SUPPLEMENTARY MATERIALS

Structural basis for human PRDM9 actions at recombination hotspots

Anamika Patel[1], John R. Horton[1], Geoffrey G Wilson[2], Xing Zhang[1], Xiaodong Cheng[1]

[1] Department of Biochemistry, Emory University School of Medicine, Atlanta, GA 30322, USA

[2] New England Biolabs, Ipswich, Massachusetts 01938, USA

Corresponding author: xcheng@emory.edu

Email addresses for all authors:

AP (anamika.patel@emory.edu)

JRH (jrhorto@emory.edu)

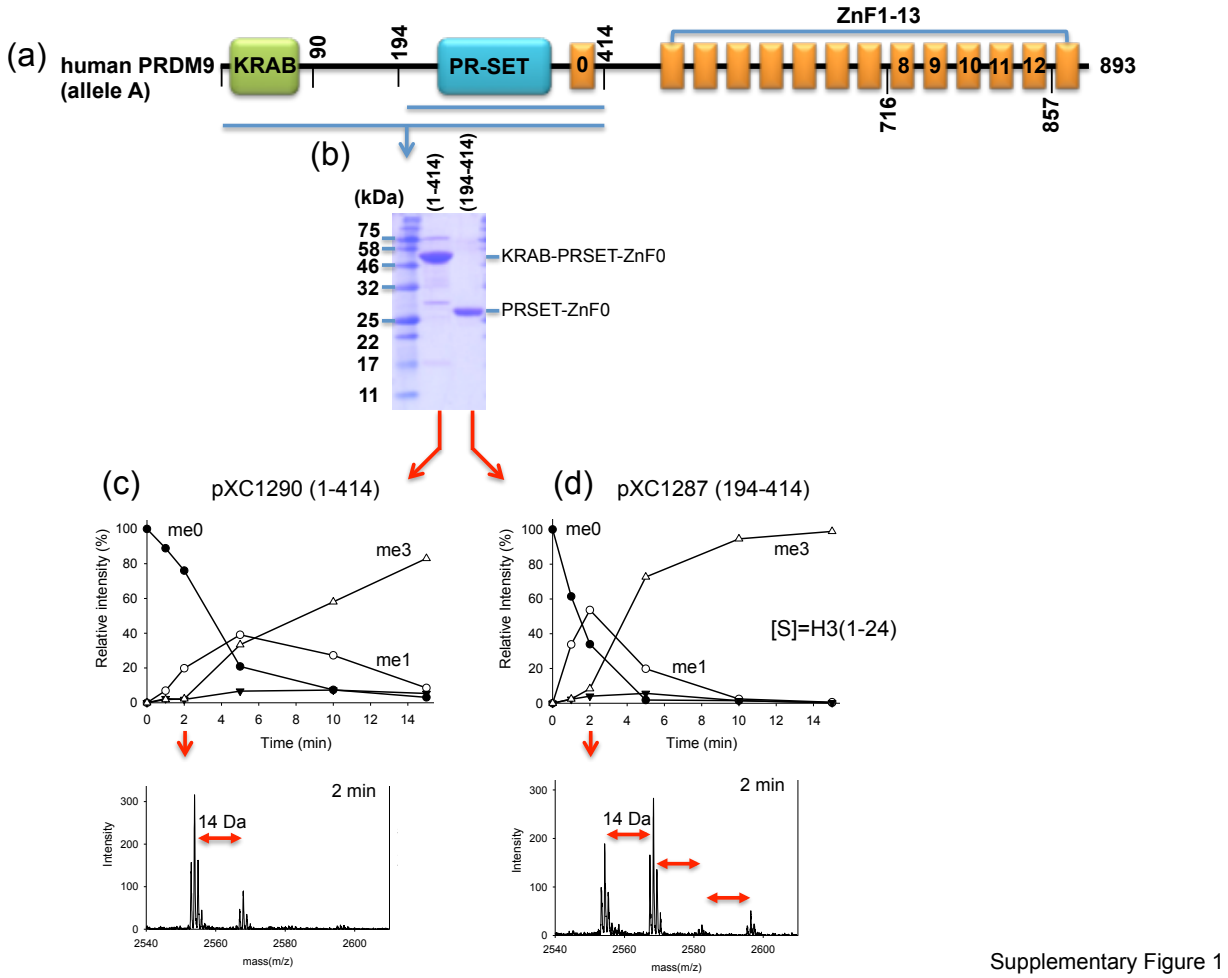
GGW (wilson@neb.com)

XZ (xzhan02@emory.edu)

XC (xcheng@emory.edu)

