

Lezione XXV-XXVI

martedì 29-XI-2011



corso di genomica

Laurea Magistrale Biotecnologia Industriale

aula 8

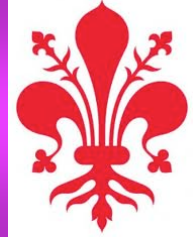
orario : Martedì ore 14.00 - 16.00

Giovedì ore 13.00 - 15.00

libro consigliato Zanichelli “Introduz. alla genomica” di
Arthur M.Lesk

D. Frezza

library of interactions with Hi C



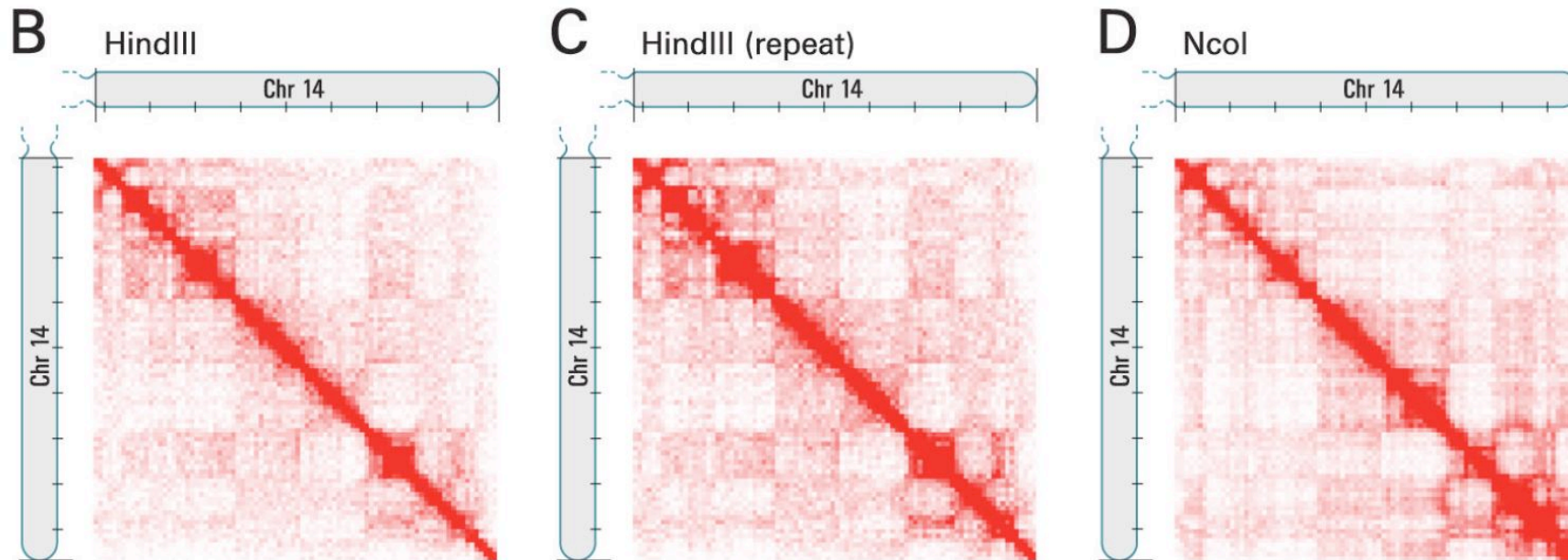
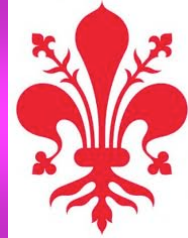
Dopo il resequencing con circa 50 reads di media

Dalla library alla matrice dividendo il genoma in regioni di ~1 Mb dove le “entry” m_{ij} sono il numero di prodotti ligati tra i loci i e j (SOM). Questa matrice riflette la media di interazioni presenti nel campione originale di cellule linfoblastoidi GM06990. La heat-map di intensità fa vedere le frequenze dei contatti (rosso più intenso).fig 1B

I controlli sostituendo HindIII e ripetendo l'analisi fig 1C,D per misurare la ripetibilità.

Esempio su un campione di library del chrms 14

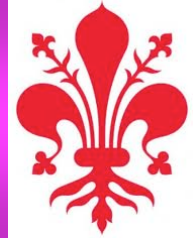
Mapping chromosome 14 interactions



Hi-C produces a genome-wide contact matrix. The submatrix shown here corresponds to intrachromosomal interactions on chromosome 14. Each pixel represents all interactions between a 1Mb locus and another 1Mb locus; intensity corresponds to the total number of reads (0-50). Tick marks appear every 10Mb.

C, D We compared the original experiment to a biological repeat using the same restriction enzyme (C, range: 0-50 reads) and to results with a different restriction enzyme (D, range: 0- 100 reads, NcoI).

Confronti di interazioni note

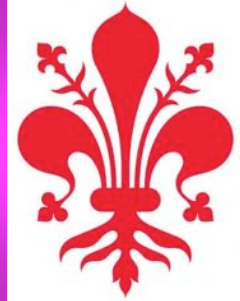


Testare se i dati sono coerenti con strutture note di organizzazioni note come territori cromosomici : - tendenza di loci distanti sullo stesso cromosoma di essere vicini in uno spazio diverso da quello del cromosoma lineare, - patterns in posizioni sub-nucleari dovuti alla tendenza di certe coppie cromosomiche a stare vicini.

Calcolo della probabilità di contatto intracromosomica $I_n(s)$ per coppie di loci separate da distanze genomiche s (distanze in bp lungo la sequenza).

Il valore decresce in modo monotono su ogni cromosoma suggerendo il comportamento polimerico in cui la distanza 3D tra i loci aumenta con la distanza genomica, ciò risulta in accordo con i risultati della FISH. A distanze oltre 200 Mb $I_n(s)$ è molto $>$ della probabilità di contatto tra chrms $\neq$

