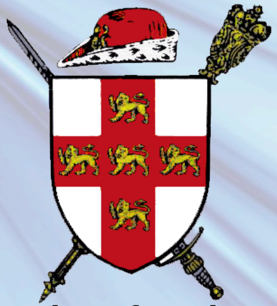
Harnessing Microbial Genomics for Epidemiological Surveillance



City of Münster & **City of York**

Dag Harmsen

University of Münster, Germany

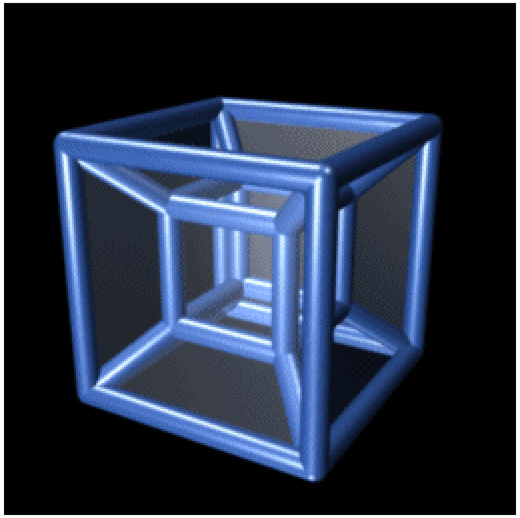
dharmsen@uni-muenster.de

Westfälische Wilhelms-Universität Münster

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

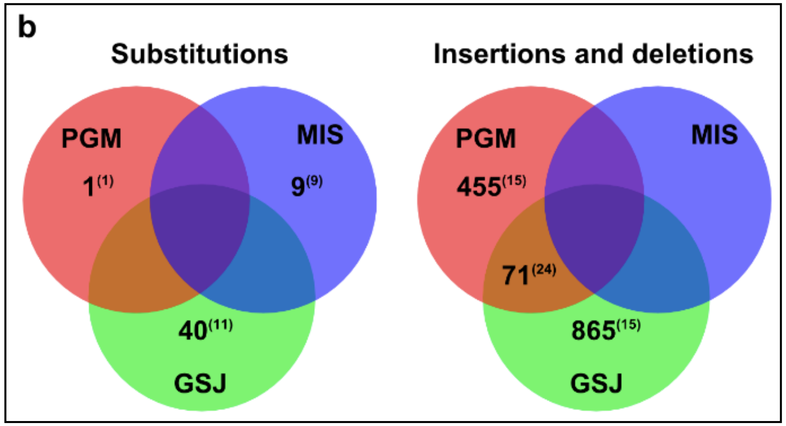


Place, Time, ‘Person’ … Type!

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

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

Current NGS Bottlenecks

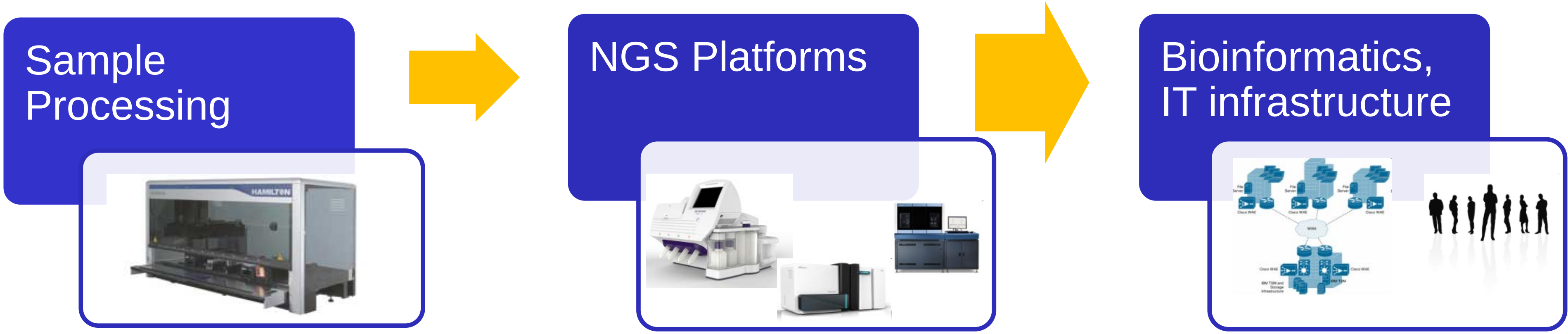
Library Prep

Template Amplification

AB **Library Builder™**

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

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 ~ *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 (≧ O(n^2), limit ≈ 30-50 genomes)**
- **- Not additive expandable, no nomenclature possible**

k-mer
without alignment

ANI
with alignment
(**A**verage **N**ucleotide **I**dentity)

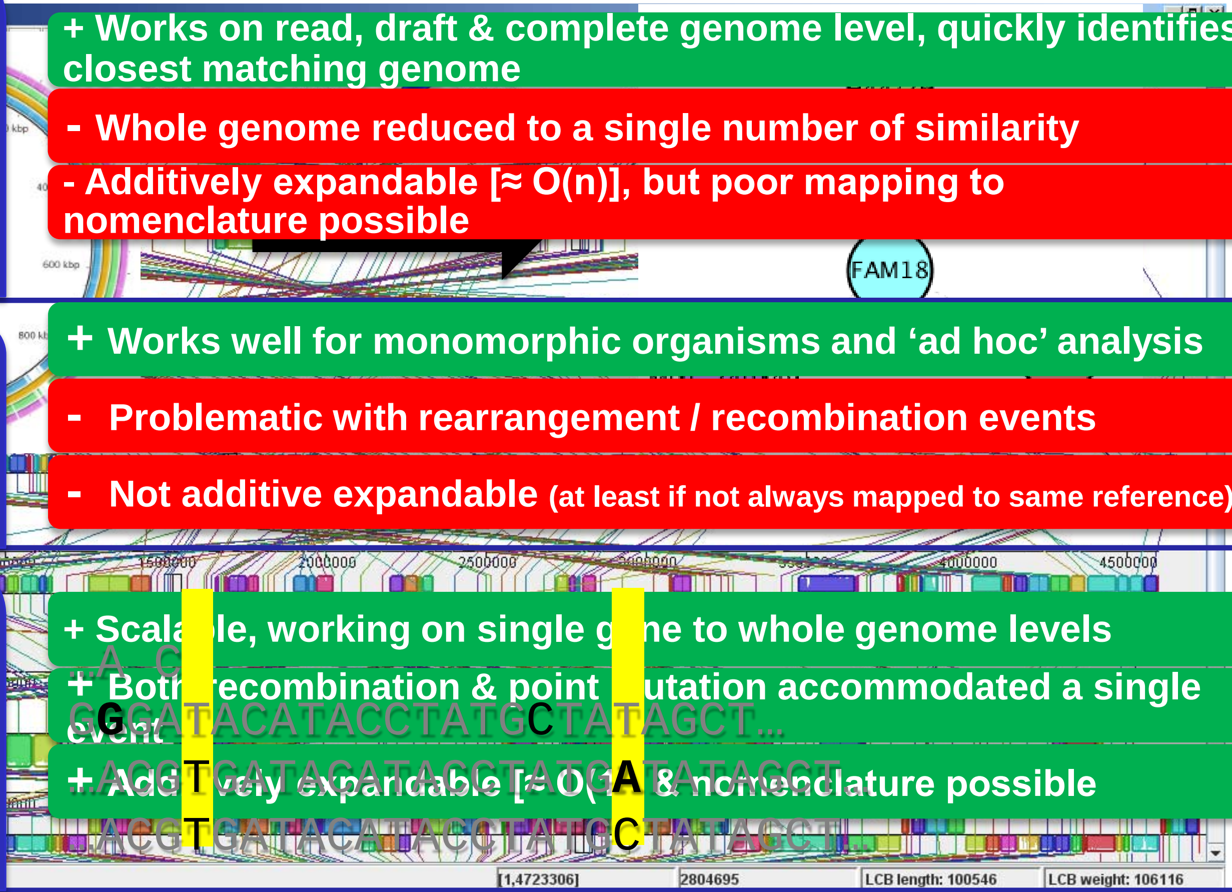
- **+ Works on read, draft & complete genome level, quickly identifies closest matching genome**
- **- Whole genome reduced to a single number of similarity**
- **- Additively expandable [≈ O(n)], but poor mapping to nomenclature possible**

