DEFINIZIONE DI UNA OPEN ACCESS POLICY PER GLI ISTITUTI SCIENTIFICI

Raccolta e analisi di letteratura scientifica su studi genomici di popolazioni umane con riguardo alla condivisione dei dati sperimentali

Nicola Milia - Istituto Italiano di Antropologia Roma, 29/12/2015

Introduzione

La condivisione dei dati sperimentali è una delle principali priorità della comunità scientifica in quanto permette di raggiungere la trasparenza e la riproducibilità delle ricerche scientifiche, di sfruttare a pieno le informazioni disponibili e di raggiungere una maggiore efficienza nell'uso delle risorse, sia umane sia economiche (Arzberger et al 2004).

Uno degli ambiti scientifici in cui la condivisione dei dati riveste un particolare interesse è quello della genomica umana. Indubbiamente, una rapida e piena disponibilità dei dati porterebbe numerosi vantaggi per il progresso di questo campo di ricerca. Tuttavia, esiste una diffusa riluttanza da parte dei ricercatori alla condivisione dei lori dati prima della loro piena valorizzazione e della pubblicazione dei relativi risultati (Contreras 2011; Williams 2013).

Inoltre, la divulgazione di dati genomici umani è ostacolata anche da problemi di natura etica. Infatti, spesso la *privacy* dei donatori non può essere garantita dal momento che le procedure di anonimizzazione non sono sufficienti a risolvere il problema. Inoltre, il riutilizzo dei dati da parte di terzi, senza il consenso del donatore non può essere controllata.

Ad oggi, solo pochi studi sono stati condotti per misurare il reale tasso di condivisione nell'ambito della genomica umana. Questi pochi studi si sono concentrati su insiemi di dati di espressione genica, attraverso l'analisi diretta di articoli di ricerca (Ochsner et al 2008, Piwowar & Chapman 2010), o attraverso l'utilizzo di procedure di *text mining* automatizzate (Piwowar 2011). Al tempo stesso, nessun dei succitati studi ha tentato di valutare l'"intelligen openess".

Secondo Boulton et al (2012), i dati sono "intelligentemente condivisi" se soddisfano i seguenti criteri: (i) reperibilità, i dati devono essere facilmente rintracciabili; (ii) accessibilità, i dati devono essere condivisi in modo che possano essere ottenuti facilmente; (iii) valutabilità, i dati devono essere accompagnati con una corretta informazione al fine di consentire una valutazione della loro qualità e affidabilità; (iv) utilizzabilità, i dati devono essere disponibili in formati che ne permettano un facile utilizzo ed essere corredati da un corretta informazioni accessorie (metadati).

Lo scopo di questo lavoro è stato valutare la condivisione di *dataset* relativi a dati genomici umani, in particolar modo di quelli prodotti utilizzando le due piattaforme di genotipizzazione *microarray* più comunemi (Illumina e Affymetrix). A tal fine, abbiamo selezionato dal database Pubmed, utilizzando parole chiave ad hoc, un totale di 100 articoli che contenessero set di dati genomici ed epigenetici originali. I *dataset* così selezionati sono stati analizzati per valutarne il grado di condivisone e l'"*Intelligent openness*".

I risultati ottenuti sono discussi alla luce delle strategie messe in campo dai diversi *stakeholders* della ricerca al fine di incrementare la disponibilità di dai genomici umani.

Materiali e Metodi

Abbiamo effettuato un'accurata analisi su un totale di 100 studi condotti sul genoma umano (SNPs), sull'epigenetica (CpG) e sull'espressione genica (trascrizione), contenenti dati ottenuti attraverso l'utilizzo di microarray di genotipizzazione. I lavori sono stati selezionati utilizzando due specifiche parole chiave: (i)

Illumina *human genome*; (ii) Affymetrix *human genome*. Solo articoli indicizzati fino a Dicembre 2013 sono stati presi in considerazione.

I lavori non pertinenti alle popolazioni umane, contenenti dati già pubblicati o prodotti attraverso tecnologie di *Next Generation Sequencing* (NGS), articoli di rassegna o meta-analisi sono stati esclusi dallo studio. Al fine di valutare l'"*intelligent openness*" degli studi abbiamo raccolto una serie di informazioni, come descritto nella Tabella 1.

