4.B). On the other hand, cells transfected with GFP grow slower than non-transfected cells and the empty vector control, even without induction of expression (Figure 2-4.C). Absence of GFP expression was confirmed through cell counting using a Countess II-FL equipped with a green filter, and with FACS cell sorting. In both cases, less than 1% of cells express GFP in the absence of doxycycline.

2.2.4 Doxycycline dose determination

The effect of doxycycline in protein expression was quantified using the GFP control cell line. The results of the doxycycline curve can be seen in Figure 2-5 and Table 2-1, which show the results of two independent tests of the reagent on the cells. Similar GFP expression could be seen under the microscope for doxycycline concentrations of 10 and 50 ng/mL added directly to the growth medium. Higher concentrations, over 100ng/mL, have a negative effect on vitality of the cells, evident by the change in morphology of the culture.

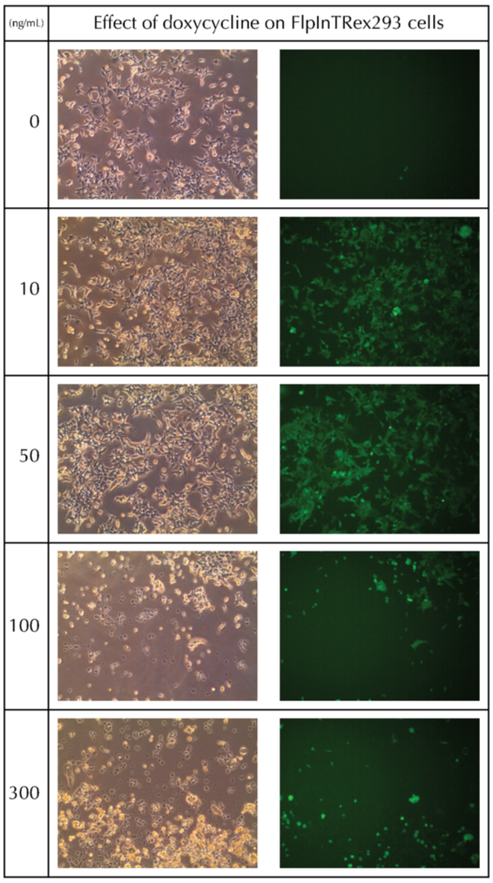


Figure 2-5. Effect of doxycycline on GFP expression and cell morphology.

Left: Bright-field. Right: Cells observed under blue light. Concentrations up to 50 ng/mL of doxycycline have no noticeable negative effects on the cell. Higher concentrations, over 100 ng/mL decrease cell viability and cause changes in cell morphology.

Detailed quantification of the effect of doxycycline on expression showed that addition of amounts as small as 1 ng/mL already induced expression in over half of the cells. The fraction of cells expressing GFP was above 80 % for all concentrations above 3 ng/mL. For concentrations between 10 ng/mL and 50 ng/mL, the fraction increases to 91–98 % in all samples (Table 2-1). Based on these results, I decided to use 10 ng/mL of doxycycline as the lowest effective dose at which over 90 % of the cells express GFP while preserving their viability.

Table 2-1. Expression of GFP in FITR293 cells with different concentrations of doxycycline. Highlighted is the selected concentration for the experiments.

Doxycycline (ng/mL)	Fraction of cells expressing GFP	Vitality (%)
0	0.00	100
1	0.59	95
3	0.80	93
5	0.97	95
10	0.91	95
30	0.98	92

2.2.5 Location of FRT site in the FITR293 genome

Whole genome sequencing of the control GFP cell line yielded an average coverage of 37X. By selecting discordant reads mapping to the pFRT/lacZeo plasmid I was able to identify its insertion site into chromosome 12, located in the ERC1 gene (ENSG00000082805), most likely inside an intron. The ERC1 has 30 exons, 25 splice variants, 15 of which could be affected by the introduction of the plasmid. The region is 0.5 Mb long. The longest transcript is 9 kb and it is not affected by the insertion, which is likely to be between positions 1.200.935 and 1.222.857 (Human genome assembly GRCh38.p13). It was not possible to identify reads mapping exactly to the insertion site of the plasmid because the provider of the cell line does not provide detailed information about the restriction enzyme used to linearize the pFRT/lacZeo plasmid, and there are homologous sequences in the pFRT/lacZeo and the pcDNA5/FRT/TO plasmids, which causes ambiguities in the mapping of the paired reads.

2.2.6 Generation of a library of random sequences in FITR293 cells

A pool of random sequences as described in Figure 2-1 was cloned into pcDNA5/FRT/TO plasmids (see Methods, page 91). The pool of plasmids was successfully used for transfection of the FITR293 cells. The cells are morphologically indistinguishable from the parental cell line and they have approximately the same doubling time, judging from the time required between passages. Some signs of stress, such as large, granulated cells were visible after all transfections, regardless of the insert.

The starting library was sequenced using Illumina MiSeq. The sequences were dereplicated (Methods, page 97) to obtain a database of all unique sequences present in the library of cells. The resulting database was composed of 3708 different clones. The distribution of predicted peptide lengths in the library matches the expected distribution of a library of random sequences of this length (Figure 2-6). This means that it matches the expected probability of obtaining stop codons at every position in the sequence without having obtained a stop codon in the previous positions, defined as the probability mass function of a geometric distribution with *p*=3/64.

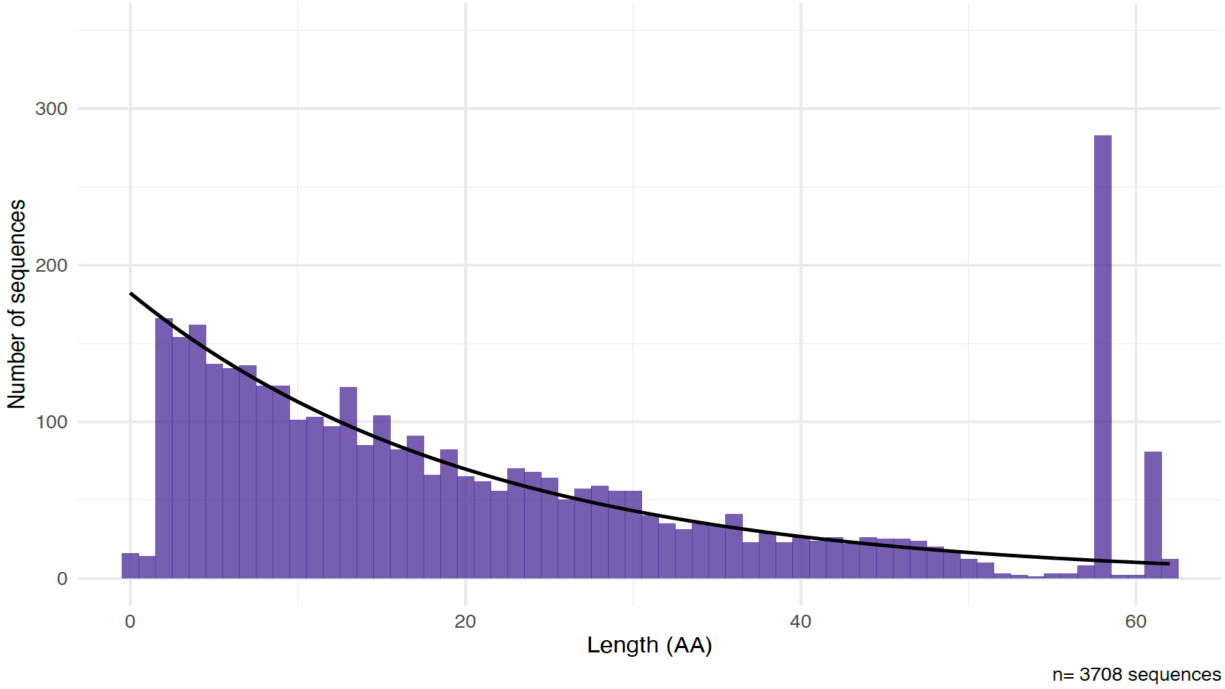


Figure 2-6. Distribution of predicted peptide lengths in the library.
The black line corresponds to the probability mass function of a geometric distribution with p=3/64, considering the probability of having a stop codon at each position without having one in any previous one in the sequence before.

Other molecular features of the nucleotide and predicted peptide sequences also match the expected distributions of values for random sequences. For example, GC content of both the full-length reads and the corresponding predicted ORFs is narrowly distributed around 50%, with means of 49.7 and 51.9, respectively (Figure 2-7A). When looking at the average frequency of each nucleotide at each position only in the random part of the sequence, there is a slight bias towards a higher content of thymine in all positions, closer to 0.38 instead of the expected 0.25. This bias results in a slightly lower frequency of adenine and guanine, but does not seem to be strong enough to affect other features of the library.

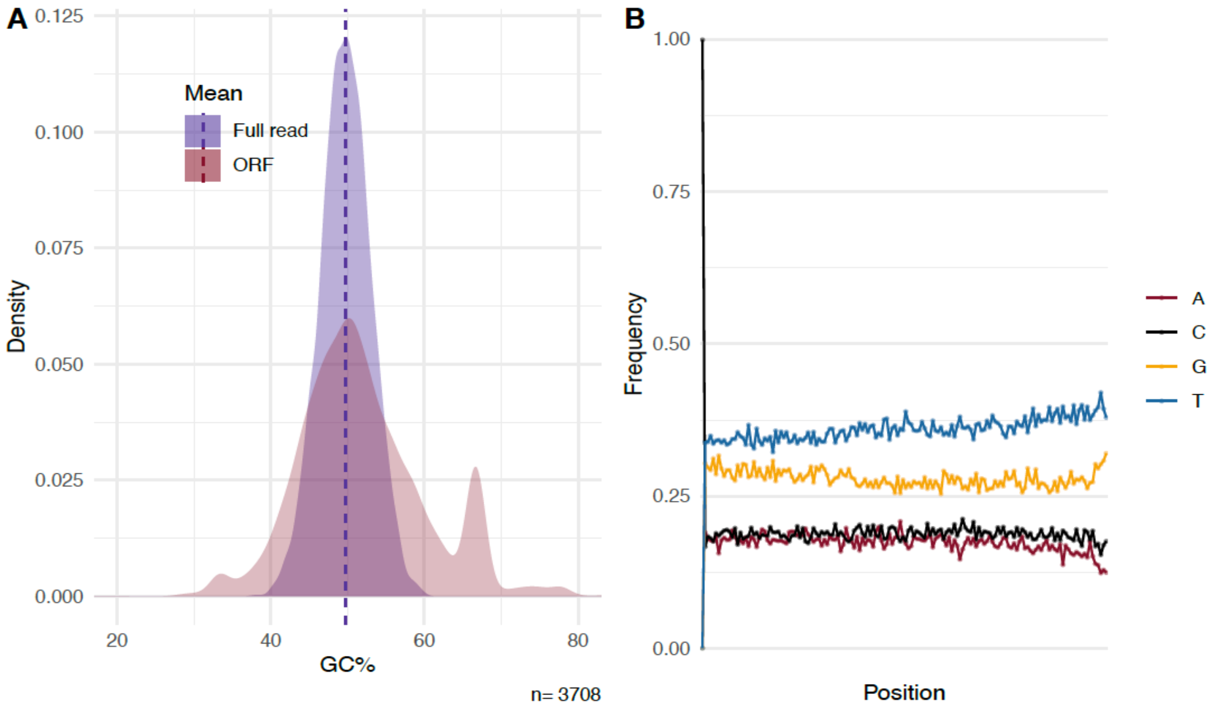


Figure 2-7. GC content of sequences in the library of random sequences.
A. GC content distribution for full-length reads (blue, mean = 49.72%) and predicted ORFs in the database (red, mean = 51.94%). B. Average frequency of nucleotides at each position of the random part of the sequence (150 nucleotides).

The main two features studied for the predicted peptides were intrinsic disorder and aggregation propensity. Intrinsic disorder was calculated as the average intrinsic disorder score of all residues in a sequence, calculated with the -short option of IUPred2A. Disorder scores are calculated based on features of each amino acid and their predicted interactions. Length is an important factor in these calculations and the software used cannot reliably assign disorder scores for peptides shorter than 30 residues. However, this analysis provides a good idea of the correlation between length and intrinsic disorder, and of GC content and intrinsic disorder (Figure 2-8).

In the case of aggregation energy, a commonly used predictor of the likelihood that a sequence will aggregate in the cell and form amyloids, there is also a clear correlation with length (Figure 2-9). More sequences in the longest peptide length categories are predicted to be prone to aggregation.

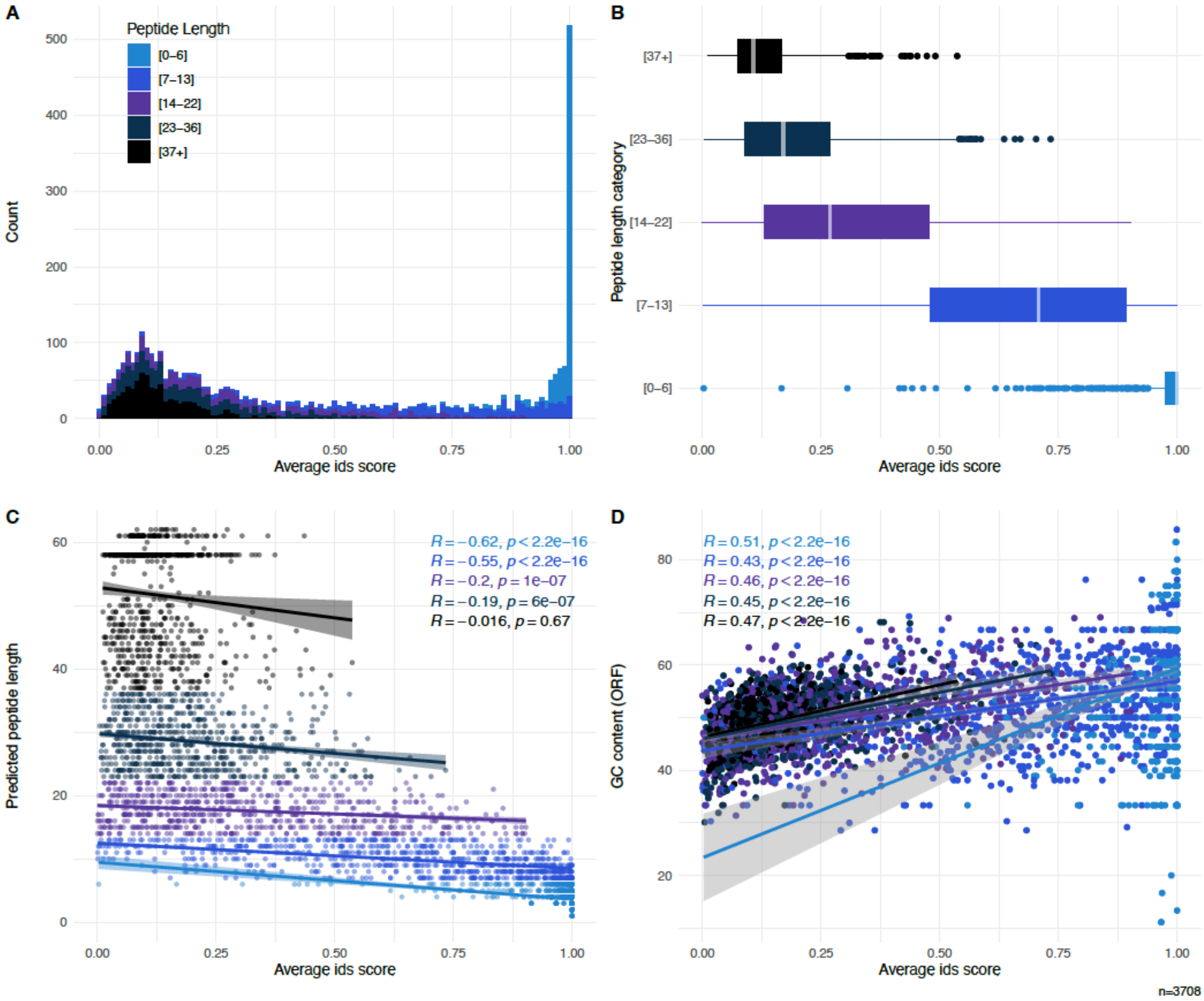


Figure 2-8. Intrinsic disorder of predicted peptides in the database, binned by length.
A. Distribution of average intrinsic disorder scores. B. Boxplot of average intrinsic disorder scores for peptides in each length category. C. Correlation between peptide length and intrinsic disorder. D. Correlation between GC content and intrinsic disorder.

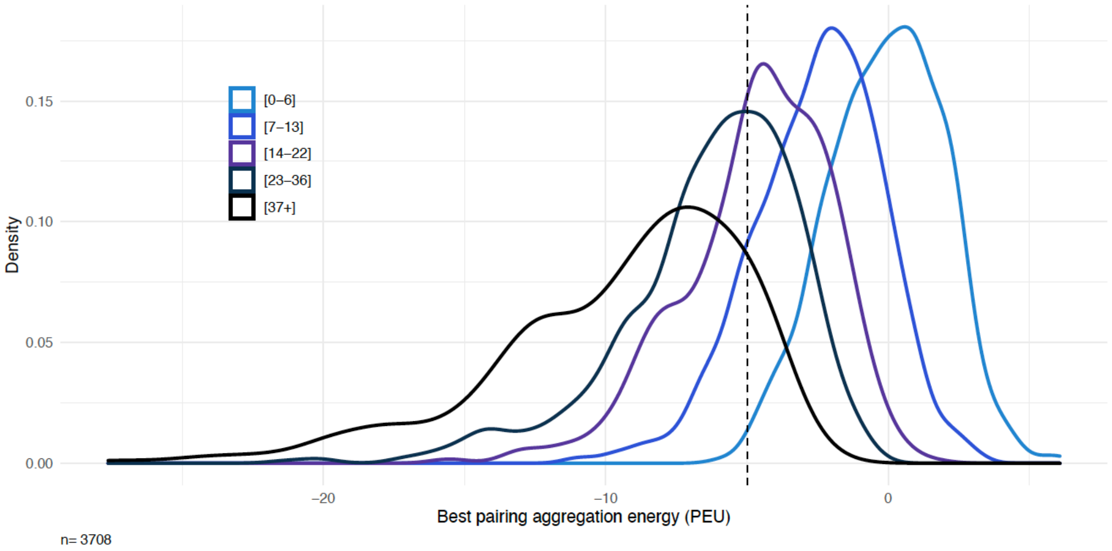


Figure 2-9. Distribution of aggregation energies for predicted peptides in the database.
Aggregation energies are calculated with the software PASTA2.0, and expressed as "pasta aggregation units" (PEU). Peptides with PEU of -5 or lower are considered to be prone to aggregation.

2.2.7 Sampling and sequencing

For the experiment, the library was passaged three times after thawing, before seeding 10 flasks with the same number of cells. Flasks 1 to 5 were assigned to the control group, which was passaged without addition of doxycycline to the medium. Flasks 6 to 10 were assigned to the induced treatment group, and doxycycline was added to these flasks at each passage to a final concentration of 10 ng/mL. The resulting 10 populations were maintained in parallel passaging them every two days for a total of 10 passages. At every passage, samples of the population were obtained and flash frozen with liquid nitrogen to be preserved until DNA extraction. As an indirect control of protein expression, a population of GFP-control cells was assigned to each group and treated in the same way as the library. GFP expression was visually confirmed for the induced group at each time point throughout the experiment, while no expression was detected in the control group.

All samples were sequenced in one Illumina NextSeq run. Samples had on average 696,779 paired-end reads. Unfortunately, sequencing failed for four samples: three in the control group at the first time point (replicates 2, 3 and 4), and one in the induced group at the fifth time point (replicate 6). Three samples were available from time point 0 (seeding), and they were used as controls for the initial analyses instead of the failed control samples, effectively replacing replicates 2, 3, and 4 of the control group in time point 1 with samples 1, 2, and 3 of time point 0. Merging, trimming and selection of valid reads was done for all sequencing files successfully. Between 85 % and 92 % of reads were accepted after trimming, and over 99 % of all trimmed forward reads contained the defined flanking sequence (clean reads).

2.2.8 Mapping and DESeq2 analyses

On average, 86 % of clean forward reads were successfully mapped to the database (SD = 3.6 %). The remaining 14% (SD= 3.5 %) were not found in the original database sequenced with MiSeq, suggesting that the library was not sequenced to saturation. This could be because the coverage in the MiSeq sequencing was not high enough for detection of clones with lower frequencies. Mapped reads were used to generate count tables, which were, in turn used to find the fold-change of each clone in the database throughout the experiment.

Count tables for each time point were compared to time point number 1 using an experimental design formula in DESeq2. An initial multifactorial analysis, combining *Time* and *Treatment* as explanatory variables (Analysis 1, design = ~Treatment + Time), revealed that only 74 clones (2 %) in the library behave differently between the control and the induced groups (Table 2-2, Analysis 1). 40 out of those 74 clones can be detected as significantly different between treatments even when comparing the first 2 timepoints, i.e., they showed differences from the start of the experiment. When analysing the samples in the induced and control groups separately, however, one can see that over 48 % and 40 % of clones differ in frequency in the cell population between the last and the first time point of the experiment, respectively (Table 2-2, Analyses 2 and 3). Interestingly, this means that regardless of the addition of doxycycline, almost half of the clones have a consistent frequency change across replicates (see section below).

Table 2-2. Summary of results for different DESeq2 analyses.
Analyses were done for 10 time points (T01–T10) with 5 replicates per time point, except for T05-6 (Induced). The results of analysis including both time and treatment as explanatory variables are highlighted.

	Analysis	No. of replicates	No. of clones	% Non-Significant$^{(1)}$	% Significant$^{(1)}$	% UP$^{(1)}$	% DOWN$^{(1)}$
1	Time + Treatment	99	3708	98.00	2.0	0.78	1.21
2	Group: Induced	49	3670	51.55	48.45	11.66	36.78
3	Group: Control	50	3660	59.59	40.41	5.98	34.43
4	Time only (mixed treatments) $^{(2)}$	99	3708	50.13	49.87	6.88	42.99

(1) $p_{adj} < 0.05$. Significant fold change between T10 and T1 of the experiment.
(2) For analyses combining both treatments, minimum p_{adj} required for assignment was 0.01.

Out of the 1780, 1479, and 1849 significant clones detected in analyses 2, 3 and 4, respectively (Table 2-2), 1089 clones, corresponding to 61 %, 74 % and 59 %, are found in all analyses with the same trend of change in frequency. There are only 6 clones that show contradictory trends between the treatments. From the remaining clones, 836 are found in analysis 4 and at least one of the others (Figure 2-10).

