

# Technical questions: Where to ask and how to get answers

Bari J. Ballew, PhD  
September 16, 2019

- Have you ever had a question regarding bioinformatics/data science?
- Where did you go for answers?
- Did you get a reply?
  
- This talk will cover:
  1. Where/whom to ask
  2. How to ask


# Where/whom can I ask?

- **Intramural resources:** many groups at NIH that provide support, including (*not an exhaustive list!*):
  - CBIIT
  - NIH Library
  - NIH HPC Group
  - Additional resources listed [here](#)
  - NIH bioinformatics slack channel: <https://nih-byob.slack.com>
- **Non-interactive resources**
- **Interactive resources**

# Non-interactive resources

- Manual/readme/documentation

bedtools v2.29.0 » [next](#) [index](#)



Bedtools is a fast, flexible toolset for genome arithmetic.

**Bedtools links**

- Issue Tracker
- Source @ GitHub
- Old Releases @ Google Code
- Mailing list @ Google Groups
- Queries @ Biostar
- Quinlan lab @ UU

## bedtools: a powerful toolset for genome arithmetic

Collectively, the **bedtools** utilities are a swiss-army knife of tools for a wide-range of genomics analysis tasks. The most widely-used tools enable *genome arithmetic*: that is, set theory on the genome. For example, **bedtools** allows one to *intersect*, *merge*, *count*, *complement*, and *shuffle* genomic intervals from multiple files in widely-used genomic file formats such as BAM, BED, GFF/GTF, VCF. While each individual tool is designed to do a relatively simple task (e.g., *intersect* two interval files), quite sophisticated analyses can be conducted by combining multiple bedtools operations on the UNIX command line.

**bedtools** is developed in the [Quinlan laboratory](#) at the [University of Utah](#) and benefits from fantastic contributions made by scientists worldwide.

## Tutorial

We have developed a fairly comprehensive [tutorial](#) that demonstrates both the basics, as well as some more advanced examples of how bedtools can help you in your research. Please have a look.

## Important notes

- As of version 2.28.0, bedtools now supports the CRAM format via the use of [htslib](#). Specify the reference genome as-

# Non-interactive resources

- Manual/readme/documentation
- FAQs/vignettes/tutorials

## Introduction to dplyr

When working with data you must:

- Figure out what you want to do.
- Describe those tasks in the form of a computer program.
- Execute the program.

The dplyr package makes these steps fast and easy:

- By constraining your options, it helps you think about your data manipulation challenges.
- It provides simple “verbs”, functions that correspond to the most common data manipulation tasks, to help you translate your thoughts into code.
- It uses efficient backends, so you spend less time waiting for the computer.

This document introduces you to dplyr's basic set of tools, and shows you how to apply them to data frames. dplyr also supports databases via the dbplyr package, once you've installed, read `vignette("dbplyr")` to learn more.