Genome-wide mapping & SNP calling

- **+ 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 gene by gene allele typing

(cgMLST or MLST+)

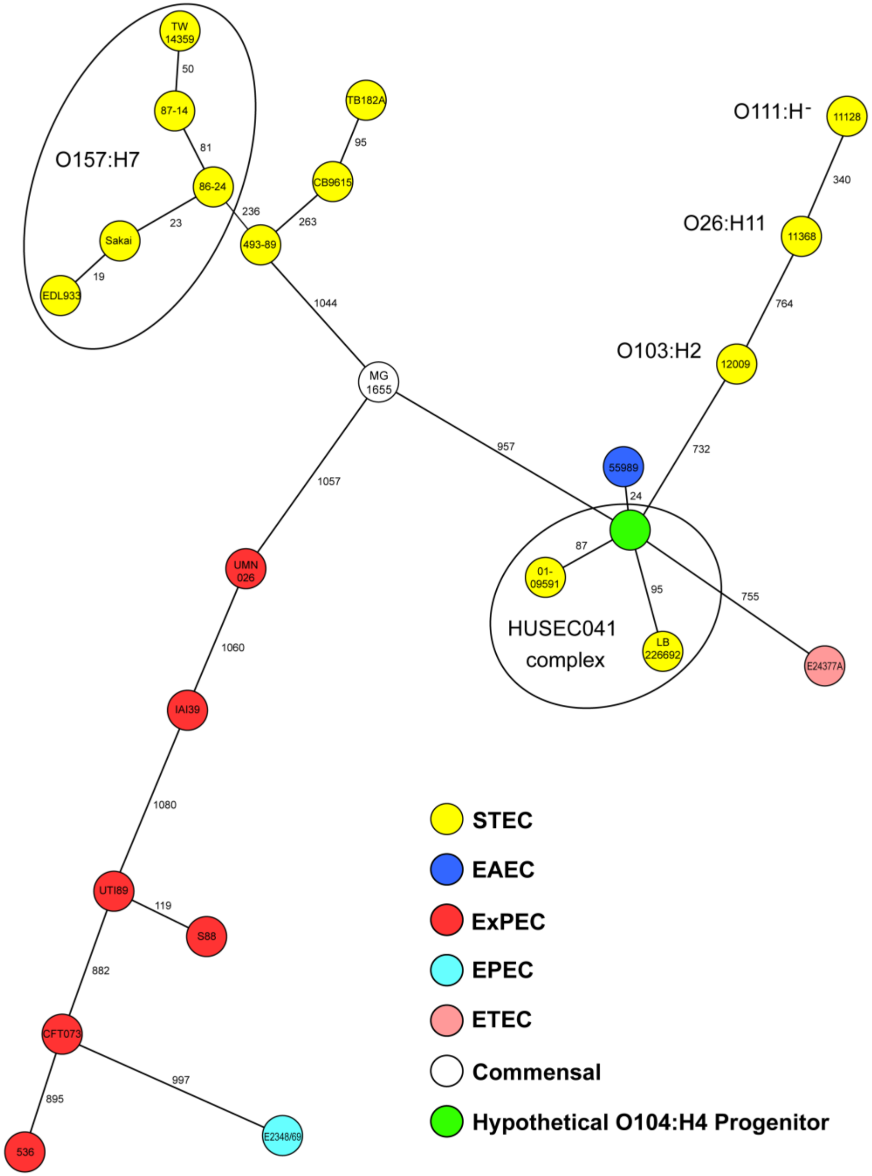


- **+ Scala le, working on single g ne to whole genome levels**
- **+ Both recombination & point utation accommodated a single event**
- **+ Add vely expandable [≈ O(& nomenclature possible**

SNP, single nucleotide polymorphism; cgMLST, core genome multi locus sequence typing; n, number of isolates in database.

Rapid ‚*Ad hoc*' NGS - *E. coli* O104:H4 Outbreak

(Germany May/June, 2011)



