Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes

Lees J.A., Vehkala M. et al., 2016 In Review

Journal Club
Triinu Kõressaar
16.03.2016

Introduction

Bacterial genomes – how to find genetic associations with phenotypic variation?

New method SEER proposed - identifies sequence elements that are significantly enriched in a phenotype of interest

Applicable to even tens of thousands of genomes

Uses raw reads or *de novo* assembled contigs

Stand-alone pipeline

SEER is used to identify resistance determinants of *Streptococcus pneumoniae* for several antibiotics

SEER implements and combines three key insights:

1. an efficient scan of all possible k-mers with a distributed string mining algorithm

2. an appropriate alignment-free correction for clonal population structure

3. a fast and fully robust association analysis of all counted k-mers

Scan of k-mers (1/2):

Kmers allow simultaneous discovery of both short genetic variants and entire genes associated with a phenotype

Longer k-mers provide higher specificity but less sensitivity than shorter k-mers

An efficient implementation that allows counting and testing simultaneously at all k-mers at lengths over 9 bases long

Scan of k-mers (2/2):

Three different methods to count kmers in samples:

1. For very large studies or for counting directly from reads rather than assemblies, an implementation of distributed string mining (DSM) which limits maximum memory usage per core, but requires a large cluster to run

2) Sets up to 5,000 sample assemblies, a single core version of FSM-lite (Frequency based String Mining) is implemented

3) DSK (disk streaming of k-mers), very low memory usage; single kmer length is counted in each sample individually, results are then combined

Correction of clonal population structure of bacterial populations

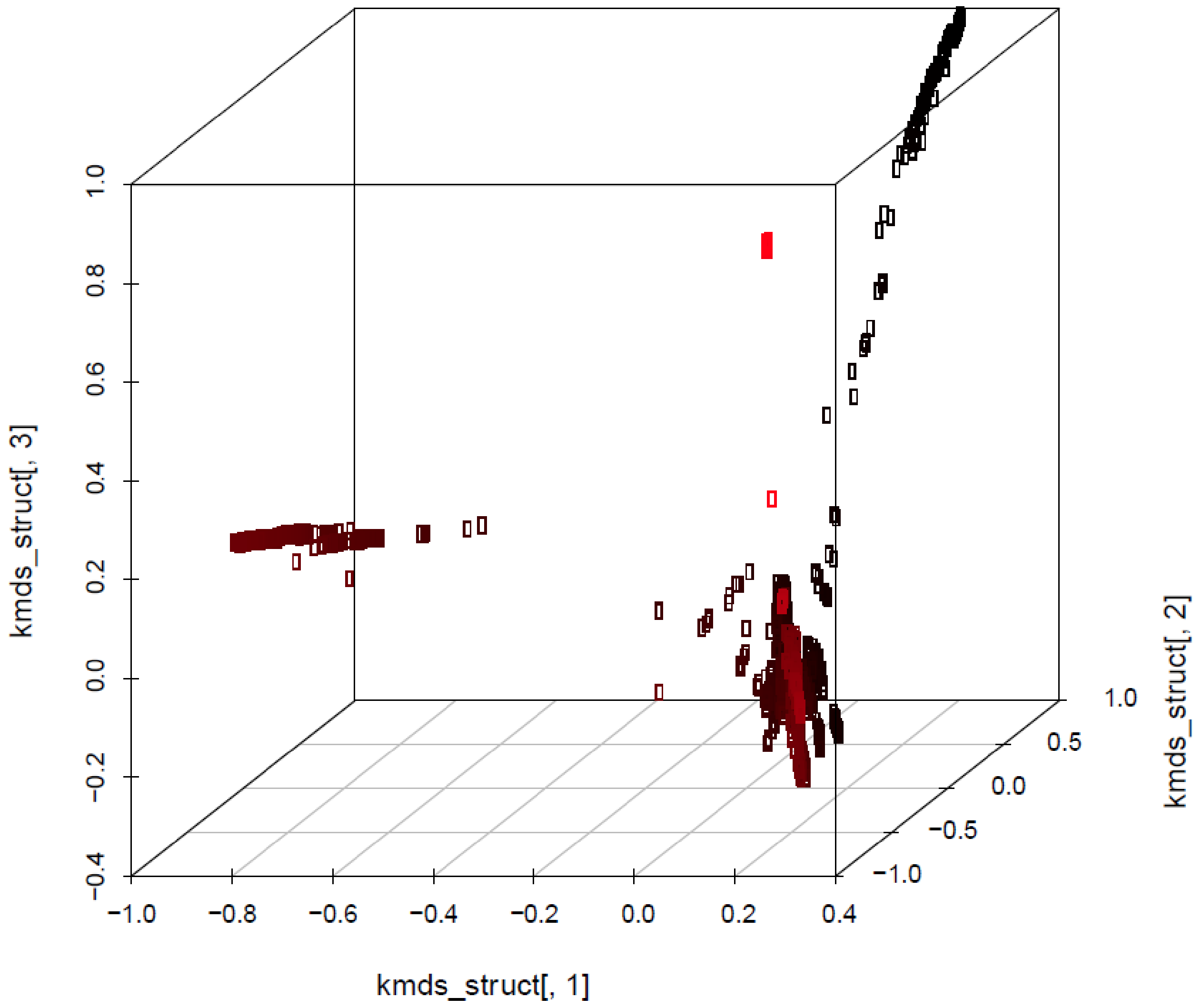
Distance matrix is constructed from a random subsample of kmers (between 0.1% and 1% of kmers appearing in between 5-95% of isolates is taken)

