Lezioni di genetica medica

Specializzazioni

2014

Programma, parte 1
genetica umana generale

- **Il genoma umano: geni ed organizzazione**
- **Next generation sequencing (NGS), l'exoma**
- **Eterogeneità clinica ed eterogeneità genetica**
- **Penetranza ed espressività, anticipazione**
- **Omozigosità ed eterozigosità composta**
- **Aploinsufficienza**
- **Meccanismo dello splicing e sue alterazioni**
- **Classi di mutazioni puntiformi, transizione e trasversione, conservative, missenso, nonsenso, nonstop**
- **Inserzioni, delezioni con frame-shift e non, duplicazioni, conversione genica**
- **Significato patologico delle varie classi di variazioni del DNA: alleli equivalente, amorfo, ipomorfo, ipermorfo, neomorfo e antimorfo**
- **Nomenclatura delle variazioni genetiche e refertazione**

Programma, parte 2:
la consulenza e le cromosopatie

- **La consulenza genetica: rischio riproduttivo dipendente ed indipendente dal partner**
- **Diagnostica prenatale e presintomatica**
- **L'analisi del cariotipo e i bandeggi, la FISH**
- **Cariotipo molecolare mediante arrayCGH**
- **Aneuploidie negli aborti e rischio di ricorrenza**
- **Triploidia da doppio corredo paterno o materno, tetraploidia**
- **Il cromosoma X e la sua inattivazione, regioni PAR**
- **Trisomie autosomiche e dei cromosomi sessuali**
- **Le monosomie, la sindrome di Turner**
- **Delezioni cromosomiche, inversioni paracentriche e pericentriche**
- **Traslocazioni sbilanciate e bilanciate, robertsoniane, markers cromosomici**
- **Delezioni e duplicazioni submicroscopiche (s. di Williams, s. di DiGeorge, s. Cri du Chat)**

Programma, parte 3
genetica medica mendeliana

- Malattie mendeliane monoalleliche con mutazioni *de novo* (craniosonostosi, acondroplasia)
- Malattie mendeliane monoalleliche a trasmissione autosomica dominante (neurofibromatosi, s. di Marfan, rene policistico, osteogenesi imperfetta)
- Malattie mendeliane monoalleliche legate al cromosoma X (distrofia muscolare di Duchenne e Becker, emofilia, ritardi mentali legati all'X)
- Malattie mendeliane bialleliche a trasmissione autosomica recessiva (fibrosi cistica, alfa e beta talassemia, amiotrofia spinale, emocromatosi, glicogenosi)

Programma, parte 4, eccezioni all'eredità mendeliana

- Mutazioni dinamiche in regioni non codificanti (X-fragile, distrofia miotonica) e codificanti (corea di Huntington, atassie spino-cerebellari)
- Mutazioni in regioni cromosomiche con imprinting (Prader-Willi, sindrome di Angelman, Beckwith-Wiedemann, Silver-Russel) disomia uniparentale
- Mutazioni del DNA mitocondriale (MERFF, MELAS, LHON, KS, s. di Leigh)
- Predisposizione genetica
- Caratteri multifattoriali
- studi GWAs

Testi consigliati

- Neri-Genuardi
  Genetica umana e medica
  Editore Elsevier Masson
- Moncharmont-Leonardi
  Patologia Generale
  Editore Idelson Gnocchi
- Strachan-Read
  Genetica Molecolare Umana
  Editore Zanichelli
- Lewis
  Genetica Umana
  Editore Piccin
- Sito web **http://www.vincenzonigro.it** **(glossario)**

1953

2001

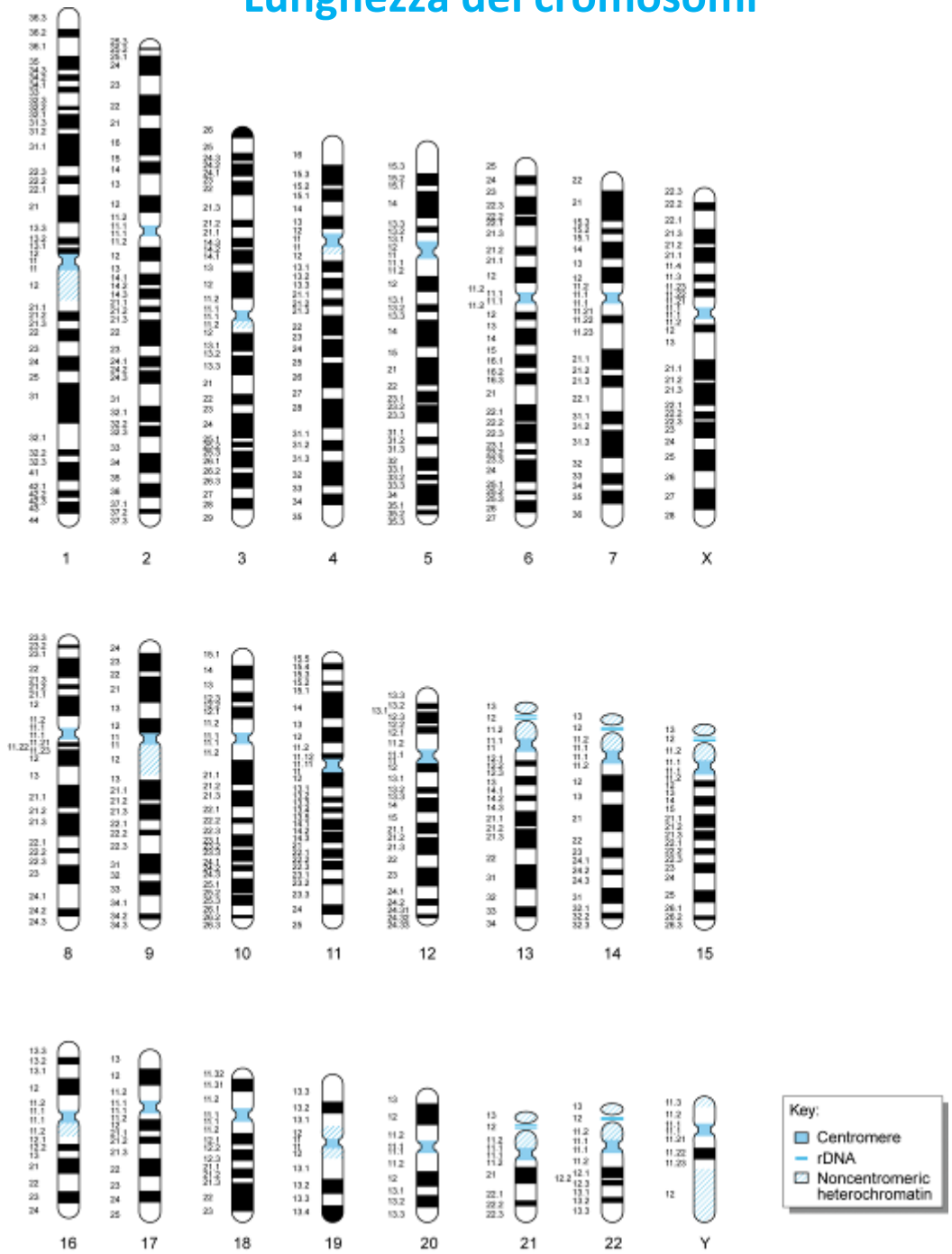
- **Il genoma** è l'intero patrimonio genetico di un organismo vivente
- Il genoma è "scritto" in un composto chimico chiamato DNA (DeoxyriboNucleic Acid, acido desossiribonucleico) a cui si aggiungono informazioni epigenetiche
- La grandezza totale del genoma umano aploide è di 3.070.000.000 basi di cui 2.843.000.000 sono di eucromatina
- Il DNA è identico per tutte le cellule di un individuo ed è contenuto quasi tutto nel nucleo, con l’eccezione del DNA mitocondriale

Counting on the Genome

3.2 billion	Approximate number of chemical nucleotide base pairs in the human genome
2.7 billion	Cost of Human Genome Project, in U.S. dollars
120 million	Estimated cost of 1000 Genomes Project, in U.S. dollars
3.3 million	Estimated number of SNPs per human genome
2.4 million	Number of base pairs in largest known human gene, dystrophin
23 500	Estimated number of human genes
10 000	Approximate current cost, in U.S. dollars, of sequencing one person's genome
2500	Number of genomes targeted for sequencing by 1000 Genomes Project
1000 or fewer	Expected cost in near future, in U.S. dollars, of sequencing one human genome

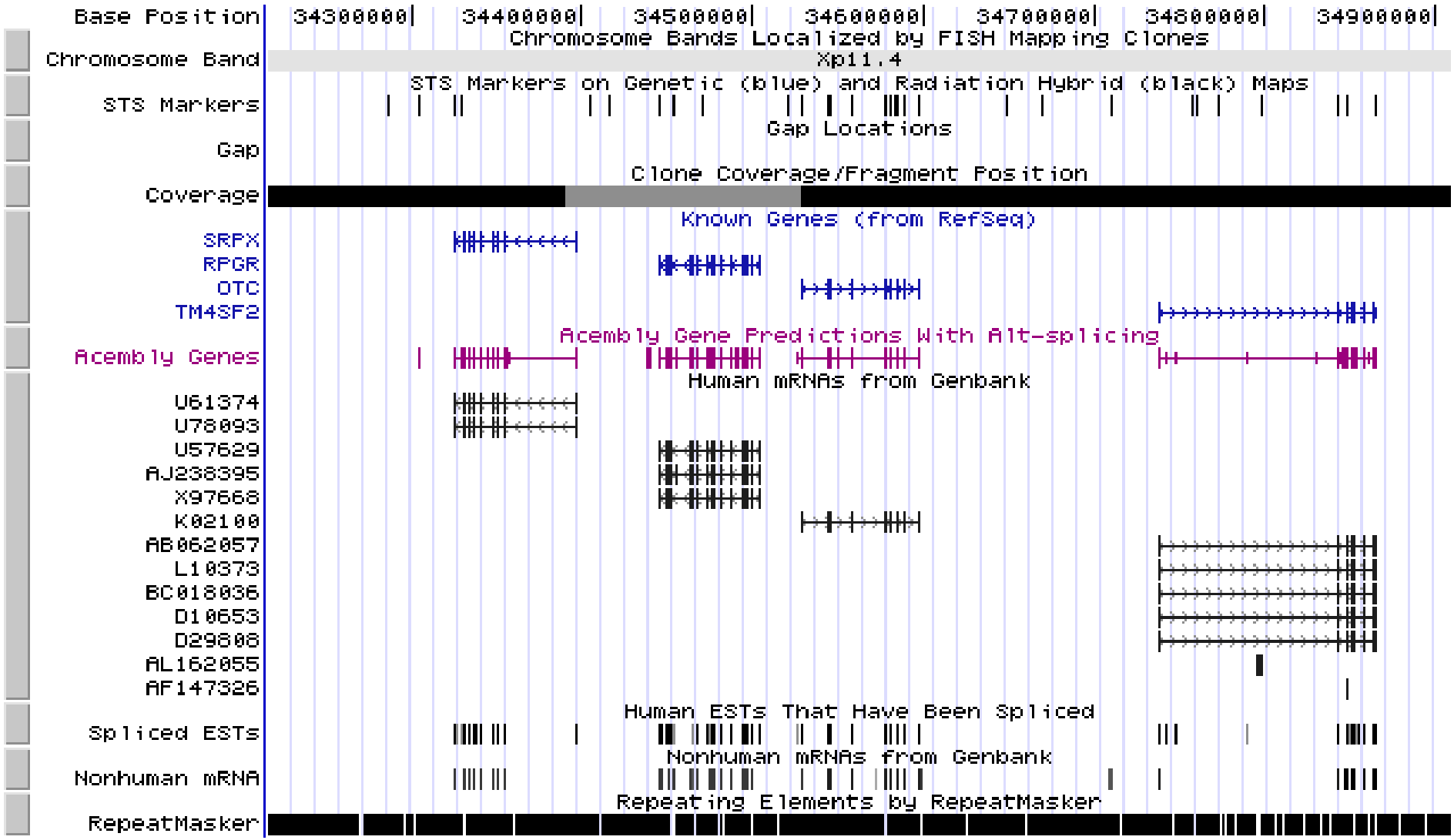
SNPs indicates single-nucleotide polymorphisms.

Lunghezza dei cromosomi

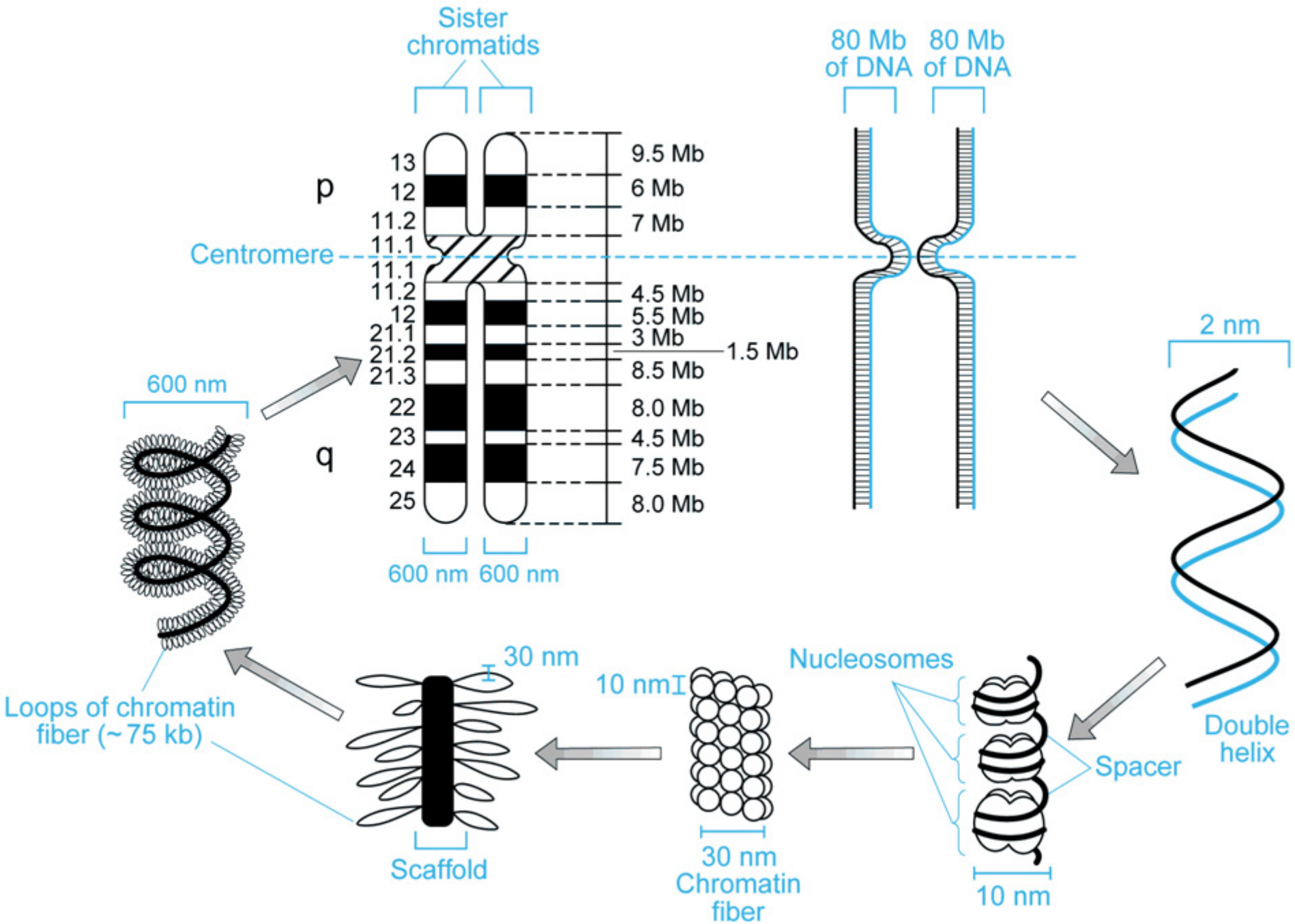


	totale	eucromatina
1	245,203,898	218,712,898
2	243,315,028	237,043,673
3	199,411,731	193,607,218
4	191,610,523	186,580,523
5	180,967,295	177,524,972
6	170,740,541	166,880,540
7	158,431,299	154,546,299
8	145,908,738	141,694,337
9	134,505,819	115,187,714
10	135,480,874	130,710,865
11	134,978,784	130,709,420
12	133,464,434	129,328,332
13	114,151,656	95,511,656
14	105,311,216	87,191,216
15	100,114,055	81,117,055
16	89,995,999	79,890,791
17	81,691,216	77,480,855
18	77,753,510	74,534,531
19	63,790,860	55,780,860
20	63,644,868	59,424,990
21	46,976,537	33,924,742
22	49,476,972	34,352,051
X	152,634,166	147,686,664
Y	50,961,097	22,761,097

