Three’s a Crowd-Source
The Collaborative Nature of Genome Annotation

Monica C Munoz-Torres1,7 | mcmunoz@lbl.gov | @monimunozto
Nathan Dunn1, Chris Mungall1, Seth Carbon1, Heiko Diezte1, Colin Diesh2, Deepak Unni2, Ian Holmes1,2, Christine Elsik1, Suzanna Lewis1
1Lawrence Berkeley National Laboratory, Genomics Division, Berkeley, CA. 2University of California Berkeley, Bioengineering, Berkeley, CA.

Unlike the more highly polished genomes of earlier projects, recent projects usually have lower coverage that confront researchers with more frequent assembly errors and annotation of genes across multiple scaffolds. Automated genome annotations must be curated to resolve discrepancies, providing clarity and validation.

MANUAL GENOME ANNOTATION EDITING
To curate automated genome annotations researchers gather and evaluate all available evidence using quality-control metrics to corroborate or modify gene predictions. This resolves discrepancies, providing clarity and validation to the gene model hypothesis.

Manual genome annotation allows us to:
• Identify elements that best represent the underlying biology.
• Eliminate elements that reflect the systemic errors of automated genome analyses.
• Determine functional roles through comparative analysis of well studied, phylogenetically similar genome elements using literature, databases, and the researcher’s experience.

Web Apollo: COLLABORATIVE GENOMIC ANNOTATION EDITING PLATFORM
Apollo is now a browser-based annotation tool integrated with the JBrowse genome browser to allow distributed research communities to collaborate performing manual gene annotations. Web Apollo allows users to create and modify transcript and exon structures via intuitive gestures while flagging potential problems.

Web Apollo also provides dynamic access to genomic analysis results from both UCSC and Chado databases as well as database storage of user-created annotations. All user-created sequence annotations are automatically uploaded to a server, ensuring reliability. The server provides synchronized updates over multiple browser instances, so annotation edits are immediately visible to all users who are working on the same region.

CROWD-SOURCING GENOME CURATION
Unlike earlier genome projects, recent sequencing projects are riddled by a number of factors that confront researchers with additional work to correct for more frequent assembly errors and annotate genes split across multiple contigs.

It is impossible for a single individual to fully curate a genome with precise biological fidelity. Beyond the problem of scale, curators need second opinions and insights from colleagues with domain and gene family expertise, an inherently collaborative task.

Working together, researchers are able to:
• Distribute problem solving
• Mine collective intelligence
• Access improved quality data
• Process work in parallel

We continuously train and support hundreds of geographically dispersed scientists to perform biologically supported manual annotations using Web Apollo.

These scientific community efforts bring together domain-specific and natural history expertise that would otherwise remain disconnected.

Breaking down large amounts of data into manageable portions and mobilizing groups of researchers to extract the most accurate representation of the biology from all available data distills invaluable knowledge from genome analysis.

EXAMPLE: Understanding the evolution of sociality. Comparing seven ant genomes provided a better understanding of evolution and organization of insect societies at the molecular level. Insights were drawn mainly from six core aspects of ant biology: Alternative morphological castes, Division of labor, Chemical Communication, Alternative social organization, Social immunity, and Mutualism. Libbrecht et al. 2013. Genome Biology, 14:212

CONTINUED IMPROVING GENOMIC ANNOTATION AND CURATION
Continuing to improve genomic annotation and curation will involve increasing the researchers’ efficiency by providing a suite of integrated curation tools, and increasing the effective population of researchers by providing universally accessible tools.

Our goals are: 1) to create a federated environment combining structural, functional, transcriptomic, proteomic, and phenotypic data; and 2) to train the next generation of researchers in making the most effective use of these tools.

IMPROVING GENOMIC ANNOTATION AND CURATION
Continuing to improve genomic annotation and curation will involve increasing the researchers’ efficiency by providing a suite of integrated curation tools, and increasing the effective population of researchers by providing universally accessible tools. To this end we are in the process of modeling a federated environment that combines gene structural and functional data, transcriptomic, proteomic, and phenotypic data.

Web Apollo enables collaboration in real-time, assisting to distill valuable knowledge from genome analysis. Collaborations encourage second opinions and insights from colleagues with diverse domain and gene family expertise.

We continually train and support hundreds of geographically dispersed scientists to perform biologically supported manual annotations using Web Apollo.

These scientific community efforts bring together domain-specific and natural history expertise that would otherwise remain disconnected.

Breaking down large amounts of data into manageable portions and mobilizing groups of researchers to extract the most accurate representation of the biology from all available data distills invaluable knowledge from genome analysis.

EXAMPLE: Understanding the evolution of sociality. Comparing seven ant genomes provided a better understanding of evolution and organization of insect societies at the molecular level. Insights were drawn mainly from six core aspects of ant biology: Alternative morphological castes, Division of labor, Chemical Communication, Alternative social organization, Social immunity, and Mutualism. Libbrecht et al. 2013. Genome Biology, 14:212