Solution structure of *Gaussia* Luciferase with five disulfide bonds and identification of a putative coelenterazine binding cavity by heteronuclear NMR

Nan Wu[$], Naohiro Kobayashi[$], Kengo Tsuda, Satoru Unzai, Tomonori Saotome, Yutaka Kuroda[*] and Toshio Yamazaki[*]

[$] Equal contribution

[*] Correspondence: YK: ykuroda@cc.tuat.ac.jp and TY: toshio.yamazaki@riken.jp

Supplementary Methods

Liquid Chromatography-Mass Spectrometry (LC-MS)

The trypsin digestion of GLuc was performed by incubating 250 μL of the sample at 37°C for 24 hours in 50 mM Tris-HCl (pH 7.0) with 0.05 mg/mL of GLuc, 0.24 mg/mL of Trypsin, 2 M of urea and 1 mM of CaCl2. After the incubation, 100 μL of the sample was aliquoted for the LC-MS measurement.

LC-MS analysis was performed using a Shimadzu LCMS-8040 triple quadruple mass spectrometer (Kyoto, Japan) equipped with a Shimadzu model LC-20AD liquid chromatography system (Kyoto, Japan). Chromatography was done by binary gradient system with 0.1% formic acid (A) and methanol (B). The sample was loaded onto a YMC-Triart C18 (3 μm) 2.1 mm I.D. x 50 mm column equilibrated with 0.1% formic acid with flow rate at 0.2 mL/min. The injection volume was 10 μL. The elution program was as follows: 0→1 min, 0% B; 1→4 min, 0→100% B; 4→6 min, 100% B; 6→7 min, 100→0% B; 7→10 min, 0% B.

45 **Supplementary Tables**

46

47

z (ionic valence)	m/z (mass to charge ratio)
4	1065.75
5	852.80
6	710.83
7	609.43
8	533.37

48

49

50

51

52 **Supplementary Table 1 | The m/z values of the peptide fragment (m =4259.01) of GLuc**
53 **obtained by the enzymatic digestion of trypsin.** The fragment identity was identified using
54 Peptide Mass (https://web.expasy.org/peptide_mass/). Taking into account the formation of
55 intramolecular disulfide bonds, the molecular weight of the largest peptide fragment was
56 4259.01 and was identified as DLEPMEQFIAQVDLCVDCTTGCLK (D106-K129) +
57 GCLICLSHIK (G55-K64) + AGCTR (A50-R54).

58

59

60

61

62

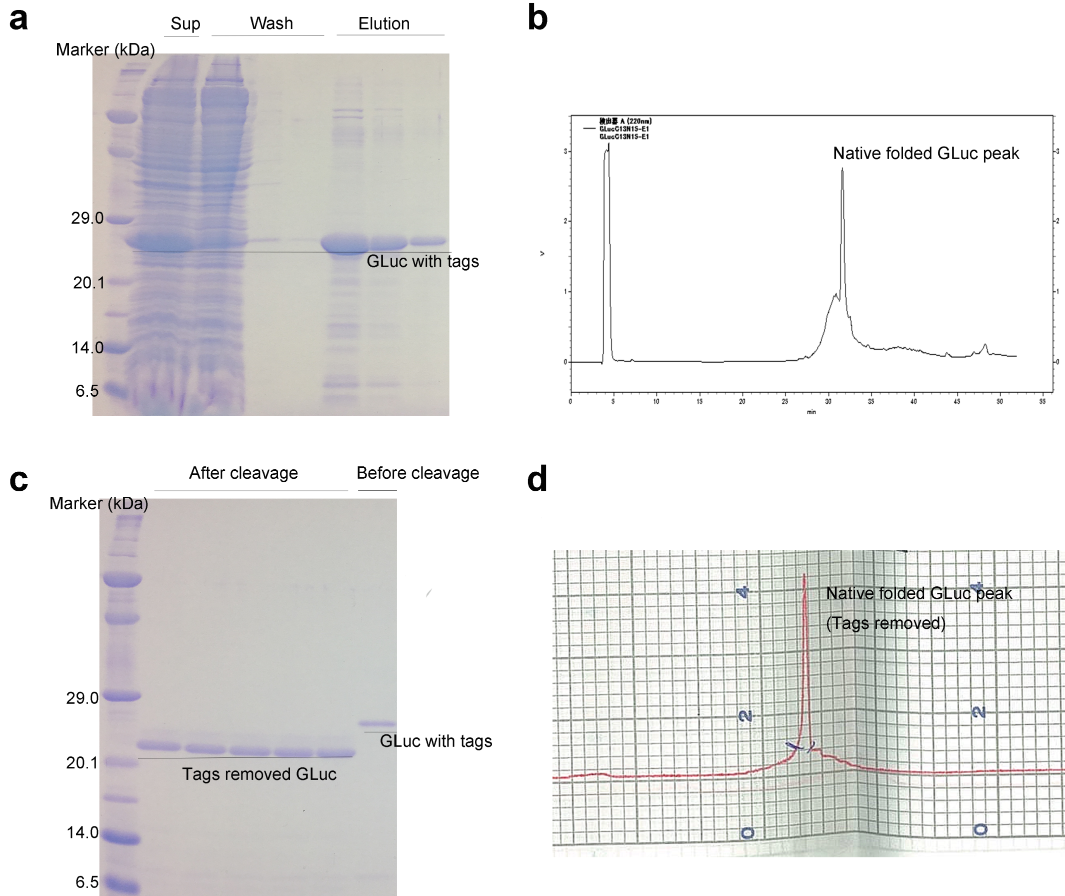
63

64

65

66 **Supplementary Figures**

67



68

69 **Supplementary Fig. 1 | Expression and Purification of GLuc (labeled with ^{13}C and ^{15}N)**

70 The cultured *E.coli* cells were harvested by centrifugation and the pellet was sonicated. GLuc

71 was purificated by applying the supernatant fraction to a Nickel NTA column followed with