Phylogenetic Analysis of EHEC 0104:H4

Method

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

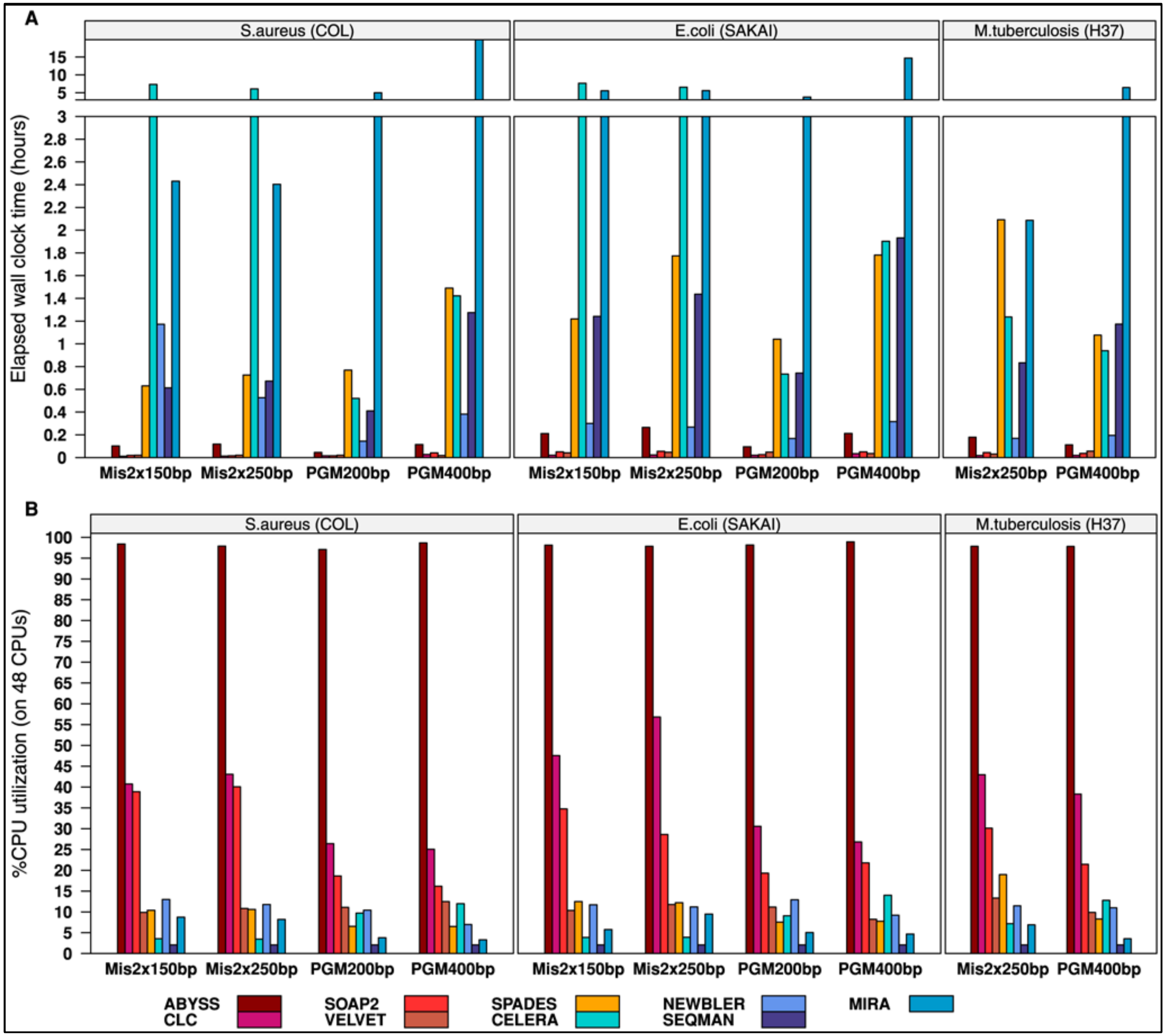
Results

- 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

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

Mellmann *et al*. (2011). *PLoS One*. **6:** e22751 [PubMed]

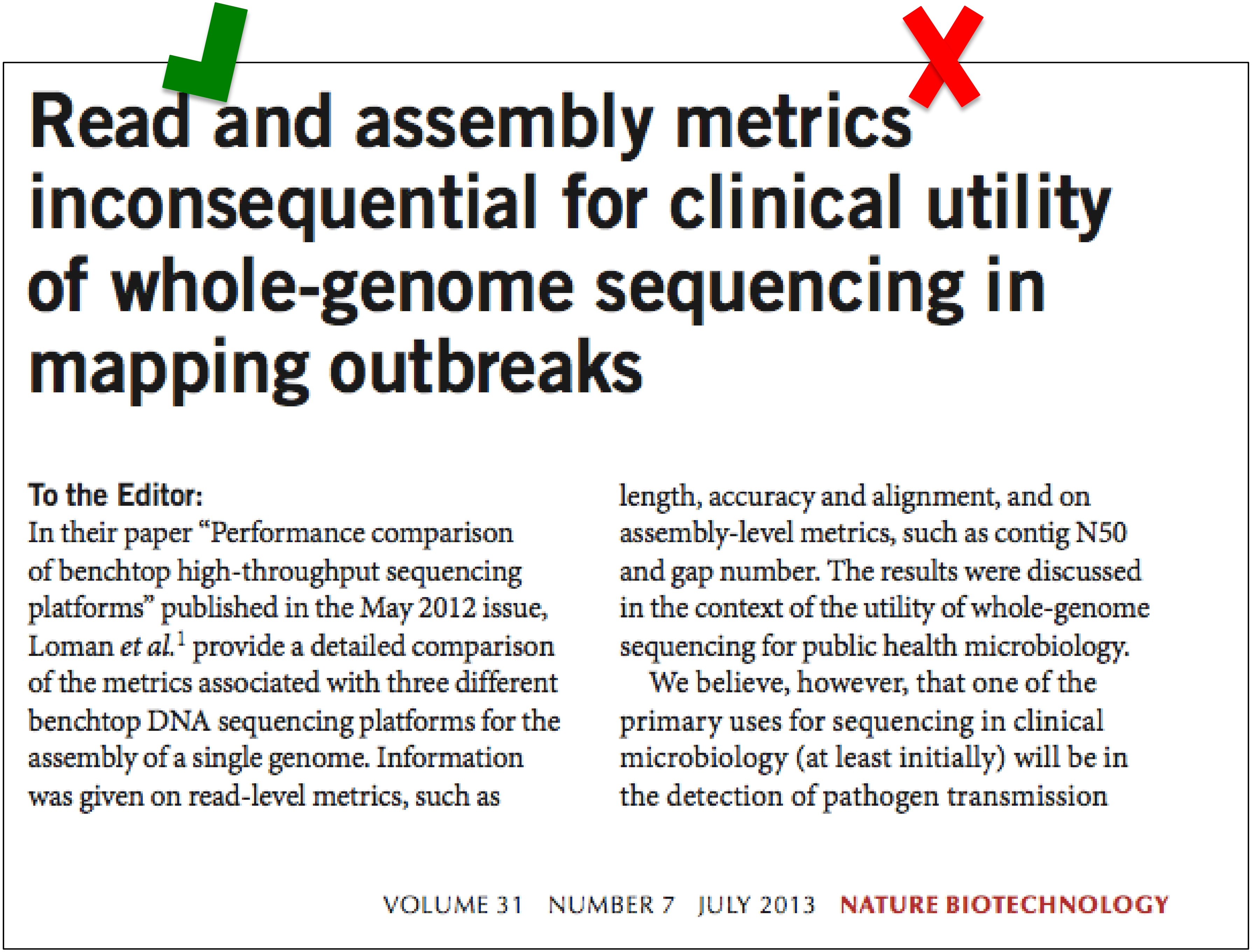
GABenchToB: A Genome Assembly Benchmark Tuned on Bacteria and Benchtop Sequencers



Based on the elapsed wall clock time (**A**, in hours) and the total CPU utilization (**B**, in percent and relative to the 48 available CPU cores of the executing compute host). With regard to the CPU utilization, all assemblies have been instructed via proper parameterization to make maximal use of the 48 available CPU cores. The only exceptions to this were SEQMAN, which does not support parallelization, and CELERA, which due to configuration constraints has altering concurrency and multi-threading parameters for different internal processes. For DBG assemblers only run time and CPU utilization of the single assemblies with the best performing k-mer parameter are shown and not the summation of the full k-mer optimization procedure (for **SPADES** and **CLC** this is equivalent).

Detailed analysis of the effects of different **coverage** and of different kmer szes [for de Bruijn graph assemblers only]!

Jünemann *et al.* (2014). *PLoS One* **9:** e107014 [PubMed].



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.*[1] 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.

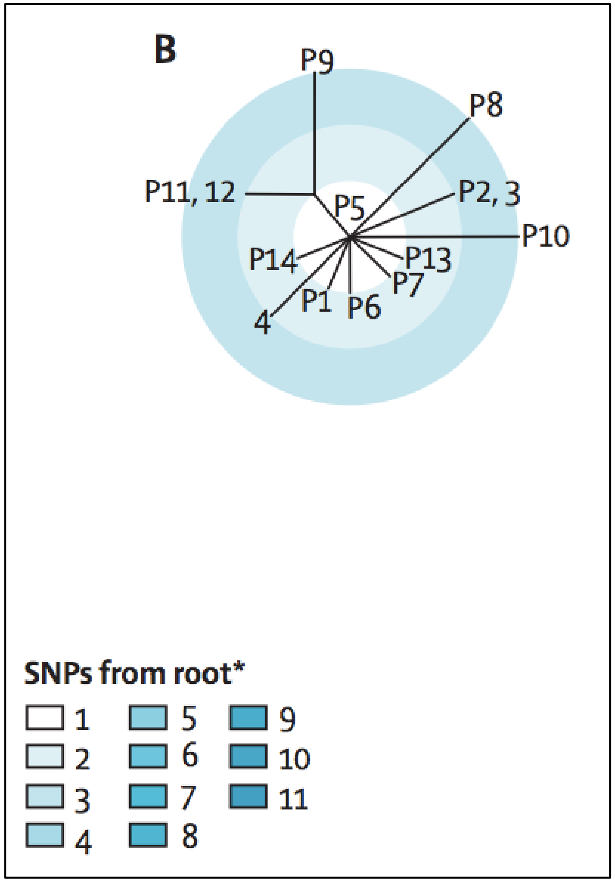
We believe, however, that one of the primary uses for sequencing in clinical microbiology (at least initially) will be in the detection of pathogen transmission

VOLUME 31 NUMBER 7 JULY 2013 NATURE BIOTECHNOLOGY

Harris *et al.* (2013). *Nature Biotechnology* **31:** 592 [PubMed].

