V1 Processing of Biological Data

Leistungspunkte/Credit points: 5 (V2/Ü1)

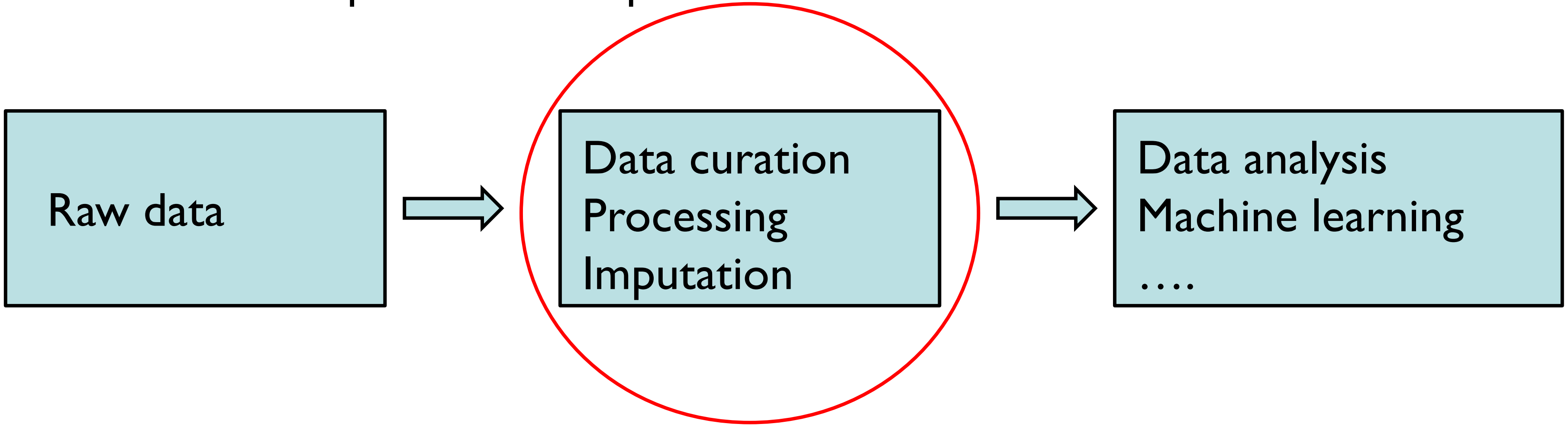
This course is taught in English language.

The material (from books and original literature) are provided online at the course website:

https://www-cbi.cs.uni-saarland.de/teaching/ws-1819/special-topic-lecture-bioinformatics-processing-of-biological-data/

Topics to be covered:

This course will discuss the handling of different sorts of biological data, often on the example of recent publications.



Raw data

Data curation
Processing
Imputation

Data analysis
Machine learning
….

Tutorial

We will handout 6 **bi-weekly assignments**.
Groups of up to two students can hand in a solved assignment.

Send your **solutions** by e-mail to the responsible tutors
until the time+date indicated on the assignment sheet.

The **bi-weekly tutorial** on Monday 12.45 am – 2.15 pm (same room, time is negotiable) will discuss the assignment solutions.

On demand, the tutors may also give some advice for solving the new assignments.

Schein conditions

The successful participation in the lecture course („Schein“)
will be certified upon fulfilling

- Schein condition 1 : ≥ 50% of the points for the assignments

- Schein condition 2 : pass **final written exam** at end of semester

The **grade** on your „Schein“ will equal that of your final exam.

Everybody who took the final exam (and passed it or did not pass it)
and those who have missed the final exam
can take the **re-exam** at the beginning of WS17/18.

Planned lecture - overview

V1: bacterial data (*S. aureus*): clustering / PCA (R. Akulenko)

V2: bacterial data/DNA methylation: prediction of missing values (BEclear, R. Akulenko)

V3: differential gene expression, detection of outliers (A. Barghash)

V4: MS proteomic data, imputation, normalization (D. Nguyen), protein arrays (M. Pedersen)

V5: peak detection, breathomics (AC Hauschild)

V6: shape detection, processing of kidney tumor MRI scans (Vera Bazhenova)

V7: genomic sequences, SNPs (M. Hamed, K. Reuter, Ha Vu Tran)

V8: functional GO annotations (M. Hamed, Ha Vu Tran)

V9: curve fitting, data smoothing (AKSmooth …)

V10: protein X-ray structures: titration states, hydration sites, multiple side chain and ligand conformations, superposition … protein-protein complexes: crystal contacts, interfaces, …

V11: analysis of MD simulation trajectories: correlation of snapshots, remove CMS motion

V12: multi-variate analysis

V13: integrative analysis of multidimensional data sets (D. Gaidar, M. Nazarieh)

Data preprocessing

Data preprocessing is one of the most critical steps in data mining.

Data preprocessing methods are divided into 4 categories:

- Data cleaning
- Data integration
- Data transformation
- Data reduction

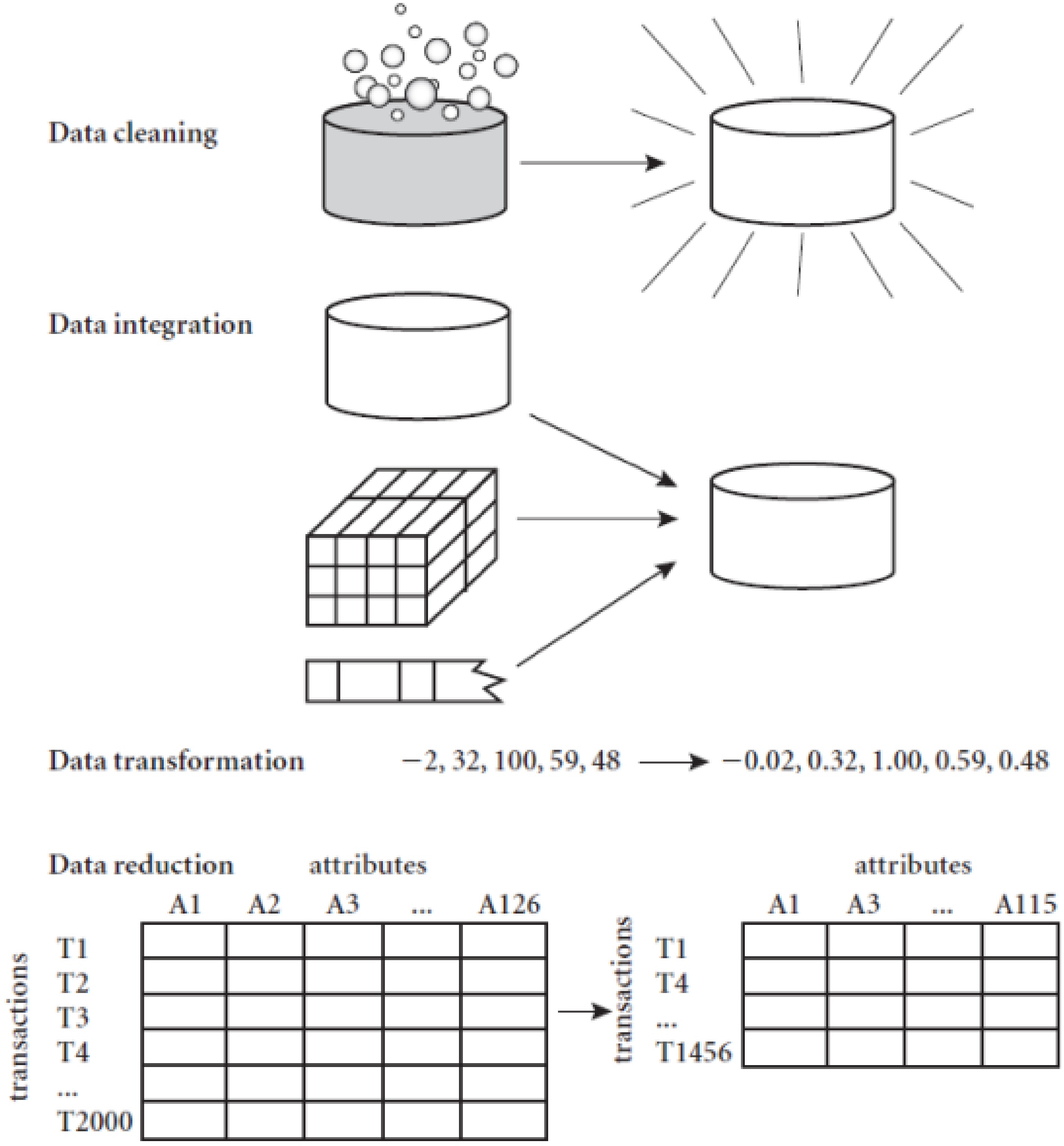


Figure 2.1 Forms of data preprocessing.

Data Mining: Know It All by Ian H. Witten et al. Publisher: Morgan Kaufmann (*2008)*

Data preprocessing

- Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.

- Data integration: using multiple databases, data cubes, or files.

- Data transformation: normalization and aggregation.

- Data reduction: reducing the volume but producing the same or similar analytical results.
    - Data discretization: part of data reduction, replacing numerical attributes with nominal ones.

Data Mining: Know It All by Ian H. Witten et al. Publisher: Morgan Kaufmann (*2008)*

Data summarization: quantile plot

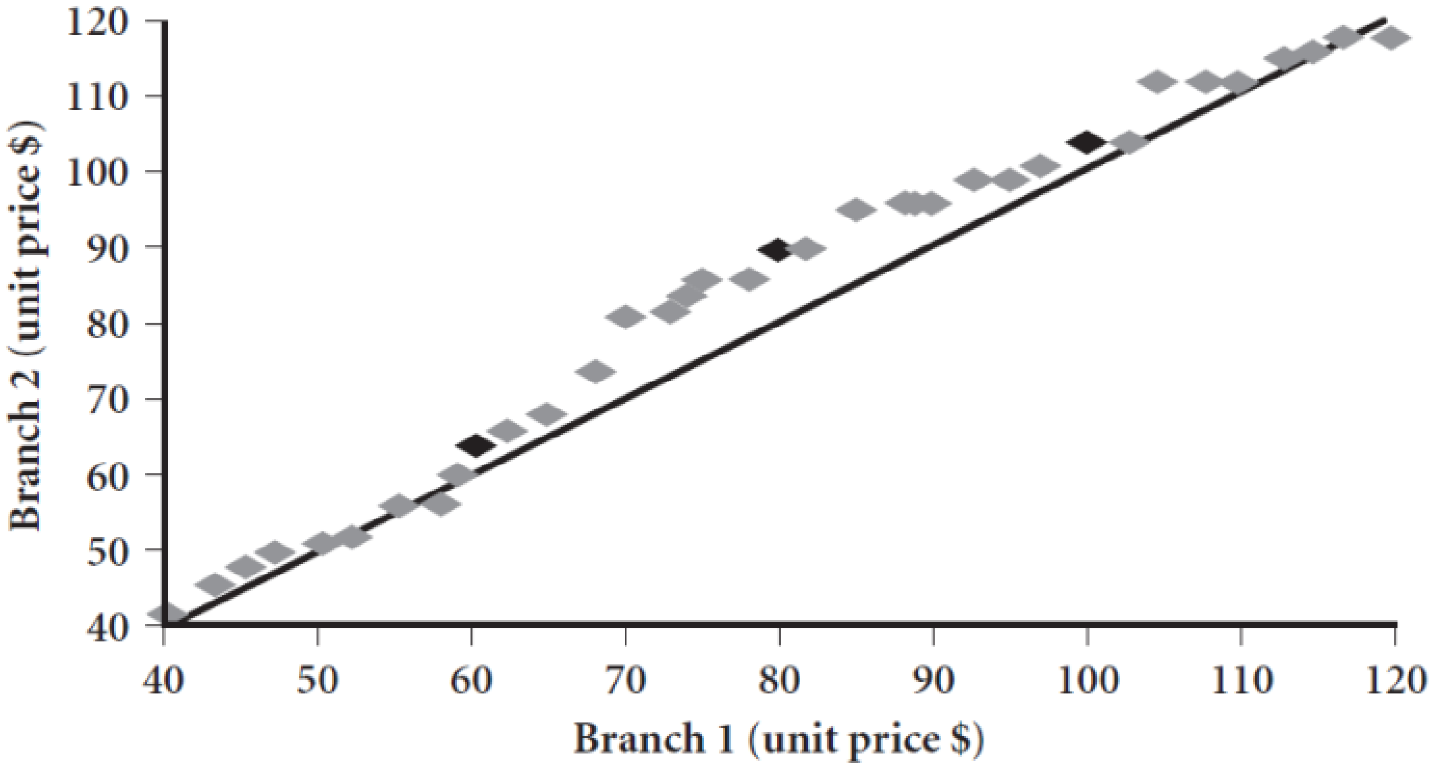


Figure 2.6 A quantile-quantile plot for unit price data from two different branches.

Interpretation: Branch 2 has – on average – higher unit prices.

