Genomics and Clustering

Mara Barucco, PhD Student

Pisa, 17Set2014

UNIVERSITÀ DI PISA

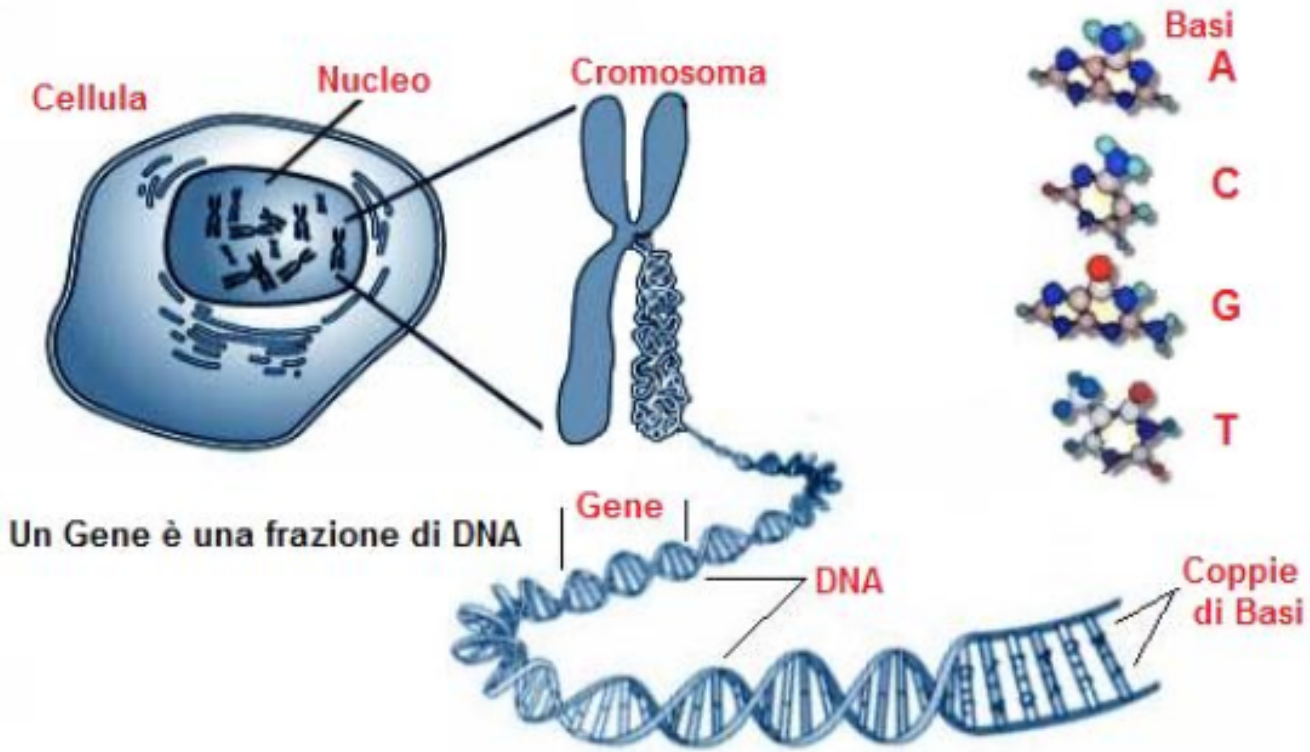
NOVARTIS

Sommario

- Parte I: Genomica e Epidemiologia
  - Introduzione alla genomica e alla ricerca sui vaccini
  - Epidemiologia e studio sulla struttura di popolazione di un patogeno
- Parte II: Clustering
  - Metodi di clustering: K-means, PCA, gerarchici
- Parte III: Applicazioni
  - Applicazione del clustering allo studio sulla struttura di popolazione di un patogeno
  - Sviluppi futuri

Il ruolo della genomica

- La genomica studia la struttura, il contenuto, la funzione e l'evoluzione del genoma degli organismi viventi, ovvero di tutto il patrimonio genetico.



Il ruolo della genomica

- La genomica studia la struttura, il contenuto, la funzione e l'evoluzione del genoma degli organismi viventi, ovvero di tutto il patrimonio genetico.

- Nasce negli anni '80, con la ricerca per il sequenziamento di interi genomi:
  - Nel 1980 viene sequenziato il genoma del virus fago Φ–X174.
  - Nel 1995 Fleischmann R.D. *et al.* sequenziano il primo batterio, Haemophilus influenzae, un genoma di notevoli dimensioni (1,8 10^6 basi).
  - Dal 1990 al 2003 molti laboratori lavorano per sequenziare per la prima volta il genoma umano (3,3 10^9 basi).

UNIVERSITÀ DI PISA NOVARTIS

Il ruolo della genomica

- La genomica studia la struttura, il contenuto, la funzione e l'evoluzione del genoma degli organismi viventi, ovvero di tutto il patrimonio genetico.

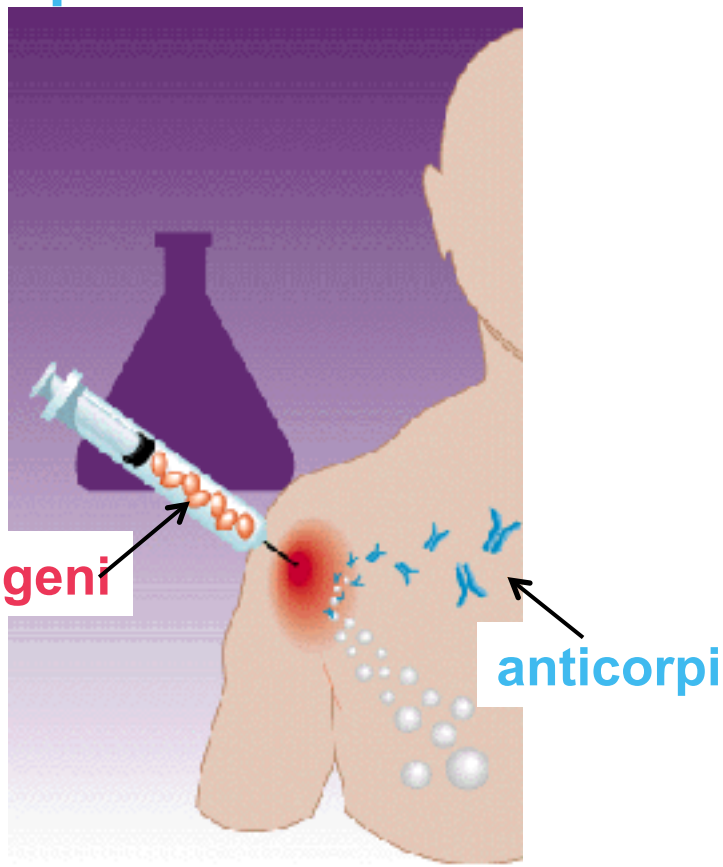
- Nasce negli anni '80, con la ricerca per il sequenziamento di interi genomi:
  - Nel 1980 viene sequenziato il genoma del virus fago Φ–X174.
  - Nel 1995 Fleischmann R.D. *et al.* sequenziano il primo batterio, Haemophilus influenzae, un genoma di notevoli dimensioni (1,8 10^6 basi).
  - Dal 1990 al 2003 molti laboratori lavorano per sequenziare per la prima volta il genoma umano (3,3 10^9 basi).

- Oggi il sequenziamento di un genoma batterico richiede qualche ora, quello di un genoma umano richiede qualche giorno.

