Giorgio Bernardi

La distribuzione dei geni nel genoma umano

Riassunto

Il genoma umano è un mosaico di isocore, cioè di lunghe regioni di DNA (>300 kb; 1 kb equivale a 1000 paia di basi), che sono caratterizzate da una bassa eterogeneità composizionale. Le isocore appartengono ad un piccolo numero di famiglie che coprono un'ampia gamma di GC (GC è il rapporto molare di guanina+citosina nel DNA). Una compartimentazione di questo tipo è molto diffusa negli eucarioti ed è differente ma stabile nelle diverse classi di vertebrati (mammiferi, anfibi, etc.). In tutti i vertebrati esistono buone correlazioni tra i livelli di GC di sequenze codificanti e non codificanti ed anche tra i livelli di GC delle diverse posizioni (I, II, III) dei codoni.

La distribuzione dei geni è bimodale nei vertebrati, la densità genica essendo molto alta nel piccolo "genome core" (o nucleo del genoma, che nel genoma umano corrisponde alle famiglie di isocore più ricche in GC, H2 e H3), ed è invece molto bassa nel vasto "genome desert" (o deserto del genoma, che è formato dalle altre famiglie di isocore, L1, L2 e H1). La distribuzione dei geni nei vertebrati è correlata ad importanti caratteristiche strutturali, funzionali ed evoluzionistiche.

La compartimentazione del genoma

Trenta anni fa abbiamo scoperto che il genoma bovino (escludendo i DNA satelliti, che consistono di corte sequenze ripetute in serie) è un mosaico di segmenti di DNA appartenenti a diverse famiglie composizionali (Filipski et al., 1973). Questa osservazione iniziale fu rapidamente estesa ad altri eucarioti. I genomi di vertebrati a sangue caldo mostrarono essenzialmente le caratteristiche composizionali appena descritte per il genoma bovino (Thiery et al., 1976). Ad esempio, nel genoma umano furono individuate e fisicamente isolate cinque famiglie di molecole di DNA (escludendo i DNA satellite e ribosomale). Queste famiglie furono chiamate L1, L2, H1, H2 e H3, secondo un ordine crescente di GC; le prime tre comprendevano circa l'88% del DNA e le ultime due circa il 12 % (Fig. 1).