72 washing (three times) and elution (three times). An aliquot (0.05% of 1 liter culture) from the

73 the supernatant (Sup) after sonication, washing and elution was collected for the SDS-PAGE

74 analysis (a). The eluted GLuc was dialyzed and purified by reverse phase HPLC in order to

75 collect the main peak that was the natively folded GLuc fraction (b). The collected peak

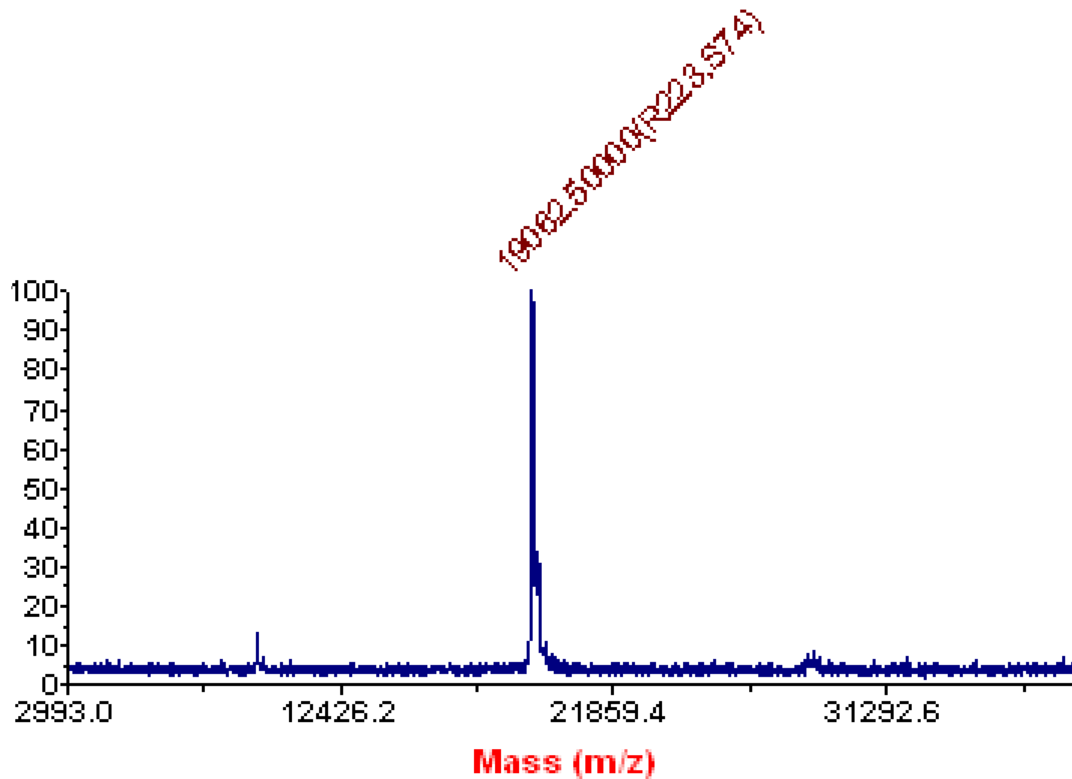
76 portion was freeze dryed and then dissolved in MilliQ water for Factor Xa cleavage. An

77 aliquot of the samples before and after (five parallel experiments) cleavage was analyzed by a

78 SDS-PAGE (c). The tags removed GLuc was then purified again through a 2nd passage to the

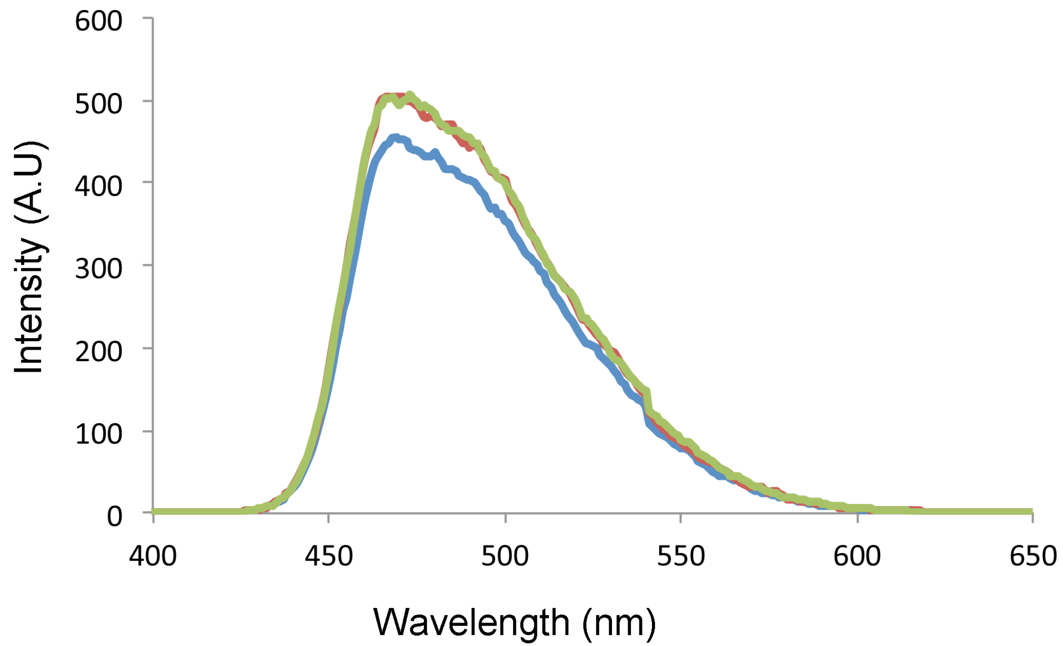
79 reverse phase HPLC (d). The main peak was collected in order to keep all GLuc sample

80 have a uniform native fold. Futher purification details are given in [1].