Data Mining: Know It All by Ian H. Witten et al. Publisher: Morgan Kaufmann (*2008)*

ZBI ZENTRUM FÜR BIOINFORMATIK

UNIVERSITAS SARAVIENSIS UNIVERSITÄT DES SAARLANDES

Whole Genome Sequence Typing and Microarray Profiling of Methicillin-Resistant *Staphylococcus aureus* isolates

(1) Classification of MSSA / MRSA *S. aureus* strains in Saarland (PLoS ONE 2012)

(2) DFG Germany-Africa project (J. Clin. Microbiol. 2016; Sci. Reports 2017)

Co-workers

(1) Ruslan Akulenko, Ulla Ruffing, Mathias Herrmann, Lutz von Müller,

(2) StaphNet Consortium led by Mathias Herrmann, funded by **DFG**

Pilot study: classification of resistant *Staphylococcus aureus* strains

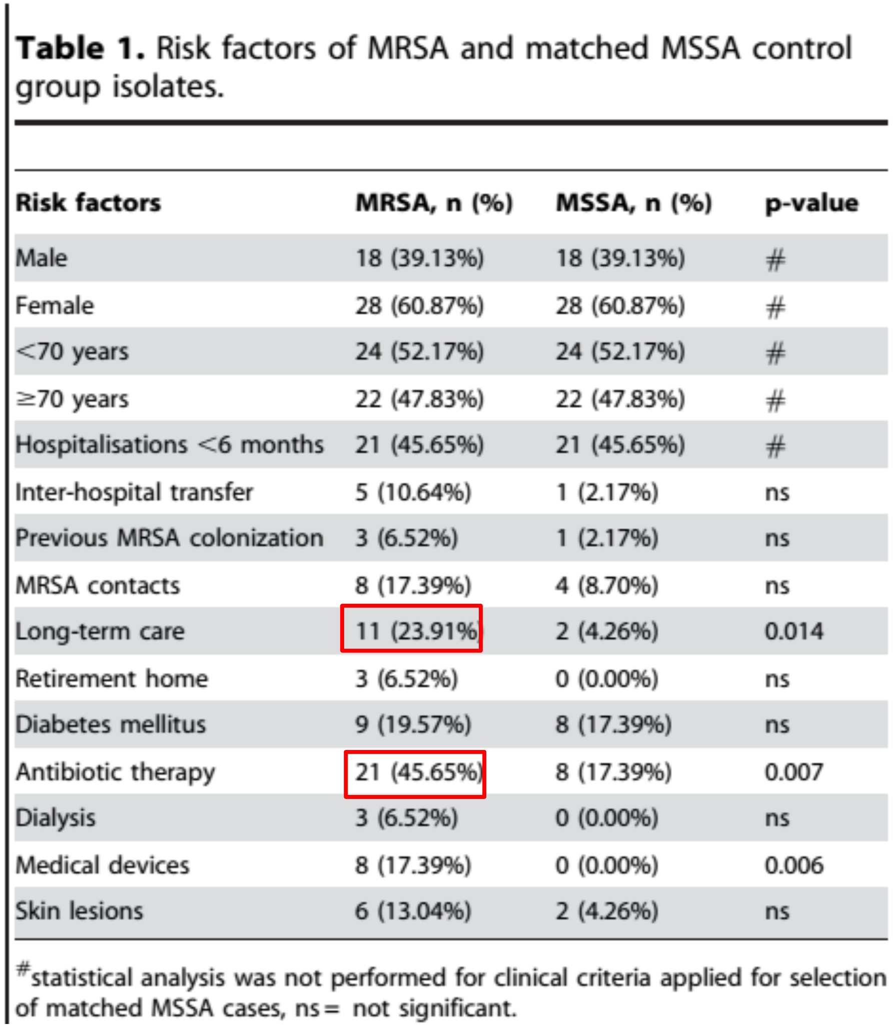
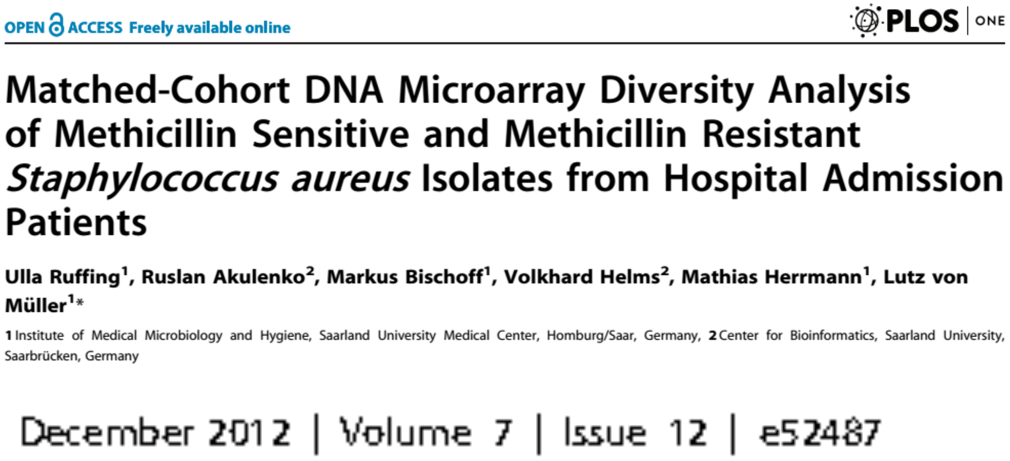


Table 1. Risk factors of MRSA and matched MSSA control group isolates.

Risk factors	MRSA, n (%)	MSSA, n (%)	p-value
Male	18 (39.13%)	18 (39.13%)	#
Female	28 (60.87%)	28 (60.87%)	#
<70 years	24 (52.17%)	24 (52.17%)	#
≥70 years	22 (47.83%)	22 (47.83%)	#
Hospitalisations <6 months	21 (45.65%)	21 (45.65%)	#
Inter-hospital transfer	5 (10.64%)	1 (2.17%)	ns
Previous MRSA colonization	3 (6.52%)	1 (2.17%)	ns
MRSA contacts	8 (17.39%)	4 (8.70%)	ns
Long-term care	11 (23.91%)	2 (4.26%)	0.014
Retirement home	3 (6.52%)	0 (0.00%)	ns
Diabetes mellitus	9 (19.57%)	8 (17.39%)	ns
Antibiotic therapy	21 (45.65%)	8 (17.39%)	0.007
Dialysis	3 (6.52%)	0 (0.00%)	ns
Medical devices	8 (17.39%)	0 (0.00%)	0.006
Skin lesions	6 (13.04%)	2 (4.26%)	ns

#statistical analysis was not performed for clinical criteria applied for selection of matched MSSA cases, ns = not significant.



OPEN ACCESS Freely available online

PLOS ONE

Matched-Cohort DNA Microarray Diversity Analysis of Methicillin Sensitive and Methicillin Resistant *Staphylococcus aureus* Isolates from Hospital Admission Patients

Ulla Ruffing[1], Ruslan Akulenko[2], Markus Bischoff[1], Volkhard Helms[2], Mathias Herrmann[1], Lutz von Müller[1]*

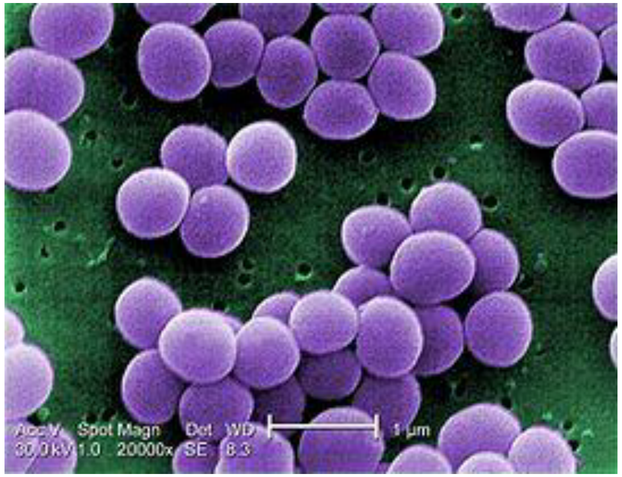
1 Institute of Medical Microbiology and Hygiene, Saarland University Medical Center, Homburg/Saar, Germany, 2 Center for Bioinformatics, Saarland University, Saarbrücken, Germany

December 2012 | Volume 7 | Issue 12 | e52487

Aim: classify MRSA / MSSA according to gene repertoire

Methycillin sensitive/resistant *Staphylococcus aureus* (MSSA/MRSA)

MSSA

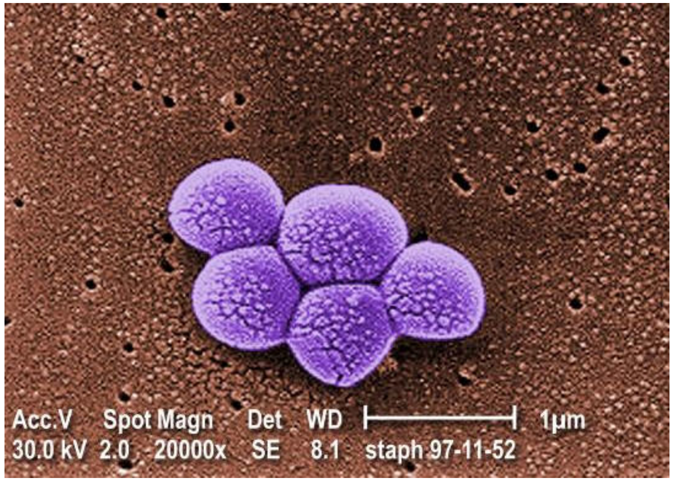


anaerobic Gram-positive coccal bacterium,

frequently part of the normal skin flora,

60% of population are carriers

MRSA


Acc.V Spot Magn Det WD 1µm
30.0 kV 2.0 20000x SE 8.1 staph 97-11-52

any strain of *S. aureus* with **resistance** to beta-lactam antibiotics:

- penicillins;
- cephalosporins;

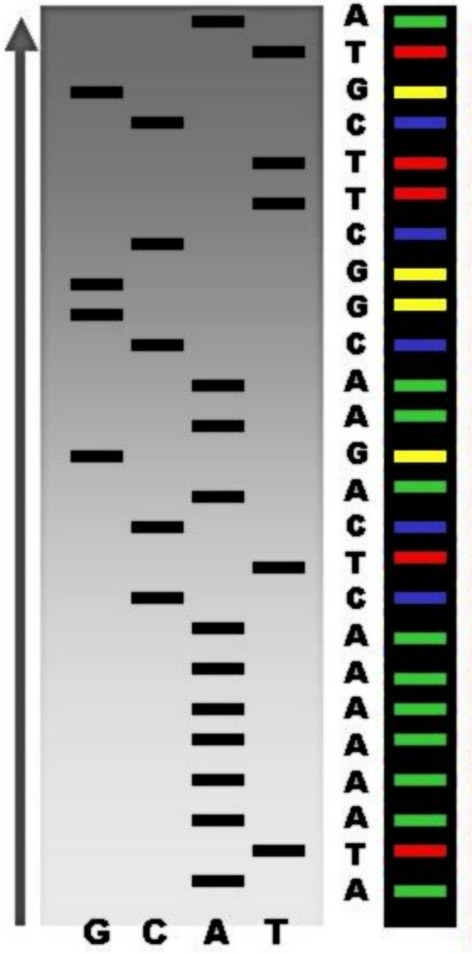
Need to classify MRSA strains to detect infections, prevent transmission

routine: Characterize MRSA by Spa-typing

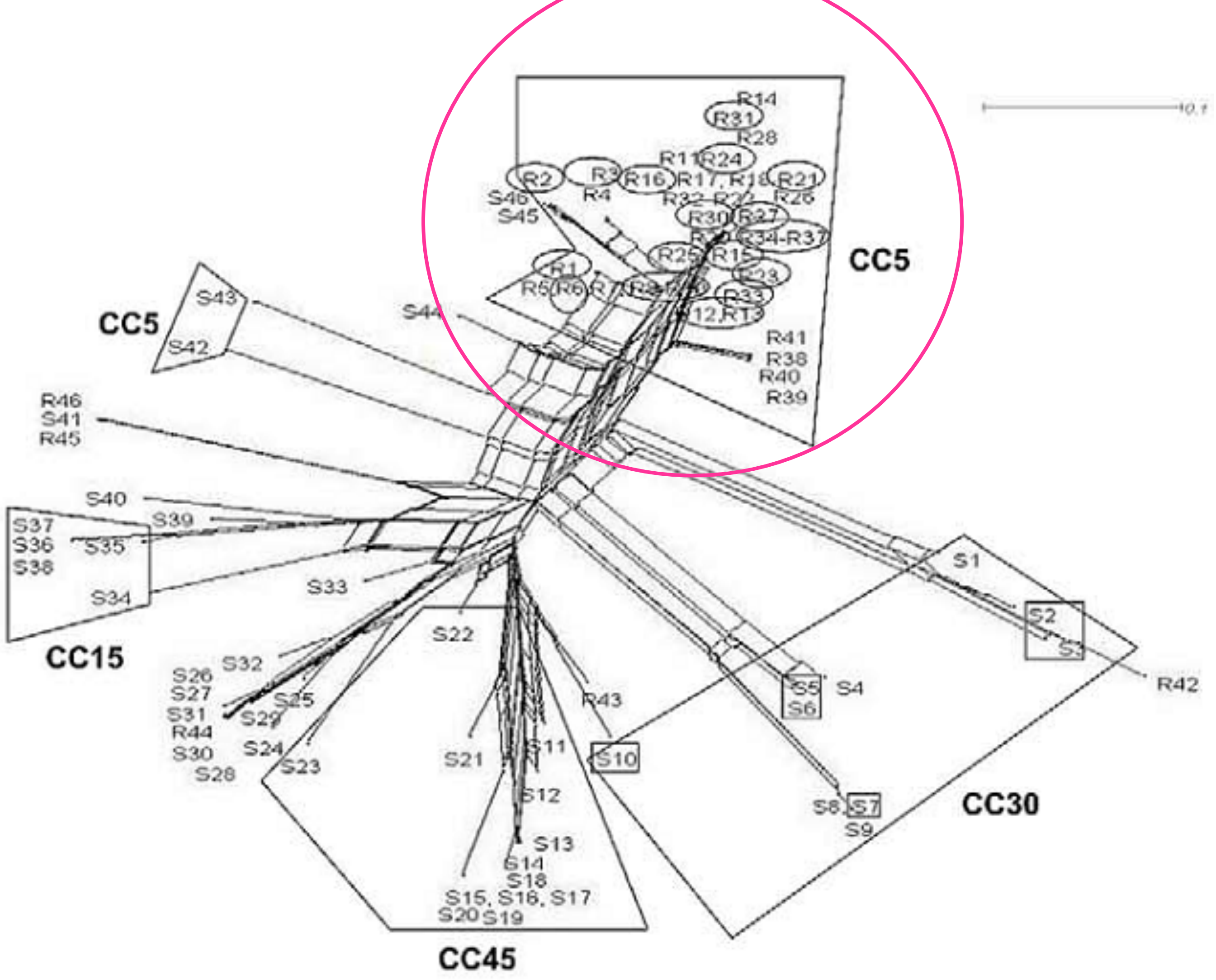
- DNA preparation of polymorphic X-region of ***staphylococcus* protein A** from *S. aureus* (Spa)
- amplify by PCR
- sequencing assignment using Ridom StaphType software

