

Introduzione al corso di bioinformatica e analisi dei genomi AA 2014-2015

Docente: Silvia Fuselli
fss@unife.it

Fonti e testi di riferimento

- Dan Graur: <http://nsmn1.uh.edu/dgraur/> >courses > bioinformatics for biologists
- Introduzione alla Genomica, A.M. Lesk, Zanichelli
- Bioinformatica, Pascarella e Paiardini, Zanichelli
- Grazie alla dottoressa Silvia Ghirotto per aver condiviso il materiale del corso tenuto da lei l'anno passato

Programma preliminare

- PARTE 1
 - I. Dimensioni ed organizzazione dei genomi
 - a. Procarioti
 - b. Eucarioti (genoma umano)
 - II. Evoluzione dei genomi
 - III. Metodologie di analisi dei genomi con particolare approfondimento dei metodi di Next generation sequencing (NGS)
 - IV. Metodologie di analisi di modelli di espressione*
- PARTE 2
 - I. Banche dati biologiche
 - II. Confronto di sequenze: allineamenti a coppie e allineamenti multipli, ricerche di sequenze in banche dati
 - III. Bioinformatica e next generation sequencing (dott. Telatin)

Le due protagoniste del corso: Genomica e Bioinformatica

Qualche definizione (wikipedia)..

Genomics

Genomics is a discipline in genetics that applies recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the function and structure of genomes (the complete set of DNA within a single cell of an organism)

Bioinformatics

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics and engineering to study and process biological data.

Gli obiettivi della genomica

La genomica è lo studio della struttura, del contenuto e dell'evoluzione dei genomi.

Non solo determinazione di sequenze, ma analisi di ESPRESSIONE, FUNZIONE e INTERAZIONE di geni e proteine.

Scopi condivisi dai Progetti genoma:

- **Sviluppo di un database e di un' interfaccia di ricerca**
- **Ottenere e combinare mappe fisiche e genetiche del genoma**
- **Generare e ordinare sequenze genomiche di geni espressi**
- **Identificare e annotare tutti i geni codificati da un determinato genoma**
- **Accumulare dati funzionali (caratteristiche biochimiche e fenotipiche dei geni)**
- **Caratterizzare la diversità di sequenza del DNA**
- **Fornire le risorse per eseguire le comparazioni tra genomi**

- Sviluppo di un database e di un' interfaccia di ricerca

Accesso all'enorme quantità di dati prodotti, congiunto all'aggregazione di dati prodotti da istituti di ricerca diversi: necessità di una piattaforma condivisa e aggiornabile

1001 Genomes
A Catalog of *Arabidopsis thaliana* Genetic Variation



Home Data Providers Accessions Tools Software Data Center About Help desk

Data Providers

Name ▾	Abbreviation ▾	Contact ▾	Collaborators ▾	Funding ▾
DOE Joint Genome Institute	JGI	Len Pennacchio	Joshua Heazlewood	DOE
Max Planck Institute for Developmental Biology	MPI	Detlef Weigel	Richard Clark (Utah) Karl Schmid (Hohenheim)	Max Planck Society, DFG
Salk Institute	Salk	Joe Ecker	Michael Egholm Tim Harkins (454) Dan Rokhsar (JGI) Olivier Loudet	Applied Biosystems, Chapman Foundation, Roche, National Science Foundation
Sainsbury Laboratory	SL	Jonathan Jones	Richard Mott (WTCGH)	
University of Lausanne	UNIL	Christian Hardtke		
Gregor Mendel Institute of Molecular Plant Biology	GMI	Magnus Nordborg		Austrian Academy of Sciences
Wellcome Trust Center for Human Genetics	WTCGH	Richard Mott	Paula Kover	BBSRC
Waksman Institute of Microbiology, Rutgers	WIM	Todd Michael		Charles and Johanna Busch Memorial Fund
Bielefeld University, CeBiTec	CeBiTec	Bernd Weisshaar	Thomas Altmann, IPK Gatersleben	UniBi & IPK
University of Warwick	Warwick	Eric Holub	Robin Allaby (Warwick) Jim Beynon (Warwick) Murray Grant (Exeter)	
University of Chicago	Chicago	Joy Bergelson		NIH
Monsanto Company	Monsanto	Mitchell Sudkamp	Todd Michael	Monsanto Research and Development

© Copyright 2008-2012 Max Planck Institute for Developmental Biology, Tübingen. <http://www.weigelworld.org> Disclaimer / Impressum Data Usage Policy

