Introduzione alla Genomica

Definizioni

- Genetica: è la scienza che studia i geni, l'ereditarietà e la variabilità degli organismi
- Genomica: è la disciplina che studia profili genetici su ampia scala per il genoma di una data specie

Quindi:

- Mentre la genetica guarda ai singoli geni, uno alla volta
- la genomica cerca di fotografare l'immagine complessiva, esaminando l'insieme dei geni (e delle caratteristiche genomiche) come un sistema completo

Pietre miliari nello studio del genoma

- 1869 – Viene isolata la “nucleina”, l’allora sconosciuto DNA
- 1953 – Watson e Crick scoprono la struttura del DNA
- 1961 – Viene compreso il codice genetico per la sintesi delle proteine
- 1977 – sviluppo del metodo di sequenziamento Sanger
- 1990 – inizio del progetto genoma umano
- 1995 – Sequenziato il primo genoma batterico (Haemophilus influenzae)
- 2000 – sequenza del genoma di *Drosophila*
- 2001 – rilascio del primo draft del genoma umano
- 2003 – completamento del progetto genoma umano

Origini della genomica

- Il termine genomica è nato negli anni 80 per indicare una disciplina dedita a:
  - *Mapping,*
  - *Sequencing,*
  - *Caracterizing genomes*

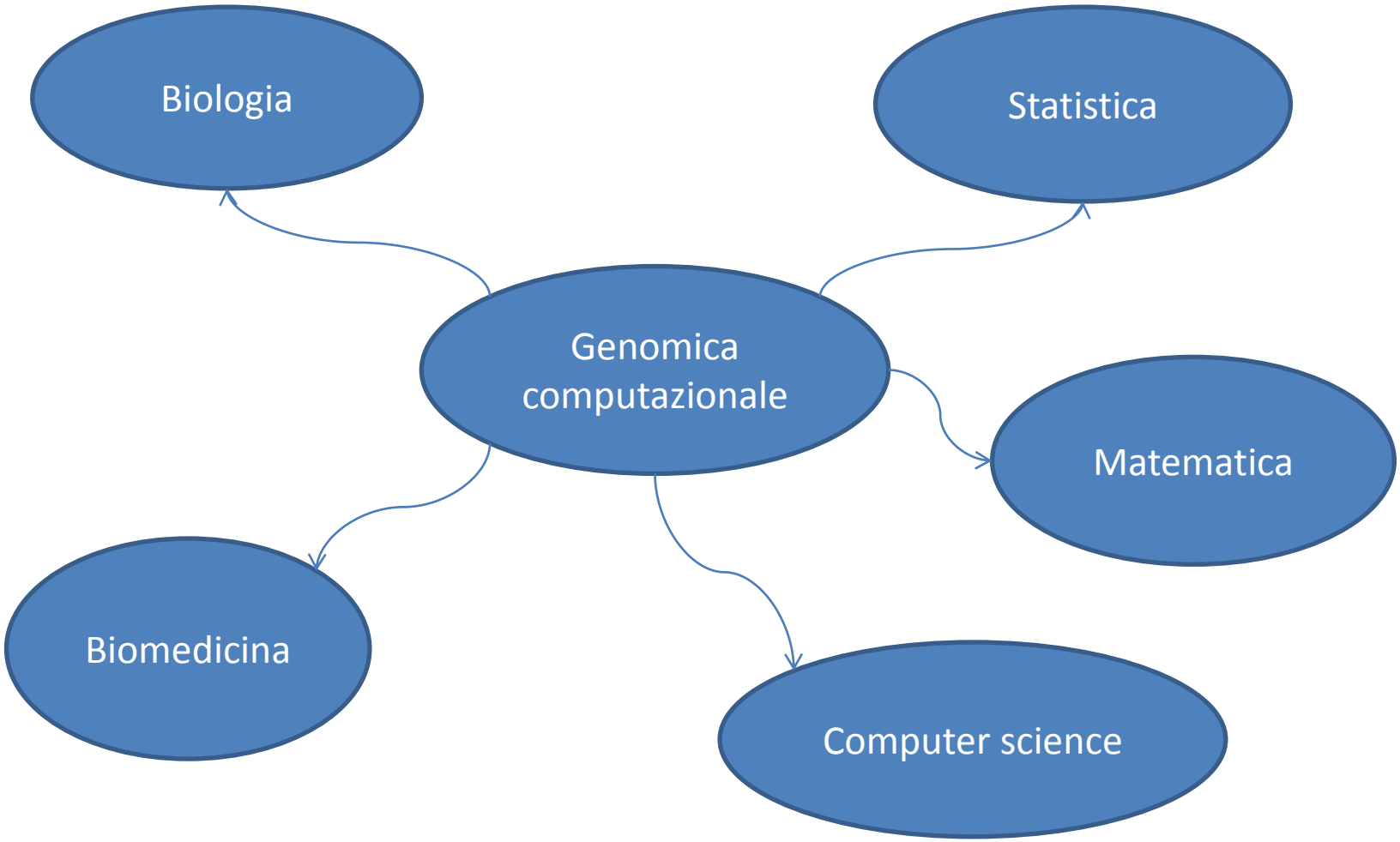
Sviluppo della genomica

- Le origini e lo sviluppo della genomica coincidono inoltre con lo sviluppo di:
  - Progetti genoma per varie specie biologiche tra cui l'uomo
  - Lo sviluppo di tecnologie per l'analisi parallela di un elevato numero di geni, sequenze ecc…:
    - Sequenziatori capillari
    - Microarray
    - Next Generation Sequencing (NGS)
    - …

Genomica e bioinformatica

- L’introduzione dei computer e metodi computazionali nei laboratori di biologia molecolare è uno dei fattori chiave nello sviluppo della genomica.
- Lo sviluppo dell’automazione in laboratorio e di tecnologie di analisi in grado di produrre grandi quantità di dati da un singolo esperimento ha dato spinta allo sviluppo della Bioinformatica.

Genomica Computazionale

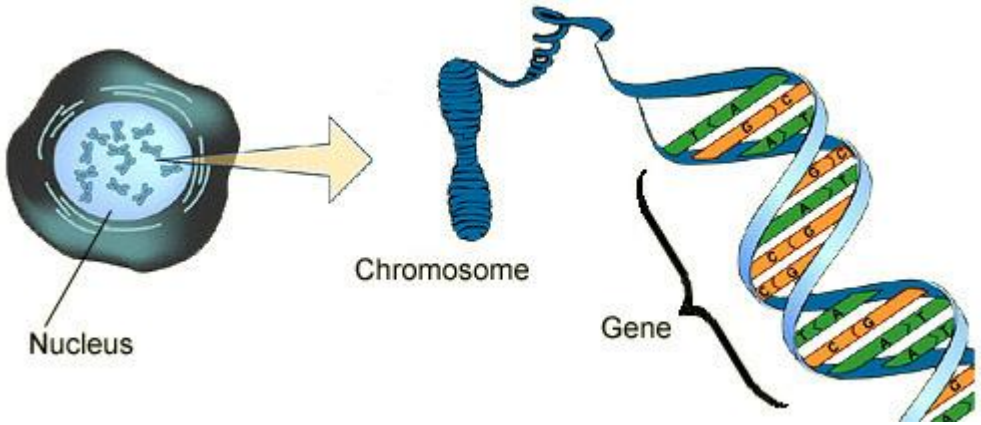


Campi della Genomica

- Genomica strutturale
  - Ricostruzione (assemblaggio) dei dati di sequenza genomica
  - Identificazione dei geni
  - Costruzione di mappe genetiche
- Genomica funzionale
  - Funzione biologica dei geni
  - Regolazione
  - Analisi della variabilità
  - …
- Genomica comparativa
  - Comparazione delle sequenze genomiche per determinare relazioni funzionali o evoluzionistiche tra genomi di organismi diversi

Cos'è un genoma

- Il termine genoma si riferisce al DNA contenuto una singola cellula (cellula aploide nel caso di un organismo diploide) di un organismo (inclusi i suoi geni). ➔ il DNA codifica l'informazione ereditaria di un organismo



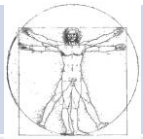
http://www.bbc.co.uk/staticarchive/678f62dce35d0fc7ef2333d6d3bfbf53744374ff.jpg

Cos’è un genoma

- Il DNA degli organismi superiori è organizzato in cromosomi (23 paia nell’uomo)
- Non tutto il DNA codifica per proteine
- Alcuni geni esistono in copie multiple (duplicazione dà origine a famiglie geniche)

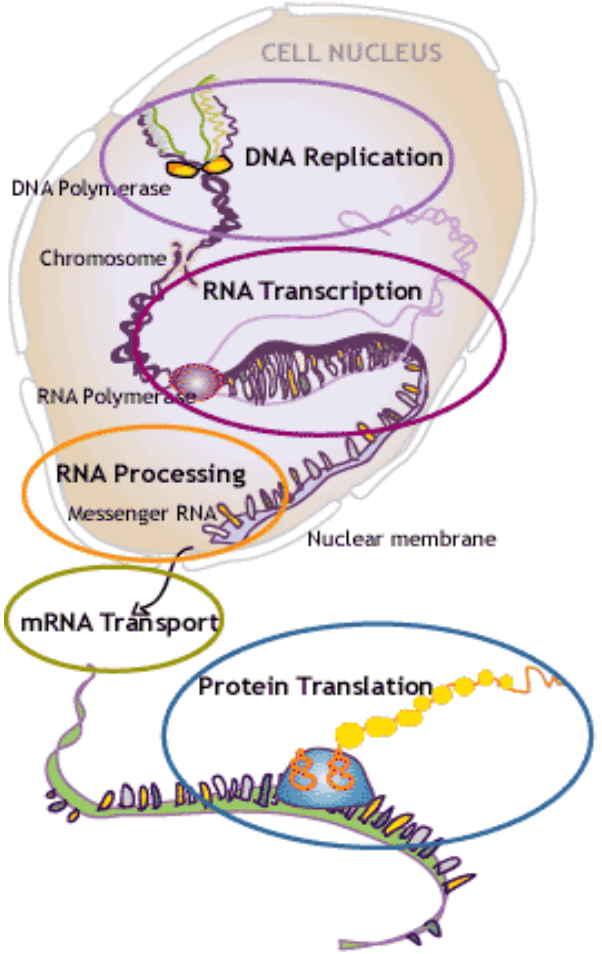
➔ Dalle dimensioni del genoma non è quindi possibile stimare facilmente la quantità di informazione traducibile in sequenza proteica

Dimensione dei genomi

	Specie	# cromosomi	# geni	Lunghezza (bp)
	Uomo	23 paia	22-25,000	3.1 miliardi
	Topo	20 paia	22-30,000	2.7 miliardi
	Pesce palla	22 paia	~ 30,000	365 milioni
	Anopheles gambie	3 paia	14,000	289 milioni
	Drosophila melanogaster	4 paia	14,000	137 milioni
	C. elegans	6 paia	19,000	97 milioni
	Escherichia coli	1	5,000	4.1 milioni
	Arabidopsis thaliana	5 paia	~28,000	125 milioni
	Vitis vinifera	19 paia	~30,000	486 milioni

Geni e genomi

- Geni codificati dal genoma contengono l'informazione per la produzione delle proteine richieste per il funzionamento della cellula

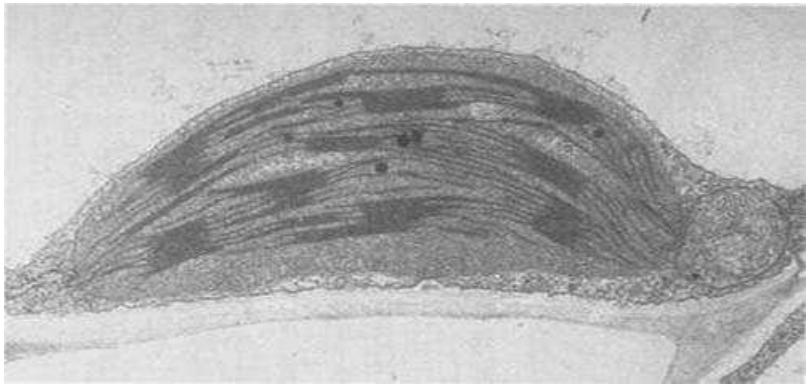


Geni → Proteine → Caratteri

Caratteri: l'aspetto, la fisiologia, il comportamento, la capacità di combattere le infezioni, la predisposizione a malattie, ecc… di un organismo

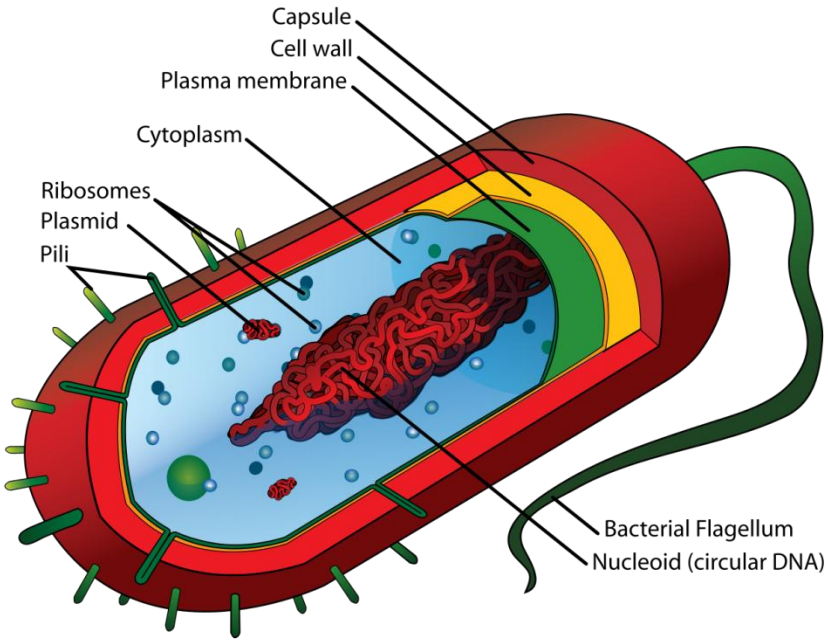
Geni extracromosomali

- Mitocondri e cloroplasti contengono molecole di DNA che portano un certo numero limitato di geni (tRNA, rRNA, enzimi facenti parte dei complessi respiratori mitocondriali)



Il genoma procariotico

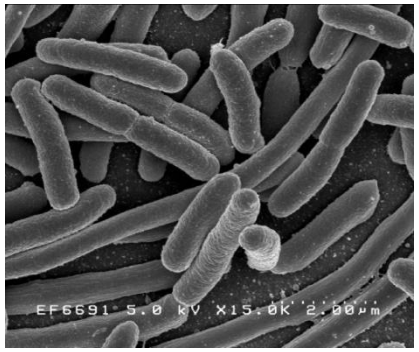
- Il genoma di un procariote è composto di una singola molecola di DNA a doppio filamento (generalmente < 5Mbp)
- Cellule procariotiche possono avere anche dei plasmidi
- Geni codificanti proteine non hanno introni
- La quantità di DNA non codificante è molto ridotta rispetto agli eucarioti



Plasmidi

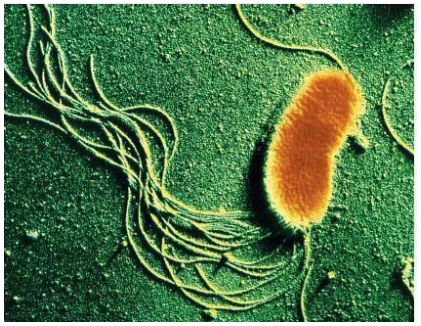
- I plasmidi sono molecole circolari di DNA a doppio filamento che sono separate dal DNA cromosomale
- Le dimensioni variano da 1 a 250 Kbp.
- In una singola cellula sono presenti da una singola copia a centinaia di copie dello stesso plasmide

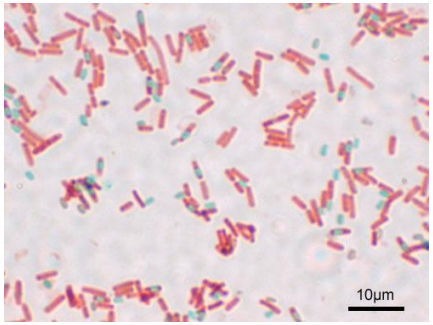
Organismi modello procariotici



Escherichia coli

Pseudomonas fluorescens





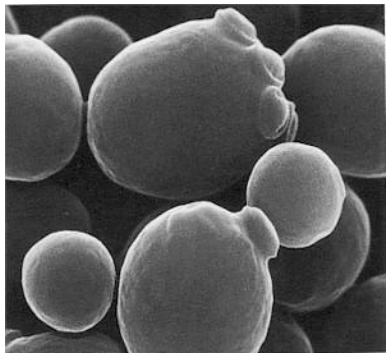
Bacillus subtilis

Mycoplasma genitalium

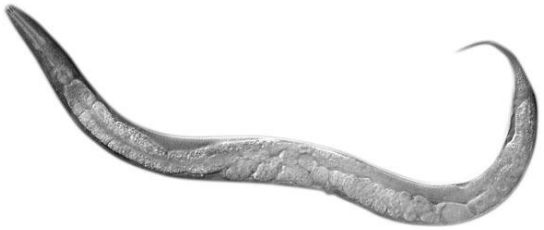
Genoma degli eucarioti

- La maggior parte del DNA è compartimentato nel nucleo e organizzato in cromosomi
- Piccole quantità di DNA sono presenti negli organelli come mitocondri e cloroplasti
- Geni all'interno di singoli cromosomi appartengono spesso a famiglie e possono essere paraloghi derivati da duplicazione che sono poi andati incontro a divergenza di sequenza (e funzione)
- All'interno del genoma possono essere presenti pseudogeni (geni inattivi)
- La maggior parte del genoma generalmente non codifica per proteine

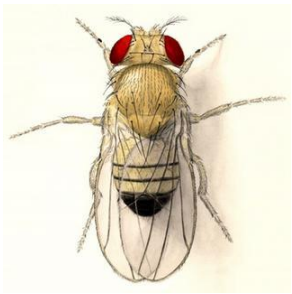
Organismi modello eucariotici



Saccharomyces cerevisiae



Caenorhabditis elegans



Drosophila melanogaster

Arabidopsis thaliana

Il progetto genoma umano (HGP)

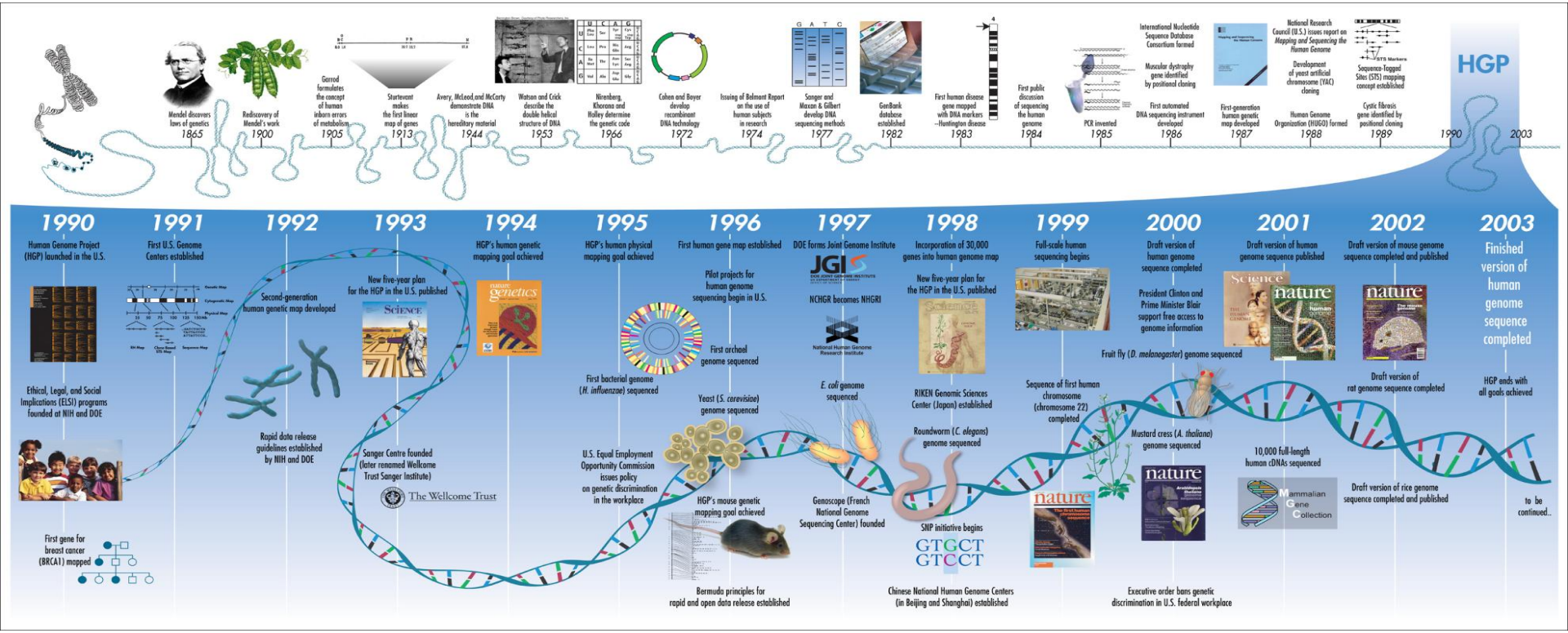
- Progetto internazionale
- Coinvolti 20 istituzioni da 6 differenti paesi (China, France, Germany, Japan, UK and USA)

Finalità del progetto genoma humano

- Generare la sequenza del genoma umano
- Identificare i geni in esso contenuti
- Memorizzare queste informazioni in un database
- Migliorare i tool per l'analisi dei dati
- Trasferire le tecnologie al settore privato
- Affrontare le problematiche etiche, legali e sociali che possono derivare dal progetto

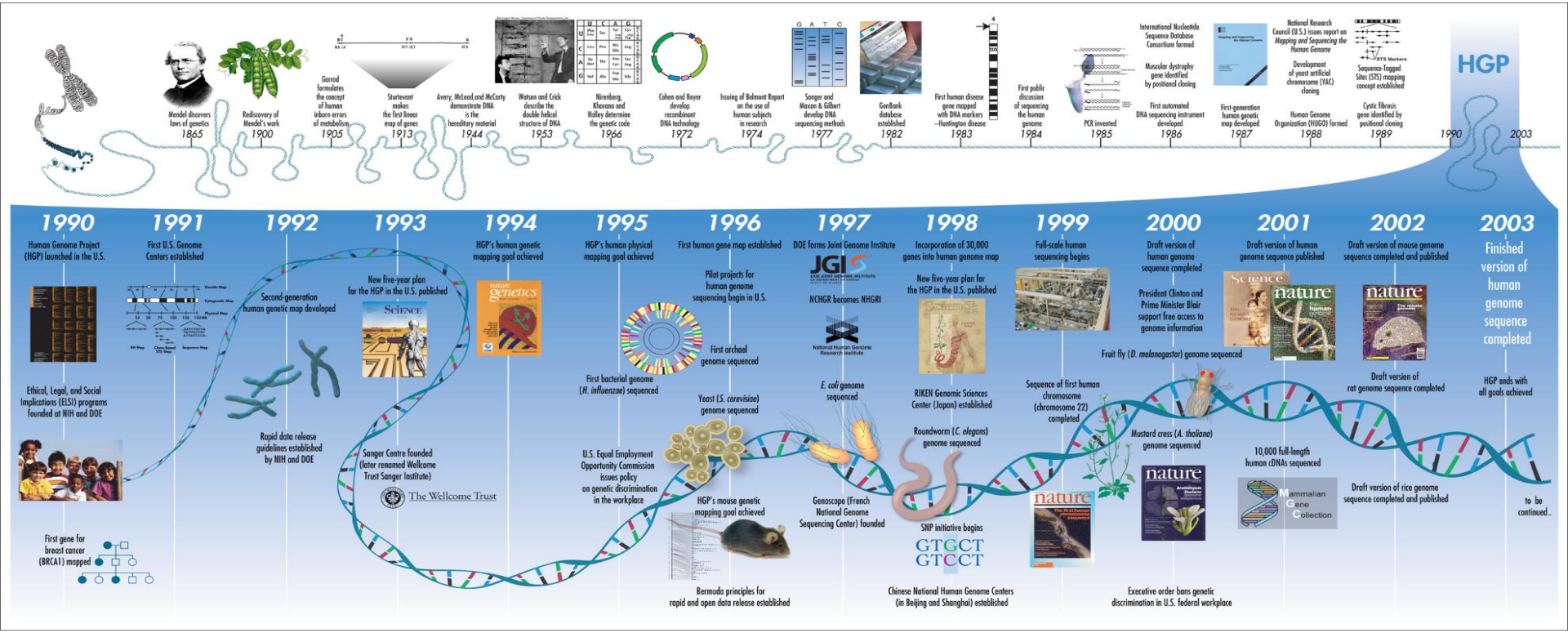
Il progetto genoma umano (HGP)

L'HGP è nato nel 1986, iniziato nel 1990 e ha richiesto 13 anni per essere completato.



Il progetto genoma umano (HGP)

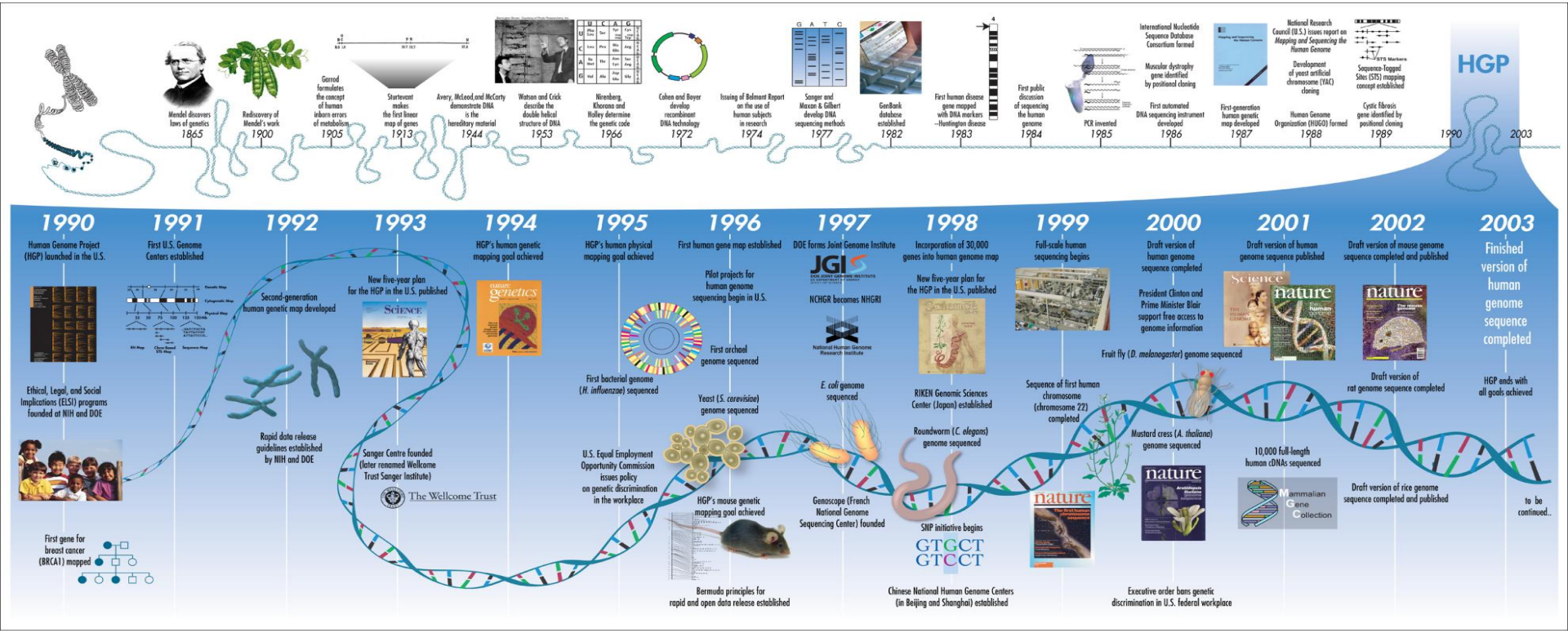
L'HGP è nato nel 1986, iniziato nel 1990 e ha richiesto 13 anni per essere completato.



