Population genetic modeling of genomic variation in *Streptococcus pneumoniae*

Pekka Marttinen, Nicholas J. Croucher, Michael U. Gutmann, Jukka Corander, William P. Hanage*

October 2020

Aalto University
School of Science

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

VE RI TAS

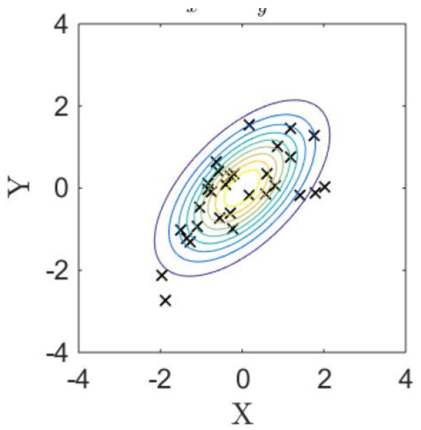
Imperial College
London

Synopsis

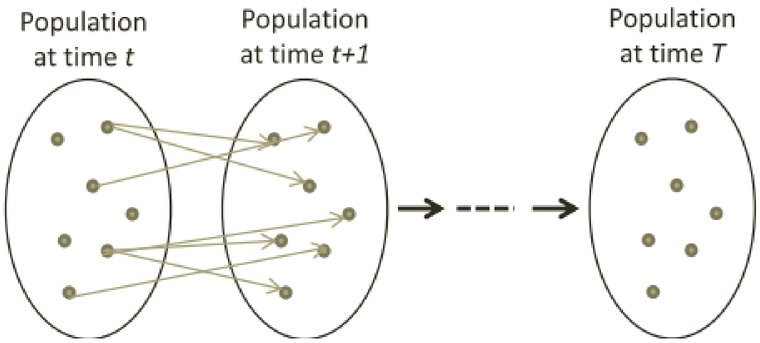
- **Background:** the number and size of bacterial genome collections is increasing rapidly.
- **Issue:** understanding how genomes evolve to produce the patterns observed in the data sets is incomplete.
- **Our goal:** to increase understanding on the evolutionary processes that shape the bacterial genomes.
- **Results:** We present a simulation model that helps to understand some high-level summaries in a collection of 616 *Streptococcus pneumoniae* whole genomes.

Simulation-based modeling

- Statistical inference, the common way
  - Assume some likelihood: *p(data|parameters)*
  - Learn *parameters* that best fit the *data*
  - Example: bivariate normal distribution



- Sometimes likelihood can not be defined or computed, but simulating data from the model is possible
  - Example: population genetics



- Applications: genetics, economics, material physics, ...

Overview

- Summary
- Biological concepts
- Background
- Data
- Model & Model fitting
- Results and conclusions

Concepts (1/2)

- **Core genome:** collection of genes shared by all genomes of a bacterial species

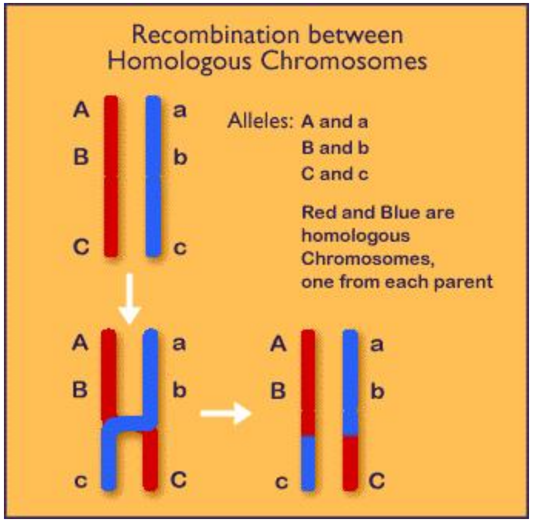
- **Accessory genome:** collection of genes present in some but not all genomes of a species
  - For example only 11% of all *Escherichia coli* genes are core.

	Gene 1	Gene 2	Gene 3	…	Gene K
Strain 1	0	0	1		
Strain 2	1	1	1		
Strain 3	0	1	1		
…					
Strain N					

Gene presence-absence matrix

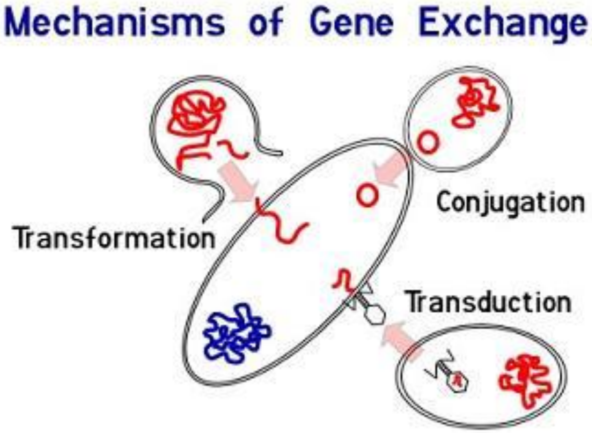
Concepts (2/2)

- **Recombination** shuffles bits of DNA between different chromosomes.


