Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Astringency Reduction by Anthocyanidin Reductase Knockout: CRISPR Guide Design for Sorghum and Grape

Priyanca Ford

scovillesplice@gmail.com

Scoville Splice LLC https://orcid.org/0000-0003-0395-1752

Research Article

Keywords: anthocyanidin reductase, astringency, proanthocyanidins, CRISPR-Cas9 guide design, Vitis vinifera, Sorghum bicolor, gene array, Iyengar Flavor Index

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11875891/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

SCOVILLESPLICE.FARM

SCOVILLE SPLICE RESEARCH SERIES

Astringency Reduction by Anthocyanidin Reductase Knockout: CRISPR Guide Design for Sorghum and Grape

Computed CRISPR-Cas9 guide designs that cut either side of the catalytic Tyr codon in the grape ANR gene and in every ANR copy of a sorghum chromosome 6 array.

P.I. Ford

Scoville Splice LLC · ORCID: 0000-0003-0395-1752

DOI: 10.5281/zenodo.24511334 License: CC BY 4.0

Patents: US 64/172,827; US 64/188,197; US 64/192,062; US 64/199,105 Figshare: figshare/35015818

Harvard Dataverse: DVN/VPB0PK OSF: osf.io/fv58q Hugging Face: capsaicinoid-data GitHub: paper-08

Website: astringency-reduction

Package record. 179 claims traced to a file or a cited source (135 computed, 26 definition, 3 design assumption, 11 design target, 4 literature value). 16 tables, 4 figures, 43 data files. Data fingerprint `e01499f1c3db`. Rerun every result with `bash run_all.sh`.



Grape cross-section. Scoville Splice LLC image library.

ABSTRACT

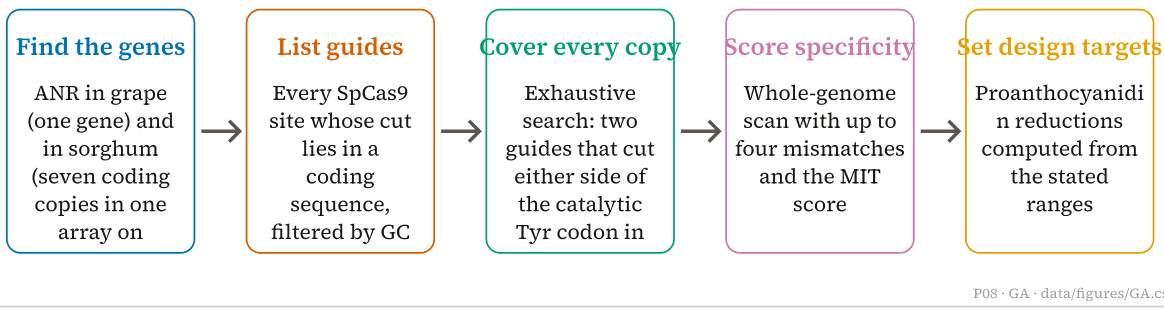
Condensed tannins (proanthocyanidins) bind salivary proteins and give astringency to wine grapes and sorghum grain. Anthocyanidin reductase (ANR) makes the flavan-3-ol units of these tannins. This paper computes CRISPR-Cas9 guide designs that knock out ANR in grape and in sorghum. Grape has one ANR gene, *VvANR* (GeneID 100232981), on the unplaced scaffold NW_003724316.1. Two guides cut either side of its catalytic Tyr codon and remove 80 coding bases. Sorghum carries 7 coding ANR copies in one array on chromosome 6. An exhaustive search over 764 candidate guides shows that 6 guides in three pairs are the fewest that cut either side of the catalytic Tyr codon in every copy. Two added guides cover the ANR-like transcript LOC8074755, which gives 8 guides in four groups. The eight guides have perfect sites only inside the array, and their MIT scores outside the array run from 89.7 to 99.8. Design targets set proanthocyanidin reductions of 83.33% to 98.33%. All results are computed. No experimental data are reported in this paper.

Keywords: anthocyanidin reductase · astringency · proanthocyanidins · CRISPR-Cas9 guide design · Vitis vinifera · Sorghum bicolor · gene array · Iyengar Flavor Index

KEY POINTS

- *VvANR* (GeneID 100232981) lies on the unplaced scaffold NW_003724316.1; two guides cut either side of its catalytic Tyr codon.
- Sorghum ANR is an array of 7 coding copies on chromosome 6; 6 guides in three pairs are the fewest that cut every copy.
- Two added guides, SbANR-G10 and SbANR-G11, cut the ANR-like transcript LOC8074755 and give 8 guides in four groups.
- The MIT scores of the eight array guides outside the array run from 89.7 to 99.8.
- Design targets set proanthocyanidin reductions of 83.33% to 98.33%; all results are computed.

Marks under numbers: computed · design target · design assumption · literature value · definition. The evidence table in Appendix A lists the source file of every one.



Graphical abstract. The design in five steps: the ANR genes are found, guides are listed, a search covers every copy, specificity is scored, and design targets are set. Drawn from the steps file. Data: `data/figures/GA.csv`; script: `code/03_array.py`.

1. Introduction

Condensed tannins, also called proanthocyanidins (PAs), bind proline-rich proteins strongly (Hagerman and Butler, 1981), and their binding of salivary proteins gives astringency to wine grapes and sorghum products. Anthocyanidin reductase (ANR)

turns anthocyanidins into 2,3-cis-flavan-3-ols, which supply the units of PAs (Xie et al., 2003). In Arabidopsis the ANR gene is *BAN* (Devic et al., 1999). In sorghum, tannin production also depends on TAN1, a WD40-repeat protein; natural non-tan alleles carry a 1-bp deletion (*tan1-a*) or a 10-bp insertion (*tan1-b*), both frameshifts (Wu et al., 2012). The sorghum ANR enzyme used as a reference is encoded by Sobic.006G227200 (Lewis et al., 2023).

This paper computes guide designs for ANR knockout in grape (*Vitis vinifera*) and sorghum (*Sorghum bicolor*). Each design uses two guides that cut either side of the catalytic Tyr codon, so the piece between the cuts is removed. The paper reports the ANR gene models, the guides with their genome-wide specificity, a search for the fewest guides that cover every ANR copy of sorghum, an added design for the ANR-like transcript LOC8074755, two guides for the TAN1 gene, and design targets for PA reduction. Sorghum ANR is an array of copies, so the guide set that covers all of them is the main computed result.

Astringency and bitterness are two axes of the Iyengar Flavor Index (IFI; 10.5281/zenodo.24511320). The IFI scores one compound against one reference receptor assay. Its astringency axis (IFI-A) uses tannic acid at 25.0 µM as the reference. Its bitterness-suppression axis (IFI-B) uses GIV3727 at the bitter receptor TAS2R31, with a reference of 7.15 µM, the mean of 6.4 and 7.9 µM (Slack et al., 2010). This paper assigns no IFI score to any compound; the axes name what the tannin reduction is aimed at.

2. Methods

2.1 Computational basis

Gene models are the RefSeq annotations of the grape assembly GCF_000003745.3 (12X; Jaillon et al., 2007) and the sorghum assembly GCF_000003195.3 (BTx623, NCBIv3; Paterson et al., 2009). Gene names in this paper are our names with the RefSeq locus in every mention (`inputs/gene_names.csv`); protein sequences are the UniProt entries of `inputs/input_sources.csv`. Steps 1 to 4 run on Python 3.9 or newer with the standard library. The research scripts ran on Python 3.12.16 with Biopython 1.88 and NumPy 2.5.5, and the scanner was compiled with `cc -O2` (`code/LIBRARIES.txt`).

2.2 Protein identity and the catalytic motif

Proteins were aligned globally with BLOSUM62, gap open -10 and gap extend -0.5. Identity over alignment columns is the identical residues divided by all alignment columns. Identity over the shorter sequence is the identical residues divided by the

length of the shorter protein. The catalytic pair is the YxxxK motif (Tyr and Lys four residues apart). In every target the catalytic Tyr is the Tyr of the YxxxK motif that aligns with Tyr168 of SbANR4 (UniProt A0A1B6PNE4).

2.3 The sorghum array

All annotated genes between 57,100,000 and 57,300,000 on NC_012875.2 (the array window) were listed from the RefSeq annotation. A gene was called ANR-type when its product name is anthocyanidin reductase or when its coding model is at least 40% identical to SbANR4 (LOC8066842 copy A). Non-coding features were translated in three frames and aligned locally to SbANR4 (LOC8066842 copy A).

2.4 Candidate guides

Candidates are SpCas9 protospacers of 20 bases with an NGG PAM on either strand whose cut lies in a coding sequence, with GC content of 40 to 70%, no TTTT and no run of 5 identical bases. The cut lies 3 bases from the PAM. The cut position is the first plus-strand base after the cut; its coding index counts from the first base of the start codon, and exons are numbered within each copy.

2.5 Scan and scoring

Each guide was scanned against the whole assembly with our own C scanner (`research_record/scripts/guide_design/scan.c`): strict NGG PAM, both strands, up to 4 mismatches. The MIT score of Hsu et al. (2013) was implemented as follows. The score of one site is the product of $(1 - w)$ over its mismatch positions, where w is the position weight (Table S8), times $1 / (((19 - \bar{d}) / 19) \times 4 + 1)$ for sites with more than one mismatch, where $\bar{d}$ is the mean distance between mismatches, times $1 / n^2$ for n mismatches, times 100. The guide score is 100 / (100 + the sum of the site scores) × 100. The MIT score outside the array sums only the sites outside the array window; the score over all sites sums every imperfect site, including those in the neighbouring copies. Perfect sites are read with their PAM from the genome.

2.6 Flanking pairs and the cover search

Two guides flank a target when both have a perfect site in its coding sequence, one cutting before the first base of the catalytic Tyr codon and the other after its last base, at most 700 coding bases apart. The exhaustive search starts from every set of guides that have perfect sites in more than one copy, completes each set with copy-specific guides and prunes with a valid bound, so a set size is excluded only after the whole space is searched (`research_record/scripts/guide_design/s3b_exact_flank.py`). Complete

six-guide sets were then counted and ranked by the lowest MIT score outside the array among their guides.

2.7 Consequence of a deletion

The two cuts are joined without added or lost bases. The coding bases between the cuts are removed, the edited sequence is translated from its first base, and the residues before the first stop codon are counted. When the shifted frame of SbANR7 (LOC8066846) runs past the end of its coding sequence, the stored sequence up to the end of its annotated mRNA is used (`inputs/guide_design/cds/`).

2.8 The eighth target

LOC8074755 is a transcribed pseudogene (RefSeq XR_002454302.1) with 11 exons and a spliced length of 2,359 nt. Its translation holds two ANR-like parts: part 1 (frame 0, spliced bases 19 to 1440) and part 2 (frame 2, bases 1302 to 2234). The same flanking rule was applied to each part. The fewest added guides were found with an integer programme (HiGHS) and checked by brute force over every single guide and every pair (`research_record/scripts/eighth_target/`).

2.9 Design targets

The reduction of a tissue is 1 − target / wild-type value, taken at the lowest and the highest assumed wild-type value. Wild-type ranges are design assumptions with no source; targets are design targets (`inputs/design_targets.csv`). The sensitivity analysis moves the wild-type value over its range with the target fixed (`code/04_reductions.py`).

3. Results

3.1 The ANR genes of grape and sorghum

Grape has one RefSeq ANR gene, *VvANR* (GeneID 100232981; NP_001267885.1), with 6 exons, a coding sequence of 1,017 bases and a protein of 338 residues. It lies on the unplaced scaffold NW_003724316.1 of the 12X assembly, a scaffold of 75,112 bases. The translated coding sequence matches the RefSeq protein in 334 of 338 alignment columns. Table 1 gives the identities: VvANR (GeneID 100232981) is 63.8% identical to BAN over alignment columns and 64.2% over the shorter sequence. SbANR4 (LOC8066842 copy A) is 44.4% and 45.3% identical to BAN, and 48.1% and 48.8%

identical to VvANR (GeneID 100232981). The catalytic YxxxK motif sits at Tyr168–Lys172 in VvANR (GeneID 100232981), Tyr169–Lys173 in BAN and Tyr168–Lys172 in SbANR4 (LOC8066842 copy A).

Sorghum ANR is an array. The array window holds 26 annotated genes, and 6 of them are ANR-type: SbANR1 (LOC110436259), SbANR2 (LOC8066840), SbANR3 (LOC8066841), the merged RefSeq model LOC8066842, SbANR6 (LOC8066844) and SbANR7 (LOC8066846). The merged model LOC8066842 holds two tandem copies, SbANR4 (copy A) and SbANR5 (copy B), so the array has 7 coding copies (Table 2, Figure 1). SbANR4 (LOC8066842 copy A) is the UniProt entry A0A1B6PNE4 (Sobic.006G227200), identical in 341 of 341 columns. The other six copies are 66.9% to 90.9% identical to it. The SbANR1 (LOC110436259) protein is identical in 345 residues to UniProt A0A1B6PND3, which Lewis et al. (2023) call SbFNR2 and suggest may act as an ANR-like isozyme.

The array also holds 3 non-coding ANR-like features, LOC8066836, LOC8074755 and LOC8074756, whose best reading frames are 82%, 71% and 78% identical to SbANR4 (LOC8066842 copy A) over the aligned residues (Table S5). Five annotated anthocyanidin reductase genes lie outside the array on three other chromosomes; they are 29.2% to 33.0% identical to SbANR4 (LOC8066842 copy A) over alignment columns, below the 40% rule (Table S7). All perfect sites of the eleven array guides and of the two-copy guides lie inside the array window.

