

# **One Disrupting Technology Fits it All - Towards Standardized Bacterial Whole Genome Sequencing for Global Surveillance**

**Dag Harmsen**

University of Münster, Germany

[dharmen@uni-muenster.de](mailto:dharmen@uni-muenster.de)

# Short CV – Dag Harmsen

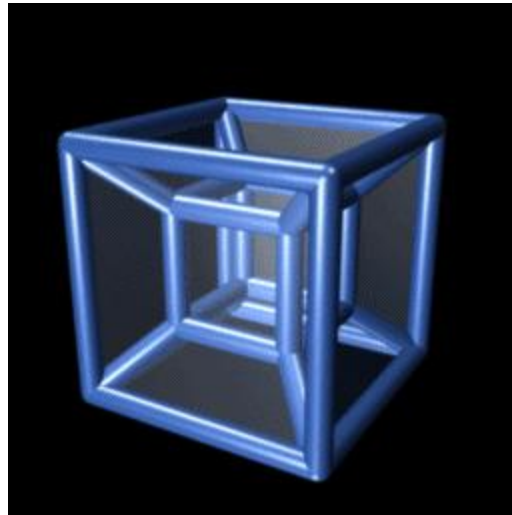


- 1983–1991, Medical Schools University Antwerpen, Belgium and Univ. Würzburg, Germany
- 1991, **MD** and doctoral thesis (‘Nucleo-capsid gene of *Coronavirus HCV-229E*, V. ter Meulen)
- 1992 – 2000, Institute of Hygiene & Microbiology (J. Heesemann, *Candida spp.* & *Aspergillus spp.* & M. Frosch, *Mycobacterium spp.*) / Internal Medicine (K. Kochsiek) / Institute of Virology (V. ter Meulen), Univ. Würzburg, Germany
- 1998, Inclusion in the German **Medical Microbiology and Infectious Epidemiology** specialist register
- 2000-2001, Head of R&D CREATOGEN diagnostics, CREATOGEN AG, Augsburg, Germany
- 2002-2004, Institute of Hygiene (H. Karch, *MRSA*), University of Münster, Germany
- 2003, co-founder and shareholder of a bioinformatics company (**Ridom** GmbH, Münster, Germany)
- 2004– , Full Professorship Department of Periodontology, Univ. Münster, Germany
- July 2005 – July 2008, Temporary Head of the Department of Periodontology, Univ. Münster
- August 2008 – , Head of Research Department of Periodontology
- July 2005 – , Member of the Executive Board of the **International Committee on Systematics of Prokaryotes** (ICSP) of the ‘International Union of Microbiological Societies’ (IUMS)
- June 2007 – July 2012, Member of the **ASM Professional Development Committee**
- October 2012 - , **ASM Ambassador** in **Germany**
- **ECDC & EFSA** technical advisor for genotyping and NGS/WGS
  
- Scientific interests: molecular diagnostic, epidemiology, and phylogeny of microorganisms; applied bioinformatics in microbiology

# Commercial Disclosure

**Dag Harmsen** is co-founder and partial owner of a bioinformatics company (Ridom GmbH, Münster, Germany) that develops software for DNA sequence analysis. Recently Ridom and Ion Torrent/Thermo Fisher (Waltham, MA) partnered and released SeqSphere+ software to speed and simplify whole genome based bacterial typing.

# Fourth Dimension Needed for More Specific Surveillance



**Place, Time, 'Person' ... Type!**

# Fourth Dimension Reloaded

## Next Generation Sequencing - Bench-top Machines



Ion Torrent Personal Genome Machine (PGM)