Recombination between Homologous Chromosomes
A
B
C
a
b
c
Alleles: A and a
B and b
C and c
Red and Blue are homologous Chromosomes, one from each parent
A
B
c
a
b
C
A
B
c
a
b
C

http://members.cox.net/amgough/Fanconi-genetics-genetics-primer.htm

- **Horizontal gene transfer** permits the exchange of DNA between different species

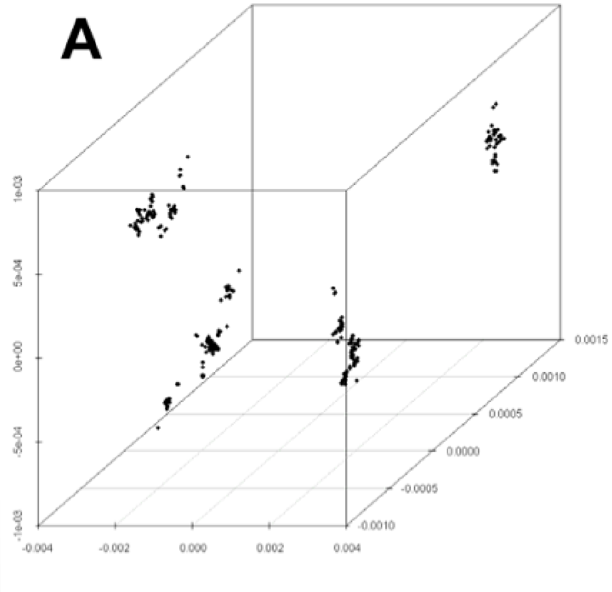


http://science.kennesaw.edu/~jdirnber/Bio2108/Lecture/LecBiodiversity/GeneTransfer1.jpg

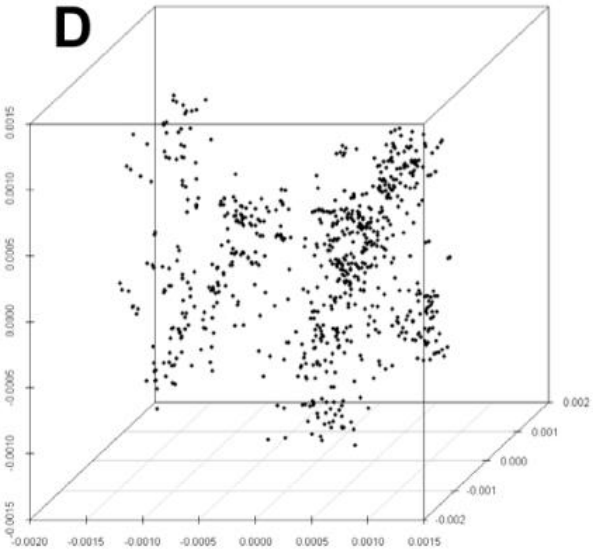
Background

- Fraser et al. (2007) presented a model for the core genome showing how recombination holds a population together

Low recombination rate



High recombination rate



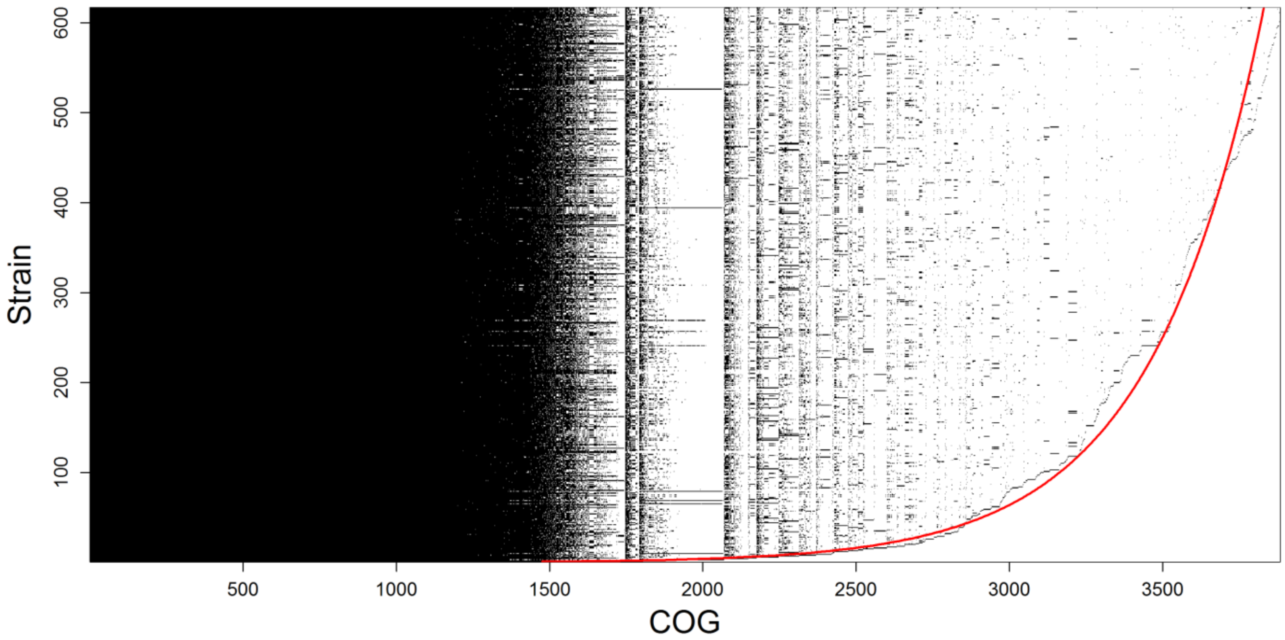
Overview

- Summary
- Biological concepts
- Background
- Data
- Model & Model fitting
- Results and conclusions

