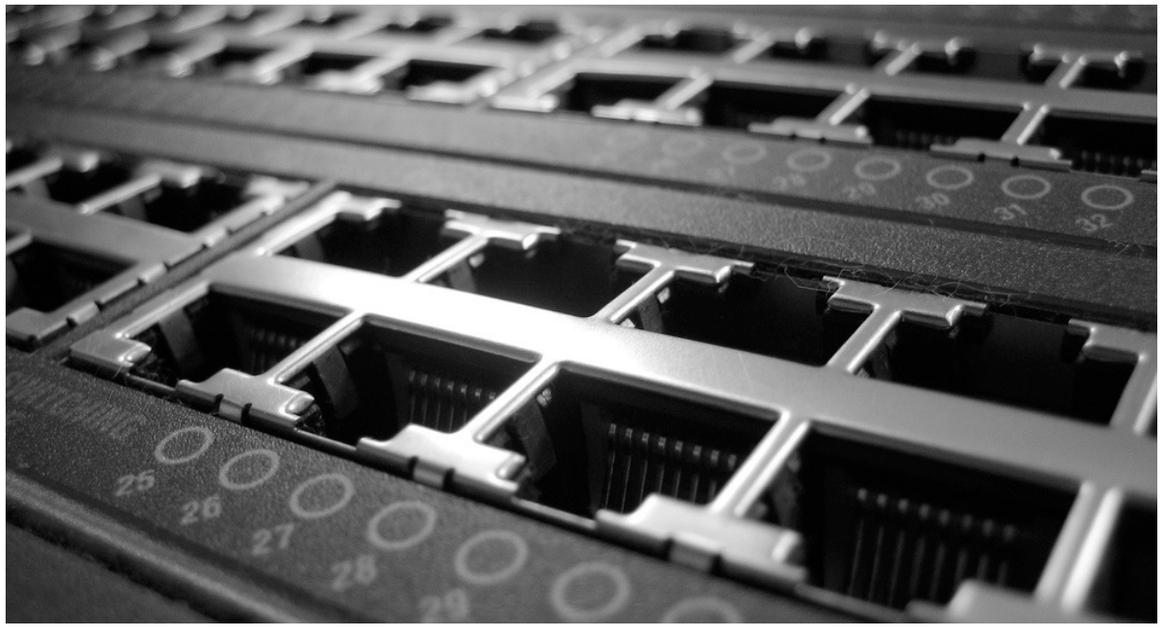


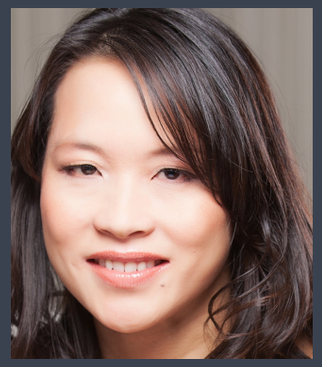
Adventures in Bioinformatics Training

Mary Piper, (Meeta Mistry) & Radhika Khetani

Harvard Chan Bioinformatics Core

Webpage: http://bioinformatics.sph.harvard.edu/training/

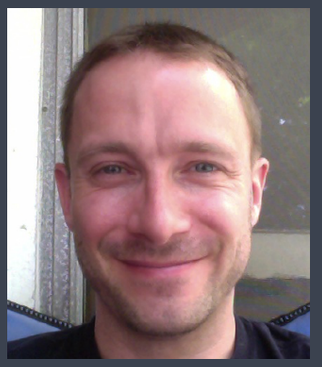
Email: hbctraining@hsph.harvard.edu



Shannan Ho Sui

John Hutchinson

Brad Chapman



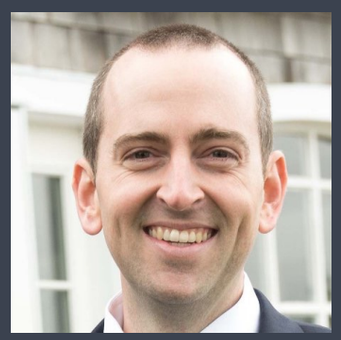
Rory Kirchner

Meeta Mistry

Radhika Khetani

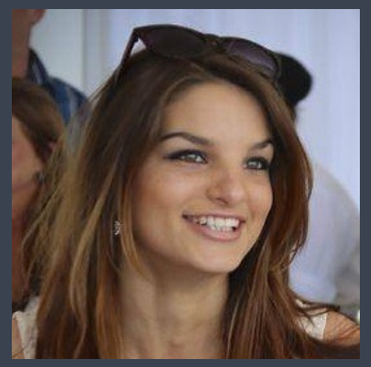
Mary Piper

Lorena Pantano



Michael Steinbaugh

Victor Barrera



Kayleigh Rutherford

Peter Kraft

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

HARVARD STEM CELL INSTITUTE

HARVARD CATALYST
THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

HARVARD MEDICAL SCHOOL

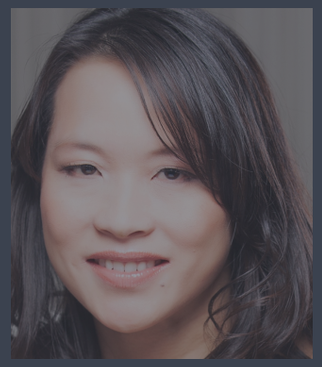
Consulting

- **Transcriptomics:** RNA-seq, small RNA-seq, scRNA-Seq
- **Epigenetics:** ChIP-seq, genome-wide methylation, ATAC-Seq
- **DNA Variation:** WGS, resequencing, exome-seq and CNV studies
- Functional enrichment analysis
- Experimental design
- Grant support

