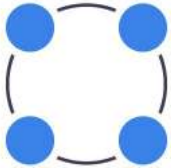


Analyzing strain mixtures in metagenomic series data

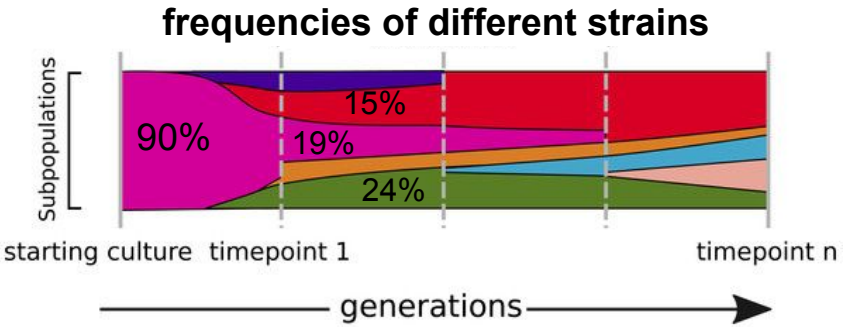
Margarita Akseshina

Scientific advisor: Sergey Nurk

Problem

Input:

- metagenomic series data
  (reads for time/spatial series)
- reference genome / draft assembly



Output:

- the number of related strains
- relative abundances across the samples
- (*) genotypes of strains

Pipeline Outline

1. SNP selection:
    a. calling
    b. filtering
    c. clustering and subsampling

2. Applying cancer tools

    Clomial [Zare et al, 2014]
    PhyloWGS [Deshwar et al, 2015]
    LICHeE [Popic et al, 2015]

    … or something else…

Step 1.b : SNP filtering

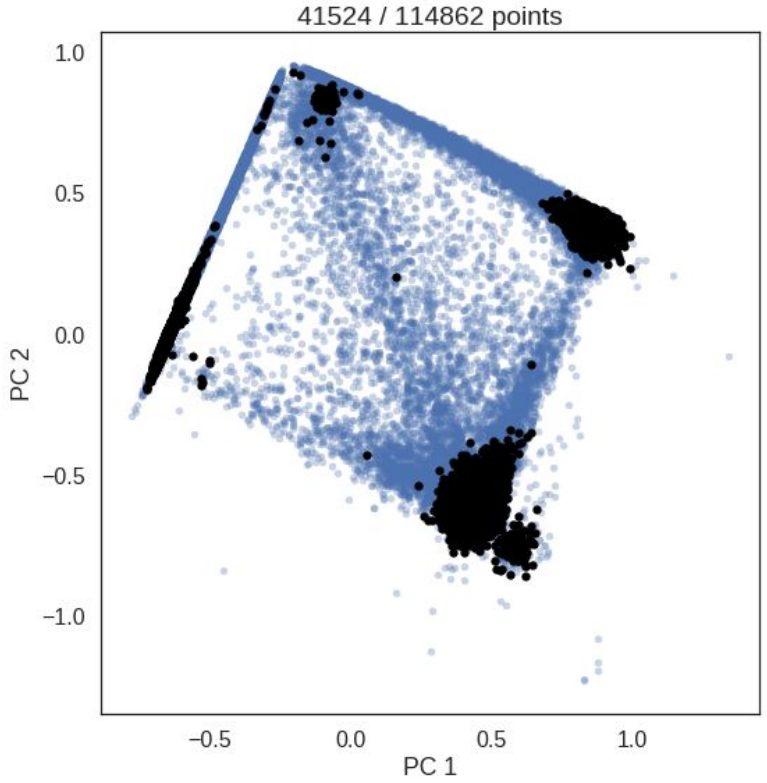
Bad SNPs are:

- copy number variations
- sites with bad coverage
- …

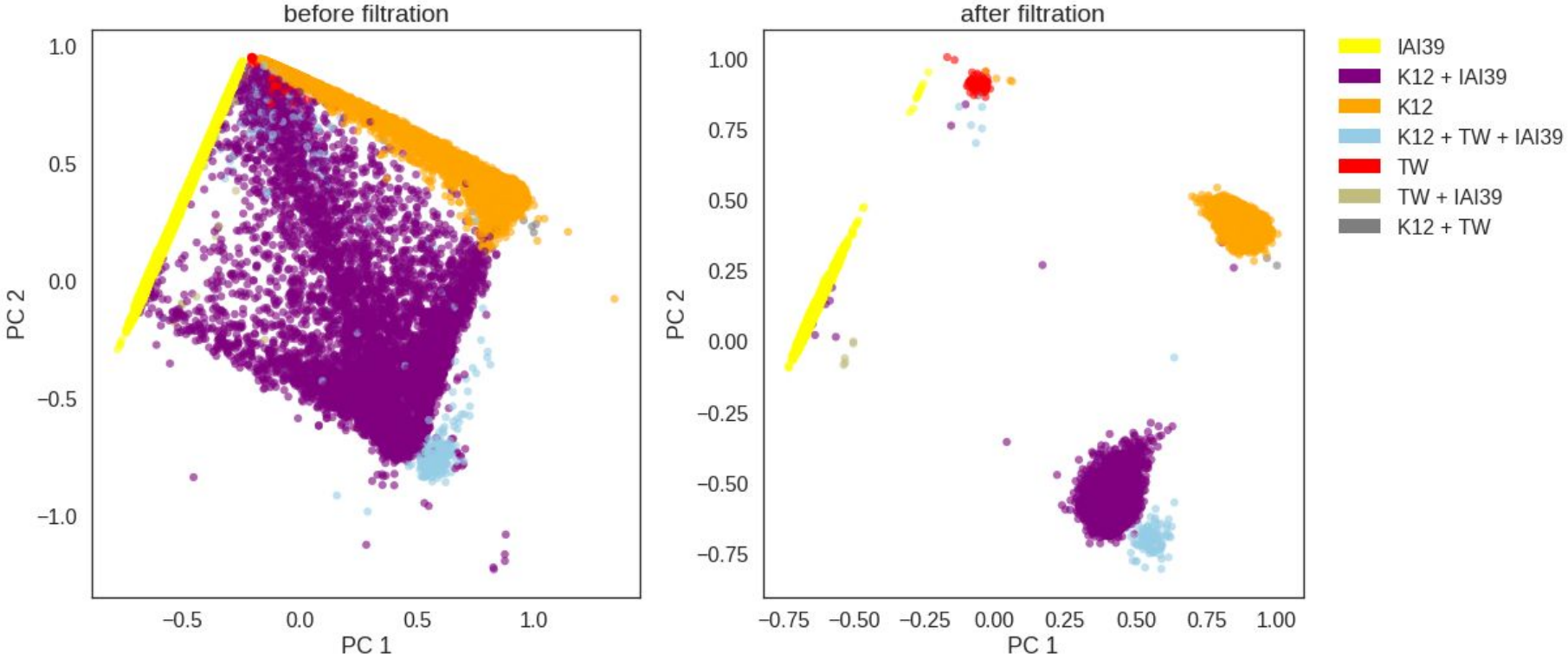
Stupid heuristics so far:

coverage between 15th and 85th percentiles in at least 80% samples

Did we get reasonable results?



