2-2017

The GEP: Crowd-Sourcing Big Data Analysis with Undergraduates

Sarah C.R. Elgin
Washington University in St. Louis, selgin@wustl.edu

Charles Hauser

Teresa Holzen

Christopher Jones

Adam Kleinschmit

See next page for additional authors

CellPress

Scientific Life

The GEP: Crowd-Sourcing Big Data Analysis with Undergraduates

Sarah C.R. Elgin,[1,*] Charles Hauser,[2] Teresa M. Holzen,[3] Christopher Jones,[4] Adam Kleinschmit,[5] Judith Leatherman,[6] and The Genomics Education Partnership[7]

The era of 'big data' is also the era of abundant data, creating new opportunities for student–scientist research partnerships. By coordinating undergraduate efforts, the Genomics Education Partnership produces high-quality annotated data sets and analyses that could not be generated otherwise, leading to scientific publications while providing many students with research experience.

Current technology has allowed massive amounts of data to be collected in many fields, including genomics, anatomy, ecology, astronomy, and so on. Typically, after analysis to answer the motivating question, the data are put into publicly accessible storage. Many of these data sets still contain useful, unmined information, creating an opportunity for expanded investigations. We have developed one such system for taking advantage of public genomic data sets, by developing data analysis tools and providing them via the Internet to allow undergraduates to engage in research. This system of coordinating 'massively parallel' undergraduate efforts can be broadly applied to other fields, providing benefits to the scientific community, the scientists directing the study, and the students themselves.

Launched in 2006, the Genomics Education Partnership (GEP[i]) brings undergraduates into genomics research. The consortium currently includes over 100 faculty members from diverse schools (see 'Contributing Authors' section). GEP students have contributed to improving the underlying DNA sequence quality and manually annotating selected regions of several *Drosophila* genomes. While helping students learn the basics of eukaryotic gene structure and genome organization, the process also introduces students to large genomics databases and bioinformatics tools, strengthens their appreciation of evolution, immerses them in scientific inquiry, encourages critical thinking, and leads some to pursue graduate work and/or bioinformatics careers. The improved DNA sequence and careful annotations they generated served as the foundation in an analysis of the comparative evolution of megabase domains (a gene-rich heterochromatic domain versus a euchromatic domain), with high confidence in the findings [1].

Such student 'crowd-sourcing' efforts are scientifically valuable. In our recent study comparing *Drosophila melanogaster* with three other *Drosophila* species, GEP students working between 2007 and 2012 improved 3.8 Mb of DNA from *Drosophila mojavensis* and *Drosophila grimshawi*, closing 72 gaps and adding 44 468 bp of sequence. Students then annotated ~8 Mb of DNA, modeling 1619 isoforms of 878 genes across three species. Whereas 58% of the final gene models agreed with the GLEAN-R gene predictions, 42% did not. Careful analysis of the findings indicates that human reconciliation of conflicting data is currently superior for accuracy, albeit significantly slower. The resulting publication, which examines the repeat characteristics (e.g., transposon density) and evolution of the genes (e.g., gene size, codon bias, and gene movement) in a heterochromatic domain, has 1014 co-authors, including 940 undergraduates [1].

The GEP project management process is presented in Figure 1. For projects such as this to be fruitful, it is necessary that the problem be one that can be subdivided, with each student (or small group) having specific responsibilities. It is also important to provide students with a standard analysis protocol, as well as leading questions and/or tools that enable students to check their work. In the GEP, students working on different species of *Drosophila* aim to construct gene models that are best supported by the available evidence. That evidence includes sequence similarity to the annotated proteins of the well-annotated reference *D. melanogaster*; results from *ab initio* and extrinsic gene finders; and all available modENCODE RNA-Seq data for the species. This information and other custom data are provided to students through a local instance of the UCSC Genome Browser (Figure 2). Students must evaluate and reconcile multiple lines of potentially contradictory evidence to construct a gene model that they can defend and use in subsequent explorations. Large numbers of participants enable the GEP to replicate annotations, with experienced students (and occasionally staff) doing a final reconciliation of any conflicting results [2]. In our recent analysis of ~2.1 Mb of the *D. biarmipes* D element, GEP students produced 610 gene models, ~74% in complete congruence with the final reconciled gene models (W. Leung, unpublished data, 2015).

