Electronic Supplementary Information 1

Efficient Enzyme-catalyzed Production of Diosgenin: Inspired by the Biotransformation Mechanisms of Steroid Saponins in *Talaromyces Stollii* CLY-6

Leiyu Cheng[1,2,3#], Han Zhang[1#], Haiyang Cui[4, 5], Mehdi D. Davari[4], Bin Wei[2], Wenya Wang[2,3]*, Qipeng Yuan[1,2]*

[1]Beijing Advanced Innovation Center for Soft Matter Science and Engineering, Beijing University of Chemical Technology, Beijing 100029, PR China; No.15 North Third Ring East Road, Chao yang District, Beijing 100029, P. R. China.

[2]State Key Laboratory of Chemical Resource Engineering, College of Life Science and Technology, Beijing University of Chemical Technology, Beijing 100029, P. R. China; No.15 North Third Ring East Road, Chao yang District, Beijing 100029, P. R. China.

[3]Amoy-BUCT Industrial Bio-technovation Institute, Beijing University of Chemical Technology, Amoy, 361022, China; No.126 Xinyuan Road, Haicang District, Amoy 361028, P. R. China.

[4]Institute of Biotechnology, RWTH Aachen University, Worringerweg 3, Aachen 52074, Germany.

[5]DWI-Leibniz Institute for Interactive Materials, Forckenbeckstraße 50, Aachen 52056, Germany.

[#]**These authors contributed to work equally.**

***Corresponding author.**
E-mail address:
Wangwy@mail.buct.edu.cn (Wenya Wang)
Yuanqp@mail.buct.edu.cn (Qipeng Yuan)
Fax: 010-64437610
Tel: 010-64421335

1. Strain isolation and screening

Strain isolation was performed with the medium containing 20 g/L agar and 20 g/L E-DZW (the crude extract of the tubers of *Dioscorea zingiberensis C.H. Wright*). Sample of strains mixture was prepared by mixing 10 mL sterile water with 1 g E-DZW and then diluted, evenly spread on the isolation medium. After being cultured under aerobic conditions at 30 ℃ for 6 days, colonies on the plate were then picked up based on the morphology, followed by cultivation in potato dextrose agar (PDA) medium. Finally, 6 unknown fungi were obtained and tentatively named as Fungus CLY-1, Fungus CLY-2, Fungus CLY-3, Fungus CLY-4, Fungus CLY-5, and Fungus CLY-6.

These isolated fungi and other 6 purchased filamentous fungi were used for diosgenin production from E-DZW by microbial biotransformation. The medium (pH 5.5) containing 3 g/L glucose, 8 g/L yeast extract, 3 g/L powder (100-mesh) of dried *Dioscorea zingiberensis C.H. Wright* tubers (P-DZW), 1 g/L $(NH_4)_2SO_4$, 3 g/L KH_2PO_4, 2 g/L $MgSO_4 \cdot 7H_2O$, and 0.8‰ (v/v) Tween 80 was prepared for the submerged culture of fungi at 160 rpm, 30 ℃. After 5 days' fermentation, the culture supernatants were collected for bioconversions (20 mL supernatant) using E-DZW (10 mg) as the substrate at pH 5.0, 50 ℃ for 72 hours. After bioconversions, the supernatants were removed by centrifugation at 8000 rpm for 20 min; then, the obtained products were analyzed by High-Performance Liquid Chromatography (HPLC) (Shimadzu LC-20AT, Japan). The fungus with the highest diosgenin yield was selected for further experiments.

2. Microorganism identification

Molecular taxonomy has been applied to classify fungal species since the 1990s [1], and it has been becoming a powerful method for microorganism identification. Internal Transcribed Spacer (ITS) region and 18S ribosomal DNA (18S rDNA) have been regarded as the potential barcodes for most fungi due to the high degree of interspecific variability. Additionally, several protein-coding areas, such as *β*-tubulin and

calmodulin genes, have become compelling molecular data that could be accepted as the secondary barcodes for the rapid identification of specimens [2]. Thus, the combination of different DNA markers was considered to be an efficient approach for species-level identification of specific fungi [3].

Genomic DNA extraction of Fungus CLY-6 was conducted using a Rapid Fungi Genomic DNA Isolation Kit supplied by Sangon Biotech (Shanghai, China). Polymerase chain reactions (PCR) were carried out in a Pro-Flex PCR System (Thermo Scientific, USA) for DNA amplification by using the oligonucleotides as primer sequences (**Table S1)**. The resultant PCR products, purified by using DNA Gel Extraction Kit (Omega Biotek, Germany), were sequenced by a POP-7™ Polymer for 3730XL DNA Analyzers (Applied Biosystems, USA). The sequenced barcodes were then analyzed by using the Basic Local Alignment Search Tool (BLAST) in the GenBank (http://www.ncbi.nem.nih.gov), and the obtained data sets were aligned in the Clustal W (version 2.1) program. Maximum-likelihood (ML) trees [4] of the aligned data sets were further constructed using MEGA software (version 6.0.6) [5] with the Kimura 2-parameter model to calculate evolutionary distances. Bootstrap values were calculated from 1000 replicate runs using the routines in the MEGA software.

3. Genome sequencing and gene annotation

Genomic DNA extraction of Fungus CLY-6 was conducted by using a Rapid Fungi Genomic DNA Isolation Kit supplied by Sangon Biotech (Shanghai, China). Then, the fragmentation of genome DNA was carried out by using an M220™ Focused-ultrasonicator (Covaris, USA) for DNA library construction. Afterwards, the end-filling and linker-adding of DNA fragments was performed using an NEB Next® Ultra™ II DNA Library Prep Kit for Illumina® (New England Biolabs GmbH, UK) to obtain a high-quality DNA library for genome sequencing. Subsequently, the genome DNA was sequenced using an Illumina MiSeq 4000 platform at the Institute of Microbial Epidemiology, Military Medical Research Institute (Beijing, China). Quality evaluation of the raw sequencing data was performed using FastQC [6]. The Trimmomatic [7] was

used to improve the accuracy of genome data, then stitched by SPAdes [8]. After supplementing the gap to the contigs obtained by splicing with GapFiller [9], the correction of editing errors and the insertion and deletion of small fragments during sequence-splicing was performed using PrInSeS-G [10], resulting in an assembled genome sequence. Next, the genetic elements, such as ORF, tRNA, rRNA, etc., were predicted by Prokka [11]. Comparison of protein sequences was performed by NCBI Blast+ [12] using the databases such as CDD (https://www.ncbi.nlm.nih.gov/cdd/), NCBI COG (https://www.ncbi.nlm.nih.gov/COG/), NCBI NR (http://ncbi.nlm.nih.gov/), PFAM (http://pfam.xfam.org/), Swiss-Prot (https://www.uniprot.org/statistics/Swiss-Prot) and TrEMBL (https://www.uniprot.org/statistics/TrEMBL), etc. to obtain functional annotation information.

4. Native enzyme production

The inoculum was obtained by inoculating *Talaromyces stollii* CLY-6 in the seed medium (potato dextrose) shaking at 160 rpm, 30 °C for 36 h. The substrate of solid-state fermentation in each Erlenmeyer flask (500 mL) consisted of 7 g bran and 3 g P-DZW, which was moistened with nutrient ion solution (2 g/L $(NH_4)_2SO_4$, 2 g/L KH_2PO_4, 2 g/L $MgSO_4 \cdot 7H_2O$, pH 5.5) to make the initial moisture level of the medium at 75% before autoclaving. The prepared inoculum was then transferred into mediums, and the fungus was cultivated at 30 °C for 7 days. The secreted proteins were extracted from the medium by adding distilled water and shaking at 160 rpm, 40 °C for 40 min. The supernatant (crude enzyme) was obtained by suction filtration, followed by centrifugation at 10000 rpm, 4 °C for 20 min.

5. Enzyme purification

The first enzyme purification step was performed by adding ammonium sulfate into the supernatant slowly to 75% of ammonium sulfate, then placing the solution in a refrigerator at 4 °C overnight. The precipitates were collected by centrifugation at 10000 rpm at 4 °C for 15 min, and then re-dissolved in 20 mM sodium phosphate buffer

(pH 6.0), followed by removal of insoluble particles with 0.22 μm filter. An ultrafiltration centrifuge tube with 10 kDa interception (Millipore, USA) was used for protein desalination and enrichment at 4000 rpm, 4 °C. Hereafter, the enzyme purification was carried out using an ÄKTA Purifier HPLC system (GE Healthcare Life Sciences, USA) at 4 °C. The desalted crude protein (pH 6.0) was first applied to a Q-Sepharose Fast Flow column (5 mL, Bersee, China) and eluted by NaCl solution with the concentration of 0.05, 0.1, 0.15, 0.2, 0.25, 0.3 and 0.5 M for 2, 2, 4, 4, 5, 4, and 4 column volumes (CVs), respectively. The active fractions were pooled and desalted, and re-purified by a DEAE-Sepharose Fast Flow column (5 mL, Bersee, China), followed by protein elution by NaCl solution with the concentration of 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, and 0.5 M for 2, 2, 4, 4, 3, 3, 3 and 3 CVs, respectively. The Superdex^{TM} 200 Increase 10/300 GL (GE Healthcare, USA), pre-equilibrated with elution buffer (50 mM sodium phosphate, 150 mM NaCl, pH 7.0) was used for final purification, and the protein was eluted at a flow rate of 0.3 mL/min. Protein purity was evaluated by 10% sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE). Protein concentration was measured by bicinchoninic acid protein assay kit (Pierce Chemicals, USA) with bovine serum albumin (Sigma, Germany) as the standard.

6. Protein identification

The purified protein band was excised from the SDS-PAGE gel for sample preparation following reported procedure [13]. Samples were analyzed by ABI 5800 Matrix-Assisted Laser Desorption/Ionization Time of Flight (MALDI-TOF/TOF) Plus mass spectrometer (Applied Biosystems, USA) with the mass spectrum scanning range from 700 to 3500 Da of peptide mass fingerprinting (PMF). The top ten peaks with high intensities were selected for analysis by secondary ion mass spectrometry. Then, the mass spectrometry data from MALDI-TOF/TOF was analyzed by a GPS Explorer V3.6 software (Applied Biosystems, USA) and a Mascot V2.3 (Matrix Science Ltd., U.K.) with the following search parameters: using annotated carbohydrate-active enzymes (CAZy) **(ESI 2)** as database; trypsin as the digestion enzyme; one missed cleavage site;

fixed modifications of Carbamidomethyl (C); partial modifications of Acetyl (Protein N-term), Deamidated (NQ), Oxidation (M); 100 ppm for precursor ion tolerance and 0.5 Da for fragment ion tolerance. The proteins were successfully identified based on 95% or a higher confidence interval of their scores in the Mascot V2.3 search engine.

Protein analysis was also performed using Liquid Chromatograph Mass Spectrometer (LC-MS/MS, Eksigent nanoLC-Ultra™ 2D System, AB SCIEX, USA). The mass spectrometry data was acquired using a Triple TOF 5600 System (AB SCIEX, USA) fitted with a Nanospray III source (AB SCIEX, USA). The ion spray voltage is 2.5 kV, the curtain gas is 30 PSI, the nebulizer gas is 5PSI, and the interface heater temperature is 150°C. For information-dependent acquisition (IDA), survey scans were acquired in 250 ms, and as many as 35 product ion scans were collected if they exceeded a threshold of 150 counts per second (counts/s) with a 2^+ to 5^+ charge-state. Survey scans were acquired in 250 ms under the information-dependent acquisition mode (IDA, Information Dependent Analysis) with the total cycle time as 2.5 s. The proteins were identified based on 95% or a higher confidence interval of their scores in the Mascot V2.3 based on the acquired MS and MS/MS spectra. The search parameters were as follows: using predicted CAZY library as database (**ESI 2**); the trypsin was the enzyme for digestion; the maximum permissible leakage site is 2; the fixed modifications is Carbamidomethyl (C); the partial modifications were Deamidated (NQ) and Oxidation (M); the fragment ion tolerance is ± 0.6 Da.

