Rendiconti
Accademia Nazionale delle Scienze detta dei XL
Memorie di Scienze Fisiche e Naturali
123° (2005), Vol. XXIX, t. I, pp. 7-17

GIORGIO BERNARDI*

Una visione ultra-darwiniana dell'evoluzione del genoma**

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 (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 correlazioni positive tra i livelli di GC delle sequenze codificanti e delle sequenze non codificanti contigue.

La distribuzione dei geni è bimodale nei vertebrati, la densità genica è molto alta nel piccolo *genome core* (che nel genoma umano corrisponde alle famiglie di isocore più ricche in GC), ed è invece molto bassa nel vasto *genome desert* (che è formato dalle altre famiglie di isocore). La distribuzione dei geni nei vertebrati è correlata ad importanti caratteristiche strutturali, funzionali ed evoluzionistiche.

Per quanto riguarda l'evoluzione del genoma dei vertebrati è importante notare che il *genome core* è andato incontro ad una transizione composizionale (arricchimento in GC) per far fronte all'emergenza della omeotermia dei vertebrati a sangue caldo, mentre il *genome desert* non è cambiato in termini di composizione in basi. L'evoluzione del genoma dei vertebrati è discussa in relazione alla teoria neo-selezionista.

Parole chiave: isocore, geni, cromatina, cromosomi, vertebrati

La compartimentazione del genoma

Circa trenta anni fa, si è visto 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 è stata rapidamente estesa ad altri eucarioti. In particolare, si

* Socio dell'Accademia. Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Napoli. E-mail: bernardi@szn.it

** Prolusione per l'inaugurazione del 223° Anno Accademico. Roma, 14 marzo 2005, "Sala Igea", Palazzo Mattei, Piazza dell'Enciclopedia Italiana 4, Roma.

è visto che i genomi dei vertebrati a sangue caldo mostravano essenzialmente le caratteristiche composizionali appena descritte per il genoma bovino (Thiery et al., 1976). Ad esempio, nel genoma umano sono state individuate ed isolate fisicamente cinque famiglie di molecole di DNA (escludendo i DNA satellite e ribosomale). Queste famiglie, caratterizzate da un ordine crescente di GC, sono state chiamate L1, L2, H1, H2 e H3; le prime tre comprendevano circa l'85% del DNA e le ultime due circa il 15% (Fig. 1). Ricerche successive hanno dimostrato che le molecole di DNA inizialmente studiate derivavano da ampie regioni di DNA (di oltre 300 kb; Macaya et al., 1976) che sono state chiamate isocore (regioni con uguale composizione).

L'esistenza e le caratteristiche delle isocore sono state confermate e visualizzate venticinque anni più tardi (Pavlicek et al., 2002), quando si è resa disponibile la prima sequenza del genoma umano (Lander et al., 2001; Venter et al., 2001). Studi ulteriori (Costantini et al., 2006), usando una finestra di 100 kb sulla sequenza completa, hanno rivelato che le isocore hanno una lunghezza media di 1 Mb (megabase; 1 milione di basi) e una deviazione standard minore o uguale a 1% GC (per la maggior parte del genoma) o a 2% GC per una minoranza (circa il 15%), che si trova prevalentemente in regioni ricche di GC.

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. 2). È interessante notare che i pattern composizionali delle sequenze codificanti (che rappresentano solo l'1-2% del genoma nella maggior parte dei vertebrati) sono simili a quelli dei genomi corrispondenti. Entrambi i pattern composizionali equivalgono a «fenotipi del genoma» (vedi Fig. 2). Questo è un concetto nuovo rispetto al fenotipo classico, che è rappresentato da forma e funzione o, in termini molecolari, dalle proteine e dal loro livello di espressione. Le differenze mostrate nella Fig. 2 sono molto importanti perché implicano un cambiamento composizionale in una parte importante del genoma (circa il 15%), cambiamento che è avvenuto in coincidenza con la comparsa della omeotermia nei vertebrati a sangue caldo (questo punto sarà discusso successivamente).

È 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).

