



**HAL**  
open science

## Biochemical characterization of a glycosyltransferase Gtf3 from *Mycobacterium smegmatis*: a case study of improved protein solubilization

Mahfoud Bakli, Loukmane Karim, Nassima Mokhtari-Soulimane, Hafida Merzouk, Florence Vincent

### ► To cite this version:

Mahfoud Bakli, Loukmane Karim, Nassima Mokhtari-Soulimane, Hafida Merzouk, Florence Vincent. Biochemical characterization of a glycosyltransferase Gtf3 from *Mycobacterium smegmatis*: a case study of improved protein solubilization. 3 Biotech, 2020, 10 (10), 10.1007/s13205-020-02431-x . hal-03163519

**HAL Id: hal-03163519**

**<https://amu.hal.science/hal-03163519v1>**

Submitted on 29 Nov 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Article title: Biochemical characterization of a glycosyltransferase Gtf3 from *Mycobacterium smegmatis*: A**  
2 **case study of improved protein solubilization.**

3

4 **Abstract**

5 Glycosyltransferases (GTs) are widely present in several organisms. These enzymes specifically transfer sugar  
6 moieties to a range of substrates. The processes of bacterial glycosylation of the cell wall and their relations with  
7 host-pathogen interactions have been studied extensively, yet the majority of mycobacterial GTs involved in the  
8 cell wall synthesis remain poorly characterized. Glycopeptidolipids (GPLs) are major class of glycolipids present  
9 on the cell wall of various mycobacterial species. They play an important role in drug resistance and host-  
10 pathogen interaction virulence. Gtf3 enzyme performs a key step in the biosynthesis of triglycosylated GPLs.  
11 Here we describe a general procedure to achieve expression and purification of recombinant protein Gtf3 from  
12 *Mycobacterium smegmatis* using an *E. coli* expression system. We reported also a combined bioinformatics and  
13 biochemical methods to predict aggregation propensity and improve protein solubilization of recombinant Gtf3.

14

15 **Keywords:** glycosyltransferase, expression and purification of recombinant protein, protein solubilization,  
16 *Mycobacterium smegmatis*.

## 17 **Introduction**

18 During the four last decades, the number of recombinant proteins used for several academic, medical and  
19 industrial applications has increased dramatically (Warne and Mahler 2018). This engineering field has been  
20 growing essentially due to considerable progress in available sequenced genomes, and to biotechnology and  
21 strategy developments in achieving high level protein expression. It ranges from expression vector design to  
22 final application (Vandermies and Fickers 2019; Kushwaha and Salis 2015; Rosano and Ceccarelli 2014), during  
23 which several obstacles may be encountered. Some problems are related to intrinsic physicochemical features  
24 such as protein conformation, stability, and structural flexibility, and others related to experimental procedures  
25 such as expression and purification (Deller et al. 2016; Young et al. 2012). Nevertheless, protein structure  
26 prediction tools have presently become sufficiently robust to provide valuable insight into the structures, even  
27 with uncrystallized proteins (Thayer 2016). Because some insoluble proteins contain residues that decrease  
28 their solubility (aggregation hotspots), several new methods were developed to predict hotspots and  
29 hydrophobic patches without a crystal structure, with the goal of solubilizing these expressed proteins (Matsui  
30 et al. 2017). The case study concerns the glycosyltransferase from *Mycobacterium smegmatis*, *Gtf3*. *Gtf3* gene  
31 belongs to glycopeptidolipids (GPLs) biosynthetic locus containing three ORFs, *Gtf1*, *Gtf2*, and *Gtf3* (Jeevarajah  
32 et al. 2002). *Gtf3* enzyme performs a key step in the biosynthesis of triglycosylated forms of GPLs (Jeevarajah et  
33 al. 2002; Billman-Jacobe 2004; Deshayes et al. 2005; Mukherjee and Chatterji 2005). GPLs are found in outer  
34 layers of the *mycobacterial* cell wall. They are produced by nontuberculous mycobacteria (Brennan and Crick  
35 2007; Schorey and Sweet 2008). Furthermore, several physiological processes are affected by presence or lack of  
36 GPLs in the mycobacterial *cell* wall, such as motility or biofilm formation, host-pathogen interactions,  
37 intracellular survival strategies, and virulence. This ultimately influences the clinical outcomes and the disease  
38 manifestations (Gutiérrez et al. 2018). GTs catalyze glycosylation reactions involving the transfer of a glycosyl  
39 group from an activated sugar moiety (NDP-donor) onto a broad variety of acceptor molecules (proteins, lipids,  
40 nucleic acids or oligosaccharides) (Lairson et al. 2008). Functionally, GTs are subdivided into retaining or  
41 inverting enzymes according to the stereochemistry of the substrates and products (Schuman et al. 2006).  
42 Structurally, GTs adopt one of the three folds, termed GT-A, GT-B, and GT-C. GT-B enzymes comprise two  
43  $\beta/\alpha/\beta$  Rossmann-like domains that face each other. Between this domains is located the active site containing  
44 residues which are involved in leaving group departure. Generally, the reaction catalyzed by these enzymes is  
45 metal ion independent (Schmid et al. 2016). Triglycosylated GPLs result from the addition of an extra rhamnosyl  
46 residue. Moreover, the function of *Gtf3* gene was not precisely determined, although genetic studies reported that

47 *gtf3* is involved in adding the 3,4-di-*O*-methyl-rhamnose to the terminal 3,4-di-*O*-methyl rhamnose and it was  
48 also involved in adding 3-*O*-Me-Rhamnose (Deshayes et al. 2005; Miyamoto et al. 2006). Gtf3 enzyme belongs  
49 to the CAZy *GTI* superfamily, sharing characteristics of the *GT-B* structural fold and *inverting* catalytic  
50 mechanism (Lairson et al. 2008). Herein are presented a combined bioinformatics and biochemical methods to  
51 predict aggregation possibility and improve solubilization of expressed Gtf3.

52

53 **Materials and methods**

54 **Bioinformatics study**

55 **Hydrophobic Cluster Analysis (HCA)**

56 Putative glycosyltransferase, Gtf3 from *Mycobacterium smegmatis* MC2 155 strain (gi|23345078) belongs to the  
57 glycosyltransferase class (GTs; EC 2.4) and GT1 family according to Carbohydrate Active Enzyme (CAZy)  
58 Database classification (Lombard et al. 2014). Enzymes classification is based on amino acid sequence  
59 similarities ([www.cazy.org/](http://www.cazy.org/)). Hydrophobic Clusters Analysis (HCA) was performed on Gtf3 using MeDor  
60 program (Lieutaud et al. 2008).

