Cell

Supplemental Information

Molecular Mechanism of V(D)J Recombination from Synaptic RAG1-RAG2 Complex Structures

Heng Ru, Melissa G. Chambers, Tian-Min Fu, Alexander B. Tong, Maofu Liao, and Hao Wu

SUPPLEMENTAL INFORMATION

Molecular Mechanism of V(D)J Recombination from Synaptic RAG1-RAG2 Complex Structures

Heng Ru, Melissa G. Chambers, Tian-Min Fu, Alexander B. Tong, Maofu Liao,* and Hao Wu*

*Correspondence to wu@crystal.harvard.edu and maofu_liao@hms.harvard.edu

Inventory of Supplemental Information

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Cloning, Protein Expression and Purification

We first screened expression levels of mouse, zebrafish and xenopus RAG1 and RAG2 in baculovirus-mediated insect cell and 293T cell expression systems. Similar to previous observations (Lapkouski et al., 2015), mouse RAG1 (mRAG1) and mRAG2 expressed in insect cells were largely aggregated, while those expressed in 293T cells gave relatively low yields. Expression experiments on xenopus RAGs also did not result in high yields. In contrast, zebrafish RAG1 (zRAG1) and zRAG2 both gave higher expression levels when expressed in insect cells and showed little aggregation. In detail, the genes encoding zRAG1 (271-1031, from the RING domain to near the C-terminus) and zRAG2 full-length were subcloned into the modified vector pFastbac1 with an N-terminal 6xHis-MBP tag and a tobacco etch virus (TEV) protease cleavage site (ENLYFQ/G, where '/' indicates the cutting site). The plasmids containing the genes of interest were transformed into the DH10Bac competent cells and the recombinant bacmid DNAs were isolated and verified according to the instructions of the Bac-to-Bac® baculovirus expression system manual (Life Technologies). Sf9 insect cells were transfected to generate the recombinant baculovirus and the titer of the baculovirus was amplified by infecting Sf9 insect cells. RAG1 and RAG2 proteins with an N-terminal 6xHis-MBP tag were separately overexpressed by infecting Sf9 cells with the recombinant baculovirus. Human HMGB1 (constructs 1-166) was subcloned into vector pET26b with a C-terminal 6xHis tag and the protein was overexpressed in BL21 (DE3) RIPL cells by induction using 0.2 mM IPTG at 30 °C for 3 hours when the OD_{600} reached 0.8.

The Sf9 cells that expressed RAG1 and RAG2 proteins were harvested by centrifugation at 2,000 rpm for 20 min. The cell pellets were re-suspended in a lysis buffer containing 20 mM HEPES at pH 7.5, 500 mM NaCl, 10% glycerol, 0.5 mM tris(2-carboxyethyl)phosphine (TCEP) and a protease inhibitor cocktail, and homogenized by ultra-sonication. For RAG1, a final concentration of 5 mM $MgCl_2$ or 10 mM $CaCl_2$ was added to stabilize the protein. The cell lysate was clarified by ultracentrifugation at 42,000 rpm at 4 °C for 2 hours. The supernatant containing the target protein was incubated with amylose resin (New England Biolabs) that was pre-equilibrated with the lysis buffer for 1 hour at 4 °C. After incubation, the resin-supernatant mixture was poured into a column and the resin was washed with the lysis buffer. The proteins were eluted by the lysis buffer supplemented with 20 mM maltose, and further purified by ion exchange chromatography. The proteins were concentrated and loaded onto a Superdex 200 *10/300* GL column (GE Healthcare) that was pre-equilibrated with the buffer containing 20 mM HEPES at pH 7.5, 150 mM NaCl, 0.5 mM TCEP and divalent metal ions (Mg^{2+} or Ca^{2+}) when necessary. Peak fractions from the gel filtration were collected and the tagged proteins were concentrated and quantified by the A280 method.

To reconstitute the Apo-RAG complex, His-MBP tagged RAG1 and RAG2 were combined with a molar excess of RAG2 and H3K4Me3 peptide was added with the same molar ratio to RAG2 to activate and stabilize the complex as shown previously (Liu et al., 2007). The mixture was incubated at 4 °C for half an hour and TEV protease was added to remove the His-MBP tag from both RAG1 and RAG2. The cleaved mixture was concentrated and applied to a Superdex 200 gel filtration column, pre-equilibrated with the buffer containing 20 mM HEPES at pH 7.5, 150 mM NaCl, 0.5 mM TCEP and divalent metal ions (Mg^{2+} or Ca^{2+}) when necessary. The Apo-RAG complex could be separated from TEV protease, His-MBP tag, and the excess RAG2. The peak fractions were pooled and concentrated. The concentration was also measured by the A280 method.