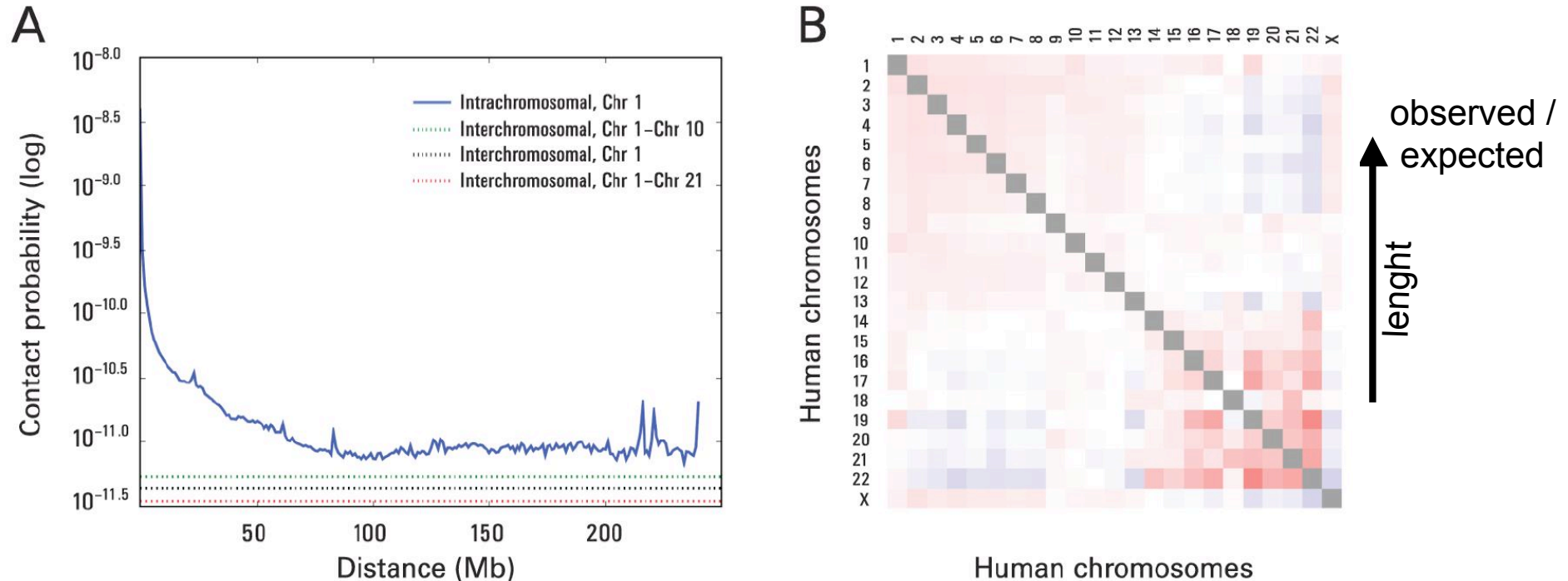
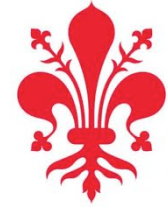
Contatti frequenti di siti lontani: esistono territori chrms



Se le frequenze di contatti osservati sono maggiori di quella casuali ciò implica l'esistenza di territori chrms.

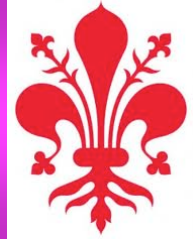
Probabilità di contatti intercromosomici (fig 2B) mostra che i chrms corti più ricchi di geni interagiscono preferenzialmente di più, ciò corrisponde a studi con FISH dove si vede frequente colocalizzazione p.es. crms 11-12, invece il crms 18 tende ad interagire poco e con FISH localizza spesso vicino alla periferia del nucleo.

fig 2 chromosome territories organization



The presence and organization of chromosome territories. **(A)** Probability of contact decreases as a function of genomic distance on chromosome 1, eventually reaching a plateau at ~90Mb (blue). The level of interchromosomal contact (black dashes) differs for different pairs of chromosomes; loci on chromosome 1 are most likely to interact with loci on chromosome 10 (green dashes) and least likely to interact with loci on chromosome 21 (red dashes). Interchromosomal interactions are depleted relative to intrachromosomal interactions. **(B)** Observed/expected number of interchromosomal contacts between all pairs of chromosomes. Red indicates enrichment, and blue indicates depletion (up to twofold). Small, gene-rich chromosomes tend to interact more with one another.

Aumento di risoluzione interna ai crms



Analisi a maggior risoluzione su cromosomi singoli per identificare regioni che associano preferenzialmente.

Normalizzazione della matrice per eliminare il bias dei contatti frequenti dovuti alla vicinanza.

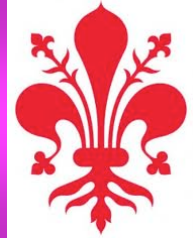
Definizione: dividere ogni entry nella matrice di contatto per la misura media della probabilità dei contatti genomici totali per i loci a quella distanza (SOM)

La matrice normalizzata mostra grandi blocchi di interazioni arricchite o impoverite rispetto alla media attesa con un pattern a plaid rosso/blue (fig 3B)

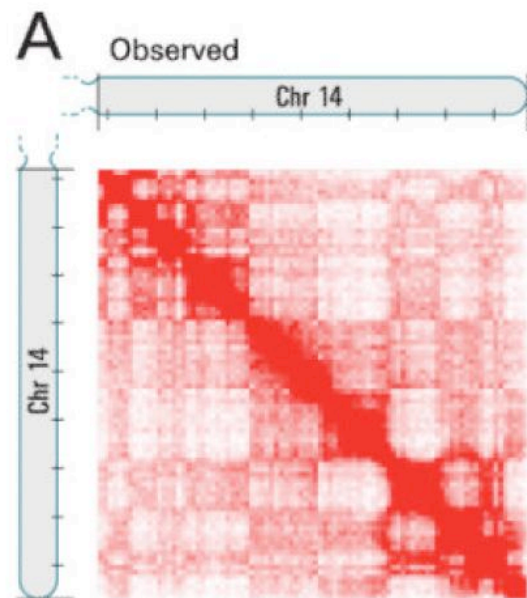
Se due loci in regioni da 1 Mb sono contigui condivideranno i confini ed avranno profili correlati.

Perciò è ottenuta la nuova matrice C in cui c_{ij} è la correlazione di Pearson tra la riga i th e la colonna j th di M^* . Questo ha (rinforzato) acuito/evidenziato il pattern a plaid (fig 3C). Il 71% delle entries ha una significatività con $p \leq 0.05$.

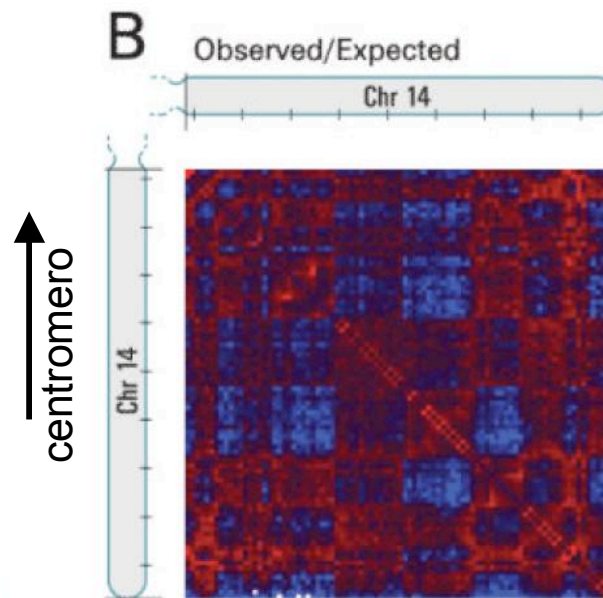
assegnazione di loci a regioni crms fig 3A-B-C



