Introduction: Identifying Mutations and Studying Microbial Genome Evolution with *breseq*

Jeffrey E. Barrick

Department of Molecular Biosciences

THE UNIVERSITY OF TEXAS AT AUSTIN

http://barricklab.org

@barricklab

Workshop Introduction

- When is *breseq* the right tool?
  - Installation
  - Basic usage
  - Input: references and reads
  - Output: HTML, GenomeDiff, etc.
- Analysis examples: Lenski LTEE
- Using *breseq* in research and education: **The other speakers in this workshop!**
- Online tutorials and workshop survey

When is *breseq* the right tool?

Deatherage, D. E., Barrick, J. E. (2014) **Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using *breseq***. *Methods Mol. Biol.* **1151**: 165–188. https://doi.org10.1007/978-1-4939-0554-6_12

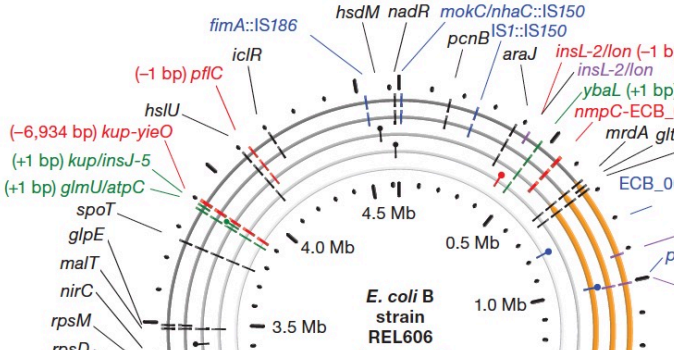
https://barricklab.org/breseq https://github.com/barricklab/breseq

- 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

Some uses of *breseq*

Genetics

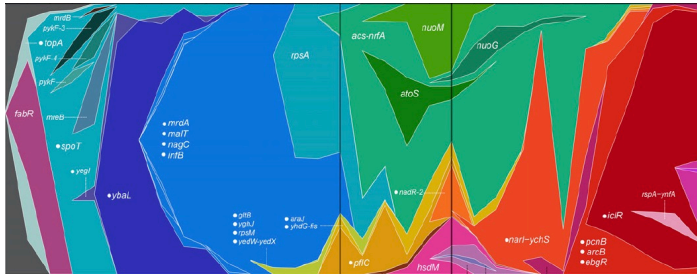
- Mechanisms of antibiotic resistance
- Mapping suppressor mutations



Barrick *et al.* (2009) *Nature*

Experimental evolution

- Rates/nature of genome evolution
- Genetic diversity in populations



Maddamsetti *et al.* (2015) *Genetics*

Biotechnology

- Verifying engineered plasmids/genomes
- Understanding beneficial mutations that arise during adaptive laboratory evolution

Installing *breseq*

breseq 0.35.4 documentation » **breseq** Manual

breseq

Next topic
Introduction

This Page
Show Source

Quick search
Go

breseq Manual

Contents:

Latest release
v0.35.6
c7cf8df
Compare

breseq v0.35.6

Edit

jeffreybarrick released this 25 days ago

- Fixed compatibility with GenBank reference files produced by Prokka and NCBI PGAP, and with GFF3 files produced by PGAP.

Assets 5

breseq-0.35.6-Linux-x86_64.tar.gz	13.7 MB
breseq-0.35.6-MacOSX-10.9+.tar.gz	13.9 MB
breseq-0.35.6-Source.tar.gz	12.4 MB
Source code (zip)	
Source code (tar.gz)	

https://github.com/barricklab/breseq/releases

Can be used on Linux, Mac OSX, and Windows machines; and in the Galaxy web platform.

Options to download and install by compiling from source code or using precompiled binaries.

Requires R and bowtie2.

Installing *breseq*

breseq 0.35.4 documentation » **breseq** Manual

breseq

Next topic
Introduction

This Page
Show Source

Quick search
Go

breseq Manual

Contents:

- Introduction
  - Citing **breseq**
- Installation
  - Install external dependencies
  - Method 1. Binary download
  - Method 2.
    - Insta
    - Insta
    - Insta
  - Method 3.
  - Installing
  - Installing
  - Troublesh
- Usage Summar
  - **breseq**
  - Command
  - Command
- Test Drive
  - 1. Downlo
    - Refe
    - Rea
  - 2. Run **bre**
  - 3. Open **b**

BIOCONDA

Navigation

User Docs
Contributing to Bioconda
Developer Docs
Tutorials

Bioconda @ Github
Package Index
chat on gitter

Quick search
Go

recipe breseq

A computational pipeline for finding mutations relative to a reference sequence in short-read DNA re-sequencing data.

Homepage: https://github.com/barricklab/breseq
License: GPL / GPL-3.0
Recipe: /breseq/meta.yaml

package breseq

downloads 18k | container none

Versions: ▶ 0.35.6-0, 0.35.5-1, 0.35.5-0, 0.35.4-0, 0.35.3-0, 0.35.2-0, 0.35.1-0, 0.35.0-0, 0.34.1-0, ...

Depends: bowtie2 >=2.0.0,!=2.0.3,!=2.0.4,!=2.3.1, libgcc-ng >=9.3.0, libstdcxx-ng >=9.3.0, r-base, zlib >=1.2.11,<1.3.0a0

Required By:

Installation

With an activated Bioconda channel (see 2. Set up channels), install with:

```
conda install breseq
```

Recommended method

BIOCONDA®

http://bioconda.github.io/index.html

Basic *breseq* usage

breseq 0.35.4 documentation » **breseq** Manual

Basic *breseq* command

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

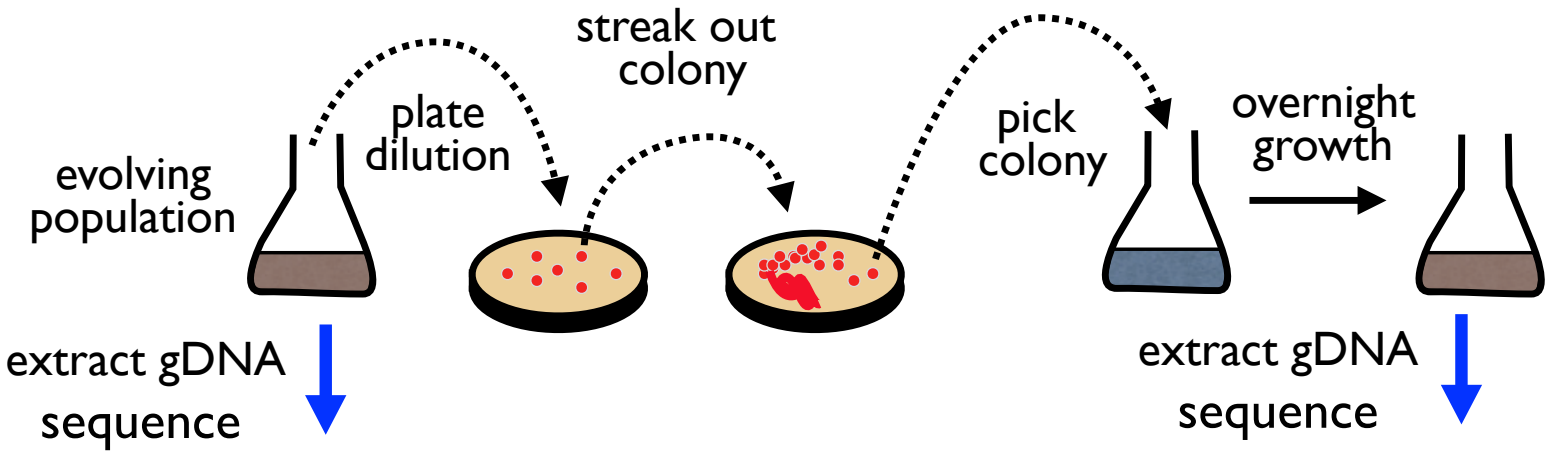
References can be in GenBank, GFF3, or FASTA format.

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

Read files can be gzipped: `reads_1.fastq.gz`

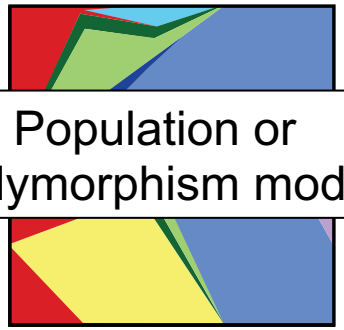
Speed up execution by using multiple cores: `-j 8`

Two main types of samples



Every read could be from any individual.

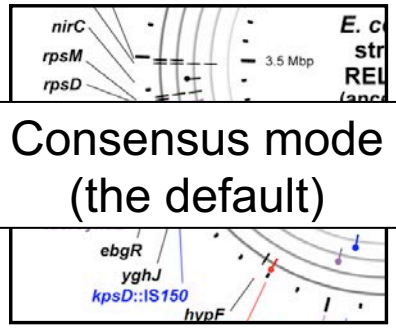
All reads are from a single clone.



Population or Polymorphism mode

```
$ breseq -p ...
```

breseq



Consensus mode (the default)

```
$ breseq ...
```

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
  - If you are using an assembly that has multiple contigs use **`-c`** instead of **`-r`** for specifying the contig reference:

```
$ breseq -c contigs.gbk reads_1.fastq reads_2.fastq
```

  - You may need to iteratively improve the assembly and annotation to get the best results. See **`gdtools APPLY`**.

Downloading a reference from NCBI

⚠️ Be sure you download a GenBank file that has both <u>features</u> and the <u>sequence</u>!

