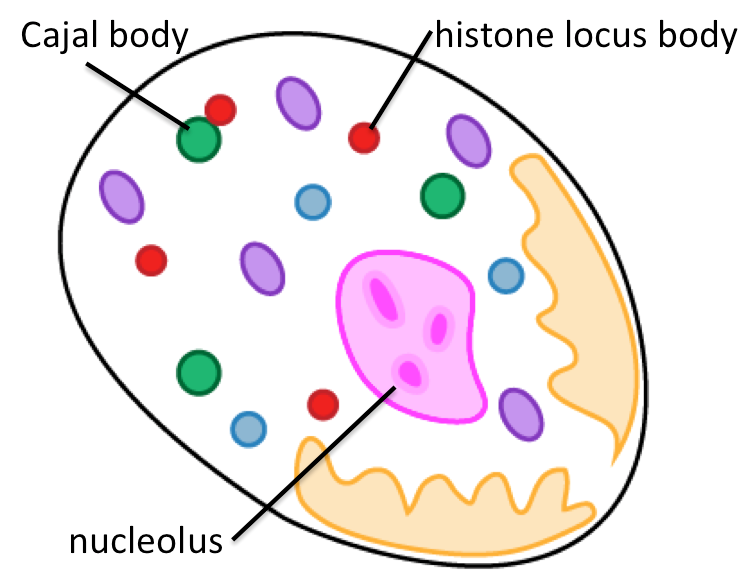
**Research Strategy**

**(A) Significance**

The eukaryotic cell exhibits a remarkable level of compartmentalization imparted by internal membranes that delimit the nucleus and cytoplasmic organelles. Yet the nucleus is filled with a variety of small organelle-like structures, called nuclear bodies, which lack membranes. Nuclear bodies are often composed of multiple proteins and RNAs thought to be held together by multiple, weak binding interactions. Medical evidence implicates nuclear bodies in the aging process and in a wide range of human diseases, including cancer, spinal muscular atrophy and ([1](#_ENREF_1), [2](#_ENREF_2)). Experimental evidence increasingly points to important functions for nuclear bodies in genome function, though many remain to be discovered. For example, nucleoli are the sites of rDNA transcription, rRNA processing and modification, and RNP assembly ([3](#_ENREF_3)). Cajal bodies (CBs) participate in the efficient assembly of snRNPs, required for the splicing of pre-mRNA, and this function is essential in vertebrate embryos ([4-6](#_ENREF_4)). Histone Locus Bodies (HLBs) arise on active histone genes in S-phase and are involved in transcription termination and histone mRNA 3’ end processing. Each of these bodies thereby contains specific RNAs as well as trans-acting factors involved in RNA processing. The respective genes encoding these RNAs often occur in repeats on disparate chromosomes; these genes seem to cluster in the 3D space of the nucleus at the sites of nuclear body formation. Crucial, unanswered questions about nuclear bodies are: How do they arise? How do they contribute to chromosomal positioning within nuclei? And how do they impact genome function?



**Fig 1 Nuclear bodies**

Taking advantage of increased sensitivity and resolution in fluorescence light microscopy as well as emerging molecular markers, numerous nuclear bodies have been identified. The common feature is the existence of punctate staining after immunofluorescent labeling of protein or RNA component. When appropriately stained, their appearance ranges from spherical to interconnected, irregular foci. For puncta to be considered bodies, an unwritten rule is that the apparent diameter of the object be greater than the wavelength of light (i.e. > 250nm). Although morphology is the starting point for defining a nuclear body, a lack of precise definitions creates ambiguity. Because their contents are not enclosed by lipid bilayers, bodies are in direct contact with their surrounding environment. Thus, molecular constituents exchange rapidly with nucleoplasm, and the bodies themselves are highly dynamic. Due to limits of resolution, fluorescent images of bodies can be hard to interpret; it is not clear whether signals come from mature, complete nuclear bodies or whether these could be sub-complexes. It is therefore hard to discern the point of body origin in time or space, especially if the cell is caught in the act of assembling a body. In particular, the size and molecular complexity of nuclear bodies vary throughout the cell cycle. What we therefore lack, but urgently need, is precise and dynamic information on the composition and chromosomal location of nuclear bodies.

The goal of this project is to define nuclear bodies as a set of molecular signatures, based on specific interactions among DNA, RNA and proteins and independent of imaging. In addition to nucleoli, CBs and HLBs, we plan to investigate nuclear speckles, PML bodies, and gems. Current details of the composition and function of each body are given in the Approach section, as this background justifies our experimental design. The significance of this proposal for the 4D Nucleome is the opportunity to generate comprehensive genome-wide datasets for a number of nuclear bodies and to integrate these with other datasets on chromatin profiles (e.g. comprehensive datasets generated through the ENCODE project) and data on chromosomal domains and positioning derived from mapping chromosomal contact maps (e.g. by HiC and related techniques ongoing in other labs). These integrative aspects will be achieved through the collaborative efforts of the Neugebauer and Gerstein labs, the latter of which will leverage expertise in advanced data analysis approaches designed to tackle these specific questions. In particular, we will be able to derive network relationships among nuclear bodies and genomic loci. Taken together, we believe that this project will contribute original concepts and tools to the study of nuclear and genome organization.

**(B) Innovation**

The notion of studying nuclear bodies through molecular signatures, as an alternative to imaging and image analysis, is highly original. Although the concept is reminiscent of the use of chromatin signatures to describe genomic regions (e.g. enhancers or transcription start sites), the latter genetic elements are not tied to cellular morphology. The molecular signatures we refer to include the crosslinking patterns of specific nuclear body proteins on RNA (iCLIP) and DNA (ChIP-Seq). Following proper validation, RNA-protein crosslinks can indicate the presence of a given nuclear body in cells, while ChIP-Seq can reveal the chromosomal locations at which nuclear bodies form. The nature of the data requires advanced bioinformatics, which ensure a properly processed and analyzed dataset as well as integration with other available datasets, such as ChIP-Seq of histone marks and chromosome contact maps (e.g. HiC). This aspect is extremely interesting, as the chromatin environment of nuclear bodies has simply not been addressed.

The feasibility of this proposal is buoyed by the combined and complementary expertise of the co-PIs, Neugebauer and Gerstein. The Neugebauer lab has a long history of working on CBs, nuclear speckles, and HLBs, having used advanced fluorescence light microscopy methods – FRET, 3D reconstruction from wide field deconvolution and confocal microscopy in mammalian tissue culture cells and zebrafish embryos – to characterize their composition and function. Neugebauer has recently applied RNA-protein crosslinking to these nuclear bodies, finding that patterns of RNA-protein crosslinking can accurately report on RNA composition ([7](#_ENREF_7)). For example, iCLIP of a crucial protein component of CBs (coilin) yielded a list of hundreds of small non-coding RNAs that were shown through validation to be present in these bodies. Moreover, the demonstration that proteins, like coilin, lacking predictable RNA binding domains display interaction with RNA under 0 Å UV-crosslinking conditions allows us to add them to the growing list of RNA interacting proteins identified in other scenarios ([8](#_ENREF_8), [9](#_ENREF_9)). The Gerstein lab has extensive experience in developing and applying computational tools for analysis and interpretation of genomic data generated by modern sequencing technologies. Gerstein will leverage the set of efficient tools we developed to conduct integrative analysis of the ChIP-Seq and RNA-Seq data generated in the project. 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, which all play roles in RNA processing and thereby might be expected to “talk” to one another. Networks may reflect trafficking of components, shared dependency on components, or coordinated responses to perturbation. The planned experiments, including e.g. temporal sampling, will provide opportunities to detect such functional relationships computationally.

