



Data Carpentry Genomics

Introduction

- Experimental data
- Workshop overview

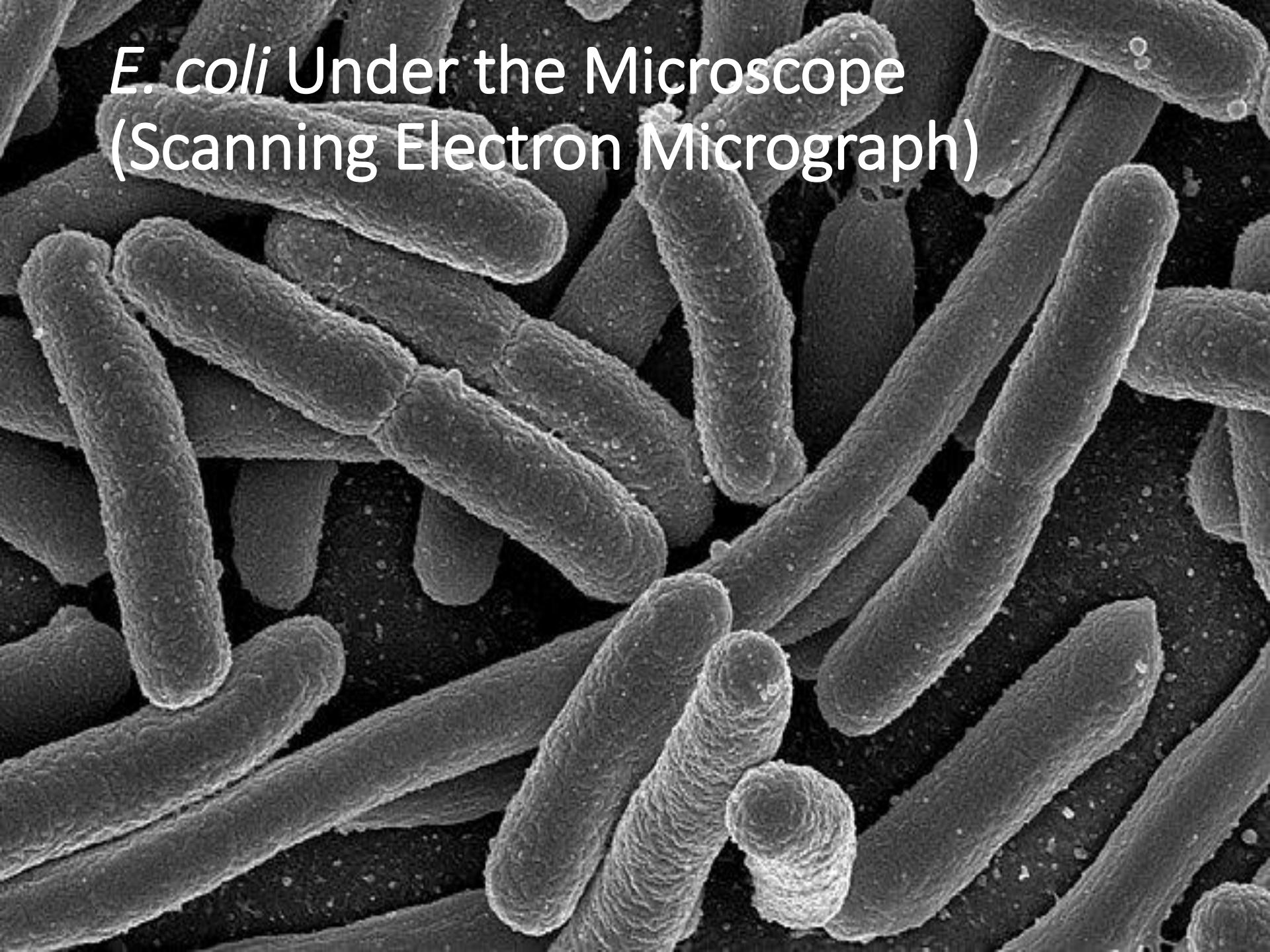


Data Carpentry Genomics

Introduction

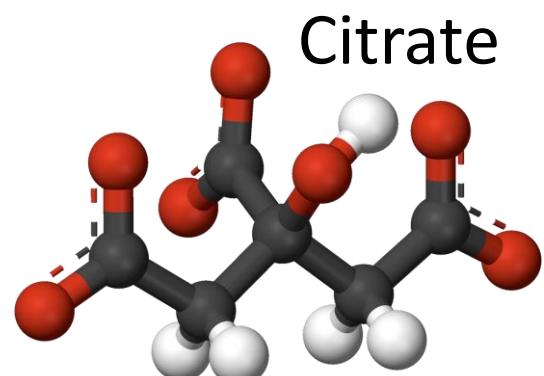
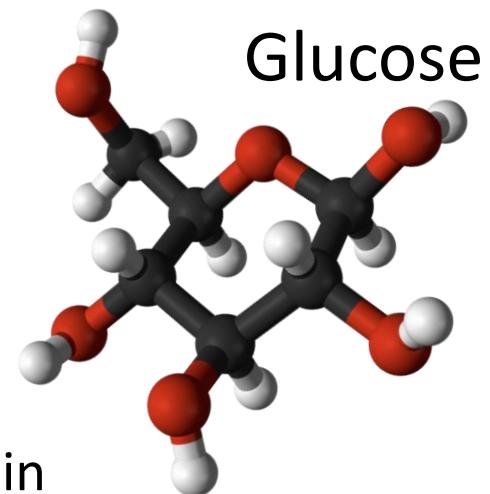
- Experimental data
- Workshop overview

E. coli Under the Microscope (Scanning Electron Micrograph)



E. coli (*Escherichia coli*)

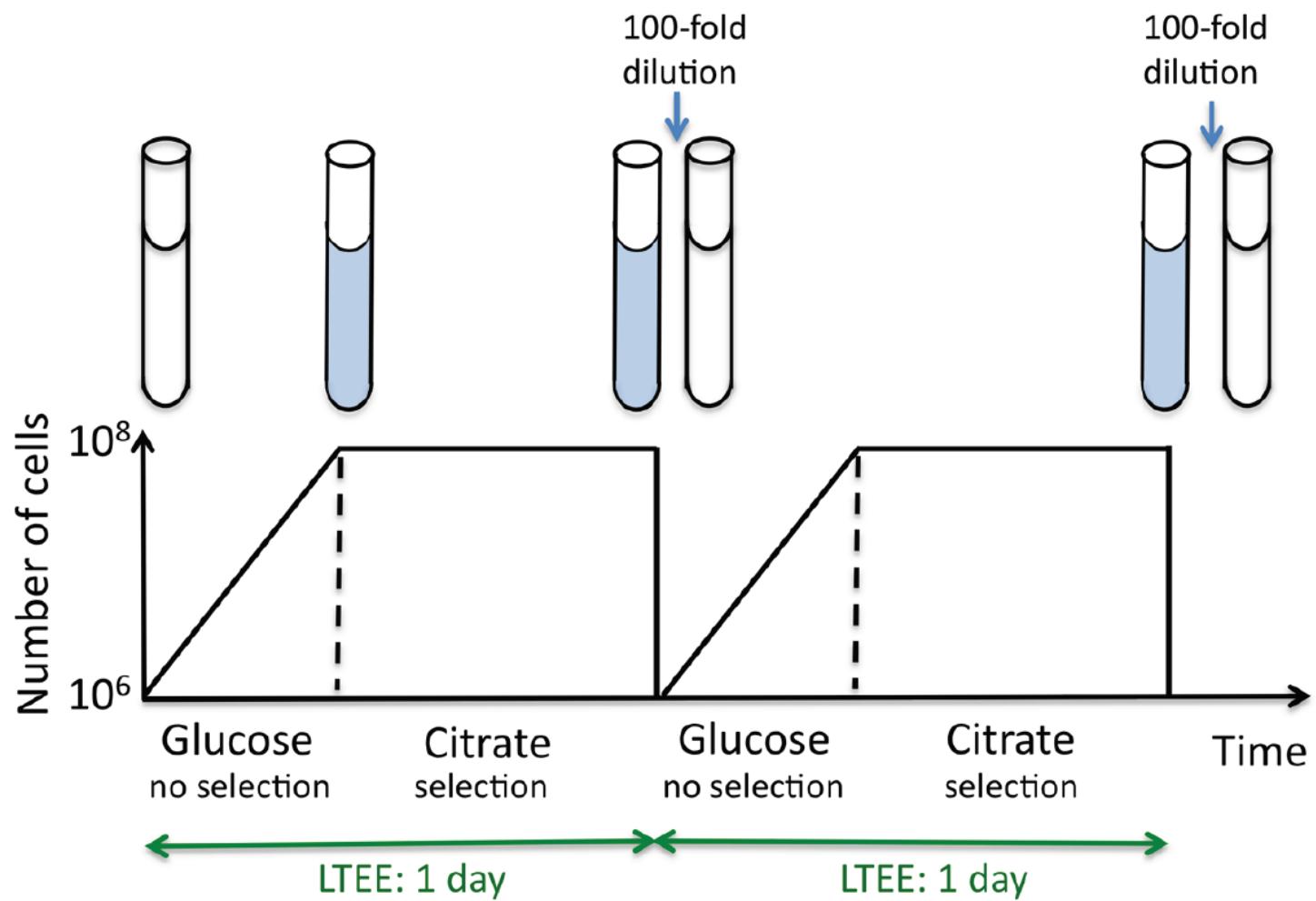
- **Prokaryotes:** single copy of their circular chromosome per cell
- **Aerobic AND Anaerobic metabolism:**
 - Uses oxygen when available, using the same glycolysis and the citric acid cycle pathways found in eukaryotes to generate energy from carbon-containing molecules like **glucose**.
 - **Citrate** can only be metabolized by *E. coli* in the absence of oxygen.
- **Easy to grow and maintain**
 - Short reproduction time ≈20 min