Il codice genomico

Una ovvia domanda riguarda la possibile esistenza di correlazione tra la composizione delle sequenze codificanti e delle sequenze non codificanti contigue. La risposta è affermativa (Fig. 3A). In realtà, il codice genomico (come è stata definita

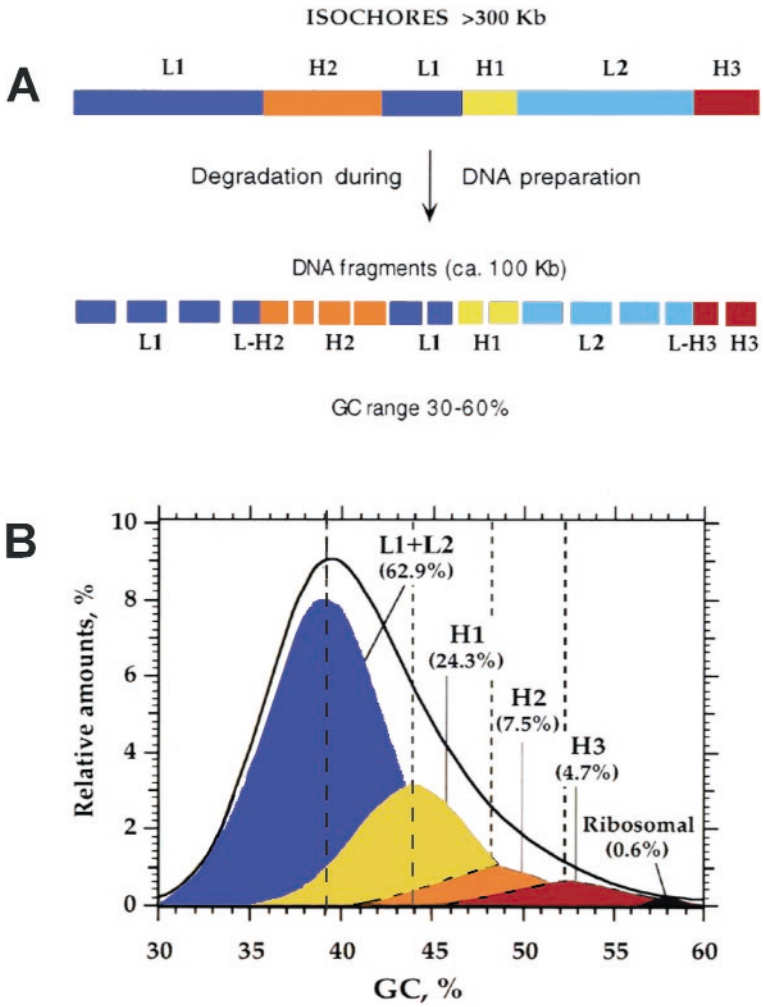


Fig. 1 – **A.** Schema dell'organizzazione delle isocore nel genoma umano. Questo genoma, che rispecchia quello della maggior parte dei mammiferi, è un mosaico di lunghi segmenti di DNA, le isocore, le quali sono piuttosto omogenee dal punto di vista della composizione e possono essere suddivise in un piccolo numero di famiglie: leggere, o povere in GC (L1 e L2), e pesanti, o ricche in GC (H1, 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 di CsCl del DNA umano è risolto nelle sue componenti, ossia in frammenti di DNA derivati da ognuna delle famiglie di isocore (L1, L2, H1, H2 e H3). I livelli modali di GC delle famiglie di isocore sono indicati sulle ascisse (linee verticali discontinue). Le quantità relative dei componenti del DNA sono indicate, mentre i DNA satelliti (che costituiscono solo una piccola percentuale del genoma umano) non sono rappresentati. (Da Zoubak et al., 1996).

