Identification of protein partners in Mycobacteria using a single step affinity purification method

Przemysław Płociński ‡ *, Daniel Laubitz ‡ *, Dominik Cysewski ‡, Krystian Stoduś ‡, Katarzyna Kowalska ‡ and Andrzej Dziembowski ‡ ¶ #.

From the ‡ Institute of Biochemistry and Biophysics Polish Academy of Sciences, Pawińskiego 5a, 02-106 Warszawa, Poland; ¶ Department of Genetics and Biotechnology, Faculty of Biology, University of Warsaw, Warsaw, Poland

* P.P and D.L. contributed equally to this work

Correspondence to:

Institute of Biochemistry and Biophysics Polish Academy of Sciences, Pawińskiego 5a, 02-106 Warszawa, Poland. Tel.: +48 22 5922033; Fax: +48 22 5922190; E-mail: andrzejd@ibb.waw.pl

ABSTRACT

Tuberculosis is a leading cause of death in developing countries. Efforts are being made to both prevent its spreading and improve curability rates. Understanding the biology of the bacteria causing the disease, *Mycobacterium tuberculosis* (*M. tuberculosis*), is thus vital. We have implemented improved screening methods for protein-protein interactions based on affinity purification followed by high-resolution mass spectrometry. This method is attractively applicable to both medium- and high-throughput studies aiming to characterize protein-protein interaction networks of tubercle bacilli. From four tested epitopes, FLAG, eGFP, Protein A, and hemagglutinin, the eGFP tag was found most useful based on easily monitored expression and as a simultaneous tool for sub-cellular localization studies. It presents a relatively low background with cost effective purification.

RNA polymerase subunit A (RpoA) was used as a model for investigation of a large protein complex. When used as a bait, it co-purified with all remaining RNA polymerase core subunits as well as many accessory proteins. The amount of RpoA strongly correlated with the amount of quantification peptide used as part of the tagging system in this study (SH), making it applicable for semi-quantification studies. Interactions between the components of the RpoA-eGFP protein complex were further confirmed using protein cross-linking.

Dynamic changes in the composition of protein complexes under induction of UV damage were observed when UvrA-eGFP expressing cells, treated with UV light were used to co-purify UvrA interaction partners.

1. Introduction

Mycobacterium tuberculosis, the causative agent of tuberculosis (TB), is a deadly human pathogen, and its infection is an epidemic in many developing countries. The latest WHO report states that 8.7 million new cases of TB occurred in 2011. Moreover, TB threatens the lives of HIV positive individuals, killing 4,30,000 of HIV-positive patients in 2011. Emergence of multidrug-resistant (MDR-TB) and totally drug-resistant (TDR-TB) strains of *M. tuberculosis* creates an urgent need for profound investigation of the tubercle bacilli's physiology and pathogenicity. Understanding its biology is fundamental for developing new effective strategies to combat TB. Genomic and proteomic methods are being utilized to broaden this knowledge and to understand the network of protein-protein interactions for a variety of organisms, including pathogenic bacteria, to understand the regulation and dynamics of important cellular functions and processes including DNA replication, transcription, and virulence.

Recent proteogenomic analysis identified 3,176 proteins from *M. tuberculosis* representing c.a. 80% of its total predicted number of genes (1). Protein-protein interaction studies, which are crucial for understanding of many biological processes, are currently not performed to a satisfactory extent. Most often, protein-protein interactions are determined by researches only for very specific biological processes, and global protein-protein interaction networks of only a few model organisms have been investigated based on medium- or high-throughput experiments. These organisms include *Mycoplasma pneumoniae (2), Helicobacter pylori (3)*, *Saccharomyces cerevisiae* (4, 5), and *Drosophila melanogaster* (6). Analysis of protein-protein interaction in human pathogens will ultimately aid in better understanding of their biology and aid therapeutic discovery.

The most comprehensive protein-protein interaction network for the *Mycobacterium tuberculosis* proteome was built using the bacterial two-hybrid (B2H) system (7). The B2H as well as the yeast two-hybrid (Y2H) systems are the most commonly used tools to study protein-protein interactions. They are powerful techniques, but intrinsically carry major limitations. A large caveat is that the screening is far from physiological conditions with a high rate of false positive and negative results (8). To increase the number of genes encoding potentially interactive protein partners, the two-hybrid system was modified to incorporate three different genes, allowing independent

expression and interaction of mycobacterial proteins in *E. coli*. This three-hybrid system was used for the RD1 complex of *M. tuberculosis* (9). However, this method can decipher only tri-protein complexes, establishing that its reliability does not reach global and complex protein-protein interactions and it must be supported by other techniques. There is also a dedicated two-hybrid assay available based on reconstitution of murine dihydrofolate reductase, called mycobacterial protein fragment complementation (M-PFC) assay, which allows to study protein-protein interactions in *M. smegmatis* host. This clearly gives an advantage of studying protein complex formation under physiological conditions and the method was successfully implemented both for soluble as well as for membrane proteins (10, 11). Additionally, computer analysis of the interactome (derived from the STRING 8.0 database) was used to analyze communication between a drug environment and resistome to identify the most plausible paths that triggered the emergence of drug resistance (12).

Here, we propose a single epitope affinity purification (AP) technique combined with LC-MS/MS as a screening method to study protein-protein interactions specifically in *Mycobacterium*. To determine the handiest epitope we designed four constructs containing four different fusion tags to be tested with targeted proteins. For further experiments we chose FLAG, hemagglutinin (HA), protein A (ProtA), and enhanced green fluorescent protein (eGFP) epitopes. We employed a LAP (localization and affinity purification) method coupled with tandem mass spectrometry (LC-MS/MS), an efficient tool to investigate protein-protein interactions in living cells under close to physiological conditions (13). This method typically produces a number of qualitative and descriptive results. Moreover, we provide evidence that chemical cross-linking followed by mass spectrometry is applicable to native mycobacterial complexes to decipher direct contact sites between identified subunits.

The most sensitive and reliable tag for protein-protein interaction and protein complexes analysis in mycobacteria was employed to determine subunits of the comprehensive, stable, and well described in other microorganisms DNA-dependent RNA polymerase. We also used this tag to describe, for the very first time in *Mycobacterium*, dynamic changes in UvrABC protein complex after UV irradiation. We strongly believe that the experimental system along with computational and informatics strategies (reviewed recently by Nesvizhskii (14)), holds capability to aid in

understanding *M. tuberculosis* biology. It will also assist in deciphering cross-talk between pathogen and host, hopefully elucidating weak points of interactions to which drugs may be targeted.

2. Materials and methods

2.1 Vectors and constructs - We designed a suite of vectors with identical backbones based on a pKW08 vector (15). Four different epitopes containing either haemagglutinin (HA), FLAG, protein A (ProtA), or enhanced green fluorescent protein (eGFP) were chosen. The gene encoding the protein of interest was separated from the epitope sequence by the cassette encoding a tobacco etch virus (TEV) protease cleavage site, followed by SH-quant peptide, and a 6-nucleotide spacer (Fig. 1A and 1B). This design allows our cassettes to be used for mass spectrometry-based qualitative analysis and absolute quantification of protein complex components by adding defined amounts of an isotope-labeled heavy version of the SH-quant peptide (AADITSLY[Lys(13C6; 15N2)]; SH-quant*) to the sample (16). The amino acids sequence of each tag was back translated into the DNA sequence using a *Mycobacterium smegmatis* codon usage table, and the nucleotide sequences of the designed tags were submitted for commercial synthesis (GenScript, USA; Integrated DNA Technology, USA). Respective sequences were introduced to the modified pKW08 plasmid to produce vectors suitable for tagging of genes of interest as described in the cloning section.