Pairwise distance matrix D is performed, each element being equal to a sum over all m sampled k-mers

$$d_{ij} = \sum_{m} \|k_{im} - k_{jm}\|$$

where k_{im} is 1 if the m-th sampled k-mer is present in sample i, and 0 otherwise

on which multidimensional scaling (MDS) is performed, projecting these distances into three dimensions. The normalised eigenvectors of each dimension are used as covariates in the regression model.



Supplementary figure 1: Plot of the k-mer distances projected into three dimensions by MDS for the Chewapreecha *et. al.* study of 3069 Thai carriage *S. pneumoniae* samples. Shade from black to red is by y-coordinate (2nd MDS component).

Statistics (1/2)

To binary phenotype data, for each kmer, logistic regression is used

$$\log\left(\frac{\boldsymbol{y}}{\boldsymbol{I}-\boldsymbol{y}}\right) = \mathbf{X}\boldsymbol{\beta}$$

y – binary outcome vector
X – design matrix, where
column 1 – vector of ones, intercept
column 2 – absence or present of kmer 0/1
column 3,4.. – eigenvectors of MDS projection, user-supplied categorical covariates, user-supplied quantitative covariates

BFGS algorithm is used to maximise the log likelihood

Statistics (2/2)

To continuous data, linear model is used $\boldsymbol{Y} = \boldsymbol{X\beta}$

Squared distance U(β) $U(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X\beta}\|^2$ is minimised using the BFGS algorithm

Wald statistics (statistic of χ^2 distribution) is used with the null hypothesis (β=0) of no association

$$W = \frac{\beta_1}{\mathrm{SE}(\beta_1)}$$

The basal cut off for significance $p < 0.05$
Bonferroni corrected $1\text{x}10^{-8}$ (based on *S.pneumoniae* genome analyzes)

SEER outputs effect size, direction, std error

Interpreting significant kmers

Kmers are filtered if they appear less than 1% or more than 99% of samples or are over 100 bases long

Significant kmers are required $\beta_1 > 0$

SEER outputs effect size, direction, std error

Kmer homology search with BLAT against well annotated reference sequence and annotated draft assemblies

Or kmers assembled using Velvet to better search for gene clusters

BLAT mapping and SNP calling with bcftools

Time and memory requirement

In case of 3069 simulated genomes, DSM finds kmers with 2hrs 38min on 16cores and uses 1Gb RAM

