Summary of National Center for Genome Analysis Support (NCGAS) 2018-2019 *de novo* Transcriptome Workflow and Workshops

Sheri A. Sanders
Thomas G. Doak
Carrie L. Ganote
Bhavya Papudeshi

Indiana University
PTI Technical Report

Nov 23, 2019

Intent:
The National Center for Genome Analysis Support (NCGAS) held a workshop entitled "*de novo* Assembly of Transcriptomes using HPC Resources" on three occasions: April 30th, 2018 through May 1st, 2018; October 1st through October 3rd, 2019; and April 29th, 2019 through May 1st, 2019.

These workshops were in serving NCGAS's mission of enabling the biological research community to analyze, understand, and make use of the genomic information now available by packaging our now seven years of experience assisting with *de novo* transcriptome assemblies and running High Performance Computing (HPC) resources into a documented, easily approachable workflow for our users. The workshop covered common questions and problems that our users have had in HPC (such as job handling, resource availability, data management, and troubleshooting) and in the construction of transcriptomes (such as software choices, combination of assemblies, and downstream analyses). The three-day workshops also highlighted the available resources for US scientists, concentrating heavily on available XSEDE resources for analyses, visualization, and archiving of data.

While the primary goal of the workshops was to provide advanced training on a common methodology, NCGAS also sought to implement an internal workflow for converting knowledge and experience gained from working on a breadth of projects with our users into easily transferable products (workflow/workshop).

Methodology:
1) Compile of our previous presentations, scripts, and tickets on the topics
NCGAS has spent years presenting on transcriptome assembly and analyses at numerous national conferences (i.e. Galaxy, Plant and Animal Genome) and guest lectures (i.e. Bethune-Cookman, Mount Desert Island Biological Laboratories Environmental Genomics Workshop). While independently desperate, the library of presentations, demos, and lectures compiled over the last seven years of NCGAS covered most topics in *de novo* assembly. NCGAS's work with Keithanne Mockaitis has resulted in an established workflow for assembly. As such, we had scripts pre-made for running almost all of the pertinent software, and hundreds of tickets covering the most common problems in working in HPC. This pool of resources served as the material to create the workflow and workshop without much additional developmental effort.

2) Clean up, annotate, and test to create best practices workflow
We took our scripts, made them as generic and readable as possible for the two main machines NCGAS clients utilize - IU's Carbonate and PSC/XSEDE's Bridges. All steps to the multi-software assembly were organized and linked to form the workflow. Two major considerations were made here. First, that this was not meant to be "push button". We want users to get familiar with the job scripts and commands while taking away the stress of building them from scratch. Common methodology was made the default, but links and documentation for other situations (polyploidy, stranded sequencing, etc.) is listed in documentation.

Second, despite lowering the learning curve to get started running the software, we wanted to preserve best practices - specifically that multiple parameters and multiple assemblers should be used to account for individual software biases. As a result, four assemblers (SOAP *de novo*, Trinity, TransABySS, and Velvet/Oases) were included with several kmers for each assembler, resulting in 19 individual assemblies. These were then set up to be merged by concordance via the software EviGene.

This step resulted in an easy-to-run, easy-to-read, easy-to-modify workflow that followed current best practices.

3) Test workflow on our own work with clients

During the first three months after development, this workflow was used in house to test for bugs, usability, etc. on client projects NCGAS was contracted to complete. This step resulted in minor adjustments of the workflow.

Evaluation:

The workflow was developed by one team member and handed off to the other team members to work with. Direct feedback was handled internally but allowed for testing on multiple projects.

4) Test workflow with independently working clients

After guest lecturing at Mount Desert Island Biological Laboratory's Environmental Genomics course, NCGAS received several requests for assistance with assembly of *de novo* transcriptomes using Illumina data. We used this opportunity to beta test our workflow with students at least semi-familiar with command line. These individuals got support, experience, and training, while NCGAS got feedback on the workflow.

Evaluation:

Direct feedback was solicited from clients in the form of tickets and informal survey. Comments on design and general organization of the workflow were solicited during presentation of the workflow at a demo at Plant and Animal Genomes 2018.

5) Design workshop around presentations and workflow

Once we had the workflow solidified, we built a workshop around the use of it. We ordered previous talks and demos into a logical order to cover the material and common problems from our library of tickets. We used publicly available data as a test set for using the assembly and talked about our experience with downstream analyses of the transcriptomes we have generated from this workflow.

This step resulted in a scaling up of knowledge transfer, further documentation of the workflow and surrounding topics, and conversion of the information into digital resources that we can point new clients to and low effort development of an in-person workshop.

Evaluation:

Pre- and post-surveys for the workshop were designed to measure comfort levels with included topics, comment on the successful and superfluous aspects of the workshop and suggest future topics for workshops.

