Single and collective loci contributions in discrimination among disease outcomes: a study on whole genome epidemiology of Non typeable Haemophilus influenzae

Mara Barucco

UNIVERSITÀ DI PISA

Non-Typeable Haemophilus influenzae (NTHi) strains cause a broad spectrum of diseases

- *Haemophilus influenzae* is a human opportunistic **bacterial pathogen**.
- It can cause **many different diseases** among the others:
  - bacterial **otitis media**
  - chronic obstructive pulmonary disease (COPD)
  - **invasive** diseases (as meningitides and septicemia).
- It is classified in six distinct capsular serotypes (a, b, c, d, e, and f) and one **non capsulated**.
- Nontypeable *Haemophilus influenzae* (NT*Hi*) **lack the** outer surface **capsule**.
- The huge **genomic variability** of NT*Hi* is exploited by the pathogen to adapt to the host environment, and makes hard the identification of **cross-protective** candidate antigens.
- NT*Hi* strains colonize the human nasopharynx, lungs, mucosal epithelium, or middle ear.
- NTHi invasive strains are fatal in more than 15% of the cases.

The aim of my PhD project is

- To investigate possible **relations** between **epidemiological features** with **genomes and genes functionality** features.
- The identification of **genes or their polymorphisms** that correlate with epidemiologic features.
- To reveal relevant peculiarities of the NT*Hi* **population structure.**

Past and present	First part:	**Data acquisition** extracting relevant information from sequencing data
	Second part:	Statistical analysis **setting up** the **analysis workflow**
Future	Third part:	Data modeling and biological discussion

Summary

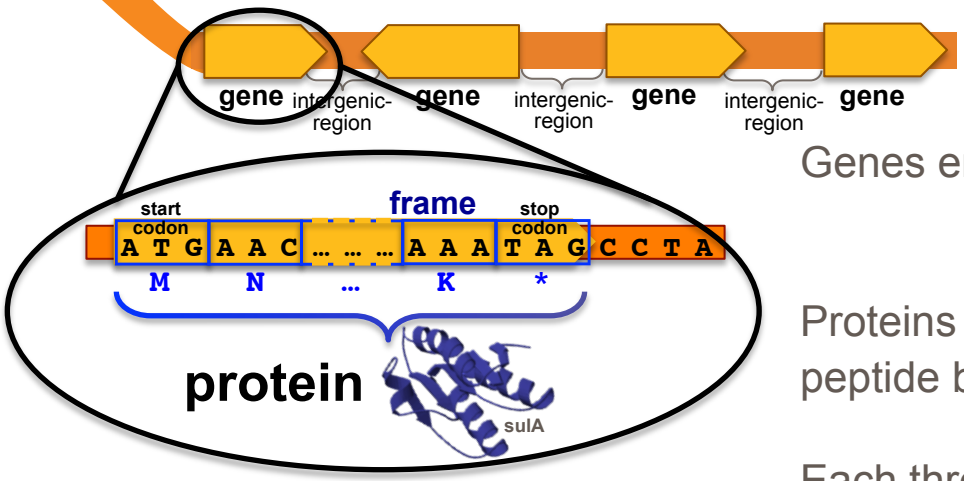
- **Introduction**: genetic fundamental concepts.
- **Data acquisition:** extracting relevant information from sequencing data.
- The **panel**: 103 NTHi bacterial samples from invasive diseases, otitis associated and carriage.
- Statistical **analysis** of data.
- **Preliminary results**.
- **Conclusions** and perspectives.

Information flows from genome to proteins through genes

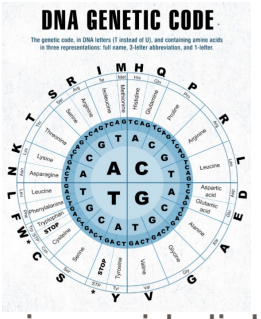
The genome of a bacterium is the whole DNA molecules of the bacterium.

It could be described by a unique long sequence of 4 nitrogen-containing nucleo-bases.

On this sequence there are genes spaced apart by intergenic regions.



Genes encode proteins.



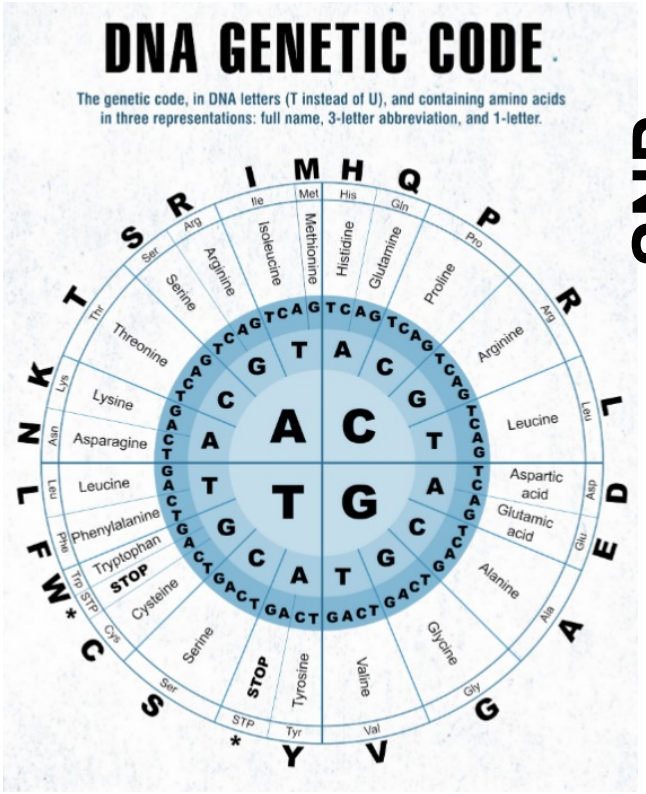
Proteins are polymers of amino-acids linked by peptide bonds to form the polypeptide chain.