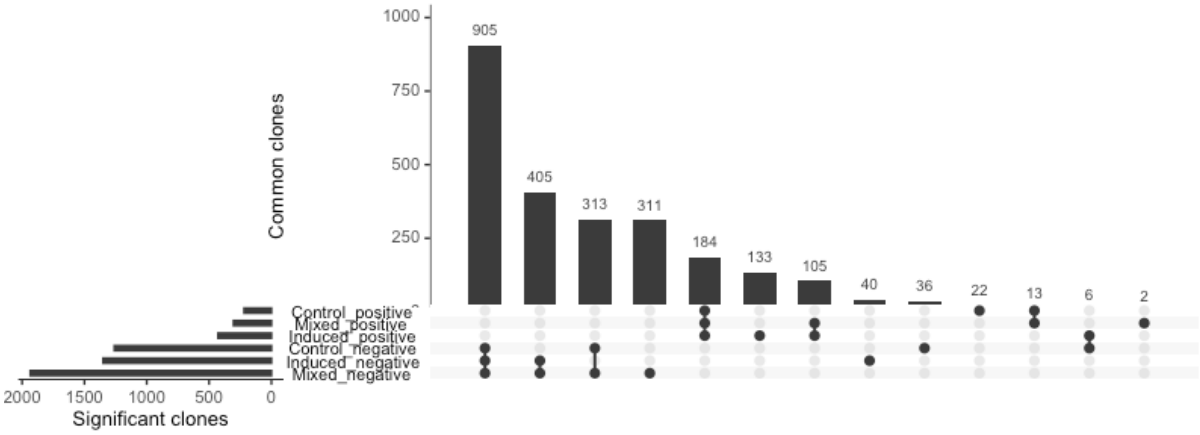


Figure 2-10. Comparison between significant clones detected in the different analyses.
The total number of clones in each category is shown on the bar plot on the left. The intersections between the different groups are represented by the bottom plot, and the number of clones in each intersection is shown on the top bar plot.

Regardless of the reason for the lack of a major difference between control and induction, it is safe to say that the samples without induction do not serve their purpose as a control. However, it is still possible to analyse these samples as a parallel experiment with constitutive sequence expression. With this in mind, the two experiments, with and without doxycycline, were further analysed in parallel, and also combining all samples as replicates at each time point (Table 2-2). The clones could be assigned to three different groups depending on whether they increased, decreased or did not change significantly in frequency between the first and last time points of the experiment. The groups were called UP, DOWN and Non-Significant (NS), respectively, and a flag was added to each clone in the database to indicate the group assignment. Between 34 % and 43 % of the sequences decrease in frequency in the population during the experiment, between 6 % and 12 % increase in frequency, and between 40 % and 50 % show no significant changes for each of the groups of samples analysed. Interestingly, these percentages of clones assigned to each group are within the range of those assigned to each group in the bacteria library experiments described in chapter 1.

The analyses of all samples combined (Analysis 4 in Table 2-2), regardless of the treatment, shows that there are some clones changing in frequency throughout the experiment. The progression in the change of frequency of clones can be seen in Figure 2-11, where the mean base counts for each gene are plotted against the log2 of the fold change, and coloured if the fold change is significant in either direction. As time progresses, the changes in fold change increase in magnitude, which indicates that the trends used for the group assignment are real.

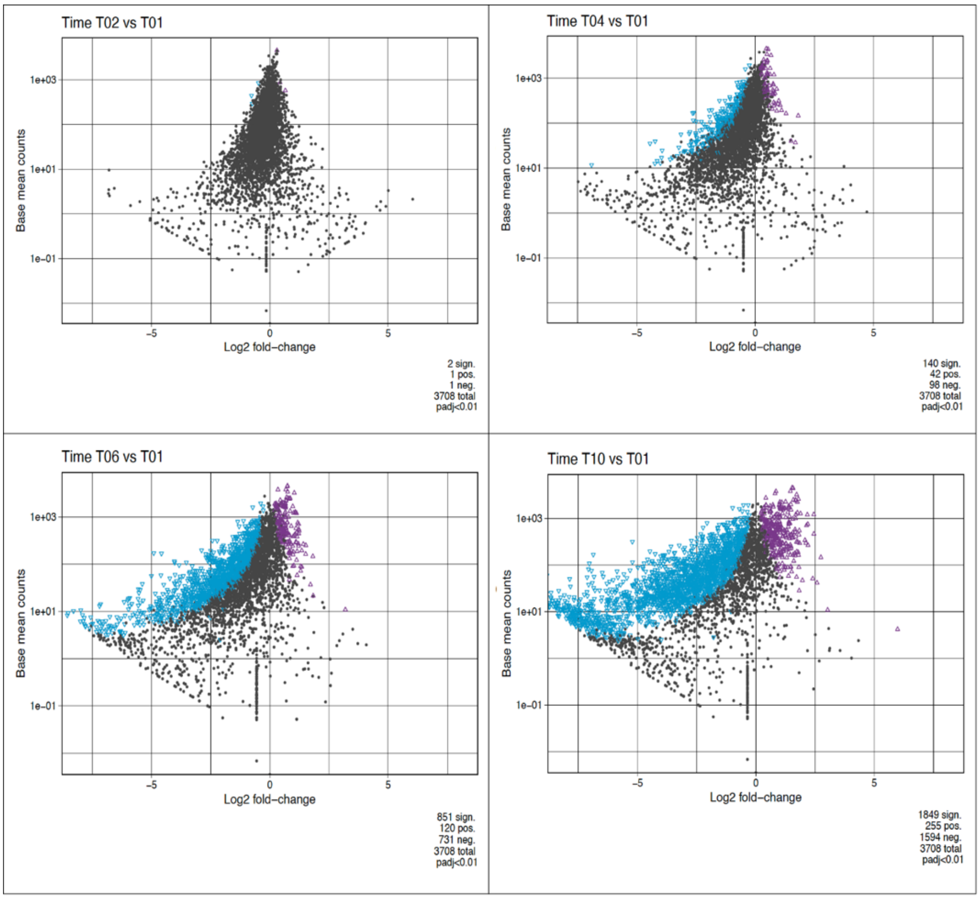


Figure 2-11. Base mean counts vs. log2 fold-change plots.

Shown here are timepoints 2, 4, 6 and 10 compared to timepoint 1 combining all samples regardless of treatment. Purple: Significant, positive change in frequency. Blue: Significant, negative change in frequency. Black: Non-significant change in frequency. p_{adj}*<0.05.*

Using the group assignment flag, it was possible to compare some molecular features of the sequences in each group against the full database. Interestingly, unlike the clear correlation of length with group assignment that was observed for the bacteria library (see chapter 1), none of the features evaluated seem to be a determining factor for assignment of a sequence to a specific group. For all features studied—length, GC content, average intrinsic disorder, and aggregation propensity—the distributions of sequences are indistinguishable of each other (Figure 2-12).

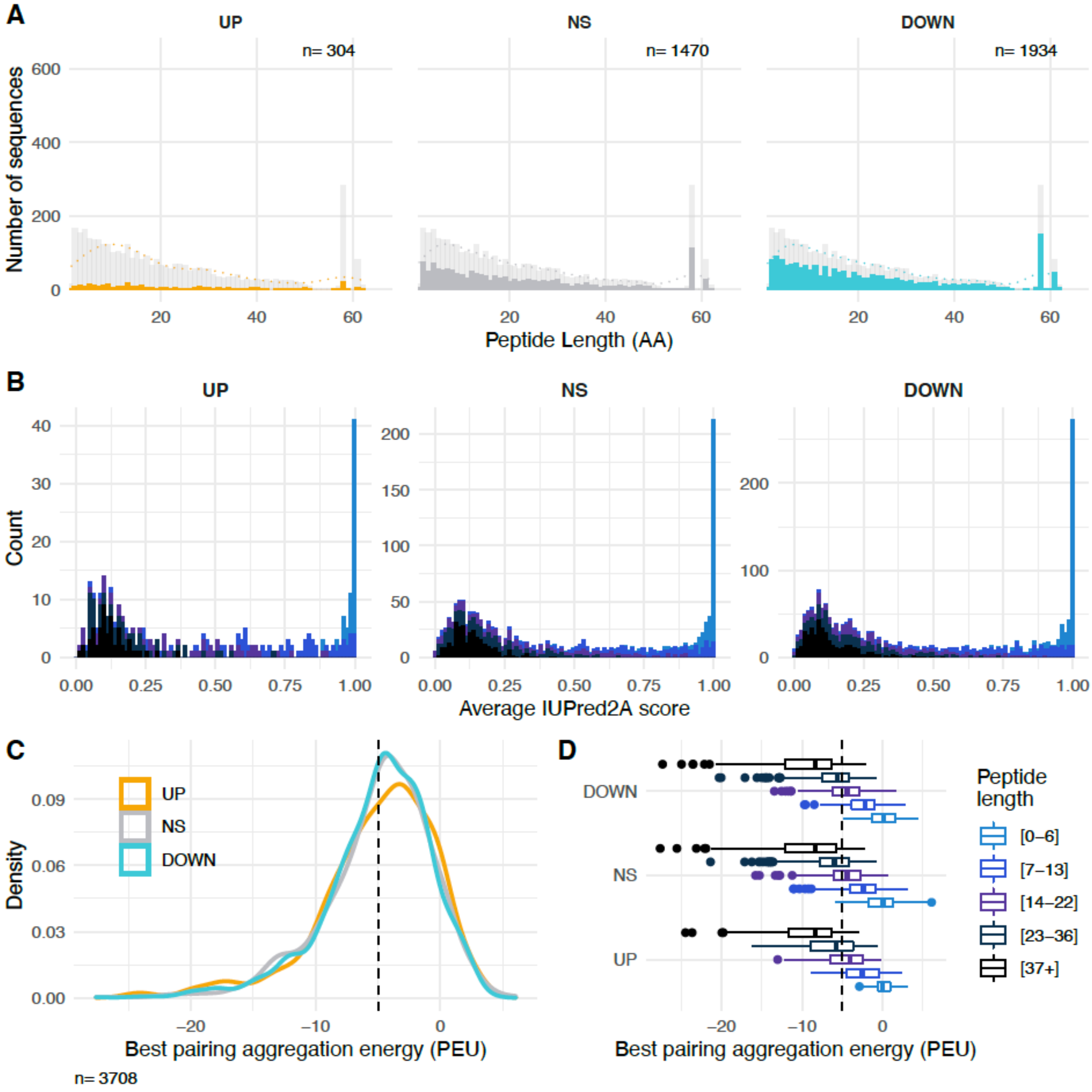


Figure 2-12. Features of sequences in each group.
A. Sequence length distribution in each group. B. Average intrinsic disorder scores for sequences in each group. C. Aggregation energies for all sequences in each group and D. divided by length categories

2.2.9 Possible reasons for the lack of regulation of the induction

In principle, there are two possible reasons for getting such similar results between the induced and uninduced samples: either induction did not work in this system and the changes of frequency are the product of drift, or expression of all sequences was permanently induced and the changes of frequency are a result of an effect of each peptide on the cell that expresses it. The first explanation is unlikely, given that the sensitivity of the cells to doxycycline was previously confirmed using a GFP control cell line, where even 1 ng/mL of doxycycline in the medium was enough to induce expression of GFP in at least 50 % of the cells (See Figure 2-3). Furthermore, a GFP control cell line was generated as a positive control in a transfection at the same time as the library. Two flasks of this cell line—one with and one without doxycycline—were maintained throughout the experiment, passaged at the same time and with the same reagents, and visually inspected for GFP expression at each passage. GFP expression could be confirmed at every time point only in the flask with doxycycline added.

Following the counts of individual clones in different replicates throughout the experiment indicates that the observed changes in frequency are not only the product of drift (Figure 2-13). All replicates follow the expected trends both for the clones identified as significantly different between treatments (Figure 2-13, A and B) and the ones that behave the same in both treatments (Figure 2-13, C and D).

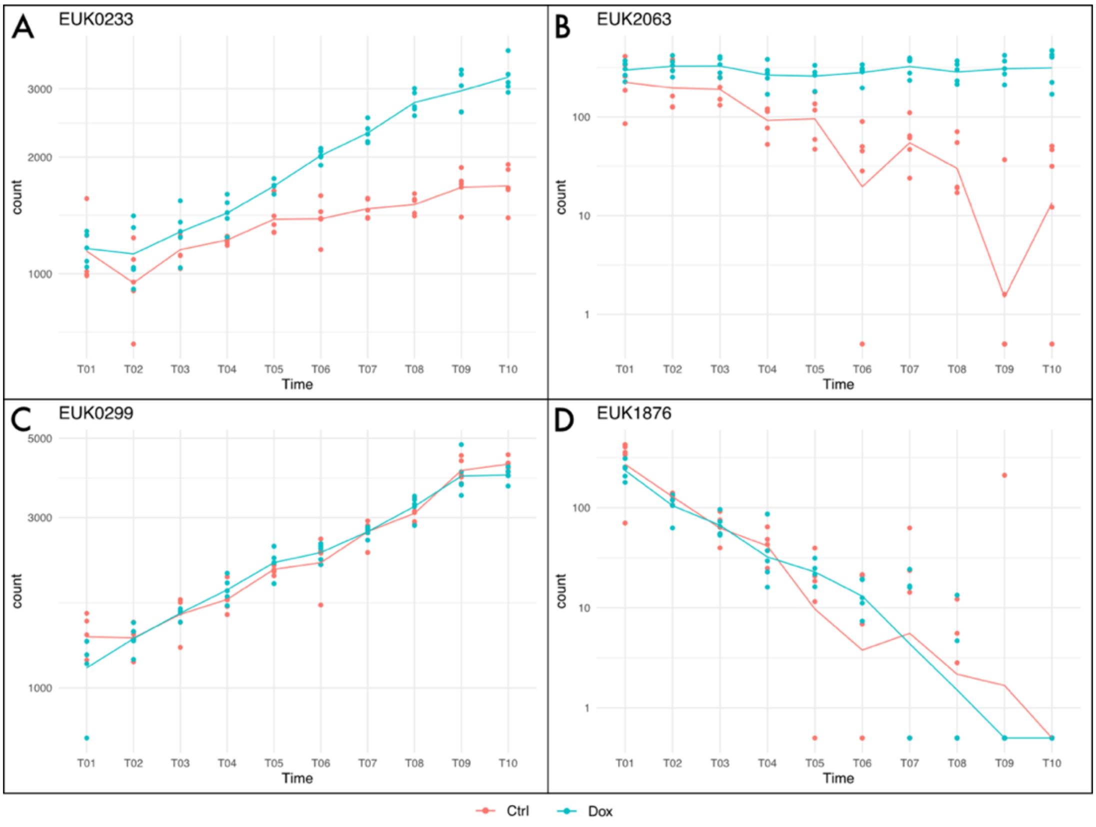


Figure 2-13. Example trajectories of clones in the library.

A–B Top two significant clones with different effects between treatments (p-adjusted <0.05). C–D Top two significant clones increasing and decreasing in frequency through the experiment regardless of treatment (p-adjusted <0.05).

Relative measurements of the random sequences' RNA levels supports the more likely alternative, that the sequences are permanently expressed. Cells were incubated with different concentrations of doxycycline to induce expression of the library, total RNA was extracted from each sample, retrotranscribed and amplified using specific primers binding 37 bp upstream the random sequence and directly on the 6 x Histidine tag downstream the random part of the sequences. This experiment was done independently of the time course experiment. The amount of RNA was normalised before retrotranscription, and three different housekeeping genes were used as reference for the amount of cDNA. The results indicate that expression levels of the random sequences in the library do not change as expected with the addition of different concentrations of doxycycline (Figure 2-14). In particular, the signal detected without doxycycline added suggests an unexpectedly high baseline expression of the library clones.

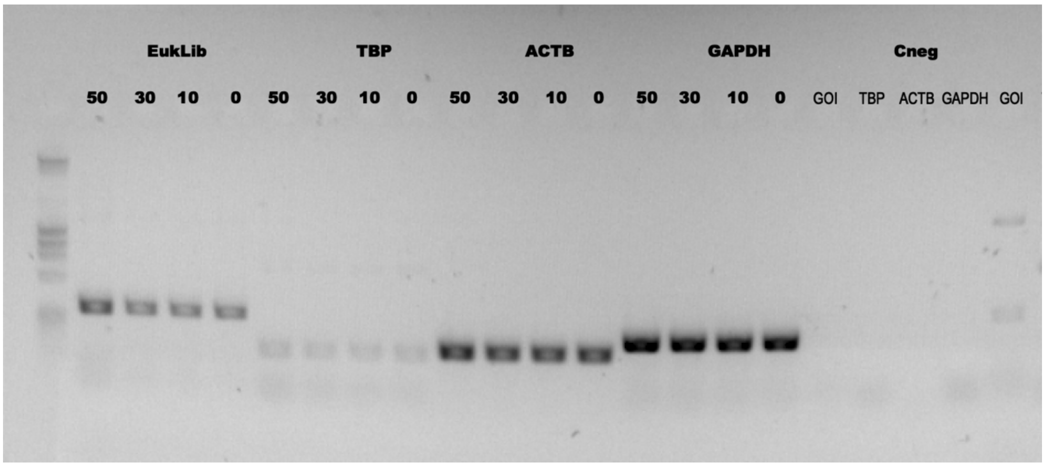


Figure 2-14. RT-PCR of the random sequence library (EukLib) and three housekeeping genes (TBP, Actin B and GAPDH). Cells were cultured with four different concentrations of doxycycline (50, 30, 10, or 0 ng/mL), RNA was extracted for cDNA synthesis and PCR. Cneg: Negative control using water instead of DNA as template.

A common source of error for these types of experiment is contamination of the medium with tetracycline or doxycycline and a common source for this is the FBS. However, all reagents used throughout this experiment were certified tetracycline-free, and such a contamination would have been detected in the GFP control cell line with amounts as small as 1 ng/mL (Figure 2-3). I have thus discarded this hypothesis as a possible explanation of the results.

A second possible source of error in the correct functioning of the induction in the system are mutations introduced into the plasmid by the bacteria in which they are amplified. If a mutation has altered the tetR binding site on the CMV promoter, repression would be diminished or completely nullified. In order to test this hypothesis, I sequenced both the empty plasmid and the library plasmid used for transfection of the cells. In both cases, I was able to confirm that the plasmid sequence had not been altered in the process of amplifying the plasmid in *E. coli*.

The third hypothesis explored is the possibility that more than one plasmid got inserted into the genome of the cell, either by successive recombination with the FRT site, or by spurious recombination into other sites. The presence of multiple competing binding sites for the repressor tetR could reduce the strength of repression and allow for "leaky" expression. Some evidence supporting this hypothesis was the presence of higher molecular weight products of the amplicon sequencing PCR for some of the samples. To test this, I isolated single cells from the population through cell sorting, allowed them to grow as populations of a single clone and sequenced their inserts with the Sanger method. From the chromatograms, I confirmed that 25 out of the 69 clones that could be isolated could contain two or more inserts (Figure 2-15). However, it can also not be excluded that the sorting that should have created single cells did not fully work as expected, although precautions were applied (see Methods).

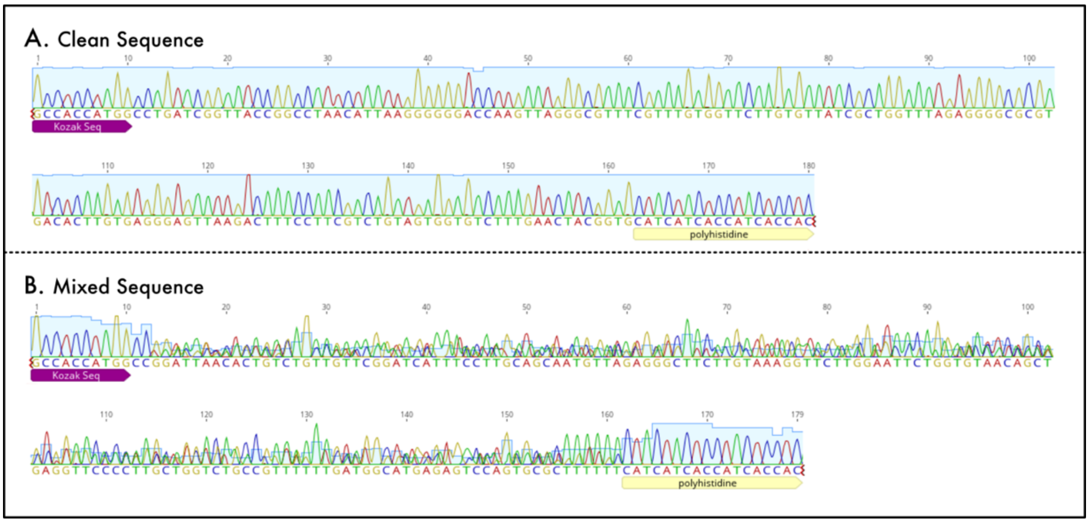


Figure 2-15. Example of single clone sequencing results.
Cells with a single plasmid insertion into the genome yield a "clean: chromatogram with clear peaks at each position.

It is hard to determine whether the high fraction of clones (36 %) with multiple insertions reflects their real frequency in the population due to the low survival rate of the sorted Flp-In T-REx 293 cells. These cells did not grow well at very low densities, and the recovery rate from the single cell sorting was extremely low (about 14 %). It is expected that depositing a single cell in a well of a 96-well plate would impose a great stress on them.

2.3 Discussion

The goal of this study was to use a library of random sequences expressed in a eukaryotic cell line to determine their tolerance to the expression of novel RNA and/or peptides. The random artificial sequences would allow us to get an idea of the fate of sequences that have not been under selection, or have been under weak selective pressures, as is the case of non-coding regions of the genome.

I successfully generated a library of random sequences in a Flp-In™ T-REx™ 293 (FITR293) cell line, derived from HEK293 cells, commonly used as a heterologous expression system. I chose the FITR293 cell line because of the possibility to integrate the expression plasmid into the genome and to induce the expression of the library with the addition of doxycycline. Given that there is little information publicly available about the cell line itself, my first task was to characterise it. I generated growth curves for the original cell line, as well as for each of the derived cell lines that I transfected. This showed that all transfected cells grow at a slower rate than the original cell line, but there is not much variation between the different inserts used for transfection. Whole genome sequencing of the cells showed that the FRT site that serves for the targeted insertion into the genome for this particular lot is located in a transcriptionally active region and does not seem to truncate any existing gene. Using a GFP plasmid, I generated control cell lines for my experiments that allowed me to standardise and visually monitor the induction of expression using doxycycline. Sequencing the library showed that it has all the features expected from a random sequence library and the expected correlation between nucleotide features such as GC content and length with peptide features such as intrinsic disorder and aggregation propensity.

