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.

Sweet Heat: In Silico Design of Fruit-Specific Brazzein Expression in Capsicum annuum

Priyanca Ford

scovillesplice@gmail.com

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

Research Article

Keywords: Capsicum annuum, brazzein, sweet protein, E8 promoter, codon adaptation index, signal peptide, CRISPR-Cas9 guide design, knock-in cassette

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11875875/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

Sweet Heat: In Silico Design of Fruit-Specific Brazzein Expression in Capsicum annuum

Three computed expression cassettes of the sweet protein brazzein under the tomato E8 promoter, with a chromosome 2 knock-in site, a two-guide pair and repair templates for Capsicum annuum.

P.I. Ford

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

DOI: 10.5281/zenodo.24511316 License: CC BY 4.0 **Patents:** US 64/172,827; US 64/188,197; US 64/192,062; US 64/199,105

Figshare: figshare/35015803 OSF: osf.io/cp6qk Hugging Face: capsaicinoid-data GitHub: paper-03 Website: sweet-heat-brazzein

Package record. 182 claims traced to a file or a cited source (137 computed, 21 definition, 18 design assumption, 1 design target, 5 literature value). 23 tables, 8 figures, 75 data files. Data fingerprint `279e903f4f2c`. Rerun every result with `bash run_all.sh`.



Phoenix fruit. Scoville Splice LLC image library.

ABSTRACT

Brazzein is a sweet protein of 54 residues (UniProt P56552). This paper describes a computed design that expresses it in the fruit of *Capsicum annuum* under the tomato E8 promoter. The design has three cassettes of 1,623, 1,626 and 1,617 bp. Each joins a 1,121-bp E8 promoter piece (GenBank X13437), a signal-brazzein reading frame and a 256-bp NOS terminator. The

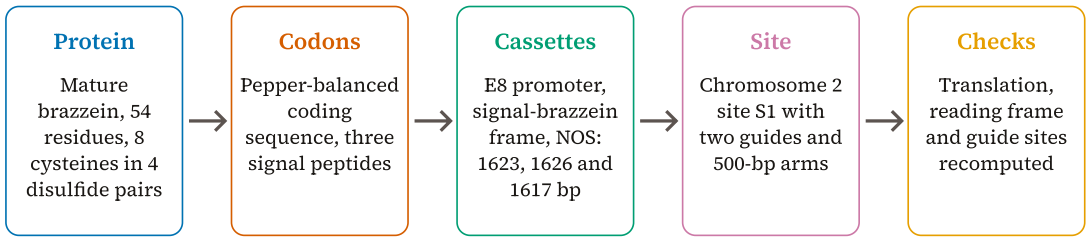
reading frames carry the signal of a pepper defensin (Q43413), an Arabidopsis endochitinase (Q9M2U5) or a pepper proteinase inhibitor (O49146) in front of mature brazzein. Codons follow pepper usage, with a codon adaptation index of 0.860 to 0.897 and a GC content of 36.5 to 39.4%. The insertion site S1 lies on chromosome 2 (NC_061112.1). Two guides, each with 1 perfect genome site, cut 101 bp apart, and the cassette replaces the sequence between the cuts. Repair templates with 500-bp homology arms are 2,617 to 2,626 bp long. Translation, reading frames and guide sites were recomputed with a second code path. All results are computed. No experimental data are reported in this paper.

Keywords: Capsicum annuum · brazzein · sweet protein · E8 promoter · codon adaptation index · signal peptide · CRISPR-Cas9 guide design · knock-in cassette

KEY POINTS

- Three cassettes of 1,617 to 1,626 bp join the E8 promoter, one of three signals with mature brazzein, and the NOS terminator.
- Pepper-balanced codons give a codon adaptation index of 0.860 to 0.897, and none of the listed motifs occurs in the reading frames.
- Two guides at the chromosome 2 site S1 cut 101 bp apart; each has 1 perfect genome site (MIT scores 64.0 and 70.3).
- Repair templates of 2,617 to 2,626 bp carry 500-bp arms; 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 brazzein protein, balanced pepper codons with three signals, three cassettes, the chromosome 2 site with its guide pair, and the recomputed checks. Drawn from the steps file. Data: `data/figures/GA.csv`; script: `code/05_tables.py`.

1. Introduction

Brazzein is a small sweet protein from the plant *Pentadiplandra brazzeana*. It was reported as a high-potency, thermostable sweet protein (Ming and Hellekant, 1994), and its solution structure was reported (Caldwell et al., 1998). It is reported to dock mainly to the T1R2 subunit of the sweet taste receptor (Walters and Hellekant, 2006), a TAS1R2/TAS1R3 heteromer (Li et al., 2002). Brazzein is the reference compound of the sweet axis of the Iyengar Flavor Index (IFI-S; 10.5281/zenodo.24511320), with a reference EC50 of 0.004

micromolar. The IFI scores one compound against one reference assay, so this paper uses it only to name the reference.

E8 is an ethylene-activated tomato gene, and its 5' flanking region is reported to give ripening-related expression (Deikman and Fischer, 1988). The pepper defensin J1-1 (UniProt Q43413) is reported to be expressed in ripe pepper fruit (Meyer et al., 1996), and its record lists fruit as the tissue. These two elements are the starting point for a fruit-expressed brazzein design in pepper.

This paper designs three cassettes, one for each of three secretion signals, and a knock-in site on chromosome 2 of the UCD10Xv1.1 reference with a two-guide pair. The host design is Phoenix (SS-0096), the cultivar of 10.5281/zenodo.24511300 with the highest heat design target, 2,956,310 SHU (Scoville heat units). The same site is used in 10.5281/zenodo.24511332 and 10.5281/zenodo.24511340. The paper reports the protein facts of brazzein, the signal sequences, the pepper-balanced codon design, the cassette sequences with their coordinates, the site search, the guide pair with its genome-wide site counts, the repair templates and the agreement of independent code paths.

2. Methods

2.1 Computational basis

Genome positions refer to the RefSeq assembly GCF_002878395.1 (*Capsicum annuum* UCD10Xv1.1; Hulse-Kemp et al., 2018). Protein and signal sequences come from stored UniProt flat files. Scripts run on Python 3.9 or newer with the standard library only; the C scanner is in `code/guide_design/`. The research record holds the research scripts, their raw outputs and the second implementations used for cross-checks.

2.2 Brazzein and signal records

Mature brazzein is UniProt P56552 (entry version 93); its chain, cysteine positions and disulfide pairs are read from the record. The average mass is the sum of the IUPAC average residue masses plus 18.01528 Da for water; each formed disulfide removes two hydrogen atoms of 1.00794 Da. The three signals are the signal features of Q43413, Q9M2U5 and O49146 (entry versions 96, 162 and 69); the residues at −3 and −1 are read from each signal sequence (`code/01_protein.py`).

2.3 Codon usage and balanced back-translation

The pepper codon table counts every codon of the 48,931 clean coding sequences of the RefSeq coding-sequence file (ATG start, terminal stop, only A, C, G and T, length divisible by three, no internal stop); it holds 21,189,723 codons. The relative adaptiveness w of a codon is its count divided by the count of the most-used codon of the same amino acid. The codon adaptation index (CAI; Sharp and Li, 1987) of a coding sequence is the geometric mean of w over its codons, without Met, Trp and stop codons: exp((1 / n) × sum of ln w). A second code path takes the n-th root of the product, and the two agree to three decimals. A one-per-gene count (the longest coding sequence of each gene) gives the same most-used codon for 19 of 20 amino acids. The pepper coding sequences average 42.3% GC and 37.5% GC at the third codon position (GC3). The terminal stop codon is TGA, the most used in pepper (shares of TGA, TAA and TAG: 41.5%, 33.7% and 24.8%).

For each amino acid the usable codons are those with w of at least 0.5. A codon is drawn with probability proportional to its pepper count. A draw is repaired, for up to 300 rounds, by redrawing codons that overlap one of 15 listed motifs (ATTTA, AATAAA, TTTT, AAAAAA, and the sites of BsaI, BsmBI, EcoRI, BamHI, HindIII, SalI, NcoI, SacI and XhoI) or a run of five identical bases. A draw is kept when no listed motif is left and the GC of the reading frame lies within 3 points of the pepper average; the window is 6

points for the Q9M2U5 and O49146 frames, whose signals are rich in Leu, Ser and Thr. Each frame has 3,000 seeded draws, and the kept draw with the highest CAI is the design. The seed is 24112026 for the Q43413 frame. Its mature codons are fixed in the other two frames, whose signal codons use the seeds 2439 and 2436. Every design is scanned for 17 motifs (Table S3, Table S4).

2.4 Promoter and terminator

The promoter piece is the 5' region of the tomato E8 gene (GenBank X13437; 3,042 bp; cultivar Cherry), bases 1 to 1,121. The record places the TATA box at 1,055 to 1,063, the transcription start at 1,087 and the E8 start codon at 1,122. The piece ends one base before that start codon, so its 5' leader, from the transcription start to the end of the piece, is 35 nt. Two further tomato records of the E8 promoter region, DQ317599 (2,175 bp) and AF515784 (2,191 bp), were compared with X13437 bases 1 to 1,086: by exact string search against DQ317599, and by a global alignment against the start of AF515784 with free end gaps in AF515784 (scores for a match, a mismatch and a gap column: 2, -3 and -4). The NOS terminator is 256 bp, bases 4022 to 4277 of AF485783.

2.5 Cassette assembly and checks

A cassette is the promoter piece, the signal-brazzein reading frame and the NOS terminator in this order on the plus strand. The script checks that the reading frame translates to the signal plus mature brazzein, that its only stop codon is the last codon, that the first ATG after the transcription start is the first base of the reading frame, that the frame is the longest ATG-to-stop frame on either strand, and it lists nine cloning sites (EcoRI, BamHI, HindIII, XbaI, SacI, BsaI, BsmBI, SmaI/XmaI and KpnI) by position (`code/03_cassettes.py`). The cassettes were compared with the sequences written by the research scripts (3 of 3). The length sums were repeated with the bc calculator (1,623, 1,626 and 1,617), and a second script translated each reading frame with its own genetic code and aligned the mature part (global alignment, BLOSUM62, gap open −11, extend −1) to a freshly fetched copy of P56552 (54 of 54).

2.6 Site search

The search covers the 12 placed chromosomes. Windows of 3,000 bp are tested every 500 bp. A window must keep 5,000 bp from every annotated gene or pseudogene on both sides, hold at most 10% repeat-masked (soft-masked) bases and no N, and lie in a gap of 13 to 60 kb between two protein-coding genes. For each gap the window with the lowest repeat fraction, nearest the gap centre, is kept. The expression of the two neighbours is from the public placenta series PRJNA789050: reads of *Capsicum chinense* SL08 at the stages 4, 24, 34, 44 and 54 days post anthesis (DPA), one library per stage, quantified with salmon against the *C. annuum* coding sequences, in transcripts per million (TPM) summed per gene. The insertion point is the midpoint of the window (Table S10).

2.7 Guide design, arms and repair templates

Every SpCas9 site with a strict NGG PAM on either strand within 110 bp either side of the insertion point was listed. A candidate passes with GC of 40 to 70% (no TTTT, no run of five identical bases). Positions are read from the window sequence (1,101 bp), and the cut lies 3 bp from the PAM. Our own C program scans the genome for every site with a strict NGG PAM, on both strands, with up to 4 mismatches; it lists 448 sites for the two chosen guides. Mismatch positions are numbered 1 to 20 from the PAM-distal end. The MIT score of a guide (our implementation of Hsu et al., 2013) is 100 / (100 + the sum of its site scores) × 100. The score of one site with 1 or more mismatches is the product of $(1 - w)$ over its mismatch positions, where w is the position weight (20 position weights from 0 to 0.851; Table S15), times $1 / (((19 - \bar{d}) / 19) \times 4 + 1)$ when there is more than one mismatch, where $\bar{d}$ is the mean distance between mismatches, times $1 / n^2$ for n mismatches, times 100. Sites with one to three mismatches were annotated with the RefSeq annotation as coding exon, gene body or between genes. A pair has one cut on each side of the insertion point, 20 to 120 bp apart; pairs of guides with exactly one perfect site are ranked by the lower MIT score

and then by the smaller distance. The homology arms are the 500 bp that end at the first cut and the 500 bp that start after the second cut. A repair template is the left arm, the cassette and the right arm. A second scanner written in numpy, which shares no code with the C program, repeated the genome scan (221 and 227).

3. Results

3.1 Mature brazzein

Table 1 and Figure 2 give the mature chain: 54 residues with 8 cysteines at 4, 16, 22, 26, 37, 47, 49 and 52, joined in 4 disulfide pairs (4-52, 16-37, 22-47 and 26-49). The record gives an average mass of 6,498 Da; computed from the sequence it is 6,498.32 Da for the unmodified chain and 6,490.26 Da with the four disulfides formed. The record lists pyroglutamate at residue 1. The chain has no methionine or tryptophan, so every codon of the chain enters the codon adaptation index.

Table 1. Mature brazzein. Facts of the mature chain from the stored UniProt record, with the average mass computed from the sequence (IUPAC average residue masses plus water). The disulfide pairs use each of the eight cysteines once. Data: `data/tables/T01.csv`; *script:* `code/01_protein.py`.

Property	Value
UniProt accession	P56552
Protein name in the record	Defensin-like protein
Source organism	Pentadiplandra brazzeana (Oubli fruit tree)
Mature chain (residues)	1 to 54
Length (residues)	54
Sequence	QDKCKKVYENYPVSKCQLANQCNYDCKLDKHARSGECFYDEKRNLQCICDYCEY
Cysteine positions	4 16 22 26 37 47 49 52
Disulfide pairs in the record	4-52 16-37 22-47 26-49
Record average mass (Da)	6498
Own average mass, unmodified (Da)	6498.32
Own average mass with four disulfides formed (Da)	6490.26
N-terminal pyroglutamate listed at residue	1

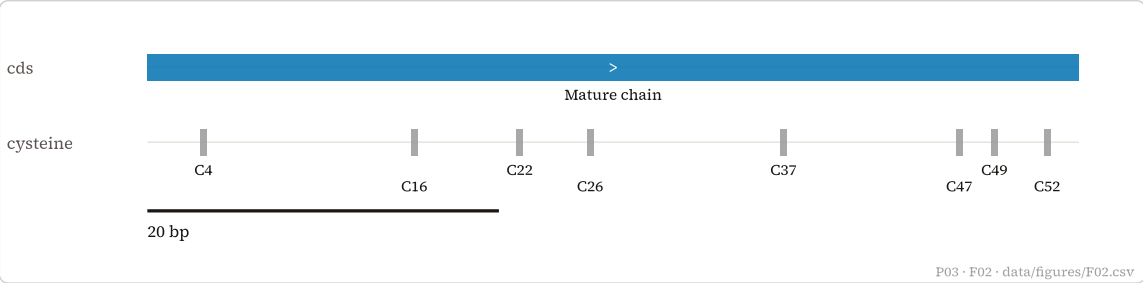


Figure 2. Mature brazzein and its cysteines. The 54-residue mature chain and the eight cysteines (C4, C16, C22, C26, C37, C47, C49, C52) that form four disulfide pairs. Positions read from the UniProt record P56552. Data: `data/figures/F02.csv`; script: `code/01_protein.py`.

3.2 Three secretion signals

Table 2 lists the signals. The Q43413 signal of the pepper defensin J1-1 is 27 residues, the Q9M2U5 signal of the Arabidopsis endochitinase EP3 is 28 residues and the O49146 signal of the pepper wound-induced proteinase inhibitor 2 is 25 residues. The Q43413 signal carries evidence code ECO:0000269, PubMed 8883377; the other two carry ECO:0000255. The residues next to the expected cleavage site are Ala at -3 and -1 for Q43413, Thr at -3 and Ala at -1 for Q9M2U5 and Val at -3 and Ala at -1 for O49146. Alanine at the last position and a small residue at the third last are the usual pattern at signal peptidase sites (von Heijne, 1986). The Q43413 record lists fruit as its tissue, and the O49146 record lists pericarp. The first residue of mature brazzein follows each signal directly.