Each three bases encode an amino-acid.

Some genome mutations modify the encoded proteins

Polymorphisms of genomes

A polymorphism is a mutation of a DNA sequence.



Original sequence in a gene

GAATATATGAATGAAGAG

E Y M N E E

SNPs

Mutated sequence with Single Nucleotide Polymorphism synonymous

GAATATATGAACGAAGAG

E Y M N E E

Mutated sequence with Single Nucleotide Polymorphism non-synonymous

GAATATATGAAGGAAGAG

E Y M K E E

INDELS

Mutated sequence with INSERTION frameSHIFT=+1

GAAATATATGAATGAAGAG

E I Y E * R

Mutated sequence with DELETION frameSHIFT=-1

GAATATTGAATGAAGAG

E Y * M K

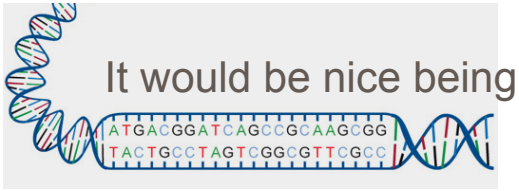
Mutated sequence with INDEL frameSHIFT=0

GAAATATATAATGAAGAG

E I Y N E E

Mutations and the genetic exchange can improve or inhibit the corresponding gene function and consequently increase or decrease the **fitness** of the corresponding bacterium.

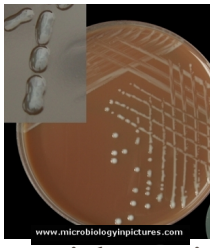
Whole genome sequence information is extracted from bacteria via sequencing process



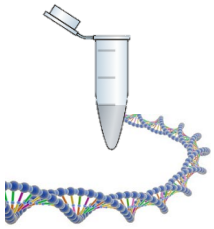
It would be nice being able to read the whole genome sequence “letter” by “letter”.

But it is not possible.

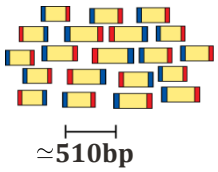
Sequencing technologies enable only few bases reading.



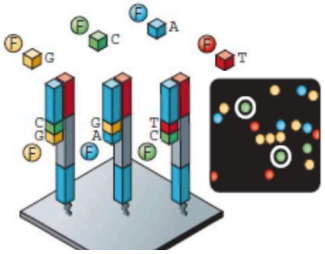
bacterial growing



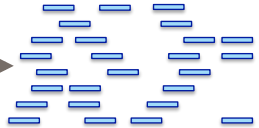
genomes extraction



DNA fragmentation

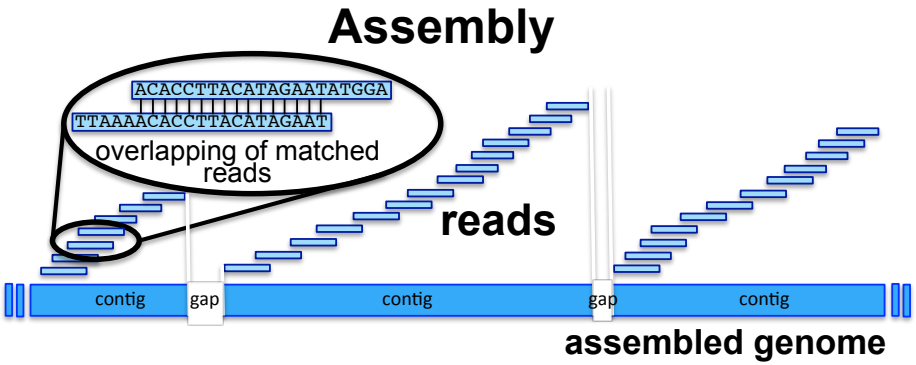


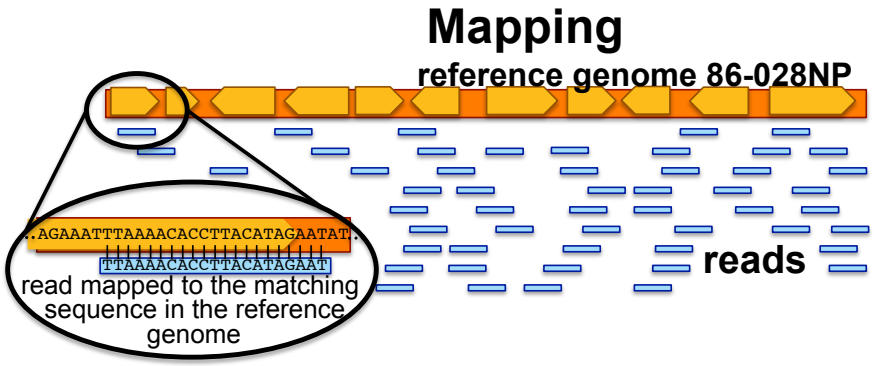
sequencing
on chip with fluorophores



reads library

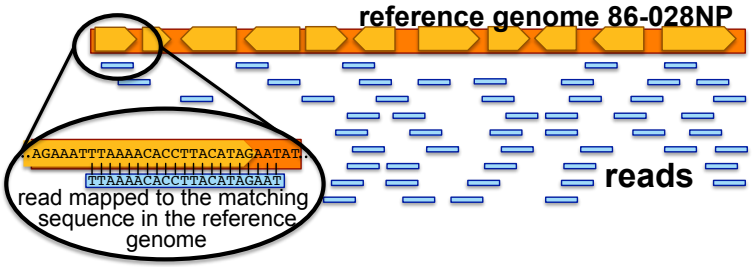
Two way to investigate the whole genome of a bacterium:





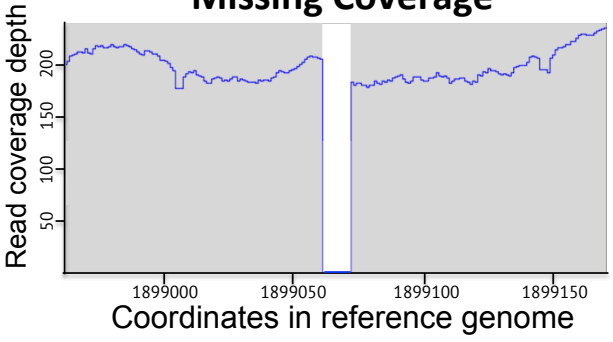
And reconstructed via mapping algorithms*

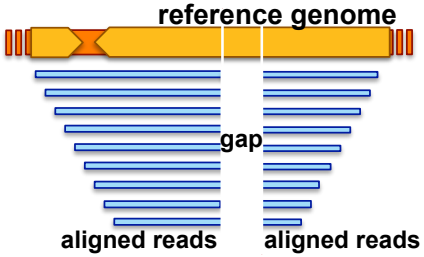
The reads of the isolates obtained from the sequencing processing, subject of our study, are **mapped to** a **reference closed and annotated** genome.



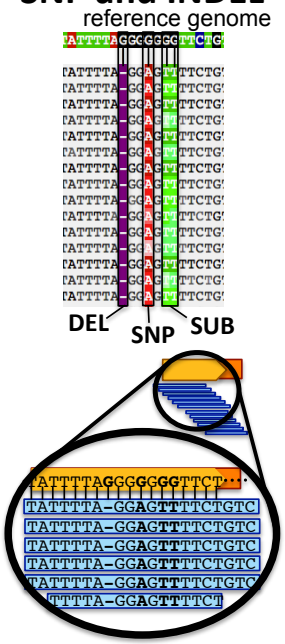
Polymorphisms' examples

Missing Coverage





SNP and INDEL



Polymorphisms considered

coverage
- **MC**: missing coverage (fraction of ORF or intergenic length)

INDELs
- **INS**: insertion length
- **DEL**: deletion length
- **SHIFTLEN**: length of insertions - deletions
- **SHIFT**: 1 if the ORF is frame-shifted

SNPs
- **SUB**: multiple base substitution length
- **RA_NONSYN**: count of non-synonymous SNPs over region length
- **RA_SYN**: count of synonymous SNPs over region length
- **RA_IG**: count of intergenic SNPs over region length

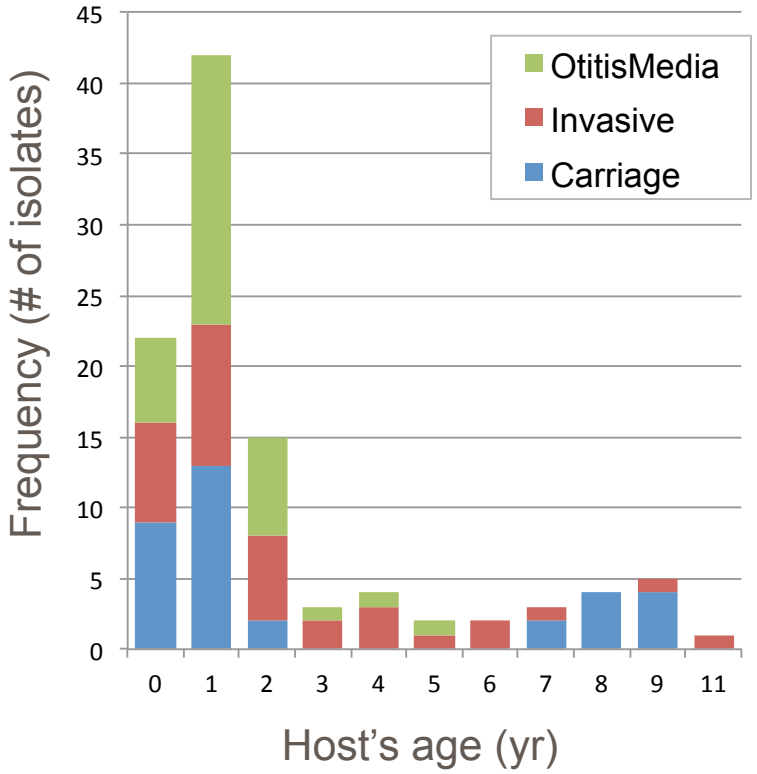
arrangements
- **JC**: new junction

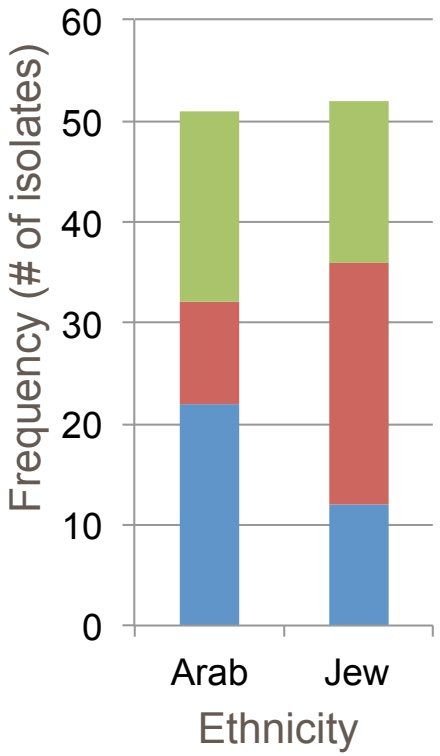
amplifications
- **AMP**: presence of duplications/amplifications

