Challenges and Inconsistencies in Type II CRISPR-Associated Nuclease Subtype Classification

Ariel Gispan, Nurit Meron, Idit Buch, Anat London Drori, Rachel Diamant

EmendoBio

emendo bio — NEXT GENERATION CRISPR MAKING ANY GENE TARGETABLE

INTRODUCTION

CRISPR-associated nucleases were first found and classified as a component of the bacterial immune defense system, designed to combat foreign DNA[1,2]. The discovery of SpCas9, and its repurpose as a genome editing tool led the path for the discovery of additional distinct nucleases with diverse properties that are employed in a variety of applications. Initial categorization of CRISPR-related nucleases was based on a narrow pool of nucleases from limited origins and could not anticipate the heterogeneity of nucleases that is known today. The current subtype classification is based on several methodologies[3,4], however, no differential weight was assigned to each classification method, and in several cases, the discrepancy between the methods resulted in subjective and/or arbitrary classifications. In the current study we employed three different commonly used classification methods of type II nucleases, namely: Loci architecture, Cas9 phylogeny and Cas1 phylogeny for analysis of a large-scale nuclease database. For each classification method, the distribution of nucleases by subtype was studied and the agreement between the methods was measured. HMM profiles of the HNH motif were built iteratively and the matching of the nucleases to each profile was determined. Finally, the relationship between subtype classification, HNH profile and nuclease function was also investigated. Out of the 9520 nucleases analyzed, about 30% were inconsistently classified. A similarity matrix revealed a diverse correlation between the methods used. In some cases, the nucleases did not fit with any of the established subtype classifications and showed a novel and unique loci architecture pattern. Nucleases of the same subtype showed preference to diverse HNH profiles. Some nucleases of the same subtype classification and the same HNH profile showed diverse activity in mammalian cells. Overall, these findings demonstrate the challenges in CRISPR-associated nuclease classification. They also question the present paradigm of affiliating nucleases into distinct allegedly homogenous groups with shared properties and functions. Accordingly, we propose that newly discovered and/or engineered nucleases should be carefully characterized prior to being confined with existing classifications.

METHODS

Data Collection

Bacterial assembled samples were downloaded from Mgnify[5] and JGI[6,7]. CRISPRCasFinder[8] was used to detect CRISPRs and CAS genes. Proteins within 3Kb of the CRISPR array, larger than 800 amino acids with RuvC and HNH domains (HHsuite hhalign[9]) were selected, resulting in 9520 nuclease containing contigs.

Loci Architecture Analysis

A list of known type-II subtype gene patterns (including order and direction) was constructed. Prodigal[10] was used to predict protein ORF. Each protein was compared to a set of CRISPR gene profiles from CRISPRCasTyper[11]. New patterns and genes were added to the pattern list and the profiles set, respectively, resulting in the mapping of each contig to a specific loci architecture pattern.

Phylogeny Analysis

Phylogenetic trees for both Cas9 annotated protein sequences[4] and Cas1 sequences were constructed, one tree per known subtype. A database of HMM profiles was created using HHsuite, one profile per clade. Each protein was compared to this database and assigned to the nearest clade, resulting in a new annotation and a score based on the distance to the nearest clade, as calculated by Makarova *et al.*[4]

HNH Analysis

Multiple sequence alignment was performed on a sample of 150 nuclease sequences with muscle[12]. The sub-sequence aligned to the SpCas9 HNH conserved domain (positions 837 – 864) was extracted for each sequence. Then, domain clusters were generated with K-means algorithm, using the HMM profile (HHsuite) as group centroid. The value of K=8 was selected with the elbow method, over the mean match score of all the sequences.

In vitro depletion assay by TXTL

Depletion of PAM sequences *in vitro* was followed as described by Maxwell *et al.*[13]

Activity in human cells on endogenous genomic targets:

Nucleases (NUCs) were assayed for their ability to edit specific genomic locations in human cells. To this end, each nuclease was transfected into HeLa cells together with sgRNA designed to target specific location in the human genome. NGS analysis was used to calculate the percentage of editing events in each target site.