61 **Prediction of secondary and tertiary structures**

62 Secondary structures of Gtf3 were predicted with Phyre2 web server  
63 ([www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index](http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index)) which uses a protein remote homology detection  
64 methods to build three-dimensional models (Kelley et al. 2015). The resulting model of Gtf3 was submitted  
65 to molecular visualization system, Pymol (Bramucci et al. 2012) in order to predict the position of cysteine  
66 residues.

67 **Prediction of oligomeric state**

68 Oligomeric state of Gtf3 was predicted using web servers ROBETTA ([new.robetta.org](http://new.robetta.org)) and SWISS-MODEL  
69 ([swissmodel.expasy.org](http://swissmodel.expasy.org)). These programs predict the homo-oligomer structure of protein of interest from an  
70 amino acid sequence (Kim et al. 2004; DiMaio et al. 2011; Biasini et al. 2014).

71 **Multiple alignment**

72 Characterized rhamnosyltransferase sequences from the CAZy GT1 superfamily  
73 ([http://www.cazy.org/GT1\\_characterized.html](http://www.cazy.org/GT1_characterized.html)) related to Gtf3 protein sequence were retrieved from National  
74 Center for Biotechnology Information (NCBI) database using BLAST program. Multiple alignment of these  
75 amino acid sequences were generated using the CLUSTALW software with default parameters and visualized by  
76 Bioedit program.

77 **Biochemical study**

78 **Expression of the recombinant protein**

79 The Gtf3 gene was cloned into pDEST<sup>TM</sup>17 expression vector (Invitrogen) in frame with a sequence coding for  
80 an N-terminal polyhistidine tag (His-tag). *E. coli* Rosetta<sup>TM</sup> (DE3) pLysS competent cells (Novagen) were  
81 transformed by heat shock with 10 ng/μl of plasmid DNA carrying ampicillin and chloramphenicol resistance  
82 genes in addition to *Gtf3*. Transformed bacteria were precultured overnight at 37°C with shaking (220 rpm) in

83 300 ml of LB Broth Miller growth medium (Fisher Scientific) supplemented with antibiotics ampicillin (100  
84  $\mu\text{g/ml}$ ) and chloramphenicol (34  $\mu\text{g/ml}$ ). 60 ml of the saturated culture were then transferred into 6 L of Super  
85 Broth medium, which contains per liter: 32 g tryptone, 20 g yeast extract, 5 g NaCl and 5 ml 1 N NaOH, and  
86 cultured at 37°C with shaking (220 rpm) up to an optical density (OD 600 nm) of 0.8 prior to induction with 0.5  
87 mM of isopropyl- $\beta$ -D-thiogalactopyranoside (IPTG). After 4 hours of incubation at 30°C and shaking (220 rpm),  
88 cells were harvested by centrifugation at 4000 rpm for 30 min at 4°C. Pellets were resuspended in 200 ml of  
89 lysis buffer (containing 50 mM Tris pH7.5, 10 mM imidazole, 150 mM NaCl, 5 mM  $\beta$ -mercaptoethanol, 0.25  
90 mg/ml lysozyme, 0.1% Tween 20, and 20.5% glycerol) and incubated in the presence of DNase (10  $\mu\text{g/ml}$ ) for  
91 30 min under gentle shaking at 4 °C and then sonicated on ice for 5 min. Cell debris were pelleted and discarded  
92 after 30 min centrifugation at 14,000 rpm at 4 °C. Supernatants containing the soluble proteins were filtered  
93 through 0.45  $\mu\text{m}$  Durapore filters (Millipore) and kept on ice for further steps.

#### 94 **Protein purification**

95 The recombinant protein Gtf3 present in the supernatant was purified under native conditions using Akta Xpress  
96 fast protein liquid chromatography (Amersham, Biosciences). Firstly, filtered supernatant was loaded onto a His-  
97 Trap HP 5-ml column (GE Healthcare, Cat. No. 17-5248-02) which was equilibrated with Buffer A (50 mM Tris  
98 pH 7.5, 10 mM imidazole, 150 mM NaCl, 5 mM  $\beta$ -mercaptoethanol) at a flow rate of 1  $\text{ml}\cdot\text{min}^{-1}$  (Ren et al.  
99 2013). After protein binding, the column was washed with 10 column volumes (CVs) of Buffer A prior to  
100 elution with Buffer B (50 mM Tris pH 7.5, 500 mM imidazole, 150 mM NaCl, 5 mM  $\beta$ -mercaptoethanol). Peak  
101 fractions containing the His-tagged recombinant Gtf3 protein were selected based on the profile obtained by  
102 SDS-PAGE, and were then pooled. The protein concentration was estimated at 280 nm using a NanoDrop<sup>TM</sup>  
103 1000 spectrophotometer (Thermo Fisher Scientific). Pooled fractions of Gtf3 (46 kDa) were dialyzed against  
104 buffer containing 50 mM Tris pH 7.5, 150 mM NaCl, and 5 mM DTT, overnight at 4 °C in appropriate dialysis  
105 cassette and then concentrated to a final volume of 6 ml using a centrifugal concentrator, Centricon of 30 kDa  
106 cut-off (Amicon). Recombinant Gtf3 was further purified by size exclusion chromatography (SEC). 6 ml of  
107 protein sample was loaded onto a HiLoad 26/60 Superdex 75 pg (GE Healthcare) column at a flow rate of 1.5  
108  $\text{ml}\cdot\text{min}^{-1}$ , which was equilibrated with a buffer containing 50 mM Tris pH 7.5, 150 mM NaCl, and 5 mM of  
109 reducing agent, DTT (dithiothreitol) (Ren et al. 2013). Purified Gtf3 protein (46 kDa) was collected and  
110 concentrated using Centricon of 30 kDa cut-off (Amicon), and its concentration determined using a NanoDrop<sup>TM</sup>  
111 1000 spectrophotometer (Thermo Fisher Scientific). The purity of Gtf3 protein was assessed by SDS-PAGE, and  
112 the identity of each band was confirmed by mass spectrometry (MS).

