A Background

Our research is focused on identifying and understanding the mechanisms that ensure the faithful inheritance of genetic and epigenetic information. Every cell cycle the entire genome must be copied in its entirety within the confines of S-phase. Duplication of the genome, once and only once per cell cycle, is accomplished by licensing and activating DNA replication origins in distinct phases of the cell cycle[?]. The origin recognition complex (ORC) bound to defined origin sequences in the budding yeast, *S. cerevisiae*, functions with Cdc6 and Cdt1 to facilitate the loading of a double hexamer of the Mcm2-7 complex onto the DNA[?]. Helicase loading is restricted to G1 and, in effect, 'licenses' the origin for activation in the coming S-phase[?]. DDK and CDK promote the recruitment of additional proteins including Cdc45 and the GINS complex culminating in the assembly of the CMG holohelicase complex[?]. Activation of the CMG holohelicase complex results in the local unwinding of DNA at the origin, priming of DNA synthesis by Pol α/primase and the onset of bi-directional DNA replication.

Helicase progression and coupling with the replisome at each fork is a tightly regulated process controlled, in part, by Rad53 and Mrc1 phosphorylation[?]. Helicase uncoupling can be induced by a number of mechanisms including limiting dNTP pools[? ?], chemical inhibition of polymerases[? ?], and by lesions that block the replisome[? ?] but not the helicase. If unchecked, helicase uncoupling can lead to significant DNA unwinding resulting in depletion of RPA and catastrophic replication failure[? ?]. For the majority of these cases, Rad53 functions as a 'dead man's switch' for limiting helicase uncoupling; however, we found in the absence of replication priming by Pol α/primase that the CMG helicase complex uncoupled from origin proximal DNA following activation in the absence of DNA synthesis[?]. Regardless of Rad53 activation status, the helicases proceed to unwind $\sim$1 kb of DNA from the origin suggesting the existence of additional mechanisms to limit helicase uncoupling.

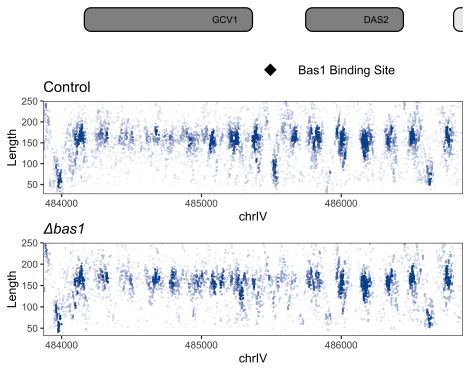

Figure 1. GCOP of the *GCV1* locus in Control and *bas1*Δ cells. Gene bodies are depicted in light and dark gray based on their orientation. Well-phased fragment centers at $\sim$150 bp represent sequences protected by nucleosomes and smaller fragments represent other DNA binding factors. The footprint at the Bas1 binding site is lost in the *bas1*Δ mutant.

A core tenet of the cell cycle is that the entire genome must be duplicated once and only once during S-phase. Failure to completely copy the genome may result in a catastrophic failure to segregate the chromosomes and may ultimately lead to chromothripsis[?] via the formation of chromosome bridges and micronuclei[?]. Increasing evidence suggests that there are mechanisms to resolve and repair unreplicated gaps during mitosis that may occur from termination defects of two converging forks[?] or at the latest replicating sequences[?]. A consequence of uncoupling the helicase from DNA synthesis at origins[?] is the generation of short unreplicated gaps at replication origins that must be resolved and repaired prior to completing mitosis. We will be able to identify and define the mechanism(s) required to resolve these unreplicated gaps and preserve genomic integrity.

Replication-coupled chromatin assembly is critical for preserving epigenetic memory via the local inheritance of parental histone $(H3\text{-}H4)_2$ tetramers[?]. However, the local chromatin environment influences the extent of local inheritance, with late replicating heterochromatin domains exhibiting the highest levels of local histone $(H3\text{-}H4)_2$ tetramer recycling[?]. While we know many of the histone chaperones involved in the deposition of nascent and parental histone $(H3\text{-}H4)_2$ tetramers[?], we are just now beginning to understand how they facilitate deposition to either the leading or lagging strands[? ? ?] which can result in extreme examples of asymmetric histone inheritance as observed in *Drosophila* male germ line stem cells[? ?]. We will define the spatiotemporal kinetics of nucleosome deposition for a number of key histone chaperones.