RESULTS

Nuclease distribution is different for each subtype classification method

Cas9 phylogenetic classification

Subtype	Count	Percentage
II-A	3651	38.3%
II-B	123	1.3%
II-C	5746	60.4%

Cas1 phylogenetic classification

Subtype	Count	Percentage
II-A	4001	42.0%
II-B	106	1.1%
II-C	4655	48.9%
Missing Cas1	758	8.0%

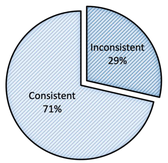
Loci architecture classification

Loci architecture	Count	Percentage	
Cas9,Cas1,Cas2	3948	41.5%	II-C
Cas9,Cas1,Cas2,Csn2	3498	36.7%	II-A
Cas9	541	5.7%	
Cas9,Cas1	316	3.3%	
Cas9,RhuM,Cas1,Cas2	168	1.8%	
Cas9,PAR,Cas1,Cas2	124	1.3%	
Cas9,Cas2	97	1.0%	
Cas9,Cas1,Cas2,Cas4	74	0.8%	II-B
Cas9,Csn2	72	0.8%	
Cas9,Cas1,Csn2	69	0.7%	
Cas9,Cas2R,Cas1R	60	0.6%	
Cas9,Cas2,Cas1R	52	0.5%	
Cas1,Cas2,Csn2,Cas9	29	0.3%	
Cas9,Cas2,Csn2	21	0.2%	
Other	**451**	**4.7%**	

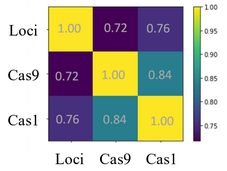
- In some cases, the nucleases did not fit with any of the established subtype classifications and showed novel loci architecture patterns.
- Some loci architecture patterns included genes that have not been previously reported in the context of CRISPR operon, such as RhuM and PAR.

Inconsistencies in subtype classification

Total nucleases analyzed

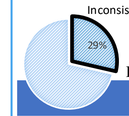


Similarity matrix



Subtype distribution in consistently classified nucleases

Subtype	Count	Percentage
II-A	2966	43.6%
II-B	74	1.1%
II-C	3759	55.3%
total	6799	100%



Leading patterns of inconsistent nuclease classification by the three methods

Loci architecture	Cas9 classification	Cas1 classification	Count	Percentage
Cas9,Cas1,Cas2,Csn2 (II-A)	II-C	II-A	528	19.4%
Cas9,Cas1	II-C	II-C	228	8.4%
Cas9,RhuM,Cas1,Cas2	II-C	II-C	168	6.2%
Cas9,Cas1,Cas2 (II-C)	II-C	II-A	144	5.3%
Cas9,Par,Cas1,Cas2	II-C	II-C	124	4.6%
Cas9,Cas2R,Cas1R	II-C	II-C	60	2.2%
Cas9,Cas1,Csn2	II-A	II-A	59	2.2%
Cas9,Cas2,Cas1R	II-C	II-C	52	1.9%
Cas9,Cas1,Cas2 (II-C)	II-A	II-A	39	1.4%
Cas9,Cas1	II-C	II-A	37	1.4%
Cas9,Cas1	II-A	II-A	29	1.1%
Cas1,Cas2,Csn2,Cas9	II-C	II-A	22	0.8%
other			1231	45.2%
Total			2721	**100%**

Examples of inconsistently classified nucleases

Accession #	Loci architecture	Cas1 classification	Cas9 classification
WP_009293010.1	Cas9,RhuM,Cas1,Cas2	II-C	II-C
ACD99347.1	Cas9,Cas1	II-C	II-C
WP_007837560.1	Cas9,RhuM,Cas1,Cas2	II-C	II-C
WP_012962169.1	Cas9,Cas1,Cas2,Csn2 (II-A)	II-A	II-C
WP_013073784.1	Cas9,Cas1,None,Cas2	II-C	II-C
WP_005791619.1	Cas9,RhuM,Cas1,Cas2	II-C	II-C
WP_003065552.1	Cas9,Cas1,Cas2,Cas2	II-A	II-A
WP_013362995.1	Cas9,Cas1,Csn2	II-A	II-A
WP_013852048.1	Cas9,Cas1,Cas2,Csn2 (II-A)	II-A	II-C
WP_004292911.1	Cas9,Par,Cas1,Cas2	II-C	II-C

