Variant-calling

New strain, who dis…

Outline of presentation and workshop

Slideshow

1. Overview of the Breseq program
2. Evidence for mutations
3. Breseq's interpretation of evidence into mutations

Hands-on workshop

1. Bash basics
2. Trial run of Breseq using virtual env

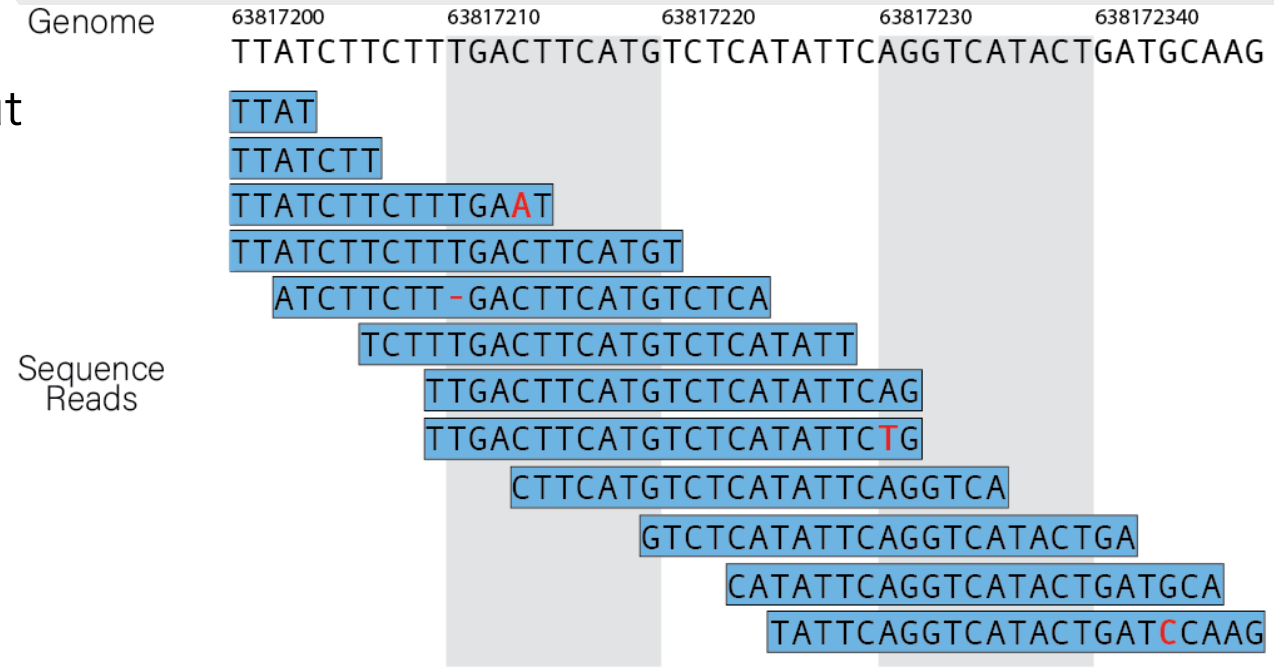
Introduction to Breseq

Short-read (Illumina variant-calling).

Tested mostly with bacteria but has been known to work with simple eukaryotic genomes (e.g., fungi).

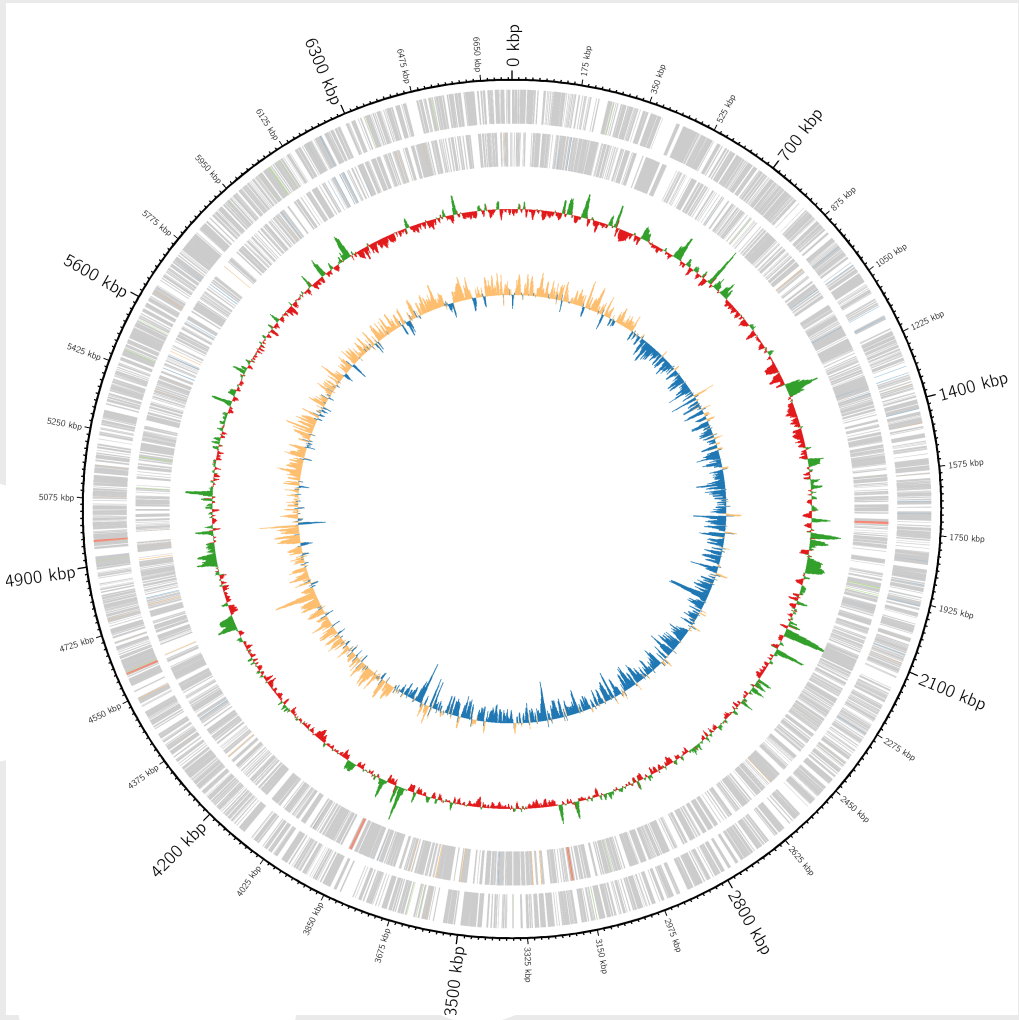
Output is designed to be human-readable, but difficult to parse.

