

# AMR Genomics: Reads to Reports

Finlay Maguire (Dalhousie University)

# Overview

- Quality control
- Trimming/Error Correction
- Genome Assembly
- Predicting AMR Genes from Assemblies
- Standardising Output
- Workflows

Things that won't be covered:

- Read-based analyses
- Metagenomics
- Many alternative tools!

# Materials

The screenshot shows a GitHub repository page for 'fmaguire / amr\_training\_workshop\_practical'. The repository is public and has 1 unwatch, 0 stars, and 0 forks. The main content area displays a list of files and folders, each with a commit message and timestamp. The files include '0.raw\_data', '1.quality\_control', '2.assembly', '3.amr\_gene\_detection', '4.hAMRnization', '5.simple\_automation', '6.workflows', '.gitignore', 'LICENSE.txt', and 'README.md'. The 'README.md' file is selected, showing its content: 'JPI-AMR PHA4GE MRC-CLIMB-BIG-DATA AMR Genomics Training Workshop'. On the right side, there are sections for 'About' (practical exercise for the JPI-AMR/PHA4GE/CLIMB AMR Genomics Training Workshop), 'Releases' (no releases published), 'Packages' (no packages published), and 'Languages' (HTML 99.8%, Shell 0.2%).

fmaguire / amr\_training\_workshop\_practical (Public) Unwatch 1 Star 0 Fork 0

<> Code Issues Pull requests Actions Projects Wiki Security Insights ...

master Go to file Add file Code About

fmaguire Fix indices 13 seconds ago 3

0.raw_data	Add practical overview	11 minutes ago
1.quality_control	Add practical overview	11 minutes ago
2.assembly	Add practical overview	11 minutes ago
3.amr_gene_detection	Add practical overview	11 minutes ago
4.hAMRnization	Add practical overview	11 minutes ago
5.simple_automation	Fix indices	13 seconds ago
6.workflows	Fix indices	13 seconds ago
.gitignore	Add practical overview	11 minutes ago
LICENSE.txt	Add practical overview	11 minutes ago
README.md	Fix indices	13 seconds ago

