Optimizing NGS-based Data Analysis Training

Mary Piper, Harvard Chan Bioinformatics Core

Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Abstract

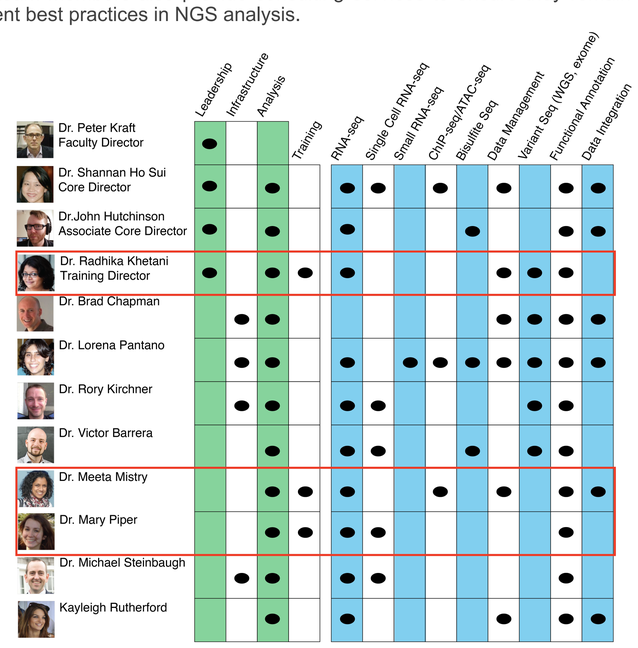
Objectives: The Harvard Chan Bioinformatics Core provides consulting and training services to the Harvard community. Three years ago, we initiated a new training program designed to empower researchers not only to understand next generation sequencing, but also to perform analysis of their own data. Our goal has been to develop local experts in the community, within labs and departments, who can help address the bottleneck in high throughput data analysis.

Methods & Conduct: Our training curriculum is guided by class survey results and observed trends from ongoing consulting requests to meet the needs of our community. The evolution of the program has progressed from teaching NGS overview workshops using Galaxy to focusing on the command line through introductory workshops for Linux, R, and differential expression, as well as, an in-depth NGS analysis course spanning 6-weeks and covering multiple NGS methods. Over the last three years we have modified not only the content, but also the methods for organizing the courses. We based these changes on trial and error, feedback from exit survey results, classroom observations and collaboration with external groups.

Impact: These changes, big and small, have resulted in higher levels of attendance, lower rates of attrition, increased workshop satisfaction, and easier troubleshooting in the classroom. I will present on the methods we have used to target appropriate audiences, the techniques for improved material design for aiding retention and facilitating hands-on tutorials, and our strategies for better evaluation and generation of feedback.

Introduction

The training team is part of the Harvard Chan Bioinformatics Core and includes three analysts with time devoted to material development, community outreach, and training. Training team members also provide consulting services to ensure they remain up-to-date with current best practices in NGS analysis.

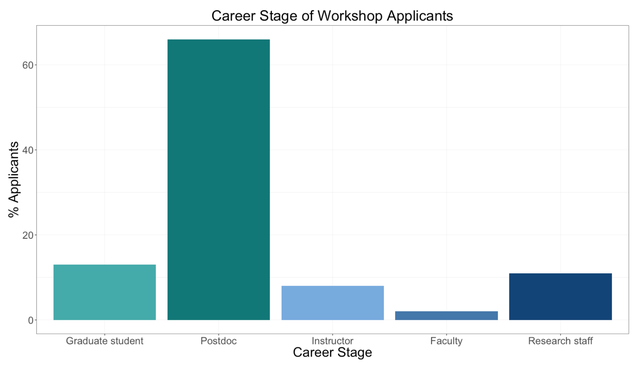


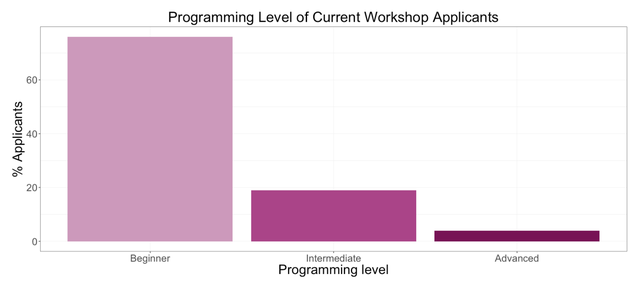
MISSION: To expand knowledge of NGS methods in the community and to develop local expertise to help address the bottleneck in NGS analysis.

The training program offers **short 1-3 day workshops** and a **10-12 day in-depth course**. The short workshops focus on a topic related to NGS analysis (e.g. Unix, R, RNA-Seq, ChIP-Seq, etc.), while the in-depth course includes:

- Unix & High-Performance Computing
- NGS data analysis (RNA-Seq, ChIP-Seq, Variant calling, scRNA-Seq)
- Statistical analysis using R
- Functional analysis
- Git/GitHub
- Markdown/R Markdown

The program bas been optimized for biomedical researchers with little to no exposure to programming; participants are predominantly early-career researchers. However, we have recently expanded to include advanced-level topics as part of the ***"Current Topics in Bioinformatics"*** workshop series. This series includes monthly, half-day, skill-building sessions with the goal of providing continuing education for participants who have attended our beginner workshops.





