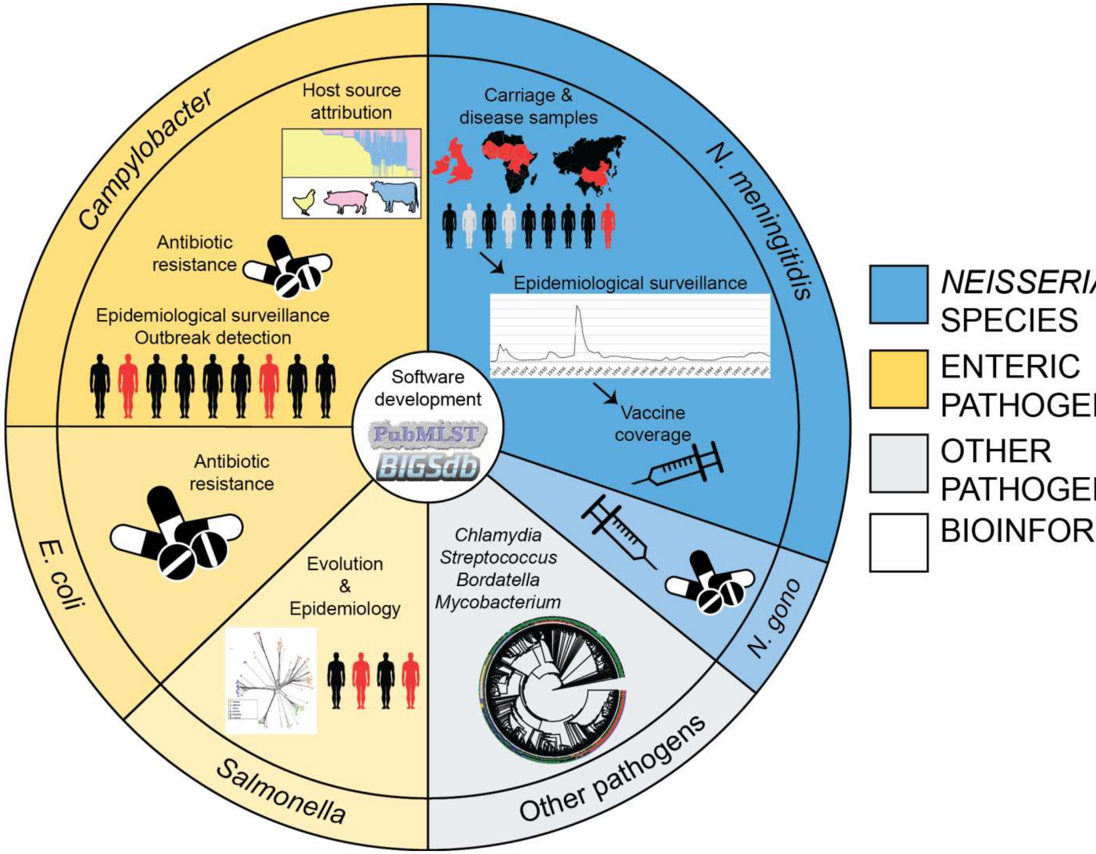
Genomics: the power of WGS as a research and public health tool on a global scale and the Global Meningitis Genome Library initiative

Martin Maiden

Department of Zoology, Peter Medawar Building

UNIVERSITY OF OXFORD

MaidenLab: Bacterial population genomics



https://www.zoo.ox.ac.uk/maiden-lab; https://pubmlst.org

Acknowledgements

aiden Laboratory

olly Bratcher
mes Bray
anny Diallo
dile Harrison
orothea Hill
elissa Jansen van ensburg
eith Jolley
enny MacLennan
artin Maiden
harlene Rodrigues
argaret Varga

Brueggemann Laboratory

Angela Brueggemann
Melissa Jansen van Rensburg

NIHR MEVacP Fellows

Asmaa Shaboum
Naima Nasir
Adeniyi Aderoba
Grace Mzumara
Lilian C. Ibe
Vitalis F. Feteh

