UNIVERSITY OF TRENTO - Italy

Department of Cellular, Computational and Integrative Biology - CIBIO

PhD Program in Biomolecular Sciences

33rd Cycle

Translational modulation through CRISPR-Cas-mediated genome editing

Tutor:
Professor Alessandro Quattrone
Department CIBIO - University of Trento

Advisor:
Gianluca Petris
Medical Research Council (MRC) - Cambridge

PhD candidate:
Chiara Ambrosini

Academic year 2019/2020

Declaration of Authorship

I, Chiara Ambrosini, confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Trento, 3rd September 2021

TABLE OF CONTENTS

ABSTRACT

More than 300 human conditions, ranging from cancer predisposition to developmental and neurological mendelian disorders, are caused by haploinsufficiency (HI), a genetic condition by which mutational inactivation of a single allele leads to reduced protein levels and is enough to produce the disease phenotype. Therefore, translational enhancement of the spare allele could exert a therapeutic effect.

Here we propose a novel approach for the potential rescue of haploinsufficiency disease loci based on the insertion of specific single nucleotide changes in the Kozak sequence. Since this sequence controls translation by regulating start codon recognition, we aimed at identifying and introducing specific nucleotide variations to enhance translation and rescue haploinsufficiency. To do so, we used CRISPR-Cas base editors, able to generate single nucleotide changes in genomic DNA without the need of a donor DNA and without creating double-strand breaks.

We performed a high-throughput screening to evaluate the strength of the Kozak sequences of 231 haploinsufficient genes. We compared the translational efficiency of each wild-type sequence to that of several variants using FACS-seq, which combines fluorescence-activated cell sorting and high-throughput DNA sequencing. We thus selected 5 candidate genes (*PPARGC1B, FKBP6, GALR1, NRXN1,* and *NCF1*) and several nucleotide variations able to up-regulate translation. Finally, we used CRISPR-Cas base editors to reproduce the most efficient variants of *NCF1* in a cell model relevant for the associated haploinsufficient disease and verified the increase of protein levels.

This study proposes a novel therapeutic strategy to rescue haploinsufficiency and sheds new insights into the regulatory mechanisms underlying the translational process. On a broader level, the possibility of modulating gene expression by acting exclusively on translation expands the CRISPR-Cas genome editing applications.

INTRODUCTION

1. EUKARYOTIC TRANSLATION

Protein synthesis

Translation in eukaryotes is a complex and crucial process that consists of four steps: initiation, elongation, termination, and ribosome recycling.

Translation initiation is a rate-limiting step that requires several eukaryotic initiation factors (eIFs) and during which the 80S elongation-competent ribosome is assembled[1]. In **Fig. 1** the main points of cap-dependent translation initiation are summarized. First, a 43S preinitiation complex (PIC) is formed in a reaction favored by eIF1, 1A, 3, and 5. The 40S small ribosomal subunit joins the ternary complex (TC), which in turn is composed of eIF2 (a GTPase), GTP, and Met-tRNA$_i$[2] (**Fig. 1A**). The eukaryotic mRNA complexed with eIF4F, formed by the scaffold protein eIF4G, the cap-binding protein eIF4E, a DEAD-box RNA helicase eIF4A, and poly-A binding protein (PABP) (**Fig. 1B**), binds to the 43S PIC. Next, the complex moves along the 5'UTR until it finds the suitable AUG starting codon, thanks to the ATP-dependent unwinding of secondary structures that could hamper translation operated by eIF4A (stimulated by eIF4B) and other helicases such as Dhx29 and Ddx3/Ded1[3] (**Fig. 1C**). One triplet at a time enters the peptidyl (P) decoding site of the small ribosomal subunit searching for complementarity with the anticodon in the Met-tRNA$_i$. In particular, eIF5 helps the scanning ribosome to discriminate the correct AUG, by activating eIF2-driven GTP hydrolysis only if the scanning is paused for a sufficiently long time[4,5]. After AUG recognition, eIF2 catalyzes GTP hydrolysis and dissociates from the complex, the Met-tRNA$_i$ is released into the P site of the small subunit, the scanning process ends and the 48S initiation complex (IC) is formed[6,7]. eIF1 is another factor that ensures the fidelity of the starting codon recognition, for example by inhibiting premature GTP hydrolysis by eIF2[8]. Its dissociation is required to start codon recognition, together with most eIFs. This converts the PIC in a closed conformation (**Fig. 1D**). eIF2-GDP is recycled to eIF2-GTP by the guanine nucleotide exchange factor eIF2B for a new round of initiation. The large 60S subunit joins the complex promoted by eIF5B-GTP and the 80S initiation complex (IC) is ready for the elongation phase[9].

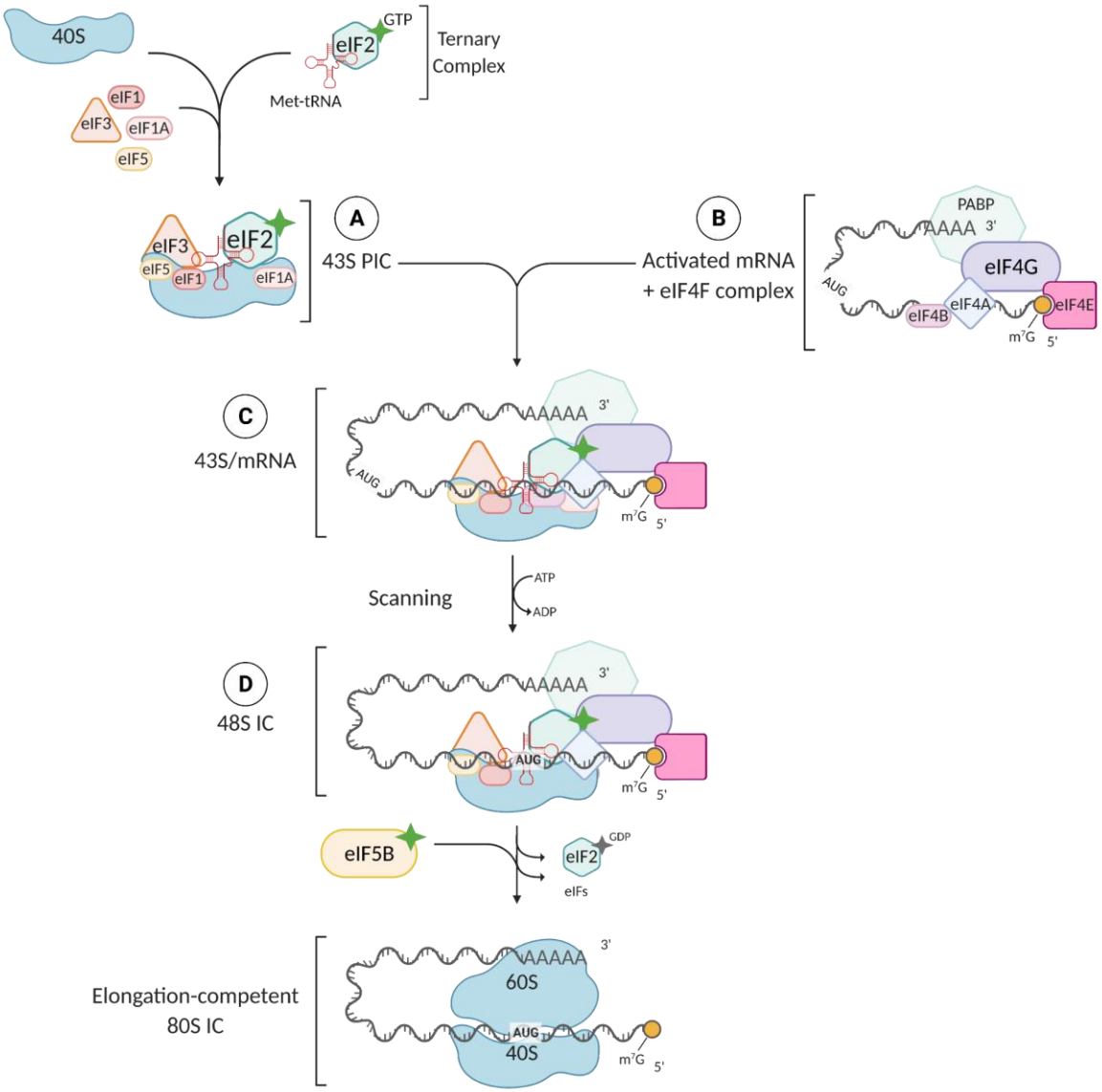


Figure 1 **The main steps of cap-dependent translation initiation in eukaryotes. A.** In the first step, the 43S preinitiation complex (PIC) is formed by the ternary complex (TC), composed of eIF2-GTP and Met-tRNA, and the 40S ribosomal subunit with eIF1, 1A, 3 and 5; **B.** The eIF4F complex, comprising eIF4A, B, E, G, and polyA-binding protein (PABP), binds to the mRNA, that reaches a circularized closed-loop structure; **C.** In the scanning process, the 43S PIC is loaded onto the mRNA and scans the 5'UTR until it finds the start codon AUG; **D.** when the AUG is decoded, most eIFs are released, 48S initiation complex (IC) is formed, and eIF5B-GTP favors the joining of the 60S ribosomal subunit to reach the 80S elongation-competent IC. *Adapted from Ambrosini et al., 2021*[9]

After translation initiation, the Met-tRNA$_i$ is bound to the ribosome's peptidyl (P) site, with the start codon base-paired with its anticodon, and the next codon of the mRNA ORF is in the aminoacyl (A) site of the ribosome. Elongation is the second step of translation, in which the progressive synthesis of the nascent peptide is carried out[10]. It starts with the decoding phase, in which the next aminoacyl-tRNA is selected based on complementarity with the codon present in the A site of the ribosome. The aminoacyl-tRNAs are delivered to the ribosome in a ternary complex with the eukaryotic elongation factor eEF1A bound

to GTP. When the codon in the A site base pairs with its anticodon, eEF1A-GDP is released after GTP hydrolysis, and the aminoacyl-tRNA is accommodated into the A site. eEF1A-GDP is then recycled to eEF1A-GTP by the exchange factor eEF1B. In the following phase, peptide bond formation is catalyzed by the peptidyl transferase center (PTC). Next, the mRNA-tRNA complex must be translocated so that the now deacylated tRNA is in the E site, ready to be released, the peptidyl-tRNA is in the P site, and the next codon of the mRNA is in the A site. This is made possible by the binding of eEF2-GTP, which hydrolyzes GTP facilitating translocation so that the complex is now ready for another round of elongation[7,9,11–13].

The elongation ends when a stop codon (UAG, UGA, or UAA) reaches the A site of the ribosome. Thanks to eukaryotic release factors (eRF), translation termination takes place. In particular, eRF1 (class I factor) catalyzes the recognition of the stop codon and the hydrolysis of the ester bond of the peptidyl-tRNA in the P site, while the GTPase eRF3 (class II factor) stimulates the release of the polypeptide[11,14].

In the last step of translation, ribosome recycling, the deacylated tRNA and the mRNA are released, and the ribosome subunits are dissociated to start another round of initiation[7].

Translational regulation

Being a crucial energy-consuming process, protein synthesis is tightly regulated in the cell to guarantee homeostasis and survival. Translational control plays a predominant role in gene expression regulation. Indeed, in a work by Schwanhausser et al., the authors quantified mRNA and protein abundance and turnover of more than 5000 genes by parallel metabolic pulse labelling, and found that mRNA levels can explain only around 40% of the protein levels fluctuation[15].

The main feature of translational regulation is the ability to rapidly react to potential environmental stress stimuli by re-shaping the proteome[16,17]. Consequently, this layer of gene expression modulation plays a fundamental role in the control of protein expression spatially and temporally, and in processes such as embryonic development, in which transcription is mostly inactive, or in stem cell proliferation[9,18]. The mechanisms evolved to regulate translation can be broadly divided into those acting at the level of global translation, and those at the level of specific mRNAs. Initiation is considered the rate-limiting step of translation since most of the general controls act on this phase[19]. The most common way to rapidly influence the rate of the process is the phosphorylation of the

initiation factors involved. One of the most important examples is the phosphorylation of eIF2 on the serine 51 of its α subunit, which is the central step in the integrated stress response (ISR)[20]. There are four different eIF2α kinases, each activated by different stresses: GCN2 (amino acid starvation), PERK (unfolded proteins in the ER), PKR (double-stranded RNA), and HRI (heme deprivation)[21]. eIF2α phosphorylation causes inhibition of eIF2B exchange factor activity, resulting in decreased GDP to GTP exchange and reduced formation of the 43S preinitiation complex, thus hampering eIF2 ability to bind to Met-tRNA$_i$ in the ternary complex. The result is a decrease of global translation initiation which allows cells to respond to the stress stimuli by saving the energy spent in protein synthesis and rapidly reprogram gene expression[22,23]. At the same time, eIF2α phosphorylation allows the translation of specific subsets of mRNAs that contain one or more upstream open reading frames (uORFs), and can help to ameliorate the cellular stress or, if homeostasis is not restored and the stress stimulus is too strong or prolonged in time, will ultimately lead to apoptosis[16].

Another example of a highly regulated initiation factor is eIF4E, a cap-binding protein that recruits the small ribosomal subunit to the 5' end of the mRNA. The family of eIF4E binding proteins (4EBPs, namely 4E-BP1, 4E-BP2, and 4E-BP3[24]) competes with the scaffold protein eIF4G for eIF4E binding, therefore inhibiting the formation of eIF4F complex and collectively translation initiation. Phosphorylation of 4EBPs by mTORC1 kinase diminishes 4EBPs ability to bind eIF4E, thus promoting translation[16,25].

In the case of mechanisms acting at the level of specific mRNAs, translation regulation often relies on the action of *cis*-elements, sequences present mainly in the 5' and 3' untranslated regions (UTR), that constitute the binding sites for *trans*-acting factor. The concerted action of these elements contributes to set the translational efficiency of the mRNAs, enhancing or hindering the recruitment of ribosomes on the mRNA.

The *cis*-elements present in the 3'UTR are usually responsible for the binding of *trans*-acting factors for the modulation of transcripts localization, degradation, and translation[26]. Such *trans*-factors are, for example, RNA-binding proteins (RBPs) or ribonucleoprotein (RNP) particles containing micro-RNAs (miRNAs)[16].

The regulatory *cis*-elements in the 5'UTR, instead, control translation by regulating mostly translation initiation, and in particular ribosomal scanning on the mRNA or the recognition of the starting codon by the translational machinery[9,26]. Examples of such elements are: the Kozak sequence, impacting translational efficiency by regulating the

recognition of the start codon (*see section 2*); the upstream open reading frames (uORFs), associated with an inhibited translation of the downstream ORF[27,28]; internal ribosomal entry sites (IRESs), complex RNA structures originally discovered in viruses[29] but commonly found in a subset of cellular mRNAs that can promote cap-independent translation[30,31] (**Fig. 2**). This work is focused on the Kozak sequence, its role in promoting translation, and the opportunities given by its modification to influence gene expression.

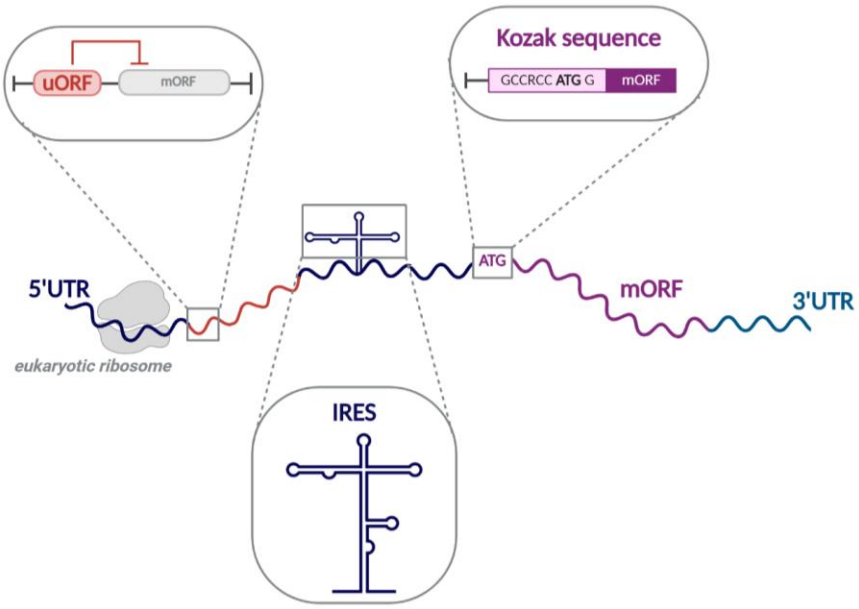


***Figure 2* Main *cis*-elements influencing translation initiation of eukaryotic mRNA.** Upstream open reading frames (uORFs) are inhibitory elements able to hamper translation of the downstream main ORF (mORF); internal ribosome entry sites (IRESs) can promote cap-independent translation; the Kozak sequence is the most favorable nucleotide context around the starting codon of a protein. *Adapted from Ambrosini et al., 2021*[9]

1. KOZAK SEQUENCE

The Kozak consensus sequence was defined for the first time in the 1980s as the optimal nucleotide context around the starting codon of a protein[5,32,33]. The contribution of each nucleotide position was demonstrated experimentally by Marilyn Kozak[34], whose pioneering work contributed to expanding the current knowledge about eukaryotic translation. It is now widely accepted that the Kozak sequence plays a fundamental role in translational regulation. Its impact is exemplified by hereditary and sporadic diseases in which translation efficiency is affected because of point mutations near the AUG starting codon (reviewed in[35] and[36]). For example, a T/C polymorphism in position -1 with respect

to the starting codon within the Kozak sequence of the *CD40* gene is associated with higher *CD40* expression and with Graves' disease[37,38].

Another example of the impact of the Kozak sequence can be found in its role in response to stress stimuli in combination with uORFs. For example, mRNAs preferentially translated in basal conditions present the inhibitory uORF in a poor Kozak sequence and the main ORF in a strong one. Contrarily, genes that must be repressed under stress conditions have a strong nucleotide context around the AUG of the uORF (uAUG)[9,39]. *GADD34*, for instance, presents two uORFs in its 5' UTR, both with a suboptimal Kozak context, which in basal conditions lead to its translation inhibition. In response to stress stimuli, however, both uORFs are bypassed and *GADD34* is preferentially expressed[40]. Therefore, the strength of the Kozak sequence is used to differentially regulate genes expressed in normal conditions versus genes necessary under stress stimuli.

Kozak sequence discovery and first studies

In the 1970s, Marilyn Kozak postulated the "scanning mechanism" of translation initiation, a model by which ribosomes bind to the 5' end of the mRNA and then start to scan it linearly until the recognition of the first AUG codon[6]. To prove this model, Kozak analyzed the sequences of more than 100 eukaryotic messengers, but she noticed that in 10% of the cases the scanning mechanism was not respected, meaning that translation started from a downstream AUG. In this work, the researcher noticed that the region flanking the starting codon was not random, but instead could represent a consensus sequence: [A/G]XXAUGG (the starting methionine is underlined). The first and most important feature of the Kozak consensus sequence is the presence of a purine in position -3 (the A of the starting codon being position +1) and of guanine in position +4.

Based on this evidence, Kozak formulated a reviewed ribosome scanning mechanism, according to which the sequence flanking the first AUG determines if all ribosomes or only a subset initiate translation at that starting codon. In particular, the first AUG in line usually presents the -3 and +4 purines (a favorable Kozak sequence) and therefore is positively recognized for translation initiation; however, if the first AUG does not present a favorable context (i.e., it has a pyrimidine in position -3) the 40S subunit is still able to recognize it, but due to the suboptimal context, it will skip it and continue scanning until it finds the next AUG in line with a more favorable Kozak sequence[33,9].

In 1986 Kozak analyzed the effects of single base substitutions around the AUG of a cloned preproinsulin gene, using site-directed mutagenesis[34]. She identified ACCAUGG as the optimal flanking region, providing experimental evidence of the importance of the hotspot positions identified before. In this influential work, the researcher discussed the presence of upstream AUG initiators that inhibit at some level translation from the main ORF. She pointed out that these sequences (that we now know to be uORFs) are usually in a poor Kozak context, in agreement with the scanning mechanism of initiation. However, she argued the possibility that, when the uORF is in an optimal Kozak context, the 40S subunit ribosome might be able to remain attached to the mRNA, reconstitute the ribosome and reinitiate at a downstream AUG.

With later works, Kozak enlarged the compilation of mRNAs analyzed from different organisms. Collectively, her findings point out an expanded Kozak sequence (GCCGCC[A/G]CCAUGG)[41] (**Fig. 3**), where the positions -3 and +4 remain the ones influencing most the strength of a Kozak consensus: pyrimidines in those positions determine a "weak" AUG flanking region, while purines signal a favorable and "strong" translation initiation site[9].

Novel insights on Kozak sequence

After the work of Marylin Kozak, many advancements have been made towards understanding the mechanism by which Kozak sequence influences translation.

One of the aspects investigated is the specific interaction between the positions flanking the starting AUG and the ribosomal components, and how certain nucleotides in the hotspot positions correlate with strong translational efficiency. For example, Pisarev and colleagues showed by cross-linking studies that the Kozak context serves to stabilize the ribosomal 48S preinitiation complex, avoiding dissociation by eIF1[42]. In particular, the interaction of eIF2α with a purine in position -3 is stronger than with a pyrimidine, therefore G or A in that position causes a stronger resistance to dissociation of the complex. The G in position +4, instead, creates a link with $AA_{1818–1819}$ in helix 44 of 18S ribosomal RNA (rRNA). Interestingly, Kozak had already hypothesized such interaction with 18S rRNA, assuming an analogy with the role of the Shine-Dalgarno (SD) sequence in prokaryotic translation[34]. SD is a purine-rich sequence in bacterial mRNAs, located 5-8 nucleotides upstream of the starting codon, which base pairs with a complementary sequence in the 16S rRNA, helping small subunit recruitment and translational start site selection[5].

More recently, Simonetti and colleagues confirmed these observations with cryoelectron microscopy structures of the late-stage 48S initiation complexes with two mRNAs (β-globin and H4)[43]. In particular, they described a strong interaction between eIF1A and the mRNA in the β-globin 48S, where the conserved eIF1A Trp70 is trapped by the A_{1819} from h44 18S rRNA and the G+4 of β-globin Kozak sequence; on the other hand, H4 presents a pyrimidine in position +4, that results in a much weaker interaction with eIF1A. Interestingly, the eIF1 gene itself is characterized by a weak Kozak sequence which modulates its expression. The poor AUG context establishes an autoregulatory negative feedback loop, maintaining balanced eIF1 levels[8,44]. Regarding the purine in position -3, present in both mRNAs, Simonetti and colleagues showed several contacts with ribosomal proteins and initiation factors, for example with domain 1 (D1) of eIF2α[8,43,44].

Another aspect investigated in more recent years regards the role of the 3'-end portion of the Kozak sequence, which is located in the coding region of the protein. Kozak highlighted the importance of guanine in position +4[5], while she reported that changes in positions +5 and +6 do not impact AUG recognition[45]. Controversial theories have been proposed on the role of these nucleotide positions. For example, it was hypothesized that the G+4 was required for optimal translation as a result of the requirement for certain amino acids at the N-terminus of the nascent protein, and not because of efficient initiation site recognition[46]. According to this interpretation, the G+4 is statistically easier to be found because the most common second amino acids encoded are Alanine (Ala) and Glycine (Gly), both coded by G-starting codons. These two residues are also the two smallest amino acids, a feature required for the common post-translational removal of the N-terminal Methionine[47]. In another work from Niimura and colleagues, the authors analyzed the base biases after the starting codon in the genome of seven eukaryotes[48] and observed that the second codon is the most biased among all the positions, with GCG (encoding Ala) as the most frequent codon in human. In the same work, the author argues that the preference in residue usage cannot explain the bias in position +6, since the third nucleotide hardly impacts the amino acid sequence. In a more recent comparative computational analysis, Volkova and colleagues found that G+4 is important when guanine is the purine in position -3 of the Kozak sequence; on the other hand, the Ann<u>AUG</u> context was found to be enriched in Serine residues. The authors suggest that the translational efficiency may depend on many cooperating factors, such as the 5' end of the Kozak, the nucleotides in the 3' part, and, in some cases, the amino acid encoded by the second codon[49]. Finally, Kozak herself argued against attributing the conservation of G+4 to a requirement for certain amino acids,

because the G+4 positive effect on translation was demonstrated also in assays limited to translational initiation[5,45].

In recent years, another interesting aspect investigated is the non-AUG translation and its dependency on Kozak sequence. Indeed, despite AUG is the largely predominant starting codon, translation can start at near-cognate codons (most commonly CUG, followed by GUG and ACG). Recently, ribosomal profiling aimed at mapping TISs genome-wide revealed that these events have a much higher frequency than what was originally thought[50]. Interestingly, non-AUG translation appears to be induced in response to certain stimuli. For example, BiP (a chaperone that senses misfolded proteins) expression is controlled by two non-AUG uORFs, which are upregulated under stress conditions[51]. Moreover, a systematic study calculating the efficiency of non-AUG translation demonstrated that certain near-cognate codons are more efficient than a fraction of canonical AUGs located in a poor Kozak sequence. In the same study, the authors demonstrated that the efficiency of non-AUG codons is more dependent on the strength of the Kozak sequence with respect to canonical AUG start sites[52].

In the last decade, many groups focused their effort on experimentally measuring the efficiency of libraries of regulatory sequence variants by employing fluorescence-activated cell sorting (FACS) combined with high-throughput sequencing. This approach has been used to screen transcriptional regulatory sequences in *E.coli*[53,54] or yeast[55,56], or 3' UTR sequences impact on translation[57,58]. The latter two studies share the basic principle of using a library of thousands of mutated translation modulation sequences to regulate the expression of a fluorescent protein in a reporter vector. The library is then transformed into the desired host and the regulatory sequences are subdivided according to their activity. Deep sequencing of the subdivided populations gives quantitative information about the influence of each sequence analyzed. This strategy has been applied also to the study of the Kozak sequence. In particular, Dvir and colleagues used a library of two thousand 10bp-long sequences located before the starting codon of a fluorescent reporter in yeast and analyzed the impact of different Kozak sequences, mRNA secondary structures, and uORFs[59].

One of the most important works that applied this principle to investigate eukaryotic translation is the one from Noderer and colleagues, who introduced the term FACS-seq to describe the combination of cell sorting and deep sequencing[60]. In this study, the authors analyzed the translation initiation efficiency of all the possible mammalian Kozak sequences, building a library of 65.536 sequences by randomizing 8 positions in the AUG

flanking region (-6 to -1, +4, and +5). Contrary to the works mentioned above, the reporter system used was a bicistronic vector, in which the library controls the expression of EGFP, and red fluorescent protein (RFP) is translated from the same transcript using an IRES and used for normalization. Viral particles containing the library were used to transduce the cells, which were then isolated into 20 subpopulations (FACS gates or bins) measuring the EGFP/RFP levels. Deep sequencing of the cells in the 20 bins allowed the authors to gain novel insights into the Kozak sequence. The importance of positions —3 and +4 was confirmed, however, the analysis also found that the —2, —4, and +5 positions influence protein expression. Interestingly, Noderer and colleagues also reported significant cooperation between combinations of nucleotides present in hotspot positions: +4 and +5 positions have been reported to be strongly correlated. Moreover, positions —2 and —4, previously reported as not so important in TIS recognition, were found to correlate with the —3 position. In summary, the high-efficiency Kozak sequence identified was RYMRMVAUGGC (where Y = U or C, M = A or C, R = A or G, and V = A, C, or G)[60].

The interest in the role of Kozak sequence in modulating translation is increasing, as demonstrated by the number of studies published in the last years spanning different aspects of this subject[52,61–63].

In conclusion, the Kozak model is still valid nowadays, and the latest works corroborate the notion that the AUG flanking region is a predominantly conserved feature among lower and higher eukaryotes that has a crucial role in the regulation of translation, each nucleotide of which contributing to the overall translational efficiency[9] (**Fig. 3**).

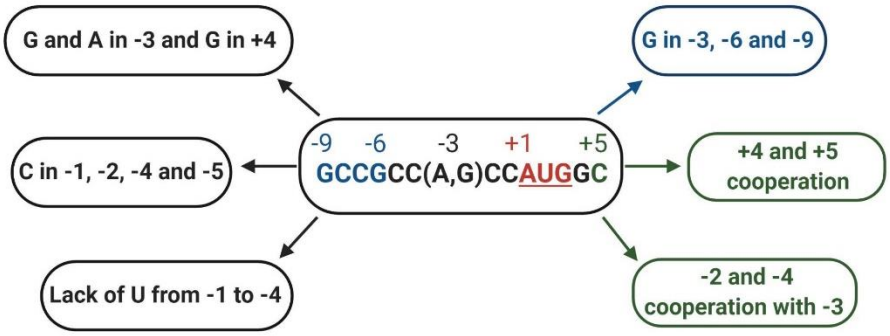


Figure 3 **The Kozak sequence.** The central panel illustrates the favorable AUG flanking region for mRNA translation initiation defined by Kozak in vertebrates. The three black panels report the first characteristics described in lower eukaryotes later confirmed in higher eukaryotes and vertebrates. The blue panel shows the last characteristics found by Kozak. The two dark green panels report the discoveries about Kozak sequences of the last decade. *Adapted from Ambrosini et al., 2021*[9]

Kozak sequence manipulation for the modulation of gene expression

Given its major role in regulating translation, the Kozak sequence has been recognized as a potential target for modulating gene expression.

Manipulation of the Kozak sequence can be aimed at disrupting it, inducing gene silencing of target genes[9]. For example, two groups in 2020 used CRISPR-Cas genome editing tools to selectively mutate the ATG starting codon of targeted genes[64,65]. Wang and colleagues targeted the programmed cell-death protein (PD-1) in mouse embryos, obtaining a PD-1 knock-out mouse[64].

Alternatively, the Kozak sequence can be targeted to obtain an enhancement in translational strength. This principle has been recently applied to improve the yield and product quality of bispecific antibodies (bsAbs), that recognize two different epitopes on one or two different antigens[66]. The authors created a library of Kozak sequences by randomizing 4 positions before and one after the ATG and screened them to find variants able to upregulate the expression of heavy chains (HCs) and light chains (LCs). By regulating the expression of these components, they increased bispecific assembly by more than two-fold over control and observed a higher product purity[9].

These attempts show that the impact of the Kozak sequence on translational regulation has been increasingly recognized and understood.