Fig 2 Molecular Signatures of Nuclear Bodies. (schematic)

**(C) Approach**

The proposed work uses known proteins as molecular handles with which to probe RNA and DNA interactions in living cells. We have chosen proteins of interest, based on their unambiguous localization to the nuclear body of interest and, where available, prior evidence of nucleic acid binding. To determine RNA and DNA interactions, each protein will be employed for UV crosslinking, immunopurification and RNA-Seq (iCLIP) as well as formaldehyde crosslinking followed by ChIP-Seq. In order to avoid biases that could arise from the use of antibodies with different affinities and levels of specificity, we will employ a single tag (GFP) to which excellent antibodies have been developed for purification ([10](#_ENREF_10)). 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 ([10](#_ENREF_10)). First, endogenous regulatory elements (enhancers, promoters, introns, etc) within each BAC and the integration of a low copy number (usually only 1) into the genome ensure that the tagged protein is expressed at endogenous or near endogenous levels. Second, the GFP tag is amenable to imaging, which is an important validation tool for this study. Third, the portability of BACs ensures that we can study any cell line we choose. Our previous work, which established a precedent for molecular signatures of CBs, relied on coilin-GFP expressed from a BAC in human HeLa and mouse P19 cell lines ([7](#_ENREF_7)). The iCLIP datasets from these two cell lines were remarkably similar, suggesting that the molecular signatures of nuclear bodies may not be cell type specific. Indeed, chromosome contact maps derived from HiC showed a high degree of similarity in 7 different cell lines ([11](#_ENREF_11)). Therefore, we feel confident our data will be generally applicable and choose to focus on human cell lines (HeLa and GM12878) that are the focus of many genome-wide studies, e.g. in the context of the ENCODE project. We have also previously established BAC transformed stable cell lines for the SR proteins present in nuclear speckles ([7](#_ENREF_7), [12-14](#_ENREF_12)). These, like many RNA binding proteins, autoregulate their levels such that the “third allele” provided on the BAC contributes to, but does not exceed, endogenous protein levels ([12](#_ENREF_12), [15](#_ENREF_15)).

A key feature of our proposal is the use of high-throughput technologies that allow the identification RNA targets of RBPs in a genome-wide manner. Previous data implicating nuclear body proteins in nucleic acid binding was taken into account in the selection of proteins to study; however the recent observation that proteins without recognizable RNA binding domains can and do bind RNA in cells leads us to include additional proteins ([8](#_ENREF_8), [9](#_ENREF_9)). Indeed, the CB protein coilin has no annotated RNA binding domain but binds hundreds of RNA targets ([7](#_ENREF_7)). Methods for identifying direct interactions between RNA and protein *in vivo* rely on cross-linking of cells with ultraviolet (UV) light and provide a relatively unbiased, global approach for the derivation of target RNAs as well as *in vivo* binding sites ([16](#_ENREF_16), [17](#_ENREF_17)). UV light penetrates cells and induces covalent bonds between RNA bases and amino acid side chains at a distance of only several ångströms; the covalent bond permits stringent washing. Unlike formaldehyde, UV light does not induce protein-protein crosslinks, making UV preferable for addressing direct binding. UV crosslinking is relatively inefficient, but that can be overcome by using sufficient starting material. In crosslinking and immunoprecipitation (CLIP), fragmentation of the purified RNA permits sequencing of so-called “tags”. A further advance in cloning strategy and library preparation in iCLIP permits identification of the actual crosslink site with a resolution of a few nucleotides ([18](#_ENREF_18)). A variation of CLIP, called PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation), utilizes incorporation photo-reactive ribonucleoside analogs, such as 4-thiouridine (4-SU) and 6-thioguanosine (6-SG) into RNA transcripts synthesized in living cells ([19](#_ENREF_19)). The advantage of PAR-CLIP is that it may increase the efficiency of cross-linking ([20](#_ENREF_20)); however a disadvantage seems to be elevated background over that detected by iCLIP ([21](#_ENREF_21)). We have optimized iCLIP as well as ChIP protocols for the study of RNA binding proteins in our lab ([7](#_ENREF_7), [15](#_ENREF_15), [22](#_ENREF_22)).

**Specific Aim 1. Develop molecular signatures for nuclear bodies**

Employ UV crosslinking and immunoprecipitation followed by RNAseq (iCLIP) to catalog RNA-protein interactions that define nuclear bodies at the molecular level. Formaldehyde crosslinking and ChIP-Seq will capture their chromosomal locations. The “molecular signature” of each of the indicated nuclear bodies is the combined RNA and DNA interaction profile for critical protein components. Below, we describe the specific experiment for each body listed. Although approaches for the study of each body are separately described, the experimental steps towards there completion will be performed in parallel.

**Aim 1a. Cajal bodies (CBs) and Gems**

**Aim 1b. Histone Locus Bodies (HLBs)**

**Aim 1c. Nucleoli**

**Aim 1d. Nuclear speckles**

**Aim 1e. PML bodies**

**Specific Aim 2. Use molecular signatures to track nuclear body dynamics and function in cell cycle**

Perform CLIP and ChIP throughout the cell cycle to identify time periods at which nuclear bodies form and function. Metabolic labeling of RNA, positive selection of RNA, followed by RNAseq will reveal function of nuclear bodies upon selected gene disruptions. Time series in the context of the following experiments will be performed:

**Aim 2a. Cell synchronization**

**Aim 2b. Metabolic labeling of RNA**

**Aim 2c. Disruption of nuclear bodies**

**Specific Aim 1. Develop molecular signatures for nuclear bodies**

Here we will perform a series of RNA crosslinking experiments, using immunoprecipitation of nuclear body proteins following UV crosslinking (iCLIP) and formaldehyde crosslinking (ChIP-Seq), to identify RNA and DNA sequences associated with each body.

**Specific Aim 3. Create a computational framework for the characterization of nuclear bodies.**

We will first analyze ChIP-Seq and CLIP-Seq data to identify the molecular interactions of protein components of nuclear bodies with RNA and with chromosomal loci. To do this, we will employ our widely used ChIP-Seq analysis tools, PeakSeq and MUSIC. Leveraging our experience in ChIP-Seq analysis, we will develop a CLIP-Seq analysis pipeline that is sensitive to key relevant aspects of nuclear bodies, including distinction of nascent and mature RNA, mapping to repetitive RNAs, and correction for RNA expression levels. We will then integrate our data with publicly available CLIP-Seq, ChIP-Seq (e.g. comprehensive surveys of histone modifications), protein-protein interactions, and HiC data to characterize nuclear bodies as molecular networks, as opposed to optically imaged bodies, and explore their dynamic assembly across the cell cycle and the differential compositions of these bodies between organisms.