Table 1. Protein identity of the anthocyanidin reductases. Global alignment with BLOSUM62 and gap penalties of -10 (open) and -0.5 (extend). Identity is given over all alignment columns and over the length of the shorter sequence. Data: `data/tables/T01.csv`*; script:* `code/01_identity.py`*.*

Pair of proteins	Lengths (aa)	Identical residues	Alignment columns	Identity over alignment columns (%)	Identity over the shorter sequence (%)
VvANR (Q5FB34) and Arabidopsis BAN (Q9SEV0)	338 and 340	217	340	63.8	64.2
SbANR4 (LOC8066842 copy A; A0A1B6PNE4) and Arabidopsis BAN (Q9SEV0)	341 and 340	154	347	44.4	45.3
SbANR4 (LOC8066842 copy A; A0A1B6PNE4) and VvANR (Q5FB34)	341 and 338	165	343	48.1	48.8

Table 2. The seven coding ANR copies of the sorghum array. RefSeq proteins, catalytic Tyr residue (the Tyr of the YxxxK motif that aligns with Tyr168 of SbANR4 (LOC8066842 copy A)), identity to SbANR4 (LOC8066842 copy A) and genomic span. SbANR5 (LOC8066842 copy B) is counted from its own start codon. Data: `data/tables/T02.csv`*; script:* `code/03_array.py`*.*

Copy	RefSeq protein	Length (aa)	Catalytic Tyr	Identical to SbANR4 (LOC8066842 copy A; UniProt A0A1B6PNE4)	Identity (%)	Coding exons	Genomic span on chr6
SbANR1 (LOC110436259)	XP_021318468.1	345	171	254 of 346	73.4	6	57,170,614-57,172,695
SbANR2 (LOC8066840)	XP_021318471.1	340	167	251 of 342	73.4	6	57,185,003-57,186,639
SbANR3 (LOC8066841)	XP_002447157.1	340	166	276 of 342	80.7	6	57,198,509-57,200,190
SbANR4 (LOC8066842 copy A)	XP_002447158.2	341	168	341 of 341	100.0	7	57,202,991-57,206,408
SbANR5 (LOC8066842 copy B)	XP_002447158.2	341	168	310 of 341	90.9	6	57,206,424-57,207,962

Copy	RefSeq protein	Length (aa)	Catalytic Tyr	Identical to SbANR4 (LOC8066842 copy A; UniProt A0A1B6PNE4)	Identity (%)	Coding exons	Genomic span on chr6
SbANR6 (LOC8066844)	XP_021319469.1	343	169	263 of 345	76.2	5	57,210,414-57,211,916
SbANR7 (LOC8066846)	XP_021319468.1	354	179	239 of 357	66.9	5	57,246,237-57,248,002

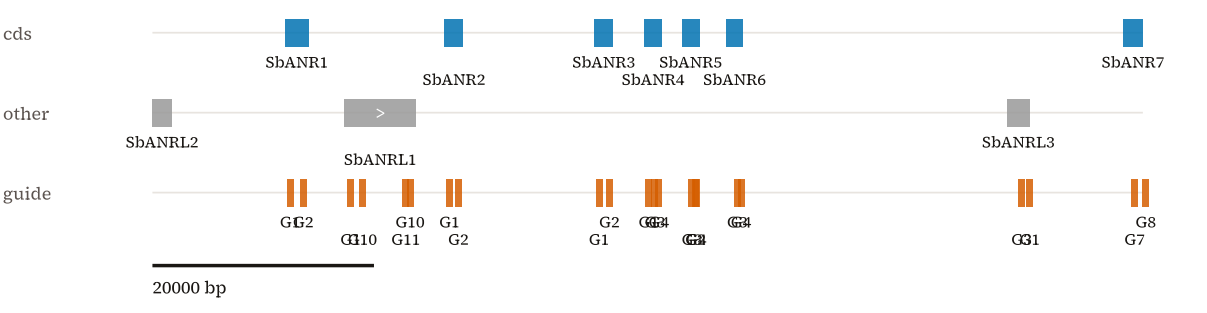


Figure 1. The sorghum chromosome 6 ANR array and the perfect sites of the eight array guides. Coding copies (cds), non-coding features (other) and the perfect sites of guides G1 to G4, G7, G8, G10 and G11 along NC_012875.2 from 57,170,614 to 57,248,132; the scale bar is in base pairs. Coordinates are RefSeq gene coordinates and the first base of each 23-base site from the stored genome scan. Data: `data/figures/F01.csv`; script: `code/03_array.py`.

3.2 Grape guides

Table 3 and Figure 2 list the grape guides. VvANR-gA (`TCAGCTTGATGGGACAGGTC`, PAM TGG) cuts exon 3 at coding index 431, and VvANR-gB (`TCAGCTAGTGTCTTGGAGGC`, PAM AGG) cuts exon 4 at coding index 511. Each guide has 1 perfect site in the genome, and its nearest imperfect sites have four mismatches (6 and 5 sites); the MIT scores are 98.7 and 99.4. The two cuts are 189 genomic bases apart and remove 80 coding bases, which shifts the reading frame from residue 144. The removed piece holds the Tyr168 codon. The edited coding sequence reaches a stop codon after 148 residues, against 338 residues in the unedited sequence.

3.3 Sorghum TAN1 guides

SbTAN1 (LOC8076027; XP_002452749.2) lies on chromosome 4 (NC_012873.2). SbTAN1-gA cuts exon 1 at coding index 91 and SbTAN1-gB cuts exon 1 at coding index 258 (Table 3, Figure 2). The MIT scores are 96.6 and 95.9. The cuts are 167 coding bases apart, so the deletion shifts the reading frame from residue 31, as the natural *tan1* alleles do (Wu et al., 2012). The RefSeq protein XP_002452749.2 and UniProt C5Y0X7 share 310 identical residues.

3.4 What one guide or a few guides reach in the sorghum array

The candidate list holds 764 guides with a cut in the coding sequence of a copy. Of these, 680 have a perfect site in one copy, 73 in two, 7 in three and 4 in four, so the best guides reach 4 of the 7 copies. A copy is cut when two guides have perfect sites either side of its catalytic Tyr codon. Figure 3 shows the best result for each number of guides: 4 copies for two or three guides, 6 for four or five guides, and all 7 for 6 guides. The exhaustive search covers every set of three, four and five guides (1, 1,511 and 51,558 nodes) and reaches its first complete sets at size six: 1,540 optimal cores after 6,751,183 nodes. The cores complete to 124,186 complete six-guide sets (174,348 without the cap on the distance between the cuts). If each copy needs only one perfect site, 3 guides are enough (4,841 sets), but then most copies have a single cut and no designed deletion. The counts are in Table S10.

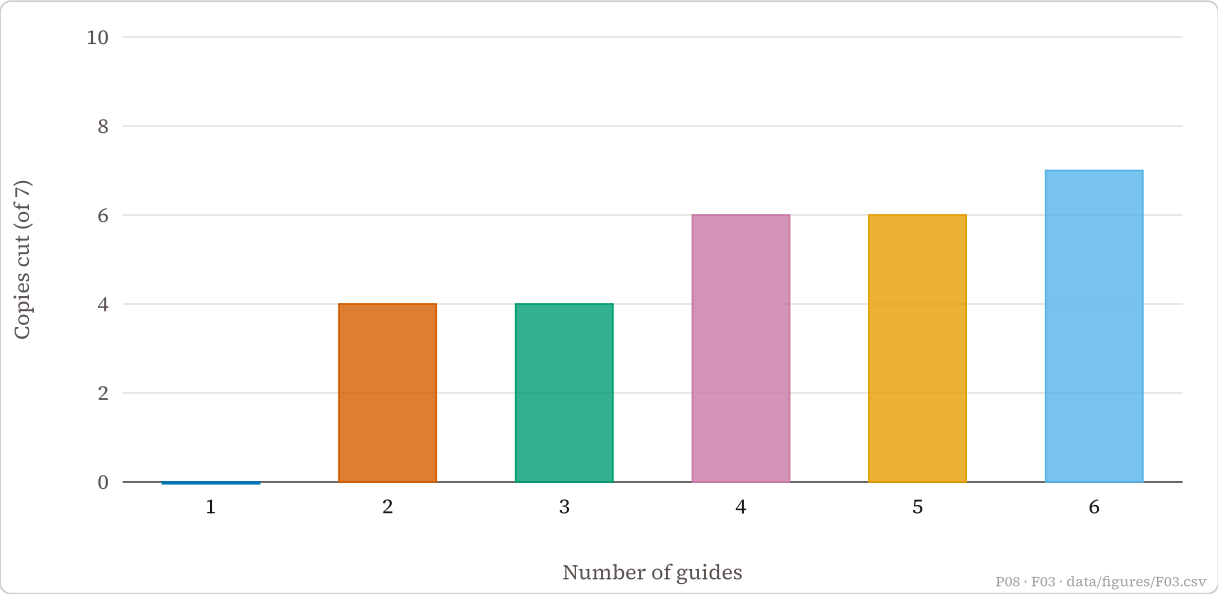


Figure 3. Copies of the sorghum array cut by the best set of k guides. The largest number of the seven coding copies in which two chosen guides both have a perfect site either side of the catalytic Tyr codon, for the best set of 1 to 6 guides. Exhaustive search over all 764 candidate guides. Data: `data/figures/F03.csv`; script: `code/03_array.py`.

3.5 The six array guides

The six guides form three pairs (Table 4, Table 5). SbANR-G1 and SbANR-G2 flank SbANR1 (LOC110436259), SbANR2 (LOC8066840) and SbANR3 (LOC8066841): they remove 471, 468 and 468 coding bases in frame, and the edited sequences reach a stop codon after 188, 184 and 184 residues. SbANR-G3 and SbANR-G4 flank SbANR4 (LOC8066842 copy A), SbANR5 (LOC8066842 copy B) and SbANR6 (LOC8066844): they remove 271, 271 and 268 coding bases, which shifts the reading frame, and the edited sequences reach a stop codon after 197, 197 and 139 residues. SbANR-G7 and SbANR-G8 flank SbANR7

(LOC8066846): they remove 514 coding bases, and the shifted frame reaches its first stop codon after 207 residues, inside the annotated 3′ UTR. The pair SbANR-G7 with SbANR-G9 is an alternative for SbANR7 (LOC8066846): it removes 27 coding bases in frame (Table S1).

Every perfect site of these guides lies inside the array window: the six guides have 19 perfect sites, spread over 77,225 bases (Table S2). Outside the array the nearest imperfect sites of these guides have three or four mismatches, and the sites with three mismatches number 1 for SbANR-G2, 5 for SbANR-G3, 1 for SbANR-G4 and 1 for SbANR-G8. SbANR-G3 has the lowest MIT score outside the array (89.7); the highest value of that lowest score over all complete six-guide sets is also 89.7, reached by 5,136 sets. SbANR-G1 has perfect sites in the non-coding features LOC8074755 and LOC8074756, and SbANR-G3 in LOC8074756 (Table S2, Table S4).

Table 3. Guides for grape ANR, sorghum TAN1 and the two-copy option. Perfect sites, sites at 1 to 4 mismatches outside the chromosome 6 array window, MIT score and cut positions. The cut position is the coding index of the first base after the cut; exons are numbered within each copy. Data: `data/tables/T03.csv`*; script:* `code/02_guides.py`.

Guide	Target	Protospacer and PAM	Perfect sites	Sites at 1 / 2 / 3 / 4 mismatches	MIT score	Cut (first base after the cut)
VvANR-gA	grape ANR, NP_001267885.1, scaffold NW_003724316.1	TCAGCTTGATGGGACAGGTC TGG	1	0 / 0 / 0 / 6	98.7	VvANR (GeneID 100232981) exon 3, base 431
VvANR-gB	grape ANR, NP_001267885.1, scaffold NW_003724316.1	TCAGCTAGTGTCTTGGAGGC AGG	1	0 / 0 / 0 / 5	99.4	VvANR (GeneID 100232981) exon 4, base 511
SbTAN1-gA	sorghum TAN1 (LOC8076027), XP_002452749.2, chr4	GCGAGCGCGTAGATCGAGTG CGG	1	0 / 0 / 0 / 22	96.6	SbTAN1 (LOC8076027) exon 1, base 91
SbTAN1-gB	sorghum TAN1 (LOC8076027), XP_002452749.2, chr4	TGAACTGGAGCTTGGTGGGT GGG	1	0 / 0 / 5 / 30	95.9	SbTAN1 (LOC8076027) exon 1, base 258
SbANR-gA	sorghum ANR, LOC8066842 copies A and B, chr6	GTTCAGAGTTCCCCTGACAG CGG	2	0 / 0 / 0 / 5	99.3	SbANR4 (LOC8066842 copy A) exon 3, base 314; SbANR5 (LOC8066842 copy B) exon 3, base 314
SbANR-gB	sorghum ANR, LOC8066842 copies A and B, chr6	GGAGAGGCAGTTTGGGACGC TGG	2	0 / 0 / 0 / 32	96.7	SbANR4 (LOC8066842 copy A) exon 4, base 632; SbANR5 (LOC8066842 copy B) exon 4, base 632

Table 4. The eight guides of the sorghum array design. Guides G1 to G4 and G7 and G8 cut the seven coding copies; G10 and G11 cut the ANR-like transcript LOC8074755. Sites are counted outside the array window; the MIT score is given outside the window and over all sites. Data: `data/tables/T04.csv`*; script:* `code/02_guides.py`*.*