E. coli long-term evolution experiment (LTEE)



E. coli long-term evolution experiment (LTEE)

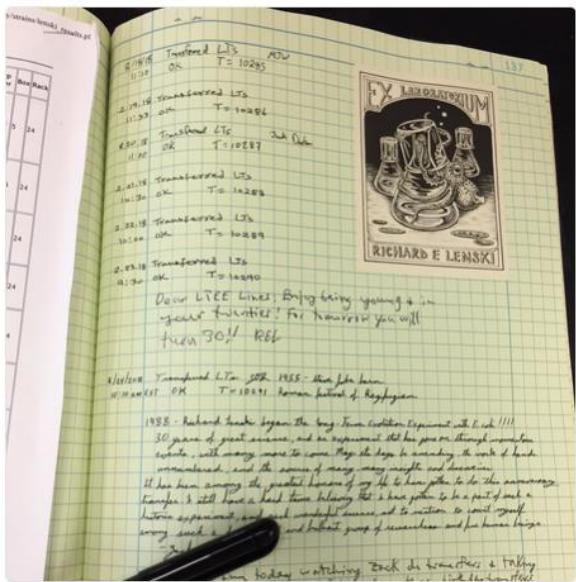




Richard E. Lenski
@RELenski

Follow

30 years of the #LTEE are in the notebook. The experiment started 24-Feb-1988 & here it is 24-Feb-2018. Postdoc extraordinaire Zack Blount did transfers & I placed bookplate to commemorate. Thanks to everyone who has helped. Now here's to the next 30 years. Evolve, LTEE, evolve!



8:11 AM - 24 Feb 2018

433 Retweets 1,040 Likes



The data in Blount et al. 2012 is restricted to populations from *one* of those initial populations.