3. HAPLOINSUFFICIENCY

Haploinsufficiency (HI) is a molecular mechanism in which the mutational inactivation of one allele is sufficient to produce the disease phenotype[67]. It is known that most loss-of-function (LoF) mutations in humans are recessive, meaning that one single allele of those genes (called haplosufficient, **Fig. 4A**) is enough to maintain the normal phenotype. According to the metabolic theory of dominance, this happens because of the redundancy of the genetic information, so that the effects of such mutations are "masked" by the wild-type allele[68,69]. This is the case for most enzyme-coding genes, for which a heterozygous LoF leading to ~50% reduction in its dosage will not impact the metabolic outcome, since that enzyme is placed in a chain of reactions that will likely compensate for the decrease in one of its components[70]. On the other hand, for haploinsufficient genes, the heterozygous LoF mutation causes a dominant phenotype, and a single allele is not enough for the normal

function[71] (**Fig. 4B**). There are some hypotheses about the mechanisms leading to HI[72]. However, the two most common theories are the balance hypothesis and the insufficient amount hypothesis[69]. The balance theory explains the HI of subunits of macromolecular complexes, in which precise stoichiometry between all the components must be maintained[73]. In this case, overexpression of the gene would have the same detrimental effect as its HI, since it would cause an equally impacting imbalance[69]. For example, HI of *IGF1R* (insulin-like growth factor I receptor) causes familial short stature[74]. Its protein product can assemble in homodimers and heterodimers, meaning that the loss of half of its expression results in an amplified decrease in active assembled receptors[75]. The insufficient amount hypothesis states that HI relies directly on the reduced level of the protein expressed in the heterozygous. An example is given by *PAX6*, a transcription factor (TF) whose HI in the eye causes aniridia (absence of the iris). If its levels are above a certain threshold during eye development, *PAX6* will bind to an enhancer triggering a positive feedback loop, which leads to an amplification of its expression. On the other hand, if this threshold is not reached, *PAX6* levels are not sufficient for the normal development of the eye[70,76]. Interestingly, in two recent studies, the authors have tried to rescue *PAX6* HI by suppressing its nonsense mutations, treating a mouse model even after birth. The treatment was shown to rescue different aspects of eye development[77,78], demonstrating the possibility to act on the dosage of target haploinsufficient genes to restore the normal phenotype.

More than 300 loci have been reported in the literature as LoF-intolerant and therefore haploinsufficient. Not surprisingly, these genes are enriched in subunits of macromolecular complexes, such as the ribosome, or involved in transcription regulation, as shown by gene-set enrichment analysis[75]. HI is associated with many different disorders, including mental retardation, neurological or developmental disorders, immunodeficiency, limb malformations, and tumorigenesis[79].

Apart from those annotated in literature, a number of algorithms have been developed to predict haploinsufficiency in given loci, following different principles. Some studies calculate the probability of a gene to be haploinsufficient by integrating genetic, transcriptional, and protein-protein interaction features[80,81]. More recently, Steinberg and colleagues created a new machine learning approach that combines information from ENSEMBL, functional annotations from the Encyclopaedia of DNA Elements Consortium (ENCODE), and the NIH Roadmap Epigenomics Project to predict haploinsufficiency[82].

This approach, called HiPred, gives a score to each human gene to quantify the probability of HI without suffering from study biases for which well-studied genes are over-represented in gene networks[80].

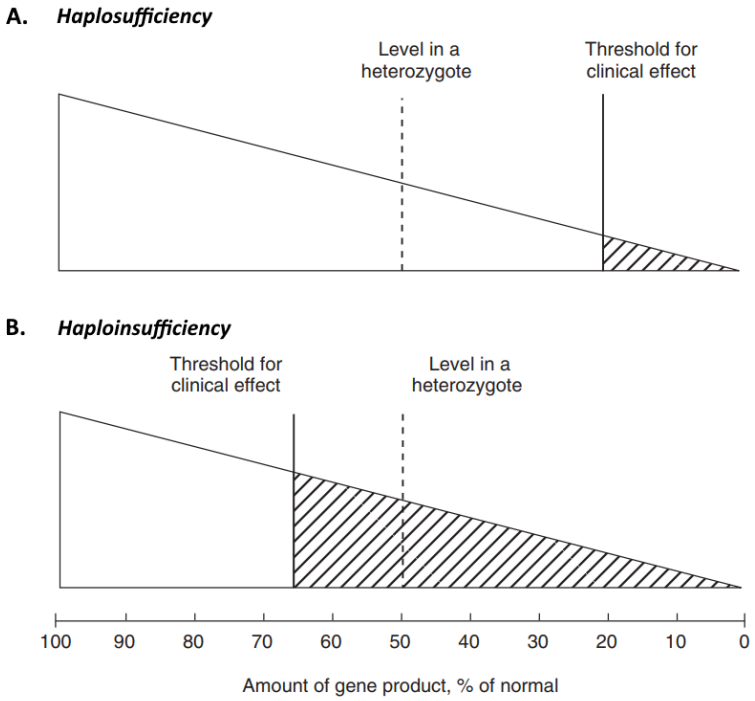


Figure 4 **Representation of the effects of mutations that decrease the protein product of a gene. A.** Haplosufficient gene: the effects are noticeable only when the decrease of the function of both alleles combined drops below 20% of normal. Loss-of-function mutations will be recessive. **B.** Haploinsufficient gene: the effects become noticeable when the level of function is below 65% of normal. The resulting condition will be dominant. *Adapted from Read, 2017*[71]

1. THE CRISPR-CAS SYSTEM

CRISPR-Cas origins and mechanism of action

CRISPR (clustered regularly interspaced short palindromic repeats)-Cas is nowadays the technology of choice in the genome-editing field due to its versatility, specificity, and simplicity in target design. This tool has revolutionized the field, as shown by the remarkable speed at which this technology is evolving. Despite our ability to harness CRISPR-Cas to manipulate genomes started almost a decade ago, its first discovery in bacteria dates back to 1987, when Ishino and colleagues noticed a particular repeated sequence downstream of the *iap* gene in *E.coli* K-12 cells, unknown in the sequence databases available at that time[83]. The same sequences were then observed in many other bacteria and archaea organisms, suggesting that it could exert a general conserved function[84,85]. Comparative studies of the CRISPR regions (as it was defined by Jensen in 2002[86]) in many organisms, allowed to postulate the common features of the arrays: they are short palindromic sequences present in intergenic regions, interspersed by unique non-repeated regions called spacers. Four genes were found to be conserved and adjacent to the CRISPR arrays and called Cas (CRISPR-associated) genes[86,87].

In 2005, three independent studies showed that the spacers are homologous to sequences found in bacteriophages and DNA invading plasmids[88–90]. Barrangou and colleagues then experimentally proved for the first time the biological function of CRISPR as a prokaryotic RNA-mediated, DNA-targeting immune system[91] in *Streptococcus thermophilus*.

CRISPR-Cas immune system mechanism of action includes three steps: (i) adaptation, (ii) expression, and (iii) interference[92]. During adaptation, foreign DNA fragments (protospacers) are cut out after the recognition of a short motif (the protospacer-adjacent motif, PAM)[93] and then inserted into the CRISPR array, becoming spacers. In the expression stage, the CRISPR array is transcribed as a long precursor transcript (pre-CRISPR RNA, or pre-crRNA), then processed in mature CRISPR RNAs (crRNAs) that are assembled to Cas proteins to form CRISPR-RNP (crRNP) complexes[94]. At the interference step, the invading nucleic acid is cleaved by Cas nuclease protein, guided by the crRNA[95].

CRISPR-Cas as a genome-editing tool

CRISPR-Cas systems are classified into six types (I-VI), that can be categorized into two broad classes: class 1 (including type I, III, and IV) is formed by multi-protein effector complexes, in which the crRNA is associated with multiple Cas proteins; class 2 (type II, V, VI), on the other hand, possesses one single effector protein[96]. Most genome editing applications developed so far rely on class 2 CRISPR, and in particular on type II. In this CRISPR-Cas type, Cas9 is the endonuclease responsible for recognizing and cleaving the foreign DNA thanks to a dual RNA guide, formed by the crRNA and a trans-activating crRNA (tracrRNA)[97,98].

Two pioneering works in 2012 made it possible to harness CRISPR-Cas as a genome-editing tool. In the first one, the authors demonstrated that Cas9 can cleave the foreign DNA using its two catalytic domains (RuvC and HNH) *in vitro*[99]; the second one demonstrated that the tracrRNA:crRNA duplex can be engineered to create a single-guide RNA (sgRNA), allowing to reprogram CRISPR-Cas9 to target a sequence of interest[100]. Thanks to these milestone discoveries, for which Jennifer Doudna and Emmanuelle Charpentier were awarded the Nobel prize in Chemistry in 2020, it is now possible to target any sequence, next to an appropriate PAM, in the genome by changing the sequence of the spacer in the sgRNA[98].

The most well-studied CRISPR-Cas type II system is that from *Streptococcus pyogenes* (SpCas9). SpCas9 is able to recognize NGG or NAG PAM sequences, but many Cas proteins with different PAM preferences have been developed, to expand the possible loci to target[101]. Advancements in the field have also allowed overcoming, at least in part, one of the major concerns related to Cas9 and CRISPR-Cas systems in general: the off-target cleavage of genomic DNA, due to the fact that Cas9 can tolerate mismatches between the protospacer and the sgRNA. Among the strategies developed to tackle this problem are the generation of high-fidelity Cas variants[102,103] and the limitation of Cas expression into the cells[104,105].

CRISPR-Cas has been successfully used for a number of applications[106]. Gene knock-out (KO) or the insertion of precise modification in the genome are made possible by exploiting the fate of the genomic DNA after Cas9 has induced a double-strand break (DSB) at the target locus (**Fig. 5**). DSBs catalyze the endogenous cellular DNA repair machinery, leading to different processes. Cells preferentially repair the damage by non-homologous end-joining repair (NHEJ), an error-prone repair that causes gene disruption through the

introduction of random insertion or deletions. For these characteristics, NHEJ offers the possibility to generate targeted knock-out in the cells. On the other hand, homology-directed repair (HDR) can insert a precise edit in the DNA thanks to the addition of a donor DNA to the system[107]. However, this process is highly inefficient in mammalian cells and is largely limited to the late S and G2 phases of the cell cycle[108].

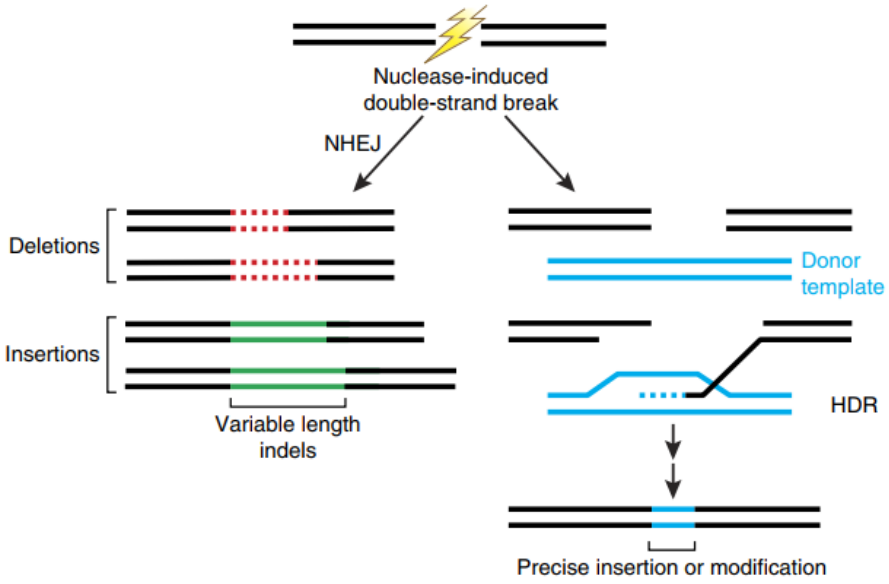


***Figure 5* Repair pathways after nuclease-induced DSB.** DSBs can be repaired by non-homologous end-joining (NHEJ) repair, that induces insertion/deletions of variable length, or by homology-directed repair (HDR), that in presence of a double-strand donor DNA can insert precise point mutations or insertions *Adapted from Sander and Joung, 2014*[109]

Besides precision editing on the genome, the CRISPR-Cas system has been extensively used also for other applications. Just some examples include: transcriptional modulation effects through the fusion of a dead-Cas9 (dCas9, with both catalytic domains inactivated) to transcriptional effector proteins acting either as repressors or activators (CRISPR-i and CRISPR-a, respectively[110,111]); high-throughput screening using libraries of sgRNAs targeting the desired set of genomic sequences to study genomic functions[112]; generation of animal models[106,112].

Two of the latest applications of CRISPR-Cas are base editing (*see section 5*) and prime editing. The latter involves the fusion of a Cas9 nickase (with one of the two domains inactivated) to an engineered reverse transcriptase. This complex is able to write new genetic information into the target locus, including nucleotide conversions, small insertions, and deletions with the help of a prime editing guide RNA (pegRNA), which determines the target sequence and encodes also for the desired edit[113].

5. CRISPR-CAS BASE EDITORS

In 2016, David Liu's lab developed base editors, a genome editing method that allows inserting precise point mutations in the genomic DNA, without creating DSBs and without the need for a donor DNA or to rely on the cellular HDR machinery[114]. Base editors are composed by the fusion of a partially active Cas9 (Cas9 nickase D10A or Cas9n) with a base modification enzyme (deaminase) able to perform single nucleotide substitutions in the DNA[115].

Base editors take advantage of the programmable capacity of CRISPR-Cas to direct the deaminase enzyme to the desired locus in the genome[116]. The Cas9 engagement of the target locus and DNA strand separation position the deaminase, which induces the transition within a window inside the sgRNA target sequence[117]. Each base editor has a different activity window in which it performs efficient point mutations, typically 4 - 5 nucleotides wide. There are two classes of base editors: cytosine base editors (CBE), able to convert C-G into T-A base pair, and adenine base editors (ABE), able to perform the substitution from A-T to G-C base pair[115].

In both base editors, after the binding of Cas9, the hybridization of the sgRNA spacer to the target site generates an ssDNA R-loops in the PAM-containing strand, exposing the target nucleotides to the action of the deaminase (**Fig. 6**). In CBE, the cytidine deaminase domain catalyzes the deamination of cytosines to uridines, which are read by polymerases as thymines[118] (**Fig. 6A**). The deaminated base becomes the target of cellular DNA repair machinery, in particular of uracil DNA glycosylase, able to remove the uracil base and form an abasic site. After that, the repair can be error-free (coming back to the initial situation) or error-prone, leading to base scrambling and indels formation. The addition of Uracil Glycosylase Inhibitor (UGI) to the system, blocks the base excision by Uracil-N-glycosylase (UNG), increasing base editing efficiency and product purity. On the other hand, ABE deaminates the target adenine into inosine, which is read by the polymerase as guanine[119] (**Fig. 6B**). Inosine is also a substrate for cellular repair machinery-induced excision. However, this action is less efficient than uracil excision in mammalian cells[120], as shown by the fact that inhibiting the glycosylase responsible for inosine removal did not substantially increase the performances of ABE[119]. In both base editors, the fusion with a Cas9 nickase (Cas9n D10A) allows the nicking of the non-edited strand. This tricks the cells into repairing the non-edited strand using the edited strand as a template[114,119,121]. As a result, upon DNA replication the desired conversion is inserted in the genomic DNA[116].

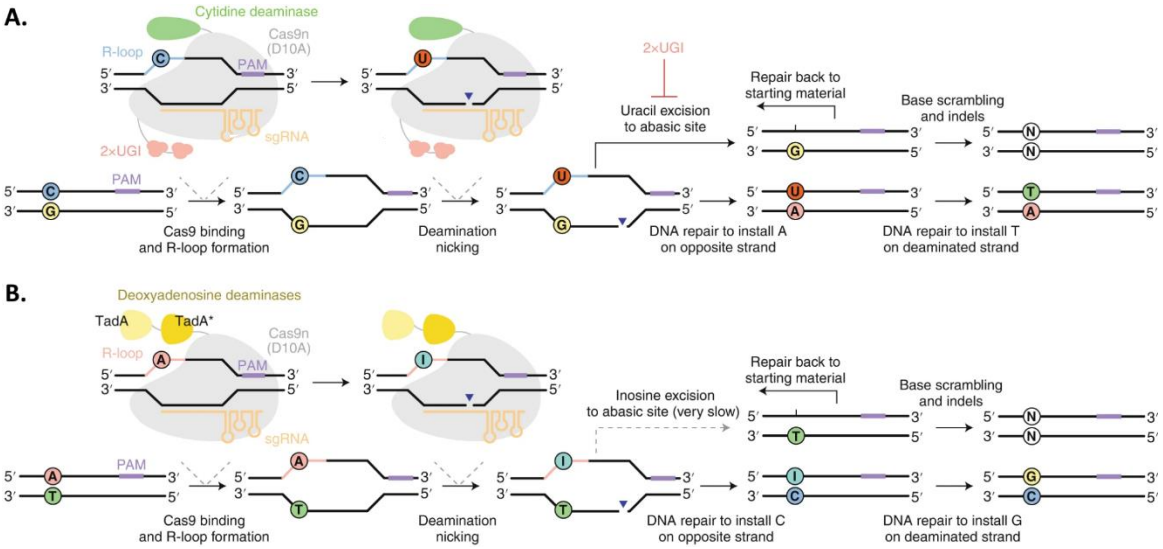


Figure 6 **Mechanism of action of base editors. A.** Cytosine base editors (CBE) catalyze the conversion from C-G to T-A in genomic DNA. After sgRNA hybridization with the target locus and Cas9n binding, the cytidine deaminase converts the C in the R-loop to uracil (U). Uracil-N-glycosylase excises the converted base forming an abasic site that is either corrected (going back to the starting material) or leads to base scrambling and indels. Uracil Glycosylase Inhibitor (UGI) hampers this activity enhancing CBE efficiency. Cas9n nicking of the non-edited strand favors the DNA repair machinery to install A in the opposite strand. Upon DNA replication the corrected conversion is installed. **B.** Adenine base editors (ABE) converts A-T into G-C. The deoxyadenosine deaminases domains convert adenine into inosine (excised to form an abasic site but at much lower levels). Cas9n nicking of the non-edited strand favors installing C on the opposite strand, leading to corrected G-C base pairs upon DNA replication. *Adapted from Anzalone et al., 2020*[122]

Base editor beginnings

The first CBE developed with all the components described above (APOBEC1–Cas9 nickase–UGI fusion) was BE3[114]. APOBEC1 is a deaminase enzyme able to deaminate ssDNA but not dsDNA[123]. In the same years, Nishida and coworkers developed an alternative CBE, Target-AID, which uses a different deaminase (CDA1) and presents some differences in the window of action with respect to BE3[121].

Since this first effort, Liu and colleagues have extensively improved base editing efficiency and product purity using several strategies. First, different linkers increased efficiency in the fourth generation of CBE (BE4). Secondly, the possibility of indels, given by the fact that an abasic site is in proximity to the nick induced by Cas9n, was reduced by fusing the bacteriophage Mu-derived Gam protein (Mu-GAM) to BE4 to generate BE4-Gam[118]. Finally, in the attempt to tackle the low expression of base editors that can reduce their activity, Liu's lab improved CBEs by optimizing codon usage and nuclear localization sequences (NLS) to create BE4max[124]. In the same work, the authors also developed

AncBE4max, a base editor with improved protein expression in mammalian cells obtained by ancestral sequence reconstruction from hundreds of APOBEC homologs[124].

Soon after the development of CBE, Liu and coworkers realized the necessity of a base editor able to reproduce the other type of transition mutation (A-T to G-C), which is the most common type of pathogenic point mutation in the ClinVar database (47% of the cases)[115]. However, contrary to cytidine deaminases, there was no known adenosine deaminase enzyme capable of accepting ssDNA as a template. Therefore, Liu's lab evolved a deaminase enzyme starting from the *E.coli* tRNA adenosine deaminase enzyme, TadA. Through directed evolution and protein engineering, the researchers selected TadA*, which, fused with Cas9n, lead to correct A-T to G-C conversion, although with low efficiency[119]. After the observation that TadA natively operated as a homodimer, in which one of the monomers is endogenously provided by WT TadA, Liu and colleagues generated a construct with two adenosine deaminase domains, one WT non-catalytic monomer, and one TadA* evolved monomer. After different rounds of evolution, ABE7.10 was generated with TadA* domain presenting 14 amino acid substitutions with respect to the WT TadA[119]. After this first ABE version, ABEmax was obtained in 2018 by improving NLS (nuclear localization signal) and codon usage, together with BE4max and AncBE4max[124]. Two novel ABE versions were obtained in 2020: ABE8e was created by evolving ABEmax, improving the deamination kinetics of 590-fold[125]; ABE8 was obtained by Gaudelli and colleagues by evolving ABE7.10 to have a wider window of action (positions 3-10) and higher activity[126]. The 8th generation of ABE also supports the use of a smaller architecture that uses a single evolved TadA* monomer. This modification allows reducing the size of the editor by about 500 bp, facilitating the generation of vectors carrying this system[127].

Advancements in base editing

Consistent work has been made in the last few years to improve base editors, under multiple points of view.

Undesired byproducts. One of the main concerns about genome editing are off-target effects (OT). In the case of base editing, the undesired effects are complicated by the presence of the deaminase domain. The undesired byproducts induced by base editors can be divided into those generated at the target site in the genome and those at off-target genomic sites[122]. At the desired locus in the genome, base editors can induce (i) undesired

transversions, (ii) bystander effects, or (iii) indels[122]. (i) Transversion mutations are observed mostly as a result of the error-prone DNA repair machinery and impact the base editing product purity. To tackle this obstacle, Liu and coworkers fused UGI to the system and noticed that using 2 UGI modules or changing the linkers architecture helped in further reducing such effects (**Fig. 6A**). UGI has also been fused in *trans* with BE3, leading to an additional improvement in product purity[128]. The ability of CBE to induce different types of conversion has been harnessed in base editing applications that differ from installing single point mutations. Bassik and Chang groups, indeed, generated base editing-mediated targeted mutagenesis in mammalian cells (targeted activation-induced deaminase (AID)-mediated mutagenesis[129] and CRISPR-X[129,130]).

(ii) Bystander effects are defined as conversions happening in the protospacer sequence but at positions different from the targeted one[115]. This effect arises from the fact that other C or A may be present in the R-loop created at the target site, and therefore be the substrate for the deaminase enzyme. Many bystander edits are likely to cause minor effects, especially if targeting non-coding regions or if the conversion leads to a silent mutation (by using CBE or ABE, it has been calculated that this is the case in 53% of the cases[115]). However, to solve the problem in which bystander editing could have detrimental effects, base editors with narrower windows of action have been generated[131].

(iii) Indels can be generated at the target site as a consequence of the abasic site created by the uracil or inosine excision next to the nick induced by Cas9n. For CBE, as stated before, this problem has been tackled by the addition of the bacteriophage Mu-GAM protein. ABE, instead, induces significantly fewer indels (usually below 1%[115]), due to the fact that inosine removal is much less efficient in mammalian cells[120].

Base editors byproducts at off-target sites include deamination occurring on the DNA and can be distinguished in two types: Cas-dependent or Cas-independent OT effects. The first type is a consequence of Cas binding to non-targeted genomic loci. Strategies to overcome this problem include fusion of base editors to high-fidelity Cas variants[132–134], truncation of the sgRNA[133], and reduction of cells exposure to base editors by RNPs delivery[132]. Cas-independent OT effects are due to the presence of the deaminase domain, which can randomly edit nucleotides in the DNA at non-targeted loci. This effect, which usually happens at a low level but is widespread in the genome, was detected with CBE but not with ABE-mediated base editing[135,136]. High-fidelity deaminase variants have been developed to minimize this effect[137,138].

The last undesired effect induced by base editing is represented by off-target editing on the RNA, mediated by the action of deaminase enzymes that natively operate on ssRNA[139,140]. In this case, too, many base editor variants have been rationally evolved to minimize random RNA deamination, such as the V106W mutant[141].

Targeting limitations. To efficiently perform base editing, two critical factors must be taken into account: the presence of a PAM sequence and the base editor window of action. PAM sequence depends on the Cas domain of the base editor. To maximize the possible sequences to target in the genome, several different base editors have been developed by fusing the deaminase to engineered Cas variants[142], or Cas different than SpCas9, such as Cas12[125,143] or *S.aureus* Cas9[126,131]. Moreover, recently Walton and colleagues developed a near PAMless SpCas9 variant, named SpRY (recognizing NRN PAM sequence), and demonstrated its efficiency when used as base editor domain[144].
The editing window is defined as the positions in the protospacer sequence that support maximal base editing. This depends on the interaction between the R-loop and the deaminase itself[122]. The window of action can change as a result of different Cas or different deaminases, which modify the architecture of the complex[131,145].

Newly-developed base editors. The huge interest in base editing is demonstrated by the number of novel editors developed in the last few years. In 2020, different groups described the generation of dual base editors, containing both cytidine and adenine deaminase domains, therefore able to install concurrently C-to-T and A-to-G conversions[146–149]. Moreover, CRISPR-Cas base editors have been successfully used to perform genome-wide functional screenings[150–153]. Finally, the first base editors able to generate C-to-G transversion have been developed[154,155].

Collectively, the rapid expansion of the field leads to a constantly growing list of base editors to choose from, with different PAM preferences and windows of action, so that the probability to target a precise locus in the genome is continuously increasing.

Gene therapy with base editors

The many advancements in the efficiency and specificity of base editing, together with the delivery methods, open novel possibilities for therapeutic approaches with base editors. For each disease, the success of the therapy requires finding the perfect combination between

base editor, sgRNA, and delivery method, which in turn will be dictated by the type of tissue targeted by the therapy. The delivery methods that have been exploited for base editing applications have been extensively reviewed by Porto and colleagues[156]. Briefly, we can divide the main methods into direct delivery of base editor and sgRNA, viral vectors, and ribonucleoprotein (RNPs).

Direct delivery of base editor and sgRNA by means of cationic lipids, electroporation, or direct injection. This approach has the advantage of being cost and time-effective, moreover, there are nearly no size restrictions or risks of integration in the genome. The transient expression also ensures a reduced chance of off-target events due to prolonged base editor presence in the cells[117]. A possible drawback of this method is that delivering DNA or mRNA could be immunogenic. However, there are examples of successful delivery of base editors via direct injection of naked plasmid DNA into the tail vein of a mouse model of tyrosinemia[157]. More recently, Musunuru and colleagues delivered the mRNA of a base editor together with the sgRNA as lipid nanoparticles in cynomolgus monkeys, to generate a loss-of-function mutation in *PCSK9*, whose upregulation is responsible for familial hypercholesterolemia. The authors observed efficient base editing and consequent reduction of PCSK9 blood levels, with no deleterious immune response to the treatment, apart from a rise in AST and ALT (aspartate and alanine aminotransferases) resolved within 1-2 weeks[158]. Electroporation *ex vivo* of base editors and sgRNAs has also been used as a delivery method. One of the most recent successes in delivering base editors with this system was described by Liu's group. The researchers delivered a custom base editor and the sgRNA through electroporation into hematopoietic stem and progenitor cells (HSPCs) *ex vivo* from patients with Sickle Cell Disease and showed improved hematological parameters in mice models of the disease transplanted with the edited cells[159].

Viral vectors commonly used for genome editing are adenovirus (AdV), adeno-associated virus (AAV), lentivirus, and retrovirus[156]. Of these, the most promising appears to be the use of AAV, since they are minimally immunogenic or toxic and they allow transient gene expression[160,161]. However, the size of the base editor is not compatible with the low packaging capacity of AAVs[106]. To overcome this challenge, split-intein base editors have been developed, by dividing the editor into N-terminal and C-terminal fragments fused to respective halves of the fast splicing split-inteins. Few examples of diseases targeted by

delivering split base editors include amyotrophic lateral sclerosis[162], phenylketonuria[163], and Niemann–Pick disease[164].

Ribonucleoprotein (RNP) delivery ensures DNA-free base editing since the base editor purified protein and the guide RNA molecule are encapsulated in a macromolecular complex. RNP delivery is characterized by a faster editing activity, and since it remains in the cells for a short timeframe, it ensures lower OT effects. As an example, Liu and colleagues delivered BE3-sgRNA RNP to the inner ear of postnatal mice[165].

AIM OF THE THESIS

Haploinsufficiency occurs when the mutational inactivation of one allele results in reduced expression of the functional protein to a level that is not sufficient to sustain its physiological role. Translational enhancement of the remaining allele could exert a therapeutic effect. Such enhancement should derive from a fine-tuning of translational efficiency and the ideal increase in protein production should not exceed double of initial levels, to avoid detrimental effects derived from the overexpression.

Despite the relevance of HI in human disease, no study has been conducted to systematically target the translational layer of regulation of haploinsufficient genes to rescue the lack of correct protein products.

The aim of this thesis is to propose a novel approach to potentially rescue HI disease loci by specifically inducing single nucleotide changes in the Kozak sequence, reported to control translational efficiency. We performed a high throughput screening of the Kozak sequences of 231 haploinsufficient genes and respective variants to identify the specific nucleotide variations able to enhance the translation of each gene. To reproduce these changes, we used CRISPR-Cas base editors that allow the induction of specific single nucleotide conversions in genomic DNA. This approach aims to create an increase in the translation efficiency of the selected genes, sufficient to rescue the loss of protein but avoiding its overexpression.

Moreover, our approach can be applied in principle to all genes that would benefit from an increase in translation efficiency, being the Kozak sequence a conserved feature of all human genes.

RESULTS

Base editing-mediated Kozak optimization enhances translation in a reporter system

In order to demonstrate the feasibility of translational regulation through CRISPR-Cas base editing, we first sought to perform a proof of principle of translational enhancement in a reporter system. In particular, we decided to up-regulate EGFP expression from a bicistronic reporter vector (pWPT-/GCCACC-mEGFP-IRES-mCherry[166], from now on referred to as pWPT-EGFP-IRES-mCherry). Since the EGFP Kozak sequence is already optimized to maximize its expression in target cells, we first created a suboptimal version of EGFP. Specifically, we inserted a C/T mutation in position -1 with respect to the ATG starting codon (EGFP-C-1T): a nucleotide variation found to down-regulate translational efficiency and to be particularly relevant in clinical cases of certain diseases, as previously mentioned[37] (**Fig. 7A**). After transiently transfecting HEK293T cells, we found that this single nucleotide change significantly reduced EGFP translation efficiency, as observed by western blot (**Fig. 7B**) and FACS analysis (**Fig. 7C, D, E**).