GEP faculty embed this research challenge where appropriate in their curriculum, generally in the laboratory portion of a genetics or molecular biology course, in a dedicated genomics laboratory course, or through independent study. Such course-based undergraduate research experiences (CURE or CRE) are more accessible for students who might not seek out a traditional apprentice-style research experience [3], thus promoting inclusive excellence. Courses also enable us to provide research experiences for more students. Each GEP faculty member decides on the preliminary training needed for their class,

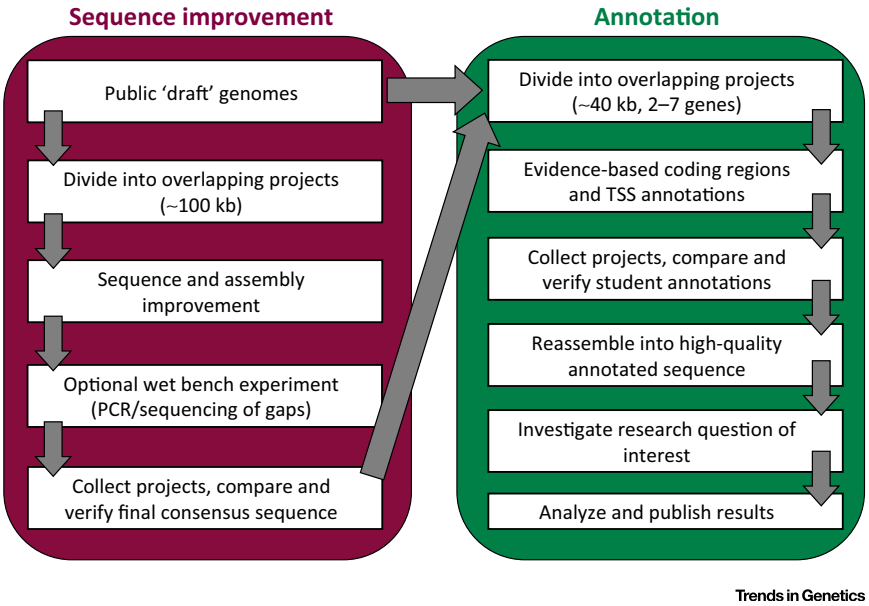

Figure 1. Flowchart of the Genomics Education Partnership (GEP) Research Process. The draft *Drosophila* genome assemblies and raw sequence data are obtained from NCBI. GEP staff at Washington University in St Louis (WUSTL) analyze these assemblies to identify regions of interest (e.g., Muller F and D element scaffolds). These regions are partitioned into overlapping projects at the appropriate size [currently ~100 kb for sequence improvement and ~40 kb (from two to seven genes) for annotation]. GEP faculty members claim the number of projects appropriate for their class. On completion, GEP students submit their projects (with a detailed report) to WUSTL. For quality-control purposes, each project is completed by at least two groups working independently and then reconciled by experienced undergraduate students. These reconciled projects are then reassembled to create a large domain (~1–3 Mb) of high-quality annotated sequence, which is then used in the final analyses and subsequent publications in the scientific literature.

creating their own curriculum or selecting from a collection of shared materials on the GEP website. Faculty members coach students throughout the ongoing research, and direct their subsequent explorations, which vary depending on the class learning objectives.

Assessment of pre- and postcourse quiz performances show that participating students increase their knowledge of eukaryotic genes and genomes and gain insight into, and appreciation for, the scientific process. In fact, GEP students and undergraduates who have spent a summer in a research lab exhibited similar responses to a survey on science learning and attitudes [4,5]. Survey comments indicate that most students appreciate the hands-on approach to learning about genes and/or genomes, and ~85% are enthusiastic about the opportunity to contribute to a genuine research project. Part of their motivation stems from the fact that their work has meaning beyond the classroom. Most students present and defend their work through a poster or oral presentation, often locally and occasionally at regional and/or national conferences.

