**The Genomics Education Alliance (GEA): Towards an Comparative Genomics CURE template**

**Draft v3, SCR Elgin & W Leung, 7/21/2020**

The GEA is a group of life science educators with experience in engaging students in genomics-based Classroom-based Undergraduate Research Experiences (CUREs). A national survey of faculty teaching genomics CUREs conducted by the GEA indicated that online bioinformatics lessons, example courses, and faculty workshops are the most needed resources in the near future to support faculty who want to teach CUREs in genomics. Here we present work on a guide or template intended to aid faculty considering implementing a CURE involving comparative genomic analyses. The template considers multiple learning objectives, entry and exit points, bioinformatics analysis frameworks, potential wet-lab extensions, and assessment strategies. We aim to facilitate efforts by faculty who want to build their own genomics CURE using our optimized resources.

There are many scientific questions that can be addressed by analysis of sequenced genomes, from the initial characterization of a newly sequenced genome (number of genes, types and distribution of repetitious sequences, presence of genes for specific biochemical pathways, etc.) to comparative studies using genomes that are recently diverged or distant, varying in complexity from phage to man. A CURE can be designed around the characterization of a newly sequenced genome or can be based on genomes already sequenced and assembled, available from NCBI. As of July 9, 2020, the genome assemblies for [71 Drosophila species](https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/Drosophila) are available through NCBI, and [34 of these genomes](https://www.ncbi.nlm.nih.gov/assembly/?term=drosophila%5Borgn%5D+latest_refseq%5Bfilter%5D) (Figure 1) are curated and maintained by the NCBI Reference Sequence (RefSeq) Database (Table 1). Drosophila will be used as an example here, but there are many other possibilities.

The quality of the genome assembly, the availability of RNA-Seq data, and the availability of a well-annotated reference genome will dictate the types of scientific questions that can be addressed. Depending on the evolutionary distance of available assembled genomes, one can examine the evolution of a given gene or of a cluster of genes in a given biological pathway or structure; look at the evolution of a chromosome or sub-domain; search for regulatory motifs; look at the history of Transposable Element (TE) acquisition or for evidence of lateral gene transfer; and address many other questions of interest.

The strategies discussed below are based on the experience of the Genomics Education Partnership (GEP) to engage students in genomics research based on engaging them first in *de novo* annotation of genes in a species of Drosophila not previously annotated beyond the computer level. The protocol requires students to learn to use the UCSC browser to examine the available evidence in detail, and to use BLAST to generate additional information. This approach does not require the student to develop specific bioinformatics skills (such as use of a command line, tools in R, etc.), but it could be modified to do so. The basic process enables students learn about genes/genomes and about how to utilize large datasets, as well as engaging them in the reconciliation of multiple lines of evidence to create defendable gene models, thus developing their analytical thinking skills. **A major advantage is that genomics research can be done entirely online, or can be expanded to include wet bench work if desired.**