UNIVERSITÀ DI PISA NOVARTIS

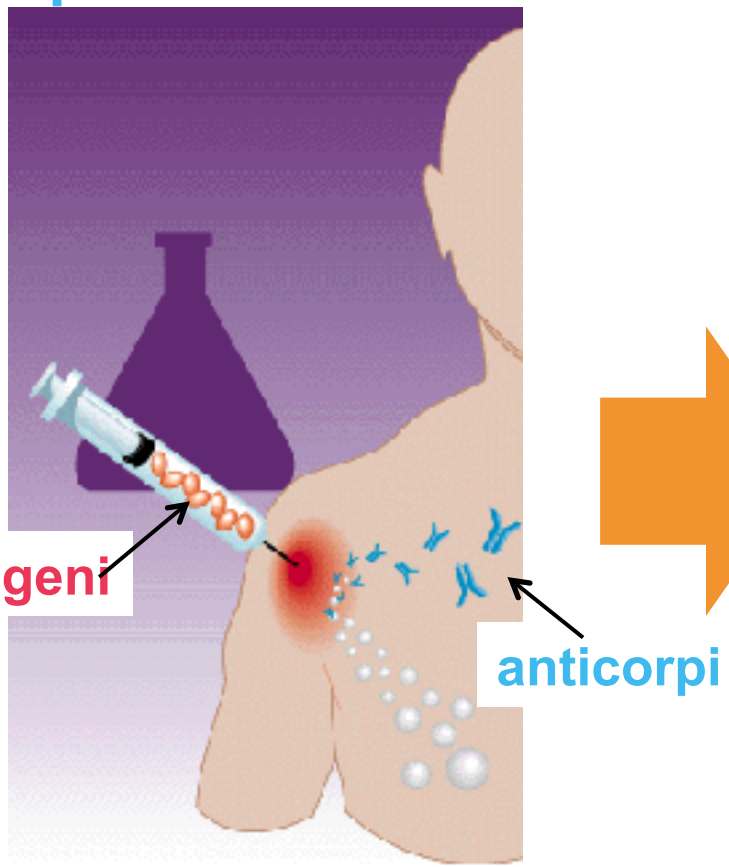
La ricerca sui vaccini

Il **vaccino** genera
la **risposta immunitaria**

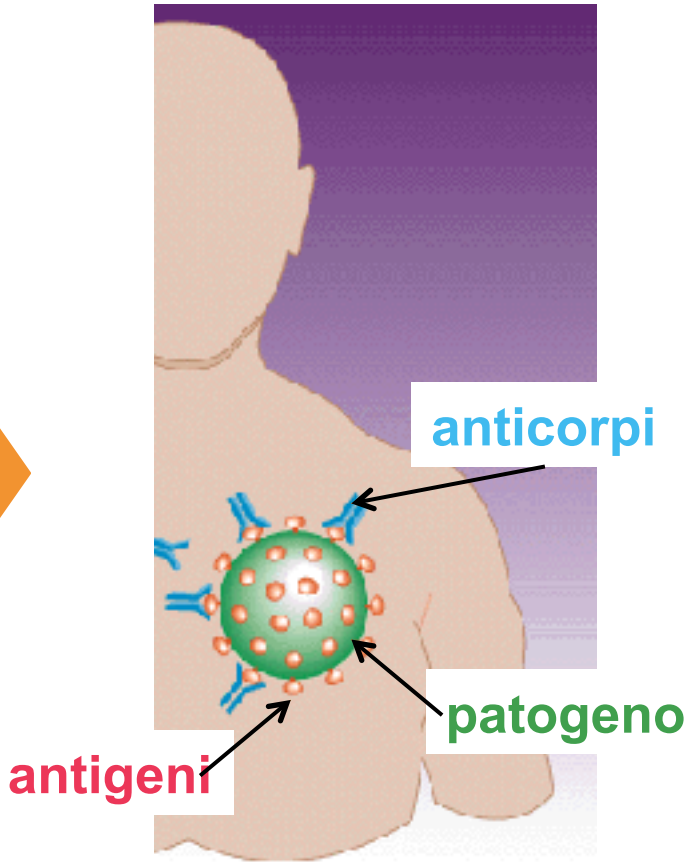


La ricerca sui vaccini

Il **vaccino** genera
la **risposta immunitaria**

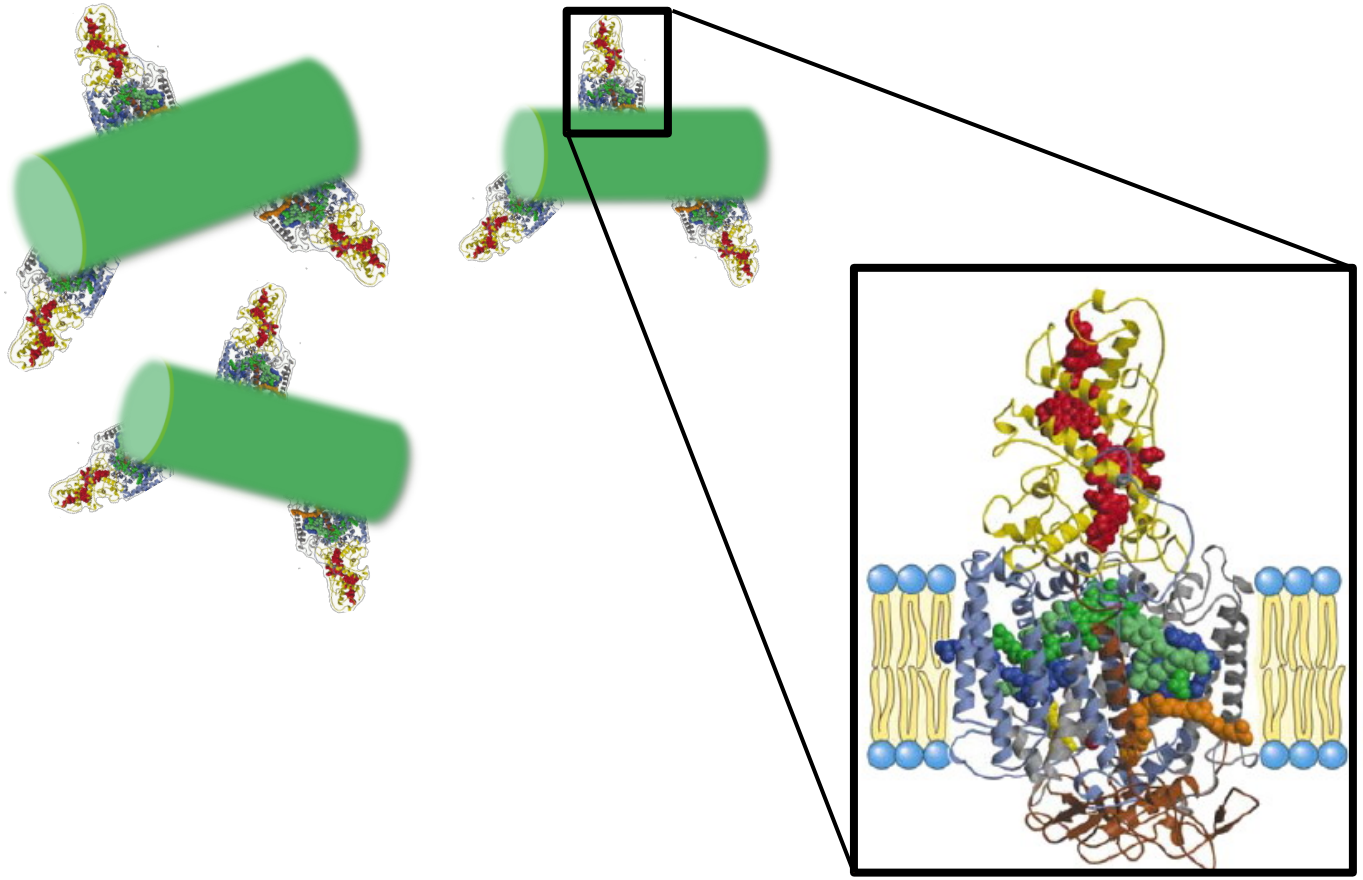


Contatto col **patogeno**



La ricerca sui vaccini

Le proteine di superficie dei patogeni sono dei buoni candidati per i vaccini.



Il ruolo della genomica nella ricerca sui vaccini: la reverse vaccinology

Com'è cambiato il modo di fare ricerca sui vaccini?

In Passato

- Microbi (virus o batteri) attenuati, inattivati o morti, purificati.
- Tossine o parti di microbi ottenuti con la tecnica del DNA ricombinante

Oggi

Il ruolo della genomica nella ricerca sui vaccini: la reverse vaccinology

Com'è cambiato il modo di fare ricerca sui vaccini?

In Passato

- Microbi (virus o batteri) attenuati, inattivati o morti, purificati.
- Tossine o parti di microbi ottenuti con la tecnica del DNA ricombinante

Oggi

- Dal genoma individuare le proteine che il microbo "produce" (esprime).
- Tra queste riconoscere le proteine di superficie e verificarne l'immunogenicità.

Il ruolo della genomica nella ricerca sui vaccini: la reverse vaccinology

Com'è cambiato il modo di fare ricerca sui vaccini?

In Passato

- Microbi (virus o batteri) attenuati, inattivati o morti, purificati.
- Tossine o parti di microbi ottenuti con la tecnica del DNA ricombinante

Problemi

- Proteine con sequenza spesso variabile, difficili da esprimere e/o purificare in grandi quantità.
- Solo pochi antigeni venivano presi in considerazione: quelli più abbondanti.
- Più complesso controllarne e limitarne la virulenza.

Oggi

- Dal genoma individuare le proteine che il microbo "produce" (esprime).
- Tra queste riconoscere le proteine di superficie e verificarne l'immunogenicità.

Il ruolo della genomica nella ricerca sui vaccini: la reverse vaccinology

Com'è cambiato il modo di fare ricerca sui vaccini?

In Passato

- Microbi (virus o batteri) attenuati, inattivati o morti, purificati.
- Tossine o parti di microbi ottenuti con la tecnica del DNA ricombinante

Problemi

- Proteine con sequenza spesso variabile, difficili da esprimere e/o purificare in grandi quantità.
- Solo pochi antigeni venivano presi in considerazione: quelli più abbondanti.
- Più complesso controllarne e limitarne la virulenza.

Oggi

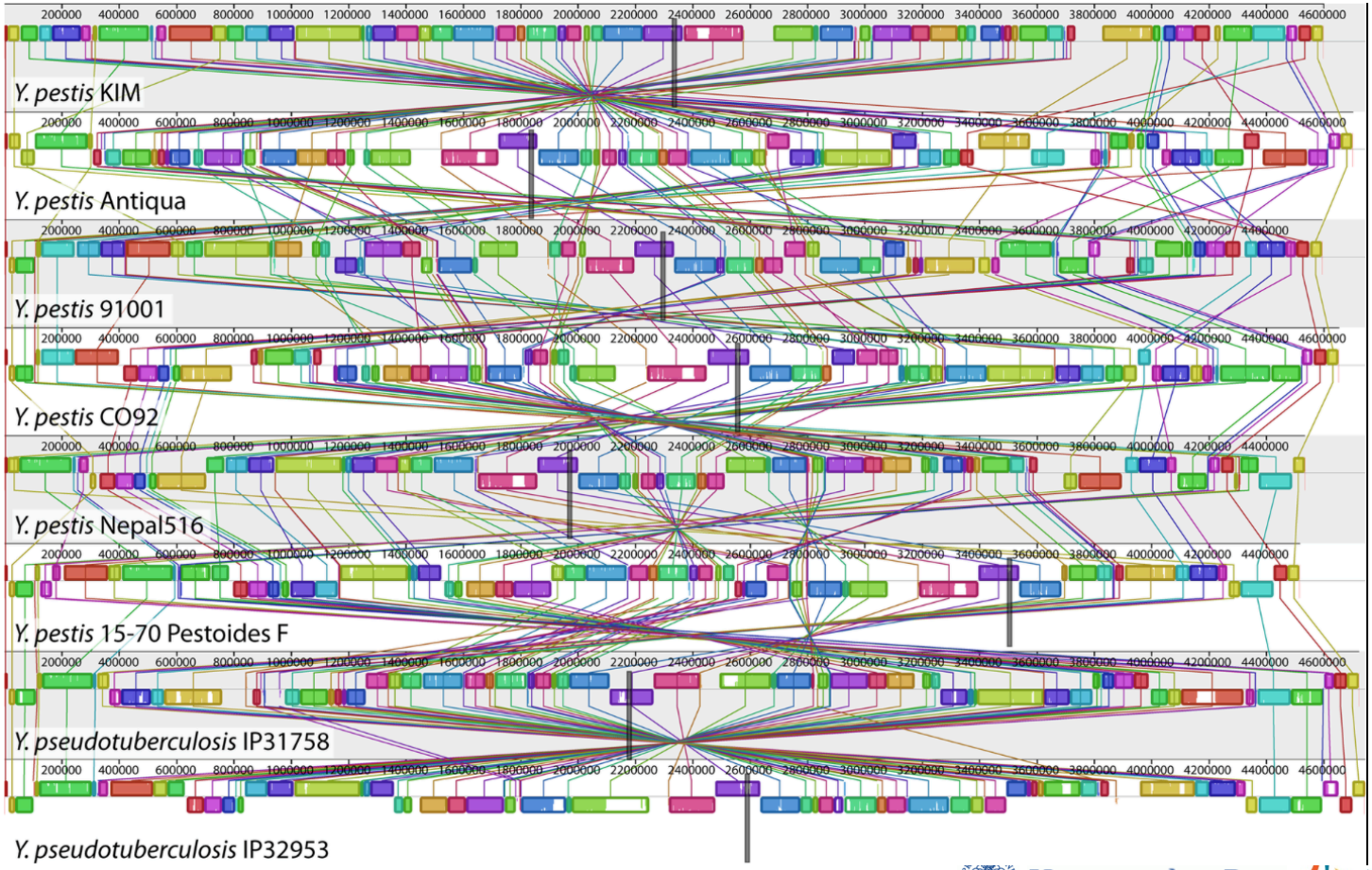
- Dal genoma individuare le proteine che il microbo "produce" (esprime).
- Tra queste riconoscere le proteine di superficie e verificarne l'immunogenicità.