Supplemental Figures S1-S8

Supplemental Table S1

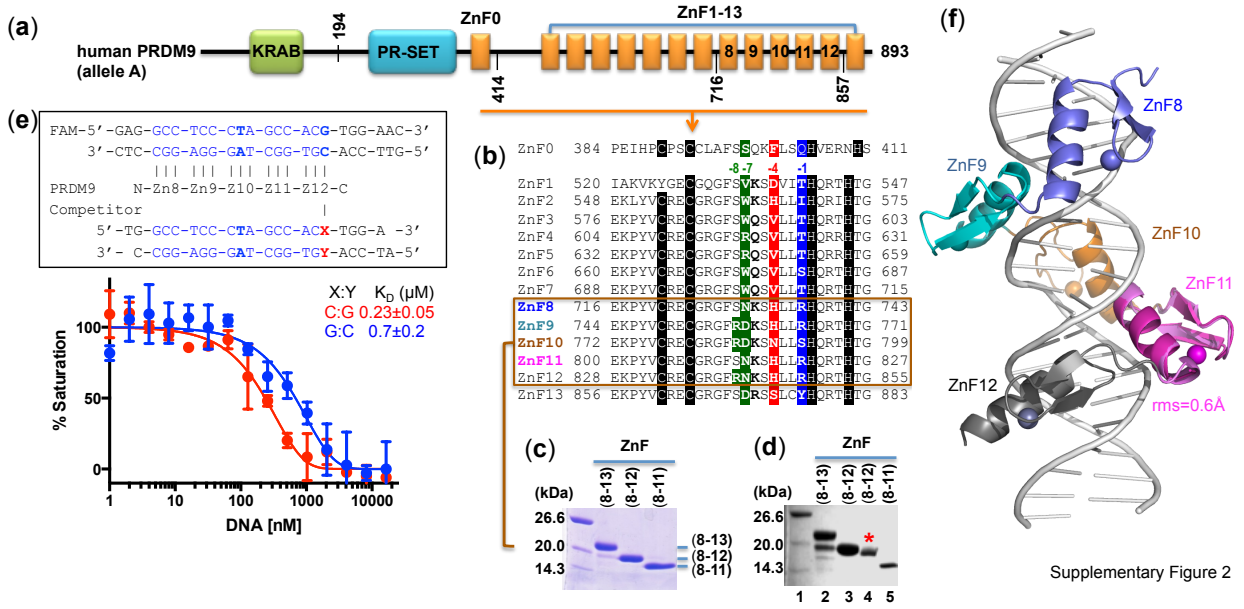


Supplementary Figure 1

Supplemental Figure S1. PRDM9 N-terminal domain harbors histone H3 lysine 4 methyltransferase activity

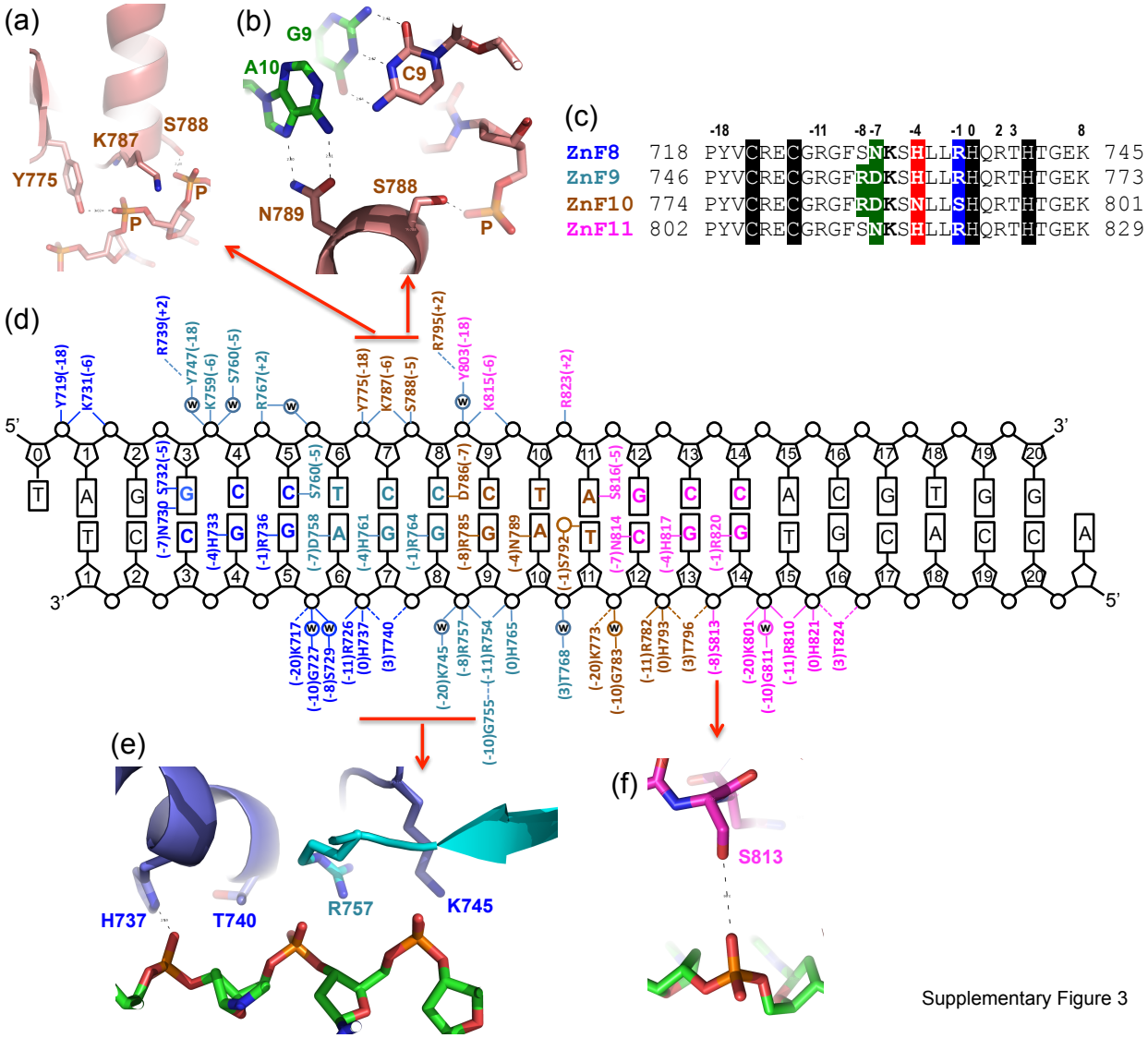
(**a**) Human PRDM9 contains an N-terminal putative KRAB domain, followed by the PR-SET domain. (**b**) SDS-PAGE gel showed the purified recombinant proteins (residues 1-414 and residues 194-414) used in this study. The two N-terminal fragments of human $PRDM9_A$ encompassing residues 1-414 (pXC1290) and residues 194-414 (pXC1287) were PCR sub-cloned into PGEX-6p1 vector, from the initial plasmid containing full length of human $PRDM9_A$ in pDEST15 (Baudat et al. 2010). Both constructs were expressed as GST tagged fusion proteins in *Escherichia coli* BL21 (DE3) Codon-plus RIL. Cells were grown into LB media at 37°C until the OD_{600} reached to 0.5; at that point the temperature was lowered to 16°C, and 0.2 mM IPTG was added to induce expression. Cells were harvested by centrifugation and resuspended into lysis buffer containing 20 mM Tris (pH 7.5), 700 mM NaCl, 5% glycerol, 0.5 mM TCEP and 0.1 mM PMSF. Cells were lysed by sonication and clarified by centrifugation. Protein was purified