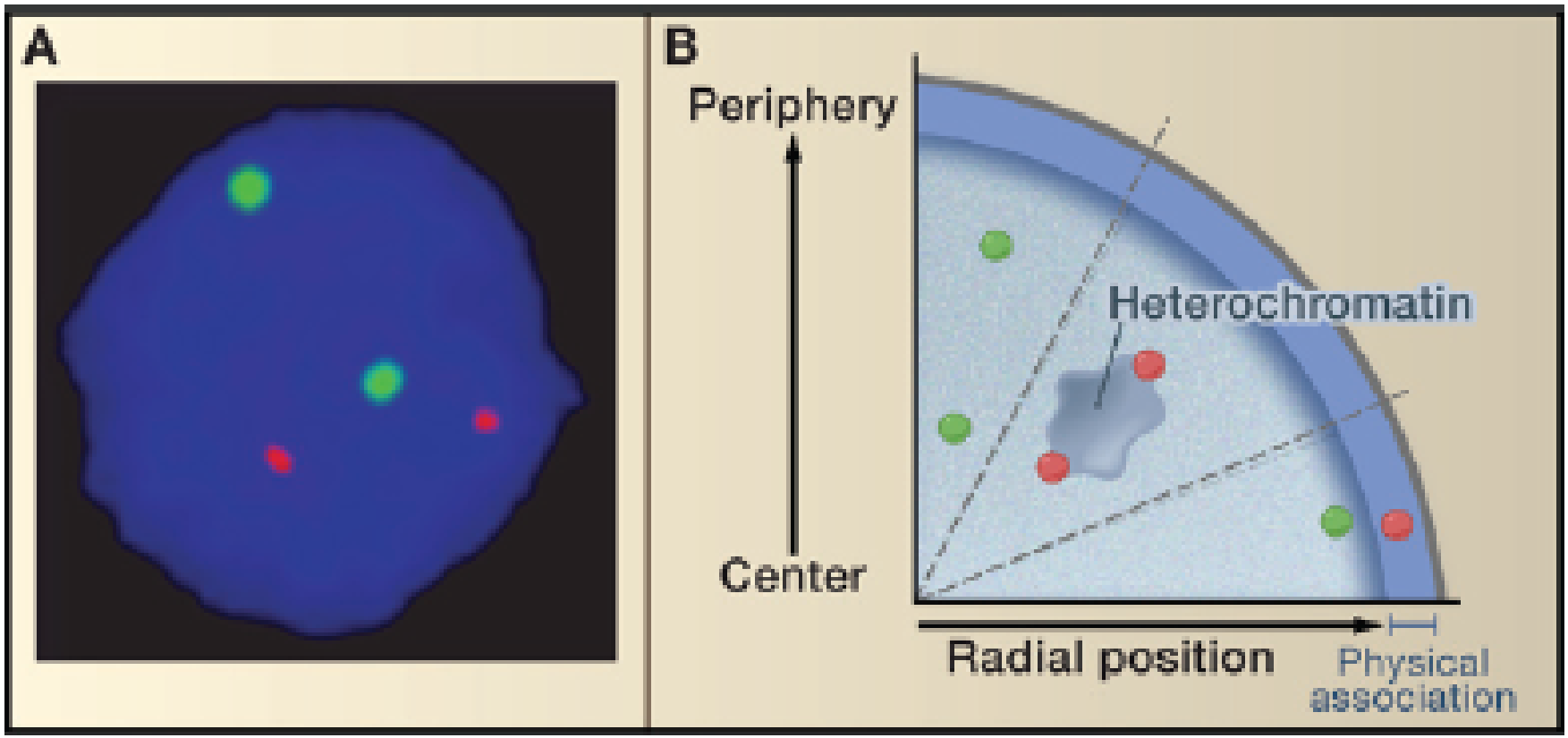
UCSC Genome Browser



Screenshot from University of California at Santa Cruz
http://genome.ucsc.edu

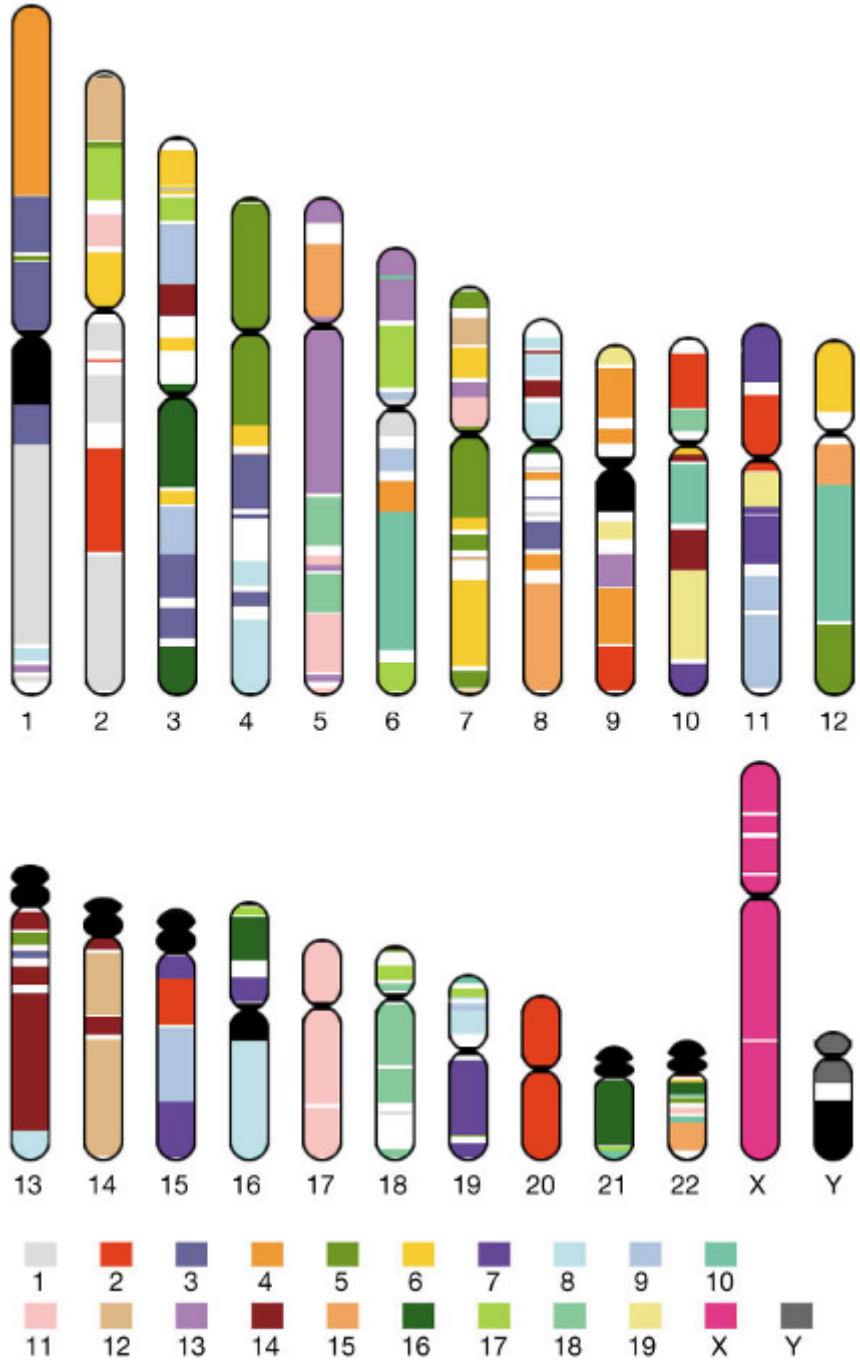

Sister chromatids
80 Mb of DNA
80 Mb of DNA
p
q
13
12
11.2
11.1
11.1
11.2
12
21.1
21.2
21.3
22
23
24
25
Centromere
9.5 Mb
6 Mb
7 Mb
4.5 Mb
5.5 Mb
3 Mb
1.5 Mb
8.5 Mb
8.0 Mb
4.5 Mb
7.5 Mb
8.0 Mb
600 nm
600 nm
600 nm
2 nm
Double helix
Spacer
Nucleosomes
10 nm
30 nm
Chromatin fiber
30 nm
Scaffold
Loops of chromatin fiber (~75 kb)

La distribuzione dei geni nel nucleo non è casuale

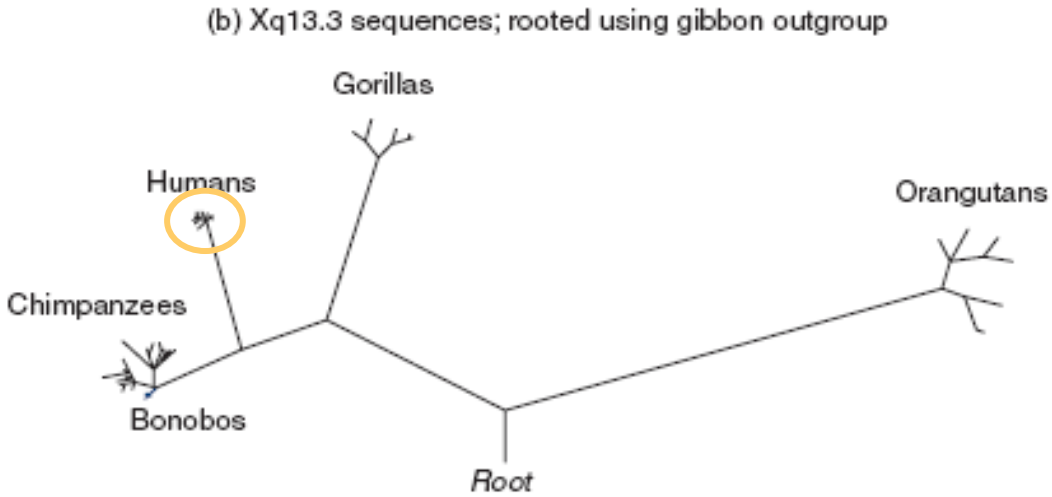


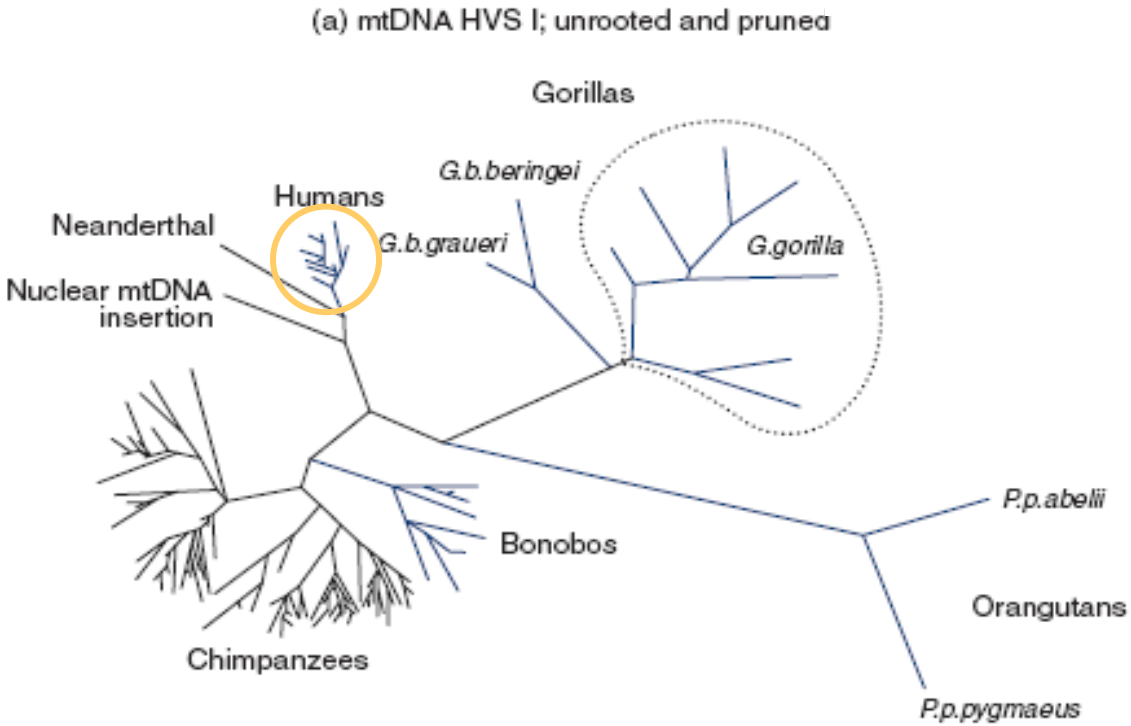
Conservazione della posizione dei geni tra uomo e topo