Ridom Spa Server

Spa-types:	Repeats:	Total strains:	Strain records:	Strain countries:
17897	762	398228	165914	135



Results from Spa-typing: splits graph



Unrouted tree generated with
www.splitstree.org

MSSA strains labeled S__
MRSA strains labeled R__

For MSSA, *spa*-typing allowed for good discrimination of patient isolates.

However, the majority of MRSA isolates clustered into clonal complex **CC5/t003.**

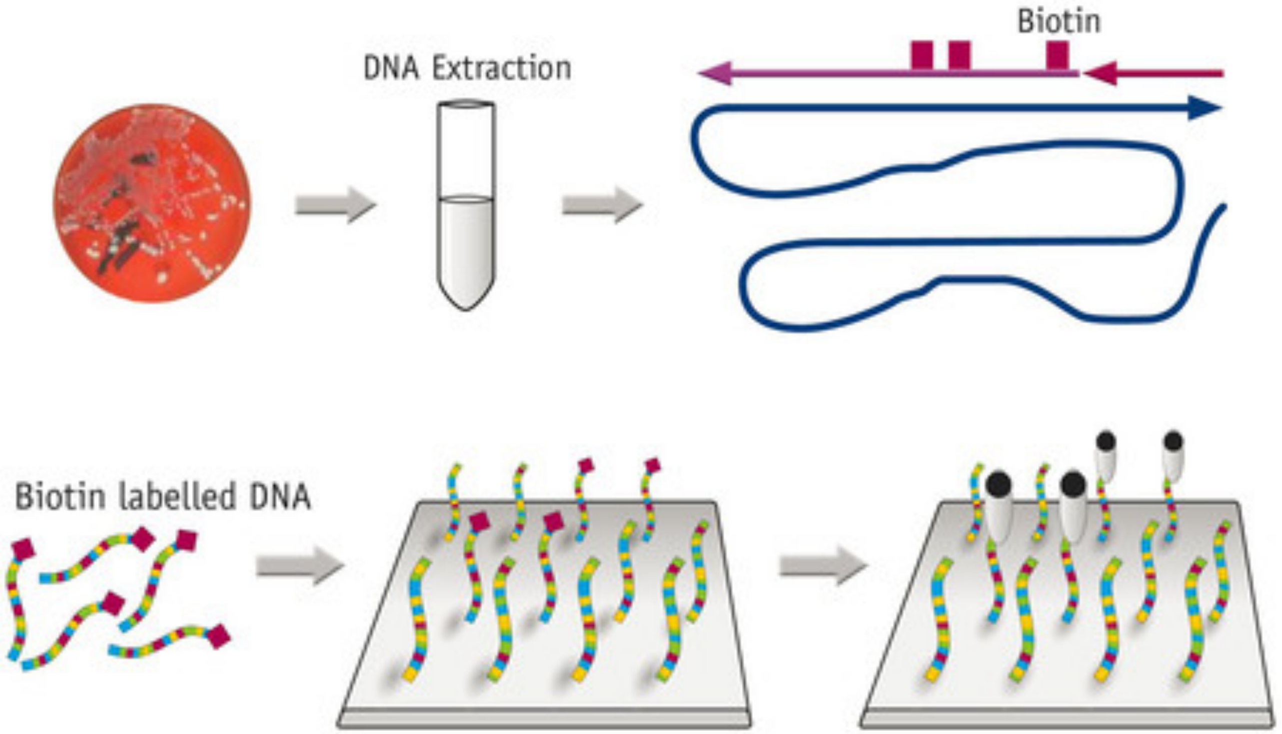
This hampers sub-classification by *spa*-typing

DNA microarray (IdentiBAC – Alere)

Microarray contains 334 DNA probes for genes/regions that are clinically relevant and/or relevant for clonal typing

alere-technologies.com

DNA microarray principle



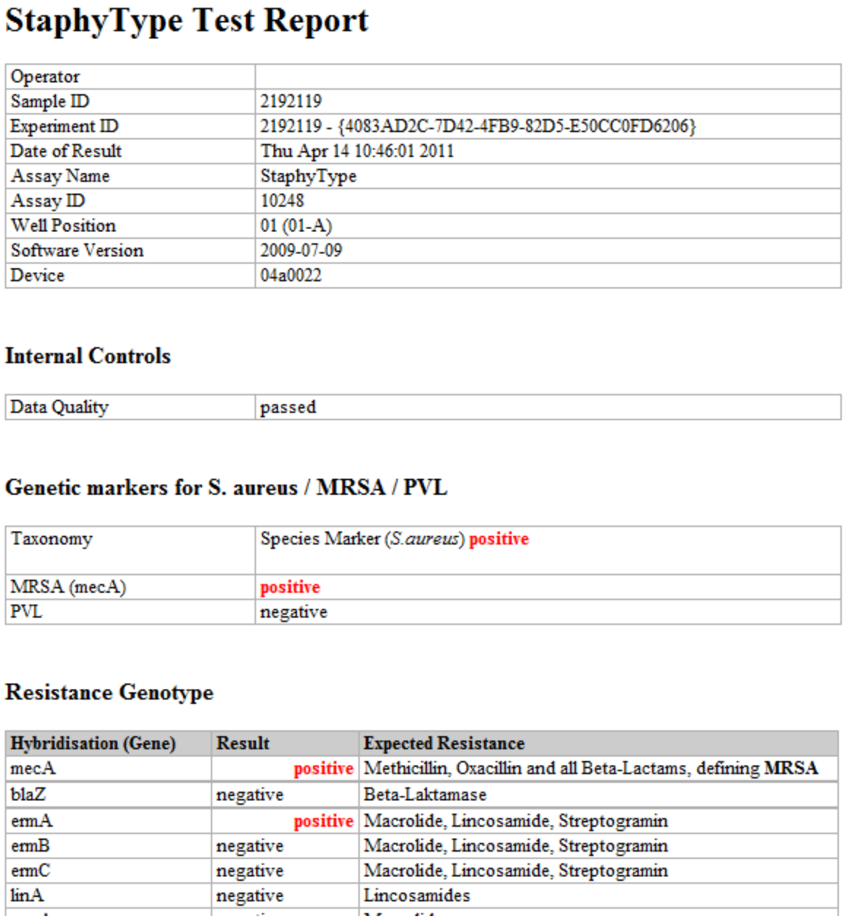
The extracted RNA free genomic DNA from the bacterial overnight culture is internally biotin labelled through a set of antisense primers.

The resulting single stranded and biotin labelled amplicons are hybridized to a set of discriminative probes that are covalently bound onto the microarrays.

The biotin labelled DNA bound to the probes on the array is subsequently stained.

alere-technologies.com

Process microarray data (334 probes)



StaphyType Test Report

Operator	
Sample ID	2192119
Experiment ID	2192119 - {4083AD2C-7D42-4FB9-82D5-E50CC0FD6206}
Date of Result	Thu Apr 14 10:46:01 2011
Assay Name	StaphyType
Assay ID	10248
Well Position	01 (01-A)
Software Version	2009-07-09
Device	04a0022

Internal Controls

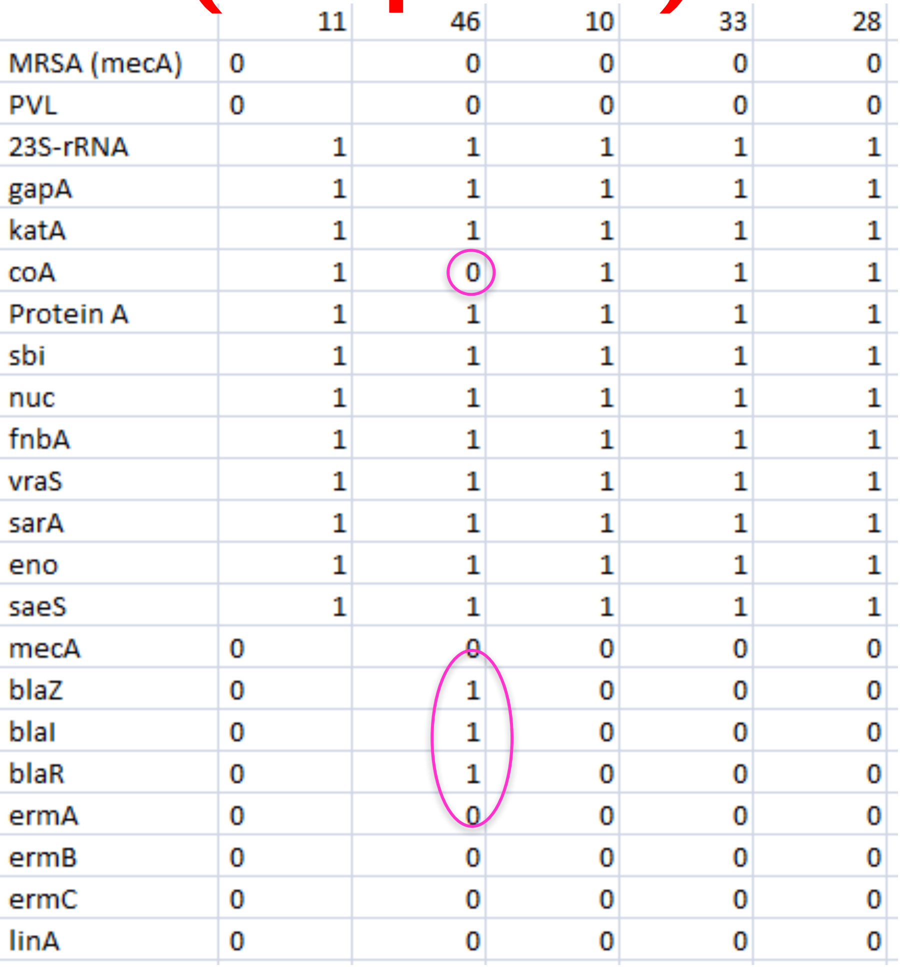
Data Quality	passed

Genetic markers for S. aureus / MRSA / PVL

Taxonomy	Species Marker (*S.aureus*) **positive**
MRSA (mecA)	**positive**
PVL	negative

Resistance Genotype

Hybridisation (Gene)	Result	Expected Resistance
mecA	**positive**	Methicillin, Oxacillin and all Beta-Lactams, defining **MRSA**
blaZ	negative	Beta-Laktamase
ermA	**positive**	Macrolide, Lincosamide, Streptogramin
ermB	negative	Macrolide, Lincosamide, Streptogramin
ermC	negative	Macrolide, Lincosamide, Streptogramin
linA	negative	Lincosamides