Supplementary Fig. 2 | MALDI-TOF mass of ^{15}N labeled GLuc expressed in *E.coli*. Freeze dryed ^{15}N labeled GLuc was dissolved in a matrix solution (1mL matrix solution contains: 10 mg sinapic acid, 500 μL acetonitrile, 100 μL 1% TFA and 400 μL MilliQ) at a final protein concentration of 5 μM. One microliter (containnig 5 pmol of GLuc) was loaded on the MALDI-TOF plate for measurement. The calculated mass of ^{15}N labeled GLuc is 19055.8 Da.

94



95 **Supplementary Fig. 3 | Bioluminescence activity of ^{15}N labeled GLuc expressed in *E.coli*.**

96 Bioluminescence was initiated by adding 30 µL of coelenterazine to a final concentration of 3

97 µM to a 2 mL 0.01 µM GLuc solution in 50 mM Tris–HCl buffer pH 8.0 and 50 mM NaCl.

98 The spectra were measured at 20 °C using a JASCO FP-8000 fluorescence spectrophotometer

99 with a scan speed of 20000 nm/min and an emission bandwidth of 5 nm. The activity

100 measurement was repeated three times (shown in red, blue and green, respectively).

101

102

103

104

105

106

107

108

```
                  ....|....|....|....|....|....|....|....|....|....|
                      5   10   15   20   25   30   35   40   45   50
GLuc sequence     KPTENNEDFNIVAVASNFATTDLDADRGKLPGKKLPLEVLKEMEANARKA
TALOS+            ccccccccccchhhhhhccccccccccccccccccchhhhhhhhhhhhhc
Consensus 85%     cccccccc--eee--ccc-----ccc-cccccccccchhhhhhhhhhhhhc
NOE flexible      ≈≈≈≈≈≈≈≈≈≈≈
H/D 20min                       * *                   *    **
H/D 18h
Packing status                      ○○○○○○○○○○○○○○○○○○○○○○○○○○○○○○○○
Domain(repeats)                             111111111111111111111111

                  ....|....|....|....|....|....|....|....|....|....|
                     55   60   65   70   75   80   85   90   95  100
GLuc sequence     GCTRGCLICLSHIKCTPKMKKFIPGRCHTYEGDKESAQGGIGEAIVDIPE
TALOS+            cecchhhhhhhcceechhhhhhccccccccccccccccccccceeecccc
Consensus 85%     cccc--eee--ccccc-c----ccccc---ccccc-ccccc---ee-ccc
NOE flexible                                     ≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈
H/D 20min                **** *      *          *          *
H/D 18h                  + ++
Packing status    ○●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●●
Domain(repeats)   11111111111111111111111111111111111111111111111222

                  ....|....|....|....|....|....|....|....|....|....|
                    105  110  115  120  125  130  135  140  145  150
GLuc sequence     IPGFKDLEPMEQFIAQVDLCVDCTTGCLKGLANVQCSDLLKKWLPQRCAT
TALOS+            cchhhhcchhhhhhhhhhhhccccchhhhhhhhcccchhhhhhhccccccc
Consensus 85%     cccccccchhhhhhhhhh---hh--hhhh-------hhhhhhh------h
NOE flexible                                                     ≈
H/D 20min         *     *     ********       **
H/D 18h                        +  + ++
Packing status    ●●●●●●●●●●●●●●●●●●●●●●●○○○○○○○○○○○○○○○○○○○○○○○○○○○
Domain(repeats)   22222222222222222222222222222222222222222222222222

                  ....|....|....|....|
                    155  160  165  170
GLuc sequence     FASKIQGQVDKIKGAGGD
TALOS+            cccccccccccccccccc
Consensus 85%     hhhhhh--hh----cccc
NOE flexible      ≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈≈
H/D 20min
H/D 18h
Packing status    ○
Domain(repeats)   222222222222222222


h = 4-turn helix (α helix). Min length 4 residues.
e = extended strand in parallel and/or anti-parallel β-sheet conformation. Min length 2 residues.
c = coil (residues which are not in any of the above conformations).
```

109

110 **Supplementary Fig. 4 | Residue-level structural characterization of GLuc.** TALOS+

111 shows GLuc's secondary structure calculated from the chemical shifts using the TALOS+

112 server [2]. Consensus 85% indicates GLuc secondary structure predicted using seven publicly

113 available predictors with 85% consensus [3]. NOE flexible and marked with "≈" shows highly

114 flexible residues identified by ^{1}H-^{15}N heteronuclear NOE experiment. H/D 20min and H/D