Table 1. For the 34 Drosophila genomes in the NCBI Reference Sequence Database, the number of contigs range from 18–34,033; the contig N50 ranges from 15,132–33,427,555 bp; the contig L50 ranges from 2–2,212; and the number of gaps ranges from 0–16,198. (The contig N50 is defined as the contig size and larger which accounts for half of the total size of the genome assembly. The number of contigs needed to reach the N50 threshold is defined as the contig L50.) All 34 species have RNA-Seq datasets that can be used for structural gene annotation. (See Supplemental Table 1 for the detailed statistics for each Drosophila assembly.)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Species | NCBI Accession Number | Number of Contigs | Contig N50 | Contig L50 | Number of Gaps |
| *Drosophila melanogaster* | GCF\_000001215.4 | 2,442 | 21,485,538 | 3 | 572 |
| *Drosophila mauritiana* | GCF\_004382145.1 | 381 | 22,120,385 | 4 | 28 |
| *Drosophila sechellia* | GCF\_004382195.1 | 440 | 19,907,079 | 4 | 38 |
| *Drosophila simulans* | GCF\_000754195.2 | 9,975 | 451,931 | 74 | 2,356 |
| *Drosophila yakuba* | GCF\_000005975.2 | 13,441 | 115,562 | 275 | 5,318 |
| *Drosophila erecta* | GCF\_003286155.1 | 94 | 22,146,549 | 3 | 0 |
| *Drosophila ficusphila* | GCF\_000220665.1 | 9,152 | 275,894 | 135 | 3,398 |
| *Drosophila suzukii* | GCF\_000472105.1 | 24,878 | 24,954 | 2,021 | 16,198 |
| *Drosophila biarmipes* | GCF\_000233415.1 | 7,856 | 474,639 | 90 | 2,333 |
| *Drosophila takahashii* | GCF\_000224235.1 | 9,703 | 126,259 | 358 | 3,970 |
| *Drosophila eugracilis* | GCF\_000236325.1 | 7,568 | 224,458 | 146 | 2,622 |
| *Drosophila rhopaloa* | GCF\_000236305.1 | 34,033 | 19,484 | 1,870 | 11,214 |
| *Drosophila elegans* | GCF\_000224195.1 | 8,403 | 212,818 | 203 | 2,974 |
| *Drosophila kikkawai* | GCF\_000224215.1 | 8,343 | 209,056 | 197 | 3,202 |
| *Drosophila serrata* | GCF\_002093755.1 | 1,356 | 942,627 | 38 | 0 |
| *Drosophila bipectinata* | GCF\_000236285.1 | 8,675 | 149,088 | 228 | 3,175 |
| *Drosophila ananassae* | GCF\_003285975.2 | 235 | 6,212,830 | 7 | 0 |
| *Drosophila pseudoobscura* | GCF\_009870125.1 | 72 | 30,706,867 | 3 | 2 |
| *Drosophila persimilis* | GCF\_003286085.1 | 432 | 5,212,974 | 7 | 0 |
| *Drosophila miranda* | GCF\_003369915.1 | 222 | 11,978,448 | 7 | 118 |
| *Drosophila guanche* | GCF\_900245975.1 | 15,012 | 294,367 | 128 | 1,506 |
| *Drosophila subobscura* | GCF\_008121235.1 | 56 | 11,370,518 | 4 | 24 |
| *Drosophila obscura* | GCF\_002217835.1 | 5,623 | 114,814 | 395 | 3,688 |
| *Drosophila willistoni* | GCF\_000005925.1 | 20,358 | 180,217 | 279 | 5,520 |
| *Drosophila arizonae* | GCF\_001654025.1 | 17,526 | 15,132 | 2,212 | 14,348 |
| *Drosophila mojavensis* | GCF\_000005175.2 | 11,874 | 121,517 | 358 | 5,033 |
| *Drosophila navojoa* | GCF\_001654015.2 | 22,848 | 26,679 | 1,317 | 9,035 |
| *Drosophila hydei* | GCF\_003285905.1 | 217 | 3,367,158 | 9 | 0 |
| *Drosophila virilis* | GCF\_003285735.1 | 222 | 8,697,263 | 7 | 0 |
| *Drosophila novamexicana* | GCF\_003285875.2 | 269 | 3,158,326 | 18 | 0 |
| *Drosophila albomicans* | GCF\_009650485.1 | 18 | 33,427,555 | 2 | 7 |
| *Drosophila innubila* | GCF\_004354385.1 | 561 | 3,071,820 | 17 | 198 |
| *Drosophila grimshawi* | GCF\_000005155.2 | 24,157 | 91,192 | 447 | 6,717 |
| *Drosophila busckii* | GCF\_011750605.1 | 323 | 1,002,992 | 27 | 229 |