### 113 **Solubilization of protein aggregates**

114 A linear carbohydrate-based polymer of 5 kDa named NVoy (or NV10) was prepared in line with the  
115 manufacturer's instructions (Expedeon) (Guild et al. 2011). Purified Gtf3 was mixed with 5-fold mass excess of  
116 NVoy (e.g., 1 mg/ml protein with 5 mg/ml polymer) and transferred into microdialysis cassette (Thermo Fisher  
117 Scientific). In order to determine the optimal condition for Gtf3 solubilization, several buffers at different pH  
118 and salt concentrations were screened for the dialysis step (**Table 1**). Absorbance of each Gtf3-containing buffer  
119 was measured at 340 nm using a Varian Cary Scan 50 spectrophotometer to assess the degree of protein  
120 solubilization. OD<sub>340</sub> was used to measure light scattering and thus to estimate the precipitation.

### 121 **SDS-PAGE**

122 Purified protein profile of Gtf3 was obtained according to the method of Laemmli (Laemmli 1970). Fractions  
123 were loaded onto a 12% polyacrylamide gel. Gel electrophoresis was run at 300 V for 15 min with Mini-  
124 PROTEAN II (Bio-Rad). Gels were subsequently stained with Coomassie R250 (Thermo scientific) and the  
125 image was captured using a ImageQuant TL software (GE Healthcare) based on densometric parameters scan  
126 (GE Healthcare) (Ali et al. 2012). Molecular weights of the protein bands were calculated based on the  
127 molecular weight marker (Bio-Rad).

### 128 **In-gel digestion and mass spectrometry analysis (MS)**

129 Protein gel bands of interest were excised and then digested overnight at 37°C by 12.5 µg/ml of trypsin  
130 (Promega) in 50 mM ammonium bicarbonate (NH<sub>4</sub>HCO<sub>3</sub>) (Sigma). The peptides were extracted with 25 mM  
131 ammonium bicarbonate for 15 min, dehydrated with acetonitrile (ACN) (Sigma), incubated with 5% formic acid  
132 (Aldrich) with shaking for 15 min. Drying of samples was performed again with ACN was accomplished via  
133 vacuum centrifugation (Ali et al. 2012). The pellets were resuspended in formic acid / acetonitrile / H<sub>2</sub>O (volume  
134 proportion, 35/50/15%). 1 µl of the peptides suspension was mixed 3 µl of 2, 5-dihydroxybenzoic acid (DHB)  
135 matrix and spotted onto the MALDI-TOF target. The air-dried samples were then analyzed on a MALDI-TOF  
136 MS (Bruker Daltonics) for identification.

### 137 **Multiangle Static Light Scattering (MALS) / refractometry characterization**

138 The integrity and quaternary structure of purified Gtf3 with and without NVoy polymer were analyzed by the  
139 combination of UV spectrophotometry, multiangle static light scattering (MALS), and refractometry, coupled  
140 on-line with an analytical size exclusion chromatography (SEC) column. Analytical SEC was carried with an  
141 HPLC-Alliance 2695 system (Waters) on a 15-ml KW-804 column (Shodex) at a flow rate of 0.5 ml.min<sup>-1</sup>, UV  
142 absorbance was detected using photodiode array detector (2996; Waters), MALS detection was performed using

143 a MiniDawn Treos (Wyatt Technology), and refractometry measurement was achieved with a differential  
144 refractometer (Optilab rEX; Wyatt Technology) (Veesler et al. 2009). Indeed, Multiangle static light scattering  
145 (MALS) measures the absolute molecular weight of injected sample and is connected to a Quasi-Elastic Light  
146 scattering detector (QELS/DLS), Dynapro Wyatt, for the measurement of hydrodynamic radius (Rh). These two  
147 detectors are coupled to an HPLC (High Performance Liquid Chromatography) system that comprises two main  
148 types of detectors, a UV-visible detector that measures the light absorption by the sample at the exit of the  
149 column and a differential refractometer. Optilab rEX Wyatt measures the variation of the refractive index (RI) of  
150 the solution at the exit of the column, which allows determining of the protein sample concentration in a similar  
151 manner with the UV-visible detector. In order to compare the oligomerization status of the Gtf3 purified protein  
152 and Gtf3 with and without NVoy polymer, 30  $\mu$ l of each sample at a concentration of 8.46 mg/ml and 3.81  
153 mg/ml were injected onto the KW804 column (Shodex). Before both MALS and CD measurements, samples  
154 were filtered through 0.22  $\mu$ m pore size Millex syringe filter (Millipore Corp) and used buffers were identical to  
155 gel filtration buffer (see above). The oligomery of Gtf3 was calculated using the software program provided by  
156 the manufacturer.

#### 157 **Circular dichroism (CD)**

158 The CD spectrum was recorded on a Jasco J-180 spectropolarimeter, deconvolved using CDNN *CD spectra*  
159 software. In addition, the percentages of  $\beta$ -strand,  $\alpha$ -helix, turns, and random coils of Gtf3 protein were  
160 determined by the *CDNN CD spectra software*. CD spectra of purified (final concentration 0.2 mg/ml) and  
161 solubilized Gtf3 (in 300  $\mu$ l of a buffer containing 20 mM HEPES pH 7.5, 150 mM NaCl, 5 mM  $\beta$ -  
162 mercaptoethanol) were achieved at 20°C in the wavelength range of 190-260 nm. Data processing was done with  
163 the Dichroweb software ([www.dichroweb.cryst.bbk.ac.uk/html/process.shtml](http://www.dichroweb.cryst.bbk.ac.uk/html/process.shtml)) (Vincentelli et al. 2004).

#### 164 **Crystallogensis**

165 The purpose of the screening of crystallization conditions is to determine all the physical conditions  
166 (temperature, volume of the drop, reservoir, etc.) and chemical conditions (protein concentration, type and  
167 concentration of precipitating agent, pH, etc.), that will induce the formation of some crystalline nuclei, and then  
168 their growth. We carried out a screening by nanodroplet method using the 20°C vapour diffusion technique  
169 (Sulzenbacher et al. 2002). For this purpose, we used GREINER 96-well crystallization plates (Greiner Bio-one),  
170 containing three drop *wells per* reservoir *well* at volumes of 100, 200, and 300 nl for each condition. Greiner  
171 plate reservoirs were filled with a TECAN robot and dispensing nanoliter droplets was performed by a Cartesian  
172 robot (Dolzan et al. 2004). The Wizard Screen I (Emerald Bio-Structures), Stura and MDL commercial kits were



173 used for screening of Gtf3 crystallization conditions with NVoy polymer. This crystallogensis method using  
174 nano-drop robotics was previously described by Sulzenbacher *et al.* (Sulzenbacher et al. 2002).