Data (1/2)

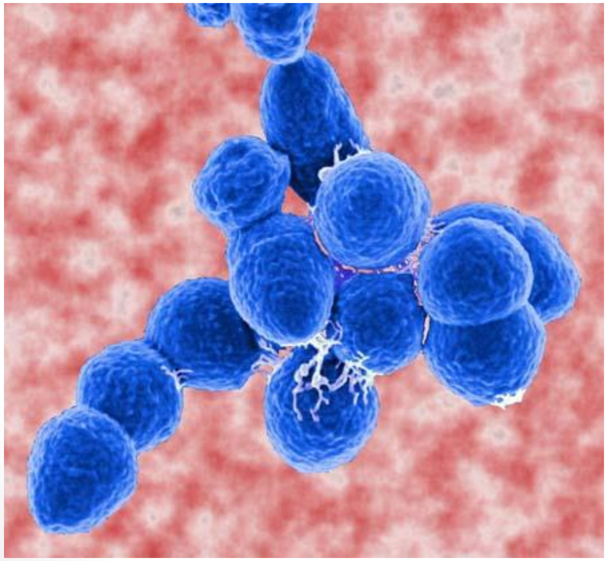
- 616 *Streptococcus pneumoniae* strains sampled in Massachusetts
- Gene presence-absence matrix
- Sequence alignments at core genes (COGs)
- Croucher et al. (2013)

Left-ordered COG presence-absence matrix

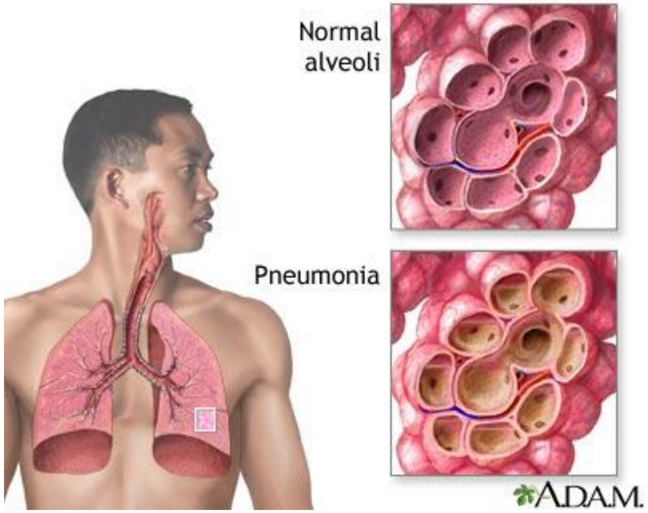


Streptococcus pneumoniae

- Lives in human upper respiratory system
- Multidrug resistant strains exist
- Infections
  - Pneumoniae
  - Meningitis
  - Etc...



http://sitemaker.umich.edu/mc13/bacterial_meningitis_causative_organism



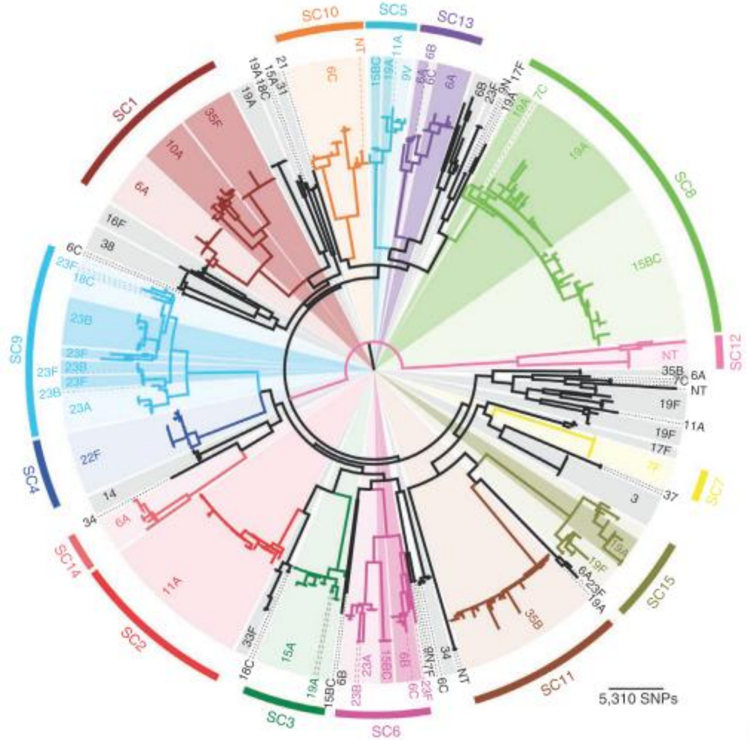
http://www.beltina.org/health-dictionary/pneumococcal-pneumonia-symptoms-treatment.html

Data (2/2)

