PNA-screen - initial analysis

I received the data and started to do some analysis. The total amount of downregulation at 10 uM is:

a.	K12:	231/585 - 39%
b.	UPEC:	418/585 - 71%

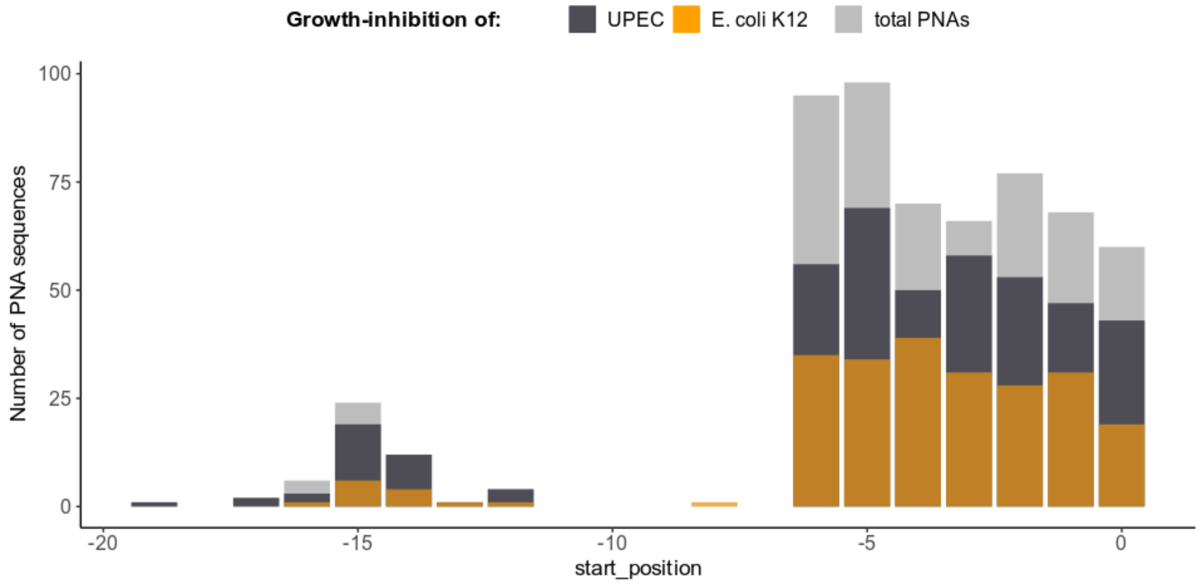
I first generated, for all PNA sequences different PNA-specific and gene-specific attributes:

Gene/PNA specific	Feature group	Feature name	Description	Feature type	Number of features
PNA	Sequence	sequences_onehot_encoded	One-hot encoded sequences (4 nt * 9mers = 36)	categorical	36
PNA	Distance to CDS start	distance_start_cds	Distance to start codon (bp)	numeric	1
PNA	Thermodynamic	PNA_GC_content	GC content of gRNA (%)	numeric	1
PNA	Thermodynamic	PNA_purine_content	purine content of PNA (%)	numeric	1
PNA	Thermodynamic	PNA_longest_purine_stretch	Longest purine stretch PNA	numeric	1
PNA	Thermodynamic	PNA_melting_temp	melting temperature between mRNA and PNA	numeric	1
PNA	Thermodynamic	homopolymers	Length of longest consecutive nucleotides (nt)	numeric	1
PNA	Thermodynamic	SC_bases	nr of consecutive self-matching bases	numeric	1
PNA	Thermodynamic	Mw	Molecular weight	numeric	1
PNA	off-targets	off_targets_tot	nr of total off-target sites (up to 2 mm)	numeric	3
PNA	off-targets	off_targets_tir	nr of TIR off-target sites (up to 2 mm)	numeric	3
gene	Expression level	gene_expression	expression level (in TPM) during exponential growth	numeric	1
gene	Pathway	kegg_pw	# of kegg pathways	numeric	1
gene	Operon information	operon_downstream_genes	The number of downstream genes in the same operon	numeric	1
gene	Operon information	ess_gene_operon	The number of downstream essential genes in the same operon	numeric	1
gene	Gene info	gene_GC_content	GC content of targeting gene (%)	numeric	1
gene	Gene info	gene_length	Gene length (bp)	numeric	1
gene	Gene info	sec_structure_TIR	secondary structure TIR (delta G)	numeric	1
			Total		56

1. PNA-specific features

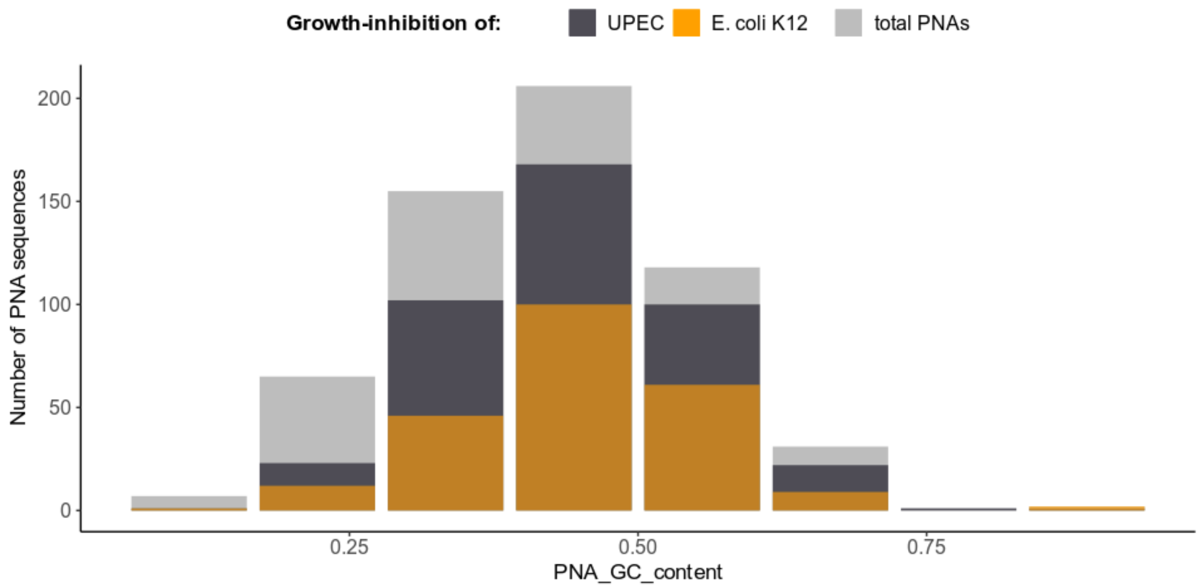
I checked the PNA sequences for different features. I saw that there some (30) PNAs that have >4 bases of self-complementary. I guess for some genes it was necessary to take them. Below I show several bar-plots which show the total number of PNAs (grey), the amount of PNAs that inhibited growth at 10 uM in UPEC (darkgrey) and K12 (orange). The x-axis is used to separate by different features: GC-content, Tm, purine-content, longest purine-stretch, longest homopolymer-stretch of PNA, self-complementary bases, and start position of PNA (rel. To CDS star, 0 is A of AUG). The plots are shown here:

Start position:



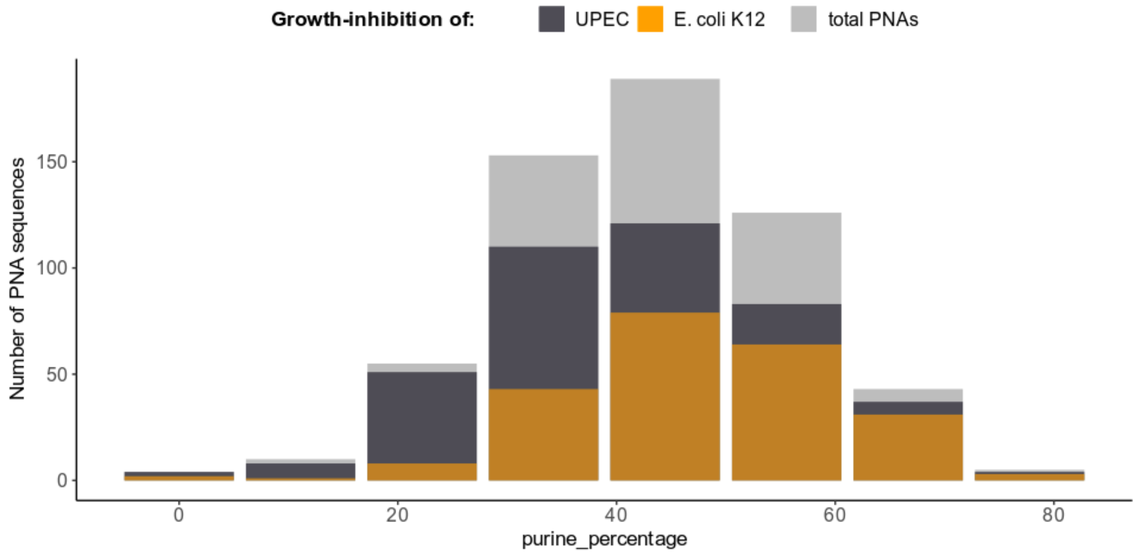
The position of the PNA seems to not have a clear effect on the PNA effect. Even the few SD-targeting PNAs seem to work very well and have high success rates.

GC-content:



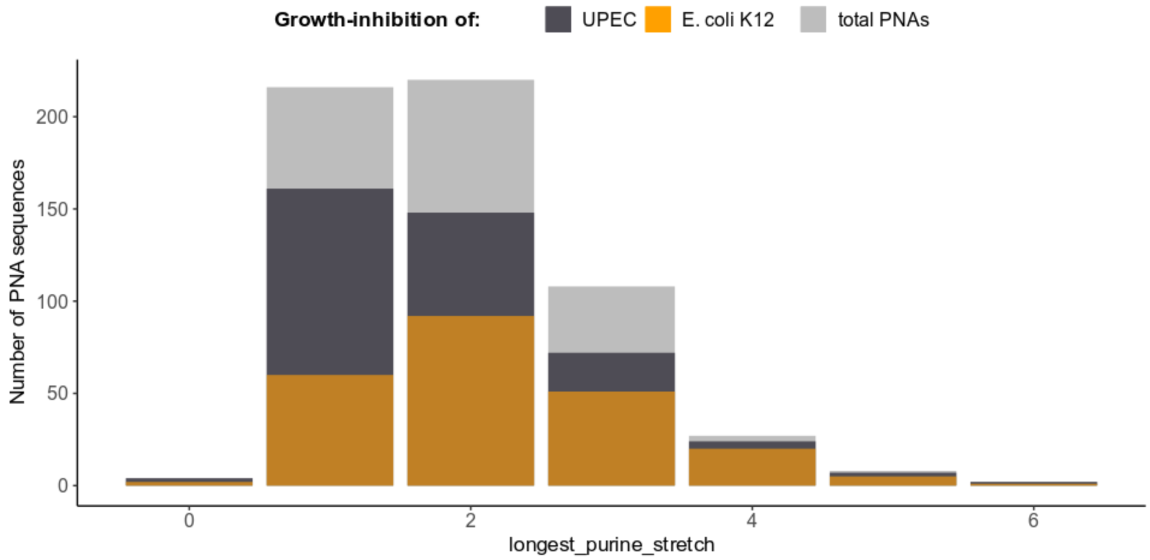
Lower relative GC contents seem to lead to smaller effects.

Purine percentage:



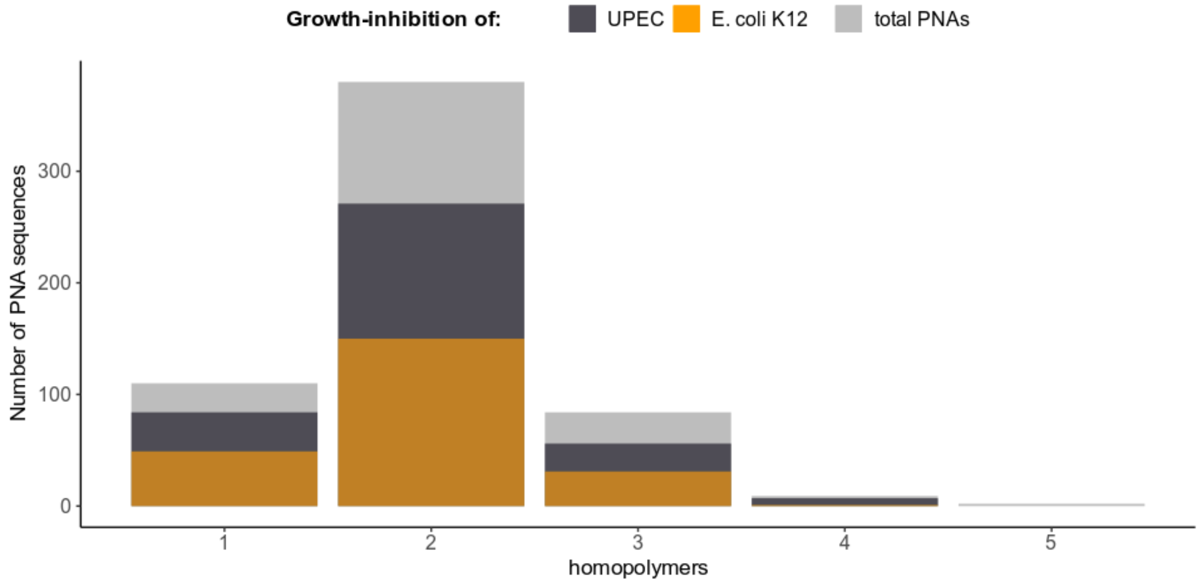
Purine percentage seems to have no effect.

