Bioinformatics analysis of silencing the extra repeats in CRISPR-Cas systems

Maximilian Feussner

Ph.D. student

Group of Zasha Weinberg

Leipzig University

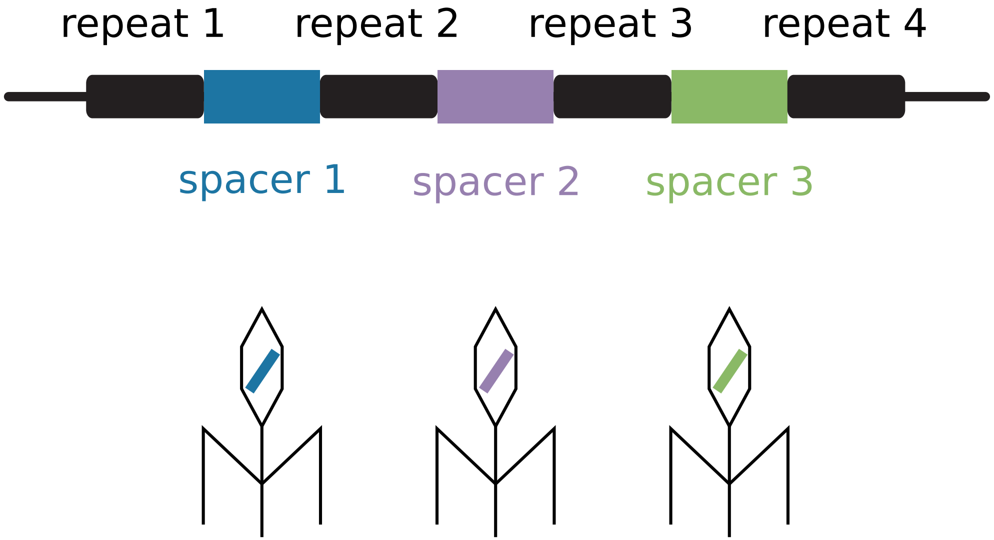
Germany

CRISPR-Cas

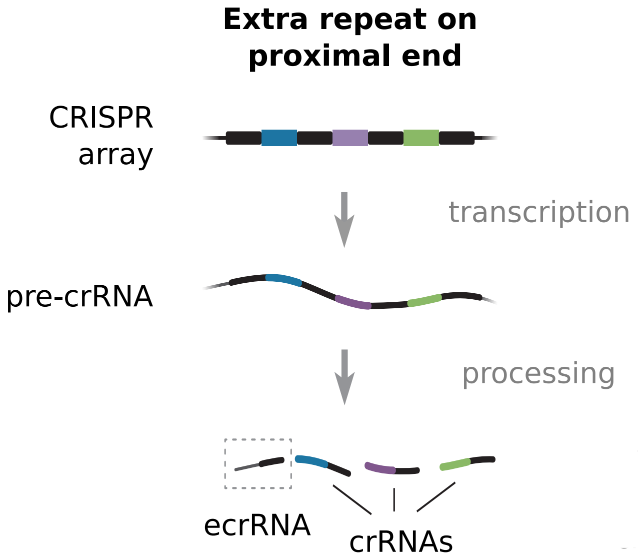
- … is an adaptive immune system for bacteria and archaea
- … has many important biotech applications
    - Genome editing to cure genetic diseases
    - Tool to silence specific genes for experiments
    - Many more
- Understanding how it works can help to exploit it

Clustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats

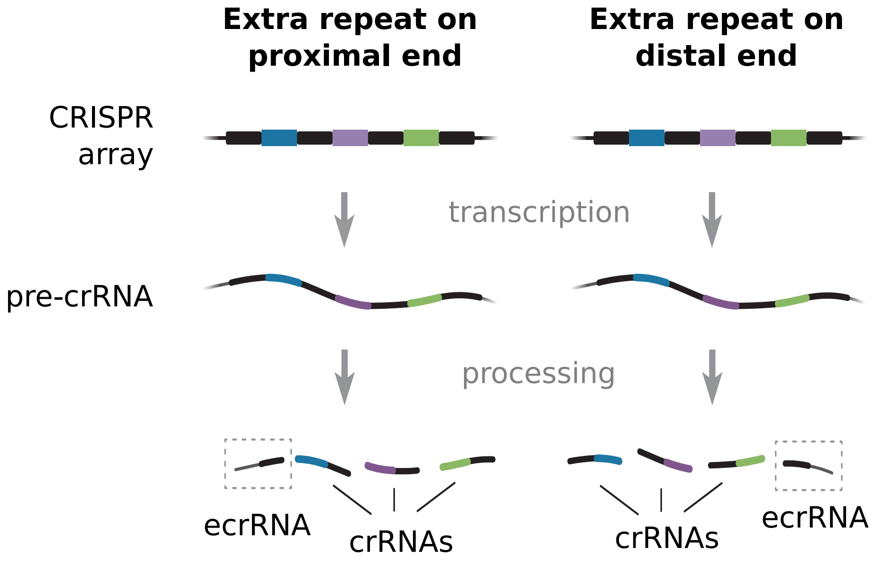
CRISPR array



There is always an extra repeat



There is always an extra repeat

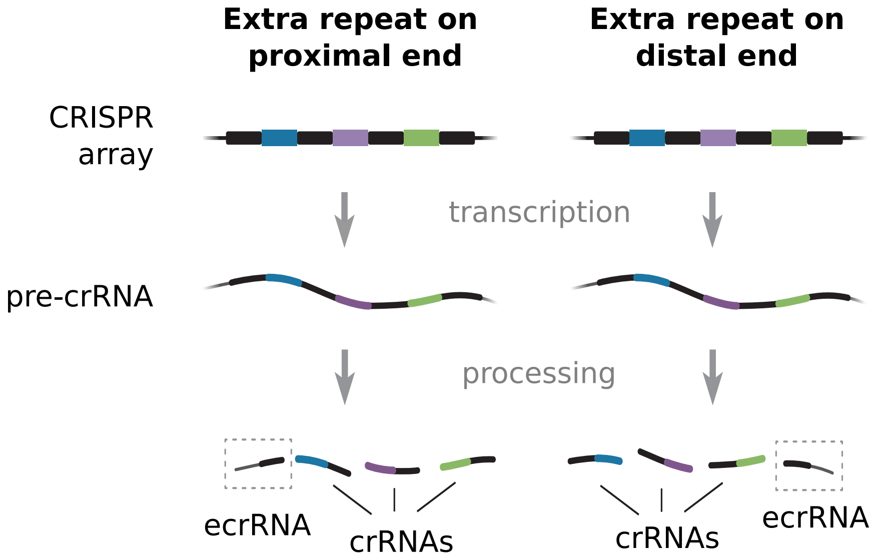


Extra repeat on proximal end

Extra repeat on distal end

Sub-types: II-A/C and VI-B I-E

There is always an extra repeat



Processing is based on sequence and secondary structure of repeat

There is always an extra

Problems:

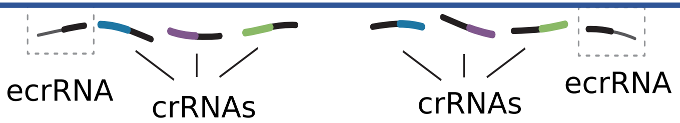
- ecrRNA does not target an invader
- At best waste of resources

ecrRNA
crRNAs
crRNAs
ecrRNA

There is always an extra

Problems:

- ecrRNA does not target an invader
- At best waste of resources



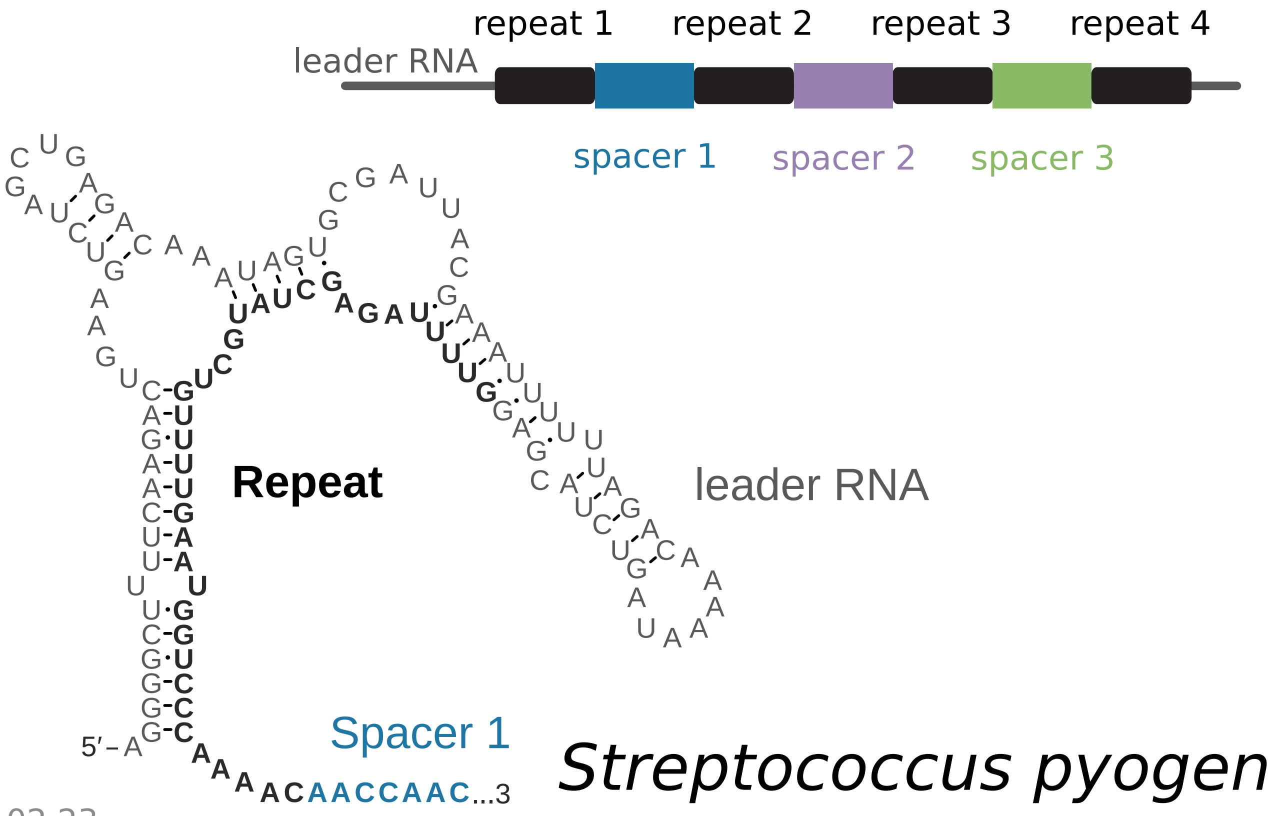
→ **It could be useful for the bacteria to prevent the processing of the extra repeat**