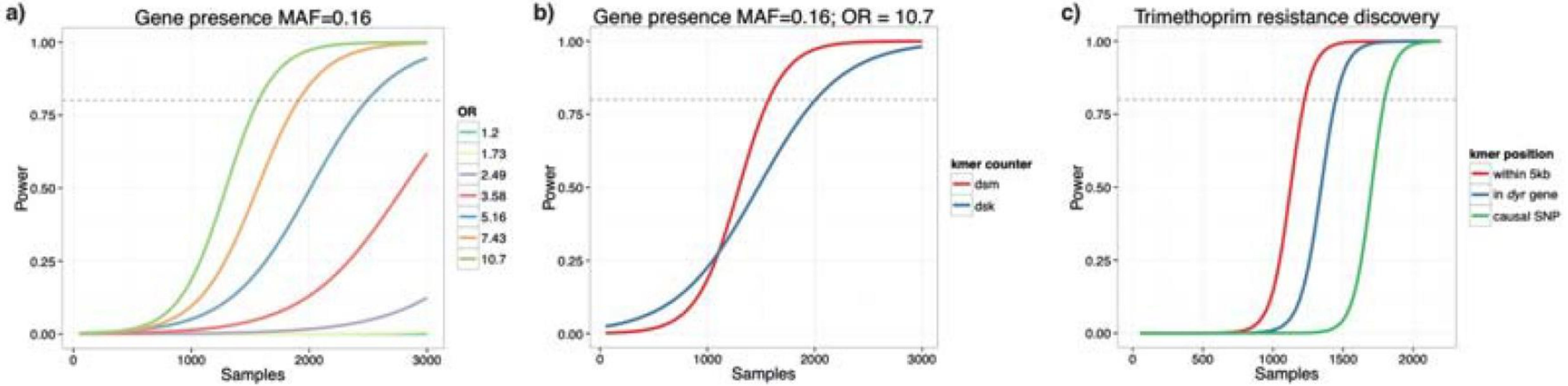
Single core versioon of DSM – 675 *S.pyogenes* genomes took 3hrs 44min and used 22.3Gb.

Calculation of covariates (300,000 kmers, 3069 genomes) using 16 cores, took 6hrs 42min and 8.33Gb RAM

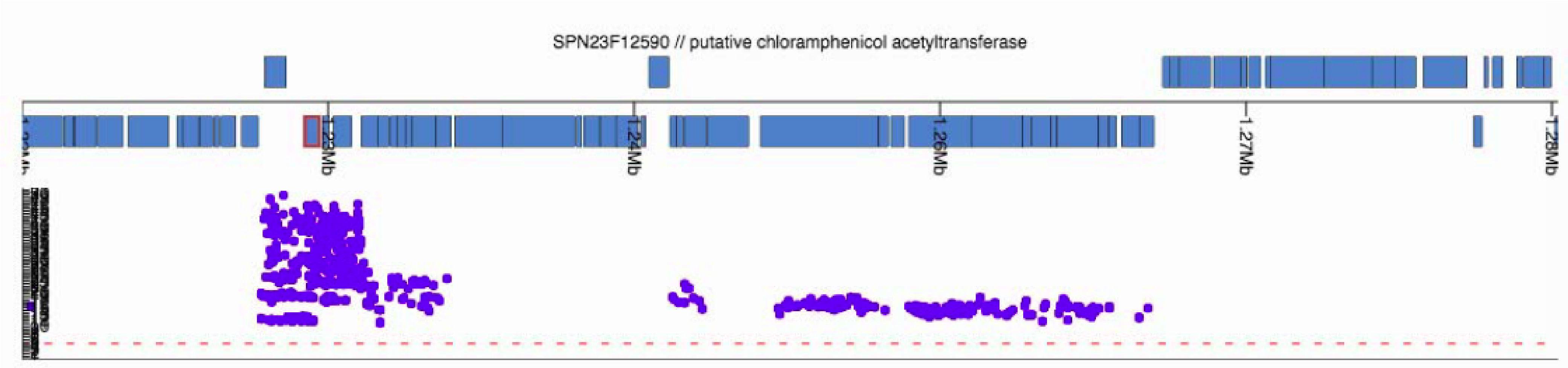
Application to simulated data

3,069 simulated *Streptococcus pneumoniae* genomes (program ALF)
Included accumulation of SNPs, indels, gene loss, recombination events estimated from real data.
Accessory gene was associated with phenotype over a range of odds-ratios