wellcome

NIHR | National Institute for Health Resear

UKaid
from the British people

Meningitis Research Foundation

Department of Health & Social Care

Medawar
pathogen research

BIG DATA INSTITUTE

UNIVERSITY OF OXFORD

ecdc
EUROPEAN CENTRE FOR DISEASE PREVENTION AND CONTROL

SEVENTH FRAMEWORK PROGRAMME

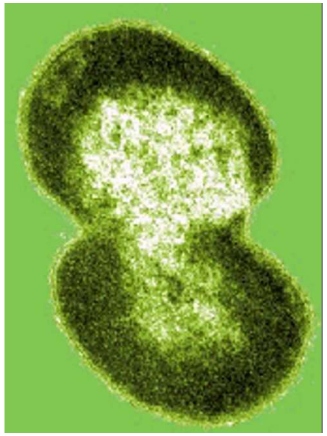
The elements of infectious disease

Place

Time

Person

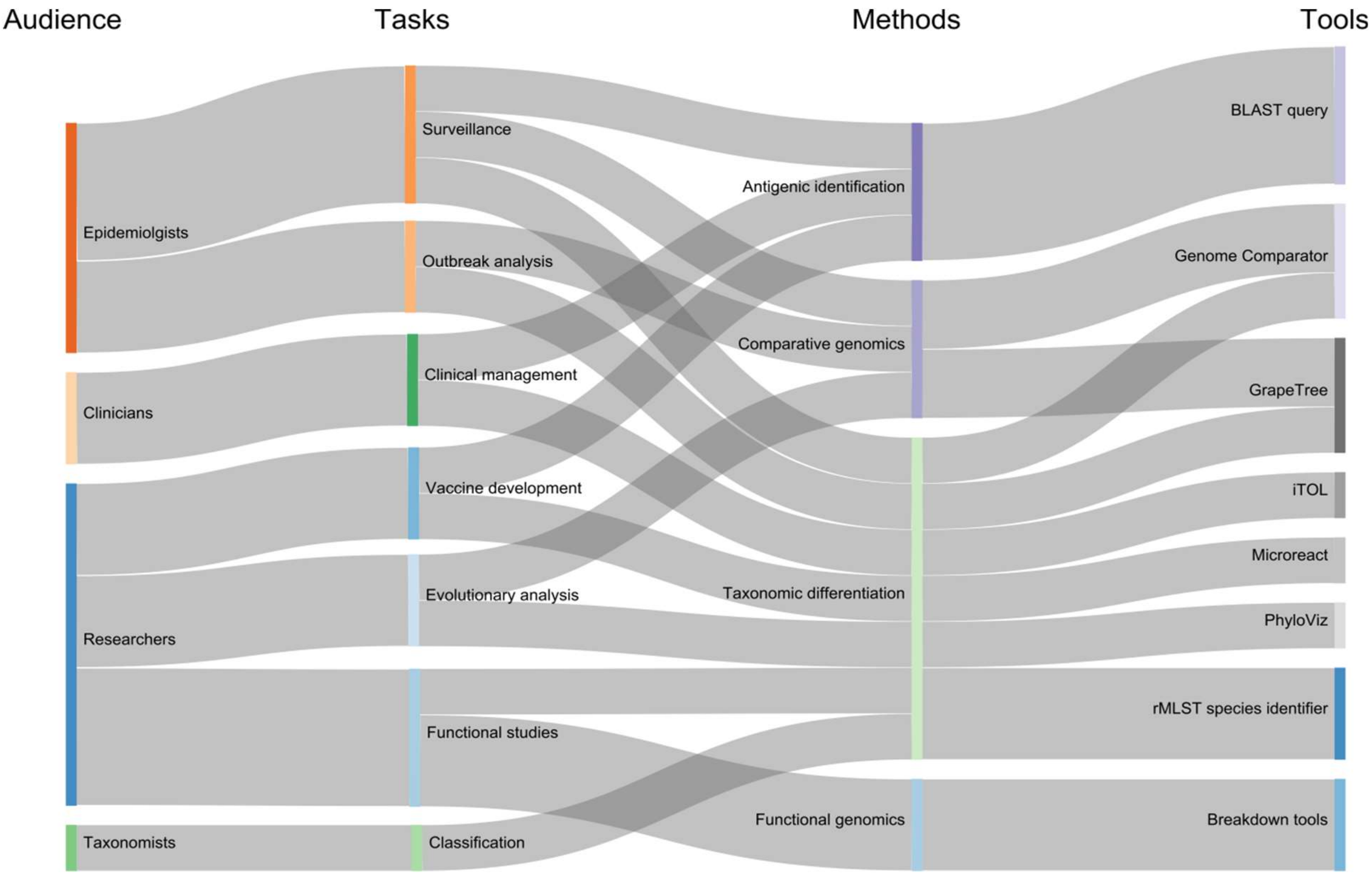
Pathogen



Pathology

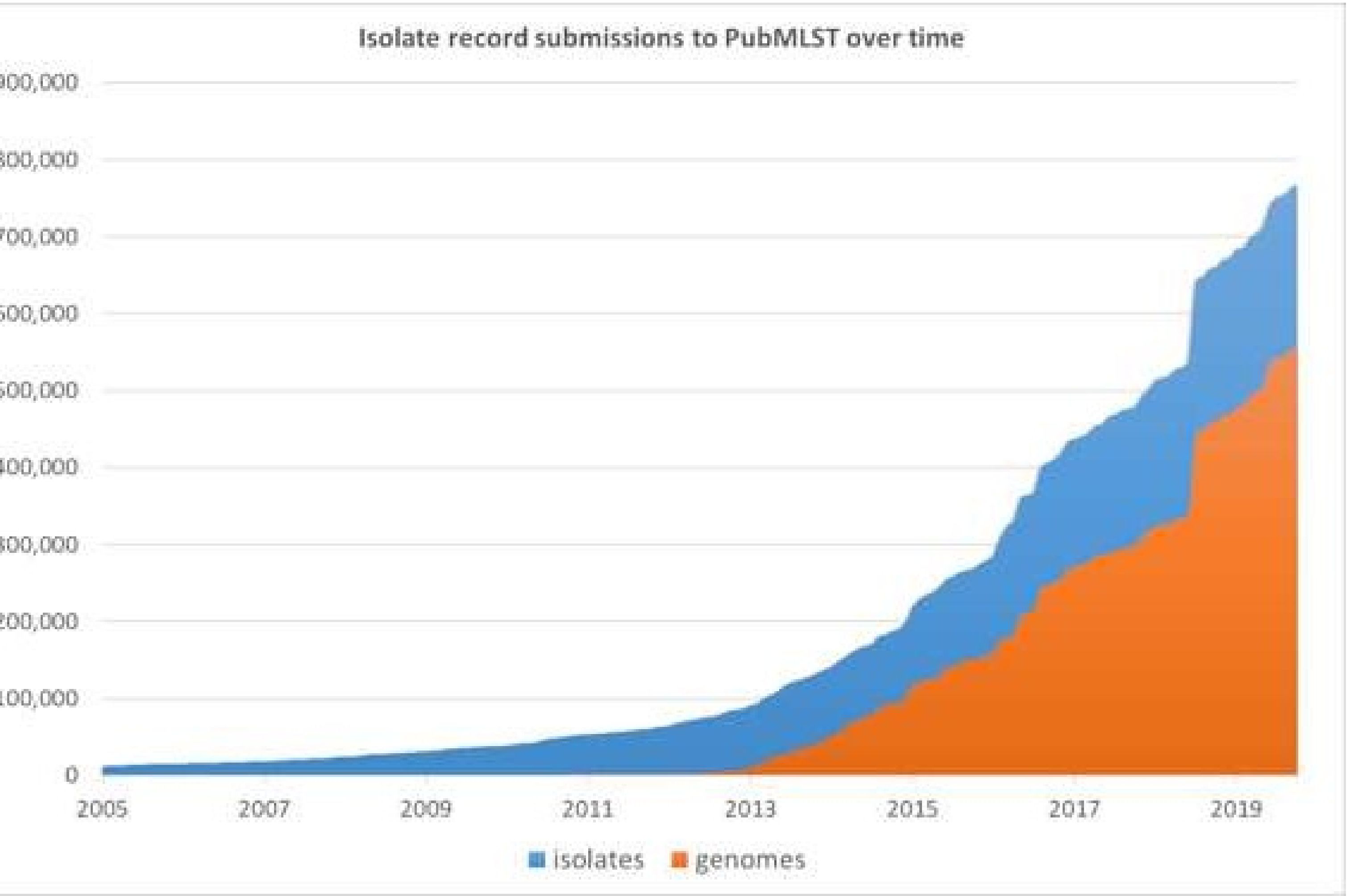
Charlotte Cleverley-Bisman
http://www.charlottecleverleybisman.com/

The potential of genome data



Jolley, K. A., Bray, J. E., and Maiden, M. C. J. (2018) Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* **3**, 124

The growth in genome data over time



There is an ever-increasing volume of isolate data being submitted to databases such as PubMLST.org with an increasing proportion of genome data.

Jolley, K. A., Bray, J. E., and Maiden, M. C. J. (2018) Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* **3**, 124

Genome data: Some requirements and challenges

Requirements

- Comprehensive (representative)
- High (known) data quality
- Availability (Open access)
- Data interpretation
- Information dissemination, sharing and visualization
- Long-term sustainability

Challenges

- Ownership, ethical use (Nagoya)
- Data standards & quality metrics
- Privacy; Patents; Papers
- Assembly, integration, annotation
- Scale and the right information for the right audience
- Hosting and funding

Building a Global Genome Library: MRF Funding from July 2019-2022*

Core concept

- The creation of a 'Global Meningitis Genome Library (GMGL)', which contains globally representative whole genome sequence data for each of the four main bacterial causes of meningitis:
    - *Neisseria meningitidis* (the meningococcus);
    - *Streptococcus pneumoniae* (the pneumococcus);
    - *Streptococcus agalactiae* (GBS); and
    - *Haemophilus influenzae* (type b, Hib)
- This library will build on the success of the Meningitis Research Foundation Meningococcal Genome Library (MRF-MGL).

*Made possible by charitable donations from Pfizer, GSK, and Sanofi Pasteur.

Building a Global Genome Library

Meningit
Research
Foundati

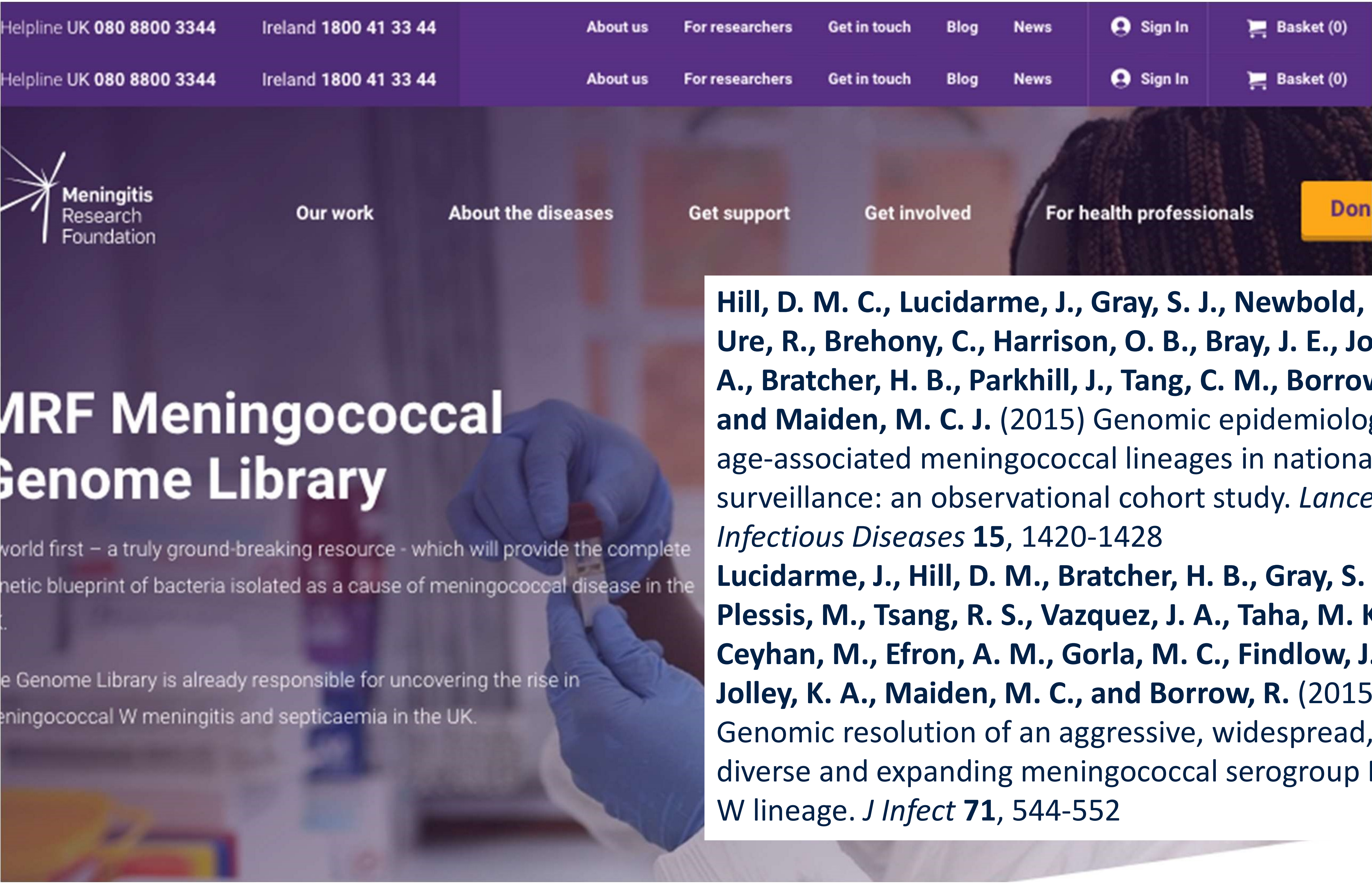
Funding award by the MRF Trustees (July 2019) for:

i. the integration of the GMGL within PubMLST.org;

ii. the assembly, data checking and upload of genome sequences;

iii. curation of individual databases;

iv. the development of interfaces for data integration and display.



