Public Health England

Protecting and improving the nation's health

Hierarchical clustering of core genome MLST data for rapid assessment of the genetic relatedness of *S. aureus*.

Bruno Pichon, Michel Doumith, Neil Woodford, Angela Kearns

Background

Public Health England invested in NGS

NGS for microbiology reference service
- streamlined workflows
- validated automated bioinformatics pipelines
- cost effective

SNP based approach for micro-epidemiology investigations
<u>but</u>
- lack of standardisation
- not portable
- dependent of reference genome

Objective: explore core genome MLST for strain comparison

cgMLST scheme

N315 reference (NC_002745) chromosome used to extract putative ORFs. (excluded : known mobile elements, insertion sequences, multi-copy genes…

BLAST analyses on 54 publicly available complete genomes (NCBI RefSeq)

Criteria:

100% conserved

>85% identity, 100% coverage

single hit per genome

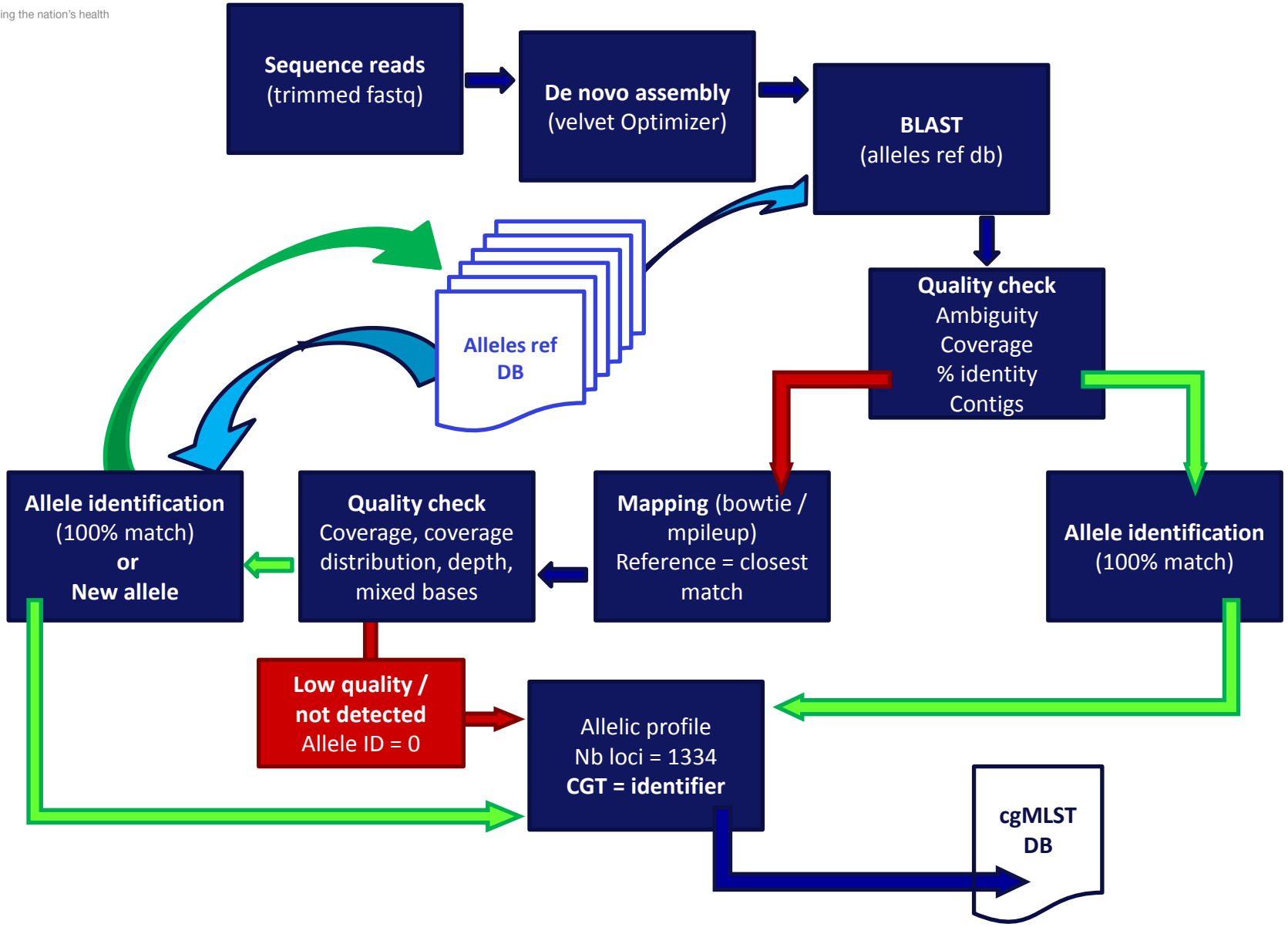
Core genome => 1334 loci

In house reference database: allele sequences, allelic profiles and CGTs

Public Health England

Protecting and improving the nation's health

Bioinformatics pipeline



Public Health England

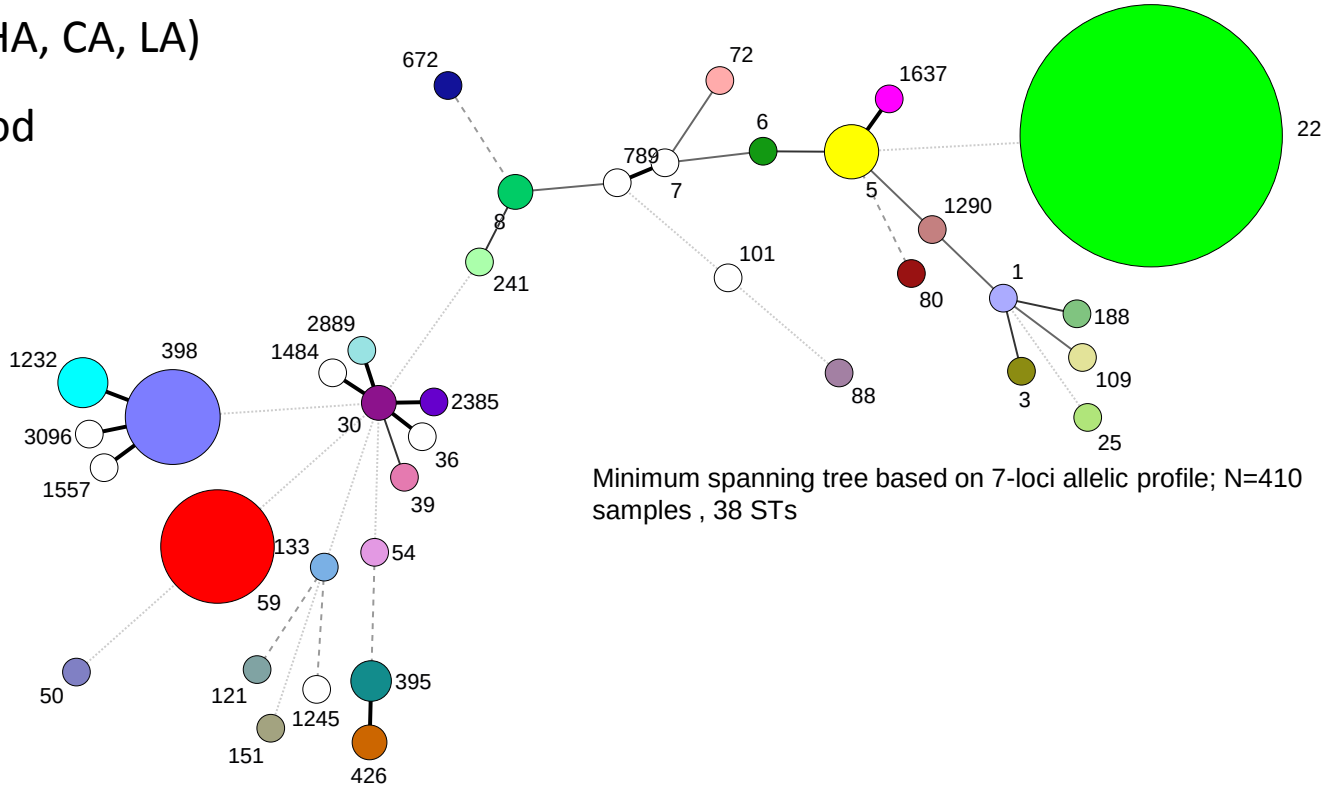
Protecting and improving the nation's health

Genome origins

54 complete genomes

356 newly sequenced genomes

- MSSA and MRSA (HA, CA, LA)
- human, animal, food
- genetic diversity
- 38 lineages