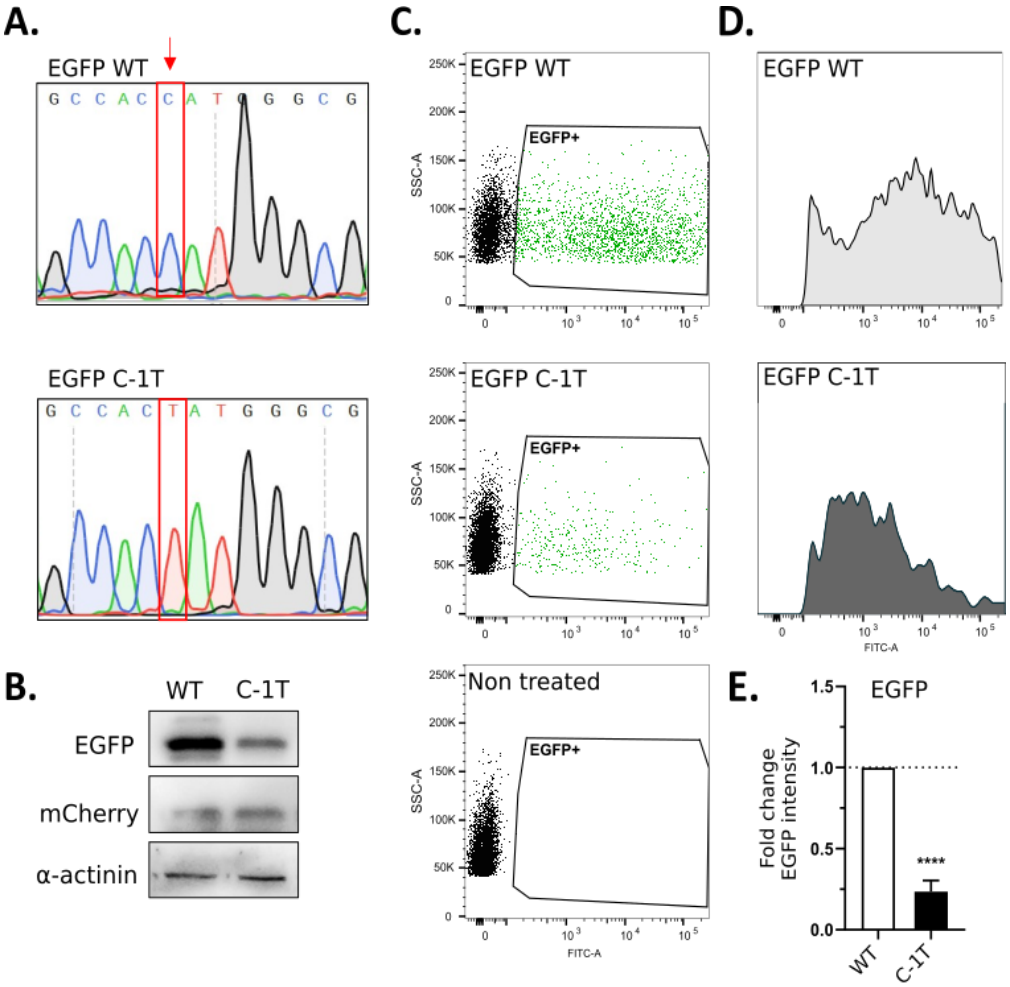


***Figure 7* A single nucleotide change in the EGFP Kozak sequence significantly impairs fluorescence.** **A.** Sanger sequencing chromatograms representing the wild type (WT) and the mutated EGFP version (C-1T), with a single variation in position -1 of the Kozak sequence. **B.** Western blot analysis of EGFP and mCherry expression in HEK293T cells transiently transfected with EGFP WT or EGFP C-1T plasmids. **C.** Representative FACS dot plots of cells 3 days after transient transfection. The EGFP+ gate was defined on the basis of the non treated cells (untransfected, *lower panel*). **D.** Representative histograms of cells in the EGFP+ gate drawn in C. **E.** FACS analysis of cells transiently transfected with the respective plasmids. The data are normalized over EGFP WT and are reported as mean ± SD of n=3 biological replicates. Statistically significant differences were calculated by unpaired t-test.

The inserted C/T mismatch can be corrected by adenine base editors, which can perform the substitution from A-T to G-C base pair. For this reason, we designed a sgRNA to target the previously inserted mutation and corrected it using ABE7.10 and ABEmax base editors. 3 days after the co-transfection of EGFP-C-1T, base editor, and sgRNA in HEK293T cells, the locus of interest was PCR amplified and Sanger sequenced. The base editing efficiency was analyzed with EditR software (*see methods*). The analysis revealed a modest percentage of T-to-C substitution (〜13% mean) (**Fig. 8A, B**).

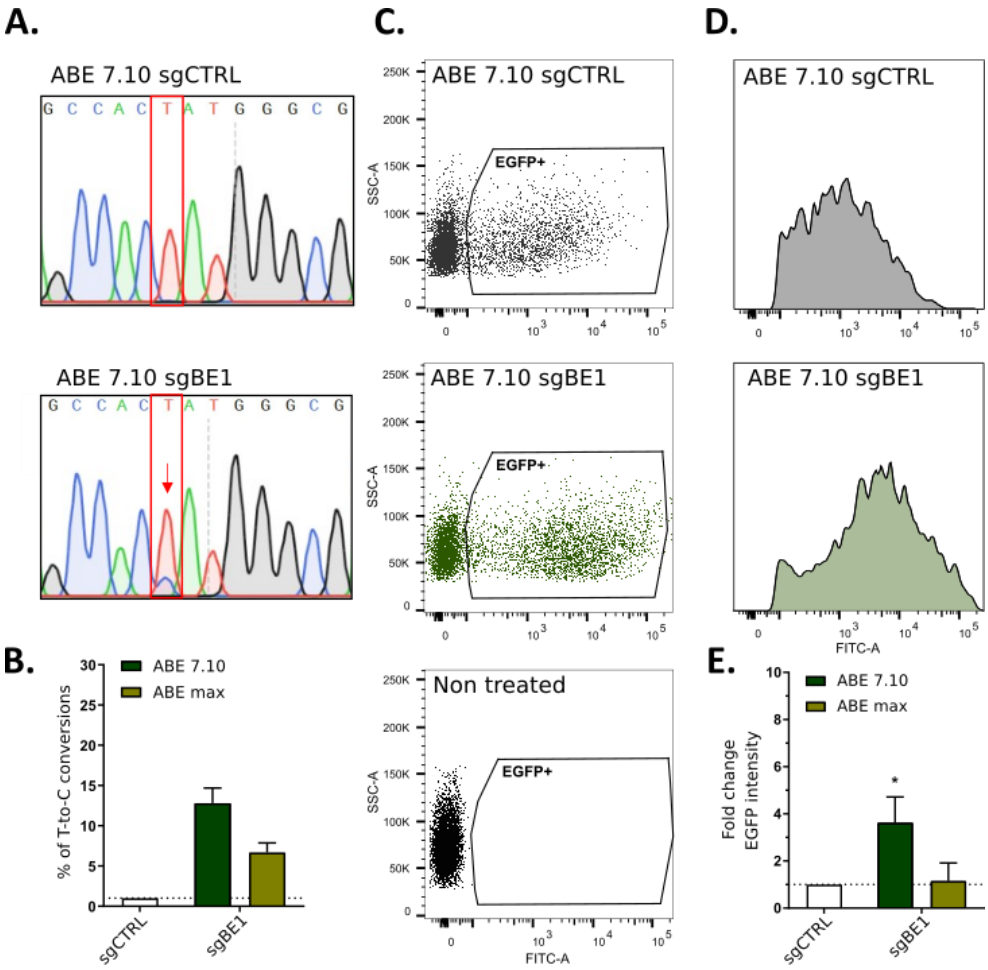


Figure 8 **Restoration of the correct Kozak sequence by base editing causes EGFP translational up-regulation. A.** Representative Sanger sequencing chromatograms of HEK293T cells edited with ABE7.10 base editor and sgBE1, in comparison with ABE7.10 in combination with a scrambled sgRNA (sgCTRL). **B.** The graph represents the percentage of correct T-to-C conversion achieved by transient transfection of base editors (ABE7.10 or ABEmax) and guide RNAs (sgBE1 or sgCTRL), analyzed with EditR software. The data are reported as means ± SD of n=3 biological replicates. **C.** Representative FACS dot plots of cells treated as in B., 3 days after transfection. The EGFP+ gate was defined on the basis of the non treated cells (untransfected, *lower panel*). **D.** Representative histograms of HEK293T cells in the EGFP+ gate drawn in C. **E.** FACS analysis of EGFP expression in cells transfected with the base editors (ABE7.10 and ABEmax) and sgCTRL or sgBE1. The data are reported as mean ± SD of n=4 biological replicates. Statistically significant differences were calculated by unpaired t-test.

Remarkably, this low efficiency was sufficient to obtain an increase in EGFP translation, as demonstrated by FACS analysis 3 days after transfection. In particular, ABE 7.10 was able to significantly increase the fluorescence intensity in the edited sample with respect to the control, of ∼3-fold (**Fig. 8C, D, E**).

This data confirms that the Kozak sequence has an impact on translational efficiency and even a single substitution can reduce EGFP expression up to 4-fold. Moreover, it demonstrates the possibility of modulating translation by editing the Kozak sequence using CRISPR-Cas base editors.

Design and generation of the Kozak variants library

The possibility of modulating translation efficiency by acting exclusively on the Kozak sequence holds great potential in haploinsufficient diseases, in which the protein expression of a given gene is reduced due to mutational inactivation of a single allele. For this reason, we aimed at performing a high throughput screening of Kozak variants of haploinsufficient genes to find modifications able to up-regulate gene expression. In particular, we screened wild-type (WT) Kozak sequences of annotated HI genes and compared them with respective variants. Indeed, we aimed to identify the specific set of variations able to up-regulate each WT Kozak sequence. We created a non-biased library of Kozak mutants, without considering the canonical sequence described in the literature to be the most performing (CC(A/G)AUGGCG). Moreover, we decided to expand the number of positions investigated on the Kozak to explore the possibility that the positions after the starting methionine in the coding sequence could have an impact on the Kozak strength.

There are more than 300 annotated haploinsufficient loci in humans[79,167], and many more have been identified by using algorithms able to predict haploinsufficiency[80–82]. To build our library, we decided to take into account the HI genes present in the most recent literature

annotation[167], together with some genes described as having a high HI prediction (HIPred score) according to the most recent work[82]. We discarded the genes associated exclusively with cancer or tumorigenesis, as our approach of translational enhancement is best suited for monogenic disorders. A complete list of the 231 HI genes and the associated diseases investigated in this work can be found in **Table S1**. Starting from this list, we built a library of Kozak sequence variants based on the following principles (**Fig. 9**):

- We defined the Kozak width from the nucleotide -4 to the nucleotide +7 (the A of the starting codon ATG being nt +1, e.g. NNNN_ATG_NNNN);
- We created variants bearing conversions reproducible with the originally developed base editors (transitions);
- Each variant bears one kind of transition at a time, meaning that no variant bears two types of transition mutations at the same time.

This led to a number of 5539 variants present in the library (4838 unique).

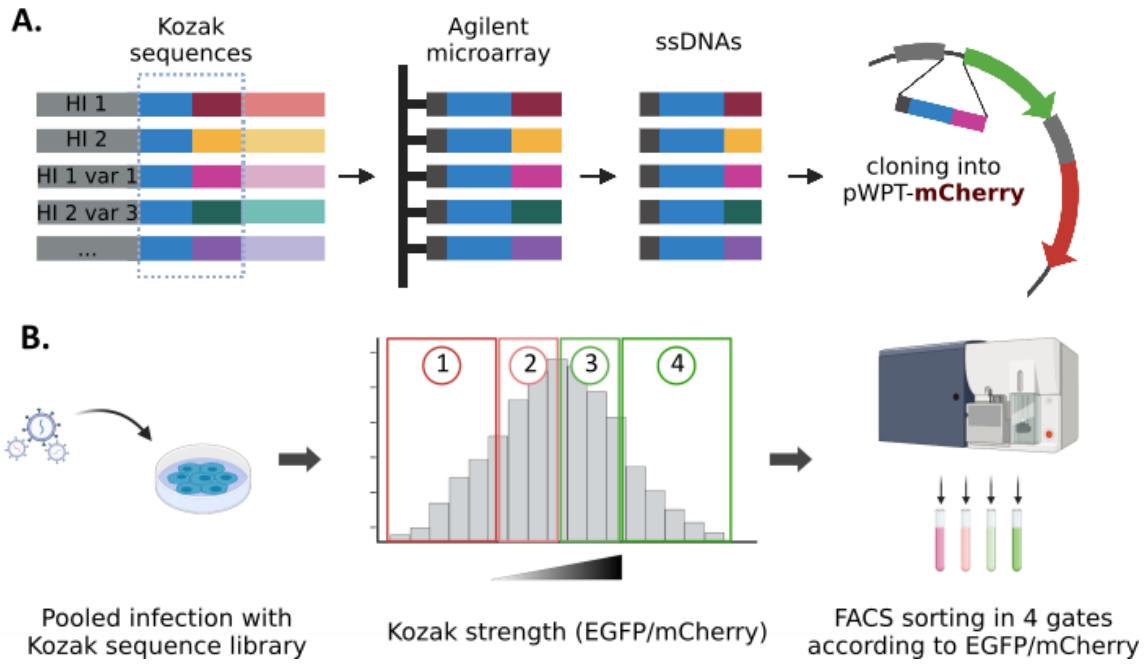


Figure 9 **Schematic representation of the library generation and high-throughput screening. A.** The Kozak variants were designed as oligonucleotides bearing the overhangs to be cloned in the destination vector. The oligos were synthesized on a custom microarray. The library was cloned in place of the EGFP Kozak sequence in a bicistronic reporter vector. **B.** The Kozak sequence library was used to transduce HEK293T cells. Transduced cells were sorted according to their EGFP/mCherry ratio as a measure of Kozak strength. The 4 gates were drawn so that each gate contained 25% of the total population. A*dapted from Oikonomou et al., 2014*[57]

The Kozak sequence variants were synthesized as oligonucleotides on a custom Agilent 244K microarray designed for this purpose. As destination vector, we chose the lentiviral bicistronic reporter previously used for the proof-of-concept[166]. After ligation, the library would replace the EGFP Kozak sequence, directing EGFP expression (**Fig. 9A**). In the reporter vector, mCherry expression is regulated by an IRES, therefore it is translated by the same transcript, and used as a reference. The obtained reporter bearing the library was used to transduce HEK293T cells, which were then cell sorted in 4 gates according to the translational efficiency (EGFP/mCherry) (**Fig. 9B**). To avoid EGFP background signal caused by random reconstitution of the digested vector during ligation or by inefficient digestion of the destination vector, we created an alternative plasmid by replacing the EGFP Kozak sequence of pWPT-EGFP-IRES-mCherry with 5 stop codons, creating a pWPT-mCherry (**Fig. 10A**). The oligonucleotides library was cloned inside the pWPT-mCherry, and after ligation, we checked the expression of the two fluorescent proteins by FACS analysis (**Fig. 10A, B, C**).

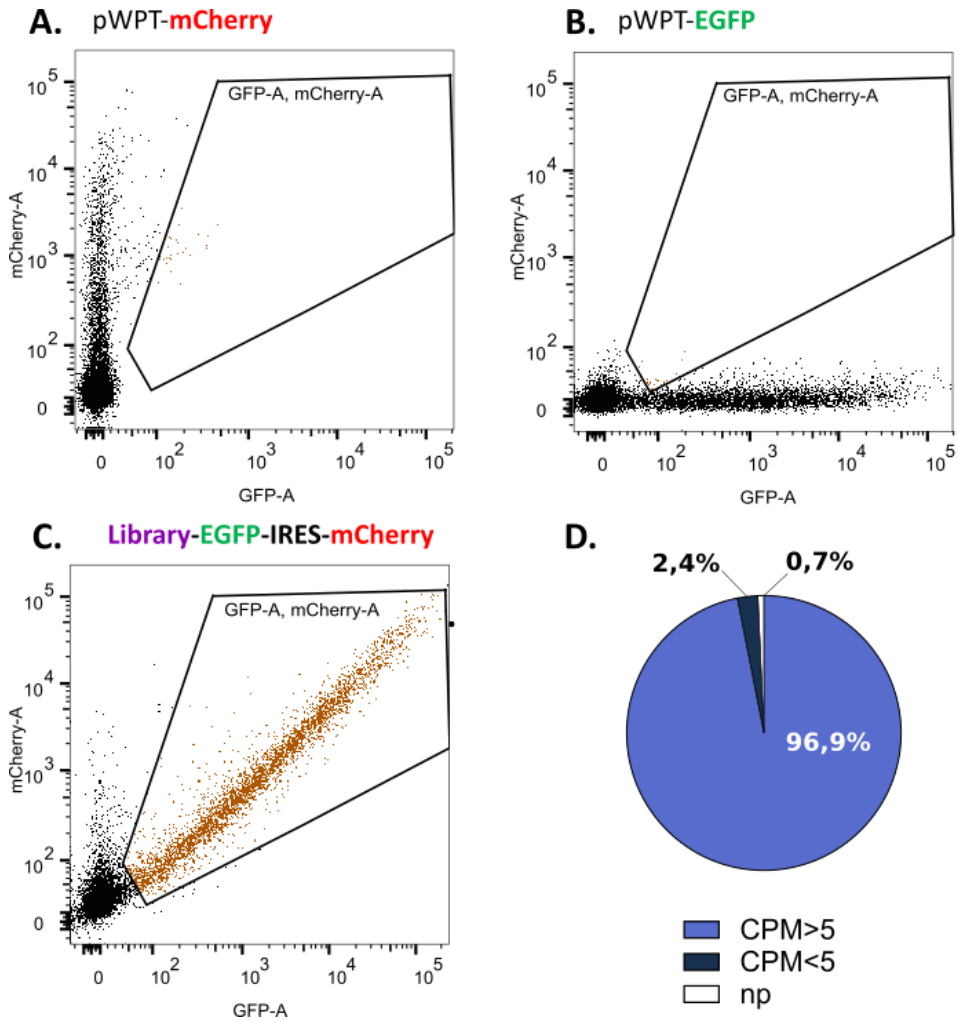


Figure 10 **Lentiviral vectors used in the screening. A., B.**, **C.**, FACS analysis of HEK293T cells transduced with the reporter vectors. **A.** pWPT-mCherry was used as destination vector for the library cloning and was created by inserting 5 stop codons in place of the EGFP Kozak sequence; **B.** pWPT-EGFP was used as control of EGFP expression and was generated by creating a 367nt deletion in the coding sequence of mCherry; **C.** Library-EGFP-IRES-mCherry: the library of Kozak sequence variants was cloned in place of the stop codons of pWPT-mCherry. **D.** Pie charts representing the percentage of sequences identified in the library by deep sequencing. CPM>5: sequences present in the library and well-represented; CPM<5: sequences present but insufficiently represented; np: sequences not present in the library.

The resulting vector bearing the Kozak library (**Fig. 10C**) was deep-sequenced to check for a good representation of all the variants. The analysis revealed that 96,9% of the Kozak sequences were present and well-represented (CPM≥5, see methods), 2,4% were present but insufficiently represented, while only 0,7% were completely absent from the library (**Fig. 10D**).

High-throughput Kozak strength evaluation

To obtain a quantitative measurement of the translational efficiency, we performed a variation of the previously described FACS-seq analysis[60]. The high-throughput screening was carried out into two rounds of cell sorting:

- In the first round, $5x10^6$ mCherry positive cells were sorted to ensure 1000X library coverage (**Fig. 11**);
- In the second round, the resulting mCherry positive cells were sorted according to their EGFP/mCherry ratio in 4 bins of different fluorescence intensity ratios (**Fig. 12)**.

The reporter bearing the library was packaged inside lentiviral particles and then used to transduce HEK293T mammalian cells. The transduction at low MOI (multiplicity of infection) ensures that each cell receives one copy of the vector. 3 days post-transduction, we analyzed the expression of the fluorescent proteins (**Fig. 11**). FACS analysis assessed 23,1% of mCherry positive cells (the reporter internal control) (**Fig. 11C**). mCherry-positive cells were sorted and seeded to achieve expansion and full recovery.

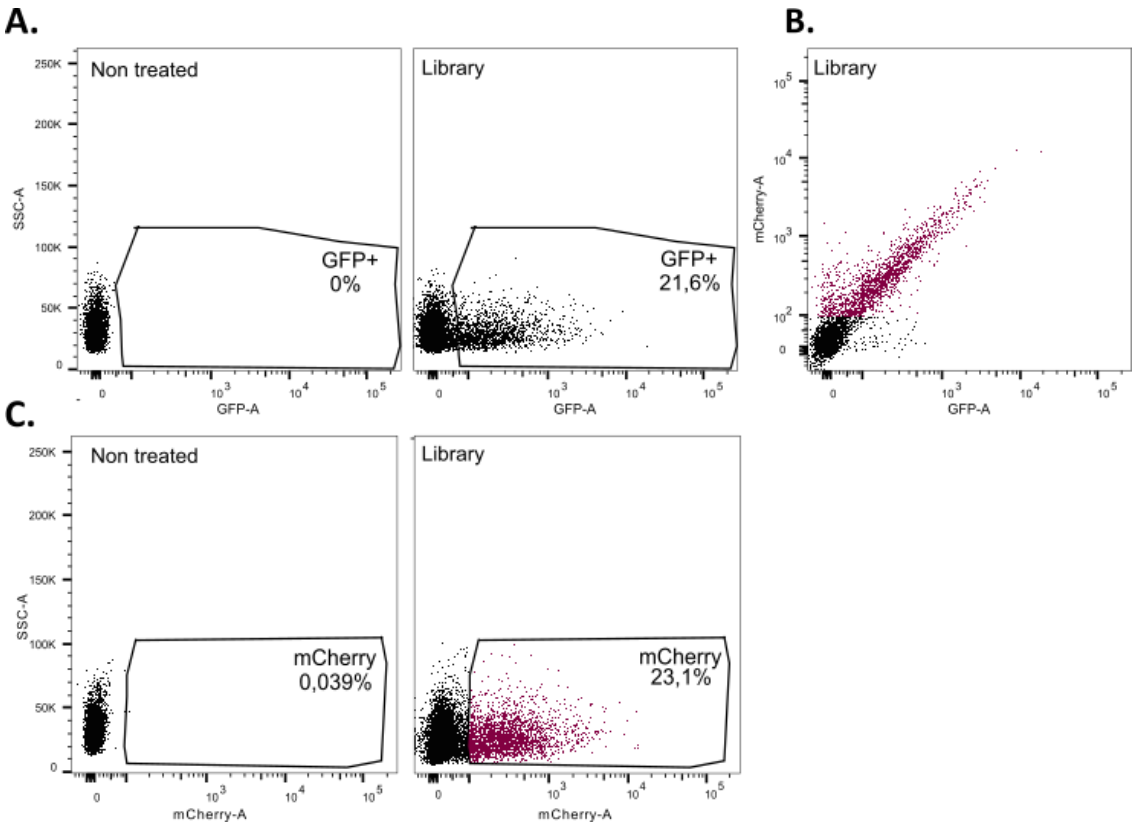


Figure 11 **FACS-seq first round of sorting. A., B., C.** FACS analysis of HEK293T cells 3 days post-transduction of the Kozak sequence library. **A.** EGFP expression of the transduced cells. EGFP+ gate was defined on the basis of non treated cells (untransfected, *left panel*). **B.** Dot plot of the transduced cells expressing mCherry and EGFP. **C.** mCherry expression of the transduced cells. $5x10^6$ mCherry-positive cells (23,1% of the total) were sorted. mCherry gate was defined on the basis of non treated cells (untransfected, *left panel*).

48 hours later (5 days post-transduction), EGFP and mCherry expression were assessed again (**Fig. 12A, B, C**). 92,6% of the sorted cells were mCherry positive, confirming the validity of the first sorting step and allowing us to proceed with the second round (**Fig. 12C**). To measure the strength of the Kozak sequence variants, we divided the population of interest (mCherry-positive cells) into 4 gates according to the ratio between EGFP and mCherry expression (EGFP/mCherry). The gates were created so that each bin contained 25% of the total population of interest (**Fig. 12D**). $1,25x10^6$ cells were sorted for each gate to maintain 1000X library coverage. After the sorting, a sample from each sorted bin was run again to check for product purity. The analysis revealed that maximal purity was obtained in the top left and top right gates (numbers 1 and 4), while some contamination was present in the central gates (~40%), which are narrower and more difficult to discriminate for the cell sorter. However, no contamination of the middle gates in the highest and lowest Kozak strength gates was observed, ensuring consistent results (**Fig. 12E**).

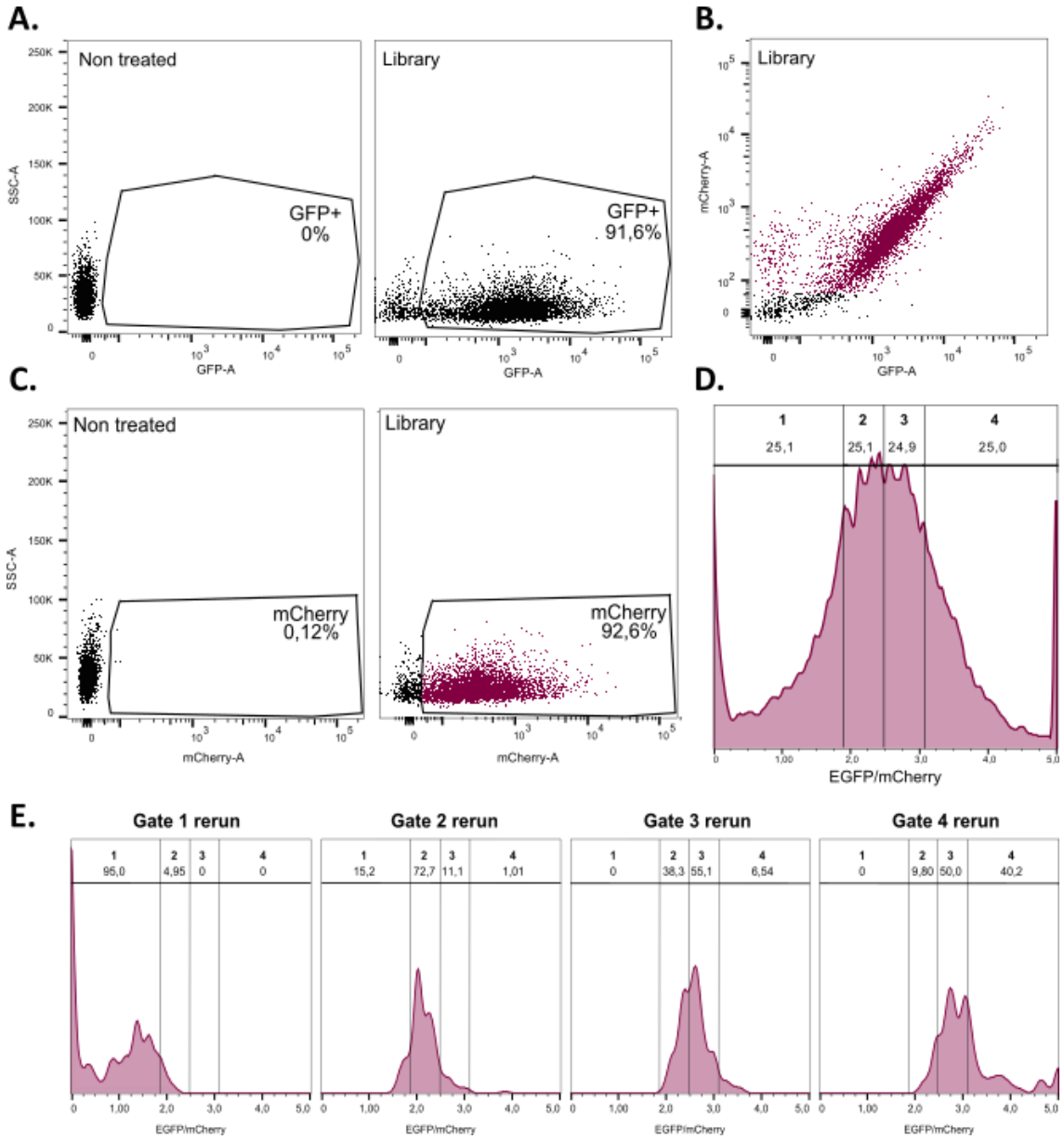


Figure 12 **FACS-seq second round of sorting. A., B., C.** FACS analysis of HEK293T cells 5 days post-transduction of the Kozak sequence library. **A.** EGFP expression of the transduced cells. EGFP+ gate was defined on the basis of non treated cells (untransfected, *left panel*). **B.** Dot plot of the cells transduced with the library expressing mCherry and EGFP. **C.** mCherry expression of the transduced cells. mCherry gate was defined on the basis of non treated cells (untransfected, *left panel*). **D.** mCherry-positive cells from the gate drawn in C. were divided into 4 gates according to EGFP/mCherry expression. $1{,}25 \times 10^6$ cells were sorted for each gate. **E.** After sorting, a purity check was performed by rerunning a small sample from the isolated fractions. Each graph represents one of the 4 fractions analyzed.

After sorting, the Kozak sequence region from the cells collected in each of the bins was PCR-amplified. Deep sequencing analyses of all fractions allowed us to compare the strength of each HI WT Kozak to its variants. We obtained 89 WT and 409 variants, which represent the sequences that passed the statistical analysis (**Fig. 13A**). The heatmap shows that each sequence (row) is significatively present in one gate (column) and decreases progressively in the adjacent ones, demonstrating the validity of the method. As shown in the heatmap, many WT sequences (lower panel) are present in higher gates, meaning that those Kozak efficiently promote translation. This result was expected, since we screened WT Kozak from human genes, and it is known that most mammalian mRNAs are characterized by a Kozak sequence close to the optimal one[168–170].

