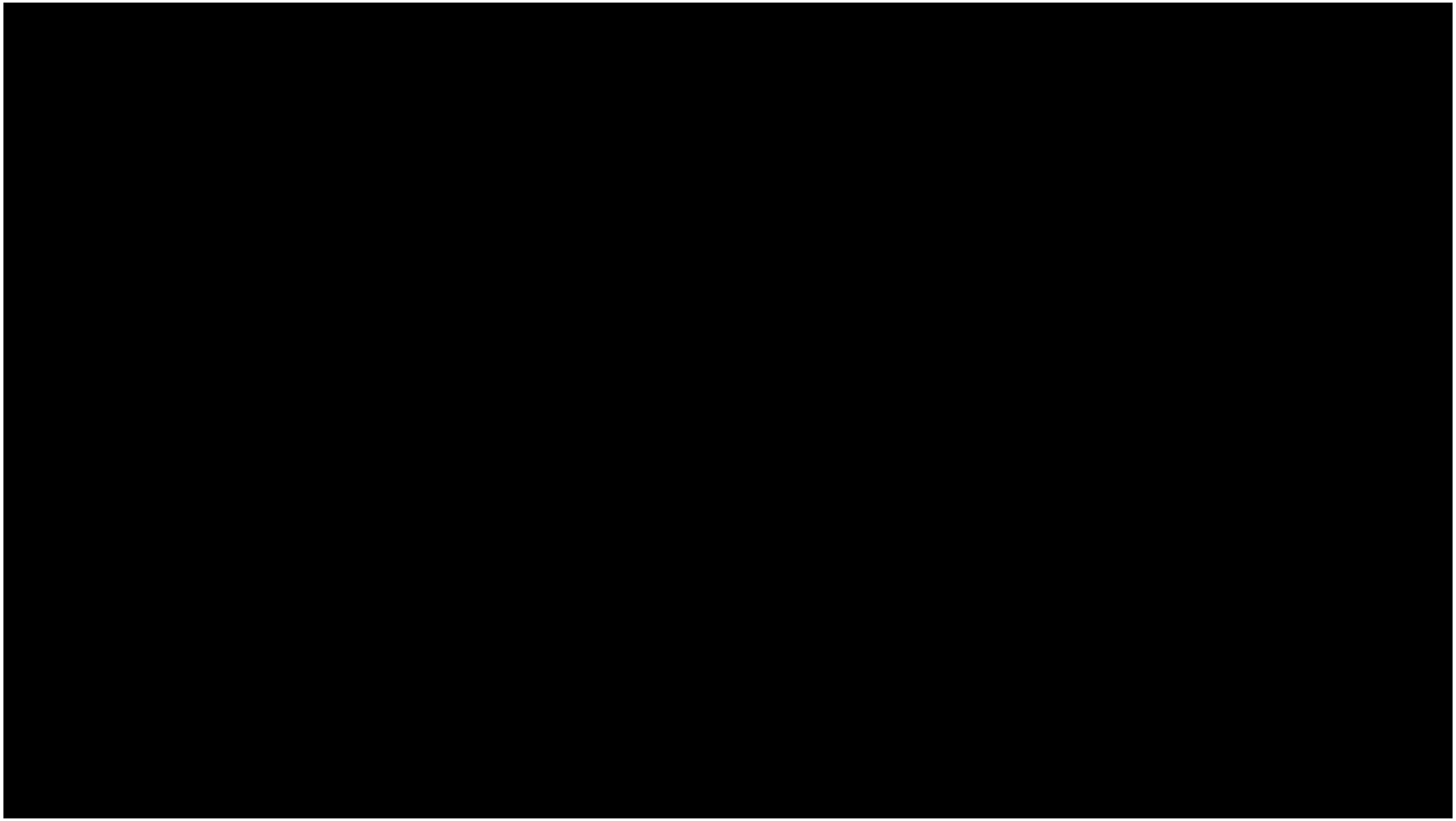
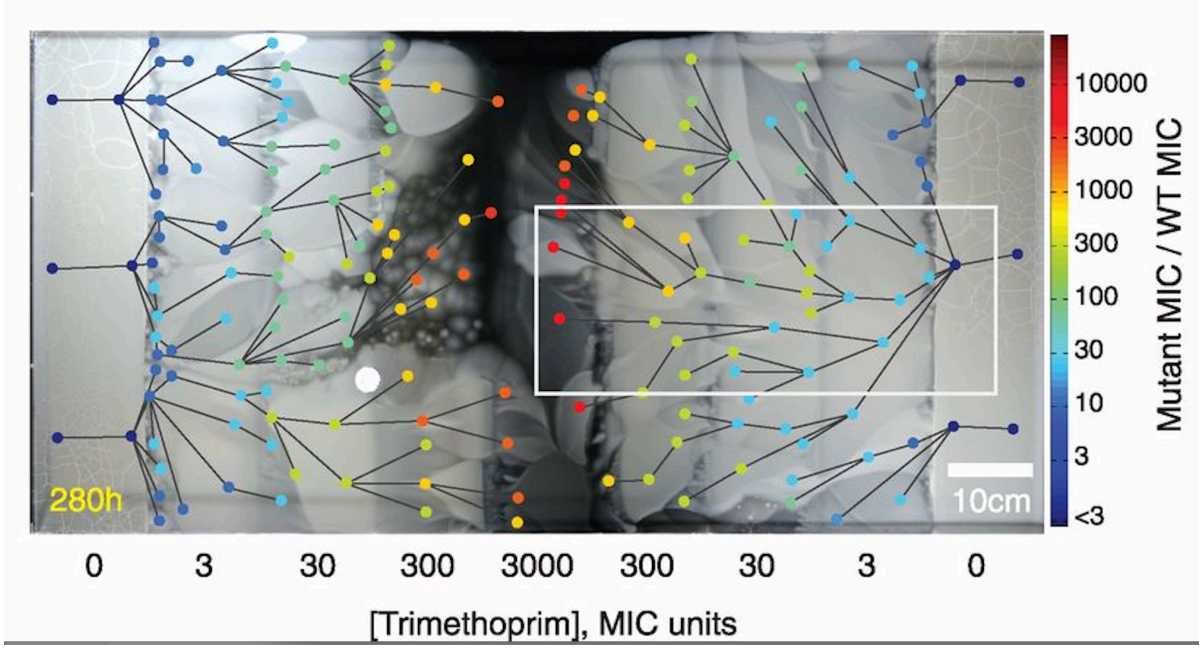
Variation Service in PATRIC

