To: Dr. Titus Brown
From: STEM E.D., LLC
Re: NGS Summer Workshop Evaluation
Date: August 30, 2012. FINAL.

A pre-workshop evaluation of the NGS Summer Workshop 2012 - Analyzing Next-Generation Sequencing Data was conducted on June 4, with a post-workshop evaluation occurring June 15, 2012. Observations were also conducted at the start, middle, and end of the workshop. In all, 25 participants completed the pre-survey and 23 completed both the pre- and post-surveys.

EXECUTIVE SUMMARY OF RESULTS

We summarize evaluation results below. Following this summary, we provide: 1) A description of scales used to measure participant characteristics, including indices of validity, reliability, and overall scores, and 2) results for qualitative and close-ended questions related to participant impressions of the workshop. In summary, we found that:

1. Scores on the Perception of Computational Ability scale were calculated for both the pre- and post-workshop surveys. Results from the Wilcoxon Signed Ranks Test indicate that pre- and post-workshop results are statistically different ($Z = -4.116$, $p \leq 0.001$), with higher post-workshop scores. *This indicates that participants perceived greater computational ability after engagement in the workshop.*
2. Scores on the Computational Understanding – Sequencing Data scale were calculated for both the pre- and post-workshop surveys. Results from the Wilcoxon Signed Ranks Test indicate that pre- and post-workshop results are statistically different ($Z = -4.111$, $p \leq 0.001$), with higher post-workshop scores. *This indicates that participants perceived greater understanding after engagement in the workshop.*
3. Scores on the Python Coding Ability scale were calculated for both the pre- and post-workshop surveys. Results from the Wilcoxon Signed Ranks Test indicate that pre- and post-workshop results are statistically different ($Z = -4.374$, $p \leq 0.001$), with higher post-workshop scores. *This indicates that participants perceived greater coding ability after engagement in the workshop.*
4. *Participants were generally very satisfied with the workshop. On average, participants rated the workshop components as Good-Very Good.*
5. *Participants generally felt the workshop met their needs and would overwhelmingly recommend it to others.*
6. Participants were generally positive about the workshop in their open-ended comments. Suggestions for improvement include: more time on RNA sequencing and differential expression/data, less focus on why tools are not good, more focus on basics of programming, scripts and/or UNIX early on, and more details about daily activities.

WORKSHOP EVALUATION SCALES AND RESULTS

Scales used for the NGS Summer Workshop were modified from scales used in an earlier evaluation (Software Carpentry); scale statistics were calculated to ensure validity and reliability of scales. Participants completed a identical pre- and post-workshop surveys containing three scales: Perception of Computational Ability – Sequencing Data, Computational Understanding, and Python Coding Ability. Respondents also completed several demographic questions and responded to two open-ended questions covering workshop content. The surveys contained

questions about expectations (pre) and overall perceptions (post) of the workshop. Scale analyses were performed on post-instruction results except where noted.

Demographics

Demographic data are reported for pre-workshop respondents. One participant declined to respond to demographic questions. Three other pre-workshop survey respondents did not provide age data, although 25 respondents completed remaining demographic questions. Participants ranged in age from 23-57, with an average age of 35.6 ± 8.6 years. Participants were 65% male (n=17), with the remainder female (n=9) and no transgendered participants. Participants were also 81% Caucasian (n=21), with one African/African American/Black participant, two Asian participants, and one mixed race Caucasian-Asian participant. One Caucasian participant also identified as Latino. Finally, the academic status of participants included, graduate students (n=9), M.S.-holding professionals (n=1), and Ph.D.-holding professionals (n=14). One respondent chose "Other" for professional status but did not provide an explanation.

Perception of Computational Ability – Sequencing Data

A new scale measuring perception of computational ability and containing eight Likert-type items corresponded to the various abilities taught in the workshop; these items are all unique from the computational ability scale used in prior evaluations. Participants were asked to rate their ability in each as No Ability, Low Ability, Intermediate Ability, or High Ability. A score of 1 implies No Ability, while a score of 4 implies High Ability.