Vantaggi

- Individuare un ampio spettro di proteine candidate ad essere antigeni (anche quelle che prima risultavano mascherate magari perchè poco abbondanti).
- Possibilità di trovare vaccini anche per specie difficilmente "coltivabili" in laboratorio.

L'epidemiologia

- Identificare e caratterizzare la struttura di una popolazione diventa quindi una risorsa chiave per lo sviluppo di nuovi vaccini.



Esempio di studi recenti: Clustering nello studio di popolazioni batteriche

De Chiara *et al.* hanno studiato i metodi evolutivi di un patogeno: Non-Typeable *Haemophilus influenzae* (NT*Hi*)

Dati:

- Librerie di sequenze di campioni con diverse origini geografiche e che portano a diverse malattie

n° genomi	Origine	Malattia
32	Finlandia	Portatori
16	Finlandia	Otitis Media
19	Spagna	COPD
22	Tutto il mondo	Varie
8	Genomi pubblici	Varie

Genome sequencing of disease and carriage isolates of nontypeable Haemophilus influenzae identifies discrete population structure, De Chiara et al., PNAS, **2014**

Esempio di studi recenti: Clustering nello studio di popolazioni batteriche

De Chiara *et al.* hanno studiato i metodi evolutivi di un patogeno: Non-Typeable *Haemophilus influenzae* (NT*Hi*)

Dati:

- Librerie di sequenze di campioni con diverse origini geografiche e che portano a diverse malattie

n° genomi	Origine	Malattia
32	Finlandia	Portatori
16	Finlandia	Otitis Media
19	Spagna	COPD
22	Tutto il mondo	Varie
8	Genomi pubblici	Varie

Obiettivi:

- Studiare come si evolve e differenzia il genoma di questo batterio.
- Individuare eventuali correlazioni tra il tipo di malattia causata dal patogeno e i suoi caratteri genetici.

Genome sequencing of disease and carriage isolates of nontypeable Haemophilus influenzae identifies discrete population structure, De Chiara et al., PNAS, **2014**

Esempio di studi recenti: Clustering nello studio di popolazioni batteriche

De Chiara *et al.* hanno studiato i metodi evolutivi di un patogeno: Non-Typeable *Haemophilus influenzae* (NT*Hi*)

Dati:

- Librerie di sequenze di campioni con diverse origini geografiche e che portano a diverse malattie

n° genomi	Origine	Malattia
32	Finlandia	Portatori
16	Finlandia	Otitis Media
19	Spagna	COPD
22	Tutto il mondo	Varie
8	Genomi pubblici	Varie

Obiettivi:

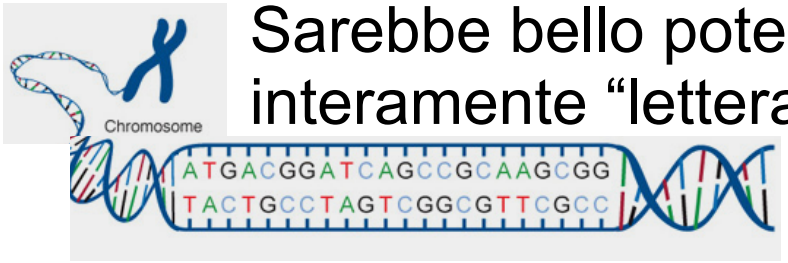
- Studiare come si evolve e differenzia il genoma di questo batterio.
- Individuare eventuali correlazioni tra il tipo di malattia causata dal patogeno e i suoi caratteri genetici.

Metodi:

- Stabilire un grado di somiglianza tra i diversi campioni (o una distanza) per poter raggruppare tra loro quelli più simili o per poter "ri"-costruire un albero filogenetico che descriva una possibile storia evolutiva del genoma del patogeno.

Genome sequencing of disease and carriage isolates of nontypeable Haemophilus influenzae identifies discrete population structure, De Chiara et al., PNAS, **2014**

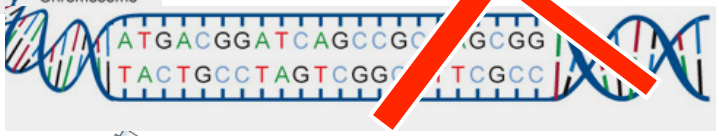
I dati ottenuti dal sequenziamento: dal sequencing all'assembly

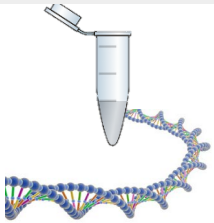


Sarebbe bello poter leggere la sequenza del DNA interamente "lettera" per "lettera"

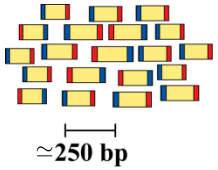
I dati ottenuti dal sequenziamento: dal sequencing all'assembly

Sarebbe bello poter leggere la sequenza del DNA interamente "lettera" per "lettera"

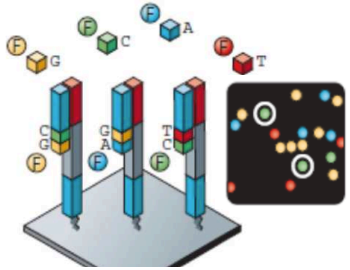




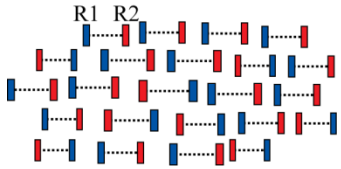
estrazione del DNA



frammentazione DNA



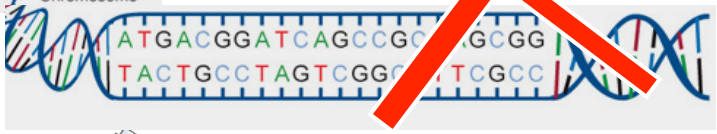
sequenziamento
su chip e con flourofori

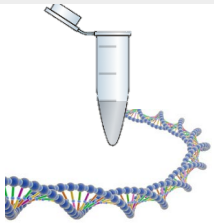


librerie di sequenze
accoppiate

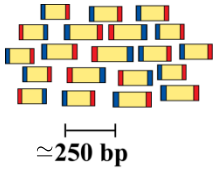
I dati ottenuti dal sequenziamento: dal sequencing all'assembly

Sarebbe bello poter leggere la sequenza del DNA interamente "lettera" per "lettera"

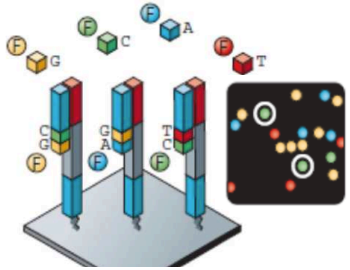




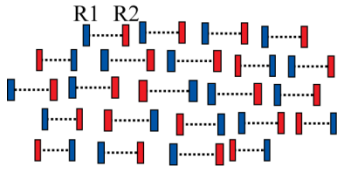
estrazione del DNA



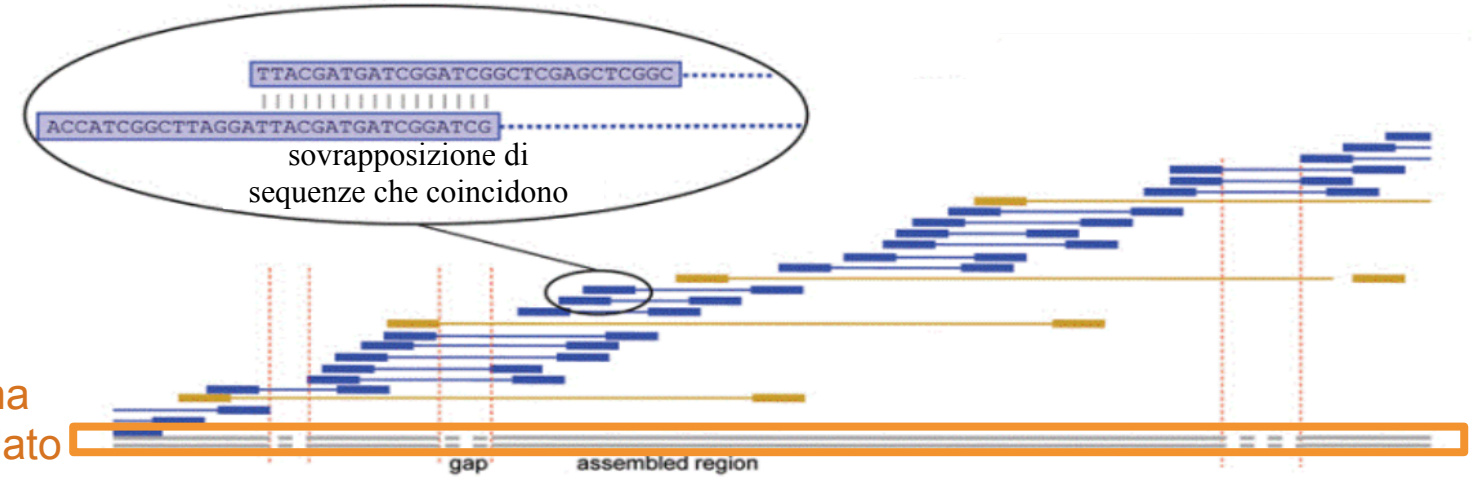
frammentazione DNA



sequenziamento
su chip e con flourofori



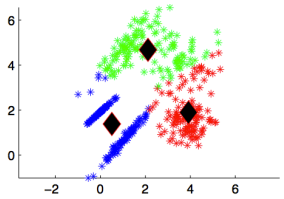
librerie di sequenze
accoppiate

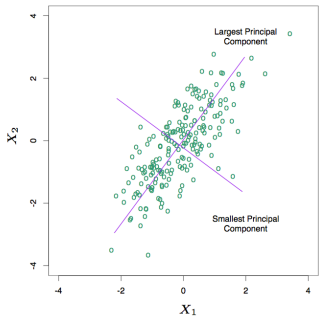


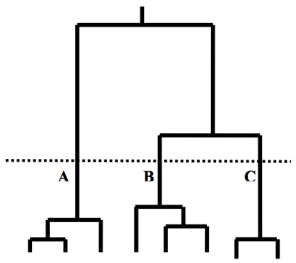
genoma
assemblato

UNIVERSITÀ DI PISA NOVARTIS

Metodi di clustering







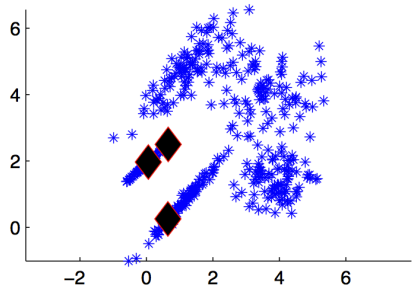
- K-means
- Riduzione dimensionale
  - Principal Component Analysis
  - Spectral Clustering
- Gerarchici
  - Agglomerativi
  - Divisivi