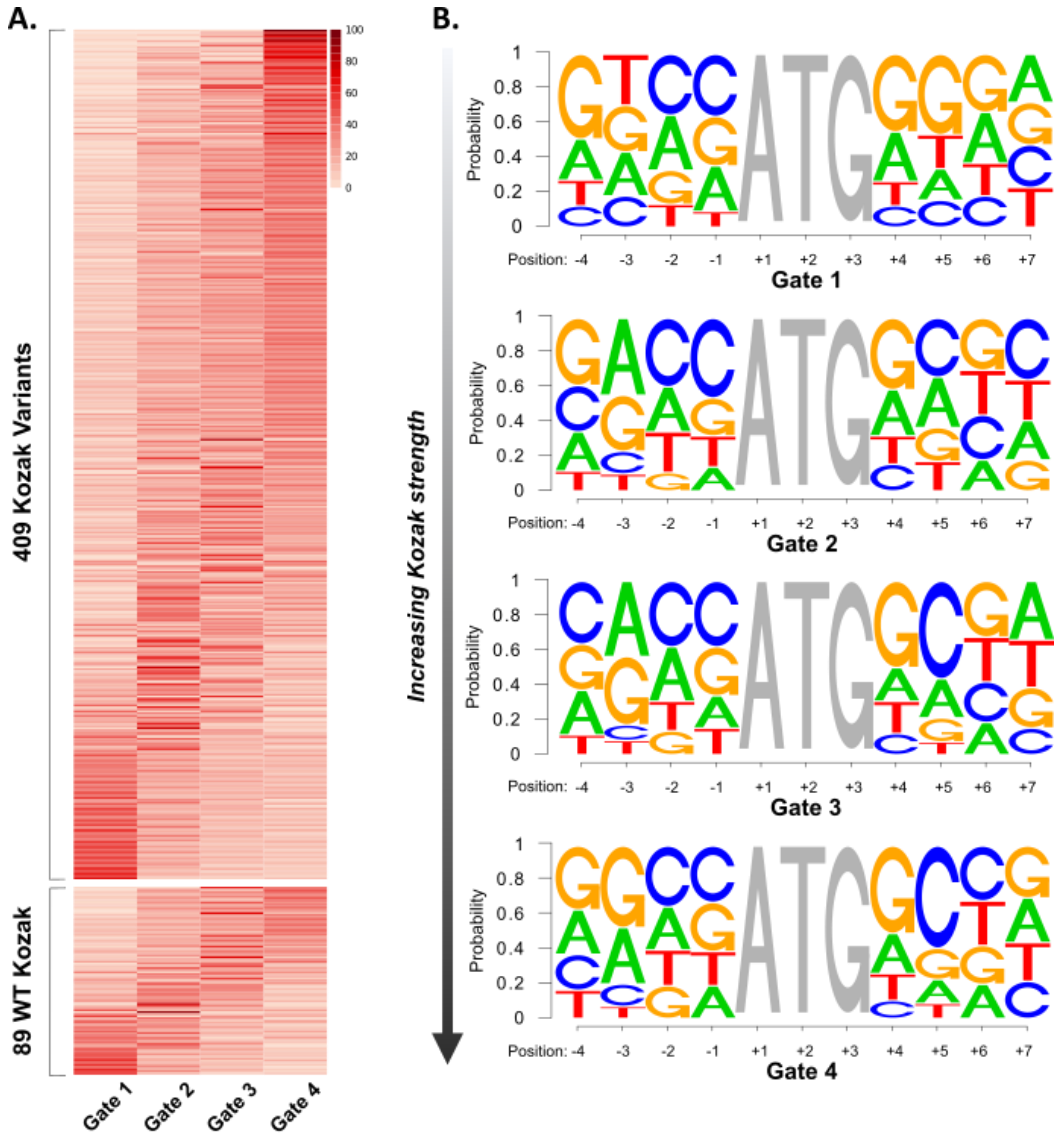


Figure 13 **FACS-seq analysis. A.** The heatmap represents the distribution of the candidate HI genes and variants which passed the statistical analysis. In the upper panel, the Kozak variants are represented. In the lower panel, the WT Kozak sequences of the HI genes are shown. Each column corresponds to one of the four gates, while each row stands for one of the Kozak sequences. **B.** Logo representation of the Kozak sequences extracted from each of the four gates. In each panel, on the x-axis are represented the positions along the Kozak sequence (with A of ATG being position +1), and on the y-axis, the probability of occurrence of each nucleotide is shown. Gate 1 (upper panel) represents the lowest translational efficiency, while gate 4 (lower panel) corresponds to the most performing Kozak sequences.

We then generated a logo for each of the four gates representing the nucleotide frequency at each position of the Kozak sequence (**Fig. 13B**, *see methods*). We analyzed the motifs and compared them with the optimal Kozak described in the literature for mammals CC(A/G)AUGGCG, where a purine in position -3 is considered to be the most important for strong translational efficiency, while a pyrimidine is associated with evident leaky scanning[171]. We noticed that in Gate 1 (**Fig. 13B**, upper panel), thymine was overall the most represented nucleotide in position -3. This is in agreement with the literature since Gate 1 includes the least performing Kozak sequences. Moving onwards with the gates, adenine or guanine become the predominant nucleotides in that position, in agreement with increased Kozak strength. Moreover, a G-stretch can be observed in the consensus of Gate 1 after the ATG. This stretch gradually disappears with the increasing translational efficiency, until cytosine becomes predominant in position +4, a feature of high-performing Kozak sequences in mammals, as confirmed by the literature[48].

Intending to find variants able to up-regulate the corresponding WT, we calculated an Expected Value (EV) for each Kozak sequence (*see methods*). Subsequently, we subtracted the EV of each WT sequence from the EV of the corresponding variants, and we selected only the variants with maximal distance from the respective WT (i.e., EV(VAR)-EV(WT)>0,50). After this step, we obtained 47 WT and 149 variants which represent the possible hits to be validated (**Table S2**). We selected from this list the 5 HI genes and their corresponding variants (Var) with the best overall scores for the screening validation. These candidate genes (*PPARGC1B*, *FKBP6*, *GALR1*, *NRXN1*, and *NCF1*) are reported in **Figure 14** with the distribution of the respective WT and variant Kozak sequences across the 4 gates.

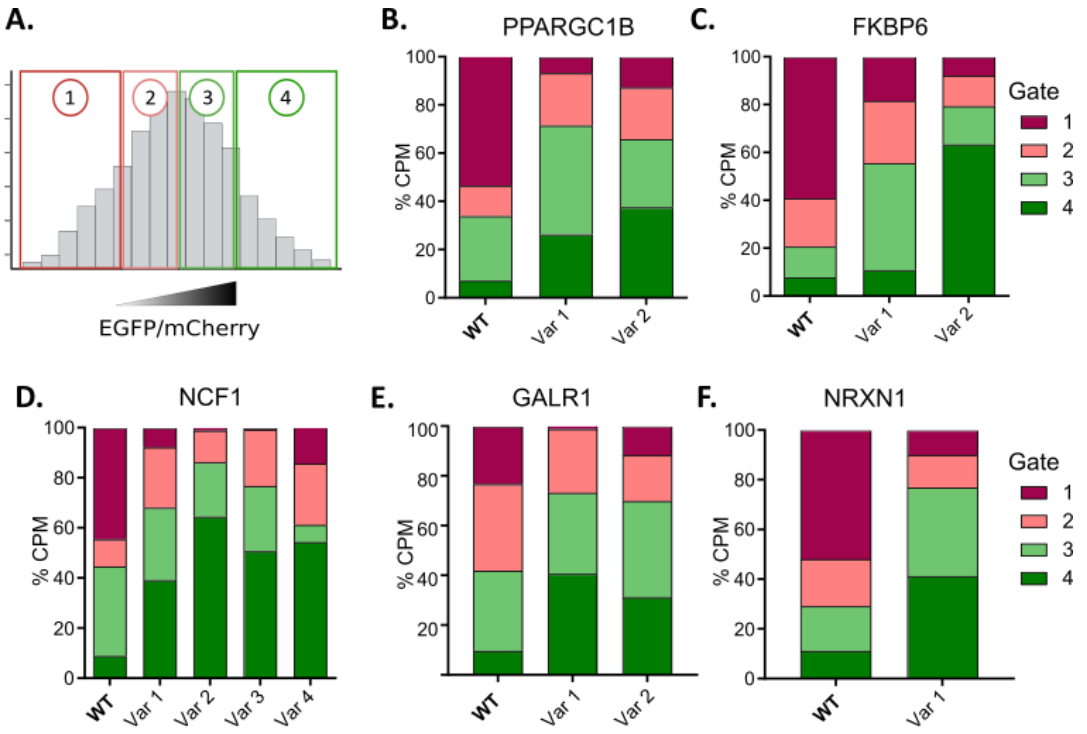


Figure 14 **FACS-seq hits selection. A.** Schematic representation of the library population divided into 4 gates according to EGFP/mCherry expression; **B., C., D., E., F.** Percentage of the count per million reads (CPM) in the 4 gates of the wild type (WT) and the respective variants (Var) of the 5 selected genes.

Additionally, we repeated the high-throughput screening performing a biological replicate with a second oligonucleotide library (Library B) synthesized and cloned independently from the first one, to corroborate our findings (**Fig. 15**). We checked for sequence representation (**Fig. 15A**) and processed the sequencing data as described above. The 5 hit genes selected from the first screening passed again the statistical analysis. Moreover, all the variants identified in the first replicate were confirmed to up-regulate translation of the corresponding WT, albeit with different levels (**Fig. 15B-F**).

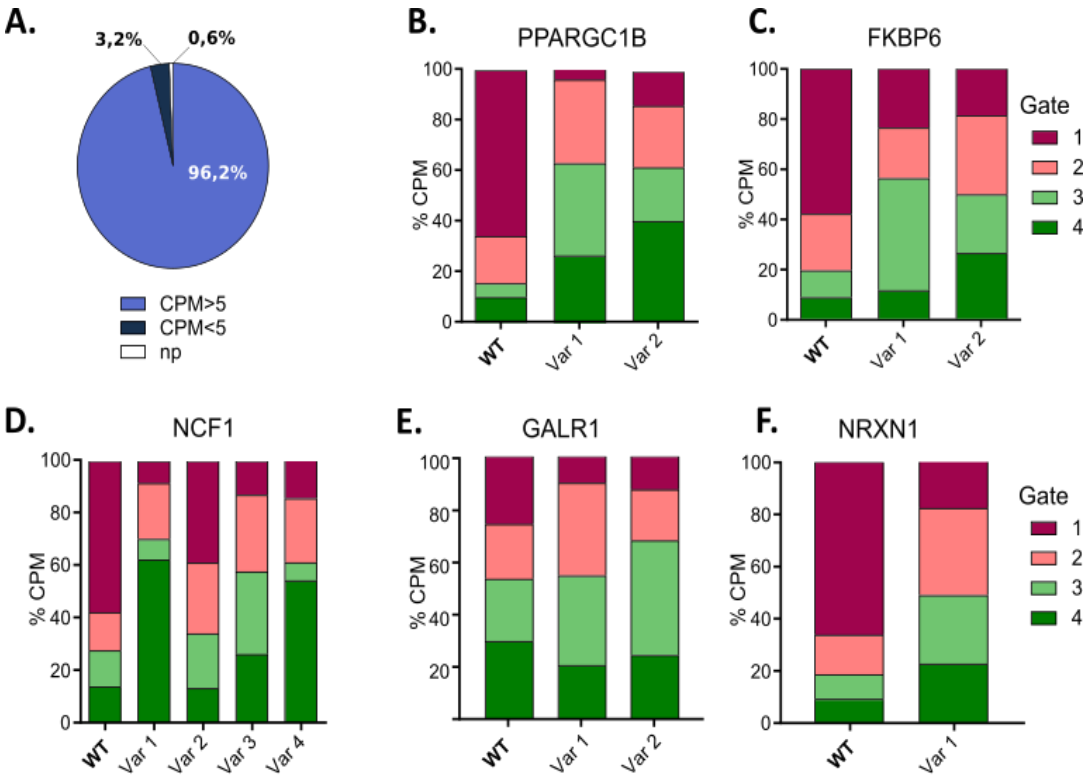


Figure 15 **Library B analysis and hits distribution. A.** Pie charts representing the percentage of sequences identified in Library B by deep sequencing. CPM>5: sequences present in the library and well-represented; CPM<5: sequences present but insufficiently represented; np: sequences not present in the library. **B., C., D., E., F.** Percentage of the count per million reads (CPM) in the 4 gates of the wild type (WT) and the respective variants (Var) of the 5 selected genes in Library B.

Validation of the translational up-regulation of the selected hits in the reporter system

For the validation of the selected hits, we cloned each of the Kozak sequences (WT and Var of the five selected genes) in place of the EGFP Kozak sequence in the pWPT-EGFP-IRES-mCherry reporter vector, creating one plasmid for each Kozak. For each gene, we transiently transfected HEK293T cells with the respective WT and Var, and we analyzed the fluorescence and the protein expression by high content image analysis and by western blot (EGFP under the control of the different Kozak sequences, and mCherry as a transfection control) three days after the transfection (**Fig. 16-20**).

The first gene considered was the Peroxisome proliferator-activated receptor γ coactivator 1β (*PPARGC1B*), which encodes for PGC-1β and belongs to a family of transcriptional coactivators[172,173]. It plays a key role in the regulation of energy metabolism, for example inducing the expression of genes implicated in oxidative phosphorylation in a variety of tissues, including in muscle cells[174]. *PPARGC1B* is one of the top-scoring genes according

to the HiPred score of haploinsufficiency prediction[82] and is reported to be downregulated in type 2 diabetes and obesity[175–177]. We validated two Kozak Var from the screening, and we found that both of them are capable of enhancing *PPARGC1B* Kozak strength (**Fig. 16**). In particular, high content image analysis of the transfected cells revealed a mean translational enhancement of 36,4% (Var 1) and 28,4% (Var 2) with respect to the WT (**Fig. 16B, C**). These results were corroborated by western blot analysis (**Fig. 16D, E**). In the western blot, we also noticed a shift in the EGFP band of the variants with respect to the WT (**Fig. 16D**).

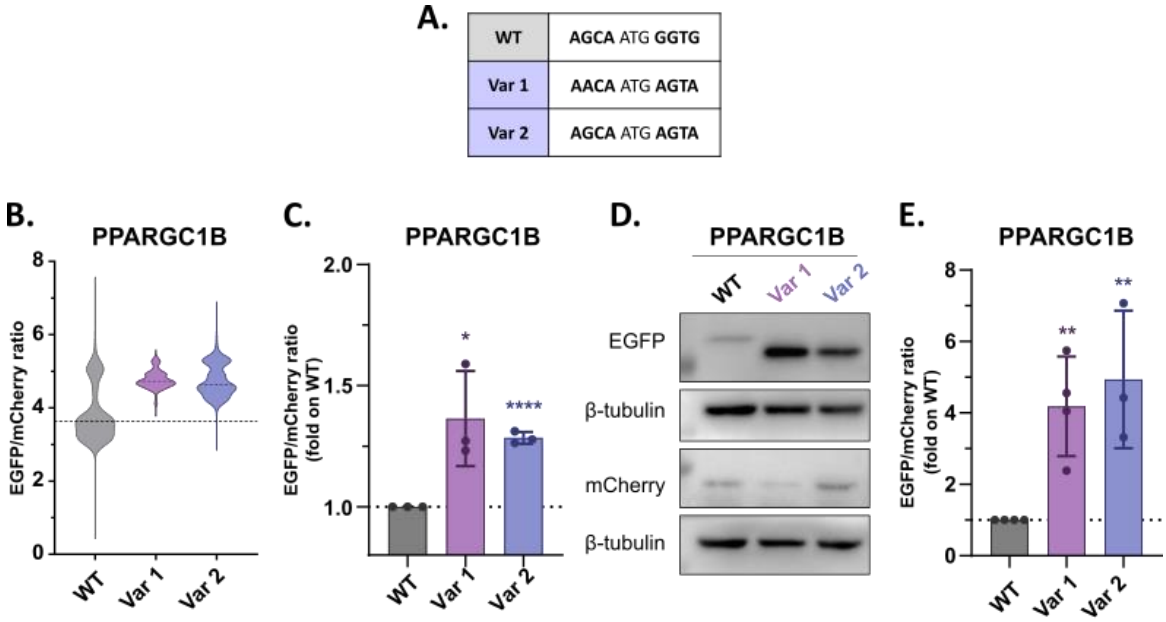


***Figure 16 PPARGC1B* validation**. HEK293T cells were transiently transfected with EGFP-IRES-mCherry bearing *PPARGC1B* WT, Var 1, and Var 2 Kozak sequence and analyzed 3 days post-transfection. **A.** *PPARGC1B* WT and variants Kozak sequences. **B., C.** *Translational enhancement analyzed as EGFP/mCherry expression by high content image analysis.* **B.** The violin plots report the distribution of the data from n=3 biological replicates. The dashed line indicates the population median; **C.** the histogram represents the median of the populations analyzed by high content image analysis from the 3 biological replicates. **D., E.** *Increased EGFP/mCherry expression analyzed by immunoblotting.* **D.** One of the blots is shown; **E.** Western blot quantification. Translational strength was calculated as EGFP/mCherry ratio, each normalized over the respective housekeeping. Data are means ± SD from n=3/4 biological replicates. Statistically significant differences were calculated by unpaired t-test of each variant versus the corresponding WT.

FKBP6 is a prolyl isomerase that plays a role in homologous chromosome pairing in meiosis. Its haploinsufficiency causes Williams-Beuren syndrome (WBS)[178], a developmental disorder characterized by dysmorphic facial features, growth retardation, and congenital heart disease. *FKBP6* is located in a chromosomal region (7q11.23) subjected to hemizygous chromosomal microdeletion in WBS[179]. Of the two variants

validated, only Var 2 was efficient in up-regulating translation, with a mean increase of 13,1% over the WT (**Fig. 17B, C**). Immunoblotting analysis confirmed the data revealing a significant difference only with Var 2 Kozak sequence (**Fig. 17D, E**). Again, the shift in the EGFP band was observed (**Fig. 17D**).

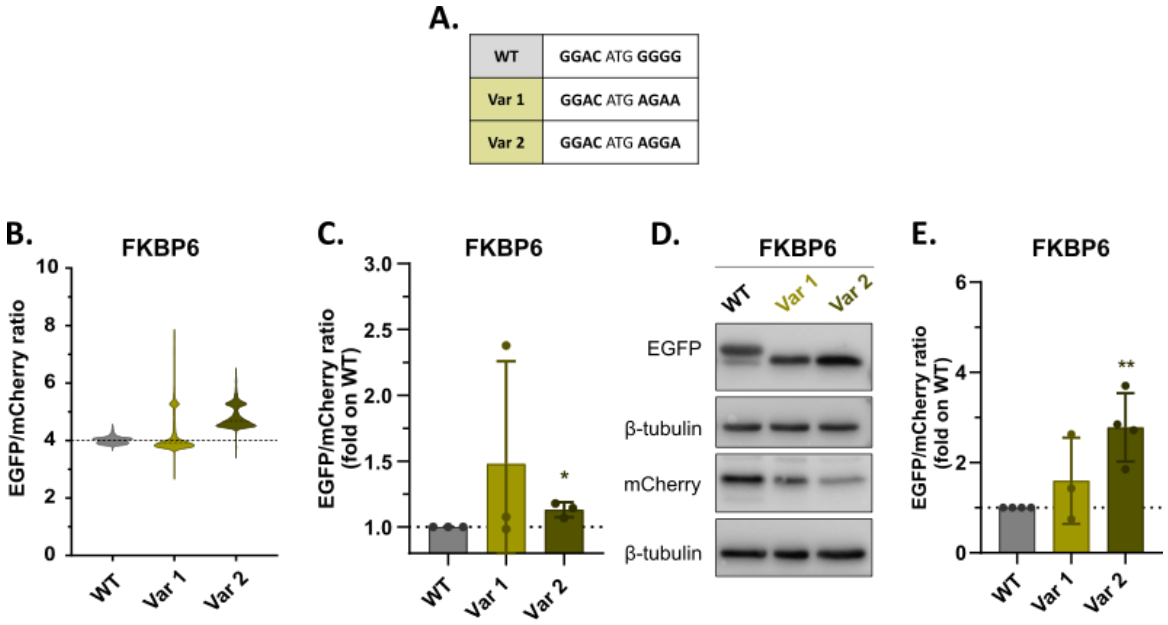


***Figure 17 FKBP6* validation**. HEK293T cells were transiently transfected with EGFP-IRES-mCherry bearing FKBP6 WT, Var 1, and Var 2 Kozak sequence and analyzed 3 days post-transfection. **A.** *FKBP6* WT and variants Kozak sequences. **B., C.** *Translational enhancement analyzed as EGFP/mCherry expression by high content image analysis.* **B.** The violin plots report the distribution of the data from n=3 biological replicates. The dashed line indicates the population median; **C.** the histogram represents the median of the populations analyzed by high content image analysis from the 3 biological replicates. **D, E.** *Increased EGFP/mCherry expression analyzed by immunoblotting.* **D.** One of the blots is shown. **E.** Western blot quantification. Translational strength was calculated as EGFP/mCherry ratio, each normalized over the respective housekeeping. Data are means ± SD from n=3/4 biological replicates. Statistically significant differences were calculated by unpaired t-test of each variant versus the corresponding WT.

GALR1 is one of the three G-protein-coupled receptors (GALR1-3) of galanin, a neuropeptide that exerts a wide variety of physiological functions in the central and peripheral nervous system[180]. *GALR1* is one of the genes located in a chromosomal region deleted in the 18q deletion (18q-) syndrome, a congenital disorder with a prevalence of approximately 1/40.000 live births[181]. Congenital Aural Atresia (CAA), an ear malformation, is one of the most common features of this syndrome. *GALR1* haploinsufficiency was demonstrated to be a primary cause of the CAA phenotype in 18q- patients[182–184]. Both variants identified in the screening (Var 1 and Var 2) were able to enhance EGFP/mCherry expression, by 27%, and 19,6% with respect to the WT (**Fig. 18 A, B**). This result was confirmed by the immunoblotting analysis (**Fig. 18 C, D**).

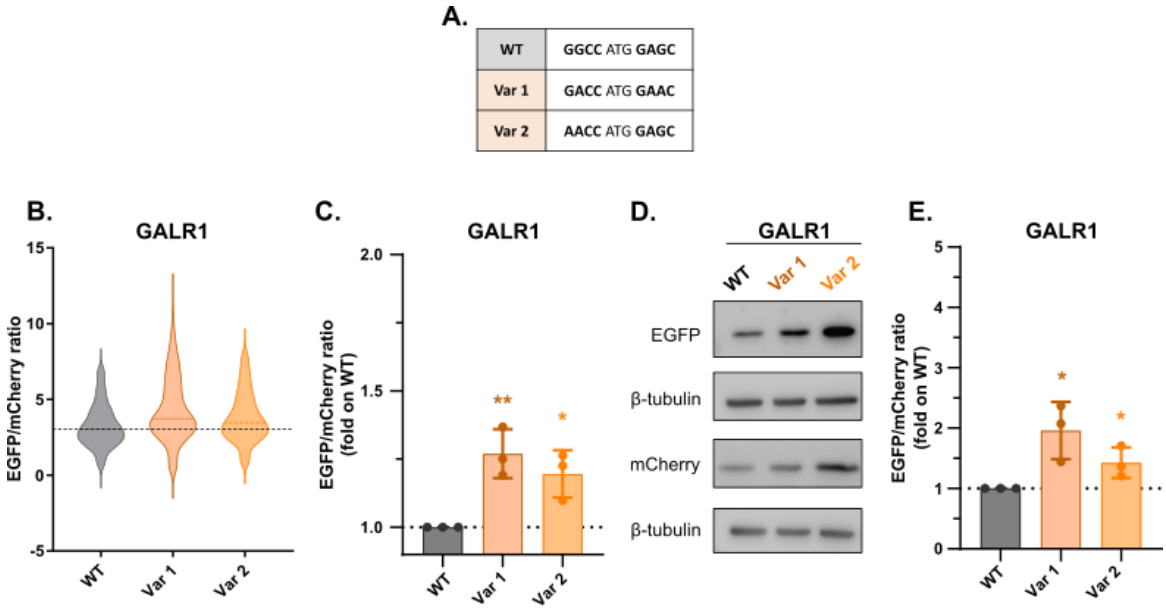


***Figure 18 GALR1* validation**. HEK293T cells were transiently transfected with EGFP-IRES-mCherry bearing GALR1 WT, Var 1, and Var 2 Kozak sequence and analyzed 3 days post-transfection: **A.** *GALR1* WT and variants Kozak sequences. **B., C.** *Translational enhancement analyzed as EGFP/mCherry expression by high content image analysis.* **B.** The violin plots report the distribution of the data from n=3 biological replicates. The dashed line indicates the population median; **C.** the histogram represents the median of the populations analyzed by high content image analysis from the 3 biological replicates. **D., E.** *Increased EGFP/mCherry expression analyzed by immunoblotting.* **D.** One of the blots is shown. **E.** Western blot quantification. Translational strength was calculated as EGFP/mCherry ratio each normalized over the respective housekeeping. Data are means ± SD from n=3 biological replicates. Statistically significant differences were calculated by unpaired t-test of each variant versus the corresponding WT.

NRXN1 encodes for two membrane proteins, NRXN1-α and -β, that belong to the family of neurexins and play a critical role in synapse formation and neurotransmission[185,186]. Haploinsufficiency of both *NRXN1* isoforms has been linked to several neurodevelopmental disorders, such as schizophrenia[187], autism spectrum disorder[186,188], intellectual disability, and growth retardation[189]. Moreover, it was shown that *C.elegans* lacking *nrx-1*, the orthologue of human neurexin, display defective behaviors such as explorative capacity. This phenotype was rescued in strains expressing human NRXN1-α or -β[190]. Our study focused on NRXN1-α, the longest isoform, that shares only the C-term with NRXN1-β. Heterozygous deletions in the first exons of the gene are reported to cause developmental delays and ASD, demonstrating that downregulation of *NRXN1α* alone is sufficient to manifest the haploinsufficient phenotype[191,192]. *NRXN1* Var 1 induced 12,8% up-regulation of the protein (**Fig. 19B, C**), confirmed by immunoblotting (**Fig. 19D, E**).

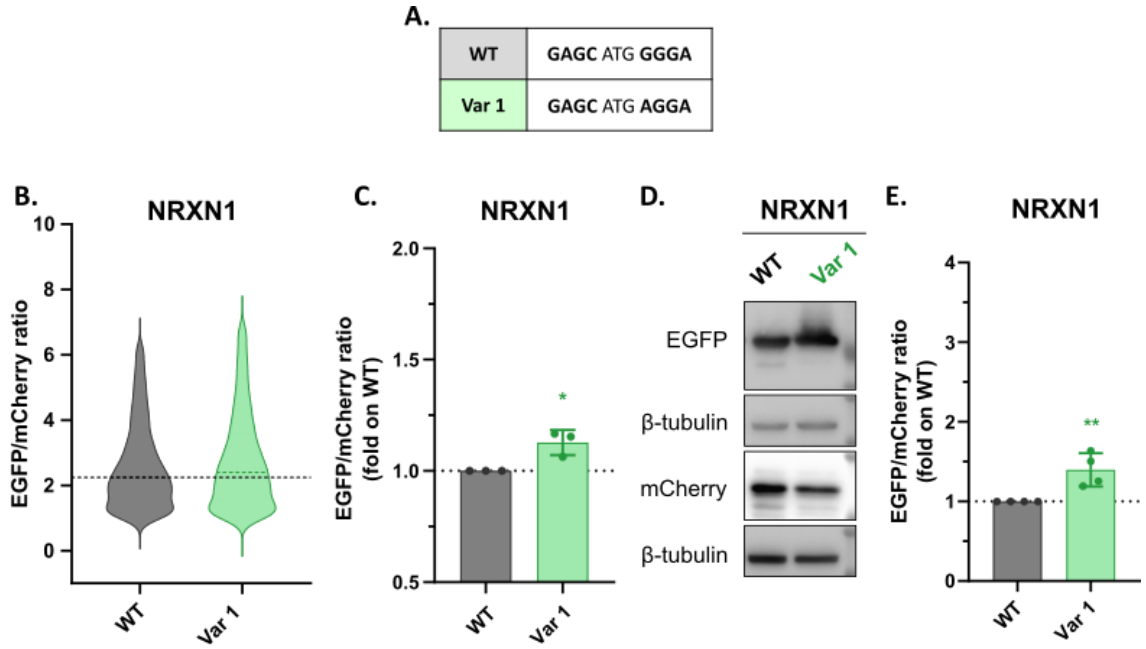
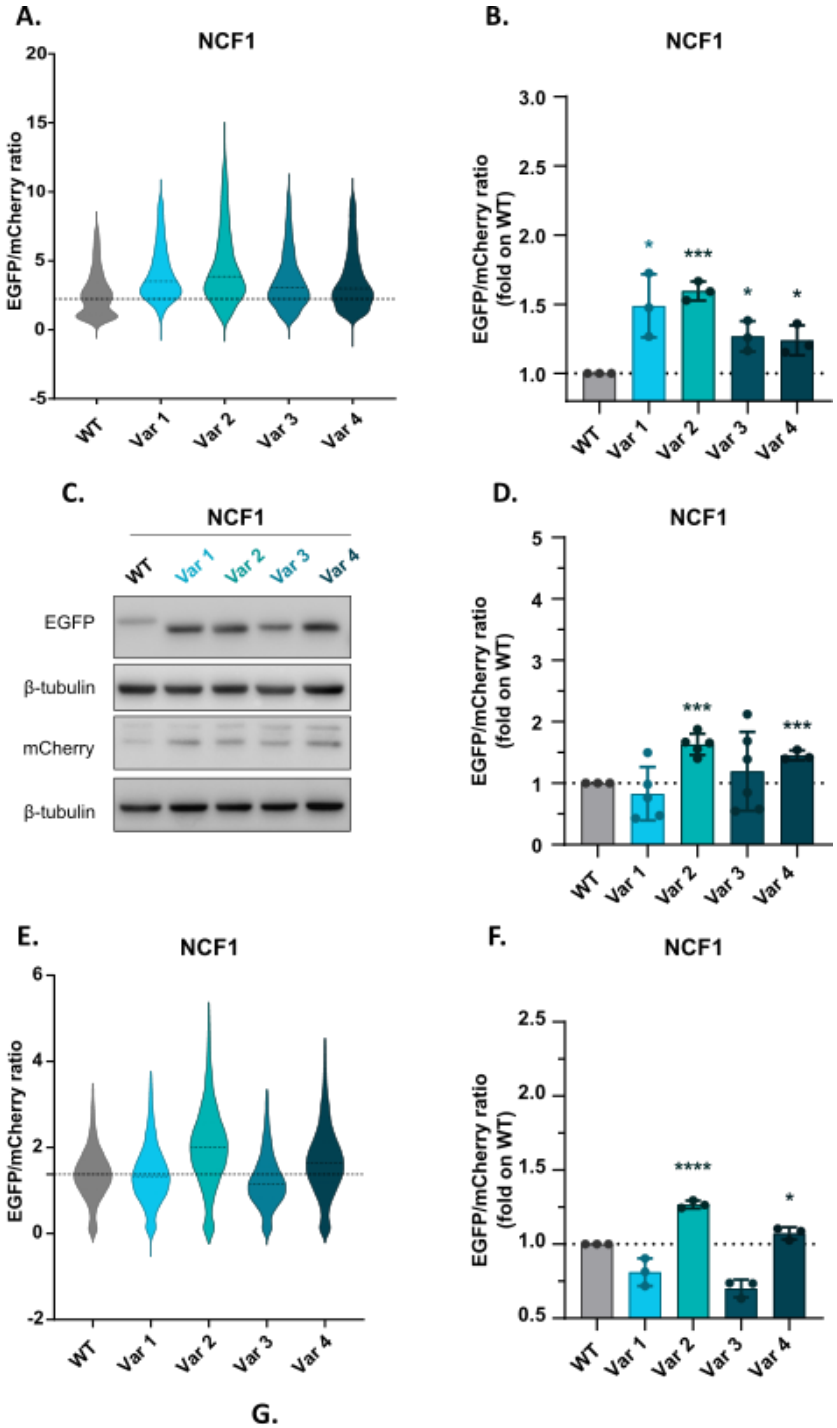


***Figure 19 NRXN1* validation**. HEK293T cells were transiently transfected with EGFP-IRES-mCherry bearing NRXN1 WT and Var 1 Kozak sequence and analyzed 3 days post-transfection: **A.** *NRXN1* WT and variants Kozak sequences. **B., C.** *Translational enhancement analyzed as EGFP/mCherry expression by high content image analysis.* **B.** The violin plots report the distribution of the data from n=3 biological replicates. The dashed line indicates the population median; **C.** the histogram represents the median of the populations analyzed by high content image analysis from the 3 biological replicates. **D., E.** *Increased EGFP/mCherry expression analyzed by immunoblotting.* **D.** One of the blots is shown. **E.** Western blot quantification. Translational strength was calculated as EGFP/mCherry ratio each normalized over the respective housekeeping. Data are means ± SD from n=4 biological replicates. Statistically significant differences were calculated by unpaired t-test of each variant versus the corresponding WT.