Table 1. Factor loadings, communalities, and item response averages for Perception of Computational Ability Scale - Sequencing Data.

Items	Factor Loadings	Communalities
Python scripting	.558	.312
Bash shell scripting	.548	.301
Cloud computing	.673	.453
Amazon EC2	.750	.563
Installing and running bwa	.866	.750
Installing and running velvet	.780	.608
Querying mappings	.792	.628
Evaluating assemblies	.715	.512

Factor Analysis. Exploratory factor analysis of responses to Perceptions of Computational Ability – Sequencing Data items was undertaken to investigate the presence of one ability scale. Two scales emerged based on eigenvalue analysis, while one strong scale was suggested by scree plot analysis. Based on this, confirmatory analysis for one scale was conducted.

Note that this analysis is preliminary given the small sample size, although the Kaiser-Meyer-Olkin measure of sampling adequacy for the eight-item scale was 0.729, above the 0.6 value recommended for factor analysis. The data set also met other minimum conditions necessary for factor analysis. First, the majority of items correlated with at least one other item at

a level over 0.3, indicating that a factor structure could be expected to emerge. The Bartlett's Test of sphericity was significant (χ^2 (28) = 84.67, p<0.001). Communalities for all items were above 0.3, indicating shared variance with other items. Given these data, factor analysis was performed on all eight items. A single scale explains 51.6% of the variance in the data and all items yielded factor loadings well over 0.32, suggesting the presence of a single scale (Table 1).

Reliability analysis. Cronbach's alpha was calculated to establish internal consistency of items. The scale has high reliability, with a calculated alpha for the sample of 0.861. Despite the small sample, this is above the recommended minimum of 0.70 for analysis of individual scores.

Results. Scores on the Perception of Computational Ability – Sequencing Data scale were calculated for the pre- and post-workshop survey. Overall scale scores were calculated as averages across all eight items. Data were inappropriate for parametric tests, so related samples nonparametric tests were run on matched data (n=23). Results from the Wilcoxon Signed Ranks Test indicate that pre- and post-workshop results are statistically different ($Z = -4.116$, $p \leq 0.001$). Statistical significance and mean results (Table 2) indicate that perceptions of computational ability increased as a result of instruction.

Table 2. Pre and Post Workshop Scale Averages

Scale	Pre	Post
Perception of Computational Ability – Sequencing Data	1.35 ± 0.34	2.84 ± 0.40
Computational Understanding	2.23 ± 0.69	3.26 ± 0.44
Python Coding Ability	1.98 ± 0.56	3.06 ± 0.86

Computational Understanding

One item in the 11-item Computational Understanding scale exhibited no variance (4 for all cases) post-workshop; this prohibits the performance of a factor analysis. As a result, the scale was evaluated via pre-workshop data. These items were similar to items used in prior evaluations, although concepts were aligned with those taught in this workshop. Participants were asked to indicate the extent to which they understood specific concepts, with ratings of Strongly Disagree, Disagree, Agree, and Strongly Agree. A score of 1 implies low understanding (Strong Disagreement), while a score of 4 implies high understanding (Strong Agreement).

Factor Analysis. Exploratory factor analysis of responses to pre-workshop Conceptual Understanding items was undertaken to evaluate the presence of one or more scales. Three scales emerged based on eigenvalue analysis, while one strong scale was suggested by scree plot analysis. Based on this, confirmatory analysis for one scale was conducted. Note that this analysis is preliminary given the small sample size, although the Kaiser-Meyer-Olkin measure of sampling adequacy of 0.69, above the 0.6 value recommended for factor analysis. The data set met other minimum conditions necessary for factor analysis. First, the majority of items correlated with at least one other item at a level over 0.3, indicating that a factor structure could be expected to emerge. The Bartlett's Test of sphericity was significant (χ^2 (55) = 146.4, p<0.001). Communalities for many items were above 0.3, indicating shared variance with other items. Given these data, confirmatory factor analysis was performed on all eleven items.