We developed and continue to extend genome-wide chromatin occupancy profiling (GCOP)[? ? ?] to explore how the local chromatin structure facilitates transcription and DNA replication. GCOPs provide a factor agnostic near nucleotide resolution view of proteins bound to DNA. Briefly, total chromatin is digested with micrococcal nuclease (MNase) and all of the recovered DNA fragments are subjected to paired-end next-generation sequencing. Importantly, the size of the protected fragment reveals whether it was protected by a nucleosome ($\sim$147 bp) or a DNA-binding factor ($<$50 bp). This assay is factor agnostic and reveals specific footprints for more than 70% of the yeast DNA-binding factors[?]. This specificity is demonstrated for the transcription factor Bas1 (**Figure 1**) where a clear loss of protected small DNA fragments is observed at a Bas1 binding site in *bas1*Δ cells.

B Recent Research Progress

Our work over the last four years has described, at nucleotide resolution, the chromatin occupancy dynamics associated with the initiation of DNA replication[?,?], gene expression in response to an environmental stress[?], DNA repair by non-homologous end joining[?], and the re-assembly of chromatin behind the replication fork[?].

Helicase activation in the absence of DNA replication To investigate the transient chromatin changes that occur at origins of DNA replication during initiation, we generated a conditional allele of the largest subunit of Pol α, *cdc17-ts-FRB*, that blocked replication priming at the non-permissive conditions[?]. At the onset of S-phase, in the absence of priming, the Cdc45-Mcm2-7-GINS holohelicase (CMG) complex activated and proceeded to unwind origin proximal DNA and disrupt chromatin organization for approximately 1 kb surrounding each activated origin (**Figure 2**). The CMG helicase stalled at sequences with elevated GC-content suggesting that a replisome competent for replication is required for processivity. A consequence of activating the CMG helicase in the absence of priming is that the origin DNA between the stalled helicases may re-anneal to form duplex DNA and that once priming is restored the stalled CMG helicases are oriented to progress away from the origin potentially leaving unreplicated gaps in their wake. Consistent with this hypothesis, we found that cells are able to recover from the restrictive conditions, but experience a prolonged delay in G2/M and reduced copy number at replication origins. Ultimately, the cells are viable and fully recover suggesting that there are mechanisms to resolve these unreplicated lesions.

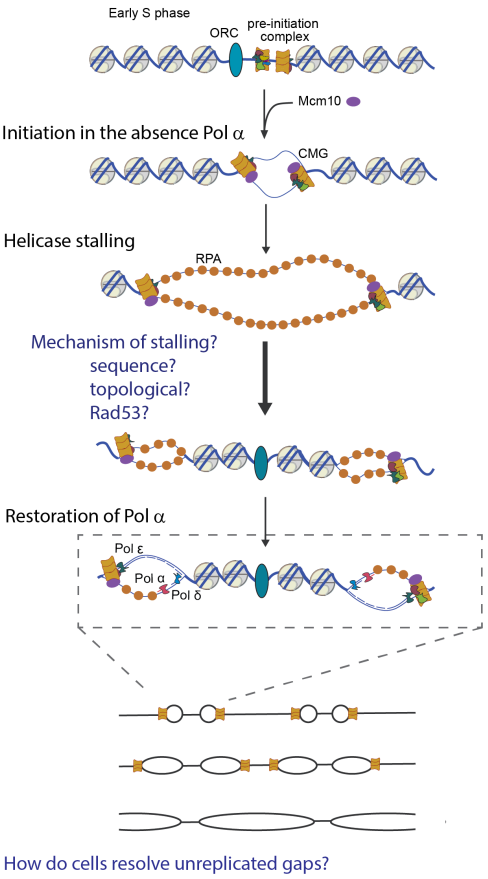

Figure 2. Model for helicase activation and stalling in the absence of active DNA replication. Following restoration of DNA replication priming, cells are able to complete S-phase but leave unreplicated gaps at the origins and exhibit a pronounced G2/M delay suggesting active mechanisms to resolve the unreplicated gaps.

Origin chromatin dynamics through multiple cell cycles We generated GCOPS for synchronized yeast cells as they progressed through two complete cell cycles[?]. With ten minute resolution we were able to visualize the chromatin occupancy dynamics at origins and the sequences surrounding origins as they progressed through consecutive cell cycles. We were able to observe the downstream re-positioning of the +1 nucleosome flanking origins and an increase in protected fragments corresponding to assembly of the pre-replicative complex in G1. Models for origin efficiency have included the strength of ORC binding[?], the numbers of Mcm2-7 loaded at origins[?] and the rate limiting concentrations of origin firing components[?,?]. In an unbiased manner, we found that the strongest correlation between chromatin occupancy at the ACS and origin efficiency occurred in early S-phase with the formation of the Cdc45-Mcm2-7-GINS (CMG) complex supporting the role of rate limiting factors in establishing the replication program.