Longest purine-stretch:



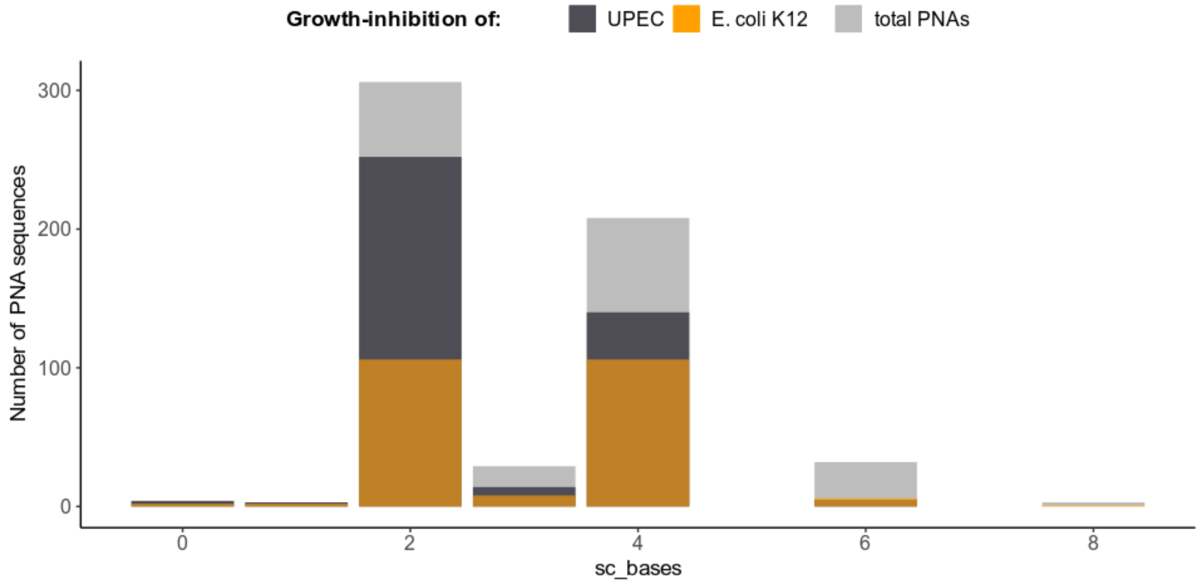
Same for the longest purine stretch. Even longer purine stretches seem to not really have a negative effect on PNA-effects.

Longest homopolymer-stretch:



Also nothing interesting.