Presented by Fangfang Xia

AMR Evolution in Action



MEGA-plate: Visualizing Antibiotic Landscape



- Mutated *dnaQ* strains showed increased rates of mutations
- Mutator phenotype emerged at least six times independently
- Highly resistant mutants may be trapped
- Growth restored by compensatory mutations

Spatiotemporal microbial evolution on antibiotic landscapes , Kishony, 2016

Comparative Genomics Goals

- Identify genetic variations among organisms
- Associate genetic variations with phenotypes
- Hypothesize mechanisms that connect genotypes and phenotypes

Genotype–Phenotype Map

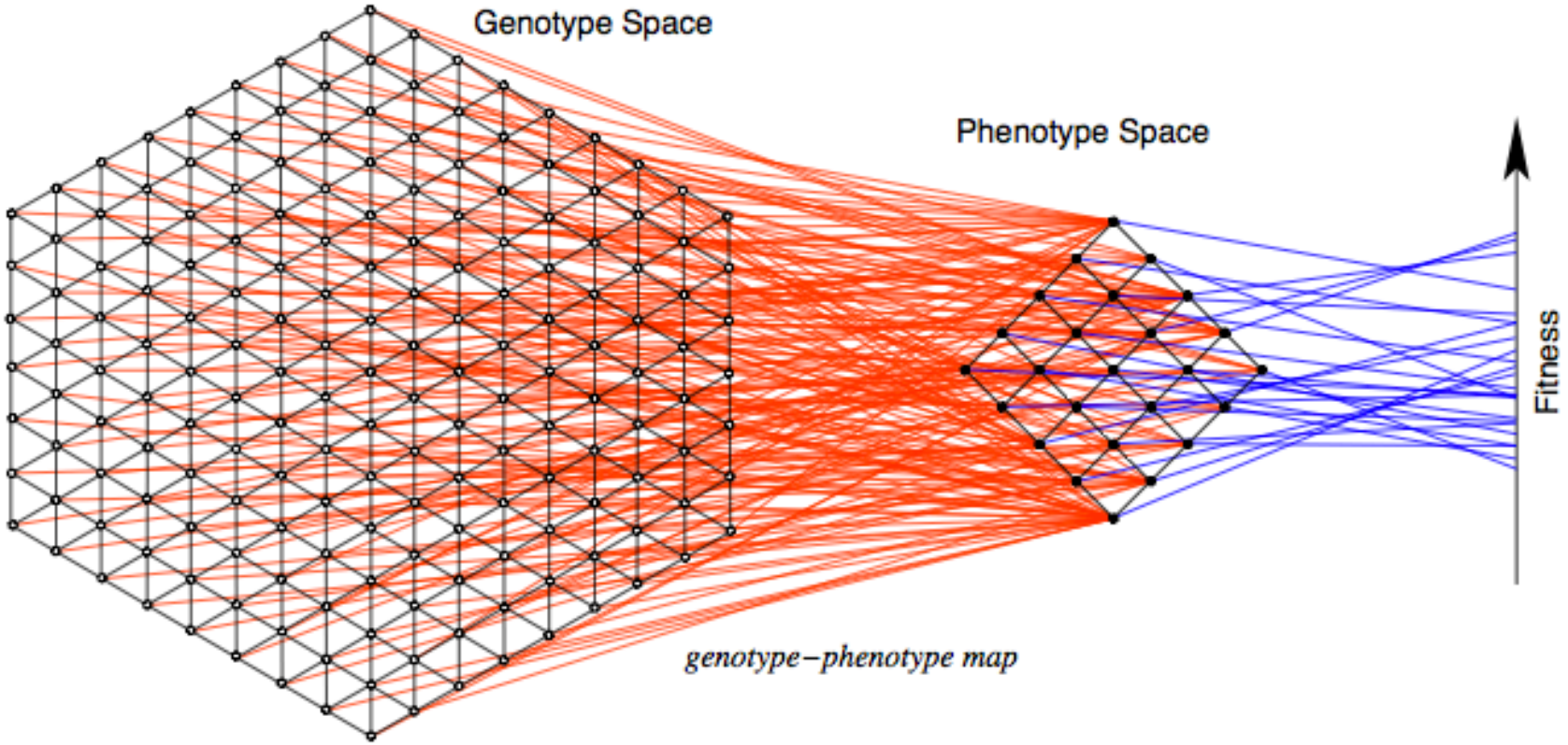
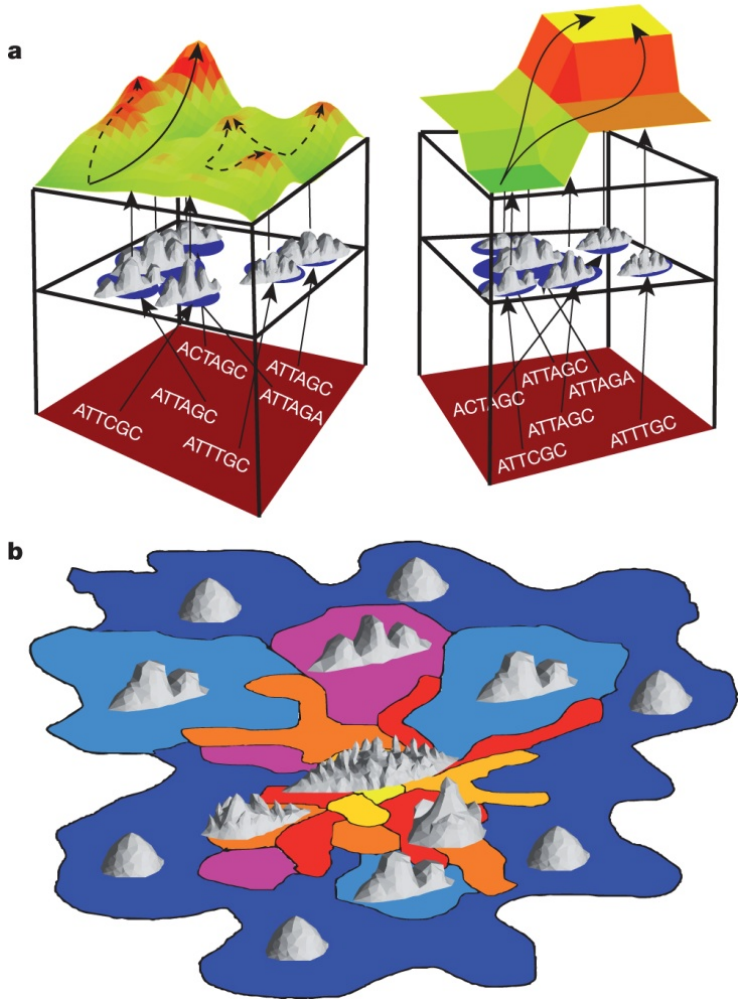