from cell extract using Glutathione Sepharose 4B (GE healthcare) as GST-tag affinity chromatography and GST-tagged protein eluted from the beads with buffer containing 100 mM Tris (pH 8.0), 500 mM NaCl, 5% glycerol, 0.5 mM TCEP and 20 mM reduced glutathione. The protein was diluted ~2.5X with 20 mM Tris-HCl (pH 7.5), 5% (v/v) glycerol, and 0.5 mM TCEP to reach ~200 mM NaCl and loaded onto HiTrap-Q column (GE Healthcare). Most of the N-terminal fragments were eluted out using a linear gradient of NaCl from 200 mM to 1 M. The eluted protein from the Q-column was subjected to proteolytic cleavage by ~100 μg of precission protease at 4°C overnight and reloaded on to Glutathione Sepharose 4B to remove cleaved GST tag. Protein was collected as flow through and loaded on Superdex-200 (16/60) column as final step of purification. (**c-d**) Mass spectrometry analysis of methylation kinetics by the two fragments (residues 1-414 and 194-414). (Top panels) The relative amount of each peptide species over the full time courses of the reactions, expressed as a percentage of the sum of intensity of all related peaks. (Bottom panels) Representative spectra at 2 min. The mass peaks for unmodified substrate and mono-, di-, and trimethylated products are increased by 14 Da for each methylation. Methyltransferase assay were conducted by mixing [E] (1 μM) with histone peptide H3 (residues 1-24) ([S]=100 μM) and *S*-adenosyl-L-methionine (AdoMet) [500 μM]. The reaction was incubated in 50 mM Tris (pH 8.5), 50 mM NaCl, 5% glycerol, and 2 mM DTT at room temperature and quenched at various time points by the addition of trifluoroacetic acid (TFA) to 0.5%. The quenched samples were diluted 1:4 with α-cyano-4-hydroxycinnamic acid solution prepared in 30% acetonitrile, 70% water and 0.1% TFA. MALDI TOF mass spectrometry was performed on a Burker Ultra FlexII TOF/TOF instrument (Biochemistry Department, Emory University) operated in reflection mode. Final spectra were averaged from 500 shots/position at 4 different positions. Methylation kinetics data were generated by integrating differentially methylated peptide peaks to give the relative quantity of each methylation state.