Relies mostly on Bowtie2: mapping software



Breseq takes two inputs: reads (FASTQ) and a reference genome

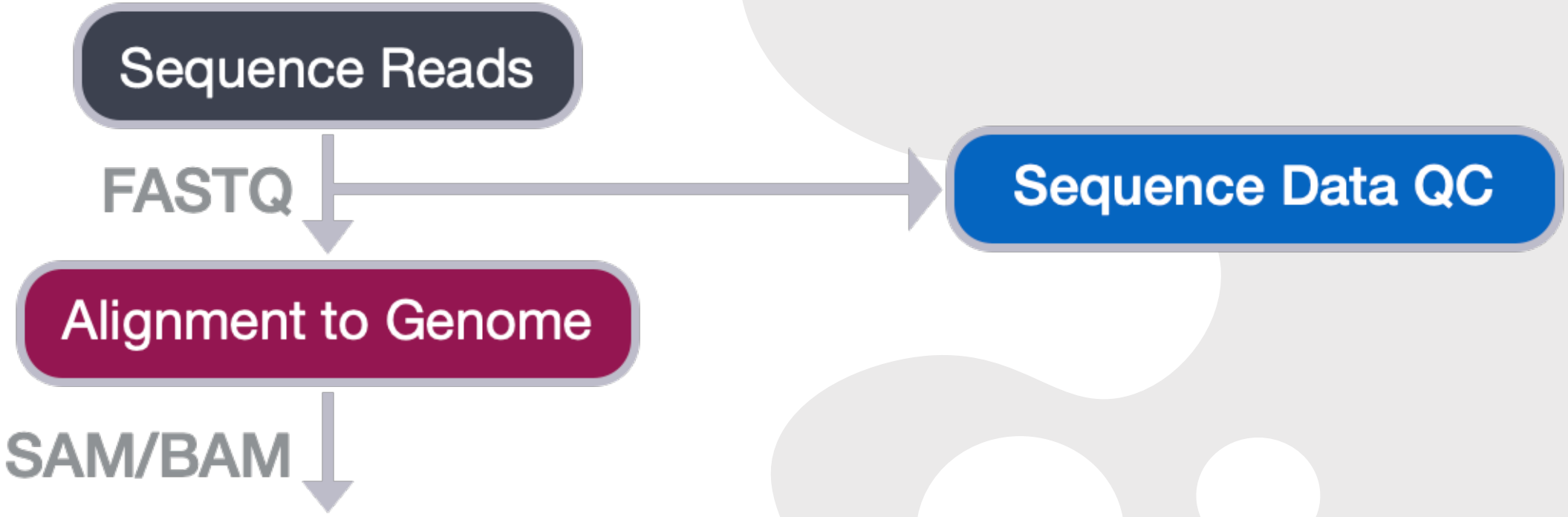
```
@M04743:199:000000000-CGG4F:1:1101:16145:1655 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGAAGGTGGCAAGCGTTGTTCGGATTCACTGGGCGTACAGGGAGCGTAGGCGGTTGGGTAAGCCCTCCGTGAAATCTCCGGG
+
ABCCCFFFCADBGGGGGGGGGGHHGHGGFHGHHHGHGGGAFFHGGGGGHHHHHHHHGGGGGHHGGGGGGGGGHGGEGGGGGHHHHHHHHGGHGGGHHHHHGGG
@M04743:199:000000000-CGG4F:1:1101:18938:1729 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGCGCAGGCTGTTTTGTAAGTCAGATGTGAAATCCCCGAG
+
BBBBBFFFBBBBGGGGGGGGGFHHHHHGGHGGGGGGGGGGGHHGGEGFHHHHHHHHGGGGHFHGGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHGGGG
@M04743:199:000000000-CGG4F:1:1101:13893:1760 1:N:0:233
GGTGCCAGCAGCCGCGGTACTACGTAGGGTGCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGTTCGTAGGCGGTTTGTCGCGTCGTTTGTGAAAACCCGGGG
+
BBBBBFFFB4CCGGGGGGGGCFFHGHHHGGHGGGGGGGGGAFGHGG?EFHFEHHHHHGGGGFHFHFGHGGHGG3EEEGGGHHEHGGGGGGDHHEHGHGGGGGG
F9FFFFFFFFFFFEFFBBBBFEB;-@DFB-BBBFFFFEFF/EBBEFFF/BADFFDFFF.;
@M04743:199:000000000-CGG4F:1:1101:14830:1795 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGTAGGTGGCAAGCGTTGTCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGTTCTTTAAGTCAGTGGTGAAATACAGCCG
+
ABBABFBFB?AAEE?EGEFCGGHHFFHGEHFFHHGHGGGCFHHGEEGGDFGDHHHGGGFGDGHGGFEGFGGDFGGGGGHHFFFBGFH34FGBFFHGHHHGHFFC
9BD?99-9/9@-BD.;ADFFFBF///BBF:FFFFFFED?DFDFF?A.
@M04743:199:000000000-CGG4F:1:1101:14968:1984 1:N:0:233
AGTGCCAGCCGCCGCGGTAATACGTAGGTGGCAAGCGTTGTCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGTTCTTTAAGTCAGTGGTGAAATACAGCCG
+
BBBBBFFFBABBGGGGGGGGGGHHGFHGHHGHHHGHGGGCFHHGGEGGHHHHHHGGGGHHHHHGHGGGGGGGHGGGGGHHHHHHHHHHHHHHGFFHHHHHHHHG
FCHHHGGHHHHHHHHHHHHHHHHHHHHHFHHHHGFHHGEGGFHHGHGGGFEGG9FGGAEGGGGAFDGEFFGGFFFBFEFFFFFFFFFFFFFFFFFF>DFDBFFI
@M04743:199:000000000-CGG4F:1:1101:12706:2099 1:N:0:233
TGTGCCAGCCGCCGCGGTAATACGGAGGGAGCTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGTTTTTTAAGTCAGAGGTGAAAGCCCGGGG
+
BCCCCFFFCCCCGGGGGGGGGGHHEGGGGDFGGHHHGGGGGHHGGGGFHHGHHHHHGGGGHHHGGGGGHHHGHGGGGGGGGACGHHHHHHGHHGHHFHHHGGG
BFFFFFFFF9FFFFFFFFFFFFFF/
@M04743:199:000000000-CGG4F:1:1101:13747:2260 1:N:0:233
CGTGCCAGCCGCCGCGGTAATACGAAGGGGGCTAGCGTTGTTCGGAATTACTGGGCGTAAAGAGTTCGTAGGCGGTTTGTCGCGTCGTTTGTGAAAACCCGGGG
+
CCCCCCFFCABCGGGGGGGGGGGHHFCEGDGGGGHHHGGGEFHHGGGFFHHFHHHHGGGGGHH@GHHHGGHGGHGGGGGGFH</>CFCGGGGHHHHHFHGGGGGG
A@@FFFFFFFFFFFFFBF9C;=CF.@;CDFFFFFBDFFFFFF?BEFFFFFFFFFFFFFFFFF?
@M04743:199:000000000-CGG4F:1:1101:20151:2263 1:N:0:233
TGTGCCAGCCGCCGCGGTAATACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGCGCAGGCTGTTTTGTAAGTCAGATGTGAAATCCCCGAG
+
BBBBBFFFBAADGGGGGGGGGGHHHHHGGHGGGGGGGGGGGHHGGDFFHHHHHHHHGGGGGHGHGGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHGGFG
@M04743:199:000000000-CGG4F:1:1101:17232:2363 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGGAGGGGGCTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGATCGGAAAGTCAGAGGTGAAATCCCAGGG
+
BBBBBFFFBBBBGGGGGGGGGGHHGDGGGGGGGHHHGGG0FGHGGEGFHHHHHHHHGGGGHHHGGGGGHHHGHGGGGGGGGGHHHHHHHGHHHGHHHHGHGG
```

