**Community Genome Annotation Training (CGAT) database** *(tentative name in honor of GCAT from Davidson)*.

**Introduction:**

There are several examples of successful creation/organization of knowledge by larger community rather than individual experts (for example, FoldIt – protein folding game <http://fold.it/portal/> ; Wikipedia – reported to have fewer errors than Encyclopedia Britannica). We will try to create an environment/database that facilitates community-based genome annotation and helps train students.

**Training:** Genomics Education Partnership (GEP; <http://gep.wustl.edu> ) is a community of researchers, educators and students working on annotation of eukaryotic genomes. Currently, student-submitted annotations are reconciled by trained annotators providing feedback to faculty at the end of the school year. In most cases, the feedback comes too late to benefit individual student. One of the challenges for students in learning to annotate is lack of immediate feedback. A small number of tutorials are available on the GEP website, but additional practice opportunities with immediate feedback would be extremely helpful, especially, if a large number of students is to be involved. Currently available Gene Model Checker (GEP website, custom software) is a great tool for catching many annotation errors and eliminating implausible gene models, but does not give students information about the correct model. CGAT database could be used by students after the Gene Model Checker as a training tool (and/or the next level of quality control).

**Community annotation:** It is possible that community effort will result in accurate annotation without upfront input from the expert (expert annotation being the standard for comparison of community annotation), at least for some genes. We will test this possibility and will try to understand characteristics of the genes/genomes amenable to community annotation, as well as the level of annotation experience required to produce quality annotation by community member (unpaid annotator)

**Goal:** Create a community genome annotation training database that

1. helps train genome annotators by

-providing immediate feedback to upon submission of completed annotation;

-guiding the choice of locations/genes to be annotated using individual’s feedback/success in prior annotation submission and community feedback;

-providing community feedback/hints/discussion (may not work – see FoldIt website forum post: <http://fold.it/portal/node/991548>)

2. creates accurate annotation for eukaryotic genes via community-input efforts

**Current state:** prototype is built and tested by developer, tried only by 1 instructor, will/can be tested by users in F2014; housed on the iplant server.

Database has good backend, but not a pretty front end and is missing detailed feedback/hints.

Student who developed it (Ryan) – graduated; to continue work at Cal Poly, we’d need 1/2month summer salary for CS supervisor (Alex Dekhtyar) and some money to pay a CS student to do the work (typically $15/hour, no more than 20 hours/week). Alternatively, can be handed off to Wilson, or Wilson’s clone, or somebody else in the GEP community with some programming expertise.

**Database description:**

**Data:**

The database will contain the following types of data:

Assignments (location in the genome to be annotated) - difficulty level? Annotation problem/challenge category? – input by GEP faculty

Submission - annotator-generated information for a specific assignment, including gene name, # of isoforms, for each isoform, coordinates of coding exons (CDS) and coordinates of transcript.

Annotation – information for a given assignment compiled based on cumulative submissions, included list/distribution of common errors, difficulty level; expert-generated “most-likely correct” set of coordinates

Annotator/user/student – creates submissions; database stores responses to survey questions, level of expertise, error types, school, faculty supervisor…

**Users:**

Students learning to annotate will

Obtain/choose an assignment/location to annotate,

Submit annotation (gene name, isoform names, GFF or list of exons)

Get feedback on submitted annotation (# of existing submissions for the given assignment/location, % of total submissions that match current submission, % “expert’s” annotations that match current submission, type of error- if available or “most likely correct”)

Resubmit/revise submission for each assignment

View cumulative individual activity report (how many genes are successfully annotated, types of genes/difficulty level,? – earn “badges, stars, points” for success, white belt vs. black belt)

Experts/GEP faculty will

Submit “reference” annotation

Analyze all submissions to categorize errors and provide feedback message for each error type; categorize annotation by difficulty level, type of challenge

Evaluate success of community annotation – e.g. how many submissions are required to have “most likely correct” consensus from the community; what kinds of genes/annotations can be successfully annotated by community efforts.

Correlate student responses to surveys, number of submissions/resubmissions with success.

Outside the database: use error categories to improve training materials/approach.

**Issues of concern:**

1. Flybase is updated every two month – reference genome changes! GEP solution – version freeze; keep a link to frozen version or let the more accurate annotation emerge as Flybase is updated? Solution: attach flybase version to all submissions, have “expert” level user re-release assignments that need to be re-annotated based on changes in later FlyBase releases.

2. ?

**Plan/timeline:**

Fall 2012 CS students at Cal Poly produce the first version. This version will contain “annotation sandbox” - tool for training that contains expert- generated correct annotations and meets goal #1 above.

Winter 2013 CS students and A.Goodman test/debug the initial version and work on “community annotation” (goal 2) functionality.

Spring 2013 Cal Poly BIO441 class tests the database – NOT DONE, was not ready.

Summer 2013 present at GEP Alumni Workshop and release to the GEP community.