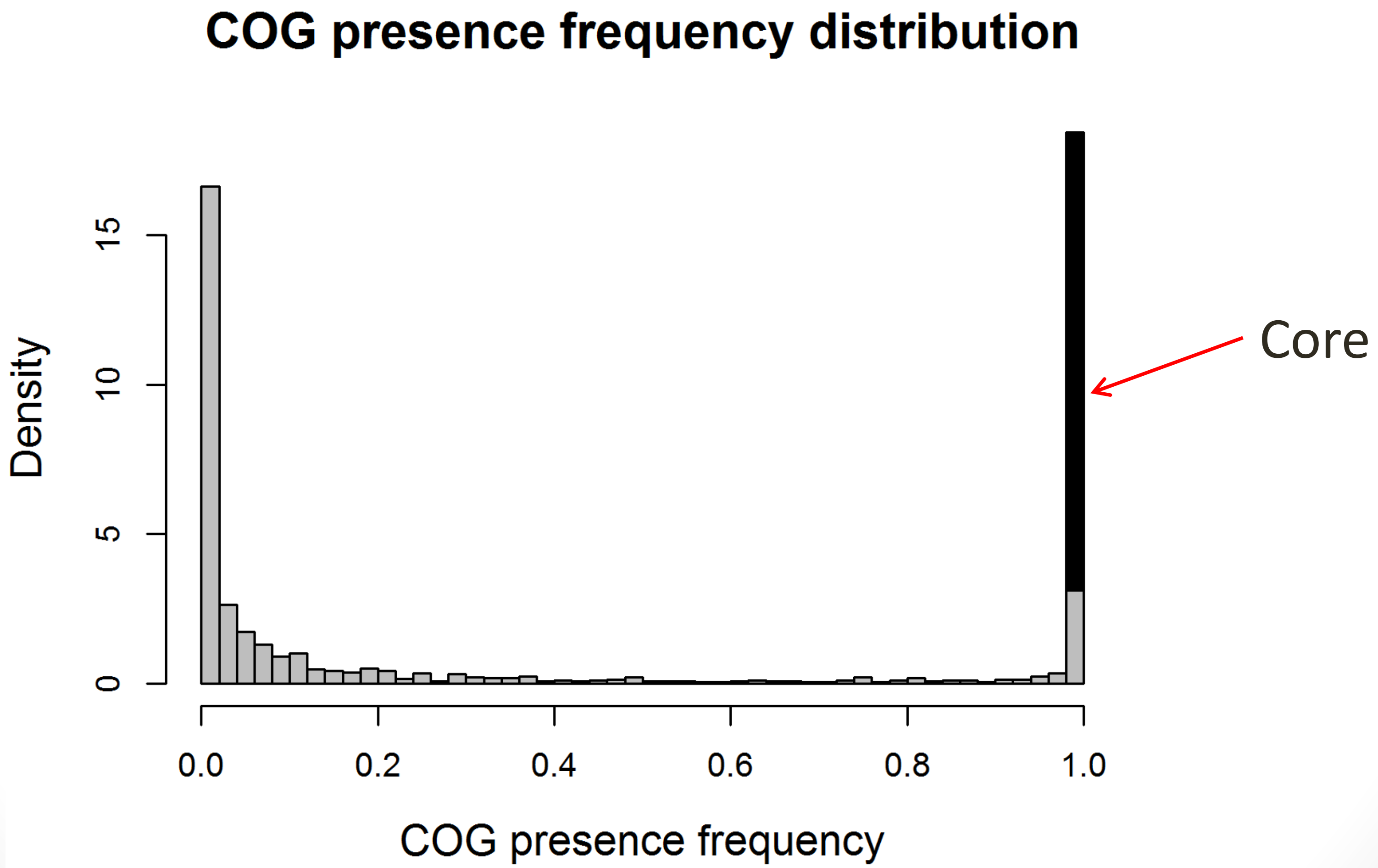
- Sequence alignments for core genes

	Core gene 1				Core gene 2					...			Core gene G			
Strain 1	A	C	G	G	A	-	T	C	C							
Strain 2	A	C	C	G	A	C	T	C	C							
...																
Strain N																

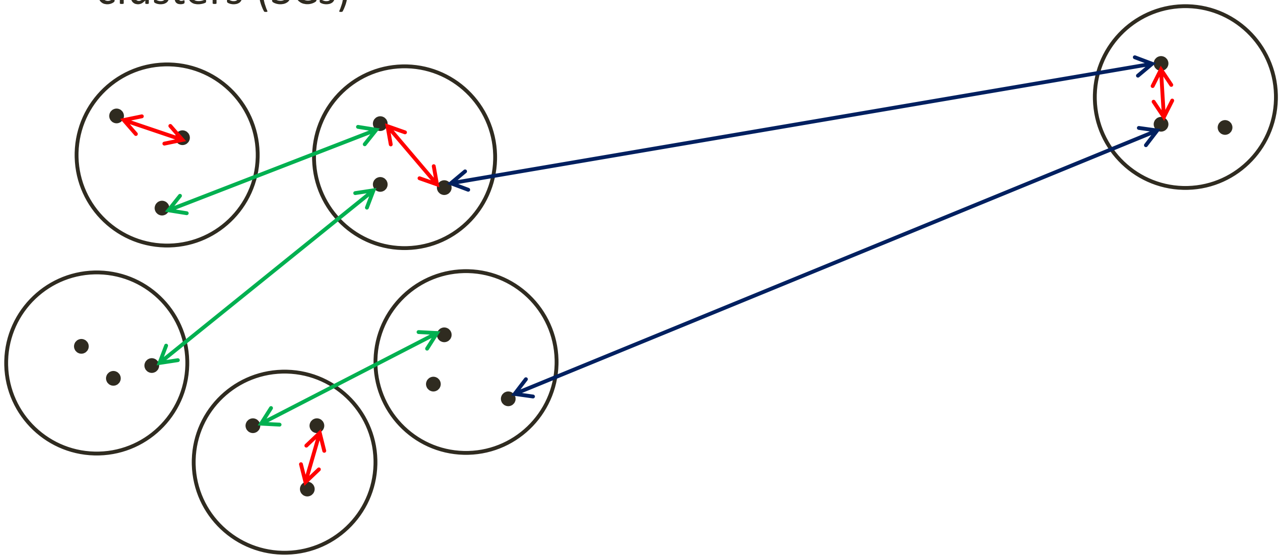
- Phylogenetic tree can be estimated using the core genome
- 15 distinct strain clusters can be identified