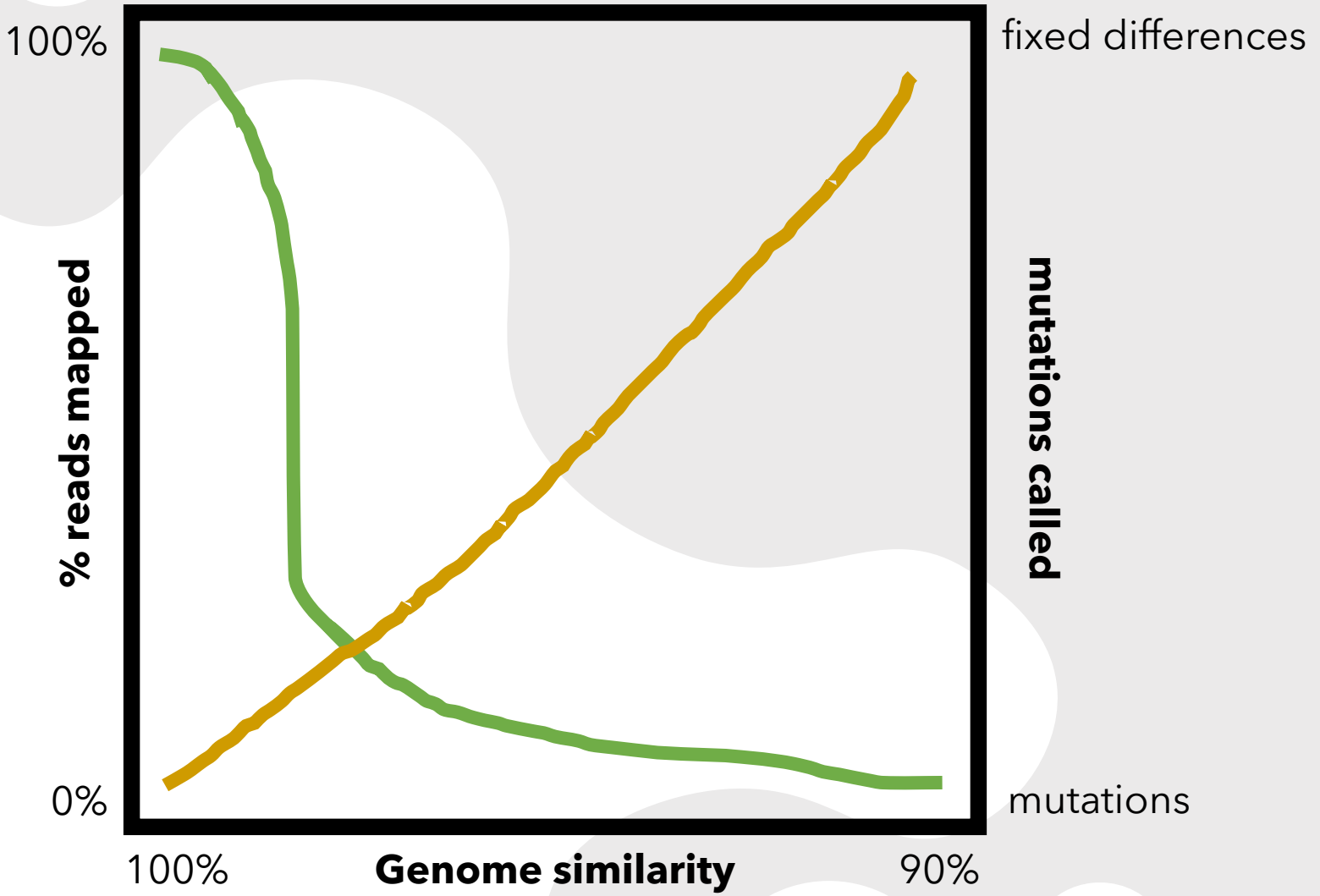


Breseq interprets mapping information from Bowtie2

Breseq generally requires strain-level similarity between the reference genome and input sequenced reads.

What happens when a more-diverged genome is used?