Helpline UK **080 8800 3344** Ireland **1800 41 33 44** About us For researchers Get in touch Blog News Sign In Basket (0)

Helpline UK **080 8800 3344** Ireland **1800 41 33 44** About us For researchers Get in touch Blog News Sign In Basket (0)

MRF Meningococcal
Genome Library

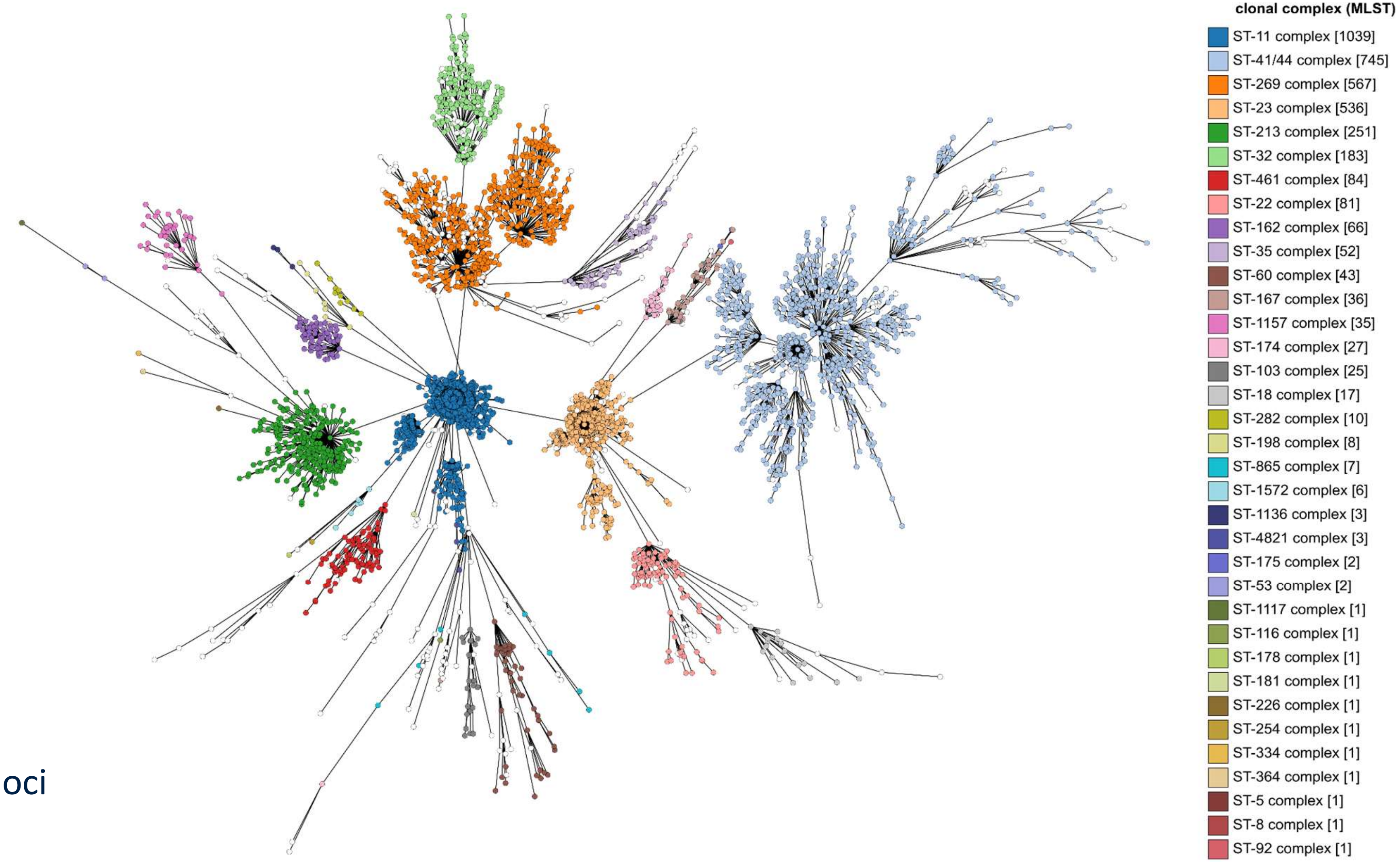
vorld first – a truly ground-breaking resource - which will provide the complete
netic blueprint of bacteria isolated as a cause of meningococcal disease in the
(.

e Genome Library is already responsible for uncovering the rise in
eningococcal W meningitis and septicaemia in the UK.

Hill, D. M. C., Lucidarme, J., Gray, S. J., Newbold,
Ure, R., Brehony, C., Harrison, O. B., Bray, J. E., Jo
A., Bratcher, H. B., Parkhill, J., Tang, C. M., Borrov
and Maiden, M. C. J. (2015) Genomic epidemiolo
age-associated meningococcal lineages in nationa
surveillance: an observational cohort study. *Lance*
Infectious Diseases **15**, 1420-1428

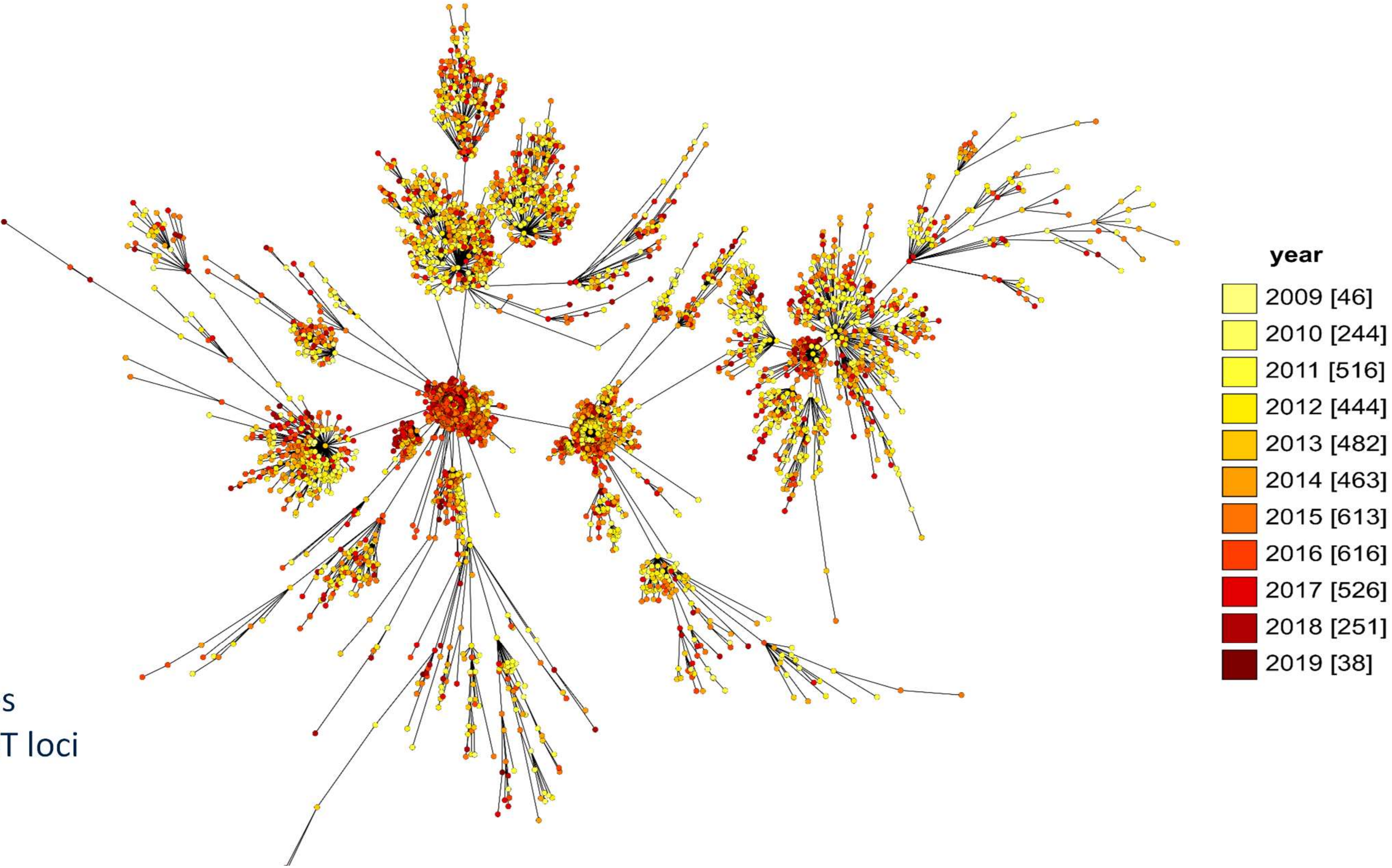
Lucidarme, J., Hill, D. M., Bratcher, H. B., Gray, S.
Plessis, M., Tsang, R. S., Vazquez, J. A., Taha, M. K
Ceyhan, M., Efron, A. M., Gorla, M. C., Findlow, J.
Jolley, K. A., Maiden, M. C., and Borrow, R. (2015
Genomic resolution of an aggressive, widespread,
diverse and expanding meningococcal serogroup
W lineage. *J Infect* **71**, 544-552