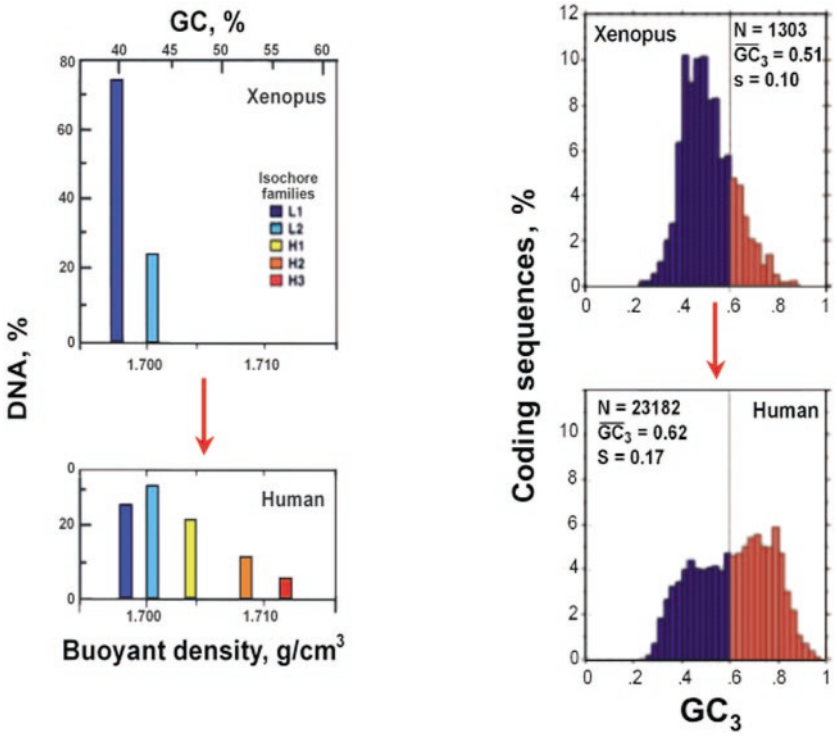


Fig. 2 – **A.** Famiglie di isocore da *Xenopus* e uomo ottenute da centrifugazione in gradiente di densità. **B**. Pattern composizionali di sequenze codificanti (rappresentati da valori di GC_3 calcolati per sequenza codificata; GC_3 è la percentuale di GC nella terza posizione dei codoni) per *Xenopus* e uomo. (Modificata da Bernardi, 1995).

tale correlazione; da non confondere col codice genetico) comprende anche le correlazioni che sussistono tra gli esoni e gli introni dei geni (Fig. 3B). Infine, esistono correlazioni tra la composizione di sequenze codificanti, l'idrofobicità e la struttura secondaria (aperiodica, ad elica e a foglietto) delle proteine codificate (Fig. 4).

È 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 concetto che le sequenze non codificanti siano «junk DNA» (Ohno, 1972) e che le «sequenze ripetute intersperse», come le Alu e le LINE, siano «selfish DNA» non può essere riconciliato con il codice genomico (Doolittle e Sapienza, 1980; Orgel e Crick, 1980). Infatti, se la composizione in basi delle sequenze codificanti è sotto selezione (vedi di seguito), lo devono essere anche le sequenze non codificanti, in quanto la loro composizione è correlata con quella delle sequenze codificanti.

