Ministero dello Sviluppo Economico

Ricevuta di presentazione

per

Brevetto per invenzione industriale

Domanda numero: 102022000006119

Data di presentazione: 29/03/2022

DATI IDENTIFICATIVI DEL DEPOSITO

Ruolo	Mandatario
Depositante	paolo di giovine
Data di compilazione	29/03/2022
Riferimento depositante	BI5666R-PDG-fp
Titolo	Method for implementing the design of synthetic nucleic acid molecules for gene therapies in rare diseases.
Carattere domanda	Ordinaria
Esenzione	NO
Accessibilità al pubblico	NO
Numero rivendicazioni	18
Autorità depositaria	

PRIVACY

Autorizzo il trattamento dei dati personali, inseriti all'interno del deposito, ai sensi del GDPR (Regolamento UE 2016/679) e del Decreto Legislativo 30 giugno 2003, n. 196 "Codice in materia di protezione dei dati personali"

RICHIEDENTE/I

Natura giuridica	Persona giuridica
Denominazione	UNIVERSITÀ DEGLI STUDI DI FERRARA
P.IVA/CF	80007370382
Tipo Società	le universita'
Nazione sede legale	Italia
Comune sede legale	Ferrara (FE)
Indirizzo	Via Ariosto
Civico	35

CAP	44121
Telefono	
Fax	
Email	
Pec	
Quota percentuale	100.0%

DOMICILIO ELETTIVO

Cognome/R.sociale	Società Italiana Brevetti S.p.A.
Indirizzo	piazza di Pietra 39
Cap	00186
Nazione	Italia
Comune	Roma (RM)
Telefono	06 - 695441
Fax	06 - 69544810
Email\PEC	UIBM.prenov16@pec.sib.it

MANDATARI/RAPPRESENTANTI

Cognome	Nome
Di Giovine	Paolo
Barbaro	Gaetano
Concone	Emanuele Enrico Maria
De Benedetti	Jacopo
De Benedetti	Fabrizio
De Giorgi	Michele
Demegni	Giovanni Nicolò
Germinario	Claudio
Manfrin	Marta
Manna	Sara
Moscone	Francesca

Papa	Elisabetta
Pietri	Simona
Pozzato	Matteo
Predazzi	Valentina
Romano	Giuseppe
Rondano	Davide
Santoro	Sofia
Soldatini	Andrea

INVENTORI

Cognome	Nome	Nazione residenza
Ferlini	Alessandra	Italia
Rossi	Rachele	Italia

CLASSIFICAZIONI

Sezione	Classe	Sottoclasse	Gruppo	Sottogruppo
G	16	B		

NUMERO DOMANDE COLLEGATE

DOCUMENTAZIONE ALLEGATA

Tipo documento	Riserva	Documento
Descrizione in inglese	NO	BI5666R_descr.pdf.p7m hash: daa5512dc185b42486dc7b6dd4b83e99
Rivendicazioni in inglese	NO	BI5666R_claims.pdf.p7m hash: cde3f72dcc896fe2bbe82f95f4ddd107
Lettera di Incarico	NO	BI5666R_ LI.pdf.p7m hash: c5810d129a065d2cea8ed9ee82eb7727

Riassunto in Inglese	NO	BI5666R_abs.pdf.p7m hash: fa872bdd6541ac445efe510cb925b40e
Disegni	NO	BI5666R_drws.pdf.p7m hash: 24c08743c6e6f08a76e817253ac042e1
Riassunto	SI	hash:
Disegni	SI	hash:
Descrizione in italiano*	SI	hash:
Rivendicazioni	SI	hash:

PAGAMENTI

Tipo	Identificativo	Data
Bollo	01191998005052	10/03/2021

ESENZIONI INDICATE

Esenzione su diritti e tasse	DM 02/04/2007 - art. 2: esonero dal pagamento dei diritti di deposito e di trascrizione relativamente ai brevetti per invenzioni industriali, e modelli di utilita' a vantaggio di: Universita'; Amministrazioni Pubbliche aventi fra i loro scopi istituzionali finalita' di ricerca; Amministrazioni della Difesa; Amministrazioni delle Politiche Agricole, alimentari e forestali.

DOVUTO

Gli importi indicati non tengono conto delle eventuali esenzioni applicabili

Importo Tasse:	€ 410,00
Importo Imposta Bollo:	€ 20,00

NOTE

RIASSUNTO

La presente invenzione è legata a metodi che possono essere usati per identificare codoni critici e progettare molecule sintetiche di acido nucleico di geni causanti malattia. Le sequenze sintetiche di acido nucleico possono essere progettate da una sequenza di acido nucleico di riferimento, per esempio, per ottimizzare l'espressione eterologa della sequenza di acido nucleico in un particolare tessuto di un organism ospitante. Alternativamente, sequenze sintetiche di acido nucleico possono essere progettate de novo per codificare un polypeptide desiderato. dette sequenze sintetiche di acido nucleico possono esser eusate per esempio in terapie geniche o alter applicazioni terapeutiche.

Metodo per implementare la progettazione di molecole di acido nucleico sintetico per terapie geniche in malattie rare.

DESCRIZIONE

Ambito tecnico

La presente invenzione si riferisce a metodi usati per identificare codoni critici che possono essere necessari nella progettazione di effettive e appropriate molecole di acido nucleico sintetico con capacità di traslazione applicando il calcolo CUB in specie, tessuti e geni specifici della malattia. Le molecole ottimizzate di acido nucleico sintetico sono pensate per essere usate per la terapia genica o altre applicazioni terapeutiche, oppure per una produzione in larga scala di proteine, reagenti per la ricerca ricombinante e strumenti molecolari in un organismo ospitante. La presente invenzione include ma non è limitata a nuovi parametri, definiti dagli Autori, trattandosi di specie, tessuti e geni di interesse CUB trend, che possono essere usati per l'implementazione e l'ottimizzazione degli algoritmi già esistenti e dei software per la progettazione di molecole sintetiche.

Stato dell'arte

Il codice genetico consiste di 64 triplette di codoni che codificano 20 aminoacidi e tre codoni d'arresto, questi ultimi sono stati riconosciuti dal macchinario transazionale come interruttori della sintesi di proteine. Con l'eccezione di due aminoacidi, triptofano e metionina, che sono codificati da un unico codone, tutti gli altri aminoacidi riconoscono codoni sinonimi multipli basati su due, tre, quattro o sei ridondanze di tripletta, un fenomeno noto come degenerazione del codone. C'è l'intrigante evidenza che la ridondanza del codice genetico abbia giocato un ruolo evoluzionario cruciale nel permettere alla sintesi di proteine di trasformare il mondo RNA nel mondo delle proteine. Per ragioni non completamente comprese, alcuni codoni diventano poco utilizzati, un fenomeno noto come bias di utilizzo del codone (CUB), oppure tende a sparire (extreme CUB) durante l'evoluzione. Nonostante il CUB sia stato ampiamente studiato in varie categorie di geni (ontologia genetica o mappe interactome) e attraverso le specie, il suo

valore evolutivo è ancora incerto. L'originale teoria neutra sull'evoluzione molecolare potrebbe non applicarsi alla selezione di codoni e la pressione mutazionale e la selezione naturale potrebbero aver giocato un ruolo maggiore nel contribuire all'uso dei codoni. Infatti, nonostante solo in pochi casi come per la cheratina e qualche gene ribosomiale o mitocondriale, è stato identificato un extreme CUB negli umani e negli scimpanzé paragonati ad altri mammiferi, suggerire il suo valore evolutivo nei processi gioca un ruolo esclusivo in una linea eucariota specifica.

Esiste un largo consenso sul concetto che la scelta di un codone sinonimo influisce sull'efficienza traslazionale della proteina, sul livello di espressione sulla struttura, e sulla funzione, una nozione che ha suggerito la designazione di codoni ottimali e codone ottimizzazione, che è un processo di routine utilizzato in biologia sintetica per aumentare l'espressione delle proteine. Tuttavia, c'è poco consenso tra i vari algoritmi di ottimizzazione dei codoni, e le metriche correntemente usate potrebbero non essere appropriate per tutti i geni.

Il valore delle variazioni sinonime nel genoma umano e il loro effetto sulle malattie ereditarie e largamente sconosciuto. Interpretare il loro impatto funzionale sui geni è difficile, se non impossibile, senza saggi dedicati funzionali.

Gli strumenti in silicio sono al momento usati per decifrare variazioni sinonime, ma sono inaccurati. Per di più, cambiamenti sinonimi sono completamente ignorati e filtrati da output di dati genomici, un fatto che causa l'omissione degli stessi nella scoperta e validazione della variazione patogena e la mancanza di potenziali nuovi geni di malattia o l'identificazione di genotipi patogeni.

In termini di energia, un CUB estremo viene predetto con il mantenimento basso della richiesta di energia per la traslazione delle proteine, stando al principio massimo dell'entropia, che può portare al progressivo aumento di CUB durante l'evoluzione e attraverso le specie. Questo trend evidenzia certi percorsi funzionali che possono essere stati di priorità energetica (supponendo bias verso codoni preferiti) via selezione naturale e possono essersi verificati in famiglie di geni con particolare rilevanza in una data linea. Per esempio, in genomi con alto contenuto GC (come l'Homo sapiens, HSA), che possono innescare

cambiamenti SNP, l'extreme CUB si verifica frequentemente e si pensa che riduca il rischio del verificarsi di variazioni nonsenso.

Il ruolo dei codoni "rari" (extremely biased) attraverso l'evoluzione è ancora controverso. Nei batteri, i codoni rari vicino all'estremità 5' facilitano la rimozione della repressione della traslazione e sono considerati "rampe autostradali" per innescare e accelerare la traslazione di proteine di più di 60 volte, con un ruolo chiave nella regolazione del traffico ribosomiale. Sinergicamente, le proprietà di ripiegamento del mRNA ricco di codoni rari vicino all'estremità 5 incrementa la velocità di traslazione, come nelle cellule che si dividono rapidamente. Al contrario, alcuni codoni usati frequentemente hanno l'effetto opposto, rallentando l'efficienza della traslazione. Alla luce di quanto detto, c'è ancora necessità di provvedere metodo e modi di migliorare il bias di utilizzo dei codoni e le sue applicazioni.

