

UNIVERSITA' DEGLI STUDI DI PAVIA

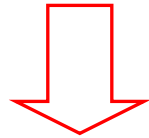
Dipartimento di Scienze  
Sanitarie Applicate e Psicocomportamentali

# **The prediction of classical HLA alleles from SNP Data in the Nuoro population**

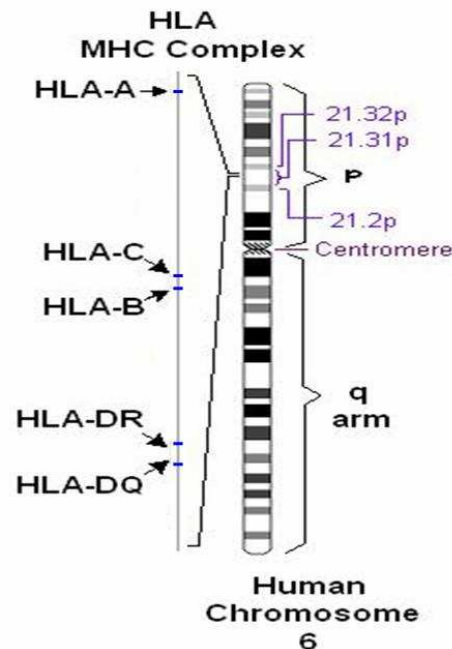
Roberta Pastorino

# Human Leukocyte Antigen

**Sclerosi multipla:** malattia infiammatoria demielinizzante del SNC  
Più geni contribuiscono a dare suscettibilità alla SM



**HLA** (Human Leukocyte Antigen) (complesso di geni sul cromosoma 6)



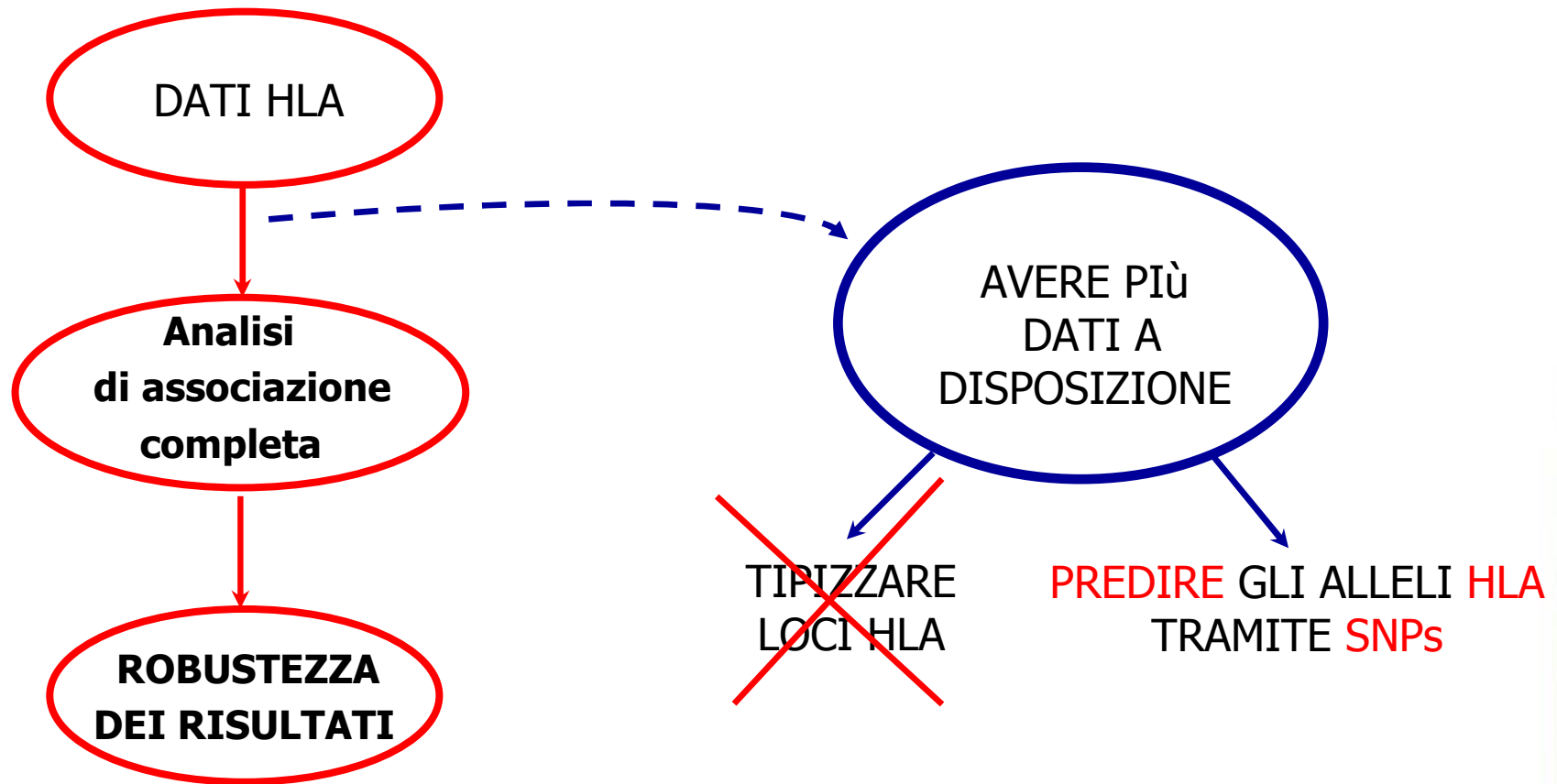
Ci siamo focalizzati su :