The *E. coli* cells expressing human HMGB1 with a C-terminal 6xHis tag were re-suspended in the buffer containing 50 mM Na_2HPO_4 and 10 mM KH_2PO_4 at pH 7.4, 500 mM NaCl, and 2.7 mM KCl. The cells were lysed by ultra-sonication and cell debris was removed by centrifugation at 17,000 rpm at 4 °C for half an hour. The supernatant was applied to a column pre-packed

with Ni-NTA resin (Qiagen). The column was subsequently washed sequentially using lysis buffer supplemented with 20 mM and 50 mM imidazole, respectively. The protein was eluted using lysis buffer containing 300 mM imidazole. The eluted protein was further purified by ion exchange and gel filtration chromatography.

The DNAs we employed in this study were synthesized as oligos (Integrated DNA Technologies). Double stranded or nicked DNA was generated by mixing the corresponding oligos in buffer containing 20 mM HEPES at pH 7.5 and 150 mM NaCl, and annealed by a temperature gradient. To reconstitute the PC, Apo-RAG was combined with nicked 12-RSS and 23-RSS intermediates and HMGB1 in the presence of Ca^{2+}, with the approximate Apo-RAG: 12-RSS: 23-RSS: HMGB1 ratio of 2: 1: 1: 3. Similarly, the SEC was reconstituted with 12-RSS and 23-RSS products and HMGB1 using the same molar ratio mentioned above in the presence of Mg^{2+}. The mixtures were further purified on a Superdex 200 gel filtration column pre-equilibrated with buffer containing 20 mM HEPES at pH 7.5, 150 mM NaCl, 0.5 mM TCEP and divalent metal ions (Ca^{2+} or Mg^{2+}) when necessary, during which excess HMGB1 and DNA were removed. The peak fractions were pooled, concentrated, and quantified by SDS-PAGE using the Apo-RAG complex as a reference.

Urea PAGE

The proteins in various RAG/DNA samples were heated at 98 °C for 5 min to inactivate the RAG enzyme, followed by digestion using proteinase K at 60 °C for at least 1 hour. The samples were then mixed 1: 1 with denaturing nucleic acid loading buffer containing 8 M urea. They were heated at 98 °C for 10 min and separated on a 15 % (w/v) denaturing polyacrylamide gel (Acrylamide: Bis-Acrylamide Ratio is 19:1, Bio-Rad) in the 1 x *Tris-borate-EDTA* (TBE) electrophoresis buffer at 150 V for half an hour. The gel was washed in deionized water, stained by SYBR® Safe DNA dye (Life Technologies) and visualized by fluorescence emission at ~470 nm.

Electron Microscopy (EM) Data Acquisition

Negative-stain EM grids were prepared following an established protocol (Booth et al., 2011) with minor modifications. Specifically, 2.5 µl of purified RAG complex was applied to glow-discharged EM grids covered by a thin layer of continuous carbon film, and the grids were washed and stained with 2% (w/v) uranyl formate. Negatively stained EM grids were imaged on a Tecnai T12 electron microscope (FEI company) operated at 120 kV at a nominal magnification of 67,000 x using a 4k x 4k CCD camera (UltraScan 4000, Gatan), corresponding to a pixel size of 1.68 Å on the specimen.

For cryo-EM, 2.5 µl of purified RAG complexes at a concentration of ~0.4 mg/ml were applied to a glow-discharged Quantifoil holey carbon grid (1.2/1.3, 400 mesh). Grids were blotted for 2.5 s with ~85 % humidity and plunge-frozen in liquid ethane using a Cryoplunge 3 System (Gatan). Cryo-EM images were collected at liquid nitrogen temperature on a TF30 Polara electron microscope (FEI company) operated at 300 kV and equipped with a K2 Summit direct electron detector (Gatan). Images were recorded in super-resolution counting mode following an established protocol (Li et al., 2013). Specifically, images were acquired at a nominal magnification of 31,000 x, corresponding to a pixel size of 1.23 Å on the specimen and 0.615 Å for super-resolution images. The dose rate was set to be 8.2 counts (corresponding to 9.9 electrons) per physical pixel per second. The total exposure time of each image stack was 6 s, leading to a total accumulated dose of 41 electrons per $Å^2$ fractionated into 30 frames (200 ms per frame). All images were recorded in a defocus range between 1.2 and 2.4 µm, using a semi-automated acquisition program UCSFImage4 (by Xueming Li).