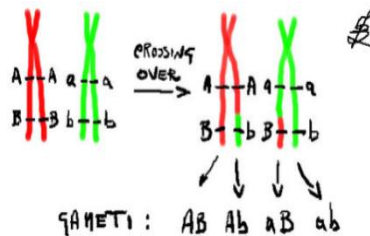
<http://www.1000genomes.org/participants>

Piuttosto che una semplice forma di archivio di sequenze o di altri dati, la maggior parte dei siti è organizzata su banche dati, di tipo relazionale, comprendenti programmi o software per la ricerca e l'analisi bioinformatica in tempo reale (direttamente on-line)

- **Ottenere e combinare mappe fisiche e genetiche del genoma**

Posizione dei geni nel genoma può essere determinata sia in base alla distanza fisica (distanza in paia di basi), sia rispetto alla posizione relativa, definita dalle frequenza di **ricombinazione***.

*Per ricombinazione genetica si intende ogni processo attraverso il quale, a partire da un genotipo, si ottengono nuove combinazioni di alleli rispetto a quelle iniziali.

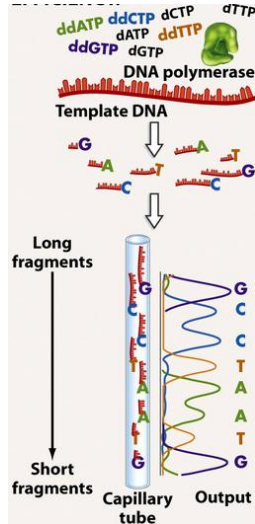


Queste informazioni sono necessarie per confrontare tra loro i genomi di diverse specie e per correlare i dati fenotipici con i dati genetici, in particolare nel caso di mappaggio di loci legati ad uno specifico tratto fenotipico (es. malattia)

- Generare e ordinare sequenze genomiche di geni espressi

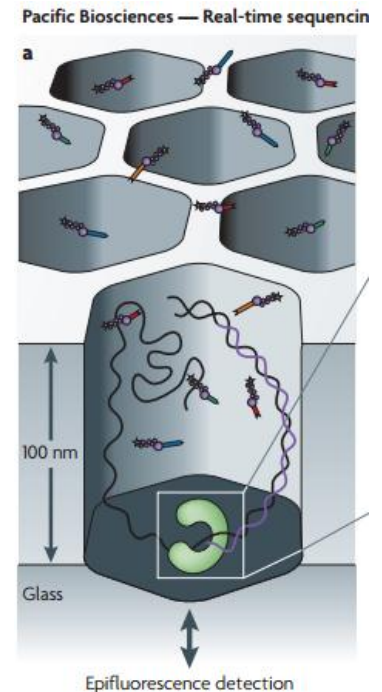
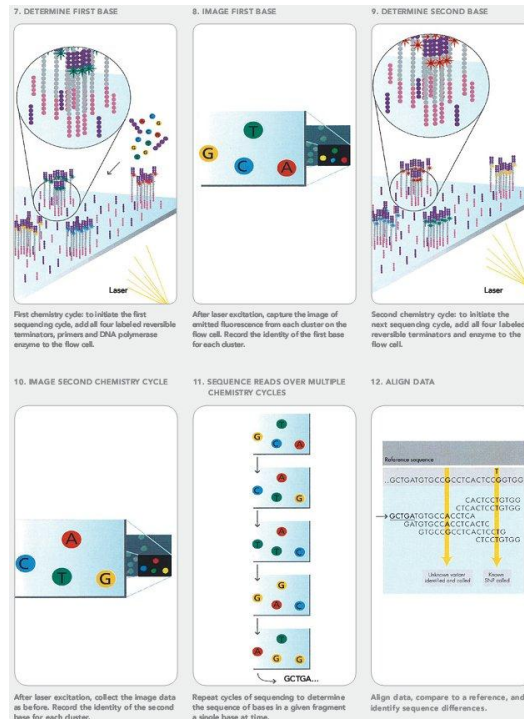
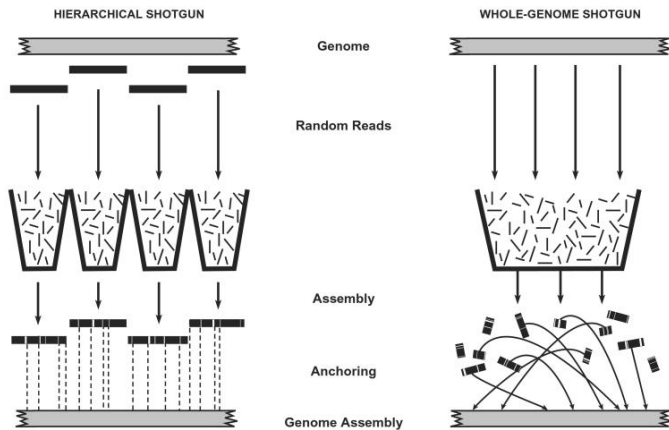
1. Sequenziamento genetico su larga scala:

metodologie
di prima
(Sanger seq)



seconda
NGS

e terza generazione
single molecule



2. Sequenziamento cloni di cDNA utilizzando mRNA come stampo per identificare GENI ESPRESSI o le sequenze che delimitano gli ESONI.

EST: *expressed sequence tag*

- **Identificare e annotare tutti i geni codificati da un determinato genoma**

L'ottenimento della sequenza completa di un genoma è solo il primo passo verso la caratterizzazione del suo contenuto genico. Bisogna identificare I GENI all'interno della sequenza con una combinazione di strategie sperimentali e bioinformatiche:

allineamento di sequenze genomiche e dei cDNA

ricerca di sequenze simili a quelle già identificate in altri genomi

applicazione di software che riconoscono caratteristiche del DNA associato a geni (Open Reading Frame, siti di inizio e fine trascrizione)

ANNOTAZIONE: la sequenza viene messa in relazione con tutti i dati di funzione ed espressione genica disponibili, con la presenza di fenotipi mutanti della proteina associata e con dati comparativi riguardanti proteine omologhe di altre specie

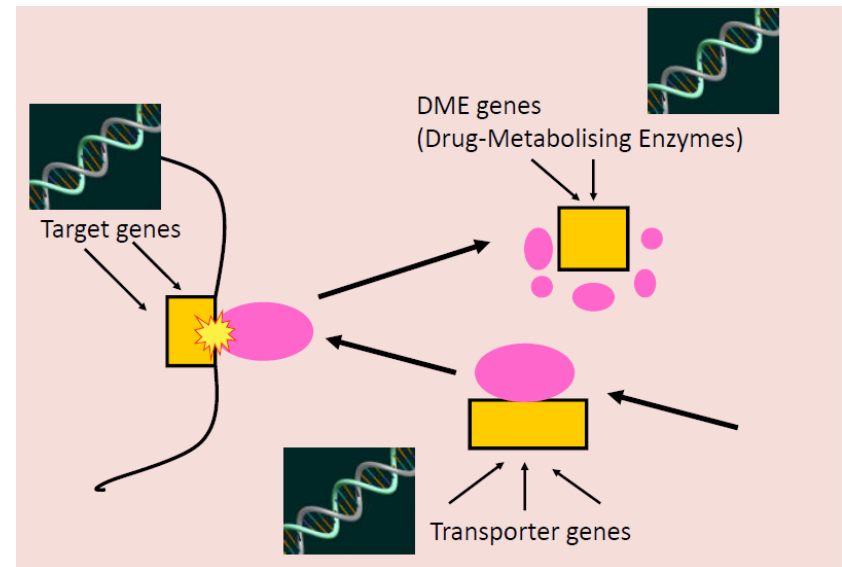
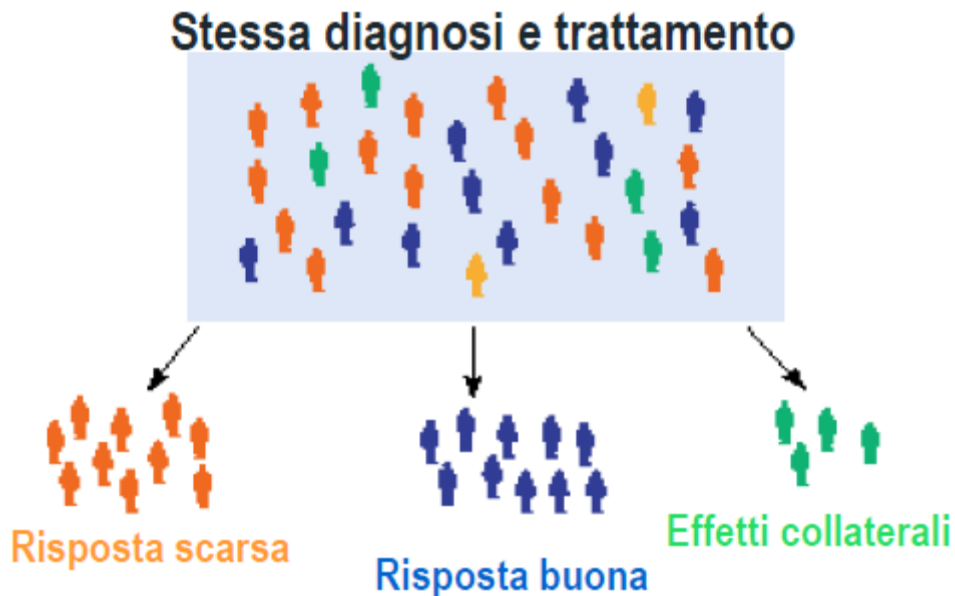
- **Accumulare dati funzionali (caratteristiche biochimiche e fenotipiche dei geni)**