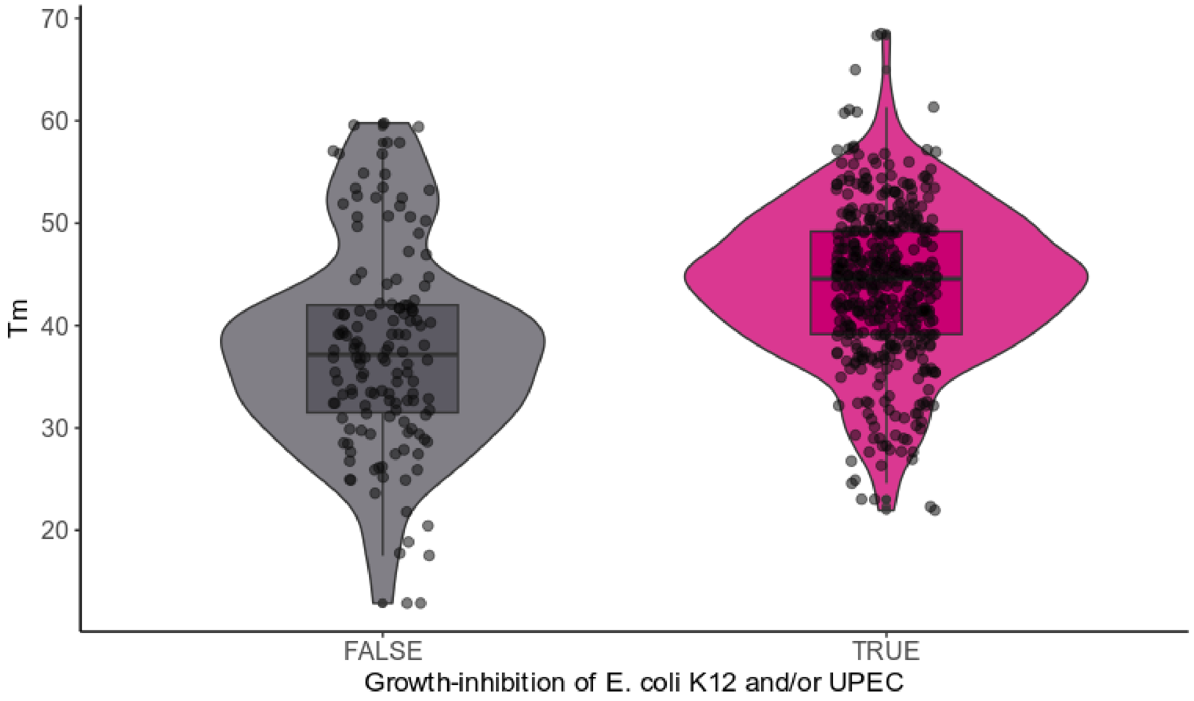
# self-complementary bases



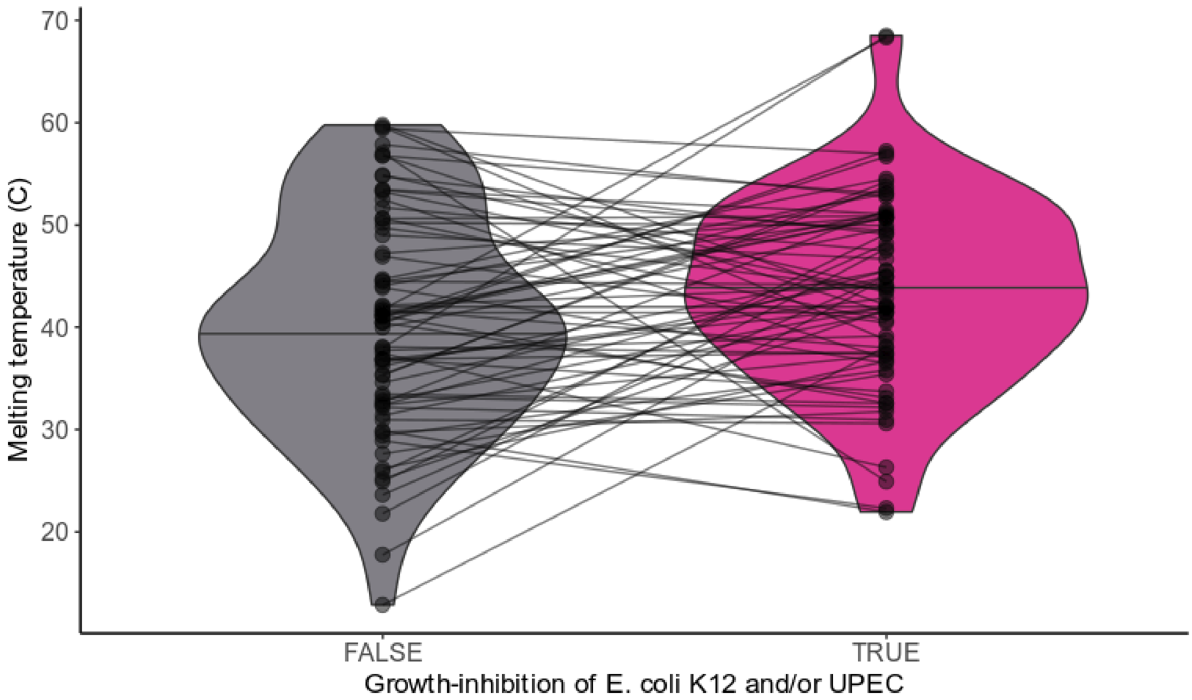
You can see that the ones with 6 self-complementary bases have lower % of growth inhibition to the respective gene.

Now I plot all the PNAs which have depleted either UPEC or K12 with the melting temperature, and different off-target frequencies to see whether there is some trend.

Tm

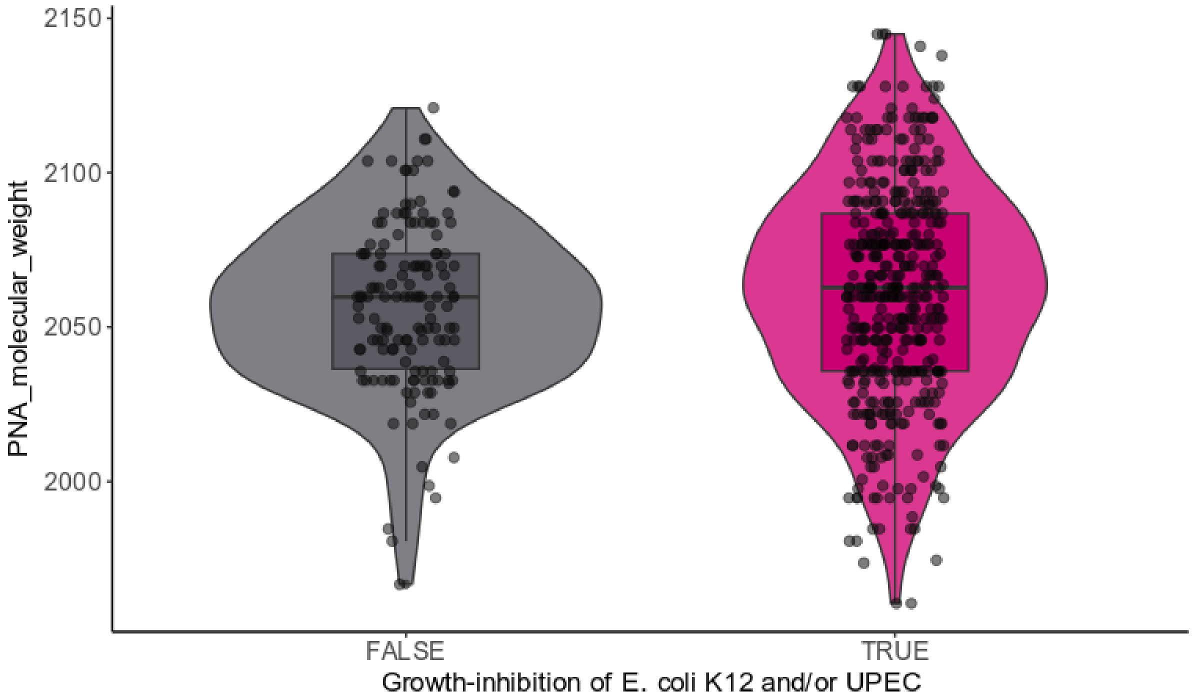


There seems to be a relationship between melting temperature and Growth-inhibition. Higher Tms go with higher PNA efficiencies.
Next, I looked at only the genes which have one effective and one not-effective PNA and see in a paired way, which PNAs belong together:



At lower Tms (around 20 degrees or lower) it is probably better to choose an alternative PNA with higher Tms.