Image Processing

SAMUEL (Simplified Application Managing Utilities for EM Labs) is a suite of python scripts that

facilitate image processing and assemble computation methods by utilizing the function modules in various software packages including SPIDER (Frank et al., 1996), Python Image Library, and SciPy. SAM scripts were used in all processing steps before 3D classification and refinement. SamViewer, an EM image analysis program written with wxPython, was used for 2D image display, particle picking, and interactive particle screening. Negative-stain EM images were 2 x 2 binned to a pixel size of 3.36 Å for further processing. Dose fractionated super-resolution image stacks collected using K2 Summit direct electron detector were binned 2 x 2 to a pixel size of 1.23 Å, and then subjected to whole-frame motion correction (Li et al., 2013). A sum of all 30 frames of each image stack was used for further processing. Defocus values were determined using CTFFIND3 (Mindell and Grigorieff, 2003).

Particle picking was performed using a semi-automated procedure. Initially ~2,000 particles were interactively picked to generate ~10 2D class averages, from which ~4 distinct averages with high signal-to-noise ratio were selected as templates to calculate particle positions using an exhaustive template matching procedure ('samautopick.py'). The particles calculated from each image were screened interactively to remove bad particles that show apparently wrong features or no visible feature. The resulting particles were then grouped using a reference-free clustering procedure ('samtree2dv2.py'), and the bad particles in each group were removed interactively. These final selected particles were boxed out, normalized to have a mean of 0 and a standard deviation of 1 ('sampilboxparticle.py'), and then subjected to further 2D and 3D analysis. For 2D classification, particles were first corrected for contrast transfer function (CTF) by flipping the phase using 'ctfapply' (by Xueming Li), and subjected to 10 cycles of correspondence analysis, *k*-means classification and multi-reference alignment using SPIDER operations ('samclasscas.py'). 3D initial models were generated with 2D class averages by SPIDER 3D projection matching refinement ('samrefine.py'), starting from a density of isotropic Gaussian distribution.

3D classification and refinement were carried out in RELION (Scheres, 2012). The 3D initial model was low-pass filtered to a resolution of 60 Å and used as the starting model for 3D classification. The homogeneous classes were either subjected to another round of 3D classification, or iteratively refined to a higher resolution using the 'auto-refine' procedure in RELION. At the end of 'auto-refine', in order to remove the effects of accumulated radiation damage and severe initial motion (Li et al., 2013), 3D refinement was continued using the particles summed from frame 3-16 in each image stack, which improved the map quality and resolution by ~0.1 Å. All refinements follow the gold-standard procedure, in which two half data sets were refined independently. RELION 'post-processing' was used to estimate resolution based on the Fourier shell correlation (FSC) = 0.143 criterion, after correcting for the effects of a soft shape mask using high-resolution noise substitution (Chen et al., 2013). In maps that show clear densities for NBD and nonamer DNA, a loose oval mask was applied around this region to exclude the active site region, and half map FSC values were calculated for resolution estimation of the NBD/nonamer portion of the density. Local resolution variations were estimated from two half data maps using ResMap (Kucukelbir et al., 2014). Amplitudes of the final maps were corrected by applying a negative B-factor using RELION 'post-processing'.

Model Building and Refinement

As the structural models of the mRAG1-RAG2 complex (PDB ID: 4wwx) (Kim et al., 2015) and the NBD-DNA complex (PDB ID: 3gna) (Yin et al., 2009) were available, we generated homology models of zRAG1 and zRAG2 on the SWISS-MODEL server http://swissmodel.expasy.org (Biasini et al., 2014). The models were first fitted into B-factor sharpened 3.3 Å resolution symmetrized map reconstructed from mixed SEC and PC particles in UCSF Chimera (Pettersen et al., 2004) and manually adjusted in COOT (Emsley et al., 2010). Because the map was most similar to the SEC map, we included cleaved RSS and coding end mimic in the model building. The atomic model for the mixed map was first refined in real space