Two hypotheses are possible

1) Interfering secondary structure between leader and extra repeat

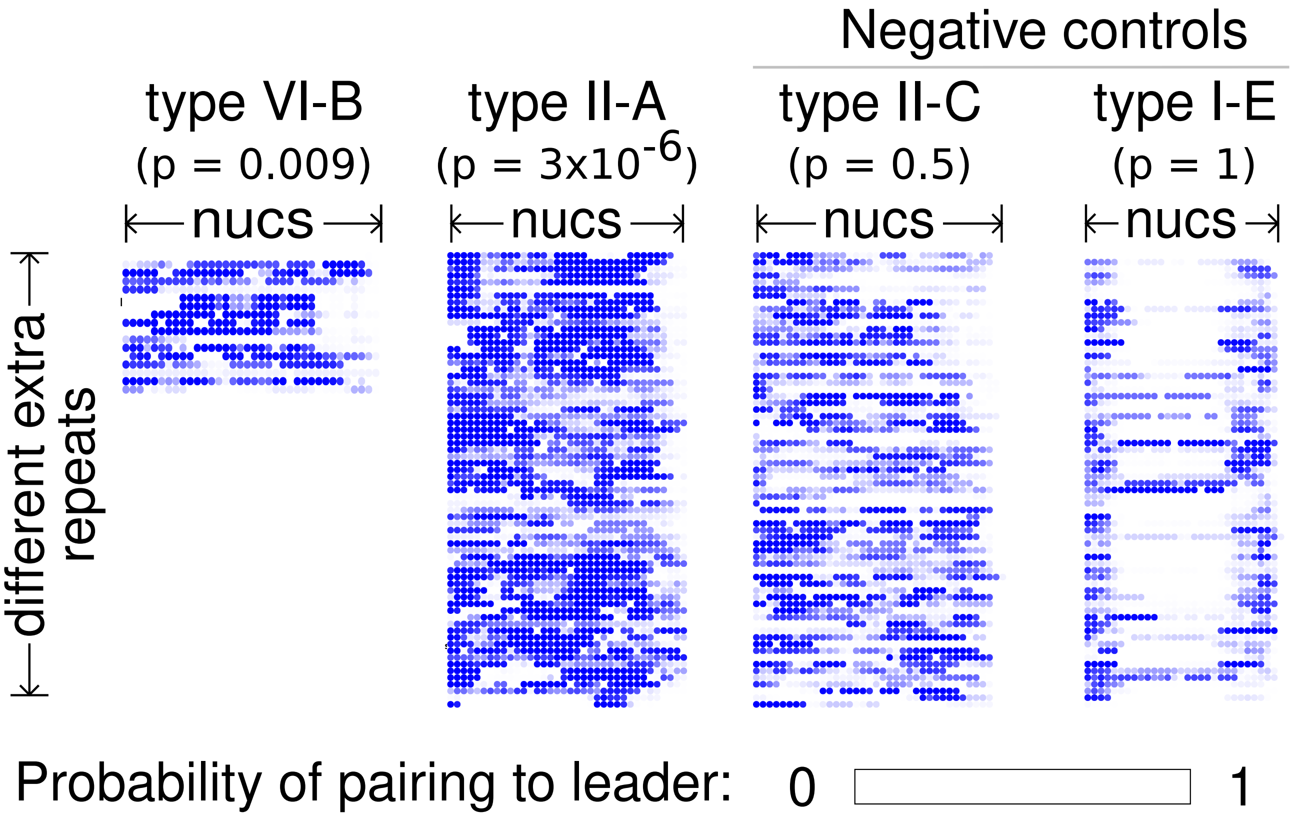
2) Mutations in the extra repeat

Interfering secondary structure: type II-A and VI-B CRISPR-Cas systems

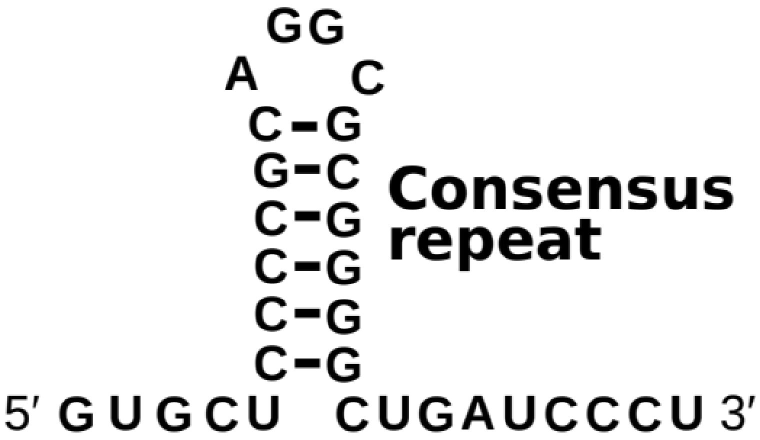


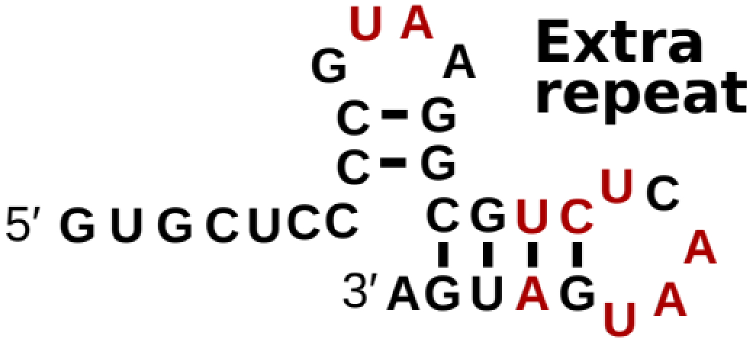
Streptococcus pyogenes (type II-A)

II-A and VI-B show strong base pairing between leader and repeat

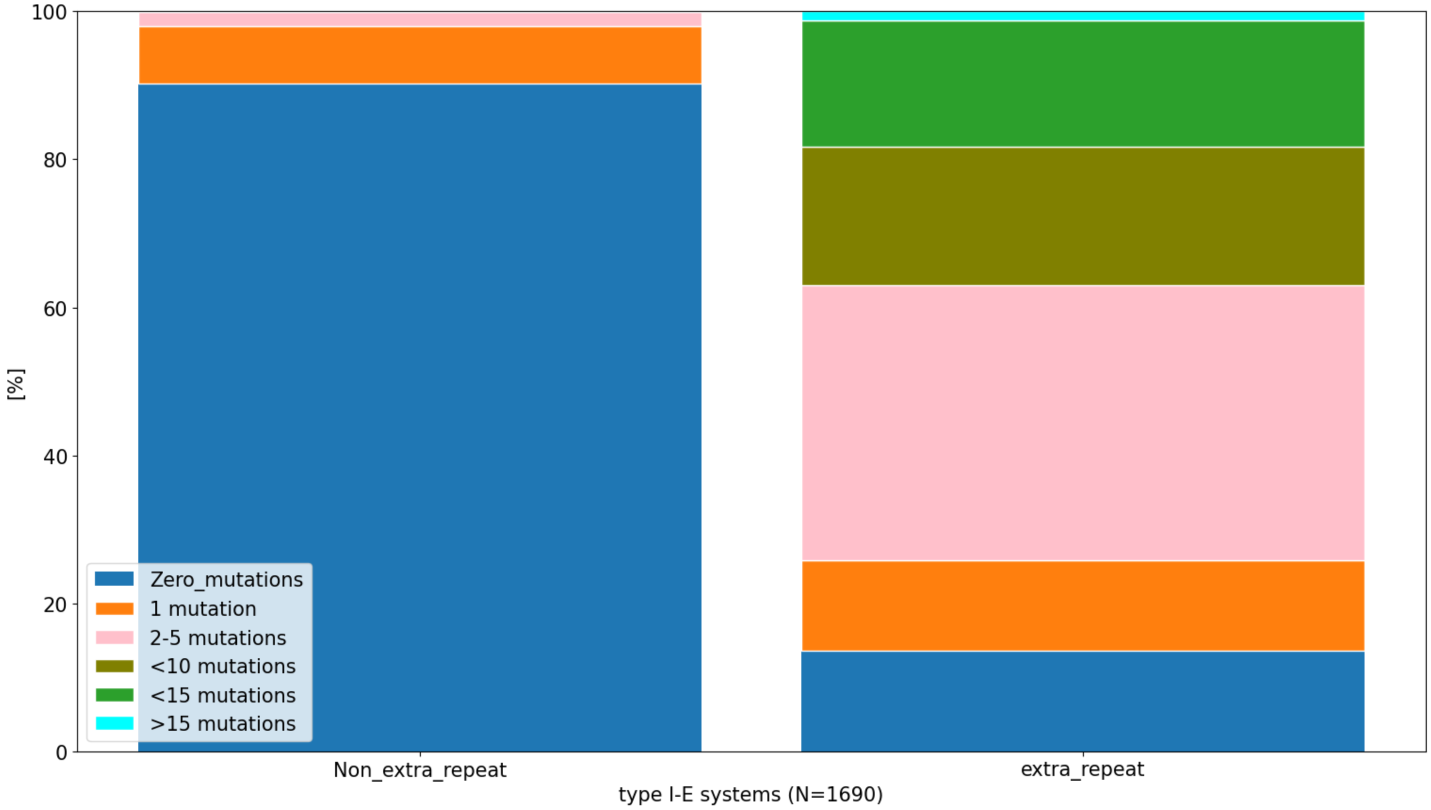


Mutations disable functional repeat





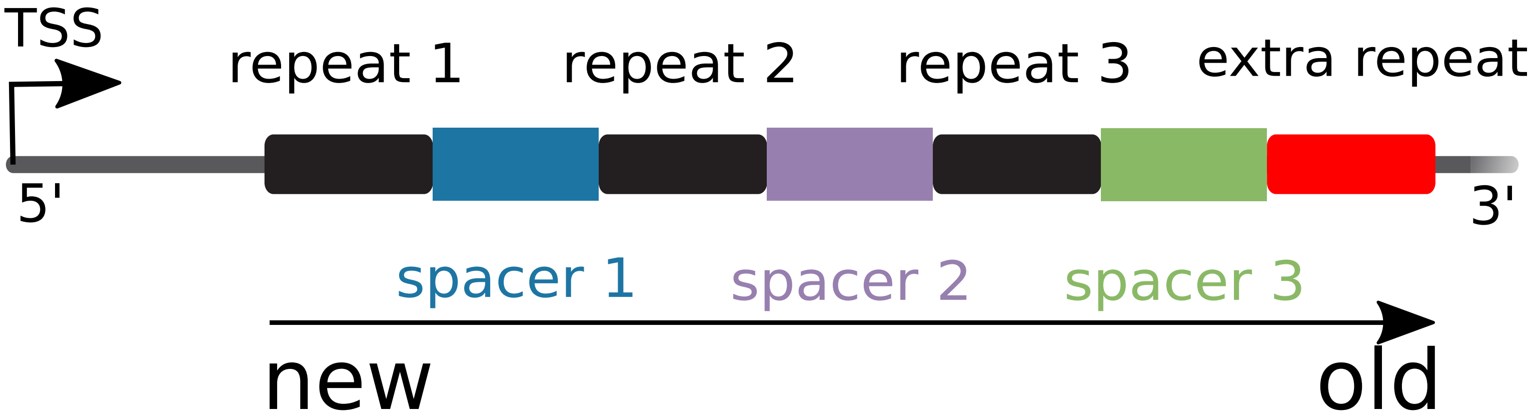
Mutations occur more often in extra repeats