7. Enzyme over-expression

The signal peptide sequences of Rhase-TS (first 22 residues) and Gluase-TS (first 19 residues), predicted by SignalP-5.0 Server (http://www.cbs.dtu.dk/services/SignalP/), were deleted. After a 6×His-tag was added in C terminal, the genes of Rhase-TS and Gluase-TS were optimized according to the codon preference of *P. pastoris*, followed by sequence synthesis by Sangon Biotech (Shanghai, China). The synthetic sequences have been submitted into the NCBI database (Rhase-TS: Genebank accession no. **MT779018**; Gluase-TS: Genebank accession no. **MT779019**). The

primers (Rhase-TS-F: 5'-GGGAAAGAATTCTTTCCATCCTATGAGCCA-3', Rhase-TS-R: 5'-GGGAAAGCGGCCGCGTGGTGATGGTGGTGGTG-3'; Gluase-TS-F: 5'-GGGAAAGAATTCTACTCTCCACCAGCTTA-3', Gluase-TS-R: 5'-GGGAAAGCGGCCGCGTGGTGGTGATGGTGGTG-3') were designed for gene amplification. The *Eco*RI and *Not*I restriction sites were introduced into the genes by amplification with the designed primers in a Pro-Flex PCR System (Thermo Scientific, USA). The amplified genes were purified by using DNA Gel Extraction Kit (Omega Biotek, Germany). The recombinant plasmids (*p*PIC9k-Rhase-TS and *p*PIC9k-Gluase-TS) (*p*PIC9k vector: Invitrogen, USA) were constructed via enzyme digestion by *Eco*RI and *Not*I restriction enzyme (Trans Gene Biotech, China) and ligation by T_4 DNA ligase (Trans Gene Biotech, China). After linearized by *Sal*I enzyme (Trans Gene Biotech, China), the recombinant plasmids were electrically transformed into GS115 (Invitrogen, Carlsbad, CA, USA), followed by screening on histidine deficient medium (SD/-His with agar, Coolaber, China). Enzyme expression was performed in buffered methanol-complex medium (BMMY) shaking at 30 °C, 220 rpm. After 5 days of culture, cells were removed by centrifugation at 10000 rpm for 10 min. Enzyme purification was performed as described previously [14]. The obtained recombinant enzymes were analyzed by 10% SDS-PAGE.

8. Enzyme characterization

Different pH buffers were prepared as previously described for enzyme pretreatment and reaction [14]. Enzyme activities at pH ranging from 2.0 to 11.0 were measured at 60 ℃ to investigate the pH effect. Enzymes were pre-incubated in pH ranging from 2.0 to 11.0 for 4 h at 4 °C, and then the residual activities were measured to study the pH stability of Rhase-TS or Gluase-TS. Effect of temperature on enzymes was evaluated by measuring activity in pH 4.5 at temperatures ranging from 10 ℃ to 90 ℃. Moreover, residual enzyme activities of Rhase-TS or Gluase-TS after pretreated at temperatures ranging from 0 to 80 ℃ for 1 h were measured to assess enzyme thermostability. To evaluate the effect of metal ions and chemicals on enyzmes, the

activities were measured in the presence of 1 mM or 10 mM Na^{+}, K^{+}, Ca^{2+}, Fe^{2+}, Fe^{3+}, Cu^{2+}, Mg^{2+}, Mn^{2+}, Zn^{2+}, Co^{2+}, dithiothreitol (DTT), sodium dodecyl sulfate (SDS), ethylene diamine tetraacetic acid (EDTA) and *β*-mercaptoethanol. To better understand the enzyme properties, tolerances of Rhase-TS or Gluase-TS against organic solvents were estimated by measuring the enzyme activity in the presence of 0.1, 0.2, 0.5, 1, 2, 3, 5, 8 ,10, 15, 20, 25, 30 % (v/v) methanol, ethanol or DMSO. Substrate specificity was studied by incubating *p*-nitroglycosides or *o*-nitroglycosides with Rhase-TS or Gluase-TS at pH4.5, 60 ℃.

9. Analytical methods

The thin-layer chromatography (TLC) analysis was performed on a silica gel plate (T225100F, Synthware, China). The $CHCl_3$: CH_3OH: H_2O (60:27:5, v/v) was used as developing solvent, and spots of compounds were appeared by spraying with 20 % (v/v) H_2SO_4, followed by heating at 80 °C for 5 min. The HPLC system was used for the quantification of steroid saponins. The isocratic elution was performed with 10 % water (solvent A) and 90 % methanol (solvent B) by using a C_{18} column (5 μm, 250×4.6 mm, Shimadzu, Japan) at a flow rate of 1 mL/min for the analysis of steroidal saponins with the UV detector at 203 nm. The rhamnose and glucose were analyzed using a Suger-Pak1 Column (10 μm, 300 mm×6.5 mm, Waters, USA) at 80 °C, 0.5 mL/min by a differential refractive index detector. The M_w and molecular formula of compounds dissolved in methanol were identified by mass spectrometry (MS) (Agilent G6500 Series Q-TOF).

Table S1. Primer pairs for sequencing DNA barcodes of *Talaromyces stollii* CLY-6

Region	Primer	Primer Sequence 5'to 3'	Direction	Reference
Internal Transcribed Spacer (ITS)	ITS1	TCCGTAGGTGAACCTGCGG	Forward	3
	ITS4	TCCTCCGCTTATTGATATGC	Reverse	
18S ribosomal DNA (18S rDNA)	NS1	GTAGTCATATGCTTGTCTC	Forward	15
	NS6	GCATCACAGACCTGTTATTGCCTC	Reverse	
β-tubulin	Bt2a	GGTAACCAAATCGGTGCTGCTTTC	Forward	3
	Bt2b	ACCCTCAGTGTAGTGACCCTTGGC	Reverse	
Calmodulin	CMD5	CCGAGTACAAGGAGGCCTTC	Forward	16
	PoCMDR1	CGCCAATCGAGGTCATGACATGG	Reverse	

Table S2. Microbial biotransformation of E-DZW.

Microorganism name	Depositary authority	Deposit number	Diosgenin yield* (%)
Aspergillus carbonarius	CICC[a]	41254	10.2±0.5
Trichoderma viride	CICC	13038	7.8±0.4
Penicillium sclerotiorum	CGMCC[b]	3.5676	3.4±0.3
Trichoderma reesei	CGMCC	3.3711	ND[c]
Trichoderma harzianum	CGMCC	3.6604	1.1±0.2
Penicillium sp.	CGMCC	3.4426	2.1±0.3
Fungus CLY-1	None	None	ND
Fungus CLY-2	None	None	ND
Fungus CLY-3	None	None	ND
Fungus CLY-4	None	None	ND
Fungus CLY-5	None	None	ND
Fungus CLY-6	CGMCC	3.16013	72.5±1.2

[a] CICC, China Center of Industrial Culture Collection.

[b] CGMCC, China General Microbiological Culture Collection Center.

[c] ND, Not detected.

* The diosgenin yield (%) was defined as the ratio of actual diosgenin output to theoretical diosgenin output.

Table S3. The features of raw sequencing data and improved one of *Talaromyces stollii* CLY-6*

	Raw sequencing data		Improved raw sequencing data	
	Number (bp)	**Ratio (%)**	**Number**	**Ratio (%)**
Total Reads Count	7872558	-	7485848	-
Total Bases	2248854514	-	1905578505	-
Average Read Length	285.66	-	254.56	-
Q10 Bases [a]	2202551015	97.94%	1901461575	99.78%
Q20 Bases	2099855535	93.37%	1879588635	98.64%
Q30 Bases	1923236030	85.52%	1790340303	93.95%
N Bases [b]	38544	0.00%	8	0.00%
GC Bases	1067411216	47.46%	901776625	47.32%

* The raw data from sequencing contains low-quality sequences and sequences with connectors; thus, original data must be filtered to get clean data for ensuring the quality of information analysis. The result showed that the number of genome's total bases decreased from 2248854514 to 1905578505, the ratio of Q30 Bases (one out of 1,000 bases will be misidentified) increased from 85.52% to 93.95%, and the number of unknown bases (N Bases) decreased from 38544 to 8 by optimization, indicating the high quality of the data for subsequent genome assembling.

[a] Q(n) Bases, the bases with Phred quality score above n.

[b] N Bases, unknown bases.

Table S4. Information of assembled genome sequence of *Talaromyces stollii* CLY-6*

Nuclear genome	Number
Total length (bp)	34462971
N Bases [a] (bp)	890
Max length (bp)	1329510
Average length (bp)	142409
Contig number (bp)	242
GC Ratio (%)	46

[a] N Bases, unknown bases.

* The assembled genome consists of 34462971 bp containing 242 contigs with a GC ratio (%) of 0.46.

Table S5. Overview of predicted genes in *Talaromyces stollii* CLY-6*

Predicted gene	Counts
All genes number	12428
Length >=500 bp	11303
Length >=1000 bp	8091
Max length (bp)	23691
Min length (bp)	120
All genes length (bp)	18513138
Mean length (bp)	1489.63

* Results showed that a total of 12428 genes were predicted, and coding regions occupy almost 53.72% of the genome with a mean length of 1489.63 bp.

Table S6. Overview of annotated genes in *Talaromyces stollii* CLY-6

Database	Number of Unigenes	Percentage (%) [a]
Annotated in CDD	9242	74.36
Annotated in KOG	6511	52.39
Annotated in NR	12256	98.62
Annotated in PFAM	8687	69.9
Annotated in Swissprot	8451	68
Annotated in TrEMBL	12212	98.26
Annotated in GO	8715	70.12
Annotated in KEGG	3975	31.98
Annotated in at least one database	12259	98.64
Annotated in all database	3156	25.39
Total Unigenes	12428	100

[a] The ratio of the number of genes annotated in the database to the total number of genes (12428)

Table S7. Purification of Rhase-TS from *Talaromyces stollii* CLY-6.

Purification steps	Total protein (mg)	Total activity (U)[a]	Specific activity (U/mg)[b]	Yield (%)	Purification (fold)
Crude extract	1545.2	1882.8	1.2	100	1
Ammonium sulfate precipitation	1056.5	1249.4	1.2	66.4	1
Q-sepharose FF	10.3	835.6	81.1	44.4	67.6
DEAE-sepharose FF	5.8	583.8	100.7	31	83.9
Suerdex™ 200 Increase 10/300 GL	1.2	167.5	139.6	8.9	116.3

[a] One unit (U) of enzyme activity was defined as the amount of protein to convert 1 μmol S_3 per 1 hour.

[b] Specific activity was described as the activities per mg of protein.

Table S8. Purification of Gluase-TS from *Talaromyces stollii* CLY-6.

Purification steps	Total protein (mg)	Total activity (U)[a]	Specific activity (U/mg)[b]	Yield (%)	Purification (fold)
Crude extract	1545.2	46.4	0.03	100	1
Ammonium sulfate precipitation	1056.5	32.5	0.03	70.04	1
Q-sepharose FF	40.3	13.2	0.33	28.45	11
DEAE-sepharose FF	10.5	5.7	0.54	12.28	18
Suerdex™ 200 Increase 10/300 GL	1.8	2.2	1.22	4.74	40.67

[a] One unit (U) of enzyme activity was defined as the amount of protein to convert 1 μmol S_6 per 1 hour.