Supplemental Figure S2. PRDM9 C-terminal ZnF domain harbors DNA binding activity

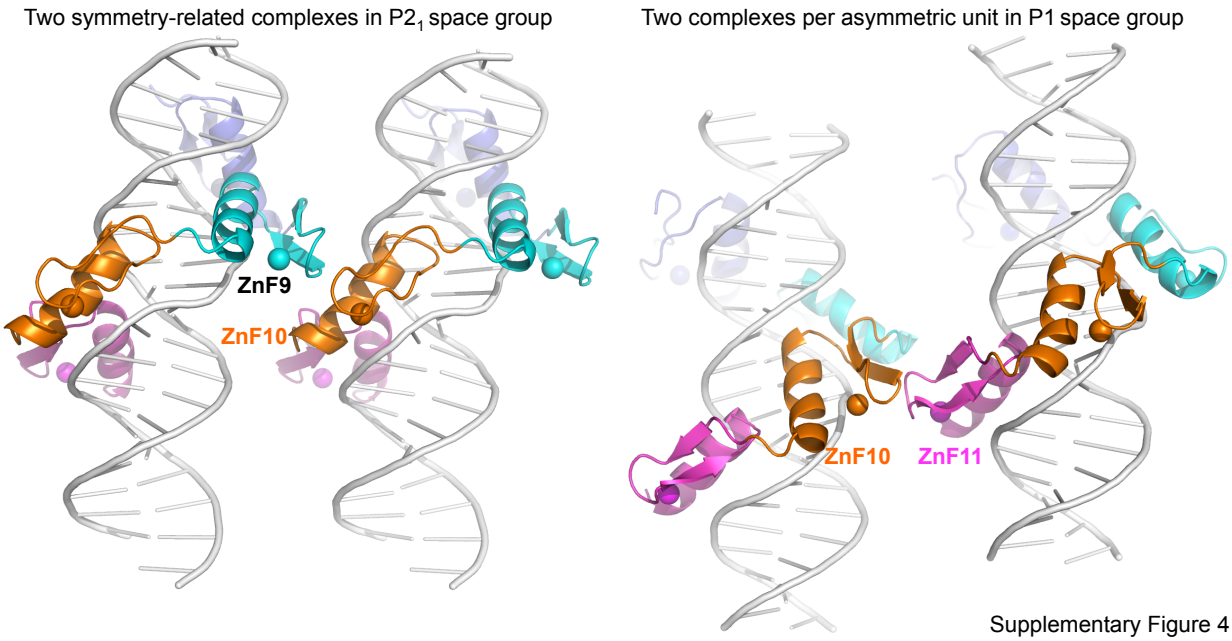
(**a**) Human PRDM9 contains a C-terminal tandem array of 13 repetitive C2H2 zinc fingers. (**b**) Sequence alignment of C2H2 ZnF with variations at positions -1, -4, -7 and -8, which correspond to the 6, 3, -1, and -2 residues, respectively, in the previous nomenclature (Wolfe et al. 2000). (**c**) SDS-PAGE gel showed the purified recombinant proteins (ZnF8-11, 8-12 and 8-13) used in this study. (**d**) SDS-PAGE showed the purified ZnF8-13 (lane 2), ZnF8-12 (lane 3), crystals of ZnF8-12 in complex with DNA washed 3-4 times with mother liquor (lane 4) and ZnF8-11 (lane 5). (**e**) DNA competition assay. ZnF8-12 [50nM] incubated with 5 nM FAM-labeled 24-bp THE1B DNA sequence in 20 mM Tris (pH 7.5), 300 mM NaCl, 5% glycerol, 25 μM $ZnCl_2$ and 0.5 mM TCEP, change in fluorescence polarization signal measured at increasing concentration of unlabeled DNA (ranging from 0 to 16.38 μM) containing the same THE1B sequence as well as with one base pair change (G:C in green to C:G in red) in the sequence as indicated. The data fit to the equation ([mP] = [maximum mP- baseline mP] × exp [-0.69/K_D] × [C]) + [baseline mP], where mP is millipolarization and [C] is the concentration of competitor DNA) in Graph-pad prim software (version 6.0) to estimate the K_D. (**f**) A model of the 5 ZnF fragment (ZnF8-12) in complex with DNA. We generated the fifth ZnF by superimposing ZnF8 and ZnF11, who share identical amino acid sequence (see panel b) with a root mean squared deviation of 0.6 Å and recognize the same GCC sequence. We placed ZnF9 after the relocated ZnF8 for the location of ZnF12. ZnF9 and ZnF12 differ only at -7 position (D vs. N) but recognizes two different sequences, TCC (ZnF9) and ACG (ZnF12).



Supplementary Figure 3

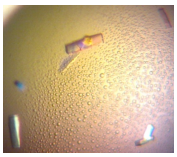
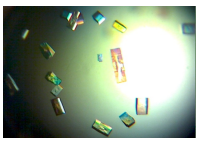
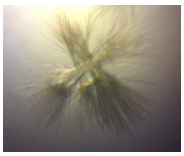
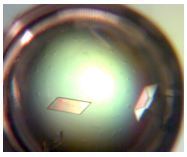
Supplemental Figure S3. Structure of hPRDM9$_A$ ZnF8-11-DNA complex

(**a**) K787 located at -6 position of ZnF10 bridges between two DNA phosphate groups. K787E variant (allele L9/L24) has reduced DNA binding affinity (see Fig. 3). (**b**) S788, located at -5 position of ZnF10, interacts with a DNA backbone phosphate group. (**c**) Amino acid sequence alignment of ZnF8 to 11. We note that the residues at -1, -4, -7, and -8 positions are the only changes among the four ZnFs. (**d**) Schematic ZnF8-11–DNA interactions. The negative numbers in parentheses refer to the amino acid positions within each ZnF as defined in panel **c**. (**e**) R757, located at -8 position of ZnF9, bridges between two DNA phosphate groups. (**f**) S813, located at -8 position of ZnF11, interacts with a DNA backbone phosphate group. S813R variant (allele-L13) has enhanced DNA binding affinity (see Supplemental Fig. S7).

