# Strain/species identification in metagenomes using genome-specific markers.

Tu, He and Zhou. 2014 *Nucleic Acids Research*

Journal Club
Triinu Kõressaar
25.04.2014

# Introduction (1/2)

Shotgun metagenome sequencing– relatively cheap, fast and high-throughput

The characterization and identification of known microorganisms at strain/species level

Lack of high-resolution tools

# Introduction (2/2)

Basic methods for analysing metagenomic samples:
16S rDNA sequencing – low resolution

Sequencing all DNA in a sample – BLAST reads against database (of genomes or clade-specific genes)
or assembling reads into contigs

Current method – k-mer based approach – GSMer – Genome-Specific Markers

# Data resources

Both finished and draft sequences from:
NCBI GenBank
Human Microbiome Project Data Analyst and Coordination Center (HMPDACC)
In total 5390 strains

Four mock community metagenomes (21 bacterial strains representing 18 genera, two even mock, two staggered mock communities) from NCBI SRA.

Finished genomes: JGI IMG Web site, 302 genomes

Human genome

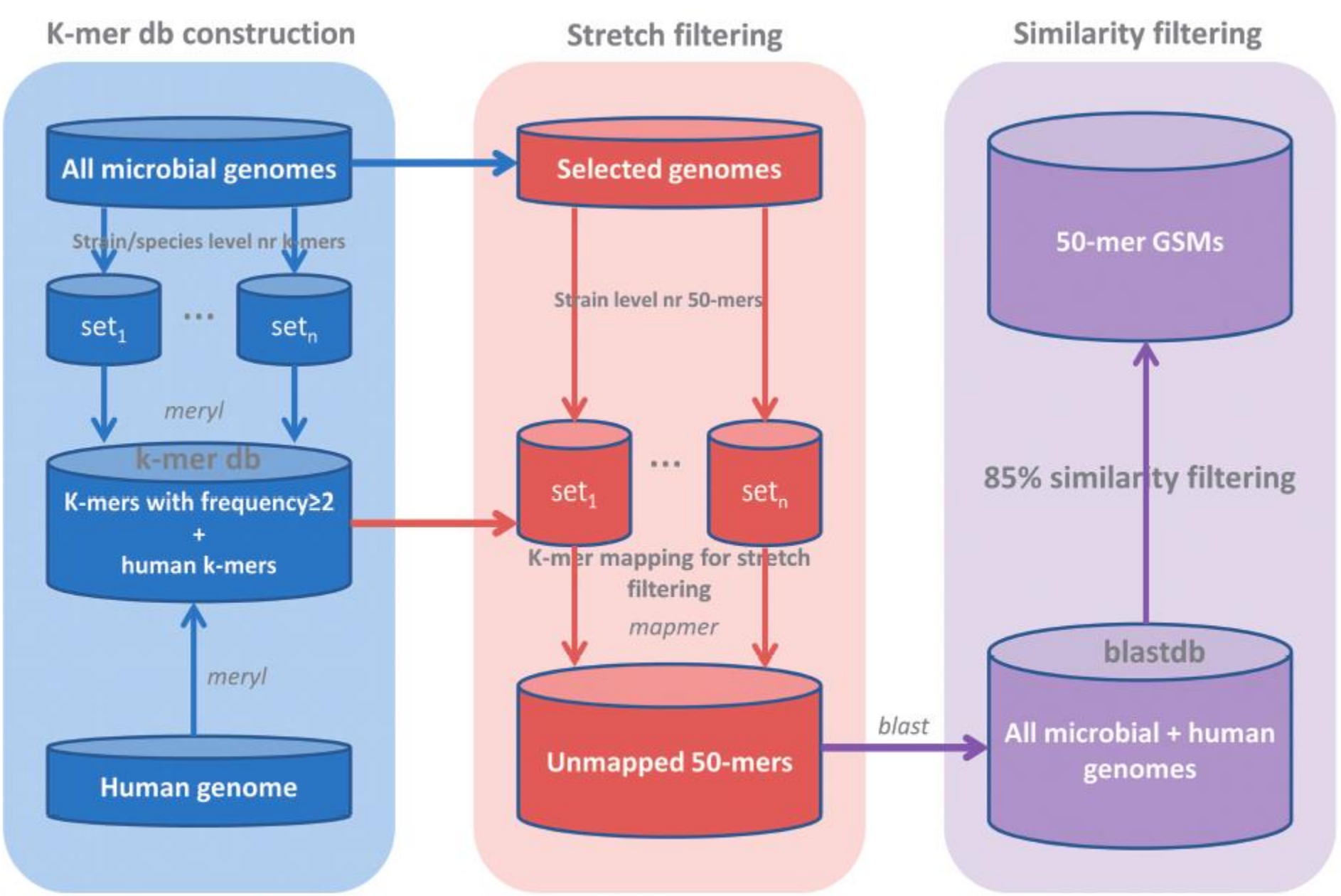



**Figure 1.** Flowchart of GSM identification processes. First, *k*-mer database (db) construction. *K*-mer db representing k-mers that show up in two o more microbial strains and all human genome *k*-mers were constructed by meryl program. *K*-mer sizes from 18 to 20 were selected. Second, 50-me GSMs were generated for selected strains/species. GSMs were then mapped with the *k*-mer db, and mapped GSMs were filtered. Third, all GSM were searched against all microbial genomes by BLAST, and GSMs having 85% identity with non-target GSMs were also filtered.

