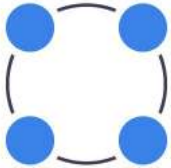


Applying cancer subpopulations analysis methods to metagenomic series

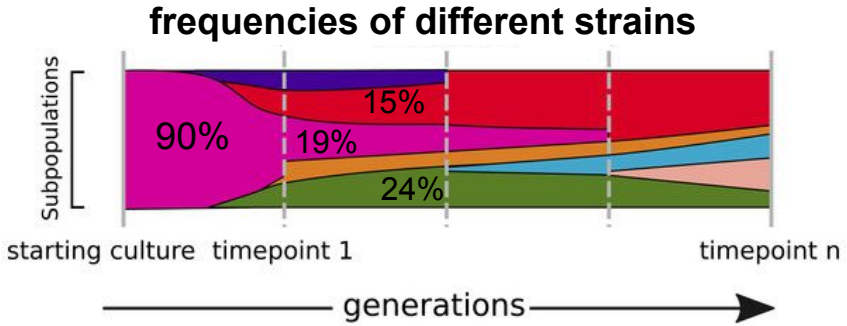
Maria Chernigovskaya
Margarita Akseshina

Scientific advisor: Sergey Nurk

Problem

Input:

- metagenomic series data
- reference genome



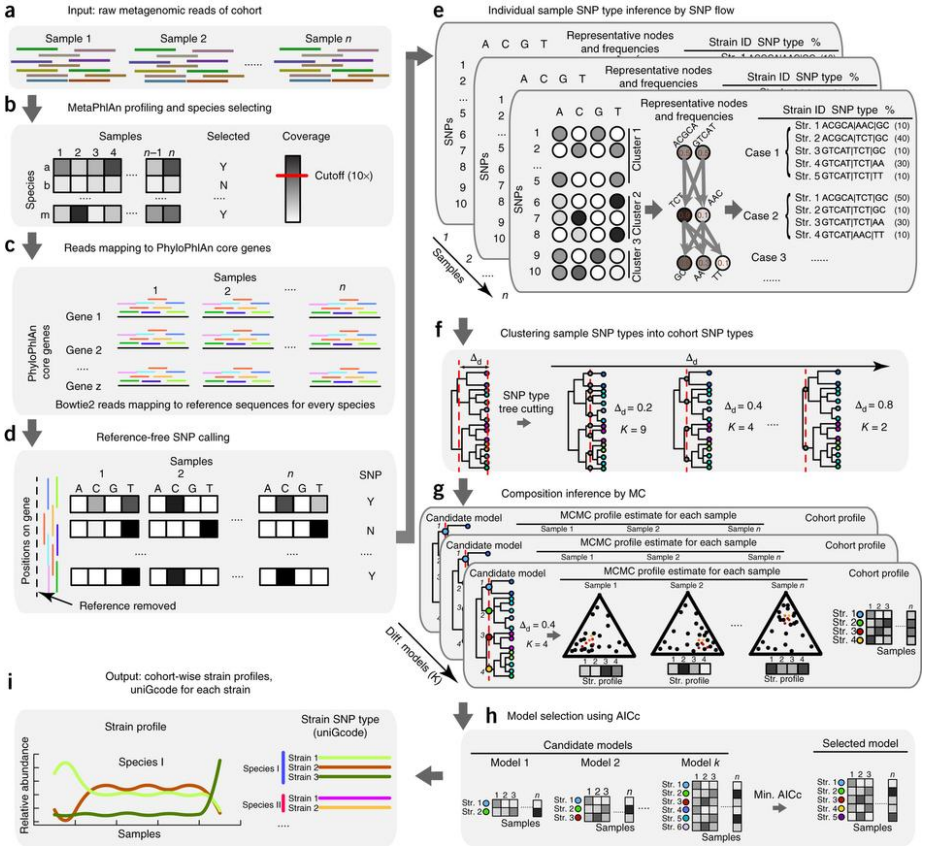
Output:

- the number of related strains
- their relative abundances across the samples

ConStrains [Luo et. al, 2015]