Tabella 1 *Termini utilizzati per la valutazione del'"Intelligent openness" (Boulton, 2012) e le relative informazioni raccolte durante l'analisi dei lavori.*

Condizioni per un "Intelligent Openess"	**Definizione**	**Informazioni raccolte**
Reperibilità (Findability)	I dati sono molto semplici da rintracciare	Modalità di accesso alla pubblicazione (*open access* o a pagamento) Disponibilità dei dati in un database pubblico Presenza del numero identificativo del *dataset* Uso di standard di metadati per l'indicizzazione (titolo dell'articolo, autori, *abstract*, parole chiave)
Accessibilità (*Accessibility*)	I dati devono essere facilmente rintracciabili ed devo essere disponibili in una forma che ne permetta il loro riutilizzo	Condivisione in database primari Condivisione in altri database (es. Database privati o delle diverse istituzioni) Condivisione nei materiali supplementari Condivisione dei dati dopo una specifica richiesta
Valutabilità (*Assessability*)	Nel testo dell'articolo devono essere riportate tutta le informazioni che permettano il riutilizzo e la valutazione dell'affidabilità dei dati	**Informazioni sul campionamento** Descrizioni sul metodo di campionamento Descrizioni dei soggetti da cui deriva il campione biologico utilizzato Descrizioni dei criteri utilizzati per la scelta del campione Descrizioni dei metodi utilizzati per stabilire se i soggetti abbiano i requisiti richiesti **Informazioni sulla genotipizzazione** Tipo di materiale biologico utilizzato Descrizione delle metodiche di genotipizzazione utilizzate o riferimento a standard e/o procedure già descritte altrove. Descrizione delle procedure utilizzate per la trasformazione dei dati Descrizione delle procedure utilizzate per valutare la qualità dei dati o riferimento a standard e/o procedure già descritte altrove.
Usabilità (*Useability*)	I dati devono essere condivisi in un formato che ne permetta il loro riutilizzo, anche per scopi diversi. E' inoltre richiede una corretta informazione di fondo e metadati. L'usabilità dei dati dipenderà anche coloro che desiderano utilizzarli.	Utilizzo di formati standard per la condivisione dei dati Utilizzo di formati standard per i metadati Informazioni sulle modalità di accesso ai dati e sui termini per il loro riutilizzo*

* Valutato solo per i dati che presentano un accesso condizionato da alcune condizioni.

La scelta delle informazioni utilizzate per determinare l'effettiva valutabilità dei dati sono quelle proposte in *Minimum Information About a Microarray Experiment* (Miame) (Brazma et al., 2001) e in *Minimum Information about a Genotyping Experiment* (MIGen) (Huang et al., 2011).

Inoltre, abbiamo analizzato il nostro set di dati suddividendolo in base alla tipologia di studio: (i) studi di associazione (GWAS); (ii) studi di espressione genica; (iii) studi sulla metilazione del genoma; (iv) studi di natura evoluzionistica e (v) studi di medicina legale. Questi diversi tipi di studi sono stati classificati in base agli scopi e le finalità riportate nei singoli articoli.

Risultati e conclusioni

I risultati ottenuti da questo studio rappresentano il primo tentativo mai realizzato volto a valutare l'"*intelligent openness*" dei dati di ricerca nel campo della variabilità del genoma umano e rappresenta un'indagine preliminare condotta su un piccolo campione.

Il nostro campione finale è composto da 100 *dataset*, riportati in altrettanti studi. Nel complesso, abbiamo osservato un basso tasso di condivisione in quanto solo il 15% degli *dataset* erano pienamente accessibili senza nessuna restrizione (Figura 1). Il 3% dei *dataset* risultava accessibile solo a determinate condizioni chiaramente esplicitate nella relativa pagina del database dove tali dati erano sottomessi. Il 2% è risultato potenzialmente disponibile su richiesta all'autore dell'articolo mentre un'ulteriore 2% è risultato disponibile solo in parte (essendo il risultato di un *LD pruning*). Abbiamo anche riscontrato un caso in cui il *dataset* era ancora sotto embargo con data di rilascio posta al 30 giugno del 2015.

