OPEN ACCESS Freely available online

PLOS | ONE

# Effects of Mycobacteria Major Secretion Protein, Ag85B, on Allergic Inflammation in the Lung

**Yusuke Tsujimura[1], Hiroyasu Inada[2], Misao Yoneda[3], Tomoyuki Fujita[4], Kazuhiro Matsuo[5], Yasuhiro Yasutomi[1,6]***

**1** Laboratory of Immunoregulation and Vaccine Research, Tsukuba Primate Research Center, National Institute of Biomedical Innovation, Tsukuba, Ibaraki, Japan, **2** Department of Pharmaceutical Sciences, Suzuka University of Medical Science, Suzuka, Mie, Japan, **3** Department of Pathologic Oncology, Institute of Molecular and Experimental Medicine, Faculty of Medicine, Mie University Graduate School of Medicine, Tsu, Mie, Japan, **4** Research Laboratories, Kyoto R&D Center, Maruho Co., Ltd, Chudoji, Shimogyo-ku, Kyoto, Japan, **5** Research and Development Department, Japan BCG Laboratory, Kiyose, Tokyo, Japan, **6** Department of Immunoregulation, Mie University Graduate School of Medicine, Tsu, Mie, Japan

## Abstract

Many epidemiological studies have suggested that the recent increase in prevalence and severity of allergic diseases such as asthma is inversely correlated with *Mycobacterium bovis* bacillus Calmette Guerin (BCG) vaccination. However, the underlying mechanisms by which mycobacterial components suppress allergic diseases are not yet fully understood. Here we showed the inhibitory mechanisms for development of allergic airway inflammation by using highly purified recombinant Ag85B (rAg85B), which is one of the major protein antigens secreted from *M. tuberculosis*. Ag85B is thought to be a single immunogenic protein that can elicit a strong Th1-type immune response in hosts infected with mycobacteria, including individuals vaccinated with BCG. Administration of rAg85B showed a strong inhibitory effect on the development of allergic airway inflammation with induction of Th1-response and IL-17and IL-22 production. Both cytokines induced by rAg85B were involved in the induction of Th17-related cytokine-production innate immune cells in the lung. Administration of neutralizing antibodies to IL-17 or IL-22 in rAg85B-treated mice revealed that IL-17 induced the infiltration of neutrophils in BAL fluid and that allergen-induced bronchial eosinophilia was inhibited by IL-22. Furthermore, enhancement of the expression of genes associated with tissue homeostasis and wound healing was observed in bronchial tissues after rAg85B administration in a Th17-related cytokine dependent manner. The results of this study provide evidence for the potential usefulness of rAg85B as a novel approach for anti-allergic effect and tissue repair other than the role as a conventional TB vaccine.

**Citation:** Tsujimura Y, Inada H, Yoneda M, Fujita T, Matsuo K, et al. (2014) Effects of Mycobacteria Major Secretion Protein, Ag85B, on Allergic Inflammation in the Lung. PLoS ONE 9(9): e106807. doi:10.1371/journal.pone.0106807

**Editor:** Yoshihiko Hoshino, National Institute of Infectious Diseases, Japan

**Received** April 18, 2014; **Accepted** August 6, 2014; **Published** September 5, 2014



**Data Availability:** The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files.

**Funding:** This work was supported by Health Science Research Grants from the Ministry of Health, Labor, and Welfare of Japan and the Ministry of Education, Culture, Sports, Science, and Technology of Japan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have no commercial or financial conflict of interest. The authors also have no competing interest in Maruho Co., Ltd. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

* Email: yasutomi@nibio.go.jp

## Introduction

Epidemiological studies showed that treatments with bacterial and viral products might be effective therapeutic strategies for suppressing the development of allergic responses [1–3]. Administration of mycobacteria, including *Mycobacterium bovis*-Bacillus Calmette Guerin (BCG), has been thought to be effective for preventing the development of asthma by induction of Th1-type immune responses [4], regulatory T (Treg) cells [5,6] and NKT cells [7,8]. On the other hand, recent data have revealed that *Mycobacterium tuberculosis* infection induced not only IFN-γ but also IL-17, which promotes granuloma organization followed by neutrophil recruitment, and IL-22, which promotes regeneration and protects against tissue damage [9]. In addition, vaccination with the mycobacteria-secreted immunogenic protein Ag85A had important links with Th1/Th17 cell induction and Treg cell reduction [10]. However, the role of mycobacteria-mediated Th17-related cytokines in allergic asthma remains unknown.

The airway epithelium and innate immune cells are considered to be essential controllers of inflammatory, immune and regenerative responses to allergens that contribute to asthma pathogenesis [11]. Dysfunction of the epithelium leading to chronic injury was suggested to be a consequence of sustained airway inflammation that is associated with Th2-driven adaptive immunity [12]. Tissue homeostasis at exposed surfaces of the lung is regulated by Th17-related cytokines, especially IL-22, in the innate immune system [13]. Therefore, the functional and structural maintenance of tissue might be necessary to induce both innate and adaptive immunity.

One immunogenic protein that can induce a strong Th1-type immune response in hosts sensitized by BCG is thought to be Ag85B. Ag85B is one of the most dominant protein antigens secreted from all mycobacterial species and has been shown to

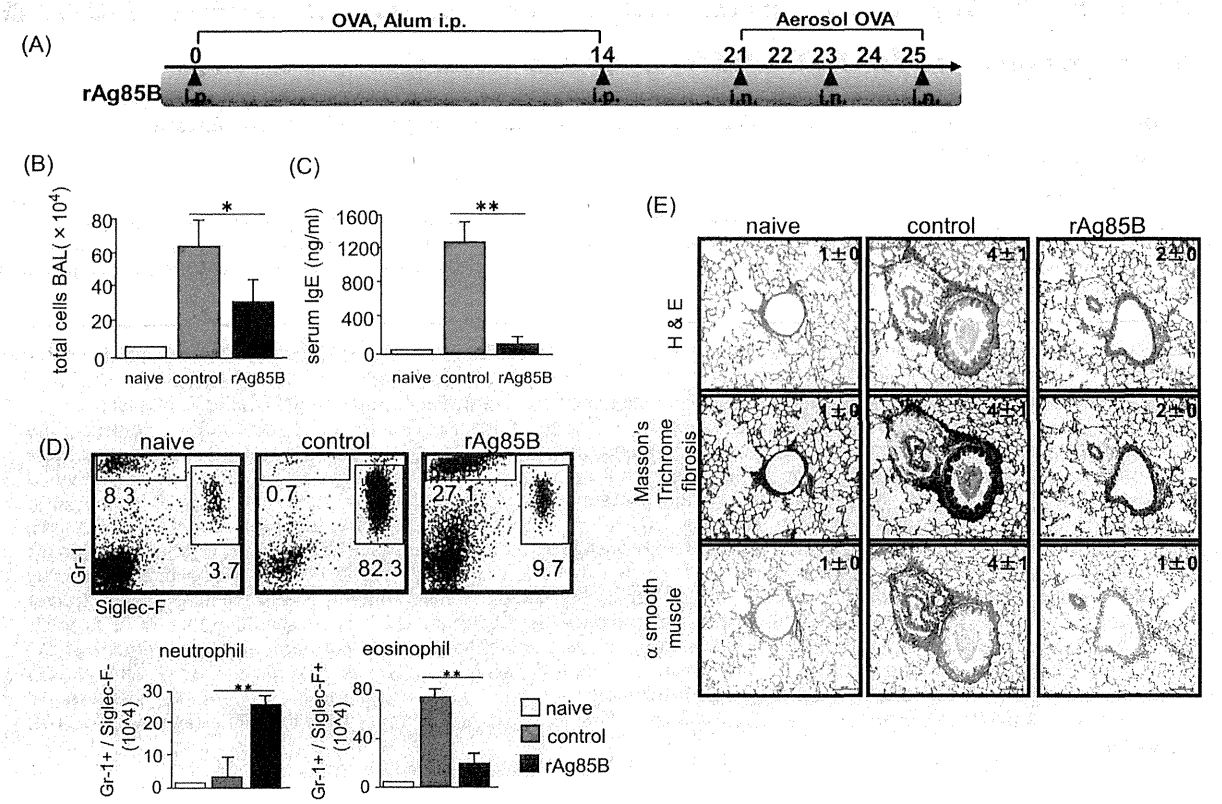


**Figure 1. Functions of rAg85B in allergic inflammation.** Experimental design used to investigate the effects of rAg85B on OVA-induced allergic lung inflammation (*A*). BALB/c mice were intraperitoneally immunized with OVA on days 0 and 14. On days 21 to 25 after the first immunization, mice were exposed to aerosolized 5% OVA for 20 min. Three hours prior to OVA inhalation, the mice were i.p. (100 μg; days 0 and 14) and i.n. (20 μg; days 21, 23, and 25) administered rAg85B. One day after the last challenge, the BAL cells were counted (*B*) and OVA-specific serum IgE concentrations were determined by ELISA (*C*). Flow cytometry of BAL cells from naïve or OVA sensitized BALB/c mice treated with PBS or rAg85B, stained with anti-Gr-1 and anti-Siglec-F. Numbers adjacent to outlined area indicate percent of eosinophils (Gr-1$^{dull}$, Siglec-F$^{+}$), and neutrophils (Gr-1$^{+}$, Siglec-F$^{neg}$) (*D*). Formalin-fixed tissue sections were stained with hematoxylin and eosin to visualized cell recruitment (upper row, scale bar, 100 mm), Masson's trichrome (center row, scale bar, 100 mm), and α-smooth muscle actin (lower row, scale bar, 50 mm). Numbers in quadrants indicate the score scale from 0 to 5 in each. (*E*). Data are representative of at least three independent experiments. (*P<0.05, **P<0.01 compared with OVA control. error bars, s.d.; n = 6 mice).
doi:10.1371/journal.pone.0106807.g001

induce substantial Th cell proliferation and vigorous Th1 cytokine production in humans and mice [14]. In addition, we have reported the possibility of using Ag85B DNA as an immunological strategic tool to induce both Th1 and Treg cells in immunotherapy for atopic dermatitis and allergic asthma [15,16].

In the present study, we found that highly purified recombinant Ag85B protein (rAg85B) had suppressive effects depending on induction of Th1 immune responses in a mouse model of allergic lung inflammation. Remarkably, rAg85B administration also promoted IL-17 and IL-22 production in both Th17 cells in lymph nodes (LNs) and various innate immune cells such as gamma delta T (γδT) cells, NKp46$^{+}$ cells, lymphoid tissue inducer (LTi)-like cells, and CD11c$^{+}$ cells in BAL fluid. More interestingly, Th17-related cytokines induced by rAg85B were involved in enhancement of the expression of genes related to maintenance of tissue homeostasis. This is the first report demonstrating that mycobacteria major secreting protein Ag85B plays an important role in the regulation of allergic airway inflammation by inducing not only a Th1-response but also recruitment of an IL-17 and/or IL-22-producing Th cell subset in LNs and innate immune BAL cells in a manner dependent on Th17-related cytokines in order to retain tissue integrity.

## Materials and Methods

### Animal and Ethic Statement

Specific pathogen-free BALB/c mice (six-week-old, female) were purchased from CLEA Japan. All of the experiments in this study were performed in accordance with the Guidelines for Animal Use and Experimentation, as set out by the National Institute of Biomedical Innovation. The protocol was approved by the Animal Welfare and Animal Care Committee of the National Institute of Biomedical Innovation (Permit Number: DS23-8R2). All animal procedures were used to minimize animal pain and suffering.

### Experimental protocol

BALB/c mice were intraperitoneally immunized with 10 μg ovalbumin (OVA) with 1 mg aluminum hydroxide on days 0 and 14. On days 21 to 25 after the first immunization, mice were exposed to aerosolized 5% OVA for 20 min. Three hours prior to

OVA inhalation, the mice were intraperitoneally (i.p.) (100 μg; days 0 and 14) and intranasal ly (i.n.) (20 μg; days 21, 23, and 25) administered rAg85B. OVA-sensitized Balb/c mice were challenged intranasally with PBS, rAg85B, rAg85B plus 5 μg anti-IL-17 Abs and/or 10 μg anti-IL-22 Abs (R&D Systems) with the same time course as that of rAg85B i.n. administration. The isotype-matched control antibody for neutralization experiments was set using normal goat IgG control (R&D systems).

## Recombinant protein Ag85B production

Plasmids containing the Ag85B gene were transformed into *E. coli* TG1. The expressed inclusion body (IB) was harvested from the disrupted cell pellet by a homogenizer with lysis buffer (30 mM sodium phosphaste, 100 mM NaCl, 5 mM EDTA and 0.5% Triton X-100). This IB of Ag85B was unfolded in 8 M urea and refolded by dilution to 0.4 M urea. The urea in the refolding buffer was removed by anion exchange chromatography using 20 mM Tris buffer and 20 mM Tris buffer with 1 M NaCl (pH 8.5). The refolded Ag85B was loaded on a cation exchange column, and crude Ag85B was passed through the resin using 50 mM NaOAc buffer and 50 mM NaOAc buffer with 1 M NaCl (pH 6.0). Finally, Ag85B was purified by anion exchange chromatography using 20 mM Tris buffer and 20 mM Tris buffer with 1 M NaCl (pH 7.6).

## Endotoxin test

The endotoxin value of Ag85B was measured by Kinetic turbidimetric LAL assay kit (Lonza). Test was carried out according to the manufacture's instruction. The endotoxin value was measured kinetically on ELISA after mixing sample and LAL reagent and was calculated automatically according to standard curve. Purified Ag85B had a purity of >95% analyzed by SDS-PAGE and contaminated less than 0.02 EU/mg of endotoxin. Protein quantitation was carried out by UV spectroscopy at 280 nm.

## Isolation and analysis of lymph node and BAL cells

BAL cells were prepared according to a published protocol [16]. Single cell suspensions from BAL fluid and mediastinal lymph nodes (MLNs) were obtained by crushing through cell strainers. Cells were stained with antibodies to the following markers: CD3, CD4, CD8, CD19, CD11b, CD11c, CD25, γδ TCR, NKp46, Gr-1, Siglec-F, CD127, IFN-γ, IL-4, Foxp3, IL-17 and IL-22 (BD). For analysis of intracellular cytokine production, cells were stimulated directly by incubation for 5 h with 50 ng/ml PMA and 750 ng/ml ionomycin (Sigma-Aldrich) at 37°C and with 10 μg/ml brefeldin A (eBioscience) added in the last 3 h. Flow cytometry data collection was performed on a FACS Calibur (BD). Files were analyzed using CellQuest Software (BD).

## Quantification of cytokines and chemokines

Concentrations of cytokines and chemokines in BAL fluid and culture supernatants of OVA-restimulated lymph node cells were determined by ELISA using commercial kits from R&D Systems. Twenty-four hours after the last OVA sensitization, MLNs and BAL fluid were harvested. MLNs were cultured with 50 μg/ml OVA, and cytokines in the culture supernatant were determined 48 h after incubation. The BAL fluid were measured directly.

## Lung histology

The organs were removed and placed in 4% buffered paraformaldehyde (PFA) overnight. Excess paraformaldehyde was removed by incubation in fresh PBS. Fixed tissues were incubated at 4°C in 70% ethanol. PFA-fixed lung sections were stained with hematoxylin and eosin, Masson's trichrome, and α-smooth muscle actin. Peribronchial infiltrates, fibrosis, and smooth muscle hyperplasia were assessed by a semiquantitative score (0–5) by a pathologist.

