Quadram Institute Science Health Food Innovation

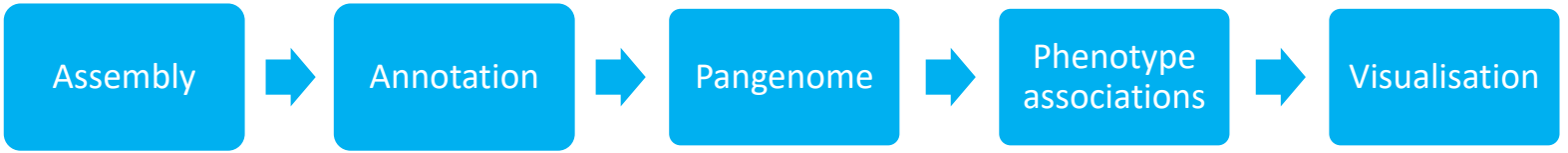
Pangenomes and MGEs

Andrew Page PhD

Step 1

Hire a bioinformatician

Reads to MGEs via Pangenome analysis



Skesa
Shovill
Spades

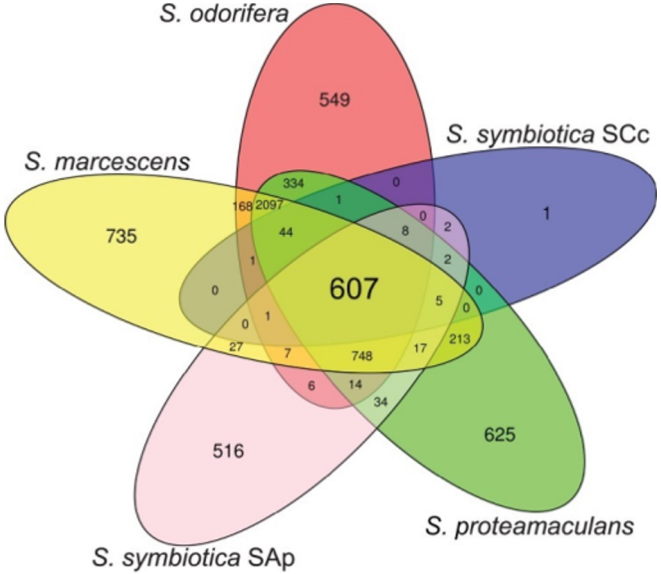
Prokka
Interproscan
NCBI PGAP

Roary
Panaroo

Scoary
pyseer

Anvi'o
PanX
Phandango

What is a pan genome?



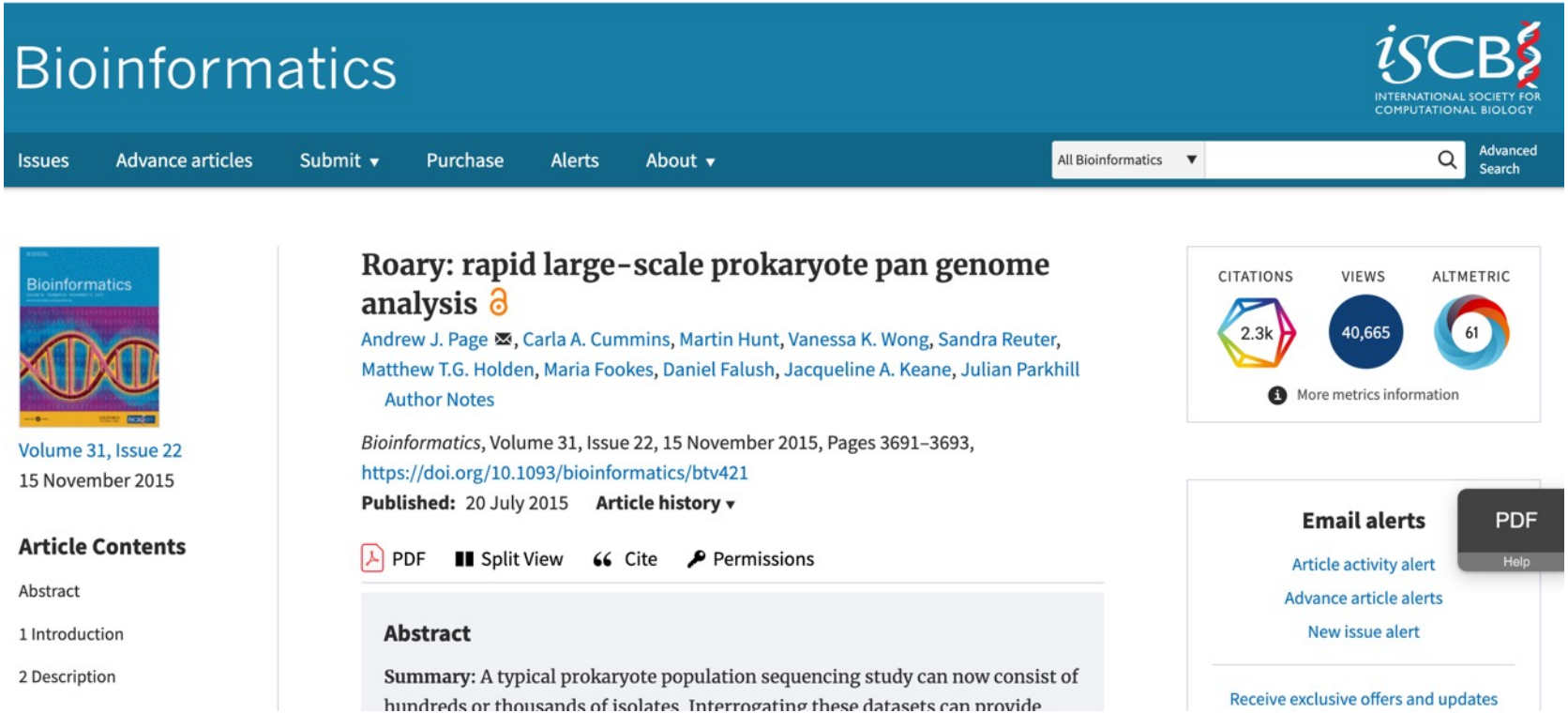
The core genome contains genes shared by all strains within a clade.
The accessory genome is made up of genes shared by a subset of the strains.

Vernikos,G. et al. (2014) Ten years of pan-genome analyses. Curr. Opin. Microbiol., 23C, 148–154.
http://openi.nlm.nih.gov/detailedresult.php?img=3471834_pone.0047274.g001&req=4

What can you do with a Pan Genome?

- High level overview of large datasets
- Identify novel mobile elements and large structural variants
- Presence/absence of drug resistance or virulence genes
- Identify contamination
- Draw a phylogenetic tree using just the core genes, without the need for a reference.
- Open or closed pan genome
- Identify Mobile Genetic Elements

Roary: for bacterial pangenome analysis



Bioinformatics

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Issues | Advance articles | Submit | Purchase | Alerts | About | All Bioinformatics | Advanced Search

Volume 31, Issue 22
15 November 2015

Article Contents

Abstract

1 Introduction

2 Description

Roary: rapid large-scale prokaryote pan genome analysis

Andrew J. Page, Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T.G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, Julian Parkhill

Author Notes

Bioinformatics, Volume 31, Issue 22, 15 November 2015, Pages 3691–3693,
https://doi.org/10.1093/bioinformatics/btv421

Published: 20 July 2015 **Article history**

PDF | Split View | Cite | Permissions

Abstract

Summary: A typical prokaryote population sequencing study can now consist of hundreds or thousands of isolates. Interrogating these datasets can provide