**Aim 3a: Apply analytical tools for ChIP-Seq and RNA-Seq data.**

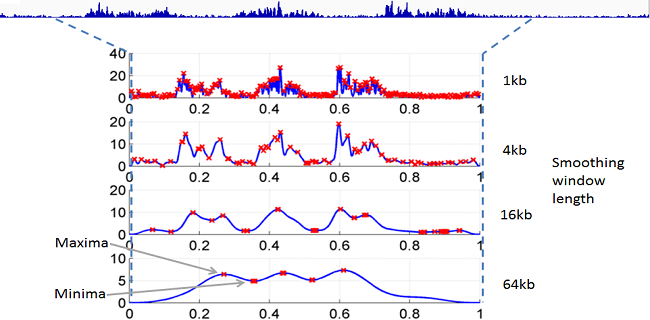
ChIP-Seq is the mainstream experimental method for genome-wide identification of transcription factor (TF) binding and chromatin modification sites, and will be fundamental to our aim to characterize the interactions of nuclear bodies with chromatin. We developed two ChIP-Seq peak calling tools: PeakSeq ([23](#_ENREF_23)) and MUSIC ([24](#_ENREF_24)). PeakSeq is a highly versatile and widely used tool for identification of TF binding sites in ChIP-Seq type experiments and, in particular, is one of the standard peak calling programs used by the ENCODE and modENCODE consortia for numerous ChIP-Seq datasets ([25](#_ENREF_25)). PeakSeq constructs local thresholds using input signals from genomic DNA without an enrichment process to simulate the null process for the background. PeakSeq then locates protein binding regions by identifying peaks that are significantly enriched relative to the background signal. MUSIC is a recently introduced signal processing approach for identification of enriched regions in ChIP-Seq experiments. Utilizing multiscale decomposition of the ChIP-Seq signal profile in conjunction with mappability correction, MUSIC allows identification of broad enrichment domains ([24](#_ENREF_24)). In addition, we introduced a probabilistic model-based method, Target Identification from Profiles (TIP) that can be used to identify a given TF’s target genes based on ChIP-Seq data ([26](#_ENREF_26)). We have applied machine-learning methods that integrate multiple genomics features to classify human regulatory regions from ENCODE data of more than 100 transcription factor binding sites. In particular, we were able to identify potential enhancers from regions classified as gene-distal regulatory modules ([27](#_ENREF_27)). This achievement signals our ability to identify regulatory modules in the context of nuclear body formation and function. For example, intergenic DNA regions have been implicated in the formation of nucleoli and histone locus bodies ([28](#_ENREF_28), [29](#_ENREF_29)); we anticipate being able to detect these and other regions important for these and other nuclear bodies from the datasets generated. Moreover, the hypothesis that nuclear bodies may regulate the expression of the genes on which they form (e.g. do CBs regulate snRNA gene transcription?) will be investigated experimentally in Aim 2. The analysis carried out in this aim may independently reveal regulatory genomic regions distal to these “target” genes.

For RNA-Seq analysis, we have developed RSEQtools, a computational package that enables expression quantification of annotated RNAs and identification of splice sites and gene models ([30](#_ENREF_30)). In addition, we have developed IQseq, a computationally efficient method to quantify isoforms for alternatively spliced transcripts ([31](#_ENREF_31)). Comparisons between RNA-Seq samples and other genome-wide data can be facilitated in part by our Aggregation and Correlation Toolbox (ACT), which is a general purpose tool for comparing genome signal tracks ([32](#_ENREF_32)). An important challenge in RNA-Seq analysis is detecting unannotated transcription that may be hard to distinguish from noise. This topic has been central to many of our expression analysis tools ([33-38](#_ENREF_33)). Our Database of Annotated Regions with Tools (DART) package contains tools for identifying unannotated genomic regions that are enriched for transcription, as well as a framework for storing and querying this information ([39](#_ENREF_39)). We have also developed specific tools to identify types of transcripts that are difficult to detect using standard analysis pipelines. We introduced FusionSeq, a pipeline to detect transcripts that arise due to trans-splicing or chromosomal translocations ([41](#_ENREF_41), [42](#_ENREF_42)). We also developed PseudoSeq to address the issue of quantification of pseudogene expression, which is difficult to separate from the transcription of parent genes with similar sequences ([43](#_ENREF_43)). Another major area of interest in RNA-Seq analysis is linking expression variation to genotype. Our AlleleSeq tool ([44](#_ENREF_44)) combines diploid genomic information with RNA-Seq data to identify transcripts showing allele specific expression.

**Fig 8 Tools for RNA-Seq analysis**



**Fig 7 ChIP-Seq Analysis through MUSIC**



We have extensive experience conducting integrated analyses of large sets of RNA-seq data, primarily through the ENCODE project ([45](#_ENREF_45)). We played a lead role in the analysis of model organisms and human transcriptome studies within the consortium ([36](#_ENREF_36), [46](#_ENREF_46), [47](#_ENREF_47)). We have also conducted extensive studies of the relationship between ChIP-Seq data for localization of transcription factors and histone modifications and gene expression through RNA-Seq ([48](#_ENREF_48), [49](#_ENREF_49)). Currently, we are active participants of the Brainspan project, which profiles RNAs in different parts of the human brain (http://www.brainspan.org) and we lead the data integration and analysis component of the data management and resource repository for the NIH extracellular RNA communication program consortium (http://commonfund.nih.gov/Exrna/).

**Aim 3b*.*****Develop a flexible and statistically powerful CLIP-Seq analysis tool (iCAT)**

CLIP-Seq is a family of powerful experimental tools to investigate protein-RNA interactions. In particular, iCLIP is a method that maps protein-RNA interaction sites to resolution that approaches a single base ([50)](https://docs.google.com/a/yale.edu/document/d/19UyKllc8LJOkXjFlTmuZgfBIT36_z6VKofW-cf1t-b4/edit#heading=h.46r0co2). However, few tools exist to analyze iCLIP [[MRS: Okay to delete the introductory sentences above, which I wrote]]. To address this problem, we will develop a flexible and statistically powerful iCLIP-Seq Analysis Tool (iCAT). The iCAT pipeline begins with a read preprocessing module to effectively remove library primers and reads duplicated due to PCR amplification. The pipeline proceeds by considering the challenge of mapping posed by extensive short truncated reads. To do this, iCAT employs a two-step mapping process: first it finds uniquely mappable reads and then uses these to guide allocation of the ambiguous reads. Specifically, during this step besides mapping to the genome, iCAT will also consider the known exon-exon junctions to maximize the mapping probability to accurately locate the binding sites. To better capture the binding affinity, we will further consider effects of other genomic features, like gene expression and GC content by effective covariant correction. In addition, iCAT also contains a peak annotation module to provide annotations to the peak sites. We will also perform motif analysis based on the claimed binding sites. Finally, our iCAT pipeline also allows differential binding analysis by quantitatively evaluating the mapping of multiple datasets.

*Computational efforts to rescue the mapping of short reads.* While iCLIP enables the identification of cross-linking sites at single nucleotide resolution ([50)](https://docs.google.com/a/yale.edu/document/d/19UyKllc8LJOkXjFlTmuZgfBIT36_z6VKofW-cf1t-b4/edit#heading=h.46r0co2), it also introduces several distinct computational challenges that necessitate the development of new pipelines for accurate binding site identification. For example, iCLIP utilizes a modified cDNA cloning procedure to capture truncated cDNAs that result in the presence of primers within the read sequences. In addition, PCR amplification bias can result in limited library complexity, with thousands of sequences arising from a single cDNA. Furthermore, it has been reported that up to 85% of the cDNAs in iCLIP experiment are truncated ([51)](https://docs.google.com/a/yale.edu/document/d/19UyKllc8LJOkXjFlTmuZgfBIT36_z6VKofW-cf1t-b4/edit#heading=h.2lwamvv), resulting in large number of short reads that are difficult to map within the human genome. In view of these difficulties, we first propose a preprocessing step that contains primer removal and PCR amplification control based on our extensive previous experience in developing small RNA processing pipelines. After preprocessing, a large proportion of reads are often less than 30bp in length, making it difficult to uniquely map these reads to the genome or transcriptome. Many previous studies remove the multimappable reads or randomly/equally allocate them to numerous mapped sites ([19,](https://docs.google.com/a/yale.edu/document/d/19UyKllc8LJOkXjFlTmuZgfBIT36_z6VKofW-cf1t-b4/edit#heading=h.3j2qqm3) [51,](https://docs.google.com/a/yale.edu/document/d/19UyKllc8LJOkXjFlTmuZgfBIT36_z6VKofW-cf1t-b4/edit#heading=h.2lwamvv) [52)](https://docs.google.com/a/yale.edu/document/d/19UyKllc8LJOkXjFlTmuZgfBIT36_z6VKofW-cf1t-b4/edit#heading=h.111kx3o). Such schemes may be problematic as they either discard a large percentage of sequenced reads or adopt a heuristic approach that may lead to a large number of false positives. Instead, we will implement a two-step mapping algorithm. In step one, we only map reads that can map uniquely to the genome (e.g. nascent snRNAs with templated 3’ end extensions) or the transcriptome (mature RNA including exon-exon junctions in the case of protein-coding genes). In step two, the remaining reads will be further mapped by allowing multi-hits, but weighting the mapping by using uniquely mapped reads as guides. By this two-step mapping approach, we will rescue short reads for downstream analysis in a dynamical [[MRS: word choice?]] manner for better biological interpretation.