Figure 1. Fitness and Genotype-Phenotype map

Landscapes and Effective Fitness, Peter Stadler, 2002

PATRIC

Conceptual interpretation of genotype–fitness map



Layers:

- fitness space
- morphological variational space
- genetic variational space

Idealized two-dimensional representation of the tooth model parameter space

Salazar-Ciudad & M Marín-Riera *Nature* **000**, 1–4 (2013) doi:10.1038/nature12142

PATRIC

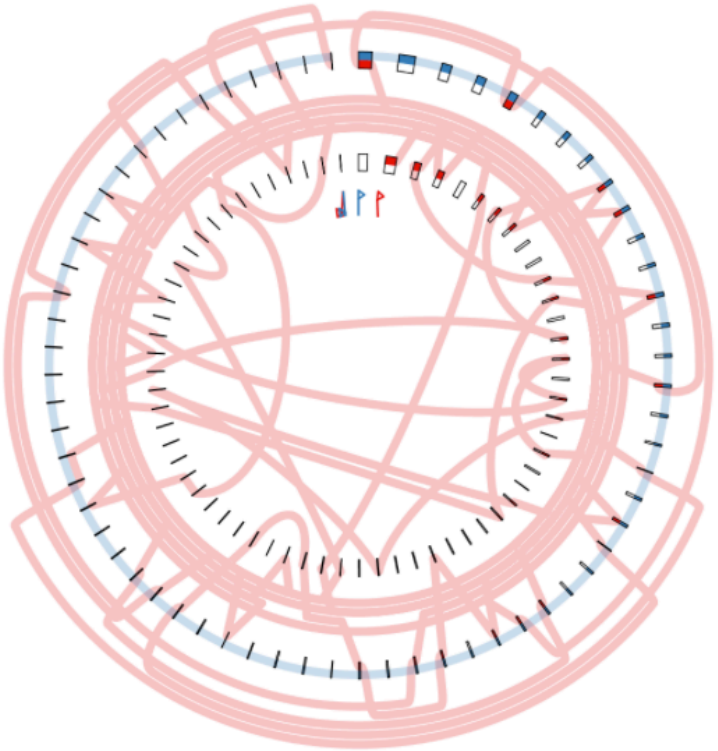
Levels of Genetic Variations

- Protein family (gains, losses, transfers, CNV)
- Protein sequence similarity

- DNA sequence similarity
- Genome rearrangements
- DNA local variations
  - SNPs, MNPs, indels, intergenic variations