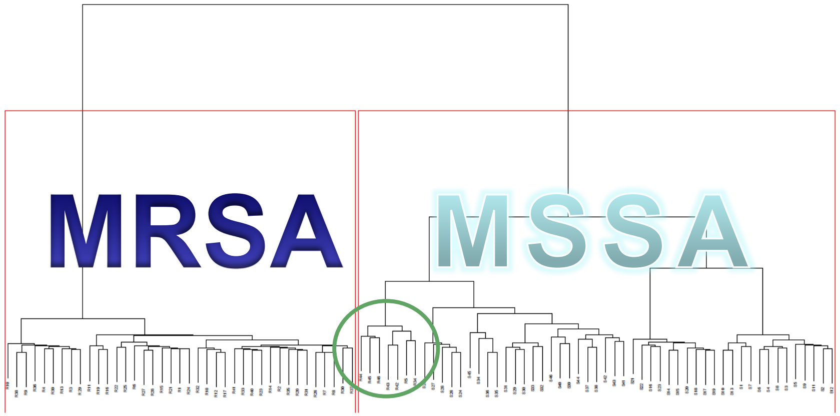
	11	46	10	33	28
MRSA (mecA)	0	0	0	0	0
PVL	0	0	0	0	0
23S-rRNA	1	1	1	1	1
gapA	1	1	1	1	1
katA	1	1	1	1	1
coA	1	0	1	1	1
Protein A	1	1	1	1	1
sbi	1	1	1	1	1
nuc	1	1	1	1	1
fnbA	1	1	1	1	1
vraS	1	1	1	1	1
sarA	1	1	1	1	1
eno	1	1	1	1	1
saeS	1	1	1	1	1
mecA	0	0	0	0	0
blaZ	0	1	0	0	0
blaI	0	1	0	0	0
blaR	0	1	0	0	0
ermA	0	0	0	0	0
ermB	0	0	0	0	0
ermC	0	0	0	0	0
linA	0	0	0	0	0

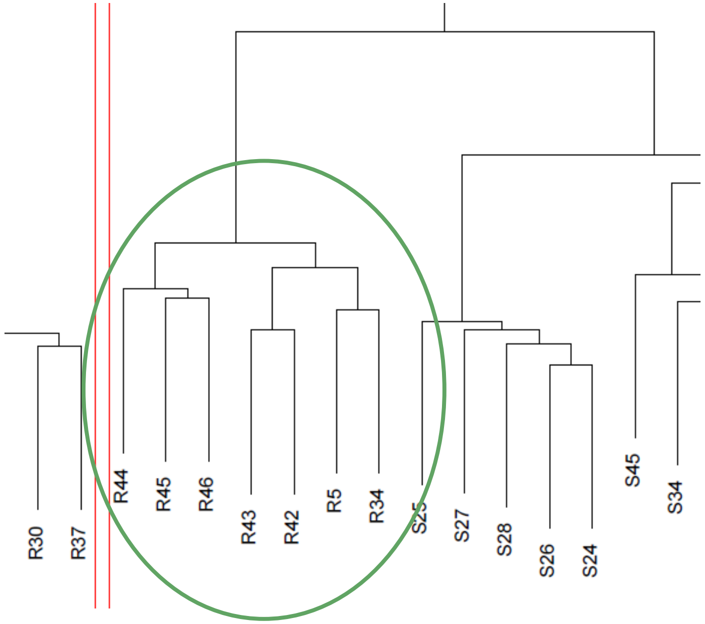
Simple idea: Compute **Euclidian distance** between samples

$$\|a - b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$$

Other distances are possible, also weighted distances, where some probes get higher weights.

Hierarchical agglomerative clustering based on MA data





Hierarchical clustering:

(1) Calculate pairwise distance matrix for all samples to be clustered.

(2) Search distance matrix for two most similar samples or clusters (initially each cluster consists of a single sample).

If several pairs have the same separation distance, a predetermined rule is used to decide between alternatives.

(3) The two selected clusters are merged to produce a new cluster that now contains at least two objects.

(4) The distances are calculated between this new cluster and all other clusters.

(5) Repeat steps 2–4 until all objects are in one cluster.

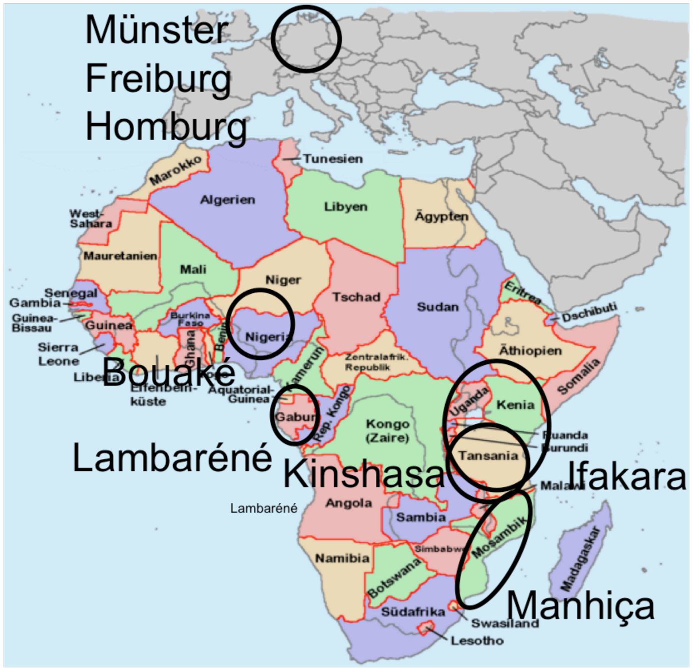
Clustering based on Euclidian distance yields almost perfect separation between MSSA/MRSA

except the encircled resistant samples

S. aureus in Germany vs. Africa: StaphNet

6 study sites each collected 100 isolates of healthy volunteers and 100 of blood culture or clinical infection sites

→ 1200 isolates



Aim

microbiological and molecular characterization of African *S. aureus* isolates

by DNA microarray analysis including clonal complex analysis

supplemented by Whole Genome Sequencing

What does the microarray measure?

Naively, one can interpret the microarray result as

1 : gene is present in the strain

0 : gene is not present in the strain

However**, false negative** non-detections of particular targets may occur due to **non-binding** of the sample amplicon to the microarray's probe or primer oligonucleotide due to **polymorphisms** in the respective target gene.

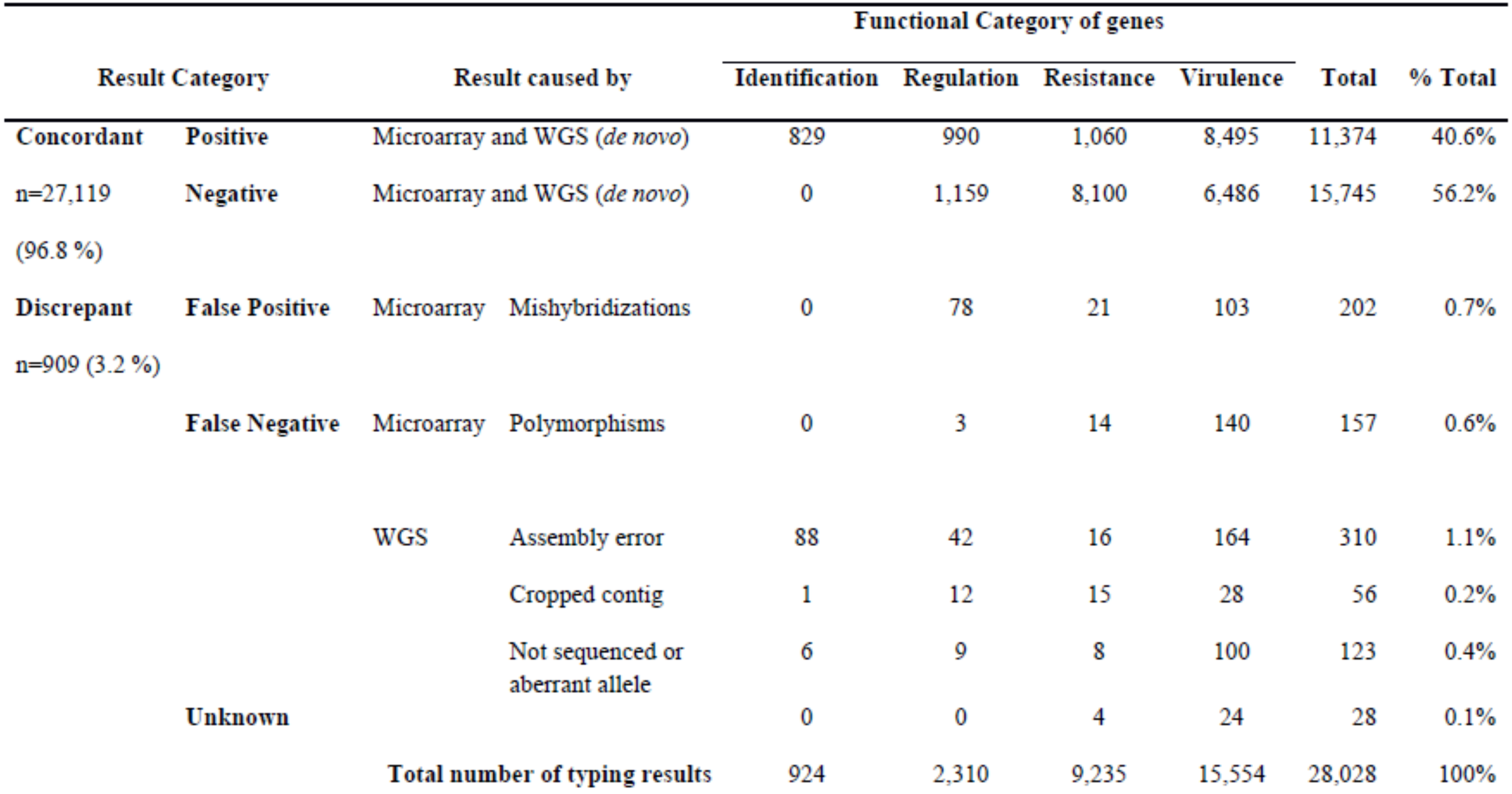
On the other hand, **false positive results** may occur between highly similar probe and amplicon sequences, e. g. between agrI and agrIV.

→ check MA results by whole genome sequencing

Strauss et al. J Clin Microbiol (2016)

MA assignment to CCs confirmed by whole-genome sequencing

154 *S. aureus* isolates (182 target genes) from Germany-vs-Africa study

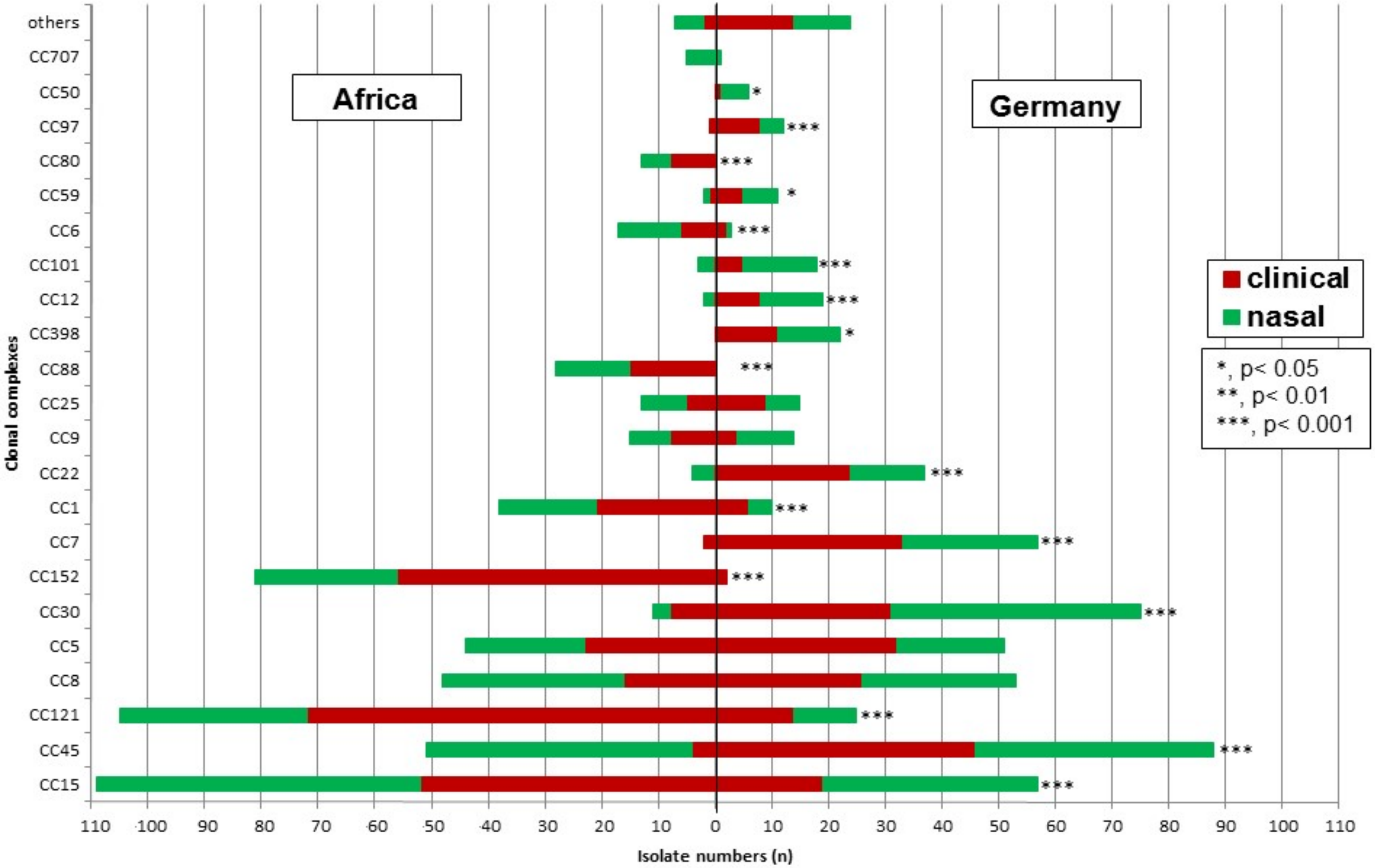


				Functional Category of genes					
Result Category		Result caused by		Identification	Regulation	Resistance	Virulence	Total	% Total
Concordant	**Positive**	Microarray and WGS (*de novo*)		829	990	1,060	8,495	11,374	40.6%
n=27,119	**Negative**	Microarray and WGS (*de novo*)		0	1,159	8,100	6,486	15,745	56.2%
(96.8 %)									
Discrepant	**False Positive**	Microarray	Mishybridizations	0	78	21	103	202	0.7%
n=909 (3.2 %)									
	False Negative	Microarray	Polymorphisms	0	3	14	140	157	0.6%
		WGS	Assembly error	88	42	16	164	310	1.1%
			Cropped contig	1	12	15	28	56	0.2%
			Not sequenced or aberrant allele	6	9	8	100	123	0.4%
	Unknown			0	0	4	24	28	0.1%
		Total number of typing results		924	2,310	9,235	15,554	28,028	100%