*Covariant correction for better binding affinity estimation.* RNA-protein interaction sites are usually identified by performing an enrichment assessment of the mapped reads. Multiple factors may affect this process. For instance, some parts of the human genome are highly repetitive, so a smaller number of mapped reads is expected even after rescuing the very short reads. This mappability issue should be considered in the enrichment analysis. In addition, gene expression levels affect the expected number of crosslinking sites. It is possible for a highly expressed transcript with lower binding affinity to generate more iCLIP reads than less abundant transcripts with higher binding affinity. Furthermore, the sequenced reads from an iCLIP experiment may display GC or context bias. We propose a multivariate regression approach for better covariant correction. Considering the over-dispersed nature of iCLIP reads across the genome, we will compare several distributions that can handle over-dispersion, including negative binomial, generalized Poisson, or beta-binomial distributions, to choose the best fitting model for the iCLIP dataset. In the end, a p-value is provided for enrichment interpretation after correction.

*Functional interpretation module for the identified binding sites*. An important question to answer is whether the RNA-protein interaction happens on the nascent or mature RNA molecules. To address this question, for each identified binding site, we will include a module in our pipeline to look back into our mapped read files to search for reads that span the intron-exon or exon-exon junctions. The number of these two types of reads will be provided in the output file to uncover the exact timing of the interaction. In addition, we will provide additional supporting information for users to interpret each binding sites, such as whether they are located in introns, UTRs, exons, or other genomic features.

*Motif analysis from identified binding sites.* Binding motifs are critical to understanding the binding specificities and affinities of proteins of interest. After calling binding sites, our iCAT pipeline will employ external software, such as DREME and MEME, for accurate motif discovery. We will also provide motif discovery in specific regions as requested by users.

*Differential binding analysis.* Observing differential occupation of binding sites in iCLIP-Seq experiments across different biological samples is key to understanding the biological roles of protein-RNA interactions. Hence our iCAT pipeline will allow input of several datasets and build statistical methods to quantitatively measure how interactions change with time.

**Aim 3c. Create BodyMapper, a tool to characterize the components, interactions and dynamics of nuclear bodies**

To create comprehensive molecular pictures of nuclear bodies, we will develop BodyMapper. This tool have three components. Module 1 will integrate, ChIP-Seq, CLIP-Seq and protein-protein interaction data to characterize the molecular components of nuclear bodies. Module 2 will use Hi-C data to investigate the 3-dimensional organization of nuclear bodies using chromosome contact maps. Finally, Module 3 will evaluate differential states of nuclear bodies, including the dynamics of assembly during the cell cycle, and could be used to investigate differences in nuclear body interactions between normal and diseased tissue. To develop BodyMapper, we will apply our extensive experience in building network tools and in analyzing large regulatory networks [[MRS: Need to increase preliminary data!]].

*Module 1: Molecular signatures of nuclear bodies.* The first module of BodyMapper will begin by using ChIP-Seq and CLIP-Seq experiments for core nuclear body proteins to identify genomic loci and RNA molecules that associate with the body (see Aim 3a-3b). For nuclear bodies, such as Cajal Bodies, for which the core factor interacts with both RNA molecules and the genomic loci from which they are derived, we will consider these high confidence itneractions. We will then use publically available data to build an expanded list of associated and putatively associated molecular factors. We will use ChIP-Seq data from the ENCODE consortium and the Roadmap Epigenomics Mapping Consortium to investigate whether nuclear bodies are particularly associated with binding of chromatin features like transcription factor binding as well as histone modifications, *and will build machine learning models to associate nuclear bodies with their chromatin context* [[MRS: include italicized text?]]. We will also identify putative protein components of nuclear bodies through publically available protein-protein interaction databases [[MRS: which databases to cite?]].

*Module 2: Investigate 3-dimensional nuclear body organization using chromosome contact maps.* The key defining feature of nuclear bodies is their focal 3-dimensional structure. This leads to the enticing hypothesis that the factors associated with core nuclear body proteins themselves associate in 3D. For example, our recent study of the Cajal Body factor Coilin showed that it targets 216 genomic loci, including many histone and U snRNA genes on different chromosomes (7). Separate microscopy studies showed that both Cajal Bodies and U snRNAs cluster particularly at mid-interphase, suggesting that Coilin and its target genes may associate in 3D in a cell cycle dependent manner (53, please check citations!).

Module 2 of BodyMapper will investigate the 3-dimensional association of genomic loci that interact with nuclear bodies, using chromosome contact maps generated by the 4D nucleome consortium using techniques like Hi-C. The genome is known to be organized in 3-dimensional domains, so we will first assess whether nuclear body-associated loci tend to be located within the same domains. We will then use distance constraints from chromosome contact maps to cluster our nuclear body associated genes. For this analysis, we will take advantage of the ChIP-Seq data for hundreds of transcription factors that are available in HeLa cells from the ENCODE consortium, most of which are not thought to associate in three dimensional bodies. This will allow us to calibrate our search for interactions of genomic loci for nuclear bodies. We will also use distance constraints from chromosome contact maps to construct networks from our nuclear body associated loci. We will highlight loci with high degree and betweenness centrality as potentially important components of nuclear bodies. Finally, just as enhancers are well known to affect gene expression through long-range association with gene promoters, we will investigate the potential regulatory effects of nuclear body association with genomic loci.

*Module 3: Investigate nuclear body dynamics*. Many nuclear bodies, including Cajal Bodies, are known to associate at particular phases of the cell cycle and some may have different properties in disease states. We will examine the dynamics of nuclear body association through differential analysis of body components and 3-dimensional associations between time points and using dynamic network theory \cite{[doi:10.1016/j.physrep.2012.03.001](http://dx.doi.org/10.1016/j.physrep.2012.03.001)}. We will begin by using modules 1 and 2 to construct lists of molecular components and 3-D chromosomal contacts of our nuclear bodies at different time points. We will then search for differential components between time points. These differences, especially in 3-D contacts, should correspond to known microscopic observations of nuclear body formation. Finally we will model both common and differential interactions as edges of a dynamic network. Applying dynamic network theory to the temporal evolution of nuclear body interactomes may shed light on the mechanism of the assembly of these bodies, for instance, whether there are multiple pathways as a result of stochasticity or the assembly follows a single pathway via a hierarchical order.

