



*Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni*

**VALUTAZIONE DI
COLLEZIONI
PUBBLICHE DI SNP E
VARIANTI GENICHE
PATOLOGICHE E
METODICHE PER LA
LORO SELEZIONE ED
ESTRAZIONE**

R. Cassandra, Mario R. Guarracino

RT-ICAR-NA-2013-6

Novembre 2013



Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR)
– Sede di Napoli, Via P. Castellino 111, I-80131 Napoli, Tel: +39-0816139508, Fax: +39-
0816139531, e-mail: napoli@icar.cnr.it, URL: www.na.icar.cnr.it



Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

VALUTAZIONE DI COLLEZIONI PUBBLICHE DI SNP E VARIANTI GENICHE PATOLOGICHE E METODICHE PER LA LORO SELEZIONE ED ESTRAZIONE¹

R. Cassandra², Mario R. Guarracino²

Rapporto Tecnico N.:
RT-ICAR-NA-2013-6

Data:
Novembre 2013

¹ Rapporto tecnico del laboratorio di Genomica, Trascrittomica e Proteomica GTP

² High Performance Computing and Networking Institute Italian National Research Council
Via P. Castellino, 111, 80131, Napoli (Italy)

I rapporti tecnici dell'ICAR-CNR sono pubblicati dall'Istituto di Calcolo e Reti ad Alte Prestazioni del Consiglio Nazionale delle Ricerche. Tali rapporti, approntati sotto l'esclusiva responsabilità scientifica degli autori, descrivono attività di ricerca del personale e dei collaboratori dell'ICAR, in alcuni casi in un formato preliminare prima della pubblicazione definitiva in altra sede.

PROGETTO PON_02_00619_3461281

VALUTAZIONE DI VARIANTI GENICHE PER LO STUDIO DI PATOLOGIE A TRASMISSIONE EREDITARIA, ATTRAVERSO L'ANALISI SU LARGA SCALA DI SEQUENZE GENOMICHE

ATTIVITA' 4.1 – VALUTAZIONE DI COLLEZIONI PUBBLICHE DI SNP E
VARIANTI GENICHE PATOLOGICHE E METODICHE PER LA LORO
SELEZIONE ED ESTRAZIONE

1



UNIONE EUROPEA
Fondo Europeo di Sviluppo Regionale



Sommario

1.	SOMMARIO ATTIVITA'	3
2.	L'IMPORTANZA DELL'INDIVIDUAZIONE DI SNPs NELLA RICERCA GENETICA	3
2.1.	IDENTIFICAZIONE DI SNPs PER LO STUDIO DI MALATTIE COMPLESSE	4
2.2.	SINGLE NUCLEOTIDE POLYMORPHISM E TAG SNPs	5
3.	INTRODUZIONE ALLE BACHE DATI BIOLOGICHE	8
3.1.	INTERROGAZIONE DELLE BANCHE DATI BIOLOGICHE	12
3.1.1.	INTERROGAZIONE SRS	13
3.1.2.	INTERROGAZIONE ENTREZ	13
3.1.3.	INTERROGAZIONE ACNUC	13
3.1.4.	INTERROGAZIONE ACEDB	13
4.	L'IMPORTANZA DELLE BANCHE DATI BIOLOGICHE NELLA RICERCA E ANALISI DI SNPs	14
4.1.	SNP DATABASE	14
4.2.	SNP DETECTION E PREDIZIONE DEGLI EFFETTI	21
4.3.	RIPRODUZIONE DEL NUMERO DI VARIAZIONI	24
4.4.	SNPs CHE CAUSANO MALATTIE	26
5.	UTILIZZO DI UNA BANCA DATI: ESEMPIO DI RICERCA SU CNV DATABASE	33
6.	DISCUSSIONI	40



1. SOMMARIO ATTIVITA'

Il Progetto Genoma Umano, ha reso disponibile la sequenza del DNA umano, rivelando che ogni individuo mostra il 99.5% d'identità genetica rispetto ad un qualsiasi altro individuo preso a caso nella popolazione. Il restante 0.5% di DNA è soggetto a variabilità individuale e mostra cambiamenti all'interno della popolazione; la somma di queste differenze costituisce la variabilità interindividuale. Questa è caratterizzata principalmente da variazioni di sequenza definite polimorfismi, vale a dire la presenza ad un dato locus di due o più alleli, presenti con una frequenza maggiore (>1%) di quella che potrebbe essere mantenuta da una mutazione¹. Lo studio della variabilità interindividuale rappresenta una sfida per la medicina moderna soprattutto nella prospettiva di poter curare il malato in maniera sempre più specifica e sicura, individuando il trattamento terapeutico più efficace.

In particolare lo studio delle varianti polimorfiche è diventato determinante nella comprensione dei meccanismi alla base della suscettibilità alle diverse patologie multifattoriali, tra cui rientrano malattie comuni quali l'asma, la psoriasi, il diabete, l'obesità, e le malattie cardiovascolari.

Sono stati individuati 46 banche dati WEB suddivise in 4 categorie:

- Categoria **SNP Databases**;
- Categoria **SNP detection and effect prediction** (Scoperta SNP e predizione degli effetti);
- Categoria **Copy number variation databases** (Database che riproducono il numero di variazioni);
- Categoria **Disease-causing variations** (Variazioni che causano malattie);

2. L'IMPORTANZA DELL'INDIVIDUAZIONE DI SNPs NELLA RICERCA GENETICA

Gli SNPs, sostituzioni di un singolo nucleotide, rappresentano la più grande fonte di variabilità interindividuale nel genoma dato che lo 0,5% di porzione variabile di sequenza è responsabile non solo delle differenze fenotipiche tra gli individui, ma soprattutto delle differenze in termini di predisposizione e resistenza alle malattie comuni.

In passato è stata formulata l'ipotesi CD = CV hypothesis "common disease/common variant"² per la quale le mutazioni (evento eccezionale) determinano le malattie rare (patologie mendeliane)

¹ Novelli e Giardina, 2003

² Becker, 2003



UNIONE EUROPEA
Fondo Europeo di Sviluppo Regionale



mentre gli SNPs (frequenti nel genoma) determinano la suscettibilità genetica alle malattie complesse. Le varianti polimorfiche sono alla base dell'eziologia patologica di molte malattie e andrebbero pertanto studiate su scala popolazionale piuttosto che su scala familiare. L'introduzione di innovativi studi genotipici su larga scala (WGA, Whole Genome Association Study) ha permesso l'identificazione di un nuovo repertorio di loci di suscettibilità di malattie complesse, con funzione fino ad oggi sconosciuta, caratterizzati da elevate frequenze alleliche e basso rischio relativo supportando maggiormente l'ipotesi $CD = CV^3$.

2.1. IDENTIFICAZIONE DI SNPs PER LO STUDIO DI MALATTIE COMPLESSE

Lo studio delle patologie complesse negli ultimi anni è passato dall'analisi specifica di un singolo locus selezionato all'analisi simultanea di più loci situati su cromosomi differenti.

Per identificare le regioni di suscettibilità alle patologie complesse, si utilizzano due approcci differenti: l'analisi di linkage e lo **studio di associazione**⁴.

In breve l'analisi di linkage è un'analisi di segregazione familiare che permette di assegnare un gene o un locus ad una determinata regione cromosomica definita da un insieme di marcatori polimorfi (i marcatori più utilizzati per questo tipo di analisi sono i microsatelliti, piccole sequenze di ripetute in tandem disperse uniformemente in tutto il genoma). Il linkage può essere studiato in famiglie estese che presentano una ricorrenza per la patologia in esame.

Se un microsatellite mappa vicino ad un gene-malattia in un determinato punto del genoma, ci si aspetta che tutti i membri affetti da quella patologia all'interno di una famiglia, ereditino lo stesso allele⁵ marcatore in linkage con l'allele responsabile del fenotipo patologico, specificatamente in quella famiglia.

Quando le famiglie studiate sono abbastanza estese da poter dimostrare che la co-localizzazione tra il gene-malattia e il marcatore non è un evento casuale, allora i due loci (malattia e marcatore) sono detti in linkage⁶. Così definito, il linkage indica che il gene coinvolto nella patologia in esame,

³ Hemmink et al, 2008

⁴ Studio di Associazione: Uno studio di associazione consiste nel confrontare la frequenza del fattore genetico (alleli, genotipi o aplotipi) in un gruppo di individui affetti rispetto ad un gruppo di individui non affetti. Lo studio di associazione caso-controllo può essere influenzato da diversi fattori come ad esempio il "mescolamento" di più popolazioni. La popolazione dei controlli dovrebbe essere scelta per essere il più possibile simile alla popolazione dei casi per tutti i possibili fattori confondenti (es. età, sesso, etnia, etc)

⁵ Allele: In genetica si definisce allele ogni variante di sequenza di un gene. Il genotipo di un individuo relativamente ad un gene è il corredo di alleli che egli si trova a possedere. In un organismo diploide, in cui sono presenti due copie di ogni cromosoma, il genotipo è dunque costituito da due alleli. Due cromosomi omologhi possiedono gli stessi geni, ma diverse forme alleliche (ad esempio, ognuno dei due possiede il gene che controlla il colore del bocciolo, ma ogni allele determinerà un colore diverso).

⁶ Risch, 2000



mappa vicino ad un marcatore e si delimita una regione minima entro la quale si procede per la ricerca del gene candidato, mediante clonaggio posizionale. Ne consegue che è possibile diagnosticare la presenza di una malattia senza conoscere effettivamente la mutazione o il gene coinvolto.

Nello studio delle patologie complesse si utilizza il linkage non-parametrico per il quale lo studio può essere effettuato su coppie di fratelli, su famiglie estese o su intere popolazioni⁷ senza che siano definiti dei parametri a priori (come il tipo di ereditarietà, la penetranza, la segregazione etc etc).

Rispetto all'analisi di linkage, gli studi di associazione confrontano la frequenza di un allele, o di un genotipo, in un campione di persone non imparentate tra loro e la frequenza del medesimo allele in un campione di controlli sani (Figura 1).

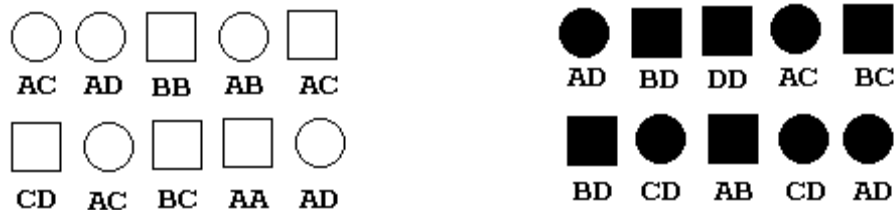


Figura 1: A sinistra sono visibili i controlli (oggetti con sfondo bianco) e a destra i casi (oggetti con sfondo nero). L'allele D è più frequente nei casi (riportati a destra) rispetto ai controlli (riportati a sinistra).

L'ipotesi alla base di uno studio di associazione è che la presenza di polimorfismi genetici (in particolare gli SNPs) sia correlata all'aumento o alla diminuzione del rischio di sviluppare patologie complesse; esistono varianti alleliche con ruolo di predisposizione alle malattie e varianti alleliche con un ruolo protettivo più frequenti negli individui sani.

Le difficoltà riscontrate in uno studio di associazione, soprattutto per quanto concerne la riproducibilità dei risultati ottenuti, derivano dalle differenze tra le popolazioni in esame, dalla disomogeneità dei metodi utilizzati nella definizione del fenotipo e dall'eterogeneità genetica della patologia in esame.

2.2. SINGLE NUCLEOTIDE POLYMORPHISM E TAG SNPs

Recenti ricerche (Genetic variation mapping project) hanno dimostrato che il 65-85% del genoma umano è organizzato in **blocchi "inscindibili"** di oltre 10.000 paia di basi. Ciascuno di questi blocchi può contenere 12 o più SNPs che tenderanno a rimanere "vicini" e quindi a trasmettersi

⁷ Strachan e Read, 1999



UNIONE EUROPEA
Fondo Europeo di Sviluppo Regionale



insieme, con il passare delle generazioni. Pertanto, per definire una mappa di queste variazioni, possiamo immaginare il nostro DNA non più come una serie di singoli punti (i nucleotidi del DNA) ma come una serie di blocchi ciascuno dei quali contiene fino a 10000 nucleotidi (e soprattutto i vari SNPs tra questi identificati): Il genoma è ereditato in segmenti e non in punti.