Genome Rearrangement

NNB4-ART-DAR vs NNB4-ART-MS



126 vs 71 contigs

DAR-901 vs MS0193

50 vs 52 contigs

PATRIC

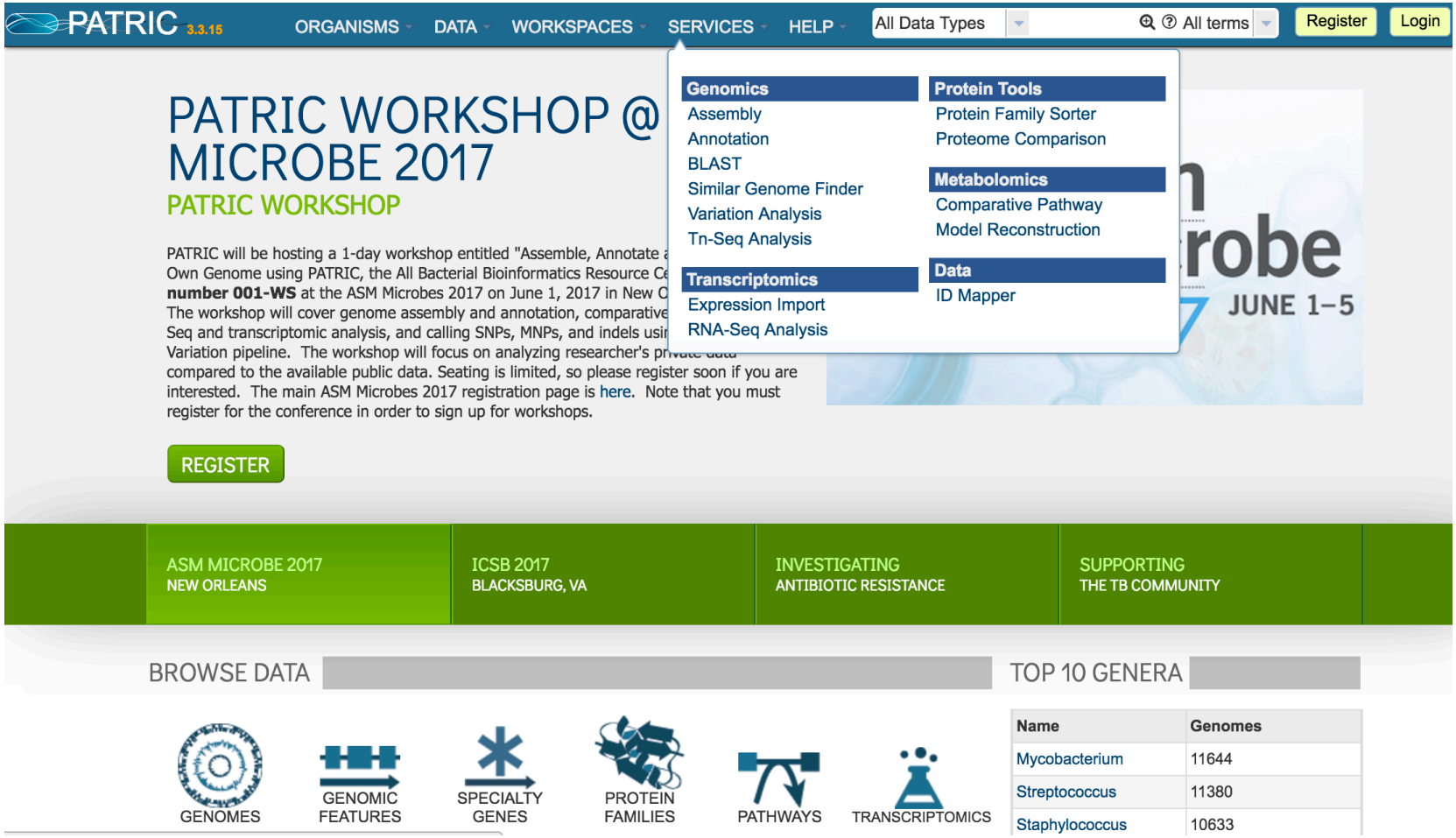
DNA Sequence Similarity

- Separating genuine sequence difference from artifacts
  - There can be more variations between assemblies of the same strain than different strains

DNA-diff comparison of pairs of assemblies

Sample	Ref \ Qry	D-a6	D-ra	D-sp	M-a6	M-ra	M-sp	SA-M	SB-M	NBB4	H37R
DAR-901_a6_scaffolds	D-a6	0.9998	0.9948	0.9988	0.9955	0.9843	0.9988	0.9987	0.9984	0.3029	0.0419
DAR-901_ray_filtered_scaffolds	D-ra	0.9977	0.9996	0.9992	0.9956	0.9854	0.9993	0.9992	0.9990	0.3058	0.0422
DAR-901_spades_filtered_scaffolds	D-sp	0.9953	0.9927	0.9998	0.9933	0.9825	0.9972	0.9969	0.9969	0.3072	0.0427
MS0193_a6_scaffolds	M-a6	0.9957	0.9928	0.9968	1.0000	0.9825	0.9971	0.9969	0.9967	0.3003	0.0415
MS0193_ray_filtered_scaffolds	M-ra	0.9976	0.9953	0.9992	0.9956	0.9996	0.9994	0.9992	0.9990	0.3077	0.0423
MS0193_spades_filtered_scaffolds	M-sp	0.9975	0.9951	0.9994	0.9958	0.9850	0.9998	0.9993	0.9992	0.3083	0.0427
Subsample_A_MS0193_spades_filtered_scaffolds	SA-M	0.9977	0.9953	0.9995	0.9959	0.9852	0.9996	0.9998	0.9994	0.3083	0.0428
Subsample_B_MS0193_spades_filtered_scaffolds	SB-M	0.9977	0.9953	0.9994	0.9958	0.9850	0.9996	0.9994	0.9998	0.3082	0.0428
Mycobacterium_chubuense_NBB4	NBB4	0.2797	0.2817	0.2832	0.2768	0.2790	0.2833	0.2831	0.2831	1.0000	0.0499
Mycobacterium_tuberculosis_H37Rv	H37R	0.0550	0.0557	0.0560	0.0545	0.0545	0.0560	0.0560	0.0560	0.0692	1.0000

PATRIC Services



Variation Job Submission

Variation Analysis

Identify and annotate sequence variation (SNP, MNP, indel) relative to a closely related reference genome.

Paired read library

READ FILE 1

SRR3146372_1.fastq.gz

READ FILE 2

SRR3146372_2.fastq.gz

Single read library

READ FILE

Selected libraries

Place read files here using the arrow buttons.

- P(SRR31..tq.gz, SRR31..tq.gz)
- P(SRR31..tq.gz, SRR31..tq.gz)
- P(SRR31..tq.gz, SRR31..tq.gz)