- Affordable
- **Speed**
- Simple workflow



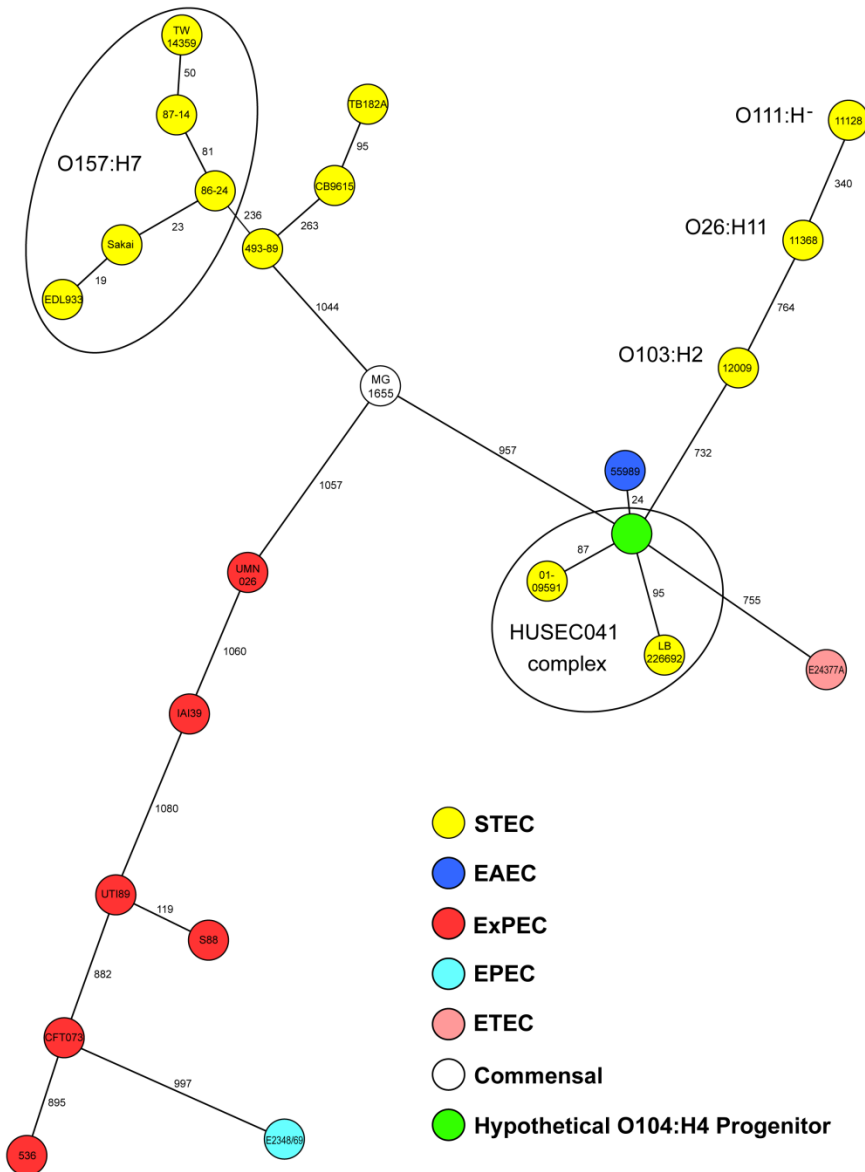
Roche/454 GS Junior



illumina **MiSeq** Personal Sequencing System

# Rapid ,Ad hoc‘ NGS - *E. coli* O104:H4 Outbreak

(Germany May/June, 2011)



## Phylogenetic Analysis of EHEC O104:H4

### Method

- By 'quick and dirty' hybrid reference mapping & *de novo* assemblies of WGS data & BIGSdb **core genome MLST (cgMLST)**
- n = 1.144 core genome genes and minimum-spanning tree

### Results

- Strain LB226692 (outbreak 2011) and strain 01-09591 (2001 German isolate causing historic HUS outbreak) belong to the HUSEC041 complex
- Both strains are only distantly related to commonly isolated EHEC serotypes

# Rapid NGS/WGS Applications in Clinical & Public Health Microbiology

## Applications

**'Ad hoc' epidemiology**

**Diagnostics**

**Therapeutics**

**Global surveillance, early warning & outbreak detection**

## Details

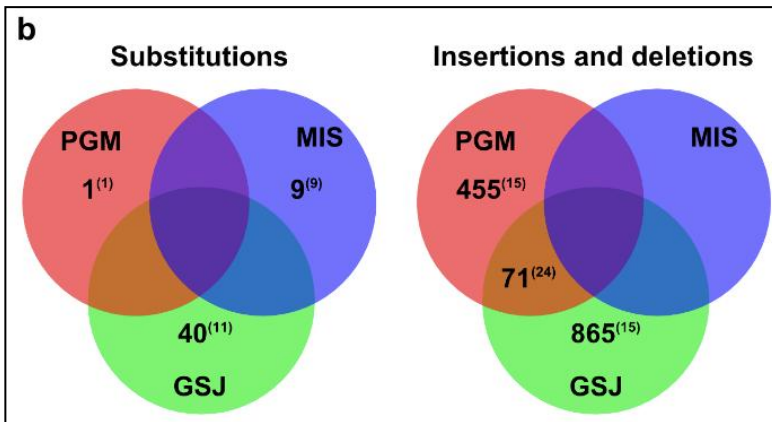
---

- Introduction of benchtop Next Generation Sequencing (NGS) machines, enables small- and medium-sized laboratories ('democratizing of NGS') to perform 'ad hoc' genomic prospective epidemiology
- Speciation / identification & pathogenicity profiling
- Molecular diagnostic screening tests
- Ultra-deep sequencing for pathogen discovery from human tissues (e.g., hemorrhagic viruses)
  
- Susceptibility profiling
- Vaccine preventability
- Reverse vaccinology (rationale vaccine design)
- Non-targeted new drug detection
  
- Standardized Whole Genome Sequencing [WGS] NGS for detection of transmission between individuals
- Outbreak detection, i.e., establishing the spread of particular strains locally, regionally or cross-border
- Longer-term and evolutionary studies to identify the emergence of particularly pathogenic or virulent variants

# It's the Consensus

## Genome-wide Gene by Gene *de novo* Consensus Accuracy

### Venn diagram of *de novo* consensus accuracy for PGM, MiSeq and GSJ



**PGM**, Ion Torrent Personal Genome Machine **300bp**;  
**MiSeq**, Illumina MiSeq **2x 250bp PE**;  
**GSJ**, 454 GS Junior with **GSJ Titanium** chemistry;  
bp, base pairs

### Details

- Consensus **errors** were analyzed for **4,632 coding NCBI Sakai reference genes** retrieved from **MIRA *de novo* assemblies** using **SeqSphere+** for all 3 platforms
- Number of variants confirmed by **bidirectional Sanger sequencing** indicated in parentheses
- Validation of the **8 substitution** and **15 indel** variants identified using all 3 NGS platforms, suggested that either the Sakai strain experienced micro-evolutionary changes or the genome sequence deposited in 2001 contains sequencing errors



