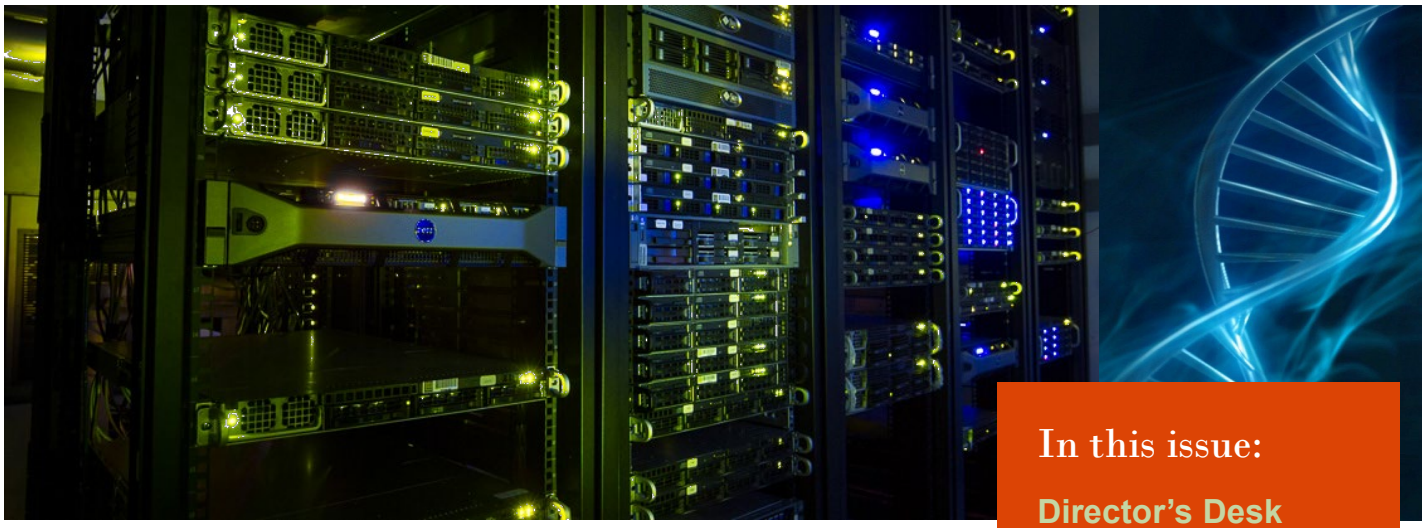
Center for Genome Research & Biocomputing

# TRANSCRIPT *The Next Generation*

FALL 2018, ISSUE 1



In this issue:



## From the CGRB Director's Desk

Hello everybody! With this first issue of "Transcript – the Next Generation" (TNG), the CGRB is re-awakening the regular newsletter "Transcript" that was published (on paper) by the CGRB in the 90's. Our goal is to publish TNG quarterly, to provide an additional, regular means for us to communicate with the CGRB community. This is intended to strengthen how we deliver on our mission of increasing the competitiveness of genome-related and data-driven life and environmental sciences across the campus.

Our approach to our mission is focused around our core values of Innovation, Collaboration, and Customer Service. By definition, delivering on these values requires strong communication with our community members. When we innovate, we need to know that new services and activities that we may plan and implement will make a substantial difference to the research programs of many community members.

*(continued on page 2)*

Oregon State University

CENTER FOR GENOME RESEARCH & BIOCOMPUTING

*(Director's Desk continued)*

Effective collaboration relies on our community knowing the full spectrum of ways we can strengthen your individual research projects, and knowing how we may customize our services and activities to better address your needs. Customer service means working hand-in-hand with each of you to make sure that each project that we assist you with is as successful as possible; this again is only possible with open and effective communication.

With each successive issue of TNG, we will be highlighting a wide variety of activities of the CGRB and its community, including new services and initiatives, that we think you may be excited to hear about. We are also eager to hear from you all about topics you would like to see highlighted in TNG, or guest articles you would like to contribute.

To kick-off TNG, here is a very brief list of some new services and activities that we are excited about.

- Cheaper RNASeq libraries (<$50) on the Lexogen platform
- Pacific Biosciences sequencing upgrade (v6.0) delivering 15-60 Gb per SMRT cell
- Super-fast GPU-computing, including deep learning, on the IBM POWER8 and POWER9 platforms

The following are some new initiatives just starting up:

- Set up of Hi-C genome scaffolding service
- Transition to high density, high speed data storage on the BeeGFS platform
- REDCap web application for securely collecting and managing health data, including protected health information.
- Data Science for the Public Good, combining public service with experiential learning to bring data science and its benefits to rural communities

*- Brett Tyler*

# WORKSHOPS

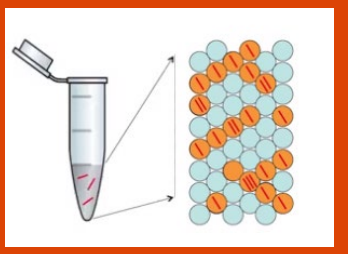
**Winter Term 2019**

Python I & II
RNA Seq I & II

For additional details and registration information see cgrb.oregonstate.edu/training/workshops

# Droplet Digital PCR

Droplet digital PCR from Bio-Rad "provides ultrasensitive and absolute nucleic acid quantification. It is particularly useful for low-abundance targets, targets in complex backgrounds, allelic variants (SNPs), and for monitoring subtle changes in target levels that cannot be detected with real-time PCR." CGRB has a service spotlight on the technology and Bio-Rad has more detailed application information on their web site. For a tour or more information contact Anne-Marie Girard.