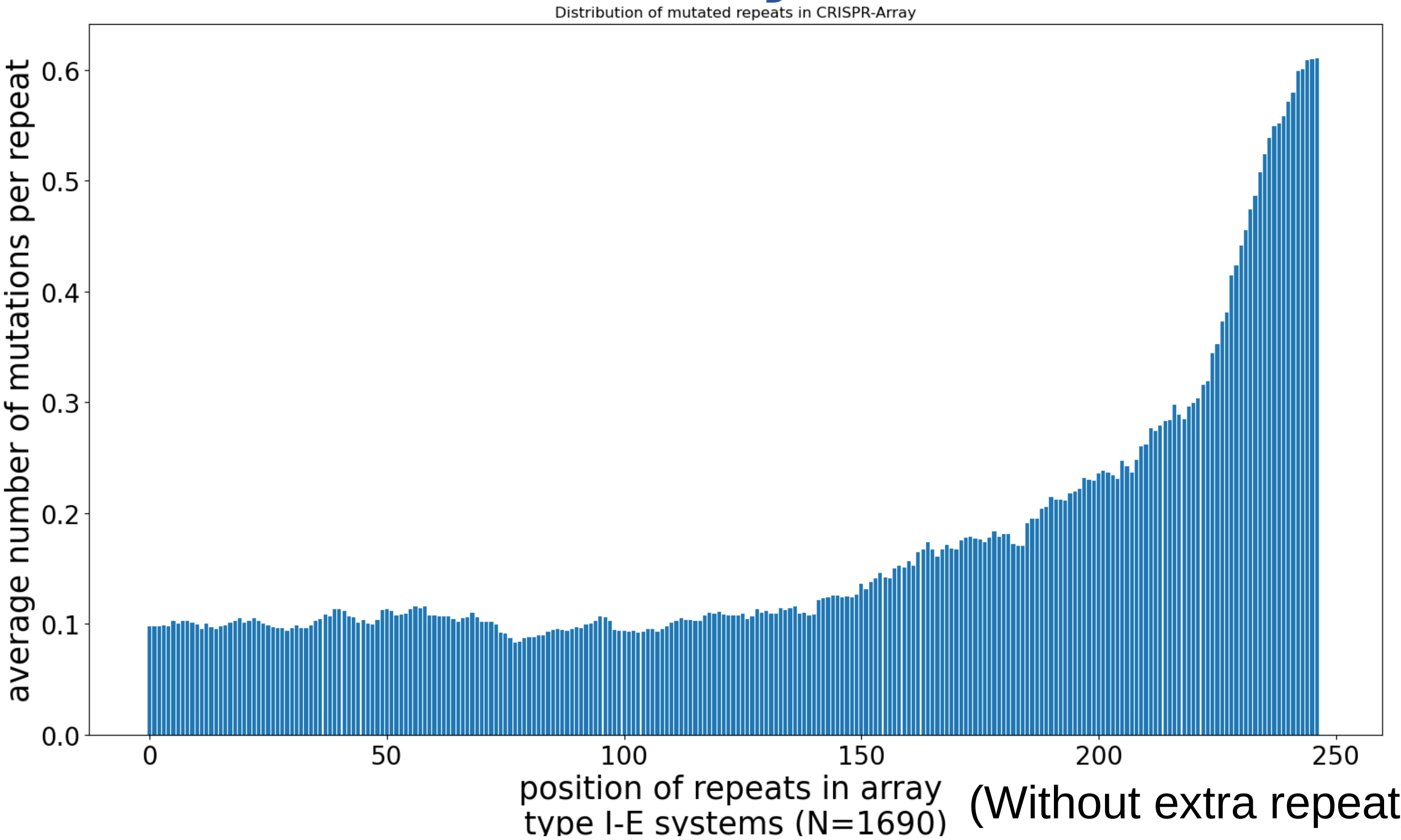
38852 non extra repeats and 1690 extra repeats