→ **97% agreement of MA and WGS**

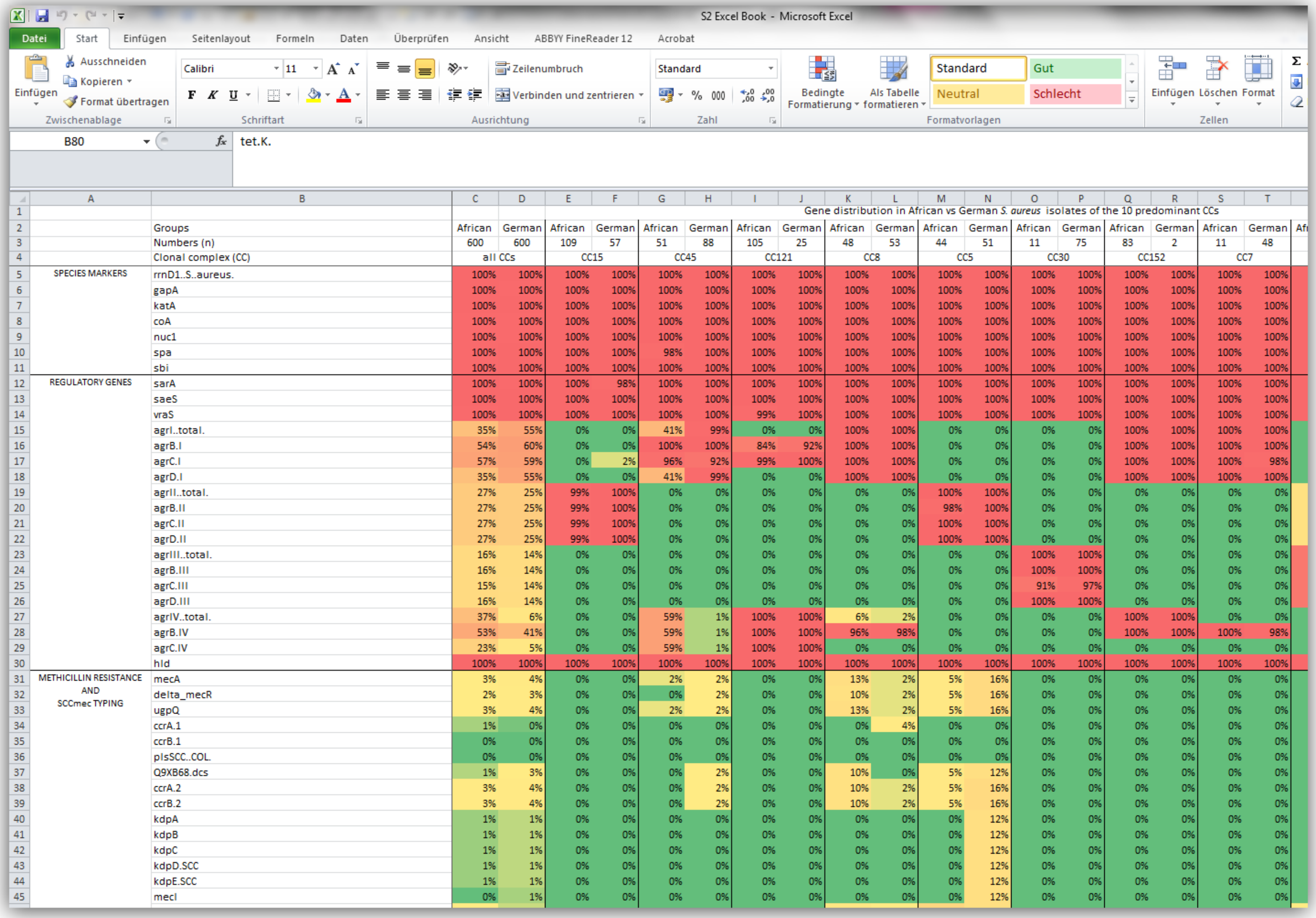
Strauss et al. J Clin Microbiol (2016)

Distribution of clonal complexes



Some clonal complexes (CC) are more prevalent in Africa, others predominant in Germany.

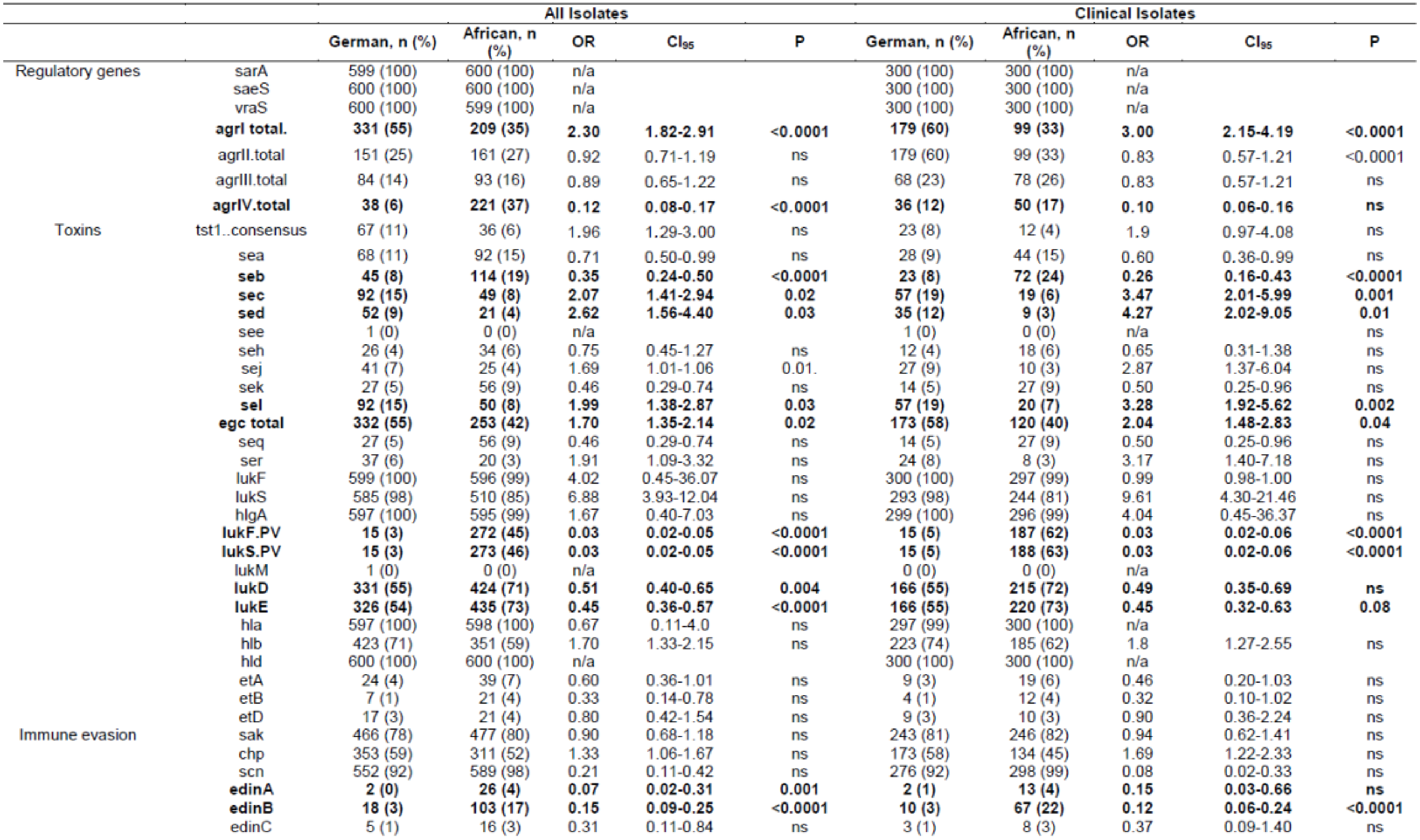
Activitity of individual probes for CCs



B80 fx tet.K.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1					Gene distribution in African vs German *S. aureus* isolates of the 10 predominant CCs																
2		Groups	African	German	African	German	African	German	African	German	African	German	African	German	African	German	African	German	African	German	Af
3		Numbers (n)	600	600	109	57	51	88	105	25	48	53	44	51	11	75	83	2	11	48	
4		Clonal complex (CC)	all CCs		CC15		CC45		CC121		CC8		CC5		CC30		CC152		CC7		
5	SPECIES MARKERS	rrnD1..S..aureus.	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
6		gapA	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
7		katA	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
8		coA	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
9		nuc1	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
10		spa	100%	100%	100%	100%	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
11		sbi	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
12	REGULATORY GENES	sarA	100%	100%	100%	98%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
13		saeS	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
14		vraS	100%	100%	100%	100%	100%	100%	99%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
15		agrI..total.	35%	55%	0%	0%	41%	99%	0%	0%	100%	100%	0%	0%	0%	0%	100%	100%	100%	100%	
16		agrB.I	54%	60%	0%	0%	100%	100%	84%	92%	100%	100%	0%	0%	0%	0%	100%	100%	100%	100%	
17		agrC.I	57%	59%	0%	2%	96%	92%	99%	100%	100%	100%	0%	0%	0%	0%	100%	100%	100%	98%	
18		agrD.I	35%	55%	0%	0%	41%	99%	0%	0%	100%	100%	0%	0%	0%	0%	100%	100%	100%	100%	
19		agrII..total.	27%	25%	99%	100%	0%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	0%	0%	
20		agrB.II	27%	25%	99%	100%	0%	0%	0%	0%	0%	0%	98%	100%	0%	0%	0%	0%	0%	0%	
21		agrC.II	27%	25%	99%	100%	0%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	0%	0%	
22		agrD.II	27%	25%	99%	100%	0%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	0%	0%	
23		agrIII..total.	16%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	
24		agrB.III	16%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	
25		agrC.III	15%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	91%	97%	0%	0%	0%	0%	
26		agrD.III	16%	14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	
27		agrIV..total.	37%	6%	0%	0%	59%	1%	100%	100%	6%	2%	0%	0%	0%	0%	100%	100%	0%	0%	
28		agrB.IV	53%	41%	0%	0%	59%	1%	100%	100%	96%	98%	0%	0%	0%	0%	100%	100%	100%	98%	
29		agrC.IV	23%	5%	0%	0%	59%	1%	100%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	
30		hld	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	
31	METHICILLIN RESISTANCE	mecA	3%	4%	0%	0%	2%	2%	0%	0%	13%	2%	5%	16%	0%	0%	0%	0%	0%	0%	
32	AND	delta_mecR	2%	3%	0%	0%	0%	2%	0%	0%	10%	2%	5%	16%	0%	0%	0%	0%	0%	0%	
33	SCCmec TYPING	ugpQ	3%	4%	0%	0%	2%	2%	0%	0%	13%	2%	5%	16%	0%	0%	0%	0%	0%	0%	
34		ccrA.1	1%	0%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	
35		ccrB.1	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	
36		plsSCC..COL.	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	
37		Q9XB68.dcs	1%	3%	0%	0%	0%	2%	0%	0%	10%	0%	5%	12%	0%	0%	0%	0%	0%	0%	
38		ccrA.2	3%	4%	0%	0%	0%	2%	0%	0%	10%	2%	5%	16%	0%	0%	0%	0%	0%	0%	
39		ccrB.2	3%	4%	0%	0%	0%	2%	0%	0%	10%	2%	5%	16%	0%	0%	0%	0%	0%	0%	
40		kdpA	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	
41		kdpB	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	
42		kdpC	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	
43		kdpD.SCC	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	
44		kdpE.SCC	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	
45		mecI	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	12%	0%	0%	0%	0%	0%	0%	

Imbalance of hybridizing resistance genes?



