Continuation/Research Progression Projects Form (7)

Required for projects that are a continuation/progression in the same field of study as a previous project.

This form must be accompanied by the previous year's abstract and Research Plan/Project Summary.

Student's Name(s) Saniya Gaitonde

To be completed by Student Researcher: List all components of the current project that make it new and different from previous research. The information must be on the form; use an additional form for previous year and earlier projects.

Components	Current Research Project	Previous Research Project: Year: 2019
1. Title	Bioinformatic Investigation of the Peculiarities of Long Intron Splicing in Hominidae	Bioinformatic Interrogation of the Evolutionary Conservation of Recursive Motifs Mediating Multi-Step RNA Splicing
2. Change in goal/ purpose/objective	The objective is to investigate the potential of recursive sequences in Hominidae genomes, using Drosophila data as control, in order to gain insight into the effect of alternative splicing on evolutionary developmental variation and implicated RNA splicing disorders.	The objective was to characterize known recursive sequences in Drosophila based on patterns of evolutionary conservation.
3. Changes in methodology	None	None
4. Variable studied	Hominidae (Great Ape) genomes: human (H. sapiens), common chimpanzee (P. troglodytes), and bonobo (P. paniscus)	Drosophila genomes: D. melanogaster, D. yakuba, D. simulans, and D. pseudoobscura
5. Additional changes	None	None

Attached are:

☑ Abstract and Research Plan/Project Summary, Year 2019

I hereby certify that the above information is correct and that the current year Abstract & Certification and project display board properly reflect work done only in the current year.

Saniya Gaitonde — Student's Printed Name(s)

Signature

09/01/2019 — Date of Signature (mm/dd/yy)

Bioinformatic Interrogation of the Evolutionary Conservation of Recursive Motifs Mediating Multi-Step RNA Splicing

Saniya J. Gaitonde | W. Tresper Clarke High School | 2019

Research Plan

a. RATIONALE:

RNA splicing is an important step in the life cycle of mRNA during which transcripts are matured through the excision of intronic sequences, usually in a single lariat unit in a two-step catalytic reaction. The process of recursive splicing, in which introns are removed in multiple splicing reactions was first discovered in the *Ultrabithorax* (*Ubx*) gene of *Drosophila melanogaster*, and thereafter in an additional small subset of introns in the species through novel computational approaches. Recursive sites have been found to occur in most very long fly introns, including many genes involved in morphogenesis and development, and tend to occur near the midpoints of introns. It is observed that fly introns with recursive sites are spliced more accurately than comparably sized non-recursive introns. Although only *D. melanogaster* has been extensively characterized, the most important aspects of the cellular, molecular, and developmental biology of *Drosophila* species have been evolutionarily well-conserved. Thus, in addition to providing an extensive resource for the study of the relationship between sequence and phenotypic diversity, the genomes of these species provide an excellent model for studying how conserved functions are maintained in the face of sequence divergence. The purpose of this investigation is to characterize known recursive sequences in *Drosophila* based on patterns of evolutionary conservation.

b. RESEARCH QUESTIONS:

1. What do the recursive sequences of each species look like?
2. How many recursive sites are present in each species?
3. Is the location of these sites conserved across species?
4. How many sites are enriched in introns, exons, and non-coding regions?

HYPOTHESIS:

Motifs for recursive sites, which mediate multi-step RNA splicing, are enriched and conserved in the introns of *Drosophila* genes with key developmental functions, consistent with known evolutionary relationships.

ENGINEERING GOAL:

The overarching goal of this project is to offer novel computational data on the potential of recursive splicing as an alternative pre-mRNA transcript maturation mechanism based on its evolutionary enrichment patterns in the *Drosophila* genus.

EXPECTED OUTCOMES:
It is expected that each species' scores and sequences will be similar and consistent with known evolutionary relationships, with dominant enrichment in introns, providing evidence of conservation.

c. PROCEDURES:
Recursive splicing can be detected by computationally parsing genomic DNA for recursive sequences. A python program was written to scan and score genomic sequences for matches to recursive sites, using other known regulatory sites as controls. The script will accept a 26 base pair position-specific scoring matrix (PSSM) file which represents a recursive motif numerically, in addition to FASTA file genomes of *D. melanogaster*, *D. yakuba*, *D. simulans*, and *D. pseudoobscura*, individually. All genome samples were obtained from the publically available UCSC database. For each sample, the script will then output a BED file delineating 26 base pair segments of the FASTA within which there is potential for the PSSM's recursive motif to fall, along with the DNA strand from which the gene was transcribed and the score of the region's match to the recursive PSSM, which can be used for several analytics.