The extra repeat is the oldest repeat

Sub-type I-E

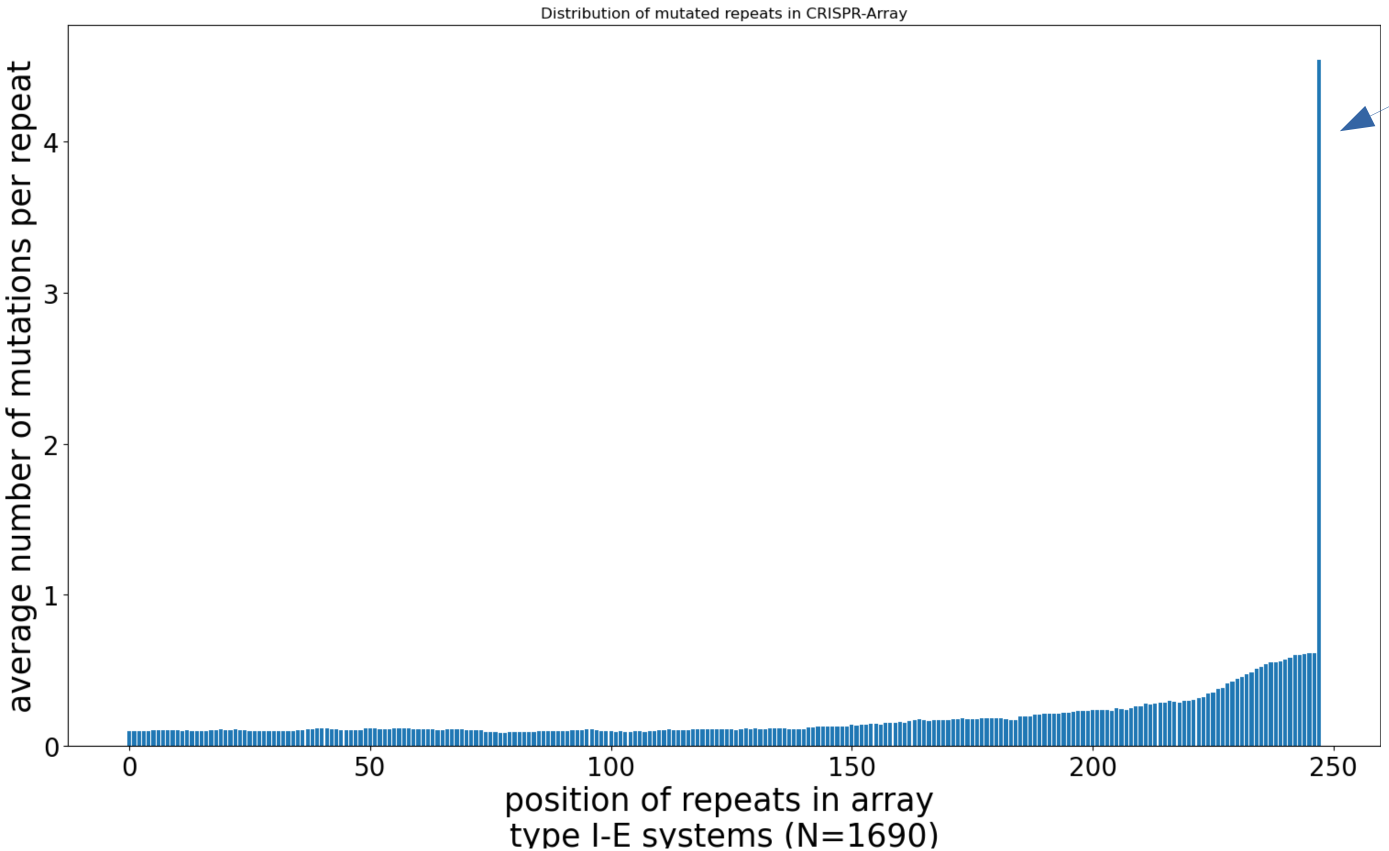


Mutations increase to the end of the array



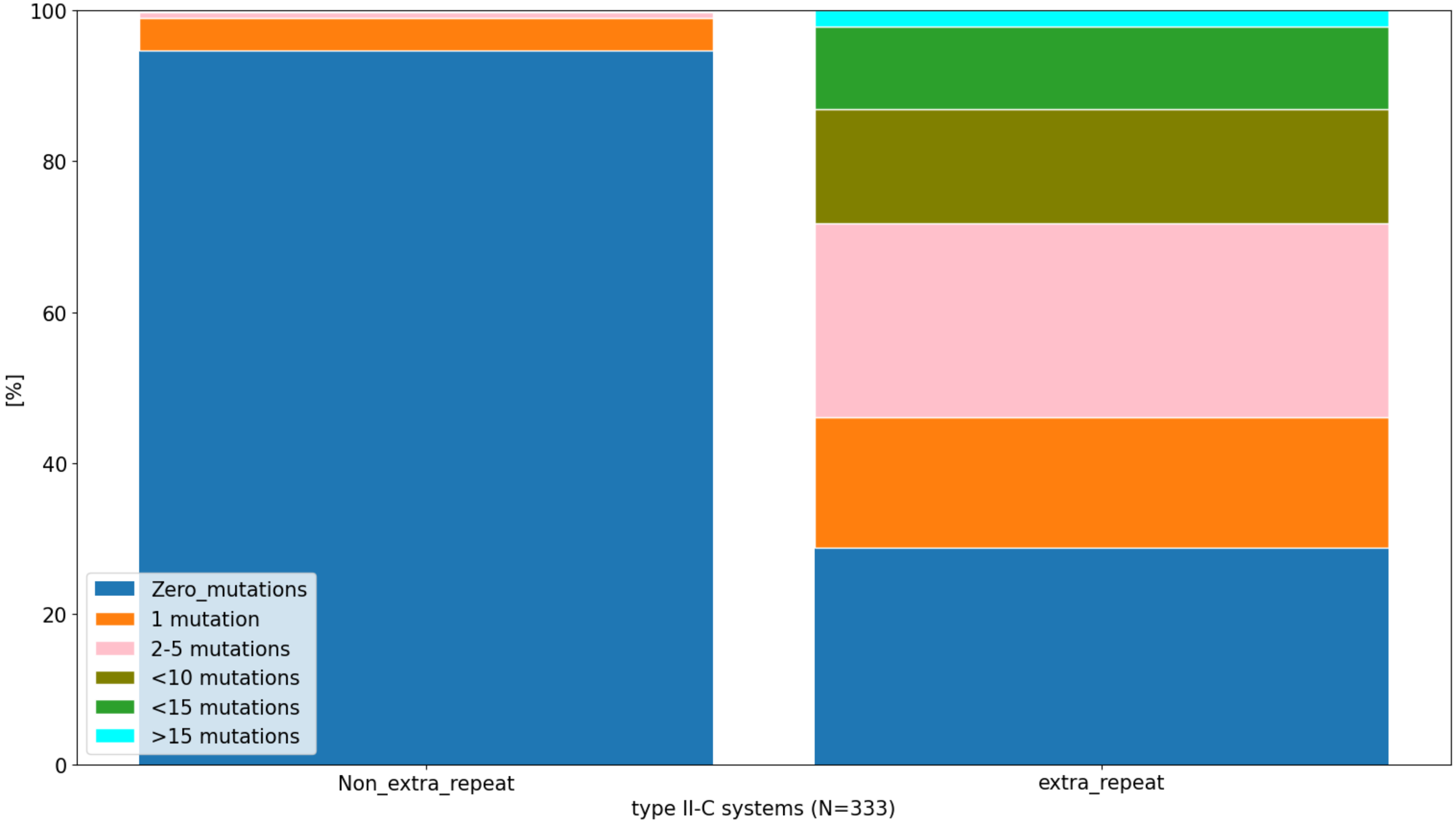
(Without extra repeat)

