bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Negative Cooperativity between Gemin2 and RNA

Determines RNA Selection and Release of the SMN Complex in snRNP Assembly

Running title: **Negative Cooperativity in snRNP Assembly**

Hongfei Yi, Li Mu, Congcong Shen, Xi Kong, Yingzhi Wang, Yan Hou and Rundong Zhang*

Department of Ophthalmology,

State Key Laboratory of Biotherapy,

West China Hospital,

Sichuan University and Collaborative Innovation Center for Biotherapy,

Chengdu 610041, P. R. China

*Corresponding author

Phone: 86-18200541808

Email: rundongzhang@scu.edu.cn

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ABSTRACT

The assembly of snRNP cores, in which seven Sm proteins, D1/D2/F/E/G/D3/B, form a ring around snRNAs, is the early step of spliceosome formation and essential to eukaryotes. It is mediated by the PMRT5 and SMN complexes sequentially in vivo. The deficiency of SMN causes neurodegenerative disease spinal muscular atrophy (SMA). How the SMN complex assembles snRNP cores in the second phase is largely unknown, especially how the SMN complex achieves stringent RNA specificity, ensuring seven Sm proteins assemble only around snRNAs, by requiring an extra 3'-adjacent stem-loop (SL) in addition to a nonameric Sm site RNA (PuAUUUNUGPu) on which snRNP cores can spontaneously form without chaperons in vitro. Moreover, how the SMN complex is released from snRNP cores is unknown. Here we show that Gemin2 of the SMN complex and RNA allosterically and mutually inhibit each other's binding to SmD1/D2/F/E/G, coupling RNA selection with the SMN complex's release. Using crystallographic and biochemical approaches, we found that Gemin2 constrains the horseshoe-shaped SmD1/D2/F/E/G in a physiologically relevant, narrow state, which prefers the snRNP-code (both the Sm site and 3'-SL)-containing RNA for assembly. Moreover, the assembly of RNA widens SmD1/D2/F/E/G, causes Gemin2's release allosterically and allows SmD3/B to join. By structural analysis we further propose a structural mechanism for the allosteric conformational changes. These findings provide deeper insights into the SMN complex's mode of action and snRNP assembly, and facilitate potential therapeutic studies of SMA.

.

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

INTRODUCTION

Small nuclear ribonucleoprotein particles (snRNPs) are major building blocks of the spliceosome, which carries out precursor mRNA splicing in eukaryotes. All snRNPs share a common feature: seven Sm (D1, D2, F, E, G, D3 and B) or Sm-like proteins (Lsm2-8) form a ring around a segment of the small nuclear RNA (snRNA) after which the snRNP is named. Correspondingly, the snRNPs can be divided into two classes: Sm-class snRNPs (U1, U2, U4 and U5 snRNPs for the major spliceosome, and U11, U12, U4atac and U5 for the minor spliceosome) and Sm-like-class snRNPs (U6 and U6atac snRNPs) [1, 2]. In addition, their assemblies also take different pathways. While Sm-like-class snRNPs are assembled completely inside the nucleus and without the assistance of assembly chaperons (will not be discussed hereafter), Sm-class snRNPs are assembled in both the nucleus and cytoplasm, and are mediated by a number of assembly chaperons [2, 3]. After being transcribed in the nucleus, precursor snRNAs (pre-snRNAs) are exported into the cytoplasm, where seven Sm proteins are assembled on the Sm site, PuAUUUNUGPu, of the RNAs to form snRNP cores (Sm cores). Proper assembly of the Sm core is pre-required for hypermethylation of snRNA's cap and import into the nucleus. After import into the nucleus, Sm-class snRNPs are maturated by further modification of RNA and joining of proteins specific to individual snRNP before they participate in pre-mRNA splicing.

Sm core assembly is a pivotal step of snRNP biogenesis and essential for eukaryotes [4]. Early studies established that Sm core assembly can occur spontaneously in vitro by mixing the three Sm hetero-oligomers, SmD1/D2, SmF/E/G and SmD3/B, with snRNA, or even oligoribonucleotide containing just the nanomeric Sm site [5, 6]. The reaction takes a stepwise fashion. SmD1/D2 and SmF/E/G bind RNA to form a stable subcore, and then SmD3/B joins to form a highly stable Sm core [5]. Interestingly, inside cells, Sm core assembly is mediated by a number of assembly chaperon proteins, classified into two complexes, the PRMT5 (protein arginine methyltransferase 5 complex, including 3 proteins: PRMT5, WD45 and pICln) and SMN complexes (survival motor neuron complex, including 9 proteins: SMN, Gemin2-8, and unrip) in vertebrates [2, 7]. Since cells contain many RNAs which have

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sequences resembling the nanomeric Sm site, Sm core can potentially assemble on many illicit RNAs and cause deleterious consequence. These assembly chaperons, especially the SMN complex [8], are believed to confer highly specific Sm core assembly, ensuring Sm proteins to assemble exclusively on cognate snRNAs, which contain both the nanomeric Sm site and a 3'-adjacent stem-loop (SL), altogether termed as the snRNP code [9].

These two complexes perform assembly chaperoning roles in consecutive phases. In the first phase, PRMT5/WD45 methylate the C-terminal arginine residues of SmD3, SmB and SmD1, which is believed to enhance the interactions between Sm proteins and SMN [10, 11]. pICln recruits SmD1/D2 and SmF/E/G to form a ring-shaped 6S complex, which pre-arranges the 5 Sm proteins in the finally assembled order and simultaneously prevents the entry of any RNAs to the RNA-binding pocket [12]. In addition, pICln also binds SmD3/B [12]. In the second phase, the SMN complex accepts SmD1/D2/F/E/G (5Sm) and SmD3/B and releases pICln [12]. Gemin2 is the acceptor of 5Sm [13, 14]. SMN binds Gemin2 by its N-terminal Gemin2-binding domain (Ge2BD, residues 26-62) [13]. Both SMN and Gemin2 are highly conserved in eukaryotes [15]. Either *smn* or *Gemin2* gene knockout causes early embryonic death in vertebrates, indicating the essential roles of the SMN complex in eukaryotic cells [16, 17]. Moreover, the deficiency of SMN causes human neurodegenerative disease spinal muscular atrophy (SMA), emphasizing the pathophysiological relevance of the Sm-core assembly pathway [18-20]. Therefore, understanding the mechanism of Sm core assembly, especially at the second phase, is of great importance because of both its fundamental role in gene expression and its potential application in SMA therapy. SMN also interacts with Gemin8 by its C-terminal self-oligomized YG box [21] and Gemin8 further binds Gemin6/7 and Unrip, but their roles are poorly understood [21-23]. Gemin3 contains a DEAD box domain and is thought to be a putative RNA helicase [24]. Gemin4 usually forms a complex with Gemin3, but its role is unknown [25]. Gemin5 is the component to initially bind pre-snRNAs and deliver them to the rest of the SMN complex for assembly into the Sm core [26], and is currently

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

considered to be the protein conferring the RNA assembly specificity by direct recognition of the snRNP code [27-32].

Recent structural studies of some assembled and intermediate complexes of Sm core assembly have provided great insights into the mechanisms of this complicated process. The assembled structures of U1 snRNPs and U4 snRNP cores explain how the nanomeric Sm site RNA interacts specifically with the seven Sm proteins [33-37]. The structures of the 6S complex (human 5Sm plus fly pICln), the 8S complex (6S plus fly Gemin2/SMN-N-terminal domain) and the later phase human SMN(26-62)/Gemin2/5Sm complex (hereafter we will refer to it as the 7S complex for brevity because it is equivalent to the 7S complex reported earlier which contains additional segments of SMN[12]) provide detailed insights into the mechanisms of the first phase, the transition from the first phase to the second phase, and the initial state of the second phase[13, 14]. Just recently, the structures of Gemin5's N-terminal WD domain complexed with oligoribonucletide containing the Sm site explain the mode of interaction between Gemin5 and RNA [29-31].

Despite these advances in understanding the mechanisms of Sm core assembly, there are still many important questions unanswered or not well explained, especially in the second phase. The first question is how the SMN complex determines RNA assembly specificity. This is the central question of Sm core assembly because it is the reason why these chaperons have evolved and exist. Although current knowledge considers that Gemin5 is the right protein by direct binding to the snRNP code and this model is partially supported by some experimental data [27-32], there are several paradoxical observations this model cannot explain. First, Sm core assembly is a highly conserved pathway in all eukaryotes, but there is no homolog of Gemin5 in many lower eukaryotes [15, 32]. Second, recent structural and biochemical studies showed that the RNA-binding specificity of Gemin5 is only able to recognize part of the Sm site, AUUU, not to mention the full feature of the snRNP code [29-31]. Third, Gemin5 can bind promiscuous RNAs, i.e., U1-tfs, the truncated U1 pre-snRNAs lacking the Sm site and the following SL [29]. These paradoxes suggest that the specificity mechanism has not been answered yet. The second significant

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

question is how the SMN complex is released from the Sm core. In the spliceosome, the mature snRNPs do not contain any component of the SMN complex [38, 39], but most proteins of this complex have been observed to enter the nucleus and concentrate on Cajal bodies (CBs) [26]. Moreover, our previous 7S complex structure and biochemical tests show that Gemin2 tightly binds to 5Sm[13]. How the SMN complex comes off the mature Sm cores has been completely unknown.

In this study, we examined closely the assembly reactions in the second phase, from the initial state of the 7S complex formation to the completion of the Sm core, by a combination of crystallographic and biochemical approaches. We found that Gemin2 is the protein conferring Sm core assembly specificity by a negative allosteric mechanism. It constrains 5Sm in a narrow, physiologically relevant state, which selects the cognate snRNAs, containing the snRNP code, to assemble into the Sm subcore. snRNAs' assembly widens 5Sm, unexpectedly causing Gemin2's release, and also allowing SmD3/B to join to form the Sm core. Further structural analysis reveals the structural mechanism for the negative allosteric conformational changes. These results provide deeper insights into the second phase of Sm core assembly, answer the above two basic questions, and facilitate therapeutic studies of SMA.

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

RESULTS

The narrow conformation of 5Sm bound by Gemin2 is not an artifact from crystal packing

In the second phase of Sm core assembly, the crystal structure of the 7S complex we determined previously is an initial state [13]. It reveals how Gemin2 binds 5Sm. Interestingly, we also observed that the conformation of 5Sm in the 7S complex is narrow compared with the mature snRNP core structures [33, 35-37](Fig. 1a-b). However, it is unknown whether the narrowness of 5Sm in the 7S complex is a real, physiologically relevant state or just an artefact arising from crystal packing, because in the crystal lattice of the 7S complex a second Gemin2's C-terminal domain (CTD) is located right in between SmD1 and SmG contacting both (Fig. 1c) and crystal packing inducing artificial conformations is well documented [40]. It is very likely that the second Gemin2's CTD pulls SmD1 and SmG close to each other and artificially induces the narrowness of 5Sm. In the first phase of Sm core assembly, there are available structures of two complexes [14], the 6S complex, in which pICln binds 5Sm in a ring shape, and the 8S complex, in which Gemin2/SMNΔC bind to the peripheral side of 5Sm in the 6S complex. In both these complexes, the conformations of 5Sm are also narrow, and more precisely, even narrower than that in the 7S complex, as the Cα-Cα distances between N37.SmD1 and N39.SmG are 25.5 Å in the 6S complex and 26.2 Å in the 8S complex, versus 27.4 Å in the 7S complex. The narrowness of 5Sm in both these complexes is because the narrow-sized pICln (occupying only the angular space of one and a half Sm proteins) contacts SmD1 and SmG [14], and therefore cannot provide any clue for the conformation of 5Sm in the 7S complex where pICln is absent. Moreover, the interfaces between Sm and Sm-like proteins are relatively pliable because they can form hexamers, heptamers and even octamers [41-43]. It is not implausible that 5Sm bound by Gemin2 in the 7S complex is in a wide conformation as it is in the final Sm core. So, to study the mechanism of Sm core assembly in the second phase, it is necessary to identify first whether the conformation of 5Sm bound by Gemin2 is narrow or wide.