Inizio del progetto

Il progetto genoma umano (HGP)

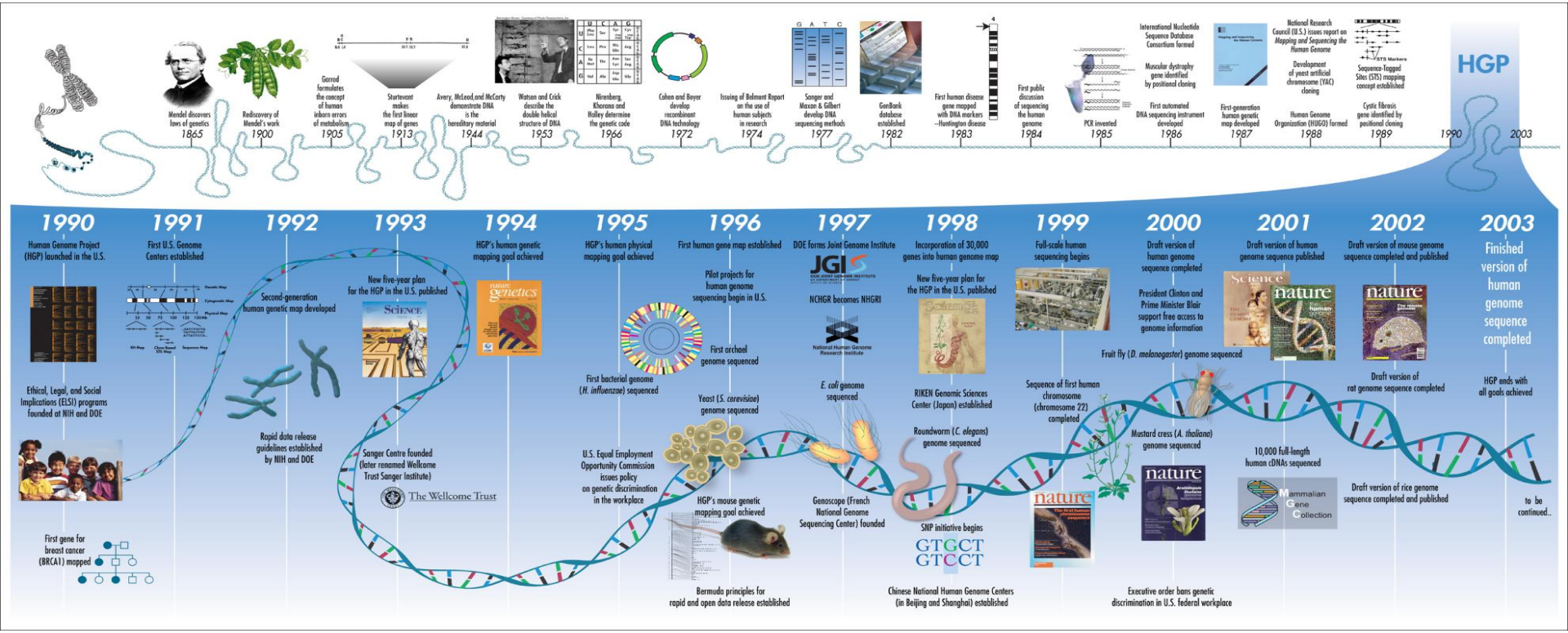
L'HGP è nato nel 1986, iniziato nel 1990 e ha richiesto 13 anni per essere completato.



- Completamento della mappa genetica
- Genetic Privacy Act: viene proposto per regolamentare la raccolta, l'analisi, lo stoccaggio e l'uso del DNA e dell'informazione genetica.

Il progetto genoma umano (HGP)

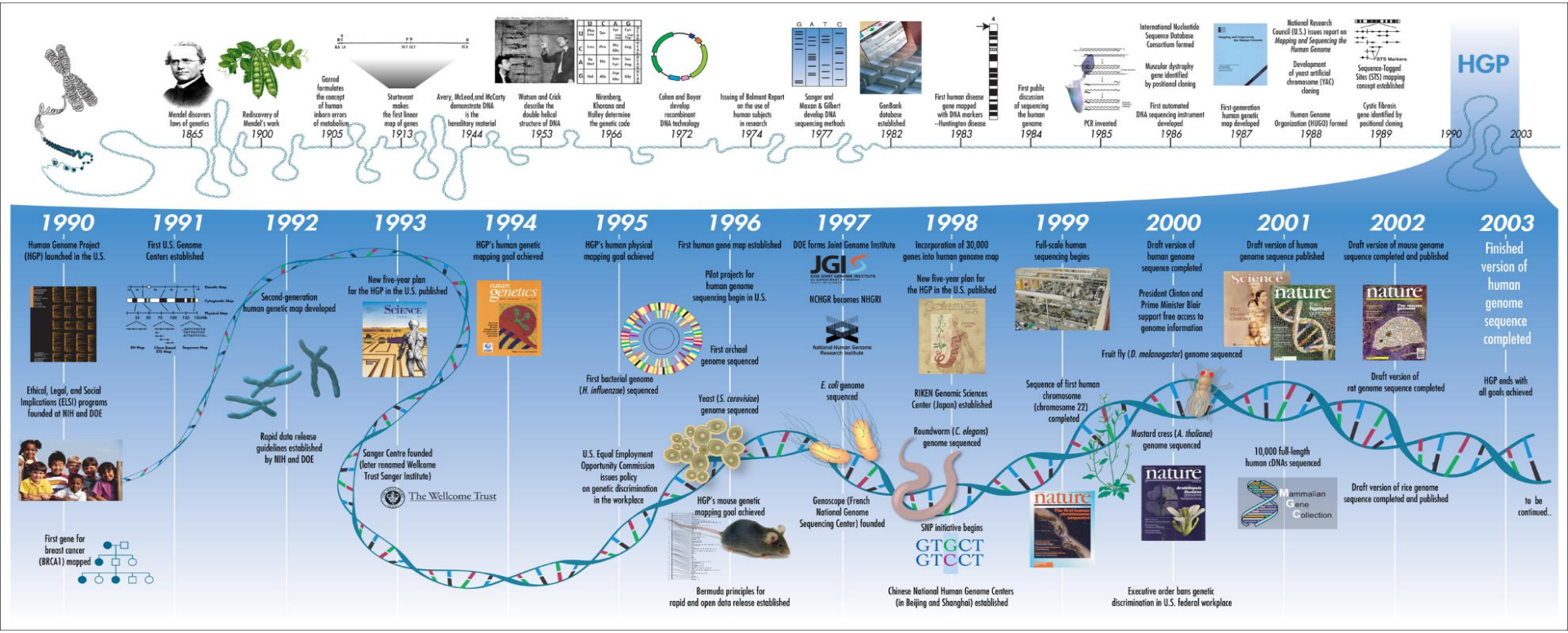
L'HGP è nato nel 1986, iniziato nel 1990 e ha richiesto 13 anni per essere completato.



- Nel 1998 parte un progetto parallelo da parte di Celera Genomics

Il progetto genoma umano (HGP)

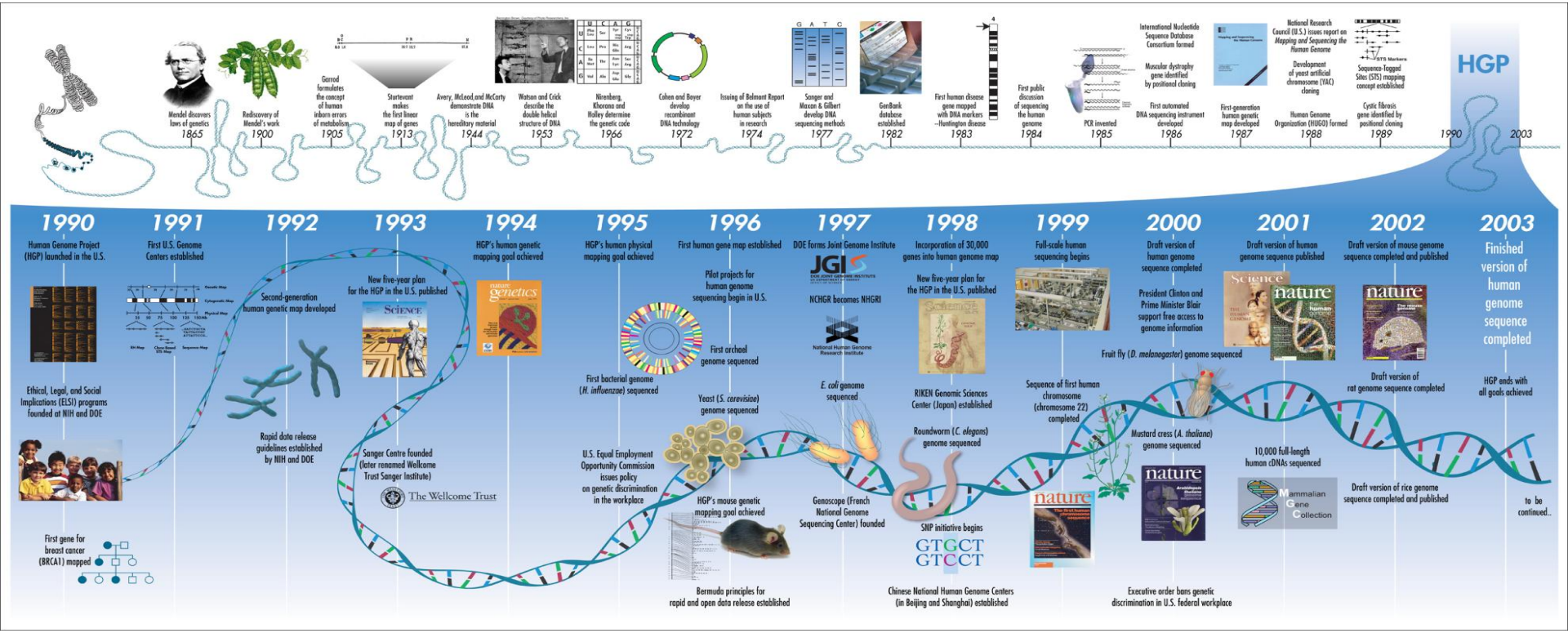
L'HGP è nato nel 1986, iniziato nel 1990 e ha richiesto 13 anni per essere completato.



•Completato il sequenziamento del primo cromosoma (22)

Il progetto genoma umano (HGP)

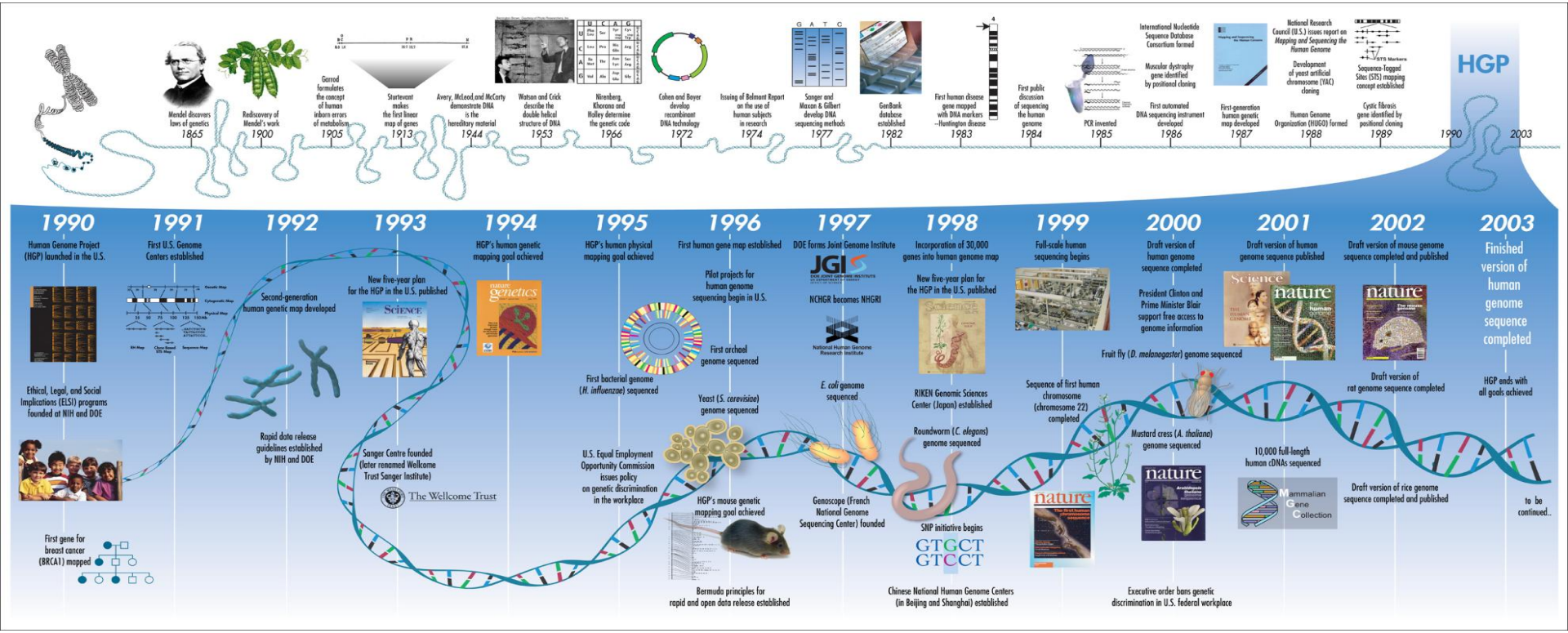
L'HGP è nato nel 1986, iniziato nel 1990 e ha richiesto 13 anni per essere completato.



- Completato il primo working draft

Il progetto genoma umano (HGP)

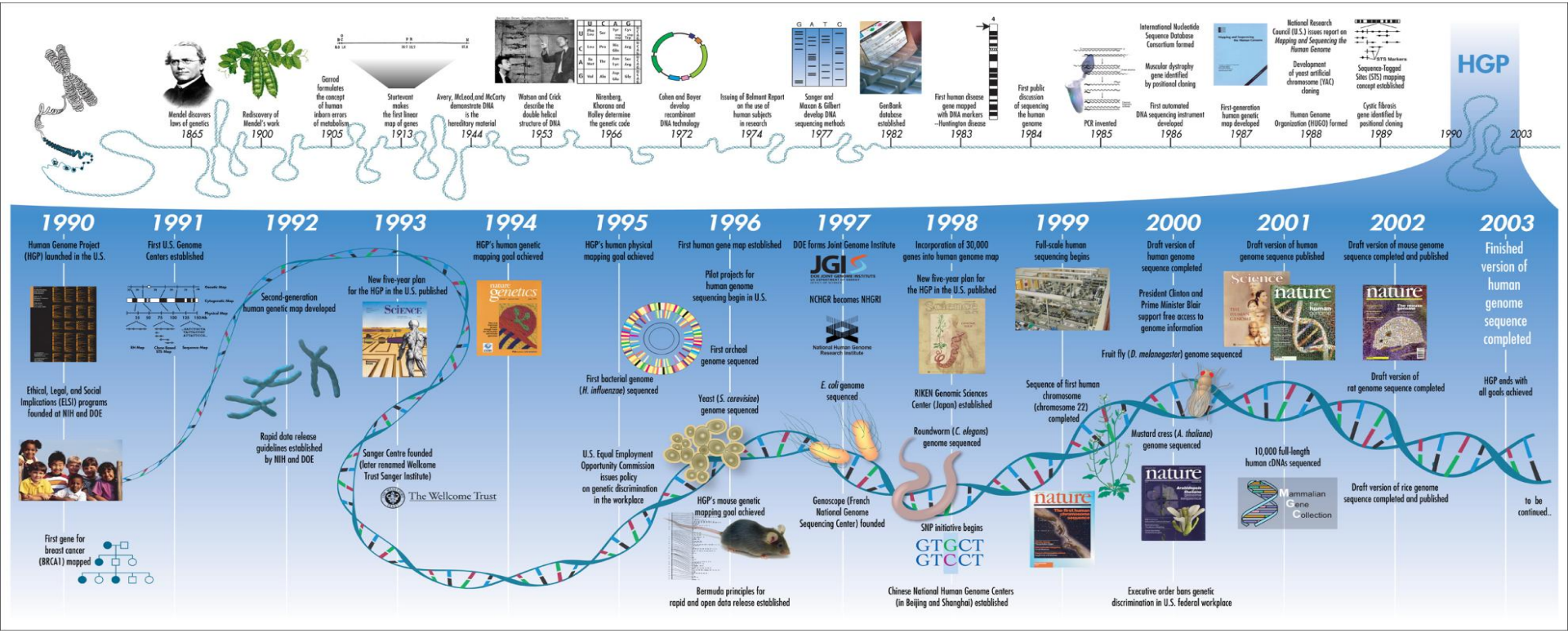
L'HGP è nato nel 1986, iniziato nel 1990 e ha richiesto 13 anni per essere completato.



- Pubblicato il draft dell'HGP
- Nel contempo anche Celera pubblica il "proprio" genoma

Il progetto genoma umano (HGP)

L'HGP è nato nel 1986, iniziato nel 1990 e ha richiesto 13 anni per essere completato.



Il progetto viene completato

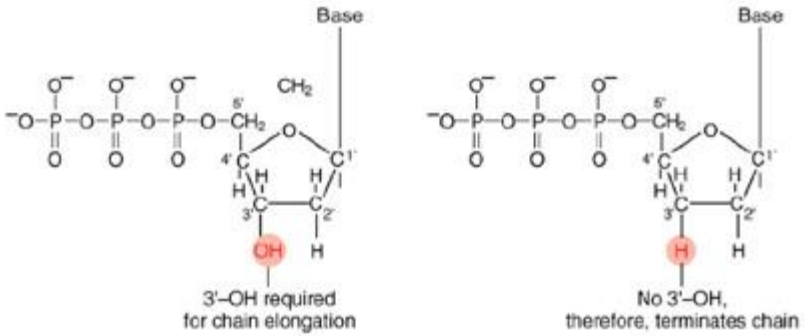
Il progetto genoma umano (HGP)

- Il progetto è stato dichiarato completato nel 2003
- Il costo complessivo del progetto è stato di 3 miliardi di dollari
- In realtà la sequenza genomica viene continuamente rifinita sia a livello di chiusura dei “gap” che della struttura (siamo al 20° rilascio denominato GRChg38)

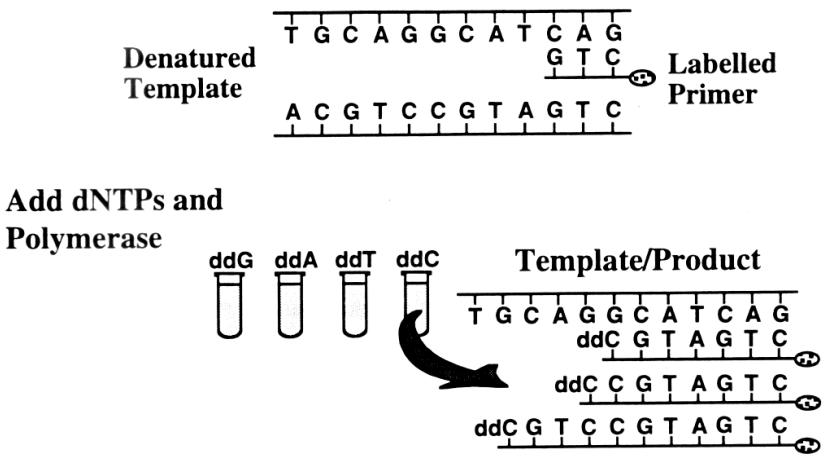
La sequenza del genoma rappresenta un genoma composito

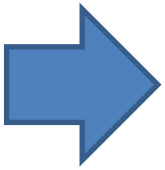
- Differenti sorgenti sono state utilizzate per il sequenziamento originale:
  - Campioni di sangue (femmine) o sperma (maschi) da un ampio numero di donatori
  - Etnicamente diversificati
  - Selezionato casualmente un numero limitato di campioni anonimi dal pool dei donatori

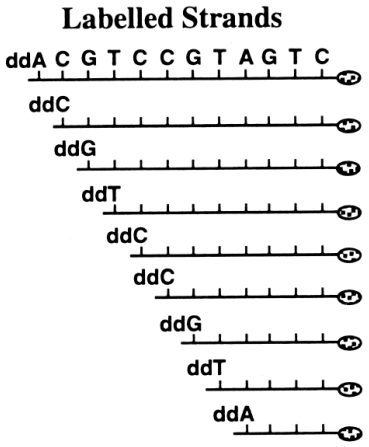
Sequenziamento Sanger (1977)



Si basa su didesossi dei 4 nucleotidi: quando un didesossi viene inserito la catena non può essere estesa





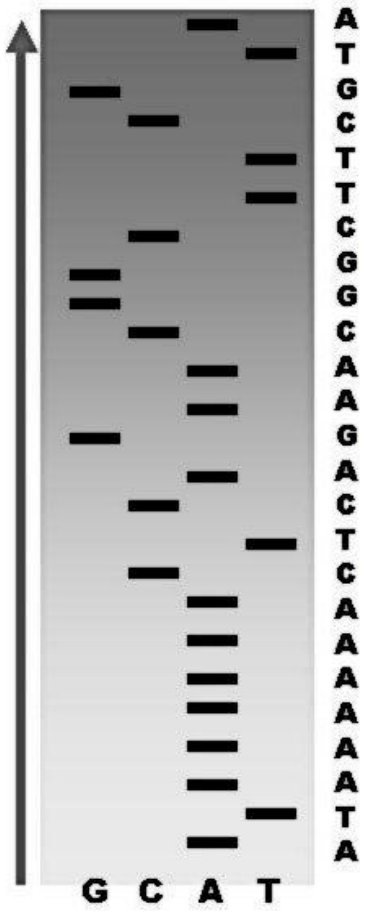


Una copia del templato viene polimerizzata in presenza di dNTP + uno dei 4 didesossi

Si generano frammenti di diversa lunghezza a seconda della posizione in cui viene inserito il didesossi.

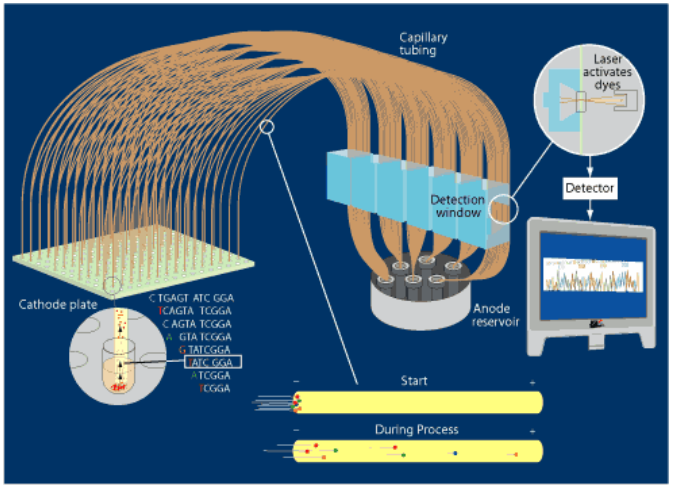
Sequenziamento Sanger

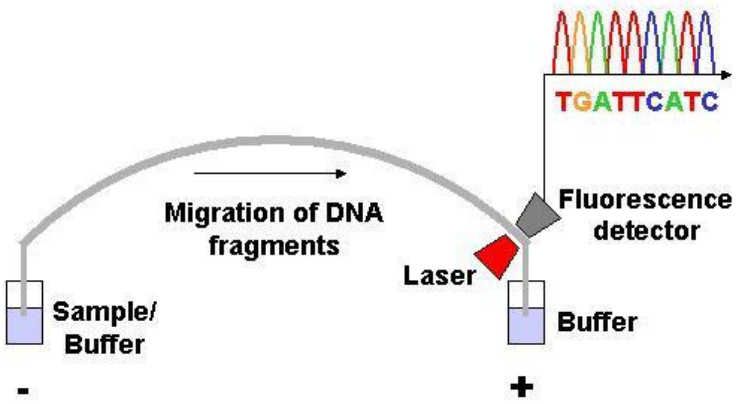
- È possibile separare tramite il prodotto di ciascuna delle 4 reazioni su una diversa lane di un gel di poliacrilamide sulla base delle dimensioni dei frammenti.



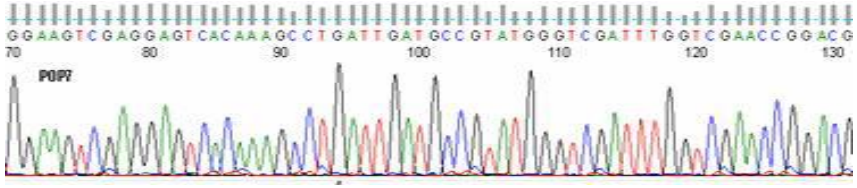
Vengono generate sequenze da circa 800nt

Sequenziamento Sanger su capillare





Utilizza 4 differenti terminatori fluorescenti per le 4 basi. I frammenti vengono separati da elettroforesi capillare. Le basi vengono rilevate automaticamente da un laser/detector e registrate come un elettroferogramma.



Sequenziatori moderni consentono di analizzare 96 campioni/sequenze in parallelo.

Produce sequenze di 800bp in lunghezza

http://www.scg.ubc.ca/wp-content/uploads/2006/08/sequencing2.gif

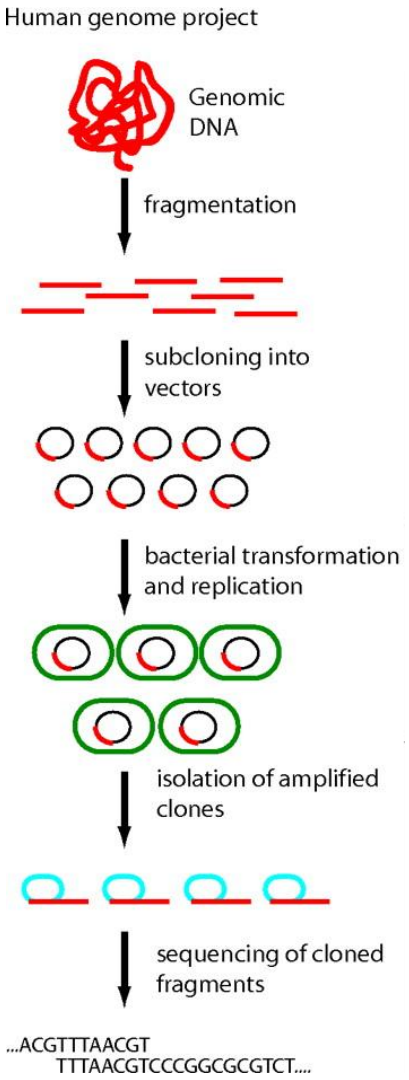
Sequenziamento del genoma umano

- 96 campioni
- 800 nt/campione
- 1.600.000 nt/giorno

- Il genoma umano è 3 miliardi di basi
- La copertura minima per avere una analisi accurata è 8X = 24 miliardi di basi
- 24E+9 nt / 1.6E+6 nt/giorno = 15.000 giorni

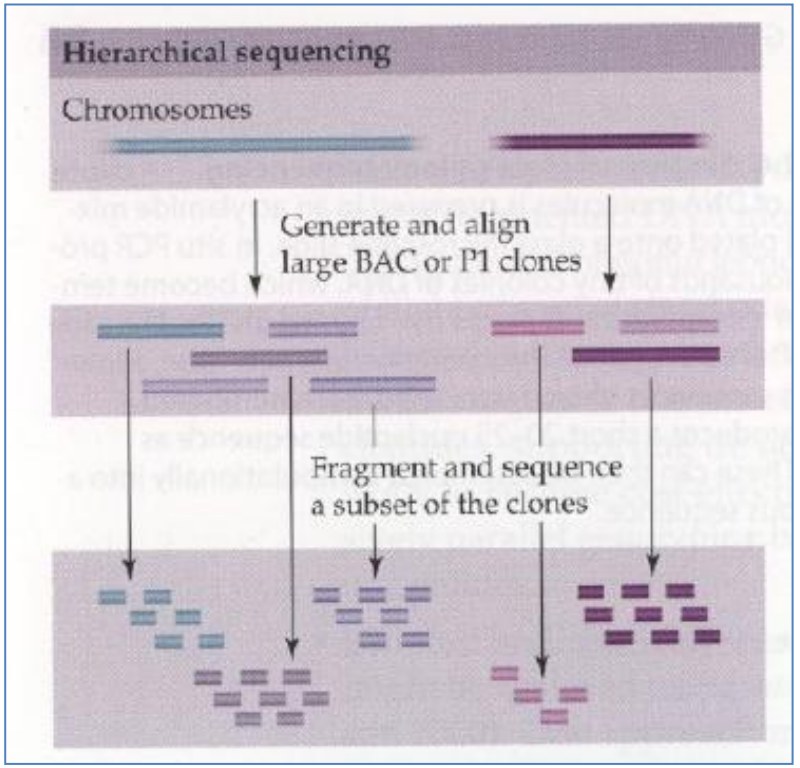
Come è stato sequenziato il genoma umano?

- Tecnologia di sequenziamento consente di ottenere la sequenza di circa 800bp alla volta
- Il DNA genomico deve essere frammentato in piccoli pezzi per il sequenziamento per poi essere ricomposto come un gigantesco puzzle
- Frammenti di 150-350 kb vengono inseriti in cromosomi batterici artificiali (BAC) che vengono trasformati in cellule batteriche e replicati
- I cloni vengono frammentati in subcloni di dimensioni più piccole (4000-6000 bp) e reinseriti in batteri per venire amplificati
- DNA viene estratto dalle colonie e
- Sequenziato tramite metodo **Sanger**



Sequenziamento gerarchico

- Il genoma viene frammentato in frammenti grandi inseriti in cloni BAC → possono venire facilmente ancorati alla rispettiva posizione sulla mappa genetica
- Da ogni clone vengono generati dei subcloni più piccoli di dimensione adatte al sequenziamento
- Dai frammenti sequenziati ricostruisco la sequenza di ciascun clone BAC che ancorato alla mappa genetica mi permette di ricostruire la sequenza degli interi genomi



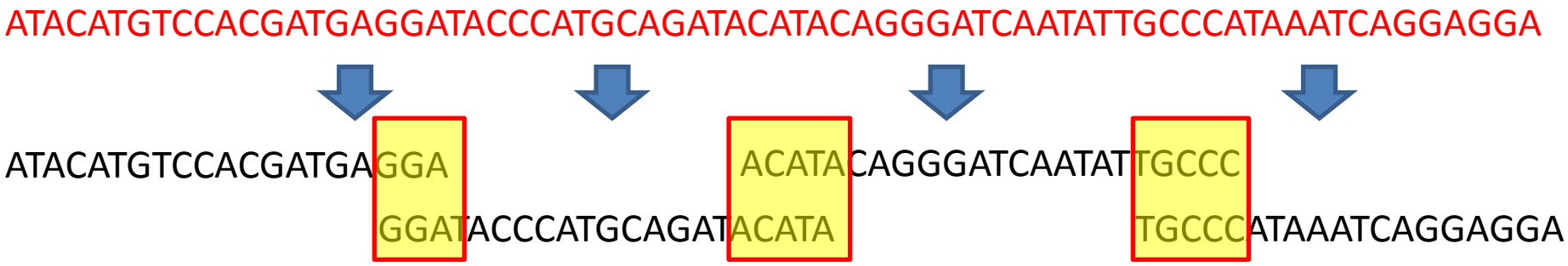
Il problema computazionale viene ridotto al problema della ricostruzione della sequenza di ciascun clone BAC (non dell'intero genoma)

Il principio è quello di ottenere una serie di frammenti sovrapponibili di DNA che possono venire connesse in una mappa continua.