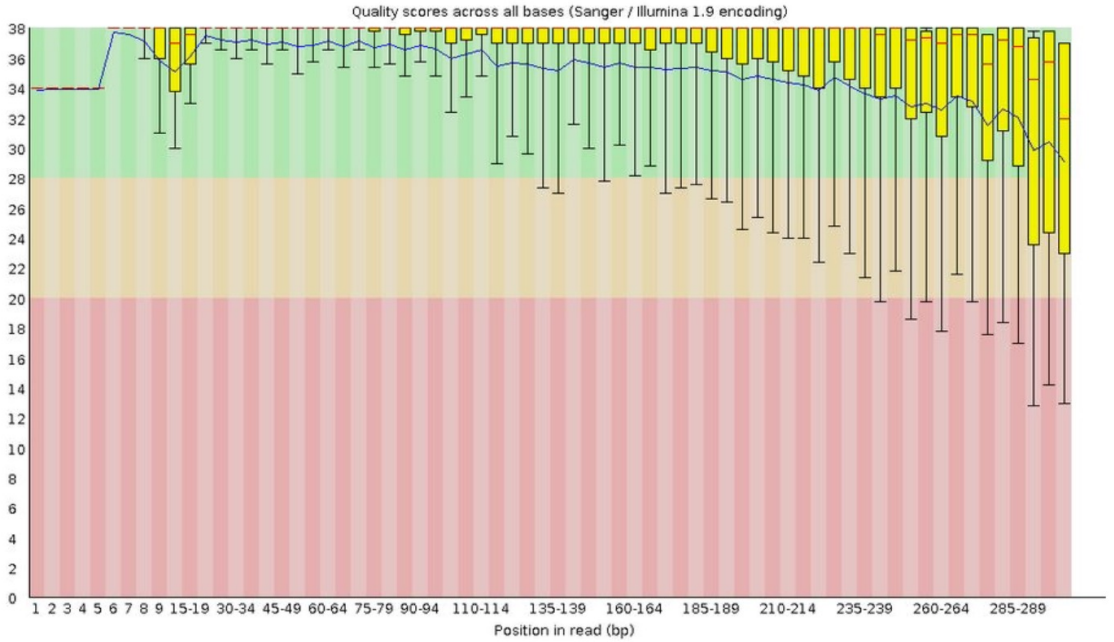

100%
fixed differences
% reads mapped
mutations called
0%
mutations
100%
Genome similarity
90%


PA14_RS21720
PA14_RS21850
PA14_RS22670
PA14_RS22840
PA14_RS23350
PA14_RS24070
PA14_RS25330
PA14_RS26415
PA14_RS26650
PA14_RS26890
PA14_RS27055
PA14_RS27265
PA14_RS27775
1
2
3
4
5
6
7
8
9
Sample #
mutation
insertion
none
transition
transversion

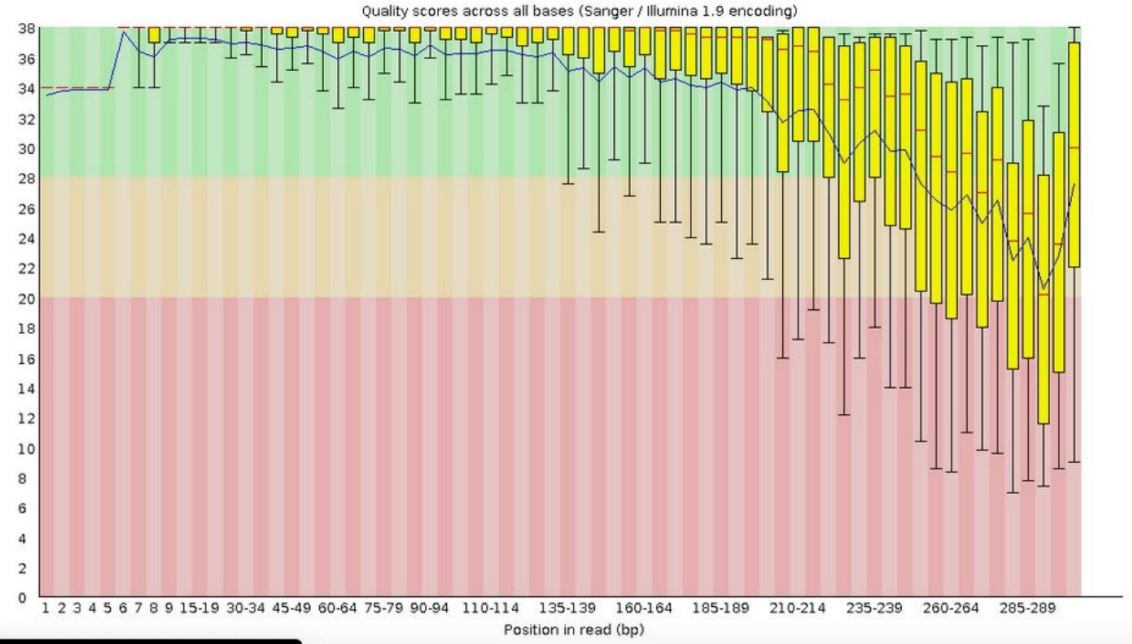
Evidence for mutations: Read alignment (RA)

```
CTTTATAGAGCATAAGCAGCAGCGCAACACCCTTAT mutation

CTTTATAGAGCATA---AGCAGCGCAACACCCTTAT reference

CTTTATAGAGCATA---A                   read 1 →   no end trimming
CTTTATAGAGCATA---AGCAG               read 2 →
      AGAGCATAAGCAGCAGCGCAA          read 3 →
          CATAAGCAGCAGCGCAACACCCTTAT read 4 ←
                AGCAGCGCAACACCCTTAT  read 5 ←
                   AGCGCAACACCCTTAT  read 6 ←

CTTTATAGAGCATaa                      read 1 → with end trimming
CTTTATAGAGCATAagcag                  read 2 →
      agagCATAAGCAGCAGCGCaa          read 3 →
          cATAAGCAGCAGCGCAACACCCTTAT read 4 ←
                agcagcGCAACACCCTTAT  read 5 ←
                   agcGCAACACCCTTAT  read 6 ←
```