Inoltre è stato verificato che nel sito rappresentante un singolo blocco, sono disponibili tutte le regioni in cui è presente il fenomeno del linkage disequilibrium (LD). Il linkage disequilibrium indica la tendenza, tra specifici alleli relativi a due o più loci strettamente associati, a trovarsi insieme sullo stesso tratto cromosomico, in una popolazione con una frequenza maggiore rispetto a quella attesa sulla base delle singole frequenze alleliche. L'osservazione che l'allele malattia è in disequilibrium con l'allele marcatore consente di restringere notevolmente la regione entro la quale ricercare il gene causativo. E questa è la grande differenza che distingue il linkage dal linkage disequilibrium. Il linkage infatti prende vantaggio dalla ricombinazione all'interno di una famiglia, il linkage disequilibrium invece prende vantaggio da molti eventi di ricombinazione che accadono storicamente all'interno di una popolazione. Gli alleli in disequilibrium costituiscono infatti un particolare aplotipo ancestrale, perché trasmesso lungo la discendenza da un comune progenitore. Per questo motivo il linkage disequilibrium è maggiore in popolazioni omogenee, cioè originate da un nucleo di individui fondatori come la popolazione sarda o finlandese. Il linkage disequilibrium è un importante strumento per individuare regioni cromosomiche di limitata ampiezza in cui si collocano i geni per una data malattia (mappaggio ad alta risoluzione) e si avvale dell'analisi molecolare di varianti alleliche (per lo più di SNPs o STRs) che costituiscono aplotipi in soggetti tra loro apparentemente non imparentati. Infatti è prevedibile che pazienti che hanno ereditato lo stesso segmento cromosomico, definito dal medesimo aplotipo, abbiano ereditato anche la stessa mutazione in esso contenuto. Questa è la ragione per la quale il linkage ha una risoluzione di 10-20 cM⁸ mentre l'associazione con approccio tramite linkage disequilibrium ha una maggior risoluzione (0.1-0.2 cM).

Nello studio delle malattie multifattoriali il LD viene applicato per identificare, negli affetti per una specifica patologia, delle regioni cromosomiche ancestrali (aplotipi) definite da marcatori genetici che si trovano in vicinanza del locus/gene predisponente al fenotipo patologico. Si presuppone che tali aplotipi siano conservati (dopo essere stati introdotti nella popolazione) per un certo numero di generazioni (anche se le ricombinazioni tendono a ridurne l'estensione). All'interno di ciascun blocco di LD sono presenti fino a 70 SNPs che non soggetti a ricombinazione che pertanto tendono a rimanere vicini nello stesso locus e sono ereditati insieme.

⁸ cM: Il centimorgan (cM) è l'unità di misura della distanza genetica tra 2 loci. È impiegato nelle mappe genetiche (mappe cromosomiche calcolate attraverso l'utilizzo delle frequenze di ricombinazione). Due loci che presentano una frequenza di ricombinazione dell'1% sono definiti distanti 1 cM. Un centiMorgan equivale a un'unità di mappa (u.m.) (1cM = 1 u.m.). Un Morgan equivale a 100 cM e quindi a 100 u.m.



Ministero dell'Istruzione,
dell'Università e della Ricerca



Ministero
dello Sviluppo Economico





UNIONE EUROPEA
Fondo Europeo di Sviluppo Regionale



La presenza di questi blocchi di disequilibrium ha evidenti vantaggi: ha permesso, nelle analisi di associazione, di studiare non più un numero indefinito di polimorfismi all'interno di ogni regione, ma quegli SNPs (definiti appunto TAG - segnale) (Figura 2) necessari a identificare il blocco di DNA in disequilibrium ed il progetto HapMap inoltre ne ha garantito l'identificazione, la localizzazione e la genotipizzazione nelle popolazioni già menzionate.

Il ragionamento che porta a considerare **Tag SNPs**⁹ invece che i singoli SNPs non è difficile da comprendere, possiamo schematizzare il tutto in pochi passi. Alleli di SNPs associati definiscono l'**aplotipo**¹⁰. Gran parte delle regioni cromosomiche sono caratterizzate da aplotipi molto rari (frequenza max 5%). Tali regioni contengono diversi SNPs ma quelli che definiscono l'unicità dell'aplotipo sono chiamati Tag SNPs.

⁹ Tag SNPs: Un tag SNP è un single nucleotide polymorphism (SNP) in una regione del genoma con un alto "disequilibrio di associazione" (l'associazione non random di alleli a due o più loci). E' possibile identificare variazioni genetiche senza genotipizzare ogni SNP in una regione cromosomica. I tag SNP sono utili in studi di associazione su interi genomi in cui centinaia o migliaia di SNPs devono essere genotipizzati rispetto all'intero genoma. Per questa ragione, l'International HapMap Project spera di utilizzare i tag SNPs per scoprire geni responsabili per diverse patologie.

¹⁰ Aplotipo: Insieme di sequenze relative ad una definita regione genomica riportanti un set di polimorfismi completamente coincidenti rispetto ad un riferimento.





UNIONE EUROPEA
Fondo Europeo di Sviluppo Regionale



modo tale da consentire l'uso dei dati stessi (e il loro aggiornamento) da parte di applicazioni software. Una banca dati biologica raccoglie informazioni e dati che possono essere derivati dalla letteratura o da analisi effettuate in laboratorio (analisi *in vitro* o *in vivo*) oppure attraverso applicazioni di analisi bioinformatiche, dette analisi *in silico* (viene utilizzato il termine “in silico”, in quanto i processori dei calcolatori sono costituiti da silicio) e dalla letteratura scientifica. Le banche dati sono progettate come contenitori costruiti per immagazzinare dati in modo efficiente e razionale al fine di renderli facilmente accessibili a tutti gli utenti: ricercatori, medici, studenti, etc. Ogni banca dati è caratterizzata da un elemento biologico centrale che costituisce l'oggetto intorno al quale viene costruita la *entry* principale della banca dati. Una *entry* di una banca dati di sequenze nucleotidiche potrebbe contenere, oltre alla sequenza di una molecola di DNA, il nome dell'organismo cui la sequenza appartiene, la lista degli articoli che riportano dati su quella sequenza, le caratteristiche funzionali (cioè si tratta di un gene o di una sequenza non codificante) e ogni altra informazione ritenuta di interesse. Di seguito si presentano due esempi di *entry*:

- In una banca dati di sequenze di acidi nucleici l'elemento centrale è la sequenza nucleotidica di DNA o RNA a cui si associano annotazioni con le quali si classifica l'elemento come ad esempio il nome della specie, la funzione, le referenze bibliografiche, ecc.
- In una banca dati dei promotori eucariotici l'elemento centrale è il promotore. Ogni *entry* racchiude quindi le informazioni che caratterizzano l'elemento, cioè gli attributi dell'elemento centrale.

Per definire la struttura di una banca dati si definiscono gli attributi e il formato con cui queste informazioni verranno organizzate. La maggior parte delle banche dati biologiche possono essere usate dalla comunità scientifica in formato *flat-file*, cioè un file sequenziale in cui ogni classe di formazione è riportata su una o più linee consecutive identificate da un codice a sinistra che caratterizza gli attributi annotati sulla linea (Figura 3).





Display Settings: GenBank

Homo sapiens tumor susceptibility gene 101 (TSG101), mRNA

NCBI Reference Sequence: NM_006292.3

[FASTA](#) [Graphics](#)

Go to:

```

LOCUS       NM_006292                1562 bp    mRNA    linear   PRI 25-NOV-2012
DEFINITION  Homo sapiens tumor susceptibility gene 101 (TSG101), mRNA.
ACCESSION   NM_006292
VERSION     NM_006292.3  GI:332000018
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1 (bases 1 to 1562)
AUTHORS    Kameyama,T., Suzuki,H. and Mayeda,A.
TITLE      Re-splicing of mature mRNA in cancer cells promotes activation of
            distant weak alternative splice sites
JOURNAL    Nucleic Acids Res. 40 (16), 7896-7906 (2012)
PUBMED    22675076
REMARK     GeneRIF: The results provide evidence for a two-step splicing
            pathway of the TSG101 mRNA in which the initial constitutive
            splicing removes all 14 authentic splice sites, thereby bringing
            the weak alternative splice sites into close proximity.
REFERENCE   2 (bases 1 to 1562)
AUTHORS    Gu,R.J., Wang,S.C., Sun,G., Zhuang,B.W. and Liu,D.L.
TITLE      [Expression and significance of tumor susceptibility gene 101 in
            hepatocellular carcinoma tissues]
JOURNAL    Xi Bao Yu Fen Zi Mian Yi Xue Za Zhi 28 (7), 738-740 (2012)
PUBMED    22768867
REMARK     GeneRIF: The expression of TSG101 in HCC is higher than that in
            corresponding non-cancer tissues and the expression level is
            closely correlated with TNM stage and metastasis of HCC.
REFERENCE   3 (bases 1 to 1562)
AUTHORS    Horgan,C.P., Hanscom,S.R., Kelly,E.E. and McCaffrey,M.W.
TITLE      Tumor susceptibility gene 101 (TSG101) is a novel binding-partner
            for the class II Rab11-FIPs
JOURNAL    PLoS ONE 7 (2), E32030 (2012)
PUBMED    22348143
REMARK     GeneRIF: identified TSG101 as a novel FIP4-binding protein, which
            can also bind FIP3. alpha-helical coiled-coil regions of both
            TSG101 and FIP4 mediate the interaction with the cognate protein
REFERENCE   4 (bases 1 to 1562)
AUTHORS    Nagashima,S., Takahashi,M., Jirintai,S., Tanaka,T., Nishizawa,T.,

```

Figura 3: Esempio di flat file.

Questo formato è molto utilizzato perché è molto leggibile e analizzabile con programmi che estraggono dalla banca dati informazioni specifiche. Prima tutte le banche dati biologiche erano in formato *flat-file*, oggi invece si usano i DBMS ovvero i Database Management System per disegnare banche dati sempre più complesse. Le banche dati biologiche vengono solitamente suddivise in due categorie principali:

- **banche dati primarie** (es. banche dati nucleotidiche, proteiche);
- **banche dati derivate.**



Nelle banche dati primarie sono presenti solo le informazioni minime necessarie da associare ai dati per identificarli al meglio. Le banche dati derivate contengono invece insiemi di dati omogenei che possono derivare da banche dati primarie, ma rivisti e annotati con varie informazioni che danno un valore aggiunto alla banca dati stessa. Ogni banca dati poi può essere:

- **Non Curata;**
- **Curata.**

Le banche dati non curate contengono i dati grezzi così come sono forniti da chi li ha ottenuti, o con annotazioni da sistemi automatici. Le banche dati curate presentano informazioni che sono verificate, confrontate con quelle di altre banche dati, opportunamente corrette (o per lo meno con segnalazione di possibili errori e conflitti con altri dati).

Con il crescere dei dati si è reso necessario adottare DBMS e si è abbandonato l'utilizzo dei *flat file*. Inoltre con l'avvento del web 2.0 è possibile accedere a informazioni tra loro correlate (*cross-referencing*) attraverso link ipertestuali, uno schema di *cross-referencing* tra banche dati è visibile in figura 4.