2.2 Cloning - Our cloning strategy is based on a sequence and ligation independent cloning (SLIC) method (17). Briefly, all inserts intended for cloning were amplified using a pair of 50 nucleotide primers, where the first 30 nucleotides overlap with the vector's compatible ends. The other 20 nucleotides complement the insert. First, the pKW08 vector was linearized with BamHI and HindIII restriction enzymes according to the manufacturer's protocols (DoubleDigest tool, Fermentas ThermoScientific). The primary insert containing the SH-quant and epitope tags was cloned into the vector using the universal forward primer F1 and a tag specific primer. The BamHI and HindIII sites were restored at the 5'end of the tag. Subsequently, the pKW08 vector containing the respective tag was prepared by BamHI/HindIII restriction digestion. Genes encoding bait proteins for protein complex purification were amplified via PCR using the appropriate primer pairs. The 18-nucleotide sequence containing the Shine-Dalgarno box (GGAGGAG) was introduced into the forward primer sequence upstream of the

START codon of the bait protein sequence. Full primer sequences are presented in supplemental Table S1. As part of the 7FP collaborative project *SysteMTb* founded by EC, we can access the Gateway Entry Clone Library (Pathogen Functional Genomics Resource Center, J. Craig Venter Institute, Sponsored by NIAID), which comprises 3295 cloned ORFs from *Mycobacterium tuberculosis* H37Rv supplemented with 430 unique cloned ORFs from *Mycobacterium tuberculosis* CDC1551, for a total of 3725 validated entry clones. All entry clones are flanked with *att* sites, allowing design of universal primers for the entire library, where 30 nucleotides overlaps with the vector's cloning compatible ends and 26 nucleotides can homologously recombine with the vector's *att* sequences. When C-terminal tagging is required, the Shine-Dalgarno box can be constructed via 4 transitions mutations (A→G) in the *attB1* site (gtacAAaAAagttgcccat → gtacGGaGGagttgcccat). N-terminal tagging requires only a STOP codon introduction between the 30 nucleotides and the *attB2* site.

The touchdown PCR (TD-PCR) protocol was used to increase specificity, sensitivity, and yield of PCR products. Phusion High-Fidelity DNA Polymerase (Finnzymes, Thermo Scientific) according to the manufacturer's protocol (200 µM each dNTP, 1x Phusion HF Buffer, 0.5 µM of each primer, 0.02 U/µl Phusion DNA Polymerase, 3-8% DMSO) was used to optimize insert amplification. Annealing temperature started at 60^0C, decreasing 1^0C every cycle, until 50^0C was reached. This 50^0C temperature was constant for the subsequent 25 cycles (98^0C for 10 s, 50^0C for 30 s, 72^0C for 2 min). To ensure complete extension of the PCR products, reactions were incubated for an additional 7 min at 72^0C, and then held at 4^0C.

To clone an amplified insert, 100 ng of linearized vector and 200 ng of PCR product were mixed and treated with 0.5 U of T4 DNA polymerase (BioLabs) in buffer G (Fermentas) at room temperature for 10 min. The reaction was stopped by adding 1/10 volume of 10 mM dATP and incubated on ice for 5 min. The annealing reaction was performed at 37^0C for 30 min, and kept on ice for transformation or stored at -20^0C.

Typically, 150 µl of chemically competent MH1 *E. coli* cells were transformed with the SLIC mixture. Bacterial cells were incubated with the SLIC mixture on ice for 30 min and then subjected to a heat-shock at 42^0C for 90 sec in a water bath, followed by 2 min on ice. Next, cells were incubated at 37^0C in 850 µl of SOB medium for 1 hour, permitting expression of transferred antibiotics resistance. Cells were pelleted, the medium was

reduced to 100 - 200 µl, and the cells were plated on LB plates containing hygromycin B (hygroGold, Invivogen) at a final concentration of 200 µg/ml.

2.3 Bacterial strains and growth conditions - The mycobacterial strains used in this study include *M. smegmatis* mc²155 and *M. bovis* BCG Danish strain 1331 (SSI, Copenhagen, Denmark). Strains were cultured in Middlebrook 7H9 broth supplemented with sodium chloride, albumin, dextrose, and catalase (ADC). For transformation of mycobacterial cells, appropriate parental strains were grown to exponential phase (OD_{600}= 0.6 - 0.8. Cells were then collected by centrifugation (4800 x g, 10 min, 4°C), washed three times with cold 10% glycerol, and transformed via electroporation (25 µF, 1000 Ω, 2500 V). Bacteria were recovered in 5 ml of fresh media for 3 hours at 37°C before plating. Transformants were selected on 7H11 solid media supplemented with ADC and hygromycin (50 µg/ml). To induce recombinant protein production, tetracycline was supplied in the growth media at a final concentration of 50 ng/ml. Cultures were grown in the presence of inducer for 3 and 48 hours for *M. smegmatis* and *M. bovis* BCG, respectively. Growth was monitored by optical density measurements at 600 nm.

2.4 Protein complex purification **-** Mycobacterial cells were collected by centrifugation (15 min at 4800 x g, 4°C) and resuspended in 9 ml of cold sonication buffer containing 50 mM Tris pH 8.0, 100 mM NaCl, 1 mM DTT, 2 mM phenylmethylsulfonyl fluoride (PMSF, Sigma-Aldrich), 25 U/ml Benzonase (Sigma-Aldrich,) and 0.5% Triton X-100 (Sigma-Aldrich). The buffer was supplemented with protease inhibitors (2 µM Pepstatin A, 2 µg/ml Chymostatin, 0.6 µM Leupeptin, 1 mM Benzamidine HCl, and 0.1 M PMSF). Cells were transferred into the conical 50 ml tubes and sonicated in the Diagonade sonication system in a cooled water bath (4°C) at high power (300 W) for 90 cycles with 45s ON and 30s OFF for each cycle. Cell debris was removed by centrifugation (20 min, 4800 x g, 4°C) and cleared whole cell lysates were transferred to new 15 ml conical tubes where 40 µl of the tag-specific resin was added: anti-GFP sepharose (prepared as described below) (18), anti-FLAG agarose (Sigma-Aldrich), anti-HA agarose (Sigma-Aldrich), or IgG Sepharose (GE Healthcare), for respective tagging systems. Samples were incubated for 2 hours in a cold room with slow (6-8 rpm) end-to-end rotation. The beads were recovered on a polypropylene Poly-Prep Chromatography Column (Bio-Rad). Flow through was collected, and for GFP tagged samples, the fluorescence of GFP unbound to the beads was measured as described below. The columns with resin and

captured proteins were washed 2 times with 10 ml of IPP150 buffer containing 10 mM Tris pH 8.0, 150 mM NaCl, and 0,1% Triton X-100 (Sigma-Aldrich), followed by 2 washes with TEV buffer (10 mM Tris pH 8.0, 150 mM NaCl, 0.5 mM EDTA, 1 mM DTT). For tags containing sites recognized by TEV proteases (FLAG, eGFP, and ProtA), Twenty microliters of TEV protease (cloned, expressed, purified, and successfully used in our lab (19)) was added to 430 µl TEV buffer and applied to the column to cleave off the bait protein from the beads, leaving the GFP tag on the column. The TEV cleavage was performed at 4^0C overnight. Purified proteins were collected into 1.5 ml Eppendorf tubes, and columns were washed with TEV buffer into a 900 µl final volume. HA-tagged proteins were eluted from the column by 400 µl 0.2 M glycine-HCl (pH 2.5) into Eppendorf vials containing 50 µl of 1 M Tris buffer (pH 8.0) for neutralization. The final 900 µl volume was adjusted with TEV buffer. Collected samples were mixed vigorously and divided into 2 equal parts. The bait protein with its interacting partners was precipitated by adding pyrogallol red-molybdate, PRM (0.05 mM pyrogallol red, 0.16 mM sodium molybdate, 1 mM sodium oxalate, 50 mM succinic acid, pH 2.5; all from Sigma-Aldrich) reagent in 1/4 of the original volume and vigorously mixed for 30 sec followed by incubation at room temperature for at least 1 hour. Precipitated proteins were spun down (25 min at 21000 x g, RT) and the supernatant removed. One sample set was submitted for LC-MS/MS analysis, and the second was resolved on an SDS-PAGE gel. The overall workflow is presented on Figure 1D.