Many research projects have been successfully integrated into a CURE format [6,7]. For example, the University of Texas at Austin recently reported that engaging freshmen in a three-semester CURE[ii] results in significantly higher retention in STEM, and higher graduation rates [8]. Most of the science being done in the Texas program is based on projects led by, and centered around, the research interests of the faculty. Developing a CURE for 10–40 students around the research of an individual local faculty member is a widespread approach, applicable across the STEM disciplines [6]. Other CUREs take advantage of remote operation of sophisticated instruments available through the national laboratories or other facilities, or analyze a local problem (e.g., the operation of a LEED-certified building or the waste stream at the campus cafeteria). There are several national projects in addition to the GEP. Perhaps the largest is SEA-PHAGES, which involves students in plaque purification and characterization of novel locally isolated phage, followed by genome sequencing and annotation[iii]. Investigations that benefit from collection and coordinated analysis of an array of data are especially good topics for a CURE.

Faculty participating in national research projects, such as the GEP, clearly benefit as well. The central organization sets up and maintains a website so that projects, curriculum, and other resources can be shared among the whole group. Joint assessment, drawing on the large pool of students, is also carried out. Faculty attend webinars during the year and summer workshops that help them stay up-to-date in a rapidly changing field, develop new curriculum, and work on publications in the scientific and science education literatures. The project also enables them to provide a research experience for a greater proportion of their students, an objective for many schools [9].

The diverse GEP membership allows us to assess the impact of different institutional characteristics (e.g., 2/4 year, public/private, large/small, selective/open, minority or Hispanic serving) on student performance. We find no significant correlation between institutional characteristics and student success (as judged by quiz scores and a science learning and attitude survey). We do find a positive correlation between the amount of time spent on the GEP project and students achieving the full benefits of a research experience [2]. Students need time to master the tools and gain familiarity with the system; they can then begin to ask and address their own questions about the genes and genome under study.

