# SIGMA

## Strain-level Identification of Genomes from Metagenomic Analysis for Biosurvillence

Tae-Hyuk Ahn, Juanjuan Chai and Chongle Pan, 2014

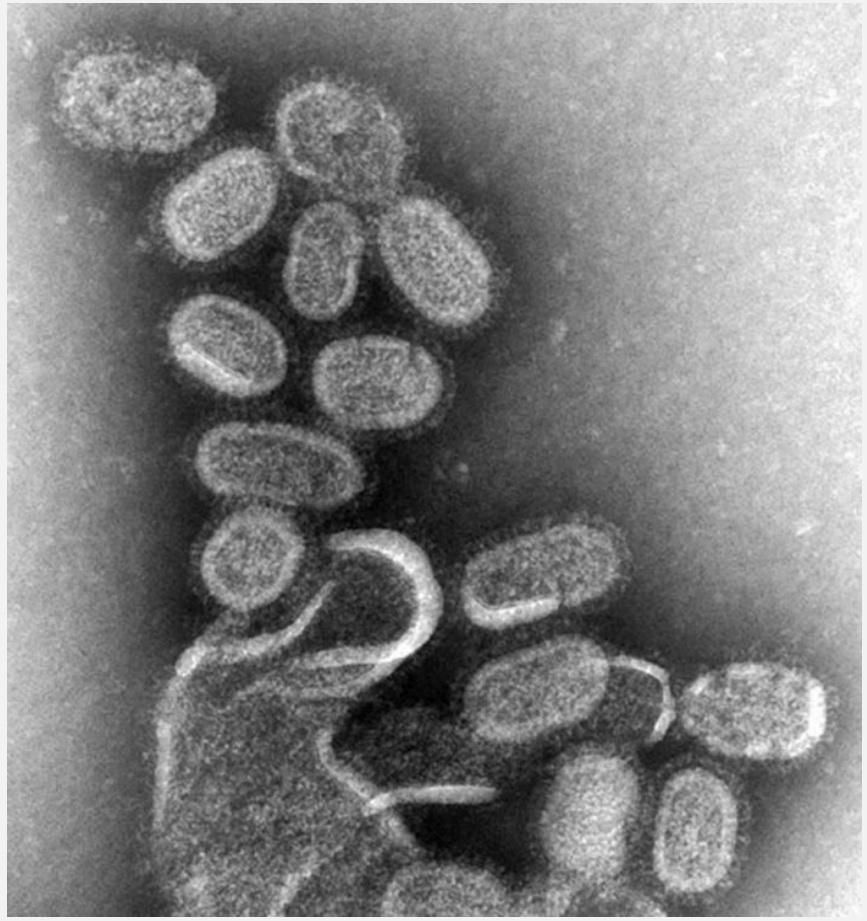
Presenter - Alon Tzur

# Outline

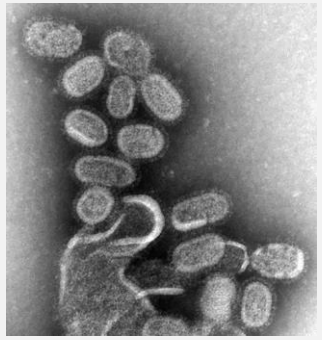
- The problems
- SIGMA
- Results
- Conclusion
- Discussion

# The problem

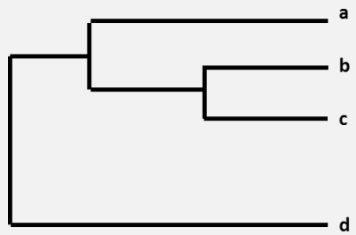
- Influenza virus



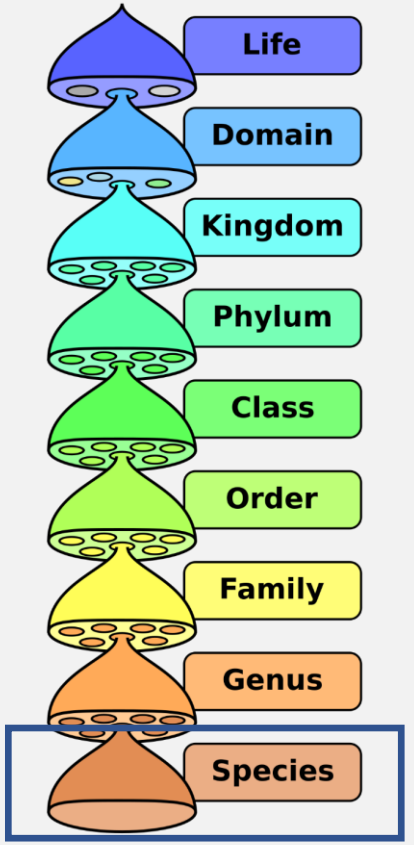
# The problem



- Influenza virus
- 4 Species: A, B, C, D
- We could imagine the phylogeny as simple as

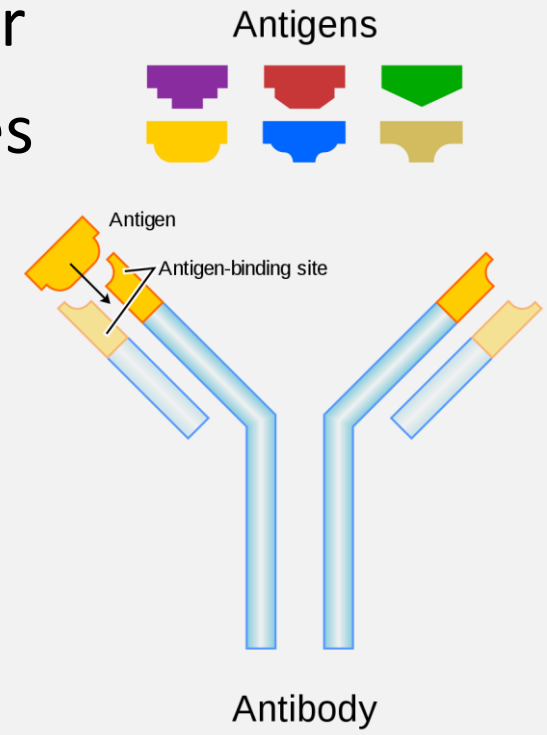



- But that’s not the case in real life



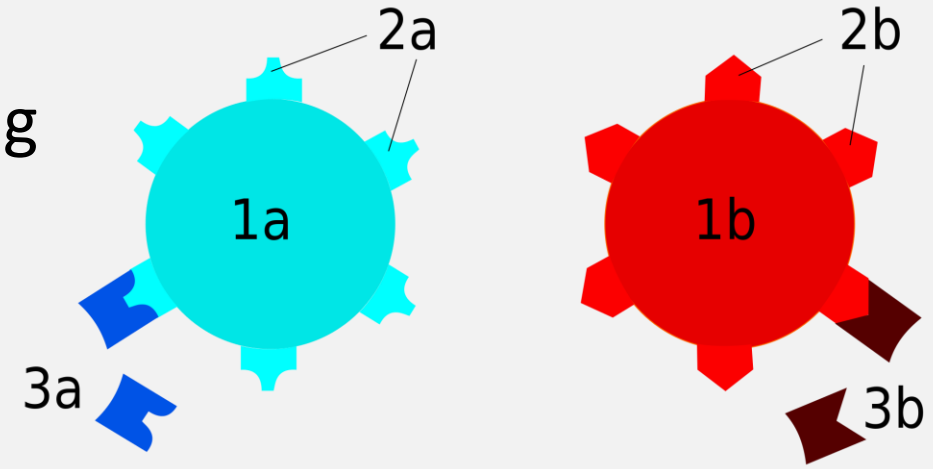
# The problem - Serotypes

- The human immune system “remembers” diseases
- When a pathogen is tracked, antibodies are produced and spread
- The antibodies are binding to antigens of the intruder
- We would expect the body to fight flu at most 4 times
- But that’s not the case…