6) Revision of workshop and workflow

The surveys before and after the workshop were used to make changes to the cirriculum and the tools included. We added annotation handling and more downstream analyses (KEGG Pathways) in response to common requests. We held one-on-one consulting session in the second iteration as they were requested, however, these were not taken advantage of, and we opted to not offer these sessions going forward. Instead this time was used to include more material on the whole. We also recorded and made available videos of all lectures, which are now available on YouTube.

Evaluation:
Pre- and post-surveys for the workshop each time so we can continue to track the effects of changes.

Recruitment for Workshop:

NCGAS contacted all current and previous clients and all participants of previous workshops NCGAS via our mailing list. Information was posted via the NCGAS twitter and Facebook page, Indiana University's IT News and Events page as well as the Evolution Directory List Serve. We also contacted The IU Biology department and the XSEDE Campus Champions list directly.

All applicants were directed to a survey (Appendix A) allowing them to provide information on their background, projects, and status (faculty, staff, PhD. student, etc.). All applicants were independently ranked by all NCGAS staff on the following criteria:

- Appropriateness of project
- Appropriateness of learning goals in relations to this workshop
- EPSCoR status

The ranks were averaged across the four reviewers and the 40 top-scoring applicants were offered space in the workshop each time. Several had to decline each round for various reasons and were reconsidered in subsequent iterations of the workshop.

Use of NSF Money:

NSF money was used from our grant (ABI-1458641/ABI-1759906) in two ways:

- Development - NSF funding of NCGAS activities and travel directly resulted in the development of the pool of presentations, lectures, demos, scripts, and tickets that formed the basis of the workflow/workshop.

- Participant Travel (Spring 2018 workshop only)- NSF funding for participant travel to workshops was used to support 26 student/staff travel vouchers. The workshop itself was free of charge, as NCGAS is already supported by NSF funds

and we did not want participants to use their NSF grant money to pay for time already covered by the NSF.

Products and Outcomes:

The primary products of the work were the workflow and the workshop. The workflow allows us to streamline work to publication and posters, as well as presentations and websites on the tool. The workshop provides contact with our clients, generates more clients, and transfers knowledge in a quantifiable way.

1) Publications - Internal use of workflow

Petek, M, Zagorscak, M, Ramsak, Z, Sanders, S, Tseng, E, Zouine, M, Coll, A, Gruden K. (in review) Cultivar-specific transcriptome and pan-transcriptome reconstruction of tetraploid potato. GigaScience. Available at https://www.biorxiv.org/content/10.1101/845818v2.abstract.

Cinel, S.D., Taylor, S.J. (2019) Prolonged bat call exposure induces a broad transcriptional response in the male fall armworm (Spodoptera frugiperda; lepidoptera: Noctuidae) brain. Frontiers in Behavior Neuroscience. https://doi.org/10.3389/fnbeh.2019.00036

Wong, J.M., Gaitan-Espitia, JD, Hofmann, GE. (2019) Transcriptional profiles of early stage red sea urchins (Mesocentrotus franciacanus) reveal differential regulation of gene expression across development. Marine Genomics. https://doi.org/10.1016/j.margen.2019.05.007

Rivera-García L, R Rivera-Vicéns, A Veglia, NV Schizas (2019) De novo transcriptome assembly of the digitate morphotype of Briareum asbestinum (Octocorallia: Alcyonacea) from the southwest shelf of Puerto Rico. Marine Genomics https://doi.org/10.1016/j.margen.2019.04.001

Veglia A, Hammerman NM, Rivera Vicens RE, NV Schizas (2018). De novo transcriptome assembly of the coral Agaricia lamarcki (Lamarck's sheet coral) from mesophotic depths in southwest Puerto Rico. Marine Genomics https://doi.org/10.1016/j.margen.2018.08.003

Yang, T., L. Fang, S. Sanders, S. Jayathi, G. Rajan, R. Ppdicheti, S.K. Thallapuranam, K. Mockaitis, F. Medina-Bolica. (2017). Stillbenoid prenyltransferases define key steps in the diversification of peanut phytoalexins. Journal of Biochemistry. Retrieved from http://www.jbc.org/content/early/2017/11/17/jbc.RA117.000564

2) Dissertaions using workflow

Wong, JM. (2019) Investigating the response of sea urchin early developmental stages to multiple stressors related to climate change. University of California, Santa Barbara. https://search.proquest.com/docview/2311653028?pq-origsite=gscholar.

Rivera-Garcia, L. (2019) Comparative transcriptomeics of the two distinct morphologies of the Carribean octocoral Briareum asbestinum. University of Puerto Rico Mayaguez. https://scholar.uprm.edu/handle/20.500.11801/2447

3) Presentations using workflow

Mansfield, C., Tseng, C., Sanders, S., Custer, TW, Custer, CM, Matson, CW. (2019) Genetic diversity comparison of tree swallow populations in the Great Lakes region using RNA-sequencing. SETAC North America 40th Annual Meeting.

Song, J., Brill, R.W, McDowell, J. (2019) 'Investigating local adaptation and placticity of an estuarine-dependent teleost, Spotted Seatrout (Cyanoscion nebulosus). In American Fisheries Society and The Wildlife Society 2019 Join Annual Conference. Retrieved from https://afs.confex.com/afs/2019/meetingapp.cgi/Paper/40622.