		All Isolates					Clinical Isolates				
		German, n (%)	African, n (%)	OR	CI_{95}	P	German, n (%)	African, n (%)	OR	CI_{95}	P
Regulatory genes	sarA	599 (100)	600 (100)	n/a			300 (100)	300 (100)	n/a		
	saeS	600 (100)	600 (100)	n/a			300 (100)	300 (100)	n/a		
	vraS	600 (100)	599 (100)	n/a			300 (100)	300 (100)	n/a		
	agrI total.	**331 (55)**	**209 (35)**	**2.30**	**1.82-2.91**	**<0.0001**	**179 (60)**	**99 (33)**	**3.00**	**2.15-4.19**	**<0.0001**
	agrII.total	151 (25)	161 (27)	0.92	0.71-1.19	ns	179 (60)	99 (33)	0.83	0.57-1.21	<0.0001
	agrIII.total	84 (14)	93 (16)	0.89	0.65-1.22	ns	68 (23)	78 (26)	0.83	0.57-1.21	ns
	agrIV.total	**38 (6)**	**221 (37)**	**0.12**	**0.08-0.17**	**<0.0001**	**36 (12)**	**50 (17)**	**0.10**	**0.06-0.16**	**ns**
Toxins	tst1..consensus	67 (11)	36 (6)	1.96	1.29-3.00	ns	23 (8)	12 (4)	1.9	0.97-4.08	ns
	sea	68 (11)	92 (15)	0.71	0.50-0.99	ns	28 (9)	44 (15)	0.60	0.36-0.99	ns
	seb	**45 (8)**	**114 (19)**	**0.35**	**0.24-0.50**	**<0.0001**	**23 (8)**	**72 (24)**	**0.26**	**0.16-0.43**	**<0.0001**
	sec	**92 (15)**	**49 (8)**	**2.07**	**1.41-2.94**	**0.02**	**57 (19)**	**19 (6)**	**3.47**	**2.01-5.99**	**0.001**
	sed	**52 (9)**	**21 (4)**	**2.62**	**1.56-4.40**	**0.03**	**35 (12)**	**9 (3)**	**4.27**	**2.02-9.05**	**0.01**
	see	1 (0)	0 (0)	n/a			1 (0)	0 (0)	n/a		ns
	seh	26 (4)	34 (6)	0.75	0.45-1.27	ns	12 (4)	18 (6)	0.65	0.31-1.38	ns
	sej	41 (7)	25 (4)	1.69	1.01-1.06	0.01.	27 (9)	10 (3)	2.87	1.37-6.04	ns
	sek	27 (5)	56 (9)	0.46	0.29-0.74	ns	14 (5)	27 (9)	0.50	0.25-0.96	ns
	sel	**92 (15)**	**50 (8)**	**1.99**	**1.38-2.87**	**0.03**	**57 (19)**	**20 (7)**	**3.28**	**1.92-5.62**	**0.002**
	egc total	**332 (55)**	**253 (42)**	**1.70**	**1.35-2.14**	**0.02**	**173 (58)**	**120 (40)**	**2.04**	**1.48-2.83**	**0.04**
	seq	27 (5)	56 (9)	0.46	0.29-0.74	ns	14 (5)	27 (9)	0.50	0.25-0.96	ns
	ser	37 (6)	20 (3)	1.91	1.09-3.32	ns	24 (8)	8 (3)	3.17	1.40-7.18	ns
	lukF	599 (100)	596 (99)	4.02	0.45-36.07	ns	300 (100)	297 (99)	0.99	0.98-1.00	ns
	lukS	585 (98)	510 (85)	6.88	3.93-12.04	ns	293 (98)	244 (81)	9.61	4.30-21.46	ns
	hlgA	597 (100)	595 (99)	1.67	0.40-7.03	ns	299 (100)	296 (99)	4.04	0.45-36.37	ns
	lukF.PV	**15 (3)**	**272 (45)**	**0.03**	**0.02-0.05**	**<0.0001**	**15 (5)**	**187 (62)**	**0.03**	**0.02-0.06**	**<0.0001**
	lukS.PV	**15 (3)**	**273 (46)**	**0.03**	**0.02-0.05**	**<0.0001**	**15 (5)**	**188 (63)**	**0.03**	**0.02-0.06**	**<0.0001**
	lukM	1 (0)	0 (0)	n/a			0 (0)	0 (0)	n/a		
	lukD	**331 (55)**	**424 (71)**	**0.51**	**0.40-0.65**	**0.004**	**166 (55)**	**215 (72)**	**0.49**	**0.35-0.69**	**ns**
	lukE	**326 (54)**	**435 (73)**	**0.45**	**0.36-0.57**	**<0.0001**	**166 (55)**	**220 (73)**	**0.45**	**0.32-0.63**	**0.08**
	hla	597 (100)	598 (100)	0.67	0.11-4.0	ns	297 (99)	300 (100)	n/a		
	hlb	423 (71)	351 (59)	1.70	1.33-2.15	ns	223 (74)	185 (62)	1.8	1.27-2.55	ns
	hld	600 (100)	600 (100)	n/a			300 (100)	300 (100)	n/a		
	etA	24 (4)	39 (7)	0.60	0.36-1.01	ns	9 (3)	19 (6)	0.46	0.20-1.03	ns
	etB	7 (1)	21 (4)	0.33	0.14-0.78	ns	4 (1)	12 (4)	0.32	0.10-1.02	ns
	etD	17 (3)	21 (4)	0.80	0.42-1.54	ns	9 (3)	10 (3)	0.90	0.36-2.24	ns
Immune evasion	sak	466 (78)	477 (80)	0.90	0.68-1.18	ns	243 (81)	246 (82)	0.94	0.62-1.41	ns
	chp	353 (59)	311 (52)	1.33	1.06-1.67	ns	173 (58)	134 (45)	1.69	1.22-2.33	ns
	scn	552 (92)	589 (98)	0.21	0.11-0.42	ns	276 (92)	298 (99)	0.08	0.02-0.33	ns
	edinA	**2 (0)**	**26 (4)**	**0.07**	**0.02-0.31**	**0.001**	**2 (1)**	**13 (4)**	**0.15**	**0.03-0.66**	**ns**
	edinB	**18 (3)**	**103 (17)**	**0.15**	**0.09-0.25**	**<0.0001**	**10 (3)**	**67 (22)**	**0.12**	**0.06-0.24**	**<0.0001**
	edinC	5 (1)	16 (3)	0.31	0.11-0.84	ns	3 (1)	8 (3)	0.37	0.09-1.40	ns

OR: odds ratio ; ratio of events to non-events

CI_{95} : confidence interval

Antibiotic resistance

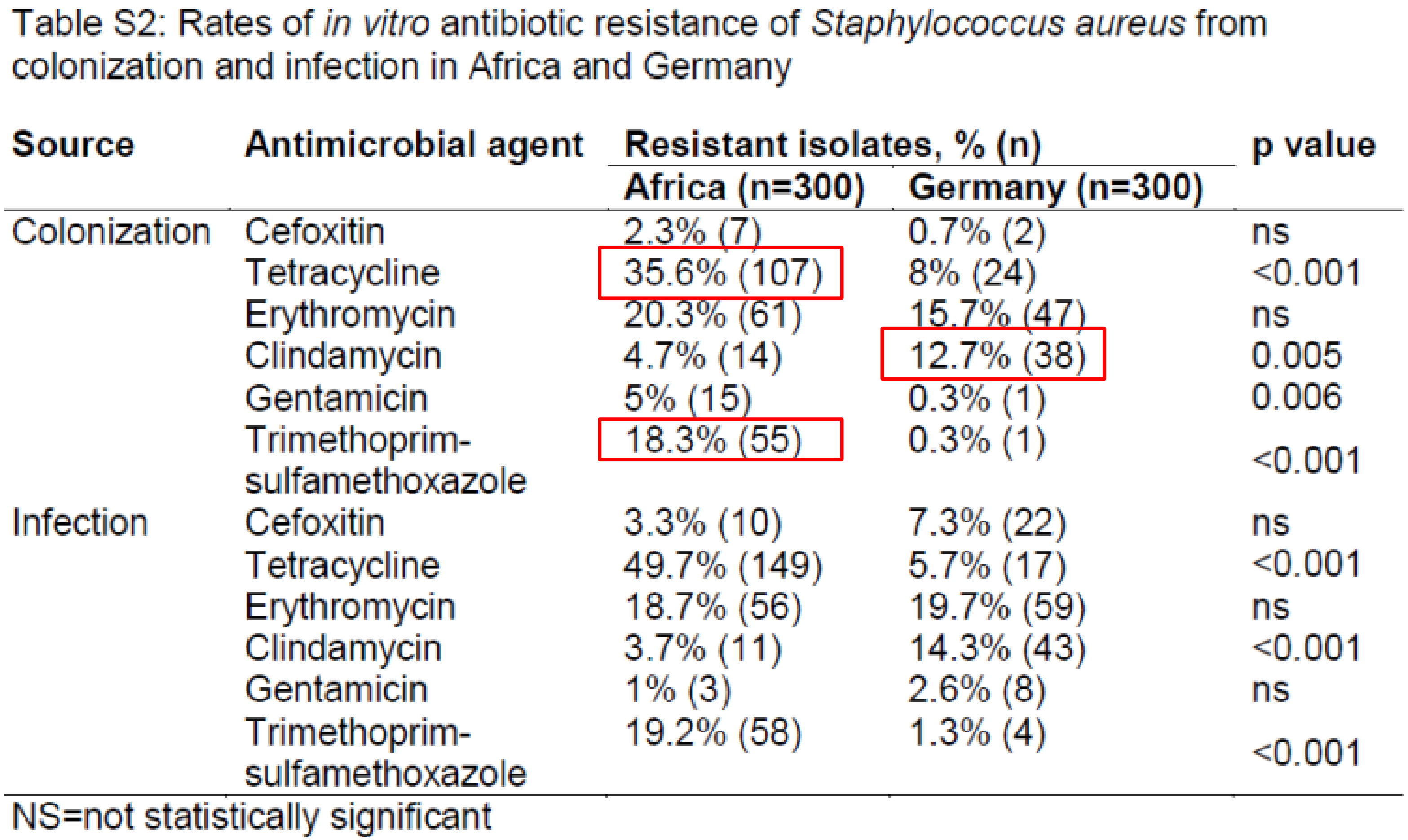


Table S2: Rates of *in vitro* antibiotic resistance of *Staphylococcus aureus* from colonization and infection in Africa and Germany

Source	Antimicrobial agent	Resistant isolates, % (n)		p value
		Africa (n=300)	Germany (n=300)	
Colonization	Cefoxitin	2.3% (7)	0.7% (2)	ns
	Tetracycline	35.6% (107)	8% (24)	<0.001
	Erythromycin	20.3% (61)	15.7% (47)	ns
	Clindamycin	4.7% (14)	12.7% (38)	0.005
	Gentamicin	5% (15)	0.3% (1)	0.006
	Trimethoprim-sulfamethoxazole	18.3% (55)	0.3% (1)	<0.001
Infection	Cefoxitin	3.3% (10)	7.3% (22)	ns
	Tetracycline	49.7% (149)	5.7% (17)	<0.001
	Erythromycin	18.7% (56)	19.7% (59)	ns
	Clindamycin	3.7% (11)	14.3% (43)	<0.001
	Gentamicin	1% (3)	2.6% (8)	ns
	Trimethoprim-sulfamethoxazole	19.2% (58)	1.3% (4)	<0.001

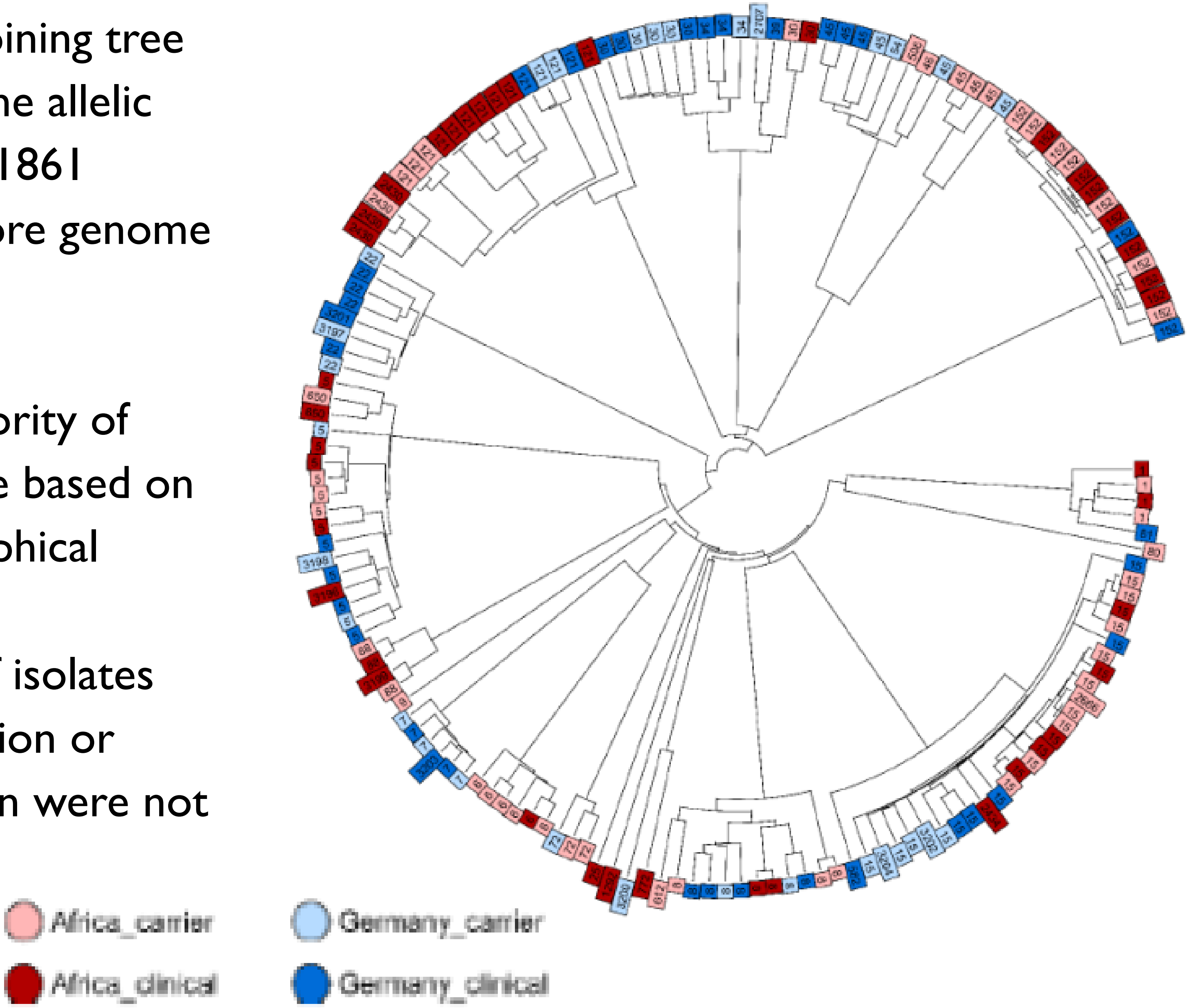
NS=not statistically significant

The majority of resistance genes were equally distributed among isolates from Africa and Germany. Striking differences in phenotypic resistance could be observed for tetracycline and trimethoprim-sulfamethoxazole with a larger proportion of resistant isolates in the African population, and clindamycin, with resistance more prevalent among German isolates

