**Research Strategy**

**(A) Significance**

Cellular functions are often compartmentalized in membrane-delimited organelles. In contrast, numerous cellular "bodies" support essential functions, by concentrating components above surrounding plasm without membrane barriers. Recent evidence suggests that liquid phase separation by low complexity proteins observed primarily *in vitro* underlies the formation and maintenance of cellular bodies (1). How body formation occurs in a cellular environment crowded with many potentially interactive molecules is unclear. Moreover, cellular bodies tend to contain multiple low-complexity proteins. Although total protein concentration within bodies is similar to surrounding plasm (2), most bodies are highly enriched in RNA (3). Many RNA binding proteins bind short, often repeated RNA sequences, making RNA one of the most multivalent polymers in the cell. Low complexity domains and unstructured regions are often found in RNA binding proteins that drive liquid phase separation under normal conditions and protein aggregation in disease states, such as amyotrophic lateral sclerosis (ALS) (4). The RNAs bound by these proteins *in vivo* are largely unknown. What role does RNA play in the formation of these dynamic cellular structures *in vivo*? Does RNA determine the cellular position of bodies? Can RNA modulate normal and aberrant structures made by low complexity proteins?

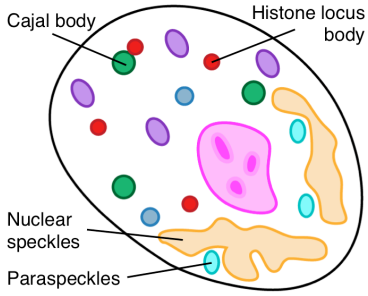



**Fig 1 Nuclear bodies organize gene expression in 3D space.**

In the cell nucleus, increasing evidence points to roles for bodies in genome function (1). Gene loci from distinct chromosomes cluster at or near nuclear bodies, raising the possibility that they determine chromosomal positioning and impact gene expression (5). Among many important nuclear bodies, four are critically important for mRNA biogenesis and are the focus of this proposal (**Fig 1**). Cajal bodies (CBs) mediate the efficient assembly of spliceosomal snRNPs, required for pre-mRNA splicing, and this is an essential function in vertebrate embryos (6-8); Histone Locus Bodies (HLBs) arise on active histone genes in S-phase and are involved in histone mRNA 3′ end processing (7, 9); nuclear speckles concentrate essential pre-mRNA splicing factors near highly expressed intron-containing genes (10); paraspeckles influence gene expression by retaining specific mRNAs in the nucleus and sequestering specific transcription factors (11, 12). Each of these bodies contains specific RNAs and trans-acting regulatory factors, which exchange rapidly with surrounding nucleoplasm (13). Nuclear bodies vary in number, size, morphology and composition during cell cycle, development, and upon experimental perturbation. Therefore, research on nuclear bodies *in vivo* has been mostly limited to interphase cells, leaving questions of their inheritance and potential function as epigenetic features unaddressed. How do nuclear bodies target to distinct chromosomal loci? How do different nuclear bodies remain distinct from one another, if they form by similar physical principles? We propose that RNA-protein interactions drive this specificity.

Our proposal addresses the hypothesis that RNA is central to the formation and function of nuclear bodies. We will use tissue culture cells and zebrafish embryos to image nuclear bodies and, in parallel, monitor the interactions of nuclear body-specific proteins with RNA and DNA at molecular resolution. We will:

- identify RNAs and chromosomal loci that interact with low complexity nuclear body proteins, using unbiased genome- and transcriptome-wide approaches.
- determine whether nascent RNA, produced by transcription of these loci, leads to nuclear body formation, gene clustering, and changes in gene expression.
- develop novel bioinformatics tools that exploit well-annotated genomic information to aid the detection, characterization and analysis of all cellular bodies.

The achievement of our aims will be significant, because we will discover RNAs and gene loci associated with nuclear bodies, shed new light on how RNA regulates nuclear body formation together with low complexity proteins, and determine how nuclear bodies impact nuclear organization during cell cycle and embryogenesis. The development of these new methodologies, which overcome current obstacles in the field, will enable further insights into the molecular mechanisms of nuclear body formation and function *in vivo*.

**(B) Innovation**

Most research on nuclear bodies has relied on imaging, which requires prior knowledge of morphology, components, and genomic sites of localization. One goal of this project is to define nuclear bodies through a set of specific molecular interactions among DNA, RNA and low complexity RNA binding proteins. This novel, alternative approach will identify new RNA components of nuclear bodies and the chromosomal sites where they assemble. The methods complement imaging and compensate for the fact that nuclear bodies are nearly

impossible to purify. Instead, we take advantage of the RNA richness of nuclear bodies and their proximity to chromatin, through crosslinking of living cells and purification of key nuclear body-specific proteins. Covalently linked RNA and DNA will be the templates for the preparation of next generation sequencing libraries. The methods employed here – CLIP and ChIP-Seq – are already well established. CLIP is typically used to identify RNA targets of RNA binding proteins (14). ChIP is commonly used to identify transcription factor binding sites or histone modification profiles. Neither method has been exploited to systematically define and track 3D structures in nuclear space. Moreover, the proteins we propose to profile are not typical RNA binding proteins or chromatin interactors; they are nuclear body proteins that have low complexity regions favoring hydrogel formation (15-17). A recent study by the Neugebauer lab achieved preliminary proof of principle, by showing that the low complexity CB scaffolding protein, coilin, binds directly to specific RNAs and associates with the genes that encode them (18, 19).

We refer to the detected set of molecular interactions specific for a given nuclear body as a "body map" (**Fig 2)**. Body maps leverage well-annotated information – genomes and transcriptomes – to define nuclear body composition and dynamics at molecular resolution and are complementary to imaging. Body maps are reminiscent of chromatin signatures, which can be used to describe genomic regions (e.g. enhancers or transcription start sites) and have relevance to the functional organization of the cell nucleus. We will integrate these molecular data with other datasets on chromatin profiles (e.g. comprehensive datasets generated through the ENCODE project) and on chromosomal domains and positioning (e.g. chromosomal contact maps). The chromatin environment of nuclear bodies is an understudied potential source of regulation. Finally, the identification of a defined set of interactions comprising a body map will establish a technique that is broadly applicable to all cellular bodies, economical, and provides information that cannot be gleaned through microscopy or biochemistry. Further innovation is seen in the development of CLIP and ChIP-Seq analysis tools suited to identifying body maps, applying them to dynamic cellular events like cell cycle, and extending this information to other genome-wide data.

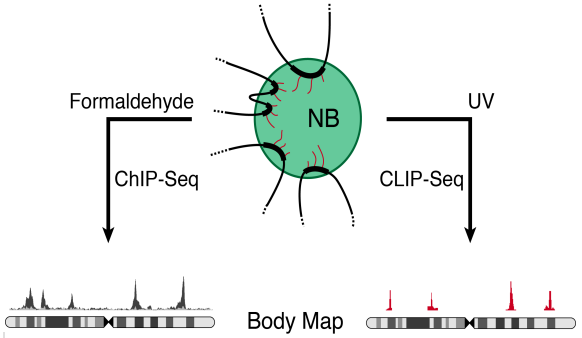



**Fig 2 A body map** defines a nuclear bodies (NB) of interest through a set of interactions between NB-specific proteins and chromosomal loci (black) and RNA (red), detected by ChIP and CLIP. Body maps inform on composition and molecular proximity, complementary to imaging.

## (C) Approach

We will use both microscopy and body maps to determine whether the formation of CBs, HLBs, nuclear speckles, and paraspeckles is transcription dependent and to observe for the first time how chromosomal loci are brought together within the nucleus in 4D. We will establish and validate body maps for these nuclear bodies in mammalian cell lines, applying them to an analysis of nuclear body assembly, maturation and disassembly throughout cell cycle. We will complement this initial series of experiments with a study of zebrafish embryos, in which the rapid reductive cleavage divisions take place in the absence of transcription; this is followed by a period of zygotic genome activation (ZGA), affording the opportunity to observe the biogenesis of nuclear bodies before during and after transcriptional onset in the embryo at the level of morphology and molecular interactions.

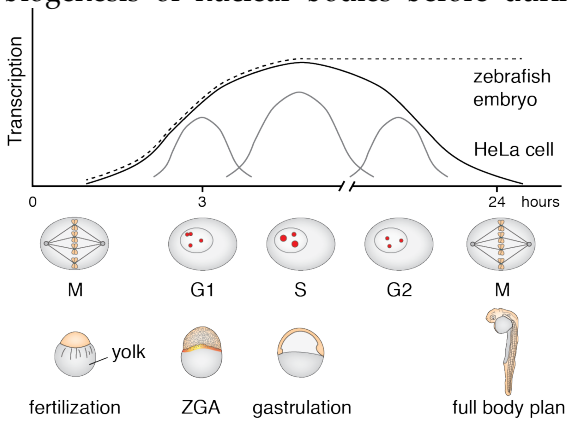
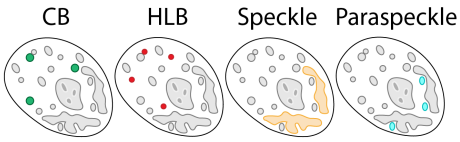



**Fig 3 HeLa cell cycle and zebrafish** embryogenesis enable analysis of nuclear bodies in the context of natural transcriptional programs, RNA-protein interactions, and gene positions.

*Choice of model systems.* A key feature of the parallel use of these two systems is that, in each, there is a phase of natural transcriptional inactivity (M phase and fertilization), followed by a phase of transcriptional onset (G1 and ZGA) and further waves of gene expression as gene expression programs unfold (**Fig 3**). Interestingly, it takes 24 hours to progress from fertilization to the full body plan in zebrafish embryos, the same amount of time it takes for one cell cycle in HeLa cells. The transcriptional changes that take place in the embryo, therefore, occur on a backdrop of rapid cell divisions. Transcription inhibitors, such as α-amanitin, can be injected into the embryo to prevent transcriptional onset. The opportunity to analyze nuclear body formation during natural programs of transcriptional activation and shut-off allows us to identify principles of nuclear body dynamics independent of the mechanics of cell division. Among human cell lines, we chose HeLa cells, an ENCODE Tier2 cell line

with abundant genomic tools, for which a vast literature on nuclear bodies exists. We wish to produce high quality data that can be integrated with other datasets (e.g. ChIP-seq data on chromatin marks, chromosome capture data) to elucidate genome function and generate further testable hypotheses. We chose zebrafish embryos, which the Neugebauer lab pioneered as a model system for the analysis of nuclear architecture due to the external development of the embryos, their transparency and hence suitability for microscopy, and the ability to inject 1-cell embryos with perturbing agents: function-blocking morpholinos, drugs, nucleoside analogs, fluorescent protein-tagged constructs, and fluorescently labeled RNAs (6, 20, 21). As a vertebrate, zebrafish is increasingly used as a disease model, including ALS (22), with clearly recognized orthologues of key proteins and a higher rate of antibody cross-reactivity than other model organisms. This combination of features is ideal for the proposed study.