# Current NGS Bottlenecks

## Library Prep



AB Library Builder™

## Template Amplification



Ion Chef™

Ion Torrent

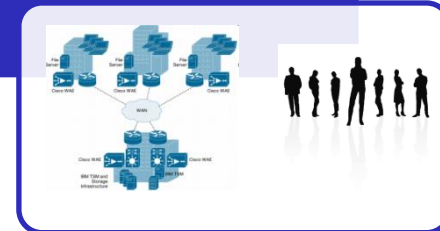
Sample Processing



NGS Platforms



Bioinformatics, IT infrastructure



NeoPrep™

done on the NGS machine

Illumina




NuGen Mondrian



PE NGS Express

Third party



# Read and assembly metrics inconsequential for clinical utility of whole-genome sequencing in mapping outbreaks

## To the Editor:

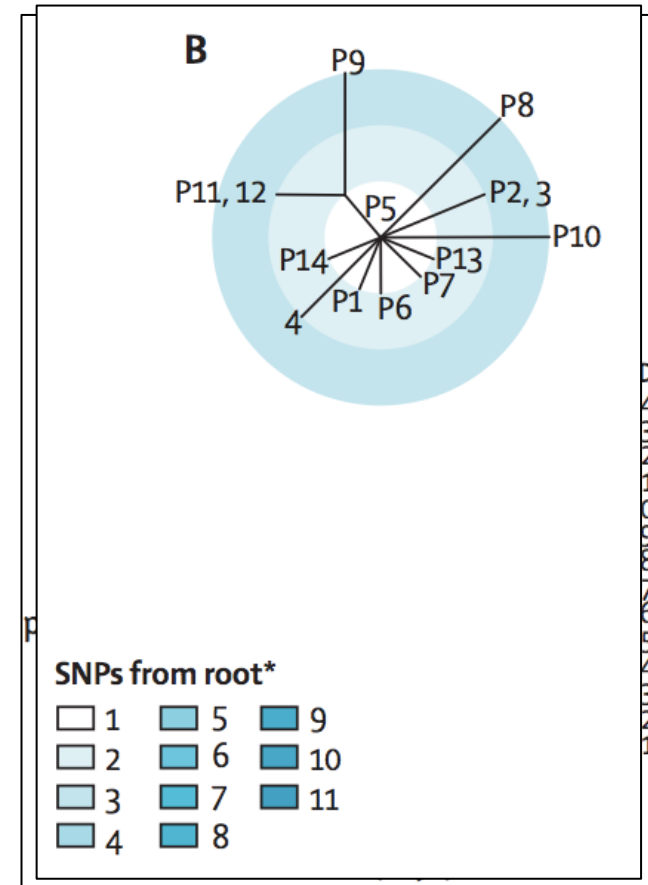
In their paper “Performance comparison of benchtop high-throughput sequencing platforms” published in the May 2012 issue, Loman *et al.*<sup>1</sup> provide a detailed comparison of the metrics associated with three different benchtop DNA sequencing platforms for the assembly of a single genome. Information was given on read-level metrics, such as

length, accuracy and alignment, and on assembly-level metrics, such as contig N50 and gap number. The results were discussed in the context of the utility of whole-genome sequencing for public health microbiology.

We believe, however, that one of the primary uses for sequencing in clinical microbiology (at least initially) will be in the detection of pathogen transmission