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Search

Advanced Help

GenBank Send to:

Escherichia coli B str. REL606, complete sequence

NCBI Reference Sequence: NC_012967.1

FASTA Graphics

Go to:

```
LOCUS       NC_012967             4629812 bp    DNA     circular CON 13-MAY-2021
DEFINITION  Escherichia coli B str. REL606, complete sequence.
ACCESSION   NC_012967
VERSION     NC_012967.1
DBLINK      BioProject: PRJNA224116
            BioSample: SAMN02603421
            Assembly: GCF_000017985.1
KEYWORDS    RefSeq.
SOURCE      Escherichia coli B str. REL606
  ORGANISM  Escherichia coli B str. REL606
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 4629812)
  AUTHORS   Jeong,H., Barbe,V., Vallenet,D., Choi,S.-H., Lee,C.H., Lee,S.-W.,
            Vacherie,B., Yoon,S.H., Yu,D.-S., Cattolico,L., Hur,C.-G.,
            Park,H.-S., Segurens,B., Blot,M., Schneider,D., Studier,F.W.,
            Oh,T.K., Lenski,R.E., Daegelen,P. and Kim,J.F.
  CONSRTM   International E. coli B Consortium
  TITLE     Complete genome sequence of Escherichia coli (B) REL606
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 4629812)
  AUTHORS   Daegelen,P., Vallenet,D., Barbe,V., Cattolico,L. and Segurens,B.
  TITLE     Direct Submission
  JOURNAL   Submitted (24-AUG-2007) UMR 8030 CNRS, Inserm, Genoscope
```

Change region shown

Customize view

- Abbreviated view
- Customize

Basic Features

- All features
- Gene, RNA, and CDS features only

Display options

- Show sequence
- Show reverse complement
- Show gap features

Update View

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Related information

Assembly

BioProject

BioSample

Open

Select

Click

Downloading a reference from NCBI

⚠️ Be sure you download a GenBank file that has both <u>features</u> and the <u>sequence</u>!

```
     gene            4629102..4629788
                     /gene="yjtD"
                     /locus_tag="ECB_RS22810"
                     /old_locus_tag="ECB_04279"
     CDS             4629102..4629788
                     /gene="yjtD"
                     /locus_tag="ECB_RS22810"
                     /old_locus_tag="ECB_04279"
                     /EC_number="2.1.1.-"
                     /inference="COORDINATES: similar to AA
                     sequence:RefSeq:NP_710140.2"
                     /note="Derived by automated computational analysis using
                     gene prediction method: Protein Homology."
                     /codon_start=1
                     /transl_table=11
                     /product="tRNA/rRNA methyltransferase"
                     /protein_id="WP_001223167.1"
                     /translation="MRITIILVAPARAENIGAAARAMKTMGFSELRIVDSQAHLEPAT
                     RWVAHGSGDIIDNIKVFPTLAESLHDVDFTVATTARSRAKYHYYATPVELVPLLEEKS
                     SWMSHAALVFGREDSGLTNEELALADVLTGVPMVADYPSLNLGQAVMVYCYQLATLIQ
                     QPTKSDTTADQHQLQALRERVMALLTTLAVADDIKLVDWLQQRLGLLEQRDTAMLHRL
                     LHDIEKNITK"
ORIGIN
        1 agcttttcat tctgactgca acgggcaata tgtctctgtg tggattaaaa aaagagtgtc
       61 tgatagcagc ttctgaactg gttacctgcc gtgagtaaat taaaatttta ttgacttagg
      121 tcactaaata ctttaaccaa tataggcata gcgcacagac agataaaaat tacagagtac
      181 acaacatcca tgaaacgcat tagcaccacc attaccacca ccatcaccat taccacaggt
      241 aacggtgcgg gctgacgcgt acaggaaaca cagaaaaaag cccgcacctg acagtgcggg
      301 cttttttttc gaccaaaggt aacgaggtaa caaccatgcg agtgttgaag ttcggcggta
      361 catcagtggc aaatgcagaa cgttttctgc gggttgccga tattctggaa agcaatgcca
      421 ggcaggggca ggtggccacc gtcctctctg cccccgccaa aatcaccaac cacctggtgg
      481 cgatgattga aaaaaccatt agcggccagg atgctttacc caatatcagc gatgccgaac
      541 gtatttttgc cgaacttttg acgggactcg ccgccgccca gccgggattc ccgctggcgc
      601 aattgaaaac tttcgtcgat caggaatttg cccaaataaa acatgtcctg catggcatta
      661 gtttgttggg gcagtgcccg gatagcatca acgctgcgct gatttgccgt ggcgagaaaa
      721 tgtcgatcgc cattatggcc ggcgtattag aagcgcgcgg tcacaacgtt accgttatcg
      781 atccggtcga aaaactgctg gcagtggggc attacctcga atctaccgtc gatattgctg
      841 agtccacccg ccgtattgcg gcaagtcgca ttccggctga tcacatggtg ctgatggcag
```

Scroll down until you see **ORIGIN**.

There should be a nucleotide sequence here!

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).

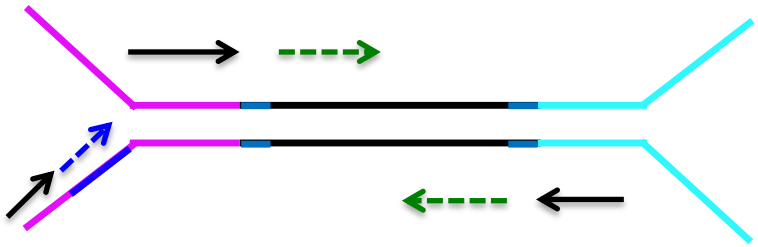
Read terminology

Fragment library (input DNA sample)

Library prep

Sequencing library
Double-stranded or Y- **adaptors** added

DNA sequencing



External Barcode and (optional) **Internal Barcode** (6–12 bases each) allow many samples can be run in one space (lane).

*Data is **demultiplexed** during processing.*

Primers **Reads** (100–1000+ bases)

FASTQ quality and trimming

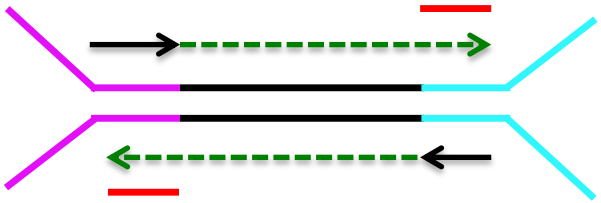
Check the quality of your FASTQ data FastQC.app

- Have internal barcodes been removed?
- Do I want to trim low-quality bases?

⚠️ Be careful to trim adaptors from your reads

(*breseq* requires >90% of a read's length to map)

Readthrough into adaptors *is especially common with new longer Illumina reads!*



Programs that can help:

`fastp, trimmomatic, cutadapt`

```
$ breseq -j 8 -r REL606.gbk SRR030255_1.fastq.gz SRR030255_2.fastq.gz
```

36 minutes later… Open **`output/index.html`**

HTML Output

breseq version 0.35.6 revision c7cf8df53bcd
mutation predictions | marginal predictions | summary statistics | genome diff | command line log

Predicted mutations

evidence	position	mutation	annotation	gene	description
RA	380,188	A→C	F239L (TTT→TTG)	*araJ* ←	predicted transporter
RA	475,292	+G	coding (14/1677 nt)	*ybaL* ←	predicted transporter with NAD(P)-binding Rossmann-fold domain
RA	649,391	T→A	I471F (ATC→TTC)	*mrdA* ←	transpeptidase involved in peptidoglycan synthesis (penicillin-binding protein 2)
RA	683,496	A→C	V65G (GTT→GGT)	*nagC* ←	DNA-binding transcriptional dual regulator, repressor of N-acetylglucosamine
JC JC	969,836	IS*150* (+) +3 bp	coding (810-812/2283 nt)	*pflB* ←	pyruvate formate lyase I
RA	1,329,516	C→T	H33Y (CAC→TAC)	*topA* →	DNA topoisomerase I
JC JC	1,544,289	IS*150* (–) +3 bp	coding (150-152/1536 nt)	*xasA* ←	predicted glutamate:gamma-aminobutyric acid antiporter
JC JC	1,733,647	IS*150* (–) +3 bp	coding (683-685/1413 nt)	*pykF* →	pyruvate kinase
RA	1,976,879	T→G	intergenic (-57/-76)	*yedW* ← / → *yedX*	predicted DNA-binding response regulator in two-component system with YedV/hypothetical protein
RA	2,082,685	G→A	A494V (GCT→GTT)	*yegI* ←	hypothetical protein
RA	2,499,315	G→A	intergenic (-110/-179)	*maeB* ← / → *talA*	malic enzyme/transaldolase A
RA	3,045,069	G→T	T312N (ACC→AAC)	*yghJ* ←	predicted inner membrane lipoprotein
RA	3,248,957	A→T	D764E (GAT→GAA)	*infB* ←	translation initiation factor IF-2
MC JC	3,289,962	Δ16 bp	coding (96-111/4554 nt)	*gltB* →	glutamate synthase, large subunit
RA	3,339,158	A→C	intergenic (+22/-4)	*yhdG* → / → *fis*	tRNA-dihydrouridine synthase B/DNA-binding protein Fis
RA	3,370,027	T→A	K117M (AAG→ATG)	*rpsM* ←	30S ribosomal protein S13
RA	3,424,910	G→A	M1M (ATG→ATA) †	*nirC* →	nitrite transporter
RA	3,483,047	C→A	R455S (CGC→AGC)	*malT* →	transcriptional regulator MalT
RA	3,762,741	A→T	K662I (AAA→ATA)	*spoT* →	bifunctional (p)ppGpp synthetase II/ guanosine-3',5'-bis pyrophosphate 3'-pyrophosphohydrolase
RA	3,875,632	$(T)_{7\to8}$	intergenic (-66/+287)	*glmU* ← / ← *atpC*	bifunctional N-acetylglucosamine-1-phosphate uridyltransferase/glucosamine-1-phosphate acetyltransferase/F0F1 ATP synthase subunit epsilon
RA	3,893,551	+G	intergenic (+6/-50)	*kup* → / → *insJ-5*	potassium transporter/IS150 hypothetical protein
MC JC	3,894,997	Δ6,934 bp	IS*150*-mediated	*rbsD–[yieO]*	*rbsD, rbsA, rbsC, rbsB, rbsK, rbsR, [yieO]*
RA	4,100,655	C→T	M192I (ATG→ATA)	*hslU* ←	ATP-dependent protease ATP-binding subunit
RA	4,126,706	$(T)_{8\to7}$	coding (342/879 nt)	*pflC* →	pyruvate formate lyase II activase
RA	4,560,632	T→C	Y131C (TAC→TGC)	*hsdM* ←	DNA methylase M

Unassigned missing coverage evidence

			seq id	start	end	size	←reads	reads→	gene	description
*	*	÷	REL606	546953–547700	555934–555877	8178–8982	20 [18]	[16] 19	*[insB-6]–[ECB_00513]*	[insB-6],insA-6,nmpC,ybcR,ybcS,ybcT,ybcU,ECB_00510,nohB,ECB_00512,[ECB_00513]
*	*	÷	REL606	2031675–2031718	2054970–2054943	23226–23296	21 [17]	[18] 21	*[manB]–[cpsG]*	[manB],manC,insB-14,insA-14,wbbD,wbbC,wzy,wbbB,wbbA,vioB,vioA,wzx,rmlC,rfbA,rfbD,rfbB,galF,wcaM,wcaL,wcaK,wzxC,wcaJ,[cpsG]

Unassigned new junction evidence

		seq id	position	reads (cov)	reads (cov)	score	skew	freq	annotation	gene	product
*	?	REL606	= 547699	NA (NA)	80 (1.360)	37/70	0.2	NA	noncoding (1/768 nt)	*IS1*	repeat region
	?	REL606	555924 =	NA (NA)					coding (1209/2346 nt)	*ECB_00513*	conserved hypothetical protein

breseq version 0.35.6 revision c7cf8df53bcd
mutation predictions | marginal predictions | summary statistics | genome diff | command lin