SNPs often shared by multiple strains

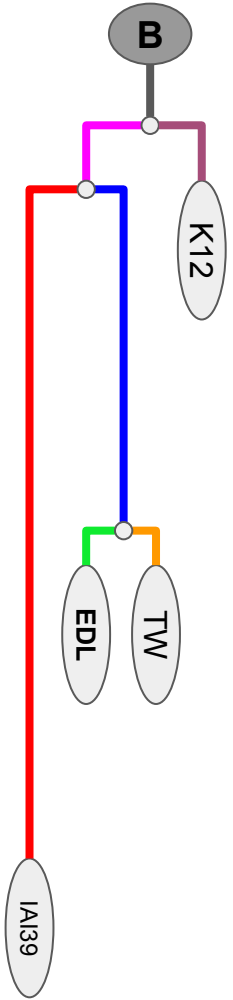


Troubles

Phylogeny inconsistency

=> can’t apply phylogeny based tools 

Group	Size
K12 + EDL + TW	1743
K12 + IAI39	1449

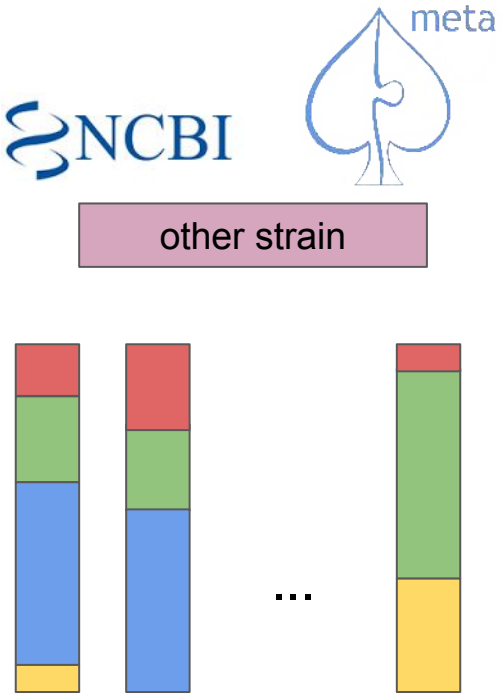


Troubles

External reference

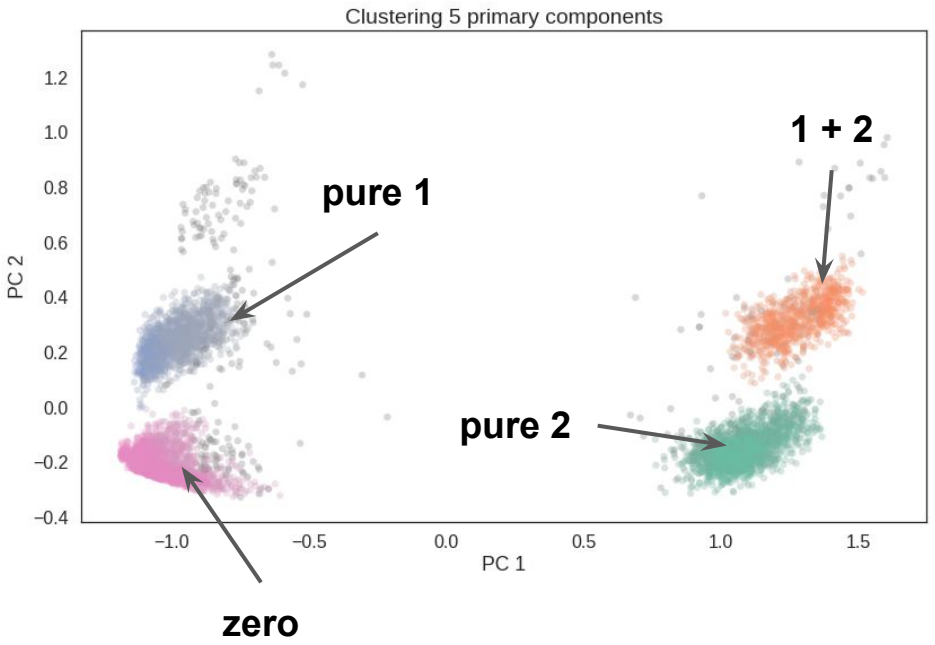
Clomial has one important constraint:
there should be positive amount
of “reference” subclone in every sample

Naive idea:
“mix” in some reads simulated from available
“reference”



Alternative “brute force” algorithm

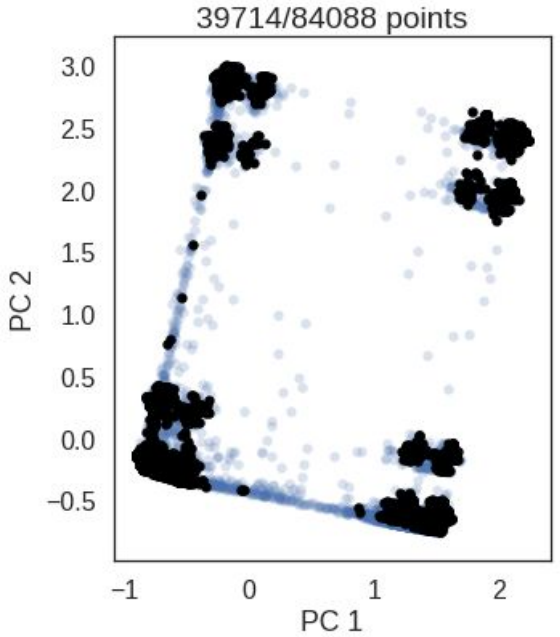
1. Calculate centers of clusters (medians)
2. Find and remove cluster which center is almost zero
3. Find clusters which centers can’t be sum of others (mark them “pure”)
4. Decompose other cluster centers by sum of pure clusters centers