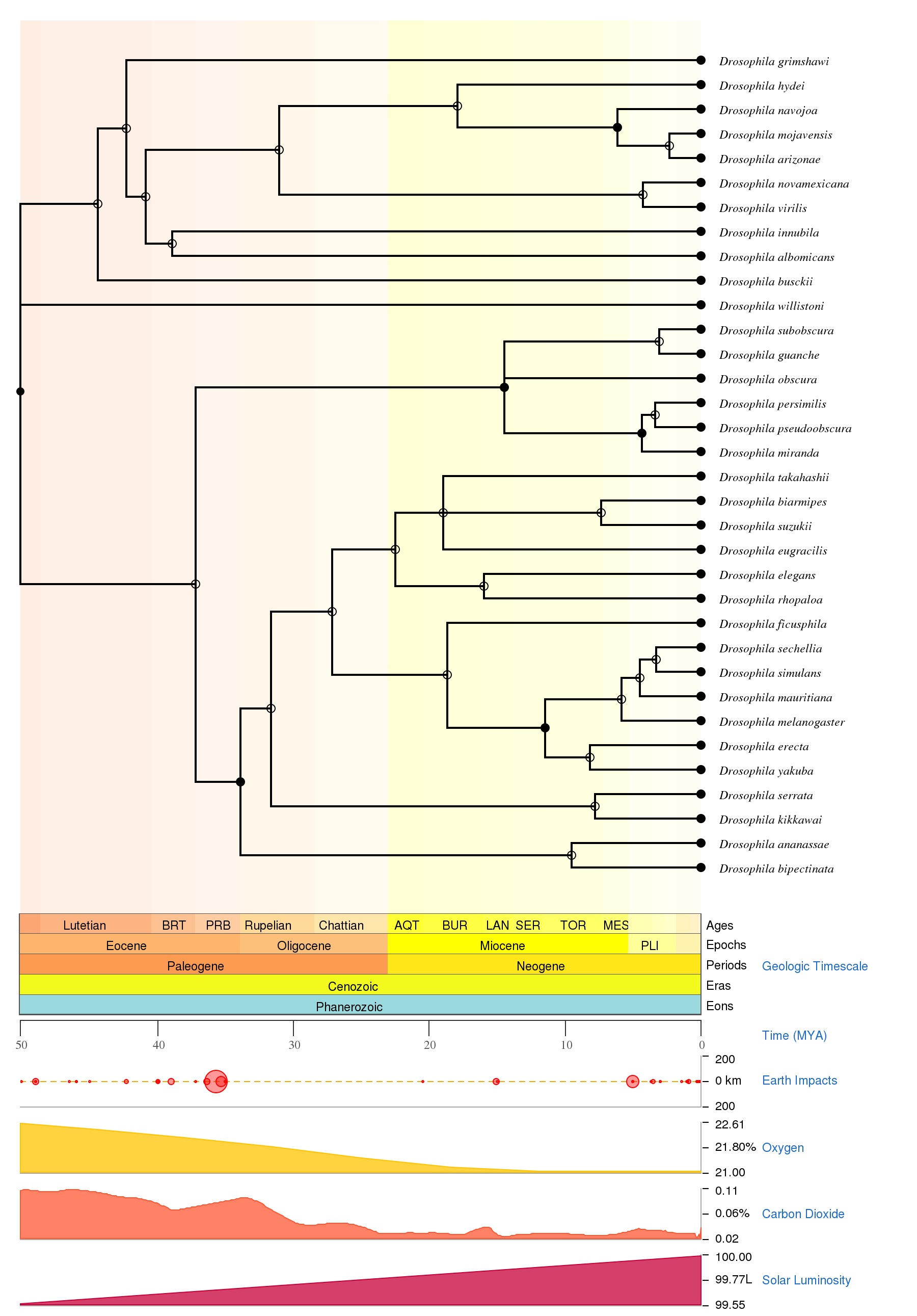


Figure 1. Phylogenetic tree for the 34 Drosophila genomes in the NCBI RefSeq database. The phylogenetic relationships among these Drosophila species were obtained from the Timetree database (Kumar S et al., 2017).

Starting from a newly sequenced genome

For prokaryotic genomes, NCBI provides the [Prokaryotic Genome Annotation Pipeline (PGAP)](https://www.ncbi.nlm.nih.gov/genome/annotation_prok/) that can quickly annotate a genome and create the annotation files for submission to GenBank (Tatusova T et al., 2016). Tools such as [DNA Master and Starterator](https://seaphages.org/software/) can be used by students in the manual annotation of prokaryotic genomes (Pope and Jacobs-Sera, 2018).