Sommario dell'invenzione

La presente invenzione è basata sull'inattesa scoperta che l'approccio basato sulla malattia genica aiuta a identificare codoni critici, che possono giocare un ruolo in eventi di mutazione genica, interpretazione di variazioni sinonime e progettazione dell'algoritmo per l'ottimizzazione dei codoni.

Gli autori della presente invenzione, come divulgato in dettaglio nella sezione sperimentale della presente applicazione, hanno comparato innovativamente i valori di utilizzo dei codoni tra geni causanti-malattia (DC) e non-causanti-malattia (NDC) come espresso in specifici tessuti, e attraverso mammiferi, perciò utilizzare un approccio basato sulla malattia, per esplorare i valori CU e il comportamento bias di utilizzo dei codoni (CUB), per poter esplorare l'influenza di questi fattori su valori CU.

Pertanto, gli Autori della presente invenzione per la prima volta hanno identificato 3 nuove metriche utili per calcolare il CUB in geni specifici causanti rare malattie per poter ridefinire e integrare i parametri usati per la progettazione di geni sintetici:

1. Per poter ottenere un gene CUB, è necessario un calcolo tra le specie (mammiferi) per controllare la conservazione del codone in quel gene

specifico. Questo trend di conservazione varia a seconda del tipo di gene. Questa analisi deve essere fatta confrontando la conservazione del codone di quel gene specifico tra i mammiferi.

2. Per conferire un robusto significato statistico, gruppi di geni che causano le stesse malattie legate ai tessuti sono confrontati tra i mammiferi, raccogliendo le sequenze di una pluralità di geni non causanti malattia esclusivamente o preferenzialmente espresse nello stesso tessuto e organismo come un controllo di gruppo;
3. Comparare i valori CU tra i geni di gruppo di controllo (non causanti malattie) e gruppi di geni causanti malattie rare.

La presente invenzione offre anche un metodo che ridefinisce i criteri attualmente usati nella preparazione di prodotti di geni sintetici che possono essere proteine, reagenti ricombinanti, e strumenti molecolari che in fase di deposito della presente invenzione erano in ordine: riposizionare i codoni con l'incremento di quelli ricchi in GC (guanina e citosina) e ridurre quelli ricchi in AT (adenina, timina) contenuto delle 3 metriche di cui sopra; correggere la struttura dell'mRNA secondario e terziario riposizionando alcuni codoni considerati perturbanti la disponibilità delle molecole di mRNA per la traslazione.

Un primo oggetto della presente invenzione è un metodo implementato per computer per la determinazione del valore del bias di uso del codone di un selezionato gene causante malattia in uno o più tessuti di un organismo.

Un ulteriore oggetto della presente invenzione è un metodo implementato per computer per progettare una molecola sintetica di acido nucleico di un selezionato gene causante malattia espressa in uno o più tessuti di un organismo.

Un ulteriore oggetto della presente invenzione è un metodo per preparare una molecola sintetica di acido nucleico di un selezionato gene causante malattia.

Un ulteriore oggetto della presente invenzione è una proteina, un reagente ricombinante o uno strumento molecolare comprendente la sequenza di acido nucleico ottimizzata per il codone ottenibile dall'eseguimento dei metodi della presente invenzione.

Un ulteriore oggetto della presente invenzione è l'uso di codoni prioritarizzati come determinato nei metodi della presente invenzione, in un metodo per la valutazione dello sviluppo di una malattia correlata alla mutazione di un selezionato gene.

Un ulteriore oggetto della presente invenzione è un programma per computer comprendente le istruzioni che, quando il programma è eseguito da un computer, comporta che il computer svolga passaggioi passaggi dei metodi quivi divulgati.

Vantaggi aggiuntivi e/o forme di realizzazione della presente invenzione saranno evidenti dalla successiva descrizione dettagliata.

Breve descrizione dei disegni

La presente invenzione e la seguente descrizione dettagliata delle preferibili realizzazioni della stessa possono essere meglio comprese con riferimento alle seguenti figure:

Figura 1. Valori CU nei geni del muscolo, della pelle e dei reni nell'*Homo sapiens.* Diagrammi di calore sono stati generati usando il pacchetto R gplots. I file sono stati raggruppati basandosi sulla distanza euclidea. La codifica dei colori varia dal blu scuro al rosso con valori CU da bassi ad alti rispettivamente. I raggruppamenti gerarchici di uso del codone sinonimo in tutti i geni studiati in differenti tessuti (muscolo, pelle, rene) sono stati generati in HSA. Il grafico del calore mostra che il raggruppamento di codoni usato frequentemente (in rosso) e dei codoni usati raramente (blu scuro) varia enormemente tra geni e tessuti. Nei geni muscolari, i codoni estremamente degeneri (valori di CU bassi, colore chiave blu scuro, o valori CU alti, colore chiave rosso scuro) sono raggruppati strettamente in termini di genere e di tipo di codone, mentre valori intermedi di CU (colore chiave azzurro o giallo) sono più dispersi negli alberi (pannello A). Tra i geni muscolari, solo il DMD non mostra CUB estremo, dal momento che non compaiono punti rossi (corrispondenti a valori più alti di CU) (figura 1A). Le impronte CUB dei geni della pelle (pannello B) mostra una prevalenza di valori CU bassi di 726 codoni (blu scuro), con pochi dispersi, non raggruppati, punti giallo-rossi distribuiti non omogeneamente (valori di CU alti e intermedi). I raggruppamenti CU sono meno definiti se comparati al muscolo. Nei geni dei reni (pannello C), le impronte CUB differiscono dagli altri due gruppi di geni. La

stragrande maggioranza dei geni hanno valori CU bassi o intermedi (raggruppamenti di punti giallo-blu ampi e diffusi) con un raggruppamento di alti valori di CU strettamente raggruppati e legati a UMOD, BSND, SLC22A8, MIOX, AQP6, PKD1, SLC12A3 e geni GGACT. Interessante in fatto che tutti questi geni sono DC.

Figura 2. Valori di CU tra i mammiferi. I diagrammi di calore sono stati generati usando il pacchetto R gplots. I file sono stati raggruppati basandosi sulla distanza euclidea. Il colore chiave dei valori CU varia dal blu al rosso con valori del CUB rispettivamente bassi e alti. I valori CU in tutti i 60 geni tessuto-specifici tra 15 specie mammifere nell'albero filogenetico dei metazoi sono mostrati nei pannelli A, B, e C, e i valori CU dei singoli geni in tutte le specie sono mostrati nei pannelli D, E, ed F. Impronte CUB tesuto-specifiche sono molto evidenti con un trend conservato tra i mammiferi (pannelli A, B, C). I codoni CAG, AAG, CAC, GAC, GAG, AUC, AAC, UAC, UGC, e UUC sono quelli usati più frequentemente (punti rossi) in tuti i tessuti e tra i mammiferi, mentre i codoni UUA, CUA, UCG, CGU, CUU, GUA, CGA, AUA, UCA, UUG e GCG sono i più rari (punti blu) nei geni muscolari e della pelle, ma non dei reni. I CUB del muscolo (pannello A) e della pelle (pannello B) hanno un’impronta simile, sebbene i geni muscolari hanno molti più codoni con valori CU più bassi paragonati ai geni della pelle; i geni dei reni mostrano un raggruppamento molto diverso.

I pannelli D, E, ed F mostrano valori CU del gene che raggruppa tutti i mammiferi. Benché un chiaro raggruppamento è scarsamente visibile, i tessuti hanno differenti impronte CUB caratterizzate da diversi dendrogrammi legati al gene. Infatti, 13/30 geni nel muscolo (pannello D), 4/40 geni nella pelle (pannello E) e 8/20 geni nei reni (pannello F) mostrano raggruppamenti di alti valori CU. I tipi dei codoni variano anche di conseguenza a valori CUdal momento che i valori alti di CU sono raggruppati nei gruppi di geni elencati di cui sopra. Questa scoperta sostiene che i valori CU hanno un’impronta legata al tessuto che è ancora mantenuta e che raggruppa tutti i mammiferi, e che alcuni valori CU gene-specifici e tipi di codoni sono osservabili.

Figura 3. Comparazione di valori relativi (CUB) di uso di codoni sinonimi in geni causanti malattia (DC) e non causanti malattia (NDC) tra le specie.

I diagrammi di calore sono stati generati usando il pacchetto R gplots. I file sono stati raggruppati in base alla distanza euclidea. Il colore chiave varia dal blu al rosso con valori del CUB rispettivamente da bassi ad alti. Abbiamo raggruppato i geni basandoci sulla loro propensione ad essere il sito di variazioni patogene (mutazioni) causanti rare malattie (geni causanti malattia o DC). I pannelli da A ad F mostrano tutti i tipi di codoni senza raggruppamento gerarchico, elencati nello stesso ordine in tutti i geni e tessuti, e sotto-raggruppati in DC e NDC. I valori CU in questi pannelli mostrano che il più frequente o il più raro tipo di codone sono sovrapponibili nei gruppi di geni.