Training

- Introduction to **command line** (Unix) and **high-performance computing**
- **Introductory R** and **differential gene expression** analysis
- **In-depth course**: Unix & R, RNA-Seq, ChIP-Seq, and variant calling
- Monthly, short workshops on various bioinformatics topics (free)

http://bioinformatics.sph.harvard.edu/



Shannan Ho Sui

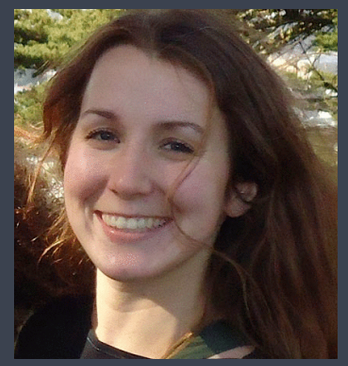
John Hutchinson

Brad Chapman

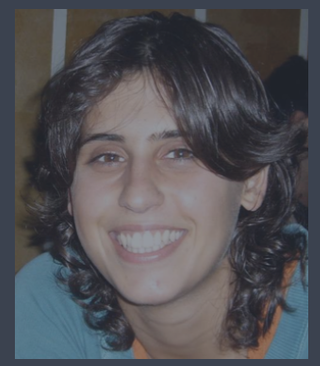
Rory Kirchner

Meeta Mistry

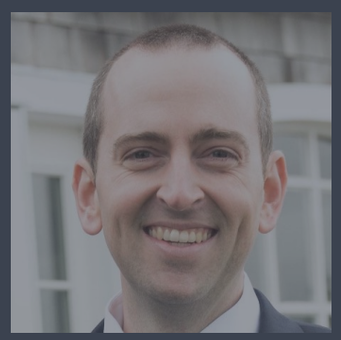
Radhika Khetani



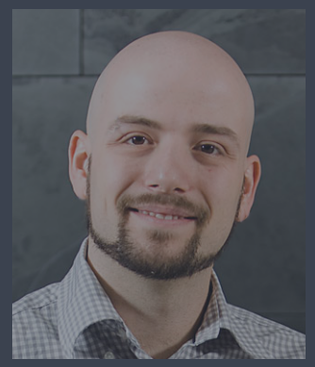
Mary Piper



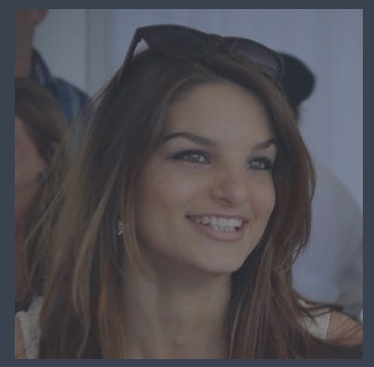
Lorena Pantano



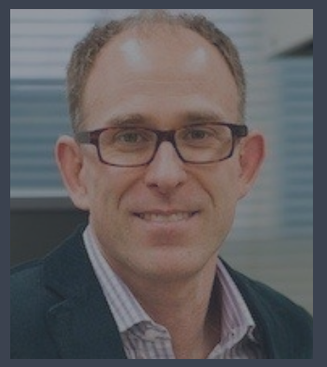
Michael Steinbaugh



Victor Barrera



Kayleigh Rutherford



Peter Kraft

The “core” training team @ the Core

About HBC's training program

- Short workshops and longer bootcamp-style courses
- 50+ workshops/courses of varying duration since December 2014
- Funded by FTEs from multiple groups at Harvard
- Training restricted to groups providing funding
- Sequence analysis and basic computational skills
  - *Transitioned away from Galaxy to command-line based training in 2016*
- Small registration fee for attendance
  - *Transitioned away from free in 2016*

About HBC's training program

1. Workshops

These workshops vary in duration from half-day to 3 days and are aimed at biologists interested in expanding their knowledge of Next-Generation Sequence (NGS) data analysis, and basic computational skills for working with big or small datasets.

- **Next-Generation Sequencing**

  Short workshops on RNA-Seq, ChIP-Seq to introduce basic concepts of Next-Generation Sequencing (NGS) analysis. The goal of these workshops are to enable researchers to design their studies appropriately and perform preliminary data analyses using best practices.

- **Basic Skills for Data Analysis**

  Topics include R, data visualization using R, Linux shell, High-Performance Computing (HPC), version control, Research Data Management, etc.

http://bioinformatics.sph.harvard.edu/training/

About HBC's training program

2. In-Depth Next Generation Sequencing Analysis Courses

These intensive courses run for 8-12 days and are aimed at bench biologists interested in learning **how to perform independent, best practice NGS-based analyses.** Topics include:

- Unix & High-Performance Computing (O2)
- NGS data analysis (RNA-Seq, ChIP-Seq, Variant calling)
- Statistical analysis using R
- Functional analysis