I cromosomi umani qui sono raffigurati con i 21 colori dei cromosomi murini



La diversità genetica umana è bassa



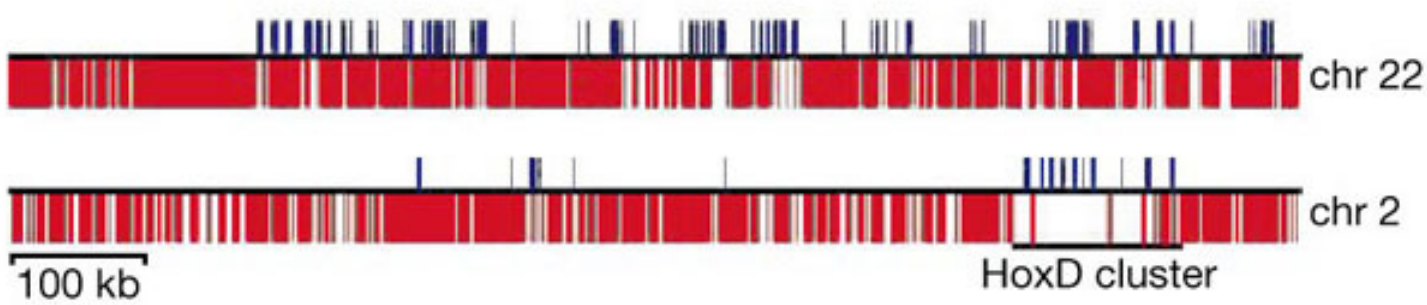


dimensione dei geni

- Gli esoni interni hanno una dimensione media di 145 bp
- Gli esoni più corti richiedono enhancer di splicing
- Il gene della distrofina è il più lungo del genoma umano (2.220.000)
- Il gene della titina ha il maggiore numero di esoni 324

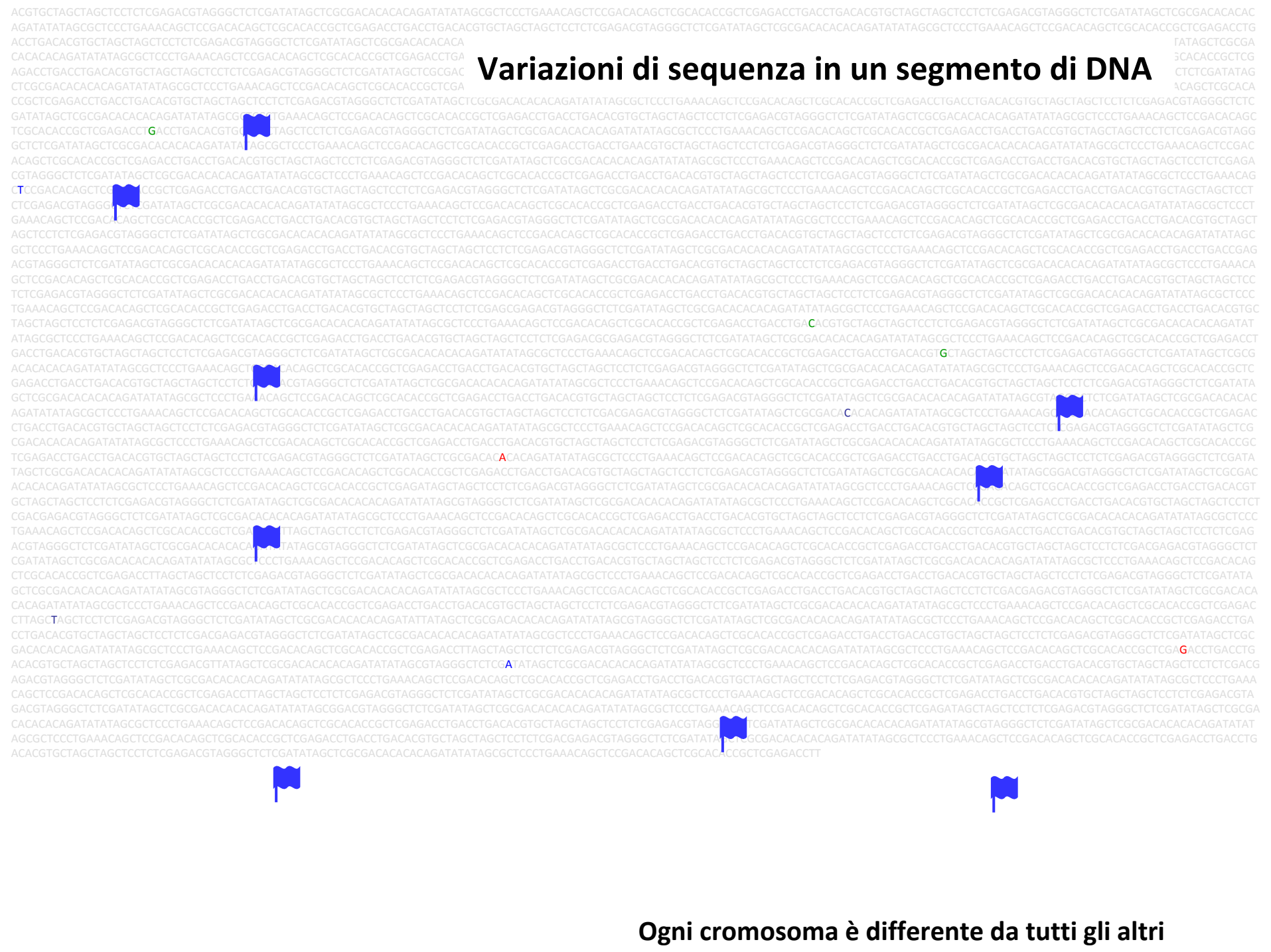
Le sequenze ripetute sono almeno il 50% del genoma

- elementi trasponibili (LINEs, SINEs, LTR retrotrasposoni, Trasposoni a DNA)
- pseudogeni processati
- singole ripetizioni
- duplicazioni segmentali
- blocchi di sequenze ripetute in tandem
- sono assenti nei loci HOX



rosso = elementi ripetuti, blu = esoni

Variazioni di sequenza in un segmento di DNA

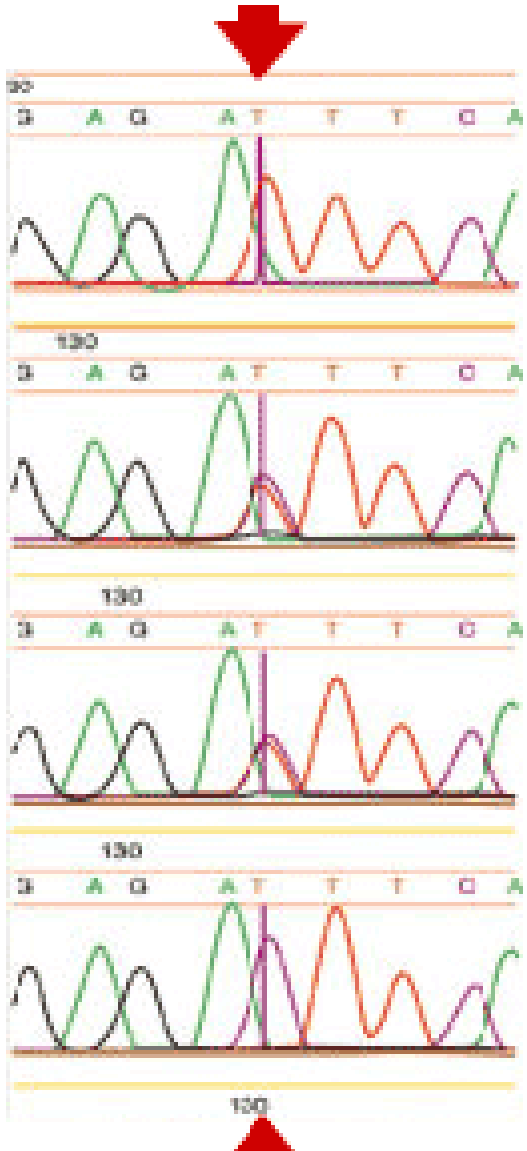


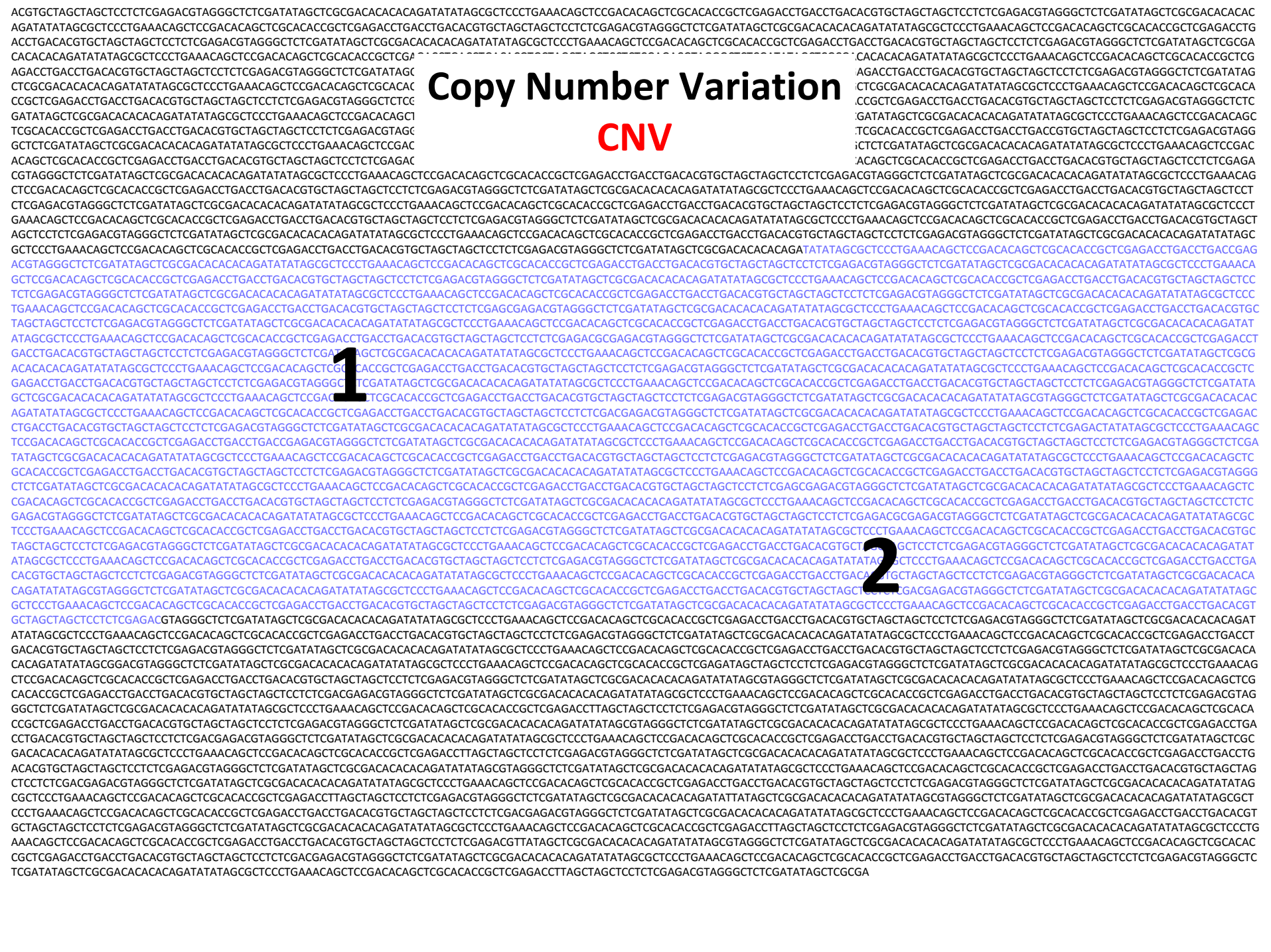
Ogni cromosoma è differente da tutti gli altri

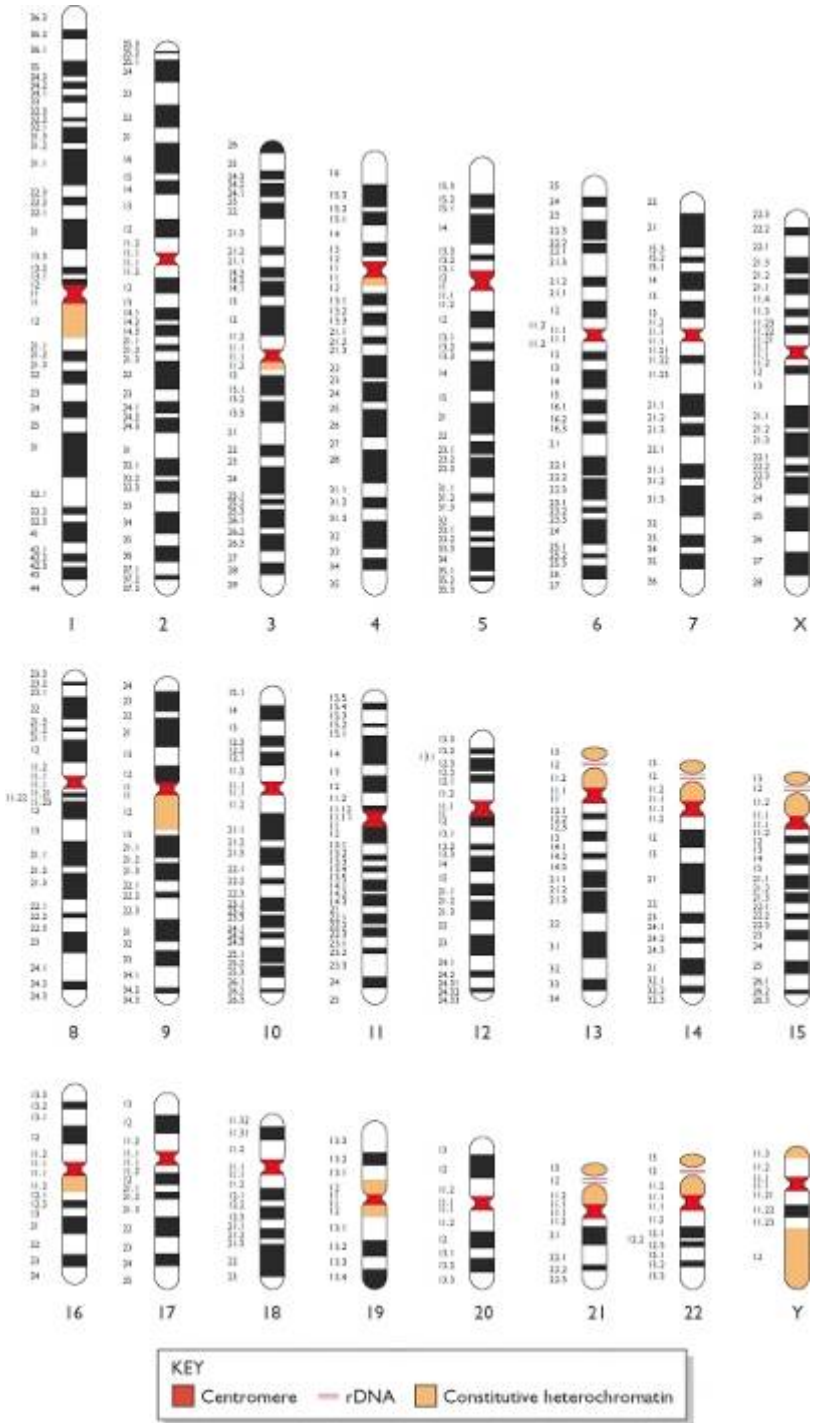
SNPs

single nucleotide polymorphisms

- Variazioni naturali che esistono tra le sequenze di qualsiasi cromosoma con un frequenza di almeno l'1% degli individui
- Consistono in sostituzioni di singoli nucleotidi, altre più rare consistono in delezioni o inserzioni di singoli nucleotidi
- Un SNP è identificato mediante sequenziamento del DNA di differenti cromosomi in individui diversi
- I due alleli possono essere identici (in omozigosi; T/T o C/C) o differenti (eterozigosi T/C o C/T) nel sito polimorfico