For eukaryotic genomes, the first step is to create a genome browser for the newly assembled genome. [G-OnRamp](http://g-onramp.org/) provides a user-friendly, web-based platform for biologists to construct genome browsers (Liu et al. 2019). The Galaxy workflows provided by G-OnRamp produce UCSC Assembly Hubs and JBrowse/Apollo genome browsers with evidence tracks from sequence alignments, *ab initio* gene predictors, RNA-Seq data, and repeat finders. The input required is the assembled target genome; highly desirable inputs are proteins and/or transcripts from an informant genome, and RNA-Seq data for the target genome.

This step can be done by the instructor prior to student involvement, or could be done by the students in an advanced class. Genome Browsers produced by G-OnRamp have been shown to facilitate the annotation of novel species by undergraduates (Sargent et al. 2020).

TEMPLATE: Beginning Students

Annotating a gene: Learning how to use a genome browser and annotating a gene is an excellent way for freshman/sophomore students taking a first course in biology or genetics to learn about eukaryotic gene structure. Curriculum to guide the student through this process has been developed by members of the GEP and has been posted as a CourseSource article (Laakso et al. 2017). Students who complete this six-unit curriculum (typically scheduled over three 3-hr lab periods in a course that includes lectures on eukaryotic gene structure) are ready to join the research process.

Learning goals (from Laakso et al 2017)**:** Students will understand the structure of a eukaryotic gene, and how sequences within a gene impact the gene product. Students will understand the relationships among DNA, pre-mRNA, mRNA, and protein. Students will also learn how to use a genome browser, and to interpret the evidence tracks produced by multiple bioinformatics tools (e.g., genome sequence, RNA-Seq data) to gain an in-depth understanding of eukaryotic genes.

Data sources: Lessons based on the *Drosophila melanogaster tra* gene are provided by GEA (<https://gea.qubeshub.org/lessons/eukaryotic-genes-modules>)

Background:

How are genomes sequenced?

Introduction to the databases (GEO, SRA, NCBI)

Optional: How to explore the possibilities for setting up your own research questions

Key Scientific Process Skills:

Gathering data/making observations

Analyzing data

Interpreting results/data

Displaying/modeling results/data

Bioinformatic analysis

Learning to use a genome browser

Accessing databases

Using modeling and simulation

Assessment

Assessment of individual student performance

Answer: fill in the blank question(s)

Answer: short answer question(s)

Create a diagram, drawing, figure, etc.

Interpret data

TEMPLATE: Intermediate/advanced Students

Research annotation of a genome: students with a solid foundation in molecular genetics and some prior introduction to genome browsers can start with an introduction to using BLAST and a protocol walk-through for gene annotation. Here we provide a BLAST lesson using the human leptin gene as an example, and an annotation walk-through designed for use with species of Drosophila. The latter includes custom software, “Gene Model Checker,” that enables students to check their model to see if it conforms to the basic rules of molecular biology (e.g., using an appropriate start and stop codons, maintaining an open reading frame, etc.); it also specifies a protocol for students to submit their gene models to a central data collection site. This work flow will need to be modified to be specific for the type of annotation and species under study.

**Lesson: Introduction to BLAST**

Learning goals: This lesson aims to introduce students to the use of the Basic Local Alignment Search Tool (BLAST) to identify related sequences and compare similarity between them. The lesson begins with introducing students to the use of NCBI databases to explore diseases and associated genes, including how to retrieve a desired sequence from NCBI that is then used to demonstrate the use of BLAST to find related sequences.

Data Sources: The BLAST introduction using the human leptin gene is available through the GEA at <https://gea.qubeshub.org/lessons/intro-blast>. Results for the leptin example BLAST search have been cached in the CyVerse Data Store to allow for rapid access to the BLAST results when implementing with large-enrollment courses.

Background and follow-up

Human genetic disabilities, including the power of diagnosis

GEP material on Fragile X

Advanced classes may follow up with an examination of the underlying algorithm for similarity searching.

