Workshop Format

We’ll be going through these topics today:

- Installation
- Basic usage: common options/modes
- Interpreting HTML output

Feel free to jump in with questions or ask them the chat!

I recommend you open a browser to this page:
https://barricklab.org/breseq

You can click through the same results I will be viewing there.
This download offers several options for using the official university wordmark and the What Starts Here tagline. The elements can be used in several formats, allowing you to choose one that best fits with your design needs.

The wordmark and tagline should not be changed or manipulated in any way. When reproduced in color the wordmark should always appear in The University of Texas at Austin's signature color, burnt orange (PMS 159). No other color is acceptable, with the exception of all black for use on faxes, memos and in newspapers, and all white for reverses on dark backgrounds.
When is *breseq* the right tool?

- You have short-read NGS resequencing data.
- Your reference genome is *haploid*.
  - Bacteria, Archaea, Phages, Plasmids, Haploid yeast
- You expect few genetic differences from the reference (a few to <1,000) in each sample.
- It’s important that you identify all mutations.
- You are comfortable with using the terminal a little.
  - Changing directories, copying files, running a command


[https://barricklab.org/breseq](https://barricklab.org/breseq)  [https://github.com/barricklab/breseq](https://github.com/barricklab/breseq)
Antibiotic Resistance Reversal: \textit{breseq} Analysis of Experimental Evolution, Compared with FACS Competition Assays of Relative Fitness

\textbf{Joan Slonczewski}

\textit{Kenyon College}

Identifying Adaptive Paths in Host-Plasmid Coevolution Using \textit{breseq}

\textbf{Olivia Kosterlitz}

\textit{University of Washington}
Workshop Presentations

Decoding Evolution-In-Action in Classroom Experiments That Simulate Infection Biology Using \textit{breseq}

\textbf{Vaughn Cooper}

\textit{University of Pittsburgh}

\textbf{ALEdb: A Living High-Quality Database of Mutations from Adaptive Evolution Experiments Powered by \textit{breseq}}

\textbf{Adam Feist}

\textit{University of California, San Diego}
Introductory Topics

– Installation
  • Different methods
  • Common problems

– Basic usage
  • Most important options/modes
  • Different types of reference files

– Interpreting HTML output
  • Compare output generated using different options
  • Understanding and evaluating predictions
Installing *breseq*

Like many bioinformatics tools, *breseq* requires a Unix-like environment with a command line.

**Linux and MacOS**
- Open the terminal

**Windows**
- Install *WSL* (Windows Subsystem for Linux). Then you should be able to follow the Linux instructions!
- Also possible to use *Cygwin* (but not as straightforward)

---

**Web-Base Galaxy Platform**
- Available from the Galaxy Toolshed
Installing *breseq*

• **Easiest way**
  – Install *miniconda* then
    
    `$ conda -c bioconda breseq`

• **Harder way**
  – Download binary for Linux or MacOS X
  – You must also install *bowtie2* and *R*
  – You need to set your `$PATH`

• **Hardest way**
  – Download source code or clone GitHub repository
  – Follow the instructions in the DEVELOPER text file
  – Requires you to have a C++ compiler, dev version of libz with headers, autoconf, automake, etc., installed.
Most Common Install Problems

• You get a message like this, or *breseq* has an error because it can’t find an installed bowtie2 or R:

```
zsh: command not found: breseq
```

  – One of these commands is not in your `$PATH$
  – Great explanation if you don’t understand: [https://astrobiomike.github.io/unix/modifying_your_path](https://astrobiomike.github.io/unix/modifying_your_path)

• You are on a computer cluster and get errors when generating output plots or empty plots
  – Installed R does not have graphics capability
  – Try installing your own version (using miniconda, for example) or ask the system administrator.
Crafting your *breseq* command

**Basic *breseq* command**

\$ breseq -r reference.gbk reads_1.fastq reads_2.fastq

References (-r) can be in GenBank, GFF3, or FASTA format.

Multiple read files can be used. Paired/unpaired are treated the same.

Multiple reference files can be used:
- -r genome.fasta -r plasmid.gff3

Read files can be gzipped: reads_1.fastq.gz

Speed up execution by using multiple threads: -j 8

View common options

\$ breseq

View all options

\$ breseq -h or \$ breseq --help
There are three overall modes for running breseq…

- Population or Polymorphism mode
  - Every read could be from any individual.
  - `$ breseq -p`

- Consensus mode (the default)
  - All reads are from a single clone.
  - `$ breseq ...`

- Targeted mode
  - Reads are from only a subset of the genome
  - `$ breseq -t`

Important! Each mode has different assumptions/options.
Reference file considerations

• **Microbes (<20Mb):** download GenBank or GFF3 files with both DNA sequence and features.

• **Important:** having transposable elements annotated leads to better predictions!

• **What do I do if there is no reference?**
  – *de novo* assemble and annotate your own
  – **Recommendation:** Unicycler **PROKKA**
  – You may need to iteratively improve the assembly and annotation to get the best results. You could use gdtools APPLY (see advanced workshop).
Specifying reference sequences

You can have three types of references:

1. **Normal** (-r, --reference)
   - Call all kinds of mutations. Each sequence is a different episome.

2. **Contig** (-c, --contig-reference)
   - This is a de novo assembly. Treat all of the sequences in this file as if they are from the same episome (e.g., one chromosome)
   - This improves calling deletions by uniformly assigning a read-depth of coverage across

3. **Junction-only** (-s, --junction-only-reference)
   - I am searching for where part of this sequence was inserted into my genome. Don’t want breseq call mutations in this sequence.
   - Example: integration cassette / transposon
**Read file considerations**

**Sequencing technology**
- Can work with any FASTQ
- Best results with short-read data (< 1000 bases)
- Not appropriate for long-read data (Nanopore, PacBio, etc.) In this case, you should de novo assemble and then compare assemblies.