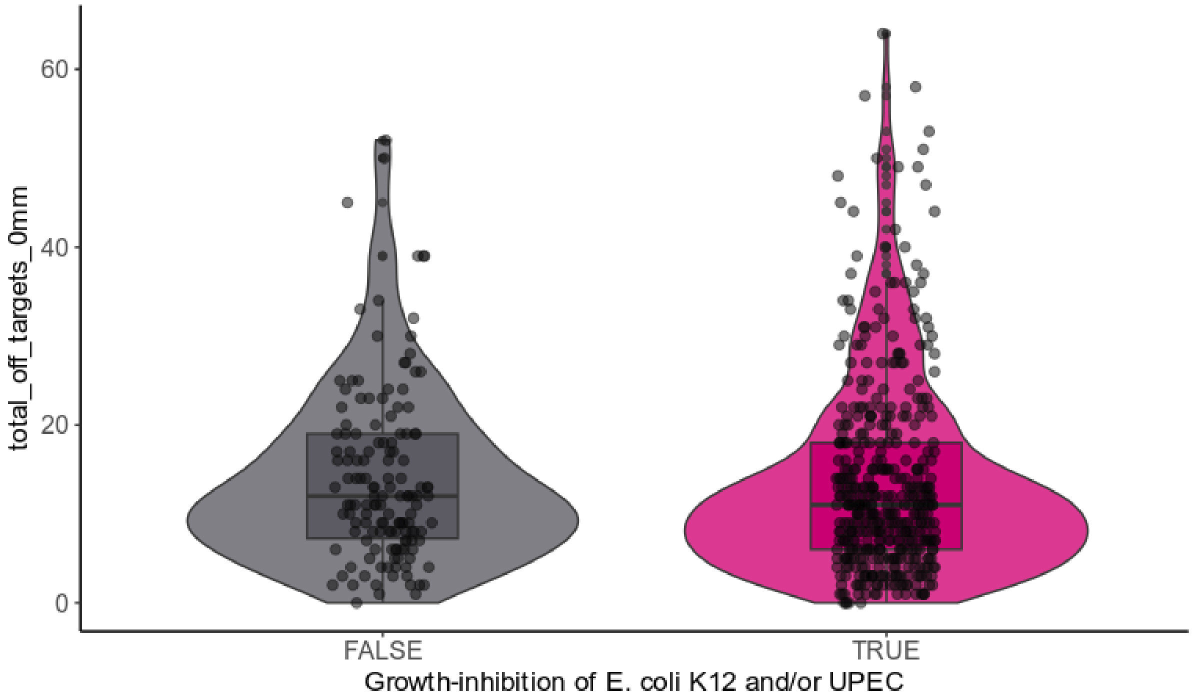
Molecular weight



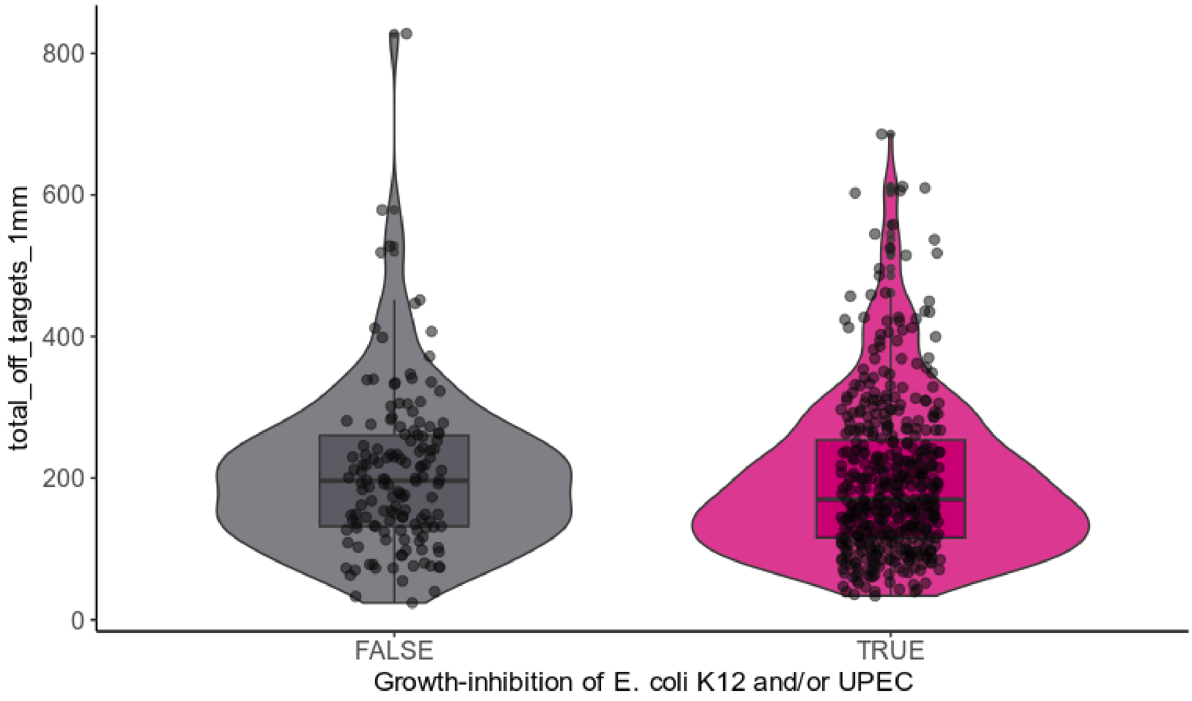
No effect of Mw.

Off-targets total

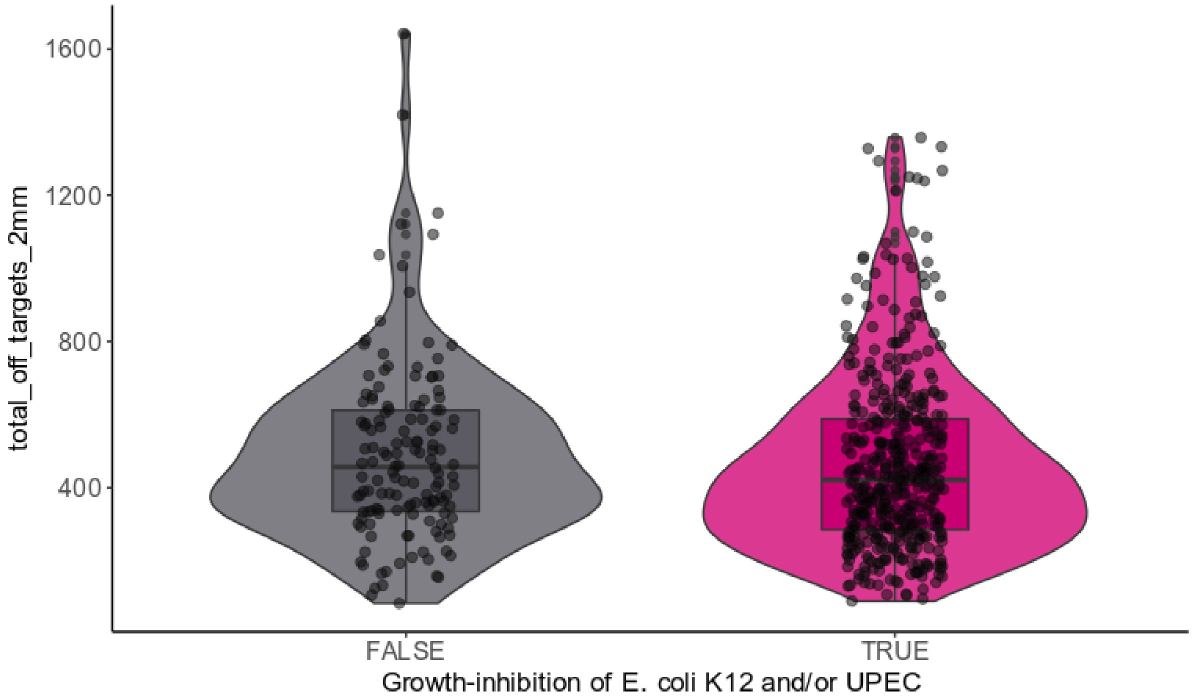
0 mismatches:



1 mismatch:



2 mismatches:

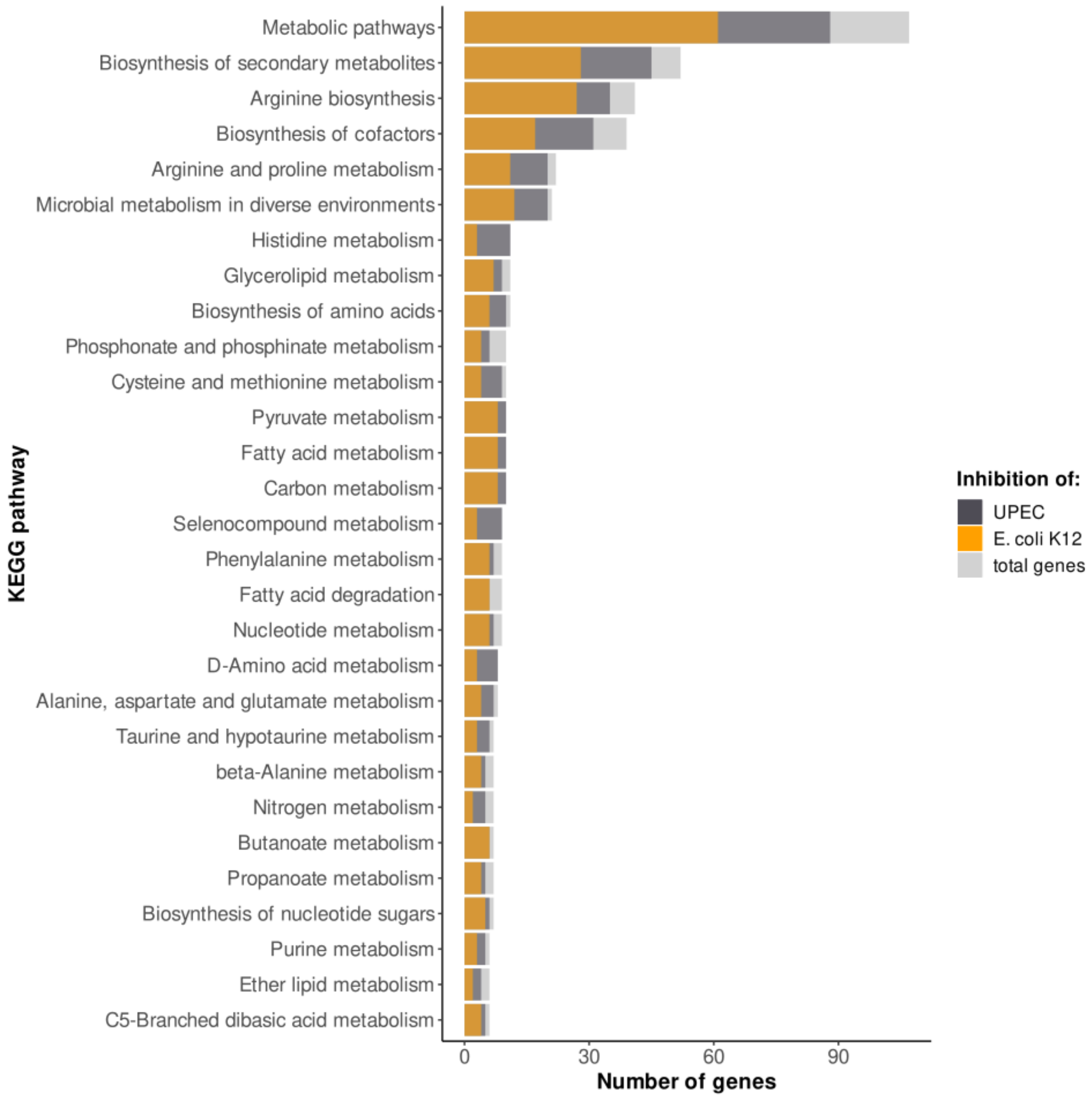


Seems like Non-effective PNAs have a trend to have higher amounts of off-targets in total (in any region). I also have checked off-targets in the TIR only but saw no effects.

2. Gene-specific features

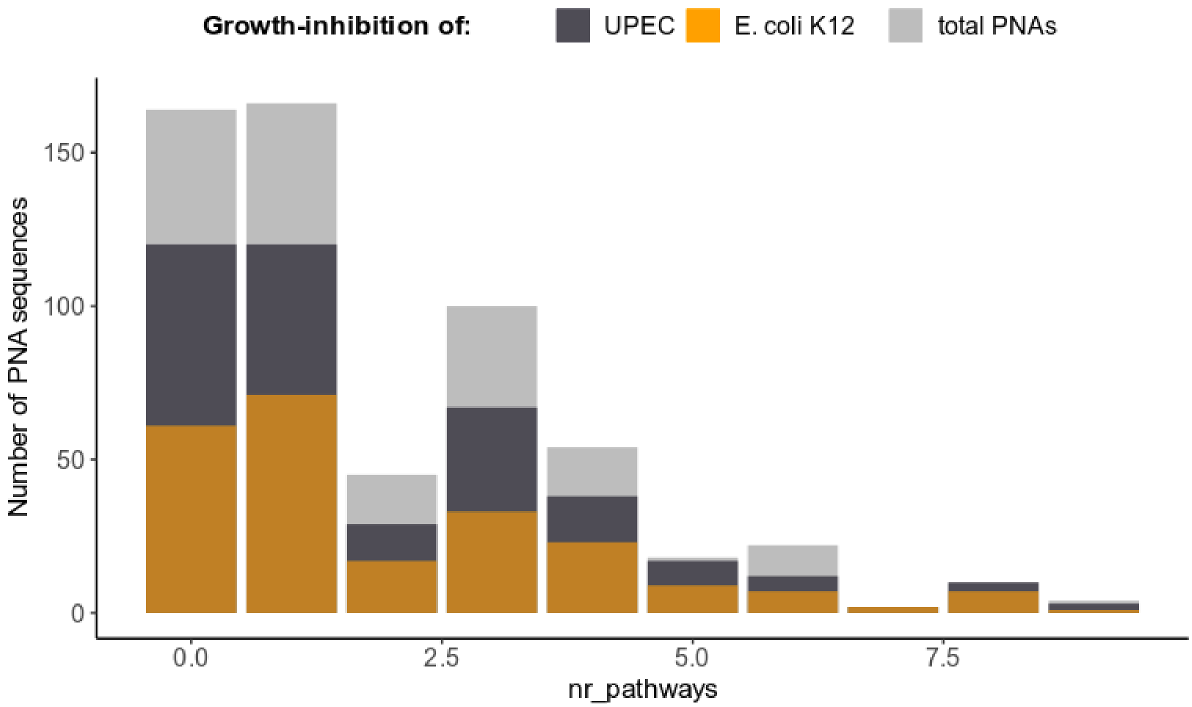
I looked at gene-specific effects and checked whether they are directly related to PNA efficiency. I looked into KEGG pathways, operon structure, secondary structure, and gene expression of the target gene.