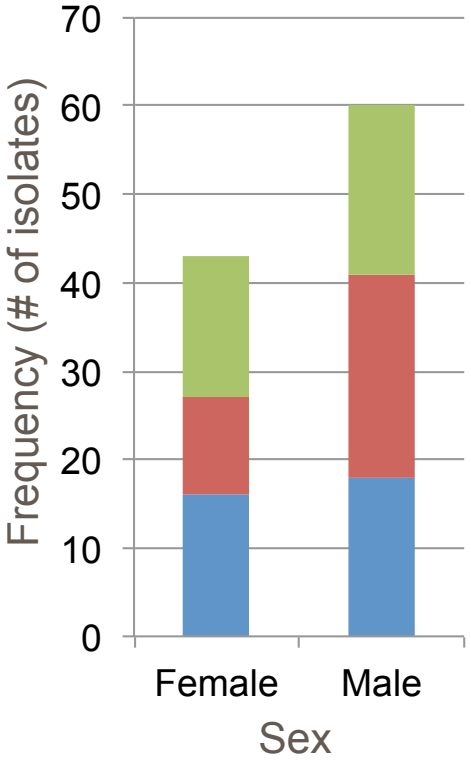
The panel: 103 genomes*

35 otitis media, 34 invasive disease, 34 healthy individuals

The strains were isolated from infants and children, from different ethnicity, males and females.



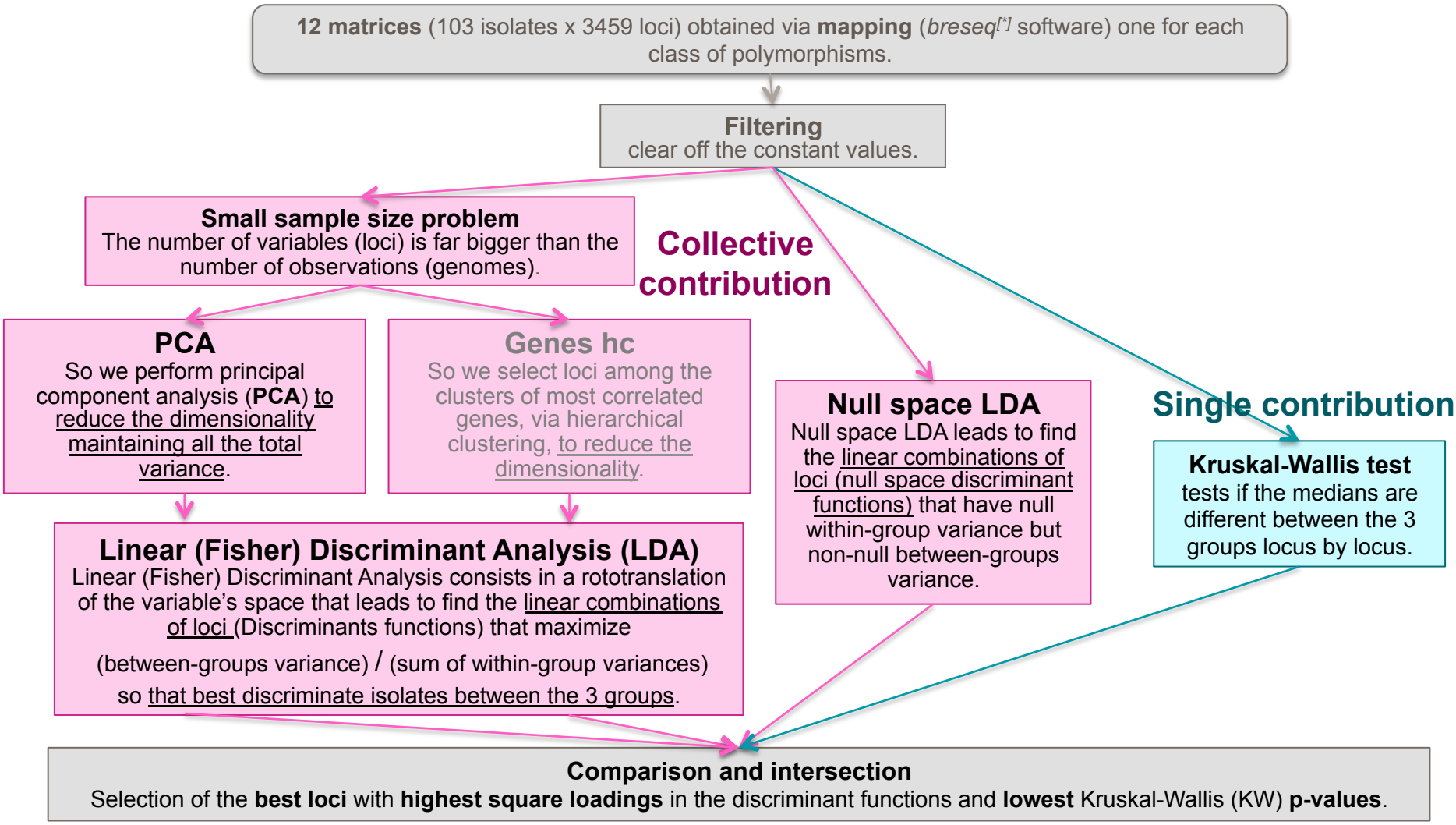




*isolated from a pediatric hospital in **Israel** between 2005-2012 by **Ron Dagan** and his lab group.