## Quantitative real-time PCR

RNA was isolated from whole lung tissue using mechanical homogenization and TRIzol reagent (Invitrogen) according to the manufacturer's instructions. RNA concentrations were measured with a Nanodrop ND 1000 (Nucliber). Omniscript reverse tanscriptase was used according to the protocol of the manufacturer (QIAGEN) for the production of cDNA in a reaction volume of 20 ul. Primers for quantitative real-time RT-PCR were designed with the Universal ProbeLibrary Assay Design Center (Roche Applied Science). Reactions were run on an RT-PCR system (LightCycler 480; Roche Applied Science) Samples were normalized to b-actin and displayed as fold induction over naïve or untreated controls unless otherwise stated.

## TLR/NLR ligand screening

The presence of TLR and NLR ligands were tested on recombinant human embryonic kidney 293 (HEK293) cell lines

**Table 1.** Effects of rAg85B to Toll-Like and NOD-Like Receptor.

| receptor | No ligand | rAg85B | control (+) |
|---|---|---|---|
| mTLR2 | 0.1±0.0 | 0.1±0.0 | 2.3±0.1 |
| mTLR3 | 0.1±0.0 | 0.1±0.0 | 2.4±0.2 |
| mTLR4 | 0.1±0.0 | 0.1±0.0 | 2.7±0.1 |
| mTLR5 | 0.1±0.0 | 0.1±0.0 | 2.7±0.1 |
| mTLR7 | 0.1±0.0 | 0.1±0.0 | 2.1±0.0 |
| mTLR8 | 0.1±0.0 | 0.1±0.0 | 2.3±0.1 |
| mTLR9 | 0.1±0.0 | 0.1±0.0 | 2.6±0.1 |
| mNOD1 | 0.1±0.0 | 0.1±0.0 | 1.7±0.1 |
| mNOD2 | 0.2±0.0 | 0.1±0.0 | 1.6±0.1 |

The results are provided as optical density values (650 nm).
The values represent the means and standard deviations of three screenings.
TLR/NLR ligand screening were performed by InvivoGen, as described in Methods.
doi:10.1371/journal.pone.0106807.t001

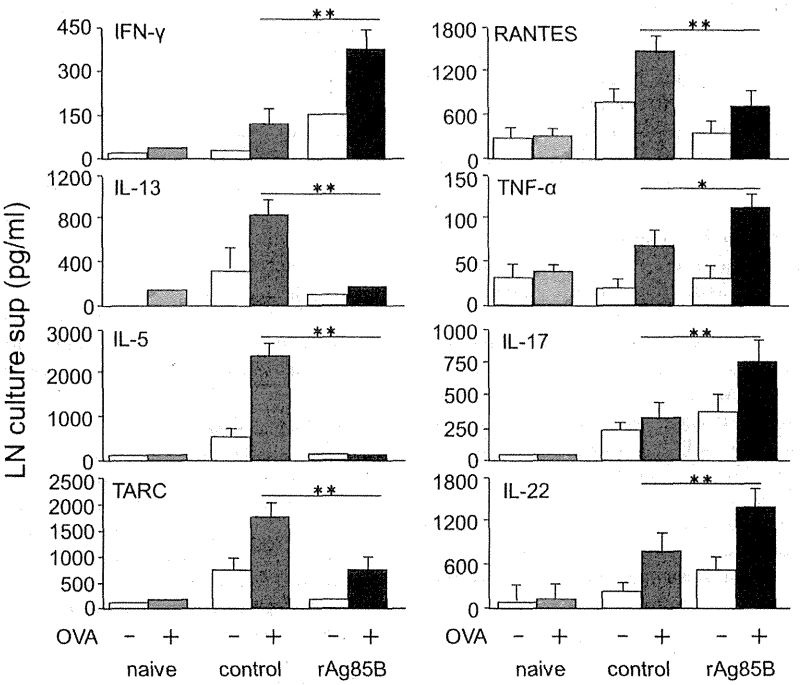


**Figure 2. Administaration of rAg85B induced immune deviation from a Th2-response towards a Th1, Th17-related response in OVA-stimulated LN cells.** OVA-immunized (i.p., day0 and 14) and sensitized (5% aerosolized-OVA, day21 to 25) BALB/c mice were challenged with PBS or rAg85B protein (i.p. (100 μg; days 0 and 14) and i.n. (20 μg; days 21, 23, and 25)). At 24 h after the last OVA sensitization, mediastinal lymph nodes (MLNs) from naïve or OVA sensitized BALB/c mice treated with PBS or rAg85B, were harvested. MLNs were cultured with OVA (50 μg/ml), and cytokines in the culture supernatant were determined 48 h after incubation by ELISA. Data are representative of at least three independent experiments (*P<0.05, **P<0.01 compared with OVA control. error bars, s.d.; n = 6 mice).
doi:10.1371/journal.pone.0106807.g002

which utilize a nuclear factor-kB inducible SEAP (secreted embryonic alkaline phosphatase) reporter gene as the read-out. These HEK293-derived cells are functionally expressing a given TLR or NOD gene from human or mouse. A recombinant HEK293 cell line for the reporter gene only was used as negative control. Positive control ligands are heat-killed *Listeria monocytogenes* (HKLM) for TLR2, Poly(I:C) for TLR3, Lipopolysaccharide (LPS); K12 for TLR4, Flagellin for TLR5, CL097 for TLR7, CL075 and poly(dT) for TLR8, CpG ODN for TLR9, C12-iEDAP for NOD1, and L18-MDP for NOD2. rAg85B (10 μg/mL) was added to the reaction volume. TLR/NLR ligand screening were performed by InvivoGen.

### Statistical analysis

Data are shown as means±SD. Statistical significance of differences between the OVA-control group and rAg85B-treated group was assessed by the non-parametric Mann-Whitney U-test. Statistical comparisons between groups of rAg85B+isotype control and rAg85B+neutralization antibody were performed using the non-paramatric Kruskal-Wallis H-test.

## Results

### Effects on allergic inflammation by administration of rAg85B

To investigate the role of rAg85B in pulmonary allergic inflammation, we examined the frequently used mouse model of ovalbumin (OVA)-induced allergic lung inflammation. The mice were intraperitoneally (i.p.) (days 0 and 14) and intranasally (i.n.) (days 21, 23, and 25) administered with rAg85B (Fig. 1*A*). The purity of rAg85B was evaluated by silver staining of SDS-PAGE gel (Fig. S1) and the Limulus Amebocyte Lysate (LAL) assay (less than 0.02 EU (endotoxin units)/ml). Furthermore rAg85B was not contaminated with any TLR/NLR binding immune stimulants (Table 1). Twenty-four hours after the final OVA challenge, inflammatory cell recruitment into the lungs was analyzed. The OVA-induced allergic manifestation was suppressed with a decrease in the total number of bronchoalveolar lavage (BAL) cells and serum IgE level in the rAg85B-administered mice (Fig. 1*B, 1C*). A marked reduction in eosinophil (Gr-1(+)/Siglec-F(+)) infiltration was observed by flow cytometric (FACS) analysis of BAL in rAg85B-administered mice (Fig. 1*D*). In association with decreased eosinophilia, neutrophil (Gr-1(+)/Siglec-F(-) ) recruitment was seen in rAg85B-administered mice (Fig. 1*D*). These results were confirmed by histopathological observation of hematoxylin and eosin (H&E) staining (Fig. 1*E*). Mice administered rAg85B showed inhibition of infiltration of cells. (Fig. 1*E*). Lung sections were also stained with Masson's trichrome to evaluate fibrosis, and stained with α-smooth muscle actin. Sizes of both the peribronchial smooth muscle area and lung fibrosis area were increased in OVA-sensitized control mice; however, mice administered rAg85B showed strong suppression of both fibrosis and α-smooth muscle actin expression as well as reduction in inflammation severity assessed by H&E staining (Fig. 1*E*). These observations indicated that rAg85B has a critical function of regulating airway inflammation in a mouse model of allergen-induced asthma. Moreover, rAg85B i.n. administration induced wound repair including suppression of both fibrosis and α-smooth

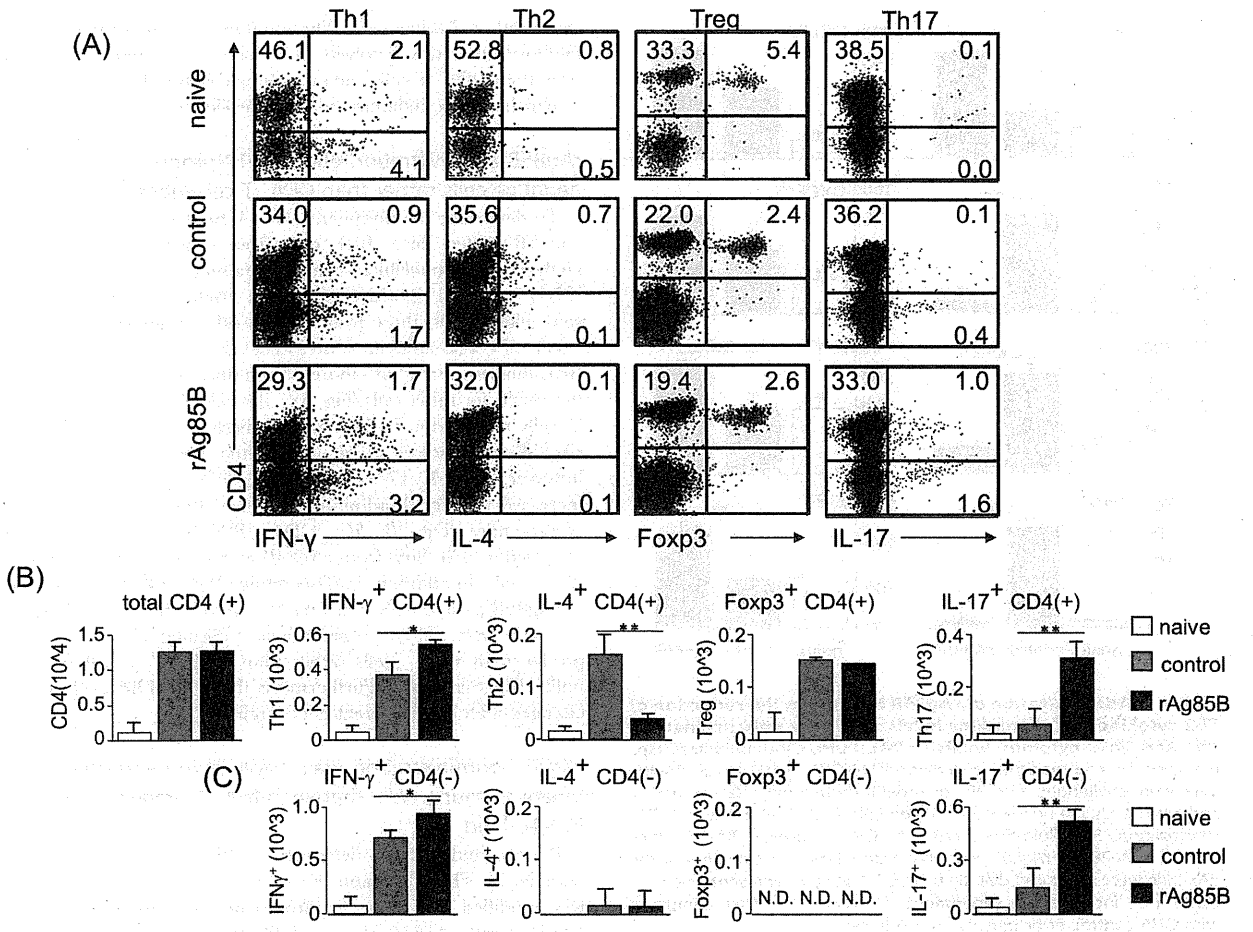


**Figure 3. IFN-γ and IL-17-producing CD4 T cell subsets proliferated in lymph nodes after rAg85B administration.** OVA-immunized (i.p., day0 and 14) and sensitized (5% aerosolized-OVA, day21 to 25) BALB/c mice were challenged with PBS or rAg85B protein (i.p. (100 μg; days 0 and 14) and i.n. (20 μg; days 21, 23, and 25)). At 24 h after the last OVA sensitization, mediastinal lymph nodes (MLNs) from naïve or OVA sensitized BALB/c mice treated with PBS or rAg85B, were harvested. MLNs were stimulated with ionomycin and PMA for 5 h, and with brefeldin A added in the last 3 h. Flow cytometry of stimulated MLNs from naïve (upper), PBS-treated (middle) and rAg85B protein-treated (lower) OVA-sensitized mice stained with specific antibodies indicated marker. Numbers in quadrants indicate percent of cells in each (*A*). Absolute numbers of various cell populations (above graphs) in lymph nodes (*B, C*). Data are representative of three independent experiments (*P<0.05, **P<0.01 compared with OVA control. error bars, s.d.; n = 6 mice).
doi:10.1371/journal.pone.0106807.g003

muscle actin expression. Incidentally, previous data showed that either i.p. (days 0, 14) or i.n. (days 21, 23, and 25) challenge with rAg85B did not induce strong suppression of Th2-response in OVA-sensitized mice (data not shown).

## Immune deviation from a Th2-response towards a Th1, Th17-related response by rAg85B administration

We next assessed the production of OVA-specific cytokines in lymph node cells after *in vitro* stimulation with OVA (Fig. 2). Cells from mediastinal lymph nodes (mLNs) were stimulated *in vitro* with OVA and the production of various types of cytokines was assessed. The level of the Th1 cytokine IFN-γ in culture supernatants of cells from rAg85B-administered mice was increased. On the other hand, the levels of Th2 cytokines IL-5 and IL-13 in culture supernatants of cells from rAg85B-administered mice were lower than those in culture supernatants of cells from control mice. Similarly, mice administered rAg85B showed inhibition of production of the CCL5 (RANTES) and the thymus- and activation-regulated chemokine CCL17 (TARC), which contribute to allergic inflammation. Production of IL-17, IL-22 and TNF-α was also enhanced in culture supernatants of OVA-stimulated mLN cells from rAg85B-administered mice. These results suggested that Th1 and Th17 cytokines are crucial factors in the suppressive effect of rAg85B on airway inflammation.