2.5 UV damage induction - M. bovis BCG strain expressing the Rv1638/eGFP fusion protein was grown exponentially and induced with 50 ng/ml tetracycline as described above. After induction, cells were spun (4800 x g, 10 min. at RT), washed once with freshly prepared M9 minimal media, and spun again. For each condition, cell pellet from 500 ml cultures was suspended in 10 ml of minimal media, transferred to a Petri dish (ø 15 cm), placed on ice, and irradiated with a Philips 15 W TUV lamp emitting UV at 254 nm with a final UV dose of 4.5 mJ/cm^2 (20). After exposure, bacteria were immediately transferred to 37^0C with moderate shaking and snap frozen in liquid nitrogen to halt UV-damage recovery at times of 0, 1, 5, 15, and 30 minutes post exposure. Protein complexes were purified from each sample using the GFP-trap and protocol described above.

2.6 Anti-GFP sepharose beads preparation - The anti-GFP nanobodies coupled sepharose beads were specifically prepared for this work. To obtain antibodies against Green Fluorescent Protein (GFP), the amino acid sequence of Chain C of the GFP minimizer nanobody (NCB Accession Number 3K1K_C[1]) (21) was back-translated to its DNA coding sequence. Codons were optimized to ensure efficient expression in *E. coli*. A PelB leader sequence was introduced in front of the GFP minimizer for export to the bacterial periplasm and ensure proper folding of the nanobody. The resulting DNA coding sequence was subsequently ligated in frame with the pET28PP vector, which allows the addition of a HisTag (6x) at the C-terminus for easier purification. The construct was transformed into *E. coli* BL21-CodonPlus®-RIL and propagated overnight in LB liquid media containing kanamycin (50 µg/ml) and chloramphenicol (37.5 µg/ml) at 37^0C. Bacterial cultures were diluted 1:50 in autoinduction media (Formidium Super Broth Base including Trace elements) used for large-scale protein expression, and incubated at 18°C for 48 hrs with aeration in an orbital agitator (150 rpm). Cells were collected by centrifugation (10 min, 5000 x g, 4^0C) and lysed by sonication (Branson 250, 40%, 15 min) in 20 mM Tris (pH 8.0) based buffer containing 500 mM NaCl, 20 mM imidazole and 10 mM 2-mercaptoethanol. The crude cell lysate was clarified by centrifugation (45 min, 119046 x g, 4^0C) and supernatant was loaded onto a 5ml Ni-NtA cartridge column (Qiagen). Unbound material was washed from columns with 10 column volumes (CV) of lysis buffer followed by 10 CV same buffer with 1 M NaCl. Pure protein was eluted from the affinity column by using 5CV of elution buffer of 500 mM NaCl and 600 mM imidazole. Affinity purification was followed by gel filtration in PBS buffer containing 500 mM NaCl, using a Superdex 75 column (GE Healthcare). Subsequently, purified GFP nanobodies were coupled with cyanogen bromide-activated-Sepharose 4 Fast Flow (Sigma-Aldrich) beads. For coupling, sepharose was washed with cold 1 mM HCl for 30 minutes (200 ml per 1 g beads), followed by distilled water (10 bead volumes), and suspended in coupling buffer (PBS with 500 mM NaCl). Purified nanobodies were added to the solution for overnight coupling and stored in the cold room. Unbound ligand was washed away by several washes with coupling buffer, and unreacted groups on sepharose were blocked overnight incubation at 4°C with 200 mM glycine. The blocking agent was removed and beads were extensively washed with coupling buffer. Finally,

[1] The amino acid sequence of this protein can be accessed through the NCBI Protein Database under NCBI Accession Number (3K1K_C).

beads were washed with 0.1 M NaAc (pH 4.0), followed by 500 mM NaCl and 100 mM Tris (pH 8.0), and stored in buffer containing 20 mM Tris (pH 8.0), 500 mM NaCl, and 0.025% sodium azide as a preservative.

2.7 Gel electrophoresis - Pelleted proteins were resuspended in loading buffer (10 µl of water, 4 µl of the NuPage LDS Sample Buffer (Invitrogen) and 1 µl of 1 M DTT (Sigma-Aldrich)), boiled for 5 min, and resolved on 4-12% gradient NuPage Bis-Tris gel (Invitrogen) using MES Running Buffer (Invitrogen) at 125 V. PageRuler Prestained Protein Ladder (Fermentas) was used as a molecular weight standard. Gels were stained with Coomasie for 2 hours and distained overnight.

2.8 Sample preparation, mass spectrometry, and peptide/protein identification - Protein pellets were dissolved in 50 µl of 100 mM NH_4HCO_3 and subjected to a standard procedure of trypsin digestion: proteins were reduced with 10 mM DTT for 30 min at 56°C, alkylated with 55 mM iodoacetamide in darkness for 45 min at room temperature, and digested overnight with 10 ng/µl trypsin. The resulting peptide mixtures were applied to RP-18 pre-columns of an HPLC system (Waters) using water containing 0.1% trifluoroacetic acid as the mobile phase, and transferred to a nano-HPLC RP-18 column (internal diameter 75 µM, Waters) using an acetonitrile gradient (0 – 35% ACN in 160 min) in the presence of 0.1% trifluoroacetic acid at a flow rate of 250 nl/min. The column outlet was coupled directly to the ion source of an Orbitrap Velos mass spectrometer (Thermo Scientific). A blank run ensured absence of cross-contamination from preceding samples.

The mass spectrometer was operated in a data-dependent mode to automatically switch between Orbitrap-MS and LTQ-MS/MS acquisition. Survey full-scan MS spectra (from m/z 300 to 2000) were acquired in the Orbitrap with a resolution of R = 15,000 at m/z 400 (after accumulation to a target of 1,000,000 charges in the LTQ). The method used allowed sequential isolation of the most intense ions (up to five, depending on signal intensity) for fragmentation on the linear ion trap using collision induced dissociation at a target value of 30,000 charges. Target ions selected for MS/MS were dynamically excluded for 60 sec. Chromatographic peak apex detection triggered data dependent scans (expected peak width: 5 s, minimal signal threshold: 10,000 counts) with phase method activated and triggering window set to 30%. General mass spectrometry

conditions were as follows: electrospray voltage, 1.8 kV, no sheath, and auxiliary gas flow. Ion selection threshold was 10,000 counts for MS/MS, and an activation Q-value of 0.22 and activation time of 30 ms were also applied.

The raw files were processed, including peaklist generation, using the MaxQuant (v1.3.0.5) computational proteomics platform and default parameters were used. The fragmentation spectra were searched using Andromeda search engine integrated into MaxQuant platform, against an *M. smegmatis* mc^2 155 protein database available at the CMR website (http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi, 6878 entries, v15.1, Oct 15, 2004) or against *M. bovis* BCG database (www.patricbrc.org, NC_008769, 3952 entries). The databases were in-house modified to contain randomized sequences of all entries to control for false-positive identifications during analysis using the Andromeda search engine. The error ranges for the first and main searches were 20 ppm and 6 ppm, respectively, with 2 missed cleavages. Carabamidomethylation of cysteines was set as a fixed modification, and oxidation and protein N-terminal acetylation were chosen as variable modifications for database searching. Minimal peptide length was set at 7 aa. Both peptide and protein identifications were filtered at a 1% false discovery rate and were thus not dependent on the peptide score. Enzyme specificity was set to trypsin, allowing cleavage of N-terminal proline. A "common contaminants" database (incorporated in MaxQuant software) containing commonly occurring contaminations (keratins, trypsin etc.) was employed during MS runs.

Bioinformatics analysis was performed using the Perseus tool (v1.3.0.4, Cox J., Max Planck, 2012). Contaminants and random protein identification were excluded. Proteins identified by less than two peptides were excluded from results, except SH, the quantification peptide. Peptides and proteins identification details including scores are provided in supplemental Tables S6 and S7.