[b] Specific activity was described as the activities per mg of protein.

Table S9. Matched peptides of protein 10806 and 3425 from *Talaromyces stollii* CLY-6 by MALDI-TOF/TOF analysis.

Protein	Observed mass	Theoretical mass	Calculated mass	ppm	Score	Peptide
	1866.0626	1865.0553	1864.9418	60.9	134	K. LSSDGTTPSILVLDYER. N
	1396.8051	1395.7978	1395.7147	59.6	74	R. NVEGYATFQVIR. S
10086	1343.7961	1342.7888	1342.7033	63.7	56	K. FNLVYAYALNR. W
	1235.7146	1234.7073	1234.6346	58.9	63	K. ISPYASGYHLK. A
	911.5198	910.5126	910.4661	51	29	K. GGQVFTFR. Q
	982.5625	981.5552	981.4880	68.5	67	K. TVYVGNSSR. N
	999.5364	998.5292	998.4603	68.9	68	R. GVAMAEEHR. G
	1109.6451	1108.6378	1108.5625	67.9	76	K. EINQHVDVR. G
	1328.7667	1327.7594	1327.6732	65.0	97	K. NVDGALPLTGSER. F
3425	1434.8944	1433.8871	1433.7991	61.4	110	K. GVDVQLGPVAGPLGR. A
	1478.7800	1477.7727	1477.6738	66.9	80	K. HYIGYEQEHFR. Q
	1620.0240	1619.0167	1618.9155	62.5	73	R. GKGVDVQLGPVAGPLGR. A
	1834.0537	1833.0464	1832.9268	65.2	173	R. LSLGAGEEAQWTTTLTR. R
	2097.2554	2096.2481	2096.1154	63.3	155	K. VAGIEVPQLYVSLGGPSDAPK. V

Table S10. Matched peptides of protein 10806 from *Talaromyces stollii* CLY-6 by LC-MS/MS analysis.

Start	End	Observed mass	Theoretical mass	Calculated mass	ppm	Score	Peptide
49	77	1081.8956	3242.665	3242.6459	6	64	K. LSSDGTTPSILVLDYERNVEGYATFQVIR. S
94	115	1207.5785	2413.1424	2413.1293	5	85	R. ALLDNYMADGPLPLAAAMDTYR. I
182	188	379.7196	757.4247	757.4235	2	56	R. IWGVGAR. T
355	387	903.4389	3609.7264	3609.7733	-13	1	K. MYSSPLTNSSALQDFLLGTNPLPVSVDGSRRDR. I
434	439	362.7108	723.4069	723.4028	6	48	K. VQQPPR. T
523	533	672.3599	1342.705	1342.7033	1	63	K. FNLVYAYALNR. W
616	633	862.4261	1722.8376	1722.8325	3	38	R. AGALAFSNASSASGWAQK. I
634	644	618.3251	1234.6356	1234.6346	1	46	K. ISPYASGYHLK. A
645	657	459.2249	1374.6529	1374.6528	0	24	K. AAFHANDSTNAKK. L
711	738	1079.2813	3234.8222	3234.807	5	29	R. YVLGIQPLVPGFKQWQIVPQTLGLQWAR. G
739	745	375.2071	748.3997	748.398	2	24	R. GSHPVPR. G
786	814	806.4184	3221.6443	3221.6292	5	31	K. GYQVSGCTVVNNSTFLVKGGQVFTFRQTK.

Table S11. Effect of metal ions and reagents on Rhase-TS and Gluase-TS.

Metal ions or reagents	Relative activity*±SD[a] (%) at:			
	1 mM		10 mM	
	Rhase-TS	Gluase-TS	Rhase-TS	Gluase-TS
Control	100±3.24	100±4.52	100±2.17	100±4.29
Na^{+}	96.56±2.45	98.53±3.19	94.76±2.85	98.28±2.18
K^{+}	98.45±3.12	97.51±2.98	102.34±3.96	101.71±3.29
Ca^{2+}	103.27±2.98	101.51±4.21	100.18±3.21	102.57±4.15
Fe^{2+}	100.04±2.79	98.52±3.28	95.34±3.19	90.57±3.87
Fe^{3+}	100.58±3.20	101.23±3.49	103.92±3.42	96.28±3.63
Mg^{2+}	107.34±2.59	120.51±4.22	100.32±3.26	123.71±3.29
Mn^{2+}	101.56±3.78	149.25±3.29	94.65±4.18	144.85±4.12
Zn^{2+}	100.72±4.56	130.35±3.53	97.93±2.18	127.71±4.17
Cu^{2+}	97.54±5.23	86.57±4.21	98.82±5.42	65.71±3.21
Co^{2+}	96.56±4.26	101.34±3.98	94.76±4.87	102.28±3.93
SDS	98.54±3.45	96.56±4.18	42.56±4.72	68.28±2.97
DTT	94.92±3.16	96.82±3.29	98.25±3.21	101.85±4.27
EDTA	106.34±5.76	102.58±3.29	97.43±3.19	97.54±4.21
β-Mercaptoethanol	101.17±4.75	99.28±4.16	98.29±2.96	100.6±3.19

* Relative activity was described as the percentage of control activity.

[a] SD, Standard deviation.

Table S12. Substrate specificity of Rhase-TS and Gluase-TS.

Substrate	Relative activity of Rhase-TS±SD[a] (%)	Relative activity [c] of Gluase-TS±SD (%)
p-Nitrophenyl-*β*-D-galactopyranoside	ND[b]	ND
o-Nitrophenyl-*β*-D-galactopyranoside	ND	ND
p-Nitrophenyl-*α*-D-galactopyranoside	ND	ND
o-Nitrophenyl-*β*-D-xylopyranoside	ND	95.54±3.43
p-Nitrophenyl-*β*-D-xylopyranoside	ND	74.32±4.12
p-Nitrophenyl-*α*-L-arabinopyranoside	ND	ND
p-Nitrophenyl-*α*-D-mannopyranoside	ND	ND
p-Nitrophenyl-*β*-D-glucopyranoside	ND	ND
p-Nitrophenyl-*α*-D-glucopyranoside	ND	ND
o-Nitrophenyl-*β*-D-glucopyranoside	ND	100
p-Nitrophenyl-*α*-L-rhamnopyranoside	ND	ND

[a] SD, Standard Deviation.

[b] ND, Not detected.

[c] Activity of Gluase-TS against *o*-Nitrophenyl-*β*-D-glucopyranoside was defined as 100%.

Table S13. The typical ^{1}H chemical shifts (δ in ppm) of final product framework*

Chemical shifts	Peak type and H number
0.81	s, 6H
0.81	d, 3H
0.99	d, 3H
1.05	s, 3H
1.98-2.04	m, 2H
2.22-2.35	m, 4H
3.40	t, 1H
3.48-3.60	m, 2H
4.43	q, 1H
5.37	d, 1H

***Note:** ^{1}H NMR of the final product was carried out with the conditions of 600 MHz; s: single peak; d: double peaks; m: multiple peaks; t: triple peaks; q: quaternary peaks. The chemical structure of the final product was identified as diosgenin by comparison with those reported in literatures[17-19].

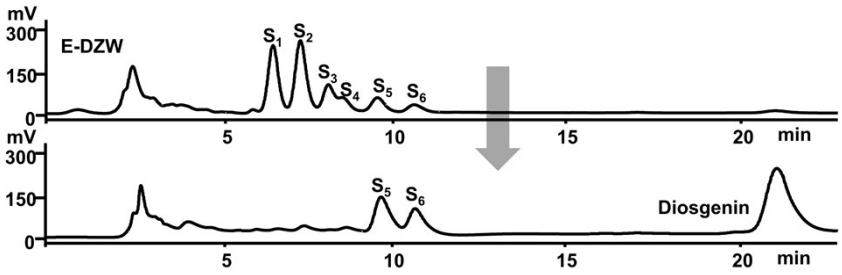


Figure S1. HPLC analysis of microbial transformation of E-DZW by *Talaromyces stollii* CLY-6; results showed that substrate S_1, S_2, S_3, and S_4 in E-DZW were converted with final product as a mixture of S_5, S_6, and diosgenin (yield: 72.5%) within 72 h. Diosgenin yield was defined as the percentage of theoretical yield.

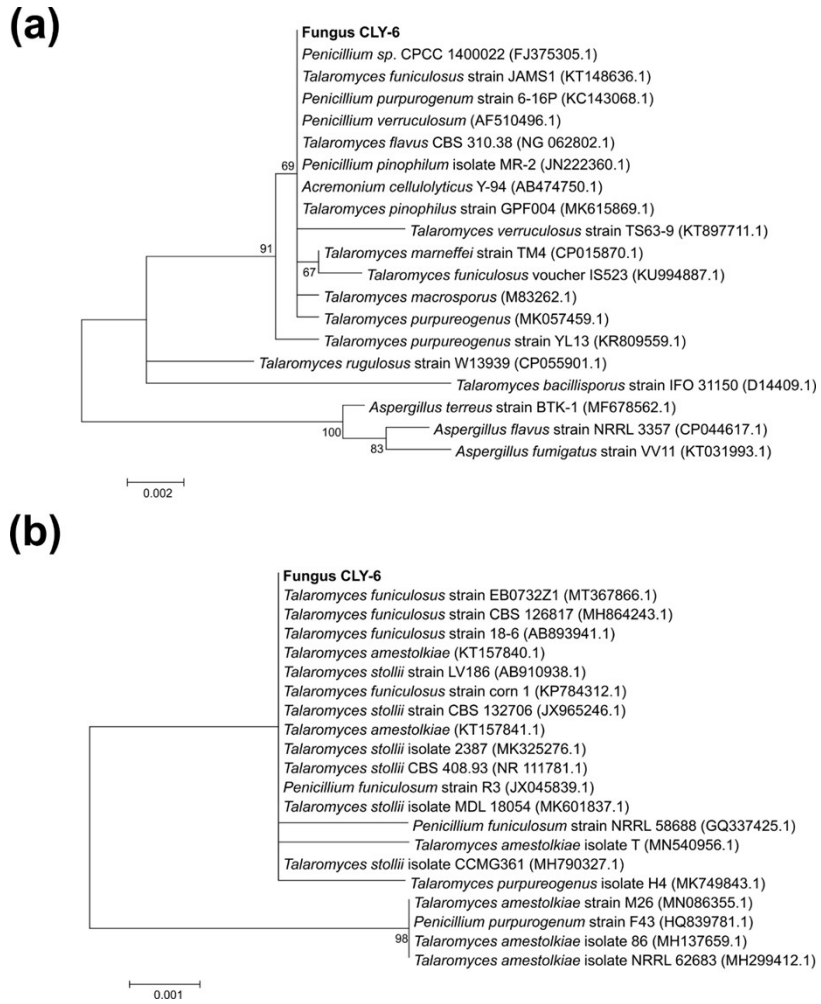


Figure S2. The Maximum-likelihood (ML) trees of **(a)** 18S ribosomal DNA and **(b)** Internal Transcribed Spacer constructed by using MEGA software (version 6.0.6) with Kimura 2-parameter model for the calculation of evolutionary distances. Bootstrap values (1000 replicate runs) greater than 50% are listed. Results showed that *Talaromyces* and *Penicillium* species were grouped in a single cluster, and Fungus CLY-6 was well-placed together with *Talaromyces funiculosus*, *Talaromyces pinophilus*, *Talaromyces amestolkiae*, *Talaromyces stollii*, and *Penicillium purpurogenum* etc.