## $CD4^+$ T cells producing IFN-γ and IL-17 were increased in mediastinal lymph nodes by rAg85B administration

We next examined Th cell responses in the mouse asthma model by intracellular staining analysis. mLN cells were stimulated with or without PMA and ionomycin, and cell fractions were analyzed by intracellular cytokine staining. Stained $CD4^+$ T cells producing IFN-γ or IL-17 were increased in mice administered rAg85B, whereas IL-4-secreting cells were decreased in those mice (Fig. 3*A*, *3B*). On the other hand, rAg85B administration was not associated with the induction of Treg cells, which express Foxp3

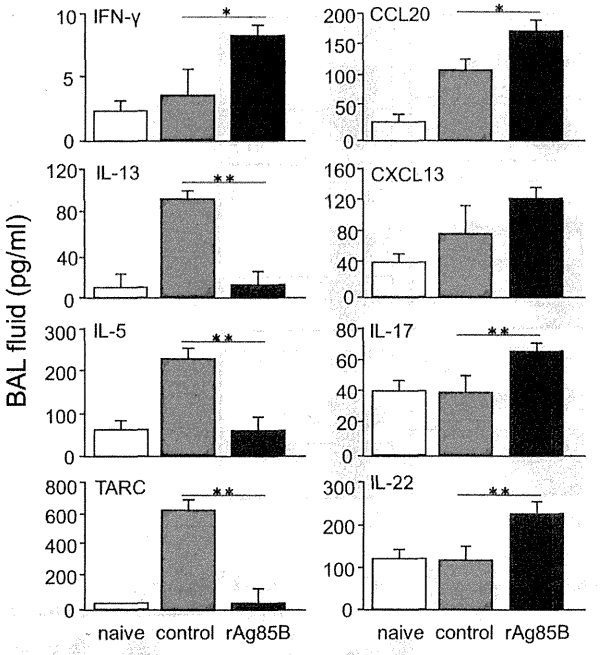


**Figure 4. Administration of rAg85B resulted in the reduction of Th2 cytokine and chemokine levels and in the enhancement of Th1 and Th17 cytokine levels in BAL fluid.** OVA-immunized (i.p., day0 and 14) and sensitized (5% aerosolized-OVA, day21 to 25) BALB/c mice were challenged with PBS or rAg85B protein (i.p. (100 μg; days 0 and 14) and i.n. (20 μg; days 21, 23, and 25)). At 24 h after the last OVA sensitization, BAL fluid from naïve or OVA sensitized BALB/c mice treated with PBS or rAg85B, were harvested. Levels of cytokines in the BAL fluid were measured directly by ELISA. Data are representative of at least three independent experiments (*P<0.05, **P<0.01 compared with OVA control. error bars, s.d.; n = 6 mice).
doi:10.1371/journal.pone.0106807.g004

and CD25, in LNs (Fig. 3*A*, *3B*, S2). These results were the same for not only the fraction of $CD4^+$ T cells but also the fraction of $CD4^-$ cells producing cytokines in mLNs (Fig. 3*C*). These results suggested that rAg85B administration was involved in the induction of IFN-γ or IL-17-producing $CD4^+$ T cells and $CD4^-$ cells in LNs.

## Mice administered rAg85B showed reduction in levels of Th2 cytokines and chemokines levels and increase in levels of Th1 and Th17 cytokines in BAL

The pathogenesis of asthma is associated with many cell types and several molecular/cellular pathways in the lung. Therefore, we investigated whether rAg85B administration regulates various cytokines associated with the pathogenesis of allergic inflammation in BAL fluid. Control mice in which allergic inflammation developed showed increased production of Th2 cytokines and chemokines in BAL fluid, such as IL-13, IL-5 and TARC. Mice administered rAg85B showed inhibition of the induction of IL-13, IL-5 and TARC (Fig. 4). Furthermore, enhancement of IFN-γ, IL-17 and IL-22 production was observed in BAL fluid from mice administered rAg85B. Production of chemokines secreted from non-T cells, CCL20 and CXCL13, was also increased in BAL fluid from rAg85B-administered mice. The chemokine CCL20 is thought to be associated with the recruitment of Th17 lymphocytes and LTi-like or NK-like cells [17,18], and CXCL13 is a chemokine ligand of C-X-C motif receptor 5 (CXCR5) that is expressed on Lti-like cells. These findings suggested that rAg85B administration was involved in induction of immune responses from both $CD4^+$ T cells and other innate cells in BAL fluid of mice in which allergic inflammation has developed.

## rAg85B administration elicits IL-17-producing CD4-negative cells rather than $CD4^+$ T cell subsets in BAL fluid

To determine the peripheral Th cell population in the lungs of rAg85B-treated mice, BAL cells from experimental mice were analyzed by intracellular cytokine staining. The percentages of IFN-γ and IL-17-positive cells from rAg85B-administered mice were higher than those from control cells in agreement with the results of FACS analysis of mLN cells (Fig. 5*A*), and Treg cells in BAL fluid from rAg85B-administered mice were also the same as the results for mLN cells (Fig. 5*A*). The absolute number of $CD4^+$ T cells stained for IL-17 was not increased in BAL cells from rAg85B-administered mice, unlike the results for mLN cells; however, total IL-17-secreting cells, $CD4^-$ $IL\text{-}17^+$ cells, were increased in rAg85B-administered mice compared with those in control mice (Fig. 5*B*, *5C*). $CD4^-$ IFN-γ-producing cells were observed in BAL fluid from rAg85B-administered mice as same to $CD4^+$ cells. In addition, IL-4-secreting $CD4^+$ cells were decreased in rAg85B-treated mice, whereas IL-4-producing $CD4^-$ cells were not observed. These observations indicated that IL-17 was produced by $CD4^-$ cells rather than by $CD4^+$ T cells in BAL, unlike IFN-γ and IL-4. Furthermore, the types of BAL cells greatly changed after rAg85B treatment (Figure S3).

## rAg85B administration was involved in recruitment of innate immune cells that secrete IL-17-related cytokines in BAL fluid

Recent studies have demonstrated that IL-17 was not only secreted by Th17 cells and the source of Th17-related cytokines was modified in various environmental conditions [19]. Mice administered rAg85B showed infiltration of CD4-negative immune cells, which secreted IL-17 cytokine in BAL fluid (Fig. 4, Fig. 5). From these findings, we next investigated the proportions of infiltrating $CD4^-$ cells that produce IL-17, including γδT cells, $IL\text{-}7R^+$ $Lin^-$ cells (LTi-like cells), $CD3^-$ $NKp46^+$ cells and $CD11c^+$ cells, in BAL fluid from experimental mice. OVA-sensitized BALB/c mice administered rAg85B, but not mice administered PBS, showed an increased number of innate immune cells in BAL fluid (Fig. 6*A*). The percentages of $CD4^+$ and $CD8^+$ T cells in BAL fluid from rAg85B-administered mice were similar to those in BAL fluid from control mice. However, the percentages of γδT cells, LTi-like cells, $NKp46^+$ cells, and $CD11c^+$ cells in BAL fluid from rAg85B-administered mice were higher than those in BAL fluid from control mice (Fig. 6*A*). Since innate immune cells, which secrete IL-17 and related cytokines, IL-22, were thought to be induced by rAg85B administration, we next explored the source of IL-17-related cytokines in BAL fluid. Small numbers of IL-17-producing γδT cells, LTi-like cells and $CD11c^+$ cells were observed (Fig. 6*D, 6F, 6G*), while production of IL-17 from CD8+ T cells and $NKp46^+$ cells was not detected (Fig. 6*C, 6E*). In the present study, a Th17-related cytokine, IL-22, was also detected in BAL fluid from mice administered rAg85B. (Fig. 4). All of the cells from BAL secreting Th17-related cytokines, including $CD4^+$ T cells, γδT cells, $NKp46^+$ cells, LTi-like cells and $CD11c^+$ cells, that were examined in this study showed IL-22 production in mice administered rAg85B (Fig. 6*B, 6D, 6E, 6F, 6G*). On the other hand, production of IL-17 from $NKp46^+$ cells and $CD11c^+$ cells were not detected (Fig. 6*E, 6G*). Although it is now known that NKT cells, alveolar macrophages and neutrophils might also

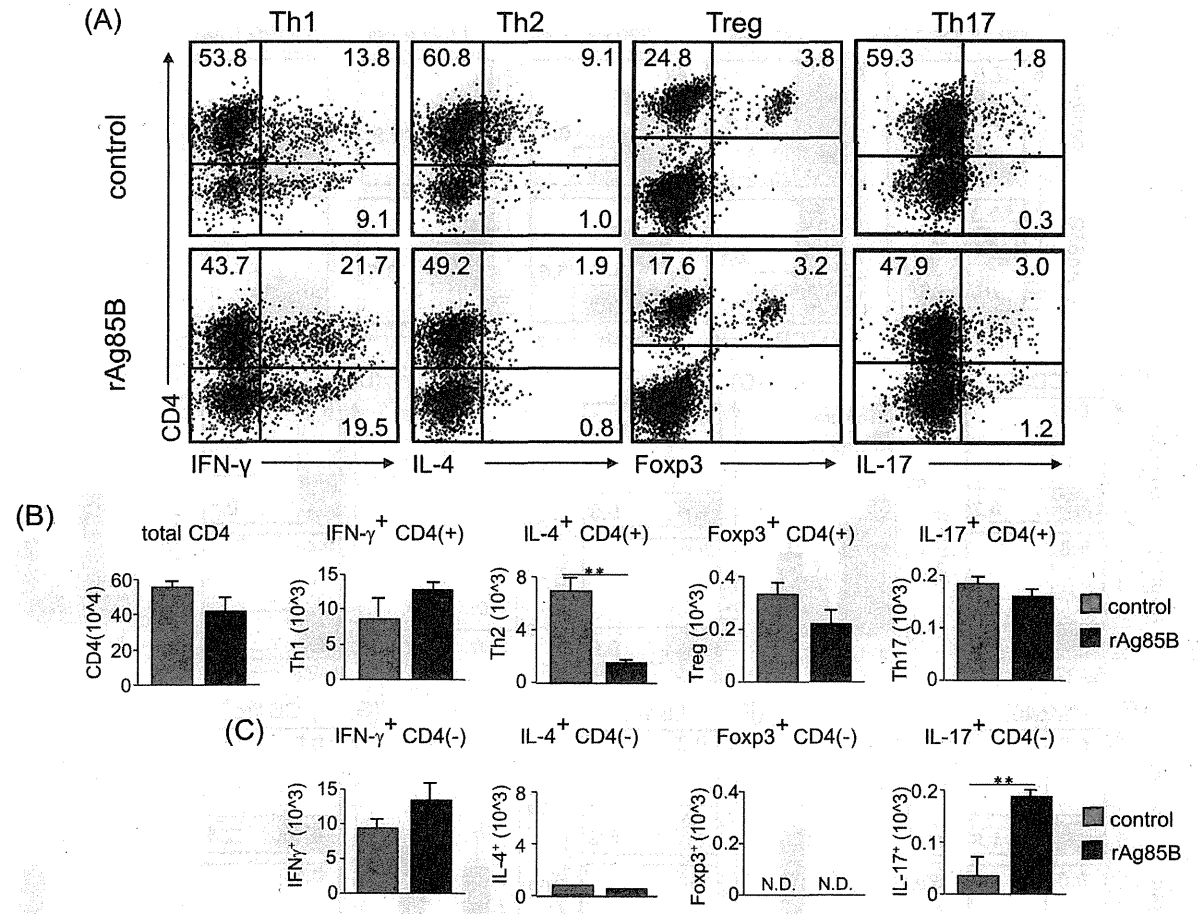


**Figure 5. IFN-γ and IL-17-producing CD4-negative cell subsets proliferated in BAL fluid after rAg85B administration.** OVA-immunized (i.p., day0 and 14) and sensitized (5% aerosolized-OVA, day21 to 25) BALB/c mice were challenged with PBS or rAg85B protein (i.p. (100 μg; days 0 and 14) and i.n. (20 μg; days 21, 23, and 25)). At 24 h after the last OVA sensitization, BAL fluid from naïve or OVA sensitized BALB/c mice treated with PBS or rAg85B, were harvested. BAL cells were stimulated with ionomycin and PMA for 5 h, and with brefeldin A added in the last 3 h. Flow cytometry of stimulated BAL cells from PBS-treated (upper) and rAg85B protein-treated (lower) OVA-sensitized mice stained with specific antibodies indicated marker. Numbers in quadrants indicate percent of cells in each (*A*). Absolute numbers of various cell populations (above graphs) in BAL fluid (*B*, *C*). Data are representative of three independent experiments (**P<0.01 compared with OVA control. error bars, s.d.; n = 6 mice).
doi:10.1371/journal.pone.0106807.g005

produce IL-17 in certain conditions, the numbers of these IL-17-secreting cells in BAL fluid from rAg85B-administered mice showed little or no change compared with those in the control group in our experimental setting. Remarkably, these responses induced by rAg85B were observed in allergic animals but not in naïve ones (data not shown).

## Functions of IL-17 and IL-22 in rAg85B-administered mice

We next investigated the importance of Th17-related cytokines by using neutralizing antibodies (Abs) to IL-17 and IL-22 in rAg85B-administered experimental mice. Administration of neutralizing Abs to IL-17 and IL-22 did not show any systemic inhibitory effects induced by rAg85B as a result of IgE production (Fig. 7*A*). Furthermore, neutralization of IL-17 and IL-22 did not restore the functions of rAg85B with immune deviation from a disease-promoting Th2 response towards a Th1 response, whereas inhibition of TARC production regulated by rAg85B was reversed by neutralizing IL-22 Abs treatment (Fig. 7*B*). These results suggested that IL-17 and IL-22 induced by rAg85B have little or no systemic inhibitory effect on the development of allergic inflammation in the lung. Neutralization of IFN-γ at the challenge phase also had little or no suppressive effect on serum IgE expressions and eosinophilia induced by rAg85B treatment. (data not shown). The number of infiltrating cells in BAL fluid were also not changed in mice administered neutralizing Abs to IL-17 and IL-22 (Fig. 7*C*); however, fractions of infiltrating cells in BAL fluid were different. Neutralization of IL-17 by IL-17-specific Abs prevented neutrophil infiltration by rAg85B administration in the airway, and this preventive effect on infiltration of neutrophils was partial in IL-22-specific Abs administered mice (Fig. 7*D*). Eosinophilia suppression by rAg85B administration was reversed by neutralizing IL-22 Abs treatment (Fig. 7*D*). These results parallel previous observations of the specificity of IL-17 and IL-22 effects [20]. Enhancement of innate immune cell recruitment induced by rAg85B was fully reversed by neutralizing IL-17 Abs treatment, and this rAg85B effect was partially reversed by administration of neutralizing IL-22 Abs in γδT cells (Fig. 7*D*). These results showed that Th17-related cytokines induced by rAg85B have pivotal roles in innate immune cell recruitment in BAL fluid and in severity of