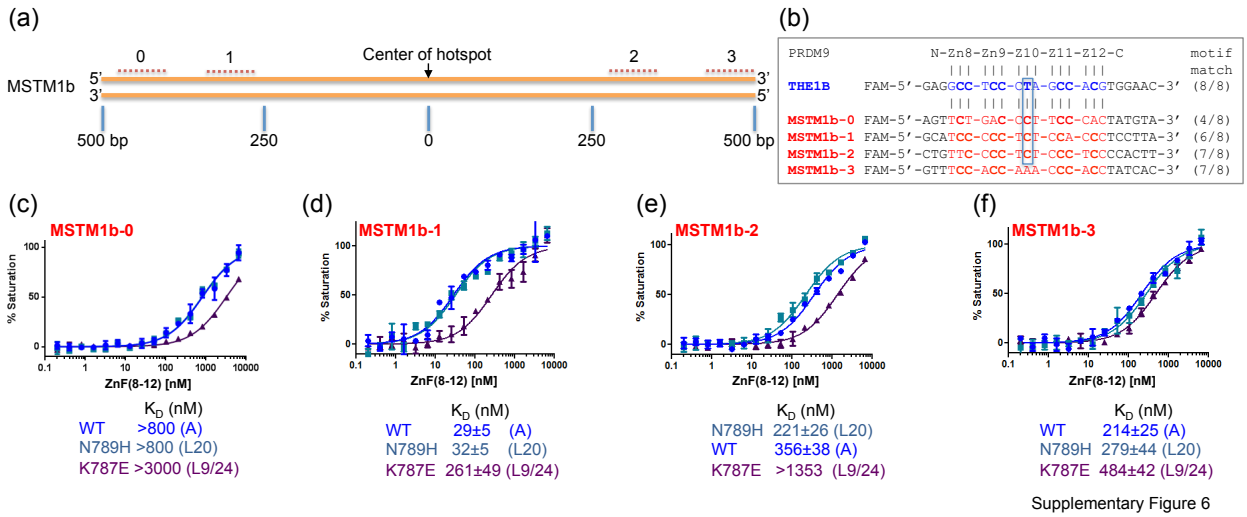


Supplemental Figure S4. Crystal packing contacts of two complexes in the space group $P2_1$ (left) and P1 (right).

bp	ZF8--9-10-11-12	Crystals	Diffraction
15+1	5′-TGCCTCCCTAGCCACG -3′ 3′- CGGAGGGATCGGTGCA-5′	NO	
16+1	5′-TGGCCTCCCTAGCCACG -3′ 3′- CCGGAGGGATCGGTGCA-5′	NO	
17+1	5′-TGGCCTCCCTAGCCACGT -3′ 3′- CCGGAGGGATCGGTGCAA-5′	NO	
18	5′-GCCTCCCTAGCCACGTGG-3′ 3′-CGGAGGGATCGGTGCACC-5′	small plates/rods (panels a and b)	poor (~15Å)
18+1	5′-TGCCTCCCTAGCCACGTGG -3′ 3′- CGGAGGGATCGGTGCACCA-5′	NO	
19	5′-GGCCTCCCTAGCCACGTGG-3′ 3′-CCGGAGGGATCGGTGCACC-5′	Needle clusters	
19+1	5′-TGCCTCCCTAGCCACGTGGA -3′ 3′- CGGAGGGATCGGTGCACCTA-5′	Needle-like	
20	5′-GGCCTCCCTAGCCACGTGGA-3′ 3′-CCGGAGGGATCGGTGCACCT-5′	Needle clusters (panel c)	poor (~20Å)
20+1 CA	5′-TGGCCTCCCTAGCCAC**G**TGGA -3′ 3′- CCGGAGGGATCGGTG**C**ACCTA-5′	suitable for data collection (panels d and e)	2 Å
20+1 GA	5′-TGGCCTCCCTAGCCAC**C**TGGA -3′ 3′- CCGGAGGGATCGGTG**G**ACCTA-5′	NO	
20+1 GC	5′-TGGCCTCCCTAGCCAC**CG**GGA -3′ 3′- CCGGAGGGATCGGTG**GC**CCTA-5′	NO	