Guide	Protospacer and PAM	GC (%)	Perfect sites	Sites at 1 / 2 / 3 / 4 mismatches outside the array	MIT score (outside / all sites)	Perfect site in
SbANR-G1	AAGACGACGGTCAGGAACCC CGG	60	6	0 / 0 / 0 / 20	97.0 / 29.0	SbANR1 (LOC110436259); SbANRL1 (LOC8074755); SbANR2 (LOC8066840); SbANR3 (LOC8066841); SbANR4 (LOC8066842 copy A); SbANRL3 (LOC8074756)
SbANR-G2	GTCTGCCCCGTCCTCACCGT CGG	70	4	0 / 0 / 1 / 30	94.9 / 52.1	SbANR1 (LOC110436259); SbANR2 (LOC8066840); SbANR3 (LOC8066841); SbANR5 (LOC8066842 copy B)
SbANR-G3	GCGCGTGGTCCTGACCTCGT CGG	70	4	0 / 0 / 5 / 30	89.7 / 28.2	SbANR4 (LOC8066842 copy A); SbANR5 (LOC8066842 copy B); SbANR6 (LOC8066844); SbANRL3 (LOC8074756)
SbANR-G4	AACTGCCTCTCCCTTCTATC AGG/CGG	50	3	0 / 0 / 1 / 16	96.4 / 47.9	SbANR4 (LOC8066842 copy A); SbANR5 (LOC8066842 copy B); SbANR6 (LOC8066844)
SbANR-G7	TACCCAGGCTGGCGGCCTCT CGG	70	1	0 / 0 / 0 / 4	99.7 / 99.6	SbANR7 (LOC8066846)
SbANR-G8	TACGGAGACGTAATCCAGTA CGG	45	1	0 / 0 / 1 / 1	99.8 / 99.8	SbANR7 (LOC8066846)
SbANR-G10	CTGCGTGTCCAAGGTGCTCC TGG	65	2	0 / 0 / 3 / 19	94.0 / 49.1	SbANRL1 (LOC8074755)
SbANR-G11	CTCTGACGCGAGGTTCAAAC CGG	55	1	0 / 0 / 0 / 3	99.7 / 99.7	SbANRL1 (LOC8074755)

Table 5. Deletions between the two cuts of each guide pair. A clean join is assumed. Coding bases removed, reading frame, first residue affected and the number of residues before the first stop codon of the edited sequence; copy SbANR4 (LOC8066842 copy A) has no stop codon in its stored sequence. Data: `data/tables/T05.csv`*; script:* `code/02_guides.py`*.*

Target	Guide pair	Coding bases removed	Frame	First residue affected	Genomic bases between the cuts	Catalytic codon between the cuts	Residues before the first stop codon (edited)
VvANR (GeneID 100232981)	VvANR-gA + VvANR-gB	80	frameshift	144	189	yes	148
SbTAN1 (LOC8076027)	SbTAN1-gA + SbTAN1-gB	167	frameshift	31	167	n/a	183
SbANR1 (LOC110436259)	SbANR-G1 + SbANR-G2	471	in frame	44	1167	yes	188
SbANR2 (LOC8066840)	SbANR-G1 + SbANR-G2	468	in frame	41	800	yes	184
SbANR3 (LOC8066841)	SbANR-G1 + SbANR-G2	468	in frame	40	930	yes	184
SbANR4 (LOC8066842 copy A)	SbANR-G3 + SbANR-G4	271	frameshift	128	367	yes	197
SbANR5 (LOC8066842 copy B)	SbANR-G3 + SbANR-G4	271	frameshift	128	395	yes	197
SbANR6 (LOC8066844)	SbANR-G3 + SbANR-G4	268	frameshift	130	354	yes	139
SbANR7 (LOC8066846)	SbANR-G7 + SbANR-G8	514	frameshift	173	1006	yes	207
SbANRL1 part 1 (LOC8074755)	SbANR-G1 + SbANR-G10	404	frameshift	78	1065	yes	86
SbANRL1 part 2 (LOC8074755)	SbANR-G11 + SbANR-G10	273	in frame	481	376	yes	220

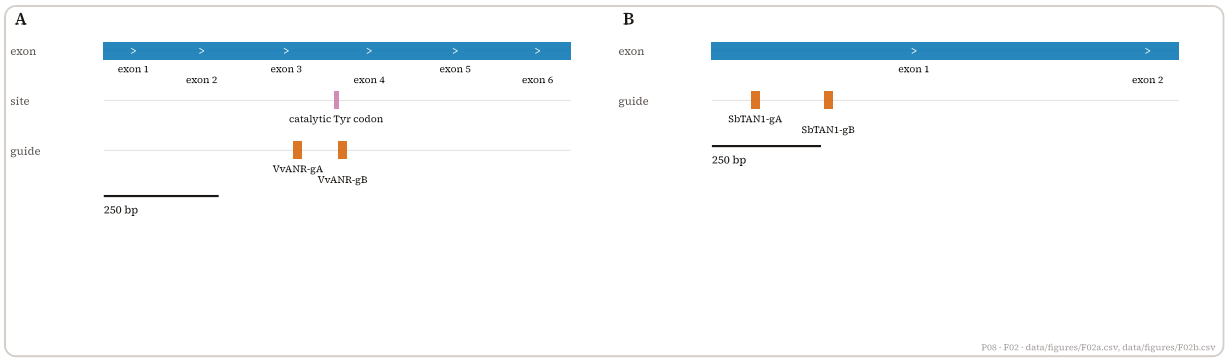


Figure 2. Guide sites in the grape ANR and sorghum TAN1 coding sequences. A, VvANR (GeneID 100232981). B, SbTAN1 (LOC8076027). Exons, the catalytic Tyr codon (VvANR) and the two guides of each gene along the coding sequence; the scale bar is in base pairs. Positions are coding-sequence coordinates computed from the exon coordinates. Data: `data/figures/F02a.csv`, `data/figures/F02b.csv`; script: `code/02_guides.py`.

3.6 The ANR-like transcript LOC8074755

LOC8074755 holds two ANR-like parts. In part 1 the catalytic Tyr codon is a His codon (CAC; motif HCVSK), and in part 2 the motif is YCVSK with the Tyr intact. Cutting both parts needs added guides: of the 124,186 complete six-guide sets, 0 flank part 1 and 0 flank part 2. The search added 258 candidates that cut only in LOC8074755 (155 in part 1 and 133 in part 2); their perfect sites all lie inside the array, and 0 lie outside it. With the six guides kept, part 1 needs one added guide (47 single guides work), part 2 needs two added guides (no single guide works; 3,119 pairs work) and both parts need two (140 pairs; Table S6).

Among the 140 pairs, the added pair has the highest lowest MIT score outside the array. SbANR-G10 (`CTGCGTGTCCAAGGTGCTCC`, PAM TGG) has 2 perfect sites, both in LOC8074755 (one in each part), and a score of 94.0. SbANR-G11 (`CTCTGACGCGAGGTTCAAAC`, PAM CGG) has 1 perfect site, in part 2, and a score of 99.7. SbANR-G1 with SbANR-G10 flanks part 1: the pair removes 404 spliced bases, which shifts the reading frame, and the stretch reaches a stop codon after 86 residues. SbANR-G11 with SbANR-G10 flanks part 2: the pair removes 273 spliced bases in frame. The design is now 8 guides in four groups: the three pairs of the six guides, and SbANR-G10 and SbANR-G11, which act on LOC8074755 together with SbANR-G1. The 22 perfect sites of the eight guides lie within the same 77,225 bases.

3.7 The two-copy option and the independent check

For the sorghum gene named in the literature, Sobic.006G227200, which is SbANR4 (LOC8066842 copy A), the pair SbANR-gA and SbANR-gB is enough. Both guides have 2 perfect sites, one in SbANR4 (LOC8066842 copy A) and one in SbANR5 (LOC8066842 copy

B), with scores of 99.3 and 96.7. The pair removes 318 coding bases in frame in each copy and leaves the other 5 copies intact. Our recomputed guide-by-target matrix equals the stored matrix of the cover search in 90 of 90 cells (Table S4). An independent second script, which shares no code, coordinates or scan output with the first, finds the same perfect-site lists for the guides SbANR-G1 to SbANR-G4 and SbANR-G7 to SbANR-G9 (6, 4, 4, 3, one, one and one) and for SbANR-G10 and SbANR-G11 (2 and one).

3.8 Design targets for proanthocyanidin reduction

Table 6 and Figure 4 give the design targets. For grape berry skin the target is below 50 µg catechin equivalents per gram of fresh weight from an assumed wild-type range of 300 to 600, which is a reduction of 83.33% to 91.67%. For grape berry seed the target is below 0.5 mg per gram of fresh weight from 8 to 12, a reduction of 93.75% to 95.83%. For tannin-type sorghum grain the target is below 0.05% of dry weight from 0.5 to 3%, a reduction of 90.00% to 98.33%. The reduction rises with the wild-type value because the target is fixed (Table S9). These are design targets for the knockouts described above; no IFI score is assigned to any compound.

Table 6. Design targets for proanthocyanidin reduction. Wild-type ranges are design assumptions with no source; targets are design targets. Reduction = 1 - target / wild-type value, at the lowest and highest wild-type value. Data: `data/tables/T06.csv`*; script:* `code/04_reductions.py`*.*

Tissue	Unit	Assumed wild-type range	Design target	Reduction at the lowest wild-type value (%)	Reduction at the highest wild-type value (%)
Grape berry skin	ug catechin equivalents/g fresh weight	300 to 600	below 50	83.33	91.67
Grape berry seed	mg catechin equivalents/g fresh weight	8 to 12	below 0.5	93.75	95.83
Sorghum grain (tannin type)	% of dry weight	0.5 to 3	below 0.05	90.00	98.33

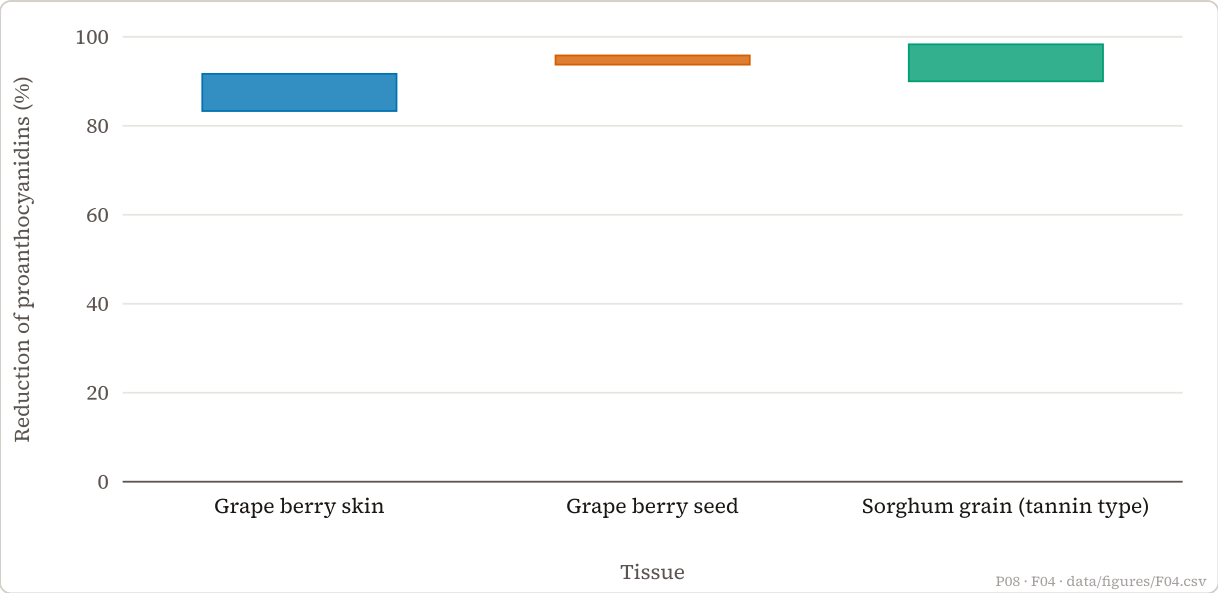


Figure 4. Reduction at the design targets across the assumed wild-type range. Each bar spans the reduction (1 minus target divided by wild-type value) from the highest to the lowest assumed wild-type value of Table 6, at the design target of that tissue. One-at-a-time analysis of the wild-type value with the target held at its design value. Data: `data/figures/F04.csv`; script: `code/04_reductions.py`.

4. Limitations

Specificity is the MIT score over strict NGG sites with up to four mismatches; on-target activity scores, NAG sites and bulged sites are not computed. Which ANR copies are expressed in each tissue is outside this paper, so copy identity rests on annotation and protein identity. The 22 perfect sites of the eight array guides lie within 77,225 bases, so deletions or rearrangements between cuts are possible. Reading-frame results assume a clean join. LOC8074755 is annotated as a transcribed pseudogene, and whether either part is translated is outside this paper. Wild-type ranges are design assumptions with no source.

5. Data and code availability

Everything behind this paper is in the data package: the claims table, the data for every table and figure, the scripts that regenerate them, the inputs, the research record and a backup. `bash run_all.sh` reruns every step and compares each regenerated file with the stored file; `bash verify.sh` checks the checksums. The quick steps start from the stored genome scan in `inputs/guide_design/hits.tsv`; `bash run_all.sh --full` repeats the guide design from the genome files. The paper and data are licensed under

CC BY 4.0 and the code under the MIT licence. The public input datasets are listed in `inputs/input_sources.csv`.

Public records of this paper: Zenodo 10.5281/zenodo.24511334; Figshare: figshare/35015818; Harvard Dataverse: DVN/VPB0PK; OSF: osf.io/fv58q; Hugging Face: capsaicinoid-data; GitHub: paper-08; Website: astringency-reduction.

Intellectual Property. The engineered constructs, cultivar designs, and methodologies described in this paper are the subject of pending United States patent applications (USPTO Provisional Application Nos. 64/172,827; 64/188,197; 64/192,062; and 64/199,105).