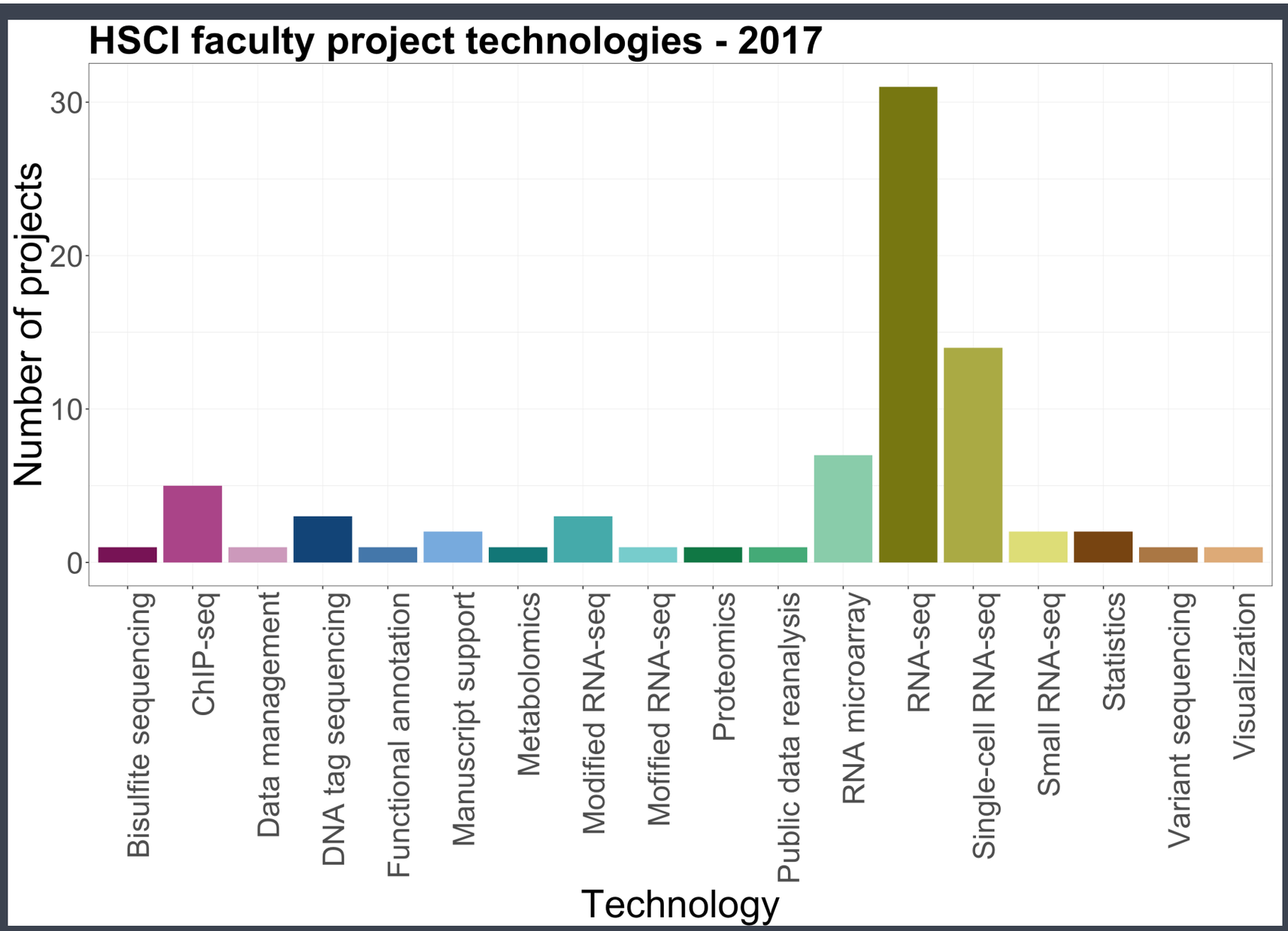
No prior NGS or command line expertise is required for our workshops or courses unless explicitly stated.

http://bioinformatics.sph.harvard.edu/training/

How do we decide on topics?

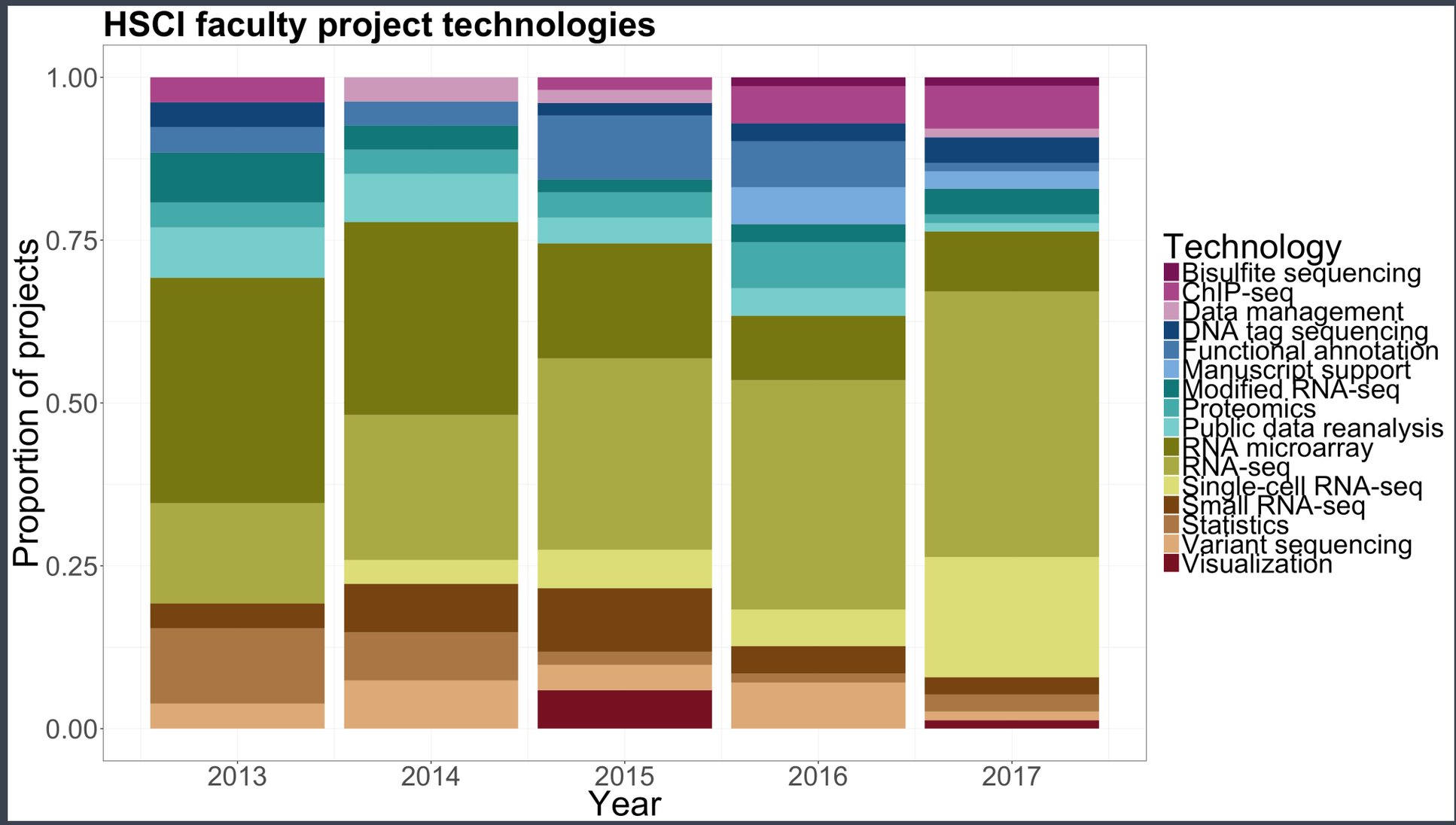
- Consulting services
- Exit surveys at workshops