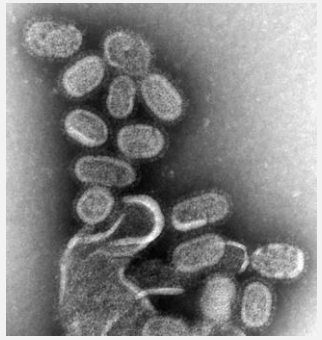


# The problem - Serotypes cont.

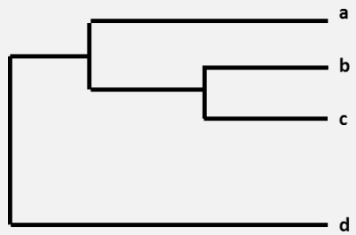
- **Serotype** is the collection of all antigens, of some **strain**.
- **Strain** is a sub-specie classification
  - Distinguished by some characteristics from other **strains**
- **Antibody** can deal with some **serotype**, but not with another
- **Strains/Serotypes** are the reason we have flu more than 4 times during lifetime
- They are the result of mutations and swapping of genetic components



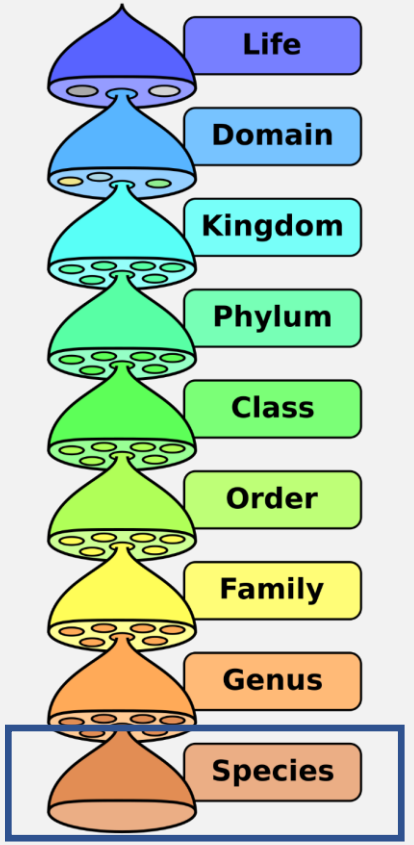
# The problem



- Influenza virus
- 4 Species: A, B, C, D
- We could imagine the phylogeny as simple as

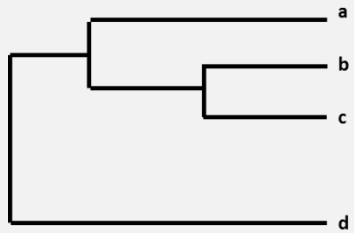



- But it’s more like

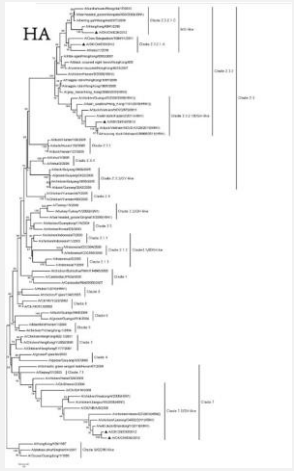


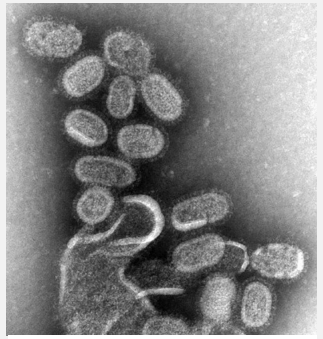
# The problem

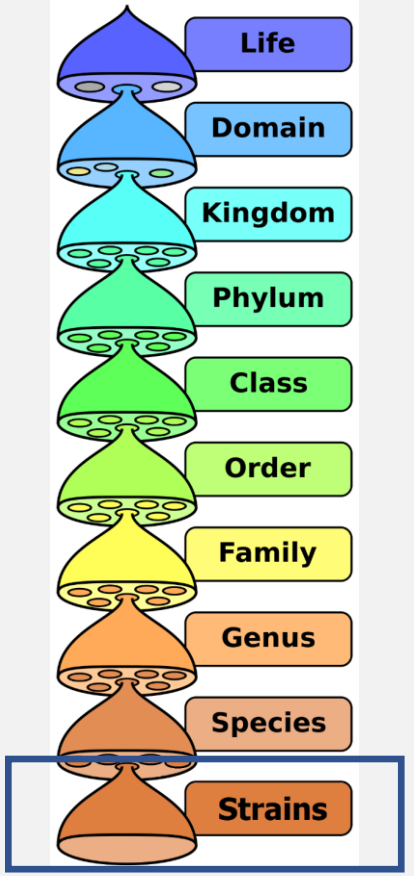
- Influenza virus
- 4 Species: A, B, C, D
  - Many strains: H1N1, H3N2, H1N2…
- We could imagine the phylogeny as simple as




- But it’s more like

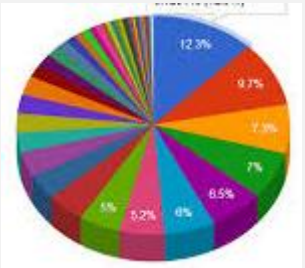


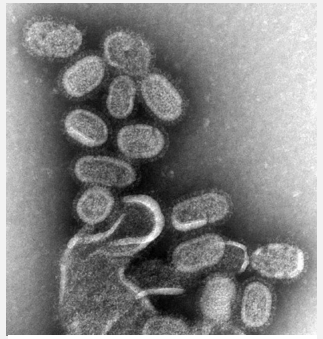



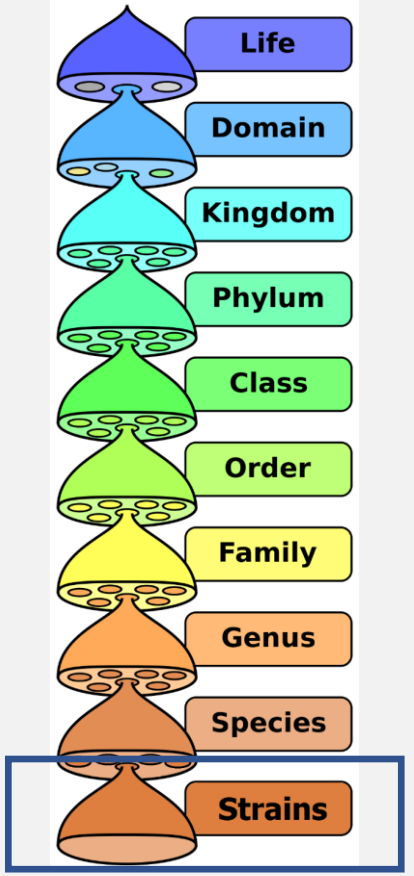


# The problem cont.

- **Biosurvillence** is the domain of predicting and reacting to pandemics
- The methods we have seen until now, show us the species distribution within a metagenomic sample
- We need to map the sample in a **strain-level**, meaning - higher resolution than just species