Due gruppi di codoni di colore chiave blu scuro e rosso scuro si verificano in tutti i 6 pannelli, indicando che l'assoluta frequenza del tipo di codone è simile anche tra geni e mammiferi, possibilmente supportando una tendenza evoluzionaria del CUB. Nondimeno, la variabilità valori più alti di CU può essere visibile in codoni con frequenza intermedia (colore chiave da azzurro a giallo) che può differenziarsi tra gruppi di geni. I geni muscolari NDC (pannello B) hanno un numero più basso di valori intermedi di CU (colore chiave giallo), seguiti dai geni della pelle DC (pannello C) e i geni dei reni (pannello F). CAG è il codone usato più frequentemente in tutti i geni e tra i mammiferi. I pannelli G-L mostrano il raggruppamento gerarchico di valori di CU nelle stesse categorie di gruppi. Impronte CUB chiaramente riconoscibili possono essere osservate nei geni NDC e DC. Questo è maggiormente evidente nei geni muscolari (pannelli G e H) e parzialmente nei geni dei reni (pannelli K e L). I geni muscolari DC hanno raggruppamenti compatti di codoni estremamente frequenti (AAG, CAG, GAG) e codoni estremamente rari (UGG, UUA, CUA), una tendenza conservata tra i mammiferi e con 774 gruppi di codoni (in termini di distanza dall'albero). I colori blu scuro e rosso (codoni più rari e frequenti) predominano in geni muscolari NDC, con pochi codoni con valori intermedi (colore chiave giallo). Questo suggerisce che un forte CUB si è verificato nei geni NDC. Di conseguenza, i geni muscolari NDC mostrano valori di CU più alti e più bassi raggruppati insieme, suggerendo una possibile tendenza evoluzionistica diversa. I geni della pelle NDC e DC (pannelli I-J) mostrano simili impronte CUB con poche differenze. I geni della pelle NDC mostrano nel complesso valori di CU più bassi (pannello J,

parte superiore) e viceversa più codoni con valori di CU intermedi (pannello J, parte inferiore) paragonati ai geni DC. Le impronte CUB dei geni del rene DC e NDC (pannelli K-L) differiscono grandemente dagli altri due gruppi dal momento che i raggruppamenti gerarchici sono opposti. Anche se la conservazione di valori di CU si verifica anche tra i mammiferi, la gerarchia dei dendrogrammi del gene del rene mostra un antenato comune per i valori di CU intermedi e bassi e non due distinte linee (valori di CU alti e bassi) come osservabile nei geni muscolari e della pelle. Le due impronte CUB del rene sono in qualche modo simili; comunque, i geni DC mostrano un più alto numero di valori CU molto bassi.

Figura 4. I 5 codoni estremamente degeneri, usati differentemente dall'*Homo sapiens* in geni DC vs NDC, e tra i tessuti. I grafici sulla sinistra (A, B, C, D, E) mostrano la frequenza di codoni sull'asse delle x e il numero dei geni che usano quello specifico codone (con la frequenza annotata nell'asse delle x) sull'asse delle y. I grafici sulla destra (F, G, H, I, L) mostrano i valori di CU in tutti i mammiferi studiati sull'asse delle x e i valori di uso dei codoni nei tessuti genetici sull'asse delle y. Cinque codoni erano usati più differentemente nell'*HSA*, CGU (Arg), CCA (Pro), GAC (Asp), GAU (Asp), e GUA (Val). Le barre rosa indicano i geni DC, le barre blu i geni NDC. CGU (A, F) è il codone usato meno di frequente nei geni muscolari NDC; CCA (B, G) è il codone usato meno di frequente nei geni muscolari DC; GAC (C, H) è il codone usato più di frequente nei geni della pelle DC; GAU (D, I) è il 798° codone usato più frequentemente nei geni della pelle NDC; GUA (E, L) è il codone usato meno di frequente nei geni del rene DC. Comparando i valori di CU di questi 5 codoni estremamente degeneri tra gruppi di geni DC e NDC e tra i mammiferi, può essere apprezzata la tendenza verso una pesante estremizzazione dei codoni durante l'evoluzione. Durante l'evoluzione, CGU e CCA hanno iniziato ad essere usati di più nei geni muscolari DC, GAC ha iniziato ad essere più utilizzato nei geni della pelle DC, GAU ha iniziato ad essere più utilizzato nei geni della pelle NDC, e GUA ha iniziato ad essere più utilizzato nei geni del rene NDC. Questo suggerisce uno specifico codone, orientato alla malattia CUB, apparentemente conservato tra i mammiferi.

Figura 5. Uso del codone del gene DMD negli umani e approccio "mapping-on-codon". Pannello A valori di uso del codone DMD umano e

percentuali di mutazione. Le barre rappresentano i valori CU nei geni *DMD* nell'*HSA*. Sull'asse delle x sono elencati i tipi di codone e i relativi aminoacidi, sull'asse delle y sono riportati i valori di CU. Le barre rosse rappresentano i 4 codoni *DMD* usati raramente, UCG, CCG, ACG E GCG, in base al nostro taglio, che è basato sulla ridondanza del codone di 2, 3, 4 e 6 triplette (vedi Metodi).

In cima alle barre (lato destro), sono riportati i numeri di mutazioni missenso e nonsenso verificatesi ai relativi codoni *DMD*. Tutti i codoni sono ancora usati dal gene umano *DMD*, e la frequenza del verificarsi della mutazione non è collegata ai valori di uso del codone. Esempi sono UAU, che è il codone più utilizzato, ma con solo 36 mutazioni "mappate", oppure CGA (Arg) che è usata raramente ma ha 114 mutazioni "mappate".

Pannello B. Mappatura delle mutazioni missenso e nonsenso della *DMD* umana sui tipi di codoni. Sull'asse delle x, ci sono i codoni e i relativi aminoacidi, sull'asse delle y, c'è il numero di mutazioni che si sono verificate e "mappate". Secondo il nostro taglio le barre rosse rappresentano i 4 codoni usati raramente nel gene umano *DMD*. In cima alle barre c'è la percentuale di valori CU 822. Il numero delle mutazioni verificatesi e i valori CU non sono strettamente collegati. Esempi sono CAG, che è il codone mutato "più frequentemente" (56%), con alti valori di CU, e UGC (Cys), raramente sito di mutazioni, ma con valori CU molto alti.

Figura 6. L'analisi della correlazione di Spearman tra geni DC e NDC nei tessuti genetici del muscolo, della pelle e del rene di *HSA.* Il test ha dimostrato che i valori CU dei geni DC e NDC si correlano significativamente nel muscolo nella pelle e nel rene ($p<0.05$).

Figura 7. I diagrammi di calore sono stati generati utilizzando il pacchetto R gplots. I file sono stati raggruppati in base al sistema della distanza euclidea. Il codice dei colori varia dal blu scuro al rosso con valori di CU rispettivamente da alti a bassi. L'impronta CUB e i valori di CU tra le espressione di geni *HSA* alto-, medio-, e basso. I geni DC NDC erano considerati dipendenti dal loro livello di espressione. Le impronte CUB hanno una forte similitudine, il che significa che infatti, raggruppare i geni secondo il loro livello di espressione produce una tendenza di valori CU. I codoni AAC, GAC, UGC, UAC, CAC, UUC, AUC, AAG,

GAG e CAG sono usati più frequentemente sia in geni alto-espressi che in geni medio-espressi mentre GUG è presente solo in geni alto-espressi. I codoni GAC, CAC, UUC, CAG, UAC, UGC, AAG, AUC, GAG, AAC, ACC, GGC, GUC, UCC e GCG sono usati più frequentemente in geni espressi di basso livello. Pochi codoni hanno valori di CU più bassi come UCC (Ser), ACC (Thr), GGC (Gly), GUC (Val) e GCG (Ala) in geni basso-espressi. Alcuni geni DC alto-espressi hanno più codoni con valori di CU più alti. Come *DYS, LMNA* e *DES* (muscolo), *UMOD* e *PKD1* (rene) e *FGFR3* (pelle). Nei geni medio-espressi la tendenza è opposta, con alcuni geni NDC che mostrano valori di CU più alti come *MLPF, TNNC2, TMEM3BA* (muscolo) e *NCLZ2, MCX* (rene). È interessante notare che UAA è il codone-stop più utilizzato nei geni alto-espressi, dal momento che induce il termine della traduzione con maggiore velocità e accuratezza a livello ribosomiale e può essere letta sia da fattori di rilascio eRF1 che eRF2. UAG e UGA hanno frequenza simile in tutti i tessuti genetici e livelli di espressione.

Figura 8. Il diagramma di flusso schematico di un metodo secondo il corpo della presente invenzione.

Glossario

L'uso delle forme singolari "un", "uno", "una", "il", "lo", "la" includono riferimenti plurali a meno che il contesto non indichi diversamente. Per esempio, riferimenti a "polinucleotide" include una pluralità di polinucleotidi, riferimenti a "substrato" include una pluralità di detti substrati, riferimenti a "una variante" includono una pluralità di varianti, ecc.

Dove è riportata una gamma di valori, deve essere compreso che ogni valore intero intervenuto, e ogni frazione dello stesso, tra i limiti riportati in alto e in basso di quella serie è inoltre specificamente divulgato, insieme a ogni sottoserie tra detti valori. I limiti superiori e inferiori di ogni serie possono essere indipendentemente inclusi nella serie, o esclusi da essa, e ogni serie in cui uno, nessuno o entrambi i limiti sono inclusi sono anche comprese nell'invenzione. Dove un valore discusso ha limiti inerenti (per esempio, dove un componente può essere presente a una concentrazione da 0 a 100%, o dove il pH di una soluzione

acquosa può variare da 1 a 14), quei limiti inerenti sono specificatamente discussi.

Dove un valore è esplicitamente riportato, deve essere compreso che valori che sono della stessa quantità o ammontare del valore riportato rientrano anche nello scopo dell'invenzione. Dove è riportata una combinazione, ogni sotto-combinazione degli elementi di quella combinazione è specificamente discusso e rientra nello scopo dell'invenzione. Al contrario, dove elementi diversi o gruppi di elementi sono discussi individualmente, combinazioni dello stesso sono anche discusse. Dove qualsiasi elemento di un'invenzione è descritto come avente una pluralità di alternative, esempi di quell'invenzione in cui ogni alternativa è esclusa singolarmente, o in qualsiasi combinazione con le altre alternative, sono discusse con la presente (più di un elemento di un'invenzione può avere queste esclusioni, e tutte le combinazioni di elementi aventi queste esclusioni sono discussi nella presente).