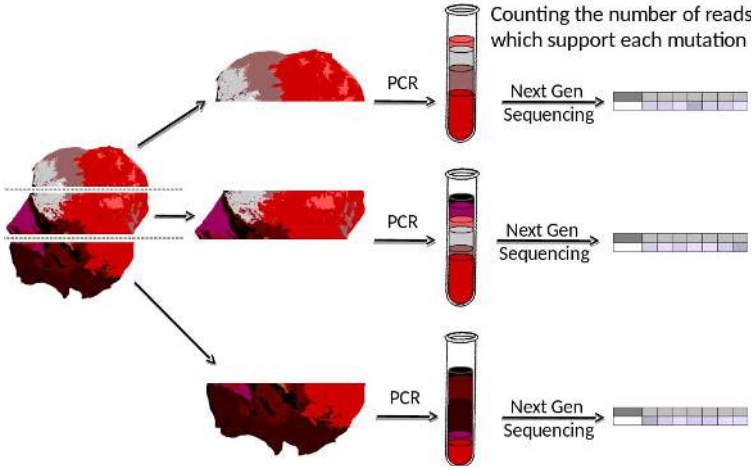
Questionable computational model!

Could not reproduce results from the paper



Subpopulation analysis in multiple tumor samples

- **Clomial** [Zare et al, 2014]
  - Matrix decomposition with EM
- PyClone [Roth et al, 2014]
- **PhyloWGS** [Deshwar et al, 2015]
  - MCMC on phylogenies
- AncesTree [Beerenwinkel et al, 2015]
- LICHeE [Popic et al, 2015]
- ...



General approach for *de novo* strain analysis

species reference

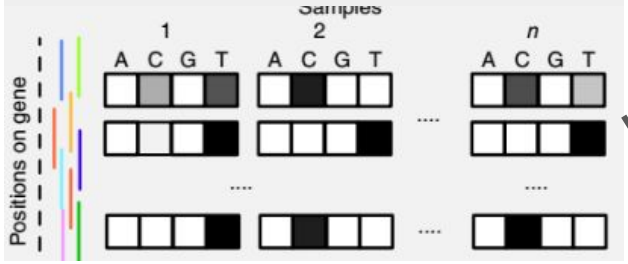
reads from sample 1

reads from sample 2

...

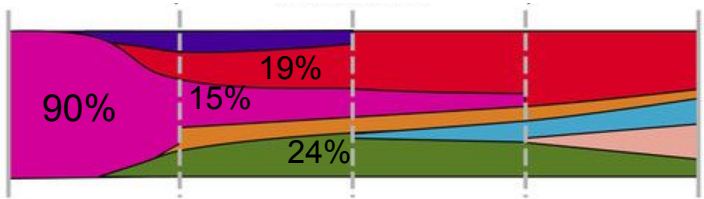
reads from sample N

SNPs:



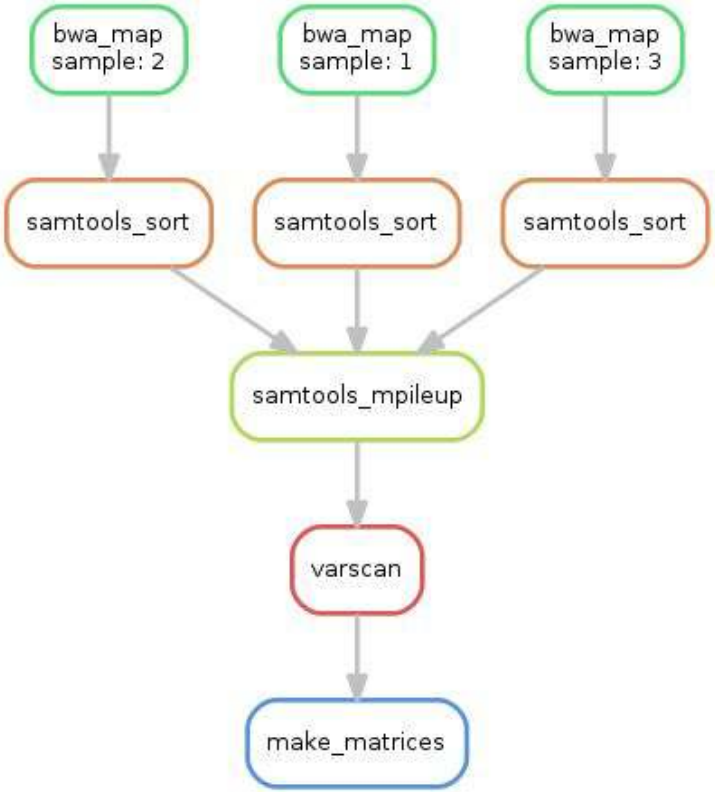
MAGIC

genotypes and frequencies of different strains



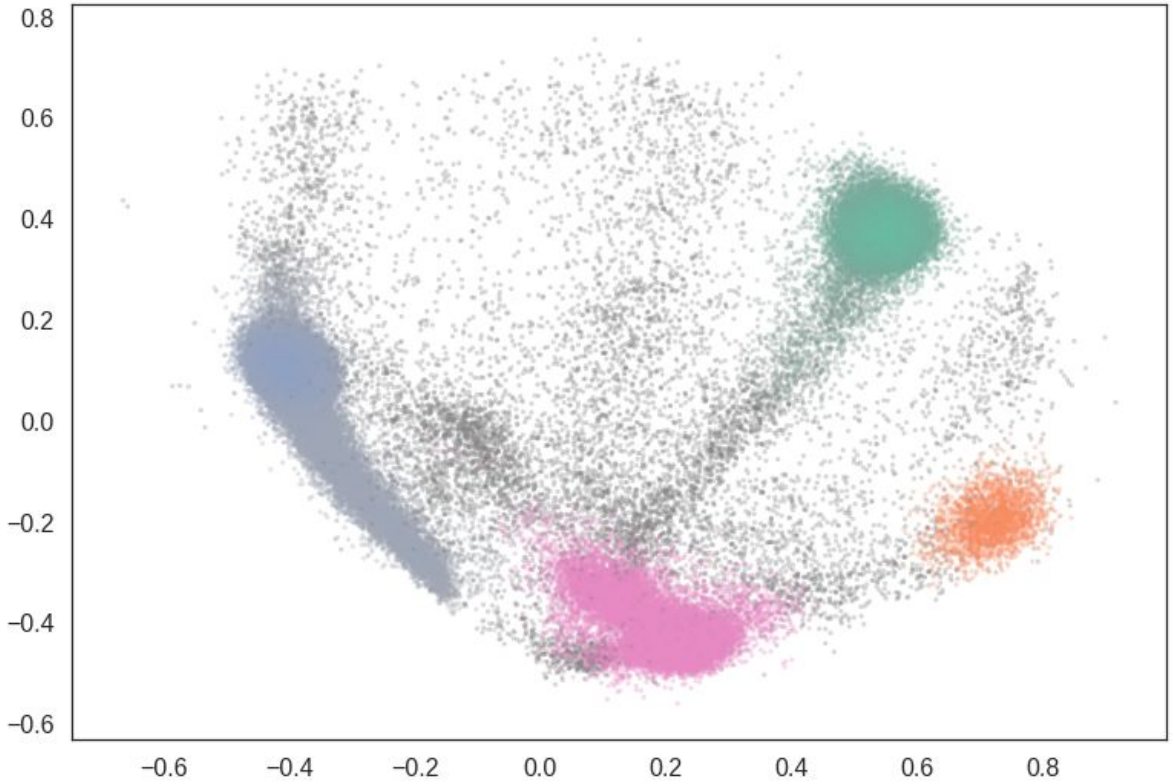
Step 1: SNPs calling

- BWA + samtools mpileup + VarScan
  wrapped in SnakeMake



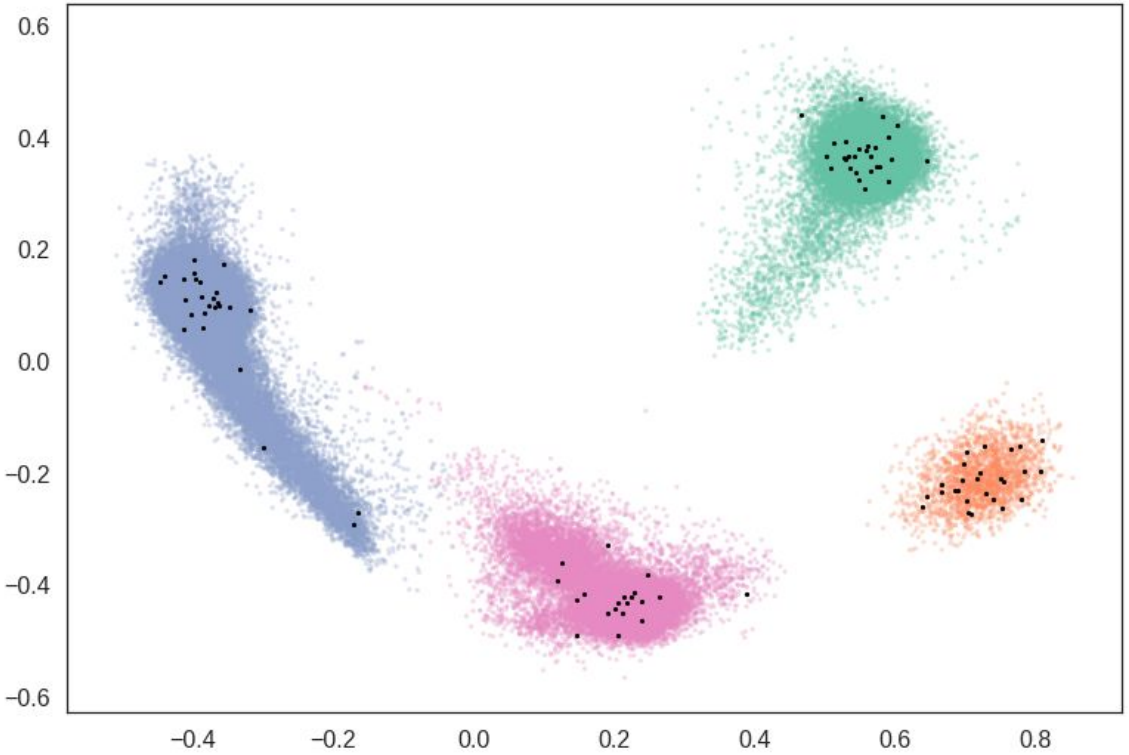
Problem:
Clomial and PhyloWGS can’t work with **too many SNPs**

Step 2: SNPs clusterisation



HDBSCAN

Step 3: SNPs subsampling



Step 4: run cancer tools

- **PhyloWGS** finished only on simple data with high coverage and produced poor results
- **Clomial** results are in good agreement with ground truth