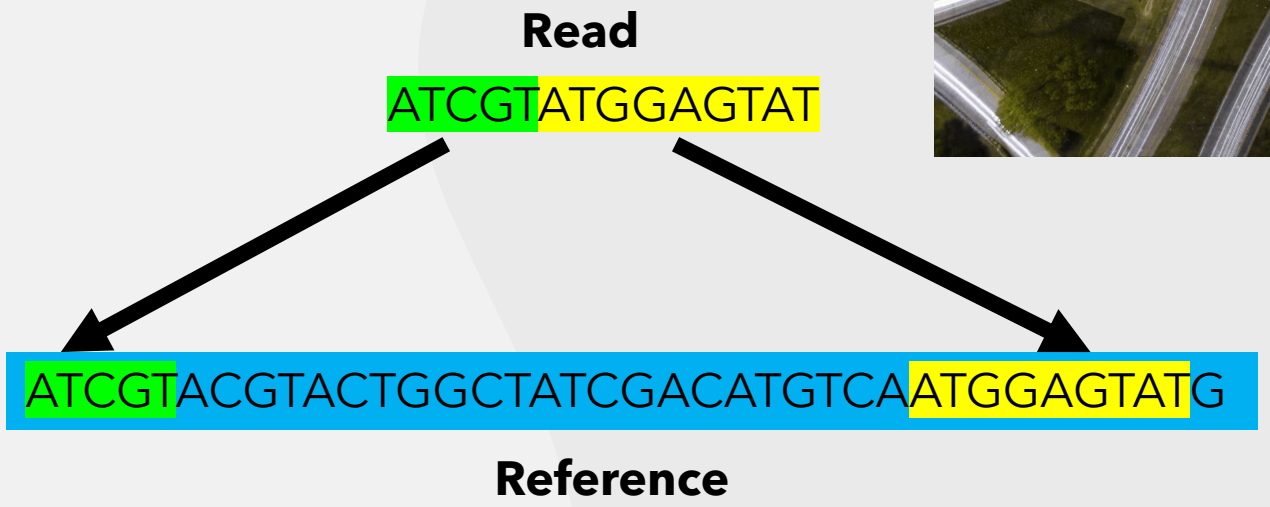
Forward reads



Reverse reads



Evidence for mutations: New Junctions (JC)



Evidence for mutations: Missing coverage (MC)

Predicted mutation

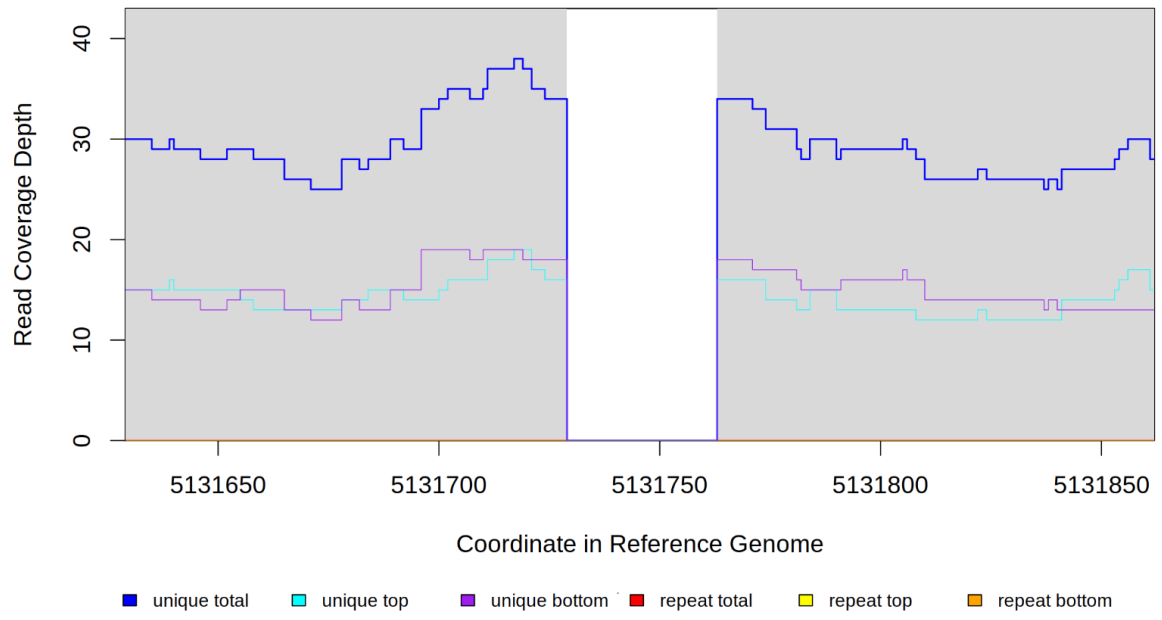
evidence	seq id	position	mutation	annotation	gene	
MC JC	NC_008463	5,131,729	Δ34 bp	pseudogene (288-321/1180 nt)	*PA14_RS23490* ←	acyl-CoA dehydrogenase family protein

Missing coverage evidence...

			seq id	start	end	size	←reads	reads→	gene	
*	*	÷	NC_008463	5131729	5131762	34	34 [0]	[0] 34	*PA14_RS23490*	acyl-CoA dehydrogenase family protein

New junction evidence

		seq id	position	reads (cov)	reads (cov)	score	skew	freq	annotation	gene	
*	?	NC_008463	= 5131728	0 (0.000)	30 (0.910)	20/260	0.3	100%	pseudogene (322/1180 nt)	*PA14_RS23490*	acyl-CoA dehydrogenase family protein
–	?	NC_008463	5131763 =	0 (0.000)					pseudogene (287/1180 nt)	*PA14_RS23490*	acyl-CoA dehydrogenase family protein