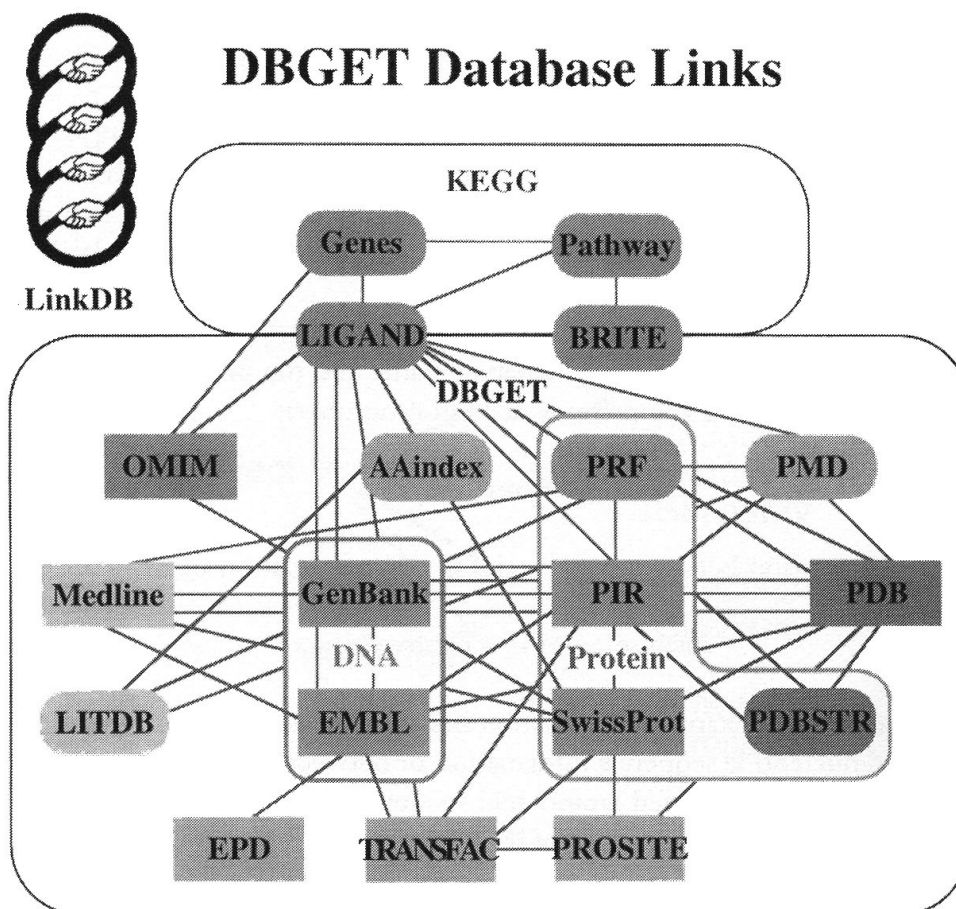


Figura 4: Schema di *cross-referencing* tra banche dati biologiche.

Negli ultimi tempi sono state anche implementate banche dati in formato XML.

3.1. INTERROGAZIONE DELLE BANCHE DATI BIOLOGICHE

Lo scopo di interrogare una banca dati è quello di ottenere informazioni da esse, attraverso sistemi informatici, e da altre banche dati cui è correlata. Uno dei principali problemi legati alle banche dati biologiche è quello della nomenclatura. Non esiste uno standard nell'assegnazione di nomi ai geni; uno stesso gene può avere diversi nomi (Es. TRF2 è anche noto come TLP o TLF), o uno stesso nome può individuare diversi geni (Es. TRF sta per TBP *Related Factor* ma anche per Transferrina o ancora per *Telomeric Repeat Binding Factor*). Occorre quindi un modo per individuare univocamente i geni e le proteine, e per gestire la grande quantità di informazioni ad essi legate: Nelle banche dati primarie ogni elemento (gene, sequenza, ecc) è individuato univocamente da un **accession number**. Per realizzare l'estrazione di dati esistono vari sistemi fra cui i più efficienti

sono **SRS** ed **ENTREZ**. Altri sistemi altrettanto validi sono **ACNUC** e **AceDB**. L'interrogazione di una banca dati può avvenire in maniera banale, inserendo il nome ricercato in una finestra di tipo text-search oppure tramite la sottomissione di *forms* in cui inserire varie informazioni sulla nostra ricerca. La logica di criterio è quella booleana che effettua intersezioni (operatore AND), somme (operatore OR), ed esclusioni (operatore BUT NOT), di insiemi di dati.

3.1.1. INTERROGAZIONE SRS

SRS (Sequence Retrieval System) è un sistema per la ricerca e l'estrazione di dati biologici via web. Esso consente di interrogare più banche dati differenti purché abbiano almeno un'informazione comune. SRS inoltre consente la navigazione attraverso varie banche dati sfruttando il *cross-referencing*. Può essere installato su diversi server e interagire con altri server SRS o altre banche dati, con pochi accorgimenti.

3.1.2. INTERROGAZIONE ENTREZ

Entrez è un sistema disponibile via web per la ricerca e l'estrazione di dati da banche dati di sequenze nucleotidiche o proteiche, dalla banca dati bibliografica Medline, dalla banca dati delle malattie mendeliane OMIM, o da risorse gnomiche. Tramite Entrez è anche possibile esplorare la classificazione degli organismi come riportata in Taxonomy o su ogni altra banca dati specializzata sviluppata all'NCBI. Entrez, a differenza di SRS, è una shell chiusa in cui non è possibile scaricare via internet, o ottenere un software che gestisce l'intero sistema, né è possibile duplicare il sito su altri computer, né installare proprie banche dati personali. Per effettuare la ricerca bisogna scegliere una categoria e poi usare gli operatori logici AND, OR, BUT NOT. Si può usare la funzione Limits per limitare la ricerca ad alcuni criteri. Il comando History visualizza tutti i risultati di una query relativi ad una categoria, che possono essere salvati col comando text.

3.1.3. INTERROGAZIONE ACNUC

ACNUC è un sistema disponibile su mainframe con sistemi operativi linux o VMS. Consente l'estrazione dei dati dalle banche dati di sequenze di acidi nucleici (EMBL o GenBank) o proteiche (SWISSPROT). Si possono ricercare dati di una sola categoria per volta. I dati possono essere selezionati coi comandi *Select* o *Find*. Coi comandi *Names*, *Short* e *Info* si possono visualizzare o stampare i risultati ottenuti con select o find. ACNUC ha il vantaggio di poter estrarre sottosequenze omogenee definite attraverso le Feature tables.

3.1.4. INTERROGAZIONE ACEDB

AceDB era stato sviluppato inizialmente per la gestione dei dati di mappaggio e sequenziamento del genoma *Caenorhabditis elegans*. Oggi è adottato per altri progetti genomici. AceDB comprende programmi per la strutturazione in formato AceDb di nuove banche dati per l'interrogazione e



UNIONE EUROPEA
Fondo Europeo di Sviluppo Regionale



l'analisi dei dati in AceDB. Si può scaricare il pacchetto con questi programmi per ricercare dei dati o anche per aggiornare il database via web.

4. L'IMPORTANZA DELLE BANCHE DATI BIOLOGICHE NELLA RICERCA E ANALISI DI SNPs

L'annotazione nelle banche dati di eventi generativi di mutazioni e polimorfismo è di rilevante importanza sia per studi di genetica di popolazione sia per studi di associazione fra mutazione e fenotipi con diversificate manifestazioni cliniche. Oltre 9 milioni di SNPs sono raccolti in appositi database. La maggior parte di questi SNPs sono stati identificati mediante sovrapposizione di sequenze durante le fasi del progetto genoma umano o successivamente (progetto HAPMAP¹¹).

4.1. SNP DATABASE

Per studiare la variabilità popolazionale in modo coordinato sono stati creati diversi DB come illustrato in tabella 1. Le due banche dati più utilizzate tra quelle elencate sono:

- Il database HGVbase, che annota tutti i dati derivati da studi di variabilità popolazionale;
- Il database dbSNPs che annota dati di SNPs, ma anche polimorfismi di regioni e mutazioni associate all'insorgenza di una specifica patologia.

¹¹ Per ulteriori dettagli e informazioni sul progetto consultare <http://hapmap.ncbi.nlm.nih.gov/>.



DB NAME	DESCRIPTION	BIBLIOGRAPHY
ALFRED	The ALlele FREquency Database (ALFRED) is designed to make allele frequency data on human population samples readily available for use by the scientific and educational communities.	1. Cheung KH, Miller PL, Kidd JR, Kidd KK, Osier MV, Pakstis AJ. "ALFRED: a Web-accessible allele frequency database". Pac Symp Biocomput 2000.:639-50.
Barley SNP Database	This online database contains information from a project at the SCRI to mine wheat and barley genes for SNPs which were mapped in barley crosses.	1. Rostoks N, Mudie S, Cardle L, Russell J, Ramsay L, Booth A, Svensson JT, Wanamaker SI, Walia H, Rodriguez EM, Hedley PE, Liu H, Morris J, Close TJ, Marshall DF, Waugh R (2005) - "Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress." Mol. Genet. Genomics 1-13
dbSNP	Database of single nucleotide polymorphisms (SNPs) and multiple small-scale variations that include insertions/deletions, microsatellites, and non-polymorphic variants.	1. Stephen T. Sherry, Minghong Ward, and Karl Sirotkin - dbSNP— Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation - 10.1101/gr.9.8.677 Genome Res. 1999. 9: 677-679
dbQSNP	A database of SNPs in human promoter regions with allele frequency information determined by single-strand conformation polymorphism-based methods.	1. Tahira T, Okazaki Y, Miura K, Yoshinaga A, Masumoto K, Higasa K, Kukita Y, Hayashi K. (2006) QSNPlite, a software system for quantitative analysis of SNPs based on capillary array SSCP analysis. Electrophoresis. 27: 3869-3878.
ENSEMBL	The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.	

FESD II	FESD is a web-based integrated database for selecting sets of SNPs in putative functional elements in human gene.	1. Hyo Jin Kang, Kyoung Oak Choi, Byung-Dong Kim, Sangsoo Kim and Young Joo Kim. FESD: a Functional Element SNPs Database in human, <i>Nucleic Acids Research</i> , 2005, Vol. 33, Database issue D518-D522
Forensic SNP Information	This site is intended to provide general information on single nucleotide polymorphism (SNP) markers that may be of interest in human identification applications. Many of these markers come from The SNP Consortium (TSC) efforts or are already present in the NCBI dbSNP database.	
F-SNP	F-SNP database provides integrated information about the functional effects of SNPs obtained from 16 bioinformatics tools and databases. The functional effects are predicted and indicated at the splicing, transcriptional, translational, and post-translational level. As such, the F-SNP database helps identify and focus on SNPs with potential pathological effect to human health.	<ol style="list-style-type: none"> 1. The F-SNP database has been described in the 2008 database issue of <i>Nucleic Acid Research</i>, "F-SNP: Computationally Predicted Functional SNPs for Disease Association Studies" by Phil H. Lee and Hagit Shatkay. 2. Detailed information on the F-SNP Scoring approach has been described in <i>Bioinformatics</i> 25(8): 1048-1055 (2009), "An integrative scoring system for ranking SNPs by their potential deleterious effects" by Phil H. Lee and Hagit Shatkay. 3. The preliminary version of the F-SNP database has been used in the research work, "Two Birds, One Stone: Selecting Functionally Informative Tag SNPs for Disease Association Studies" by Phil H. Lee and Hagit Shatkay, in the Proceedings of the 7th Workshop on Algorithms in Bioinformatics (WABI 2007), Springer series Lecture Notes in Bioinformatics, LNBI 4645, pp. 61-72, and in "Ranking Single Nucleotide Polymorphisms by Potential Deleterious Effects", by P. H. Lee and H. Shatkay in Proc. of the Annual Symp. of the American Medical Informatics Association (AMIA'08). November, 2008.