Papudeshi, B., Chafin, T., Sanders, S., Ganote, C., Reshetnikov, A., Sokolov, S., Doak, T., Pummil, J.F., Douglas, M.R., Douglas, M. (2019) 'Genome and transcriptome analysis of fish tapeworm Nippotaenia percotti through scientific collaboration between research labs and national cyberinfrastructure.' In American Fisheries Society and The Wildlife Society 2019 Join Annual Conference. Retrieved from https://afs.confex.com/afs/2019/meetingapp.cgi/Paper/39888.

Papudeshi, B., Sanders, S., Ganote, C., Doak, T., Chafin, T., Reshetnikov, A., Sokolov, S., Pummil, J., Douglas, M., Douglas, M. (2019) 'The Genome of Fish Tapeworm Nippotaenia percotti as a Potential Bookmark for Gene Loci that Facilitates Anthropogenic Infection.' Plant and Animal Genome XXVII.

Sanders, S., Papudeshi, B., Ganote, C., Doak, G.T. Mansfield, C., Tseng, C. Y., Custer, T., Custer, C., Matson, C. (2019) 'Population Genetics of Tree Swallows, in Collaboration with NCGAS'. Plant and Animal Genome XXVII.

Ganote, C., Sanders, S., Wu, L., Doak, T., & Mockaitis, K. (2018). Solving the challenges of complex genome analysis collaborations on-line using XSEDE resources. In Plant and Animal Genomics 2018, San Diego, CA. Retrieved from http://hdl.handle.net/2022/21903

Sanders, S., Podicheti, R., Yang, T., Fang, L., Jayanthi, S., Rajan, G., Kumar, T. K. S., Medina-Bolivar, F., & Mockaitis, K. (2018). Stilbenoid prenylation pathway discovery in peanut using targeted transcriptomics. In Plant and Animal Genomics 2018, San Diego, CA. Retrieved from http://hdl.handle.net/2022/21902

Sanders, S., & Pfrender, M. (2017). *de novo* assembly and annotation of Ambystoma laterale and Ambystoma jeffersonianum transcriptomes: the first steps

toward investigating polyploid salamander expression. In Evolution Conference. Retrieved from https://scholarworks.iu.edu/dspace/handle/2022/21599

4) Presentations on Workflow

Sanders, S., Papudeshi, B., Ganote, C. Doak, T. (2019) Transcriptomes are easy start points in genomic research and NCGAS can help. In American Fisheries Society and The Wildlife Society 2019 Join Annual Conference. Retrieved from https://afs.confex.com/afs/2019/meetingapp.cgi/Paper/40300.

Sanders, S., Papudeshi, B., Ganote, C., Doak, T. (2019) NCGAS Makes Robust Transcriptome Assembly even easier with added features to an accessible de no transcriptome assembly workflow, in Plant and Animal Genome Conference 2019. Available at: https://plan.core-apps.com/pag_2019/abstract/6e070c0ebde6a5b908f9f08fb2eecd05 (Accessed: 28 January 2019).

Sanders, S., Ganote, C., & Doak, T. (2017). *de novo* Transcriptome Assembly. In Mount Desert Island Biological Laboratory. Retrieved from https://scholarworks.iu.edu/dspace/handle/2022/21645

Sanders, S., Ganote, C., Papudeshi, B., Mockaitis, K., & Doak, T. (2018). NCGAS makes robust transcriptome analysis easier with a readily usable workflow following *de novo* assembly best practices. In Plant and Animal Genomics 2018, San Diego, CA. Retrieved from http://hdl.handle.net/2022/21904

5) Websites for available data

GitHub: https://github.com/NCGAS/Transcriptome-assembly-workshop-2018; https://github.com/NCGAS/Transcriptome-Assembly-Workshop-Fall-2018; https://github.com/NCGAS/Transcriptome-Assembly-Workshop-Spring-2019

Workflow Documentation Page: http://ncgas.org/WelcomeBasket_Pipeline.php

YouTube Playlist of Videos: https://www.youtube.com/playlist?list=PLqi-7yMgvZy_IaAiPG89AX2cQH2JY4Ifo

6) Beta test commentary excerpts - full comments in Appendix B

"The use of this pipeline has saved me tons of time from having to figure out the script for each assembly program and it is VERY easy to use, especially for a person like myself who barely understands Linux!"