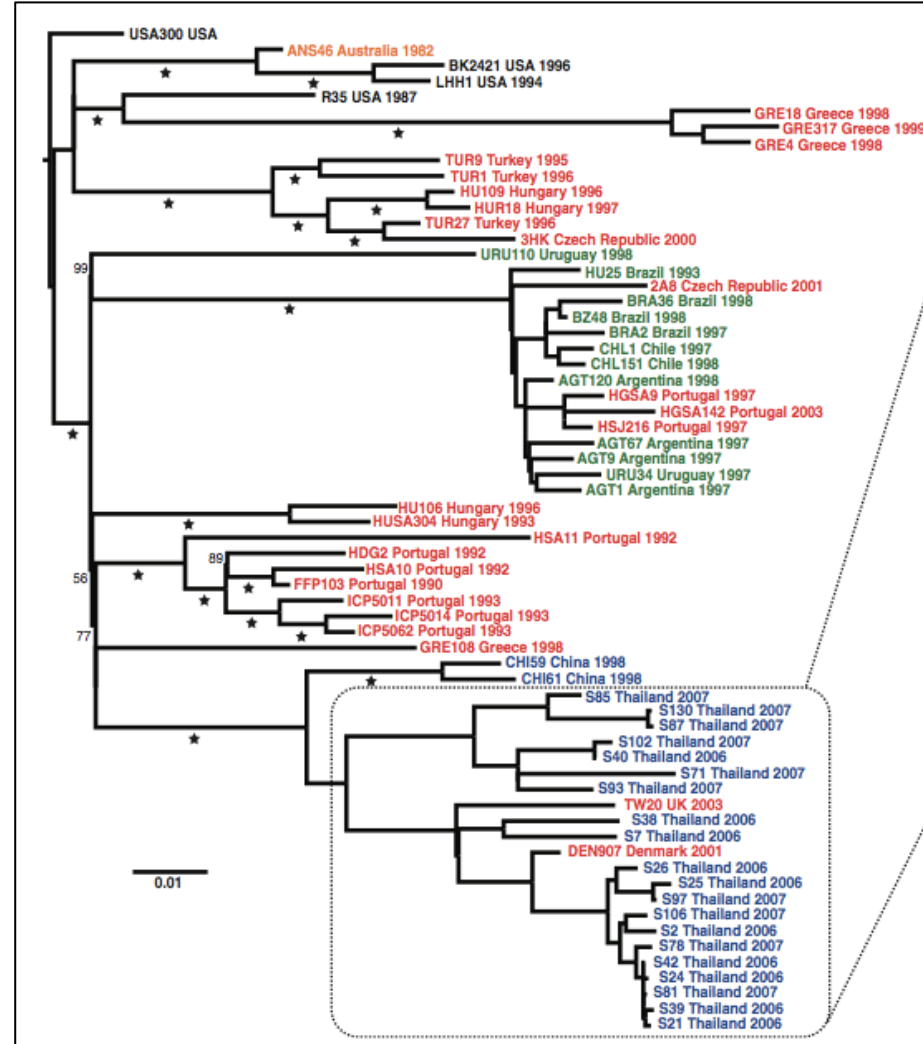
# Mapping & SNP Calling

- MRSA outbreak on a special care baby unit in 6 month period, 2011 UK - **Harris et al.** (2013). *Lancet Infect Dis.* **13**: 130 [[PubMed](#)]
- 15 outbreak (**ST 2371**) and 9 control isolates re-sequenced on Illumina HiSeq & MiSeq and Ion Torrent PGM
- reads were **mapped against** the chromosome of an **EMRSA-15 reference** (HO 50960412; accession number HE681097; **ST 22**, i.e. SLV of ST 2371) and discriminatory single nucleotide polymorphisms (**SNPs**) were identified in the shared core genome of all 24 isolates (majority base needed to be present in at least 75% of reads on each strand → *consensus*)
- all platforms clearly discriminated outbreak from the 9 non-outbreak isolates (with an average of 13,154 SNP differences between both groups for MiSeq and 13,297 SNPs for PGM)
- all platforms identified a total of 23 SNPs among the 15 outbreak isolates
- no strong temporal signature of sequential patient transmission (due to repeated transmission of staff member or slow mutation rate and short outbreaks?)



# Mapping & SNP Calling II

- high-resolution view of the epidemiology and microevolution of a dominant lineage (**ST 239**) of methicillin-resistant *Staphylococcus aureus* (MRSA)
- reads were **mapped** for each isolate **against TW20 reference (ST 239)** and discriminatory single nucleotide polymorphisms (**SNPs**) were identified in the shared core genome
- reveals the global geographic structure within the lineage, its intercontinental transmission through four decades, and the potential to trace person-to-person transmission within a hospital environment
- **Both studies are not comparable!**



The



# genome project

by informaticians, for informaticians

## Goal:

Develop algorithms that scale to arbitrarily large datasets

## Design requirements:

1. Must handle data streams
2. Compute cost to add new genome must be  $\sim O(1)$

'n+1' problem

## Examples:

1. Multiple sequence alignment via profile-HMM
2. Phylogenetic placement on reference tree
3. Bloom filters

## Emerging challenge:

Deleting all the redundant data

n, number of isolates in database

# Surveillance & Phylogeny

'Molecular Typing Esperanto' by Standardized Genome Comparison

## Multiple Genome Alignment

(e.g., progressive Mauve)

- Difficult to interpret with draft genomes
- Computational intensive ( $\geq O(n^2)$ , limit  $\approx 30-50$  genomes)
- Not additive expandable, no nomenclature possible

k-mer  
without alignment

+ Works on read, draft & complete genome level, quickly identifies closest matching genome

- Whole genome reduced to a single number of similarity
- Additively expandable [ $\approx O(n)$ ], but poor mapping to nomenclature possible

ANI

with alignment  
(Average Nucleotide Identity)

+ Works well for monomorphic organisms and 'ad hoc' analysis

- Problematic with rearrangement / recombination events
- Not additive expandable (at least if not always mapped to same reference)