GVS (Genome Variation Server)	<p>The Genome Variation Server (GVS), fed by a local database, enables rapid access to human genotype data found in dbSNP, and provides tools for analysis of genotype data. The current release of genotype data found in the GVS database is that of dbSNP build 137 (June 2012). The variation locations are mapped to the human genome reference sequence of February 2009 (UCSC hg19, NCBI build 37). This GVS database contains 11.8 million variations with corresponding genotype data. To be included in our database, a variation must have genotype data, and it must be uniquely mapped to the human genome by dbSNP. As most submitters to dbSNP report double genotypes for X and Y chromosome variations, we put double genotypes in our database, and changed single genotypes to (homozygous) double genotypes. If a genotype on the Y chromosome was reported to be heterozygous, we omitted it. We have not corrected the frequencies for male X chromosome genotypes.</p>	
JSNP Database	<p>SNP Database Network of three databases which have been developed by the millennium project has been released. These databases are Database of Japanese Single Nucleotide Polymorphism for Geriatric Research (Tokyo Metropolitan Geriatric Medical Center), Human Mitochondrial Genome Single Nucleotide Polymorphism Database (Gifu International Institute of Biotechnology) and Protein Polymorphism Database (National Institute of Radiological Sciences).</p>	<p>1. JSNP: a database of common gene variations in the Japanese population <i>Nucleic Acids Research</i>, 30:158-162, 2002 PubMed 2. Haga H, Yamada R, Ohnishi Y, Nakamura Y, Tanaka T. Gene-based SNP discovery as part of the Japanese Millennium Genome Project : identification of 190,562 genetic variations in the human genome. <i>Journal of Human Genetics</i>, 2002;47(11):605-610 PubMed</p>
GWAS Central	<p>GWAS Central provides a centralized compilation of summary level findings from genetic association studies, both large and small. We actively gather datasets from public domain projects, and encourage direct data submission from the community.</p>	

<p>HGMD</p>	<p>The Human Gene Mutation Database (HGMD®) represents an attempt to collate known (published) gene lesions responsible for human inherited disease.</p>	<ol style="list-style-type: none"> 1. Stenson PD, Ball EV, Mort M, Phillips AD, Shaw K, Cooper DN. The Human Gene Mutation Database (HGMD) and Its Exploitation in the Fields of Personalized Genomics and Molecular Evolution. Curr Protoc Bioinformatics Unit 1.13, 2012. 2. David N. Cooper, Peter D. Stenson and Nadia A. Chuzhanova: The human gene mutation database (HGMD) and its exploitation in the study of mutational mechanisms. Current Protocols in Bioinformatics, Unit 1.13, 2005.+ 3. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeyasinghe S, Krawczak M, Cooper DN: Human Gene Mutation Database (HGMD®): 2003 update. Hum Mutat 21(6): 577-581, 2003. 4. Krawczak M, Ball EV, Stenson PD, Cooper DN: HGMD: The human gene mutation database. Chapter 8 in Bioinformatics; Databases and Systems. Ed. SI Letovsky. pp 99-104, Kluwer Academic Publishers, Boston, 1999. 5. Krawczak M, Cooper DN: The human gene mutation database (HGMD). Genome Digest 3: 7-8, 1996.
<p>MGI</p>	<p>MGI is the international database resource for the laboratory mouse, providing integrated genetic, genomic, and biological data to facilitate the study of human health and disease.</p>	
<p>PhenCode</p>	<p>PhenCode is a collaborative project to better understand the relationship between genotype and phenotype in humans. It connects human phenotype and clinical data in various locus-specific mutation databases (LSDBs) with data on genome sequences, evolutionary history, and function in the UCSC Genome Browser.</p>	

PolymiRTS	PolymiRTS (Polymorphism in microRNAs and their TargetSites) is a database of naturally occurring DNA variations in microRNA (miRNA) seed regions and miRNA target sites. MicroRNAs pair to the transcripts of protein-coding genes and cause translational repression or mRNA destabilization. SNPs and INDELs in miRNAs and their target sites may affect miRNA-mRNA interaction, and hence affect miRNA-mediated gene repression.	1. Bhattacharya A, Ziebarth JD, Cui Y. PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. Nucleic Acids Res. 2013; doi: 10.1093/nar/gkt1028. 2. Ziebarth JD, Bhattacharya A, Chen A, Cui Y. PolymiRTS Database 2.0: linking polymorphisms in microRNA target sites with human diseases and complex traits. Nucleic Acids Res. 2012;40 (D1):D216-221. PMID: 22080514.
SeattleSNPs	SeattleSNPs is funded as part of the National Heart Lung and Blood Institute's (NHLBI) Programs for Genomic Applications (PGA). The SeattleSNPs PGA is focused on identifying, genotyping, and modeling the associations between single nucleotide polymorphisms (SNPs) in candidate genes and pathways that underlie inflammatory responses in humans.	1. SeattleSNPs. NHLBI Program for Genomic Applications, SeattleSNPs, Seattle, WA (URL: http://pga.gs.washington.edu)
SCAN DB	SCAN is a large-scale database of genetics and genomics data associated to a web-interface and a set of methods and algorithms that can be used for mining the data in it.	
SNP500Cancer Database	The goal of the SNP500Cancer project is to resequence 102 reference samples to find known or newly discovered single nucleotide polymorphisms (SNPs) which are of immediate importance to molecular epidemiology studies in cancer. SNP500Cancer provides a central resource for sequence verification of SNPs.	
SNAP (SNP Annotation and Proxy Search)	SNAP finds proxy SNPs based on linkage disequilibrium, physical distance and/or membership in selected commercial genotyping arrays. Pair-wise linkage disequilibrium is pre-calculated based on phased genotype data from the International HapMap Project. Information about the genotyping arrays is based on data published by the vendors. SNAP can also generate linkage disequilibrium plots, like the one shown at the right. To generate plots, click on the Plots tab above and select plotting options.	1. Johnson, A. D., Handsaker, R. E., Pulit, S., Nizzari, M. M., O'Donnell, C. J., de Bakker, P. I. W. - SNAP: A web-based tool for identification and annotation of proxy SNPs using HapMap - Bioinformatics, 2008 24(24):2938-2939

SNP Control Database	The SNP control database currently contains genome wide one million SNPs of 700 samples to evaluate the appropriate thresholds of quality controls and to use the high quality SNP data for case-control study. Allele frequencies and genotype frequencies and Hardy Weinberg Equilibrium test results and annotations of each SNP and SNP call rate are stored.	
SNPedia	SNPedia is a wiki investigating human genetics. We share information about the effects of variations in DNA, citing peer-reviewed scientific publications. It is used by Promethease to analyze and help explain your DNA.	
SNPper	Retrieve known SNPs by position or by association with a gene; save, filter, analyze, display or export SNP sets; explore known genes using names or chromosome positions.	<ol style="list-style-type: none"> 1. A. Riva, I. S. Kohane, A SNP-centric database for the investigation of the human genome, BMC Bioinformatics 2004, 5:33 2. A. Riva, I. S. Kohane, SNPper: retrieval and analysis of human SNPs, Bioinformatics, 2002 18: 1681-1685. 3. A. Riva, I. S. Kohane, A Web-Based Tool to Retrieve Human Genome Polymorphisms from Public Databases, Proc. AMIA Symp. 2001:558-562
SNP@WEB	SNP@WEB is a web-based catalog of databases and tools for SNP studies. Currently, SNP@WEB collected ~90 SNP resources classified into eight categories (SNP acquisition, annotation, tagSNP, haplotype, population, mutability, database, and SNP effect). SNP@WEB is developed as a Wiki (real-time editable website) system, it supports fully open curation and communication among SNP researchers.	
Tagger	Tagger is a tool for the selection and evaluation of tag SNPs from genotype data such as that from the International HapMap Project. It combines the simplicity of pairwise tagging methods with the efficiency benefits of multimarker haplotype approaches.	1. P.I.W. de Bakker, R. Yelensky, I. Pe'er, S.B. Gabriel, M.J. Daly, D. Altshuler (2005) Efficiency and power in genetic association studies. Nature Genetics. 37: 1217-1223

Tabella 1: Elenco delle banche dati biologiche che contengono collezioni di SNP.



UNIONE EUROPEA
Fondo Europeo di Sviluppo Regionale



4.2. SNP DETECTION E PREDIZIONE DEGLI EFFETTI

Una volta individuate le differenze fra le due sequenze nucleotidiche (quella del paziente e quella di un individuo sano) è necessario valutarne gli effetti a livello della proteina, in quanto i vari tipi di mutazione possono avere effetti molto diversi. La predizione degli effetti di una mutazione *missense* a livello della struttura della proteina è un processo complesso, in cui bisogna tenere conto di tutte le caratteristiche degli amminoacidi della catena proteica e delle loro reciproche interazioni. Attualmente sono disponibili banche dati in cui è possibile estrarre questa informazione e alcuni software in grado di elaborare tutti questi dati e fornire una previsione di come la mutazione influenzerà la struttura tridimensionale della proteina.



DB NAME	DESCRIPTION	BIBLIOGRAPHY
GenEpi Toolbox	These tools have been collected during several recent genetic epidemiological studies and have proven to be very useful to interpret the often surprising and unexpected findings of GWA studies. Since many of these tools may not be familiar to genetic epidemiologists who are not intensively involved in bioinformatics, we offer here a collection of bioinformatic tools addressing many different issues that a researcher in the field of genetic epidemiology may encounter. This should help to avoid not to see the wood for the trees and get lost in space. Besides several types of “classic” tools for the analysis of the effects of single SNPs and mutations, we report also tools for the identification of known and unknown regulatory elements, data and literature, analysis of large genomic loci, tools for the annotation and selection of candidate SNPs, and, finally, several general databases.	1. Coassin S., Brandstätter A., Kronenberg F.: Lost in the space of bioinformatic tools: a constantly updated survival guide for genetic epidemiology. The GenEpi Toolbox. <i>Atherosclerosis</i> 209: 321-335 (2010)
pfSNP	Potentially functional SNP Search Engine.	1. Lee et al. pfSNP: An integrated potentially functional SNP resource that facilitates hypotheses generation through knowledge syntheses. <i>Human Mutation</i> , 2011 Jan;32(1):19-24. doi: 10.1002/humu.21331
PolyPhen-2	PolyPhen-2 (Polymorphism Phenotyping v2) is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations.	1. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. <i>Nat Methods</i> 7(4):248-249 (2010).

Pupasuite 3	PupaSuite is an interactive web-based SNP analysis tool that allows for the selection of relevant SNPs within a gene, based on different characteristics of the SNP itself, such as validation status, type, frequency/population data and putative functional properties (pathological SNPs, SNPs disrupting potential transcription factor binding sites, intron/exon boundaries...). Also, PupaSuite provides information about LD parameters (based on genotype data from HapMap) and identifies haplotype blocks and tag SNPs (using the Haploview software).	<ol style="list-style-type: none"> 1. Conde L, Vaquerizas JM, Dopazo H, Arbiza L, Reumers J, Rousseau F, Schymkowitz J and Dopazo J (2006) PupaSuite: finding functional SNPs for large-scale genotyping purposes Nucl Acids Research, 2006, 34: W621-W625 2. Conde, L., Vaquerizas, J.M., Ferrer-Costa, C., Orozco, M. & Dopazo, J. (2005) PupasView: a visual tool for selecting suitable SNPs, with putative pathologic effect in genes, for genotyping purposes Nucleic Acids Research 2005 33 (Web Server issue):W501-W505; 3. Conde, L., Vaquerizas, J.M., Santoyo, J., Al-Shahrour, F., Ruiz-Llorente, S., Robledo, M. & Dopazo, J. (2004) PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level Nucleic Acids Research 32 (Web Server issue), W242-W248.
SNAP	SNAP is a method for evaluating effects of single amino acid substitutions on protein function.	1. Yana Bromberg and Burkhard Rost - SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Research, 2007, Vol. 35, No. 11 3823-3835 (PubMed)
SNP2Prot	A tool to map human DNA variation onto proteins.	

Tabella 2: Elenco delle banche dati utilizzate per SNP detection.

4.3. RIPRODUZIONE DEL NUMERO DI VARIAZIONI

Le mutazioni sono cambiamenti rari, improvvisi e casuali del patrimonio ereditario di un individuo che determinano caratteristiche genotipiche non presenti nei genitori. Una mutazione si definisce quindi come una modificazione stabile ed ereditabile del materiale genetico. La mutazione può avvenire a diversi livelli, può modificare l'assetto dell'intero genoma a causa di una variazione nel numero di cromosomi (mutazione genomica), può riguardare un singolo cromosoma a causa di variazioni a livello della sua struttura (mutazione cromosomica), oppure può riguardare un solo gene (mutazione genica). Per questo è fondamentale l'esistenza di banche dati che partendo da una sequenza (un gene) ricerchi tutte le possibili variazioni (mutazioni) di quel gene conosciute in letteratura. In tabella 2 sono elencate le principali banche dati che cercano di riprodurre il numero di variazioni partendo da una sequenza nota.

DB NAME	DESCRIPTION	BIBLIOGRAPHY
DGV	The objective of the Database of Genomic Variants is to provide a comprehensive summary of structural variation in the human genome. We define structural variation as genomic alterations that involve segments of DNA that are larger than 50bp. The content of the database is only representing structural variation identified in healthy control samples. The Database of Genomic Variants provides a useful catalog of control data for studies aiming to correlate genomic variation with phenotypic data. The database is continuously updated with new data from peer reviewed research studies. We always welcome suggestions and comments regarding the database from the research community.	Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. Detection of large-scale variation in the human genome. Nat Genet. 2004 Sep;36(9):949-51.
ECARUCA	ECARUCA is a database which collects and provides cytogenetic and clinical information on rare chromosomal disorders, including microdeletions and microduplications. ECARUCA aims to be a database that is easily accessible for all participants and encourages information exchange as well as exchange of technical knowledge. ECARUCA wants to improve patient care and collaboration between genetic centres in the field of clinical cytogenetics. ECARUCA collects the results of cytogenetic research & the accompanying clinical features, but NOT the patient material used for the analysis; this stays in the centre where the research was carried out.	

