REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggesstions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any oenalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY) 06-05-2016	2. REPORT TYPE Final Report	3. DATES COVERED (From - To) 1-Oct-2011 - 31-Oct-2014

4. TITLE AND SUBTITLE
Genome Sequencing to Enable a Model Salamander for Tissue Regeneration Research

5a. CONTRACT NUMBER
W911NF-11-1-0475

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
1620BM

6. AUTHORS
Stephen R. Voss

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Kentucky
500 South Limestone Street
109 Kinkead Hall
Lexington, KY 40526 -0001

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
60727-LS.3

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

Salamanders are important vertebrate model organisms in regenerative medicine. In this application, we proposed to enhance research resources for the Mexican axolotl (Ambystoma mexicanum) by innovating an approach tosequence their large 32,000,000,000 base pair genome. We successfully developed methods to isolate and sequence whole axolotl chromosomes, and published a manuscript in a leading journal that describes our approach and initial characterization of the axolotl genome. Thus, we accomplished the objectives of our project, laying the ground work for developing a full axolotl genome assembly.

15. SUBJECT TERMS
Genome Sequencing Regeneration Mexican axolotl

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Stephen Voss
a. REPORT UU	b. ABSTRACT UU	c. THIS PAGE UU	UU		19b. TELEPHONE NUMBER 859-257-9888

Report Title

Genome Sequencing to Enable a Model Salamander for Tissue Regeneration Research

ABSTRACT

Salamanders are important vertebrate model organisms in regenerative medicine. In this application, we proposed to enhance research resources for the Mexican axolotl (Ambystoma mexicanum) by innovating an approach tosequence their large 32,000,000,000 base pair genome. We successfully developed methods to isolate and sequence whole axolotl chromosomes, and published a manuscript in a leading journal that describes our approach and initial characterization of the axolotl genome. Thus, we accomplished the objectives of our project, laying the ground work for developing a full axolotl genome assembly.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received		Paper
05/06/2016	2.00	Melissa C. Keinath, Vladimir A. Timoshevskiy, Nataliya Y. Timoshevskaya, Panagiotis A. Tsonis, S. Randal Voss, Jeramiah J. Smith. Initial characterization of the large genome of the salamander Ambystoma mexicanum using shotgun and laser capture chromosome sequencing, Scientific Reports, (11 2015): 0. doi: 10.1038/srep16413
TOTAL:		1

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received	Paper
TOTAL:	

Number of Papers published in non peer-reviewed journals:

(c) Presentations

December 2015- Invited Seminar – Southeastern Louisiana University, “Axolotl Story: Enabling an Endangered Species for Human Health and Disease Research”.

December 2015- December 2015- Invited Seminar – Southeastern Louisiana University, “Genetics and Genomics of Salamander Paedomorphosis”.

September 2015- “Salamander limb regeneration”, Course Instructor, Stem Cells and Regenerative Medicine, MBL Woods Hole, MA.

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper

TOTAL:

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received		Paper
08/18/2015	1.00	Melissa C. Keinath, Vladimir A. Timoshevskiy, Nataliya Y. Timoshevskaya, Panagiotis A. Tsonis, S. Randal Voss, Jeramiah J. Smith. Initial Characterization of the large genome of the Mexican Axolotl (Ambystoma mexicanum) using laser capture and whole chromosome sequencing, IN submission (06 2015)
TOTAL:		**1**

Number of Manuscripts:

Books

Received Book

TOTAL:

Received Book Chapter

TOTAL:

Patents Submitted

Patents Awarded

Awards

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Claudia Arenas	0.01	
Mellisa Keinath	0.11	
Qingchao Qiu	0.11	
Nour Al Haj Baddar	0.11	
FTE Equivalent:	**0.34**	
Total Number:	**4**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Jeramiah J Smith	0.11	
Stephen R Voss	0.00	No
FTE Equivalent:	**0.11**	
Total Number:	**2**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:...... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):...... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:...... 0.00

Names of Personnel receiving masters degrees

NAME

Total Number:

Names of personnel receiving PHDs

NAME

Total Number:

Names of other research staff

NAME PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

Final Report: The objective of this project was to originate an approach to sequence whole salamander chromosomes and verify the approach by sequencing one of the 14 Mexican axolotl chromosomes. We accomplished this objective. In December 2015, we published a paper in the journal Scientific Reports that details our approach and results, which include assemblies for more than one chromosome. These assemblies allowed us to place a total of 2062 previously characterized genes and ESTs on the two smallest chromosomes and identify genomic sequences that flank these genes. Comparative genomic analyses performed in this paper reveal strong conservation of gene content between salamander and other vertebrates (as expected) and lend further support to the idea that our assemblies represent single chromosomes.