## Genome-wide mapping & SNP calling

## Genome-wide gene by gene allele typing

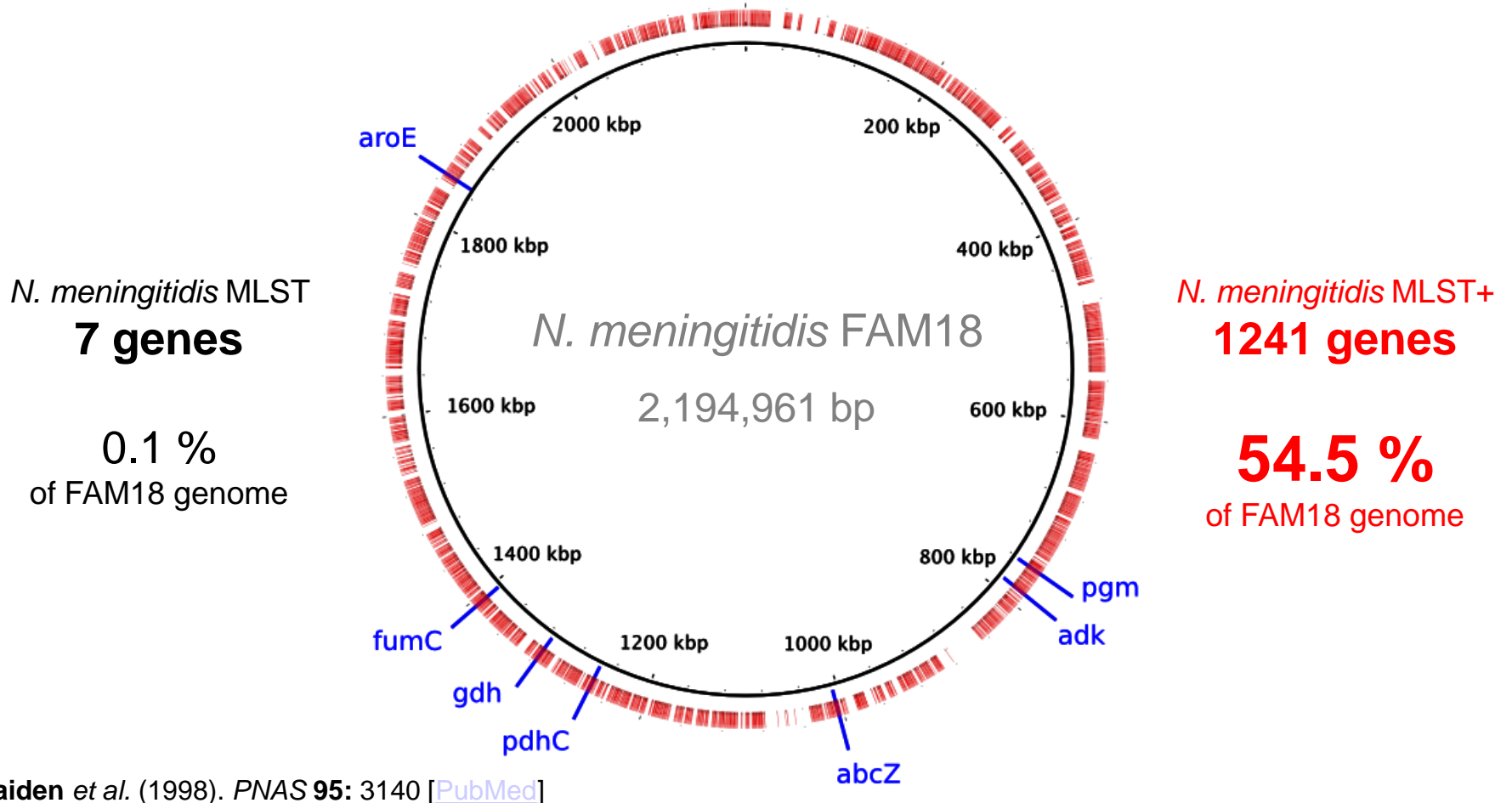
(cgMLST or MLST+)

+ Scalable, working on single gene to whole genome levels

+ Both recombination & point mutation accommodated a single event

+ Additively expandable [ $\approx O(n)$ ] & nomenclature possible

# The Next Generation Sequencing Typing+



## MLST

- 5-7 housekeeping genes
- Sequence type (ST) and Clonal complex (CC)
- Public nomenclature