□ HLA-A HLA-CW HLA-B → loci di classe I

□ HLA-DR HLA-DQ → loci di classe II

Chiarire il ruolo dei geni HLA  
nella popolazione isolata di Nuoro

# Studio di Associazione



# Predizione Alleli HLA

Esistono due articoli di riferimento:

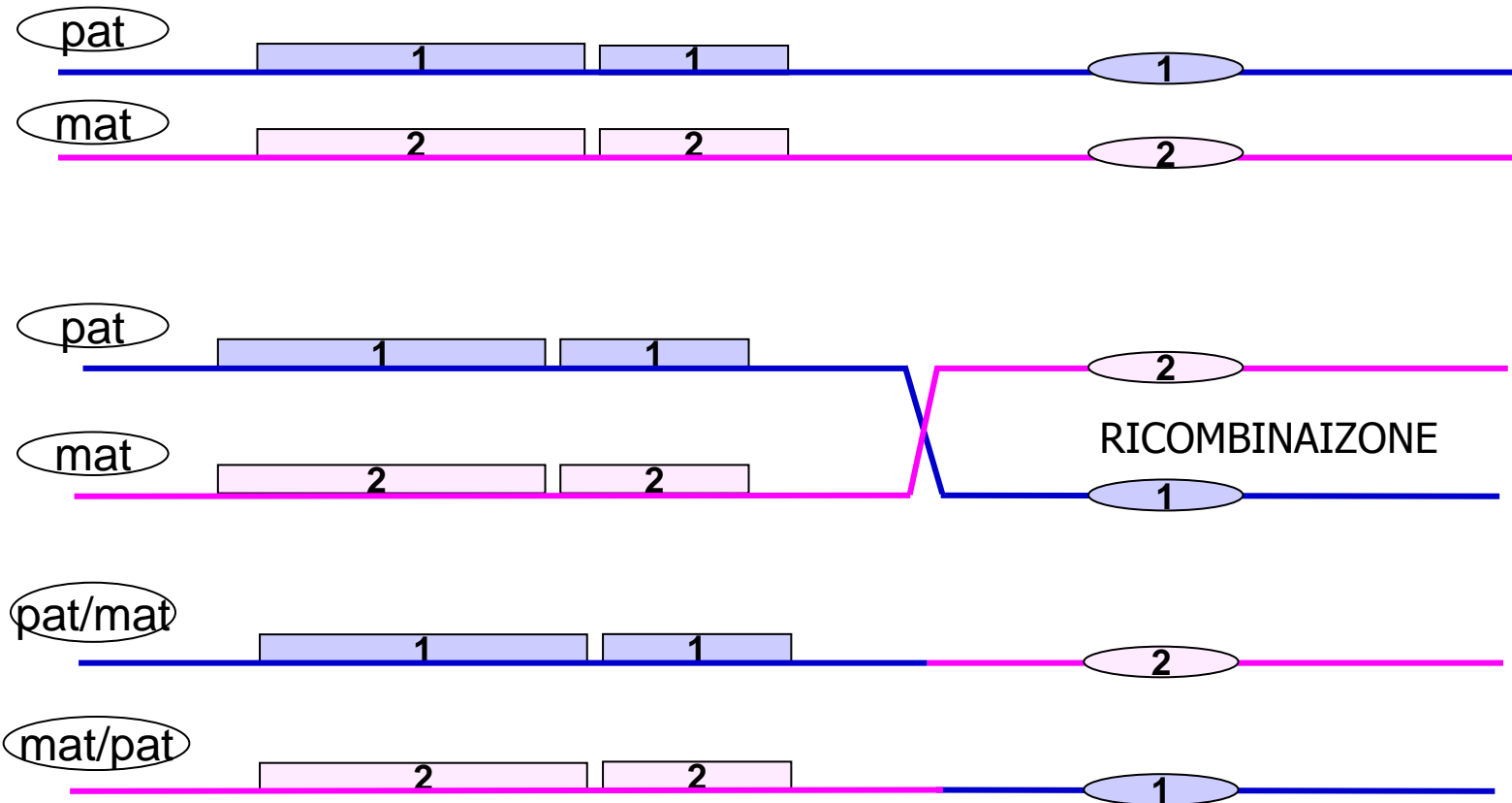
- ❑ *Bakker et al (2006) A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet. 38: 1166-72.*
- ❑ *Leslie S, Donnelly P, McVean G. (2008) A statistical method for predicting classical HLA alleles from SNP data. Am J Hum Genet. 82:48-56.*
- ❑ Predizione non univoca per popolazioni diverse