Modeling gene expression from chromatin occupancy data We have had a long standing collaboration with Dr. Alex Hartemink (Duke) to model gene expression from changes in chromatin occupancy. We used exposure to the heavy metal cadmium to induce a stress response in yeast cells and simultaneously captured chromatin occupancy and gene expression data[?]. We were able to identify chromatin based signatures for transcriptional activation or repression from the occupancy differences in nucleosomes and or DNA binding factors. From these signatures, we were able to generate predictive models of gene expression that rivaled models based on chromatin modification data. We also developed RoboCOP, a multivariate state space model that integrates chromatin accessibility data (MNase-seq, ATAC-seq, DNase-seq) and sequence, to jointly compute a robust probability estimate for nucleosome and transcription factor occupancy[?].

Chromatin assembly behind the replication fork We developed nascent genome-wide chromatin occupancy profiling to assess the spatiotemporal dynamics of chromatin assembly at nucleotide resolution behind the DNA replication fork[?]. Every cell division, the complex regulatory landscape that defines the epigenetic state of the

cell must be disassembled and re-assembled in the wake of the replication fork. We identified locus specific differences in the assembly and maturation of chromatin behind the fork with the histone variant H2A.Z being predictive of genes with rapid maturation kinetics. We also observed differential chromatin assembly kinetics at replication origins depending on whether they underwent initiation or were passively replicated by forks from distal origins. Finally, we found that some DNA binding proteins were transiently associated with nascent chromatin following passage of the fork which may contribute to gene regulation if not evicted.

Chromatin dynamics associated with double-strand break and repair We investigated the chromatin dynamics following induction of a site specific double-strand break and the subsequent repair of the break by non-homologous end joining (NHEJ)[?] . We found that the nucleosomes flanking the break were immediately evicted in an Mre11-dependent manner and that nucleosomes more distal to the break were translocated away from the site of damage unless blocked by a transcription factor. We allowed the cells to repair the genetic lesion by NHEJ and asked whether the cells were able to repair the epigenetic lesion (disrupted nucleosome organization) without passage through S-phase. We found that S-phase was not required to restore the precise nucleosome organization at the repaired lesion and that replication independent chromatin assembly mechanisms must exist to restore the epigenetic landscape at sites of DNA damage and repair.

Collaborative work We have been fortunate to participate in a number of exciting collaborations based, in part, on our expertise in molecular biology, genomics and DNA replication. In collaboration with Dr. Chris Counter (Duke) we used maximum depth sequencing[?] to identify initiating oncogenic mutations in RAS alleles with extreme sensitivity ($\sim$1 mutation/million templates)[?] . Together with Dr. Robert Duronio (UNC), we characterized the replication timing of sequences throughout the *Drosophila* genome from developing wing imaginal disc cells and the role of repressive chromatin marks (H3K9me3) in establishing the late DNA replication program[?] . Finally, with Dr. Antonio Bedalov (FHCRC) we were able to demonstrate that active transcription of a RNA pol II non-coding transcript at the rDNA was able to push and relocalize the Mcm2-7 complex away from the initial loading site into a region of decreased nucleosome occupancy[?] . The relocalization of the helicase by the transcription machinery resulted in increased rDNA origin efficiency. This work supports prior work from our group[?] and the Remus lab[?] demonstrating plasticity in origin selection by transcription mediated displacement of the Mcm2-7 helicase.

C Overview of Future Research

We will continue to work on and address major questions in DNA replication, chromatin assembly, and gene regulation. This work will provide insights into the mechanisms that ensure genomic integrity and the inheritance of epigenetic information. In addition, we will continue to build robust probabilistic models that shine light into the 'black box' of chromatin-mediated gene regulation. We are committed to rigorous and reproducible research. All experiments are performed at a minimum in biological replicate and robust statistical approaches are utilized to assess significance. To ensure the reproducibility of our bioinformatic approaches, all of our analyses from raw data to figure are scripted and publicly available on Duke's GitLab.

C.1 Chromatin assembly behind the fork

Passage of the DNA replication fork through chromatin results in the transient disassembly of nucleosomes at the fork and their re-assembly on the two nascent daughter strands behind the fork. Chromatin assembly on nascent DNA is a complex and regulated process that is critical for preserving epigenetic information[? ?] . A major question is understanding how specific histone chaperones contribute to and facilitate the spatiotemporal kinetics of nucleosome deposition and chromatin assembly behind the replication fork.