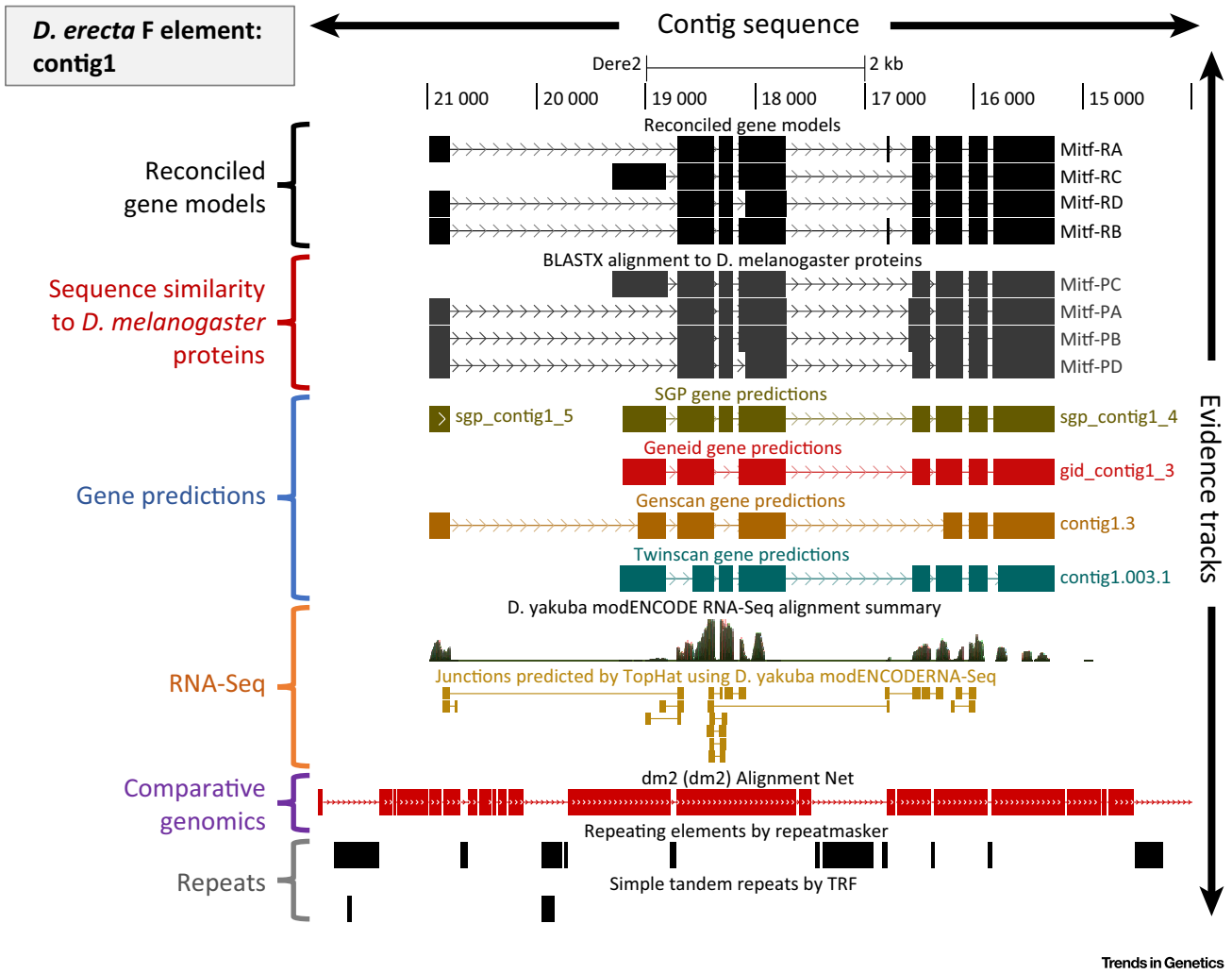

Figure 2. A Genomics Education Partnership (GEP) UCSC Genome Browser Mirror View of the *Mitf* Gene on the *Drosophila erecta* F Element. The Genome Browser provides student annotators with a workspace where they can visualize all of the available computational and experimental evidence. The available evidence tracks include sequence similarity to *Drosophila melanogaster* protein sequences, predictions from multiple gene finders, RNA-Seq read coverage and splice junction predictions from TopHat, whole-genome alignments against other *Drosophila* species, and repeats identified by RepeatMasker and Tandem Repeats Finder (TRF). Note the discrepancies among the four computational gene predictions, the lack of RNA-Seq evidence for isoform RC first exon, and the small exon in isoforms RA and RB, suggested by the RNA-Seq and TopHat tracks. In this case, the student annotators were able to resolve these contradictory lines of evidence and produce gene annotations for four different isoforms of the putative *Mitf* ortholog in *D. erecta*, as shown on the 'Reconciled Gene Models' custom track.

Having a centrally organized national experiment such as the GEP collaborative has been a win-win experience for us, the GEP faculty. In implementing this CURE, we have provided our students with rich learning experiences, while also generating useful scientific information that would be prohibitively expensive to generate by traditional means (i.e., locally with full-time research scientists). Bioinformatics is particularly well suited for a CURE, because infrastructure costs are low (computers with Internet access being the only requirement), and 24/7 access can be provided with no safety concerns, a circumstance that lends itself to peer instruction. We believe that our approach is applicable to many other studies utilizing comparative genomics in other species. Toward this end, we are working with members of the Galaxy Project (led by J. Goecks, George Washington University) to develop G-OnRamp, a system that facilitates creation of a genome browser for any eukaryotic genome.

Genome annotation and analysis is just one of many studies that can benefit from careful collection of many data points by undergraduates (see [6] for many different examples). We suggest that STEM education reform efforts could be profoundly enhanced by establishing a suite of national experiments in a variety of disciplines, enabling more faculty, especially

those at primarily undergraduate institutions (PUIs) with limited research resources, to engage in such a project. We anticipate that the development of G-OnRamp, together with our existing curriculum and tools, will facilitate the development of additional CURE projects in genomics. However, the strategy is clearly applicable beyond genomics. We hope that readers in many fields will think creatively about how their own research projects might benefit from educational involvement such as we describe. The solution to many data acquisition and/or data-mining problems may be the students currently enrolled in undergraduate laboratories and classrooms across the country.