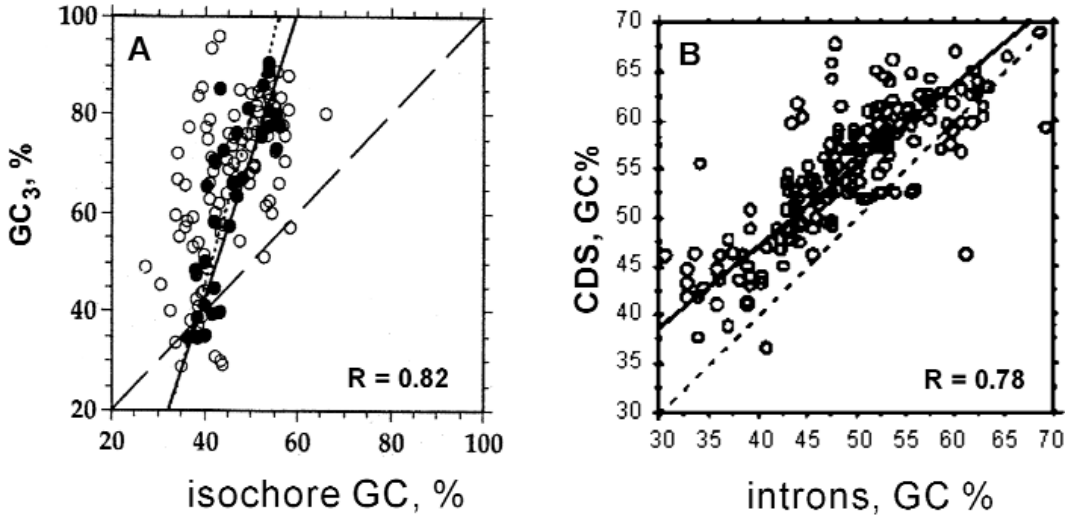


Fig. 3 – **A.** Correlazioni tra il GC_3 di geni umani e il livello di GC di frazioni di DNA, o YACs (cromosomi artificiali di lievito), in cui i geni sono localizzati (punti scuri), o di sequenze 3' contigue (punti chiari; da Zoubak et al., 1996). **B**. Correlazioni tra il livello di GC di sequenze codificanti umane (CDS) e degli introni corrispondenti. (Modificata da Clay et al., 1996).

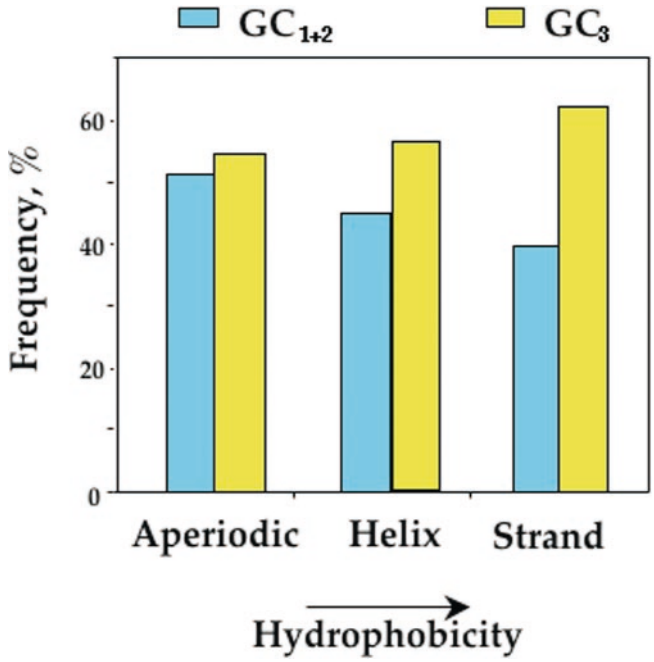
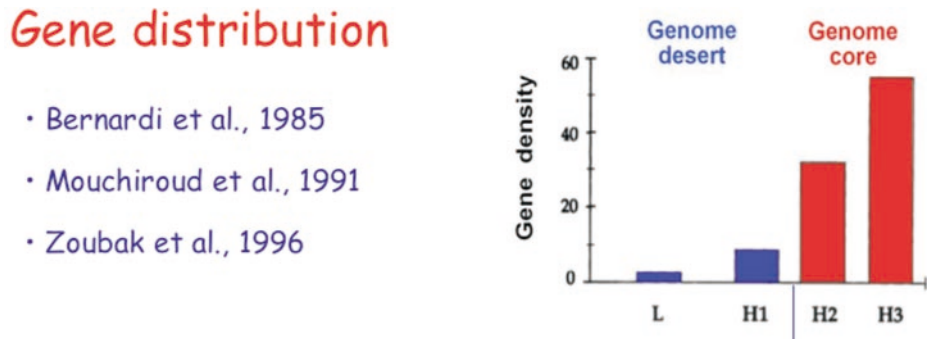


Fig. 4 – Istogramma delle frequenze di GC_3 e GC_{1+2} in tre strutture secondarie delle proteine, ordinate secondo il valore crescente di idrofobicità. (Da D'Onofrio et al., 2002).

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. 5). La prima indicazione circa la sorprendente non uniformità della distribuzione dei geni risale a venti anni fa (Bernardi et al., 1985). Successive ricerche (Mouchiroud et al., 1991; Zoubak et al., 1996), hanno permesso di identificare due spazi genici, un *genome core* (corrispondente alle famiglie di isocore più ricche in GC, le famiglie H2 e H3) caratterizzato da un'elevata densità genica, ed un *genome desert* (corrispondente alle famiglie di isocore più povere in GC, le famiglie L1, L2 e H1), in cui si trova una bassa densità genica (Fig. 5). Il *genome desert* e il *genome core* rappresentano rispettivamente l'85% ed il 15% del genoma, mentre il numero dei geni è approssimativamente lo stesso nei due spazi genici, i quali sono associati con differenti proprietà strutturali e funzionali (vedi Fig. 5).