Tabella 3: Elenco delle banche dati utilizzate per la ricerca del numero di variazioni.

4.4. SNPs CHE CAUSANO MALATTIE

La motivazione più importante nello studio di SNP è quella di comprendere gli effetti che una singola mutazione possono avere sull'organismo. Il primo elemento da considerare, per valutare gli effetti di una mutazione genica, è quale parte del gene coinvolga. Mutazioni che colpiscano la regione codificante hanno effetti sulla traduzione determinando un'alterazione della proteina, mutazioni alla giunzione fra introne ed esone possono interferire con il riconoscimento dei siti di *splicing* causando la produzione di un trascritto maturo alterato, mentre mutazioni nelle regioni non tradotte (UTRs) e negli introni sono spesso neutre, o in alcuni casi alterano le funzioni regolatrici di queste regioni. Le mutazioni che causano più frequentemente effetti fenotipici, e per questo le più note, sono quelle che colpiscono la regione codificante; ci concentreremo perciò su di esse. Nel caso di una mutazione silente (Fig. 6 b), la sostituzione di una base provoca un cambiamento nel codone senza un cambiamento dell'aminoacido codificato, perché il codice genetico è degenerato (Fig. 5) e uno stesso aminoacido può essere codificato da più di un codone. Per esempio, se una mutazione modifica il codone CUA in CUG, poiché entrambi i codoni codificano per lo stesso aminoacido (leucina), l'aminoacido che viene inserito nella proteina non cambia. Per questo motivo una mutazione **silente** non ha effetto a livello della proteina e quindi del fenotipo. Nel caso di una mutazione *missense*, la sostituzione di una base di un codone con una base diversa provoca il cambiamento del significato del codone (Fig. 6 c). Per esempio, se il codone AGC, che codifica per l'aminoacido serina, diventa AGA, che codifica per l'aminoacido arginina, durante la sintesi proteica verrà inserita una arginina invece di una serina. La sostituzione di un aminoacido con un altro nella proteina può quindi comprometterne la funzione. Sulla base del tipo di aminoacido sostituito, le mutazioni missense si dividono in:

- **sostituzioni conservative:** il nuovo aminoacido ha caratteristiche simili a quello sostituito (mutazione neutra);
- **sostituzioni non conservative:** il nuovo aminoacido ha caratteristiche diverse da quello sostituito [ad es. GAG (glu) GTG (val), come nell'anemia falciforme].

Seconda lettera

		U	C	A	G				
U	UUU	Phe (F)	UCU	Ser (S)	UAU	Tyr (Y)	UGU	Cys (C)	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu (L)	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp (W)	G
C	CUU		CCU		CAU	His (H)	CGU		U
	CUC	Leu (L)	CCC	Pro (P)	CAC		CGC	Arg (R)	C
	CUA		CCA		CAA	Gln (Q)	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU		ACU		AAU	Asn (N)	AGU	Ser (S)	U
	AUC	Ile (I)	ACC	Thr (T)	AAC		AGC		C
	AUA		ACA		AAA	Lys (K)	AGA	Arg (R)	A
	AUG	Met (M)	ACG		AAG		AGG		G
G	GUU		GCU		GAU	Asp (D)	GGU		U
	GUC	Val (V)	GCC	Ala (A)	GAC		GGC	Gly (G)	C
	GUA		GCA		GAA	Glu (E)	GGA		A
	GUG		GCG		GAG		GGG		G

Figura 5: Il codice genetico. La tabella illustra le corrispondenze tra i diversi codoni e gli amminoacidi codificati.

Le conseguenze di una mutazione *missense* sono più o meno gravi a seconda delle caratteristiche chimico-fisiche dei due amminoacidi coinvolti. Le caratteristiche degli amminoacidi presenti nella struttura primaria di una proteina ne determinano le strutture secondarie e terziaria per mezzo di interazioni ioniche ed idrofobiche. A sua volta, il corretto ripiegamento della catena proteica ne determina la funzionalità a livello biologico. A seconda della posizione dell'amminoacido sostituito e del suo ruolo strutturale nella proteina, può essere fondamentale, per il mantenimento delle



funzioni della proteina, che esso abbia un certo ingombro sterico o un certo grado di polarità. Nel caso dell'anemia falciforme, a causa di una mutazione *missense* nel gene della beta-globina, un aminoacido polare e quindi solubile in acqua (acido glutammico) è sostituito da un aminoacido apolare insolubile (valina); questo causa la produzione di una catena beta globinica anormale con conseguente solubilità alterata dell'emoglobina, che a sua volta provoca la caratteristica forma dei globuli rossi allungati a falce.

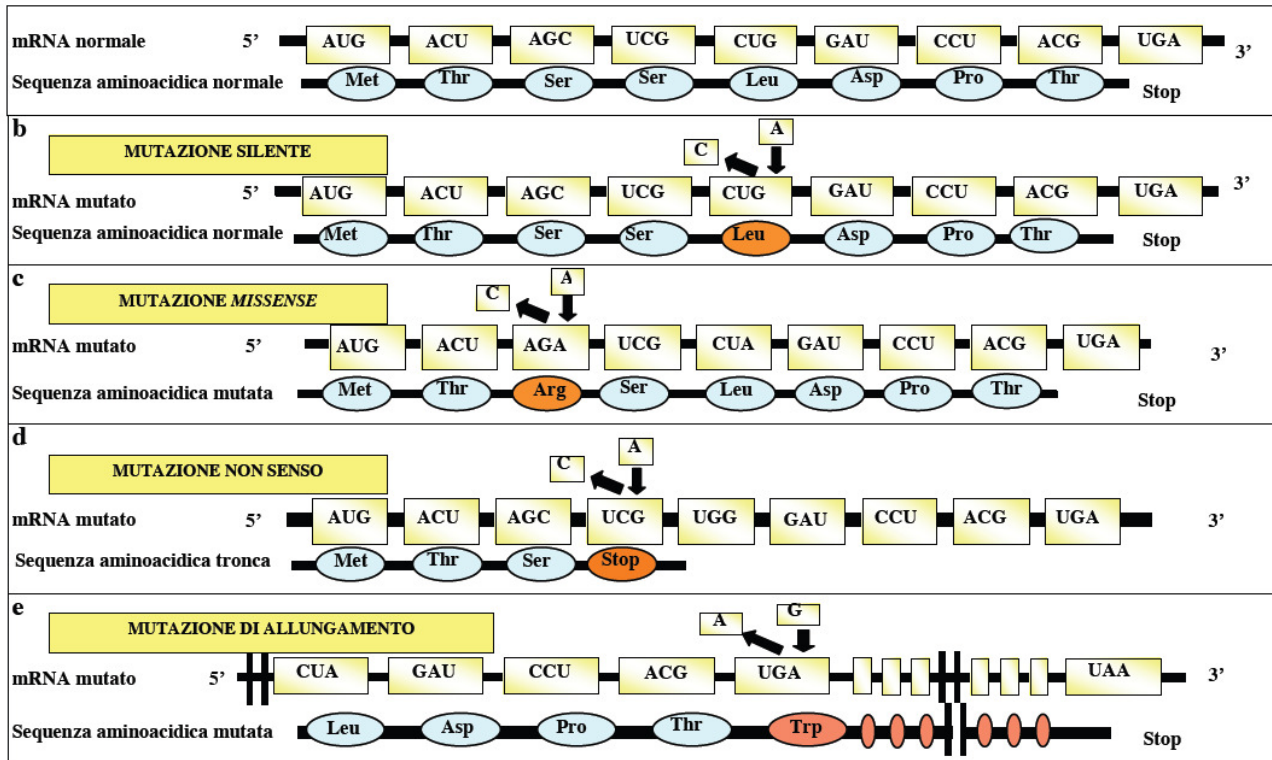


Figura 6: Conseguenze di una mutazione genica da sostituzione di basi in un esone.

(a) Sequenza nucleotidica e aminoacidica normali;

(b) se la sostituzione di basi, pur modificando la sequenza nucleotidica, non modifica il significato del codone, essa determina una mutazione silente;

(c) si ha una mutazione missense, se la sostituzione di basi nel DNA provoca un cambiamento nel significato del codone, per cui viene inserito un aminoacido diverso al posto di quello selvatico (normale). Le mutazioni missense producono proteine della stessa lunghezza della proteina normale;

(d) se la sostituzione di basi determina un cambiamento da un codone "senso" che codifica per un aminoacido ad un codone di terminazione (detto anche codone non senso o di stop), insorge una mutazione non senso. La traduzione della proteina mutata viene pertanto bloccata prematuramente, in corrispondenza della tripletta non senso formatasi per





UNIONE EUROPEA
Fondo Europeo di Sviluppo Regionale



mutazione. Si forma pertanto una proteina più corta di quella selvatica, che in genere è incapace di svolgere la propria funzione;

(e) si origina una mutazione di allungamento, quando la tripletta non senso del gene selvatico viene mutata in una tripletta “senso”, che codifica per un aminoacido. Ciò determina quindi la sintesi di una proteina mutata più lunga di quella selvatica. La sintesi della proteina mutata termina, quando si incontra un codone di stop, presente naturalmente nel DNA del gene, a valle della mutazione di allungamento. Per semplicità, è riportata la sequenza dell’RNA messaggero, anche se la mutazione è avvenuta nel DNA.

Una mutazione genica da sostituzione di basi può trasformare un codone senso in un codone non senso (mutazione non senso), cioè in un segnale di termine della sintesi proteica (Fig. 6 d). Per esempio, si verifica una mutazione non senso se il codone AAG, che codifica per lisina, diventa UAG, che è un codone di stop della sintesi proteica. La sintesi proteica, quindi, terminerà precocemente in corrispondenza del codone non senso che si è formato per mutazione. La proteina sintetizzata è quindi incompleta e, nella maggior parte dei casi, non sarà funzionale. Un esempio di malattia monogenica associata a questo tipo di mutazione è una forma di beta talassemia diffusa in Sardegna, causata da una mutazione che genera un codone di stop nel primo esone del gene per la beta globina, con conseguente assenza della proteina. Una mutazione di allungamento (Fig. 6 e), si origina quando una tripletta nonsense presente nel gene selvatico, a seguito di una sostituzione di base, diventa una tripletta “senso”, vale a dire una tripletta che codifica per un aminoacido. In questo caso, essendo stata eliminata dalla mutazione la tripletta di stop, la traduzione dell’ mRNA proseguirà con formazione di una proteina più lunga rispetto a quella selvatica, generalmente instabile. Delezioni e inserzioni di basi nella regione codificante di un gene provocano invece scivolamento del sistema di lettura del codice e sono per questo chiamate mutazioni *frameshift*. Allo stesso modo, l’inserzione o la delezione di basi in un gene, purché esse non siano tre o in numero multiplo di tre, determina una mutazione *frameshift*, ovvero uno scivolamento della fase di lettura del gene (Fig. 7). Tutti i codoni a valle della mutazione cambiano e, quindi, da quel punto in poi vengono incorporati nella proteina in formazione amminoacidi diversi, che danno luogo a una proteina del tutto anomala e quasi sempre non funzionale. Se le basi inserite o cancellate sono tre o un multiplo di tre, vi è aggiunta o perdita di uno o più aminoacidi nella proteina, con conseguenze anche molto gravi, come nel caso della fibrosi cistica (CF, cystic fibrosis), dove la mutazione più comune, ΔF 508, determina la perdita della fenilalanina in posizione 508 nella proteina CFTR (*cystic fibrosis transmembrane regulator*). Una delezione di un intero tratto di un gene causa la perdita di una parte della proteina codificata e può alterarne gravemente la struttura tridimensionale, le proprietà chimiche e quindi funzionalità. Anche in questo caso, la delezione può portare ad uno scivolamento della fase di lettura del codice (*frameshift*), determinando oltre alla perdita di una parte della proteina anche l’alterazione della sequenza amminoacidica della proteina residua a valle della delezione.



