nature ecology & evolution

SUPPLEMENTARY INFORMATION

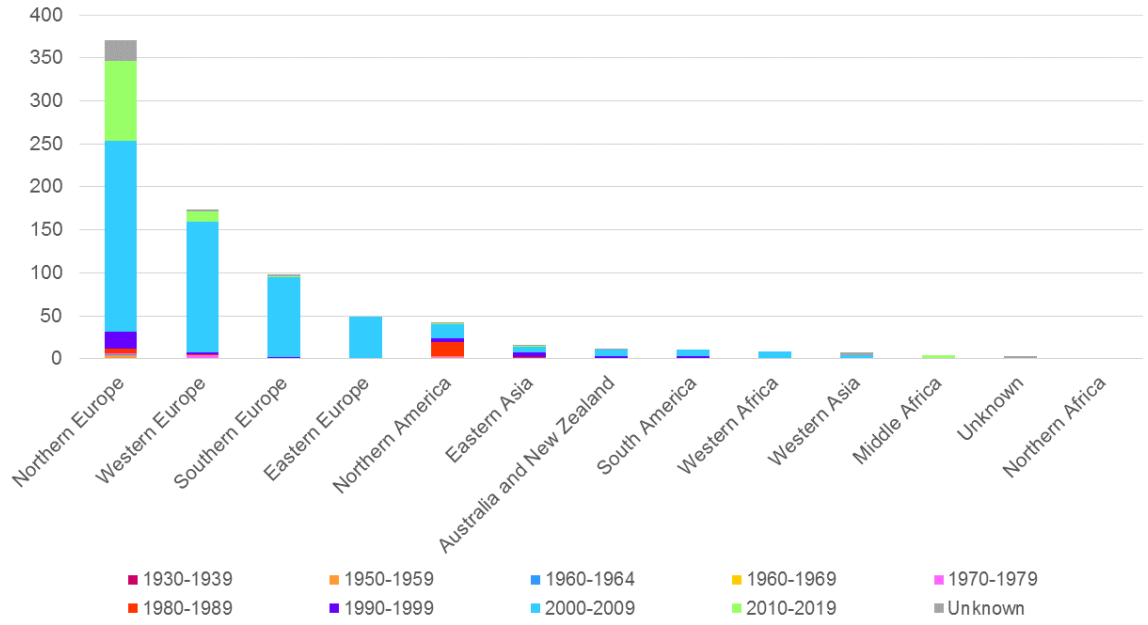
https://doi.org/10.1038/s41559-018-0617-0

In the format provided by the authors and unedited.

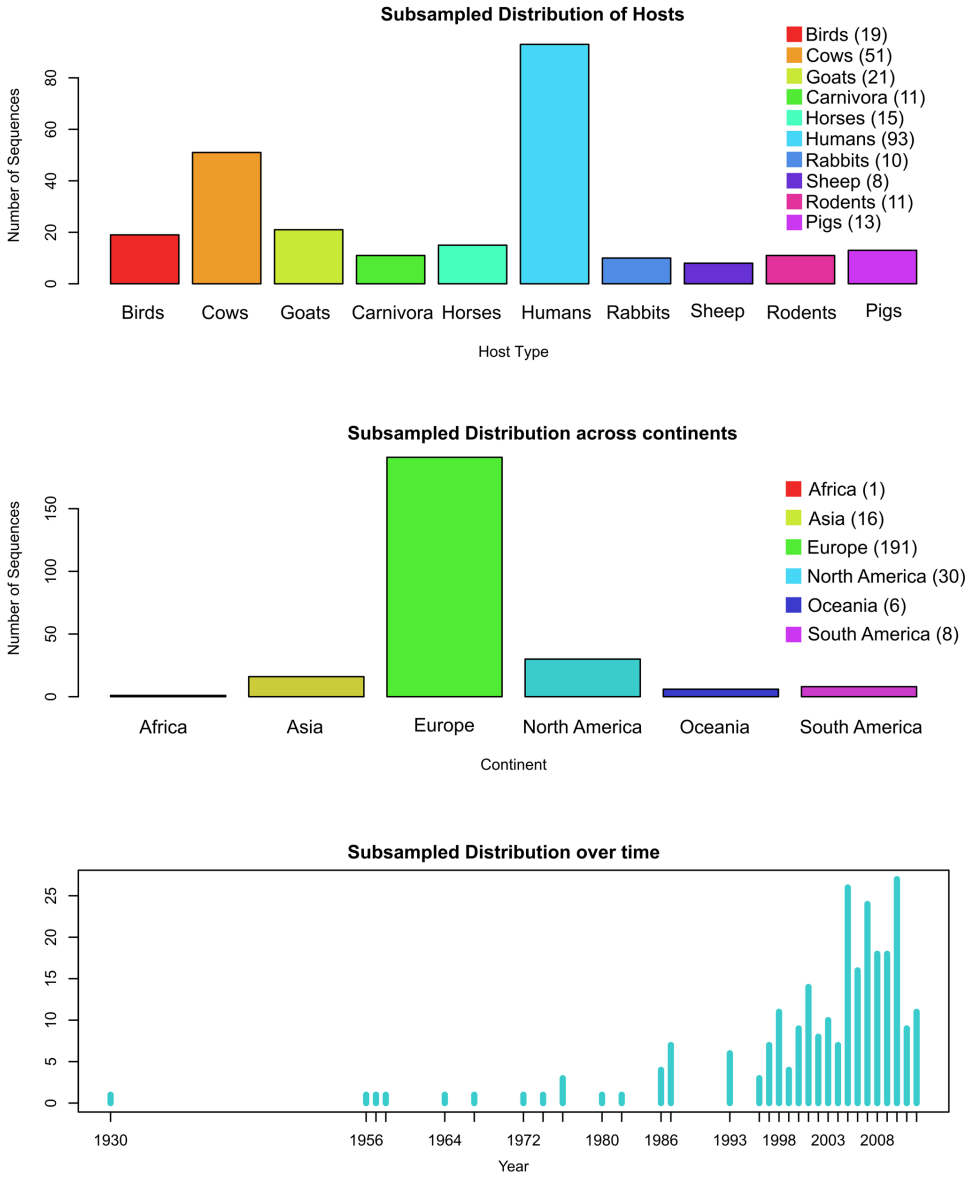
Gene exchange drives the ecological success of a multi-host bacterial pathogen

Emily J. Richardson[1,18,19], Rodrigo Bacigalupe[1,19], Ewan M. Harrison[2,19], Lucy A. Weinert[3,19], Samantha Lycett[1], Manouk Vrieling[1], Kirsty Robb[4], Paul A. Hoskisson[4], Matthew T. G. Holden[5], Edward J. Feil[6], Gavin K. Paterson[7], Steven Y. C. Tong[8,9], Adebayo Shittu[10], Willem van Wamel[11], David M. Aanensen[12,13], Julian Parkhill[14], Sharon J. Peacock[15], Jukka Corander[14,16,17], Mark Holmes[3] and J. Ross Fitzgerald[1]*

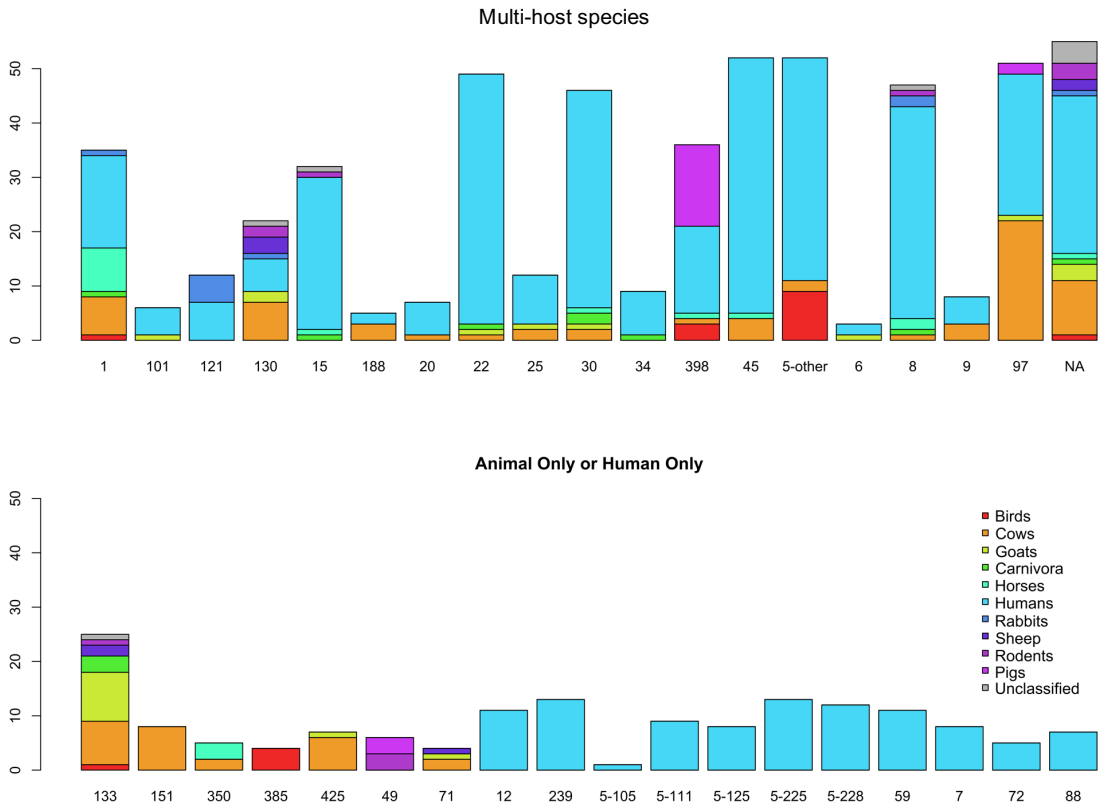
[1]The Roslin Institute, Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, UK. [2]Department of Medicine, University of Cambridge, Cambridge, UK. [3]Department of Veterinary Medicine, University of Cambridge, Cambridge, UK. [4]University of Strathclyde, Glasgow, UK. [5]School of Medicine, University of St Andrews, St Andrews, UK. [6]Milner Centre for Evolution, University of Bath, Bath, UK. [7]Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, UK. [8]Victorian Infectious Disease Service, The Royal Melbourne Hospital and The University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia. [9]Menzies School of Health Research, Darwin, Australia. [10]Department of Microbiology, Obafemi Awolowo University, Ile-Ife, Nigeria. [11]Department of Medical Microbiology and Infectious Diseases, Erasmus MC, Rotterdam, The Netherlands. [12]Centre for Genomic Pathogen Surveillance, Hinxton, UK. [13]Department of Infectious Disease Epidemiology, Imperial College London, London, UK. [14]Wellcome Trust Sanger Institute, Hinxton, UK. [15]London School of Hygiene and Tropical Medicine, London, UK. [16]Helsinki Institute for Information Technology, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland. [17]Department of Biostatistics, University of Oslo, Oslo, Norway. [18]Present address: Institute of Microbiology and Infection, University of Birmingham, Birmingham, UK. [19]These authors contributed equally: Emily J. Richardson, Rodrigo Bacigalupe, Ewan M. Harrison, Lucy A. Weinert. *e-mail: Ross.Fitzgerald@ed.ac.uk