2.9 Protein cross-linking, mass spectrometric analysis, and cross-links validation - For protein complexes cross-linking, we chose the DNA-directed RNA polymerase, where the alpha subunit (rpoA, MSMEG_1524) was fused with the C-terminal GFP tag. The purification procedure was as described above, with the TEV cleavage buffer changed to a 10 mM HEPES (pH 8.0) based buffer. Purified protein complexes eluted form the column were subjected to cross-linking. We used bis(sulfosuccinimidyl) suberate (BS3) as a cross-linker (Thermo Scientific) with an 8-carbon spacer arm (11.4 Å) according to

the manufacturer's protocol. Heavy (d4)- and light (d0)- versions of BS3 reagent were dissolved in DMSO and mixed at a 1:1 ratio immediately before use. The d0/d4 mixture was used to induce stable and selective chemical cross-links between lysine (K) residues available on surfaces of purified proteins to fix potential interactions between protein partners. 50 mM of BS3 (d0/d4) mixture was added at a final concentration of 2 mM to purified proteins and incubated for 15 min at 4°C. The reaction was stopped by adding 10 µL of 3 M Tris solution (pH 8.0). Samples were precipitated with PRM as already described. Subsequently, proteins were digested overnight with 10 ng/ml trypsin (Promega) in 100 mM ammonium bicarbonate buffer at 37°C. Peptides were reduced in 10 mM dithiotreitol (DTT) for 30 min at RT and alkylated in 55 mM iodoacetamide for 20 min at room temperature. Finally, trifluoroacetic acid was added at a final concentration of 0.1%.

To determine protein compositions of cross-linked samples, we used MaxQuant software (as described above). To search for cross-linked peptides, we used pLink (pFind Studio) (22). The following parameters were set: precursor mass tolerance 50 ppm, fragment mass tolerance 20 ppm, cross-linker light [d0]-BS3 and heavy [d4]-BS3 (cross-linking sites K and protein N terminus, xlink mass-shift 138.0680796, monolink mass-shift 156.0786442), isotope shift 4.0247 Da, fixed modification C 57.02146, and enzyme trypsin.

We used .mgf files (Mascot Generic Files generated from .raw files by Mascot Distiller) and a protein database containing proteins found in preceding MaxQuant search. All looplinks and monolikns were excluded from our obtained results. Only inter- or intra-molecular cross-links were taken for further analysis. Molecular graphics and analyses were performed with the UCSF Chimera package. Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS 9P41GM103311).

2.10 Microscope evaluation – The eGFP tag allows subcellular visualization of proteins of interest. *M. smegmatis* mc^2155 expressing the RpoA-eGFP fusion protein was used as a model for testing GFP localization. The strain was grown in culturing media described above to an OD_{600} of 0.6 - 0.8. Protein production was induced by adding 50 ng/ml tetracycline for 3 hours. Cells were collected by centrifugation (15 min at 4800 x g, 4^0C) and washed with PBS. As a counterstain, nuclei were stained with DAPI at 0.5 µg/ml for

10 minutes at RT. Cells were washed again to remove excess dye. Slides were closed with fluorescent Mounting Medium (Dako). A IX81 fluorescence microscope (Olympus) fitted with a PLANAPO 100x/1.35 oil immersion objective and appropriate filter sets (Semrock) was used for bright-field and fluorescence microscopy, and images were acquired using an Orca R^2 camera (Hamamatsu) and the Excellence software package. Images were processed using ImageJ 1.46r and Adobe Photoshop CS4 software.

2.11 Fluorescence intensity measurement **-** In order to estimate the approximate efficiency of binding of eGFP-tagged bait proteins to the anti-GFP beads, fluorescence intensity was exploited. Briefly, cell lysates derived from recombinant *M. smegmatis* strains expressing eGFP fusion proteins were made, pre-cleared by centrifugation, and flow-through after binding to the column was diluted 1:1 in IPP150 buffer and transferred onto 96 well black solid plate (Nunc, Thermo Scientific). Cell lysate from the *M. smegmatis* mc^2 155 parent was used as a background control. Lysates were prepared from approximately the same number of cells as measured by cell pellet weight. Fluorescence counts were measured using Beckman Coulter DTX 800/880 Multimode Detector and Multimode Detection Software. Excitation at 485 nm and emission at 535 nm was used with a data integration time of 1 second. The relative binding efficiency was calculated by dividing the fluorescence intensity of flow-through by the intensity of the lysate before binding to the column, multiplied by 100%.

3. Results

Affinity tags serve as selective and efficient tools for protein purification, and are applicable for purification of native protein complexes. Out of many available tags, we examined four that we chose based on a predicted usefulness for high throughput purification and analysis of the protein complexes in mycobacteria. To simplify the method and ensure co-purification of the most possible interacting partners including weak and transient interactions, we performed single step purification. We chose FLAG tag (23), HA tag, and the protein A IgG-binding domain (24, 25), which all are popular tags. Our fourth tag was based on eGFP protein and designed with the GFP-binding beads for the purpose of these experiments. The tags we investigated are known to interact with appropriate/respective affinity resins coupled to specific antibodies. We restricted our study to affinity tags that could be eluted under relatively mild conditions,

ensuring that we pull down protein complexes to analyze intracellular interactions. The FLAG, eGFP and ProtA tags contain the TEV protease cleavage site (Fig. 1), making protease cleavage a favorable method for elution of protein complexes. Because the HA tag was not provided with protease cleavage site, we had an opportunity to test an alternative elution method. We used the most effective, mildly denaturing elution buffer 0.2 M glycine (pH 2.5). Applied low pH disrupts most antibody-antigen interactions and this elution method was particularly effective.

3.1 Comparison of tags for protein complexes purification **-** To compare efficiency and specificity of protein complex purification using chosen tags, we selected 8 proteins from *Mycobacterium smegmatis*. Those proteins are implicated in different metabolic pathways (purine/pyrimidine metabolism, glycolysis/gluconeogenesis, pentose phosphate pathway) and fulfill varying cellular functions (e.g. RNA synthesis, glycolysis/gluconeogenesis, and recombination). All selected genes encoding selected proteins are summarized in Table 1. Each gene was expressed in *M. smegmatis* in fusion with all four tags, totaling 32 combinations.

Each of the tagged genes was expressed and resulting protein complexes were purified on a specific resin followed by LC-MS/MS and computational analysis. Just prior to precipitation with PRM reagent, samples were divided. Half of each sample was loaded on a Tris-glycine SDS-PAGE gel (Figure 2) and the other half submitted for LC-MS/MS analysis. All identifications, including prey calculated intensities of prey are presented in supplemental Table 2.

LC-MS/MS experiments include high levels of contaminations created mainly by non-specific interactions of proteins with the resin used for affinity purification. For complex peptide mixtures in cell lysates, co-elution may complicate biological evaluation of results. Between most common contaminating proteins we found chaperons, heat-shock proteins, ribosomal proteins, and other proteins nonspecifically bound to the purified protein complexes and to the resin. To find a tag most useful in mycobacterial pull-downs and applicable for high throughput experiments, we attempted to achieve balance between low number of total identified proteins (low background) without losing the real binding partners. Figure 3A shows the total number of identified proteins specific for the bait proteins and compares the number of obtained identifications with different tags purified on tag-specific resins. The lowest number of binding candidates

was observed with FLAG and eGFP tags. Using HA and ProtA tags we obtained much longer lists of detected proteins (supplemental Table S1). Thus, based on the total number identified proteins, FLAG and eGFP tags seem to be serviceable with the lowest resin specific background. We also discerned the number of exclusive proteins identified by MaxQuant software. The 'exclusive proteins' term represents all proteins specific for both the tag/resin and the bait, indicating a combination of real interactors and a protein background specific for the particular tag/resin. We observed that the number of purified proteins depends not only on the tag, but also on the bait. For example, MSMEG0358, the beta subunit of the ribonucleotide-diphosphate reductase, was purified with the highest number of both total and exclusive identifications independent of tag used. Of note, all tags selected for this work were cloned into and expressed from the same vector using identical induction conditions of the tetR08 promoter.