README.md

## JPI-AMR PHA4GE MRC-CLIMB-BIG-DATA AMR Genomics Training Workshop

HTML 99.8%  
Shell 0.2%

github.com/fmaguire/amr\_training\_workshop\_practical

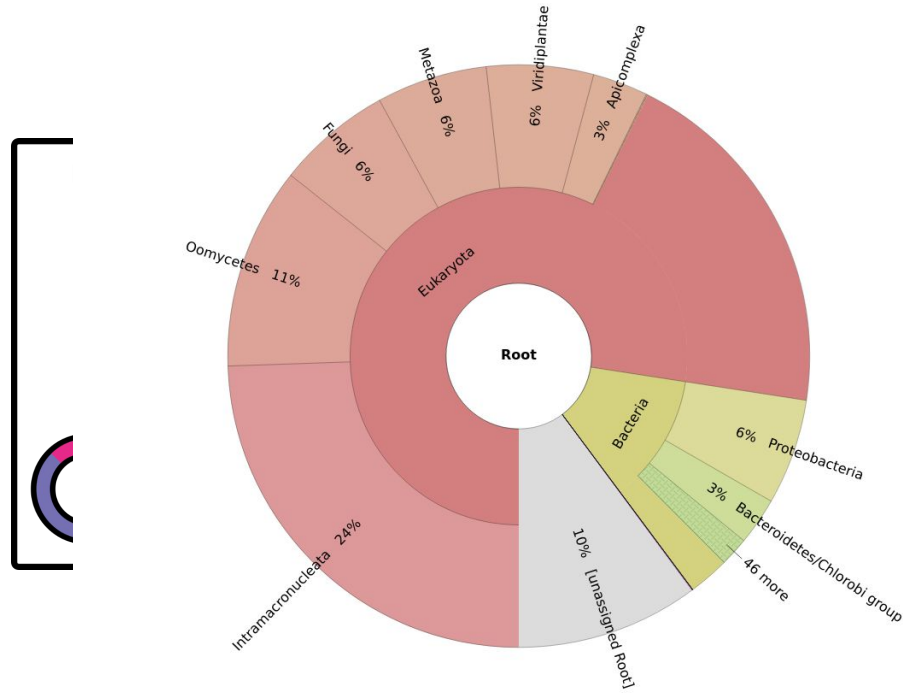
# Setting Yourself Up

- Bioinformatics relies heavily on the UNIX shell:
  - Linux
  - Mac's OSX
  - Windows Subsystem Linux (WSL)
- Package managers and containers make installation easier
- Environments prevent tools getting in each other's way
- Bioconda provides both ([docs.conda.io/en/latest/miniconda.html](https://docs.conda.io/en/latest/miniconda.html) + [bioconda.github.io](https://bioconda.github.io))

```
conda create -n amr fastp shovill ncbi-amrfinderplus hAMRionization
conda activate amr
```

# Garbage In - Garbage Out: Quality Control

- Positive and Negative Controls
- Contamination checks (e.g., kraken2 + krona)
- Sequencing quality checks
  - Quality scores
  - Over-representation
  - N's



# Tidying Up Your Reads: Trimming/Error Correction

## Before filtering

total reads:	5.512546 M
total bases:	689.068250 M
Q20 bases:	664.553262 M (96.442299%)
Q30 bases:	641.367221 M (93.077459%)
GC content:	37.854518%

## After filtering

### Filtering result

reads passed filters:	5.403942 M (98.029876%)
reads with low quality:	105.014000 K (1.905000%)
reads with too many N:	3.590000 K (0.065124%)
reads too short:	0 (0.000000%)

```
fastp --in1 s1.fq.gz --out1 s1_trimmed.fq.gz
```

```
_trimmed.fq.gz
```

## Adapters

### Adapter or bad ligation of read1

The input has little adapter percentage (~0.007310%), probably it's trimmed before.

Sequence	Occurrences
all adapter sequences	627

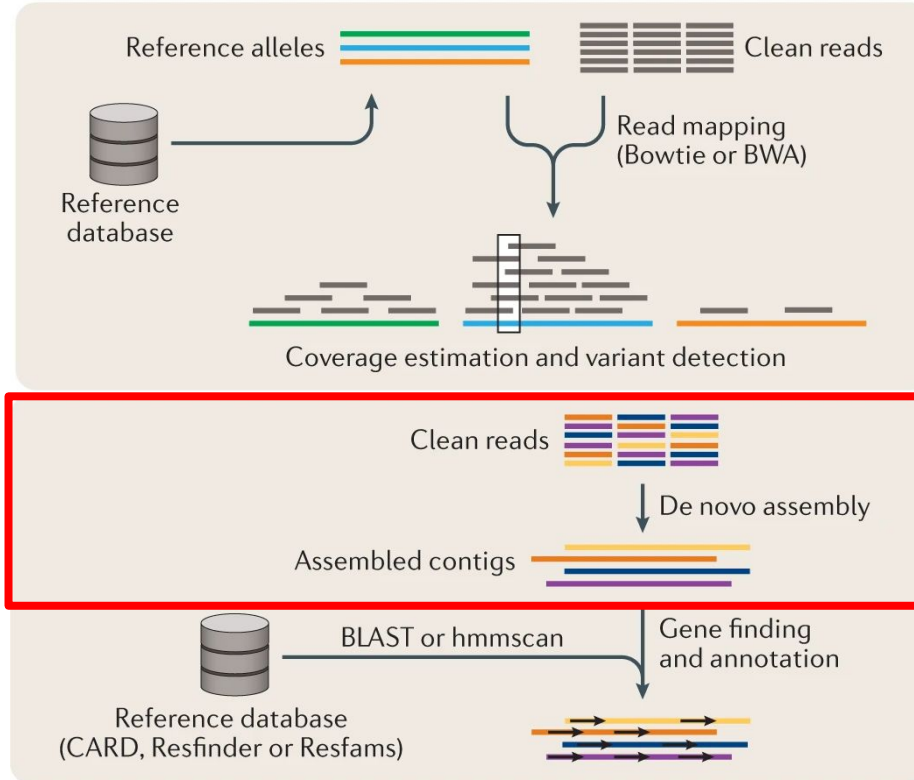
### Adapter or bad ligation of read2

The input has little adapter percentage (~0.007310%), probably it's trimmed before.