NCF1 (Neutrophil cytosolic factor-1) encodes for $p47^{phox}$, one of the components of phagocyte NADPH oxidase[193]. This enzyme in granulocytes and macrophages is responsible for the formation of superoxide and the respiratory burst, necessary for the microbicidal activity[194]. Defects in any NADPH oxidase components result in Chronic Granulomatous Disease (CGD), a primary immunodeficiency with recurrent bacterial and fungal infections[195]. The most common autosomal mutation of $p47^{phox}$ linked to the disease is a homozygous deletion of a dinucleotide repeat, leading to the complete absence of the protein[195,196]. However, there is a huge heterogeneity in the remaining CGD cases due to *NCF1* mutations[197]. Moreover, $p47^{phox}$ haploinsufficiency has been found to cause reduced respiratory burst activity in a mice model[198], and, in CGD carriers, it is proposed to be a susceptibility factor for inflammatory bowel disease[199,200]. Four variants emerged from the screening (**Fig. 20G**), all of which resulted in increased translational efficiency as analyzed by high content image analysis (**Fig. 20A, B**): 48,9% for Var 1, 59,6% for Var 2, 26,9% for Var 3, and 23,8% for Var 4 (**Fig. 20B**). Western blot analysis evidenced Var 2 and Var 4 as significantly able to enhance protein expression (**Fig. 20C, D**).



G.

WT	AGTC ATG GGGG
Var 1	AGTC ATG AAAA
Var 2	AGTC ATG GAAA
Var 3	AGTC ATG AGAA
Var 4	AGTC ATG GGAA

***Figure 20 NCF1* validation in HEK293T**. **A., B., C., D.** Cells were transiently transfected with EGFP-IRES-mCherry bearing NCF1 WT, Var 1, Var 2, Var 3, and Var 4 Kozak sequence and analyzed 3 days post-transfection: **A., B.** *Translational enhancement analyzed as EGFP/mCherry expression by high content image analysis.* **A.** The violin plots report the distribution of the data from n=3 biological replicates. The dashed line indicates the population median; **B.** the histogram represents the median of the populations analyzed by high content image analysis from the 3 biological replicates. **C., D.** *Increased EGFP/mCherry expression analyzed by immunoblotting.* **C.** One of the blots is shown. **D.** Western blot quantification. Translational strength was calculated as EGFP/mCherry ratio each normalized over the respective housekeeping. Data are means ± SD from n=3/5 biological replicates. Statistically significant differences were calculated by unpaired t-test of each variant versus the corresponding WT. **E., F.** Cells were transduced at low MOI with lentiviral particles of EGFP-IRES-mCherry bearing NCF1 WT, Var 1, Var 2, Var 3, and Var 4 Kozak sequence and analyzed 3 days post-transduction by flow cytometry. **E.** The violin plots report the distribution of the data from n=3 biological replicates. The dashed line indicates the population median; **B.** the histogram represents the median of the populations analyzed by high content image analysis from the 3 biological replicates. **G.** *NCF1* WT and variants Kozak sequences.

Collectively, these data demonstrate the validity of our Kozak screening and data analysis approach and highlight at least one variant for each of the selected genes that can significantly enhance Kozak strength, as demonstrated by the validation with two independent and orthogonal techniques.

We then decided to better investigate the shift in the EGFP band of some Kozak variants that was observed in NCF1, PPARGC1B, and FKBP6 western blot analysis. We noticed that the shift is probably due to a change in the third AA encoded by the protein, caused by nucleotide variations in the Kozak sequence. In particular, all the Kozak variants that encode for Lysine (K) as the third AA, showed a band shift in the western blot analysis. *PPARGC1B*, *FKBP6*, and *NCF1* WT Kozak encode, instead, for Glutamic acid (E), and present a higher band. Finally, *NRXN1* and *GALR1* encode for the same AA in both WT and Var, therefore the band is at the same height (**Fig. 21**).

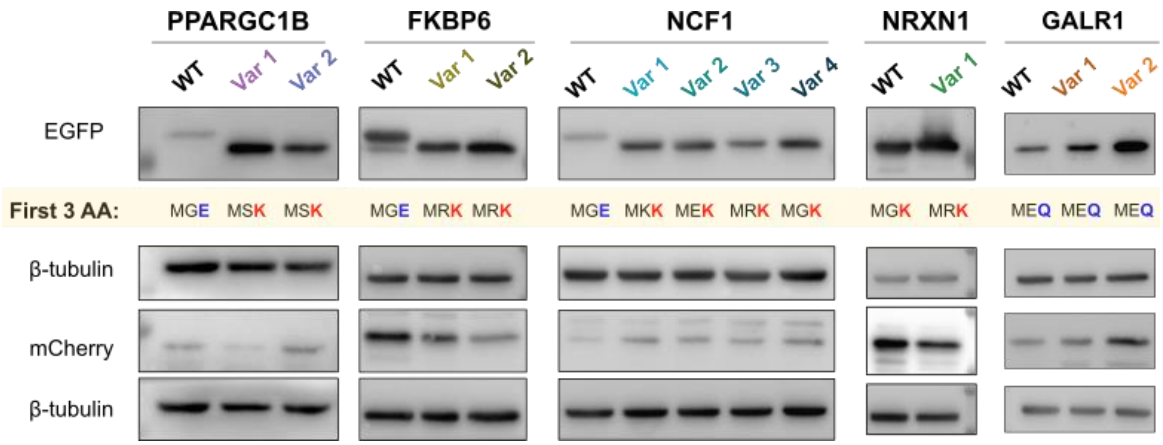


***Figure 21* Western blots of the 5 hit genes and respective encoded amino acids (AA).** Highlighted in yellow, the first 3 AA encoded by EGFP as a result of the variations inserted in the Kozak sequence. (M=Methionine; G=Glycine; E=Glutamic Acid; R=Arginine; Q=Glutamine; K=Lysine).

Next, we decided to validate the results obtained for *NCF1* since it was the gene for which we reached the highest translational up-regulation. Before introducing the variants in the endogenous locus, we sought to validate the candidate nucleotide conversions with a third approach. For this reason, we produced lentiviral particles of each Kozak variant-bearing reporter and transduced HEK293T cells at low MOI (reproducing the protocol used for the library transduction) and we analyzed the EGFP/mCherry ratio with flow cytometry three days post-transduction (**Fig. 20E, F**). The results confirmed what was observed so far in transient transfection, with Var 2 and Var 4 being the best performing variants in all the validations performed. Western blot analysis confirmed the increased EGFP/mCherry protein expression ratio with the two candidate variants (data not shown). Finally, we further confirmed *NCF1* translation enhancement in another cell model, U2OS (**Fig. 22**). In this system, Var 2 enhanced translation by 30,2% and Var 4 by 44,7% (**Fig. 22A, B**). Immunoblotting analysis confirmed the statistically significant increase operated by these two variants, reinforcing our previous data (**Fig. 22C, D**).

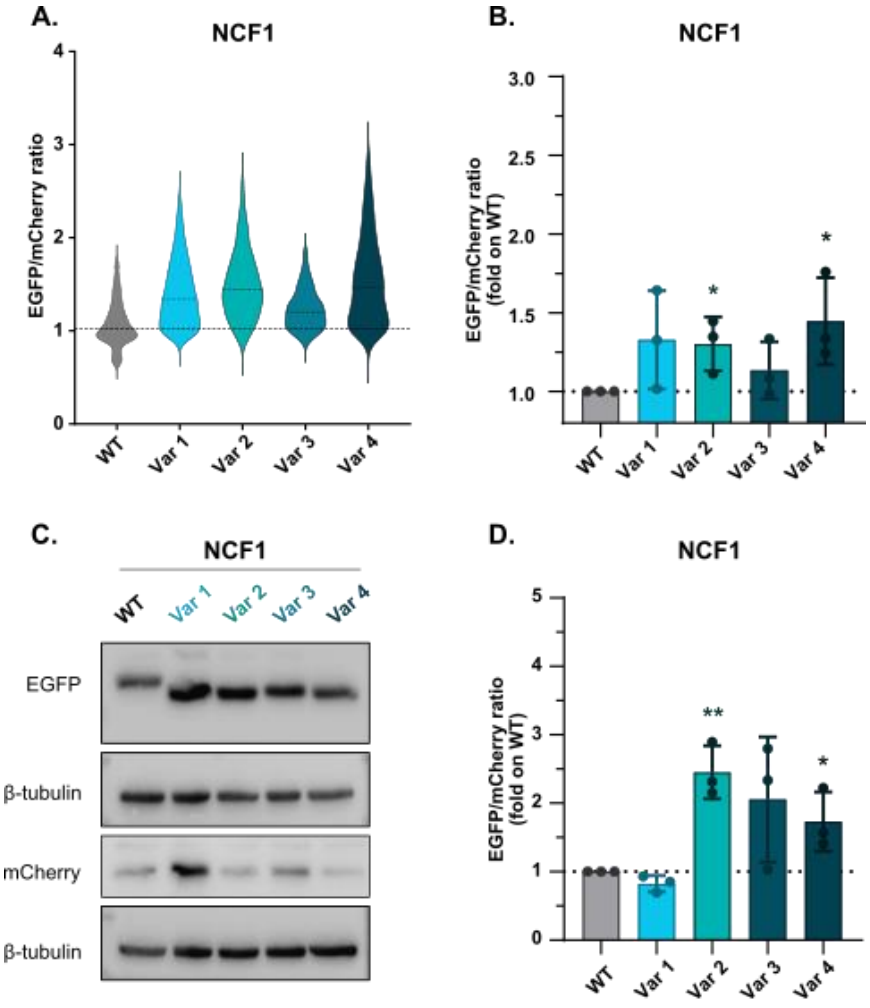


***Figure 22 NCF1* validation in U2OS cells. A., B., C., D.** Cells were transduced at low MOI with lentiviral particles of EGFP-IRES-mCherry bearing NCF1 WT, Var 1, Var 2, Var 3, and Var 4 Kozak sequence and analyzed 3 days post-transduction. **A., B.** *Translational enhancement analyzed as EGFP/mCherry expression by high content image analysis.* **A.** The violin plots report the distribution of the data from n=3 biological replicates. The dashed line indicates the population median; **B.** the histogram represents the median of the populations analyzed by high content image analysis from the 3 biological replicates. **C., D.** *Increased EGFP/mCherry expression analyzed by immunoblotting.* **C.** One of the blots is shown; **D.** Western blot quantification. Translational strength was calculated as EGFP/mCherry ratio each normalized over the respective housekeeping. Data are means ± SD from n=3 biological replicates. Statistically significant differences were calculated by unpaired t-test of each variant versus the corresponding WT.

Base editing-mediated Kozak modification of the NCF1 endogenous locus enhances translation

Next, we aimed at performing base editing in the *NCF1* endogenous locus to reproduce the validated conversions in the Kozak sequence and measure the increase in protein translation. In particular, we focused on inserting Variants 2 and 4, which resulted in a significant increase in translation in all the validation techniques employed. The conversions to reproduce Var 2 and Var 4 are G-to-A, a substitution that can be inserted by cytosine base editors (CBE). We tested the two latest developed CBEs (BE4max and AncBE4max), for which the window of action can be assessed between positions 4-8 (counting the PAM as positions 21-23). Given the fact that Var 2 and Var 4 differ only for adenine in position +2 (**Fig 23A**), we designed three sgRNAs targeting the *NCF1* Kozak sequence with the three target guanines at the limit of the base editor window of action (**Fig. 23B**).

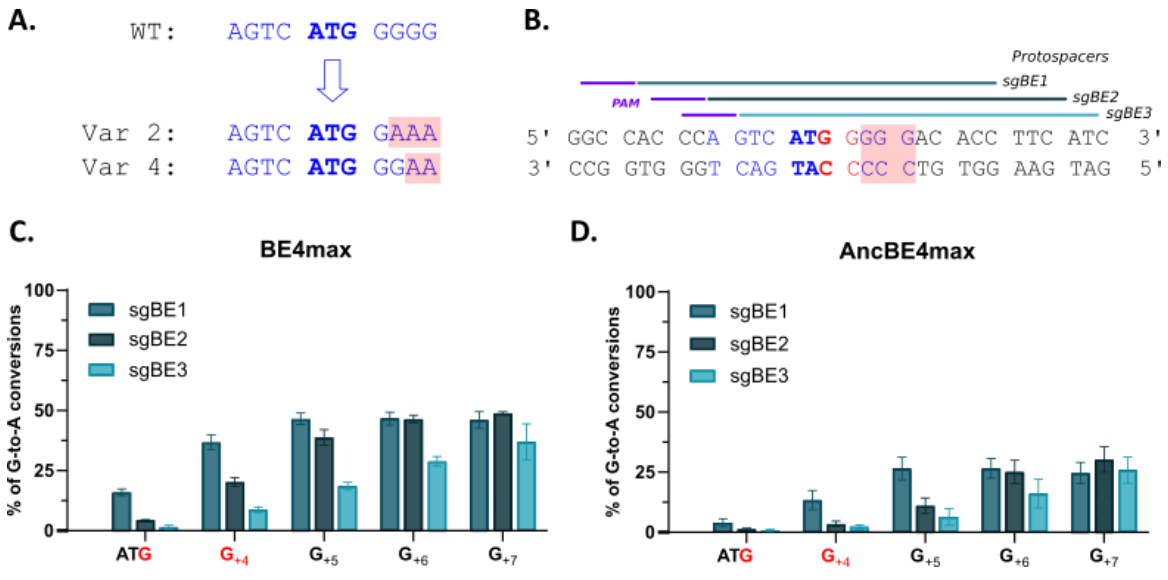


***Figure 23 NCF1* Kozak sequence base editing in HEK293T cells. A.** Schematic representation of the *NCF1* Wild type (WT), Var 2, and Var 4 Kozak sequences. The starting codon is in bold blue; the nucleotide changes in the variants are highlighted in pink; **B.** *NCF1* target locus, with the Kozak sequence in blue, the starting codon in bold, the target guanines highlighted in pink, and the potential bystanders in red. Protospacer-adjacent motif (PAM) sequences are drawn in violet, with the respective protospacers; **C., D.** Editing efficiency at target and bystander (in red) guanines analyzed with EditR software 5 days post-transient transfection of either base editor BE4max (**C.**) or AncBE4max (**D.**) and the respective sgRNAs. Editing efficiencies are calculated as % of editing with the listed sgRNAs minus % of editing with the scrambled sgCTRL. Data are means ± SD from n=3 independent experiments.

First, we performed base editing in HEK293T cells to choose the best performing sgRNA. Cells were transiently transfected with either BE4max or AncBE4max in combination with each sgRNAs. **Figure 23C., D.** shows the percentage of corrected G-to-A conversions (y-axis) for each position in the *NCF1* Kozak sequence (x-axis, with the A of ATG being position +1), analyzed by Sanger sequencing 5 days post-transfection. BE4max was the most efficient editor, with an average of correct conversions of 45% in the desired positions (G_{+5}, G_{+6}, and G_{+7}) with sgBE1 and sgBE2, and of 28% with sgBE3 (**Fig. 23C**). However, all the sgRNAs caused bystander editing also in G_{+4} and the guanine of the starting codon (ATG) and for this reason, BE4max was discarded.

AncBE4max efficiency was lower, with an overall mean percentage of correct conversions of ~25% for sgBE1, 22% for sgBE2, and 16% for sgBE3 (**Fig. 23D**). Despite this, the editing was significantly more precise, with no bystander effects except for sgBE1 in position G_{+4}. For this reason, we decided to proceed with AncBE4max and sgBE2, since it was the combination that gave us maximal efficiency and precision of editing.

We then evaluated the system in Raji cells, a B lymphocyte cell line derived from Burkitt's lymphoma, that constitutively expresses the gene of interest. We electroporated Raji cells with AncBE4max and sgBE2, obtaining an editing efficiency in the bulk population lower than 30% in the best-edited position, as analyzed by Sanger sequencing 5 days after electroporation (**Fig 24A**). For this reason, we then decided to isolate single clones from the edited cells. We selected two cell lines that reproduce the desired nucleotide changes displayed in Fig. **23A**, hereafter referred to as Var 2 and Var 4 cells (**Fig. 24B, C**). In particular, in both cell lines, G_{+7} was completely converted to A, editing efficiency in G_{+6} was ~30%, while G_{+5} was partially edited (~17%) for Var 2 but left unedited for Var 4, reproducing the desired variants (**Fig. 23A, 24C**). No bystander editing was observed (**Fig. 24B**).

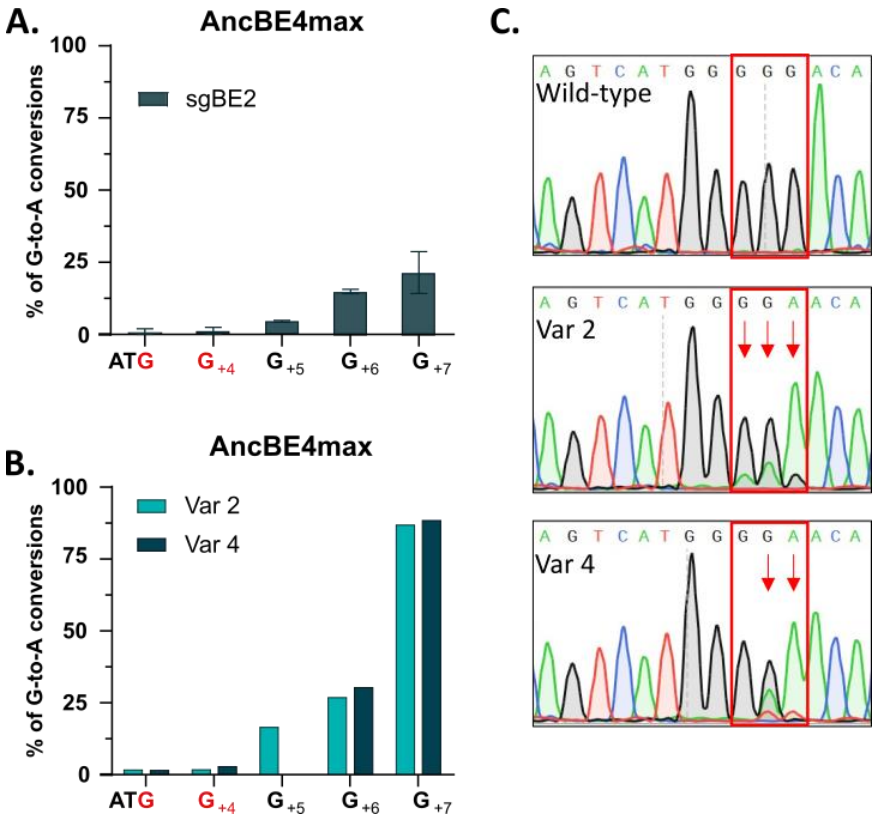


***Figure 24 NCF1* Kozak sequence base editing in Raji cells. A.** Editing efficiency in the bulk population at target and bystander (in red) guanines, analyzed with EditR software 5 days post-electroporation of AncBE4max and sgRNABE2. Data are means ± SD from n=3 biological replicates. **B.** Editing efficiency in the two cell lines isolated from the bulk population (Var 2 and Var 4) at target and bystander (in red) guanines. Editing efficiencies are calculated as % of editing with the listed sgRNAs minus % of editing with the scrambled sgCTRL. **C.** Sanger sequencing chromatograms of *NCF1* Kozak sequence in Raji WT, Var 2, and Var4 cells.

Next, we evaluated the expression of $p47^{phox}$, the protein encoded by *NCF1*, in the edited cells. Western blot analysis revealed an increase of $p47^{phox}$ with both variants as compared to the WT (**Fig. 25A, B**). In particular, Var 2 increased protein expression by 69,2% and Var 4 by 49,7% (**Fig. 25B**). Moreover, we analyzed *NCF1* mRNA expression in the three cell lines and found that they were unchanged, strongly supporting the idea that the increase in protein expression is the result of an enhancement in the translation of *NCF1* (**Fig. 25C**). To confirm this, we performed a sucrose gradient fractionation in Raji cells Var 2, Var 4, and WT cells. First, we identified the fractions containing the polysomes (representing the actively translating ribosomes) by western blot analysis of two polysome markers: RPS6 (40S ribosomal protein S6), and RPL26 (60S ribosomal protein L26) (**Fig. 25D**). We then measured the Translational Efficiency (TE) of *NCF1* by qPCR, by calculating the ratio between the fold change of *NCF1* in polysomes (fractions 8-9) and in the total RNA (fractions 4-9) (**Fig. 25E**).

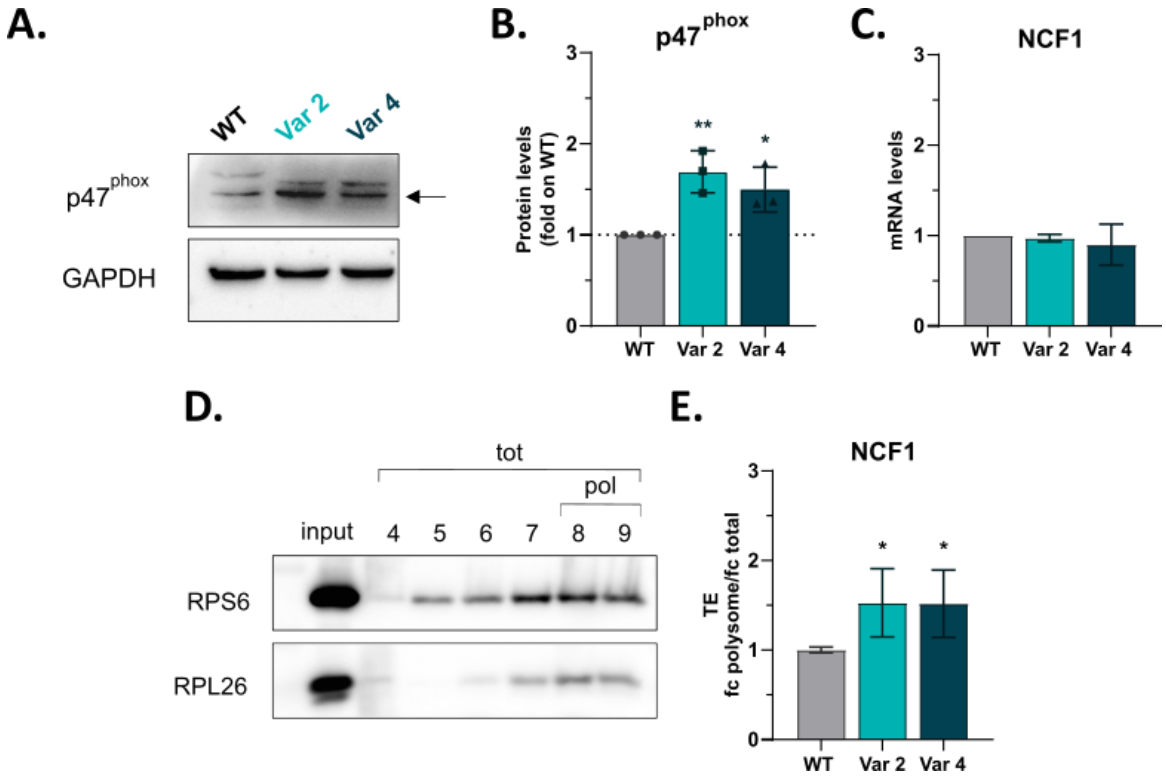


***Figure 25 NCF1* translational enhancement validation. A.** Western blot analysis of p47^{phox} expression in Raji cells (WT, Var 2, and Var 4). A representative blot result is shown. The 47KDa band corresponding to p47^{phox} is indicated by the arrow. **B.** Western blot quantification. p47^{phox} expression was normalized on the housekeeping and the fold change on the WT expression is shown. Data are means ± SD from n=3 biological replicates. **C.** qPCR of *NCF1* on WT, Var 2, or Var 4 cells. Data are means ± SD from n=3 biological replicates. **D.** Representative western blot of two polysomal markers (RPS6 and RPL26) in the fractions isolated by sucrose gradient fractionation. The input is the cellular cytoplasmic lysate loaded on the sucrose gradient. tot=fractions corresponding to the total RNA; pol=fractions selected as polysomes and used in E. **E.** Translational efficiency (TE) quantification of *NCF1* in Var 2 and Var 4 with respect to the WT cells. TE is the ratio between polysomal (fractions 8-9) and total (fractions 4-9) RNA changes (fc=fold change) measured by qPCR. Data are means ± SD from n=3 independent experiments. Statistically significant differences were calculated by unpaired t-test of each variant versus the WT.

This experiment showed that the increase in protein expression is due to an increased level of mRNA loaded on the polysomes, and therefore actively translated.

Collectively, these results showed the feasibility of targeting the Kozak sequence with base editors and that reproducing the identified variants in the endogenous locus causes translational up-regulation of the target gene.

DISCUSSION

Haploinsufficient diseases are caused by the mutational inactivation of one of the two alleles of a gene, which leads to reduced protein levels that are not sufficient to guarantee the physiological function. Although HI is responsible for hundreds of human diseases, few attempts have been made at rescuing protein expression of haploinsufficient genes with therapeutic purposes. In one example, Hsiao and colleagues increased the expression of *SCN1A*, whose HI is responsible for Dravet syndrome, by using compounds blocking an antisense-non-coding-RNA that inhibits the target gene[201]. In some attempts, CRISPR-Cas-based genome editing has been employed to rescue HI. In these works, the authors used CRISPR-a, a CRISPR-Cas genome editing approach aimed at enhancing gene expression at the transcriptional level. To do this, a catalytically inactive Cas9 (dCas9) is fused to transcriptional activators (such as VP64, VP16, p65, and RTA), to induce overexpression of target genes[202]. Colasante and colleagues, for example, used CRISPR-a to upregulate *SCN1A* transcription in Dravet syndrome[203]. In another study, the authors focused on two endogenous genes responsible for the obesity phenotype in mice (*Sim1* and *Mc4r*), by targeting with CRISPR-a their promoter or enhancer sequences[204]. These studies clearly demonstrate that HI can be rescued by using gene editing approaches to target the remaining functional allele. However, this approach does not represent a permanent cure for the disease. In fact, CRISPR-a does not install a permanent mutation in the genome to enhance protein production, meaning that its use in therapy would require multiple rounds of treatment, hard to be transferred to the clinic.

Here we propose a novel approach for the regulation of gene expression aimed at rescuing haploinsufficiency, relying on CRISPR-Cas base editors. In particular, we focused on the modification of the Kozak sequence, which has been reported to control translational efficiency. Our system provides several advantages: firstly, base editors can install permanent nucleotide conversions in the genomic DNA; secondly, nucleotide modifications on the Kozak sequence allow to target virtually all HI genes by inducing small, controlled increases in translational efficiency. This is crucial because the final goal is to achieve an increase in protein levels sufficient to rescue the HI phenotype but not high enough to create imbalances and be incompatible with the physiology of the cells. Finally, this approach relies entirely on the translational machinery, regardless of the transcriptional efficiency at which the target genes are expressed.

As stated before, the Kozak sequence has been already targeted by genome editing approaches aimed at modulating translational efficiency. However, no attempt has been made to manipulate it with the aim of upregulating protein production with therapeutic purposes.

To demonstrate the feasibility of the proposed approach, we performed a proof-of-concept experiment focused on enhancing EGFP translation from a reporter vector. First, we created a suboptimal version of the EGFP Kozak sequence by mutating a single nucleotide (C-1T) and showed that this conversion alone is able to downregulate EGFP intensity up to 4-fold. This step was crucial in demonstrating the impact of the Kozak sequence on translational efficiency (**Fig. 7**). Secondly, we corrected this nucleotide variation with base editors and observed a significant increase in EGFP expression, clearly providing a proof-of-principle that base editors can be used to selectively mutate single nucleotides in the Kozak sequence (**Fig. 8**). The results of this pilot experiment indicated that the proposed approach could be suitable to modulate the expression of HI disease genes.