How do we decide on topics?

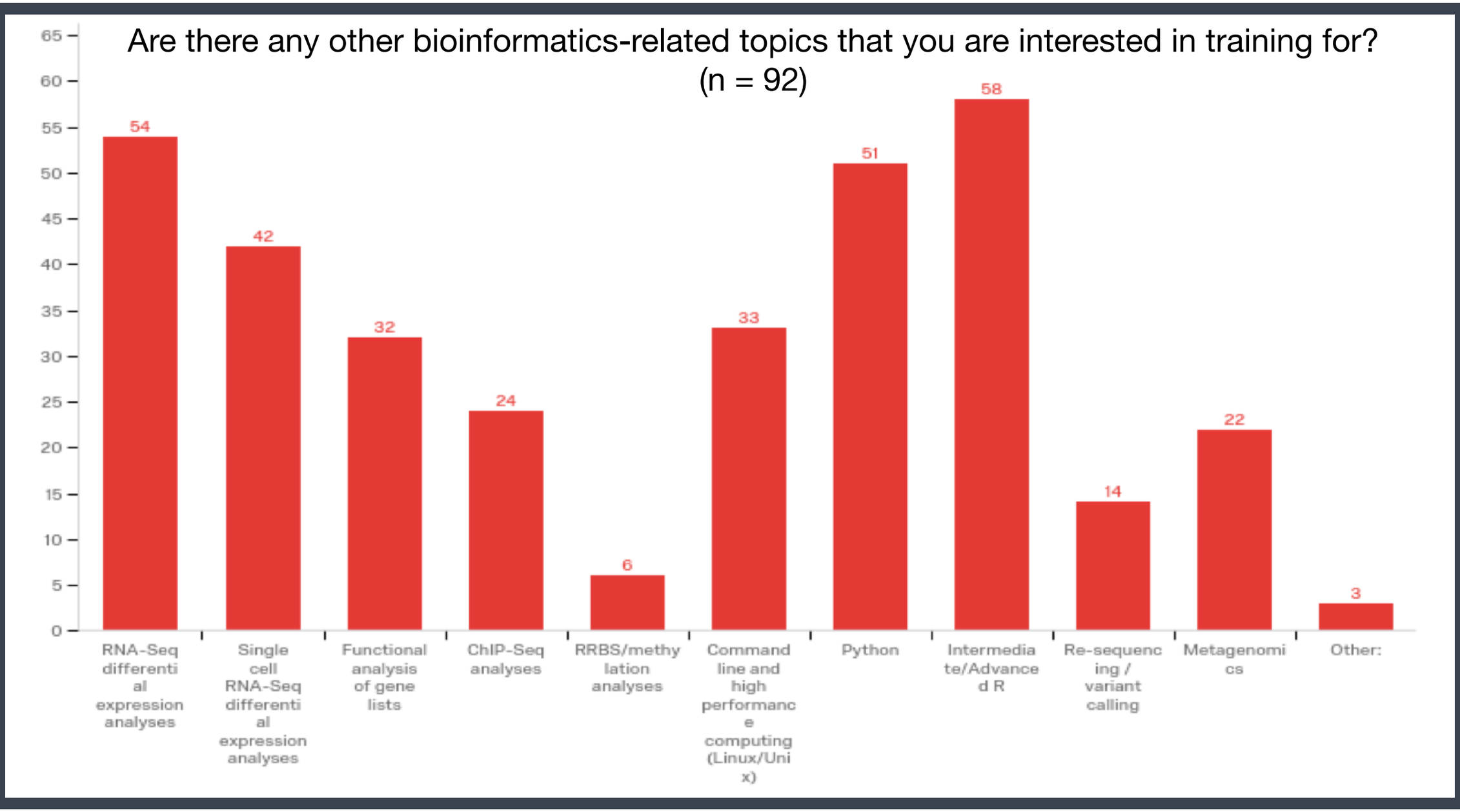


Created by John Hutchinson

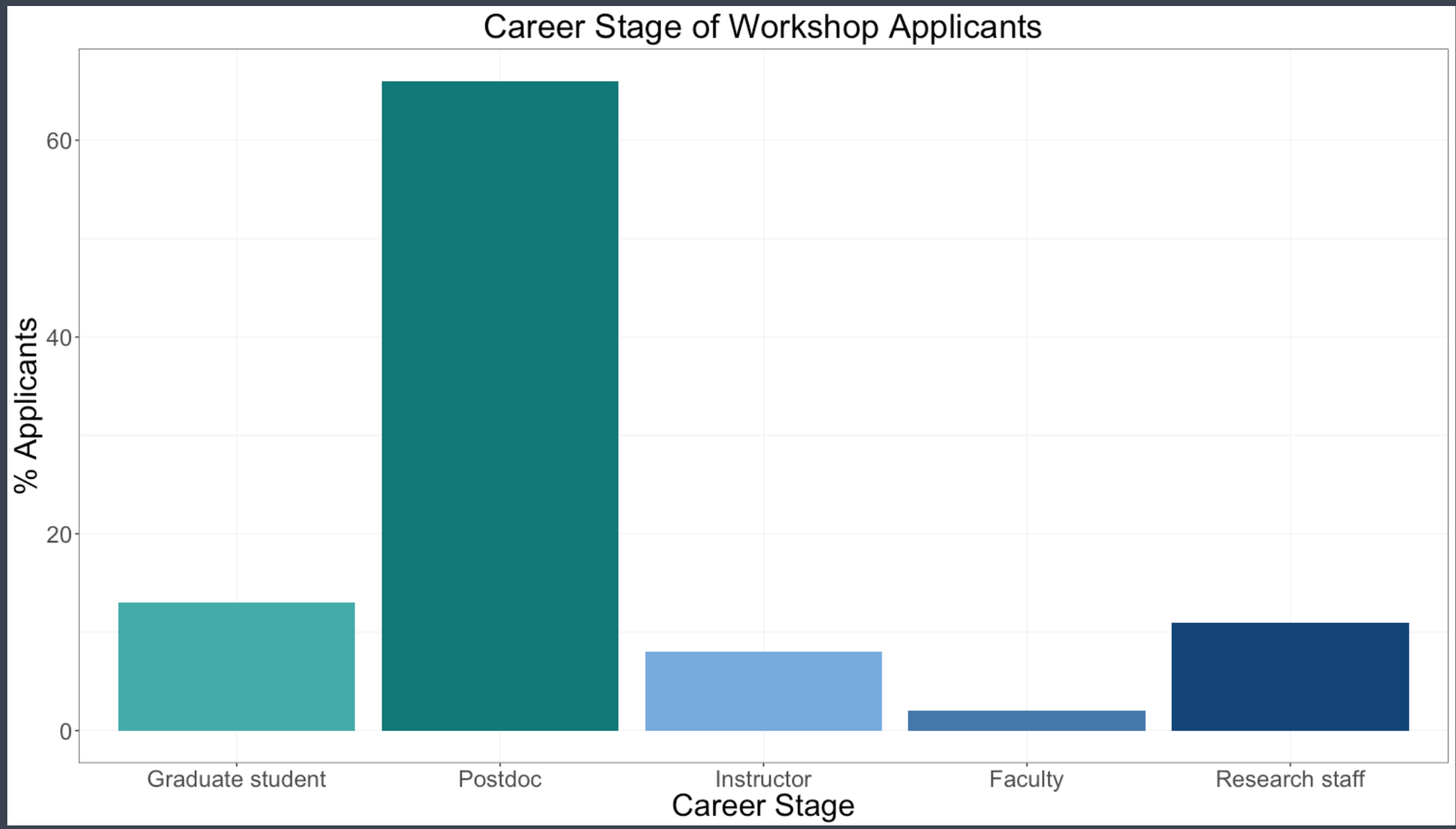