A meno che diversamente previsto, tutti i termini tecnici e scientifici usati qui hanno lo stesso significato come comunemente intese da una delle abilità ordinarie nell'arte della genetica, della bioinformatica e di progettazione genica. Ogni metodo e materiale simile o equivalente a quelli qui descritti può essere usato nella pratica o in fase di test delle realizzazioni dell'invenzione, anche se alcuni metodi e materiali sono esemplificati da quelli qui discussi.,

Il bias di uso del codone: come qui usato, il termine "bias di uso del codone", o semplicemente "uso del codone", si riferisce alle differenze nella frequenza del verificarsi di un particolare codone come opposto ad altri codoni sinonimi, nella codificazione del DNA, per la codificazione di un aminoacido all'interno di un organismo. Un bias di uso del codone può essere espresso come una misurazione quantitativa del tasso al quale un particolare codone è usato nel genoma di un particolare organismo, per esempio, se comparato ad altri codoni che codificano lo stesso aminoacido. Negli oggetti elencati in questo documento sono considerati o prioritizzati codoni la cui frequenza di uso è statisticamente differente paragonata agli altri codoni sinonimi, sia alti che bassi.

Vari metodi sono conosciuti da quelli di competenza nell'arte di determinare l'uso del metodo indice di adattamento del codone (CAI), che è essenzialmente

una misurazione della distanza di uso del codone di un gene all'uso del codone di un set predefinito di geni altamente espressi.

Sharp e Li (1987) Nucleic Acids Res. 15:1281-95. Quindi, il bias di uso del codone include le frequenze relative dell'uso dei codoni che codificano lo stesso aminoacido ("codoni sinonimi"). Un bias può verificarsi naturalmente; per esempio, il bias del codone nel genoma di un organismo riflette l'uso complessivo di codoni sinonimi all'interno di tutti i geni in quell'organismo. Un bias può anche essere usato in un algoritmo computazionale, dove, per esempio, può essere usato per determinare la frequenza relativa con cui codoni sinonimi differenti sono selezionati per essere utilizzati nella progettazione di una sequenza polinucleotidica. Similarmente, la frequenza "relativa" di ogni elemento della sequenza usato per codificare un polipeptide all'interno di una sequenza nucleotidica con la quale quell'elemento della sequenza è utilizzato per codificare una caratteristica del polipeptide, diviso dal numero degli accadimenti all'interno del polipeptide in una data cornice di lettura di caratteristiche cheche possono essere codificate da quell'elemento della sequenza.

Il bias di uso del codone può essere dedotto da una tavola di uso del codone per una particolare espressione di un organismo ospitante. Le tavole di uso dei codoni sono prontamente reperibili per molte delle espressioni di organismi ospiti. Vedi. E.g., Nakamura et al. (2000) Nucleic Acids Res. 28:292 (Database di Uso del Codone- versioni aggiornate disponibili a kazusa.or.jp/codon).

I termini "tavola di uso del codone", oppure "tavola del bias del codone), oppure "tavola della frequenza del codone" sono usate indifferentemente e descrivono una tavola che correla ogni codone che può essere usato per codificare un particolare aminoacido con le frequenze con cui ogni codone è usato per codificare quell'aminoacido in un organismo specifico, all'interno di una specifica classe di geni all'interno di quell'organismo, o all'interno di uno o più polinucleotidi sintetici.

Frequenza assoluta del codone: come qui utilizzato, il termine "frequenza assoluta del codone" si riferisce alla frequenza con cui appare un codone relativa al numero totale di codoni sinonimi all'interno di un polinucleotide o set di polinucleotidi in una data cornice di lettura (e.g., una cornice di lettura che è usata

per codificare un polipeptide di interesse). Similarmente, la frequenza “assoluta” di ogni elemento della sequenza usato per codificare un polipeptide all’interno di un polinucleotide è la frequenza con cui quell’elemento della sequenza è usato per codificare una caratteristica (e.g., aminoacido, coppia di aminoacidi, ecc.) del polipeptide, diviso dal numero di occorrenze all’interno del polipeptide di caratteristiche della stessa misura di quelle che potrebbero essere codificate da quell’elemento della sequenza.

Cornice di lettura aperta: come qui utilizzato, il termine “cornice di lettura aperta” si riferisce a tutte le possibili sequenze di polinucleotidi che possono essere usate per codificare uno specifico polipeptide, attraverso la variazione dei codoni usati per codificare aminoacidi all’interno del polipeptide.

Sostituzione del codone: come qui utilizzato, il termine “sostituzione del codone” si riferisce all’alterazione di una sequenza codificante nucleotidica attraverso il cambiamento di uno o più codoni che codificano uno o più aminoacidi di un polipeptide codificato, senza alterare la sequenza di aminoacidi del polipeptide codificato.

Ottimizzazione del codone: Come qui utilizzato, il temine “ottimizzazione del codone” si riferisce ai processi messi in atto per modificare una sequenza codificatrice esistente, oppure per la progettazione di una sequenza codificatrice in prima istanza, per esempio, per migliorare la traduzione in una cellula o organismo ospite di una molecola di RNA trascritto, trascritta dalla sequenza codificante, o per migliorare la trascrizione di una sequenza codificante. L’ottimizzazione del codone include, ma non è limitata, a processi che includono la selezione di codoni per la sequenza codificante per adattarsi alla preferenza di codone di espressione dell’organismo ospite. L’ottimizzazione del codone include anche, per esempio, il processo a cui a volte ci si riferisce come “armonizzazione del codone”, in cui i codoni di una sequenza di codoni che vengono riconosciuti come codoni a basso-utilizzo nell’organismo di origine sono alterati a codoni che vengono riconosciuti come a basso utilizzo nella nuova espressione ospite. Questo processo può permettere ai polipeptidi espressi di piegarsi normalmente tramite l’introduzione di pause naturali e appropriate durante la traduzione/estensione. Birkholtz et al. (2008) Malaria J. 7:197-217.

Modificare: Come qui utilizzati, i termini “modificare” o “alterare”, o qualsiasi altra forma dello stesso, significano modificare, alterare, riposizionare, cancellare, sostituire, rimuovere, variare, o trasformare.

Molecola di acido nucleico: come qui utilizzato, il termine “molecola di acido nucleico” si riferisce a una forma polimerica di nucleotidi, che può includere filamenti di RNA senso e anti-senso, cDNA, DNA genomico, e forme sintetiche e polimeri misti di cui sopra. Un nucleotide può riferirsi a un ribonucleotide, deossiribonucleotide, o una forma modificata dei due tipi di nucleotide. Una “molecola di acido nucleico” come qui usato è sinonimo con “acido nucleico” e “polinucleotide”. Una molecola di acido nucleico ha generalmente la lunghezza di 10 basi, a meno che diversamente specificato. Il termine include forme a filamento singolo o doppio di DNA. Una molecola di acido nucleico può includere uno o entrambi tra nucleotidi verificatisi naturalmente o modificati, legati insieme da legami nucleotidici verificatisi naturalmente o non-naturalmente.

Proteina/polipeptide: i termini “proteina” e “polipeptide” sono qui usati indifferentemente. I termini si riferiscono alla catena molecolare contigua di aminoacidi legati attraverso legami peptidici. I termini non si riferiscono a una specifica lunghezza del prodotto. Perciò, “peptidi”, “oligopeptidi”, e “proteine” sono inclusi nella definizione di polipeptide. I termini includono polipeptidi contenenti modifiche co- e/o post-traslazionali del polipeptide fatti in vitro o in vivo; per esempio e senza limitazione: glicosilazione, acetilazione, fosforilazione, PEGilazione e solfatazione. Inoltre, frammenti di proteine, analoghi (compresi gli aminoacidi non codificati nel codice genetico: e.g., omocisteina, ornitina, p-acetilfenilalanina, D-amminoacidi e creatina), mutanti naturali o artificiali, varianti, proteine di fusione, residui derivatizzati (per esempio, alchilazione di gruppi amminici, acetilazione o esterificazione di gruppi carbossilici), e combinazioni di qualsiasi di cui sopra sono inclusi nel significato di polipeptide.

Come qui utilizzato, il termine “percentuale di identità di sequenza” può riferirsi al valore determinato dall’allineamento di due o più sequenze (e.g., sequenze di acido nucleico e sequenze di aminoacidi) su una finestra di confronto, in cui la porzione della sequenza nella finestra di confronto può comprendere addizioni o rimozioni) per l’allineamento ottimale delle due sequenze. La percentuale è

calcolata dalla determinazione del numero di posizioni alle quali il nucleotide identico o il residuo aminoacido si verifica in entrambe le sequenze per produrre il numero di posizioni abbinate, dividendo il numero delle posizioni abbinate per il numero totale di posizioni nella finestra di confronto, e moltiplicando il risultato per 100 per produrre la percentuale di identità di sequenza. Metodi per l'allineamento delle sequenze per il confronto sono ben conosciute nel campo.

Sintetico: come qui utilizzato in riferimento a una sequenza nucleotidica (o molecola di acido nucleico comprendente una sequenza nucleotidica sintetica), il termine "sintetico" si riferisce a una sequenza che è progettata (e.g., in silicio), per esempio, con lo scopo si esprimere un polipeptide codificato di interesse. Il termine "nucleotide sintetico" include anche il prodotto della manifattura di una molecola di acido nucleico per mezzo di oligonucleotidi sintetizzati chimicamente attraverso metodologie in vitro o in vivo conosciute da coloro i quali sono esperti nella sintesi di geni, o attraverso combinazioni di metodi in vitro o in vivo.

Descrizione dettagliata

Nella seguente, saranno descritte diverse realizzazioni dell'invenzione. È inteso che le caratteristiche delle varie realizzazioni possono essere combinate, laddove è compatibile. In generale, realizzazioni susseguenti saranno discusse solo in relazione alle differenze con quelle precedentemente descritte.

Come precedentemente menzionato, un primo oggetto della presente invenzione è rappresentato da un metodo implementato al computer per determinare il valore del bias di uso del codone di un selezionato gene causante malattia.

In una forma di realizzazione questo metodo comprende i seguenti passaggi:

(i) raccogliere le sequenze di uno o più geni causanti-malattia espressi in uno o più tessuti di un organismo;

(ii) raccogliere le sequenze di una pluralità di geni non-causanti-malattia espressi nello stesso tessuto e organismo dei geni nel passaggiopassaggio (i);

(iii) determinare il calcolo indipendente della frequenza di uso del codone per i 19 aminoacidi essenziali (metionina e triptofano esclusi) in ogni gene raccolto negli passaggio (i) e (ii);

(iv) comparare la frequenza di uso del codone determinata nello passaggio (iii) in modo da ottenere il valore del bias di utilizzo del codone.