Sequence	Occurrences
all adapter sequences	627

Note: reads in repository are simulated to be simple; trimming isn't doing much to them

# Turning Reads Into a Genome



# De novo Assembly

**Assembly theory**

vs

**Practical assembly**



Similar patterns. Millions of pieces. Missing pieces. Damaged pieces.  
And you don't know the right answer!

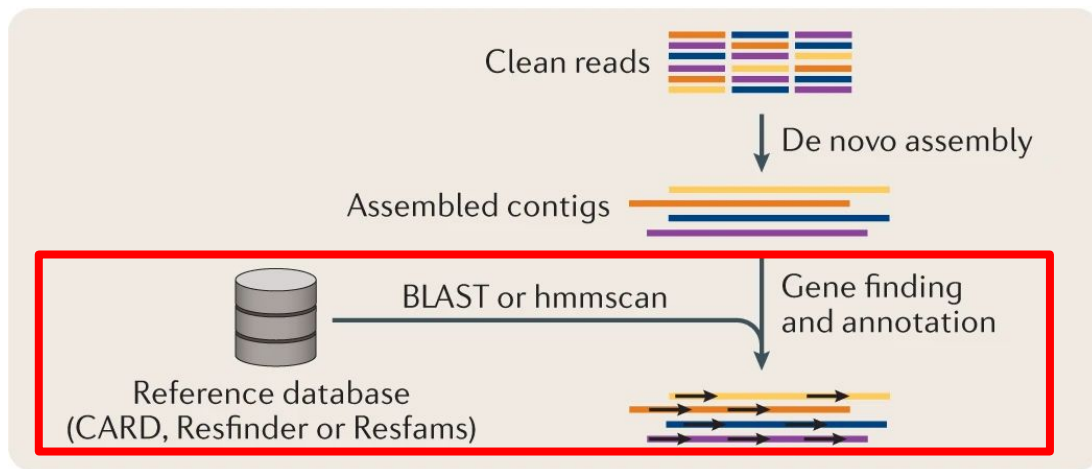


# Shovill

- Estimate genome size (mash)
- Downsample reads (seqtk) to ~100x
- Trim reads (trimmomatic)
- Error correct reads (lighter)
- Stitch overlapping reads (Flash)
- Assemble reads (SPAdes)
- Correct mistakes (BWA-MEM + Pilon)

```
shovill --R1 sampleA_R1_trimmed.fq.gz --R2 sampleA_R2_trimmed.fq.gz \  
--outdir sampleA_assembly
```