CITATIONS 2.3k | VIEWS 40,665 | ALTMETRIC 61

More metrics information

Email alerts

Article activity alert

Advance article alerts

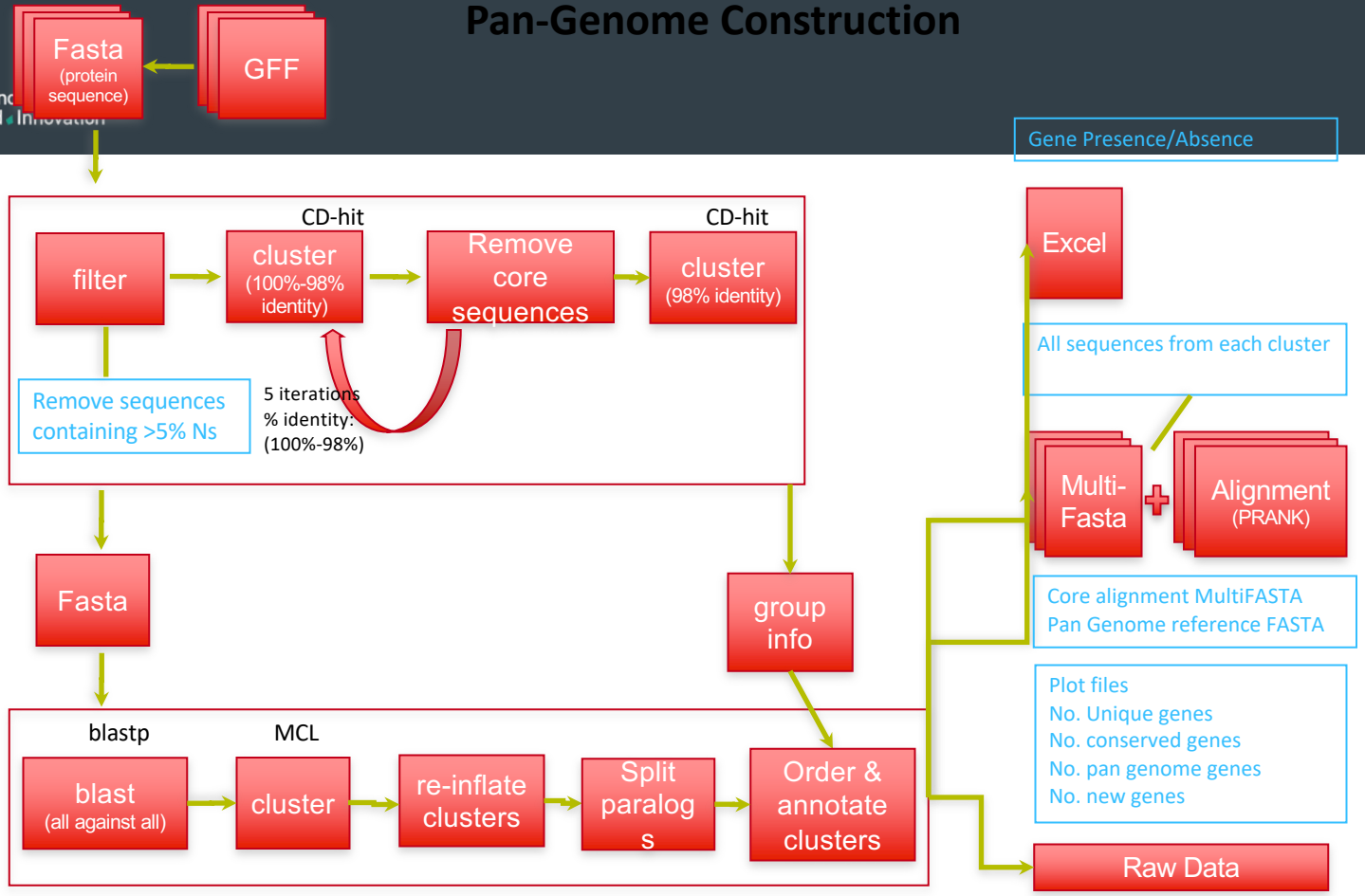
New issue alert

Receive exclusive offers and updates

PDF Help

*** Other software is available, but as I wrote this one, it's the example I will use.**

Pan-Genome Construction



Inputs and outputs

Inputs: Annotated de novo assemblies (GFF files)

Outputs:

- Spreadsheet with presence and absence of genes
- Multi-FASTA alignment of core genes so you can build a tree without a reference
- Multi-FASTA alignments for each gene
- Plots for the open/closed genome, unique genes
- Output format can be used with multiple downstream programs
- QC report from Kraken to help identify suspect samples

How to install and run it

conda install -c bioconda roary

*roary *.gff*

Splitting paralogs

Cluster with paralogs

```
Genome1_Seq1, Genome1_Seq2, Genome2_Seq3
```

Go back to the assemblies and look at 5 genes on either side

```
..->GeneB->GeneC->Genome1_Seq1->GeneD->GeneE->..
..->GeneV->GeneW->Genome1_Seq2->GeneX->GeneY->..
..->GeneB->GeneC->Genome2_Seq3->GeneD->GeneE->..
```

Split into 2 clusters

```
Genome1_Seq1, Genome2_Seq3
```

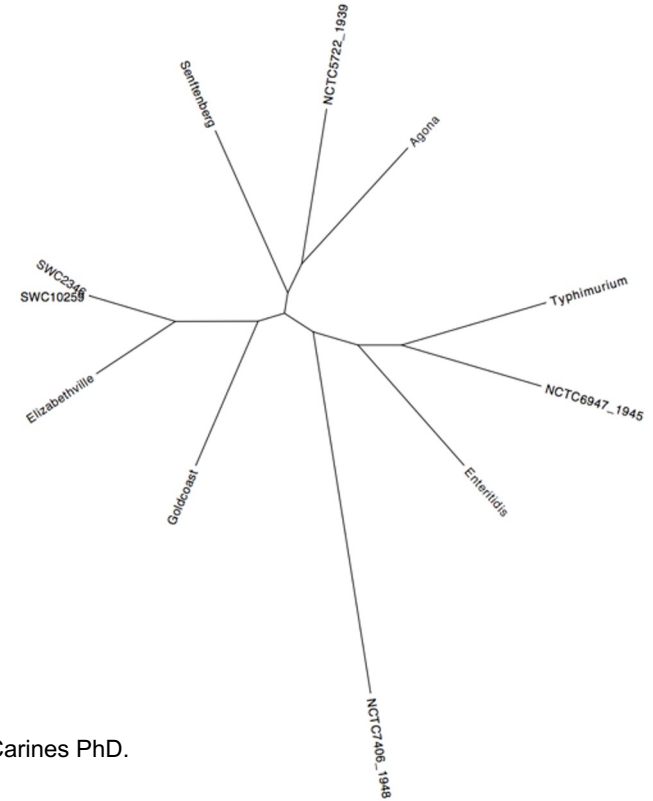
```
Genome1_Seq2
```

What can you do with Roary?

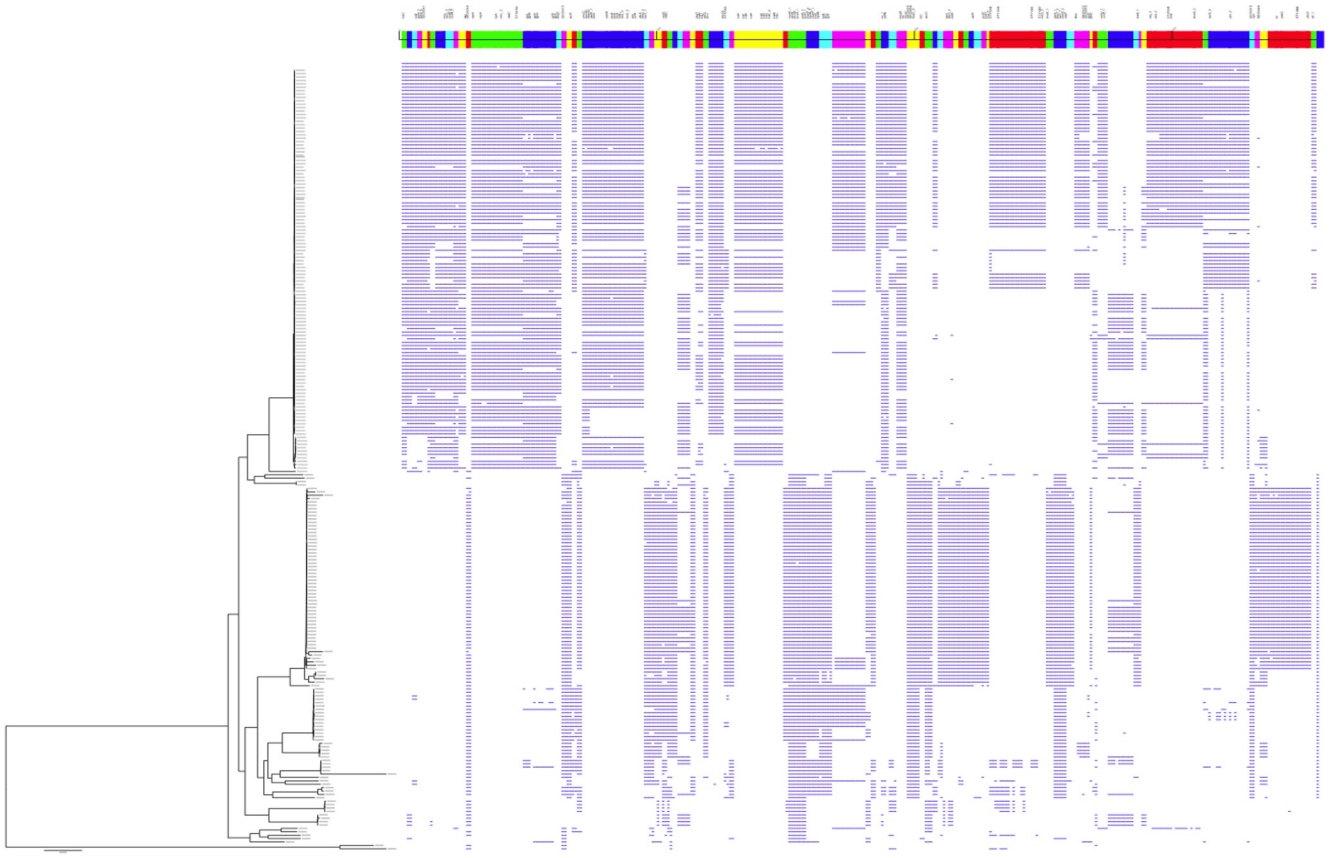
Reference free trees

How does Salmonella Weltevreden fit with other NCTC Salmonellas?