MRF-MGL at a glance - genotypes



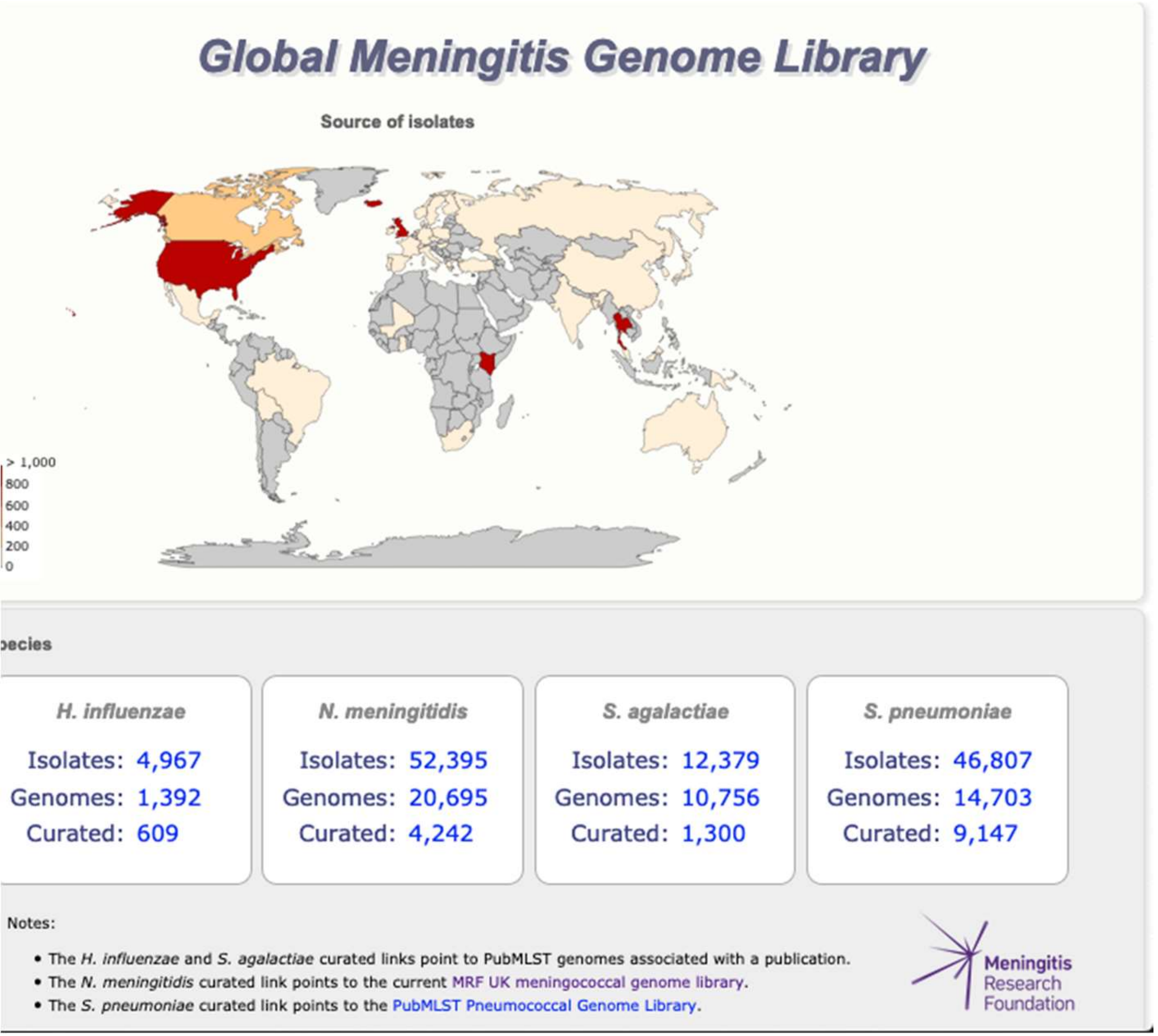
242 isolates
605 cgMLST loci

MRF-MGL at a glance – over time



242 isolates
605 cgMLST loci

https://pubmlst.org/gmgl/



Global Meningitis Genome Library

Source of isolates

pecies

H. influenzae	N. meningitidis	S. agalactiae	S. pneumoniae
Isolates: 4,967	Isolates: 52,395	Isolates: 12,379	Isolates: 46,807
Genomes: 1,392	Genomes: 20,695	Genomes: 10,756	Genomes: 14,703
Curated: 609	Curated: 4,242	Curated: 1,300	Curated: 9,147

Notes:

- The *H. influenzae* and *S. agalactiae* curated links point to PubMLST genomes associated with a publication.
- The *N. meningitidis* curated link points to the current MRF UK meningococcal genome library.
- The *S. pneumoniae* curated link points to the PubMLST Pneumococcal Genome Library.

Meningitis Research Foundation

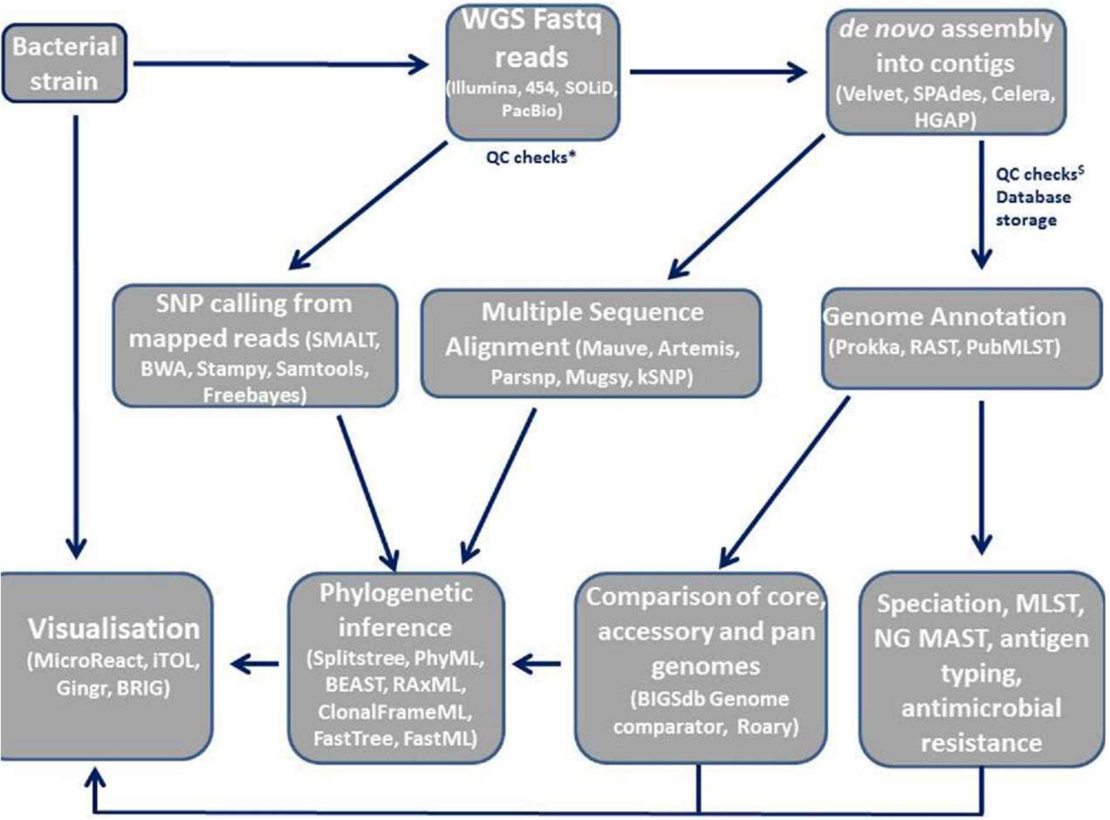
- Coordinated genome library for four major meningitis and sepsis causing bacteria.
- Open access curated isolate records and genomes, includes:
  - MRF Meningitis Genome Library;
  - Pneumococcal Genome Library.
- Linked genome and provenance and phenotype data
- Web/API access