# Spotlight: Blouin Lab

*Adult* Biomphalaria glabrata *snails* *Photo provided by Blouin Lab*

Our lab has two main areas of research, both of which involve studying traits under selection and the genes involved in response to selection. Firstly, we study how salmon adapt to hatcheries. We used microsatellite genotyping services to build a multi-generation pedigree of all the steelhead that spawned in the Hood River over 19 years. That pedigree showed that fish of hatchery origin are less fit than hatchery fish, and that the difference results from rapid adaptation to captivity. We are now studying the traits under selection in hatcheries that cause that difference. This work includes RNAseq to identify networks of genes that are differentially expressed between types of fish, and microsatellite genotyping to keep track of families in our experimental hatchery work.

We also study the snail intermediate host for schistosomiasis, a parasitic disease of humans in the tropics. Humans become infected by contact with water in which infected snails are shedding the parasite. We are exploring genetic methods for breaking the cycle of transmission at the snail stage. Towards that end, we have been using whole-genome sequencing for genome-wide association studies (GWAS) to identify regions of the snail genome in which allelic variation controls resistance to infection by the parasite. A new PacBio assembly of the snail genome created at the CGRB has been essential to this work.

To date we have identified a new class of pathogen recognition proteins that regulate shedding of the parasite. Using the Illumina metagenomics sequencing protocol, we showed that allelic variation at this locus also affects the microbiome of the snail. This raises the intriguing possibility that these genes control schistosomes, not directly, but indirectly by altering the microbiome environment experienced by the invading parasite. All of the methods mentioned above were done at the CGRB, and we continue to rely on their services as we continue both lines of research.

Michael Blouin, Professor
Integrated Biology
Cordley Hall

## Out of the Past

*From the July 1992 edition of the* CGRB transcript *newsletter*

*The move of the Center for Gene Research and Biotechnology office to the new Agriculture and life Sciences (ALS) Building is now complete. The space provides for quite efficient and very pleasant new quarters located on the third floor on the east side of the building, near the elevator (Room 3021). For the present, we are also home for the administrative aspects of the Molecular/Cellular Biology graduate program, and we would like to welcome Maureen Eburne to our staff as she will now be filling in for both MCB and Genetics. We are very lucky to have been given such a wonderful setting for our permanent home.*

# techtips

## *CGRB's SGE batch queueing system*

by Shawn O'Neil, PhD, Advanced Cyberinfrastructure Teaching Facility Manager

The CGRB research cyberinfrastructure runs a batch queueing system called SGE (Sun of Grid Engine, replacing the no-longer-supported Sun Grid Engine). This system allows users to submit "jobs"–program commands or scripts, along with information about the CPU and RAM resources needed to run the command. A dedicated scheduler organizes these jobs, and distributes them to computers with the right resources at the right time.

The CGRB is not alone in its use of a batch scheduler; most high-performance computing (HPC) clusters work similarly. What sets the CGRB apart are the tools we develop to help researchers make effective use of this powerful resource. These include SGE_Batch, SGE_Array, and helper tools such as SGE_Avail and SGE_Plotdir.

SGE's built-in job submission tool, qsub, requires that researchers write a job submission script specifying various options using a specialized syntax, as well as the commands to be run. SGE_Batch, the CGRB's first customized submission tool, automates this process by allowing the researcher to specify just the command and relevant job parameters directly on the command line.

While SGE_Batch is great for submitting individual jobs, HPC clusters support running many jobs in parallel. SGE supports large job sets via "array jobs," but the syntax for using these in both qsub and SGE_Batch requires yet more cumbersome scripting. Many a researchers' first instinct is thus to repeatedly call SGE_Batch, but this degrades performance for the jobs and the cluster because the scheduler is unaware that the jobs can be treated as a set. SGE_Array makes it easy to submit large sets of jobs properly, by just specifying a file of commands to run.

Additional options include the ability to throttle how many jobs may run simultaneously, as well as to submit and "hold" job sets until another job or job set completes.

Given that jobs can only run when the requested resources are available, two common questions arise: 1) How can I find out what resources are available to me? 2) How can I find out what resources a recent job actually used? SGE_Avail answers the first question by collating information about the status of the cluster in an easy-to-read format. SGE_Plotdir answers the second, by analyzing job logs (produced by SGE_Batch and SGE_Array) and reporting time and RAM used.

Batch queuing systems are obtuse, but they nevertheless provide the computational horsepower needed for modern data analysis. Making these resources accessible to the OSU research community is a top priority of the CGRB. In fact, CGRB Analyst Ed Davis is currently investigating the Python-based Snakemake for job-wrangling of entire pipelines. Perhaps someday, a full RNA-seq analysis or genome annotation will be a simple call to SGE_Snake!

### BUG Winter Term 2019

Jan 9: Dana Gibbon, CGRB
*Wading through GATK's best practices and more*
Jan 23: Molly Megraw, Botany & Plant Pathology
Feb 6: Justin Preece, Jaiswal Lab
Feb 20: Noor Al-Bader, Jaiswal Lab
Mar 6: Parul Gupta, Jaiswal Lab
Mar 20: Kelly Vining, Horticulture

BUG is at noon in ALS 3005

BUG
Bioinformatics User Group

SPRING CONFERENCE
April 19, 2019

FALL CONFERENCE
TBA

Center for Genome Research & Biocomputing
3021 Ag & Life Sciences Bldg.
Oregon State University
Corvallis, OR 97331
541-737-3347
cgrb.oregonstate.edu
@CGRBOregonState

CENTER FOR GENOME RESEARCH & BIOCOMPUTING

Comments and questions about this newsletter can be sent to Jeannine Cropley, jeannine@cgrb.orgonstate.edu