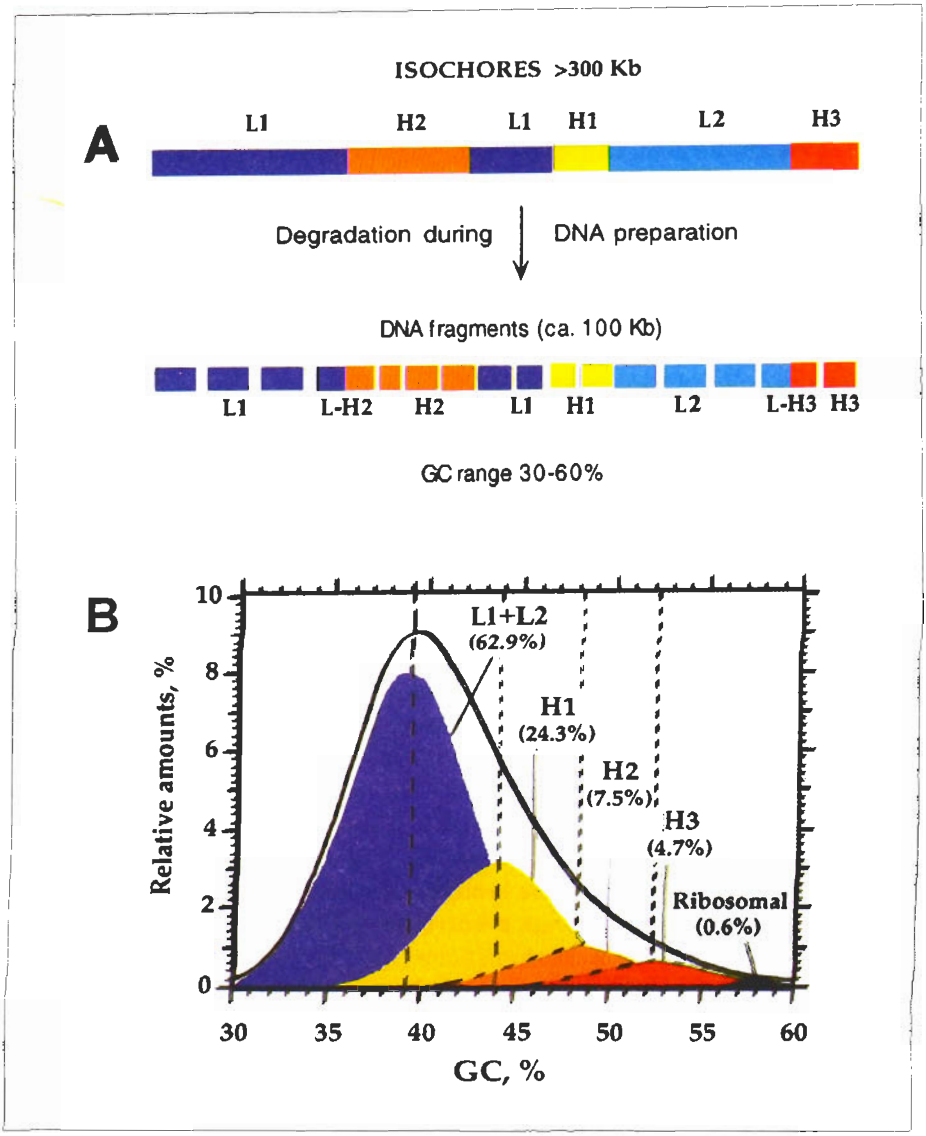


Figura 1. A. Schema dell'organizzazione delle isocore nel genoma umano. Questo genoma, che rispecchia il genoma della maggior parte dei mammiferi, è un mosaico di lunghi segmenti di DNA, le isocore, che sono abbastanza omogenei dal punto di vista della composizione e possono essere suddivise in un piccolo numero di famiglie, povere in GC (L1, L2 e H1), e ricche in GC (H2 e H3). Le isocore vengono degradate durante la preparazione del DNA in frammenti [di 50-100 kb]. Lo spettro di GC di queste molecole di DNA del genoma umano è estremamente ampio, dal 30% al 60% (da Bernardi, 1995). **B.** Il profilo in CsCl del DNA umano è risolto nei suoi componenti, ossia in frammenti di DNA derivati da ognuna delle famiglie di isocore. I livelli modali di GC delle famiglie di isocore sono indicati sulle ascisse (linee verticali interrotte). Le quantità relative dei componenti del DNA sono indicati. I DNA satelliti (che costituiscono solo una piccola percentuale del genoma umano) non sono rappresentati. (Da Zoubak et al., 1996).

Ricerche successive definirono le ampie regioni di DNA (di oltre 300 kb; Macaya et al., 1976) che furono chiamate isocore (per regioni dalla composizione uguale). L'esistenza e le caratteristiche delle isocore sono state confermate e visualizzate venticinque anni più tardi (Pavlicek et al., 2002) non appena fu disponibile la prima sequenza del genoma umano (Lander et al., 2001; Venter et al., 2001; vedi Fig. 2).