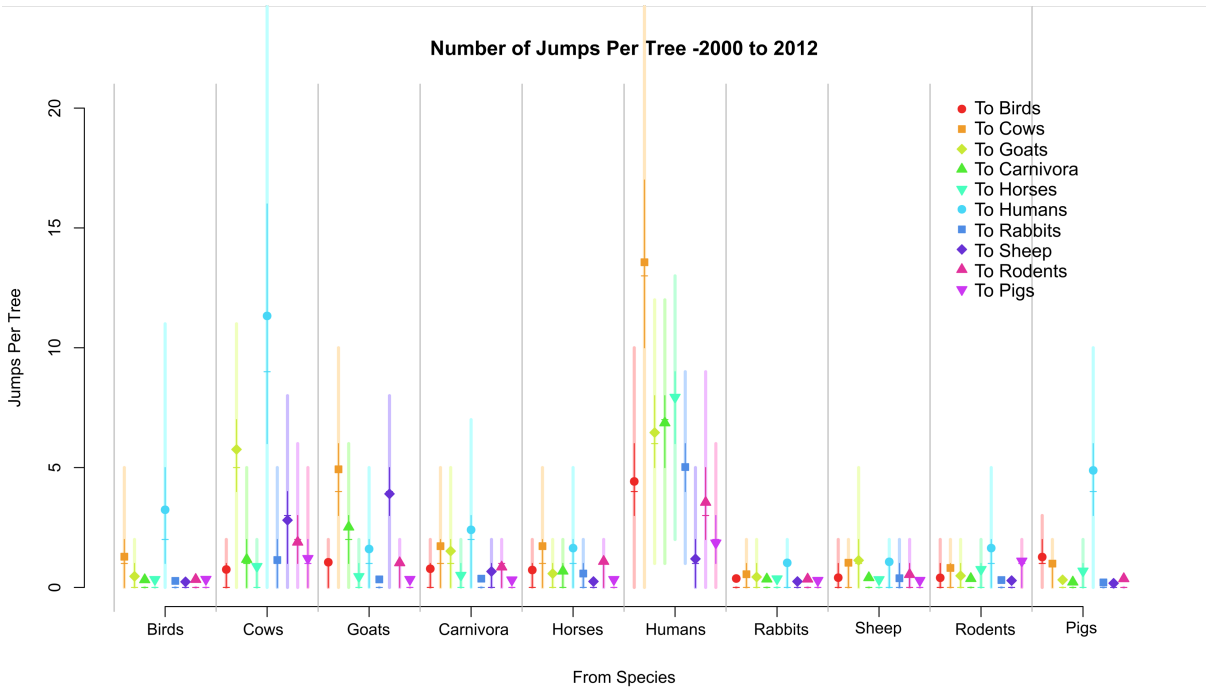
Supplementary Figure 1. Distribution of isolates by geographical region and date of isolation.



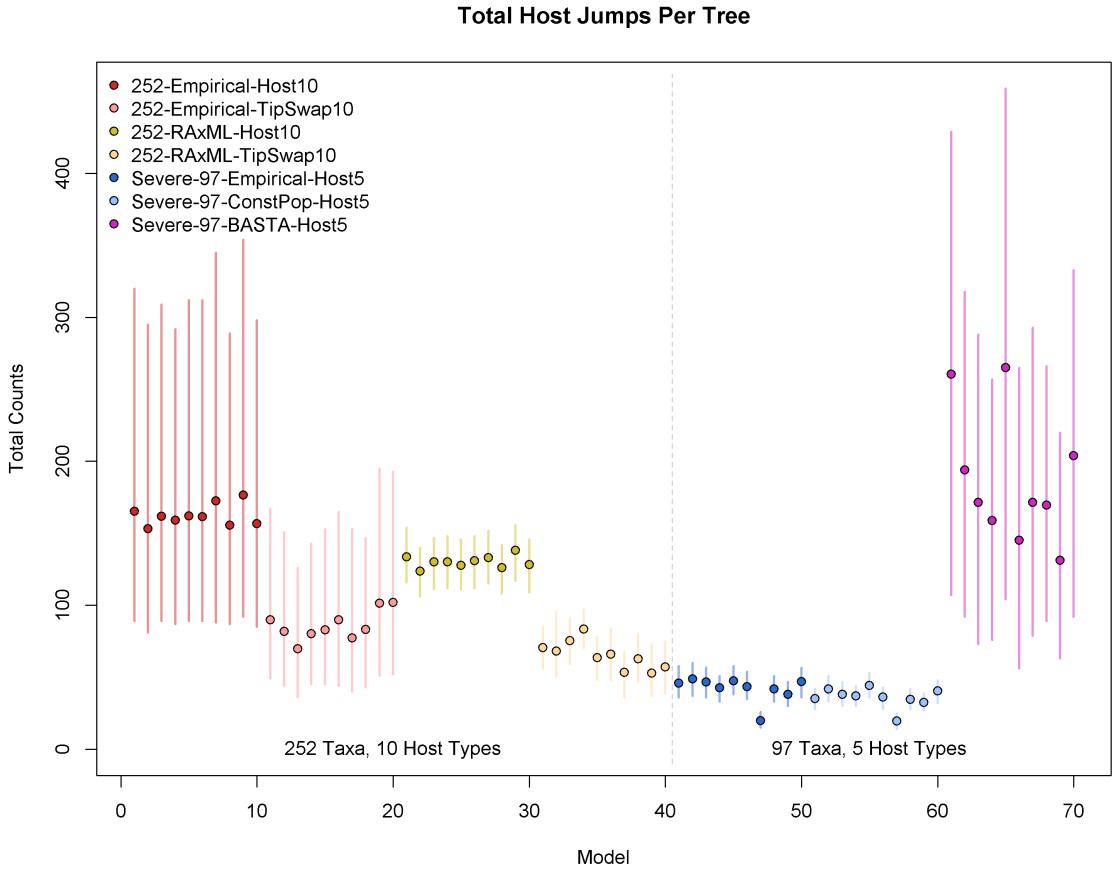
Supplementary Figure 2. Distribution of *S. aureus* genome sequences in subsampled data set for BEAST analysis. The unclassified sequences were omitted from the original set of 696 sequences, and the remainder were sub-sampled in order to reduce over-representation of some of the host-types (especially human and cow) but retain spatial and temporal diversity. 10 random subsamples were drawn from the data set (with replacement), with a maximum of 5 sequences per host-type per continent per year of sampling. Each subsample had the same distribution sequences for hosts, continents and years.



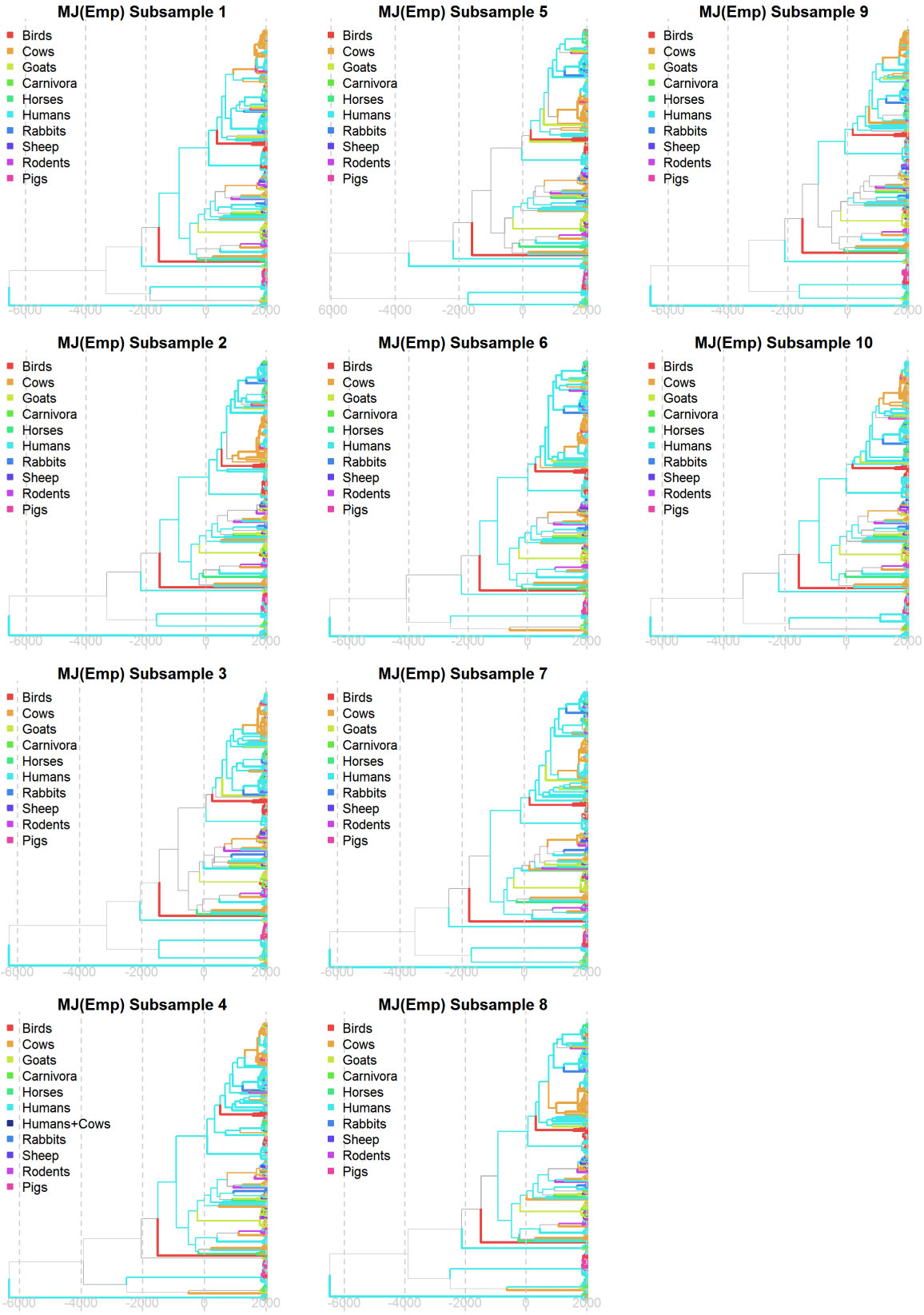
Supplementary Figure 3. *S. aureus* clonal complex composition by host species. Distribution of the number of sequences in the original set of 696 sequences by host type and clonal complex, including sequences not assigned to a clade (NA) and sequences not classified into a main host type (unclassified, grey).