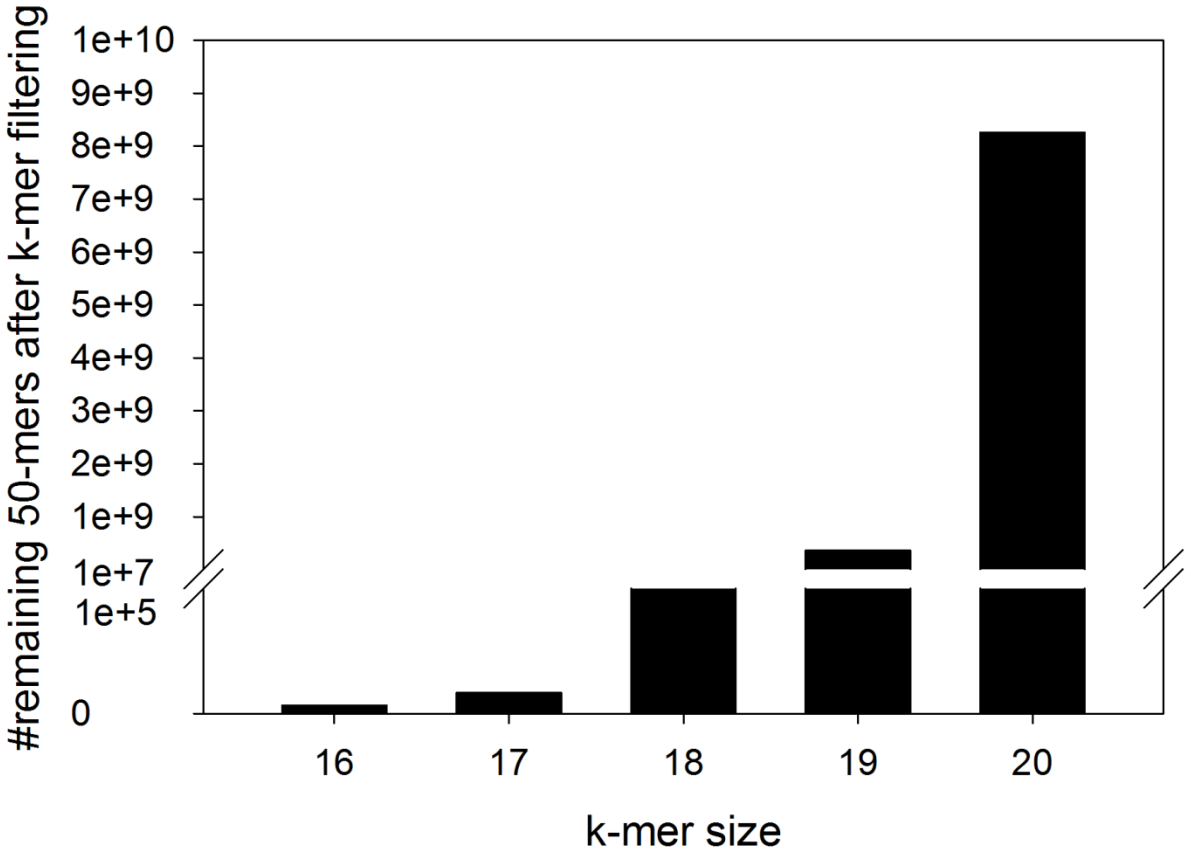



**Fig. S1** Number of candidate GSMs when different k-mer sizes were used for continuous stretch filtering.

# Results

5,390 microbial strains – 4,088 strains could have >=50 strain-specific GSM-s identified

2,548 – 18mer
1,161 – 19mer
384 – 20mer

A total 8,770,321 strain-specific GSMs – 68.5% located within genes, 18.9% within intergenic regions, 9.8% overlapped between gene and intergenic regions, 2.7% were from unannotaed genomes

2005 species (4933 strains) - 11,736,360 GSMs
1,872 – 18mer
198 – 19mer
48 -20mer

63% within genes, 17.2% within intergenic regions, 8.8% overlapped between a gene and intergenic regions, 11% unannotated genomes