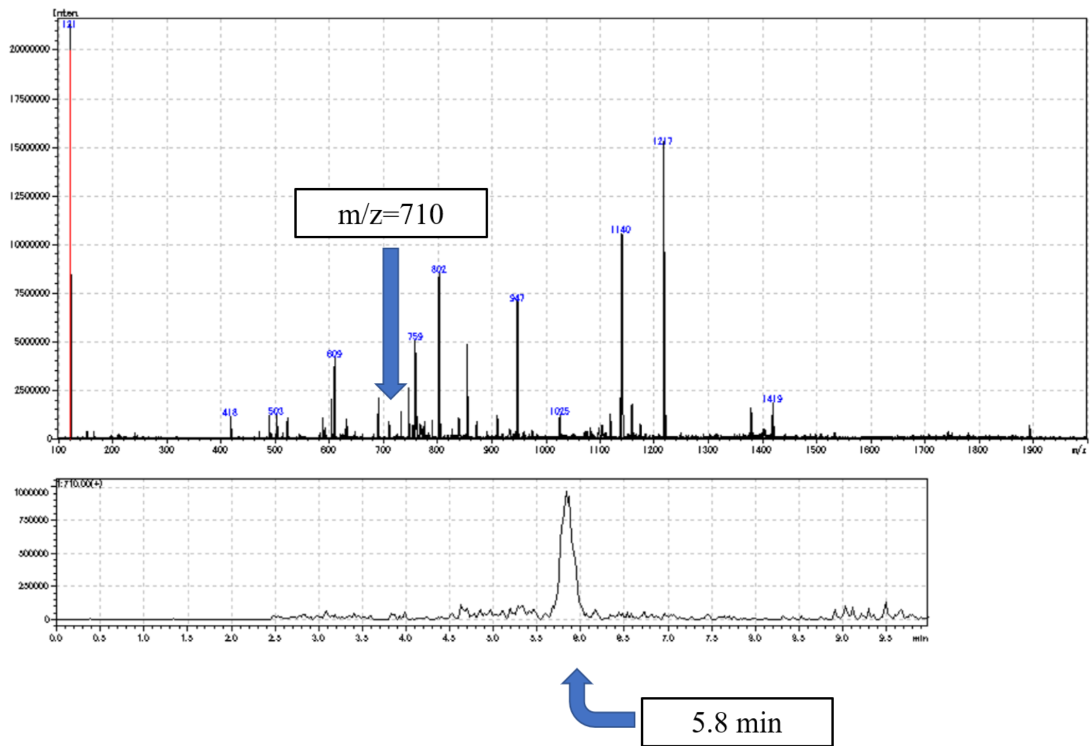
115 18h respectively shows residues that retained their signal after 20 minutes (marked with "*")

116 and 18 hours (marked with "+") in D_2O. Packing status shows the tightly packed moiety

117 (marked with "●") and the loosely packed moiety (marked with "○"), respectively. Domain

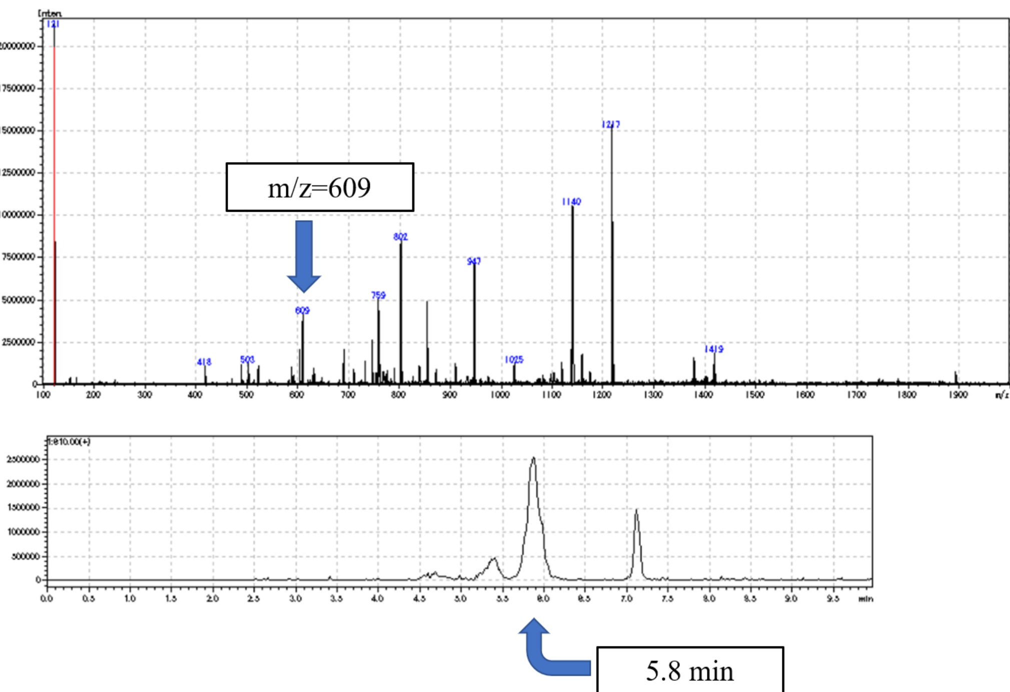
118 (repeats) indicates the two homologous repeats of GLuc that identified in the reference [4].

z=6



119

z=7



120

121

122 **Supplementary Fig. 5 | LC-MS spectra and selected ion chromatograms of the peptide**

123 **fragment (m=4259) of GLuc obtained by the enzymatic digestion of trypsin.** The

124 experimental values of m/z coincided with the computed values (z=6 and z=7), and the

125 retention times of the LC peaks of the two different ionic valences were identical.