E. coli long-term evolution experiment (LTEE)



Data Carpentry Genomics data set

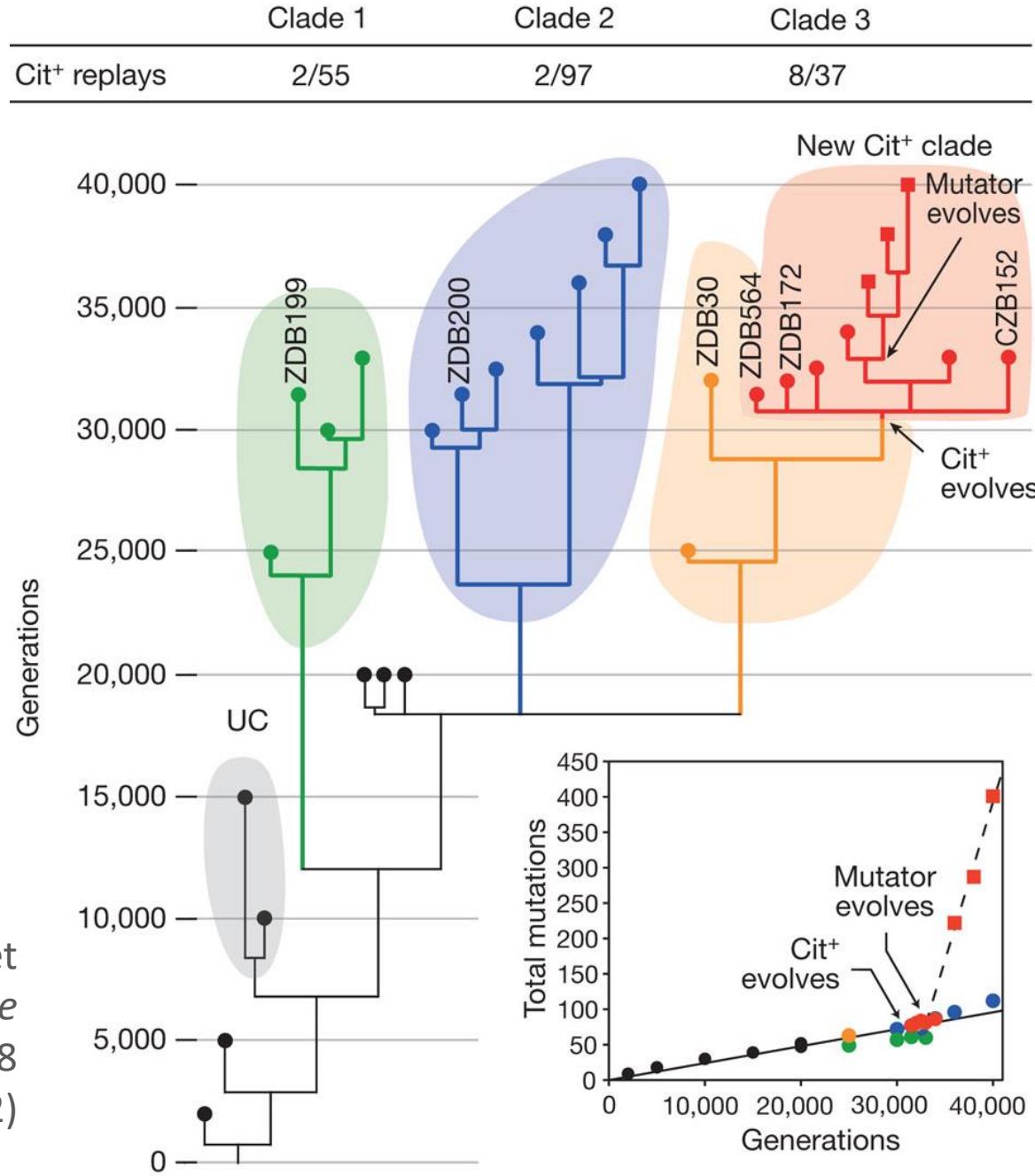


- Growth medium contains small amount of glucose (food source) and citrate (helps *E. coli* obtain iron from medium)
- Around 33,000 generations (≈ 15 years into the experiment!), *E.coli* population could grow aerobically on citrate (Cit^+)
- Remarkable: inability to grow aerobically on citrate (Cit^-) is characteristic of *E. coli* as a species

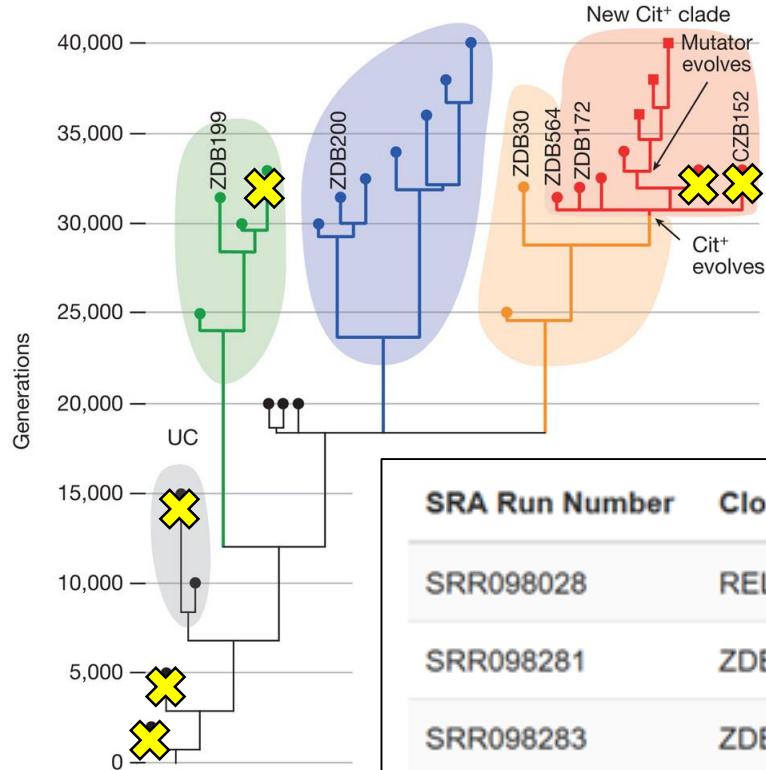
“Back in 2008, this was really frustrating for those of us who wanted to know the molecular details. I'm sure this was frustrating for Blount and Lenski and their colleagues as well. We had no idea how difficult it would be to sort out the mutations. Now we know, and the photo of Zachary Blount with all his Petri dishes drives home the point.”