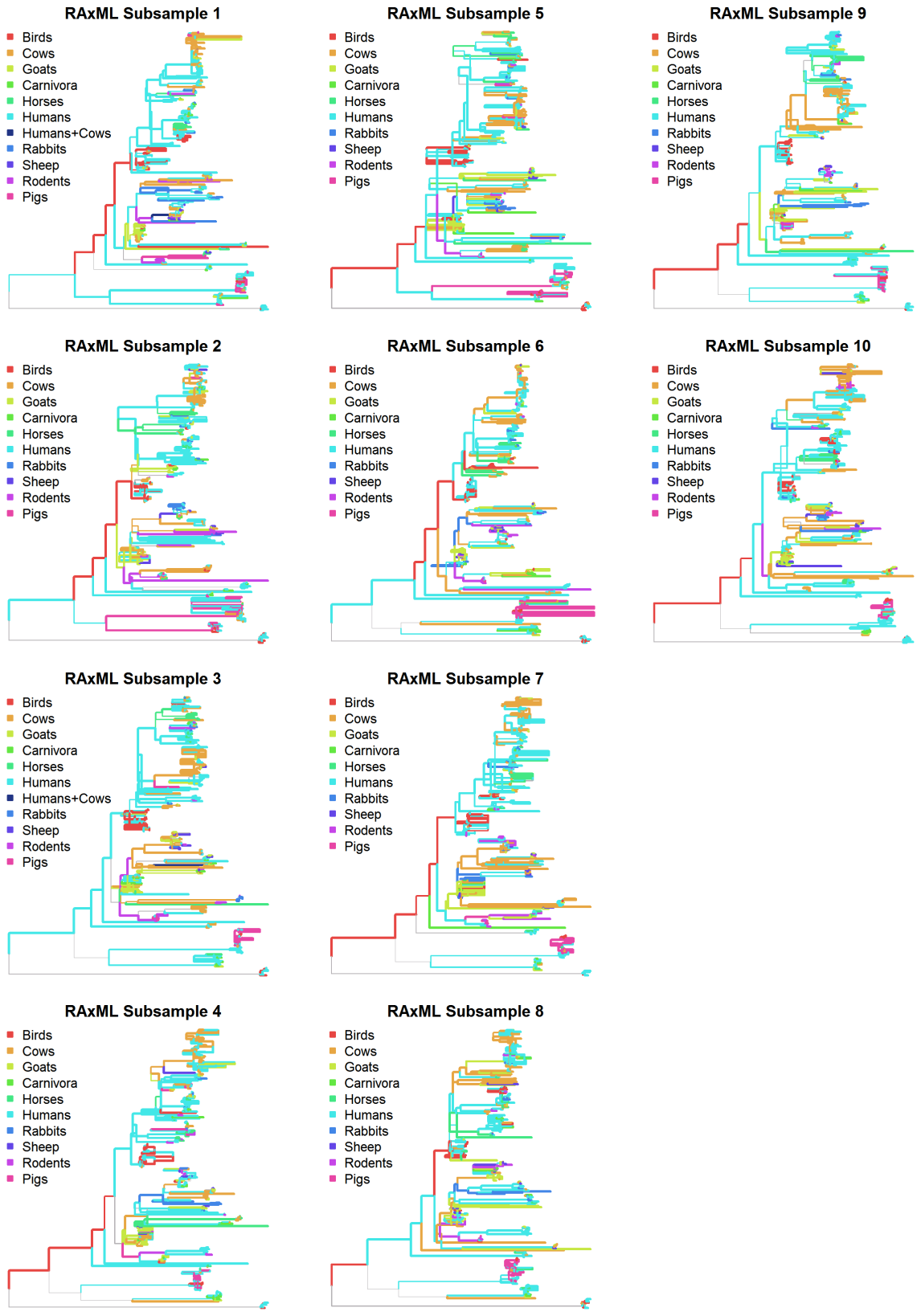
Supplementary Figure 4. Number of host jumps per tree as distributions from all subsamples and trees. The points represent the mean values and the small horizontal bars are the medians, the darker lines are the interquartile range and the pale background lines are the whisker limits.



Supplementary Figure 5. Comparison of total number of host jumps between sets of 10 subsamples and models. 'TipSwap' refers to models in which the tip labels are permuted (randomised) during the MCMC steps. The total number of host jumps across the trees are similar for the time-scaled trees and RAxML trees (red and yellow), however the number of jumps inferred under BASTA (purple) seem to be much higher than the corresponding discrete traits analysis on the same data (blues).

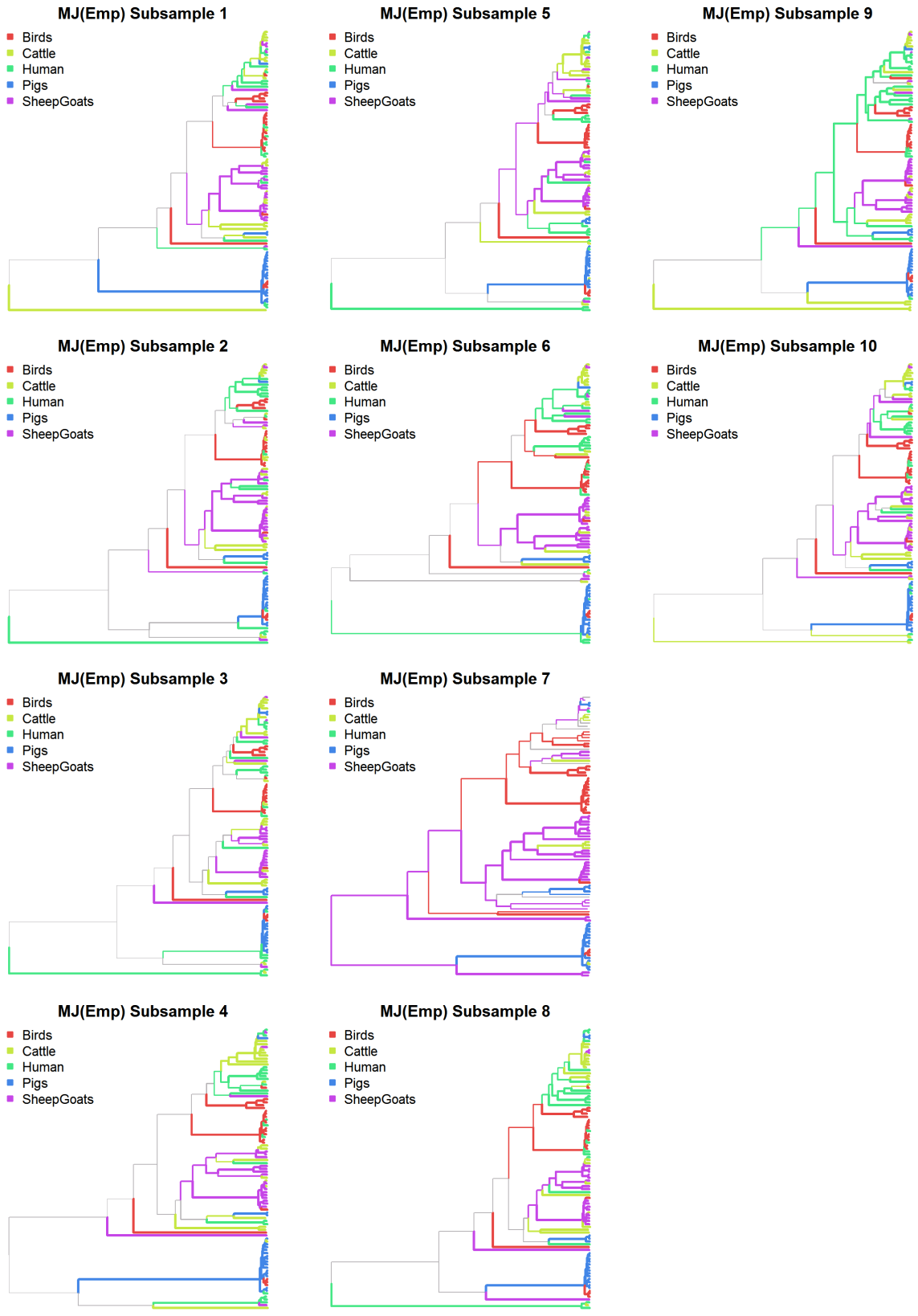


Supplementary Figure 6. BEAST analysis with 10 subsamples, 252 taxa, MCC Trees with 10 Host-type reconstruction using Discrete Trait Markov Jumps asymmetric model (empirical trees).