3.2 Background evaluation **-** In AP-MS studies, determining noise, false positives, and false negatives is necessary to distinguish true interactions from contaminants. Sequential purification steps (e.g. Tandem Affinity Purification) may decrease these unwanted results, but at a risk of losing both weak and transient interactions. We have analyzed nonspecifically binding proteins, commonly associated with all tested baits. We established averaged intensities values characteristic for each nonspecifically binding protein for each of 4 tags. The highest number of background proteins identified in all 8 proteins was observed with HA-tag experiments (76 proteins) and the lowest with the eGFP-tag (25 proteins). Nonspecific binders for FLAG and ProtA experiments were 33 and 32, respectively (Fig. 4, supplemental Table S3). Interestingly, we found only 8 proteins identified in all 32 analyzed samples. These include 4 ribosomal proteins, 2 chaperons, a reductase, and a transcription termination factor (supplemental Table S3). Additionally, samples were examined by SDS-PAGE gel electrophoresis visualized by Coomassie staining (Figure 2). The ProtA and HA tagged samples were enriched compared to the FLAG tagged samples, correlated to the total number of identified proteins presented in Fig. 3. The protein enrichment in eGFP-tagged experiments is higher than in FLAG, but considering the amount of background resin-bound proteins (Fig. 3B), an eGFP tag and respective resin provides relatively low background and high specificity with high protein enrichment. We thus conclude that the eGFP tag combines the desirable features mentioned, offers easy ways to monitor binding efficiency by measuring GFP fluorescence (Table 2), can be used directly for localization experiments

(Fig. 6). For further experiments, we chose the eGFP protein tag for practical application in mycobacterial proteomic experiments.

3.3 Protein complexes identified by AP-MS approaches - All proteins used as baits for affinity purification experiments in *M. smegmatis* purified on specific affinity resins and were identified as dominant proteins in respective samples (Supplemental table S2). After removing the common contaminants as well as bead specific contaminants it was possible to see complex formation for most of them. Some of them like the RNA polymerase complex were predictable, others were completely novel, and they will need further studies to understand the biological meaning of formation of such complexes in the mycobacterial cell. For instance the MSMEG_0358, annotated as beta subunit of ribonucleoside-diphosphate reductase co-purified with significant amounts of MSMEG_1960 and MSMEG_1961, both conserved hypotheticals, and MSMEG_1476, signal peptide peptidase. These proteins were found exclusively in every purification of MSMEG_0358, regardless of tagging system. There was also substantial increase in the amount of MSMEG_6284, cyclopropane-fatty-acyl-phospholipid synthase in those samples. Another bait MSMEG_1666, predicted to be RNA polymerase sigma factor SigJ, specifically co-purified with MSMEG_4121, gntR transcriptional regulator. Finally, MSMEG_3086-eGFP, predicted as triosephosphate isomerase (Tpi) when used as a bait, co-purified with MSEMG_3085 phosphoglycerate kinase (Pgk) from the same operon, but not vice versa. Pgk on the other hand was found to form complex with MSMEG_4248, 1-acylglycerol-3-phosphate O-acyltransferase and MSMEG_2340, a hypothetical protein with limited similarity to isopentyl pyrophosphate isomerase.

3.~~3~~4 Purification of DNA-directed RNA polymerase protein complex **-** Bacterial RNA polymerase is a well characterized enzyme comprised of five core subunits (α, α, β, β', ω), which bind accessory proteins like sigma factors to form functional holoenzyme (26, 27). The structure of the *E. coli* core enzyme is available and importantly shares sequence similarity with mycobacterial homologues (α - 54,9%, β - 56,8%, β' - 55,0%, ω - 30,1%, *E. coli* to *M. smegmatis*). Since the structure and composition of the RNA polymerase complex is known, it is often used as a model for purifying protein complexes and thus RNA polymerase alpha subunit (RpoA; α), was chosen as a target in

our study. It allows to evaluate purification and accuracy of detection of the RNA polymerase subunits.

The four core subunits of RNA polymerase co-purified with RpoA fused to all four tested tags. Tagged RpoA with RpoB and C subunits were detected with high signal intensity (Fig. 5). The lowest intensity of subunits was found with FLAG tag. RpoZ, the smallest subunit of the holoenzyme, was detected with lowest intensity, but its sequence coverage was the same for all tags. Data presented in Figure 5 includes hits remaining after filtering out the first 40 proteins with highest intensity. Excluded proteins were classified as contaminants (supplemental Table S3). This method placed all known holoenzyme components in the top 10 of the hit list. The proteins with the highest abundance and with best enrichment *vs.* background located in the top right-hand corner of each scatter. The RpoA abundance was in agreement with the abundance of SH quant peptide which is helpful with determining the number of molecules of RpoA in each sample.

Due to a fast growth rate, well established methods for genetic manipulation, and biosafety level 1 requirements, *M. smegmatis* is one of the best organisms to study cellular mechanisms of its pathogenic cousin *M. tuberculosis*. However, *M. smegmatis* is a nonpathogenic mycobacteria (except for extremely immunodeficient individuals) with a genome approximately 1.7 times larger than *M. tuberculosi*s. In moving to a model closer to *M. tuberculosis*, we tested our eGFP tag procedure in vaccine *Mycobacterium bovis* BCG Danish strain, with the similarity to *M. tuberculosis* with approximately 99.9%, at the genetic level. Both *M. bovis* BCG and *M. tuberculosis* are member strains of the *M. tuberculosis* complex (28). Thus, successful application of our method to BCG may suggest that the same approach can be used in virulent *M. tuberculosis* with only little modification. Additionally, with high genetic similarity between the two, we expect little or no difference between native complexes, justifying *M. bovis* as a more optimal model organism to study protein-protein interactions for *M. tuberculosis*.

To test our method in *M. tuberculosis* closer cousin, the coding sequence of *M. tuberculosis* RpoA (Rv3457c) was cloned into a pKW08-eGFP vector and transformed into *M. bovis* BCG by electroporation. The RNA polymerase protein complex was obtained by the same exact method as used for *M. smegmatis*. This time we used both C-terminal tagging and N-terminal tagging to be able to answer if that is going to influence

the purification outcome. Additionally, we prepared a strain expressing eGPF-tagged SigB (Rv2710, identical with BCG_2723) - one of the less abundant subunits found in the complex to see if that protein is going to be capable of pulling down the entire complex as well. Results presented in Table 3 show that all core subunits were nicely purified and detected by MS with high sequence coverage and intensity values. In all cases, we detected all the RNA polymerase core subunits and two sigma factors, SigA (also referred as MysA, RpoD) and SigB, in contrast to four sigma factors detected in *M. smegmatis*: SigA, sigma-70, sigma-F, and SigB (Figure 5). When SigB was used as a bait it was purified as dominating protein in the sample, but again it pooled down all the core RNA polymerase components and did not significantly affected the ratio between the other subunits.

Since the eGFP tag can be applied to determine the subcellular localization of targeted proteins, we determined the RpoA-eGFP fusion localization within mycobacterial cells. RNA polymerase is known to exhibit affinity toward DNA, so it was not surprising to find that RpoA-eGFP co-localized with the mycobacterial chromosome (Figure 6). Fusion protein localization might also suggest its functionality and that it may be involved with a holoenzyme. Moreover, induction of eGFP fusion protein production can be discerned under a microscope, an added useful feature.

3.54 Analysis of the RNA polymerase subunit interaction using chemical cross-linking - Affinity purification is a front lining method for analyzing protein-protein interactions and topology of complexes by chemical cross-linking. Cross-linking converts non-covalent interactions between proteins surfaces into artificial covalent bonds. Cross-linking along with MS analysis can support modeling and aid in solving structures of complexes. Since the protein complex purification method based on the eGFP protein fusion/resin has proven efficient with relatively low background, we decided to test its application for cross-linking experiments. Because a 3D structure of the core RNA polymerase enzyme is available for *E. coli* (NCB Accession Number 3LU0[2]), and most of the mycobacterial components of this enzyme share high amino acid sequence similarity, we used protein cross-linking for ascertaining interactions between the homologous mycobacterial enzyme subunits. We employed the BS3 cross-linker, which

[2] The atomic coordinates for the crystal structure of this protein complex are available in the Molecular Modeling Database http://www.ncbi.nlm.nih.gov/Structure/mmdb/mmdb.dtl under MMDB Number 3LU0

is reactive towards amine groups and is designed with an 11.4 Å spacer arm, which allows chemical cross-linking of two neighboring lysine (K) residues and/or the N-terminal amino acid within reach of the spacer arm. Several cross-links were identified with pLink software (pFind Studio) (supplemental Table S4), and sample cross-links were then projected into the 3D structure model mentioned previously to assess proximity by measuring the distances between the *E. coli* amino acids in Chimera software. Products from cross-linking proved that mycobacterial core enzymes share high sequence and structure homology to their homologues in *E. coli* since many of the cross-links were separated by less than 20 Å and were positioned on the contact surface between two different protein subunits (e.g. cross-link *M. smegmatis* RpoC K827 - RpoB K184, in *E. coli* repectively RpoC D751 - RpoB K164 distance = 19.527 Å; *M. smegmatis* RpoA K153 - RpoB K837, in *E. coli* E162 - T927 distance = 16.867 Å and data not shown). Our experiments also confirmed the usefulness of cross-linking in assigning real interacting partners identified initially by AP-MS. Information obtained from cross-linking may indicate the structures of multiprotein complexes (29) and help to identify the contact surface between two proteins.