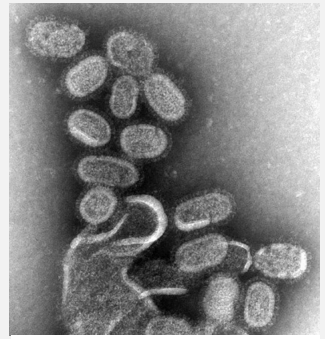


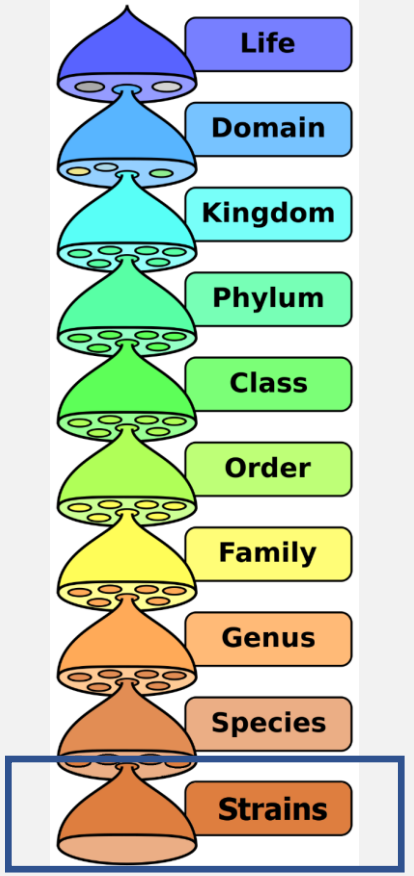




# The problem cont.

- **Example** - In order to determine if water are safe, it is tested for several known pathogens, including E. coli
- E. coli is a species
- Most of E. coli strains are harmless to human
- Therefore, high abundance of E. coli doesn't necessarily imply that the water are polluted



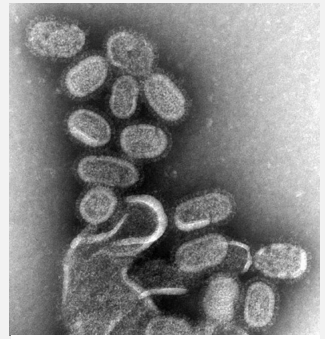


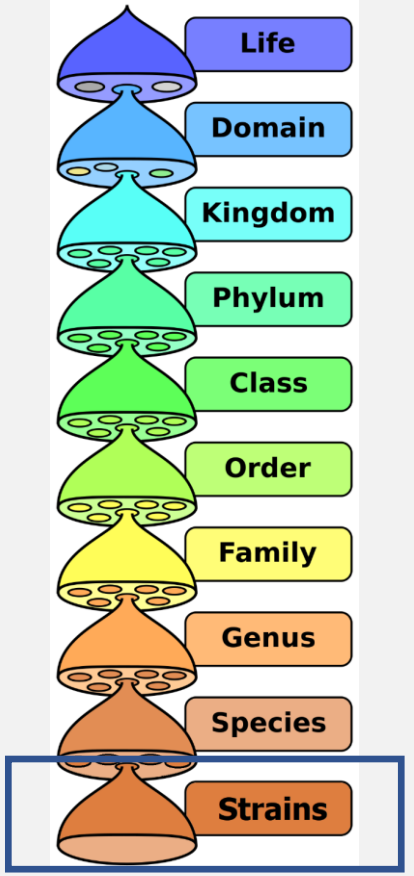
# The problem cont.

- The classic method to determine strains in a sample:
  - Isolation
  - Culturing
  - Genotyping

- That would consume time, and may lead to ambiguous results.

- In biosurvillence - the rapidness is crucial.



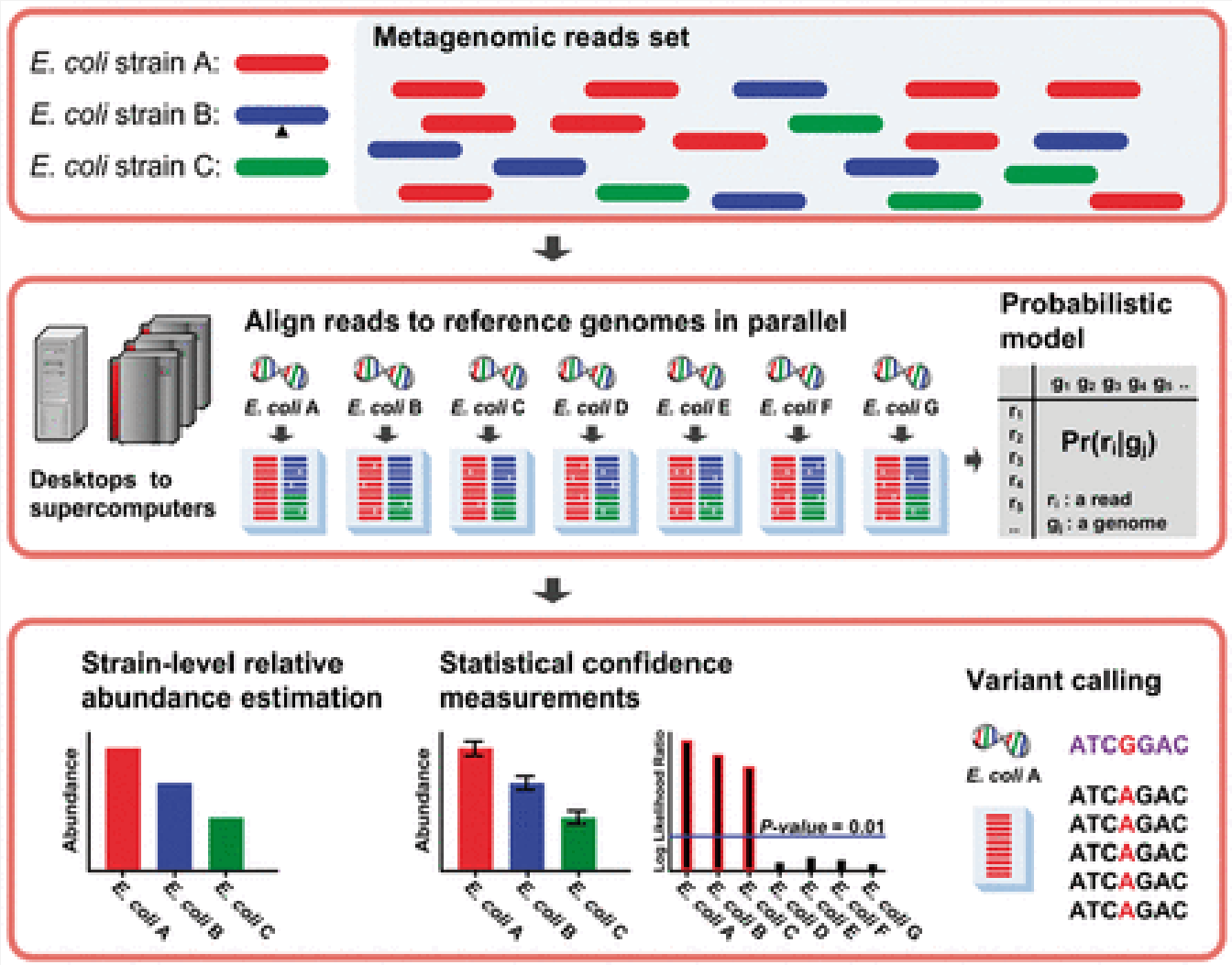


# The problem cont.

- There are several approaches for sensitive and specific identification of pathogens in metagenomic sample:
  - Clade specific marker gene approach - MetaPhlAn
    - Pro: Doesn’t require significant amount of computation
    - Con: Not sensitive enough to distinguish strain-level
  - Read mapping approach
    - Pro: Sensitive enough to distinguish strain-level, for referenced pathogens with variety of documented strains
    - Con: Can’t detect novel pathogens