Through sampling and sequencing of the library during a time-course experiment of about 40 cell divisions, I was able to monitor clones expressing different sequences as their frequency changed through the experiment. Assigning clones to groups according to this change resulted in three groups of sequences, increasing, decreasing or not changing significantly in the population. The percentage of sequences assigned to each group was comparable to what was reported for bacteria in Neme et. al 2017 and chapter 1 of this thesis. It is notable that nothing indicates that there is a correlation between the assignment of a sequence to a specific group and its length, GC content, intrinsic disorder, or aggregation propensity. Furthermore, it is interesting that over half of the sequences are kept in the populations without significant changes in frequency throughout the experiment.

An unexpected result of this study is that the repression of expression mediated by the TetON system in the cell line does not seem to work for this library. External sources of error, such as reagents and manipulation, were tested and rejected as possible causes. Therefore, it must be a property of the library itself causing the permanent expression.

One possibility is that, due to the scaling up of transfection to obtain as many different clones as possible using large amounts of plasmid, more than one plasmid was inserted into the genome of the cells. A preliminary analysis sorting individual cells and sequencing them indicates that this could be the case. Since the repression of expression depends on Tet repressor molecules binding to the CMV promoter, multiple insertions diminish repression by decreasing the number of available Tet repressor molecules in the cell. This could result in incomplete repression, or permanent induction of expression of up to a third of the clones in the library. Unfortunately, with the available data it is not possible to rule out the possibility that this result is an artifact product of more than one cell inadvertently sorted into the same well. However, this effect alone is insufficient to explain why only 2 % of the sequences show a difference between the induced and non-induced samples.

The puzzling inability of controlling expression in this system requires further study in order to rule out all possible causes. Here, I will discuss two more hypotheses that could explain these results, but that would require further testing and follow-up studies. The first one is that the peptides expressed by some of the clones could be affecting the system by, for example, binding to the tetR molecules and—either directly or allosterically—reducing their affinity to the binding site on the promoter (Goeke et al., 2012; Klotzsche et al., 2005). This would require, first, that the clones produce enough of these molecules to affect most of the culture; second, that the cell expressing such molecules somehow released them into the medium; and third, that the other cells somehow uptake the peptides into the nucleus of the cell. Although it is unlikely that all of these conditions are fulfilled, it would be simple to test this hypothesis by either “transplanting” conditioned medium where the library is growing to a culture of the GFP cells, or spiking the library culture with GFP cells directly.

A second and, admittedly, more speculative hypothesis is that random sequences could be encoding regulatory sequences that initiate their own transcription. Experiments done in bacteria and yeast have shown that random sequences can easily act as promoters and initiate transcription of reporter genes (de Boer et al., 2020; Yona et al., 2018). In the case of yeast, a striking 83 % of all random sequences used in the experiment yielded expression of a reporter gene, and the authors of the study suggest that the proportion could be even higher in mammals, where promoters are more variable. There are also well documented cases of regulatory sequences shifting between being promoters and enhancers (Carelli et al., 2018).

If it is true that most random sequences can initiate their own transcription, even at low levels, the results presented here can be explained. It would also mean that the effect of sequences as short as the ones presented here is strong enough to drive changes in frequency in a population of cells. Traditional analyses of transcriptomes tend to remove very short transcripts from their analyses due to the likelihood of spurious transcription. However, this most likely results in loss of relevant information to understand physiologic and evolutionary processes. Although the importance of studying short proteins has been stressed in several publications (Mackowiak et al., 2015; Orr et al., 2020; Storz et al., 2014), we still have ways to go in order to reach a full understanding of the mechanisms that drive evolution of shorter sequences.

The hypothesis presented here would also help to explain findings of pervasive expression of most of the genome in mice, humans, yeast and fruit flies (Berretta & Morillon, 2009; Jacquier, 2009). Expression of short sequences that could have a negative effect on the cells would expose them to selection to be removed from the genome. Simultaneously, the high tolerance of the cells to random sequences would allow them to accumulate mutations in a neutral way. Although these two claims might sound initially contradictory, the vast sequence space occupied by even short sequences would permit selection to act, while keeping the total set of sequences close to being random through accumulation of neutral mutations. A good example of this is the fact that random sequences assigned to each group in this experiment seem to be a random sample of the starting population. The sheer size of the possible combinations of unique sequences—2.04x10^{90}, in our study—combined with the robustness of the cells to disruptions and the lower information content of regulatory sequences in eukaryotes makes the probability of rare events much larger than it would be otherwise.

The results presented here open the door to further research possibilities and improved models of the birth of *de novo* genes. It is important to be aware, however, that there are several points that should be evaluated carefully, and could be improved in future research, which will be discussed below.

First, it is necessary to test similar expression systems in other types of cells or organisms. HEK293 cells have been shown to have genomic instabilities that result in mutations or changes in the expression profile that, in turn, generate changes in expression profiles and even the phenotype of the cells (Stepanenko & Dmitrenko, 2015). Although they are commonly used for heterologous expression of proteins, they are far from being a model organism for experimental evolution. The stress of transfection, even with methods that have large viability rates, can also cause differences in transcriptome and genotype, which could be biasing the results.

Another possible source of bias is the fact that the starting population of cells in the library is not a completely homogeneous one. The fact that expression could not be regulated with the TetON system inevitably resulted in a heterogeneous starting population in which not all clones were found in the same proportion. This is unlikely to have altered the results presented here, since individual clones across replicates show a consistent behaviour and variance of expression levels is not large enough to attribute the trends to drift. However, it does mean that clones that had a strong negative effect on the cells could not be detected, since they must have decreased in frequency beyond detection in the first amplification round of the cells

Further consideration should also be given to the design of the random sequences inserted in the cells. In this study, I included a Kozak sequence and optimized codons in the flanking sequences according to human codon usage. It would be interesting to see whether a sequence design without a Kozak sequence has any effects on the patterns found here.

As far as I know this is the first experimental study of the effect of expression of random sequences on eukaryotic cells. It is of course only a preliminary analysis that cannot encompass the complexity of eukaryotic organisms. But it shows that organisms at different levels of complexity might have different strategies to deal with the expression of non-coding sequences. The types of sequences tolerated by bacteria (see Chapter 1) and by eukaryotic cells are different, with bacteria being seemingly more tolerant to shorter sequences. It is difficult to know whether this tolerance is maintained at the organism level in multicellular organisms, particularly those with adaptive immune systems (Bekpen et al., 2018). However, the variation provided by random sequences could be a rich source of innovation. The results shown in this chapter suggest that eukaryotic cells are very robust to the expression of random sequences, as expected from cells that have long non-coding regions frequently transcribed spuriously.

Chapter 3. Expression of mouse de novo genes in a human cell line

3.1 INTRODUCTION

After the discovery of *de novo* genes, their study has continued towards functional characterization of potential or confirmed candidates. A few examples of this are the study of *MDF1* in *S. cerevisiae* (Li et al., 2014), fish anti-freeze glycoproteins in codfish (Baalsrud et al., 2018); the discovery of the effect of the genes *saturn* and *goddard* in Drosophila (Gubala et al., 2017; Lange et al., 2021); and several examples of orphan in *Arabidopsis*, such as the ones described in (Arendsee et al., 2014).

More recently, Xie and others identified 110 orphan genes with translation evidence and possible *de novo* origins, and characterised the important role of one of them in regulating the oestrous cycle of the house mouse (Xie et al., 2019). A follow up study using knock-out mice for two other candidates in the list found that these novel genes are involved in transcriptional pathways related to development (Xie et al., 2020). Functions of orphan, taxonomically restricted, and *de novo* genes are varied and difficult to predict from sequence alone. This is true in particular for *de novo* genes, given their lack of sequence homology to any functional protein. Furthermore, these genes tend to have low expression levels, meaning that their effects on phenotype are likely to be small (Schmitz et al., 2020).

Despite knowing that at least some *de novo* genes are functional and may play important roles in the organisms where they are found, there is still much that we ignore about how these genes are integrated into their metabolic and regulatory networks. As mentioned before in this thesis, accurate identification of *de novo* genes is made difficult by the need to find syntenic non-coding regions in related species that might already have long times of divergence between them, and without population genomic data, it is not possible to tell whether nascent genes have already been fixed in a population. Lacking clear boundaries in the continuum between the birth of a gene and its integration into the genetic repertoire of a species, makes it difficult to study the process (Vakirlis et al., 2018).

When looking at the function of novel genes from an evolutionary perspective, one could expect protein or RNA products of novel genes to interact with already existing regulatory and metabolic networks in the organism, and to provide a fitness benefit that would help it become fixed in a population. In reality, it is not clear how these sequences interact with other extant genes or even if they initially provide any fitness advantage to the cells. Investigation into how young genes, protogenes or orphan genes integrate into cellular networks suggest that this process occurs fairly rapidly, within 14 MY (Abrusan, 2013). Recent work has also found that putative *de novo* genes in mouse are preferentially located near enhancer sequences, which may facilitate their integration into regulatory networks (Majic & Payne, 2020).

Bioinformatics approaches to understanding how *de novo* genes might interact with the regulatory and metabolic networks of an organism are limited by our knowledge of the complex interactions and regulatory feedback loops of an organism. In the case of model organisms with smaller genomes, such as *E. coli* or *S. cerevisiae*, it could be possible to generate predictions using machine learning or other computational approaches. However, for organisms with larger, more complex genomes, it might be impossible to predict what a new sequence does, or even if it has any function at all. Attempts to experimentally characterize *de novo* genes are still far between, and it is expected that much more should be done in this area in coming years (Bornberg-Bauer & Heames, 2019; Wu & Zhang, 2013). With this in mind, in this chapter I have started to investigate the effects of individual novel sequences being expressed in a eukaryotic cell line.

To achieve this, I selected candidates from a list of putative *de novo* genes from mouse (Xie et al., 2019), amplified them from biological material, and inserted them for expression into a human cell line—HEK293 cells using the Flp-In T-REx system (for a complete description of the expression system, see Chapter 2). Given that these genes are already expressed in mice and appear to be well tolerated, it is to be expected that the expression of these foreign sequences would be well tolerated by the human cell line. Using a transcriptomics approach, I investigated whether the young genes from mice have any effect on the transcriptome of the cells that might indicate interactions with their regulatory of metabolic networks.

Only 3 of the chosen genes could be analysed for this project, but I found that the expression of these mouse *de novo* genes was indeed well tolerated by the human cells, while the effects on the transcriptome were remarkably small. For the three sequences that were cloned, less than 5 genes could be identified as differentially expressed.

The results presented in this chapter provide some insight into what could happen to protogenes in an intermediate step of the process towards becoming a gene.

3.2 Results

3.2.1 Selection of candidates and generation of FITR293 cell lines

Initially, ten candidate genes out of a list of putative *de novo* genes identified in mouse (Xie et al., 2019) were selected to be amplified from mouse samples and transfected into the FITR293 cells. Selected candidates were chosen from those that have the shortest ORFs, between 144 and 501 nucleotides long, to match the short lengths that emerging *de novo* genes are expected to have. All selected candidates have proteomics evidence associated to them in the mouse and diverse levels of intrinsic disorder and hydrophobicity. They also have transcriptomic evidence in a wide range of tissues and developmental stages in the mouse. Each candidate was assigned a code Mdng (Mouse *de novo* gene) that will be used throughout the rest of this chapter (Table 3-1).

Table 3-1. List of selected candidate genes. Modified from (Xie et al., 2019).
Highlighted rows correspond to genes successfully transfected into FITR293 cells for this project.

Code	Gene_ID	Strand	Coding exon number	Exon number	ORF length	Protein length	ISD	Hydrophobic clusters	Amplified from
Mdng01	ENSMUSG00 000052075	-	3	3	441	147	0.61	0.3265	Testis cDNA
Mdng02[(1)]	ENSMUSG00 000053181	+	1	2	501	167	0.71	0.2994	Head cDNA
Mdng03	ENSMUSG00 000054057	-	1	3	471	157	0.05	0.758	Head cDNA
Mdng04	ENSMUSG00 000063254	+	1	1	366	122	0.01	0.8197	Embryo cDNA
Mdng05	ENSMUSG00 000074215	-	1	1	333	111	0.60	0.3514	Embryo cDNA
Mdng06	ENSMUSG00 000078444	-	1	3	144	48	0.16	0.6667	Testis cDNA
Mdng07[(2)]	ENSMUSG00 000078518	-	3	3	429	143	0.33	0.5035	Oviduct cDNA
Mdng08	ENSMUSG00 000078640	-	2	3	165	55	0.21	0.5818	Embryo cDNA
Mdng09	ENSMUSG00 000072983	-	2	2	399	133	0.42	0.4511	Testis cDNA
Mdng10	ENSMUSG00 000079261	+	4	4	465	155	0.30	0.4129	Testis cDNA
Mdng11	ENSMUSG00 000054450	+	1	3	273	91	0.01	0.8681	Embryo cDNA
Mdng12	ENSMUSG00 000056089	+	1	1	465	155	0.32	0.4258	DNA
Mdng13	ENSMUSG00 000073000	+	1	2	303	101	0.04	0.6832	Embryo cDNA
Mdng14	ENSMUSG00 000094690	-	2	2	303	101	0.58	0.4356	Testis cDNA

(1) Gene characterized in (Xie et al., 2020).
(2) Gene characterized in (Xie et al., 2019).

Using available DNA and RNA samples, specific primers were used to amplify all candidates from mouse tissues specified in Table 3-1, and from mouse genomic DNA directly in the case of Mdng12, which has only one coding exon, and for which no appropriate cDNA sample was available. A second PCR step was done in order to add restriction enzyme sites to the amplified products for cloning and amplification of the plasmids in *E. coli*. Purified plasmids were sequenced and seven of them, containing the gene sequence without mismatches to the reference mouse genome (Mdng1, 2, 6, 7, 10, 11, and 12) were used for transfection of FITR293 cells. Mismatches between the cloned sequences and the reference genome could be artefacts produced during the two separate PCR amplifications done to obtain enough product for cloning. Sequences obtained were compared with reported SNP data from mouse populations on the MGD website (Bult et al., 2019) [accessed on 2021.05.25], but no matches were found.

Due to low success rate of transfection, likely caused by difficulties during seeding of the cells in 96-well plates (see Methods), only three out of the 7 candidates yielded viable clones. The remaining results and discussion in this chapter will deal with these three clones: Mdng2, Mdng7, and Mdng10. Coincidentally, Mdng7 corresponds to the selected candidate *Gm13030* used for in-depth analysis in (Xie et al., 2019), and Mdng2 corresponds to gene *A830005F24Rik* for which knock-out mice were generated for RNAseq analyses in (Xie et al., 2020). It is worth noting that, even though I could only generate cell lines for the expression of three out of the fourteen candidates, the fact that it was possible to amplify all but one of the products from cDNA provides further support to the transcription of these genes in different tissues or developmental stages.

Taking a closer look at the three candidate mouse genes inserted in the FITR293 cells (Table 3-2), there are no BLAST hits for the nucleotide sequences of Mdng2 and 7 besides the gene itself in *Mus musculus*. For Mdng10, a hit could be found with low coverage but high identity in a syntenic region of the rat genome on chromosome 15 (location: chr15: 20,772,832-20,772,948, assembly Rnor_6.0). ENSEMBL (Howe et al., 2021) reports orthologues of all three genes in other mouse species, but no hits were found with the BLAST search. Phenotypes associated with Mdng2, reported in (Xie et al., 2020) include anomalies in heart, kidney and seminal vesicle morphology in knockout mice, while no associated phenotypes have been reported for the other two genes. Finally, all candidates were compared with the CAMP (antimicrobial peptide database) through CAMPsign (Waghu et al., 2016), which runs a BLAST on a database of known antimicrobial peptides. No hits were found for any of the candidates.

Table 3-2. Selected genes for which expression cell lines were successfully generated.

	Mdng2	**Mdng7**	**Mdng10**
Protein ID	ENSMUSP00000069912	ENSMUSP00000101431	ENSMUSP00000107457
Transcript ID	ENSMUST00000065465	ENSMUST00000105805	ENSMUST00000111826.3
Gene ID	ENSMUSG00000053181	ENSMUSG00000078518	ENSMUSG00000079261
Location (GRCm39)	chr13: 48,667,046-48,668,335	chr4: 138,598,303-138,601,275	chr14: 46,616,924-46,621,065
Gene Name	A830005F24Rik	Gm13030	Gm15217
BLASTn hits	Only self	Only self	Self-hit

			Rattus norvegicus (93.4% identity, 42% coverage)
ENSEMBL orthologues	*Mus musculus castaneus* *Mus spretus* *Mus caroli*	*Mus spretus* *Mus caroli*	*Mus spretus* *Mus caroli* *Mus pahariand* *Mus spicilegus*
Associated phenotypes	Enlarged heart, abnormal kidney morphology (enlarged kidney), abnormal seminal vesicle morphology	None	None
Domain prediction (INTERPRO)	3 disordered regions	None	None

The three cell lines generated with the mouse *de novo* genes were morphologically indistinguishable from the control ones transfected with GFP and the empty pcDNA5/FRT/TO vector. Their growth speed did not differ noticeably from the control or the non-transfected cells, with doubling times of around 24 hours for cultures seeded at 20% confluence. Presence of the inserted genes in the cells was confirmed by sequencing the specific PCR products from genomic DNA extractions. Correct function of induction of expression when adding doxycycline to the medium was confirmed visually for the cells transfected with GFP, and via relative quantification using RT-PCR for one of the candidate genes, Mdng10 (Figure 3-1). For this, RNA was extracted from cells exposed to 4 different concentrations of doxycycline, diluted to the same concentration in all samples, and retrotranscribed. Specific primers were used to amplify Mdng10 and 2 housekeeping genes, and the results visualised in an agarose gel. Although this approach is not strictly quantitative, it does provide an indication that there is a low level of basal expression even without the addition of doxycycline to the medium, and that the expression levels increase with the addition of doxycycline.

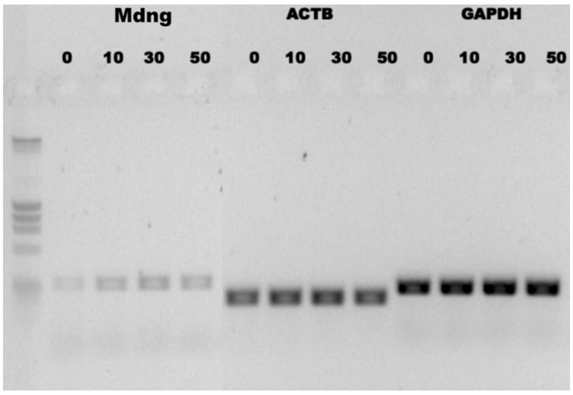


Figure 3-1. Relative quantification of candidate gene Mdng10.
Doxycycline was added to a final concentration between 0 and 50 ng/mL, RNA was extracted and retrotranscribed, and specific primers were used on the cDNA to amplify the candidate gene and two housekeeping genes (Actin B and GAPDH). Increased expression of the candidate gene can be seen with higher concentration of doxycycline.

3.2.2 Mouse de novo genes expressed in FITR293 cells

Cells from each of the 3 cell lines, and from the original FITR293 cells without transfection, were seeded in 24 replicate wells. Doxycycline was added to 12 of those wells to a final concentration of 50 ng/mL. Cells were sampled 36 hours after seeding, and total RNA was extracted and used for RNAseq in a single Illumina NextSeq run. After generating fastq files, merging the results of the sequencing lanes and trimming reads, between 84.62 and 95.75 % paired reads were accepted for analysis.

An average of 5 million reads were obtained for each replicate sample. However, batch effects were identified from the sequencing output. Replicates 5 to 8 of all treatments, corresponding to columns 5 to 8 on the sequencing plates, had 10 times less reads for all cell lines. In addition to this, replicates 1 and 2 of Mdng10 corresponding to the control group have between 10 and 100 times more reads than all other samples in the run (Figure 3-2).

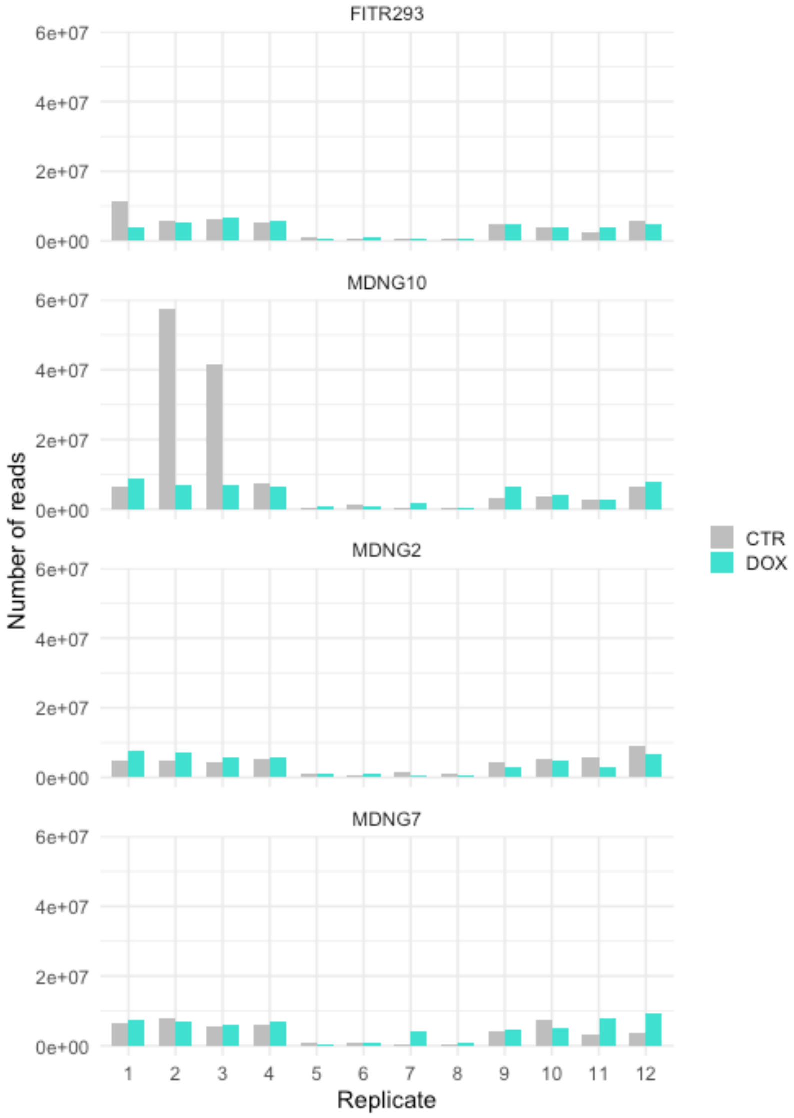


Figure 3-2. Total number of reads per replicate.