Used mainly for **population genetics & evolutionary studies**

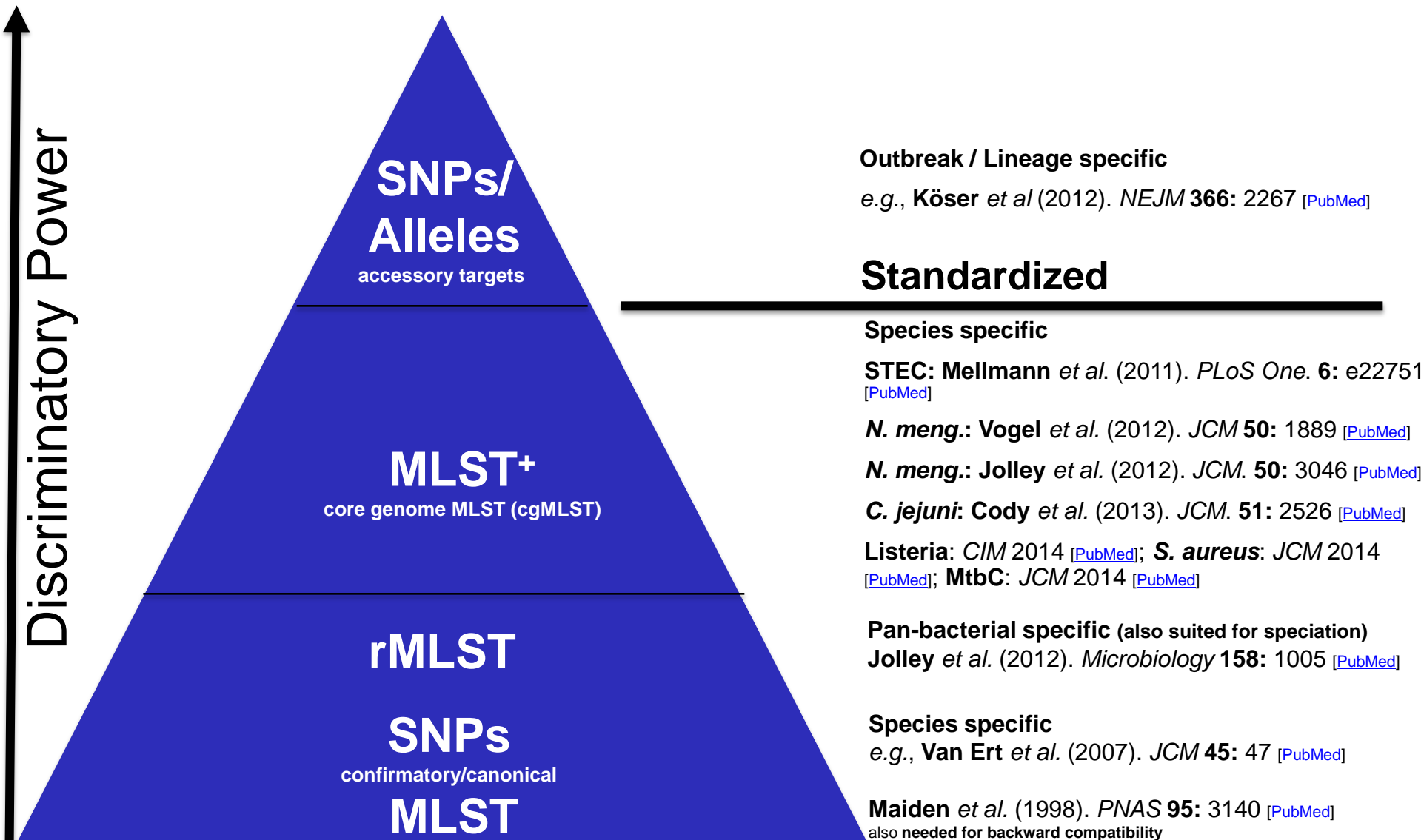
## MLST+/cgMLST

- Hundreds/thousands of 'core genome' genes
- Scalable, portable and under-standable
- Public, additive expandable nomenclature

Higher discrimination power for **outbreak investigation**



# Standardized Hierarchical Microbial Typing



**Hierarchical microbial typing approach.** From bottom to top with increasing discriminatory power. MLST, multi locus sequence typing; rMLST, ribosomal MLST; SNP, single nucleotide polymorphism; cgMLST, core genome MLST.



# One Disruptive Technology Fits it All - Genomic Surveillance



- BIGSdb**

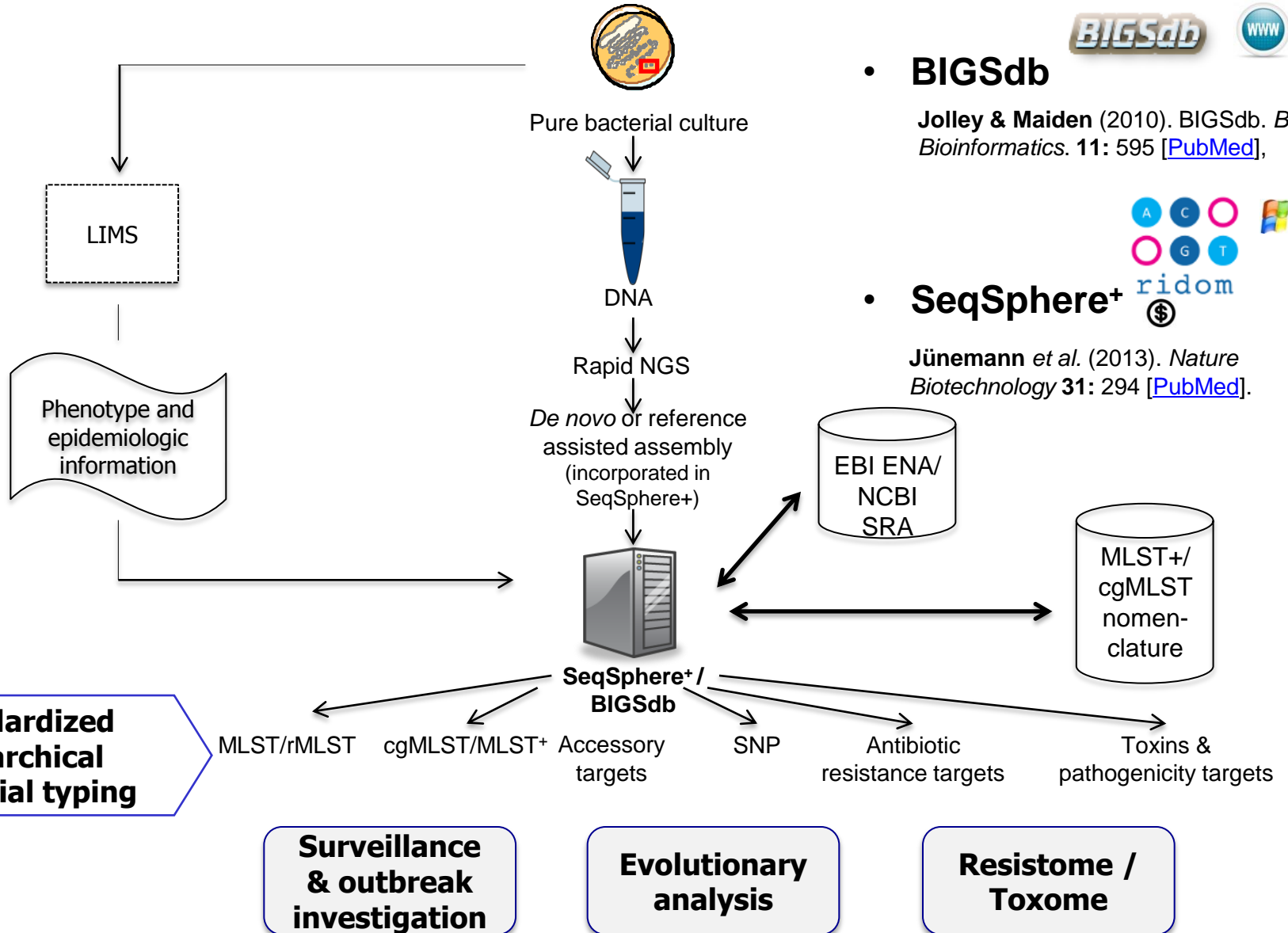
Jolley & Maiden (2010). BIGSdb. *BMC Bioinformatics*. 11: 595 [[PubMed](#)],