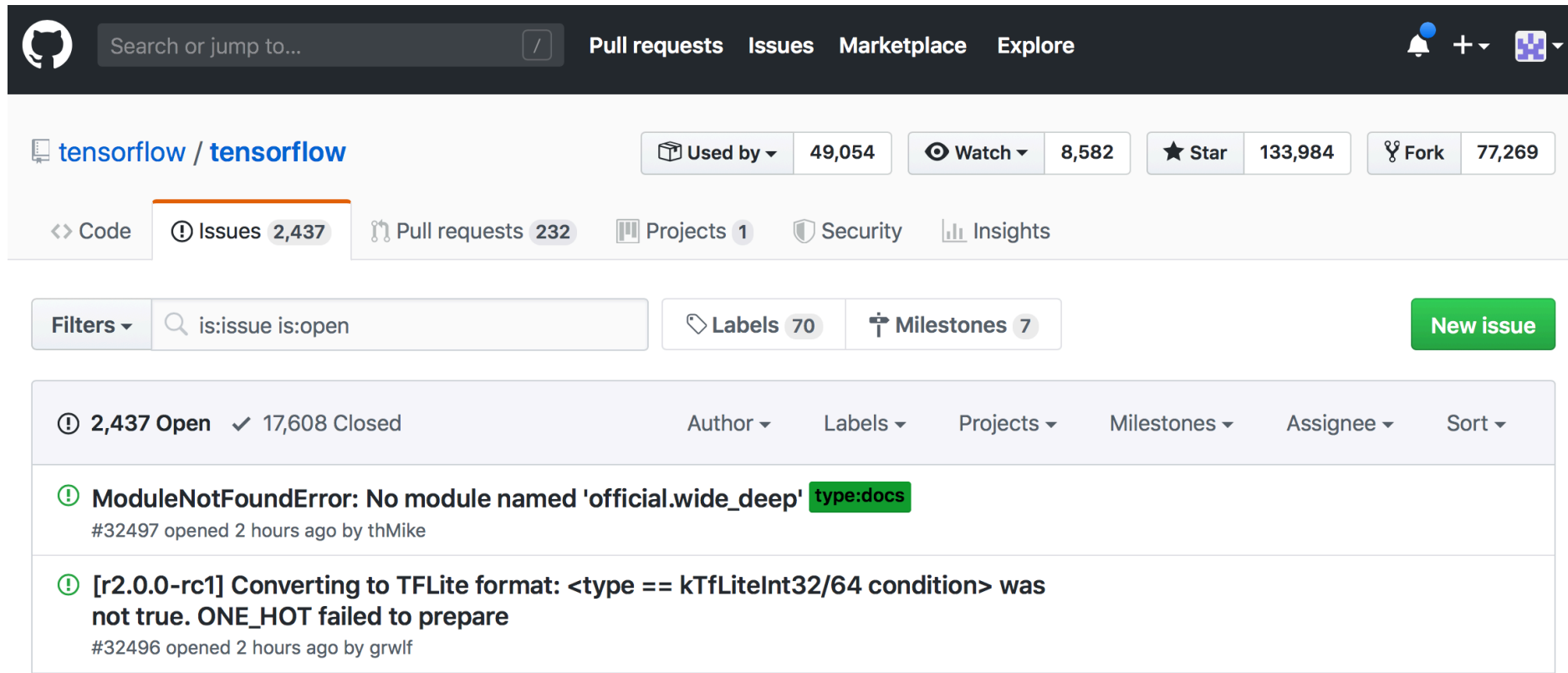
## Data: nycflights13

To explore the basic data manipulation verbs of dplyr, we'll use `nycflights13::flights`. This dataset contains all 336776 flights that departed from New York City in 2013. The data comes from the US [Bureau of Transportation Statistics](#), and is documented in `?nycflights13`

```
library(nycflights13)
dim(flights)
#> [1] 336776    19
flights
#> # A tibble: 336,776 x 19
```

# Interactive resources

- GitHub/Bitbucket issues pages



The screenshot shows the GitHub interface for the tensorflow/tensorflow repository. The top navigation bar includes a search bar, navigation links for Pull requests, Issues, Marketplace, and Explore, and user notification icons. Below the repository name, statistics for Used by (49,054), Watch (8,582), Star (133,984), and Fork (77,269) are displayed. The main navigation bar highlights the Issues section with 2,437 issues, alongside links for Code, Pull requests (232), Projects (1), Security, and Insights. A filter bar shows 'is:issue is:open' and 'Labels 70' with a 'New issue' button. The issue list table has columns for status, author, labels, projects, milestones, assignee, and sort. Two issues are visible: #32497 (ModuleNotFoundError) and #32496 (Conversion to TFLite format error).

Issue ID	Author	Labels	Projects	Milestones	Assignee	Sort
#32497	thMike	type:docs				
#32496	grwlf					

# Interactive resources

- GitHub/Bitbucket issues pages
- Forums
  - <https://www.biostars.org>
  - <https://stackoverflow.com>

The screenshot shows the Biostars forum interface. At the top, there are navigation tabs for various categories: LATEST, OPEN, RNA-SEQ, CHIP-SEQ, SNP, ASSEMBLY TUTORIALS, TOOLS, JOBS, FORUM, PLANET, and ALL. The Biostars logo is prominently displayed, along with the tagline "BIOINFORMATICS EXPLAINED". A user profile for "bari.ballew" with 180 reputation is shown, along with a "Logout" button and links for "about", "faq", and "rss". Below the navigation, there are icons for "Community", "Messages", "Votes", "My Posts", "My Tags", "Following", "Bookmarks", and "New Post". A search bar is present with a "Live search" input and a "Classic search" button. The main content area displays a list of posts with their respective statistics (votes, answers, views) and tags. The first post is titled "Snakemake FastQC MissingOutputFiles ErrorM" with 3 votes, 1 answer, and 75 views. The second post is "Anyone knows how to plot gene model like this?" with 0 votes, 0 answers, and 16 views. The third post is "Biopython Bio.motifs: How to create a motif object with aligned sequences" with 0 votes, 0 answers, and 7 views. On the right side, there is a "Recent Votes" section listing various topics and a "Recent Locations" section with an "All" link.