The current genome assembly covers >26 Gb of the salamander genome, with an N50 scaffold length of 19.7 Kb. Ongoing efforts are aimed at further improving contiguity, including the integration of chromosome sequencing information and ultradense linkage mapping information, and building and sequencing HiC libraries. The work accomplished under this grant will greatly accelerate genomic and genetic work in the tissue regeneration field because the Mexican axolotl is one of the primary model organisms.

Technology Transfer

Summary of Major Accomplishments in Year 1

During the first year of the project we optimized protocols for isolating, spreading, and capturing chromosomes for library construction and DNA sequencing. We found that 20 chromosomes (dyads) yields approximately 1.5 micrograms of amplified DNA, an amount that should be sufficient for library prep and sequencing via Illumina HiSeq technology. We contacted scientists at the Craig Venter Institute (CVI) because we had heard through a third party that they had developed expertise in making libraries for DNA sequencing in cases where the amount of starting material is limiting. Also, they quoted a very low price to make the libraries because they wanted to engage us in a lasting partnership. We sent CVI two DNA samples under the idea that they would make two libraries using different approaches, with the goal of identifying the best approach to move forward. Neither approach they tried was successful and thus after investing 4 months of time in this collaboration, we decided to try a different vendor. We were not charged by CVI for this collaboration. We isolated additional dyads and delivered these to Hudson Alpha Genomic Services Laboratory for library preparation and sequencing. We are currently awaiting the sequencing results from this experiment.

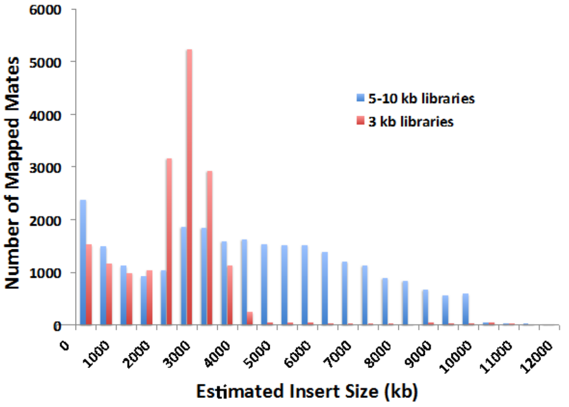

Figure 1 – Observed fragment length distributions for mate pair libraries, based on mapping to assembled BAC sequences. Observed distributions match expected distributions for these "long" fragment libraries.

	3KB	10KB
Total #reads	8628692	17297824
Reads mapped to transcriptome	10955	25433
Total bp (reads)	706436083	1244522971
Mapped bp	24283	52829
Avg read length	81.9	71.9
% of reads mapped to transcriptome	0.13%	0.15%
% of transcripts mapped	21.3%	31.0%

Table 1 – Read mapping statistics for alignments between 3 kb and 5-10 kb fragment libraries and the salamander transcriptome (genes).

Concurrent with this aspect of the project, we generated whole genome sequence data for three different "long" fragment libraries (3kb, 5-10 kb and 40kb). The 3kb and 5-10kb mate pair libraries were produced and sequenced by Ion Torrent for free. They did not charge us because they too would like to be involved in our project and they value our skills in assessing new products that they are currently developing. In this case, they are developing new ways to make 3 and 10kb jumping libraries for Illumina sequencing. By mapping these fragments to our previously assembled BAC sequences, we verified that the libraries are of high quality and validated the fragment size distributions. These libraries have also yielded new scaffolding information that has resolved a few remaining linkages among previously assembled BAC contigs that were not resolved by shotgun sequencing alone.