A single scale explains 43.7% of the variance in the data and all items yielded factor loadings over 0.32, suggesting the presence of a single scale (Table 3).

Table 3. Factor loadings and communalities for Computational Understanding Scale

Items	Factor Loadings	Communalities
I understand what "cd" means.	.521	.271
I understand what "bwa" means.	.790	.625
I know what "oases" does.	.433	.187
I understand what "#! /bin/bash" means.	.640	.410
I know how to write a simple script in Python.	.469	.220
I know how to run R.	.672	.452
I know what Tophat does.	.746	.556
I am familiar with the DEGexp function.	.434	.189
I know what "grep" does.	.859	.737
I know what Samtools is.	.731	.534
I know how Cufflinks and Tophat are related.	.792	.628

Reliability analysis. Cronbach's alpha was calculated to establish internal consistency of items. The scale has high reliability, with a calculated alpha for the sample of 0.866. Despite the small sample, this is above the recommended minimum of 0.70 for analysis of individual scores.

Results. Scores on the Computational Understanding scale were calculated for both the pre- and post-workshop surveys. Overall scale scores were calculated as averages across all 11 items. Data were inappropriate for parametric tests, so related samples nonparametric tests were run on matched data (n=23). Results from the Wilcoxon Signed Ranks Test indicate that pre- and post-workshop results are statistically different ($Z = -4.111$, $p \leq 0.001$). Statistical significance and mean results (Table 2) indicate that computational understanding increased as a result of instruction.

Python Coding Ability

A scale measuring Python coding ability and containing 22 Likert-type items was modified from the scale used in prior evaluations, itself modified from Askar and Davenport (2009), to focus on Python and to reflect concepts taught in the workshop. Participants were asked to indicate the extent to which they agreed with statements about their Python coding ability, with ratings of Strongly Disagree, Disagree, Agree, and Strongly Agree. A score of 1 implies low ability (Strong Disagreement), while a score of 4 implies high ability (Strong Agreement).

Factor Analysis. Scree plot analysis suggests the presence of a single scale, prompting use of confirmatory factor analysis of responses to post-instruction Python Coding Ability items to confirm one ability scale. Two poorly loading items, related to using the built-in help functions and motivation were removed; the poor loading for the evaluation item was observed in an

earlier evaluation). Confirmatory analysis was performed on the remaining 20 items.

Note that this analysis is preliminary given the small sample size, although the Kaiser-Meyer-Olkin measure of sampling adequacy of 0.665, above the 0.6 value recommended for factor analysis. The data set met other minimum conditions necessary for factor analysis. First, the majority of items correlated with at least one other item at a level over 0.3, indicating that a factor structure could be expected to emerge. The Bartlett's Test of sphericity was significant (χ^2 (190) = 402.2, p<0.001). Communalities for most items were above 0.5, indicating shared variance with other items.

A single scale explains 51.7% of the variance in the data and all items yielded factor loadings over 0.32, suggesting the presence of a single scale (Table 4).

Table 4. Factor loadings and communalities for Python Coding Ability Scale

Items	Factor Loadings	Communalities
I can write syntactically correct Python statements.	.817	.668
I understand the language structure of Python.	.693	.480
I can write logically correct blocks of code using Python	.804	.646
I can write a Python program that displays a greeting message.	.757	.573
I can write a Python program that computes the average of three numbers.	.700	.490
I can write a Python program that computes the average of any set of numbers.	.752	.565
I can write a small Python program to solve a problem that is familiar to me.	.865	.748
I can make use of a pre-written function if given a clearly labeled declaration of the function.	.774	.598
I cannot complete a programming project unless someone else helps me get started.	-.752	.565
I can debug a long and complex program that I have written.	.755	.571
I can comprehend a long, complex multi-file program.	.785	.616
I cannot complete a programming project unless someone shows me how to solve the problem first.	-.772	.596
I cannot complete a programming project unless I have the language reference manual.	-.716	.513
I can complete a programming project if I have a lot of time to complete the program.	.530	.281
I can find ways of overcoming the problem if I get stuck at a point on a programming project.	.711	.506
I can come up with a suitable strategy for a given programming project in a short time.	.611	.374