LATEST OPEN RNA-SEQ CHIP-SEQ SNP ASSEMBLY TUTORIALS TOOLS JOBS FORUM PLANET ALL »

**Biostars** — BIOINFORMATICS EXPLAINED —

bari.ballew • 180 | Logout about • faq • rss

Community Messages Votes My Posts My Tags Following Bookmarks New Post

Live search: start typing... or Classic search

Limit to: all time <prev • 75,166 results • page 1 of 2506 • next > Sort by: update

**3** votes **1** answer **75** views **Snakemake FastQC MissingOutputFiles ErrorM**  
snakemake fastqc  
written 17 hours ago by lasejourny • 10 • updated just now by bari.ballew • 180

**0** votes **0** answers **16** views **Anyone knows how to plot gene model like this?**  
gene model visualization  
written 26 minutes ago by louiesxscape • 0

**0** votes **0** answers **7** views **Biopython Bio.motifs: How to create a motif object with aligned sequences**  
python alignment motif  
written 45 minutes ago by kinetic • 0

**Recent Votes**

- C: Sequence duplication levels in de-novo assemblies
- C: PacBio RSII reads
- A: Alignment and mapping
- A: Limma for metabolomics data
- A: Sequence duplication levels in de-novo assemblies
- A: Best strategy to get GO terms for a proteome?
- C: Dealing with two batches in RNA-seq

**Recent Locations** • All »

# Interactive resources

- GitHub/Bitbucket issues pages
- Forums
  - <https://www.biostars.org>
  - <https://stackoverflow.com>
- NIH listserv: <https://list.nih.gov>
  - BIOINFORMATICS-SIG-L
  - NIH-DATASCIENCE-L