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To test this, we used a crystallographic approach and attempted to pack the 7S complex and its derivatives in different crystal lattices to avoid the above packing contacts. After we failed at obtaining a different crystal of the original 7S complex by trying different solution conditions, we made Complex A, a 7S complex with a short version of SmD1, SmD1s, in which the nonessential C-terminus (residues 83-119) was truncated, replacing SmD1. Complex A formed crystals in different crystal lattice and its packing is significantly different from the previous 7S complex (Table S1 and Fig. S1), however, Gemin2's CTD is still located in between the SmD1 and SmG of another complex and the distance between the SmD1 and SmG is little altered. We also made several other 7S complex variations, among which the most significant one is Complex B, a derivative of Complex A without SmG and with Gemin2ΔN39 (the N-terminal residues 1-39 are truncated to further test the effect of the N-tail) replacing Gemin2 (Fig. 1d-e, S1 and Table S1). In the crystal lattice, although SmD1 is still in contact with Gemin2's CTD of a second complex, due to the absence of SmG, SmE at the other end of the crescent Sm hetero-oligomer is far away from the second Gemin2's CTD for interaction, eliminating the influence of crystal packing on the curvature of the Sm proteins (Fig. 1e). However, the curvature of D1/D2/F/E is little different from that of the original complex as indicated by no increase of the Cα-Cα distance between the most conserved residues Asn37 of SmD1 and Asn55 of SmE (28.6 Å in Complex B vs. 29.4 Å in the previous 7S complex) (Fig. 1b,d). So, this observation indicates that the narrow conformation of 5Sm bound by Gemin2 is not caused by crystal packing artifact, but is a physiologically relevant one. We term this conformation as the ground state.

The narrowness of 5Sm is not caused by Gemin2's N-terminal tail.

Since in the previous structure of the 7S complex the N-terminal tail (N-tail, residues 22-31) of Gemin2 is located inside the central RNA-binding pocket, it is possible that this N-tail induces the narrowness of the 5Sm. To test this possibility, we reconstituted Complex C, a derivative of Complex A with Gemin2ΔN39 replacing Gemin2, for crystallization. The crystal of Complex C had the same space group as Complexes A and B and similar unit cell parameters to them (Table S1). However, the curvature

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of 5Sm is still the same as that of the previous complex (Data not shown). In addition, we also created Complex B (see above) for crystallization. The absence of Gemin2's N-tail did not change the curvature of D1/D2/F/E (Fig. 1d-e). These data demonstrate that Gemin2's N-tail does not play a role in the narrowness of 5Sm and the narrowness of 5Sm is caused by the rest part (residues 40-280) of Gemin2.

Gemin2's N-tail flips dynamically and plays a minor inhibitory role in RNA binding

In addition, surprisingly, in one 7S complex crystal structure with the full-length Gemin2 (Complex A), we observed that there was no electron density inside the RNA-binding pocket of 5Sm (Fig. 1f, Table S1 and Fig. S1), in contrast to the previous complex structure where Gemin2's N-tail is inside the RNA-binding pocket [13]. Complex A was crystallized under a condition similar to the previous 7S complex, but has a slightly different crystal packing (Table S1 and Fig. S1). Checking the components of crystal sample by SDS-PAGE, we saw the band of Gemin2 keeping the original full-length size, indicating no degradation (Fig. 1g). So it is only reasonable to explain that Gemin2's N-tail was located outside and flexible in the crystal of Complex A. These data indicate that Gemin2's N-tail may not be located firmly in one place; instead its positions may be quite dynamic. Consistent with this was that the peak of 7S (containing the full-length Gemin2) eluted earlier than SMN(26-62)/Gemin2ΔN39/5Sm (7SΔN39) in gel filtration chromatography (GFC) (Fig. S2), indicating that Gemin2's N-tail flips outside the pocket and increase the complex's size.

In the previous study, an inhibitory role of Gemin2's N-tail on snRNAs' binding to 5Sm was observed [13]. However, the experiments were carried out by mixing separate SMN_{Ge2BD}/Gemin2 (or SMN_{Ge2BD}/Gemin2ΔN39) and 5Sm at various ratios, and this way could not faithfully mimic the physiological state, in which 5Sm binds to Gemin2/SMN in an equal stoichiometry. In this study, we used preformed 7S and 7SΔN39 to examine snRNA binding. Using electrophoresis mobility shift assay (EMSA), we observed that 7S or 7SΔN39 formed Sm subcore with U4 snRNA as its concentration increases, while the highest concentration of 7SΔN39 tested could not bind the negative control, U4ΔSm

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

RNA, in which the nonameric Sm site was replaced by AACCCCCGA (Fig. 1h). 7SΔN39 formed more subcore with U4 snRNA than 7S, but its binding efficiency was only about 2-fold higher than that of 7S (Fig. 1h). We concluded that Gemin2's N-tail has a minor inhibitory role in 7S binding to snRNAs. This conclusion is consistent with the crystallographic and chromatographic observations and the dynamic nature of Gemin2's N-tail.

Binding to 5Sm in 7SΔN39 needs more RNA features than the Sm site

In in vitro experiments, mixing D1/D2, F/E/G and a 9-nucleotide Sm-site RNA (9nt), AAUUUUUGA, could readily produce a stable Sm subcore [6]. We wondered whether a preformed 7S with 5Sm in the narrow state would similarly accept a Sm-site RNA. Since Gemin2's N-tail can flip outside the RNA-binding pocket of 7S and does not play a major role in snRNA binding, to simplify our analysis, we used Gemin2ΔN39 instead of the full-length Gemin2 to perform RNA binding experiments. Furthermore, to facilitate analysis of complex components by taking advantage of purified proteins and RNAs, we adopted GFC instead of EMSA. Three parameters are generally monitored for the formation of a RNA-protein complex: peak elution volume (position), ratio of OD_{260nm} to OD_{280nm} ($OD_{260/280}$), and SDS-PAGE followed by silver staining or Coomassie brilliant blue (CBB) staining. Using the 9nt, AAUUUUUGA, to incubate with D1/D2 and F/E/G followed by GFC separation, we observed the formation of Sm subcore (the peak at 14.37 ml with $OD_{260/280}$ over one) (Fig. S3, b-d), which is consistent with the early report [6]. To better detect RNA by SDS-PAGE and silver staining, we used a longer RNA (37nt) containing the Sm site at its 3' end (3' Sm) to perform the same experiment and observed the formation of Sm subcore (Fig. S3, a-c,e). Surprisingly, however, using the same 3' Sm RNA to incubate with the preformed 7SΔN39 complex, we could only see the 7SΔN39 peak (13.78 ml) but no formation of any RNA-protein complex (Fig. 2a); similar observation was made by using a negative control, U4ΔSm snRNA (Fig. 2b). This indicated that the Sm-site RNA, even with additional single-stranded RNA at its 5'-end, cannot bind to 5Sm when the latter is bound by Gemin2ΔN39/SMN_{Ge2BD}. In contrast, using a middle-sized, 3' fraction of human U4 snRNA (U4 snRNA), which has SLs flanking the Sm site and had

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

proved to assemble into the Sm core [44], to incubate with 7SΔN39, we observed the formation of Sm subcore (peak at 13.31 ml), albeit in a small percentage, from GFC separation (Fig. 2c). These observations indicate that the narrow conformation of 5Sm bound by Gemin2 plays a restrictive role in RNA binding. The 7SΔN39 state can only bind the normal U4 snRNA to a limited degree, and cannot bind the Sm-site RNA at all.

This surprising observation triggered us to ask what RNA feature the 7SΔN39 can recognize. Does it match the snRNP code previously identified by cell-based experiments [9]? To make a systematic study, we designed several derivatives of U4 snRNA by linearizing or deleting the SL at either side of the Sm site, one at a time. We used the same procedure as above to test the binding of these RNA variants to 7SΔN39. At first, we changed the 3'-SL of U4 snRNA to a linear single strand (U4-3'ss). The GFC trace showed three peaks eluted at 13.11, 13.80 and 14.48 ml (Fig. 2d), which were the 7SΔN39-RNA complex, 7SΔN39 and RNA respectively. This observation suggests that single-stranded RNA at the 3' end of the Sm site can still bind to 7SΔN39. When the 3'SL was completely removed (U4-3'Δ), however, the RNA did not bind to 7SΔN39 as showed by the absence of an earlier peak than the 7SΔN39 peak (~13.82 ml) (Fig. 2e). This result indicates that the presence of RNA at the 3' side of the Sm site, either single- or double-stranded, is critical for the formation of a 7SΔN39-RNA subcore.

When we linearized the 5'-SL of the U4 snRNA (U4-5'ss), we observed that besides the RNA peak (15.67 ml), a new RNA-containing peak, because of its $OD_{260/280}$ over one, came early at about 13.70 ml (Fig. 2f). The silver staining result also proved the presence of RNA in the peak. Surprisingly, we observed a subtle difference of the OD_{260} and OD_{280} peak positions (13.65 and 13.71 ml respectively). This indicated that there might be two peaks of similar sizes coming out at about 13.7 ml, one being 7SΔN39, and the other being a complex containing RNA. But the identity of the latter complex was perplexing. When we used the 5'SL-deleted form of U4 snRNA (U4-5'Δ) to do the experiment, we observed an striking GFC profile (Fig. 2g): besides the RNA peak at 17.49 ml, there were two peaks in front, one OD_{280} peak at 14.10 ml, containing higher protein/RNA ratio, and the other OD_{260} peak at 14.29

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ml, containing higher RNA/protein ratio. The early peak was very likely 7SΔN39, which should elute at about 13.7 ml, but the overlapping with the later peak shifted its precise OD_{280} value. The observation that the RNA-containing peak eluted later than 7SΔN39 was very surprising in that the size of 7SΔN39 if bound by RNA would likely be no less than that of 7SΔN39 alone and would generally elute no later than 7SΔN39. It must not be RNA alone because RNA alone came out only at 17.48 ml (Fig. S4c). At this point, there would be two possibilities: either the binding of RNA to 7SΔN39 changes the conformation and reduces its hydrophilic volume or there was a loss of protein components upon the binding of RNA to 7SΔN39, logically Gemin2/SMN_{Ge2BD}. Although the latter was buttressed by further experiments in which the Sm subcores reconstituted from these RNAs with D1/D2, F/E/G were eluted at similar positions to their corresponding OD_{260} peaks described above (Fig. S4a-e), at this time we were unable to make a clear distinction. Anyway, these observations showed that the 5' RNA of the Sm site is not required for RNA binding, but a single-stranded RNA at the 3' side of the Sm site seems necessary and sufficient to bind to 7SΔN39. To further confirm this conclusion, we used a minimal RNA containing only the Sm site and 3' single-stranded RNA (U4-5'Δ-3'ss) to perform the assay. As we expected, an RNA-protein complex formed at 13.97 ml (Fig. S4f).

The release of Gemin2 during Sm core assembly

From the above experiments, three types of RNAs, (1)U4, in which 2 SLs tightly flank the Sm site, (2)the Sm site plus a 3'SL and (3)the Sm site plus a 3' single strand, were observed to bind to 7SΔN39, but they seemed to behave differently in terms of their expected Sm subcore sizes and protein components. To better monitor whether Gemin2ΔN39/SMN_{Ge2BD} is released, to which extent, and at which step of Sm-core assembly, we created full-length U4 snRNA (flU4) and several derivatives, the assembly of which into Sm cores would potentially elute earlier and have a better separation from 7SΔN39 as well as Gemin2ΔN39/$SMN_{Ge2BD.}$ At first, the negative control, flU4ΔSm RNA, incubated with 7SΔN39 for GFC, had no RNA-containing complex formed, but only the separate RNA peak (12.93 ml) and 7SΔN39 peak (13.75 ml) (Fig. 3a). In contrast, incubating flU4 snRNA with 7SΔN39, we observed that a small peak

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

containing RNA appeared the earliest at 12.26 ml. This peak fraction contained all 5 Sm proteins and Gemin2ΔN39, but the stoichiometry of Gemin2ΔN39 to 5Sm was less than 1:1 (Fig. 3b. compare lanes 11B and 14A). These experiments showed that Gemin2ΔN39/SMN_{Ge2BD} have started to dissociate from 5Sm when U4 snRNA binds to 5Sm.

In the previous section, we used U4-5'ss or U4-5'Δ RNA for binding assay and noticed an aberrant peak of RNA-protein complex, which was suspected to be the Sm subcore without Gemin2ΔN39/SMN_{Ge2BD} bound. However, the elution positions of both the Sm subcores were too close to that of 7SΔN39 to see the absence of Gemin2ΔN39. We wondered if the absence of an adjacent SL at the 5' side of the Sm site (type 2 RNA) could cause a complete release of Gemin2ΔN39/SMN_{Ge2BD} at the step of Sm subcore formation. To test it, we made a derivative of flU4 snRNA, flU4-spacer, to insert a room between the Sm site and its 5' SL by replacing the 3 nucleotides, GGC, intimately 5' adjacent to the Sm site, with CCG. The incubation of flU4-spacer with 7SΔN39 gave rise to a Sm subcore with a complete removal of Gemin2ΔN39 (Fig. 3c, lanes 11A-12A), indicating that a free or SL-free 5' end of the Sm site did cause a complete release of Gemin2ΔN39/SMN_{Ge2BD} from the Sm subcore. But the presence of free RNA (peak at 12.97 ml, also see Fig. S5b) and free 7SΔN39 (Fig. 3c, lanes 13B-14B) indicated that the formation of the subcore was in an equilibrium with the reactants.