Laurence A. Moran (Professor Emeritus in the Department of Biochemistry at the University of Toronto)

ZD Blount et
al. *Nature*
489, 513–518
(2012)

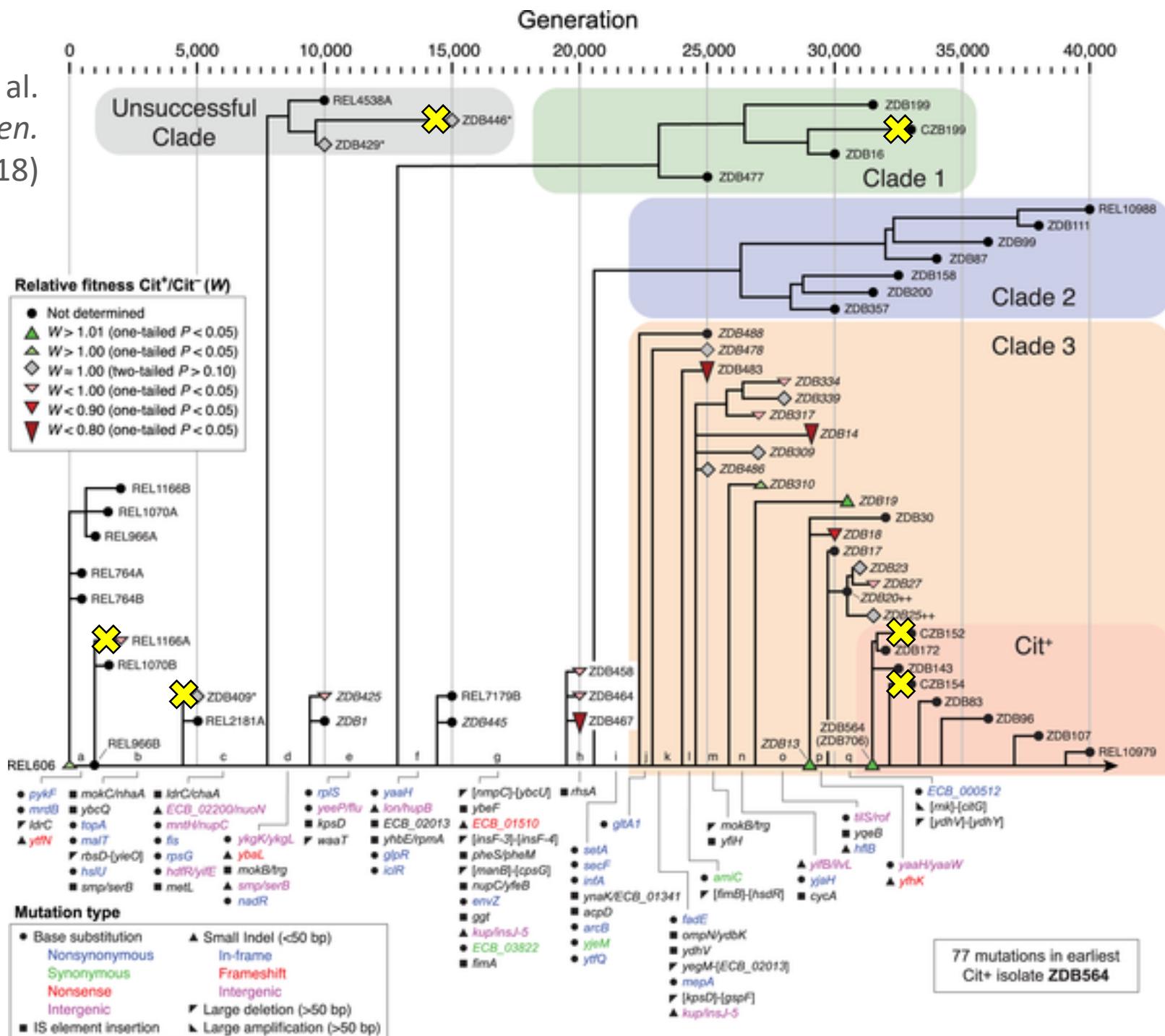


DC Genomics data set



For the purposes of this workshop we're going to be working with 6 of the sequence reads from this experiment. We also made up genome sizes for each of the strains, to look at the relationship between Cit status and genome size. **The genome sizes are not real data!!**

| SRA Run Number | Clone | Generation | Cit | GenomeSize |
|----------------|----------|------------|---------|------------|
| SRR098028 | REL1166A | 2,000 | Unknown | 4.63 |
| SRR098281 | ZDB409 | 5,000 | Unknown | 4.6 |
| SRR098283 | ZDB446 | 15,000 | Cit- | 4.66 |
| SRR097977 | CZB152 | 33,000 | Cit+ | 4.8 |
| SRR098026 | CZB154 | 33,000 | Cit+ | 4.76 |
| SRR098027 | CZB199 | 33,000 | Cit- | 4.59 |