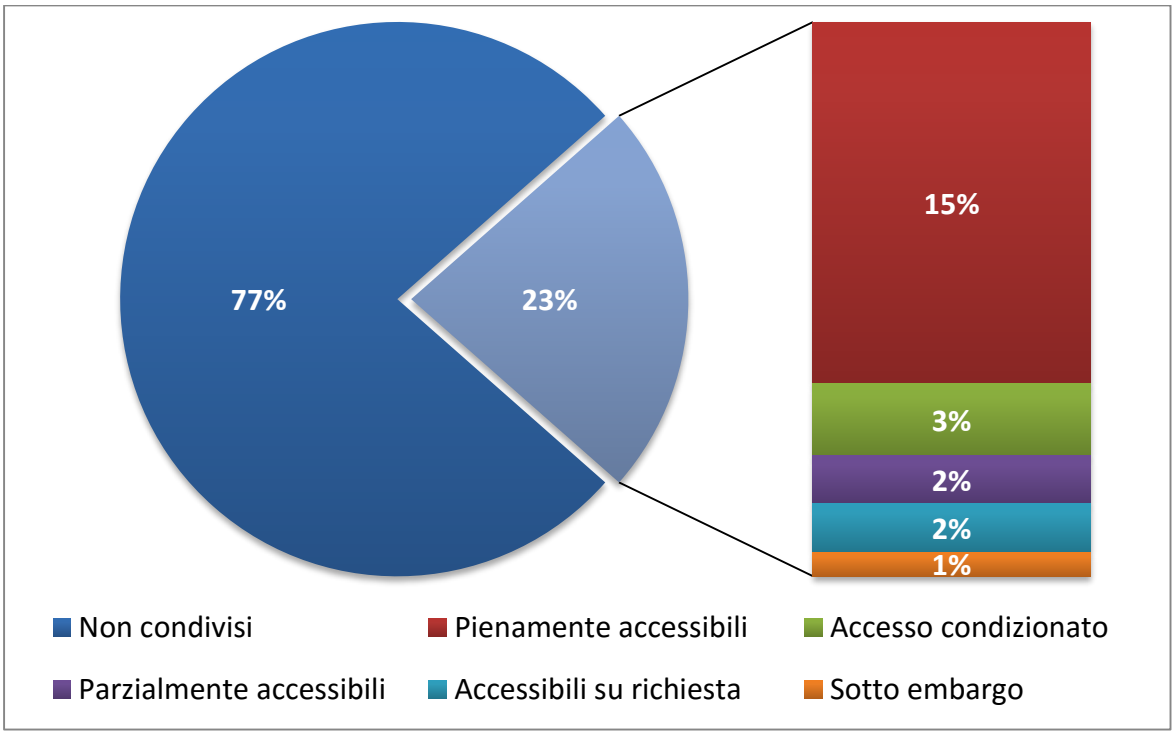

Figura 1 *Grado e modalità di condivisione dei dati nell'ambito degli studi genomici umani.*

Analizzando i nostri risultati sulla base della tipologia di studio, abbiamo osservato che i ricercatori che si occupano di analisi dell'espressione genica sono più inclini a condividere i propri dati (Figura 2). Infatti, il

64,3% (9 su 14) dei *dataset* è risultato liberamente disponibile, mentre 2 set di dati (14,2%) sono risultati o accessibili a determinate condizioni o ancora sotto embargo. Questo tasso di condivisione è notevolmente più alto rispetto a quanto riscontrato in precedenti studi, sia attraverso l'analisi diretta di 397 lavori (47% di tasso di condivisione) (Ochsner et al 2008, Piwowar & Chapman 2010) sia attraverso l'utilizzo di una procedura di *text-mining* automatizzata effettuata su più di 11.000 lavori (45% del tasso di condivisione) (Piwowar 2011). Naturalmente, il numero relativamente basso di lavori di espressione genica che abbiamo analizzato in questa sede, rispetto agli altri due studi, può aver contribuito a generare queste differenze.