Predicted mutations

evidence	position	mutation	annotation	gene	
RA	380,188	A→C	239L (TTT→TTG)	*araJ* ←	predicted trans
RA	475,292	+G	coding (14/1677 nt)	*ybaL* ←	predicted trans
RA	649,391	T→A	I471F (ATC→TTC)	*mrdA* ←	transpeptidase
RA	683,496	A→C	V65G (GTT→GGT)	*nagC* ←	DNA-binding tr
JC JC	969,836	IS*150* (+) +3 bp	coding (810-812/2283 nt)	*pflB* ←	pyruvate forma
RA	1,329,516	C→T	H33Y (CAC→TAC)	*topA* →	DNA topoisom
JC JC	1,544,289	IS*150* (–) +3 bp	coding (150-152/1536 nt)	*xasA* ←	predicted gluta
JC JC	1,733,647	IS*150* (–) +3 bp	coding (683-685/1413 nt)	*pykF* →	pyruvate kinas
RA	1,976,879	T→G	intergenic (-57/-76)	*yedW* ← / → *yedX*	predicted DNA
RA	2,082,685	G→A	A494V (GCT→GTT)	*yegI* ←	hypothetical pr
RA	2,499,315	G→A	intergenic (-110/-179)	*maeB* ← / → *talA*	malic enzyme/transaldolase A
RA	3,045,069	G→T	T312N (ACC→AAC)	*yghJ* ←	predicted inner membrane lipoprotein
RA	3,248,957	A→T	D764E (GAT→GAA)	*infB* ←	translation initiation factor IF-2
MC JC	3,289,962	Δ16 bp	coding (96-111/4554 nt)	*gltB* →	glutamate synthase, large subunit
RA	3,339,158	A→C	intergenic (+22/-4)	*yhdG* ← / → *fis*	tRNA-dihydrouridine synthase B/DNA-binding protein Fis
RA	3,370,027	T→A	K117M (AAG→ATG)	*rpsM* ←	30S ribosomal protein S13
RA	3,424,910	G→A	M1M (ATG→ATA) †	*nirC* →	nitrite transporte
RA	3,483,047	C→A	R455S (CGC→AGC)	*malT* →	transcriptional re
RA	3,762,741	A→T	K662I (AAA→ATA)	*spoT* →	bifunctional (p)p
RA	3,875,632	$(T)_{7\to8}$	intergenic (-66/+287)	*glmU* ← / ← *atpC*	bifunctional N-ac synthase subunit epsilon
RA	3,893,551	+G	intergenic (+6/-50)	*kup* → / → *insJ-5*	potassium transp
MC JC	3,894,997	Δ6,934 bp	IS*150*-mediated	*rbsD–[yieO]*	*rbsD, rbsA, rbsC*
RA	4,100,655	C→T	M192I (ATG→ATA)	*hslU* ←	ATP-dependent
RA	4,126,706	$(T)_{8\to7}$	coding (342/879 nt)	*pflC* →	pyruvate format
RA	4,560,632	T→C	Y131C (TAC→TGC)	*hsdM* ←	DNA methylase

Unassigned missing coverage evidence

			seq id	start	end	size	←reads	reads→	gene	description
*	*	÷	REL606	546953–547700	555934–555877	8178–8982	20 [18]	[16] 19	*[insB-6]–[ECB_00513]*	[insB-6],insA-6,nmpC,ybcR,ybcS,ybcT,ybcU,ECB_00510,nohB,ECB_00512,[ECB_00513]
*	*	÷	REL606	2031675–2031718	2054970–2054943	23226–23296	21 [17]	[18] 21	*[manB]–[cpsG]*	[manB],manC,insB-14,insA-14,wbbD,wbbC,wzy,wbbB,wbbA,vioB,vioA,wzx,rmlC,rfbA,rfbD,rfbB,galF,wcaM,wcaL,wcaK,wzxC,wcaJ,[cpsG]

Unassigned new junction evidence

		seq id	position	reads (cov)	reads (cov)	score	skew	freq	annotation	gene	product
*	?	REL606	= 547699	NA (NA)	80 (1.360)	37/70	0.2	NA	noncoding (1/768 nt)	*IS1*	repeat region
	?	REL606	555924 =	NA (NA)					coding (1209/2346 nt)	*ECB_00513*	conserved hypothetical protein

Mutations (fully predicted)

- Base substitutions
- Small indels
- IS element insertions
- Large deletions

Evidence for other genetic differences that can't be fully resolved

RA = Read Alignment Evidence

breseq version 0.35.6 revision c7cf8df53bcd
mutation predictions | marginal predictions | summary statistics | genome diff | command line log

Predicted mutations

evidence	position	mutation	annotation	gene
RA	380,188	A→C	F239L (TTT→TTG)	*araJ* ←
RA	475,[illegible]	+G	coding (14/1677 nt)	*ybaL* ←
RA	649,391	T→A	I471F (ATC→TTC)	*mrdA* ←
RA	683,496	A→C	V65G (GTT→GGT)	*nagC* ←
JC JC	969,836	IS*150* (+) +3 bp	coding (810-812/2283 nt)	*pflB* ←
RA	1,329,516	C→T	H33Y (CAC→TAC)	*topA* →
JC JC	1,544,289	IS*150* (–) +3 bp	coding (150-152/1[illegible] nt)	*xasA* ←
JC JC	1,733,647	IS*150* (–) +3 bp	coding (683-685/1413 nt)	*pykF* →
RA	1,976,879	T→G	intergenic (-57/-76)	*yedW* ← / → *ydX*
RA	2,082,685	G→A	A494V (GCT→GTT)	*yegI* ←
RA	2,499,315	G→A	intergenic (-110/-179)	*maeB* ← / → *talA*
RA	3,045,069	G→T	T312N (ACC→AAC)	*yghJ* ←
RA	3,248,957	A→T	D764E (GAT→GAA)	*infB* ←
MC JC	3,289,962	Δ16 bp	coding (96-111/4554 nt)	*gltB* →
RA	3,339,158	A→C	intergenic (+22/-4)	*yhdG* → / → *fis*
RA	3,370,027	T→A	K117M (AAG→ATG)	*rpsM* ←
RA	3,424,910	G→A	M1M (ATG→ATA) †	*nirC* →
RA	3,483,047	C→A	R455S (CGC→AGC)	*malT* →
RA	3,762,741	A→T	K662I (AAA→ATA)	*spoT* →
RA	3,875,632	$(T)_{7\to8}$	intergenic (-66/+287)	*glmU* ← / ← *atpC*
RA	3,893,551	+G	intergenic (+6/-50)	*kup* → / → *insJ-5*
MC JC	3,894,997	Δ6,934 bp	IS*150*-mediated	*rbsD–[yieO]*
RA	4,100,655	C→T	M192I (ATG→ATA)	*hslU* ←
RA	4,126,706	$(T)_{8\to7}$	coding (342/879 nt)	*pflC* →
RA	4,560,632	T→C	Y131C (TAC→TGC)	*hsdM* ←

Unassigned missing coverage evidence

			seq id	start	end	size	←reads	reads→	gene	
*	*	÷	REL606	546953–547700	555934–555877	8178–8982	20 [18]	[16] 19	[insB-6]–[ECB_00513]	[ins
*	*	÷	REL606	2031675–2031718	2054970–2054943	23226–23296	21 [17]	[18] 21	[manB]–[cpsG]	[ma G]

Unassigned new junction evidence

		seq id	position	reads (cov)	reads (cov)	score	skew	freq	annotat		
*	?	REL606	= 547699	NA (NA)	80 (1.360)	37/70	0.2	NA	noncoding (1		
	?	REL606	555924 =	NA (NA)					coding (1209/2346 nt)	*ECB_00513*	conserved hypothetical protein

Predicted mutation

evidence	seq id	position	mutation	annotation	gene	description
RA	REL606	380,188	A→C	F239L (TTT→TTG)	*araJ* ←	predicted transporter

Read alignment evidence...

	seq id	position	ref	new	freq	score (cons/poly)	reads	annotation	genes	product
*	REL606	380,188	0	A	C	100.0%	129.8 / NA	40	F239L (TTT→TTG)	*araJ*	predicted transporter