Fig. 1

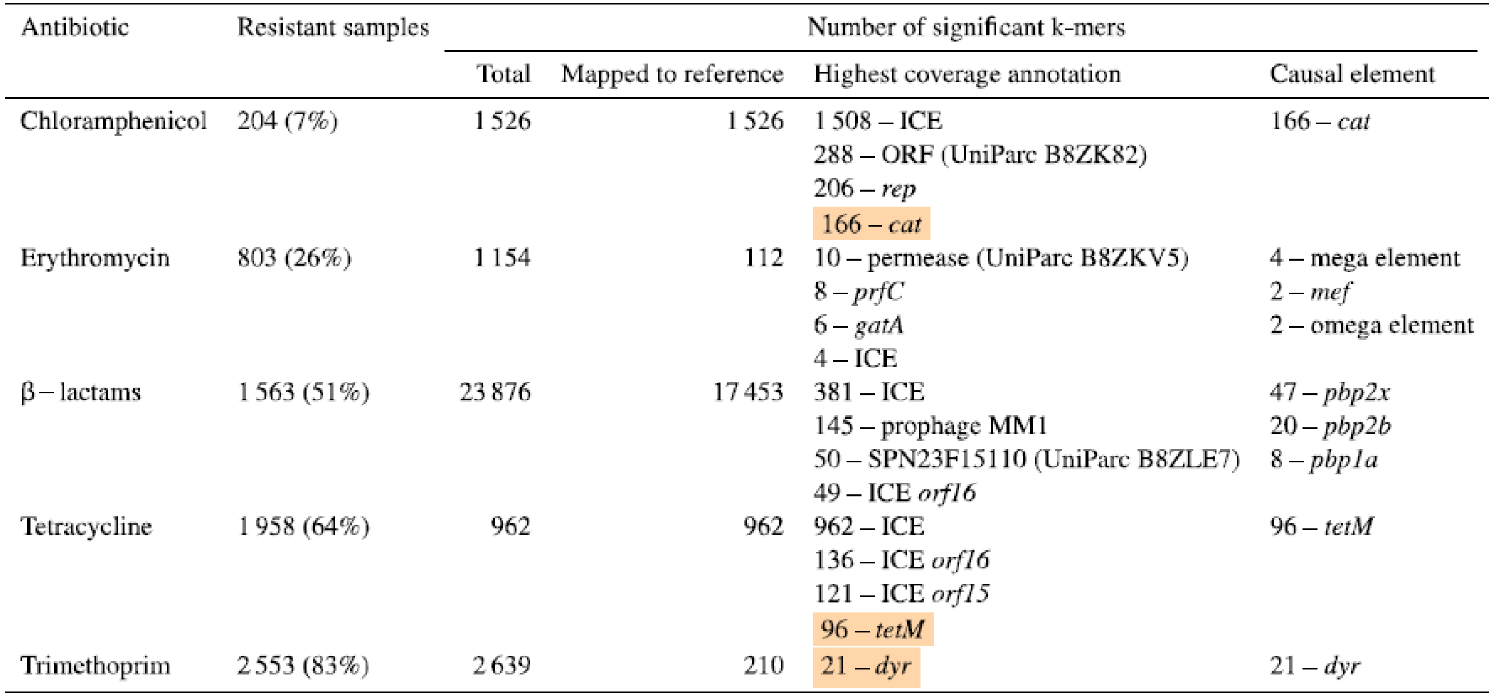


Conformation of known resistance mechanisms in a large population of *S. pneumonia*.



Supplementary figure 4: JScandy view of ATCC 700669 reference genome (blue blocks at top genes on forward and reverse strands) and Manhattan plot of start positions of the 1 508 of 1 526 k-mers significantly associated with chloramphenicol resistance which map to the integrative conjugative element (ICE) Tn*5253*. The hits are all in within the ICE, and the most significant hits cluster around the *cat* gene (which is outlined in red).