“If this pipeline was not available, I would have most likely used only package and at one kmer size for my assembly, and it would have probably taken me just as long to figure out and run.“

7) Workshop Results

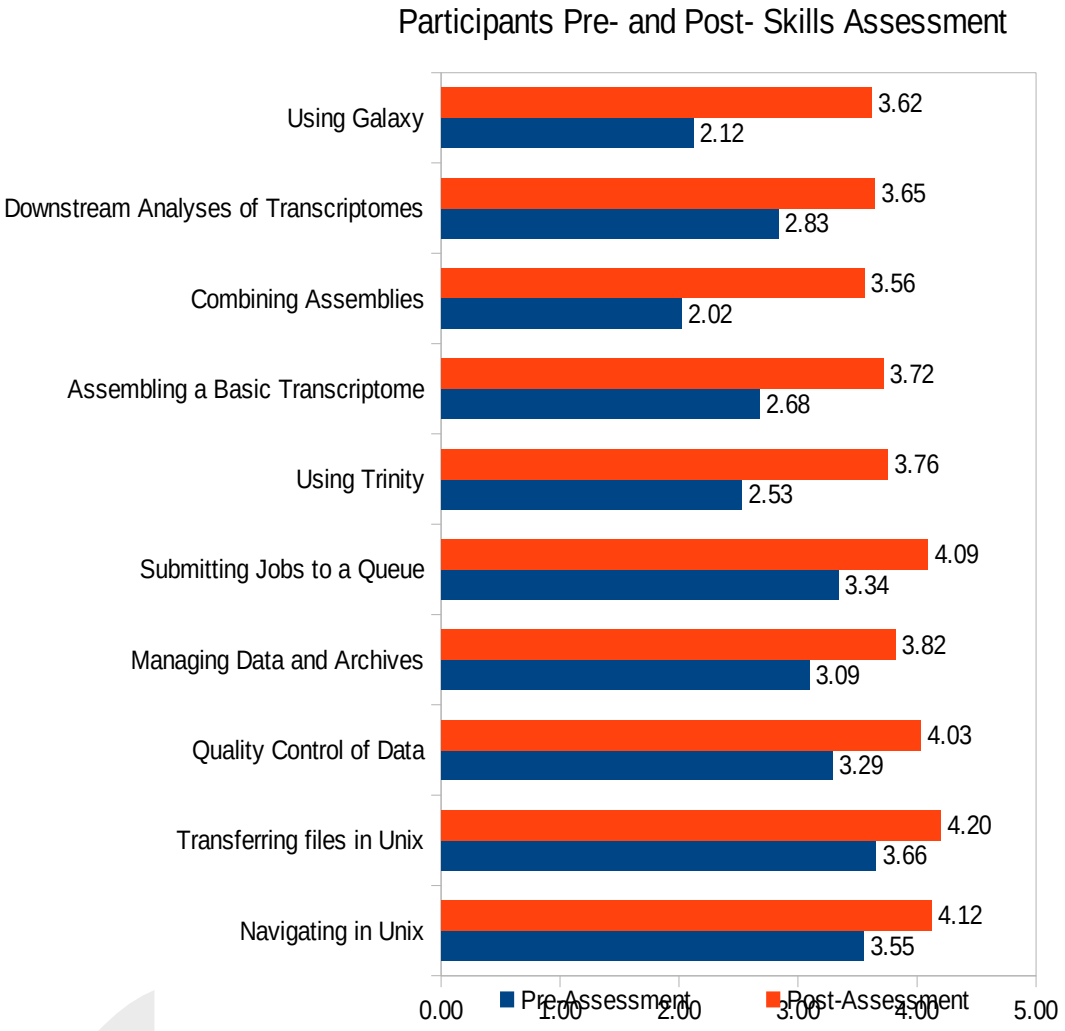
- Interest Level – We had 69, 44, and 34 applicants for the 30 seats in the three workshops respectively. These applicants came from a total of 69 institutions in 32 states/1 territory (16 EPSCoR states/1 territory), and seven foreign countries.

- Workshop Attendance – We've had a total of 82 people attend the workshop from 43 institutions in 26 states/1 territory (11 EPSCoR states/1 territory), as well as one student from the UK. The sex ratio was about even at 33 male and 32 female participants (of the 65 reported on the survey). There were 36 white non-hispanic, 4 Hispanic, 17 Asian, and 3 African American participants (of the 60 reported on the survey).

- Efficacy – In general, the format for the workshop was well received, with only 4 participants stating they would prefer a non-in person format (specifically live webinars). **Overall, the comfort level of the participants increased 30% in general transcriptome assembly.** When queried about ten individual skills before the workshop, five were in the "no hands-on experience" range and five were in the "can run limited examples" range. The average skill level was 2.91 out of 5. After the two-day workshop, six of these skills were on average in the "ability to run limited examples", and four were in the "ability to run realistic examples" range, with an average skill level of 3.86 out of 5 – a full 14% more confident in their skill set and on average able to run real data.

Comfort in overall transcriptome assembly - 3.73 to 5.83 on 7-point scale (increase of 30% in comfort level)

Comfort level in each skill – on a 5-point Likert scale as defined below:
1 – No previous experience of knowledge
2 – Knowledge of its function, but no hands-on experience
3 – Ability to run very limited examples, such as small data sets and tutorials
4 – Ability to run more realistic examples, such as real data
5 – Ability to troubleshoot tasks for myself and others



- Select Comments on Efficacy of workshop – In free commentary, it appears our workshop was well received and perceived as a useful two days in boosting confidence, knowledge, and size. Many participants were surprised to learn about XSEDE (one had heard of it prior to the course) and all the resources available to them through NSF. Many were surprised (and happy) to hear about the breadth of services NCGAS provided as well - we've received 24 new allocations from participants of the workshops.