How do we decide on topics?



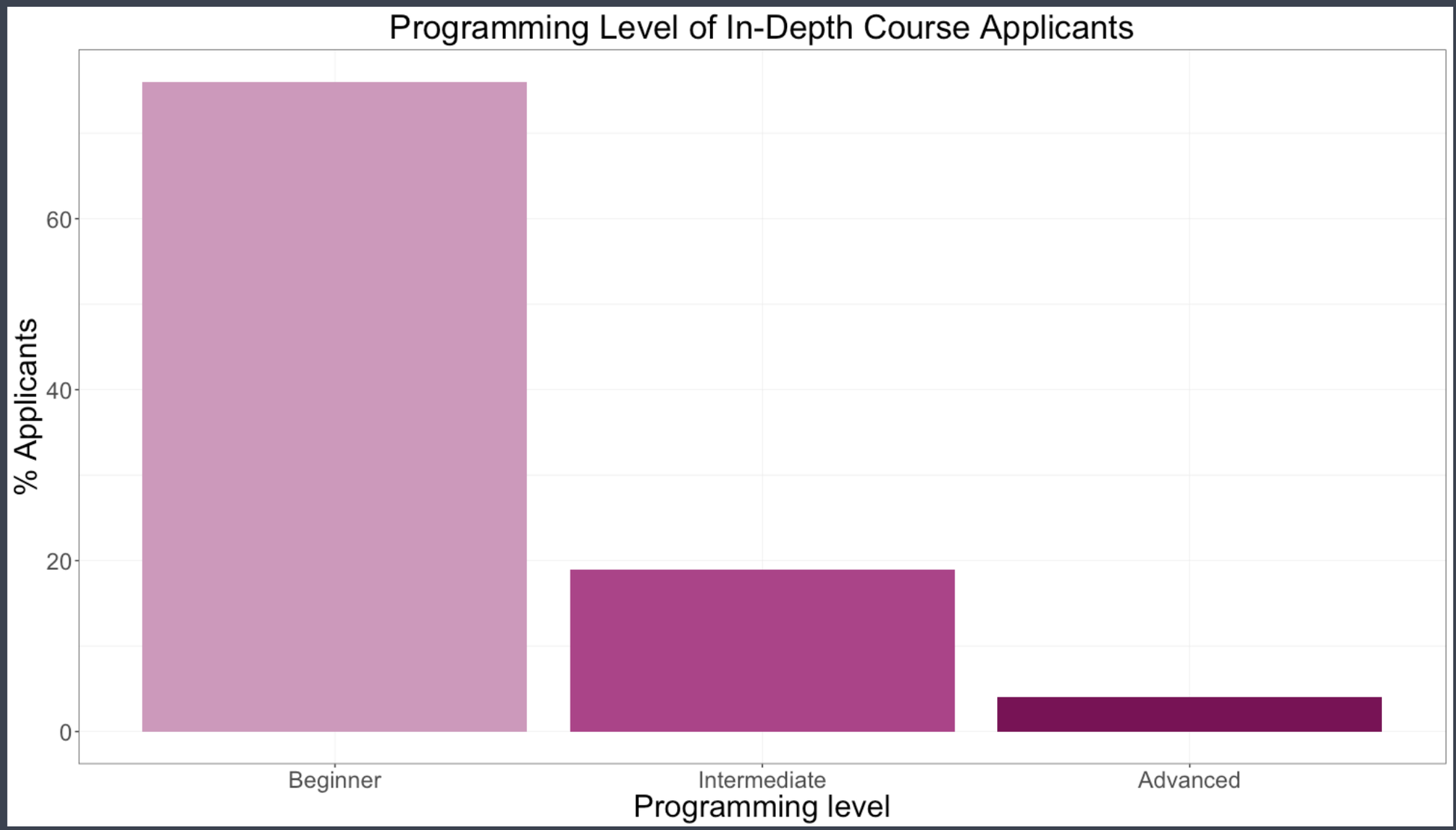
How do we decide on topics?