Reads supporting (aligned to +/- strand): **ref** base A (0/0); **new** base C (18/22); **total** (18/22)

```
CAGCACCATCCCTAGCCCAACTAACATCATAATAAAGGTCATCGCCGTTTCCGAAAAACCGGAAAT  >  REL606/380155-380220
                                   |
caacACCATCCCTAGCCCAACTAACATCATAATCaa                                <  1:4049041/33-1 (MQ=38)
cAGCACCATCCCCAGCCCAACTAACATCATAATCaa                                >  2:3988152/1-36 (MQ=38)
 aGCACCATCCCTAGCCCAACTAACATCATAATCAAg                               >  2:2846392/1-36 (MQ=255)
 aGCACCATCCCTAGCCCAACTAACATCATAATCAAg                               >  1:1710600/1-36 (MQ=255)
 aGCACCATCCCTAGCCCAACTAACATCATAATCAAg                               <  2:478931/36-1 (MQ=255)
   cACCATCCCTAGCCCAACTAACATCATAATCAAGGt                             <  2:3427822/36-1 (MQ=255)
   cACCATCCCTAGCCCAACTAACATCATAATCAAGGt                             >  2:2868696/1-36 (MQ=255)
     ccATCCCTAGCCCAACTAACATCATAATCAAGGTCa                           <  1:613685/36-1 (MQ=255)
     ccATCCCTAGCCCAACTAACATCATAATCAAGGTCa                           <  1:3037406/36-1 (MQ=255)
     ccATCCCTAGCCCAACTAACATCATAATCAAGGTCa                           <  1:2129644/36-1 (MQ=255)
      cATCCCTAGCCCAACTAACATCATAATCAAGGTCAt                          >  2:113100/1-36 (MQ=255)
      cATCCCTAGCCCAACTAACATCATAATCAAGGTCAt                          >  2:2149235/1-36 (MQ=255)
      cATCCCTAGCCCAACTAACATCATAATCAAGGTCAt                          >  1:309277/1-36 (MQ=255)
       aTCCCTAGCCCAACTAACATCATAATCAAGGTCATc                         >  1:1146898/1-36 (MQ=255)
       aTCCCTAGCCCAACTAACATCATAATCAAGGTCATc                         <  2:880957/36-1 (MQ=37)
         cccTAGCCCAACTAACATCATAATCAAGGTCATCGc                       <  2:3039312/36-1 (MQ=255)
         cccTAGCCCAACTAACATCATAATCAAGGTCATCGc                       <  1:1074806/36-1 (MQ=255)
         cccAAGCCCAACTAACATCATAATCAAGGTCATCGc                       <  1:2828983/36-1 (MQ=37)
          ccTAGCCCAACTAACATCATCATCAAGGTCATCGcc                      >  2:904743/1-36 (MQ=37)
           cTAGCCCAACTAACATCATAATCAAGGTCATCGCCg                     <  1:2956106/36-1 (MQ=255)
            taacccAACTAACACCATAATCAAGGTCATCGCCGt                    <  2:2953907/33-1 (MQ=25)
            tAGCCCAACTAACATCATAATCAAGGTCATCGC Gt                    >  1:2693636/1-36 (MQ=37)
             aGCCCAACTAACATCATAATCAAGG CATCGCCGtt                   >  2:1240813/1-36 (MQ=37)
               cccAACTAACATCATAATCAAGGTCATCGCCGTTTc                 <  2:465024/36-1 (MQ=255)
                ccAACTAACATCATAATCAAGGTCATCGCCGTTTcc                >  2:136753/1-36 (MQ=255)
                 cAACTAACATCATAATCAAGGCCATCGCCGTTTCCg               >  1:430752/1-36 (MQ=38)
                  aacCaacaTCATAATCAAGGTCATCGCCGTTTCCGa              <  1:566412/36-1 (MQ=37)
                  aaCTAACATCATAATCAAGGTCATCGCCGTTTCCGa              <  1:4078565/36-1 (MQ=255)
                  aaCTAACATCATAATCAAGGTCATCGCCGTTTCCGa              <  1:1319202/36-1 (MQ=255)
                   aCTAACATCATAATCAAGGTCATCGCCGTTTCCGaa             >  1:2592025/1-36 (MQ=255)
                   aCTAACATCATAATCAAGGTCATCGCCGTTTCCGaa             >  2:422670/1-36 (MQ=255)
                    cTAACATCATAATCAAGGTCATCGCCGTTTCCGaaa            <  2:3759532/36-1 (MQ=255)
                    cTAACATCATAATCAAGGTCATCGCCGTTTCCGaaa            <  1:2197499/36-1 (MQ=255)
                     tAACATCA AATCAA GTCATCGCCGTTTCCGaaaa           <  2:1187588/36-1 (MQ=21)
                      aaCATCATAATCAAGGTCATCGCCGTTTCCGaaaaa          <  2:2434349/36-1 (MQ=255)
                       aCATCATAATCAAGGTCATCGCCGTTTCCGAAAAAc         <  2:1690209/36-1 (MQ=255)
                        catcatAATCAAGGTCATCGCCGTTTCCGAAAAAcc        <  1:2225790/36-1 (MQ=255)
                        catcatAATCAAGGTCATCGCCGTTTCCGAAAAAcc        >  2:3173081/1-36 (MQ=255)
                        catcatAATCAAGGTCATCGCCGTTTCCGAAAAAcc        <  1:1341065/36-1 (MQ=255)
                         atcatAATCAAGGTCATC CCGTTTCCGAAAA CCg       >  2:270940/1-36 (MQ=25)
                          catAATCAAGGTCATCGCCGTTTCCGAAAAACCGGa      >  1:1008139/1-36 (MQ=255)
```

MC = Missing Coverage Evidence

breseq version 0.35.6 revision c7cf8df53bcd
mutation predictions | marginal predictions | summary statistics | genome diff | command line log

Predicted mutations

evidence	position	mutation	annotation	gene	description
RA	380,188	A→C	F239L (TTT→TTG)	*araJ* ←	
RA	475,292	+G	coding (14/1677 nt)	*ybaL* ←	
RA	649,391	T→A	I471F (ATC→TTC)	*mrdA* ←	
RA	683,496	A→C	V65G (GTT→GGT)	*nagC* ←	
JC JC	969,836	IS*150* (+) +3 bp	coding (810-812/2283 nt)	*pflB* ←	
RA	1,329,516	C→T	H33Y (CAC→TAC)	*topA* →	
JC JC	1,544,289	IS*150* (–) +3 bp	coding (150-152/1536 nt)	*xasA* ←	
JC JC	1,733,647	IS*150* (–) +3 bp	coding (683-685/1413 nt)	*pykF* →	
RA	1,976,879	T→G	intergenic (-57/-76)	*yedW* ← / → *yed*	
RA	2,082,685	G→A	A494V (GCT→GTT)	*yegI* ←	
RA	2,499,315	G→A	intergenic (-110/-179)	*maeB* ← / → *talA*	
RA	3,045,069	G→T	T312N (ACC→AAC)	*yghJ* ←	
RA	3,248,957	A→T	D764E (GAT→GAA)	*infB* ←	
MC JC	3,289,962	Δ16 bp	coding (96-111/4554 nt)	*gltB* →	
RA	3,339,158	A→C	intergenic (+22/-4)	*yhdG* → / → *fis*	
RA	3,370,027	T→A	K117M (AAG→ATG)	*rpsM* ←	
RA	3,424,910	G→A	M1M (ATG→ATA) †	*nirC* →	
RA	3,483,047	C→A	R455S (CGC→AGC)	*malT* →	
RA	3,762,741	A→T	K662I (AAA→ATA)	*spoT*	
RA	3,875,632	$(T)_{7\to8}$	intergenic (-66/+287)	*glmU* ← / ← *atpC*	
RA	3,893,551	+G	intergenic (+6/-50)	*kup* → / → *insJ-S*	
MC JC	3,894,997	Δ6,934 bp	IS*150*-mediated	*rbsD–[yieO]*	
RA	4,100,655	C→T	M192I (ATG→ATA)	*hslU* ←	
RA	4,126,706	$(T)_{8\to7}$	coding (342/879 nt)	*pflC* →	
RA	4,560,632	T→C	Y131C (TAC→TGC)	*hsdM* ←	

Predicted mutation

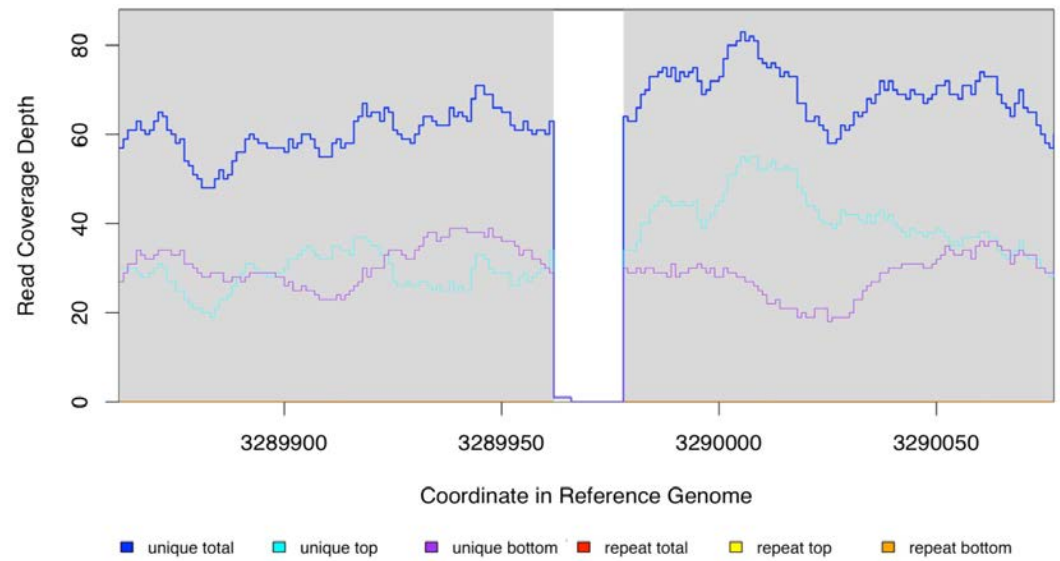
evidence	seq id	position	mutation	annotation	gene	description
MC JC	REL606	3,289,962	Δ16 bp	coding (96-111/4554 nt)	*gltB* →	glutamate synthase, large subunit

Missing coverage evidence...

			seq id	start	end	size	←reads	reads→	gene	description
*	*	÷	REL606	3289962	3289977	16	63 [1]	[0] 64	*gltB*	glutamate synthase, large subunit

New junction evidence

		seq id	position	reads (cov)	reads (cov)	score	skew	freq	annotation	gene	product
*	?	REL606	= 3289961	1 (0.020)	62 (1.050)	39/70	0.1	99.2%	coding (95/4554 nt)	*gltB*	glutamate synthase, large subunit
	?	REL606	3289978 =	0 (0.000)					coding (112/4554 nt)	*gltB*	glutamate synthase, large subunit



Unassigned missing coverage evidence

			seq id	start	end	size	←reads	reads→	gene	
*	*	÷	REL606	546953–547700	555934–555877	8178–8982	20 [18]	[16] 19	[insB-6]–[ECB_00513]	[in
*	*	÷	REL606	2031675–2031718	2054970–2054943	23226–23296	21 [17]	[18] 21	[manB]–[cpsG]	[m

Unassigned new junction evidence

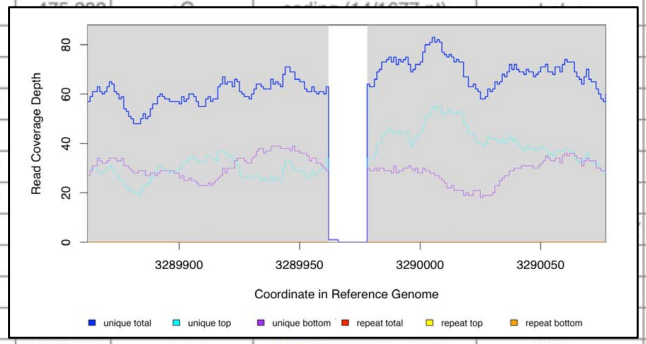
		seq id	position	reads (cov)	reads (cov)	score	skew	freq	annotation	gene	product
*	?	REL606	= 547699	NA (NA)	80 (1.360)	37/70	0.2	NA	noncoding (1/768 nt)	*IS1*	repeat region
	?	REL606	555924 =	NA (NA)					coding (1209/2346 nt)	*ECB_00513*	conserved hypothetical protein

JC = New Junction Evidence

breseq version 0.35.6 revision c7cf8df53bcd
mutation predictions | marginal predictions | summary statistics | genome diff | commar