3.6~~5~~ Analysis of protein-protein complexes under changing growth conditions **-** Ideally, the method used in our study should translate to *M. tuberculosis* to investigate cellular processes. It is well established that the composition of various protein complexes may differ during various growth conditions or under stresses. In this study, we decided to use UvrA, a protein involved in a process of DNA damage repair system (NER), and also well conserved between the bacterial species. UvrA is known to be in complex with UvrB, where a UvrA-UvrA dimer binds UvrB to form a DNA integrity scanning complex, $UvrA_2B$ or $UvrA_2B_2$ (30). The complex undergoes structural rearrangement and dissociates whenever it identifies helical distortions induced by a mismatched DNA sequence (31). This enables recruitment of other proteins needed to complete repair. We expressed *M. tuberculosis* UvrA (*Rv1638*), which is identical to BCG_1676 from *M. bovis* BCG, in BCG to determine complex formation after DNA damage induction with UV light. We highlight that UvrA is in complex with UvrB in cells as expected. We were also able to monitor complex dissociation during the DNA damage repair and re-association after the repair process is completed (see Table 4). Polymerization of newly synthesized DNA fragments is performed by DNA polymerase I, and we observed an enrichment of DNA polymerase I after 5 minutes post induction of UV damage. Five minutes after UV

irradiation, we observed the dissociation of the UvrAB complex, and 25 minutes later, the UvrAB complex was again detectable. This suggests that the kinetics of UvrA-B dissociation is similar to kinetics observed in *E. coli* (30), despite *E. coli*'s 20 minutes (average) doubling time in comparison to 16-20 hours for *M. bovis* BCG. This result adds a dynamic capability to our method.

4. Discussion

Affinity purification coupled with mass spectrometry is used to identify proteins and their interacting partners. The first step is efficient purification of protein complexes with, ideally, no or little background. Optimizing this method to improve efficiency and breadth of interactions discovered would help gain understanding of pathogen biology. Several affinity tags are now used to facilitate isolation of proteins with their partners. Based on the nature of the affinity tag and its target, we can distinguish several systems: protein – immobilized molecular ligand (hexahistidine – metal) (32), protein-protein (calmodulin binding peptide – calmodulin) (33), and subsystem protein – antibody (FLAG – anti-FLAG) (23). A large number of affinity tags and specific binding resins are commercially available. Selecting the best for both protein bait and organism is indeed a key step for a successful experiment. Importantly, a properly chosen affinity tag allows proteins to be purified using generalized protocols (34), which is an important parameter in large scale and high-throughput experiments. As Jordan Lichty and colleagues summarized, the ideal affinity tag should be characterized by efficient, high yield protein purification, used with any protein without losing function, placed at any position (C- or N-terminus), used in any host or expression system, can be easily used to detect the recombinant protein, and should bind and be eluted from an inexpensive resin (34). Using affinity tags fused to a protein of interest allows production, isolation, and accurate identification of interacting partners in the native system. Protein insolubility, conformation, stability, structural flexibility, and purification yield and recovery are challenges that must be resolved in these experiments. Carefully chosen affinity tags and the relevant purification protocol, specific resin, and elution method mitigate aforementioned problems. In high-throughput experiments, affinity tag and purification method should be versatile, serviceable, and inexpensive. The most popular affinity tags and proteases used for tag removal have been detailed elsewhere (35). We decided to test four different tags with 8 mycobacterial proteins expressed in a

commonly used, non-pathogenic, laboratory strain *M. smegmatis* mc^2 155. All tags are a method via which to bind the protein to a resin with immobilized antibodies that recognize a specific epitope. We have shown that affinity tags can be used for protein purification from mycobacterial species, and interacting protein partners are detectable. The purity and background signal do vary. As described before (34) and from our data, highest purity with lowest quantities were obtained by using a FLAG tag. ProtA and HA tags yielded a large amount of interacting material, but with high resin-specific background. We focused on the eGFP tag, which merges the high protein enrichment of ProtA and HA protocols with relatively low background as seen with the FLAG tag.

In our hands, background detected in experiments exploiting the GFP tag was the lowest. Additionally, one of the important features of this experimental setup is easiness of detection of tagged recombinant proteins. With the exception of GFP, all examined tags need special attention to visualize recombinant proteins within cells and to determine their sub-cellular localization and expression levels. Only cells expressing protein fused with eGFP can be directly used for microscopy. Our eGFP tag allows binding efficiency measuring of tagged protein to the respective affinity resin (Table 2).

The eGFP tag is full length enhanced green fluorescent protein, which may impact structure or solubility of the tagged protein within cells. It is often detectable during overexpression of recombinant proteins when missfolded proteins aggregate and form inclusion bodies (36). For this reason, expression of recombinant proteins in our system was relatively low and protein fusions were not visible as thick bands after gel electrophoresis of the cellular lysates, as it commonly is for overexpressed proteins (data not shown). When the eight different proteins we fused to eGFP were purified, aggregation was neither detected during purification nor in inclusion bodies by microscopy. Moreover, the GFP tag can dually be used for protein localization studies. This feature can control for protein aggregation and for localization screening of purified proteins.

To the list of GFP tag advantages, we can add the ability to quickly and accurately control expression of protein fused to eGFP, the high recovery ratio from the anti-GFP resin, and very low cost of purification. Cost is a critical parameter when an affinity tag and appropriate resin is selected, particularly for high-throughput experiments. We compared the price of purification for different affinity resins. In-house preparation of

the anti-GFP affinity resin, as it was done for this work, dramatically decreases expenses (supplemental Table S5).

Deciphering of the protein-protein interactions might be very helpful to improve our understanding of biology and pathogenesis of *Mycobacterium*. To aid this quest, we compare four different affinity tags commonly used for affinity purification and evaluate their potential use in high throughput experiments in mycobacterial model. Although, the two-hybrid and three-hybrid systems have been used successfully (7, 9, 37, 38) similarly efficient assays need to be developed for use in the relevant native organism. Based on our data, we strongly advocate the use of the eGFP-based affinity tags for protein purification, identification of protein-protein interactions in both small- and large-scale experiments in mycobacterial cells. Potential targets from the list of preys co-purified in the AP-MS experiments can be then confirmed by alternative techniques. Moreover, chemical cross-linking can be helpful to increase the confidence of the interaction and assigning the contact surfaces between the proteins forming a complex. This is especially valuable for structural studies on complexes with known homology to already characterized complexes isolated from other organisms. This was the case for RNA polymerase in our studies. Numerous cross-links between *M. smegmatis* subunits not only mapped to the *E. coli* model (39), but there was a number of crosslinks between the core RNA polymerase subunits and the sigma factors, which provides additional information that can be useful for modeling of this essential large protein complex.

We also tried to address the issue of tagging at which terminus C- or N- of RpoA will allow for more efficient purification of RNA polymerase complex from *M. bovis* BCG (Table 3). We found almost exactly the same purification efficiency regardless of placement of the eGFP tag. However this certainly cannot be treated as absolute true for every protein and some proteins will require tag to be placed on specific terminus to allow complex formation in living cells. That may be the case why we were able to observe complex formation between the two glycolitic enzymes TpiA and Pgk only in one combination, not the other. This two enzymes were found to be closely linked in many other organisms. In *Thermatoge maritima* they were found as covalently linked

fusion proteins able to form multimeric bifunctional complex (40). It is highly possible that the Pgk needs it's C-terminus to be tag free to be able to interact with and pull down TpiA.