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 [*gave nearly identical SNP lists*]
- 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?*)

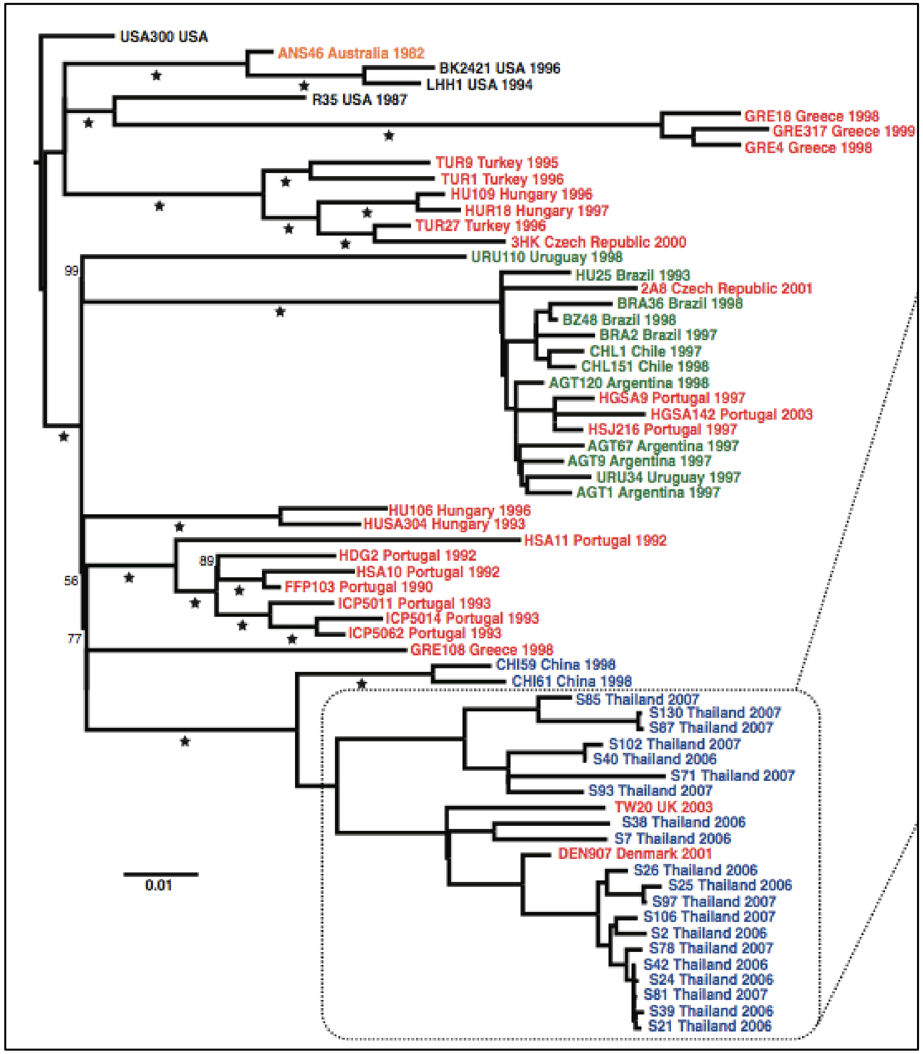


Harris *et al.* (2013). *Nature Biotechnology* **31:** 592 [PubMed].

Mapping & SNP Calling II

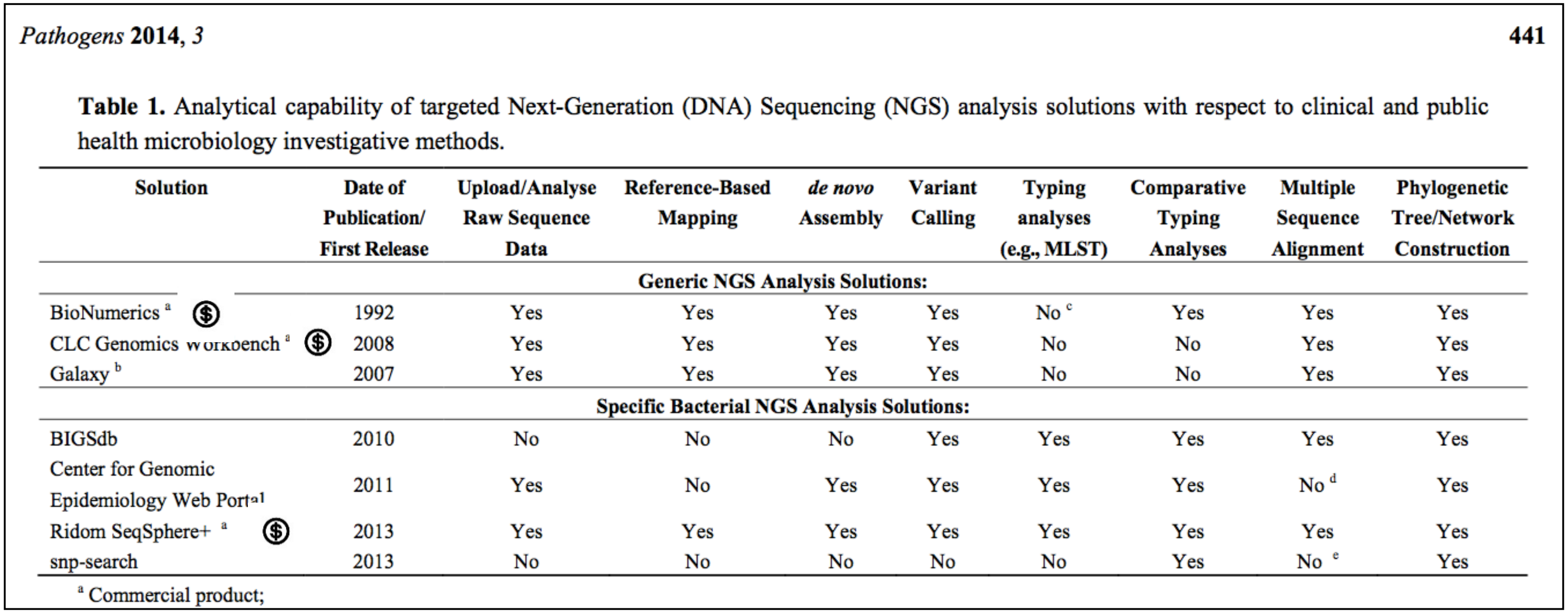
- high-resolution view of the epidemiology and microevolution of a dominant lineage (**ST 239**) of methicillin-resistant Staphylococcus aureus (MRSA)
- reads were **mapped** for each isolate **against TW20 reference** (**ST 239**) and discriminatory single nucleotide polymorphisms (**SNPs**) were identified 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 instantly comparable due to different reference genomes used!**



Harris *et al.* (2010). *Science* **327:** 469 [PubMed].