Phylogenetic tree based on WGS data of 154 strains

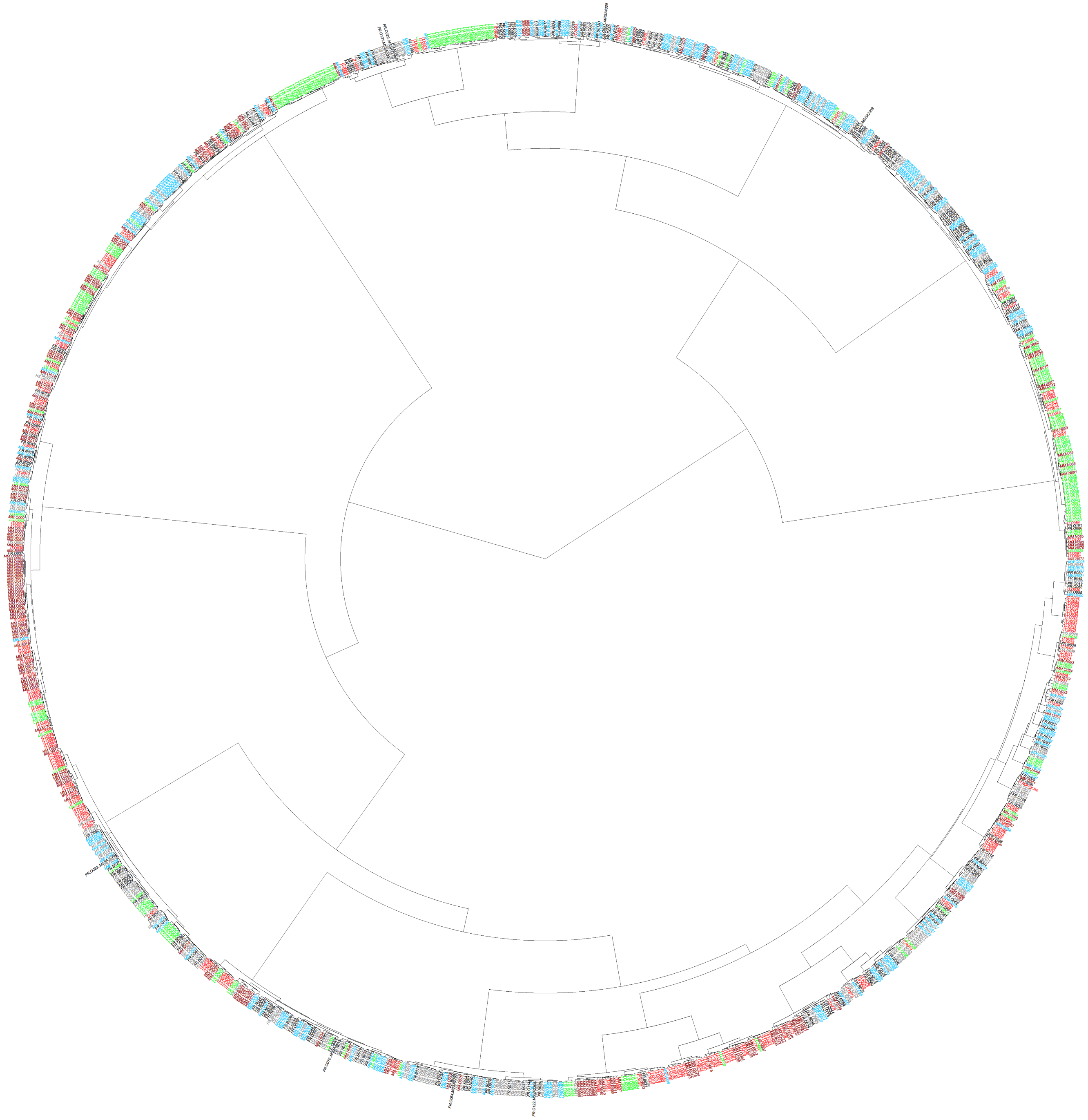
neighbor-joining tree based on the allelic profiles of 1861 *S. aureus* core genome features.

-> the majority of clusters are based on the geographical region.
Clusters of isolates from infection or colonization were not detected



Clustering of all 1200 microarray samples is not handy

Can’t see too much

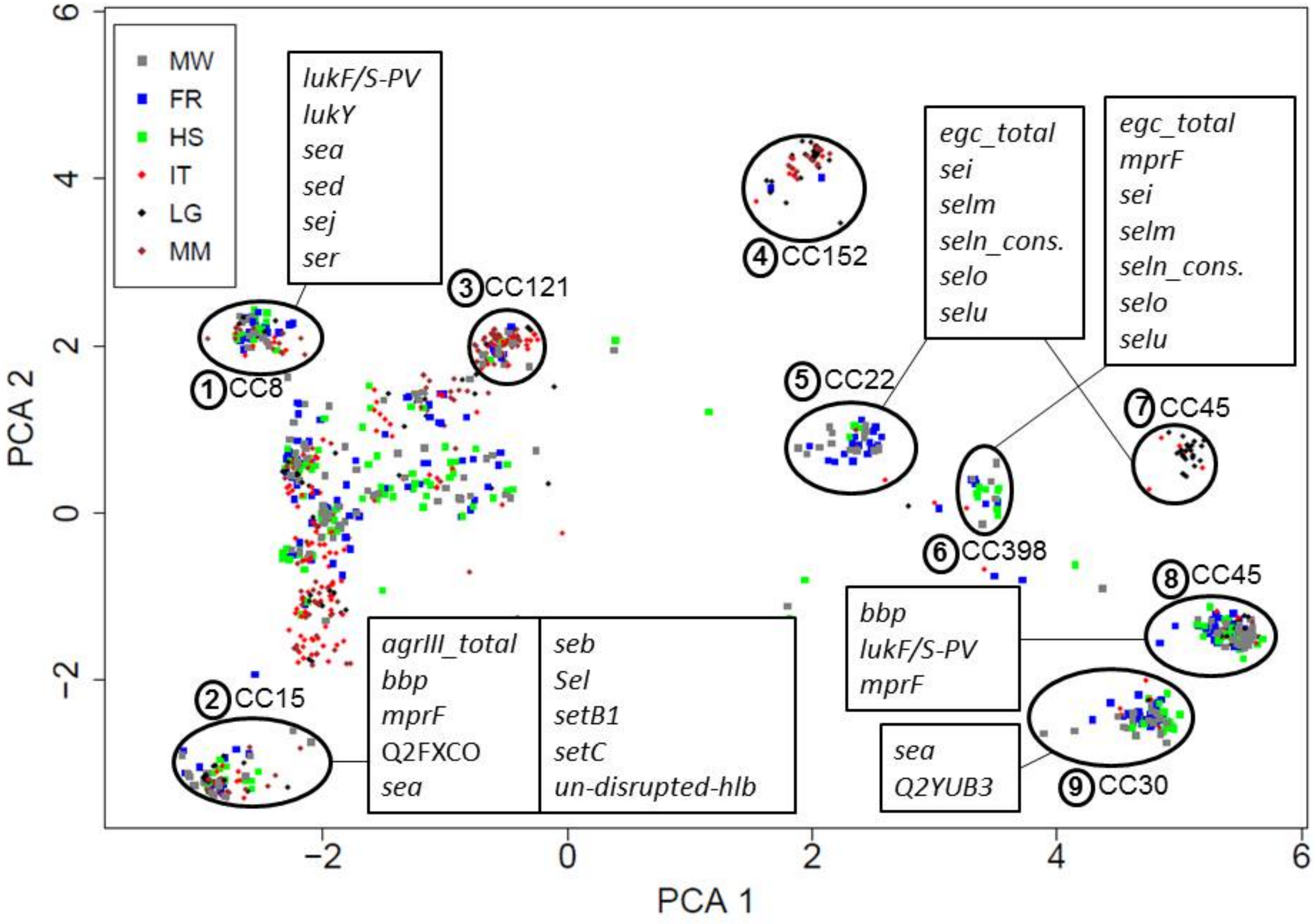


Principle component analysis of 1200 strains

Input data: binary matrix of MA data; dimension 1200 x 334 probes

PCA identifies local gene clusters that are characteristic

for particular clonal complexes



Color code:

6 different sites

Marked in boxes:

Characteristic genes present in this cluster.

PCA- intro

PCA is the most popular multivariate statistical technique.
It is used by almost all scientific disciplines.

It is likely also the oldest multivariate technique.

Its origin can be traced back to Pearson, Cauchy, Jordan, Cayley etc

This part of the lecture is based on the article
"Principal component analysis" by Herve Abdi & Lynne J. Williams in
WIREs Computational Statistics, 2, 433-459 (2010)

PCA- intro

PCA analyzes a data table **X** representing observations described by several dependent variables, which are, in general, inter-correlated.

The goal of PCA is to extract the important information from the data table and express this information as a set of new orthogonal variables called **principal components**.

We will consider a data table **X** of *I* observations and *J* variables.

The elements are x_{ij}.

The matrix **X** has rank *L* where $L \leq \min [I,J]$

PCA- preprocessing data entries

In general, the data table will be **preprocessed** before the analysis.

The columns of **X** are **centered** so that the **mean** of each column is equal to 0.

$$x_{ij} \rightarrow x_{ij} - \mu_j$$

If in addition, each element of **X** is divided by $\sqrt{I}$ or $\sqrt{I-1}$,
the matrix Σ = $\mathbf{X}^T\mathbf{X}$ is a covariance matrix,

$$\Sigma = \left[(\mathbf{X}-\mu)^T(\mathbf{X}-\mu) \quad \right]$$

and the analysis is referred to as **covariance PCA**.

PCA- preprocessing data entries

In addition to centering, when the variables are measured with different units, it is customary to **standardize** each variable to **unit norm**.

This is obtained by dividing each variable by its norm (i.e. the square root of the sum of all squared elements of this variable) $\sqrt{\sum_i (x_i)^2}$, which is equivalent to dividing it by its standard deviation (except dividing by *n* vs *n*-1).

In this case, the analysis is referred to as a **correlation PCA** because, then, then matrix $\mathbf{X}^T\mathbf{X}$ is a correlation matrix.

We will make use of the fact that the matrix **X** has a **singular value decomposition (SVD)**

$$\mathbf{X} = \mathbf{P}\Delta\mathbf{Q}^T$$

What is a SVD?

Insert: review of eigenvalues

A vector **u** that satisfies $\mathbf{A}\,\mathbf{u} = \lambda\,\mathbf{u}$

or $(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{u} = 0$

is an **eigenvector** of this matrix **A**.

The scalar value λ is the **eigenvalue** associated with this eigenvector.

For example, the matrix $\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}$ has the eigenvectors

$u_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ with eigenvalue $\lambda_1 = 4$.

Test $2 \cdot 3 + 3 \cdot 2 = 4 \cdot 3$; $2 \cdot 3 + 1 \cdot 2 = 4 \cdot 2$

and

$u_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ with eigenvalue $\lambda_1 = -1$.

Test $2 \cdot (-1) + 3 \cdot 1 = (-1) \cdot (-1)$; $2 \cdot (-1) + 1 \cdot 1 = (-1) \cdot 1$

Insert: review of eigenvalues

For most applications we normalize the eigenvectors so that their length is equal to 1, i.e.

$$\mathbf{u}^T\mathbf{u} = 1$$

Traditionally, we put the set of eigenvectors of **A** in a matrix denoted by **U**.

Then, each column of **U** contains an eigenvector of **A**.

The eigenvalues are stored as diagonal elements of a diagonal matrix Λ .