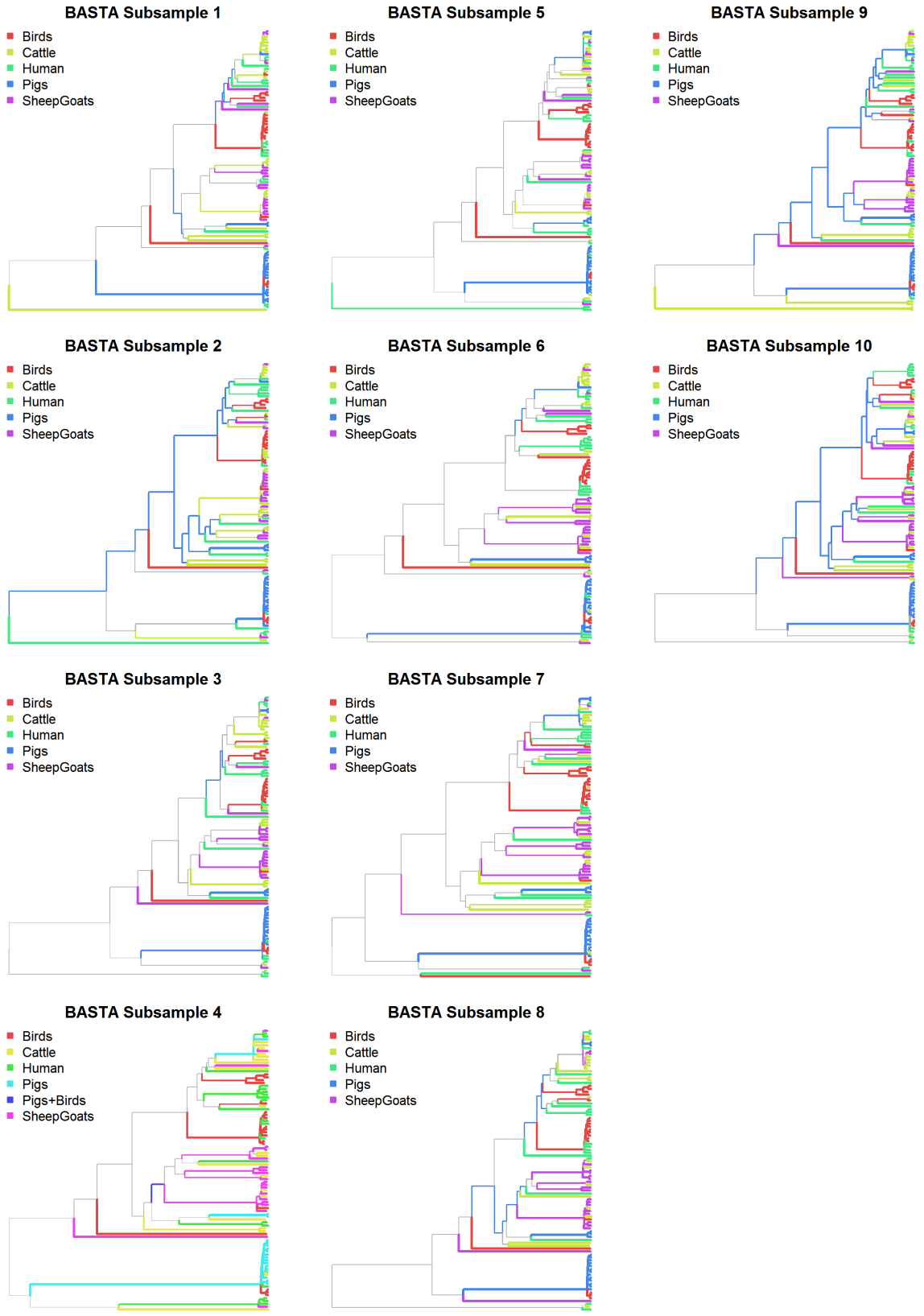


Supplementary Figure 7. BEAST analysis with 10 subsamples, 252 taxa, RAxML (Maximum Likelihood) Trees with 10 Host-type reconstruction using Discrete Trait Markov Jumps asymmetric model

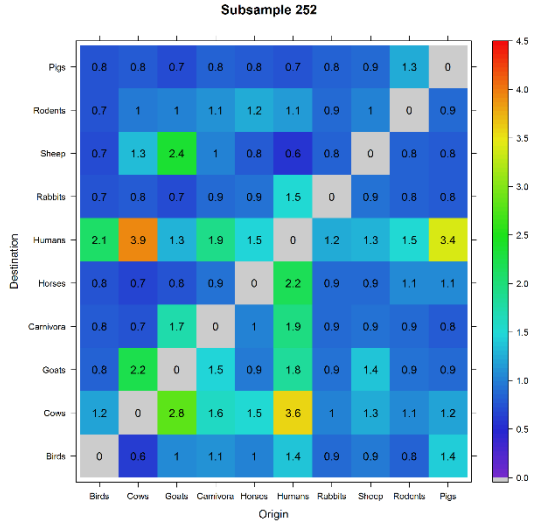
Supplementary Figure 8. 10 severe subsamples, 97 taxa, BEAST MCC Trees, with 5 Host-type reconstruction using Discrete Trait Markov Jumps asymmetric model (joint inference).



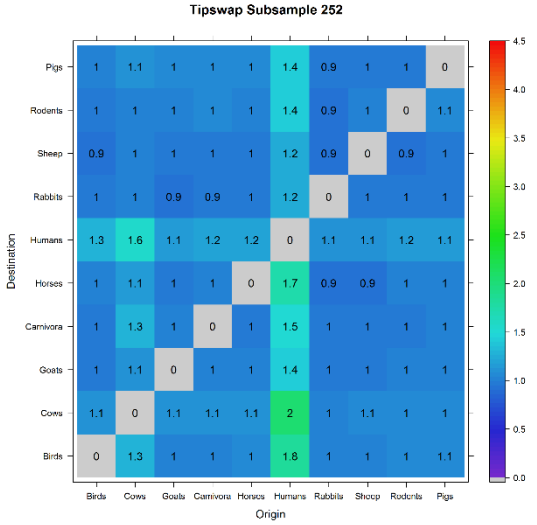
Supplementary Figure 9. 10 severe subsamples, 97 taxa, BEAST MCC Trees, with 5 Host-type reconstruction using Discrete Trait Markov Jumps asymmetric model (empirical trees).



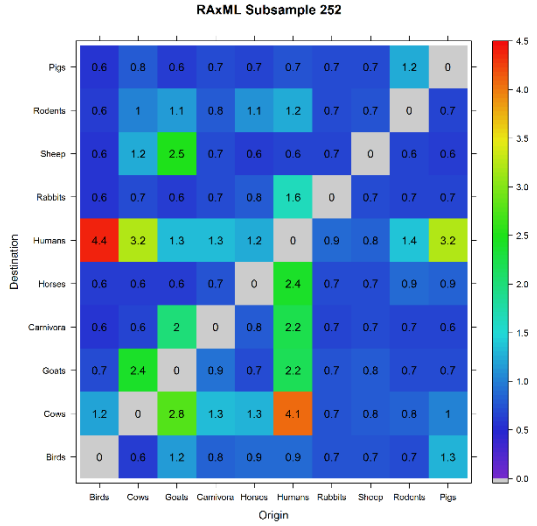
Supplementary Figure 10. 10 severe subsamples, 97 taxa, BEAST-BASTA MCC Trees, with 5 Host-type reconstruction (approximate structured coalescent).