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 meta-



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

Fig. 5 – Distribuzione dei geni nel genoma umano. Sono riportate le proprietà strutturali e funzionali associate ad ogni spazio genico.

fasici, le regioni dense di geni delle isocore H2 e H3 sono prevalentemente localizzate nelle regioni telomeriche, mentre le regioni povere in geni si trovano per lo più vicino ai centromeri (Fig. 6). Inoltre le regioni cromosomiche più ricche in geni sono posizionate al centro del nucleo interfasico, mentre le regioni più povere in geni sono addensate contro la membrana nucleare (Fig. 7). Infine, una differenza importante tra le due regioni è la struttura della cromatina: «aperta» nelle regioni ricche in geni, «chiusa» nella regioni povere in geni.

Quest'ultima scoperta si ricollega al fatto che le regione dense in geni sono caratterizzate da un alto livello di trascrizione rispetto alle regioni povere in geni.

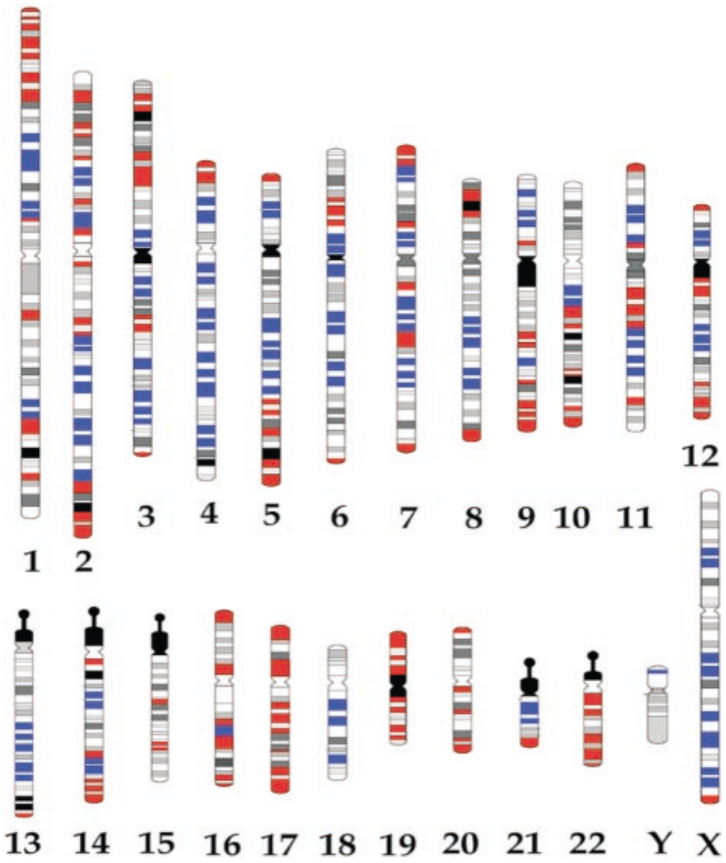


Fig. 6 – Identificazione delle bande cromosomiche più povere e più ricche in GC. Il cariotipo umano, ad una risoluzione di 850 bande, mostra le bande cromosomiche che contengono le isocore più povere in GC (L1+, bande blu) e più ricche in GC (H3+, bande rosse). Le prime corrispondono alle bande G(iemsa), le seconde alle bande R(everse). Le bande «intermedie» (il 50% delle 850 bande che sono mediamente ricche in GC ed in geni) sono mostrate in bianco per le bande R $H3^-$ ed in grigio (secondo la ripartizione di Francke, 1994) per le bande G $L1^-$. (Modificata da Federico et al., 2000).