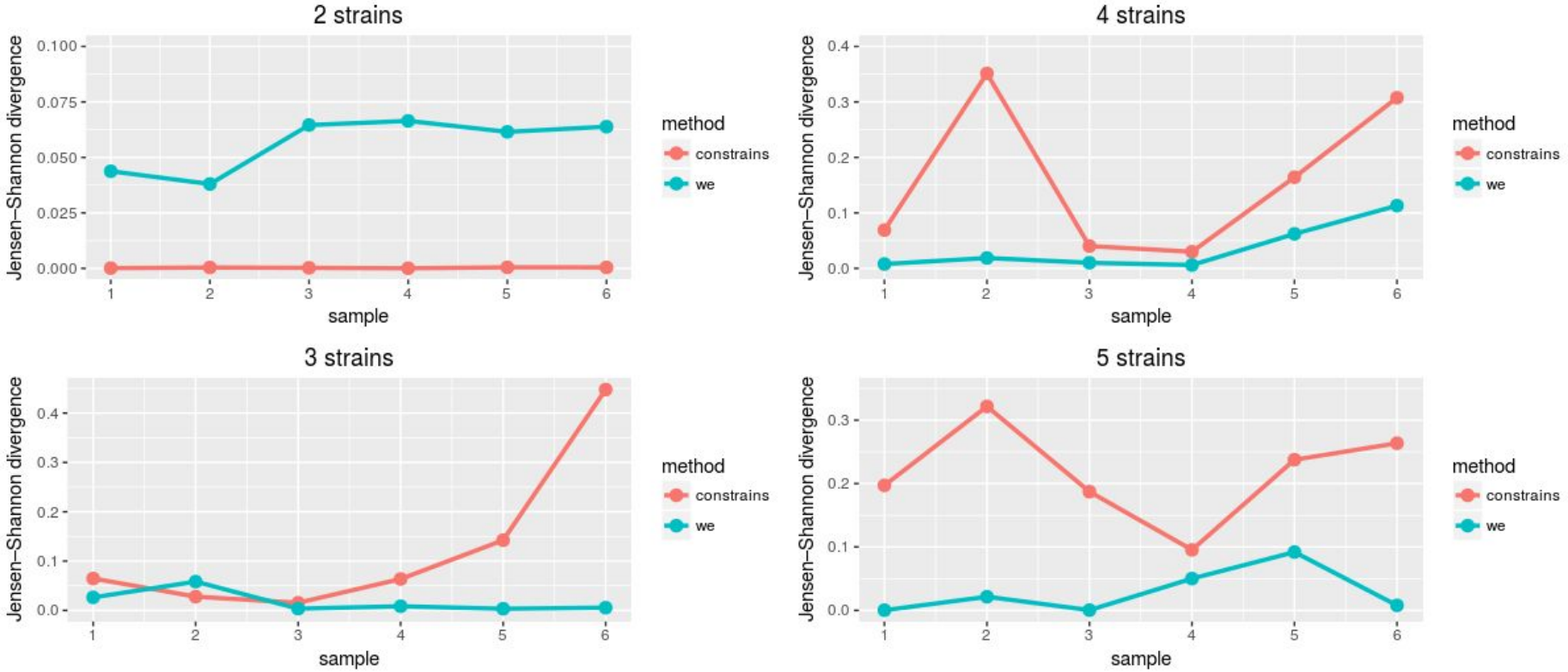
Further results concern Clomial

Datasets

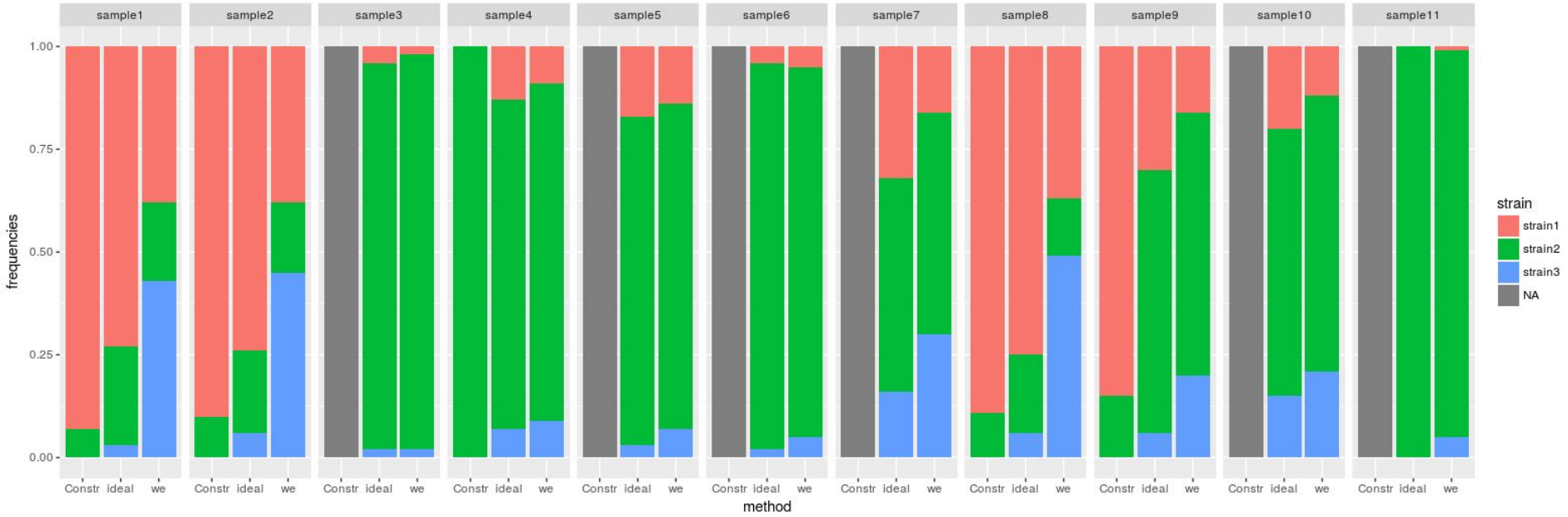
- Synthetic data
  - 4 datasets: 2 - 5 strains, 6 samples in each
  - frequencies from ConStrains paper
- Real data
  - Infant gut [Sharon et al, 2013]
  - known abundances of the *Staphylococcus epidermidis* strains

Benchmarking: count Jensen-Shannon divergence between predicted and real frequencies

Results: synthetic data



Results: infant gut



ConStrains found 2 out of 3 strains and failed in 6 out of 11 samples

Results

- SNP detection pipeline
- SNP clustering and subsampling
- Benchmarking framework
- Clomial and PhyloWGS were applied to metagenomic series data
- Clomial produced reasonable results, beating the ConStrains baseline

Further plans

- Improve current pipeline
  - determine number of strains
  - study the correctness of resulting genotypes
- Integrate with existing metagenomic pipelines (MIDAS, metaPhlan)
- Test stability with rare strains contamination
- Come up with reference-free analysis
- Try other cancer tools (AncesTree, LICHeE, ...)

Thank you!