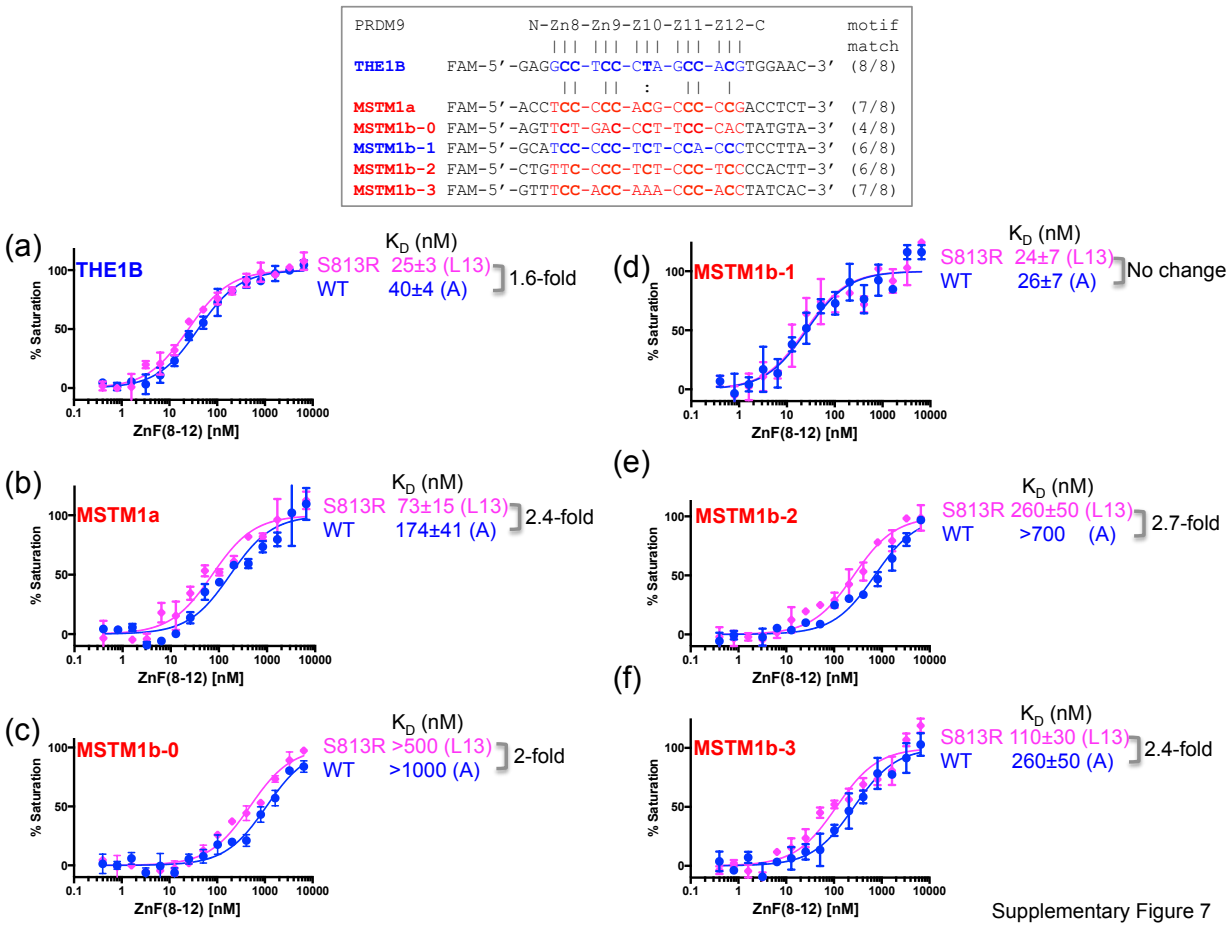
(a)	(b)	(c)	(d)	(e)
				
(Sitting drop) 0.1 M Bis-tris (pH 5.5) 25% PEG 3350 0.2 M NaCl	(Hanging drop) 0.1 M Bis-tris (pH 5.5) 25% PEG 3350 0.2 M NaCl	(Sitting drop) 0.1 M Bis-tris (pH 5.5) 25% PEG 3350	(Sitting drop) 60 mM Bis-tris propane, 40 mM Citric acid (pH 6.4) 20% PEG 3350	(Hanging drop) 60 mM Bis-tris propane, 40 mM Citric acid (pH 5.5) 18% PEG 3350

Supplemental Figure S5. Summary of DNA sequences used for crystallization, crystals observed and quality of X-ray diffractions. Five examples of crystals and corresponding conditions were shown (panels **a** to **e**).