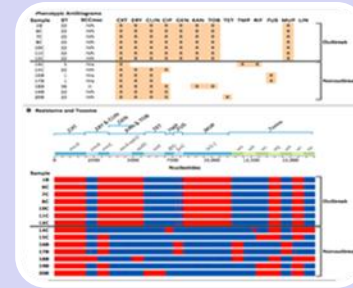
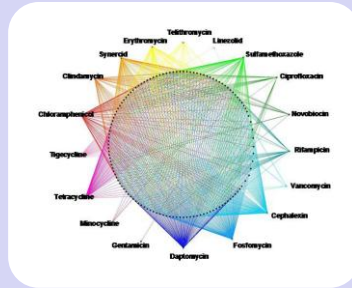
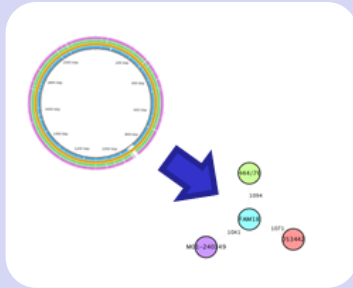
- SeqSphere+**



Jünemann *et al.* (2013). *Nature Biotechnology* 31: 294 [[PubMed](#)].



# Outlook



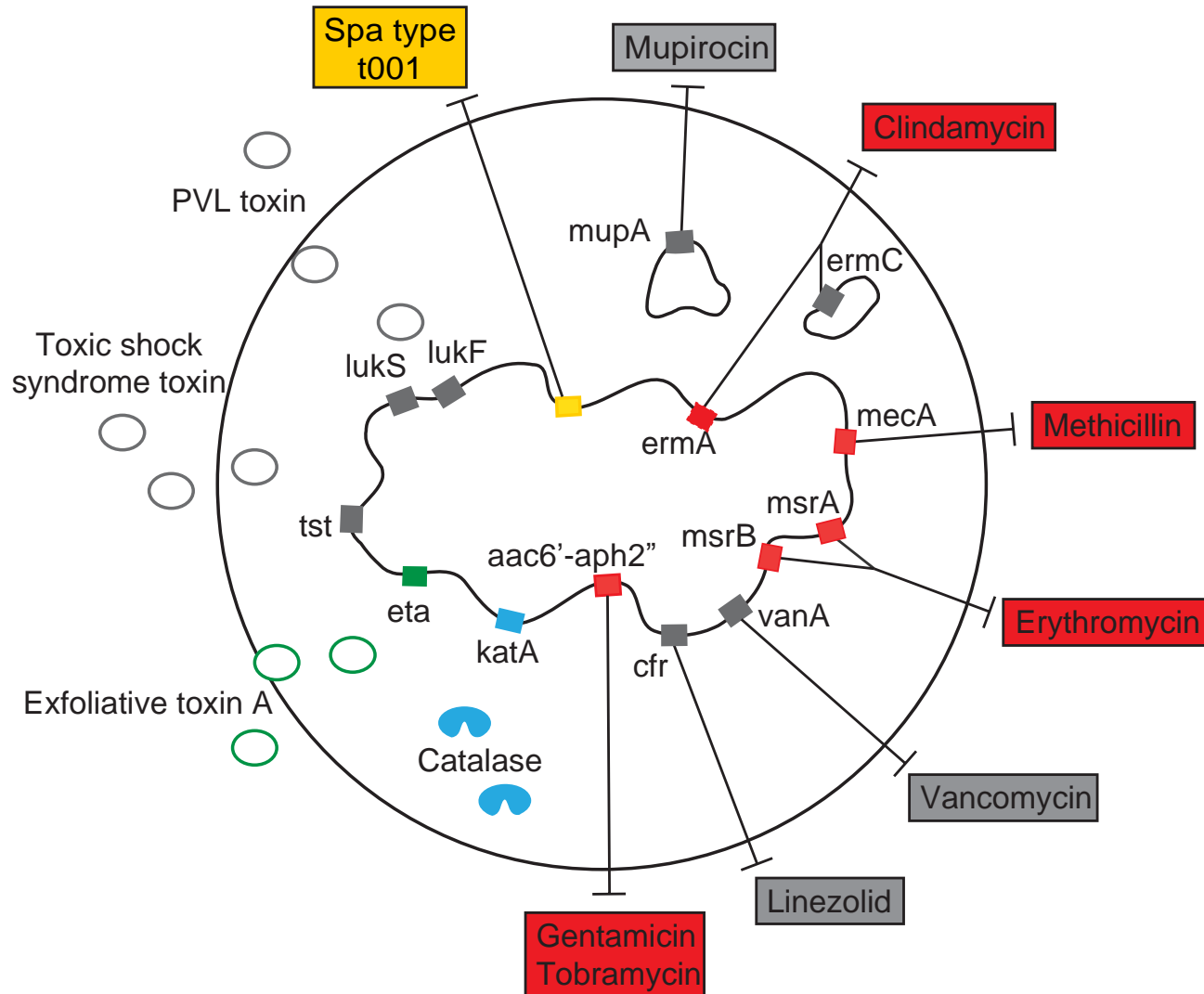
**Standardize  
WGS Typing  
with  
MLST+**