Overall, BodyMapper will combine data for molecular interactions and chromosome contact maps to build a microscopy-independent molecular picture of nuclear bodies, including their dynamic formation. This pipeline will be applicable to any nuclear body whose central factor has ChIP-Seq data, and particularly to those with both ChIP-Seq and CLIP-Seq data. Because of the availability of ChIP-Seq data for over a hundred chromatin-associated proteins, the BodyMapper’s use of chromosome contact maps to investigate associated between genomic loci bound by chromatin factors may enable discovery of additional proteins that form nuclear bodies [[MRS: Should we cut this last sentence?]].

Old Aims 3c/3d.

**Aim 3c. Characterize nuclear bodies via integration of functional genomic data.**

To identify the location of various nuclear bodies, we propose to integrate ChIP-Seq data of defining protein specific to a particular nuclear body (for example coilin to Cajal body (CB)) with genomic resources generated by the ENCODE consortium and the Roadmap Epigenomics Mapping Consortium. Previous analysis of coilin showed the enrichment of ChIP-Seq peaks snRNA genes and histone genes ([7](#_ENREF_7)), further integration of chromatin features like transcription factor binding as well as histone modifications will enable us to investigate the location not only in terms of genome annotation, but the co-localization of nuclear bodies with various transcriptional machineries (e.g. Pol II and other regulatory factors), as well as the co-localization of various nuclear bodies of interest, for instance HLBs and CBs.

Apart from locating various nuclear bodies in terms of the 1D genomics coordinates, we further plan to study their spatial location with respect to the 3D genome. Our Coilin ChIPseq data identified 216 gene loci bound by coilin, including many repetitive histone and U snRNA genes on disparate chromosomes ([7](#_ENREF_7)). Yet HeLa cells have only 3-4 CBs in mid-interphase. This strongly suggests that coilin target genes cluster as the cell cycle proceeds, and an earlier FISH study is consistent with this ([53](#_ENREF_53)). It would be extremely interesting to know if this clustering, representative of a nuclear body, is detectable within chromosome contact maps! This will be done by further integrating the ChIP-Seq data with 3D physical contact maps to be generated by the 4D nucleome consortium using techniques like Hi-C. Contact maps reveal how chromosomes are organized into domains and territories. We plan to investigate the location of nuclear bodies with respect to different domains and territories. Such integration further allows us to investigate how distant parts of the genome co-localize with particular nuclear bodies. As many distal connections are responsible for distant regulatory elements like enhancers, we plan to investigate if nuclear bodies play any role in the regulatory processes.

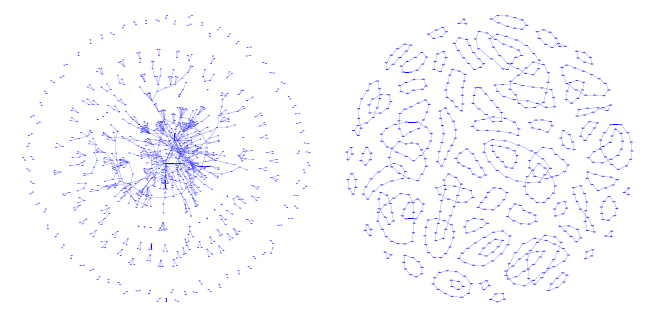
The iCLIP and ChIPseq data feed back on each other, because iCLIP often identifies RNAs that serve to nucleate nuclear bodies co-transcriptionally. Thus, in analyzing the iCLIP data for a number of nuclear body proteins, we can identify genome-wide all the annotated transcribed RNAs that the protein binds to and that potentially determine the chromosomal location of the nuclear body. One can similarly identify novel un-annotated transcribed regions or TARs (transcriptionally active regions ([34](#_ENREF_34))) of the genome, which may be transcribed into RNA and localized in the nuclear body of interest. These predictions can be queried by looking into the ChIP-Seq data and further tested (see Aims 1 & 2). [[MG: how to we integrate this look back ref.]] Additionally, we can make use of the 3D physical contact maps for further spatial orientation. Using genes that are known to be physical part of a nuclear body (U2 and U3 for the Cajal body), we can identify genic DNA regions in the genome, which are physically linked to the same nuclear body. By comparing the genes whose DNA is associated with the body with the transcribed RNAs specific to a nuclear body (identified from iCLIP), we also can distinguish the RNAs that are transcribed at the nuclear body from those that are transcribed in a different location but transported to the nuclear body of interest. Some genomic regions associated with a nuclear body could also contain genes, which are not transcribed which might be activated in response to a change in cellular state and may warrant further testing.

We have described how we intend to identify the DNA regions (and their constituent genes) and the transcribed RNAs that can be associated with different nuclear bodies. Furthermore with the use of protein-protein-interaction networks (PPIs) we intend to identify other as yet unidentified proteins that physically co-associate with the known proteins specific to nuclear bodies. We can compare these proteins to see if they correspond to the DNA genes and RNAs that we identify as being associated with the same nuclear body. Thus we intend to create a comprehensive initial molecular catalog of the constituent members of nuclear bodies at the DNA, RNA and protein level where possible.

**Aim 3d. 4D network analysis**

*Preliminary results* Network representations can be applied consistently to many different types of biological data and therefore can be used as a powerful framework for integration of heterogeneous datasets ([54-56](#_ENREF_54)). We have developed many novel approaches for studying biological networks. We developed methods to construct and analyze the regulatory networks of human and model organisms ([46](#_ENREF_46), [48](#_ENREF_48), [57](#_ENREF_57), [58,](#_ENREF_58) ) based on ENCODE and modENCODE datasets. We integrated regulatory networks with gene expression to uncover different types of functional modules ([59-62](#_ENREF_59)). We constructed and analyzed hierarchical regulatory networks in model organism and humans ([48](#_ENREF_48), [63-66](#_ENREF_63)), discovering that that the hierarchy rather than centrality ("hubiness") better reflects the importance of regulators. We built a multi-layered network that incorporated information from heterogeneous data sources such as protein-protein interactions and metabolic, phosphorylation, signaling, genetic, and regulatory networks ([67](#_ENREF_67)) and we used networks to improve our understanding of genomic variants ([68](#_ENREF_68)) We also introduced several software tools for network analysis including Topnet, ([69](#_ENREF_69)) tYNA ([70](#_ENREF_70)) and PubNet ([71](#_ENREF_71)).

Capitalizing on the uniformly processed and matched experimental data obtained by ENCODE and modENCODE consortia, we have performed a series of comparative studies across distant metazoan phyla. A comparative analysis of human, worm, and fly revealed remarkable conservation of general properties of regulatory networks. ([72](#_ENREF_72)). We discovered co-expression modules shared in animals and enriched in their developmental genes. To examine the degree of conservation on how chromatin features affect gene expression, we constructed a ‘universal model’ for quantitative prediction of coding and non-coding gene expression levels from chromatin features at the promoter. The model is based on a single set of organism-independent parameters and, in the three model organisms, achieved accuracy comparable to the organism specific models ([47](#_ENREF_47)) We performed a multi-organism comparison of pseudogenes and found that they are much more lineage specific than protein-coding genes, reflecting the different genome remodeling processes in each organism ([73](#_ENREF_73)). We also introduced a framework to quantify differences between networks and by comparing matching networks across organisms, found a consistent ordering of rewiring rates of different network types ([74](#_ENREF_74)) We developed a new comparative genomics tool, OrthoClust, for simultaneously clustering data across multiple species ([75](#_ENREF_75)) This integrates co-association networks of individual species utilizing the orthology relationships of genes between species.