Sequencing, trimming, mapping and gene assignment results are summarized in Table 3-3. In order to get rid of the batch effects, replicates 5 to 8 for all samples were removed from downstream analyses, and replicates 1 and 2 of the Mdng10 treatment were subsampled to reduce the number of reads by 10-fold to match the average of all other files. After removing these outliers, the average number of reads was 5.7 million (± 1.8 million).

Table 3-3. Summary of Trimming, Mapping and Alignment results for reads after removal of outliers.

	Mean number of reads (SD)	**Mean Trimmed Read Pairs (%)**	**Mean Alignment (%)**	**Mean Concordant Reads (%)**	**Mean Assigned Reads (%)**
FITR293-CTR	5.8 x 10^6 (±2.6 x 10^6)	91.55	97.65	88.88	68.00
FITR293-DOX	4.8 x 10^6 (±1.2 x 10^6)	91.03	97.71	89.27	68.82
Mdng2-CTR	5.5 x 10^6 (±1.5 x 10^6)	92.31	97.88	88.37	66.80
Mdng2-DOX	5.4 x 10^6 (±1.8 x 10^6)	92.15	97.67	88.92	69.64
Mdng7-CTR	5.6 x 10^6 (±1.7 x 10^6)	92.40	97.64	88.86	68.98
Mdng7-DOX	6.7 x 10^6 (±1.5 x 10^6)	93.31	97.91	89.87	69.93
Mdng10-CTR	5.0 x 10^6 (±2.0 x 10^6)	92.49	97.84	89.16	69.10
Mdng10-DOX	6.3 x 10^6 (±1.9 x 10^6)	92.33	97.88	89.47	68.98
Total	*5002121*	*92.20*	*97.77*	*89.10*	*68.78*

When mapping the reads to the human genome (*Homo sapiens* assembly GRCh38) the recombined plasmids' expected sequence from the Flp-In™ T-REx™ system was added as an extra chromosome, and each mouse *de novo* gene was inserted at the plasmid's multiple-cloning site. Overall alignment rate was over 94 % for all samples and over 84 % of all aligned pairs were aligned concordantly exactly one time (one best hit with both reads in the expected direction).

To assign the mapped reads to the corresponding human genes and the genes on the plasmid, the GTF file downloaded from ENSEMBL with gene annotations for the human genome was joined with a GTF annotation file generated for the artificial plasmid construct. Although most reads were mapped with the approach described above, only between 59 and 76.2 % of alignments could be unambiguously assigned to annotated genes. This is likely to be due to multiple mapping of reads to paralogs, or the pervasive transcription of intergenic sequences.

Paired-end reads for each sample were also mapped to the plasmid construct separately in order to confirm that the induction of expression worked correctly. Results of this mapping were as expected, with clear differences in the expression of the MNDGs between the non-induced and induced treatments (Figure 3-3). Thus, confirming that the induction of the expression worked for all samples at least at the level of transcription. Leaky expression of the genes is expected and can, indeed, be observed from our results as well. Low basal levels of expression can be detected for all genes, in particular for Mdng10, while expression after induction with doxycycline increases 5.8, 4.2, and 2.4 times, respectively for Mdng2, 7 and 10.

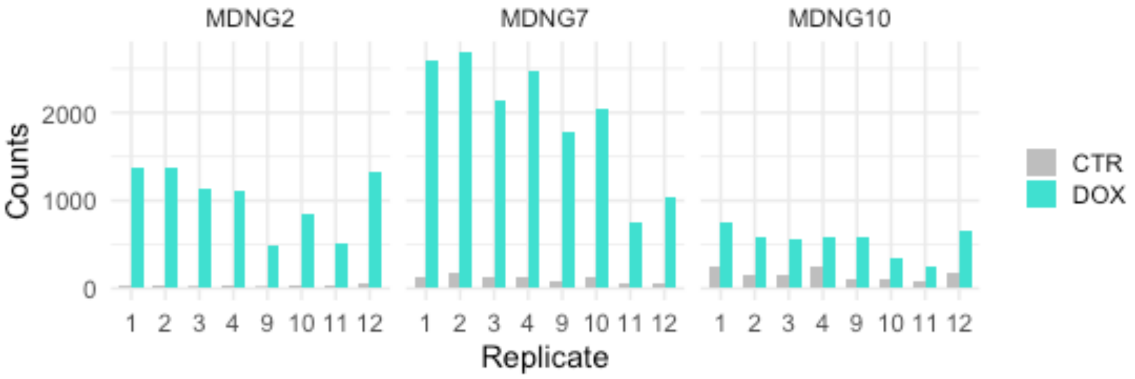


Figure 3-3. Number of raw counts assigned to each Mdng in the different replicates.
A clear difference in expression levels can be observed between the control samples (CTR) and the ones with addition of Doxycycline (DOX)

3.2.3 Few differentially expressed genes following expression of Mdng

Having confirmed that the induction of expression worked well in all cell lines generated, I studied the effect of the expression of the mouse *de novo* genes on the transcriptome of the cells. Using the count tables generated for each sample, I performed a differential gene expression analysis using DESeq2. Having identified batch effects likely caused during library preparation and sequencing, replicates were included in the experimental design formula to account for effects due to location of the samples in the plates.

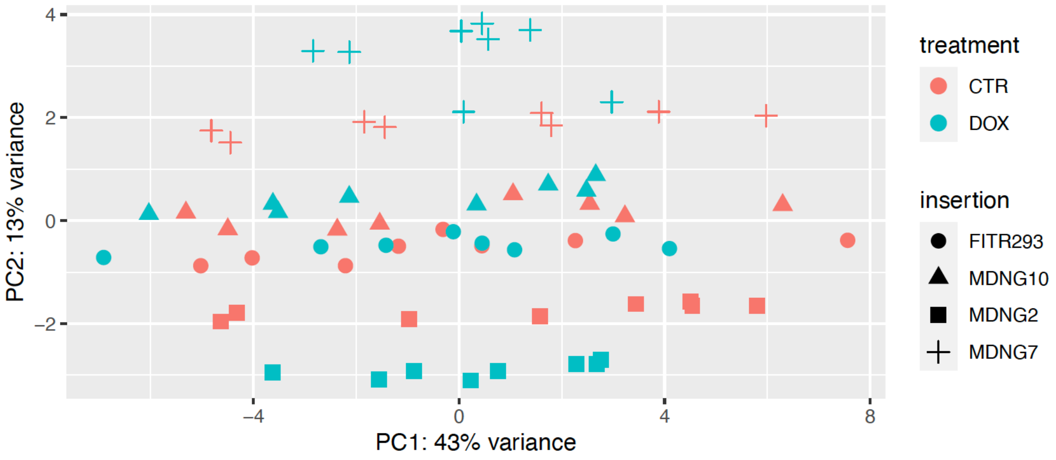


Figure 3-4. Plot of principal components 1 and 2, explaining 43 % and 13 % of variance between samples.
CTR: Control samples, DOX: Induced samples with doxycycline. Shapes correspond to each cell line. Samples and treatments separate along PC2, most clearly for Mdng2 and Mdng7. Mdng10 and FITR293 samples cluster together. There is a high level of variance between replicates along PC1.

Samples from different cell lines and treatments were distinguishable from each other in a PCA for Mdng2 and Mdng7, while Mdng10 tends to be more similar to the non-transfected cells (Figure 3-4). This analysis also revealed that there is considerable variance between replicates of the same cell line and treatment.

Results of the differential gene expression analyses are shown in Figure 3-5. There were very few differentially expressed genes following the expression of the "foreign" mouse *de novo* genes in the human cell line (Table 3-4).

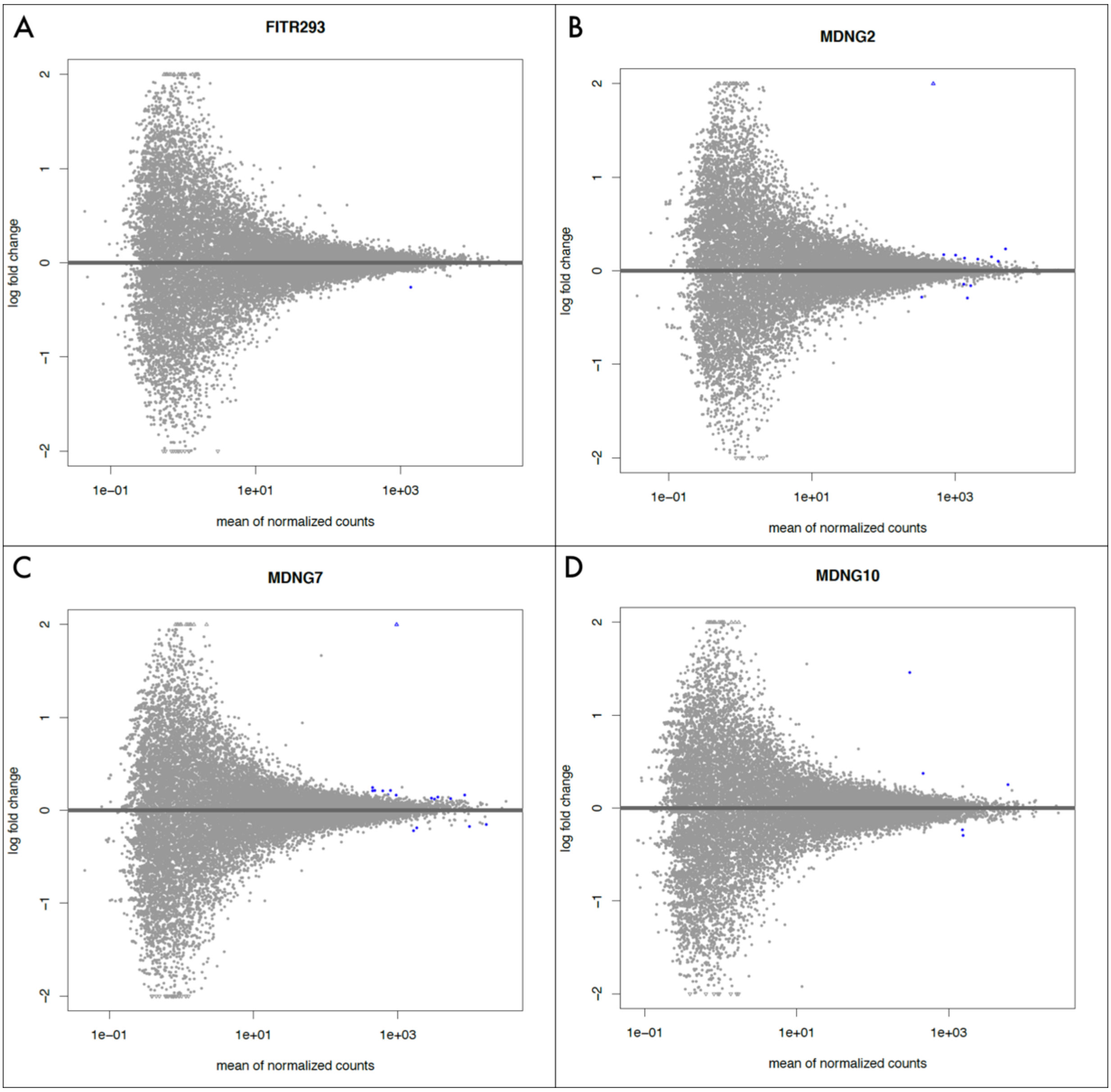


Figure 3-5. MA plots of the four cell lines comparing gene expression before and after addition of doxycycline to the medium.
Blue dots are differentially expressed genes (p_{adj}<0.05). A. Non-transfected Flp-In T-REx 293 cell line. B. Mdng2. C. Mdng7. D. Mdng10. For the Mdng, the gene with the highest fold-change corresponds to the Mdng itself.

First, looking at the non-transfected FITR293 cells to identify the effect of doxycycline on the cells, only one gene shows significant differential expression between the induced and non-induced treatments: UQCRFS1 (Ubiquinol-cytochrome c reductase, ENSG00000169021), part of the mitochondrial electron transport chain which drives oxidative phosphorylation. It shows a small 0.8-fold change in the cells to which doxycycline was added, and no other DEGs could be identified.

For all three Mdng the most significantly differentially expressed gene is the Mdng itself, followed in second or third place by the *UQCRFS1* gene found to be repressed by the addition of doxycycline. In all cases, the fold change of the *UQCRFS1* gene is comparable to that in non-transfected cells. Cells expressing Mdng2 and Mdng10 have a small increase in expression of a lncRNA associated with X chromosome inactivation—*XIST* (X Inactive Specific Transcript). There are other effects on mitochondrial genes associated with transcription and respiration. For Mdng10, there is also a DEG associated with iron metabolism.

Table 3-4. Complete list of differentially expressed genes for the four cell lines evaluated.
Normalised counts are base mean counts, Fold change is 2^log2foldChange, and padj is BH-corrected p-value, as obtained from DESeq2 results. Highlighted are the mouse de novo genes themselves.

Type	Gene ID	Gene Symbol	Gene/Protein name	Mean counts	Fold change	Adjusted p-value
FITR293	ENSG00000169021	*UQCRFS1*	Ubiquinol-cytochrome c reductase, Rieske iron-sulfur polypeptide 1	1388.27	0.84	7.23E-03
Mdng2	*ENSMUSG00000053181*	*Mdng2*	*RIKEN cDNA A830005F24 gene*	*498.16*	*41.86*	*0.00E+00*
	ENSG00000169021	*UQCRFS1*	Ubiquinol-cytochrome c reductase, Rieske iron-sulfur polypeptide 1	1495.52	0.82	8.12E-11
	ENSG00000229807	*XIST*	X Inactive Specific Transcript	5107.83	1.18	4.81E-07
	ENSG00000167978	*SRRM2*	Serine/arginine repetitive matrix protein 2	3248.42	1.11	1.42E-02
	ENSG00000198840	*MT-ND3*	Mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 3	1663.73	0.90	1.49E-02
Mdng7	*ENSMUSG00000078518*	*Mdng7*	*Gm13030*	*963.16*	*13.54*	*5.96E-41*
	ENSG00000210082	*MT-RNR2*	Mitochondrially Encoded 16S rRNA	9858.18	0.89	1.82E-06
	ENSG00000169021	*UQCRFS1*	Ubiquinol-cytochrome c reductase, Rieske iron-sulfur polypeptide 1	1652.64	0.86	3.01E-05
	ENSG00000231500	*RPS18*	Ribosomal protein S18	8464.77	1.12	4.73E-02
	ENSG00000198804	*MT-CO1 (COI, COX1, MTCO1)*	Mitochondrially encoded cytochrome c oxidase I	16872.82	0.90	4.73E-02
Mdng10	*ENSMUSG00000079261*	*Mdng10*	*Gm15217*	*307.51*	*2.75*	*7.62E-60*
	ENSG00000169021	*UQCRFS1*	Ubiquinol-cytochrome c reductase, Rieske iron-sulfur polypeptide 1	1549.48	0.82	3.03E-08
	ENSG00000229807	*XIST*	X Inactive Specific Transcript	6045.54	1.19	1.28E-03
	ENSG00000087086	*FTL*	Ferritin light chain	1519.74	0.85	1.28E-03

3.2.4 High sample dispersion impedes identification some DEGs

Since there have been reports that high dispersion in RNAseq samples may make the identification of differentially expressed genes difficult (Ching et al., 2014; Xie et al., 2020), I took a look at the dispersion fitted by DESeq2 for the samples (Figure 3-6). Samples for all genes have considerably large dispersions, as high as 0.5 for genes with small count numbers. According to the power analyses reported in (Xie et al., 2020), this means that these data do not provide sufficient statistical power to reliably identify differentially expressed genes with fold changes below 1.3 for genes with less than around 512 counts. Indeed, the only differentially expressed genes with a lower number of counts that could be identified for this study are the Mdng themselves, which have much larger fold-changes.

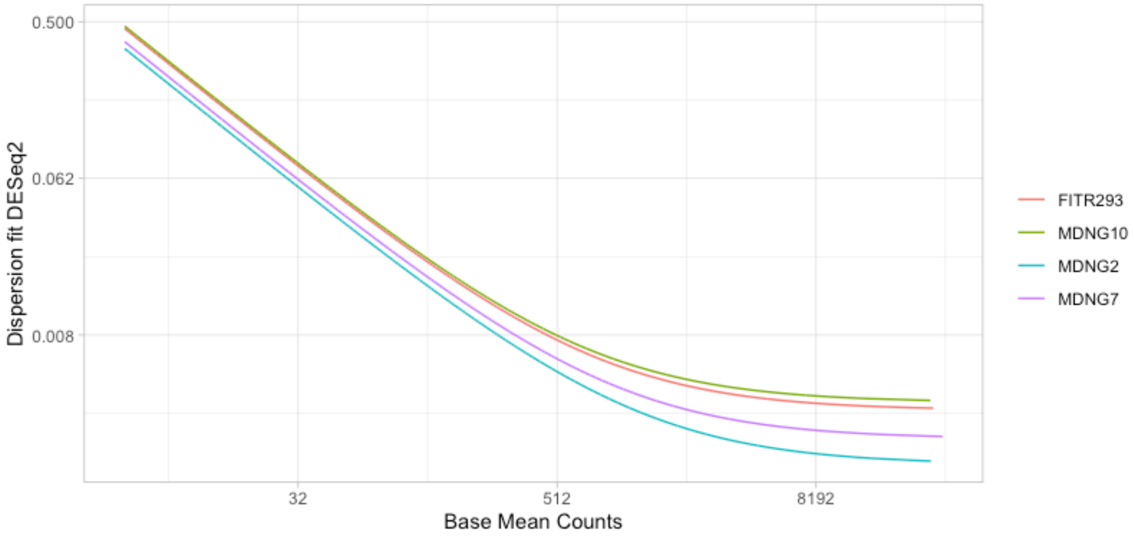


Figure 3-6. Fitted dispersion for count data of genes with more than 5 mean counts.
Dispersion values calculated by DESeq2 are related to mean and variance of counts for each gene.

3.3 Discussion

In this chapter, I generated 3 cell lines based on human HEK293 cells, expressing putative mouse *de novo* genes, and performed a transcriptomics analysis on them. The goal was to identify the effects that a completely new gene can have over the transcriptome of a different organism. This, in turn, could help us obtain a better understanding of the way in which novel genes that begin to be expressed in a cell could interact with its transcriptional and regulatory networks. By inserting novel genes from mouse in a human cell line, one could study the effect of sequences that are already known to be tolerated by a eukaryotic cell. Such a sequence could be used as a proxy for a *de novo* gene in the early stages of birth, but after having passed the filters imposed by selection on toxic ORF products. Given the stable expression of the tested genes in mouse species, my expectation was that their expression of the novel genes would not have a negative impact on the human cells.

Indeed, that is what I found from the results presented in this chapter. The transformed cells showed no morphological differences or slower growth and the expression of the mouse genes had an almost negligible effect on the transcriptome of the cells. In all cases there were 3 or less differentially expressed genes besides the mouse gene itself and UQCRFS1, which was identified as differentially expressed in all cells to which the inducer, doxycycline, was added. It is well known that doxycycline affects protein synthesis in prokaryotes, and in eukaryotes; including mitochondria protein synthesis in the latter (Nguyen et al., 2014; Sanchez et al., 2020). The effects were found even at the low dosage used for induction in this study. Two of the other differentially expressed genes—*MT-CO1* and *MT-ND3*, downregulated in the Mdng2 and Mdng7 cells, respectively—are mitochondrial genes, also part of the oxidative phosphorylation pathway. *RPS18* and *MT-RNR2*, are a riboprotein and a mitochondrially encoded rRNA, respectively.

There are 3 genes out of the 8 that were found to be differentially expressed in the cells expressing Mdng that do not seem to be related to any known effect of the antibiotics used. The first one was found to be slightly upregulated in cells expressing Mdng2 and Mdng10: *XIST* (X Inactive Specific Transcript), a lncRNA that has known functions in X chromosome inactivation, inflammation and stress responses (Wang et al., 2021). The second one, *SRRM2*, was upregulated in Mdng2 cells, and it is part of the spliceosome (Ilik et al., 2020). Finally, Mdng10 cells showed decreased expression of *FTL* (Ferritin light chain), a gene involved with iron metabolism (Muhoberac & Vidal, 2019).

All differentially expressed genes, besides the *Mdng* themselves, have very small fold changes. It is interesting that all the DEGs detected, except for *XIST* and *FTL*, could be related to effects of the antibiotics used for induction of expression of selection, as responses to stress. The information provided by the transcriptomics analyses alone is not enough to derive reliable conclusions about the effect of the mouse genes on the HEK293 cells. The dispersion of the data among different replicates was also higher than what is required to detect differentially expressed genes with small fold changes, so it is uncertain whether the effects of the genes are completely absent or just below the power of detection. Besides the batch effects found in this analysis, this dispersion could also be due to the cells being in different stages of the cell cycle. No attempt to synchronise the cell cycle was done in this experiment.

In either case, this work has shown that novel genes from a eukaryotic organism could be well tolerated by another one. If one tries to extrapolate the results to the process of *de novo* gene birth, these examples could provide information about the behaviour of protogenes in intermediate stages of their way to becoming a gene. In this scenario, such protogenes would not provide any initial fitness advantage, but they would be kept in a population without having a significant effect on the cells. This behaviour gives support to the hypothesis that they could be retained and even fixed in the population as “frozen accidents” (Schmitz et al., 2018).

Another interesting implication of these results is that these genes expressed in HEK293 cells are unlikely to interact with central nodes in the regulatory networks of the cells—usually composed by older genes (Abrusan, 2013). Given the long divergence times between mice and human, the lack of an effect of the *Mdng* expression on the transcriptome of the HEK293 cells could be explained if they only interact with younger genes.

The results presented in this chapter agree with evidence that novel (random) sequences expressed in a cell are well tolerated (Tretyachenko et al., 2017), and could provide some support to the hypothesis that *de novo* genes that stay in a population are, in principle, neutrally evolving, and could be the result of “frozen accidents” or neutral ORFs that escaped the initial purging by selection (Ruiz-Orera et al., 2018; Schmitz et al., 2018; Wu & Zhang, 2013).