In preliminary experiments, we have focused on the role of the Caf-1 complex which is involved in the replication-dependent deposition of nascent histone $(H3\text{-}H4)_2$ tetramers during S-phase[? ?] . In yeast cells, loss of *CAC1*, which encodes the largest subunit of the Caf-1 complex results in defects in silencing[?] , increased sensitivity to DNA damage[?] , and elevated cryptic transcription[?] . However, despite these phenotypes there are very little observed differences in the steady state distribution and occupancy of nucleosomes throughout the genome[?] . We used our nascent GCOPs to characterize the spatiotemporal dynamics of chromatin assembly behind the replication fork in wild type and *cac1*Δ cells (**Figure 3**). We found that fully mature chromatin (40 minute chase) was nearly indistinguishable between wild type and *cac1*Δ cells. Despite the similarities in mature chromatin we observed dramatic differences in the kinetics of chromatin maturation. Wild type cells were able to

rapidly re-establish nucleosome occupancy and organization (10 min chase); however, the chromatin of *cac1*Δ cells was significantly more disorganized during the pulse and early chase periods. Strikingly, we observed that the deposition rate of individual nucleosomes was heterogeneous, with nucleosomes being deposited with either 'fast' or 'slow' kinetics.

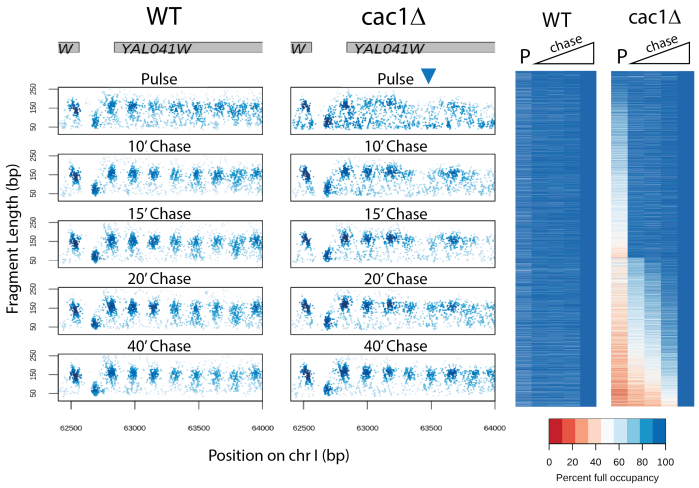

Figure 3. Nascent GCOPs reveal heterogeneous nucleosome deposition in *cac1*Δ cells. Nucleosome assembly dynamics for the *YAL041W* locus are shown for wild type and *cac1*Δ cells. In addition to the delayed assembly, the blue triangle denotes a nucleosome that exhibits a 'slow' deposition. The heatmap on the right depicts the fraction of nucleosome occupancy relative to mature chromatin for ∼50,000 nucleosomes over the assembly time course (pulse, 10, 15, 20, 40 minute chase).

We are interested in understanding the mechanisms by which individual nucleosomes are deposited with heterogenous kinetics. For example, the 'slow' nucleosomes may represent nascent histone $(H3\text{-}H4)_2$ tetramers that are deposited considerably after passage of the replication fork by replication independent mechanisms, or perhaps they represent $(H3\text{-}H4)_2$ tetramers that are deposited with the fork, but fail to incorporate two dimers of H2A/H2B to complete and stabilize the histone octamer. We will use ChIP-exo with antibodies directed against H3K56Ac (nascent) or H3K4Me3 (mature) to precisely map the location and deposition of nascent and mature histone octamers and $(H3\text{-}H4)_2$ tetramers in synchronized cells following release into S-phase to begin to distinguish among these models. Alternatively, biochemical and structural studies suggest that the Caf-1 complex forms a scaffolding structure with DNA via the K/E/R and WHD domains of Cac1 which, while not required for the *in vitro* deposition of $(H3\text{-}H4)_2$ tetramers, show defects in replication coupled nucleosome assembly assays[??]. We will delete the K/E/R and WHD domains of Cac1 to test the hypothesis that the DNA interactions of Cac1 may promote the deposition of $(H3\text{-}H4)_2$ tetramers in sequences that are recalcitrant to nucleosome formation[?] to ensure proper nucleosome phasing.

We hypothesize that the observed *cac1*Δ phenotype of elevated cryptic transcription[?] is not due to differences in steady state chromatin architecture, but rather arises from the marked differences in chromatin assembly during S-phase. For example, the increase in cryptic transcripts observed in *cac1*Δ cells may be due to the transient exposure of cryptic promoters at the locations of 'slow' nucleosome deposition during chromatin maturation. We will explicitly test this hypothesis by using TSS-seq to capture the 5' capped end of nascent mRNAs[??]. We expect to observe increased cryptic transcription initiating from the locations of 'slow' nucleosomes.