# Outline

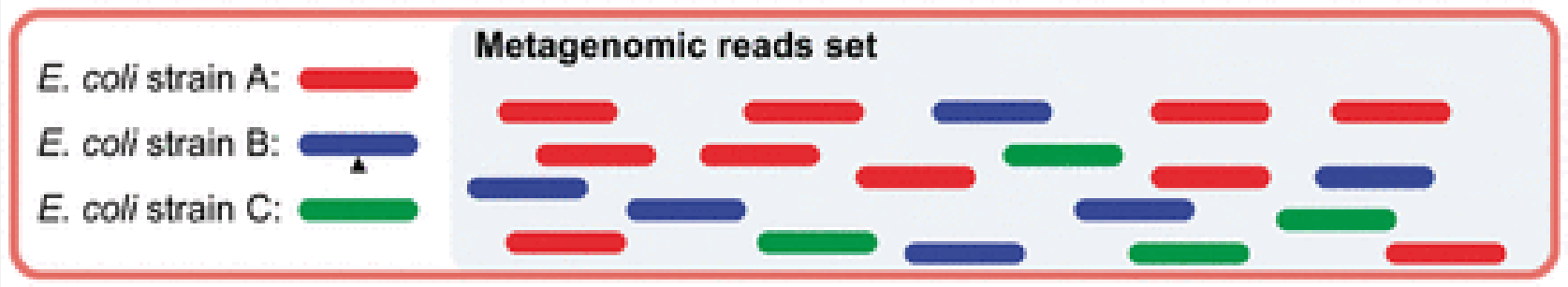
- The problems
- SIGMA
- Results
- Conclusion
- Discussion



# SIGMA

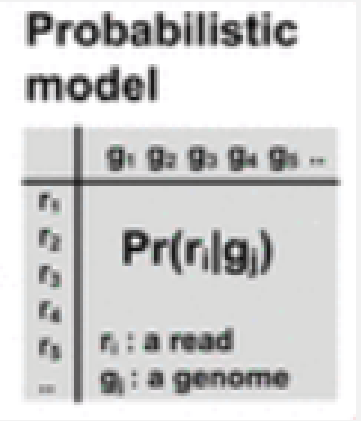
- Set U>0 as the upper bound of mismatches
  - Between some read and some reference genome

- Map the reads of the metagenomic dataset onto a **reference genomes database**



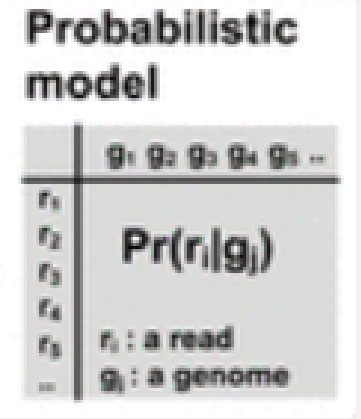
# SIGMA

- Denote:
  - $\sigma = 5\%$ - The uniform probability for a mismatch, hyper-parameter.
  - $r_i$ - the ith read
  - $g_j$ - the jth genome reference
  - $\mathrm{l_i} = |r_i|$ - The length of the ith read.
- Align every read $r_i$ with every genome reference $g_j$, find $z_{ij}$
  - $z_{ij}$ - The number of mismatches for the alignment of $r_i$ with $g_j$
- Calculate the Q matrix, where $Q_{ij} = \Pr(r_i | g_j)$
  - If $z_{ij} \leq U$, set $\Pr(r_i | g_j) = \sigma^{z_{ij}}(1-\sigma)^{l_i - z_{ij}}$
  - If $z_{ij} > U$, set $\Pr(r_i | g_j) = 0$



# SIGMA

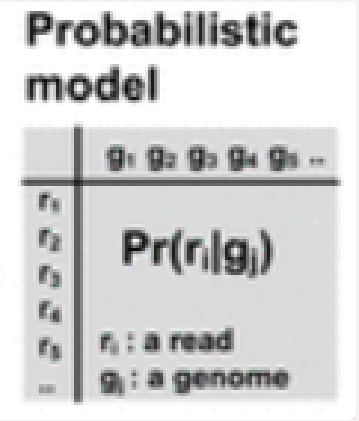
- Denote $G = (\Pr(g_1), \Pr(g_2), \dots)$
  - $\Pr(r_i, g_j) \overset{Conditional\ probability\ definition}{\triangleq} \overbrace{\Pr(r_i|g_j)}^{Q_{ij}} \cdot \Pr(g_j)$
    $= Q_{ij} \cdot \Pr(g_j)$
  - $\Pr(r_i) \overset{Marginal\ distribution}{\triangleq} \sum_j \Pr(r_i, g_j) = \sum_j Q_{ij} \Pr(g_j)$




- Reminder: we want to estimate $\Pr(g_j)$ for every j, those will estimate the relative abundance for every reference genome.

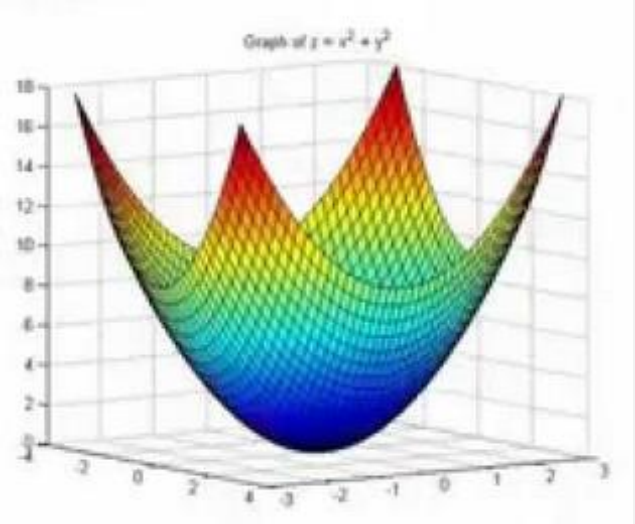
# SIGMA

- We will use Maximum-Likelihood Estimation to find $\Pr(g_j)$.
- We want to maximize the likelihood to see our reads, so:
  - $\max L(G; r_1, r_2, \dots, r_n) = \max \Pr(r_1, r_2, \dots, r_n)$
  - $\overset{Independence\ assumpation\ of\ r_i}{\widehat{=}}\ \max \prod_{i=1}^{n} Pr(r_i)$
  - $\overset{Assignment\ from\ previous\ page}{\widehat{=}}\ \max \prod_{i=1}^{n} \sum_j Q_{ij} \Pr(g_j)$
- We will perform log transformation on the output, and transform it to minimum objective function
  - $\min_{G \in \mathbb{R}^m} -\Sigma_{i=1}^{n} ln \sum_j Q_{ij} \Pr(g_j) = \min_{G \in \mathbb{R}^m} -\Sigma_{i=1}^{n} ln\,(< Q_i, G\ >)$, for G as distribution