using Phenix.real_space_refine (Adams et al., 2010). The map was then placed into an artificial unit cell with P1 symmetry and converted to MTZ format using Phenix.map_to_structure_factors (Adams et al., 2010). The resulting reflection file was used to perform maximum likelihood phased refinement using PHENIX (Adams et al., 2010) with secondary structure restraints, and reference model or Ramachandran restraints iteratively. Similar procedures were used in model building and refinement of the symmetrized SEC and PC models. For refinement against the 3.3 Å resolution map, mRAG1 and mRAG2 crystal structures were used as reference models. Refinements of the 3.4 Å resolution SEC model and the 3.7 Å resolution PC model used the 3.3 Å structure as the reference model. For refinement of the 4.6 Å resolution non-symmetrized PC model, the 3.7 Å resolution PC model and the NBD-DNA complex crystal structure were used as reference models, without Ramachandran restraints. The unambiguous fitting of the DNA sequences was facilitated by two factors. First, the nonamer sequences need to be bound to the NBDs, which fixed the orientation of the bound DNAs in these structures. Second, because all 100 and 122 nucleotides in nicked 12-RSS and 23-RSS intermediates are counted for in the non-symmetrized PC, only one DNA registration can accommodate all the nucleotides. Third, the DNA registration we determined for PC was transfer to the SEC structure. However, due to the limits in resolutions, we could not distinguish purine bases from pyrimidine bases in the DNA structures.

The final structures were validated using MolProbity (Chen et al., 2010). While most side chain densities were clear, some acidic side chains were not as well defined in the cryo-EM maps, likely contributed by their radiation sensitivity (Allegretti et al., 2014; Bartesaghi et al., 2014). Difference maps between cryo-EM volumes and atomic models were calculated using PHENIX (Adams et al., 2010). All molecular representations were generated in PyMOL (http://www.pymol.org) (Delano, 2002) and UCSF Chimera (Pettersen et al., 2004). Sequence alignments were performed in ClustalW2 (Larkin et al., 2007) and displayed using the online server of Espript 3.0 (Robert and Gouet, 2014).

Table S1. Cryo-EM Structure Determination and Model Validation, Related to Figure 1

EM data collection	**SEC**	**PC**	
Voltage (kV)	300	300	
Electron dose ($e^-/Å^2$)	41	41	
Defocus average (range) (μm)	1.86 (1.3-2.3)	1.95 (1.4-2.4)	
Number of images	650	550	
Number of particles	108,544	46,440	
Models†	**Symmetrized SEC**	**Symmetrized PC**	**Non-Symmetrized PC**
Particle number for the final map	63,853	42,486	14,129
Resolution (Å)	3.4	3.7	4.6
Map sharpening B-factor ($Å^2$)	-110	-135	-156
Number of protein residues	1,802	1,802	1,946
Number of nucleotides	116	124	222
Number of atoms	16,666	16,847	19,997
Geometric parameters (r.m.s.d.)			
Bond length (Å)	0.005	0.005	0.009
Bond angle (°)	1.5	0.9	1.4
Ramachandran statistics			
Favored regions (%)	93.6	94.6	93.6
Allowed regions (%)	6.2	5.2	5.6
Disallowed regions (%)	0.2	0.2	0.9
Rotamer outliers (%)	9.2	5.5	2.5
MolProbity score	2.9	2.6	2.5
Clashscore	16.2	12.8	20.1

Table S2. Human RAG Disease Mutations on Residues Involved in RSS-Binding and Dimer Closure, Related to Figure 3

Mutations at the NBD are omitted as they have been analyzed previously (Yin et al., 2009). The data are taken from (Kim et al., 2015; Lee et al., 2014; Schatz and Swanson, 2011; Sobacchi et al., 2006). SCID: severe combined immunodeficiency; CID: combined immunodeficiency.

Human mutations	Mouse residues	Zebrafish residues	Domain location	Residue function	Related disease
RAG1					
R474S/H/C	R471	R490	DDBD	Heptamer binding	Leaky SCID
S480G	S477	S496	DDBD	Heptamer binding	Enteropathy
R507W	R504	R523	DDBD	Spacer DNA binding	CID with granuloma/autoimmunity
H612R	H609	H629	RNH	Heptamer binding and interface of closed conformation	Leaky SCID
N650	N647	N669	RNH	Heptamer binding	Atypical SCID
R737H	R734	R756	ID/ZnB	Coding end binding	Omenn syndrome
R776Q/W	R773	R795	ID/ZnB	Coding end binding	SCID
R841W/Q	R838	R860	ID/ZnB	Interface of closed conformation	CID with granuloma/autoimmunity
N855I	N852	N874	ID/ZnB	Heptamer recognition	SCID with maternofetal engraftment
R973H/C	R970	R992	RNH	Heptamer binding	Omenn syndrome
R975W/Q	R972	R994	RNH	Heptamer binding	CID with granuloma/autoimmunity
Q981R/P	Q978	Q1000	CTD	Heptamer recognition	SCID with γδ T cells
K992E/R	K989	K1011	CTD	Spacer binding	Omenn syndrome
RAG2					
R39G	R39	R39	WD40	Coding end binding	SCID

SUPPLEMENTAL REFERENCES

Adams, P.D., Afonine, P.V., Bunkoczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W.*, et al.* (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr *66*, 213-221.