The inheritance of epigenetic state behind the fork is facilitated by the local recycling and deposition of parental $(H3\text{-}H4)_2$ tetramers to the nascent DNA strands. A number of replisome components function as histone chaperones at the fork including Mcm2[???], Pol ϵ[?], Pol α[?], and RPA[?]. Interestingly, these chaperones work to balance the distribution of parental $(H3\text{-}H4)_2$ tetramers between the two daughter strands formed by leading and lagging strand synthesis. Mcm2, for example, facilitates the transfer of a parental $(H3\text{-}H4)_2$ tetramer to the lagging strand while in a *mcm2-3A* mutant there is a relative increase in parental $(H3\text{-}H4)_2$ tetramers on the leading strand[?]. We have recently adopted our nascent GCOP assay to provide strand specific data and can readily observe the preferential deposition of $(H3\text{-}H4)_2$ tetramers and histone octamer formation on the leading strand in *mcm2-3A* cells (**Figure 4**). In contrast to prior ChIP based methodologies (*e.g.* eSPAN[??]), our methodology is not dependent on cell synchronization and release into hydroxyurea, provides a quantitative single locus view of strand-specific deposition throughout the genome, and importantly reveals how the nucleosomes ultimately 'fill-in' the gaps on the lagging or leading strands during chromatin maturation. A caveat is that we are just observing differences in nucleosome occupancy on the leading and lagging strands and are not able to distinguish between parental and nascent histones. However, it is clear from the loss of nucleosome occupancy on the lagging strand that the rate of nascent $(H3\text{-}H4)_2$ tetramer deposition on the lagging strand is insufficient to keep up with the preferential deposition of parental histones to the leading strand. We will continue to probe the different histone chaperones alone and in combination (*e.g. cac1*Δ *mcm2-3A*) to characterize the different kinetics of nucleosome

deposition on the leading and lagging strands and how they ultimately achieve steady state levels of nucleosome occupancy and whether the differential kinetics of nucleosome deposition for the leading and lagging strands results in the propagation of different chromatin landscapes between the daughter strands of DNA.

C.2 DNA replication and genome integrity

Coupling of helicase activity with the replisome and active replication is critical for maintaining genomic stability. Uncoupling of the helicase from active DNA replication can result in the generation of ssDNA tracts which are not only susceptible to DNA damage[?], but can also lead to replication catastrophe via the sequestration of RPA[???]. We induced helicase uncoupling at the origin using a very tight conditional allele of *CDC17*(*POL1*), *cdc17-ts-FRB*, to prevent RNA priming by Pol α/primase[?]. As the *cdc17-ts-FRB* cells entered S-phase in the absence of DNA replication, the CMG helicase complex was activated and proceeded to unwind approximately 1 kb of DNA surrounding the origin before the holohelicase complex stalled.

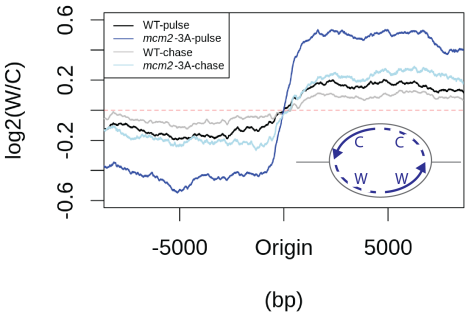

Figure 4. Strand specific nascent GCOPs reveal preferential deposition of nucleosomes on the leading strand in a *mcm2-3A* mutant. Specifically, we recover more nucleosome fragments associated with the nascent leading strand which maps to the Watson strand on the right of the origin and to the Crick strand on the left of the origin.