Data summaries (1/2)



Observed population structure



14 equidistant strain clusters (SCs)

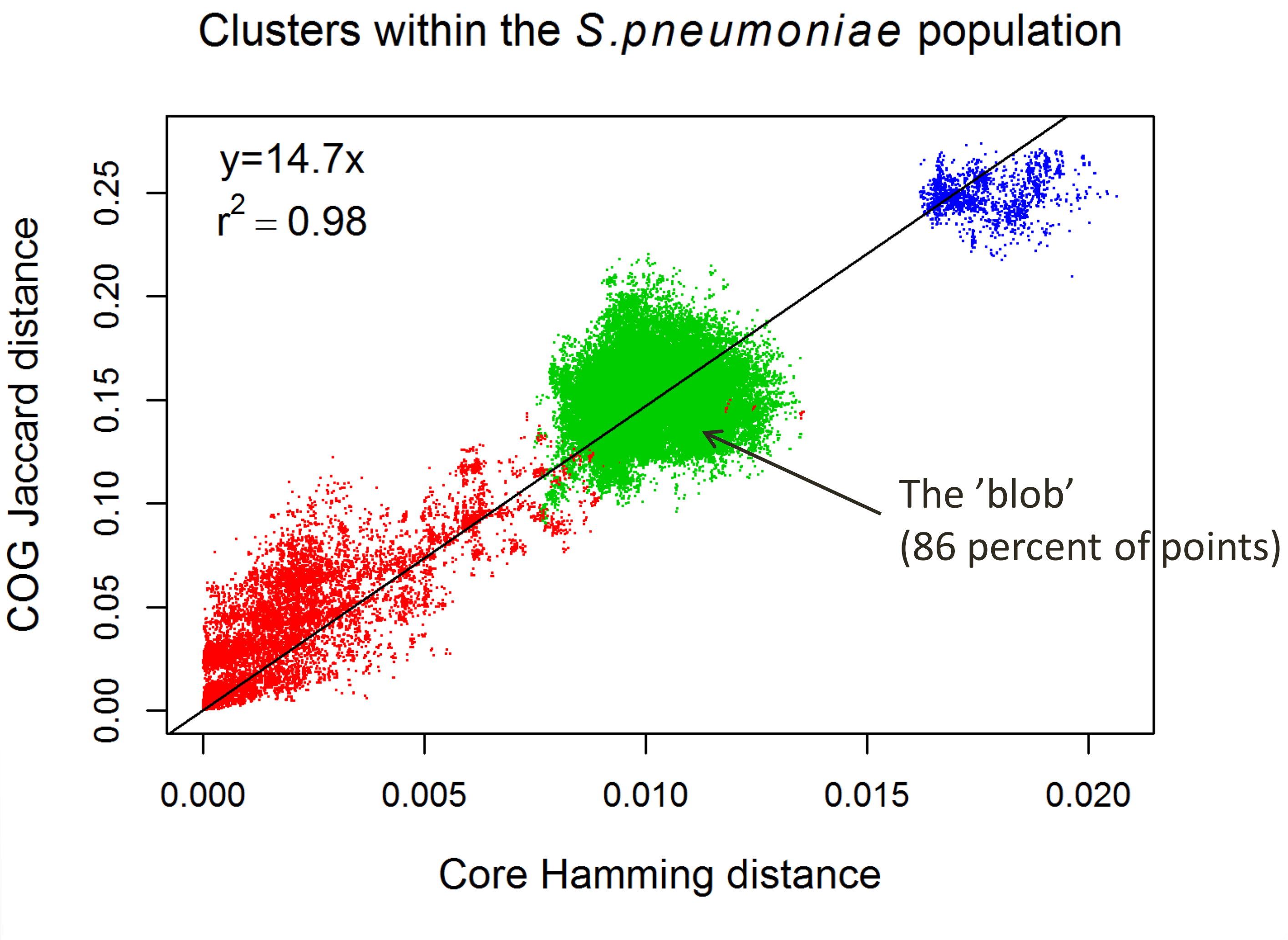
One divergent strain cluster, SC12

↔ within-cluster distances

↔ between-cluster distances (excl. SC12)

↔ distances between SC12 and other clusters

Data summaries (2/2)