*Nuclear bodies and choice of proteins.* The proposed work uses known nuclear body proteins as molecular handles to identify RNA and chromatin with which nuclear bodies interact. We have chosen four nuclear bodies, about which we have sufficient knowledge to select protein components (**Fig 4**). All are known to contain high concentrations of specific RNAs (though additional RNAs may be discovered). All selected nuclear body proteins are low complexity and concentrated in the indicated nuclear body (**Fig 4**). Prior evidence of nucleic acid binding was considered. However the recent observation that proteins without recognizable RNA binding domains bind RNA in cells (23, 24) leads us to include additional proteins. Indeed, the CB protein coilin is a low complexity protein with no annotated RNA binding domain, yet it binds hundreds of RNA targets (18). Low complexity proteins are believed to contribute to nuclear body formation by promoting liquid phase separation (15, 16) and have the potential to serve as interaction platforms. Therefore, amino acid composition and annotated motifs were also considered. The properties of each protein are described in Aim 1A.




| | CB | HLB | Speckle | Paraspeckle |
|---|---|---|---|---|
| Coilin | + | +/− | − | − |
| WRAP53, TGS1 | + | − | − | − |
| NPAT, FLASH | − | + | − | − |
| SRSF1&7 | − | − | + | − |
| RBM14, FUS | − | − | − | + |

**Fig 4 Nuclear bodies and proteins to be analyzed.** Two proteins per NB, suitable for imaging and body mapping by iCLIP and ChIP-Seq, are listed. Specificity for the indicated NB is denoted by (+). Coilin is found in HeLa HLBs, due to CB/HLB overlap in this cell type. See Aim 1 for details of each NB and protein.

*Methodologies. In vivo* crosslinking strategies are well suited to protocols that aim for a high degree of specificity through high stringency washing. We will employ two modes of cross-linking: UV crosslinking for RNA target identification (CLIP) and formaldehyde for identification of interacting chromosomal loci (ChIP-Seq). In CLIP, UV light penetrates cells and induces covalent bonds between RNA bases and amino acid side chains at a distance of only several Å. In both CLIP and ChIP, crosslinked antigen is immunopurified and libraries are prepared for sequencing on the Illumina HiSeq2000 platform (14, 25). Two forms of CLIP are proposed here. In iCLIP, UV crosslinking at 254nm samples all RNAs in the cell; iCLIP exploits the inability of reverse transcriptase to pass the crosslinked nucleotide (nt), enabling identification of the crosslink site at nt resolution (26, 27). After crosslinking, RNA-protein adducts are cut from the gel; this adds a purification step and reduces background. In PAR-CLIP, metabolic labeling of RNA with 4-thio-U (4SU) enables 365nm UV crosslinking of only those transcripts made during the labeling period (14, 28). Thus, PAR-CLIP will be used to distinguish newly made transcripts (e.g. nascent RNA) from long-lived RNAs, using short 4SU labeling pulses in HeLa and 4SU injection into zebrafish embryos (21, 29). These crosslinking strategies will be complemented by total RNA-Seq and RNA-Seq of metabolically labeled RNA for normalization and analysis of transcription.

*Analysis*. The analysis of iCLIP and ChIP-Seq datasets requires advanced bioinformatics to ensure a properly processed and analyzed dataset as well as integration with other valuable datasets, such as ChIP-Seq of histone marks and chromosome contact maps (e.g. HiC). The Gerstein lab will leverage expertise in developing and applying computational tools for analysis and interpretation of genomic data generated by modern sequencing technologies. A comprehensive iCLIP analysis tool called **iCAT**, which can be applied to all prominent CLIP methods (iCLIP, PAR-CLIP, and HITS-CLIP), will be developed with the aid of our datasets and enhance analysis. The Gerstein lab will also develop **BodyMapper**, a research approach to integrate ChIP-Seq, iCLIP and RNA-Seq data on nuclear bodies. The planned analysis includes inquiry into the chromatin environment, including the possibility for intra- and inter-chromosomal interactions that likely take place in nuclear bodies. Finally, we will seek to identify networks of interactions among the chosen nuclear bodies, many of which play roles in RNA metabolism and might be expected to "talk" to one another. Networks may reflect trafficking of components, shared dependency on components, or coordinated dynamics. The planned experiments, including e.g. temporal sampling, will provide opportunities to detect functional relationships computationally. Constant interactions between the Neugebauer and Gerstein labs ensure constructive feedback and innovation in implementing these approaches to the 3D and 4D organization of the cell nucleus.

**AIM 1. Identify RNAs and chromosomal loci associated with low complexity nuclear body proteins**

To identify RNA interactors of nuclear body proteins and associated chromosomal loci, we will carry out iCLIP and ChIP-Seq in HeLa cells and validate the results by fluorescence microscopy. We will develop analysis tools specifically designed to detect, quantitate and integrate the data. A powerful way of representing the data is in terms of a body map (see **Fig 2**) of each nuclear body, which is the combined RNA and DNA interaction profile for critical protein component(s).

**1A. Profile low complexity nuclear body proteins in cycling HeLa cells using iCLIP and ChIP-Seq.**

*Establishment of BAC transgenic cell lines.* In order to avoid biases that could arise from the use of antibodies with differing affinities and specificities, we will employ a single GFP tag to which specific antibodies have been developed for purification (30). The GFP tag will be introduced at the N- or C-terminus through recombineering of bacterial artificial chromosomes (BACs), and stable cell lines will be derived following BAC transformation. This system offers several other advantages (30). First, endogenous regulatory elements (enhancers, promoters, introns, etc) within each BAC and integration of a low copy number (usually only 1) into the genome ensure that tagged proteins are expressed at or near endogenous levels. Second, the GFP tag is amenable to imaging. Third, the portability of BACs permits us to study any cell line we choose. Our previous work, which established a precedent for CB and HLB body maps, relied on coilin-GFP expressed from a BAC in HeLa and P19 cell lines (18). In the event that recombineering is unsuccessful or the tagged protein is mislocalized, we will evaluate commercially available antibodies (all presently listed in vendor catalogs) for suitability to iCLIP and ChIP.

*iCLIP, and ChIP-Seq.* Each protein will be subjected to a simple pilot test, based on the early steps of the iCLIP protocol, to determine whether UV-dependent RNA-protein crosslinking is detectable. This test allows us to quickly and economically screen for proteins appropriate for iCLIP. In addition to the proteins listed in Fig 4, other candidate proteins can be pursued. If possible, we will identify two proteins for each nuclear body for which iCLIP and ChIP-Seq will be carried out in two biological replicates. Single-end 75bp reads will be obtained on Illumina HiSeq 2000 at the Yale Center for Genome Analysis. A threshold of ≥0.9 (rank-based Spearman correlation coefficient) will be used to decide whether additional replicates are necessary.

*Total RNA-Seq.* Here we will develop a specific method tailored to our downstream analysis, in which we wish to relate our iCLIP data to transcript abundance. In iCAT (Aim 1B), RNA-Seq data is desired to perform expression correction to derive apparent binding affinities. Therefore, we will purify total RNA from HeLa cells and subject it to the same library preparation protocol as the iCLIP samples (immunoprecipitation excluded). This RNA will be a faithful representation of the transcriptome sampled by UV crosslinking and will contain RNAs of different classes and sizes (e.g. pre-mRNA, mRNA, rRNA, small non-coding RNAs). In iCLIP library preparation, RNA is enzymatically fragmented before adapter ligation and PCR. cDNA is size selected by gel purification to protect against loss of particular RNAs based on size. In our transcriptome data, we will control for size biases with "spike in" RNA size standards (e.g. 50,100, 200, 500, and 1,000 nt long); the resulting reads will be used as internal controls and for normalization. In addition, we will optimize total RNA-Seq to allow rRNA removal by standard methods (Ribo-Zero, RiboMinus); this will allow us to sequence fewer lanes but carries the potential danger of depleting snoRNAs. We will determine the effect of Ribo-Zero depletion on snoRNA abundance, also using the "spike ins" for normalization.

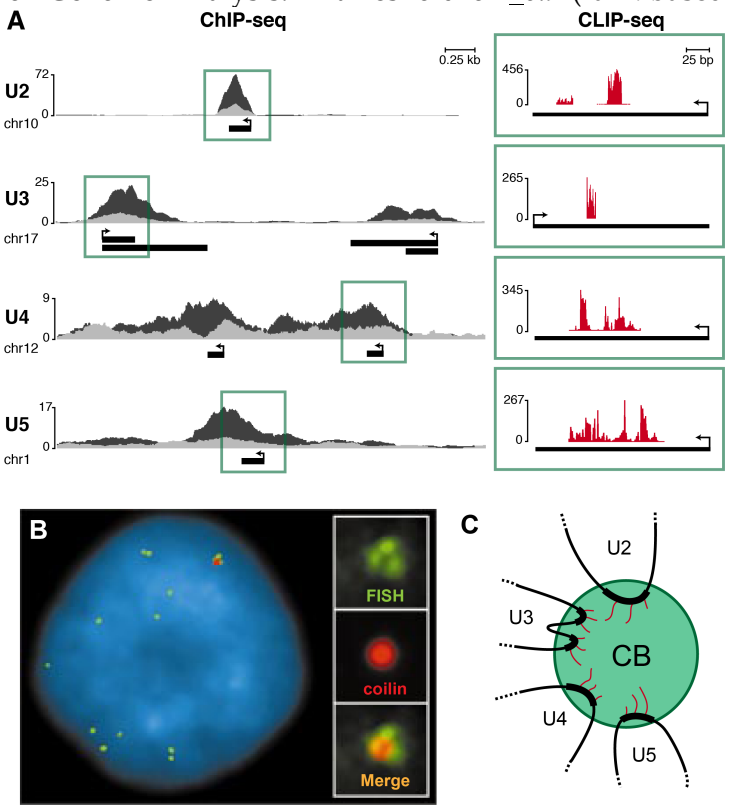



**Fig 5 Preliminary CB body map** based on (A) coilin CLIP and ChIP-Seq [18]. (B) A prior study showed by FISH that multiple U snRNA loci cluster at or near CBs [36]. (C) Schematic illustrating the model that nascent snRNA (red) binding by coilin nucleates CBs at active U snRNA genes (black), which cluster.

**Cajal bodies (CBs).** Most somatic nuclei contain 2-4 CBs, which require the low complexity protein coilin for assembly and maintenance (6, 31). Coilin is conserved yet contains no annotated protein motifs or domains (19). CBs contain high concentrations of all nuclear small non-coding RNAs: spliceosomal snRNAs, snoRNAs en route to the nucleolus, and specialized snoRNAs (scaRNAs) that guide snRNA modifications in CBs (18). CBs are the sites of snRNP and snoRNP maturation (18, 32, 33), suggesting that concentration of factors within CBs makes RNP assembly more efficient (6, 34). Using iCLIP, we showed that coilin actually binds snRNAs, suggesting that coilin binding to nascent snRNA accounts for CB occurrence at U snRNA gene loci (35-37), and may contribute to later snRNP recruitment to CBs (18). Furthermore, coilin ChIP-Seq revealed robust detection of >200 U snRNA genes, which are repeated and occur on disparate chromosomes. By microscopy, CBs occur at or near U snRNA genes, and their limited numbers at interphase suggests that chromosomal loci cluster together within the CB. Evidence supporting gene clustering (36) is incomplete, because the FISH probes were mixed (**Fig 5B**). The combined coilin iCLIP and ChIP-Seq data represent a preliminary body map (**Fig 5**), and the observed set of interactions corresponds to the description of CBs by imaging and functional studies.

We propose to profile TGS1 and WRAP53 (TCAB1, WDR79) by iCLIP and ChIP. TGS1 is 853aa long and contains unstructured with low-complexity regions. We showed that TGS1, which hypermethylates 5′ end caps on RNA, is a strong coilin interactor (18), creating potential for networks of interactions. WRAP53 is present in telomerase RNP and scaRNPs, owing to its interactions with the CAB box RNA element (38-42). WRAP53 contains six WD40 repeats and low complexity regions and is required for CB integrity (38). Coilin ChIP-Seq signal is not present on genes harboring scaRNAs or snoRNAs, consistent with evidence that these RNAs concentrate in CBs after transcription (18). Therefore, detection of TGS1 and WRAP53 at U snRNA genes by ChIP would further validate the presence of the intact CBs at these chromosomal loci.