Predicted mutation

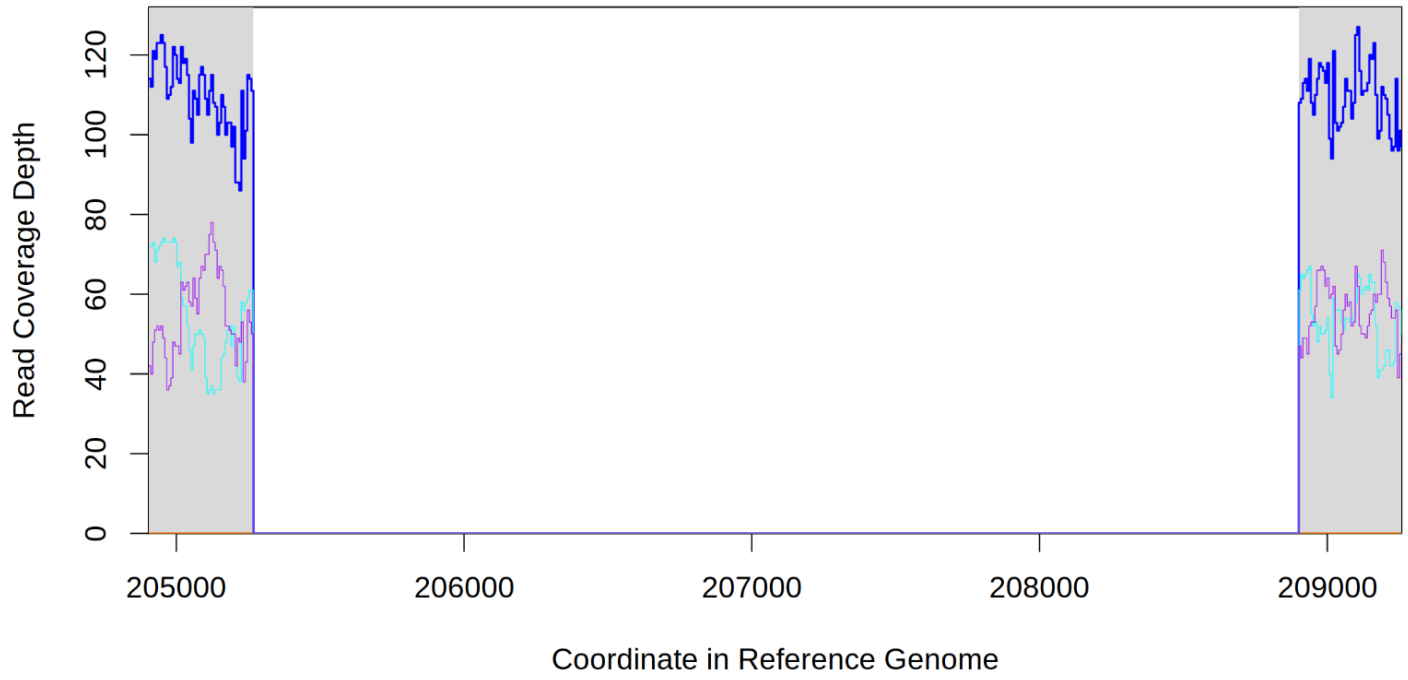
evidence	seq id	position	mutation	annotation	gene	
MC JC	NZ_CP064350	205,267	Δ3,634 bp		*[IT766_RS01045]–IT766_RS01055*	*[IT766_RS01045], IT766_RS01050, IT766_RS01055*

Missing coverage evidence...

			seq id	start	end	size	←reads	reads→	gene	
*	*	÷	NZ_CP064350	205267	208900	3634	111 [0]	[0] 108	*[IT766_RS01045]–IT766_RS01055*	[IT766_RS01045],IT766_RS01050,IT766_RS0105

New junction evidence

		seq id	position	reads (cov)	reads (cov)	score	skew	freq	annotation	gene
*	?	NZ_CP064350	= 205266	0 (0.000)	108 (0.930)	40/290	0.6	100%	coding (1962/1962 nt)	*IT766_RS01045*
	?	NZ_CP064350	208901 =	0 (0.000)					intergenic (-2/+254)	*IT766_RS01055/IT766_RS01060*



Mutation Prediction

Base substitutions

RA evidence = SNP or SUB mutation

Base substitution mutations are called from RA evidence. When only a single base is affected, **breseq** calls a base substitution (SNP) mutation. When multiple base substitutions occur adjacent to each other or in conjunction with indels (see below), **breseq** calls a substitution (SUB) mutation.

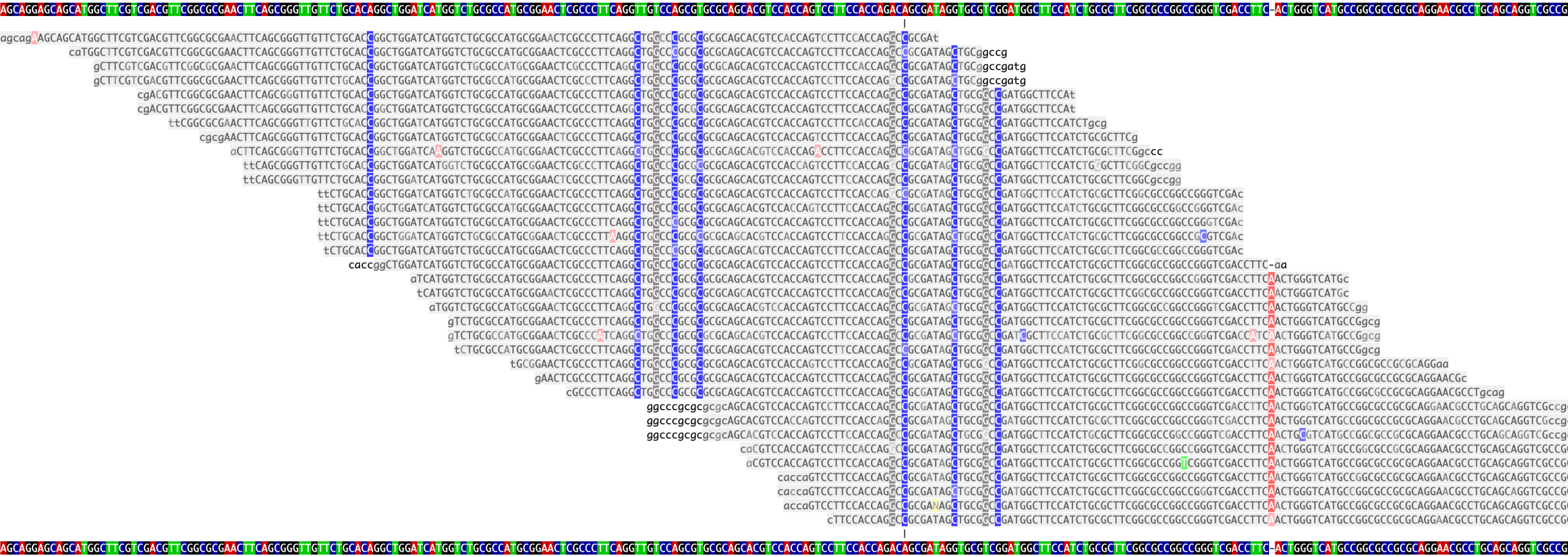
Predicted mutation

evidence	seq id	position	mutation	annotation	gene	description
RA	NC_008463	96,393	A→C	pseudogene (241/723 nt)	*tagH* ←	type VI secretion system-associated FHA domain protein TagH