Took less than 10 minutes to get a reference free tree made from 3370 core genes (3,034,596 bases)

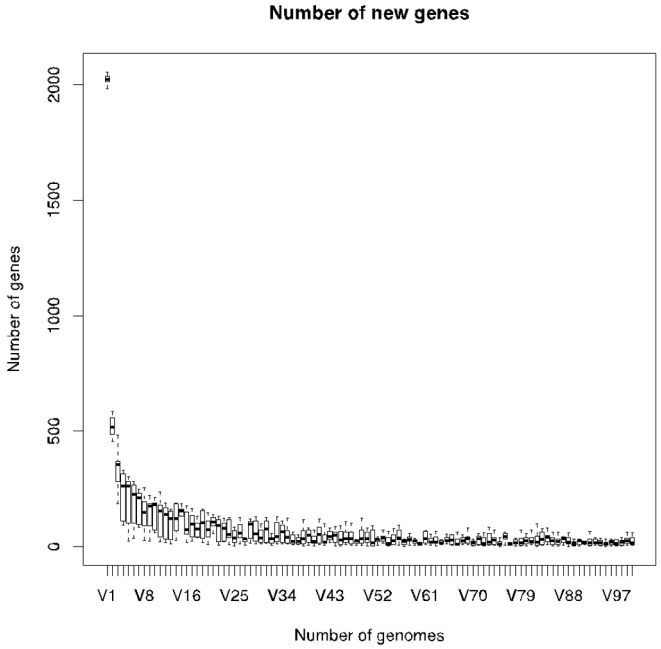


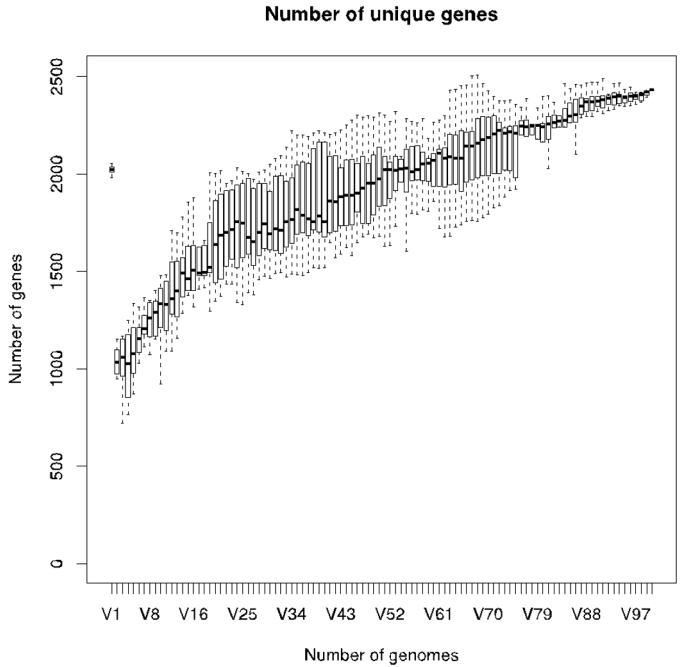
S. Welteverden and S. Elizabethville from FX as part of Carines PhD.
Everything else from NCTC 3000.



Phandango output of presence and absence of genes in accessory genome.
S. Weltevreden & public S. enterica genomes

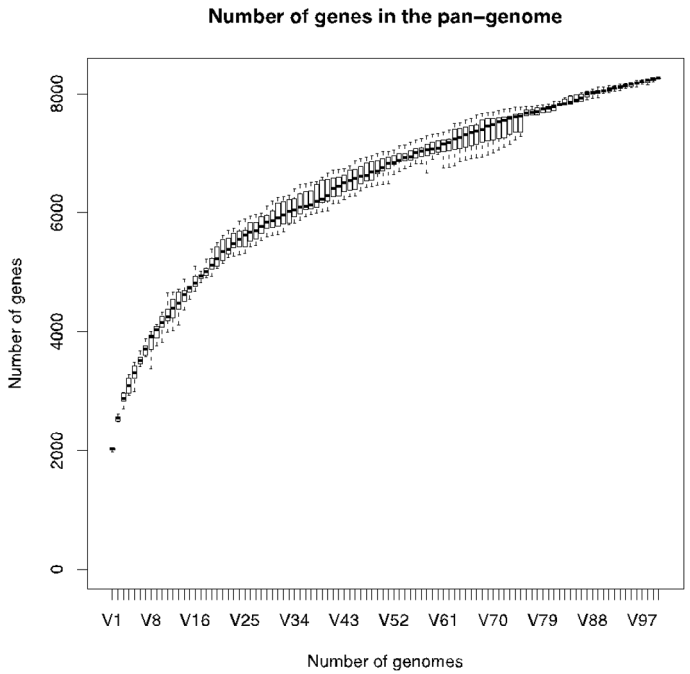
New and unique genes

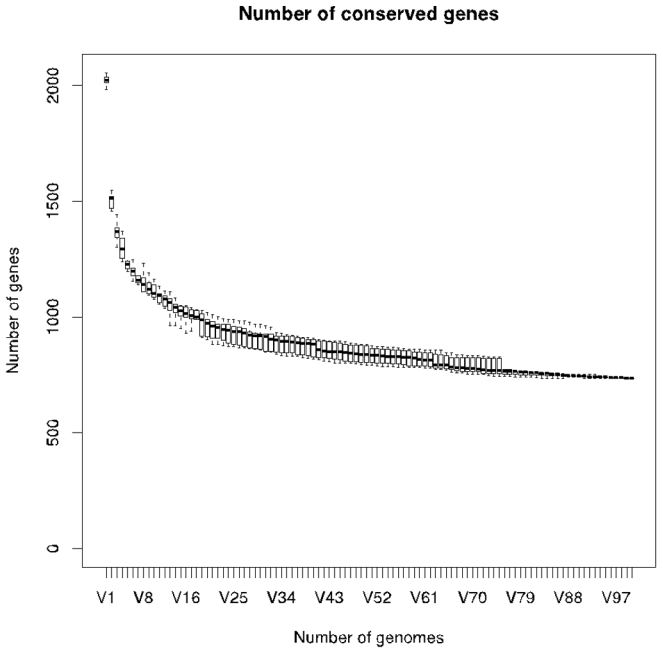




Subset of 100 S. pneumoniae from Nick Crouchers NatGen paper

Open/Closed Pan Genome





Subset of 100 S. pneumoniae from Crouchers NatGen paper.

What are we missing by using short reads

Salmonella Typhi

- 44 S. typhi have been sequenced on PacBio and on Illumina, so we can calculate the limitations of using highly fragmented Illumina assemblies.
- Illumina assemblies have ~5% less predicted genes than PacBio
- <1% of genes split over Illumina assembly contig boundaries

Run roary with 17 Staph aureus samples

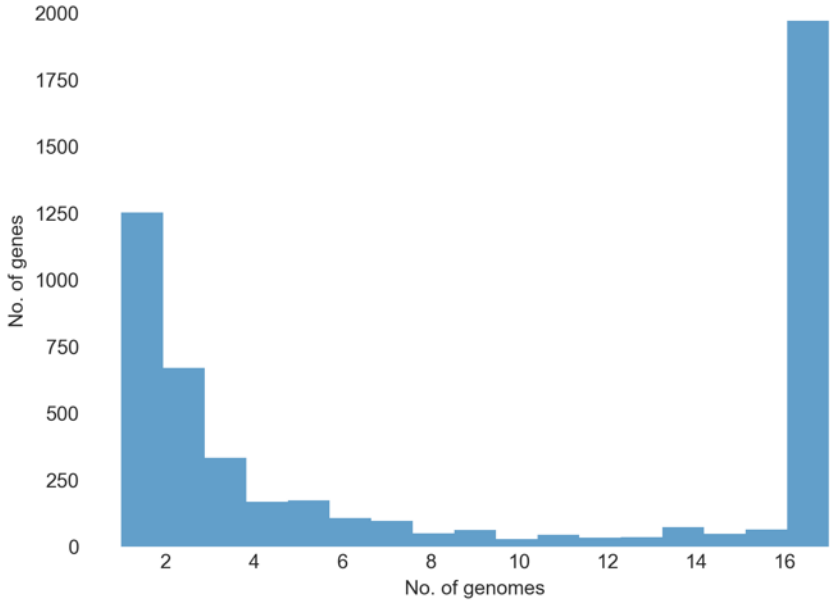
I randomly selected some samples from NCTC. No idea inadvance what is there or whats interesting…