175

## 176 **Results**

### 177 **Bioinformatics study**

#### 178 **Hydrophobic Cluster Analysis (HCA)**

179 Medor program is used mainly to identify, from the protein primary sequence, the two-dimensional folding  
180 signatures (secondary structures), and also to visualize structured, non-structured and/or poorly structured  
181 globular regions in the protein. In addition, it gives the analysis of hydrophobic clusters. Hydrophobic clusters of  
182 Gtf3 are distributed along the protein sequence (**Fig. 1**). These hydrophobic clusters are derived from the  
183 formation of secondary structures,  $\alpha$ -helix and  $\beta$ -sheets. Although hydrophobic amino acids (V, I, L, M, Y, W,  
184 F) belong mainly to regular secondary structures and participate to the densely packed cores of globular  
185 domains, some proteins have exposed hydrophobic patches, which are stabilized by interactions (i.e., either with  
186 partner proteins or to form oligomers) (Bitard-Feildel et al. 2018). Among 422 amino acids of the full-length  
187 Gtf3 protein, 130 amino acids are hydrophobic (30%). Furthermore, based on known protein structures sharing  
188 sequence similarities with Gtf3, the prediction of secondary structures performed by the Phyre2 program showed  
189 that there are probably also hydrophobic patches on the surface of Gtf3 protein that can generate intermolecular  
190 interactions and form aggregates (**Fig. 2 A & B**). Altogether, the high percentage of hydrophobic residues of the  
191 sequence and the significant hydrophobic patches on the surface of Gtf3 predict likely an aggregation propensity  
192 and a low solubilization of this protein when overexpressed.

#### 193 **Prediction of secondary structures**

194 The top ranking structural model of Gtf3 found by Phyre2 is Vancosaminyltransferase GtfD of *Amycolatopsis*  
195 *orientalis* (PDB code: 1rrv) with 100.0% of confidence and 22% of sequence identity. The result of this  
196 predicted model shows the presence of secondary structures mainly  $\alpha$ -helix and  $\beta$ -sheets (**Fig. 2 A**). The primary  
197 sequence of Gtf3 protein contains 6 cysteines of which 4 are on the surface and exposed to the solvent according  
198 to Phyre2 prediction (**Fig. 2 C & D**). The presence of exposed cysteine residues may lead to *intermolecular*  
199 disulfide bridges, requiring the usage of reducing agents, such as  $\beta$ -mercaptoethanol and DTT (dithiothreitol)  
200 throughout expression, purification, and biochemical characterization of Gtf3 to prevent its aggregation.

#### 201 **Prediction of oligomeric state**

202 Sequence analysis using ROBETTA and SWISS-MODEL web servers predicted Gtf3 to be a monomeric  
203 protein. The top ranking structural model found in this prediction was glycosyltransferase GtfA from  
204 [Actinoplanes teichomyceticus](#) and [Amycolatopsis orientalis](#) (PDB code: 3H4I) with 30.42% of sequence identity  
205 and 90% of sequence coverage.

## 206 **Multiple alignment**

207 Multiple sequence alignment analysis has been achieved with Gtf3 homologs from GT1 superfamily (CAZy  
208 classification) which have been characterized to have a rhamnosyltransferases activity (**Fig. 3**). This analysis  
209 revealed the existence of fairly conserved motif (HHxxAG) among GTs sequences and was superimposed on the  
210 motif of the closest model to Gtf3 protein according to Phyre2 secondary structure prediction program which is  
211 HHxxAGT. In the structural model, this motif has been reported to be localized in a loop between the domains  
212 and is involved in the interaction with the donor nucleotide-sugar, TDP-*epi*-vancosamine (GtfA) (Mulichak et al.  
213 2003). Therefore the conserved motif can interact with the nucleotide diphosphate of the donor substrate of Gtf3  
214 to NDP-3,4-di-*O*-Me-Rhamnose and NDP-3-*O*-Me-Rhamnose. Furthermore, another motif G(T/S)RGD was  
215 highly conserved throughout the rhamnosyltransferases sequences and was suggested to be the potential catalytic  
216 base in GtfD enzyme (Chen et al. 2009). Multiple sequence alignment revealed also that Asp 348 was the  
217 negatively charged residue which was highly conserved among these sequences of the same family. This residue  
218 could be involved in the inversion catalytic mechanism by a nucleophilic attack at the binding site to the  
219 acceptor substrate of the Gtf3 N-terminal domain. Absolutely conserved hydrophobic residues have been  
220 identified, Leu 24, Gly 337, Pro 339, Leu 341, and Gly 362, which could have a structural role in the active site  
221 of Gtf3 enzyme and/or in protein oligomerization.

## 222 **Protein expression and purification**

223 As indicated in **Fig. 4**, the recombinant Gtf3 protein was overexpressed and detected in both insoluble and  
224 soluble fractions. Then, it was purified by His-Trap HP 5-ml column affinity chromatography followed by size  
225 exclusion chromatography (SEC). The SEC chromatogram showed 2 peaks, corresponding to proteins eluted at  
226 113 ml, which is the void volume ( $V_0$ ) and 170 ml, as shown by the elution profile (**Fig. 5**). Fractions containing  
227 proteins were separated and analyzed by SDS-PAGE (**Fig. 6 A**). We used ExPASy server to compute the  
228 theoretical isoelectric point ( $pI$ ) and molecular weight (MW) of recombinant Gtf3, which are 5.96 and 46.02  
229 kDa, respectively. SDS-PAGE displayed abundant protein bands with apparent MW of 46 kDa, corresponding to  
230 the calculated mass of recombinant Gtf3. Abundant protein bands were excised from the gel and submitted to  
231 mass spectrometry (MS) analysis, confirming that the detected protein corresponds to Gtf3. This result  
232 demonstrates that both 1<sup>st</sup> and 2<sup>nd</sup> peaks of gel filtration (**Fig. 5**) contain Gtf3. The 1<sup>st</sup> peak being eluted in the  
233 void volume, means that Gtf3 in this peak is highly aggregated. The 2<sup>nd</sup> peak corresponds to a mass of 275 kDa  
234 as compared to gel filtration calibration (**Fig. 6 B and C**), which is approximately equivalent to six fold the mass  
235 of Gtf3 ( $6 \times 46 \text{ kDa} = 276 \text{ kDa}$ ). Therefore, Gtf3 is very likely a hexamer at the outlet of gel filtration with a low

236 proportion of aggregates. This result is consistent with Hydrophobic Cluster Analysis prediction. The quantity of  
237 produced Gtf3 was estimated using a NanoDrop<sup>TM</sup> 1000 spectrophotometer (Thermo Fisher Scientific) and  
238 yielded 3.17 mg per liter of bacterial culture. Our results indicate that our expression system is functional,  
239 although it might need further optimization but it is not our focus in this study. Protein profiles were analyzed  
240 using ImageQuant<sup>TM</sup> TL software to determine the relative abundance of each band. The purity of Gtf3 in the  
241 different fractions was greater than 90%, and was considered sufficiently pure for downstream biochemical  
242 characterization (**Table 2**).