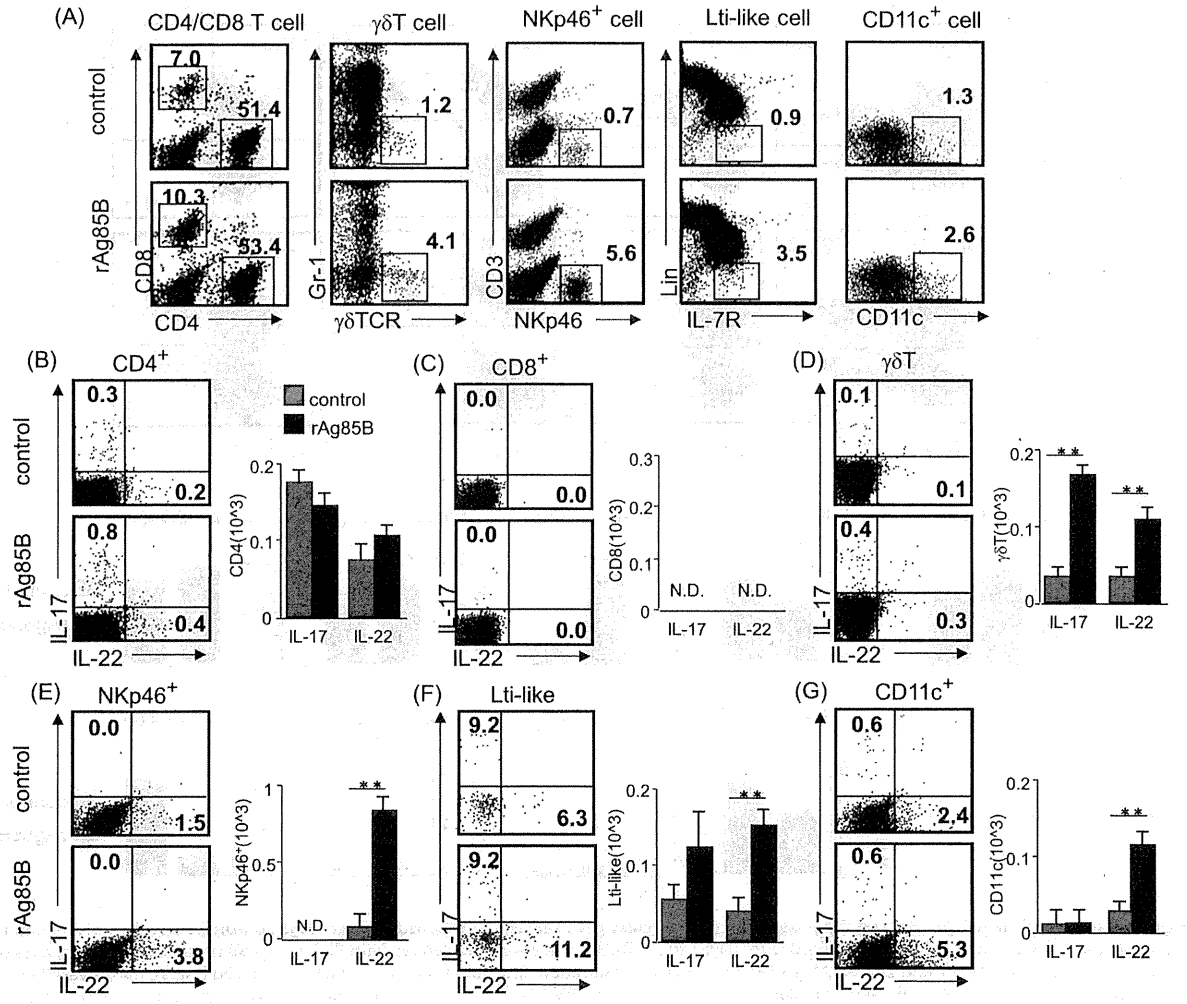


**Figure 6. Innate immune cells that secrete Th17-related cytokines are induced by rAg85B administration in BAL fluid.** OVA-immunized (i.p., day0 and 14) and sensitized (5% aerosolized-OVA, day21 to 25) BALB/c mice were challenged with PBS or rAg85B protein (i.p. (100 μg; days 0 and 14) and i.n. (20 μg; days 21, 23, and 25)). At 24 h after the last OVA sensitization, BAL fluid from naïve or OVA sensitized BALB/c mice treated with PBS or rAg85B, were harvested. BAL cells were stimulated with ionomycin and PMA for 5 h, and with brefeldin A added in the last 3 h. Flow cytometry of BAL cells from PBS-treated (upper) and rAg85B protein-treated (lower) OVA-sensitized mice stained with anti-CD3, anti-CD4, anti-CD8, anti-Gr-1, anti-γδ TCR, anti-NKp46, anti-CD11c, anti-CD127 (IL-7R) and Lineage specific marker (CD3, CD19, Gr-1, CD11b, CD11c). Numbers in quadrants indicate percent of cells in each (*A*). Intracellular IL-17 and IL-22 staining in indicated cells by flow cytometry (dot plots) and absolute numbers of those cell populations (side graphs) in the BAL fluid (*B, C, D, E, F, G*). Data are representative of at least two independent experiments (**P<0.01 compared with OVA control. error bars, s.d.; n = 6 mice).
doi:10.1371/journal.pone.0106807.g006

lung inflammation but not in regulated systemic allergic inflammation involving Th responses.

## Administration of rAg85B promoted Th17-related innate responses in the lung

Our data suggested an important link between rAg85B and airway innate immune cells producing IL-17 and IL-22 that contributed to the homeostasis expression of Th17-related cytokine response genes. However, these two cytokines induced by rAg85B administration did not clearly show inhibitory effects on systemic allergy responses (Fig. 7*A*, 7*B*). From these findings, we next explored the relationship between the roles of airway innate immune cells and wound repair in mice that received i.n. administration of rAg85B. Mice that received IL-17 or IL-22 or both neutralizing antibodies showed a marked induction of fibrosis and actin staining but incomplete cancellation of rAg85B suppressive effects at the same levels as those in OVA control mice (Fig. 8*A*). Histological findings suggested that IL-17 and IL-22 induced by rAg85B i.n. administration were partially involved in regulation of local tissue allergic inflammation. The inhibition of rAg85B effects by neutralizing Abs of IL-17 and IL-22 to allergic inflammation was partial; however, tissue repair in lungs was seen in rAg85B-administered mice by histopathological examination. These results led us to hypothesize that IL-17 and IL-22 induced by rAg85B induced local tissue remodeling/repair molecules. To confirm this, the induction of tissue homeostasis-related gene expression in rAg85B-administered mice was examined by real-time RT-PCR. Rb2, Cyclin D1 and c-Myc are associated with

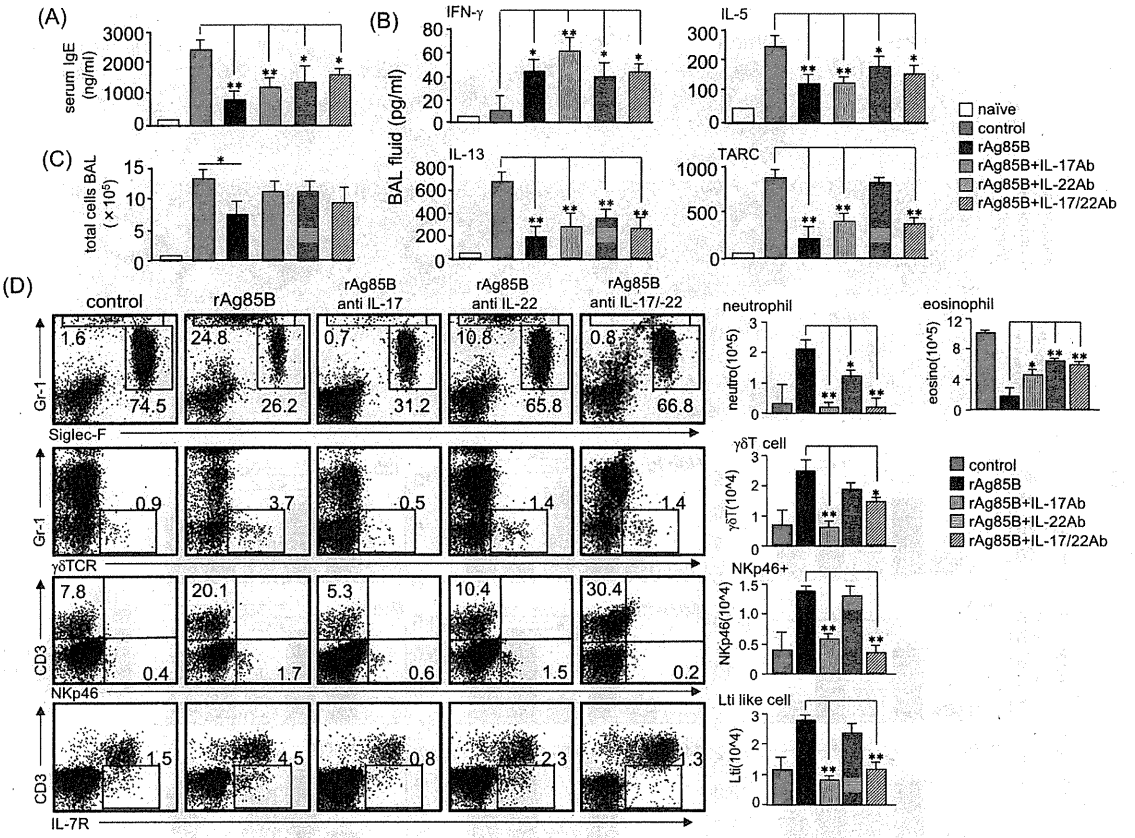


**Figure 7. Neutralization of Th17-related cytokines inhibits cell recruitment to the lung but does not change cytokine and chemokine production.** OVA-sensitized BALB/c mice (5% aerosolized-OVA, day21 to 25) were challenged intranasally with PBS (control), rAg85B (rAg85B + isotype-matched control antibody (Ab)), rAg85B plus neutralizing IL-17 (rAg85B + IL-17Ab) or IL-22 (rAg85B + IL-22Ab), or a combination of both antibodies (rAg85B +IL-17/22Ab) on days 21, 23, and 25. The isotype control was treated with the same time course as neutralization Ab i.n. administration. One day after the last challenge, OVA-specific serum IgE concentration and levels of cytokines and chemokines in BAL fluid were determined by ELISA (*A, B*). BAL cells from naïve or OVA sensitized BALB/c mice treated with PBS or rAg85B with/without neutralization Ab were counted (*C*), and were stained with anti-Gr-1, anti-Siglec-F, anti-gd TCR, anti-CD3, anti-NKp46, and anti-CD127 for flow cytometric analysis. Numbers adjacent to outlined area indicate percent of eosinophils (Gr-1$^{dull}$, Siglec-F$^{+}$), neutrophils (Gr-1$^{+}$, Siglec-F$^{neg}$), γδT cells (Gr-1$^{neg}$, γδTCR$^{+}$), NKp46$^{+}$ cells (CD3$^{neg}$, NKp46$^{+}$), LTi like cells (CD3$^{neg}$, L-7R$^{+}$), and absolute numbers of those cell populations (side graphs) in BAL fluid (*D*). Data are representative of at least two independent experiments (*P<0.05, **P<0.01 compared with rAg85B+isotype control challenged group. error bars, s.d.; n = 6 mice).
doi:10.1371/journal.pone.0106807.g007

wound healing, tissue repair and remodeling including proliferative molecules [21]. Muc1 [22,23], matrix metalloproteinase 13 (MMP13) [24], and the extracellular matrix proteins decorin and dermatopontin [25] produce protective mucus. Lymphotoxin-beta (Ltb) is a molecule related to signaling in stromal cells to produce factors that organize lymphoid cells into lymph nodes [26]. The transcription of Reg3γ is involved in tissue repair and antimicrobial responses [27]. The expression of these genes involved in innate immune response-mediated signaling was significantly enhanced in the lungs of rAg85B-administered mice (Fig. 8*B*). The increases in mRNA levels of all molecules other than Reg3γ and dermatopontin were inhibited by treatment with neutralizing Abs of IL-17 (Fig. 8*B*). On the other hand, the expression of mRNA of molecules enhanced by rAg85B administration was decreased after treatment with IL-22 neutralizing Abs except for Rb2, Cyclin D1, c-Myc, Mmp13 and Muc1 (Fig. 8*B*). These results suggested that IL-17 and IL-22 induced by rAg85B administration affected induction of pulmonary innate response. In conclusion, IL-17 induced by rAg85B administration induced the expression of various types of wound healing, tissue repair and remodeling molecules. Interestingly, IL-22 in rAg85B-immunized mice induces the expression of molecules mainly associated antimicrobial responses such as Reg3γ, decorin, dermatopontin and Ltb. In summary, Th1 and Th17 cells are induced in regional lymph nodes by administration of rAg85B; however, Th17 cells are not induced in BAL unlike in Th1 cells. IL-17 is produced by innate immune cells with IL-22 production. IL-17 and IL-22 are important in not only anti-allergic effects, such as eosinophil inhibition, but also wound healing and tissue repair in the lung (Fig. 9).

## Discussion

Results of several experimental studies on mycobacteria involving mycobacterial antigens in mouse models of allergic airway inflammation have been reported. In murine asthma models, intranasal administration of BCG suppressed asthma manifestations probably through Th1 response [2,4], Treg cells [5,6,28], or NKT cells [7,8,29]. In our experimental setting using rAg85B protein, we did not find any detectable effect or

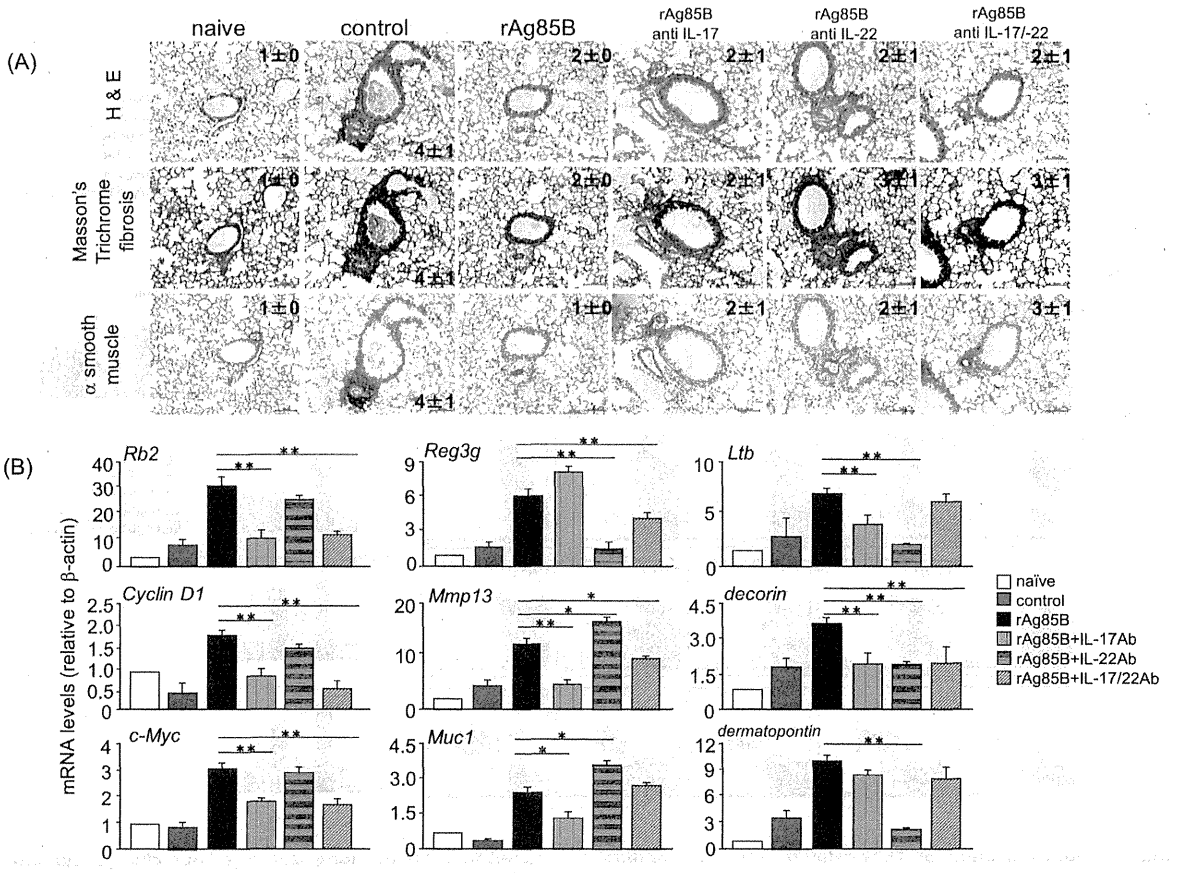


**Figure 8. Ag85B administration promotes Th17-related innate responses in the lung.** OVA-sensitized BALB/c mice (5% aerosolized-OVA, day21 to 25) were challenged intranasally with PBS (control), rAg85B (rAg85B + isotype control), rAg85B plus neutralizing IL-17 (rAg85B + IL-17Ab) or IL-22 (rAg85B + IL-22Ab), or a combination of both antibodies (rAg85B +IL-17/22Ab) on days 21, 23, and 25. The isotype-matched control antibody was treated with the same time course as neutralization Ab i.n. administration. Lungs from naïve or OVA sensitized BALB/c mice treated with PBS or rAg85B with/without neutralization Ab were sampled one day after the last challenge for histological analysis and quantification of mRNA levels. Lung sections were stained with hematoxylin and eosin (left row, scale bar, 100 mm), Masson's trichrome (center row, scale bar, 100 mm), α-smooth muscle actin (right row, scale bar, 50 mm). Numbers in quadrants indicate the score scale from 0 to 5 in each (*A*). Real-time RT-PCR was performed for the indicated molecules expression on RNA isolated from individual mice lungs (*B*). Data are representative of at least two independent experiments (*P<0.05, **P<0.01 compared with rAg85B+isotype control challenged group. error bars, s.d.; n = 6 mice).
doi:10.1371/journal.pone.0106807.g008

substantial change in numbers of both NKT cells and Treg cells in BAL fluid from rAg85B-administered mice. Moreover, microarray analysis revealed that the gene expression pattern of splenocytes stimulated with rAg85B and that of splenocytes stimulated with BCG were very different (data not shown). This discrepancy in the effects of vaccination with BCG and vaccination with rAg85B might be related to the factors affecting immune responses. BCG contains many essential components to induce early immune response such as glycolipid and DNA, whereas rAg85B is a single immunogenic protein.