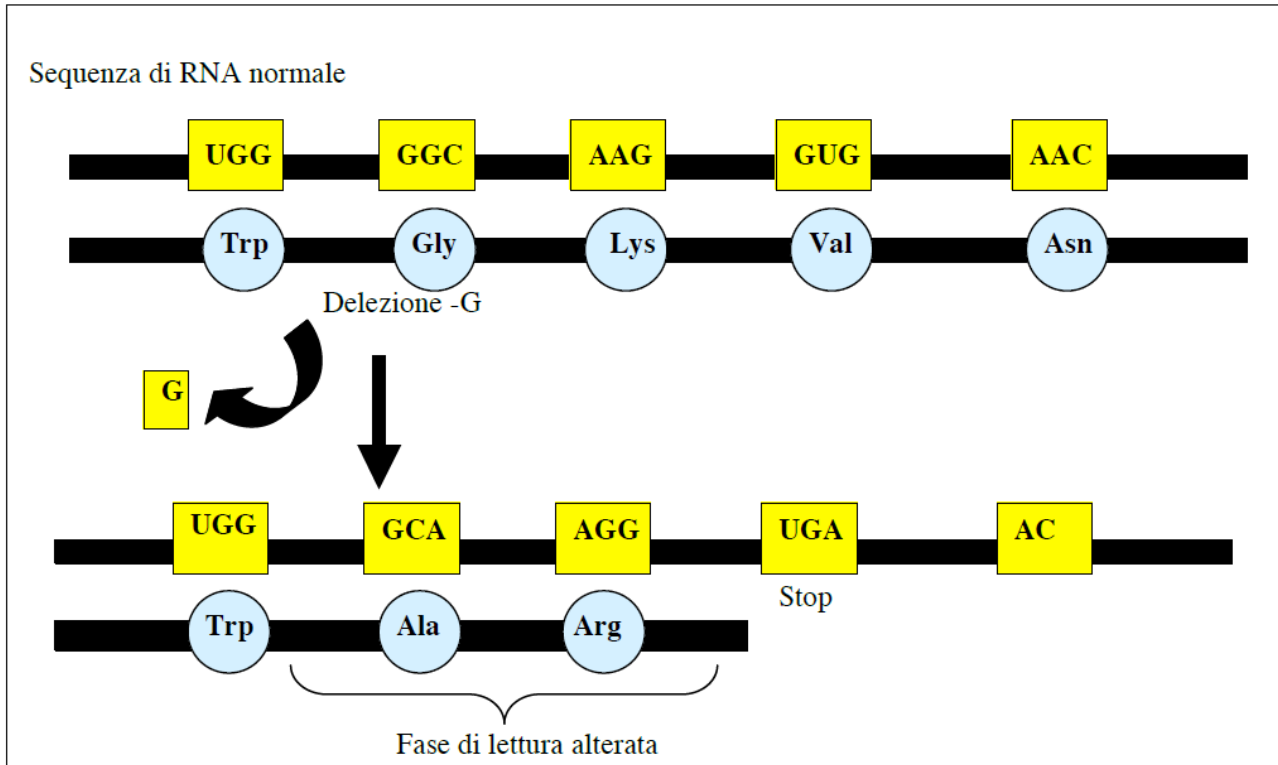


Figura 7: Mutazione frameshift: la delezione di una guanina sposta il sistema di lettura e di conseguenza all'aminoacido glicina si sostituisce l'alanina. Ulteriore conseguenza dello slittamento è la formazione della tripletta UGA, che corrisponde a un codone di stop. La proteina risultante sarà perciò accorciata.

Una mutazione in un gene può anche determinare una variazione quantitativa (non qualitativa!) della proteina; questo si verifica quando la mutazione cade in una regione regolatrice (ad es. la sequenza del promotore) e determina l'alterazione della trascrizione del gene. Una mutazione puntiforme (da sostituzione oppure da inserzione/delezione di basi), sia in una regione codificante che in una non codificante, può dare origine a un RFLP (polimorfismo di lunghezza dei frammenti di restrizione), oppure ad un SNP (polimorfismo di singoli nucleotidi)



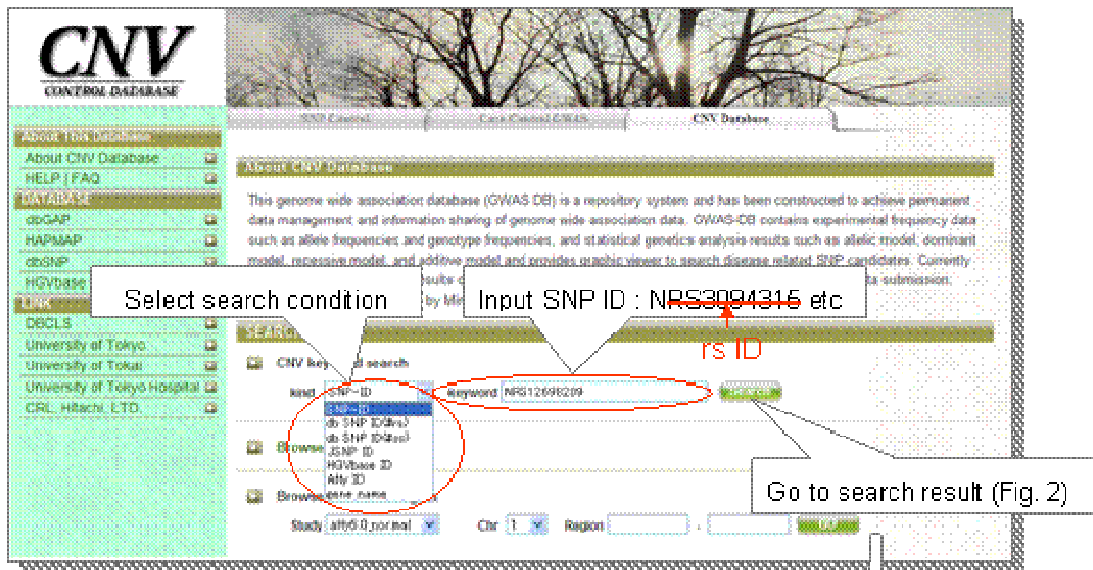
DB NAME	DESCRIPTION	BIBLIOGRAPHY
COSMIC	<p>All cancers arise as a result of the acquisition of a series of fixed DNA sequence abnormalities, mutations, many of which ultimately confer a growth advantage upon the cells in which they have occurred. There is a vast amount of information available in the published scientific literature about these changes. COSMIC is designed to store and display somatic mutation information and related details and contains information relating to human cancers. Some key features of COSMIC are: Contains information on publications, samples and mutations. Includes samples which have been found to be negative for mutations during screening therefore enabling frequency data to be calculated for mutations in different genes in different cancer types. Samples entered include benign neoplasms and other benign proliferations, in situ and invasive tumours, recurrences, metastases and cancer cell lines.</p> <p>The mutation data and associated information is extracted from the primary literature and entered into the COSMIC database. In order to provide a consistent view of the data a histology and tissue ontology has been created and all mutations are mapped to a single version of each gene. The data can be queried by tissue, histology or gene and displayed as a graph, as a table or exported in various formats.</p>	<p>1. S Bamford, E Dawson, S Forbes, J Clements, R Pettett, A Dogan, A Flanagan, J Teague, PA Futreal ,MR Stratton and R Wooster The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website - British Journal of Cancer (2004) 91, 355 – 358</p>
DMuDB	<p>The Diagnostic Mutation Database (DMuDB) was established in 2005 by NGRL as a repository of diagnostic variant data, to support the diagnostic process in UK genetic testing laboratories. NGRL developed DMuDB in response to a need amongst UK laboratories for an easy and secure way to share variant data in order to support the interpretation of new variants and improve the quality and consistency of diagnoses. Access to DMuDB has now been extended to non-UK laboratories through a partnership with EMQN.</p>	
HGVS database list	<p>The Society maintains comprehensive lists of databases.</p>	

Locus Specific database list	Based on various online resources and direct submissions of LSDBs	
MutDB	The goal of MutDB is to annotate human variation data with protein structural information and other functionally relevant information, if available. The mutations are organized by gene. Click on the alphabet below to go alphabetically through the list of genes. If you want to search for a specific target or disease, search below (the search may take up to a minute).	
Universal Mutation Database	The collection of these mutations will be critical for researchers and clinicians to establish genotype/phenotype correlations. Other fields such as molecular epidemiology will also be developed using these new data. Consequently, the future lies not in simple repositories of locus-specific mutations but in dynamic databases linked to various computerized tools for their analysis and that can be directly queried on-line. To meet this goal, we devised a generic software called UMD (Universal Mutation Database).	1. Beroud C, Collod-Beroud G, Boileau C, Soussi T, Junien C. UMD (Universal mutation database): a generic software to build and analyze locus-specific databases. Hum Mutat 2000; 15: 86-94.

Tabella 4: Elenco delle banche dati utilizzate per la ricerca di SNP che causano malattie.

5. UTILIZZO DI UNA BANCA DATI: ESEMPIO DI RICERCA SU CNV DATABASE

STEP 1: Nel primo step come mostrato nella prima parte della figura bisogna selezionare la condizione di ricerca e inserire come input l'ID dello SNP.



Link to region map (Fig. 6)

Fig. 1.

CNV Search Result : Search keyword (NRS12698209)

Study ID	Study Name	Variation ID	Chr	Position Start-End	gene
affy6.0_normal3	affy6.0_normal_control3	var_318	7	157798385 - 157916205	
affy6.0_normal5	affy6.0_normal_control5	var_1150	7	157805816 - 157916205	
affy6.0_normal	affy6.0_normal_control	var_5085	7	157810338 - 157810356	
affy6.0_AD	affy6.0_normal_AD	var_10152	7	157810336 - 157810356	

Fig. 2.

STEP 2: Una volta ottenuti i risultati, come nella seconda parte della figura, per raggiungere la mappa della regione in cui è presente lo SNP bisogna selezionare il relativo link.



Fig. 3

STEP 3: Una seconda modalità per la ricerca di SNP è quella di utilizzare la “geografia” del genoma



UNIONE EUROPEA
Fondo Europeo di Sviluppo Regionale

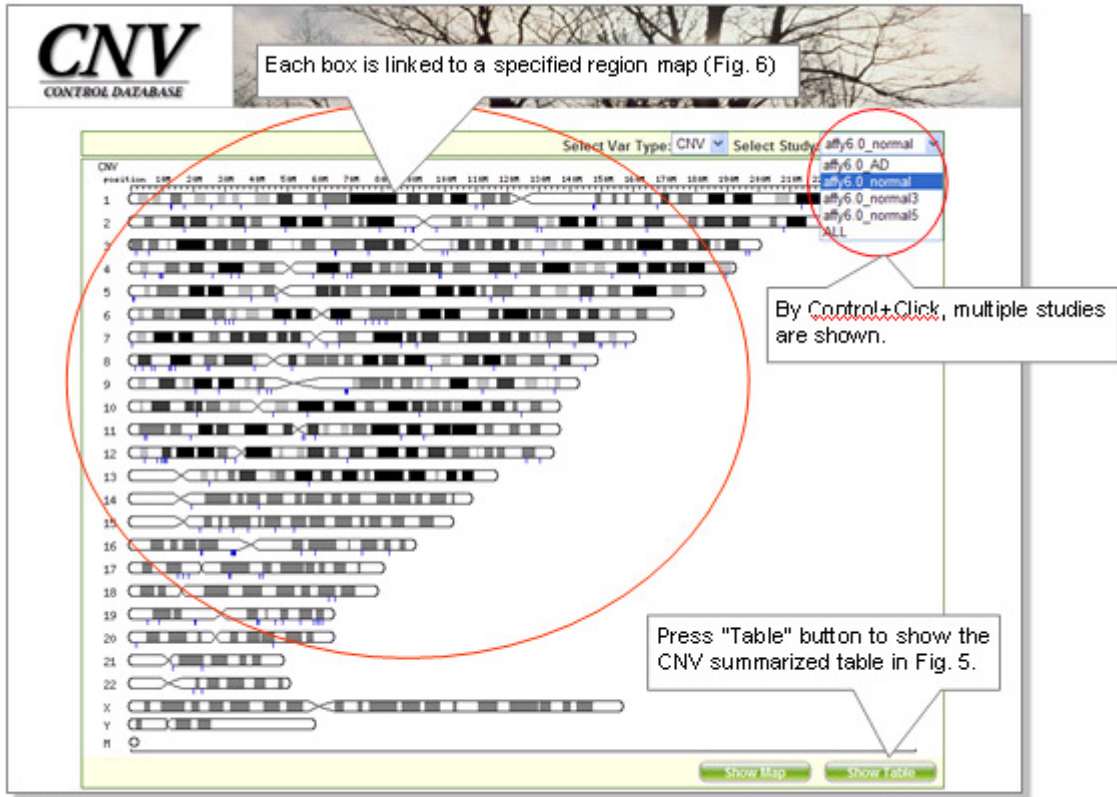


Fig. 4

STEP 4: Ogni box è collegato ad una specifica regione della mappa.