**Histone Locus Bodies (HLBs)** contain factors essential for the synthesis and processing of histone mRNAs. Replication-dependent histone mRNAs are not polyadenylated but rather require U7 snRNP, SLBP, and specific subunits of the polyadenylation machinery to recognize and cleave the primary transcript at the 3′ end (43). Histone genes occur in clusters on different chromosomes and in species-specific numbers. During the G1/S transition, replicative histone genes become active, and HLB numbers increase (44-46). We identified a preliminary HLB body map from our coilin iCLIP and ChIP-Seq data (**Fig 6**). In HeLa cells, CBs and HLBs are merged, explaining why coilin ChIP-Seq robustly identified individual genes within the two major human histone gene clusters (HIST1 and HIST2). Because U7 snRNA is a major target of coilin by CLIP, U7 snRNP likely links coilin to active histone genes (18). This data agrees with the previous suggestion that histone loci can cluster with U snRNA genes at the composite HLB+CB body (**Fig 6**). Although HLBs and CBs overlap in many transformed cell lines, HLBs are distinct from CBs for three reasons: 1) they have different functions, 2) they are physically separate in stem cells, fly (*D. melanogaster*) and fish (*D. rerio*) cells (9, 20, 47), and 3) HLBs do not depend on coilin; instead the protein FLASH is required for HLB integrity (48, 49). FLASH seems to provide a platform for interactions among histone mRNA, U7 snRNP and the cleavage machinery (50). Therefore, we will profile NPAT (1427aa) and/or FLASH (1982aa). Both proteins are enormous and almost entirely comprised of low complexity regions. We will compare the ChIP-Seq patterns with coilin ChIP-Seq to verify that coilin's recognition of histone genes (**Fig 6**) reflects joined HLBs + CBs.

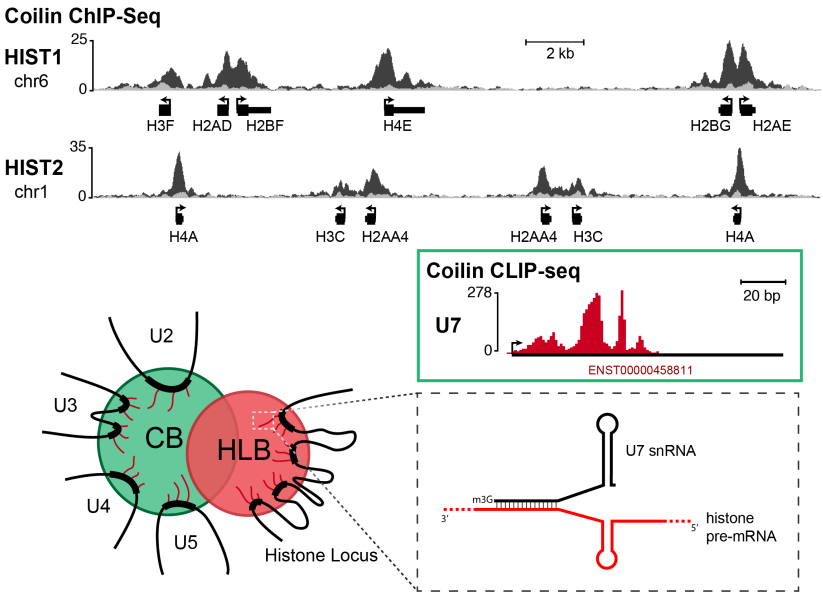



**Fig 6 Preliminary HLB body map** based on coilin ChIP-Seq data [18], which robustly detects histone genes within clusters on disparate chromosomes, as shown. Coilin CLIP detected U7 snRNA [18]. Schematic diagram illustrates the model that nascent histone mRNA (red) basepairs with U7 snRNA, which is in turn bound by coilin. Histone loci cluster together with U snRNA loci in the HLB+CB in HeLa cells.

**Nuclear speckles and paraspeckles** Nuclear speckles are irregularly shaped nuclear bodies that occupy chromatin-poor regions within nuclei (5). They are marked by high concentrations of SR proteins (**Fig 7**), a family of RNA binding proteins involved in all aspects of mRNA biogenesis (51). SR proteins are low-complexity proteins with RNA binding domains capable of forming hydrogels *in vitro* (17). *In vivo*, they bind pre-mRNA splicing targets (52-54) as well as the abundant, intronless lncRNA MALAT1 (metastasis associated lung adenocarcinoma transcript 1). MALAT1 is predominantly nuclear and thought to regulate cell motility and invasion of cancer cells through effects on gene expression (5). Whether SR protein interactions with MALAT1 regulate liquid phase separation *in vivo* is unknown. Concentration of SR proteins by MALAT1 in nuclear speckles regulates the phosphorylation state of SR proteins and may act as a molecular sponge to titrate their concentrations (55, 56). SR proteins and MALAT1 associate with actively transcribed protein-coding regions (52, 57-59). An adjacent nuclear structure, the paraspeckle (**Fig 7**), was found to contain a second lncRNA, NEAT1 (60, 61). NEAT1 transcription is absolutely required for paraspeckle integrity. Paraspeckles are involved in the retention of highly edited mRNAs as well as transcription factors and contain two low-complexity proteins with RNA binding domains, RBM14 and FUS, which undergo liquid phase separation and form aberrant aggregates in disease (11, 62).

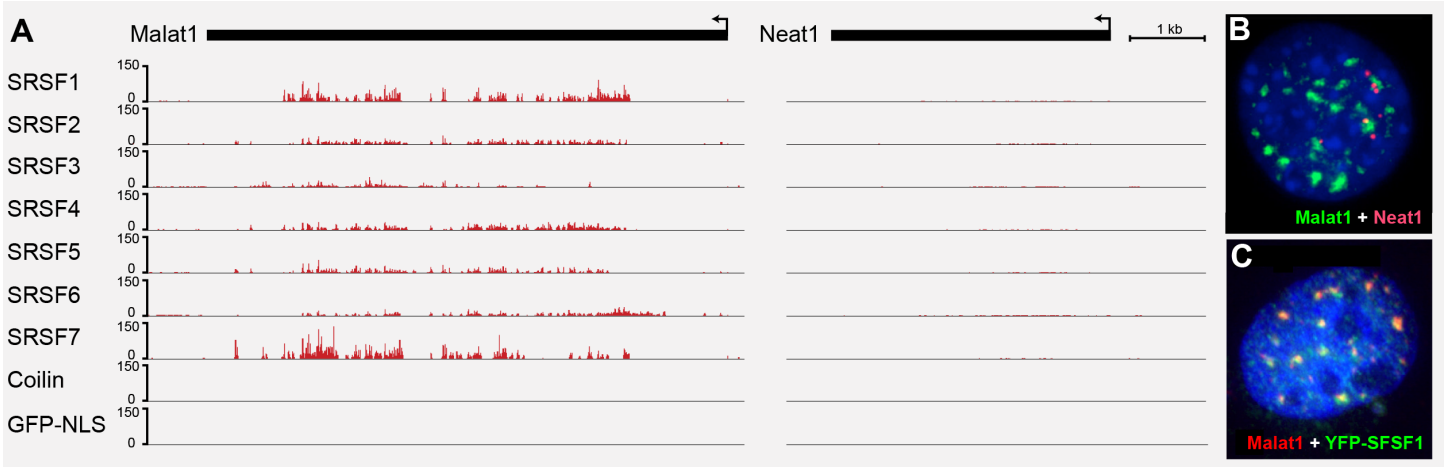



**Fig 7 Preliminary nuclear speckle body map** based on (A) robust detection of iCLIP tags for GFP-tagged SR proteins (SRSF1-7) on MALAT1 but not NEAT1 lncRNA in P19 cells. No Coilin-GFP or GFP-NLS iCLIP tags mapped to either transcript. (B) NEAT1 defines paraspeckles, while (B&C) MALAT1 and SR protein co-localization defines nuclear speckles [images from ref. 55].

We previously established BAC transformed stable cell lines (HeLa and P19) for SR proteins and showed that C-terminally tagged SR proteins are functional (52, 54, 63). We have carried out extensive iCLIP for the entire SR protein family in P19 cells (unpublished). This experiment identifies MALAT1 but not NEAT1 as a major target of all SR protein family members (**Fig 7**). In contrast, coilin-GFP or GFP-NLS crosslinks to these abundant lncRNAs were not detected, showing that MALAT1 reads are not background contaminants. We will focus on SRSF1 and SRSF7, the strongest interactors of MALAT1 in P19 cells (**Fig 6**). A remarkable number of paraspeckle proteins are interesting for our study: RBM14, FUS, PSPC-1, and NONO all contain RNA binding domains and low complexity regions, similar to SR proteins; each is important for the integrity of paraspeckles and protein-protein interactions have been mapped (11). This is very auspicious for successful iCLIP experiments with paraspeckle proteins. In addition, ChIP-seq is planned, and the data will be compared with recent detection of genomic sites of MALAT1 and NEAT1 localization (58).

**1B. Develop a flexible and statistically powerful CLIP-Seq analysis tool (iCAT).**

CLIP-Seq is a family of powerful experimental tools to investigate RNA-protein interactions. In particular, iCLIP maps RNA-protein interactions interactions sites at single nucleotide resolution (26, 64, 65), introducing distinct advantages and challenges regarding analysis. Few tools are specifically designed to analyze iCLIP data. Here we will develop a flexible and statistically powerful iCLIP-Seq Analysis Tool (iCAT) (see **Fig 8** for workflow**)**. After a read pre-processing to remove library primers and PCR duplicates, iCAT employs a two-step mapping process (see below) and considers known exon-intron,

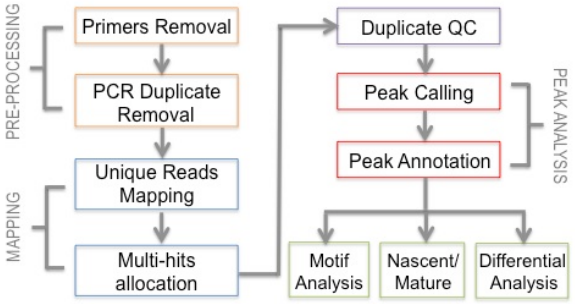



**Fig 8 iCAT, a versatile CLIP analysis tool**

exon-exon, and snoRNA-intron junctions to maximize the read mapping probability. To better assess RNA-protein binding affinities, we will further consider effects of other genomic features, such as gene expression and GC content by effective covariate correction (see below). In addition, iCAT contains a peak annotation module to provide annotations to the peak sites and will also perform motif analysis using external software (e.g. DREME and MEME). Finally, iCAT allows differential binding analysis by quantitatively evaluating the mapping of multiple datasets. Although we are capable of analyzing CLIP data with available tools (see Aim 1), iCAT will enhance the information content extracted from the data (e.g. nascent vs mature RNA) and maximize statistical power when read number is limiting (e.g. when target transcripts are rare or when limited numbers of cells or embryos are employed). Three unique modules of iCAT are discussed below:

*Computationally rescue short reads.* In iCLIP, reverse transcriptase terminates at the crosslink nucleotide, identifying the site of RNA-protein interaction (65). Because of premature termination, up to 85% of reads are short (<30bp) (64) and therefore difficult to map uniquely to the human genome. Most studies either remove the multi-mapped reads or randomly/equally allocate them to numerous mapped sites (28, 64, 66), which results in a loss in accuracy and statistical power. Instead, we will implement a two-step mapping algorithm. In step one, we only map reads that map uniquely to the genome (e.g. nascent snRNAs with templated 3′ end extensions) or the transcriptome (e.g. mature RNA including exon-exon junctions). In step two, the remaining reads are mapped by allowing multi-hits, employing a weighting process guided by the relative abundance of uniquely mapped reads through an expectation-maximization (EM) algorithm. This approach allows us to rescue the short reads for downstream analysis and better biological interpretation. In our data (Aim 1A), 78% of the reads are short; of those, ~60% are uniquely mapped. Thus, up to 40% of our reads are rescuable.