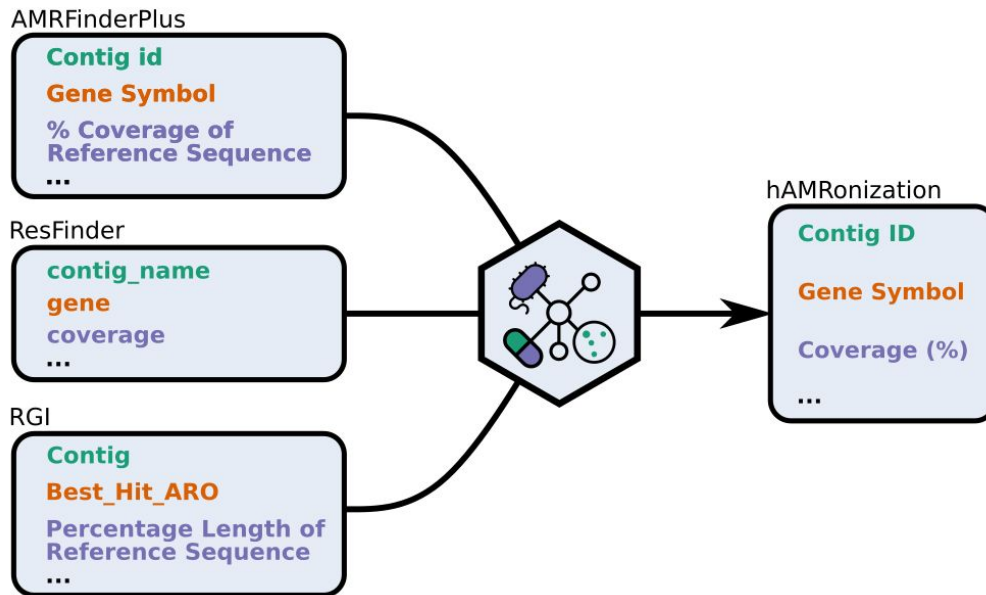
# Contig-based AMR Gene Prediction



10.1038/s41576-019-0108-4

```
amrfinder --update  
> Database version: 2021-09-30.1  
amrfinder --nucleotide sampleA_assembly/contigs.fa \  
--output amrfinderplus_results.tsv
```

# Standardising AMR Detection Results



```
hamronize amrfinderplus --analysis_software_version 3.10.16 \  
  --reference_database_version 2021-09-30.1 \  
  --input_file_name sampleA amrfinderplus_results.tsv \  
> hAMRonized_amr_report.tsv
```

# Summarising Results

Public Health Alliance for Genomic Epidemiology

Search Show Only Genomes With Hits Restore Results

abricate: config 0 amrfinderplus: config 0 csstar: config 0 resfinder.py: config 0 staramr: config 0

Genome ID	abricate	amrfinderplus	csstar	resfinder.py	staramr
ERR873305	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873306	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873307	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873308	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873309	16 hits	14 hits	14 hits	6 hits	11 hits
ERR873310	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873311	11 hits	10 hits	10 hits	6 hits	9 hits
ERR873312	11 hits	10 hits	10 hits	6 hits	9 hits

Search Results

Total hits: 0

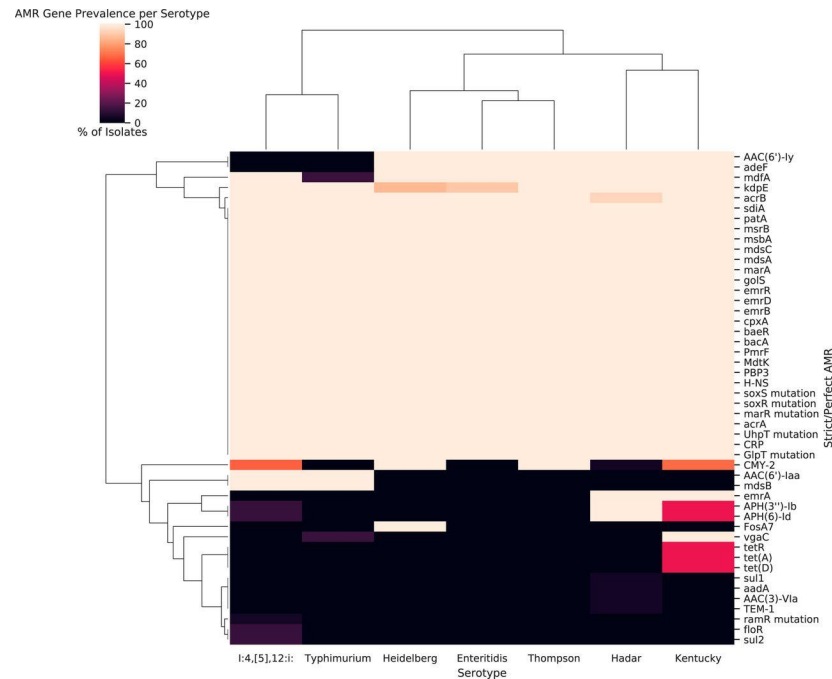
Genomes with hits: 0

Tools with hits: 0

Differential results: 0

Selected

Compare Clear



```
hamronize summarize --summary_type interactive hAMRonized_amr_report.tsv \
> amr_summary.html
```

Great, that is how we do it for 1 sample but what  
about 10-1,000s?