Statistical analysis workflow

Searching the linear combinations of loci that best discriminate genomes between 3 groups.



*_breseq_ software developed by **Deatherage, D.E.**, **Barrick, J.E.** (2014) *Methods Mol. Biol.* **1151**:165-188.

Collective effect of loci

Linear discriminant analysis via principal component analysis or via null space

I model n genomes as vectors x in a p-dimensional **vector space**.

Applying Linear Discriminant Analysis means to find v that **maximizes**

$$\frac{v^T S_B v}{v^T S_W v}$$

where:

$$S_W = \sum_{i=1}^{g} \sum_{x \in C_i} (x - \bar{x}_i)(x - \bar{x}_i)^T \qquad \bar{x}_i = \frac{1}{m_i} \sum_{x \in C_i} x$$

$$S_B = \sum_{i=1}^{g} m_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \qquad \bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$$

Solution: apply Lagrange multiplier method

Maximize $v^T S_B v$ with the constraint $v^T S_W v = 1$ is equal to resolve the eigenvalue problems:

$$S_W^{-1} S_B v = \lambda v$$

v are the Discriminant Functions.

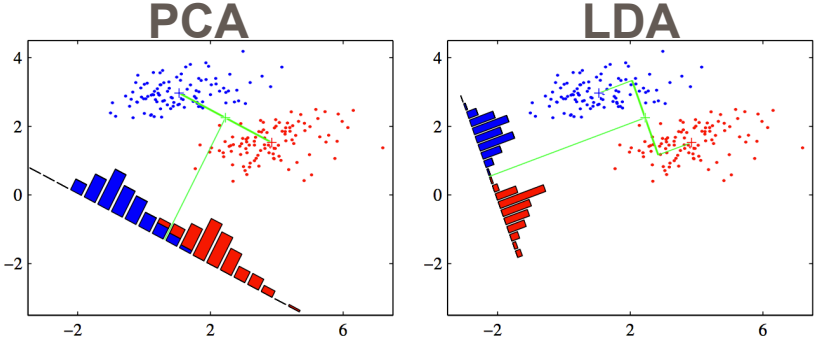
S_W must be invertible so its rank must be p, but

$$rank(S_W) \leq rank(X) \leq n \ll p$$

small sample size problem

Two solutions

1. via **dimensionality reduction**
   Principal Component Analysis (PCA)



2. via **null space LDA**
   the discriminant function are v such that
   $v^T S_W v = 0$ and $v^T S_B v \neq 0$

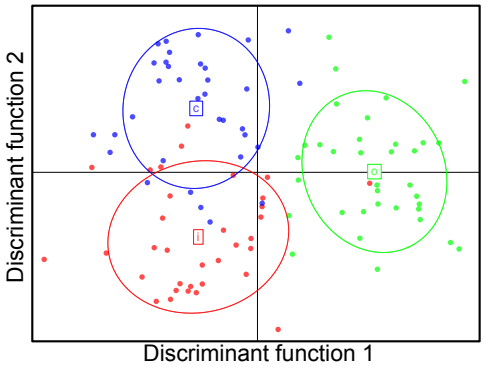
Discriminant Functions

Linear combinations of loci that discriminate between the three groups

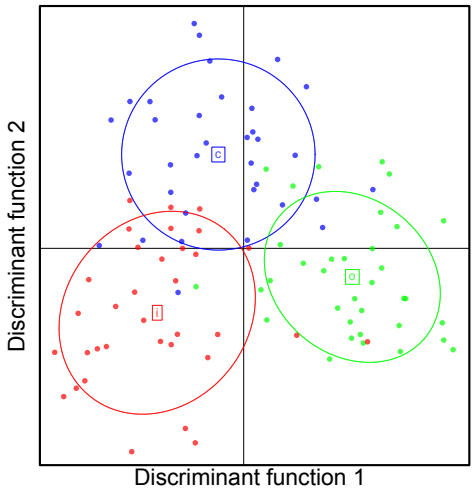
The two Discriminant functions that best discriminate between **invasive (i)**, **otitis** associated **(o)** and **carriage (c).**

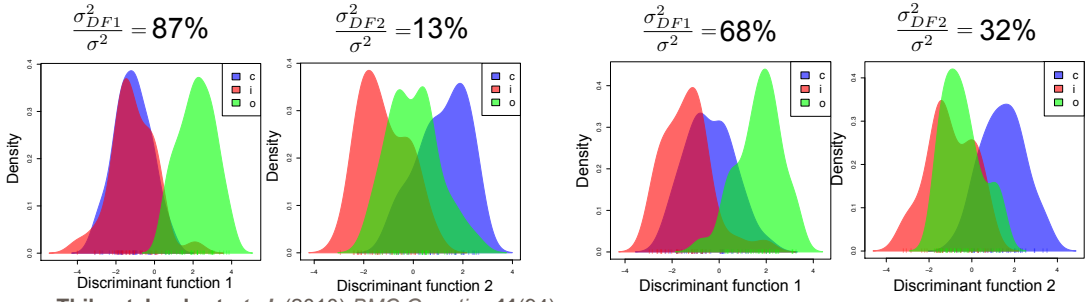
- Applying the combination of PCA and LDA[*]