Key scientific process skills

Finding and using large databases

Finding related sequences

Interpreting results/data (significance / type of relatedness)

Displaying/modeling results/data

Bioinformatics analysis

Learning to use BLAST and to interpret the results

Assessment

Formative assessments are imbedded within the lesson to lead the student to consider their search results.

A summative assessment is used to help the students reflect on the uses of databases and BLAST to explore functional relationships of genes and genomes.

**CURE: A gene annotation protocol to explore a group of genes of interest**

Learning goals: to engage the students in a research project using comparative genomics, starting with mindful gene annotation.

Data sources

Assembled genomes available through NCBI

Well-annotated reference species?

Other species at a suitable evolutionary distance for the question to be asked?

RNA-Seq data sets? (developmental stages, specific perturbations, etc.)

RNA-Seq data sets designed to identify TSS?

Other information: KEGG….

Newly assembled genome of interest?

Background and curriculum: (whether these topics are just touched on, or developed in depth, will depend on the goals of the course and the time available with students.)

The biological question being asked will dictate background materials needed to develop the scientific interest.

Detecting and interpreting genetic homology; exploration of similarity searching

Design and use of RepeatMasker

Generation and analysis of RNA-Seq data (including TSS-based assays as applicable)

Introduction to *ab initio* gene finding

Other topics depending on the evidence tracks to be used in the annotation

Issues of genome organization as appropriate (e.g., synteny)

Example curriculum can be found on the GEP webpage at <https://gep.wustl.edu/curriculum/course_materials_WU/annotation/all_annotation_materials>

Experimental design considerations:

Is the quality of the genome sequence/assembly adequate for the questions to be addressed? Which sequencing protocol?

What additional data is available?

How does the data available and questions of interest shape the annotation workflow?

[An example flow chart for overall annotation strategy is available at

<http://community.gep.wustl.edu/repository/introducing_genes/GEP_Annotation_Flowchart.pdf>

An example flow chart for determining splice sites is available at

<http://community.gep.wustl.edu/repository/introducing_genes/Determine_CDS_Flowchart.pdf>

Bioinformatics analysis

* Which genome browser to use? (UCSC, JBrowse, Apollo)
* What reference species will be included in the browser? Other evidence tracks?
* What evidence will students report to support their gene models (intron/exon structure)? How will discrepancies between data sources be resolved?
* How will student work be checked for basic consistency (open reading frame etc.) How will exceptions be reported? (Note that Gene Model Checker, a custom software that does the basic checking, is currently available (through GEP) only for Drosophila, but could be adapted to other groups of species.
* Will two independent annotations and reconciliation be required? How organized?
* What gene features beyond intron/exon structure of isoforms are students asked to examine?
* What features of the overall region will students be asked to examine (e.g. repeats, synteny etc.)
* Will additional searches be done? For example, looking for conserved motifs; if so, what tools will be used, how do they work, etc.

Molecular/genetic methods (wet bench experiments) if desired:

Look for evidence of a particular isoform that is predicted to be present in one species but not another

Look at chromosome assignment of wanderer genes in different species

How to divide the work among student groups (alt. participation pathways)

Advantages to having students working together; create partners or buddies to stimulate dialogue

Balance between synchronous (in person or online) and asynchronous work

Balance between lecture/demonstration and hands-on work from a protocol

Challenges and opportunities:

Doing the whole process in one semester

Plan for two independent annotations of each gene, then reconciliation

Consider micropublication of each gene model.

Accumulating enough data for the desired meta-analysis

Can require several years; student co-authors must stay in touch.

Recent innovation being explored: generate a micropublication for each gene, enabling student work to be published prior to the whole project meta-analysis..