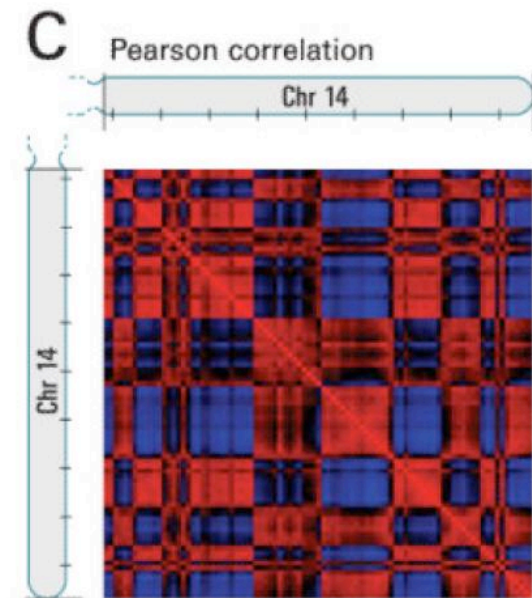
Configurazione a plaid



chrms 14 interact.



normalized pattern
in the borders



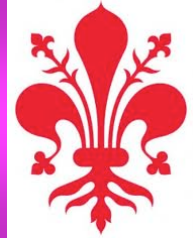
correlation matrix
sharpened

Leggenda fig 3 A-B-C



The nucleus is segregated into two compartments corresponding to open and closed chromatin. (A) Map of chromosome 14 at a resolution of 1Mb (1 tick mark = 10Mb) exhibits substructure in the form of an intense diagonal and a constellation of large blocks (three experiments combined, range: 0-200 reads). The Observed/expected matrix (B) shows loci with either more (red) or less (blue) interactions than would be expected given their genomic distance (range: 0.2 – 5). Correlation matrix (C) illustrates the correlation (red: 1, blue: -1) between the intrachromosomal interaction profiles of every pair of 1 Mb loci along chromosome 14. The plaid pattern indicates the presence of two compartments within the chromosome. The plaid is sharpened for the subtraction of border interactions.

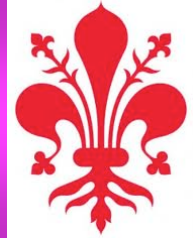
Analisi delle componenti principali



La struttura a plaid suggerisce che ogni cromosoma può essere scomposto in due sets di loci arbitrari “A” e “B” all’interno dei quali i contatti sono arricchiti ed i contatti tra i sets diminuiti. Ogni cromosoma viene suddiviso in questo modo usando l’analisi della componente principale. Per tutti tranne due crms la prima comp. principale CP corrispondeva chiaramente al pattern a plaid con valori positivi che definiscono un set ed i negativi quell’altro in fig S1. Per i crms 4 e 5 la prima CP corrispondeva alle due braccia dei crms, ma la seconda al pattern a plaid. (file S allegato sul sito)

Le entries del vettore CP riflettono le transizioni nitide tra compartimento e compartimento osservate nel pattern a plaid. Inoltre il pattern di ogni cromosoma era in accordo: le etichette A o B stavano in accordo sui diversi cromosomi.

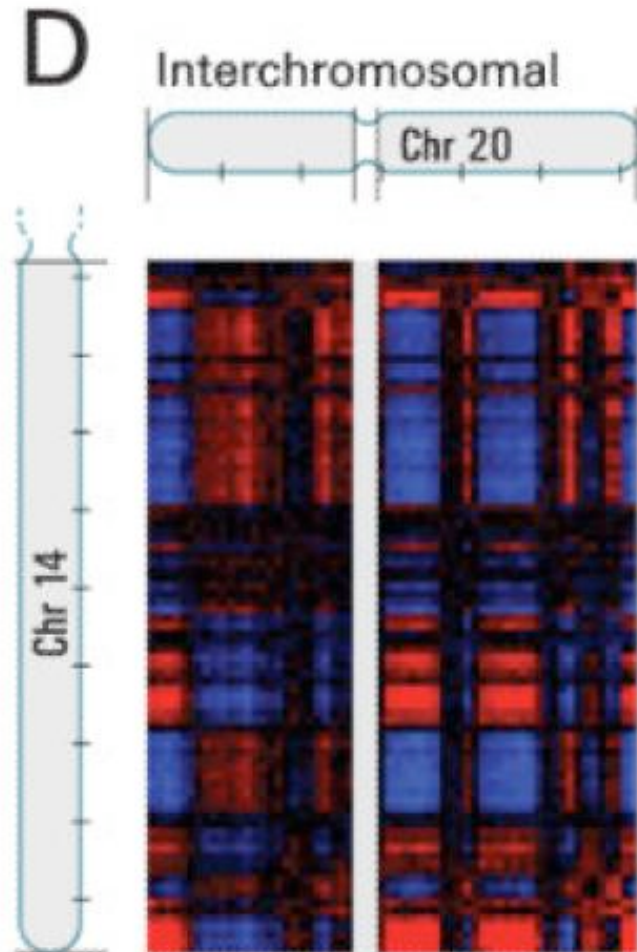
Analisi dei compartimenti A e B



Il pattern a plaid di ogni crms era congruo tra vari crms: le etichette A o B potevano essere assegnate in ogni crms in modo che su crms diversi regioni con le stesse etichette corrispondevano a profili di contatto correlati e regioni con etichette diverse corrispondevano a profili di contatto diversi. Questo implica che l'intero genoma può essere suddiviso in due compartimenti spaziali tali che la maggior parte di interazioni avvengono all'interno di ciascun compartimento più che attraverso compartimenti diversi.

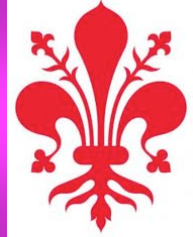
I dati Hi-C implicano che regioni tendono ad essere più vicine se appartengono allo stesso compartimento A vs B che se non vi appartengono.

Figura 3D interazioni crms 14-20



Interchromosomal correlation map for chromosome 14 and chromosome 20 (red: 0.25, blue: 0.25). The unalignable region around the centromere of chromosome 20 is indicated in grey. Each compartment on chromosome 14 has a counterpart on chromosome 20 with a very similar genome-wide interaction pattern. Contacts are not equally (randomly) distributed.