The studies already published about Mdng2 (Xie et al., 2020) and Mdng7 (Xie et al., 2019) have shown that these genes do have an effect on the transcriptome of mice. Unfortunately, I had some difficulties during the transfection and selection of the cells, which resulted in a very low efficiency of generation of the cell lines. Given the small sample of genes studied here, it is difficult to generalize these results. Several further studies are necessary to improve and refine the results presented here. First, it would be useful to attempt the transfection of murine cells with the *Mdng* in over-expression assays, for example, in order to confirm that we are not only observing a lack of response in the system. Second, it would be interesting to see whether the effects are as small when using other candidates for transfection, and maybe also with control genes that are older and probably better interconnected in the protein interaction or regulatory networks of the human cells. Also, one could expect that by synchronising cell cycle stages, for example by treatment with a double thymidine block (Chen & Deng, 2018), one could reduce the dispersion of the data and thus make the assay more sensitive to specific changes.

Section III. Materials and Methods

Materials and methods used in specific chapters in Section II are indicated with the chapter number in square brackets. If no number is indicated, methods were used in all chapters.

1. Nucleic acid extraction and purification

For plasmid extraction and purification, bacteria were grown overnight in 5 mL LB medium (37 °C, 250 rpm). 3 mL of culture were centrifuged and plasmid DNA was purified using QIAprep spin miniprep kit (QIAGEN, Cat. No. 27106) following the provider's protocol, and re-suspended in the provided buffer.

Total genomic DNA was extracted from frozen cell pellets using DNeasy blood/tissue kit (QIAGEN, Cat. No. 69504)following the provider's protocol. For cells growing in 96-well plates, cells were lysed directly on the plate after removal of growth medium, transferred to a 96-deep-well plate, and DNA was extracted using an automated TECAN Freedom Evo 150 system with the Stratec Invisorb DNA Tissue HTS 96 kit (STRATEC, Cat. No. 7032900300).

RNA was extracted from frozen cell pellets using TRIzol Reagent (Invitrogen, Cat. No. 15596018), resuspended in nuclease-free water, and treated with DNase I (Promega, Cat. No. M0303S). For extraction from plates in Chapter 3, cells growing in 12-well plates were directly lysed in the well after removal of growth media using buffer RLT plus, and RNA was extracted using RNeasy 96 plus kit (QIAGEN, Cat. No. 74181).

Purification of PCR products was done either directly from the PCR reaction using QIAquick PCR Purification Kit (QIAGEN, Cat. No. 28104), or from excised agarose electrophoresis gel bands using QIAquick Gel Extraction Kit (Cat. No. 28706).

Nucleic acid content of single samples was quantified using a NanoDrop spectrophotometer, for single samples. For samples extracted in plates, DNA content was measured using Quant-iT PicoGreen dsDNA Reagent (Invitrogen, Cat. No. P11495) with a TECAN fluorescence microplate reader.

2. Sequencing

2.1 Sanger sequencing

Sanger sequencing was done for confirmation of PCR and plasmid miniprep products using specific primers (see list below). Sequencing reactions were prepared using purified templates and sequenced in an ABI Genetic Analyzer 3130XL or 3500XL.

2.2 Amplicon sequencing [CH1, CH2]

In Chapter 1, amplicon sequencing of the library samples at each timepoint was performed using specific barcoded primers to amplify a 356 bp fragment including the random sequences in a one-step PCR using PHUSION HF master mix (Invitrogen). The cycling program consisted of an initial denaturation at 98 ˚C for 30 seconds, followed by 25 cycles of 98 ˚C for 10 seconds, 65 ˚C for 20 seconds, and 72 ˚C for 1 minute. After a final elongation step of 72 ˚C for 10 minutes, samples were purified using a Qiagen MinElute Gel Extraction kit. Concentration of samples was calculated through relative quantification in an agarose gel, using a Molecular Imager Gel Doc XR+ System with the Image Lab Software (Bio-Rad).

In Chapter 2, a sample of the library was first sequenced using the Illumina MiSeq to obtain the library. After the time-course experiment, all samples from each timepoint were sequenced together using the Illumina NextSeq with shorter read lengths. In either case, amplicon sequencing primers were designed for two-step PCR-based sequencing (see list below). The first PCR for the amplicon sequencing was performed in triplicate, using Q5 Hot Start High-Fidelity 2X Master Mix (Cat. No. M0494S, NEB), with the following cycling program: initial denaturation at 98 ˚C for 30 seconds, followed by 24 cycles of 98 ˚C for 10 seconds, 62 ˚C for 30 seconds, and 72 ˚C for 20 seconds, with a final elongation cycle at 72 ˚C for 2 minutes. PCR products were pooled, visualized on a gel and purified using double sided selection with SPRIselect (Beckmann-Coulter, Cat. No. B23318). The second PCR was done with indexed Illumina primers, and the following cycling conditions: initial denaturation at 98 ˚C for 30 seconds, followed by 5 cycles of 98 ˚C for 10 seconds, 60 ˚C for 30 seconds, and 72 ˚C for 20 seconds, with a final elongation cycle at 72 ˚C for 2 minutes. The products were purified using AmpureXP beads (Beckmann-Coulter, Cat. No. A63881).

For both chapters, each barcoded sample was quantified, and then pooled together in equal concentrations to obtain the sequencing library. The final pool was measured with the Agilent Bioanalyzer using the Agilent DNA7500 kit and measuring with the fluorescence NanoDrop 3300 using the Qubit ds DNA BR Assay Kit. Finally, the pool was diluted to a final concentration of 10 pM for the MiSeq, or 1.8 pM for the NextSeq, and sequenced with 10 % or 1 % PhiX, respectively. Paired-end sequencing was done using Illumina's MiSeq Reagent Kit v3 to get 300 bp (CH1) or 250 bp reads, or NextSeq 500 to get 150 bp reads.

2.3 Whole genome sequencing [CH2]

There is no information provided by the manufacturer regarding where the FRT site is located in the genome of the cells. In order to make sure that the integration of the plasmids will not occur in a functional part of the genome, the genome of Flp-In T-REx 293 cells successfully transfected with a pcDNA5/FRT/TO/GFP-stop plasmid was sequenced. Cells were cultured under normal conditions, 1x10^6 cells were collected in microcentrifuge tubes and centrifuged. Supernatant was removed and DNA was extracted from the flash frozen pellet. Libraries were prepared using Illumina TruSeq Nano kit, and sequenced in an Illumina NextSeq with paired-end reads of 150 nucleotides.

After generation of fastq files using the bcl2fastq software from Illumina, reads were trimmed as described below and mapped to the human genome and the three plasmids: pcDNA6™/TR, pFRT/lacZeo, and pcDNA5/FRT/TO/GFP-stop using Bowtie2. Reads mapped to the plasmids were extracted and pairs that were mapped to both a plasmid and the genome were extracted to identify the location of the insertion of each plasmid in the genome.

2.4 RNASEQ [CH3]

Library preparation was done following Illumina's TruSeq Stranded mRNA Sample Preparation HS protocol.

2.5 READ CONVERSION AND QUALITY CONTROL

Reads were demultiplexed from base call (BCL) files using the software bcl2fastq (v2.20.0.422). The resulting fastq files were examined using FastQC (v.0.11.8). Reads from all sequencing experiments were trimmed using Trimmomatic (v. 0.36). For samples sequenced with NextSeq, reads from multiple lanes were merged with the script mergelanes.sh (https://github.com/stephenturner/mergelanes).

3. Library and plasmid generation [CH2, CH3]

3.1 CLONING INTO PCDNA5/FRT/TO PLASMIDS

pcDNA5/FRT/TO plasmids used for transfection of the Flp-In T-REx 293 cells were cut simultaneously with two restriction enzymes for directional cloning—specified below for each insert—and shrimp alkaline phosphatase (NEB, Cat. No. M0371S) was added to the restriction reaction in order to avoid re-ligation of empty plasmids.

Specific PCR products used as inserts were purified as described above and cut also simultaneously with the same restriction enzymes as the plasmid. Reaction times and temperatures were determined according to the pair of enzymes used. When heat inactivation was possible, the restriction mix was incubated at the temperature required by the enzyme with the highest inactivation temperature. Otherwise, the restricted plasmid and inserts were purified as indicated above.

Ligation was done in all cases using T4 DNA ligase (NEB, M0202S), for 1 hour at room temperature (approximately 21 °C). Ligation ratios of vector to insert were 3:1 unless stated otherwise, and the appropriate amounts were calculated using the NEBioCalculator (http://nebiocalculator.neb.com/) tool online.

Ligation products were used to transform chemocompetent *E. coli* (JM109, Promega, Cat. No. L1001) following the provider's protocol. 100 µL of the transformed cells were plated onto LB-agar plates with 100 mg/mL ampicillin and allowed to grow overnight at 37 °C. The remaining cells were used to inoculate liquid cultures in LB broth with 50µg/mL ampicillin for miniprep or midiprep plasmid extraction. For long term storage, 500 freezing in 25% glycerol. Single colonies were picked from the plates, re-suspended in 50 µL of nuclease-free water and lysed by heating to 98 °C for 5 minutes. Lysates were used as templates for colony PRC using DreamTaq polymerase Green PCR Master Mix (ThermoFisher, Cat. No. K1082). PCR products were visually inspected in agarose gels, and products with the expected size were purified and used for Sanger sequencing.

3.2 Eukaryotic library (EUKL) [CH2]

A random sequence library was designed to have sequences with a region of 150 random nucleotides flanked by predetermined features. On the 5' end of the random region all sequences have a HindIII restriction site for cloning, and a Kozak sequence including a start codon for transcription and translation initiation. The selected Kozak sequence is one of the most commonly used in human cells according to the literature (REF). On the 3' end of the random region, a histidine tag (6xHis-Tag) for protein detection, a stop codon and a NotI restriction site for cloning were included. In order to increase the probability of successful expression of the tags in a human cell line, care was taken to use histidine codons with a mid- to high- frequency of use in human genes. The full length of the ORF designed in the library is 174bp (58aa), including the 150bp (50aa) random region.

The oligonucleotide pool for the library was ordered from metabion GmbH as a single stranded oligo and amplified using specific primers. To avoid enrichment of partial sequences or PCR errors, only the reverse primer was added for an initial long amplification cycle with 30 seconds annealing at 55°C and 20 minutes extension at 72 °C, following the addition of the forward primer another long amplification cycle with annealing temperature of 60 °C was followed by six two-step amplification cycles: 98 °C for 10 seconds, and 72 °C for 1 minute. Products with the right size were purified from an agarose gel, cut using HindIII-HF and NotI-HF (NEB) and ligated into a pcDNA5/FRT/TO plasmid. The ligated plasmids were cloned into chemo-competent *E. coli* JM109 (Promega). Transformed bacteria were allowed to grow overnight in LB agar plates and 50 mL LB liquid medium with 50 µg/mL ampicillin. 5 mL of the culture were used the next day to freeze stocks in glycerol. The complete remaining volume was used in a midiprep assay to obtain purified plasmids for transfection of the eukaryotic cells.

3.3 Control GFP plasmid

pcDNA5/FRT/TO GFP was a gift from Harm Kampinga (Addgene plasmid # 19444 ; http://n2t.net/addgene:19444 ; RRID:Addgene_19444). The GFP in this plasmid does not have a stop codon so I designed specific primers to amplify the GFP sequence (see list below), and add a stop codon while keeping the HindIII and BamHI restriction sites for re-cloning into the empty pcDNA5/FRT/TO plasmid.

3.4 Mouse *de novo* gene (Mdng) plasmids [CH3]

DNA and RNA samples for the amplification of the candidate genes were kindly provided by Dr. Chen Xie (MPI for Evolutionary Biology). cDNA was obtained from the available mouse RNA samples using RevertAid First Strand cDNA Synthesis Kit (ThermoFisher, Cat. No. K1622). PCR conditions were optimised for each primer pair until specific bands could be obtained. PCR products were sequenced with Sanger. Purified and confirmed PCR products for the candidate genes were cut using HindIII and BamHI (NEB, Cat. No. R3104M and R3136M), and ligated.

4. Cell culture [CH2, CH3]

Cells were counted using a Countess II-FL cell counter, and collected by centrifugation at 250 x g for 5 minutes, unless otherwise stated.

4.1 Cells

Flp-In T-REx 293 cells were grown in complete medium (CM)—DMEM High-glucose, pyruvate (Gibco, Cat. No. 41966052) supplemented with 10% tetracycline-free FBS (PAN Biotech, Cat. No. P30-3602, Lot No. P080317TC)—and antibiotics at 37 ˚C with 5 % CO2. For non-transfected cells, Zeocin and Blasticidin were added to final concentrations of 100 µg/mL and 15 µg/mL, respectively; for transfected cells, Hygromycin to a final concentration of 100µg/mL was added instead of Zeocin. Cells were allowed to grow to up to 80 % confluence before passaging according to the manufacturer's instructions. Routinely, cells were passaged in a 1:4 to 1:10 ratio. For passaging a T75 flask (75 cm^2 growth area), cells were washed once using 10mL DPBS (PAN biotech, Cat. No. P04-36500). 3 mL of 0.05 % trypsin (Gibco, Cat. No. 25300054) were used to detach the cells with incubation at 37 ˚C for 5 minutes, and 5 mL of complete medium were added to stop the reaction. 2 mL of the resulting cell suspension were transferred directly into a new flask with 12 mL of fresh growth medium to obtain a 1:4 passage ratio. The quantities of reagents were scaled up or down as necessary for containers of different size.

4.2 Growth curves [CH2]

Growth curves were obtained by counting three replicate wells of cells seeded in a 12-well plate (growth surface of 3.5cm2). For the Flp-In T-REx 293 cells, optimal seed number was first determined by testing four different seed numbers (5E3, 1.5E4, 6E4, and 1.8E4 cells) in wells with three replicates each. Subsequent growth curves were done seeding 1.5E4 cells in each well. Cells from three replicate wells were trypsinized daily for one week, and counted with trypan blue for quantification of cell viability. Medium was refreshed as needed by replacing two thirds of the volume for fresh medium whenever there was a change in pH visible by the change in coloration of the medium from pink to orange or yellow. Growth curves were made by calculating the number of live cells per square centimetre of growth surface.

4.3 Transfection

Using a control plasmid expressing GFP (pmaxGFP® Vector, Lonza), I confirmed in a preliminary experiment that transfection of these cells can be achieved with high efficiencies of up to 80% using either nucleofection, Fugene6 reagent (Promega, Cat. No. E2693) or Lipofectamine3000 reagent (ThermoFisher, Cat. No. L3000008). At the same time, I calculated low efficiency rates of integration of the plasmid into the genome (0.0036 %–0.0063 %) with either method. Cells were cultured under normal conditions until they reached 70 % confluence before transfection. On the day of transfection, medium was replaced with CM without antibiotics or Opti-MEM (Gibco, 31985062) at least one hour before the experiment.

For the eukaryotic library (CH2), three T75 flasks (estimated 1x10^7 cells/flask) were co-transfected using 30 µL Lipofectamine3000 reagent with either 20, 50 or 80 µg of plasmid, distributed between the library plasmid (pcDNA5/FRT/TO/Lib) and the Flp integrase plasmid (pOG44) in a 1:9 ratio.

For the Mdng candidates (CH3), cells growing in a 96-well plate were transfected with plasmids that had the expected insert. Transfection was done in 4 replicate wells for each insert, adding 0.2 ug of total plasmid per well, in the same ratios as indicated above, and using Fugene6 reagent in a 3:1 ratio to the total amount of DNA.

4.4 Selection of transfectants

Cells that do not integrate a pcDNA5/FRT/TO plasmid at the FRT site do not acquire resistance to Hygromycin and are killed in the course of one to two weeks after passaging. For all inserts, cells were cultured under normal conditions with the transfection reagent complex for 48 hours and then passaged in a 1:4 ratio, with selection antibiotics (Blasticidin and Hygromycin) added to the medium. Selection medium was refreshed every three to four days by replacing two thirds of the medium in the container with fresh, pre-warmed selection medium. When individual colonies of attached and dividing cells with normal morphology could be detected under the microscope, cells were trypsinized and collected in a new T75 flask for amplification and freezing.

4.5 Induction of expression

The minimum amount of doxycycline needed to induce expression was determined by generating an induction curve. Eight different concentrations of doxycycline (0, 1, 10, 20, 30, 50, 100, 300 ng/mL) were added to cells transfected with the pcDNA5/FRT/TO plasmid containing GFP. Doxycycline was added at the moment of seeding, and cells were monitored for 4 days, to control for changes of morphology, and quantify expression levels through RNA extraction and RT-PCR. Quantification of the number of cells expressing GFP was done using a Countess II FL cell counter with a green filter. Viability of the cells was also quantified using trypan blue.

4.6 Single clone isolation [CH2]

Cells grown in normal conditions were trypsinized and re-suspended in CM with antibiotics. The suspension was strained into a 5 mL Falcon™ round-bottom polystyrene test tube with cell strainer cap. A sterile Corning™ Stripwell™ Microplate was prepared with 150 µL of room-temperature medium in each well. The cell suspension was sorted into the strips using a Bio-Rad S3e™ Cell Sorter with the purity setting to ensure that only one cell was sorted per well. The sorted cells were kept in an incubator under normal conditions for 2 weeks or until colonies could be detected under the microscope.

Wells where cells could be detected were trypsinized and passaged into a new 96-well plate with CM and antibiotics. When confluence reached 70-90% in all wells, medium was removed and cells were lysed in the plate for DNA extraction. Extracted DNA was used as template for PCR with primers on the flanking regions of the random sequences, and the purified product used for Sanger sequencing. The presence of each clone in the database was confirmed using BLAST.

5. Time-course experiment with E. coli library [CH1]

I had access to a sample of the original library produced by Neme et al, 2017. Transformed *E. coli* DH10B cells were available as frozen 500 µL stocks in 20% glycerol. In order to assess the strength of the effect observed in the 2017 study, I repeated the experiment using a 100-fold dilution of the original library and a one-day sampling schedule, with samplings every three hours for a total of four samplings in 12 hours. I seeded 5 µL from the stock on 25 mL LB liquid medium with 500 µg/mL ampicillin, and allowed it to grow overnight at 37 °C with constant shaking (250 rpm). After 16 hours, 500 μL of the liquid culture were transferred into five 5 mL tubes containing 4.5 mL of LB medium with 1 mM IPTG to induce expression of the random sequences. For each cycle, 500 μL of culture from each tube were used to seed a new replicate after three hours of growth (37, 250 rpm). From the remaining bacterial culture for each replicate, 3 mL were collected and used for plasmid extraction. Extracted plasmids were eluted in 30 μL of elution buffer and stored at -20 ˚C for amplicon sequencing.

5.1 *E. COLI* AVAILABLE DATA [CH1]

In addition to sequencing data from the diluted library experiment described above, I had access to the original fastq files of the eight experiments described in Neme et al. 2017. The original experiments were done following two different sampling schedules: either a one-day course with samplings every three hours, or a four-day course with samplings every 24 hours. In either case, four timepoints were sampled. The number of replicates, cycle duration and experiment length for each of the experiments are described in the Table 1, as well as the equivalency for the experiments mentioned in Neme et al., 2017. In addition to three experiments with 10 replicates of each type of sampling schedule, there are two 4-day experiments with 5 replicates. One of them was done with a treatment control without induction with IPTG, while the other one was sequenced 5 times to capture even rare clones present at low frequencies in the population. Finally, I included the sequencing data from the diluted-library experiment described above.

Table- A. Available data. Amplicon sequencing of random sequence library in E. coli.

Experiment Number	Cycle length/ Experiment length	No. of replicates	Description (Experiment Name in Neme et al. 2017)
1	3 hours/ 1 day	10	(E1)
2	3 hours/ 1 day	10	(E2)
3	3 hours/ 1 day	10	(E3)
4	24 hours/4 days	10	(E4)
5	24 hours/4 days	10	(E5)
6	24 hours/4 days	10	(E6)
7	24 hours/4 days	5	Samples with and without IPTG (Induction control)
8	24 hours/4 days	5	Re-sequenced 5 times (Deep sequencing)
9	3 hours/ 1 day	5	Diluted library

6. Eukaryotic library time-course experiment [CH2]

The experiment was done starting from a frozen aliquot of the library with 3.5 million cells. The aliquot was thawed and seeded in a T75 flask with 14 mL of CM. After 80 % confluence was reached, cells were passaged into two T150 flasks with 20 mL of CM, plus antibiotics (Blasticidin and Hygromycin) and allowed to grow again to at least 70 % confluence. Cells were trypsinized and collected in a single tube, counted, and used to seed ten T75 flasks with 3x10^6 cells each. Doxycycline was added to a final concentration of 10 ng/mL to five of these flasks at the seeding stage. At each subsequent passage every two days, one fourth of the population of each flask was seeded into a new one, and the remaining cells were divided in four aliquots to be used in DNA, RNA and protein extractions, and for freezing. Aliquots were centrifuged, supernatant was removed, and pellets were either flash frozen in liquid nitrogen or re-suspended in freezing media and frozen as routinely. Sampling was done for 10 timepoints. Samples were divided in 2 96-well plates and DNA extraction was done as stated above.

7. Random sequence Analysis Pipeline [CH1, CH2]

7.1 Database generation [CH1, CH2]

For Chapter 1, and the preliminary MiSeq sequencing experiment in Chapter 2, trimmed reads for each experiment were merged using the software USEARCH10 (-fastq_mergepairs, -fastq_maxdiffs 30, -fastq_minmergelen 100). Since each read in a pair covers the entire random sequence, up to 30 mismatches are allowed between the paired forward and reverse reads. The fastq_mergepairs algorithm resolves discrepancies between the forward and reverse reads by comparing the quality score for the conflicting position in each read. It keeps the residue with the best quality score in the merged read. Merging the reads with this algorithm reduces the percentage of sequencing errors kept in each read. Unfortunately, it is not possible to account for PCR errors that have occurred during the library preparation.

To remove reads that do not belong to a PCR product from the plasmids in the library, a custom Perl script was used to find and save all merged reads containing pre-defined sequences up- and downstream the random sequences on the pFLAG-CTC plasmid. The pre-defined sequences were a 18bp sequence around the start codon, and the FLAG tag, including the stop codon. The reads thus selected are considered clean amplicon reads, trimmed around the pre-defined sequences, and used for all subsequent analyses.

For each sequence in each of the two libraries, ORFs were predicted using getorf from the EMBOSS suite. The ORFs used were those between a start and a stop codon (-find 3) with the default minimum size changed to 12 nucleotides—the length of the constant start sequence of the bacteria library—or 3 nucleotides—a single amino acid in the case of the eukaryotic library. The first ORF found for each sequence was taken as the transcription product of the sequence. If no starting codons could be found in the sequence, the ORF was assigned as a missing value to the database. ORFs were translated in the first frame using the program transeq from EMBOSS, and assigned as the predicted peptides for each sequence. GC contents for both the full read and the ORF were calculated using custom Perl scripts, and intrinsic disorder scores were calculated for the predicted peptides using IUPRED with both the short and long options.