Ministero dell'Istruzione,
dell'Università e della Ricerca



Ministero
dello Sviluppo Economico





UNIONE EUROPEA
Fondo Europeo di Sviluppo Regionale



Variation ID	OGV3D	Chr	Position Start-End	Variation type	cytoband	gene	comment
var 5051		4	189574080 - 189574160	CNV			
var 5052		5	7947779 - 878120	CNV		ZDHHC11B	
var 5053		5	38183243 - 38184645	CNV			
var 5054		5	46313969 - 46313969	CNV			
var 5055		5	114787066 - 114787066	CNV			
var 5056		5	118729074 - 118729332	CNV		TNFAIP8	
var 5057		5	143386876 - 143390478	CNV			
var 5058		5	146376815 - 146377264	CNV		PPP2R2B	
var 5059		6	214735 - 323970	CNV		similar to dual specificity phosphatase 22 DUSP22	
var 5060		6	26853894 - 26856009	CNV			
var 5061		6	29948930 - 30034440	CNV		HLA-A	
var 5062		6	31403956 - 31403956	CNV			
var 5063		6	31445851 - 31448786	CNV			
var 5064		6	32556076 - 32665279	CNV			
var 5065		6	32701485 - 32703061	CNV			
var 5066		6	49041029 - 49042934	CNV			
var 5067		6	65770513 - 65770513	CNV		EGFL10	
var 5069		6	67094767 - 67094767	CNV			
var 5070		6	74648953 - 74648953	CNV			
var 5071		6	77078998 - 77081385	CNV			
var 5072		6	79036117 - 79083405	CNV			
var 5073		6	81345333 - 81345333	CNV			
var 5074		7	250149 - 250149	CNV			
var 5075		7	56732287 - 56732287	CNV			
var 5076		7	62326882 - 62343621	CNV			
var 5077		7	66280703 - 66280703	CNV		TNNT1	
var 5078		7	86079324 - 86082341	CNV			
var 5079		7	90876042 - 90877744	CNV			
var 5080		7	133436439 - 133437483	CNV			
var 5081		7	141416529 - 141429438	CNV		MGAM	
var 5082		7	149191005 - 149210881	CNV		hypothetical protein LOC643641 hypothetical gene LOC401431 ATP5V0E2	
var 5083		7	149918110 - 149931660	CNV			
var 5084		7	154031763 - 154031763	CNV		DPP6	
var 5085		7	157810308 - 157810356	CNV		PTPRN2	
var 5086		8	1346442 - 1347637	CNV			
var 5087		8	3274458 - 3277329	CNV		CSMD1	
var 5088		8	6843305 - 6844636	CNV			

Link to Fig. 6

Fig. 5

STEP 5: Dettaglio della delle variazioni sulla regione selezionata.



Ministero dell'Istruzione,
dell'Università e della Ricerca



Ministero
dello Sviluppo Economico





UNIONE EUROPEA
Fondo Europeo di Sviluppo Regionale

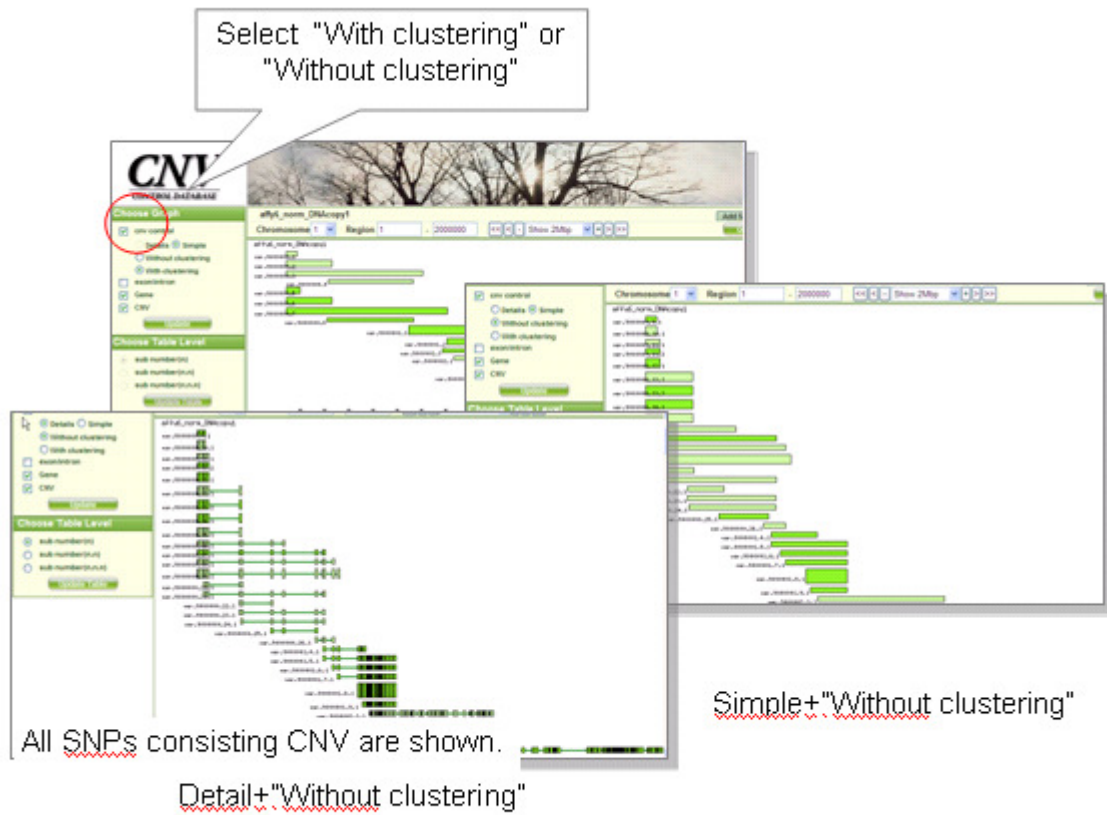


Fig. 6 (transition from fig. 2, 3, and 5)

STEP 6: Dettagli delle sequenze clusterizzate o non clusterizzate.



Ministero dell'Istruzione,
dell'Università e della Ricerca



Ministero
dello Sviluppo Economico





UNIONE EUROPEA
Fondo Europeo di Sviluppo Regionale

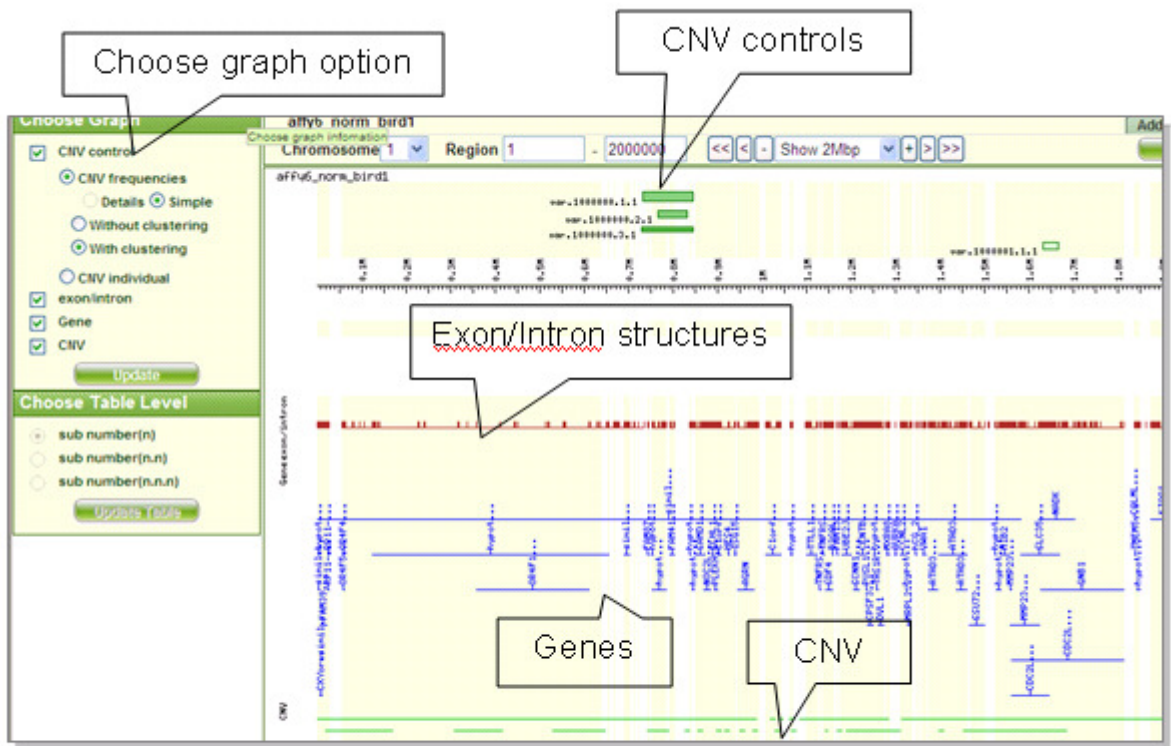


Fig. 7 (transition from Fig.2, 3, 5)

STEP 7: Grafico che mostra la struttura delle sequenze che compongono la zona selezionata.



Ministero dell'Istruzione,
dell'Università e della Ricerca



Ministero
dello Sviluppo Economico



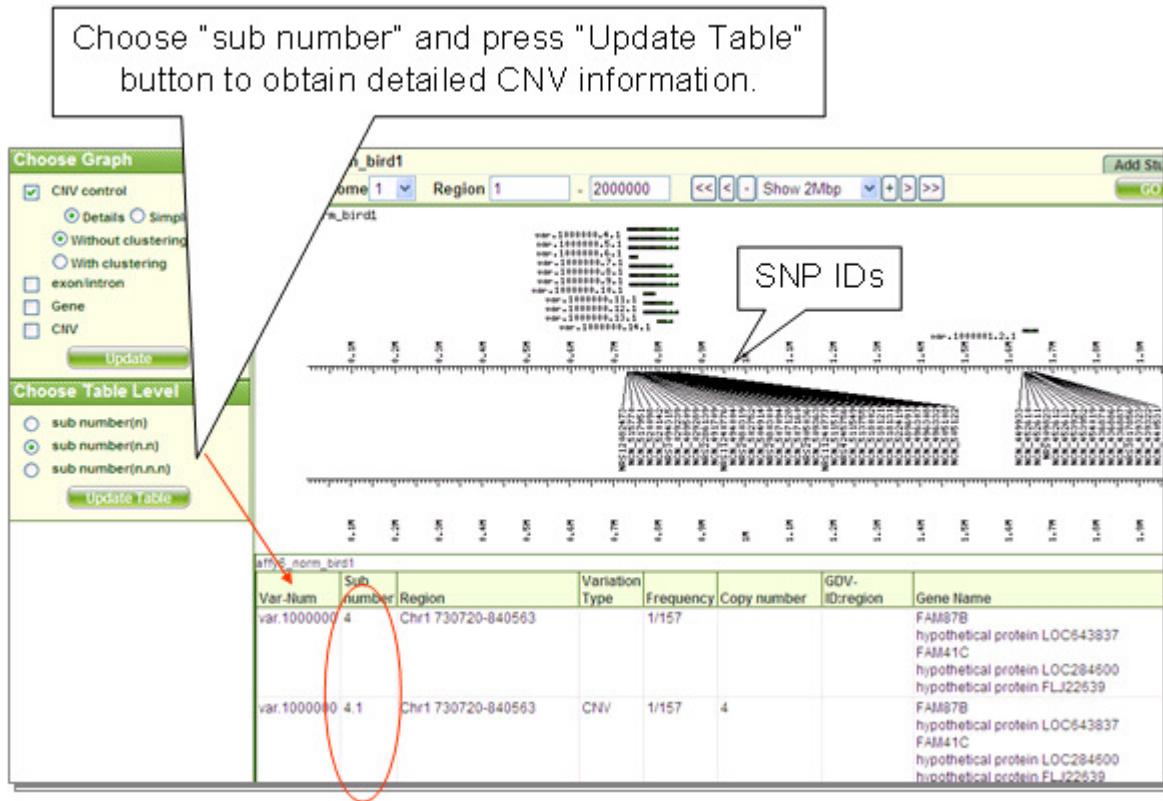


Fig. 8

STEP 8: Si sceglie il “sub number” e si seleziona “Update Table” per ottenere dettagli per l’informazione CNV.