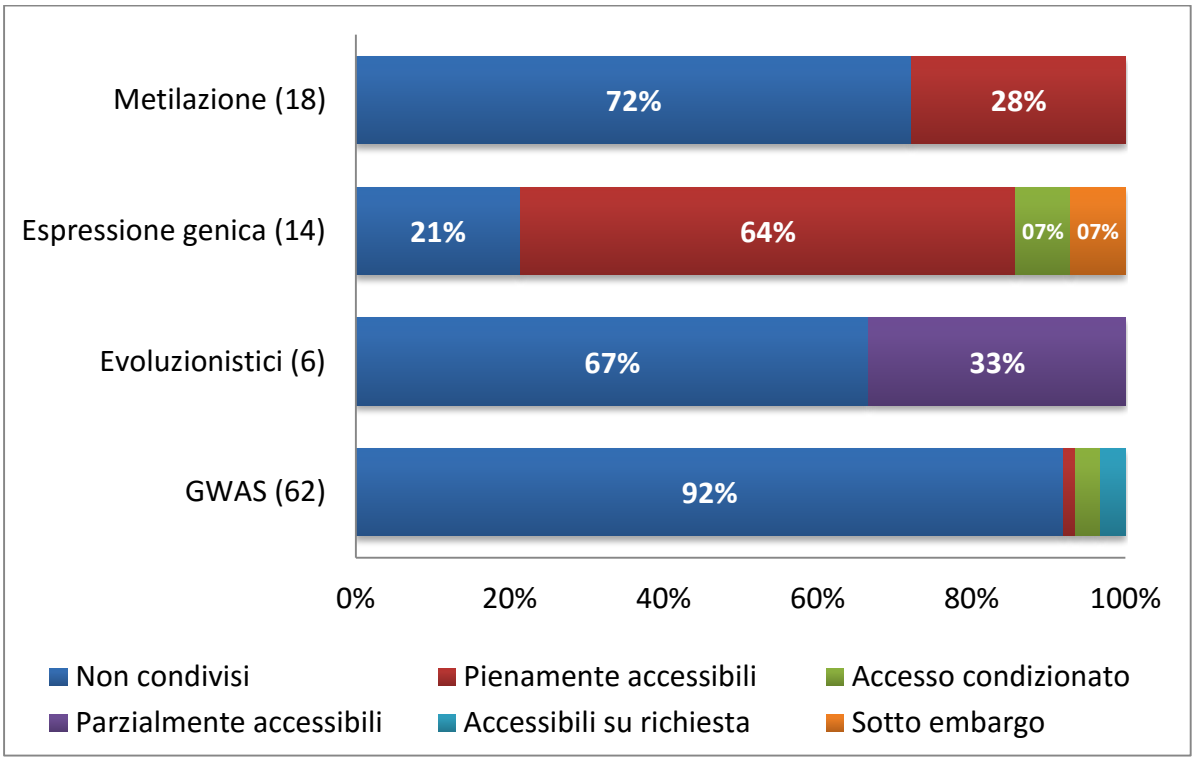

Figura 2 *Grado e modalità della condivisione dei dati nelle diverse tipologie di articoli analizzati in questo studio.*

Una tendenza opposta è stata osservata negli studi di associazione (GWAS), in cui solo un *dataset* è risultato pienamente condiviso (1,6%, su 61 *dataset*). I restanti quattro *dataset* considerati come condivisi rappresentano il 6.4% del totale e equamente suddivisi tra accessibili su richiesta e accessibili a determinate condizioni. Gli studi sulla metilazione mostrano una percentuale di condivisione del 27,8% (5 su 18), inferiore rispetto agli studi di espressione genica.

Abbiamo analizzato 5 lavori di natura evoluzionistica, in due di questi (40%) sono risultati condivisi solo una parte degli SNP totali prodotti. Infine, abbiamo trovato un solo *dataset* analizzato con finalità forensi, che è risultato essere non accessibile.

Purtroppo, il numero ristretto di *dataset* condivisi non ci ha permesso di investigare in modo approfondito tutte le questioni relative all'"*intelligent openness*" così riportiamo solo i risultati complessivi. Dei 23 set di dati, classificati come condivisi, poco più della metà (52,2%, 12 set di dati) rispettano pienamente i criteri di "*intelligent openness*". Tutti questi insiemi di dati sono facilmente trovabili e accessibili poiché sono pubblicati con una licenza di *open access*, depositati in un database pubblico (NCBI's Gene Expression Omnibus and EBI's Array Express Archive), e riportano un identificativo univoco e valido mentre al contempo sono corredati di metadato di indicizzazione standardizzati. Inoltre, sono "valutabili" in quanto forniscono

informazioni riguardanti le procedure di campionamento, genotipizzazione, trasformazione dei dati e di controllo della qualità e, infine, “utilizzabili” in quanto sono disponibili in formati standard e dispongono di metadati anch’essi in formato standard.