To generate a database of all unique sequences in the library that could be detected by the amplicon sequencing approach, all clean reads from available experiments and replicates were merged and dereplicated using USEARCH10. Dereplication was done in three rounds. In the first two rounds, sequences were sorted alphabetically, and the -fastx_uniques option was used to keep only one sequence of each type in the database, and removing singleton reads from the file, while keeping track of the number of total sequences of each type with the -sizeout option. The first round was done on each individual sequencing file, and the second one was done on a combined file merging all de-replicated files from the first round. In this way repeated sequences are deleted and an annotation is added to the read name indicating how many exact matches were present in the clean read files. This exact matching approach is prone to enrichment of PCR or sequencing errors, since any two reads with even a single nucleotide difference are kept as individual sequences in the database. Singleton reads–more likely to be PCR or sequencing errors–were removed and a third dereplication round using a clustering approach was implemented.

The third round of dereplication aimed to remove reads generated by PCR or sequencing artefacts. The clustering approach used is based on the one used for OTU validation in microbiome analyses. Reads were sorted in decreasing order of size annotation, and the -cluster_smallmem option of USEARCH10 was used with an -id of 0.97. The UCLUST algorithm used by USEARCH is a greedy clustering approach. Here, sorting by size means that high-frequency reads are used as centroids for clusters first. This strategy relies on the assumption that reads found in high frequencies are more likely to be real, and less-common, highly-similar reads are probably generated through PCR errors. The identity threshold of 0.97 allows less frequent reads with up to 5 mismatches in the 195-nucleotide sequence to join the high-frequency centroids forming the clusters. Using an additional filter of minimum cluster size of 8 reads, commonly used in microbiome amplicon sequencing analyses, removes chimeras and other artefacts from the database. The resulting libraries of unique clusters (full database, BACT_DB.fa, and EUK_DB.fa) are considered to contain the sequences of all clones present in each library.

7.2 Database quality control [CH1, CH2]

In order to make sure that the database generated contains all clones actually present in the library, and that they correspond to the expected content of a random database of sequences, I evaluated what percentage of the original reads were lost in the database generation process. To do this, I followed the percentage of reads kept and correctly merged after trimming, percentage of reads containing both forward and reverse primers, and the percentage of reads that were included in the de-replication and clustering steps. Other parameters computed were the percentage of clones reported in the 2017 analysis that could be mapped back to the new database for the bacteria library, and, most importantly, agreement of the distribution of the length of the predicted ORFs to the expected distribution of unique ORFs obtained from a set of random sequences, as described below.

7.3 Sequence features [CH1, CH2]

Several parameters were used to characterise the sequences in the complete database, as well as in the sequence groups generated after mapping of the reads to find changes in frequency:

Sequence length was calculated for each read, as well as the predicted ORF and peptide encoded by them using bash programs during the database generation. The number of peptides of each length will depend on the probability of getting a stop codon at each consecutive position, and not before. This is best described by the probability function of a geometric distribution (1-p)^(k-1)*p, where k is the number of trials, in this case, the number of positions or the length of the sequence; and p is the probability of "success" or at getting a stop codon. Multiplying this probability distribution by the number of synthesised sequences, we get the expected count of peptides of each length. In addition to this, it is possible to predict the possible number of unique sequences of each length using the exponential function 20^k, to describe the number of possible unique combinations of the 20 amino acids in a sequence of length k.

GC content was calculated as percentage of either the full read or of the predicted ORF using custom Perl scripts as the percentage of guanine (G) and cytosine (C) in a sequence relative to its length. Amino acid composition of the database and different sequence groups were calculated using the Biostrings package from Bioconductor in R. Lists of sequences from each database formatted as AAStringSets were used as input for the letterfrequency function and amino acid frequencies were plotted for each sequence correcting for length.

Intrinsic disorder was calculated using the command line version of IUPred (IUPred2A) with the "short" option. Intrinsic disorder values were averaged for each protein to obtain a single average disorder value per protein. Since very short proteins are very likely disordered–they cannot produce any secondary structures–a simple ratio of order-inducing versus disorder-inducing amino acids was also calculated as a descriptor of sequences that might be more prone to aggregating. Higher ratios of order-inducing, polar or very charged, amino acids might increase the chances of aggregation even for short peptides.

Protein aggregation propensity was calculated for each sequence using the program PASTA 2.0 on the web server of The BioComputing UP lab of the University of Padua (Italy). For each sequence, free energy for the best pairing was obtained using the default settings for peptides. The best energy pairing for self-aggregation was obtained for each sequence, and energies of -5 or less were considered indicative of a high probability of aggregation.

7.4 Mapping of reads to full database [CH1, CH2]

For chapter 1, clean reads for all replicates and timepoints in each experiment were mapped to the BACT database using a global alignment-based method from the program USEARCH10 (option -usearch_global). Since the NextSeq reads were shorter and the overlap was too short for merging, for chapter 2, trimmed forward reads were mapped onto their corresponding database. For consistency with the clustering analysis, alignments had a minimum required identity of 0.97, minimum query coverage of 0.9, and maximum one hit and 5 gaps. Hits were extracted from the search results and counted using custom bash scripts to generate count tables for each replicate in each experiment.

7.5 Frequency change determination [CH1, CH2]

Raw count tables for each experiment were used as input for statistical analyses using the package DESeq2 in R. The analysis was performed using the standard DESeq2 pipeline normally used for differential expression analysis of RNAseq samples. Sequences with less than 5 reads mapped to them were excluded from the analysis. Count data of each experiment were analysed independently using cycle number as explanatory variable.

Since replicates are not independent between timepoints, a design formula including an effect from cycle, plus an effect by replicate was used in the fitting the data. Since no differences were found between the design formula with timepoint and replicate, and the design formula with only Cycle, the latter was used for analysis. Contrasts were set between each timepoint and timepoint number 1. Results were visualised as MA plots generated by plotting the logarithm of the normalised base counts against the logarithm base two of the fold-change predicted by DESeq2. For plotting, significant data points (padj<0.05) were colour-coded according to the direction of their change in frequency.

7.6 Group assignment [CH1, CH2]

Using the result tables generated by DESeq2 for the contrasts set between the last cycle of each experiment with the first, sequences were divided in three categories for downstream analyses. Categories were assigned according to the fold-change in frequency: non- significant, significant and with a positive fold change (increase in frequency), and significant with a negative fold change (decrease in frequency). For simplicity, the three categories are called non-significant, positive and negative throughout this document. Sequences were considered significant if they had an adjusted p-value (padj) lower than 0.05 as calculated by the program. Fold-change was considered positive when larger than one and negative when lower than -1. For category assignment, a flag was added to each sequence depending on whether its fold-change was positive (> 1) or negative (< -1) and significant (padj < 0.05), or non-significant (padj > 0.05) for each experiment.

Control and induced samples were analysed both independently and together including an interaction term for treatment and time point. The results table containing average counts for each sequence was used to determine the magnitude (fold-change) and significance of their frequency change (p-value adjusted for multiple testing, padj > 0.05).

According to the DESeq2 results, individual sequences in the library were assigned to groups depending on whether they increased, decreased or did not change significantly in frequency between the first and last timepoint of the experiment.

7.7 Cross comparison between experiments [CH1, CH2]

Category flags were compared for all sequences in the database across experiments, and kept when the majority of experiments had the same flag (n=(number of experiments)/2 +1). Category assignment was done for comparisons of all experiments available, and independently for the two types of sampling schemes, i.e., every 3 hours or every 24 hours, to differentiate between sequence effects and effects due to the experimental design.

7.8 Database simulations

Simulations of random databases were generated in R by sampling with replacement the four nucleotides with equal probabilities. Sequence features were determined for the simulated databases in the same way as for the experimental database.

8. RNAseq experiment [CH3]

8.1 Mouse *de novo* genes (Mdng) Experimental design [CH3]

Candidate genes were selected from the list of putative *de novo* genes from (Xie et al., 2019), with length as the main selection criterion. From the shortest sequences, care was taken to select candidates with higher levels of expression, and a diversity of intrinsic disorder and aggregation propensity scores. Selected genes were amplified from available genetic material and plasmids obtained as described above were used for transfection of Flp-In T-REx 293 cells.

Using the cell lines produced, a transcriptomics experiment was set up with two treatments: Induced samples using 50 ng/mL of doxycycline to induce expression of each candidate gene, and a control group with no additives to the growth medium. In addition to the candidate genes, the original, non-transfected FITR293 cell line was also included in both treatments in order to assess the effect of doxycycline alone on the cells. Twelve replicates were seeded for each treatment and each cell line in 12-well plates. Sampling was done 36 hours after induction of expression with doxycycline. RNA extraction, library preparation and sequencing were done in 96 well plates with one treatment per row and 12 replicates (Table A).

Table- B. Experimental setup for RNA sequencing of Mdng-expressing and control cells.

Row	Sample (Columns 1–12)
A	FITR 293 non-transfected cells CTR (no Doxycycline)
B	FITR 293 non-transfected cells DOX (with Doxycycline)
C	Transfected cells - Mdng2 CTR (no Doxycycline)
D	Transfected cells - Mdng2 DOX (with Doxycycline)
E	Transfected cells - Mdng7 CTR (no Doxycycline)
F	Transfected cells - Mdng7 DOX (with Doxycycline)
G	Transfected cells - Mdng10 CTR (no Doxycycline)
H	Transfected cells - Mdng10 DOX (with Doxycycline)

8.2 Candidate gene characterization

Candidate gene CDS sequences were downloaded from ENSEMBL Mouse assembly (GRCm38). Specific primers were designed for a two-step amplification (Table- H). The first PCR was done to amplify the product from the corresponding cDNA or DNA sample as shown in Table 3-1. The second PCR was done to add restriction sites for HindIII and BamHI in the 5’ and the 3’ end of the sequences, respectively (Table- I). Amplified products were cloned as described.

Gene information was retrieved from ENSEMBL annotations for each candidate. Additionally, the CDS sequences were used as query for BLAST against the complete non-redundant nucleotide database of the NCBI, with default parameters. Domain prediction was done on the web interface of INTERPRO (Blum et al., 2021), and signatures of anti-microbial peptide homology were investigated through the web interface of CAMPsite (Waghu et al., 2016).

8.3 RNAseq analyses

Reads were trimmed to a minimum average quality score of 20 and 75 bases of length. In order to get rid of the batch effects, replicates 5 to 8 for all samples were removed from downstream analyses, and replicates 1 and 2 of the Mdng10 treatment were subsampled using seqtk sample (https://github.com/lh3/seqtk.git) to reduce the number of reads by 10-fold to match the average of all other files.

Mapping was done using HISAT2 (v.2.2.1) (Kim et al., 2019) . The human genome (*Homo sapiens* assembly GRCh38) FASTA and GTF files were downloaded from ENSEMBL, and the recombined plasmids' expected sequence for each of the genes from the Flp-In™ T-REx™ system was added as an extra chromosome. Mapped reads were saved in sorted BAM format using samtools (v.1.9) (Berretta & Morillon, 2009), and used for gene counting using the Subread package featureCounts (v.2.0.1) (Liao et al., 2014).

Raw count tables generated independently for each gene were used for differential gene expression analysis using DESeq2 (v.1.30.0) (Love et al., 2014). A likelihood ratio test was used to compare results using the full experimental design formula ~Replicate + Treatment and the reduced formula ~Treatment. Since no differences were identified, the reduced formula was used for the analyses presented here. Dispersions calculated by DESeq2 were extracted to analyse the degree of variance among replicates in the experiment.

In addition to mapping with HISAT2, a quick mapping of the trimmed reads was done against the plasmid sequence for each of the three genes in Geneious. The results were used to evaluate the differences of expression of the Mdng between the induced and non-induced samples.

9. Primers used in this study

Primers were designed by me, unless otherwise stated, using the primer design tools Geneious, Primer3Plus (https://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi) and OligoAnalyzer Tool from IDT (https://eu.idtdna.com/pages/tools/oligoanalyzer). Primers were ordered from Sigma-Aldrich/Merck either as lyophilised powder or resuspended in nuclease-free water. All stock solutions were resuspended to a concentration of 100 µM for storage, and diluted to a concentration of 5 µM for working stocks. Primers were stored frozen at -20 °C between uses.

Table- C. Primers used for one-step PCR amplicon sequencing of E. coli library [CH1]
These primers were designed by the authors of (Neme et al., 2017).

Primer name	Sequence
pFLAG-CTC FWD-1	AATGATACGGCGACCACCGAGATCTACAC AACCGCAT ACACTCTTTCCCTACACGACGCTCTTCCGATCT CATCATAACGGTTCTGGCAAATATTC
pFLAG-CTC FWD-2	AATGATACGGCGACCACCGAGATCTACAC AAGGCCTT ACACTCTTTCCCTACACGACGCTCTTCCGATCT T CATCATAACGGTTCTGGCAAATATTC
pFLAG-CTC FWD-3	AATGATACGGCGACCACCGAGATCTACAC AGAGTGTG ACACTCTTTCCCTACACGACGCTCTTCCGATCT GT CATCATAACGGTTCTGGCAAATATTC

pFLAG-CTC FWD-4	AATGATACGGCGACCACCGAGATCTACAC CACAAGTC ACACTCTTTCCCTACACGACGCTCTTCCGATCT CGA CATCATAACGGTTCTGGCAAATATTC
pFLAG-CTC FWD-5	AATGATACGGCGACCACCGAGATCTACAC CGTTCCTA ACACTCTTTCCCTACACGACGCTCTTCCGATCT ATGA CATCATAACGGTTCTGGCAAATATTC
pFLAG-CTC FWD-6	AATGATACGGCGACCACCGAGATCTACAC GCTTGGAT ACACTCTTTCCCTACACGACGCTCTTCCGATCT TGCGA CATCATAACGGTTCTGGCAAATATTC
pFLAG-CTC FWD-7	AATGATACGGCGACCACCGAGATCTACAC GTCAACAC ACACTCTTTCCCTACACGACGCTCTTCCGATCT GAGTGG CATCATAACGGTTCTGGCAAATATTC
pFLAG-CTC RWD-A	CAAGCAGAAGACGGCATACGAGAT AACCGGAA GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT A CTGTATCAGGCTGAAAATCTTCT
pFLAG-CTC RWD-B	CAAGCAGAAGACGGCATACGAGAT AGAGTGAC GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT TC CTGTATCAGGCTGAAAATCTTCT
pFLAG-CTC RWD-C	CAAGCAGAAGACGGCATACGAGAT CAACTGGT GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT CTA CTGTATCAGGCTGAAAATCTTCT
pFLAG-CTC RWD-D	CAAGCAGAAGACGGCATACGAGAT CGTTCGTT GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT GATA CTGTATCAGGCTGAAAATCTTCT
pFLAG-CTC RWD-E	CAAGCAGAAGACGGCATACGAGAT CTGTTCAC GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT ACTCA CTGTATCAGGCTGAAAATCTTCT
pFLAG-CTC RWD-F	CAAGCAGAAGACGGCATACGAGAT GCTTGCAA GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT TTCTCT CTGTATCAGGCTGAAAATCTTCT
pFLAG-CTC RWD-G	CAAGCAGAAGACGGCATACGAGAT GTCAACTG GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT CACTTCT CTGTATCAGGCTGAAAATCTTCT

Table- D. Primers used for sequencing of pcDNA5/FRT/TO plasmid [CH2, CH3]

Primer Name	Sequence	Product length (bp)	Annealing temperature (°C)
CMV-FW [(1)]	CGCAAATGGGCGGTAGGCGTG	338	61.425
BGH-RV [(1)]	TAGAAGGCACAGTCGAGG		
pcDNA5-2-929F	TAGAAGACACCGGGACCGAT	993	59.25
pcDNA5-2-1921R	ATAGGTCAGGCTCTCGCTGA		
pcDNA5-3-1738F	CTCGGAGGGCGAAGAATCTC	908	59.825
pcDNA5-3-2645R	GGTTTCCACTATCGGCGAGT		
pcDNA5-4-2561F	ACTGTCGGGCGTACACAAAT	784	59.425
pcDNA5-4-3344R	TTTGCTGGCCTTTTGCTCAC		
pcDNA5-5-3261F	CAGCTCACTCAAAGGCGGTA	987	59.075
pcDNA5-5-4247R	AGCCCTCCCGTATCGTAGTT		
pcDNA5-6-4208F	TGACTCCCCGTCGTGTAGAT	998	59.425
pcDNA5-6-68R	CTATGCGGCATCAGAGCAGA		
pcDNA5-7-20F	CGATCCCCTATGGTGCACTC	928	59.9
pcDNA5-7-947R	TCGGTCCCGGTGTCTTCTAT		
pcDNA5-4-2561F	ACTGTCGGGCGTACACAAAT	456	59.4
M13-R-46 [(1)]	GAGCGGATAACAATTTCACACAGG		

[(1)] *Universal sequencing primers.*

Table- E. Primers used for amplification and cloning of GFP into pcDNA5/FRT/TO/GFP-STOP, control plasmid [CH2, CH3]

Primer Name	Sequence	Product length (bp)	Annealing temperature (°C)
GFP_AddgeneF1	GCTCGTTTAGTGAACCGTCAG	973	57.675
GFP_AddgeneR1	GCCACTGTGCTGGATATCTG		
GFP_AddgeneF1	GCTCGTTTAGTGAACCGTCAG	846	59.325
GFP_Rstop	AGTGGATCCCTAGTACAGCTCGTC		

Table- F. Primers used to amplify human housekeeping genes from FITR293 cells [CH2, CH3]. Highlighted primers were also used to amplify housekeeping genes from mouse DNA and cDNA.

Gene (Accession no.)	Primer Name	Sequence	Product length (bp) (RNA | DNA)	Annealing temperature (°C)
Actin beta (ACTB) (NG_007992.1)	ACTB_JF_F2	CTGGCACCACACCTTCTACA	183 | 624	58.275
	ACTB_JF_R2	CCAGAGGCGTACAGGGATAG		
Glyceraldehyde-3-phosphate Dehydrogenase (GAPDH) (NG_007073.2)	GAPDH_JF_F2	GGACCTGACCTGCCGTCTA	248 | 352	59.925
	GAPDH_JF_R2	CCACCACCCTGTTGCTGTAG		
TATA-box binding protein (TBP) (NG_008165.1)	TBP_KK_F1 [(1)]	TGAGCCAGAGTTATTTCCTGGT	166 | 847	57.4
	TBP_JF_R2	CGTCTTCCTGAATCCCTTTAGA		

[(1)] *Designed by Koray Kasan under my supervision*

Table- G. Primers used to amplify random oligonucleotide pool for library generation [CH2].

Primer Name	Sequence	Product length (bp)	Annealing temperature (°C)
EukLib1F	GACGATGTAGGTGACGAAGC	231	57.075
VR [(1)]	ATTACCGCCTTTGAGTGAGC		

[(1)] *Universal sequencing primer.*

Table- H. Primers used for two-step amplicon sequencing PCR 1 [CH2].

Primer Name	Heterogeneity Spacer	Sequence
SeqEL-F1		ACACTCTTTCCCTACACGACGCTCTTCCGATCT CCTCCGGACTCTAGCGTTTA
SeqEL-F2	T	ACACTCTTTCCCTACACGACGCTCTTCCGATCT T CCTCCGGACTCTAGCGTTTA
SeqEL-F3	GT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT GT CCTCCGGACTCTAGCGTTTA
SeqEL-F4	CGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCT CGA CCTCCGGACTCTAGCGTTTA
SeqEL-F5	ATGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCT ATGA CCTCCGGACTCTAGCGTTTA
SeqEL2-R1		GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT CAACAGATGGCTGGCAACTA
SeqEL2-R2	A	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT A CAACAGATGGCTGGCAACTA
SeqEL2-R3	TC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT TC CAACAGATGGCTGGCAACTA

SeqEL2-R4	CTA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT CTA CAACAGATGGCTGGCAACTA
SeqEL2-R5	GATA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT GATA CAACAGATGGCTGGCAACTA

Table- I. Primers used for two-step amplicon sequencing PCR 2 [CH2]. Illumina adaptor barcoded primers provided by Dr. Sven Künzel.

Primer Name	Illumina Adaptor	Barcode	Sequence
F-1	P5	AACCGCAT	AATGATACGGCGACCACCGAGATCTACAC AACCGCAT ACACTCTTTCCCTACACG
F-2	P5	AAGGCCTT	AATGATACGGCGACCACCGAGATCTACAC AAGGCCTT ACACTCTTTCCCTACACG
F-3	P5	AGAGTGTG	AATGATACGGCGACCACCGAGATCTACAC AGAGTGTG ACACTCTTTCCCTACACG
F-4	P5	CACAAGTC	AATGATACGGCGACCACCGAGATCTACAC CACAAGTC ACACTCTTTCCCTACACG
F-5	P5	CGTTCCTA	AATGATACGGCGACCACCGAGATCTACAC CGTTCCTA ACACTCTTTCCCTACACG
F-6	P5	GCTTGGAT	AATGATACGGCGACCACCGAGATCTACAC GCTTGGAT ACACTCTTTCCCTACACG
F-7	P5	GTCAACAC	AATGATACGGCGACCACCGAGATCTACAC GTCAACAC ACACTCTTTCCCTACACG
F-8	P5	GTCACTGA	AATGATACGGCGACCACCGAGATCTACAC GTCACTGA ACACTCTTTCCCTACACG
F-9	P5	TCTCGTCA	AATGATACGGCGACCACCGAGATCTACAC TCTCGTCA ACACTCTTTCCCTACACG
F-10	P5	TTGGTACG	AATGATACGGCGACCACCGAGATCTACAC TTGGTACG ACACTCTTTCCCTACACG
F-11	P5	CGTTGGAT	AATGATACGGCGACCACCGAGATCTACAC CGTTGGAT ACACTCTTTCCCTACACG
F-12	P5	CGTTAAGC	AATGATACGGCGACCACCGAGATCTACAC CGTTAAGC ACACTCTTTCCCTACACG
F-13	P5	ACAGCTCA	AATGATACGGCGACCACCGAGATCTACAC ACAGCTCA ACACTCTTTCCCTACACG
R-A	P7	AACCGGAA	CAAGCAGAAGACGGCATACGAGAT AACCGGAA GTGACTGGAGTTCAGACG
R-B	P7	AGAGTGAC	CAAGCAGAAGACGGCATACGAGAT AGAGTGAC GTGACTGGAGTTCAGACG
R-C	P7	CAACTGGT	CAAGCAGAAGACGGCATACGAGAT CAACTGGT GTGACTGGAGTTCAGACG
R-D	P7	CGTTCGTT	CAAGCAGAAGACGGCATACGAGAT CGTTCGTT GTGACTGGAGTTCAGACG
R-E	P7	CTGTTCAC	CAAGCAGAAGACGGCATACGAGAT CTGTTCAC GTGACTGGAGTTCAGACG
R-F	P7	GCTTGCAA	CAAGCAGAAGACGGCATACGAGAT GCTTGCAA GTGACTGGAGTTCAGACG
R-G	P7	GTCAACTG	CAAGCAGAAGACGGCATACGAGAT GTCAACTG GTGACTGGAGTTCAGACG
R-H	P7	TCCTCATG	CAAGCAGAAGACGGCATACGAGAT TCCTCATG GTGACTGGAGTTCAGACG
R-I	P7	TCGACTAG	CAAGCAGAAGACGGCATACGAGAT TCGACTAG GTGACTGGAGTTCAGACG
R-J	P7	TTGCAAGC	CAAGCAGAAGACGGCATACGAGAT TTGCAAGC GTGACTGGAGTTCAGACG
R-K	P7	AGAGGTGT	CAAGCAGAAGACGGCATACGAGAT AGAGGTGT GTGACTGGAGTTCAGACG
R-L	P7	GCTACGAT	CAAGCAGAAGACGGCATACGAGAT GCTACGAT GTGACTGGAGTTCAGACG
R-M	P7	GTCAAGAG	CAAGCAGAAGACGGCATACGAGAT GTCAAGAG GTGACTGGAGTTCAGACG
R-N	P7	ATGGTAGG	CAAGCAGAAGACGGCATACGAGAT ATGGTAGG GTGACTGGAGTTCAGACG
R-O	P7	GACTTCAG	CAAGCAGAAGACGGCATACGAGAT GACTTCAG GTGACTGGAGTTCAGACG

Table- J. Primers used for Mdng amplification from mouse samples [CH3].