Who is our target audience?



Who is our target audience?



What are our objectives for this training program?

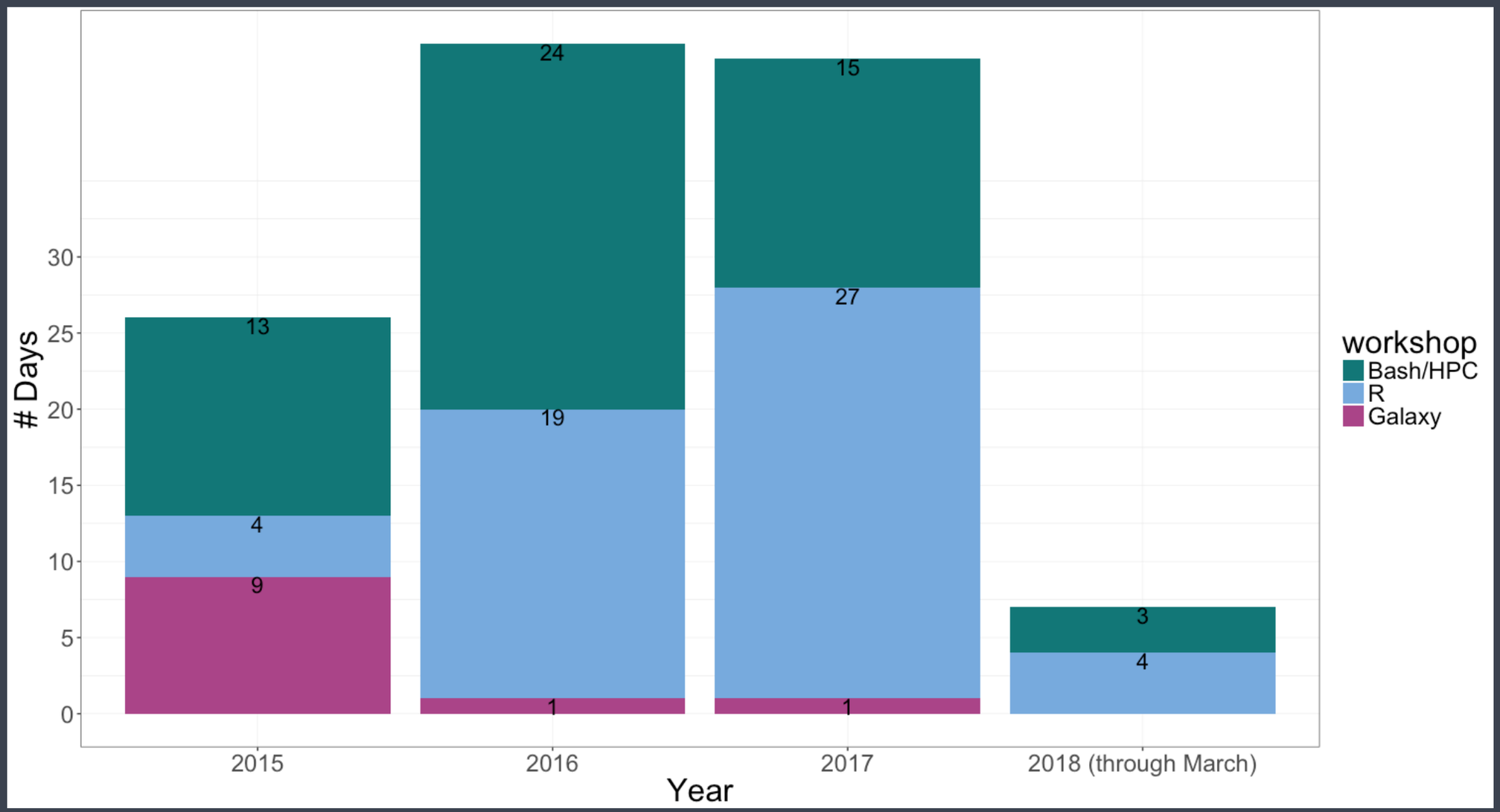
Enable students to:

- utilize best practices for experimental design and data analysis
- analyze their own data more independently
- disseminate information to their labs/groups
- remain engaged during class
- revise concepts well after the workshop/course

What methods are we using to meet the program's objectives? (1/3)

- Small class sizes (average 25 people, max 35 people) with 3 instructors
- Incoming students at similar level of computational skills
- Teach on local research computing resources that students will use for their own analysis
- Move away from a point-and-click interface like Galaxy

Training Platforms

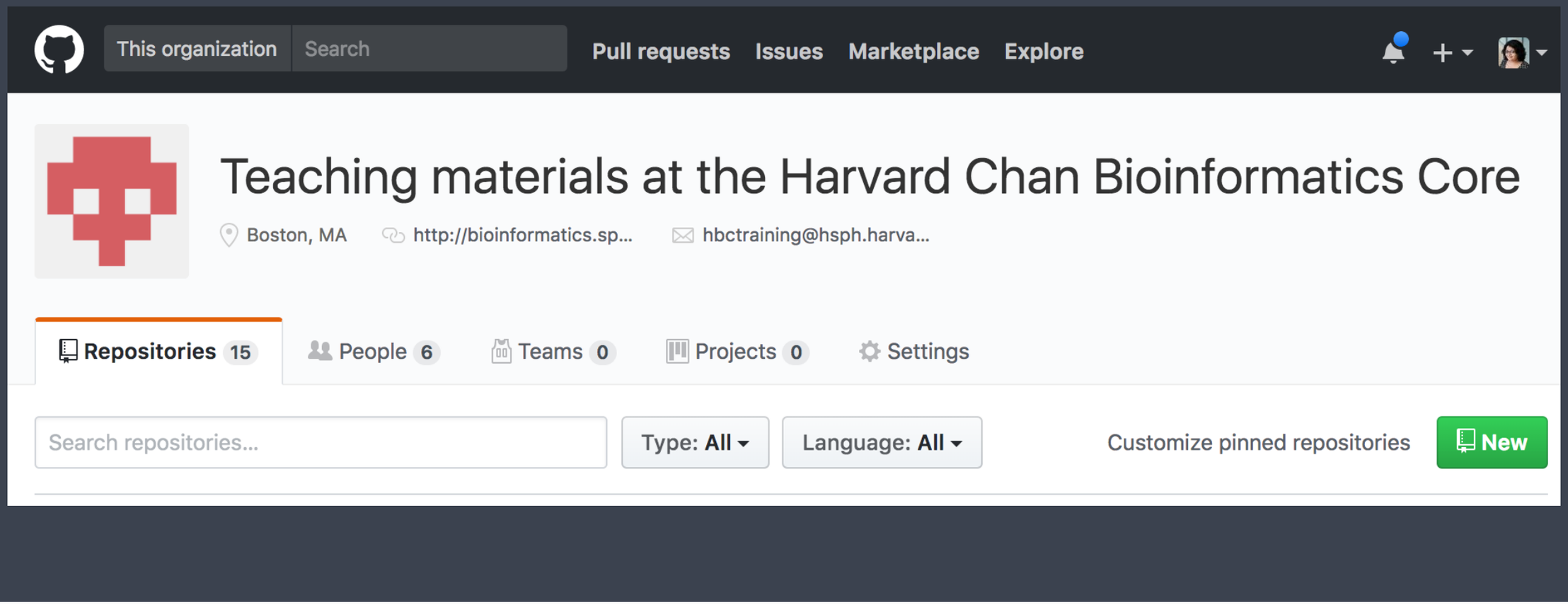


What methods are we using to meet the program's objectives? (2/3)

- In-class assessments with red/green post-its and PollEverywhere
- Registration fee to ensure motivated students and lower attrition
- 8-12 day course:
  - Application process to ensure motivation & PI buy in
  - Special topics covered by invited speakers

What methods are we using to meet the program's objectives? (3/3)

- Close coordination among team members to ensure lessons flow
- Easily accessible and open training/learning materials (mostly github-based)



Materials on Github

Workshops

Introduction to Next-Generation Sequencing (NGS) analysis series:

The goal of these workshops (2-3 days) are to enable researchers to design their NGS studies appropriately and perform preliminary data analyses.

Training topic and link to lessons	Prerequisites	Workshop Duration
R and ggplot2	None	2 days
RNA-seq data analysis using High-Performance Computing	None	2 - 3 days
R and Differential Gene Expression (DGE) analysis	None	3 days
Differential Gene Expression (DGE) analysis	R and ggplot2	1.5 days
ChIP-seq using High-Performance Computing	None	*In development*
Identifying variants in genome/exome sequencing data	None	*In development*

https://hbctraining.github.io/main/

Materials on Github

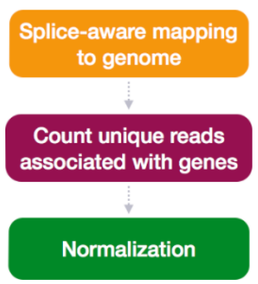
Approximate time: 60 minutes

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.



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.

STAR Aligner

To determine where on the human genome our reads originated from, we will align our reads to the reference genome using STAR (Spliced Transcripts Alignment to a Reference). STAR is an aligner designed to specifically address many of the challenges of RNA-seq data mapping using a strategy to account for spliced alignments.