126

127

128

129

130

131

132

133

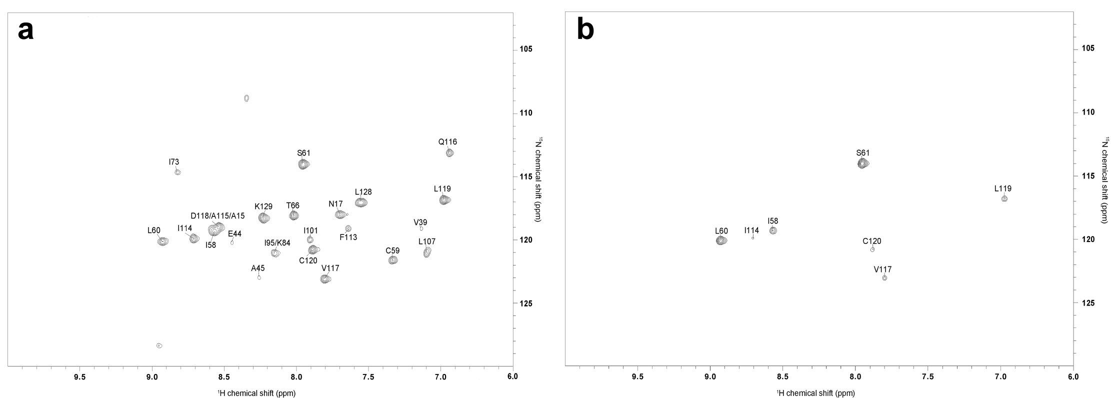
134

```
MODEL A
# No:  Chain   Z    rmsd lali nres  %id PDB  Description

MODEL B
# No:  Chain   Z    rmsd lali nres  %id PDB  Description
   1:  4ddg-A  2.5  3.5   59   399    5   MOLECULE: UBIQUITIN-CONJUGATING ENZYME E2 D2, UBIQUITIN THI
   2:  2xhi-A  2.4  3.7   52   316   10   MOLECULE: N-GLYCOSYLASE/DNA LYASE;
x  3:  4oge-A  2.4  9.3   99   977    5   MOLECULE: HNH ENDONUCLEASE DOMAIN PROTEIN;
x  4:  5ukh-A  2.3  5.6   52   321   13   MOLECULE: UNCHARACTERIZED PROTEIN;
   5:  1u9p-A  2.3  3.6   48    96    4   MOLECULE: PARC;
   6:  2rji-A  2.3  2.8   53    84    2   MOLECULE: ERYTHROCYTE BINDING ANTIGEN 175;
   7:  4d8o-A  2.3  3.0   46   509   11   MOLECULE: ANKYRIN-2;
   8:  3mzy-A  2.2  4.0   67   123    9   MOLECULE: RNA POLYMERASE SIGMA-H FACTOR;
   9:  5dic-A  2.2  2.9   51   115    4   MOLECULE: ODORANT-BINDING PROTEIN;
  10:  6e11-E  2.2  4.2   65   210    8   MOLECULE: UNKNOWN (CLAW);
x 11:  5gha-D  2.1  5.6   55   310    4   MOLECULE: SULFUR TRANSFERASE TTUA;
x 12:  5bmq-A  2.1  7.4   59   205    8   MOLECULE: ERFK/YBIS/YCFS/YNHG FAMILY PROTEIN;
x 13:  4dbg-B  2.1  8.2   50   146   10   MOLECULE: RANBP-TYPE AND C3HC4-TYPE ZINC FINGER-CONTAINING
  14:  2g7r-A  2.1  2.7   48    86   13   MOLECULE: MUCOSA-ASSOCIATED LYMPHOID TISSUE LYMPHOMA TRANSL
x 15:  6hls-A  2.1 10.5   70  1431   11   MOLECULE: DNA-DIRECTED RNA POLYMERASE I SUBUNIT RPA190;
  16:  3oao-A  2.0  4.2   47   140    9   MOLECULE: UNCHARACTERIZED PROTEIN FROM DUF2059 FAMILY;
               H1 H2 H3 H4 H5 H6 H7 H8 H9
       4ddg-A  o  -  o  -  -  -  -  o  -
       2xhi-A  o  -  o  o  -  -  -  -  -
       1u9p-A  o  -  o  o  -  -  -  -  -
       2rji-A  o  -  o  -  -  -  -  o  -
       4d8o-A  o  -  o  o  -  -  -  -  -
       3mzy-A  o  -  o  o  -  -  -  o  -
       5dic-A  -  -  o  o  -  -  -  o  -
       6e11-E  o  o  o  -  -  -  -  o  -
       2g7r-A  o  -  o  o  o  -  -  -  -
       3oao-A  o  o  o  o  -  -  o  -  -

MODEL C
# No:  Chain   Z    rmsd lali nres  %id PDB  Description
   1:  4qo5-A  2.4  4.8   73   521    8   MOLECULE: HYPOTHETICAL MULTIHEME PROTEIN;
   2:  5c4y-A  2.2  2.7   61   136    3   MOLECULE: PUTATIVE TRANSCRIPTION REGULATOR LMO0852;
   3:  2g7r-A  2.0  2.8   49    86   12   MOLECULE: MUCOSA-ASSOCIATED LYMPHOID TISSUE LYMPHOMA TRANSL
               H1 H2 H3 H4 H5 H6 H7 H8 H9
       4qo5-A  -  o  o  o  o  -  o  -  -
       5c4y-A  -  o  o  o  o  -  o  -  -
       2g7r-A  o  -  o  o  o  -  -  -  -

MODEL D
# No:  Chain   Z    rmsd lali nres  %id PDB  Description

MODEL E
# No:  Chain   Z    rmsd lali nres  %id PDB  Description
   1:  5k78-B  2.8  3.7   61   351   10   MOLECULE: RNA LARIAT DEBRANCHING ENZYME, PUTATIVE;
   2:  2r18-A  2.0  2.3   41   129   10   MOLECULE: CAPSID ASSEMBLY PROTEIN VP3;
               H1 H2 H3 H4 H5 H6 H7 H8 H9
       5k78-B  o  -  o  -  -  -  -  o  -
       2r18-A  -  -  o  -  -  -  -  -  o

MODEL F
# No:  Chain   Z    rmsd lali nres  %id PDB  Description
   1:  4d8o-A  2.1  3.0   47   509    9   MOLECULE: ANKYRIN-2;
   2:  2h56-A  2.0  3.1   45   218   11   MOLECULE: DNA-3-METHYLADENINE GLYCOSIDASE;
               H1 H2 H3 H4 H5 H6 H7 H8 H9
       4d8o-A  o  -  o  -  -  -  -  o  -
       2r18-A  -  -  o  -  -  -  -  -  o

MODEL G
# No:  Chain   Z    rmsd lali nres  %id PDB  Description
x  1:  1e3a-B  2.8  6.5   85   560    8   MOLECULE: PENICILLIN AMIDASE ALPHA SUBUNIT;
```

Supplementary Fig. 6 | Similar structure search of GLuc against the Protein Data Bank using Dali server. Seven GLuc models that have all five disulfide bonds were selected from the nineteen NMR-derived structures and submitted to a Dali server (http://ekhidna2.biocenter.helsinki.fi/dali/) [5] to search similar structure in the Protein Data Bank. The data outputs of Dali sever of all seven GLuc models are shown for MODEL A (representative structure), B, C, D, E, F and G, respectively. None the searches showed a similar structure for MODEL A and D. Although a few similar structures were identified for