K-means

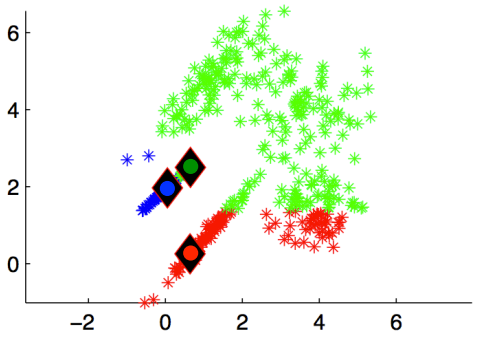
Consente di individuare esattamente K cluster sulla base della distanza euclidea tra i punti.

- Dati n punti $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$
- Scelgo a caso K punti in $\mathbb{R}^d$

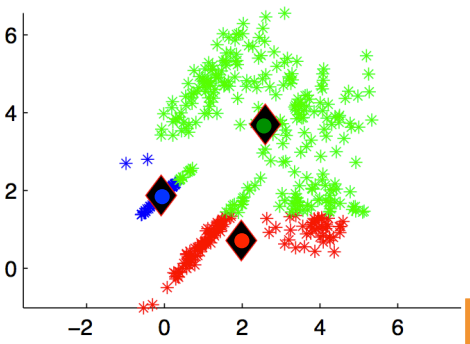
Input



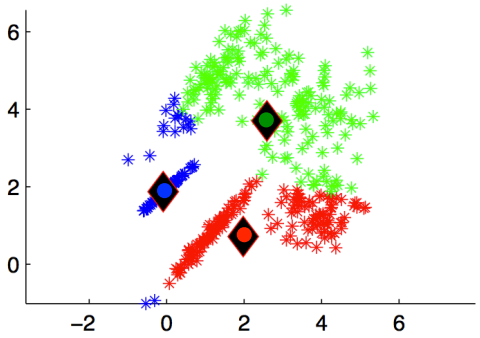
1)



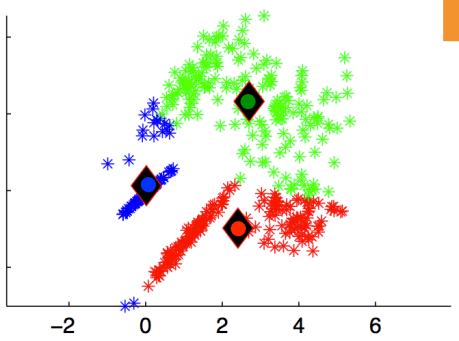
2)



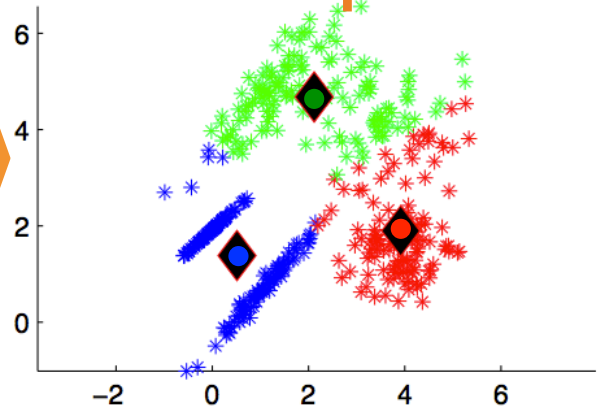
3)



4)



Output



K-means: l'algoritmo

Input:

- n punti $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$
- numero di cluster K

1. Inizializzare in modo casuale i centri $m_1^{(0)}, \ldots, m_K^{(0)}$
2. Iterare fino a convergenza i seguenti passi:
   a. Assegnare ciascun punto al centro a lui più vicino, definendo i cluster $C_1^{(i+1)}, \ldots, C_K^{(i+1)}$

$$X_s \in C_k^{(i+1)} \iff \|X_s - m_k^{(i)}\|^2 \leq \|X_s - m_l^{(i)}\|^2 \quad \forall\, l = 1, \ldots, K$$

   b. Calcolare i nuovi baricentri dei cluster

$$m_k^{(i+1)} = \frac{1}{\left|C_k^{(i+1)}\right|} \sum_{s \in C_k^{(i+1)}} X_s$$

Output: dopo M passi i cluster $C_1^{(M)}, \ldots, C_K^{(M)}$

K-means

- Con questa procedura cerco di minimizzare:

$$\min_{C_1,\ldots,C_K} \sum_{k=1}^{K} \sum_{i \in C_k} \|X_i - m_k\|^2 \qquad \text{con } m_k = \frac{1}{|C_k|} \sum_{i \in C_k} X_i$$

- È equivalente a cercare la configurazione di cluster che minimizza le distanza tra punti dello stesso cluster:

$$\min_{C_1,\ldots,C_K} \sum_{k=1}^{K} \frac{1}{|C_k|^2} \sum_{i \in C_k, j \in C_k} \|X_i - X_j\|^2$$

Osservazione: Ad ogni passo della procedura queste funzioni decrescono.

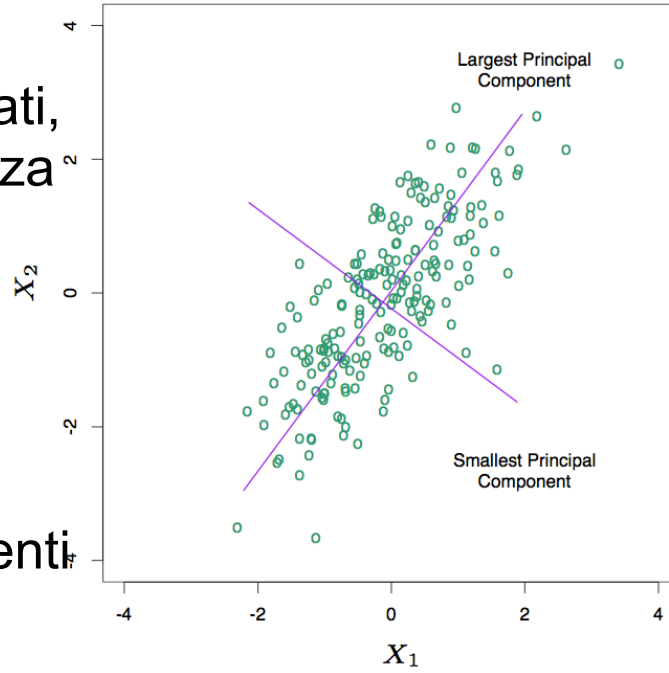
Problema: la procedura non garantisce di trovare il minimo assoluto.

→ eseguire più volte con diverse inizializzazioni dei centri.

Principal Component Analysis (PCA)

Scelta di nuovi assi che massimizzano la capacità di discriminare i punti lungo le nuove componenti.

- Dati n punti $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ ogni rotazione preserva la configurazione dei dati, vogliamo individuare la rotazione che massimizza la varianza dei punti lungo i nuovi assi.
- La prima componente principale è la direzione che massimizza la varianza, la componente principale successiva è ortogonale alla prima e massimizza la varianza residua, …
- Dei nuovi assi considero solo le prime componenti



ANALOGIA

- Nel caso 3D , dato un insieme di punti nello spazio, applicare la PCA significa individuare l'asse principale di rotazione di questo sistema.

Statistics, Data Mining, and Machine Learning in Astronomy, Zeljko Ivezic et al., **2014**

Principal Component Analysis (PCA)

Input: date n osservazioni ciascuna caratterizzata da d caratteri

$$X_1, X_2, \ldots, X_n \in \mathbb{R}^d \qquad \mathbb{X} = \begin{pmatrix} \text{——} X_1 \text{——} \\ \text{——} X_2 \text{——} \\ \vdots \\ \text{——} X_n \text{——} \end{pmatrix}$$

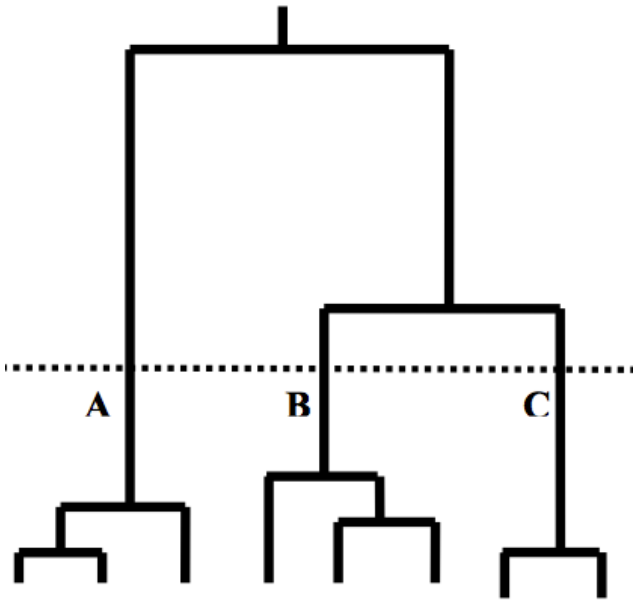
- Da questa matrice ottengo la matrice A centrata togliendo da ogni colonna la sua media
- La covarianza di A vale: $C_A = \frac{1}{n-1} A^T A$
- Cerco la rotazione R che porta il primo asse a coincidere con l'asse lungo il quale i punti hanno la massima varianza.
  Voglio massimizzare $r_1^T C_A r_1$ con la normalizzazione $r_1^T r_1 = 1$

$$\phi(r_1, \lambda_1) = r_1^T C_A r_1 - \lambda_1 (r_1^T r_1 - 1)$$

  da cui ottengo $C_A r_1 = \lambda_1 r_1$ cioè cerco il massimo autovettore di C_A
- Iterando questa procedura cerco $Y = AR$
  tale che $C_Y = R^T A^T A R = R^T C_A R$ risulti diagonale

Cluster gerarchici

- Vengono solitamente rappresentati come alberi:
  - il nodo “di partenza” è detto radice
  - tutti i dati iniziali rappresentano dei nodi terminali e sono detti foglie
  - da un nodo può uscire un solo collegamento verso l’alto (ciascun nodo ha un solo “genitore”)
- La lunghezza dei rami che separano due nodi deve essere indicativa della distanza tra gli elementi rappresentati da quei due nodi



Core Concepts in Data Analysis, Boris Mirkin, **2010**

Cluster gerarchici: agglomerativi

A partire dai dati si raggruppano via via gli elementi considerando ogni volta quelli più vicini.