** Programming level and career stage statistics were generated from survey results from 2016-2017 workshops*

Training Program Design

We designed our training program to achieve the following:

- high attendance / low attrition
- inform and provide skills to perform independent NGS analyses
- detailed feedback and assessment
- accessible, reviewable workshop materials for better retention

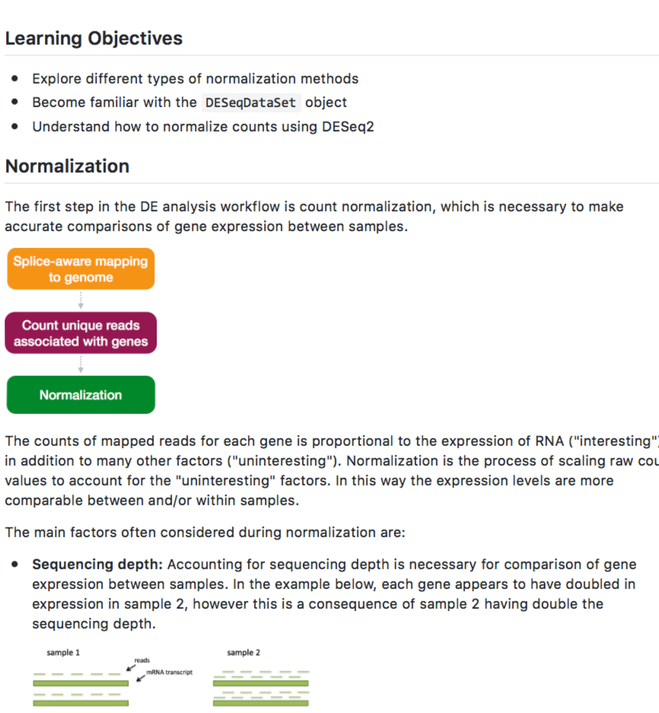
Administration:

- **Advertisement:** clear learning objectives and detailed descriptions align participant expectations with workshop goals.
- **Registration fees:** Fees of at least **$25** achieve high attendance and low attrition.
- **Class size:** Optimized class size based on workshop for appropriate student:teacher ratios (1:6 - 1:12).
- **Application process:** Short workshops are first-come-first-served, but our in-depth course has an application process to select motivated students in need of training who plan to share their knowledge with the community.

Materials:

- **Format:** We changed our materials from a lecture-based format to **Markdown-based format* to allow easier self-learning**. Theory-based material is integrated into these hands-on markdowns to aid in future review.

  ** Change to Markdown format was inspired by Software Carpentry*



Learning Objectives

- Explore different types of normalization methods
- Become familiar with the `DESeqDataSet` object
- Understand how to normalize counts using DESeq2

Normalization

The first step in the DE analysis workflow is count normalization, which is necessary to make accurate comparisons of gene expression between samples.

Splice-aware mapping to genome

Count unique reads associated with genes

Normalization

The counts of mapped reads for each gene is proportional to the expression of RNA ("interesting") in addition to many other factors ("uninteresting"). Normalization is the process of scaling raw count values to account for the "uninteresting" factors. In this way the expression levels are more comparable between and/or within samples.

The main factors often considered during normalization are:

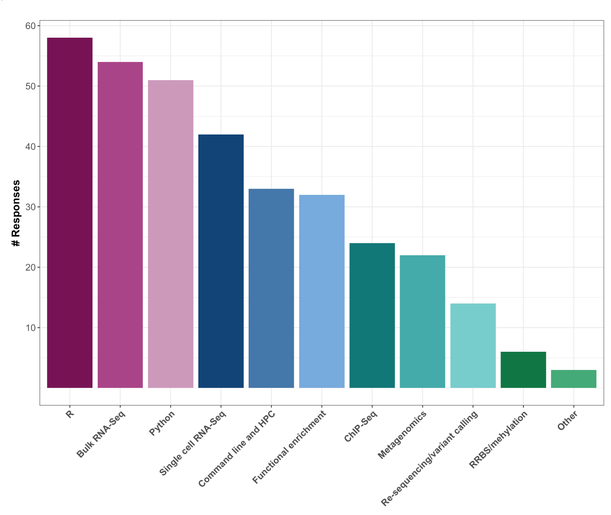
- **Sequencing depth:** Accounting for sequencing depth is necessary for comparison of gene expression between samples. In the example below, each gene appears to have doubled in expression in sample 2, however this is a consequence of sample 2 having double the sequencing depth.

sample 1 — reads — mRNA transcript — sample 2

- **Access:** All markdown lessons are available for access on GitHub (**https://hbctraining.github.io/main**).