143 on MODEL B, C, E, F and G, all of which exhibited low Z-score (<3.0, which over 6.0 could
144 be taken to be highly significant), indicating no significant structural similarity. Structures
145 marked with a "×" were firstly ignored, because they have an rmsd >4.0 are structurally
146 dissimilar to GLuc. Structures with rmsd less than 4.0 were then partially matched with GLuc
147 models over helix segments (shown in "Details", "H1"-"H9" represent nine helices of GLuc,
148 all matched helices were marked with "○" below). GLuc structure shows H3H4H5-H7H8H9
149 is critical for GLuc's structural skeleton (Fig. 3) and the putative catalytic cavity (Fig. 4), but
150 none of the identified structures matched GLuc in this areas, which strongly suggested that
151 the matches are accidental.

152

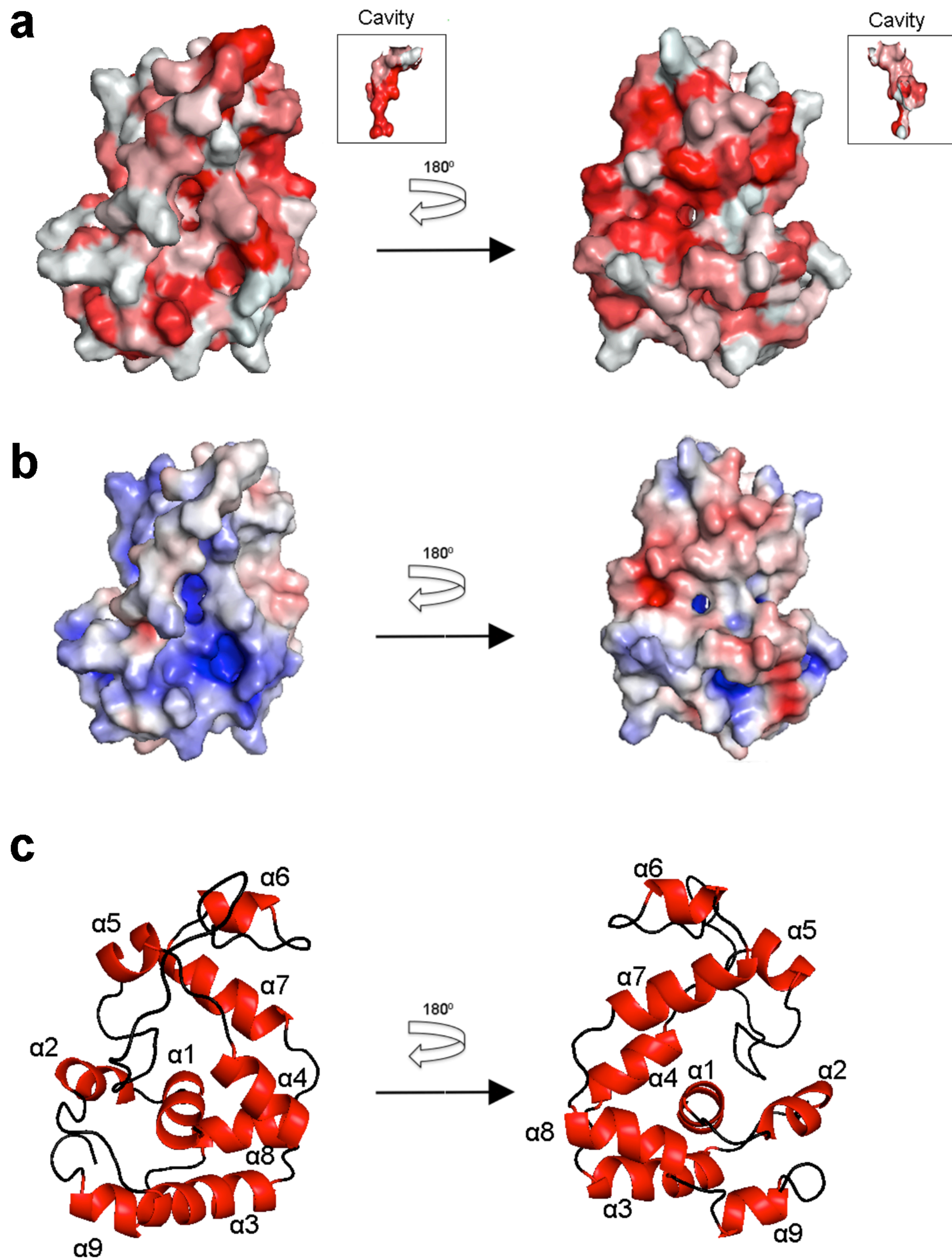


153 **Supplementary Fig. 7 | H/D exchanging ^{1}H-^{15}N HSQC spectra of GLuc.** Measurements

154 were initialed by dissolving GLuc to a final concentration of 0.2 mM in D_2O containing 50

155 mM MES buffer pH 6.0 with 2 mM NaN_3. ^{1}H-^{15}N HSQC spectra were measured after

156 incubating GLuc at 298 K for 20 minutes (a) and 18 hours (b), respectively.