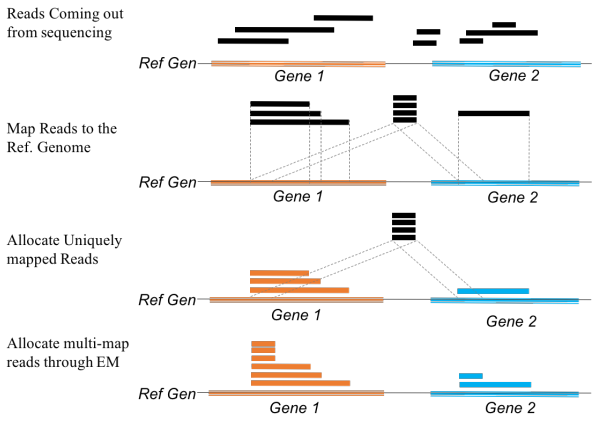



**Fig 9 Rescue of short reads through weighting.**

*Covariate correction for apparent binding affinity estimation.* RNA-protein interaction sites are usually identified by performing an enrichment assessment test of mapped reads. Multiple factors can affect this process. For instance, fewer mapped reads are expected in highly repetitive regions of the human genome even after rescuing short reads. This mappability issue should be considered in the enrichment analysis. Transcript abundance also affects the expected number of crosslinking sites. It is possible that a highly expressed transcript with lower binding affinity generates more iCLIP reads than less abundant transcripts with higher binding affinity. Furthermore, the sequenced reads from an iCLIP experiment may demonstrate sequence or context bias. We propose a multivariate regression approach for better covariate correction. Suppose $y_i$ represents the total number of reads at position $i$, and $\mathbf{X}_i$ denotes the covariate matrix, including parameters like expression, GC content, and mappability. Then we have (Eq. 1):

$$E\left[y_i\right] = \mu_i = \exp\left\{\mathbf{b}\mathbf{X}_i\right\}$$

where $\mu_i$ and $\mathbf{b}$ are the mapped reads expectation and covariate vector. Considering the over-dispersed nature of iCLIP reads across the genome, we will compare several distributions that can handle over-dispersion in the regression (Eq. 1), including negative binomial, generalized Poisson, or beta-binomial distributions, to choose the best fitting model for the iCLIP dataset. Numerical methods, such as Newton's method, can help us to estimate the parameters in (Eq. 1) by the method of maximum likelihood. In the end, a p-value is provided for enrichment interpretation after correction by (Eq. 1).

*Distinguishing nascent from mature RNA*. Whether RNA-protein interactions occur on nascent or mature RNA will be addressed in a functional interpretation module within iCAT that will identify reads spanning intron-exon, exon-exon, snoRNA-intron junctions or DNA-templated reads beyond 3′ end cleavage sites. The number of RNA processing-related reads uncovers the timing of interaction. iCAT also provides information regarding the distribution of crosslink sites (e.g. in introns, UTRs, exons, across junctions).

**1C. Create a computational framework for the investigation of nuclear bodies.**

Nuclear bodies are defined by their constellation of RNA and protein constituents, their chromosomal locations, and their focal 3D structure. This leads to the enticing hypothesis that factors associated with core proteins associate in 3D and can be used to quantitatively characterize and analyze nuclear bodies. For example, the CB protein coilin is detected by ChIP on many histone and U snRNA genes on different chromosomes, binds most snRNAs and snoRNAs as well as a limited set of proteins (18); we would like

employ this combination of features as a molecular signature – a body map (**Fig 2**) – of CBs to enable analysis in cells that are either not amenable to imaging or do not contain nuclear bodies with classical morphology (e.g. M-phase). Furthermore, we would like to discover whether chromosomal features (e.g. a particular histone mark we do not currently know about) correlate with body maps, thereby enabling us to further delve into the mechanisms of nuclear body formation and function. In this aim, we will therefore integrate ChIP-Seq and iCLIP experiments conducted in this study with publically available data on genomic features (histone modifications, transcription factors and other DNA binding proteins, nucleosome positioning), chromosome conformations (HiC), and protein-protein interactions. To do this, the Gerstein lab will develop BodyMapper, leveraging extensive experience analyzing ChIP-Seq and RNA-Seq data (67-78).

*Analytical tools for ChIP-Seq.* ChIP-Seq is a mainstream experimental method for genome-wide identification of transcription factor (TF) binding and chromatin modification sites and will be fundamental to characterizing the interactions of nuclear bodies with chromatin. The Gerstein lab developed PeakSeq (68), a versatile tool for identification of TF binding sites and a standard peak calling program used by the ENCODE and modENCODE consortia for ChIP-Seq datasets (76). The proteins we will analyze by ChIP do not have DNA-binding domains and likely do not associate directly with DNA; therefore, the characteristics of the ChIP-Seq peaks we will obtain are uncertain. This problem is addressed by MUSIC, a peak caller that performs multiscale decomposition of ChIP signals to enable simultaneous and accurate detection of enrichment at a range of narrow and broad peak breadths (68, 78). Binding to genes with similar sequence characteristics (e.g. snRNA genes for CBs, histone genes for HLBs) will be aided by MUSIC's improved mappability correction.

*Analytical tools for RNA-Seq.* For RNA-Seq analysis, we will employ tools developed in the Gerstein lab to handle challenges in read quantification: RSEQtools, enabling expression quantification of annotated RNAs (72); tools that detect, store and query unannotated transcripts (67, 69, 74, 79-82); and methods (e.g. incRNA) that predict and analyze novel ncRNAs, which may be discovered in the course of this project (83). Thus our expertise in identifying and quantifying numerous transcript classes, including pseudogene transcripts (84-86) that share sequence repetitiveness with RNA families in nuclear bodies (e.g. snRNAs and histone RNAs), maximizes opportunities for discovery and analysis of novel RNAs present in nuclear bodies.

*Integration of ChIP-Seq and iCLIP data through BodyMapper.* To advance our mechanistic understanding of how nuclear bodies form and function, we plan to identify molecular components of nuclear bodies and explore the chromatin context of their associated DNA loci. We also want to determine the extent to which nuclear bodies organize the genome by clustering chromosomal loci. The pipeline begins by analyzing ChIP-Seq and iCLIP data to identify RNAs and DNA loci. Nucleic acid targets showing high statistical enrichment will be considered high confidence components, while moderately enriched loci will comprise a list of potential nuclear body components to be pursued by validation. To build an expanded list of associated and putatively associated molecular factors, we will mine publically available databases, including protein-protein interaction and ChIP-Seq data from the ENCODE consortium and the Roadmap Epigenomics Mapping Consortium (87, 88). Previous studies have shown that the genome is organized into domains (89-91). We will assess whether nuclear body-associated loci tend to be located within the same domains. We will use distance constraints from chromosome contact maps (Hi-C) to cluster our nuclear body associated genes, using a similar procedure to distance geometry in protein structure solution which has also recently been employed for reconstruction of chromatin 3D structure (92). One attractive method is Clustering by Fast Search and Find of Density Peaks, which automatically selects the number of clusters it identifies (93). We expect nuclear bodies to have fewer total clusters than diffuse DNA-binding factors, indicating their high 3D connectivity.

*Network Analysis.* BodyMapper will build a combined network containing the RNA and DNA interactions of nuclear body factors, chromosome contacts, and protein-protein interactions (**Fig 10**). With this network, we will search for highly connected modules within the network, which may represent key components for body formation or function. With this network, we will search for highly connected modules, which may represent key components for body formation or function. Our analysis framework will build upon our previous work analyzing biological networks, particularly work integrating ChIP-Seq and RNA-Seq data with to understand transcriptional regulation from ENCODE and modENCODE data (70, 77, 94, 95). To search for key groups, we will initially use simple clique and defective-clique identification procedures (96). If these are insufficient, we will also employ well-established module identification procedures and a simple

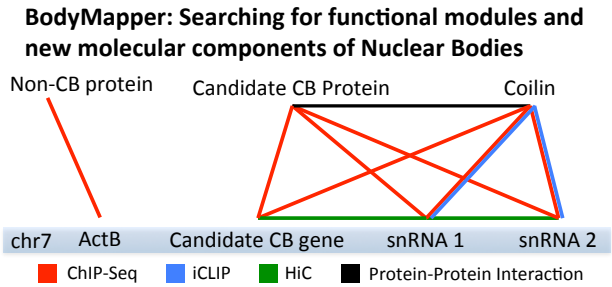


Fig 10. Network Analysis with BodyMapper

search for graph regions with high clustering coefficient (97). We will carefully consider how to weight the contributions of our heterogenous data, similar to our work analyzing combined metabolic, phosphorylation, signaling, genetic, and regulatory networks (98). Members of cliques or highly connected modules will become additional candidate components of nuclear bodies, even if they are missing direct evidence from ChIP-Seq and iCLIP of nuclear body factors (**Fig 10**). The predictions of these models will be validated with methods described throughout this proposal.

*Modeling of Nuclear Body Regulation*. Because of their potential to concentrate regulatory factors, nuclear bodies are an exciting specific context in which to apply predictive analytical frameworks that we have developed on a genome-wide scale. For example, we developed machine-learning models that use binding profiles of chemically modified histones and transcription factors at promoters to predict gene expression measured by RNA-Seq (71, 77, 99). Using a similar framework, but targeted to nuclear body-associated genes, we may observe higher predictive accuracy, or different regulatory patterns than seen more generally. Additionally, nuclear body associated regions may constitute a class of genomic regulatory regions, similar to enhancers. We will use analytical frameworks similar to our enhancer prediction tools in an attempt to identify with HeLa ChIP-Seq data chromatin signals that are particularly associated with nuclear body association (87, 100). These models may help identify nuclear body associated genes in cell lines where the molecular interactions of currently known nuclear body factors have not been investigated.

**1D. Validate CLIP, ChIP and bodymapper results through *in situ* hybridization.**

We will image proteins through their GFP tags and by standard immunostaining combined with fluorescence in situ hybridization (FISH) in order to validate the presence of RNA and DNA species detected by iCLIP and ChIP-Seq in nuclear bodies. When antigens are sensitive to denaturing FISH protocols or when RNAs are inaccessible to hybridization, injection of fluorescent *in vitro* transcribed RNAs is a viable alternative (18). From the preliminary results in Aim 1A, we already know that the U snRNA and histone genes are of interest and will be investigated in this context. It is currently unknown which, if any, chromosomal loci are clustered in nuclear speckles or paraspeckles. However, the presence of MALAT1 and NEAT1 lncRNAs on a subset of protein-coding genes suggests clustering (58), which will be investigated by FISH. ChIP-Seq results from Aim 1A and predicted associations from BodyMapper will guide selection of gene loci for analysis. This is crucial, because inferences from independent chromosome capture experiments could reveal chromosomal loci in nuclear bodies that are currently unknown to us. For examples, some loci may not be detectable by ChIP.