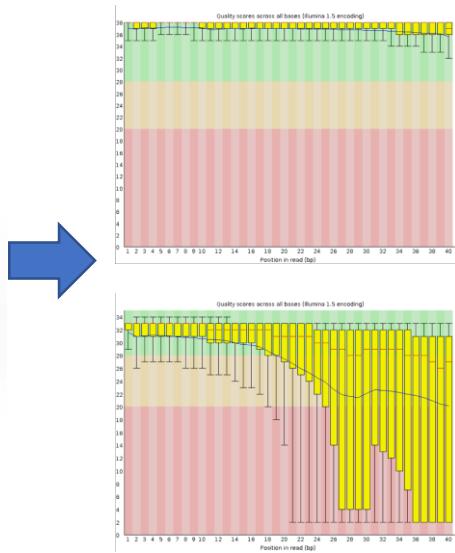
Data Carpentry Genomics Introduction

- Experimental data
- Workshop overview

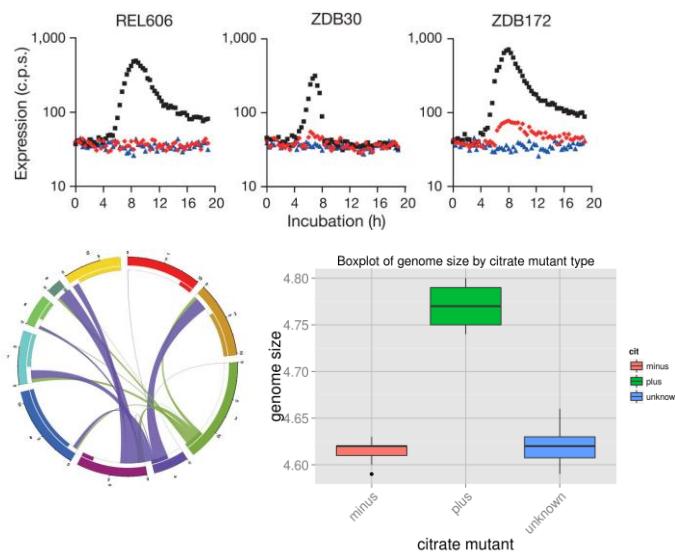
Set up project



Process data



Analyze & visualize

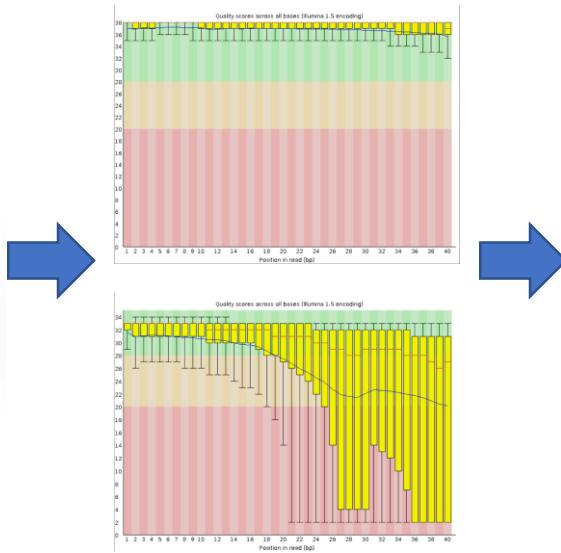


Set up project

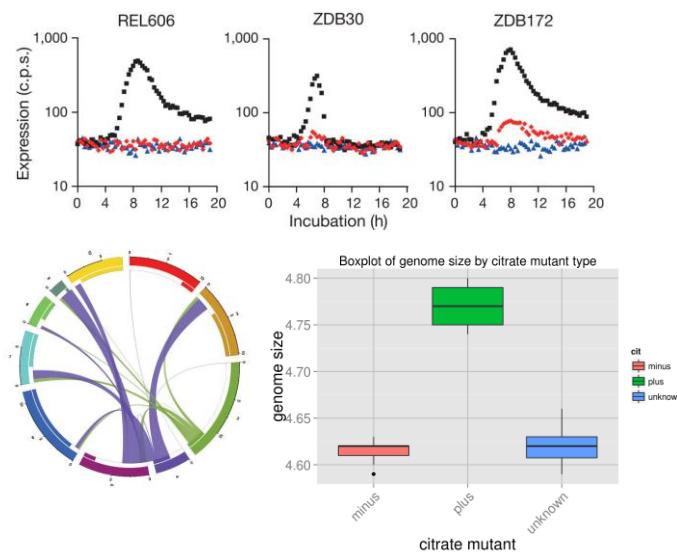


- ① Tuesday morning:
Project organization
- Metadata files
 - Open data

Process data



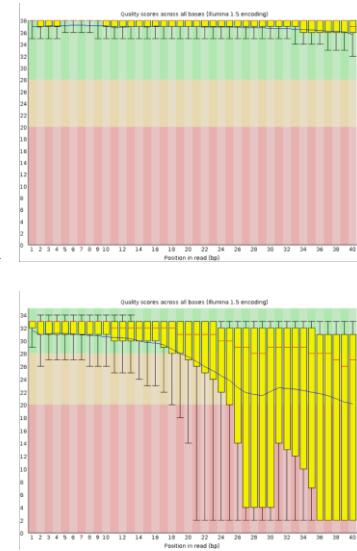
Analyze & visualize



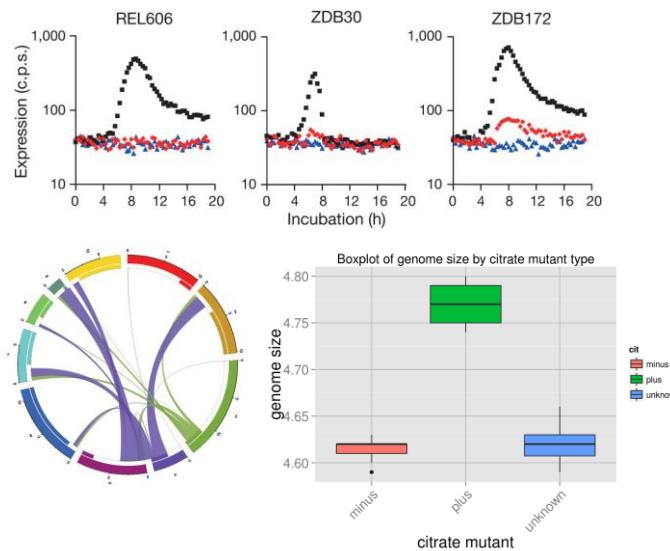
Set up project



Process data



Analyze & visualize



```
c:\MINGW64\U\Users\cvanghe\Documents
cvanghe@LAPTOP-GUSR6SCE MINGW64 ~/Documents
$ cd Documents
bash: cd: Documents: No such file or directory
cvanghe@LAPTOP-GUSR6SCE MINGW64 ~/Documents
$ ls
amazon-web-server/ 'My Videos'@
articles/          notes/
desktop.ini          personal-projects/
MINTIME/            presentations/
misc/              Programs/
Mobaxterm_v9.4/      R/
'My Music'@        unix-practice/
'My Pictures'@      cvanghe@LAPTOP-GUSR6SCE MINGW64 ~/Documents
$ |
```