Table 2. Three secretion signals. The three signals placed in front of mature brazzein, one per cassette. Evidence codes are those of the UniProt records. Data: `data/tables/T02.csv`; script: `code/01_protein.py`.

Accession	Protein	Source	Signal (aa)	Signal sequence	Evidence in the record	Tissue in the record
Q43413	Defensin J1-1	Capsicum annuum (Capsicum pepper)	27	MAGFSKVVATIFLMMLLVFATDMMAEA	ECO:0000269|PubMed:8883377	Fruit
Q9M2U5	Endochitinase EP3	Arabidopsis thaliana (Mouse-ear cress)	28	MLTPTISKSISLVTILLVLQAFSNTTKA	ECO:0000255	
O49146	Wound-induced proteinase inhibitor 2	Capsicum annuum (Capsicum pepper)	25	MAVPKEVSFLASLLVLGILLLHVDA	ECO:0000255	Pericarp

3.3 Pepper-balanced codon designs

Table 3, Figure 3 and Figure 4 give the three reading frames. They are 246, 249 and 240 nt long (Q43413, Q9M2U5 and O49146) and encode 81, 82 and 79 residues before the stop codon. The CAI is 0.860, 0.892 and 0.897; against the one-per-gene counts it is 0.859, 0.889 and 0.893. Their GC and GC3 are 39.4% GC and 48.8% GC3 (Q43413), 36.5% GC and 37.3% GC3 (Q9M2U5) and 37.9% GC and 35.0% GC3 (O49146), close to the pepper average of 42.3% GC. None of the 17 listed motifs occurs in the three frames. The Q43413 frame has 30 kept draws of 3,000, the Q9M2U5 frame 999 and the O49146 frame 2,990 (Table S4).

The mature chain alone has CAI 0.861, 38.1% GC and 48.2% GC3 with balanced codons. Choosing the most-used pepper codon for every residue gives 1.000, 24.4% GC and 1.8% GC3, and the sequence holds TTTT at 2 positions and AAAAAA at 2 positions. The balanced design changes 27 of the 54 codons (Table S2); every codon and its weight is in Table S1. The reading frames share the 54 mature codons and the 421 bp from the first mature codon to the end of each cassette.

Table 3. Codon designs. Reading frames and mature coding sequences back-translated with pepper codon usage. CAI is the codon adaptation index against the pepper codon table; GC3 is the GC share at the third codon position. The two mature-chain rows compare the most-used and the balanced codon choice. Data: `data/tables/T03.csv`*; script:* `code/02_codon_design.py`*.*

Design	ORF (nt)	Protein (aa)	CAI	GC (%)	GC3 (%)	Listed motifs absent
Q43413	246	81	0.860	39.4	48.8	yes
Q9M2U5	249	82	0.892	36.5	37.3	yes
O49146	240	79	0.897	37.9	35.0	yes
mature, most-used codons	168	54	1.000	24.4	1.8	no
mature, balanced codons	168	54	0.861	38.1	48.2	yes

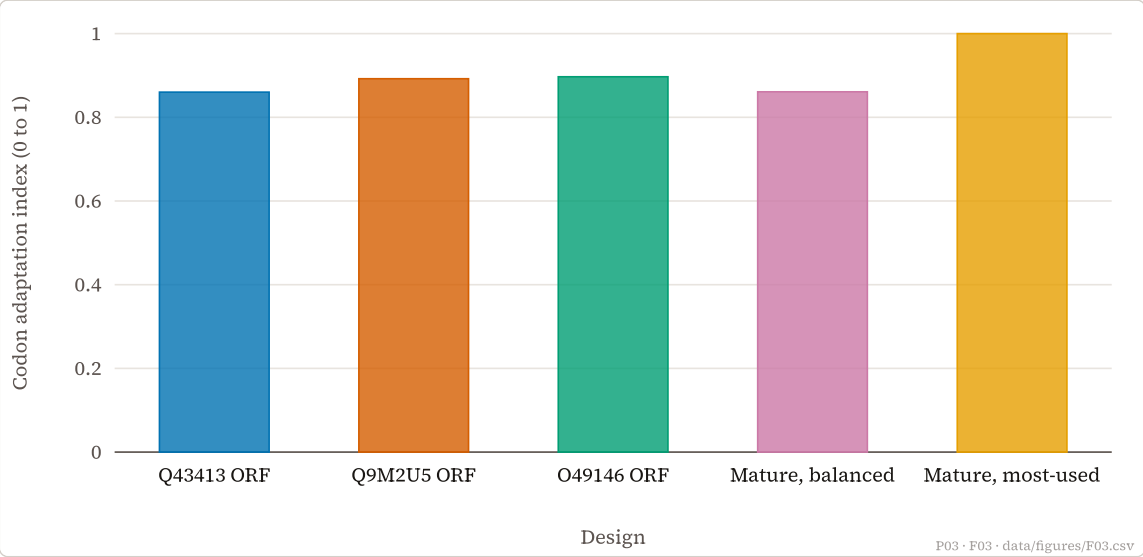


Figure 3. Codon adaptation index of the designs. Index of the three balanced reading frames and of the mature chain with balanced and with most-used codons, against the pepper codon table. Geometric mean of the relative adaptiveness of every codon (Sharp and Li, 1987). Data: `data/figures/F03.csv`; script: `code/02_codon_design.py`.



Figure 4. GC content of the coding designs. GC content of the pepper coding-sequence average, the mature chain with most-used codons and with balanced codons (bars 2 and 3), and the three reading frames. Share of G and C bases. Data: `data/figures/F04.csv`; script: `code/02_codon_design.py`.

3.4 The three cassettes

Table 4 and Figure 1 give the cassettes: 1,623 bp with the Q43413 signal, 1,626 bp with Q9M2U5 and 1,617 bp with O49146, with 29.0%, 28.5% and 28.7% GC. In the Q43413 cassette the promoter piece occupies positions 1 to 1,121, the signal 1,122 to 1,202 (81 nt), mature brazzein 1,203 to 1,364 (162 nt), the stop codon 1,365 to 1,367 and the NOS terminator 1,368 to 1,623 (Table 5; blocks of the other two cassettes are in Table S5). The leader of 35 nt holds no ATG, so the first ATG after the transcription start is at position 1,122. The designed frame is the longest ATG-to-stop frame on either strand of each cassette (81, 82 and 79 residues; the next longest in the Q43413 cassette is 68). The cassettes hold none of 9 listed cloning sites (Table S7). The three cassettes are identical outside the signal: the promoter piece is the same in all three, and so are the mature codons, the stop codon and the NOS terminator (Table S6).

Table 4. The three cassettes. Each cassette is the 1,121-bp E8 promoter piece, the signal-brazzein reading frame and the 256-bp NOS terminator. The repair template adds a 500-bp homology arm on each side. Data: `data/tables/T04.csv`*; script:* `code/03_cassettes.py`.

Variant (signal)	Signal (aa)	ORF (nt)	Cassette (bp)	Cassette GC (%)	Repair template (bp)
Q43413	27	246	1623	29.0	2623
Q9M2U5	28	249	1626	28.5	2626
O49146	25	240	1617	28.7	2617

Table 5. Blocks of the Q43413 cassette. Coordinates in the 1,623-bp cassette. The 5' leader lies inside the promoter piece. Blocks of the other two cassettes are in the supplementary tables. Data: `data/tables/T05.csv`*; script:* `code/03_cassettes.py`.

Block	Start	End	Length (bp)	Source
E8 promoter and 5' leader (X13437 1-1121)	1	1121	1121	tomato E8 gene record X13437
5' leader (transcript start to ATG)	1087	1121	35	X13437 1087-1121
Signal Q43413 (27 aa)	1122	1202	81	UniProt Q43413 residues 1-27
Mature brazzein (54 aa)	1203	1364	162	UniProt P56552 residues 1-54
Stop TGA	1365	1367	3	terminal stop codon
NOS terminator (256 bp)	1368	1623	256	NOS terminator part

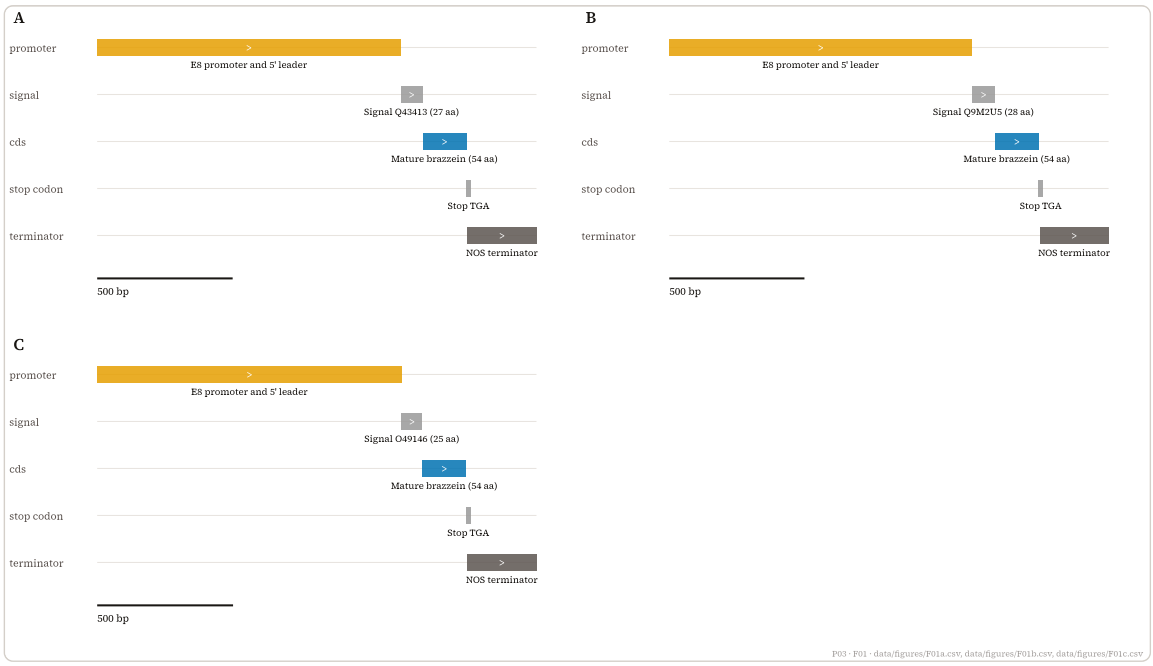


Figure 1. The three expression cassettes. A, Q43413 cassette (1,623 bp). B, Q9M2U5 cassette (1,626 bp). C, O49146 cassette (1,617 bp). E8 promoter piece, signal, mature brazzein, stop codon and NOS terminator along the cassette; the scale bar is in base pairs. Block coordinates computed from the part lengths. Data: `data/figures/F01a.csv`, `data/figures/F01b.csv`, `data/figures/F01c.csv`; script: `code/03_cassettes.py`.

3.5 The E8 promoter piece

Table S8 lists the E8 records. X13437 is 3,042 bp and its first 1,086 bases occur unchanged in DQ317599 starting at position 1,090 (1,086 of 1,086 bases identical). Aligned to AF515784 the same bases give 1,081 of 1,086 columns (99.5%). The three records of three tomato cultivars therefore agree over the promoter piece used here.

3.6 Site S1 on chromosome 2

Table 7 and Figure 8 describe the site. The annotation lists 40,503 genes and pseudogenes, 27,958 protein-coding. The search found 38,660 gaps between neighbouring genes; 10,543 are 13 to 60 kb long and 5,328 of those have protein-coding genes on both sides (Table S9). A window with at most 10% repeat-masked bases fits in 1,129 gaps, and 9 of those windows hold no repeat-masked base. S1 is the window 131,429,594 to 131,432,593 on NC_061112.1 (3,000 bp) with a repeat-masked fraction of 0.0. The insertion point is 131,431,093. The left neighbour is LOC107859470, 38,000 bp from the insertion point, and the right neighbour is LOC107858201, 8,292 bp away; the gap between them is 46,291 bp, and 12 protein-coding genes lie within 100 kb of the window centre. The two neighbours are expressed in placenta at all five stages: the left at 7.1 to 15.2 TPM and the right at 8.6 to 20.4 TPM.

Table 7. Site S1. The 3,000-bp window on chromosome 2 and its two neighbouring genes. TPM is transcripts per million in the public placenta series, five stages, one library per stage. Data: `data/tables/T07.csv`; script: `code/04_site_guides.py`.

Item	Value
Chromosome (RefSeq)	NC_061112.1
Window	131,429,594-131,432,593
Insertion point	131,431,093

Item	Value
Repeat-masked fraction of the window	0.0
Left neighbour (LOC ID)	LOC107859470
Left neighbour distance from the insertion point (bp)	38,000
Right neighbour (LOC ID)	LOC107858201
Right neighbour distance from the insertion point (bp)	8,292
Protein-coding genes within 100 kb of the window centre	12
Left neighbour placenta TPM, lowest of five stages	7.1
Left neighbour placenta TPM, highest of five stages	15.2
Right neighbour placenta TPM, lowest of five stages	8.6
Right neighbour placenta TPM, highest of five stages	20.4

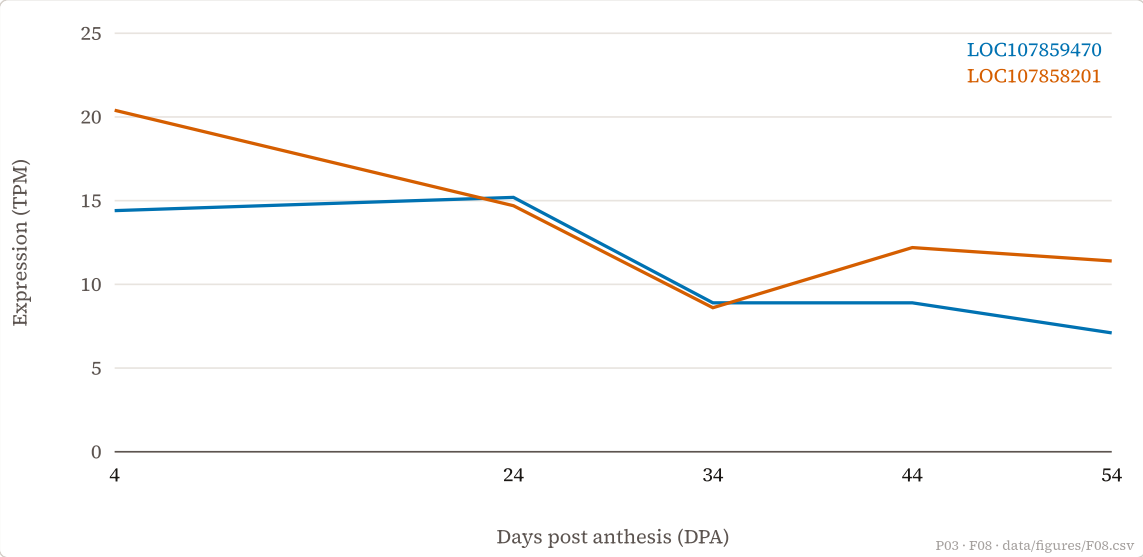


Figure 8. Placenta expression of the two genes next to S1. Transcripts per million of LOC107859470 and LOC107858201 in the public placenta series, one library per stage. Per-transcript quantification summed per gene. Data: `data/figures/F08.csv`; script: `code/04_site_guides.py`.