To test the assembly of the type 3 RNA, we used an further derivative of flU4 snRNA, flU4-spacer-3'ss, which linearized the 3' SL on the basis of flU4-spacer. The incubation of flU4-spacer-3'ss with 7SΔN39 generated a gel filtration profile similar to flU4, in which a small fraction of Sm subcore formed, to part of which Gemin2ΔN39 was still bound (Fig. 3d).

As we proved, 7SΔN39 has a narrow conformation of 5Sm, which is in conflict with SmD3/B binding. Does the binding of RNA to 7SΔN39 expand the SmD1-G opening to allow SmD3/B to join? What about Gemin2 release upon Sm-core assembly? To test these, we incubated 7SΔN39, flU4 snRNA and D3(1-75)/B(1-91) (the nonessential C-terminal tails of both are truncated), and subjected the mixture

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

to GFC. The earliest and also highest peak (12.31 ml) contained RNA and all 7 Sm proteins but no Gemin2ΔN39 (Fig. 3e). The band of Gemin2ΔN39 appeared at about 15.5-16.5 ml on SDS-PAGE, consistent with the position of Gemin2ΔN39 alone (Fig. S3f), indicating that the Gemin2ΔN39 was in a free state. Furthermore, few 7SΔN39 complex components at the positions 13.5-14.5 ml (Fig. 3e, lanes 13B and 14A) indicated that almost all Sm pentamer was assembled into the Sm core. These results showed that Sm-core assembly goes to completion upon the joining of SmD3/B to the 7SΔN39-RNA complex, and simultaneously causes a complete release of Gemin2ΔN39/SMN_{Ge2BD}. The incubation of 7SΔN39, U4 snRNA and D3(1-75)/B(1-91) also gave rise to a similar conclusion (Fig. S5c). For flU4-spacer, which seems more efficient in forming the Sm subcore than flU4 (Fig. 3b-c), the addition of D3/B would drive the Sm core formation to a completion as in the case of flU4 snRNA.

The incubation of flU4-spacer-3'ss with both 7SΔN39 and D3(1-75)/B(1-91) also produced the Sm core and caused Gemin2ΔN39 to dissociate, but the assembly did not proceed to a completion, as indicated by the presence of free RNA (peak at about 12.7 ml), 7SΔN39 (lanes 13B-14B) and D3/B (lanes 16B-17A) (Fig. 3f). This indicated that RNAs containing the Sm site and 3' single strand, although can assemble into the Sm core, are less efficient substrates than RNAs containing the Sm site and 3'-SL.

The snRNP code assembles into 5Sm of 7SΔN39 more efficiently

The Sm site plus either a 3'-SL or a single-stranded RNA can bind the 7S complex and assemble into the Sm core. But the above experiments suggested that they might have different efficiency. To directly compare assembly efficiency, we performed a competition study by incubating 7SΔN39 with equal molar amount of flU4-spacer and U4-5'Δ-3'ss. The large difference of their RNA sizes makes their subcore formation visible on SDS-PAGE. The major fractions of Sm subcore containing flU4-spacer appeared on lanes 11A-12B, whereas the major fractions of Sm subcore containing U4-5'Δ-3'ss on lanes 13B-14B, which overlapped 7SΔN39 and made precise quantification impossible (Fig. 4a). In spite of this, a simple comparison of the darkness of the Sm proteins showed that the Sm subcore containing flU4-

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

spacer dominated, indicating that the snRNP code is more efficient than the Sm site plus a 3' single strand in subcore formation. In addition, we swapped the 5' portions of the RNAs and performed a competition study by incubating 7SΔN39 with equal molar amount of flU4-spacer-3'ss and U4-5'Δ. This time, the major fractions of the Sm subcore containing flU4-spacer-3'ss became weak (lanes 11A-12A), while the Sm subcore containing U4-5'Δ became dark (lanes 13B-15A) (Fig. 4b). This result confirmed that the snRNP code assembles more efficiently than the Sm site plus a 3' single strand. We also incubated 7SΔN39 with equal amount of the two RNAs, U4-5'Δ-3'ss and U4-5'Δ, which were identical in length and had no 5' extra portion (Fig. S4g). Consistent with our anticipation, the front peak of OD_{260} appeared at 14.25 ml, close to the peak of the Sm subcore containing U4-5'Δ (14.29 ml), while the free RNA appeared at 16.88 ml, close to the peak of free U4-5'Δ-3'ss (16.62 ml) instead of the peak of free U4-5'Δ (17.48 ml), indicating that more U4-5'Δ was assembled into the Sm subcore.

In addition, to compare the assembly efficiency of the final Sm core, we made a competition analysis by incubating equal molar amount of flU4-spacer-3'ss and U4-5'Δ with both 7SΔN39 and D3(1-75)/B(1-91). The peak of the Sm core containing flU4-spacer-3'ss appeared at about 12.0 ml, whereas the peak of the Sm core containing U4-5'Δ at about 14.5 ml. Gemin2ΔN39 came later, at about 15.5 ml (Fig. 4c). Comparing the darkness of 7 Sm proteins in lanes 13B-15B with that in lanes 11A-12B, we could estimate that the assembly of Sm core on U4-5'Δ was 2-fold more than on flU4-spacer-3'ss. This showed that Sm-core assembly is more efficient on the snRNP-code RNA than on the Sm site in the middle of a linear RNA. This result is consistent with the previous report, in which a des-stem RNA (equivalent to the Sm site plus a 3' single strand) was microinjected into the cytoplasm of Xenopus oocytes and its assembly efficiency into the Sm core was reduced by 2 folds [9].

The assemblies of the two types (types 1 and 2) of RNAs containing the snRNP code, with or without the 5' adjacent SL of the Sm site, into the Sm core have little difference, as demonstrated by the incubation of equal molar amount of flU4 and U4-5'Δ with 7SΔN39 and D3(1-75)/B(1-91) followed by GFC (Fig. 4d). Similar amount of Sm cores were observed to assemble on flU4 and U4-5'Δ.

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Gemin2 serves as a negative allosteric modulator of Sm core assembly

Superposition of 7SΔN39 with U4 Sm core [37] (Fig. S6) or U1 Sm core [36] (Data not shown) on SmF/E/G reveals that there is no clash of Gemin2's N-terminal domain (NTD) with RNA, and on SmD1/D2 reveals that there is no clash of Gemin2's CTD with RNA too. This indicates that Gemin2ΔN39 and RNA are not spatially exclusive. However, the binding of Gemin2ΔN39 on the periphery of 5Sm inhibits RNA assembly onto the central RNA-binding pocket of 5Sm, allowing only the cognate, the-snRNP-code-containing RNAs preferably to assemble into the Sm subcore. Moreover, the binding of cognate RNAs to 5Sm causes a "narrow-to-wide" conformational change of the latter, which decreases the binding affinity of Gemin2 to 5Sm and causes Gemin2 to dissociate from the Sm subcore. Therefore, in addition to the previously identified role of binding 5Sm, Gemin2 serves as a negative allosteric modulator in Sm core assembly, coupling RNA assembly specificity with Gemin2's release.

Structural basis for Gemin2's negative allosteric modulation

Building an RNA Sm site model in the narrow conformation of 5Sm in 7SΔN39 reveals why the Sm-site RNA alone cannot assemble into the RNA-binding pocket. In contrast to a circular shape inside the mature Sm cores [37], the Sm-site RNA is elliptical with two bases (Ura4 and Ura5) bulging out at the SmD1-SmG opening (Fig. 5a). The evenly distributed negative charges on the phosphate backbone of RNA are constrained into such a narrow and unbalanced conformation that the conformation of RNA in 7S must be highly unstable (Fig. 5b); as a result the RNA tends to either dissociate from the 5Sm (causing no binding) or splay the ellipse into a circle (causing a narrow-to-wide switch of 5Sm) if extra binding energy is provided from other parts of the RNA to hold the RNA inside the RNA-binding pocket of 5Sm. The 5' side of RNA has little contact with 5Sm and therefore plays little role in RNA assembly with the exception of a 5'-adjacent SL, which may sterically interfere with the binding of the Sm site RNA into the central RNA-binding pocket of 5Sm. In contrast, the 3' side of RNA can form electrostatic interactions with many positively charged residues on the interior face of 5Sm and therefore is essential for RNA

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

binding to 7S (Fig. 5c and S7). A 3' SL can provide more electrostatic interactions with 5Sm and therefore is preferred for Sm core assembly to a single strand.

To better understand the structural mechanism of Gemin2's release upon 5Sm splay, we re-processed the coordinate of the previous 7S complex and obtained a quality-improved structure as indicated by reduced R and R_{free} values (from 25.7% and 33.1% to 22.4% and 29.7% respectively) (Table S1). There are two significant improvements in the interface between Gemin2 and 5Sm: (1) on the Gemin2's NTD-SmF/E surface, the last 2 residues of the 3-residue linker (residues 63-65) between α1 and β1 of Gemin2 has more hydrogen bonding interactions with SmF (Fig. S8, a-b). Overall, Gemin2's NTD is like two sticks, one being α1(residues 49-62) and the other being the linker's last 2 residues plus β1 (residues 64-69), connected by a short joint (residue 63). Each Sm protein of SmE/F provides a set of interacting parts, one helix and one strand to interact with each stick of Gemin2's NTD. (2) on the Gemin2's CTD-SmD1/D2 surface, a 3_{10}-helix is rebuilt on Gemin2's C-terminus (residues 270-280), which provides more interactions with SmD2 (Fig. S8, c-f). Overall, the interface of Gemin2's CTD to contact SmD1/D2 is highly rigid and can be viewed as a rock-solid surface.

In our previous analysis, we used full-length sequences of Sm proteins for superposition of 5Sms in 7S and U1 snRNP and suggested that the interfaces within each of the Sm sub-complexes, SmD1/D2 and SmF/E/G, are rigid, but the SmD2-SmF interface is widened in U1 snRNP [13]. However, this suggestion cannot explain Gemin2's release because the portion of Gemin2 connecting SmD2 and SmF is a flexible loop. To make a better analysis, in this study, 48 residues' main chain atoms of each Sm protein in the 7S complex and U4 snRNP core, which are from the conserved and less variable β sheet, are used for superposition and comparison of overall conformational change by root mean square deviation (RMSD) (Fig. S9a). While the RMSDs of the five Sm proteins are relatively small if superposed individually (0.52, 0.39, 0.48, 0.51 and 0.80 Å for the main chain atoms of D1, D2, F, E and G respectively), the RMSDs of their adjacent Sm proteins increase. For example, when D2 is superposed, the RMSDs of D1 and F are 0.54 and 1.64 Å respectively. When F is superposed, the RMSDs of D2 and

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

E are 1.82 and 1.87 Å respectively. And when E is superposed, the RMSDs of F and G are 1.69 and 0.83 Å respectively (Fig. S9b). This indicates that although there are overall increased conformational changes between all the neighboring Sm proteins, the conformational changes of D2-F and F-E are more substantial.

Superposition of SmD2 reveals that upon RNA binding there is a conformational shift of β1-Loop2-β2 and β3-Loop4-β4 of SmD1 toward the RNA 3'-SL, reducing the interactions between SmD1 and Gemin2's CTD (Fig. 6a-b). Superposition of SmF reveals that upon RNA binding the N-terminal helix of SmE moves toward SmF (i.e., the Cα-Cα distance between I18.SmE and G38.SmF reduces 2.3 Å) while the N-terminal helix of SmF moves away from the center of 5Sm (i.e., the Cα-Cα distance between N12 and G38 of SmF increases 1.7 Å) (Fig. 6c-e). These anisotropic movements of the helixes upon 5Sm's splay make the two "sticks" of Gemin2's NTD unable to interact with SmF and SmE simultaneously, therefore losing affinity between Gemin2's NTD and SmF/E (Fig. 6e). Upon RNA's assembling into and widening 5Sm, Gemin2's NTD and CTD lose affinity to SmF/E and SmD1/D2 respectively, causing Gemin2 to tend to dissociate. However, RNA's and Gemin2's binding to 5Sm are mutually inhibitory, and therefore even a cognate snRNP code-containing RNA is unable to move the assembly reaction to a completion. The joining of SmD3/B, however, stabilizing Sm subcore by forming a more stable Sm core, drives Sm core assembly to a completion and causes Gemin2 to release completely.