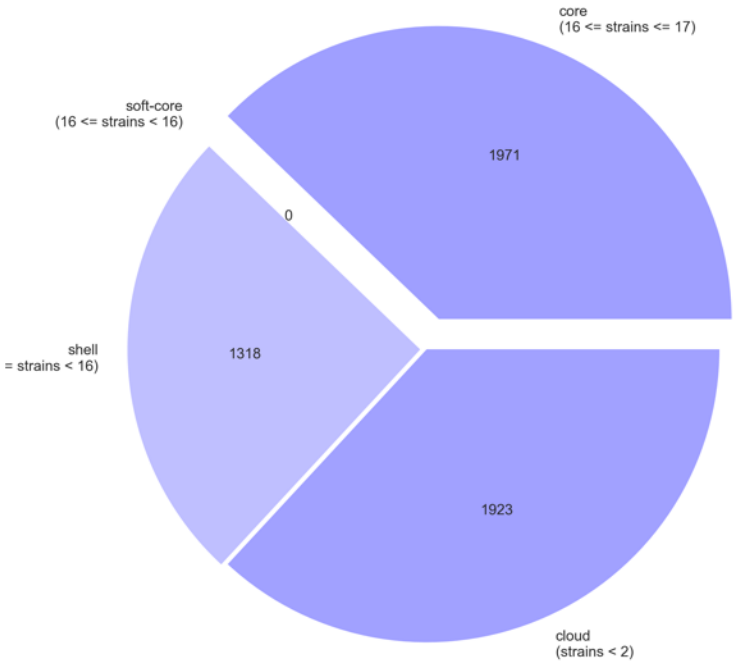
```
(base) N120294:roary_example pagea$ roary -p 4 *.gff
Use of uninitialized value in require at /Users/pagea/miniconda3/lib/perl5/site_perl/5.22.0/darwin-thread-multi-2level/Encode.pm line 61.

Please cite Roary if you use any of the results it produces:
    Andrew J. Page, Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T. G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, Julian Parkhill,
        "Roary: Rapid large-scale prokaryote pan genome analysis", Bioinformatics, 2015 Nov 15;31(22):3691-3693
    doi: http://doi.org/10.1093/bioinformatics/btv421
        Pubmed: 26198102

(base) N120294:roary_example pagea$ mv * ..
(base) N120294:roary_example pagea$ ~/code/Roary/contrib/roary_plots/roary_plots.py accessory_binary_genes.fa.newick gene_presence_absence.csv
(base) N120294:roary_example pagea$ ~/code/Roary/contrib/roary2svg/roary2svg.pl --acconly gene_presence_absence.csv > pangenome.svg
Found 17 taxa: ERS811724 ERS811726 ERS811727 ERS811728 ERS811729 ERS811730 ERS811733 ERS811737 ERS811738 ERS811740 ERS812507 ERS819824 ERS825151 ERS825158 ERS825162 ERS825165 ERS825170
Found 3241 clusters.
Box = 0.264116013576057 x 20 px
Left label = 9 chr x 15 px
Right label = 3 chr x 15 px
Writing SVG file
Done.
(base) N120294:roary_example pagea$ ls
ERS811724.gff                       ERS811738.gff                       ERS825170.gff                       core_accessory.header.embl          number_of_unique_genes.Rtab
ERS811726.gff                       ERS811740.gff                       accessory.header.embl               core_accessory.tab                  pangenome.svg
ERS811727.gff                       ERS812507.gff                       accessory.tab                       core_accessory_graph.dot            pangenome_frequency.png
ERS811728.gff                       ERS819824.gff                       accessory_binary_genes.fa           gene_presence_absence.Rtab          pangenome_matrix.png
ERS811729.gff                       ERS825151.gff                       accessory_binary_genes.fa.newick    gene_presence_absence.csv           pangenome_pie.png
ERS811730.gff                       ERS825158.gff                       accessory_graph.dot                 number_of_conserved_genes.Rtab      summary_statistics.txt
ERS811733.gff                       ERS825162.gff                       blast_identity_frequency.Rtab       number_of_genes_in_pan_genome.Rtab
ERS811737.gff                       ERS825165.gff                       clustered_proteins                  number_of_new_genes.Rtab
(base) N120294:roary_example pagea$
```

Data from: https://www.sanger.ac.uk/resources/downloads/bacteria/nctc/


2000
1750
1500
1250
1000
750
500
250
0
No. of genes
2
4
6
8
10
12
14
16
No. of genomes

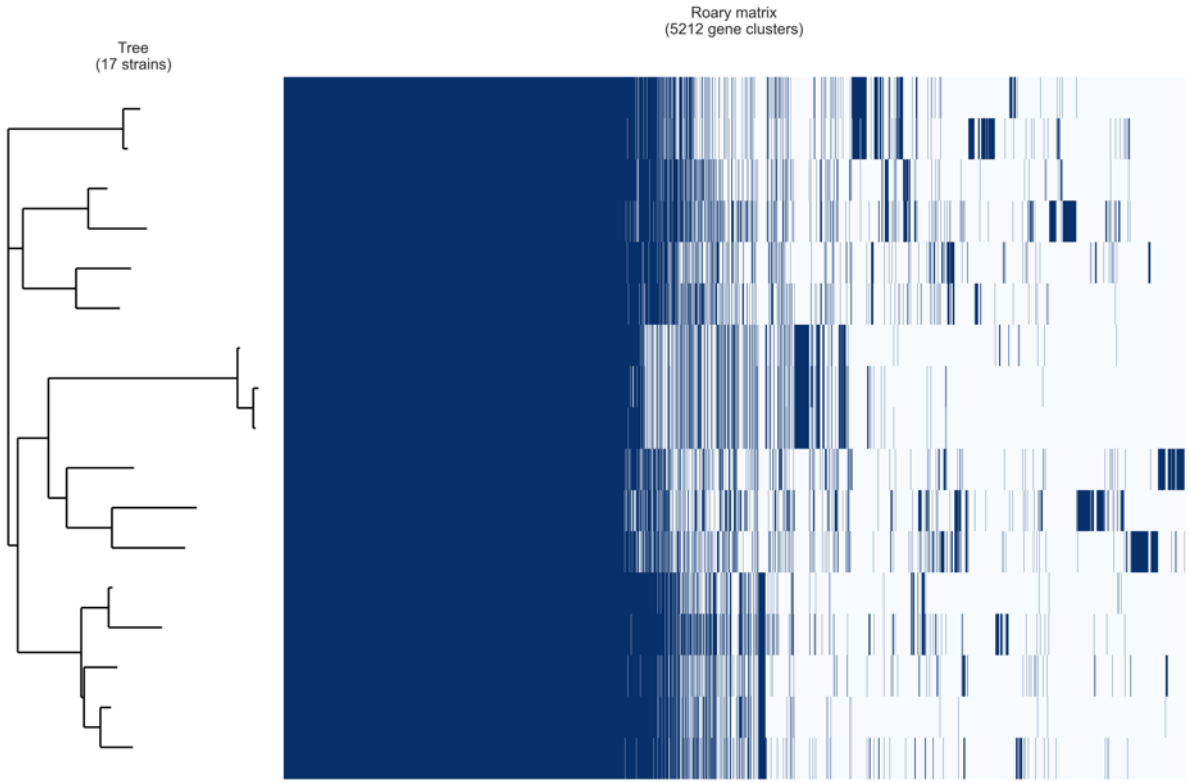

core
(16 <= strains <= 17)
soft-core
(16 <= strains < 16)
1971
0
shell
= strains < 16)
1318
1923
cloud
(strains < 2)

Manual vs Automated

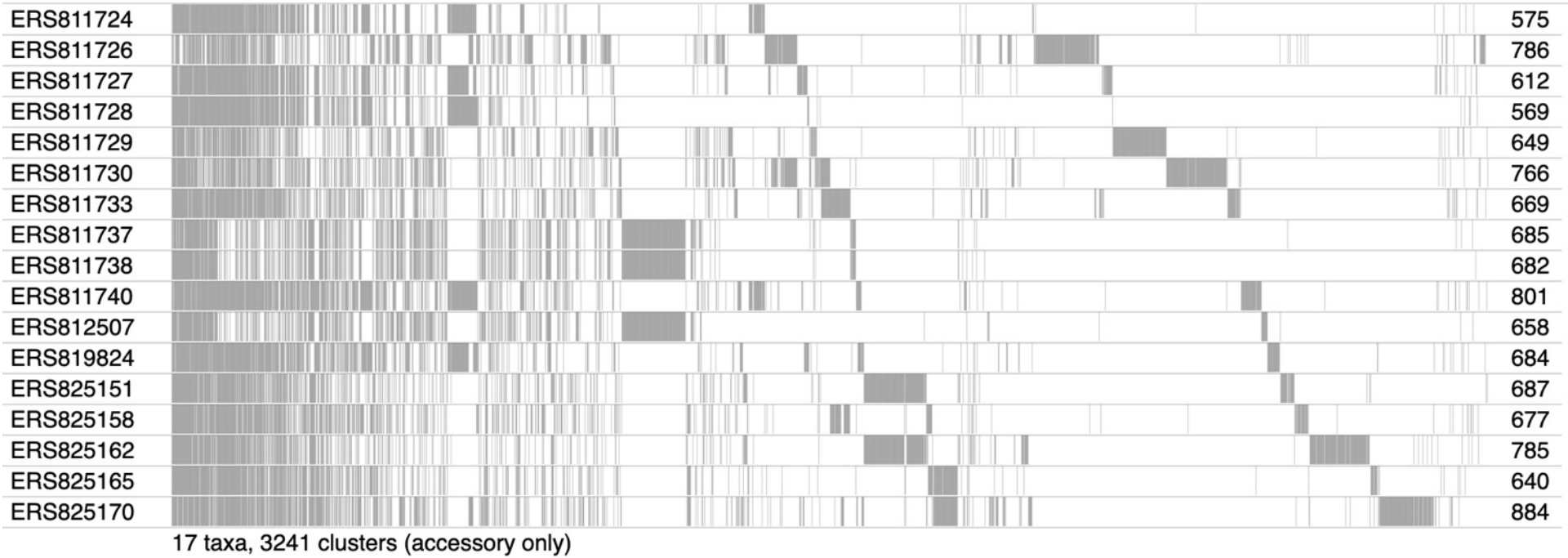
Manual
Humans can see patterns
Always eyeball data incase of obvious errors
Can go on a fishing expedition

Automated
Fast
Usually need other data (like phenotype)