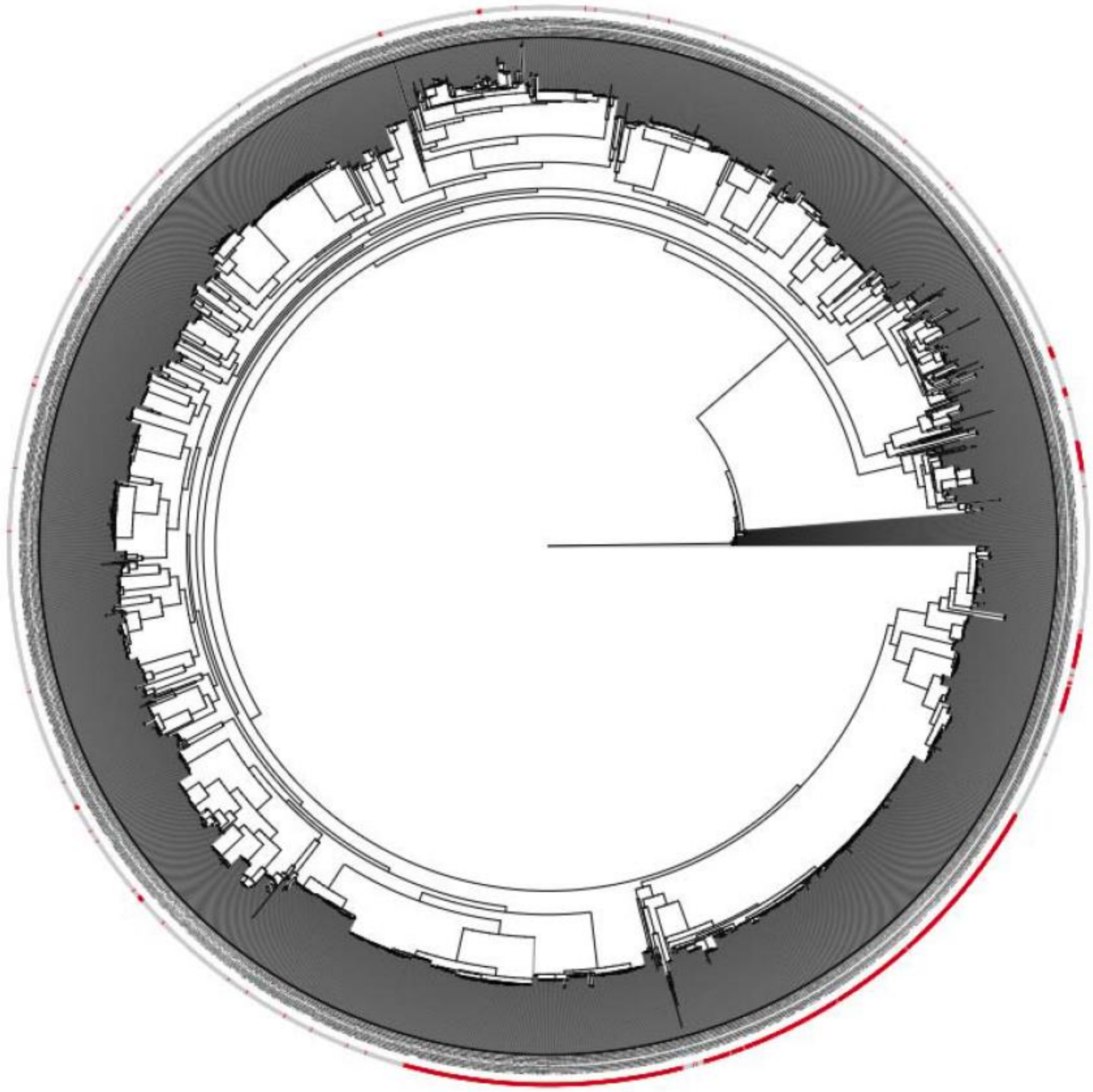
Tables



Antibiotic	Resistant samples	Number of significant k-mers			
		Total	Mapped to reference	Highest coverage annotation	Causal element
Chloramphenicol	204 (7%)	1 526	1 526	1 508 – ICE 288 – ORF (UniParc B8ZK82) 206 – *rep* 166 – *cat*	166 – *cat*
Erythromycin	803 (26%)	1 154	112	10 – permease (UniParc B8ZKV5) 8 – *prfC* 6 – *gatA* 4 – ICE	4 – mega element 2 – *mef* 2 – omega element
β – lactams	1 563 (51%)	23 876	17 453	381 – ICE 145 – prophage MM1 50 – SPN23F15110 (UniParc B8ZLE7) 49 – ICE *orf16*	47 – *pbp2x* 20 – *pbp2b* 8 – *pbp1a*
Tetracycline	1 958 (64%)	962	962	962 – ICE 136 – ICE *orf16* 121 – ICE *orf15* 96 – *tetM*	96 – *tetM*
Trimethoprim	2 553 (83%)	2 639	210	21 – *dyr*	21 – *dyr*

Table 1: Results from SEER for antibiotic resistance binary outcome on a population of 3069 *S. pneumoniae*. Significant k-mers are first interpreted by mapping to the ATCC 700669 reference genome. Up to the first four highest covered annotations are shown, and if the known mechanism is amongst these it is highlighted in orange. The ICE is the top hit in three analyses, as it carries multiple drug-resistance elements and is commonly found in multi-drug resistant strains[16]. The distribution of phenotype across the phylogeny is shown in Supplementary figure 5.

Resistance to erythromycin is conferred by multiple genes that can perform the same function



Supplementary figure 5: Neighbour joining tree from Chewapreecha *et. al.* study of 3069 Thai carriage *S. pneumoniae* samples, from a SNP alignment produced by mapping to the ATCC 700669 reference strain. Outer ring: red if resistant to Erythromycin, grey if sensitive.