The present study indicated that Th17-related immune responses induced by rAg85B administration had a suppressive effect on allergic airway inflammation, and we attributed this suppressive effect to the larger proportion of Th17-related cytokine-producing innate immune cells in BAL fluid. It has been reported that Mycobacterium antigens increased the number of γδT cells that express IFN-γ [30] or IL-17 [31], and these responses induce healing to epithelial surfaces [32]. Given the integral role of γδT cells in innate immunity, γδT cells are one of the crucial factors in the rAg85B immune regulatory functions. Interestingly, our study also showed that IL-22-producing cells in lungs of rAg85B-administered mice were $NKp46^+$ cells, LTi-like cells, γδT cells and $CD11c^+$ cells. This is the first time demonstration of an important link between the mycobacterium antigen rAg85B and IL-22-producing cells. Although we could not rule out the possibility of Th17 cytokine-producing cells other than those described here, $NKp46^+$ cells, LTi-like cells, γδT cells and $CD11c^+$ cells were thought to be rapid innate sources of IL-22, which is required in the early stage to maintain epithelial cell integrity and to suppress eosinophilia. Moreover, IL-22 can act synergistically or additively with other cytokines, including IL-17 or TNF, to promote gene expression for antimicrobial peptides, chemokines, matrix metalloproteinases, cytokines, and acute-phase proteins from epithelial cells in the lung [33]. These findings also support our results showing that simultaneous induction of these cytokines and expression of many genes may be beneficial functions of rAg85B treatment in local allergic pathology.

Pulmonary infection of mycobacteria induced not only a neutrophil-mediated response but also T cell-mediated IFN-γ

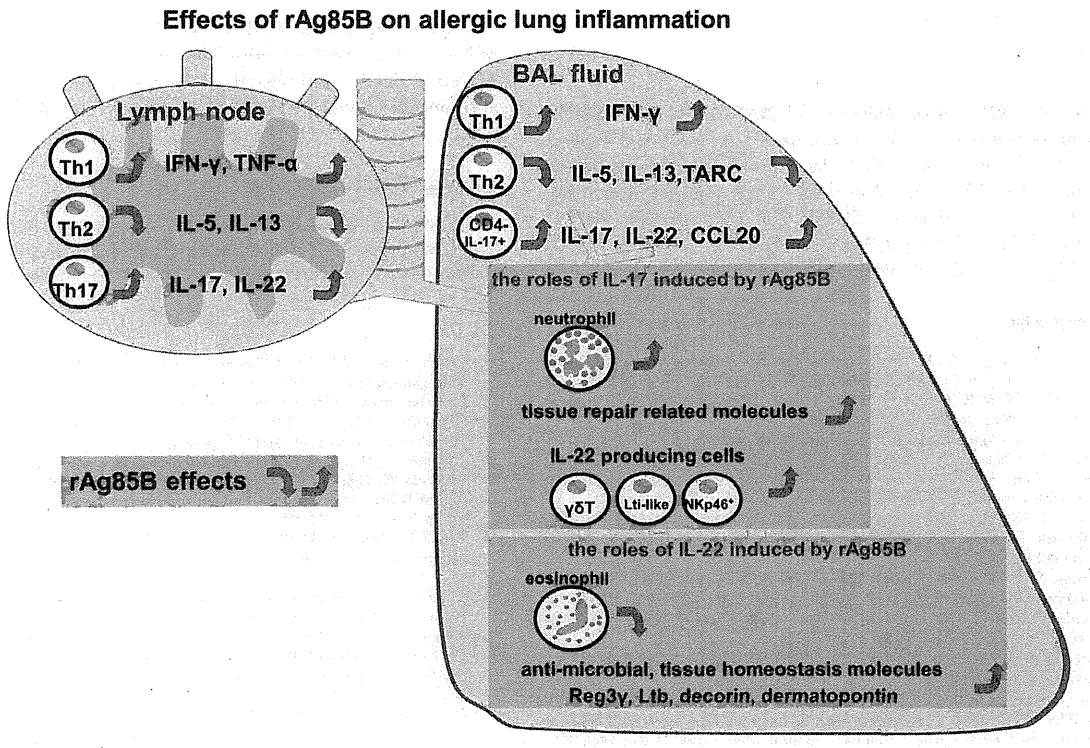


**Figure 9. Schematic illustration of the proposed effects of rAg85B in a mouse model of allergic inflammation.** IFN-γ and IL-17-producing Th cells are induced in regional lymph nodes by rAg85B challenge, however, Th17 cells do not enter the lung unlike Th1 cells. Th17-related cytokine-secreting cells in lungs from rAg85B-administered mice are innate immune cells including γδT cells, IL-7R$^+$ Lin$^-$ cells, CD3$^-$ NKp46$^+$ cells and CD11c$^+$ cells. IL-17 and IL-22 induced by rAg85B in an allergic environment have crucial roles in not only anti-allergic effects but also regulation of tissue homeostatic reactions.
doi:10.1371/journal.pone.0106807.g009

production and granuloma formation depending on IL-17 from especially γδT cells [34]. The hallmark of mycobacterial infection in the lung is granuloma formation with infiltrating neutrophils, which creates an immune microenvironment in which the infection can be controlled. On the other hand, it also provides the mycobacterium with a niche in which it can survive, modulating the immune response to ensure its survival without damage over a long period of time [35,36]. Mature granulomas include fibroblasts and extracellular matrix, which surround and separate the granulomas from the normal environment. Administration of anti-IL-17 Abs during the inhalatory challenge phase abolished the bronchial neutrophilia and the upregulation of genes related to tissue repair and homeostasis observed in rAg85B-administered mice in the present study. Neutrophils may also promote epithelial healing [37] and are now known to be rich sources of prestored and expressible proteins [38] that may directly promote wound healing [39,40]. In the present study, induction of neutrophilia and upregulation of described genes related to wound healing with suppression of tissue injury might be the mechanisms of granuloma formulation induced by mycobacteria infection.

In the present study, IFN-γ and Th17-related cytokines were key factors to regulate allergic severity in our experimental setting. Although infiltration of Th1 and Th17 cells elicited by rAg85B was induced in pulmonary lymph nodes, such effector cells were not increased in BAL fluid of mice showing anti-allergic effects of Ag85B administration. Moreover, our results suggested that the accumulation of neutrophils and IL-17 and/or IL-22-producing innate immune cells contributed to the homeostatic functions in the Th1-balanced environment induced by rAg85B administration. These findings provide a new insight into the regulatory effects of various innate immune factors induced by the mycobacteria major secretion protein rAg85B in allergic inflammation.

## Supporting Information

**Figure S1 SDS-PAGE separation and silver staining of rAg85B.** The recombinant purified Ag85B was solubilized in sample buffer to the desired concentration, and boiled for 5 min. 15 μl/well from each samples were separated on 10% SDS gel using mini-PROTEAN electrophoresis instrument (Bio-Rad Laboratories). Silver staining of the gel was performed according to the standard protocol of EzStain Silver (ATTO). Various concentration of rAg85B on the gel (12.5, 25, 50, 100, 200, and 400 ng).
(TIF)

**Figure S2 CD4$^+$ Foxp3$^+$ T cells were almost expressing CD25.** OVA-immunized (i.p., day0 and 14) and sensitized (5% aerosolized-OVA, day21 to 25) BALB/c mice were challenged with PBS or rAg85B protein (i.p. (100 μg; days 0 and 14) and i.n. (20 μg; days 21, 23, and 25)). At 24 h after the last OVA sensitization, mediastinal lymph nodes (MLNs) from naïve or OVA sensitized BALB/c mice treated with PBS or rAg85B, were harvested. MLNs cells from naïve (upper), PBS-treated (middle) and rAg85B protein-treated (lower) OVA-sensitized mice were

stained with anti-CD4 and anti-Foxp3. Cells in R1-R4 were analyzed for the expression of CD25.
(TIF)

**Figure S3 The composition of BAL cells in rAg85B administered mice.** OVA-immunized (i.p., day0 and 14) and sensitized (5% aerosolized-OVA, day21 to 25) BALB/c mice were challenged with PBS (control) or rAg85B (i.p. (100 μg; days 0 and 14) and i.n. (20 μg; days 21, 23, and 25)). One day after the last challenge, BAL cells from OVA sensitized BALB/c mice treated with PBS or rAg85B were counted. Different BAL cells populations were measured by surface staining. Flow cytometry of BAL cells stained with anti-CD3, anti-CD4, anti-CD8, anti-Gr-1, anti-γδ TCR, anti-NKp46, anti-CD11c, anti-CD127 (IL-7R) and Lineage specific marker (CD3, CD19, Gr-1, CD11b, CD11c).
(TIF)

## Author Contributions

Conceived and designed the experiments: YY. Performed the experiments: YT HI MY. Analyzed the data: YY KM. Contributed reagents/materials/analysis tools: TF. Contributed to the writing of the manuscript: YT YY.

## References


1. Adams JF, Scholvinck EH, Gie RP, Potter PC, Beyers N, et al. (1999) Decline in total serum IgE after treatment for tuberculosis. Lancet 353: 2030–2033.
2. Herz U, Gerhold K, Gruber C, Braun A, Wahn U, et al. (1998) BCG infection suppresses allergic sensitization and development of increased airway reactivity in an animal model. J Allergy Clin Immunol 102: 867–874.
3. Cavallo GP, Elia M, Giordano D, Baldi C, Cammarota R (2002) Decrease of specific and total IgE levels in allergic patients after BCG vaccination: preliminary report. Arch Otolaryngol Head Neck Surg 128: 1058–1060.
4. Choi IS, Koh YI (2002) Therapeutic effects of BCG vaccination in adult asthmatic patients: a randomized, controlled trial. Ann Allergy Asthma Immunol 88: 584–591.
5. Stassen M, Jonuleit H, Muller C, Klein M, Richter C, et al. (2004) Differential regulatory capacity of CD25+ T regulatory cells and preactivated CD25+ T regulatory cells on development, functional activation, and proliferation of Th2 cells. J Immunol 173: 267–274.
6. Robinson DS, Larche M, Durham SR (2004) Tregs and allergic disease. J Clin Invest 114: 1389–1397.
7. Cui J, Watanabe N, Kawano T, Yamashita M, Kamata T, et al. (1999) Inhibition of T helper cell type 2 cell differentiation and immunoglobulin E response by ligand-activated Valpha14 natural killer T cells. J Exp Med 190: 783–792.
8. Harada M, Magara-Koyanagi K, Watarai H, Nagata Y, Ishii Y, et al. (2006) IL-21-induced Bepsilon cell apoptosis mediated by natural killer T cells suppresses IgE responses. J Exp Med 203: 2929–2937.
9. Torrado E, Cooper AM (2010) IL-17 and Th17 cells in tuberculosis. Cytokine Growth Factor Rev 21: 455–462.
10. Griffiths KL, Pathan AA, Minassian AM, Sander CR, Beveridge NE, et al. (2011) Th1/Th17 cell induction and corresponding reduction in ATP consumption following vaccination with the novel Mycobacterium tuberculosis vaccine MVA85A. PLoS One 6: e23463.
11. Lambrecht BN, Hammad H (2012) The airway epithelium in asthma. Nat Med 18: 684–692.
12. Holgate ST (2012) Innate and adaptive immune responses in asthma. Nat Med 18: 673–683.
13. Sonnenberg GF, Fouser LA, Artis D (2011) Border patrol: regulation of immunity, inflammation and tissue homeostasis at barrier surfaces by IL-22. Nat Immunol 12: 383–390.
14. Takamura S, Matsuo K, Takebe Y, Yasutomi Y (2005) Ag85B of mycobacteria elicits effective CTL responses through activation of robust Th1 immunity as a novel adjuvant in DNA vaccine. J Immunol 175: 2541–2547.
15. Mori H, Yamanaka K, Matsuo K, Kurokawa I, Yasutomi Y, et al. (2009) Administration of Ag85B showed therapeutic effects to Th2-type cytokine-mediated acute phase atopic dermatitis by inducing regulatory T cells. Arch Dermatol Res 301: 151–157.
16. Karamatsu K, Matsuo K, Inada H, Tsujimura Y, Shiogama Y, et al. (2012) Single systemic administration of Ag85B of mycobacteria DNA inhibits allergic airway inflammation in a mouse model of asthma. J Asthma Allergy 5: 71–79.
17. Takatori H, Kanno Y, Watford WT, Tato CM, Weiss G, et al. (2009) Lymphoid tissue inducer-like cells are an innate source of IL-17 and IL-22. J Exp Med 206: 35–41.
18. Cella M, Fuchs A, Vermi W, Facchetti F, Otero K, et al. (2009) A human natural killer cell subset provides an innate source of IL-22 for mucosal immunity. Nature 457: 722–725.
19. Korn T, Bettelli E, Oukka M, Kuchroo VK (2009) IL-17 and Th17 Cells. Annu Rev Immunol 27: 485–517.
20. Schnyder B, Lima C, Schnyder-Candrian S (2010) Interleukin-22 is a negative regulator of the allergic response. Cytokine 50: 220–227.
21. Dieli F, Ivanyi J, Marsh P, Williams A, Naylor I, et al. (2003) Characterization of lung gamma delta T cells following intranasal infection with Mycobacterium bovis bacillus Calmette-Guerin. J Immunol 170: 463–469.
22. Sonnenberg GF, Nair MG, Kirn TJ, Zaph C, Fouser LA, et al. (2010) Pathological versus protective functions of IL-22 in airway inflammation are regulated by IL-17A. J Exp Med 207: 1293–1305.
23. Sugimoto K, Ogawa A, Mizoguchi E, Shimomura Y, Andoh A, et al. (2008) IL-22 ameliorates intestinal inflammation in a mouse model of ulcerative colitis. J Clin Invest 118: 534–544.
24. Planus E, Galiacy S, Matthay M, Laurent V, Gavrilovic J, et al. (1999) Role of collagenase in mediating in vitro alveolar epithelial wound repair. J Cell Sci 112 (Pt 2): 243–252.
25. Monticelli LA, Sonnenberg GF, Abt MC, Alenghat T, Ziegler CG, et al. (2011) Innate lymphoid cells promote lung-tissue homeostasis after infection with influenza virus. Nat Immunol 12: 1045–1054.
26. De Togni P, Goellner J, Ruddle NH, Streeter PR, Fick A, et al. (1994) Abnormal development of peripheral lymphoid organs in mice deficient in lymphotoxin. Science 264: 703–707.
27. Graf R, Schiesser M, Reding T, Appenzeller P, Sun LK, et al. (2006) Exocrine meets endocrine: pancreatic stone protein and regenerating protein–two sides of the same coin. J Surg Res 133: 113–120.
28. Zuany-Amorim C, Sawicka E, Manlius C, Le Moine A, Brunet LR, et al. (2002) Suppression of airway eosinophilia by killed Mycobacterium vaccae-induced allergen-specific regulatory T-cells. Nat Med 8: 625–629.
29. Taniguchi M, Harada M, Kojo S, Nakayama T, Wakao H (2003) The regulatory role of Valpha14 NKT cells in innate and acquired immune response. Annu Rev Immunol 21: 483–513.
30. Dieli F, Troye-Blomberg M, Ivanyi J, Fournie JJ, Bonneville M, et al. (2000) Vgamma9/Vdelta2 T lymphocytes reduce the viability of intracellular Mycobacterium tuberculosis. Eur J Immunol 30: 1512–1519.
31. Lockhart E, Green AM, Flynn JL (2006) IL-17 production is dominated by gammadelta T cells rather than CD4 T cells during Mycobacterium tuberculosis infection. J Immunol 177: 4662–4669.
32. Li Z, Burns AR, Miller SB, Smith CW (2011) CCL20, gammadelta T cells, and IL-22 in corneal epithelial healing. FASEB J 25: 2659–2668.
33. Guilloteau K, Paris I, Pedretti N, Boniface K, Juchaux F, et al. (2010) Skin Inflammation Induced by the Synergistic Action of IL-17A, IL-22, Oncostatin M, IL-1{alpha}, and TNF-{alpha} Recapitulates Some Features of Psoriasis. J Immunol.
34. Okamoto Yoshida Y, Umemura M, Yahagi A, O'Brien RL, Ikuta K, et al. (2010) Essential role of IL-17A in the formation of a mycobacterial infection-induced granuloma in the lung. J Immunol 184: 4414–4422.
35. Adams DO (1976) The granulomatous inflammatory response. A review. Am J Pathol 84: 164–192.
36. Sandor M, Weinstock JV, Wynn TA (2003) Granulomas in schistosome and mycobacterial infections: a model of local immune responses. Trends Immunol 24: 44–52.
37. Li Z, Burns AR, Han L, Rumbaut RE, Smith CW (2011) IL-17 and VEGF are necessary for efficient corneal nerve regeneration. Am J Pathol 178: 1106–1116.
38. Grenier A, Chollet-Martin S, Crestani B, Delarche C, El Benna J, et al. (2002) Presence of a mobilizable intracellular pool of hepatocyte growth factor in human polymorphonuclear neutrophils. Blood 99: 2997–3004.
39. Li Z, Rumbaut RE, Burns AR, Smith CW (2006) Platelet response to corneal abrasion is necessary for acute inflammation and efficient re-epithelialization. Invest Ophthalmol Vis Sci 47: 4794–4802.
40. Stirling DP, Liu S, Kubes P, Yong VW (2009) Depletion of Ly6G/Gr-1 leukocytes after spinal cord injury in mice alters wound healing and worsens neurological outcome. J Neurosci 29: 753–764.

BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 22 2014, pages 3279–3280
doi:10.1093/bioinformatics/btu510

*Structural bioinformatics*

Advance Access publication July 26, 2014

# Kotai Antibody Builder: automated high-resolution structural modeling of antibodies

Kazuo Yamashita[1], Kazuyoshi Ikeda[2], Karlou Amada[1], Shide Liang[1], Yuko Tsuchiya[2], Haruki Nakamura[3], Hiroki Shirai[4] and Daron M. Standley[1,*]

[1]WPI Immunology Frontier Research Center (IFReC), Osaka University, Suita, Osaka 565-0871, [2]National Institute of Biomedical Innovation, Ibaraki City, Osaka 567-0085, [3]Institute for Protein Research, Osaka University, Suita, Osaka 565-0871 and [4]Molecular Medicine Research Laboratories, Drug Discovery Research, Astellas Pharma Inc., Tsukuba, Ibaraki 305-8585, Japan

Associate Editor: Alfonso Valencia

**ABSTRACT**

**Motivation:** Kotai Antibody Builder is a Web service for tertiary structural modeling of antibody variable regions. It consists of three main steps: hybrid template selection by sequence alignment and canonical rules, 3D rendering of alignments and CDR-H3 loop modeling. For the last step, in addition to rule-based heuristics used to build the initial model, a refinement option is available that uses fragment assembly followed by knowledge-based scoring. Using targets from the Second Antibody Modeling Assessment, we demonstrate that Kotai Antibody Builder generates models with an overall accuracy equal to that of the best-performing semi-automated predictors using expert knowledge.

**Availability and implementation:** Kotai Antibody Builder is available at http://kotaiab.org

**Contact:** standley@ifrec.osaka-u.ac.jp

Received and revised on May 24, 2014; accepted on July 18, 2014

## 1 INTRODUCTION

Antibody variable regions constitute a unique protein module that has evolved to recognize virtually any biomolecular structure with high specificity and affinity. These properties have enabled the design of antibodies for use in the diagnosis and treatment of cancers and autoimmune and infectious diseases (Kuroda *et al.*, 2012). In addition to their clinical value, antibodies are extremely important for routine assays used in basic research. Computational modeling of antibody structure is a crucial step in engineering new antibody molecules, but there are few tools available to the general public, and accurately modeling loops in complementary determining regions (CDRs) remains an open problem. The PIGS server (Marcatili *et al.*, 2008) was validated in the first blind Antibody Modeling Assessment (Almagro *et al.*, 2011). However, prediction of the third heavy chain CDR (CDR-H3) remains difficult because of its structural diversity.

Recently, the Second Antibody Modeling Assessment (AMA-II) was held. AMA-II was divided into two stages: in stage 1, sequences were provided, and teams were assessed on the overall accuracy of their models. In stage 2, the crystal structures of the variable region lacking only CDR-H3 were provided, and groups were assessed on the accuracy of CDR-H3 loop prediction. In stage 1, the joint Osaka University Astellas (JOA) team achieved the lowest average root-mean square deviation (RMSD) for CDR-H3 (2.3 Å) and generated the most accurate models for 4 of 11 targets. In stage 2, the JOA team generated the most accurate models (with RMSDs of 1 Å or less) for 4 of 10 targets (Almagro *et al.*, 2014). However, the method used by the JOA team required much manual intervention and expert knowledge. Kotai Antibody Builder represents a fully automated but simplified implementation of the pipeline used by the JOA team (Shirai *et al.*, 2014).

*To whom correspondence should be addressed.

## 2 METHODS

Kotai Antibody Builder is composed of two main modules: MANGO and Spanner (Lis *et al.*, 2011). The MANGO module selects template structures for the framework (i.e. non-CDR) and each CDR by a sequence-based database search and rule-based heuristics, while Spanner builds loops by fragment assembly. Because CDR-H3 loops are well known to be more difficult to model than those of other CDRs, Kotai Antibody Builder provides a refinement option that includes sampling by fragment assembly followed by side-chain modeling and scoring by an empirical scoring function.

### 2.1 Framework selection

The local structure of residues 7–10 in the heavy (H) chain (here denoted 'framework motif') is diverse and can be classified into five types. In the first step of Kotai Antibody Builder, the framework motif is predicted by a statistics-based classification. Next, sequence alignment is used to find framework templates for H and light (L) chains separately. Only templates that have the same framework motif with that predicted for the query are used. Here, the CDR regions are masked so that only the framework region is aligned and scored. The H and L results are merged and sorted by sequence identity (seqID). The Molprobity software (Chen *et al.*, 2010) is used to assess if the percentage of backbone Ramachandran conformations inside the favored region is above a threshold (>85%). If there is no such model found, the selection criteria are relaxed and template models are selected by seqID, regardless of the predicted framework motif.

### 2.2 Non-H3 CDR selection

It is well known that each non-H3 CDR can be classified into one of several canonical clusters. We use the most recent definition of CDR clusters (North *et al.*, 2011), along with position-specific scoring matrix

(PSSM)-based scoring, to predict the best non-H3 CDR cluster for the query (Shirai *et al.*, 2014). For a given CDR, if the framework template and query are predicted to belong to the same cluster, the loop in the framework template is used; otherwise, the template with the highest seqID in the predicted cluster is used.

### 2.3 CDR-H3 selection

Because CDR-H3 is diverse in terms of length, sequence and structure, canonical rules have not been identified. Earlier, we developed H3 rules that partly classify CDR-H3 structures based on amino acid sequence (Kuroda *et al.*, 2008; Shirai *et al.*, 1996). Kotai Antibody Builder uses the most important of these rules (rule i), which predicts the structural class of the 'base' proximal to CDR-H3. For construction of the initial model, we use rule i if the length of loop is longer than five residues. In the optional refinement step, Spanner is used to generate 20 loop models followed by successive energy minimizations by the cosgene module in the myPresto package (Fukunishi *et al.*, 2003) and OSCAR-leap (Liang *et al.*, 2014). Here, we used a customized Spanner fragment database including only antibody structures for fast and specific CDR loop modeling. Finally, the single loop model was selected based on the OSCAR-leap score.

### 2.4 Model building

Selected loop models are grafted onto the framework template. If there are insertions/deletions in the template, Spanner is used to fix them. Side-chain modeling by OSCAR-star (Liang *et al.*, 2011) followed by short energy minimization with positional restraints on backbone atoms is also carried out (Fukunishi *et al.*, 2003). The initial calculation takes 5–10 min, whereas refinement requires an additional ~90 min.

### 2.5 Web server

Kotai Antibody Builder accepts amino acid sequences for H and L Fv regions. The resulting 3D model can be downloaded in PDB format and visualized by the JSmol viewer. The PDB IDs of templates as well as the canonical classifications of each loop are shown.

## 3 RESULTS

To assess the accuracy of Kotai Antibody Builder, we used targets 2–11 from AMA-II. In Figure 1, we show the resulting Cα RMSDs of rank-1 models submitted by the JOA team alongside rank-1 initial and refined models from the Web server. The

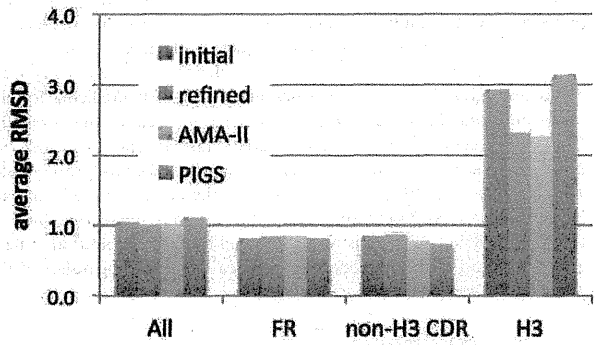


**Fig. 1.** Average Cα RMSDs of AMA-II targets 2–11. 'AMA-II' indicates rank-1 models submitted by the JOA team; 'initial' and 'refined' indicate the corresponding Web server options; 'All' and 'FR' indicate the entire Fv and framework regions for H and L chains combined

overall RMSD was ~1 Å. The refinement option was much more successful in modeling CDR-H3 loops than the protocol used to generate initial models or by the PIGS server (Marcatili *et al.*, 2008). Surprisingly, the CDR-H3 accuracy for the refined loops (2.3 Å) was equal to that of the stage-1 JOA submitted models, the generation of which required careful manual inspection. There was a slight increase (~0.1 Å) in the RMSD of the non-H3 CDR loops when the refinement option was used because of the fact that the other loops were not held rigid during the minimizations; however, we found that this slight flexibility in the non-H3 CDRs was necessary for proper modeling of the H3 loops.

## 4 CONCLUSIONS

There are few fully automated antibody modeling pipelines available to the general public, and none that we are aware of that can reach this level of accuracy for CDR-H3 loops. Thus, Kotai Antibody Builder is expected to contribute uniquely to the field of antibody structural modeling and design.

## ACKNOWLEDGEMENTS

The authors would like to thank N. Sakiyama, H. Nakagawa, E. Kanamori, K. Tsuchida, N. Tanigawa and S. Soga of Astellas Pharma for helpful discussions.

*Funding*: This work was supported by the Platform for Drug Discovery, Informatics and Structural Life Science, MEXT, Japanese Government.

*Conflict of interest*: none declared.

## REFERENCES

Almagro,J.C. *et al.* (2011) Antibody modeling assessment. *Proteins*, **79**, 3050–3066.
Almagro,J.C. *et al.* (2014) Second antibody modeling assessment (AMA-II). *Proteins*, **82**, 1553–1562.
Chen,V.B. *et al.* (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.*, **D66**, 12–21.
Fukunishi,Y. *et al.* (2003) The filling potential method: a method for estimating the free energy surface for protein-ligand docking. *J. Phys. Chem. B.*, **107**, 13201–13210.
Kuroda,D. *et al.* (2008) Structual classification of CDR-H3 revisited: a lesson in antibody modeling. *Proteins*, **73**, 608–620.
Kuroda,D. *et al.* (2012) Computer-aided antibody design. *Protein Eng. Des. Sel.*, **25**, 507–521.
Liang,S. *et al.* (2011) Fast and accurate prediction of protein side-chain conformations. *Bioinformatics*, **27**, 2913–2914.
Liang,S. *et al.* (2014) LEAP: highly accurate prediction of protein loop conformations by integrating coarse-grained sampling and optimized energy scores with all-atom refinement of backbone and side chains. *J. Comput. Chem.*, **35**, 335–341.
Lis,M. *et al.* (2011) Bridging the gap between single-template and fragment based protein structure modeling using Spanner. *Immunome Res.*, **7**, 1–8.
Marcatili,P. *et al.* (2008) PIGS: automatic prediction of antibody structures. *Bioinformatics*, **24**, 1953–1954.
North,B. *et al.* (2011) A new clustering of antibody CDR loop conformations. *J. Mol. Biol.*, **406**, 228–256.
Shirai,H. *et al.* (1996) Structural classification of CDR-H3 in antibodies. *FEBS Lett.*, **399**, 1–8.
Shirai,H. *et al.* (2014) High-resolution modeling of antibody structures by a combination of bioinformatics, expert knowledge, and molecular simulations. *Proteins*, **82**, 1624–1635.