3.7 Guide pair, arms and repair templates

Table 6 and Figure 5 give the pair. S1-g1 (GTTGGAGCGAGTACTATGCA, PAM AGG) has its protospacer at 131,430,997 to 131,431,016 and cuts after 131,430,999; S1-g2 (ACCTAATGGGTTGCAAGGAA, PAM TGG) is at 131,431,098 to 131,431,117 and cuts after 131,431,100. Both lie on the minus strand. The cuts are 94 bp left and 7 bp right of the insertion point and 101 bp apart, so the cassette replaces the 101 bp between them (Figure 5). The guides have 50% and 45% GC. Each has 1 perfect site in the genome. S1-g1 has 0, 9, 25 and 186 sites with one, two, three and four mismatches (221 sites with up to four mismatches, counting the perfect site); S1-g2 has 0, 4, 40 and 182 (227 sites) (Figure 6). The MIT scores are 64.0 and 70.3, and they equal the scores of the scan output.

S1-g1 has 34 sites with one to three mismatches: 25 are intergenic, 2 in gene bodies and 7 in coding exons. S1-g2 has 44 such sites: 43 are intergenic, 1 in a gene body and none in coding exons (Table S13, Table S14).

The sites with two mismatches (9 and 4) are all intergenic.

28 candidate guides pass the filters near the insertion point (Table S11, Figure 7); 9 of them have exactly one perfect site, at the intended position. 13 pairs of such guides have one cut on each side of the insertion point; the chosen pair ranks first, with a lowest MIT score of 64.0 (Table S12).

The window for the arms is 1,101 bp (131,430,500 to 131,431,600). The two arms are 500 bp each, with 49.0% and 45.4% GC, and hold 0 repeat-masked bases. The repair templates are 2,623 bp (Q43413), 2,626 bp (Q9M2U5) and 2,617 bp (O49146) (Table 4, Table S16). The templates carry no intact target site of either guide.

Table 6. The two guides at S1. Both guides lie on the minus strand. The cut is 3 bp from the PAM. Sites are genome-wide counts from a scan with a strict NGG PAM on both strands; 0 mismatches is the intended site. MIT is the Hsu et al. (2013) aggregate score. Data: `data/tables/T06.csv`*; script:* `code/04_site_guides.py`.

Guide	Protospacer and PAM	Strand	Protospacer position	Cut after	GC (%)	Sites at 0 / 1 / 2 / 3 / 4 mismatches	MIT score
S1-g1	GTTGGAGCGAGTACTATGCA AGG	-	NC_061112.1:131430997-131431016	131,430,999	50	1 / 0 / 9 / 25 / 186	64.0
S1-g2	ACCTAATGGGTTGCAAGGAA TGG	-	NC_061112.1:131431098-131431117	131,431,100	45	1 / 0 / 4 / 40 / 182	70.3

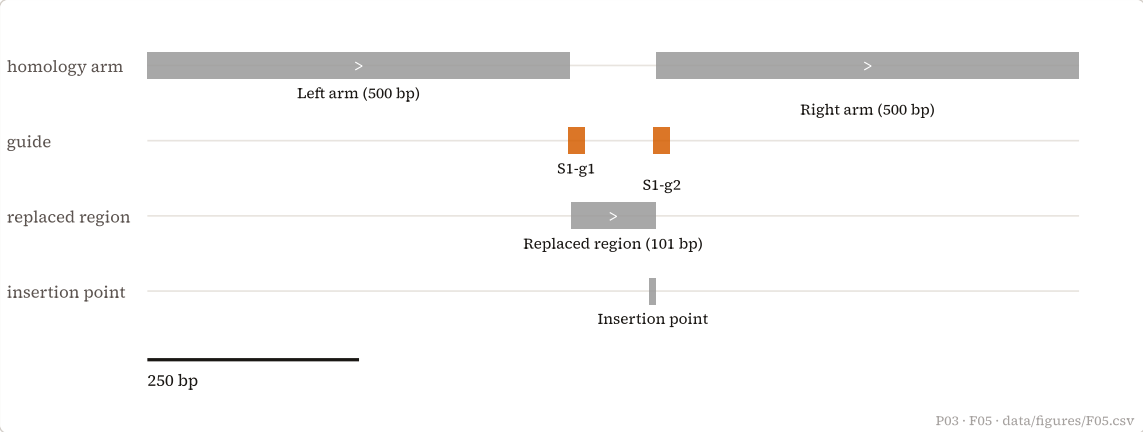


Figure 5. Site S1 and the guide pair. The 1,101-bp window on chromosome 2 (NC_061112.1:131,430,500-131,431,600): the two 500-bp homology arms, the two guides on the minus strand, the 101 bp between the cuts that the cassette replaces and the insertion point. Positions computed from the window sequence and the guide sequences. Data: `data/figures/F05.csv`; script: `code/04_site_guides.py`.

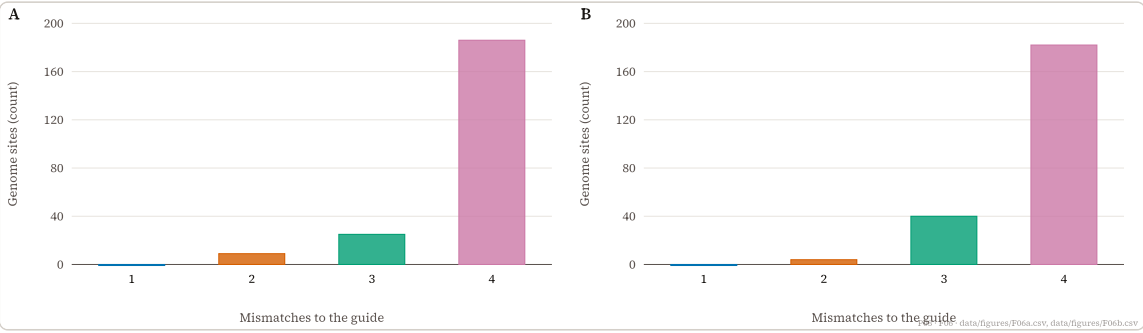


Figure 6. Genome sites by mismatch number. A, S1-g1. B, S1-g2. Sites with one to four mismatches found by the whole-genome scan (strict NGG PAM, both strands). Counts of the stored scan hits. Data: `data/figures/F06a.csv`, `data/figures/F06b.csv`; script: `code/04_site_guides.py`.

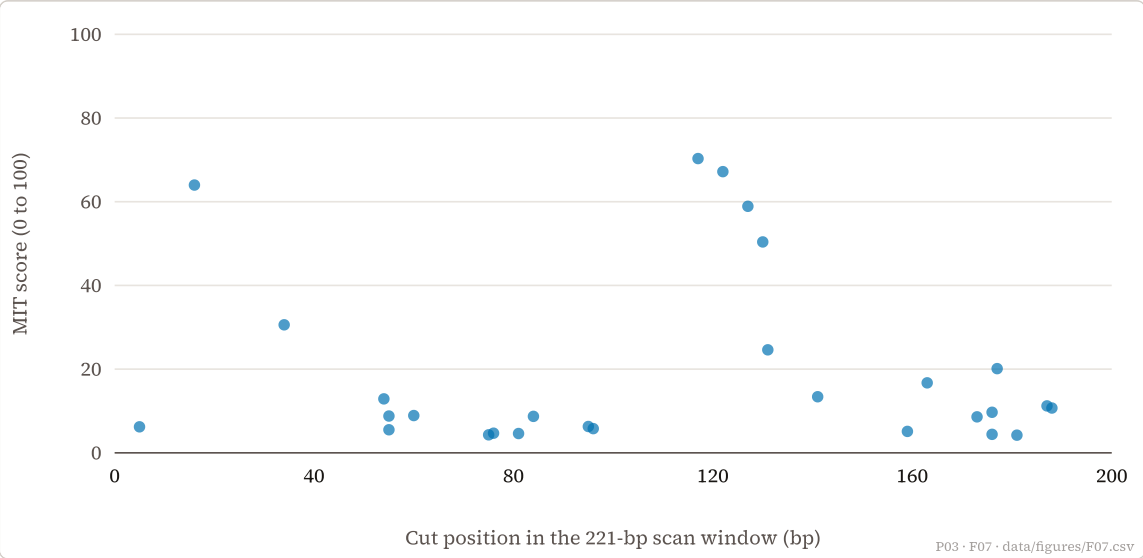


Figure 7. Specificity of the 28 candidate guides. Each point is one candidate guide at S1: its cut position within the scan window and its MIT score. Scores from the whole-genome scan. Data: `data/figures/F07.csv`; script: `code/04_site_guides.py`.

3.8 Agreement of independent computations

The C scanner and the numpy scanner list the same site sets for both guides: 221 and 227 sites. The recomputed MIT scores equal the scan output. All 3 cassettes equal the sequences written by the research scripts, and the bc calculator gives the same lengths (1,623, 1,626 and 1,617). In the second alignment, mature brazzein in each of the 3 cassettes matches P56552 in 54 of 54 residues.

4. Limitations

All results are computed. The E8 promoter is a tomato element, and the paper makes no estimate of its activity in pepper fruit or of the brazzein level. The signal cleavage sites are those of the source records; no secretion prediction was run. Site S1 and the guides refer to the UCD10Xv1.1 reference, and the sequence of the site in Phoenix is not part of the computation. Guide specificity is the MIT score; on-target activity scores are not computed. The neighbour expression has one library per stage, from *C. chinense* reads mapped to *C. annuum*. The homology arms were not searched for further genome matches.

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 cassette and repair-template sequences are FASTA files in `data/results/`. 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.24511316; Figshare: figshare/35015803; OSF: osf.io/cp6qk; Hugging Face: capsaicinoid-data; GitHub: paper-03; Website: sweet-heat-brazzein.

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 sensory panel screening, training and taster-status genotyping, brazzein sweetness in the pepper matrix and its interaction with heat, sweet-protein stability in the fruit matrix (pH, heat and storage), E8 promoter activity in pepper fruit (reporter assay, transient then stable), and a chromosome 2 site read in the lab lines with guide-pair cutting at that site.

6. References

1. Caldwell, J. E., Abildgaard, F., Džakula, Z., Ming, D., Hellekant, G., & Markley, J. L. (1998). Solution structure of the thermostable sweet-tasting protein brazzein. *Nature Structural Biology, 5*(6), 427–431. doi:10.1038/nsb0698-427

2. Deikman, J., & Fischer, R. L. (1988). Interaction of a DNA binding factor with the 5′-flanking region of an ethylene-responsive fruit ripening gene from tomato. *The EMBO Journal, 7*(11), 3315–3320. doi:10.1002/j.1460-2075.1988.tb03202.x

3. Ford, P.I. (2026). In Silico Design of Ten Capsicum annuum Cultivars Spanning 0.87 to 2.96 Million Scoville Heat Units. *Scoville Splice Research Series*. doi:10.5281/zenodo.24511300

4. 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

5. Ford, P.I. (2026). Kokumi Engineering in Capsicum: In Silico Design of γ-Glutamyl-Valyl-Glycine Accumulation in Fruit. *Scoville Splice Research Series*. doi:10.5281/zenodo.24511332

6. Ford, P.I. (2026). Cool-Burn: In Silico Design of Combined TRPV1 and TRPM8 Activation in Capsicum annuum. *Scoville Splice Research Series*. doi:10.5281/zenodo.24511340

7. 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

8. Hulse-Kemp, A. M., Maheshwari, S., Stoffel, K., et al. (2018). Reference quality assembly of the 3.5-Gb genome of *Capsicum annuum* from a single linked-read library. *Horticulture Research, 5*, 4. doi:10.1038/s41438-017-0011-0

9. Li, X., Staszewski, L., Xu, H., Durick, K., Zoller, M., & Adler, E. (2002). Human receptors for sweet and umami taste. *Proceedings of the National Academy of Sciences, 99*(7), 4692–4696. doi:10.1073/pnas.072090199

10. Meyer, B., Houlne, G., Pozueta-Romero, J., Schantz, M. L., & Schantz, R. (1996). Fruit-specific expression of a defensin-type gene family in bell pepper. Upregulation during ripening and upon wounding. *Plant Physiology, 112*(2), 615–622. doi:10.1104/pp.112.2.615

11. Ming, D., & Hellekant, G. (1994). Brazzein, a new high-potency thermostable sweet protein from *Pentadiplandra brazzeana* B. *FEBS Letters, 355*(1), 106–108. doi:10.1016/0014-5793(94)01184-2

12. Sharp, P. M., & Li, W.-H. (1987). The codon adaptation index: a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research, 15*(3), 1281–1295. doi:10.1093/nar/15.3.1281

13. von Heijne, G. (1986). A new method for predicting signal sequence cleavage sites. *Nucleic Acids Research, 14*(11), 4683–4690. doi:10.1093/nar/14.11.4683

14. Walters, D. E., & Hellekant, G. (2006). Interactions of the sweet protein brazzein with the sweet taste receptor. *Journal of Agricultural and Food Chemistry, 54*(26), 10129–10133. doi:10.1021/jf062359y

Supplementary data

The supplementary tables hold the full data behind the paper. Table S1 gives the pepper codon table with the relative adaptiveness of every codon and Table S2 the 54 balanced codons of the mature chain. Table S3 lists every setting of the back-translation and Table S4 its draws. Table S5 gives the blocks of the three cassettes, Table S6 their checks and Table S7 the cloning sites and motifs. Table S8 compares the E8 records. Table S9 and Table S10 give the site search and its settings. Table S11 lists the 28 candidate guides at S1 and Table S12 the 13 guide pairs. Table S13 lists every genome site with one to three mismatches to the chosen guides and Table S14 counts them by context. Table S15 gives the MIT position weights and Table S16 the repair templates with their junctions. 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. Pepper codon usage. Codon counts of the 48,931 clean coding sequences of Capsicum annuum UCD10Xv1.1, codons per 1,000 and relative adaptiveness (weight) for the all-isoform counts and the one-per-gene counts. Data: `data/tables/S01.csv`; script: `code/02_codon_design.py`.