Parameters

Target Genome

Acinetobacter baumannii UH81_364

OUTPUT FOLDER

Variation Demo

OUTPUT NAME

Patient_81

Reset Submit

Uploads 0·0 Jobs 46·1·0·6

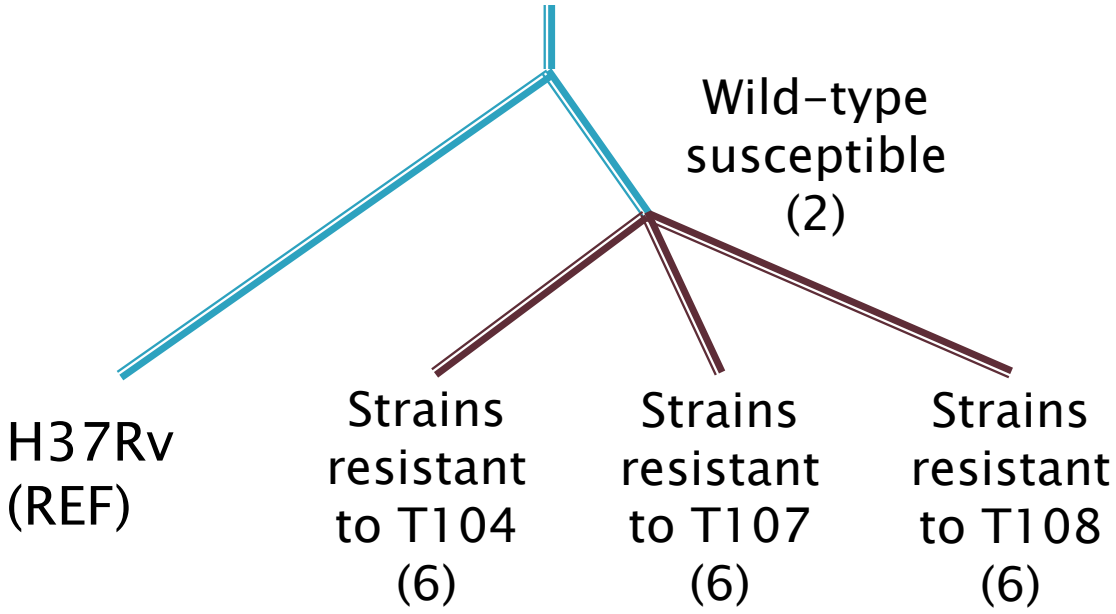
PATRIC

Variation Analysis

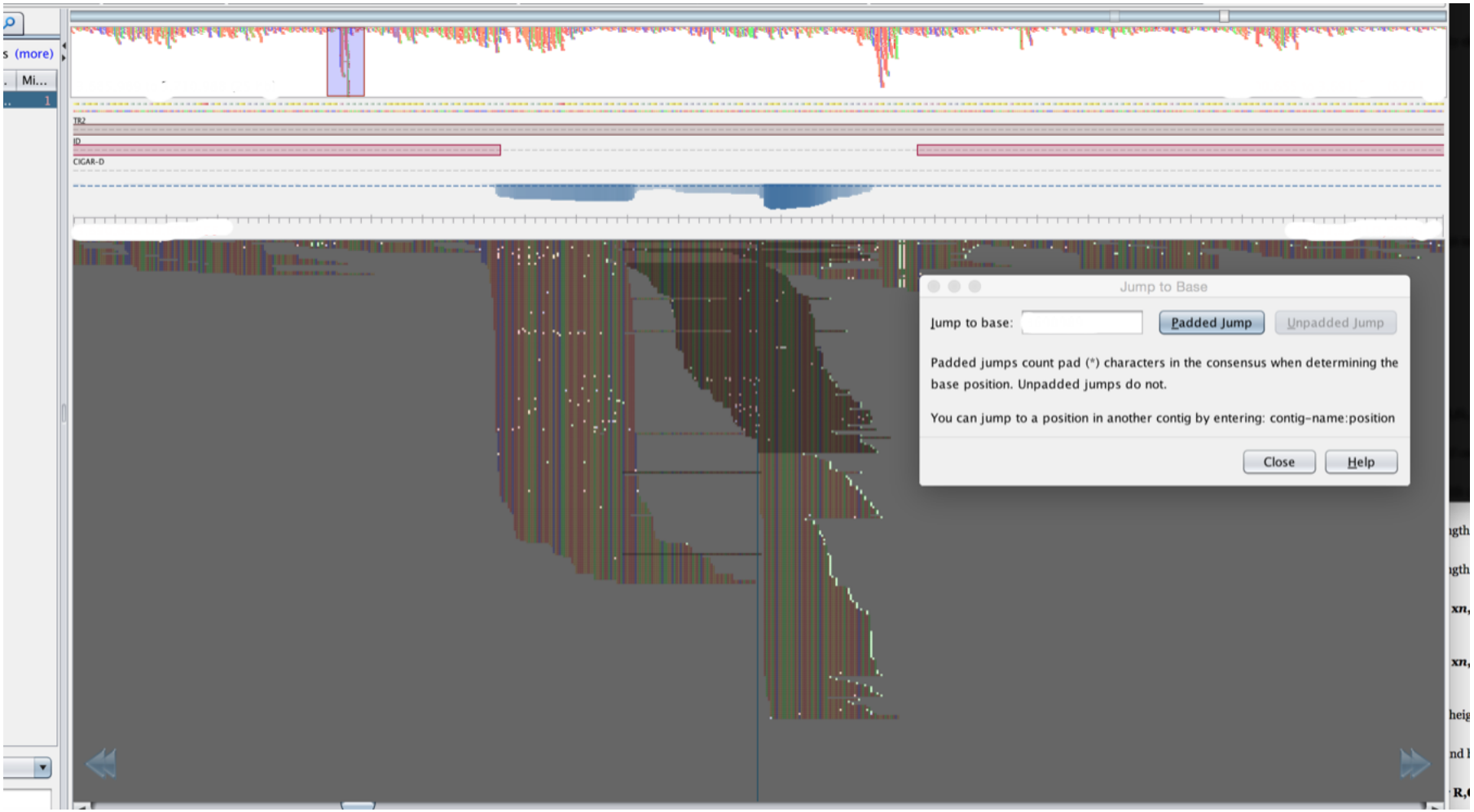
- Read mapping
- Coverage statistics
- Variant calling
- Variant annotation
- Cross-sample summary

AMR Analysis with Gyanu Lamichhane Lab

- 20 genomes
  - 2 are wild-type parent strains
  - 18 offspring strains resistant to 3 carbapenems



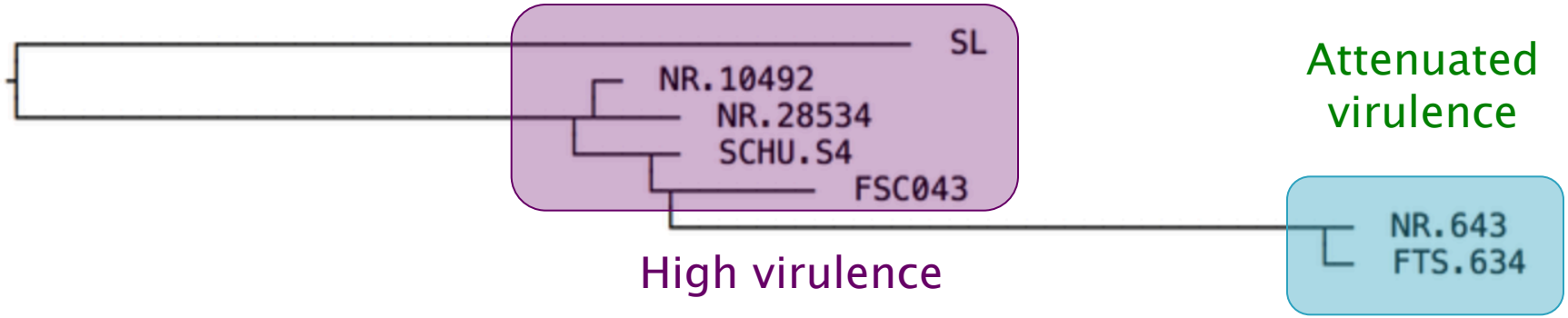
Highly Expressed Intergenic RD



PATRIC

DNA sequence comparison of *Francisella tularensis* Schu S4 substrains

- Seven substrains in two virulence groups



- Goal: Identify Regions of Differences
  - Verify the presence of virulence factors
  - Confirm the identity of the organism in stock culture