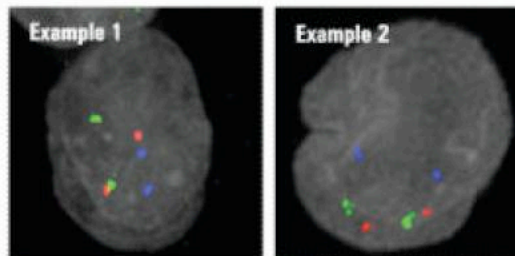
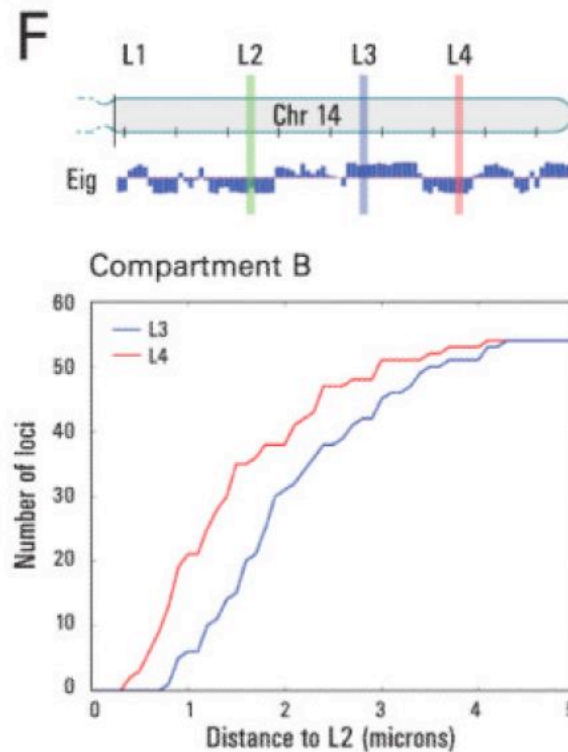
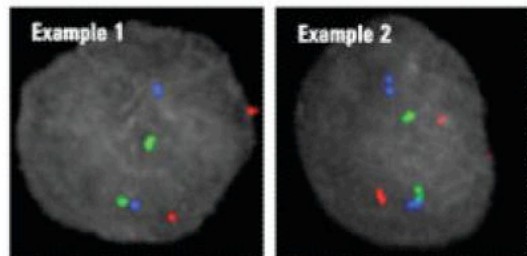
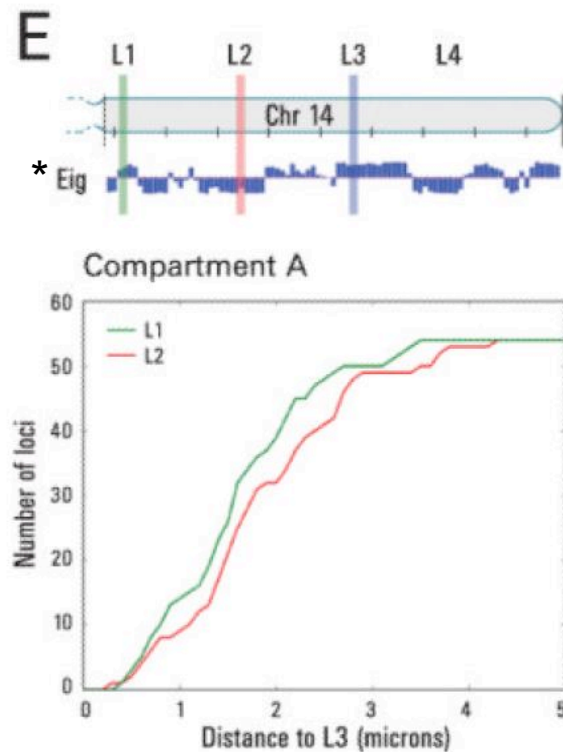
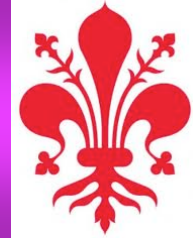
test di conferma con 3D FISH



Conferma con metodo FISH

Testati quattro loci sul crms 14 (L1, L2, L3, L4) che si alternano tra i due compartimenti (L1 e 3 in A; L2 e 4 in B). L'analisi 3D FISH mostra che L3 è più vicina ad L1 che L2 benchè L2 stia nel mezzo tra L1 ed L3 (fig 3E). Similmente L2 è più vicina ad L4 che non ad L3 (fig 3F). Risultati comparabili sono ottenuti con loci consecutivi su crms 22 (fig S2A, B). Tutte insieme queste osservazioni confermano la compartimentalizzazione spaziale del genoma dedotta dalla analisi Hi-C.

distanze tra compartimenti omologhi fig 3E-F

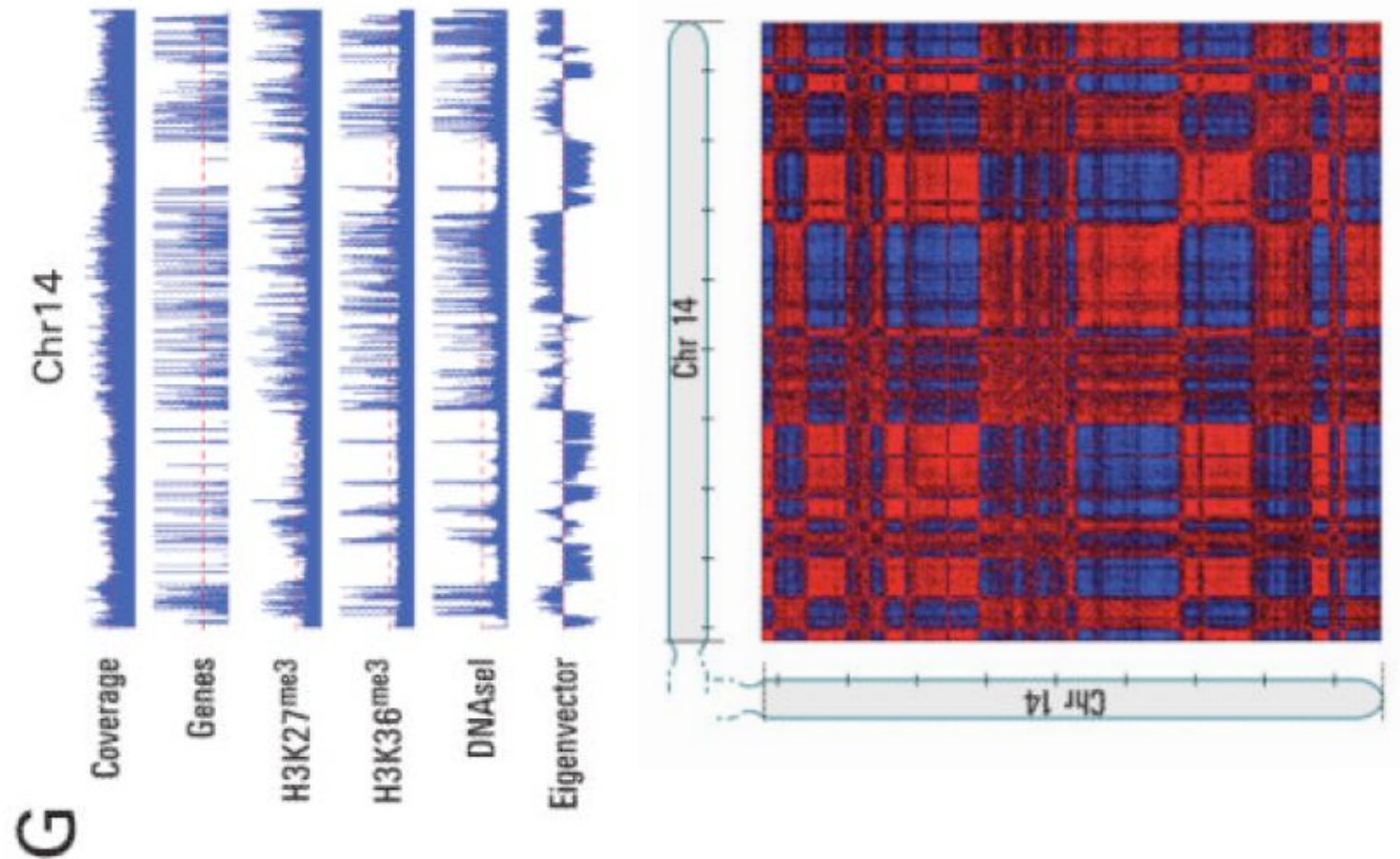
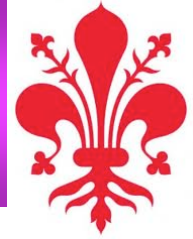