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DISCUSSION

In this study, we closely examined the assembly steps of the Sm core in the second phase, from the formation of SMN/Gemin2/5Sm, to the assembly of the Sm subcore, and finally to the completion of the Sm core by a combination of structural and biochemical approaches. We established the narrow state of 5Sm bound by Gemin2/SMN is real and discovered its physiological role. We identified Gemin2's second role in Sm core assembly in addition to being a binder of 5Sm—it serves as a negative allosteric modulator. By constraining 5Sm in a narrow conformation, Gemin2 helps 5Sm select RNA substrates, allowing preferably the cognate snRNAs, containing the snRNP code, to assemble; the assembly of RNA into the Sm subcore widens 5Sm, causing Gemin2's release and allowing SmD3/B to join. Our proposed mechanism is schematically drawn in Figure 7, the structural basis in Figures 5-6, and the energetic changes in the pathway in Fig. S10. This mechanism simultaneously provides answers to the two significant questions, how the SMN complex confers RNA assembly specificity, and how the SMN complex dissociates from the assembled Sm core. These findings cause a paradigm shift in our understanding of the mode of action of the SMN complex and snRNP assembly.

Since Sm core assembly in vitro is a spontaneous process, why eukaryotes evolve so many assembly chaperons in this process is a central question. Early studies have proved that these chaperons, especially the SMN complex, help the assembly in an exclusively specific way [8], and require RNA substrates containing both the nonameric Sm site and adjacent 3'-SL[9]. People have been long probing the assembly specificity mechanism, however, which of the chaperons plays the major role to confer RNA assembly specificity and how it does have not been solved. pICln binds 5Sm in a closed ring to form the 6S complex, preventing premature and illicit assembly. However, it prevents any RNAs including the cognate ones to bind [12, 14]. Similarly, although the N-tail of Gemin2, which we observed inside the RNA-binding pocket of 5Sm in our previous study, plays an inhibitory role in RNA-binding, it inhibits the binding of both correct and illicit RNAs [13]. Therefore, both pICln and Gemin2's N-tail are unable to serve as the specificity factor. Gemin5 of the SMN complex has long been considered as the

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

specific factor and to play the role by direct binding to the snRNP code [26-31]. This model is currently the dominating mechanism. Although Gemin5 is the first component of the SMN complex to bind precursor snRNAs and deliver them to the rest part of the SMN complex in vertebrates, this model has several drawbacks. First, it cannot explain why snRNPs assemble still efficiently in many low eukaryotes where no ortholog of Gemin5 is found [15, 32]. Second, recent structural and biochemical studies of Gemin5-RNA interactions also provided evidence against this model: (1) only part of the Sm site, AUUU, much less than the snRNP code, is critical for RNA binding to Gemin5 [29-31]; (2) Gemin5 can bind promiscuous RNAs, i.e., U1-tfs, the truncated U1 pre-snRNAs lacking the Sm site and the following SL [29]. In contrast, our finding that Gemin2 achieves the high specificity of snRNP core assembly by a negative allosteric mechanism can explain all these puzzles. First, Gemin2 is the most conserved component of the complex, from human to yeast [15, 45]. Structure-guided sequence alignment of Gemin2 homologs in various species indicates that they all have the conserved F/E binding domain and D1/D2 binding domain, therefore would bind to 5Sm in the same way as does the human Gemin2 [13]. All eukaryotes survive healthily without bothering of illicit RNAs assembly is because they all have Gemin2. Second, the minimal required RNA feature by the 7S complex is a Sm site plus a 3'-RNA, preferably a 3'-SL, matching the snRNP code identified previously [9], which is much more than the 4 nucleotides, AUUU, for binding to Gemin5. Third, our model can explain the previous in vivo and in vitro results. In 2002, using cell extracts and pull-down assays, the SMN complex was shown to assemble the major spliceosomal U-rich snRNPs, while total proteins (containing the 7 Sm proteins) also assembled other types of RNAs [8]. The test condition was the purified SMN complex containing Sm proteins. With hindsight, we can readily deduce that the 5 Sm proteins were already bound by Gemin2, in the narrow, constrained state. Finally, RNAi-mediated knockdown experiments also support our conclusion. Knockdown of Gemin2 disrupted Sm core assembly whereas knockdown of Gemin5 had little effect on Sm core assembly [46, 47]. In summary, it is the combination of 5Sm and Gemin2, instead of Gemin5, that determines RNA specificity and recognizes the snRNP code. As Gemin5 is the first protein of the SMN complex in more complexed eukaryotes to bind snRNAs, it may play a role of preliminary

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

screening of RNAs. It is interesting that under our *in vitro* experimental condition RNAs containing the snRNP code were merely about two-fold more efficient in Sm core assembly than RNAs just containing the Sm site in the middle (Fig. 4c). However, this efficiency is consistent with the previous *in vivo* experimental result, in which a radioactivity-labeled RNA containing the Sm site plus a 3' single strand was observed to assemble into the Sm core by 50% efficiency after microinjected into the cytoplasm of Xenopus oocytes and pulled-down by anti-Sm antibody [9]. This consistency further supports that our findings reflect the *in vivo* situation. As for how to understand that the SMN complex dominantly assemble the Sm core on snRNAs instead of non-cognate RNAs, there may be two additional contributing factors: (1) high abundance of snRNAs in cells compared with other non-cognate RNAs which containing just the Sm site [48, 49]; (2) in other non-cognate RNAs containing only the Sm site, the single strand 3' to the Sm site might be a binding site of other proteins, which would block the Sm core assembly.

The mechanism of the SMN complex's dissociation from the Sm core has long been a black box. Our findings provide the first model on it. Because Gemin2 is the major protein to bind Sm proteins [13] and therefore, we reason that how the SMN complex is released depends mostly on how Gemin2 is released. We had hypothesized that Gemin2 were released by a cleavage at the connecting loop between its NTD and CTD inside cells. But we found that Gemin2 is entirely intact in Hela cells when we tagged Gemin2 at both its N- and C-termini in Hela cells and checked its tags by Western blot (Data not shown). How Gemin2 can dissociate intact from the Sm core had puzzled us for a while. In this study, we found that Gemin2 is released upon cognate RNAs' assembly into 5Sm and completely upon Sm-core formation. In Hela cells, most of the SMN complex components, including Gemin2, were observed also in the nucleus, mostly in CBs [26]. How to explain this apparent disparity? First of all, we notice that the concentrations of most of the SMN complex proteins are higher in the cytoplasm than in the nucleus [26]. This tells that most of the SMN complex dissociate from the Sm core and recycle in the cytoplasm, which is consistent with our proposed model. In addition, we also know that the SMN protein contains a tudor domain, which is able to bind the methylated RG-rich tails of SmD1, D3 and B [50]. It is likely that small

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

portion of the SMN complex remain tethered to Sm cores through these interactions and follows Sm cores into the nucleus. In the CBs, the methylated arginine residues in the CB-hallmark protein coilin can bind SMN [51]. It is possible that coilin competes with the methylated RG-rich tails of SmD1, D3 and B, releasing Sm cores from the SMN complex completely.

Because of the highest conservation of Gemin2 among the SMN complex in eukaryotes, we propose that this negative allosteric mechanism mediated by Gemin2 is a fundamental mechanism of Sm-core assembly in all eukaryotes. The fact that only orthologue of Gemin2, Brr1, is found to play an important role in Sm-core assembly in the simplest eukaryote, S. cerevisiae supports this idea [15, 52], and the Gemin2 cycle mechanism might explain all the mechanism in this simplest eukaryote. This finding also provides further clues for the evolution of the Sm-core assembly chaperon system. This work would facilitate further mechanistic study of other components of the SMN complex in snRNP assembly in high eukaryotes. For example, the structure of Gemin6/7 resembles a Sm heterodimer and it has been suggested that Gemin6/7 serve as a surrogate to bind to 5Sm in the position of SmD3/B [53]. Our study makes this suggestion less likely probable because the narrow 5Sm disables any Sm-fold dimer to join, and once a cognate RNA binds to 5Sm and widens 5Sm, SmD3/B can readily join to finish the assembly efficiently. There is no reason for Gemin6/7 to bind first. So Gemin6/7 may play a different role, which awaits further investigation.

In addition, we predict that this mechanism likewise apply to the assembly of U7 snRNP core, which has a variation of Sm protein components and Sm site but is assembled by the SMN complex and requires a 3'-SL adjacent to the Sm site [54-56].

Furthermore, our finding also facilitates pathogenesis study of SMA and development of therapeutics. The demonstration of Gemin2 offering the basic mechanism of Sm-core assembly provides the possibility to develop strategies to assemble Sm cores without the SMN protein. These strategies

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

could be used to test if SMA is totally caused by deficiency of Sm-core assembly and to develop possible therapeutics targeted on Sm-core assembly.

The major method used in this study to characterize protein-RNA interaction is GFC. It is an approach commonly used in protein-protein interactions, yet much rarely used in protein-RNA interaction studies compared with EMSA. However, it is superior to EMSA when multiple-component systems are studied, as illustrated here by discovering the release of Gemin2 from Sm subcores, which escaped from many previous studies by EMSA [8, 12, 14]. So GFC is generally applicable for other protein-RNA and protein-DNA interactions, especially studying nucleotides interaction with multiple-protein components.

The interaction mechanism of Gemin2-mediated snRNP core assembly is unique in protein-RNA interactions. Unlike many other RNA-binding proteins, which enhance RNA-binding specificity by combining domains recognizing RNA directly [57, 58], Gemin2-mediated snRNP core assembly is a combination of both direct RNA recognition (sequence-dependent and sequence-independent) and indirect, allosteric effect. To our knowledge, this is the first report of negative allosteric mechanism in protein-RNA interaction.

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

MATERIALS AND METHODS

Plasmid Construction and Protein Expression and Purification

All of the plasmids used in the studies contain human complementary DNAs (cDNAs). Full-length SmD1 and SmD2 (pCDFDuet-HT-D2-D1), full-length SmF and SmE (pCDFDuet-HT-F-E), full-length SmG (pET28-HT-G) and full-length Gemin2 (pCDF-HT-Gemin2) were constructed as described before [13]. SmD1s (residues 1-82) and SmD2 (pCDFDuet-HT-D2-D1s) were constructed by replacing the full-length D1 with SmD1s in pCDFDuet-HT-D2-D1. The Sm fold portion of SmD3(residues 1-75) and SmB(residues 1-91) [pCDFDuet-HT-B(1-91)-D3(1-75)] were constructed in a single pCDFDuet vector (Novagen) with N-terminal His(6)-tag followed by Tobacco Etch Virus (TEV) cleavage site (HT) fused to SmB. Gemin2ΔN39 (pCDF-HT-Gemin2ΔN39) were constructed by deletion of the N-terminal 39 residues in pCDF-HT-Gemin2. SMN_{Ge2BD}, containing SMN residues 26–62 (pET21-HMT-SMN_{Ge2BD}), was fused with an N-terminal His(6)-tag followed by maltose binding protein (MBP) tag and TEV cleavage site in pET21 vector (Novagen).

SmD1/D2 (or SmD1s/D2) was purified by Ni-column first, followed by TEV protease cleavage, secondary pass of Ni-column, cation exchange, and gel filtration chromatography. SmF/E was purified by a similar procedure except that anion exchange was used instead. SmF/E and SmG were coexpressed and purified in the same way as SmF/E. Gemin2 and SMN_{Ge2BD} were coexpressed and purified by Ni-column first, followed by TEV protease cleavage, Ni-column, and anion exchange chromatography.

To make the heptamer of the Gemin2 (or Gemin2ΔN39)-SMN_{Ge2BD}-5Sm complex, equal molar amount of the SmD1s/D2, SmF/E/G, and Gemin2(or Gemin2ΔN39)/SMN_{Ge2BD} complexes were mixed in gel filtration buffer (20 mM Tris-HCl [pH 8.0], 150 mM NaCl, 1 mM EDTA, and 1 mM TCEP [tris(2-carboxyethyl) phosphine]) supplemented with 0.5 M NaCl, and subjected to superdex200 GFC (HiLoad 16/600 or Increase 10/300 GL, GE Healthcare Bio-Sciences, Sweden). The fractions containing all seven components were checked by SDS-PAGE, pooled and concentrated to 7–12 mg/ml, and used for

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

crystallization studies. To make the hexamer of the Gemin2ΔN39-SMN_{Ge2BD}-D1s/D2/F/E complex, equal molar amount of the SmD1s/D2, SmF/E, and Gemin2ΔN39/SMN_{Ge2BD} complexes were mixed in the same gel filtration buffer as above, and subjected to superdex200 GFC. The fractions containing all six components were checked by SDS-PAGE, pooled and concentrated to 4-5 mg/ml, and used for crystallization studies.