Examples of well-known inconsistently classified nucleases

Sample	Bacteria	Loci architecture	Loci architecture	Cas1	Cas1 distance	Cas9	Cas9 distance
NZ_CUFQ01000030	Staphylococcus_aureus	Cas9,Cas1,Cas2,Csn2	II-A	II-A	0.52	II-C	0.94
ARMAN1_contig	ARMAN1	Cas9,Cas4,Cas2,Cas1	II-C-2	II-B	0.94	II-C	2.52
Nitrospiraceae_contig	Nitrospiraceae_II-D	Cas9,Cas1,Cas2,Cas4	II-B	II-B	0.84	II-C	2.44
Streptococcus_thermophilus_LMD-9	Streptococcus_thermophilus (short)	Cas9,Cas1,Cas2,Csn2	II-A	II-A	0.49	II-C	0.72

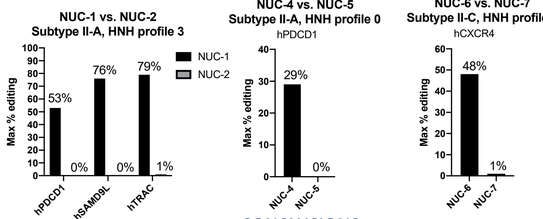
- About 30% of the nucleases analyzed were inconsistently classified by the three subtype classification methods.
- A similarity matrix revealed a diverse correlation between the methods used.

Nucleases of the same subtype show diverse HNH profiles

HNH profile analysis per subtype in consistently classified nucleases

Profile	Subtype II-A	Subtype II-B	Subtype II-C
Cluster_0	941 (31.7%)	0	1 (0.03%)
Cluster_1	0	0	276 (7.3%)
Cluster_2	0	74 (100%)	122 (3.2%)
Cluster_3	1024 (34.5%)	0	1 (0.03%)
Cluster_4	0	0	1791 (47.6%)
Cluster_5	316 (10.7%)	0	1 (0.03%)
Cluster_6	45 (1.5%)	0	1154 (30.7%)
Cluster_7	624 (21%)	0	0
None	16 (0.5%)	0	413 (11%)

Nucleases of the same subtype and HNH profile show different activity in mammalian cells



CONCLUSIONS

- Our findings demonstrate the challenges in type-II CRISPR-associated nuclease subtype classification.
- They also question the present paradigm of affiliating nucleases into distinct homogenous groups with allegedly shared properties and functions.
- Predictability of nuclease activity in mammalian cells is currently very limited and could not be anticipated based on current classifications.
- Accordingly, we propose that newly discovered and/or engineered nucleases should be carefully characterized prior to being confined with existing classifications.

REFERENCES

1. Barrangou, R., *et al.* (2007). Science.
2. Mojica, F.J.M., *et al.* (2000). Mol. Microbiol.
3. Makarova, K.S., *et al.*, (2015). Nat. Rev. Microbiol.
4. Makarova, K.S., *et al.* (2020). Nat. Rev. Microbiol.
5. Mitchell, A.L., *et al.* (2020). Nucleic Acids Res.
6. Chen, I.-M.A., *et al.* (2021). Nucleic Acids Res.
7. Mukherjee, S., *et al.* (2021). Nucleic Acids Res.
8. CRISPR-CAS++ https://crisprcas.i2bc.paris-saclay.fr/.
9. Steinegger, M., *et al.* (2019). BMC Bioinformatics.
10. Hyatt, D., *et al.* (2010). BMC Bioinformatics.
11. Russel, J., *et al.* (2020). CRISPR J.
12. Edgar, R.C. *et al.* (2004). Nucleic Acids Res.
13. Maxwell, C.S., *et al.* (2018). Methods.