# SIGMA – Objective function

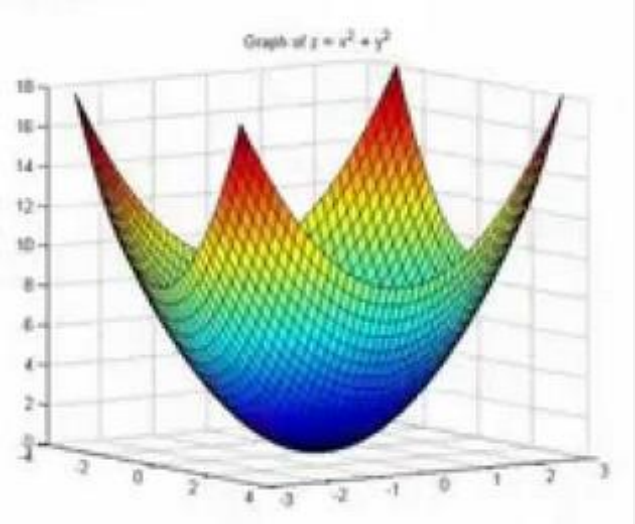
- Given the objective function $\min_{G \in \mathbb{R}^m} -\Sigma_{i=1}^{n} ln\,(< Q_i, G\ >)$
- The fact it's convex
- Our constraints (G is a distribution)
- We can use Non-Linear programming
- Specifically the primal-dual interior point method (Wachter and Biegler, 2006).



# SIGMA - NLP - Primal-dual interior point method

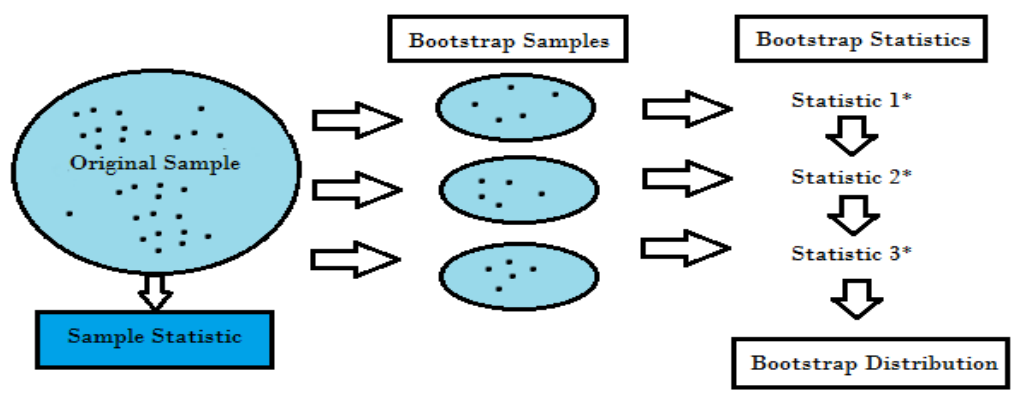
- The method can efficiently (*) find a solution for that optimization problems class:
  - The method can find a solution of constrained problem
  - The problems should satisfy the perturbed KKT conditions
  - The objective function must be convex

- (*) - Superlinear convergence rate



# SIGMA - Bootstraping

- SIGMA outputs confidence intervals for its finding.
- It does so by performing bootstrapping on the Q matrix rows, and run the NLP method many times, then taking a confidence interval for the relative abundance.
  - Bootstrapping - Synonym for just “draw oranges from a box with return”
- It is done B times, B=1000 by default. B is a hyperparameter.
- The result of the bootstrapping is B times $\tilde{Q}$ matrix, with the same dimensions as Q, but some rows may be duplicates, and others missing.



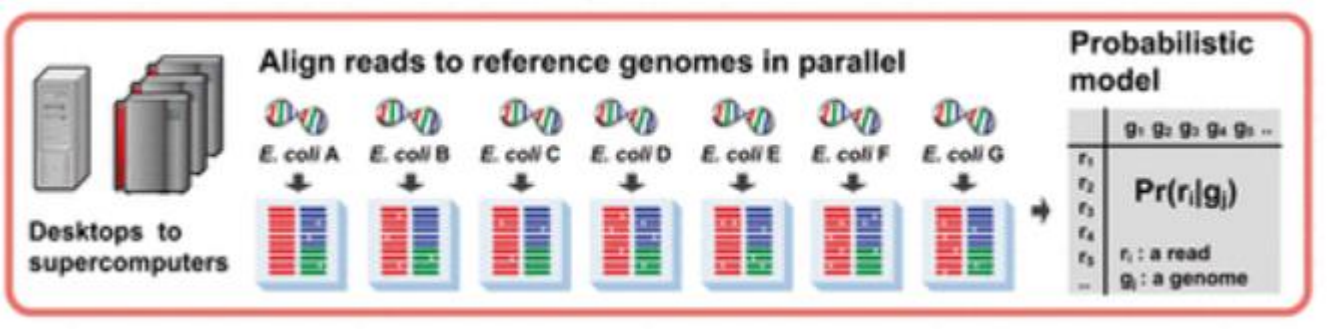
# SIGMA - Parallelization

Several SIGMA steps may be parallelized:

- Read alignments against reference genomes can be done independently
- Bootstrapping and running the NLP for each bootstrap can be done independently
- NLP evaluation step can also be parallelized

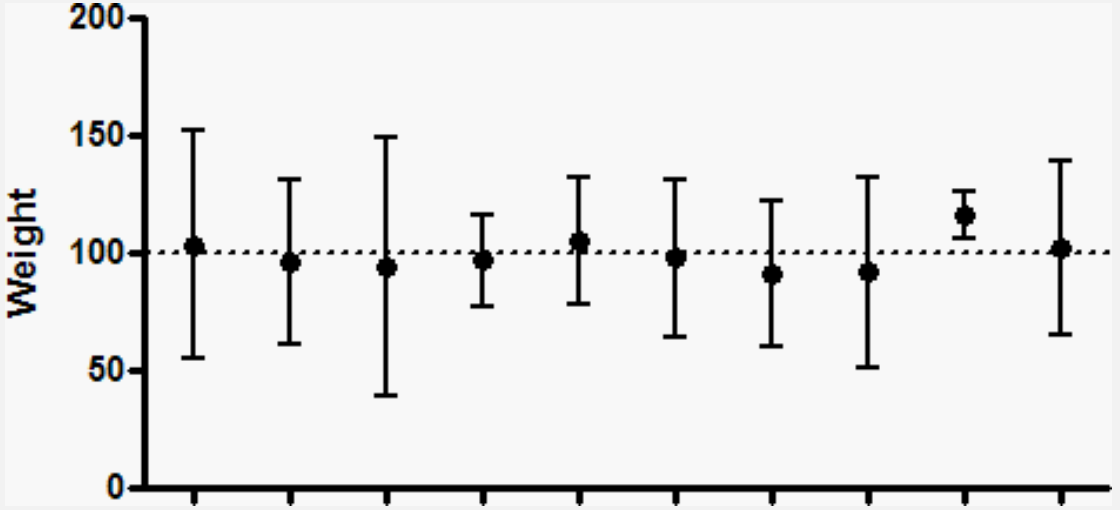
Therefore, SIGMA is suitable for running both on desktops and also on supercomputers and computer clusters.