Crystallization, Data Collection and Structure Determination

Human Gemin2-SMN_{Ge2BD}-D1s/D2/F/E/G complex (Complex A) crystals were grown in 6% PEG8000, 100 mM Tris-HCl (pH 7.5–8.2), human Gemin2ΔN39-SMN_{Ge2BD}-D1s/D2/F/E/G complex (Complex B) crystals were grown in 1% PEG8000, 100 mM Tris-HCl (pH 7.5–8.2), and human Gemin2ΔN39-SMN_{Ge2BD}-D1s/D2/F/E complex (Complex C) crystals were grown in 4% PEG8000, 100 mM Tris-HCl (pH 7.5–8.2). They were all grown by hanging-drop vapor diffusion method at 20°C within a couple of days. They all form in space group $P2_12_12_1$, but with various unit cell parameters (Table S1). The crystals were cryoprotected by gradual transfer from reservoir solution containing 10% to 40% PEG400, and frozen in liquid nitrogen. The X-ray diffraction data sets of these complex crystals were collected at beamlines BL17U1 and BL19U1 at the National Facility for Protein Science (NFPS) and Shanghai Synchrotron Radiation Facility (Shanghai, China) at wavelengths of 0.97853 and 0.97846 Å. Data were processed by HKL2000 [59]. Since the diffraction of the crystals was severely anisotropic, the data sets were reprocessed and truncated ellipsoidally by anisoscaling[60]. The structures were solved by molecular replacement with the 2.5 Å crystal structure (PDB code 3S6N) as the search model by PHASER [61] from CCP4 suite [62]. The models were improved by cycles of manual rebuilding in Coot [63] and REFMAC refinement [64]. The final data collection and refinement statistics are summarized in Table S1. The coordinates and structural factors of the three complexes, A-C, have been deposited in the Protein Data Bank under ID codes 5XJQ, 5XJR and 5XJS. The previous crystal structure of the 7S complex (PDB code 3S6N) was re-refined with reference to two related complex structures accessible in recent years, the NMR structure of the complex Gemin2 (residues 95-280)/SMN (residues 26-51) (PDB

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

code 2LEH) and the 8S complex (PDB code 4V98), containing human SmD1/D2/F/E/G and Drosophila melanogaster pICln (residues 1-180), SMN (residues 1-122) and full-length Gemin2 (residues 1-245) [14, 65] by cycles of manual rebuilding in Coot [63] and REFMAC refinement [64] in CCP4 suite [62]. The final structure has improved quality as indicated by reduced values of R and R_{free} from previous 25.7% and 33.1% to 22.4% and 29.7% respectively (Table S1) and the new coordinate has been deposited (PDB code 5XJL).

Building of the Sm site RNA model in the 7S complex

The first 7 nucleotides of the Sm site in U4 snRNP (PDB code 4WZJ) were individually saved together with their interacting Sm proteins. Each of the coordinates was then aligned with its corresponding Sm protein in the 7S complex. The 7 nucleotides were linked in Coot [63] and followed by a relaxing of conformational constrains.

In vitro RNA production and purification

With the exception of the nanomeric Sm site, AAUUUUUGA, which was chemical synthesized by Takara, all RNAs, including U4, flU4 and their derivatives (Their sequences and predicted secondary structures are in Table S2 and Fig. S11) were produced by in vitro transcription using MEGAscript kit (Ambion). The templates were made by either annealing of two complementary primers or PCR. Transcribed RNAs were separated by urea-PAGE and the gel containing the RNAs was cut and collected. RNAs were purified by phenol-chloroform extraction, followed by precipitation using ethanol. After spin-vacuum dry, the purified RNAs were dissolved in buffer containing 20mM Tris-HCl, 250 mM KCl, 2mM $MgCl_2$, pH7.5.

In vitro RNA binding, electrophoresis mobility shift assay

Binding of 7S or 7SΔN39 complex to U4 or U4ΔSm RNA was performed in buffer containing 20mM Tris-HCl (pH 8.0), 250mM NaCl, 2mM $MgCl_2$, 1 mM EDTA, and 1mM DTT. Various amounts of

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

reconstituted 7S or 7SΔN39 complex (2.5, 5, and 10 pM of each) were incubated with 50 pM of U4 or U4ΔSm RNA at 37°C for 40 min. After that, 1/10 (v/v) glycerol was added to the reaction mixture and the RNPs were analyzed by 0.4% native agarose gel electrophoresis. RNA was visualized by SYBR green (Thermo Fisher Scientific).

In vitro RNA-Protein complex assembly assay

RNA-protein complex assembly assays were performed by incubating 5Sm or 7SΔN39 with various RNAs in final volume of 500µl in assembly buffer containing 20mM Tris-HCl (pH 7.5), 250mM NaCl, 2mM $MgCl_2$, 1 mM EDTA, and 1mM DTT, with their amounts described in detail in Table S3 (control proteins or RNAs followed the same procedure). RNAs were pre-incubated at 65°C for 10 min followed by cool-down in room temperature before mixing with proteins. After incubation at 37°C for 40 min, the samples were spin down at 15,000 rpm for 5 min in a table centrifuge and applied into superdex200 Increase 10/300 GL GFC via a 500µl sample loop. The elution fractions were collected each 0.5 ml, resolved by SDS-PAGE directly (visualized by silver staining) or after concentration to 50µl (visualized by CBB staining). Their GFC positions are summarized in Table S4.

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ACKNOWLEDGEMENTS

We thank Gideon Dreyfuss and his group members at University of Pennsylvania for reading the manuscript and providing comments. We thank the staff of the beamlines BL17U1 and BL19U1 at the National Facility for Protein Science (NFPS) and Shanghai Synchrotron Radiation Facility, Shanghai, People's Republic of China, for assistance during crystal diffraction data collection. This work was supported by National Key R&D programs (No. 2017YFA0504300 and 2017YFA0505800) and National Natural Science Foundation of China (No.31570720 and 81441109). Coordinates and structural factors have been deposited in the Protein Data Bank with the accession codes 5XJQ, 5XJR, 5XJS, and 5XJL.

AUTHOR CONTRIBUTIONS

H. Yi crystallized the complexes and performed the biochemical assays. L. Mu, C. Shen, X. Kong, Y. Wang and Y. Hou participated in the project. R. Zhang conceived, designed and supervised the project, solved the crystal structures and wrote the paper.

CONFLICTS OF INTEREST

The authors declare no competing financial interest.

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

REFERENCES

1. Will, C.L. and R. Luhrmann, *Spliceosome structure and function.* Cold Spring Harb Perspect Biol, 2011. **3**(7).
2. Matera, A.G. and Z. Wang, *A day in the life of the spliceosome.* Nat Rev Mol Cell Biol, 2014. **15**(2): p. 108-21.
3. Li, D.K., et al., *SMN control of RNP assembly: from post-transcriptional gene regulation to motor neuron disease.* Semin Cell Dev Biol, 2014. **32**: p. 22-9.
4. Martin, W. and E.V. Koonin, *Introns and the origin of nucleus-cytosol compartmentalization.* Nature, 2006. **440**(7080): p. 41-5.
5. Raker, V.A., G. Plessel, and R. Luhrmann, *The snRNP core assembly pathway: identification of stable core protein heteromeric complexes and an snRNP subcore particle in vitro.* EMBO J, 1996. **15**(9): p. 2256-69.
6. Raker, V.A., et al., *Spliceosomal U snRNP core assembly: Sm proteins assemble onto an Sm site RNA nonanucleotide in a specific and thermodynamically stable manner.* Mol Cell Biol, 1999. **19**(10): p. 6554-65.
7. Chari, A., E. Paknia, and U. Fischer, *The role of RNP biogenesis in spinal muscular atrophy.* Curr Opin Cell Biol, 2009. **21**(3): p. 387-93.
8. Pellizzoni, L., J. Yong, and G. Dreyfuss, *Essential role for the SMN complex in the specificity of snRNP assembly.* Science, 2002. **298**(5599): p. 1775-9.
9. Golembe, T.J., J. Yong, and G. Dreyfuss, *Specific sequence features, recognized by the SMN complex, identify snRNAs and determine their fate as snRNPs.* Mol Cell Biol, 2005. **25**(24): p. 10989-1004.
10. Friesen, W.J., et al., *The methylosome, a 20S complex containing JBP1 and pICln, produces dimethylarginine-modified Sm proteins.* Mol Cell Biol, 2001. **21**(24): p. 8289-300.
11. Brahms, H., et al., *Symmetrical dimethylation of arginine residues in spliceosomal Sm protein B/B' and the Sm-like protein LSm4, and their interaction with the SMN protein.* RNA, 2001. **7**(11): p. 1531-42.
12. Chari, A., et al., *An assembly chaperone collaborates with the SMN complex to generate spliceosomal SnRNPs.* Cell, 2008. **135**(3): p. 497-509.
13. Zhang, R., et al., *Structure of a key intermediate of the SMN complex reveals Gemin2's crucial function in snRNP assembly.* Cell, 2011. **146**(3): p. 384-95.

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

14. Grimm, C., et al., *Structural basis of assembly chaperone- mediated snRNP formation.* Mol Cell, 2013. **49**(4): p. 692-703.

15. Kroiss, M., et al., *Evolution of an RNP assembly system: a minimal SMN complex facilitates formation of UsnRNPs in Drosophila melanogaster.* Proc Natl Acad Sci U S A, 2008. **105**(29): p. 10045-50.

16. Schrank, B., et al., *Inactivation of the survival motor neuron gene, a candidate gene for human spinal muscular atrophy, leads to massive cell death in early mouse embryos.* Proc Natl Acad Sci U S A, 1997. **94**(18): p. 9920-5.

17. Jablonka, S., et al., *Gene targeting of Gemin2 in mice reveals a correlation between defects in the biogenesis of U snRNPs and motoneuron cell death.* Proc Natl Acad Sci U S A, 2002. **99**(15): p. 10126-31.

18. Lefebvre, S., et al., *Identification and characterization of a spinal muscular atrophy-determining gene.* Cell, 1995. **80**(1): p. 155-65.

19. Fischer, U., Q. Liu, and G. Dreyfuss, *The SMN-SIP1 complex has an essential role in spliceosomal snRNP biogenesis.* Cell, 1997. **90**(6): p. 1023-9.

20. Burghes, A.H. and C.E. Beattie, *Spinal muscular atrophy: why do low levels of survival motor neuron protein make motor neurons sick?* Nat Rev Neurosci, 2009. **10**(8): p. 597-609.

21. Otter, S., et al., *A comprehensive interaction map of the human survival of motor neuron (SMN) complex.* J Biol Chem, 2007. **282**(8): p. 5825-33.

22. Carissimi, C., et al., *Gemin8 is required for the architecture and function of the survival motor neuron complex.* J Biol Chem, 2006. **281**(48): p. 37009-16.

23. Carissimi, C., et al., *Gemin8 is a novel component of the survival motor neuron complex and functions in small nuclear ribonucleoprotein assembly.* J Biol Chem, 2006. **281**(12): p. 8126-34.

24. Charroux, B., et al., *Gemin3: A novel DEAD box protein that interacts with SMN, the spinal muscular atrophy gene product, and is a component of gems.* J Cell Biol, 1999. **147**(6): p. 1181-94.

25. Charroux, B., et al., *Gemin4. A novel component of the SMN complex that is found in both gems and nucleoli.* J Cell Biol, 2000. **148**(6): p. 1177-86.

26. Yong, J., et al., *Gemin5 delivers snRNA precursors to the SMN complex for snRNP biogenesis.* Mol Cell, 2010. **38**(4): p. 551-62.

27. Battle, D.J., et al., *The Gemin5 protein of the SMN complex identifies snRNAs.* Mol Cell, 2006. **23**(2): p. 273-9.

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

28. Lau, C.K., J.L. Bachorik, and G. Dreyfuss, *Gemin5-snRNA interaction reveals an RNA binding function for WD repeat domains.* Nat Struct Mol Biol, 2009. **16**(5): p. 486-91.

29. Xu, C., et al., *Structural insights into Gemin5-guided selection of pre-snRNAs for snRNP assembly.* Genes Dev, 2016. **30**(21): p. 2376-2390.

30. Tang, X., et al., *Structural basis for specific recognition of pre-snRNA by Gemin5.* Cell Res, 2016. **26**(12): p. 1353-1356.

31. Jin, W., et al., *Structural basis for snRNA recognition by the double-WD40 repeat domain of Gemin5.* Genes Dev, 2016. **30**(21): p. 2391-2403.

32. Wahl, M.C. and U. Fischer, *The right pick: structural basis of snRNA selection by Gemin5.* Genes Dev, 2016. **30**(21): p. 2341-2344.

33. Pomeranz Krummel, D.A., et al., *Crystal structure of human spliceosomal U1 snRNP at 5.5 A resolution.* Nature, 2009. **458**(7237): p. 475-80.

34. Weber, G., et al., *Functional organization of the Sm core in the crystal structure of human U1 snRNP.* EMBO J, 2010. **29**(24): p. 4172-84.

35. Leung, A.K., K. Nagai, and J. Li, *Structure of the spliceosomal U4 snRNP core domain and its implication for snRNP biogenesis.* Nature, 2011. **473**(7348): p. 536-9.