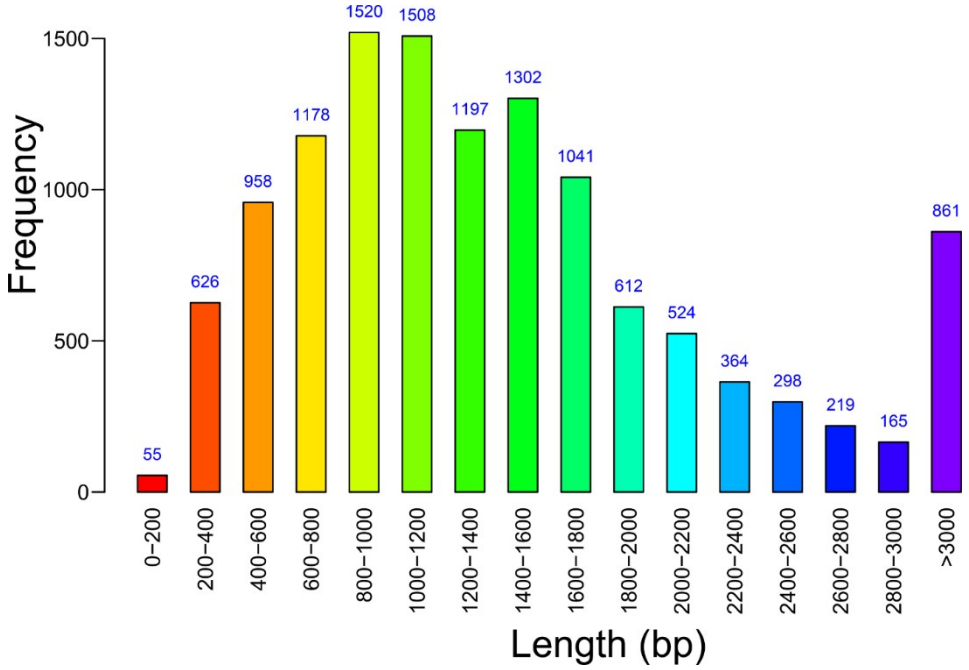


Figure S3. Gene length distribution. The horizontal axis is the gene length interval, and the vertical axis is the number of genes in the corresponding interval. Results showed that gene length distributed mainly in the range of 200~2200 bp.

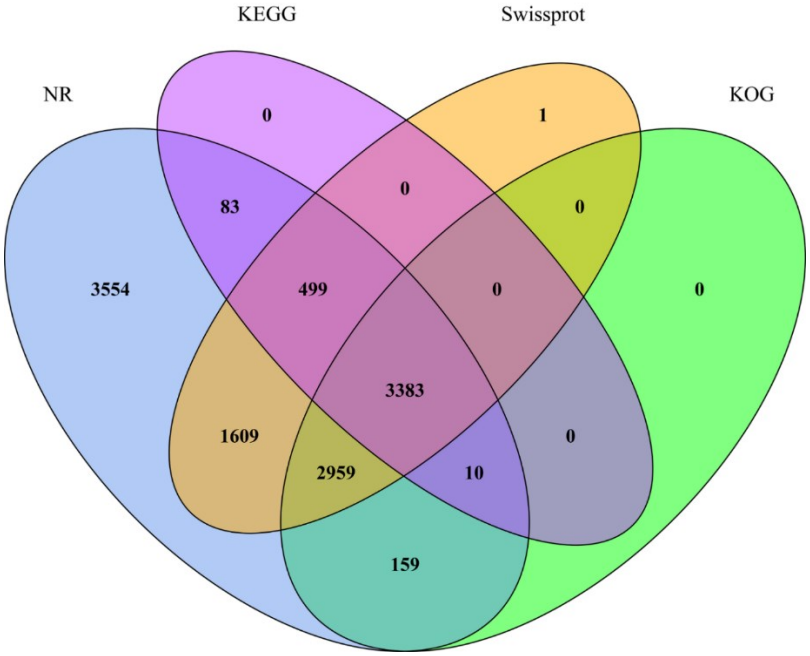


Figure S4. Venn diagram of annotated genes in databases. Green, yellow, violet and blue area represent the number of annotated genes from *Talaromyces stollii* CLY-6 in KOG, Swissprot, KEGG, and NR databases, respectively. Overlapping area represents the number of shared annotated-genes in different databases.

Figure S5. Amino acid sequence of **(a)** Rhase-TS (Genebank accession no. **MT779018**) and **(b)** Gluase-TS (Genebank accession no. **MT779019**). The gene of Rhase-TS consists of 2442 bp, encoding 814 amino acids with the pI value predicted as 4.91 by ProtParam [20], while the gene of Gluase-TS consists of 2568 bp, encoding 856 amino acid with the pI value calculated as 4.79. The signal peptides of Rhase-TS (first 22 residues) and Gluase-TS (first 19 residues) were predicted by SignalP-5.0 Server (http://www.cbs.dtu.dk/ services/SignalP/). The predicted N-glycosylation sites by NetNGlyc 1.0 Server (http://www.cbs.dtu.dk/-services/NetNGlyc/) were marked by asterisks. Matched peptides of Rhase-TS and Gluase-TS in protein identification were shown in red color.

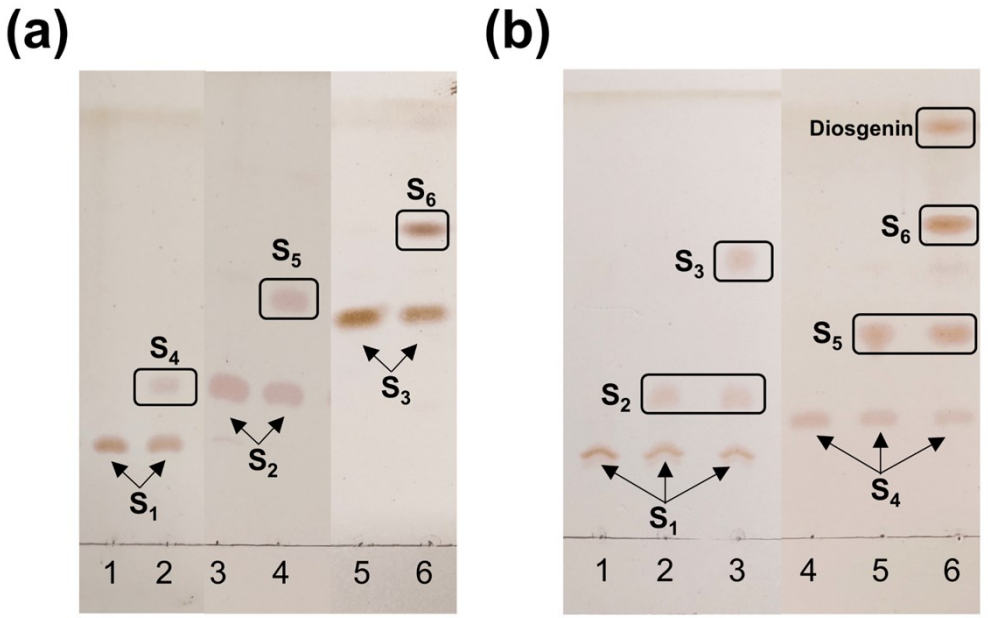


Figure S6. TLC chromatogram. **(a)**, lane 1: substrate S_1; lane 2: hydrolysate (S_4) from S_1 by Rhase-TS; lane 3: substrate S_2; lane 4: hydrolysate (S_5) from S_2 by Rhase-TS; lane 5: substrate S_3; lane 6: hydrolysate (S_6) from S_3 by Rhase-TS; **(b),** lane 1: substrate S_1; lane 2: hydrolysate (S_2) from S_1 by Gluase-TS at 24 h; lane 3: hydrolysates (S_2 and S_3) from S_1 by Gluase-TS at 24 h; lane 4: substrate S_4; lane 5: hydrolysate (S_5) from S_4 by Gluase-TS at 24 h; lane 6: hydrolysates (S_5, S_6, and Diosgenin) from S_4 by Gluase-TS at 48 h.

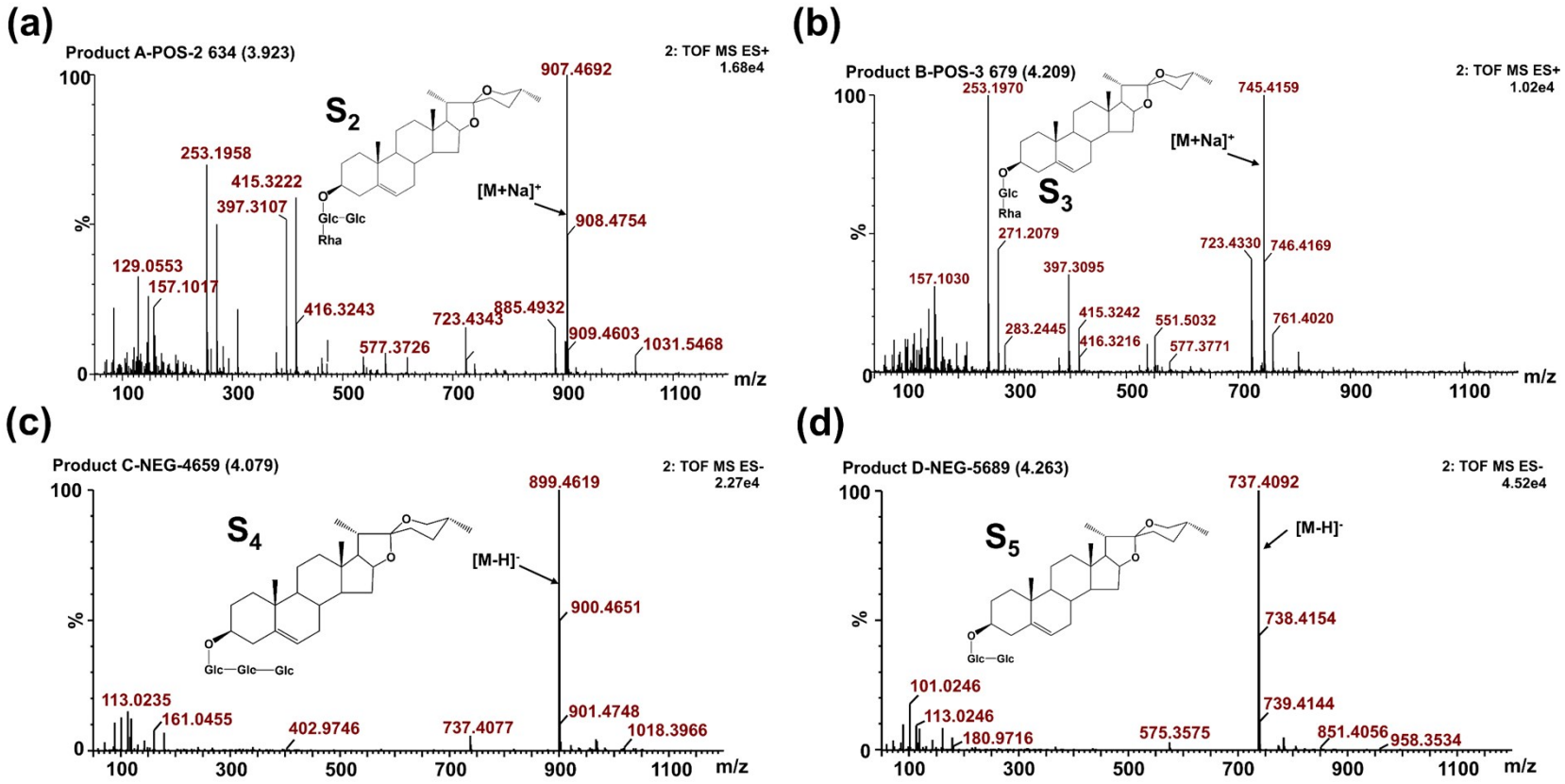


Figure S7. Secondary mass spectra of products from steroidal saponins. **(a),** Product A from S_1 by Gluase-TS, *m/z* 907.47=$[M_w+Na]^+$, M_w=885.04. Product A was identified as deltonin (S_2). **(b)**, Product B from S_1 by Gluase-TS, *m/z* 745.41=$[M_w+Na]^+$, M_w=722.90. Product B was identified as prosapogenin A (S_3). **(c),** Product C from S_1 by Rhase-TS, *m/z* 899.46=$[M_w-H]^-$, M_w=901.04. Product C was identified as diosgenin-triglucoside (S_4). **(d)**, Product D from S_2 by Rhase-TS, *m/z* 737.41=$[M_w-H]^-$, M_w=738.88. The product D was identified as diosgenin-diglucoside (S_5).