At least 300 HI diseases have been described. We decided to discard those associated with cancer and tumorigenesis since a monogenic disorder would represent a more suitable target for our approach. Therefore, we systematically screened 231 HI genes Kozak sequences and a number of respective variants, in order to find precise nucleotide conversions able to upregulate translation. We performed a high-throughput screening taking advantage of the previously described technique of FACS-seq[60], in which a bicistronic reporter vector is used to evaluate the translational efficiency of a library of selected Kozak sequences. In our system, EGFP was under the control of the Kozak sequence variants, while mCherry expression was used as a control. The results demonstrated that our system is a valid and feasible approach to measure Kozak sequence strength and to compare WT and variants of the respective genes. The logos of the consensus sequences derived from each of the four gates generated in the screening (**Fig. 13B**) showed, as expected, that sequences with a measured lower translational efficiency display a Kozak consensus sequence distant from the one described in the literature as optimal. We then validated 5 HI genes Kozak sequences (*PPARGC1B, FKBP6, GALR1, NRXN1,* and *NCF1*) and a number of respective variants (16 variants in total), chosen from the screening as able to upregulate translation, and we found a good correlation between these measurements and those observed with the high-throughput method (**Fig. 16-20**).

In our library, we did not take into account the previously reported notions about Kozak efficiency. Instead, we designed the nucleotide variations for each HI gene on the basis of the conversions that can be reproduced by the originally developed base editors (cytosine and adenine base editors, able to introduce transition mutations). Moreover, we decided to expand the number of positions in the Kozak sequence analyzed to 4 nucleotides after the ATG (NNNN_ATG_NNNN), to investigate the role of those positions inside the coding sequence on translational efficiency. The impact of the first nucleotides after the ATG on the Kozak sequence strength is highly debated. Indeed, some reports argue that the importance of those positions derives from the amino acid (AA) they encode (see section 2, *Novel insights on Kozak sequence*). Our results, however, corroborate the notion that the sequence rather than the amino acids encoded is the major determinant. For example, we validated two *FKBP6* variants from the screening. In the WT Kozak (GGAC ATG **GGGG**) the first AA after the ATG is Glycine (G). Var 1 (GGAC ATG **AGAA**) and Var 2 (GGAC ATG **AGGA**) introduce the same AA change, from Glycine to Arginine (R). However, only Var 2 was able to upregulate *FKBP6* translation, demonstrating that the translational strength depends on the sequence, and does not rely on a particular AA (**Fig. 17, 21**). Another example is given by *PPARGC1B*, for which Var 1 (AACA ATG **AGTA**) and Var 2 (AGCA ATG **AGTA**) introduce the same AA changes from the WT (AGCA ATG **GGTG**): GE to SK after the starting methionine. In this case, both variants enhanced translation but at different levels. The two variants differ only for the nucleotides before the ATG, showing, as expected, the importance of those positions (**Fig. 16**).

In the majority of the immunoblots performed to validate the effect of the variants on EGFP levels, a shift in the EGFP band was observed for the Var with respect to the WT Kozak. Interestingly, this shift can be explained by the fact that all the validated variants encode for Lysine (K) as third AA, differently from the WT. On the other hand, both the WT and the Var sequences of *NRXN1* and *GALR1* encode for the same AA, and the EGFP bands are at the same height (**Fig. 21**). It can be argued that the substitution of a nonpolar AA (Glycine or Glutamic acid for most of the WT) with a polar one (K), can produce a shift in the migration of the bands. Nevertheless, the data showed that the K at that specific position did not correlate with increased translation. Indeed, for *NCF1*, all the 4 Vars encode for K, however, only 2 out of 4 were able to upregulate translation, corroborating the idea that it is the sequence the major determinant, and not the AA.

This AA substitution did not impact the validation we performed by high-content image analysis or by Flow cytometry, since a change in the third AA of EGFP should not be able to affect its fluorescence. EGFP is indeed well tolerant to deletions or substitutions in the N- or C-terminal of the protein, being the chromophore in the central α-helix[205]. Additionally, in a study analyzing a set of EGFP mutants, it has been demonstrated that deletions or substitutions at the second or third AA of EGFP coding sequence do not impair the fluorescence[206].

Our screening has some limitations. For example, aiming at detecting small differences in translation efficiency, the long half-life of the fluorescent proteins might have influenced the differences observed. Indeed, three days after the delivery of the plasmid, residual fluorescent protein expression could impact the results of the validation, thereby leading to an underestimation of the translational increase. This problem could be addressed by using a destabilized fluorescent protein (such as EGFP-PEST[207]).

Another limit of the screening is that it doesn't consider other regulatory components and mechanisms that cooperate with the Kozak strength to promote translation, such as the length of the 5'UTR or the presence of secondary structures adjacent to the start site. Moreover, the availability of some initiation factors can vary in response to stress stimuli or as a result of mutations. For instance, eIF1, 1A, and 5, the three factors involved in efficient start codon recognition, can be post-transcriptionally modified with the result of impaired fidelity of start codon selection and preferential expression from suboptimal Kozak sequences[51,208,209]. All these features reflect mRNA and cell-specific situations that could impair our approach and will have to be investigated singularly when upregulating the HI gene at the endogenous locus level.

Finally, we decided to further validate the *NCF1* variants on the endogenous locus. *NCF1* encodes for one of the subunits of the enzyme NADPH oxidase, and its haploinsufficiency has been linked to Chronic Granulomatous Disease (CGD), a primary immunodeficiency. Different validation techniques indicated two variants as significantly able to upregulate translation (Var 2 and Var 4) (**Fig. 20,22**), therefore we used CRISPR-Cas base editors to induce their nucleotide changes in a cell line expressing the target gene. Both Var 2 and Var 4 Kozak sequences lead to an increased protein expression (**Fig. 25A, B**). Interestingly, Var 2 induced an upregulation similar to that observed in EGFP (**Fig.20D**), and it was more efficient than Var 4, as it was observed in the reporter system. The results showed that the

increase in protein production was due to a translational modulation, indeed the mRNA levels of *NCF1* were unchanged (**Fig. 25C**). Furthermore, polysomal fractionation revealed that *NCF1* was significantly more loaded onto the polysomes and therefore actively translated in the Var 2 and Var 4 cells versus the WT, which could be likely due to an increased Kozak strength and therefore number of ribosomes engaged in *NCF1* translation (**Fig. 25D, E**). Deeper studies must be carried out to understand which is the correct protein level needed to restore the physiological functions in HI models of the disease. Additionally, since *NCF1* is part of an enzymatic complex, the concentration of the other components must be taken into account as well, to balance the stoichiometry of the subunits[70]. Moreover, to move towards a therapeutic approach based on the manipulation of the Kozak sequence, the AA changes induced with the nucleotide conversions in the endogenous locus should be carefully analyzed. Despite changes in the N-term should not affect the functionality or structure of the protein, additional layers of translational regulation could be impacted, such as the removal of the starting methionine[47], or the N-end rule, a principle by which the AA in the N-term part of the protein affects its stability[210].

Nevertheless, these results clearly demonstrated that it is feasible to increase protein production by installing changes in the Kozak sequence.

On a broader level, the results of our screening identified 149 promising variants to upregulate the translation from 47 HI genes. A requirement for harnessing the identified nucleotide changes is the availability of a PAM sequence to place the target position inside the window of action of the base editor. However, the pace at which novel base editors are being developed together with the ones already available, such as the near PAMless CBE and ABE[144] increases the chance of finding a combination of sgRNA and editor to target all the HI genes identified.

The validation experiments of the 5 hits we selected indicated that the increase in translational efficiency was never above 60% of the WT. This is crucial since we aimed at inducing persistent changes in the genome to achieve small increases in gene expression. In fact, overexpression of the gene could lead to the same detrimental effect as the lack of expression. For example, overexpression of $p47^{phox}$, the protein encoded by *NCF1*, has been linked to Chronic Heart Failure (CHF)[211] and kidney injury in mice (albuminuria and kidney fibrosis)[212] due to excessive ROS production. On the other hand, small increases in translation should be in principle sufficient to rescue HI. For example, the T-1C

polymorphism in the *CD40* gene leads to increased protein production and Grave's disease (see section 2, *Kozak sequence*). It has been shown that *CD40* Kozak bearing T in -1 was characterized by a reduction in the expression of ~15–30%, proving that modest changes in translation efficiency have biological relevance[37]. Moreover, treatment of several diseases with insufficient protein production such as Cystic Fibrosis[213] and Duchenne's muscular dystrophy[214] with nonsense suppressor drugs, demonstrated that a translational increase of 25% correlated with improvement in functional studies[77].

In conclusion, we provide a novel approach to rescue haploinsufficiency mediated by the Kozak sequence manipulation. The data obtained from this study could be relevant in finding new therapeutic opportunities for a large fraction of the known HI diseases. Moreover, this approach could be targeted to other diseases where an increase in the target gene would be desirable. For instance, in the recessive disease Friedreich Ataxia, a trinucleotide repetition leads to a dramatic decrease in *FXN* protein production[215]. In this case, low levels of correct mRNA are still produced, therefore our translational enhancement approach could rescue the physiological levels.

Finally, our work takes advantage of CRISPR-Cas base editors to fine-tune gene expression acting exclusively on the level of translation, expanding genome editing applications for the treatment of genetic diseases.

MATERIALS AND METHODS

Plasmids

Guide RNAs were cloned inside pUC19 (Addgene plasmid #50005) using BbsI restriction sites as previously described[216]. The base editors used were purchased from Addgene: pCMV_ABE7.10 (#102919); pCMV_ABEmax (#112095); pCMV_BE4max (#112093); pCMV_AncBE4max (#112094).

pWPT-/C-1T-mEGFP-IRES-mCherry was obtained by designing a mutated Kozak sequence as an oligonucleotide and cloning it in pWPT-/GCCACC-mEGFP-IRES-mCherry (Addgene plasmid #49235) using EcoRI and XhoI restriction sites. In the same oligonucleotide, a PAM sequence was added to allow base editing of the Kozak sequence.

pWPT-mCherry was obtained by cloning 5 stop codons as an oligonucleotide in place of the Kozak sequence in pWPT-/GCCACC-mEGFP-IRES-mCherry.

pWPT-mEGFP was obtained by digesting pWPT-/GCCACC-mEGFP-IRES-mCherry with PstI and XmaI restriction enzymes, creating a 367 nt deletion inside the mCherry coding sequence. Blunt ends were generated with DNA Polymerase I, Large (Klenow) Fragment (NEB, M0210S).

For the validation of the hits that emerged from the screening, the Kozak sequence variants were purchased as oligonucleotides and cloned in pWPT-/GCCACC-mEGFP-IRES-mCherry using EcoRI and XhoI restriction sites.

Cell cultures

HEK293T cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM, Life Technologies); U2OS and Raji cells were cultured in RPMI 1640 medium. All media were supplemented with 10% Fetal bovine serum (FBS, Life Technologies), 1% L-glutamine, 100U/ml antibiotics (PenStrep, Life Technologies). The cells were maintained at 37°C in a 5% CO_2 humidified atmosphere.

HEK293T transfection

For FACS analysis experiments, 10^5 HEK293T cells/w were seeded into 24-well plates (Corning). After one day, cells were transfected with 4ul polyethyleneimine (PEI) per well

using 500ng of pWPT-/GCCACC-mEGFP-IRES-mCherry or pWPT/C-1T/mEGFP-IRES-mCherry. Cells were cultured for 3 days before cell detachment and analysis at FACS Canto.

For base editing experiments in HEK293T cells, 10^5 cells/w were seeded into 24-well plates (Corning) and transfected with 4ul PEI per well using 750ng of base editor plasmid, 250ng of sgRNA plasmid, and 100ng of pWPT-C-1T-mEGFP-IRES-mCherry. Cells were culture for 5 days before DNA extraction.

For High Content Screening System Operetta (PerkinElmer) analysis, 10^5 HEK293T cells/w were seeded into 24-well plates (Corning) and transfected with 4ul PEI per well using 100ng of pWPT-mEGFP-IRES-mCherry bearing either the wild type or a variant of the target Kozak sequences emerged from the high-throughput screening. 24 hours post-transfection, cells were detached and plated in a 96 well plate (Corning) (8.000 cells/well). 72 hours post-transfection, cells were analyzed at High Content Screening System Operetta (PerkinElmer). At the same time point, cells were collected for protein extraction and western blot analysis.

Raji electroporation

For base editing experiments, cells were transfected using the Neon transfection system (MPK5000) according to the manufacturer's instructions. Briefly, $7x10^5$ Raji cells/condition were harvested and washed in PBS (Invitrogen). Cells were then resuspended in 100ul of R buffer and electroporated with the following conditions: 1350V, 30s, 1 pulse. After electroporation, cells were immediately transferred to a 12-well plate (Corning) containing pre-warmed antibiotic-free medium. Cells were cultured for 5 days before DNA and protein extraction.

To increase base editing efficiency, 5 days post electroporation Raji edited cells were serially diluted to obtain single clones. Single-cell clones were picked, base editing efficiency was analyzed and the best edited clones (Var2 and Var4) were selected for further experiments.

Base editing efficiency analysis

Genomic DNA was extracted using QuickExtract DNA Extraction Solution (Epicentre). The target region was PCR-amplified using MyTaq HS RedMix 2X (Meridian Bioscience). The oligos used are listed in Table S3. PCR products were purified using NucleoSpin Gel and PCR clean-up (Macherey-Nagel), Sanger sequenced and analyzed by EditR software to evaluate base editing efficiency[217]. Editing efficiencies are shown in comparison with the percentage of editing obtained with the scrambled sgRNA (sgCTRL). When not present, editing efficiency is calculated as the percentage of editing obtained with the sgRNA used minus the percentage of editing obtained with sgCTRL.

Kozak sequences library construction

The Kozak sequence variants were synthesized as oligonucleotides on a custom Agilent 244K microarray designed for this purpose. Two libraries were synthesized with 2 different synthesis processes. Library B was synthesized with a High Fidelity Synthesis process (error rate reduced from 1/250 -1/500 nt to 1/600-1/1200 nt). The libraries were purchased as pooled unamplified lyophilized ssDNA oligonucleotides.

The oligonucleotides designed were 98nt long. 11 central nucleotides in each oligo (4 before and 4 after the ATG) represent the Kozak sequence and the variable part of each oligonucleotide. The remaining nucleotides represent the homology arms with the final reporter vector and the restriction sites of the desired enzymes for cloning.

The library was cloned inside pWPT-mCherry, to avoid background EGFP signal in case of reconstitution of the empty vector during ligation or inefficient digestion of the destination vector.

For the plasmid library, pWPT-mCherry was digested with EcoRI and XhoI and the oligonucleotides library was cloned in the linearized vector using NEBuilder® HiFi DNA Assembly Master Mix (compatible with ligation of ssDNA oligos and dsDNA assembly). In particular, 1pmol of resuspended library oligos was ligated with 100ng of digested purified vector following the manufacturer's instructions (NEBuilder). TOP10 *E.coli* were transformed with the ligation product. The colonies were scraped and DNA was purified through a Midiprep purification (Qiagen).

Lentiviral transduction

Lentiviral particles of the Kozak variants library were produced by seeding $10x10^6$ HEK293T cells into 15cm dishes. The day after the plates were transfected with 25ug of the vector together with 16,25ug psPax2 packaging vector and 8,75ug pMD2.G using PEI. After 6 hours of incubation, the medium was replaced with fresh complete DMEM. 48h later, the supernatant containing the viral particles was collected, spun down at 250 g for 5 minutes, and filtered through a 0,45 um PES filter. Lentiviral particles were concentrated by ultracentrifugation for 2 hours at 150.000 g at 4°C with a 20% sucrose cushion. The titers of the lentiviral vectors (reverse transcriptase units, RTU) were measured using the product enhanced reverse transcriptase (SG-PERT) assay as previously described[218].

FACS-seq experiments were carried out on HEK293T cells. To ensure 1000X coverage of the library, viral particles were added to $25x10^6$ HEK293T cells. The titer of viral particles used was calculated to obtain 25% of infection frequency (MOI=0,3), as validated by flow cytometry 3 days post-transduction, ensuring that transduced cells received a single copy of the vector.

For U2OS transduction, lentiviral particles of each Kozak variant were produced as described above. $5x10^4$ cells/w were seeded into 24-well plates (Corning) and the day after were transduced with 3 RTU of lentiviral vectors. 24 hours post-transduction, cells were detached and plated in a 96 well plate (Corning) (4.000 cells/well). 72 hours post-transduction, cells were analyzed at High Content Screening System Operetta (PerkinElmer). At the same time point, cells were collected for protein extraction and western blot analysis.

Fluorescence-activated cell sorting (FACS)

HEK293T cells transduced with the Kozak variants library were sorted into 4 gates according to the EGFP/mCherry ratio as a measure of Kozak strength. All sortings were performed using the FACS Aria IIIu (Becton Dickinson, BD Biosciences) using the GFP channel (488 nm excitation laser, 500 nm splitter, 530/30 nm emission filter) and the mCherry channel (561 excitation laser, splitter 600 nm, 610/20 nm emission filter) and FACS Diva Software (BD Biosciences version 8.0.2). 561nm laser(Yellow-Green) allows optimal mCherry excitation. Cells were resuspended in PBS without Ca^{++} Mg^{++} complemented with 2% BSA, 1% Pen-Strep, and 1,5mM EDTA and filtered through a 30

µm filter (Becton Dickinson). Cells were sorted at low pressure (20-25 psi) with the 100 µm nozzle.

The sorting was divided into two rounds. In the first round, 3 days after lentiviral particles transduction, $5x10^6$ HEK293T cells positive for mCherry expression were sorted to ensure 1000X coverage of the library. The sorted cells were expanded for 2 days and mCherry expression was evaluated before proceeding with the sorting. In the second round of sorting (day 5 post-transduction), mCherry-positive cells were sorted in 4 bins according to their EGFP/mCherry ratio, so that 25% of the population fell in each gate. To maintain 1000X coverage, $1,25x10^6$ cells were collected from each bin. After the sorting, a small sample of cells from each bin was re-run to check for purity.

The cell sorter FACS Aria IIIu (Becton Dickinson, BD Biosciences) was used also to analyze the EGFP/mCherry ratio of the target Kozak sequences and respective variants individually selected through the screening (3 days post transient transfection). Data were analyzed with FlowJo software (v. 10.7.1).

Deep sequencing

The library of Kozak sequences was deep sequenced before the high-throughput screening to check for proper representation of all the variants. After the cell sorting, the Kozak sequences from the 4 subpopulations were PCR amplified and deep sequenced. Briefly, genomic DNA was extracted from the subpopulations using the DNeasy Blood & Tissue Kit (Qiagen). The DNA from each population was loaded in PCR reactions with 400ng input each. A second PCR was performed for the ligation of standard Illumina adapters. The PCR products were purified with Ampure XP beads (Beckman Coulter), mixed in equimolar ratios, and sequenced with the Illumina MiSeq on an SR250 v2 flow cell (Illumina, San Diego, CA).

Reads were quality checked with FASTQC (v0.11.4) and filtered to retain only those sequences that were not shifted (starting codon: CCA/CNA). Furthermore, only reads with no mismatch in the translation starting codon (ATG/CTG) were retained. Following the extraction of the 11 nt Kozak sequences from the reads, the occurrences of expected sequences were calculated and the unexpected sequences were removed.

Sequence read counts were converted in count per million (CPM) and filtered to retain only sequences with at least 5 CPM considering all four gates. Subsequently, the Gini index was calculated for each sequence and a cutoff of 0.25 was applied. Once a WT sequence was filtered out at each filtering step, the corresponding variants sequences were filtered out too.

Considering the WT sequences distribution as the expected ones and the variant sequences distribution as the observed ones, a Chi-square goodness of fit test was performed between each pair of WT and variant sequences. Benjamini-Hochberg (BH) multiple testing correction was further applied on all p-values, and a 0.01 cut-off was set on the adjusted p-values. Finally, the Expected Value (EV) was computed for each sequence and the EV of each WT sequence was subtracted from the EV of the respective variants sequences. Only sequences with positive values were retained as possible hits to be validated.

For the consensus sequences of the gates, the distribution of EVs of both variants and WT sequences was subdivided into quartiles, each representing one of four distribution-derived gates. Sequences for each respective gate were extracted and a consensus sequence representing the nucleotide frequency in each position of the Kozak sequence was generated using the seqLogo R package (v1.52).

High Content Screening System (Operetta)

Microplates with seeded and transfected cells (Corning 96-well plate) were imaged on the High Content Screening System Operetta™ (PerkinElmer). In each well, images were acquired in 9 preselected fields with LWD 10x objective over four channels: brightfield, digital phase contrast (DPC) based on brightfield images, fluorescence with excitation filters 460-490 and 520-550 nm, and emission filters 500-550 and 560-630 nm for GFP and mCherry reporters, respectively. For the feature extraction, the images were analyzed by Harmony software version 4.1 (PerkinElmer). Briefly, individual cell nuclei were segmented in the DPC channel. Nucleus morphology, GFP, and mCherry mean intensity were quantified in the cell nuclei population. Cell reporter fluorescence intensity was corrected for background calculated per well in cell-free area using texture machine learning algorithm in brightfield imaging. Single-cell object features were extracted from each sample well. To discriminate between GFP/mCherry negative and positive cells, for

each experiment a threshold was determined on the basis of GFP/mCherry intensity frequency distribution of all samples.

Western blot

For protein extraction, cells were homogenized in RIPA buffer with a complete protease inhibitor (PI) cocktail (Roche) and quantified with a BCA (bicinchoninic acid) assay. For protein extraction from the single sucrose fractions, 10% Trichloroacetic acid (TCA) was added and mixed. After incubation overnight at -20°C, samples were centrifuged at 14.000rpm for 10 min at 4°C. Samples were washed 3 times with 1ml ice cold Acetone and centrifuged at 14.000 rpm for 5 min. Pellets were resuspended in 30ul of RIPA buffer + PI.

Protein lysates were resolved on SDS-PAGE and transferred to PVDF membrane. Membrane blocking was carried out with 5% milk (BioRad)-TBS-T for one hour. Incubation with the primary antibodies was performed overnight at 4°C. Incubation with the secondary antibodies was performed for 1 hour at room temperature. The following antibodies were used: mouse anti-EGFP (sc-9996, Santa Cruz); rabbit anti-Cherry (PA5-34974, Thermo Fisher Scientific); mouse anti-beta tubulin (sc-53140, Santa Cruz); mouse anti-alpha-actinin (sc-17829, Santa Cruz); goat anti-$p47^{phox}$ (PA1-9073, Thermo Fisher Scientific); rabbit anti-RPS6 (5G10, Cell Signaling); rabbit anti-RPL26 (ab59567, Abcam); secondary anti-mouse IgG HRP (sc-2005, Santa Cruz); secondary anti-rabbit IgG HRP (#31460, Thermo Fisher Scientific); secondary anti-goat IgG HRP (ab97100, Abcam). Blots were imaged with the Uvitec Alliance Mini imaging system (UVITEC, Cambridge, UK) after incubation with ECL Prime or Select detection reagent (GE Healthcare, Buckinghamshire, UK). The intensity of the bands was quantified by densitometry using the Image J analysis program.

Polysome profiling

Polysomal profiling was performed according to previously described protocols[219]. Briefly, the cells were treated with cycloheximide and then lysed in 700uL of cold lysis buffer. The lysate was centrifuged at 13.000g for 10min at 4°C to pellet cell debris and loaded on a linear 15%–50% [w/v] sucrose gradient. The lysate was then centrifuged in an SW41Ti rotor (Beckman) at 40.000rpm for 1 h 40 min at 4°C in a Beckman Optima Optima XPN-100 Ultracentrifuge. Fractions of 1 mL of volume were then collected monitoring the absorbance at 254 nm with the UA-6 UV/VIS detector (Teledyne Isco).

Total RNA extraction

Raji cells were pelleted and lysed in 1ml of Trizol (Thermo Fisher) per 5×10^6 cells. For polysomal RNA extraction, sucrose fractions corresponding to polysomes, and total RNA were pooled together and lysed in 1ml of Trizol (Thermo Fisher).

Chloroform was added corresponding to ⅕ of the total volume after 15min incubation at RT. Samples were centrifuged at 12.000g for 15min at 4°C. The formed aqueous phase was transferred to a new tube and 1 ml of isopropanol was added. After 1h incubation at -80°C, samples were centrifuged at 12.000g for 10 min at 4°C, the supernatant was removed, and pellets were washed with 1 ml of 70% ethanol. Finally, samples were centrifuged at 5000g for 10 min at 4°C, the supernatant was removed, and the pellet was air-dry for 5-10 min before being dissolved in 20 µl DEPC-treated water.

Quantitative real-time PCR

For *NCF1* qPCR, 1 µg of total RNA was reverse transcribed using the RevertAid RT kit (ThermoFisher, K1619) following manufacturers' instructions. Sybr-green qPCR was performed as follows: 20ng template cDNA, ExcelTaq™ 1X Q-PCR Master Mix (SYBR, NO ROX, SMOBIO), and 0,4µM of each primer in a reaction volume of 15 µl. The qPCR reaction was performed on a CFX96 real-time PCR Detection System (Bio-Rad Laboratories) with the following cycling conditions: 95°C for 2min, followed by 40 cycles at 95°C for 15 sec, 60°C for 60 sec. HPRT1 expression was used as a reference. The ΔΔCq method was used to calculate the relative mRNA levels of each gene.

For polysome fractionation analysis, the gene-specific Translation Efficiency (TE) was calculated as the ratio between the fold change at the polysomal level and the fold change at the total level of the gene of interest as described before[219].

Statistical Analyses

For high Content Screening System (Operetta) analysis, the violin plots report the distribution of the data from three biological replicates. The dashed line indicates the median of the population. For western blot quantification and base editing efficiency analysis, the data were normalized over the wild type of each respective gene and are reported as mean ± SD (standard deviation) of at least three biological replicates, as

indicated in the figure legends. Statistical significance was determined by unpaired two-tailed t-test, (comparing each variant to the corresponding wild type), as indicated in the figure legends ($*p < 0.05$, $**p < 0.01$, $***p < 0.001$, $****p < 0.0001$).

BIBLIOGRAPHY

1. Jackson RJ, Hellen CUT, Pestova TV. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol*. 2010;11(2):113-127.

2. Hershey JWB, Sonenberg N, Mathews MB. Principles of Translational Control. *Cold Spring Harb Perspect Biol*. 2019;11(9). doi:10.1101/cshperspect.a032607

3. Hinnebusch AG. The scanning mechanism of eukaryotic translation initiation. *Annu Rev Biochem*. 2014;83:779-812.

4. Das S, Maitra U. Functional significance and mechanism of eIF5-promoted GTP hydrolysis in eukaryotic translation initiation. *Progress in Nucleic Acid Research and Molecular Biology*. Published online 2001:207-231. doi:10.1016/s0079-6603(01)70018-9

5. Kozak M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*. 2005;361:13-37. doi:10.1016/j.gene.2005.06.037

6. Kozak M. How do eucaryotic ribosomes select initiation regions in messenger RNA? *Cell*. 1978;15(4):1109-1123.

7. Kapp LD, Lorsch JR. The molecular mechanics of eukaryotic translation. *Annu Rev Biochem*. 2004;73:657-704.

8. Maag D, Fekete CA, Gryczynski Z, Lorsch JR. A Conformational Change in the Eukaryotic Translation Preinitiation Complex and Release of eIF1 Signal Recognition of the Start Codon. *Molecular Cell*. 2005;17(2):265-275. doi:10.1016/j.molcel.2004.11.051

9. Ambrosini C, Garilli F, Quattrone A. Reprogramming translation for gene therapy. *Prog Mol Biol Transl Sci*. 2021;182:439-476.

10. Schuller AP, Green R. Roadblocks and resolutions in eukaryotic translation. *Nat Rev Mol Cell Biol*. 2018;19(8):526-541.

11. Dever TE, Green R. The elongation, termination, and recycling phases of translation in eukaryotes. *Cold Spring Harb Perspect Biol*. 2012;4(7):a013706.

12. Dever TE, Dinman JD, Green R. Translation Elongation and Recoding in Eukaryotes. *Cold Spring Harb Perspect Biol*. 2018;10(8). doi:10.1101/cshperspect.a032649

13. Rodnina MV, Wintermeyer W. Protein Elongation, Co-translational Folding and Targeting. *Journal of Molecular Biology*. 2016;428(10):2165-2185. doi:10.1016/j.jmb.2016.03.022

14. Jackson RJ, Hellen CUT, Pestova TV. Termination and post-termination events in eukaryotic translation. *Fidelity and Quality Control in Gene Expression*. Published online 2012:45-93. doi:10.1016/b978-0-12-386497-0.00002-5

15. Schwanhäusser B, Busse D, Li N, et al. Global quantification of mammalian gene expression control. *Nature*. 2011;473(7347):337-342.

16. Sonenberg N, Hinnebusch AG. Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. *Cell*. 2009;136(4):731-745. doi:10.1016/j.cell.2009.01.042

17. Hernández G, Osnaya VG, Pérez-Martínez X. Conservation and Variability of the AUG Initiation Codon Context in Eukaryotes. *Trends Biochem Sci*. 2019;44(12):1009-1021.

18. Macdonald P. Diversity in translational regulation. *Current Opinion in Cell Biology*. 2001;13(3):326-331. doi:10.1016/s0955-0674(00)00215-5

19. Origins and evolution of the mechanisms regulating translation initiation in eukaryotes. *Trends Biochem Sci*. 2010;35(2):63-73.

20. Pakos-Zebrucka K, Koryga I, Mnich K, Ljujic M, Samali A, Gorman AM. The integrated stress response. *EMBO reports*. 2016;17(10):1374-1395. doi:10.15252/embr.201642195

21. Harding HP, Zhang Y, Zeng H, et al. An integrated stress response regulates amino acid metabolism and resistance to oxidative stress. *Mol Cell*. 2003;11(3):619-633.

22. Merrick WC, Pavitt GD. Protein Synthesis Initiation in Eukaryotic Cells. *Cold Spring Harb Perspect Biol*. 2018;10(12). doi:10.1101/cshperspect.a033092

23. Wek RC. Role of eIF2α Kinases in Translational Control and Adaptation to Cellular Stress. *Cold Spring Harbor Perspectives in Biology*. 2018;10(7):a032870. doi:10.1101/cshperspect.a032870

24. So L, Lee J, Palafox M, et al. The 4E-BP-eIF4E axis promotes rapamycin-sensitive growth and proliferation in lymphocytes. *Sci Signal*. 2016;9(430):ra57.