### 243 **Secondary structure of Gtf3**

244 We wanted to confirm bioinformatic prediction and quality of the purified Gtf3, so we performed circular  
245 dichroism spectroscopy (CD) analysis. The experimental spectrum is typical of a protein organized into  $\beta$ -sheets  
246 (largely negative ellipticity between 216 and 222 nm) and  $\alpha$ -helices (positive ellipticity between 180 and 200  
247 nm) with a more or less noisy spectrum in this measurement range; these results are compared to the CD  
248 reference spectrum (**Fig. 7**). This result seems to be in agreement with the predictions of secondary structures.

249 To determine the proportion of each type of secondary structure, the experimental spectrum was analyzed in its  
250 elementary components and deconvoluted with the Dichroweb software. The CD results showed 49% of  $\alpha$ -helix,  
251 25% of  $\beta$ -sheets, 8% of loops and 20% of disordered structures. This CD experimental result was also in  
252 agreement with the Phyre2 secondary structure prediction (**Fig. 1 and 2**).

### 253 **Oligomerization status of Gtf3**

254 Both *in silico* prediction and the SEC result showed a tendency of Gtf3 to aggregate, although the protein keeps  
255 its folding integrity. In order to prevent this aggregation we mixed purified Gtf3 with NVoy (with 5-fold mass  
256 excess than Gtf3). This latter is a long polymer of 5 kDa, which has been shown to bind surface hydrophobic  
257 regions of target proteins and prevent the aggregation without affecting their active site (Klammt et al. 2011). In  
258 addition, we screened several buffers with different pH and salt concentrations, in the presence of NVoy, to  
259 optimize solubilization of Gtf3 prior to analysis (**Supplementary Material, Table S1 and S2**). The selected and  
260 used buffer in expression, purification and biochemical characterization experiments of Gtf3 was (Tris pH 7.5  
261 and 150 mM NaCl). SEC-MALS (Size Exclusion Chromatography coupled to MultiAngle static Light  
262 Scattering) analysis allowed us to measure the masses of the Gtf3 protein coupled to NVoy polymer (**Fig. 8**).  
263 Gtf3 in absence of NVoy displayed the mass of 230 kDa, corresponding likely to pentameric protein. This result  
264 is different to that obtained by size exclusion chromatography (SEC) in which the aggregate of Gtf3 was a  
265 hexamer (**Fig. 5**). This discrepancy is due to protein properties (e.g., geometry) and molecules in the buffer (e.g.,

266 salt concentration) may interfere with determining the real molecular weight. For instance, a fibrillar protein  
267 (elongated shape) and a globular protein of the same mass will not behave the same way through the column  
268 (Breton et al. 2006; Burgess 2018). However, in the presence of NVoy, Gtf3 had the molecular weight of 92.77  
269 kDa. This molecular weight is unlikely a dimer ( $2 \times 46$  kDa) because of the presence of NVoy. Thus, It is likely  
270 a monomer with nine molecules of NVoy polymer ( $46 + 5 \times 9$  kDa). In addition, this result is in accordance with  
271 the oligomeric state prediction.

## 272 **Crystallogenesis**

273 We performed the crystallization tests of Gtf3, purified in the presence of NVoy, at a concentration of 10.61  
274 mg/ml. Gtf3 complexed with the NVoy polymer formed small and rod shaped crystals. They were obtained at a  
275 concentration of 8.23 mg/ml in a solution of the wizard kit of 2.5 M NaCl, 0.1 M Tris pH7 and 0.2 MgCl<sub>2</sub> (**Fig.**  
276 **9**). Our result indicates that, in addition to being highly beneficial in preventing aggregates, NVoy is not  
277 interfering with crystallogenesis steps. Moreover, obtaining crystals of Gtf3 in the presence of Nvoy confirms  
278 the MALS results indicating that NVoy polymer contributes to stabilizing Gtf3 protein in monomeric form  
279 which is monodisperse in crystallization buffer solution. In order to obtain larger crystals, we optimized  
280 crystallization conditions by varying pH values and concentration of the precipitating agent. However, we were  
281 unable to have better results.

## 282 **Discussion**

283 In the present study, Gtf3 from *Mycobacterium smegmatis* was expressed using *E. coli* expression system.  
284 Expressed recombinant Gtf3 protein was more abundant in the insoluble fraction than in the soluble fraction  
285 (**Fig. 4**). In addition, Gtf3 protein has a great tendency to aggregate. Indeed, one of the contributing factors to  
286 protein aggregation is the interaction of the exposed hydrophobic patches. We therefore chose to investigate the  
287 use of the NVoy polymer on expression of Gtf3 in *E. coli*. Moreover, NVoy polymer was observed not to block  
288 the synthesis and to favor solubility of recombinant expressed proteins in the reaction mixture of wheat germ  
289 cell-free expression system (Guild et al. 2011). This polymer was added to the purified Gtf3 because the protein  
290 of interest, tends to aggregate and form an oligomeric aggregate. This was predicted by hydrophobic cluster  
291 analysis (presence of hydrophobic patches). This tendency has also been shown by the biochemical  
292 characterization of the oligomerization state. The experimental characterization by SEC and MALS allowed to  
293 reveal that the expressed recombinant protein Gtf3 alone in *E. coli* was an oligomer and that NVoy polymer  
294 dissociated the oligomer into monomer. This was in agreement with the prediction of oligomeric state.  
295 Nonetheless, further experiments are needed to determine the Gtf3/NVoy ratio. The result indicates that NVoy

296 helps to solubilize and stabilize Gtf3 in monomeric forms unlike the pentamer/hexamer form obtained in the  
297 absence of NVoy. Altogether, our result suggests that NVoy polymer interacts with surface exposed hydrophobic  
298 patches on the protein, thereby limiting nonspecific interactions which can cause the aggregation and dissociate  
299 the aggregate of Gtf3 protein.