Tools for Surveillance & Phylogeny



Pathogens **2014**, *3* **441**

Table 1. Analytical capability of targeted Next-Generation (DNA) Sequencing (NGS) analysis solutions with respect to clinical and public health microbiology investigative methods.

Solution	Date of Publication/ First Release	Upload/Analyse Raw Sequence Data	Reference-Based Mapping	*de novo* Assembly	Variant Calling	Typing analyses (e.g., MLST)	Comparative Typing Analyses	Multiple Sequence Alignment	Phylogenetic Tree/Network Construction
Generic NGS Analysis Solutions:									
BioNumerics [a] ($)	1992	Yes	Yes	Yes	Yes	No [c]	Yes	Yes	Yes
CLC Genomics Workbench [a] ($)	2008	Yes	Yes	Yes	Yes	No	No	Yes	Yes
Galaxy [b]	2007	Yes	Yes	Yes	Yes	No	No	Yes	Yes
Specific Bacterial NGS Analysis Solutions:									
BIGSdb	2010	No	No	No	Yes	Yes	Yes	Yes	Yes
Center for Genomic Epidemiology Web Portal [a]	2011	Yes	No	Yes	Yes	Yes	Yes	No [d]	Yes
Ridom SeqSphere+ [a] ($)	2013	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
snp-search	2013	No	No	No	No	No	Yes	No [e]	Yes

[a] Commercial product;

Nomenclature
No
No
No
Yes
No
Yes
No

Wyres *et al.* (2014). WGS analysis and interpretation in clinical and public health microbiology laboratories: what are the Requirements and how do existing tools compare? *Pathogens* **3:** 437 [doi:10.3390/pathogens3020437].

3rd vs. 2nd Generation Sequencing

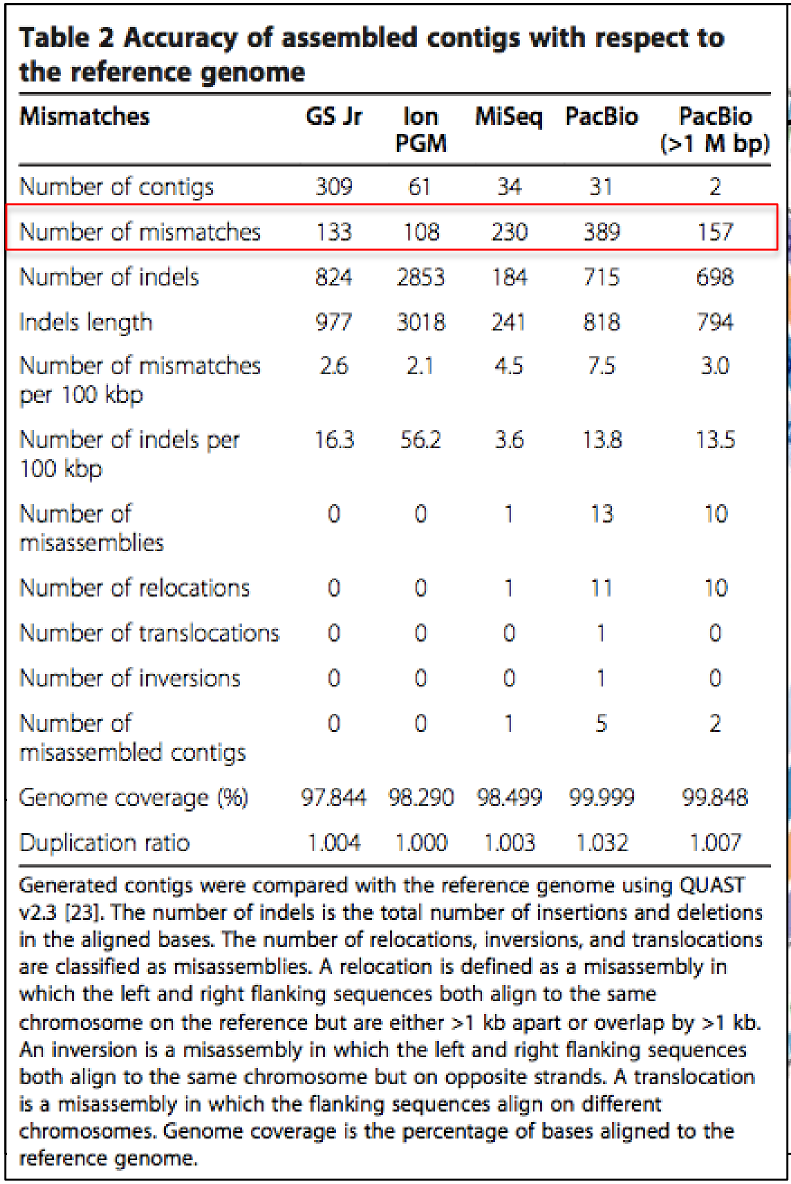
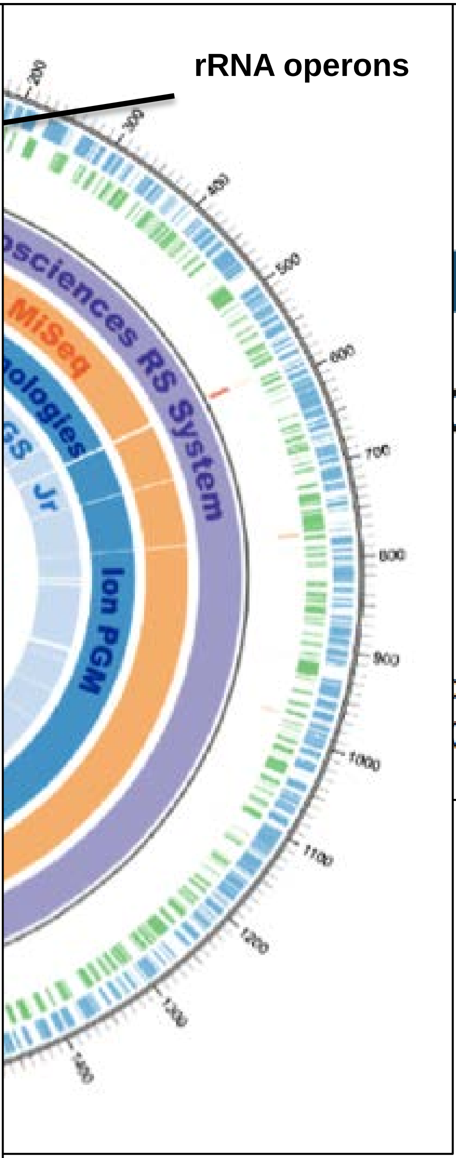
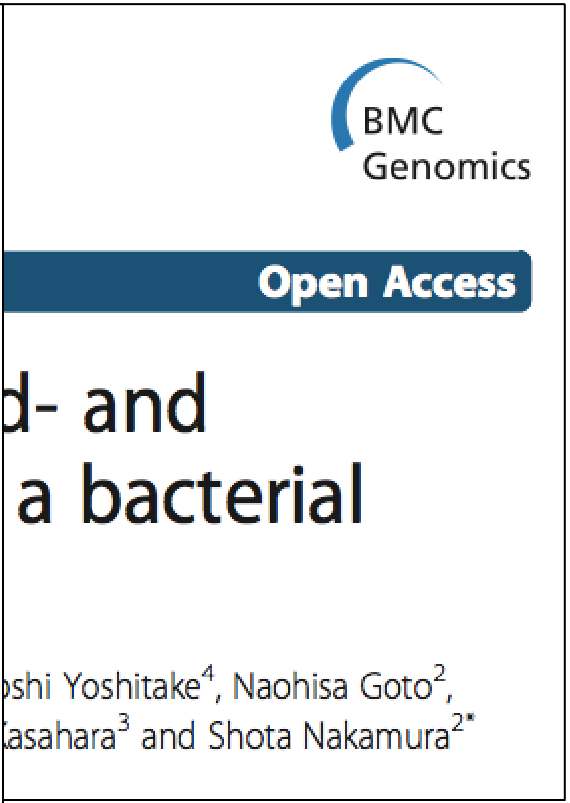