Experimental status. All results reported in this paper are computational. Experimental work on these designs is under way at the authors' laboratory facilities in Montecito, California, and Geneva, Switzerland. Those data are not part of this paper, are shared only with commercial partners under agreement, and are planned for publication next year.

Active experiments for this design include a proanthocyanidin baseline (grape, sorghum, pepper) with a pepper ANR check, resequencing of every target locus in the lab lines, the astringency reference assay (salivary protein precipitation), sensory panel screening, training and taster-status genotyping, and transcript and ANR-like activity analysis of gene LOC8074755 together with its added guides.

6. References

1. Devic, M., Guilleminot, J., Debeaujon, I., Bechtold, N., Bensaude, E., Koornneef, M., Pelletier, G., & Delseny, M. (1999). The *BANYULS* gene encodes a DFR-like protein and is a marker of early seed coat development. *The Plant Journal, 19*(4), 387–398. doi:10.1046/j.1365-313x.1999.00529.x

2. Ford, P.I. (2026). The Iyengar Flavor Index (IFI): A Receptor-Normalized Standard for Cross-Modal Flavor Measurement. *Scoville Splice Research Series*. doi:10.5281/zenodo.24511320

3. Hagerman, A. E., & Butler, L. G. (1981). The specificity of proanthocyanidin-protein interactions. *Journal of Biological Chemistry, 256*(9), 4494–4497. doi:10.1016/S0021-9258(19)69462-7

4. Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., ... & Zhang, F. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology, 31*(9), 827–832. doi:10.1038/nbt.2647

5. Jaillon, O., Aury, J.-M., Noel, B., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature, 449*(7161), 463–467. doi:10.1038/nature06148

6. Lewis, J. A., Zhang, B., Harza, R., Palmer, N., Sarath, G., Sattler, S. E., Twigg, P., Vermerris, W., & Kang, C. (2023). Structural similarities and overlapping activities among dihydroflavonol 4-reductase, flavanone 4-reductase, and anthocyanidin reductase offer metabolic flexibility in the flavonoid pathway. *International Journal of Molecular Sciences, 24*(18), 13901. doi:10.3390/ijms241813901

7. Paterson, A. H., Bowers, J. E., Bruggmann, R., et al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature, 457*(7229), 551–556. doi:10.1038/nature07723

8. Slack, J. P., Brockhoff, A., Batram, C., et al. (2010). Modulation of bitter taste perception by a small molecule hTAS2R antagonist. *Current Biology, 20*(12), 1104–1109. doi:10.1016/j.cub.2010.04.043

9. Wu, Y., Li, X., Xiang, W., et al. (2012). Presence of tannins in sorghum grains is conditioned by different natural alleles of *Tannin1*. *Proceedings of the National Academy of Sciences, 109*(26), 10281–10286. doi:10.1073/pnas.1201700109

10. Xie, D.-Y., Sharma, S. B., Paiva, N. L., Ferreira, D., & Dixon, R. A. (2003). Role of anthocyanidin reductase, encoded by *BANYULS* in plant flavonoid biosynthesis. *Science, 299*(5605), 396–399. doi:10.1126/science.1078540

Supplementary data

The supplementary tables hold the full data behind the paper. Table S1 gives the deletion consequences of every designed pair and the alternatives. Table S2 lists the perfect sites of the 15 guides with their cut positions. Table S3 lists every guide pair that cuts either side of a catalytic codon. Table S4 gives the best site of each guide in each coding copy and non-coding feature. Table S5 describes the non-coding ANR-like features. Table S6 gives the fewest guides for the eighth target. Table S7 lists the ANR-type genes outside the array. Table S8 gives the MIT position weights. Table S9 gives the sensitivity of the reductions. Table S10 gives the search counts. The scores of all candidate guides are in `data/results/candidate_scores.csv` and `data/results/candidate_scores_LOC8074755.csv`. Every table is stored as a CSV file in `data/tables/`, and the scripts that write them are in `code/`. The file map below lists every data, input, code and research-record file.

Table S1. Deletion consequences of the designed pairs and alternatives. Every pair listed in Table 5, the pair SbANR-gA with SbANR-gB in the two copies of LOC8066842, and the pair SbANR-G7 with SbANR-G9 in SbANR7 (LOC8066846). Data: `data/tables/S01.csv`*; script:* `code/02_guides.py`*.*

Guide pair	Target	Coding bases removed	Frame	First residue affected	Genomic bases between the cuts	Catalytic codon between the cuts	Residues before the first stop (original)	Residues before the first stop (edited)
VvANR-gA + VvANR-gB	VvANR (GeneID 100232981)	80	frameshift	144	189	yes	338	148
SbANR-gA + SbANR-gB	SbANR4 (LOC8066842 copy A)	318	in frame	105	414	yes	341	235
SbANR-gA + SbANR-gB	SbANR5 (LOC8066842 copy B)	318	in frame	105	442	yes	341	235
SbTAN1-gA + SbTAN1-gB	SbTAN1 (LOC8076027)	167	frameshift	31	167	n/a	356	183
SbANR-G1 + SbANR-G2	SbANR1 (LOC110436259)	471	in frame	44	1167	yes	345	188
SbANR-G1 + SbANR-G2	SbANR2 (LOC8066840)	468	in frame	41	800	yes	340	184
SbANR-G1 + SbANR-G2	SbANR3 (LOC8066841)	468	in frame	40	930	yes	340	184
SbANR-G3 + SbANR-G4	SbANR4 (LOC8066842 copy A)	271	frameshift	128	367	yes	341	197
SbANR-G3 + SbANR-G4	SbANR5 (LOC8066842 copy B)	271	frameshift	128	395	yes	341	197
SbANR-G3 + SbANR-G4	SbANR6 (LOC8066844)	268	frameshift	130	354	yes	343	139
SbANR-G7 + SbANR-G8	SbANR7 (LOC8066846)	514	frameshift	173	1006	yes	354	207
SbANR-G7 + SbANR-G9	SbANR7 (LOC8066846)	27	in frame	173	116	yes	354	345
SbANR-G1 + SbANR-G10	SbANRL1 part 1 (LOC8074755)	404	frameshift	78	1065	yes	474	86
SbANR-G11 + SbANR-G10	SbANRL1 part 2 (LOC8074755)	273	in frame	481	376	yes	311	220

Table S2. Perfect sites of the 15 guides. Every genome site with no mismatch to a guide, with the first base after the cut and its coding index in the target model. Data: `data/tables/S02.csv`*; script:* `code/02_guides.py`.

Guide	Contig	Start of the 23-mer	Strand	Feature	Cut (first base after)	Coding index of the cut	Exon
VvANR-gA	NW_003724316.1	6,629	+	VvANR (GeneID 100232981)	6,646	431	3
VvANR-gB	NW_003724316.1	6,829	-	VvANR (GeneID 100232981)	6,835	511	4
SbANR-gA	NC_012875.2	57,203,587	-	SbANR4 (LOC8066842 copy A)	57,203,593	314	3
SbANR-gA	NC_012875.2	57,206,925	-	SbANR5 (LOC8066842 copy B)	57,206,931	314	3
SbANR-gB	NC_012875.2	57,204,001	-	SbANR4 (LOC8066842 copy A)	57,204,007	632	4
SbANR-gB	NC_012875.2	57,207,367	-	SbANR5 (LOC8066842 copy B)	57,207,373	632	4
SbTAN1-gA	NC_012873.2	62,315,712	-	SbTAN1 (LOC8076027)	62,315,718	91	1
SbTAN1-gB	NC_012873.2	62,315,879	-	SbTAN1 (LOC8076027)	62,315,885	258	1
SbANR-G1	NC_012875.2	57,170,728	+	SbANR1 (LOC110436259)	57,170,745	132	1
SbANR-G1	NC_012875.2	57,176,180	+	SbANRL1 (LOC8074755)	57,176,197	234	1
SbANR-G1	NC_012875.2	57,185,108	+	SbANR2 (LOC8066840)	57,185,125	123	1
SbANR-G1	NC_012875.2	57,198,611	+	SbANR3 (LOC8066841)	57,198,628	120	1
SbANR-G1	NC_012875.2	57,203,093	+	SbANR4 (LOC8066842 copy A)	57,203,110	120	1
SbANR-G1	NC_012875.2	57,237,495	-	SbANRL3 (LOC8074756)	57,237,501		
SbANR-G2	NC_012875.2	57,171,895	+	SbANR1 (LOC110436259)	57,171,912	603	4
SbANR-G2	NC_012875.2	57,185,908	+	SbANR2 (LOC8066840)	57,185,925	591	4
SbANR-G2	NC_012875.2	57,199,541	+	SbANR3 (LOC8066841)	57,199,558	588	4
SbANR-G2	NC_012875.2	57,207,318	+	SbANR5 (LOC8066842 copy B)	57,207,335	594	4
SbANR-G3	NC_012875.2	57,203,645	+	SbANR4 (LOC8066842 copy A)	57,203,662	383	3
SbANR-G3	NC_012875.2	57,206,983	+	SbANR5 (LOC8066842 copy B)	57,207,000	383	3
SbANR-G3	NC_012875.2	57,211,077	+	SbANR6 (LOC8066844)	57,211,094	389	3

Guide	Contig	Start of the 23-mer	Strand	Feature	Cut (first base after)	Coding index of the cut	Exon
SbANR-G3	NC_012875.2	57,236,738	-	SbANRL3 (LOC8074756)	57,236,744		
SbANR-G4	NC_012875.2	57,204,012	+	SbANR4 (LOC8066842 copy A)	57,204,029	654	4
SbANR-G4	NC_012875.2	57,207,378	+	SbANR5 (LOC8066842 copy B)	57,207,395	654	4
SbANR-G4	NC_012875.2	57,211,431	+	SbANR6 (LOC8066844)	57,211,448	657	4
SbANR-G7	NC_012875.2	57,246,957	-	SbANR7 (LOC8066846)	57,246,963	518	3
SbANR-G8	NC_012875.2	57,247,952	+	SbANR7 (LOC8066846)	57,247,969	1032	5
SbANR-G9	NC_012875.2	57,247,062	+	SbANR7 (LOC8066846)	57,247,079	545	4
SbANR-G10	NC_012875.2	57,177,245	+	SbANRL1 (LOC8074755)	57,177,262	638	4
SbANR-G10	NC_012875.2	57,181,540	+	SbANRL1 (LOC8074755)	57,181,557	1714	9
SbANR-G11	NC_012875.2	57,181,175	-	SbANRL1 (LOC8074755)	57,181,181	1441	8

Table S3. Every guide pair that cuts either side of a catalytic codon. Pairs of guides with perfect sites on either side of the catalytic codon of one target, at most 700 coding bases apart. Data: `data/tables/S03.csv`*; script:* `code/02_guides.py`*.*

Target	Upstream guide	Downstream guide	Coding bases removed	Removed mod 3	Genomic bases between the cuts
SbANR1 (LOC110436259)	SbANR-G1	SbANR-G2	471	0	1167
SbANR2 (LOC8066840)	SbANR-G1	SbANR-G2	468	0	800
SbANR3 (LOC8066841)	SbANR-G1	SbANR-G2	468	0	930
SbANR4 (LOC8066842 copy A)	SbANR-G1	SbANR-G4	534	0	919

Target	Upstream guide	Downstream guide	Coding bases removed	Removed mod 3	Genomic bases between the cuts
SbANR4 (LOC8066842 copy A)	SbANR-G1	SbANR-gB	512	2	897
SbANR4 (LOC8066842 copy A)	SbANR-G3	SbANR-G4	271	1	367
SbANR4 (LOC8066842 copy A)	SbANR-G3	SbANR-gB	249	0	345
SbANR4 (LOC8066842 copy A)	SbANR-gA	SbANR-G4	340	1	436
SbANR4 (LOC8066842 copy A)	SbANR-gA	SbANR-gB	318	0	414
SbANR5 (LOC8066842 copy B)	SbANR-G3	SbANR-G2	211	1	335
SbANR5 (LOC8066842 copy B)	SbANR-G3	SbANR-G4	271	1	395
SbANR5 (LOC8066842 copy B)	SbANR-G3	SbANR-gB	249	0	373
SbANR5 (LOC8066842 copy B)	SbANR-gA	SbANR-G2	280	1	404
SbANR5 (LOC8066842 copy B)	SbANR-gA	SbANR-G4	340	1	464

Target	Upstream guide	Downstream guide	Coding bases removed	Removed mod 3	Genomic bases between the cuts
SbANR5 (LOC8066842 copy B)	SbANR-gA	SbANR-gB	318	0	442
SbANR6 (LOC8066844)	SbANR-G3	SbANR-G4	268	1	354
SbANR7 (LOC8066846)	SbANR-G7	SbANR-G8	514	1	1006
SbANR7 (LOC8066846)	SbANR-G7	SbANR-G9	27	0	116
SbANRL1 part 1 (LOC8074755)	SbANR-G1	SbANR-G10	404	2	1065
SbANRL1 part 2 (LOC8074755)	SbANR-G11	SbANR-G10	273	0	376
VvANR (GeneID 100232981)	VvANR-gA	VvANR-gB	80	2	189

Table S4. Best site of each guide in each target. Mismatches of the best site of each guide in each coding copy and non-coding feature (0 = perfect site, - = none within four mismatches). Guides G1 to G4 and G7 to G11 are the array guides; gA and gB are SbANR-gA and SbANR-gB. Data: `data/tables/S04.csv`*; script:* `code/03_array.py`*.*