*10086–10098* *Nucleic Acids Research, 2014, Vol. 42, No. 15* *Published online 25 July 2014*
*doi: 10.1093/nar/gku681*

# Quantifying sequence and structural features of protein–RNA interactions

**Songling Li*, Kazuo Yamashita, Karlou Mar Amada and Daron M. Standley***

Laboratory of Systems Immunology, WPI Immunology Frontier Research Center, Osaka University, Osaka 565–0871, Japan

Received March 28, 2014; Revised July 11, 2014; Accepted July 14, 2014

## ABSTRACT

**Increasing awareness of the importance of protein–RNA interactions has motivated many approaches to predict residue-level RNA binding sites in proteins based on sequence or structural characteristics. Sequence-based predictors are usually high in sensitivity but low in specificity; conversely structure-based predictors tend to have high specificity, but lower sensitivity. Here we quantified the contribution of both sequence- and structure-based features as indicators of RNA-binding propensity using a machine-learning approach. In order to capture structural information for proteins without a known structure, we used homology modeling to extract the relevant structural features. Several novel and modified features enhanced the accuracy of residue-level RNA-binding propensity beyond what has been reported previously, including by meta-prediction servers. These features include: hidden Markov model-based evolutionary conservation, surface deformations based on the Laplacian norm formalism, and relative solvent accessibility partitioned into backbone and side chain contributions. We constructed a web server called aaRNA that implements the proposed method and demonstrate its use in identifying putative RNA binding sites.**

## INTRODUCTION

Many biological processes require specific interactions between protein and RNA molecules. Protein–RNA interactions coordinate the flow of genetic information from transcription to translation at various levels (1–3). Protein and RNA molecules can fold together to form stable subunits of molecular machines such as the ribosome (4) or spliceosome (5) and also form transient complexes, such as target-specific ribonucleases (6) and helicases (7). Like proteins, RNA molecules can adopt myriad structural conformations, a consequence of which is a great variety of protein–RNA interaction motifs. Understanding the underlying principles of these interactions is a non-trivial task since there are far fewer solved structures of protein–RNA complexes than there are known interactions and RNA structure determination poses a unique set of challenges (8). Nevertheless, the growth rate of structurally determined protein–nucleotide complexes has continued to rise over the last decade. Therefore, there is a need to establish methods that can reliably translate such structural data into predictive models.

Computational methods for the prediction of RNA binding sites on proteins make use of various features. A number of methods are based on sequence information, including: PiRaNhA (9), which uses position-specific scoring matrices (PSSMs), inherent binding propensities of interface residues, solvent accessibility and hydrophobicity; BindN+ (10), which uses side chain pKa, hydrophobicity, the molecular masses of residues and evolutionary information captured by PSSMs; PRBR (11), which uses predicted secondary structure, conservation of residue physicochemical properties and residue-dependent charge-polarity and hydrophobicity; SRCPred (12), which uses PSSMs and global amino acid composition (GAC) to predict di-nucleotide binding propensities. Methods that make use of structural information include: KYG (13), which combines residue-based binding propensities, spatially close residue doublets, and sequence profiles; DRNA (14), which performs alignment with known complex structures and scores targets with a statistical energy function and OPRA (15), which uses accessible-surface-weighed residue binding propensities calculated from known binding interfaces.

Sequence-based predictors are usually shortsighted, due to their fragmented view of a binding site; a sliding window can only capture a continuous segment of sequential residues, thereafter neglecting correlation between sequentially distant but spatially close residues. In contrast, structure-based predictors can reach higher specificity but usually at a cost of sensitivity (16). Structure-based methods generally attempt to recall geometric features from known protein–RNA complexes and fit these to geometric features of query proteins. Due to the large degree of freedom in-

*To whom correspondence should be addressed. Tel: +81 6 6879 4264; Fax: +81 6 6879 4272; Email: standley@ifrec.osaka-u.ac.jp
Correspondence may also be addressed to Songling Li. Tel: +81 6 6879 4924; Fax: +81 6 6879 4272; Email: slli@ifrec.osaka-u.ac.jp

troduced by protein folding from 1D sequence to 3D structure and the limited number of training structures, geometric features of RNA-binding proteins have not been exhaustively explored, resulting in lower sensitivity as compared with sequence-based methods. As a consequence of these tradeoffs, we aimed to develop a method that could optimally utilize both sequence and structural features of RNA-binding proteins in order to accurately quantify their contributions to protein–RNA molecular recognition.

To this end, we have made use of several established and novel features. In addition to the sequence features used previously in the SRCPred method (12), we included hidden Markov model (HMM)-based evolutionary conservation (EC) scores to better evaluate conservation. We adopted an algorithm that collects positional amino acid occurrence from reweighed alignments acquired through HMM-based comparisons (17). We found that the HMM-based EC feature provided a more straightforward measure of EC than the previously described PSSM-based feature (12). For structural features, we made use of local relative accessible surface area, which we developed in a novel way and mapped onto patches of spatially neighboring residues in order to capture information from spatially close residues. Finally, we represented molecular structure by using the Laplacian norm (LN) (18). LN is a structural descriptor that measures surface convexity/concavity over different length scales. By tuning the granularity, the LN could be made tolerant to structural deviations among RNA binding surfaces, while still being sensitive enough to distinguish binding surfaces from non-binding ones. Consequently, both sensitivity and specificity of the predictor could be achieved.

In summary, we present an RNA binding site predictor using various features that outperforms sequence- or structure-only predictors. Importantly, the proposed method makes use of structural features even for sequence-only input through in-line homology modeling and is robust with respect to typical input noise levels that occur in the homology modeling phase. The proposed method has been implemented as a web service called aaRNA at http://sysimm.ifrec.osaka-u.ac.jp/aarna/, and is expected to enhance functional annotation of putative RNA-binding proteins at the residue level.

## MATERIALS AND METHODS

### Dataset and contact profile

Protein–RNA complexes with a resolution better than 3.0 Å and solved by X-ray crystallography were downloaded from the Protein Data Bank (PDB) (19) in May 2013. Only protein chains with at least 30 resolved residues and no <3 residues in RNA contact were considered. We also required the RNA partner chain or chains to be at least 3 nucleotides long. Protein chains that include only carbon alpha (Cα) atoms were discarded. Redundancy among protein chains was reduced by clustering them using BLASTCLUST at a 25% sequence identity threshold. From each protein chain cluster, we selected a representative with the largest number of RNA contacts. A protein–RNA contact was defined structurally, and only contacts between protein and RNA chains within the same biological unit (BU) were considered. For a structure with a single BU, an amino acid and ribonucleotide residue pair was considered to be in contact when their minimum distance was <3.5 Å. For structures with multiple BUs, individual units were separated following '**REMARK 350**' records in their respective PDB files. Contacts taking place at the interface of two nearby BUs were ignored, because these interactions might not be functionally relevant. A di-nucleotide contact is defined as a nucleotide in contact together with one of its flanking nucleotides on either direction. We neglected structurally unresolved protein residues because neither contact nor structure could be defined. Protein chains were partitioned into a 'non-ribosomal' and a 'full' (ribosome-including) dataset; the former included no protein chains interacting with bulky ribosomal RNAs (longer than 200 bps), resulting in 141 and 205 protein chains, respectively. Two supplementary files listing all PDB protein and RNA chains for contact included in the dataset can be found at Supplementary Materials. In the non-ribosomal dataset, 2899 out of 43 863 residues were RNA binding, while in the full dataset, 5493 out of 51 781 residues were RNA binding. With the full dataset, we further checked hydrogen bonds in the structures using HBPLUS (20), and analyzed contact preferences between protein residues and RNA nucleotides. Note that in order to identify hydrogen bonds correctly, a profile indicating atoms of hydrogen donator and acceptor in nucleotides must be prepared. In particular, the O3' atom of ribose can only serve as an acceptor after forming a phosphodiester bond.

### Artificial neural network training and testing

Binary and di-nucleotide classifiers were built by using an artificial neural network algorithm implemented in the Stuttgart Neural Network Simulator suite (http://wolf.bms.umist.ac.uk/naccess). A three-layered fully connected network was constructed and trained via a standard back propagation protocol. The number of nodes in the first network layer equaled the number of training features (to be described below), and the last output layer contained either 1 or 16 output nodes (ranging from 0 to 1 in value) for binary and di-nucleotide classifiers, respectively. In the later case, the 16 output nodes are due to the 16 possible di-nucleotide combinations: [A|C|G|U] x [A|C|G|U]. The number of nodes in the middle hidden layer was tuned to optimize the performance, resulting in a five-node layer. The performance of the model was assessed by a five-fold cross-validation scheme. Instead of randomly sampling an equal number of 'positive' (i.e. RNA binding) and negative (i.e. non-binding) inputs, a more stringent method was used. A nearly equal number of protein chains (28 or 29 for the RB141 dataset and 41 for the RB205 dataset) were randomly allocated into five subsets, with each subset containing only residues from the assigned chains. Then, three out of five subsets were iteratively selected to train the network. To avoid over-fitting, training was halted when an early stopping criterion was satisfied based on evaluation by one of the subsets left out. Finally, the last remaining subset was tested by the network to estimate the performance. All protein chains were tested after shuffling train-

ing and testing datasets in this way. This approach to training and testing resulted in a far larger proportion of negative residues than if the datasets were artificially partitioned into an equal number of positive and negative sites. Tests returned numerical values of binding propensities from 0 to 1. Receiver operating characteristic (ROC) analysis was then carried out on these propensities by varying cutoff values, above which a prediction was considered as binding. For the di-nucleotide classifier, because the best performance was achieved under different cutoffs for different di-nucleotide types, output binding propensities were not directly comparable among different classes. Therefore binding propensity raw scores (prediction scores returned by the neural network) for different di-nucleotide types were transformed into precision values calculated by using corresponding prediction scores as the cutoff during ROC analysis. In order to test the stability of model, we repeated each training cycle five times by reinitializing the network with different random number seeds before training.

### Training features based on protein sequence

Three sequence features (21-bit coding, GAC and PSSM) were taken from our previous work (12) as a starting point. A sliding window of size $2N+1$, which corresponded to a center residue and its $N$ nearest sequence neighbors on either side, was used to scan protein sequences. Since the window was moved by 1 in each step, neighboring windows shared $2N$ residues. A 21-bit sparse coding method was used to encode window fragments into their amino acid composition in sequence order. Each residue was represented by a 21-bit long string. The first 20 bits were used to label specific amino acids types. For each of the 20 common amino acid types, only one bit position was set to 1, and the rest were set to 0. The last bit was set to 1 for vacant sequence positions or non-standard amino acids. Next, a 20-column GAC vector was used, which represented the abundance of each amino acid type in a protein sequence. Last, evolutionary profiles based on PSSMs were computed by the PSI-BLAST program with an E-value threshold of 1E-3 and three iterations against NCBI's NR database. Raw PSSM values $V$ were normalized by a logistic operator:

$$\frac{1}{1+e^{-1*V}}.$$

In addition, we evaluated protein EC with a method combining HMM–HMM comparison and position-specific amino acid frequency calculated with weights from multiple sequence alignments, as described in (17). Homologs were identified by searching a pre-clustered HMM database of UniProt protein sequences at a 20% similarity level (HHsuite Uniprot20 Database) by using the HHblits program (21) with two iterations, an E-value cutoff of 1E-3 and the -realign option turned on. The filtering options in HHblits pairwise sequence identities were turned off in order to include all possible homologous sequences in the database. After database searching, multiple-sequence-alignment (MSA) files in a3m format were transformed into a2m format using the HHblits reformat.pl utility with the '-M first' option, which turned all residues in the first sequence (the query sequence) into a match status. For a multiple sequence alignment matrix $\{A_{i=1\ldots L}^{m=1\ldots M}\}$ of size $M \times L$, where $M$ (rows) is the number of collected protein sequences and $L$ (columns) is the index of residue positions in the query sequence, each matrix element could be one of the 20 amino acid types or a gap type, marked as '-'. To alleviate biased sampling due to different numbers of similar HMMs deposited in the database, a normalized Hamming distance was used to assess similarities among homologous sequences. The weight of each collected homologous sequence in the amino acid frequency calculation was calibrated according to its sequence similarity to the rest of the homologs. Only amino acids in upper case (HMM match status) at alignment columns were taken into consideration. For any two aligned protein sequences $S_m$ and $S_n$ (where $m \neq n$, and $1 <= m, n <= M$), the Hamming distance (22) was defined as the number of positions for which the corresponding amino acids were different. The Hamming distance was normalized by dividing by the alignment length. If the normalized Hamming distance was smaller than a pre-defined distance threshold (0.3), which means the similarity was greater than 0.7, then their pairwise weight $W_{m,n}$ is set to 1; otherwise it was set to 0. When $m$ and $n$ were equal, $W_{m,n}$ was set to 1. For each sequence in the MSA profile, the corresponding weight $W_m$ was given by the inverse of the sum of all pairwise weights (including with self), as follows:

$$W_m = \frac{1.0}{1.0 + \sum_{n=1, n\neq m}^{M} W_{m,n}}, \text{ where} 1 \leq m \leq M. \quad (1)$$

The larger the number of close neighbors (above a similarity threshold of 0.7 by default) one could find for a sequence, the lower the weight of the amino acid occurrence in the aligned column. The contribution of an aligned sequence to the occurrence of a given type of amino acid in a given column of the multiple sequence alignment could then be calibrated by its weight as:

$$f_i(A) = \frac{1}{\lambda + M_{\text{eff}}} \left[ \frac{\lambda}{q} + \sum_{m=1}^{M} W_m \times \delta(A_i^m, A) \right], \quad (2)$$

which represents the frequency of an amino acid type $A$ in column $i$ of the alignment with reweighed sequences. In Equation (2), $i$ indicates a sequence position from 1 to $L$, $A$ indicates the residue type at position $i$ and $M_{\text{eff}} = \sum_{m=1}^{M} W_m$ is the effective number of sequences in the alignment after reweighting. The term $\delta(A_i^m, A)$ is the Kronecker delta function, which equals 1 when $A_i^m = A$; otherwise it returns 0. A pseudo-count term ($\lambda$) is used to regularize the data for the finite number of sequences, and was set to 0.5 by default. After reweighing, statistics on the occurrences of different amino acids at each alignment column of the query sequence were collected. Residue conservation values (weighed occurrence frequencies scaled from 0 to 1) in the query sequence were then normalized by the largest value of that sequence.

### Training features based on protein structure

Protein residues that form a continuous interface with RNA are not necessarily close in terms of the primary sequence. Therefore, we investigated features that encode spatial constraints among residues. For simplicity, we used carbon alpha (Cα) atoms as reference points to calculate Euclidean distances between residue pairs. Structural neighbors of a target residue were represented as a list in order of increased Cα/Cα atom distance from the target.

To be able to discriminate RNA-binding residues from non-binding residues, we used the LN on multiple scales to represent deformation of the protein structure. This method has previously been used to compare and classify proteins from different families (18), and we found that it worked well for characterizing RNA binding sites as well. Simply speaking, LN measures a weighed distance between the Cartesian coordinates of each residue and the coordinate centers of its neighboring residues (except the two to which the residue is covalently bonded). Importantly, LN coordinates are invariant to translation and rotation.