We also mapped the 3 and 5-10 kb libraries to our collection of cDNA contigs (Table 1; Figure 2). We found that 20% and 30% of our contigs have one or more reads in the 3 and 5-10 kb libraries respectively, and thus a total of > 40% of all contigs are covered by at least one of the libraries. These observed percentages meet our expectations. For example, assuming that the number of coding bases is similar between salamander and human, one would estimate that roughly .2% of the genome should be coding, which approximates the percentage of cDNA contigs with mapped reads (0.13 and .15% from the 3 and 5-10 kb libraries, especially if you consider that intron/exon breaks will tend to disrupt these sort of alignments

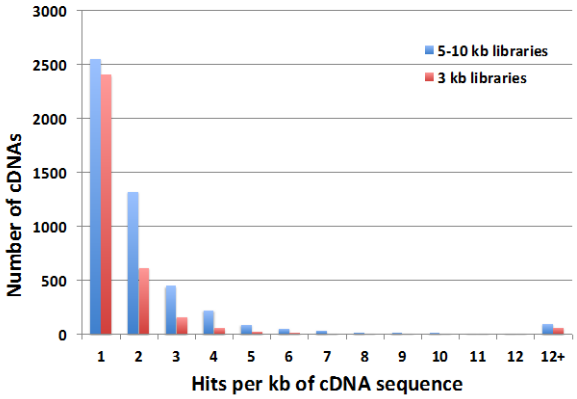

Figure 2 – Distribution of hits from 3 kb and 5-10 kb libraries when aligned to the salamander transcriptome. Most aligned genes have a single hit, while smaller numbers have more hits, consistent with expected patterns expected for non-repetitive sequence. Only a small fraction of mappings are consistent with repetitive/ duplicated sequence.

A 40kb (fosmid ditag) library was also generated and evaluated; QC analyses were performed by mapping the resulting reads to assembled BACs and the axolotl transcriptome. The results show that the library is of high quality and yields useful long-range scaffolding information. Fosmid pools that were produced for this library have also been preserved as freezer stocks and thus present additional genomic resources that can be tapped for targeted sequencing studies. This library was generated from start-up funds to Jeramiah Smith and thus our budget was spared this expense too.

In summary, we accomplished a lot in the first year without having to expend much money. This will obviously change in the second year as we scale up for deep sequencing runs that will allow us to assemble a whole salamander chromosome. We anticipate that sequencing will be completed in the next year and we should also begin genome assembly and annotation. Thus, we are on schedule to accomplish the objectives of the project.

Summary of Major Accomplishments in Year 2

We optimized protocols for amplifying salamander chromosomes and generated pilot sequencing runs. These pilot runs demonstrate that our whole genome amplification (WGA) approach generates specific sequence information from targeted chromosomes, validate previous genetic studies, and further resolve the structure of the salamander genome. Thus far we have generated 30Gb of sequence from three target chromosomes. These sequences validate our prediction that the NOR-containing chromosome (Chromosome 2) corresponds to linkage group 3. While this was previously inferred on the basis of association with the *white* mutant with NOR length variants and linkage mapping of the *white* locus, this association is made abundantly clear by direct comparisons between WGA sequence and the sequences that have been anchored to the salamander genetic map. These experiments also reveal that a smaller linkage group (LG13: a candidate target for this project) is physically linked to linkage group 3 (Chr 2) and verify previous studies indicating that another small linkage group (LG11) is physically linked to a larger linkage group (together comprising Chr3).

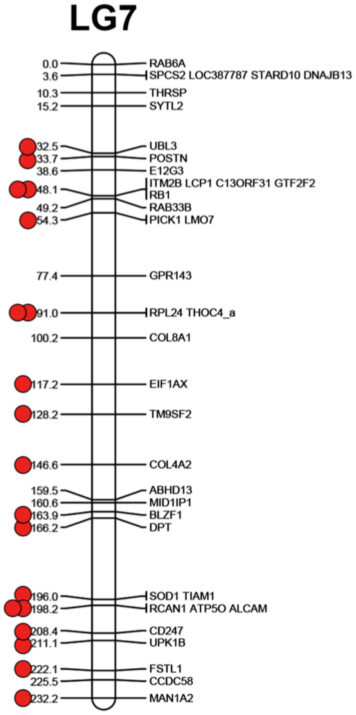

Figure 1 – Genes sampled from sequencing of material derived from a single dyad of the chromosome targeted by this project (Chr14/LG7).

We have identified a third, small and physically distinguishable, chromosome as the target of the current study. Our initial pilot runs have generated ~3X sequence coverage of this chromosome, robustly identify the linkage group that corresponds to this chromosome and provide sampling of the majority of the loci that are known to reside on the chromosome (Figure 1). We are beginning production sequencing runs of this target chromosome that include several technological advances that were gained through our pilot runs. We anticipate that sequencing runs will be completed next quarter and that all sequence resources will be in place to initiate assembly by the last quarter of 2013.