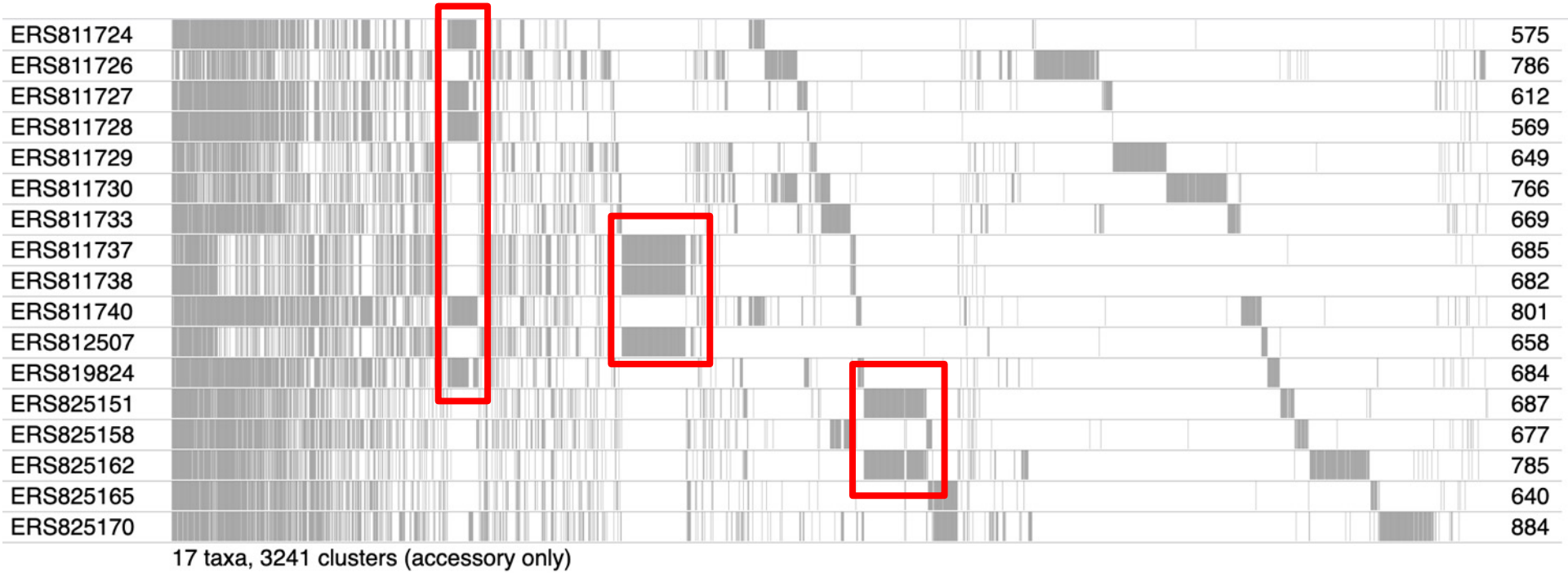
Manual analysis



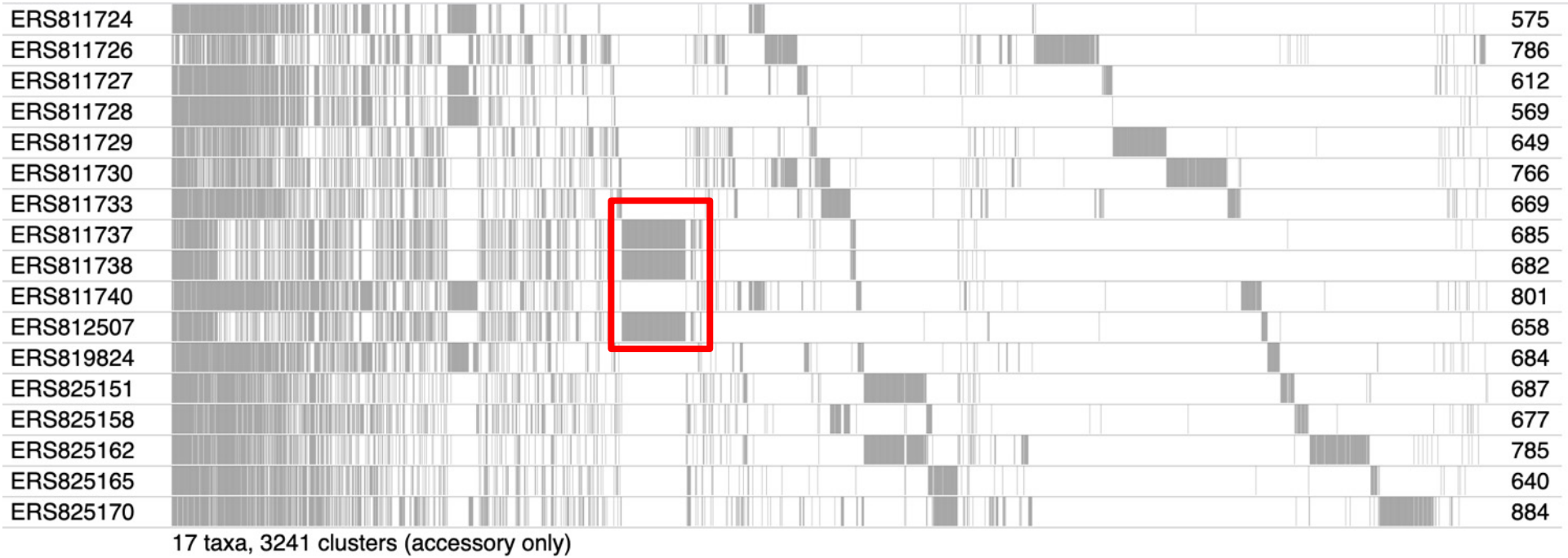
Just the accessory



Just the accessory – interesting commonalities



Lets pull out one



Gene presence & absence spreadsheet (Excel or R)

Gene	Non-unique Gene nan	Annotation	No. isolat	No. sequenc	Avg sequences per isola	Genome Fragme	Order within Fragme	Acce
group_3054	seo	enterotoxin O	3	3	1	1	3549	
group_489		Hypothetical protein	3	3	1	1	3508	
isdA		iron-regulated heme-iron binding protein	3	3	1	1	2364	
group_3019		transcriptional activator rinB-like protein	4	4	1	1	2766	
group_3058		transcriptional activator RinB family protein	4	4	1	1	2115	
group_2186		putative lipoprotein	4	4	1	1	2514	
group_963		lipoprotein	4	4	1	1	2515	
group_3022		phage protein	3	3	1	1	2060	
group_3027		ORF060	3	3	1	1	1546	
group_3026		Phage antirepressor protein	3	3	1	1	1547	
group_3023		phage protein	3	3	1	1	2059	
group_3024		phage protein	3	3	1	1	2058	
group_3025		ORF065	3	3	1	1	2057	
group_242		Phage antirepressor protein	4	4	1	1	1548	
dpnM		DNA adenine methylase	3	3	1	1	2560	
group_3064		helix-turn-helix family protein	3	3	1	1	2565	
group_3063		prophage L54a, antirepressor	3	3	1	1	2568	
group_3062		putative phi PVL-like protein	3	3	1	1	2567	
group_3061		putative phi PVL-like protein	3	3	1	1	2566	
group_3067		Leukocidin S subunit	3	3	1	1	2550	
group_3066		leukocidin-protein 1	3	3	1	1	2551	
group_3057	map_3	MHC class II antigen-like protein	3	3	1	1	2919	

Basic excel filtering

Can see cluster present in our 3 genomes

ERS811737	ERS811738	ERS811740	ERS812507	ERS819824	ERS825151	ERS8
NCTC7445_01806	NCTC7446_01802		NCTC2669_01787			
NCTC7445_01799	NCTC7446_01795		NCTC2669_01780			
NCTC7445_01027	NCTC7446_01025		NCTC2669_01015			
NCTC7445_01434	NCTC7446_01431		NCTC2669_01416			
NCTC7445_01986	NCTC7446_01983	NCTC8399_00482	NCTC2669_01966			
NCTC7445_01466	NCTC7446_01462		NCTC2669_01448			
NCTC7445_01465	NCTC7446_01461		NCTC2669_01447			
NCTC7445_01452	NCTC7446_01448		NCTC2669_01434			
NCTC7445_01458	NCTC7446_01454		NCTC2669_01440			
NCTC7445_01457	NCTC7446_01453		NCTC2669_01439			
NCTC7445_01453	NCTC7446_01449		NCTC2669_01435			
NCTC7445_01454	NCTC7446_01450		NCTC2669_01436			
NCTC7445_01455	NCTC7446_01451		NCTC2669_01437			
NCTC7445_01456	NCTC7446_01452		NCTC2669_01438			
NCTC7445_02018	NCTC7446_02015		NCTC2669_01999			
NCTC7445_02016	NCTC7446_02013		NCTC2669_01997			
NCTC7445_02014	NCTC7446_02011		NCTC2669_01995			
NCTC7445_02013	NCTC7446_02010		NCTC2669_01994			
NCTC7445_02012	NCTC7446_02009		NCTC2669_01993			
NCTC7445_02022	NCTC7446_02019		NCTC2669_02003			
NCTC7445_02021	NCTC7446_02018		NCTC2669_02002			
NCTC7445_01952	NCTC7446_01948		NCTC2669_01933			
NCTC7445_00315	NCTC7446_00313		NCTC2669_00314			
NCTC7445_00314	NCTC7446_00312		NCTC2669_00313			
NCTC7445_00769	NCTC7446_00766		NCTC2669_00755			
NCTC7445_00767	NCTC7446_00764		NCTC2669_00754			
NCTC7445_00761	NCTC7446_00758		NCTC2669_02013			
NCTC7445_00766	NCTC7446_00763					
NCTC7445_00344	NCTC7446_00342		NCTC2669_00343			
NCTC7445_02041	NCTC7446_02038	NCTC8399_02165				
NCTC7445_02040	NCTC7446_02037	NCTC8399_02164				
NCTC7445_02039	NCTC7446_02036	NCTC8399_02163				
NCTC7445_02036	NCTC7446_02033	NCTC8399_02160				