The ability to parallelize SIGMA is important for biosurvillance, because performance and rapidness is crucial.



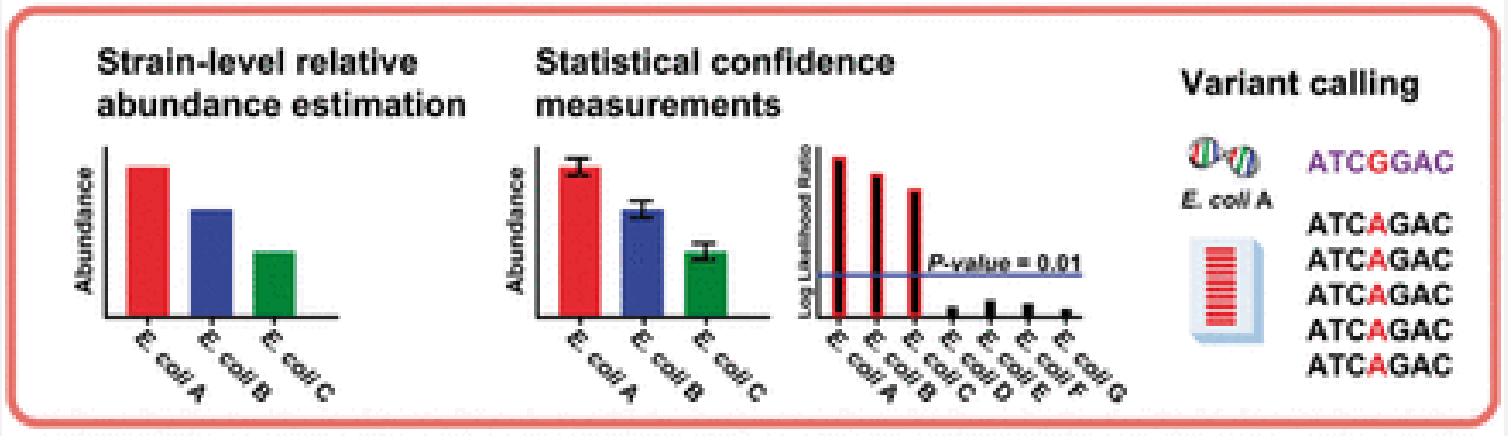
# SIGMA -Statistical view

- SIGMA will calculate **confidence intervals** for its findings.
- It will use the B times G distributions out of the bootstrapping process.
- It will use $1 - \alpha$ as the confidence level



# SIGMA -Statistical view

- **Hypothesis testing** - SIGMA calculates 2 MLEs for user-selected genomes:
  - Under null hypothesis: $\Pr(g_j) = 0$
    (reference genome column is dropped from Q matrix)
  - Under alternative hypothesis: $\Pr(g_j) \in (0,1]$
- Then it outputs the log likelihood ratio $-2\ln(\frac{L_0}{L_1})$
- That would help us to determine if the reference genome is likely to be present in the metagenomic sample



# Outline

- The problems
- SIGMA
- Results
- Conclusion
- Discussion

# Results

SIGMA was compared with 3 taxonomic classification algorithms:

- MEGAN
- Pathoscope
- MetaPhlAn

# Results - Genome identification

It was tested with the following samples

- 5 Genomes synthetic community, to test the taxonomic resolution of the classification.

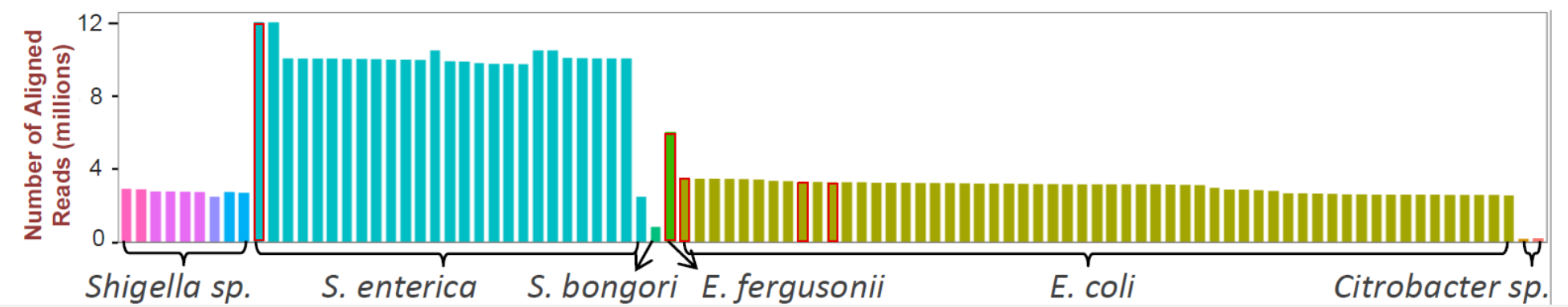
Reads simulated from:

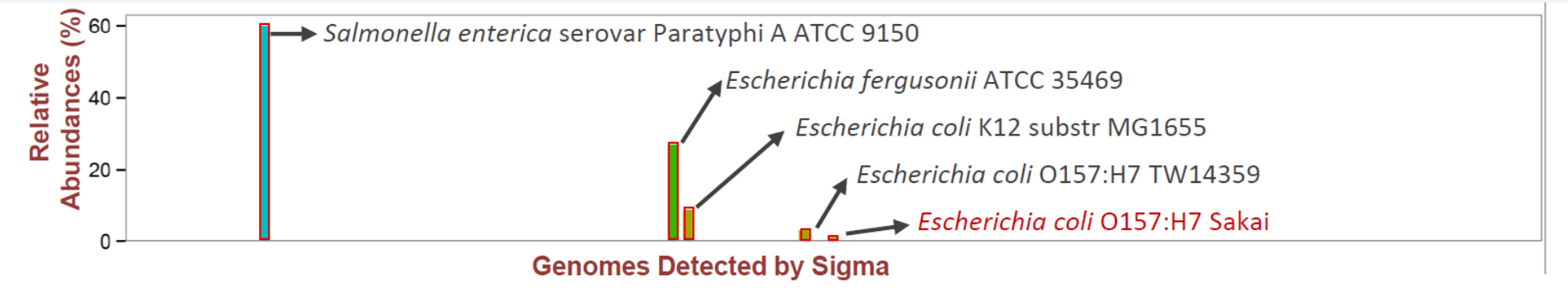
- E. coli O157:H7 Sakai - 1%
- E. coli O157:H7 TW14359 (Same **serotype** as the previous) - 3%
- E. coli K12 (Same **species** as the previous ones) - 9%
- E. fergusonii (Same **genus** as the previous ones) - 27%
- Salmonella enterica serovar Paratyphi A, strain ATCC9150 (Same **family** as the previous ones) - 60%

# Results - Genome identification

- The algorithms input was 20 million simulated reads
- The reference database supplied to the algorithms contained 2266 reference genomes
- The database contained 58 other E. coli strains