Contributing Authors

The full list of authors and affiliations is as follows: Anna Allen, Howard University; Consuelo Alvarez, Longwood University; Sara Anderson, Minnesota State University Moorhead; Gaurav Arora, Gallaudet University; Cindy Arrigo, New Jersey City University; Andrew Arsham, Bemidji State University; Cheryl Bailey, Mount Mary University; Daron Barnard, Worcester State University; Ana Maria Barral, National University; Chris Bazinet, St John's University; Dale Beach, Longwood University; James E. J. Bedard, University of the Fraser Valley, BC; April Bednarski, Washington University in St Louis; John Braverman, Saint Joseph's University; Jeremy Buhler, Washington University in St Louis; Martin Burg, Grand Valley State University; Hui-Min Chung, University of West Florida; Paula Croonquist, Anoka-Ramsey Community College; Scott Danneman, Anoka-Ramsey Community College; Randall DeJong, Calvin College; Justin R. DiAngelo, Penn State Berks; Robert Drew, University of Massachusetts Dartmouth; Robert Drewell, Clark University; Chunguang Du, Montclair State University; Sondra Dubowsky, McLennan Community College; Todd Eckdahl, Missouri Western State University; Heather Eisler, University of the Cumberlands; Julia Emerson, Amherst College; Amy Frary, Mount Holyoke College; Donald Frohlich, University of St Thomas (Houston); Thomas Giarla, Siena College; Anya Goodman, California Polytechnic State University San Luis Obispo; Shubha Govind, City College, CUNY; Elena Gracheva, Washington University in St Louis; Adam Haberman, University of San Diego; Amy Hark, Muhlenberg College; Shan Hays, Western State Colorado University; Arlene Hoogewerf, Calvin College; Laura Hoopes, Pomona College; Carina Howell, Lock Haven University of Pennsylvania; Diana Johnson, George Washington University; M. Logan Johnson, Notre Dame College; Lisa Kadlec, Wilkes University; Marian Kaehler, Luther College; Jacob Kagey, University of Detroit Mercy; Jennifer Kennell, Vassar College; Cathy Silver Key, North Carolina Central University; Melissa Kleinschmit, Trinidad State Junior College; Nighat Kokan, Cardinal Stritch University; Olga Ruiz Kopp, Utah Valley University; Meg Laakso, Eastern University; Wilson Leung, Washington University in St Louis; David Lopatto, Grinnell College; Christy MacKinnon, University of the Incarnate Word; Mollie Manier, George Washington University; Elaine Mardis, Washington University Genome Institute; Juan C. Martinez-Cruzado, University of Puerto Rico at Mayaguez; Luis Matos, Eastern Washington University; Amie Jo McClellan, Bennington College; Gerard McNeil, York College - City University of New York; Evan Merkhofer, Mount Saint Mary College; Hemlata Mistry, Widener University; Elizabeth Mitchell, McLennan Community College; Nathan T. Mortimer, Illinois State University; John Mullican, Washburn University; Jennifer Leigh Myka, Gateway Community & Technical College; Alexis Nagengast, Widener University; Paul Overvoorde, Macalester College; Don Paetkau, Saint Mary's College - Indiana; Leocadia Paliulis, Bucknell University; Susan Parrish, McDaniel College; Celeste Peterson, Suffolk University; Jeff Poet, Missouri Western State University; Johanna M. Porter-Kelley, Winston-Salem State University; Mary Lai Preuss, Webster University; James Price, Utah Valley University; Nicholas Pullen, University of Northern Colorado; Laura Reed, University of Alabama Tuscaloosa; Nick Reeves, Mt. San Jacinto College, Menifee Valley Campus; Gloria Regisford, Prairie View A&M University; Catherine Reinke, Linfield College; Dennis Revie, California Lutheran University; Srebrenka Robic, Agnes Scott College; Jennifer A. Roecklein-Canfield, Simmons College; Ryan Rogers, Wentworth Institute of Technology; Anne Rosenwald, Georgetown University; Michael R. Rubin, University of Puerto Rico at Cayey; Takrima Sadikot, Washburn University; Jamie Sanford, Ohio Northern University; Maria Santisteban, University of North Carolina at Pembroke; Kenneth Saville, Albion College; Stephanie Schroeder, Webster University; Christopher Shaffer, Washington University in St Louis; Karim Sharif, Massasoit Community College; Mary Shaw, New Mexico Highlands University; Matthew Skerritt, Corning Community College; Diane Sklensky, Lane College; Chiyedza Small, Medgar Evers College, CUNY; Sheryl Smith, Arcadia University; Mary Smith, North Carolina Agricultural & Technical State University; Robert Snyder, State University of New York at Potsdam; Eric Spana, Duke University; Rebecca Spokony, Baruch College; Aparna Sreenivasan, California State University Monterey Bay; Joyce Stamm, University of Evansville; Justin Thackeray, Clark University; Jeffrey S. Thompson, Denison University; Chau-Ti Ting, National Taiwan University; Melanie Van Stry, Lane College; Leticia Vega, Barry University; Matthew Wawersik, College of William and Mary; Colette Witkowski, Missouri State University; Cindy Wolfe, Southwest Baptist University; Michael Wolyniak, Hampden-Sydney College; James Youngblom, California State University Stanislaus; Brian Yowler, Geneva College; Leming Zhou, University of Pittsburgh