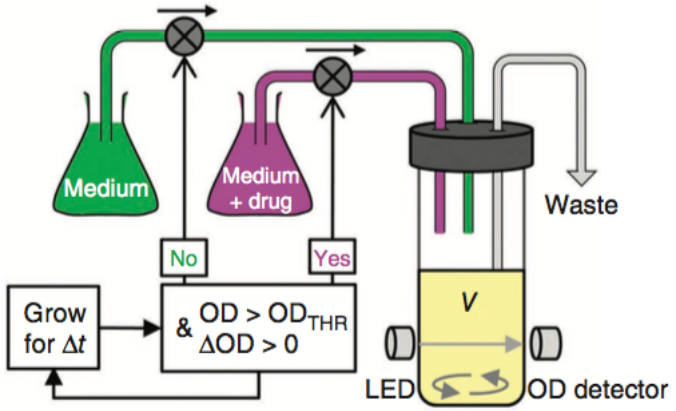
Separating Genuine Regions of Differences from Assembly Artifacts

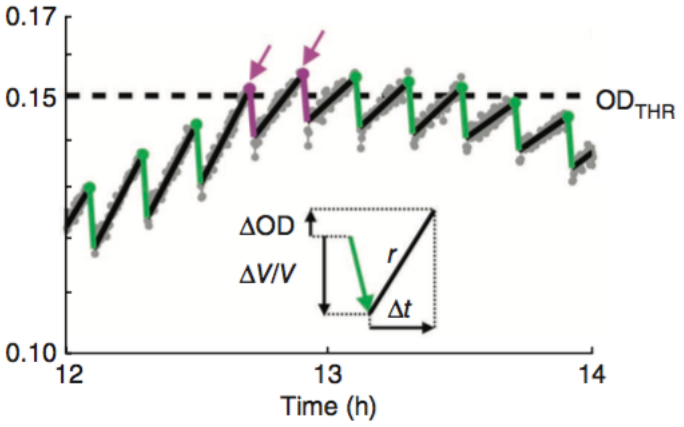
AJ749949 REF		FSC043	NR-10492	NR-28534	SL	FTS-634	NR-643	Type	
3195*	-	1_0* -	1_1 C	1_1 C	1_0* -	1_0* -	1_0* -	IND	Repeat region (957 nt, NR-10492_1:1:957)
3195*	-	1_0* -	1_2 C	1_2 C	1_0* -	1_0* -	1_0* -	IND	Likely assembly artificats caused by repeats:
3195*	-	1_0* -	1_3 T	1_3 T	1_0* -	1_0* -	1_0* -	IND	The mapped reads that extend from SCHUS4 ref:
...*	-	...* -	... .	... .	...* -	...* -	...* -	IND	13 lines
3195*	-	1_0* -	1_17 G	1_17 G	1_0* -	1_0* -	1_0* -	IND	The locus from 3105...4084 has mobile element
3195*	-	1_0* -	1_18 A	1_18 A	1_0* -	1_0* -	1_0* -	IND	Of these 50 there are 4 truncated ISFTU1s, so
3195*	-	1_0* -	1_19 G	1_19 G	1_0* -	1_0* -	1_0* -	IND	In other words, the reads used for the assembl
3195*	-	1_0* -	1_20 T	1_20 T	1_1 T	1_0* -	1_0* -	IND	
3195*	-	1_0* -	1_21 A	1_21 A	1_2 A	1_0* -	1_0* -	IND	
3195*	-	1_0* -	1_22 T	1_22 T	1_3 T	1_0* -	1_0* -	IND	
...*	-	...* -	... .	... .	... .	...* -	...* -	IND	70 lines
3195*	-	1_0* -	1_93 T	1_93 T	1_74 T	1_0* -	1_0* -	IND	
3195*	-	1_0* -	1_94 T	1_94 T	1_75 T	1_0* -	1_0* -	IND	
3195*	-	1_0* -	1_95 A	1_95 A	1_76 A	1_0* -	1_0* -	IND	
3196	A	1_0* -	1_96 A	1_96 A	1_77 A	1_0* -	1_0* -	IND	
3197	A	1_0* -	1_97 A	1_97 A	1_78 A	1_0* -	1_0* -	IND	
3198	G	1_0* -	1_98 G	1_98 G	1_79 G	1_0* -	1_0* -	IND	
--									
1419874	A	1_1417461 A	1_1418732 A	1_1416876 A	1_1417139 A	1_1416598 A	1_1416737 A		
1419875	T	1_1417462 T	1_1418733 T	1_1416877 T	1_1417140 T	1_1416599 T	1_1416738 T		
1419876	G	1_1417463 G	1_1418734 G	1_1416878 G	1_1417141 G	1_1416600 G	1_1416739 G		
1419877	C	1_1417464 C	1_1418735 C	1_1416879 C	1_1417142 C	1_1416601 T	1_1416740 T	SNP	Nonsyn (Cct -> Tct, P -> S)
1419878	C	1_1417465 C	1_1418736 C	1_1416880 C	1_1417143 C	1_1416602 C	1_1416741 C		
1419879	T	1_1417466 T	1_1418737 T	1_1416881 T	1_1417144 T	1_1416603 T	1_1416742 T		
1419880	A	1_1417467 A	1_1418738 A	1_1416882 A	1_1417145 A	1_1416604 A	1_1416743 A		
--									
1423159	C	1_1420746 C	1_1422017 C	1_1420161 C	1_1420424 C	1_1419883 C	1_1420022 C		
1423160	A	1_1420747 A	1_1422018 A	1_1420162 A	1_1420425 A	1_1419884 A	1_1420023 A		
1423161	C	1_1420748 C	1_1422019 C	1_1420163 C	1_1420426 C	1_1419885 C	1_1420024 C		
1423162	A	1_1420749 A	1_1422020 A	1_1420164 A	1_1420427 A	1_1419886 G	1_1420025 G	SNP	Nonsyn (Agt -> Ggt, S -> G)
1423163	G	1_1420750 G	1_1422021 G	1_1420165 G	1_1420428 G	1_1419887 G	1_1420026 G		
1423164	T	1_1420751 T	1_1422022 T	1_1420166 T	1_1420429 T	1_1419888 T	1_1420027 T		
1423165	A	1_1420752 A	1_1422023 A	1_1420167 A	1_1420430 A	1_1419889 A	1_1420028 A		