Table 2 Accuracy of assembled contigs with respect to the reference genome

Mismatches	GS Jr	Ion PGM	MiSeq	PacBio	PacBio (>1 M bp)
Number of contigs	309	61	34	31	2
Number of mismatches	133	108	230	389	157
Number of indels	824	2853	184	715	698
Indels length	977	3018	241	818	794
Number of mismatches per 100 kbp	2.6	2.1	4.5	7.5	3.0
Number of indels per 100 kbp	16.3	56.2	3.6	13.8	13.5
Number of misassemblies	0	0	1	13	10
Number of relocations	0	0	1	11	10
Number of translocations	0	0	0	1	0
Number of inversions	0	0	0	1	0
Number of misassembled contigs	0	0	1	5	2
Genome coverage (%)	97.844	98.290	98.499	99.999	99.848
Duplication ratio	1.004	1.000	1.003	1.032	1.007

Generated contigs were compared with the reference genome using QUAST v2.3 [23]. The number of indels is the total number of insertions and deletions in the aligned bases. The number of relocations, inversions, and translocations are classified as misassemblies. A relocation is defined as a misassembly in which the left and right flanking sequences both align to the same chromosome on the reference but are either >1 kb apart or overlap by >1 kb. An inversion is a misassembly in which the left and right flanking sequences both align to the same chromosome but on opposite strands. A translocation is a misassembly in which the flanking sequences align on different chromosomes. Genome coverage is the percentage of bases aligned to the reference genome.





BMC Genomics

Open Access

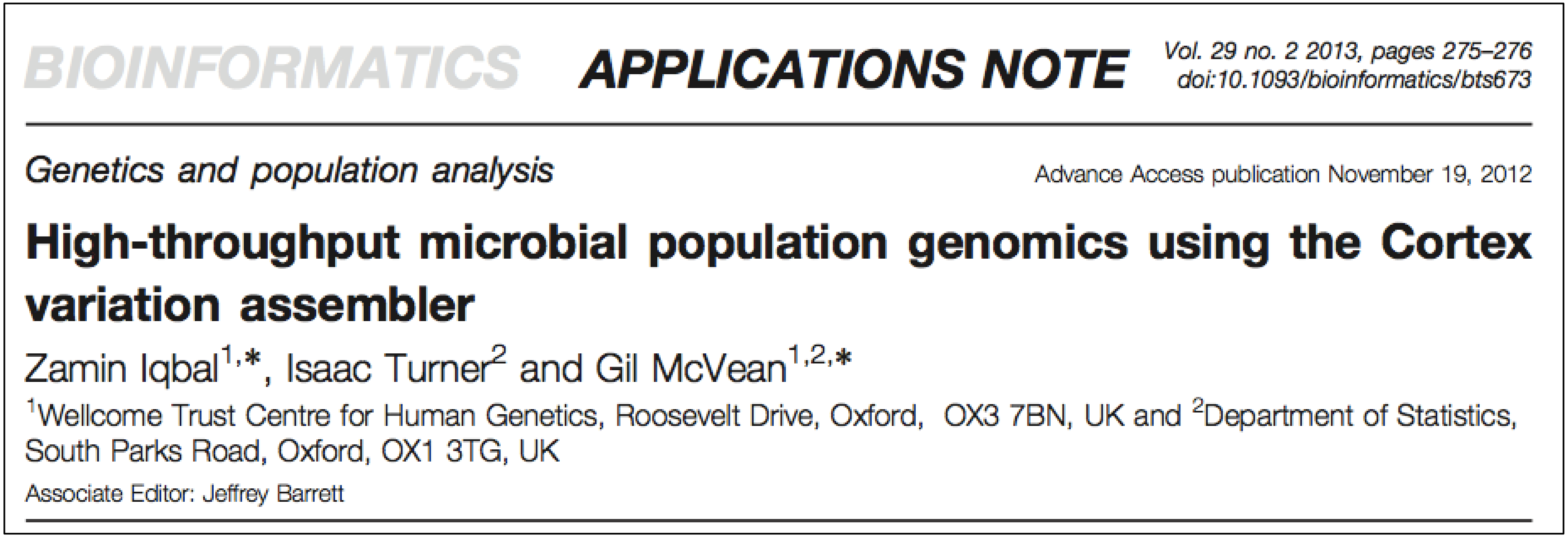
d- and
a bacterial

shi Yoshitake[4], Naohisa Goto[2],
asahara[3] and Shota Nakamura[2*]

3rd generation sequencer improvements
- better de novo assemblies
- more complete genomes

Miyamoto *et al.* (2014). *BMC Genomics* **15:** 699 [PubMed].

De *novo* Assembly and SNP Calling



BIOINFORMATICS APPLICATIONS NOTE

Vol. 29 no. 2 2013, pages 275–276
doi:10.1093/bioinformatics/bts673

Genetics and population analysis

Advance Access publication November 19, 2012

High-throughput microbial population genomics using the Cortex variation assembler

Zamin Iqbal[1,*], Isaac Turner[2] and Gil McVean[1,2,*]

[1]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK and [2]Department of Statistics, South Parks Road, Oxford, OX1 3TG, UK