Minimum spanning tree based on 7-loci allelic profile; N=410 samples , 38 STs

Public Health England

Protecting and improving the nation's health

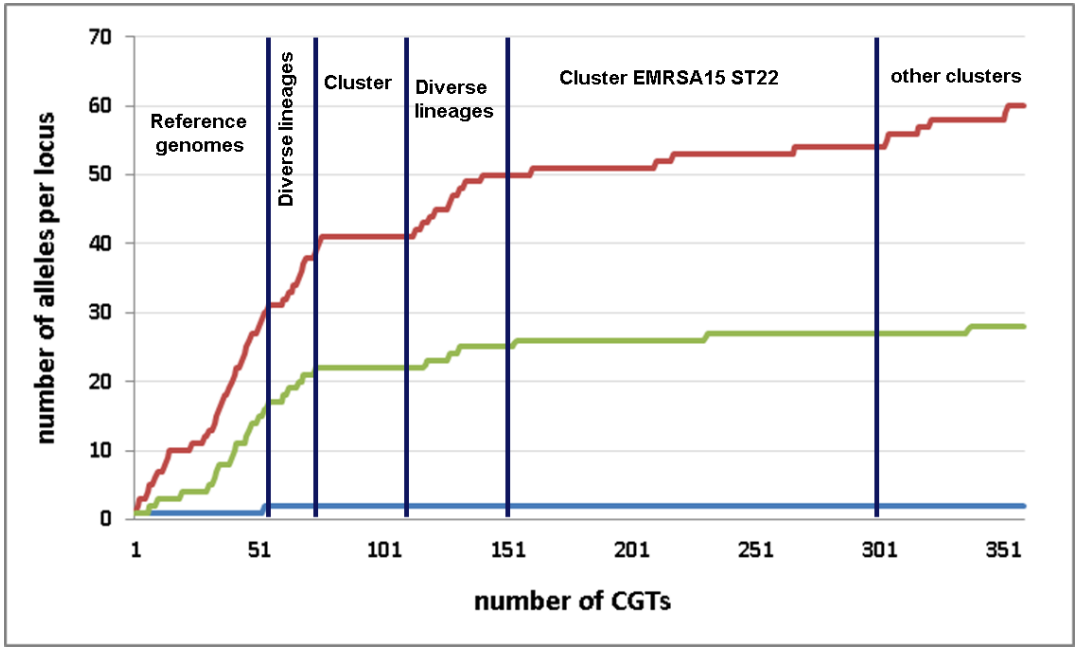
cgMLST analysis

410 genomes tested

355 CGTs identified

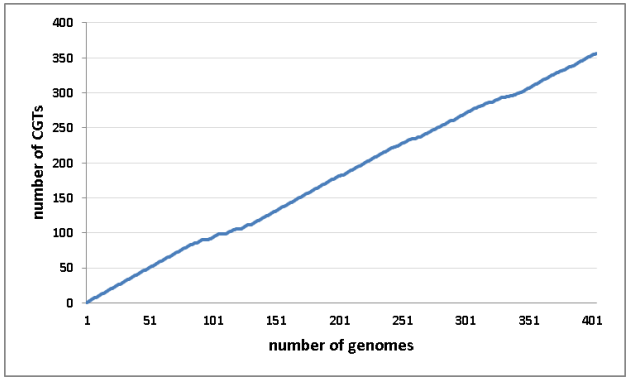
Allelic variation: median 28 alleles per locus

min = 2 (sa1282 - DNA-binding protein HU),
max = 60 (sa2039 - pump efflux acr family)



cgMLST for molecular epidemiology ?

cgMLST very discriminative



Reproducible

- 100 % concordance on repeat analysis from set of fastq files
- Repeated sequencing (25 x same strain => error rate of 0.03%)

<u>But</u>

How to assess genetic relatedness between CGTs ?

