Analysis of MOlecular VAriance (AMOVA)

Alessandra Nardi

alenardi@mat.uniroma2.it

Vediamo come gli elementi base dell'Analisi delle Varianza possano essere utilizzati per studiare la variabilità di dati molecolari in popolazioni diverse

Per meglio comprendere le caratteristiche della metodologia faremo riferimento ad una analisi su dati reali tratta dal lavoro

Polymorphisms of the COL1A2, CYP1A1 and HS1,2 Ig enhancer genes in the Tuaregs from Libya

Cristina Martinez-Labarga et al.
Annals of Human Biology, 2007; 34(4) : 425−436

Background: Restriction fragment length polymorphisms (RFLPs) of the COL1A2 and CYP1A1 and short tandem repeats of HS1,2 Ig enhancer genes are proving to be useful markers for describing human populations and thus are of interest for anthropogenetic research. Moreover, they can provide useful information in identifying alleles and haplotypes associated with particular forms of common diseases or for pharmacogenomics studies.

Aim: The first objective of this study was to define the genetic structure of Libyan Tuaregs and to establish the degree of genetic homogeneity amongst the El Awaynat and Tahala groups.

L'AMOVA, introdotta nel 1992 da Laurent Excoffier, riprende la logica dell'analisi della varianza classica per applicarla allo studio della variabilità genetica.

Rispetto alla classica ANOVA

- La variabile risposta è adesso un aplotipo (dal greco haploos= singolo o semplice) cioè la combinazione di varianti alleliche lungo un cromosoma o segmento cromosomico contenente loci strettamente associati tra di loro, e che in genere, vengono ereditati insieme
- Non è ovviamente possibile assumere un modello Normale

Resta invece valida l'idea di scomporre la variabilità complessiva dei dati in variabilità all'interno delle popolazioni e tra le popolazioni.

Attenzione perché , in questo ambito, il termine *popolazione* viene utilizzato per indicare un insieme di individui della stessa specie che si incrociano tra di loro, possedendo così un pool genetico comune.

A partire dai dati individuali viene cotruita una matrice di *distanze* tra individui e da questa matrice verranno ricavati i dati sulle diverse fonti di variabilità

Dato un aplotipo, questo viene rappresentato attraverso il vettore

$\mathbf{p}^t = (p_1, p_2, \ldots, p_s)$

i cui elementi sono 1 o 0 a seconda che siano state riscontrate delle differenze rispetto ad un aplotipo base, scelto come riferimento.

La differenza tra l'individuo j e l'individuo k viene definita come

$(\mathbf{p}_j - \mathbf{p}_k) = [(p_{1j} - p_{1k}), (p_{2j} - p_{2k}), \ldots, (p_{sj} - p_{sk})]$

per arrivare al calcolo della distanza Euclidea

$\delta_{jk} = (\mathbf{p}_j - \mathbf{p}_k)^t (\mathbf{p}_j - \mathbf{p}_k)$

che nel nostro caso conta i siti in cui è stata osservata una differenza.

Definizioni alternative di distanza possono essere considerate ad esempio pesando in modo diverso i singoli siti.

A partire dalla matrice delle distanze viene stimata la varianza totale, quella all'interno delle popolazioni e quella tra popolazioni sfruttando la relazione

$\sum_i (x_i - \bar{x})^2 = \frac{\sum_i \sum_j (x_i - x_j)^2}{2n}$

Stimate le diverse componenti della varianza, viene costruita una statistica F simile nella logica alla classica statistica F dell'ANOVA

Resta la difficoltà di definire un modello di riferimento data la particolare natura della variabile risposta

Come diretta conseguenza la scomposizione della devianza non è più funzionale alla verifica d'ipotesi sui valori attesi

L'interpretazione della varianza *within population* come stimatore della varianza casuale resta legata alla possibilità di assumere individui omogenei all'interno delle popolazioni

Mancando l'assunzione di un modello Normale la distribuzione della statistica test F sotto l'ipotesi nulla viene generata attraverso metodi di permutazione.

Sotto H_0 infatti le osservazioni saranno *scambiabili*; si generano allora permutazioni casuali dei dati originari e per ciascuna delle permutazioni ottenute si calcola il valore della statistica F. L'istogramma corrispondente ci fornisce una stima simulata della distribuzione campionaria.

AMOVA analysis showed no variation between the two villages, since the proportion of genetic variance that could be attributed to differences between the two populations was -1.45% ($p = 0.57$), indicating that genetic variance within these populations was larger than that between them. These data demonstrate the high genetic homogeneity of the Libyan Tuaregs, at least using the statistical resolution of the AMOVA.

In order to provide a clearer picture of COL1A2, CYP1A1 and HS1,2 Ig enhancer allele and haplotype frequency distributions in various human groups distributed over a wide geographic area, comparisons with other African, European and Asian populations were carried out by analysis of molecular variance (AMOVA) and genetic distance analysis.

È possibile estendere l'analisi creando una struttura gerarchica che prevede le popolazioni riunite in *Gruppi*

Ad esempio nel Gruppo Europeo troviamo le popolazioni Italiana, Tedesca e Spagnola

A questa struttura gerarchica corrisponde un'ulteriore decomposizione della varianza in varianza tra gruppi, varianza tra popolazioni all'interno dei gruppi e varianza all'interno delle popolazioni

Analyses of molecular variance (AMOVA) based on allele and haplotype frequencies in various populations.

Group of populations	Genetic markers	Source of variation		
		Among groups	Among populations within groups	Whithin populations
Ecuadorian	*COL1A2*	7.58	2.29	90.13
Native Americans				
Colorado				
Cayapa				
South Saharan Africans				
Benin				
Saharan Africans	*CYP1A1*	36.79	1.41	61.80
Libyan Tuaregs				
Europeans				
Germany/Italy				
Spain				
East Asians	*HS1,2-A*	9.30	0.81	91.51
Japan/Indonesia/Mongolia				
West Asians				
Turkey				