I cannot complete a programming project unless I can call someone for help if I get stuck.	-.796	.633
I can rewrite confusing portions of code to be more readable.	.534	.285
I can find a way to concentrate on my program, even when there are many distractions around me.	.342	.117
I can write a program that someone else can comprehend.	.718	.516

Reliability analysis. Cronbach's alpha was calculated to establish internal consistency of items. The scale has a very high reliability, with a calculated alpha for the sample of 0.942. Despite the small sample, this is above the recommended minimum of 0.70 for analysis of individual scores. Note that this value was calculated after reverse coding of four negatively loaded items, as is appropriate for the alpha metric.

Results. Scores on the Python Coding Ability scale were calculated for both the pre- and post-workshop surveys. Overall scale scores were calculated as averages across all 20 items. Data were inappropriate for parametric tests, so related samples nonparametric tests were run on matched data (n=23). Results from the Wilcoxon Signed Ranks Test indicate that pre- and post-workshop results are statistically different ($Z = -4.374$, $p \leq 0.001$). Statistical significance and mean results (Table 2) indicate that Python coding ability increased as a result of instruction.

Open-Ended Questions

Two open-ended questions related to the concepts covered in the workshop were completed pre- and post-workshop (Table 5). These questions were analyzed for evidence of increased understanding of workshop concepts, specifically as reflected in the complexity of participant responses. Response rates were much higher post-workshop for individual questions, suggesting greater ability to respond to questions; as one participant put it pre-workshop, "*[I] cannot do this yet.*" In general, complexity of responses increased from pre- to post-workshop. Further analysis of open-ended responses for accuracy is possible with collaboration between evaluators and workshop facilitators.

Table 5. Open-Ended Response Rates

Open-Ended	Pre Response Rate*	Post Response Rate*
Suppose that you are using Illumina to sequence DNA from a mouse sample that should have genetic differences from the mouse reference genome. Discuss one or more approaches you would take to analyze the data, as well as your expected sensitivity and specificity to SNPs and indels. Include in your discussion how much you will miss, and how much you find that will be wrong.	68%	91%

Suppose that you are doing a transcriptome expression analysis of a non-model system for which you do not have a good reference genome. Discuss a sequencing strategy, challenges you expect to face, the expected sensitivity and specificity of your analysis, and what kinds of additional large-scale and/or computational data sets you could use to help explore your transcriptome data set.	76%	87%

*Response rates are for participants completing surveys only: 25 for pre- and 23 for post-workshop data.

Observations

The workshop was observed four times for one to two hour periods: once on 6/5/12, twice on 6/8/12, and once on 6/15/12. Observations indicate that facilitators struck a good balance between lecture and student engagement in actual programming. During observations, little connection was made between computational science and biological science, although this may reflect the limited number of observations.

Overall Workshop Impressions

Participants were asked to respond to close-ended questions related to overall impressions of the workshop (Table 6). Participants also responded to open-ended questions about their expectations for the workshop (pre-survey) and their perceptions of the workshop (post-survey).

Close-Ended Questions. Eleven questions asked participants to rate components of the workshop, as well as the overall workshop, on a 5-point Likert scale of Very Poor-Poor-Adequate-Good-Very Good. A 1 corresponds to a Very Poor rating and a 5 corresponds to a Very Good Rating. Participants were also asked if the workshop met their needs on a 4-point scale (4=very well), if they learned as expected from the workshop, if understanding of computational science changed, and if they would recommend the workshop.