Analysis of E.coli strains Larry Smarr's data

- Crohn's disease patient
- 21 samples with large amount of E.coli 10-70% of sample diversity

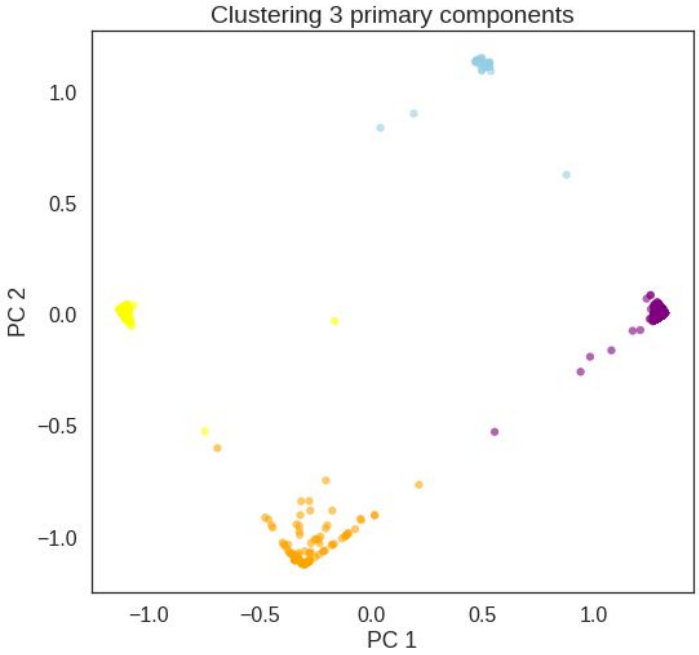
- *Soon realized that deal with a lot of strains*
- *Did not trust the clustering*

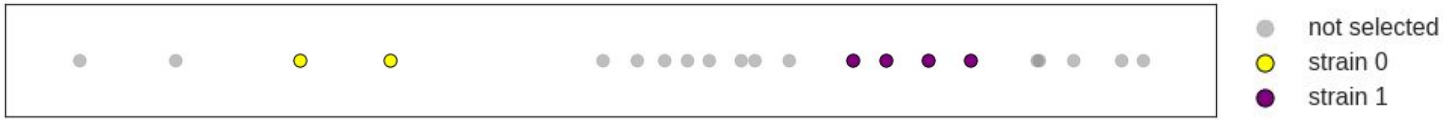


Trick to simplify the data

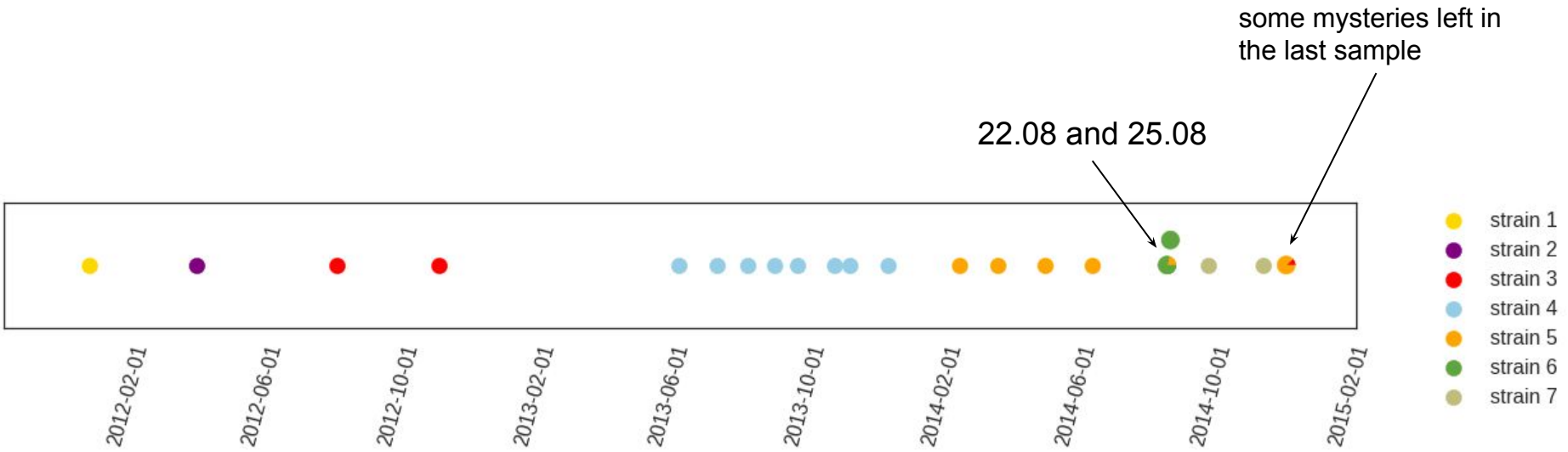
1. Select few samples
   (with presumably few strains)
2. Perform brute force analysis on them
3. Repeat steps 1-2 for other subsets
   and merge results

Almost all samples were pure!!!