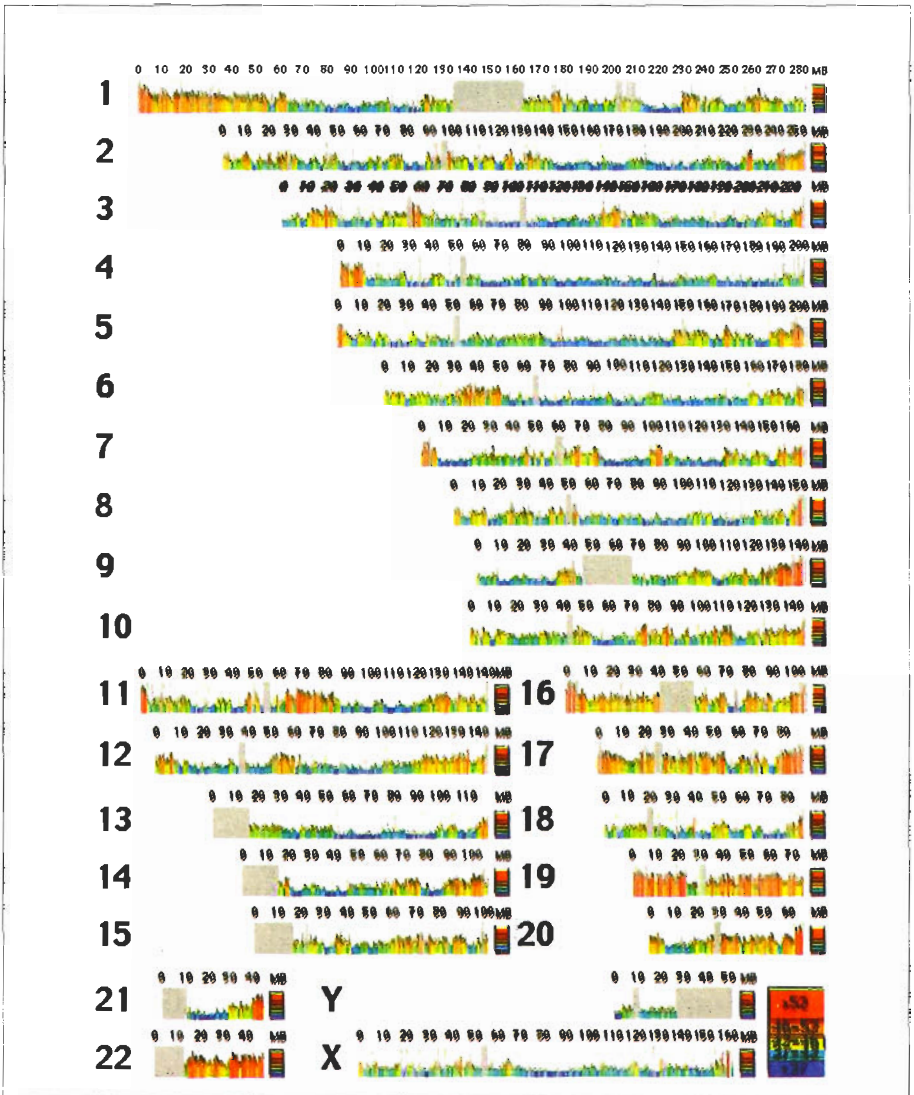


Figura 2. Mappa composizionale dei cromosomi umani. I livelli di GC che vanno dal blu ultramarino (isocore più povere in GC) al rosso scarlatto (isocore più ricche in GC). (Da Costantini et al., 2005).

I fenotipi del genoma

In contrasto con i vertebrati a sangue caldo, i vertebrati a sangue freddo mostrano una eterogeneità composizionale meno evidente, in quanto le isocore più ricche in GC non raggiungono i livelli di quelle dei vertebrati a sangue caldo (Fig. 3). E' interessante notare che i pattern composizionali delle sequenze codificanti (che rappresentano solo l'1-2% del genoma della maggior parte dei vertebrati) sono simili a quelli dei genomi corrispondenti. Entrambi i pattern composizionali equivalgono a "fenotipi del genoma" (vedi Fig. 3). Questo è un concetto nuovo rispetto al fenotipo classico, che è rappresentato da forma e funzione o, in termini molecolari, dalle proteine e dalla loro espressione.

E' importante sottolineare che l'organizzazione in isocore del genoma non è limitata ai soli vertebrati, ma è molto diffusa tra gli eucarioti in quanto è stata osservata anche nelle piante, nei tripanosomi, etc. (vedi Bernardi, 2004).