**1E. Test whether perturbing nuclear bodies disrupts gene clustering and expression.**

We will disrupt nuclear bodies through depletion or deletion of essential nuclear body components and test for effects on mRNA biogenesis through RNA-Seq, using protocols elaborated above. Effects on transcription will be ascertained through RNA-Seq. Additionally, we will determine the effects of nuclear body loss on the clustering of chromosomal loci using multi-color DNA FISH. CBs and HLBs will be perturbed through the deletion of both alleles of coilin, using an established HeLa cell line (101). Because HLBs may be resistant to coilin depletion, we will test for the presence of HLBs in the coilin-/- cell line by immunostaining. If HLBs remain, these cells will additionally be depleted of FLASH (see above), using RNA interference; the key question for HLBs is whether repeated histone gene loci on different chromosomes are clustered in the presence and/or absence of HLBs. Paraspeckles will be perturbed through depletion of NEAT1, NONO, RBM14 and/or FUS, each of which is required for paraspeckle integrity (11). Speckles are difficult to perturb by depletion of either MALAT1 or SR proteins. However, if gene clustering is observed in Aim1D, we test whether gene clustering can be reduced or abolished by MALAT1 or SR protein knockdown. Clustered genes detected in Aim 1D will be localized relative to one another by FISH, imaging and downstream image analysis.

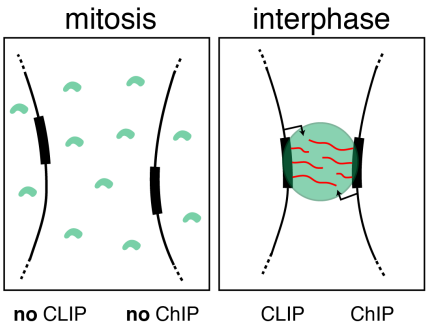



**Fig 11 Nuclear body dynamics in cell cycle may be determined by transcription.** Transcription is off in mitosis, and NBs appear disassembled (green). NBs are mature in interphase, transcription is on, and nascent RNA (red) is present. Possible expectations for body maps shown below panels.

**AIM 2. Determine whether nascent transcripts specify nuclear body assembly during cell cycle.**

This aim uses existing knowledge as well as new findings from Aim 1 to test expectations of our hypothesis that nuclear bodies form through interactions of NB-specific low complexity proteins with nascent RNAs and that nuclear bodies thereby cluster different chromosomal loci within the 3D space of the nucleus over time (**Fig 11**).

*Rationale.* Transcription of a particular locus has places numerous RNA ligands for nuclear body components at genes (7, 102). Morphological detection of CBs and paraspeckles is absolutely dependent on transcription (5, 103, 104), but it is unclear whether subcomplexes or small bodies remain. Whether HLBs are transcription-dependent is controversial (102, 105). Interestingly, transcription inhibition does not cause speckles to disperse but rather to round up (10); whether these speckles have different molecular features is an open question we will address. In addition to exploring through cell cycle and transcription inhibitors, we will use metabolic labeling to profile transcription through cell cycle (remarkably, there is no existing study) and use PAR-CLIP to determine protein interactions with newly synthesized RNA.

**2A. Test dependence of nuclear bodies on cell cycle through cell synchronization.**

Changes in nuclear body morphology across the cell cycle are well known (3, 5, 7, 10, 106). CBs in human tissue culture cells disassemble before mitosis and reassemble during G1 (**Fig 12**). The molecular events underlying assembly and disassembly are completely unknown and are highly relevant to our understanding of the 3D and 4D organization of nuclei. Because transcription is shut off in mitosis, roles for transcriptional activity in the formation and maintenance of nuclear bodies have been proposed. In addition to the nucleating potential of nascent RNA (see **Fig 11**), transcription factors and changes in chromatin state may independently contribute to nuclear body integrity (105, 107). These issues are difficult to address, partly because of the limitations of imaging during nuclear body formation and partly because we are blind to changes in nuclear body composition during assembly and disassembly.

We will pursue two experimental approaches: First, we will create body maps at the 4 major phases of the cell cycle (M, G1, S, G2) to *i*) determine whether known changes in the appearance of nuclear bodies by microscopy translate into changes in body maps and *ii*) compare body maps at mitosis with the transcription inhibition experiment in Aim 3b. This will address whether nuclear body disassembly at mitosis could be purely due to transcriptional shut off (**Fig 11**) or whether other regulatory mechanisms contribute. Second, we will sample the cell cycle in a higher density time course to *i*) determine whether body maps capture nuclear body assembly and disassembly events and *ii*) use the time resolved data to develop BodyMapper4D (Aim 2B).

*Experimental plan and preliminary results.* We previously established synchronization of HeLa cells by double thymidine block (**Fig 12**), and our coilin ChIP data indicate that CB and HLB body maps may disappear at mitosis, reform in G1 and S, and disassemble in G2 (18). Coilin ChIP signals were low on histone (HLB) and U snRNA (CB) genes at mitosis, suggesting a role for transcription. Interestingly, coilin ChIP signals were 3-4 fold higher only 1 hour after S, suggesting rapid transcriptional activation and/or nuclear body assembly. The cell cycle dependence of coilin ChIP signals on U snRNA genes raises the possibility of previously unknown transcription regulation. Therefore, HeLa cells will be synchronized as shown and cell cycle position will be validated by FACS analysis, as before (**Fig 12 inset**). We will choose 2-4 different proteins representing HLBs and CBs (e.g. coilin, WRAP53 and FLASH) for ChIP-Seq and iCLIP in 2 replicates and analyze the data with iCAT and BodyMapper. We will perform total RNA-Seq at each time point. Moreover, metabolic labeling of RNA, followed by 4SU-RNA-Seq (see Aim 2C) will be performed to determine global transcriptional activity in M, G1, S, and G2.

To determine the time period with the greatest rate of change in body maps, we plan to obtain higher density time points. The above data set on M, G1, S, and G2 will be used to determine the cell cycle phase with the greatest alteration in body maps. To capture dynamics, we would like to concentrate additional time points during the period of greatest change. Note that M lasts only 1 hour, while S lasts 8, G2 lasts 3 and G1 lasts at least 12 hours (**Fig 12**). We plan to generate single replicates of ChIP-Seq and iCLIP, and analyze the single replicates, together with their neighbors, as duplicates

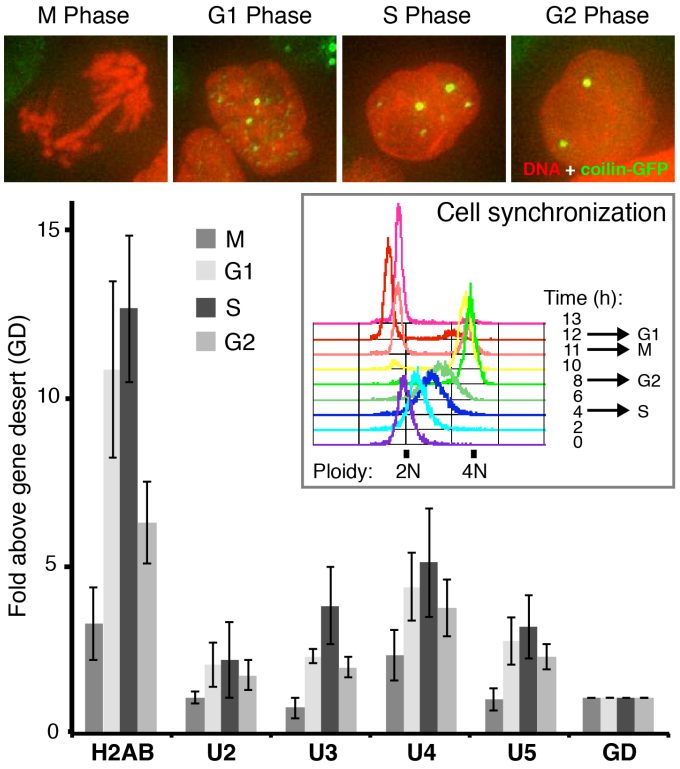



**Fig 12 Cell cycle dependence of CBs and coilin ChIP signals** on histone and U snRNA genes in HeLa cells [18]. Inset, cell synchronization achieved by double thymidine block, with corresponding images above. Below, coilin ChIP.

along a sliding window in the time axis. This should make efficient use of valuable replicates and support the analysis of nuclear body dynamics over time by BodyMapper4D. For future applications, it will be important to determine the analytical power of the data collected, in terms of time resolution and replicate number.

**2B. Extend BodyMapper to analyze time series data.**

*Create BodyMapper4D investigate nuclear body dynamics and association between nuclear bodies*. All nuclear bodies are dynamic. We will evaluate differential states of nuclear bodies, including their dynamics over time, specifically during cell cycle. Molecular changes in nuclear bodies and potential interactions between them are highly interesting. Through BodyMapper4D, we will characterize nuclear bodies as dynamic molecular networks. We will examine the dynamics of nuclear body association through differential analysis of body components and 3D associations between time points (Aim 2A). We will begin by using BodyMapper (Aim 1C) to construct lists of molecular components and networks of nuclear bodies at different time points. We will identify differential components and network characteristics between time points. In our network module identification, some nuclear body-associated modules will likely be stable over time, while others may be required for formation of 3D foci. We will use the temporal clustering of components (e.g. across the cell cycle) as an additional factor to add components to putative bodies (as in BodyMapper). We will further formalize comparisons of network characteristics between time points using dynamic network theory (108). Finally, we will investigate the association of nuclear bodies by constructing joint molecular networks from multiple ChIP/iCLIP/HiC experiments from the same cell cycle phase. Unbiased module finding will indicate possible associations between nuclear bodies; we will compare these modules to those identified from interaction networks that are restricted to a single body. These findings will be tested by fluorescence microscopy.

**2C. Test nuclear body and nuclear organization dependence on transcription through perturbation.**

*Experimental plan*. The drug α-amanitin will be used to inhibit RNA polymerase II in HeLa cells as described (109). Body maps will be derived from iCLIP and ChIP data and compared to data already generated in the absence of α-amanitin (Aim 1). To accompany analysis by iCAT, total RNA-Seq will be performed (Aim1). To address transcriptional activity, we will employ 4-thio-uridine (4SU) metabolic labeling of RNA (29); a 30-minute pulse of will be administered to cells in the presence or absence of α-amanitin. Newly transcribed RNA, indicative of active transcription, will be biotinylated, purified on streptavidin beads and prepared for RNA-Seq according to published protocols established in our lab (21, 29). Reads mapped to the nuclear genome will be normalized to mitochondrial RNA reads, because transcription of the mitochondrial genome is insensitive to α-amanitin (21, 110). Body maps will be compared to nuclear body morphology +/- α-amanitin as determined by fluorescence imaging.

**2D. Identify nascent RNAs associated with nuclear body proteins through metabolic labeling.**

*Experimental plan*. 4SU pulse labeling of synchronized cells (Aim2A) as well as cycling cells +/- α-amanitin (Aim 2C) will be followed by PAR-CLIP to identify newly synthesized RNAs associated with nuclear body proteins. This strategy allows us to distinguish between nascent RNAs that pass through nuclear bodies after transcription. For example, snoRNAs and snRNAs bind coilin and concentrate in CBs after transcription (18, 33, 111). Thus, we expect to detect snRNAs in coilin PAR-CLIP but not snoRNAs. This experiment will provide an unprecedented glimpse of the dynamic association of RNA with nuclear bodies.