Read alignment evidence...

	seq id	position	ref	new	freq	score (cons/poly)	reads	annotation	genes	product
*	NC_008463	96,393	0	A	C	100.0%	132.7 / NA	35	pseudogene (241/723 nt)	*tagH*	type VI secretion system-associated FHA domain protein TagH

Reads supporting (aligned to +/- strand): **ref** base A (0/0); **new** base C (23/12); **total** (23/12)



Predicted mutation

evidence	seq id	position	mutation	annotation	gene	description
RA	NC_008463	927,917	T→C	I56T (ATC→ACC)	*PA14_RS04340* →	CHASE domain-containing protein

Read alignment evidence...

	seq id	position	ref	new	freq	score (cons/poly)	reads	annotation	genes	product
*	NC_008463	927,917	0	T	C	100.0%	130.4 / NA	34	I56T (ATC→ACC)	*PA14_RS04340*	CHASE domain-containing protein

Reads supporting (aligned to +/- strand): **ref** base T (0/0); **new** base C (14/20); **total** (14/20)

Mutation Prediction

Short insertions and deletions

RA or JC evidence = INS, DEL, or SUB mutation

For single-base insertions and deletions, RA evidence with gap characters is used to call mutations as in the case of base substitutions. For longer insertions and deletions, for which missing coverage evidence may not exist, these events may be predicted solely on the basis of new junctions joining them.

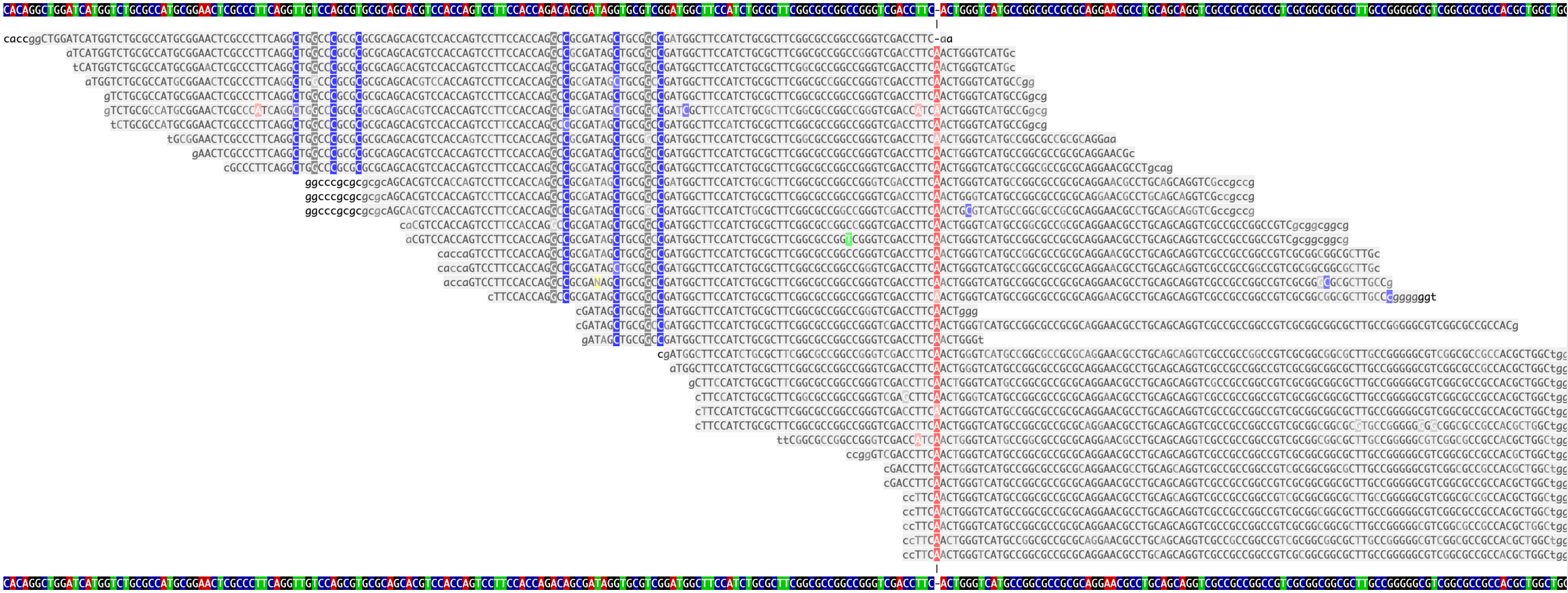
Predicted mutation

evidence	seq id	position	mutation	annotation	gene	description
RA	NC_008463	96,452	+A	pseudogene (182/723 nt)	*tagH* ←	type VI secretion system-associated FHA domain protein TagH

Read alignment evidence...

	seq id	position	ref	new	freq	score (cons/poly)	reads	annotation	genes	product
*	NC_008463	96,451	1	.	A	100.0%	107.8 / NA	36	pseudogene (183/723 nt)	*tagH*	type VI secretion system-associated FHA domain protein TagH

Reads supporting (aligned to +/- strand): **ref** base . (0/0); **new** base A (20/16); **total** (20/16)