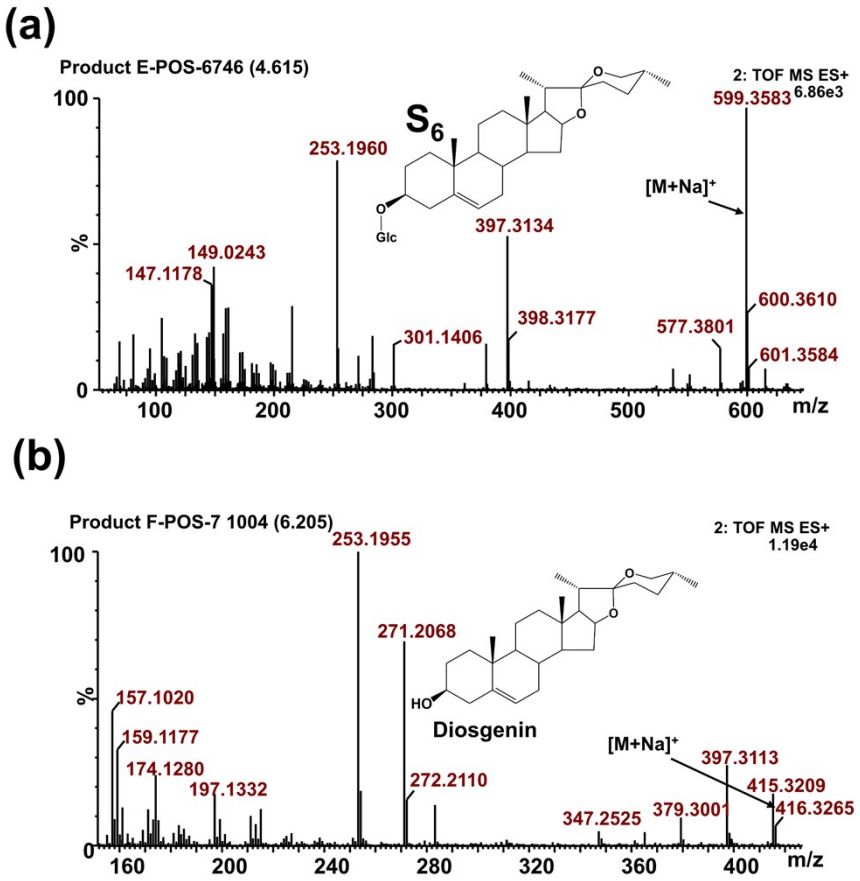
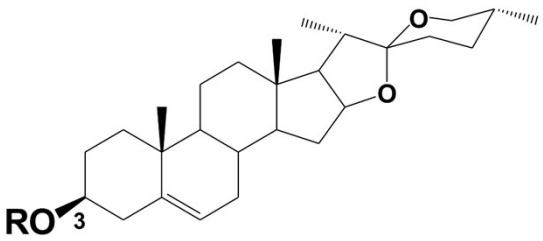


Figure S8. Secondary mass spectra of products from steroidal saponins. **(a),** Product E from S_3 by Rhase-TS, *m*/*z* 599.36=[M_w+Na]$^+$, M_w 576.76. Product E was identified as trillin (S_6). **(b)**, Product F from E-DZW by Rhase-TS and Cel-TL4, *m*/*z* 415.32=[M_w+H]$^+$, M_w=414.62. Product F was identified as diosgenin.



R (C-3)	Name
—Glc(1,4)—Rha; Glc(1,2)—Rha	Dioscin
—Glc(1,4)—Rha	Progenin II
—Glc(1,4)—Ara; Glc(1,2)—Rha	Poliphyllin I
—Glc(1,4)—Ara	Compound A
—H	Diosgenin

Figure S9. Chemical structures of dioscin, progenin II, polyphyllin I, compound A (diosgenin-3-O-*α*-L-arabinofuranosyl-(1→4)-*β*-D-glucopyranoside), and their aglycone (diosgenin).

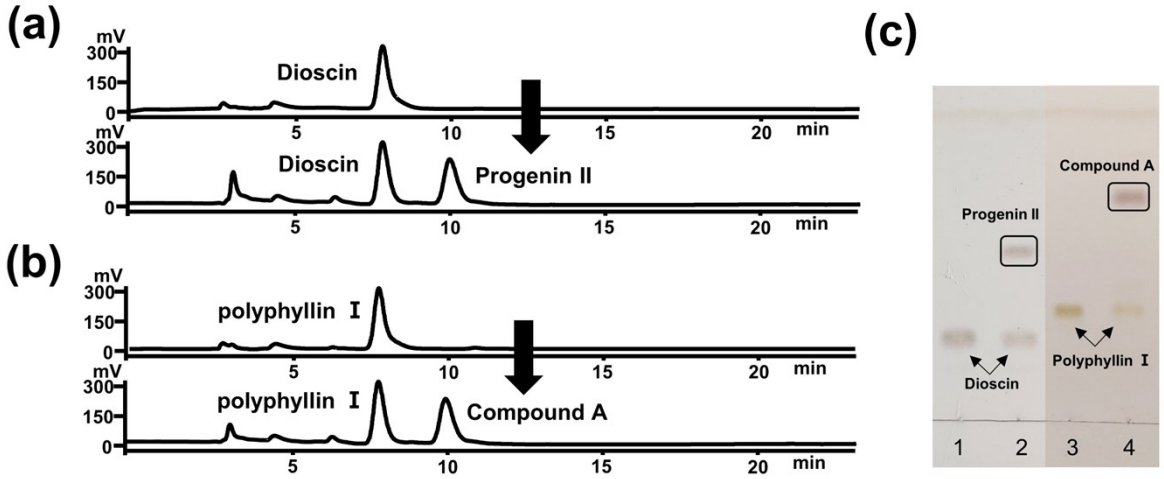


Figure S10. (a), HPLC chromatogram of dioscin hydrolyzation by Rhase-TS; **(b)**, HPLC chromatogram of dioscin polyphyllin Ⅰ hydrolyzation by Rhase-TS; reactions were carried out by incubating recombinant Rhase-TS (0.35 mg/mL, 10 μL, specific activity: 141.3 U/mg against S_3) with substrate (2 mM) in 2 mL 50 mM sodium phosphate buffer (pH 4.5) at 60 ℃ for 1 h. Compound A is diosgenin-3-O-*α*-L-arabinofuranosyl-(1→4)-*β*-D-glucopyranoside. Rhase-TS could hydrolyze all substrates with the terminal 1,2-rhamnose released. **(c),** TLC chromatogram. lane 1: substrate dioscin; lane 2: hydrolysate (progenin II) from dioscin by Rhase-TS; lane 3: substrate polyphyllin Ⅰ; lane 4: hydrolysate (compound A, diosgenin-3-O-*α*-L-arabinofuranosyl-(1→4)-*β*-D-glucopyranoside) from polyphyllin Ⅰ by Rhase-TS.

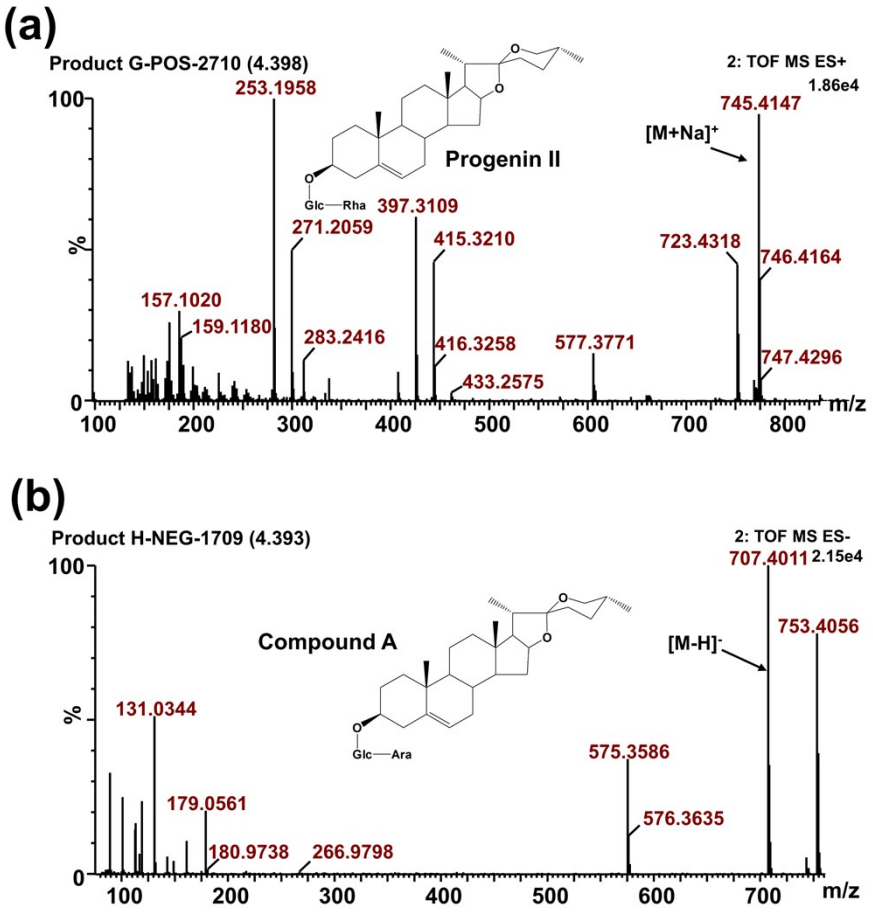


Figure S11. Secondary mass spectra of products from steroidal saponins. **(a)**, Product G from dioscin by Rhase-TS, *m*/*z* 745.41=[M_w+Na]$^+$, M_w=722.9. Product G was identified as progenin II. **(b)**, Product H from polyphyllin Ⅰ by Rhase-TS, *m*/*z* 707.40=[M_w-H]$^-$, M_w=708.87. Product H was identified as diosgenin-3-O-*α*-L-arabinofuranosyl-(1→4)-*β*-D-glucopyranoside (Compound A).