Interesting annotation

Phage
Mobile element
Pathogenicity island protein

Gene	Non-unique Gene nam	Annotation	No. isola
group_3054	seo	enterotoxin O	
group_489		Hypothetical protein	
isdA		iron-regulated heme-iron binding protein	
group_3019		transcriptional activator rinB-like protein	
group_3058		transcriptional activator RinB family protein	
group_2186		putative lipoprotein	
group_963		lipoprotein	
group_3022		phage protein	
group_3027		ORF060	
group_3026		Phage antirepressor protein	
group_3023		phage protein	
group_3024		phage protein	
group_3025		ORF065	
group_242		Phage antirepressor protein	
dpnM		DNA adenine methylase	
group_3064		helix-turn-helix family protein	
group_3063		prophage L54a, antirepressor	
group_3062		putative phi PVL-like protein	
group_3061		putative phi PVL-like protein	
group_3067		Leukocidin S subunit	
group_3066		leukocidin-protein 1	
group_3057	map_3	MHC class II antigen-like protein	
group_2941	xerD_1	integrase	
group_2502		abi-like family protein	
group_719		putative lipoprotein	
group_2981		membrane protein	
group_2324		pahtogenicity island protein	
group_2327		Exfoliative toxin A/B	
group_825		pathogenicity island protein	
group_494		mobile element-associated protein	
group_826		pathogenicity island protein	
group_824		pathogenicity island protein	
group_300		transposase	

Lets look at that pathogenicity island

Gene	Non-	Annotation	No. isolat	No. sequenc	Avg sequences per isola	Genome Fragme	Order within Fragme	Accessory Fragme	Accessory Order with Fragme
group_2945		pathogenicity island protein	3	3	1	1	1632	1	897
group_2946		hypothetical bovine pathogenicity island protein	3	3	1	1	1631	1	898
group_2948		pathogenicity island protein	3	3	1	1	1629	1	900
group_2951		pathogenicity island protein	3	3	1	1	3011	1	1013
group_3049		pathogenicity island protein	3	3	1	1	827	2	149

Accessory fragment: a graph of gene clusters is created based on which genes are beside each other on contigs in the assemblies. Core genes are removed. Disjointed graphs are created, and graphs are walked. These pathogenicity island proteins are on the same accessory fragment (graph) and have consecutive numbers indicating they are probably usually found together in assemblies.

Lets look at that pathogenicity island

Gene	Non-	Annotation	No. isolat	No. sequenc	Avg sequences per isola	Genome Fragme	Order within Fragme	Accessory Fragme	Accessory Order with Fragme
group_2945		pathogenicity island protein	3	3	1	1	1632	1	897
group_2946		hypothetical bovine pathogenicity island protein	3	3	1	1	1631	1	898
group_2948		pathogenicity island protein	3	3	1	1	1629	1	900
group_2951		pathogenicity island protein	3	3	1	1	3011	1	1013
group_3049		pathogenicity island protein	3	3	1	1	827	2	149

V	W	X	Y
ERS811737	ERS811738	ERS811740	ERS812507
NCTC7445_00339	NCTC7446_00337		NCTC2669_00338
NCTC7445_00340	NCTC7446_00338		NCTC2669_00339
NCTC7445_00342	NCTC7446_00340		NCTC2669_00341
NCTC7445_00354	NCTC7446_00352		NCTC2669_00353
NCTC7445_01786	NCTC7446_01782		NCTC2669_01766

Accessory fragment: a graph of gene clusters is created based on which genes are beside each other on contigs in the assemblies. Core genes are removed. Disjointed graphs are created, and graphs are walked. These pathogenicity island proteins are on the same accessory fragment (graph) and have consecutive numbers indicating they are probably usually found together in assemblies.

Eyeball region in 1 sample in Artemis

Eyeball region in 1 sample

All genes on same strand

Annotation of island starts with helix-turn-helix, toxin genes slightly upstream

Toxin genes just upstream

NCTC7445_00363	402773	403750		staphylococcal exotoxin,superantigen-like protein,Domain of unknown function (DUF1954)
NCTC7445_00364	404111	404815		exotoxin,superantigen-like protein 5,Domain of unknown function (DUF1954)
NCTC7445_00365	405262	405957		exotoxin 1,Toxic shock syndrome toxin-1 precursor,superantigen-like protein 7,Domain of unknown function (DUF1954)
NCTC7445_00366	406298	406996		exotoxin 5,SPE C,superantigen-like protein,Domain of unknown function (DUF1954)
NCTC7445_00367	407359	408042		superantigen-like protein, exotoxin 14,Toxic shock syndrome toxin-1 precursor,superantigen-like protein,Domain of unknown function (DUF1954)
NCTC7445_00368	408304	409860		Type I restriction enzyme EcoprrI M protein,Type I restriction enzyme EcoKI M protein,N5-glutamine S-adenosyl-L-methionine-dependent methyltr
NCTC7445_00369	409853	411085		type I restriction modification DNA specificity domain-containing protein,EcoKI restriction-modification system protein HsdS,Type I restricti
NCTC7445_00370	411468	412166		exotoxin,superantigen-like protein,Domain of unknown function (DUF1954)
NCTC7445_00371	412185	413672		exported protein
NCTC7445_00372	413778	414086	c	exported protein

Get the sequence

Highlight and copy

Automated analysis with Scoary

```
,TSS
ERS811724,0
ERS811726,0
ERS811727,0
ERS811728,0
ERS811729,0
ERS811730,0
ERS811733,0
ERS811737,1
ERS811738,1
ERS811740,0
ERS812507,1
ERS819824,0
ERS825151,0
ERS825158,0
ERS825162,0
ERS825165,0
ERS825170,0
```

```
(base) N120294:roary_example pagea$ scoary -t traits.csv -g gene_presence_absence.csv
==== Scoary started ====
Command: /Users/pagea/miniconda3/bin/scoary -t traits.csv -g gene_presence_absence.csv
Reading gene presence absence file
Creating Hamming distance matrix based on gene presence/absence
Building UPGMA tree from distance matrix
Reading traits file
Finished loading files into memory.


==== Performing statistics ====
-- Filtration options --
Individual (Naive):    0.05
Collapse genes:    False


Tallying genes and performing statistical analyses
Gene-wise counting and Fisher's exact tests for trait: TSS
100.00%
Adding p-values adjusted for testing multiple hypotheses

Storing results: TSS
Calculating max number of contrasting pairs for each nominally significant gene
100.00%
Storing results to file


==== Finished ====
Checked a total of 5212 genes for associations to 1 trait(s). Total time used: 1 seconds.
No warnings were recorded.
```

Pick same 3 samples

Which genes aren't in the 3 of interest?

Gene	N	Annotation	Numbe	Number_neg_present	Number_pos_not_present	Number_neg_not_present	Sensitivity	Specificity	Odds_ra	Naive_p	Bonferroni	Benjamini_H
ebh_2		LPXTG-motif cell wall anchor domain-containing protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799
esaC_1		EsaC protein within ESAT-6 gene cluster	0	14	3	0	0	0	0	0.00147059	1	0.020632799
esxB		ESAT-6/Esx family secreted protein EsxB	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_1215		Hypothetical protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799
isdA_2		iron-regulated heme-iron binding protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_1288		Phosphoesterase family protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_1290		Exotoxin	0	14	3	0	0	0	0	0.00147059	1	0.020632799
lukD		Leukotoxin LukD	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_1403		gamma-hemolysin component B	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_1460		lipoprotein, putative	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_148		Putative cytosolic protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_150		membrane spanning protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_1612		Hypothetical protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_1851		FIG01108246: hypothetical protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_1862		Putative cytosolic protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799
essC		FtsK/SpoIIIE family protein, putative secretion system component EssC/YukA	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_1866		Hypothetical protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_1901		Hypothetical protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799
xerD_1		DNA integration/recombination/invertion protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799
mrpA		Na( ) H( ) antiporter subunit A	0	14	3	0	0	0	0	0.00147059	1	0.020632799
bceS		Two-component sensor histidine kinase BceS	0	14	3	0	0	0	0	0.00147059	1	0.020632799
macB_1		ABC transporter ATP-binding protein VraF	0	14	3	0	0	0	0	0.00147059	1	0.020632799
vraG		Bacitracin export permease protein BceB	0	14	3	0	0	0	0	0.00147059	1	0.020632799
fmtA_1		FmtA protein involved in methicillin resistance; affects cell wall cross-linking and amidation	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_1918		Hypothetical protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799
pheT_2		tRNA-binding domain-containing protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_1989		comK family protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_1990		putative staphylococcal protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799
arg		Arginase	0	14	3	0	0	0	0	0.00147059	1	0.020632799
bmr3		Multidrug resistance protein B	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_2025		acetyltransferase	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_2057		small heat shock protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799
aur		Zinc metalloproteinase precursor / aureolysin	0	14	3	0	0	0	0	0.00147059	1	0.020632799
isaB		immunodominant antigen B	0	14	3	0	0	0	0	0.00147059	1	0.020632799
hlgA_1		Leukocidin S subunit lukE	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_2230		Leukocidin S subunit	0	14	3	0	0	0	0	0.00147059	1	0.020632799
cbiX		sirohydrochlorin ferrochelatase	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_410		staphylococcal enterotoxin, putative	0	14	3	0	0	0	0	0.00147059	1	0.020632799
tst_2		Exotoxin	0	14	3	0	0	0	0	0.00147059	1	0.020632799
ssp		Extracellular ECM and plasma binding protein Emp	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_446		acyltransferase family protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799
group_450		FIG01107916: hypothetical protein	0	14	3	0	0	0	0	0.00147059	1	0.020632799