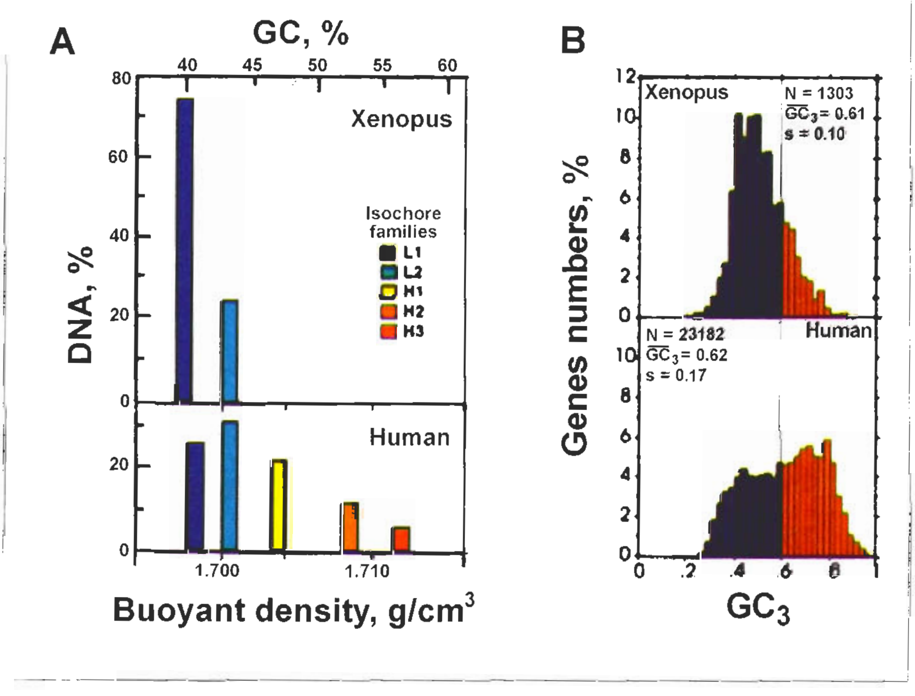


Figura 3. A. Famiglie di isocore da *Xenopus* e uomo dedotte da centrifugazione in gradiente di densità. **B.** Pattern composizionali di sequenze codificanti (rappresentati da valori di GC_3 calcolati per sequenza codificata. GC_3 è il valore del GC nella terza posizione dei codoni) per *Xenopus* e uomo. (Modificata da Bernardi, 1995).

Il codice genomico

Una ovvia domanda riguarda la possibile esistenza di una correlazione tra la composizione delle sequenze codificanti e delle sequenze non codificanti contigue. La risposta è affermativa. Infatti, le sequenze codificanti ricche in GC sono contigue a sequenze non codificanti ricche in GC, mentre le sequenze codificanti povere in GC sono contigue a sequenze non codificanti povere in GC (Fig. 4A). In realtà, il codice genomico (come furono definite tali correlazioni) va oltre, in quanto correlazioni composizionali sussistono anche tra gli esoni e gli introni dei geni (Fig. 4B) e tra le posizioni I,II,III dei codoni (Fig. 4C). Queste ultime vengono definite correlazioni universali, in