  Responses to "Favorite Part":
  "This was the first workshop of its kind that I have attended and one of the most useful workshops in my doctoral degree. So, I received everything and think that everything is very important and directly applicable to my research."

  "Fantastic workshop. I learned how to do in 2 days what I have been trying to do on my own for more than 5 months (make a Trinity pipeline -> annotation -> EdgeR analysis)"

“Successfully completing this exercise gave me the confidence to tackle and use published bioinformatics pipelines that I would otherwise have been too intimidated to try and run.”

“You guys really know how to handle this type of workshop and I think the teacher to student ratio was spot on. I've been to some of these things where we get through less than this in much more time. Of course that was all on command line and not Galaxy, but you guys are seriously working magic! Galaxy is sooooooooooo nice!”

“Being able to run through the entire pipeline myself in a streamlined fashion - I have attended workshops where there was no flow to what we were learning and everything seemed very disjointed, and in those instances, it has always been very hard to follow what we are actually supposed to be doing, but that was not the case here.”

“I liked having the pipeline- hearing how everything works together and the pros/cons of different analyses was incredibly helpful. Reading about each program is daunting and honestly I do not understand the technicalities of each. Having a working model was a huge booster.”

“Leaving feeling that it is possible to do de novo assembly myself. “

Responses to “Most surprising thing you learned”:
“The resources that are available to NSF researchers. I've been struggling on my own with my limited coding knowledge for quite some time. NCGAS is an absolutely amazing resource and cuts down on wasted time for me. It all makes analyzing transcriptomes and genomes much easier, faster and more efficient, which contributes to the scientific community overall.”

“The number of NSF-funded super-computing resources available to researchers; the presenters did a great job of publicizing such resources”

“you being available for anyone when they need some help to move to the next step and availability to answer the questions related our independent work as well”

“The new program that we discussed, "evigene" is a concept that I had discussed with some of my associates, and to see it actually developed and in action is extremely exciting. “

“Well designed activities could help users of computer clusters to get a better understanding of job submission.”

“A well-organized pipeline is a beneficiary for data archive and troubleshooting. “

"The wrapper scripts and the pipelines were incredibly impressive and well-engineered."

"The most surprising thing to me is that it is now possible to do the complete assembly remotely, and I am not restricted by working at a small college without computing capabilities. "

- Internal comments on workshop structure efficacy - Having five instructors in the room at all times was critical. We had the NCGAS main staff present - Tom Doak (Manager), Carrie Ganote (Bioinformatic Analyst), Bhavya Papudeshi (Bioinformatic Analyst), Sheri Sanders (Bioinformatic Analyst). Having one instructor presenting and the rest moving around to prevent any participants from falling behind kept everyone on track and made sure everyone was able to complete the exercises. Building in obvious outputs to some activities (graphical output to screens for Galaxy and Data Transfer) facilitated this further, as staff could walk around the room and easily survey when everyone had hit a checkpoint. We plan to add another instructor to help this go more smoothly in the future.

  NCGAS has developed a close relationship with the rest of the HPC center at IU over the years, which added an additional twist to this workshop that is often missing from similar workshops. HPC topics learned from working with the HPC center were integrated into the analysis, emphasizing the importance of clusters as useful tools that can be used more efficiency with more training/knowledge. Novice job handling and data management can add unnecessary time to analyses and burden on systems, but these topics are seldom taught explicitly to biological users. Covering this material was appreciated by participants, as several users sighted these topics as the most useful.

  The close association with HPC staff was also helpful in allowing NCGAS to pull in system administration staff to quickly fix a system-side issue immediately (while staying on schedule) and talk to the participants when technical questions arose. The HPC staff also provided us with knowledgeable guides for tours of the IU Data Center (a much appreciated activity) and helpful comments during the evening chats over dinner. Lightly including the HPC personnel in the workshops helped expose biologists to conversing with and humanizing the computer scientists.

- Changes through Iterations – After the first workshop, we added a consulting time by request. It was discontinued after one workshop, as it was not taken advantage of. Instead we used the extra time to add more material on annotation, pathway analysis, and other downstream analyses which was well received.

Summary of Results and Future Plans:

We were able to synthesize and package our previous knowledge and work into a workflow for *de novo* transcriptome assembly and a workshop based around it. The

three-day workshops were well attended and received. We were able to elevate the skill set across a diversity of HPC and bioinformatic skills for a diverse set of students, staff, and professors across the country. We also introduced almost all of the participants to XSEDE resources to power their new skills, toured the data center while talking about what it takes to run and manage these machines, and introduced them to some HPC staff to help provide the biologists with a better comfort level in working with HPC systems/staff.