25. Korets SB, Czok S, Blank SV, Curtin JP, Schneider RJ. Targeting the mTOR/4E-BP pathway in endometrial cancer. *Clin Cancer Res*. 2011;17(24):7518-7528.

26. Gebauer F, Preiss T, Hentze MW. From Cis-Regulatory Elements to Complex RNPs and Back. *Cold Spring Harb Perspect Biol*. 2012;4(7):a012245.

27. Silva J, Fernandes R, Romão L. Translational Regulation by Upstream Open Reading Frames and Human Diseases. *The mRNA Metabolism in Human Disease*. Published online 2019:99-116. doi:10.1007/978-3-030-19966-1_5

28. Chen H-H, Tarn W-Y. uORF-mediated translational control: recently elucidated mechanisms and implications in cancer. *RNA Biol*. 2019;16(10):1327-1338.

29. Jang SK, Davies MV, Kaufman RJ, Wimmer E. Initiation of protein synthesis by internal entry of ribosomes into the 5' nontranslated region of encephalomyocarditis virus RNA in vivo. *J Virol*. 1989;63(4):1651-1660.

30. Macejak DG, Sarnow P. Internal initiation of translation mediated by the 5' leader of

a cellular mRNA. *Nature*. 1991;353(6339):90-94.

31. Komar AA, Hatzoglou M. Cellular IRES-mediated translation: the war of ITAFs in pathophysiological states. *Cell Cycle*. 2011;10(2):229-240.

32. Nakagawa S, Niimura Y, Gojobori T, Tanaka H, Miura K-I. Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Research*. 2007;36(3):861-871. doi:10.1093/nar/gkm1102

33. Kozak M. Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. *Nucleic Acids Research*. 1981;9(20):5233-5252. doi:10.1093/nar/9.20.5233

34. Kozak M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell*. 1986;44(2):283-292.

35. Wolf A, Caliebe A, Thomas NST, et al. Single base-pair substitutions at the translation initiation sites of human genes as a cause of inherited disease. *Human Mutation*. 2011;32(10):1137-1143. doi:10.1002/humu.21547

36. Roos D, de Boer M. Mutations in cis that affect mRNA synthesis, processing and translation. *Biochim Biophys Acta Mol Basis Dis*. 2021;1867(9):166166.

37. Jacobson EM, Concepcion E, Oashi T, Tomer Y. A Graves' disease-associated Kozak sequence single-nucleotide polymorphism enhances the efficiency of CD40 gene translation: a case for translational pathophysiology. *Endocrinology*. 2005;146(6):2684-2691.

38. Sultan CS, Weitnauer M, Turinsky M, et al. Functional association of a CD40 gene single-nucleotide polymorphism with the pathogenesis of coronary heart disease. *Cardiovasc Res*. 2020;116(6):1214-1225.

39. Baird TD, Palam LR, Fusakio ME, et al. Selective mRNA translation during eIF2 phosphorylation induces expression of IBTKα. *Mol Biol Cell*. 2014;25(10):1686-1697.

40. Lee Y-Y, Cevallos RC, Jan E. An upstream open reading frame regulates translation of GADD34 during cellular stresses that induce eIF2alpha phosphorylation. *J Biol Chem*. 2009;284(11):6661-6673.

41. Kozak M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Research*. 1987;15(20):8125-8148. doi:10.1093/nar/15.20.8125

42. Pisarev AV, Kolupaeva VG, Pisareva VP, Merrick WC, Hellen CUT, Pestova TV. Specific functional interactions of nucleotides at key -3 and +4 positions flanking the initiation codon with components of the mammalian 48S translation initiation complex. *Genes Dev*. 2006;20(5):624-636.

43. Simonetti A, Guca E, Bochler A, Kuhn L, Hashem Y. Structural Insights into the Mammalian Late-Stage Initiation Complexes. *Cell Rep*. 2020;31(1):107497.

44. Ivanov IP, Loughran G, Sachs MS, Atkins JF. Initiation context modulates autoregulation of eukaryotic translation initiation factor 1 (eIF1). *Proc Natl Acad Sci U S A*. 2010;107(42):18056-18060.

45. Kozak M. Recognition of AUG and alternative initiator codons is augmented by G in position 4 but is not generally affected by the nucleotides in positions 5 and 6. *The EMBO Journal*. 1997;16(9):2482-2492. doi:10.1093/emboj/16.9.2482

46. Xia X. The 4G Site in Kozak Consensus Is Not Related to the Efficiency of Translation Initiation. *PLoS ONE*. 2007;2(2):e188. doi:10.1371/journal.pone.0000188

47. Boissel JP, Kasper TJ, Bunn HF. Cotranslational amino-terminal processing of cytosolic proteins. Cell-free expression of site-directed mutants of human hemoglobin. *J Biol Chem*. 1988;263(17). Accessed August 10, 2021. https://pubmed.ncbi.nlm.nih.gov/3372535/

48. Niimura Y. Comparative analysis of the base biases at the gene terminal portions in seven eukaryote genomes. *Nucleic Acids Research*. 2003;31(17):5195-5201. doi:10.1093/nar/gkg701

49. Volkova OA, Kochetov AV. Interrelations between the nucleotide context of human start AUG codon, N-end amino acids of the encoded protein and initiation of translation. *J Biomol Struct Dyn*. 2010;27(5). doi:10.1080/07391102.2010.10508575

50. Ingolia NT. Ribosome profiling: new views of translation, from single codons to genome scale. *Nature Reviews Genetics*. 2014;15(3):205-213. doi:10.1038/nrg3645

51. Kearse MG, Wilusz JE. Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes Dev*. 2017;31(17):1717-1731.

52. Diaz de Arce AJ, Noderer WL, Wang CL. Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons. *Nucleic Acids Res*. 2018;46(2):985-994.

53. Kinney JB, Murugan A, Callan CG Jr, Cox EC. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci U S A*. 2010;107(20):9158-9163.

54. Kosuri S, Goodman DB, Cambray G, et al. Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc Natl Acad Sci U S A*. 2013;110(34):14024-14029.

55. Sharon E, Kalma Y, Sharp A, et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol*. 2012;30(6):521-530.

56. Lubliner S, Regev I, Lotan-Pompan M, Edelheit S, Weinberger A, Segal E. Core promoter sequence in yeast is a major determinant of expression level. *Genome Res*. 2015;25(7):1008-1017.

57. Oikonomou P, Goodarzi H, Tavazoie S. Systematic Identification of Regulatory

Elements in Conserved 3′ UTRs of Human Transcripts. *Cell Reports*. 2014;7(1):281-292. doi:10.1016/j.celrep.2014.03.001

58. Shalem O, Sharon E, Lubliner S, et al. Systematic Dissection of the Sequence Determinants of Gene 3' End Mediated Expression Control. *PLOS Genetics*. 2015;11(4):e1005147. doi:10.1371/journal.pgen.1005147

59. Dvir S, Velten L, Sharon E, et al. Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. *Proc Natl Acad Sci U S A*. 2013;110(30):E2792-E2801.

60. Noderer WL, Flockhart RJ, Bhaduri A, et al. Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Mol Syst Biol*. 2014;10:748.

61. Cuperus JT, Groves B, Kuchina A, et al. Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. *Genome Research*. 2017;27(12):2015-2024. doi:10.1101/gr.224964.117

62. Decoene T, Peters G, De Maeseneire SL, De Mey M. Toward Predictable 5′UTRs in Saccharomyces cerevisiae: Development of a yUTR Calculator. *ACS Synthetic Biology*. 2018;7(2):622-634. doi:10.1021/acssynbio.7b00366

63. Karollus A, Avsec Ž, Gagneur J. Predicting mean ribosome load for 5'UTR of any length using deep learning. *PLoS Comput Biol*. 2021;17(5). doi:10.1371/journal.pcbi.1008982

64. Wang X, Liu Z, Li G, et al. Efficient Gene Silencing by Adenine Base Editor-Mediated Start Codon Mutation. *Mol Ther*. 2020;28(2):431-440.

65. Chen S, Xie W, Liu Z, et al. CRISPR Start-Loss: A Novel and Practical Alternative for Gene Silencing through Base-Editing-Induced Start Codon Mutations. *Mol Ther Nucleic Acids*. 2020;21:1062-1073.

66. Blanco N, Williams AJ, Tang D, et al. Tailoring translational strength using Kozak sequence variants improves bispecific antibody assembly and reduces product-related impurities in CHO cells. *Biotechnology and Bioengineering*. 2020;117(7):1946-1960. doi:10.1002/bit.27347

67. Torgerson T, Ochs H. Genetics of Primary Immune Deficiencies. *Stiehm's Immune Deficiencies*. Published online 2014:73-81. doi:10.1016/b978-0-12-405546-9.00003-0

68. Kacser H, Burns JA. THE MOLECULAR BASIS OF DOMINANCE. *Genetics*. 1981;97(3-4):639-666. doi:10.1093/genetics/97.3-4.639

69. Deutschbauer AM, Jaramillo DF, Proctor M, et al. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics*. 2005;169(4):1915-1925.

70. Johnson AF, Nguyen HT, Veitia RA. Causes and effects of haploinsufficiency. *Biological Reviews*. 2019;94(5):1774-1785. doi:10.1111/brv.12527

71. Read AP. Haploinsufficiency. *eLS*. Published online 2017:1-5. doi:10.1002/9780470015902.a0005489.pub2

72. Veitia RA. Exploring the etiology of haploinsufficiency. *Bioessays*. 2002;24(2):175-184.

73. Papp B, Pál C, Hurst LD. Dosage sensitivity and the evolution of gene families in yeast. *Nature*. 2003;424(6945):194-197. doi:10.1038/nature01771

74. Fang P, Schwartz ID, Johnson BD, et al. Familial short stature caused by haploinsufficiency of the insulin-like growth factor i receptor due to nonsense-mediated messenger ribonucleic acid decay. *J Clin Endocrinol Metab*. 2009;94(5):1740-1747.

75. Veitia RA, Caburet S, Birchler JA. Mechanisms of Mendelian dominance. *Clin Genet*. 2018;93(3):419-428.

76. Bhatia S, Bengani H, Fish M, et al. Disruption of Autoregulatory Feedback by a Mutation in a Remote, Ultraconserved PAX6 Enhancer Causes Aniridia. *The American Journal of Human Genetics*. 2013;93(6):1126-1134. doi:10.1016/j.ajhg.2013.10.028

77. Gregory-Evans CY, Wang X, Wasan KM, Zhao J, Metcalfe AL, Gregory-Evans K. Postnatal manipulation of Pax6 dosage reverses congenital tissue malformation defects. *J Clin Invest*. 2014;124(1):111-116.

78. Wang X, Gregory-Evans K, Wasan KM, Sivak O, Shan X, Gregory-Evans CY. Efficacy of Postnatal In Vivo Nonsense Suppression Therapy in a Pax6 Mouse Model of Aniridia. *Molecular Therapy - Nucleic Acids*. 2017;7:417-428. doi:10.1016/j.omtn.2017.05.002

79. Dang VT, Kassahn KS, Marcos AE, Ragan MA. Identification of human haploinsufficient genes and their genomic proximity to segmental duplications. *Eur J Hum Genet*. 2008;16(11):1350-1357.

80. Steinberg J, Honti F, Meader S, Webber C. Haploinsufficiency predictions without study bias. *Nucleic Acids Res*. 2015;43(15):e101.

81. Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet*. 2010;6(10):e1001154.

82. Shihab HA, Rogers MF, Campbell C, Gaunt TR. HIPred: an integrative approach to predicting haploinsufficient genes. *Bioinformatics*. 2017;33(12):1751-1757.

83. Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *J Bacteriol*. 1987;169(12):5429-5433.

84. Groenen PM, Bunschoten AE, van Soolingen D, van Embden JD. Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method. *Mol Microbiol*. 1993;10(5):1057-1065.

85. Hoe N, Nakashima K, Grigsby D, et al. Rapid molecular genetic subtyping of serotype M1 group A Streptococcus strains. *Emerg Infect Dis*. 1999;5(2):254-263.

86. Jansen R, van Embden JDA, Gaastra W, Schouls LM. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol*. 2002;43(6):1565-1575.

87. Makarova KS, Aravind L, Grishin NV, Rogozin IB, Koonin EV. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res*. 2002;30(2). doi:10.1093/nar/30.2.482

88. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Soria E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol*. 2005;60(2):174-182.

89. Pourcel C, Salvignol G, Vergnaud G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology*. 2005;151(Pt 3):653-663.

90. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*. 2005;151(Pt 8):2551-2561.

91. Barrangou R, Fremaux C, Deveau H, et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*. 2007;315(5819):1709-1712.

92. Oost J van der, van der Oost J, Westra ER, Jackson RN, Wiedenheft B. Unravelling the structural and mechanistic basis of CRISPR–Cas systems. *Nature Reviews Microbiology*. 2014;12(7):479-492. doi:10.1038/nrmicro3279

93. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*. 2009;155(Pt 3):733-740.

94. Brouns SJJ, Jore MM, Lundgren M, et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*. 2008;321(5891):960-964.

95. Barrangou R, Horvath P. A decade of discovery: CRISPR functions and applications. *Nature Microbiology*. 2017;2(7). doi:10.1038/nmicrobiol.2017.92

96. Makarova KS, Wolf YI, Iranzo J, et al. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nature Reviews Microbiology*. 2020;18(2):67-83. doi:10.1038/s41579-019-0299-x

97. Garneau JE, Dupuis M-È, Villion M, et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*. 2010;468(7320):67-71.

98. Deltcheva E, Chylinski K, Sharma CM, et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*. 2011;471(7340):602-607.

99. Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc Natl Acad Sci U S A*. 2012;109(39):E2579-E2586.

100. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*. 2012;337(6096):816-821.

101. Kleinstiver BP, Prew MS, Tsai SQ, et al. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature*. 2015;523(7561):481-485.

102. Casini A, Olivieri M, Petris G, et al. A highly specific SpCas9 variant is identified by in vivo screening in yeast. *Nat Biotechnol*. 2018;36(3):265-271.

103. Kleinstiver BP, Pattanayak V, Prew MS, et al. High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature*. 2016;529(7587):490-495.

104. Petris G, Casini A, Montagna C, et al. Hit and go CAS9 delivered through a lentiviral based self-limiting circuit. *Nat Commun*. 2017;8:15334.

105. Montagna C, Petris G, Casini A, et al. VSV-G-Enveloped Vesicles for Traceless Delivery of CRISPR-Cas9. *Molecular Therapy - Nucleic Acids*. 2018;12:453-462. doi:10.1016/j.omtn.2018.05.010

106. Wang F, Qi LS. Applications of CRISPR Genome Engineering in Cell Biology. *Trends in Cell Biology*. 2016;26(11):875-888. doi:10.1016/j.tcb.2016.08.004

107. Doudna JA, Charpentier E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science*. 2014;346(6213):1258096.

108. Yeh CD, Richardson CD, Corn JE. Advances in genome editing through control of DNA repair pathways. *Nat Cell Biol*. 2019;21(12):1468-1478.

109. Sander JD, Joung JK. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotechnol*. 2014;32(4):347-355.

110. Larson MH, Gilbert LA, Wang X, Lim WA, Weissman JS, Qi LS. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat Protoc*. 2013;8(11):2180-2196.

111. Gilbert LA, Horlbeck MA, Adamson B, et al. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell*. 2014;159(3):647-661. doi:10.1016/j.cell.2014.09.029

112. Evers B, Jastrzebski K, Heijmans JPM, Grernrum W, Beijersbergen RL, Bernards R. CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. *Nat Biotechnol*. 2016;34(6):631-633.

113. Anzalone AV, Randolph PB, Davis JR, et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*. 2019;576(7785):149-157.

114. Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*. 2016;533(7603):420-424.

115. Rees HA, Liu DR. Base editing: precision chemistry on the genome and transcriptome of living cells. *Nat Rev Genet*. 2018;19(12):770-788.

116. Hess GT, Tycko J, Yao D, Bassik MC. Methods and Applications of CRISPR-Mediated Base Editing in Eukaryotic Genomes. *Mol Cell*. 2017;68(1):26-43.

117. Schatoff EM, Zafra MP, Dow LE. Base editing the mammalian genome. *Methods*. 2019;164-165:100-108.

118. Komor AC, Zhao KT, Packer MS, et al. Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C:G-to-T:A base editors with higher efficiency and product purity. *Science Advances*. 2017;3(8):eaao4774. doi:10.1126/sciadv.aao4774

119. Gaudelli NM, Komor AC, Rees HA, et al. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature*. 2017;551(7681):464-471.

120. Lau AY, Wyatt MD, Glassner BJ, Samson LD, Ellenberger T. Molecular basis for discriminating between normal and damaged bases by the human alkyladenine glycosylase, AAG. *Proc Natl Acad Sci U S A*. 2000;97(25):13573-13578.

121. Nishida K, Arazoe T, Yachie N, et al. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science*. 2016;353(6305). doi:10.1126/science.aaf8729

122. Anzalone AV, Koblan LW, Liu DR. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nature Biotechnology*. 2020;38(7):824-844. doi:10.1038/s41587-020-0561-9

123. Harris RS, Petersen-Mahrt SK, Neuberger MS. RNA Editing Enzyme APOBEC1 and Some of Its Homologs Can Act as DNA Mutators. *Molecular Cell*. 2002;10(5):1247-1253. doi:10.1016/s1097-2765(02)00742-6

124. Koblan LW, Doman JL, Wilson C, et al. Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat Biotechnol*. 2018;36(9):843-846.

125. Richter MF, Zhao KT, Eton E, et al. Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity. *Nat Biotechnol*. 2020;38(7):901.

126. Gaudelli NM, Lam DK, Rees HA, et al. Directed evolution of adenine base editors with increased activity and therapeutic application. *Nat Biotechnol*. 2020;38(7):892-900.

127. Grünewald J, Zhou R, Iyer S, et al. CRISPR DNA base editors with reduced RNA off-target and self-editing activities. *Nat Biotechnol*. 2019;37(9):1041-1048.

128. Wang L, Xue W, Yan L, et al. Enhanced base editing by co-expression of free uracil DNA glycosylase inhibitor. *Cell Res*. 2017;27(10):1289-1292.

129. Ma Y, Zhang J, Yin W, Zhang Z, Song Y, Chang X. Targeted AID-mediated

mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. *Nat Methods*. 2016;13(12):1029-1035.

130. Hess GT, Frésard L, Han K, et al. Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. *Nat Methods*. 2016;13(12):1036-1042.

131. Kim YB, Komor AC, Levy JM, Packer MS, Zhao KT, Liu DR. Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. *Nat Biotechnol*. 2017;35(4):371-376.

132. Rees HA, Komor AC, Yeh W-H, et al. Improving the DNA specificity and applicability of base editing through protein engineering and protein delivery. *Nat Commun*. 2017;8:15790.

133. Lee JK, Jeong E, Lee J, et al. Directed evolution of CRISPR-Cas9 to increase its specificity. *Nat Commun*. 2018;9(1):3048.

134. Liang P, Sun H, Sun Y, et al. Effective gene editing by high-fidelity base editor 2 in mouse zygotes. *Protein Cell*. 2017;8(8):601-611.

135. Kim D, Lim K, Kim S-T, et al. Genome-wide target specificities of CRISPR RNA-guided programmable deaminases. *Nat Biotechnol*. 2017;35(5):475-480.

136. Liang P, Xie X, Zhi S, et al. Genome-wide profiling of adenine base editor specificity by EndoV-seq. *Nat Commun*. 2019;10(1):67.

137. Doman JL, Raguram A, Newby GA, Liu DR. Evaluation and minimization of Cas9-independent off-target DNA editing by cytosine base editors. *Nat Biotechnol*. 2020;38(5):620-628.

138. Yu Y, Leete TC, Born DA, et al. Cytosine base editors with minimized unguided DNA and RNA off-target events and high on-target activity. *Nat Commun*. 2020;11(1):2052.

139. Zhou C, Sun Y, Yan R, et al. Off-target RNA mutation induced by DNA base editing and its elimination by mutagenesis. *Nature*. 2019;571(7764):275-278.

140. Grünewald J, Zhou R, Garcia SP, et al. Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors. *Nature*. 2019;569(7756):433-437.

141. Rees HA, Wilson C, Doman JL, Liu DR. Analysis and minimization of cellular RNA editing by DNA adenine base editors. *Sci Adv*. 2019;5(5):eaax5717.

142. Miller SM, Wang T, Randolph PB, et al. Continuous evolution of SpCas9 variants compatible with non-G PAMs. *Nat Biotechnol*. 2020;38(4):471-481.

143. Li X, Wang Y, Liu Y, et al. Base editing with a Cpf1–cytidine deaminase fusion. *Nature Biotechnology*. 2018;36(4):324-327. doi:10.1038/nbt.4102

144. Walton RT, Christie KA, Whittaker MN, Kleinstiver BP. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science*.

2020;368(6488):290-296.

145. Huang TP, Zhao KT, Miller SM, et al. Circularly permuted and PAM-modified Cas9 variants broaden the targeting scope of base editors. *Nat Biotechnol*. 2019;37(6):626-631.

146. Li C, Zhang R, Meng X, et al. Targeted, random mutagenesis of plant genes with dual cytosine and adenine base editors. *Nat Biotechnol*. 2020;38(7):875-882.

147. Sakata RC, Ishiguro S, Mori H, et al. Base editors for simultaneous introduction of C-to-T and A-to-G mutations. *Nature Biotechnology*. 2020;38(7):865-869. doi:10.1038/s41587-020-0509-0

148. Zhang X, Zhu B, Chen L, et al. Dual base editor catalyzes both cytosine and adenine base conversions in human cells. *Nat Biotechnol*. 2020;38(7):856-860.

149. Grünewald J, Zhou R, Lareau CA, et al. A dual-deaminase CRISPR base editor enables concurrent adenine and cytosine editing. *Nat Biotechnol*. 2020;38(7):861-864.

150. Kweon J, Jang A-H, Shin HR, et al. A CRISPR-based base-editing screen for the functional assessment of BRCA1 variants. *Oncogene*. 2020;39(1):30-35.

151. Jun S, Lim H, Chun H, Lee JH, Bang D. Single-cell analysis of a mutant library generated using CRISPR-guided deaminase in human melanoma cells. *Commun Biol*. 2020;3(1):154.

152. Hanna RE, Hegde M, Fagre CR, et al. Massively parallel assessment of human variants with base editor screens. *Cell*. 2021;184(4):1064-1080.e20.

153. Functional interrogation of DNA damage response variants with base editing screens. *Cell*. 2021;184(4):1081-1097.e19.

154. Koblan LW, Arbab M, Shen MW, et al. Efficient C•G-to-G•C base editors developed using CRISPRi screens, target-library analysis, and machine learning. *Nat Biotechnol*. Published online June 28, 2021. doi:10.1038/s41587-021-00938-z

155. Kurt IC, Zhou R, Iyer S, et al. CRISPR C-to-G base editors for inducing targeted DNA transversions in human cells. *Nat Biotechnol*. 2021;39(1):41-46.

156. Porto EM, Komor AC, Slaymaker IM, Yeo GW. Base editing: advances and therapeutic opportunities. *Nat Rev Drug Discov*. 2020;19(12):839-859.

157. Song C-Q, Jiang T, Richter M, et al. Adenine base editing in an adult mouse model of tyrosinaemia. *Nature Biomedical Engineering*. 2020;4(1):125-130. doi:10.1038/s41551-019-0357-8

158. Musunuru K, Chadwick AC, Mizoguchi T, et al. In vivo CRISPR base editing of PCSK9 durably lowers cholesterol in primates. *Nature*. 2021;593(7859):429-434.

159. Newby GA, Yen JS, Woodard KJ, et al. Base editing of haematopoietic stem cells rescues sickle cell disease in mice. *Nature*. 2021;595(7866):295-302.

160. Tong S, Moyo B, Lee CM, Leong K, Bao G. Engineered materials for in vivo delivery of genome-editing machinery. *Nat Rev Mater*. 2019;4:726-737.

161. Robert MA, Chahal PS, Audy A, Kamen A, Gilbert R, Gaillet B. Manufacturing of recombinant adeno-associated viruses using mammalian expression platforms. *Biotechnol J*. 2017;12(3). doi:10.1002/biot.201600193

162. Lim CKW, Gapinske M, Brooks AK, et al. Treatment of a Mouse Model of ALS by In Vivo Base Editing. *Mol Ther*. 2020;28(4):1177-1189.

163. Villiger L, Grisch-Chan HM, Lindsay H, et al. Treatment of a metabolic liver disease by in vivo genome base editing in adult mice. *Nat Med*. 2018;24(10):1519-1525.

164. Levy JM, Yeh W-H, Pendse N, et al. Cytosine and adenine base editing of the brain, liver, retina, heart and skeletal muscle of mice via adeno-associated viruses. *Nat Biomed Eng*. 2020;4(1):97-110.

165. Yeh W-H, Chiang H, Rees HA, Edge ASB, Liu DR. In vivo base editing of post-mitotic sensory cells. *Nat Commun*. 2018;9(1):2184.

166. Ferreira JP, Overton KW, Wang CL. Tuning gene expression with synthetic upstream open reading frames. *Proc Natl Acad Sci U S A*. 2013;110(28):11284-11289.

167. Han X, Chen S, Flynn E, Wu S, Wintner D, Shen Y. Distinct epigenomic patterns are associated with haploinsufficiency and predict risk genes of developmental disorders. *Nat Commun*. 2018;9(1):2138.

168. Kozak M. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J Mol Biol*. 1987;196(4). doi:10.1016/0022-2836(87)90418-9

169. Blanco N, Williams AJ, Tang D, et al. Tailoring translational strength using Kozak sequence variants improves bispecific antibody assembly and reduces product-related impurities in CHO cells. *Biotechnol Bioeng*. 2020;117(7):1946-1960.

170. Benitez-Cantos MS, Yordanova MM, O'Connor PBF, et al. Translation initiation downstream from annotated start codons in human mRNAs coevolves with the Kozak context. *Genome Res*. 2020;30(7):974-984.

171. Kozak M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*. 2005;361. doi:10.1016/j.gene.2005.06.037

172. Lin J, Puigserver P, Donovan J, Tarr P, Spiegelman BM. Peroxisome proliferator-activated receptor gamma coactivator 1beta (PGC-1beta), a novel PGC-1-related transcription coactivator associated with host cell factor. *J Biol Chem*. 2002;277(3):1645-1648.

173. Christian Riehle EDA. PGC-1 Proteins and Heart Failure. *Trends Cardiovasc Med*. 2012;22(4):98.

174. St-Pierre J, Lin J, Krauss S, et al. Bioenergetic analysis of peroxisome proliferator-activated receptor gamma coactivators 1alpha and 1beta (PGC-1alpha and PGC-1beta) in muscle cells. *J Biol Chem*. 2003;278(29):26597-26603.

175. Hoeks J, Hesselink MK, Russell AP, et al. Peroxisome proliferator-activated receptor-gamma coactivator-1 and insulin resistance: acute effect of fatty acids. *Diabetologia*. 2006;49(10). doi:10.1007/s00125-006-0369-2

176. Patti ME, Butte AJ, Crunkhorn S, et al. Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of PGC1 and NRF1. *Proc Natl Acad Sci U S A*. 2003;100(14):8466.

177. Andersen G. Evidence of an association between genetic variation of the coactivator PGC-1 and obesity. *Journal of Medical Genetics*. 2005;42(5):402-407. doi:10.1136/jmg.2004.026278

178. Meng X, Lu X, Morris CA, Keating MT. A Novel Human GeneFKBP6Is Deleted in Williams Syndrome. *Genomics*. 1998;52(2):130-137. doi:10.1006/geno.1998.5412

179. Metcalfe K, Simeonov E, Beckett W, Donnai D, Tassabehji M. Autosomal dominant inheritance of Williams???Beuren syndrome in a father and son with haploinsufficiency for FKBP6. *Clinical Dysmorphology*. 2005;14(2):61-65. doi:10.1097/00019605-200504000-00002

180. Webling KEB, Runesson J, Bartfai T, Langel Ü. Galanin Receptors and Ligands. *Frontiers in Endocrinology*. 2012;3. doi:10.3389/fendo.2012.00146

181. Cody JD, Ghidoni PD, DuPont BR, et al. Congenital anomalies and anthropometry of 42 individuals with deletions of chromosome 18q. *Am J Med Genet*. 1999;85(5). doi:10.1002/(sici)1096-8628(19990827)85:5<455::aid-ajmg5>3.0.co;2-z

182. Dostal A, Nemeckova J, Gaillyova R, et al. Identification of 2.3-Mb gene locus for congenital aural atresia in 18q22.3 deletion: a case report analyzed by comparative genomic hybridization. *Otol Neurotol*. 2006;27(3):427-432.

183. Veltman JA, Jonkers Y, Nuijten I, et al. Definition of a critical region on chromosome 18 for congenital aural atresia by arrayCGH. *Am J Hum Genet*. 2003;72(6):1578-1584.

184. Margarit E, Morales C, Rodríguez-Revenga L, et al. Familial 4.8 MB deletion on 18q23 associated with growth hormone insufficiency and phenotypic variability. *Am J Med Genet A*. 2012;158A(3):611-616.

185. Castronovo P, Baccarin M, Ricciardello A, et al. Phenotypic spectrum of NRXN1 mono- and bi-allelic deficiency: A systematic review. *Clin Genet*. 2020;97(1):125-137.

186. Ishizuka K, Yoshida T, Kawabata T, et al. Functional characterization of rare NRXN1 variants identified in autism spectrum disorders and schizophrenia. *J Neurodev Disord*. 2020;12(1):25.

187. Levinson DF, Shi J, Wang K, et al. Genome-wide association study of multiplex schizophrenia pedigrees. *Am J Psychiatry*. 2012;169(9):963-973.

188. Tromp A, Mowry B, Giacomotto J. Neurexins in autism and schizophrenia—a review of patient mutations, mouse models and potential future directions. *Molecular Psychiatry*. 2021;26(3):747-760. doi:10.1038/s41380-020-00944-8

189. Béna F, Bruno DL, Eriksson M, et al. Molecular and clinical characterization of 25 individuals with exonic deletions of NRXN1 and comprehensive review of the literature. *Am J Med Genet B Neuropsychiatr Genet*. 2013;162B(4):388-403.