Codon	Amino acid	Count	Per 1,000 codons	Weight (all isoforms)	Weight (one per gene)	Most used
GCA	A	472,744	22.31	0.805	0.805	
GCC	A	191,973	9.06	0.327	0.339	
GCG	A	107,887	5.09	0.184	0.196	
GCT	A	586,966	27.70	1.000	1.000	yes
TGC	C	159,264	7.52	0.641	0.619	
TGT	C	248,540	11.73	1.000	1.000	yes
GAC	D	304,673	14.38	0.366	0.369	
GAT	D	831,642	39.25	1.000	1.000	yes
GAA	E	793,707	37.46	1.000	1.000	yes
GAG	E	593,043	27.99	0.747	0.740	
TTC	F	340,134	16.05	0.618	0.634	
TTT	F	550,394	25.97	1.000	1.000	yes
GGA	G	464,233	21.91	1.000	1.000	yes
GGC	G	197,680	9.33	0.426	0.426	
GGG	G	233,089	11.00	0.502	0.497	
GGT	G	440,061	20.77	0.948	0.955	
CAC	H	160,156	7.56	0.460	0.469	
CAT	H	348,318	16.44	1.000	1.000	yes
ATA	I	327,848	15.47	0.568	0.561	
ATC	I	265,707	12.54	0.460	0.465	
ATT	I	577,574	27.26	1.000	1.000	yes
AAA	K	686,219	32.38	1.000	1.000	yes
AAG	K	664,516	31.36	0.968	0.959	
CTA	L	225,098	10.62	0.430	0.429	
CTC	L	224,288	10.58	0.428	0.433	
CTG	L	252,008	11.89	0.481	0.446	
CTT	L	523,918	24.73	1.000	0.978	yes
TTA	L	318,504	15.03	0.608	0.612	
TTG	L	523,154	24.69	0.999	1.000	
ATG	M	523,728	24.72	1.000	1.000	yes
AAC	N	329,785	15.56	0.504	0.515	
AAT	N	654,770	30.90	1.000	1.000	yes

Codon	Amino acid	Count	Per 1,000 codons	Weight (all isoforms)	Weight (one per gene)	Most used
CCA	P	379,990	17.93	0.961	0.993	
CCC	P	127,591	6.02	0.323	0.331	
CCG	P	97,162	4.59	0.246	0.263	
CCT	P	395,423	18.66	1.000	1.000	yes
CAA	Q	453,106	21.38	1.000	1.000	yes
CAG	Q	332,725	15.70	0.734	0.677	
AGA	R	377,025	17.79	1.000	1.000	yes
AGG	R	279,674	13.20	0.742	0.740	
CGA	R	125,087	5.90	0.332	0.325	
CGC	R	76,530	3.61	0.203	0.196	
CGG	R	86,411	4.08	0.229	0.216	
CGT	R	153,701	7.25	0.408	0.399	
AGC	S	218,183	10.30	0.434	0.442	
AGT	S	363,350	17.15	0.723	0.740	
TCA	S	467,466	22.06	0.931	0.949	
TCC	S	222,116	10.48	0.442	0.453	
TCG	S	128,860	6.08	0.257	0.276	
TCT	S	502,378	23.71	1.000	1.000	yes
ACA	T	363,160	17.14	0.922	0.918	
ACC	T	168,920	7.97	0.429	0.436	
ACG	T	99,286	4.69	0.252	0.259	
ACT	T	393,909	18.59	1.000	1.000	yes
GTA	V	258,205	12.19	0.449	0.452	
GTC	V	206,697	9.75	0.359	0.359	
GTG	V	339,372	16.02	0.590	0.605	
GTT	V	575,377	27.15	1.000	1.000	yes
TGG	W	263,999	12.46	1.000	1.000	yes
TAC	Y	218,203	10.30	0.554	0.570	
TAT	Y	394,196	18.60	1.000	1.000	yes

Table S2. Codons of the mature brazzein chain. The 54 balanced codons shared by the three cassettes, with the most-used pepper codon for comparison. Data: `data/tables/S02.csv`*; script:* `code/02_codon_design.py`*.*

Position	Residue	Balanced codon	Weight of the balanced codon	Most-used codon	Same codon
1	Q	CAG	0.734	CAA	no
2	D	GAT	1.000	GAT	yes
3	K	AAG	0.968	AAA	no
4	C	TGT	1.000	TGT	yes
5	K	AAG	0.968	AAA	no
6	K	AAA	1.000	AAA	yes
7	V	GTG	0.590	GTT	no
8	Y	TAC	0.554	TAT	no

Position	Residue	Balanced codon	Weight of the balanced codon	Most-used codon	Same codon
9	E	GAA	1.000	GAA	yes
10	N	AAC	0.504	AAT	no
11	Y	TAT	1.000	TAT	yes
12	P	CCA	0.961	CCT	no
13	V	GTT	1.000	GTT	yes
14	S	TCT	1.000	TCT	yes
15	K	AAG	0.968	AAA	no
16	C	TGT	1.000	TGT	yes
17	Q	CAA	1.000	CAA	yes
18	L	TTG	0.999	CTT	no
19	A	GCT	1.000	GCT	yes
20	N	AAC	0.504	AAT	no
21	Q	CAA	1.000	CAA	yes
22	C	TGT	1.000	TGT	yes
23	N	AAC	0.504	AAT	no
24	Y	TAT	1.000	TAT	yes
25	D	GAT	1.000	GAT	yes
26	C	TGC	0.641	TGT	no
27	K	AAG	0.968	AAA	no
28	L	TTG	0.999	CTT	no
29	D	GAT	1.000	GAT	yes
30	K	AAG	0.968	AAA	no
31	H	CAT	1.000	CAT	yes
32	A	GCT	1.000	GCT	yes
33	R	AGG	0.742	AGA	no
34	S	TCT	1.000	TCT	yes
35	G	GGA	1.000	GGA	yes
36	E	GAG	0.747	GAA	no
37	C	TGC	0.641	TGT	no
38	F	TTC	0.618	TTT	no

Position	Residue	Balanced codon	Weight of the balanced codon	Most-used codon	Same codon
39	Y	TAC	0.554	TAT	no
40	D	GAT	1.000	GAT	yes
41	E	GAA	1.000	GAA	yes
42	K	AAG	0.968	AAA	no
43	R	AGG	0.742	AGA	no
44	N	AAT	1.000	AAT	yes
45	L	TTG	0.999	CTT	no
46	Q	CAG	0.734	CAA	no
47	C	TGT	1.000	TGT	yes
48	I	ATT	1.000	ATT	yes
49	C	TGC	0.641	TGT	no
50	D	GAT	1.000	GAT	yes
51	Y	TAT	1.000	TAT	yes
52	C	TGC	0.641	TGT	no
53	E	GAG	0.747	GAA	no
54	Y	TAT	1.000	TAT	yes

Table S3. Codon design parameters. Every setting of the balanced back-translation and the pepper reference numbers it uses. Data: `data/tables/S03.csv`*; script:* `code/02_codon_design.py`*.*

Parameter	Value	Meaning
usable_weight_threshold	0.5	A codon is usable when its weight is at least this value
draws	3000	Random draws per reading frame
repair_rounds	300	Redraws of a codon that overlaps a listed motif, per draw
gc_window_Q43413	3	Allowed distance of the reading-frame GC from the pepper average (points), Q43413
gc_window_other_signals	6	Same window for Q9M2U5 and O49146
pepper_cds_gc_percent	42.3	Average GC of the pepper coding sequences, from the codon counts
pepper_cds_gc3_percent	37.5	Average GC at the third codon position
clean_coding_sequences	48931	Coding sequences used for the codon table and the stop-codon counts

Parameter	Value	Meaning
stop_codon	TGA	Most used terminal stop codon in pepper
stop_TGA_percent	41.5	Share of TGA among terminal stop codons
stop_TAA_percent	33.7	Share of TAA
stop_TAG_percent	24.8	Share of TAG
codon_total	21189723	Codons counted in the codon table
most_used_codons_agree_amino_acids	19	Amino acids (of 20) with the same most-used codon in the all-isoform and the one-per-gene count
codons_changed_most_used_to_balanced_mature	27	Mature codons that differ between the most-used and the balanced design (of 54)
motifs_checked	15	Listed motifs checked during the back-translation, plus runs of five identical bases
motifs_reported	17	Motifs scanned and reported for every coding design

Table S4. Back-translation draws. Seeded draws for each reading frame, the number kept after the filters and the rejections by reason. Data: `data/tables/S04.csv`*; script:* `code/02_codon_design.py`*.*

Design	Seed	Draws	GC window (points)	Draws kept	Rejected: GC outside window	Rejected: motif left
Q43413	24112026	3000	3	30	2970	0
Q9M2U5	2439	3000	6	999	2001	0
O49146	2436	3000	6	2990	10	0

Table S5. Cassette blocks. Coordinates of every block in the three cassettes. Data: `data/tables/S05.csv`*; script:* `code/03_cassettes.py`*.*

Variant	Block	Kind	Start	End	Length (bp)	Source
Q43413	E8 promoter and 5' leader (X13437 1-1121)	promoter	1	1121	1121	tomato E8 gene record X13437
Q43413	5' leader (transcript start to ATG)	utr	1087	1121	35	X13437 1087-1121
Q43413	Signal Q43413 (27 aa)	signal	1122	1202	81	UniProt Q43413 residues 1-27
Q43413	Mature brazzein (54 aa)	cds	1203	1364	162	UniProt P56552 residues 1-54
Q43413	Stop TGA	stop codon	1365	1367	3	terminal stop codon
Q43413	NOS terminator (256 bp)	terminator	1368	1623	256	NOS terminator part
Q9M2U5	E8 promoter and 5' leader (X13437 1-1121)	promoter	1	1121	1121	tomato E8 gene record X13437
Q9M2U5	5' leader (transcript start to ATG)	utr	1087	1121	35	X13437 1087-1121
Q9M2U5	Signal Q9M2U5 (28 aa)	signal	1122	1205	84	UniProt Q9M2U5 residues 1-28
Q9M2U5	Mature brazzein (54 aa)	cds	1206	1367	162	UniProt P56552 residues 1-54
Q9M2U5	Stop TGA	stop codon	1368	1370	3	terminal stop codon
Q9M2U5	NOS terminator (256 bp)	terminator	1371	1626	256	NOS terminator part
O49146	E8 promoter and 5' leader (X13437 1-1121)	promoter	1	1121	1121	tomato E8 gene record X13437
O49146	5' leader (transcript start to ATG)	utr	1087	1121	35	X13437 1087-1121
O49146	Signal O49146 (25 aa)	signal	1122	1196	75	UniProt O49146 residues 1-25
O49146	Mature brazzein (54 aa)	cds	1197	1358	162	UniProt P56552 residues 1-54
O49146	Stop TGA	stop codon	1359	1361	3	terminal stop codon
O49146	NOS terminator (256 bp)	terminator	1362	1617	256	NOS terminator part

Table S6. Cassette checks. Translation, reading-frame, start-codon and length checks of the three cassettes. Data: `data/tables/S06.csv`*; script:* `code/03_cassettes.py`*.*

Variant	Check	Result
Q43413	translation of the reading frame equals signal + brazzein	True
Q43413	amino acids translated (without stop)	81
Q43413	stop codons in the frame	1
Q43413	stop codon is the last codon	True
Q43413	promoter + reading frame + NOS = cassette length	True
Q43413	reading frame length = 3 x (signal + 54) + 3	True
Q43413	leader length (nt)	35
Q43413	ATGs in the leader	0
Q43413	first ATG after the transcript start (position)	1122
Q43413	first ATG equals the first base of the reading frame	True
Q43413	longest ATG-to-stop frame on either strand (aa)	81
Q43413	longest frame is the designed frame	True
Q43413	second longest frame (aa)	68
Q9M2U5	translation of the reading frame equals signal + brazzein	True
Q9M2U5	amino acids translated (without stop)	82
Q9M2U5	stop codons in the frame	1
Q9M2U5	stop codon is the last codon	True
Q9M2U5	promoter + reading frame + NOS = cassette length	True
Q9M2U5	reading frame length = 3 x (signal + 54) + 3	True
Q9M2U5	leader length (nt)	35
Q9M2U5	ATGs in the leader	0
Q9M2U5	first ATG after the transcript start (position)	1122
Q9M2U5	first ATG equals the first base of the reading frame	True
Q9M2U5	longest ATG-to-stop frame on either strand (aa)	82
Q9M2U5	longest frame is the designed frame	True
Q9M2U5	second longest frame (aa)	42
O49146	translation of the reading frame equals signal + brazzein	True
O49146	amino acids translated (without stop)	79

Variant	Check	Result
O49146	stop codons in the frame	1
O49146	stop codon is the last codon	True
O49146	promoter + reading frame + NOS = cassette length	True
O49146	reading frame length = 3 x (signal + 54) + 3	True
O49146	leader length (nt)	35
O49146	ATGs in the leader	0
O49146	first ATG after the transcript start (position)	1122
O49146	first ATG equals the first base of the reading frame	True
O49146	longest ATG-to-stop frame on either strand (aa)	79
O49146	longest frame is the designed frame	True
O49146	second longest frame (aa)	69
all	promoter part identical in the three cassettes	True
all	mature codons, stop and NOS identical in the three cassettes	True
all	length of that shared tail (bp)	421

Table S7. Cloning sites and motifs in the cassettes. Positions (1-based) of nine common cloning sites and of two sequence motifs in each cassette. Data: `data/tables/S07.csv`*; script:* `code/03_cassettes.py`*.*

Variant	Site or motif	Count	Positions
Q43413	EcoRI GAATTC	0	
Q43413	BamHI GGATCC	0	
Q43413	HindIII AAGCTT	0	
Q43413	XbaI TCTAGA	0	
Q43413	SacI GAGCTC	0	
Q43413	BsaI GGTCTC / GAGACC	0	
Q43413	BsmBI CGTCTC / GAGACG	0	
Q43413	SmaI / XmaI CCCGGG	0	
Q43413	KpnI GGTACC	0	
Q43413	ATTTA (AU-rich instability) in the cassette	10	181 292 312 331 335 446 521 721 1496 1539
Q43413	AATAAA (polyadenylation signal) in the cassette	7	86 408 435 462 530 756 1387
Q9M2U5	EcoRI GAATTC	0	

Variant	Site or motif	Count	Positions
Q9M2U5	BamHI GGATCC	0	
Q9M2U5	HindIII AAGCTT	0	
Q9M2U5	XbaI TCTAGA	0	
Q9M2U5	SacI GAGCTC	0	
Q9M2U5	BsaI GGTCTC / GAGACC	0	
Q9M2U5	BsmBI CGTCTC / GAGACG	0	
Q9M2U5	SmaI / XmaI CCCGGG	0	
Q9M2U5	KpnI GGTACC	0	
Q9M2U5	ATTTA (AU-rich instability) in the cassette	10	181 292 312 331 335 446 521 721 1499 1542
Q9M2U5	AATAAA (polyadenylation signal) in the cassette	7	86 408 435 462 530 756 1390
O49146	EcoRI GAATTC	0	
O49146	BamHI GGATCC	0	
O49146	HindIII AAGCTT	0	
O49146	XbaI TCTAGA	0	
O49146	SacI GAGCTC	0	
O49146	BsaI GGTCTC / GAGACC	0	
O49146	BsmBI CGTCTC / GAGACG	0	
O49146	SmaI / XmaI CCCGGG	0	
O49146	KpnI GGTACC	0	
O49146	ATTTA (AU-rich instability) in the cassette	10	181 292 312 331 335 446 521 721 1490 1533
O49146	AATAAA (polyadenylation signal) in the cassette	7	86 408 435 462 530 756 1381

Table S8. E8 promoter records. The tomato E8 record used for the promoter piece and two further tomato records compared with it. Data: `data/tables/S08.csv`*; script:* `code/03_cassettes.py`*.*

Item	Value
X13437 length (bp)	3042
X13437 organism and cultivar	Solanum lycopersicum, cv. Cherry
X13437 TATA box (positions)	1055-1063
X13437 transcription start (position)	1087
X13437 E8 start codon (position)	1122

Item	Value
Promoter piece used (X13437 bases)	1-1121
Promoter piece length (bp)	1121
5' leader length, transcript start to the base before the E8 ATG (nt)	35
DQ317599 length (bp)	2175
DQ317599 cultivar	Jinfeng 1
X13437 bases 1-1086 found unchanged in DQ317599 at position	1090
X13437 bases 1-1086 identical to DQ317599 (bases)	1086 of 1086
AF515784 length (bp)	2191
AF515784 cultivar	ZHONGSHU NO.5
X13437 bases 1-1086 aligned to AF515784, identical columns	1081 of 1086
same, percent identical	99.5

Table S9. Site search. Genes of the assembly and the gaps that pass each filter of the site search. Data: `data/tables/S09.csv`*; script:* `code/04_site_guides.py`*.*

Step	Count
Genes and pseudogenes on the 12 chromosomes	40,503
of which protein-coding	27,958
Gaps between neighbouring genes	38,660
Gaps of 13 to 60 kb	10,543
Gaps of 13 to 60 kb with protein-coding genes on both sides	5,328
Gaps holding a 3,000-bp window with 10% or less repeat-masked bases	1,129

Table S10. Site search and guide settings. Every setting of the site search, the guide filters and the pair rule. Data: `data/tables/S10.csv`*; script:* `code/04_site_guides.py`*.*

Parameter	Value	Meaning
chromosomes	12 placed chromosomes (NC_061111.1 to NC_061122.1)	Placed chromosomes of the assembly
distance_from_genes_bp	5000	Window keeps at least this distance from every annotated gene or pseudogene on both sides
window_bp	3000	Window length
step_bp	500	Step between the tested windows

Parameter	Value	Meaning
repeat_masked_fraction_max	0.1	Largest share of soft-masked bases in the window
maximum_gap_between_neighbours_bp	60000	Largest gap between the neighbouring genes
minimum_gap_between_neighbours_bp	13000	Gap that holds a window and the two distances (2 x 5,000 + 3,000)
neighbours	both protein-coding	Biotype required for the two flanking genes
windows_with_N	0	No N allowed in the window
guide_gc_range_percent	40 to 70	Guide filter
guide_filters	no TTTT, no run of five identical bases	Guide filter
guide_scan_window_bp	220	Guides listed within 110 bp either side of the insertion point
pair_rule	both guides unique, one cut each side of the insertion point, cuts 20 to 120 bp apart, highest lowest-MIT then smaller distance	Rule used to choose the pair
arm_length_bp	500	Homology arm length each side
mismatches_scanned	4	Mismatches allowed in the genome scan
pam	strict NGG	PAM read from the genome, both strands

Table S11. All 28 candidate guides at S1. Every SpCas9 guide with a strict NGG PAM within 110 bp either side of the insertion point that passes the GC and repeat filters, with its genome-wide site counts and MIT score. Data: `data/tables/S11.csv`*; script:* `code/04_site_guides.py`.