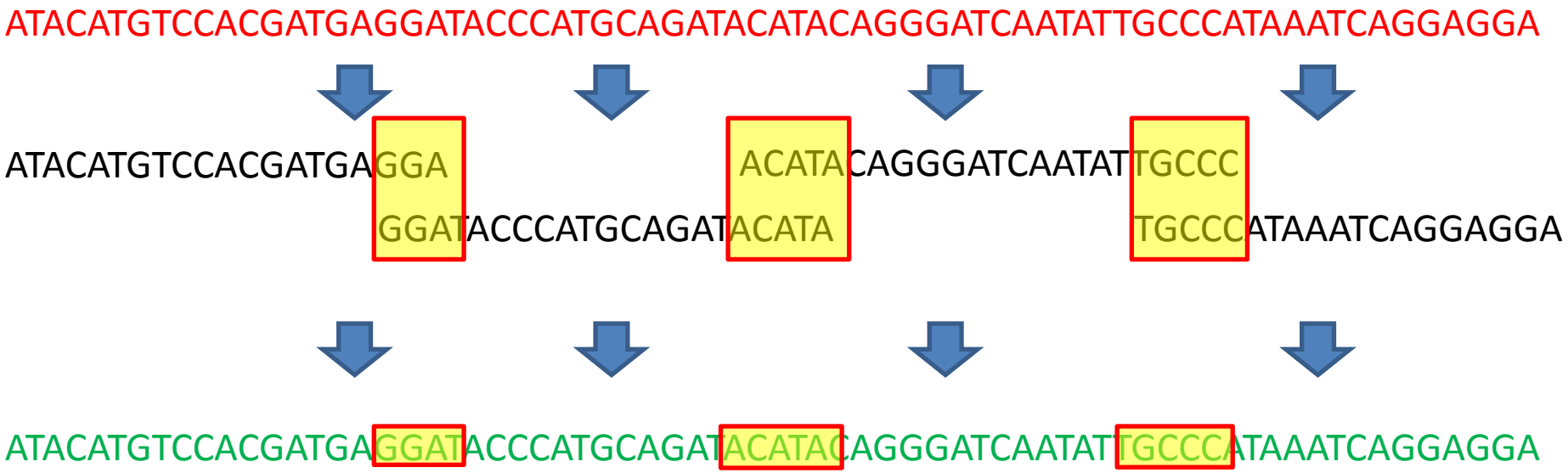
ATACATGTCCACGATGAGGATACCCATGCAGATACATACAGGGATCAATATTGCCCATAAATCAGGAGGA

Il principio è quello di ottenere una serie di frammenti sovrapponibili di DNA che possono venire connesse in una mappa continua.

Il principio è quello di ottenere una serie di frammenti sovrapponibili di DNA che possono venire connesse in una mappa continua.



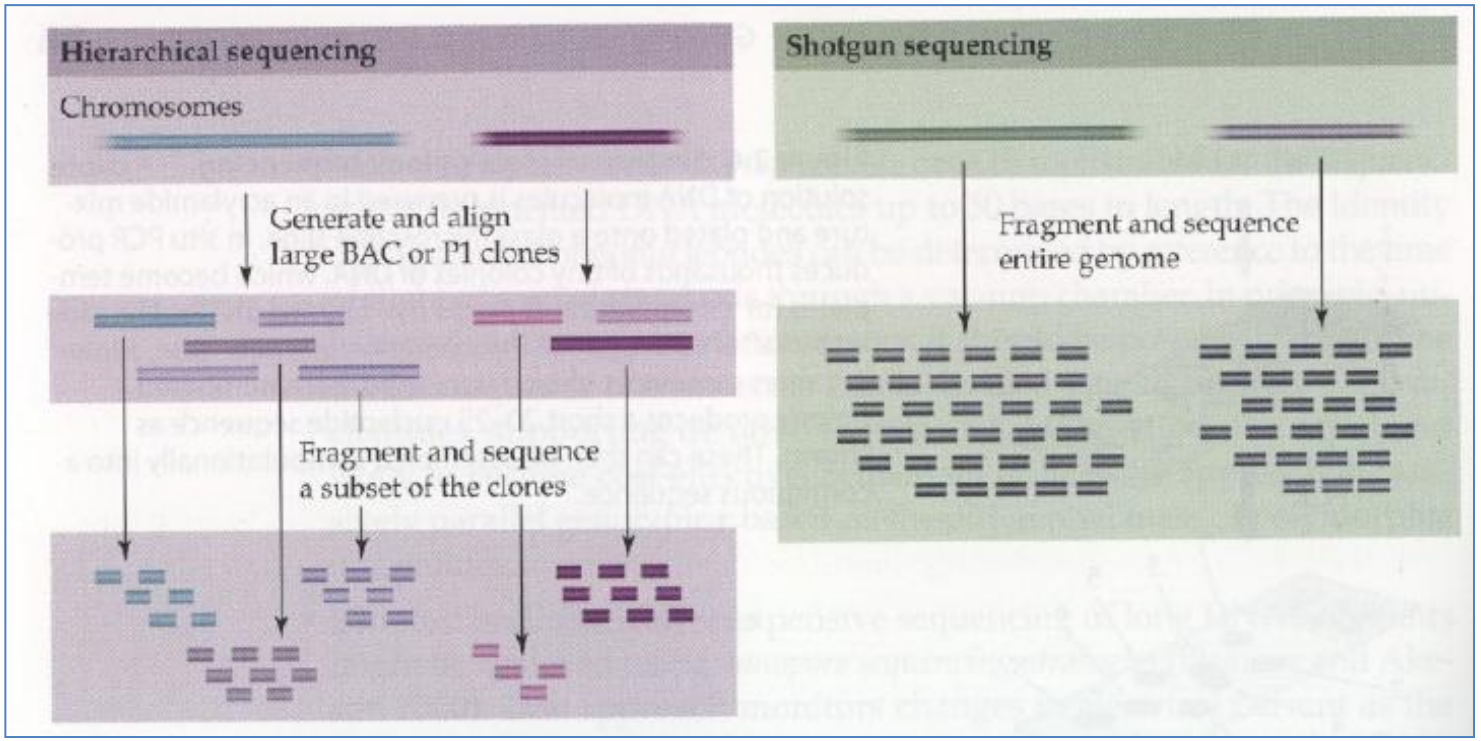
Il principio è quello di ottenere una serie di frammenti sovrapponibili di DNA che possono venire connesse in una mappa continua.



Celera (1998)

- Celera corporation (Craig Venter)
- Si propone di sequenziare il genoma umano per 300 milioni di dollari in 3 anni
- Basato su shotgun sequencing
- *Draft* del genoma annunciato nel giugno del 2000 battendo sul tempo di qualche mese l'HGP

Sequenziamento gerarchico vs. sequenziamento shotgun

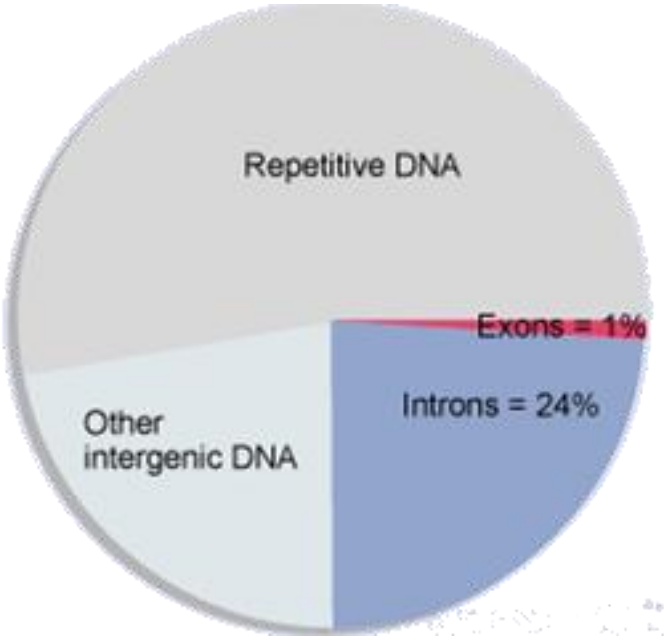


Approccio shotgun più semplice da un punto di vista della preparazione delle librerie genomiche (più veloce e meno costoso) ma è molto più esigente da un punto di vista computazionale!!!

E' stato possibile solo con lo sviluppo di metodi bioinformatici più avanzati e una potenza computazionale più elevata.

Il genoma umano

- 3.2 miliardi di basi
- Distribuito in 22 autosomi più i cromosomi sessuali
- Sequenze codificanti proteine costituiscono solo il 1-2% circa del genoma umano (circa 26.000 geni)
- Sequenze ripetute costituiscono più del 50% del genoma



Sequenze ripetute

Sequenze ripetute comprendono oltre il 50% del genoma:

- Elementi trasponibili, o ripetizioni intersperse incluse LINE e SINE
- Pseudogeni retrotrasposti
- Ripetizioni di corti oligomeri
- Duplicazioni segmentali (~10 - 300kb)
- Blocchi di ripetizioni in tandem (incluse famiglie geniche)

Element	Size (bp)	Copy number	Fraction of genome %
Short Interspersed Nuclear Elements (SINEs)	100-300	1.500.000	13
Long Interspersed Nuclear Elements (LINEs)	6000-8000	850.000	21
Long Terminal Repeats	15.000 -110.000	450.000	8
DNA Transposon fossils	80-3000	300.000	3

Caratteristiche dei geni umani

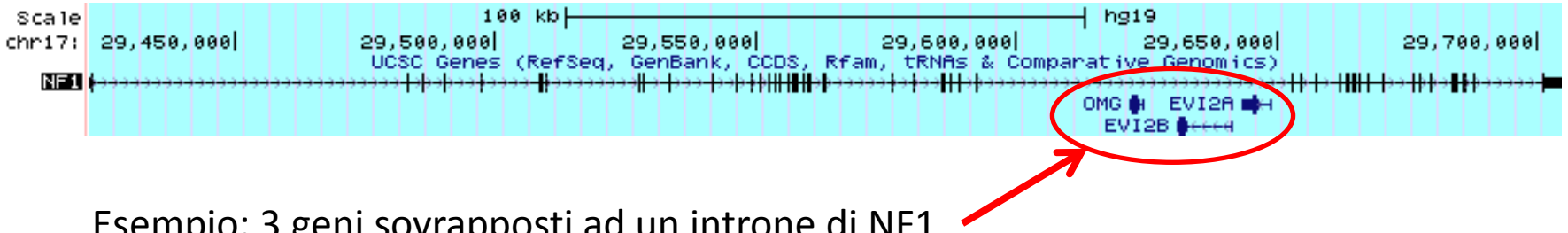
- Elevata variabilità nel numero e lunghezza degli introni.
- Geni codificanti proteine possono arrivare fino a >2.4Mbp di lunghezza (distrofina)
- Lunghezza media dei geni: ~ 8,000 bp
- Numero medio di esoni per gene: 9.8
- Numero massimo di esoni per gene: 363
- Lunghezza media degli esoni: ~300 bp
- Lunghezza media degli introni: ~2,000 bp
- Geni monoesonici: ~8%

Dimensione dei geni umani

Human protein	Size of protein (no. of amino acids)	Size of gene (kb)	No. of exons	Coding DNA (%)	Average size of exon (bp)	Average size of intron (bp)
SRY	204	0.9	1	94	850	–
b-Globin	146	1.6	3	38	150	490
Type VII collagen	2928	31	118	29	77	190
p53	393	39	10	6	236	3076
Huntingtin	3144	189	67	8	201	2361
CFTR	1480	250	27	2.4	227	9100
Dystrophin	3685	2400	79	0.6	180	30,770

Geni parzialmente sovrapposti

- La densità genica varia ampiamente da cromosoma a cromosoma e tra differenti regioni dello stesso cromosoma
- In regioni ad alta densità geni possono essere parzialmente sovrapposti e generalmente trascritti in senso opposto (9% dei geni)



Esempio: 3 geni sovrapposti ad un introne di NF1

Principali categorie funzionali

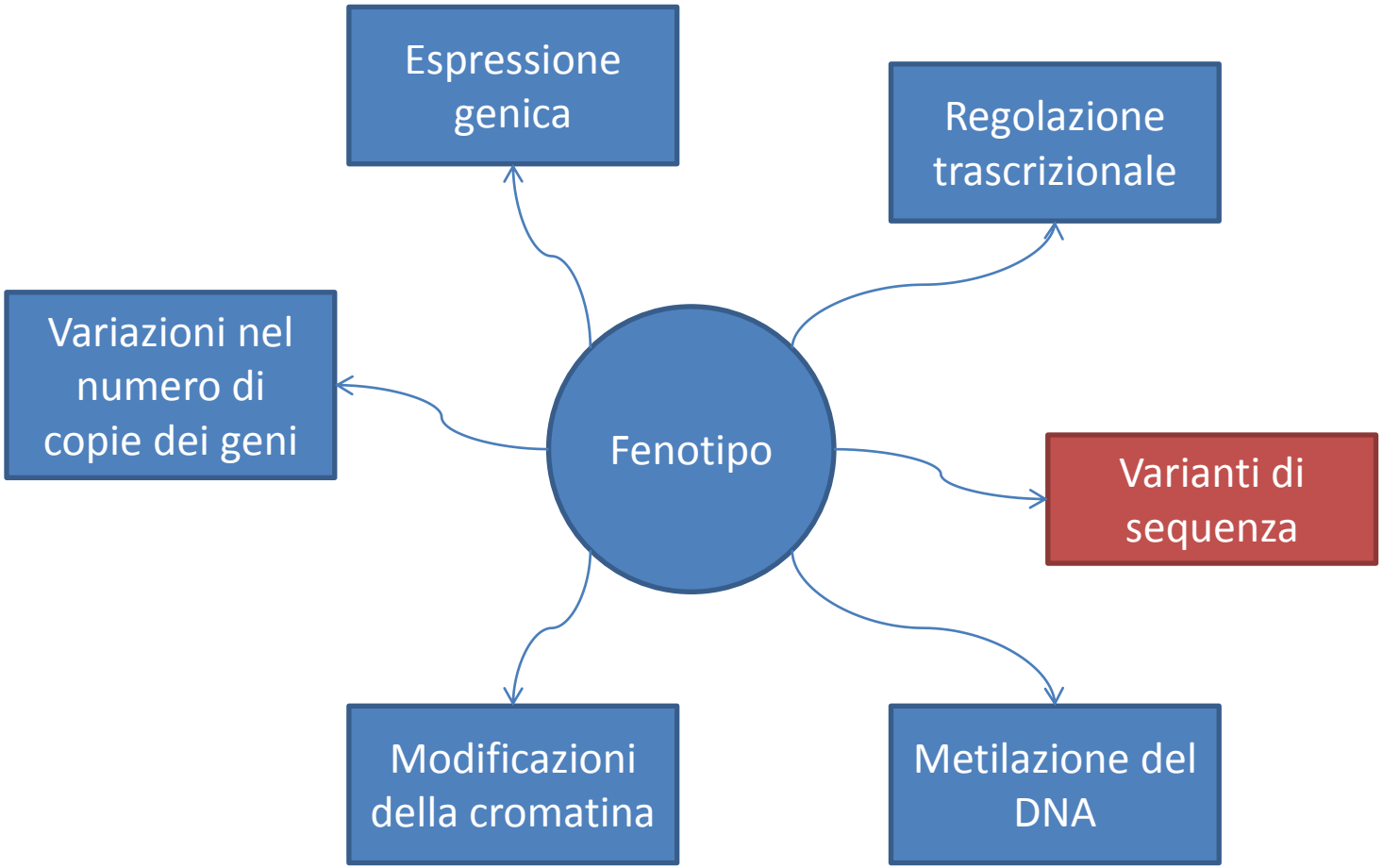
Function	Number	%
Nucleic acid binding	2207	14.0
DNA binding	1656	10.5
DNA repair protein	45	0.2
DNA replication factor	7	0.0
Transcription factor	986	6.2
RNA binding	380	2.4
Structural protein of ribosome	137	0.8
Translation factor	44	0.2
Transcription factor binding	6	0.0
Cell Cycle regulator	75	0.4
Chaperone	154	0.9
Motor	85	0.5
Actin binding	129	0.8
Defense/immunity protein	603	3.8
Enzyme	3242	20.6
Peptidase	457	2.9
Endopeptidase	403	2.5
Protein kinase	839	5.3
Protein phosphatase	295	1.8
Enzyme activator	3	0.0

Function	Number	%
Apoptosis inhibitor	132	0.8
Signal transduction	1790	11.4
Receptor	1318	8.4
Transmembrane receptor	1202	7.6
G-protein link receptor	489	3.1
Olfactory receptor	71	0.0
Storage protein	7	0.0
Cell adhesion	189	1.2
Structural protein	714	4.5
Cytoskeletal structural protein	145	0.9
Transporter	682	4.3
Ion channel	269	1.7
Neurotransmitter transporter	19	0.1
Ligand binding or carrier	1536	9.7
Electron transfer	33	0.2
Cytochrome P450	50	0.3
Tumor suppressor	5	0.0
Unclassified	4813	30.6
Total	**15683**	**100.0**

Quasi metà dei geni umani ha funzione ignota

Perché conoscere la sequenza del genoma è importante

- Per studiare la variabilità nell'uomo
- Per studiare l'espressione genica
- Per studiare le relazioni evoluzionistiche tra l'uomo e gli altri organismi
- Per identificare correlazioni tra l'informazione contenuta nel genoma e la suscettibilità e predisposizione alle malattie e le risposte ai farmaci (farmacogenomica)


Espressione genica
Regolazione trascrizionale
Variazioni nel numero di copie dei geni
Fenotipo
Varianti di sequenza
Modificazioni della cromatina
Metilazione del DNA

Differenze tra il genoma di diversi individui

- Fratelli sono 99.98% identici con regioni codificanti identiche al 99.99999%.
- Individui non imparentati sono al 99.8% identici con regioni codificanti identiche al 99.9999%.
- L'identità con gli scimpanzé è del 98%.
- L'identità con topo è del 90%.
- Con piante si arriva comunque ad un 40% circa di identità con le sequenze codificanti.

Variabilità genetica

- Il genoma umano contiene approssimativamente 10 milioni di polimorfismi, varianti genetiche che occorrono con una frequenza di 1% o più nella popolazione

- Gran parte di questi polimorfismi sono SNP (single nucleotide polymorphisms)

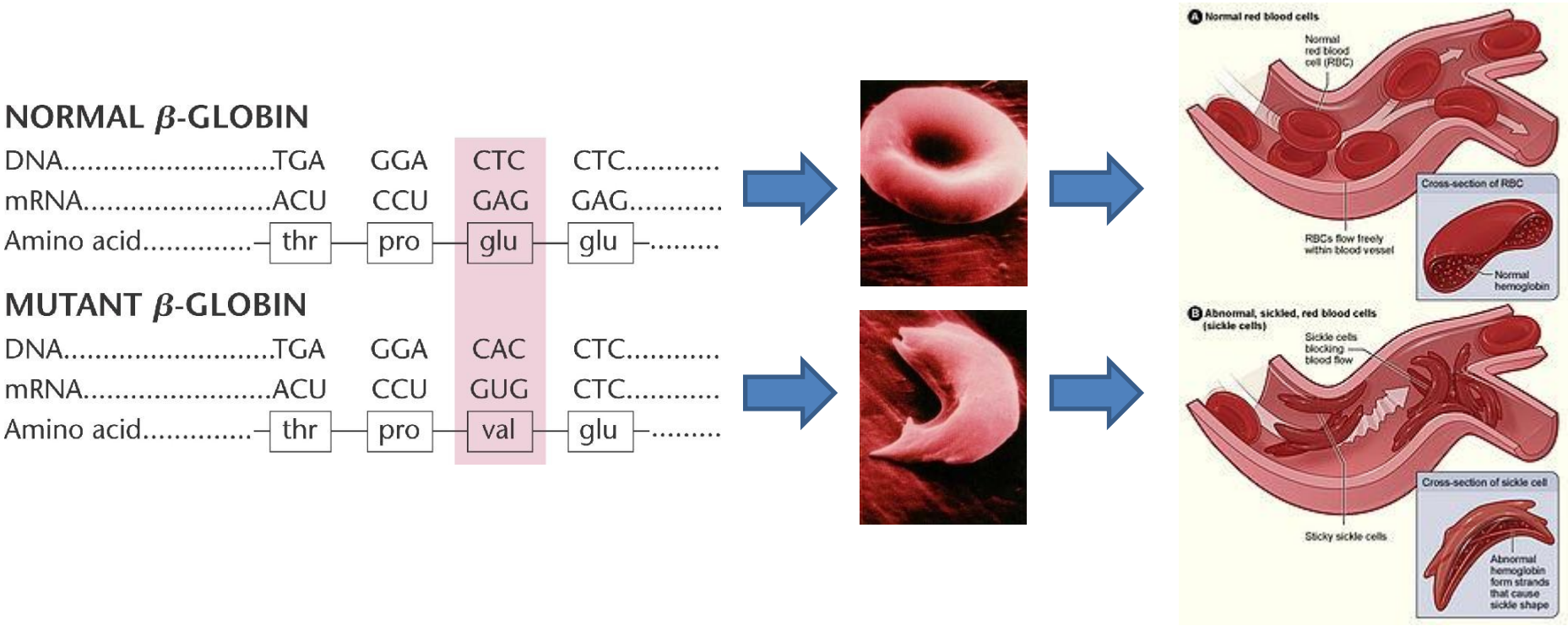
- Questi polimorfismi contribuiscono alla nostra individualità e influenzano la suscettibilità a varie malattie.

Motivazioni per studiare la variabilità genetica nell'uomo

- Studiare l'evoluzione della specie umana
- Comprendere la genetica alla base delle malattie, soprattutto quelle complesse (cardiovascolari, diabete, neurodegenerative…)
- Consentire l'applicazione di trattamenti farmaceutici personalizzati per il singolo individuo

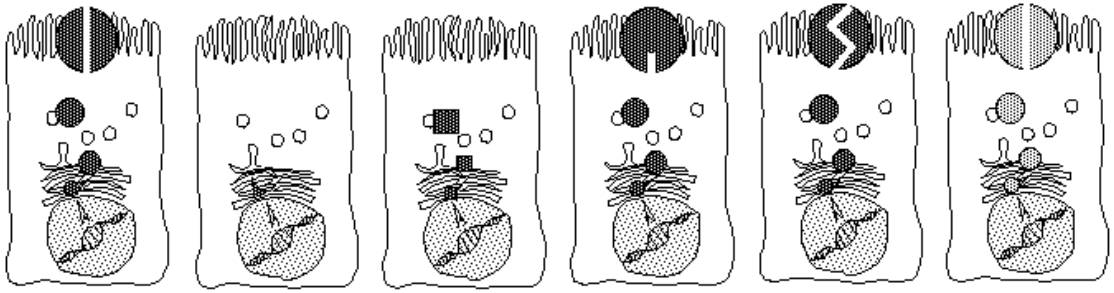
Anemia falciforme

Una singola mutazione nell'emoglobina porta le cellule rosse del sangue ad assumere la caratteristica forma a falce.



Fibrosi cistica

Mutazioni nel gene CFTR come codoni di stop prematuri, frameshift, delezioni o mutazioni missenso possono portare ad una mancata sintesi o ad una alterata funzionalità della proteina.

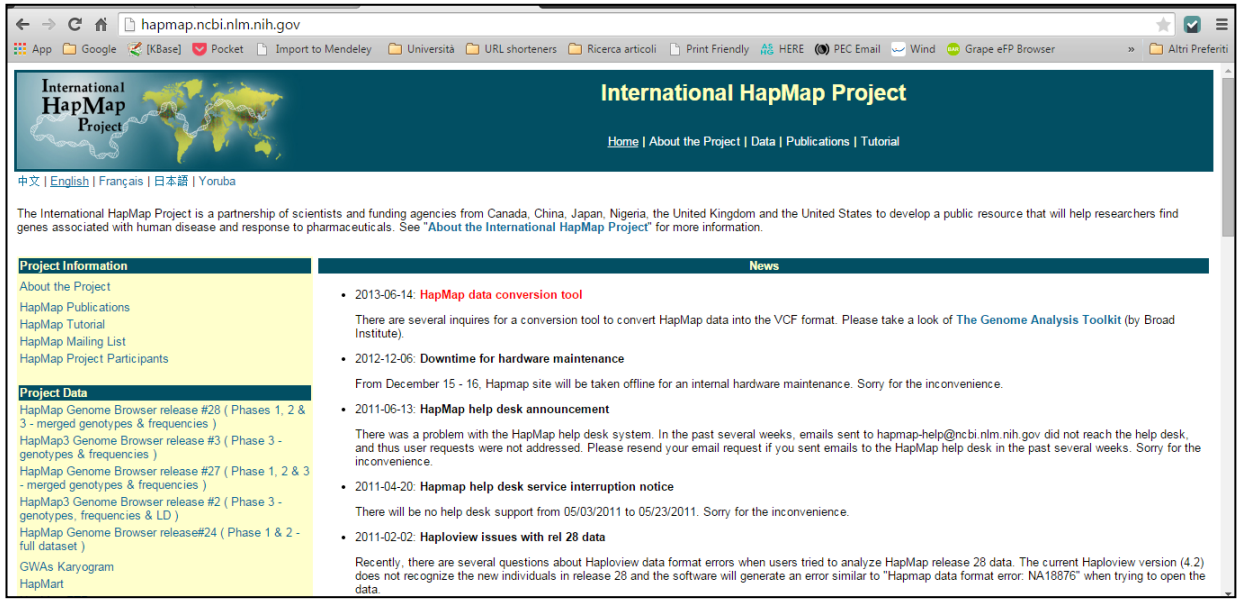


Normal	I	II	III	IV	V
	No synthesis	Block in processing	Block in regulation	Altered conductance	Reduced synthesis
	Nonsense G542X Frameshift 394delTT Splice junction 1717-1G→A	Missense N1303K AA deletion ΔF508	Missense G551D	Missense R117H R347P	Missense A455E Alternative splicing 3849+10kbC→T

HapMap project

Progetto internazionale, ha lo scopo di sviluppare una risorsa pubblica finalizzata ad aiutare nello studio dei geni associati con le malattie umane e la risposta ai farmaci.

Genotipizzati 1.6 milioni di SNP in 1184 individui da 11 popolazioni e sequenziate 10 regioni da 100 Kb in 692 di questi individui



http://hapmap.ncbi.nlm.nih.gov/

La corsa per il *Genoma a $1,000*

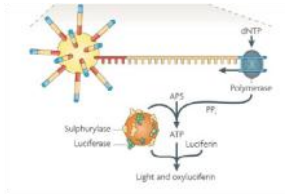
- L'HGP è costato 3 miliardi di dollari
- Con lo sviluppo di tecnologie bioinformatiche più avanzate applicabili allo *shotgun sequencing* Celera ha ridotto i costi a 300 milioni di dollari
- Nel 2007: 1-2 milioni di dollari per il sequenziamento del genoma di James Watson
- Il progresso negli ultimi anni delle tecnologie di sequenziamento sta consentendo di raggiungere il traguardo del genoma a 1000 dollari rendendo quindi possibile il sequenziamento di ciascun individuo ➔ il collo di bottiglia è l'interpretazione dei dati.

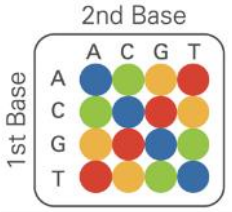
Next Generation Sequencing (NGS)

- 2005 – 454 Life Sciences acquired by Roche Diagnostics; based on emPCR and pyrosequencing. Massive parallel Sequencing by Synthesis.
- 2007 – SOLiD (Applied Biosystems): based on Sequencing By Ligation technology.
- 2007 – Solexa acquired by Illumina:

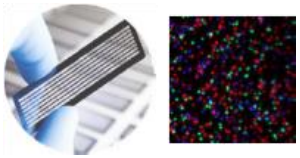
  Based on Sequencing by Synthesis with cleavable fluorescent didesoxy-nucleotides.
- 2011 – Pacific Biosciences: single molecule real time sequencing (SMRT)
- 2005 (founded) – Oxford Nanopore: nanopore based sequencing of single DNA molecules.