We will identify the mechanism(s) that function as a 'dead man's switch' to limit helicase progression in the absence of DNA replication. Helicase progression may be limited by sequence, topological constraints or Rad53-mediated phosphorylation. We found that the CMG helicase complex stalled in regions of elevated GC content relative to the origins of DNA replication[?]. We will explicitly test the role of sequence features in limiting helicase movement by inserting synthetic sequences of approximately 1 kb with differing GC-content adjacent to an early efficient origin. Alternatively, topological constraints may limit helicase progression in the absence of active DNA replication. We will manipulate the levels of yeast topoisomerases, Top1 and Top2, by chemical inhibition, conditional depletion (*e.g.* degron or anchor away), and over-expression and assess helicase progression. Rad53 has been linked to helicase speed and fork progression[?] as well as limiting helicase uncoupling[??]. Notably, we did not observe a difference in helicase progression in the absence of *MRC1* and activated Rad53[?]. Similarly, early reports on DNA unwinding in Pol α and primase mutants also observed significant unwinding in the absence of DNA replication[?], suggesting that Pol α/primase may be required to recruit Rad53 to the replisome to regulate progression. In support of this hypothesis the Remus group demonstrated that Rad53 regulated helicase uncoupling is independent of its interaction with Cdc45 and Mrc1[?]. We will use ChIP-seq to assess the recruitment of Rad53[?] to sites of helicase uncoupling induced by preventing Pol α mediated priming. If we observe a dependence on Pol α for the recruitment of Rad53, we will dissect the interaction. Together, these experiments will provide valuable insights into the mechanism(s) that regulate helicase progression in the absence of DNA replication.

A consequence of the CMG-helicase complexes stalling on either side of the origin is that upon restoration of priming activity, the helicases will be oriented to unwind DNA away from the origin, potentially leaving a stretch of unreplicated duplex DNA at the origin. Consistent with this hypothesis, cells with restored priming experienced a prolonged delay in G2/M due to the presence of unreplicated DNA at the origin[?]. However, this delay is transient and the cells are eventually able to recover and resume normal cell cycling, suggesting that mechanism(s) must exist to deal with short unreplicated gaps. It is also worth noting that the unreplicated gaps we observe at origins are structurally similar to DNA intermediates that would be produced from defects in helicase unloading and replication termination upon fork convergence[?]. Our ability to readily produce these intermediates at specific locations in the genome will enable us to readily identify the ensemble of factors involved in their resolution.

We will identify the mechanism(s) by which the cells are able to resolve unreplicated gaps in their chromosomes. We hypothesize that alternative helicases will be recruited to the unreplicated duplex DNA to facilitate unwinding and synthesis. We will start with a targeted genetic screen of non-replicative helicases (*e.g.* Pif1, Rrm3, Sgs1, Srs2, Mph1) involved in DNA repair and maintenance. We will analyze the ability of cells (*e.g. cdc17-ts-FRB pif*Δ) to recover from the transient loss of replication priming. Our preliminary data indicate that cells with short unreplicated gaps at origin sequences require Pif1 for viability and that in the absence of Pif1 they accumulate in early anaphase with a failure to segregate their chromosomes. This is reminiscent of elegant *in vitro* work from the Labib group which found that Pif1 was required for the termination of DNA replication at

converging forks[?]. We will continue to screen additional helicases as there will likely be multiple factors involved in resolving these structures. We will use live imaging (LoS Lew) to quantify the delay in progression through mitosis for each of the mutant strains as well as to observe specific mitotic defects like anaphase bridges. Together, these experiments will provide important insights into how the cell responds to and repairs unreplicated gaps which could arise from defects in initiation or termination of DNA replication.

C.3 Gene regulation

A major challenge in genome biology is deciphering the complex regulatory code embedded within the chromatin landscape to model and predict gene expression. Gene regulatory networks have been inferred from the analysis of gene expression across a collection of yeast transcription factor deletion mutants[? ?]. Despite the power of these approaches, it is difficult to determine if the observed interactions are direct or indirect. Similarly, ChIP based approaches to identify the binding sites of specific TFs are prone to false positives[?] and only reveal potential direct interactions for one factor at a time. To overcome these shortcomings, we generated GCOPs from 201 yeast strains harboring individual deletions of transcription regulators. This factor agnostic approach reveals not just the loss of chromatin occupancy at *bona fide* binding sites for specific factors (**Figure 1**), but also reveals indirect changes in chromatin including the loss and gain of other regulatory factors and altered patterns of nucleosome occupancy and organization reflective of transcription levels. We are able to recapitulate known gene expression networks solely from the changes in chromatin occupancy and nucleosome organization (**Figure 5**). We are also able to capture chromatin perturbations that are not associated with differential gene expression nor would they have been captured in prior gene expression based regulatory networks. These locus specific chromatin changes independent of gene expression may serve to prime the locus for a subsequent or future transcriptional response[?]. Together with the Hartemink group at Duke (LoS Hartemink) we are now generating larger and more complex regulatory networks that integrate chromatin perturbations with gene expression.