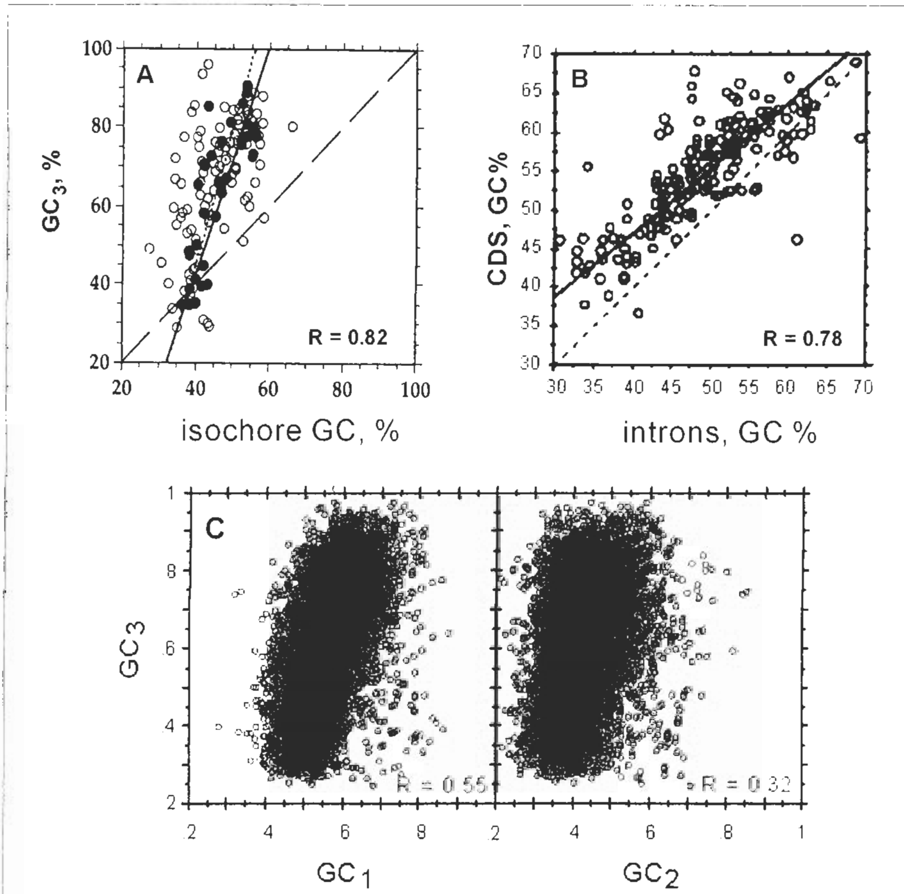


Figura 4. A. Correlazioni tra i valori medi di GC_3 di geni umani e il livello di GC di frazioni di DNA, o YACs (cromosomi artificiali del lievito), in cui i geni sono localizzati (punti scuri), o di sequenze 3' contigue (punti chiari; da Zoubak et al., 1996). **B.** Correlazioni tra il livelli di GC di sequenze codificanti umane (CDS) e degli introni corrispondenti. (Modificata da Clay et al., 1996). **C.** Correlazioni tra GC_3 e GC_1 o GC_2 per 20.148 geni umani. (Modificata da Jabbari et al., 2003).

quanto sono presenti in tutti gli organismi, dai procarioti ai mammiferi (D'Onofrio e Bernardi, 1992). Inoltre, esistono correlazioni tra la composizione di sequenze codificanti da un lato e l'idrofobicità e la struttura secondaria (aperiodica, ad elica e a foglietto) delle proteine codificate dall'altro (Fig. 5).

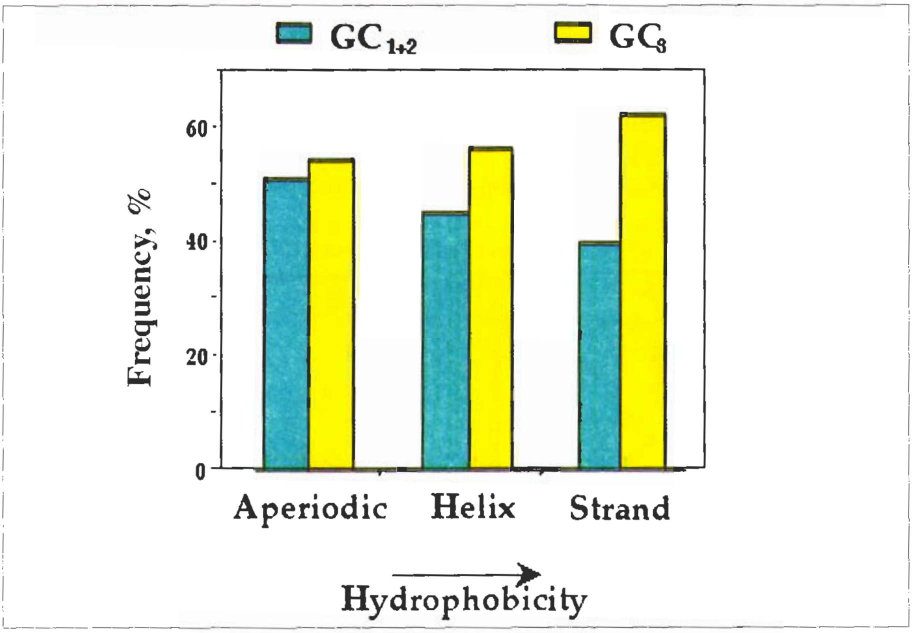
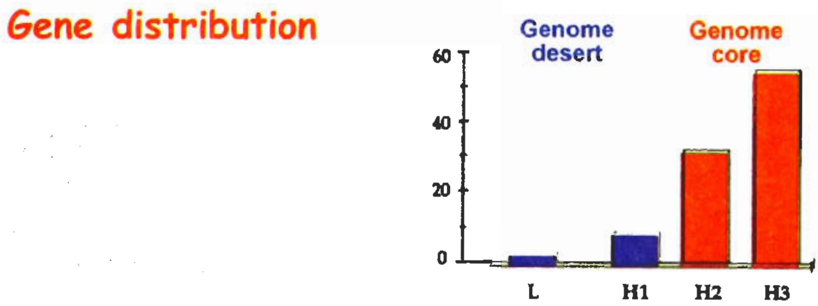