300 Concerning GTs structures, difficulties with high-level expression, purification, and crystallization, as well as the  
301 ratio of loops vs secondary elements (which is high in GTs) hamper the resolution of crystal structure of these  
302 enzymes (Breton et al. 2006; Schmid et al. 2016). GTs have very high donor and acceptor substrate specificities  
303 and are in general limited to the establishment of one glycosidic linkage. In addition, interactions between the  
304 sugar-nucleotide donor and a few protein residues seem to determine the specificity of the glycosyltransferases  
305 for their donor substrate. It has been reported that most of the loops, which are involved in donor substrate  
306 binding, are highly flexible and induce conformational changes (open and closed active conformations) (Qasba  
307 et al. 2005). This feature leads to a low electron density, thus limiting the detailed description of the catalytic  
308 domains (Schmid et al. 2016). Nevertheless, crystal structure of Gtf3 protein from *Streptococcus parasanguinis*  
309 has been reported (Zhu et al. 2011). Besides, this recombinant Gtf3 protein from *S. parasanguinis* was  
310 expressed, purified, and crystallized in *E. coli* in a soluble fraction (Zhu et al. 2013). The *crystal structure of this*  
311 *native Gtf3* was solved in a *tetrameric* form sharing a structural similarities with GTs from GT4, GT5, and GT20  
312 subfamilies (Zhu et al. 2011). Additionally, Zhu *et al.* have identified the key residues and domains involved in  
313 UDP- or UDP-glucose substrate binding and in Gtf3 function and oligomerization, respectively (Zhu et al.  
314 2011). Despite the low homology between Gtf3 from *M. smegmatis* belonging to GT1 family and Gtf3 from *S.*  
315 *parasanguinis* belonging to non-classified GT family (CAZy classification), oligomerization status can be  
316 compared between these proteins using these biochemical quality control methods. Beside further structural  
317 studies, several important scientific questions remain unanswered in this case study and require future  
318 investigations concerning, i.e., a functional characterization to determine interaction of nucleotides (TDP, GDP,  
319 UDP, and ADP) with Gtf3 as well as to study the interaction with bivalent ions such as MgCl<sub>2</sub>, MnCl<sub>2</sub> by  
320 conducting fluorescence quenching experiments. These experiments will allow us to determine the dissociation  
321 constants (K<sub>D</sub>) for each potential ligand, and provide information on the enzymatic kinetics of Gtf3. The result of  
322 this study will allow designing the methylated nucleotide-rhamnose donor in order to carry out experiments in  
323 solution, allowing determination of which rhamnose is transferred to GPLs. Furthermore, it will contribute to  
324 determine the biological function of Gtf3, and provide a better understanding of the catalytic mechanism of this  
325 enzyme.

326 This biochemical study includes size exclusion chromatography (SEC), circular dichroism (CD),  
327 MALS/UV/refractometry/SEC analysis, and crystallogenesis. Quality control procedures were carried out to,  
328 respectively; assess molecular weight, secondary structure, aggregation status, and homogeneity of Gtf3. In  
329 addition, these methods can be applied to any protein, which is a prerequisite in post-genomics era, dealing  
330 mainly with proteins having an unknown function. This will contribute to face the challenges related to different  
331 applications of proteins expressed in clinical, biotechnology, scientific research, and industry.

332

### 333 **Acknowledgments**

334 We would like to thank Dr. Yves Bourne from AFMB laboratory facilities, Dr. Badreddine Douzi and Dr.  
335 Renaud Vincentelli for support in protein expression and purification, Dr. Silvia Spinelli for crystallization  
336 facilities as well as Stéphanie Blangy and Dr. David Veessler for support in MALS/UV/refractometry/SEC.

337

338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365

## References

- Ali ZM, Bakli M, Fontaine A, Bakkali N, Hai VV, Audebert S, Boublik Y, Pagès F, Remoué F, Rogier C (2012) Assessment of *Anopheles* salivary antigens as individual exposure biomarkers to species-specific malaria vector bites. *Malaria journal* 11 (1):439
- Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic acids research* 42 (W1):W252-W258
- Billman-Jacobe H (2004) Glycopeptidolipid synthesis in mycobacteria. *Current Science*:111-114
- Bitard-Feildel T, Lamiable A, Mornon JP, Callebaut I (2018) Order in disorder as observed by the “hydrophobic cluster analysis” of protein sequences. *Proteomics* 18 (21-22):1800054
- Bramucci E, Paiardini A, Bossa F, Pascarella S (2012) PyMod: sequence similarity searches, multiple sequence-structure alignments, and homology modeling within PyMOL. *BMC bioinformatics* 13 (4):S2
- Brennan PJ, Crick DC (2007) The cell-wall core of *Mycobacterium tuberculosis* in the context of drug discovery. *Current topics in medicinal chemistry* 7 (5):475-488
- Breton C, Šnajdrová L, Jeanneau C, Koča J, Imberty A (2006) Structures and mechanisms of glycosyltransferases. *Glycobiology* 16 (2):29R-37R
- Burgess RR (2018) A brief practical review of size exclusion chromatography: Rules of thumb, limitations, and troubleshooting. *Protein expression and purification* 150:81-85
- Chen Y-L, Chen Y-H, Lin Y-C, Tsai K-C, Chiu H-T (2009) Functional characterization and substrate specificity of spinosyn rhamnosyltransferase by in vitro reconstitution of spinosyn biosynthetic enzymes. *Journal of Biological Chemistry* 284 (11):7352-7363
- Deller MC, Kong L, Rupp B (2016) Protein stability: a crystallographer's perspective. *Acta Crystallographica Section F: Structural Biology Communications* 72 (2):72-95
- Deshayes C, Laval F, Montrozier H, Daffe M, Etienne G, Reyat JM (2005) A glycosyltransferase involved in biosynthesis of triglycosylated glycopeptidolipids in *Mycobacterium smegmatis*: impact on surface properties. *J Bacteriol* 187 (21):7283-7291. doi:10.1128/JB.187.21.7283-7291.2005
- DiMaio F, Leaver-Fay A, Bradley P, Baker D, André I (2011) Modeling symmetric macromolecular structures in Rosetta3. *PloS one* 6 (6)



366 Dolzan M, Johansson K, Roig-Zamboni V, Campanacci V, Tegoni M, Schneider G, Cambillau C (2004) Crystal  
367 structure and reactivity of YbdL from *Escherichia coli* identify a methionine aminotransferase function.  
368 FEBS letters 571 (1-3):141-146