Target	G1	G2	G3	G4	G7	G8	G9	G10	G11	gA	gB
SbANR1 (LOC110436259)	0	0	1	-	-	-	-	1	-	-	-
SbANR2 (LOC8066840)	0	0	2	-	-	-	-	4	-	-	-
SbANR3 (LOC8066841)	0	0	1	4	-	-	-	4	-	-	-
SbANR4 (LOC8066842 copy A)	0	2	0	0	-	-	-	3	-	0	0
SbANR5 (LOC8066842 copy B)	1	0	0	0	-	-	-	4	-	0	0
SbANR6 (LOC8066844)	1	2	0	0	-	-	-	-	-	-	2
SbANR7 (LOC8066846)	4	2	2	-	0	0	0	-	-	-	-
SbANRL2 (LOC8066836)	1	1	1	2	-	-	-	3	-	-	4
SbANRL1 (LOC8074755)	0	2	2	4	4	-	-	0	0	-	4
SbANRL3 (LOC8074756)	0	1	0	1	-	-	-	-	-	4	2

Table S5. Non-coding ANR-like features of the array. Local alignment of the best reading frame of each feature to SbANR4 (LOC8066842 copy A; BLOSUM62, gap -10 and -0.5). Data: `data/tables/S05.csv`*; script:* `code/03_array.py`*.*

Feature	Type	Genomic span on chr6	Spliced length (nt)	Copy A residues aligned	Identical / aligned residue pairs	Identity (%)	Stop codons in the whole frame
LOC8066836	pseudogene	57158598-57160357	1299	8-177	138/169	82	3
LOC8074755	transcribed pseudogene	57175964-57182390	2359	2-340	239/336	71	12
LOC8074756	pseudogene	57235753-57237804	1056	1-340	265/338	78	2
LOC110436542	lncRNA	57241677-57242420	657	278-341	33/64	52	4

Table S6. Fewest guides for the eighth target. Fewest guides that cut either side of the catalytic codon in every target, with free choice and with the six array guides kept (integer programme, optimality proven). Data: `data/tables/S06.csv`*; script:* `code/03_array.py`*.*

Targets to cut either side of the catalytic codon	Fewest guides (free choice)	Fewest guides with the six kept	Guides added to the six
the seven copies	6	6	0
the seven copies and part 1	7	7	1
the seven copies and part 2	7	8	2
the seven copies and both parts	7	8	2

Table S7. ANR-type annotated genes outside the array. Annotated anthocyanidin reductase genes elsewhere in the sorghum genome and their identity to SbANR4 (LOC8066842 copy A) over alignment columns. Data: `data/tables/S07.csv`*; script:* `code/03_array.py`*.*

Gene	Contig	Genomic span	Strand	RefSeq protein	Length (aa)	Identical to SbANR4 (LOC8066842 copy A), alignment columns	Identity (%)
LOC110431240	NC_012879.2	57,603,657-57,613,061	-	XP_021305772.1	433	147 of 453	32.5
LOC8060651	NC_012871.2	72,605,623-72,615,957	-	XP_002463150.1	341	113 of 357	31.7
LOC8069347	NC_012879.2	55,415,930-55,417,809	-	XP_021304687.1	355	123 of 373	33.0

Gene	Contig	Genomic span	Strand	RefSeq protein	Length (aa)	Identical to SbANR4 (LOC8066842 copy A), alignment columns	Identity (%)
LOC8079005	NC_012871.2	72,605,514-72,608,325	-	XP_002463149.2	387	118 of 404	29.2
LOC8083550	NC_012876.2	59,193,165-59,195,564	+	XP_002444518.1	355	117 of 367	31.9

Table S8. MIT position weights. Weight of a mismatch at each position of the protospacer, counted from the PAM-distal end (Hsu et al., 2013). Data: `data/tables/S08.csv`*; script:* `code/02_guides.py`*.*

Position from the PAM-distal end	Weight of a mismatch
1	0
2	0
3	0.014
4	0
5	0
6	0.395
7	0.317
8	0
9	0.389
10	0.079
11	0.445
12	0.508
13	0.613
14	0.851
15	0.732
16	0.828
17	0.615

Position from the PAM-distal end	Weight of a mismatch
18	0.804
19	0.685
20	0.583

Table S9. Sensitivity of the reductions to the wild-type value. Reduction at the mid-range, lowest and highest assumed wild-type value, with the target held fixed. Data: `data/tables/S09.csv`*; script:* `code/04_reductions.py`*.*

Tissue	Reduction at the mid-range wild-type value (%)	Reduction at the lowest wild-type value (%)	Reduction at the highest wild-type value (%)	Mid-range wild-type value
Grape berry skin	88.89	83.33	91.67	450
Grape berry seed	95.00	93.75	95.83	10
Sorghum grain (tannin type)	97.14	90.00	98.33	1.75

Table S10. Search counts. Counts from the candidate list, the exhaustive cover search, the array census and the eighth-target search. Data: `data/tables/S10.csv`*; script:* `code/03_array.py`*.*

Metric	Value	Meaning
candidates with cut in coding sequence	764	candidate protospacers (GC 40 to 70 percent, no TTTT, no run of five) whose cut lies in the coding sequence of a copy
candidates with perfect sites in 1 copies	680	candidates with a perfect coding-sequence cut site in exactly 1 of the seven copies
candidates with perfect sites in 2 copies	73	candidates with a perfect coding-sequence cut site in exactly 2 of the seven copies
candidates with perfect sites in 3 copies	7	candidates with a perfect coding-sequence cut site in exactly 3 of the seven copies

Metric	Value	Meaning
candidates with perfect sites in 4 copies	4	candidates with a perfect coding-sequence cut site in exactly 4 of the seven copies
perfect site cover minimum	3	fewest guides that give each copy one perfect site
perfect site cover sets	4841	sets of that size
search nodes k3	1	nodes searched for sets of 3 guides
complete cores k3	0	optimal cores completing to at most 3 guides
search nodes k4	1511	nodes searched for sets of 4 guides
complete cores k4	0	optimal cores completing to at most 4 guides
search nodes k5	51558	nodes searched for sets of 5 guides
complete cores k5	0	optimal cores completing to at most 5 guides
search nodes k6	6751183	nodes searched for sets of 6 guides
complete cores k6	1540	optimal cores completing to at most 6 guides
flanking cover minimum	6	fewest guides that give every copy two flanking perfect sites
optimal cores	1540	optimal cores (sets of multi-copy guides)
complete six guide sets	124186	complete six-guide sets (cap of 700 coding bases)
best worst guide mit	89.7	highest worst-guide MIT score outside the window over all complete six-guide sets
sets reaching best mit	5136	six-guide sets that reach that worst-guide MIT score
optimal cores no cap	2772	optimal cores without the cap of 700 coding bases
complete six guide sets no cap	174348	complete six-guide sets without the cap
genes in window	26	annotated genes in NC_012875.2:57,100,000-57,300,000
window genes lncRNA	1	genes of type lncRNA in the window

Metric	Value	Meaning
window genes protein coding	22	genes of type protein coding in the window
window genes pseudogene	2	genes of type pseudogene in the window
window genes transcribed pseudogene	1	genes of type transcribed pseudogene in the window
eighth target extra candidates	258	added candidates that cut only in LOC8074755
candidates cutting part1	155	candidates with a perfect-site cut in part 1 of LOC8074755
candidates cutting part2	133	candidates with a perfect-site cut in part 2 of LOC8074755
added candidates perfect outside array	0	added candidates with a perfect site outside the array
single additions part1	47	single added guides that flank part 1 when the six are kept
single additions part2	0	single added guides that flank part 2 when the six are kept
pair additions part2	3119	pairs of added guides that flank part 2 when the six are kept
single additions both	0	single added guides that flank both parts when the six are kept
pair additions both	140	pairs of added guides that flank both parts when the six are kept
six guide covers rebuilt	124186	complete six-guide covers rebuilt in the second method
six guide covers flanking part1	0	of those covers, sets that flank part 1
six guide covers flanking part2	0	of those covers, sets that flank part 2
loc8074755 stretch identical to C5YGL7	344	residues of the LOC8074755 part-1 translation identical to UniProt C5YGL7 (SbFNR1)

Metric	Value	Meaning
loc110436259 identical to A0A1B6PND3	345	residues of the SbANR1 (LOC110436259) protein identical to UniProt A0A1B6PND3 (SbFNR2)
grape cds vs refseq identical	334	alignment columns identical between the translated grape coding sequence and the RefSeq protein
grape cds vs refseq columns	338	alignment columns compared for the grape protein
tan1 refseq vs uniprot identical	310	residues of UniProt C5Y0X7 identical to the RefSeq protein XP_002452749.2
six array guides perfect sites	19	perfect sites of the six array guides
six array guides site span	77225	bases from the first to the last perfect site of the six array guides (57170728 to 57247952)
eight array guides perfect sites	22	perfect sites of the eight array guides
eight array guides site span	77225	bases from the first to the last perfect site of the eight array guides (57170728 to 57247952)
coverage cells compared	90	guide-by-target cells compared with the stored matrix of the cover search
coverage cells equal	90	cells equal to the stored matrix

Supplementary data: file map

Path	Contents
`data/tables/T01.csv`	Table 1: Protein identity of the anthocyanidin reductases
`data/tables/T02.csv`	Table 2: The seven coding ANR copies of the sorghum array
`data/tables/T03.csv`	Table 3: Guides for grape ANR, sorghum TAN1 and the two-copy option
`data/tables/T04.csv`	Table 4: The eight guides of the sorghum array design
`data/tables/T05.csv`	Table 5: Deletions between the two cuts of each guide pair
`data/tables/T06.csv`	Table 6: Design targets for proanthocyanidin reduction
`data/tables/S01.csv`	Table S1: Deletion consequences of the designed pairs and alternatives
`data/tables/S02.csv`	Table S2: Perfect sites of the 15 guides

Path	Contents
data/tables/S03.csv	Table S3: Every guide pair that cuts either side of a catalytic codon
data/tables/S04.csv	Table S4: Best site of each guide in each target
data/tables/S05.csv	Table S5: Non-coding ANR-like features of the array
data/tables/S06.csv	Table S6: Fewest guides for the eighth target
data/tables/S07.csv	Table S7: ANR-type annotated genes outside the array
data/tables/S08.csv	Table S8: MIT position weights
data/tables/S09.csv	Table S9: Sensitivity of the reductions to the wild-type value
data/tables/S10.csv	Table S10: Search counts
data/figures/GA.csv	Data behind the graphical abstract
data/figures/F01.csv	Data behind Figure 1
data/figures/F02a.csv	Data behind Figure 2
data/figures/F02b.csv	Data behind Figure 2
data/figures/F03.csv	Data behind Figure 3
data/figures/F04.csv	Data behind Figure 4
data/results/guide_table.csv	The 15 guides: sequence, PAM, GC content, perfect sites, sites at 1 to 4 mismatches (whole genome and outside the array window) and MIT scores
data/results/guide_sites.csv	Every perfect site of every guide with the first base after the cut, the target model and the coding index and exon of the cut
data/results/flanking_pairs.csv	Every pair of guides whose cuts lie either side of a catalytic codon, with the coding bases removed
data/results/pair_deletions.csv	Consequence of the designed deletions: bases removed, frame, first residue affected, residues before the first stop codon
data/results/mit_weights.csv	MIT position weights (Hsu et al., 2013)
data/results/identity.csv	Protein identity of VvANR, Arabidopsis BAN and SbANR4: identical residues, alignment columns and percentages
data/results/catalytic_motifs.csv	YxxxK motifs of VvANR, Arabidopsis BAN and SbANR4 with Tyr and Lys residue numbers
data/results/array_copies.csv	The seven coding copies of the sorghum array: protein, length, catalytic Tyr, identity to SbANR4, exons and genomic span
data/results/coverage_matrix.csv	Best site of each guide in each coding copy and non-coding feature, with the mismatch positions
data/results/coverage_check.csv	Comparison of the recomputed coverage matrix with the stored matrix of the cover search
data/results/design_summary.csv	Counts from the candidate list, the cover search, the array census and the eighth-target search, each with its source file
data/results/eighth_target_minimum.csv	Fewest guides for the seven copies and the two parts of LOC8074755, with free choice and with the six array guides kept
data/results/other_anr_genes.csv	Annotated ANR-type genes outside the array with identity to SbANR4

Path	Contents
`data/results/candidate_scores.csv`	Perfect sites and MIT scores of the 764 candidate guides of the seven copies
`data/results/candidate_scores_LOC8074755.csv`	Perfect sites and MIT scores of all 1,022 candidate guides including the 258 added for LOC8074755
`data/results/pa_reductions.csv`	Reduction of proanthocyanidins at the design targets for the lowest and highest assumed wild-type value
`data/results/pa_reduction_sensitivity.csv`	Reduction at the mid-range, lowest and highest assumed wild-type value
`data/results/ifi_references.csv`	Reference values of the IFI astringency and bitterness axes
`data/figures/*`	Data behind the figures (one CSV per figure or panel; GA.csv holds the five boxes of the graphical abstract)
`data/wet_lab_active.csv`	Experiments active in the authors' laboratories for this paper: register identifier (WL-01 to WL-84), experiment name and stage code (S0 paperwork, S1 calibration and cheap decisive tests, S2 transient and in vitro mechanism tests, S3 stable lines, S4 tests on engineered material, S5 optional and long-lead)
`inputs/gene_names.csv`	Names used in this paper with RefSeq loci, proteins and chromosomes
`inputs/ifi_reference_values.csv`	Reference values of the astringency and bitterness axes of the Iyengar Flavor Index
`inputs/design_targets.csv`	Assumed wild-type ranges and design targets for proanthocyanidins
`inputs/input_sources.csv`	Public genome assemblies and UniProt entries
`inputs/proteins/*`	UniProt protein sequences (grape ANR, Arabidopsis BAN, sorghum ANR, TAN1, SbFNR1, SbFNR2)
`inputs/guide_design/guides.csv`	The 15 guides with sequence and PAM
`inputs/guide_design/hits.tsv`	Stored genome scan: every site with up to four mismatches to each of the 15 guides
`inputs/guide_design/regions.csv`	Coordinates of the coding copies, non-coding features and the array window
`inputs/guide_design/gene_models.json`	Exon coordinates of every target
`inputs/guide_design/catalytic_tyr.csv`	Catalytic Tyr residue of each ANR target
`inputs/guide_design/cds/*`	Spliced coding sequences of every target and the 3' sequence of SbANR7 (LOC8066846) to the end of the mRNA
`code/01_identity.py`	Protein identities and catalytic motifs
`code/02_guides.py`	Guide table, cut positions, flanking pairs, deletion consequences and gene maps
`code/03_array.py`	Array copies, coverage matrix, search counts and figure data
`code/04_reductions.py`	Reduction arithmetic and sensitivity
`code/05_full_rerun.py`	Full-mode rerun of the guide-design scripts from the genome files
`code/heavy/NEEDS.txt`	What full mode needs: genome files, libraries and run time