As mentioned above, these studies have also led to significant advances in our ability to sequence individual chromosomes. We are working directly with Genomic Services Laboratory at Hudson Alpha (Huntsville, AL) to increase the fidelity of WGA sequence. Through these efforts we have found that the inclusion of a short leader sequence dramatically improves cluster calling and base accuracy as reads traverse low complexity sequence derived from WGA amplification adapters. These findings have spurred a collaboration with Rubicon Genomics targeted at improving these same aspects of WGA library preparation.

We have also completed two other studies in support of the assembly of our target chromosome, specifically aimed at resolving mid-range linkages (4-40 kb):

1) We have outsourced the development of a fosmid library by Lucigen Corporation, providing blood drawn from the female salamander that is being used for BAC development. This project provides resolution of linkages at the 40kb scale, and appears to provide excellent representation of genic regions (Figure 2).

2) We have generated pilot sequence data for the salamander genome (~0.2X) depth using long read technology recently released by Pacific Biosciences. We are currently evaluating the utility of this resource, but it appears that this technology will significantly improve our ability to bridge repetitive sequences in genic regions. This project provides resolution of linkages at the 2-10kb scale (mean 4.5kb).

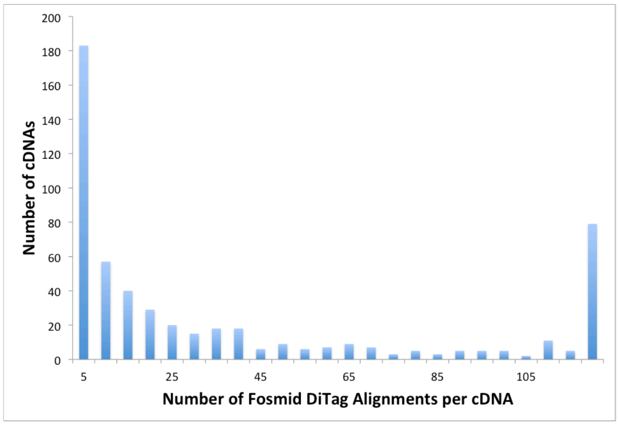

Figure 2 – Sequenced fosmid ditags provide representation of a large number of genic regions. Several known salamander genes (cDNAs) are represented by one or more fosmid end sequence.

Summary of Major Accomplishments in Year 3

We have collaborated with Rubicon Genomics and the Genomic Services Laboratory at Hudson Alpha in order to optimize the PicoPlex whole genome amplification kit to incorporate a modified leader that permits cluster identification, Illumina-specific priming sequences, and multiplex barcodes. Besides improving our ability to sequence the amplified chromosome fragments, including library preparation in our amplification protocol decreased the overall sequencing cost and has permitted us to generate additional sequence data that improve our ability to assemble our target chromosomes. Through beta testing of this kit, we were able to sequence two sample pools of 10 Nucleolus Organizer Region (NOR) containing dyads covering 89.34% of LG3. In addition to the larger pools of NOR-containing dyads, we sequenced 2 pools of 5 NOR-containing chromosomes, 2 single NOR-containing chromosomes, 3 other single chromosomes, and a human control provided by Rubicon Genomics.

Among the single dyads sequenced, reads from one sample were enriched on two different LGs, 4 and 13, and corresponds roughly to the fourth largest chromosome in the *Ambystoma* karyotype. The sampling of markers suggests that portions of each LG correspond to one chromosome. The reads mapped to LG4 were clustered downstream of a large marker-free interval in the distal 138 cM of the linkage group. A previous study predicted the linkage of these segments to form a conserved synteny with segments of *Xenopus tropicalis* chr9 and chicken chr 3. In addition to providing LG-specific libraries, the samples from these tests yielded the smallest fraction of alignments to the human genome (>7% each). Using the same kit, we were able to sequence 12 single dyad samples of the two smallest chromosomes in the *Ambystoma* karyotype. Four of those samples specifically mapped to and recovered 73% of the genetic markers in LG15 and 17. Two other samples specifically mapped to LG14 and recovered 78.9% of LG14 genetic markers.