369 Guild K, Zhang Y, Stacy R, Mundt E, Benbow S, Green A, Myler PJ (2011) Wheat germ cell-free expression  
370 system as a pathway to improve protein yield and solubility for the SSGCID pipeline. Acta  
371 Crystallographica Section F: Structural Biology and Crystallization Communications 67 (9):1027-1031

372 Gutiérrez AV, Viljoen A, Ghigo E, Herrmann J-L, Kremer L (2018) Glycopeptidolipids, a double-edged sword  
373 of the *Mycobacterium abscessus* complex. Frontiers in microbiology 9

374 Jeevarajah D, Patterson JH, McConville MJ, Billman-Jacobe H (2002) Modification of glycopeptidolipids by an  
375 O-methyltransferase of *Mycobacterium smegmatis*. Microbiology 148 (10):3079-3087.  
376 doi:doi:10.1099/00221287-148-10-3079

377 Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE (2015) The Phyre2 web portal for protein  
378 modelling, prediction and analysis. Nature protocols 10 (6):845-858. doi:10.1038/nprot.2015.053

379 Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. Nucleic  
380 acids research 32 (suppl\_2):W526-W531

381 Klammt C, Perrin MH, Maslennikov I, Renault L, Krupa M, Kwiatkowski W, Stahlberg H, Vale W, Choe S  
382 (2011) Polymer-based cell-free expression of ligand-binding family B G-protein coupled receptors  
383 without detergents. Protein Sci 20 (6):1030-1041. doi:10.1002/pro.636

384 Kushwaha M, Salis HM (2015) A portable expression resource for engineering cross-species genetic circuits and  
385 pathways. Nature communications 6:7832

386 Laemmli UK (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. nature  
387 227 (5259):680-685

388 Lairson L, Henrissat B, Davies G, Withers S (2008) Glycosyltransferases: structures, functions, and mechanisms.  
389 Annual review of biochemistry 77

390 Lieutaud P, Canard B, Longhi S (2008) MeDor: a metaserver for predicting protein disorder. BMC genomics 9  
391 (2):S25

392 Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes  
393 database (CAZy) in 2013. Nucleic acids research 42 (D1):D490-D495

394 Matsui D, Nakano S, Dadashpour M, Asano Y (2017) Rational identification of aggregation hotspots based on  
395 secondary structure and amino acid hydrophobicity. Scientific reports 7 (1):9558

396 Miyamoto Y, Mukai T, Nakata N, Maeda Y, Kai M, Naka T, Yano I, Makino M (2006) Identification and  
397 characterization of the genes involved in glycosylation pathways of mycobacterial glycopeptidolipid  
398 biosynthesis. *J Bacteriol* 188 (1):86-95. doi:10.1128/JB.188.1.86-95.2006

399 Mukherjee R, Chatterji D (2005) Evaluation of the role of sigma B in *Mycobacterium smegmatis*. *Biochemical*  
400 *and biophysical research communications* 338 (2):964-972

401 Mulichak AM, Losey HC, Lu W, Wawrzak Z, Walsh CT, Garavito RM (2003) Structure of the TDP-epi-  
402 vancosaminyltransferase GtfA from the chloroeremomycin biosynthetic pathway. *Proceedings of the*  
403 *National Academy of Sciences* 100 (16):9238-9243

404 Qasba PK, Ramakrishnan B, Boeggeman E (2005) Substrate-induced conformational changes in  
405 glycosyltransferases. *Trends in biochemical sciences* 30 (1):53-62

406 Ren B, Pham TM, Surjadi R, Robinson CP, Le T, Chandry P, Peat TS, McKinstry WJ (2013) Expression,  
407 purification, crystallization and preliminary X-ray diffraction analysis of a lactococcal bacteriophage  
408 small terminase subunit. *Acta Crystallographica Section F: Structural Biology and Crystallization*  
409 *Communications* 69 (3):275-279

410 Rosano GL, Ceccarelli EA (2014) Recombinant protein expression in *Escherichia coli*: advances and challenges.  
411 *Frontiers in microbiology* 5:172

412 Schmid J, Heider D, Wendel NJ, Sperl N, Sieber V (2016) Bacterial glycosyltransferases: challenges and  
413 opportunities of a highly diverse enzyme class toward tailoring natural products. *Frontiers in*  
414 *microbiology* 7:182

415 Schorey JS, Sweet L (2008) The mycobacterial glycopeptidolipids: structure, function, and their role in  
416 pathogenesis. *Glycobiology* 18 (11):832-841

417 Schuman B, Alfaro JA, Evans SV (2006) Glycosyltransferase structure and function. In: *Bioactive*  
418 *Conformation I*. Springer, pp 217-257

419 Sulzenbacher G, Gruez A, Roig-Zamboni V, Spinelli S, Valencia C, Pagot F, Vincentelli R, Bignon C, Salomoni  
420 A, Grisel S (2002) A medium-throughput crystallization approach. *Acta Crystallographica Section D:*  
421 *Biological Crystallography* 58 (12):2109-2115

422 Thayer KM (2016) Structure prediction and analysis of neuraminidase sequence variants. *Biochemistry and*  
423 *Molecular Biology Education* 44 (4):361-376

424 Vandermies M, Fickers P (2019) Bioreactor-Scale Strategies for the Production of Recombinant Protein in the  
425 Yeast *Yarrowia lipolytica*. *Microorganisms* 7 (2):40

426 Veesler D, Blangy S, Siponen M, Vincentelli R, Cambillau C, Sciara G (2009) Production and biophysical  
427 characterization of the CorA transporter from *Methanosarcina mazei*. Analytical biochemistry 388  
428 (1):115-121

429 Vincentelli R, Canaan S, Campanacci V, Valencia C, Maurin D, Frassinetti F, Scappucini-Calvo L, Bourne Y,  
430 Cambillau C, Bignon C (2004) High-throughput automated refolding screening of inclusion bodies.  
431 Protein Science 13 (10):2782-2792

432 Warne NW, Mahler H-C (2018) Challenges in Protein Product Development, vol 38. Springer,

433 Young CL, Britton ZT, Robinson AS (2012) Recombinant protein expression and purification: a comprehensive  
434 review of affinity tags and microbial applications. Biotechnology journal 7 (5):620-634

435 Zhu F, Erlandsen H, Ding L, Li J, Huang Y, Zhou M, Liang X, Ma J, Wu H (2011) Structural and functional  
436 analysis of a new subfamily of glycosyltransferases required for glycosylation of serine-rich  
437 streptococcal adhesins. Journal of Biological Chemistry 286 (30):27048-27057

438 Zhu F, Wu R, Zhang H, Wu H (2013) Structural and biochemical analysis of a bacterial glycosyltransferase. In:  
439 Glycosyltransferases. Springer, pp 29-39

440

441 **Figure captions:**

442 **Figure 1. Hydrophobic Cluster Analysis "HCA" plot of Gtf3 protein sequence.** The hydrophobic residues  
443 are grouped into clusters. The secondary structures are represented ( $\alpha$ -helix and  $\beta$ -sheets). Red stars, squares and  
444 lozenges represent, respectively, prolines, serines/threonines and glycines residues.