Genomica Funzionale: metodologie destinate all'individuazione delle caratteristiche biochimiche, cellulari e fisiologiche di tutti i prodotti genici possibili

Proteomica: metodi per l'identificazione delle proteine espresse e delle interazioni tra proteine

Genomica strutturale: individuazione della struttura terziaria delle proteine presenti nelle cellule

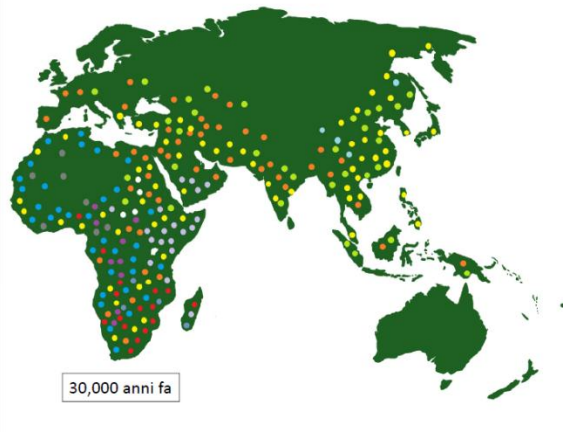
Farmacogenomica: interazioni tra farmaci, proteine di trasporto e metaboliti



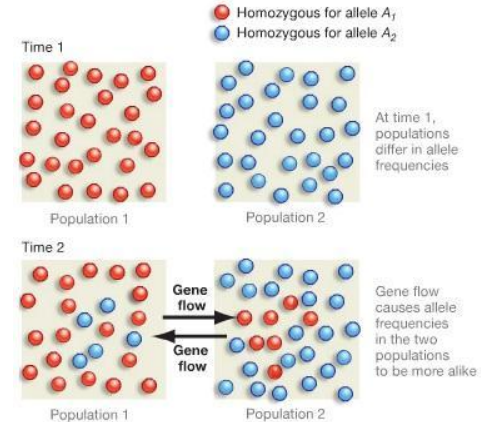
- Caratterizzare la diversità di sequenza del DNA

Identificazione dei polimorfismi genetici (SNP, indel,..) che caratterizzano i genomi

Livello di polimorfismo ci dice qualcosa sulla storia demografica della specie di cui stiamo studiando il genoma



Gene Flow



Bioinformatics

NSF – introduction

Large databases that can be accessed and analyzed with sophisticated tools have become central to biological research and education. The information content in the genomes of organisms, in the molecular dynamics of proteins, and in population dynamics, to name but a few areas, is enormous. Biologists are increasingly finding that the management of complex data sets is becoming a bottleneck for scientific advances. Therefore, **bioinformatics** is rapidly become a key technology in all fields of biology.

NSF – mission statement

The present bottlenecks in **bioinformatics** include the education of biologists in the use of advanced computing tools, the recruitment of computer scientists into this evolving field, the limited availability of developed databases of biological information, and the need for more efficient and intelligent search engines for complex databases.

Molecular Bioinformatics

Molecular Bioinformatics involves the use of computational tools to discover new information in complex data sets (from the **one-dimensional** information of DNA through the **two-dimensional** information of RNA and the **three-dimensional** information of proteins, to the **four-dimensional** information of evolving living systems).

Bioinformatics (Oxford English Dictionary):

The branch of science concerned with information and information flow in biological systems, esp. the use of computational methods in genetics and genomics.

The field of science in which **biology**, **computer science** and **information technology** merge into a single discipline

Biologists

collect molecular data:
DNA & Protein sequences,
gene expression, etc.

Bioinformaticians

Study biological questions by
analyzing molecular data

Computer scientists

(+Mathematicians, Statisticians, etc.)
Develop tools, softwares, algorithms
to store and analyze the data.