KEGG pathways:
For each targeted gene, I checked for the KEGG pathways they belong to (in K12). I first created a plot showing all pathways with the total number of genes in our data belonging to the pathway (grey), the number of effective target genes belonging to the pathway in K12 (orange) and UPEC (dark grey):



Some PWs seem to have higher % of efficient targets (e.g. FA metabolism).

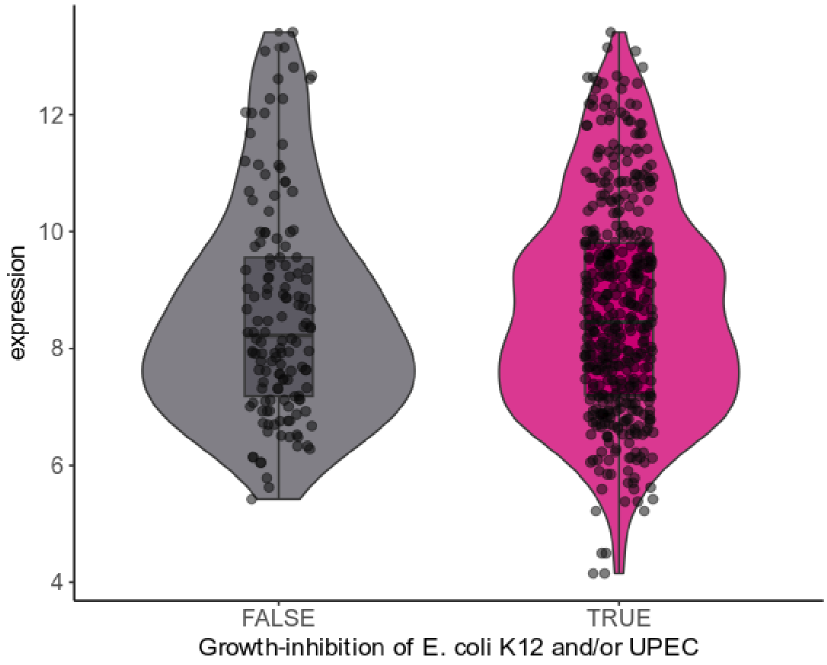
Next, I looked for all genes, how many pathways they are part of:



Maybe a slight trend indicating that if a gene belongs to more pathways at the same time, it might be more important → more essential → PNAs are more effective.

Gene expression:

I took an RNA-Seq dataset from the literature to get expression values for all the genes (from K12). The expression is in log TPM normalized:

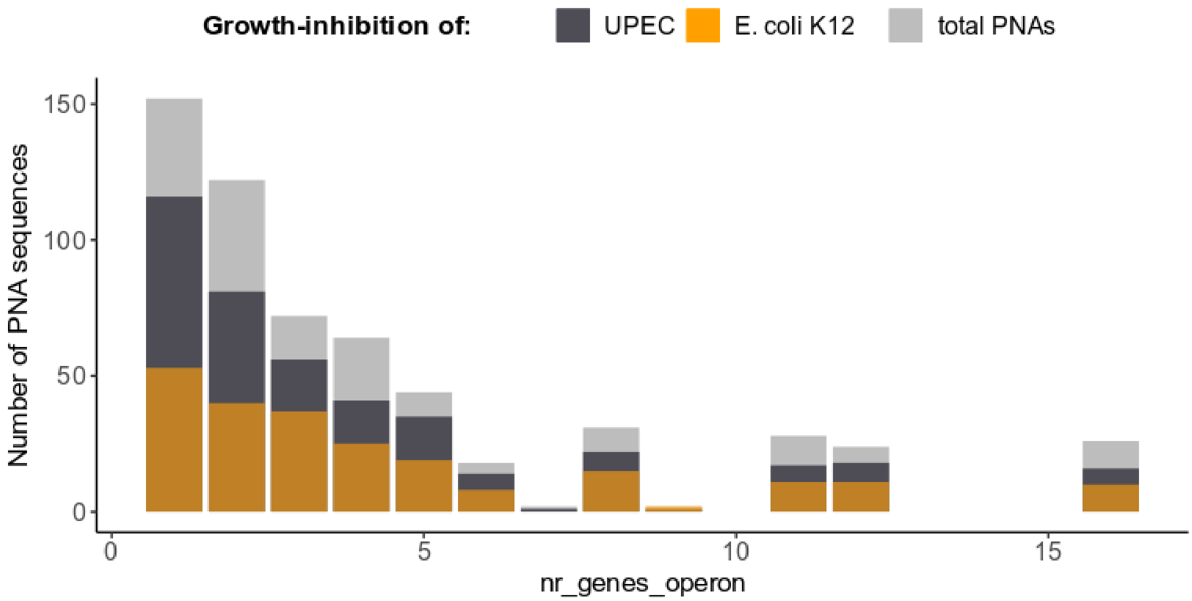


No difference can be seen (as we have seen in UPEC before)