36. Kondo, Y., et al., *Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition.* Elife, 2015. **4**.

37. Li, J., et al., *Re-refinement of the spliceosomal U4 snRNP core-domain structure.* Acta Crystallogr D Struct Biol, 2016. **72**(Pt 1): p. 131-46.

38. Hegele, A., et al., *Dynamic protein-protein interaction wiring of the human spliceosome.* Mol Cell, 2012. **45**(4): p. 567-80.

39. Zhou, Z., et al., *Comprehensive proteomic analysis of the human spliceosome.* Nature, 2002. **419**(6903): p. 182-5.

40. Layten, M., V. Hornak, and C. Simmerling, *The open structure of a multi-drug-resistant HIV-1 protease is stabilized by crystal packing contacts.* J Am Chem Soc, 2006. **128**(41): p. 13360-1.

41. Mund, M., et al., *Structure of the LSm657 complex: an assembly intermediate of the LSm1-7 and LSm2-8 rings.* J Mol Biol, 2011. **414**(2): p. 165-76.

42. Naidoo, N., et al., *Crystal structure of Lsm3 octamer from Saccharomyces cerevisiae: implications for Lsm ring organisation and recruitment.* J Mol Biol, 2008. **377**(5): p. 1357-71.

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

43. Toro, I., et al., *Archaeal Sm proteins form heptameric and hexameric complexes: crystal structures of the Sm1 and Sm2 proteins from the hyperthermophile Archaeoglobus fulgidus.* J Mol Biol, 2002. **320**(1): p. 129-42.

44. Yong, J., et al., *snRNAs contain specific SMN-binding domains that are essential for snRNP assembly.* Mol Cell Biol, 2004. **24**(7): p. 2747-56.

45. Cauchi, R.J., *SMN and Gemins: 'we are family' ... or are we?: insights into the partnership between Gemins and the spinal muscular atrophy disease protein SMN.* Bioessays, 2010. **32**(12): p. 1077-89.

46. Shpargel, K.B. and A.G. Matera, *Gemin proteins are required for efficient assembly of Sm-class ribonucleoproteins.* Proc Natl Acad Sci U S A, 2005. **102**(48): p. 17372-7.

47. Feng, W., et al., *Gemins modulate the expression and activity of the SMN complex.* Hum Mol Genet, 2005. **14**(12): p. 1605-11.

48. Busch, H., et al., *SnRNAs, SnRNPs, and RNA processing.* Annu Rev Biochem, 1982. **51**: p. 617-54.

49. Cech, T.R. and J.A. Steitz, *The noncoding RNA revolution-trashing old rules to forge new ones.* Cell, 2014. **157**(1): p. 77-94.

50. Tripsianes, K., et al., *Structural basis for dimethylarginine recognition by the Tudor domains of human SMN and SPF30 proteins.* Nat Struct Mol Biol, 2011. **18**(12): p. 1414-20.

51. Boisvert, F.M., et al., *Symmetrical dimethylarginine methylation is required for the localization of SMN in Cajal bodies and pre-mRNA splicing.* J Cell Biol, 2002. **159**(6): p. 957-69.

52. Noble, S.M. and C. Guthrie, *Transcriptional pulse-chase analysis reveals a role for a novel snRNP-associated protein in the manufacture of spliceosomal snRNPs.* EMBO J, 1996. **15**(16): p. 4368-79.

53. Ma, Y., et al., *The Gemin6-Gemin7 heterodimer from the survival of motor neurons complex has an Sm protein-like structure.* Structure, 2005. **13**(6): p. 883-92.

54. Pillai, R.S., et al., *Unique Sm core structure of U7 snRNPs: assembly by a specialized SMN complex and the role of a new component, Lsm11, in histone RNA processing.* Genes Dev, 2003. **17**(18): p. 2321-33.

55. Kolev, N.G. and J.A. Steitz, *In vivo assembly of functional U7 snRNP requires RNA backbone flexibility within the Sm-binding site.* Nat Struct Mol Biol, 2006. **13**(4): p. 347-53.

56. Tisdale, S., et al., *SMN is essential for the biogenesis of U7 small nuclear ribonucleoprotein and 3'-end formation of histone mRNAs.* Cell Rep, 2013. **5**(5): p. 1187-95.

57. Ban, T., et al., *Structural mechanisms of RNA recognition: sequence-specific and non-specific RNA-binding proteins and the Cas9-RNA-DNA complex.* Cell Mol Life Sci, 2015. **72**(6): p. 1045-58.

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

58. Helder, S., et al., *Determinants of affinity and specificity in RNA-binding proteins.* Curr Opin Struct Biol, 2016. **38**: p. 83-91.

59. Otwinowski, Z. and W. Minor, *Processing of X-ray diffraction data collected in oscillation mode.* Methods Enzymol, 1997. **276**: p. 307-26.

60. Strong, M., et al., *Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis.* Proc Natl Acad Sci U S A, 2006. **103**(21): p. 8060-5.

61. McCoy, A.J., et al., *Phaser crystallographic software.* J Appl Crystallogr, 2007. **40**(Pt 4): p. 658-674.

62. Potterton, E., et al., *A graphical user interface to the CCP4 program suite.* Acta Crystallogr D Biol Crystallogr, 2003. **59**(Pt 7): p. 1131-7.

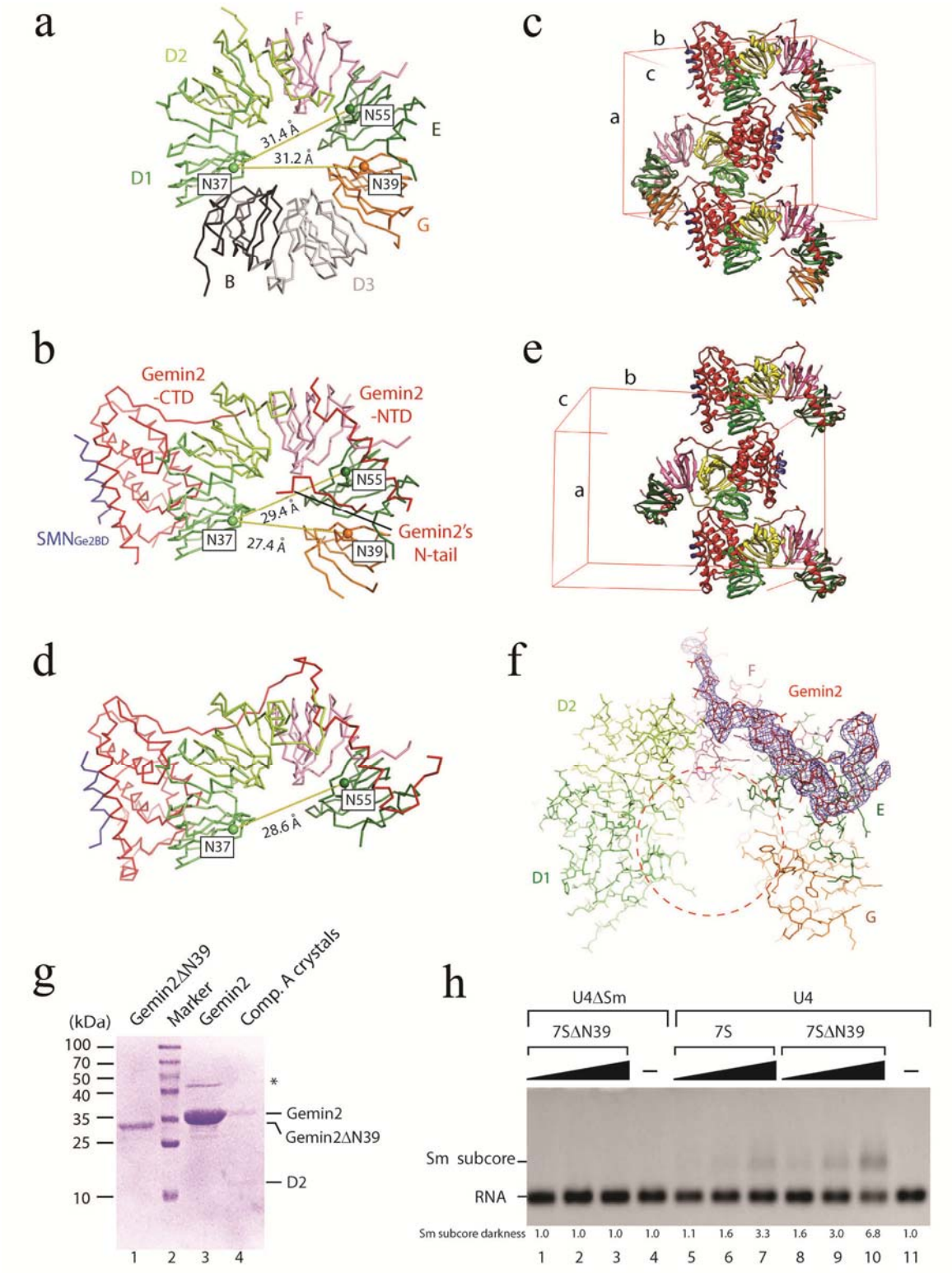
63. Emsley, P. and K. Cowtan, *Coot: model-building tools for molecular graphics.* Acta Crystallogr D Biol Crystallogr, 2004. **60**(Pt 12 Pt 1): p. 2126-32.

64. Murshudov, G.N., et al., *REFMAC5 for the refinement of macromolecular crystal structures.* Acta Crystallogr D Biol Crystallogr, 2011. **67**(Pt 4): p. 355-67.

65. Sarachan, K.L., et al., *Solution structure of the core SMN-Gemin2 complex.* Biochem J, 2012. **445**(3): p. 361-70.

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

FIGURES



bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 1. The narrow conformation of Sm heteropentamer bound by Gemin2/SMN$_{Ge2BD}$ is not caused by crystal packing as well as Gemin2's N-tail. (a) In the mature Sm core represented by U1 snRNP (PDB code 4PJO), the heteropentamer SmD1/D2/F/E/G is in a wide conformation, as indicated by the Cα-Cα distances between N37.SmD1 and N39.SmG, and between N37.SmD1 and N55.SmE. (b) In the previous 7S complex (PDB code 3S6N), the Sm heteropentamer is in a narrow conformation. (c) But in the crystal lattice, the opening between SmD1 and SmG of one 5Sm is occupied by Gemin2's CTD of a second complex, which may induce the narrowness of 5Sm. (d) In the structure of Complex C, which has no SmG and no Gemin2's N-tail, the conformation the Sm proteins is still narrow. (e) In the crystal lattice of Complex C, only SmD1 contacts Gemin2's CTD of a second complex. (f) Gemin2's N-tail is outside the RNA-binding pocket in the crystal structure of Complex A, in contrast to the previous 7S structure (PDB code 3S6N). Only Gemin2's N-terminus and 5Sm are shown. SigmaA-weighted 2Fo-Fc electron density maps (blue meshes) are contoured at 1.1σ of Gemin2 N-terminus (residues 1-76). Red circles indicate the RNA-binding pocket. The seven Sm proteins, D1, D2, F, E ,G, D3 and B, are colored in green, lemon, pink, dark green, orange, light grey and dark grey respectively. Gemin2 and SMN$_{Ge2BD}$ are colored in red and blue respectively. Unit cells for the previous 7S complex and Complex C are showed in (c) and (e). (g) The Gemin2 in Complex A crystals was intact. Crystals of complex A were picked up and subjected to SDS-PAGE and CBB staining. Full-length Gemin2 and Gemin2ΔN39 were used as controls. The asterisk indicates impurity. (h) Gemin2's N-tail accounts for only about 2-fold of U4 RNA inhibition. Three concentrations (2.5, 5 and 10 nM) of reconstituted 7S or 7SΔN39 complex were pre-incubated with U4 or U4ΔSm snRNA at 37°C for 40 min and subjected to electrophoresis mobility shift assay. The positions of the free RNA and 5Sm assembled on the RNA are indicated. The levels of Sm subcore were quantitated using imaging software and normalized. See also Fig. S1-2.