**BIOINFORMATICS-SIG-L Home Page**

## BIOINFORMATICS-SIG-L@LIST.NIH.GOV

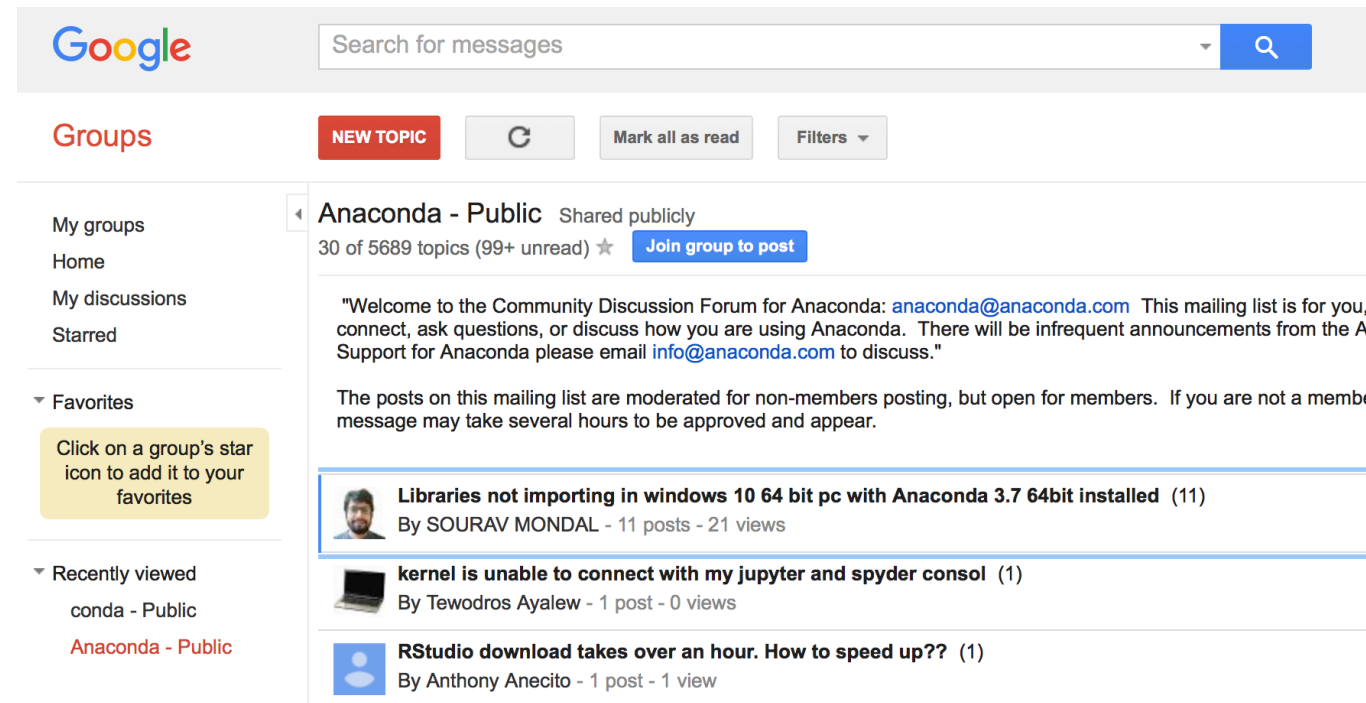
### Latest Messages

<a href="#">Walk-In Consult with HPC staff Wed 18 Sep</a>	NIH HPC Staff <staff@HPC.NIH.GOV>	Wed, 11 Sep 2019 08:10:02 -0400
<a href="#">NCI Clinical Bioinformatician - Apply by September 13</a>	Guiton, Jordan (NIH/NCI) [E] <jordan.guiton@NIH.GOV>	Tue, 10 Sep 2019 18:11:02 +0000
<a href="#">Register for September webinars on NCI-funded informatics tools Single-cell Genome Viewer, GAIL, and more!</a>	Wasel, Paul (NIH/NCI) [C] <paul.wasel@NIH.GOV>	Fri, 6 Sep 2019 18:02:48 +0000
<a href="#">NIH Figshare: A New NIH Data Publication Platform now available</a>	Sean Davis <seandavi@GMAIL.COM>	Fri, 6 Sep 2019 13:11:18 -0400
<a href="#">Re: Using LASER with hg38?</a>	McGaughey, David (NIH/NEI) [E] <david.mcgaughey@NIH.GOV>	Tue, 3 Sep 2019 13:39:57 +0000



# Interactive resources

- GitHub/Bitbucket issues pages
- Forums
  - <https://www.biostars.org>
  - <https://stackoverflow.com>
- NIH listserv: <https://list.nih.gov>
  - BIOINFORMATICS-SIG-L
  - NIH-DATASCIENCE-L
- User groups
  - Specific to tool



The screenshot shows the Google Groups interface for the 'Anaconda - Public' group. At the top, there is a search bar for messages and navigation buttons like 'NEW TOPIC', 'Mark all as read', and 'Filters'. The group name 'Anaconda - Public' is displayed as 'Shared publicly' with 30 of 5689 topics (99+ unread) and a 'Join group to post' button. A welcome message is visible, stating: "Welcome to the Community Discussion Forum for Anaconda: [anaconda@anaconda.com](mailto:anaconda@anaconda.com). This mailing list is for you to connect, ask questions, or discuss how you are using Anaconda. There will be infrequent announcements from the Anaconda Support for Anaconda please email [info@anaconda.com](mailto:info@anaconda.com) to discuss." Below this, a moderation notice states: "The posts on this mailing list are moderated for non-members posting, but open for members. If you are not a member, your message may take several hours to be approved and appear." A list of recent posts is shown, including: "Libraries not importing in windows 10 64 bit pc with Anaconda 3.7 64bit installed (11)" by SOURAV MONDAL (11 posts, 21 views), "kernel is unable to connect with my jupyter and spyder consol (1)" by Tewodros Ayalew (1 post, 0 views), and "RStudio download takes over an hour. How to speed up?? (1)" by Anthony Anecito (1 post, 1 view). On the left sidebar, there are navigation options for 'My groups', 'Home', 'My discussions', and 'Starred', along with a 'Favorites' section and a 'Recently viewed' list containing 'conda - Public' and 'Anaconda - Public'.

# How to ask

- Do your due diligence
- Ask a colleague
- Post on a forum

# Due diligence

- GOOGLE! Your problem is probably not unique
  - Search all non-interactive resources (manuals, FAQs, etc)
  - Search previous posts to interactive resources (forums, issues, etc)
- Use the information gleaned to do some troubleshooting

# Ask a colleague

Original

I tried to run bcftools merge, but got an error. What do I do?

Better

A large empty rectangular box with a dark green border, intended for a better version of the question.

# Ask a colleague: Describe the problem

- Your environment (OS, program version, dependencies)
- The exact command/series of commands you ran

## Original

I tried to run bcftools merge, but got an error. What do I do?