Missing coverage (MC)



Frame shift (SHIFT)

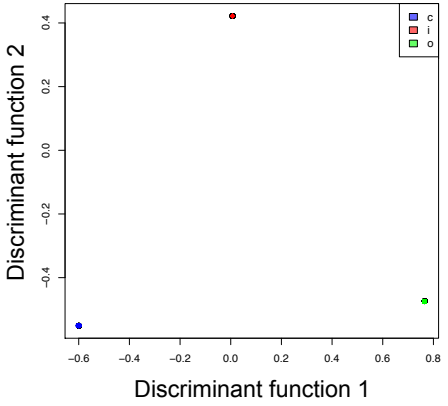




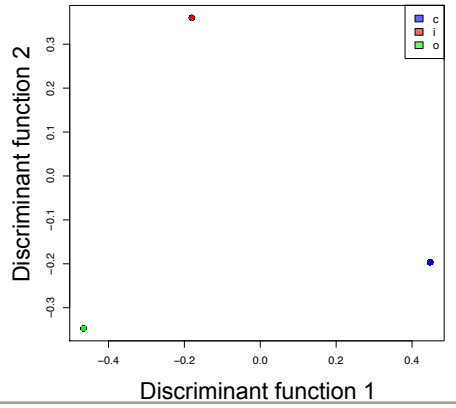
Thibaut Jombart *et al.* (2010) *BMC Genetics* **11**(94).

- Applying null space LDA

Missing coverage (MC)



Frame shift (SHIFT)



Single effect

Kruskal-Wallis test locus by locus

Kruskal-Wallis is a non parametric test based on ranks values.

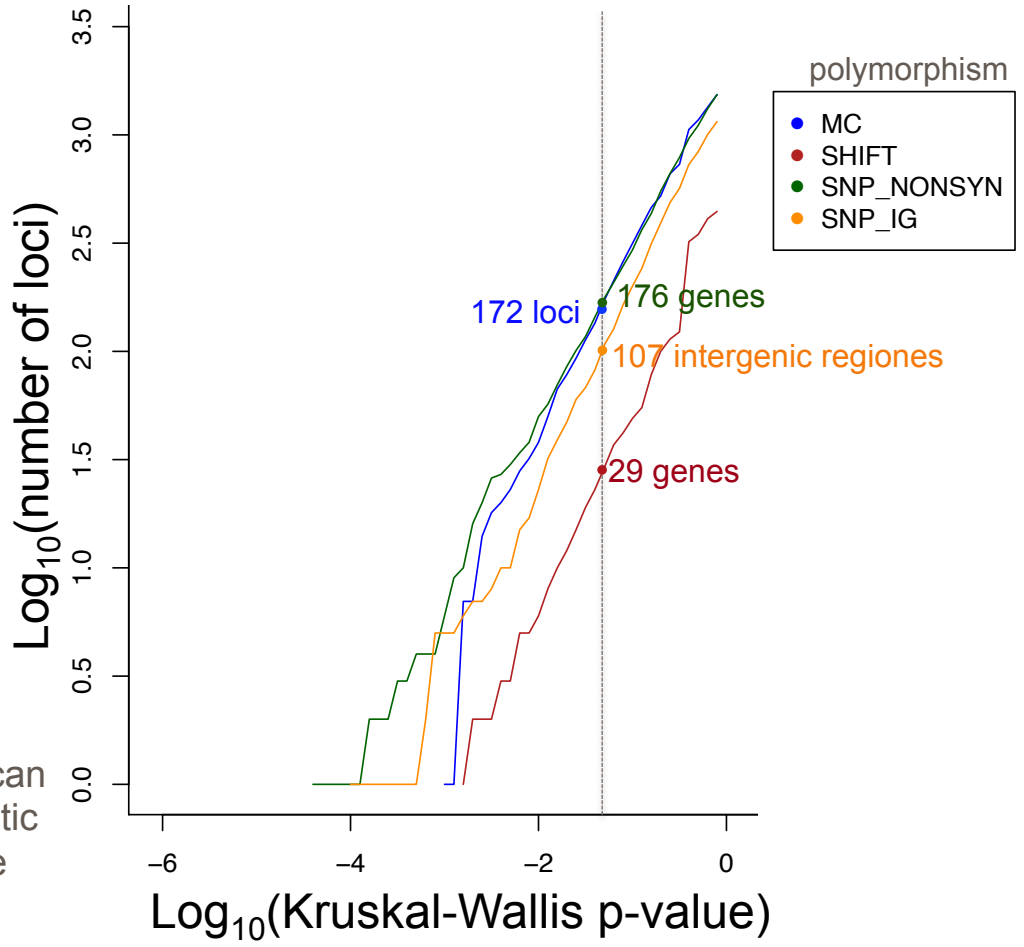
Null hypothesis:
the distributions among the g groups are the same,

equivalently $k \sim \chi^2_{g-1}$ with

$$k = \frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i} \frac{(r_{ij}-\bar{r})^2}{(\sum_{i=1}^{g} n_i)-1}}$$

r_{ij} rank of the j-th element of the i-th group
n_i number of elements in the i-th group

Applying the KW **p-value threshold** of **0.05** we can determine the list of loci having a pattern of genetic variation putatively associated to one of the three groups of isolates (invasive, otitis or carriage).