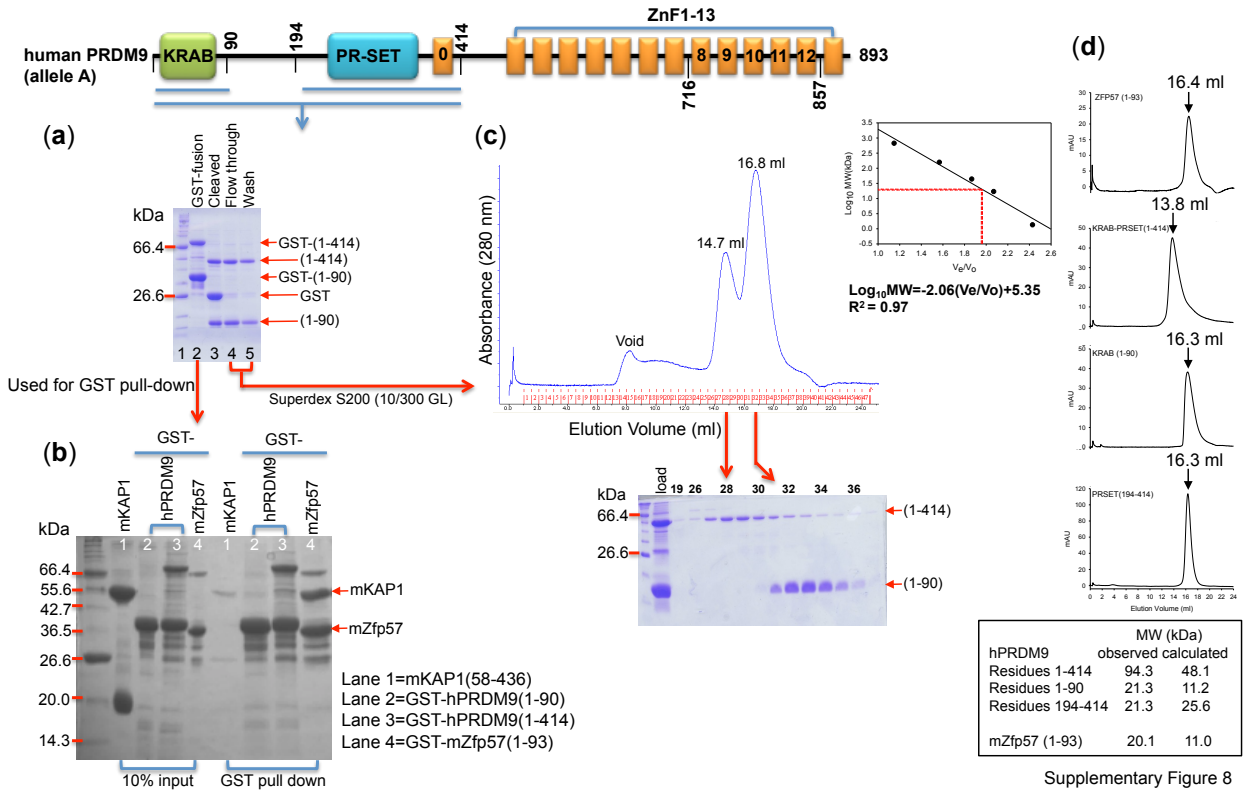


Supplemental Figure S6. Hotspot MSTM1b on chromosome 1q42.3

(**a**) Schematic showing the relative positions of identified hotspot sequences, MSTM1b-0 (Neumann and Jeffreys 2006) and MSTM1b-1, -2, and -3 (this study), within ±0.5 kb of the hotspot center (Jeffreys et al. 2005). (**b**) Four identified MSTM1b sequences aligned with THE1B sequence. We note that the conserved T:A base pair THE1B sequence has been changed to C:G base pair in three out of four MSTM1b sequences (boxed). (**c-f**) Comparison of DNA binding affinities of three different alleles, WT (allele-A), N789H (allele-L20), and K787E (allele-L9/L24), with oligos containing putative MSTM1b hostspot sequences.



Supplemental Figure S7. Comparison of DNA binding affinities of S813R (allele-L13**)** and WT (allele-A) in the context of ZnF8-12 with oligos containing hostspot sequences for (**a**) THE1B, (**b**) MSTM1a, and (**c-f**) MSTM1b-0 to -3.



Supplemental Figure S8. The N-terminal KRAB domain of hPRDM9 does not interact with KAP1, but potentially forms a dimer

(**a**) SDS-PAGE gel showed the partially purified GST-fusion of hPRDM9 fragment encompassing residues 1-414 (pXC1290). A low molecular weight band (residues 1-90) was co-purified (lane 2) after GST and HiTrap-Q columns. Precission protease (purified in house) cleaved GST tag from both bands (lane 3). The cleaved GST tag was removed by passing through Glutathione Sepharose 4B (lanes 4 and 5). (**b**) PRDM9 belongs to the family of KRAB domain containing ZnF transcription factors (Fig. 1a) (Collins et al. 2001; Huntley et al. 2006; Meylan et al. 2011; Liu et al. 2013). The KRAB domain from various proteins associates directly with KAP1 (Krüppel-associated protein 1) co-repressor (Friedman et al. 1996; Kim et al. 1996; Moosmann et al. 1996; Quenneville et al. 2011; Quenneville et al. 2012), which is also known as TRIM28 (tripartite-motif-containing 28), an essential regulator of genomic imprinting (Messerschmidt et al. 2012)). However, hPRDM9 KRAB domain did not interact with mouse KAP1 (residues 58-436; pXC1237) in a GST pull-down experiment, whereas mZfp57 KRAB domain did. The sequence identity is 96% between human and mouse KAP1 fragment used in the study (only 14 out of 379 residues are different and 5 of them are conserved changes) (not

shown). The mKAP1 protein [10 μM] (purified in house) incubated with 10 μM of GST tagged hPRDM9 (1-90 or 1-414). GST tagged mZfp57 KRAB domain (1-93) used as a control. Glutathione Sepharose 4B beads used to pull down GST tagged protein followed by washing with buffer containing 50 mM Tris (pH 7.5), 500 mM NaCl, 5 % glycerol, 5 mM β-mercaptoethanol and 1 μM $ZnCl_2$, boiled with 20 μL of 1X loading dye and subjected to run on SDS-PAGE. (**c**) Another interesting observation, that the increased hotspot strength variance in individuals heterozygous for PRDM9 (Pratto et al. 2014), led us to ask whether the presence of a second PRDM9 allele can influence hotspot strength of double-strand breaks by protein dimerization. In mouse, increased variance in hotspot strength was also observed in heterozygous animals relative to homozygotes (Pratto et al. 2014). For transcription factors, sequence changes at DNA binding sites explain partially differential transcription factors occupancy (Kasowski et al. 2010; Reddy et al. 2012). Many transcription factors are in homo- and hetero-dimerization (such as NF-κB (Hoffmann et al. 2006; Oeckinghaus and Ghosh 2009)). We asked whether hPRDM9 N-terminal KRAB domain could form dimer in solution. Elution profile of Superdex 200 (10/300 GL) column (GE Healthcare) illustrated that the KRAB domain (residues 1-90) or the KRAB-containing fragment (residues 1-414) has the apparent molecule weights (MW) twice of calculated MW, whereas the fragment devoid of the KRAB domain (residues 194-414) has the monomer size (see panel **d**). (**d**) Elution profiles of four consecutive runs on a Superdex-200 (10/300 GL) column with 20 mM Tris (pH 7.5), 300 mM NaCl, 5% glycerol, and 0.5 mM TCEP and containing (from top to bottom) mouse Zfp57 KARB domain (residues 1-93), hPRDM9 N-terminal fragment (residues 1-414), KRAB domain (residues 1-90), and PR-SET catalytic domain (residues 194-414). The inset shows the standardization of the size exclusion column using a protein marker kit (Biorad) at the time PRDM9 proteins were profiled using the same buffer.