E) L3 blue più vicino ad L1 (verde) che ad L2 (rosso) benchè si trovi nel mezzo
F) L2 rosso più vicino ad L4 che ad L3

L1 ed L3 su Eig mappano su regioni simili così come L2 ed L4

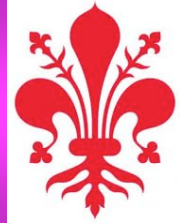
- Eigenvectors of adjacency matrices are useful as measures of centrality or of status. Misure sui margini di confini nelle matrici.
- Vedi fig 3 G

distanze tra compartimenti omologhi fig 3G



Corrispondenza con diverse analisi dei confini tra pattern a plaid e compartimenti, interazioni, distribuzione di geni e cromatina aperta

alta correlazione con 3D FISH

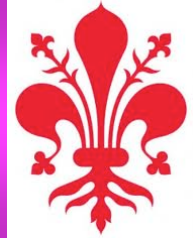


Più in generale la correlazione tra il numero di “reads” Hi-C m_{ij} e le misure di distanza tra i e j ottenute con FISH indicano che il conto dei “read” Hi-C può servire come misuratore per la distanza.

Dopo analisi dettagliata dei dati Hi-C abbiamo notato che coppie di loci nel compartimento B mostravano frequenze più alte di interazione ad una distanza genomica data rispetto a coppie di loci del compartimento A (fig S4).

Questo indica che il compartimento B è più densamente impacchettato. I dati della FISH sono in accordo con queste osservazioni: i loci del compartimento B mostravano tendenza più alta a localizzazioni spaziali vicine.

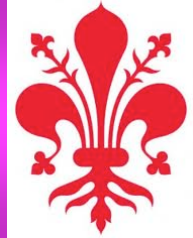
corrispondenza dei compartimenti



Per vedere se i compartimenti spaziali corrispondessero a strutture note del genoma li abbiamo confrontati nelle nostre mappe di correlazione da 1 Mb a conformazioni note genetiche ed epigenetiche. I compartimenti A corrispondevano con la presenza di geni (Spearman's $\rho = 0.431$ $p < 10^{-137}$, alta espressione (wide genome mRNA) Spearman's $\rho = 0.476$ $p < 10^{-145}$ [fig S5] e cromatina accessibile per la sensibilità alla DNaseI, Spearman's $\rho = 0.651$ p trascurabile.

Il compartimento A mostra anche arricchimento di marcatori della cromatina per l'attivazione (H3K36 trimetil. Spearman's $\rho = 0.601$ $p < 10^{-296}$) e repressione H3K27 trimetil. $\rho = 0.282$ $p < 10^{-56}$. Ripetuta l'analisi a maggior risoluzione di 100 Kb (fig 3G) dava una forte conferma con tutte le altre forme genetiche ed epigenetiche, la correlazione crollava solo per il marcatore di repressione H3K27 (Spearman's $\rho = 0.046$ $p < 10^{-15}$). Da questa analisi concludiamo che il compartimento A è fortemente associato con cromatina aperta, accessibile, attivamente trascritta.