Table 6. Overall Workshop Impressions

Workshop Components	**Average Score (ideal score)**	**% Yes**
Day 1: EC2, UNIX, and BLAST	4.43 (5)	NA
Day 2: Mapping	4.38 (5)	NA
Day 3: Assembly	4.48 (5)	NA
Day 4: ChIP-Seq	4.33 (5)	NA
Day 5: Statistics and Plotting	3.76 (5)	NA
Day 6, part a: More ChIP-Seq	3.95 (5)	NA
Day 6, part b: Genome Assembly and Annotation	4.15 (5)	NA
Day 6, part c: Current and Future Sequencing Technology	3.71 (5)	NA
Day 7: mRNAseq and Alternative Splicing	4.14 (5)	NA

Day 8: STACKS and Reduced Representation Sequencing	3.95 (5)	NA
Overall Workshop Rating	4.48 (5)	NA
Meet Needs?	3.3 (4)	NA
Learn What Hoped to Learn?	NA	87%
Computational Understanding Change?	NA	96%
Recommend Workshop?	NA	96%

Results. Post-workshop responses to questions about the efficacy of the workshop indicate that participants were generally very satisfied with the workshop (Table 6). On average, participants rated the workshop components as Good-Very Good. The one exception (Current and Future Sequencing Technology) was rated Adequate-Good. This was one of three concepts covered on Day 6 and workshop leaders may want to reduce content covered over that single day.

Participants generally felt the workshop met their needs and would overwhelmingly recommend it to others. Participants gave the workshop a rating of 4.48 out of 5, indicating the workshop met their needs well to very well. Eighty-seven percent (n=20) of participants felt their computational understanding changed, and 96% (n=22) felt they learned what they hoped to learn and would recommend the workshop to others.

Qualitative Data: Participant Expectations and Perceptions. On the pre-survey, participants responded to a prompt: "*Please provide any additional comments about your expectations for the workshop below*." Participants also responded to a similar post-survey question: "*Please provide any additional suggestions or comments about the workshop below.*", and were given opportunities to comment to each of the three yes-no questions in Table 6.

Results. Five participants provided comments about expectations on the pre-survey (Table 7). These comments related to interest in analyzing sequencing data, use of multiple data sets, and increasing understanding of tools. Given their brevity, pre-workshop expectations are provided:

- *A big improvement of understanding for computational biology and design of NGS experiments*
- *I hope to learn how to map data, determine quality of the data & compare the expression levels of multiple data sets*
- *Understanding use of tools and gain ability to work independently on my data, getting started and learning how programs and techniques work is the starting point*
- *I can learn basic skill on how to analyze the next generation sequencing data and get related knowledge*
- *I hope not to be so completely unprepared to analyze my sequence data*

Ten overall post-survey responses expressed either satisfaction with the workshop or suggestions for improvement. In general, comments were positive, for examples:

- *[I learned] much more than I ever knew I would*
- *I feel much more confident in my computational abilities*
- *The instructors for this course were so enthusiastic and friendly.*

Suggestions for the workshop included adding more time on RNA sequencing (n=1) and differential expression/data (n=2), less focus on why tools are not good (n=1), more focus on basics of programming, scripts and/or UNIX early on (n=3). One participant suggested providing more details about daily activities, specifically: *Provide a schedule of what material we will be covering each day in advance. Provide activities to work on in class to reinforce learning of concepts. Provide more detailed account of what we expect to accomplish with a certain task (i.e. what we are starting with, what the program does, and what we will end up with) - make it VERY obvious.*

Table 7. Perception and Expectations Response Rates

Prompt/Question	**N responses**
Please provide any additional comments about your expectations for the workshop. (pre)	5
Please provide any additional suggestions or comments about the workshop. (post)	10

REFERENCES

Askar, P. & Davenport, D. (2009). An investigation of factors related to self-efficacy for Java programming among Engineering Students. Turkish Online Journal of Educational Technology, 8(1), 26-32.