Medawa
pathogen research

BIG DATA INSTITUTE

UNIVERSITY OF
OXFOR

Gene-by-gene genome analysis and omics



Bacterial strain

WGS Fastq reads (Illumina, 454, SOLiD, PacBio)

QC checks*

de novo assembly into contigs (Velvet, SPAdes, Celera, HGAP)

QC checks$^{\$}$ Database storage

SNP calling from mapped reads (SMALT, BWA, Stampy, Samtools, Freebayes)

Multiple Sequence Alignment (Mauve, Artemis, Parsnp, Mugsy, kSNP)

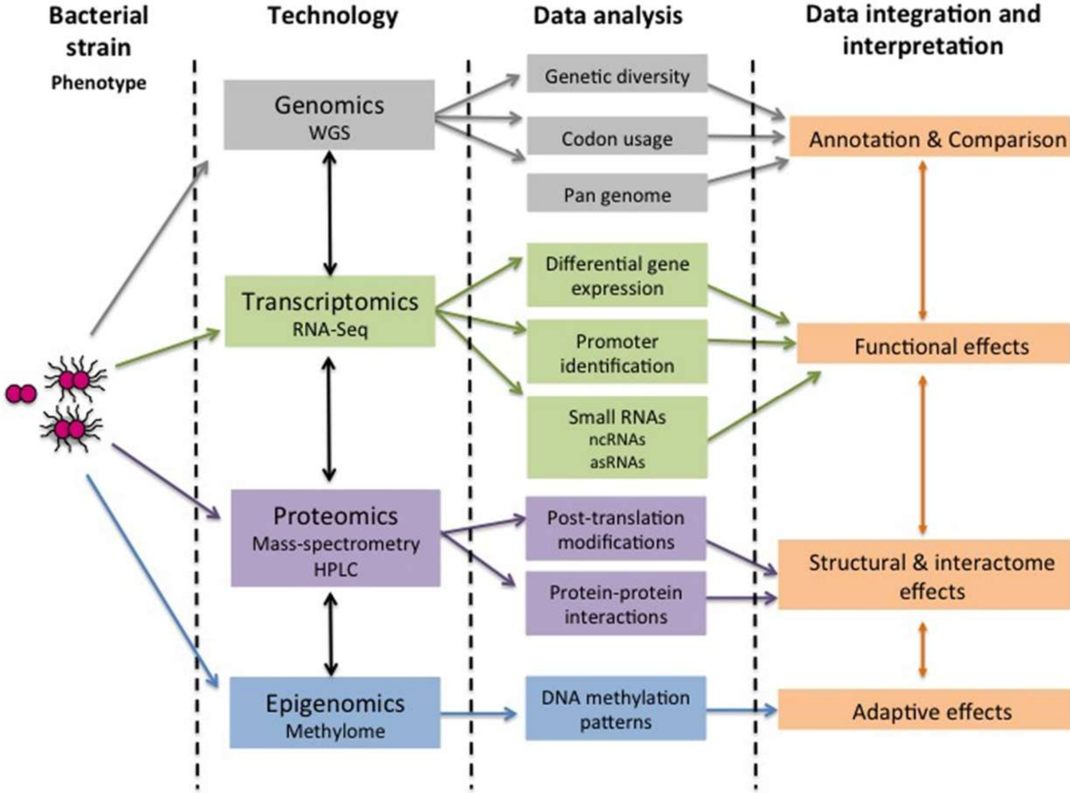
Genome Annotation (Prokka, RAST, PubMLST)

Visualisation (MicroReact, iTOL, Gingr, BRIG)

Phylogenetic inference (Splitstree, PhyML, BEAST, RAxML, ClonalFrameML, FastTree, FastML)

Comparison of core, accessory and pan genomes (BIGSdb Genome comparator, Roary)

Speciation, MLST, NG MAST, antigen typing, antimicrobial resistance



Bacterial strain

Phenotype

Technology

Data analysis

Data integration and interpretation

Genomics
WGS

Transcriptomics
RNA-Seq

Proteomics
Mass-spectrometry
HPLC

Epigenomics
Methylome

Genetic diversity

Codon usage

Pan genome

Differential gene expression

Promoter identification

Small RNAs
ncRNAs
asRNAs

Post-translation modifications

Protein-protein interactions

DNA methylation patterns

Annotation & Comparison

Functional effects

Structural & interactome effects

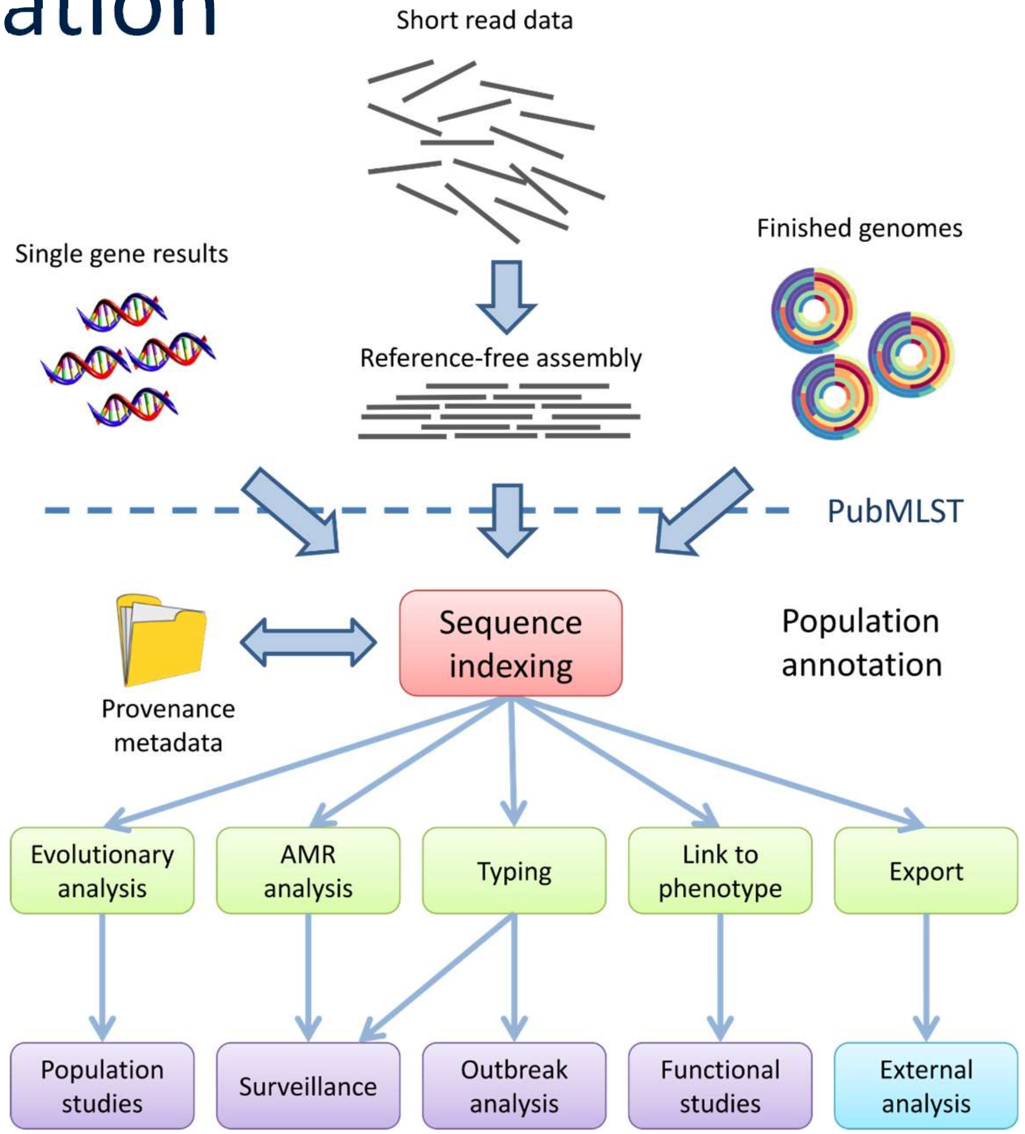
Adaptive effects

Harrison, O. B., Schoen, C., Retchless, A. C., Wang, X., Jolley, K. A., Bray, J. E. & Maiden, M. C. J. (2017). *Neisseria* genomics: current status and future perspectives. *Pathog Dis* 75.