## Option 1: Lots of Typing and Mistakes

```
fastp raw_reads_1 > trimmed_reads_1  
shovill trimmed_reads_1 > assembly_1  
amrfinderplus assembly_1 > amr_predict_1  
hamronize amr_predict_1 > hamronized_1
```

```
fastp raw_reads_2 > trimmed_reads_2  
shovill trimmed_reads_2 > assembly_2  
amrfinderplus assembly_2 > amr_predict_2  
hamronize amr_predict_2 > hamronized_2
```

```
fastp raw_reads_3 > trimmed_reads_3  
shovill trimmed_reads_3 > assembly_3  
amrfinderplus assembly_3 > amr_predict_3  
hamronize amr_predict_3 > hamronized_3
```

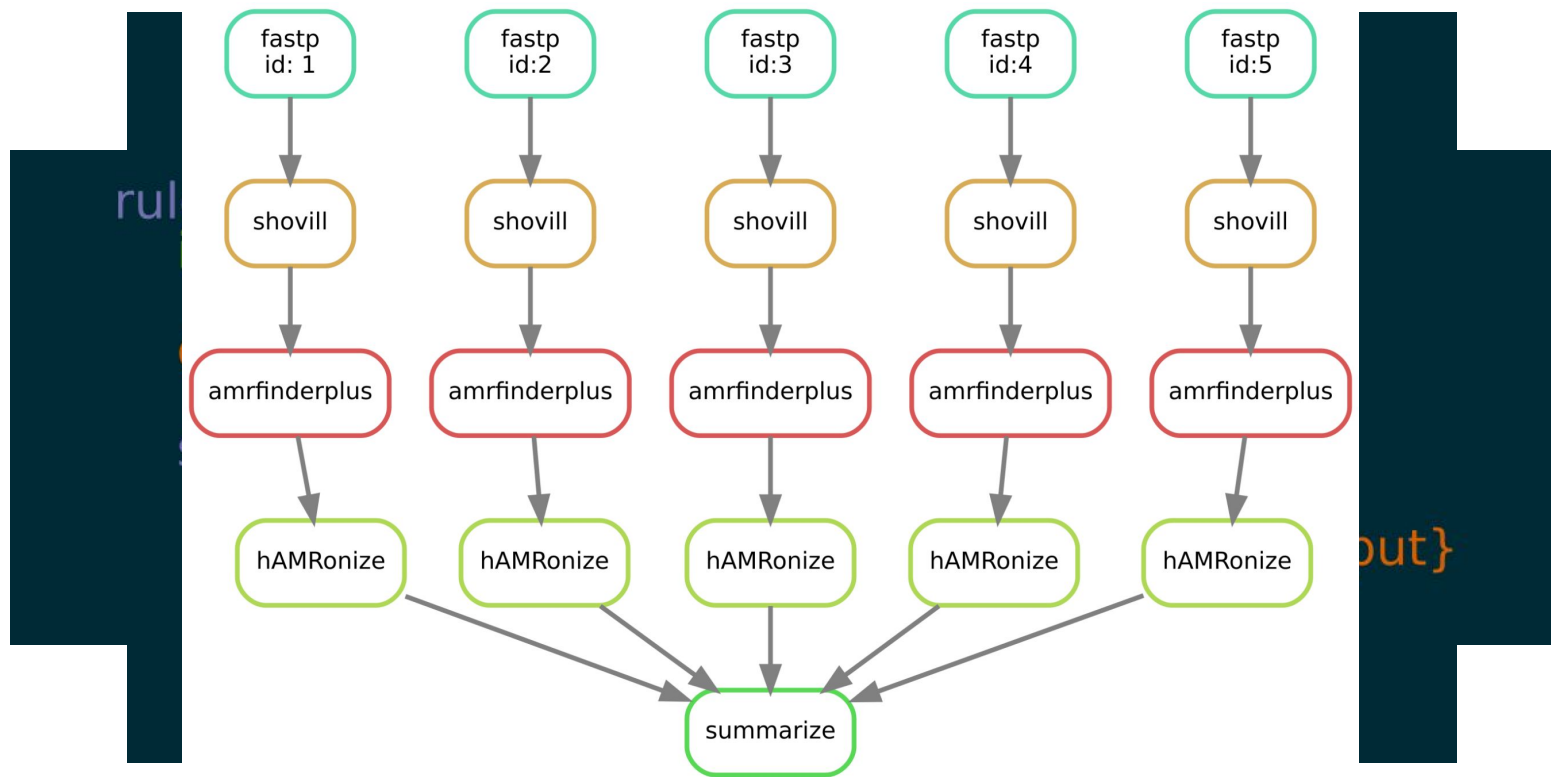
## Option 2: Bash Loop

```
fastp raw_reads_1 > trimmed_reads_1  
shovill trimmed_reads_1 > assembly_1  
amrfinderplus assembly_1 > amr_predict_1  
hamronize amr_predict_1 > hamronized_1
```

```
for sample in $(seq 1 10000); do  
  fastp raw_reads_$sample > trimmed_reads_$sample  
  shovill trimmed_reads_$sample > assembly_$sample  
  amrfinderplus assembly_$sample > amr_predict_$sample  
  hamronize amr_predict_$sample > hamronized_$sample  
done
```

```
shovill trimmed_reads_3 > assembly_3  
amrfinderplus assembly_3 > amr_predict_3  
hamronize amr_predict_3 > hamronized_3
```