Potential implementation strategy using resources provided by FlyBase and NCBI:

1. Identify the list of genes of interest

* FlyBase Gene Group List: <https://flybase.org/lists/FBgg/>
  + Example: Polycomb Group Complexes Gene Group ([FBgg0000309](https://flybase.org/reports/FBgg0000309.html))
* FlyBase Pathway Report List: <https://flybase.org/lists/FBgg/pathways>
  + Example: Insulin-like Receptor Signaling Pathway ([FBgg0000910](https://flybase.org/reports/FBgg0000910.html))
* Gene Ontology:
  + Example: Biological Process - [chromatin silencing](http://flybase.org/cgi-bin/cvreport.pl?id=GO%3A0006342)
    - Gene List for [GO:0006342](http://flybase.org/cgi-bin/fbcvquery.pl?mode=subtreehits&cvterm=GO:0006342%3EFBgn)
  + Example: Biological Process - [histone deacetylation](http://flybase.org/cgi-bin/cvreport.pl?id=GO%3A0016575)
    - Gene List for [GO:0016575](http://flybase.org/cgi-bin/fbcvquery.pl?mode=subtreehits&cvterm=GO:0016575%3EFBgn)

2. Identify orthologs of *D. melanogaster* genes in other Drosophila species

* [Putative orthologs of *D. melanogaster* genes](https://ftp.ncbi.nlm.nih.gov/gene/DATA/special_requests/gene_orthologs_supplemental.gz) in other insects provided by the NCBI RefSeq database
  + Include ortholog assignments for [30 Drosophila species in the RefSeq dataset](https://ncbiinsights.ncbi.nlm.nih.gov/2020/04/22/flies-are-a-buzzing-in-refseq/)

3. Create student annotation projects based on the locations of the putative RefSeq

orthologs, assigning the student to annotate all genes and other features found in the

region of xxxxxx bp to yyyyyy bp in the species of interest.

•  If possible use the genome browsers for multiple Drosophila species provided by the GEP UCSC Genome Browser or the NCBI Genome Data Viewer; otherwise create appropriate browser (e.g. using G-OnRamp).

**Assessment**

How do we know the students have met the learning objectives?

Assessment approaches:

Graded problem sets associated with curriculum; demonstrate understanding in the use of a tool or resource

Lab meetings where students present challenges (graded on clear presentation of the problem, 3-5 ppt slides)

Students write an entire or partial formal manuscript on their gene models; overview of genome region

Instructions

Rubric

Teaching tips

Students do oral PowerPoint presentations on their gene models; overview of genome region

Instructions

Rubric

Teaching tips

What content to add as options?

How we want the project to function – one project that everyone works on together and publishes as a group, or a distributed project where individual schools look at related but distinct groups of genes, publishing independently? The former is necessary if a large number of annotations are needed, defined by the number of genes to be examined times the number of species to be included in the study,

Discussion question: What research questions come to your mind after reading this framework? Are you interested in a particular biochemical or developmental pathway? Or a particular biological structure? What questions might be addressed by looking at the evolution of the genes involved? Similarly, if you are interested in a particular feature of genome organization, epigenetic mechanisms, etc. what might be learned by looking at the evolution of genes subject to that influence?

# References

Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol. 2017;34(7):1812-1819. doi:10.1093/molbev/msx116

Laakso MM, Paliulis LV, Croonquist P, et al. An undergraduate bioinformatics curriculum that teaches eukaryotic gene structure. CourseSource. 2017. doi:10.24918/cs.2017.13

Liu Y, Sargent L, Leung W, Elgin SCR, Goecks J. G-OnRamp: a Galaxy-based platform for collaborative annotation of eukaryotic genomes. Bioinformatics. 2019;35(21):4422-4423. doi:10.1093/bioinformatics/btz309

Pope WH, Jacobs-Sera D. Annotation of Bacteriophage Genome Sequences Using DNA Master: An Overview. Methods Mol Biol. 2018;1681:217-229. doi:10.1007/978-1-4939-7343-9\_16

Sargent L, Liu Y, Leung W, et al. G-OnRamp: Generating genome browsers to facilitate undergraduate-driven collaborative genome annotation. PLoS Comput Biol. 2020;16(6):e1007863. doi:10.1371/journal.pcbi.1007863

Tatusova T, DiCuccio M, Badretdin A, et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 2016;44(14):6614-6624. doi:10.1093/nar/gkw569