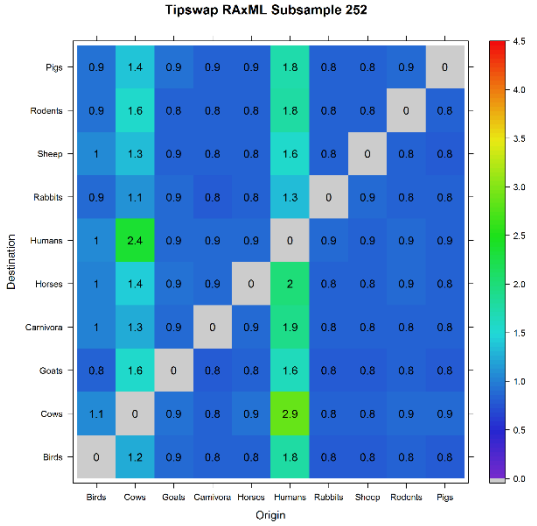
252 taxa subsamples 1-10



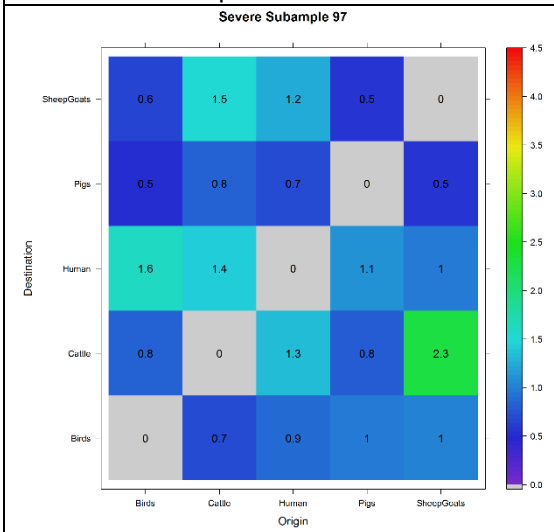
252 taxa subsamples 1-10 with tipswap



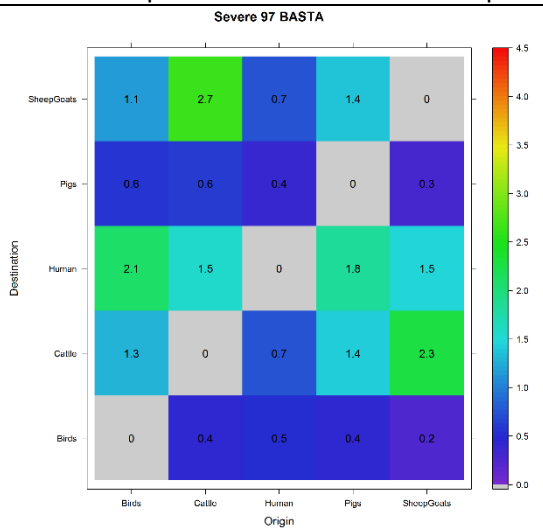
252 taxa subsamples with RAxML trees



252 subsamples with RAxML trees and tipswap

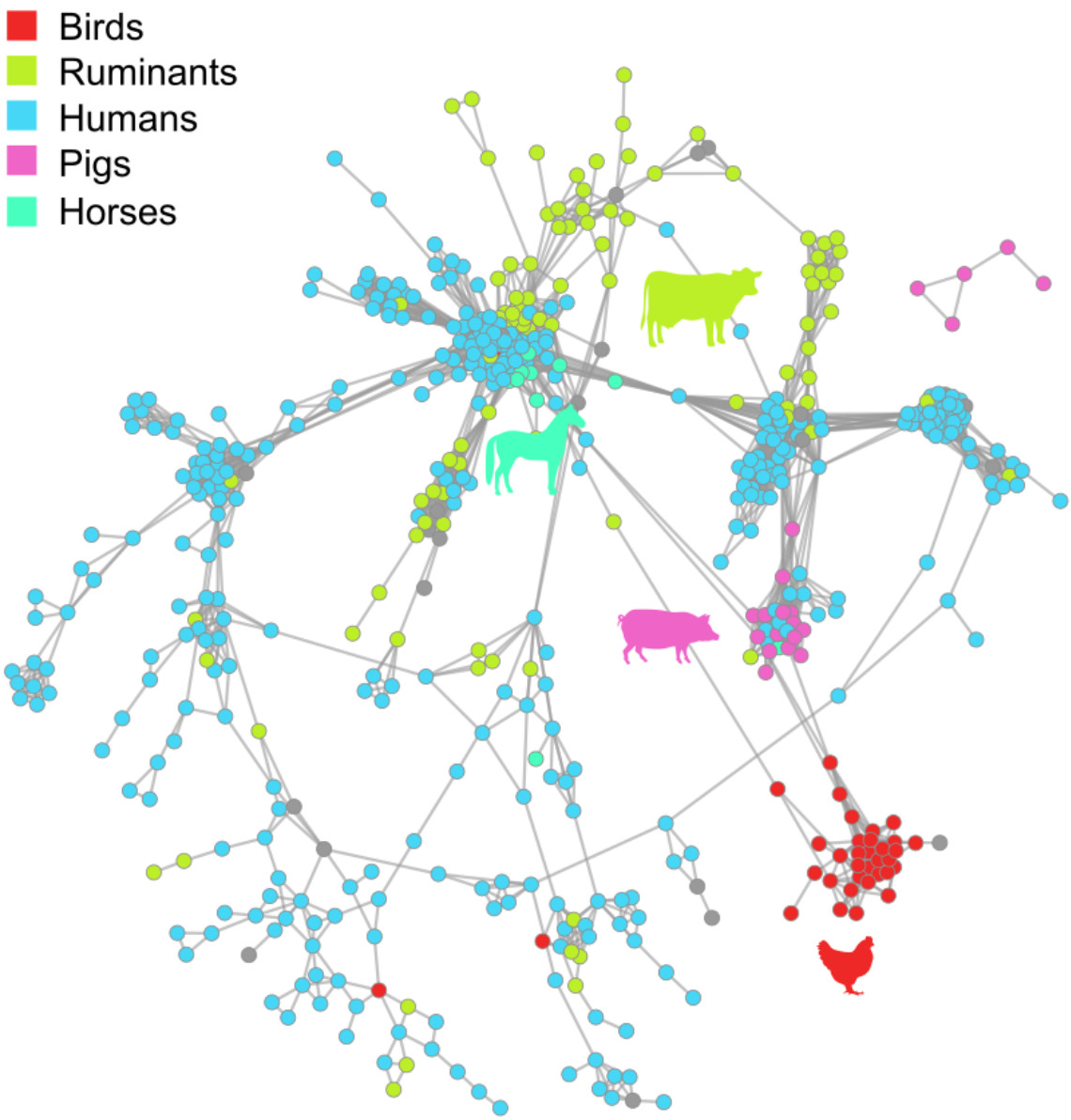


97 taxa severe subsamples empirical and constant population (no difference at this scale)

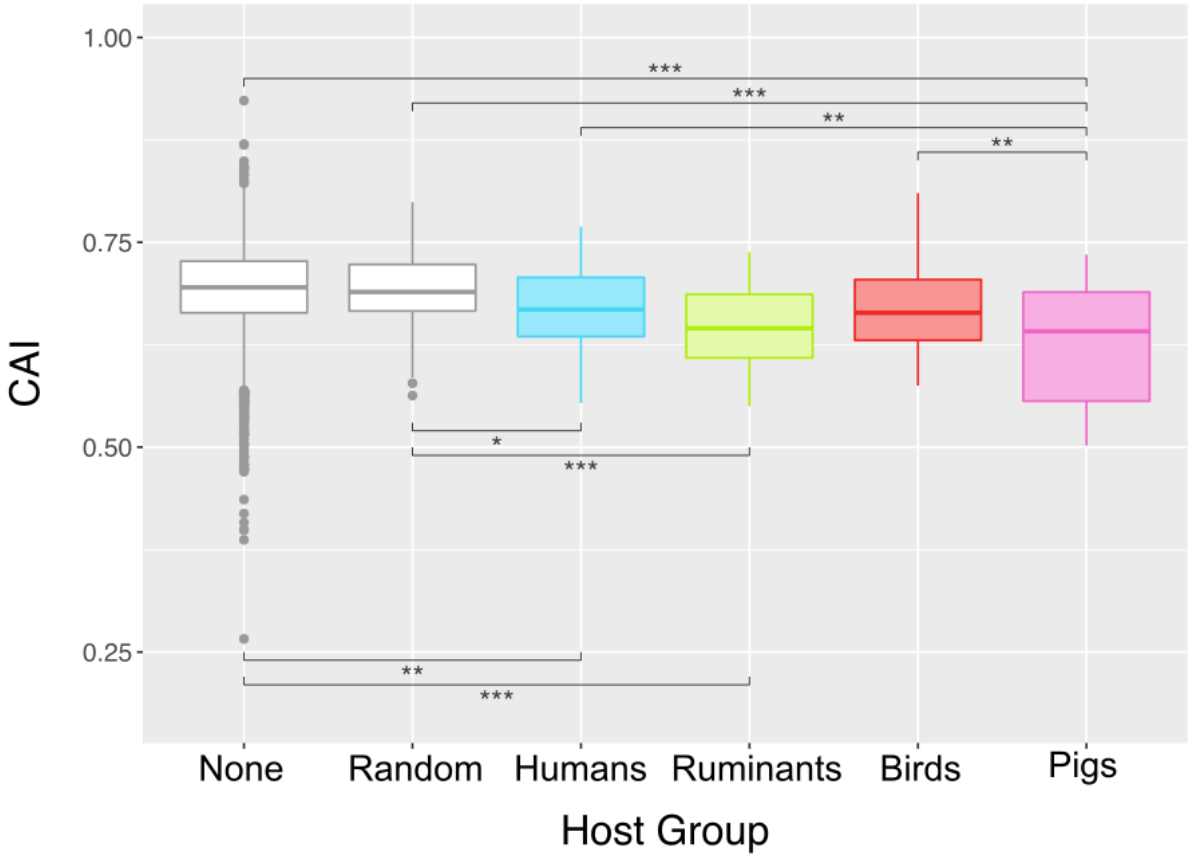


97 taxa severe subsamples BASTA

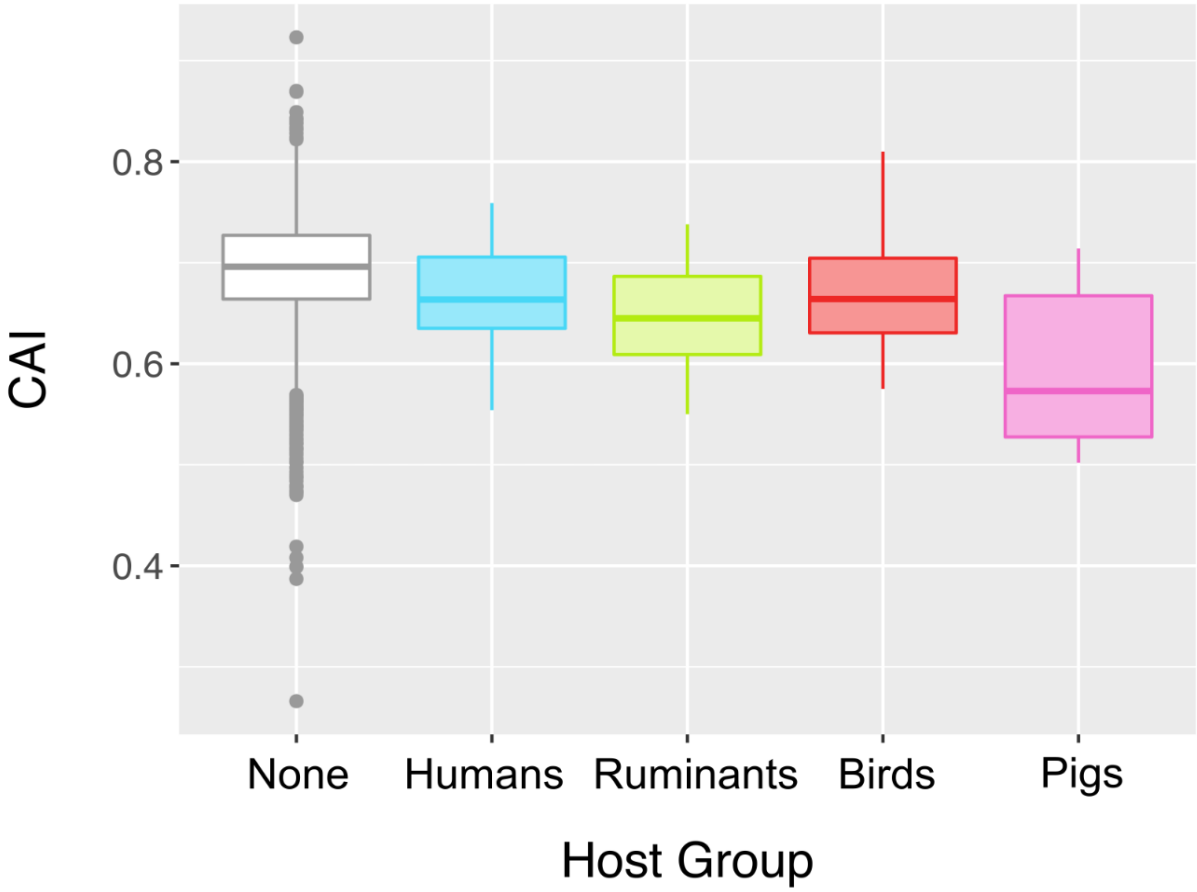
Supplementary Figure 11. Relative Rate Matrices for Host JumpTransitions