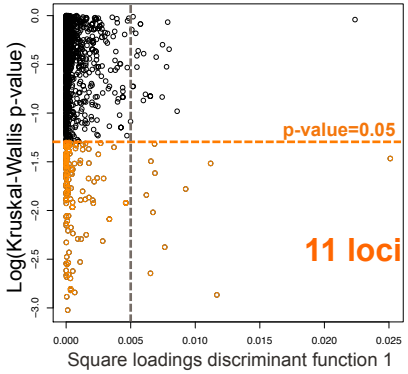
Comparison and intersection

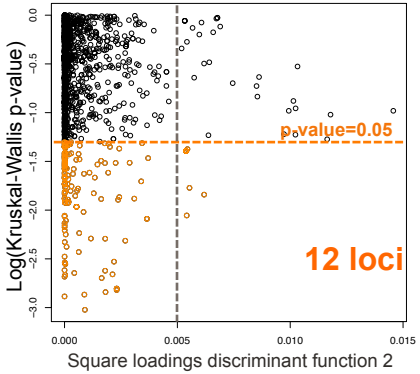
Between Kruskal-Wallis p-values and Discriminant Functions square loadings

Intersecting the collective and single approaches I want to select those loci that significantly contribute both individually and dominating a collective effect.

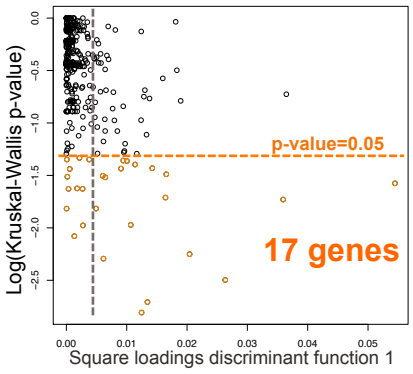
- Applying LDA via PCA

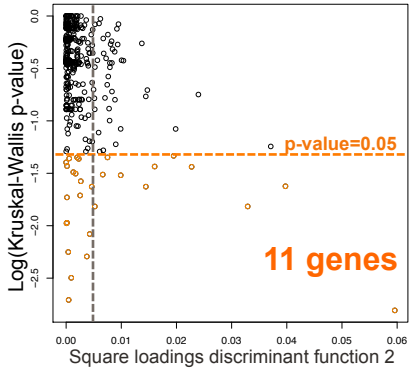
Missing coverage (MC)





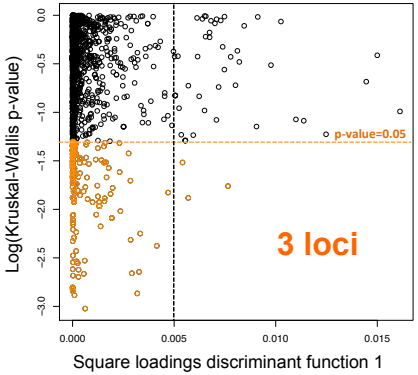
Frame shift (SHIFT)

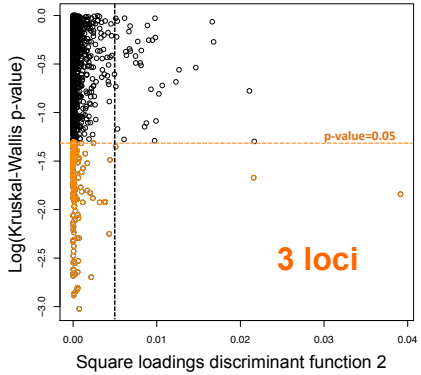




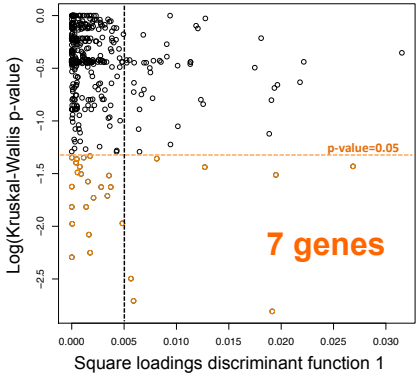
- Applying null space LDA

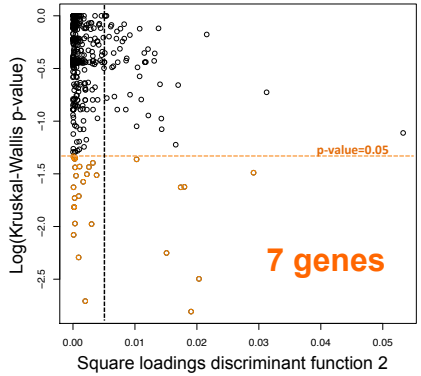
Missing coverage (MC)





Frame shift (SHIFT)