Path	Contents
`research_record/scripts/p08_extract_inputs.py`	Assembles the stored inputs from the genome files and the stored scan outputs
`research_record/scripts/guide_design/*`	Grape, TAN1 and sorghum guide-design scripts and the scanner as run (run order in README_run_order.txt)
`research_record/scripts/model/*`	Protein identity scripts as run
`research_record/scripts/eighth_target/*`	Eighth-target scripts as run (LOC8074755 model, candidates, cover search, scoring, second method and final check)
`research_record/outputs/guide_design/*`	Raw outputs of the guide-design scripts: gene models, candidate lists, scan hits, scores, cover-search results
`research_record/outputs/model/*`	Raw outputs of the protein identity scripts and their protein sequences
`research_record/outputs/eighth_target/*`	Raw outputs of the eighth-target scripts: model, candidates, scan hits, scores, cover results
`research_record/second_script_rerun/*`	Independent second scripts (genome regular-expression search, brute-force site counts, deletion and translation) with their outputs

Cite this paper

Ford, P.I. (2026). Astringency Reduction by Anthocyanidin Reductase Knockout: CRISPR Guide Design for Sorghum and Grape. Scoville Splice Research Series. https://doi.org/10.5281/zenodo.24511334

```
@misc{Ford2026_P08,
  author = {Ford, P.I.},
  title = {Astringency Reduction by Anthocyanidin Reductase Knockout: CRISPR Guide Design for Sorghum and Grape},
  year = {2026},
  publisher = {Scoville Splice LLC},
  series = {Scoville Splice Research Series},
  doi = {10.5281/zenodo.24511334},
  url = {https://doi.org/10.5281/zenodo.24511334}
}
```

Also in `CITATION.cff`, `citation.bib` and `citation.ris` in the data package.

Related papers in this series

ID	Title	DOI
P04	The Iyengar Flavor Index (IFI): A Receptor-Normalized Standard for Cross-Modal Flavor Measurement	10.5281/zenodo.24511320
P06	Zero-Bitter Brassica: In Silico Design of Myrosinase Knockout in Brassica oleracea	10.5281/zenodo.24511326

Appendix A. Evidence table

Claim ID	Claim	Value	Kind	Source file	Location
P08-C001	RefSeq GeneID of the grape ANR gene	100232981	definition	`inputs/gene_names.csv`	row VvANR, column locus
P08-C002	RefSeq scaffold that carries the grape ANR gene	NW_003724316.1	definition	`inputs/gene_names.csv`	row VvANR, column contig
P08-C003	VvANR, VvANR-gA with VvANR-gB, coding bases removed	80	computed	`data/results/pair_deletions.csv`	target=VvANR; pair VvANR-gA + VvANR-gB; column coding_bases_removed
P08-C004	Coding ANR copies in the sorghum array	7	computed	`data/results/array_copies.csv`	seven rows, one per coding copy
P08-C005	Chromosome of the sorghum ANR array	6	definition	`inputs/gene_names.csv`	rows SbANR1 to SbANR7, column chromosome
P08-C006	candidate protospacers (GC 40 to 70 percent, no TTTT, no run of five) whose cut lies in the coding sequence of a copy	764	computed	`data/results/design_summary.csv`	metric=candidates_with_cut_in_coding_sequence

Claim ID	Claim	Value	Kind	Source file	Location
P08-C007	Fewest guides that cut either side of the catalytic Tyr codon in every copy	6	computed	`data/results/design_summary.csv`	metric=flanking_cover_minimum
P08-C008	Guides of the sorghum array design	8	computed	`data/results/guide_table.csv`	rows SbANR-G1, G2, G3, G4, G7, G8, G10 and G11 (eight guides)
P08-C009	Lowest MIT score outside the array among the eight array guides	89.7	computed	`data/results/guide_table.csv`	guide=SbANR-G3; column mit_outside_window
P08-C010	Highest MIT score outside the array among the eight array guides	99.8	computed	`data/results/guide_table.csv`	guide=SbANR-G8; column mit_outside_window
P08-C011	Lowest reduction at the design targets	83.33	design target	`data/results/pa_reductions.csv`	tissue=Grape berry skin; column reduction_at_lowest_wild_type_percent
P08-C012	Highest reduction at the design targets	98.33	design target	`data/results/pa_reductions.csv`	tissue=Sorghum grain; column reduction_at_highest_wild_type_percent
P08-C013	Length of the tan1-a deletion	1-bp	literature value	`literature`	Wu et al. 2012, doi:10.1073/pnas.1201700109
P08-C014	Length of the tan1-b insertion	10-bp	literature value	`literature`	Wu et al. 2012, doi:10.1073/pnas.1201700109
P08-C015	IFI-A reference EC50 of tannic acid (micromolar)	25.0	definition	`inputs/ifi_reference_values.csv`	row IFI-A, column reference_ec50_um
P08-C016	IFI-B reference IC50 of GIV3727 at TAS2R31 (micromolar), the mean of two assays	7.15	definition	`inputs/ifi_reference_values.csv`	row IFI-B, mean of the two assays
P08-C017	IC50 of GIV3727 against acesulfame K at TAS2R31 (micromolar)	6.4	literature value	`literature`	Slack et al. 2010, doi:10.1016/j.cub.2010.04.043
P08-C018	IC50 of GIV3727 against saccharin at TAS2R31 (micromolar)	7.9	literature value	`literature`	Slack et al. 2010, doi:10.1016/j.cub.2010.04.043

Claim ID	Claim	Value	Kind	Source file	Location
P08-C019	Grape assembly used	GCF_000003745.3	definition	`inputs/input_sources.csv`	row Vitis vinifera reference genome 12X
P08-C020	Sorghum assembly used	GCF_000003195.3	definition	`inputs/input_sources.csv`	row Sorghum bicolor reference genome BTx623 NCBIv3
P08-C021	Minimum Python version of the quick steps	3.9	definition	`code/LIBRARIES.txt`	line Quick mode
P08-C022	Python version of the research scripts	3.12.16	definition	`code/LIBRARIES.txt`	line Research scripts as run
P08-C023	Biopython version	1.88	definition	`code/LIBRARIES.txt`	line Full mode
P08-C024	NumPy version	2.5.5	definition	`code/LIBRARIES.txt`	line Full mode
P08-C025	Gap-open penalty of the alignments	-10	definition	`research_record/outputs/model/model_inputs.csv`	row alignment_gap_open
P08-C026	Gap-extend penalty of the alignments	-0.5	definition	`research_record/outputs/model/model_inputs.csv`	row alignment_gap_extend
P08-C027	Catalytic Tyr residue of SbANR4 (UniProt A0A1B6PNE4)	168	computed	`data/results/catalytic_motifs.csv`	row Sorghum ANR (SbANR4), column tyr_residue
P08-C028	Start of the array window on NC_012875.2	57,100,000	definition	`inputs/guide_design/regions.csv`	row array_window, column start
P08-C029	End of the array window on NC_012875.2	57,300,000	definition	`inputs/guide_design/regions.csv`	row array_window, column end

Claim ID	Claim	Value	Kind	Source file	Location
P08-C030	Identity threshold for an ANR-type gene (percent of alignment columns)	40	definition	`research_record/scripts/guide_design/s1b_enumerate.py`	line 46: ANR-type when the protein is at least 0.40 identical
P08-C031	Length of the protospacer (bases)	20	definition	`inputs/guide_design/guides.csv`	column protospacer_20nt
P08-C032	Lowest GC content allowed (percent)	40	definition	`research_record/outputs/guide_design/s2_output.txt`	candidate rule of s2_candidates.py
P08-C033	Highest GC content allowed (percent)	70	definition	`research_record/outputs/guide_design/s2_output.txt`	candidate rule of s2_candidates.py
P08-C034	Length of the identical-base run that is excluded	5	definition	`research_record/outputs/guide_design/s2_output.txt`	candidate rule of s2_candidates.py
P08-C035	Distance of the cut from the PAM (bases)	3	definition	`code/02_guides.py`	Cut convention in the header
P08-C036	Mismatches allowed in the scan	4	definition	`research_record/scripts/guide_design/scan.c`	comment of the scanner

Claim ID	Claim	Value	Kind	Source file	Location
P08-C037	Largest number of coding bases between the two cuts of a flanking pair	700	definition	`research_record/outputs/guide_design/s4d_output.txt`	line P08_MAXDEL
P08-C038	Exons of LOC8074755	11	definition	`research_record/outputs/eighth_target/e1_output.txt`	line exons 11
P08-C039	Spliced length of LOC8074755 (nt)	2,359	definition	`research_record/outputs/eighth_target/e1_output.txt`	line spliced length 2359
P08-C040	First spliced base of part 1	19	computed	`research_record/outputs/eighth_target/e1b_output.txt`	frame0 stretch
P08-C041	Last spliced base of part 1	1440	computed	`research_record/outputs/eighth_target/e1b_output.txt`	frame0 stretch
P08-C042	First spliced base of part 2	1302	computed	`research_record/outputs/eighth_target/e1b_output.txt`	frame2 stretch

Claim ID	Claim	Value	Kind	Source file	Location
P08-C043	Last spliced base of part 2	2234	computed	`research_record/outputs/eighth_target/e1b_output.txt`	frame2 stretch
P08-C044	Coding exons of VvANR	6	computed	`research_record/outputs/guide_design/p08_guides_output.txt`	line Vitis NP_001267885.1: 6 exons
P08-C045	Coding sequence of VvANR (bases)	1,017	computed	`research_record/outputs/guide_design/p08_guides_output.txt`	line Vitis NP_001267885.1: CDS 1017 nt
P08-C046	Protein length of VvANR (aa)	338	computed	`research_record/outputs/guide_design/p08_guides_output.txt`	line Vitis NP_001267885.1: 338 aa
P08-C047	Length of scaffold NW_003724316.1 (bases)	75,112	computed	`research_record/outputs/guide_design/p08_array_content_output.txt`	line NW_003724316.1

Claim ID	Claim	Value	Kind	Source file	Location
P08-C048	Alignment columns identical between the translated VvANR coding sequence and the RefSeq protein	334	computed	`research_record/outputs/guide_design/p08_checks_output.txt`	line identical over alignment 334 of 338 columns
P08-C049	Alignment columns compared for VvANR	338	computed	`research_record/outputs/guide_design/p08_checks_output.txt`	line identical over alignment 334 of 338 columns
P08-C050	Identity of VvANR to Arabidopsis BAN over alignment columns (percent)	63.8	computed	`data/results/identity.csv`	comparison=VvANR (Q5FB34) and Arabidopsis BAN (Q9SEV0); column identity_over_columns_percent
P08-C051	Identity of VvANR to Arabidopsis BAN over the shorter sequence (percent)	64.2	computed	`data/results/identity.csv`	comparison=VvANR (Q5FB34) and Arabidopsis BAN (Q9SEV0); column identity_over_shorter_percent
P08-C052	Identity of SbANR4 to BAN over alignment columns (percent)	44.4	computed	`data/results/identity.csv`	comparison=SbANR4 (LOC8066842 copy A; A0A1B6PNE4) and Arabidopsis BAN (Q9SEV0); column identity_over_columns_percent
P08-C053	Identity of SbANR4 to BAN over the shorter sequence (percent)	45.3	computed	`data/results/identity.csv`	comparison=SbANR4 (LOC8066842 copy A; A0A1B6PNE4) and Arabidopsis BAN (Q9SEV0); column identity_over_shorter_percent
P08-C054	Identity of SbANR4 to VvANR over alignment columns (percent)	48.1	computed	`data/results/identity.csv`	comparison=SbANR4 (LOC8066842 copy A; A0A1B6PNE4) and VvANR (Q5FB34); column identity_over_columns_percent
P08-C055	Identity of SbANR4 to VvANR over the shorter sequence (percent)	48.8	computed	`data/results/identity.csv`	comparison=SbANR4 (LOC8066842 copy A; A0A1B6PNE4) and VvANR (Q5FB34); column identity_over_shorter_percent
P08-C056	Catalytic Tyr residue of VvANR	168	computed	`data/results/catalytic_motifs.csv`	uniprot=Q5FB34; motif YPASK; column tyr_residue