**AIM 3. Profile nuclear bodies in zebrafish embryos before, during and after zygotic genome activation**

The Neugebauer lab has pioneered the use of zebrafish as a model system for nuclear architecture. We have identified CBs and HLBs, addressed their dynamics by imaging and have profiled the timing of transcriptional onset through metabolic labeling, identifying the first 592 genes transcribed at the 256-cell stage (6, 20, 21). Before this time, the genome is transcriptionally silent, offering a distinct biological context in which to examine the transcriptional dependence of nuclear bodies and the clustering of chromosomal loci (see **Fig 4**). Orthologs of all of the nuclear body proteins and lncRNAs discussed here are present in zebrafish. We are in a unique position to apply all of the described approaches to zebrafish embryos.

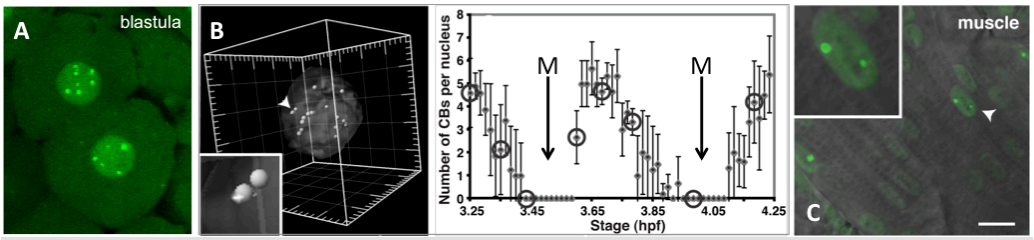



**Fig 13. CB number in zebrafish embryos** varies in cell cycle and development [20]. Confocal sections of blastula (A) pre-ZGA and (B) ZGA, and differentiated muscle cells at 30hpf (C). Live cell imaging and quantitative analysis (B) shows that CBs break down immediately before mitosis (M) and reform within 5 minutes. CBs visualized with YFP-coilin.

### A. Determine the timing of nuclear body appearance relative to transcriptional onset.

*Experimental Plan and Preliminary Results.* We have complete knowledge of early transcription in zebrafish embryos (21, 112). Transcription of histone and snRNA genes begins at the 256-cell stage. We have previously shown that CBs are numerous in the early blastula, even in the absence of transcription, differing from tissue culture cells (**Fig 13**). We presume that, like *Xenopus*, zebrafish may amplify CBs during oogenesis (3). This also necessitates an examination of the constituents of each nuclear body, which can vary between pluripotent, differentiated and germ cells. Here we will image each nuclear body, using fluorescence 2P confocal microscopy and timed embryos. Expression of fluorescently tagged proteins and fluorescent *in vitro* transcribed RNAs will be used as imaging tools, as previously described (6, 20). Using U2 snRNA as a marker of CBs and U7 snRNA as marker of HLBs, we found that HLBs arise at the 128-256 cell stage, when histone gene transcription is activated (**Fig 14A**), supporting our proposal that HLBs are nucleated by gene transcription. Furthermore, α-amanitin treatment leads to abolition of HLBs (**Fig 14B**). Using similar approaches, we will image nuclear speckles and paraspeckles and compare their appearance to the transcription of MALAT1 and NEAT1, which are maternally provided (data not shown). We will investigate the dependence on transcription through α-amanitin treatment, as above.

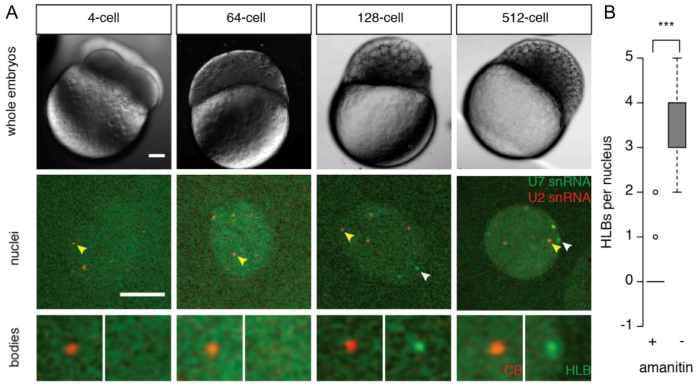



**Fig 14. Zebrafish HLBs** (U7, green) appear when histone genes become transcriptionally active (A), in contrast to CBs (red). HLBs disappear upon transcription inhibition (B).

### B. Establish CLIP in zebrafish embryos and analyze CBs, HLBs, nuclear speckles, and paraspeckles

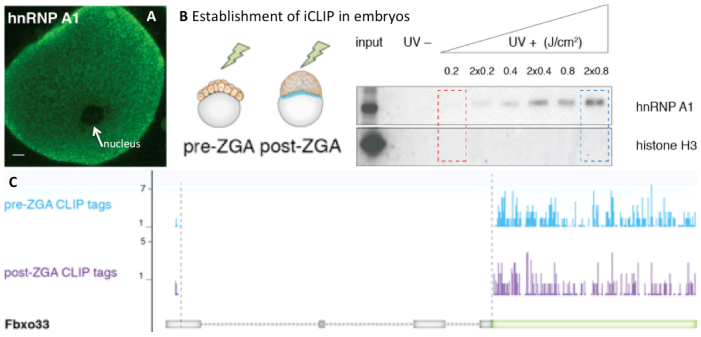



**Fig 14 Establishment of iCLIP in zebrafish embryos**

We established iCLIP in zebrafish embryos (**Fig 13**), utilizing the abundant low-complexity RNA binding protein hnRNP A1. In particular, UV dose had to be carefully adjusted to optimize crosslinking and produce high quality libraries (**Fig 13B**). HnRNP A1 is exclusively nuclear in pre-ZGA embryos (**Fig 13A**), likely due to its binding of 3'UTR sequences in maternal RNA (**Fig 13C**). We will use GFP and/or FLAG-tagging of CB, HLB, speckle and paraspeckle proteins (Aim 1) for iCLIP analysis in zebrafish embryos. RNA-Seq data will be used for iCAT analysis (Aim 1B). These data will reveal the dynamic RNA content of nuclear bodies relative to ZGA.

To determine whether RNAs associated with zebrafish embryo nuclear body proteins are maternally inherited or rather newly transcribed from the zygotic genome, we will employ metabolic labeling of zygotic RNA, previously established in our lab for zebrafish embryos (21). We will use PAR-CLIP (see Aim 2D) to specifically crosslink and identify newly transcribed RNAs. Controls include α-amanitin treatment.

### C. Determine chromosomal sites of nuclear body formation consequences of functional perturbation

We will carry out ChIP-Seq experiments in zebrafish embryos according to published protocols (113). These data will reveal candidate loci at which nuclear bodies assemble. In addition, candidate genes established in Aim1 will be investigated. These will be validated through FISH, and BodyMapper will be implemented to establish body maps and investigate potential chromatin and transcription factor signatures that may participate in nuclear body formation and function.

For functional perturbation, we will initially use morpholinos to deplete nuclear body proteins and antisense oligos to deplete lncRNAs (114). In future, genome engineering with CRISPR/Cas will be important. Cellular phenotypes will be evaluated by confocal microscopy and developmental phenotypes will be determined (6, 20). We will perform RNA-Seq to determine whether these disruptions lead to alterations in gene expression programs. Rescue of depletion phenotypes through injection of *in vitro* transcribed m- and lnc-RNAs will be important controls as well as tools for expression of mutant proteins and RNAs. Finally, we will determine whether these disruptions as well as transcription inhibition lead to a loss of the gene clustering using FISH.

## References


1. **Brangwynne CP.** 2013. Phase transitions and size scaling of membrane-less organelles. The Journal of cell biology **203:**875-881.
2. **Handwerger KE, Cordero JA, Gall JG.** 2005. Cajal bodies, nucleoli, and speckles in the Xenopus oocyte nucleus have a low-density, sponge-like structure. Mol Biol Cell **16:**202-211.
3. **Gall JG.** 2000. Cajal bodies: the first 100 years. Annu Rev Cell Dev Biol **16:**273-300.
4. **King OD, Gitler AD, Shorter J.** 2012. The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease. Brain research **1462:**61-80.
5. **Mao YS, Zhang B, Spector DL.** 2011. Biogenesis and function of nuclear bodies. Trends Genet **27:**295-306.
6. **Strzelecka M, Trowitzsch S, Weber G, Luhrmann R, Oates AC, Neugebauer KM.** 2010. Coilin-dependent snRNP assembly is essential for zebrafish embryogenesis. Nature structural & molecular biology **17:**403-409.
7. **Machyna M, Heyn P, Neugebauer KM.** 2013. Cajal bodies: where form meets function. Wiley Interdiscip Rev RNA **4:**17-34.
8. **Matera AG, Wang Z.** 2014. A day in the life of the spliceosome. Nature reviews. Molecular cell biology **15:**108-121.
9. **Liu JL, Murphy C, Buszczak M, Clatterbuck S, Goodman R, Gall JG.** 2006. The Drosophila melanogaster Cajal body. J Cell Biol **172:**875-884.
10. **Spector DL, Lamond AI.** 2011. Nuclear speckles. Cold Spring Harb Perspect Biol **3**.
11. **Hennig S, Kong G, Mannen T, Sadowska A, Kobelke S, Blythe A, Knott GJ, Iyer KS, Ho D, Newcombe EA, Hosoki K, Goshima N, Kawaguchi T, Hatters D, Trinkle-Mulcahy L, Hirose T, Bond CS, Fox AH.** 2015. Prion-like domains in RNA binding proteins are essential for building subnuclear paraspeckles. J Cell Biol **210:**529-539.
12. **Hirose T, Virnicchi G, Tanigawa A, Naganuma T, Li R, Kimura H, Yokoi T, Nakagawa S, Benard M, Fox AH, Pierron G.** 2014. NEAT1 long noncoding RNA regulates transcription via protein sequestration within subnuclear bodies. Mol Biol Cell **25:**169-183.
13. **Dundr M, Hebert MD, Karpova TS, Stanek D, Xu H, Shpargel KB, Meier UT, Neugebauer KM, Matera AG, Misteli T.** 2004. In vivo kinetics of Cajal body components. J Cell Biol **164:**831-842.
14. **Anko ML, Neugebauer KM.** 2012. RNA-protein interactions in vivo: global gets specific. Trends Biochem Sci **37:**255-262.
15. **Han TW, Kato M, Xie S, Wu LC, Mirzaei H, Pei J, Chen M, Xie Y, Allen J, Xiao G, McKnight SL.** 2012. Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies. Cell **149:**768-779.
16. **Kato M, Han TW, Xie S, Shi K, Du X, Wu LC, Mirzaei H, Goldsmith EJ, Longgood J, Pei J, Grishin NV, Frantz DE, Schneider JW, Chen S, Li L, Sawaya MR, Eisenberg D, Tycko R, McKnight SL.** 2012. Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. Cell **149:**753-767.
17. **Kwon I, Xiang S, Kato M, Wu L, Theodoropoulos P, Wang T, Kim J, Yun J, Xie Y, McKnight SL.** 2014. Poly-dipeptides encoded by the C9orf72 repeats bind nucleoli, impede RNA biogenesis, and kill cells. Science **345:**1139-1145.
18. **Machyna M, Kehr S, Straube K, Kappei D, Buchholz F, Butter F, Ule J, Hertel J, Stadler PF, Neugebauer KM.** 2014. The Coilin Interactome Identifies Hundreds of Small Noncoding RNAs that Traffic through Cajal Bodies. Molecular cell **56:**389-399.
19. **Machyna M, Neugebauer KM, Stanek D.** 2015. Coilin: The first 25 years. RNA Biol **12:**590-596.
20. **Strzelecka M, Oates AC, Neugebauer KM.** 2010. Dynamic control of Cajal body number during zebrafish embryogenesis. Nucleus **1:**96-108.
21. **Heyn P, Kircher M, Dahl A, Kelso J, Tomancak P, Kalinka AT, Neugebauer KM.** 2014. The earliest transcribed zygotic genes are short, newly evolved, and different across species. Cell Rep **6:**285-292.
22. **Yu Y, Chi B, Xia W, Gangopadhyay J, Yamazaki T, Winkelbauer-Hurt ME, Yin S, Eliasse Y, Adams E, Shaw CE, Reed R.** 2015. U1 snRNP is mislocalized in ALS patient fibroblasts bearing NLS mutations in FUS and is required for motor neuron outgrowth in zebrafish. Nucleic Acids Res **43:**3208-3218.
23. **Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, Krijgsveld J, Hentze MW.** 2012. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell **149:**1393-1406.
24. **Baltz AG, Munschauer M, Schwanhausser B, Vasile A, Murakawa Y, Schueler M, Youngs N, Penfold-Brown D, Drew K, Milek M, Wyler E, Bonneau R, Selbach M, Dieterich C, Landthaler M.** 2012. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Molecular cell **46:**674-690.
25. **Konig J, Zarnack K, Luscombe NM, Ule J.** 2011. Protein-RNA interactions: new genomic technologies and perspectives. Nat Rev Genet **13:**77-83.
26. **Konig J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J.** 2010. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nature structural & molecular biology **17:**909-915.
27. **Huppertz I, Attig J, D'Ambrogio A, Easton LE, Sibley CR, Sugimoto Y, Tajnik M, Konig J, Ule J.** 2014. iCLIP: protein-RNA interactions at nucleotide resolution. Methods **65:**274-287.
28. **Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jr., Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T.** 2010. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell **141:**129-141.