Evolutionary paths to antibiotic resistance under dynamically sustained drug selection

Erdal Toprak[1,6], Adrian Veres[2,6], Jean-Baptiste Michel[1,3], Remy Chait[1], Daniel L Hartl[4] & Roy Kishony[1,5]

Antibiotic resistance can evolve through the sequential accumulation of multiple mutations[1]. To study such gradual evolution, we developed a selection device, the 'morbidostat', that continuously monitors bacterial growth and dynamically regulates drug concentrations, such that the evolving population is constantly challenged[2–5]. We analyzed the evolution of resistance in *Escherichia coli* under selection with single drugs, including chloramphenicol, doxycycline and trimethoprim. Over a period of ~20 days, resistance levels increased dramatically, with parallel populations showing similar phenotypic trajectories. Whole-genome sequencing of the evolved strains identified mutations both specific to resistance to a particular drug and shared in resistance to multiple drugs. Chloramphenicol and doxycycline resistance evolved smoothly through diverse combinations of mutations in genes involved in translation, transcription and transport[3]. In contrast, trimethoprim resistance evolved in a stepwise manner[1,6], through mutations restricted to the gene encoding the enzyme dihydrofolate reductase (*DHFR*)[7,8]. Sequencing of *DHFR* over the time course of the experiment showed that parallel populations evolved similar mutations and acquired them in a similar order[9].





Demo Data

- Reference genome:
  - Escherichia coli str. K-12 substr. MG1655
- Reads:
  - BioSample: SAMN00723035

SRR396518.fastq.gz	237 MB	7/26/16, 3:57:00 AM
SRR396519.fastq.gz	264 MB	7/26/16, 4:00:00 AM
SRR396520.fastq.gz	218 MB	7/26/16, 4:02:00 AM
SRR396521.fastq.gz	271 MB	7/26/16, 4:05:00 AM

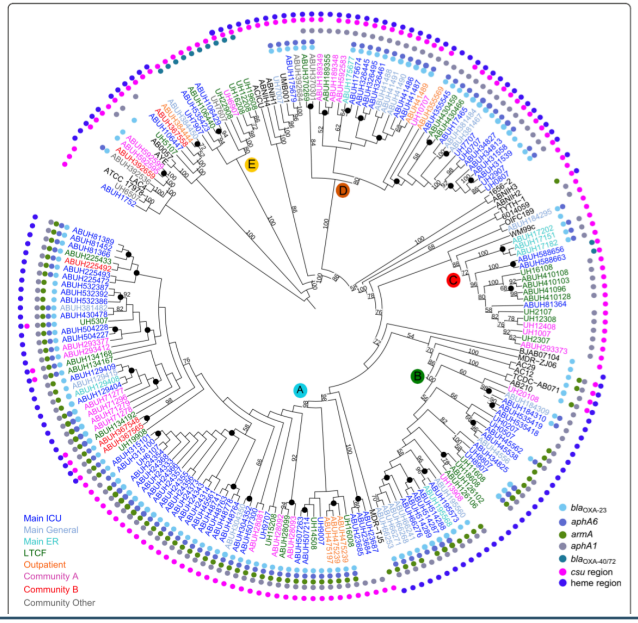
Wright *et al. Genome Medicine* (2016) 8:26
DOI 10.1186/s13073-016-0279-y