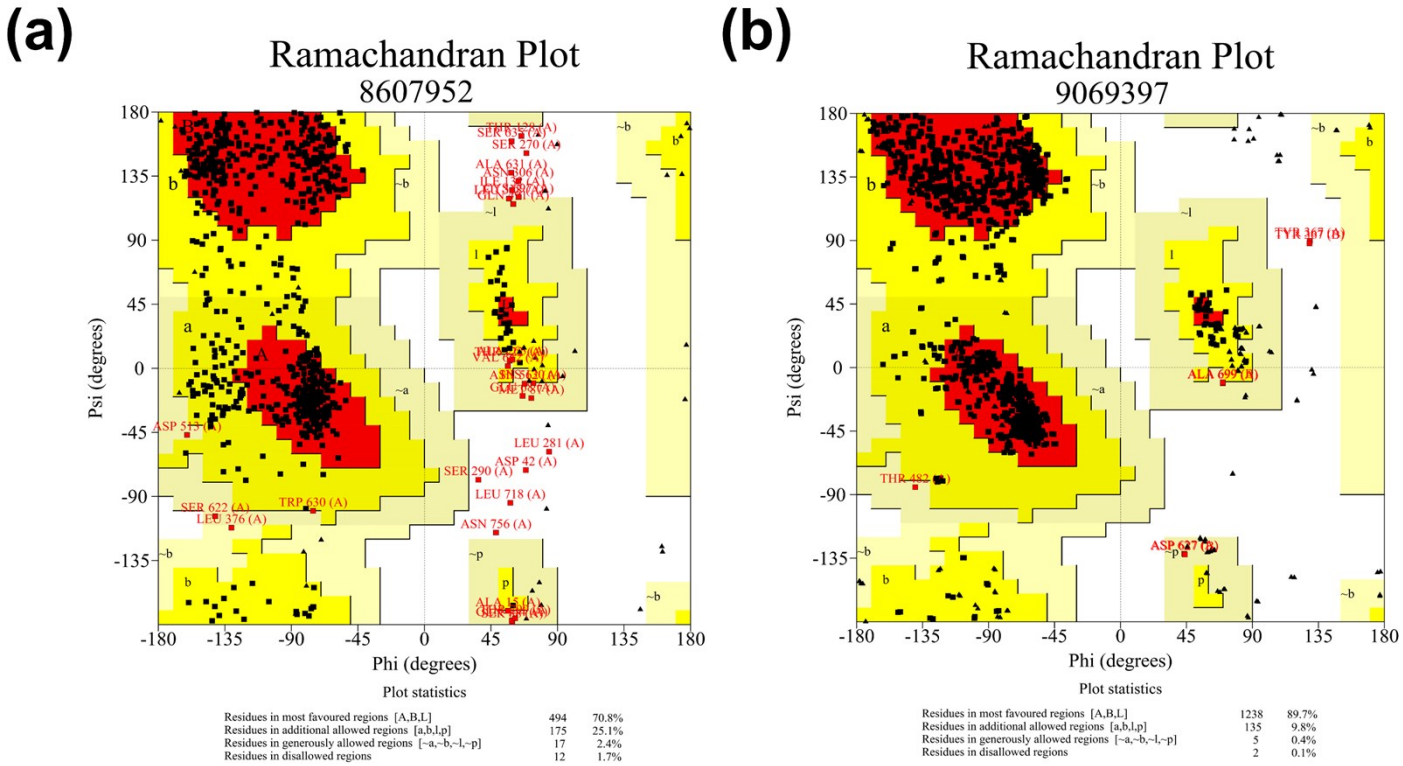


Figure S12. Ramachandran's plots of Rhase-TS and Gluase-TS models by PROCHECK [21]; the residue ratios in allowed regions of Rhase-TS and Gluase-TS models were 98.3% and 99.9%, respectively.

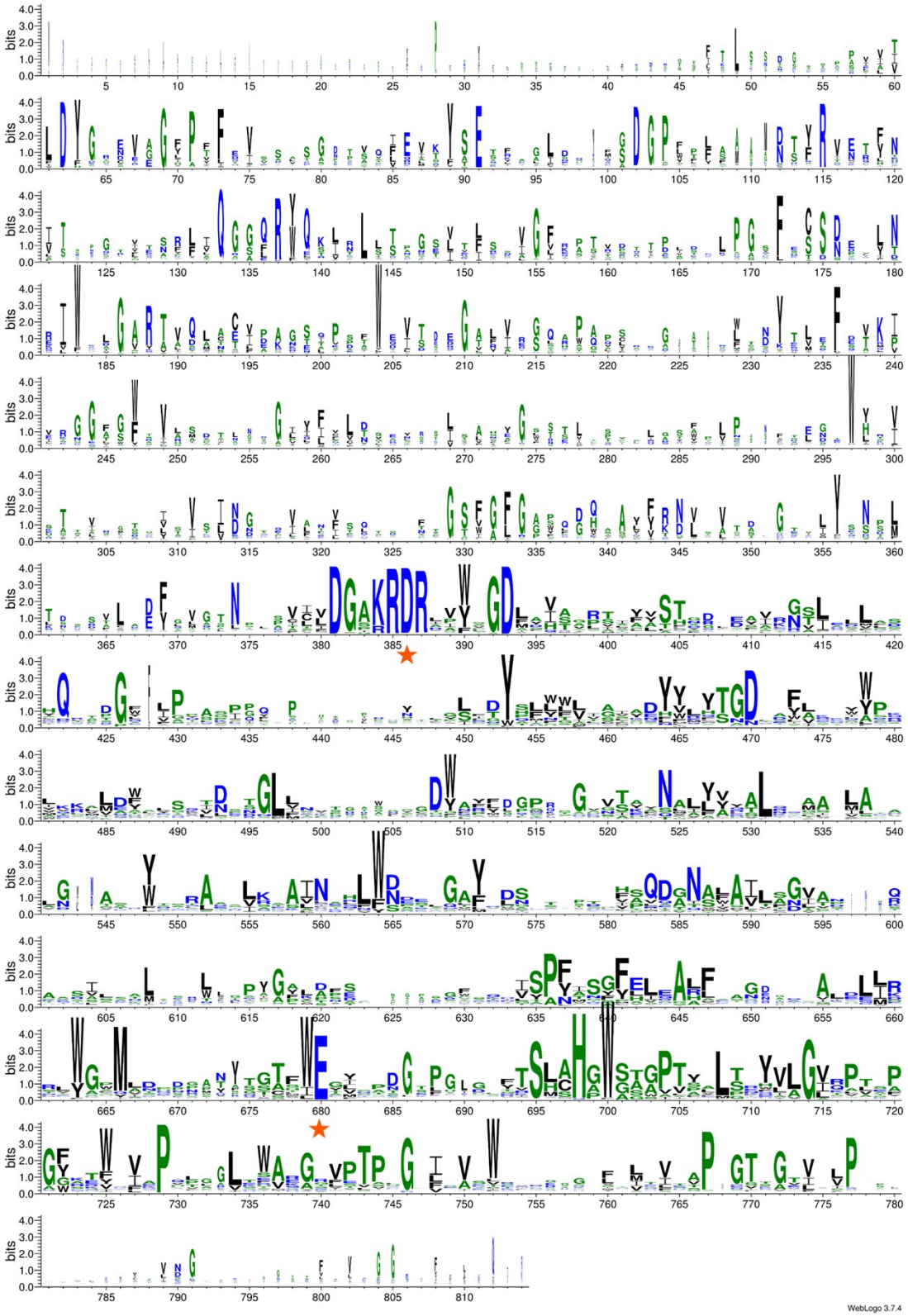


Figure S13. Sequence logo by WebLogo 3 using the homologous sequences of Rhase-TS obtained by Consensus Finder[22]. The putative catalytic residues, Asp 386 (nucleophile) and Glu 680 (acid/base), were marked as red asterisks.

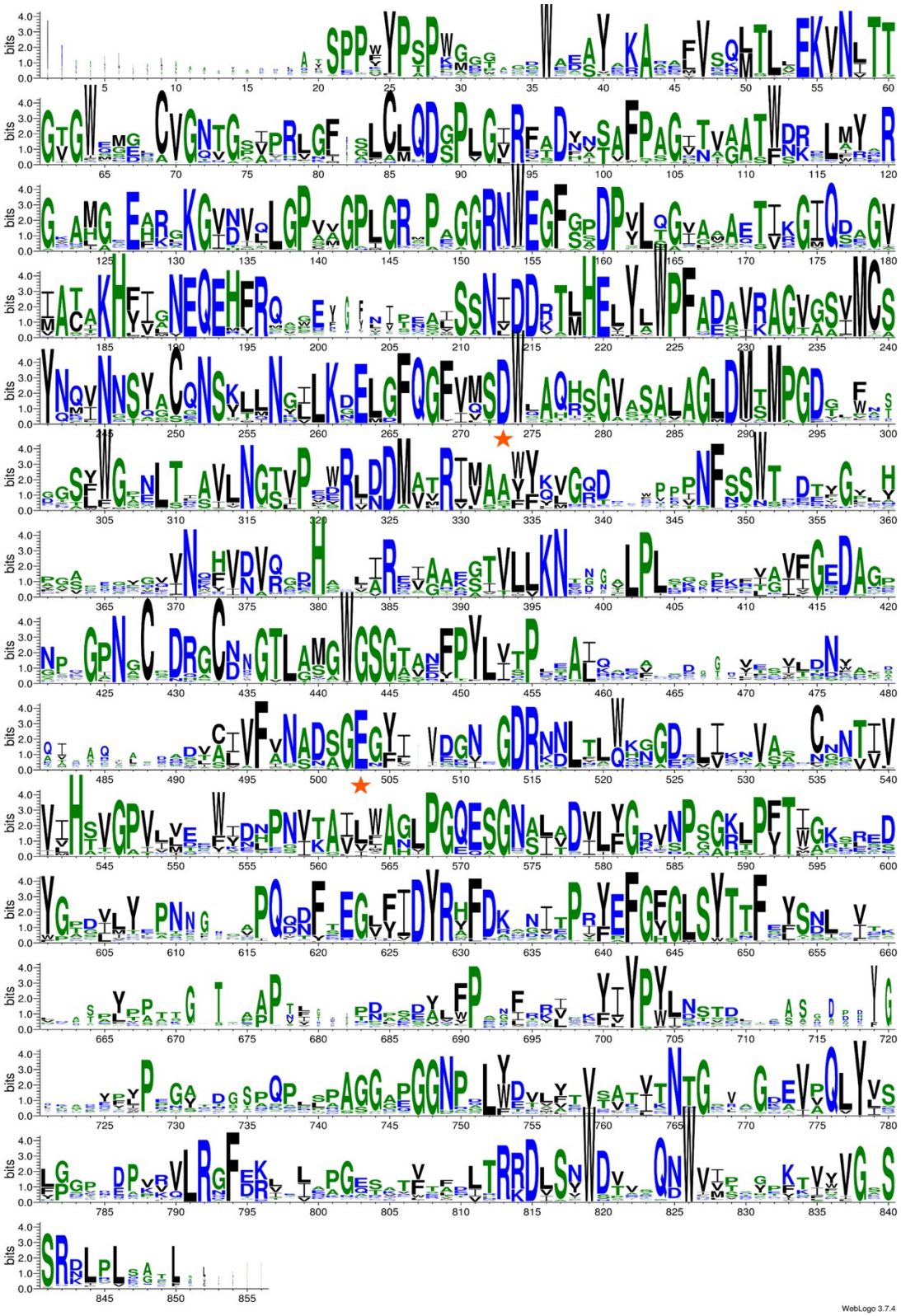


Figure S14. Sequence logo by WebLogo 3 using the homologous sequences of Gluase-TS obtained by Consensus Finder[22]. The putative catalytic residues, Asp273 (nucleophile) and Glu 503 (acid/base), were marked as red asterisks.