Guide	PAM	Strand	Start	End	Cut after	Cut relative to the insertion point (bp)	GC (%)	Perfect sites	1 mm	2 mm	3 mm	4 mm	MIT score
TACTATGCAAGGCAAGAACC	AGG	-	131430986	131431005	131430988	-105	45.0	1	13	40	27	95	6.2
GTTGGAGCGAGTACTATGCA	AGG	-	131430997	131431016	131430999	-94	50.0	1	0	9	25	186	64.0
GGGCAAACAGAGCTTACAGT	TGG	-	131431015	131431034	131431017	-76	50.0	1	3	8	10	74	30.6
GGCAGTGAAGCCTACGGCAT	GGG	-	131431035	131431054	131431037	-56	60.0	30	14	15	26	67	12.9
GCTCTGTTTGCCCATGCCGT	AGG	+	131431022	131431041	131431038	-55	60.0	2	25	15	17	30	5.5
GGGCAGTGAAGCCTACGGCA	TGG	-	131431036	131431055	131431038	-55	65.0	27	18	14	16	13	8.8
TGAGTGGGCAGTGAAGCCTA	CGG	-	131431041	131431060	131431043	-50	55.0	28	18	3	8	152	8.9
GAGAGCCATGCAGGATGAGT	GGG	-	131431056	131431075	131431058	-35	55.0	30	29	11	35	303	4.3
AGAGAGCCATGCAGGATGAG	TGG	-	131431057	131431076	131431059	-34	55.0	27	32	12	6	37	4.7
CACTGCCCACTCATCCTGCA	TGG	+	131431048	131431067	131431064	-29	60.0	29	29	9	7	42	4.6
CCTACAGAAGAGAGCCATGC	AGG	-	131431065	131431084	131431067	-26	55.0	28	20	23	9	17	8.7
CCTGCATGGCTCTCTTCTGT	AGG	+	131431062	131431081	131431078	-15	55.0	28	19	20	10	30	6.3

Guide	PAM	Strand	Start	End	Cut after	Cut relative to the insertion point (bp)	GC (%)	Perfect sites	1 mm	2 mm	3 mm	4 mm	MIT score
CTGCATGGCTCTCTTCTGTA	GGG	+	131431063	131431082	131431079	-14	50.0	29	21	18	13	68	5.8
ACCTAATGGGTTGCAAGGAA	TGG	-	131431098	131431117	131431100	7	45.0	1	0	4	40	182	70.3
TATCCACCTAATGGGTTGCA	AGG	-	131431103	131431122	131431105	12	45.0	1	0	1	11	100	67.2
ACCATTCCTTGCAACCCATT	AGG	+	131431094	131431113	131431110	17	45.0	1	0	2	16	141	58.9
ATTCCTTGCAACCCATTAGG	TGG	+	131431097	131431116	131431113	20	45.0	1	0	4	36	230	50.4
CCCTAAATCTATCCACCTAA	TGG	-	131431112	131431131	131431114	21	40.0	1	3	13	28	384	24.6
CCCATTAGGTGGATAGATTT	AGG	+	131431108	131431127	131431124	31	40.0	1	3	17	128	1214	13.4
AATGGGTTGTGAAGAGTGGC	AGG	-	131431140	131431159	131431142	49	50.0	2	20	59	36	61	5.1
ACCTAATGGGTTGTGAAGAG	TGG	-	131431144	131431163	131431146	53	45.0	2	19	37	41	177	16.7
GCCACTCTTCACAACCCATT	AGG	+	131431140	131431159	131431156	63	50.0	2	17	33	19	84	8.6
ACTCTTCACAACCCATTAGG	TGG	+	131431143	131431162	131431159	66	45.0	2	22	33	24	122	4.4
CCTAAACCTAGCCACCTAAT	GGG	-	131431157	131431176	131431159	66	45.0	11	14	36	40	118	9.7
CCCTAAACCTAGCCACCTAA	TGG	-	131431158	131431177	131431160	67	50.0	12	10	32	25	91	20.1
TCACAACCCATTAGGTGGCT	AGG	+	131431148	131431167	131431164	71	50.0	3	23	38	54	63	4.2
CCCATTAGGTGGCTAGGTTT	AGG	+	131431154	131431173	131431170	77	50.0	11	11	29	15	102	11.2
CCATTAGGTGGCTAGGTTTA	GGG	+	131431155	131431174	131431171	78	45.0	12	10	32	25	108	10.7

Table S12. Guide pairs around the insertion point. Pairs of unique guides with one cut on each side of the insertion point and cuts 20 to 120 bp apart, ranked by the lower MIT score and then by the smaller distance. Data: `data/tables/S12.csv`*; script:* `code/04_site_guides.py`*.*

Rank	Left guide	Right guide	Bp between cuts	MIT left	MIT right	Lowest MIT
1	GTTGGAGCGAGTACTATGCA	ACCTAATGGGTTGCAAGGAA	101	64.0	70.3	64.0
2	GTTGGAGCGAGTACTATGCA	TATCCACCTAATGGGTTGCA	106	64.0	67.2	64.0
3	GTTGGAGCGAGTACTATGCA	ACCATTCCTTGCAACCCATT	111	64.0	58.9	58.9
4	GTTGGAGCGAGTACTATGCA	ATTCCTTGCAACCCATTAGG	114	64.0	50.4	50.4
5	GGGCAAACAGAGCTTACAGT	ACCTAATGGGTTGCAAGGAA	83	30.6	70.3	30.6
6	GGGCAAACAGAGCTTACAGT	TATCCACCTAATGGGTTGCA	88	30.6	67.2	30.6
7	GGGCAAACAGAGCTTACAGT	ACCATTCCTTGCAACCCATT	93	30.6	58.9	30.6
8	GGGCAAACAGAGCTTACAGT	ATTCCTTGCAACCCATTAGG	96	30.6	50.4	30.6
9	GGGCAAACAGAGCTTACAGT	CCCTAAATCTATCCACCTAA	97	30.6	24.6	24.6
10	GTTGGAGCGAGTACTATGCA	CCCTAAATCTATCCACCTAA	115	64.0	24.6	24.6
11	GGGCAAACAGAGCTTACAGT	CCCATTAGGTGGATAGATTT	107	30.6	13.4	13.4
12	TACTATGCAAGGCAAGAACC	ACCTAATGGGTTGCAAGGAA	112	6.2	70.3	6.2
13	TACTATGCAAGGCAAGAACC	TATCCACCTAATGGGTTGCA	117	6.2	67.2	6.2

Table S13. Genome sites with 1 to 3 mismatches. Every genome site within three mismatches of the two guides, with its position (first base of the 23-mer on the plus strand) and whether it lies in a coding exon, a gene body or between genes. Data: `data/tables/S13.csv`*; script:* `code/04_site_guides.py`*.*

Guide	Chromosome	Position	Strand	Mismatches	Context
S1-g1	NC_061115.1	173194420	+	2	intergenic
S1-g1	NC_061116.1	134828286	-	2	intergenic
S1-g1	NC_061116.1	218935709	-	2	intergenic

Guide	Chromosome	Position	Strand	Mismatches	Context
S1-g1	NC_061118.1	165036267	-	2	intergenic
S1-g1	NC_061119.1	169465744	-	2	intergenic
S1-g1	NC_061119.1	173967871	+	2	intergenic
S1-g1	NC_061122.1	4468398	-	2	intergenic
S1-g1	NW_025819819.1	3284	-	2	intergenic
S1-g1	NW_025831415.1	1146	+	2	intergenic
S1-g1	NC_061111.1	44426183	-	3	intergenic
S1-g1	NC_061111.1	72655998	+	3	intergenic
S1-g1	NC_061112.1	104479437	+	3	CDS exon
S1-g1	NC_061113.1	26646009	-	3	CDS exon
S1-g1	NC_061113.1	136084254	+	3	intergenic
S1-g1	NC_061113.1	208972083	+	3	gene body
S1-g1	NC_061113.1	245245345	-	3	intergenic
S1-g1	NC_061114.1	113230540	-	3	intergenic
S1-g1	NC_061114.1	206789517	-	3	CDS exon
S1-g1	NC_061116.1	39652438	+	3	intergenic
S1-g1	NC_061116.1	170794372	+	3	gene body
S1-g1	NC_061116.1	209390479	-	3	CDS exon
S1-g1	NC_061117.1	223230111	+	3	intergenic
S1-g1	NC_061119.1	64841300	+	3	intergenic
S1-g1	NC_061120.1	120001713	+	3	intergenic
S1-g1	NC_061121.1	10261942	-	3	CDS exon
S1-g1	NC_061121.1	14310259	-	3	CDS exon
S1-g1	NC_061121.1	47421687	-	3	CDS exon
S1-g1	NC_061122.1	89404435	+	3	intergenic
S1-g1	NC_061122.1	113427446	-	3	intergenic
S1-g1	NC_061122.1	211131588	+	3	intergenic
S1-g1	NW_025833286.1	310	-	3	intergenic
S1-g1	NW_025833759.1	623	+	3	intergenic
S1-g1	NW_025852459.1	1342	+	3	intergenic

Guide	Chromosome	Position	Strand	Mismatches	Context
S1-g1	NW_025868305.1	1170	-	3	intergenic
S1-g2	NC_061112.1	138498822	+	2	intergenic
S1-g2	NC_061113.1	245245538	-	2	intergenic
S1-g2	NC_061119.1	169465937	-	2	intergenic
S1-g2	NW_025887908.1	1720	+	2	intergenic
S1-g2	NC_061111.1	38737577	+	3	intergenic
S1-g2	NC_061111.1	44426330	-	3	intergenic
S1-g2	NC_061112.1	131431187	-	3	intergenic
S1-g2	NC_061112.1	138498868	+	3	intergenic
S1-g2	NC_061113.1	136084061	+	3	intergenic
S1-g2	NC_061113.1	136084107	+	3	intergenic
S1-g2	NC_061113.1	136084153	+	3	intergenic
S1-g2	NC_061113.1	154423248	-	3	intergenic
S1-g2	NC_061113.1	245245492	-	3	intergenic
S1-g2	NC_061114.1	30046648	+	3	intergenic
S1-g2	NC_061114.1	43469694	+	3	intergenic
S1-g2	NC_061114.1	113230641	-	3	intergenic
S1-g2	NC_061115.1	42707155	-	3	intergenic
S1-g2	NC_061115.1	75694610	-	3	intergenic
S1-g2	NC_061115.1	123405592	+	3	intergenic
S1-g2	NC_061115.1	203083218	+	3	intergenic
S1-g2	NC_061116.1	39652291	+	3	intergenic
S1-g2	NC_061116.1	218935810	-	3	intergenic
S1-g2	NC_061117.1	132546369	+	3	intergenic
S1-g2	NC_061118.1	165036414	-	3	intergenic
S1-g2	NC_061118.1	165036460	-	3	intergenic
S1-g2	NC_061119.1	94398671	-	3	intergenic
S1-g2	NC_061119.1	132517958	-	3	intergenic
S1-g2	NC_061119.1	189825813	+	3	intergenic
S1-g2	NC_061120.1	69081235	-	3	gene body

Guide	Chromosome	Position	Strand	Mismatches	Context
S1-g2	NC_061120.1	74155773	-	3	intergenic
S1-g2	NC_061121.1	15983298	-	3	intergenic
S1-g2	NC_061122.1	4468545	-	3	intergenic
S1-g2	NC_061122.1	116914788	-	3	intergenic
S1-g2	NW_025819819.1	3385	-	3	intergenic
S1-g2	NW_025824845.1	1342	+	3	intergenic
S1-g2	NW_025831415.1	999	+	3	intergenic
S1-g2	NW_025833759.1	476	+	3	intergenic
S1-g2	NW_025840679.1	1003	-	3	intergenic
S1-g2	NW_025842160.1	1259	+	3	intergenic
S1-g2	NW_025851898.1	2230	+	3	intergenic
S1-g2	NW_025852100.1	1113	+	3	intergenic
S1-g2	NW_025852459.1	1195	+	3	intergenic
S1-g2	NW_025880601.1	2386	-	3	intergenic
S1-g2	NW_025883384.1	38410	+	3	intergenic

Table S14. Sites with 1 to 3 mismatches by context. Counts of the sites listed in the previous table by mismatch number and genomic context. Data: `data/tables/S14.csv`*; script:* `code/04_site_guides.py`*.*

Guide	Mismatches	Intergenic	Gene body	Coding exon	Total
S1-g1	1	0	0	0	0
S1-g1	2	9	0	0	9
S1-g1	3	16	2	7	25
S1-g1	1 to 3	25	2	7	34
S1-g2	1	0	0	0	0
S1-g2	2	4	0	0	4
S1-g2	3	39	1	0	40
S1-g2	1 to 3	43	1	0	44

Table S15. MIT position weights. Mismatch weights of the Hsu et al. (2013) score used for every MIT value. Data: `data/tables/S15.csv`*; script:* `code/04_site_guides.py`*.*

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

Table S16. Repair templates. Left arm, cassette and right arm of each repair template with the 40 bases around each junction (the vertical bar marks the junction). Data: `data/tables/S16.csv`*; script:* `code/04_site_guides.py`*.*

Variant	Left arm (bp)	Cassette (bp)	Right arm (bp)	Template (bp)	Template GC (%)	Left junction (20 + 20)	Right junction (20 + 20)
Q43413	500	1623	500	2623	35.9	TAGCCTGGTTCTTGCCTTGC|CTAGAAGGAATTTCACGAAA	ATCTATGTTACTAGATCGGG|CTTGCAACCCATTAGG
Q9M2U5	500	1626	500	2626	35.6	TAGCCTGGTTCTTGCCTTGC|CTAGAAGGAATTTCACGAAA	ATCTATGTTACTAGATCGGG|CTTGCAACCCATTAGG
O49146	500	1617	500	2617	35.8	TAGCCTGGTTCTTGCCTTGC|CTAGAAGGAATTTCACGAAA	ATCTATGTTACTAGATCGGG|CTTGCAACCCATTAGG