454 LIFE SCIENCES

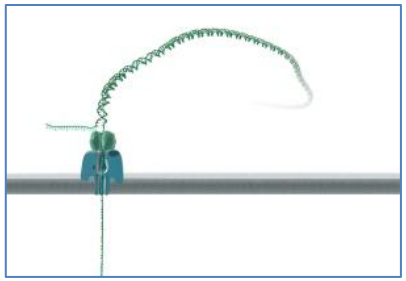




illumina®

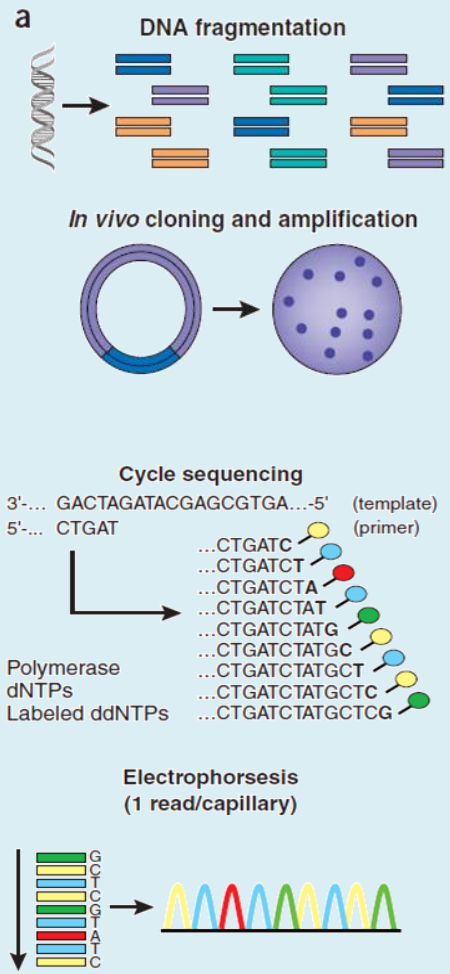


PACIFIC BIOSCIENCES™



Sanger vs NGS

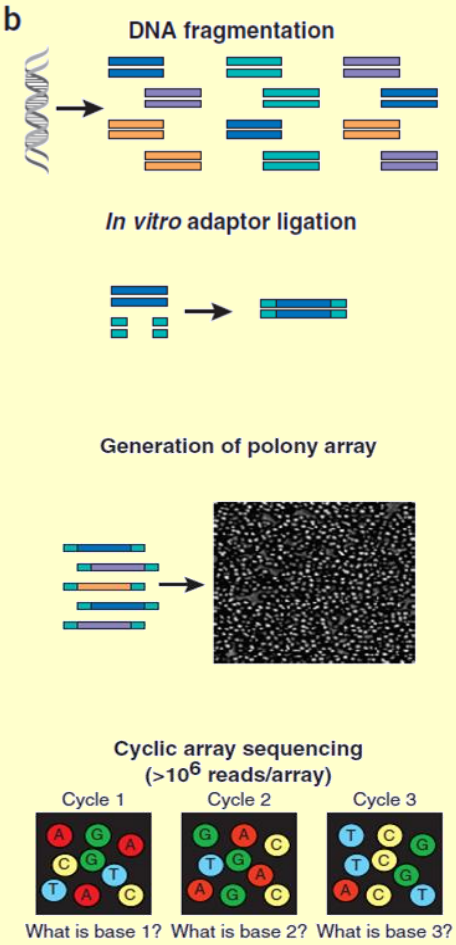
Sanger



Preparazione dei campioni più semplice e veloce per NGS (no clonaggio richiesto)

NGS consente una **elevata parallelizzazione** (fino a centinaia di milioni di reazioni di sequenza in parallelo)

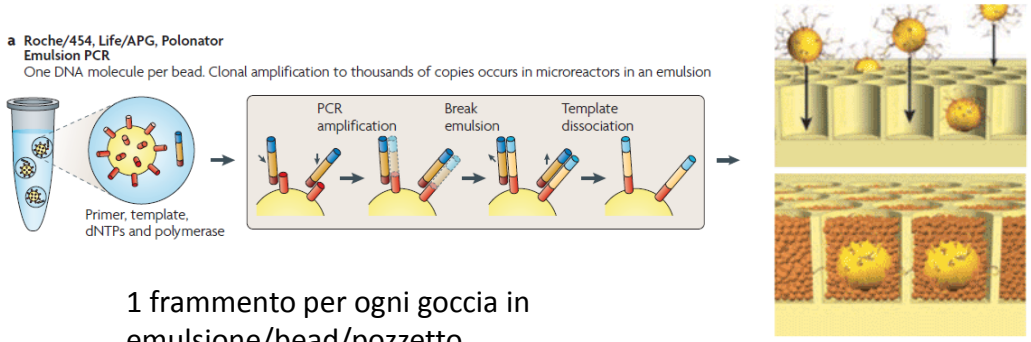
NGS



Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. Nature biotechnology, 26(10), 1135–45. doi:10.1038/nbt1486

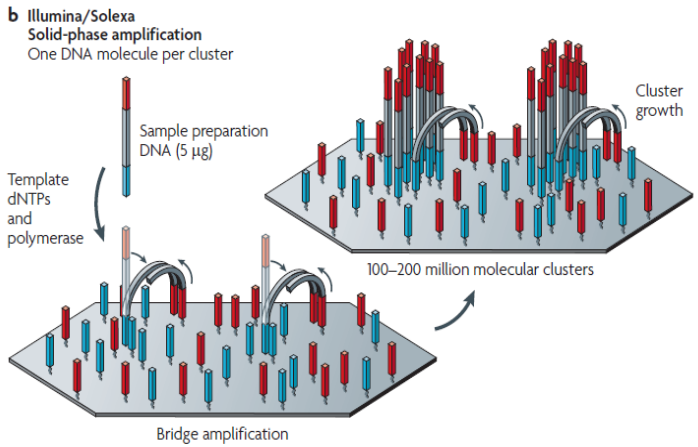
Parallelizzazione del sequenziamento

454



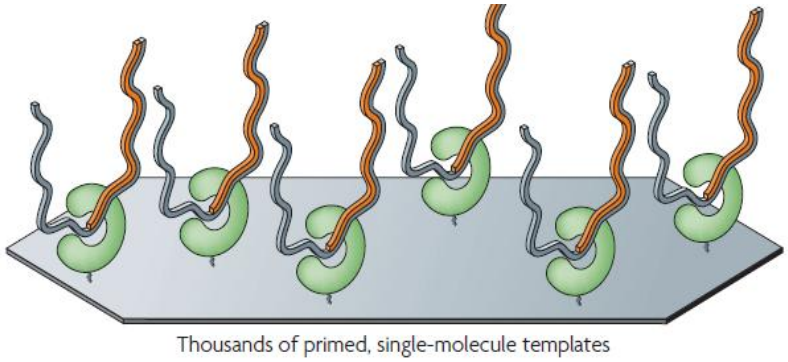
1 frammento per ogni goccia in emulsione/bead/pozzetto

Illumina



Clones of fragments are confined in clusters generated by bridge PCR (each cluster correspond to a fragment/sequence). Millions of clusters are generated per each lane

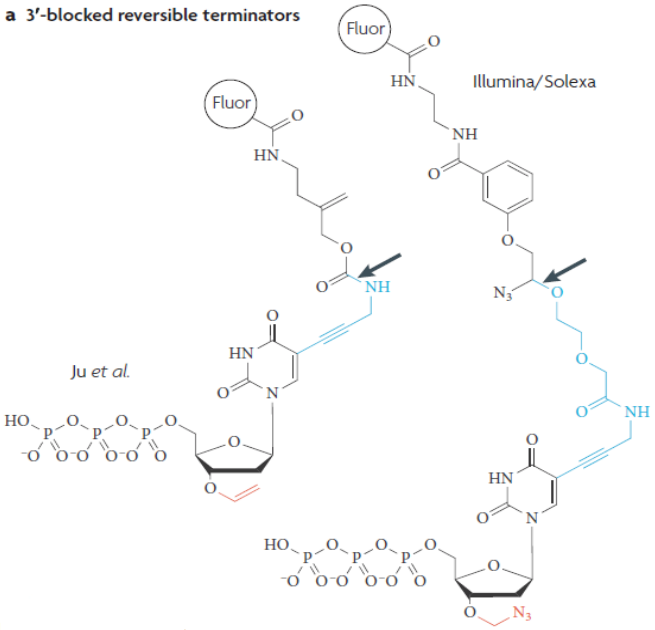
PacBio

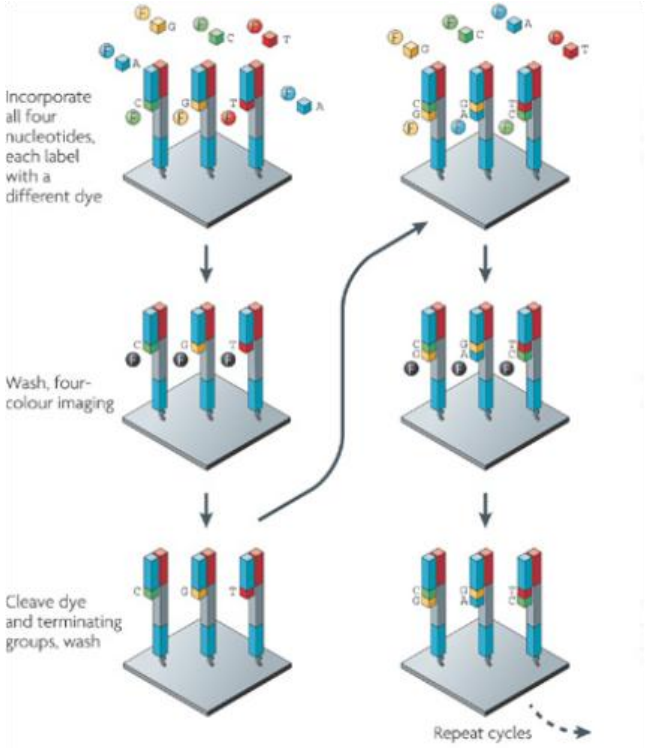


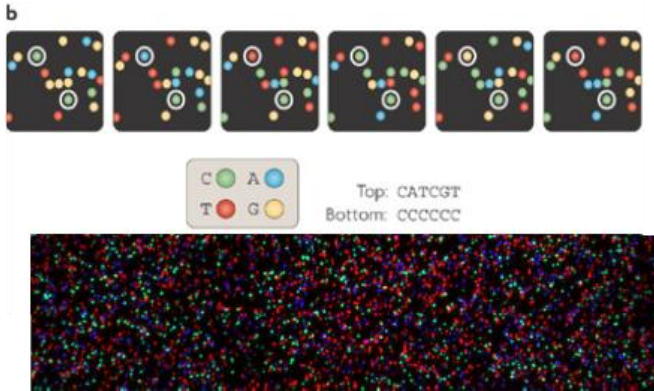
Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature, 456(7218), 53–9. doi:10.1038/nature07517
Metzker, M. L. (2010). Sequencing technologies - the next generation. Nature reviews. Genetics, 11(1), 31–46. doi:10.1038/nrg2626

illumina

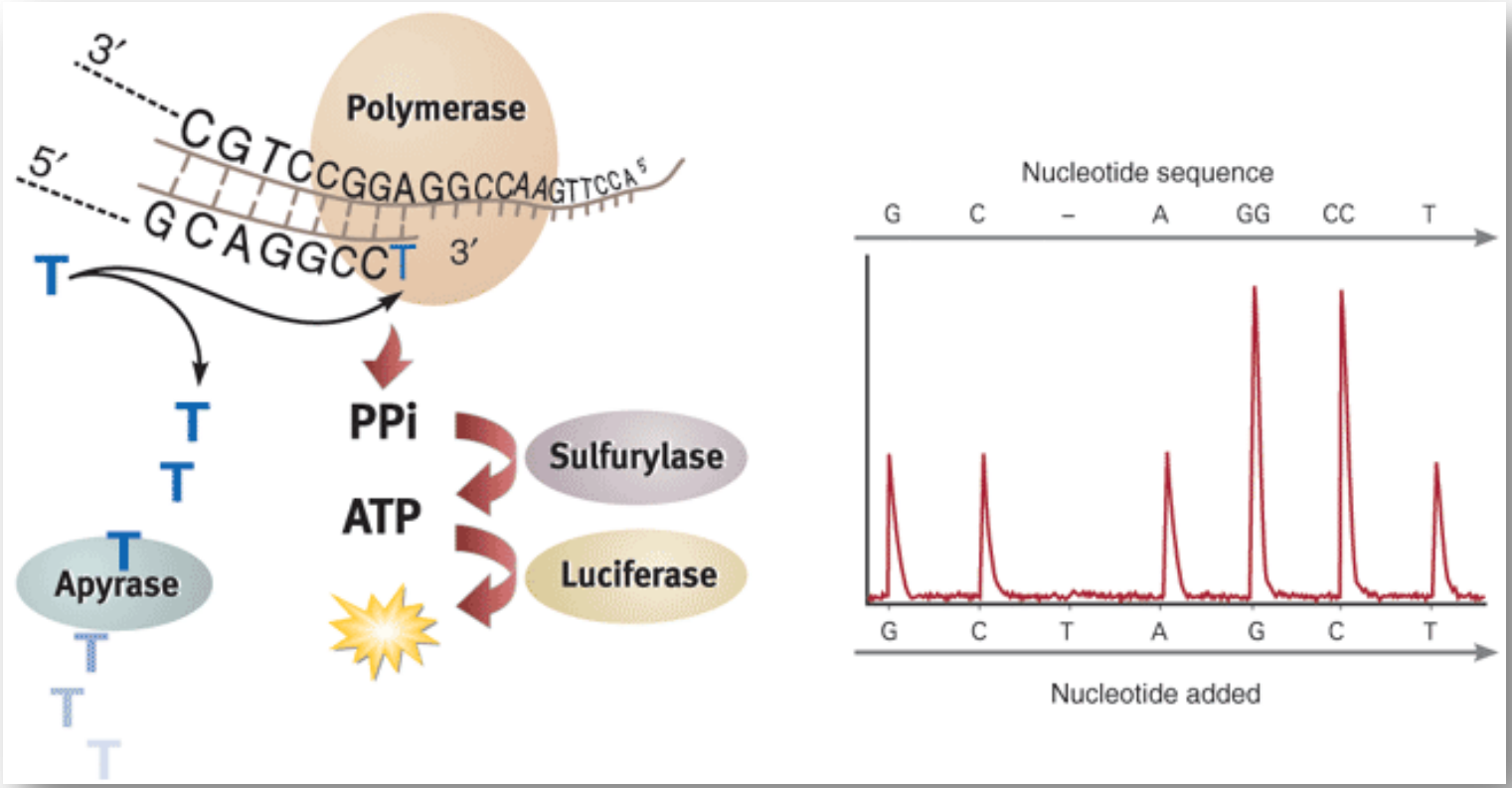
Actual sequencing is performed by addition of labeled reversible terminators (Sequencing by Synthesis; SBS). 1 base is added/read per cicle for all the fragments. The block and phluorophore is then removed and another cycle starts up to the desired read lenght.





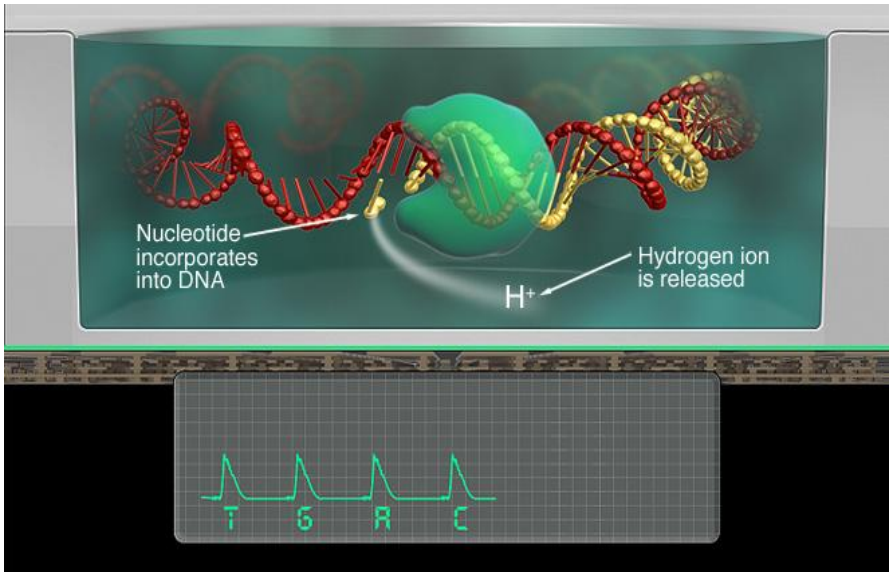


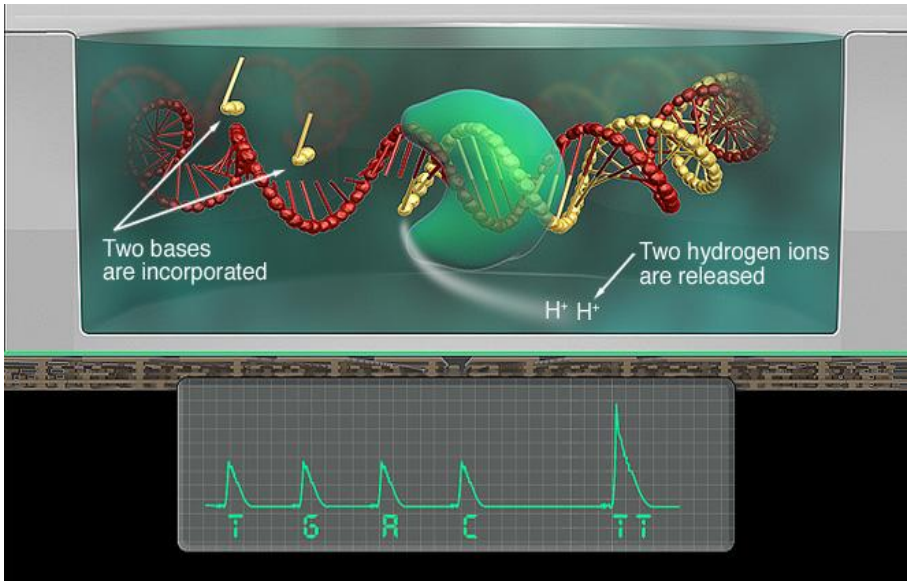
454



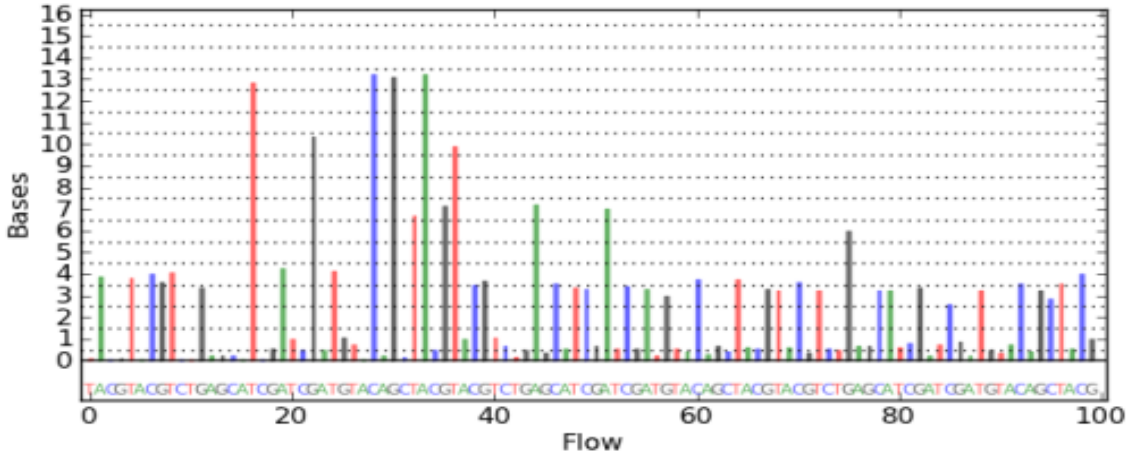
Utilizza pirosequenziamento: rilascio di gruppo fosfato viene rilevato come emissione di fluorescenza emessa da una luciferasi

Ion proton

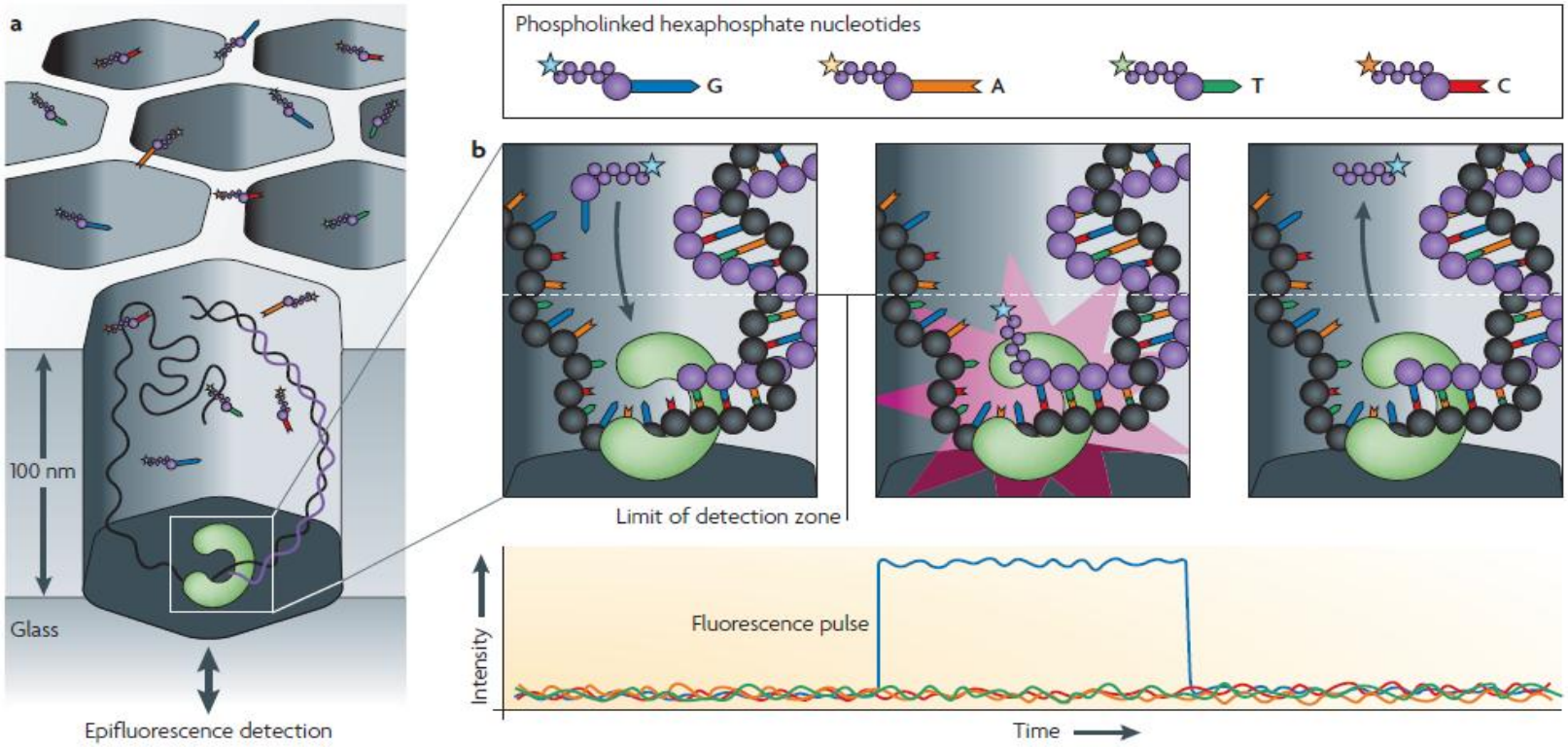




Misura il rilascio di protoni conseguente all'incorporazione di 1 o più nucleotidi nella copia del DNA sintetizzata a partire dal templato da sequenziare



PacBio

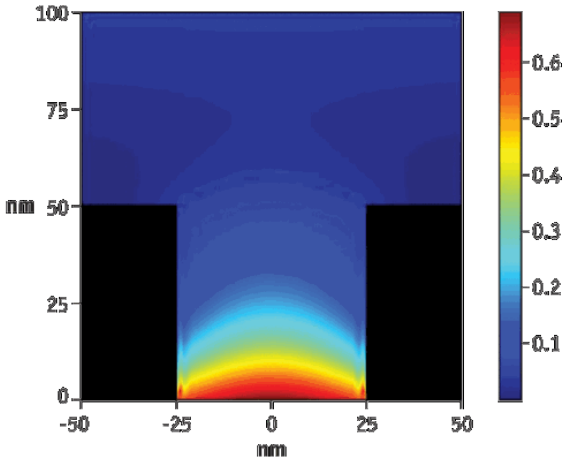


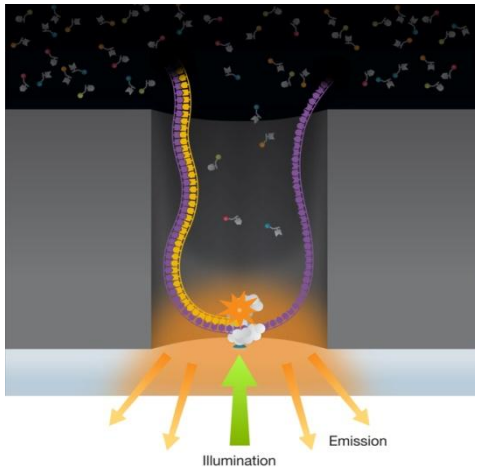
Metzker, M. L. (2010). Sequencing technologies - the next generation. Nature reviews. Genetics, 11(1), 31–46. doi:10.1038/nrg2626

Sequenziamento di una singola molecola di DNA per “pozzetto”.
Viene misurata la fluorescenza emessa dal nucleotide fluorescente al momento dell’incorporazione.
Basato su Zero Mode Waveguide.

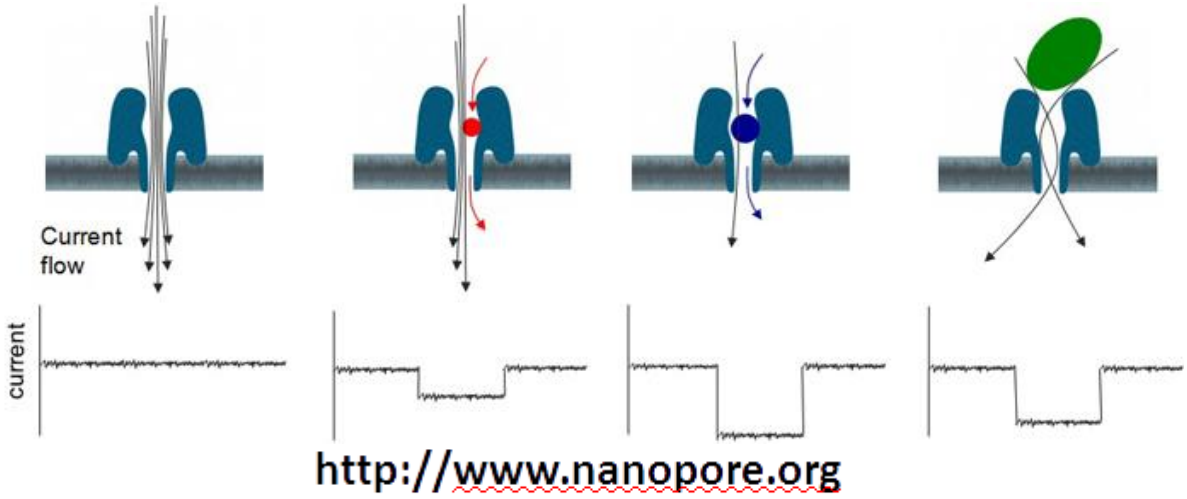
Zero Mode Waveguide

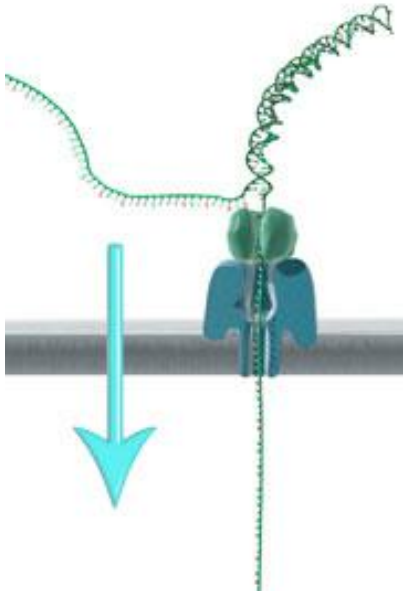
•I fori sulla *zero mode waveguide* hanno dimensioni comparabili alla lunghezza d'onda del laser utilizzato per eccitare il fluoroforo.
•Questo causa una attenuazione della luce che riesce a penetrare solo i primi 20-30 nm della guida d'onda ottica.
•Il volume su cui avviene la detection è nell'ordine degli zeptolitri (1E-21).
•Il nucleotide marcato in fluorescenza viene quindi rilevato solo al momento dell'incorporazione da parte della polimerasi che si trova sul fondo del "pozzetto" mentre i nucletodi liberi non vengono rilevati.