Associate Editor: Jeffrey Barrett

Abstract

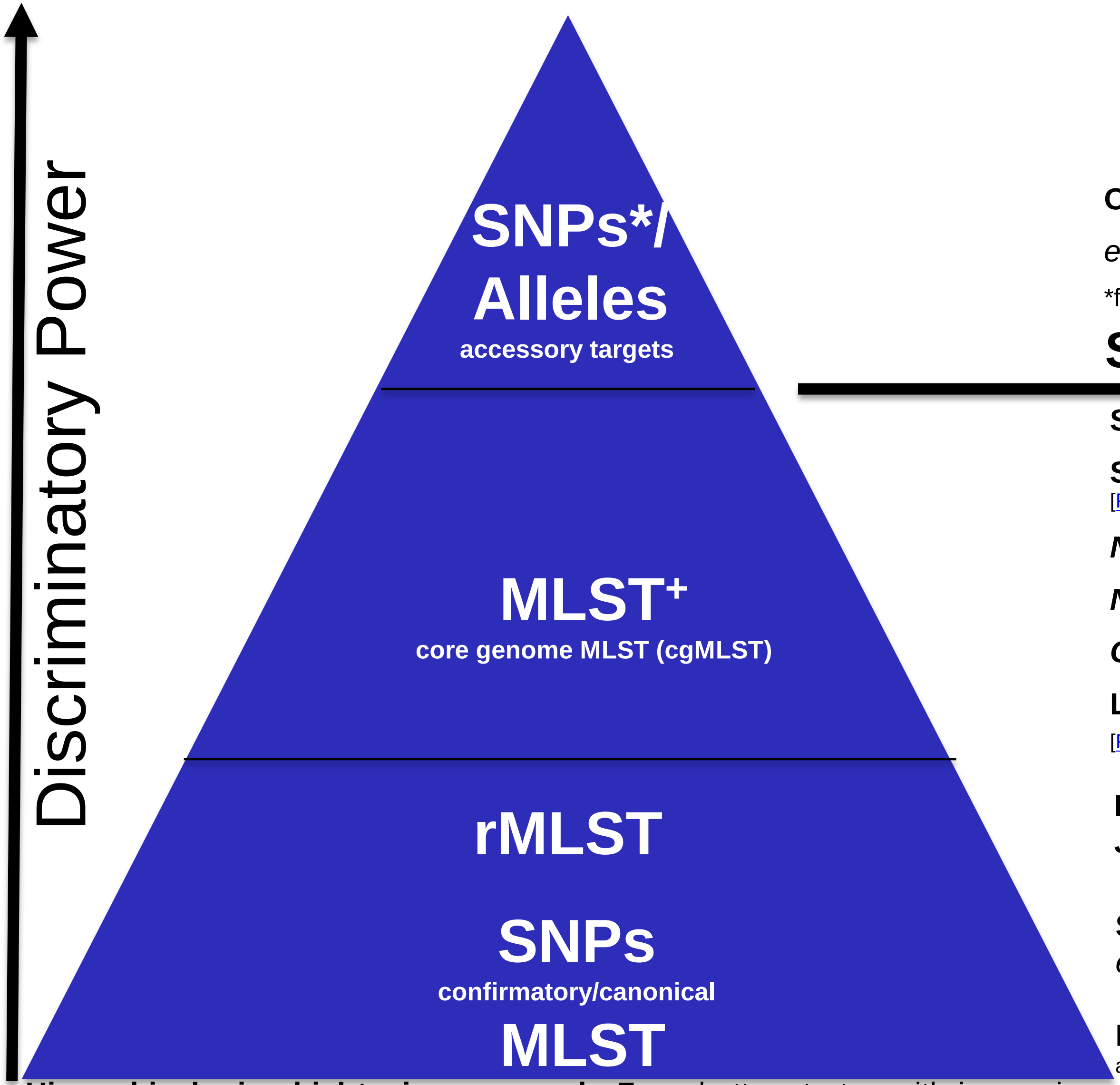
SUMMARY: We have developed a software package, Cortex, designed for the analysis of genetic variation by *de novo* assembly of multiple samples. This allows direct comparison of samples without using a reference genome as intermediate and incorporates discovery and genotyping of single-nucleotide polymorphisms, indels and larger events in a single framework. We introduce pipelines which simplify the analysis of microbial samples and increase discovery power; these also enable the construction of a graph of known sequence and variation in a species, against which new samples can be compared rapidly [*Cortex memory-use scales linearly with number of kmers and samples*]. We demonstrate the ease-of-use and power by reproducing the results of studies using both long and short reads.

AVAILABILITY: http://cortexassembler.sourceforge.net (GPLv3 license).

CONTACT: zam@well.ox.ac.uk, mcvean@well.ox.ac.uk

Iqbal *et al.* (2013). *Bioinformatics* **29:** 275 [PubMed].

Standardized Hierarchical Microbial Typing



Outbreak / Lineage specific

e.g., **Köser** *et al* (2012). *NEJM* **366:** 2267 [PubMed]

*from ***de novo* assembled** and/or mapped genomes

Standardized

Species specific

STEC: Mellmann *et al*. (2011). *PLoS One*. **6:** e22751 [PubMed]

***N. meng.*: Vogel** *et al.* (2012). *JCM* **50:** 1889 [PubMed]

***N. meng.*: Jolley** *et al.* (2012). *JCM*. **50:** 3046 [PubMed]

***C. jejuni*: Cody** *et al.* (2013). *JCM*. **51:** 2526 [PubMed]

Listeria: *CIM* 2014 [PubMed]; ***S. aureus***: *JCM* 2014 [PubMed]; **MtbC**: *JCM* 2014 [PubMed]

Pan-bacterial specific (also suited for speciation)
Jolley *et al.* (2012). *Microbiology* **158:** 1005 [PubMed]

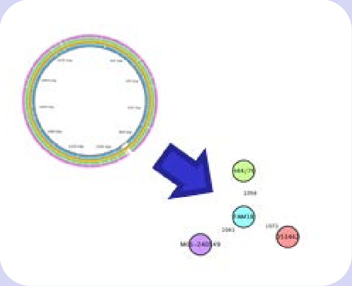
Species specific
e.g., **Van Ert** *et al.* (2007). *JCM* **45:** 47 [PubMed]

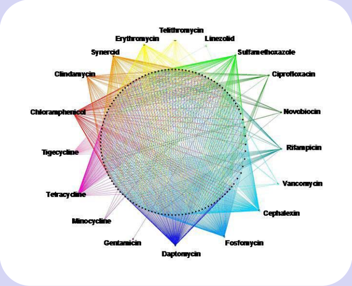
Maiden *et al.* (1998). *PNAS* **95:** 3140 [PubMed]
also **needed for backward compatibility**

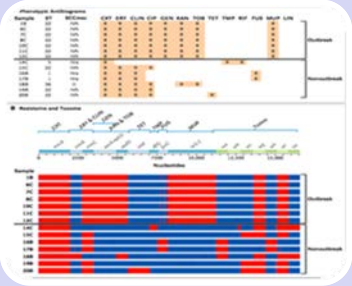
Hierarchical microbial typing approach. From bottom to top with increasing discriminatory power. MLST, multi locus sequence typing; rMLST, ribosomal MLST; SNP, single nucleotide polymorphism; cgMLST, core genome MLST.

For hierarchical microbial typing see also: **Maiden** *et al.* (2013). *Nature Rev. Microbiol.* **11:** 728 [PubMed].

Outlook







Standardization

WGS Typing

(cgMLST/MLST

+ [?])

From genotype to phenotype

(resistome, pathogenome & toxome analysis)