The PubMLST platform: data analysis, curation & dissemination

- Data generation:
  - individual gene sequences;
  - *de novo* assembled draft genomes;
  - finished genomes.
- Deposition into PubMLST (BIGSDB):
  - annotation of loci;
  - linkage to provenance/phenotype information (metadata);
  - comparative analyses.
- Data reporting, sharing, and export.

atcher, H. B., Corton, C., Jolley, K. A., Parkhill, J., and aiden, M. C. (2014) A gene-by-gene population genomics tform: *de novo* assembly, annotation and genealogical alysis of 108 representative *Neisseria meningitidis* nomes. *BMC Genomics* **15**, 1138



Short read data

Single gene results

Reference-free assembly

Finished genomes

PubMLST

Provenance metadata

Sequence indexing

Population annotation

Evolutionary analysis

AMR analysis

Typing

Link to phenotype

Export

Population studies

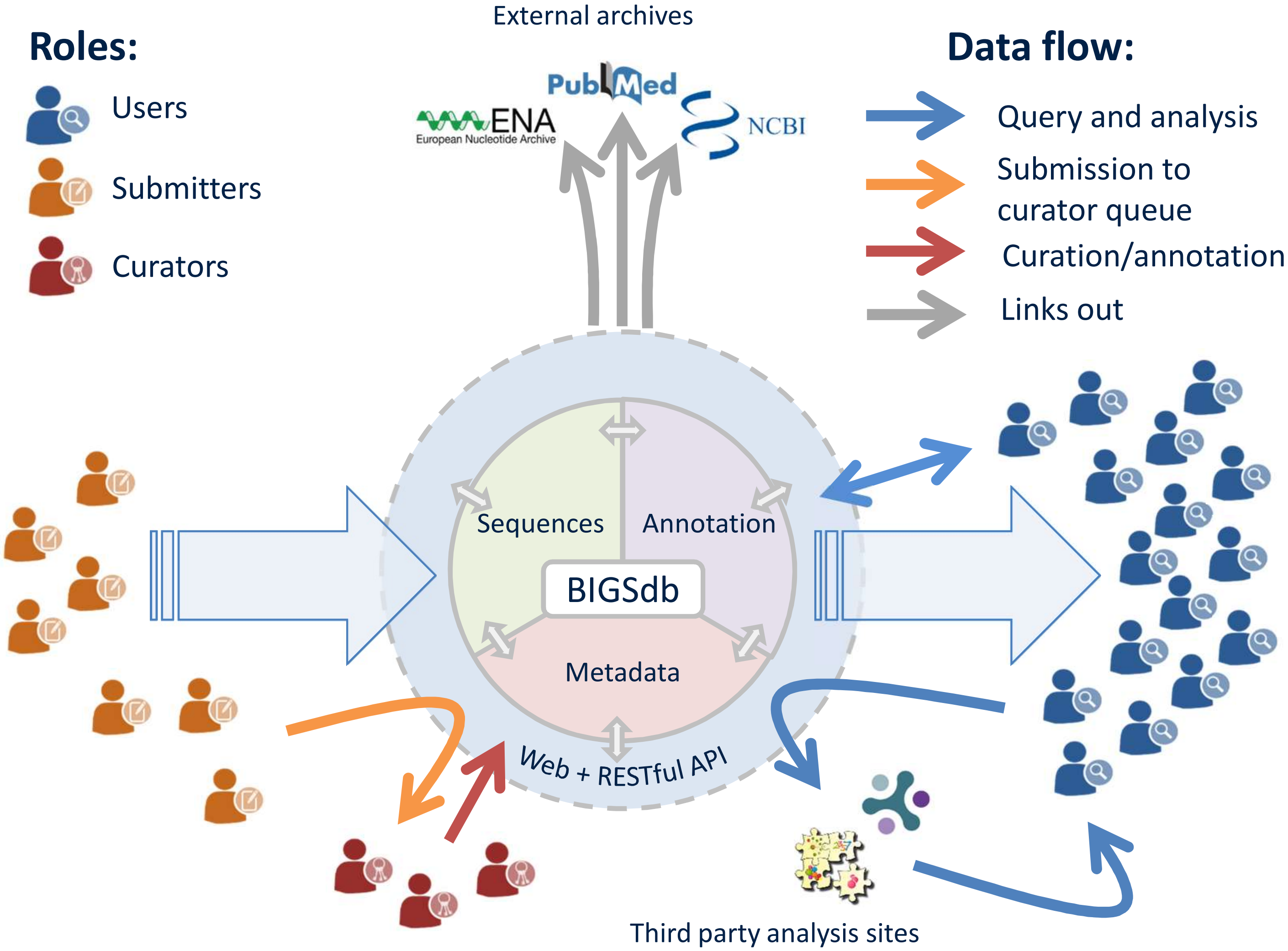
Surveillance

Outbreak analysis

Functional studies

External analysis

The PubMLST platform: dataflows



Roles:

- Users
- Submitters
- Curators

External archives

Data flow:

- Query and analysis
- Submission to curator queue
- Curation/annotation
- Links out

Sequences

Annotation

BIGSdb

Metadata

Web + RESTful API

Third party analysis sites

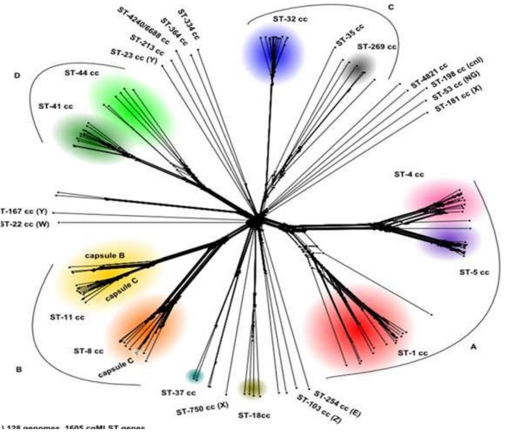
Jolley K.A., Chan M.S., Maic **M.C.** (2004) mlstdbNet - distributed multi-locus sequ typing (MLST) databases. *BM Bioinformatics*. 5, 86.

Jolley K.A., Maiden M.C. (20 BIGSdb: Scalable analysis of bacterial genome variation a the population level. *BMC Bioinformatics*. **11,** 595.

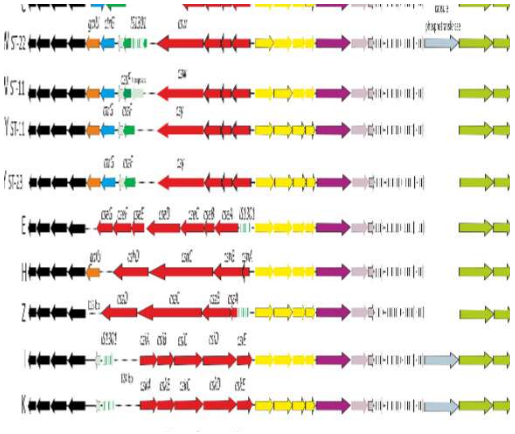
Jolley K.A., Maiden M.C. (20 Using MLST to study bacteri variation: prospects in the genomic era. *Future Microbi* 623-30.