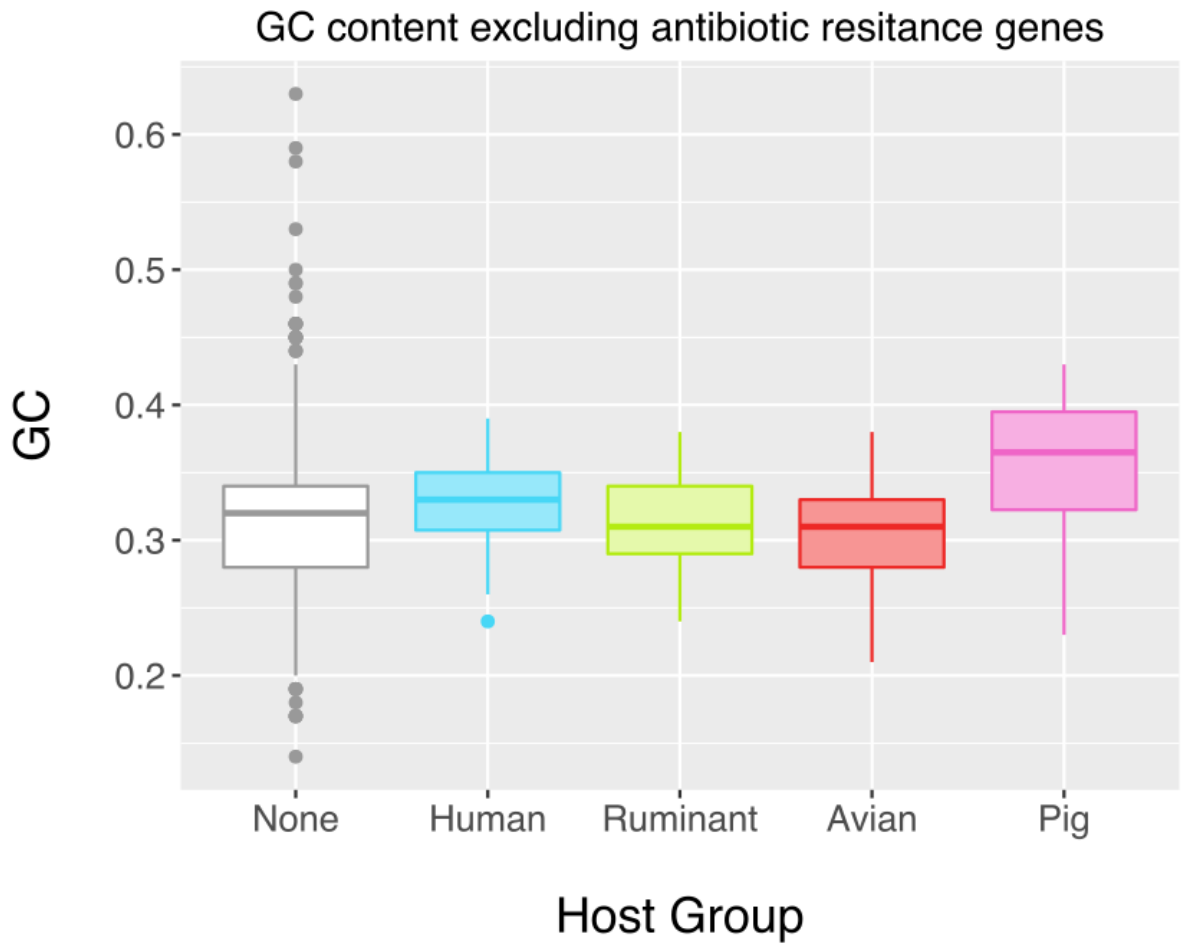
Supplementary Figure 12. Accessory genome correlation network based on pairwise distance matrix with antibiotic resistance determinants removed. As a result of removing antibiotic associated genes the pig genomes from ST97 and ST49 no longer cluster with pig ST398. Further, the horse genomes no longer cluster in a separate group together, rather they are more closely-allied to a large multi-host associated cluster. Edges are cut off with a distance threshold of 0.5 (all edges represent relationship with 50% identity or greater). The length of the edges is weighted by distance (the shorted the edge the more closely related the accessory genomes are).



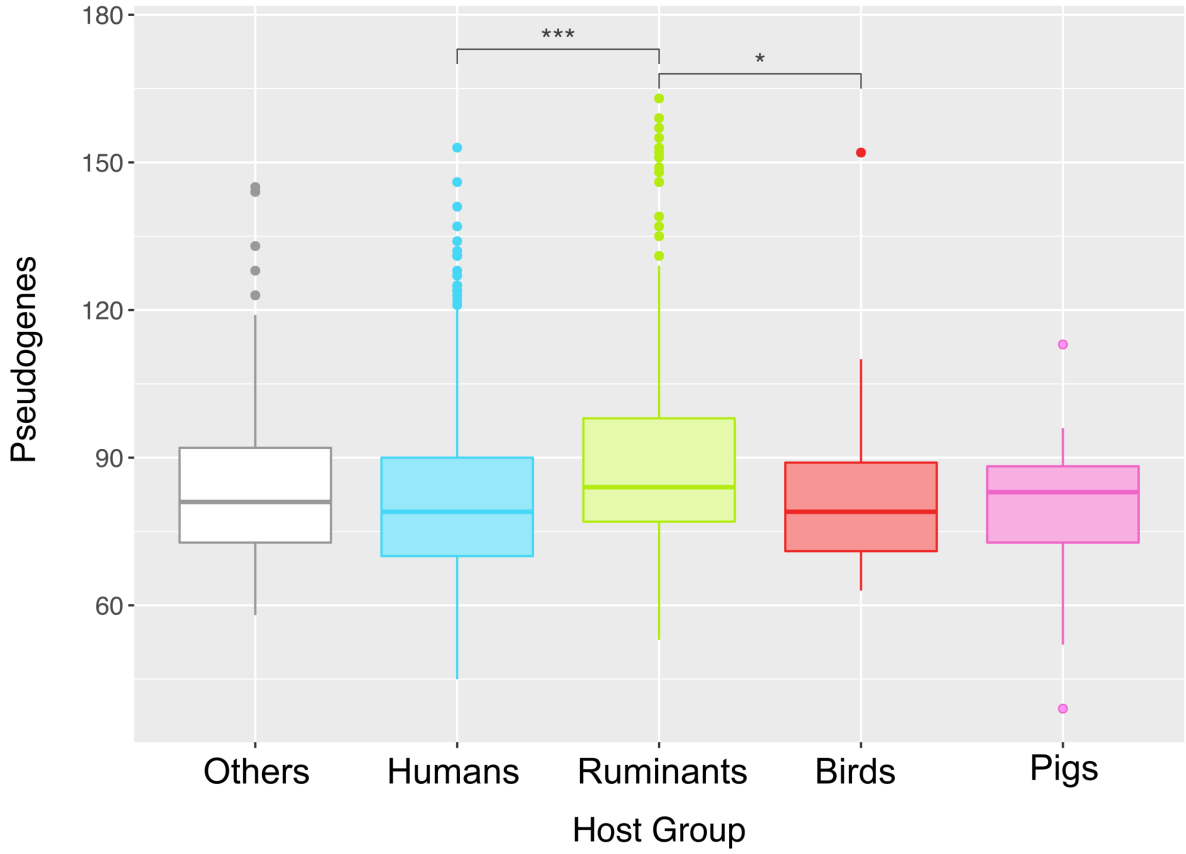
Supplementary Figure 13. Comparison of codon adaptation index (CAI) values for host-specific accessory genomes. Host group 'None' represents all other genomes (ie not associated with the explicitly stated host groups). 'Random' represents a control of five randomly selected sets of 50 core genes from the 'None' group, to demonstrate that host-specific accessory genome CAI values are not different by random chance. P-values are shown as *, ** and ***, representing p-values of 0.05, 0.01 and <0.001 respectively.



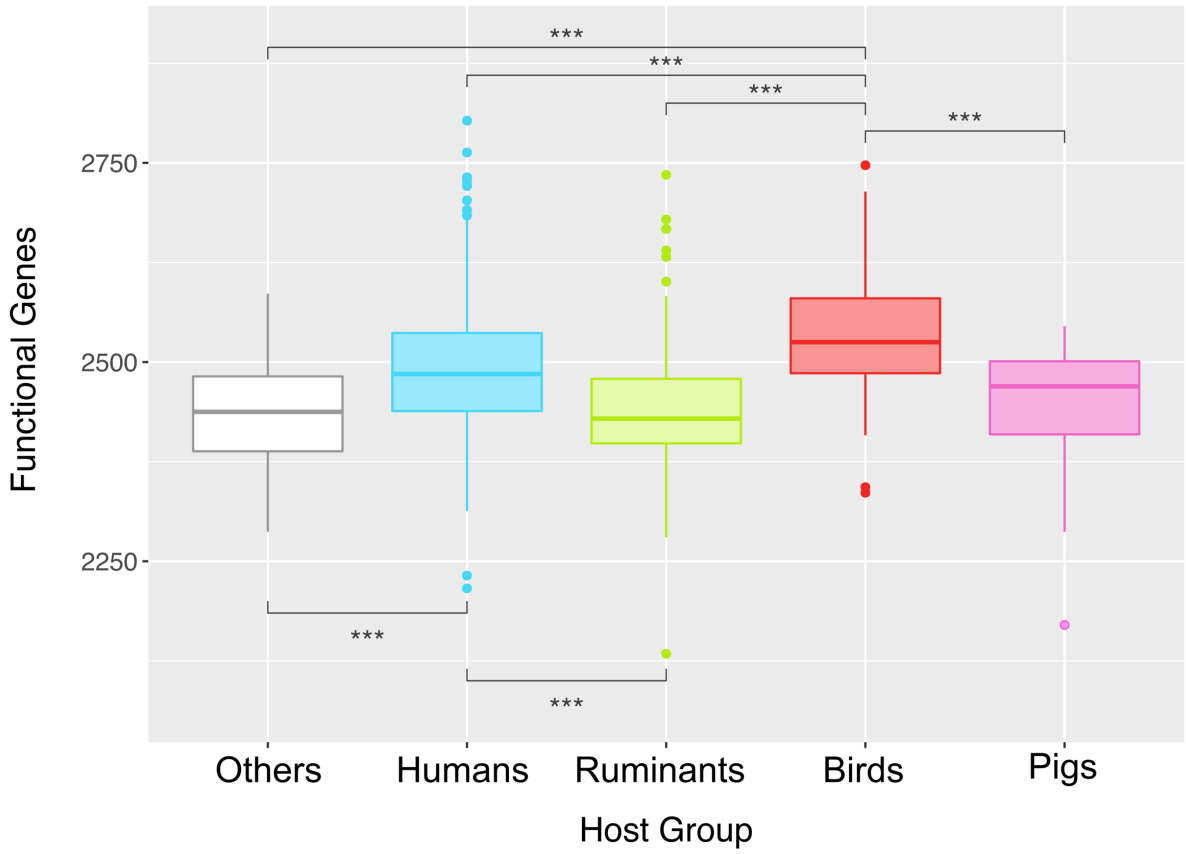
Supplementary Figure 14. Comparison of codon adaptation index (CAI) values for host jump associated accessory genomes after removal of antibiotic resistance genes.