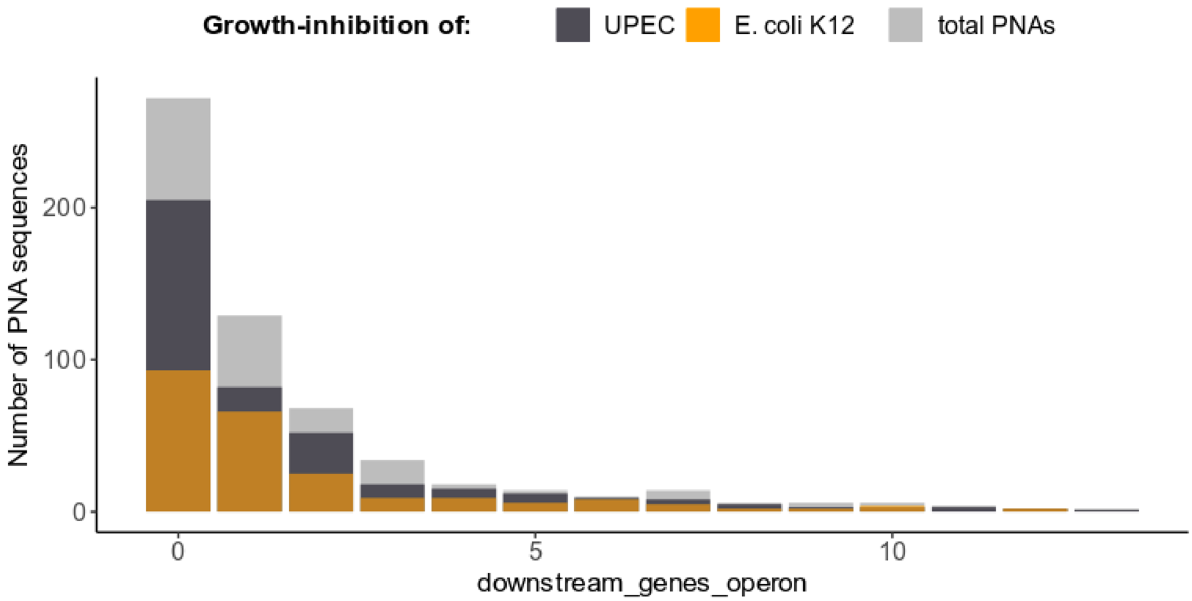
Operon effects:
I used regulonDB data for operons and generated different attributes for the genes:

- Nr of genes in the operon that the gene belongs to
- Nr of genes downstream in same operon (bc we saw before that downstream genes can be affected)
- Nr of essential genes downstream in same operon

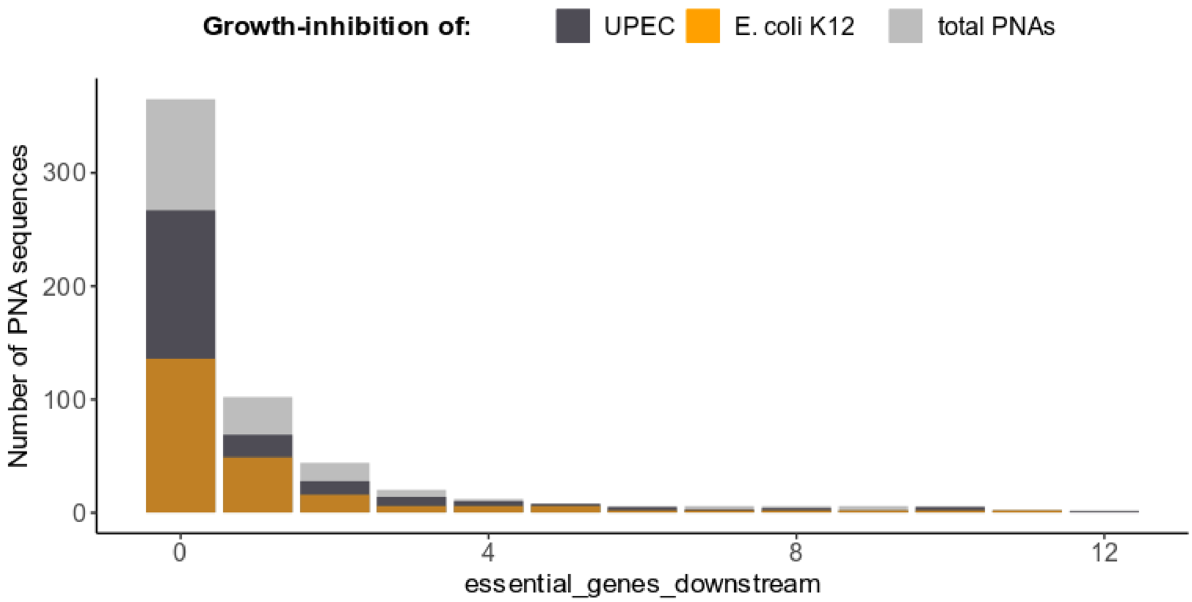
Nr of genes in operon:



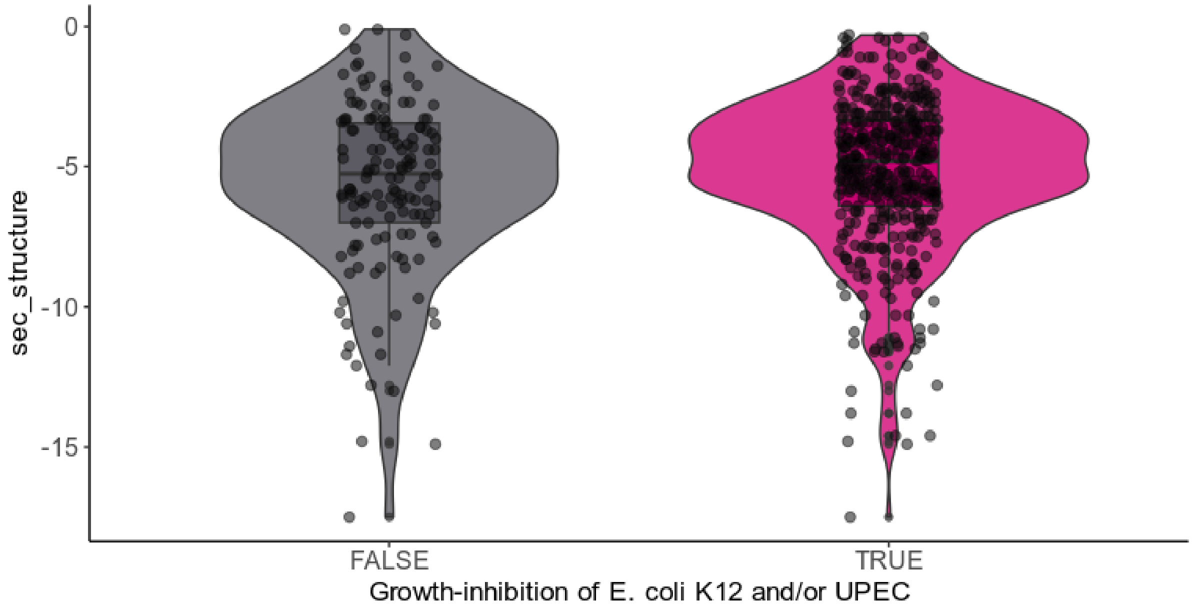
Nr of genes downstream in same operon:



Nr. of ess. Genes downstream in operon:



Secondary structure at TIR:
I calculated the secondary structure (in delta G) of TIRs of the target genes:


sec_structure
FALSE
TRUE
Growth-inhibition of E. coli K12 and/or UPEC