Answer on timeline

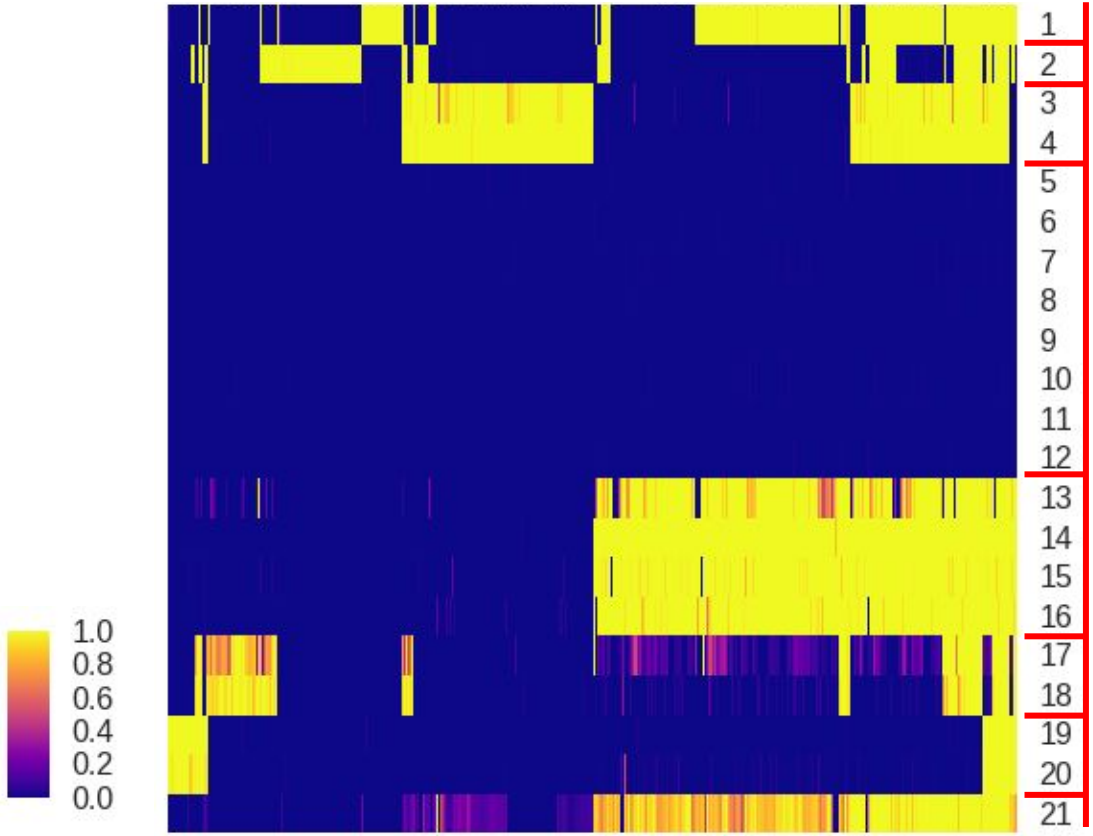


More sensible way

Plot heatmap of SNPs

Rows - samples

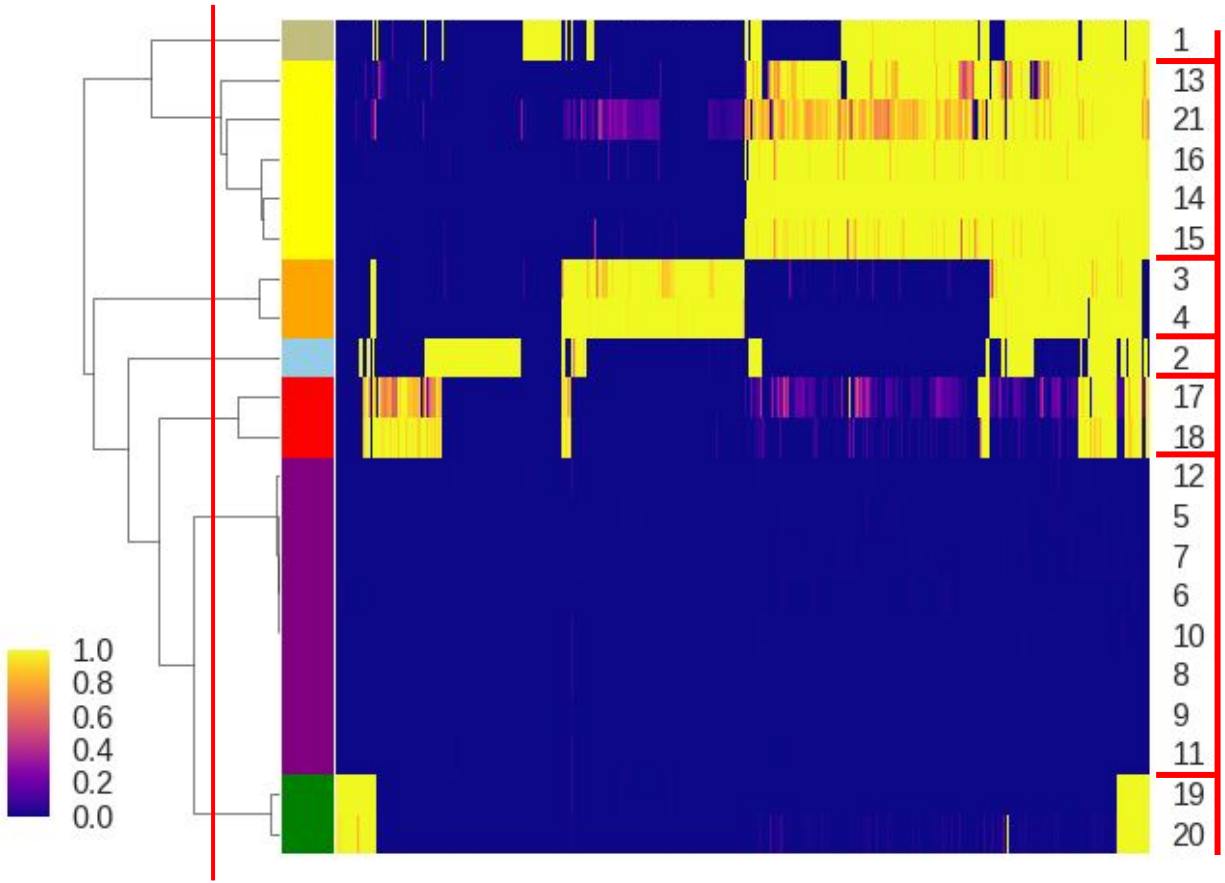
Columns - sites with SNPs clustered by frequencies



Biclustering

Results consistent with previous analysis

SNPs identified by MIDAS [Nayfach et al, 2016], but same results for our SNP calling/filtering pipeline