But the extra repeat is shows more mutations than expected



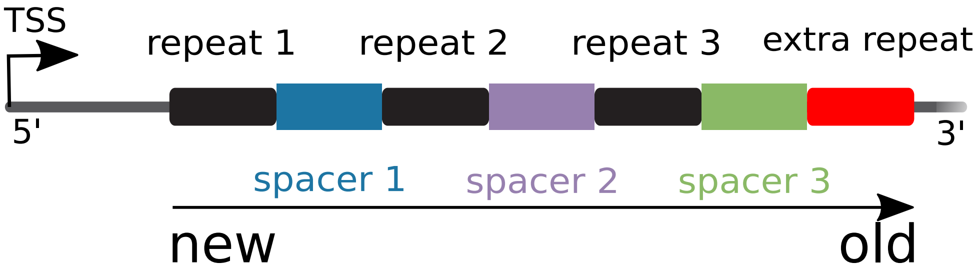
Extra repeat

Similar patterns in sub-type II-C

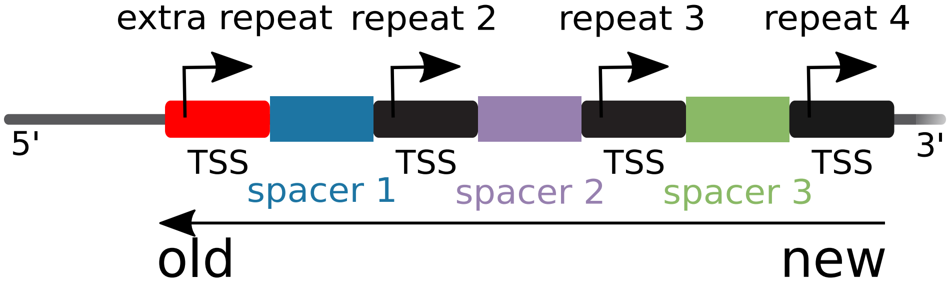


The extra repeat is the oldest repeat

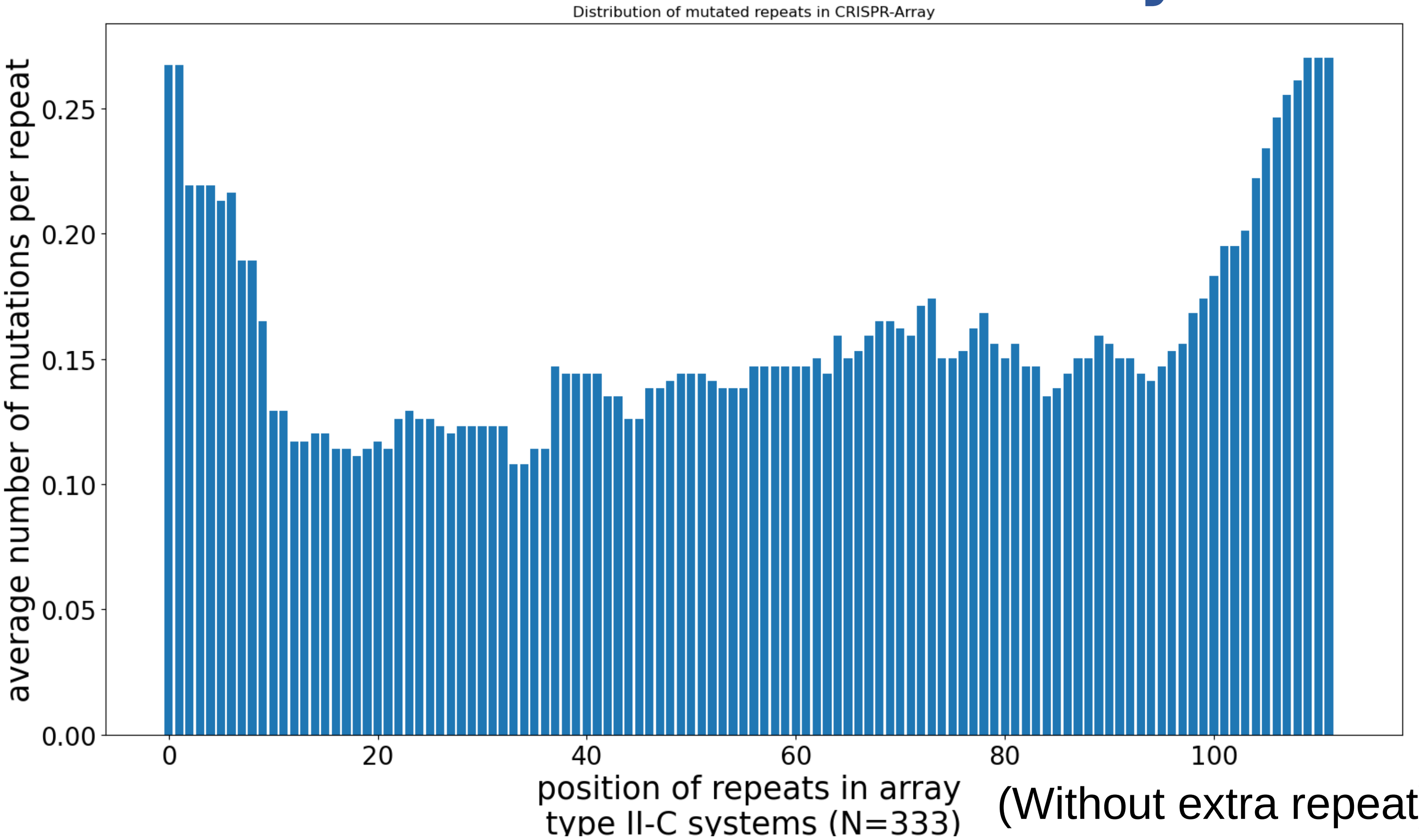
Sub-type I-E



Sub-type II-C

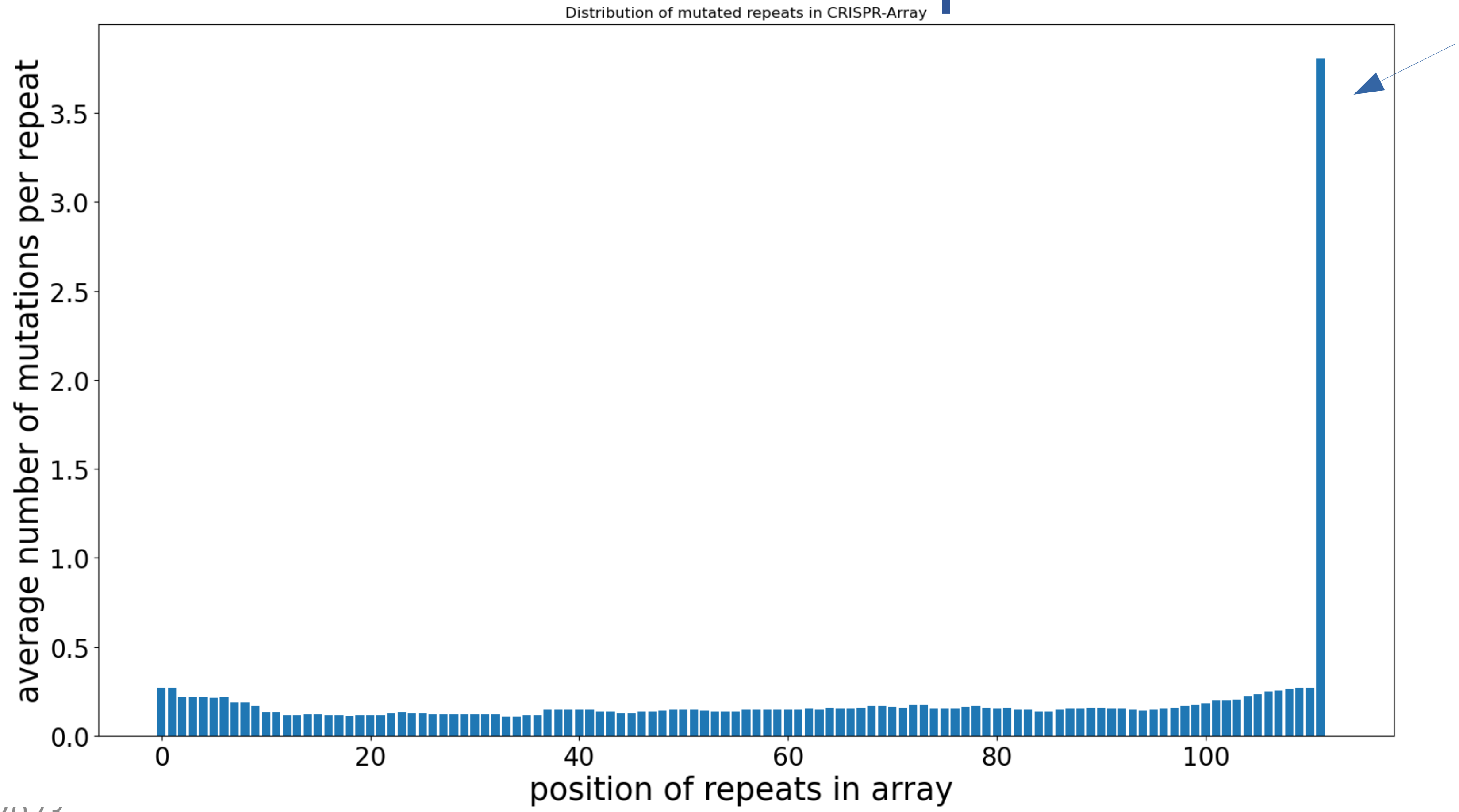


Mutations seems to occur at start and end of the array



(Without extra repeat)

Again the extra repeat is shows more mutations than expected



Extra repeat

What about CRISPR-arrays where the extra repeat is not disabled?