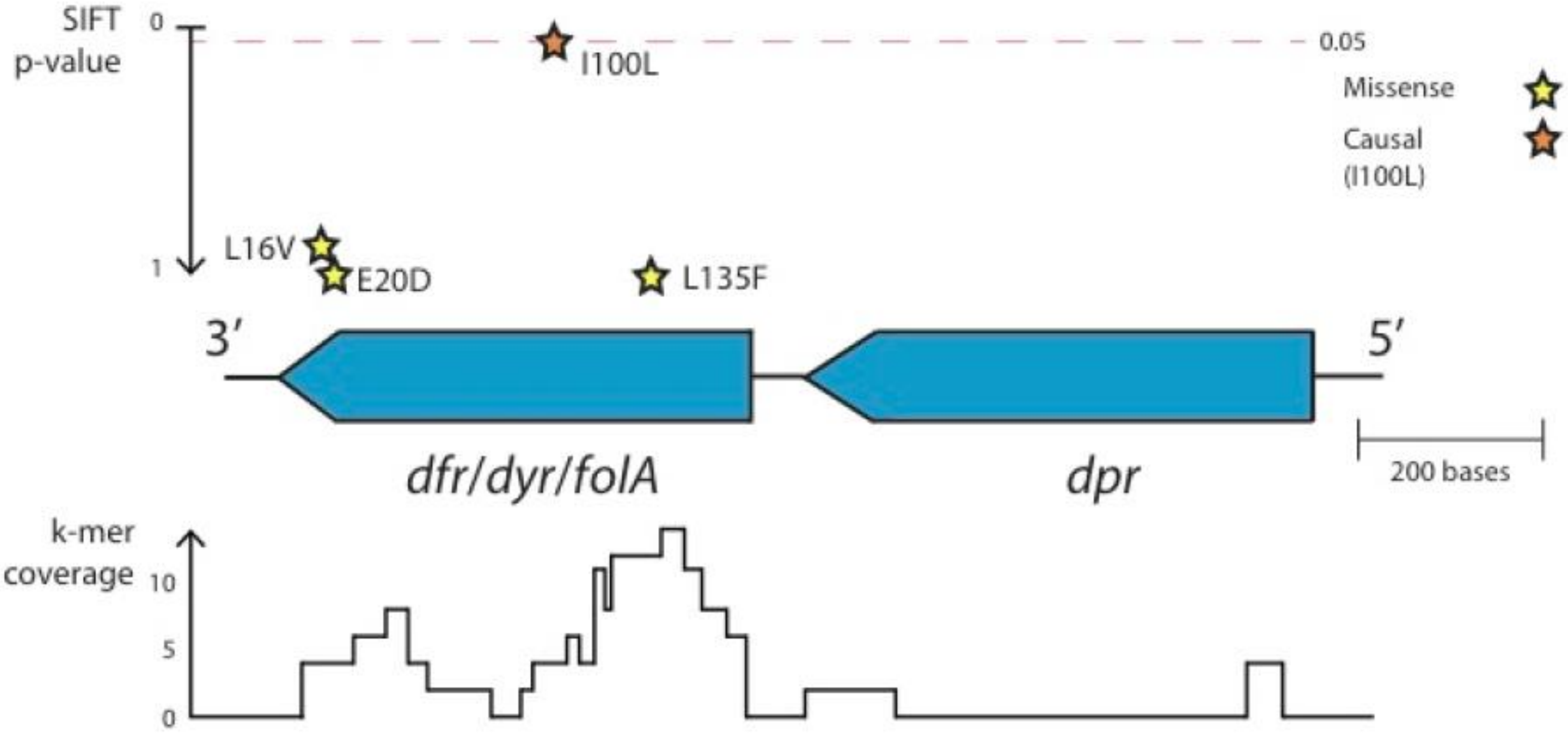


Fig. 2: Fine mapping trimethoprim resistance. The locus pictured contains 72 significant k-mers, the most of any gene cluster. Coverage over the locus is pictured at the bottom of the figure. Shown above the genes are high quality missense SNPs, plotted using their p-value for affecting protein function as predicted by SIFT.

Conclusion

SEER is capable of finding bacterial sequence elements associated with a range of phenotypes

SEER is reference sequence independent

SEER is able to analyze very large sample sizes