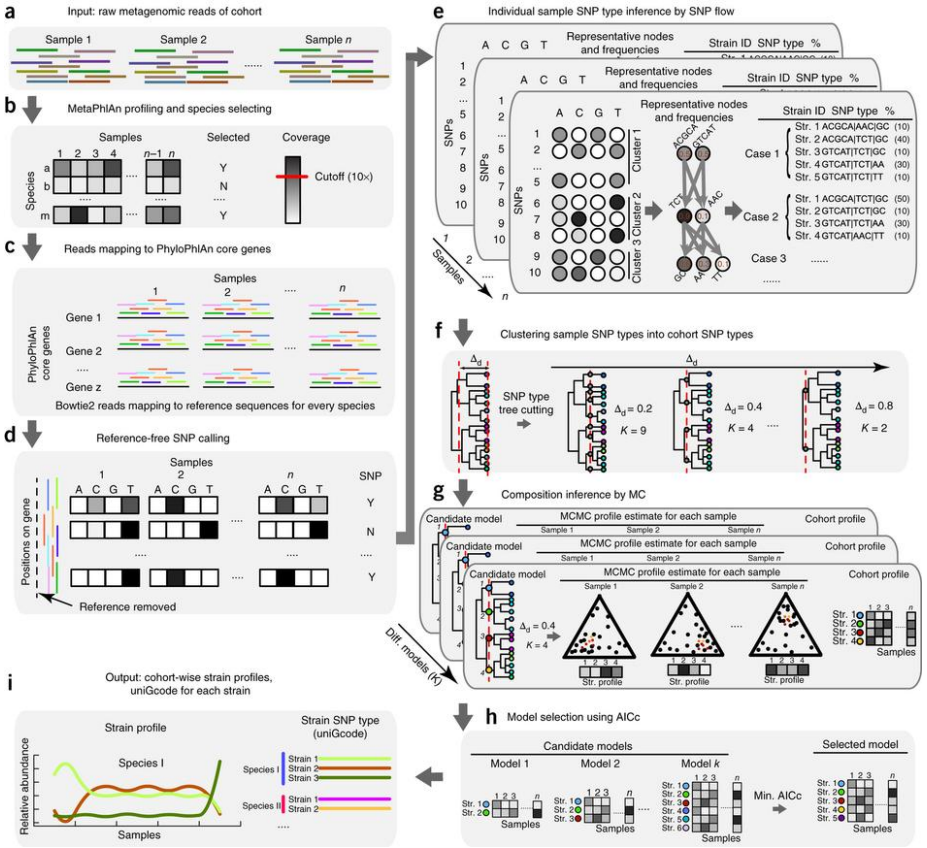


Thank you!

ConStrains [Luo et. al, 2015]

Questionable computational model!

Could not reproduce results from the paper



SNP-based *de novo* strain analysis

species reference

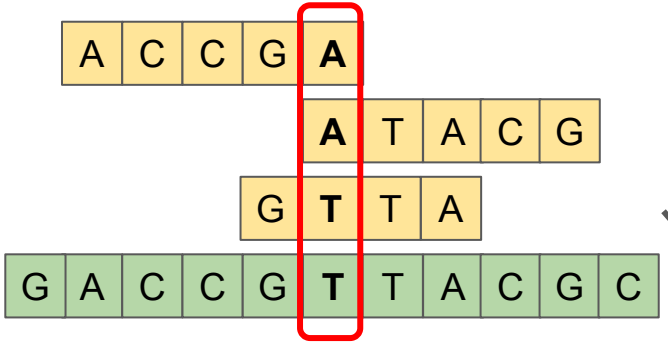
reads from sample 1

reads from sample 2

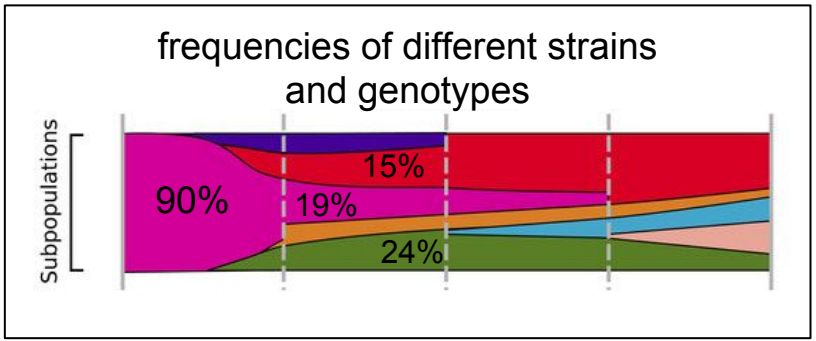
...

reads from sample N

SNPs for every sample:

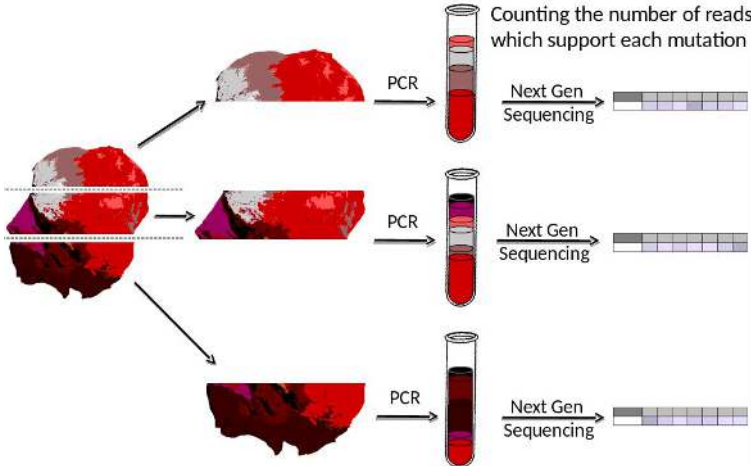


MAGIC (different algorithms)