- We hypothesize that some extra repeats are used to express crRNA-like regulators of genes
- We identified promising candidates but we’re waiting for experimental insight
- We’re working on other sub-types

Leipzig University
Zasha Weinberg

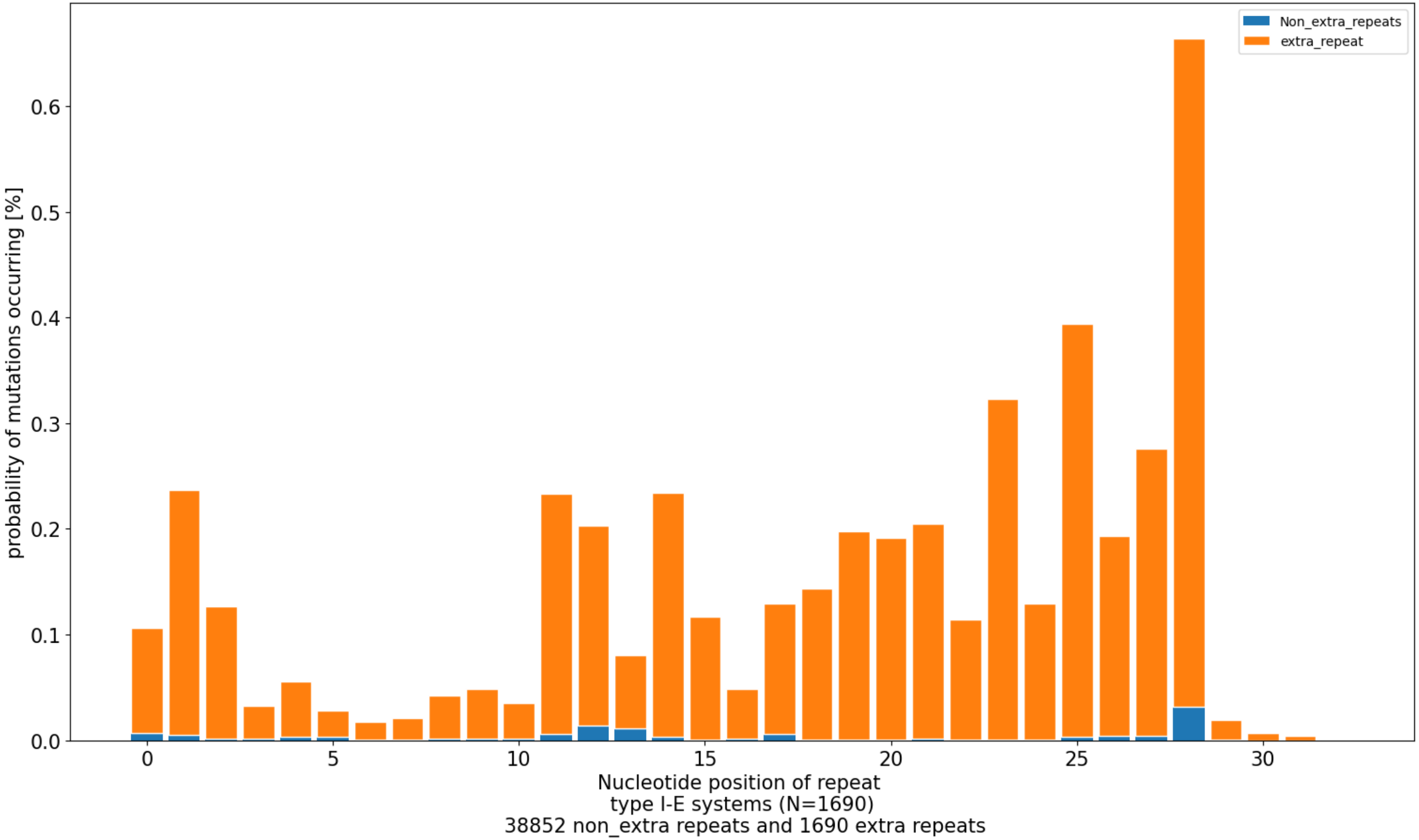
University of Freiburg
Omer Alkhnbashi
Alexander Mitrofanov