Predicted mutations

evidence	position	mutation	annotation	gene	
RA	380,188	A→C	F239L (TTT→TTG)	*araJ* ←	predicted t
RA	[illegible]	[illegible]	[illegible]	[illegible]	predicted t
RA	[illegible]	[illegible]	[illegible]	[illegible]	transpeptic
RA	[illegible]	[illegible]	[illegible]	[illegible]	DNA-bindi
JC JC	[illegible]	[illegible]	[illegible]	[illegible]	pyruvate f
RA	[illegible]	[illegible]	[illegible]	[illegible]	DNA topoi
JC JC	[illegible]	[illegible]	[illegible]	[illegible]	predicted g
JC JC	[illegible]	[illegible]	[illegible]	[illegible]	pyruvate k
RA	[illegible]	[illegible]	[illegible]	[illegible]	predicted l
RA	[illegible]	[illegible]	[illegible]	[illegible]	hypothetic
RA	[illegible]	[illegible]	[illegible]	[illegible]	malic enzy
RA	[illegible]	[illegible]	[illegible]	[illegible]	predicted i
RA	3,248,957	A→T	D764E (GAT→GAA)	*infB* ←	translation
MC JC	[illegible]	Δ16 bp	coding (96-111/4554 nt)	*gltB* →	glutamate
RA	3,339,158	A→C	intergenic (+22/-4)	*yhdG* → / → *fis*	A-dihy
RA	3,370,027	T→A	K117M (AAG→ATG)	*rpsM* ←	30S ribosc
RA	3,424,910	G→A	M1M (ATG→ATA) †	*nirC* →	nitrite tran
RA	3,483,047	C→A	R455S (CGC→AGC)	*malT* →	transcriptic
RA	3,762,741	A→T	K662I (AAA→ATA)	*spoT* →	bifunctiona
RA	3,875,632	$(T)_{7\to8}$	intergenic (-66/+287)	*glmU* ← / ← *atpC*	bifunctiona subunit ep
RA	3,893,551	+G	intergenic (+6/-50)	*kup* → / → *insJ-5*	potassium
MC JC	3,894,997	Δ6,934 bp	IS*150*-mediated	*rbsD–[yieO]*	*rbsD, rbsA*
RA	4,100,655	C→T	M192I (ATG→ATA)	*hslU* ←	ATP-deper
RA	4,126,706	$(T)_{8\to7}$	coding (342/879 nt)	*pflC* →	pyruvate f
RA	4,560,632	T→C	Y131C (TAC→TGC)	*hsdM* ←	DNA meth



Unassigned missing coverage evidence

			seq id	start	end	size	←reads	reads→	gene	
*	*	÷	REL606	546953–547700	555934–555877	8178–8982	20 [18]	[16] 19	*[insB-6]–[ECB_00513]*	[insB-6],insA-6,
*	*	÷	REL606	2031675–2031718	2054970–2054943	23226–23296	21 [17]	[18] 21	*[manB]–[cpsG]*	[manB],manC,i

Unassigned new junction evidence

		seq id	position	reads (cov)	reads (cov)	score	skew	freq	annotation	
*	?	REL606	= 547699	NA (NA)	80 (1.360)	37/70	0.2	NA	noncoding (1/768 nt)	
	?	REL606	555924 =	NA (NA)					coding (1209/2346 nt)	

Predicted mutation

evidence	seq id	position	mutation	annotation	gene	description
MC JC	REL606	3,289,962	Δ16 bp	coding (96-111/4554 nt)	*gltB* →	glutamate synthase, large subunit

Missing coverage evidence...

			seq id	start	end	size	←reads	reads→	gene	description
*	*	÷	REL606	3289962	3289977	16	63 [1]	[0] 64	*gltB*	glutamate synthase, large subunit

New junction evidence

		seq id	position	reads (cov)	reads (cov)	score	skew	freq	annotation	gene	product
*	?	REL606	= 3289961	1 (0.020)	62 (1.050)	39/70	0.1	99.2%	coding (95/4554 nt)	*gltB*	glutamate synthase, large subunit
	?	REL606	3289978 =	0 (0.000)					coding (112/4554 nt)	*gltB*	glutamate synthase, large subunit

```
GGGTTCCCGCAGAGCCTGGGGGAGGTTCACGATAT------------------------------------  >  REL606/3289927-3289961
-----------------------------------CTTGAGAGGGATAACTGTGGTTTCGGCCTGATCGC  >  REL606/3289978-3290012

GGGTTCCCGCAGGGCCTGGGGGAGGTTCACGATATC                                     <  2:3623510/36-1
GGGTTCCCGCAGAGCCTGGGGGAGGTTCACGATATC                                     >  2:3380704/1-36
 GGTTCCCGCAGAGCCTGGGGGAGGTTCACGATATCT                                    >  2:3460319/1-36
 GGTTCCCGCAGAGCCTGGGGGAGGTTCACGATATCT                                    >  1:131137/1-36
   TTCCCGC GAGCCTGGGGGAGGTTCACGATATCTTG                                  <  1:1353443/36-1
   TTCCCGCAGAGCCTGGGGGAGGTTCACGATATCTTG                                  <  1:3014326/36-1
    TCCCGCAGAGCCTGGGGGAGGTTCACGAT TCTTGA                                 >  1:3346641/1-36
    TCCCGCAGAGCCTGGGGGAGGGTCACGATATCTTGA                                 >  1:689686/1-36
    TCCCGCAGAGCCAGGGGGAGGTTCACGATATCTTGA                                 <  2:657798/36-1
     CCCGCAGAGCCTGGGGGAGGTTCACGATATCTTGAG                                <  1:2511982/36-1
      CCGCAGAGCCTGGGGGAGGTTCACGATATCTTGAGA                               <  1:2025880/36-1
      CCGCAGAGCCTGGGGGAGGTTCACGATATCTTGAGA                               <  2:3369418/36-1
        GCAGAGCCTGGGGGAGGTTCACGATATCTTGAGAGG                             >  1:2431606/1-36
         CAGAGCCTGGGGGAGGTTCACGATATCTTGAGAGGG                            >  2:2593722/1-36
         CAGAGCCTGGGGGAGGTTCACGATATCTTGAGAGGG                            <  2:2252332/36-1
          AGAGCCTGGGGGAGGTTCACGATATCTTGAGAGGGA                           <  2:892776/36-1
          AGAGCCTGGGGGAGCTTCACGATAT TTGAGAGGGA                           >  1:2634884/1-36
           GAG CTGGGGGAGGTTCACGATATCTTGAGAGGGAT                          <  2:55338/36-1
           GAG CTGGGGGAGGTTCACGATATCTTGAGAGGGAT                          <  2:2034737/36-1
           GAGCCTGGGGGAGGTTCACGGTATCTTGAGAGGGAT                          >  2:16872/1-36
           GAGCCTGGGGGAGGTTCACGATATCTTGAGAGGGAT                          >  2:223924/1-36
           GAGCCTGGGGGAGGTTCACGATATCTTGAGAGGGAT                          <  1:187041/36-1
             GCCTGGGGGAGGTTCACGATATCTTGAGAGGGATAA                        <  1:1339127/36-1
             GCCTGGGGGAGGTTCACGATATCTTGAGAGGGATAA                        >  2:2339905/1-36
              CCTGGGGGAGGTTCACGATATCTTGAGAGGGATAAC                       >  2:2339574/1-36
                TGGTGGAGGTTCACGATATCTTGAGAGGGATAACTG                     >  1:1082563/1-36
                TGGGGGAGGTTCACGATAT TTGAGAGGGATA CTG                     >  1:2546169/1-36
                TGGGGGAGGTTCACGATATCTTGAGATGGATAACTG                     >  1:1973857/1-36
                TGGGGGAGGTTCACGATATCTTGAGAGGGATAACTG                     >  2:1448864/1-36
                TGGGGGAGGTTCACGATATCTTGAGAGGGATAACTG                     >  1:732312/1-36
                 GGGGGAGGTTCACGATATCTTGAGAGGGATAACTGT                    <  2:2707039/36-1
                 GGGGGAGGTTCACGATATCTTGAGAGGGATAACTGT                    >  1:2303795/1-36
                 GGGGGAGGTTCACGATATCTTGAGAGGGATAACTGT                    >  1:2120882/1-36
                 GGGGGAGGTTCACGATATCTTGAGAGGGATAACTGT                    >  1:372974/1-36
                  GGGGAGGTTCACGATATCTTGAGAGGGATAACTGTG                   <  2:2173522/36-1
                    GGAGGTTCACGATATCTTGAGAGGGATAACTGTGGT                 <  2:2610370/36-1
                    GGAGGTTCACGATATCTTGAGAGGGATAACTGTGGT                 <  2:2779388/36-1
                    GGAGGTTCACGATATCTTGAGAGGGATAACTGTGGT                 >  2:2438260/1-36
```

Summary Statistics

breseq version 0.35.6 revision c7cf8df53bcd
mutation predictions | marginal predictions | summary statistics | genome diff | command line log

Read File Information

	read file	reads	bases	passed filters	average	longest	mapped
errors	SRR030255_1	4,092,676	147,336,336	98.7%	36.0 bases	36 bases	95.3%
errors	SRR030255_2	4,103,100	147,711,600	98.9%	36.0 bases	36 bases	93.9%
	total	**8,195,776**	**295,047,936**	98.8%	36.0 bases	**36** bases	94.6%

Reference Sequence Information

		seq id	length	fit mean	fit dispersion	% mapped reads	description
coverage	distribution	REL606	4,629,812	60.6	3.1	100.0%	Escherichia coli strain REL606.
		total	**4,629,812**			100.0%	

fit dispersion is the ratio of the variance to the mean for the negative binomial fit. It is =1 for Poisson and >1 for over-dispersed data.

New Junction Evidence

Junction Candidates Tested

option	limit	actual
Number of alignment pairs examined for constructing junction candidates	≤ 100000	100047
Coverage evenness (position-hash) score of junction candidates	≥ 2	≥ 2
Test this many junction candidates (n). May be smaller if not enough passed the coverage evenness threshold	100 ≤ n ≤ 5000	60
Total length of all junction candidates (factor times the reference genome length)	≤ 0.1	0.001

Junction Skew Score Calculation

reference sequence	pr(no read start)
REL606	0.48689

Reference Sequence Coverage

breseq version 0.35.6 revision c7cf8df53bcd
mutation predictions | marginal predictions | summary statistics | genome diff | commar

Read File Information

	read file	reads	bases	passed filters	average	longest	ma
errors	SRR030255_1	4,092,676	147,336,336	98.7%	36.0 bases	36 bases	9
errors	SRR030255_2	4,103,100	147,711,600	98.9%	36.0 bases	36 bases	9
	total	8,195,776	295,047,936	98.8%	36.0 bases	36 bases	9

Reference Sequence Information

		seq id	length	fit mean	fit dispersion	% mapped reads	des
coverage	distribution	REL606	4,629,812	60.6	3.1	100.0%	Esc
		total	4,629,812			100.0%	

fit dispersion is the ratio of the variance to the mean for the negative binomial fit. It is =1 for Poisson and >1 for over-dispersed data.