445  
446 **Figure 2. Secondary structure prediction of Gtf3 protein.** The protein structure prediction was performed  
447 using Phyre2 web server and visualized using Pymol 2.3 molecular graphics software. **A.** Secondary structure  
448 prediction of Gtf3 protein showing  $\alpha$ -helix (red),  $\beta$ -sheets (yellow) and loops (green). **B.** Surface of the Gtf3  
449 model predicted by the Phyre2 showing hydrophobic patches (orange regions). **C&D.** Structure surface from  
450 different sides of Gtf3 model predicted by the Phyre2 program. The yellow color represents the cysteines  
451 exposed to the solvent.

452  
453 **Figure 3. Multiple amino acid sequence alignment of rhamnosyltransferases from CAZy GT1 family.** GTs  
454 shown in the analysis are as follows (Uniprot accession number in parentheses): putative glycosyltransferase,  
455 Gtf3 (AAN28688), putative glycosyltransferase (ADU85989), putative glycosyltransferase (ADU86026), RtfA  
456 (AAD44209), putative glycosyltransferase (AAC71701), rhamnosyltransferase (AAC71702), possible glycosyl  
457 transferase (CAB05415), rhamnosyltransferase chain B (AAG06866), probable NDP-rhamnosyltransferase  
458 (AAG23268), L-rhamnosyltransferase (ABL09968), elloramycin glycosyltransferase (CAC16413), and glycosyl  
459 transferase (CAJ42338). The indicated species between brackets are in the order: *Mycobacterium smegmatis*,  
460 *Dactylosporangium aurantiacum*, *Dactylosporangium aurantiacum*, *Mycobacterium avium*, *Mycobacterium*  
461 *avium*, *Mycobacterium avium*, *Mycobacterium tuberculosis*, *Pseudomonas aeruginosa*, *Saccharopolyspora*  
462 *spinosa*, *Streptomyces echinatus*, *Streptomyces olivaceus*, and *Streptomyces steffisburgensis*. Identical amino  
463 acids are shown with a black background, and similar residues are shown with a gray background. The Clustal  
464 consensus symbols indicate the amount of conservation (\*: Exact, ': Conserved Substitution, .: Semi-  
465 conserved substitution).

466  
467 **Figure 4. SDS-PAGE gel of expressed and purified recombinant Gtf3 protein.** Gtf3 recombinant protein  
468 samples from *M. smegmatis* expressed in *E. coli* expression system and purified by affinity chromatography  
469 were separated on 12% SDS-PAGE and post-stained with Imperial™ Protein Stain (Thermo Scientific). Lane L,  
470 molecular weight marker (BenchMark™ Protein Ladder). Lane 1, insoluble fraction (pellet). Lane 2, soluble  
471 fraction (supernatant). Lane 3, the recombinant Gtf3 purified by His-Trap HP 5-ml column affinity  
472 chromatography. Arrow shows Gtf3 protein bands.

473  
474 **Figure 5. Gel filtration chromatography elution profile of Gtf3 protein.** Gel filtration of Gtf3 on an HiLoad  
475 26/60 Superdex 75 pg column. The 1st peak (fractions A2-A8) represents the aggregated protein and the 2nd  
476 peak (fractions B6-D6) hexameric Gtf3. Elution volume of Gtf3 protein was 170 ml.

477  
478 **Figure 6. a.** SDS-PAGE gel of purified recombinant Gtf3 protein stained with Coomassie blue. Gtf3 purified by  
479 both affinity and gel filtration chromatographies, were separated on 12% SDS-PAGE and post-stained with  
480 Imperial™ Protein Stain (Thermo Scientific). Five microliter of each collected fraction were loaded per well.  
481 Lane L, molecular weight marker (Pierce Unstained Protein MW Ladder). Other wells, Peak fractions (Lanes 1–  
482 9) collected from gel filtration chromatography. Lane 1 corresponds to pooled fractions from the 1st peak. Lane  
483 2 corresponds to pooled fractions (A9-B5) between the 1st and the 2nd peak. Lanes 3-9 correspond to fractions  
484 from the 2nd peak. MW: molecular weight. kDa: kiloDalton. **b.** Calibration of gel filtration column. Protein  
485 standards of known molecular weight were used to calibrate the 26/60 Superdex 75 pg gel filtration column; i.e.,  
486 blue dextran (2000 kDa; peak 1), albumin (67 kDa; peak 2), ovalbumin (43 kDa peak 3), chymotrypsinogen (25  
487 kDa; peak 4) and ribonuclease A (13.7 kDa; peak 5). **c.** Linear regression analysis of the gel filtration calibration.  
488  $K_{av} = (V_e - V_o) / (V_T - V_o)$ ,  $V_e$  is the elution volume,  $V_T$  and  $V_o$  are the total liquid volume (320 ml) and the void  
489 volume of the column (113 ml), respectively.

490  
491 **Figure 7. Circular dichroism (CD) spectrum of Gtf3 protein.**

492  
493 **Figure 8. MALS/UV/refractometry/SEC analysis of Gtf3 protein using a KW-804 column.** The left y-axis  
494 represents the molar mass, and the the right y-axis represents the absorbance at 280 nm according to the retention  
495 volume of the column (x-axis). The colored plots represent the measured molecular weights of pentameric Gtf3  
496 (red line), monomeric Gtf3 + NVoy (blue line). Values of the measured masses at the volume corresponding to  
497 the base of each peak are reported according to the same color scheme.

498

499 **Figure 9. Crystals of Gtf3 protein.** Crystals of Gtf3 with NVoy polymer in a drop of crystallization plates  
500 (Greiner Bio-one), obtained in conditions: 2.5 M NaCl, 0.1 M Tris pH 7 and 0.2 M MgCl<sub>2</sub>.