Supplementary Table 1. Summary of X-ray diffraction and refinement

Protein	hPRDM9 ZnF8-12	hPRDM9 ZnF8-12	hPRDM9 ZnF8-12	hPRDM9 ZnF8-12
DNA	TG**GCCTCCCTAGCCACG**TGGA C**CGGAGGGATCGGTGC**ACCTA	TG**GCCTCCCTAGCCACG**TGGA C**CGGAGGGATCGGTGC**ACCTA	TAG**GCCTCCCTAGCCACG**TGG TC**CGGAGGGATCGGTGC**ACCA	TGAG**GCCTCCCTAGCCACG**TG CTC**CGGAGGGATCGGTGC**ACA
# of Crystals	(1)	(4)	(1)	(1)
PDB	-	5EI9	5EGB	5EH2
Space group	P1	P1	$P2_1$	P1
Cell dimensions (Å)	a=45.2, b=56.8, c=76.5	a=44.3, b=55.7, c=76.0	a=34.6, b=79.2, c=68.4	a=45.2, b=57.7, c=76.2
(°)	α=90.0, β=104.6, γ=94.7	α=89.8, β=75.5, γ=86.5	α=90, β=95.4, γ=90	α=90.1, β=75.3, γ=83.3
X-ray source	(SERCAT) APS 22-ID	APS 22-ID	Cu Kα	APS 22-ID
Wavelength (Å)	1.2814 (Zn edge)	1.0000	1.5418	1.0000
Resolution (Å) *	32.89-2.39 (2.48-2.39)	32.98-1.92 (2.02-1.92)	29.64-1.97 (2.04-1.97)	32.75-2.05 (2.12-2.05)
[A] R_{merge}	0.075 (0.423)	0.160 (0.948)	0.071 (0.712)	0.042 (0.456)
[B] <I/σI>	19.45 (3.1)	16.39 (2.6)	25.3 (1.9)	13.63 (2.2)
Completeness (%)	89.9 (55.7)	99.9 (99.6)	86.1 (37.6)	90.4 (76.8)
Redundancy	8.2 (5.3)	24.9 (13.8)	7.7 (4.5)	1.8 (1.6)
Obs. Reflections	212,976	1,311,673	171,468	73,602
Unique reflections	51,497 (1,598) (50,898 have I^+ and I^-)	52,634 (5,283)	22,386 (985)	41,966 (3,547)
Mean FOM (SAD)	0.451			
Refinement (ASU)	(Two complexes)	(Two complexes)	(One complex)	(Two complexes)
Resolution (Å)	2.39	1.92	1.97	2.05
No. reflections	25,197	52,506	22,258	41,841
[C] R_{work} / [D] R_{free}	0.205 / 0.242	0.215 / 0.239	0.193 / 0.219	0.201 / 0.230
No. Atoms				
Protein	1,804	1,774	934	1,815
DNA	1,709	1,709	865	1,710
Zn	8	8	4	8
Waters	7	169	167	122
B Factors ($Å^2$)				
Protein	68.3	48.7	36.7	58.6
DNA	85.6	62.8	43.6	70.3
Zn	76.5	47.9	37.7	55.9
Waters	44.8	48.3	37.6	55.4
R.m.s. deviations				
Bond lengths (Å)	0.006	0.009	0.010	0.009
Bond angles (°)	0.9	1.1	1.1	1.2
All atom clash score	10.9	7.1	7.6	7.1
Ramachandran plot	97.7 %	98.2 %	98.2 %	97.7 %
Additional allowed	2.3 %	1.8 %	1.8 %	2.3 %
Cβ deviation	1	0	0	0

*Values in parenthesis correspond to highest resolution shell; [A] $R_{merge} = \Sigma | I - <I>| / \Sigma I$, where I is the observed intensity and <I> is the averaged intensity from multiple observations; [B] <I/σI> = averaged ratio of the intensity (I) to the error of the intensity (σI); [C] $R_{work} = \Sigma | Fobs - Fcal | / \Sigma | Fobs |$, where Fobs and Fcal are the observed and the calculated structure factors, respectively; [D] R_{free} was calculated using a randomly chosen subset (5%) of the reflections not used in refinement.