You will get random noise – more samples help (and more diverse)

Which genes are only in the 3 of interest

group_2917		proline rich protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2918		membrane protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2919	emrI	transporter protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2920		N-acetyltransferase family protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2921		N-acetyltransferase	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2922		DNA-binding protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2923		TfoX domain-containing protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2924		FIG01108246: hypothetical protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2925	yecD	isochorismatase	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2926		staphylocoagulase	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2927		glutamine amidotransferase class-I	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2928		Putative cytosolic protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2929	essC	essC	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2930		type VII secretion effector	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2931		hypothetical protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2932		putative transposase	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2933		membrane protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2934		hypothetical protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2935		hypothetical protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2936		membrane protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2937		Putative cytosolic protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2938		Repetitive hypothetical protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2940		membrane protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2941	xerD	integrase	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2942		membrane spanning protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2943		putative regulatory protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2944		Helix-turn-helix domain	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2945		pathogenicity island protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2946		hypothetical bovine pathogenicity island protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2948		pathogenicity island protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2949		abortive infection bacteriophage resistance protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2951		pathogenicity island protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2952		exotoxin	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2954		exotoxin	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
set1		exotoxin 1	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
set5		exotoxin 5	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
set14		superantigen-like protein, exotoxin 14	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2958		type I restriction modification DNA specificity domain-containing protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2959		exotoxin	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2960		exported protein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2961	lpl2_	putative lipoprotein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799
group_2962	lpl9_	staphylococcal tandem lipoprotein	3	0	0	14	100	100	inf	0.00147059	1	0.020632799

Expanding your fishing expedition

Goal: binary traits (yes/no, case/control, present/absent)

- **Metadata**
  - Geographical signal
  - Age/gender, sampling location or bodysite
  - Timeframe (years, decades, seasons etc…)
- **Genotype**
  - AMR genes predicted
  - Virulence genes
  - Plasmid rep/inc types
- **Phenotype**
  - Anti-biogram
  - Growth curves,….. or anything else you find experimentally

Using Genotypic information

Run all assemblies through abricate with NCBI AMR database

abricate --db ncbi ERS825151.fa > ERS825151.fa.amr.tab

Combine all AMR results into a single file:
abricate --summary *.amr.tab

#FILE	NUM_FOUND	ant(6)-Ia	ant(9)-Ia	aph(3')-IIIa	blaI	blaPC1	blaR1	blaZ	erm(A)	fosD	mecA	mecR1	sat4	tet(38)	tet(K)	tet(M)
ERS811724.fa.amr.tab	1	.	.	.	.	.	.	.	.	.	.	100.00	.	.		
ERS811726.fa.amr.tab	2	.	.	.	.	.	.	100.00	.	.	.	100.00	.	.		
ERS811727.fa.amr.tab	1	.	.	.	.	.	.	.	.	.	.	100.00	.	.		
ERS811728.fa.amr.tab	1	.	.	.	.	.	.	.	.	.	.	100.00	.	.		
ERS811729.fa.amr.tab	1	.	.	.	.	.	.	.	.	.	.	100.00	.	.		
ERS811730.fa.amr.tab	1	.	.	.	.	.	.	.	.	.	.	100.00	.	.		
ERS811733.fa.amr.tab	2	.	.	.	.	.	.	100.00	.	.	.	100.00	.	.		
ERS811737.fa.amr.tab	2	.	.	.	.	.	.	100.00	.	.	.	100.00	.	.		
ERS811738.fa.amr.tab	2	.	.	.	.	.	.	100.00	.	.	.	100.00	.	.		
ERS811740.fa.amr.tab	1	.	.	.	.	.	.	.	.	.	.	100.00	.	.		
ERS812507.fa.amr.tab	2	.	.	.	.	.	.	100.00	.	.	.	100.00	.	.		
ERS819824.fa.amr.tab	1	.	.	.	.	.	.	.	.	.	.	100.00	.	.		
ERS825151.fa.amr.tab	2	.	.	.	.	.	.	100.00	.	.	.	100.00	.	.		
ERS825158.fa.amr.tab	5	.	.	100.00	.	100.00	100.00	100.00	.	.	.	100.00	.	.		
ERS825162.fa.amr.tab	2	.	.	.	.	.	.	100.00	.	.	.	100.00	.	.		
ERS825165.fa.amr.tab	5	.	.	100.00	.	100.00	100.00	100.00	.	.	.	100.00	.	.		
ERS825170.fa.amr.tab	14	100.00;100.00	100.00	100.00;45.16	100.00;96.33;100.00;99.74;100.00				100.00;100.00;100.00;30.26;100.00					100.00;100.00;1		
.00	100.00	100.00	55.46	100.00;100.00	100.00	99.93	100.00									

https://github.com/tseemann/abricate

Virulence & plasmids

Virulence
abricate --db vfdb ERS825165.fa > ERS825165.fa.vfdb.tab
abricate --summary *.vfdb.tab > vfdb.summary.tsv

Plasmids
abricate --db plasmidfinder ERS811728.fa > ERS811728.fa.plasmidfinder.tab

#FILE	NUM_FOUND	rep16_1_CDS8(pSAS)	rep16_2_CDS6(pSJH101)	rep19_10_rep(pWBG746)	rep19_7_repA(SAP019A)	rep20_1_ORF1(EDINA)	rep20_3_rep(pTW20)	rep21_10_rep(pKH14)	rep21_14_re
p(pKH21)	rep21_3_pS0385-3	rep5_1_rep(pMW2)	rep5_5_rep(pRJ6)	rep7_14_rep(MSSA476)	repUS20__rep(pAVX)	repUS5__CDS20(pETB)			
ERS811724.fa.plasmidfinder.tab	0	.	.	.	.	.	.	.	.
ERS811726.fa.plasmidfinder.tab	0	.	.	.	.	.	.	.	.
ERS811727.fa.plasmidfinder.tab	0	.	.	.	.	.	.	.	.
ERS811728.fa.plasmidfinder.tab	0	.	.	.	.	.	.	.	.
ERS811729.fa.plasmidfinder.tab	0	.	.	.	.	.	.	.	.
ERS811730.fa.plasmidfinder.tab	3	.	99.87	100.00	.	100.00	.	.	.
ERS811733.fa.plasmidfinder.tab	3	.	99.87	100.00	.	100.00	.	.	.
ERS811737.fa.plasmidfinder.tab	3	.	100.00	.	100.00	.	.	.	.
ERS811738.fa.plasmidfinder.tab	3	.	100.00	.	100.00	.	.	.	.
ERS811740.fa.plasmidfinder.tab	2	.	.	.	.	100.00	.	.	.
ERS812507.fa.plasmidfinder.tab	3	.	100.00	.	100.00	.	.	.	.
ERS819824.fa.plasmidfinder.tab	1	.	.	.	.	.	.	.	.
ERS825151.fa.plasmidfinder.tab	3	.	.	.	.	.	.	.	.
ERS825158.fa.plasmidfinder.tab	5	20.97	.	.	.	.	100.00	29.12	74.40

13.52	.	.	.	.	100.00	100.00	100.00		

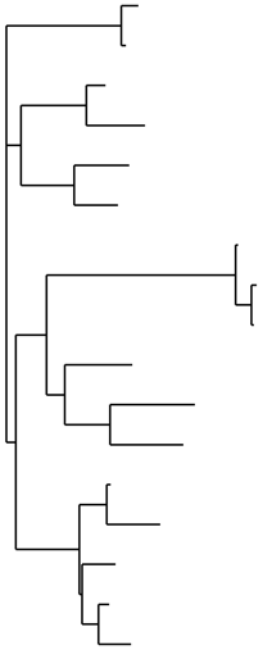
Quadram Institute
Science Health
Food Innovation

Phylogenetic Trees

Phylogenetic trees

Roary provides a rough and ready tree so you can visualise immediately.
Based on gene presence & absence of the accessory.

You can do a lot better!