Subpopulation analysis in multiple tumor samples

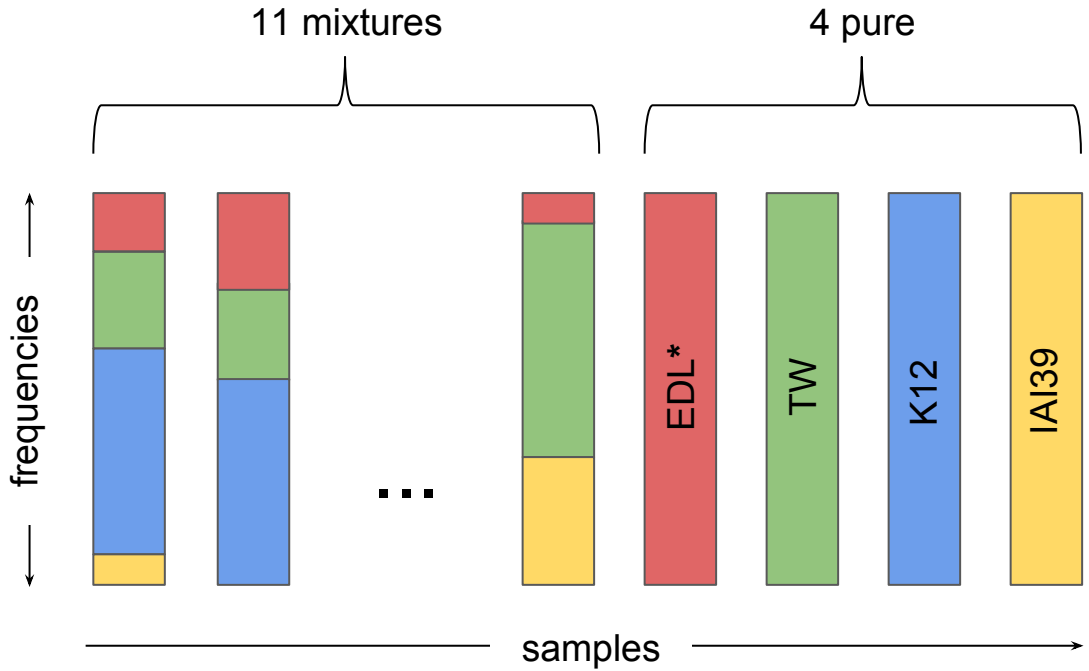
- **Clomial** [Zare et al, 2014]
  matrix decomposition with EM
- **PhyloWGS** [Deshwar et al, 2015]
  MCMC on phylogenies
- **LICHeE** [Popic et al, 2015]
  evolutionary constraint network
- PyClone [Roth et al, 2014]
- AncesTree [Beerenwinkel et al, 2015]
- ...



here we have **normal cells as reference**

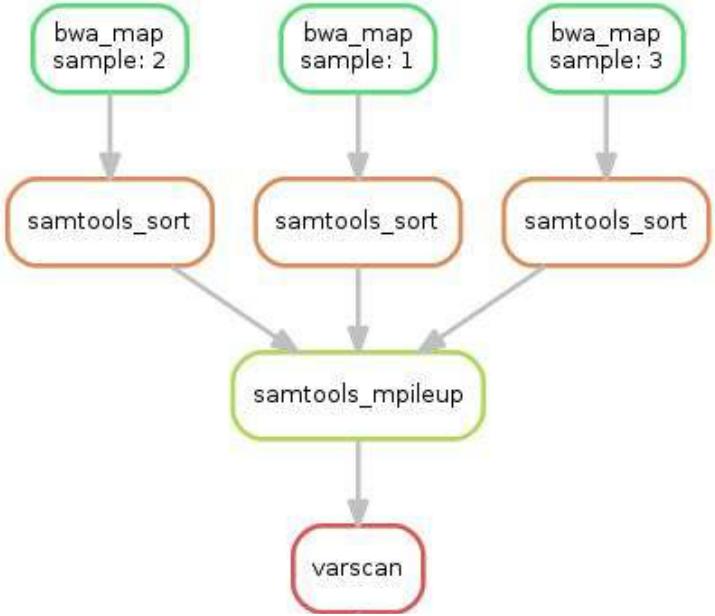
Example synthetic dataset

- 4 E.Coli strains (EDL, TW, K12, IAI39)
- “real life” frequencies in 11 samples
- 4 “pure” samples
- reference* is in data (EDL)



Step 1.a : SNP calling

- BWA + samtools mpileup + VarScan
  wrapped in SnakeMake



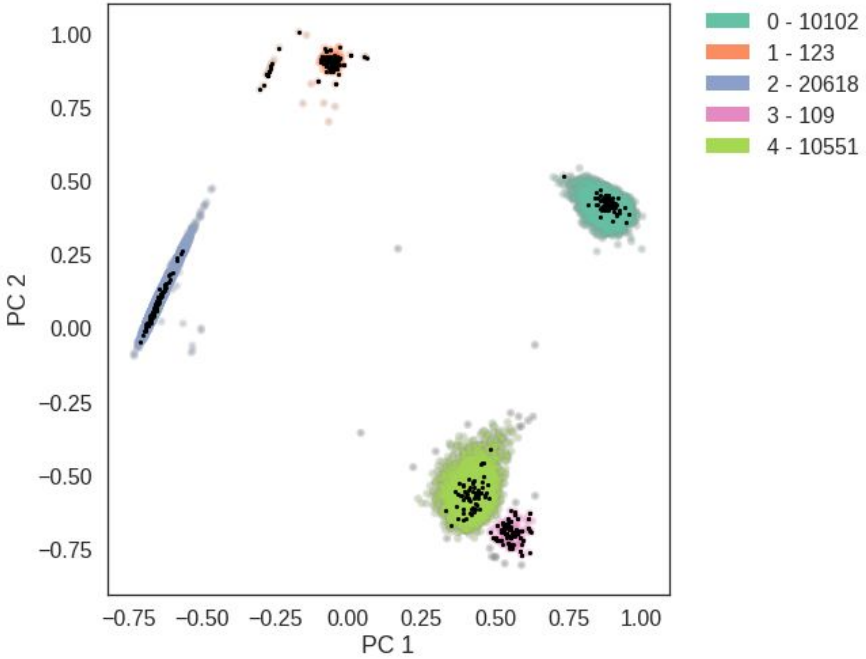
Problem:

Too many SNPs

We want to **filter them**

Step 1.c : SNP clustering and subsampling

- HDBSCAN [McInnes et al, 2017] or KMeans
- subsampling from clusters depending on their sizes