29. **Windhager L, Bonfert T, Burger K, Ruzsics Z, Krebs S, Kaufmann S, Malterer G, L'Hernault A, Schilhabel M, Schreiber S, Rosenstiel P, Zimmer R, Eick D, Friedel CC, Dolken L.** 2012. Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. Genome research **22:**2031-2042.
30. **Poser I, Sarov M, Hutchins JR, Heriche JK, Toyoda Y, Pozniakovsky A, Weigl D, Nitzsche A, Hegemann B, Bird AW, Pelletier L, Kittler R, Hua S, Naumann R, Augsburg M, Sykora MM, Hofemeister H, Zhang Y, Nasmyth K, White KP, Dietzel S, Mechtler K, Durbin R, Stewart AF, Peters JM, Buchholz F, Hyman AA.** 2008. BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. Nat Methods **5:**409-415.
31. **Liu JL, Wu Z, Nizami Z, Deryusheva S, Rajendra TK, Beumer KJ, Gao H, Matera AG, Carroll D, Gall JG.** 2009. Coilin is essential for Cajal body organization in Drosophila melanogaster. Mol Biol Cell **20:**1661-1670.
32. **Stanek D, Neugebauer KM.** 2004. Detection of snRNP assembly intermediates in Cajal bodies by fluorescence resonance energy transfer. J Cell Biol **166:**1015-1025.
33. **Boulon S, Verheggen C, Jady BE, Girard C, Pescia C, Paul C, Ospina JK, Kiss T, Matera AG, Bordonne R, Bertrand E.** 2004. PHAX and CRM1 are required sequentially to transport U3 snoRNA to nucleoli. Molecular cell **16:**777-787.
34. **Klingauf M, Stanek D, Neugebauer KM.** 2006. Enhancement of U4/U6 small nuclear ribonucleoprotein particle association in Cajal bodies predicted by mathematical modeling. Mol Biol Cell **17:**4972-4981.
35. **Frey MR, Matera AG.** 1995. Coiled bodies contain U7 small nuclear RNA and associate with specific DNA sequences in interphase human cells. Proc Natl Acad Sci U S A **92:**5915-5919.
36. **Frey MR, Bailey AD, Weiner AM, Matera AG.** 1999. Association of snRNA genes with coiled bodies is mediated by nascent snRNA transcripts. Curr Biol **9:**126-135.
37. **Jacobs EY, Frey MR, Wu W, Ingledue TC, Gebuhr TC, Gao L, Marzluff WF, Matera AG.** 1999. Coiled bodies preferentially associate with U4, U11, and U12 small nuclear RNA genes in interphase HeLa cells but not with U6 and U7 genes. Mol Biol Cell **10:**1653-1663.
38. **Mahmoudi S, Henriksson S, Weibrecht I, Smith S, Soderberg O, Stromblad S, Wiman KG, Farnebo M.** 2010. WRAP53 is essential for Cajal body formation and for targeting the survival of motor neuron complex to Cajal bodies. PLoS Biol **8:**e1000521.
39. **Richard P, Darzacq X, Bertrand E, Jady BE, Verheggen C, Kiss T.** 2003. A common sequence motif determines the Cajal body-specific localization of box H/ACA scaRNAs. Embo J **22:**4283-4293.
40. **Cristofari G, Adolf E, Reichenbach P, Sikora K, Terns RM, Terns MP, Lingner J.** 2007. Human telomerase RNA accumulation in Cajal bodies facilitates telomerase recruitment to telomeres and telomere elongation. Molecular cell **27:**882-889.
41. **Tycowski KT, Shu MD, Kukoyi A, Steitz JA.** 2009. A conserved WD40 protein binds the Cajal body localization signal of scaRNP particles. Molecular cell **34:**47-57.
42. **Deryusheva S, Gall JG.** 2013. Novel small Cajal-body-specific RNAs identified in Drosophila: probing guide RNA function. RNA **19:**1802-1814.
43. **Marzluff WF, Wagner EJ, Duronio RJ.** 2008. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. Nat Rev Genet **9:**843-854.
44. **Ma T, Van Tine BA, Wei Y, Garrett MD, Nelson D, Adams PD, Wang J, Qin J, Chow LT, Harper JW.** 2000. Cell cycle-regulated phosphorylation of p220(NPAT) by cyclin E/Cdk2 in Cajal bodies promotes histone gene transcription. Genes & development **14:**2298-2313.
45. **Wei Y, Jin J, Harper JW.** 2003. The cyclin E/Cdk2 substrate and Cajal body component p220(NPAT) activates histone transcription through a novel LisH-like domain. Molecular and cellular biology **23:**3669-3680.
46. **Bongiorno-Borbone L, De Cola A, Vernole P, Finos L, Barcaroli D, Knight RA, Melino G, De Laurenzi V.** 2008. FLASH and NPAT positive but not Coilin positive Cajal Bodies correlate with cell ploidy. Cell Cycle **7:**2357-2367.
47. **Ghule PN, Dominski Z, Lian JB, Stein JL, van Wijnen AJ, Stein GS.** 2009. The subnuclear organization of histone gene regulatory proteins and 3' end processing factors of normal somatic and embryonic stem cells is compromised in selected human cancer cell types. J Cell Physiol **220:**129-135.
48. **Barcaroli D, Bongiorno-Borbone L, Terrinoni A, Hofmann TG, Rossi M, Knight RA, Matera AG, Melino G, De Laurenzi V.** 2006. FLASH is required for histone transcription and S-phase progression. Proceedings of the National Academy of Sciences of the United States of America **103:**14808-14812.
49. **Barcaroli D, Dinsdale D, Neale MH, Bongiorno-Borbone L, Ranalli M, Munarriz E, Sayan AE, McWilliam JM, Smith TM, Fava E, Knight RA, Melino G, De Laurenzi V.** 2006. FLASH is an essential component of Cajal bodies. Proceedings of the National Academy of Sciences of the United States of America **103:**14802-14807.
50. **Sabath I, Skrajna A, Yang XC, Dadlez M, Marzluff WF, Dominski Z.** 2013. 3'-End processing of histone pre-mRNAs in Drosophila: U7 snRNP is associated with FLASH and polyadenylation factors. RNA **19:**1726-1744.
51. **Shepard PJ, Hertel KJ.** 2009. The SR protein family. Genome Biol **10:**242.
52. **Sapra AK, Anko ML, Grishina I, Lorenz M, Pabis M, Poser I, Rollins J, Weiland EM, Neugebauer KM.** 2009. SR protein family members display diverse activities in the formation of nascent and mature mRNPs in vivo. Mol Cell **34:**179-190.

53. **Anko ML, Muller-McNicoll M, Brandl H, Curk T, Gorup C, Henry I, Ule J, Neugebauer KM.** 2012. The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. Genome Biol **13:**R17.
54. **Anko ML, Morales L, Henry I, Beyer A, Neugebauer KM.** 2010. Global analysis reveals SRp20- and SRp75-specific mRNPs in cycling and neural cells. Nature structural & molecular biology **17:**962-970.
55. **Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, Freier SM, Bennett CF, Sharma A, Bubulya PA, Blencowe BJ, Prasanth SG, Prasanth KV.** 2010. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Molecular cell **39:**925-938.
56. **Tripathi V, Song DY, Zong X, Shevtsov SP, Hearn S, Fu XD, Dundr M, Prasanth KV.** 2012. SRSF1 regulates the assembly of pre-mRNA processing factors in nuclear speckles. Molecular biology of the cell **23:**3694-3706.
57. **Engreitz JM, Sirokman K, McDonel P, Shishkin AA, Surka C, Russell P, Grossman SR, Chow AY, Guttman M, Lander ES.** 2014. RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. Cell **159:**188-199.
58. **West JA, Davis CP, Sunwoo H, Simon MD, Sadreyev RI, Wang PI, Tolstorukov MY, Kingston RE.** 2014. The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. Molecular cell **55:**791-802.
59. **Ji X, Zhou Y, Pandit S, Huang J, Li H, Lin CY, Xiao R, Burge CB, Fu XD.** 2013. SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase. Cell **153:**855-868.
60. **Bond CS, Fox AH.** 2009. Paraspeckles: nuclear bodies built on long noncoding RNA. J Cell Biol **186:**637-644.
61. **Clemson CM, Hutchinson JN, Sara SA, Ensminger AW, Fox AH, Chess A, Lawrence JB.** 2009. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Molecular cell **33:**717-726.
62. **Courchaine E, Neugebauer KM.** 2015. Paraspeckles: Paragons of functional aggregation. J Cell Biol **210:**527-528.
63. **Pabis M, Neufeld N, Steiner MC, Bojic T, Shav-Tal Y, Neugebauer KM.** 2013. The nuclear cap-binding complex interacts with the U4/U6.U5 tri-snRNP and promotes spliceosome assembly in mammalian cells. Rna **19:**1054-1063.
64. **Sugimoto Y, Konig J, Hussain S, Zupan B, Curk T, Frye M, Ule J.** 2012. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. Genome Biol **13:**R67.
65. **Konig J, Zarnack K, Luscombe NM, Ule J.** 2012. Protein-RNA interactions: new genomic technologies and perspectives. Nat Rev Genet **13:**77-83.
66. **Granneman S, Kudla G, Petfalski E, Tollervey D.** 2009. Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. Proceedings of the National Academy of Sciences of the United States of America **106:**9613-9618.
67. **Wang Z, Gerstein M, Snyder M.** 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet **10:**57-63.
68. **Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB.** 2009. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol **27:**66-75.
69. **Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A, Rozowsky JS, Gerstein MB, Wahlestedt C, Hayashizaki Y, Carninci P, Gingeras TR, Mattick JS.** 2011. The reality of pervasive transcription. PLoS Biol **9:**e1000625; discussion e1001102.
70. **Cheng C, Yan KK, Hwang W, Qian J, Bhardwaj N, Rozowsky J, Lu ZJ, Niu W, Alves P, Kato M, Snyder M, Gerstein M.** 2011. Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. PLoS Comput Biol **7:**e1002190.
71. **Cheng C, Min R, Gerstein M.** 2011. TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. Bioinformatics **27:**3221-3227.
72. **Habegger L, Sboner A, Gianoulis TA, Rozowsky J, Agarwal A, Snyder M, Gerstein M.** 2011. RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. Bioinformatics **27:**281-283.
73. **Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N, Bhardwaj N, Rubin M, Snyder M, Gerstein M.** 2011. AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol **7:**522.
74. **Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Roder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigo R, Gingeras TR.** 2012. Landscape of transcription in human cells. Nature **489:**101-108.
75. **Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, Yan KK, Dong X, Djebali S, Ruan Y, Davis CA, Carninci P, Lassman T, Gingeras TR, Guigo R, Birney E, Weng Z, Snyder M, Gerstein M.** 2012. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome research **22:**1658-1667.