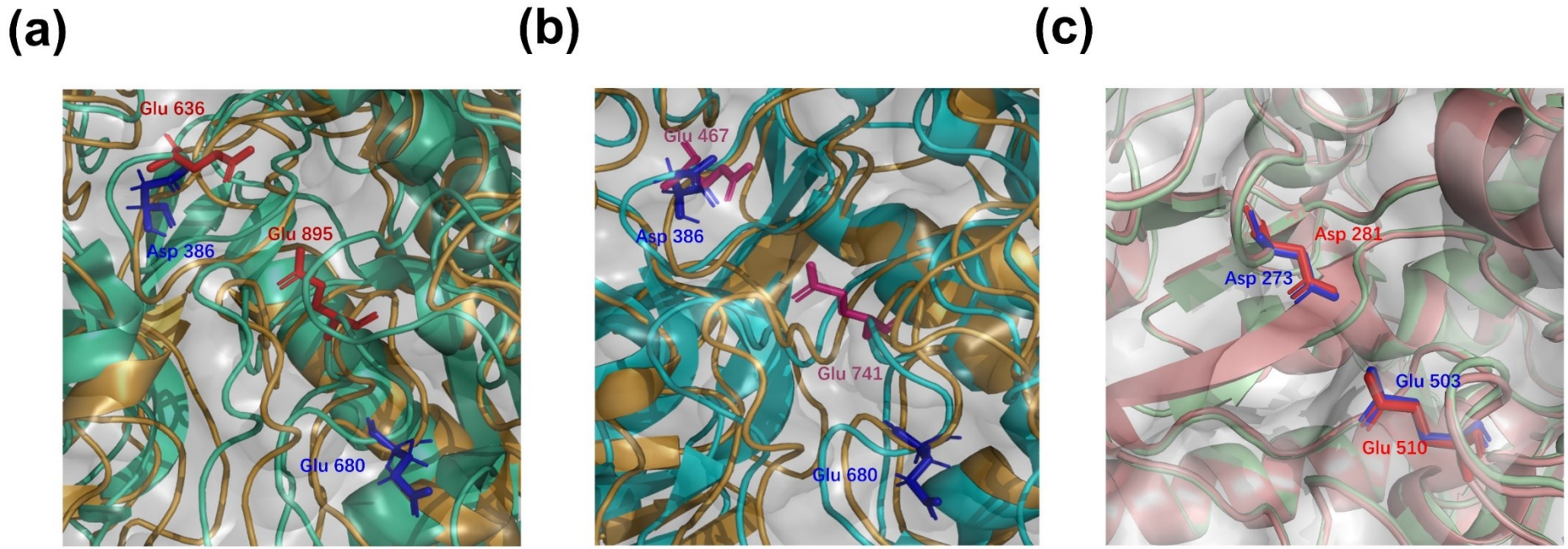


Figure S15. Structure superpositions of Rhase-TS and Gluase-TS models with templates in PyMol. **(a),** Rhase-TS model and template (PDB ID:3w5m, 1.8 Å, Chain A) [23] showed a close structural similarity (TM-score: 0.893) [24] and were colored in yellow and green cyan, respectively. The catalytic residues of template (Glu636, Glu895) were displayed in red, and the putative catalytic residues of Rhase-TS (**Asp386, Glu680**) were displayed in blue. **(b),** Rhase-TS model and template (PDB ID:6gsz, 1.38 Å, Chain A) [25] showed a close structural similarity (TM-score: 0.903) and were colored in yellow and cyan, respectively. The catalytic residues of template (Glu467, Glu741) were displayed in pink, and the putative catalytic residues of Rhase-TS (**Asp386, Glu680**) were displayed in blue. **(c),** Gluase-TS model and template (PDB ID: 5FJI, 1.95 Å, Chain A) [26] were colored in green and salmon. The catalytic residues of template (Asp281, Glu510) were displayed in red, and the putative catalytic residues of Gluase-TS (**Asp273, Glu503**) were displayed in blue.

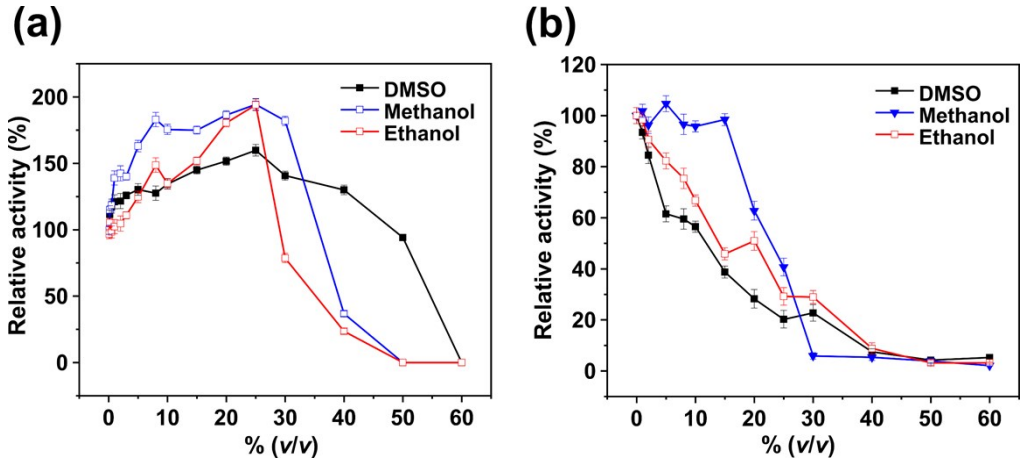


Figure S16. Effect of organic solvents (methanol, ethanol, and DMSO) on **(a)** Rhase-TS and **(b)** Gluase-TS. Relative activity was defined as the percentage of control. The improved activity of Rhase-TS at the concentration of organic solvents lower than 25% (v/v) (methanol, ethanol, and DMSO) might be attributed to the improvement of hydrophobic substrate's solubility, leading to the increase of matter-transferring frequency between substrate and enzyme. Gluase-TS could be more stable in methanol than in DMSO or ethanol at a concentration of lower than 25% (v/v).

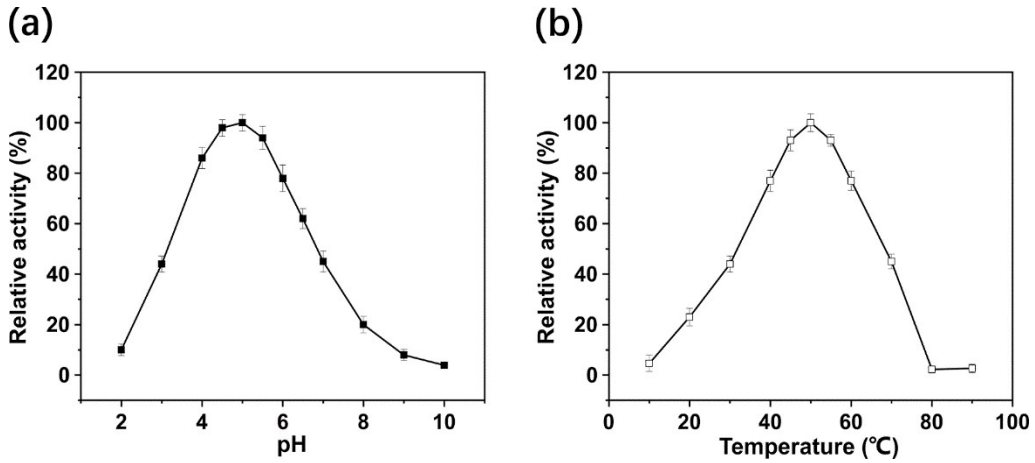


Figure S17. (a), Effect of temperature on Cel-TL4 (optimum temperature: 50 ℃). **(b),** Effect of pH on Cel-TL4 (optimum pH: 5.0). Activity assay of Cel-TL4 was measured by using trillin (S_6) as substrate. One unit (U) was defined as the amount of enzyme required to generate 1 μmol diosgenin per hour. Relative activity was defined as the percentage of maximum activity of Cel-TL4. All assays were performed in triplicate.

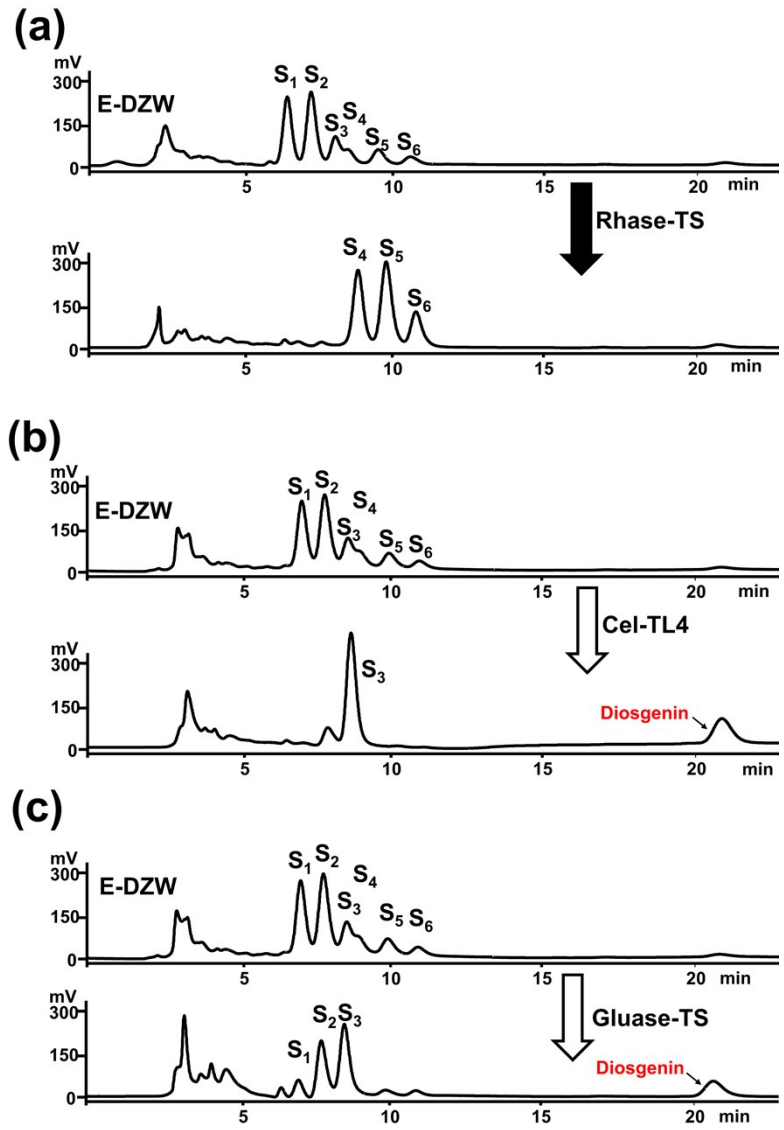


Figure S18. (a), HPLC chromatogram of E-DZW (10 mg) hydrolyzation by Rhase-TS (10 uL, specific activity: 141.3 U/mg against S_3) at pH 4.5, 60 ℃. Substrate S_1, S_2, and S_3 in E-DZW were completely hydrolyzed into S_4, S_5, and S_6, respectively, within 12 h. **(b)**, HPLC chromatogram of E-DZW (10 mg) hydrolyzation by Cel-TL4 (0.2 mg, specific activity: 10.5 U/mg against S_6) at pH 4.5, 50 ℃ for 24 h. SubstrateS_1 and S_2 in E-DZW were almost hydrolyzed into S_3 (yield: 92.4%). Substrate S_4, S_5, and S_6 in E-DZW were completely hydrolyzed into diosgenin (yield: 15.8%). **(c)**, HPLC chromatogram of E-DZW (10 mg) hydrolyzation by Gluase-TS (150 uL, 1.35 mg/mL, specific activity: 1.4 U/mg against S_6) at pH 4.5, 60 ℃ for 24 h with the yield of S_3 and diosgenin as 21.5% and 14.1%, respectively.