## Obiettivi:

- 1) Mettere a punto il metodo statistico
- 2) Predire gli alleli HLA nella popolazione di Nuoro

# Concetti Chiave

- **Linkage disequilibrium (LD):** *nonindependence, at a population level, of the alleles carried at different positions in the genome*



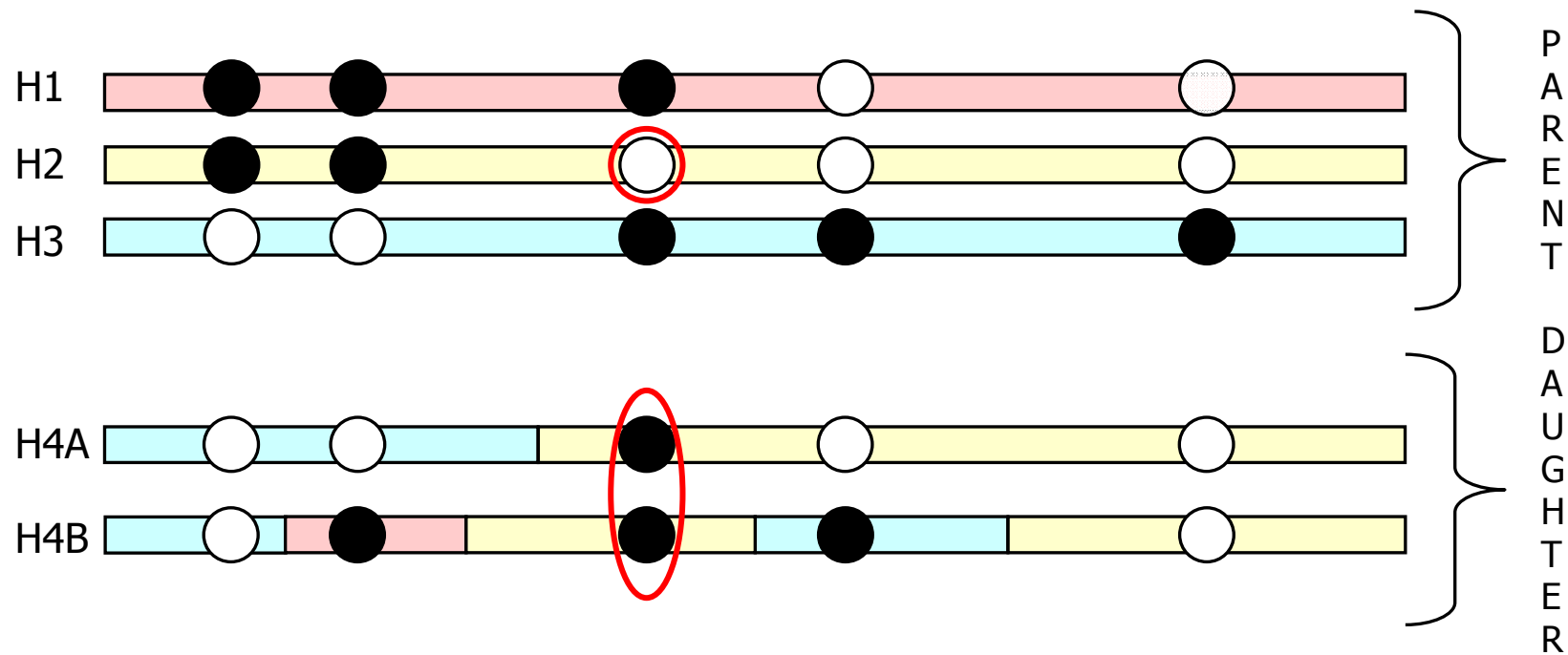
# Concetti Chiave

- Sequenze di DNA di diversi individui allo stesso locus non sono indipendenti
- Alto LD tra SNPs e loci HLA
- In presenza di **mutazione** e **ricombinazione** possiamo affermare che il background genetico di individui che hanno lo stesso allele ad un dato locus sono più simili tra di loro rispetto a individui con differenti alleli

**Come modellizziamo questa similarità tra gli individui?**

# Modello di Li and Stephens (2003)

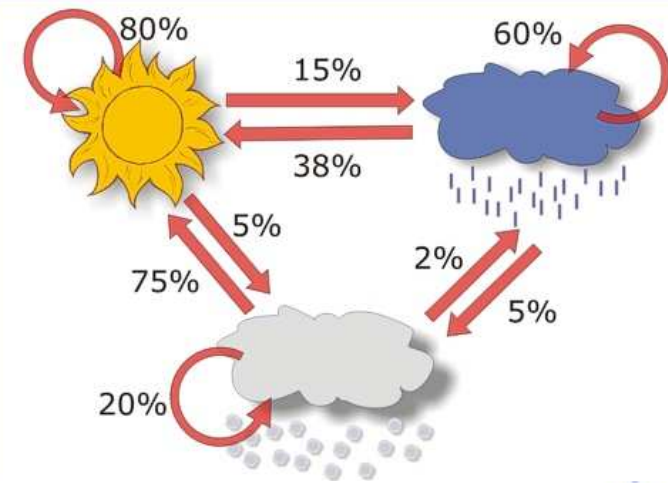
- Probabilità che una nuova sequenza  $h^{n+1}$  sia ottenuta da una popolazione da cui abbiamo già estratto  $h^1 \dots h^n$



# Processo di Markov

- Processo di copiatura modellizzato con un processo di Markov

- Modello matematico di processo stocastico dove l'evoluzione al tempo  $t+1$  dipende solo dallo stato del sistema al tempo  $t$ 
  - Insieme  $S$  di  $N$  stati
  - Insieme  $A$  di  $M$  possibili emissioni
  - Matrice  $P$  di transizione

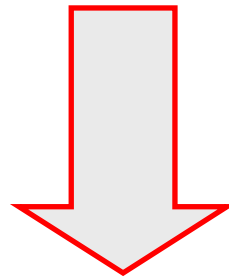


# Processo di Markov

- Nel nostro caso la nuova sequenza “daughter” è l’osservazione (emissione) mentre gli stati sono costituiti dalle sequenze “parents” utilizzate nel processo di copiatura ai vari loci.
- Stati sono sconosciuti → **Modelli markoviani nascosti (HMM)**
- **Probabilità di transizione:** ricombinazione
- **Probabilità di emissione:** mutazione