Preferibilmente, il metodo della presente invenzione comprende un ulteriore passaggio in cui, per poter ottenere un gene CUB, è necessario un calcolo tra le specie (mammiferi) per controllare la conservazione del codone in quel gene specifico, laddove il gene specifico è il gene selezionato causante-malattia. Questa tendenza di conservazione varia a seconda del tipo di gene. Questa analisi deve essere svolta comparando la conservazione del codone di quel gene specifico tra i mammiferi. Detto calcolo tra le specie è utilizzato nel metodo per ottenere il valore del bias di uso del codone.

In una forma di realizzazione dell'invenzione detto organismo è un mammifero. In un'altra realizzazione dell'invenzione, detti geni causanti-malattia sono raccolti da diversi mammiferi differenti. Preferibilmente, detti mammiferi sono selezionati da *R. ferrumequinum* (pipistrello ferro di cavallo maggiore), *M. musculus* (topo), *F. catus* (gatto), *C. lupus familiaris* (cane), *E. caballus* (cavallo), *B. taurus* (bovino), *M. murinus* (lemure topo grigio), *G. variegatus* (colugo della Sonda), *C. jacchus* (uistitì comune), *M. mulatta* (macaco), *N. leucogenys* (gibbone), *P. abelii* (orangotango), *G. gorilla* (gorilla), *P. troglodytes* (scimpanzé) e *H. sapiens* (umano).

In una ulteriore forma di realizzazione questo metodo è un metodo implementato per computer per la progettazione di una molecola sintetica di acido nucleico di un selezionato gene causante-malattia, che comprende i seguenti passaggi:

(i) raccogliere le sequenze di uno o più geni causanti-malattia esclusivamente o preferibilmente espressi in uno o più tessuti di un organismo;

(ii) raccogliere le sequenze di una pluralità di geni non-causanti-malattia esclusivamente o preferibilmente espressi nello stesso tessuto ed organismo dei geni nello passaggio (i);

(iii) determinare il calcolo indipendente della frequenza di utilizzo del codone per i 19 aminoacidi essenziali (metionina e triptofano esclusi) in ogni gene raccolto negli passaggio (i) e (ii);

(iv) comparare la frequenza di utilizzo del codone determinata nello passaggio (iii) quindi ottenendo il valore del bias di utilizzo del codone, per identificare codoni a comparsa (codoni tessuto-specifici e codoni gene-specifici) per poter dare priorità ai codoni più usati in modo diverso nel gene e nel/nei tessuto/i di interesse;

(v) progettare una molecola di acido nucleico di detto gene causante-malattia, modificando la struttura secondaria o terziaria di detto gene utilizzando i codoni prioritizzati nello passaggio (iv).

Secondo il punteggio delle banche dati pubbliche nel descrivere o valutare l'espressione di un gene in un tessuto, “esclusivamente” è usato quando un gene è espresso solo in quel particolare tessuto, mentre “altamente” è usato quando un gene è espresso anche in altri tessuti ma (risulta) sostanzialmente inferiore rispetto al tessuto considerato, “preferenzialmente” se il gene è espresso in altri tessuti ma il tessuto considerato mostra un’espressione più alta. Detti tessuti sono per esempio il muscolo, la pelle, il rene o qualsiasi altro tessuto coinvolto in malattie suscettibili con terapia genica.

In una forma di realizzazione, detta malattia è selezionata da distrofie muscolari, miopatie congenite, malattia tubulointestinale del rene, rene policistico di tipo 1, ipercheratosi epidermolitica, displasia ectodermica.

In alcune forme di realizzazione dette malattie sono tipi di malattia rara (RD), per esempio legate al muscolo, alla pelle, al rene o altre mutazioni del gene dei tessuti. Malattie rare del muscolo includono per esempio distrofie muscolari e miopatie congenite, i cui geni causanti sono espressi in maniera predominante e alta nei muscoli scheletrici. Da cui, lo scopo del metodo è l’ottimizzazione della sequenza nucleotidica di un gene selezionato espresso in uno o più tessuti, che è coinvolto in una particolare malattia, in particolare tutte le malattie suscettibili con terapia genica.

La pluralità di geni nei passaggi (i) e/o (ii) significa un gruppo di geni dei quali si conosce il coinvolgimento o meno nella malattia del gene selezionato e che sono espressi, preferibilmente in maniera esclusiva, altamente o preferenzialmente espressi, nello stesso tessuto dello stesso organismo del gene selezionato. Preferibilmente, il numero di geni raccolti nei passaggi (i) e/o (ii) è il più alto numero disponibile nella banca dati per la malattia selezionata, oppure per esempio dal 70 al 99%, preferibilmente dall'80 al 99%.

Il bias di utilizzo del codone può essere determinato secondo la procedura di calcolo conosciuta nel campo, per esempio può essere determinata dal metodo di indice di adattamento del codone (CAI), che è essenzialmente una misurazione della distanza dell'utilizzo del codone di un gene all'utilizzo del codone di un set predefinito di geni altamente espressi, usando in questo caso come set di dati di riferimento i geni causanti-malattia e/o geni non-causanti-malattia, il CAI sarà in questo caso modificato introducendo una correzione basata sul set di dati di riferimento usati e piò essere determinato un bias di uso del codone migliorato. Eppure, secondo l'invenzione il CAI sarà modificato utilizzando come set di dati di riferimento predefinito i geni causanti-malattia e/o non-causanti-malattia, e successivamente introducendo una correzione basata sul diverso set di dati di riferimento e può essere determinato un bias di utilizzo del codone migliorato.

In una forma di realizzazione, detto passaggio (iv), di dare la priorità ai codoni più differentemente usati è effettuato raggruppando i valori di frequenza di utilizzo del codone ottenuti nello passaggio (iii) usando un algoritmo di raggruppamento gerarchico con un valore-p minore dello 0.05 oppure applicando un'analisi prioritaria in termini di percentuali CUB e selezionando i codoni meno/più utilizzati per le specie, il tessuto e il gene di interesse.

In un'altra forma di realizzazione dell'invenzione, detto passaggio (v), di modificare la struttura secondaria o terziaria di detto gene causante-malattia, è svolto tramite la sostituzione e/o rimozione dei codoni meno utilizzati e più utilizzati, ottenuti nel passaggio (iv), nella sequenza di detto gene causante-malattia.

Un ulteriore oggetto della presente invenzione è un metodo per preparare una molecola di acido nucleico di un selezionato gene causante-malattia,

comprendendo i passaggi discussi secondo ogni realizzazione qui discussa e un ulteriore passaggio (vi) di sintetizzare di una molecola di acido nucleico comprendente la sequenza di acido nucleico ottimizzata per il codone del passaggio (v).

Il passaggio ulteriore (vi) rispetto al precedentemente descritto metodo di sintetizzazione di una molecola di acido nucleico comprendente o consistente nella sequenza di acido nucleico ottimizzata per il codone progettata secondo il precedente passaggio (v), è svolto con procedure e protocolli per la sintetizzazione di molecole di acido nucleico di una data sequenza nota nell'arte.

In una forma di realizzazione della presente invenzione il metodo comprende i seguenti passaggi:

(i) raccogliere le sequenze di uno o più geni causanti-malattia esclusivamente o preferenzialmente espressi in uno o più tessuti di un organismo;

(ii) raccogliere le sequenze di una pluralità di geni non-causanti-malattia esclusivamente o preferenzialmente espressi nello stesso tessuto e organismo dei geni nello passaggio (i);

(iii) determinare il calcolo indipendente della frequenza di utilizzo del codone per i 19 aminoacidi essenziali (escluse metionina e triptofano) in ogni gene raccolto nei passaggio (i) e (ii);

(iv) comparare la frequenza di utilizzo del codone determinata nel passaggio (iii) in modo da ottenere il valore del bias di utilizzo del codone, per identificare codoni a comparsa (codoni tessuto-specifici e codoni gene-specifici) per poter prioritizzare i codoni usati più diversamente usati nel gene e nel/nei tessuto/i di interesse;

(iv-a) ridurre i codoni rari (eccetto i codoni "a comparsa" identificati nei passaggi precedenti);

(iv-b) aumentare il contenuto di guanina (G) e citosina (C) senza intaccare i codoni "a comparsa", ciò significa che se la sostituzione di una A con una G eliminerà un codone "pop up", questo non dovrebbe essere fatto;

(v) progettare una molecola di acido nucleico di detto gene causante-malattia, modificando la struttura secondaria o terziaria di detta molecola di acido nucleico senza eliminare nessun codone a comparsa.

In una forma di realizzazione il metodo della presente invenzione è un metodo per preparare un prodotto genico sintetico che potrebbero essere le proteine, reagenti ricombinanti, e strumenti molecolari. La sintesi di questi prodotti comprendenti o consistenti nella sequenza di acido nucleico ottimizzata per il codone progettata secondo qualunque realizzazione dei metodi qui discussi. Procedure e protocolli per la sintetizzazione di qualsiasi delle sopra-citate molecole di una data sequenza sono conosciute nel campo.

La molecola di acido nucleico sintetica ottenuta con questi metodi può essere usata per esempio nei trattamenti di malattie suscettibili con terapia genica o in esprimenti in vitro per valutare l'espressione di detta molecola di acido nucleico.

Un altro oggetto della presente invenzione è l'uso dei codoni prioritizzati come ottenuti nel passaggio (iv) di entrambi i metodi descritti nella presente invenzione, in un metodo per la valutazione dello sviluppo di una malattia legata alla mutazione di un selezionato gene. In alcune realizzazioni, le sequenze di acido nucleico estratte da geni causanti-malattia che codificano una trascrizione tradotta o non tradotta e/o da geni non-causanti-malattia che codificano una trascrizione tradotta o non tradotta possono essere importate (e.g., individualmente importate da una banca dati) in un programma software implementato per computer che è capace di ottimizzare la sequenza di codifica secondo i metodi qui discussi.

In un ulteriore aspetto l'invenzione è programma per computer comprendente istruzioni che, quando il programma è eseguito da un computer, comportano che il computer svolga i passaggi del metodo secondo ogni realizzazione qui discussa, per esempio un programma che gira su un web server.