Hierarchical clustering of cgMLST data

Based on pairwise distance between allelic profiles

Distance measured in allelic difference (AD)

Clustering on 9 levels of AD: 3, 10, 25, 50, 75, 100, 150, 200 and 300

At each level , clusters are assigned to a unique identifier by analogy to cluster reference database

Allelic profiles are assigned to cluster addresses: concatenation of 9 cluster IDs plus CGT ID

In house reference databases (clusters and cluster addresses)

	ID	AD 300	AD 200	AD 150	AD 100	AD 75	AD 50	AD 25	AD 10	AD 3	CGT
ST5	N315	1	1	1	1	1	1	1	1	1	1
	Mu3	1	1	1	1	1	2	1	1	1	2
ST1	MSSA476	2	1	1	1	1	1	1	1	1	7
	MW2	2	1	1	2	1	1	1	1	1	8

Cluster detection

410 genomes, 38 STs (MLST 7 loci)

Clustering level (AD)	Number of clusters
300	32
200	42
150	59
100	83
75	99
50	121
25	212
10	259
3	285
CGT	355

Public Health England

Protecting and improving the nation's health

Outbreak investigations

51 independents case cluster reports of EMRSA15 referred form ICTs

Isolates N=151

Characterized by *spa* typing and PFGE

130 CGTs

Cluster addresses populated into meta database for analysis

Definitive link: same CGT or same cluster at AD 3

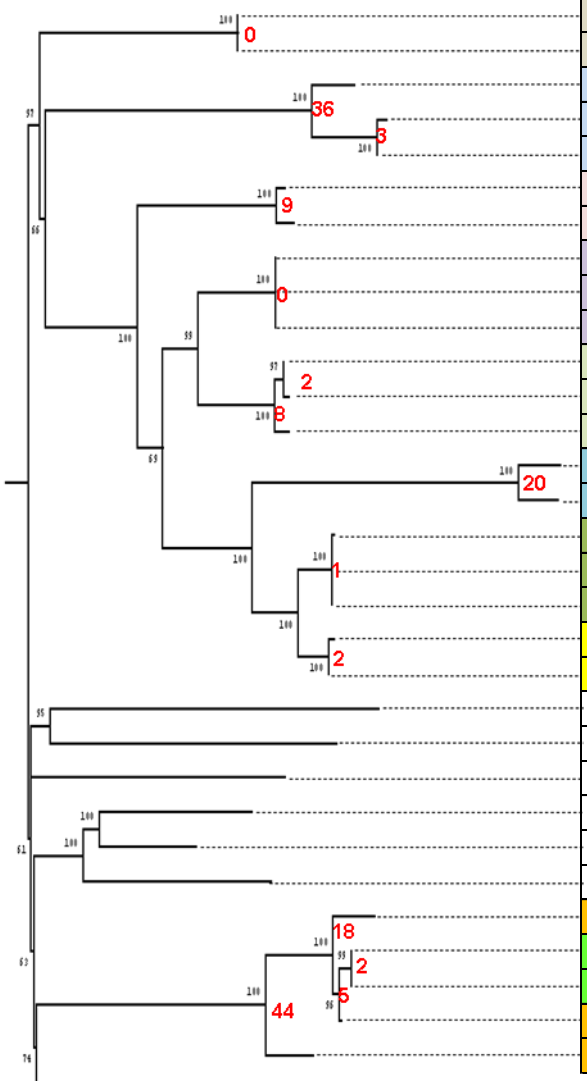
Probable link : same cluster at AD 10

Questionable: same cluster at AD 25

Unrelated: ≥ AD 50

			Cluster addresses									
Incident	*spa*	PFGE	AD 300	AD 200	AD 150	AD 100	AD 75	AD 50	AD 25	AD 10	AD 3	CGT
3	t032	L	5	1	1	1	1	1	6	1	1	147
3	t032	L	5	1	1	1	1	1	6	1	2	148
5	t022	G	5	1	1	1	1	1	7	1	1	150
5	t022	G	5	1	1	1	1	1	7	1	1	264
5	t022	G	5	1	1	1	1	1	7	1	2	261
4	t032	R	5	1	1	1	1	1	7	2	1	234
4	t032	R	5	1	1	1	1	1	7	2	1	234
4	t032	R	5	1	1	1	1	1	7	2	1	234
-	t032		5	1	1	1	1	1	12	1	1	180
-	t032		5	1	1	1	1	1	14	1	1	185
-	t032		5	1	1	1	1	1	16	1	1	191
7	t032	S	5	1	1	1	1	1	20	1	1	200
7	t032	S	5	1	1	1	1	1	20	1	1	206
7	t032	S	5	1	1	1	1	1	20	1	1	209
8	t032	P	5	1	1	1	1	1	20	2	1	208
8	t032	P	5	1	1	1	1	1	20	2	1	208
-	t032		5	1	1	1	1	1	24	1	1	207
6	t032	H	5	1	1	1	1	1	35	1	2	274
6	t032	H	5	1	1	1	1	1	35	1	3	275
10	t032	F	5	1	1	1	2	1	1	1	1	182
10	t032	F	5	1	1	1	2	1	1	1	1	182
9	t032	F	5	1	1	1	2	1	1	1	1	188
9	t032	F	5	1	1	1	2	1	1	1	2	189
9	t032	F	5	1	1	1	2	1	1	2	1	190
1	t2033	N	5	1	1	1	2	2	1	1	1	233
1	t2033	N	5	1	1	1	2	2	1	1	1	233
2	t032	A	5	1	1	1	2	4	1	1	1	270
2	t032	A	5	1	1	1	2	4	1	1	1	272
2	t032	B	5	1	1	1	2	4	1	2	1	271
-	t032		5	1	1	1	2	5	1	1	1	304
-	t032		5	1	1	1	4	1	1	1	1	216