Nanopore

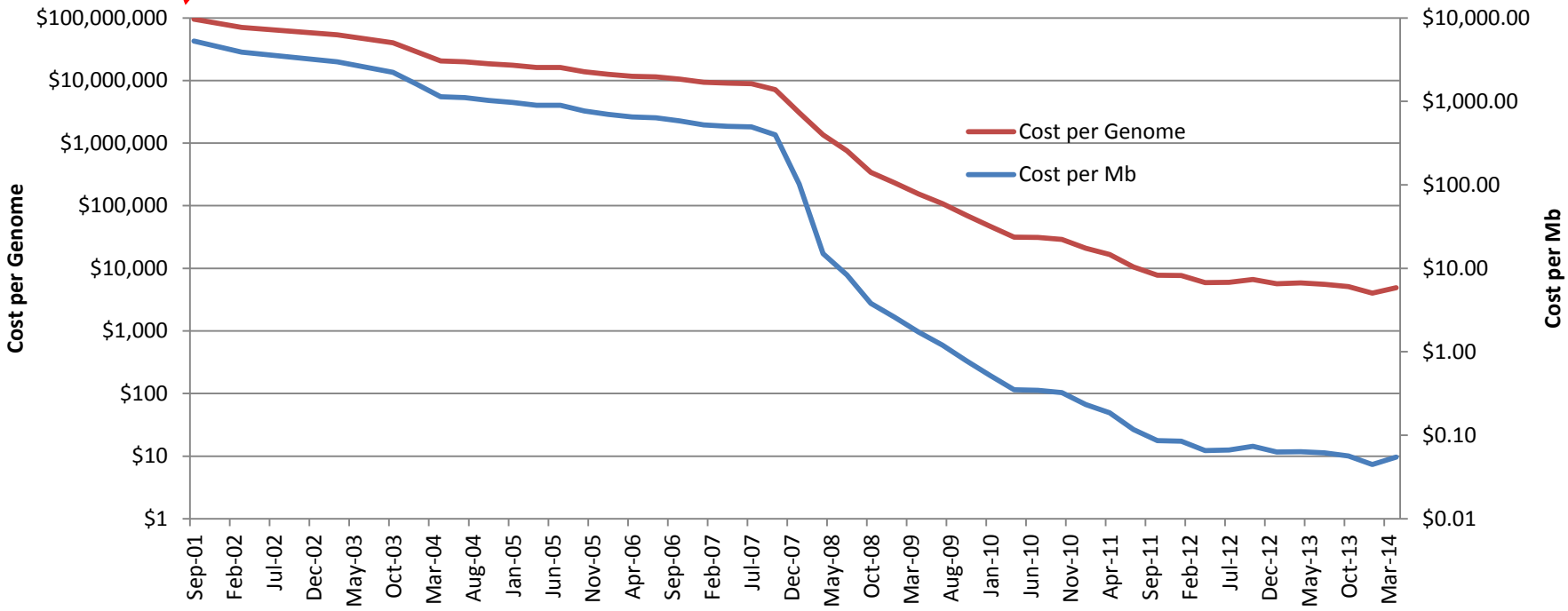




- Easy sample preparation (no labeling)
- Fast
- Cheap
- Long sequences (important to determine aplotypes); aiming to 100,000 bases sequences
- Complete transcripts sequenced
- Almost solves the assembly problem

Diminuzione del costo del sequenziamento del genoma

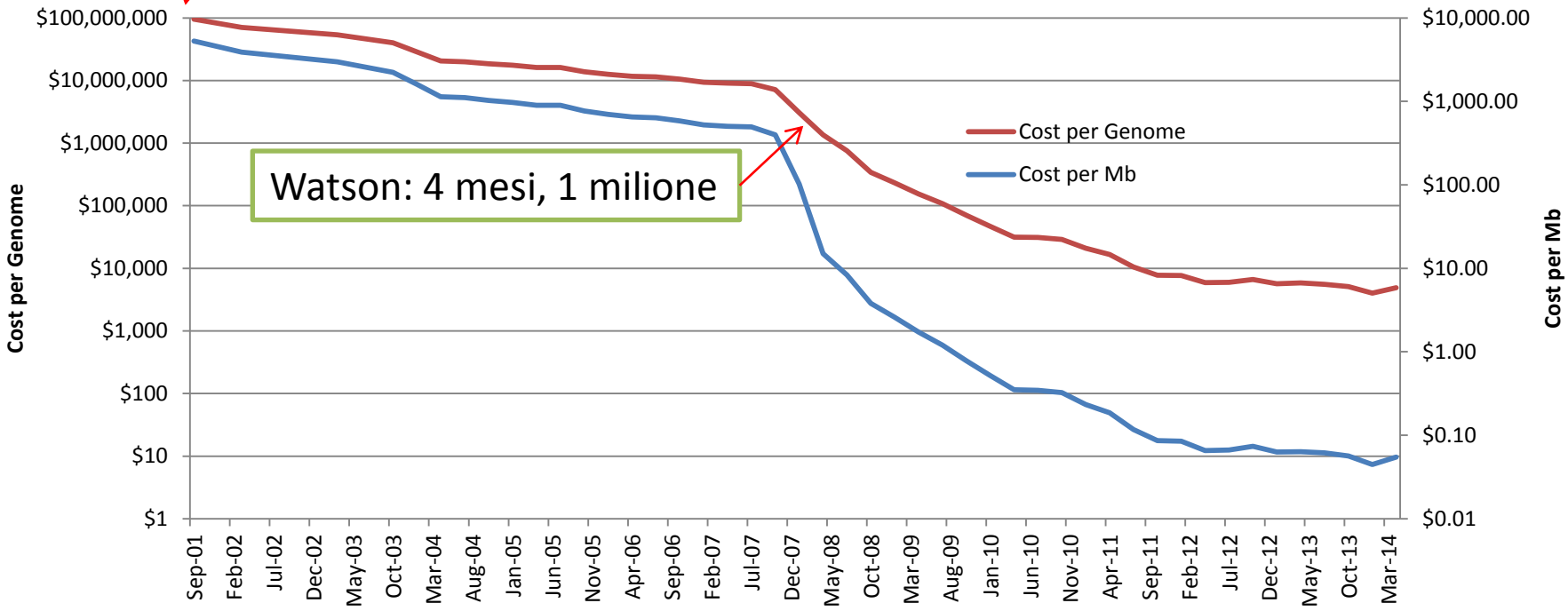
HGP: 13 anni, miliardi di dollari
Celera: 3 anni, centinaia di migliaia di dollari



Costo del genoma oggi è alla portata del sequenziamento del genoma di ciascun individuo:
→ "Genomica personalizzata" → Medicina di precisione

Diminuzione del costo del sequenziamento del genoma

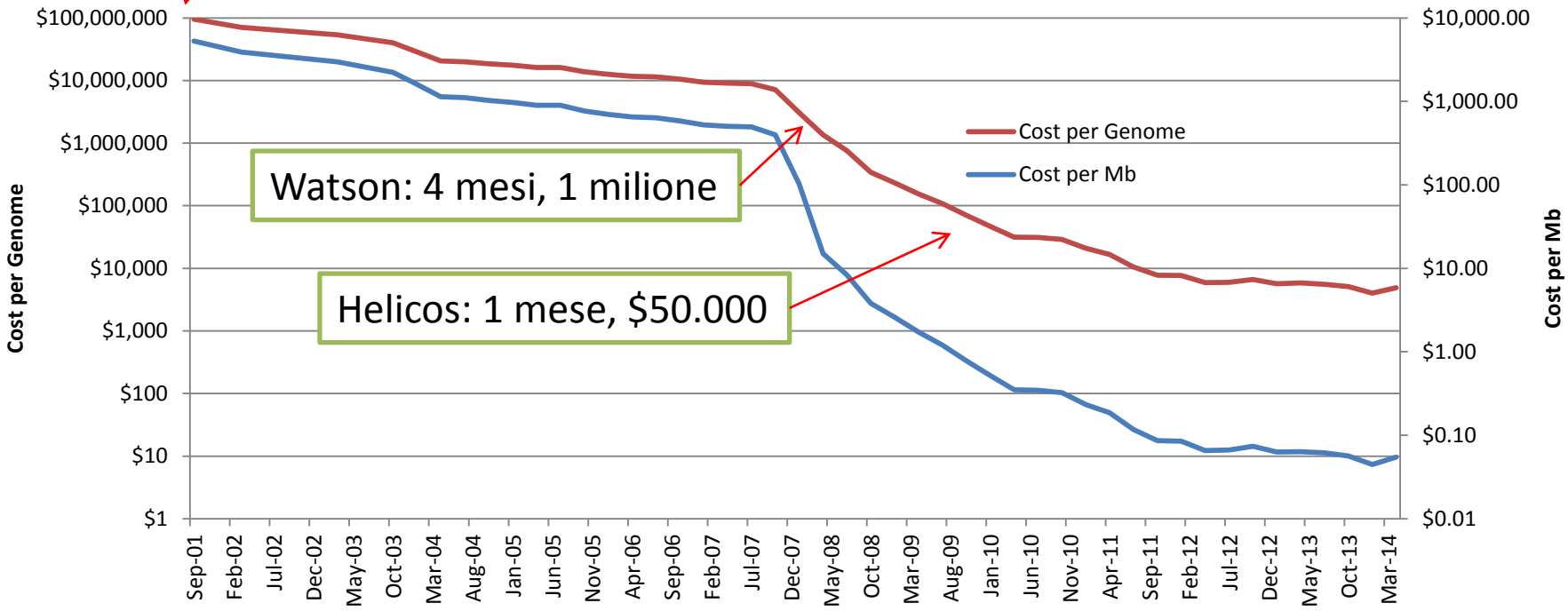
HGP: 13 anni, miliardi di dollari
Celera: 3 anni, centinaia di migliaia di dollari



Costo del genoma oggi è alla portata del sequenziamento del genoma di ciascun individuo:
→ "Genomica personalizzata" → Medicina di precisione

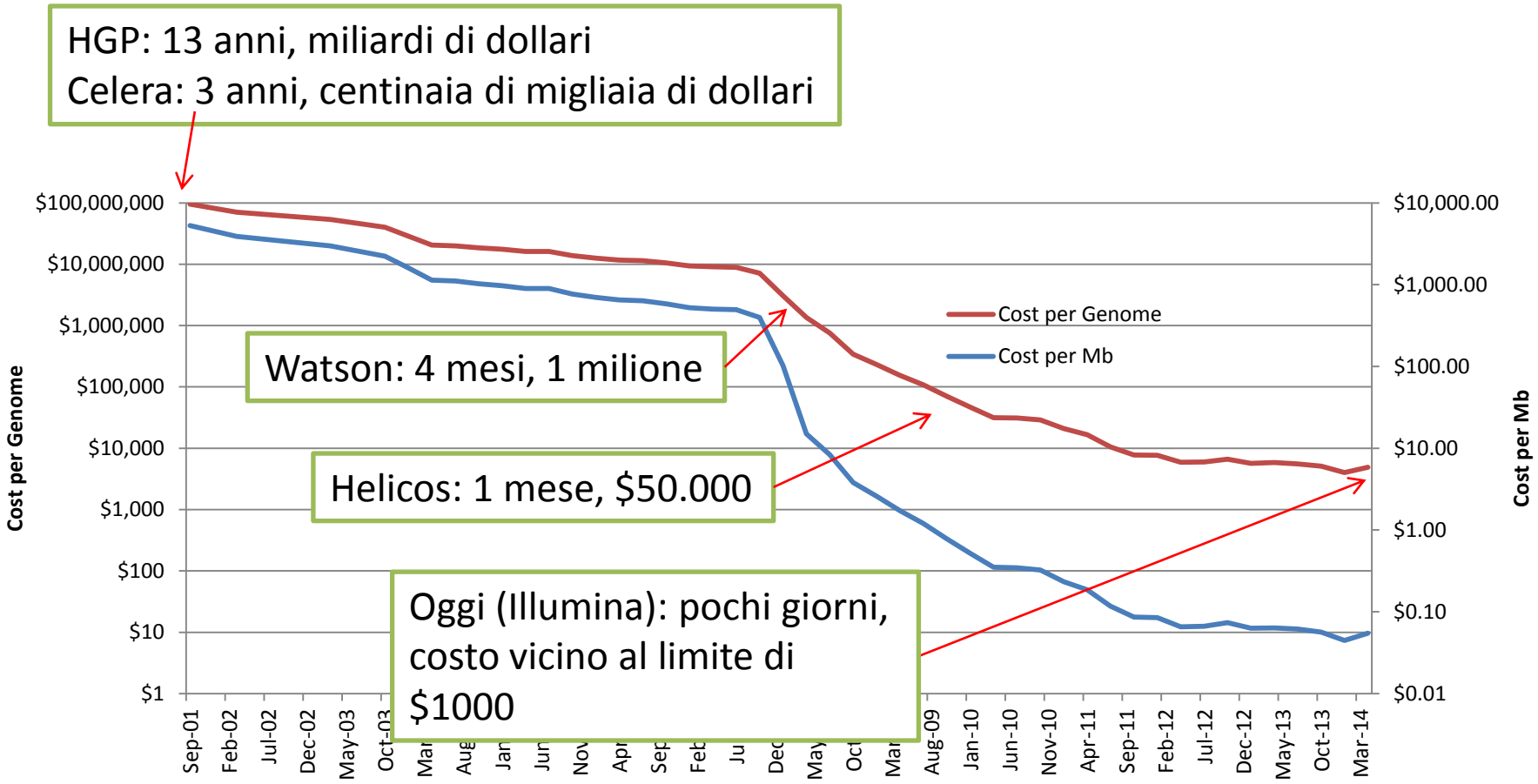
Diminuzione del costo del sequenziamento del genoma

HGP: 13 anni, miliardi di dollari
Celera: 3 anni, centinaia di migliaia di dollari



Costo del genoma oggi è alla portata del sequenziamento del genoma di ciascun individuo:
→ “Genomica personalizzata” → Medicina di precisione

Diminuzione del costo del sequenziamento del genoma



Costo del genoma oggi è alla portata del sequenziamento del genoma di ciascun individuo:
→ "Genomica personalizzata" → Medicina di precisione

Cosa significa

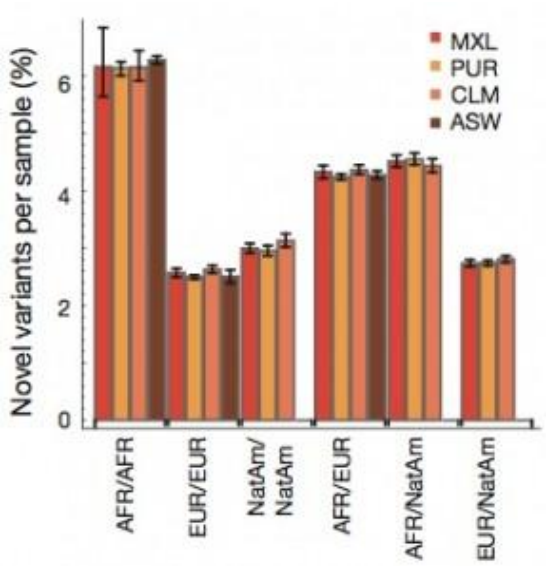
- Democratizzazione del sequenziamento: qualsiasi laboratorio sarà a breve in grado di sequenziare un genoma
- Il costo del sequenziamento ormai sta raggiungendo gli stessi livelli di altri test diagnostici come test prenatali o test genetici per la predisposizione a malattie: destinato a sostituire tutti gli altri test genetici

1000 Genomes project

2008 (-2012 phase1; 2014 phase3)

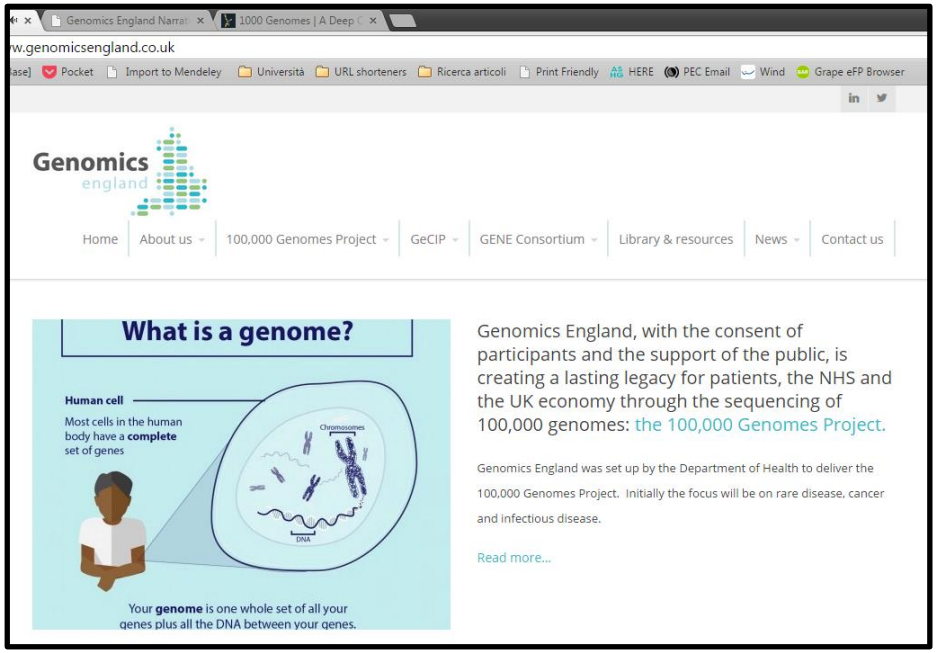
- Progetto finalizzato a costruire una risorsa per capire il contributo genetico alle malattie
- Basato su nuove tecnologie di sequenziamento
- Sequenziati circa 2500 individui da 25 popolazioni da Europa, Africa, Asia e Americhe

•Il goal è stato quello di caratterizzare anche varianti rare presenti in meno dell'1% dei cromosomi umani sottorappresentate negli studi precedenti.
•Dal 3% al 6% di varianti nuove identificate per popolazione.
•Varianti rare sono interessanti tendenzialmente arricchite in varianti "funzionali".
•Utili per studi su contributo genetico alle malattie.

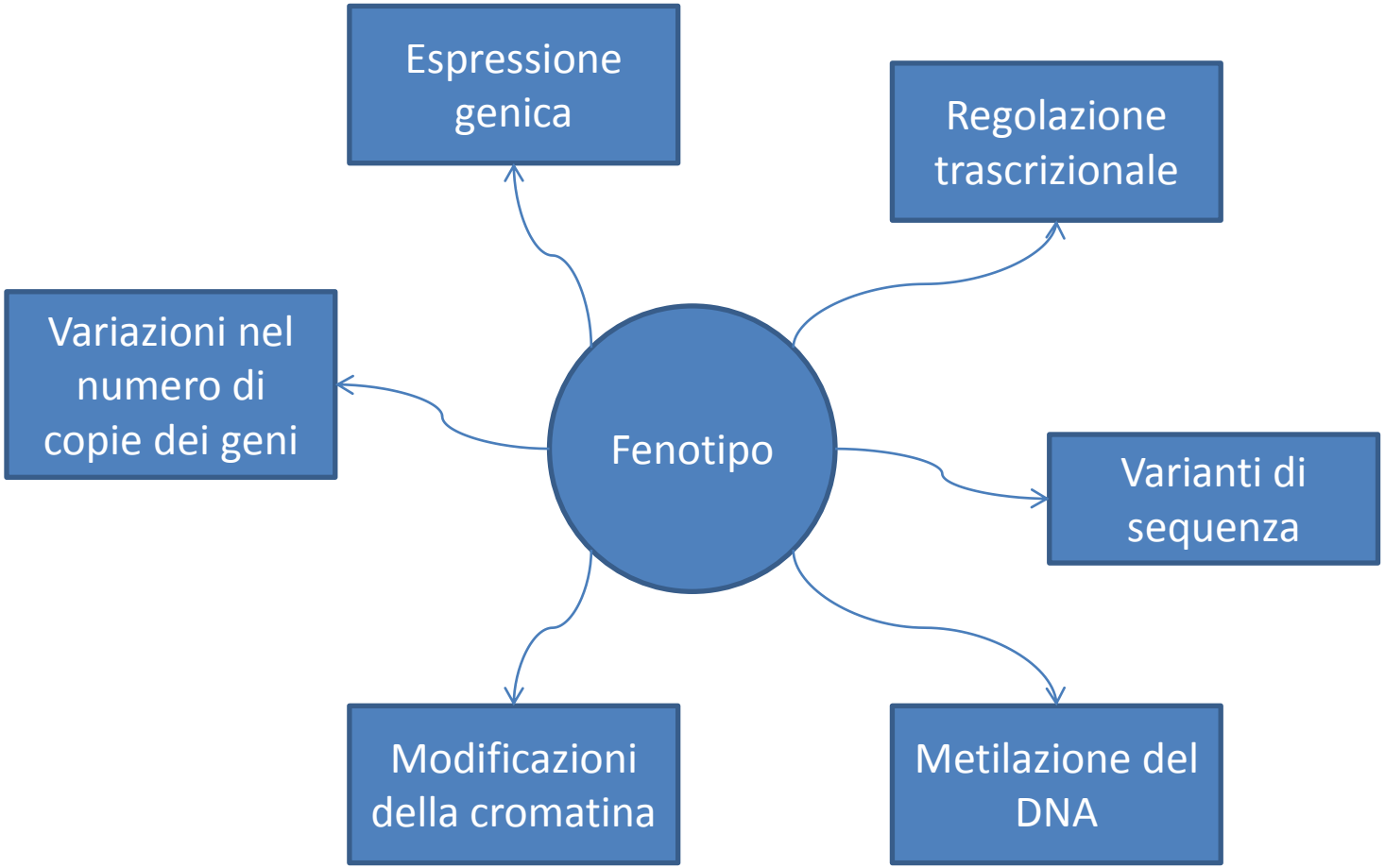


Genomics England

- Progetto annunciato dal primo ministro inglese nel 2012
- Finalità: sequenziare il genoma di 100.000 individui entro il 2017
- Individui corredati da misure fisiologiche, dati clinici, storico di record medici.
- Finalità:
  - Creare un programma etico e trasparente basato sul consenso
  - Portare benefici a pazienti e avviare un servizio di medicina genomica
  - Consentire nuove scoperte scientifiche
  - Dare avvio allo sviluppo ad una industria per la genomica in UK



http://www.genomicsengland.co.uk/

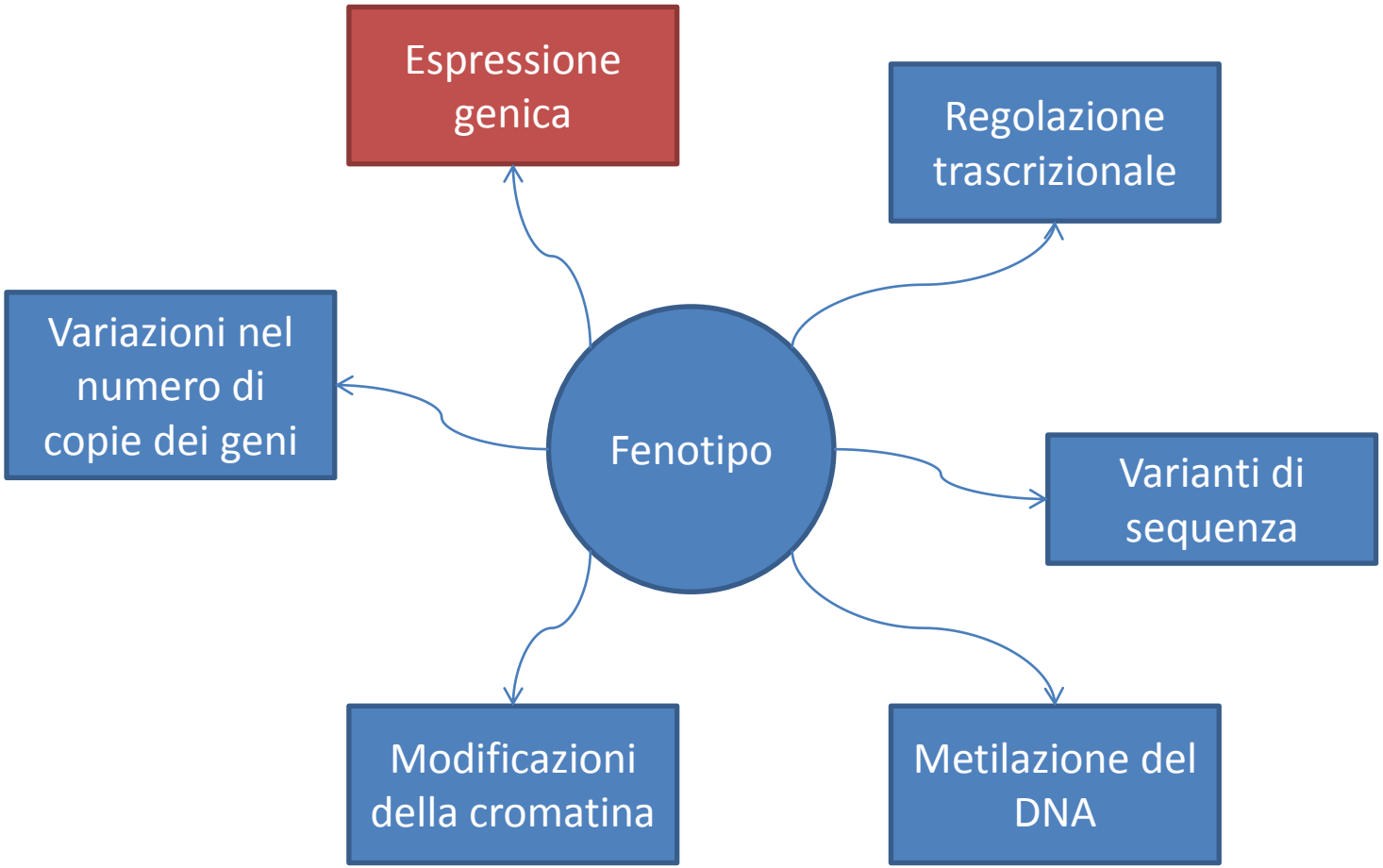

Espressione genica
Regolazione trascrizionale
Variazioni nel numero di copie dei geni
Fenotipo
Varianti di sequenza
Modificazioni della cromatina
Metilazione del DNA

ENCODE

Il progetto ENCODE ha avuto come finalità la costruzione di un catalogo degli elementi funzionali nel genoma umano incluse regioni trascritte, elementi regolatori, modificazioni del DNA e delle proteine legate al DNA.

Anche se solo l'1% codifica per proteine il progetto ha messo in evidenza che più dell'80% del genoma è funzionalmente "attivo".

2012


Espressione genica
Regolazione trascrizionale
Variazioni nel numero di copie dei geni
Fenotipo
Varianti di sequenza
Modificazioni della cromatina
Metilazione del DNA

Transcriptome

- *The complete set of all the mRNAs of an organism/cell type in a given moment.*

- *Transcriptome is dinamic and changes continuously depending on the particular conditions considered. Different conditions correspond to different expression profiles.*

➔ *Transcriptomics: the study of transcriptome; the analysis of transcriptome in different conditions allows to infer which genes are putatively involved in a give developmental process, stress response…*

Post-genomics

- The study of the transcriptome has become especially important in the post-genomics era:
  - As many genome sequences have been released during the last years it has become important to be able to infer function for the genes encoded by the genomic sequence.