Acknowledgments

The GEP was originally supported by the Howard Hughes Medical Institute through a Professors grant to S.C.R.E. (#52007051) and is currently funded by

NSF IUSE grant #1431407, with continuing support from Washington University in St Louis. The GEP-Galaxy project is funded by NIH BD2K grant 1R25GM119157.

Resources

[i] http://gep.wustl.edu
[ii] https://cns.utexas.edu/fri
[iii] http://seaphages.org

[1]Washington University, St Louis, MO, USA
[2]Bioinformatics Program, St. Edwards University, Austin, TX 78704, USA
[3]Biology Department, Mount Mary University, Milwaukee, WI 53222, USA
[4]Department of Biological Sciences, Moravian College, Bethlehem, PA 18018, USA
[5]Department of Biology, Adams State University, Alamosa, CO 81101, USA
[6]Department of Biological Sciences, University of Northern Colorado, Greeley, CO 80639, USA
[7]See Contributing Authors.

*Correspondence: selgin@wustl.edu (Sarah C.R. Elgin).

http://dx.doi.org/10.1016/j.tig.2016.11.004

References

1. Leung, W. *et al.* (2015) Drosophila Muller F elements maintain a distinct set of genomic properties over 40 million years of evolution. *G3* 5, 719–740
2. Shaffer, C.D. *et al.* (2014) A course-based research experience: how benefits change with increased investment in instructional time. *CBE Life Sci. Educ.* 13, 111–130
3. Bangera, G. and Brownell, S. (2014) Course-based undergraduate research experiences can make scientific research more inclusive. *CBE Life Sci. Educ.* 13, 602–606
4. Lopatto, D. *et al.* (2008) Undergraduate research. Genomics Education Partnership. *Science* 322, 684–685
5. Shaffer, C.D. *et al.* (2010) The Genomics Education Partnership: successful integration of research into laboratory classes at a diverse group of undergraduate institutions. *CBE Life Sci. Educ.* 9, 55–69
6. National Academy of Sciences, Engineering, and Medicine (2015) *Integrating Discovery-Based Research into the Undergraduate Curriculum: Report of A Convocation,* National Academies Press, (Washington, D.C)
7. Elgin, S.C.R. *et al.* (2016) Insights from a convocation: integrating discovery-based research into the undergraduate curriculum. *CBE Life Sci. Educ.* 15, fe2
8. Rodenbusch, S.E. *et al.* (2016) Early engagement in course-based research increases graduation rates and completion of science, engineering, and mathematics degrees. *CBE Life Sci. Educ.* 15, ar20
9. Lopatto, D. *et al.* (2014) A central support system can facilitate implementation and sustainability of a Classroom-based Undergraduate Research Experience (CURE) in genomics. *CBE Life Sci. Educ.* 13, 711–723