cgMLST vs SNP



	Cluster addresses									
Incident	AD 300	AD 200	AD 150	AD 100	AD 75	AD 50	AD 25	AD 10	AD 3	CGT
1	5	1	1	1	2	2	1	1	1	233
1	5	1	1	1	2	2	1	1	1	233
2	5	1	1	1	2	4	1	2	1	271
2	5	1	1	1	2	4	1	1	1	272
2	5	1	1	1	2	4	1	1	1	270
3	5	1	1	1	1	1	6	1	1	147
3	5	1	1	1	1	1	6	1	2	148
4	5	1	1	1	1	1	7	2	1	234
4	5	1	1	1	1	1	7	2	1	234
4	5	1	1	1	1	1	7	2	1	234
5	5	1	1	1	1	1	7	1	1	264
5	5	1	1	1	1	1	7	1	1	150
5	5	1	1	1	1	1	7	1	2	261
6	5	1	1	1	1	1	35	1	2	274
6	5	1	1	1	1	1	35	1	3	275
7	5	1	1	1	1	1	20	1	1	209
7	5	1	1	1	1	1	20	1	1	206
7	5	1	1	1	1	1	20	1	1	200
8	5	1	1	1	1	1	20	2	1	208
8	5	1	1	1	1	1	20	2	1	208
-	5	1	1	1	4	1	1	1	1	216
-	5	1	1	1	2	5	1	1	1	304
-	5	1	1	1	1	1	14	1	1	185
-	5	1	1	1	1	1	12	1	1	180
-	5	1	1	1	1	1	24	1	1	207
-	5	1	1	1	1	1	16	1	1	191
9	5	1	1	1	2	1	1	1	2	189
10	5	1	1	1	2	1	1	1	1	182
10	5	1	1	1	2	1	1	1	1	182
9	5	1	1	1	2	1	1	1	1	188
9	5	1	1	1	2	1	1	2	1	190

Public Health England

Protecting and improving the nation's health

Outbreak investigations -summary

Microbiology investigations	No Inc	cgMLST clustering	SNP range	Outliers	SNP range
Definitive same spa-type / PFGE	8	Definitive (unique CGT)	0-6	2 (AD >25)	68-235
Definitive same spa-type / PFGE	22	Definitive (<3 AD)	0-9	5 (>AD 50)	37-245
Probable: same spa-type / PFGE variants	7	Probable (<10 AD)	2-20	1 (>AD50)	ND
Questionable same spa-type / multiple PFGE profile	7	Questionable (> 25AD)	11-50	1 (>AD50)	180
Not confirmed	7	Not confirmed > 50 AD	87-180	-	

Conclusions

Pros:

- cgMLST scheme is very discriminative
- cluster addresses enabled strains comparison
- resolve micro-epidemiology of pandemic clone
- national surveillance
- ? basis for nomenclature

Cons:

- dependent of quality sequencing
- lower discrimination power than SNP based approach
- threshold definition
- scalability
- need of central databases for inter-laboratories usage

Future works

Validation : testing additional genomes from various lineages

Robustness : various sources of sequences

Portability: external collaborations

Public Health England

Protecting and improving the nation's health

Acknowledgments

AMRHAI
Mark Ganner
Lauren Harwin
Sharla McTavish

Information Communication and Technology
Francesco Giannoccaro

Genomic Services and Development Unit

Applied Bioinformatics and Laboratory Informatics