RISK & SAFETY:
There are no risks and no safety precautions necessary for an investigation of this nature.

DATA ANALYSIS:
Using the BED file outputs, the average match score will be calculated for each sample's low, middle, and high scoring regions. The BED files will then each be inputted into WebLogo, a software which will materialize the appearances of the recursive sequences in each species by score to determine whether they look similar. Next, the BED files will be uploaded to the online LiftOver tool which will convert the genomic coordinates of each species' sequences to a common reference. The converted BED files will then be executed in the pre-programmed python LiftOver script which will return the total number of motif matches in each species and how many overlapped in both directions for any combination of two species, revealing the extent to which the location of recursive sites has been conserved. Finally, the annotation overlap python script will be used to determine how many sites are enriched in introns, exons, and non-coding regions.

d. BIBLIOGRAPHY:

Evolution of genes and genomes on the Drosophila phylogeny. (2007). *Nature, 450*(7167), 203-218. doi:10.1038/nature06341

Pai, A. A., Paggi, J. M., Yan, P., Adelman, K., & Burge, C. B. (2018). Numerous recursive sites contribute to accuracy of splicing in long introns in flies. *PLOS Genetics, 14*(8). doi:10.1371/journal.pgen.1007588

Pandey, U. B., & Nichols, C. D. (2011). Human Disease Models in Drosophila melanogaster and the Role of the Fly in Therapeutic Drug Discovery. *Pharmacological Reviews, 63*(2), 411-436. doi:10.1124/pr.110.003293

No changes were made to this Research Plan.

OFFICIAL ABSTRACT and CERTIFICATION

Bioinformatic Interrogation of the Evolutionary Conservation of Recursive Motifs Mediating Multi-Step RNA Splicing

Saniya Gaitonde
W. Tresper Clarke High School, Westbury, New York, USA

Recursive splicing is an RNA maturation phenomenon wherein transcripts are matured during multiple splicing reactions, first discovered to be an increasingly accurate mechanism in long Drosophila melanogaster introns of genes involved in morphogenesis and development. The Drosophila genus is an excellent model for studying how functions are maintained in the face of sequence divergence due to molecular aspects being evolutionarily well-conserved. The purpose of this investigation was to characterize known recursive sequences in Drosophila based on patterns of evolutionary conservation. It was hypothesized that motifs for recursive sites, which mediate multi-step RNA splicing, are enriched and conserved in Drosophila introns of genes with key developmental functions, consistent with known evolutionary relationships. A python script that scans and scores genomic sequences for matches to recursive sites, using other known regulatory sites as controls, was used to scan the genomes of D. melanogaster along with three related species. The potential of recursive sites in these species was then evaluated based on probability to share appearance with known recursive motifs, scoring of matches, conservation of location, and enrichment in various coding regions. The general similarity across all four species, along with a dominant conservation in intronic regions, supported the hypothesis. D. yakuba sequences were significantly more similar to D. melanogaster than D. simulans and D. pseudoobscura. Future work would entail similar analyses with mammalian genomes to gather data relative to developmental variation and implicated RNA splicing disorders.

Category
Pick one only — mark an "X" in box at right

- Animal Sciences
- Behavioral & Social Sciences
- Biochemistry
- Biomedical & Health Sciences
- Biomedical Engineering
- Cellular & Molecular Biology ■
- Chemistry
- Computational Biology & Bioinformatics
- Earth & Environmental Sciences
- Embedded Systems
- Energy: Sustainable Materials and Design
- Engineering Mechanics
- Environmental Engineering
- Materials Science
- Mathematics
- Microbiology
- Physics & Astronomy
- Plant Sciences
- Robotics & Intelligent Machines
- Systems Software
- Translational Medical Sciences

1. As a part of this research project, the student directly handled, manipulated, or interacted with (check ALL that apply):

   ☐ human participants ☐ potentially hazardous biological agents

   ☐ vertebrate animals ☐ microorganisms ☐ rDNA ☐ tissue

2. I/we worked or used equipment in a regulated research institution or industrial setting: ☐ Yes ■ No

3. This project is a continuation of previous research. ☐ Yes ■ No

4. My display board includes non-published photographs/visual depictions of humans (other than myself): ☐ Yes ■ No

5. This abstract describes only procedures performed by me/us, reflects my/our own independent research, and represents one year's work only ■ Yes ☐ No

6. I/we hereby certify that the abstract and responses to the above statements are correct and properly reflect my/our own work. ■ Yes ☐ No

This stamp or embossed seal attests that this project is in compliance with all federal and state laws and regulations and that all appropriate reviews and approvals have been obtained including the final clearance by the Scientific Review Committee.