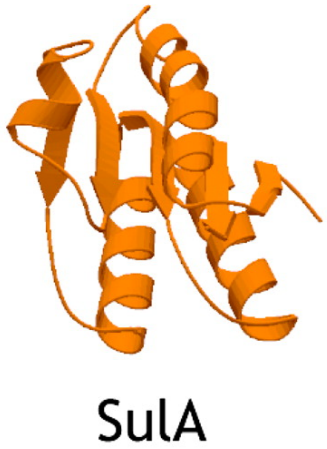




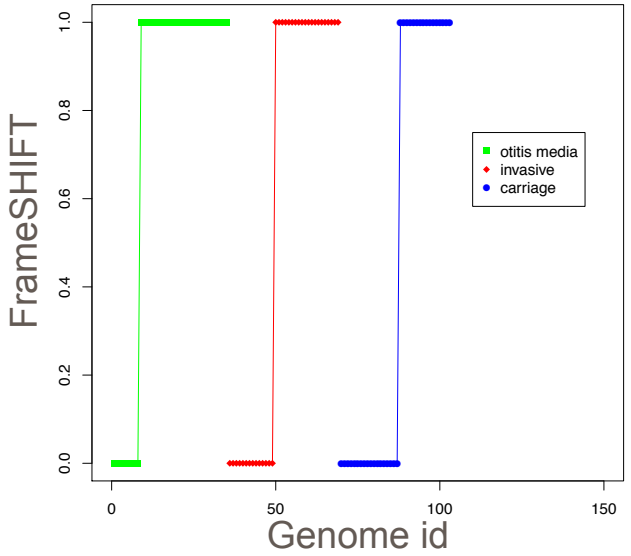
The selected loci are the ones with the highest square loading DFs and the smallest KW test p-values.

Example of one selected locus

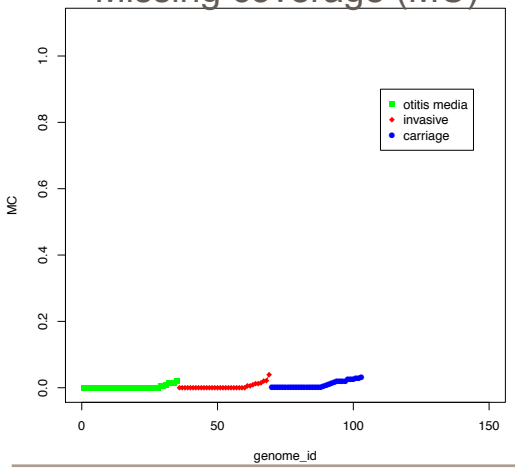
SulA gene is more often frame-shifted among otitis genomes than carriage or invasive ones.



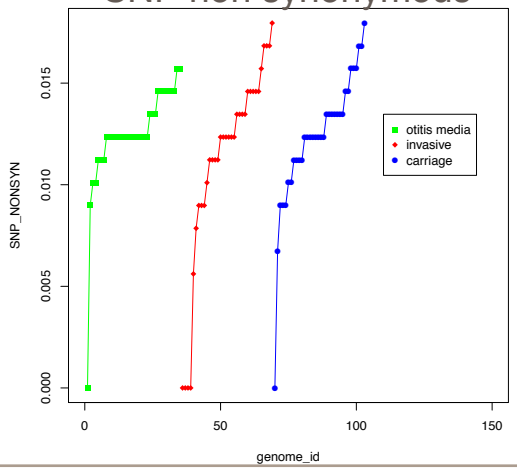
NTHI1379 cell division inhibitor SulA (sulA)
start:**1308002** stop:**1308892**
KW p-value: **0.0365**
Selected by: LDA via PCA, null space LDA



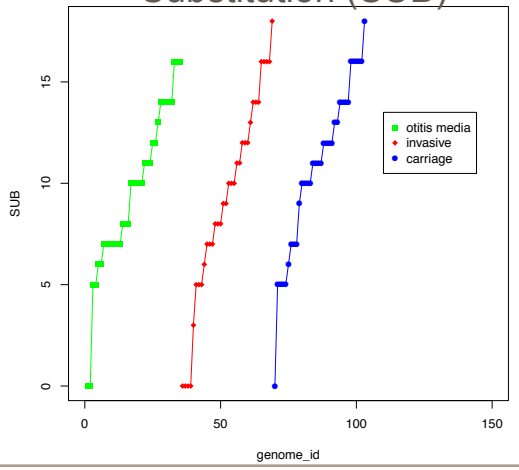
Missing coverage (MC)



SNP non synonymous



Substitution (SUB)



Conclusion and perspective

- We enlarged the analysis of NTHi whole genomes to a variety of polymorphisms (12 genetic variability marker), including INDELs and rearrangements.
- This analysis allowed us to identify polymorphisms in genic and inter-genic regions of NTHi genomes as good candidates to discriminate between being associated with otitis media, invasive disease or carriage.
- **NTHi genetic diversity between carriage and disease isolates is limited, making difficult the identification of genetic features correlated with disease outcome.**
- **Anyway, signals that possibly identify these signatures can be determined by our analysis (the example of SulA locus).**

Next steps will be:

- To study the power prediction of the resulting discriminant functions.
- To overcome the small sample size problem enlarging the sample size including other genome sequences that will probably be available during the next year.
- To apply this study to other species with higher number of genomes in the public domain. For example, in Neisseria meningitidis serogroup b (MenB) epidemiology, evident genetic differences between carriage and invasive strains are known, so it could be interesting to validate my procedure through this species.

Thanks to

- Gabriella De Angelis,
- Monica Moschioni,
- Silvia Guidotti,
- Giulia Torricelli,
- Nicola Pacchiani,
- Mariagrazia Pizza,
- Stefano Censini,
- Marco Soriani,
- Alessandro Muzzi,
- Prof. Dino Leporini
- Fabiola Blengio
- Andrea Cavallone
- Mattia Dalsass
- Federica Martina

And all of you for your attention