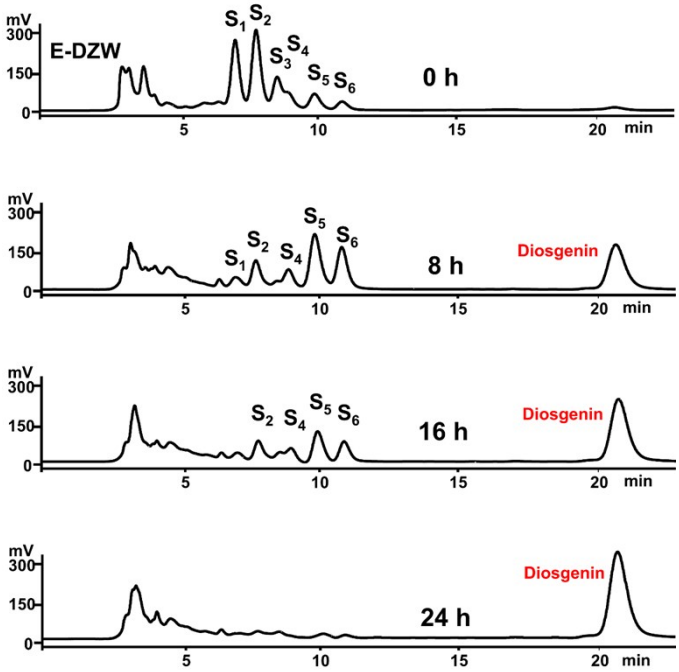


Figure S19. HPLC chromatogram of E-DZW (120 mg/mL) hydrolyzation by Rhase-TS and Cel-TL4 at pH 5.0, 52 ℃. Steroid saponins in E-DZW were almost hydrolyzed into diosgenin within 24 h with a yield of 96.5%.

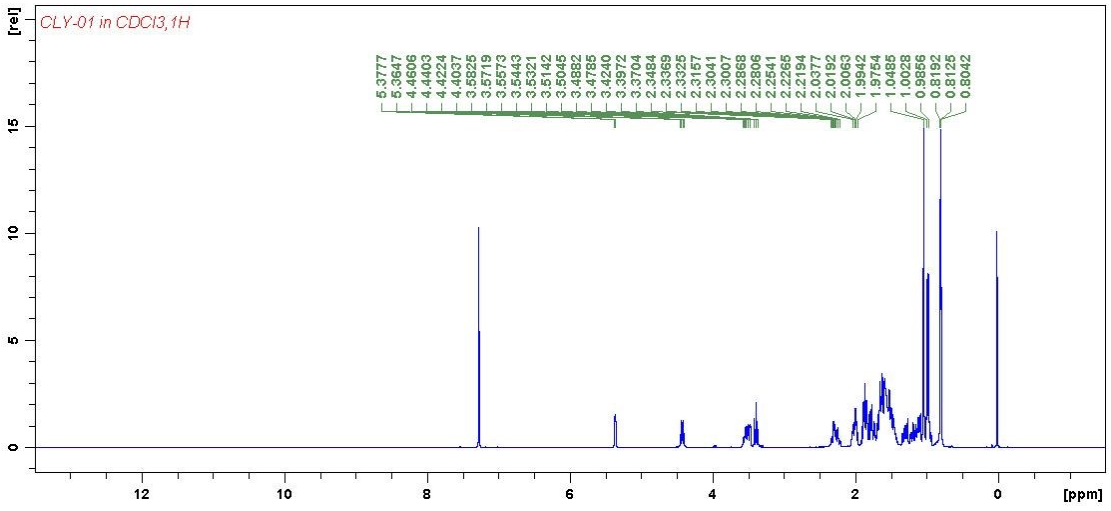
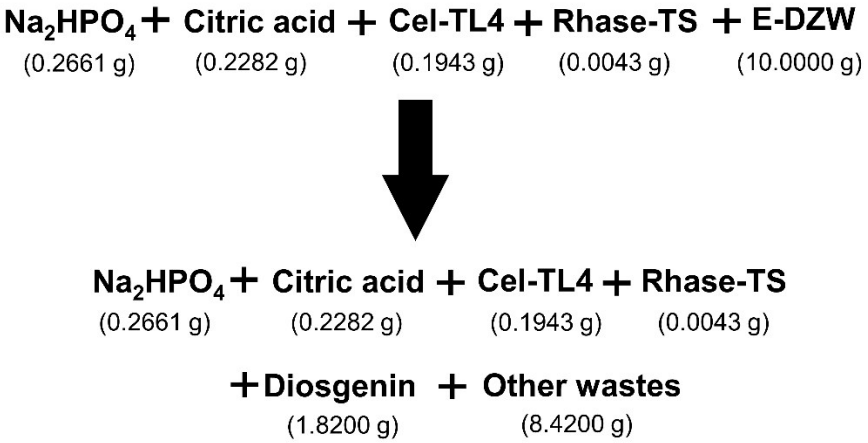
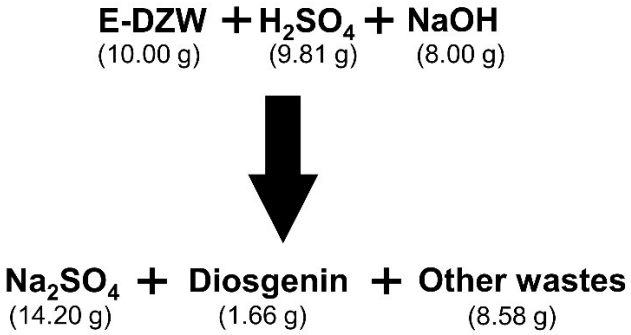


Figure S20. ^{1}H NMR spectrum of the final product (diosgenin).



Total amount of reactants: 0.2661 g+0.2282 g+0.1943 g+0.0043 g+10.0000 g=10.6929 g
Amount of final product (diosgenin): 1.82 g
Amount of waste: 0.2661 g+0.2282 g+0.1943 g+0.0043 g+8.4200 g=9.1129 g
E-Factor=Amount of waste/Amount of product=9.1129/1.8200=5.007

Figure S21. E-factor calculation for the production of diosgenin by enzymatic catalysis.



Total amount of reactants: 10.00 g+9.81 g+8.00 g=27.81 g
Amount of final product (diosgenin): 1.66 g
Amount of waste: 14.2 g+8.58 g=22.78 g
E-Factor=Amount of waste/Amount of product=22.78/1.66=13.72

Figure S22. E-factor calculation for the production of diosgenin by sulfuric acid hydrolysis.

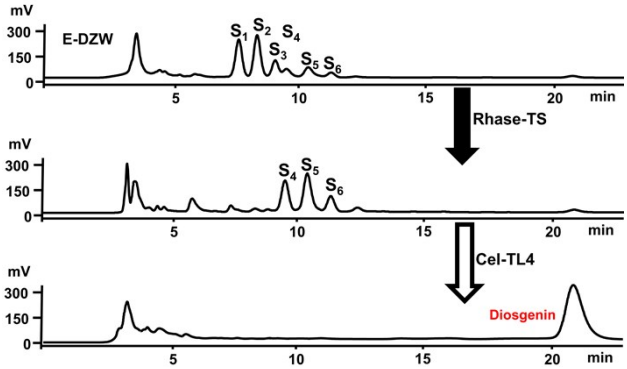


Figure S23. HPLC chromatogram of E-DZW hydrolyzation by two-step enzymatic hydrolysis. E-DZW was first hydrolyzed by Rhase-TS at pH 4.5, 60 ℃ for 10 h, and then hydrolyzed into diosgenin by Cel-TL4 at pH 5.0, 50 ℃ for 25 h with diosgenin yield of 98.5%.

Reference

1. T. D. Bruns, T. J. White and J. W. Taylor, *Annu. Rev. Ecol. Syst.*, 1991, **22**, 525-564.
2. C. M. Visagie, Y. Hirooka, J. B. Tanney, E. Whitfield, K. Mwange, M. Meijer, A. S. Amend, K. A. Seifert and R. A. Samson, *Stud. Mycol.*, 2014, **78**, 63-139.
3. A. D. Tekpinar and A. Kalmer, *Nova Hedwigia*, 2019, **109**, 187-224.
4. H. White, *Econometrica*, 1982, 1-25.
5. K. Tamura, G. Stecher, D. Peterson, A. Filipski and S. Kumar, *Mol. Biol. Evol.*, 2013, **30**, 2725-2729.
6. R. M. Leggett, R. H. Ramirez-Gonzalez, B. J. Clavijo, D. Waite and R. P. Davey, *Front. Genet.*, 2013, **4**, 288.
7. A. M. Bolger, M. Lohse and B. Usadel, *Bioinformatics*, 2014, **30**, 2114-2120.
8. A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev and P. A. Pevzner, *J. Comput. Biol.*, 2012, **19**, 455-477.
9. M. Boetzer and W. Pirovano, *Genome Biol.*, 2012, **13**, R56.
10. A. Massouras, K. Hens, C. Gubelmann, S. Uplekar, F. Decouttere, J. Rougemont, S. T. Cole and B. Deplancke, *Nat. Methods*, 2010, **7**, 485-486.
11. T. Seemann, *Bioinformatics*, 2014, **30**, 2068-2069.
12. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, *Nucleic Acids Res.*, 1997, **25**, 3389-3402.
13. H. Katayama, T. Nagasu and Y. Oda, *Rapid Commun. Mass Spectrom.*, 2001, **15**, 1416-1421.
14. L. Cheng, H. Zhang, H. Liang, X. Sun, X. Shen, J. Wang, W. Wang, Q. Yuan, H.-I. Ri and T.-M. Kim, *Process Biochem.*, 2020, **90**, 81-88.
15. Y. Fu, C. Li, J. Zhu, L. Zhang, Y. Wang, Q. Chen, L. Xu, S. Zhang, Y. Fang and T. Liu, *Biochem. Syst. Ecol.*, 2020, **93**, 104186.

16. L. Xian, F. Wang, X. Luo, Y. Feng and J. Feng, *PLoS One*, 2015, **10**, e0121531.
17. H. Yang, H. Yin, Y. Shen, G. Xia, B. Zhang, X. Wu, B. Cai and J. Tam, *J. Clean. Prod.*, 2016, **131**, 10-19.
18. Y. Chen, Y. Dong, Y. Chi, Q. He, H. Wu and Y. Ren, *Steroids*, 2018, **136**, 40-46.
19. L. Lei, W. Zhang, Z. Gao, L. Liu and X. Yu, *Eur. Food Res. Technol.*, 2018, **244**, 1447-1452.
20. E. Gasteiger, C. Hoogland, A. Gattiker, M. R. Wilkins, R. D. Appel and A. Bairoch, in *The proteomics protocols handbook*, Springer, 2005, pp. 571-607.
21. R. A. Laskowski, M. W. Macarthur, D. S. Moss and J. M. Thornton, *J. Appl. Crystallogr.*, 1993, **26**, 283-291.
22. B. J. Jones, C. N. E. Kan, C. Luo and R. J. Kazlauskas, *Methods Enzymol.*, 2020, **643**, 129-148.
23. Z. Fujimoto, A. Jackson, M. Michikawa, T. Maehara, M. Momma, B. Henrissat, H. J. Gilbert and S. Kaneko, *J. Biol. Chem.*, 2013, **288**, 12376-12385.
24. Y. Zhang and J. Skolnick, *Nucleic Acids Res.*, 2005, **33**, 2302-2309.
25. P. Pachl, J. Škerlová, D. Šimčíková, M. Kotik, A. Křenková, P. Mader, J. Brynda, J. Kapešová, V. Křen and Z. Otwinowski, *Acta Crystallogr. D*, 2018, **74**, 1078-1084.
26. J. Agirre, A. Ariza, W. A. Offen, J. P. Turkenburg, S. M. Roberts, S. McNicholas, P. V. Harris, B. McBrayer, J. Dohnalek and K. D. Cowtan, *Acta Crystallogr. D*, 2016, **72**, 254-265.