Applicazioni dell'analisi del trascrittoma

- Identificare geni espressi in differenti tipi cellulari
- Analizzare come i livelli di espressione cambiano in differenti stadi di sviluppo
- Analizzare come i livelli di espressione cambiano durante lo sviluppo di una malattia
- Analizzare le relazioni tra gruppi di geni
- Identificare i processi cellulari a cui partecipano i geni

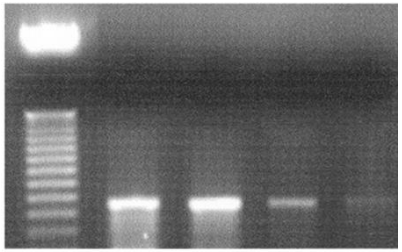
Central "*Assumption*" of Gene Expression analysis

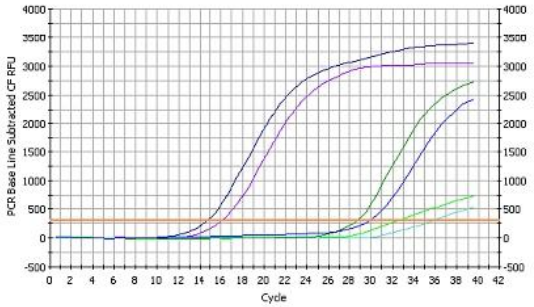
- The level of a given mRNA is positively correlated with the expression of the associated protein.
  - Higher mRNA levels mean higher protein expression, lower mRNA means lower protein expression
- Other factors:
  - Protein degradation, mRNA degradation, polyadenylation, codon preference, translation rates, alternative splicing, translation lag…
- *This is relatively obvious, but worth emphasizing*

Gene expression analysis

Before the "omics" technologies

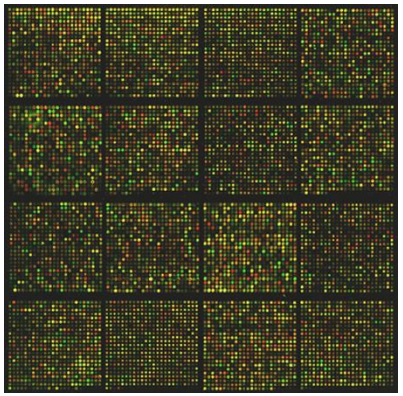
- One or few genes analyzed per time by semiquantitative or quantitative PCR

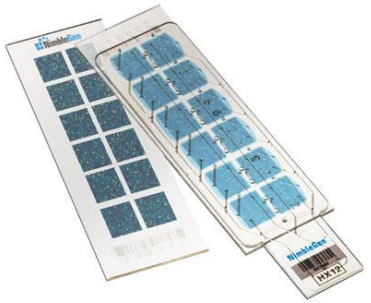




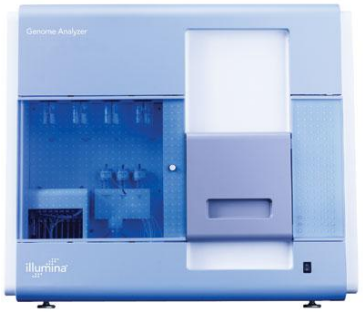
Now

- From few thousands of genes to complete transcriptomes analyzed in a single experiment





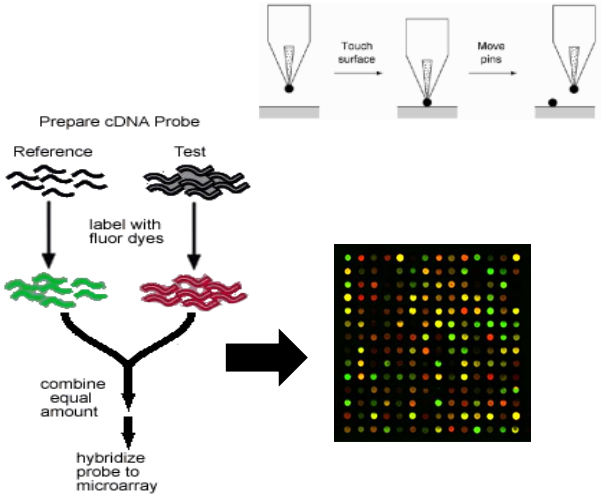
Microarray

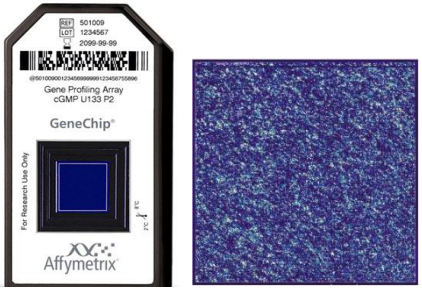


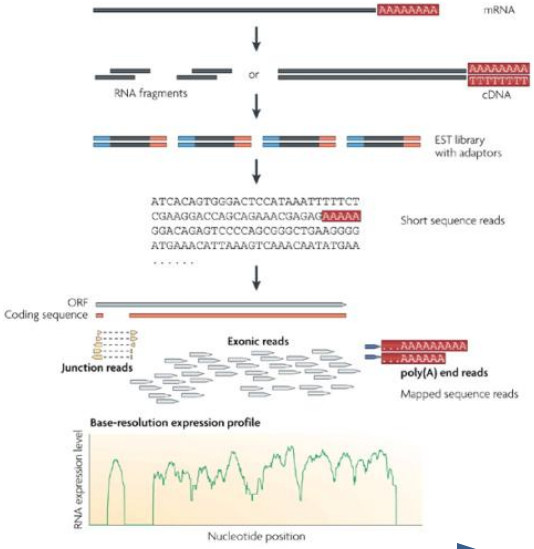
```
@ILLUMINA-C3C24B_0053:6:1:1429:17486#0/1
GTTCTACCAGAACTTGCCTGACCTTCTGCATCAGTGGAT
+ILLUMINA-C3C24B_0053:6:1:1429:17486#0/1
bb^aca`b^^``babbbabbbbbbb`bbcbbbbbbabbb
@ILLUMINA-C3C24B_0053:6:1:1429:12585#0/1
GGTTGACTGGACATCTAGGGGTAAAGCACTGTTTCGGTG
+ILLUMINA-C3C24B_0053:6:1:1429:12585#0/1
b`b`bbabbbbbbbbbabbbb_bbbbbbbbbbbbbbbYb
```

Next Generation Sequecing (NGS)

Evoluzione delle tecnologie di analisi del trascrittoma







1995- Sviluppati i primi microarray basati su spotting di molecole di cDNA

2002- High density oligo microarrays

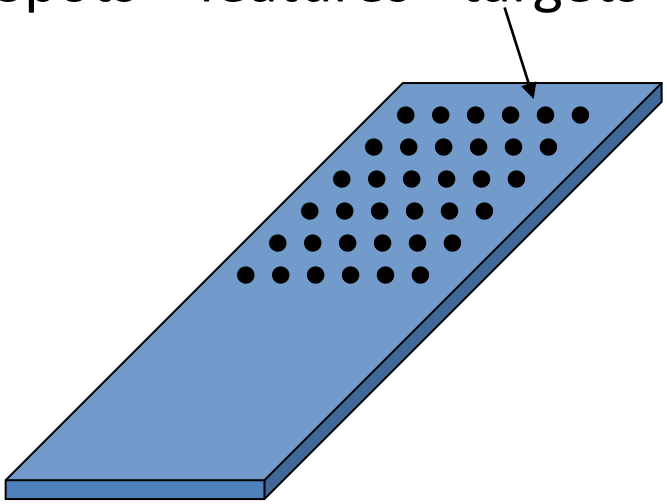
2008- RNA-Seq: sequenziamento dei messaggeri basato su tecnologie NGS

Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray- Schena et. al.

Microarray

Sonde a DNA o RNA complementari ai geni di cui vogliamo l'espressione genica sono immobilizzate per deposizione o sintetizzate direttamente su una superficie solida (vetro, silicio).

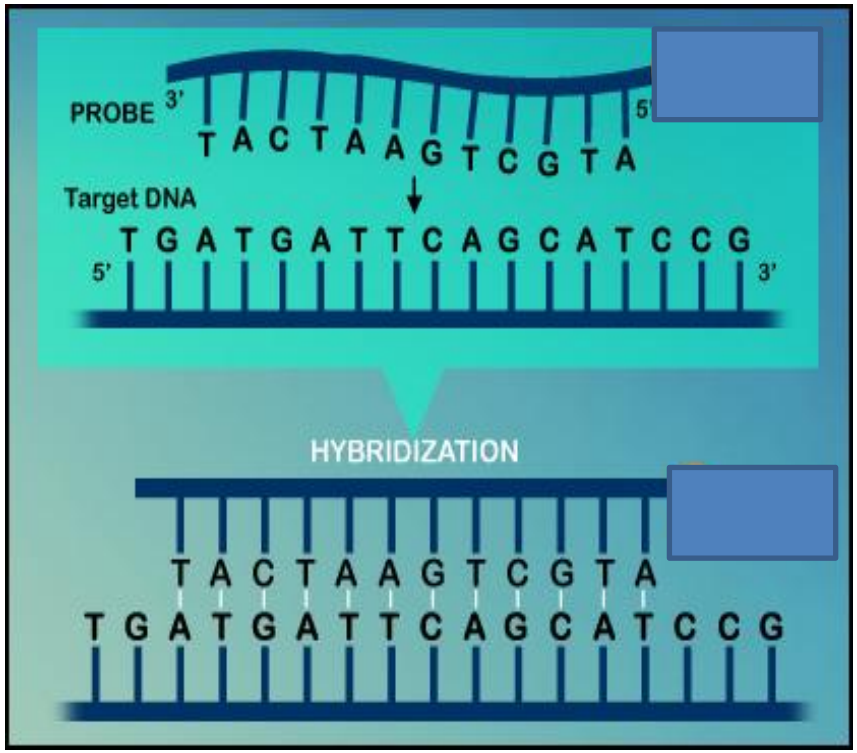
Spots – features - targets

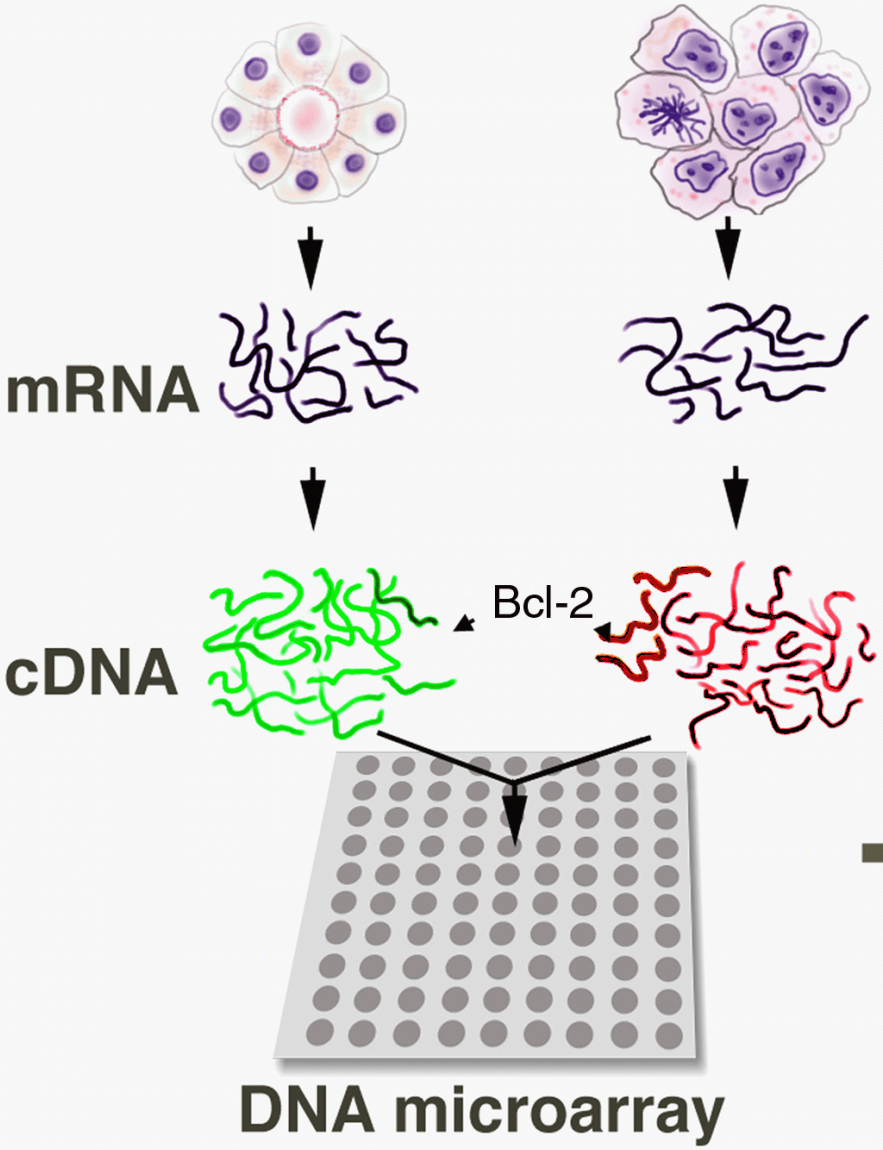


Fino a milioni di sonde su un singolo array. E' possibile rappresentare l'intero set di geni umani su un singolo chip

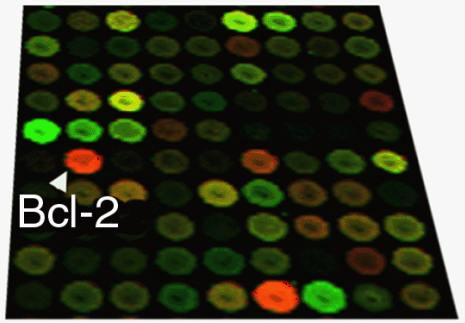
Principio del microarray

- Microarray si basa su ibridazione tra una sonda immobilizzata a singolo filamento con un DNA target complementare denaturato (singolo filamento) → tendono naturalmente ad appaiarsi in un doppio filamento



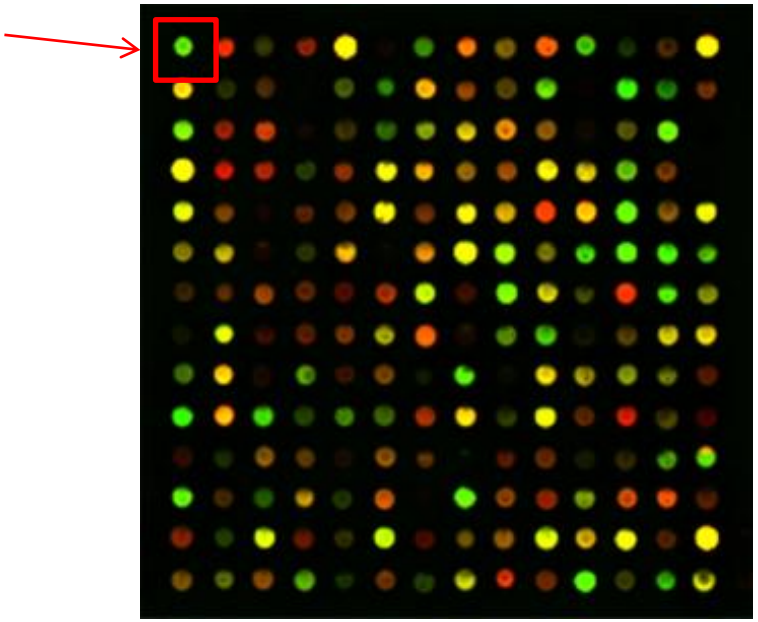


mRNA estratti da diversi campioni vengono marcati con fluorofori di diverso colore e ibridati alle sonde immobilizzate sull'array. Gli mRNA andranno a legarsi alla sonda complementare al gene da cui sono stati trascritti



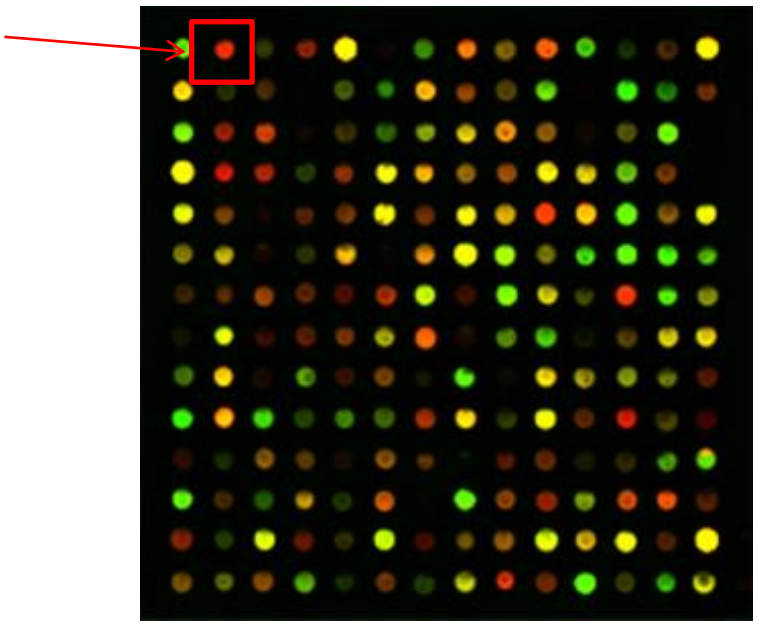
Ogni spot rappresenta un gene diverso.

Se lo spot corrispondente ad un determinato gene si illumina di verde significa che quel gene è più espresso nel campione marcato con il verde



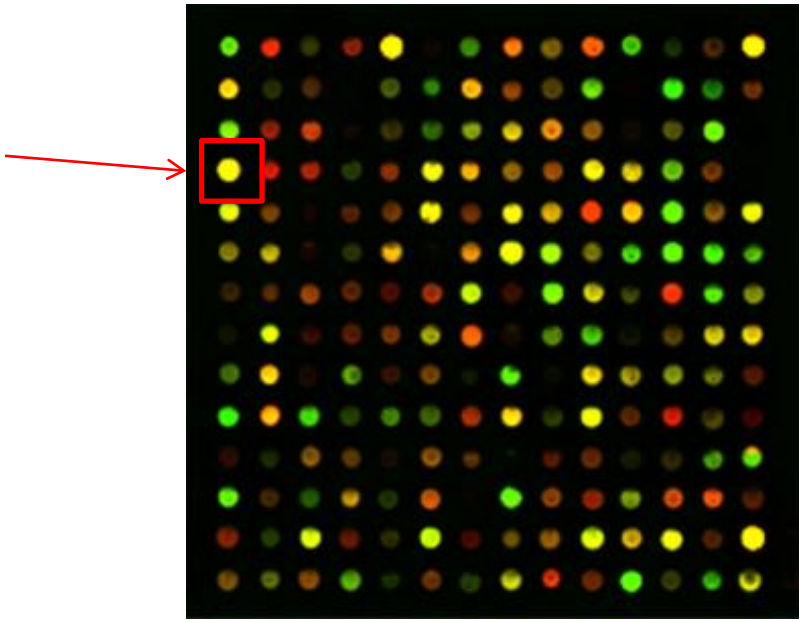
Ogni spot rappresenta un gene diverso.

Se lo spot corrispondente ad un determinato gene si illumina di rosso significa invece che quel gene è più espresso nel campione marcato con il rosso



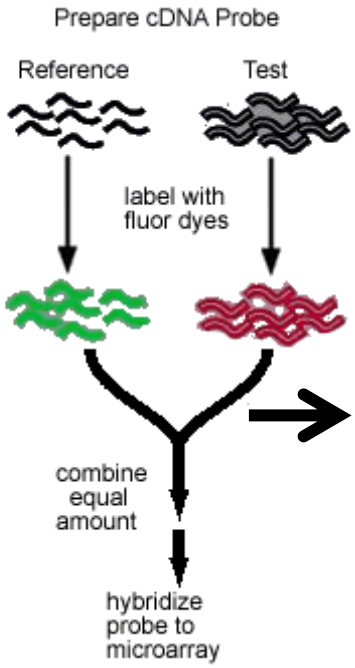
Ogni spot rappresenta un gene diverso.

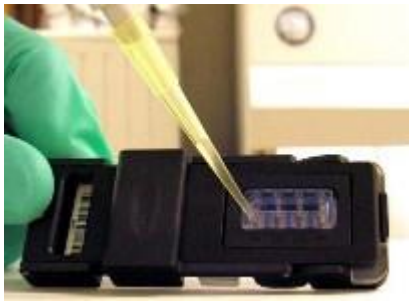
Se lo spot corrispondente ad un determinato gene si illumina di giallo significa invece che quel gene è espresso in ugual misura nei due campioni



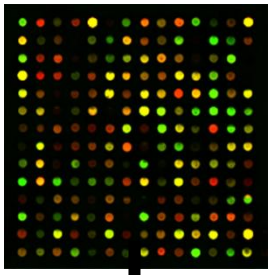
Microarray

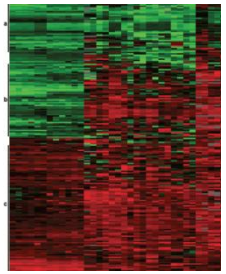
I valori di fluorescenza vengono registrati tramite uno scanner e convertiti in valori numerici che possono essere ulteriormente analizzati



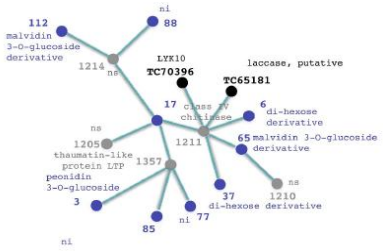


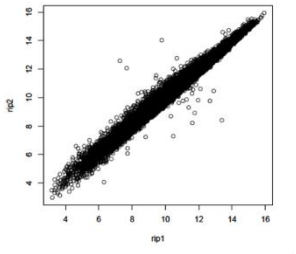
Scanning



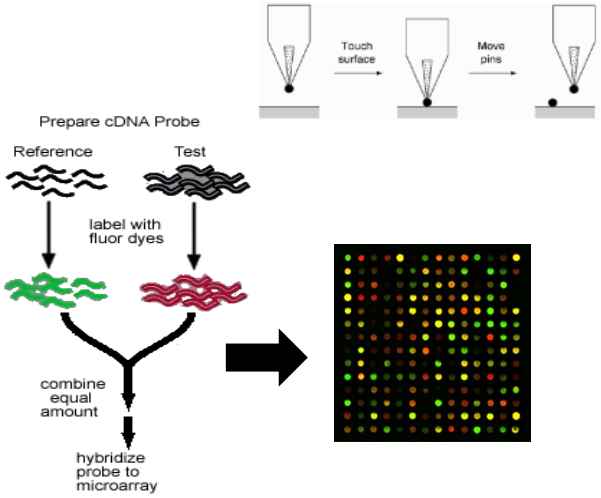


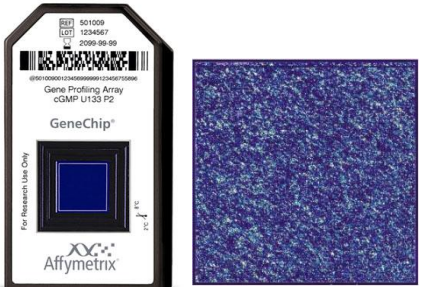
Data analysis

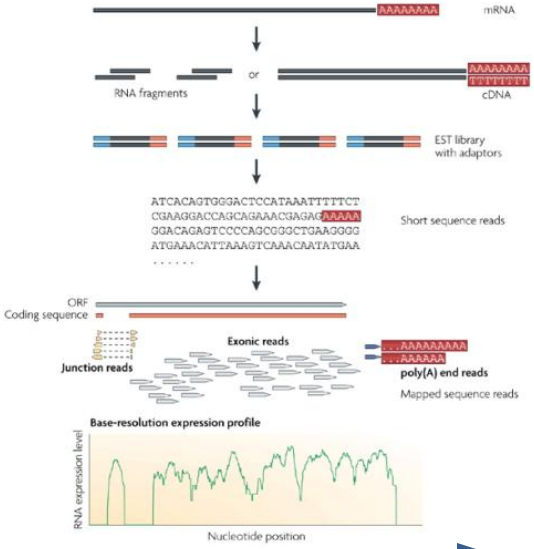




Evoluzione delle tecnologie di analisi del trascrittoma







1995- Sviluppati i primi microarray basati su spotting di molecole di cDNA

2002- High density oligo microarrays

2008- RNA-Seq: sequenziamento dei messaggeri basato su tecnologie NGS

Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray- Schena et. al.

Analisi RNASeq

Campioni di interesse

Tessuto normale **Tessuto tumorale**

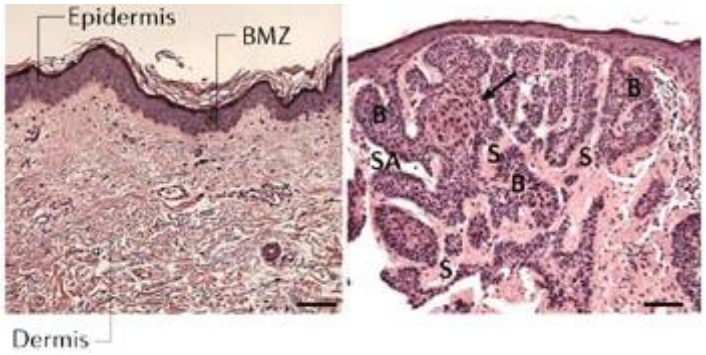
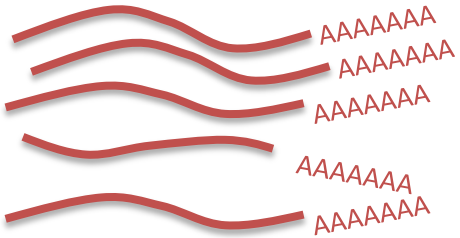
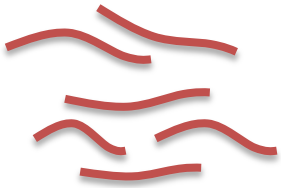


Immagine modificata da:
http://www.nature.com/nrc/journal/v6/n4/full/nrc1838.html

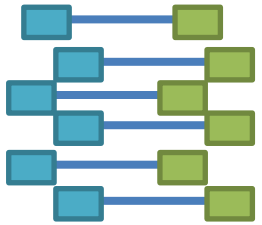
Isolamento dell'RNA/mRNA



Frammentazione chimica

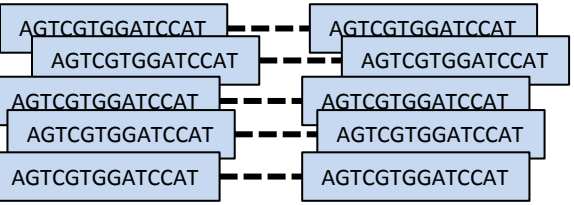


Conversione a cDNA e ligazione degli adattatori



Sequenziamento

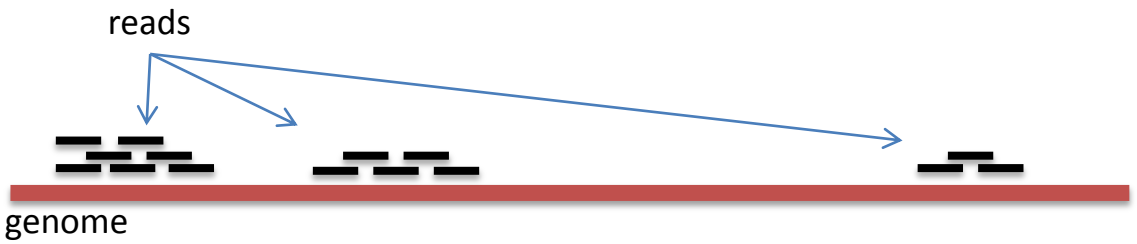
Milioni di sequenze



Principi dell'analisi RNASeq

- **Quanto più un gene è espresso tante più molecole di mRNA saranno presenti nel campione e quindi tante più sequenze verranno generate.**
- **E' quindi possibile utilizzare il numero di sequenze corrispondenti a ciascun gene per ottenere una misura del livello di espressione del gene.**
- I valori di espressione ottenuti dall'RNA-Seq deriva dalla conta diretta delle read che mappano su un gene: **misura digitale**

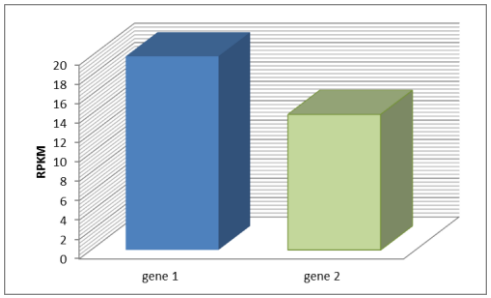
Protocollo di analisi dati RNA-Seq



Allineamento su un genoma di riferimento

Assegnamento delle read ai geni annotati

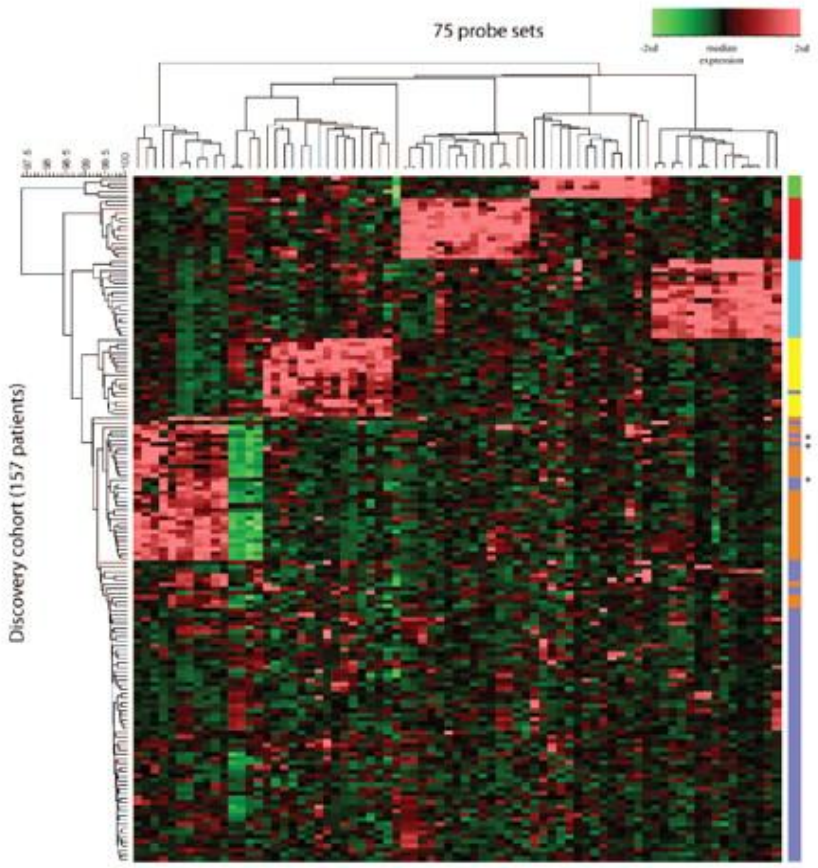
Rilevazione di eventuali geni “nuovi” non annotati



Quantificazione dell’espressione e analisi statistica

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, *5*(7), 621-8. doi: 10.1038/nmeth.1226.