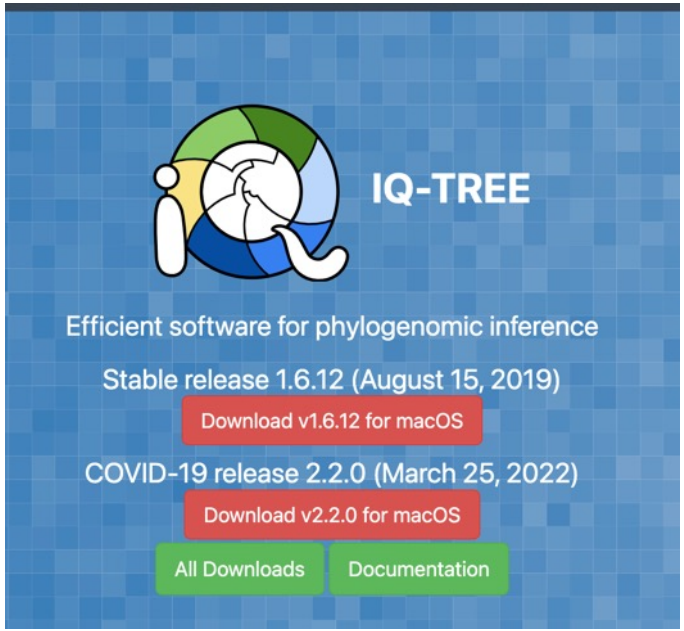
Tree building software

IQ-Tree (version 2) is the most popular phylogenetic tree program
- Fast, can run on a laptop
- Good defaults
- Takes a multi-FASTA alignment (output of roary)

Other software
- RAxML
- FastTree
- beast, revBayes, phylobayes
- rapidnj
- mpboot

Tree building models haven't changed much in decades, its just the packaging that's different.

Software not to use: TREEFINDER



http://www.iqtree.org/

Core gene tree

Core gene tree, lets you see stable* context (*recombination excluded)
Take all core genes, align, SNPs give you phylogenetic signal
Excludes intergenic regions

```
roary –e *.gff
```

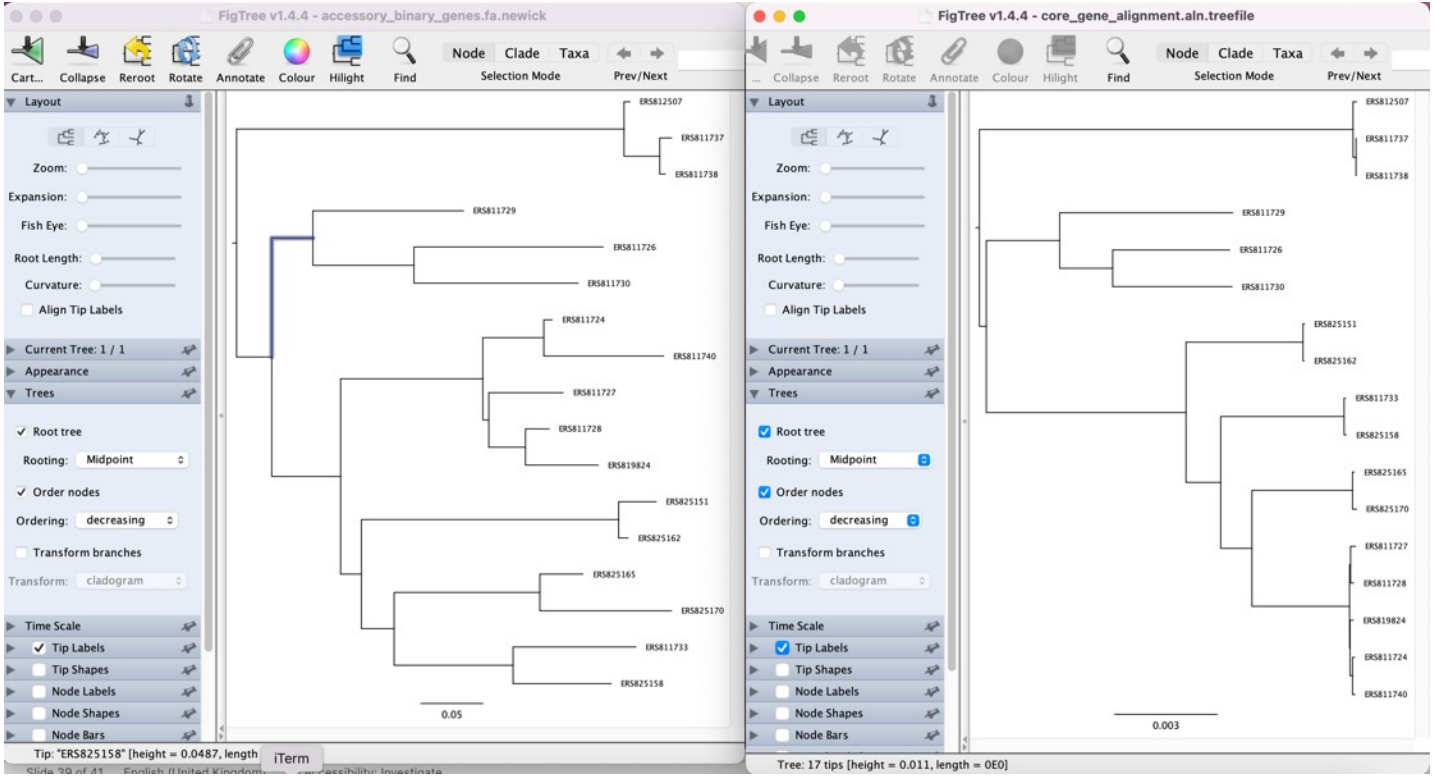
Build a core gene tree

```
iqtree –s core_gene_alignment.aln
```

If its too slow (or want a quicker look) run snp-sites first to keep only bases with variation (*less accurate):

```
snp-sites core_gene_alignment.aln
```

Figtree

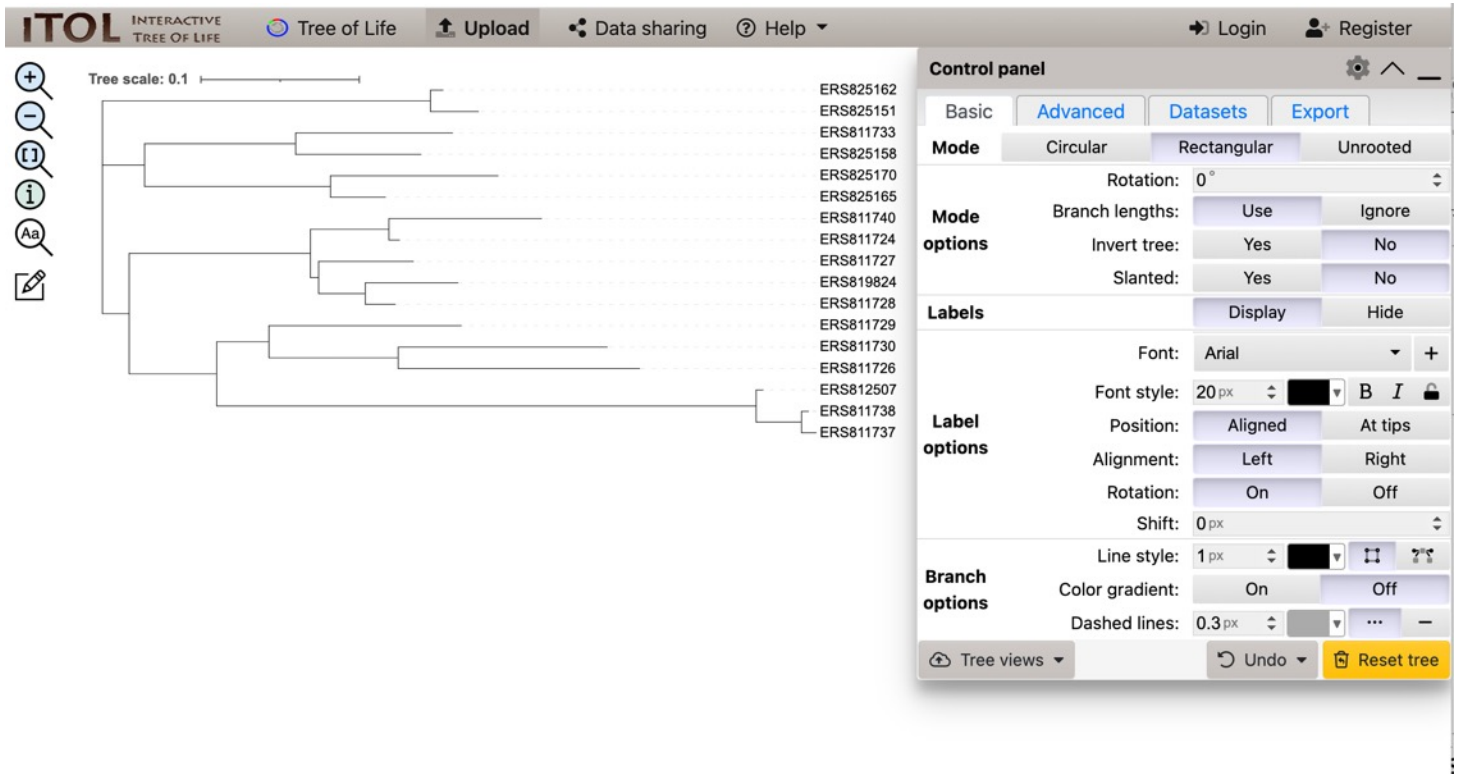


Accessory

Core gene tree

http://tree.bio.ed.ac.uk/software/figtree/

iTOL



https://itol.embl.de/

What next?

After writing these slides I've realised we need a better tool for pulling out MGEs, so watch this space….