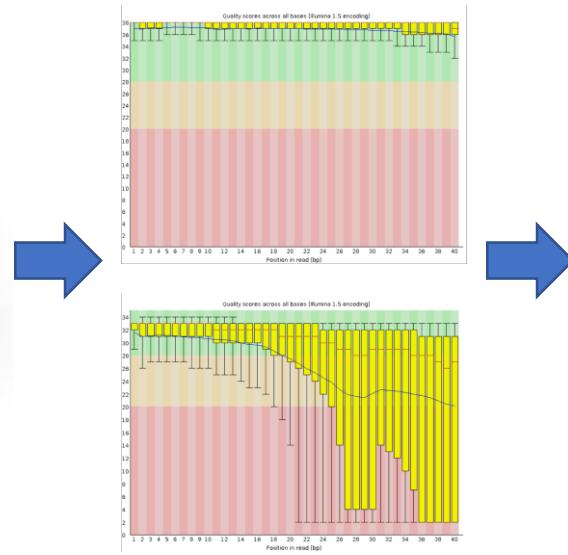
② Tuesday afternoon: Unix command line

- Fast,
- Repeatable &
- Reproducible data processing

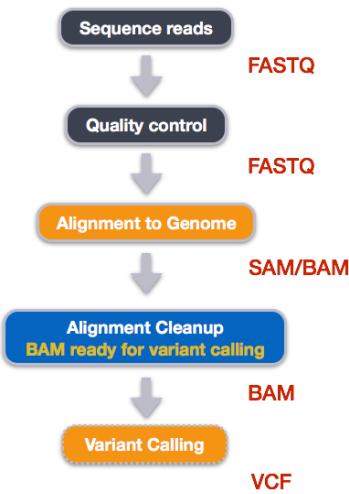
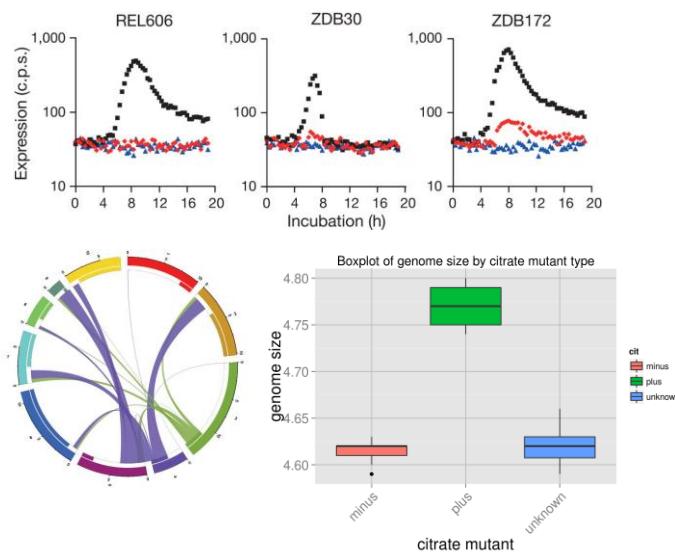
Set up project



Process data



Analyze & visualize



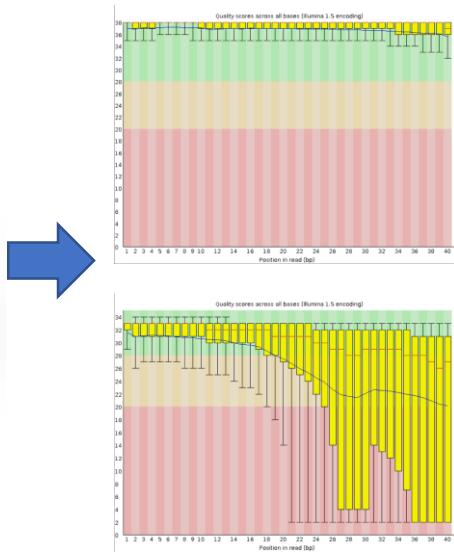
③ Wednesday morning & afternoon: Sequencing pipeline

- Quality control of read data
- Alignment to genome
- Variant calling

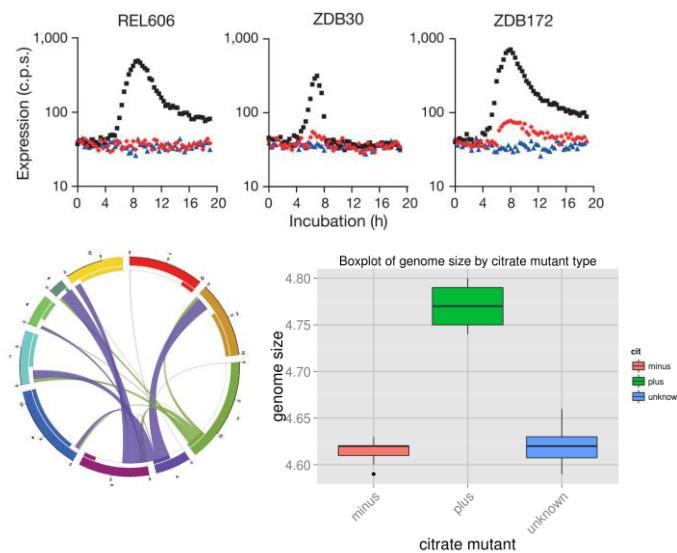
Set up project



Process data



Analyze & visualize



④ Wednesday afternoon:

Data analysis with R

- Manipulate data tables
- dplyr
- Make plots

References



<http://www.evo-ed.com>

- https://commons.wikimedia.org/wiki/File:EscherichiaColi_NIAID.jpg
- https://en.wikipedia.org/wiki/E._coli_long-term_evolution_experiment#/media/File:Lenski%27s_12_long-term_lines_of_E._coli_on_25_June_2008.jpg
- Richard E. Lenski Twitter feed:
<https://twitter.com/RELenski/status/967431687260065792>
- https://www.researchgate.net/figure/Petri-plates-aplenty-Former-student-now-postdoc-Zachary-Blount-and-Richard-Lenski_fig2_279308783
- Laurence A. Moran blog: <https://sandwalk.blogspot.ca/2013/12/lenskis-long-term-evolution-experiment.html>
- ZD Blount et al. *Nature* 489, 513–518 (2012)
- D. Leon et al. *PLOS Genetics* (2018)
- <http://www.molecularecologist.com/2014/12/migration-routes-as-circos-plots-in-r/>