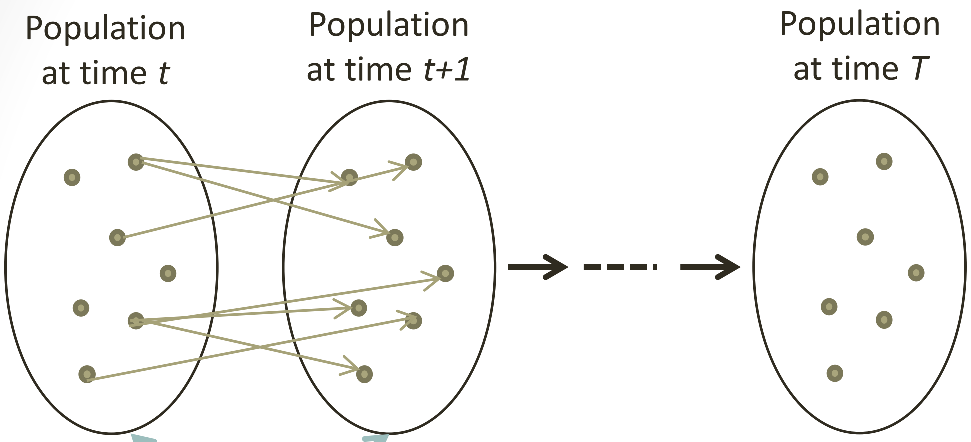
Clusters within the *S.pneumoniae* population

The ’blob’
(86 percent of points)

Overview

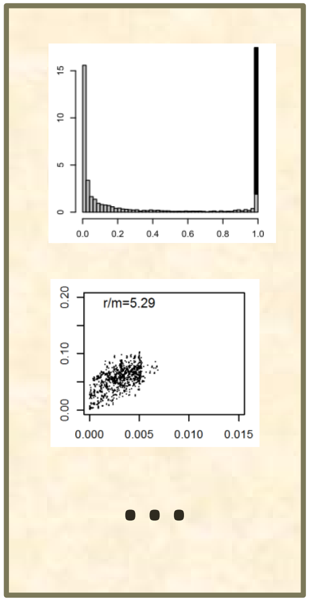
- Summary
- Biological concepts
- Background
- Data
- Model & Model fitting
- Results and conclusions

Modeling approach



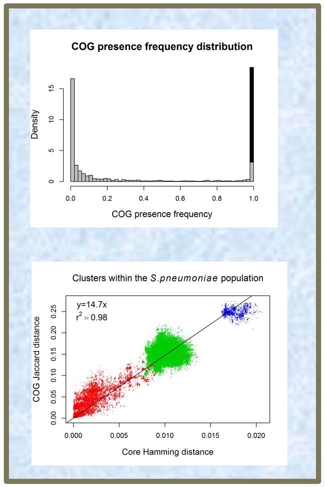
- mutation
- recombination (homol.)
- gene introduction
- gene deletion
- gene transfer

Compute summaries



Simulated summaries

Compare



Observed summaries

The model is a so-called Wright-Fisher model.

Efficient simulation-based inference, see Gutmann and Corander (2015) or Järvenpää et al. (2019).

Inference

- **Parameters affecting gene content** (deletion rate, novel gene introduction rate, gene transfer rate) fitted by matching
  - Gene frequency histogram
  - Median clonality score (see the article)

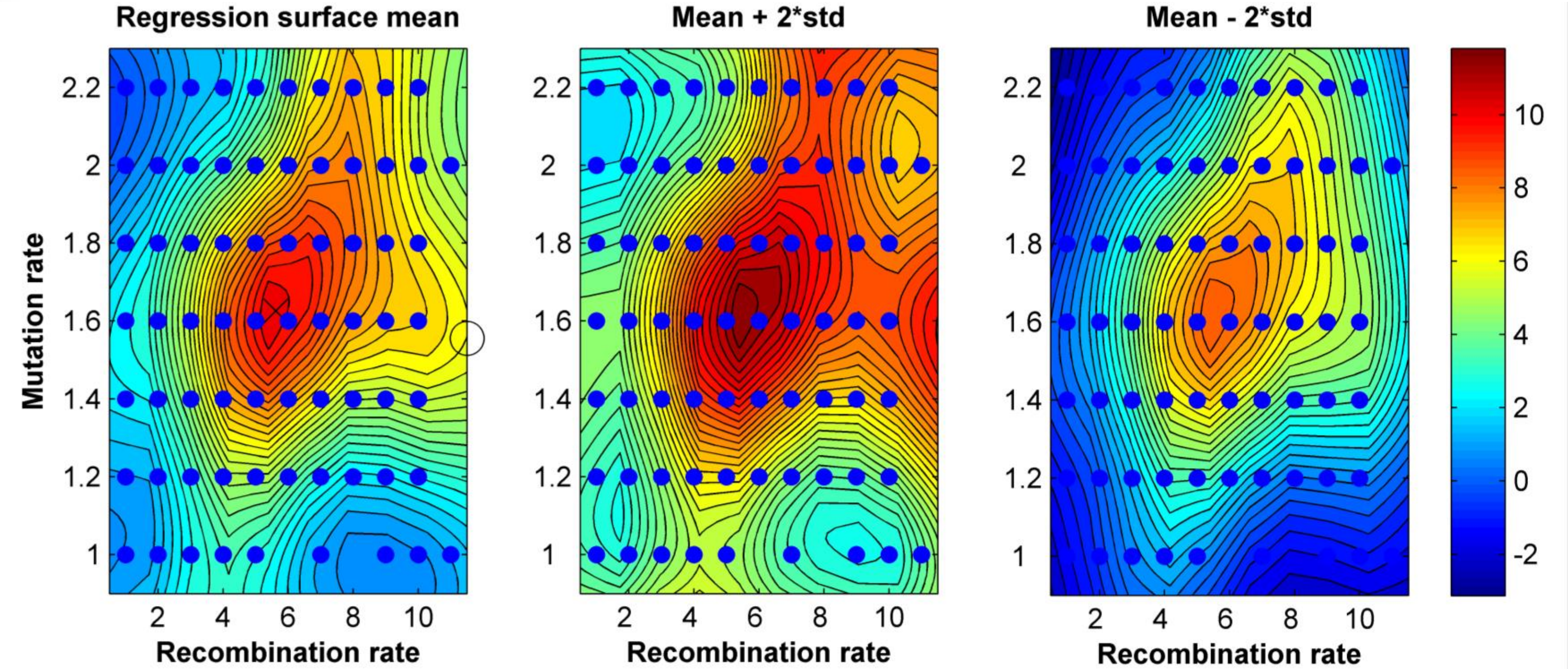
  between real data and data simulated from the model.