Copy Number Variation
CNV
1
2



CCDS IDs per chromosome	
Chromosome	**Count**
1	2,513
2	1,548
3	1,299
4	898
5	1,028
6	1,236
7	1,094
8	807
9	921
10	971
11	1,509
12	1,240
13	385
14	749
15	711
16	967
17	1,370
18	350
19	1,616
20	672
21	282
22	530
X	967
Y	53
XY	23

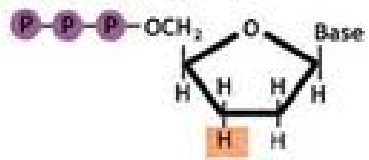
classificazione strutturale delle mutazioni

1. sostituzioni
2. piccole inserzioni, delezioni o inserzioni + delezioni contemporaneamente (indels)
3. riarrangiamenti genomici a due (delezioni, duplicazioni) o più punti di rottura (traslocazioni, inversioni ecc.)
4. copy number variations (CNV)

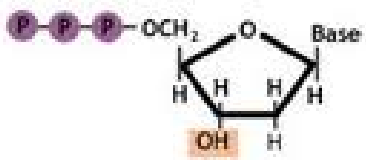
a queste quattro classi appartengono in modo indistinguibile tanto le variazioni innocue quanto le mutazioni causative di malattia

storia:
Sanger sequencing

Frederick Sanger
Nobel price 1958 and 1980
born August 13 1918, died November 19 2013



dideoxynucleotide (ddNTP)

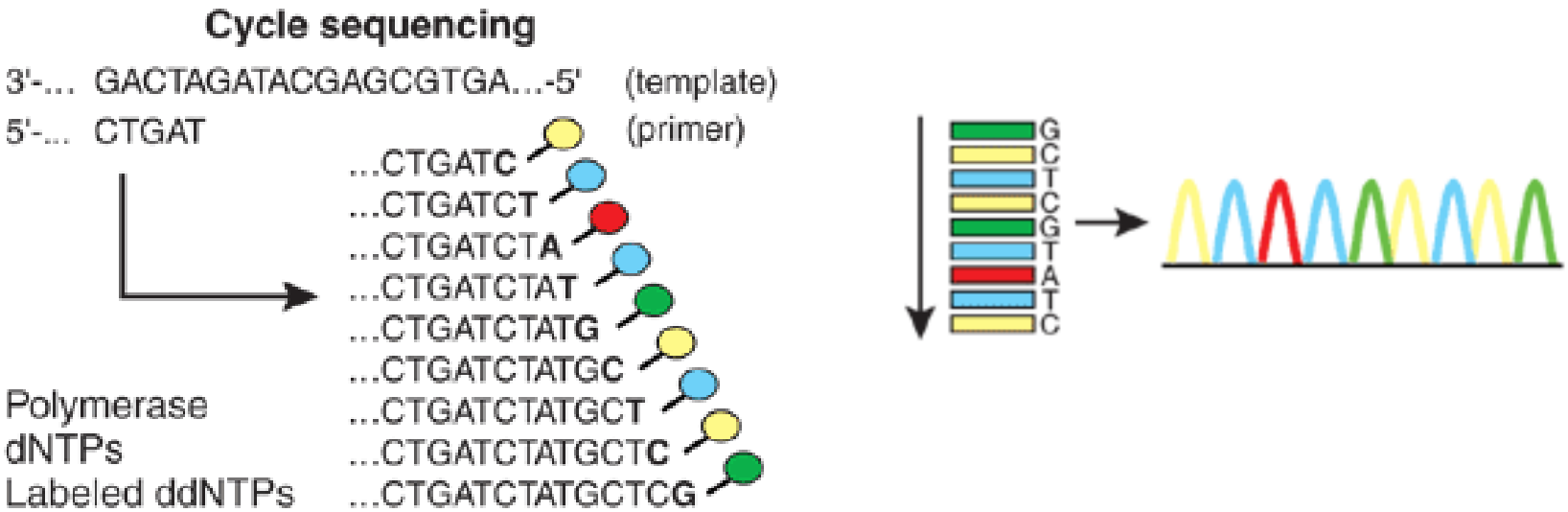


deoxynucleotide (dNTP)

I ddNTPs (ddATP, ddTTP, ddGTP, ddCTP) sono detti terminatori perché bloccano la polimerizzazione del DNA

Metodo “Sanger” modificato

- **le reazioni di sequenziamento sono ricopiature di milioni di stampi di DNA tutti identici**
- **ciascuna si blocca per l’inserimento casuale di un nucleotide fluorescente terminatore al posto di un nucleotide normale che avrebbe fatto continuare la ricopiatura**
- **la separazione dei frammenti fluorescenti per dimensione consente la decodifica**



“prima generazione”

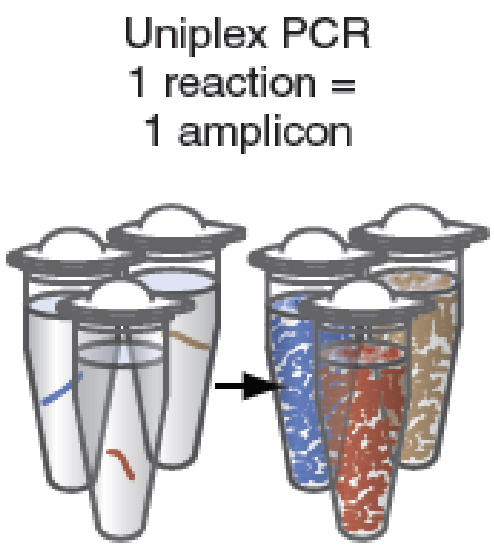
sequenziamento del DNA secondo il metodo Sanger

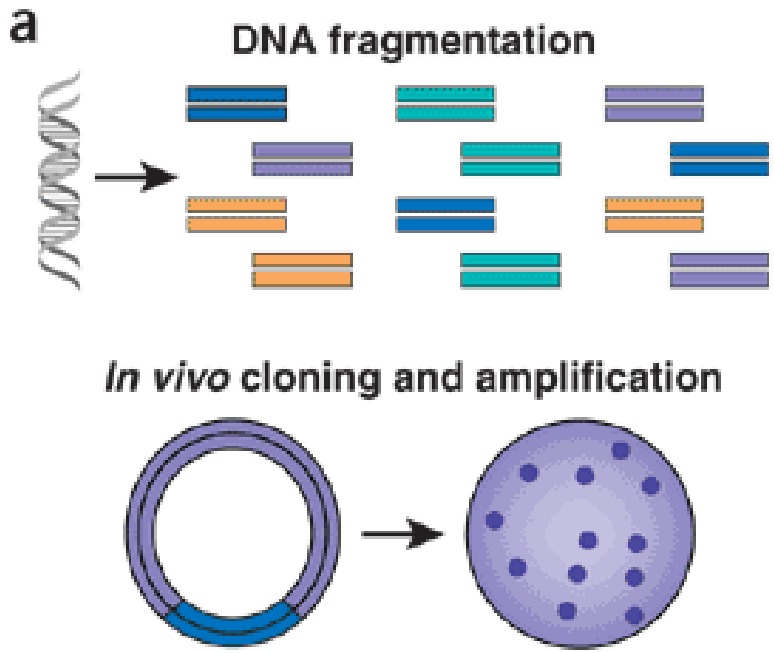
Il sequenziamento con tutte le versioni modificate del metodo Sanger ha dominato nella scienza e nell’industria per almeno 20 anni e ha consentito la lettura del genoma umano e la scoperta di oltre 2.500 malattie genetiche monogeniche

Il metodo Sanger rappresenta la tecnologia di **“prima generazione”**, mentre i nuovi metodi sono denominati **“next-generation sequencing (NGS)”**

Metodo "Sanger"

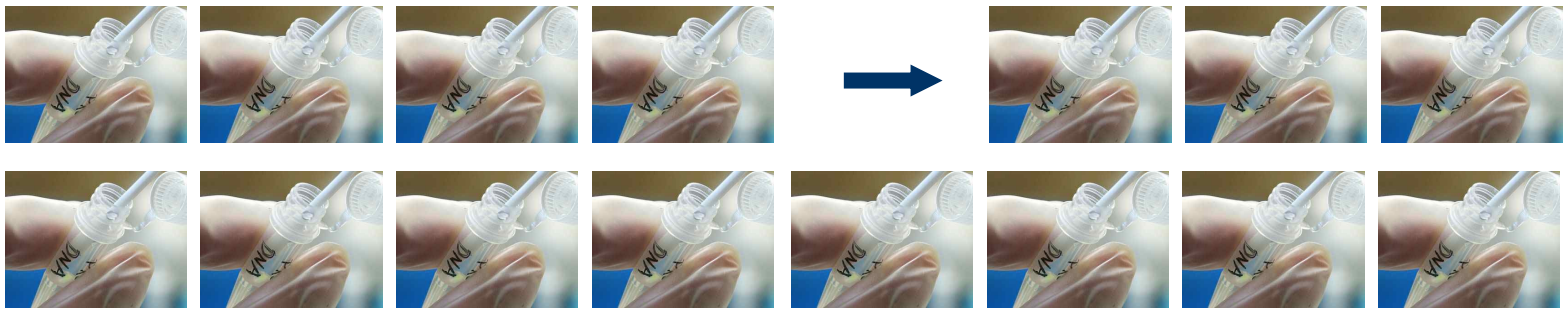
- prima si isolano segmenti di DNA chimicamente omogenei
- si replicano fino a nanogrammi di DNA
- poi partono le singole reazioni





Basato su Sanger è il gene-by-gene testing

- Troppo costoso e lungo per trovare la causa di malattie genetiche eterogenee



2ª generazione
"next-generation sequencing (NGS)"

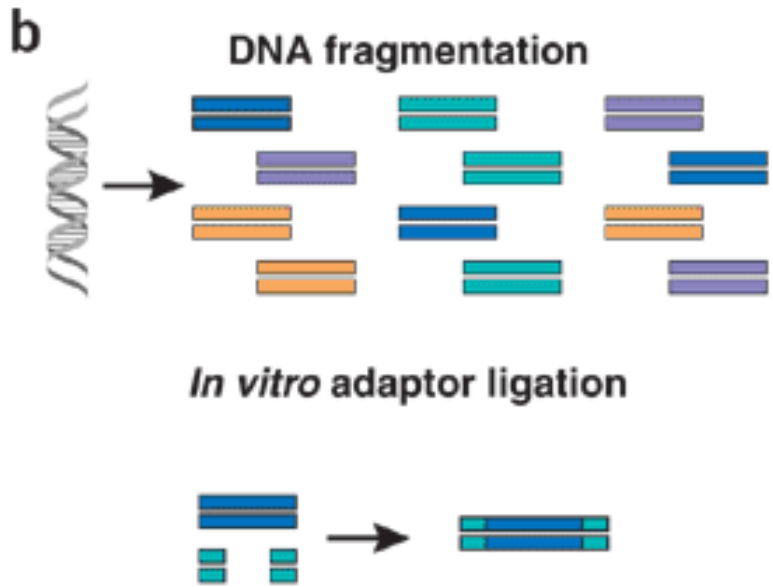
• Il principale vantaggio è poter lavorare con milioni di molecole di DNA senza doverle separare

• la possibilità tecnica di produrre un volume enorme di dati a **costi** estremamente più bassi ed in **tempi** estremamente più rapidi

• Il potenziale dell'NGS è simile ai primi tempi della PCR con il limite principale dovuto all'immaginazione

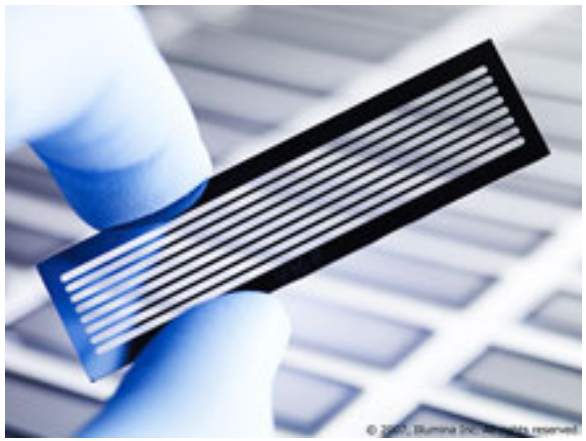
2ª generazione
"next-generation sequencing (NGS)"

- prima si frammenta il DNA casualmente
- si aggiungono adattatori comuni
- poi partono milioni di reazioni di sequenziamento individuali

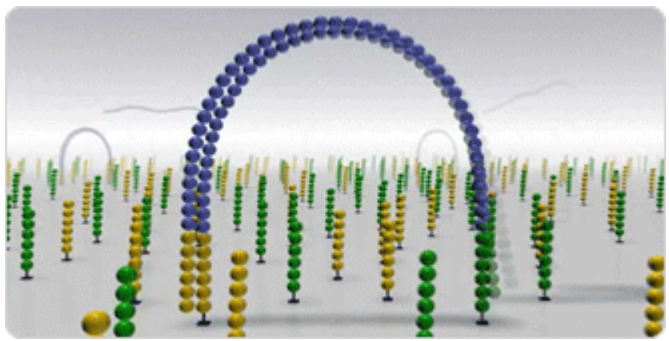


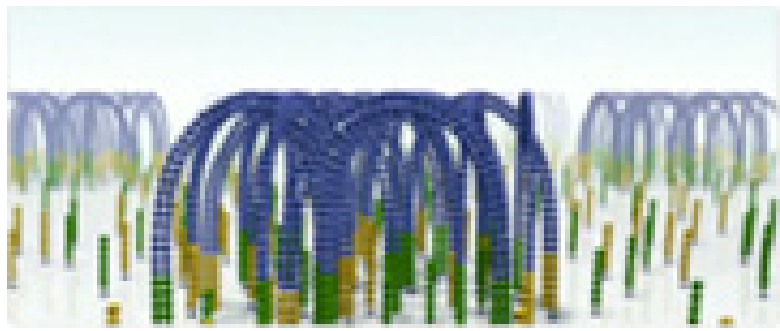
flow cell

la Solid-phase amplification può generare fino a 2.000 milioni di clusters di molecole di DNA distinte (Illumina HiSeq) per vetrino