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

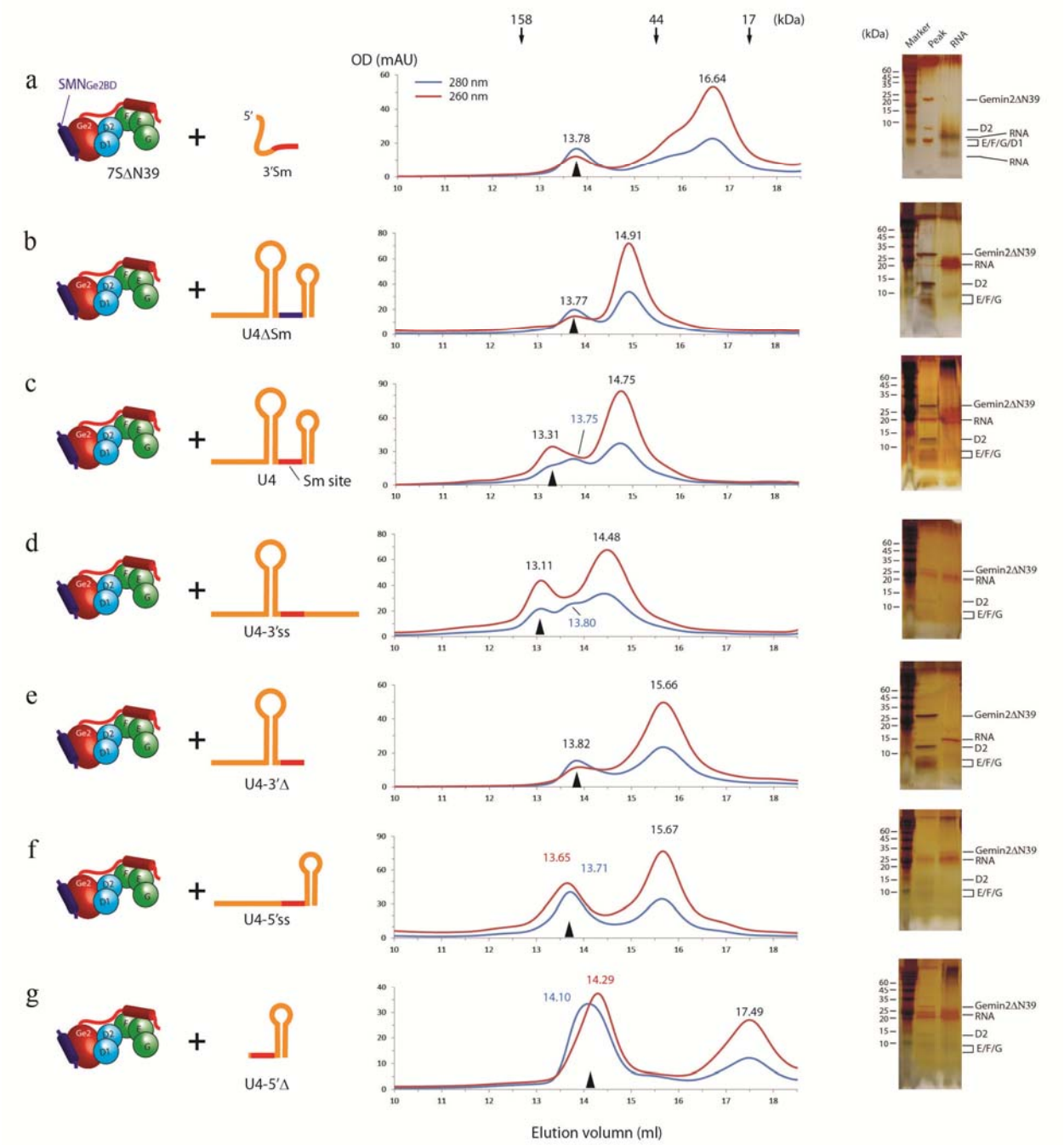


Figure 2. The 7SΔN39 complex selectively binds RNA with both the Sm site and a 3' RNA. 7SΔN39 was pre-incubated with various RNAs: 3'Sm (a), U4ΔSm (b), U4 (c), U4-3'ss (d), U4-3'Δ (e), U4-5'ss (f) or U4-5'Δ (g), and the mixture was subjected to gel filtration (middle panels). The front peak fractions and RNA inputs were analyzed by SDS-PAGE plus silver staining (right panels). The input components are showed in cartoon (left panels). See also Fig. S3-4.

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

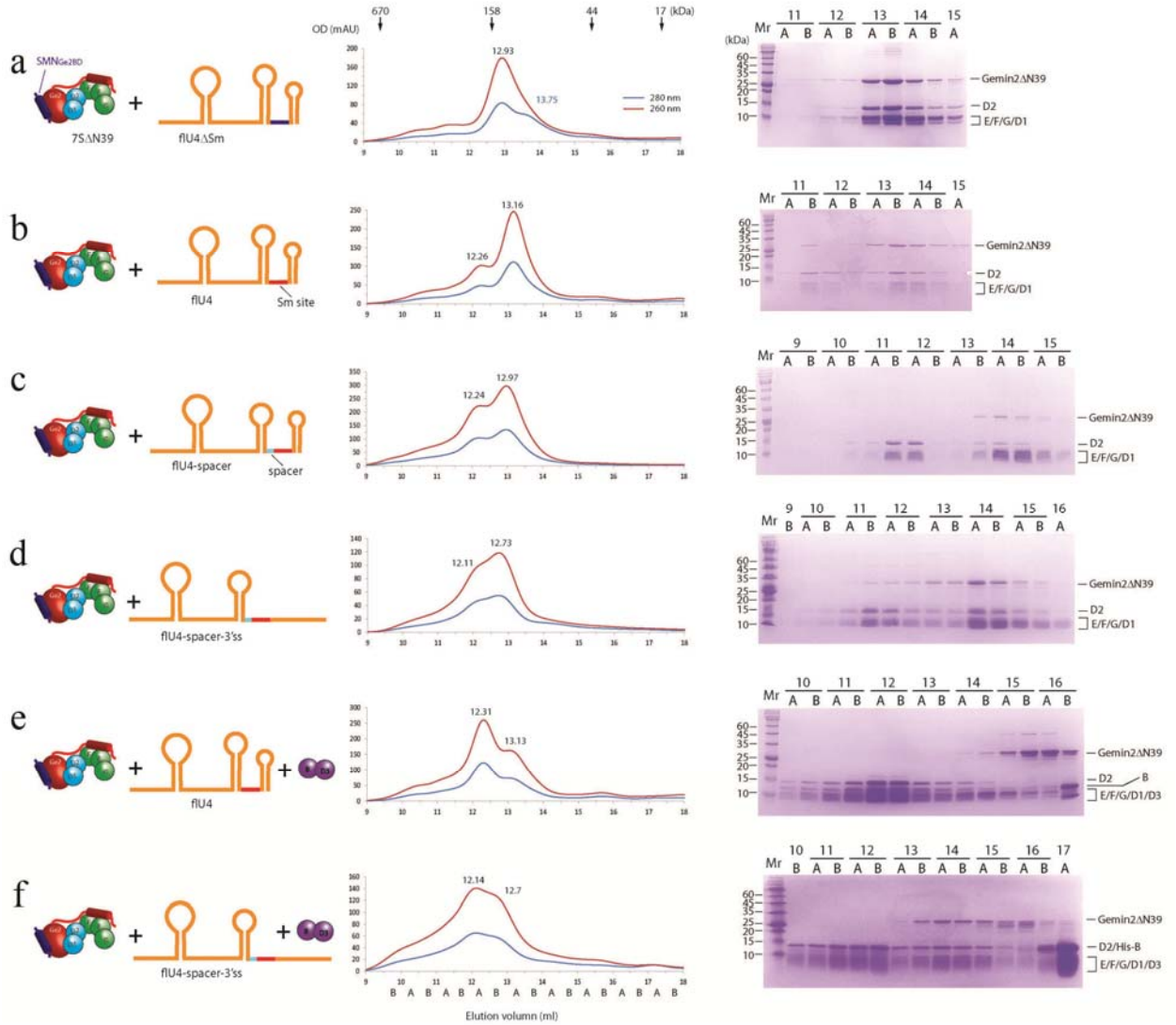


Figure 3. Release of Gemin2 during the assembly of Sm cores. 7SΔN39 was pre-incubated with various RNAs: flU4ΔSm (a), flU4 (b), flU4-spacer (c) or flU4-spacer-3'ss (d) for Sm subcore assembly, or with D3(1-75)/B(1-91) and flU4 (e) or flU4-spacer-3'ss (f) for Sm core assembly, and the mixture was subjected to gel filtration (middle panels). Individual fractions (based on elution position, A and B indicate the first and second half ml) were analyzed by SDS-PAGE plus CBB staining (right panels). The input components are showed in cartoon (left panels). See also Fig. S5.

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

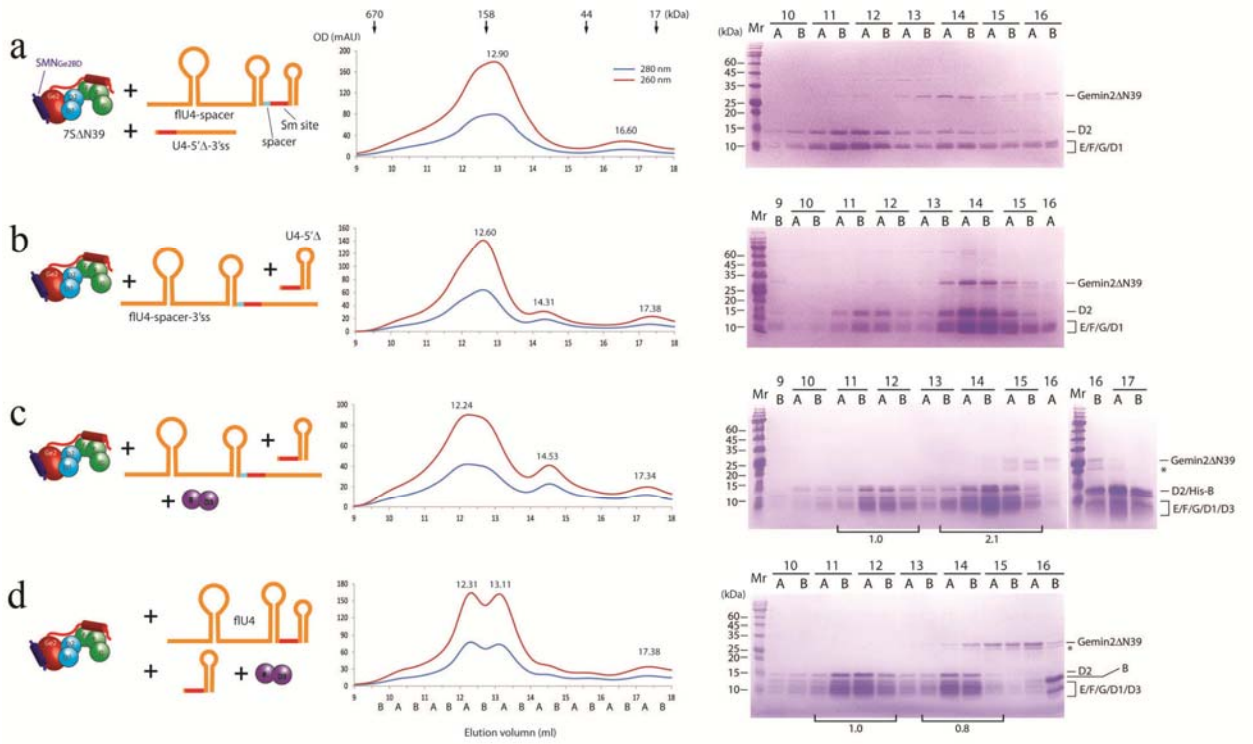


Figure 4. The snRNP code assembles into Sm core selectively. Sm subcore and core assembly assay and 7SΔN39 was pre-incubated with equal molar amount of flU4-spacer and U4-5'Δ-3'ss (a), or equal molar amount of flU4-spacer-3'ss and U4-5'Δ (b) for Sm subcore assembly. 7SΔN39 was pre-incubated with D3(1-75)/B(1-91) and equal molar amount of flU4-spacer-3'ss and U4-5'Δ (c) or equal molar amount of flU4 and U4-5'Δ (d) for Sm core assembly. The mixtures were subjected to gel filtration (middle panels). Individual fractions were analyzed by SDS-PAGE plus CBB staining (right panels). The input components are showed in cartoon (left panels). The levels of Sm core assembly were quantitated using imaging software and normalized to the front (c-d).