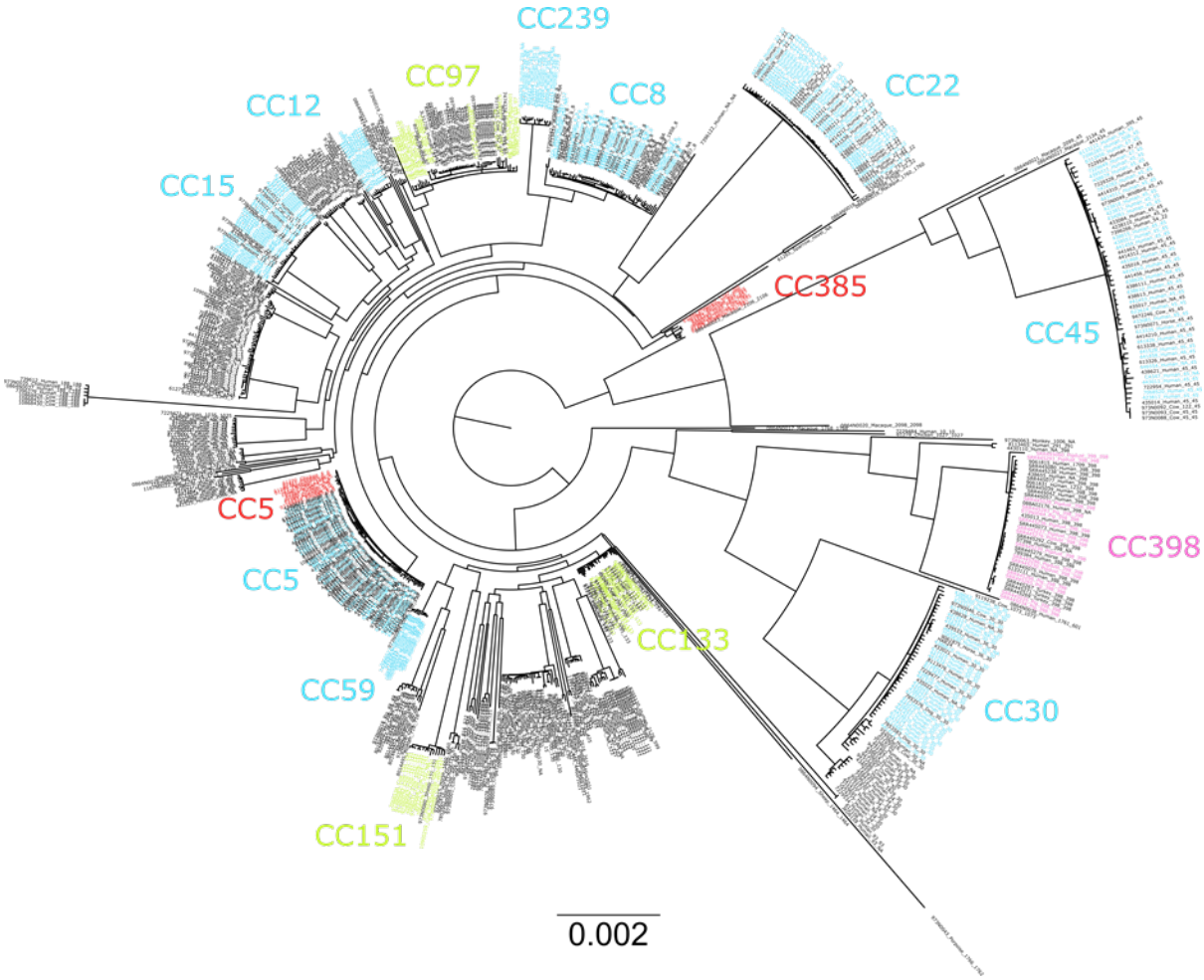
Supplementary Figure 15. GC content of accessory genomes grouped according to host-species association.



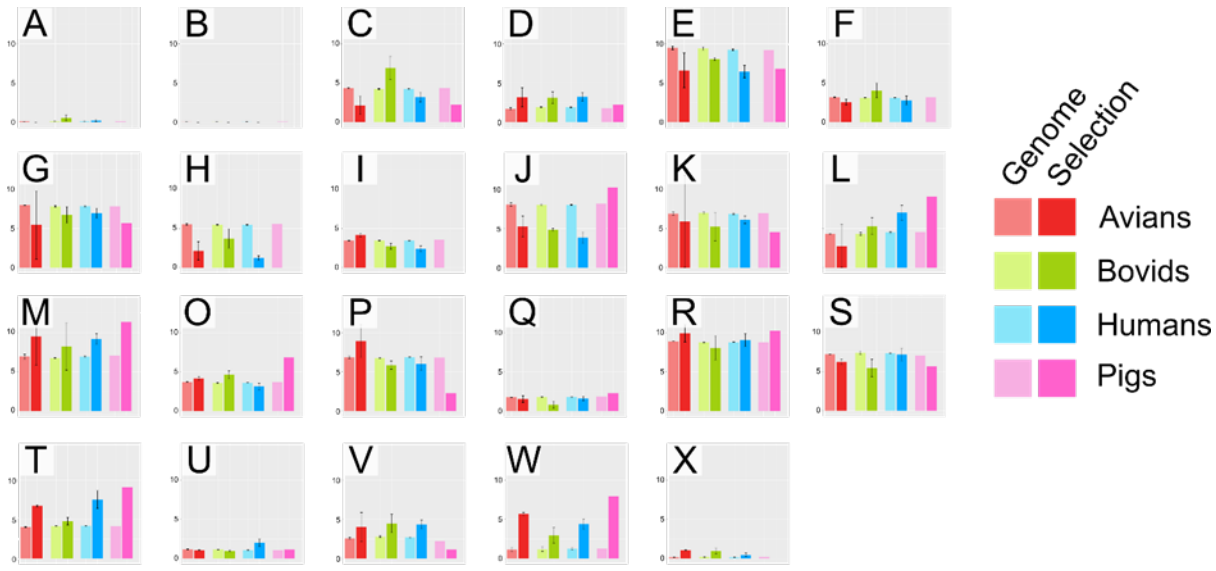
Supplementary Figure 16. Comparison of pseudogene counts by genome according to host species group. The number of pseudogenes in ruminant genomes is significantly higher than birds and human host groups. P-values are shown as *, ** and ***, representing p-values of 0.05, 0.01 and <0.001 respectively.



Supplementary Figure 17. Box and whisker plot representing the number of functional genes per genome, organized by host group. Birds functional gene count is significantly higher than all other host groups. P-values are shown as *, ** and ***, representing p-values of 0.05, 0.01 and <0.001 respectively.



Supplementary Figure 18. Isolates selected for positive selection analysis indicated by colored text with clonal complex annotated on the Maximum-likelihood tree.



A - RNA processing and modification
B - Chromatin structure and dynamics
C - Energy production and conversion
D - Cell cycle control, cell division, chromosome partitioning
E - Amino acid transport and metabolism
F - Nucleotide transport and metabolism
G - Carbohydrate transport and metabolism
H - Coenzyme transport and metabolism
I - Lipid transport and metabolism
J - Translation, ribosomal structure and biogenesis
K - Transcription
L - Replication, recombination and repair
M - Cell wall/membrane/envelope biogenesis
O - Posttranslational modification, protein turnover, chaperones
P - Inorganic ion transport and metabolism
Q - Secondary metabolites biosynthesis, transport and catabolism
R - General function prediction only
S - Function unknown
T - Signal transduction mechanisms
U - Intracellular trafficking, secretion, and vesicular transport
V - Defense mechanisms
W - Extracellular structures
X - Mobilome: prophages, transposons

Supplementary Figure 19. Differences in Clusters of Orthologous Groups for genes under positive selection in different hosts compared to the pangenome.

Supplementary Table 1. Metadata for all *S. aureus* isolates examined in the current study.

Supplementary Table 2: Substitution rate priors (x 10^{-6} nucleotides/site/year) used for dating analysis.

phylogroup	median	95%HPD		sample size	reference
ST22-A	1.27	1.15	1.41	162	{Holden, 2013 #127}
ST239	2.25	1.199	2.52	63	{Harris, 2010 #116}
CC8	1.8	1.2	2.4	174	{Strommenger, 2014 #2502}
ST225	2	1.2	2.9	73	{Nübel, 2010 #218}
CC30	1.42	1.04	1.8	87	{McAdam, 2012 #562}
ST8	1.22	0.604	1.86	387	{Uhlemann, 2014 #2312}

Supplementary Table 3: Classification of isolates into host-species groups for BEAST analysis.

Host classification*	**Number of Samples**	**Further descriptors**
Birds	19	1: corn crake 1: chaffinch 14: chicken 3: turkey
Cows	92	
Goats	22	
Carnivores	11	5: dog 3: cat 2: mongoose 1: lion
Horses	18	
Humans	477	
Rabbits	10	1: mountain hare 9: rabbit
Sheep	8	
Rodents	11	1: guinea pig 3: mara 1: capybara 6: red squirrel
Pigs	20	
Unclassified	8	3: fruit bat 1: hedgehog 1: zoo chimpanzee 2: harbour porpoise 1: tapir
total	696	

*The sequences were grouped into 10 host types with each group containing at least 5 isolates. 8 sequences were grouped as 'unclassified' and were omitted from further analyses.

Supplementary Table 4. Different approaches employed for quantification of host jump events

Name	#taxa	#hosts	Analysis type	Comments
252-Empirical-Host10	252	10	Uses a pre-computed set of posterior time scaled trees and an asymmetric discrete trait model with Markov Jumps	Main analysis
252-Empirical-TipSwap10	252	10	As above but with the host-type labels permuted	Used as 'null model' to compare with the above
252-RAxML-Host10	252	10	Uses a pre-computed set of bootstrapped trees from RAxML and an asymmetric discrete trait model with Markov Jumps	Alternative topology and ancestral reconstructions, used to compare with the main analysis
252-RAxML-TipSwap10	252	10	As above but with the host-type labels permuted	Used as 'null model' to compare with the above
Severe-97-Empirical-Host5	97	5	Uses a pre-computed set of posterior time scaled trees and an asymmetric discrete trait model with Markov Jumps	'Severe' subsampling performed in order to balance the number of taxa per host-type
Severe-97-ConstPop-Host5	97	5	Trees inferred jointly from sequences and traits using an asymmetric discrete trait model with Markov Jumps	To compare with the above
Severe-97-BASTA-Host5	97	5	Trees inferred using BASTA structured coalescent approximation	To compare with the above

Supplementary Table 5. Number of jumps between host-species groups and confidence intervals for all approaches used (as listed in Supplementary Table 4).