- **Parameters affecting core genome** (mutation rate, homologous recombination rate) by matching
  - Slope of the Jaccard vs. Hamming plot
  - Median linkage score (see the article)

Model fitting illustrated

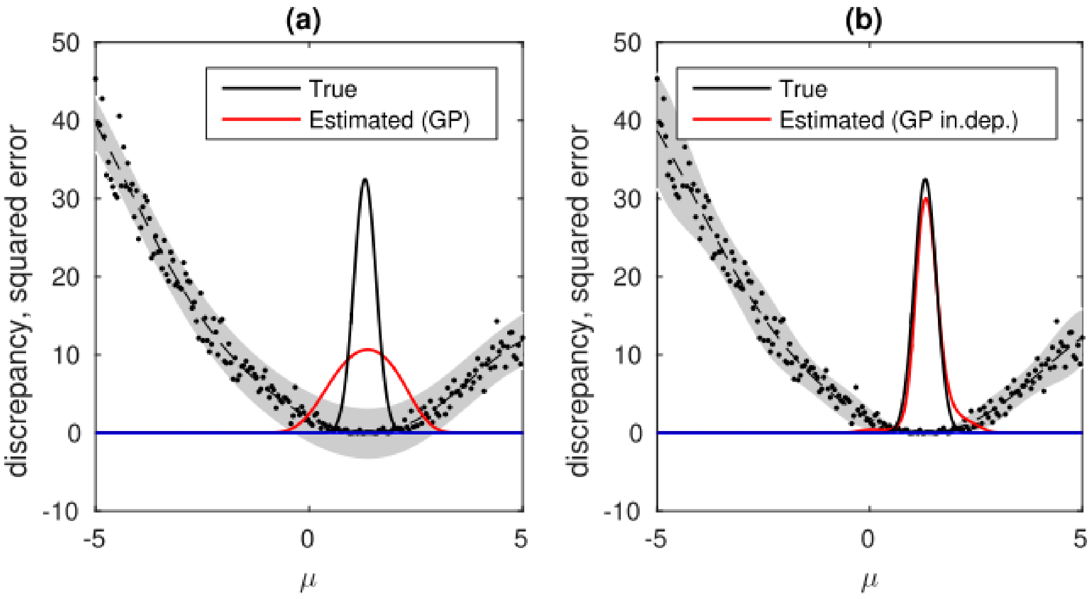
- Maximize the similarity between simulated and real data summaries. Here the similarity is defined as

$$-\log((s_{simu} - s_{\mathrm{real}})^2) - \log((l_{simu} - l_{\mathrm{real}})^2)$$