**Fig 9 something about networks (left) and (right).**

*Comparison and Dynamics*. The integration of ChIP-Seq and iCLIP data with various functional genomic resources essentially arrive at an interactome that characterize the constituent members of a nuclear body of interest. The comparison of such interactomes is of particular interest because it directly addresses important questions, such as: the differences of a nuclear body of interest between normal and disease states; how spatial organization changes in response to cellular conditions; and the degree of conservation across various species (e.g. human, mouse and zebrafish). We aim to understand how the interactome changes across tissues, cell types and organisms by quantifying the conserved and specific structures by a rewiring framework ([74](#_ENREF_74)). As the interactome comprises of multiple types of edges (for instance, protein-protein interactions, protein-RNA interactions, protein-DNA interactions), we plan to extend the multi-layer network framework we developed for comparing the similarity and difference between the interactome ([75](#_ENREF_75)). [[MG: we need a little more on what we'd do here w/ networks]]

Apart from comparing the localization and composition of a particular nuclear body in two states, we plan to further characterize the dynamics, for instance across the cell cycle. Previous results of using ChIP-Seq and iCLIP on Coilin hinted at the dependence of CB number and size on the cell cycle ([7](#_ENREF_7)). We aim to continue the thread and generalize the investigation to other nuclear bodies. Practically, we plan to use a temporal network framework to integrate the interactomes obtained at different time-point of the cell cycle \cite{[doi:10.1016/j.physrep.2012.03.001](http://dx.doi.org/10.1016/j.physrep.2012.03.001)}. Understanding the temporal evolution of the interactomes shed light on the mechanism of the assembly of various nuclear bodies, for instance, whether there are multiple pathways as a result of stochasticity or the assembly follows a single pathway via a hierarchical order.

**References**

1. Morimoto M, Boerkoel CF. The role of nuclear bodies in gene expression and disease. Biology (Basel). 2013;2(3):976-1033. Epub 2013/09/17. doi: 10.3390/biology2030976. PubMed PMID: 24040563; PMCID: 3771687.

2. McCann KL, Baserga SJ. Genetics. Mysterious ribosomopathies. Science. 2013;341(6148):849-50. Epub 2013/08/24. doi: 10.1126/science.1244156. PubMed PMID: 23970686; PMCID: 3893057.

3. Boulon S, Westman BJ, Hutten S, Boisvert FM, Lamond AI. The nucleolus under stress. Molecular cell. 2010;40(2):216-27. Epub 2010/10/23. doi: 10.1016/j.molcel.2010.09.024. PubMed PMID: 20965417; PMCID: 2987465.

4. Strzelecka M, Trowitzsch S, Weber G, Luhrmann R, Oates AC, Neugebauer KM. Coilin-dependent snRNP assembly is essential for zebrafish embryogenesis. Nature structural & molecular biology. 2010;17(4):403-9. Epub 2010/04/02. doi: 10.1038/nsmb.1783. PubMed PMID: 20357773.

5. Machyna M, Heyn P, Neugebauer KM. Cajal bodies: where form meets function. Wiley Interdiscip Rev RNA. 2013;4(1):17-34. Epub 2012/10/09. doi: 10.1002/wrna.1139. PubMed PMID: 23042601.

6. Matera AG, Wang Z. A day in the life of the spliceosome. Nature reviews Molecular cell biology. 2014;15(2):108-21. Epub 2014/01/24. doi: 10.1038/nrm3742. PubMed PMID: 24452469.

7. Machyna M, Kehr S, Straube K, Kappei D, Buchholz F, Butter F, Ule J, Hertel J, Stadler PF, Neugebauer KM. The Coilin Interactome Identifies Hundreds of Small Noncoding RNAs that Traffic through Cajal Bodies. Molecular cell. 2014;56(3):389-99. Epub 2014/12/17. doi: 10.1016/j.molcel.2014.10.004. PubMed PMID: 25514182.

8. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, Krijgsveld J, Hentze MW. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell. 2012;149(6):1393-406. Epub 2012/06/05. doi: 10.1016/j.cell.2012.04.031. PubMed PMID: 22658674.

9. 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. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Molecular cell. 2012;46(5):674-90. Epub 2012/06/12. doi: 10.1016/j.molcel.2012.05.021. PubMed PMID: 22681889.

10. 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. BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. Nat Methods. 2008;5(5):409-15. PubMed PMID: 18391959.

11. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. 2014;159(7):1665-80. Epub 2014/12/17. doi: 10.1016/j.cell.2014.11.021. PubMed PMID: 25497547.

12. Anko ML, Morales L, Henry I, Beyer A, Neugebauer KM. Global analysis reveals SRp20- and SRp75-specific mRNPs in cycling and neural cells. Nature structural & molecular biology. 2010;17(8):962-70. Epub 2010/07/20. doi: 10.1038/nsmb.1862. PubMed PMID: 20639886.

13. Sapra AK, Anko ML, Grishina I, Lorenz M, Pabis M, Poser I, Rollins J, Weiland EM, Neugebauer KM. SR protein family members display diverse activities in the formation of nascent and mature mRNPs in vivo. Mol Cell. 2009;34(2):179-90. Epub 2009/04/28. doi: S1097-2765(09)00147-6 [pii]

10.1016/j.molcel.2009.02.031. PubMed PMID: 19394295.

14. Pabis M, Neufeld N, Steiner MC, Bojic T, Shav-Tal Y, Neugebauer KM. The nuclear cap-binding complex interacts with the U4/U6.U5 tri-snRNP and promotes spliceosome assembly in mammalian cells. Rna. 2013;19(8):1054-63. Epub 2013/06/25. doi: 10.1261/rna.037069.112. PubMed PMID: 23793891; PMCID: 3708526.

15. Anko ML, Muller-McNicoll M, Brandl H, Curk T, Gorup C, Henry I, Ule J, Neugebauer KM. The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. Genome Biol. 2012;13(3):R17. Epub 2012/03/23. doi: 10.1186/gb-2012-13-3-r17. PubMed PMID: 22436691; PMCID: 3439968.

16. Konig J, Zarnack K, Luscombe NM, Ule J. Protein-RNA interactions: new genomic technologies and perspectives. Nat Rev Genet. 2011;13(2):77-83. Epub 2012/01/19. doi: 10.1038/nrg3141. PubMed PMID: 22251872.

17. Anko ML, Neugebauer KM. RNA-protein interactions in vivo: global gets specific. Trends Biochem Sci. 2012;37(7):255-62. Epub 2012/03/20. doi: 10.1016/j.tibs.2012.02.005. PubMed PMID: 22425269.

18. Konig J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nature structural & molecular biology. 2010;17(7):909-15. Epub 2010/07/06. doi: 10.1038/nsmb.1838. PubMed PMID: 20601959; PMCID: 3000544.

19. 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. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell. 2010;141(1):129-41. Epub 2010/04/08. doi: 10.1016/j.cell.2010.03.009. PubMed PMID: 20371350; PMCID: 2861495.

20. Kishore S, Gruber AR, Jedlinski DJ, Syed AP, Jorjani H, Zavolan M. Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. Genome Biol. 2013;14(5):R45. Epub 2013/05/28. doi: 10.1186/gb-2013-14-5-r45. PubMed PMID: 23706177.

21. Friedersdorf MB, Keene JD. Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. Genome Biol. 2014;15(1):R2. Epub 2014/01/08. doi: 10.1186/gb-2014-15-1-r2. PubMed PMID: 24393468; PMCID: 4053780.

22. Bieberstein NI, Straube K, Neugebauer KM. Chromatin immunoprecipitation approaches to determine co-transcriptional nature of splicing. Methods Mol Biol. 2014;1126:315-23. Epub 2014/02/20. doi: 10.1007/978-1-62703-980-2\_23. PubMed PMID: 24549674.

23. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol. 2009;27(1):66-75. Epub 2009/01/06. doi: 10.1038/nbt.1518. PubMed PMID: 19122651; PMCID: 2924752.

24. Harmanci A, Rozowsky J, Gerstein M. MUSIC: Identification of Enriched Regions in ChIP-Seq Experiments using a Mappability-Corrected Multiscale Signal Processing Framework. Genome Biol. 2014;15(10):474. Epub 2014/10/09. doi: 10.1186/PREACCEPT-9116006401338101. PubMed PMID: 25292436; PMCID: 4234855.

25. 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. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome research. 2012;22(9):1813-31. Epub 2012/09/08. doi: 10.1101/gr.136184.111. PubMed PMID: 22955991; PMCID: 3431496.

26. Cheng C, Min R, Gerstein M. TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. Bioinformatics. 2011;27(23):3221-7. Epub 2011/11/01. doi: 10.1093/bioinformatics/btr552. PubMed PMID: 22039215; PMCID: 3223362.

27. Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, Rozowsky J, Birney E, Bickel P, Snyder M, Gerstein M. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol. 2012;13(9):R48. Epub 2012/09/07. doi: 10.1186/gb-2012-13-9-r48. PubMed PMID: 22950945; PMCID: 3491392.

28. Salzler HR, Tatomer DC, Malek PY, McDaniel SL, Orlando AN, Marzluff WF, Duronio RJ. A sequence in the Drosophila H3-H4 Promoter triggers histone locus body assembly and biosynthesis of replication-coupled histone mRNAs. Dev Cell. 2013;24(6):623-34. Epub 2013/03/30. doi: 10.1016/j.devcel.2013.02.014. PubMed PMID: 23537633; PMCID: 3613982.

29. Grob A, Colleran C, McStay B. Construction of synthetic nucleoli in human cells reveals how a major functional nuclear domain is formed and propagated through cell division. Genes & development. 2014;28(3):220-30. Epub 2014/01/23. doi: 10.1101/gad.234591.113. PubMed PMID: 24449107; PMCID: 3923965.

30. Habegger L, Sboner A, Gianoulis TA, Rozowsky J, Agarwal A, Snyder M, Gerstein M. RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. Bioinformatics. 2011;27(2):281-3. Epub 2010/12/08. doi: 10.1093/bioinformatics/btq643. PubMed PMID: 21134889; PMCID: 3018817.

31. Du J, Leng J, Habegger L, Sboner A, McDermott D, Gerstein M. IQSeq: integrated isoform quantification analysis based on next-generation sequencing. PLoS One. 2012;7(1):e29175. Epub 2012/01/13. doi: 10.1371/journal.pone.0029175. PubMed PMID: 22238592; PMCID: 3253133.

32. Jee J, Rozowsky J, Yip KY, Lochovsky L, Bjornson R, Zhong G, Zhang Z, Fu Y, Wang J, Weng Z, Gerstein M. ACT: aggregation and correlation toolbox for analyses of genome tracks. Bioinformatics. 2011;27(8):1152-4. Epub 2011/02/26. doi: 10.1093/bioinformatics/btr092. PubMed PMID: 21349863; PMCID: 3072554.

33. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320(5881):1344-9. Epub 2008/05/03. doi: 1158441 [pii]

10.1126/science.1158441. PubMed PMID: 18451266.

34. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M. Global identification of human transcribed sequences with genome tiling arrays. Science. 2004;306(5705):2242-6. Epub 2004/11/13. doi: 10.1126/science.1103388. PubMed PMID: 15539566.

35. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57-63. Epub 2008/11/19. doi: 10.1038/nrg2484. PubMed PMID: 19015660; PMCID: 2949280.

36. 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. Landscape of transcription in human cells. Nature. 2012;489(7414):101-8. Epub 2012/09/08. doi: 10.1038/nature11233. PubMed PMID: 22955620; PMCID: 3684276.

37. Agarwal A, Koppstein D, Rozowsky J, Sboner A, Habegger L, Hillier LW, Sasidharan R, Reinke V, Waterston RH, Gerstein M. Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. BMC Genomics. 2010;11:383. Epub 2010/06/23. doi: 10.1186/1471-2164-11-383. PubMed PMID: 20565764; PMCID: 3091629.

38. 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. The reality of pervasive transcription. PLoS Biol. 2011;9(7):e1000625; discussion e1102. Epub 2011/07/19. doi: 10.1371/journal.pbio.1000625. PubMed PMID: 21765801; PMCID: 3134446.

39. 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. The DART classification of unannotated transcription within the ENCODE regions: associating transcription with known and novel loci. Genome research. 2007;17(6):732-45. Epub 2007/06/15. doi: 10.1101/gr.5696007. PubMed PMID: 17567993; PMCID: 1891334.

40. 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. Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. Genome research. 2011;21(2):276-85. Epub 2010/12/24. doi: 10.1101/gr.110189.110. PubMed PMID: 21177971; PMCID: 3032931.

41. 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. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol. 2010;11(10):R104. Epub 2010/10/23. doi: 10.1186/gb-2010-11-10-r104. PubMed PMID: 20964841; PMCID: 3218660.

42. 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. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome research. 2011;21(1):56-67. Epub 2010/11/03. doi: 10.1101/gr.110684.110. PubMed PMID: 21036922; PMCID: 3012926.

43. 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. The GENCODE pseudogene resource. Genome Biol. 2012;13(9):R51. Epub 2012/09/07. doi: 10.1186/gb-2012-13-9-r51. PubMed PMID: 22951037; PMCID: 3491395.

44. 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. AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol. 2011;7:522. Epub 2011/08/04. doi: 10.1038/msb.2011.54. PubMed PMID: 21811232; PMCID: 3208341.

45. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74. Epub 2012/09/08. doi: 10.1038/nature11247. PubMed PMID: 22955616; PMCID: 3439153.

46. 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. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science. 2010;330(6012):1775-87. Epub 2010/12/24. doi: 10.1126/science.1196914. PubMed PMID: 21177976; PMCID: 3142569.

47. 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. Comparative analysis of the transcriptome across distant species. Nature. 2014;512(7515):445-8. Epub 2014/08/29. doi: 10.1038/nature13424. PubMed PMID: 25164755; PMCID: 4155737.

48. 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. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012;489(7414):91-100. Epub 2012/09/08. doi: 10.1038/nature11245. PubMed PMID: 22955619; PMCID: 4154057.

49. 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. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome research. 2012;22(9):1658-67. Epub 2012/09/08. doi: 10.1101/gr.136838.111. PubMed PMID: 22955978; PMCID: 3431483.

50. Konig J, Zarnack K, Luscombe NM, Ule J. Protein-RNA interactions: new genomic technologies and perspectives. Nat Rev Genet. 2012;13(2):77-83. Epub 2012/01/19. doi: 10.1038/nrg3141. PubMed PMID: 22251872.

51. Sugimoto Y, Konig J, Hussain S, Zupan B, Curk T, Frye M, Ule J. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. Genome Biol. 2012;13(8):R67. Epub 2012/08/07. doi: 10.1186/gb-2012-13-8-r67. PubMed PMID: 22863408; PMCID: 4053741.

52. Granneman S, Kudla G, Petfalski E, Tollervey D. 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. 2009;106(24):9613-8. Epub 2009/06/02. doi: 10.1073/pnas.0901997106. PubMed PMID: 19482942; PMCID: 2688437.