Primer Name	Sequence	Product length (bp)	Annealing temperature (°C)
Mdng1_F	CAGCACAGCCTTGTTTGTTGA	659	59.875
Mdng1_R	TGTCACGGGTTTTGGAGGTC		
Mdng2_F	CAAGCTTTAAAAGGACAACATGG	594	57.05
Mdng2_R	TCAAGTGTGGCGTGTATCCT		
Mdng3_F	CTGCAGAACCTCTTCTTTGGA	597	57.375
Mdng3_R	TCCCACGGTGTGAATTATCC		
Mdng4_F	ATGTGCTGTGTGTTTACACATTT	399	55.4
Mdng4_R	AAGCAAATTAACATAGTCTGTGGTT		
Mdng5_F	ATTCTCCCTGGTGACAGGTG	391	57.9
Mdng5_R	TCTTTCTGGCCTCGATTCTG		
Mdng6_F	GAATTCGCTTGGTCTCATCC	248	57.675
Mdng6_R	GGCACCTGGTCCTCTGACT		
Mdng7_F	TGATCTCGGGGACACAGG	500	56.525
Mdng7_R	AGAAAAAATTGGCTAGACTTAAGAAAG		
Mdng8_F	AGTGTCCGCTGGAGTTGC	237	58.4
Mdng8_R	ATCCAGGAGAGCTGTTTCCA		
Mdng9_F	TCTTCCACCTGGATGACTCC	466	57.025
Mdng9_R	CGCTCATGAACTCCCAATCT		
Mdng10_F	ATGTCACATTCAAGCCACTCA	498	57.025
Mdng10_R	GGATGGATTGGTCTCCATTCT		
Mdng11_F	GGTTGCTAGGTGGGTGTGTT	547	59.05
Mdng11_R	CCTGCCACCAAACCAGATGA		
Mdng12_F	CGACGGGCTTAGATTCTGCT	619	59.975
Mdng12_R	CTTCCAGGGCTCAATGGGTT		
Mdng13_F	TCCCCTGGGACTCGAGTTAT	418	57.1
Mdng13_R	CAAATACACACAGATTCTTACTGGA		
Mdng14_F	AGGCTTGGACTCCTAATTGCAA	414	58.825
Mdng14_R	ACCAGGACAGAAACCAGCTC		

Table- K. Primers used to add restriction sites to amplified Mdng [CH3].

Primer Name	Sequence	Product length (bp)	Annealing temperature (°C)
Mdng1_F-HindIII	CATGACAAGCTTATGAGCTCCAG	479	57.775
Mdng1_R-XhoI	CATACTCGAGGGCATCCTG		
Mdng2_F-HindIII	CAAGCTTTAAAAGGACAACATGG	594	57.575
Mdng2_R-BamHI	CAAGTGTGGCGTGGATCC		
Mdng3_F-HindIII	CTGCAGAAGCTTTTCTTTGGTT	591	58.45
Mdng3_R-BamHI	TGTGAAGGATCCGCACACTC		

Mdng4_F-HindIII	GTGCTGAAGCTTTACACATTTATGC	404	58.275
Mdng4_R-BamHI	GCCGAAGGATCCAATTAACATAGT		
Mdng5_F-HindIII	ATACAGAAGCTTGACAGGTGAACA	381	57.675
Mdng5_R-BamHI	CTCGATGGATCCTGTTTCAGA		
Mdng6_F-HindIII	GTGTTGTAAGCTTCCGGGATGT	184	58.825
Mdng6_R-BamHI	GCACCGGATCCTCTGACT		
Mdng7_F-HindIII	GGACACAAGCTTGGGAAATGT	463	58.675
Mdng7_R-XhoI	GGCCTCGAGATCTGTTATCCA		
Mdng8_F-HindIII	CCGCTAAGCTTATCCTATGTTC	204	56.875
Mdng8_R-BamHI	CTCTGTGGATCCCAAAATCTTCA		
Mdng9_F-HindIII	TTCATAAGCTTCCCGCCAAT	434	57.5
Mdng9_R-BamHI	GTTTGGGGATCCAGCTAAGATA		
Mdng10_F-HindIII	GGGTAAAAGCTTATGTCACATTCAA	511	57.875
Mdng10_R-XhoI	CCTTTCTCGAGGTCTCCATTCT		
Mdng11_F-HindIII	GTTAAGCTTGATGTTGCAGCGTA	307	57.65
Mdng11_R-XhoI	GAGACTCGAGAGCTCCACAT		
Mdng12_F-HindIII	GTTAAGCTTATGCCTCCCTTGAA	504	58.075
Mdng12_R-BamHI	GAAGGATCCATGTGTTAGTTTCCA		
Mdng13_F-HindIII	GCTAAGCTTGTAAGAGCGACAA	354	56.475
Mdng13_R-XhoI	CTTAGCTCGAGTTCTTACTGGATT		
Mdng14_F-HindIII	CTCATTAAGCTTTGCATCTCATTTACA	350	58.1
Mdng14_R-BamHI	CAAGGATCCTCACAGGTTTCAA		

General discussion

With the discovery that *de novo* gene birth is a real, and important source of genetic innovation, comes the work of understanding how this process works. There are still many things unknown about the way in which novel genes transition from non-coding to coding sequences. The main goal of this thesis was to address three of such questions about the process of *de novo* gene birth. The first question was whether there are any features of a random sequence that would make it more likely to have a positive effect on the growth of prokaryotic cells (Chapter 1). The second question was whether there are any differences between prokaryotes and eukaryotes in the types and numbers of sequences that could be used as *de novo* gene material (Chapter 2). The third question was how does a novel sequence, already exposed to selection interact with the genes already present in the organism and the regulatory machinery (Chapter 3). The specific results and limitations have already been discussed in each chapter, so I will focus this discussion on how they are related, and their implications on the study of gene evolution.

It would seem that there are different requirements for whether a novel sequence is well tolerated in bacteria or in eukaryotes. The driving factor in bacteria seems to be peptide length, while no specific factor could be identified for the eukaryotic sequences. This finding matches the prediction that costs associated to expressing a gene are smaller in eukaryotes than prokaryotes. This prediction includes the costs of translation—the longer the gene, the costlier it is to express it in bacteria (Lynch & Marinov, 2015). On the other hand, the correlations between length, disorder, GC content and aggregation propensity are well known (Angyan et al., 2012; Basile et al., 2017; Li et al., 2015; Oliver & Marin, 1996). Therefore, it is not surprising that no other predicting factors could be identified.

Based on what has been shown in the literature regarding the tolerance of random sequences by cells (Neme et al., 2017; Tretyachenko et al., 2017), my general hypothesis was that cells would tolerate the expression of random sequences, even more so in the eukaryotic cells than the prokaryotic ones. This was also based on the idea that, since eukaryotic cells have more non-coding sequences, they could be better at managing spurious expression. In the case of the mouse *de novo* genes, the null hypothesis was that their expression would not have any effect on the transcriptome. The results, however only matched my initial hypotheses partially. In the case of the random libraries, the cells did indeed have a high tolerance for their expression, with over half of the sequences being maintained in the populations with no apparent negative effects on growth. However, there is no indication that the eukaryotic cells are any more tolerant to the expression of random sequences than the prokaryotic ones. For the mouse *de novo* genes expressed in a human cell line it was not possible to reject the null hypothesis. This means that either the effect is too small to be detected, or there is no effect—which in itself is surprising.

These results have further implications for the study of gene evolution. First, the fact that sequences spanning the whole range of intrinsic disorder scores could have positive or non-significant effects on cell growth suggests that intrinsic disorder is not a limiting factor for the evolution of *de novo* genes. Also, the mouse *de novo* genes had either high (Mdng2) or low (Mdng7 and Mdng10) disorder scores, and showed no significant effects in either case. Of course, there are examples of ordered and disordered functional proteins both ancient and novel. There is at least one example of a *de novo* protein having a defined secondary structure and some "rudimentary fold" (Bungard et al., 2017). What is interesting, is that, although younger genes tend to be more disordered, disorder itself is not a requirement for their birth.

This thesis also joins the voice of many authors in stressing that it is important to expand the study of protein function and evolution to short peptides (Andrews & Rothnagel, 2014; Bazzini et al., 2014; Khitun et al., 2019; Mackowiak et al., 2015; Orr et al., 2020; Storz et al., 2014). Here, peptides—between 4 and 60 residues in length—with apparent positive or negative effects on cell growth could be identified in both types of cells. It has been shown that the entire genome of at least several organisms is transcribed at some point. Many of the transcription products contain sORFs which could be translated and, as shown here, exposed to selection regardless of how small. This could mean that the entire genome is selected against coding for deleterious peptides, even when those peptides are never actually fixed in the population.

Finally, I presented in this thesis evidence that the exploration of sequence space through expression of random peptides is possible in living cells. These "junk polypeptides"—with low or null positive fitness effects—are better at exploring sequence space. This would facilitate the evolution of complex adaptations to escape local fitness peaks, in a similar way that non-adaptive mutations or pre-adaptations do (Nielly-Thibault & Landry, 2019; Pal & Papp, 2017). *De novo* genes or random sequences could serve as stepping stones in this process, which is why having more non-coding sequences provides the variability necessary to explore more possible phenotypes (Knibbe et al., 2007).

Future directions

The results presented here have, undoubtedly, generated many new questions which might prove useful in directing future efforts to study the evolution of new genes. Future studies using random sequences and heterologous expression systems should focus on addressing some of the variables that could not be controlled in the experimental work presented here.

Given that the experiments described in this thesis are in general proofs of principle, replication studies for all of them would be useful. Using different candidates, newly synthesised pools of oligonucleotides, and different cell lines or bacteria strains, would allow us to understand to what degree the results obtained here can be generalised. For example, although I have demonstrated that both random sequence libraries have the expected features of random sequences, they contain only a few thousand clones each. Since this is just a small part of the possible sequence space (2.04×10^9 sequences), it is possible that we are not observing a representative sample. It would be interesting to see whether such experiments recover the same trends described here.

The same considerations apply to the mouse *de novo* genes. It will be necessary to take a look at the effect of more candidates, and compare them to the effects of murine cells, for example. Considering that *de novo* genes are expressed in a tissue-specific manner, looking at their effects on different types of cells or developmental stages could be informative.

For the eukaryotic library, it would also be interesting to evaluate whether the selected Kozak sequence has an effect on the trends and effects seen here. Also, a more specific study of the clones that showed differences between the induced and uninduced treatments, could help to identify the reason for the lack of regulation of the tetON system, that worked well with the mouse *de novo* genes and the GFP control cells.

The datasets and genetic material generated in these projects could be used for other types of analyses as well. In particular, it would be interesting to attempt a machine learning or HMM approach to study the random sequences and their group assignments. Such approaches could be helpful to disentangle the correlations between the different features of the sequences and their effect on cell growth. Furthermore, the short peptides with potential positive or negative effects on growth could be used to inform future research on drug discovery or cell growth assays.

Concluding remarks

In his influential essay, “Evolution and tinkering” François Jacob described Evolution as a “tinkerer” that takes existing parts and pieces of existing genetic information in organisms and uses it to create novelty (Jacob, 1977). He used as examples well know and frequent evolutionary processes, and reached the conclusion that innovation could not happen “*de novo*” from random pieces of sequence without a pre-existing function. The fact that gene duplication and neo-functionalization, or mutations in regulatory sequences are indeed the most common mechanisms through which new genes are born seemed to back this statement.

But the metaphor would be incomplete if we ignored the tinkerer’s ability to take pieces of apparent ”junk” and transforming them into important elements of their creations. After all, the most elemental parts that make all matter follow the rules of physics and chemistry, and therefore, although it might be unlikely that a new type of nucleotide could arise from atoms, the probabilities of getting a functional sequence of translated amino acids are orders of magnitude higher. Why, then, would evolution not tinker with the abundance of non-coding sequences present in eukaryotic genomes as well? In fact, it is now a well-known fact that *de novo* gene birth is not only possible, but also pervasive in all domains of life.

We are just now beginning to understand the fascinating complexity and diversity of mechanisms through which genetic novelty arises. Given the sheer magnitude of the sequence space, it would seem that any mechanism that allows for its exploration, no matter how unlikely could yield functional genes.

List of Figures and Tables

1. Figures

2. Tables

Bibliography

Abrusan, G. (2013). Integration of new genes into cellular networks, and their structural maturation. *Genetics*, *195*(4), 1407-1417. https://doi.org/10.1534/genetics.113.152256

Aguilera, F., McDougall, C., & Degnan, B. M. (2017). Co-Option and *De Novo* Gene Evolution Underlie Molluscan Shell Diversity. *Mol Biol Evol*, *34*(4), 779-792. https://doi.org/10.1093/molbev/msw294

Andrews, S. J., & Rothnagel, J. A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. *Nat Rev Genet*, *15*(3), 193-204. https://doi.org/10.1038/nrg3520

Angyan, A. F., Perczel, A., & Gaspari, Z. (2012). Estimating intrinsic structural preferences of *de novo* emerging random-sequence proteins: is aggregation the main bottleneck? *FEBS Lett*, *586*(16), 2468-2472. https://doi.org/10.1016/j.febslet.2012.06.007

Arendsee, Z. W., Li, L., & Wurtele, E. S. (2014). Coming of age: orphan genes in plants. *Trends Plant Sci*, *19*(11), 698-708. https://doi.org/10.1016/j.tplants.2014.07.003

Baalsrud, H. T., Torresen, O. K., Solbakken, M. H., Salzburger, W., Hanel, R., Jakobsen, K. S., & Jentoft, S. (2018). *De Novo* Gene Evolution of Antifreeze Glycoproteins in Codfishes Revealed by Whole Genome Sequence Data. *Mol Biol Evol*, *35*(3), 593-606. https://doi.org/10.1093/molbev/msx311

Basile, W., Sachenkova, O., Light, S., & Elofsson, A. (2017). High GC content causes orphan proteins to be intrinsically disordered. *PLoS Comput Biol*, *13*(3), e1005375. https://doi.org/10.1371/journal.pcbi.1005375

Bazzini, A. A., Johnstone, T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., Fleming, E. S., Vejnar, C. E., Lee, M. T., Rajewsky, N., Walther, T. C., & Giraldez, A. J. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J*, *33*(9), 981-993. https://doi.org/10.1002/embj.201488411

Bekpen, C., Xie, C., & Tautz, D. (2018). Dealing with the adaptive immune system during *de novo* evolution of genes from intergenic sequences. *BMC Evol Biol*, *18*(1), 121. https://doi.org/10.1186/s12862-018-1232-z

Bernardi, G. (2019). The Genomic Code: A Pervasive Encoding/Molding of Chromatin Structures and a Solution of the "Non-Coding DNA" Mystery. *Bioessays*, *41*(12), e1900106. https://doi.org/10.1002/bies.201900106

Berretta, J., & Morillon, A. (2009). Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO Rep*, *10*(9), 973-982. https://doi.org/10.1038/embor.2009.181

Bird, C. P., Stranger, B. E., & Dermitzakis, E. T. (2006). Functional variation and evolution of non-coding DNA. *Curr Opin Genet Dev*, *16*(6), 559-564. https://doi.org/10.1016/j.gde.2006.10.003

Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G. A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D. H., Letunic, I., Marchler-Bauer, A., . . . Finn, R. D. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res*, *49*(D1), D344-D354. https://doi.org/10.1093/nar/gkaa977

Bornberg-Bauer, E., & Heames, B. (2019). Becoming a *de novo* gene. *Nat Ecol Evol*, *3*(4), 524-525. https://doi.org/10.1038/s41559-019-0845-y

Bornberg-Bauer, E., Schmitz, J., & Heberlein, M. (2015). Emergence of *de novo* proteins from 'dark genomic matter' by 'grow slow and moult'. *Biochem Soc Trans*, *43*(5), 867-873. https://doi.org/10.1042/BST20150089

Bult, C. J., Blake, J. A., Smith, C. L., Kadin, J. A., Richardson, J. E., & Mouse Genome Database, G. (2019). Mouse Genome Database (MGD) 2019. *Nucleic Acids Res*, *47*(D1), D801-D806. https://doi.org/10.1093/nar/gky1056

Bungard, D., Copple, J. S., Yan, J., Chhun, J. J., Kumirov, V. K., Foy, S. G., Masel, J., Wysocki, V. H., & Cordes, M. H. J. (2017). Foldability of a Natural *De Novo* Evolved Protein. *Structure*, *25*(11), 1687-1696 e1684. https://doi.org/10.1016/j.str.2017.09.006

Cai, J., Zhao, R., Jiang, H., & Wang, W. (2008). *De novo* origination of a new protein-coding gene in Saccharomyces cerevisiae. *Genetics*, *179*(1), 487-496. https://doi.org/10.1534/genetics.107.084491

Carelli, F. N., Liechti, A., Halbert, J., Warnefors, M., & Kaessmann, H. (2018). Repurposing of promoters and enhancers during mammalian evolution. *Nat Commun*, *9*(1), 4066. https://doi.org/10.1038/s41467-018-06544-z

Carvunis, A. R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B., Brar, G. A., Weissman, J. S., Regev, A., Thierry-Mieg, N., Cusick, M. E., & Vidal, M. (2012). Proto-genes and *de novo* gene birth. *Nature*, *487*(7407), 370-374. https://doi.org/10.1038/nature11184

Chen, G., & Deng, X. (2018). Cell Synchronization by Double Thymidine Block. *Bio Protoc*, *8*(17). https://doi.org/10.21769/BioProtoc.2994

Chen, S., Krinsky, B. H., & Long, M. (2013). New genes as drivers of phenotypic evolution. *Nat Rev Genet*, *14*(9), 645-660. https://doi.org/10.1038/nrg3521

Chiarabelli, C. (2006). On the Folding Frequency in a Totally Random Library of de novo Proteins Obtained by Phage Display.