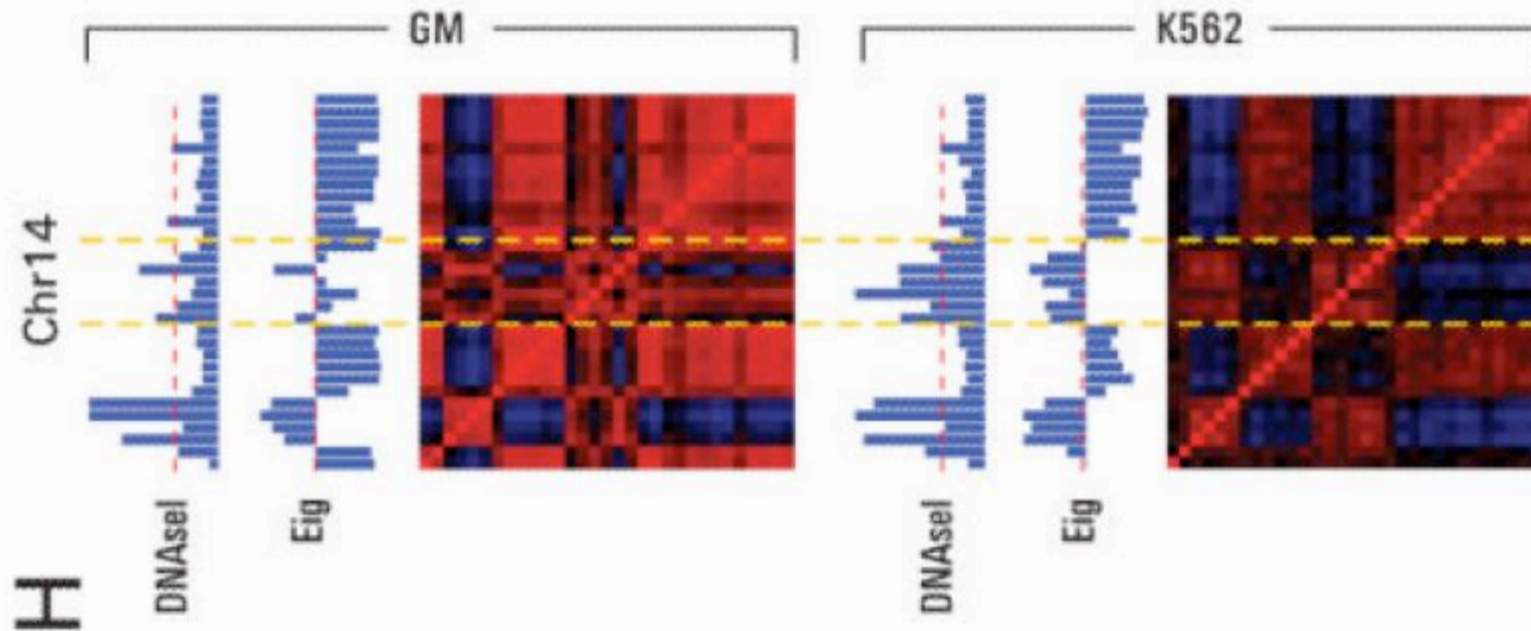
Analisi con linea eritroleucemica K562



In K562 ci sono due compartimenti simili in composizione a quelli osservati con la linea linfoblastoide GM06990, Spearman's $r = 0.732$ $p < \text{trascurabile}$ (fig S9) e mostravano alta correlazione per le condizioni della cromatina aperta e chiusa con analisi della DNaseI (Spearman's $\rho = 0.455$ $p < 10^{-154}$).

I pattern di compartimentazione tra cellule K562 e GM sono simili, ma ci sono molti loci nel compartimento aperto che stanno in quello chiuso nell'altra linea. Analizzando questi loci discordanti su cromosomi con cariotipo normale in K562 si osserva una forte correlazione tra il pattern di compartimentazione e l'accessibilità alla cromatina analizzata nello stesso tipo cellulare. Quindi anche in cellule con genoma riarrangiato la compartimentazione spaziale correla fortemente con lo stato della cromatina.

analisi su due linee GM e K562 fig 3H



DNase, Eigenvector and heat map

A 31Mb window from the chromosome 14 is shown; the indicated region (yellow dashes) alternates between the open and closed in compartment in GM06990 (top, eigenvector and heatmap), but is predominantly open in K562 (bottom, eigenvector and heatmap). The change in compartmentalization corresponds to a shift in chromatin state (DNaseI).

Messaggio: la regione è definita analogamente nei due tipi di cellule.

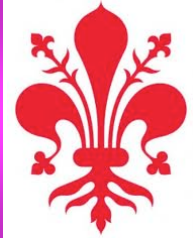
Gli stati della cromatina corrispondono alle frequenze di interazione



I risultati dimostrano che la cromatina aperta e chiusa in tutto il genoma occupano compartimenti spaziali differenti nel nucleo. Queste osservazioni si basano su singoli loci in istanze particolari di interazione; sia tra geni attivi e lontani, sia tra geni lontani ed inattivi. Ultima indagine per vedere la struttura interna dei domini di cromatina aperta e chiusa che corrisponde ai compartimenti visti nelle mappe di correlazione a plaid.

Hanno osservato con alta risoluzione il comportamento medio delle frequenze dei contatti intracromosomici come funzione della distanza genomica, calcolando la distribuzione “genome wide” $I(s)$. Il plot su scala log-log mostra efficienza tra 500Kb e ~ 7Mb con scala di probabilità di contatto $=s^{-1}$ (fig 4A). Questo range corrisponde alla grandezza nota dei domini di cromatina aperta e chiusa.

Per il DNA risoluzione analoga ad un polimero



Comportamento simile a quello della teoria dei polimeri.

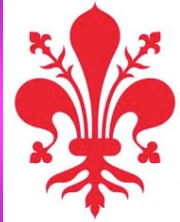
Esiste un modello “equilibrium-globule”: configurazione compatta, densamente annodata usata originariamente per descrivere un polimero in un solvente poco efficiente all’equilibrio.

Grosberg et al. hanno proposto un modello che include il DNA interfascico,

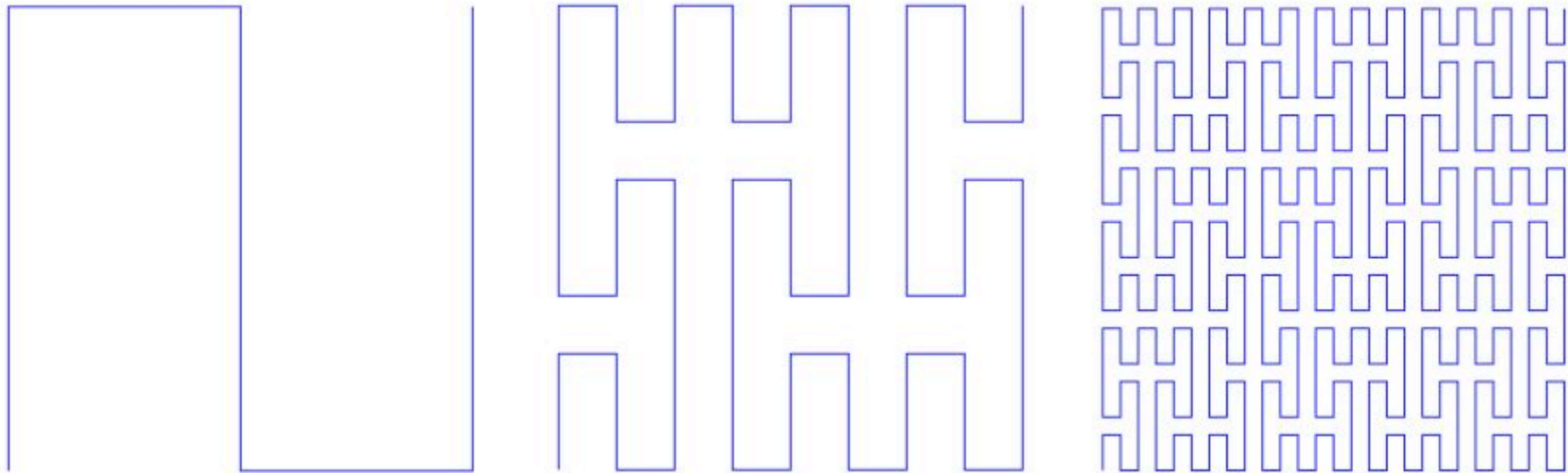
Un polimero si può organizzare non all’equilibrio in conformazione chiamata “fractal globule”. Lo stato altamente compatto è formato da un polimero non aggrovigliato quando si accartocchia in una serie di piccoli globuli in una configurazione a palline su una fibra. Le palline servono come monomeri in “rounds” successivi di arrotolamento finché rimane un singolo globulo di globulo di globulo. La struttura risultante somiglia ad una curva di Peano (figura del frattale) con traiettoria continua che riempie densamente uno spazio tridimensionale senza incrociarsi.

Descrizione utile per la cromatina perché è senza nodi e facilita l’avvolgimento e lo svolgimento durante l’attivazione genica, la repressione o il ciclo cellulare.