Claim ID	Claim	Value	Kind	Source file	Location
P08-C057	Lys residue of the catalytic motif of VvANR	172	computed	data/results/catalytic_motifs.csv	uniprot=Q5FB34; motif YPASK; column lys_residue
P08-C058	Catalytic Tyr residue of Arabidopsis BAN	169	computed	data/results/catalytic_motifs.csv	uniprot=Q9SEV0; motif YPISK; column tyr_residue
P08-C059	Lys residue of the catalytic motif of Arabidopsis BAN	173	computed	data/results/catalytic_motifs.csv	uniprot=Q9SEV0; motif YPISK; column lys_residue
P08-C060	Lys residue of the catalytic motif of SbANR4	172	computed	data/results/catalytic_motifs.csv	uniprot=A0A1B6PNE4; motif YPVSK; column lys_residue
P08-C061	annotated genes in NC_012875.2:57,100,000-57,300,000	26	computed	data/results/design_summary.csv	metric=genes_in_window
P08-C062	Annotated ANR-type genes in the array window	6	computed	research_record/outputs/guide_design/s1b_output.txt	line n genes in window: 26 ; ANR-type: six gene names
P08-C063	SbANR4 identical residues to UniProt A0A1B6PNE4	341	computed	data/results/array_copies.csv	copy=LOC8066842_A; column identical_to_copy_A
P08-C064	SbANR4 alignment columns against UniProt A0A1B6PNE4	341	computed	data/results/array_copies.csv	copy=LOC8066842_A; column alignment_columns
P08-C065	Lowest identity of a copy to SbANR4 (percent)	66.9	computed	data/results/array_copies.csv	minimum of column identity_percent over the copies other than LOC8066842_A

Claim ID	Claim	Value	Kind	Source file	Location
P08-C066	Highest identity of a copy to SbANR4 (percent)	90.9	computed	data/results/array_copies.csv	maximum of column identity_percent over the copies other than LOC8066842_A
P08-C067	Residues of SbANR1 identical to UniProt A0A1B6PND3	345	computed	data/results/design_summary.csv	metric=loc110436259_identical_to_A0A1B6PND3
P08-C068	Non-coding ANR-like features of the array	3	computed	data/tables/S05.csv	rows LOC8066836, LOC8074755, LOC8074756
P08-C069	Local identity of LOC8066836 to SbANR4 (percent)	82	computed	data/tables/S05.csv	row LOC8066836, column Identity (%)
P08-C070	Local identity of LOC8074755 to SbANR4 (percent)	71	computed	data/tables/S05.csv	row LOC8074755, column Identity (%)
P08-C071	Local identity of LOC8074756 to SbANR4 (percent)	78	computed	data/tables/S05.csv	row LOC8074756, column Identity (%)
P08-C072	Lowest identity of an ANR gene outside the array to SbANR4 (percent)	29.2	computed	data/results/other_anr_genes.csv	minimum of column identity_percent
P08-C073	Highest identity of an ANR gene outside the array to SbANR4 (percent)	33.0	computed	data/results/other_anr_genes.csv	maximum of column identity_percent
P08-C074	Exon cut by VvANR-gA	3	computed	data/results/guide_sites.csv	guide=VvANR-gA; site 1; column exon
P08-C075	Coding index of the VvANR-gA cut	431	computed	data/results/guide_sites.csv	guide=VvANR-gA; site 1; column coding_index_of_cut
P08-C076	Exon cut by VvANR-gB	4	computed	data/results/guide_sites.csv	guide=VvANR-gB; site 1; column exon

Claim ID	Claim	Value	Kind	Source file	Location
P08-C077	Coding index of the VvANR-gB cut	511	computed	data/results/guide_sites.csv	guide=VvANR-gB; site 1; column coding_index_of_cut
P08-C078	Perfect sites of VvANR-gA	1	computed	data/results/guide_table.csv	guide=VvANR-gA; column perfect_sites
P08-C079	Sites of VvANR-gA at 4 mismatches	6	computed	data/results/guide_table.csv	guide=VvANR-gA; column sites_4mm_all
P08-C080	Sites of VvANR-gB at 4 mismatches	5	computed	data/results/guide_table.csv	guide=VvANR-gB; column sites_4mm_all
P08-C081	MIT score of VvANR-gA	98.7	computed	data/results/guide_table.csv	guide=VvANR-gA; column mit_all_sites
P08-C082	MIT score of VvANR-gB	99.4	computed	data/results/guide_table.csv	guide=VvANR-gB; column mit_all_sites
P08-C083	VvANR, VvANR-gA with VvANR-gB, genomic bases between cuts	189	computed	data/results/pair_deletions.csv	target=VvANR; pair VvANR-gA + VvANR-gB; column genomic_bases_between_cuts
P08-C084	VvANR, VvANR-gA with VvANR-gB, first residue affected	144	computed	data/results/pair_deletions.csv	target=VvANR; pair VvANR-gA + VvANR-gB; column first_residue_affected
P08-C085	VvANR, VvANR-gA with VvANR-gB, residues before first stop edited	148	computed	data/results/pair_deletions.csv	target=VvANR; pair VvANR-gA + VvANR-gB; column residues_before_first_stop_edited
P08-C086	VvANR, VvANR-gA with VvANR-gB, residues before first stop original	338	computed	data/results/pair_deletions.csv	target=VvANR; pair VvANR-gA + VvANR-gB; column residues_before_first_stop_original

Claim ID	Claim	Value	Kind	Source file	Location
P08-C087	Chromosome of SbTAN1	4	definition	`inputs/gene_names.csv`	row SbTAN1, column chromosome
P08-C088	Exon cut by SbTAN1-gA	1	computed	`data/results/guide_sites.csv`	guide=SbTAN1-gA; site 1; column exon
P08-C089	Coding index of the SbTAN1-gA cut	91	computed	`data/results/guide_sites.csv`	guide=SbTAN1-gA; site 1; column coding_index_of_cut
P08-C090	Exon cut by SbTAN1-gB	1	computed	`data/results/guide_sites.csv`	guide=SbTAN1-gB; site 1; column exon
P08-C091	Coding index of the SbTAN1-gB cut	258	computed	`data/results/guide_sites.csv`	guide=SbTAN1-gB; site 1; column coding_index_of_cut
P08-C092	MIT score of SbTAN1-gA	96.6	computed	`data/results/guide_table.csv`	guide=SbTAN1-gA; column mit_all_sites
P08-C093	MIT score of SbTAN1-gB	95.9	computed	`data/results/guide_table.csv`	guide=SbTAN1-gB; column mit_all_sites
P08-C094	SbTAN1, SbTAN1-gA with SbTAN1-gB, coding bases removed	167	computed	`data/results/pair_deletions.csv`	target=SbTAN1; pair SbTAN1-gA + SbTAN1-gB; column coding_bases_removed
P08-C095	SbTAN1, SbTAN1-gA with SbTAN1-gB, first residue affected	31	computed	`data/results/pair_deletions.csv`	target=SbTAN1; pair SbTAN1-gA + SbTAN1-gB; column first_residue_affected
P08-C096	Residues of UniProt C5Y0X7 identical to the RefSeq protein	310	computed	`data/results/design_summary.csv`	metric=tan1_refseq_vs_uniprot_identical

Claim ID	Claim	Value	Kind	Source file	Location
P08-C097	candidates with a perfect coding-sequence cut site in exactly 1 of the seven copies	680	computed	`data/results/design_summary.csv`	metric=candidates_with_perfect_sites_in_1_copies
P08-C098	candidates with a perfect coding-sequence cut site in exactly 2 of the seven copies	73	computed	`data/results/design_summary.csv`	metric=candidates_with_perfect_sites_in_2_copies
P08-C099	candidates with a perfect coding-sequence cut site in exactly 3 of the seven copies	7	computed	`data/results/design_summary.csv`	metric=candidates_with_perfect_sites_in_3_copies
P08-C100	candidates with a perfect coding-sequence cut site in exactly 4 of the seven copies	4	computed	`data/results/design_summary.csv`	metric=candidates_with_perfect_sites_in_4_copies
P08-C101	Most copies reached by one guide	4	computed	`data/results/design_summary.csv`	metric candidates_with_perfect_sites_in_4_copies (largest group)
P08-C102	Copies cut by the best 2 or 3 guides	4	computed	`data/figures/F03.csv`	rows guides=2 and guides=3, column copies_cut
P08-C103	Copies cut by the best 4 or 5 guides	6	computed	`data/figures/F03.csv`	rows guides=4 and guides=5, column copies_cut
P08-C104	Copies cut by the best 6 guides	7	computed	`data/figures/F03.csv`	row guides=6, column copies_cut
P08-C105	Nodes searched for sets of 3 guides	1	computed	`data/results/design_summary.csv`	metric=search_nodes_k3
P08-C106	Nodes searched for sets of 4 guides	1,511	computed	`data/results/design_summary.csv`	metric=search_nodes_k4

Claim ID	Claim	Value	Kind	Source file	Location
P08-C107	Nodes searched for sets of 5 guides	51,558	computed	data/results/design_summary.csv	metric=search_nodes_k5
P08-C108	Optimal cores at size 6	1,540	computed	data/results/design_summary.csv	metric=optimal_cores
P08-C109	Nodes searched for sets of 6 guides	6,751,183	computed	data/results/design_summary.csv	metric=search_nodes_k6
P08-C110	Complete six-guide sets (cap of 700 coding bases)	124,186	computed	data/results/design_summary.csv	metric=complete_six_guide_sets
P08-C111	Complete six-guide sets without the cap	174,348	computed	data/results/design_summary.csv	metric=complete_six_guide_sets_no_cap
P08-C112	Fewest guides that give each copy one perfect site	3	computed	data/results/design_summary.csv	metric=perfect_site_cover_minimum
P08-C113	Sets of that size	4,841	computed	data/results/design_summary.csv	metric=perfect_site_cover_sets
P08-C114	LOC110436259, SbANR-G1 with SbANR-G2, coding bases removed	471	computed	data/results/pair_deletions.csv	target=LOC110436259; pair SbANR-G1 + SbANR-G2; column coding_bases_removed
P08-C115	LOC8066840, SbANR-G1 with SbANR-G2, coding bases removed	468	computed	data/results/pair_deletions.csv	target=LOC8066840; pair SbANR-G1 + SbANR-G2; column coding_bases_removed
P08-C116	LOC8066841, SbANR-G1 with SbANR-G2, coding bases removed	468	computed	data/results/pair_deletions.csv	target=LOC8066841; pair SbANR-G1 + SbANR-G2; column coding_bases_removed

Claim ID	Claim	Value	Kind	Source file	Location
P08-C117	LOC110436259, SbANR-G1 with SbANR-G2, residues before first stop edited	188	computed	`data/results/pair_deletions.csv`	target=LOC110436259; pair SbANR-G1 + SbANR-G2; column residues_before_first_stop_edited
P08-C118	LOC8066840, SbANR-G1 with SbANR-G2, residues before first stop edited	184	computed	`data/results/pair_deletions.csv`	target=LOC8066840; pair SbANR-G1 + SbANR-G2; column residues_before_first_stop_edited
P08-C119	LOC8066841, SbANR-G1 with SbANR-G2, residues before first stop edited	184	computed	`data/results/pair_deletions.csv`	target=LOC8066841; pair SbANR-G1 + SbANR-G2; column residues_before_first_stop_edited
P08-C120	LOC8066842_A, SbANR-G3 with SbANR-G4, coding bases removed	271	computed	`data/results/pair_deletions.csv`	target=LOC8066842_A; pair SbANR-G3 + SbANR-G4; column coding_bases_removed
P08-C121	LOC8066842_B, SbANR-G3 with SbANR-G4, coding bases removed	271	computed	`data/results/pair_deletions.csv`	target=LOC8066842_B; pair SbANR-G3 + SbANR-G4; column coding_bases_removed
P08-C122	LOC8066844, SbANR-G3 with SbANR-G4, coding bases removed	268	computed	`data/results/pair_deletions.csv`	target=LOC8066844; pair SbANR-G3 + SbANR-G4; column coding_bases_removed
P08-C123	LOC8066842_A, SbANR-G3 with SbANR-G4, residues before first stop edited	197	computed	`data/results/pair_deletions.csv`	target=LOC8066842_A; pair SbANR-G3 + SbANR-G4; column residues_before_first_stop_edited
P08-C124	LOC8066842_B, SbANR-G3 with SbANR-G4, residues before first stop edited	197	computed	`data/results/pair_deletions.csv`	target=LOC8066842_B; pair SbANR-G3 + SbANR-G4; column residues_before_first_stop_edited
P08-C125	LOC8066844, SbANR-G3 with SbANR-G4, residues before first stop edited	139	computed	`data/results/pair_deletions.csv`	target=LOC8066844; pair SbANR-G3 + SbANR-G4; column residues_before_first_stop_edited
P08-C126	LOC8066846, SbANR-G7 with SbANR-G8, coding bases removed	514	computed	`data/results/pair_deletions.csv`	target=LOC8066846; pair SbANR-G7 + SbANR-G8; column coding_bases_removed