Automated genome scanning provides…

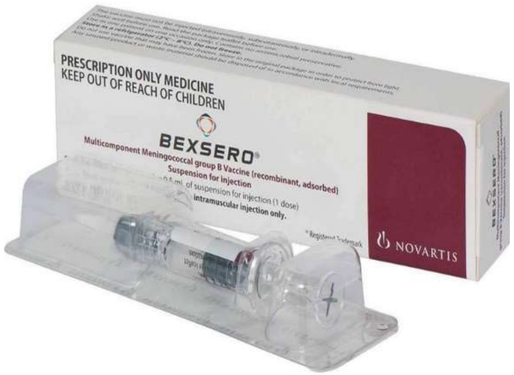
Typing/Phylogeny



Capsular group

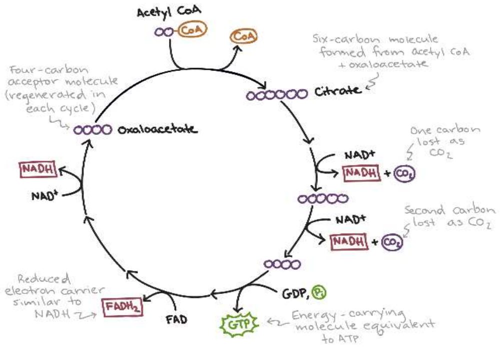


Vaccine coverage



Antibiotic resistance

Metabolic genes

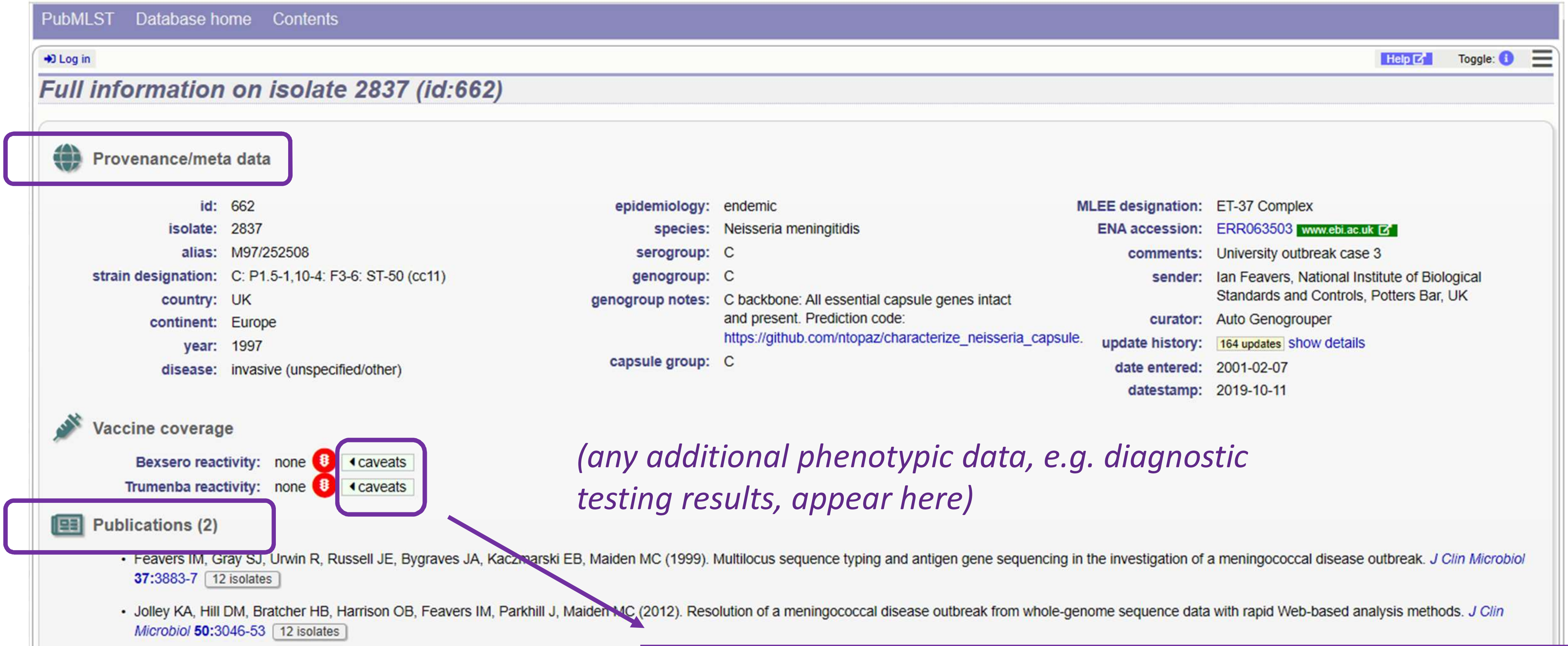


… and much, much more…

PubMLST: the ‘three Ps’, Private Data

- Most data in PubMLST are public – this is the default status.
- Uploading of private user data supported (privacy, publications, patents):
  - Facilitates analysis within context of public data;
  - User requires submitter privileges;
  - Data remain in control of the submitter until they decide to publish it.

Provenance and phenotype



Full information on isolate 2837 (id:662)

Provenance/meta data

id: 662
isolate: 2837
alias: M97/252508
strain designation: C: P1.5-1,10-4: F3-6: ST-50 (cc11)
country: UK
continent: Europe
year: 1997
disease: invasive (unspecified/other)
epidemiology: endemic
species: Neisseria meningitidis
serogroup: C
genogroup: C
genogroup notes: C backbone: All essential capsule genes intact and present. Prediction code: https://github.com/ntopaz/characterize_neisseria_capsule.
capsule group: C
MLEE designation: ET-37 Complex
ENA accession: ERR063503 www.ebi.ac.uk
comments: University outbreak case 3
sender: Ian Feavers, National Institute of Biological Standards and Controls, Potters Bar, UK
curator: Auto Genogrouper
update history: 164 updates show details
date entered: 2001-02-07
datestamp: 2019-10-11

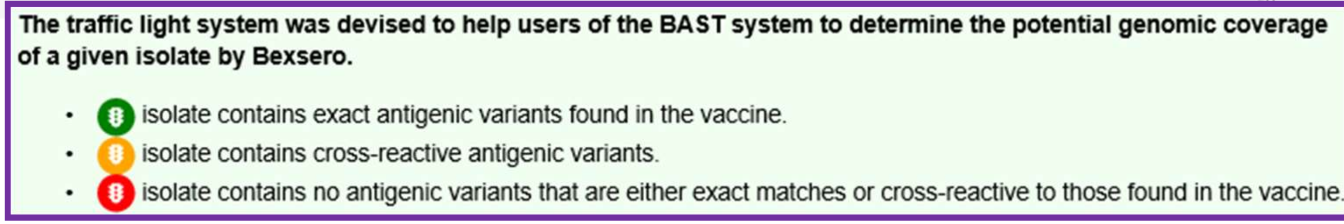
Vaccine coverage

Bexsero reactivity: none 🔴 caveats
Trumenba reactivity: none 🔴 caveats

Publications (2)

- Feavers IM, Gray SJ, Urwin R, Russell JE, Bygraves JA, Kaczmarski EB, Maiden MC (1999). Multilocus sequence typing and antigen gene sequencing in the investigation of a meningococcal disease outbreak. *J Clin Microbiol* **37**:3883-7 12 isolates
- Jolley KA, Hill DM, Bratcher HB, Harrison OB, Feavers IM, Parkhill J, Maiden MC (2012). Resolution of a meningococcal disease outbreak from whole-genome sequence data with rapid Web-based analysis methods. *J Clin Microbiol* **50**:3046-53 12 isolates

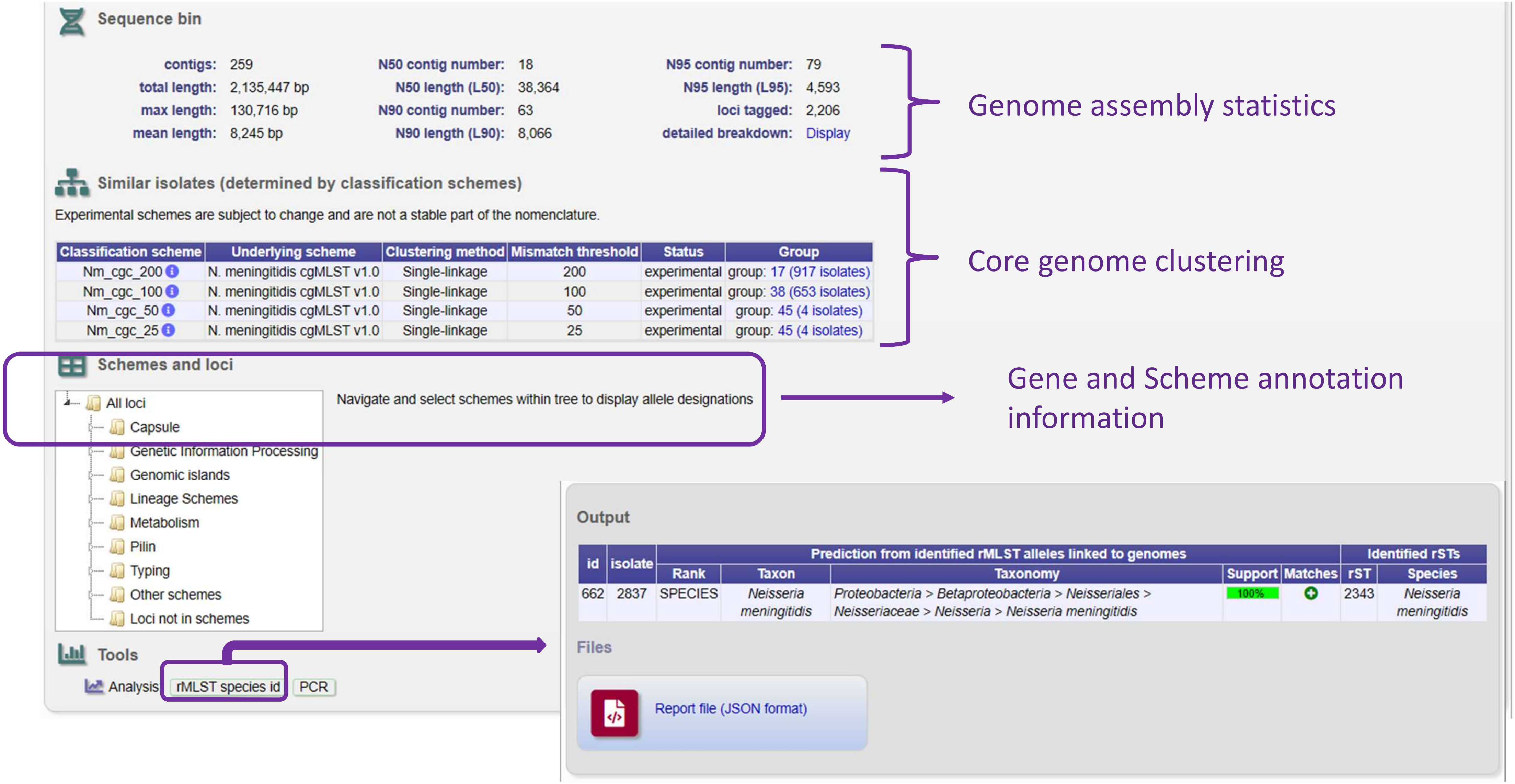
(any additional phenotypic data, e.g. diagnostic testing results, appear here)



The traffic light system was devised to help users of the BAST system to determine the potential genomic coverage of a given isolate by Bexsero.