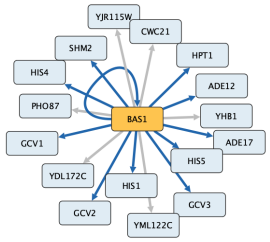

Figure 5. *BAS1* chromatin-based regulatory network. Nodes represent genes with altered chromatin in the promoter region in *bas1*Δ cells. The edges represent gene expression changes with blue being a decrease in gene expression in *bas1*Δ cells. Gray edges represent no detectable gene expression change despite a pronounced chromatin change in the promoter.

C.4 Technology development

We have continued to develop and extend our GCOP assay to describe the chromatin occupancy of nascent DNA[?] as well as strand specific nascent DNA. We would like to develop a single molecule read out of chromatin occupancy that would describe the precise chromatin occupancy status for an entire single chromosome or large chromosomal fragment with binary precision. Recent advances in next generation long read technologies (*e.g.* Nanopore, PacBio) allow the direct read out of modified bases (eg. m6A methylation)[? ? ? ?] or nucleotide analogs (*e.g.* BrdU)[? ?]. Long read sequencing coupled with the detection of m6A methylation led to the development of Fiber-seq which uses a non-specific m6A methyltransferase to methylate accessible adenine residues as a read out of chromatin occupancy[?]. Historically, the very first locus specific protein DNA occupancy maps were generated by the chemical treatment of chromatin with DMS (dimethyl sulfate) and the non-specific methylation of accessible purines[?]. More recently, DMS-seq was developed to map chromatin accessibility following the enrichment of methylated residues by immunoprecipitation and next-generation Illumina short read sequencing[?]. We propose to adapt DMS-seq for use on long read sequencing platforms with direct readout of methylated purines to provide precise chromatin occupancy maps for individual molecules of DNA. Alternatively, we will adapt Fiber-seq for the analysis of yeast chromatin. Each approach presents complex analysis challenges for interpretation of the data including the sparseness of modifications in the enzyme based Fiber-seq system and deciphering the complex signal arising from potentially methylating all exposed purines with DMS. These are not trivial challenges but our long time collaborator, Alex Hartemink, has experience developing statistical models for dealing with sparseness in single cell RNA-seq datasets[? ?] as well as developing robust sequence based models for chromatin occupancy[?] which will help guide the interpretation of the complex DMS-seq data (LoS Hartemink). A single molecule view of chromatin occupancy will allow us to address multiple fundamental questions in DNA replication, repair, and transcription including the ability to provide a quantitative assessment of ORC, helicase loading and activation at individual replication origins across individual cells.

Facilities

Laboratory

Dr. MacAlpine's laboratory occupies around 1,200 square feet in the Department of Pharmacology and Cancer Biology (PCB) within the Levine Science Research Center (LSRC). The lab space has benches for 12 investigators, a fume hood, and animal procedures. Each bench has network ports available for computers connected to the Internet. Additional space is available for large or shared equipment and tissue culture.

The environment at Duke is uniquely suited for these studies. As a member of the Duke Cancer Institute (DCI) and Center for Genome and Computational Biology (GCB), Dr. MacAlpine has access to a variety of shared resources including multiple next-generation sequencing platforms (Illumina, PacBio and Nanopore) and cluster computing environments. The campus boasts a strong and interactive cadre of faculty studying various aspects of transcriptional regulation and nucleic acid metabolism, as well as a diversity of relevant graduate programs. MacAlpine's research group is composed of students from the University Program in Genetics and Genomics, Computational Biology and Bioinformatics, Pharmacology, and Molecular Cancer Biology.

Clinical

Not applicable.

Animal

Not applicable.

Computer

The MacAlpine computer facilities are located adjacent to the MacAlpine research laboratory. These currently include an 16 processor Xenon Server with 32GB of RAM and 8TB of local storage. Dedicated Linux and Mac workstations are available for data analysis. Laser printers and scanners are also available. All computational workstations, servers, and disk space owned by GCB Investigators are managed by the GCB information technology (IT) group directed by Dr. Hilmar Lapp and staffed with four application and database developers and system administrators. The GCB provides shared resources consisting of several compute servers connected to a SAN storage system with a total capacity of about 800TB of disk storage that provides fast access to data, with built-in redundancy and daily backups to protect against data loss. Individual lab servers and disk space are integrated into the GCB infrastructure and are completely managed by the IT group.

Office

Dr. MacAlpine's office is located in the PCB Department, also within the LSRC building. His office space is immediately adjacent to his laboratory.