Figura 5. Istogramma delle frequenze di GC_3 e GC_{1+2} in tre strutture secondarie delle proteine, ordinate secondo il valore crescente di idrofobicità. (Da D'Onofrio et al., 2002).

E' da sottolineare che il codice genomico ha fornito la prima prova del fatto che il genoma degli eucarioti è un insieme integrato. Si potrebbe anche dire che il codice genomico ha dimostrato, per parafrasare Galileo, che il libro del genoma è scritto in un linguaggio matematico. Inoltre, il codice genomico non può essere riconciliato con il concetto che le sequenze non codificanti siano "junk DNA", DNA spazzatura (Ohno, 1972) e con il concetto che le "sequenze ripetute intersperse", come le Alu e le LINE, siano "selfish DNA", DNA egoista (Doolittle e Sapienza, 1980; Orgel e Crick, 1980). Infatti, se la composizione in basi delle sequenze codificanti è sotto selezione, lo devono essere anche le sequenze non codificanti in quanto la loro composizione è correlata con quella delle sequenze codificanti.

La distribuzione dei geni nelle isocore e nei cromosomi

Due proprietà molto importanti dei vertebrati (ed anche di altri eucarioti) riguardano la distribuzione dei geni nel genoma e nei cromosomi e le correlazioni di tali distribuzioni con altre caratteristiche strutturali e funzionali (Fig. 6). La prima indicazione circa la sorprendente non uniformità di questa distribuzione genica risale a venti anni fa (Bernardi et al., 1985). Successive ricerche (Mouchiroud et al., 1991; Zoubak et al., 1996) identificarono due spazi genici, un "genome core" (o nucleo del genoma, corrispondente alle famiglie di isocore più ricche in GC, H2 e H3) caratterizzato da un'elevata densità genica, ed un "genome desert" (o deserto del genoma,



Correlations with structure and function

Intron, UTR size	Large	Small
Chromatin structure	Closed	Open
GC Heterogeneity	Low	High
Gene expression	Low	High
Replication timing	Late	Early
Recombination	Low	High

Figura 6. Distribuzione dei geni nel genoma umano. Le proprietà strutturali e funzionali associate con ogni spazio del genico sono indicate.

corrispondente alle famiglie di isocore più povere in GC, L1, L2 e H1), in cui si trova una bassa densità genica (Fig. 6). Il *genome desert* e il *genome core* rappresentano rispettivamente l'88% e il 12 % del genoma, mentre il numero dei geni è approssimativamente lo stesso nei due spazi genici.