È interessante notare che, la metà dei lavori in cui è risultata esserci una "*intelligent openness*" trattano dati di espressione genica (58,3%. 7 su 12), un terzo analizza processi di metilazione (4 su 12) e uno solo è uno studio di associazione.

I risultati ottenuti suggeriscono che una gestione più controllata e standardizzata è assolutamente necessaria per incrementare il livello di condivisione dei dati. Questo, nonostante le alte percentuali di condivisione di *dataset* di espressione genica che, tuttavia, risultano più basse di quanto ci si potrebbe attendere visto che un numero sempre maggiore di riviste e enti finanziatori, ne richiedono sia la condivisione che l’ottemperanza con specifici standard di qualità quali i MIAME (Nature, editoriale)

Naturalmente, bisogna anche considerare che i dati del genoma umano sono soggetti a questioni di *privacy* e di riservatezza. Questi fattori possono aver giocano un ruolo importante nel determinare i bassi tassi di condivisione osservati, nonostante le attuali risorse digitali e i quadri normativi in materia permettono ai ricercatori di mantenere il controllo sull'accesso ai dati (tramite embargo o accesso condizionato) mentre ne assicurano la presentazione a lungo termine attraverso il deposito su archivi pubblici quali Gene Expression Omnibus e Array Express Archive.

Gli *stakeholders* della ricerca (ad es. I governi, fondatori, università, riviste) in tutto il mondo sono sempre più interessati alla conservazione dei dati, alla loro condivisione e al loro riutilizzo, soprattutto quando questi, come i dati del genoma umano, possano avere un elevato impatto potenziale sulla vita quotidiana delle persone. Ad esempio, molte Università, soprattutto nel Regno Unito e negli Stati Uniti, hanno messo a punto, per il loro personale di ricerca, specifiche politiche di gestione dei dati per regolamentarne il rilascio e la condivisione, nel rispetto dei principi dell’"*Intelligent openness*" (Mauthner e Parry 2013).

Inoltre, i dati aperti sono diventati una priorità riconosciuta da organizzazioni intergovernative come il G8, l'Organizzazione per la Cooperazione Economica e lo Sviluppo e il Consiglio europeo della ricerca.

In conclusione, abbiamo visto come un’"*Intelligent openness*" dei dati genomici umani sia effettivamente possibile. Le infrastrutture digitali, i quadri normativi, gli standard per dati e metadati sono ad oggi già disponibili. Servono tuttavia lavorare su due aspetti. Il primo è la diffusione di una cultura *“open data”* tra i ricercatori, una delle sfide più importanti che la comunità scientifica e i vari *stakeholder* della ricerca dovranno affrontare nel prossimo futuro. Sarà importante sviluppare anche standard di metadati che definiscano un quadro omogeneo di informazioni che possano permettere un riutilizzo più completo dei dati in ambito antropologico. A tale scopo, i risultati di questa analisi verranno utilizzati per ottimizzare il *form* per la raccolta dei metadati adottato nel database Anthro Digit^{data}.

Bibliografia

Arzberger P, et al. 2004. Data Science Journal 3:135-152.

Bijsterbosch M, 2013. EUDAT-DAPUR workshop, Barcelona, 25 – 26 September 2013

Boulton G, et al. 2012. Science as an open enterprise. The Royal Society, London.

Brazma A, et al. 2001. Nature Genetics 29: 365–371.

Contreras JL, 2011. Minnesota Journal of Law, Science & Technology 12:61.

Huang J, et al. 2011. Standards in Genomics Science 5:224-229.

Milia M, et al. 2012. Plos One 7:e37552.

Nature Editorial, 2002. Nature 419: 323.

Ochsner SA, et al. 2008. Nature Methods 5:991.

Piwowar HA, 2011. Plos One 6:e18657.

Piwowar HA, Chapman WW, 2010. Journal of Informetrics 4:148-156.

Williams HL, 2013. Journal of Political Economy 121:1.