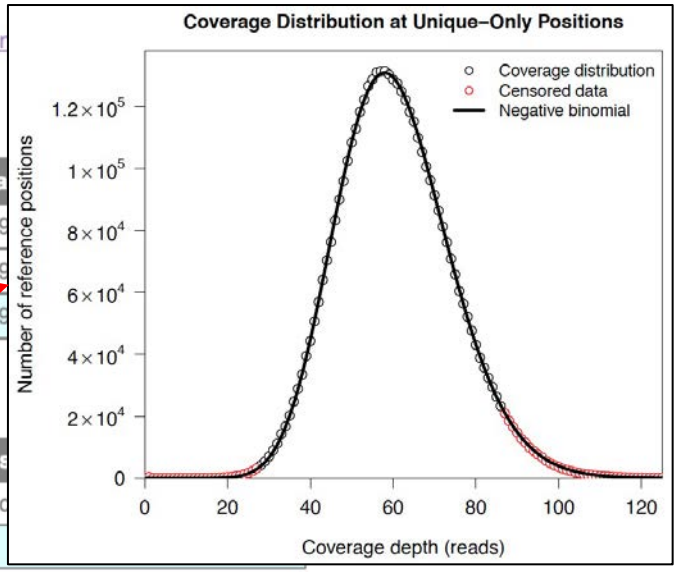
New Junction Evidence

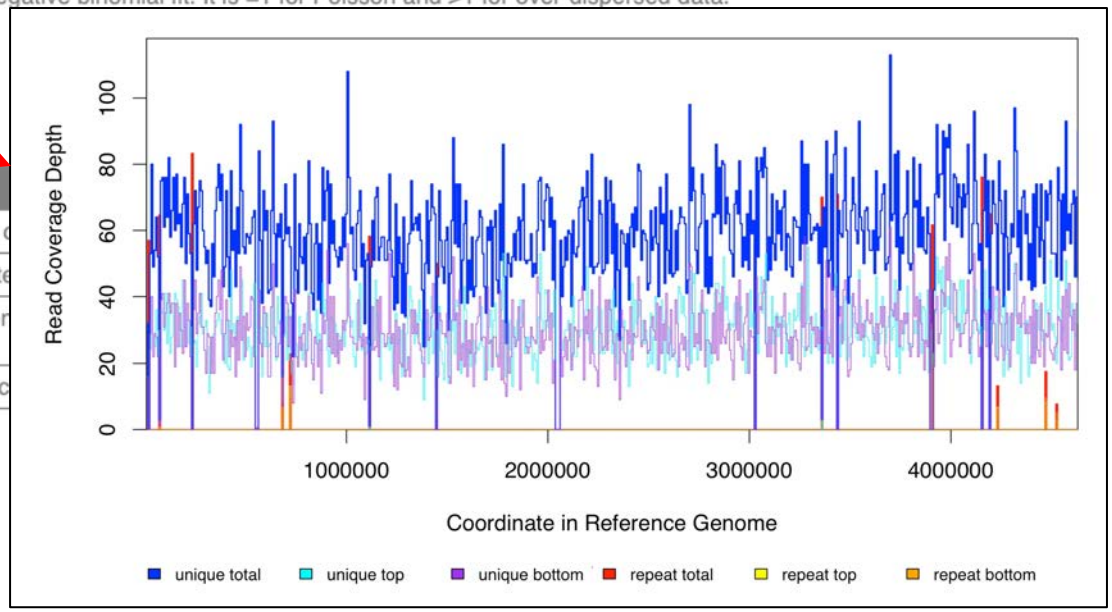
Junction Candidates Tested

option
Number of alignment pairs examined for constructing junction c
Coverage evenness (position-hash) score of junction candidate
Test this many junction candidates (n). May be smaller if not er threshold
Total length of all junction candidates (factor times the referenc

Junction Skew Score Calculation

reference sequence	pr(no read start)
REL606	0.48689

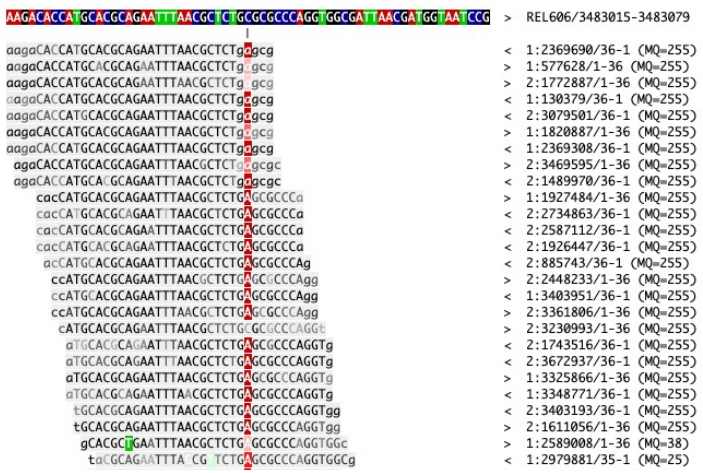




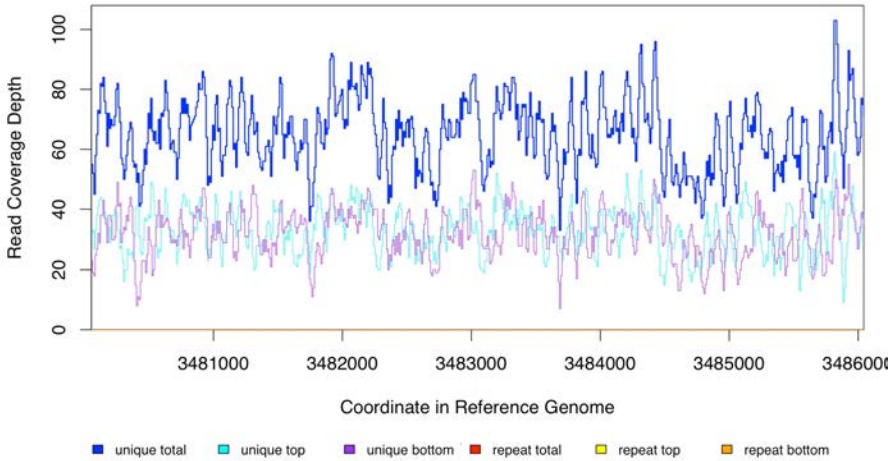
Utilities to explore output

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.

Explore aligned reads using IGV

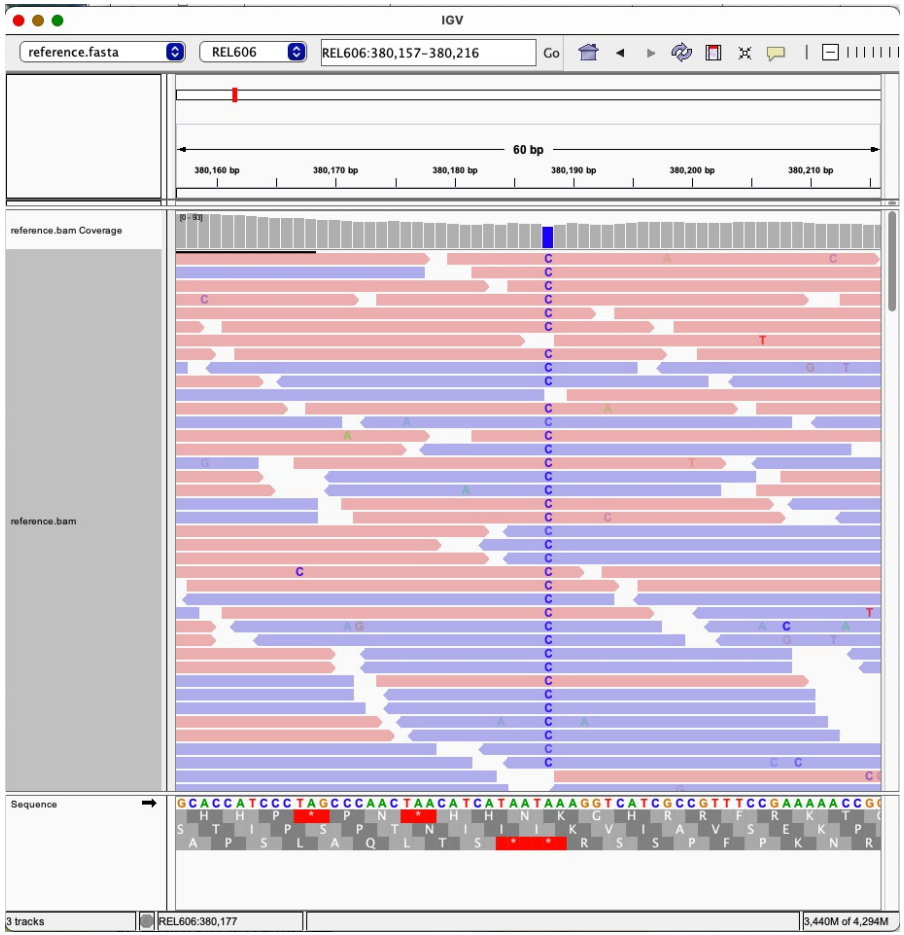
Integrative Genomics Viewer

https://software.broadinstitute.org/software/igv/

Viewing Output / Aligned Reads in the IGV

You can visualize the "raw data" (how **breseq** aligned reads to the reference genome) using the Integrative Genomics Viewer (IGV) and files located in the `data` folder created by **breseq**.

1. Install and open IGV
2. Import the reference genome sequence:
   - Click 'File', and then 'Import Genome…'
   - Fill out the requested information: 'ID', 'Name'
   - Choose the FASTA file: `data/reference.fasta`.
   - The other fields are optional.
3. Import the reference genome feature information:
   - Click 'File', and then 'Load from File…"
   - Choose the GFF3 file: `data/reference.gff3`.
4. Import the read alignments to the reference genome:
   - Click 'File', and then 'Load from File…"
   - Choose the BAM file: `data/reference.bam`.



GenomeDiff output

Machine-readable text files for further processing

```
#=GENOME_DIFF    1.0
#=CREATED   15:16:00 24 May 2021
#=PROGRAM   breseq 0.35.6 revision c7cf8df53bcd
#=COMMAND   breseq -j 8 -o tests/long_Ara-1_10000gen_4536A ...
#=REFSEQ    tests/long_Ara-1_10000gen_4536A/../data/long_tests/REL606.gbk
#=READSEQ   tests/long_Ara-1_10000gen_4536A/../data/long_tests/SRR030255_1.fastq.gz
#=READSEQ   tests/long_Ara-1_10000gen_4536A/../data/long_tests/SRR030255_2.fastq.gz
#=CONVERTED-BASES 295047936
#=CONVERTED-READS 8195776
#=INPUT-BASES    298701576
#=INPUT-READS    8297266
#=MAPPED-BASES   277772336
#=MAPPED-READS   7750270
SNP   1     29    REL606      380188      C
INS   2     32    REL606      475292      G
SNP   3     36    REL606      649391      A
SNP   4     37    REL606      683496      C
MOB   5     101,102     REL606      969836      IS150 1     3
SNP   6     41    REL606      1329516     T
MOB   7     103,109     REL606      1544289     IS150 -1    3
MOB   8     110,111     REL606      1733647     IS150 -1    3
SNP   9     46    REL606      1976879     G
SNP   10    49    REL606      2082685     A
...
```

GenomeDiff format

`output/output.gd`

Format specification provided in the *breseq* manual

What can you do with a GenomeDiff?