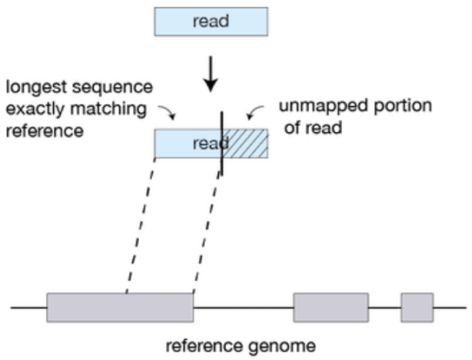
STAR Alignment Strategy

STAR is shown to have high accuracy and outperforms other aligners by more than a factor of 50 in mapping speed, but it is memory intensive. The algorithm achieves this highly efficient mapping by performing a two-step process:

1. Seed searching
2. Clustering, stitching, and scoring

Seed searching

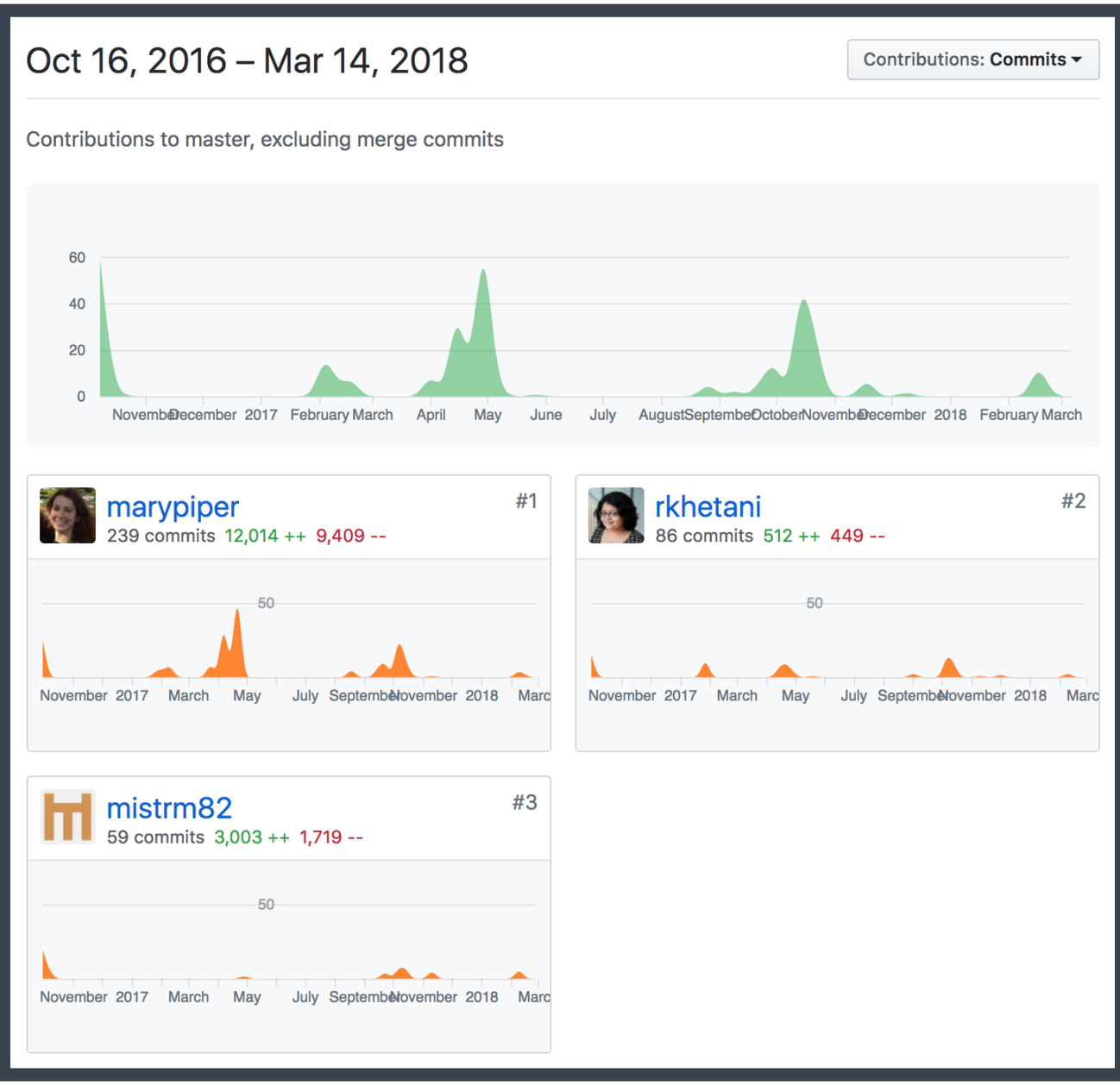
For every read that STAR aligns, STAR will search for the longest sequence that exactly matches one or more locations on the reference genome. These longest matching sequences are called the Maximal Mappable Prefixes (MMPs):



The different parts of the read that are mapped separately are called 'seeds'. So the first MMP that is mapped to the genome is called *seed1*.

https://hbctraining.github.io/main/

Materials on Github



https://hbctraining.github.io/main/

Collaborators and co-trainers

- HMS
  - HMS-RC (Research Computing)
  - RITS (Research Information Technology Solutions)
  - Countway library
  - Sarah Boswell at the Single Cell Core (Dept. of Systems Biology)
  - DBMI (Dept. of Biomedical Informatics)
- Harvard University
  - FAS-RC
  - Bob Freeman (HBS, FAS-RC, software carpentry)
  - Ista Zahn + Kareem Carr (IQSS)
- Software Carpentry + Data Carpentry
- H3ABioNet (Bioinformatics Network for The Human Heredity and Health in Africa Initiative)
- GOBLET (Global Organisation for Bioinformatics Learning, Education & Training)

Discussion Topics

- Training platforms
- Class sizes
- Varying skills among students
- Assessments methods (before class, in-class, after class)
- Long-term surveys
- Registration fee
- Open training materials
- …?
- …?
- …?