Many publications, dissertations, and presentations are coming out of this pipeline. At present, **six publications** have used the workflow (which will continue to increase), **two dissertations** have used the workflow as a substantial portion of the research, and **twelve presentations** have been given on the material. **Overall, 39 authors have publications with this workflow in less than two years.**

We have started to offer a second workshop on metagenomics using this workflow design protocol. It is currently being adjusted for a second round with revisions, to be offered again in April 2020. We are also in the process of applying a similar workflow design protocol to transfer our knowledge on genome annotation, though we are still in an early stage as we work with collaborators to develop the material. Basic skills workshops (i.e. R, python) are also underway. A separate report on the scaling of these workshops will follow.

APPENDIX A: Application for Workshop

Spring 2018 Workshop Registration Form

NCGAS is preparing a 2-day workshop in Spring 2018 to provide training in basic bioinformatics concepts, with hands-on tutorials and walkthroughs. Our staff will share our collective expertise to help you become more familiar with High Performance Computing (HPC) environments, available lab technologies to answer your experimental questions, web-based visual bioinformatics tools such as Galaxy, data management tips and troubleshooting approaches. We have room for ~30 applicants so this form will help us make tough choices should we get too many responses!

NOTE: This workshop is now at capacity. If you would like to be added to the waiting list, please feel free to fill out the form!

1) First Name	
2) Last Name	
3) Best Contact Email	
4) Institution/University/Organization Name	
5) State or province of your institution	
6) What is your current status?	Student - Undergrad Student - Masters Student - PhD Post-doc Faculty

Staff

Other

7) List any grants with which you're involved (investigator, researcher, consultant, etc):

Expand

8) Are you (or your institution) able to cover the full cost of lodging + transportation + meals during the workshop?

Yes

No

reset

9) What is the biggest challenge you've faced so far when dealing with bioinformatics?

Expand

10) Describe in 1-2 paragraphs what you hope to learn during this workshop.

Expand

11) Describe your current work in 1-2 paragraphs - are you part of a lab project, do you lead your own work, or are you studying something particularly compelling?

Expand

APPENDIX B: Beta testing Commentary

REQUEST:
Hey all!

The three of you were using my pipeline, and I was wondering if you had a second to respond to a couple questions:

1) Has this made it easier to run the software packages?

2) Got a one sentence review you want to leave?

3) Suggested Improvements?

RESPONSE:
FIRST:
Hi Sheri,

Sure thing! To be honest, I haven't finished running through the pipeline yet. We had some delays getting data, and the Thomas fire really threw us for a loop here in Santa Barbara. But now I'm finally catching back up on things!

1) Definitely. If this pipeline was not available, I would have most likely used only package and at one kmer size for my assembly, and it would have probably taken me just as long to figure out and run.

2) The pipeline has compiled and organized everything I needed, saving me time and frustration on a process that otherwise would have taken me months to disentangle.

3) The biggest problem I've had is running into virtual memory and wall time limits, as it takes some time for before it's actually hit those limits and I've been alerted the job has failed, then I adjust the parameters and try again. I've tried setting vmem to 500, but I've still run into limit problems. For some of the runs, I've actually tried breaking up the code into separate jobs (i.e., running 34, 45, and 55 separately). I'm not sure if that's the best way to go about things?
One suggestion might be to include in the READMEs, what it should look like when steps are done running (i.e., what files you should expect to have to know everything has run okay).

Thanks again for all of your help, Sheri!
Best,
Juliet

SECOND:
1) Has this made it easier to run the software packages?

Extremely! All I have to do is tell it my file names and run each program. This is so much easier than figuring out the script for each program.

2) Got a one sentence review you want to leave?
The use of this pipeline has saved me tons of time from having to figure out the script for each assembly program and it is VERY easy to use, especially for a person like myself who barely understands Linux!

3) Suggested Improvements?
So far I have none.

Heather Walsh

APPENDIX C: Workshop Schedule (Spring 2019 version)

Day 1	Activity	Lead
8:30am	Registration	All
9:00am	Introduction to NCGAS and staff	Tom
9:30am	Introduction to Clusters and Other Resources	Sheri
11:00am	BREAK	
11:15am	Optimizing Jobs	Carrie
12:15pm	Data Center Tour/Lunch	
2:15pm	Data Management and Movement	Bhavya
3:45pm	BREAK	
4:00pm	Introduction to Assembly and Pipeline	Sheri
5:05pm	BREAK	
5:10pm	Publicly Available Resources	All
6:00pm	We will have reservations at a local tavern, Nick's. We will be joining you to chat/relax, but you will be responsible for paying for your meal and drinks.	