1. Dati n punti $X_1, \ldots, X_n$ e una matrice di distanze reciproche, ciascun punto costituisce un cluster a sè
2. Si controlla quali di questi cluster sono più vicini e si "fondono" in un nuovo cluster
3. Si hanno ora n-1 cluster (tutti quelli di partenza, meno i due che sono stati fusi, più il cluster nuovo) e su questo nuovo insieme si cercano nuovamente i cluster più vicini

   …

n Si ripete la procedura fino ad ottenere un unico cluster, la radice del grafo.

Core Concepts in Data Analysis, Boris Mirkin, **2010**

Cluster gerarchici: agglomerativi

A partire dai dati si raggruppano via via gli elementi considerando ogni volta quelli più vicini.

1. Dati n punti $X_1, \ldots, X_n$ e una matrice di distanze reciproche, ciascun punto costituisce un cluster a sè
2. Si controlla quali di questi cluster sono più vicini e si "fondono" in un nuovo cluster
3. Si hanno ora n-1 cluster (tutti quelli di partenza, meno i due che sono stati fusi, più il cluster nuovo) e su questo nuovo insieme si cercano nuovamente i cluster più vicini

   …

n Si ripete la procedura fino ad ottenere un unico cluster, la radice del grafo.

Entities	Ave	Ant	Ast	Bay	Bre
Ave	0.00	**0.51**	0.88	1.15	2.20
Ant	**0.51**	0.00	0.77	1.55	1.82
Ast	0.88	**0.77**	0.00	1.94	1.16
Bay	1.15	1.55	1.94	0.00	0.97
Bre	2.20	1.82	1.16	0.97	0.00

Core Concepts in Data Analysis, Boris Mirkin, **2010**

Cluster gerarchici: agglomerativi

A partire dai dati si raggruppano via via gli elementi considerando ogni volta quelli più vicini.

1. Dati n punti $X_1, \ldots, X_n$ e una matrice di distanze reciproche, ciascun punto costituisce un cluster a sè
2. Si controlla quali di questi cluster sono più vicini e si "fondono" in un nuovo cluster
3. Si hanno ora n-1 cluster (tutti quelli di partenza, meno i due che sono stati fusi, più il cluster nuovo) e su questo nuovo insieme si cercano nuovamente i cluster più vicini

   …

n Si ripete la procedura fino ad ottenere un unico cluster, la radice del grafo.

Entities	Ave	Ant	Ast	Bay	Bre
Ave	0.00	**0.51**	0.88	1.15	2.20
Ant	**0.51**	0.00	0.77	1.55	1.82
Ast	0.88	**0.77**	0.00	1.94	1.16
Bay	1.15	1.55	1.94	0.00	0.97
Bre	2.20	1.82	1.16	0.97	0.00

Core Concepts in Data Analysis, Boris Mirkin, **2010**

Cluster gerarchici: agglomerativi

A partire dai dati si raggruppano via via gli elementi considerando ogni volta quelli più vicini.

1. Dati n punti $X_1, \ldots, X_n$ e una matrice di distanze reciproche, ciascun punto costituisce un cluster a sè
2. Si controlla quali di questi cluster sono più vicini e si "fondono" in un nuovo cluster
3. Si hanno ora n-1 cluster (tutti quelli di partenza, meno i due che sono stati fusi, più il cluster nuovo) e su questo nuovo insieme si cercano nuovamente i cluster più vicini

   …

n Si ripete la procedura fino ad ottenere un unico cluster, la radice del grafo.

Entities	Ave	Ant	Ast	Bay	Bre
Ave	0.00	**0.51**	0.88	1.15	2.20
Ant	**0.51**	0.00	0.77	1.55	1.82
Ast	0.88	**0.77**	0.00	1.94	1.16
Bay	1.15	1.55	1.94	0.00	0.97
Bre	2.20	1.82	1.16	0.97	0.00

Core Concepts in Data Analysis, Boris Mirkin, **2010**

Cluster gerarchici: agglomerativi

A partire dai dati si raggruppano via via gli elementi considerando ogni volta quelli più vicini.

1. Dati n punti $X_1, \ldots, X_n$ e una matrice di distanze reciproche, ciascun punto costituisce un cluster a sè
2. Si controlla quali di questi cluster sono più vicini e si "fondono" in un nuovo cluster
3. Si hanno ora n-1 cluster (tutti quelli di partenza, meno i due che sono stati fusi, più il cluster nuovo) e su questo nuovo insieme si cercano nuovamente i cluster più vicini

   …

n Si ripete la procedura fino ad ottenere un unico cluster, la radice del grafo.

Entities	Ave	Ant	Ast	Bay	Bre
Ave	0.00	**0.51**	0.88	1.15	2.20
Ant	**0.51**	0.00	0.77	1.55	1.82
Ast	0.88	**0.77**	0.00	1.94	1.16
Bay	1.15	1.55	1.94	0.00	0.97
Bre	2.20	1.82	1.16	0.97	0.00

Sulla base del criterio usato per stabilire la nuova distanza tra i cluster (quella calcolata dopo la fusione) si distinguono diversi metodi:

Single linkage

Average linkage

Complete linkage

UNIVERSITÀ DI PISA NOVARTIS

Core Concepts in Data Analysis, Boris Mirkin, **2010**

Cluster gerarchici: agglomerativi

A partire dai dati si raggruppano via via gli elementi considerando ogni volta quelli più vicini.

1. Dati n punti $X_1, \ldots, X_n$ e una matrice di distanze reciproche, ciascun punto costituisce un cluster a sè
2. Si controlla quali di questi cluster sono più vicini e si "fondono" in un nuovo cluster
3. Si hanno ora n-1 cluster (tutti quelli di partenza, meno i due che sono stati fusi, più il cluster nuovo) e su questo nuovo insieme si cercano nuovamente i cluster più vicini

   …

n Si ripete la procedura fino ad ottenere un unico cluster, la radice del grafo.

Entities	Ave	Ant	Ast	Bay	Bre
Ave	0.00	**0.51**	0.88	1.15	2.20
Ant	**0.51**	0.00	0.77	1.55	1.82
Ast	0.88	**0.77**	0.00	1.94	1.16
Bay	1.15	1.55	1.94	0.00	0.97
Bre	2.20	1.82	1.16	0.97	0.00

Entities	Ave	Ant	Ast	Bay	Bre
Ave			0.88	1.15	2.20
Ant			0.77	1.55	1.82
			0.82	1.35	2.01

Sulla base del criterio usato per stabilire la nuova distanza tra i cluster (quella calcolata dopo la fusione) si distinguono diversi metodi:

Single linkage

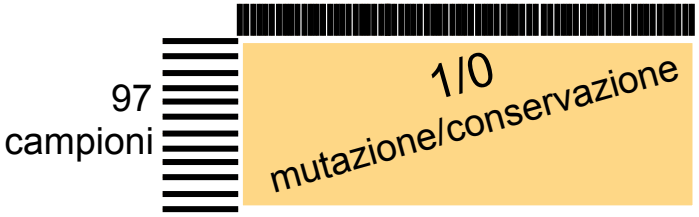
Average linkage

Complete linkage

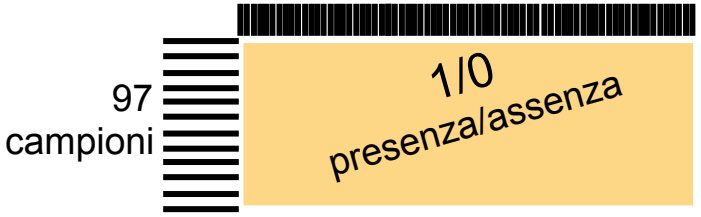
Core Concepts in Data Analysis, Boris Mirkin, **2010**

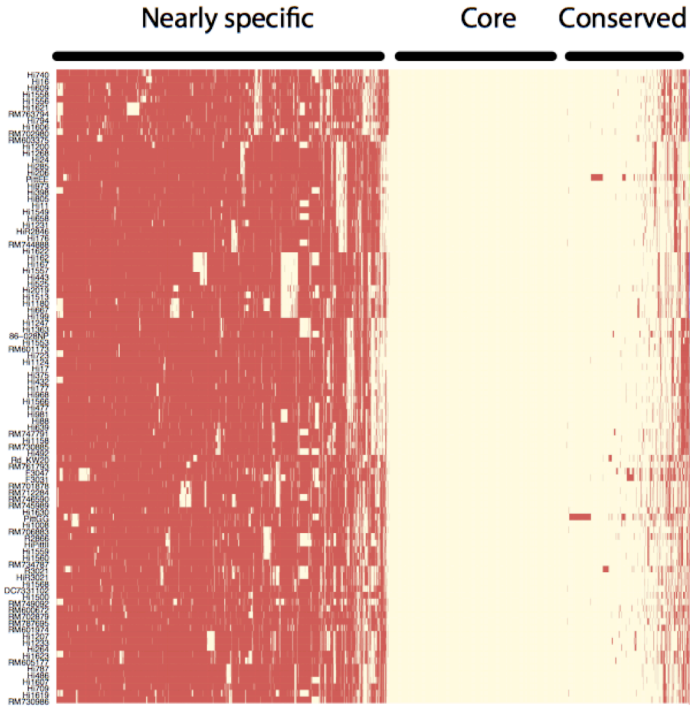
Clustering nello studio di popolazioni batteriche: METODI

coordinata delle singole basi
nel genoma di riferimento (ca.150000)



geni del genoma di riferimento
(ca.2000)