Ching, T., Huang, S., & Garmire, L. X. (2014). Power analysis and sample size estimation for RNA-Seq differential expression. *RNA*, *20*(11), 1684-1696. https://doi.org/10.1261/rna.046011.114

Daubin, V., & Ochman, H. (2004a). Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. *Genome Res*, *14*(6), 1036-1042. https://doi.org/10.1101/gr.2231904

Daubin, V., & Ochman, H. (2004b). Start-up entities in the origin of new genes. *Curr Opin Genet Dev*, *14*(6), 616-619. https://doi.org/10.1016/j.gde.2004.09.004

de Boer, C. G., Vaishnav, E. D., Sadeh, R., Abeyta, E. L., Friedman, N., & Regev, A. (2020). Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat Biotechnol*, *38*(1), 56-65. https://doi.org/10.1038/s41587-019-0315-8

Domazet-Loso, T., Brajkovic, J., & Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet*, *23*(11), 533-539. https://doi.org/10.1016/j.tig.2007.08.014

Durand, E., Gagnon-Arsenault, I., Hallin, J., Hatin, I., Dube, A. K., Nielly-Thibault, L., Namy, O., & Landry, C. R. (2019). Turnover of ribosome-associated transcripts from *de novo* ORFs produces gene-like characteristics available for *de novo* gene emergence in wild yeast populations. *Genome Res*, *29*(6), 932-943. https://doi.org/10.1101/gr.239822.118

Elena, S. F., Cooper, V. S., & Lenski, R. E. (1996). Punctuated evolution caused by selection of rare beneficial mutations. *Science*, *272*(5269), 1802-1804. https://doi.org/10.1126/science.272.5269.1802

Goeke, D., Kaspar, D., Stoeckle, C., Grubmuller, S., Berens, C., Klotzsche, M., & Hillen, W. (2012). Short peptides act as inducers, anti-inducers and corepressors of Tet repressor. *J Mol Biol*, *416*(1), 33-45. https://doi.org/10.1016/j.jmb.2011.12.009

Goodman, M. (1981). Decoding the pattern of protein evolution. *Prog Biophys Mol Biol*, *38*(2), 105-164. https://doi.org/10.1016/0079-6107(81)90012-2

Gubala, A. M., Schmitz, J. F., Kearns, M. J., Vinh, T. T., Bornberg-Bauer, E., Wolfner, M. F., & Findlay, G. D. (2017). The Goddard and Saturn Genes Are Essential for Drosophila Male Fertility and May Have Arisen *De Novo*. *Mol Biol Evol*, *34*(5), 1066-1082. https://doi.org/10.1093/molbev/msx057

Guerzoni, D., & McLysaght, A. (2016). *De Novo* Genes Arise at a Slow but Steady Rate along the Primate Lineage and Have Been Subject to Incomplete Lineage Sorting. *Genome Biol Evol*, *8*(4), 1222-1232. https://doi.org/10.1093/gbe/evw074

Hayashi, Y., Sakata, H., Makino, Y., Urabe, I., & Yomo, T. (2003). Can an arbitrary sequence evolve towards acquiring a biological function? *J Mol Evol*, *56*(2), 162-168. https://doi.org/10.1007/s00239-002-2389-y

Heames, B., Schmitz, J., & Bornberg-Bauer, E. (2020). A Continuum of Evolving *De Novo* Genes Drives Protein-Coding Novelty in Drosophila. *J Mol Evol*, *88*(4), 382-398. https://doi.org/10.1007/s00239-020-09939-z

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., . . . Flicek, P. (2021). Ensembl 2021. *Nucleic Acids Res*, *49*(D1), D884-D891. https://doi.org/10.1093/nar/gkaa942

Ilik, I. A., Malszycki, M., Lubke, A. K., Schade, C., Meierhofer, D., & Aktas, T. (2020). SON and SRRM2 are essential for nuclear speckle formation. *Elife*, *9*. https://doi.org/10.7554/eLife.60579

Jacob, F. (1977). Evolution and tinkering. *Science*, *196*(4295), 1161-1166. https://doi.org/10.1126/science.860134

Jacquier, A. (2009). The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat Rev Genet*, *10*(12), 833-844. https://doi.org/10.1038/nrg2683

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. *Genome Res*, *20*(10), 1313-1326. https://doi.org/10.1101/gr.101386.109

Keefe, A. D., & Szostak, J. W. (2001). Functional proteins from a random-sequence library. *Nature*, *410*(6829), 715-718. https://doi.org/10.1038/35070613

Keeling, P. J., & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet*, *9*(8), 605-618. https://doi.org/10.1038/nrg2386

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R., & Bosch, T. C. G. (2009). More than just orphans: are taxonomically-restricted genes important in evolution? *Trends in Genetics*, *25*(9), 404-413. https://doi.org/10.1016/j.tig.2009.07.006

Khitun, A., Ness, T. J., & Slavoff, S. A. (2019). Small open reading frames and cellular stress responses. *Mol Omics*, *15*(2), 108-116. https://doi.org/10.1039/c8mo00283e

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*, *37*(8), 907-915. https://doi.org/10.1038/s41587-019-0201-4

Klasberg, S., Bitard-Feildel, T., Callebaut, I., & Bornberg-Bauer, E. (2018). Origins and structural properties of novel and *de novo* protein domains during insect evolution. *FEBS J*, *285*(14), 2605-2625. https://doi.org/10.1111/febs.14504

Klotzsche, M., Berens, C., & Hillen, W. (2005). A peptide triggers allostery in tet repressor by binding to a unique site. *J Biol Chem*, *280*(26), 24591-24599. https://doi.org/10.1074/jbc.M501872200

Knibbe, C., Coulon, A., Mazet, O., Fayard, J. M., & Beslon, G. (2007). A long-term evolutionary pressure on the amount of noncoding DNA. *Mol Biol Evol*, *24*(10), 2344-2353. https://doi.org/10.1093/molbev/msm165

Knopp, M. (2019). *De novo* emergence of peptides that confer antibiotic resistance. https://doi.org/10.1128/mBio

Knopp, M., Babina, A. M., Gudmundsdottir, J. S., Douglass, M. V., Trent, M. S., & Andersson, D. I. (2021). A novel type of colistin resistance genes selected from random sequence space. *PLoS Genet*, *17*(1), e1009227. https://doi.org/10.1371/journal.pgen.1009227

Knowles, D. G., & McLysaght, A. (2009). Recent *de novo* origin of human protein-coding genes. *Genome Res*, *19*(10), 1752-1759. https://doi.org/10.1101/gr.095026.109

Lange, A., Patel, P. H., Heames, B., Damry, A. M., Saenger, T., Jackson, C. J., Findlay, G. D., & Bornberg-Bauer, E. (2021). Structural and functional characterization of a putative *de novo* gene in Drosophila. *Nat Commun*, *12*(1), 1667. https://doi.org/10.1038/s41467-021-21667-6

Li, D., Yan, Z., Lu, L., Jiang, H., & Wang, W. (2014). Pleiotropy of the *de novo*-originated gene MDF1. *Sci Rep*, *4*, 7280. https://doi.org/10.1038/srep07280

Li, J., Zhou, J., Wu, Y., Yang, S., & Tian, D. (2015). GC-Content of Synonymous Codons Profoundly Influences Amino Acid Usage. *G3 (Bethesda)*, *5*(10), 2027-2036. https://doi.org/10.1534/g3.115.019877

Li, Z. W., Chen, X., Wu, Q., Hagmann, J., Han, T. S., Zou, Y. P., Ge, S., & Guo, Y. L. (2016). On the Origin of *De Novo* Genes in Arabidopsis thaliana Populations. *Genome Biol Evol*, *8*(7), 2190-2202. https://doi.org/10.1093/gbe/evw164

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923-930. https://doi.org/10.1093/bioinformatics/btt656

Light, S., Basile, W., & Elofsson, A. (2014). Orphans and new gene origination, a structural and evolutionary perspective. *Curr Opin Struct Biol*, *26*, 73-83. https://doi.org/10.1016/j.sbi.2014.05.006

Lin, Y. C., Boone, M., Meuris, L., Lemmens, I., Van Roy, N., Soete, A., Reumers, J., Moisse, M., Plaisance, S., Drmanac, R., Chen, J., Speleman, F., Lambrechts, D., Van de Peer, Y., Tavernier, J., & Callewaert, N. (2014). Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nat Commun*, *5*, 4767. https://doi.org/10.1038/ncomms5767

Long, M., VanKuren, N. W., Chen, S., & Vibranovski, M. D. (2013). New gene evolution: little did we know. *Annu Rev Genet*, *47*, 307-333. https://doi.org/10.1146/annurev-genet-111212-133301

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), Article 550. https://doi.org/10.1186/s13059-014-0550-8

Luis Villanueva-Canas, J., Ruiz-Orera, J., Agea, M. I., Gallo, M., Andreu, D., & Alba, M. M. (2017). New Genes and Functional Innovation in Mammals. *Genome Biol Evol*, *9*(7), 1886-1900. https://doi.org/10.1093/gbe/evx136

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, *290*(5494), 1151-1155. https://doi.org/10.1126/science.290.5494.1151

Lynch, M., & Marinov, G. K. (2015). The bioenergetic costs of a gene. *Proc Natl Acad Sci U S A*, *112*(51), 15690-15695. https://doi.org/10.1073/pnas.1514974112

Mackowiak, S. D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., Mastrobuoni, G., Rajewsky, N., Kempa, S., Selbach, M., & Obermayer, B. (2015). Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol*, *16*, 179. https://doi.org/10.1186/s13059-015-0742-x

Majic, P., & Payne, J. L. (2020). Enhancers Facilitate the Birth of *De Novo* Genes and Gene Integration into Regulatory Networks. *Mol Biol Evol*, *37*(4), 1165-1178. https://doi.org/10.1093/molbev/msz300

McLysaght, A., & Guerzoni, D. (2015). New genes from non-coding sequence: the role of *de novo* protein-coding genes in eukaryotic evolutionary innovation. *Philos Trans R Soc Lond B Biol Sci*, *370*(1678), 20140332. https://doi.org/10.1098/rstb.2014.0332

McLysaght, A., & Hurst, L. D. (2016). Open questions in the study of *de novo* genes: what, how and why. *Nat Rev Genet*, *17*(9), 567-578. https://doi.org/10.1038/nrg.2016.78

Muhoberac, B. B., & Vidal, R. (2019). Iron, Ferritin, Hereditary Ferritinopathy, and Neurodegeneration. *Front Neurosci*, *13*, 1195. https://doi.org/10.3389/fnins.2019.01195

Nakashima, T., Toyota, H., Urabe, I., & Yomo, T. (2007). Effective selection system for experimental evolution of random polypeptides towards DNA-binding protein. *J Biosci Bioeng*, *103*(2), 155-160. https://doi.org/10.1263/jbb.103.155

Neme, R., Amador, C., Yildirim, B., McConnell, E., & Tautz, D. (2017). Random sequences are an abundant source of bioactive RNAs or peptides. *Nat Ecol Evol*, *1*(6), 0217. https://doi.org/10.1038/s41559-017-0127

Neme, R., & Tautz, D. (2016). Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to *de novo* gene emergence. *Elife*, *5*, e09977. https://doi.org/10.7554/eLife.09977

Nguyen, F., Starosta, A. L., Arenz, S., Sohmen, D., Donhofer, A., & Wilson, D. N. (2014). Tetracycline antibiotics and resistance mechanisms. *Biol Chem*, *395*(5), 559-575. https://doi.org/10.1515/hsz-2013-0292

Nielly-Thibault, L., & Landry, C. R. (2019). Differences Between the Raw Material and the Products of *de Novo* Gene Birth Can Result from Mutational Biases. *Genetics*, *212*(4), 1353-1366. https://doi.org/10.1534/genetics.119.302187

Ohno, S., Wolf, U., & Atkin, N. B. (1968). Evolution from fish to mammals by gene duplication. *Hereditas*, *59*(1), 169-187. https://doi.org/10.1111/j.1601-5223.1968.tb02169.x

Oliver, J. L., & Marin, A. (1996). A relationship between GC content and coding-sequence length. *J Mol Evol*, *43*(3), 216-223. https://doi.org/10.1007/BF02338829

Orr, M. W., Mao, Y., Storz, G., & Qian, S. B. (2020). Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Res*, *48*(3), 1029-1042. https://doi.org/10.1093/nar/gkz734

Pal, C., & Papp, B. (2017). Evolution of complex adaptations in molecular systems. *Nat Ecol Evol*, *1*(8), 1084-1092. https://doi.org/10.1038/s41559-017-0228-1

Prabh, N., & Rodelsperger, C. (2019). *De Novo*, Divergence, and Mixed Origin Contribute to the Emergence of Orphan Genes in Pristionchus Nematodes. *G3 (Bethesda)*, *9*(7), 2277-2286. https://doi.org/10.1534/g3.119.400326

Prince, V. E., & Pickett, F. B. (2002). Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet*, *3*(11), 827-837. https://doi.org/10.1038/nrg928

Rebeiz, M., & Tsiantis, M. (2017). Enhancer evolution and the origins of morphological novelty. *Curr Opin Genet Dev*, *45*, 115-123. https://doi.org/10.1016/j.gde.2017.04.006

Reinhardt, J. A., Wanjiru, B. M., Brant, A. T., Saelao, P., Begun, D. J., & Jones, C. D. (2013). *De novo* ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genet*, *9*(10), e1003860. https://doi.org/10.1371/journal.pgen.1003860

Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabido, E., Kondova, I., Bontrop, R., Marques-Bonet, T., & Alba, M. M. (2015). Origins of *De Novo* Genes in Human and Chimpanzee. *PLoS Genet*, *11*(12), e1005721. https://doi.org/10.1371/journal.pgen.1005721

Ruiz-Orera, J., Messeguer, X., Subirana, J. A., & Alba, M. M. (2014). Long non-coding RNAs as a source of new peptides. *Elife*, *3*, e03523. https://doi.org/10.7554/eLife.03523

Ruiz-Orera, J., Verdaguer-Grau, P., Villanueva-Canas, J. L., Messeguer, X., & Alba, M. M. (2018). Translation of neutrally evolving peptides provides a basis for *de novo* gene evolution. *Nat Ecol Evol*, *2*(5), 890-896. https://doi.org/10.1038/s41559-018-0506-6

Sabath, N., Wagner, A., & Karlin, D. (2012). Evolution of viral proteins originated *de novo* by overprinting. *Mol Biol Evol*, *29*(12), 3767-3780. https://doi.org/10.1093/molbev/mss179

Sanchez, G., Linde, S. C., & Coolon, J. D. (2020). Genome-wide effect of tetracycline, doxycycline and 4-epidoxycycline on gene expression in Saccharomyces cerevisiae. *Yeast*, *37*(7-8), 389-396. https://doi.org/10.1002/yea.3515

Schlotterer, C. (2015). Genes from scratch--the evolutionary fate of *de novo* genes. *Trends Genet*, *31*(4), 215-219. https://doi.org/10.1016/j.tig.2015.02.007

Schmitz, J. F., & Bornberg-Bauer, E. (2017). Fact or fiction: updates on how protein-coding genes might emerge *de novo* from previously non-coding DNA. *F1000Res*, *6*, 57. https://doi.org/10.12688/f1000research.10079.1

Schmitz, J. F., Chain, F. J. J., & Bornberg-Bauer, E. (2020). Evolution of novel genes in three-spined stickleback populations. *Heredity (Edinb)*, *125*(1-2), 50-59. https://doi.org/10.1038/s41437-020-0319-7

Schmitz, J. F., Ullrich, K. K., & Bornberg-Bauer, E. (2018). Incipient *de novo* genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat Ecol Evol*, *2*(10), 1626-1632. https://doi.org/10.1038/s41559-018-0639-7

Snel, B., Bork, P., & Huynen, M. A. (2002). Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res*, *12*(1), 17-25. https://doi.org/10.1101/gr.176501

Stepanenko, A. A., & Dmitrenko, V. V. (2015). HEK293 in cell biology and cancer research: phenotype, karyotype, tumorigenicity, and stress-induced genome-phenotype evolution. *Gene*, *569*(2), 182-190. https://doi.org/10.1016/j.gene.2015.05.065

Stepanov, V. G., & Fox, G. E. (2007). Stress-driven in vivo selection of a functional mini-gene from a randomized DNA library expressing combinatorial peptides in Escherichia coli. *Mol Biol Evol*, *24*(7), 1480-1491. https://doi.org/10.1093/molbev/msm067

Storz, G., Wolf, Y. I., & Ramamurthi, K. S. (2014). Small proteins can no longer be ignored. *Annu Rev Biochem*, *83*, 753-777. https://doi.org/10.1146/annurev-biochem-070611-102400

Tautz, D. (2014). The discovery of *de novo* gene evolution. *Perspect Biol Med*, *57*(1), 149-161. https://doi.org/10.1353/pbm.2014.0006

Tautz, D., & Domazet-Loso, T. (2011). The evolutionary origin of orphan genes. *Nat Rev Genet*, *12*(10), 692-702. https://doi.org/10.1038/nrg3053

Tretyachenko, V., Vymetal, J., Bednarova, L., Kopecky, V., Jr., Hofbauerova, K., Jindrova, H., Hubalek, M., Soucek, R., Konvalinka, J., Vondrasek, J., & Hlouchova, K. (2017). Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Sci Rep*, *7*(1), 15449. https://doi.org/10.1038/s41598-017-15635-8

Vakirlis, N., Hebert, A. S., Opulente, D. A., Achaz, G., Hittinger, C. T., Fischer, G., Coon, J. J., & Lafontaine, I. (2018). A Molecular Portrait of *De Novo* Genes in Yeasts. *Mol Biol Evol*, *35*(3), 631-645. https://doi.org/10.1093/molbev/msx315

Van Oss, S. B., & Carvunis, A. R. (2019). *De novo* gene birth. *PLoS Genet*, *15*(5), e1008160. https://doi.org/10.1371/journal.pgen.1008160

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R., Erichsen, J. T., Jasinska, A. J., Turner, J. M., Bertelsen, M. F., Murchison, E. P., Flicek, P., & Odom, D. T. (2015). Enhancer evolution across 20 mammalian species. *Cell*, *160*(3), 554-566. https://doi.org/10.1016/j.cell.2015.01.006

Waghu, F. H., Barai, R. S., Gurung, P., & Idicula-Thomas, S. (2016). CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Res*, *44*(D1), D1094-1097. https://doi.org/10.1093/nar/gkv1051

Wang, G., Sun, S., & Zhang, Z. (2016). Randomness in Sequence Evolution Increases over Time. *PLoS One*, *11*(5), e0155935. https://doi.org/10.1371/journal.pone.0155935

Wang, S., Mao, C., & Liu, S. (2019). Peptides encoded by noncoding genes: challenges and perspectives. *Signal Transduct Target Ther*, *4*, 57. https://doi.org/10.1038/s41392-019-0092-3

Wang, W., Min, L., Qiu, X., Wu, X., Liu, C., Ma, J., Zhang, D., & Zhu, L. (2021). Biological Function of Long Non-coding RNA (LncRNA) Xist. *Front Cell Dev Biol*, *9*, 645647. https://doi.org/10.3389/fcell.2021.645647

Wang, Y. W., Hess, J., Slot, J. C., & Pringle, A. (2020). *De Novo* Gene Birth, Horizontal Gene Transfer, and Gene Duplication as Sources of New Gene Families Associated with the Origin of Symbiosis in Amanita. *Genome Biol Evol*, *12*(11), 2168-2182. https://doi.org/10.1093/gbe/evaa193

White, S. H., & Jacobs, R. E. (1990). Statistical distribution of hydrophobic residues along the length of protein chains. Implications for protein folding and evolution. *Biophys J*, *57*(4), 911-921. https://doi.org/10.1016/S0006-3495(90)82611-4

White, S. H., & Jacobs, R. E. (1993). The evolution of proteins from random amino acid sequences. I. Evidence from the lengthwise distribution of amino acids in modern protein sequences. *J Mol Evol*, *36*(1), 79-95. https://doi.org/10.1007/BF02407307

Wilson, B. A., Foy, S. G., Neme, R., & Masel, J. (2017). Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis of *De Novo* Gene Birth. *Nat Ecol Evol*, *1*(6), 0146-0146. https://doi.org/10.1038/s41559-017-0146

Wilson, B. A., & Masel, J. (2011). Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol Evol*, *3*, 1245-1252. https://doi.org/10.1093/gbe/evr099

Wissler, L., Gadau, J., Simola, D. F., Helmkampf, M., & Bornberg-Bauer, E. (2013). Mechanisms and dynamics of orphan gene emergence in insect genomes. *Genome Biol Evol*, *5*(2), 439-455. https://doi.org/10.1093/gbe/evt009

Wu, D. D., & Zhang, Y. P. (2013). Evolution and function of *de novo* originated genes. *Mol Phylogenet Evol*, *67*(2), 541-545. https://doi.org/10.1016/j.ympev.2013.02.013

Xiao, W., Liu, H., Li, Y., Li, X., Xu, C., Long, M., & Wang, S. (2009). A rice gene of *de novo* origin negatively regulates pathogen-induced defense response. *PLoS One*, *4*(2), e4603. https://doi.org/10.1371/journal.pone.0004603

Xie, C., Bekpen, C., Kunzel, S., Keshavarz, M., Krebs-Wheaton, R., Skrabar, N., Ullrich, K. K., & Tautz, D. (2019). A *de novo* evolved gene in the house mouse regulates female pregnancy cycles. *Elife*, *8*. https://doi.org/10.7554/eLife.44392

Xie, C., Bekpen, C., Kunzel, S., Keshavarz, M., Krebs-Wheaton, R., Skrabar, N., Ullrich, K. K., Zhang, W., & Tautz, D. (2020). Dedicated transcriptomics combined with power analysis lead to functional understanding of genes with weak phenotypic changes in knockout lines. *PLoS Comput Biol*, *16*(11), e1008354. https://doi.org/10.1371/journal.pcbi.1008354

Yona, A. H., Alm, E. J., & Gore, J. (2018). Random sequences rapidly evolve into *de novo* promoters. *Nat Commun*, *9*(1), 1530. https://doi.org/10.1038/s41467-018-04026-w

Yu, J. F., Cao, Z., Yang, Y., Wang, C. L., Su, Z. D., Zhao, Y. W., Wang, J. H., & Zhou, Y. (2016). Natural protein sequences are more intrinsically disordered than random sequences. *Cell Mol Life Sci*, *73*(15), 2949-2957. https://doi.org/10.1007/s00018-016-2138-9

Zhang, L., Ren, Y., Yang, T., Li, G., Chen, J., Gschwend, A. R., Yu, Y., Hou, G., Zi, J., Zhou, R., Wen, B., Zhang, J., Chougule, K., Wang, M., Copetti, D., Peng, Z., Zhang, C., Zhang, Y., Ouyang, Y., . . . Long, M. (2019). Rapid evolution of protein diversity by *de novo* origination in Oryza. *Nat Ecol Evol*, *3*(4), 679-690. https://doi.org/10.1038/s41559-019-0822-5

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S., & Wang, W. (2008). On the origin of new genes in Drosophila. *Genome Res*, *18*(9), 1446-1455. https://doi.org/10.1101/gr.076588.108

Acknowledgements

This work would have not been possible without the help and support from many people. First of all, I would like to thank my supervisor, Prof. Dr. Diethard Tautz for his support and patience through these years. I could not have hoped for a better mentor throughout this process.

I want to acknowledge in particular the people who helped me with their skills and expertise in different areas:

For guiding me through my first incursion in bioinformatics, and generously sharing his knowledge when I needed it, I thank Dr. Kristian Ullrich.

Barbara Kleinhenz gave me valuable training and help in the cell culture. Ellen McConnell, Cornelia Burghardt, and Dr. Sven Künzel helped me with technical support in the laboratory and sequencing. Derk Wachsmuth and the other members of the IT group at the Max Planck Institute for Evolutionary Biology provided technical support with IT systems and programs.

Dr. Kerstin Mehnert and Angela Donner, coordinators of the IMPRS program, helped me navigate administration and formal issues.

Of course, this work would have not been possible without funding, provided by the German taxpayers through the IMPRS and the Max Planck society.

I had the opportunity to supervise three students these years, who helped me with the maintenance of the cell culture at different stages, and gave me the opportunity to improve as a teacher myself. Thank you, Koray Kazan, Jun Ishigohoka and Silvia Aldrovandi for your hard work.

For fruitful discussions, suggestions and ideas, I would like to thank all members of the Evolutionary Genetics research group, especially Dr. Chen Xie and Dr. Guy Reeves. I am also very grateful for the guidance and advice of the members of my Thesis Advisory Committee: Dr. Julien Dutheil and Prof. Dr. Philip Rosenstiel.

Last, but not least, to my friends Dr. Devika Bhave, Dr. Juan Lugo, and Alejandro Cerón: thank you for your moral support, scientific insights, and for being there to talk when I needed to. And to my mother: Rosa Castro, and my sister, Laura Fajardo: thank you for believing in me, and supporting me with your love and trust. I wish to always be a person who deserves such a wonderful family.

Declaration of Authorship

I, Rossy Johana Fajardo Castro, declare that this thesis and the work presented in it are my own and have been generated by me as the result of my own original research.

I confirm that:

1. Apart from Prof. Dr. Diethard Tautz's guidance, the content and design of this thesis are all my own work.
2. This thesis has not been submitted either partially or wholly as part of a doctoral degree to another examining body.
3. Chapter 1 (Section I) has been submitted for publication in a peer-reviewed journal.
4. This thesis has been prepared in accordance to the Rules of Good Scientific Practice of the German Research Foundation.
5. I have acknowledged all main sources of help;
6. No academic degree has ever been withdrawn from me.

Signed in Plön, Germany on November 30th, 2021.

Rossy Johana Fajardo