Day 2	Activity	Lead
9:00am	Introduction to Data	Tom
9:45am	BREAK	
10:00am	Using Galaxy	Carrie
11:30am	BREAK	
11:45am	Hands on Pipeline Use	Sheri/All
12:45pm	Lunch	
1:45pm	Hands on Pipeline Use Cont’d	Sheri/All
2:30pm	BREAK	
2:45pm	Annotation Demo	Sheri
3:45pm	Downstream Analyses Discussion	Sheri
4:15pm	BREAK	
4:30pm	KEGG Demo	All
6:00pm	We will meet at a favorite local bar, the Upstairs. They do not serve food, but there is plenty of food nearby that you can bring with you! Again, we will be joining you to chat/relax, but you will be responsible for paying for your	

	meal and drinks.	
Day 3	**Activity**	**Lead**
9:00am	DE Analysis Introduction	Sheri
10:00am	DE Demo- Galaxy	Carrie
11:00pm	BREAK	
11:15am	DE Demo- Command line	Sheri
11:45am	Final Remarks	Tom
12:00 pm	Lunch: On your own, there are a couple restaurants nearby you can check out. A few of them are listed in the Welcome document given during registration	

APPENDIX D: Pre and Post Workshop Surveys (Spring 2018 version)

NCGAS Pre-Workshop Survey

Start of Block: Informed Consent

Q10 Thank you for registering to attend the NSF-funded National Center for Genome Analysis Support (NCGAS) Spring Workshop. To inform workshop content, we ask that you participate in this short survey for the purpose of gauging the needs and experience levels of attendees. Data will also be used in evaluating workshop effectiveness and future workshop content. Your participation, as well as all survey questions, are optional and your responses are confidential. Data will be reported in the aggregate without any identifying information that you choose to, or inadvertently, disclose. Your decision to participate will not in any way affect your relationship with the NSF, the NCGAS project, the Pervasive Technology Institute, or Indiana University.

For questions about this survey, including problems with accessing the survey, please contact cesg@iu.edu and reference protocol #1804120218/exempt, approved on April 26, 2018, by the Indiana University Institutional Review Board.

Do you agree to participate?

1. Yes (1)
2. No (2)

Skip To: End of Block If Thank you for registering to attend the NSF-funded National Center for Genome Analysis Support (N... = Yes
Skip To: End of Survey If Thank you for registering to attend the NSF-funded National Center for Genome Analysis Support (N... = No

End of Block: Informed Consent

Start of Block: Default Question Block

Q1 Have you ever used command line before?

3. Yes (1)
4. No (2)
5. Not sure (3)

Q2 How comfortable are you using the computer via command line?

6. Extremely uncomfortable (1)
7. Moderately uncomfortable (2)
8. Slightly uncomfortable (3)
9. Neither comfortable nor uncomfortable (4)
10. Slightly comfortable (5)
11. Moderately comfortable (6)
12. Extremely comfortable (7)

Q3 Have you previously worked with Unix?

13. Yes (1)
14. No (2)
15. Not sure (3)

Q4 Are you or are you currently working with any bioinformaticians?

16. Yes (1)
17. No (2)

Q5 Which, if any, bioinformatics tools have you used?

_____

_____

_____

_____

_____

Q6 Do you have existing data that you plan on assembling?

18. Yes (1)
19. No (2)

End of Block: Default Question Block

Start of Block: Data questions

Q7 Please provide a brief description of the data you plan on assembling.

_____

_____

_____

_____

_____

Q8 Are there any specific issues you have encountered with your data to-date?

_____

_____

_____

_____

_____

End of Block: Data questions

Start of Block: Skill Level

Q9 What is your skill level for each of the following:

	No previous experience or knowledge (1)	Knowledge of its function, but no hands-on experience (2)	Ability to run very limited examples, such as small data sets and tutorials (3)	Ability to run more realistic examples, such as real data (4)	Ability to troubleshoot tasks for myself and others (5)
Navigating in Unix (1)	20.	21.	22.	23.	24.
Transferring files in Unix (2)	25.	26.	27.	28.	29.
Quality Control of Data (3)	30.	31.	32.	33.	34.
Managing Data and Archives (4)	35.	36.	37.	38.	39.
Submitting Jobs to a Queue (5)	40.	41.	42.	43.	44.
Using Trinity (6)	45.	46.	47.	48.	49.
Assembling a Basic Transcriptome (7)	50.	51.	52.	53.	54.
Combining Assemblies (8)	55.	56.	57.	58.	59.
Downstream Analyses of Transcriptomes (9)	60.	61.	62.	63.	64.
Using Galaxy (10)	65.	66.	67.	68.	69.

Page
Break

Q17 How comfortable are you assembling a transcriptome?