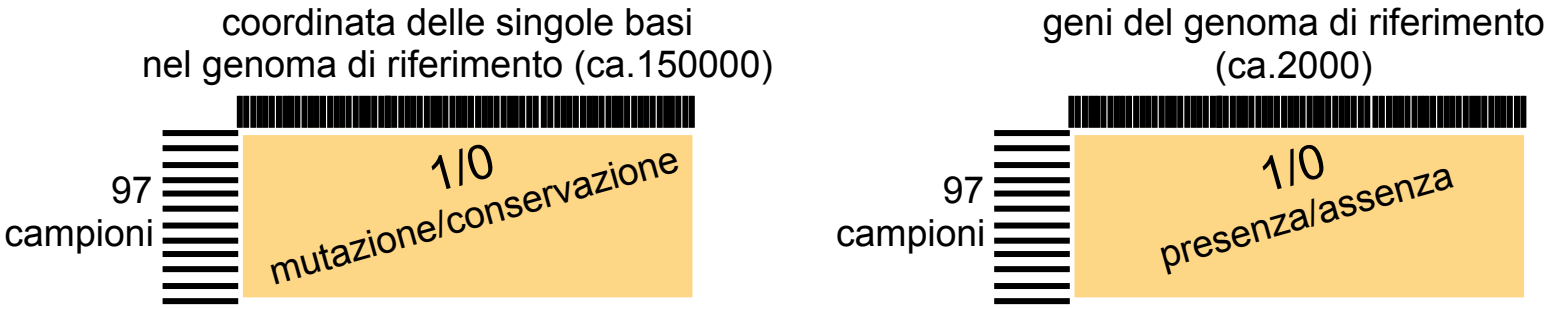




De Chiara et al, PNAS, **2014**

UNIVERSITÀ DI PISA NOVARTIS

Clustering nello studio di popolazioni batteriche: METODI



- Data la mole di dati serve un approccio che consenta di ridurre la dimensione del problema:

applicano PCA

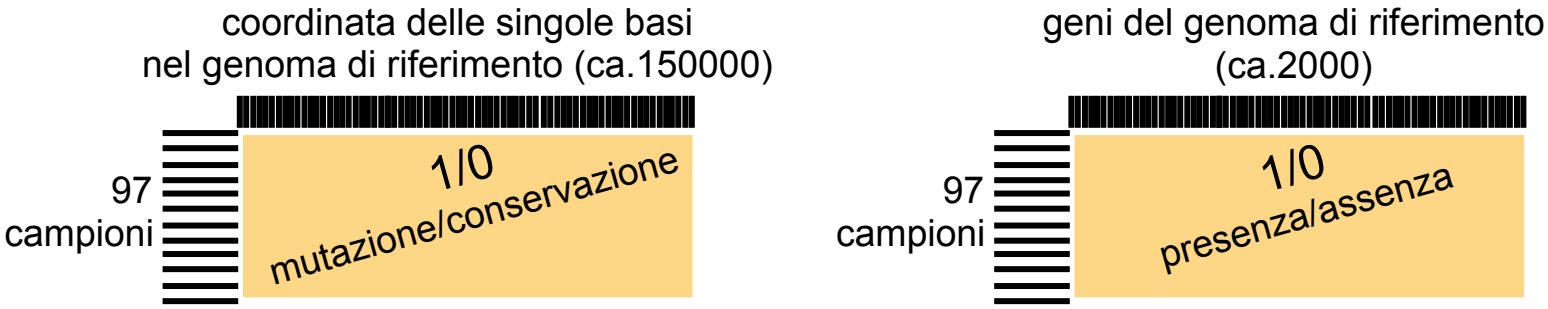
tengono 60 componenti (soglia al 90% della variabilità)

Discriminant analysis of principal components: a new method for the analysis of genetically structured populations, Jombart et al., BMC Genetics**, 2010**, 11

De Chiara et al, PNAS, **2014**

UNIVERSITÀ DI PISA NOVARTIS

Clustering nello studio di popolazioni batteriche: METODI



- Data la mole di dati serve un approccio che consenta di ridurre la dimensione del problema:

  applicano PCA

  tengono 60 componenti (soglia al 90% della variabilità)

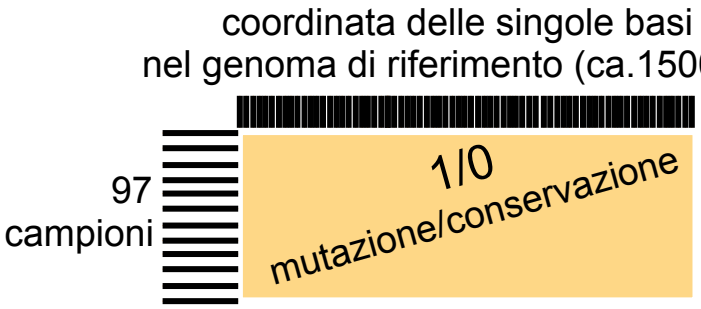
- per individuare i cluster applicano il K-means facendo variare K

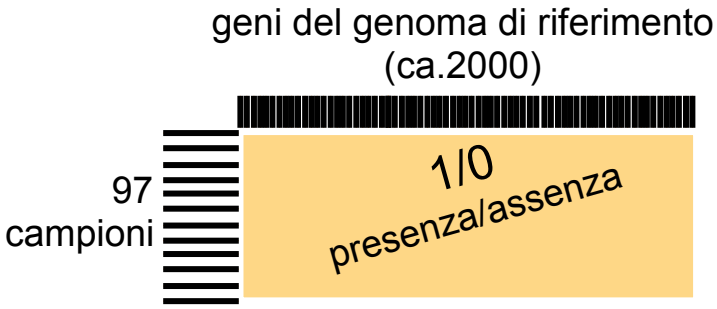
Discriminant analysis of principal components: a new method for the analysis of genetically structured populations, Jombart et al., BMC Genetics**, 2010**, 11

De Chiara et al, PNAS, **2014**

UNIVERSITÀ DI PISA NOVARTIS

Clustering nello studio di popolazioni batteriche: METODI





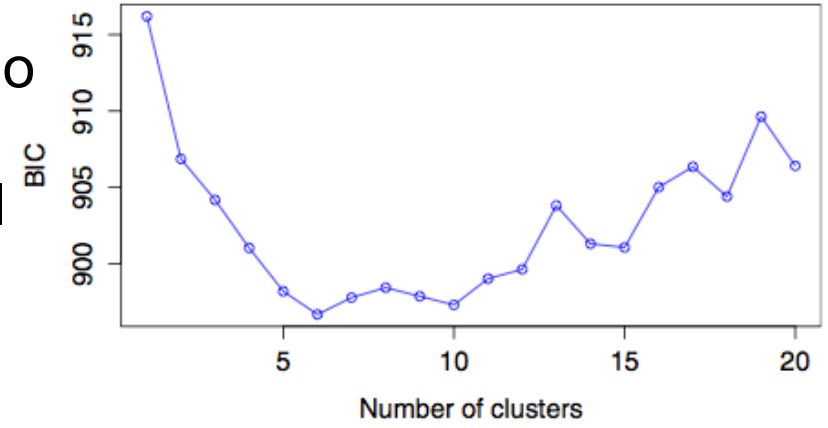
- Data la mole di dati serve un approccio che consenta di ridurre la dimensione del problema:

applicano PCA

tengono 60 componenti (soglia al 90% della variabilità)

- per individuare i cluster applicano il K-means facendo variare K
- per scegliere il miglior K usano il Bayesian Information Criterion

$$-2\ln\left[L^0(M)\right] + K\ln(n)$$

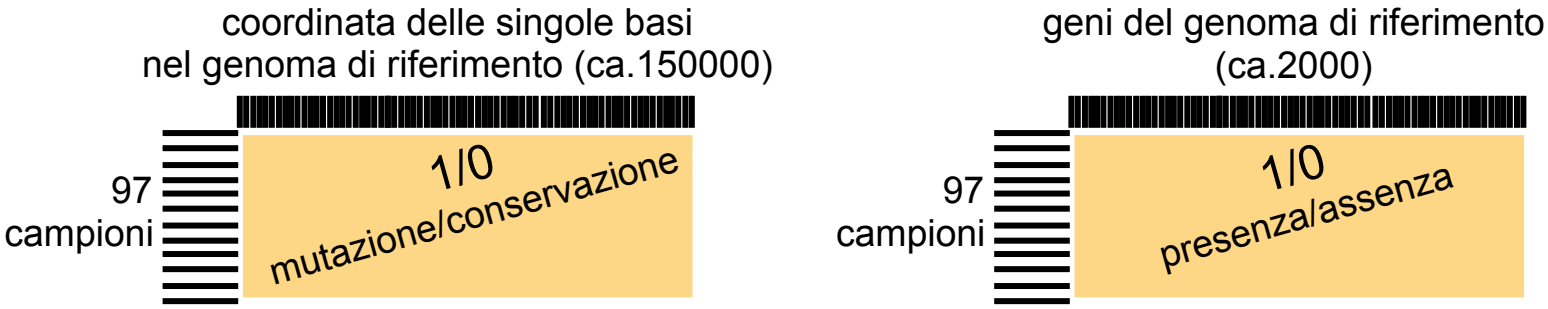


Discriminant analysis of principal components: a new method for the analysis of genetically structured populations, Jombart et al., BMC Genetics**, 2010**, 11

De Chiara et al, PNAS, **2014**

UNIVERSITÀ DI PISA NOVARTIS

Clustering nello studio di popolazioni batteriche: METODI



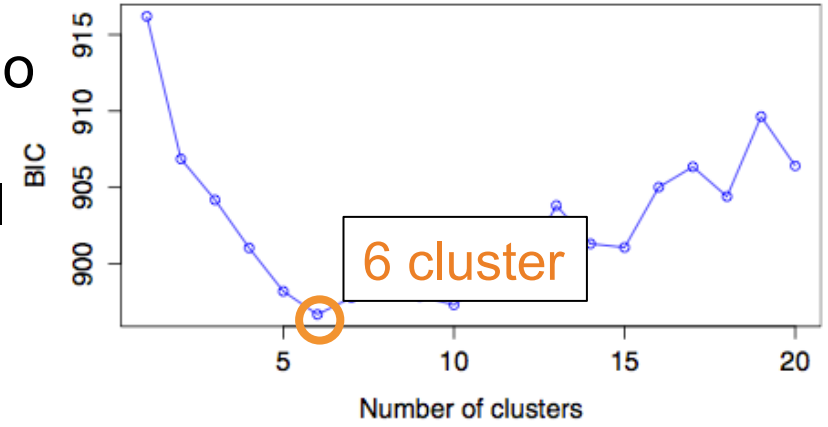
- Data la mole di dati serve un approccio che consenta di ridurre la dimensione del problema:

applicano PCA

tengono 60 componenti (soglia al 90% della variabilità)