Esempio: classificazione delle malattie sulla base dei pattern di attività genica



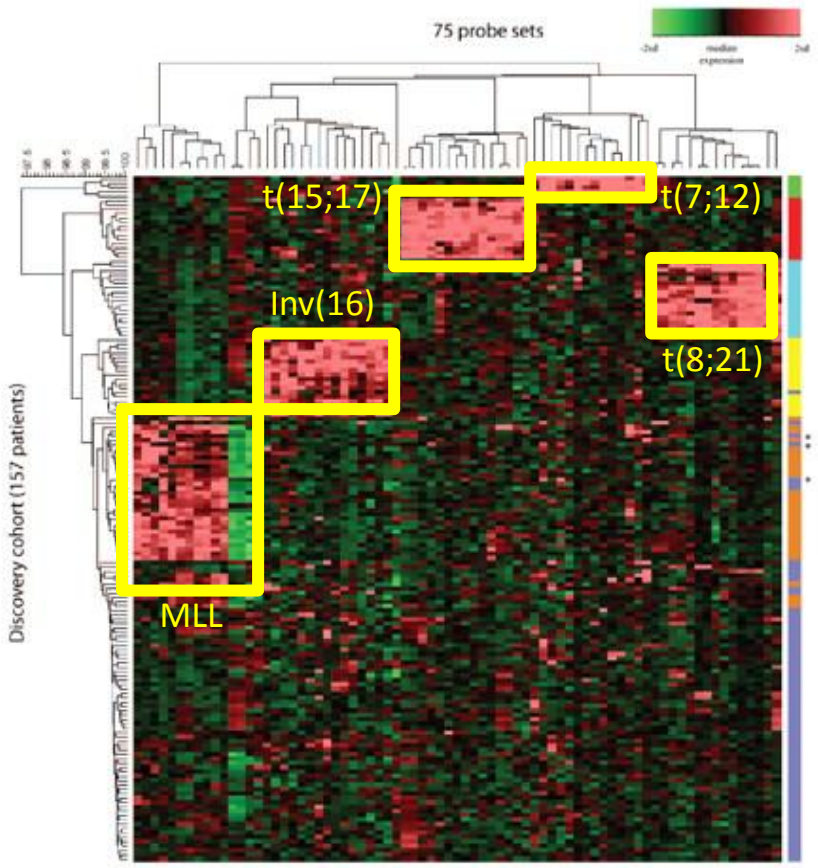
Clustering di dati di espressione di 75 geni (colonne) in 157 pazienti (righe).

Il colore rosso indica un aumento del valore di espressione nel campione tumorale rispetto al normale per quel dato gene in quel paziente.

Il colore verde indica una riduzione del valore di espressione

Brian V. Balgobind et al. Haematologica 2011;96:221-230

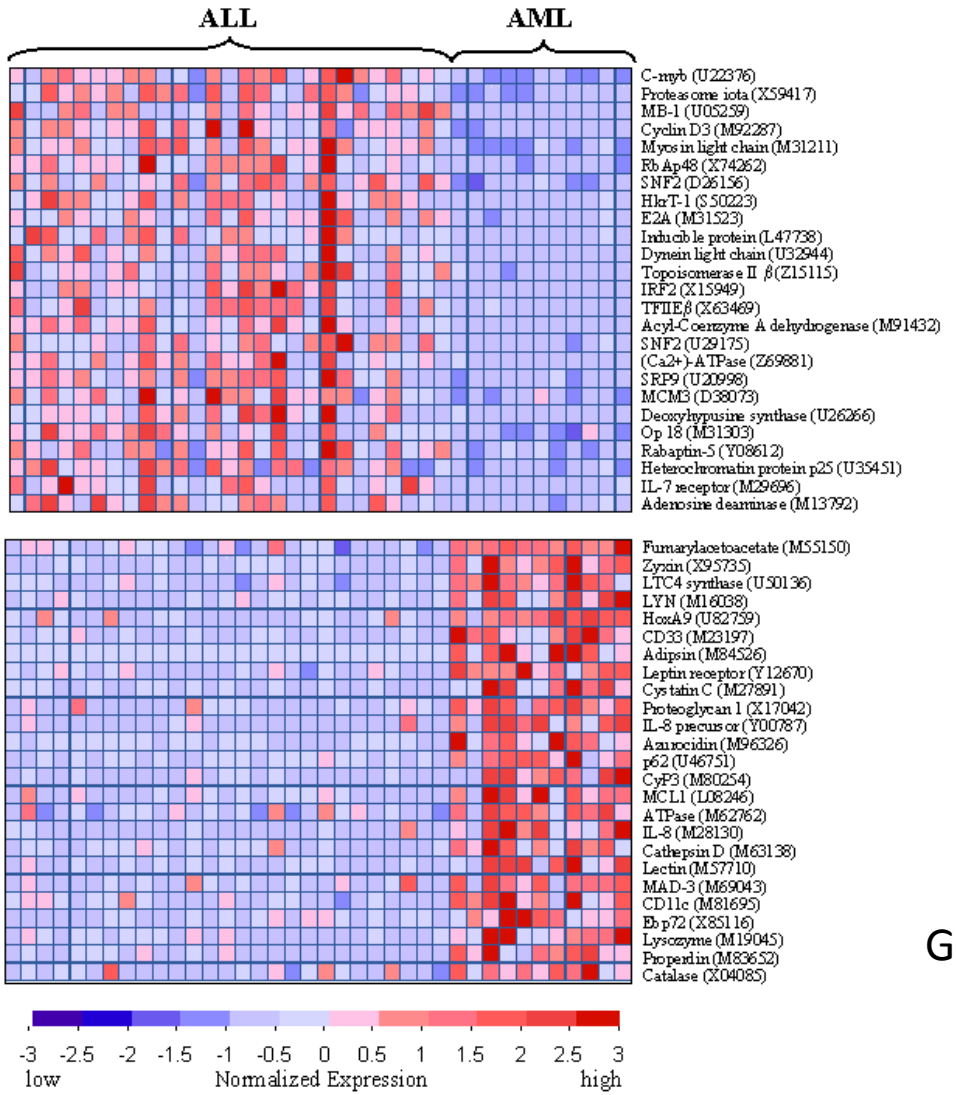
Esempio: classificazione delle malattie sulla base dei pattern di attività genica



Profili di espressione specifici sono associati a diversi gruppi di pazienti con diversi sottotipi di leucemia

Brian V. Balgobind et al. Haematologica 2011;96:221-230

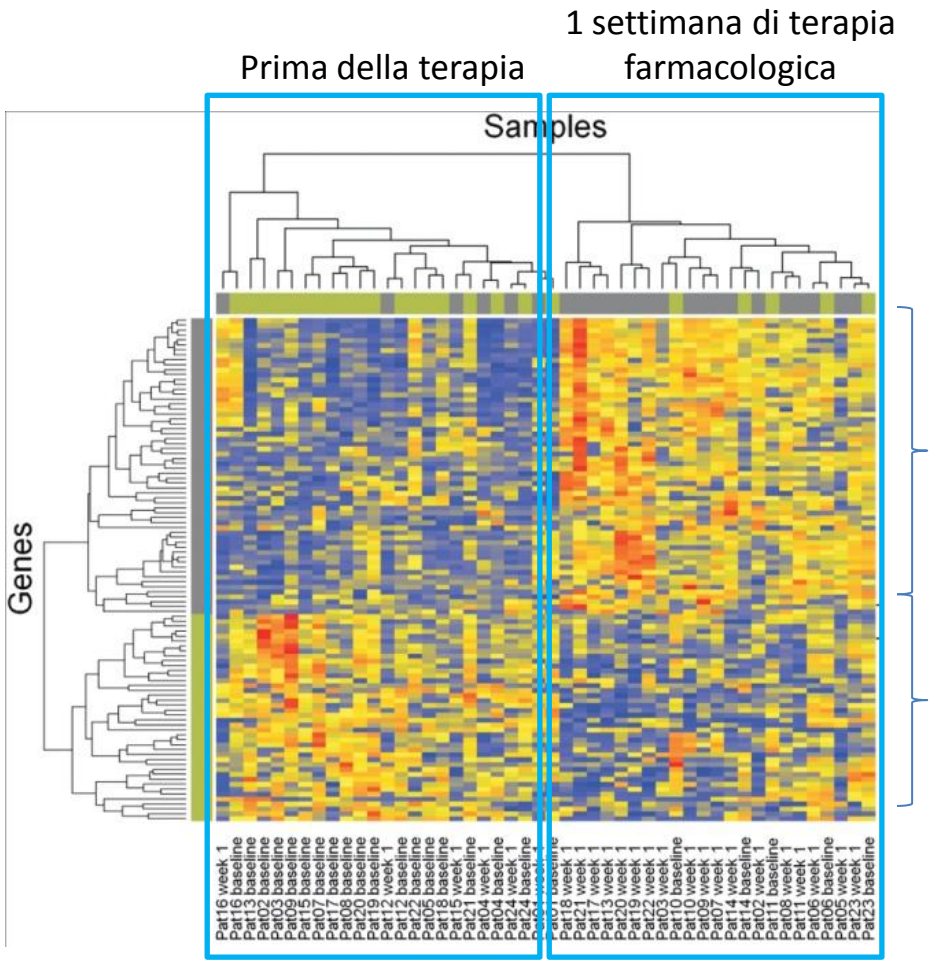
Esempio: classificazione delle malattie sulla base dei pattern di attività genica



Campioni di pazienti affetti da leucemia linfoide acuta e leucemia mieloide acuta possono essere distinti sulla base dei profili di espressione di un set di geni indotti in maniera specifica in ciascun tipo di leucemia

Golub et al., Science 286:531-537. (1999).

Esempio: caratterizzazione della risposta a terapie (farmacogenomica)



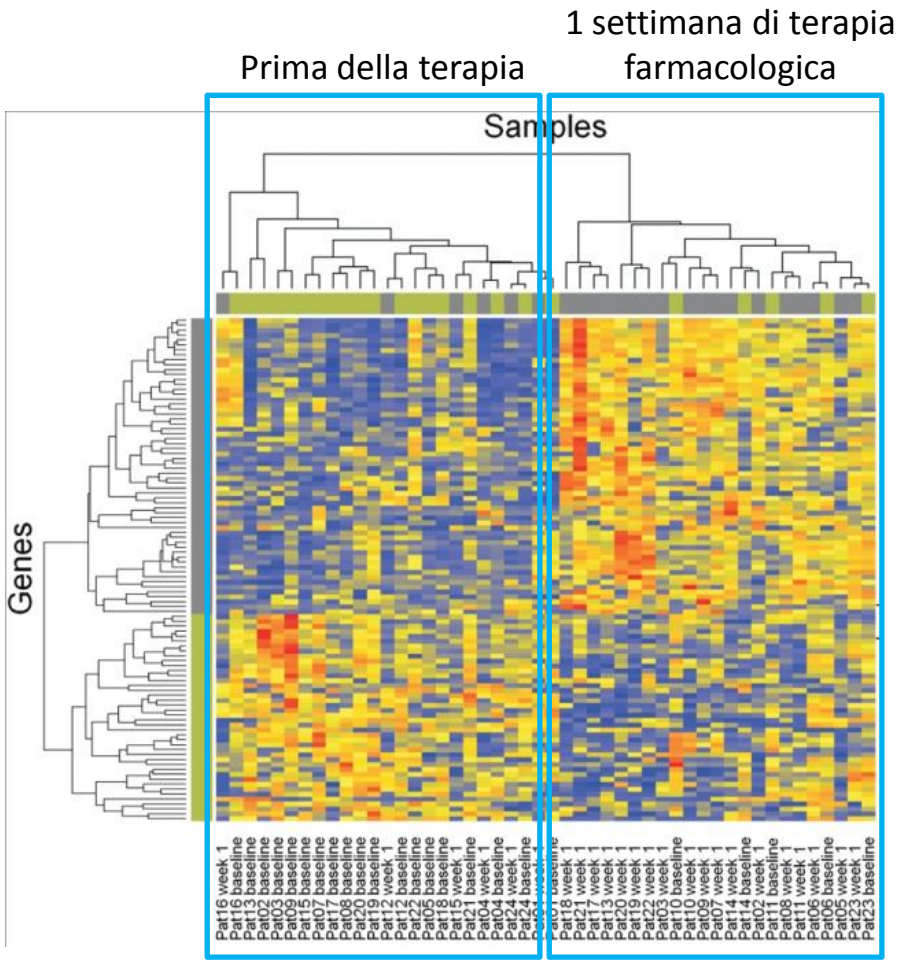
L'heatmap mostra i livelli di espressione di 102 geni indotti o repressi in risposta al trattamento con interferone-B-1a in pazienti malati di sclerosi multipla

Geni indotti dal trattamento

Geni repressi dal trattamento

The Pharmacogenomics Journal (2012) **12,** 134-146;
doi:10.1038/tpj.2010.77

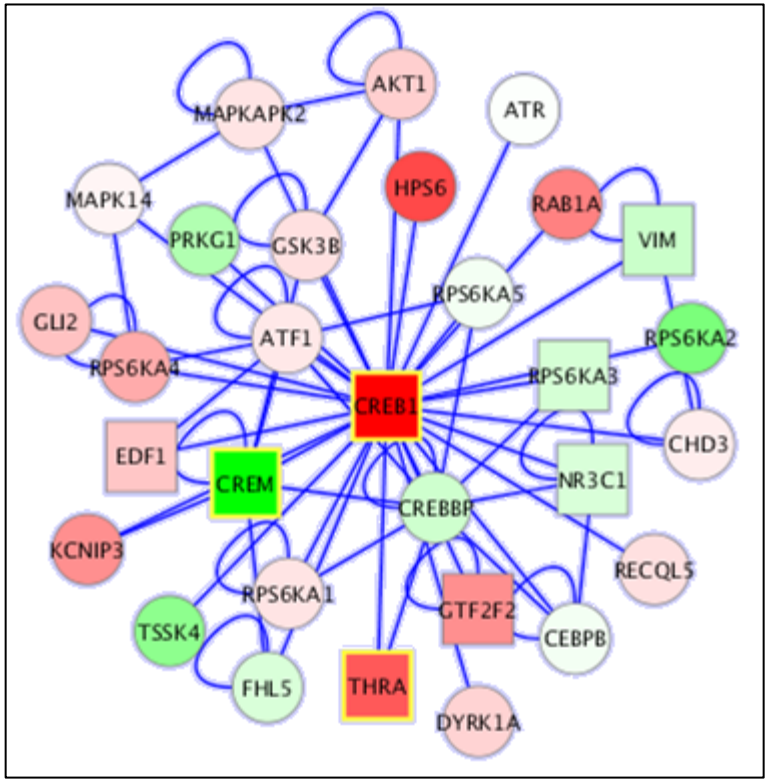
Esempio: caratterizzazione della risposta a terapie (farmacogenomica)



Utile per capire il meccanismo d'azione della terapia ed eventualmente per "prevedere" la risposta ai farmaci di diversi gruppi di pazienti → terapia personalizzata

The Pharmacogenomics Journal (2012) **12,** 134-146;
doi:10.1038/tpj.2010.77

Analisi network di dati di espressione

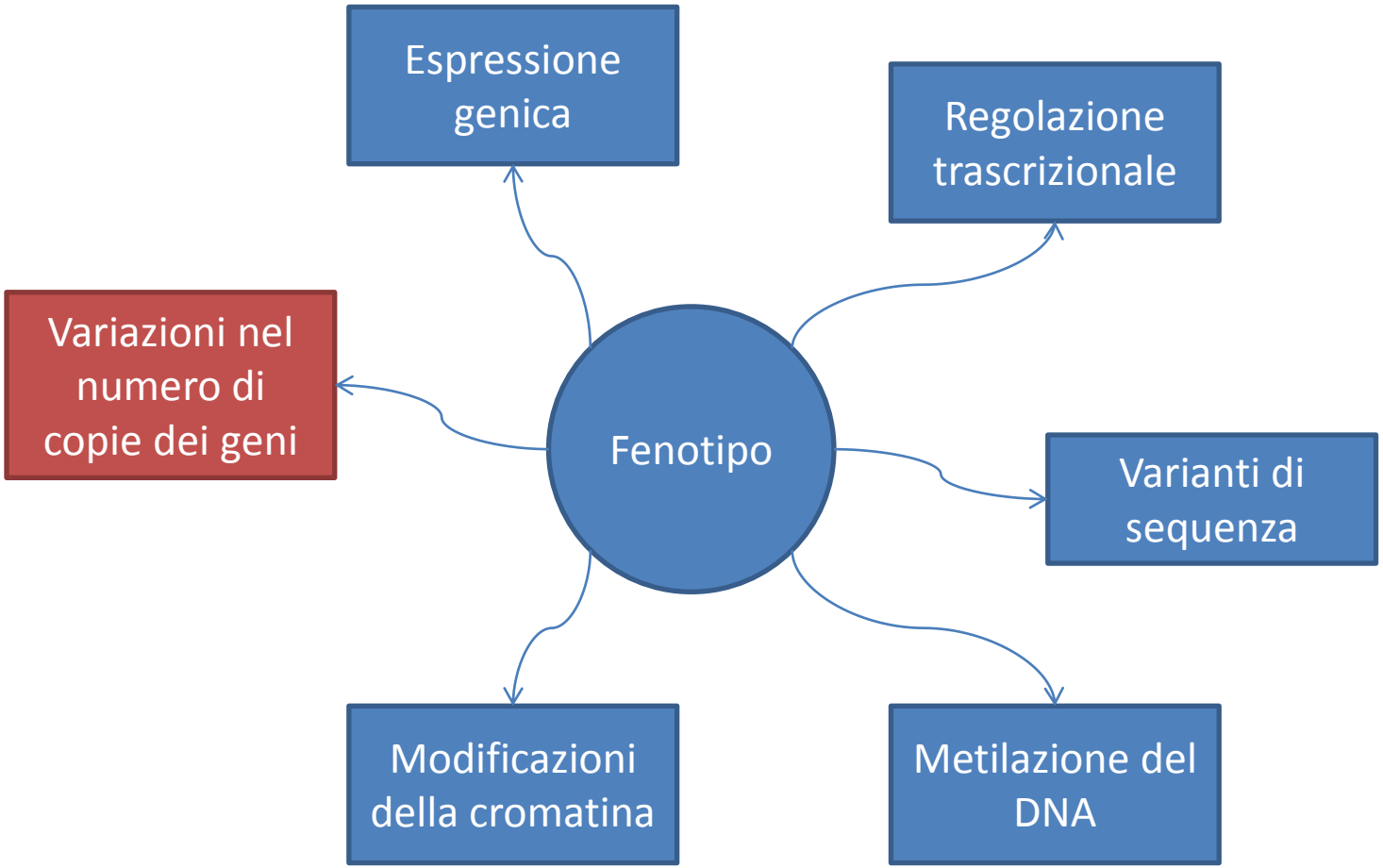


È possibile mappare su un network di interazione proteina-proteina i dati di espressione.

Geni repressi in rosso, geni indotti in verde

Il knockdown del fattore di trascrizione CREB1 in cellule di leucemia mieloide causa l'alterazione dei livelli di espressione di un set di geni che codificano per proteine che interagiscono con CREB1: indica una attività di regolazione da parte di CREB1

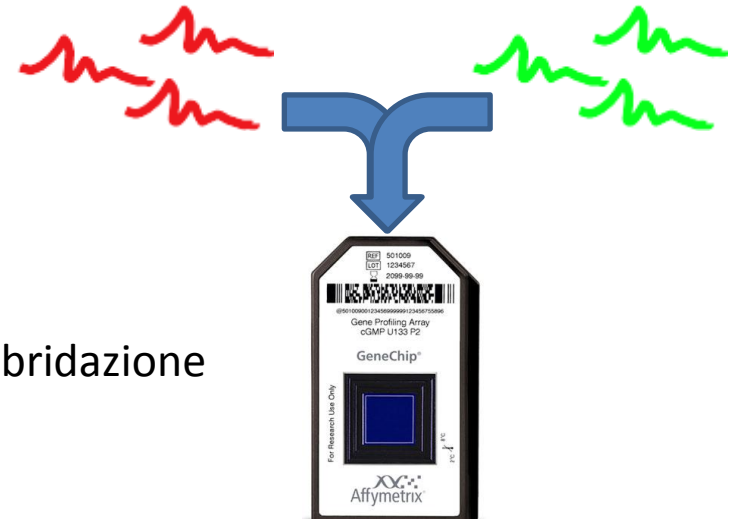
Expression profile of CREB knockdown in myeloid leukemia cells. Matteo Pellegrini,corresponding author1 Jerry C Cheng,2 Jon Voutila,2 Dejah Judelson,2 Julie Taylor,2 Stanley F Nelson,3 and Kathleen M Sakamoto4,5 BMC Cancer. 2008; 8: 264.


Espressione genica
Regolazione trascrizionale
Variazioni nel numero di copie dei geni
Fenotipo
Varianti di sequenza
Modificazioni della cromatina
Metilazione del DNA

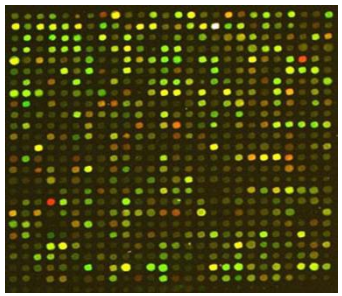
Ibridazione genomica comparativa (CGH)

Campione di DNA tumorale

Campione di DNA normale



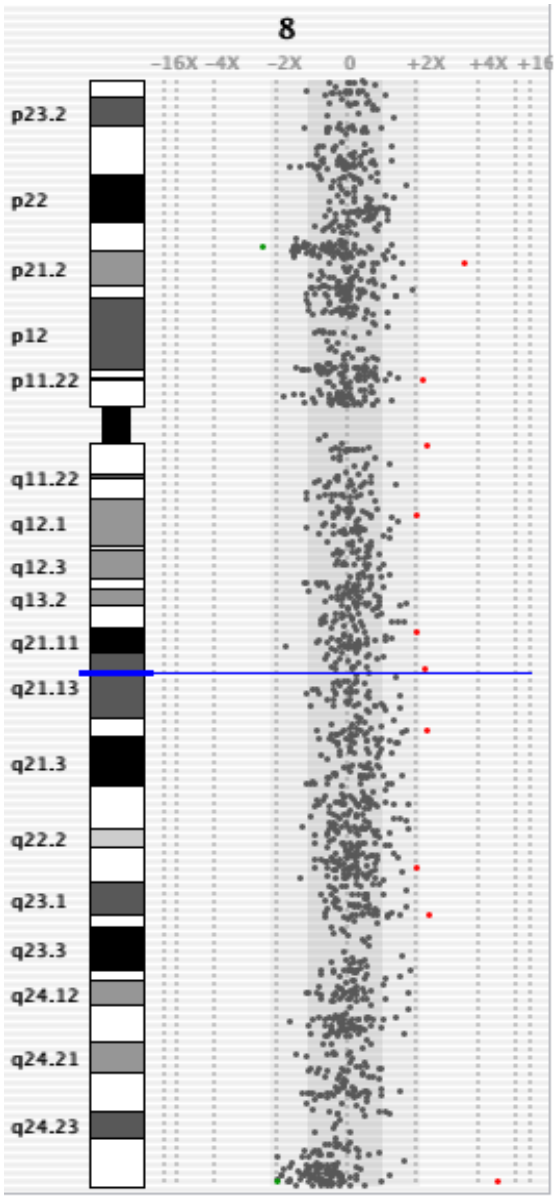
Ibridazione



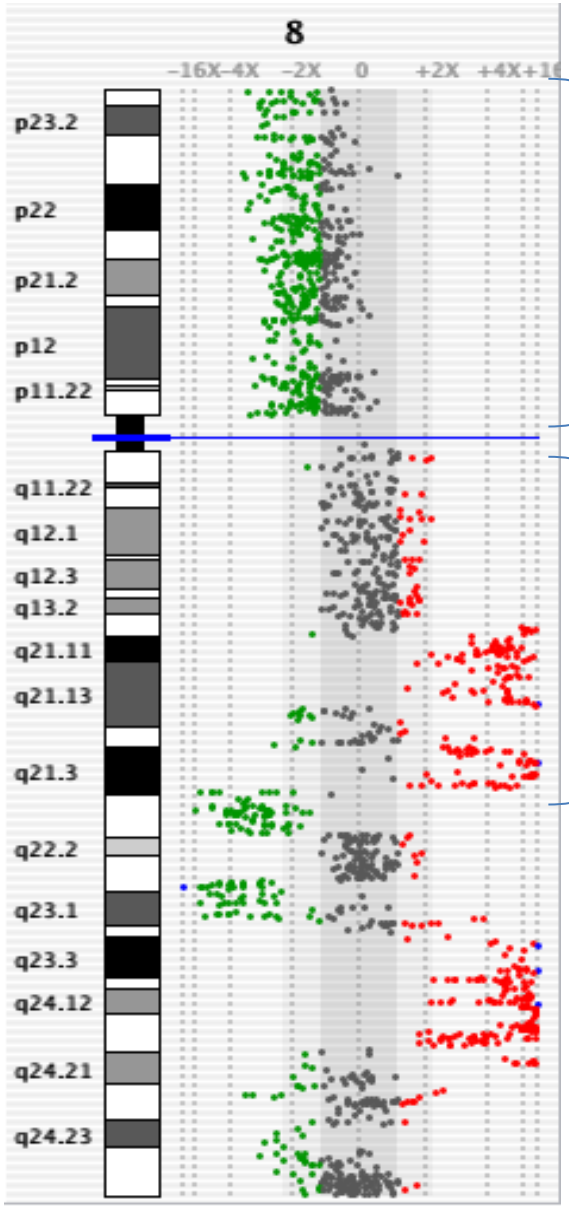
Un campione normale di riferimento e un campione tumorale (o altra condizione) vengono marcati con 2 diversi fluorofori e ibridati su un chip che rappresenta diverse regioni genomiche/geni.