70. Extremely comfortable (1)
71. Moderately comfortable (2)
72. Slightly comfortable (3)
73. Neither comfortable nor uncomfortable (4)
74. Slightly uncomfortable (5)
75. Moderately uncomfortable (6)
76. Extremely uncomfortable (7)

End of Block: Skill Level

Start of Block: Demographics

Q11 Please describe your institution/organization: *Please select all that apply.*

1. Institution located in an EPSCoR state (AL, AK, AR, DE, HI, IA, KS, KY, LA, ME, MS, MT, NE, NV, NH, NM, ND, OK, RI, SC, SD, VT, WV, WY) (1)
2. Minority-Serving Institution (MSI) (2)
3. Associate's College (all degrees are at the associate's level) (3)
4. Baccalaureate College/University (4)
5. Master's College/University (5)
6. Doctorate-Granting University (6)
7. Teaching-Focused Institution (7)
8. Research-Focused Institution (8)
9. Government Lab or Center (9)
10. High performance computing resource provider (e.g. NCSA, TACC, etc.) (10)
11. Non-Profit Organization (non-academic) (11)
12. Corporate/Industrial Organization (12)

Page Break

Q12 What is your gender? *Select all that apply.*

13. Female (1)
14. Male (2)
15. Non-Cisgender (3)
16. Other (4)
17. Prefer not to disclose (5)

Q13 What is your ethnicity?

77. Hispanic or Latino (1)
78. Not Hispanic or Latino (2)
79. Prefer not to disclose (3)

Q14 What is your race? *Please select all that apply.*

80. American Indian (1)
81. Alaska Native (2)
82. Asian (3)
83. Black or African-American (4)
84. Native Hawaiian or Other Pacific Islander (5)
85. White (6)
86. Other: (7) _____
87. Prefer not to disclose (8)

Q15 Do you consider yourself to be a person living with a disability?

88. Yes (1)
89. No (2)
90. Prefer not to disclose (3)

End of Block: Demographics

NCGAS Post-Workshop Survey

Start of Block: Informed Consent

Q13 Thank you for attending to attend the NSF-funded National Center for Genome Analysis Support (NCGAS) Spring Workshop. We ask that you participate in this short survey for the purpose of gauging your workshop experience, as well as the efficacy of the content and delivery format of the workshop of attendees. Data will also be used in evaluating content and formats of future workshop. Your participation, as well as all survey questions, are optional and your responses are confidential. Data will be reported in the aggregate without any identifying information that you choose to, or inadvertently, disclose. Your decision to participate will not in any way affect your relationship with the NSF, the NCGAS project, the Pervasive Technology Institute, or Indiana University.

For questions about this survey, including problems with accessing the survey, please contact cesg@iu.edu and reference protocol #1804120218/exempt/exempt, approved on April x, 2018, by the Indiana University Institutional Review Board.

Do you agree to participate?

91.Yes (1)

92.No (2)

Skip To: End of Block If Thank you for attending to attend the NSF-funded National Center for Genome Analysis Support (NCG... = Yes
Skip To: End of Survey If Thank you for attending to attend the NSF-funded National Center for Genome Analysis Support (NCG... = No

End of Block: Informed Consent

Start of Block: Block 1

Q12 Given your participation in the workshop, what is your skill level for each of the following:

	No previous experience or knowledge (1)	Knowledge of its function, but no hands-on experience (2)	Ability to run very limited examples, such as small data sets and tutorials (3)	Ability to run more realistic examples, such as real data (4)	Ability to troubleshoot tasks for myself and others (5)
Navigating in Unix (1)	93.	94.	95.	96.	97.
Transferring files in Unix (2)	98.	99.	100.	101.	102.
Quality Control of Data (3)	103.	104.	105.	106.	107.
Managing Data and Archives (4)	108.	109.	110.	111.	112.
Submitting Jobs to a Queue (5)	113.	114.	115.	116.	117.
Using Trinity (6)	118.	119.	120.	121.	122.
Assembling a Basic Transcriptome (7)	123.	124.	125.	126.	127.
Combining Assemblies (8)	128.	129.	130.	131.	132.
Downstream Analyses of Transcriptomes (9)	133.	134.	135.	136.	137.
Using Galaxy (10)	138.	139.	140.	141.	142.

Q3 How comfortable are you now assembling a transcriptome?

143. Extremely comfortable (1)
144. Moderately comfortable (2)
145. Slightly comfortable (3)
146. Neither comfortable nor uncomfortable (4)
147. Slightly uncomfortable (5)
148. Moderately uncomfortable (6)
149. Extremely uncomfortable (7)

Q2 What was your favorite part of the NCGAS Spring Workshop?

_____

_____

_____

_____

_____

Q4 What is the most surprising or interesting thing you learned?

_____

_____

_____

_____

_____

Q5 Which aspect(s) of the workshop did you find most useful?

_____

Q6 If you were to add one thing to the workshop, what would it be?

_____

_____

_____

_____

_____

Q7 What aspect(s) of the workshop did you find least useful?

_____

_____

_____

_____

_____

Q8 If you feel you would benefit from a more extensive workshop, what you would add?

_____

_____

_____

_____

_____

Q9 Considering the content of the workshop, which delivery format do you believe would be ***most*** effective?

150. In-person workshop (1)
151. Live webinar (2)
152. Recorded webinar (3)
153. Other: (4) _____

Q10 What other workshops would you like to see us offer in the future?

_____

_____

_____

_____

_____

End of Block: Block 1