- per individuare i cluster applicano il K-means facendo variare K
- per scegliere il miglior K usano il Bayesian Information Criterion

$$-2\ln\left[L^0(M)\right] + K\ln(n)$$

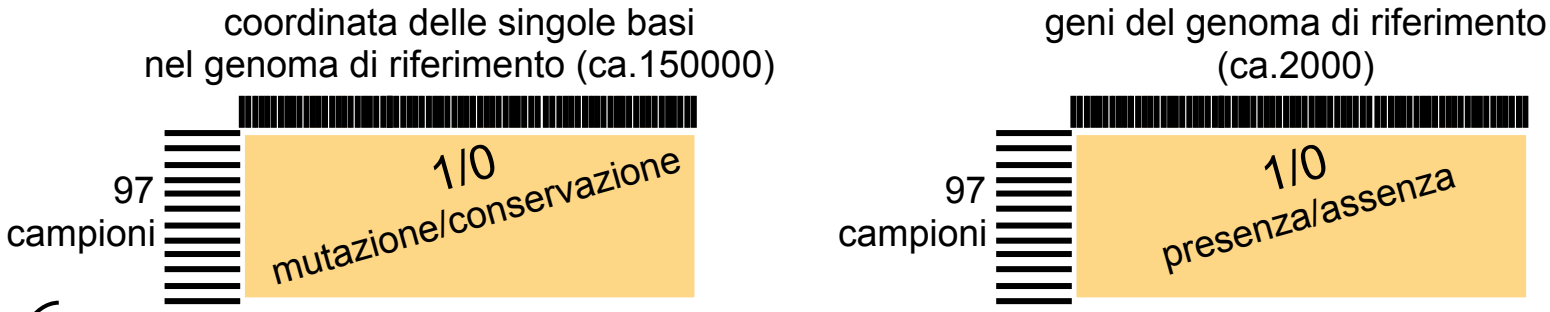


Discriminant analysis of principal components: a new method for the analysis of genetically structured populations, Jombart et al., BMC Genetics**, 2010**, 11

De Chiara et al, PNAS, **2014**

UNIVERSITÀ DI PISA NOVARTIS

Clustering nello studio di popolazioni batteriche: METODI



DAPC

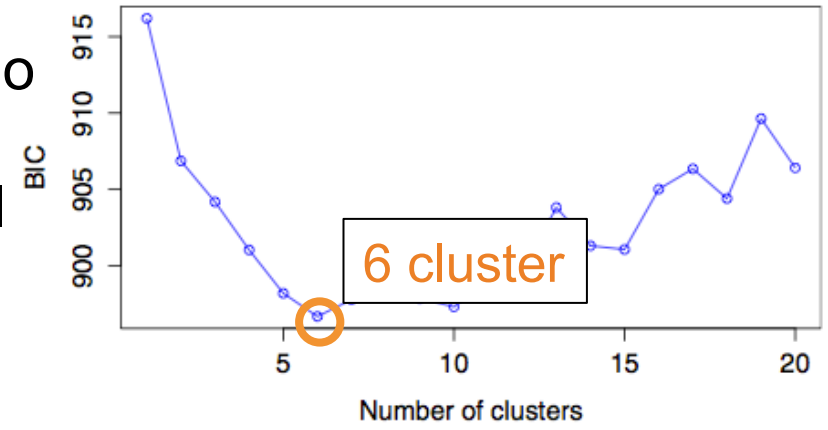
- Data la mole di dati serve un approccio che consenta di ridurre la dimensione del problema:

  applicano PCA

  tengono 60 componenti (soglia al 90% della variabilità)

- per individuare i cluster applicano il K-means facendo variare K
- per scegliere il miglior K usano il Bayesian Information Criterion

$$-2\ln\left[L^0(M)\right] + K\ln(n)$$



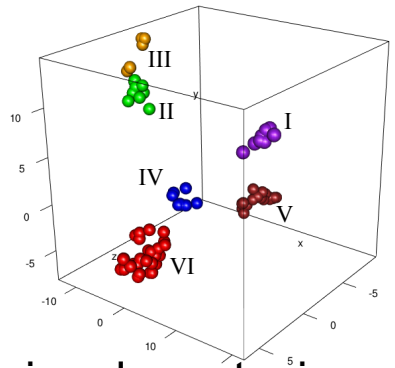
Discriminant analysis of principal components: a new method for the analysis of genetically structured populations, Jombart et al., BMC Genetics, **2010**, 11