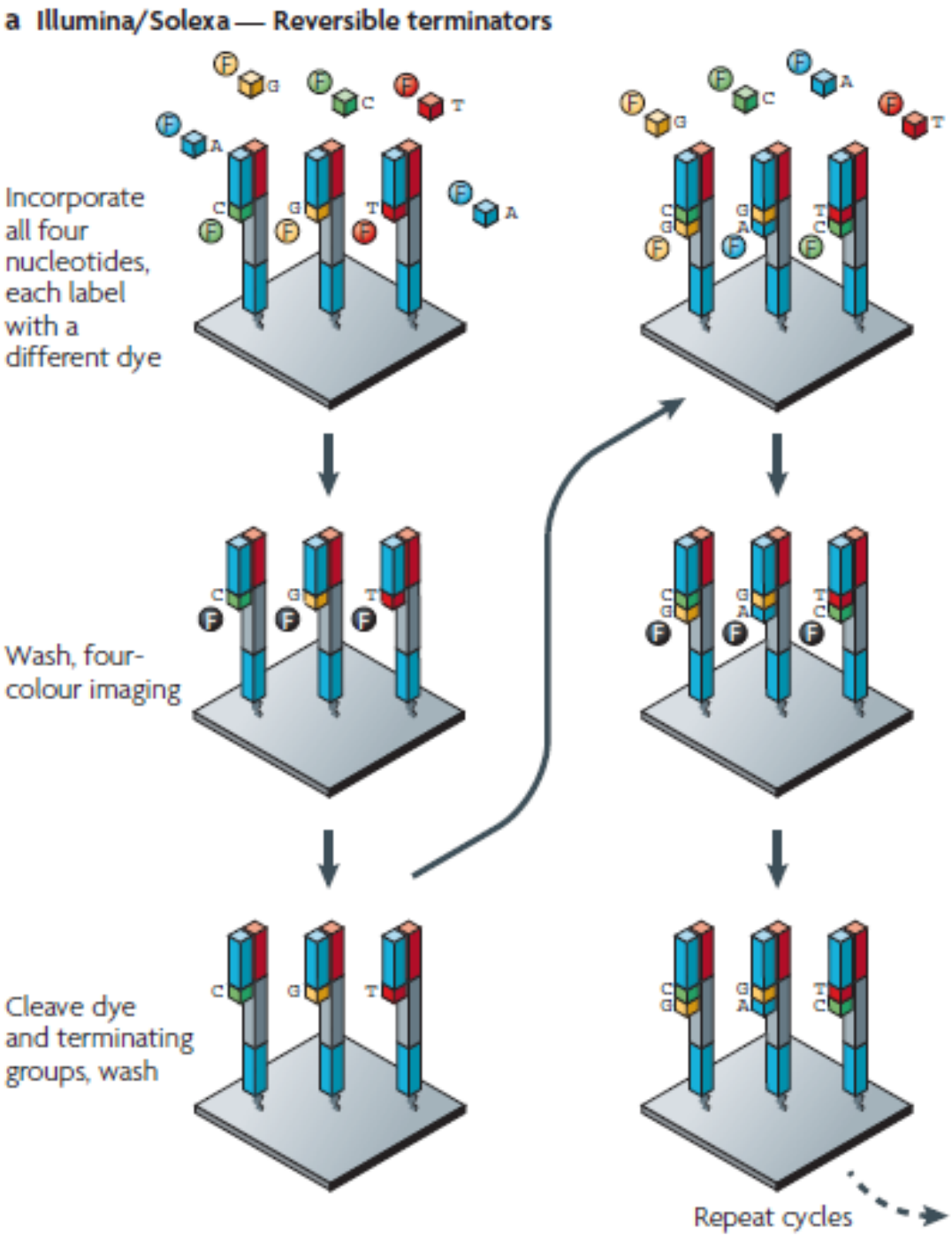
flow cell





Bridge PCR

illumina®

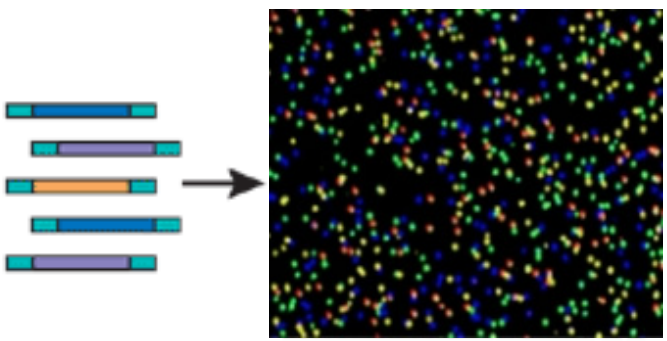


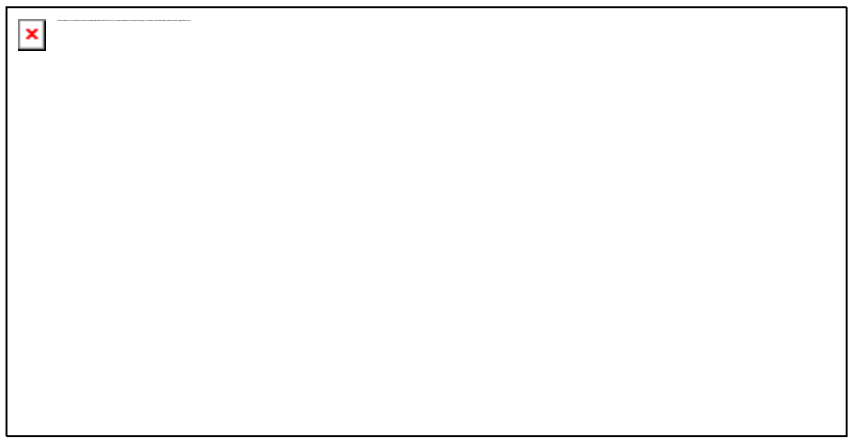
illumina®

I gruppi bloccanti attaccati al 3’ causano un ostacolo nell’incorporazione di un altro nucleotide, ma l’ostacolo è rimosso dopo ogni lettura e la reazione prosegue

Sistema Illumina

- ciascun frammento di DNA è immobilizzato su una superficie solida
- molecole di DNA spazialmente separate permettono di eseguire simultaneamente milioni di reazioni di sequenziamento
- ciascuna reazione produce una fluorescenza puntiforme che è fotografata
- la scansione di queste immagini consente di leggere la sequenza per ogni punto

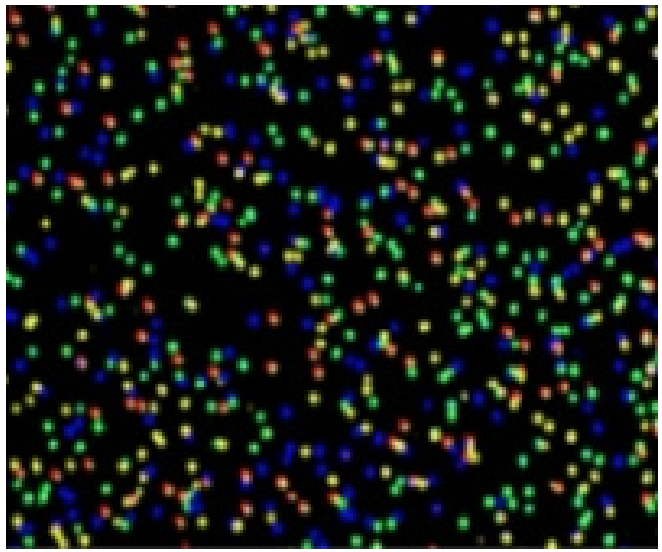


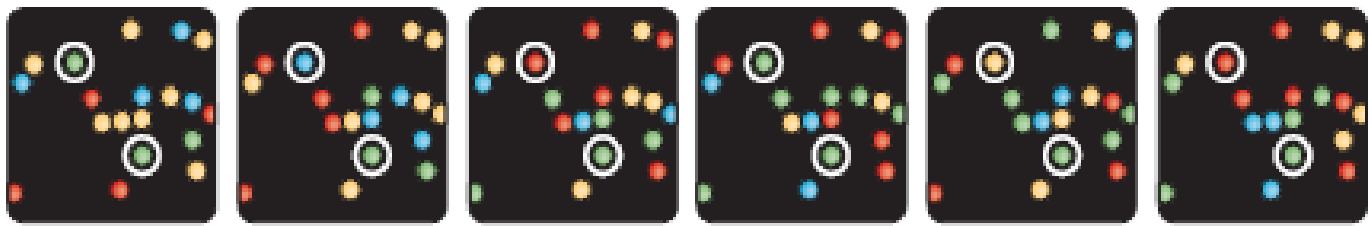


L'Illumina HiSeq2000 legge

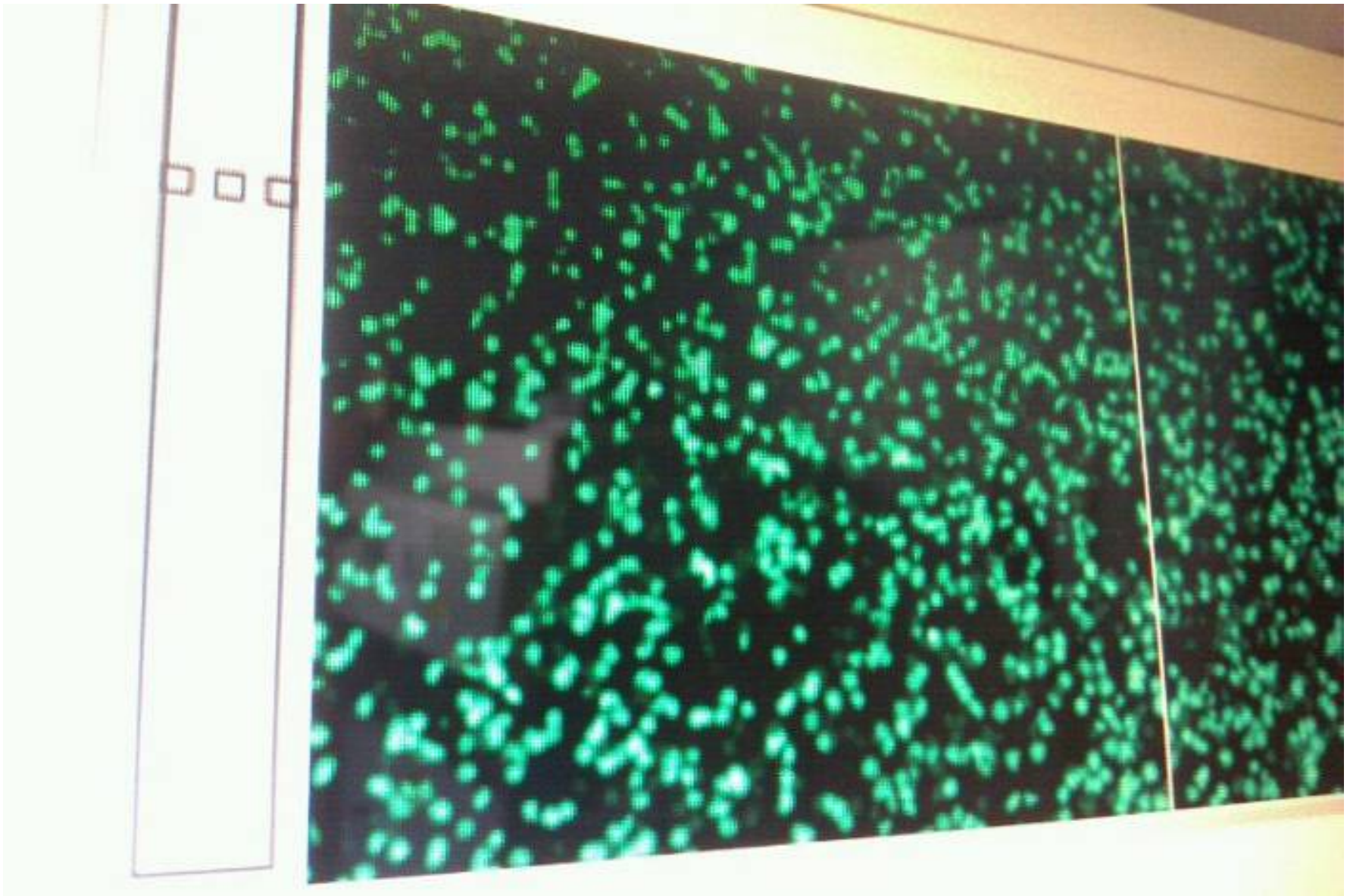
oltre 600.000.000.000 di basi

di DNA per corsa





Top: CATCGT
Bottom: CCCCCC



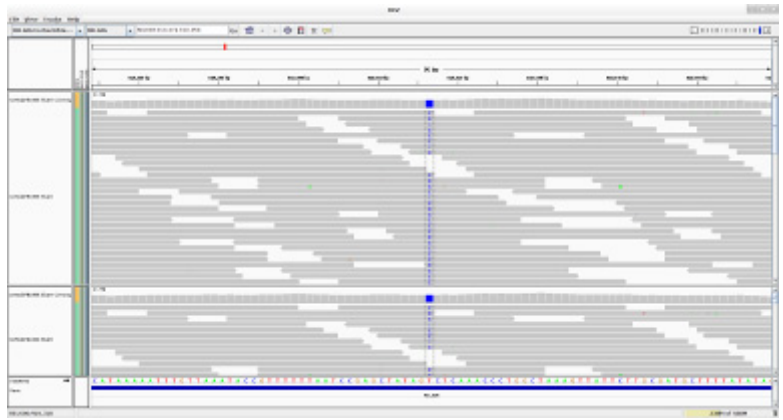
svantaggi

"next-generation sequencing (NGS)"

- la lunghezza di ciascuna sequenza è **più corta** (75-150 basi) rispetto alla tecnica Sanger (500-1000 basi)
- il numero di **errori** di lettura è molto più elevato, circa dieci volte più elevato
- ogni variazione di sequenza rilevante va confermata con metodica Sanger mirata
- esiste un complesso problema etico che riguarda le mutazioni trovate per caso (incidental findings) se queste hanno un valore predittivo

Integrative Genomics Viewer

Quali varianti del DNA possono essere perse dall'NGS?