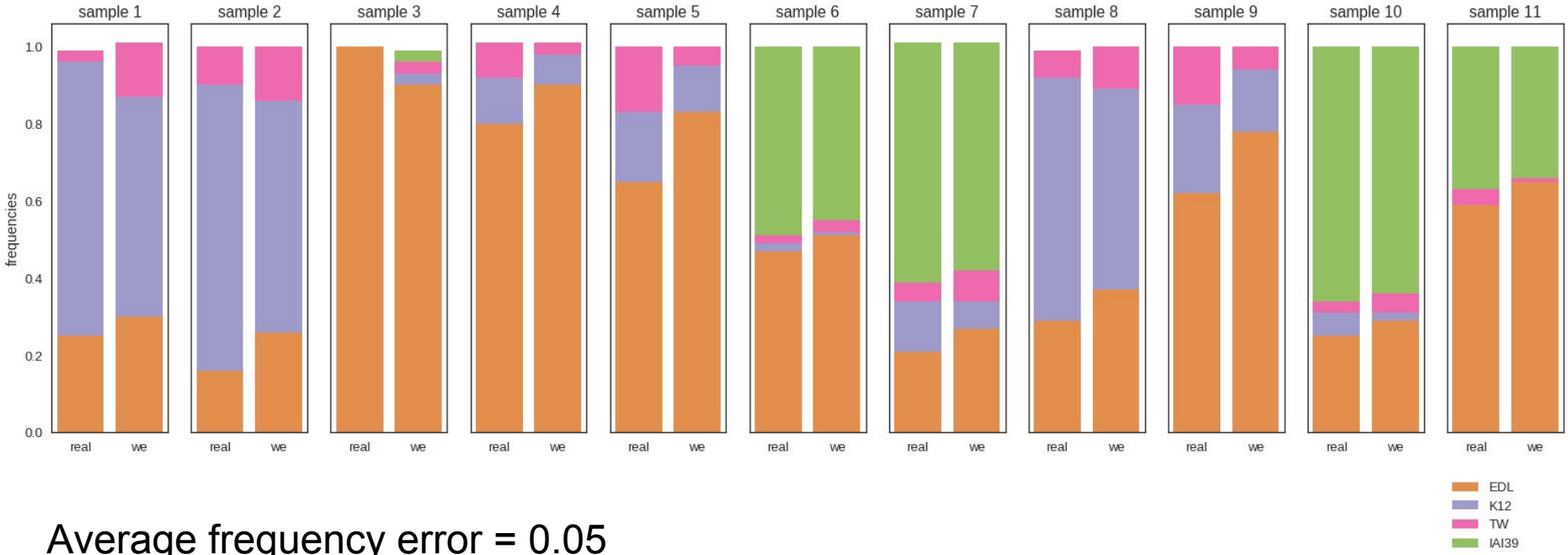


Step 2: Applying cancer tools

- **Clomial** results are in good agreement with ground truth
- **PhyloWGS** finished only on simple data with high coverage and produced poor results
- **LICHeE** requires fine tuning for reasonable clusterization