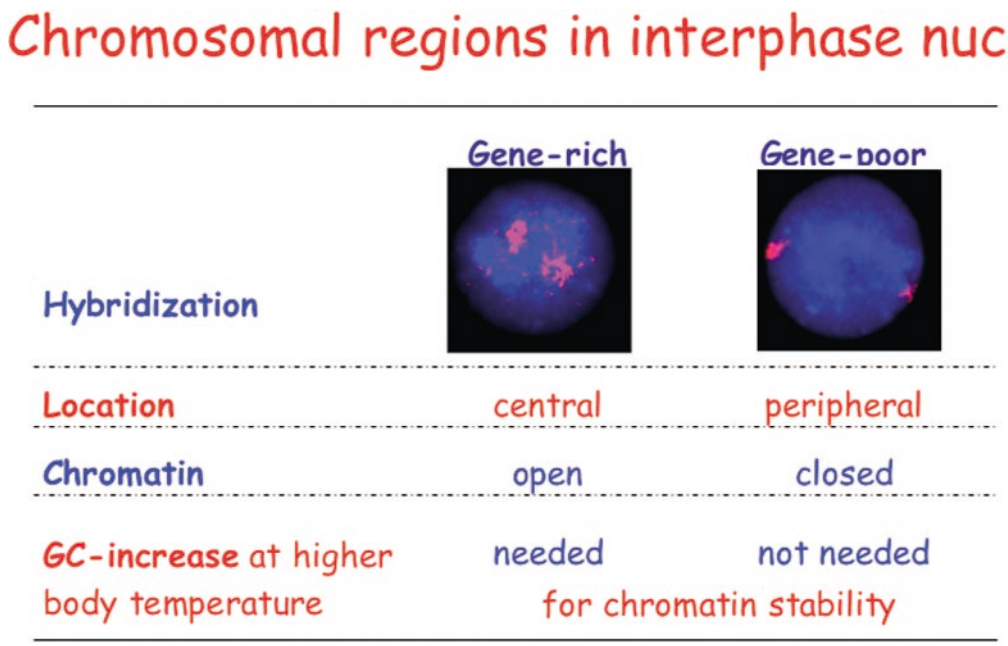


Fig. 7 – Distribuzione nucleare delle regioni cromosomiche caratterizzate da differenti livelli di GC. Regioni cromosomiche di 15-20 Mb corrispondenti alle bande più ricche e più povere in GC sono state ibridate sul DNA di nuclei interfasici. I DNA sono stati rispettivamente marcati con biotina (segnali rossi) e digoxigenina (segnali verdi). I nuclei sono stati colorati con DAPI (blu). (Da Saccone et al., 2002).

Inoltre fornisce una spiegazione al fatto che solo le regioni ricche in geni hanno subito un cambiamento composizionale (arricchimento in GC) nella transizione tra vertebrati a sangue freddo e vertebrati a sangue caldo (Fig. 8), dal momento che le due grandi classi di vertebrati conservano una composizione stabile nel tempo. La transizione composizionale è stata attribuita (Bernardi e Bernardi, 1986) alla necessità di stabilizzare termodinamicamente il DNA alle più alte temperature corporee dei vertebrati a sangue caldo. Il motivo per cui questo cambiamento abbia avuto luogo solo nel *genome core* è stato spiegato (Saccone et al., 2002) dal fatto che, mentre le regioni ricche in geni, la cui cromatina è aperta, hanno avuto bisogno di una stabilizzazione termodinamica (conseguita grazie ad un arricchimento in G e C nei vertebrati a sangue caldo), le regioni povere in geni sono stabilizzate dalla loro struttura cromatinica chiusa. È interessante notare che l'arricchimento in GC del DNA ha anche condotto ad una stabilizzazione dell'RNA e delle proteine codificate, poiché i codoni più ricchi in GC producono aminoacidi che stabilizzano termodinamicamente le proteine (Bernardi e Bernardi, 1986).