Un ulteriore oggetto della presente invenzione è un supporto di memorizzazione leggibile dal computer che comprende istruzioni che, quando eseguito da un computer, comporta che il computer svolga i passaggi del metodo secondo ogni forma di realizzazione qui discussa, per esempio una chiavetta di

memoria, un CD-ROM o un dispositivo comprendente detto supporto di memorizzazione leggibile dal computer.

Un ulteriore oggetto della presente invenzione è un metodo per identificare specie, tessuti o codoni gene-critici per metodi implementati al computer, al fine di valutare il potenziale di sviluppo di una malattia causata delle varianti della sequenza del DNA. I codoni sono definiti critici basandosi sulla loro frequenza di utilizzo, e questi sono codoni estremamente rari o di uso estremamente comune per la specie, il tessuto e il gene di interesse. L'estremizzazione dell'uso potrebbe influenzare l'efficienza di traduzione e/o il ripiegamento del peptide.

Sono sotto riportati esempi che hanno lo scopo di illustrare meglio le metodologie discusse nella presente descrizione, detti esempi non devono in alcun modo essere considerati come limitazioni della precedente descrizione e delle susseguenti rivendicazioni.

Esempi e dati sperimentali

METODI

La nostra strategia era basata sulla comparazione dei valori CU e i loro raggruppamenti gerarchici nei tessuti del muscolo, del rene e della pelle nell'*Homo Sapiens,* tra i mammiferi selezionati e nei geni DC versus i geni NDC.

Selezione delle specie, dati di sequenza e risorse per il calcolo dei valori CU

Abbiamo selezionato i seguenti 15 mammiferi dell'albero filogenetico dei metazoi: *R. ferrumequinum* (Ferro di cavallo maggiore), *M. musculus* (topo), *F. catus* (gatto), *C. lupus familiaris* (cane), *E. caballus* (cavallo), *B. Taurus* (toro), *M. murinus* (lemure topo grigio), *G. variegatus* (colugo della sonda), *C. jacchus* (uistitì comune), *M. mulatta* (macaco), *N. leucogenys* (gibbone), *P. abelii* (orangotango), *G. gorilla* (gorilla), *P. troglodytes* (scimpanzé) e *H. sapiens* (umano).

Tavola 1

SEQUENCES REFERENCES	NCBI LINK
Rhinolophus ferrumequinum (Greater horseshoe bat)	https://www.ncbi.nlm.nih.gov/assembly/GCA_004115265.3
Mus musculus (House mouse)	https://www.ncbi.nlm.nih.gov/gene/13405
Felis catus (Cat)	https://www.ncbi.nlm.nih.gov/assembly/GCF_000181335.3
Canis lupus familiaris (Dog)	https://www.ncbi.nlm.nih.gov/assembly/GCF_000002285.3/
Equus caballus (Horse)	https://www.ncbi.nlm.nih.gov/assembly/GCF_002863925.1
Bos taurus (Cattle)	https://www.ncbi.nlm.nih.gov/assembly/GCF_002263795.1
Microcebus murinus (Gray mouse lemur)	https://www.ncbi.nlm.nih.gov/nuccore/1135509992
Galeopterus variegatus (Sunda flying lemur)	https://www.ncbi.nlm.nih.gov/nuccore/640470054
Callithrix jacchus (Common marmoset)	https://www.ncbi.nlm.nih.gov/assembly/GCF_000004665.1/
Macaca mulatta (Rhesus macaque)	https://www.ncbi.nlm.nih.gov/assembly/GCF_000772875.2/
Nomascus leucogenys (Northern white-cheeked gibbon)	https://www.ncbi.nlm.nih.gov/nuccore/350542784

Pongo abelii (Sumatran orangutan)	https://www.ncbi.nlm.nih.gov/nuccore/180204958
Gorilla gorilla (Western gorilla)	https://www.ncbi.nlm.nih.gov/assembly/GCF_000151905.2/
Pan troglodytes (Chimpanzee)	https://www.ncbi.nlm.nih.gov/assembly/GCF_000001515.7/
Homo sapiens (Human)	Various genes see supp table 2, 3 and 4

La tavola 1 mostra la lista di specie e sequende di assemblaggio di genoma. Le sequenze di riferimento dell'mRNA di tutti i gruppi di geni dell' *H. sapiens* sono stati recuperate da RefSeq e GeneBank al National Center for Biotechnology Information come illustrato nelle tavole 2, 3 e 4.

Tavola 2

GENI CAUSANTI MALATTIA

	GENE	**NCBI LINK**	**RNA TS TPM***	**PROTEIN EXPRESSION (score)****	**OMIM NUMBER**	**TISSUE SPECIFICITY**
1	**DYSF**: Homo sapiens dysferlin (DYSF), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001130987.1	32,8	High	603009	predominantly
2	**CAPN3**: Homo sapiens calpain 3 (CAPN3), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000070.2	336,8	Medium	114240	Predominantly
3	**SGCB**: Homo sapiens sarcoglycan beta (SGCB), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000232.4	35,4	Medium	600900	Predominantly (the highest of two)
4	**SGCA**: Homo sapiens sarcoglycan alpha (SGCA), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000023.3	181,4	Medium	600119	Predominantly

5	**LMNA**: Homo sapiens lamin A/C (LMNA), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_170707.3	55	High	150330	All
6	**DES**: Homo sapiens desmin (DES), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001927.3	5462	High	125660	Only
7	**MYOT**: Homo sapiens myotilin (MYOT), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_006790.2	766,8	High	604103	Only
8	**ANO5**: Homo sapiens anoctamin 5 (ANO5), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_213599.2	23,6	Low	608662	One of the five
9	**COL6A1**: Homo sapiens collagen type VI alpha 1 chain (COL6A1), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001848.2	29,1	Only smooth muscle	120220	Predominantly
10	**TRIM32**: Homo sapiens tripartite motif containing 32 (TRIM32), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_012210.	5,4	Medium	602290	All
11	**DMD**: Homo sapiens dystrophin (DMD), transcript variant Dp427m, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_004006.2	34,8	Medium	300377	Only

GENI NON CAUSANTI MALATTIA

5

	GENE	**NCBI LINK**	**RNA TS TPM***	**PROTEIN EXPRESSION (score)****	**TISSUE SPECIFICITY *****
1	**ACTN3**: Homo sapiens actinin alpha 3 (gene/pseudogene) (ACTN3), transcript variant 1, coding, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001104.3	556,5	Not performed	Only

2	**MYLPF**: Homo sapiens myosin light chain, phosphorylatable, fast skeletal muscle (MYLPF), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_013292.4	6541,1	Medium	Predominantly (the highest of three)
3	**TNNC2**: Homo sapiens troponin C2, fast skeletal type (TNNC2), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_003279.2	9898,9	Medium	Only
4	**ANKRD23**: Homo sapiens ankyrin repeat domain 23 (ANKRD23), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_144994.7	495,5	Medium	Predominantly (the highest of two)
5	**LBX1**: Homo sapiens ladybird homeobox 1 (LBX1), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_006562.4	9,1	Not performed	Only
6	**LSMEM1**: Homo sapiens leucine rich single-pass membrane protein 1 (LSMEM1), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_182597	31,8	Not performed	Predominantly
7	**TMEM38A**: Homo sapiens transmembrane protein 38A (TMEM38A), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_024074.2	199,7	Medium	One of two
8	**RPL3L**: Homo sapiens ribosomal protein L3 like (RPL3L), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_005061.2	323,9	Medium	Only
9	**MYH1**: Homo sapiens myosin heavy chain 1 (MYH1), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_005963.3	2753,9	High	Only

La tavola mostra la lista dei geni Homo sapiens, prioritizzati nel tessuto muscolare scheletrico, utilizzando la Human Protein Atlas database (https://www.proteinatlas.org/) Per poter prioritizzare i geni del muscolo, abbiamo selezionato quelli con una più alta

espressione dalla lista di geni arricchiti del muscolo scheletrico della Human Protein Atlas database (https://www.proteinatlas.org/search/tissue_specificity_rna:skeletal%20muscle;Tissue%20enriched+AND+sort_by:tissue+specific+score+AND+show_columns:groupenriched). Tutti i dati (RNA, TS, TPM, and Protein expression scores) sono stati ottenuti tramite Human Protein Atlas database. *RNA TS TPM indica un livello di RNA riportato come media TPM (transcripts per million), nel tessuto di riferimento, muscolo scheletrico in questo caso. **I punteggi di espressione proteica sono basati su una stima ottimale dalla "vera" espressione proteica da un'annotazione basata sulla conoscenza nel tessuto selezionato,in questo caso il muscolo scheletrico. ***La specificità del tessuto è basata su dati trovati nel grafico nominato "HPA tissue dataset", un sub-categoria della sezione "RNA sample summary" nel sito HPA, per ogni gene.

La sezione RNA summary mostra una normale distribuzione di campioni individuali nei set di dati di analisi multiple RNA-seq visualizzati con i box plot. "Solo" è usato trascrizione di un gene presente solo nel tessuto specifico (muscolo scheletrico). "Principalmente" è usato quando la maggior parte della trascrizione di un gene è presente nel tessuto specifico (muscolo scheletrico). "Tutto/i"è usato per la trascrizione di un gene presente in tutti i tessuti.