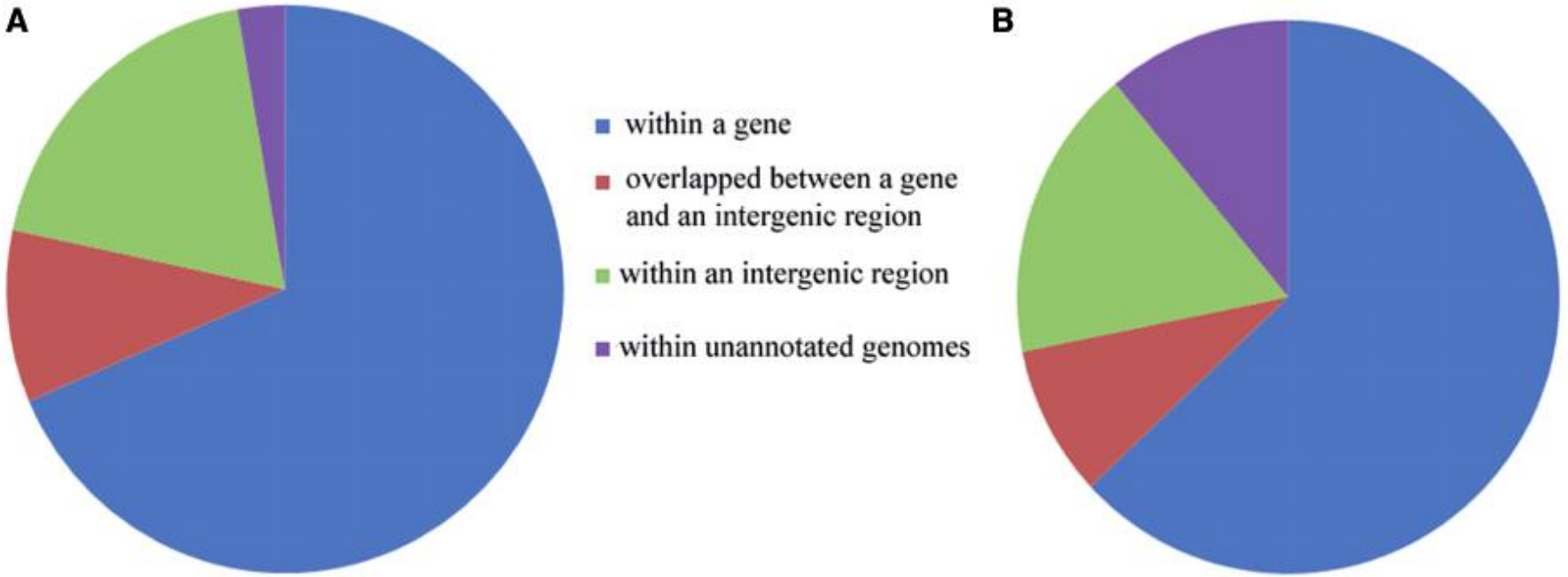



**Figure 2.** Location of the identified GSMs in the genome. (**A**) strain-specific GSMs; (**B**) species-specific GSMs. Different colors denote different locations in the genome: blue for GSMs within genes, green for GSMs within intergenic regions, red for GSMs overlapped between a gene and an intergenic region and purple for unannotated genomes.

# Specificity analysis

MEGABLAST – only perfect matches

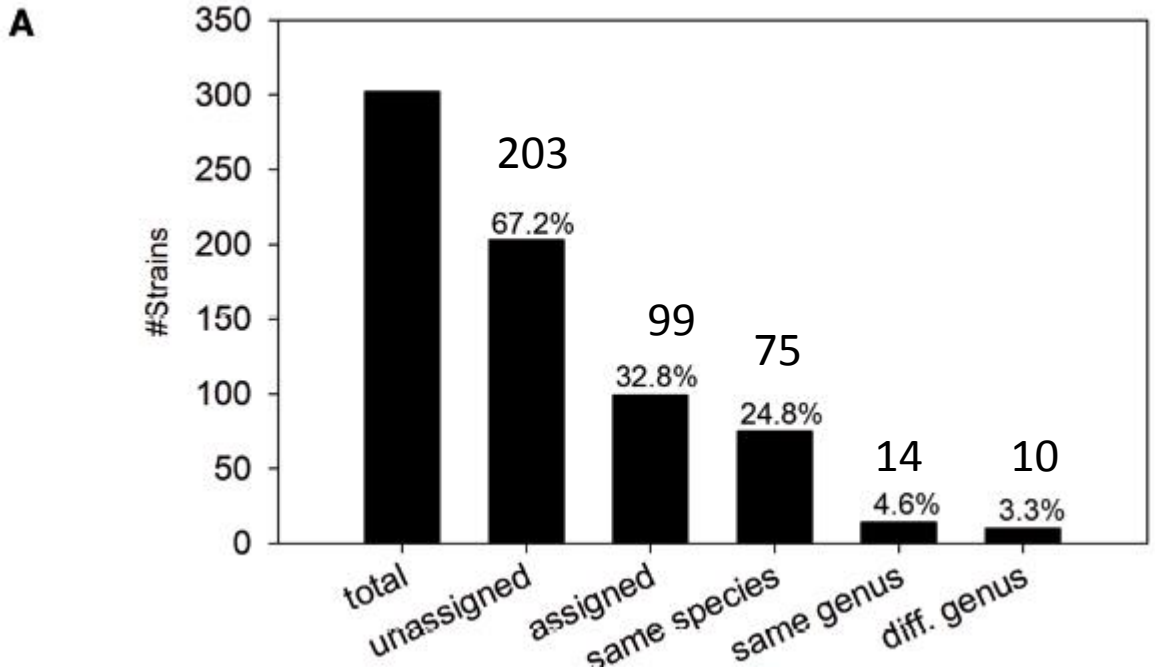
*Even mock microbial community* (16 species had GSMs available) - all 16 species were identified, no false positives.

*Staggered mock community* – 12 and 14 true-positive findings.

False negative identification was due to the low coverage of these strains

False positive (3 FP in one set) findings were because of only one mapped read for each strain

False-positive results could be removed when a cutoff of identified reads number and/or mapped GSM number were used.