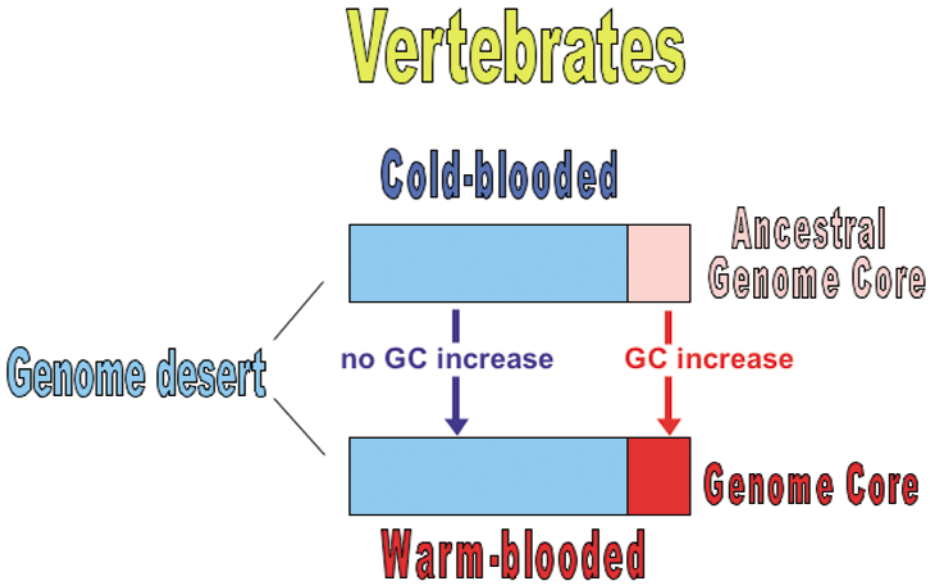


Fig. 8 – Schema della transizione composizionale mostrata dai genomi dei vertebrati a sangue freddo. Mentre il *genome desert*, povero in GC ed in geni (box blu), non ha subito alcun cambiamento composizionale, il *genome core* ancestrale dei vertebrati a sangue freddo da moderatamente ricco in GC (box rosa) è divenuto molto ricco in GC per diventare il *genome core* dei vertebrati a sangue caldo (box rosso).

La teoria neo-selezionista: una teoria ultra-darwiniana

L'esistenza di isocore, di famiglie di isocore, la conservazione ed i cambiamenti di fenotipi del genoma dei vertebrati, nonché le correlazioni composizionali sono fenomeni che non possono essere spiegati dalla teoria neutra. Questa conclusione (Bernardi e Bernardi, 1986) deve tuttavia essere riconciliata col fatto innegabile che la maggior parte delle mutazioni sono neutre, perché l'enorme maggioranza del genoma dei vertebrati non codifica. La teoria neo-selezionista (Bernardi, 2004) opera questa riconciliazione, ipotizzando che la maggior frequenza delle mutazioni GC _ AT (rispetto ad AT _ GC) porta ad una diminuzione locale di GC. Superata una certa soglia, questa conduce ad una alterazione della struttura della cromatina e, di conseguenza, ad una diminuita espressione dei geni localizzati nelle regioni affette dai cambiamenti composizionali. A sua volta, questa variazione epigenomica conduce ad una selezione negativa dei portatori di questi cambiamenti (Fig. 9). In altre parole, la teoria neo-selezionista afferma che anche i cambiamenti neutri (ammessi da Darwin come non soggetti alla selezione naturale) in definitiva sono soggetti a selezione. Per questa ragione la teoria neo-selezionista è una teoria ultra-darwininana.