Supplementary data: file map

Path	Contents
`data/tables/T01.csv`	Table 1: Mature brazzein
`data/tables/T02.csv`	Table 2: Three secretion signals
`data/tables/T03.csv`	Table 3: Codon designs
`data/tables/T04.csv`	Table 4: The three cassettes
`data/tables/T05.csv`	Table 5: Blocks of the Q43413 cassette
`data/tables/T06.csv`	Table 6: The two guides at S1
`data/tables/T07.csv`	Table 7: Site S1
`data/tables/S01.csv`	Table S1: Pepper codon usage
`data/tables/S02.csv`	Table S2: Codons of the mature brazzein chain
`data/tables/S03.csv`	Table S3: Codon design parameters
`data/tables/S04.csv`	Table S4: Back-translation draws
`data/tables/S05.csv`	Table S5: Cassette blocks
`data/tables/S06.csv`	Table S6: Cassette checks
`data/tables/S07.csv`	Table S7: Cloning sites and motifs in the cassettes
`data/tables/S08.csv`	Table S8: E8 promoter records
`data/tables/S09.csv`	Table S9: Site search
`data/tables/S10.csv`	Table S10: Site search and guide settings
`data/tables/S11.csv`	Table S11: All 28 candidate guides at S1
`data/tables/S12.csv`	Table S12: Guide pairs around the insertion point
`data/tables/S13.csv`	Table S13: Genome sites with 1 to 3 mismatches
`data/tables/S14.csv`	Table S14: Sites with 1 to 3 mismatches by context
`data/tables/S15.csv`	Table S15: MIT position weights
`data/tables/S16.csv`	Table S16: Repair templates
`data/figures/GA.csv`	Data behind the graphical abstract
`data/figures/F01a.csv`	Data behind Figure 1
`data/figures/F01b.csv`	Data behind Figure 1
`data/figures/F01c.csv`	Data behind Figure 1
`data/figures/F02.csv`	Data behind Figure 2
`data/figures/F03.csv`	Data behind Figure 3
`data/figures/F04.csv`	Data behind Figure 4
`data/figures/F05.csv`	Data behind Figure 5
`data/figures/F06a.csv`	Data behind Figure 6
`data/figures/F06b.csv`	Data behind Figure 6
`data/figures/F07.csv`	Data behind Figure 7
`data/figures/F08.csv`	Data behind Figure 8
`data/results/brazzein_protein.csv`	Properties of mature brazzein read from the UniProt record, with the computed mass
`data/results/brazzein_cysteines.csv`	The eight cysteines with their disulfide partners
`data/results/brazzein_composition.csv`	Amino-acid composition of the mature chain
`data/results/signal_peptides.csv`	The three signal peptides: source, length, sequence, residues at -3 and -1, evidence code and tissue in the record
`data/results/codon_usage_weights.csv`	Pepper codon counts, per-1,000 values and relative adaptiveness for both counting methods
`data/results/mature_codons.csv`	The 54 balanced codons of mature brazzein with their weights and the most-used codon

Path	Contents
`data/results/orf_summary.csv`	Length, codon adaptation index, GC, GC3 and motif check of each coding design
`data/results/orf_motifs.csv`	Positions of 17 listed motifs in each coding design
`data/results/design_runs.csv`	Seeds, draws, draws kept and rejections of the balanced back-translation
`data/results/codon_design_parameters.csv`	Settings and pepper reference numbers of the balanced back-translation
`data/results/brazzein_orfs.fasta`	The three signal-brazzein reading frames (balanced pepper codons)
`data/results/brazzein_mature_cds.fasta`	Mature brazzein coding sequence with balanced codons and with most-used codons
`data/results/cassette_Q43413.fasta`	Cassette with the Q43413 signal (1,623 bp)
`data/results/cassette_Q9M2U5.fasta`	Cassette with the Q9M2U5 signal (1,626 bp)
`data/results/cassette_O49146.fasta`	Cassette with the O49146 signal (1,617 bp)
`data/results/cassette_blocks.csv`	Block coordinates of the three cassettes
`data/results/cassette_summary.csv`	Length, GC and codon adaptation index of the three cassettes
`data/results/cassette_checks.csv`	Translation, reading-frame, start-codon and length checks
`data/results/cassette_sites.csv`	Cloning sites and two motifs in the cassettes
`data/results/cassette_crosscheck.csv`	Comparison of the cassettes with the sequences written by the research scripts
`data/results/promoter_records.csv`	The E8 record used and the comparison with two further tomato records
`data/results/guide_pair.csv`	The two guides: sequence, PAM, position, cut, site counts and MIT score (recomputed and from the scan output)
`data/results/site_window_checks.csv`	Checks of the window, the arms, the cuts and the insertion point
`data/results/scanner_agreement.csv`	Site counts of the two independent scanners and whether the site sets are identical
`data/results/offtarget_by_context.csv`	Sites with 1 to 3 mismatches by mismatch number and genomic context
`data/results/offtarget_sites.csv`	Every genome site with 1 to 3 mismatches to the two guides
`data/results/repair_templates.csv`	Repair templates: arm and cassette lengths, GC, junction sequences, intact guide sites
`data/results/repair_template_Q43413.fasta`	Repair template with the Q43413 cassette (2,623 bp)
`data/results/repair_template_Q9M2U5.fasta`	Repair template with the Q9M2U5 cassette (2,626 bp)
`data/results/repair_template_O49146.fasta`	Repair template with the O49146 cassette (2,617 bp)
`data/results/s1_replaced_sequence.fasta`	The 101 bp between the two cuts, replaced by the cassette
`data/results/s1_candidates.csv`	The 28 candidate guides at S1
`data/results/s1_pairs.csv`	The 13 guide pairs around the insertion point, ranked
`data/results/site_context.csv`	Window, neighbouring genes, distances, expression and list counts of site S1
`data/results/site_search_summary.csv`	Gene and gap counts at each filter of the site search
`data/results/site_search_parameters.csv`	Settings of the site search, guide filters and pair rule
`data/results/mit_position_weights.csv`	The 20 MIT position weights
`data/figures/*`	Data behind the figures (one CSV per figure or panel)

Path	Contents
`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/input_sources.csv`	Public inputs: UniProt records, GenBank records, genome, RNA-seq series, with versions and checksums
`inputs/residue_masses.csv`	IUPAC average residue masses and water, used for the mass of the brazzein chain
`inputs/uniprot/*`	UniProt records of brazzein (P56552) and of the three signal sources (Q43413, Q9M2U5, O49146)
`inputs/genbank/*`	GenBank records of the tomato E8 gene region (X13437) and of two further tomato E8 promoter records
`inputs/parts/*`	The E8 promoter piece (X13437 bases 1 to 1121) and the 256-bp NOS terminator
`inputs/codon_usage/pepper_codon_usage_table.csv`	Codon counts and weights of the pepper coding sequences (UCD10Xv1.1)
`inputs/codon_usage/pepper_stop_usage.txt`	Terminal stop codon counts of the 48,931 clean pepper coding sequences
`inputs/guide_design/S1_window_131430500-131431600.fasta`	Genome sequence of the 1,101-bp window around the insertion point on chromosome 2 (NC_061112.1, plus strand)
`inputs/guide_design/S1_arm_left_500.fasta`	Left homology arm (500 bp)
`inputs/guide_design/S1_arm_right_500.fasta`	Right homology arm (500 bp)
`inputs/guide_design/guides_to_scan.txt`	The two guide protospacers
`inputs/guide_design/hits.tsv`	Stored whole-genome scan output for the two guides: contig, position, strand, guide index, mismatches, mismatch mask
`inputs/guide_design/s1_candidates_scored.tsv`	The 28 candidate guides at S1 with site counts and MIT score from the genome scan
`inputs/guide_design/offtarget_context.tsv`	Genome sites with 1 to 3 mismatches to the two guides, annotated as coding exon, gene body or intergenic
`inputs/site_search/candidate_windows.tsv`	The 1,129 windows of 3,000 bp that passed the site search, with neighbouring genes and placenta TPM
`inputs/phoenix_design_target.csv`	Heat design target and capsaicinoid blend of the cultivar Phoenix (SS-0096)
`inputs/ifi_reference_values.csv`	Reference compounds and reference EC50 values of the sweet and heat axes of the Iyengar Flavor Index
`code/01_protein.py`	Mature brazzein and signal peptides from the UniProt records
`code/02_codon_design.py`	Codon table, balanced back-translation, adaptation index, GC and motif scans
`code/03_cassettes.py`	Cassette assembly, block coordinates, checks and E8 promoter records
`code/04_site_guides.py`	Site S1, guide pair, repair templates, scanner agreement and site context
`code/05_tables.py`	Printed tables, table index and graphical abstract data
`code/06_full_scan.py`	Full-mode genome window extraction and guide scan
`code/seqtools.py`	Genetic code, translation, FASTA and record readers, GC, codon adaptation index, motif list
`code/guide_design/scan.c`	C program: scans the genome for sites with up to four mismatches to each guide (strict NGG PAM, both strands)
`code/heavy/NEEDS.txt`	What full mode needs
`research_record/scripts/t0_stop_usage.py`	Terminal stop-codon counts of the clean pepper coding sequences
`research_record/scripts/t12_brazzein_cds.py`	Brazzein record, pepper back-translation, adaptation index and signal frames
`research_record/scripts/t5_cassette.py`	Cassette assembly with the Q43413 signal for two codon choices
`research_record/scripts/t7_three_cassettes.py`	The three cassettes, one per signal
`research_record/scripts/t6a_sites.py`	Site search: gaps between genes, windows with few repeat bases, neighbour expression
`research_record/scripts/t6b_enum_guides.py`	Lists every SpCas9 guide near the insertion points
`research_record/scripts/t6c_score_pairs.py`	Merges scan hits, MIT score, guide pairs

Path	Contents
`research_record/scripts/t6e_site_context.py`	Site S1 context: neighbours, arms, replaced sequence
`research_record/scripts/t6f_offtarget_context.py`	Context (coding exon, gene body, intergenic) of the sites with 1 to 3 mismatches
`research_record/scripts/scan.c`	C scanner used for the genome scan
`research_record/scripts/run_scan_t6.sh`	Runs the C scanner on the guide batches
`research_record/scripts/genome_common.py`	Genome fetch and annotation helpers of the research scripts
`research_record/scripts/extract_pair_inputs.py`	Extracts the stored scan hits, candidate scores, off-target context and genome window of the two guides into inputs/guide_design/
`research_record/outputs/*`	Raw outputs of the research scripts and the cassettes they wrote
`research_record/second_script_rerun/scan_numpy.py`	Second genome scanner (numpy), independent of scan.c
`research_record/second_script_rerun/run_numpy_t6.sh`	Runs the second scanner on the guide pair
`research_record/second_script_rerun/t6d_*`	Second-scanner hits, its site-set comparison with the C scanner, the guide pair file and run time
`research_record/second_script_rerun/t5_bc_cmds.sh`	Length arithmetic repeated with the bc calculator
`research_record/second_script_rerun/p03_P56552_fetch.fasta`	Freshly fetched UniProt P56552 sequence used by the second alignment
`research_record/second_script_rerun/p03_align_second_script.py`	Second implementation: translation of each cassette frame and global alignment of the mature part to P56552
`research_record/second_script_rerun/p03_align_second_script_output.txt`	Output of the second alignment

Cite this paper

Ford, P.I. (2026). Sweet Heat: In Silico Design of Fruit-Specific Brazzein Expression in Capsicum annuum. Scoville Splice Research Series. https://doi.org/10.5281/zenodo.24511316

```
@misc{Ford2026_P03,
  author = {Ford, P.I.},
  title = {Sweet Heat: In Silico Design of Fruit-Specific Brazzein Expression in Capsicum annuum},
  year = {2026},
  publisher = {Scoville Splice LLC},
  series = {Scoville Splice Research Series},
  doi = {10.5281/zenodo.24511316},
  url = {https://doi.org/10.5281/zenodo.24511316}
}
```

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

Related papers in this series

ID	Title	DOI
P01	In Silico Design of Ten Capsicum annuum Cultivars Spanning 0.87 to 2.96 Million Scoville Heat Units	10.5281/zenodo.24511300
P04	The Iyengar Flavor Index (IFI): A Receptor-Normalized Standard for Cross-Modal Flavor Measurement	10.5281/zenodo.24511320
P07	Kokumi Engineering in Capsicum: In Silico Design of γ-Glutamyl-Valyl-Glycine Accumulation in Fruit	10.5281/zenodo.24511332
P09	Cool-Burn: In Silico Design of Combined TRPV1 and TRPM8 Activation in Capsicum annuum	10.5281/zenodo.24511340
P16	Computational Redesign of Brazzein, Single-Chain Monellin and Thaumatin with ProteinMPNN, ESMFold and FoldX	10.5281/zenodo.24511358

Appendix A. Evidence table

Claim ID	Claim	Value	Kind	Source file	Location
P03-C001	Mature brazzein is UniProt P56552	P56552	definition	`inputs/input_sources.csv`	row P56552
P03-C002	Mature brazzein is 54 residues long	54	computed	`data/results/brazzein_protein.csv`	row Length (residues)
P03-C003	The stored P56552 record is entry version 93	93	definition	`data/results/brazzein_protein.csv`	row Record entry version
P03-C004	Mature brazzein has 8 cysteines	8	computed	`data/results/brazzein_protein.csv`	row Cysteine residues (count)
P03-C005	The cysteines are at residues 4, 16, 22, 26, 37, 47, 49 and 52	4, 16, 22, 26, 37, 47, 49 and 52	computed	`data/results/brazzein_protein.csv`	row Cysteine positions
P03-C006	The record lists 4 disulfide pairs	4	computed	`data/results/brazzein_protein.csv`	row Disulfide pairs (count)
P03-C007	The disulfide pairs are 4-52, 16-37, 22-47 and 26-49	4-52, 16-37, 22-47 and 26-49	computed	`data/results/brazzein_protein.csv`	row Disulfide pairs in the record
P03-C008	The record gives an average mass of 6,498 Da	6,498	computed	`data/results/brazzein_protein.csv`	row Record average mass (Da)
P03-C009	The mass computed from the sequence is 6,498.32 Da (unmodified chain)	6,498.32	computed	`data/results/brazzein_protein.csv`	row Own average mass, unmodified (Da)

Claim ID	Claim	Value	Kind	Source file	Location
P03-C010	With four disulfides formed the computed mass is 6,490.26 Da	6,490.26	computed	`data/results/brazzein_protein.csv`	row Own average mass with four disulfides formed (Da)
P03-C011	The record lists pyroglutamate at residue 1	1	computed	`data/results/brazzein_protein.csv`	row N-terminal pyroglutamate listed at residue
P03-C012	The chain has no methionine and no tryptophan residue	no methionine or tryptophan	computed	`data/results/brazzein_protein.csv`	row Methionine and tryptophan residues (count) = 0
P03-C013	Average residue masses are the IUPAC values plus 18.01528 Da for water	18.01528	definition	`inputs/residue_masses.csv`	row water
P03-C014	Forming one disulfide removes 2 hydrogen atoms of 1.00794 Da each	1.00794	definition	`code/01_protein.py`	variable mass_ss
P03-C015	The signal source Q43413	Q43413	definition	`inputs/input_sources.csv`	row Q43413
P03-C016	The Q43413 signal is 27 residues long	27	computed	`data/results/signal_peptides.csv`	row Q43413, column signal_length_aa
P03-C017	The signal source Q9M2U5	Q9M2U5	definition	`inputs/input_sources.csv`	row Q9M2U5
P03-C018	The Q9M2U5 signal is 28 residues long	28	computed	`data/results/signal_peptides.csv`	row Q9M2U5, column signal_length_aa
P03-C019	The signal source O49146	O49146	definition	`inputs/input_sources.csv`	row O49146
P03-C020	The O49146 signal is 25 residues long	25	computed	`data/results/signal_peptides.csv`	row O49146, column signal_length_aa
P03-C021	Entry versions of the stored signal records are 96, 162 and 69	96, 162 and 69	definition	`data/results/signal_peptides.csv`	column record_entry_version
P03-C022	The Q43413 signal ends in Ala at -3 and Ala at -1	Ala at -3 and -1	computed	`data/results/signal_peptides.csv`	row Q43413, columns residue_minus3 and residue_minus1
P03-C023	The Q9M2U5 signal has Thr at -3 and Ala at -1	Thr at -3 and Ala at -1	computed	`data/results/signal_peptides.csv`	row Q9M2U5
P03-C024	The O49146 signal has Val at -3 and Ala at -1	Val at -3 and Ala at -1	computed	`data/results/signal_peptides.csv`	row O49146
P03-C025	The Q43413 signal carries evidence code ECO:0000269 (protein sequencing, PubMed 8883377)	ECO:0000269, PubMed 8883377	computed	`inputs/uniprot/Q43413.txt`	SIGNAL feature evidence and RX line
P03-C026	The signals of Q9M2U5 and O49146 carry evidence code ECO:0000255 (sequence analysis)	ECO:0000255	computed	`data/results/signal_peptides.csv`	column evidence