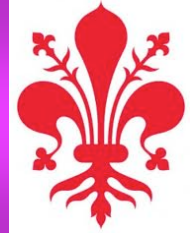
organizzazione vantaggiosa del frattale curva di Peano



In un globulo frattale regioni contigue del genoma tendono a formare settori spaziali la cui dimensione corrisponde alla grandezza della regione originale (fig 4C). Riempimento mantenendo lo spazio iniziale.



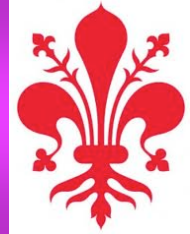
Vantaggi del modello frattale



In contrasto all'equilibrio globulare, il globulo frattale genera predizioni molto diverse che riguardano la scala di probabilità di contatto con la distanza genomica s . Il globulo all'equilibrio predice una probabilità di contatto $s^{-3/2}$, che non si osserva sui dati ottenuti con questo lavoro. L'analisi analitica per un globulo frattale scende ad s^{-1} (SOM) che corrisponde da vicino al valore ottenuto con scala prominente (-1.08).

Il modello di equilibrio e globulo frattale producono previsioni diverse anche sulle distanze tridimensionali 3D tra coppie di loci (fratt. $S^{1/2}$; equil. $S^{1/3}$). Mentre le distanze 3D con Hi-C non sono misurate direttamente, però un nuovo lavoro con 3D-FISH riporta una scala $s^{1/3}$ per distanze genomiche tra 500 Kb e 2 Mb.

Analisi con metodo MonteCarlo

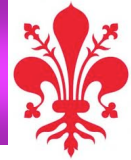


Simulazione tramite metodo MonteCarlo per costruire insiemi di globuli frattali e globuli all'equilibrio.

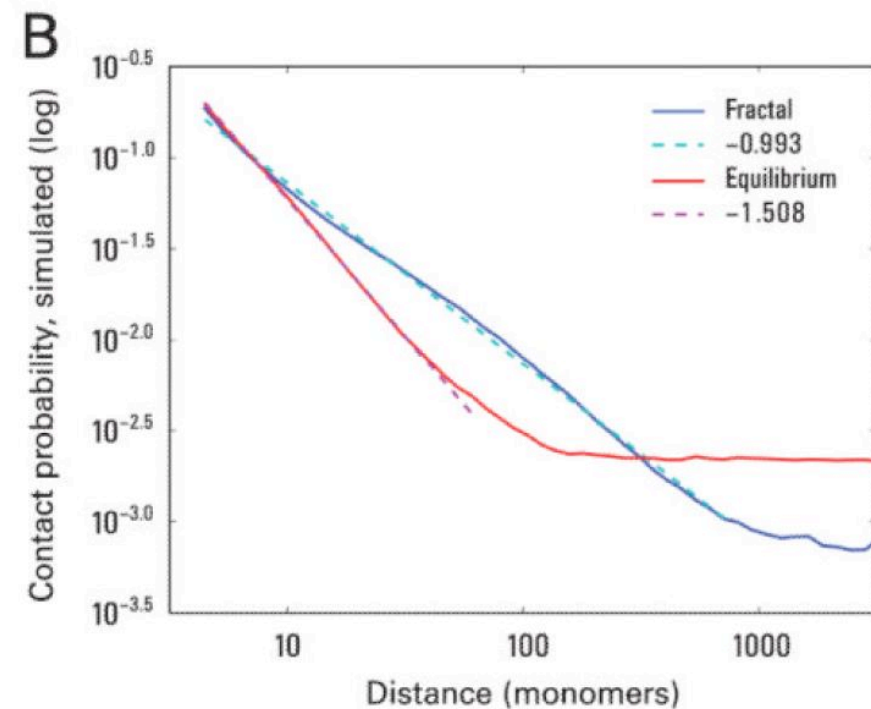
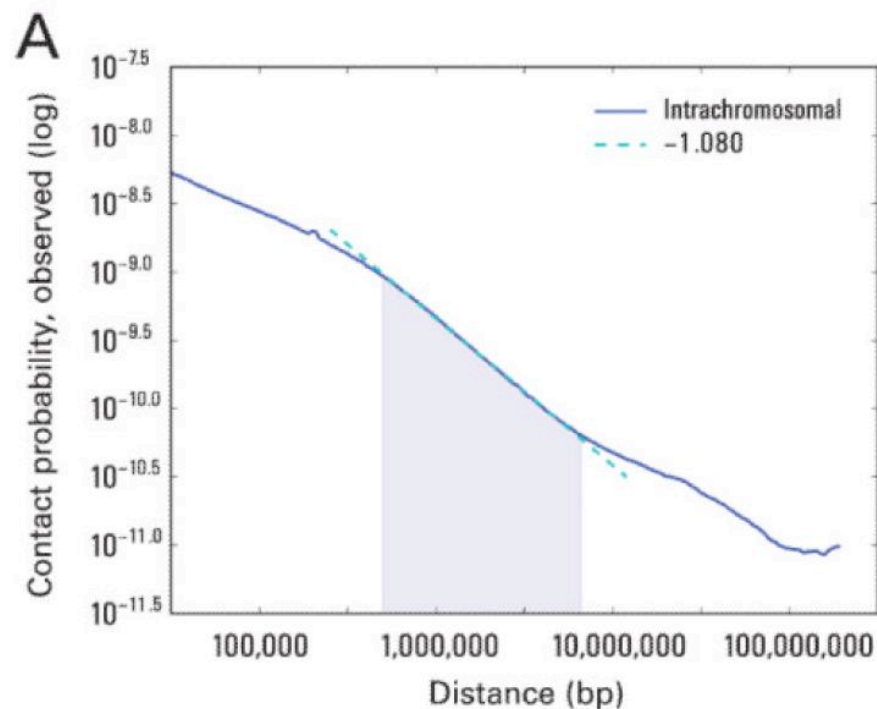
Le proprietà degli insiemi corrispondevano alle probabilità di scala di contatto derivate teoricamente (frat. s^{-1} ; equil. $s^{-3/2}$) e per la distanza 3D (frat. $s^{1/3}$; equil. $s^{1/2}$). Queste simulazioni illustrano anche l'assenza di annodamento [misurate con il polinomio teoretico del nodo di Alexander] e la formazione di settori spaziali nel globulo frattale (fig 4B)

The local packing of chromatin is consistent with the behavior of a fractal globule.

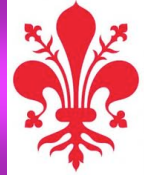
Fig 4 A-B fino a 200 “reads” letture di sequenza



(A) Contact probability as a function of genomic distance, averaged across the genome (blue) shows a power law scaling between 500kb and 7Mb (shaded region) with a slope of -1.08 (fit shown in cyan). (B) Simulation results for contact probability as a function of distance (1 monomer~6 nucleosomes~1200 bp, SOM) for equilibrium (red) and fractal (blue) globules.