Generate an HTML table comparing multiple clones/populations:

```
$ gdtools COMPARE -o compare.html -r reference.gbk input1.gd input2.gd ...
```

Convert to TSV, VCF or other formats for interchange with other programs:

```
$ gdtools ANNOTATE -o -f TSV -r reference.gbk input1.gd input2.gd ...
```

Count mutations and numbers of sites at risk for mutations:

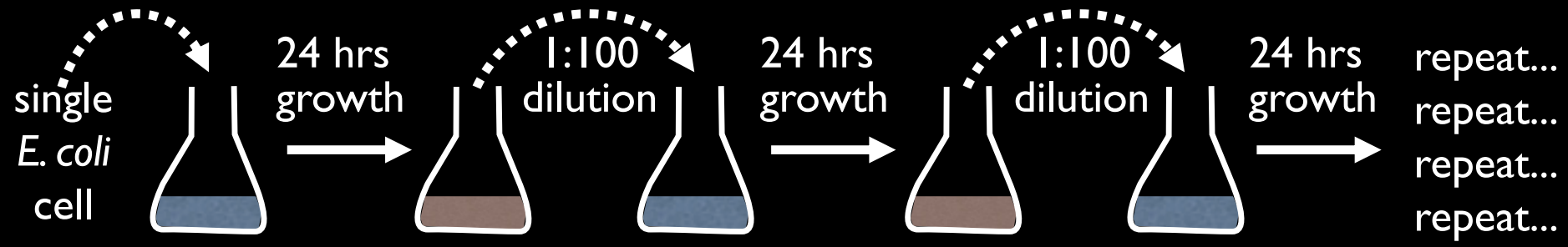
```
$ gdtools COUNT -o output.csv -r reference.gbk input1.gd input2.gd ...
```

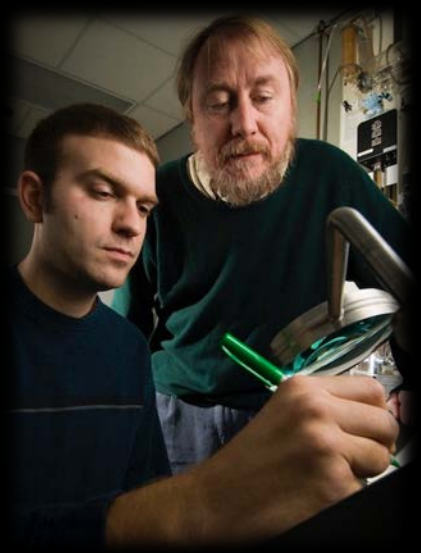
Apply the mutations to generate an updated reference sequence:

```
$ gdtools APPLY -f GENBANK -o updated.gbk -r reference.gbk input.gd
```

And more… `$ gdtools --help`

Lenski Long-Term Evolution Experiment





Richard Lenski
Michigan State

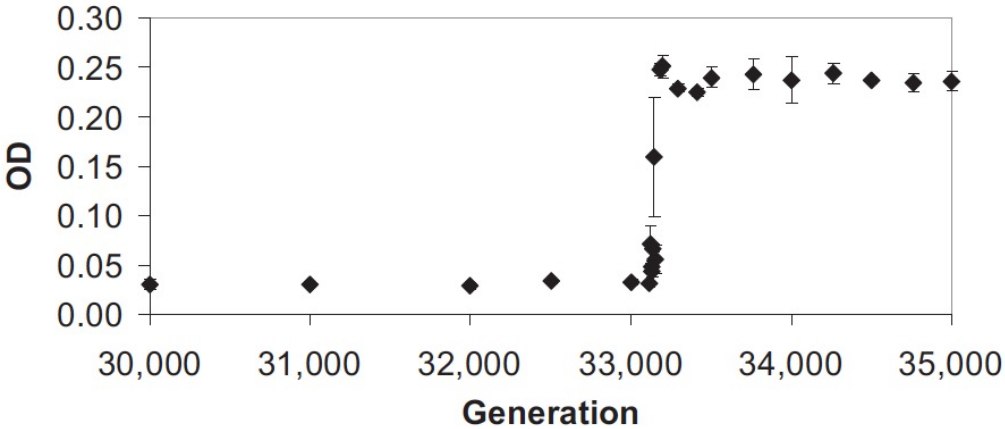
- 12 independent populations
- Deep evolutionary history
- Viable frozen "fossil record"

>73,000 generations of
E. coli growth (>30 years)!

Analysis: Causative Mutations

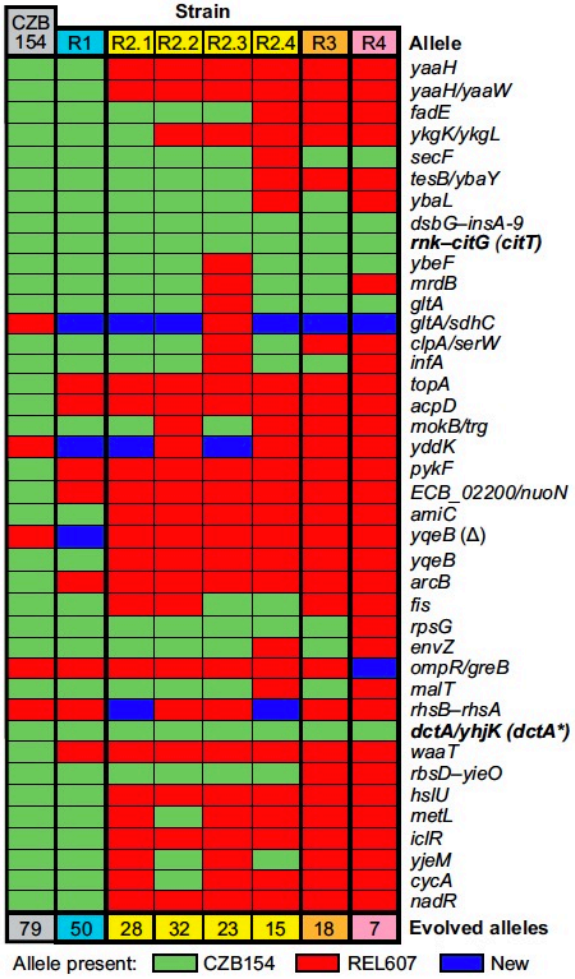
How?

Backcross and sequence: Only **two mutations** required for strong Cit+ phenotype



Citrate utilization evolved after 33,000 generations in one LTEE population

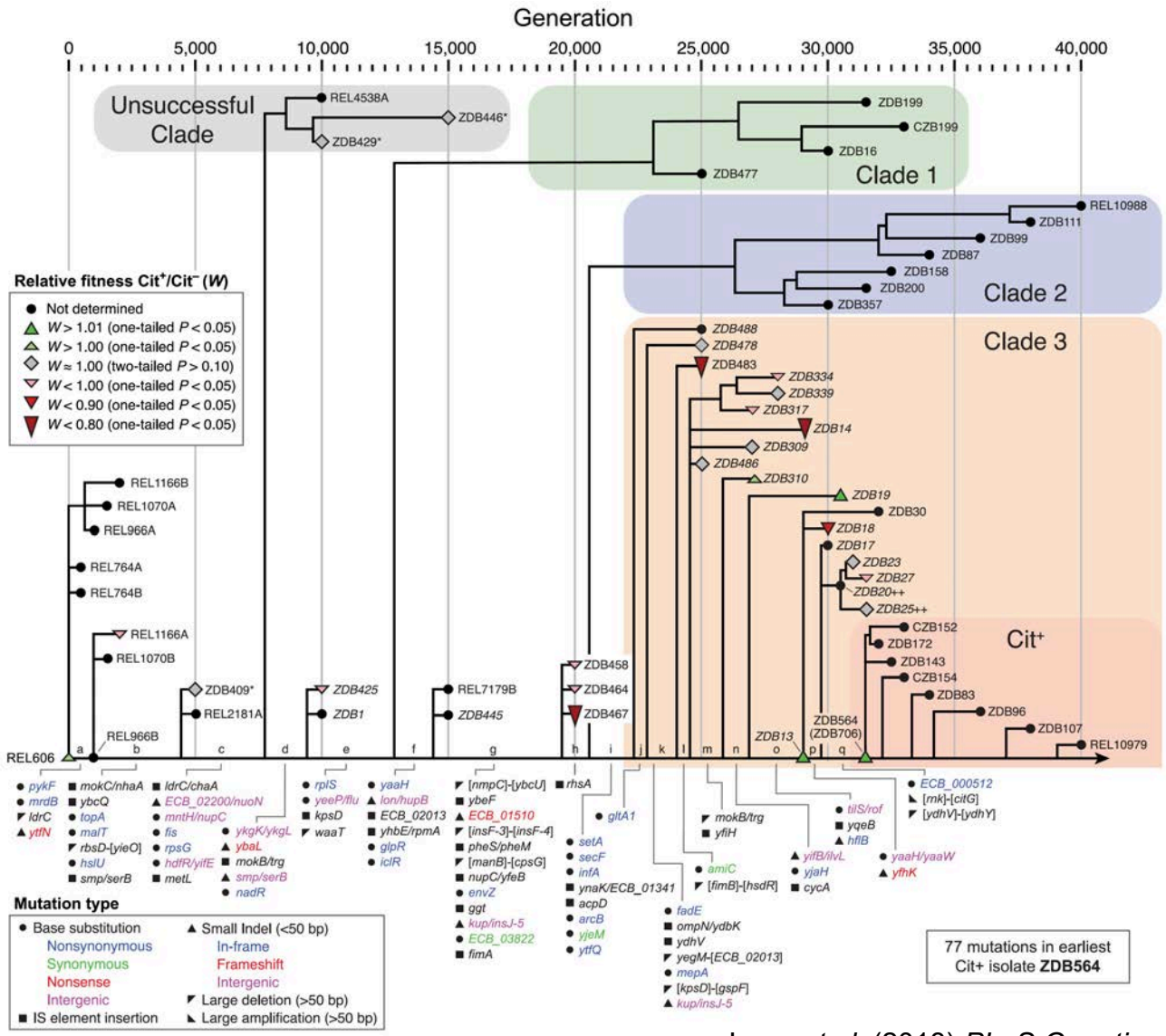
Blount *et al.* (2008) *PNAS*




```
$ gdtools COMPARE...
```

Quandt *et al.* (2014) *PNAS*

Analysis: Phylogenetic trees



Leon *et al.* (2018) *PLoS Genetics*

What mutations led to Cit+ evolution?