# Results - Genome identification



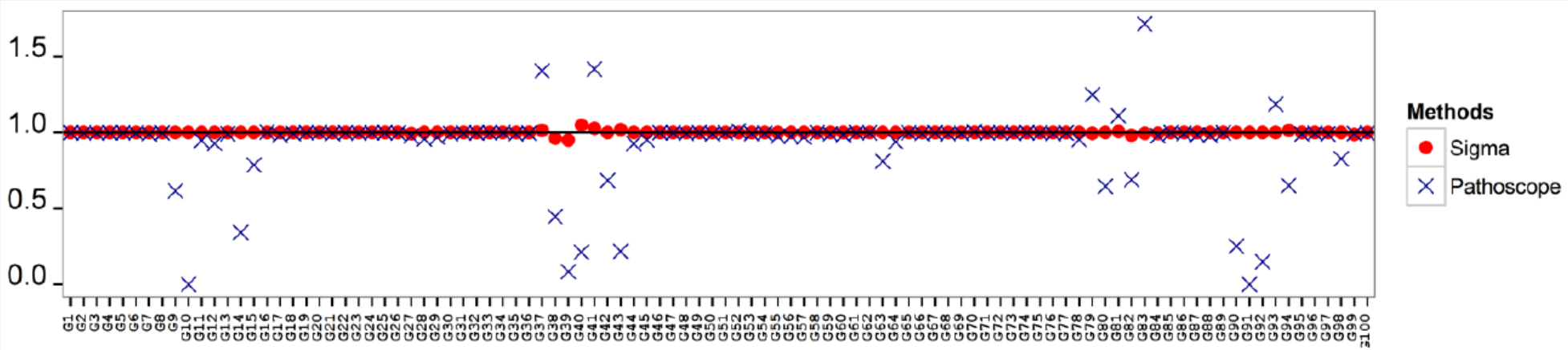
# Results - Genome identification

- SIGMA’s output relative abundance on the strain genome (Sakai) is 0.94% instead of ground truth of 1%
- Pathoscope’s output relative abundance is 0.018%
- MEGAN’s output relative abundance is 0.02%
- MetaPhlAn’s output is just 13.36% relative abundance of E. coli species

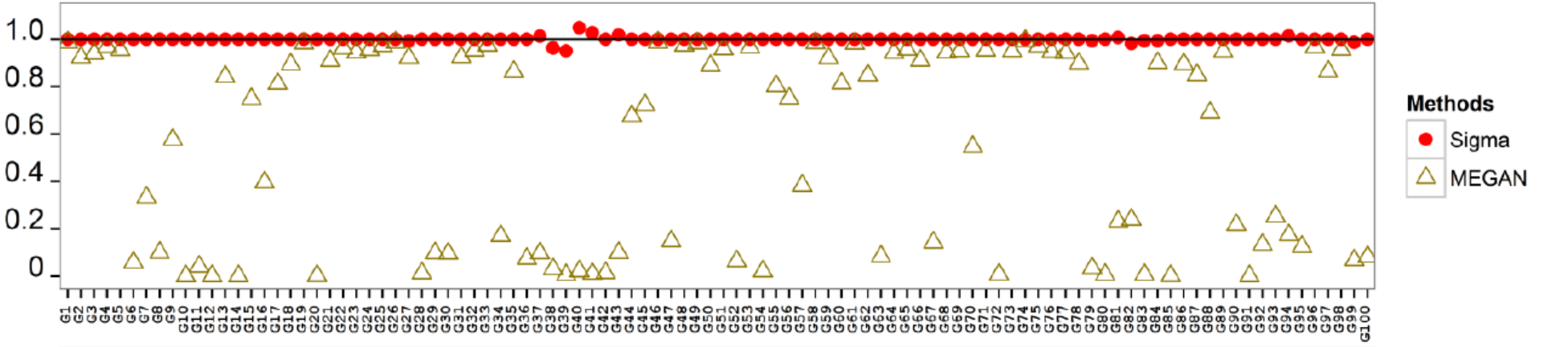
# Results - Quantification performance

- Another sample composed out of 100 diverse bacterial genomes.
- The community covered 35 classes.
- 100 million reads were simulated and sent to the 4 algorithms.

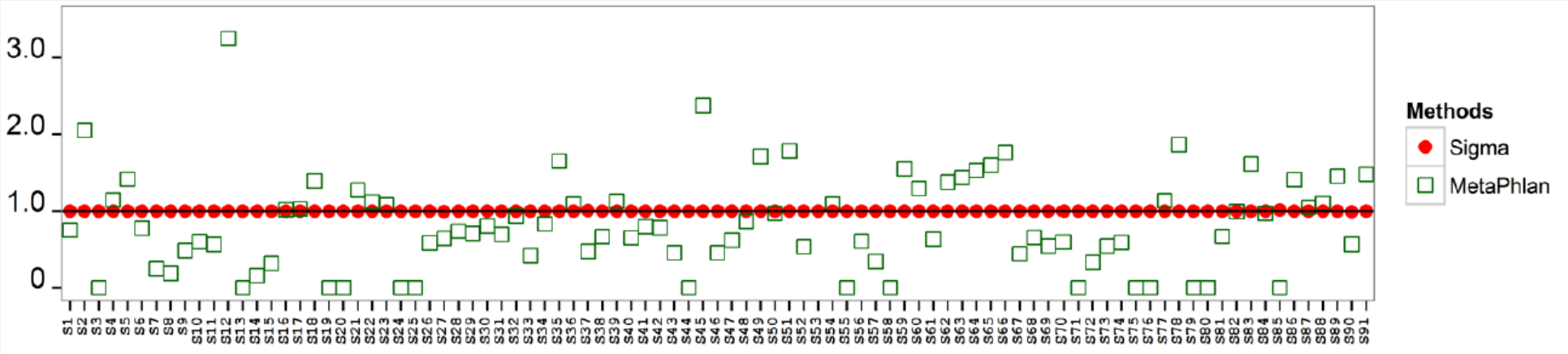
# Results - Quantification performance



# Results - Quantification performance



# Results - Quantification performance



# Results - Quantification performance

- SIGMA managed to predict the RA of **all** in a deviation of 95% - 105% of the expected RA.

# Results - Turnaround time performance

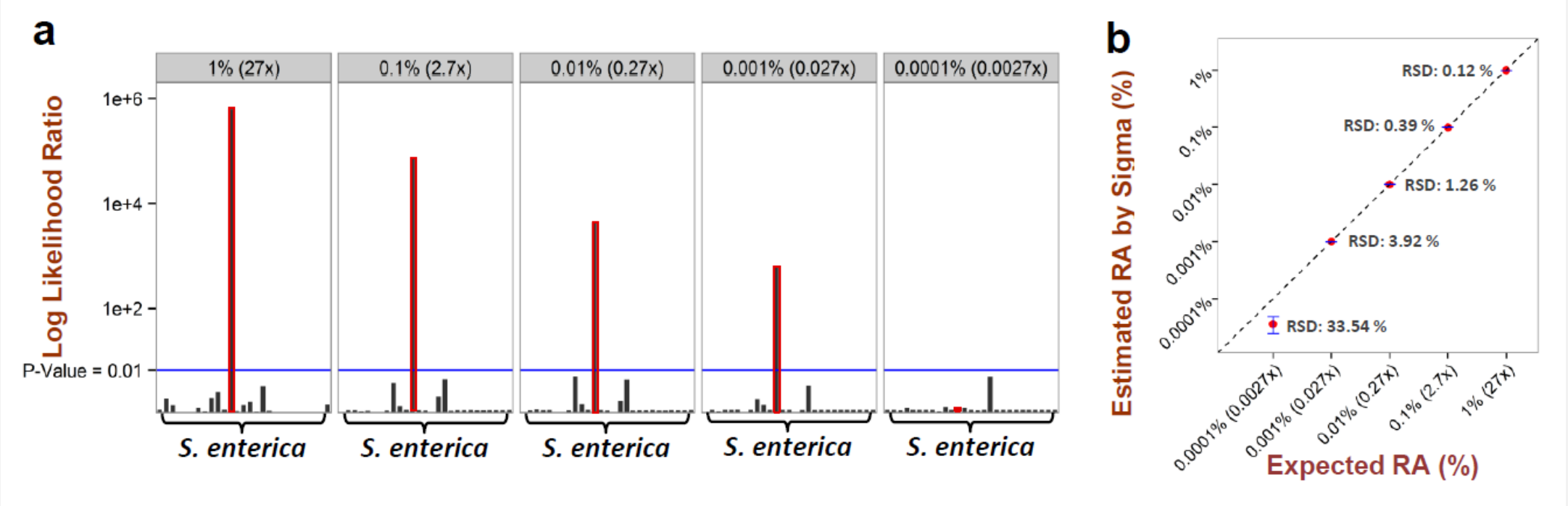
- The turnaround time test ran with the synthetic 5-genomes sample