Claim ID	Claim	Value	Kind	Source file	Location
P03-C027	The Q43413 record lists fruit as its tissue	fruit	computed	`data/results/signal_peptides.csv`	row Q43413, column tissue_in_record
P03-C028	The O49146 record lists pericarp as its tissue	pericarp	computed	`data/results/signal_peptides.csv`	row O49146, column tissue_in_record
P03-C029	The codon table uses 48,931 clean pepper coding sequences	48,931	computed	`data/results/codon_design_parameters.csv`	row clean_coding_sequences
P03-C030	The table counts 21,189,723 codons	21,189,723	computed	`data/results/codon_design_parameters.csv`	row codon_total
P03-C031	The pepper coding sequences average 42.3% GC	42.3	computed	`data/results/codon_design_parameters.csv`	row pepper_cds_gc_percent
P03-C032	The pepper coding sequences average 37.5% GC at the third codon position	37.5	computed	`data/results/codon_design_parameters.csv`	row pepper_cds_gc3_percent
P03-C033	TGA ends 41.5% of the clean pepper coding sequences, TAA 33.7% and TAG 24.8%	41.5%, 33.7% and 24.8%	computed	`data/results/codon_design_parameters.csv`	rows stop_TGA_percent, stop_TAA_percent, stop_TAG_percent
P03-C034	The most-used codon is the same for 19 of 20 amino acids in the all-isoform and the one-per-gene counts	19 of 20	computed	`data/results/codon_design_parameters.csv`	row most_used_codons_agree_amino_acids
P03-C035	A codon is usable when its relative adaptiveness is at least 0.5	0.5	design assumption	`data/results/codon_design_parameters.csv`	row usable_weight_threshold
P03-C036	Each reading frame has 3,000 random draws	3,000	design assumption	`data/results/codon_design_parameters.csv`	row draws
P03-C037	A draw has up to 300 repair rounds	300	design assumption	`data/results/codon_design_parameters.csv`	row repair_rounds
P03-C038	The GC window is 3 points for the Q43413 reading frame	3	design assumption	`data/results/codon_design_parameters.csv`	row gc_window_Q43413
P03-C039	The GC window is 6 points for the other two reading frames	6	design assumption	`data/results/codon_design_parameters.csv`	row gc_window_other_signals

Claim ID	Claim	Value	Kind	Source file	Location
P03-C040	The back-translation checks 15 listed motifs and runs of five identical bases	15	design assumption	data/results/codon_design_parameters.csv	row motifs_checked
P03-C041	The reported motif scan covers 17 motifs	17	computed	data/results/codon_design_parameters.csv	row motifs_reported
P03-C042	The seed of the Q43413 reading frame is 24112026	24112026	design assumption	data/results/design_runs.csv	row Q43413
P03-C043	The seeds of the Q9M2U5 and O49146 signals are 2439 and 2436 (2411 plus the signal length)	2439 and 2436	design assumption	data/results/design_runs.csv	rows Q9M2U5 and O49146
P03-C044	30 of 3,000 draws pass the filters for Q43413	30	computed	data/results/design_runs.csv	row Q43413, column draws_kept
P03-C045	999 of 3,000 draws pass the filters for Q9M2U5	999	computed	data/results/design_runs.csv	row Q9M2U5, column draws_kept
P03-C046	2,990 of 3,000 draws pass the filters for O49146	2,990	computed	data/results/design_runs.csv	row O49146, column draws_kept
P03-C047	The Q43413 reading frame is 246 nt (81 residues and a stop codon)	246	computed	data/results/orf_summary.csv	row Q43413, column orf_nt
P03-C048	The Q9M2U5 reading frame is 249 nt	249	computed	data/results/orf_summary.csv	row Q9M2U5, column orf_nt
P03-C049	The O49146 reading frame is 240 nt	240	computed	data/results/orf_summary.csv	row O49146, column orf_nt
P03-C050	The reading frames encode 81, 82 and 79 residues before the stop codon	81, 82 and 79	computed	data/results/orf_summary.csv	column protein_length_aa
P03-C051	The codon adaptation index of the three reading frames is 0.860 to 0.897	0.860 to 0.897	computed	data/results/orf_summary.csv	rows Q43413, Q9M2U5, O49146
P03-C052	The Q43413 reading frame has CAI 0.860	0.860	computed	data/results/orf_summary.csv	row Q43413
P03-C053	The Q9M2U5 reading frame has CAI 0.892	0.892	computed	data/results/orf_summary.csv	row Q9M2U5
P03-C054	The O49146 reading frame has CAI 0.897	0.897	computed	data/results/orf_summary.csv	row O49146
P03-C055	The product form of the index agrees with the log-sum form to three decimals for all designs	three decimals	computed	data/results/orf_summary.csv	columns cai_all_isoform_weights and cai_product_form
P03-C056	Against the one-per-gene counts the index is 0.859, 0.889 and 0.893	0.859, 0.889 and 0.893	computed	data/results/orf_summary.csv	column cai_one_per_gene_weights
P03-C057	The GC of the three reading frames is 36.5 to 39.4%	36.5 to 39.4%	computed	data/results/orf_summary.csv	column gc_percent
P03-C058	The Q43413 reading frame has 39.4% GC and 48.8% GC3	39.4% GC and 48.8% GC3	computed	data/results/orf_summary.csv	row Q43413

Claim ID	Claim	Value	Kind	Source file	Location
P03-C059	The Q9M2U5 reading frame has 36.5% GC and 37.3% GC3	36.5% GC and 37.3% GC3	computed	data/results/orf_summary.csv	row Q9M2U5
P03-C060	The O49146 reading frame has 37.9% GC and 35.0% GC3	37.9% GC and 35.0% GC3	computed	data/results/orf_summary.csv	row O49146
P03-C061	The mature chain with balanced codons has CAI 0.861, 38.1% GC and 48.2% GC3	0.861, 38.1% GC and 48.2% GC3	computed	data/results/orf_summary.csv	row mature, balanced codons
P03-C062	The mature chain with the most-used codon for every residue has CAI 1.000, 24.4% GC and 1.8% GC3	1.000, 24.4% GC and 1.8% GC3	computed	data/results/orf_summary.csv	row mature, most-used codons
P03-C063	The most-used design holds TTTT at 2 positions and AAAAAA at 2 positions	TTTT at 2 positions and AAAAAA at 2 positions	computed	data/results/orf_motifs.csv	rows mature, most-used codons
P03-C064	27 of the 54 mature codons differ between the most-used and the balanced design	27 of the 54	computed	data/results/codon_design_parameters.csv	row codons_changed_most_used_to_balanced_mature
P03-C065	None of the 17 listed motifs occurs in the three balanced reading frames	none of the listed motifs	computed	data/results/orf_motifs.csv	rows ORF Q43413, ORF Q9M2U5, ORF O49146, count 0
P03-C066	The E8 promoter piece is from GenBank X13437	X13437	definition	inputs/input_sources.csv	row X13437
P03-C067	X13437 is 3,042 bp long	3,042	computed	data/results/promoter_records.csv	row X13437 length (bp)
P03-C068	X13437 is from Solanum lycopersicum, cultivar Cherry	cultivar Cherry	computed	data/results/promoter_records.csv	row X13437 organism and cultivar
P03-C069	The X13437 record places the TATA box at 1,055 to 1,063	1,055 to 1,063	computed	data/results/promoter_records.csv	row X13437 TATA box
P03-C070	The record places the transcription start at 1,087	1,087	computed	data/results/promoter_records.csv	row X13437 transcription start
P03-C071	The record places the E8 start codon at 1,122	1,122	computed	data/results/promoter_records.csv	row X13437 E8 start codon
P03-C072	The promoter piece is X13437 bases 1 to 1,121 (1,121 bp)	1,121	computed	data/results/promoter_records.csv	row Promoter piece length (bp)
P03-C073	The 5' leader from the transcription start to the base before the E8 start codon is 35 nt with no ATG	35	computed	data/results/promoter_records.csv	row 5' leader length
P03-C074	DQ317599 is 2,175 bp long (cultivar Jinfeng 1)	2,175	computed	data/results/promoter_records.csv	row DQ317599 length (bp)
P03-C075	X13437 bases 1 to 1,086 occur unchanged in DQ317599 starting at position 1,090	1,090	computed	data/results/promoter_records.csv	row X13437 bases 1-1086 found unchanged

Claim ID	Claim	Value	Kind	Source file	Location
P03-C076	All 1,086 of 1,086 bases are identical	1,086 of 1,086	computed	`data/results/promoter_records.csv`	row identical to DQ317599
P03-C077	AF515784 is 2,191 bp long (cultivar Zhongshu 5)	2,191	computed	`data/results/promoter_records.csv`	row AF515784 length (bp)
P03-C078	X13437 bases 1 to 1,086 align to AF515784 with 1,081 of 1,086 identical columns (99.5%)	1,081 of 1,086 columns (99.5%)	computed	`data/results/promoter_records.csv`	rows aligned to AF515784 and percent identical
P03-C079	The alignment scores a match as 2, a mismatch as -3 and a gap column as -4	2, -3 and -4	definition	`code/03_cassettes.py`	function semi_global
P03-C080	The NOS terminator is 256 bp	256	computed	`data/results/cassette_summary.csv`	column terminator_bp
P03-C081	The NOS terminator is bases 4022 to 4277 of pBI121 (AF485783)	bases 4022 to 4277 of AF485783	definition	`inputs/input_sources.csv`	row AF485783
P03-C082	The Q43413 cassette is 1,623 bp	1,623	computed	`data/results/cassette_summary.csv`	row Q43413, column cassette_bp
P03-C083	The Q9M2U5 cassette is 1,626 bp	1,626	computed	`data/results/cassette_summary.csv`	row Q9M2U5, column cassette_bp
P03-C084	The O49146 cassette is 1,617 bp	1,617	computed	`data/results/cassette_summary.csv`	row O49146, column cassette_bp
P03-C085	The cassettes are 1,617 to 1,626 bp long	1,617 to 1,626	computed	`data/results/cassette_summary.csv`	column cassette_bp
P03-C086	The cassettes have 29.0%, 28.5% and 28.7% GC	29.0%, 28.5% and 28.7%	computed	`data/results/cassette_summary.csv`	column gc_cassette_percent
P03-C087	The promoter piece occupies cassette positions 1 to 1,121	1 to 1,121	computed	`data/results/cassette_blocks.csv`	row E8 promoter, variant Q43413
P03-C088	In the Q43413 cassette the signal occupies positions 1,122 to 1,202 (81 nt)	1,122 to 1,202 (81 nt)	computed	`data/results/cassette_blocks.csv`	row Signal Q43413
P03-C089	Mature brazzein occupies positions 1,203 to 1,364 (162 nt)	1,203 to 1,364 (162 nt)	computed	`data/results/cassette_blocks.csv`	row Mature brazzein, variant Q43413
P03-C090	The stop codon is at positions 1,365 to 1,367	1,365 to 1,367	computed	`data/results/cassette_blocks.csv`	row Stop TGA, variant Q43413
P03-C091	The NOS terminator occupies positions 1,368 to 1,623	1,368 to 1,623	computed	`data/results/cassette_blocks.csv`	row NOS terminator, variant Q43413

Claim ID	Claim	Value	Kind	Source file	Location
P03-C092	The three cassettes share the same 421 bp from the first mature codon to the end	421	computed	`data/results/cassette_checks.csv`	row all, shared tail length
P03-C093	The designed frame is the longest ATG-to-stop frame on either strand of each cassette (81, 82 and 79 residues)	81, 82 and 79	computed	`data/results/cassette_checks.csv`	rows longest ATG-to-stop frame
P03-C094	The next longest frame is 68 residues in the Q43413 cassette	68	computed	`data/results/cassette_checks.csv`	row Q43413 second longest frame
P03-C095	The first ATG after the transcription start is at position 1,122 in all three cassettes	1,122	computed	`data/results/cassette_checks.csv`	row first ATG after the transcript start
P03-C096	None of 9 listed cloning sites occurs in any cassette	none of 9 listed cloning sites	computed	`data/results/cassette_sites.csv`	rows with count 0
P03-C097	All 3 cassettes equal the sequences written by the research scripts	3	computed	`data/results/cassette_crosscheck.csv`	column sequences_identical
P03-C098	The length sums 1121 + 246 + 256 = 1623, 1121 + 249 + 256 = 1626 and 1121 + 240 + 256 = 1617 agree with the bc calculator	1,623, 1,626 and 1,617	computed	`research_record/outputs/t7_bc_check.txt`	lines 1 to 3
P03-C099	In the second alignment, the mature part of each of the 3 cassettes is identical to P56552 in 54 of 54 residues	54 of 54	computed	`research_record/second_script_rerun/p03_align_second_script_output.txt`	lines for Q43413, Q9M2U5 and O49146
P03-C100	Site S1 lies on NC_061112.1	NC_061112.1	definition	`inputs/guide_design/S1_window_131430500-131431600.fasta`	header
P03-C101	The S1 window is NC_061112.1:131,429,594-131,432,593	131,429,594 to 131,432,593	computed	`data/results/site_context.csv`	row Window
P03-C102	The S1 window is 3,000 bp	3,000	computed	`data/results/site_context.csv`	row Window length (bp)
P03-C103	The insertion point is 131,431,093, the midpoint of the window	131,431,093	computed	`data/results/site_context.csv`	row Insertion point
P03-C104	The S1 window holds a repeat-masked fraction of 0.0	0.0	computed	`data/results/site_context.csv`	row Repeat-masked fraction of the window
P03-C105	The left neighbour is LOC107859470 (131,375,188 to 131,393,093, plus strand)	LOC107859470	computed	`data/results/site_context.csv`	row Left neighbour (LOC ID)

Claim ID	Claim	Value	Kind	Source file	Location
P03-C106	The left neighbour is 38,000 bp from the insertion point	38,000	computed	data/results/site_context.csv	row Left neighbour distance
P03-C107	The right neighbour is LOC107858201 (131,439,385 to 131,443,927, minus strand)	LOC107858201	computed	data/results/site_context.csv	row Right neighbour (LOC ID)
P03-C108	The right neighbour is 8,292 bp from the insertion point	8,292	computed	data/results/site_context.csv	row Right neighbour distance
P03-C109	The gap between the two neighbouring genes is 46,291 bp	46,291	computed	data/results/site_context.csv	row Gap between the neighbouring genes (bp)
P03-C110	12 protein-coding genes lie within 100 kb of the window centre	12	computed	data/results/site_context.csv	row Protein-coding genes within 100 kb
P03-C111	The left neighbour is at 7.1 to 15.2 TPM in placenta at the five stages	7.1 to 15.2	computed	data/results/site_context.csv	rows Left neighbour placenta TPM
P03-C112	The right neighbour is at 8.6 to 20.4 TPM	8.6 to 20.4	computed	data/results/site_context.csv	rows Right neighbour placenta TPM
P03-C113	The placenta series has five stages: 4, 24, 34, 44 and 54 days post anthesis, one library per stage	4, 24, 34, 44 and 54	definition	data/figures/F08.csv	column dpa
P03-C114	The placenta RNA-seq series is PRJNA789050 (Capsicum chinense SL08)	PRJNA789050	definition	inputs/input_sources.csv	row RNA-seq series
P03-C115	The reference genome is Capsicum annuum UCD10Xv1.1, RefSeq assembly GCF_002878395.1	GCF_002878395.1	definition	inputs/input_sources.csv	row genome
P03-C116	The assembly annotates 40,503 genes and pseudogenes on the 12 chromosomes, 27,958 protein-coding	40,503 genes and pseudogenes, 27,958 protein-coding	computed	data/tables/S09.csv	rows 1 and 2
P03-C117	38,660 gaps lie between neighbouring annotated genes	38,660	computed	data/results/site_search_summary.csv	row Gaps between neighbouring genes
P03-C118	10,543 of the gaps are 13 to 60 kb long	10,543	computed	data/results/site_search_summary.csv	row Gaps of 13 to 60 kb
P03-C119	5,328 of those gaps have protein-coding genes on both sides	5,328	computed	data/results/site_search_summary.csv	row Gaps of 13 to 60 kb with protein-coding genes on both sides
P03-C120	1,129 gaps hold a 3,000-bp window with 10% or less repeat-masked bases	1,129	computed	data/results/site_search_summary.csv	row Gaps holding a 3,000-bp window
P03-C121	9 of those windows hold no repeat-masked base	9	computed	data/results/site_context.csv	row Windows in the list with no repeat-masked base