- Amplificazione (oncogene)
- Numero di copie normale
- Perdita di copie (tumor suppressor)

Cromosoma normale

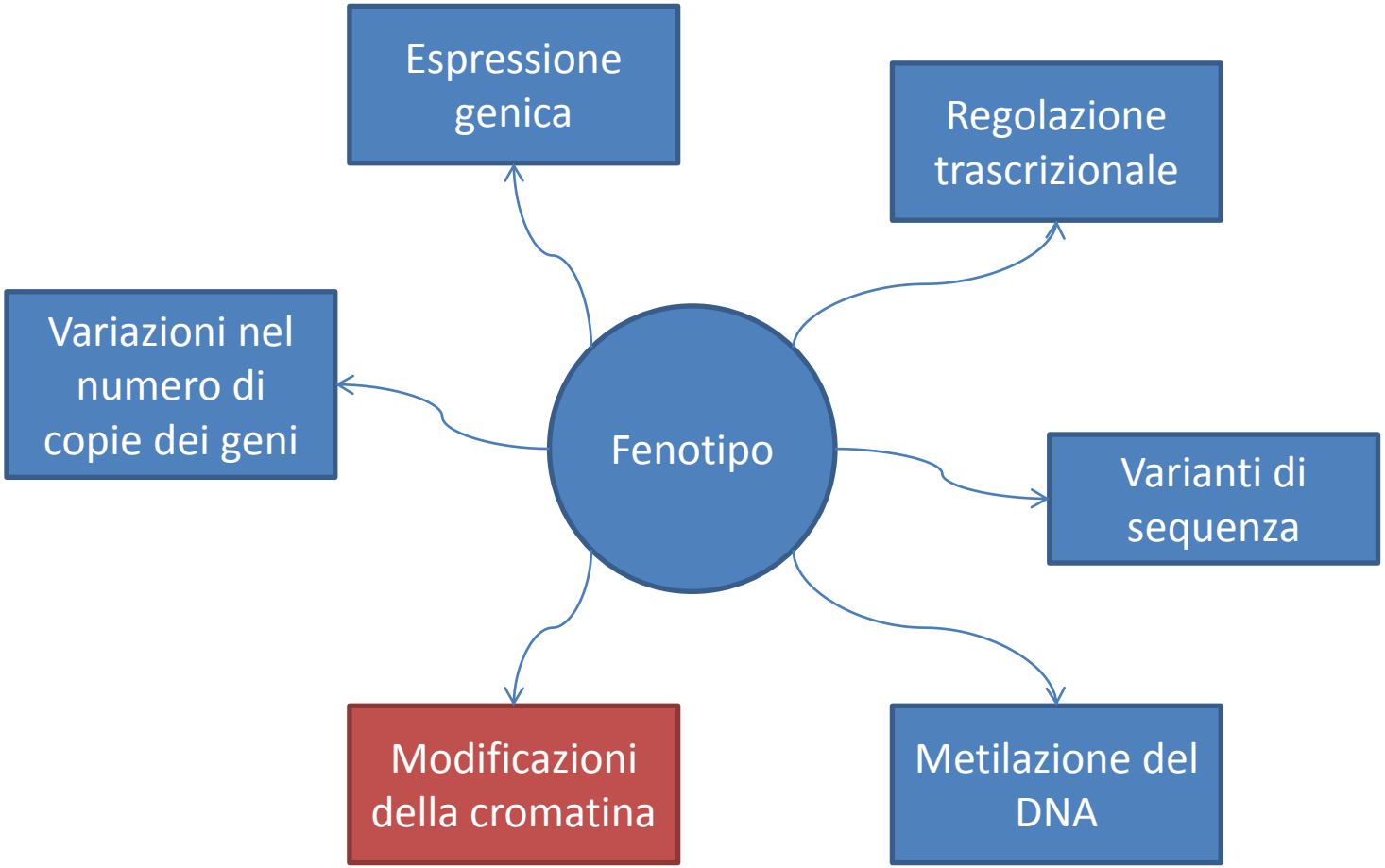


Cromosoma tumorale

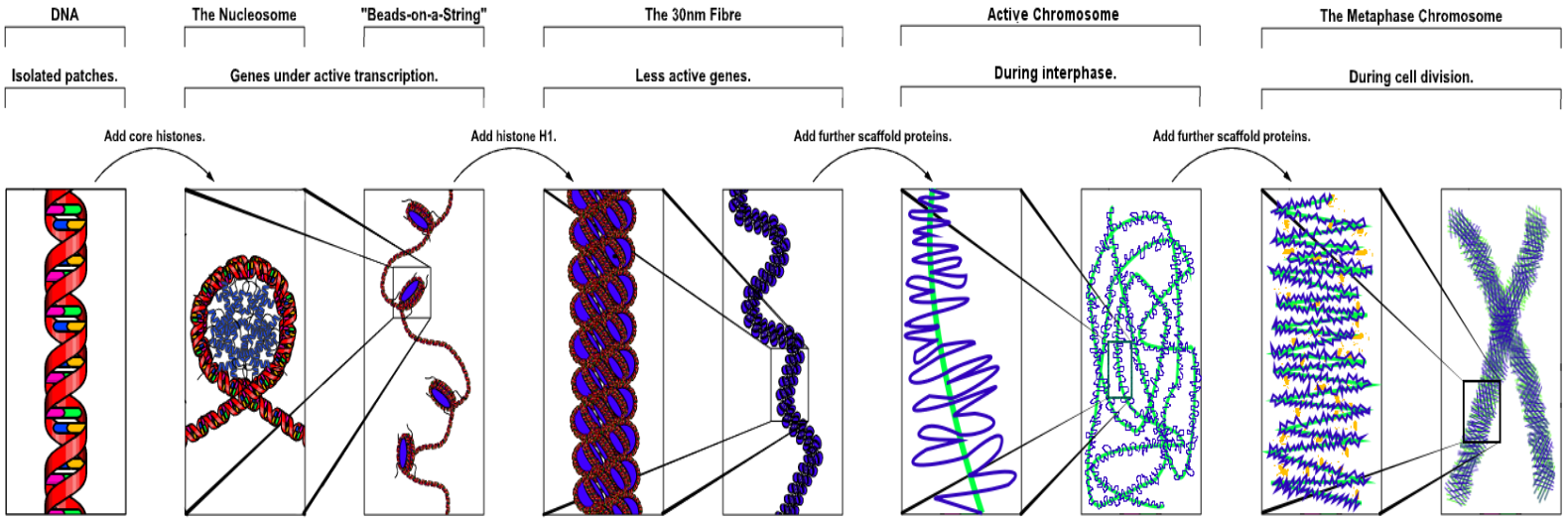


Perdita di questa regione su uno dei 2 cromatidi

Regioni amplificate


Espressione genica
Regolazione trascrizionale
Variazioni nel numero di copie dei geni
Fenotipo
Varianti di sequenza
Modificazioni della cromatina
Metilazione del DNA

Cromatina



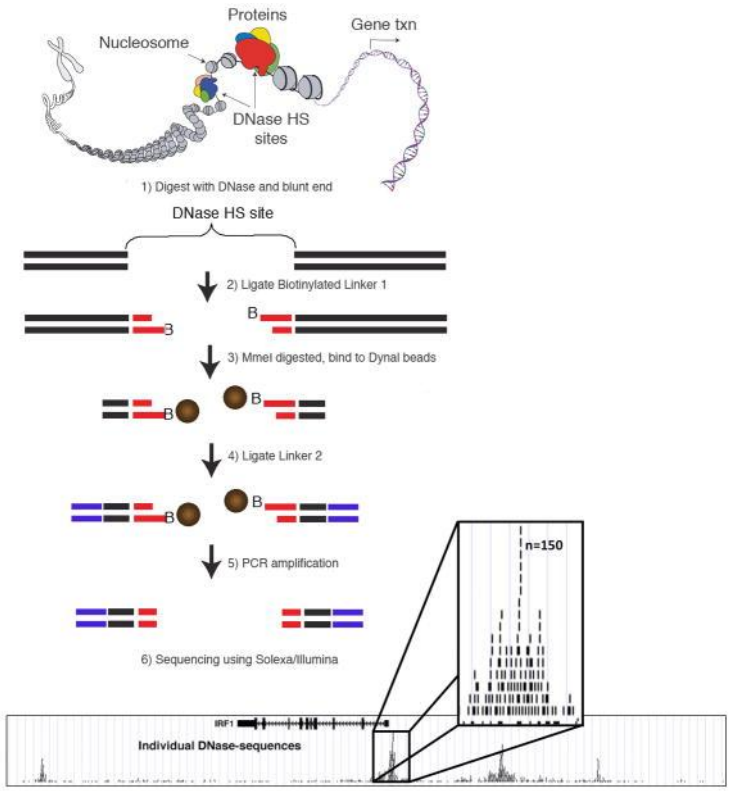
La cromatina è il complesso di proteine e DNA che costituiscono i cromosomi. E' una struttura dinamica.

I nucleosomi sono i blocchi da costruzione della struttura cromatinica.

Modificazioni istoniche e altre modificazioni come la metilazione del DNA modificano il livello di condensazione del DNA e la sua accessibilità regolando quindi anche I livelli di trascrizione

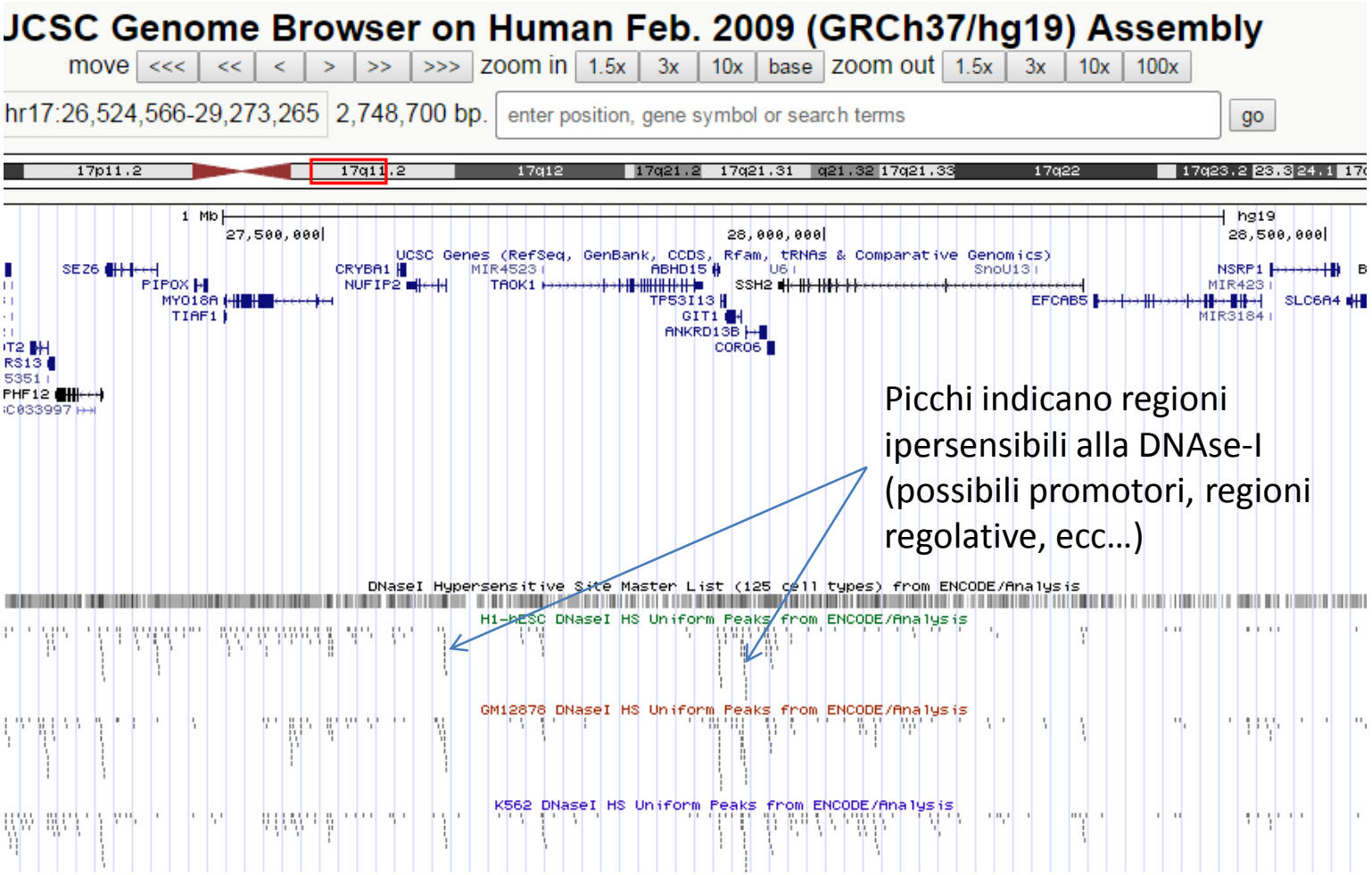
DNAse-Seq

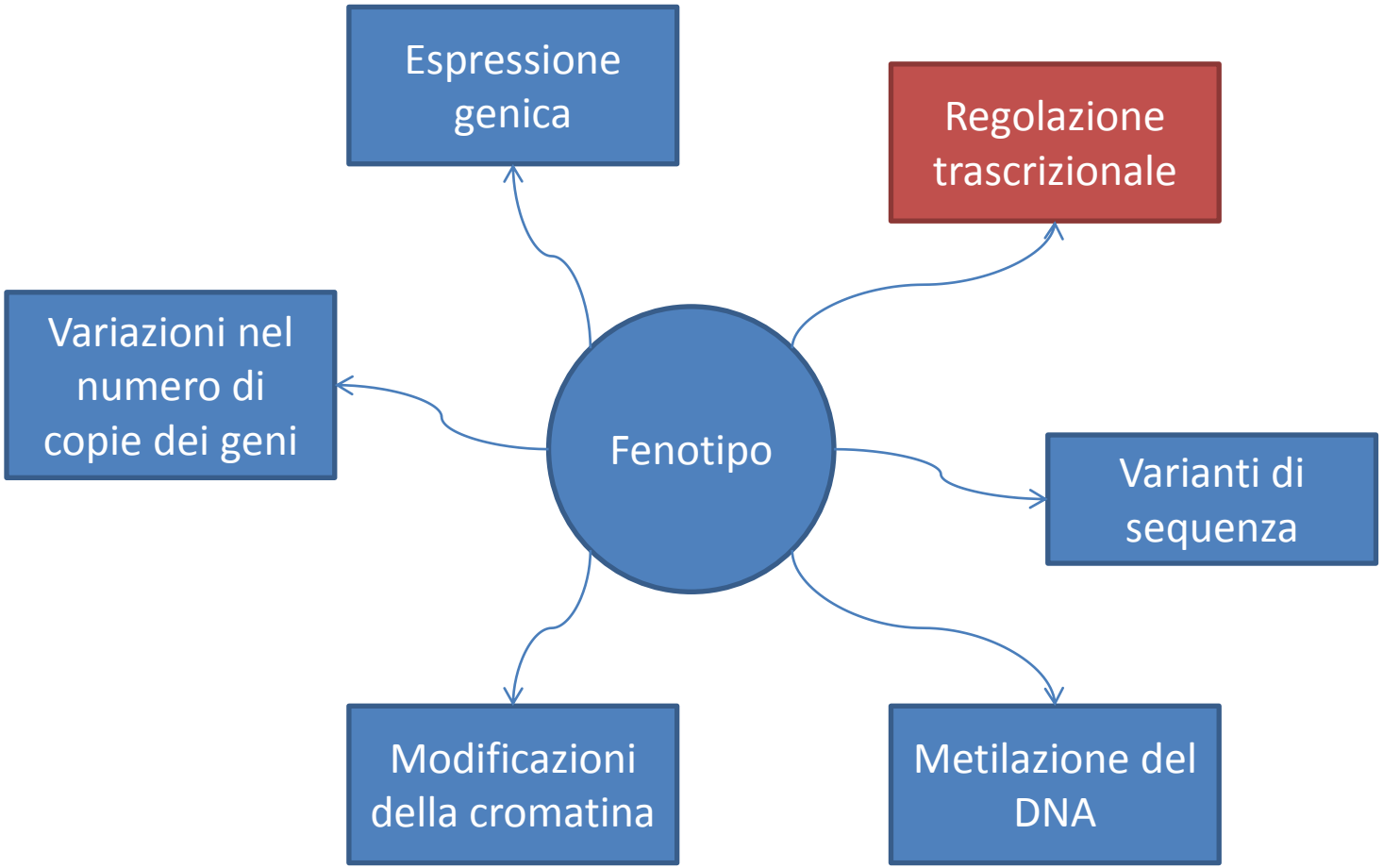
- Basato sul test di ipersensibilità (HS)del DNA alla DNAse I
- Viene sequenziata la regione adiacente ai siti tagliati dalla DNAse I
- Identifica zone accessibili del DNA:
  - Zone regolatrici (promotori, enhancer, silencer…)
  - Siti di inizio della trascrizione



DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells Lingyun Song and Gregory E. Crawford1 Cold Spring Harb Protoc. 2010 Feb; 2010(2): pdb.prot5384.

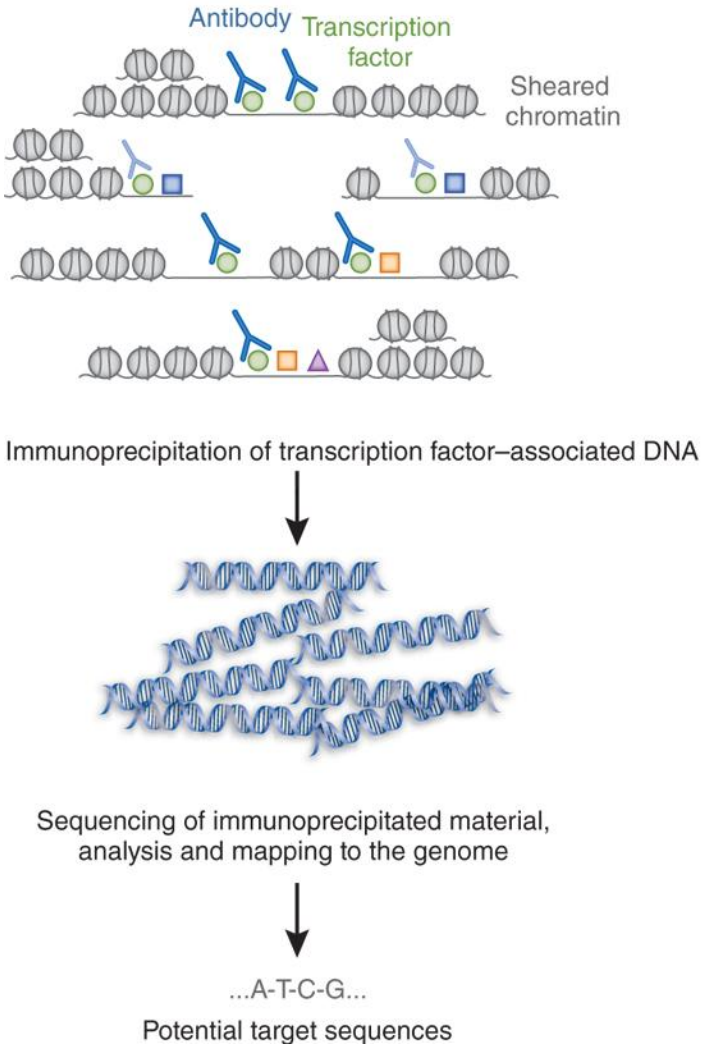
DNAse-Seq


JCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
hr17:26,524,566-29,273,265 2,748,700 bp.
enter position, gene symbol or search terms
go
17p11.2 17q11.2 17q12 17q21.2 17q21.31 q21.32 17q21.33 17q22 17q23.2 23.3 24.1
1 Mb
27,500,000 28,000,000 28,500,000 hg19
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
SEZ6 CRYBA1 MIR4523 ABHD15 U6 SnoU13 NSRP1
PIPOX NUFIP2 TAOK1 SSH2 MIR423
MYO18A TP53I13 EFCAB5 SLC6A4
TIAF1 GIT1 MIR3184
ANKRD13B
CORO6
PHF12
Picchi indicano regioni ipersensibili alla DNAse-I (possibili promotori, regioni regolative, ecc…)
DNaseI Hypersensitive Site Master List (125 cell types) from ENCODE/Analysis
H1-hESC DNaseI HS Uniform Peaks from ENCODE/Analysis
GM12878 DNaseI HS Uniform Peaks from ENCODE/Analysis
K562 DNaseI HS Uniform Peaks from ENCODE/Analysis


Espressione genica
Regolazione trascrizionale
Variazioni nel numero di copie dei geni
Fenotipo
Varianti di sequenza
Modificazioni della cromatina
Metilazione del DNA

Chip-Seq

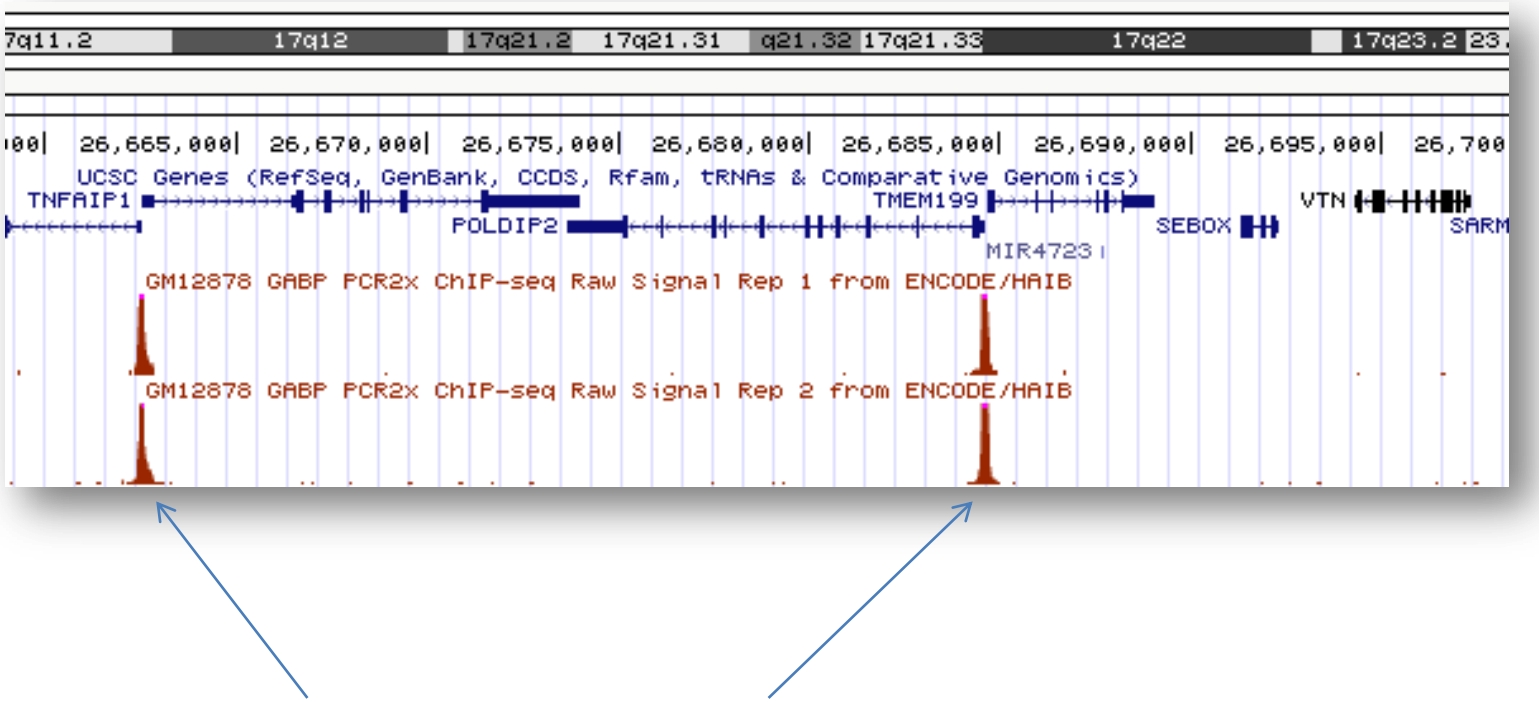
- Anticorpi specifici per fattori di trascrizione di interesse vengono utilizzati per catturare fattori di trascrizione legati al DNA per immunoprecipitazione
- Il DNA immunoprecipitato, corrispondente a siti di legame del fattore di trascrizione viene sequenziato e mappato sul genoma
- Mi permette di identificare sperimentalmente i siti di legamente di un fattore di trascrizione sull'intero genoma



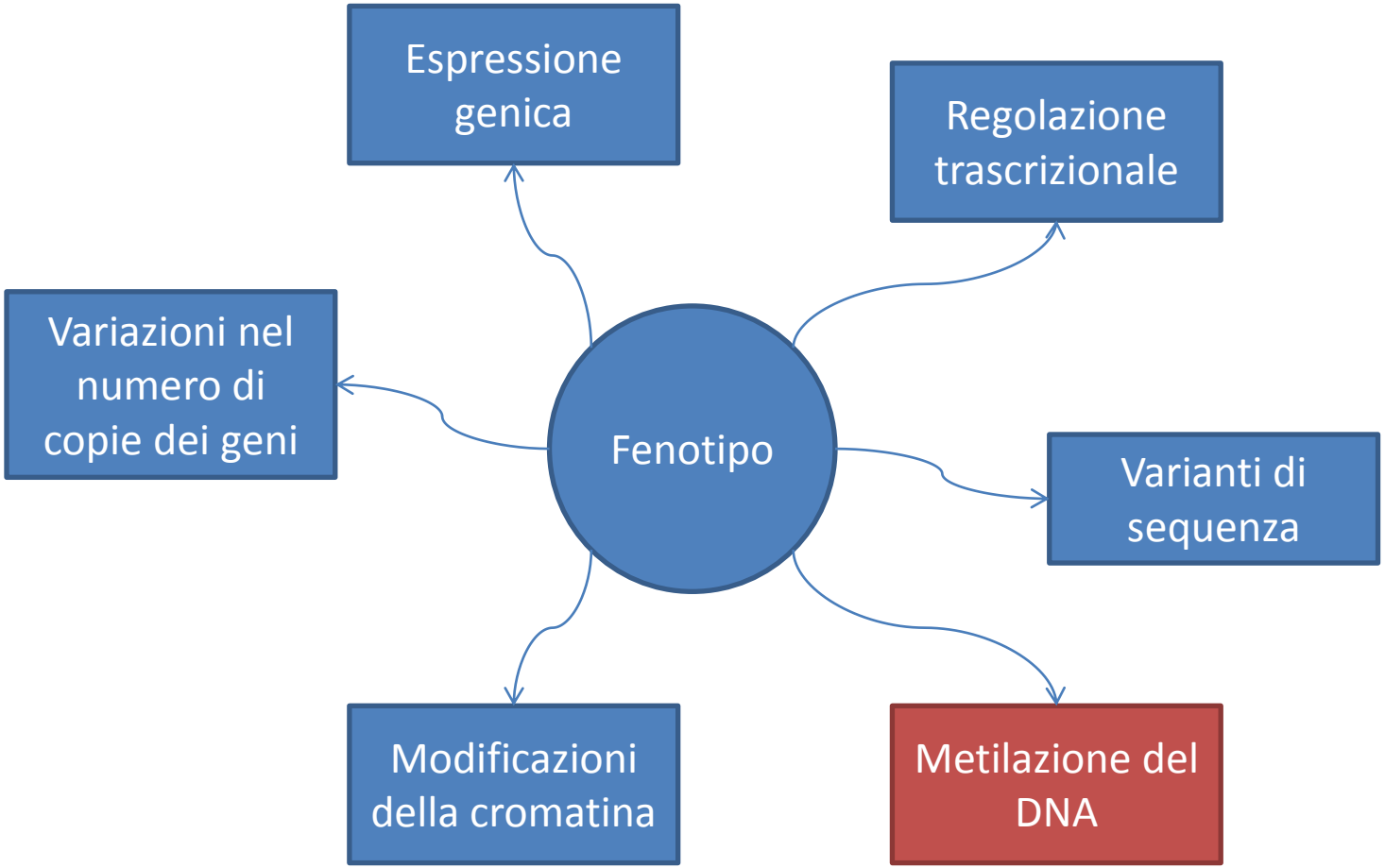
Exciting times: bountiful data to facilitate studies of*cis*-regulatory control. Anil Ozdemir & Angelike Stathopoulos. Nature Methods 8, 1016–1017 (2011) doi:10.1038/nmeth.1795

Chip-Seq

Le sequenze dei frammenti precedentemente legati ai fattori di trascrizione vengono mappate sul genoma per identificare la posizione sul genoma dei siti di legame

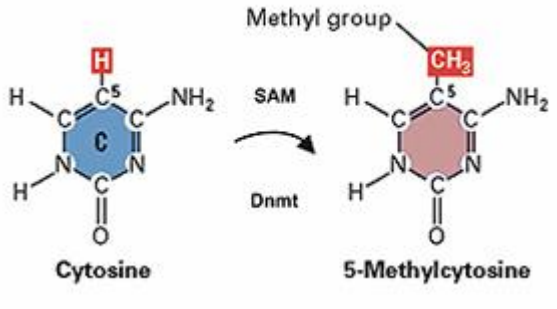


I picchi indicano che il fattore di trascrizione GABP si è legato al DNA a monte dei geni TNFAIP1, POLDIP2, TMEM199


Espressione genica
Regolazione trascrizionale
Variazioni nel numero di copie dei geni
Fenotipo
Varianti di sequenza
Modificazioni della cromatina
Metilazione del DNA

Metilazione del DNA

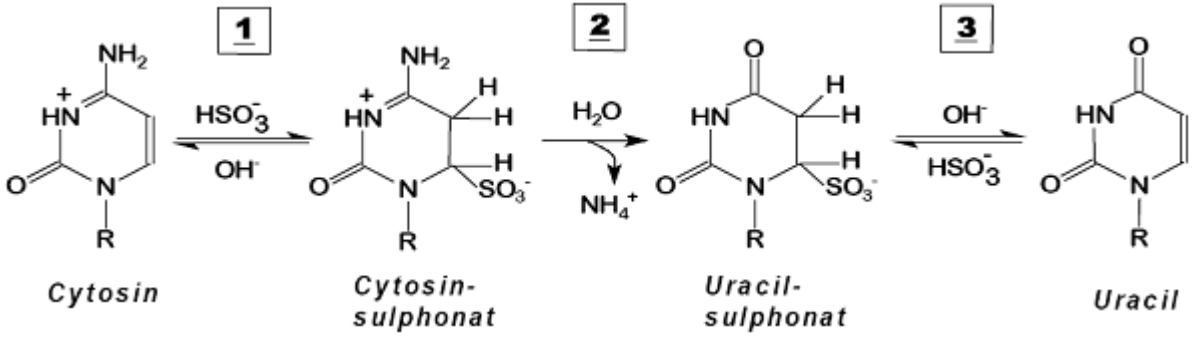
- La metilazione dei residui citosina del DNA alla posizione del carbonio 5 è un marcature epigenetico comune a molti eucarioti e viene spesso osservato nel contesto CpG o CpHpG (H=A, T, C).



- Metilazione di citosine nei promotori è generalmente associata a repressione della trascrizione.
- Metilazione di CpG incrementa nel corpo dei geni in piante e mammiferi.
- Transposoni non espressi di piante mostrano metilazione CpHpG
- In batteri: 5meC ma anche N-4-metilcitosina e N-6-metiladenina
- La metilazione avviene ad opera di **metiltransferasi**.

Conversione con il bisolfito

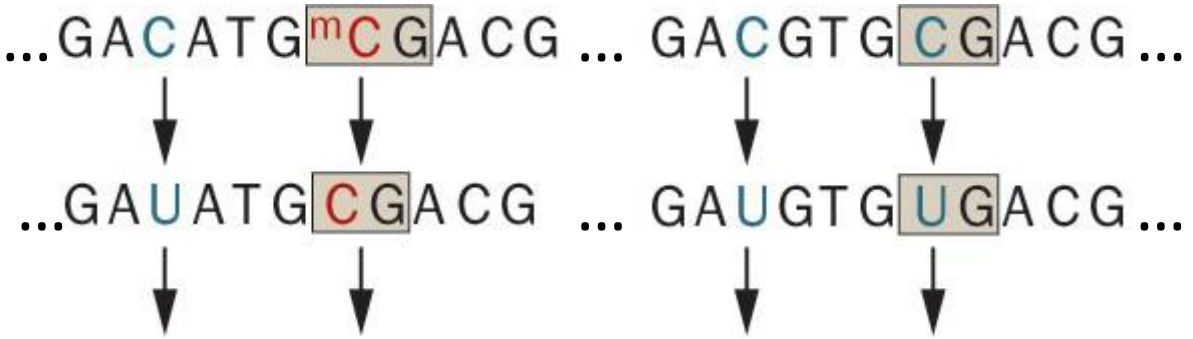
- Basato sul fatto che il sodio bisolfito deamina chimicamente molto più rapidamente i residui di citosine non metilate rispetto alle citosine metilate causando una conversione da C a U.
- Produce informazione sulla metilazione del DNA con una risoluzione di una singola base.



Methyl-Seq

DNA genomico

DNA convertito



Le citosine metilate rimangono citosine mentre le citosine non metilate vengono convertite ad uracile (viene letto come una timina T dal sequenziatore)

Sequenziamento

Le sequenze vengono allineate sul genoma di riferimento. Se ho una sostituzione da C a T significa che quella citosina non era metilata, se invece non ho sostituzione la citosina era originariamente metilata

Sequenza del DNA convertito ...GAUATGCGA...GATGTGTGACG...

Genoma di riferimento ...GAUATGCGA...GACGTGCGACG...