- 🟢 isolate contains exact antigenic variants found in the vaccine.
- 🟠 isolate contains cross-reactive antigenic variants.
- 🔴 isolate contains no antigenic variants that are either exact matches or cross-reactive to those found in the vaccine.

Genotype and quality metrics



Sequence bin

contigs:	259	**N50 contig number:**	18	**N95 contig number:**	79
total length:	2,135,447 bp	**N50 length (L50):**	38,364	**N95 length (L95):**	4,593
max length:	130,716 bp	**N90 contig number:**	63	**loci tagged:**	2,206
mean length:	8,245 bp	**N90 length (L90):**	8,066	**detailed breakdown:**	Display

Genome assembly statistics

Similar isolates (determined by classification schemes)

Experimental schemes are subject to change and are not a stable part of the nomenclature.

Classification scheme	Underlying scheme	Clustering method	Mismatch threshold	Status	Group
Nm_cgc_200	N. meningitidis cgMLST v1.0	Single-linkage	200	experimental	group: 17 (917 isolates)
Nm_cgc_100	N. meningitidis cgMLST v1.0	Single-linkage	100	experimental	group: 38 (653 isolates)
Nm_cgc_50	N. meningitidis cgMLST v1.0	Single-linkage	50	experimental	group: 45 (4 isolates)
Nm_cgc_25	N. meningitidis cgMLST v1.0	Single-linkage	25	experimental	group: 45 (4 isolates)

Core genome clustering

Schemes and loci

- All loci
  - Capsule
  - Genetic Information Processing
  - Genomic islands
  - Lineage Schemes
  - Metabolism
  - Pilin
  - Typing
  - Other schemes
  - Loci not in schemes

Navigate and select schemes within tree to display allele designations

Gene and Scheme annotation information

Tools

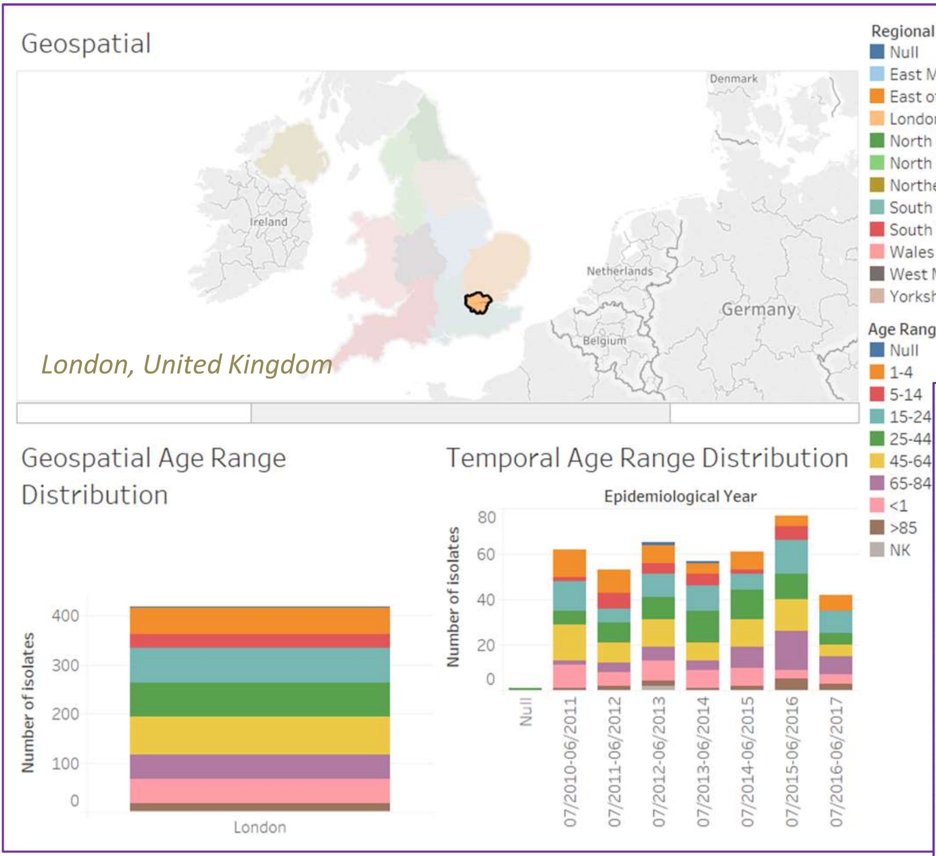
Output

id	isolate	Prediction from identified rMLST alleles linked to genomes					Identified rSTs	
		Rank	Taxon	Taxonomy	Support	Matches	rST	Species
662	2837	SPECIES	*Neisseria meningitidis*	*Proteobacteria > Betaproteobacteria > Neisseriales > Neisseriaceae > Neisseria > Neisseria meningitidis*	100%	⊕	2343	*Neisseria meningitidis*

Files

Report file (JSON format)

Data presentation: Visual analytics



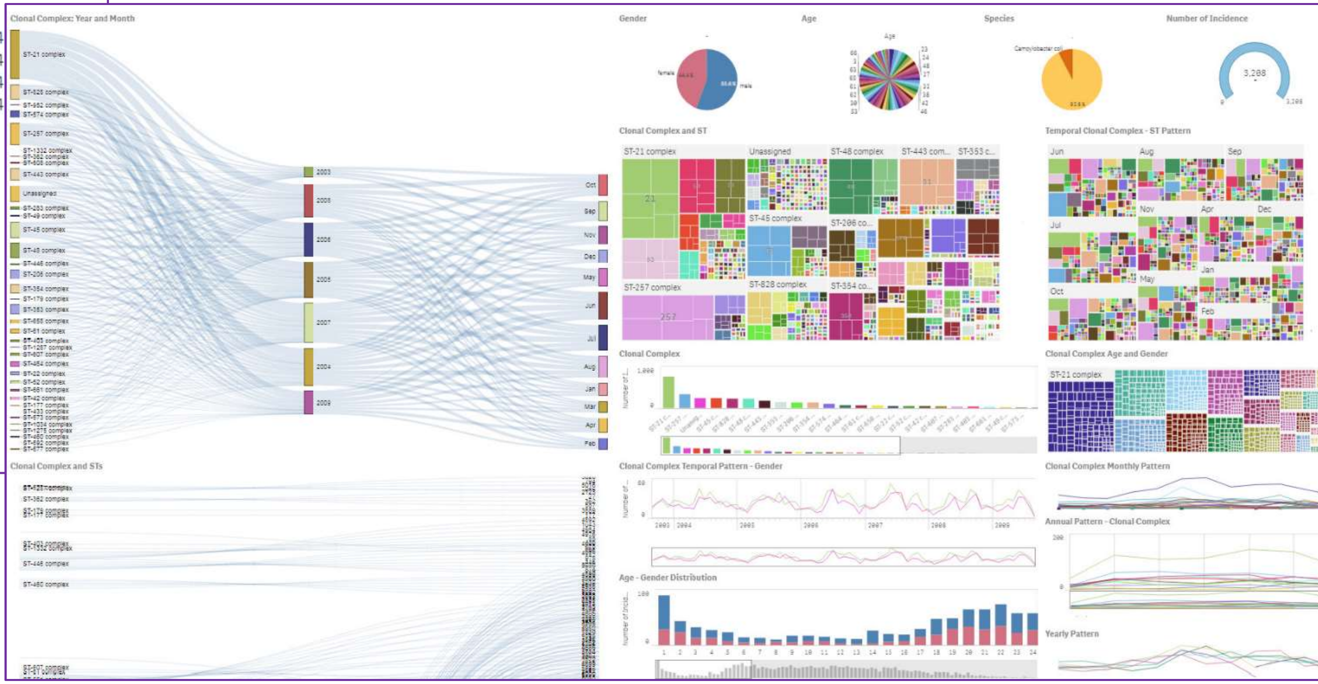
curated, genome sequenced, annotated isolates

PLUS

stratified provenance data

=

DATA VISUALIZATION



PubMLST: Longevity

- Continuous funding, 1998 – 2024.
- Software and data open source.
- New Wellcome Trust Biomedical Resource Grant:
  - £1.2M over 5 years (to Oct 2024)
    - Software development,
    - Visual analytics,
    - Infrastructure.
- External collaborations, including Institut Pasteur (Silvain Brisse)
- New server infrastructure