Allegretti, M., Mills, D.J., McMullan, G., Kuhlbrandt, W., and Vonck, J. (2014). Atomic model of the F420-reducing [NiFe] hydrogenase by electron cryo-microscopy using a direct electron detector. eLife *3*, e01963.

Bartesaghi, A., Matthies, D., Banerjee, S., Merk, A., and Subramaniam, S. (2014). Structure of beta-galactosidase at 3.2-A resolution obtained by cryo-electron microscopy. Proc Natl Acad Sci U S A *111*, 11709-11714.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T.G., Bertoni, M., Bordoli, L.*, et al.* (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res *42*, W252-258.

Booth, D.S., Avila-Sakar, A., and Cheng, Y. (2011). Visualizing proteins and macromolecular complexes by negative stain EM: from grid preparation to image acquisition. Journal of visualized experiments : JoVE.

Chen, S., McMullan, G., Faruqi, A.R., Murshudov, G.N., Short, J.M., Scheres, S.H., and Henderson, R. (2013). High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. Ultramicroscopy *135*, 24-35.

Chen, V.B., Arendall, W.B., 3rd, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr *66*, 12-21.

Delano, W.L. (2002). The PyMol Molecular Graphics System.

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta Crystallogr D Biol Crystallogr *66*, 486-501.

Frank, J., Radermacher, M., Penczek, P., Zhu, J., Li, Y., Ladjadj, M., and Leith, A. (1996). SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. J Struct Biol *116*, 190-199.

Kim, M.S., Lapkouski, M., Yang, W., and Gellert, M. (2015). Crystal structure of the V(D)J recombinase RAG1-RAG2. Nature *518*, 507-511.

Kucukelbir, A., Sigworth, F.J., and Tagare, H.D. (2014). Quantifying the local resolution of cryo-EM density maps. Nat Methods *11*, 63-65.

Lapkouski, M., Chuenchor, W., Kim, M.S., Gellert, M., and Yang, W. (2015). Assembly Pathway and Characterization of the RAG1/2-DNA Paired and Signal-end Complexes. J Biol Chem *290*, 14618-14625.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R.*, et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.

Lee, Y.N., Frugoni, F., Dobbs, K., Walter, J.E., Giliani, S., Gennery, A.R., Al-Herz, W., Haddad, E., LeDeist, F., Bleesing, J.H.*, et al.* (2014). A systematic analysis of recombination activity and genotype-phenotype correlation in human recombination-activating gene 1 deficiency. J Allergy Clin Immunol *133*, 1099-1108.

Li, X., Mooney, P., Zheng, S., Booth, C.R., Braunfeld, M.B., Gubbens, S., Agard, D.A., and Cheng, Y. (2013). Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. Nat Methods *10*, 584-590.

Liu, Y., Subrahmanyam, R., Chakraborty, T., Sen, R., and Desiderio, S. (2007). A plant homeodomain in RAG-2 that binds Hypermethylated lysine 4 of histone H3 is necessary for efficient antigen-receptor-gene rearrangement. Immunity *27*, 561-571.

Mindell, J.A., and Grigorieff, N. (2003). Accurate determination of local defocus and specimen tilt in electron microscopy. J Struct Biol *142*, 334-347.
Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem *25*, 1605-1612.
Robert, X., and Gouet, P. (2014). Deciphering key features in protein structures with the new ENDscript server. Nucleic Acids Res *42*, W320-324.
Schatz, D.G., and Swanson, P.C. (2011). V(D)J recombination: mechanisms of initiation. Annual review of genetics *45*, 167-202.
Scheres, S.H. (2012). RELION: implementation of a Bayesian approach to cryo-EM structure determination. J Struct Biol *180*, 519-530.
Sobacchi, C., Marrella, V., Rucci, F., Vezzoni, P., and Villa, A. (2006). RAG-dependent primary immunodeficiencies. Hum Mutat *27*, 1174-1184.
Yin, F.F., Bailey, S., Innis, C.A., Ciubotaru, M., Kamtekar, S., Steitz, T.A., and Schatz, D.G. (2009). Structure of the RAG1 nonamer binding domain with DNA reveals a dimer that mediates DNA synapsis. Nat Struct Mol Biol *16*, 499-508.