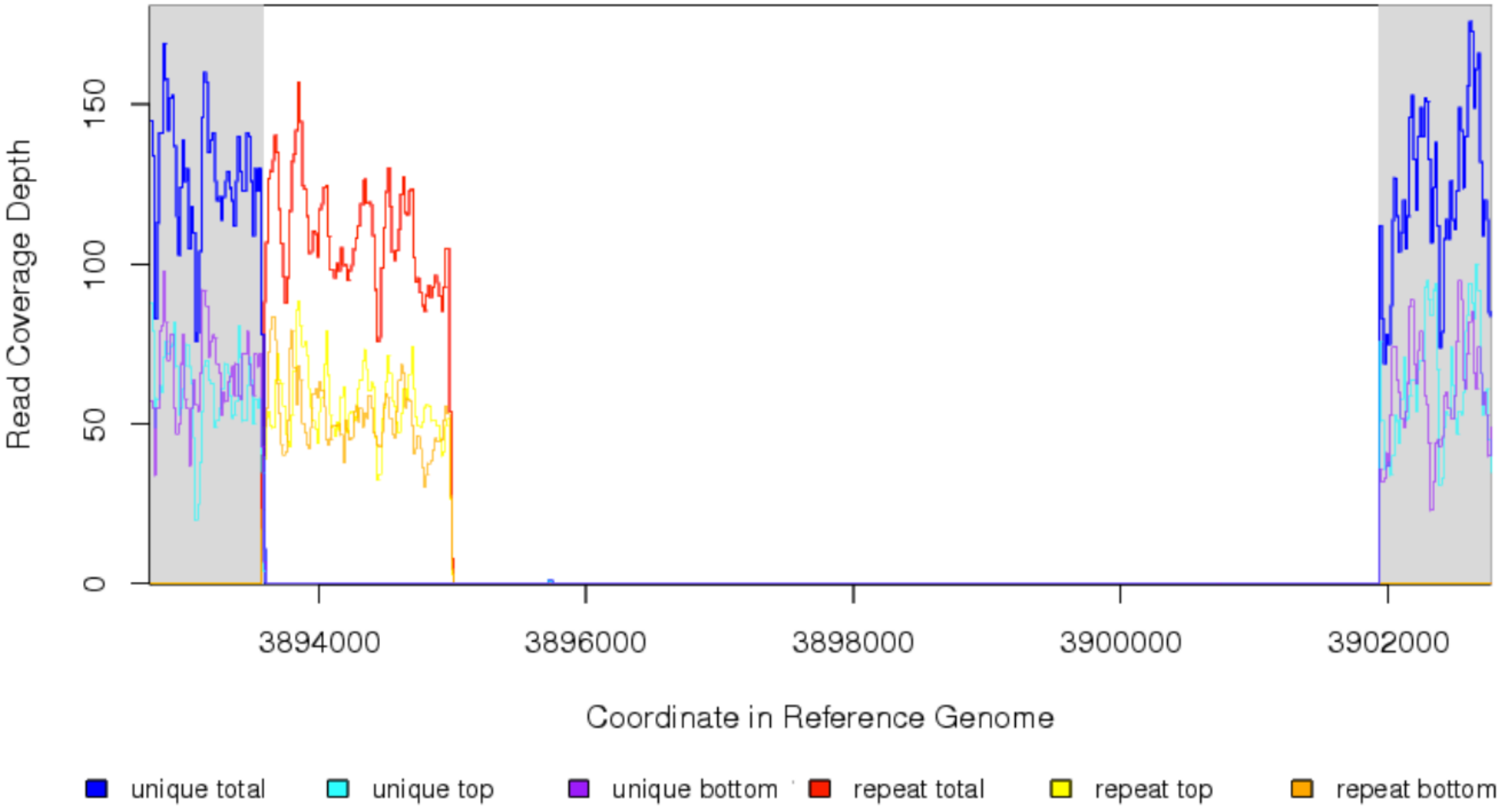


Mutation Prediction

Large deletions

MC+JC evidence = DEL mutation

Missing coverage typically indicates a large deletion event. When a junction also exists that precisely joins compatible endpoints, **breseq** predicts a deletion (DEL) mutation.


Read Coverage Depth
0
50
100
150
3894000
3896000
3898000
3900000
3902000
Coordinate in Reference Genome
unique total
unique top
unique bottom
repeat total
repeat top
repeat bottom

Mutation Prediction

Mobile element insertions

JC+JC evidence = MOB mutation

When two junctions exist that would join positions close by in the reference sequence to the ends of an annotated repeat_region, **breseq** predicts a mobile element insertion (MOB). It further tries to shift the ends of the junctions such that they align best with the ends of the mobile element.

Mutation Prediction

Duplications

JC evidence = AMP mutation

If new junction evidence connects a region of the genome to a region upstream on the same strand, then it typically indicates that the intervening bases have been duplicated and **breseq** predicts a duplication. **breseq** currently does not use evidence from changes in read coverage depth to predict copy number, so coverage should be manually examined to verify this class of mutations.

Mutation Prediction

Limitations

Even given perfect data, **breseq** cannot find some types of mutations:

- *Novel sequences, not existing in the reference*
- *Mutations in repeat regions*
- *Chromosomal inversions and rearrangements through repeat sequences*

Mutation Prediction: single-nucleotide polymorphism (SNP)

evidence	position	mutation	annotation	gene	description
RA	70,867	T→C	D92G (GAC→GGC)	*araA*	L-arabinose isomerase

evidence	position	mutation	annotation	gene	description
RA	1,298,712	T→G	intergenic (+674/-64)	*ychE/oppA*	predicted inner membrane protein/oligopeptide transporter subunit

Mutation Prediction: Substitution (SUB)

evidence	position	mutation	annotation	gene	description
RA	47,977	2 bp→AC	intergenic (+33/–)	*lambdap79*/–	hypothetical protein/–

Mutation Prediction: Insertion (INS)

evidence	position	mutation	annotation	gene	description
RA	3,893,551	+G	intergenic (+6/-50)	*kup/insJ-5*	potassium transporter/IS150 hypothetical protein

evidence	position	mutation	annotation	gene	description
RA	3,290,071	+CC	coding (205/4554 nt)	*gltB*	glutamate synthase, large subunit

Mutation Prediction: Deletion (DEL)

evidence	position	mutation	annotation	gene	description
MC JC	3,894,997	Δ6,934 bp	IS150-mediated	*rbsD–[yieO]*	*rbsD*, *rbsA*, *rbsC*, *rbsB*, *rbsK*, *rbsR*, *[yieO]*

evidence	position	mutation	annotation	gene	description
RA	1,332,148	Δ1 bp	intergenic (+131/-79)	*topA/cysB*	DNA topoisomerase I/DNA-binding transcriptional dual regulator, O-acetyl-L-serine-binding

Mutation Prediction: Mobile element insertion (MOB)

evidence	position	mutation	annotation	gene	description
JC JC	3,571,196	IS3 (–) :: +TCA (+3) bp	coding (397–399/435 nt)	*uspA*	universal stress global response regulator

evidence	position	mutation	annotation	gene	description
JC JC	4,524,522	IS186 (+) (+6) bp	coding (494–499/549 nt)	*fimA*	major type 1 subunit fimbrin (pilin)

evidence	position	mutation	annotation	gene	description
JC JC	2,736,667	Δ2 :: IS186 (+) (+9) bp	coding (818–826/1425 nt)	*ascB*	cryptic 6-phospho-beta-glucosidase

Avoid, if at all possible, and do long-read sequencing


Command Line
Candy
Bar
Candy