Supplementary Table 6. Identification of accessory genes enriched in isolates according to host-species and gain or loss of genes correlated with host-switching events.

Supplementary Table 7. Functional categories (GO terms) of genes correlated with switches into specific host-species.

Gene Ontology	Description	Total Number	Number Observed	Number Expected	P-value	GO Network	Host Group
GO:0051704	multi-organism process	21	18	12.21	0.0046	BP	Human
GO:0009405	pathogenesis	16	14	9.3	0.0097	BP	Human
GO:0044764	multi-organism cellular process	7	7	4.07	0.0205	BP	Human
GO:0065008	regulation of biological quality	7	7	4.07	0.0205	BP	Human
GO:0016772	transferase activity, transferring phosp...	9	8	4.46	0.017	MF	Human
GO:0033013	tetrapyrrole metabolic process	2	2	0.35	0.03	BP	Ruminant
GO:0033014	tetrapyrrole biosynthetic process	2	2	0.35	0.03	BP	Ruminant
GO:0003678	DNA helicase activity	2	2	0.42	0.044	MF	Ruminant
GO:0043170	macromolecule metabolic process	74	10	5.16	0.0041	BP	Birds
GO:0009307	DNA restriction-modification system	3	2	0.21	0.0129	BP	Birds
GO:0044355	clearance of foreign intracellular DNA	3	2	0.21	0.0129	BP	Birds
GO:0006259	DNA metabolic process	47	7	3.28	0.0193	BP	Birds
GO:0071704	organic substance metabolic process	88	10	6.14	0.0197	BP	Birds
GO:0006952	defense response	4	2	0.28	0.0249	BP	Birds
GO:0006807	nitrogen compound metabolic process	78	9	5.44	0.0325	BP	Birds
GO:0090304	nucleic acid metabolic process	65	8	4.53	0.0355	BP	Birds
GO:0006139	nucleobase-containing compound metabolic...	67	8	4.67	0.043	BP	Birds
GO:0003676	nucleic acid binding	108	19	11.9	0.0028	MF	Birds
GO:0004519	endonuclease activity	6	3	0.66	0.0191	MF	Birds
GO:0043565	sequence-specific DNA binding	29	7	3.19	0.0253	MF	Birds
GO:0004518	nuclease activity	7	3	0.77	0.031	MF	Birds
GO:0003677	DNA binding	94	15	10.36	0.0404	MF	Birds
GO:0006812	cation transport	7	5	1.47	0.0048	BP	Pig
GO:0006811	ion transport	9	5	1.88	0.0205	BP	Pig

Supplementary Table 8. Functional groups of pseudogenes enriched in S. aureus by host-species

Supplementary Table 9. Isolates selected for positive selection analysis according to the Online Methods.

CC45_Humans_1	CC45_Humans_2	CC45_Humans_3	CC59_Humans_1	CC59_Humans_2
434061	423817	423811	0864N0012	0864N0012
438615	423818	433081	0864N0014	0864N0014
4414311	438626	434051	0864N0015	0864N0015
441433	438674	438612	61223	7068522
4414510	441426	4395311	7229531	7229531
441458	441454	441466	7474263	7474263
613335	443011	441528	7474292	7474292
7068520	446554	7474282	M013	M013
7229649	6133112	8113456	SA40	SA40
7396285	623614	CA347	SA957	SA957

ST30_Humans_1	ST30_Humans_2	ST30_Humans_3	ST239_Humans_1	ST239_Humans_2
435023	438625	433025	434063	434063
4386610	441413	4386212	434064	434064
438662	441421	441411	438669	438669
438675	441461	4414511	441427	612111
441424	441529	4414610	612111	6133212
441453	458473	613311	6133212	613332
443015	613318	613316	613332	7065864
613315	7396260	7065830	7065864	Bmb9393
613334	8113481	8113425	JKD6008	T0131
MRSA252	8728565	8113466	T0131	TW20

CC5_Humans_1	CC5_Humans_2	CC5_Humans_3	ST239_Humans_3	ST8_Humans_1
433027	16035	18583	434063	423816
434054	435018	433024	434064	4330811
4350112	4350211	433026	6133212	433089
435011	438617	433083	613332	434062
438687	4386510	435027	7065864	438656
4395211	438673	441415	Bmb9393	446552
441412	439527	613314	JKD6008	7229483
441428	6133110	6236111	T0131	7396136
441457	613319	7229516	TW20	8113414
4415311	JH1	JH9	Z172	Newman

CC15_Humans_1	CC15_Humans_2	CC15_Humans_3	ST8_Humans_2	ST8_Humans_3
435075	4350210	4350111	433023	4238111
441435	435024	438618	435019	423814
4414612	435075	438683	4386211	434066
441469	441436	439523	4386512	443029
441522	441451	441431	438657	4465510
443017	443012	441469	4386611	789385
443028	6133211	441522	438665	8113414
623616	613331	441526	7229692	8113435
7229330	7229329	613321	7893810	8113529
7229695	7229695	7068523	COL	8113575

CC22_Humans_1	CC22_Humans_2	CC22_Humans_3	CC12_Humans	CC133_Ruminants_1
438658	4340512	4386710	0864N0092	61233
438672	4386811	4386711	441418	61240
441423	439521	4386712	441452	61243
441524	439525	4395210	443024	61248
441537	441455	441523	623619	61249
443019	4415212	441531	7229312	61252
443025	441533	441534	7229488	61258
6236110	4430111	441535	7229536	973N0009
7065844	458471	443022	8113449	973N0048
8113592	7068517	613313	8113487	ED133

CC151_Cows_1	CC151_Cows_2	973N0083	CC133_Ruminants_2	CC133_Ruminants_3
7068527	7068527	CC151_Cows_3	61234	61233
7068529	7068529	10900259	61235	61234
7068538	7068531	7068529	61237	61235
973N0058	7068538	7068531	61241	61237
973N0061	973N0061	7068538	61243	61240
973N0062	973N0062	973N0061	61244	61243
973N0068	973N0068	973N0062	61249	973N0009
973N0082	973N0076	973N0068	61258	973N0010
973N0083	973N0082	973N0076	973N0010	973N0054
RF122	973N0083	973N0082	ED133	ED133

CC97_Cows_1	CC97_Cows_2	CC97_Cows_3	CC398_Pigs_1	CC398_Pigs_2

10900246	10900246	52701	973N0044	973N0044
52704	52705	80382	SRR445028	SRR445028
52710	80379	80386	SRR445034	SRR445034
7068528	80381	80388	SRR445230	SRR445060
80382	80383	9119280	SRR445239	SRR445236
80386	80386	973N0053	SRR445265	SRR445239
80387	80388	973N0075	SRR445266	SRR445266
973N0057	CHILE5	CHILE5	SRR445281	SRR445276
973N0075	CTH54	CTH54	SRR445286	SRR445281
LMA1166B	RC7	LMA1166B	SRR445291	SRR445291
CC5_Birds_1	**CC5_Birds_2**	**CC5_Birds_3**	**CC385_Birds**	**CC398_Pigs_3**
61267	61267	61267	61263	973N0044
61270	61270	61270	61269	SRR445028
61271	61272	61274	61282	SRR445034
61274	61280	61280	61283	SRR445035
61280	61281	61281	61284	SRR445060
61281	61287	61288	61286	SRR445236
61287	61288	61289	61290	SRR445237
61288	61289	61291	8014464	SRR445266
61289	61291	973N0091	973N0056	SRR445286
973N0091	973N0091	ED98	CC385_Birds	SRR445291

Supplementary Table 10. Proportion of genes exhibiting positive selection in different host species/lineages.

CC/ST	Host	Core genome	Genes under selection	Proportion
CC12	Humans	2444	36	1.47%
CC15	Humans	2406	64	2.66%
CC22	Humans	2596	83	3.20%
CC45	Humans	2488	61	2.45%
CC59	Humans	2395	42	1.75%
CC5	Humans	2456	86	3.50%
ST239	Humans	2571	67	2.61%
ST30	Humans	2558	85	3.32%
ST8	Humans	2545	87	3.42%
CC151	Ruminants	2554	51	2.00%
CC97	Ruminants	2408	111	4.61%
CC133	Ruminants	2512	129	5.14%
CC385	Birds	2419	47	1.94%
CC5	Birds	2532	33	1.30%
CC398	Pigs	2445	42	1.72%

Supplementary Table 11. Functional categories (GO terms) of genes under positive selection in different host species

Supplementary Table 12. Distribution of anti-microbial resistance determinants according to host-species group

Supplementary Notes

Further details for the validation of Host Jump Analysis.

Figures 5-11 relate to the host jump analysis using BEAST. The number of host jumps between host-types can be estimated using a discrete trait model with Markov Jumps on phylogenetic trees. To assess the robustness of the main analysis of 10 subsamples of 252 taxa each (figure 2 and supplementary figures 2 & 4), we performed several additional analyses (10 replicates each) as indicated in the table below. These analyses include 'severe balanced' subsamples of 97 taxa each containing 18-20 taxa of 5 host-types (Birds, Cattle, Humans, Pigs, Sheep&Goats), and ancestral state and host-jumps using the BASTA approximation to the structured coalescent. BASTA can help with dealing with population sampling bias issues when trying to infer migration rates between populations, however due to the large number of sequences and host-types, and diverse nature of the main analysis we were only able to perform it on the 'severe balanced' subsamples.

Consistently with the methods employed, there is strong evidence of human-cattle transmissions, human transmission into other animal species (notably goats and the carnivora) and transmission from other animal species, particularly cows, birds and pigs into humans. Furthermore, the analyses indicate that transmissions have been ongoing for thousands of years, and humans are an important source of S. aureus for domesticated animals. Number of jumps from/to the host-types and confidence intervals can be found in Supplementary Table 5.