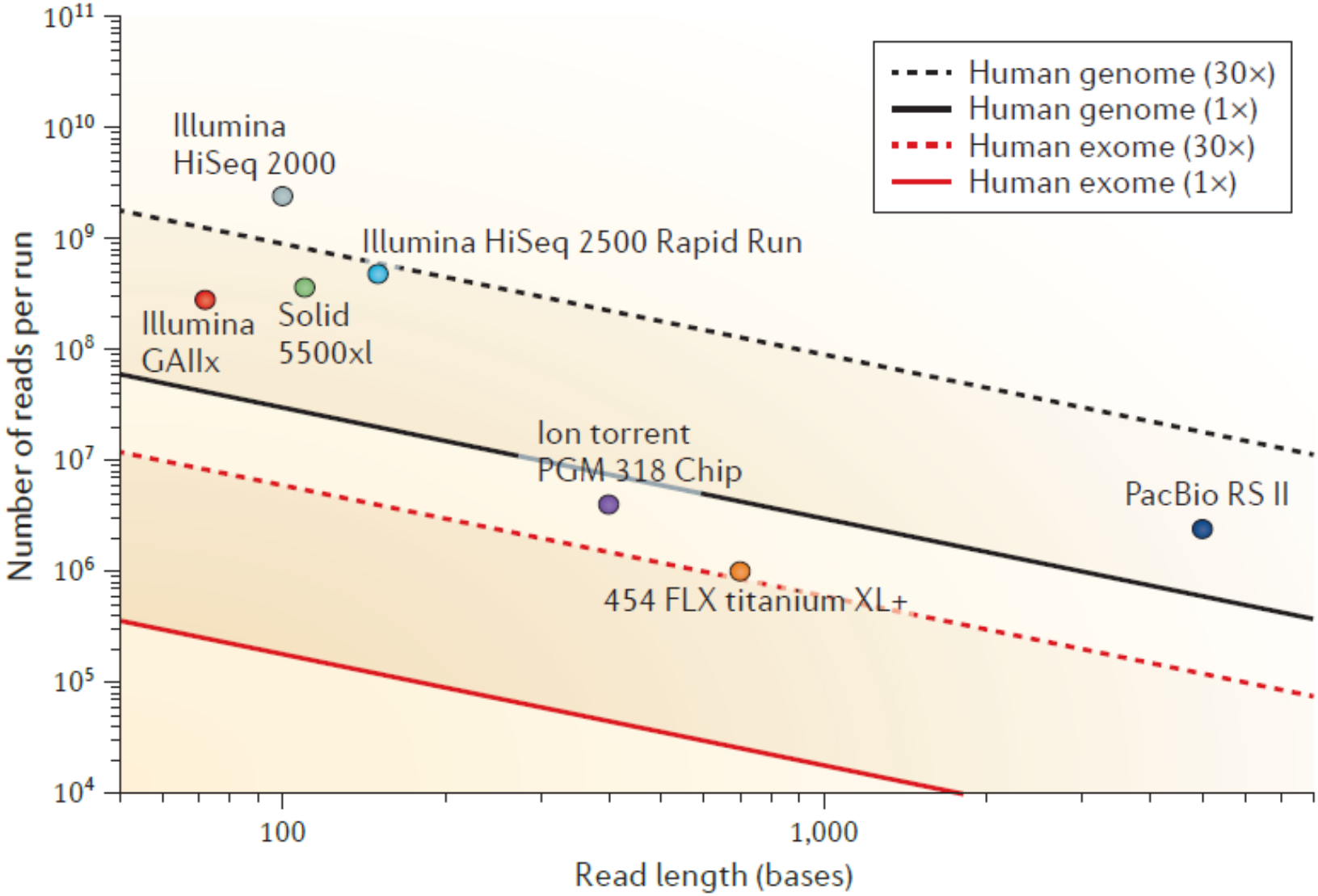


Non identificate

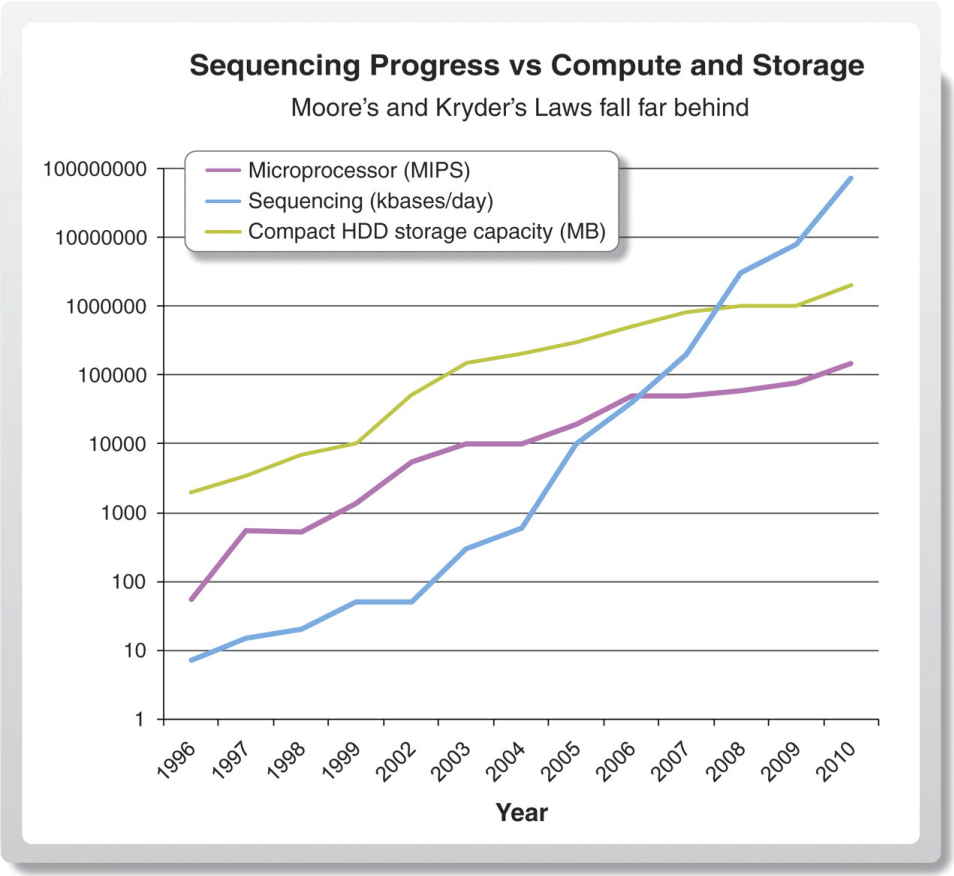
- quelle in regioni con poca/assente copertura
- inserzioni/delezioni >10-20nt
- Espansioni di triplette
- copy number variations
- traslocazioni
- inversioni

Non riconosciute

- Difetti di splicing elusivi
- Varianti in geni a funzione sconosciuta
- Varianti multiple di geni molto grandi



GAIIx, Genome Analyzer IIx; PacBio, Pacific Biosciences; PGM, personal genome machine.



La crescita delle potenzialità di sequenziamento sta superando i miglioramenti delle prestazioni della capacità di memorizzazione dei dati e di analisi computazionale

Science
AAAS

S D Kahn Science 2011;331:728-729

arricchimento “enrichment”

- tutti i geni (esoni+introni) di una regione cromosomica dove è mappata una malattia genetica da causa ignota
- tutti gli esoni di tutti i geni coinvolti in varie forme di quella malattia genetica
- tutti gli esoni di tutti i geni noti del genoma umano (esoma umano =whole exome)

meno dell’1% delle mutazioni che causano malattie genetiche cadono fuori dagli esoni

Il termine "NGS" indica tipologie di analisi molto differenti

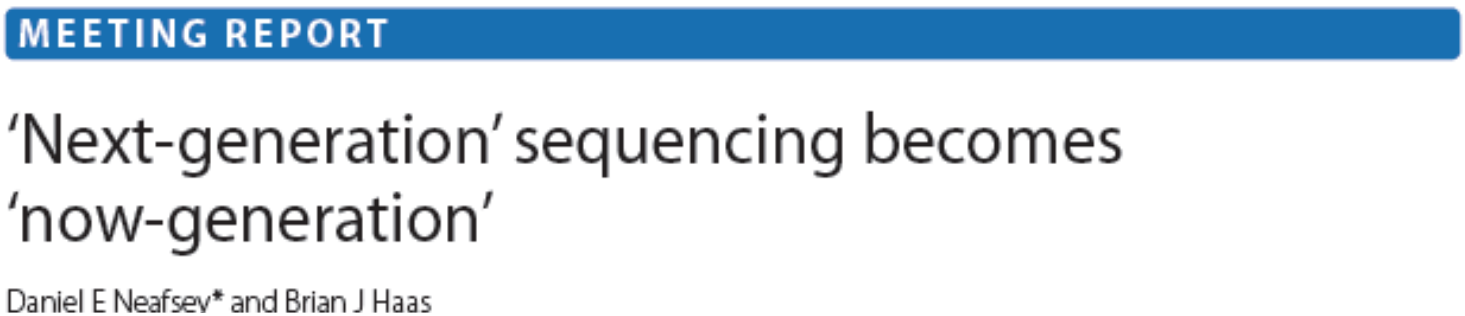
Whole genome
coverage @40 x (~ **150** Gb of sequences/sample) = 1,000 Gbytes

Whole exome
coverage @60 x (~ **8** Gb/sample) = 60 Gbytes/sample

Solo 10-15 geni noti
coverage @200 x (~ **0.1** Gb/sample) = 1 Gbytes/sample

- **Next generation sequencing** = **9.162** pubblicazioni
  - prima del 2010 erano **221**

- **Exome** = **2.491** pubblicazioni
  - prima del 2011 erano **53**
  - prima del 2010 erano **8**



nature Vol 452|17 April 2008|doi:10.1038/nature06884

LETTERS

The complete genome of an individual by massively parallel DNA sequencing

David A. Wheeler[1]*, Maithreyan Srinivasan[2]*, Michael Egholm[2]*, Yufeng Shen[1]*, Lei Chen[1], Amy McGuire[3], Wen He[2], Yi-Ju Chen[2], Vinod Makhijani[2], G. Thomas Roth[2], Xavier Gomes[2], Karrie Tartaro[2]†, Faheem Niazi[2], Cynthia L. Turcotte[2], Gerard P. Irzyk[2], James R. Lupski[4,5,6], Craig Chinault[4], Xing-zhi Song[1], Yue Liu[1], Ye Yuan[1], Lynne Nazareth[1], Xiang Qin[1], Donna M. Muzny[1], Marcel Margulies[2], George M. Weinstock[1,4], Richard A. Gibbs[1,4] & Jonathan M. Rothberg[2]†

The association of genetic variation with disease and drug response, and improvements in nucleic acid technologies, have given great optimism for the impact of 'genomic medicine'. However, the formidable size of the diploid human genome[1], approximately 6 gigabases, has prevented the routine application of sequencing methods to deciphering complete individual human genomes. To realize the full potential of genomics for human health, this limitation must be overcome. Here we report the DNA sequence of a diploid genome of a single individual, James D. Watson, sequenced to 7.4-fold redundancy in two months using massively parallel sequencing in picolitre-size reaction vessels. This sequence was completed in two months at approximately one-hundredth of the cost of traditional capillary electrophoresis methods. Comparison of the sequence to the reference genome led to the identification of 3.3 million single nucleotide polymorphisms, of which 10,654 cause amino-acid substitution within the coding sequence. In addition, we accurately identified small-scale (2–40,000 base pair (bp)) insertion and deletion polymorphism as well as copy number variation resulting in the large-scale gain and loss of chromosomal segments ranging from 26,000 to 1.5 million base pairs. Overall, these results agree well with recent results of sequencing of a single individual[2] by traditional methods. However, in addition to being faster and significantly less expensive, this sequencing technology avoids the arbitrary loss of genomic sequences inherent in random shotgun sequencing by bacterial cloning because it amplifies DNA in a cell-free system. As a result, we further demonstrate the acquisition of novel human sequence, including novel genes not previously identified by traditional genomic sequencing. This is the first genome sequenced by next-generation technologies. Therefore it is a pilot for the future challenges of 'personalized genome sequencing'.

To catalogue the genomic diversity within a single individual, a total of 106.5 million high-quality reads were generated by 454-sequencing[3], representing approximately 24.5 billion DNA bases. Reads that aligned to the genome were further filtered using stringent criteria to ensure the accuracy of mapping, resulting in 93.2 million reads aligned to reference genome sequence. The reference genome sequence was thus covered to an average depth of 7.4-fold (Fig. 1a). The alignments between the uniquely mapped reads and the reference genome were used to catalogue genetic variation in the subject's DNA, including single nucleotide polymorphisms (SNPs), small insertions and deletions (indels), and copy number variation (CNV).

The 454 base-calling software provides error estimates (*Q* values) for each base. We developed a three-step filtering process using the patterns of error and associated *Q* values from the 454 base-calling software to improve the accuracy of SNP discovery. An initial 14 million variant positions were filtered to 3.32 million putative SNPs (Table 1).

Comparison of these putative SNPs in the subject's genome with those in the dbSNP (dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP/) revealed 2.72 million in common ('known SNPs'). Approximately 99% of SNPs in dbSNP are bi-allelic. At only 10,425 positions did the subject's variant not match the variant found in dbSNP. Although some of these could represent a third allele in the population, or an error in the dbSNP polymorphism record, we conservatively estimated the false discovery rate in the known SNPs to be approximately 0.38% based on the mismatches with dbSNP.

The remaining 0.61 million SNPs were at positions not previously identified as polymorphic in dbSNP ('novel SNPs'). The known SNPs were divided almost equally between homozygous (50.2%) and heterozygous (49.8%) SNPs, whereas within the novel SNPs heterozygotes predominate (83.3%) compared with homozygotes (16.7%). Because most common alleles in human populations are already captured in dbSNP, novel variants are expected to be rare, and therefore much more likely to be found as heterozygotes.

We assessed the accuracy of the known SNPs derived from DNA sequencing by comparison with the experimental genotyping of the subject's DNA using an Affymetrix 500K microarray. Compared with a haploid reference sequence, there are four possible outcomes of SNP array genotyping: homozygous for the reference allele; homozygous for the variant (non-reference) allele; heterozygous; and assay failure. Table 2 shows the results for 494,713 markers that were successfully genotyped. The subject's DNA sequence exhibited only the reference allele at 99.4% of the markers homozygous for the reference and at 95.1% of markers homozygous for the variant. Genotyping identified 135,413 heterozygous markers of which 75.8% exhibited two alleles in the 454-reads. The lower sensitivity of detection of heterozygotes is predicted by a Poisson process of sampling DNA fragments modelled on a diploid genome (Methods). Consistent

[1]Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. [2]454 Life Sciences, Roche Diagnostics, 20 Commercial Street, Bradford, Connecticut 06405, USA. [3]Center for Ethics and Health Policy, Baylor College of Medicine, One Baylor Plaza, Houston Texas 77030, USA. [4]Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston Texas 77030, USA. [5]Department of Pediatrics, Baylor College of Medicine, One Baylor Plaza, Houston Texas 77030, USA. [6]Texas Children's Hospital, Texas Medical Center, Houston, Texas 77030, USA. †Present addresses: Molecular Imaging Systems, Carestream Health, Inc., 4 Science Park, New Haven, Connecticut 06511, USA (K.T.); Rothberg Institute for Childhood Diseases, 530 Whitfield Street, Guilford, Connecticut 06437, USA (J.M.R.).
*These authors contributed equally to this work.