Using the nucleotide excision repair protein UvrA in *M. bovis* BCG model, we proved that AP-MS based approaches are capable of detecting dynamic changes in the protein complex formation under changing circumstances. Without DNA damaging stimuli, UvrA was found to co-purify with substantial amount of its partner UvrB and induction of DNA damage caused specific reaction of the cell, trying to repair the damage made by UV irradiation, causing temporary dissociation of the UvrAB complex. Our data also indicates presence of possible additional players in the damage scanning mechanism - topoisomerase I (topo I) and a DNA helicase II annotated as UvrD2, however additional studies will have to be conducted to understand the underlying mechanisms of such interactions. One possibility would be that the DNA integrity scanning complex requires topo I and DNA helicase II to respectively relax and unwind the DNA during scanning. In eukaryotes it was shown that down modulation of topoisomerase I using antisense RNA inhibits repair of UV-induced lesions. The experiments show that topo I is actively recruiting onto genomic DNA following DNA damage by UV light, possibly acting during pre- or post-DNA damage processing (41). Similar function of topoisomerase I may be required for effective NER repair in *Mycobacteria* and possibly other prokaryotes. Interaction between UvrA, UvrB and UvrD was shown by immunoprecipitation in *E. coli* (42) and here we have just another proof that these proteins form a complex in prokaryotic cells.

Acknowledgements

The authors thank Aleksander Chlebowski for help with microscope imaging, Agata Malinowska, Jacek Olędzki and Agnieszka Fabijańska for LC-MS/MS technical help and discussion. D.L. designed construct and experiments, P.P. designed and cloned anti-GFP nanobodies, D.L. and P.P. designed and performed all experiments, collected and analyzed data, prepared tables and figures and wrote the manuscript, D.C. performed cross-link experiments, calculated and analyzed all MS data, K.K. prepared expression constructs, K.S. expressed and purified anti-GFP nanobodies and prepared anti-GFP resin, A.D. conceived and directed the studies and corrected the manuscript.

Competing interests

None declared

References

1. Kelkar DS, *et al.* (2011) Proteogenomic analysis of Mycobacterium tuberculosis by high resolution mass spectrometry. *Molecular & cellular proteomics : MCP* 10(12):M111 011627.
2. Kuhner S, *et al.* (2009) Proteome organization in a genome-reduced bacterium. *Science* 326(5957):1235-1240.
3. Rain JC, *et al.* (2001) The protein-protein interaction map of Helicobacter pylori. *Nature* 409(6817):211-215.
4. Ito T, *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America* 98(8):4569-4574.
5. Yu H, *et al.* (2008) High-quality binary protein interaction map of the yeast interactome network. *Science* 322(5898):104-110.
6. Giot L, *et al.* (2003) A protein interaction map of Drosophila melanogaster. *Science* 302(5651):1727-1736.
7. Wang Y, *et al.* (2010) Global protein-protein interaction network in the human pathogen Mycobacterium tuberculosis H37Rv. *Journal of proteome research* 9(12):6665-6677.
8. Zhou H & Wong L (2011) Comparative analysis and assessment of M. tuberculosis H37Rv protein-protein interaction datasets. *BMC genomics* 12 Suppl 3:S20.
9. Tharad M, *et al.* (2011) A three-hybrid system to probe in vivo protein-protein interactions: application to the essential proteins of the RD1 complex of M. tuberculosis. *PloS one* 6(11):e27503.
10. Singh A, Mai D, Kumar A, & Steyn AJ (2006) Dissecting virulence pathways of Mycobacterium tuberculosis through protein-protein association. *Proceedings of the National Academy of Sciences of the United States of America* 103(30):11346-11351.
11. Dziedzic R, *et al.* (2010) Mycobacterium tuberculosis ClpX interacts with FtsZ and interferes with FtsZ assembly. *PloS one* 5(7):e11058.
12. Padiadpu J, Vashisht R, & Chandra N (2010) Protein-protein interaction networks suggest different targets have different propensities for triggering drug resistance. *Systems and synthetic biology* 4(4):311-322.
13. Gingras AC, Gstaiger M, Raught B, & Aebersold R (2007) Analysis of protein complexes using mass spectrometry. *Nature reviews. Molecular cell biology* 8(8):645-654.
14. Nesvizhskii AI (2012) Computational and informatics strategies for identification of specific protein interaction partners in affinity purification mass spectrometry experiments. *Proteomics* 12(10):1639-1655.
15. Williams KJ, Joyce G, & Robertson BD (2010) Improved mycobacterial tetracycline inducible vectors. *Plasmid* 64(2):69-73.

Field Code Changed

Formatted: Font: +Headings (Cambria), 12 pt

16. Wepf A, Glatter T, Schmidt A, Aebersold R, & Gstaiger M (2009) Quantitative interaction proteomics using mass spectrometry. *Nature methods* 6(3):203-205.
17. Li MZ & Elledge SJ (2007) Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. *Nature methods* 4(3):251-256.
18. Rothbauer U*, et al.* (2008) A versatile nanotrap for biochemical and functional studies with fluorescent fusion proteins. *Molecular & cellular proteomics : MCP* 7(2):282-289.
19. Tomecki R*, et al.* (2010) The human core exosome interacts with differentially localized processive RNases: hDIS3 and hDIS3L. *The EMBO journal* 29(14):2342-2357.
20. Fabisiewicz A & Janion C (1998) DNA mutagenesis and repair in UV-irradiated E. coli K-12 under condition of mutation frequency decline. *Mutation research* 402(1-2):59-66.
21. Kirchhofer A*, et al.* (2010) Modulation of protein properties in living cells using nanobodies. *Nature structural & molecular biology* 17(1):133-138.
22. Yang B*, et al.* (2012) Identification of cross-linked peptides from complex samples. *Nature methods* 9(9):904-906.
23. Brizzard BL, Chubet RG, & Vizard DL (1994) Immunoaffinity purification of FLAG epitope-tagged bacterial alkaline phosphatase using a novel monoclonal antibody and peptide elution. *BioTechniques* 16(4):730-735.
24. Rigaut G*, et al.* (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nature biotechnology* 17(10):1030-1032.
25. Puig O*, et al.* (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* 24(3):218-229.
26. Murakami KS & Darst SA (2003) Bacterial RNA polymerases: the wholo story. *Current opinion in structural biology* 13(1):31-39.
27. Tagami S*, et al.* (2010) Crystal structure of bacterial RNA polymerase bound with a transcription inhibitor protein. *Nature* 468(7326):978-982.
28. Mostowy S, Cousins D, Brinkman J, Aranaz A, & Behr MA (2002) Genomic deletions suggest a phylogeny for the Mycobacterium tuberculosis complex. *The Journal of infectious diseases* 186(1):74-80.
29. Luo J, Fishburn J, Hahn S, & Ranish J (2012) An integrated chemical cross-linking and mass spectrometry approach to study protein complex architecture and function. *Molecular & cellular proteomics : MCP* 11(2):M111 008318.
30. Orren DK & Sancar A (1990) Formation and enzymatic properties of the UvrB.DNA complex. *The Journal of biological chemistry* 265(26):15796-15803.
31. Van Houten B, Croteau DL, DellaVecchia MJ, Wang H, & Kisker C (2005) 'Close-fitting sleeves': DNA damage recognition by the UvrABC nuclease system. *Mutation research* 577(1-2):92-117.
32. Hochuli E, Dobeli H, & Schacher A (1987) New metal chelate adsorbent selective for proteins and peptides containing neighbouring histidine residues. *Journal of chromatography* 411:177-184.
33. Stofko-Hahn RE, Carr DW, & Scott JD (1992) A single step purification for recombinant proteins. Characterization of a microtubule associated protein (MAP 2) fragment which associates with the type II cAMP-dependent protein kinase. *FEBS letters* 302(3):274-278.