I due spazi genici sono associati, come già accennato, con differenti proprietà strutturali e funzionali (vedi Fig. 6). Tra le proprietà strutturali è da notare che gli introni e le regioni non tradotte (UTR) sono lunghi nel *genome desert* e corti nel *genome core*. Nei cromosomi metafasici, le regioni dense in geni delle isocore H2 e H3 sono predominanti nelle regioni telomeriche, mentre le regioni povere in geni si trovano per lo più vicino ai centromeri (Fig. 7A). Da notare che le regioni cromosomiche più ricche in geni sono caratterizzate da una cromatina molto "aperta" e sono posizionate al centro del nucleo interfasico, mentre le regioni più povere in geni sono addensate contro la membrana nucleare (Fig. 7B).

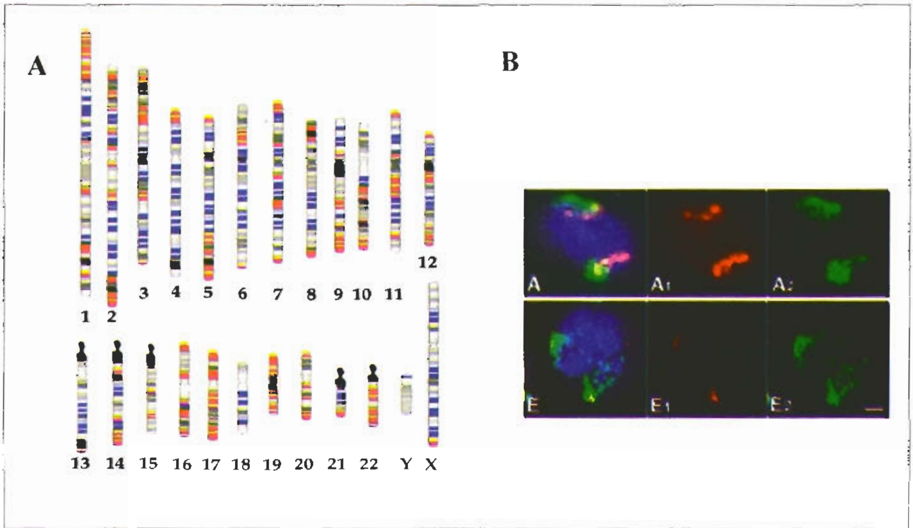


Figura 7. A. Identificazione delle bande cromosomiche più povere e più ricche in GC. Il cariotipo umano ad una risoluzione di 850 bande mostra le bande cromosomiche che contengono le isocore più povere in GC (L1+, bande blu) e più ricche in GC (H3+, bande rosse). Le prime corrispondono alle bande G(iemsa), le seconde alle bande R(everse). Le bande "intermedie" (il 50% delle 850 bande) sono mostrate in bianco per le bande R H3$^-$, in grigio (secondo la ripartizione di Francke, 1994) per le bande G L1$^-$. (Modificata da Federico et al., 2000). **B.** Distribuzione nucleare delle regioni cromosomiche caratterizzate da differenti livelli di GC. Regioni cromosomiche di 15-20 Mb corrispondenti alle bande più ricche in GC, 6p21 (A) e più povere in GC, 12q21 (E) sono state co-ibridate con le corrispondenti sonde cromosomiche. I DNA delle bande e dei cromosomi sono stati rispettivamente marcati con biotina (segnali rossi) e digoxigenina (segnali verdi). I nuclei sono stati colorati con DAPI (blu). (Da Saccone et al., 2002).

Quest'ultima scoperta è riprova non solo del fatto che le regione dense in geni sono caratterizzate da un alto livello di trascrizione rispetto alle regioni povere in geni, ma anche della necessità per la loro cromatina aperta di una stabilizzazione termodinamica che è stata conseguita grazie ad un arricchimento in G e C nei vertebrati a sangue caldo. Al contrario, le regioni povere in geni sono stabilizzate dalla loro struttura cromatinica chiusa. Altre proprietà funzionali distintive riguardano il tempo di replicazione (tardivo nel *genome desert* e precoce nel *genome core*) e la ricombinazione che è frequente nel primo caso e raro nel secondo.

In conclusione, la distribuzione bimodale dei geni rivelata dall'approccio composizionale caratteristico delle nostre ricerche è fortemente correlato ad essenziali caratteristiche strutturali, funzionali ed evolutive del genoma.