Other

Duke Genome Sequencing Shared Resource The Genome Sequencing Shared Resource (GSSR) has operated and made available to Duke researchers a variety of genome-sequencing instruments and related support equipment and analysis for more than 10 years. Its faculty director is Dr. Greg Wray, and it is staffed full-time by its associate director, Dr. Olivier Fedrigo, and six other employees: a bioinformatics specialist, an operations administrator, and four lab technicians. The GSSR has over 2,000 square feet of lab space, equipment rooms, and office space. All instruments are housed and maintained in the facility and are available on a fee-for-service and cost-recovery basis for academic researchers.

Dr. MacAlpine has been a client for several years, including for his production role in the modENCODE project, and has excellent working relationships with GSSR staff. The GSSR has a wide range of sequencing platforms available including Illumina MiSeq and HiSeq 2000 and 2500 sequencers, IonTorrent PGM and Proton sequencers, and a PacBio RS II sequencer. The Illumina instruments will provide the majority of our sequencing reads, but diverse other platforms are available for any sort of validation we might require along the way. Specifically, the GSSR provides a wide range of next-generation library preparations and sequencing services including DNA-seq, RNA-seq, ChIP-seq, MNase-seq, DNase-seq, smRNA-seq, and ATAC-seq, as well as mate-pair and targeted (exome, amplicon, gene panels) sequencing. Ancillary equipment includes a Beckman Coulter

Biomek FX liquid handling robot, a Beckman Coulter Z2 particle counter, a Genomic Solutions HydroShear fragmentation apparatus, Agilent 2100 Bioanalyzer and 2200 TapeStation QC instruments, an Apollo 324 NGS library prep system, a Qiagen TissueLyser, a Covaris E-Series focused-ultrasonicator, a Life Technologies StepOnePlus RT-PCR instrument, and two Qubit 2.0 fluorometers for nucleic acid quantitation, along with three servers for next-generation sequencing data primary analyses and two additional compute servers, one with 24 CPU-cores and 256GB RAM for secondary and tertiary bioinformatics analysis.

Duke Scalable Computing Support Center (SCSC) To support researchers with high-performance computing needs, Duke established the Scalable Computing Support Center (SCSC). The center is staffed with a variety of computational experts, providing researchers with a broad range of support from algorithm development, performance improvement, code parallelization, and data visualization. In conjunction with the Office of Information Technology (OIT), SCSC maintains a Linux cluster called the Duke Shared Cluster Resource (DSCR). The power, cooling, maintenance, and systems administration for all of the equipment in the DSCR, including researcher-owned machines, is handled by OIT staff. At present, the DSCR consists of around 4,300 CPU-cores, and these–plus 720 additional CS cores and 640 additional GCB cores–are available to the faculty on this project.

Duke Data Commons (NIH 1S10OD018164-01) Duke Research Computing conceived and has managed a large 1.5 petabyte storage installation of EMC Isilon equipment to support data-producing core facilities in the life sciences. The project, funded by the National Institutes of Health in fall 2014 (NIH 1S10OD018164-01), began operation in November 2014. This initiative was designed not merely to provide a place to put data, but is the beginning of a data commons for life sciences research, which has experienced explosive growth in data that are useful for researchers. Thus, as the project matures, the data storage equipment will serve as a foundation for building solutions to data management challenges, especially those arising from research using large, complex, multi-dimensional data. Duke researchers and their collaborators can use the data commons by using one or more of the core facilities to extract data for their projects. At present, these data are mostly molecular profiles (DNA or protein sequence and RNA analysis), though the data commons is also used for light microscopy and data under analysis by Duke's 'Omics Analysis Core. Researchers in the life sciences with extraordinary data storage requirements also use the resource, though the resource is particularly targeted to meet needs of NIH-sponsored researchers, as of course are all NIH Shared Instrumentation Grants (S10). Duke Research Computing and Duke's Office of Information Technology have created a plan for sustaining storage capacity and for developing computational resources that can be attached to the storage. For example, data storage for particularly sensitive data such as patient data regulated by HIPAA/HITECH has been isolated and attached to Duke's Protected Research Network where computational analysis of the data can be conducted safely and securely. A project underway in Duke's School of Medicine will begin to use the storage for data requiring high provenance particularly for data collected in biomedical research.

Major Equipment

The MacAlpine laboratory has all the equipment required to carry out the proposed research. Major equipment includes: a flow cytometer, radioactivity counters, ultra-centrifuges and centrifuges, laminar flow hoods, yeast incubators, fluorescent spectrometers, tissue culture equipment, cloning facilities, X-ray film processors, microscopes, including a Zeiss Axiophot, power supplies, HPLC and FPLC, ultralow and regular freezers, refrigerators. In addition, shared equipment is also available for use, including a phosphoimager, a confocal-laser microscope, a DNA sequencer, and a BIAcore.