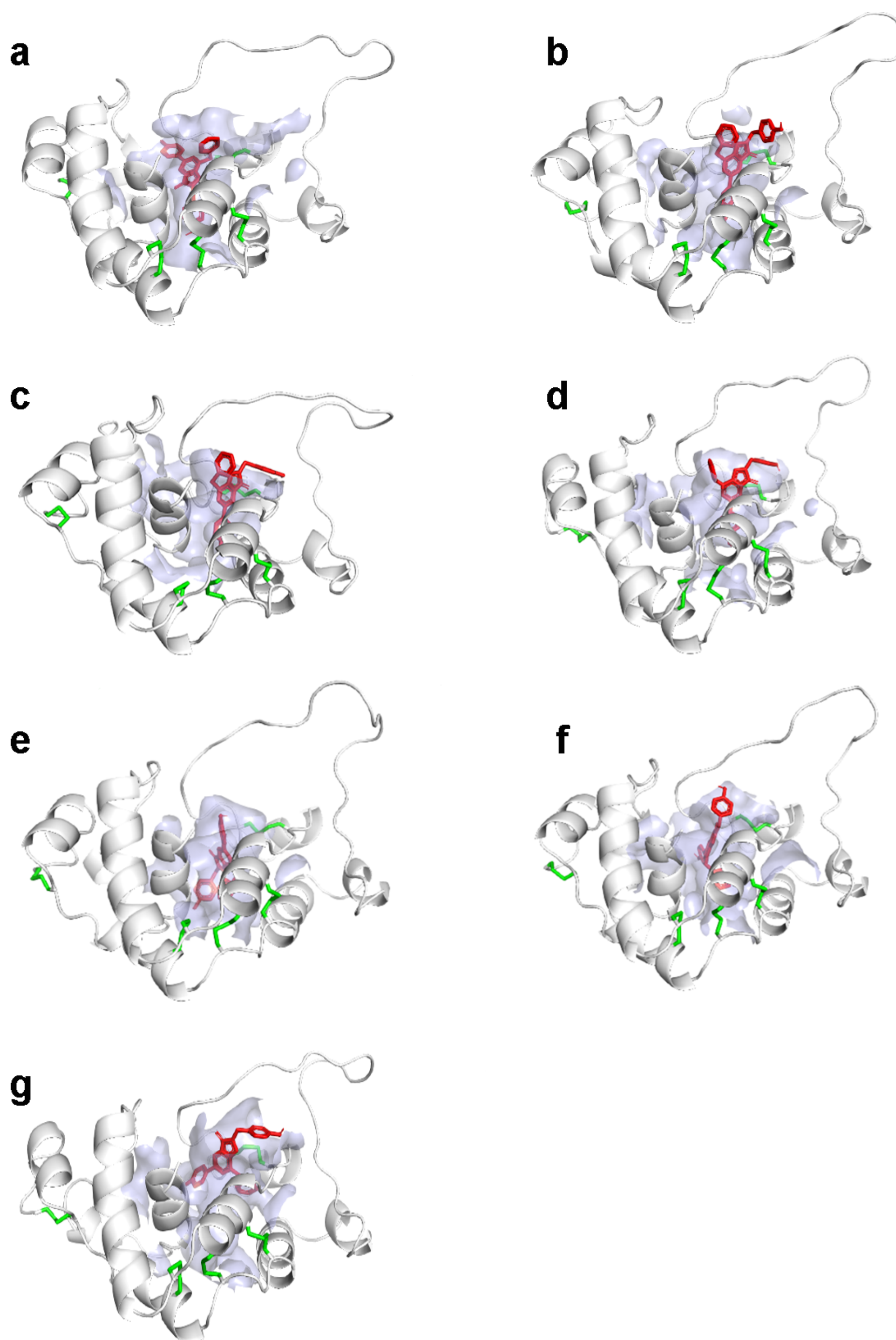
157

158 **Supplementary Fig. 8 | Surface representation of GLuc (representative structure,**
159 **residues 10-148).** (a) Hydropathy surface representation of GLuc shown by coloring the
160 molecule using "color_h" pymol script (https://pymolwiki.org/index.php/Color_h):
161 hydrophilic regions were shown in white and hydrophobic regions were shown in red. The
162 insert figures show the interior cavity. (b) Electrostatic surface representation of GLuc

163 calculated using an APBS plugin of Pymol. Positive charges are shown in blue, negative

164 charges are shown in red and neutral residues are in gray. (c) Ribbon representation of GLuc

165 from the same direction as in (a) and (b).



166

167 **Supplementary Fig. 9 | Conformations of seven best docking complexes.**

168 Seven GLuc models that are referred in Supplementary Fig. 6 were docked with
169 coelenterazine using Autodock4.2 [6]. The coelenterazine and three activity-related residues on
170 GLuc (R76, H78 and T79) were set as flexible, and all other residues were set as rigid. 50
171 docking complexes were generated during the docking simulation of each GLuc model

172 (Genetic Algorithm), and the one with lowest binding energy were selected and shown in (a)

173 ~ (g), respectively. (a) shows the docking complex of MODEL A (representative structure),

174 binding energy= -6.47 kcal/mol, inhibit constant= 17.96 μM; (b) shows the docking complex

175 of MODEL B, binding energy= -6.13 kcal/mol, inhibit constant= 32.26 μM; (c) shows the

176 docking complex of MODEL C, binding energy= -5.8 kcal/mol, inhibit constant= 56.21 μM;

177 (d) shows the docking complex of MODEL D, binding energy= -7.07 kcal/mol, inhibit

178 constant= 6.57 μM; (e) shows the docking complex of MODEL E, binding energy= -9.29

179 kcal/mol, inhibit constant= 154.09 nM; (f) shows the docking complex of MODEL F, binding

180 energy= -11.8 kcal/mol, inhibit constant= 2.26 nM; (g) shows the docking complex of