Specificity evaluation against recently sequenced genomes (302 finished genomes)

# Sensitivity, detection limit

Two major questions:

At what sequencing coverage could the microbial genome be identified by GSMs?

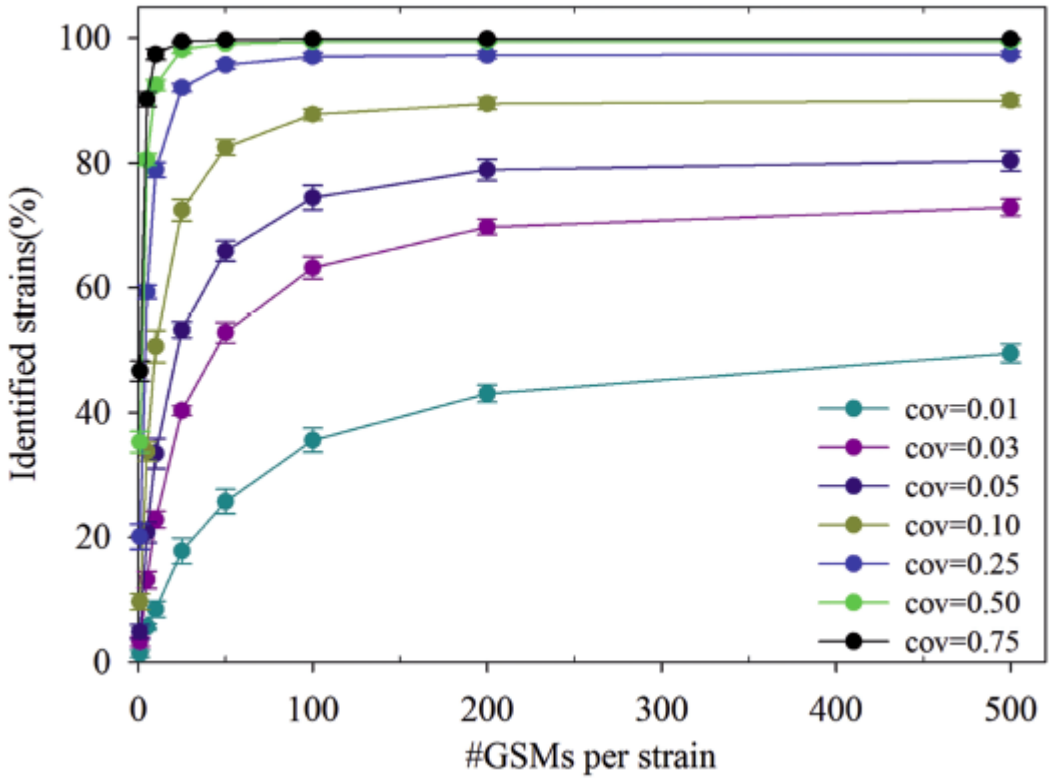
How many GSMs are required for effective identification of microbial strains/species?

Simulated metagenomes targeted 695 gut microbial genomes (generated by Grinder program), cov. 0.01-0.75, PE 100bp, 1,5,10,25,50,100,200 and 500 GSMs per strain.

# Sensitivity, detection limit

Minimum number of GSMs per strain for low coverage (<=0.25x) sequence data is 100