We have continued working with HAIB GSL and Rubicon Genomics to test a new WGA kit (PicoPLEX ™ DNA-seq) that includes a dual indexing plate. We have performed extensive testing of this kit and have identified optimal clustering and base calling parameters. Our recent test runs have generated data for several chromosomes, including those that correspond to LG2, LG5, LG6, LG9, LG10 and the half of LG4 not previously sampled. Further sequencing of the library that mapped specifically to LG2, containing the metamorphic timing QTL (*met1*) for the species, recovered 80.1% of genetic markers from LG2 and allowed us to assign LG2 to approximately the second largest chromosome of the *Ambystoma* karyotype. Additional sequencing of the single dyad that mapped to LG9, which contains the amphibian sex-determining locus, *ambysex*, recovered almost 70% of LG9 genetic markers.

In order to proceed with assembly of individual chromosomes, we have also generated ~1.5 TB of whole genome shotgun (WGS) data that will be used to improve our chromosome-specific assemblies by permitting the correction inherent base calling errors associated with sequencing of whole chromosome amplified DNA. We are currently evaluating several error correction and assembly pipelines and expect to identify optimal assembly parameters within the coming month. After completion of the initial chromosome-specific assemblies, we will incorporate additional scaffolding information from long insert libraries. To this end, we recently outsourced the development and QC of a 4 kb insert library. Sequencing of this library will be initiated at the onset of Year 4.

We have updated Sal-Site (www.ambystoma.org) to make the community aware of this project and our progress. Below, we provide examples of this progress. The figures show the mapping of short-sequence reads generated from shotgun sequencing to genetic markers of the Ambystoma linkage map. The red and blue circles present results from two separate chromosome sequencing experiments.

Summary of Major Accomplishments in Extension Year 4

As we near the end of our project, we restate the objectives of our project: Generate datasets and develop computational approaches to assemble and analyze salamander chromosomes and the overall genome. In the last year we completed sequencing and an initial assembly of chromosome 14 and submitted a manuscript detailing the assembly and analysis of this and one other chromosome (in revision, Scientific Reports). We uploaded this manuscript in support of this progress report. Our assemblies have allowed us to uniquely place a total of 2062 previously characterized genes and ESTs on the two smallest chromosomes and identify genomic sequences that flank these genes. Comparative genomic analyses reveal strong conservation of gene content between salamander and other vertebrates (as expected) and lend further support to the idea that our assemblies represent single chromosomes. As part of these sequencing efforts, we have also identified high quality libraries corresponding to an additional six chromosomes.

During the last year we also generated a whole genome shotgun sequence dataset for the axolotl (using both long- and short- insert mate pair reads). This turned out to be an incredibly valuable dataset because it facilitated error correction of reads that were generated from amplified chromosomes and it permitted the first genome-scale characterization of repetitive content. However, we have found that existing algorithm/computational platforms are incapable of assembling this large genome (in revision, Scientific Reports). To resolve this issue, we recruited a talented programmer and have implemented several modifications to an existing assembler (Sparse) that recently achieved the first successful assembly of the salamander genome. The current genome assembly covers >26 Gb of the salamander genome, with an N50 scaffold length of 19.7 Kb. Ongoing efforts are aimed at further improving contiguity, including the integration of chromosome sequencing information and ultradense linkage mapping information. We are also currently assessing the capabilities of an assembler that was recently developed recently developed to leverage extreme scale architectures (meraculous) in collaboration with the Rokhsar group at the Joint Genome Institute.

Final Report: The objective of this project was to originate an approach to sequence whole salamander chromosomes and verify the approach by sequencing one of the 14 Mexican axolotl chromosomes. We accomplished this objective. In December 2015, we published a paper in the journal Scientific Reports that details our approach and results, which include assemblies for more than one chromosome. These assemblies allowed us to place a total of 2062 previously characterized genes and ESTs on the two smallest chromosomes and identify genomic sequences that flank these genes. Comparative genomic analyses performed in this paper reveal strong conservation of gene content between salamander and other vertebrates (as expected) and lend further support to the idea that our assemblies represent single chromosomes.

The current genome assembly covers >26 Gb of the salamander genome, with an N50 scaffold length of 19.7 Kb. Ongoing efforts are aimed at further improving contiguity, including the integration of chromosome sequencing information and ultradense linkage mapping information, and building and sequencing HiC libraries. The work accomplished under this grant will greatly accelerate genomic and genetic work in the tissue regeneration field because the Mexican axolotl is one of the primary model organisms.