Helmholtz Centre For Infection Research
Chase Beisel
Anzhela Migur

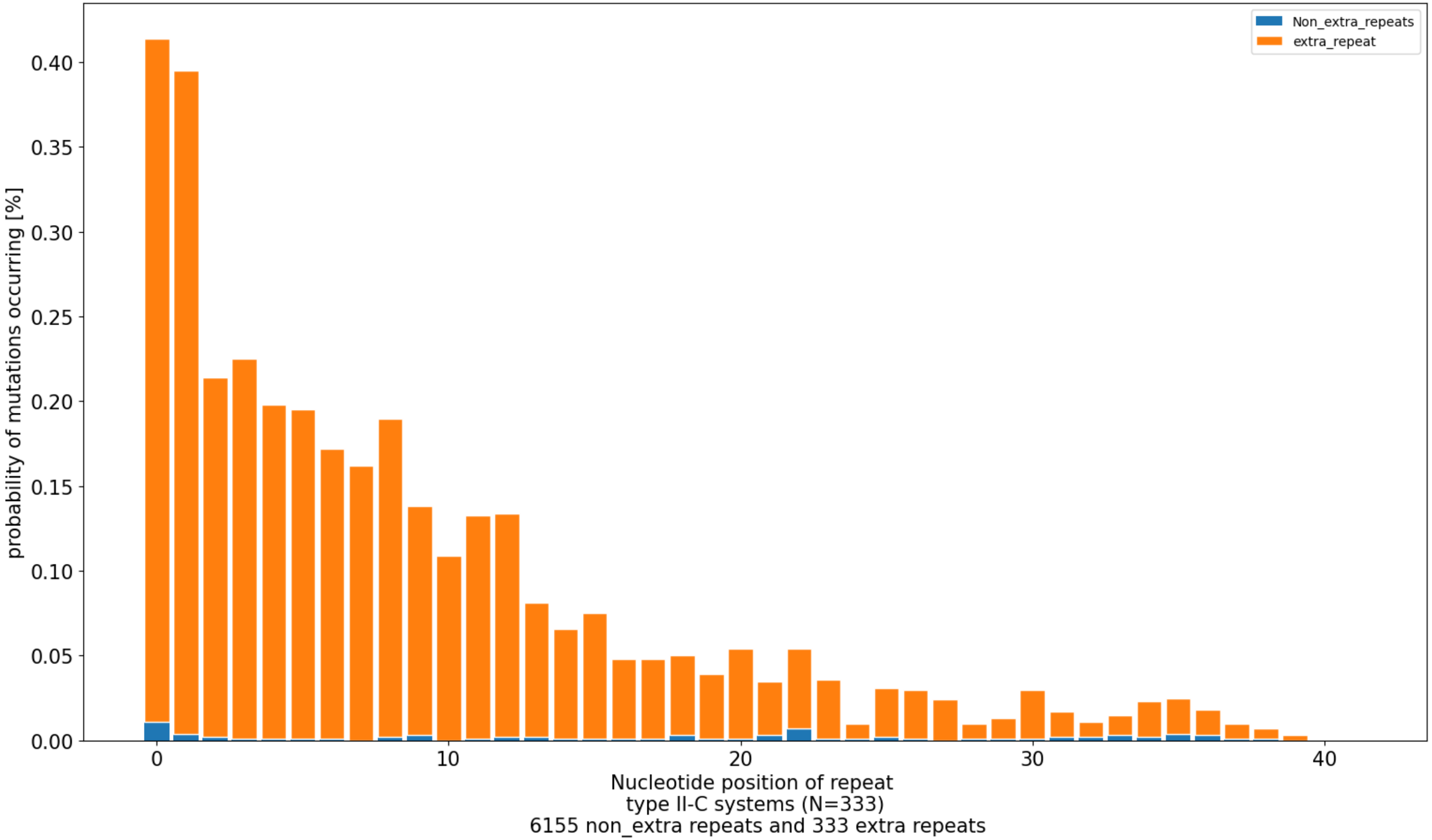
Deutsche Forschungsgemeinschaft

UNIVERSITÄT LEIPZIG

The 3’ end of the extra repeat is more mutated



Mutations are at the 5’ end of the repeat



Training and test sets

- Training Set: 38 systems (max identity=70%)
- Test Set: 30 systems (max identity=70%)
  - Max identity to training set: 50%

Score for leader-repeat interaction

- Pr(8 consecutive spanning base pairs | leader+extra repeat sequence)
  - At thermodynamic equilibrium
  - Calculated with samples from Boltzmann Distribution
- 8 was determined based on training data set

Statistical significance

- p-value:
  - Null model: leader sequence is random
    - (preserves dinucleotide frequencies)
- Combined p-values of multiple CRISPR-Cas systems
  - Fisher‘s Method

→ p-value (on test set): 3×10^{-6} (II-A), 0.009 (VI-B)

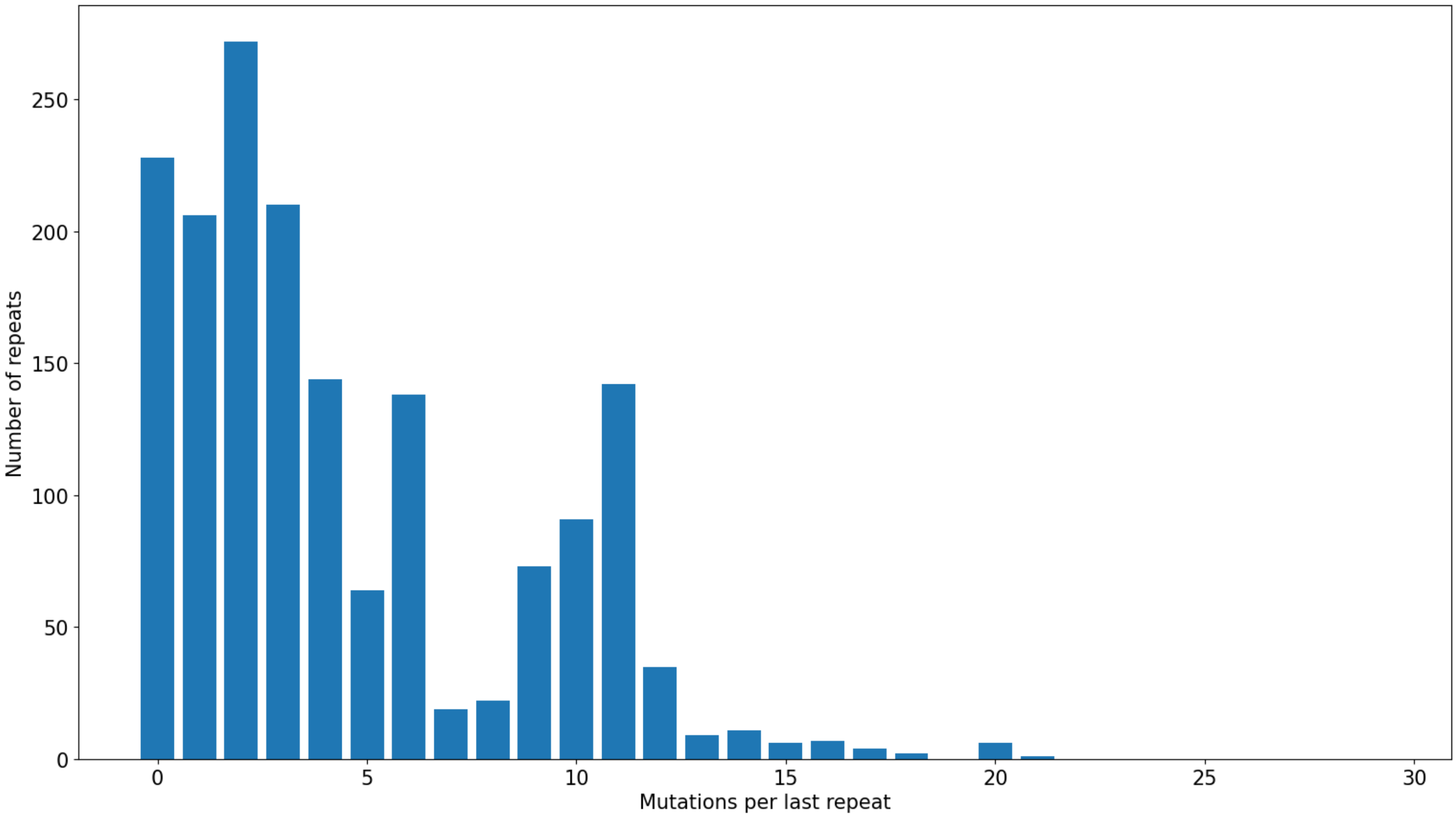
Wait: Fisher's Method assumes independence

- We clustered at 70% identity
    - <70% : too few systems
- Simulations suggest 70% is okay in practice
    - Method
        - Randomly shuffle columns in alignments of leaders
        - Calculate Fisher‘s Method p-values
        - p-value was less than 5% about 5% of the time
    - Also: shuffled in blocks of 30 alignment columns

The extra repeat may be mutated to such an extent that it cannot be detected

- BLAST to identify possible extra repeats
- p-value:
  - Comparison of BLAST scores
  - Null model: leader sequence is random
- p-value threshold:
  - 0.05 → 1% false-positives
  - 0.01 → 0.2% false-positives

Most extra repeats have mutations


Number of repeats
Mutations per last repeat