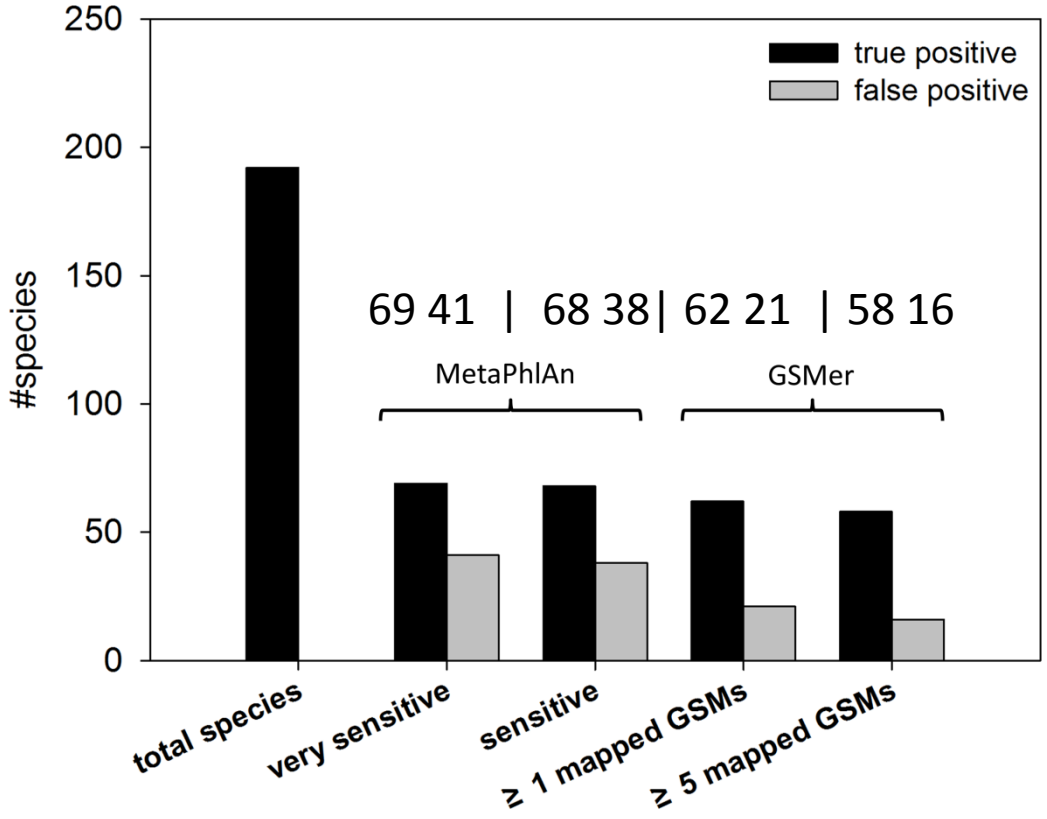
for higher coverage >0.25x 50 GSMs per genome is required




body site. (C) Sensitivity evaluation of GSMs using simulated metagenomes from 695 guts microbial strains. Simulated metagenomes at seven different coverages (0.01, 0.03, 0.05, 0.1, 0.25, 0.5 and 0.75) were searched against different number of GSMs per strain (1, 5, 10, 25, 50, 100, 200 and 500). The percentages of identified microbial strains were analyzed.

# Comparision with other approaches

MetaPhlAn (Metagenomic Phylogenetic Analysis) - maps reads against a reduced set of clade-specific marker sequences




- **Fig. S3** Comparison with MetaPhlAn at species level using synthetic metagenomes generated from 302 recently sequenced microbial genomes (192 microbial species).

# Discussion

Unique 50mers – shorter –mers could be more sensitive?
some specific 50mers could be rejected? (false-positive/negative rate)

Using fixed number of GSMs – GSMs are not guaranteed to be located over genome evenly

How many genomes had GSMs only within intergenic regions, which species.

Background database – could be more variable and larger, cannot be limited with genomes only

All bacteria are analysed together – which bacteria GSMs could'nt be found – do they have medical importance, how many of them has medical importance.

Many GSMs are located within intergenic regions – good or bad?

# Conclusions

GSMer - K-mer based approach :

•direct, rapid and accurate identification of microorganisms at the strain/species level from NGS data

• reduses the searching database ~0.05% of the whole genomes

• minimizes the noise in strain-level microbial identification

• dealing with sequencing errors  can be avoided when using many short GSMs

• 50mer is shorter than NGS reads length

Minimum of 50 GSMs per strain and 10% cutoff for mapped GSMs shall be used for positive callings for most microbial strains at >=0.25x sequencing coverage

By integrating GSM database with NGS platforms, instant detection of microbial strains/species is possible

Kraken