Genome Medicine

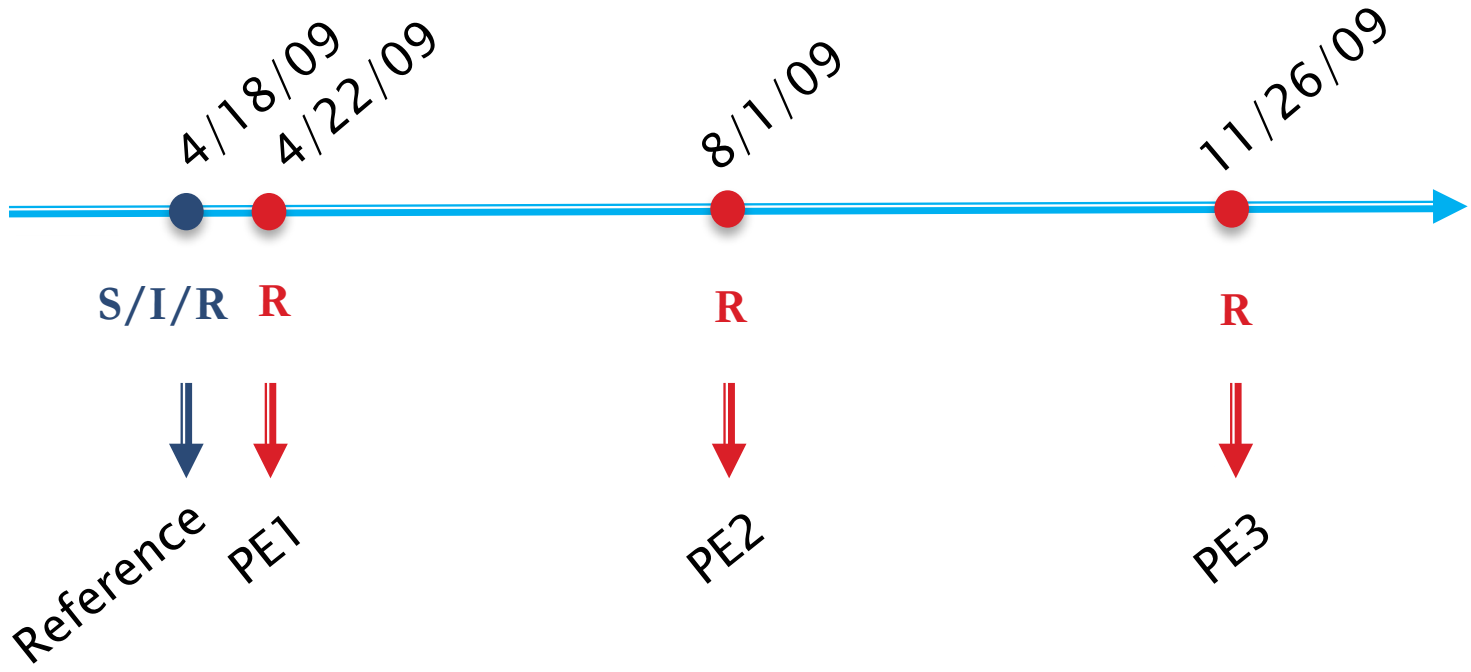
RESEARCH Open Access

Genome dynamics of multidrug-resistant *Acinetobacter baumannii* during infection and treatment

Meredith S. Wright[1], Alina Iovleva[2], Michael R. Jacobs[3,4], Robert A. Bonomo[5,6] and Mark D. Adams[1*]



Serial Samples from Patient 81



Results

- Summary
  - A total of 552 variants have been found. 293 of these variants are identified in all read libraries.
- Questions
  - Which RDs are significant?
    - frameshift, stop gain/loss, nonsyn, regulatory region
    - high fraction, high coverage
  - Which RDs are common to the later samples?

Inferred Variations by Sample

```
[1] PE1 (SRR3146360_1.fastq.gz,SRR3146360_2.fastq.gz)

    Total reads              = 12261074
    Properly mapped reads    = 11919546 (97.21%)
    Total reference bases    = 3977313
    Median base coverage     = 18
    Mean base coverage       = 63.4
    Bases with zero coverage = 462022 (11.616%)
    Bases with <=10 coverage = 1606088 (40.381%), in 111711 contiguous regions
    Raw FreeBayes variants   = 2056
    High quality variants    = 719

[2] PE2 (SRR3146370_1.fastq.gz,SRR3146370_2.fastq.gz)

    Total reads              = 10282737
    Properly mapped reads    = 9865157 (95.94%)
    Total reference bases    = 3976580
    Median base coverage     = 18
    Mean base coverage       = 67.1
    Bases with zero coverage = 436258 (10.971%)
    Bases with <=10 coverage = 1577711 (39.675%), in 114191 contiguous regions
    Raw FreeBayes variants   = 2027
    High quality variants    = 693
```

Comparison with Paper Results

- *Phylogenetic analysis revealed that a significant fraction of apparently persistent infections are in fact due to re-infection with new strains. SNVs primarily resulted in protein coding changes, and IS events primarily interrupted genes or were in an orientation such that the adjacent gene would be over-expressed.* ***Mutations acquired during infection were over-represented in transcriptional regulators, notably pmrAB and adeRS, which can mediate resistance to the last line therapies*** *colistin and tigecycline, respectively, as well as transporters, surface structures, and iron acquisition genes.*

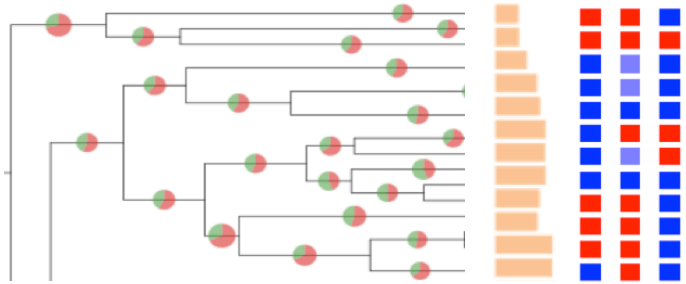
Feature Space to Compute Against

- Genetic variations
  - Protein presence/absence
  - SNPs, MNPs
  - Insertions/deletions
  - Genome rearrangements
  - Kmer counts
  - Evolutionary histories
- Machine learning
- Statistics
- Omics

Bacterial-Genome Association Studies

Genotypic Variation

- What's changed:
  gene gains/losses
- SNPs, insertions, deletions
- Genome rearrangements
- CNV in repeat regions

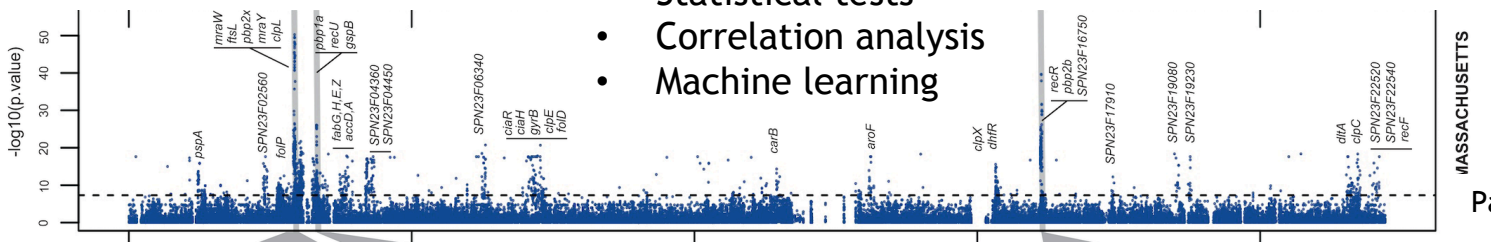


Phenotypic Variation

- Antibiotic resistance
- Differential growth rates on media types
- Survival/growth under different conditions
- Metabolomic fingerprints

Genotype-phenotype association

- Statistical tests
- Correlation analysis
- Machine learning



Parkhill et al, 2014

Curated subsystems | *Specialty gene databases* | *Transcriptome analysis* | *Metabolic and regulatory models* | *Protein – protein association graphs* | *Experimental validation*

Mechanism elucidation

Deep Learning for Phenotype Prediction