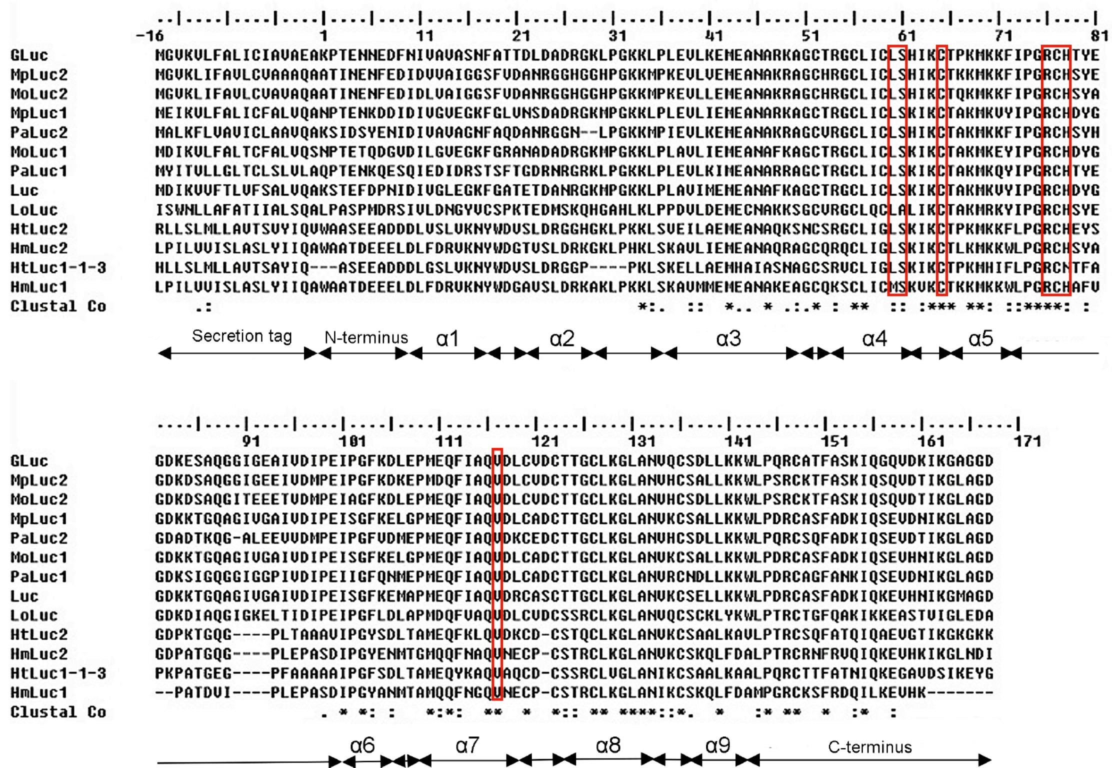
181 MODEL G, binding energy= -8.36 kcal/mol, inhibit constant= 742.04 nM. GLuc molecules

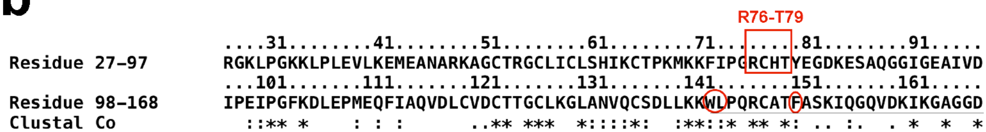
182 are shown in white; coelenterazines are shown in red; five disulfide bonds are shown in green;

183 the interior surface (cavity) that is formed by residues N10, V12, A13, V14, S16, N17, F18,

184 L60, S61, I63, K64, C65, R76, C77, H78, T79, F113, I114, V117 are shown in transparent

185 light blue.





186

187 **Supplementary Fig. 10 | Sequence alignment of luciferases.** GLuc alignment with 12

188 similar luciferases is shown in (a). GLuc indicates the GLuc's amino acid sequence. Residues

189 -16 to 0 represent a secretion tag sequence. Clustal Co indicates amino acid conservation

190 throughout the 12 sequences of BLAST-detected luciferases according to ClustalW [3]: (An

191 asterisk indicates a fully conserved residue, a column shows a highly conserved residue, and a

192 dot shows a poorly conserved residue according to ClustalW classification). Nine α helices

193 identified from the representative structure are marked at the bottom. Residues L60, S61, C65,

194 R76, C77, H78 and V117 were marked with red squares. The alignment of GLuc's

195 homologous repeats [3,4]: residues 27-97 against residues 98-168 is shown in (b). Residues
196 W143, L144, F151 [1] are marked with red circles, and the residues in the activity-associated
197 loop R76-T79 [7] are marked with a red square. The flexible C-terminus is underlined. Clustal
198 Co shows amino acid consensus generated by ClustalW (the symbol codes are the same as in
199 (a)).

200

201 **References**

202 1. Wu, N., Kamioka, T. & Kuroda, Y. A novel screening system based on VanX‐
203 mediated autolysis—Application to *Gaussia* luciferase. *Biotechnol Bioeng* **113**,
204 1413-1420 (2016).

205 2. Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. TALOS+: a hybrid method for
206 predicting protein backbone torsion angles from NMR chemical shifts. *J Biomol*
207 *NMR* **44**, 213-223 (2009).

208 3. Wu, N., Rathnayaka, T. & Kuroda, Y. Bacterial expression and re-engineering of
209 *Gaussia princeps* luciferase and its use as a reporter protein. *BBA-Proteins Proteom*
210 **1854**, 1392-1399 (2015).

211 4. Inouye, S. & Sahara, Y. Identification of two catalytic domains in a luciferase
212 secreted by the copepod *Gaussia princeps*. *Biochem Biophys Res Commun* **365**,
213 96-101 (2008).

214 5. Hasegawa, H. & Holm, L. Advances and pitfalls of protein structural alignment. *Curr*
215 *Opin in Struc Biol* **19**, 341-348 (2009).

216 6. Morris, G.M. et al. AutoDock4 and AutoDockTools4: Automated docking with
217 selective receptor flexibility. *J Comput Chem* **30**, 2785-2791 (2009).
218 7. Kim, S.B., Suzuki, H., Sato, M. & Tao, H. Superluminescent variants of marine
219 luciferases for bioassays. *Anal Chem* **83**, 8732-40 (2011).
220