De Chiara et al, PNAS, **2014**

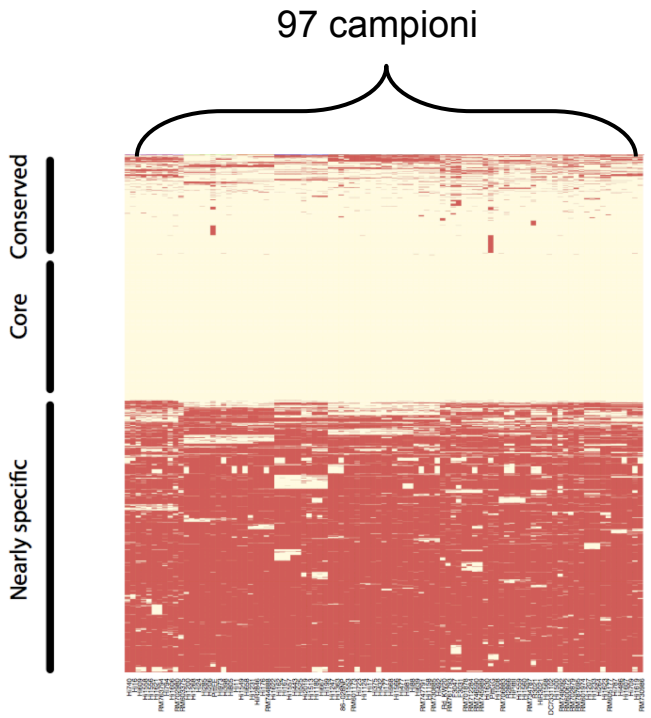
UNIVERSITÀ DI PISA NOVARTIS

Clustering nello studio di popolazioni batteriche: RISULTATI

- DAPC



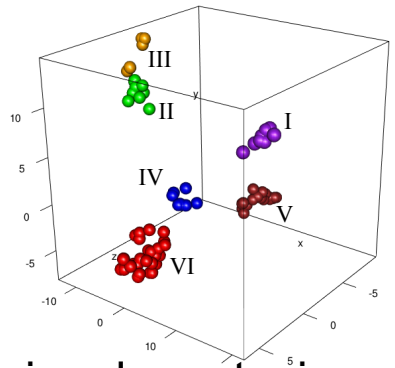
singola mutazione nel core



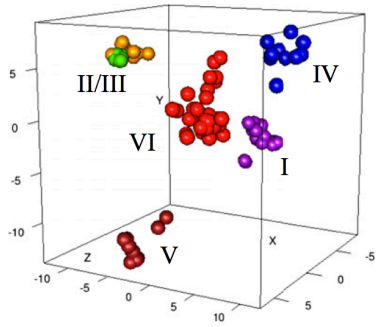
De Chiara et al, PNAS, **2014**

Clustering nello studio di popolazioni batteriche: RISULTATI

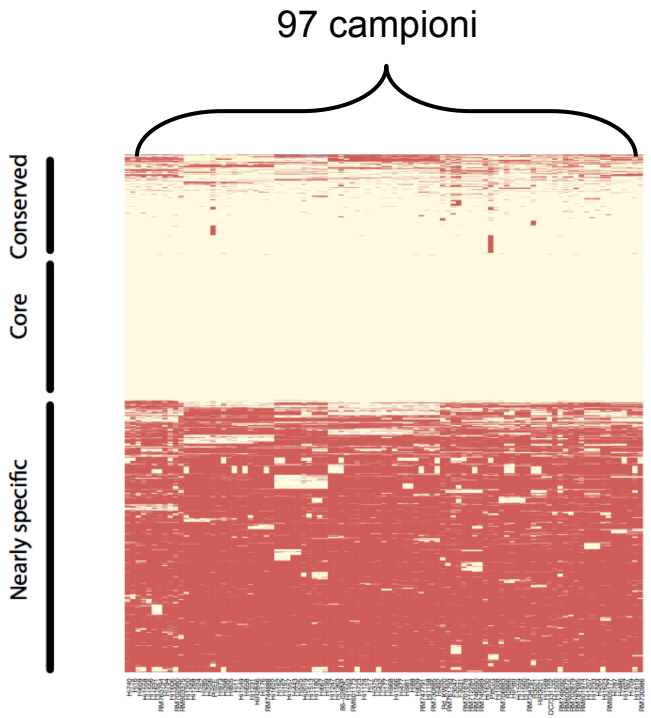
- DAPC



singola mutazione nel core



presenza/assenza geni



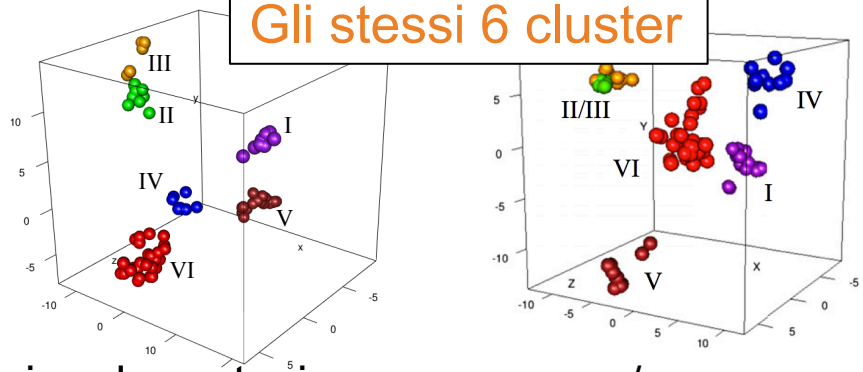
De Chiara et al, PNAS, **2014**

UNIVERSITÀ DI PISA NOVARTIS

Clustering nello studio di popolazioni batteriche: RISULTATI

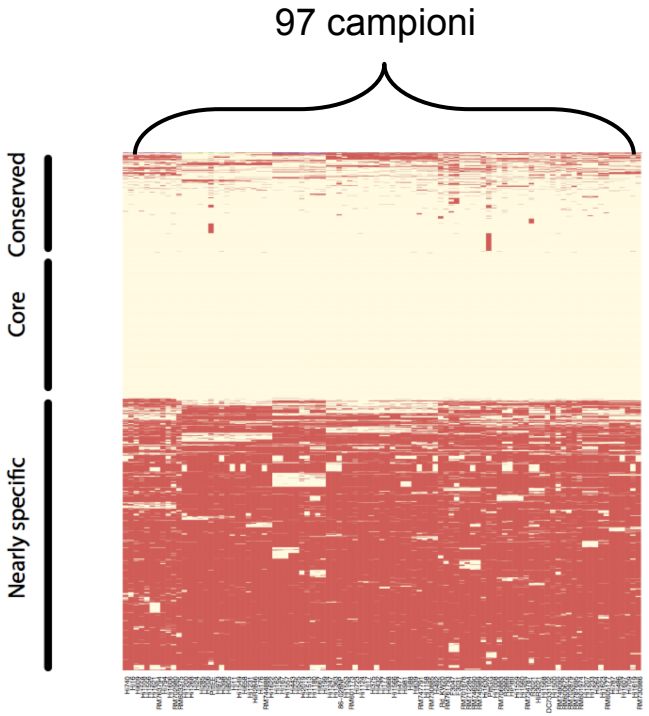
- DAPC

Gli stessi 6 cluster



singola mutazione nel core

presenza/assenza geni

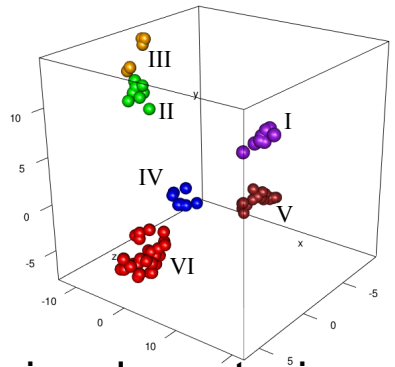


De Chiara et al, PNAS, **2014**

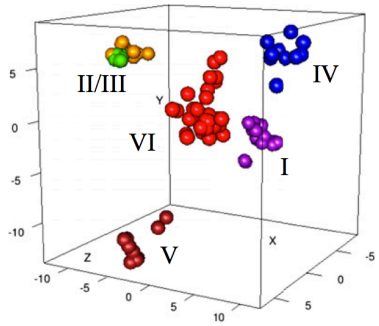
UNIVERSITÀ DI PISA NOVARTIS

Clustering nello studio di popolazioni batteriche: RISULTATI

- DAPC

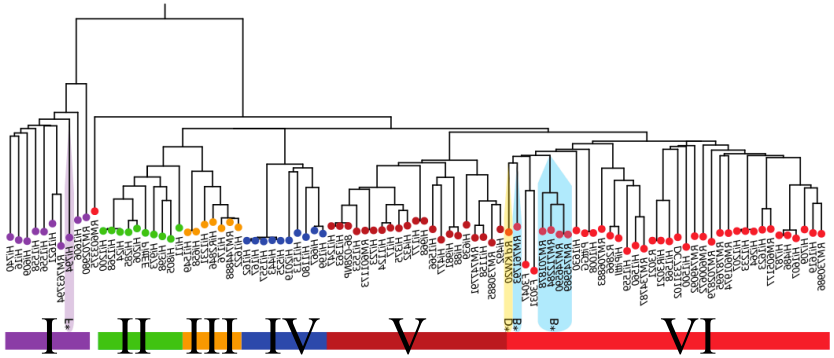


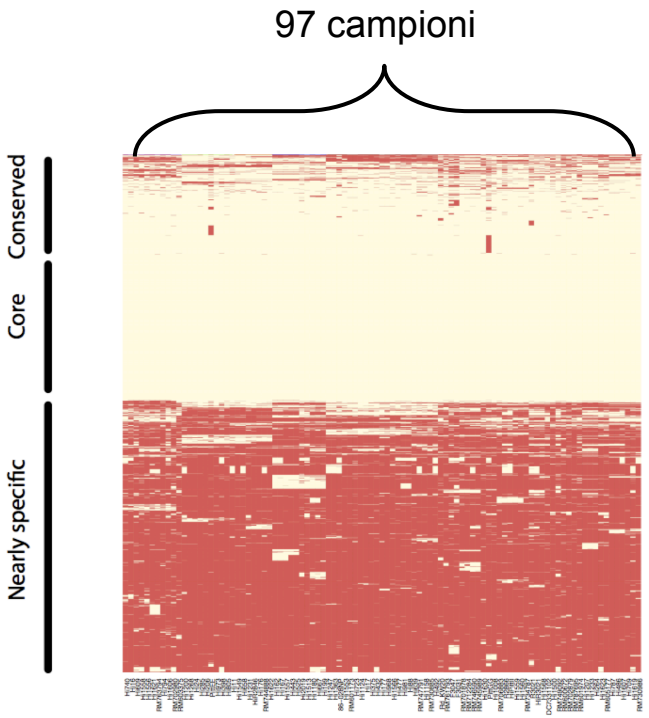
singola mutazione nel core



presenza/assenza geni

- Cluster gerarchici



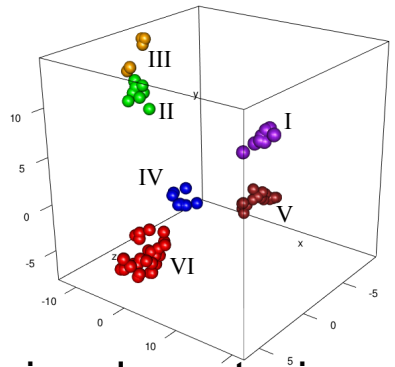


De Chiara et al, PNAS, **2014**

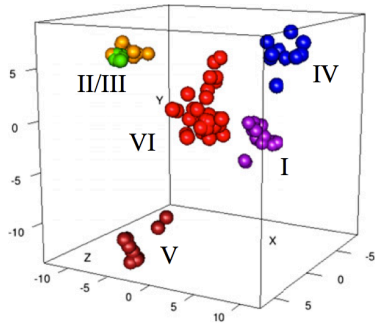
UNIVERSITÀ DI PISA NOVARTIS

Clustering nello studio di popolazioni batteriche: RISULTATI

- DAPC

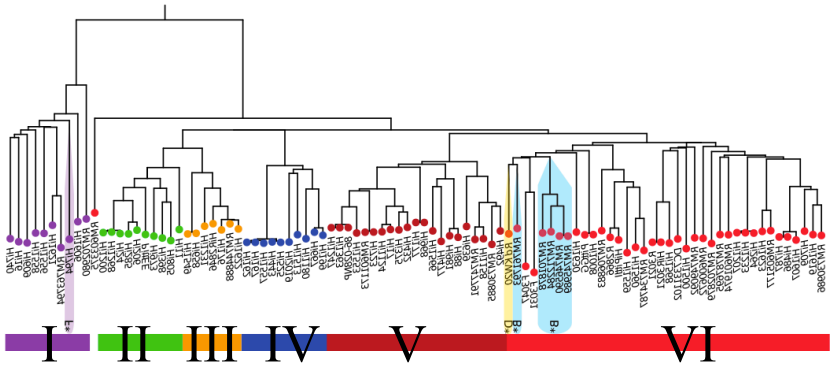


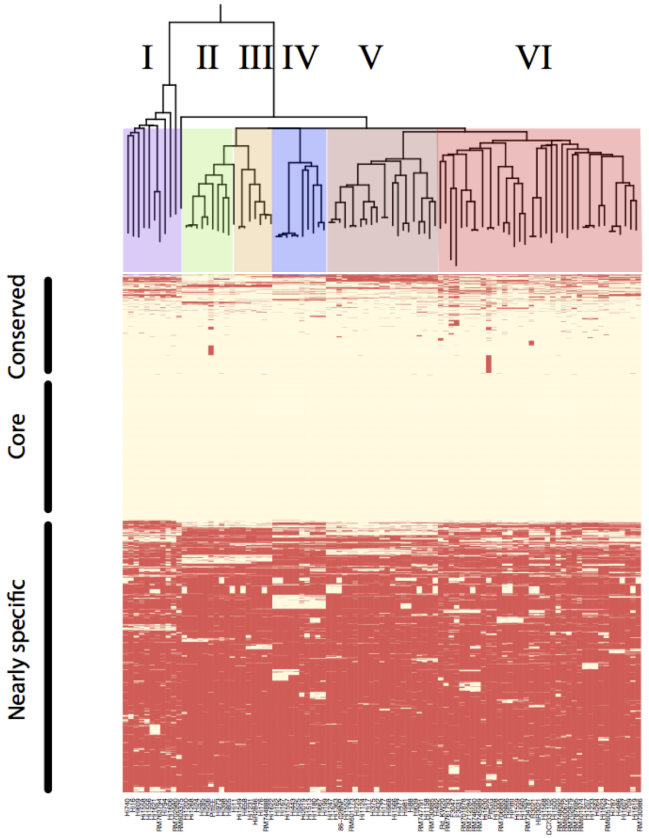
singola mutazione nel core



presenza/assenza geni

- Cluster gerarchici





De Chiara et al, PNAS, **2014**

UNIVERSITÀ DI PISA NOVARTIS

Proposte per il futuro

obiettivo: individuare la presenza di correlazioni tra i genomi e le malattie provocate dai batteri stessi.

novità:

1. Analisi di più campioni ma tutti con la stessa origine geografica e stesso periodo temporale
2. Sempre rispetto ad un genoma di riferimento, considerare più tipi di mutazioni e classificarle in dettaglio (inserzioni, delezioni, ripetizioni…)
3. Utilizzare la tecnica del mapping per individuare le mutazioni
4. Implementare codice che esegua operazioni in parallelo.

metodi:

- DAPC e cluster gerarchici agglomerativi
- Spectral clustering

Ringraziamenti

Ringrazio

- Prof. Dino Leporini
- Dott. Alessandro Muzzi

e tutti voi per l'attenzione