Tavola 3

GENI CAUSANTI MALATTIA

	GENE	NCBI LINK	RNA TS TPM*	PROTEIN EXPRESSION (score) **	OMIM NUMBER	TISSUE SPECIFICITY
1	**KRT10**: Homo sapiens keratin 10 (KRT10), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000421.3	18886	High	148080	Only
2	**KRT1**: Homo sapiens keratin 1 (KRT1), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_006121.3	15454,7	High	139350	Predominantly (one of three)
3	**DSG1**: Homo sapiens desmoglein 1 (DSG1), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001942.3	725,7	High	125670	Predominantly

4	**ALOXE3**: Homo sapiens arachidonate lipoxygenase 3 (ALOXE3), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001165960.1	36,5	Medium	607206	Predominantly (the highest of two)
5	**COL17A1**: Homo sapiens collagen type XVII alpha 1 chain (COL17A1), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000494.3	769,3	Medium / Low	113811	Predominantly
6	**FGFR3**: Homo sapiens fibroblast growth factor receptor 3 (FGFR3), transcript variant 3, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001163213.1	333,1	High	134934	Predominantly
7	**TYR**: Homo sapiens tyrosinase (TYR), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000372.4	29,8	High	604103	Only
8	**LOR**: Homo sapiens loricrin (LOR), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000427.2	532	Medium / Low	606933	Only
9	**HOXC13**: Homo sapiens homeobox C13 (HOXC13), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_017410.2	4,6	Only in hair	142976	Only
10	**KRT2**: Homo sapiens keratin 2 (KRT2), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000423.2	2600	High	600194	Predominantly (one of three)

GENI NON CAUSANTI MALATTIA

5

	GENE	**NCBI LINK**	**RNA TS TPM***	**PROTEIN EXPRESSION (score)****	**TISSUE SPECIFICITY**
1	**DCT**: Homo sapiens dopachrome tautomerase (DCT), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001922.4	126,8	Medium/ High	Only

2	**PMEL**: Homo sapiens premelanosome protein (PMEL), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001200054.1	133,4	Medium/ High	Only
3	**GSDMA**: Homo sapiens gasdermin A (GSDMA), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_178171.4	58,4	Medium/ High	Only
4	**KLK5**: Homo sapiens kallikrein related peptidase 5 (KLK5), transcript variant 2, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001077491.1	213,5	Medium	Only
5	**DMKN**: Homo sapiens dermokine (DMKN), transcript variant 2, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_033317.4	3991,5	High	Predominantly
6	**DSC1**: Homo sapiens desmocollin 1 (DSC1), transcript variant Dsc1a, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_024421.2	273	High	Predominantly (one of three)
7	**KRT77**: Homo sapiens keratin 77 (KRT77), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_175078.2	308,7	High	Predominantly (the highest of four)
8	**PLA2G4E**: Homo sapiens phospholipase A2 group IVE (PLA2G4E), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001206670.1	20,4	Medium/ High	Predominantly
9	**KRTDAP**: Homo sapiens keratinocyte differentiation associated protein (KRTDAP), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_207392.2	4647,8	Medium/ High	Predominantly
10	**MLANA**: Homo sapiens melan-A (MLANA), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_005511.1	50,4	High	Only

La tavola mostra la lista dei geni Homo sapiens, prioritizzati nei tessuti della pelle, usando il database Human Protein Atlas (https://www.proteinatlas.org/).

Per prioritizzare i geni della pelle, abbiamo selezionato quelli con espressioni più alte dalla lista dei geni arricchiti della pelle del database Human Protein Atlas

(https://www.proteinatlas.org/search/tissue_specificity_rna:skin;Tissue%20enriched+AND+sort_by:tissue+specific+score+AND+show_columns:groupenriched).

Tutti i dati (RNA, TS, TPM, punteggi di espressione proteicae specificità del tessuto) sono anch'esse state ottenute dal database Human Protein Atlas. *RNA TS TPM indica un livello di RNA riportato come media TPM (transcripts per million), nel tessuto di riferimento,in questo caso la pelle. **I punteggi di espressione proteica sono basati su una stima ottimale della "vera" espressione proteica da un annotazione basata sulla conoscenza nel tessuto selezionato,in questo caso la pelle. ***La specificità dei tessuti è basata su dati trovati nel grafico nominato "HPA tissue dataset", una sub-categoria della sezione "RNA sample summary" nel sito HPA, per ogni gene. La sezione RNA summary mostra la normale distribuzione di campioni individuali nei set di dati di analisi multiple RNA-seq visualizzati con i box plot. "Solo" è usato per la trascrizione di un gene presente solo nel tessuto specific (pelle). "Principalmente" è usato quando la maggior parte della trascrizione di un gene è presente nel tessuto specifico (pelle). "Tutto/i"è usato per la trascrizione di un gene presente in tutti i tessuti.

Tavola 4

GENI CAUSANTI MALATTIA

	GENE	LINK NCBI	RNA TS TPM*	ESPRESSIONE PROTEINA (PUNTEGGIO)**	NUMERO OMIM	SPECIFICITA' TESSUTALE
1	**UMOD**: Homo sapiens uromodulina (UMOD), variante trascrtto 2, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001008389.2	2392	Alta	191845	Unica
2	**SLC12A1**: Homo sapiens solute carrier famiglia 12 membro 1 (SLC12A1), variante trascritto 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000338.2	780	Alta	600839	Unica
3	**KCNJ1**: Homo sapiens sottofamiglia canale voltaggio-dipendente J membro 1 (KCNJ1), variante trascritto 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000220.4	210,8	Alta	600359	Unica

4	**SLC12A3**: Homo sapiens solute carrier famiglia 12 membro 3 (SLC12A3), variante trascritto 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000339.2	109,8	Media	600968	Unica
5	**NPHS2**: Homo sapiens, membro della famiglia della stomatina, podocina (NPHS2), variante trascritto 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_014625.3	93,6	Alta	604766	Unica
6	**BSND**: Homo sapiens barttin subunità beta accessoria tipo CLCNK (BSND), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_057176.2	10,2	Alta	606412	Prevalentemente (il più alto dei due)
7	**CLDN16**: Homo sapiens claudin 16 (CLDN16), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_006580.3	51,7	Media	603959	Unica
8	**PKD1**: Homo sapiens policistina 1, recettore transitorio che interagisce con il canale potenziale (PKD1), variante del trascritto 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001009944.2	2,3	Alta	601313	Tutte
9	**PKD2**: Homo sapiens polycystin 2, recettore transitorio del canale cationico potenziale (PKD2), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000297.3	54,2	Media/Bassa	173910	Tutte
10	**ATP6V0D2**: Homo sapiens ATPasi H+ trasportante la subunità V0 d2 (ATP6V0D2), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_152565.1	59,4	Media	618072	Unica

GENI NON CAUSANTI MALATTIA

5

	GENE	LINK NCBI	RNA TS TPM*	ESPRESSIONE PROTEINA (PUNTEGGIO)**	NUMERO OMIM
1	**BBOX1**: Homo sapiens gamma-butirrobetaina idrossilasi 1 (BBOX1), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_003986.2	473	Alta	Prevalentemente (il più alto dei due)
2	**SLC22A8**: Homo sapiens solute carrier famiglia 22 membro 8 (SLC22A8), variante trascritto 2, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001184732.1	339,8	Media	Unica
3	**MIOX**: Homo sapiens mio-inositolo ossigenasi (MIOX), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_017584.5	821,5	Media	Unica
4	**TMEM52B**: Homo sapiens proteina transmembrana 52B (TMEM52B), variante trascritto 2, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001079815.1	183,5	Media	Prevalentemente (il più alto dei due)
5	**TINAG**: Homo sapiens antigene della nefrite tubulointerstiziale (TINAG), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_014464.3	113,9	Bassa	Prevalentemente (uno dei due)
6	**CALB1**: Homo sapiens calbindina 1 (CALB1), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_004929.3	336,9	Alta	Prevalentemente (uno dei due)
7	**ATP6V1G3**: Homo sapiens ATPasi H+ trasportante la subunità V1 G3 (ATP6V1G3), variante trascritto 3, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001320218.1	21,3	Media	Unica
8	**AQP6**: Homo sapiens aquaporina 6 (AQP6), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001652.3	20,8	Alta	Unica

9	**FXYD4**: Homo sapiens dominio FXYD contenente ioni regolatori di trasporto 4 (FXYD4), variante trascritto 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_173160.2	127,4	Media	Unica
10	**GGACT**: Homo sapiens gamma-glutamilammina ciclotransferasi (GGACT), variante trascritto 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_033110.2	75,9	Bassa	Unica

La tavola mostra la lista di geni Homo sapiens, prioritizzati nel tessuto del rene, utilizzando il database Human Protein Atlas (https://www.proteinatlas.org/). Per prioritizzare I geni del rene abbiamo selezionato quelli con espressione più alta dalla lista di geni arricchiti del rene del database Human Protein Atlas (https://www.proteinatlas.org/search/tissue_specificity_rna:kidney;Tissue%893 20enriched+AND+sort_by:tissue+specific+score+AND+show_columns:groupenriched).

895 Tutti i dati (RNA, TS, TPM, punteggi di espressione proteica e specificità del tessuto) sono anch'essi stati ottenuti dal database Human Protein Atlas. *RNA TS TPM indica un livello di RNA riportato come media TPM (transcripts per million), nel tessuto di riferimento,in questo caso il rene. **I punteggi di espressione proteica sono basati su una stima ottimale della "vera" espressione proteica da un'annotazione basata sulla conoscenza nel tessuto selezionato,in questo caso il rene. ***La specificità del tessuto è basata su dati trovati nel grafico nominato "HPA tissue dataset", una sub-categoria della sezione "RNA sample summary" nei sito HPA,per ogni gene.

La sezione RNA summary mostra la normale distribuzione di campioni individuali nei set di dati di analisi multiple RNA-seq visualizzati con i box plot. "Solo" è usato per la trascrizione di un gene presente solo nel tessuto specifico (rene). "Principalmente" è usato quando la maggior parte della trascrizione di un gene è presente nel tessuto specifico (rene). "Tutto/i"è usato per la trascrizione di un gene presente in tutti i tessuti.

I mammiferi sono stati selezionati basandosi sul numero più alto di geni annotati.

Abbiamo selezionato geni con differenti lunghezze come l'mRNAs e, per massimizzare il numero di codoni analizzati ed evitare pregiudizio contro le sequenze brevi o parziali, sono state selezionate solo sequenze codificate di piena lunghezza.se è stata annotate più di una isoforma di splicing, sono state selezionate le isoforme più

lunghe, dal moment oche recenti studi sulla singola cellula di RNAseq hanno dimostrato che la lunghezza dell'mRNA non influenza il livenno di espressione nel tessuto isoforme.