# Option 3: Workflows!





# Option 4: Someone else's workflow!

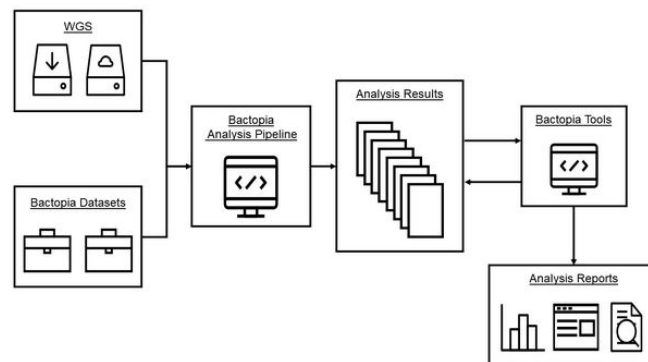
- Galaxy (user friendly and not terminal-based): Published Community Workflows (use highly rated/used)
- Snakemake (familiar python but difficult model): “Snakemake workflows” (curated best practice workflows).
- Nextflow (unfamiliar language but simple model): “nf-core” (curated best practice workflows)

The screenshot shows the GitHub repository page for 'tseemann/nullarbor'. At the top, it indicates the repository is public and has 18 watchers, 92 stars, and 36 forks. Below this are navigation tabs for Code, Issues (57), Pull requests, Actions, Projects, Wiki, and Security. The main content area shows a commit by 'andersgs' on 3 Aug 2020 with 354 comments. A file tree is visible with folders like 'bin', 'conf', 'perl5', 'plugins', and 'scripts', each with a brief description and the time since the last update.

## Bactopia

Bactopia is a flexible pipeline for complete analysis of bacterial genomes. The goal of Bactopia is to process your data with a broad set of tools, so that you can get to the fun part of analyses quicker!

Bactopia can be split into three main parts: [Bactopia Datasets](#), [Bactopia Analysis Pipeline](#), and [Bactopia Tools](#).



# Take-aways

- Bioinformatics is built around terminals
- Use conda environments or containers to install tools
- Always read the documentation and help messages for bioinformatics tools
- Robust quality control and controls are vital
- Consider hAMRonizing your AMR gene predictions
- Workflows let you efficiently and robustly run an analysis over many samples
- Don't reinvent the wheel: use high quality workflows that already exist (modify if you have to)

# Acknowledgements

- Our speakers:
  - Dr. Kara Tsang (LSHTM)
  - Dr. Mike Feldgarden (NCBI/NIH)
  - Inês Mendes (IMM/ULisboa)
- Organising committee:
  - Dr. Finlay Maguire, Dr. Andrew Page, Dr. Emma Griffiths, Prof. Mark Pallen, **Lisa Marchioretto**, Dr. Jessica Boname, Dr. Carolyn Johnson
  - JPI-AMR, PHA4GE, MRC-CLIMB-BIG-DATA
- Panelists:
  - A/Prof. Henk den Bakker (UGeorgia)
  - Dr. Anthony Underwood (CPGS)