Aspects of model fitting

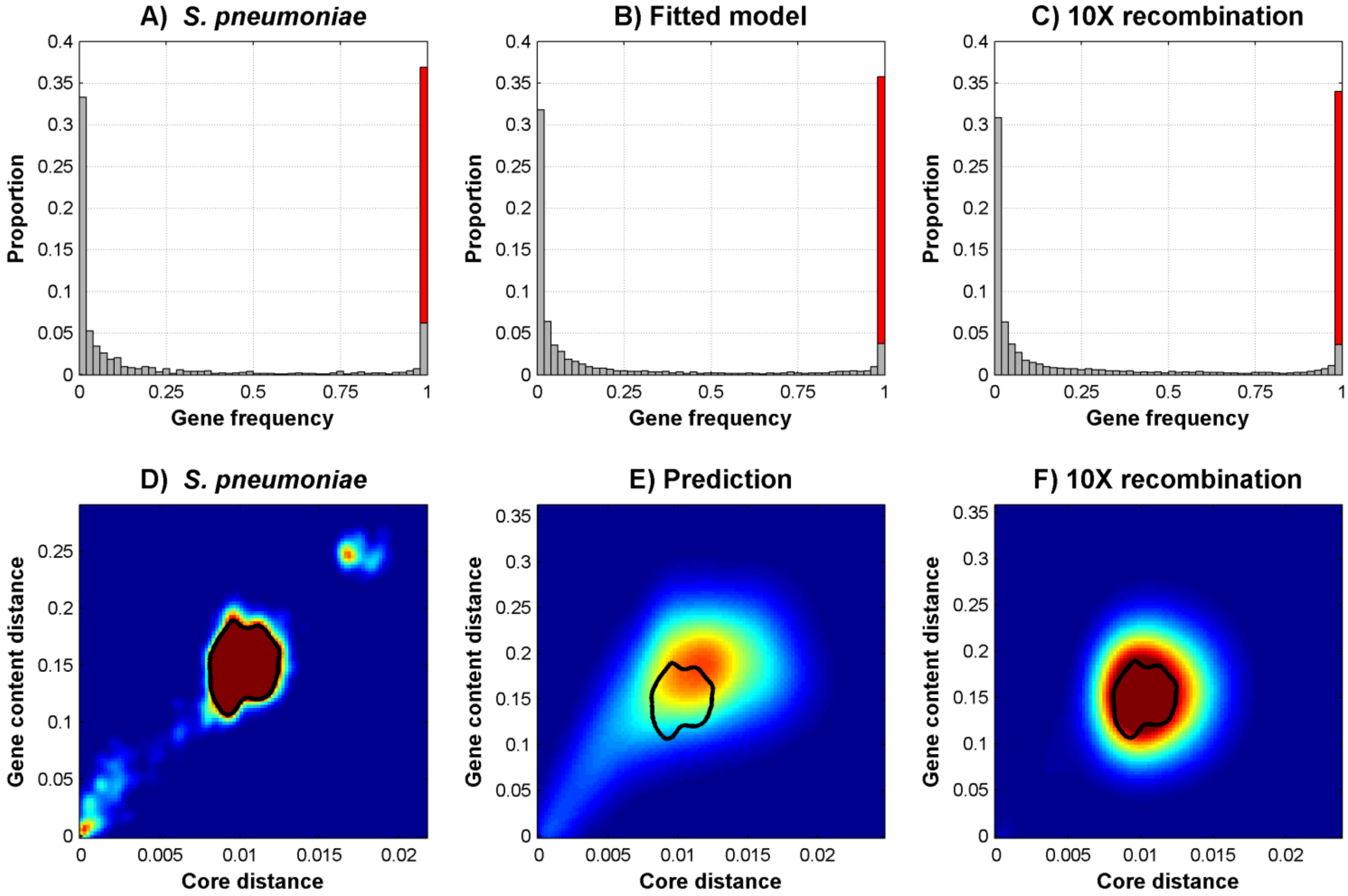
- How to select the next point to evaluate? ->Bayesian optimization.
- How to get most of the existing model evaluations? ->GP-ABC. Example below.
- Active research topics in probabilistic machine learning.



Overview

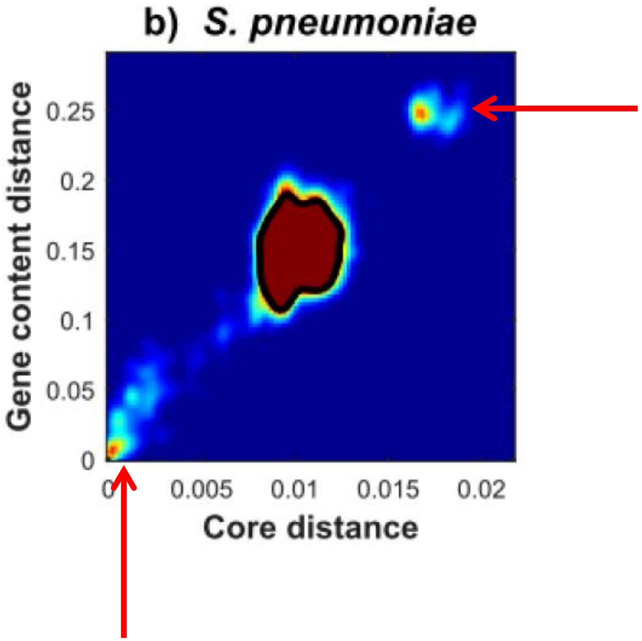
- Summary
- Biological concepts
- Background
- Data
- Model & Model fitting
- Results and conclusions

Results



A) *S. pneumoniae*
B) Fitted model
C) 10X recombination
Proportion
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.25
0.5
0.75
1
Gene frequency
D) *S. pneumoniae*
E) Prediction
F) 10X recombination
Gene content distance
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.005
0.01
0.015
0.02
Core distance

Deviations from the model



the second mode represents a divergent strain cluster that has been recognized as a distinct species.

excess of closely related strains can be explained, e.g., by selection, a recent bottleneck or biased sampling.

Conclusions

- Simulation-based modeling was found useful in helping to understand the genomic structure of a bacterial population
- The model was fitted by matching simulated and observed summary statistics
- High-level features of the observed genomic distribution emerged without explicit selection. -> Nevertheless, the extent of selection remains an open question
- The model predicted the existence of equidistant strain clusters, and this followed from an equilibrium between
  - Diverifying forces: mutation, gene deletion, introduction of genes
  - Cohesive force: recombination, gene transfer

References

- **Croucher, N.J. et al. (2013).** Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nature Genetics* **45**, 656–663.
- **Croucher, N.J. et al. (2014).** Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nature Communications*, Article number, **5471**.
- **Fraser, C. et al. (2007).** Recombination and the nature of bacterial speciation. *Science* **315,** 476–480.
- **Gutmann , M.U. and Corander, J. (2016).** Bayesian optimization for likelihood-free inference in simulator-based statistical models. *Journal of Machine Learning Research*.
- **Järvenpää, M., Gutmann, M.U., Vehtari, A. and Marttinen, P. (2019).** Efficient acquisition rules for model-based approximate Bayesian computation. *Bayesian Analysis.*
- **Marttinen, P. et al. (2015).** Recombination produces coherent bacterial species clusters in both core and accessory genomes. *Microbial Genomics*, 1, doi:10.1099/mgen.0.000038.