Roche 454 SEQUENCING

The genome of James Watson

7.4 x coverage

234 runs

24.5 billions bp

11 mutazioni causative di malattie genetiche

Table 3 | SNPs matching HGMD mutations causing disease or other phenotypes

HGMD accession	Chromosome	Coordinate	HUGO symbol	Gene name	Cytogenetic	Phenotype	Zygosity
CM003589	1	97937679	DPYD	Dihydropyrimidine dehydrogenase	1q22	Dihydropyrimidine dehydrogenase deficiency	Heterozygous
CM950484	1	157441978	FY	Duffy blood-group antigen	1q	Duffy blood group antigen, absence	Homozygous*
CM942034	4	619702	PDE6B	Phosphodiesterase 6B, cGMP-specific, rod, beta	4p16.3	Retinitis pigmentosa 40	Heterozygous
CM021718	9	36208221	GNE	UDP-N-acetylglucosamine 2-epimerase	9p	Myopathy, distal, with rimmed vacuoles	Heterozygous
CM980633	10	50348375	ERCC6	Excision repair cross-complementing rodent repair deficiency, complementation group 6 protein (CSB)	10q	Cockayne syndrome	Homozygous†
CM050716	11	76531431	MYO7A	Myosin VIIA	11q13.5	Usher syndrome 1b	Homozygous†
CM950928	12	46812979	PFKM	Phosphofructokinase, muscle	12q13.3	Glycogen storage disease 7	Homozygous*
CM032029	14	20859880	RPGRIP1	Retinitis pigmentosa GTPase regulator interacting protein 1	14q11	Cone–rod dystrophy	Heterozygous
CM984025	19	18047618	IL12RB1	Interleukin-12 receptor, beta 1	19p13.1	Mycobacterial infection	Heterozygous
CM024138	19	41014441	NPHS1	Nephrosis-1, congenital, Finnish type	19q	Congenital nephrotic syndrome, Finnish type	Heterozygous
CM910052	22	49410905	ARSA	Arylsulphatase A	22q	Metachromatic leukodystrophy	Heterozygous

Roche 454 SEQUENCING

"Whole - Genome" Seq

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Whole-Genome Sequencing in a Patient with Charcot–Marie–Tooth Neuropathy

James R. Lupski, M.D., Ph.D., Jeffrey G. Reid, Ph.D., Claudia Gonzaga-Jauregui, B.S., David Rio Deiros, B.S., David C.Y. Chen, M.Sc., Lynne Nazareth, Ph.D., Matthew Bainbridge, M.Sc., Huyen Dinh, B.S., Chyn Jing, M.Sc., David A. Wheeler, Ph.D., Amy L. McGuire, J.D., Ph.D., Feng Zhang, Ph.D., Pawel Stankiewicz, M.D., Ph.D., John J. Halperin, M.D., Chengyong Yang, Ph.D., Curtis Gehman, Ph.D., Danwei Guo, M.Sc., Rola K. Irikat, B.S., Warren Tom, B.S., Nick J. Fantin, B.S., Donna M. Muzny, M.Sc., and Richard A. Gibbs, Ph.D.

ABSTRACT

very heterogeneous genotype, genetic testing for 15/39 loci

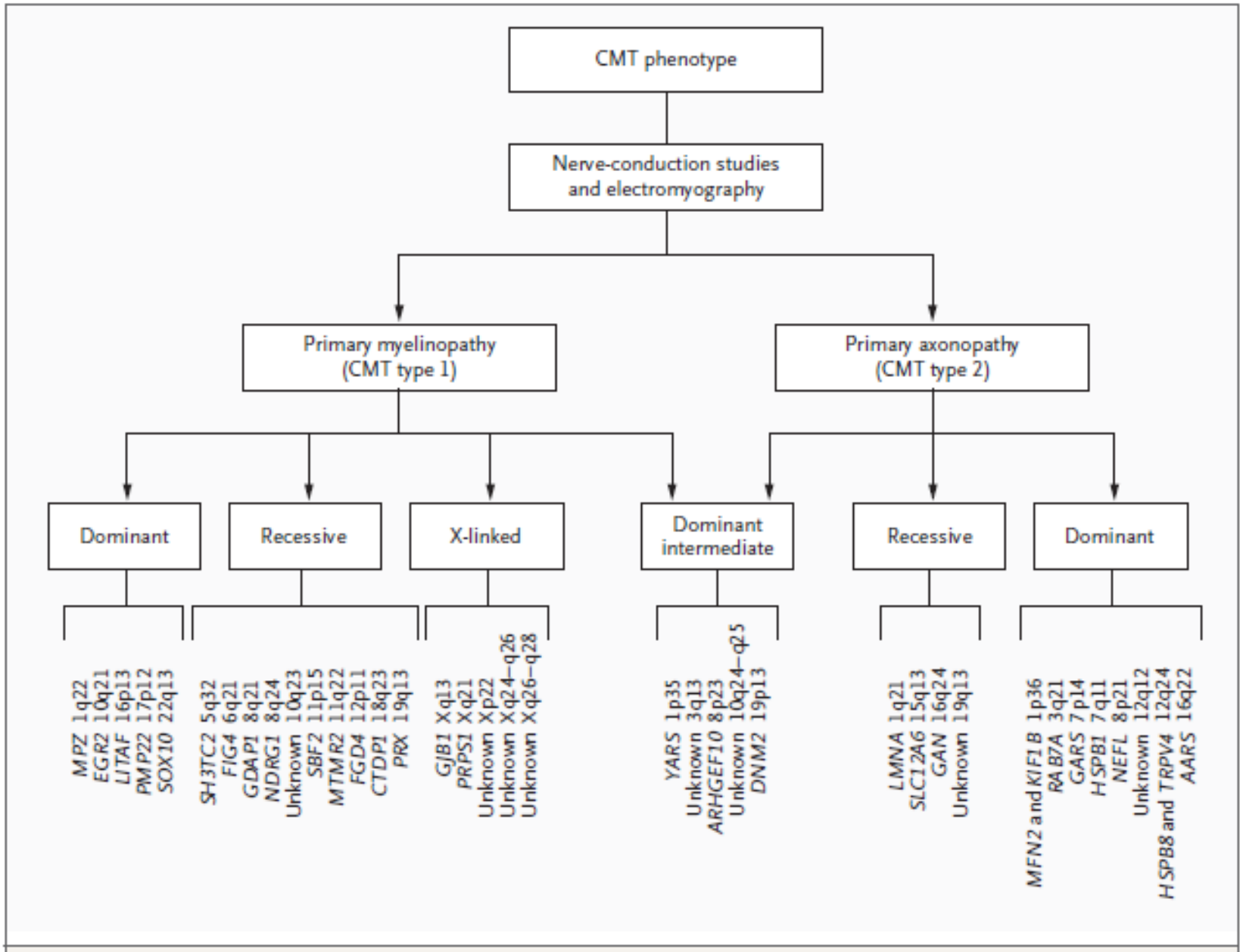
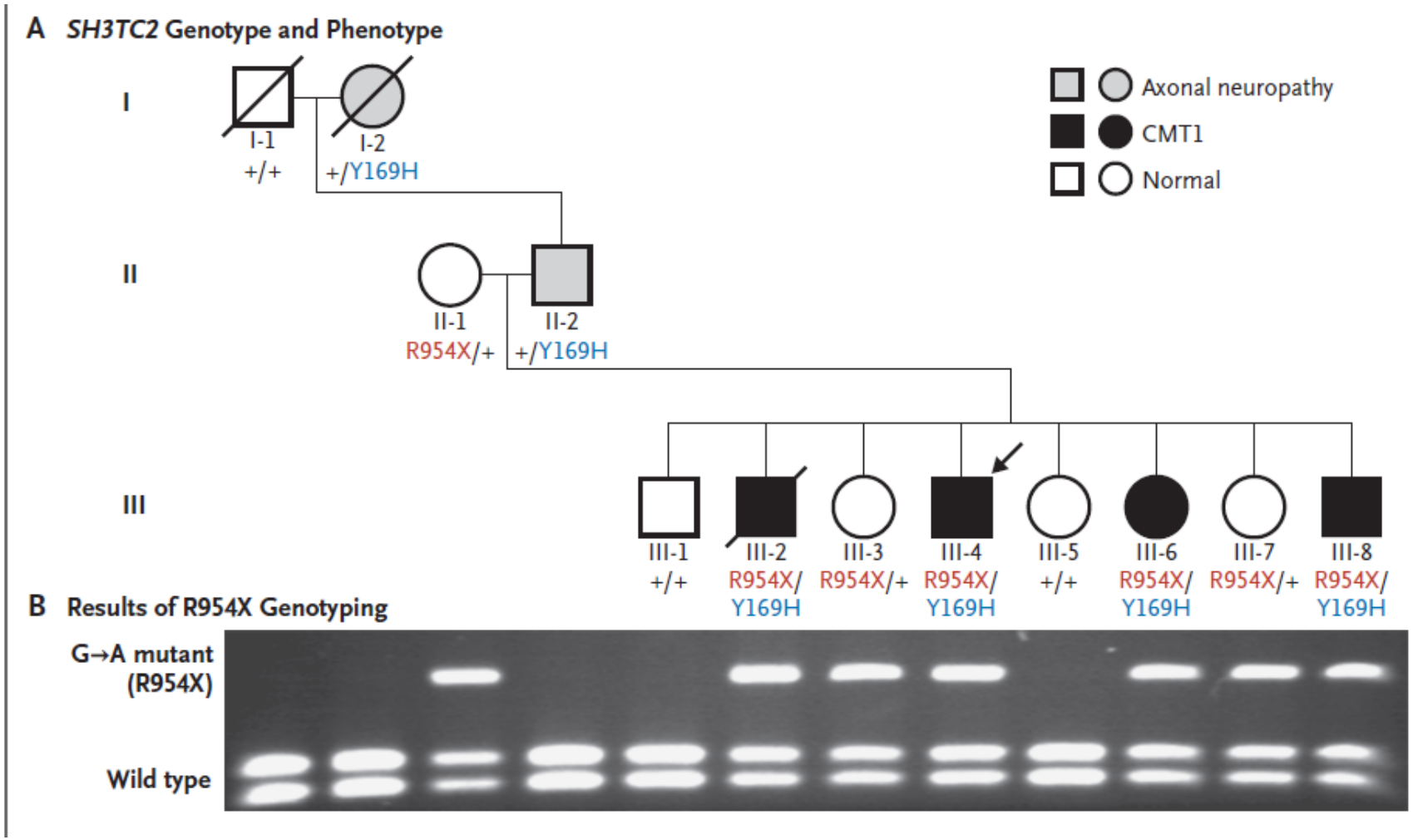


Table 4. Disease and Trait Associations of Nonsynonymous SNPs Identified in the Proband, According to the Human Gene Mutation Database.*

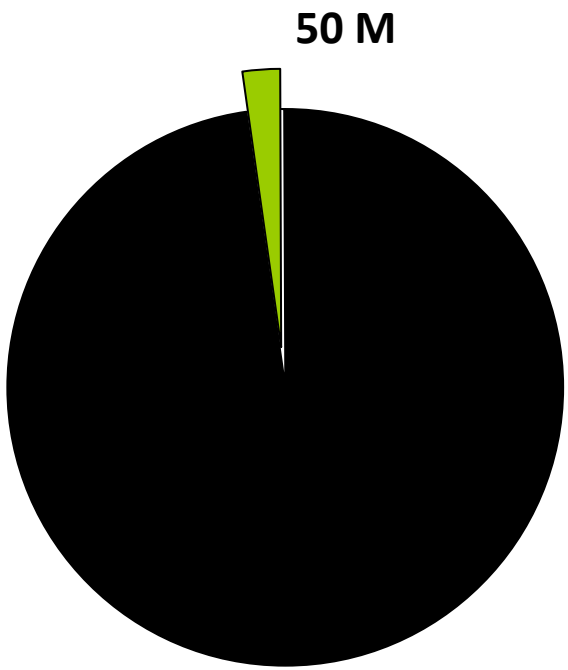
Disease or Trait Associated with Mutation	SNPs
	no. (%)
Total	159 (100)
Behavioral disorder	6 (4)
Cancer	33 (21)
Association	7
Increased risk	9
Reduced risk	3
Susceptibility	14
Complex disease	48 (30)
Mendelian disease	21 (13)
Metabolic trait	17 (11)
Pharmacogenetic trait	14 (9)
Other traits	20 (13)

* SNP denotes single-nucleotide polymorphism.

Pedigree of the family and segregation of SH3TC2 mutations

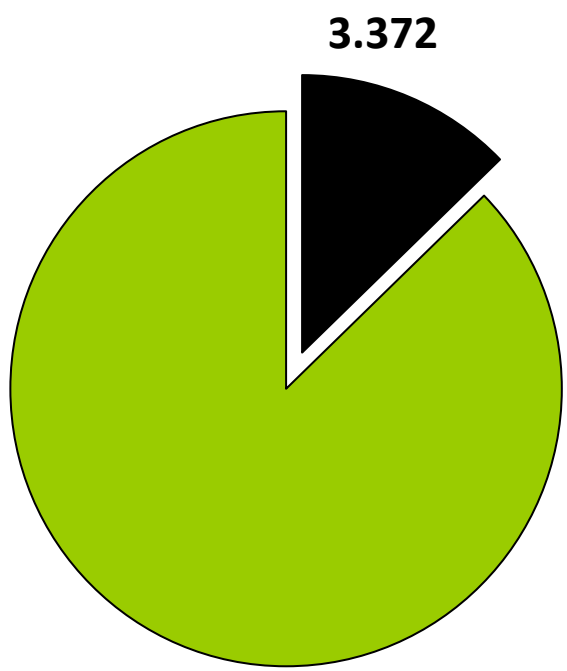


Esoni dei geni umani
(in milioni di basi)