# HMM: Forward Algorithm

- Calcolo  $P(h_{n+1} / h_1, \dots, h_n) = \sum \text{Parental.Path}$
- T osservazioni, N stati  $\rightarrow N^T$  sequenze

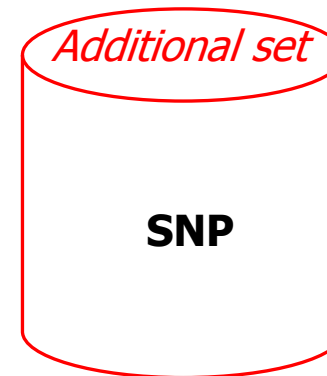
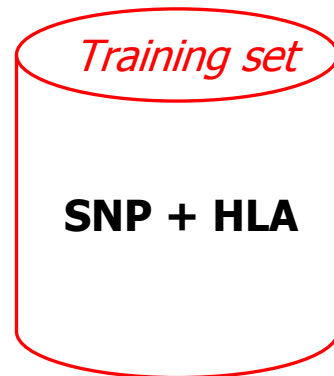


$\sum$  risolta efficientemente usando l'algoritmo forward degli HMM

$$P(h_{n+1} / h_1, \dots, h_n) = \sum_{j=1}^n f^j$$

# Algoritmo di Predizione

- Utilizza un DB di aplotipi con SNP e HLA noti (*training set*) per predire alleli HLA in un DB addizionale (*additional set*) con solo SNPs noti.



- Confrontiamo gli SNPs del *set addizionale* con quelli del *training set* con l'assunzione che se la sequenza addizionale porta un dato allele HLA esso sarà un imperfetto mosaico delle sequenze che portano lo stesso allele

# Algoritmo di Predizione

- **Stage 1:** tra gli SNPs disponibili si seleziona un set di SNPs ottimali (*prediction SNPs*) utilizzando una misura di distanza dal condizionale ottimale (*100% accuratezza*)
  - Si possono usare gli SNPs già disponibili
- **Stage 2:** utilizzando il set di SNPs predittivo, training set, additional set vengono stimati gli alleli HLA mancanti nel set addizionale

$$P(a / h^i) = \frac{P(h^i / a) * P(a)}{\sum_{b \in A} P(b) \pi(h^i / b)}$$

Diagram illustrating the HMM probability calculation for allele prediction:

- Probabilità HMM** (blue arrow) points to  $P(h^i / a)$ .
- Probabilità a priori** (red arrow) points to  $P(a)$ .
- Probabilità a posteriori** (grey arrow) points to  $P(a / h^i)$ .
- Fattore di normalizzazione** (green arrow) points to the denominator  $\sum_{b \in A} P(b) \pi(h^i / b)$ .

The final prediction is given by:

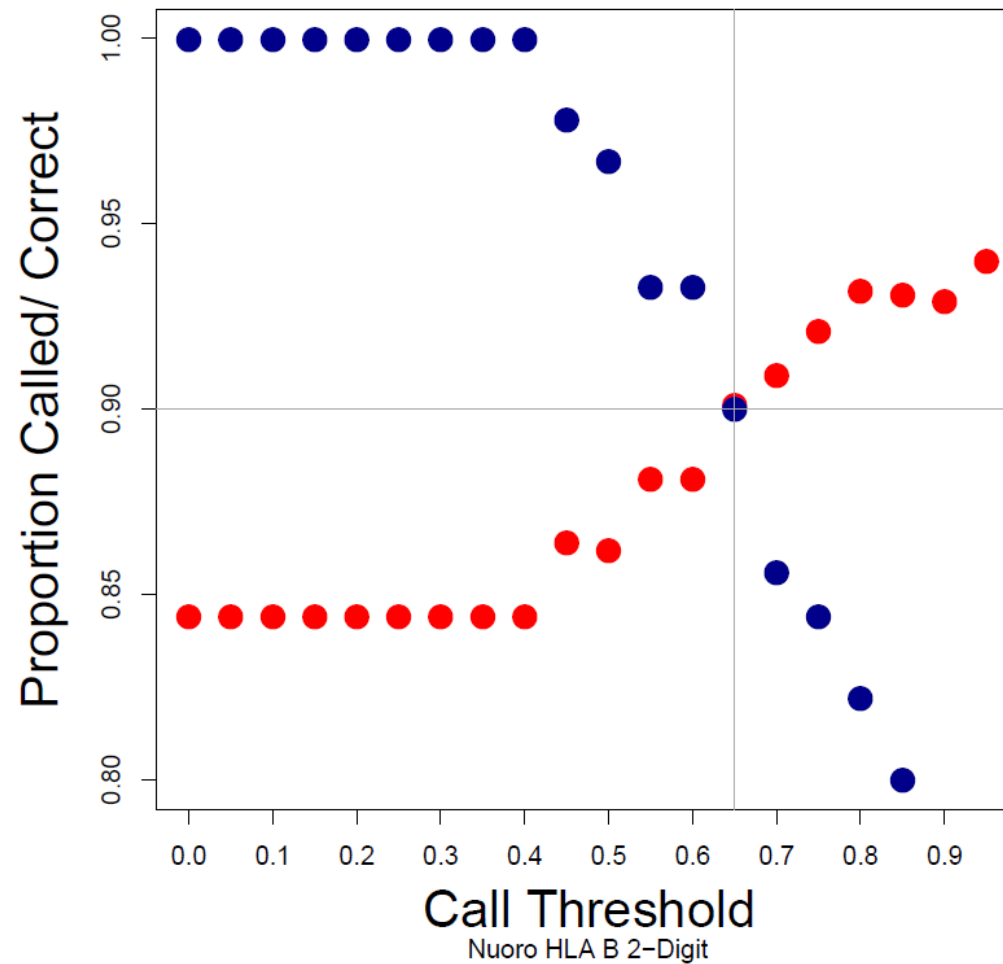
$$a^P = \arg \max P(a / h^i)$$

# Analisi

- Essenziale una valutazione della predizione
  - **Sensitività:** proporzione delle predizione corrette
  - **Specificità:** proporzione di volte in cui un dato allele, quando presente, sia correttamente predetto
  
- Dati:
  - *Training set:* 130 individui (trio)
  - *Additional set:* 45 individui (casi e controlli)
  - Loci HLA: A CW B DR DQ
  - 994 SNPs

# RISULTATI: Call Threshold

Effect of Setting a Call Threshold



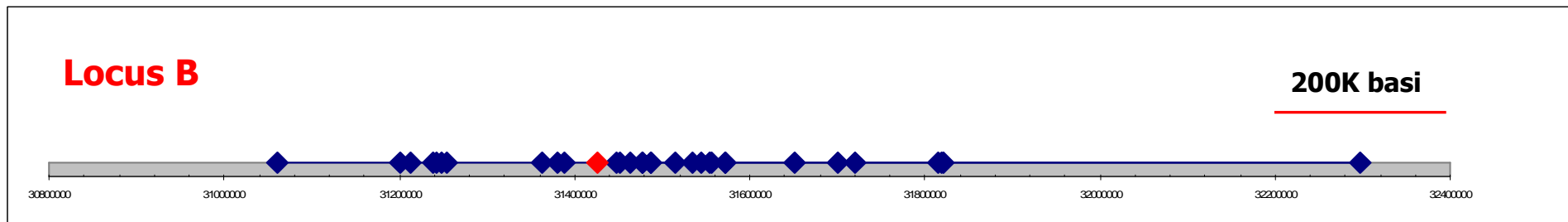
# RISULTATI: Sensitività

Diverse accuratèzze  
a seconda dei vari  
alleli

<b>Allele</b>	<b>Sensitività</b>
<b>A30</b>	1
<b>CW5</b>	0.93
<b>B18</b>	1
<b>DR3</b>	0.90
<b>DQ2</b>	0.91

# RISULTATI e CONCLUSIONI

- ❑ Risultati promettenti (sensibilità  $> 0.9$ )
- ❑ SNPs per la predizione



Lavoro orientato ad uno studio di associazione!