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

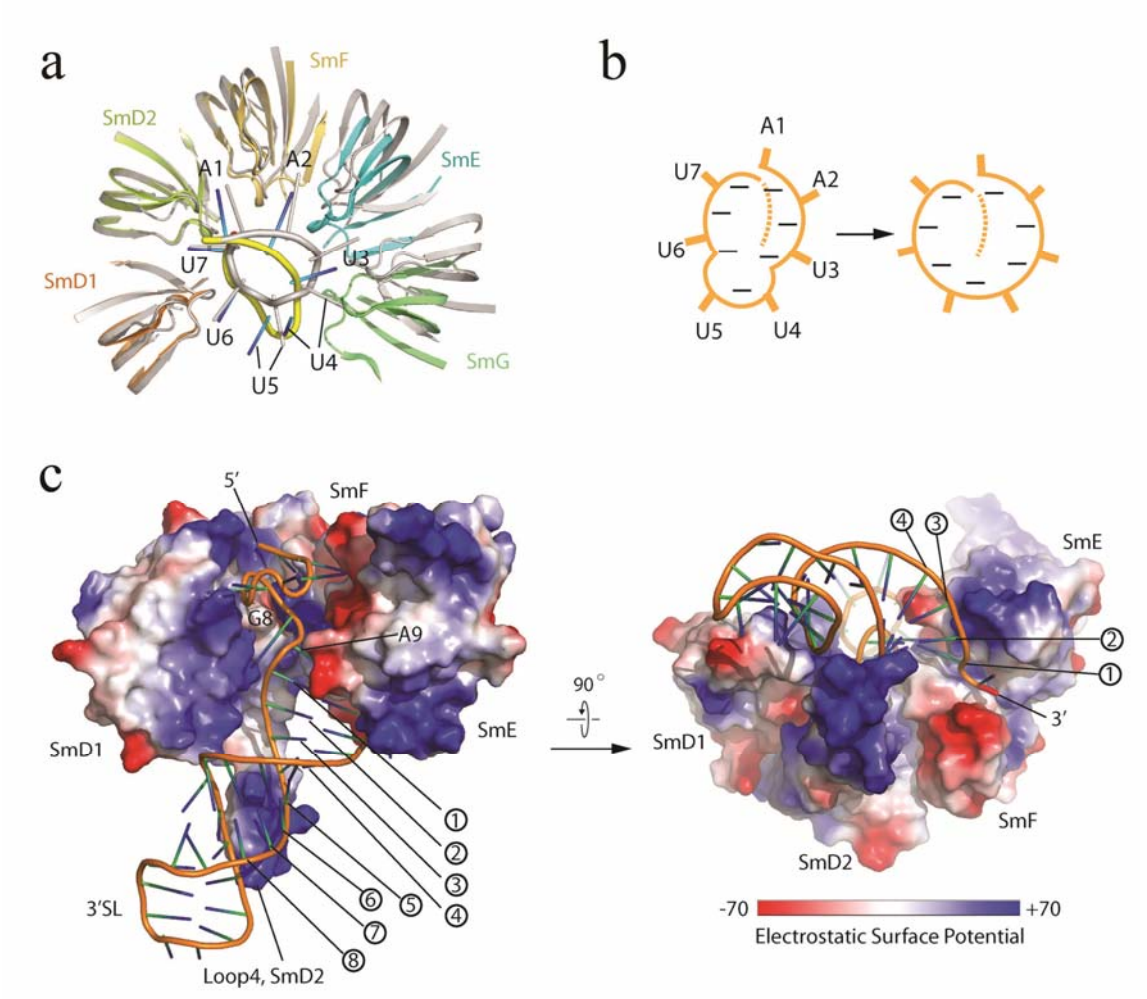
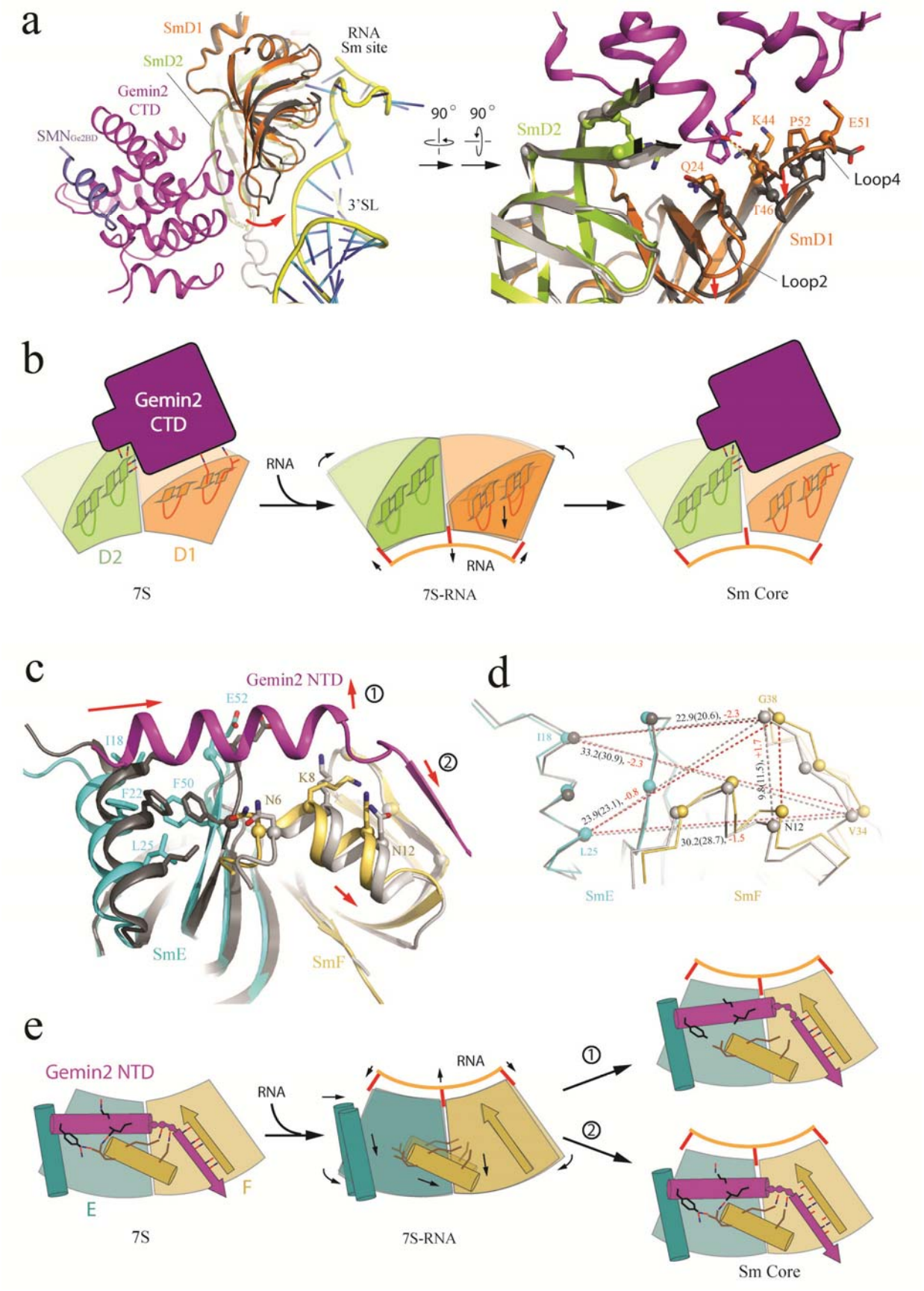


Figure 5. Structural basis for RNA selection caused by Gemin2's constraint on 5Sm. (a) The Sm site RNA model (the first seven nucleotides) manually built in the RNA-binding pocket in the 7S complex is an ellipse in contrast to a circle in the Sm core. SmD2s in the 7S complex (The five Sm proteins, D1, D2, F, E and G, are colored in orange, lemon, yellow, cyan and green respectively, Gemin2 and SMN_{Ge2BD} are omitted) and Sm core (in light grey) are superposed. The Sm site RNA model inside the 7S complex is showed in cartoon with its backbone in yellow and its bases in blue. (b) The elliptical Sm site RNA (inside 5Sm) tends to expand into a circle. (c) The 3'-adjacent SL of the Sm site provides more electrostatic interactions with the Sm proteins, with its downward strand contacting SmD1 and SmD2, and its upward strand contacting SmE, to enable a snRNP code-containing RNA to bind 5Sm in the 7S complex. See also Fig. S6-7.

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 6. Structural basis for Gemin2's release upon RNA assembly into 5Sm. (a) Reduced interactions between Gemin2's CTD and SmD1/D2 upon Sm core assembly. Superposition of SmD2s of the 7S complex and Sm core reveals a flipping of SmD1's loops 2 and 4 towards RNA's 3'-SL (red arrows), which causes Gemin2's CTD to lose most contacts with SmD1. (b) Schematic model of reducing interactions between Gemin2's CTD and SmD1/D2 upon Sm core assembly. (c) Reduced interactions between Gemin2's NTD and SmF/E upon Sm core assembly. SmFs in the 7S complex and Sm core are superposed. Gemin2's α1 would move towards SmF following the N-terminal helix of SmE as indicated by the red arrow. The second half of Gemin2's NTD would have two possible movements as indicated by red arrows labeled with ① and ②. (d) Conformational changes of SmF/E upon Sm core assembly. SmEs in the 7S complex and Sm core are superimposed. Cα-Cα distances (Å) in the 7S complex and Sm core are shown in black with the latter in brackets, and distance changes are shown in red. (e) Schematic model of reduced interactions between Gemin2's NTD and SmE/F. Hydrogen bonds are shown in red (H-acceptor) and blue (H-donor) bar pairs in schematic models (b) and (e). Gemin2 and SMN_{Ge2BD} are colored in purple and blue respectively. See also Fig. S8-9.

bioRxiv preprint doi: https://doi.org/10.1101/312124; this version posted May 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

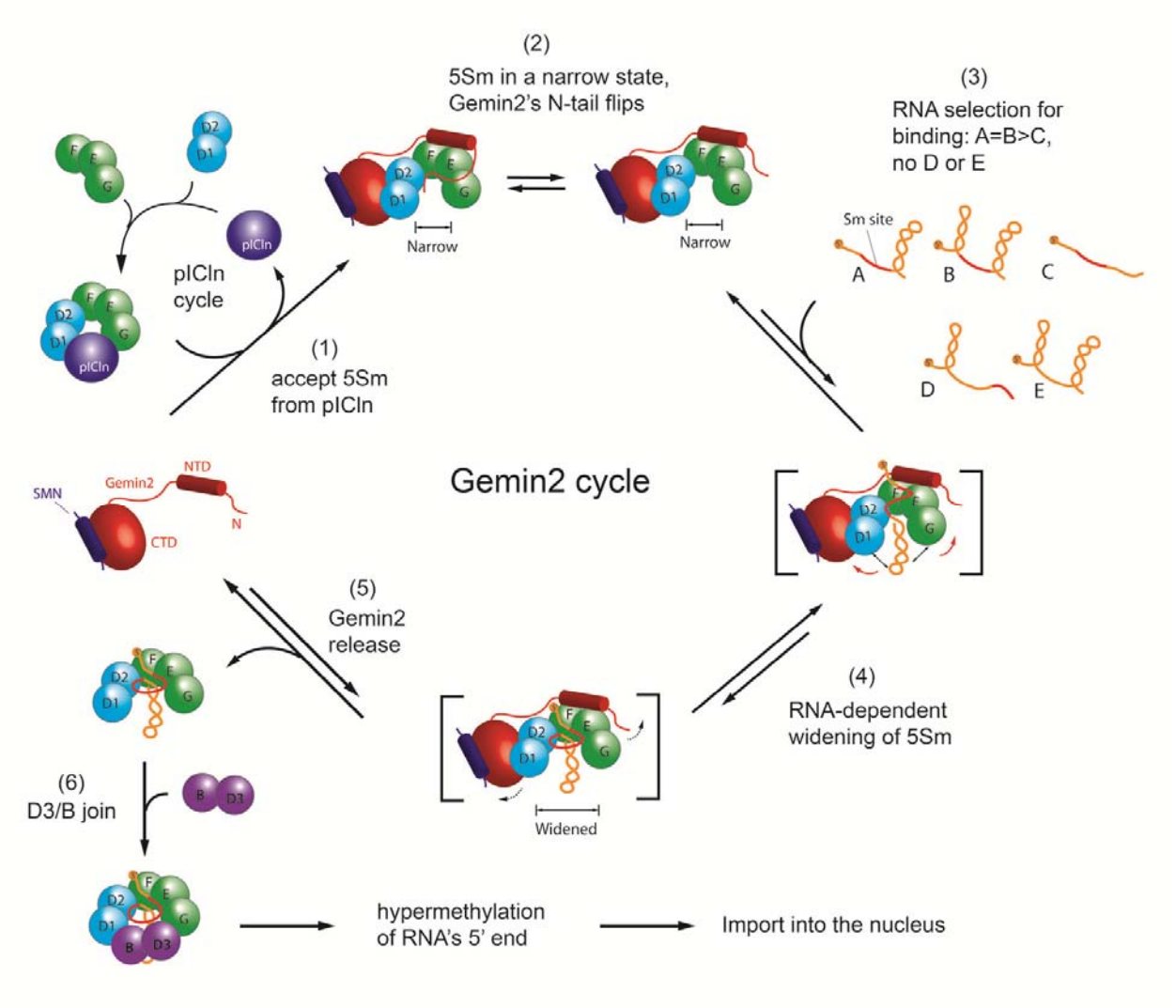


Figure 7. Schematic model of Sm core assembly. See also Fig. S10.