Further results concern Clomial

Clomial result on synthetic data



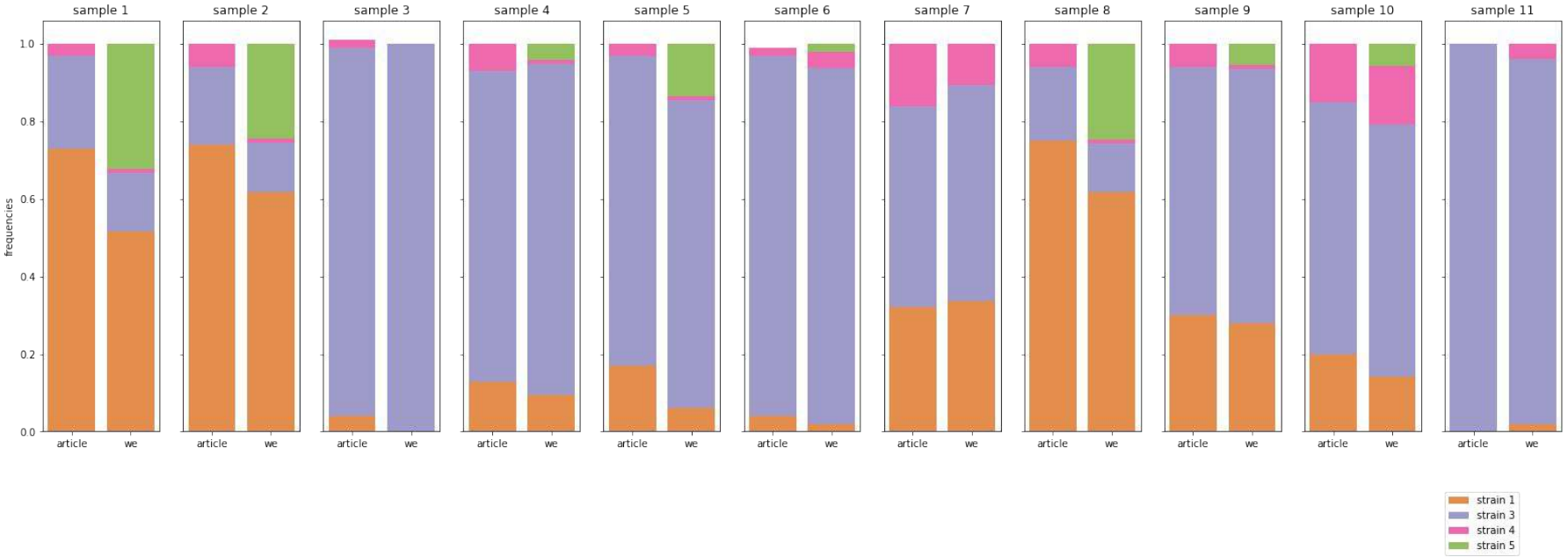
Average frequency error = 0.05

ConStrains (Luo et al, 2015) did not identify strains at all

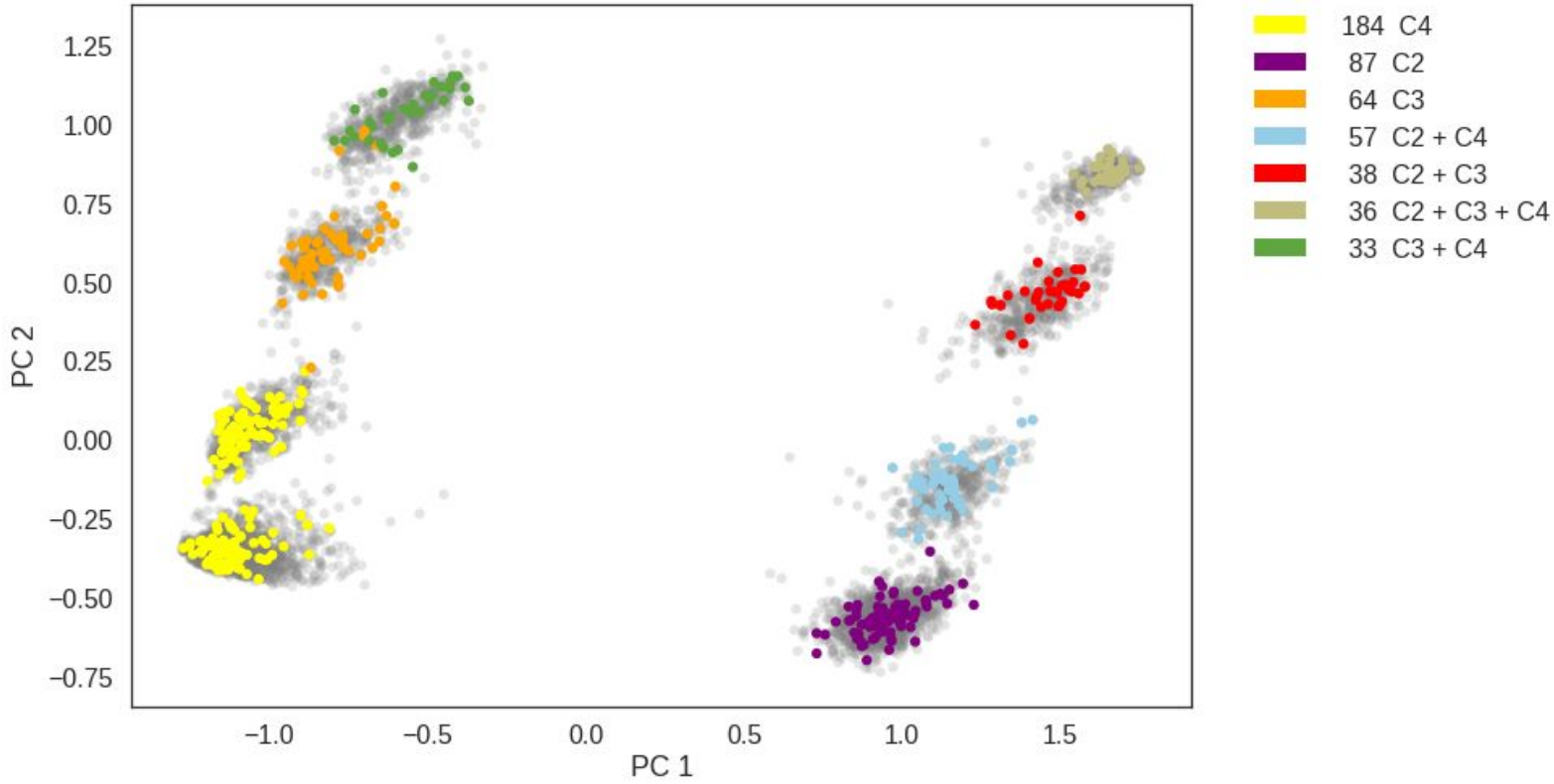
Real dataset: Infant Gut

- [Sharon et al, 2013]
- 11 samples
- “known” abundances of the *Staphylococcus epidermidis* strains

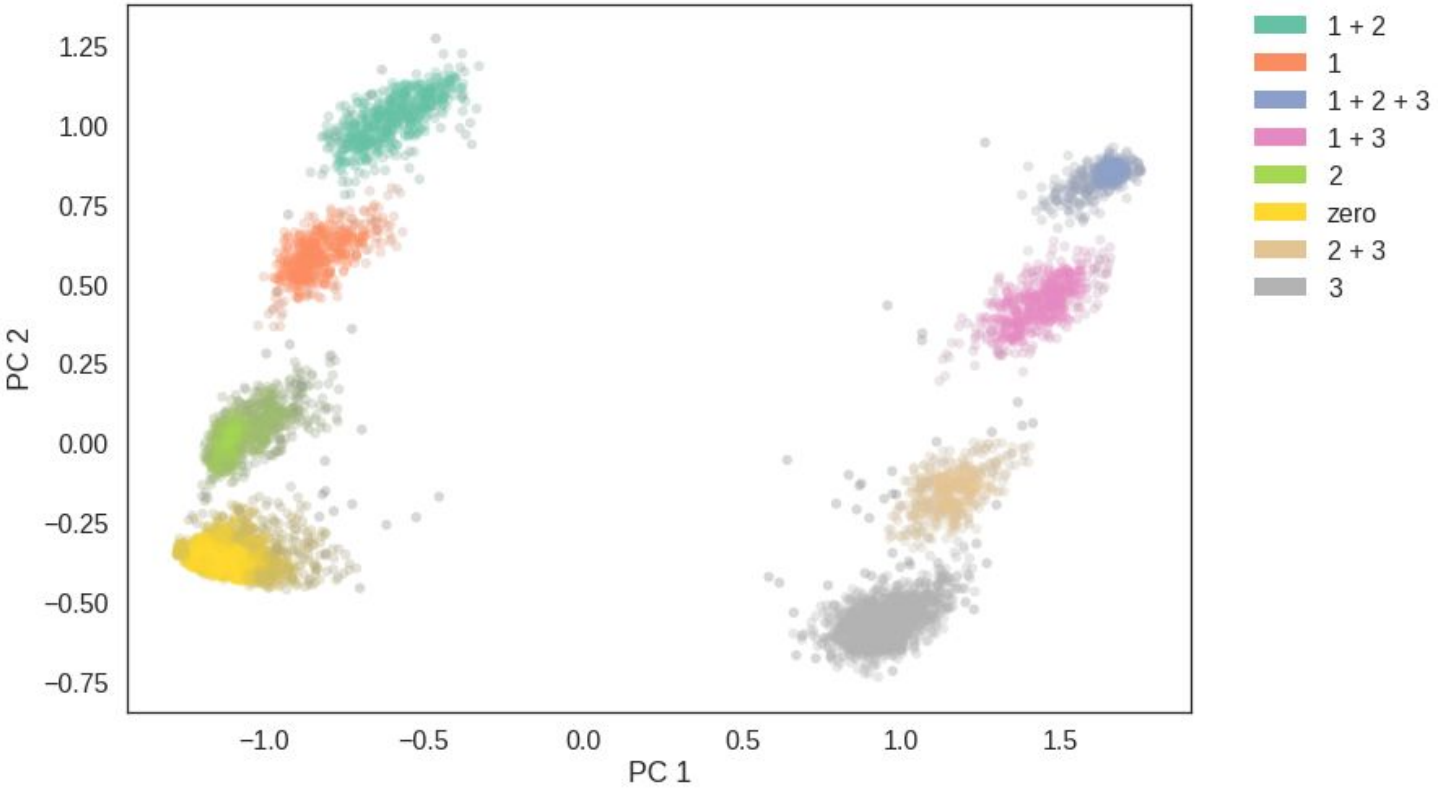
Comparison to Sharon et al 2013 results



Clomial genotypes against NCBI reference



Brute force results



Identifying samples with dominant strains

Task: find samples which have a dominant strain (frequency > 60 %)