Leggenda fig 4 C-D



The slope for a fractal globule is very nearly -1 (cyan), confirming our prediction (SOM).

The slope for an equilibrium globule is $-3/2$, matching prior theoretical expectations.

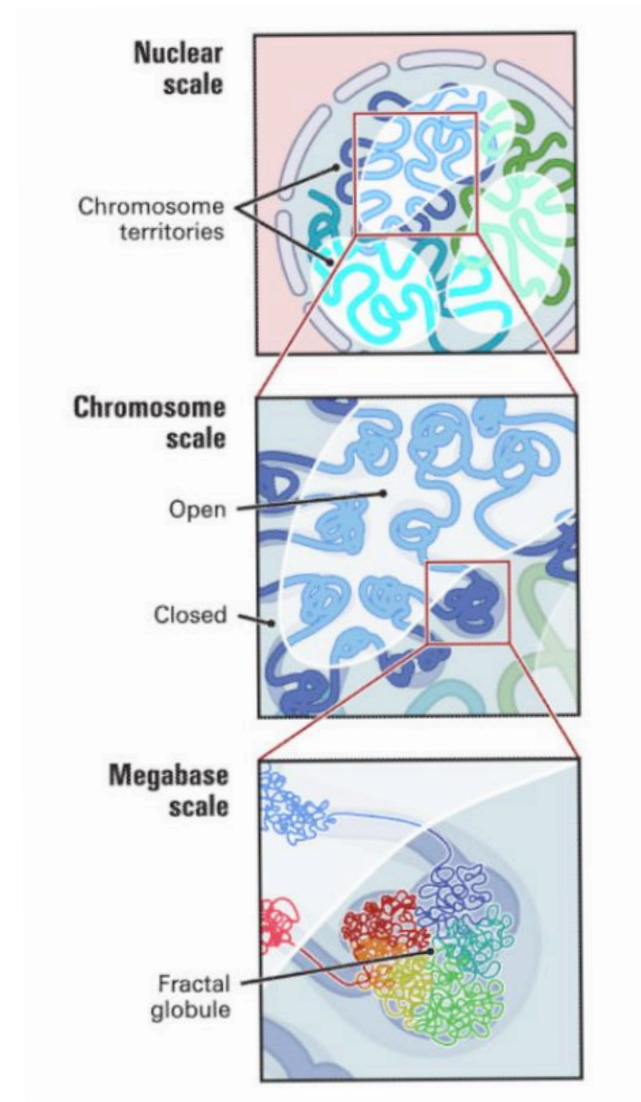
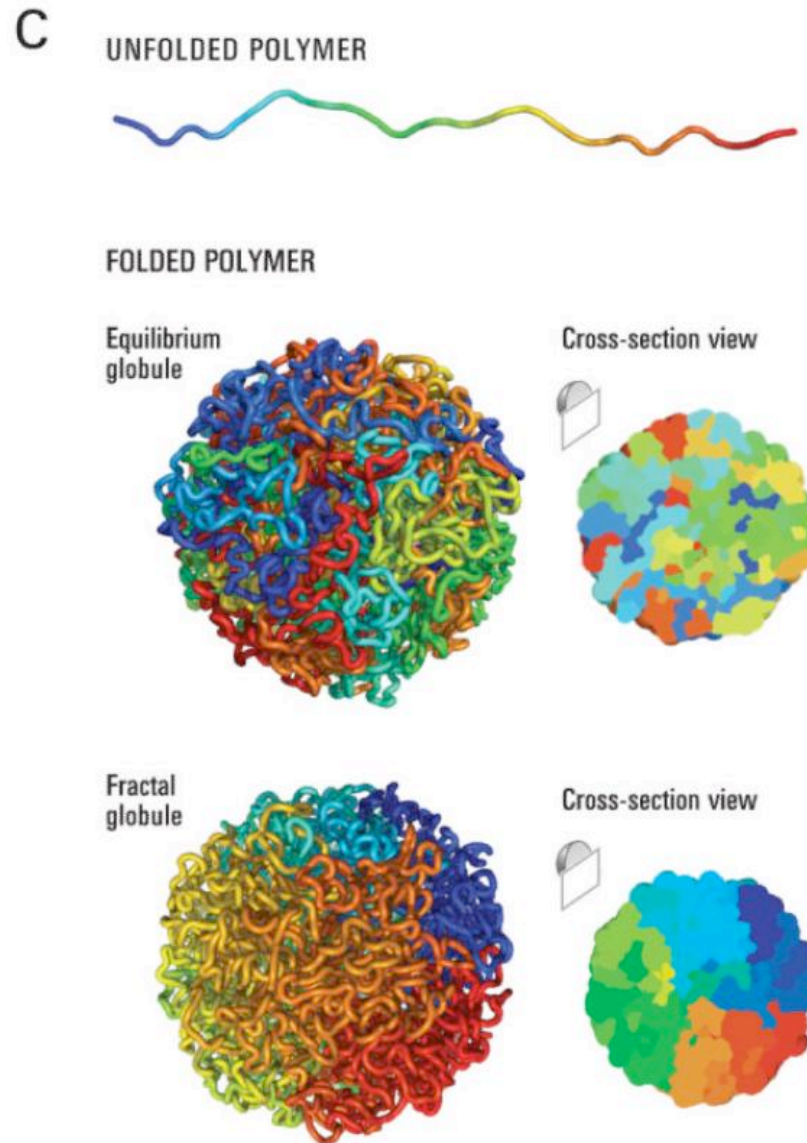
The slope for the fractal globule closely resembles the slope we observed in the genome.

(C) Top: An unfolded polymer chain, 4000 monomers (4.8 Mb) long. Coloration corresponds to distance from one endpoint, ranging from blue to cyan, green, yellow, orange, and red.

Middle: An equilibrium globule. The structure is highly entangled; loci that are nearby along the contour (similar color) need not be nearby in 3D. Bottom: A fractal globule.

Nearby loci along the contour tend to be nearby in 3D, leading to monochromatic blocks both on the surface and in cross-section. The structure lacks knots. (D) Genome architecture at three scales. Top: Two compartments, corresponding to open and closed chromatin, spatially partition the genome. Chromosomes (blue, cyan, green) occupy distinct territories. Middle: Individual chromosomes weave back-and-forth between the open and closed chromatin compartments. Bottom: At the scale of single megabases, the chromosome consists of a series of fractal globules.

figura 4 C modello



conclusioni



Il modello funziona ed il sistema di analisi è valido
Il metodo di simulazione MonteCarlo è adatto allo scopo.
Il metodo si basa sulla reiterazione di prove successive
con un algoritmo adatto che sostituisce in migliaia di
combinazioni una variabile alla volta, ma fatevelo spiegare
da chi lo usa normalmente in statistica applicata
Concettualmente è semplice, applicarlo non è difficile ma
va saputo usare.