In order to compute the LN coordinates, we first set up an $n$x3 Cartesian coordinate matrix $P$ for a protein of $n$ residues as:

$$P_i = [x_i, y_i, z_i], \text{where } i = 1...n.$$

To compute the LN, a discrete Laplace operator was defined as in (18).

$$\Omega_{ij}^P(\sigma) = \begin{cases} -\frac{\|pi-pj\|^2}{\sigma^2} \\ e \\ 0 \end{cases} \begin{matrix} \\ \text{, if} |i-j| > 1, \\ \text{, otherwise} \end{matrix} \quad (3)$$

where $pk$ is the Cartesian coordinate of a residue in protein $P$, and the parameter σ in the Gaussian kernel controls the scale. Therefore, the weight between a residue pair is distance and scale dependent. Under a given scale σ, the weights between proximate residue pairs are higher and decrease rapidly as the distance increases. The scale factor σ determines the relative importance of near and distant residue neighbors. Sequential neighbors, which contribute a large, and roughly constant term, were omitted from Equation (3) in order to highlight the distribution pattern of sequentially distant residues. This equation corresponds to the weighed adjacency matrix of an undirected graph.

The Diagonal matrix $D^P(\sigma)$ is defined as follows:

$$D_{ii}^P(\sigma) = \sum_j \Omega_{ij}^P(\sigma). \quad (4)$$

And the discrete Laplace operator $L^P(\sigma)$ of the protein $P$ can be expressed as:

$$L^P(\sigma) = I - D^P(\sigma)^{-1}\Omega^P(\sigma). \quad (5)$$

In the $L^P(\sigma)$ matrix (5),

$$\text{diagonal elements } L_{i,i}^P(\sigma) = 1,$$
$$\text{subdiagonal elements } L_{i,i+1}^P(\sigma) \text{or } L_{i+1,i}^P(\sigma) = 0,$$
$$\text{and } L_{i,j}^P(\sigma) = -\frac{\Omega_{i,j}^P(\sigma)}{\sum_{|i-k|>1} \Omega_{i,k}^P(\sigma)}, \text{where} |i-j| > 1.$$

By multiplying the discrete Laplace operator $L^P(\sigma)$ with the protein coordinate matrix $P$, the coordinates of each target residue were subtracted by the weighed center (total weights being 1) of the coordinates of all other non-covalently bonded residues in the structure, which yielded the Laplacian coordinates $Lc^P(\sigma)$ of each residue as in Equation (6). This explains the translation invariance of the Laplacian coordinates.

$$Lc_i^P(\sigma) = L_i^P(\sigma) \times P = P_i - \frac{\sum_{j}^{|i-j|>1} \left[P_j \times \Omega_{i,j}^P(\sigma)\right]}{\sum_{j}^{|i-j|>1} \Omega_{i,j}^P(\sigma)}, \text{ where } i = 1...n. \quad (6)$$

After taking the Euclidean norm of the Laplacian coordinates for each residue, the distance between the target residue and its weighed center of neighboring residues was measured. This step makes the LN invariant to rotation.

LN values can reflect geometrical features of a target residue under different scales. By defining a scale factor σ, a pseudo-sphere centered at the target residue with a σ-related radius can be envisaged. The contribution of residues outside the sphere will be almost negligible in calculating the coordinate center of neighboring residues. For a buried residue surrounded by neighboring residues, it will result in a coordinate center of neighbors close to the target residue. The more symmetric the neighbors spatial distribution, the lower the LN value. Therefore, such buried and symmetrically organized residues will have LN values close to zero. For more exposed residues, especially when localized on the extreme periphery of the structure, the coordinate center of neighbors will deviate from the target residue, which results in a larger LN value. In contrast, when a target residue is on a concave surface, it will be partially surrounded by neighbors and consequently have a smaller LN value. An illustration of LN of residues on concave and convex surfaces at a global scale is shown in Supplementary Figure S1. Note that LN values of buried residues can fluctuate above zero, and concave residues can have LN values close to zero, depending on the spatial distribution of their neighboring residues, which are visible to the target residue at a given scale.

To compensate for residue position information lost after taking Euclidean norms of Laplacian coordinates, a range of σ values was used. By varying σ, the topology of a residue could be described on various scales. A small σ measures deformation of each residue locally, with only spatially close but not covalently bound neighbors being included; a large σ will describe residue deformations on a more global level. Finally, each residue was encoded into a multidimensional vector indexed by the scale. Distance distributions between all Cα atom pairs were determined first. Distances at 0.0, 0.25, 0.5, 0.75 and 1.0 quantile positions of the distribution were used to compute LN scale indices. For each protein, LN values were re-normalized by the largest value calculated under a given scale. A sliding window, incremented by 1, was used and the LN value of a given residue was encoded in a feature vector of length 11.

We next devised a novel way to calculate the normalized accessible surface area (ASA) of residues. Absolute ASA was computed by the NACCESS program (http://www.ra.cs.uni-tuebingen.de/SNNS). The ASA calculation was carried out for each protein chain isolated from other chains. Normalization of each residue was done by dividing the ASA values of the residue in the protein structure by the corresponding value of the isolated residue. Coordinates of each residue were extracted from the protein chain. Their ASAs were calculated in the absence of all other residues. After that, total ASA (all_atom_abs), ASA of the side chain atoms (total_side_abs) and ASA of the main chain atoms (main_chain_abs) were computed and were normalized by the all_atom_abs value of the single residue by itself. In this way, the ASA feature of each residue was represented by a row vector of length 3. Each column of an ASA feature vector (all_atom_abs, total_side_abs and main_chain_abs) was then re-normalized by the largest value in that column. A neighbor list of length 11 (including the target residue itself and its 10 nearest neighbors) was used to encode the ASA of each residue.

In addition, we checked the residue composition of RNA-binding surfaces in terms of their physicochemical properties. Here, again, target residue spatial neighbors were included. The R package 'seqinr' (23) was used to translate residue neighbor sequences into 10 physicochemical features; namely, 'tiny', 'small', 'aliphatic', 'aromatic', 'polar', 'non-polar', 'charged', 'acidic', 'basic', plus the isoelectric point of the residue. A 21-residue neighbor list (including the target residue itself) was used.

Lastly, DSSP (24) predicted secondary structure was used as a feature. An eight-bit binary feature vector was used to encode different types of secondary structures defined by DSSP, namely -, B, E, G, H, I, S, T. Again, a sliding window approach, incremented by 1, was used.

### Validation by homology models and independent dataset

Performance was measured by means of ROC curves, Area Under the ROC Curve (AUC), Precision-Recall (PR) curve, Specificity [+] (also known as Precision), Specificity [−], Sensitivity, F-measurement and Matthews Correlation Coefficient (MCC), based on the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Measures used were defined are as follows:

$$\text{Sensitivity(Recall)} = \frac{TP}{TP+FN},$$
$$\text{Specificity}[-] = \frac{TN}{TN+FP},$$
$$\text{Specificity}[+](\text{Precision}) = \frac{TP}{TP+FP},$$
$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}},$$
$$F-\text{score} = \frac{2\times \text{Precision}\times \text{Sensitivity}}{\text{Precision}+\text{Sensitivity}}.$$

To validate the robustness of our method, we tested the performance of our model using homology models in place of experimentally determined RNA-bound structures. The program Spanner (25) was used to render HHpred (26) alignments into structural models. Homologous templates with different sequence similarities (the top, <100%, <90%, <50% and <30% identity) were selected to build structures.

In addition, three sequence-representative standard benchmarks (RB106 adapted from (27), RB144 adapted from (28) and RB198 adapted from (29)) constructed by the authors of predictor RNABindR 2.0 (16) were used for comparison. We did training and testing on these three benchmarks by applying our feature-coding scheme and evaluated the residue-based and protein-based performance on structure data as described previously (16). Furthermore, the performance of aaRNA and BindN+ was compared by predicting RNA binding sites on pooled and redundancy-reduced RNABindR 2.0 datasets under a 3.5 Å distance cutoff. Two runs of redundancy removal were applied by using BLASTCLUST at a 30% sequence identity. Three benchmark sets (RB106, RB144 and RB198) were merged and then clustered. After that, representative sequences (from each cluster, one chain was randomly selected) were clustered again with the sequences for aaRNA/BindN+ training. The final test dataset was composed of 46 representative chains, which did not cluster with any of the training sequences.

In addition, we tested our model on an independent dataset from study (30) by prediction and compared our performance with methods reviewed in studies (16) and (30), which included a best-performing meta-predictor built from other sequence-based predictors (PiRaNhA (9), PPRInt (31) and BindN+ (10)), and three structure-based predictors (KYG (13), DRNA (14) and OPRA (15)). Moreover, we constructed an up-to-date test benchmark by collecting protein–RNA complexes that were solved by X-ray crystallography or nuclear magnetic resonance and released between June 2013 and June 2014. RNA-contacting protein sequences were defined as mentioned in Section "Materials and Methods, Dataset and contact profile", and clustered using BLASTCLUST at a 30% sequence identity level, which resulted in 154 clusters. Redundancy between representative sequences (with the largest number of RNA contacts in individual clusters) and sequences used for training was further reduced by retaining only representative chains sharing a maximum sequence identity below 30% when compared with training sequences. Finally, 67 protein sequences (RB67 benchmark) were selected and tested by different methods in the same way as the RB44 benchmark. A complete list of RB67 dataset is available at the Supplementary Material.

### Web server

The aaRNA web server can be found at http://sysimm.ifrec.osaka-u.ac.jp/aarna/.

## RESULTS

As described in Materials and Methods, we quantified the performance of each feature using ROC and PR curves, using two datasets. In the 'non-ribosomal' dataset, ribosomal proteins were excluded; in the 'full set' ribosomal proteins were included. As a control, we use a network trained using the sequence features previously used by the SRCPred method (12). In all cases, the parameter varied in the ROC and PR curves is the cutoff value in the neural network output above which RNA binding was predicted.

### Statistics of protein–RNA interactions

The contact preferences between amino acid and ribonucleotide residues were analyzed for the full dataset, and compared with the results of previous studies (32,33). Following the 2001 work by Jones *et al.* (32), for a given amino acid type, interface propensity was measured by comparing the fraction of ASA in contact with RNA with the fraction in contact with protein. The number of non-redundant complexes has increased dramatically since 2001, which has resulted in a significant change in the propensities of protein–RNA interactions (Supplementary Figure S2C). In particular, residues Arg, His, Lys, Trp and Tyr, when found on protein surfaces, have a higher probability to mediate RNA contacts than previously reported (32). A close look at these contacts in terms of their hydrogen bonds results in a generally similar pattern to that described in a 2011 study by Gupta *et al.* (33), which serves as a validation of our representative dataset. In brief, the largest number of hydrogen bonds mediating protein–RNA contacts takes place between protein side chains and RNA backbones (NS), as shown in Supplementary Figure S2A and Table S1. Guanine and uracil are higher than background levels for RNA side chain contacts (SS+SN), while cytosine is lower. In contrast, cytosine is lower and guanine is higher than background levels in RNA backbone contacts with protein (Supplementary Figure S2B). Moreover, interaction between protein side chains and RNA side chains (SS) favors charged (either positive or negative) or polar amino acids, whereas side chains of positively charged and aromatic residues interact more frequently with RNA backbones. Notably, the backbone of glycine mediates more contacts than that of other amino acids (Supplementary Figure S2D).

The results in Supplementary Figure S2 are complementary to earlier studies by Kondo and Westhof, who carried out a classification of base pairs and pseudo pairs observed in RNA-ligand complexes in terms of interaction edges (Watson–Crick, Hoogsteen or sugar-edge) of RNA bases and the glycosidic bond orientations relative to hydrogen bonds (*cis* or *trans*) (34). In their later work (35), the authors found that five kinds of amino acid residues (Asn, Gln, Asp, Glu and Arg) were able to form pseudo pairs with bases in a coplanar manner. When the interaction took place between a peptide backbone and nucleotide bases, the base adenine (A) was the most favorable. In addition, they found that the Watson–Crick side of bases formed the majority of pseudo pairs. The Hoogsteen edge of purine bases can bind to amino acid side chains both specifically (preferring G) or non-specifically (preferring A). The sugar edge of bases, however, interacts rarely with side-chain or backbone atoms (34,35).

### Contributions from EC

The EC feature is illustrated in Figure 1A, using the class-I *Archaeoglobus fulgidus* CCA-adding enzyme bound to a tRNA fragment as an example. We found that for non-ribosomal and full datasets, the EC feature could improve the AUC by ~1.3% and 0.8%, respectively, and also resulted in a better PR curve (Supplementary Figure S3) than the control method (sequence features used in the SRCPred method (12)). In order to quantify the information contained in each feature, we used EC and PSSM separately. We took the substitution frequency of each residue to itself in the PSSM profile, and normalized the frequencies via a logistic operator. We also included the 21-bit sparse coding feature and the GAC feature. The resulting AUCs were 0.7277 and 0.7075, respectively, for the EC- and PSSM-based model on the non-ribosomal dataset, and 0.8046 and 0.7942 on the complete dataset. These values verify that the EC feature contains additional information not found in the conservation values of the PSSM. We tested different E-value thresholds (1E-3, 1E-5 and 1E-10) for building MSA profiles, from which EC values were calculated. Using different E-values, a combination of E-values, or building a PSSM-like substitution matrix with occurrence frequencies for each of the 20 amino acid types did not result in an increase in performance. Therefore, the default E-value threshold was set to 1E-3. It should be noted that, depending on the number of homologous sequences in the database, the weight calculation step could be time-consuming. We were able to greatly speed this process up, however, by parallelization. After manually inspecting many known protein–RNA complexes, we could discern a rough correlation between residue conservation and distance to the bound RNA. As shown in Supplementary Figure S4A, the mean distance between protein surface residues and their bound RNAs was inversely related to the EC values. Moreover, RNA-binding residues were more enriched in large EC values than non-binding or back-

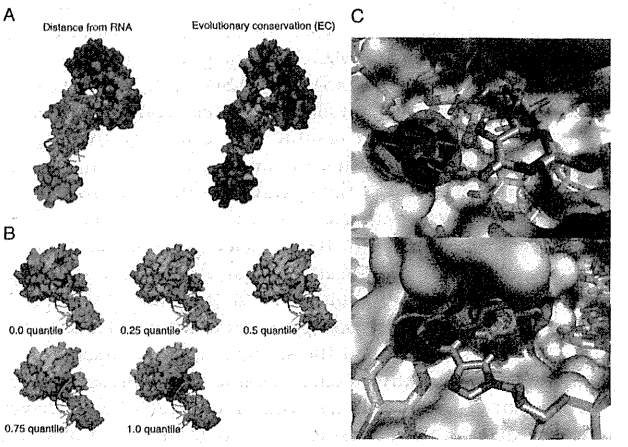


**Figure 1.** Novel features used in this work. (**A**) EC. A surface representation of the class-I *A. fulgidus* CCA-adding enzyme bound to a tRNA fragment (PDB ID: 3OVB). A distance map between protein and bound RNA with near (far) residues colored red (blue) is shown on the left. The EC value with high (low) colored red (blue) is shown on the right. (**B**) LN under a series of scales. LN values increase from blue to red. At each granularity level, warmer colors indicate convex residues, while cooler color represents concave residues. (**C**) Solvent ASA. A surface representation of RNase Cas6 (PDB ID: 4ILL) is shown. The protein makes both side-chain and backbone contacts with substrate RNA. Target residues (meshed) and nucleotides are represented by opaque sticks, connected by hydrogen bonds (dashed lines). The side chain of R268 protrudes and binds G15 (top). The backbone of Y168, which is mostly buried and forms part of a cleft, interacts with A5 (bottom). All figures of 3D structure representation in this work were generated by PyMOL Molecular Graphics System, Version 1.5, Schrödinger, LLC.