| | Alignment | Abundance Estimation | | Total |
|---|---|---|---|---|
| | Wall-Clock Time (hr) | Wall-Clock Time (hr) | Memory (GB) | Wall-Clock Time (hr) |
| Sigma | 18 | 1 | 62 | 19 |
| Pathoscope | 70 | 13 | 118 | 83 |
| MEGAN | 70 | 12 | 93 | 82 |
| MetaPhlAn | N/A | 0.2 | 1 | 0.2 |

# Results - Statistical confidence assessment

- 125 million reads of fecal metagenome dataset were mapped to 24,994 reference genomes
- SIGMA identified 135 genomes
- No salmonella genome was found in this fecal sample.
- Varying amounts of simulated reads from S. enterica Paratyphi A strain ATCC9510 were spiked into the dataset, at several RA: 1%, 0.1%, 0.01%, 0.001% and 0.0001%
- The dataset contained 26 S. enterica other strains, which served as decoys

# Results - Statistical confidence assessment



# Results - Detection of nearest genomes and strain variations

- That dataset with 1% of S. enterica Paratyphi A strain ATCC9510 is also tested against several genome references
- Ref1 test was against a genome reference database where the strain was known
- In Ref2 the strain was unknown, but the species was known
- In Ref3 the species was unknown, but the genus was known
- In Ref4 the genus was unknown, but the family was known

# Results - Detection of nearest genomes and strain variations

| Test | Identified Genome | Relative Abundance (%) | Genome Coverage (%) | High-confidence SNPs |
|---|---|---|---|---|
| [1]Ref1 | *Salmonella enterica* serovar Paratyphi A ATCC 9150 | 0.988 | 100 | 0 |
| [2]Ref2 | *Salmonella enterica* serovar Paratyphi A AKU 12601 | 0.986 | 99.95 | 190 |
| [3]Ref3 | *Salmonella bongori* NCTC 12419 | 0.031 | 11.46 | 3750 |
| [4]Ref4 | None | | | |

# Outline

- The problems
- SIGMA
- Results
- Conclusion
- Discussion

# Conclusion

- SIGMA algorithm brings
  - High performance due to parallelization
  - Statistical confidence and uncertainty quantification
  - High resolution of strain level RA estimation
  - Novel strain identification, including SNPs from the closest found strain
  - Open source

- That allows SIGMA to be a good choice for biosurvillence and strains variants research.

# Outline

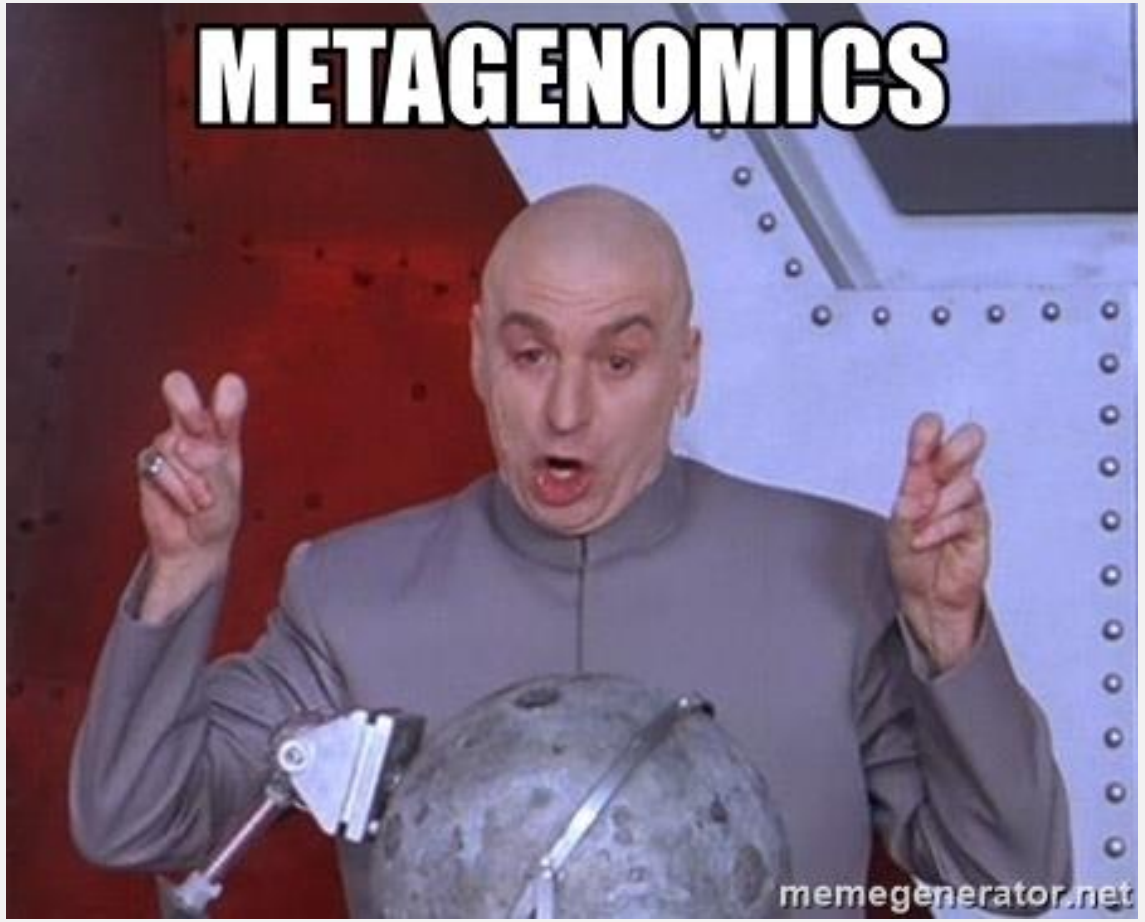
- The problems
- SIGMA
- Results
- Conclusion
- Discussion

# Discussion

- What have failed SIGMA algorithm from becoming a popular algorithm?
- Is a reference genome database based method restricting the metagenomic research?
- What is considered a good mismatch probability and how is it determined?
- What about species which couldn’t get cultured and sequenced?

# Questions?


BACTERIA
BACTERIA EVERYWHERE
memegenerator.net


METAGENOMICS
memegenerator.net