190. Calahorro F, Ruiz-Rubio M. Human alpha- and beta-NRXN1 isoforms rescue behavioral impairments of Caenorhabditis elegans neurexin-deficient mutants. *Genes Brain Behav*. 2013;12(4):453-464.

191. Zahir FR, Baross A, Delaney AD, et al. A patient with vertebral, cognitive and behavioural abnormalities and a de novo deletion of NRXN1. *Journal of Medical Genetics*. 2007;45(4):239-243. doi:10.1136/jmg.2007.054437

192. Duong L, Klitten LL, Møller RS, et al. Mutations in NRXN1 in a family multiply affected with brain disorders: NRXN1 mutations and brain disorders. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*. 2012;159B(3):354-358. doi:10.1002/ajmg.b.32036

193. El-Benna J, Dang PM-C, Gougerot-Pocidalo M-A, Marie J-C, Braut-Boucher F. p47phox, the phagocyte NADPH oxidase/NOX2 organizer: structure, phosphorylation and implication in diseases. *Exp Mol Med*. 2009;41(4):217-225.

194. Bedard K, Krause KH. The NOX family of ROS-generating NADPH oxidases: physiology and pathophysiology. *Physiol Rev*. 2007;87(1). doi:10.1152/physrev.00044.2005

195. Roos D. Chronic granulomatous disease. *Br Med Bull*. 2016;118(1):50.

196. Casimir CM, Bu-Ghanim HN, Rodaway AR, Bentley DL, Rowe P, Segal AW. Autosomal recessive chronic granulomatous disease caused by deletion at a dinucleotide repeat. *Proc Natl Acad Sci U S A*. 1991;88(7):2753-2757.

197. Roos D, de Boer M, Köker MY, et al. Chronic granulomatous disease caused by mutations other than the common GT deletion in NCF1, the gene encoding the p47phox component of the phagocyte NADPH oxidase. *Hum Mutat*. 2006;27(12). doi:10.1002/humu.20413

198. Mitchison NA, Harbord M, Hankin A, Roes J. Conditional haploinsufficiency of NCF1 (encoding p47phox), a signaling gene with a heterozygous phenotype potentially subject to natural selection. *Immunology Letters*. 2005;97(1):63-67. doi:10.1016/j.imlet.2004.09.015

199. Harbord M, Hankin A, Bloom S, Mitchison H. Association between p47phox pseudogenes and inflammatory bowel disease. *Blood*. 2003;101(8):3337-3337.

200. Heyworth PG, Cross AR, Curnutte JT. Chronic granulomatous disease. *Current*

Opinion in Immunology. 2003;15(5):578-584. doi:10.1016/s0952-7915(03)00109-2

201. Hsiao J, Yuan TY, Tsai MS, et al. Upregulation of Haploinsufficient Gene Expression in the Brain by Targeting a Long Non-coding RNA Improves Seizure Phenotype in a Model of Dravet Syndrome. *EBioMedicine*. 2016;9:257-277. doi:10.1016/j.ebiom.2016.05.011

202. Mali P, Aach J, Stranges PB, et al. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol*. 2013;31(9):833-838.

203. Colasante G, Lignani G, Brusco S, et al. dCas9-Based Scn1a Gene Activation Restores Inhibitory Interneuron Excitability and Attenuates Seizures in Dravet Syndrome Mice. *Mol Ther*. 2020;28(1):235-253.

204. Matharu N, Rattanasopha S, Tamura S, et al. CRISPR-mediated activation of a promoter or enhancer rescues obesity caused by haploinsufficiency. *Science*. 2019;363(6424). doi:10.1126/science.aau0629

205. Yang F, Moss LG, Phillips GN Jr. The molecular structure of green fluorescent protein. *Nat Biotechnol*. 1996;14(10):1246-1251.

206. Liu S-S, Wei X, Dong X, Xu L, Liu J, Jiang B. Structural plasticity of green fluorescent protein to amino acid deletions and fluorescence rescue by folding-enhancing mutations. *BMC Biochem*. 2015;16:17.

207. Corish P, Tyler-Smith C. Attenuation of green fluorescent protein half-life in mammalian cells. *Protein Eng*. 1999;12(12):1035-1040.

208. Barth-Baus D, Bhasker CR, Zoll W, Merrick WC. Influence of translation factor activities on start site selection in six different mRNAs. *Translation (Austin)*. 2013;1(1):e24419.

209. Fijałkowska D, Verbruggen S, Ndah E, Jonckheere V, Menschaert G, Van Damme P. eIF1 modulates the recognition of suboptimal translation initiation sites and steers gene expression via uORFs. *Nucleic Acids Research*. 2017;45(13):7997-8013. doi:10.1093/nar/gkx469

210. Varshavsky A. The N-end rule pathway and regulation by proteolysis. *Protein Sci*. 2011;20(8):1298.

211. Ahn B, Beharry AW, Frye GS, Judge AR, Ferreira LF. NAD(P)H oxidase subunit p47phox is elevated, and p47phox knockout prevents diaphragm contractile dysfunction in heart failure. *Am J Physiol Lung Cell Mol Physiol*. 2015;309(5):L497-L505.

212. Wang H, Chen X, Su Y, et al. p47phox contributes to albuminuria and kidney fibrosis in mice. *Kidney International*. 2015;87(5):948-962. doi:10.1038/ki.2014.386

213. Du M, Liu X, Welch EM, Hirawat S, Peltz SW, Bedwell DM. PTC124 is an orally bioavailable compound that promotes suppression of the human CFTR-G542X nonsense allele in a CF mouse model. *Proceedings of the National Academy of*

Sciences. 2008;105(6):2064-2069. doi:10.1073/pnas.0711795105

214. Kayali R, Ku J-M, Khitrov G, Jung ME, Prikhodko O, Bertoni C. Read-through compound 13 restores dystrophin expression and improves muscle function in the mdx mouse model for Duchenne muscular dystrophy. *Hum Mol Genet*. 2012;21(18):4007-4020.

215. Pallardó FV, Pagano G, Rodríguez LR, Gonzalez-Cabo P, Lyakhovich A, Trifuoggi M. Friedreich Ataxia: current state-of-the-art, and future prospects for mitochondrial-focused therapies. *Translational Research*. 2021;229:135-141. doi:10.1016/j.trsl.2020.08.009

216. Shalem O, Sanjana NE, Hartenian E, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*. 2014;343(6166):84-87.

217. Kluesner MG, Nedveck DA, Lahr WS, et al. EditR: A Method to Quantify Base Editing from Sanger Sequencing. *CRISPR J*. 2018;1:239-250.

218. Pizzato M, Erlwein O, Bonsall D, Kaye S, Muir D, McClure MO. A one-step SYBR Green I-based product-enhanced reverse transcriptase assay for the quantitation of retroviruses in cell culture supernatants. *J Virol Methods*. 2009;156(1-2). doi:10.1016/j.jviromet.2008.10.012

219. Tebaldi T, Zuccotti P, Peroni D, et al. HuD Is a Neural Translation Enhancer Acting on mTORC1-Responsive Genes and Counteracted by the Y3 Small Non-coding RNA. *Mol Cell*. 2018;71(2):256-270.e10.

APPENDIX

Table S1: List of the HI genes, respective diseases and relevant publications considered in the high-throughput screening. HP = HiPred score (0-0,88).

Gene	Disease	PMID
ADAR	Dyschromatosis symmetrica hereditaria	16536805
AHSP	Exacerbates beta-thalassemia	15545996
ALDH1A2	Facilitate posterior organ development and prevent spina bifida	11953746
ALX4	Parietal foramina (PFM)	11017806
ANK2	Arrhythmia	12571597
ANKRD11	KBG syndrome	29258554
ARID1B	Autism spectrum disorder and intellectual disability	22405089
ATP1A2	Familial hemiplegic migraine type 2	12539047
ATP2A2	Darier disease	32354065
ATP2C1	Hailey-Hailey disease (skin disorder)	23474827
ATR	Familial cutaneous telengiectasia and cancer syndrome	30159786
AUTS2	Neurodevelopmental disorders and dysmorphic features	33577136
BAG1	Impaired cartilage development and osteogenic differentiation	27633857
BAG3	Dilated cardiomyopathy (DCM)	21353195
BCL11B	Suppression of lymphomagenesis and thymocyte development	17306224
BMP4	Anophthalmia and pituitary gland hypoplasia	16835935
BMPR1A	Juvenile polyposis	23399955
BMPR2	Primary pulmonary hypertension	11115378
BUB3	Early onset of aging-related features	16600919
CD2AP	Glomerular disease susceptibility	12764198
CDC73	Ossifying fibroma of the jaws	16458039
CDKN2C	Predicted by HIPred score	HP 0,6955
CHD2	Developmental delay, intellectual disability, epilepsy	24834135
CHD5	Congenital and developmental anomalies	12592387
CHD7	CHARGE syndrome	33060836
CHRNA7	15q 13.3 deletion syndrome	24556925
COL1A1	Osteogenesis Imperfecta type 1	9067755
COL1A2	Recurrent joint subluxation or hypodontia	17898012
COL2A1	Stickler syndrome	10819645
COL3A1	Ehlers-Danlos syndrome type IV	21637106
COL5A1	Structural abnormalities of the cornea and lid	16431952
COL6A1	Bethlem myopathy	9580662
COMT	22q11.2 deletion syndrome	16848928
COPS3	Smith-Magenis syndrome	10851253
COPS5	Predicted by HIPred score	HP 0,7829
CREBBP	Rubinstein-Taybi syndrome (RSTS)	26603346
CRX	Photoreceptor degeneration, Cone-rod dystrophy	10892846
CYR61	Persistent ostium primum atrial septal defects	17023674
DLL4	Embryonic lethality due to major defects in vascular development	30116629

Gene	Disease	PMID
DMRT1	Ovotesticular Disorder of Sexual Development	29219112
DMRT2	Sex reversal	10857744
DSG1	Diseases of epidermal integrity	17194569
DSP	Woolly hair syndrome, Striate palmoplantar keratoderma	11841538
DYRK1A	Neurological defects, developmental delay	12192061
EDN3	Hirschsprung disease	19040714
EFTUD2	Mandibulofacial dysostosis and microcephaly (MFDM)	22305528
EGR1	Development of myeloid disorders	17420284
EHMT1	9q34 subtelomeric deletion syndrome	16826528
ELAVL4	Predicted by HIPred score	HP 0,7831
ELOVL4	Defective skin permeability barrier function and neonatal lethality	17311087
ENG	Hereditary hemorrhagic telangiectasia type 1	15718503
EXT1	Multiple Hereditary Exostoses (MHE) syndrome	21310272
EYA1	Branchio-oto-renal (BOR) syndrome	29552445
EYA4	Sensorineural hearing loss (SNHL)	17567890
FAS	Autoimmune Lymphoproliferative syndrome	21490157
FBLN1	Limb malformations	19922873
FBN1	Marfan syndrome	27437668
FECH	Protoporphyria	10068685
FEN1	Neuromuscular and neurodegenerative diseases	16978612
FGF10	Aplasia of lacrimal and salivar glands	15654336
FGF3	Otodental syndrome	17656375
FGF8	Several human craniofacial disorders	17448458
FGFR1	Combined pituitary hormone deficiency (CPHD), Hypogonadism	23657145
FKBP6	Williams-Beuren syndrome	15770126
FLI1	Paris-Trousseau thrombopenia	15525489
FOXC1	Axenfeld-Rieger anomaly of the anterior eye chamber	14564054
FOXC2	The lymphatic/ocular disorder Lymphedema-Distichiasis	16910099
FOXE3	Anterior segment dysgenesis similar to Peters' anomaly	11980846
FOXF1	Alveolar capillary dysplasia with misalignment of pulmonary veins	23505205
FOXG1	Impaired intellectual development	32158381
FOXL2	Blepharophimosis syndrome associated with ovarian dysfunction	11468277
FOXO1	Predicted by HIPred score	HP 0,8247
FOXP2	Speech and language impairment and oromotor dysprax	16470794
FSCN2	Autosomal dominant retinitis pigmentosa	16043865
FZD4	complex chromosome rearrangement with multiple abnormalities	17103440
GALR1	Congenital aural atresia phenotype in 18q deletion syndrome	16639285
GATA3	HDR (hypoparathyroidism, deafness and renal dysplasia) syndrome	17046739
GATA4	Congenital heart disease	10096597
GATA6	Right-Left Type Bicuspid Aortic Valve	29567669
GCH1	Malignant hyperphenylalaninemia	15241655
GCK	Maturity onset diabetes of the young	9570959
GCNT1	T lymphoma cells resistant to cell death	16778138
GDF5	Multiple-synostosis syndrome	16532400
GDNF	Hirschsprung-like intestinal obstruction and early-onset lethality	11774071

Gene	Disease	PMID
GHRL	Obesity	16204371
GLI2	Developmental anomalies	29988648
GLI3	Greig cephalopolysyndactyly and Pallister-Hall syndromes	15739154
GRIN2B	Mental retardation autosomal dominant 6	28377535
GTF2I	Williams-Beuren syndrome	29305905
GTF2IRD1	Williams Syndrome	29305905
H2AFX	Predicted by HIPred score	HP 0,7644
HIC1	Miller-Dieker syndrome	16724116
HIRA	DiGeorge syndrome	9063745
HMGA1	Predicted by HIPred score	HP 0,7112
HNF1A	Reduced serum apolipoprotein M levels	14633861
HNF1B	Maturity-onset diabetes of the young (MODY)	10720943
HNF4A	Autosomal dominant non-insulin-dependent diabetes type I	10905494
ID2	Congenital hydronephrosis	15569159
IGF1	Subtle inhibition of intrauterine and postnatal growth	15769976
IGF1R	Metabolic syndrome	28351623
IRF6	van der Woude syndrome and popliteal pterygium syndrome	
JAG1	Alagille syndrome	16575836
KANSL1	Koolen-de Vries syndrome	33361104
KCNAB2	Epilepsy in patients with 1p36 deletion syndrome	11580756
KCNH2	Long QT Syndrome	21951015
KCNQ2	Epilepsy susceptibility	12700166
KCNQ4	Nonsyndromic Progressive Sensorineural Hearing Loss	18797286
KHDRBS1	Predicted by HIPred score	HP 0,8020
KIF11	Microcephaly, chorioretinopathy	31428438
KLHL10	Infertility	15136734
KMT2D	Kabuki syndrome	31814321
KRT5	Dowling-Degos disease (DDD)	16465624
LEMD3	Osteopoikilosis	19762329
LHX4	Combined pituitary hormone deficiency 4 (CPHD4)	18073311
LIMK1	Williams syndrome (WS)	9637430
MAD2L1	Optimal hematopoiesis	17038523
MAPK10	Epileptic encephalopathy of the Lennox-Gaustaut type	
MC3R	Susceptibility to obesity	23680515
MC4R	Increased adiposity and linear growth	12851322
MED13L	MED13L Haploinsufficiency syndrome	24781760
MED15	22q11.2 deletion syndrome	23791650
MITF	Waardenburg syndrome type 2	10952390
MLLT3	neuromotor developmental delay, cerebellar ataxia, and epilepsy	16001262
MNX1	Currarino syndrome	32571425
MPZ	Charcot-Marie-Tooth disease type 1b	27344971
MSX1	Oligodontia	14630905
MSX2	Pleiotropic defects in bone growth and ectodermal organ formation	10742104
MYCN	Feingold syndrome	15821734
MYF6	Myopathy and severe course of Becker muscular dystrophy	11053684

Gene	Disease	PMID
MYH9	Hematological abnormalities	16162639
NCF1	Chronic granulomatous disease (CGD)	15626477
NF2	Polyneuropathy	17655741
NFIA	Complex nervous system malformations and urinary tract defects	17530927
NFRKB	Cellular immunodeficiency, pancytopenia, malformations	11920839
NIPBL	Cornelia de Lange Syndrome (CdLS)	25255084
NKX2-1	Choreoathetosis, hypothyroidism, pulmonary alterations	11854319
NKX2-5	Microcephaly and congenital heart disease	16470726
NLRP3	Autoimmunity	30052286
NOG	Severe liver fibrosis and hepatitis-induced carcinogenesis	15197228
NPAS3	Schizophrenia	12746393
NR2F2	Severe isolated congenital diaphragmatic hernia	24122781
NR5A2	Acute pancreatitis	29443959
NRXN1	Intellectual disability, Autism Spectrum Dis.	23533028
OPA1	Optic atrophy	16735988
OTX2	Subfertility	30261489
PAFAH1B1	Lissencephaly	28811646
PAX1	Klippel-Feil syndrome	12774041
PAX2	Renal-coloboma syndrome	14569086
PAX3	Developmental delay and autism	12070244
PAX6	Eye diseases	16866875
PAX8	Congenital hypothyroidism	22898500
PCGF2	Recognizable Syndrome with Craniofacial and Neurological Features	30343942
PHOX2B	Hirschsprung disease	12631670
PIK3R1	Insulin resistance	10829070
PITX1	Clubfoot	21775501
PITX2	Atrial fibrillation (AF)	24395921
PITX3	Posterior polar cataracts and neurodevelopmental abnormalities	16565358
PKD1	Autosomal dominant polycystic kidney disease (ADPKD)	18679710
PKD2	Autosomal dominant polycystic kidney disease	16720597
PMP22	Hereditary neuropathy with liability to pressure palsies	15955700
PPARGC1B	Predicted by HIPred score	HP 0,5935
PRM1	Infertility	11326282
PRM2	Infertility	11326282
PROX1	Lymphatic vascular defects, adult-onset obesity	16170315
PRRT2	Paroxysmal Kinesigenic Dyskinesia (PKD)	25027704
RAE1	Premature separation of sister chromatids, severe aneuploidy	16355229
RAI1	Smith-Magenis syndrome	17041942
RASA1	Capillary malformation-arteriovenous malformation (CM-AVM)	21626678
RB1	Metaphase cytogenetic abnormalities	12531801
RBPJ	Embryonic lethality and formation of arteriovenous malformations	15466160
RELN	Cognitive disruption and altered hippocampus synaptic function	16376115
RNF135	Phenotypic abnormalities	17632510
RPS17	Diamond-Blackfan anemia	22045982
RPS24	Myelodysplastic syndromes (MDS)	26408650

Gene	Disease	PMID
RPS26	Diamond-Blackfan anemia	24675553
RTN4R	Schizophrenia susceptibility	18043741
RUNX2	cleidocranial dysplasia	16270353
SALL1	Townes-Brocks syndrome	16429401
SALL4	Okihiro syndrome	16790473
SATB2	Craniofacial dysmorphologies, cleft palate	16960803
SCN1A	Severe myoclonic epilepsy, Dravet syndrome	16865694
SCN2A	Autism spectrum disorder (ASD)	31230762
SDHC	Paraganglioma	19546167
SDHD	Paragangliomas	10657297
SETD5	KBG syndrome	32793091
SF1	Mild gonadal dysgenesis	17940071
SF3B4	Nager syndrome	22541558
SHANK3	Phelan-McDermid syndrome	23758760
SHFM1	Severe mental retardation, microcephaly and deafness	17230488
SHH	Holoprosencephaly, sacral anomalies, and situs ambiguus	10852374
SIX3	Holoprosencephaly	18694563
SIX6	Bilateral anophthalmia and pituitary anomalies	10512683
SLC2A1	GLUT1 Deficiency syndrome	20129935
SLC40A1	Ferroportin disease	16135412
SLC5A3	Brain inositol deficiency	16644257
SLC9A3R1	Predicted by HIPred score	HP 0,5497
SMAD3	Thoracic aortic disease	30661052
SMAD4	Juvenile polyposis hereditary hemorrhagic telangiectasia syndrome	23090737
SMARCB1	Schwannomatosis	18285426
SMARCC1	Predicted by HIPred score	HP 0,8313
SNCA	Familial Parkinson's disease	12477695
SOX10	Waardenburg/Hirschsprung disease	11641219
SOX18	Mental retardation	17290276
SOX2	Hippocampal malformations and epilepsy	16529618
SOX8	Alpha-thalassemia-related syndrome (ART-16)	18492098
SPAST	Spastic Paraplegia 4	17035675
SPR	Dopa-responsive dystonia	15241655
SRGAP3	Severe mental retardation	12195014
STAT5A	Amelioration of IL-7-induced mortality and disease development	15870688
STXBP1	Early infantile epileptic encephalopathies	29538625
SUMO1	Nonsyndromic cleft lip and palate	17606301
SUZ12	Malformation of the bain and neural tube	19535498
SYNGAP1	Intellectual disability, Autism Spectrum Disorder, epilepsy	23161826
TBX3	Ulnar-mammary syndrome	17265068
TBX5	Congenital heart disease (CHD)	33321106
TCF12	Kallmann syndrome	32620954
TCF4	Pitt-Hopkins syndrome	17478476
TCOF1	Treacher Collins syndrome (TCS)	17552945
TERT	Short telomere syndromes	32315675

Gene	Disease	PMID
TFAP2B	Patent ductus arteriosus (PDA)	15684060
TFRC	Stressed erythropoiesis and neurologic abnormalities	25782630
TGIF1	Holoprosencephaly 4	16962354
TNXB	Ehlers-Danlos Syndrome	27297501
TPM1	Type 3 familial hypertrophic cardiomyopathy	26025024
TRPS1	Tricho-rhino-phalangeal (TRP) syndromes	11285235
TWIST1	Coronal synostosis	16540516
UBE3A	Angelman syndrome (AS)	20034088
WNT2B	Mental retardation, short stature	17351355
XRCC5	Growth retardation	16325483
ZEB2	Mowat-Wilson syndrome	32950463
ZIC2	Holoprosencephaly (HPE)	11285244

Table S2: List of the 47 HI WT Kozak sequences and 149 variants identified from the screening as able to upregulate translation from the respective gene. The hits selected for the validation are highlighted in yellow.

Gene	WT	Variants
AHSP	GCAGATGGCTC	GCGGATGGCTC
		GCAGATGGTTT
		GCAAATGGCTC
		GTAGATGGCTT
		GCAGATGGCTT
		GTAGATGGCTC
ATR	CAGCATGGGGG	CAGCATGAAGG
		CAGCATGGAGA
		CAACATGAAAG
		CAACATGAGAA
		CAACATGAGGG
		CAACATGAGAG
BCL11B	GGCAATGTCCC	GGCAATGCCCC
		GGTAATGTCCT
BMP4	GAAGATGCGAG	GGAGATGCGGG
		GAGGATGCGAG
		AAAAATGCAAA
		GAAGATGCGAA
		GAGGATGCGGG
BUB3	GCAGATGACCG	GCAGATGGCCG
		ACAGATGACCA
CHD5	GGGCATGCGGG	AGACATGCGAG
		GGACATGCGAG
CHRNA7	CAACATGCGCT	TAATATGTGCT
COL2A1	AGCCATGATTC	AGCCATGACCC
		AGCCATGACTC
COMT	GCAGATGCCGG	GCAGATGCCGA
		GCAAATGCCGA
DMRT1	CACCATGCCCA	TACTATGCCCA
		CATTATGTCCA
		TACCATGTCTA
		TACTATGTCCA
		CATCATGCCTA
DSP	CGACATGAGCT	TGATATGAGCT
DYRK1A	GACGATGCATA	GACGATGCACA
EFTUD2	CATCATGGATA	CATCATGGATG
ELOVL4	CGCGATGGGGC	CGCGATGGGAC
		CGCAATGGAGC
		CGCGATGAAGC
		CGCGATGGAGC
		CACGATGGAGC
EYA1	GTCTATGGAAA	GTCTATGGAAG

Gene	WT	Variants
FBLN1	GCCCATGGAGC	GCCCATGGAGT
		ACCCATGGAGC
		GCTTATGGAGT
		GCCTATGGAGT
		ACCCATGGAAC
		GCTCATGGAGT
		GTCTATGGAGT
FEN1	TGCCATGGGAA	TGCCATGAGAA
FGF8	CGCCATGGGCA	CACCATGAGCA
		CACCATGAACA
FGFR1	TGGGATGTGGA	TGGGATGTAGA
FKBP6	GGACATGGGGG	GGACATGAGGA
		GGACATGAGAA
		GAACATGGAAG
		GAACATGAAAG
		GAACATGGGAG
		GGACATGGAGA
		AAACATGGAAG
		AAACATGGGAG
		GAACATGGAGG
		GAACATGAGGG
		AGACATGGAAG
GALR1	GGCCATGGAGC	GGTTATGGAGC
		AACCATGGAGC
		GACCATGGAAC
		AACCATGAAGC
GATA6	GTGGATGGCCT	ATAAATGGCCT
		ATAGATGGCCT
		GTAGATGGCCT
		GCGGATGGCCT
GHRL	GGCCATGCCCT	GGTCATGCCCT
		GGCTATGTCTT
		GGTTATGTCCT
		GGCCATGCCTT
		GGCCATGTCTT
GRIN2B	GAAGATGAAGC	GAGGATGGAGC
		GAGGATGAAGC
		GGAGATGAGGC
		GAAGATGGAGC
		GGGGATGAAGC
HIC1	CTGAATGACTT	CCGAATGACTT
HNF1A	AGCCATGGTTT	AGCCATGGCCC

Gene	WT	Variants
IGF1	AGCAATGGGAA	AACAATGAGAA
KANSL1	CTGAATGGCTG	CTAAATGACTG
KRT5	CACCATGTCTC	TATTATGTCTC
MYCN	GCCGATGCCGA	GCCAATGCCGA
		ACCAATGCCAA
NCF1	AGTCATGGGGG	AGTCATGAAAA
		AGTCATGGGAA
		AGTCATGAGAA
		AGTCATGGAAA
		AGTCATGGGAG
NLRP3	GCAGATGAAGA	ACAGATGAAGA
		ACAAATGAAGA
NRXN1	GAGCATGGGGA	GAGCATGAGGA
OPA1	CGGGATGTGGC	CGGAATGTAAC
		CAAGATGTAAC
PAFAH1B1	CAAGATGGTGC	CAGGATGGTGC
		CAAGATGGCGC
		TAAGATGGTGT
PPARGC1B	AGCAATGGGTG	AGCAATGAGTA
		AACAATGAGTA
		AGCAATGGATG
		AGCAATGAATA
PROX1	AGTGATGCCTG	AGTGATGTCTG
RPS17	CACCATGGTAG	CACCATGGTAA
		CACCATGGCAG
SIX3	GTCCATGGTAT	GTCCATGGCAT
		GCCCATGGCAC
SIX6	CTCGATGTTCC	CCCGATGTTCC
		CTTGATGTTCC
		CTCGATGTTTC
SOX10	CGACATGGCGG	CGACATGACGG

Gene	WT	Variants
SMARCC1	GACGATGGCCG	GACGATGGCTG
		AACAATGGCCA
		GACAATGGCCG
		AACGATGGCCA
		AACAATGGCCG
		AACGATGACCA
		GATGATGGCTG
STXBP1	CGCCATGGCCC	CGTCATGGCCT
		TGCCATGGCTT
		CGCTATGGCCT
		TGTCATGGCCT
		TGTTATGGCCT
		CGTTATGGCCT
		CGCTATGGTTT
		CGCTATGGCCC
		TGCCATGGCCT
		TGCCATGGCTC
		CGTCATGGCTC
		CGTCATGGTCT
		TGTTATGGCTC
		CGCCATGGCTC
TBX3	GTGGATGAGCC	GCGGATGAGCC
		GTAAATGAACC
TGIF1	GAGGATGGTTC	AAAGATGGTTC
WNT2B	AGCTATGTGAG	AGCTATGTTGA
ZEB2	ATCAATGAAGC	GTCGATGGAGC
		ATCGATGGAGC
		ATCGATGGGGC
		GTCGATGGGGC
		GTCAATGAAGC

Table S3: List of the oligos used in this study.
Oligos for pWPT-/C-1T-mEGFP-IRES-mCherry cloning:

oligo XhoI-EcoRI C-1T + PAM fw	TCGAgtcccatctaactaagccacTatgggcg
oligo XhoI-EcoRI C-1T + PAM rev	AATTcgcccatAgtggcttagttagatgggac

Oligos for pWPT-mCherry cloning:

oligo XhoI-EcoRI stop fw	TCGAgtaataataatagtgataag
oligo XhoI-EcoRI stop rev	AATTcttatcactattattattac

Representative oligos for cloning the Kozak sequence of the hits in pWPT-EGFP-IRES-mCherry

oligo XhoI-EcoRI Kozak fw	TCGAgtaactaactaagcNNNNATGNNNN
oligo XhoI-EcoRI Kozak rev	AATTNNNNCATNNNNgcttagttagttac

Oligos for PCR for EditR analysis:

oligo EGFP pWPT (EF-1alfa promoter) fw	ccgagggtgggggagaac
oligo EGFP pWPT rev	agctcgtccatgccgagagtgatc
oligo NCF1 fw	AGCCTGAAGAGTCCCCAGAA
oligo NCF1 rev	CACTCTCTGATAGCTGGGCT

Oligos for qPCR:

oligo qPCR NCF1 fw	GCGAGAGCGGTTGGTGGTTC
oligo qPCR NCF1 rev	TGTAGGCCTTGATGGCGACG
oligo qPCR HPRT1 fw	TGACACTGGCAAAACAATGCA
oligo qPCR HPRT1 rev	GGTCCTTTTCACCAGCAAGCT

Oligos for library cloning and deep sequencing. The part annealing to the target sequence is in bold.

oligo Kozak library fw	tcgtgacgcgcatccaggcc
oligo Kozak library rev	caccccggtgaacagctcctc

oligo F3 fw deep sequencing	tcgtcggcagcgtcagatgtgtataagagaca**gccagaacacaggtgtcgtga**
oligo R1 rev deep sequencing	gtctcgtgggctcggagatgtgtataagagaca**gccggtggtgcagatgaactt**

Table S4: sequences of sgRNA spacers used for base editing, with the respective base editor used and the target positions. PAM sequence is highlighted in bold; the target positions are in red; the ATG starting codon is underlined.

	Protospacer	**Base editor**	**Target**
EGFP sgBE1	CCATAGTGGCTTAGTTAGAT	ABE7.10, ABEmax	**ccc**atctaactaagccacTatgggc
NCF1 sgBE1	AGGTGTCCCCCATGACTGGG	BE4max, AncBE4 max	**cca**cccagtcatggGGGacacct
NCF1 sgBE2	TGAAGGTGTCCCCCATGACT		**ccc**agtcatggGGGacaccttca
NCF1 sgBE3	ATGAAGGTGTCCCCCATGAC		**cca**gtcatggGGGacaccttcat