Totale ~ 3.100 M

Geni con mutazioni che causano
malattie umane



Totale ~ 22.000 geni

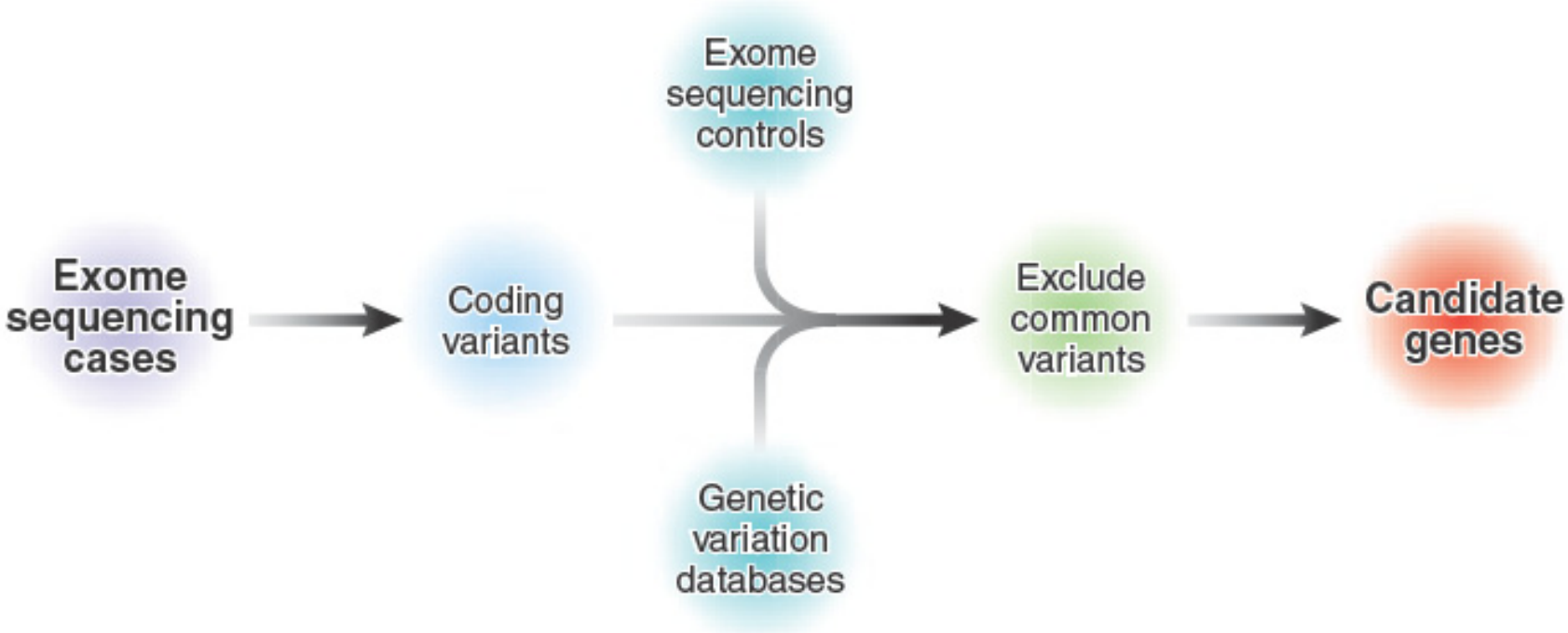
le mutazioni di sole 12 basi del DNA su 10.000 causano tutte le malattie monogeniche (mendeliane) note sinora

nature Vol 461|10 September 2009|doi:10.1038/nature08250

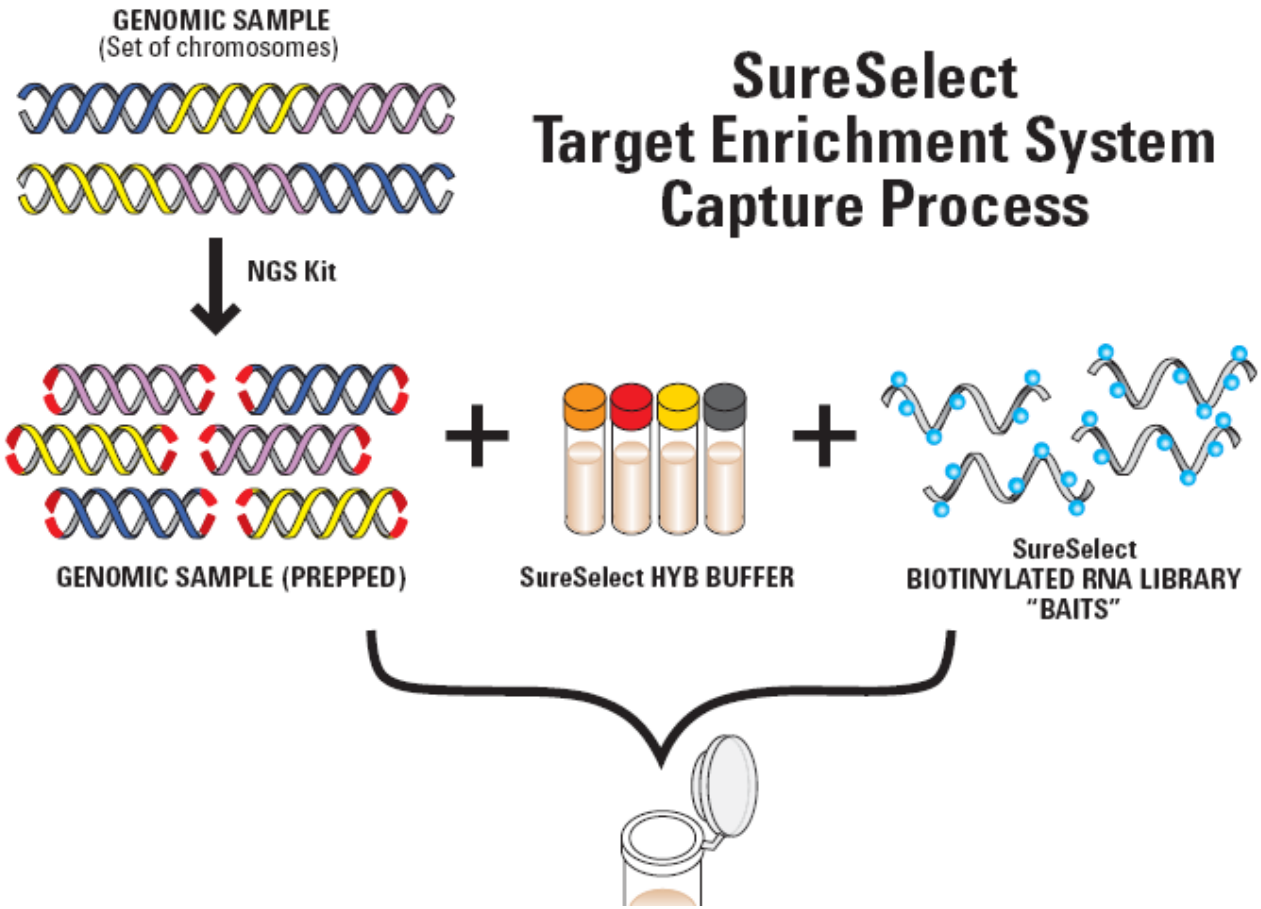
LETTERS

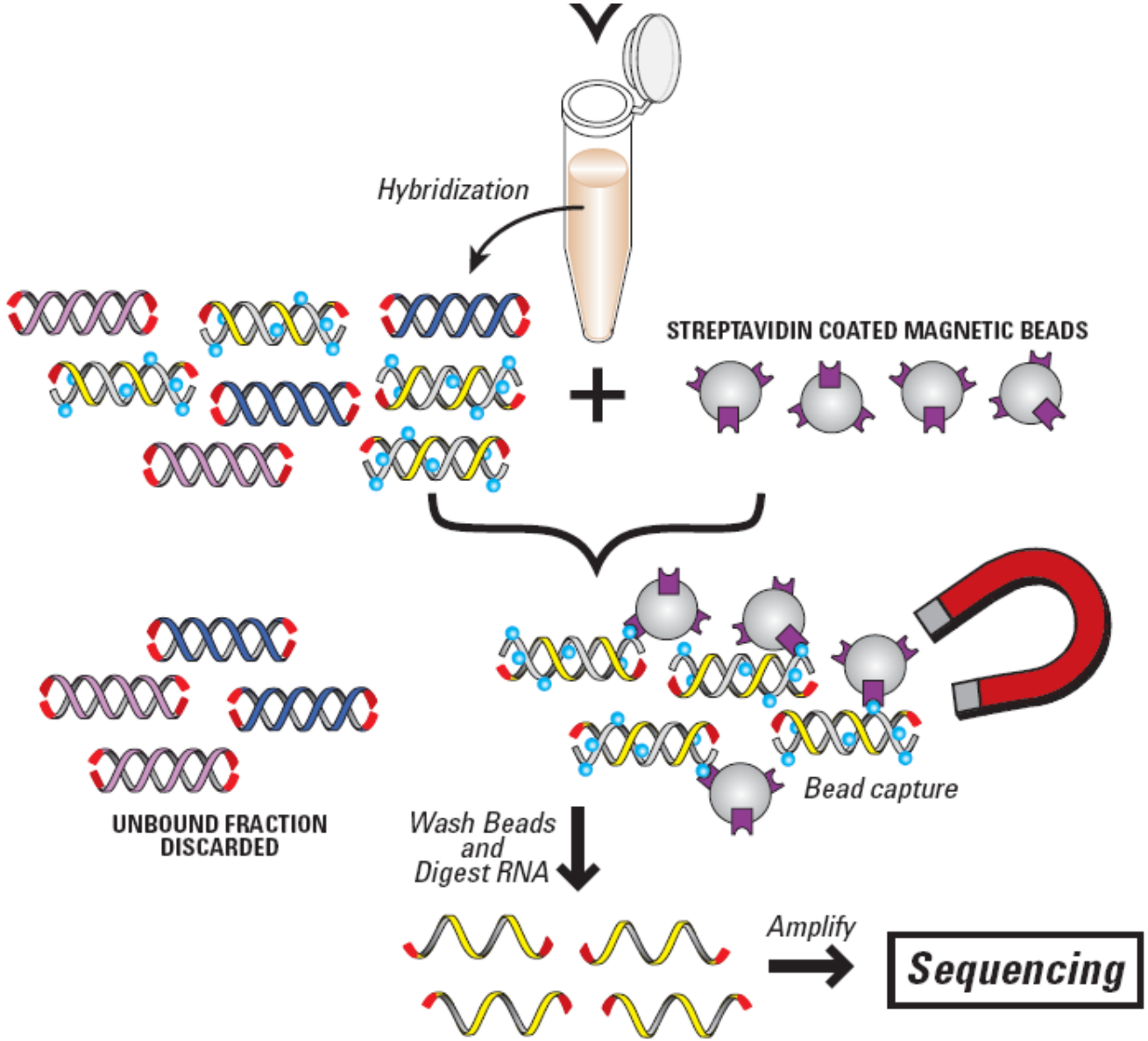
Targeted capture and massively parallel sequencing of 12 human exomes

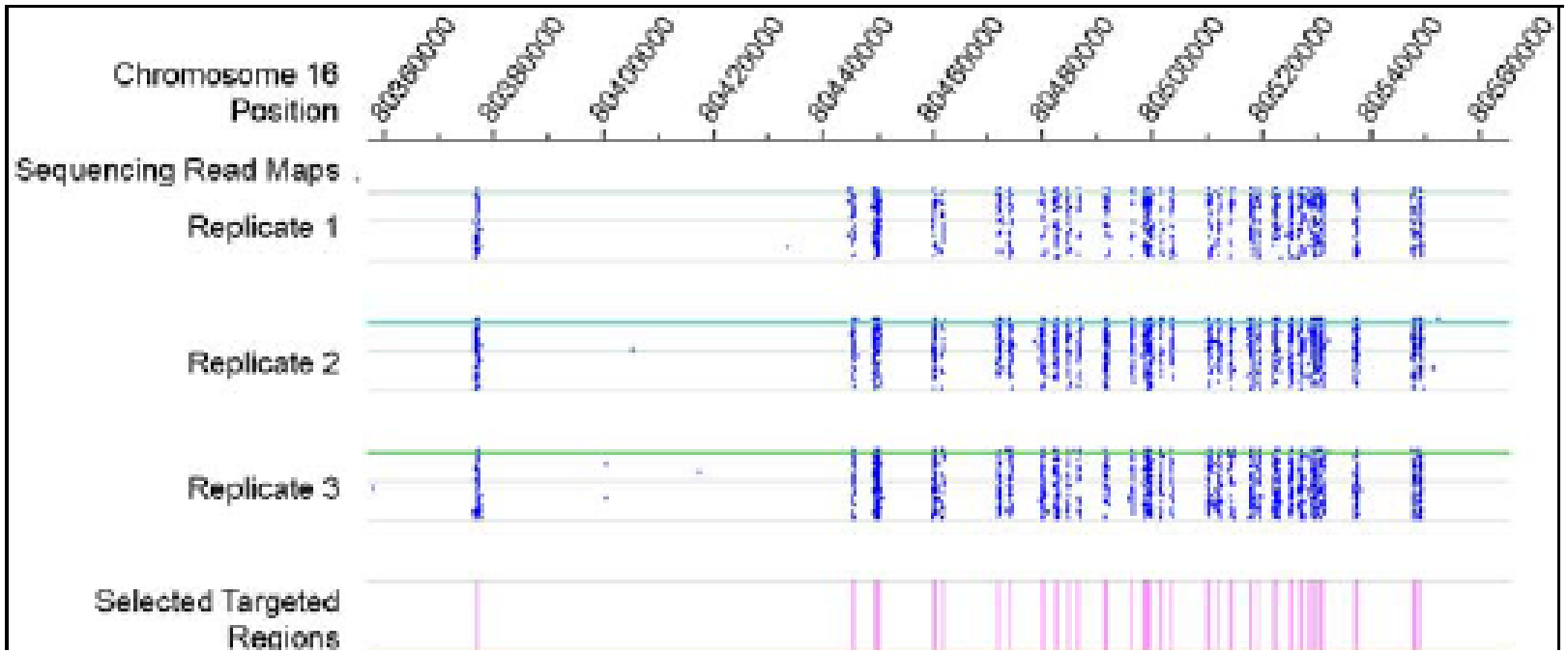
Sarah B. Ng[1], Emily H. Turner[1], Peggy D. Robertson[1], Steven D. Flygare[1], Abigail W. Bigham[2], Choli Lee[1], Tristan Shaffer[1], Michelle Wong[1], Arindam Bhattacharjee[4], Evan E. Eichler[1,3], Michael Bamshad[2], Deborah A. Nickerson[1] & Jay Shendure[1]


Exome sequencing controls
Exome sequencing cases
Coding variants
Exclude common variants
Candidate genes
Genetic variation databases

SureSelect Target Enrichment System Capture Process




Hybridization
STREPTAVIDIN COATED MAGNETIC BEADS
UNBOUND FRACTION
DISCARDED
Bead capture
Wash Beads
and
Digest RNA
Amplify
Sequencing



minigenome preparation

Come si fa a comprendere quale variazione nella sequenza del DNA possa avere un significato patologico?

	Dr. Venter's Exome	Dr. Watson's Exome
Total Number of Nonsynonymous SNPs	10,389	10,569
Number of Novel Nonsynonymous SNPs	772 (7% of total nsSNPs)	1,573 (15% of total nsSNPs)
% nsSNPs predicted to affect protein function*	14% (7,781 predicted on)	20% (3,898 predicted on)
Number of Coding Indels	739	345**

*Different prediction algorithms were used [30,33], and this may account for the difference between the two exomes.

**Indels of size 2 bp and greater were considered; 1 bp indels were discarded. If we removed 1 bp indels from Dr. Venter's exome in order to compare with Dr. Wats exome, Dr. Venter would have 423 coding indels.

doi:10.1371/journal.pgen.1000160.t004