## Better

I tried to run bcftools merge (version 1.9) on biowulf on an interactive node to combine several bgzipped VCF files as follows:

```
bcftools merge -o out.vcf file1.vcf.gz  
file2.vc.gzf file3.vcf.gz
```

# Ask a colleague: Describe the problem

- Your environment (OS, program, version information)
- The exact command/series of commands you ran
- Your input
- Your output (expected and observed)

## Original

I tried to run bcftools merge, but got an error. What do I do?

## Better

I tried to run bcftools merge (version 1.9) on biowulf on an interactive node to combine several bgzipped VCF files as follows:

```
bcftools merge -o out.vcf file1.vcf.gz  
file2.vcf.gz file3.vcf.gz
```

I got an output file containing the headers and some variant rows, but many fewer variants than were in my input files.

# Ask a colleague: Describe the problem

- Your environment (OS, program, version information)
- The exact command/series of commands you ran
- Your input
- Your output (expected and observed)
- Any error messages

## Original

I tried to run bcftools merge, but got an error. What do I do?

## Better

I tried to run bcftools merge (version 1.9) on biowulf on an interactive node to combine several bgzipped VCF files as follows:

```
bcftools merge -o out.vcf file1.vcf.gz  
file2.vc.gzf file3.vcf.gz
```

I got an output file containing the headers and some variant rows, but many fewer variants than were in my input files. I got an error message that says “Error at chr1:123: wrong number of fields in ANN\_phred.”

# Ask a colleague: Describe the problem

- Your environment (OS, program, version information)
- The exact command/series of commands you ran
- Your input
- Your output (expected and observed)
- Any error messages

**What resources should I have looked at prior to asking this question?**

## Original

I tried to run bcftools merge, but got an error. What do I do?

## Better

I tried to run bcftools merge (version 1.9) on biowulf on an interactive node to combine several bgzipped VCF files as follows:

```
bcftools merge -o out.vcf file1.vcf.gz  
file2.vc.gzf file3.vcf.gz
```

I got an output file containing the headers and some variant rows, but many fewer variants than were in my input files. I got an error message that says “Error at chr1:123: wrong number of fields in ANN\_phred.”



# Ask a colleague: Provide context

- Briefly explaining *why* you're doing something can often yield more helpful answers
  - **Original answer:** The steps to upload the files are...
  - **Better answer:** If you are looking at differential expression, it's best not to look by eye, but to use an established pipeline like xyz.

## Original

How do you upload BAMs to the UCSC Genome Browser?

## Better

I am trying to identify differently expressed transcripts from my RNA-seq data. I would like to visualize the read abundance in each of my samples in the UCSC Genome Browser to find differences. How do I upload my BAMs to the browser?

# Ask a colleague: Show what you've tried

- Describing what you've tried so far to solve your problem makes others more likely to help you
- It also helps others better understand your issue and give more helpful replies

## Original

I'm using bedtools v2.28.0 to intersect a BED with a VCF but when I run it as follows it gives me an empty file.

```
bedtools intersect -a x.vcf -b y.bed > outFile
```

## Better

I'm using bedtools v2.28.0 to intersect a BED with a VCF but when I run it as follows it gives me an empty file.

```
bedtools intersect -a x.vcf -b y.bed > outFile
```

I've checked that both my input files are not empty, and used vcf-checker to make sure the VCF file is properly formed. I tried running the same command in v2.19. I also tried switching the order (`-a y.bed -b x.vcf`) but my output file is still empty.

# Ask a colleague: Be specific

- There are a lot of different ways to do things...and they don't necessarily have the same results.
- The person providing an answer is often going to try to follow your steps and replicate the problem, so they can fix it and tell you how. Help them out!

## Original

I installed python and pandas as required by the program.

## Better

I am using python 3.7 via biowulf's module system (i.e. `module load python3`). I installed pandas with anaconda (`conda install pandas`).

# Post to a forum

- All the earlier things, plus...
  - Title
  - Text formatting (if applicable)
  - Avoid cross-posting
  - Focus on one question per post

Original

Title: Help with bedtools

Better

Title: Bedtools intersect VCF with multiple databases:  
empty output file

# Follow-up

- Try to reply to specific questions that are asked of you (e.g. “Can you post the output of conda list?”)
- If you figure out the solution yourself, consider replying to your own post, for future answer-seekers

# Asking a question seems like a lot of work!

- Yes, it is! But, doing the background reading, the troubleshooting, and framing your question correctly will often help you better understand the situation, if not solve it on your own.
- Absent this work, your question may come off as expecting others to do the heavy lifting for you. In this case it is likely your question will not be answered.

Thanks!