**Recommended depth of coverage**

- >40x for clonal samples
- >120x for population samples

More coverage is unlikely to give improvements without error correction (ex: molecular barcodes).

**Adaptor and Barcode Removal**

You must trim your reads to remove these!

Use *fastp*, *trimmomatic*, etc. You can evaluate reads with *fastqc*.

If you don’t clean this up, then they may result in reads not mapping (90% of length must be covered by the read alignment by default).
Example *breseq* input/output

Let’s look at some results!  [https://barricklab.org/breseq](https://barricklab.org/breseq)

---

**Example 1a:** Analyzing an evolved *E. coli* clone with a high quality reference sequence for its ancestor (LTEE Ara+1 50,000 generations, Clone A)

```
breseq -p -l 80 -r REL606.gbk SRR2584524.fastq.gz
```

View Results

**Example 1b:** What the results look like if you run this same clonal sample in polymorphism mode (LTEE Ara+1 50,000 generations, Clone A)

```
breseq -p -l 80 -r REL606.gbk SRR2584524.fastq.gz
```

View Results

**Example 2:** Results for another evolved clone that was sequenced with longer reads (LTEE Ara+1 50,000 generations, Clone B)

```
breseq -r REL606.gbk SRR2584534_1.fastq.gz SRR2584534_2.fastq.gz
```

View Results

**Example 3:** Analyzing the mixed population that both of these clones were isolated from (LTEE Ara+1 50,000 generations, Population)

```
breseq -j 8 -p -r REL606.gbk SRR6173952_1.fastq.gz SRR6173952_2.fastq.gz
```

View Results

**Example 4:** Results from mapping to reference genome of a closely related strain—many predictions (links removed to save disk space).

```
breseq -r NC_006913.3.MG1655.gbk SRR2584534_1.fastq.gz SRR2584534_2.fastq.gz
```

View Results

**Example 5:** Analyzing an *E. coli* cell that contains a plasmid

```
breseq -r E_coli_M1110_NC_007779.1.gbk -r GFP_Plasmid_SK04.gbk AR_E1_GTTTCG_L005_R2_001.fastq.gz AR_E1_GTTTCG_L005_R1_001_1.fastq.gz AR_E1_GTTTCG_L005_R1_001_2.fastq.gz AR_E1_GTTTCG_L005_R2_001_1.fastq.gz
```

View Results

**Example 6a:** Locating the insertion site of an integration cassette in the *A. baylyi* genome using a junction reference (best option)

```
breseq --junction-only-reference pBTK622_tdk-kanR_cassette_for_Golden_Transformation.gbk -r Acinetobacter-baylyi-ADP1-WT.gff3 G2_CCGTCC_L007_R1_001.fastq.gz
```

View Results

**Example 6b:** Same sample set using junction reference
Utilities to explore output

You can run utility subcommands from inside the main output directory of a `breseq` run. $breseq --help to see others.

$ breseq BAM2ALN
-o alignment.html
REL606:3483047-3483047

$ breseq BAM2COV
-o coverage.png
REL606:3480047-3486047

These can help with identifying copy number changes (e.g., duplications) and understanding complex structural variation.
Tutorial: Population Samples (Polymorphism Mode)

In this exercise, you will analyze two population (metagenomic) samples using breseq to track the frequencies of evolved alleles and changes in genetic diversity in population Ara-3 of the Lenski long-term evolution experiment (LTEE). As discussed in Tutorial: Clonal Samples (Consensus Mode) this population evolved citrate utilization after 31,500 generations.

Tutorial: Clonal Samples (Consensus Mode)

This tutorial expands on the Test Drive. You will analyze mutations in the genomes of multiple clones isolated from population Ara-3 of the Lenski long-term evolution experiment (LTEE). A complex mutation is present in these samples that was necessary for evolution of the aerobic citrate utilization trait (Cit+). In addition to some tips on breseq usage and examples of interpreting more complex mutations in the output, this tutorial also introduces functionality in the gtools utility command that can be used to compare and analyze mutations in an entire set of evolved genomes.

Note: This tutorial was created for the EMBO Practical Course Measuring intra-species diversity using high-throughput sequencing held 27–31 July 2015 in Oeiras, Portugal.

Warning: If you encounter any breseq or gtools errors or crashes in running this tutorial, please report issues on GitHub.

1. Download data files

First, create a directory called tutorial_clonal:

$ mkdir tutorial_clones
$ cd tutorial_clones

Reference sequence

breseq prefers the reference sequence in Genbank or GFF3 format. In this example, the
Let us know how we can help!

These slides can be downloaded at [http://barricklab.org/breseq](http://barricklab.org/breseq)

Post bug reports and issues on GitHub

Please check that you are using the newest `breseq` version first!
## Acknowledgments

### Breseq Developers

Dan Deatherage  
David Knoester  
Geoffrey Colburn  
Matt Strand  
Jordan Borges  
Aaron Reba

### Funding

NIH K99/R00  
(GM087550)  
NSF CAREER  
(CBET-1554179)  
NSF BEACON Center  
(DBI-0939454)

Thanks to many breseq users and research collaborators who have given feedback over the past decade!

Including Richard Lenski, Dominique Schneider, Olivier Tenaillon, Vaughn Cooper, Michael Desai, Yousif Shamoo, Zachary Blount, Genoscope, the Gulbenkian Institute, and members of these and many other research groups and communities.