Then we can write $\mathbf{A}\,\mathbf{U} = \mathbf{U}\,\Lambda$ or: $\mathbf{A} = \mathbf{U}\,\Lambda\,\mathbf{U}^{-1}$ (if we multiply with $\mathbf{U}^{-1}$)

This is the **eigendecomposition** of this matrix. Not all matrices have a EDC.

Insert: positive (semi-) definite matrices

A type of matrices used often in statistics are called **positive semi-definite** (PSD)

The eigen-decomposition of such matrices always exists, and has a particularly convenient form.

A matrix **A** is positive (semi-)definite, if there exists a real-valued matrix **X** and

$$\mathbf{A} = \mathbf{X}\,\mathbf{X}^T$$

Correlation matrices, covariance, and cross-product matrices are all semi-definite matrices.

The eigenvalues of PSD matrices are always positive or null.

The eigenvectors of PSD are pairwise orthogonal when their eigenvalues are different.

Insert: positive (semi-) definite matrices

This implies $\mathbf{U}^{-1} = \mathbf{U}^T$

Then we can express **A** as $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^T$ with $\mathbf{U}^T\mathbf{U} = 1$

where **U** is the matrix storing the normalized eigenvectors.

E.g. $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$ can be decomposed as

$$\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^{-1} = \begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} 4\sqrt{\frac{1}{2}} & 4\sqrt{\frac{1}{2}} \\ 2\sqrt{\frac{1}{2}} & -2\sqrt{\frac{1}{2}} \end{bmatrix} = \begin{bmatrix} 2+1 & 2-1 \\ 2-1 & 2+1 \end{bmatrix}$$

with $\begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ showing that the 2 eigenvectors are orthonormal.

Singular Value Decomposition (SVD)

SVD is a generalization of the eigen-decomposition.

SVD decomposes a rectangular matrix **A** into three simple matrices:
two orthogonal matrices **P** and **Q** and one diagonal matrix Δ.

$$\mathbf{A} = \mathbf{P}\Delta\mathbf{Q}^T$$

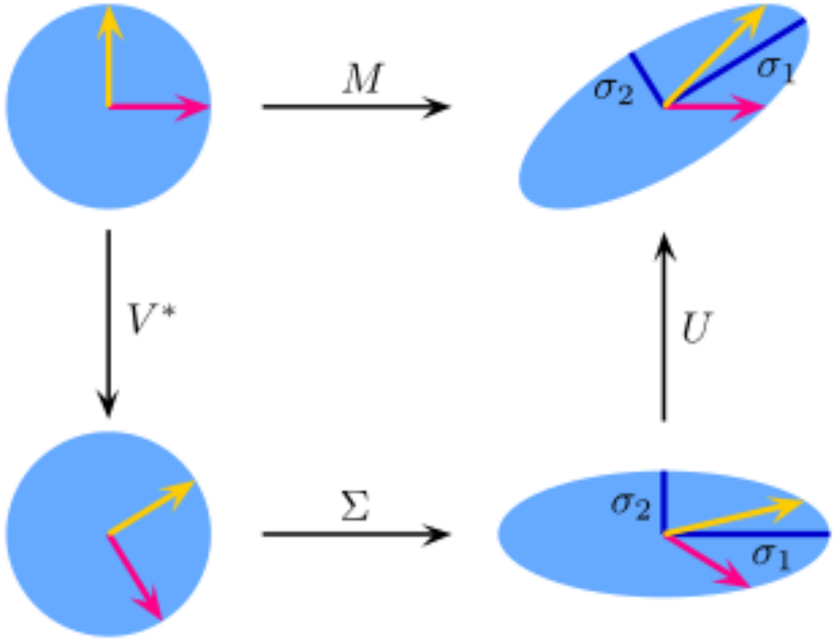
P : contains the normalized eigenvectors of the matrix $\mathbf{A}\,\mathbf{A}^T$. (i.e. $\mathbf{P}^T\mathbf{P} = \mathbf{1}$)
The columns of **P** are called *left singular vectors* of **A**.

Q : contains the normalized eigenvectors of the matrix $\mathbf{A}^T\mathbf{A}$. (i.e. $\mathbf{Q}^T\mathbf{Q} = \mathbf{1}$)
The columns of **Q** are called *right singular vectors* of **A**.

Δ : the diagonal matrix of the *singular values*. They are the square root values of the eigenvalues of matrix $\mathbf{A}\,\mathbf{A}^T$ (they are the same as those of $\mathbf{A}^T\mathbf{A}$).

Interpretation of SVD

In the special, yet common, case when **M** is an *m* × *m* real square matrix with positive determinant: **U**, **V***, and **Σ** are real *m* × *m* matrices as well. **Σ** can be regarded as a scaling matrix, and **U**, **V*** can be viewed as rotation matrices.



$$M = U \cdot \Sigma \cdot V^*$$

www.wikipedia.org

Goals of PCA

(1) Extract the most important information from the data table

→ PC1 should describe the direction along which the data contains the largest variance; PC2 is orthogonal to PC1 and describes the direction of the largest remaining variance etc

(1) Compress the size of the data set by keeping only this important information

(2) Simplify the description of the data set

(3) Analyze the structure of the observation and the variables.

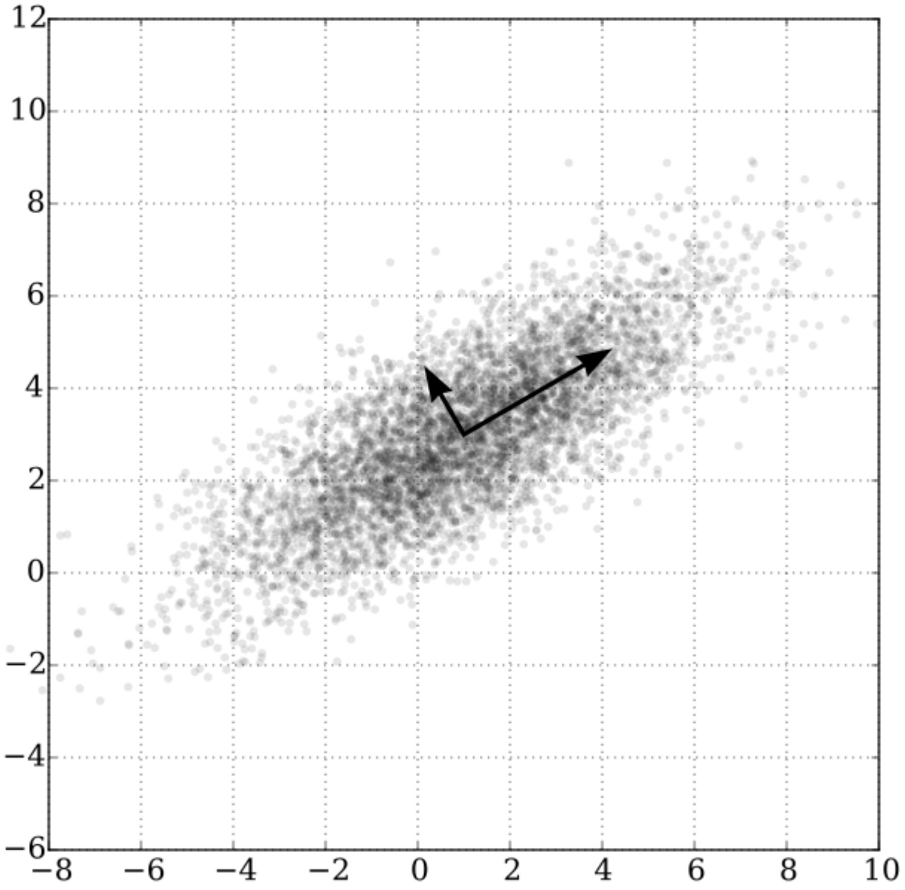
In order to achieve these goals, PCA computes new variables called principal components (PCs) as linear combinations of the original variables.

PC1 is the eigenvector of $\mathbf{X}^T\mathbf{X}$ with largest eigenvalue etc.

PCA example

PCA of a multivariate Gaussian distribution $\mathbf{X}$ centered at (1,3) with a standard deviation of 3 in roughly the (0.866, 0.5) direction and of 1 in the orthogonal direction.

The two PCA vectors shown are the eigenvectors of the covariance matrix $\mathbf{X}^T\mathbf{X}$ scaled by the square root of the corresponding eigenvalue, and shifted so that their tails are at the mean.



www.wikipedia.org

Note that shown here is the data along the original coordinates.
In a PCA plot, the data is projected onto two PCs, usually PC1 and PC2.

Deriving the components

The principal components are obtained from the SVD of **X**,

$$\mathbf{X} = \mathbf{P}\Delta\mathbf{Q}^T$$

Q contains the principal components (normalized eigenvectors of $\mathbf{X}^T\mathbf{X}$).

The *I* x *L* matrix of **factor scores**, denoted **F**, is obtained as

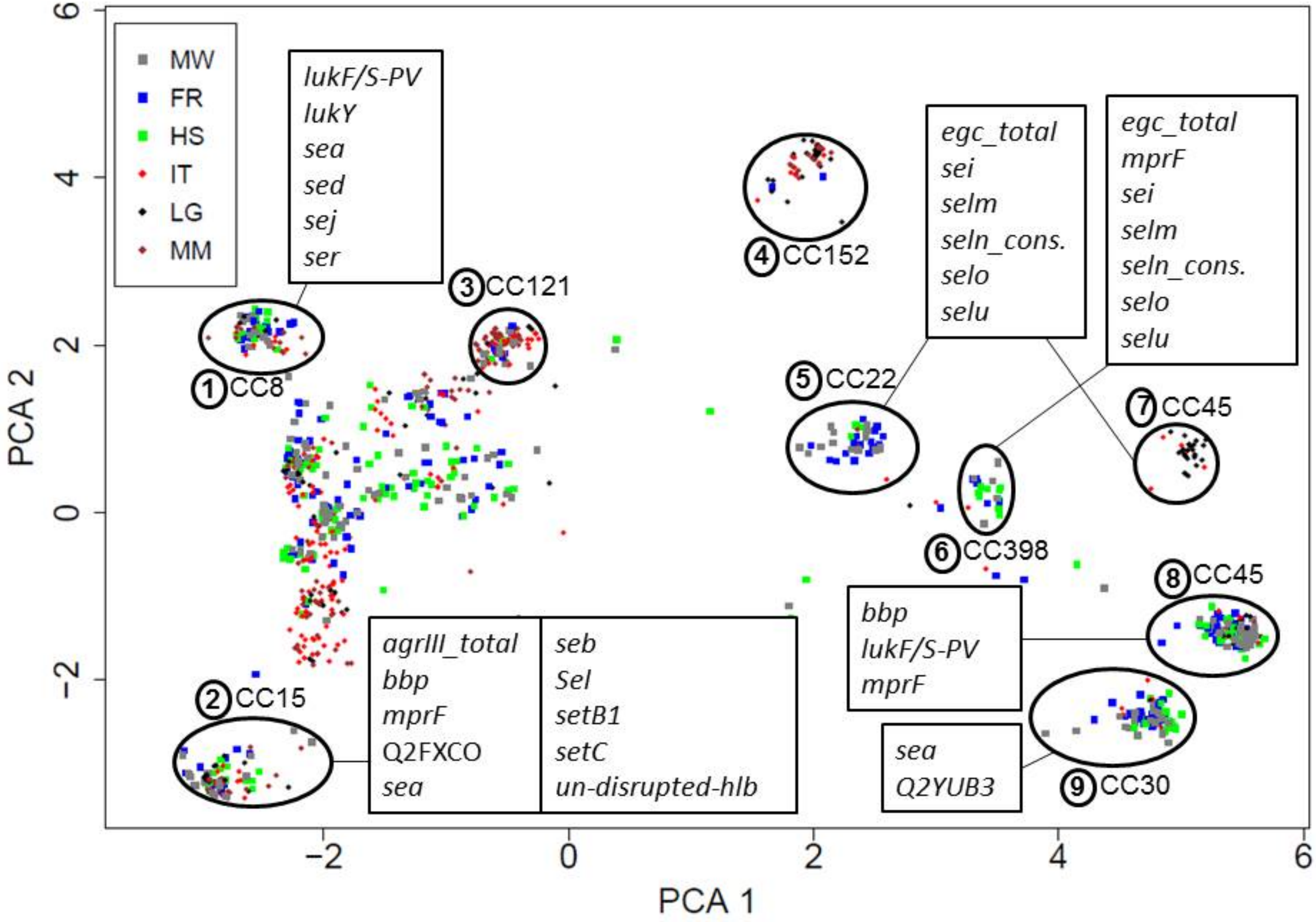
$$\mathbf{F} = \mathbf{P}\Delta = \mathbf{P}\Delta\mathbf{Q}^T\mathbf{Q} = \mathbf{X}\mathbf{Q}$$

Thus, **F** can be interpreted as a **projection matrix** because multiplying **X** with **Q** gives the values of the projections of the observations **X** on the principal components **Q**.

PCA of MA hybridization data (again)

PCA identifies local clusters that are characteristic

for particular clonal complexes



Projection (factor score) of data points on PC1

Summary

What we have covered **today**:

- Detection of DNA probes by DNA microarray
- Euclidian distance of 1/0 signals as distance measure
- Clustering of MA data
- PCA analysis of MA data

Next lecture:

- Reconstruct missing (ambiguous) data values with BEclear