Claim ID	Claim	Value	Kind	Source file	Location
P03-C122	A window keeps at least 5,000 bp from every annotated gene or pseudogene	5,000	design assumption	data/results/site_search_parameters.csv	row distance_from_genes_bp
P03-C123	The window is 3,000 bp	3,000	design assumption	data/results/site_search_parameters.csv	row window_bp
P03-C124	The step between windows is 500 bp	500	design assumption	data/results/site_search_parameters.csv	row step_bp
P03-C125	At most 10% of the window may be repeat-masked	10%	design assumption	data/results/site_search_parameters.csv	row repeat_masked_fraction_max
P03-C126	The gap between neighbours is 13 to 60 kb	13 to 60 kb	design assumption	data/results/site_search_parameters.csv	rows minimum_gap and maximum_gap
P03-C127	Guide S1-g1 is GTTGGAGCGAGTACTATGCA with PAM AGG	GTTGGAGCGAGTACTATGCA	computed	data/results/guide_pair.csv	row S1-g1
P03-C128	Guide S1-g2 is ACCTAATGGGTTGCAAGGAA with PAM TGG	ACCTAATGGGTTGCAAGGAA	computed	data/results/guide_pair.csv	row S1-g2
P03-C129	S1-g1 protospacer is at 131,430,997 to 131,431,016	131,430,997 to 131,431,016	computed	data/results/guide_pair.csv	row S1-g1, column protospacer_position
P03-C130	S1-g2 protospacer is at 131,431,098 to 131,431,117	131,431,098 to 131,431,117	computed	data/results/guide_pair.csv	row S1-g2, column protospacer_position
P03-C131	S1-g1 cuts after 131,430,999 (94 bp left of the insertion point)	131,430,999	computed	data/results/guide_pair.csv	row S1-g1, column cut_after
P03-C132	S1-g2 cuts after 131,431,100 (7 bp right of the insertion point)	131,431,100	computed	data/results/guide_pair.csv	row S1-g2, column cut_after
P03-C133	The S1-g1 cut is 94 bp left and the S1-g2 cut is 7 bp right of the insertion point	94 bp left and 7 bp right	computed	data/results/guide_pair.csv	column cut_relative_to_insertion_point_bp
P03-C134	The two cuts are 101 bp apart	101	computed	data/results/site_window_checks.csv	row distance between the cuts (bp)
P03-C135	The guides have 50% and 45% GC	50% and 45%	computed	data/results/guide_pair.csv	column gc_percent
P03-C136	Each guide has 1 perfect site in the genome	1	computed	data/results/guide_pair.csv	column perfect_sites
P03-C137	S1-g1 has 0, 9, 25 and 186 sites with 1, 2, 3 and 4 mismatches (221 sites in all)	0, 9, 25 and 186	computed	data/results/guide_pair.csv	row S1-g1
P03-C138	S1-g2 has 0, 4, 40 and 182 sites with 1, 2, 3 and 4 mismatches (227 sites in all)	0, 4, 40 and 182	computed	data/results/guide_pair.csv	row S1-g2

Claim ID	Claim	Value	Kind	Source file	Location
P03-C139	S1-g1 has 221 sites with up to 4 mismatches	221	computed	data/results/scanner_agreement.csv	row S1-g1
P03-C140	S1-g2 has 227 sites with up to 4 mismatches	227	computed	data/results/scanner_agreement.csv	row S1-g2
P03-C141	The MIT score of S1-g1 is 64.0	64.0	computed	data/results/guide_pair.csv	row S1-g1, column mit_score_recomputed
P03-C142	The MIT score of S1-g2 is 70.3	70.3	computed	data/results/guide_pair.csv	row S1-g2, column mit_score_recomputed
P03-C143	The recomputed MIT scores equal the scores of the scan output	equal	computed	data/results/guide_pair.csv	columns mit_score_recomputed and mit_score_scan_output
P03-C144	The MIT score uses 20 position weights from 0 to 0.851	20 position weights from 0 to 0.851	definition	data/results/mit_position_weights.csv	column weight
P03-C145	The scan allows up to 4 mismatches	4	design assumption	data/results/site_search_parameters.csv	row mismatches_scanned
P03-C146	S1-g1 has 34 sites with 1 to 3 mismatches	34	computed	data/results/offtarget_by_context.csv	row S1-g1, 1 to 3, total
P03-C147	Of those, 25 are intergenic, 2 lie in gene bodies and 7 in coding exons	25 are intergenic, 2 in gene bodies and 7 in coding exons	computed	data/results/offtarget_by_context.csv	row S1-g1, 1 to 3
P03-C148	S1-g2 has 44 sites with 1 to 3 mismatches	44	computed	data/results/offtarget_by_context.csv	row S1-g2, 1 to 3, total
P03-C149	Of those, 43 are intergenic, 1 lies in a gene body and none in coding exons	43 are intergenic, 1 in a gene body and none in coding exons	computed	data/results/offtarget_by_context.csv	row S1-g2, 1 to 3
P03-C150	The sites with 2 mismatches (9 for S1-g1 and 4 for S1-g2) are all intergenic	9 and 4	computed	data/results/offtarget_by_context.csv	rows mismatches 2
P03-C151	28 candidate guides lie within 110 bp of the insertion point and pass the filters	28	computed	data/results/s1_candidates.csv	number of rows
P03-C152	9 of the 28 have exactly 1 perfect site at the intended position	9	computed	data/results/s1_candidates.csv	rows with perfect_sites 1 and perfect_site_at_intended_position True
P03-C153	13 pairs of unique guides have one cut on each side of the insertion point and cuts 20 to 120 bp apart	13	computed	data/results/s1_pairs.csv	number of rows
P03-C154	The cuts must be 20 to 120 bp apart	20 to 120	design assumption	data/results/site_search_parameters.csv	row pair_rule

Claim ID	Claim	Value	Kind	Source file	Location
P03-C155	The chosen pair ranks first with a lowest MIT score of 64.0	64.0	computed	data/results/s1_pairs.csv	row rank 1
P03-C156	Candidate guides have GC 40 to 70%, no TTTT and no run of 5 identical bases	40 to 70%	design assumption	data/results/site_search_parameters.csv	rows guide_gc_range_percent, guide_filters
P03-C157	Guides are listed within 110 bp either side of the insertion point	110	design assumption	data/results/site_search_parameters.csv	row guide_scan_window_bp
P03-C158	The cut lies 3 bp from the PAM	3	definition	code/04_site_guides.py	docstring, conventions
P03-C159	The window used for the arms is 1,101 bp (NC_061112.1:131,430,500-131,431,600)	1,101	computed	data/results/site_window_checks.csv	row window length (bp)
P03-C160	The window runs from 131,430,500 to 131,431,600	131,430,500 to 131,431,600	computed	data/results/site_window_checks.csv	rows window start and window end
P03-C161	Each homology arm is 500 bp	500	design assumption	data/results/site_window_checks.csv	rows left arm (bp), right arm (bp)
P03-C162	The arms have 49.0% and 45.4% GC and no N	49.0% and 45.4%	computed	data/results/site_window_checks.csv	rows left arm GC, right arm GC
P03-C163	Both arms hold 0 repeat-masked bases	0	computed	research_record/outputs/t6e_output.txt	lines left arm and right arm
P03-C164	101 bp lie between the arms and are replaced by the cassette	101	computed	data/results/site_window_checks.csv	row bases between the arms (bp)
P03-C165	The Q43413 repair template is 2,623 bp	2,623	computed	data/results/repair_templates.csv	row Q43413
P03-C166	The Q9M2U5 repair template is 2,626 bp	2,626	computed	data/results/repair_templates.csv	row Q9M2U5
P03-C167	The O49146 repair template is 2,617 bp	2,617	computed	data/results/repair_templates.csv	row O49146
P03-C168	The repair templates are 2,617 to 2,626 bp	2,617 to 2,626	computed	data/results/repair_templates.csv	column template_bp
P03-C169	No intact target site of either guide (protospacer and NGG) remains in the 3 repair templates	no intact target site of either guide	computed	data/results/repair_templates.csv	column intact_guide_target_sites_in_template
P03-C170	The numpy scanner finds the same 221 and 227 sites as the C scanner	221 and 227	computed	data/results/scanner_agreement.csv	rows S1-g1 and S1-g2

Claim ID	Claim	Value	Kind	Source file	Location
P03-C171	The heat design target of Phoenix (SS-0096) is 2,956,310 SHU	2,956,310	design target	`inputs/phoenix_design_target.csv`	row SS-0096
P03-C172	The reference EC50 of brazzein at TAS1R2/TAS1R3 in the sweet axis of the Iyengar Flavor Index is 0.004 micromolar	0.004	definition	`inputs/ifi_reference_values.csv`	row IFI-S
P03-C173	Scripts run on Python 3.9 or newer	3.9	definition	`code/LIBRARIES.txt`	line 1
P03-C174	Brazzein was reported as a high-potency, thermostable sweet protein of Pentadiplandra brazzeana	high-potency, thermostable sweet protein	literature value	`literature`	Ming and Hellekant 1994, doi:10.1016/0014-5793(94)01184-2
P03-C175	The solution structure of brazzein was reported	solution structure	literature value	`literature`	Caldwell et al. 1998, doi:10.1038/nsb0698-427
P03-C176	Brazzein was reported to dock mainly to the T1R2 subunit of the sweet taste receptor	mainly to the T1R2 subunit	literature value	`literature`	Walters and Hellekant 2006, doi:10.1021/jf062359y
P03-C177	E8 is an ethylene-activated tomato gene whose 5' flanking region gives ripening-related expression	ripening-related expression	literature value	`literature`	Deikman and Fischer 1988, doi:10.1002/j.1460-2075.1988.tb03202.x
P03-C178	The pepper defensin J1-1 is reported to be expressed in ripe pepper fruit	expressed in ripe pepper fruit	literature value	`literature`	Meyer et al. 1996, doi:10.1104/pp.112.2.615
P03-C179	The promoter comparison uses X13437 bases 1 to 1,086 (up to the base before the transcription start)	1,086	computed	`data/results/promoter_records.csv`	rows X13437 bases 1-1086
P03-C180	The site search covers the 12 placed chromosomes (NC_061111.1 to NC_061122.1)	12	definition	`data/results/site_search_parameters.csv`	row chromosomes
P03-C181	Mismatch positions are counted from the PAM-distal end, positions 1 to 20	1 to 20	definition	`data/results/mit_position_weights.csv`	column position_from_pam_distal_end
P03-C182	The scan lists 448 sites for the two guides in all (221 and 227)	448	computed	`inputs/guide_design/hits.tsv`	number of rows

Appendix B. Glossary, abbreviations and units

Term	Meaning	Type
CAI	Codon adaptation index: the geometric mean of the relative adaptiveness of the codons of a coding sequence (0 to 1)	abbreviation
GC	Share of G and C bases in a sequence	abbreviation
GC3	GC share at the third position of the codons	abbreviation
w	Relative adaptiveness of a codon: its count divided by the count of the most-used codon of the same amino acid	term

Term	Meaning	Type
ORF	Open reading frame: the codons from a start codon to the stop codon	abbreviation
Reading frame	The run of codons from the ATG to the stop codon that is translated	term
Signal peptide	N-terminal peptide that sends a protein into the secretory pathway and is cleaved off	term
Mature brazzein	The 54-residue chain of brazzein without any signal	term
Cassette	Promoter, reading frame and terminator as one DNA unit	term
E8 promoter	Promoter of the ethylene-activated tomato fruit gene E8	term
NOS terminator	Terminator of the Agrobacterium nopaline synthase gene	abbreviation
5' leader	Transcribed sequence between the transcription start and the start codon	term
Disulfide	A bond between two cysteine residues	term
Cys	Cysteine	abbreviation
Gln	Glutamine	abbreviation
Ala	Alanine	abbreviation
Thr	Threonine	abbreviation
Val	Valine	abbreviation
Leu	Leucine	abbreviation
Ser	Serine	abbreviation
Met	Methionine	abbreviation
Trp	Tryptophan	abbreviation
Da	Dalton: unit of molecular mass	unit
bp	Base pairs	unit
nt	Nucleotides	unit
kb	Kilobases: 1,000 bases	unit
aa	Amino acids (residues)	unit
TPM	Transcripts per million: a normalised measure of transcript abundance	unit
DPA	Days post anthesis: days after the flower opens	unit
SHU	Scoville heat units: the scale of pungency	unit
IFI-S	Sweet axis of the Iyengar Flavor Index; the reference compound is brazzein at the TAS1R2/TAS1R3 receptor	term
EC50	Concentration giving half of the maximum response of a receptor assay	abbreviation
TAS1R2/TAS1R3	The sweet taste receptor, a heteromer of two taste receptor type 1 subunits	abbreviation

Term	Meaning	Type
SpCas9	Streptococcus pyogenes Cas9 nuclease	abbreviation
PAM	Protospacer-adjacent motif: the NGG sequence next to the target that SpCas9 needs	abbreviation
Protospacer	The 20-nucleotide target sequence of a guide RNA	term
Guide	A guide RNA, named here by its 20-nucleotide protospacer	term
MIT score	Hsu et al. (2013) aggregate specificity score of a guide: higher means fewer likely off-target sites	term
Mismatch	A position where a genome site differs from the guide	term
S1	The knock-in site on chromosome 2 (NC_061112.1:131,429,594-131,432,593)	term
Insertion point	The position in the site where the cassette is placed	term
Homology arm	DNA of the genome next to the cuts that is placed on each side of the cassette so that repair copies it in	term
Repair template	Left arm, cassette and right arm as one DNA piece	term
Knock-in	Placing a DNA cassette at a chosen genome position	term
Intergenic	Between genes	term
LOC ID	NCBI gene identifier of the form LOC followed by digits	term
UniProt	Protein sequence and annotation database	abbreviation
GenBank	NCBI nucleotide sequence database	abbreviation
RefSeq	NCBI reference sequence collection	abbreviation
BLOSUM62	Amino-acid substitution matrix used for the alignment	term
bc	Command-line calculator used to repeat the length sums	term
Soft-masked	Genome bases written in lower case because they are repeats	term
Gene body	The span of an annotated gene from its first to its last base	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_protein.py: 0.1 s; 2 code/02_codon_design.py: 11.0 s; 3 code/03_cassettes.py: 0.3 s; 4 code/04_site_guides.py: 0.1 s; 5 code/05_tables.py: 0.1 s

Item	Value
Input files	24 files, SHA-256 in `code/environment.json`
Data fingerprint	`279e903f4f2cbeadc53f8ce5b1be9f4ea55bf8ac79e92b9ae7048c791a045e89`

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.24511316 · 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