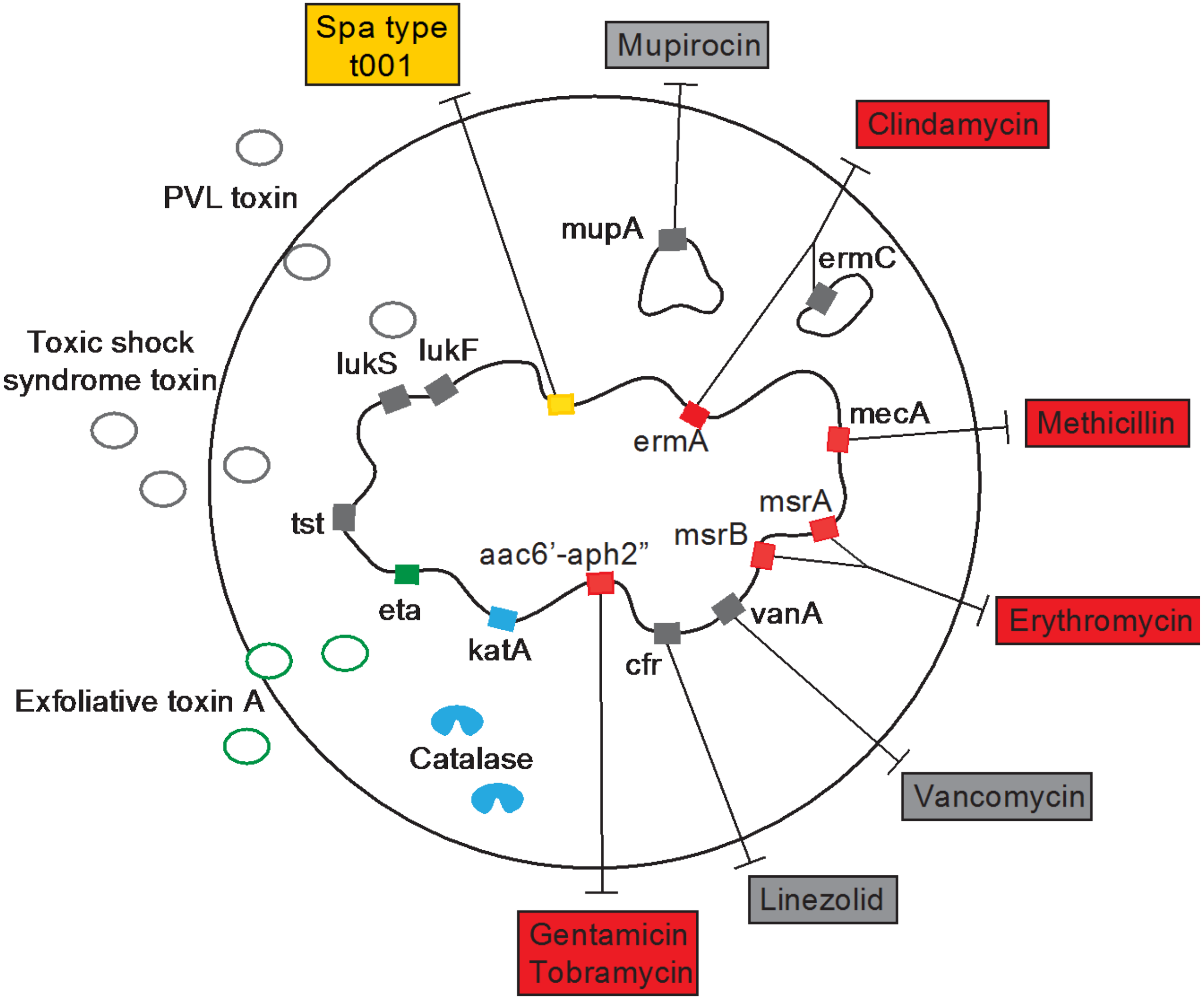
Early warning system & GIS

Plain language report

PATHO NGenTrace

http://patho-ngen-trace.eu/

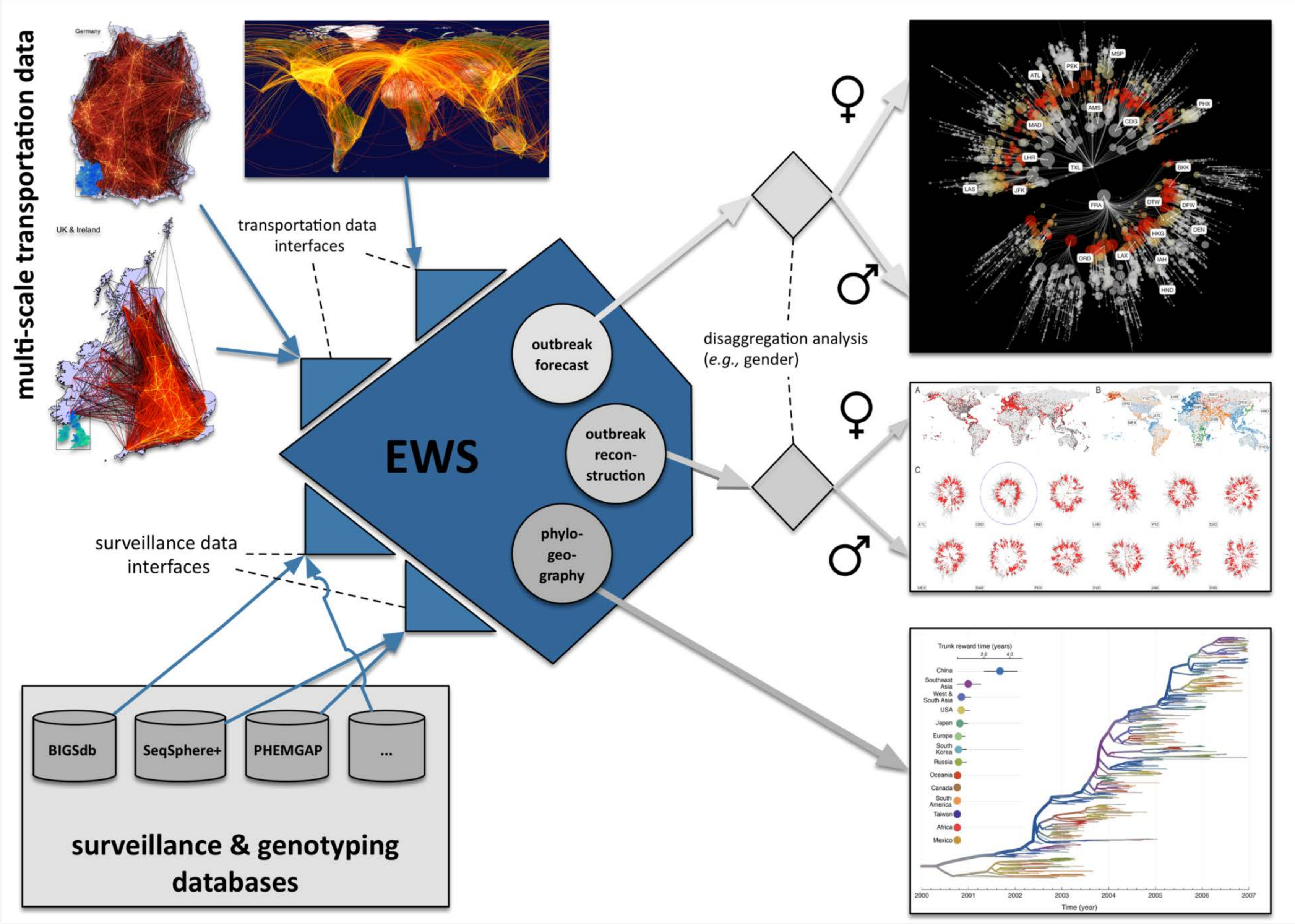
From WGS Geno- to Phenotype



Staphylococcus aureus **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.

Leopold *et al.* (2014). *JCM* **52:** 2365 [PubMed].

Predictive Models: Early Warning, Outbreak Spread, Outbreak Source location & Outbreak Reconstruction



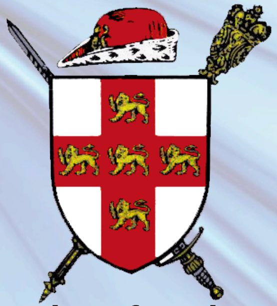
EWS; early warning system.

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

Harnessing Microbial Genomics for Epidemiological Surveillance

City of Münster

&



City of York

Dag Harmsen
University of Münster, Germany

dharmsen@uni-muenster.de

Westfälische Wilhelms-Universität Münster