Claim ID	Claim	Value	Kind	Source file	Location
P08-C127	LOC8066846, SbANR-G7 with SbANR-G8, residues before first stop edited	207	computed	`data/results/pair_deletions.csv`	target=LOC8066846; pair SbANR-G7 + SbANR-G8; column residues_before_first_stop_edited
P08-C128	LOC8066846, SbANR-G7 with SbANR-G9, coding bases removed	27	computed	`data/results/pair_deletions.csv`	target=LOC8066846; pair SbANR-G7 + SbANR-G9; column coding_bases_removed
P08-C129	Perfect sites of the six array guides	19	computed	`data/results/design_summary.csv`	metric=six_array_guides_perfect_sites
P08-C130	Bases from the first to the last perfect site of the six array guides	77,225	computed	`data/results/design_summary.csv`	metric=six_array_guides_site_span
P08-C131	Sites of SbANR-G2 at 3 mismatches outside the array	1	computed	`data/results/guide_table.csv`	guide=SbANR-G2; column sites_3mm_outside_window
P08-C132	Sites of SbANR-G3 at 3 mismatches outside the array	5	computed	`data/results/guide_table.csv`	guide=SbANR-G3; column sites_3mm_outside_window
P08-C133	Sites of SbANR-G4 at 3 mismatches outside the array	1	computed	`data/results/guide_table.csv`	guide=SbANR-G4; column sites_3mm_outside_window
P08-C134	Sites of SbANR-G8 at 3 mismatches outside the array	1	computed	`data/results/guide_table.csv`	guide=SbANR-G8; column sites_3mm_outside_window
P08-C135	Highest worst-guide MIT score over all complete six-guide sets	89.7	computed	`data/results/design_summary.csv`	metric=best_worst_guide_mit
P08-C136	Six-guide sets that reach that worst-guide MIT score	5,136	computed	`data/results/design_summary.csv`	metric=sets_reaching_best_mit

Claim ID	Claim	Value	Kind	Source file	Location
P08-C137	Complete six-guide covers rebuilt in the second method	124,186	computed	data/results/design_summary.csv	metric=six_guide_covers_rebuilt
P08-C138	Of those covers, sets that flank part 1	0	computed	data/results/design_summary.csv	metric=six_guide_covers_flanking_part1
P08-C139	Of those covers, sets that flank part 2	0	computed	data/results/design_summary.csv	metric=six_guide_covers_flanking_part2
P08-C140	Added candidates that cut only in LOC8074755	258	computed	data/results/design_summary.csv	metric=eighth_target_extra_candidates
P08-C141	Candidates with a perfect-site cut in part 1	155	computed	data/results/design_summary.csv	metric=candidates_cutting_part1
P08-C142	Candidates with a perfect-site cut in part 2	133	computed	data/results/design_summary.csv	metric=candidates_cutting_part2
P08-C143	Added candidates with a perfect site outside the array	0	computed	data/results/design_summary.csv	metric=added_candidates_perfect_outside_array
P08-C144	Single added guides that flank part 1 when the six are kept	47	computed	data/results/design_summary.csv	metric=single_additions_part1
P08-C145	Pairs of added guides that flank part 2 when the six are kept	3,119	computed	data/results/design_summary.csv	metric=pair_additions_part2
P08-C146	Pairs of added guides that flank both parts when the six are kept	140	computed	data/results/design_summary.csv	metric=pair_additions_both

Claim ID	Claim	Value	Kind	Source file	Location
P08-C147	Perfect sites of SbANR-G10	2	computed	`data/results/guide_table.csv`	guide=SbANR-G10; column perfect_sites
P08-C148	MIT score of SbANR-G10 outside the array	94.0	computed	`data/results/guide_table.csv`	guide=SbANR-G10; column mit_outside_window
P08-C149	Perfect sites of SbANR-G11	1	computed	`data/results/guide_table.csv`	guide=SbANR-G11; column perfect_sites
P08-C150	MIT score of SbANR-G11 outside the array	99.7	computed	`data/results/guide_table.csv`	guide=SbANR-G11; column mit_outside_window
P08-C151	LOC8074755_P1, SbANR-G1 with SbANR-G10, coding bases removed	404	computed	`data/results/pair_deletions.csv`	target=LOC8074755_P1; pair SbANR-G1 + SbANR-G10; column coding_bases_removed
P08-C152	LOC8074755_P1, SbANR-G1 with SbANR-G10, residues before first stop edited	86	computed	`data/results/pair_deletions.csv`	target=LOC8074755_P1; pair SbANR-G1 + SbANR-G10; column residues_before_first_stop_edited
P08-C153	LOC8074755_P2, SbANR-G11 with SbANR-G10, coding bases removed	273	computed	`data/results/pair_deletions.csv`	target=LOC8074755_P2; pair SbANR-G11 + SbANR-G10; column coding_bases_removed
P08-C154	Perfect sites of the eight array guides	22	computed	`data/results/design_summary.csv`	metric=eight_array_guides_perfect_sites
P08-C155	Bases from the first to the last perfect site of the eight array guides	77,225	computed	`data/results/design_summary.csv`	metric=eight_array_guides_site_span
P08-C156	Perfect sites of SbANR-gA	2	computed	`data/results/guide_table.csv`	guide=SbANR-gA; column perfect_sites

Claim ID	Claim	Value	Kind	Source file	Location
P08-C157	MIT score of SbANR-gA outside the array	99.3	computed	`data/results/guide_table.csv`	guide=SbANR-gA; column mit_outside_window
P08-C158	MIT score of SbANR-gB outside the array	96.7	computed	`data/results/guide_table.csv`	guide=SbANR-gB; column mit_outside_window
P08-C159	LOC8066842_A, SbANR-gA with SbANR-gB, coding bases removed	318	computed	`data/results/pair_deletions.csv`	target=LOC8066842_A; pair SbANR-gA + SbANR-gB; column coding_bases_removed
P08-C160	Copies of the array left uncut by the two-copy option	5	computed	`data/results/coverage_matrix.csv`	columns gA and gB: 5 of the 7 rows have no perfect site
P08-C161	Cells equal to the stored matrix	90	computed	`data/results/design_summary.csv`	metric=coverage_cells_equal
P08-C162	Guide-by-target cells compared with the stored matrix	90	computed	`data/results/design_summary.csv`	metric=coverage_cells_compared
P08-C163	Perfect sites of G1 in the independent second script	6	computed	`research_record/second_script_rerun/s5_verify_output.txt`	line G1 total perfect sites: 6
P08-C164	Perfect sites of G2	4	computed	`research_record/second_script_rerun/s5_verify_output.txt`	line G2 total perfect sites: 4

Claim ID	Claim	Value	Kind	Source file	Location
P08-C165	Perfect sites of G3	4	computed	`research_record/second_script_rerun/s5_verify_output.txt`	line G3 total perfect sites: 4
P08-C166	Perfect sites of G4	3	computed	`research_record/second_script_rerun/s5_verify_output.txt`	line G4 total perfect sites: 3
P08-C167	Perfect sites of G10 in the independent second script	2	computed	`research_record/second_script_rerun/a3_p08_out.txt`	line G10 perfect sites: 2
P08-C168	Design target for grape skin PA (ug catechin equivalents/g FW)	50	design target	`inputs/design_targets.csv`	row Grape berry skin, column target_below
P08-C169	Assumed wild-type range of grape skin PA (ug/g FW)	300 to 600	design assumption	`inputs/design_targets.csv`	row Grape berry skin, columns wild_type_low and wild_type_high
P08-C170	Reduction for grape skin at the lowest wild-type value (percent)	83.33	design target	`data/results/pa_reductions.csv`	tissue=Grape berry skin
P08-C171	Reduction for grape skin at the highest wild-type value (percent)	91.67	design target	`data/results/pa_reductions.csv`	tissue=Grape berry skin
P08-C172	Design target for grape seed PA (mg catechin equivalents/g FW)	0.5	design target	`inputs/design_targets.csv`	row Grape berry seed, column target_below

Claim ID	Claim	Value	Kind	Source file	Location
P08-C173	Assumed wild-type range of grape seed PA (mg/g FW)	8 to 12	design assumption	`inputs/design_targets.csv`	row Grape berry seed
P08-C174	Reduction for grape seed at the lowest wild-type value (percent)	93.75	design target	`data/results/pa_reductions.csv`	tissue=Grape berry seed
P08-C175	Reduction for grape seed at the highest wild-type value (percent)	95.83	design target	`data/results/pa_reductions.csv`	tissue=Grape berry seed
P08-C176	Design target for sorghum grain PA (percent of dry weight)	0.05	design target	`inputs/design_targets.csv`	row Sorghum grain, column target_below
P08-C177	Assumed wild-type range of sorghum grain PA (percent of dry weight)	0.5 to 3	design assumption	`inputs/design_targets.csv`	row Sorghum grain
P08-C178	Reduction for sorghum grain at the lowest wild-type value (percent)	90.00	design target	`data/results/pa_reductions.csv`	tissue=Sorghum grain
P08-C179	Reduction for sorghum grain at the highest wild-type value (percent)	98.33	design target	`data/results/pa_reductions.csv`	tissue=Sorghum grain

Appendix B. Glossary, abbreviations and units

Term	Meaning	Type
ANR	Anthocyanidin reductase: the enzyme that turns anthocyanidins into the flavan-3-ol units of condensed tannins (called BAN in Arabidopsis)	abbreviation
Astringency	The dry, puckering feel caused by condensed tannins binding salivary proteins	term
Proanthocyanidins (PA)	Condensed tannins: chains of flavan-3-ol units	term
PA	Proanthocyanidins	abbreviation
Catalytic Tyr codon	The codon of the Tyr in the YxxxK motif of the enzyme, which aligns with Tyr168 of SbANR4	term
YxxxK motif	A Tyr and a Lys four residues apart; the catalytic pair of the enzyme	term
CRISPR-Cas9	A genome-editing system in which a guide RNA directs the Cas9 nuclease to a DNA site	abbreviation
Guide	A 20-base protospacer that directs Cas9 to one site	term
Protospacer	The 20 bases of the target site that match the guide RNA	term
PAM	Protospacer adjacent motif: the NGG bases next to the protospacer that SpCas9 needs	abbreviation
SpCas9	Cas9 nuclease of Streptococcus pyogenes	abbreviation
Perfect site	A genome site with no mismatch to the guide and an NGG PAM	term
Mismatch	A base of a site that differs from the guide	term
MIT score	Specificity score of Hsu et al. (2013): 100 means no off-target sites; lower values mean more or closer off-target sites	term

Term	Meaning	Type
Flanking pair	Two guides whose cuts lie either side of the catalytic Tyr codon of one target, so that the piece between them is removed	term
Clean join	The two cut ends are joined without added or lost bases	term
In frame / frameshift	Removal of a multiple of three coding bases keeps the reading frame (in frame); any other number shifts it (frameshift)	term
Coding index	Position of a base in the coding sequence, counting from the first base of the start codon	term
Array window	NC_012875.2:57,100,000-57,300,000, the region of sorghum chromosome 6 that holds the ANR array	term
Copy A / copy B	The two ANR copies inside the merged RefSeq model LOC8066842, named SbANR4 (copy A) and SbANR5 (copy B) of LOC8066842	term
Part 1 / part 2	The two ANR-like parts of the transcribed pseudogene LOC8074755	term
TAN1	TANNIN1: a sorghum WD40-repeat protein needed for tannin production	abbreviation
IFI	Iyengar Flavor Index: scores one compound against a reference receptor assay	abbreviation
IFI-A	Astringency axis of the IFI, with tannic acid as the reference compound	abbreviation
IFI-B	Bitterness-suppression axis of the IFI, with GIV3727 at TAS2R31 as the reference	abbreviation
EC50 / IC50	Concentration giving half of the maximal effect (EC50) or half of the maximal block (IC50)	abbreviation
TAS2R31	Human bitter taste receptor 31	abbreviation
GIV3727	A bitter-taste receptor blocker	term
RefSeq	The NCBI reference sequence collection	abbreviation

Term	Meaning	Type
LOC ID	The NCBI Gene symbol LOC followed by the Gene ID, used for genes without a published name	abbreviation
UTR	Untranslated region of a transcript	abbreviation
GC	Share of G and C bases in a protospacer	abbreviation
bp	Base pair	unit
aa	Amino acid residue	unit
nt	Nucleotide	unit
Mb	Megabase (1,000,000 bases)	unit
ug/g FW	Micrograms per gram of fresh weight	unit
mg/g FW	Milligrams per gram of fresh weight	unit
% DW	Percent of dry weight	unit
uM	Micromolar	unit
BLOSUM62	Amino-acid substitution matrix used for the alignments	term
HiGHS	The solver used for the integer programme of the eighth-target search	term

Appendix C. Computational environment

Item	Value
Python	3.15.2
Platform	Darwin x86_64
Libraries	standard library only
Steps and run time	1 code/01_identity.py: 0.1 s; 2 code/02_guides.py: 0.1 s; 3 code/03_array.py: 0.1 s; 4 code/04_reductions.py: 0.1 s

Item	Value
Input files	27 files, SHA-256 in `code/environment.json`
Data fingerprint	`e01499f1c3db2445747eb8e9c97d5ff9bd48f6453affcb398acc9f4251e01165`

LEGAL NOTICES

Research use only. All results in this paper are computed designs, predictions and targets. They are not food, a dietary supplement, a drug or a consumer product, and they are not medical, dietary, agronomic, regulatory or financial advice. Concentrated capsaicinoids can cause severe pain and injury. Do not expose any person or animal to material derived from these designs.

No warranty; limitation of liability. This paper, its data and its code are provided "as is" without warranty of any kind. To the fullest extent permitted by law, Scoville Splice LLC and the authors are not liable for any claim, loss or damage arising from their use. Users are responsible for complying with all biosafety, biosecurity, plant-health, food-safety, import and export, and other laws that apply to them, and for obtaining every approval needed before any laboratory, greenhouse or field work.

Licence. The paper and the data are licensed under Creative Commons Attribution 4.0 International (CC BY 4.0); the code is licensed under the MIT licence. These licences cover copyright in the files only and grant no right under any patent, patent application, trade secret or trademark. See `legal/DISCLAIMER.txt`, `legal/TERMS_OF_USE.txt` and `legal/NOTICE.txt` in the data package.

DOI 10.5281/zenodo.24511334 · US provisional applications 64/172,827, 64/188,197, 64/192,062 and 64/199,105 · © 2026 Scoville Splice LLC and P.I. Ford · Published under CC BY 4.0 · scovillesplice.farm · ORCID: 0000-0003-0395-1752