34. Lichty JJ, Malecki JL, Agnew HD, Michelson-Horowitz DJ, & Tan S (2005) Comparison of affinity tags for protein purification. *Protein expression and purification* 41(1):98-105.
35. Young CL, Britton ZT, & Robinson AS (2012) Recombinant protein expression and purification: a comprehensive review of affinity tags and microbial applications. *Biotechnology journal* 7(5):620-634.
36. Wang H & Chong S (2003) Visualization of coupled protein folding and binding in bacteria and purification of the heterodimeric complex. *Proceedings of the National Academy of Sciences of the United States of America* 100(2):478-483.
37. Li Y, Franklin S, Zhang MJ, & Vondriska TM (2011) Highly efficient purification of protein complexes from mammalian cells using a novel streptavidin-binding peptide and hexahistidine tandem tag system: application to Bruton's tyrosine kinase. *Protein science : a publication of the Protein Society* 20(1):140-149.
38. Huang F & He ZG (2012) Characterization of a conserved interaction between DNA glycosylase and ParA in Mycobacterium smegmatis and M. tuberculosis. *PloS one* 7(6):e38276.
39. Opalka N, *et al.* (2010) Complete structural model of Escherichia coli RNA polymerase from a hybrid approach. *PLoS biology* 8(9).
40. Schurig H, *et al.* (1995) Phosphoglycerate kinase and triosephosphate isomerase from the hyperthermophilic bacterium Thermotoga maritima form a covalent bifunctional enzyme complex. *The EMBO journal* 14(3):442-451.
41. Mao Y & Muller MT (2003) Down modulation of topoisomerase I affects DNA repair efficiency. *DNA repair* 2(10):1115-1126.
42. Ahn B (2000) A physical interaction of UvrD with nucleotide excision repair protein UvrB. *Molecules and cells* 10(5):592-597.

Tables

Table 1. List of selected targets.

Sample no	Locus name	Gene definition	Gene lenght (nt)	Protein lenght (aa)
1	MSMEG0358	ribonucleoside-diphosphate reductase, beta subunit [1.17.4.1]	963	320
2	MSMEG0752	fructose-bisphosphate aldolase, class II (fbaA) [4.1.2.13]	1038	345
3	MSMEG1524	DNA-directed RNA polymerase, alpha subunit (rpoA) [2.7.7.6]	1053	350
4	MSMEG1666	RNA polymerase sigma-70 factor	915	304
5	MSMEG2136	phosphoglucomutase, alpha-D-glucose phosphate-specific (pgm) [5.4.2.2]	1635	544
6	MSMEG3021	AAA ATPase, central region	1344	447
7	MSMEG3085	phosphoglycerate kinase (pgk) [2.7.2.3]	1227	408
8	MSMEG3086	triosephosphate isomerase (tpiA) [5.3.1.1]	786	261

Table 2. Calculated binding efficiency based on eGFP fluorescence intensity detected in cell lysates vs. flow-through.

Locus name	cell lysate [fluorescence counts]	flow-through [fluorescence counts]	binding efficiency
MSMEG0358	5222428	978136	81%
MSMEG0752	2949152	336980	89%
MSMEG1524	10954260	4326100	61%
MSMEG1666	1897892	283148	85%
MSMEG2136	1763528	716720	59%
MSMEG3021	1125388	440396	61%
MSMEG3085	1021644	293248	71%
MSMEG3086	5509748	1315996	76%

Table 3. List of candidates identified after purification of *M. tuberculosis* derived RpoA (Rv3457c) tagged with C-terminal or N-terminal eGFP or SigB (Rv2710) tagged with C-terminal eGFP and expressed in *M. bovis* BCG. Sequence coverage and intensity values

are assigned by MaxQuant software. Intensity values for tagged subunits of DNA-dependent RNA polymerase are highlighted in bold.

Protein IDs	Description	Mol. weight [kDa]	Intensity		
			RpoA-CGFP	NGFP-RpoA	SigB-CGFP
BCG_0716	DNA-directed RNA polymerase. beta subunit (rpoB)	129.2	5.17E+10	4.91E+10	1.40E+10
BCG_0717	DNA-directed RNA polymerase. beta subunit (rpoC)	146.7	5.04E+10	4.85E+10	2.28E+10
Rv3457c/ BCG_3522c	DNA-directed RNA polymerase. alfa subunit (rpoA)	37.7	**3.89E+10**	**3.50E+10**	1.12E+10
BCG_1451	DNA-directed RNA polymerase. omega subunit (rpoZ)	11.8	2.32E+09	2.31E+09	5.98E+08
BCG_2716	sigma factor SigA	57.8	1.14E+09	1.04E+09	5.15E+08
Rv 2710/ BCG_2723	sigma factor SigB	36.3	1.16E+08	8.50E+07	**1.08E+10**

Table 4. **Purification of protein interactors of UvrA after UV-induced DNA damage**. UvrA was eGFP tagged at the C-terminus and expressed in *M. bovis* BCG. Samples were collected at 0, 1, 5, 15 and 30 minutes post-exposure to UV light. Intensity values are given for DNA repair associated proteins identified at listed time points only. Intensity values for tagged protein are highlighted in bold.

Protein IDs	Description	Mol. weight [kDa]	Intensity				
			UvrA 0'	UvrA 1'	UvrA 5'	UvrA 15'	UvrA 30'
Rv1638/ BCG_1676	excinuclease ABC subunit A (UvrA)	106.2	**6.98E+08**	**5.11E+08**	**6.51E+08**	**2.21E+09**	**8.45E+09**
BCG_1671	excinuclease ABC subunit B (UvrB)	78.1	1.20E+07	5.97E+06			2.56E+06
BCG_3704c	DNA topoisomerase I	102.4	1.03E+06		2.64E+06		3.08E+06
BCG_3222c	putative DNA helicase II (UvrD2)	75.6	1.58E+06				1.60E+06
BCG_1667	DNA polymerase I	98.5			4.99E+06		

Figure Legends

Figure 1. **Schematic of the strategy used to purify protein complexes and identify protein-protein interactions in mycobacteria**. (A) For expression of selected bait, pKW08-derived plasmids were engineered. Genes selected for further tests were cloned into constructed vectors, allowing fusion with specific tags. (B) To minimize differences between tags, all tags were designed similarly, containing the SH-quant, a cleavage site for TEV protease, a spacer, and a single specific epitope that terminates with a STOP codon. Recombinant proteins were expressed in mycobacterial cells and bait was purified on epitope specific resin. (C) The anti-GFP nanobodies prepared for this work were immobilized on activated sepharose beads and eGFP binding was examined by microscopy. (D) The overview of the purification procedure followed by LC-MS/MS and

protein identification by MaxQuant software. (E) eGFP can be used to visualize the sub-cellular localization of a target protein.

Figure 2. **Poliacrylamide gel electrophoresis (Novex NuPage) of protein complexes purified using different tags on specific beads**. For each tagging epitope, lanes 1 to 8 represent protein complexes purified from *M. smegmatis* mc²155 expressing tagged proteins of interest. Details are listed in Table 1.

Figure 3. **Number of identified proteins from a specific tag.** (A) Total number of proteins identified by MaxQuant software, and pulled down on FL, GFP, HA and ProtA resins. (B) Number of proteins, purified exclusively with the target on specific beads. This set contains both prey specific for tagged protein as well as proteins not present in other purifications for the same resin.

Figure 4. **Number of identified proteins found as bead-specific background.** These prey proteins were found in all 8 pull downs, regardless of target protein used as bait. The resin specific background details are listed in supplemental Table S3.

Figure 5. **Semiquantitative analysis of co-immunoprecipitation results using SH Quant peptide tagged RpoA protein as bait**. Points corresponding to subunits of the RPO complex are indicated with squares on scatter. Protein abundance was defined as the signal intensity calculated by MaxQuant software for each protein and divided by its molecular weight. Specificity was defined as the ratio of protein signal intensity measured in the bait purification to background level. A protein was arbitrary treated as background if it was found in all 8 purifications and its abundance was set as median intensity of values obtained in all purifications.

Figure 6. **Sub-cellular localization of RpoA (MSMEG1524) fused to eGFP.** Exponentially growing cells were induced with tetracycline for expression of the tagged target and counterstained with DAPI to visualize bacterial chromosomes. Arrows indicate RpoA co-localized with DNA.