Scope:

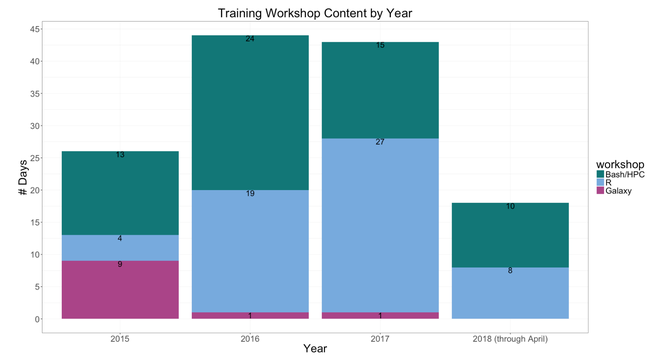
- **Topics:** We update our course offerings to reflect the needs of our community, which are determined by in-class surveys and consulting requests. Currently, the majority of interest and consults (and therefore training sessions) are for RNA-Seq.



- **External experts:** External experts expand the depth of our teaching offerings and offer a change in pace in long workshops.

Execution:

- **Platform:** The short workshop materials transitioned over time from Galaxy-based to Unix/R-based materials.



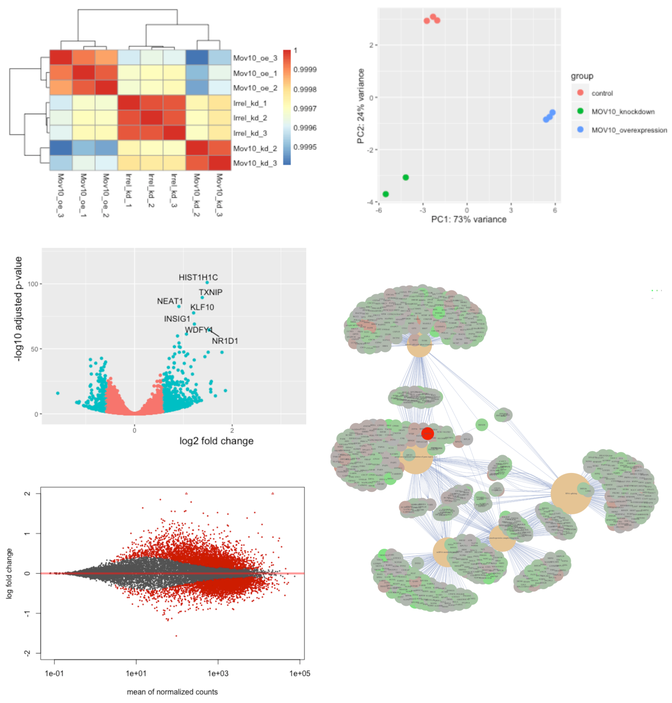
- **Modularity:** The Unix/R short workshops can be used to get an overview of an NGS workflow or can be used in a modular way to acquire skills to perform independent analyses, similar to the content in the in-depth course.

  For example, the following series of short workshops can be taken in serial to obtain skills to independently perform RNA-Seq analyses:

  - Introduction to Unix with RNA-Seq (2 days)
  - Introduction to R (2 days)
  - Differential expression analysis with DESeq2 (1 day)

Participants learn the basics of Unix and R, and, when combined with the differential expression analysis workshop, they will have the knowledge to independently perform RNA-Seq analysis and downstream functional analysis.

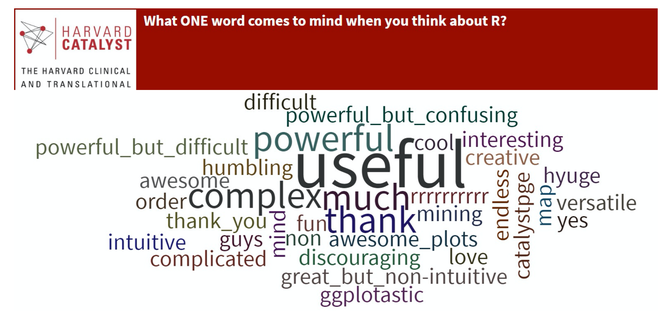
RNA-Seq Analysis Figures



Assessments:

- **Surveys:** Detailed feedback on satisfaction, areas of improvement, additional topics of interest, etc.
- **In-class polls:** Determine status of class with regard to progress, pace, etc.
- **Stickies*:** Actively monitor progress of participants using green and red stickies

  ** Use of stickies was inspired by Software Carpentry*



Future Plans

Future workshops will expand our reach to non-beginner programmers. We plan to continue offering more advanced monthly 1/2 day modules. In addition, we plan to hold an in-depth course for advanced programmers.

Based on consulting requests and survey responses, single-cell RNA-Seq is in high demand for user training. We plan to offer single-cell RNA-Seq workshops starting with the in-depth course in Fall 2018.

The HBC is located in Building 2 of the Harvard T.H. Chan School of Public Health and can be reached at:

Training: hbctraining@hsph.harvard.edu
Website: bioinformatics.sph.harvard.edu
Twitter: @bioinfocore
Training materials: **https://hbctraining.github.io/main**