Generate an alignment of genomic changes

```
$ gdtools COMPARE
  -f PHYLIP clone1.gd
  clone2.gd ....
```

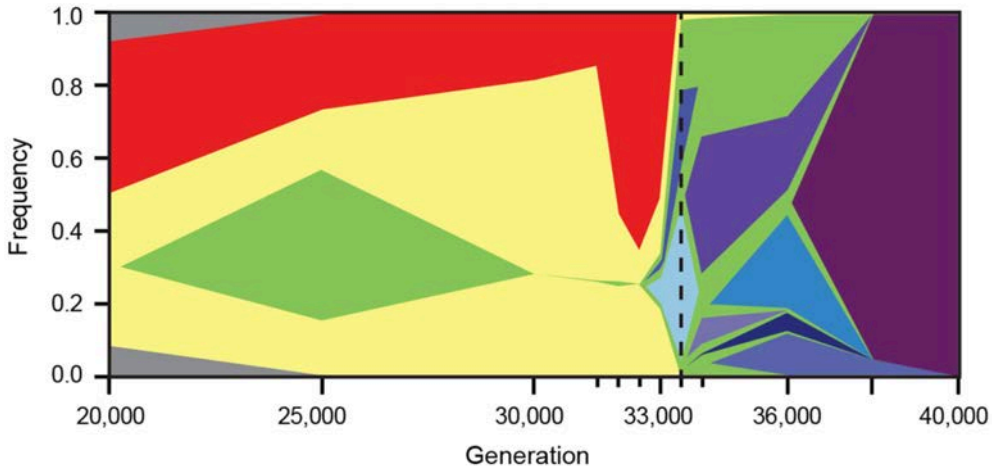
or

```
$ gdtools COMPARE
  -f FASTA clone1.gd
  clone2.gd ....
```

Build and visualize a maximum parsimony tree using PHYLIP, MEGAX, etc.

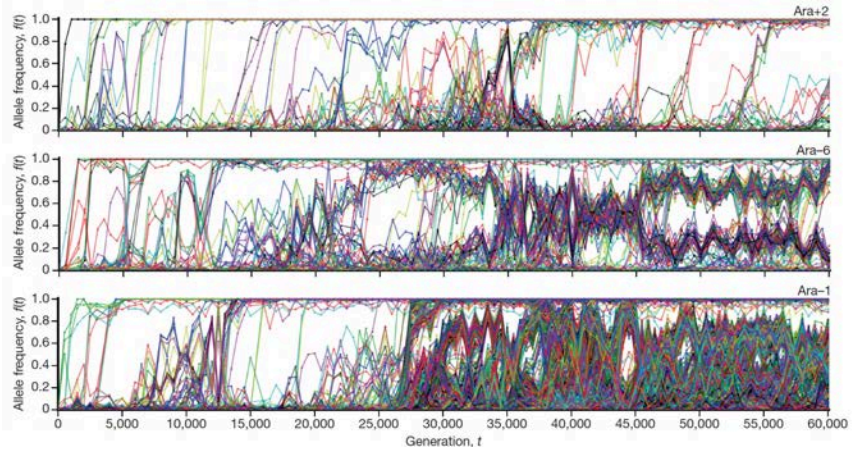
Analysis: Allele/Genotype Frequencies

Muller Plot (Genotype Frequency)



Quandt *et al.* (2015) *eLife*

Allele Frequency



Good *et al.* (2017) *Nature*

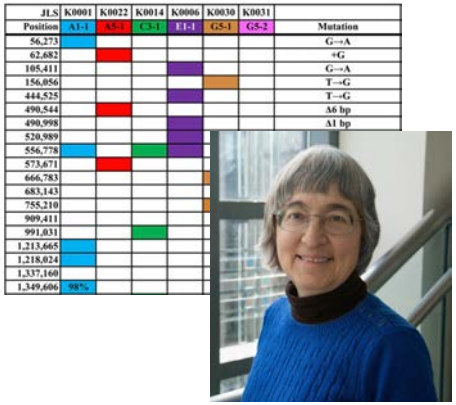
For tracking how genetic diversity evolves within populations, visualizing dynamics, selective sweeps, and stable coexistence.

```
gdtools COMPARE -f CSV pop1.gd pop2.gd ....
```

Programs/packages that can help:

R, ggplot, ggMuller, EvoFreq, MullerPlot

Workshop Presentations

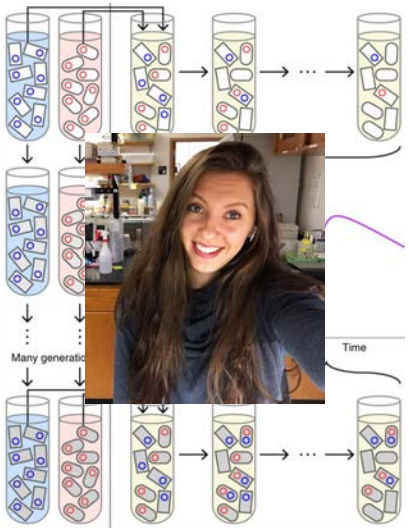


JLS	K0001	K0022	K0014	K0006	K0030	K0031	
Position	A1-1	A5-1	C3-1	E1-1	G5-1	G5-2	Mutation
56,273							G→A
62,682							+G
105,411							G→A
156,056							T→G
444,525							T→G
490,544							Δ6 bp
490,998							Δ1 bp
520,989							
556,778							
573,671							
666,783							
683,143							
755,210							
909,411							
991,031							
1,213,665							
1,218,024							
1,337,160							
1,349,606	98%						

Antibiotic Resistance Reversal: *breseq* Analysis of Experimental Evolution, Compared with FACS Competition Assays of Relative Fitness

Joan Slonczewski

Kenyon College



Identifying Adaptive Paths in Host-Plasmid Coevolution Using *breseq*

Olivia Kosterlitz

University of Washington

Workshop Presentations

EVOLVINGSTEM

Decoding Evolution-In-Action in Classroom Experiments That Simulate Infection Biology Using *breseq*

Vaughn Cooper

University of Pittsburgh

ALEdb

ALEdb: A Living High-Quality Database of Mutations from Adaptive Evolution Experiments Powered by *breseq*

Adam Feist

University of California, San Diego

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.

Table of Contents

Previous topic

Tutorial: Clonal Samples (Consensus Mode)

Next topic

Tutorial: Ultra-rare variant detection using consensus reads and targeted sequencing

This Page

Show Source

Quick search

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 **gdtools** 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 **gdtools** errors or crashes in running this tutorial, please report issues on GitHub.

Table of Contents

1. Download data files

First, create a directory called `tutorial_clonal`:

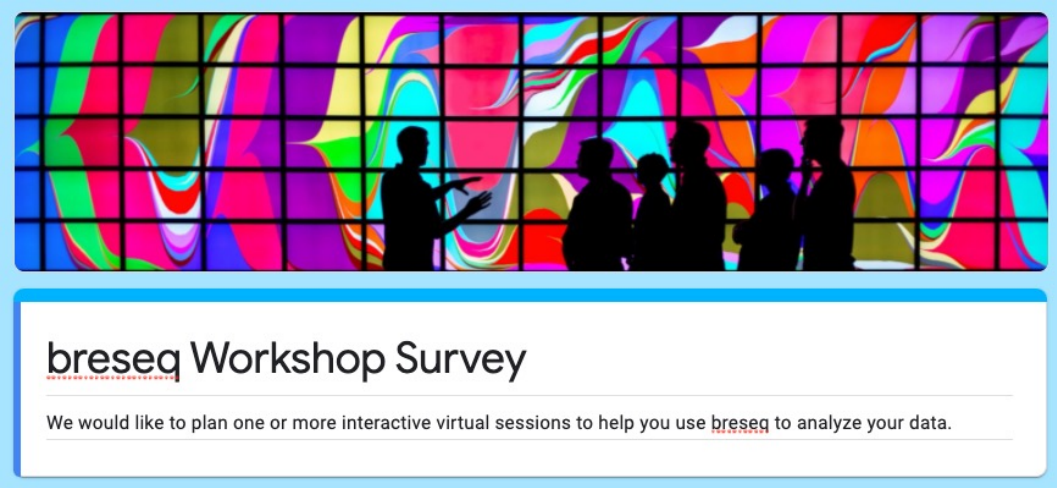
```
$ mkdir tutorial_clones
$ cd tutorial_clones
```

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



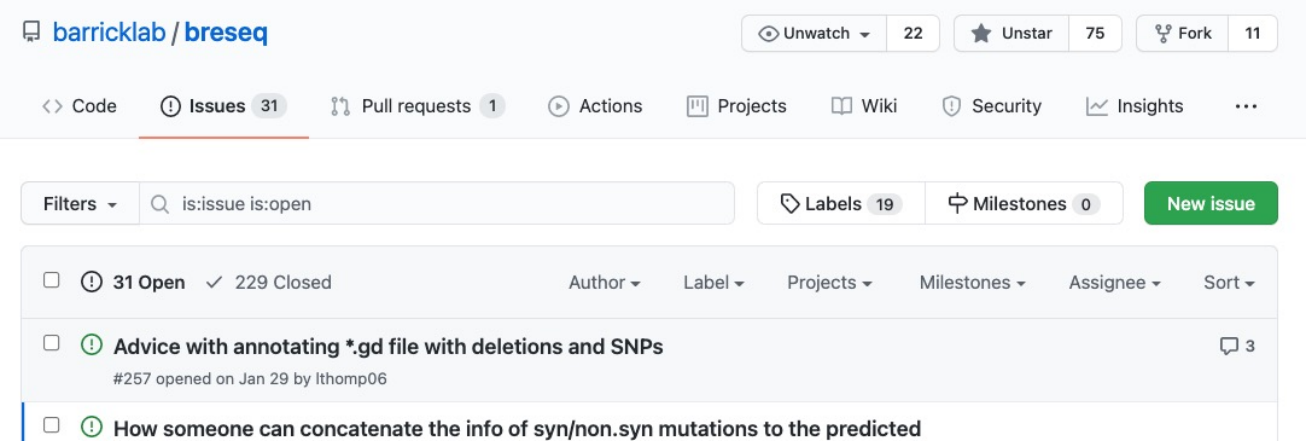
https://forms.gle/qkvkjbqCXZAhY7GW6

Interactive Workshop

- Install on your system
- Use on your data
- Help interpret output
- Provide advice on further analysis

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.