76. **Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M.** 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome research **22:**1813-1831.
77. **Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Frietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, Lacroute P, Leng J, Lian J, Monahan H, O'Geen H, Ouyang Z, Partridge EC, Patacsil D, Pauli F, Raha D, Ramirez L, Reddy TE, Reed B, Shi M, Slifer T, Wang J, Wu L, Yang X, Yip KY, Zilberman-Schapira G, Batzoglou S, Sidow A, Farnham PJ, Myers RM, Weissman SM, Snyder M.** 2012. Architecture of the human regulatory network derived from ENCODE data. Nature **489:**91-100.
78. **Harmanci A, Rozowsky J, Gerstein M.** 2014. MUSIC: Identification of Enriched Regions in ChIP-Seq Experiments using a Mappability-Corrected Multiscale Signal Processing Framework. Genome Biol **15:**474.
79. **Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M.** 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science **320:**1344-1349.
80. **Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M.** 2004. Global identification of human transcribed sequences with genome tiling arrays. Science **306:**2242-2246.
81. **Agarwal A, Koppstein D, Rozowsky J, Sboner A, Habegger L, Hillier LW, Sasidharan R, Reinke V, Waterston RH, Gerstein M.** 2010. Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. BMC Genomics **11:**383.
82. **Rozowsky JS, Newburger D, Sayward F, Wu J, Jordan G, Korbel JO, Nagalakshmi U, Yang J, Zheng D, Guigo R, Gingeras TR, Weissman S, Miller P, Snyder M, Gerstein MB.** 2007. The DART classification of unannotated transcription within the ENCODE regions: associating transcription with known and novel loci. Genome research **17:**732-745.
83. **Lu ZJ, Yip KY, Wang G, Shou C, Hillier LW, Khurana E, Agarwal A, Auerbach R, Rozowsky J, Cheng C, Kato M, Miller DM, Slack F, Snyder M, Waterston RH, Reinke V, Gerstein MB.** 2011. Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. Genome research **21:**276-285.
84. **Sboner A, Habegger L, Pflueger D, Terry S, Chen DZ, Rozowsky JS, Tewari AK, Kitabayashi N, Moss BJ, Chee MS, Demichelis F, Rubin MA, Gerstein MB.** 2010. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol **11:**R104.
85. **Pflueger D, Terry S, Sboner A, Habegger L, Esgueva R, Lin PC, Svensson MA, Kitabayashi N, Moss BJ, MacDonald TY, Cao X, Barrette T, Tewari AK, Chee MS, Chinnaiyan AM, Rickman DS, Demichelis F, Gerstein MB, Rubin MA.** 2011. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome research **21:**56-67.
86. **Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M, Reymond A, Hubbard TJ, Harrow J, Gerstein MB.** 2012. The GENCODE pseudogene resource. Genome Biol **13:**R51.
87. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature **489:**57-74.
88. **Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, Farnham PJ, Hirst M, Lander ES, Mikkelsen TS, Thomson JA.** 2010. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol **28:**1045-1048.
89. **Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL.** 2014. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell **159:**1665-1680.
90. **Dekker J, Marti-Renom MA, Mirny LA.** 2013. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet **14:**390-403.
91. **Naumova N, Imakaev M, Fudenberg G, Zhan Y, Lajoie BR, Mirny LA, Dekker J.** 2013. Organization of the mitotic chromosome. Science **342:**948-953.
92. **Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS.** 2010. A three-dimensional model of the yeast genome. Nature **465:**363-367.
93. **Rodriguez A, Laio A.** 2014. Machine learning. Clustering by fast search and find of density peaks. Science **344:**1492-1496.
94. **Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, Alves P, Chateigner A, Perry M, Morris M, Auerbach RK, Feng X, Leng J, Vielle A, Niu W, Rhrissorrakrai K, Agarwal A, Alexander RP, Barber G, Brdlik CM, Brennan J, Brouillet JJ, Carr A, Cheung MS, Clawson H, Contrino S, Dannenberg LO, Dernburg AF, Desai A, Dick L, Dose AC, Du J, Egelhofer T, Ercan S, Euskirchen G, Ewing B, Feingold EA, Gassmann R, Good PJ, Green P, Gullier F, Gutwein M, Guyer MS, Habegger L, Han T, Henikoff JG, Henz SR, Hinrichs A, Holster H, Hyman T, Iniguez AL, Janette J, Jensen M, Kato M, Kent WJ, Kephart E, Khivansara V, Khurana E, Kim JK, Kolasinska-Zwierz P, Lai EC, Latorre I, Leahey A, Lewis S, Lloyd P, Lochovsky L, Lowdon RF, Lubling Y, Lyne R, MacCoss M, Mackowiak SD, Mangone M, McKay S, Mecenas D, Merrihew G, Miller DM, 3rd, Muroyama A, Murray JI, Ooi SL, Pham**

**H, Phippen T, Preston EA, Rajewsky N, Ratsch G, Rosenbaum H, Rozowsky J, Rutherford K, Ruzanov P, Sarov M, Sasidharan R, Sboner A, Scheid P, Segal E, Shin H, Shou C, Slack FJ, Slightam C, Smith R, Spencer WC, Stinson EO, Taing S, Takasaki T, Vafeados D, Voronina K, Wang G, Washington NL, Whittle CM, Wu B, Yan KK, Zeller G, Zha Z, Zhong M, Zhou X, Ahringer J, Strome S, Gunsalus KC, Micklem G, Liu XS, Reinke V, Kim SK, Hillier LW, Henikoff S, Piano F, Snyder M, Stein L, Lieb JD, Waterston RH.** 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science **330:**1775-1787.

95. **Yan KK, Fang G, Bhardwaj N, Alexander RP, Gerstein M.** 2010. Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks. Proceedings of the National Academy of Sciences of the United States of America **107:**9186-9191.
96. **Yu H, Paccanaro A, Trifonov V, Gerstein M.** 2006. Predicting interactions in protein networks by completing defective cliques. Bioinformatics **22:**823-829.
97. **Rivera CG, Vakil R, Bader JS.** 2010. NeMo: Network Module identification in Cytoscape. BMC Bioinformatics **11 Suppl 1:**S61.
98. **Khurana E, Fu Y, Chen J, Gerstein M.** 2013. Interpretation of genomic variants using a unified biological network approach. PLoS Comput Biol **9:**e1002886.
99. **Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ, Pei B, Harmanci AO, Duff MO, Djebali S, Alexander RP, Alver BH, Auerbach R, Bell K, Bickel PJ, Boeck ME, Boley NP, Booth BW, Cherbas L, Cherbas P, Di C, Dobin A, Drenkow J, Ewing B, Fang G, Fastuca M, Feingold EA, Frankish A, Gao G, Good PJ, Guigo R, Hammonds A, Harrow J, Hoskins RA, Howald C, Hu L, Huang H, Hubbard TJ, Huynh C, Jha S, Kasper D, Kato M, Kaufman TC, Kitchen RR, Ladewig E, Lagarde J, Lai E, Leng J, Lu Z, MacCoss M, May G, McWhirter R, Merrihew G, Miller DM, Mortazavi A, Murad R, Oliver B, Olson S, Park PJ, Pazin MJ, Perrimon N, Pervouchine D, Reinke V, Reymond A, Robinson G, Samsonova A, Saunders GI, Schlesinger F, Sethi A, Slack FJ, Spencer WC, Stoiber MH, Strasbourger P, Tanzer A, Thompson OA, Wan KH, Wang G, Wang H, Watkins KL, Wen J, Wen K, Xue C, Yang L, Yip K, Zaleski C, Zhang Y, Zheng H, Brenner SE, Graveley BR, Celniker SE, Gingeras TR, Waterston R.** 2014. Comparative analysis of the transcriptome across distant species. Nature **512:**445-448.
100. **Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, Rozowsky J, Birney E, Bickel P, Snyder M, Gerstein M.** 2012. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol **13:**R48.
101. **Chen Y, Deng Z, Jiang S, Hu Q, Liu H, Songyang Z, Ma W, Chen S, Zhao Y.** 2015. Human cells lacking coilin and Cajal bodies are proficient in telomerase assembly, trafficking and telomere maintenance. Nucleic Acids Res **43:**385-395.
102. **Shevtsov SP, Dundr M.** 2011. Nucleation of nuclear bodies by RNA. Nat Cell Biol **13:**167-173.
103. **Carmo-Fonseca M, Pepperkok R, Carvalho MT, Lamond AI.** 1992. Transcription-dependent colocalization of the U1, U2, U4/U6, and U5 snRNPs in coiled bodies. J Cell Biol **117:**1-14.
104. **Mao YS, Sunwoo H, Zhang B, Spector DL.** 2011. Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs. Nat Cell Biol **13:**95-101.
105. **Salzler HR, Tatomer DC, Malek PY, McDaniel SL, Orlando AN, Marzluff WF, Duronio RJ.** 2013. A sequence in the Drosophila H3-H4 Promoter triggers histone locus body assembly and biosynthesis of replication-coupled histone mRNAs. Dev Cell **24:**623-634.
106. **Boulon S, Westman BJ, Hutten S, Boisvert FM, Lamond AI.** 2010. The nucleolus under stress. Molecular cell **40:**216-227.
107. **Grob A, Colleran C, McStay B.** 2014. Construction of synthetic nucleoli in human cells reveals how a major functional nuclear domain is formed and propagated through cell division. Genes & development **28:**220-230.
108. **Holme P, Saramaki J.** 2012. Temporal Networks. Physics Reports **519:**97-125.
109. **Stanek D, Rader SD, Klingauf M, Neugebauer KM.** 2003. Targeting of U4/U6 small nuclear RNP assembly factor SART3/p110 to Cajal bodies. J Cell Biol **160:**505-516.
110. **Reid BD, Parsons P.** 1971. Partial purification of mitochondrial RNA polymerase from rat liver. Proceedings of the National Academy of Sciences of the United States of America **68:**2830-2834.
111. **Stanek D, Pridalova-Hnilicova J, Novotny I, Huranova M, Blazikova M, Wen X, Sapra AK, Neugebauer KM.** 2008. Spliceosomal Small Nuclear Ribonucleoprotein Particles Repeatedly Cycle through Cajal Bodies. Mol Biol Cell **19:**2534-2543.
112. **Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL, Regev A, Schier AF.** 2012. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res **22:**577-591.
113. **Vastenhouw NL, Zhang Y, Woods IG, Imam F, Regev A, Liu XS, Rinn J, Schier AF.** 2010. Chromatin signature of embryonic pluripotency is established during genome activation. Nature **464:**922-926.
114. **Pauli A, Montague TG, Lennox KA, Behlke MA, Schier AF.** 2015. Antisense Oligonucleotide-Mediated Transcript Knockdown in Zebrafish. PLoS One **10:**e0139504.