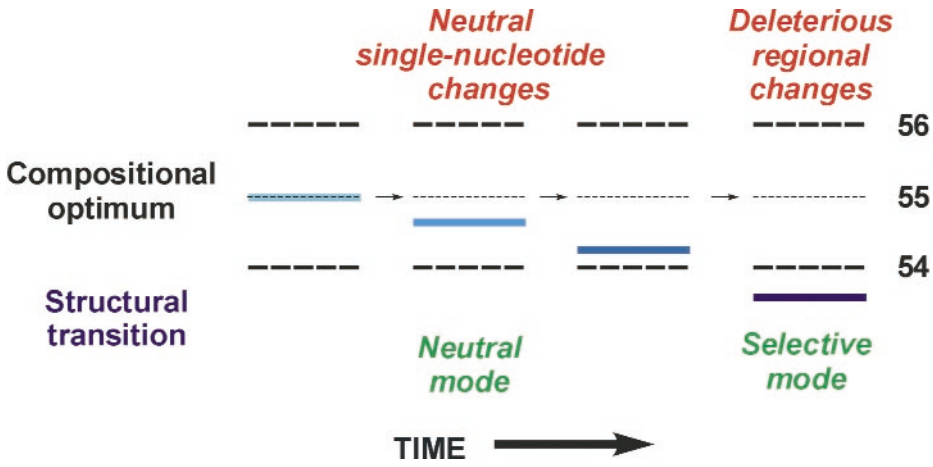


Fig. 9 – Evoluzione in funzione del tempo dei cambiamenti composizionali tipici di una regione intergenica ricca in GC di un vertebrato a sangue caldo. Sappiamo che esiste una forte tendenza all'arricchimento in AT. In una prima fase, il livello medio di GC della regione, che inizialmente coincide con un optimum di composizione (55% GC, ad esempio), diminuisce ma rimane in una zona tollerata dalla espressione dei geni (le soglie arbitrarie delle zone sono indicate dalle linee nere discontinue). In una fase successiva, il GC medio oltrepassa il livello inferiore, la cromatina subisce un cambiamento strutturale, che è deleterio per l'espressione dei geni e conduce ad una selezione negativa dei portatori e della loro discendenza. Fino a quel punto le mutazioni sono neutre o quasi neutre.

LAVORI CITATI

Bernardi G (1995) The human genome: organization and evolutionary history. *Annual Review of Genetics*, **29**, 445-476.

Bernardi G (2004) *Structural and Evolutionary Genomics. Natural Selection in Genome Evolution*. Elsevier: Amsterdam.

Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. *Journal of Molecular Evolution*, **24**, 1-11.

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. *Science*, **228**, 953-957.

Clay O, Cacciò S, Zoubak S, Mouchiroud D, Bernardi G (1996) Human coding and non-coding DNA: compositional correlations. *Molecular Phylogenetics and Evolution*, **5**, 2-12.

Costantini M, Clay O, Auletta F, Bernardi G (2006) An isochore map of human chromosomes. *Genome Research* (submitted).

D'Onofrio G, Ghosh TC, Bernardi G (2002) The base composition of the genes is correlated with the secondary structures of the encoded proteins. *Gene*, **300**, 179-187.

Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. *Nature*, **284**, 601-603.

Federico C, Andreozzi L, Saccone S, Bernardi G (2000) Gene density in the Giemsa bands of human chromosomes. *Chromosome Research*, **8**, 737-746.

Filipski J, Thiery JP, Bernardi G (1973) An analysis of the bovine genome by Cs2SO4-Ag+ density gradient centrifugation. *Journal of Molecular Biology*, **80**, 177-197.

Francke W (1994) Digitized and differentially shaded human chromosome ideograms for genomic applications. *Cytogenetics and Cell Genetics*, **6**, 206-219.

Lander ES et al. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.

Macaya G, Thiery JP, Bernardi G (1976) An approach to the organization of eukaryotic genomes at a macromolecular level. *Journal of Molecular Biology*, **108**, 237-254.

Mouchiroud D, D'Onofrio G, Aïssani B, Macaya G, Gautier C, Bernardi G (1991). The distribution of genes in the human genome. *Gene*, **100**, 181-187.

Ohno S (1972) So much "junk" DNA in our genome. *Brookhaven Symposia Biology*, **23**, 366-370.

Orgel LE and Crick FH (1980) Selfish DNA: the ultimate parasite. *Nature*, **284**, 604-607.

Pavlicek A, Paces J, Clay O, Bernardi G. (2002) A compact view of isochores in the draft human genome sequence. *FEBS Letters*, **511**, 165-169.

Saccone S, Federico C, Bernardi G (2002) Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. *Gene*, **300**, 169-178.

Thiery JP, Macaya G, Bernardi G (1976) An analysis of eukaryotic genomes by density gradient centrifugation. *Journal of Molecular Biology*, **108**, 219-235.

Venter JC et al. (2001). The sequence of the human genome. *Science*, **291**, 1304-1351.

Zoubak S, Clay O, Bernardi G (1996) The gene distribution of the human genome. *Gene*, 174, 95-102.