6. DISCUSSIONI

L'obiettivo principale nella ricerca genetica e l'individuazione di SNP è quello di trovare farmaci quanto più efficaci sul singolo individuo che presenta una patologia genetica derivante da variazione. Grazie alla recente esplosione di conoscenze nell'ambito della genetica, dovute principalmente agli sviluppi della biologia molecolare e al sequenziamento del genoma umano, la consapevolezza del ruolo dei fattori genetici come causa di malattie ereditarie è aumentata notevolmente. Per numerose malattie ereditarie monogeniche (causate da alterazioni in un singolo gene) sono stati identificati i corrispondenti "geni-malattia" (Fig. 8). Inoltre, è stata identificata un'implicazione genetica nell'eziologia di numerose malattie complesse, quali le coronaropatie, il diabete mellito, l'ipertensione e le principali psicosi (Fig. 9).

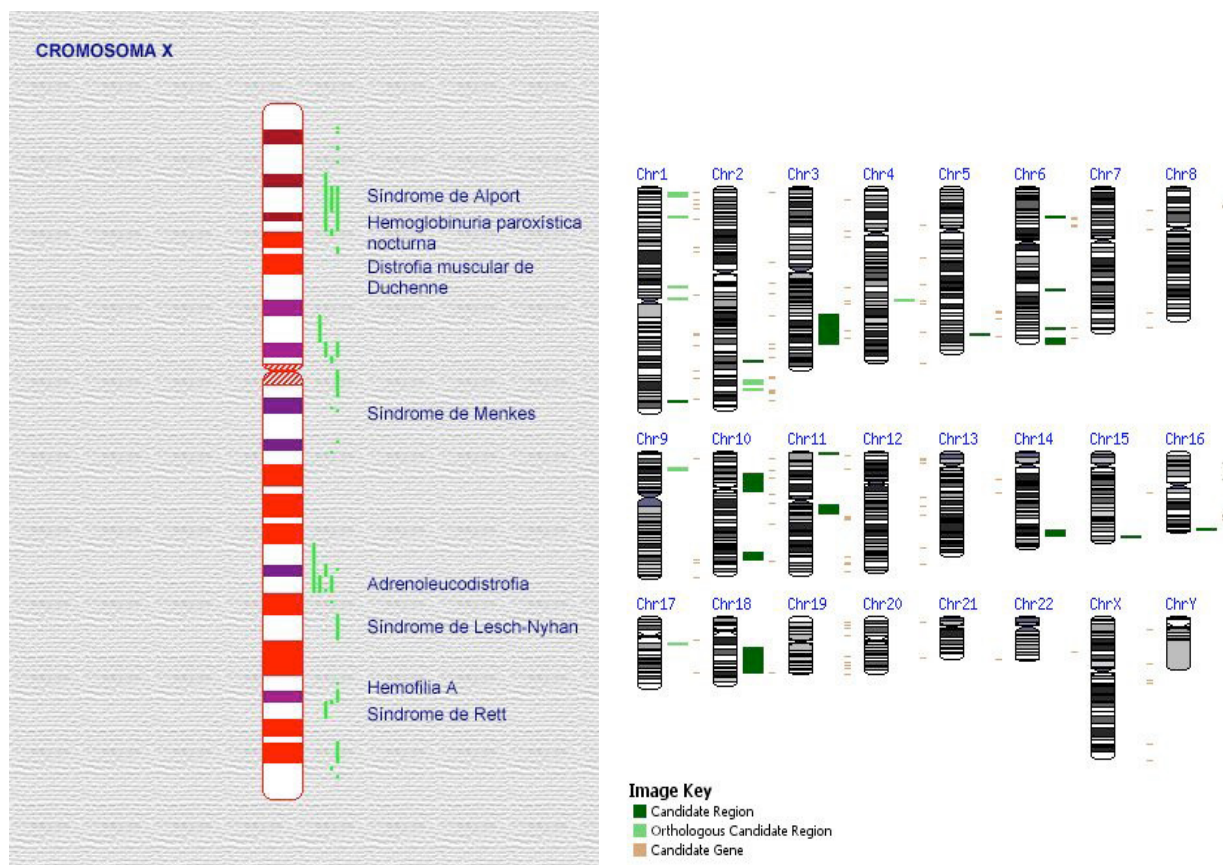


Figura 8: Alcune malattie ereditarie dovute a geni localizzati sul cromosoma X.

Figura 9: Geni e diabete mellito: sono indicati i geni e le regioni cromosomiche "candidate" come causa della patologia.



UNIONE EUROPEA
Fondo Europeo di Sviluppo Regionale



La variabilità nella risposta al trattamento farmacologico tra individui costituisce uno dei maggiori problemi clinici. Le risposte ai farmaci variano con effetti terapeutici ridotti o addirittura assenti, reazioni avverse o effetti collaterali, nonostante la somministrazione dello stesso farmaco con la stessa posologia. Questa variabilità inter-individuale veniva attribuita principalmente all'influenza di fattori non genetici come l'età, il sesso, lo stato nutrizionale, la funzionalità renale ed epatica, la dieta, l'abuso di alcool e fumo, la concomitante assunzione di altri farmaci o la presenza di comorbidità. Attualmente si ritiene che, oltre a questi fattori, sia di estrema importanza anche la componente ereditaria. I fattori genetici esercitano un ruolo importante nel determinare farmacocinetica e farmacodinamica, causando una variabilità inter-individuale nella risposta al trattamento farmacologico dovuta a mancata o solo parziale efficacia della terapia, da effetti collaterali di un determinato principio attivo o da reazioni avverse anche gravi e talvolta fatali.

La farmacogenetica nasce intorno agli anni cinquanta quando i ricercatori associarono la regolazione della risposta ai farmaci ai geni e la variabilità di reazione individuale a un certo principio attivo alle differenze genetiche. Il termine farmaco genetica venne introdotto da Friedrich Vogel nel 1959, per indicare lo studio delle singole differenze ereditarie nel metabolismo e nelle risposte ai farmaci. La farmaco genetica studia le variazioni inter-individuali nella sequenza del DNA in relazione alla risposta ai farmaci. L'applicazione pratica delle conoscenze, provenienti dalla ricerca in farmacogenetica, consiste nella possibilità di predire la risposta di un paziente ad un certo farmaco sulla base di un test genetico, per arrivare ad un'individualizzazione della terapia, "il farmaco giusto al paziente giusto nel momento giusto alla dose giusta". Dal Progetto Genoma sappiamo, come già ampiamente descritto nelle precedenti sezioni, che il DNA è identico in tutti gli individui per il 99.9 %, ma il restante 0.1% è a Singolo Nucleotide (SNPs). Sono noti più di 3 milioni di SNPs ed esistono mappe di SNPs molto dettagliate, cioè mappe relative alla posizione degli SNPs nel genoma umano. Ma ciò che è fondamentale capire da parte della ricerca è il significato della variabilità naturale nelle sequenze di DNA al fine di definire una terapia che tenga conto della unicità del genoma. Una vera e propria rivoluzione concettuale, che sostituisce la farmacologia basata sulla malattia con una terapia dell'individuo guidata dal corredo genetico, con un intervento sempre più individuale, studiato non per curare la malattia ma per guarire quel singolo paziente. Personalizzare la cura significherà, per esempio, prevedere e quindi evitare gli effetti collaterali, garantendo una migliore e più mirata efficacia del farmaco. L'altro ramo importante della ricerca farmaceutica, è la farmacogenomica, la ricerca di nuovi bersagli per i farmaci di domani. La farmacogenomica è una scienza emergente sviluppatasi dalla farmacogenetica classica e dal Progetto Genoma Umano. Può essere definita come la scienza che si interessa di come le nuove conoscenze sul genoma umano e sui suoi prodotti (RNA e Proteine) possano essere utilizzate nella scoperta e sviluppo di nuovi farmaci. Le basi si alternano nel DNA ma non è ancora chiara la funzione della maggioranza dei geni, il modo con cui lavorano per permettere le nostre funzioni



Ministero dell'Istruzione,
dell'Università e della Ricerca



Ministero
dello Sviluppo Economico





UNIONE EUROPEA
Fondo Europeo di Sviluppo Regionale



vitali. Comprendere cosa fanno i geni e come lo fanno, la loro funzione, è ciò che permetterà alla farmacogenomica di scoprire nuovi farmaci, ma soprattutto farmaci innovativi, in grado di curare malattie per le quali oggi non vi è cura disponibile. Il destino dei farmaci nell'organismo (farmacocinetica) ed i loro effetti terapeutici e tossici (farmacodinamica) sono regolati da processi complessi ai quali partecipano, cooperando, numerose proteine deputate al trasporto e al metabolismo dei farmaci, o coinvolte nel loro meccanismo di azione, a loro volta codificate da geni diversi. Nell'uomo si ritiene che la maggioranza dei geni contenga variazioni casuali della sequenza nucleotidica tra i diversi individui, sviluppatasi nel corso dell'evoluzione; quando tali variazioni avvengono nella sequenza codificante o regolatoria possono portare all'inserzione di un amminoacido diverso a livello di una specifica posizione nella proteina e conseguentemente a modificazioni della sua funzione, dei meccanismi di trascrizione e traduzione, modulando quindi i livelli di espressione dei prodotti genici (mRNA e proteine). I polimorfismi genici danno luogo a proteine con attività diversa con un possibile effetto sulla risposta farmacologica di un individuo se si tratta di geni codificanti per enzimi con diversi livelli di attività metabolica o di recettori con diversa affinità per il farmaco.

I test del DNA, basati su queste variazioni genetiche, possono predire come un paziente risponderà ad un trattamento farmacologico. I clinici potranno utilizzare questa informazione per decidere la terapia ottimale e per personalizzare il dosaggio; i benefici consistono in una ridotta incidenza di reazioni avverse, in migliori esiti clinici ed in costi ridotti. Questi test rappresentano il primo passo verso terapie paziente-specifiche. Con i test di farmacogenetica è possibile identificare molte variazioni nella struttura dei geni che codificano per enzimi del metabolismo dei farmaci, per proteine trasportatrici o proteine bersaglio (recettori, canali ionici, enzimi) di farmaci e correlarle alle variazioni inter-individuali nella risposta ai farmaci, individuando vari fattori genetici predittivi della tossicità e della risposta terapeutica al trattamento farmacologico. Un paziente con un metabolismo rapido, per esempio, può richiedere dosi più elevate e più frequenti per raggiungere le concentrazioni terapeutiche; invece un paziente con un metabolismo lento può avere bisogno di dosi più basse e meno frequenti per evitare la tossicità, specialmente nel caso di farmaci con un ristretto margine di sicurezza. Inoltre, identificando gli individui che con alta probabilità di manifestare una reazione avversa ad uno specifico trattamento farmacologico, il medico sarebbe aiutato nella scelta del farmaco e della dose ottimale per il singolo paziente, evitando un lungo processo di aggiustamento della terapia ed il rischio di tossicità. Analogamente, questi test sono potenzialmente utili nella selezione dei pazienti che con maggiore probabilità beneficerebbero dal punto di vista terapeutico di uno specifico trattamento farmacologico. I test di farmacogenetica metterebbero quindi il medico nelle condizioni di sapere a priori se un medicinale sarà tollerato bene dal suo paziente e quale dei diversi principi attivi a disposizione per curare una certa patologia avrà l'effetto migliore su una determinata persona. Attualmente la scelta del farmaco giusto avviene attraverso



Ministero dell'Istruzione,
dell'Università e della Ricerca



Ministero
dello Sviluppo Economico





UNIONE EUROPEA
Fondo Europeo di Sviluppo Regionale



una procedura per tentativi ed errori, cambiando la prescrizione fino a trovare il trattamento adatto per quella persona. Una simile procedura, tuttavia, espone il paziente ad eventi tossici. Il ricorso all'esecuzione di opportuni test genetici permetterebbe di evitare farmaci potenzialmente tossici e prescrivere terapie efficaci più tempestivamente, in altri termini affrontare in modo più efficace ed economico le malattie. In prospettiva, la farmacogenetica punta a una personalizzazione dei trattamenti, cioè a farmaci o combinazioni di farmaci che siano efficaci per ciascun paziente, a secondo del suo specifico patrimonio genetico.



*Ministero dell'Istruzione,
dell'Università e della Ricerca*



*Ministero
dello Sviluppo Economico*