**From genotype  
to phenotype  
  
(resistome,  
pathogenome &  
toxome analysis)**

**Early warning  
system & GIS**

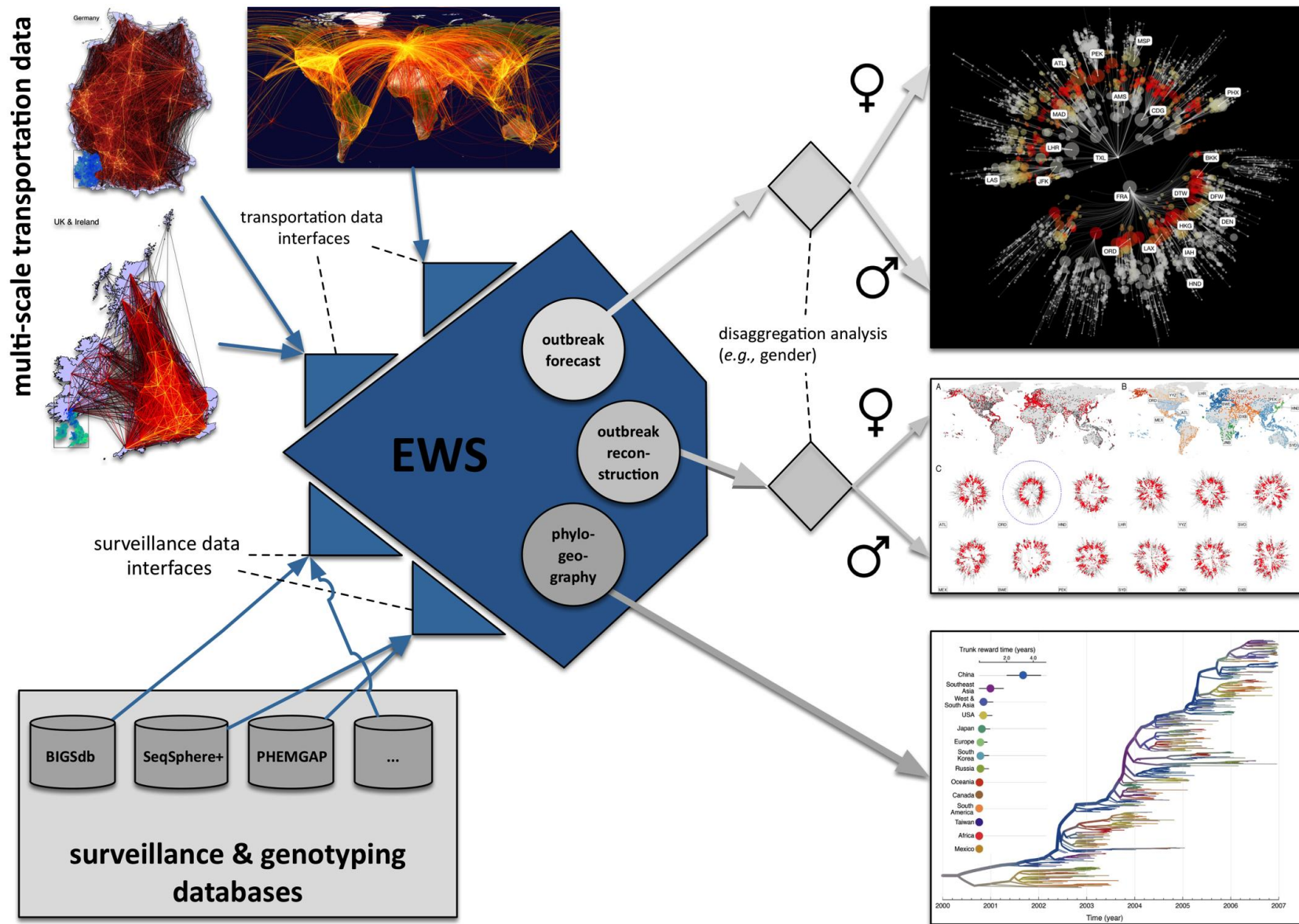
**Plain language  
report**

# From WGS Geno- to Phenotype - MRSA



**Species identification, *spa* type, antibiotic susceptibility profile and presence of toxins** can be rapidly determined by query of the WGS data. Colored squares represent genes potentially present on the chromosome and/or plasmids. The presence of genes in our cluster isolates are indicated by color: antibiotic resistance genes are shown in red, green for the toxin gene, blue for the catalase-encoding *katA*, yellow for the *spa* gene and gray indicates genes that were queried but not found.

# Predictive Models for Risk Assessment



EWS; early warning system.

Brockmann *et al.* (2013). *Science* 342: 1337 [PubMed].

# One Disrupting Technology Fits it All - Towards Standardized Bacterial Whole Genome Sequencing for Global Surveillance



**Dag Harmsen**  
University of Münster, Germany

[dharmsen@uni-muenster.de](mailto:dharmsen@uni-muenster.de)