53. Frey MR, Bailey AD, Weiner AM, Matera AG. Association of snRNA genes with coiled bodies is mediated by nascent snRNA transcripts. Curr Biol. 1999;9(3):126-35. PubMed PMID: 10021385.

54. Xia Y, Lu LJ, Gerstein M. Integrated prediction of the helical membrane protein interactome in yeast. J Mol Biol. 2006;357(1):339-49. Epub 2006/01/18. doi: 10.1016/j.jmb.2005.12.067. PubMed PMID: 16413578.

55. Jansen R, Lan N, Qian J, Gerstein M. Integration of genomic datasets to predict protein complexes in yeast. J Struct Funct Genomics. 2002;2(2):71-81. Epub 2003/07/03. PubMed PMID: 12836664.

56. Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M. Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends Genet. 2002;18(10):529-36. Epub 2002/09/28. PubMed PMID: 12350343.

57. Yan KK, Fang G, Bhardwaj N, Alexander RP, Gerstein M. 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. 2010;107(20):9186-91. Epub 2010/05/05. doi: 10.1073/pnas.0914771107. PubMed PMID: 20439753; PMCID: 2889091.

58. Cheng C, Yan KK, Hwang W, Qian J, Bhardwaj N, Rozowsky J, Lu ZJ, Niu W, Alves P, Kato M, Snyder M, Gerstein M. Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. PLoS Comput Biol. 2011;7(11):e1002190. Epub 2011/11/30. doi: 10.1371/journal.pcbi.1002190. PubMed PMID: 22125477; PMCID: 3219617.

59. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M. Genomic analysis of regulatory network dynamics reveals large topological changes. Nature. 2004;431(7006):308-12. Epub 2004/09/17. doi: 10.1038/nature02782. PubMed PMID: 15372033.

60. Cheng C, Fu X, Alves P, Gerstein M. mRNA expression profiles show differential regulatory effects of microRNAs between estrogen receptor-positive and estrogen receptor-negative breast cancer. Genome Biol. 2009;10(9):R90. Epub 2009/09/03. doi: 10.1186/gb-2009-10-9-r90. PubMed PMID: 19723326; PMCID: 2768979.

61. Yu H, Luscombe NM, Qian J, Gerstein M. Genomic analysis of gene expression relationships in transcriptional regulatory networks. Trends Genet. 2003;19(8):422-7. Epub 2003/08/07. doi: 10.1016/S0168-9525(03)00175-6. PubMed PMID: 12902159.

62. Qian J, Lin J, Luscombe NM, Yu H, Gerstein M. Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. Bioinformatics. 2003;19(15):1917-26. Epub 2003/10/14. PubMed PMID: 14555624.

63. Yu H, Gerstein M. Genomic analysis of the hierarchical structure of regulatory networks. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(40):14724-31. Epub 2006/09/28. doi: 10.1073/pnas.0508637103. PubMed PMID: 17003135; PMCID: 1595419.

64. Bhardwaj N, Kim PM, Gerstein MB. Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators. Sci Signal. 2010;3(146):ra79. Epub 2010/11/04. doi: 10.1126/scisignal.2001014. PubMed PMID: 21045205.

65. Bhardwaj N, Yan KK, Gerstein MB. Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(15):6841-6. Epub 2010/03/31. doi: 10.1073/pnas.0910867107. PubMed PMID: 20351254; PMCID: 2872381.

66. Bhardwaj N, Carson MB, Abyzov A, Yan KK, Lu H, Gerstein MB. Analysis of combinatorial regulation: scaling of partnerships between regulators with the number of governed targets. PLoS Comput Biol. 2010;6(5):e1000755. Epub 2010/06/05. doi: 10.1371/journal.pcbi.1000755. PubMed PMID: 20523742; PMCID: 2877725.

67. Khurana E, Fu Y, Chen J, Gerstein M. Interpretation of genomic variants using a unified biological network approach. PLoS Comput Biol. 2013;9(3):e1002886. Epub 2013/03/19. doi: 10.1371/journal.pcbi.1002886. PubMed PMID: 23505346; PMCID: 3591262.

68. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, Das J, Abyzov A, Balasubramanian S, Beal K, Chakravarty D, Challis D, Chen Y, Clarke D, Clarke L, Cunningham F, Evani US, Flicek P, Fragoza R, Garrison E, Gibbs R, Gumus ZH, Herrero J, Kitabayashi N, Kong Y, Lage K, Liluashvili V, Lipkin SM, MacArthur DG, Marth G, Muzny D, Pers TH, Ritchie GR, Rosenfeld JA, Sisu C, Wei X, Wilson M, Xue Y, Yu F, Dermitzakis ET, Yu H, Rubin MA, Tyler-Smith C, Gerstein M. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science. 2013;342(6154):1235587. Epub 2013/10/05. doi: 10.1126/science.1235587. PubMed PMID: 24092746; PMCID: 3947637.

69. Yu H, Zhu X, Greenbaum D, Karro J, Gerstein M. TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. Nucleic acids research. 2004;32(1):328-37. Epub 2004/01/16. doi: 10.1093/nar/gkh164. PubMed PMID: 14724320; PMCID: 373274.

70. Yip KY, Yu H, Kim PM, Schultz M, Gerstein M. The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. Bioinformatics. 2006;22(23):2968-70. Epub 2006/10/06. doi: 10.1093/bioinformatics/btl488. PubMed PMID: 17021160.

71. Douglas SM, Montelione GT, Gerstein M. PubNet: a flexible system for visualizing literature derived networks. Genome Biol. 2005;6(9):R80. Epub 2005/09/20. doi: 10.1186/gb-2005-6-9-r80. PubMed PMID: 16168087; PMCID: 1242215.

72. Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, Gardner K, Hillier LW, Janette J, Jiang L, Kasper D, Kawli T, Kheradpour P, Kundaje A, Li JJ, Ma L, Niu W, Rehm EJ, Rozowsky J, Slattery M, Spokony R, Terrell R, Vafeados D, Wang D, Weisdepp P, Wu YC, Xie D, Yan KK, Feingold EA, Good PJ, Pazin MJ, Huang H, Bickel PJ, Brenner SE, Reinke V, Waterston RH, Gerstein M, White KP, Kellis M, Snyder M. Comparative analysis of regulatory information and circuits across distant species. Nature. 2014;512(7515):453-6. Epub 2014/08/29. doi: 10.1038/nature13668. PubMed PMID: 25164757.

73. Sisu C, Pei B, Leng J, Frankish A, Zhang Y, Balasubramanian S, Harte R, Wang D, Rutenberg-Schoenberg M, Clark W, Diekhans M, Rozowsky J, Hubbard T, Harrow J, Gerstein MB. Comparative analysis of pseudogenes across three phyla. Proceedings of the National Academy of Sciences of the United States of America. 2014;111(37):13361-6. Epub 2014/08/27. doi: 10.1073/pnas.1407293111. PubMed PMID: 25157146; PMCID: 4169933.

74. Shou C, Bhardwaj N, Lam HY, Yan KK, Kim PM, Snyder M, Gerstein MB. Measuring the evolutionary rewiring of biological networks. PLoS Comput Biol. 2011;7(1):e1001050. Epub 2011/01/22. doi: 10.1371/journal.pcbi.1001050. PubMed PMID: 21253555; PMCID: 3017101.

75. Yan KK, Wang D, Rozowsky J, Zheng H, Cheng C, Gerstein M. OrthoClust: an orthology-based network framework for clustering data across multiple species. Genome Biol. 2014;15(8):R100. Epub 2014/09/25. doi: 10.1186/gb-2014-15-8-r100. PubMed PMID: 25249401.