Selezione dei geni

I geni DC sono stati prioritizzati basandosi sul loro coinvolgimento nelle malattie Mendeliane rare, con incidenza minore di 1:5000 (secondo il catalogo OMIM, www.omim.org), e con un fenotipo omogeneo e legato a tessuti/organi specifici (renale, muscolo scheletrico e pelle). Criteri di esclusione per i geni sono stati il coinvolgimento im malattie poligeniche o cancro (entrambi Mendelianie somatici) e l'assenza di specificità di tessuto (geni di manutenzione). Abbiamo poi selezionato i geni DC basandoci sulla malattia causata e la tessuto-specificità e I genoi NDC basandoci sulla loro maggiore espressione negli stessi tessuti nei quali erano espresso i geni DC (muscolo, pelle e rene).

I criteri di selezione dei geni (Tavole 2, 3 e 4)erano basati su: i) geni pienamente annotati riconosciuti come causanti malattie Mendeliane rare (geni DC), ii) geni pienamente annotati non causanti malattie Mendeliane rare (geni NDC).

Il database dle catalogo OMIM è stato usato per categorizzare i geni come DC o NDC. I geni DC dovevano essere associate con una malattia Mendeliana in uno dei tre tessuti selezionati (rene, muscolo scheletrico e pelle)in almeno 5 famiglie/pazienti riportati, quindi essere confermate nel database OMIM, e con pattern ereditati definiti (autosomico recessivo, autosomico dominante o X-linked recessivo). Evidenza del coinvolgimento del gene in cancri Mendeliani e somatici, come la suscettibilità dei geni, sono stati considerati criteri di esclusione per la selezione dei geni, dal momento che i geni del cancro Mendeliano sono spesso anche geni predisponendi al cancro, e pertento possono rappresentare fattori di confusione nel nostro studio, che si concentra solo sulle malattie Mendeliane.

I tipi di malattie rare sono legate alle mutazioni dei geni dei tessuti del muscolo, della pelle e del rene. Malattie rare del muscolo includono distrofie muscolari e miopatie congenite, i cui geni causative sono principalmente e maggiormente espresso nel muscolo scheletrico come distrofina, disferlina, onella matrice extracellulare del muscolo, come il gene del collagene 6A1. Malattie rare del rene includono gene dell'uromodulina e della policistina 1, le cui mutazioni causano Malattia renale tubulointerstiziale o rene policistico di tipo 1 rispettivamente, gene altamente espresso in questi due differenti compartimenti del rene. Infine, malattie rare della pelle includono geni della cheratina 10 e HOXC13, le cui mutazioni sono associate con l'ipercheratosi

epidermolitica e la displasia ectodermica 9, due malattie differenti in termini di fenotipo e coinvolgimento dello strato cutaneo. Le tavole 2, 3, e 4, riportano l'intera lista di geni RD con tutti i numeri OMIM corrispondenti.

Due geni , TMEM52B and PLA2G4E, non elencati del database OMIM dal moment oche non sono mai stati associate con alcuna delle malattie umane, sono stati controllati usando i database PubMed, ClinVar e DMDM e quindi esclusi per essere causativi di malattie Mendeliane.

Abbiamo selezionato i tessuti basandoci sul loro alto arricchimento genico nel database Human Protein Atlas (HPA). la nostra prioritizzazione era basata su metriche HPA usate pe ril livello di RNA (Transcripts Per Million, TPM), il punteggio di espressione proteica (alto, medio, basso, nullo)e valori di tessuto-specificità. Questi punteggi ci hanno permesso di classificarei profili di espressione di geni NDC e DC secondo la loro specificità di tessuto. Il più alto valore di espressione implica almeno livelli di mRNA quattro volte più alti nei tessuti selezionati comparata a ogni altro tessuto, mentre i punteggi proteici erano alti o medi, bassi livelli di espressione sono stati esclusi.

Database delle mutazioni

Abbiamo esaminato i database pubblici OMIM, ExAC and ClinVar insieme con i nostri database interni UNIFE per variazioni genetiche a singolo nucleotide DMD. Sono state considerate unicamente le variazioni patogeniche missenso e nonsenso, dal momento che il loro significato e la loro identificazione nei database non sono equivoci. Cambiamenti sinonimi nel gene DMD non sono considerati in questo studio dal moment oche sono quasi invariabilemnte definiti come variant dal significato incerto (VUS) ovarianti benigne, decondo le line guida dell'ACMGG.

Calcolo dei valori CU e analisi statistica

La frequenza di utilizzo del codone (CU) è stata calcolata indipendentemente per ognuno dei 19 anminoacidi considerati in ogni gruppo di geni. E' stato valutato anche l'uso dei tre codoni di stop. Metionina (AUG) and triptofano (UGG) non sono stati inclusi dal momento che sono codificati fa un'unica tripletta. Tutte le analisi statistiche sono state svolte con R-3.4.4 (R version 3.4.4. R. https://www.r-project.org/ (2018)).

La significanza statistica è stata definite come valore $P < 0.05$. Test statistici basilari e generazione di diagrammi a barre e diagrammi a scatola sono state svolte utilizzando funzioni built-in incluse con la distribuzione base di R o funzioni nel pacchetto ggplot2 (Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. (2016); 978-

3-319-24277-4).

Per un paragone di utilizzo del codone tra geni DC e NDC, tipi di tessuti e specie, un sum test Wilcox on rank è stato applicato per calcolare valori P a due code usando la funzione 'Wilcox test' in R (Bauer DF. Constructing confidence sets using rank statistics. Journal of the American Statistical Association. (1972); s*67*, 687-690), e questi sono stati visualizzati utilizzando il pacchetto R ggplot2.

La frequenza di CU in geni DC eNDC è stata anche comparata per identificare i codoni usati più differentemente nei tessuti genici, tra le specie e nei geni DC versus NDC, e i dati sono stati visualizzati indiagrammi a scatola.

Il coefficiente di correlazione Spearman di frequenza di utilizzo dei codoni nei geni DC e NDC del muscolo, della pelle e del rene nell'HSA è stato usato e visualizzato utilizzando il pacchetto R ggplot2.

La funzione heatmap.2 nel pacchetto R gplots è stata usata per il raggruppamento dei codoni e la loro visualizzazione.

Nel raggruppamento, poiché gli usi dei codoni sono dati di intervallo e non sono influenzati da outlier con valori estremamente larghi, la distanza metrica euclidea è stata selezionata per una facile implementazione e semplice interpretazione. Il raggruppamento agglomerativo gerarchico è stato svolto utilizzando il metodo standard "completo" in funzione hclust.

I geni sono sempre elencati in grafici in ordine decrescente secondo il numero 224 di esoni, e il numero di esoni è stato calcolato basandosi su annotazioni di dati scaricate dal database Ensembl Genome. I dati sono stati analizzati usando l'algoritmo di raggruppamento gerarchico o applicando un'analisi prioritaria in termini di percentuali CUB.

I codoni sinonimi che non sono affatto utilizzati dai geni, pertanto avendo un bias di codone estremo, sono stati chiamati "codoni-zero".

RISULTATI

Impronta CUB dei geni tessuto-specifici dell'Homo sapiens (HSA)

Abbiamo inizialmente verificato il raggruppamento gerarchico di uso dei codoni sinonimi in tutti i geni studiati con diversa specificità di tessuto (muscolo, pelle e rene) nell'HSA. Abbiamo osservato un raggruppamento di codoni tessuto-specifici, che abbiamo definite "imprinta CUB". I grafici dei diagrammi di calore (Figura 1) mostrano

che raggruppamenti di codoni frequentemente utilizzati (in rosso) e raramente utilizzati (blu scuro) variano grandemente tra i tessuti umani.

Nei geni del muscolo, codoni con un bias estremo (valori CU bassi, colore blu scuro, o alti valori CU, colore chiave rosso scuro) sono strettamente raggruppati in termini sia di gene che di tipo di codone, mentre valori CU intermedi (colore chiave azzurro o giallo) sono disperse negli alberi (Figura 1A).

L'impronta CUB dei geni della pelle (Figura 1B) è caratterizzata da una predominanza di valori CU bassi (punti gialli) con pochi punti rossi disperse e un chiaro raggruppamento di codoni più rari (punti blu scuro), mentre i pochi codoni con valori CU intermedi o alti sono distribuiti in maniera disomogenea tra i geni con grandi distanze nell'albero. I raggruppamenti di codoni generate sono anche meno definiti se paragonati al muscolo e al rene.

Nei geni del rene, l'impronta CUB differisce dagli altri due gruppi di geni e la gerarchia dei codoni è più definite (solo due grandi linee, Figura 1C). I codoni con valori intermedi sono raggrippati con tendenze legate al gene, come visible nei geni PKD2, KCNJ1 e MIOX. La stragrande maggioranza dei geni hanno valori CU bassi o intermedi con ampi e diffusi raggruppamenti di punti blu. Un piccolo gruppo di geni (UMOD, BSND, SLC22A8, MIOX, AQP6, PKD1, SLC12A3, and GGACT) mostrano alti valori CU dei codoni UGA, UAC, UUC, AUC, GAC, AAC, AAG, GAG, UGC, CAC, e CAG, appartenenti a linee ben definite nel raggruppamento (impronta CUB gene-codone specifica).

Guardando al raggruppamento dei codoni, i codoni più utilizzati, CAG, AAG, UGC, GAG, UAC, UUC, CAC, AAC, AUC e GAC, hanno raggruppamento identico e si sovrappongono nedi geni del muscolo e del rene ma non nei geni della pelle (Figura 1A e 1C, lato sinistro). Tra i geni dei muscoli, solo DMD non mostra un CUB estremo, dal moment oche non si verificano punti rossi (Figura 1A), con l'unica ovvia eccezione dell'unico codone di stop UAG.

Guardando al raggruppamento dei geni, i dendrogrammi dei geni del muscolo e del rene mostrano impronte riconoscibili (gerarchia vertical nelle mappe di calore nella Figura 1A e 1C), dal momento che i geni con valori CU alti o bassi (arricchiti con punti rossi o blu) sono chiaramente raggruppati (13 geni nella Figura 1A, e 8 geni nella Figura C, lato sinistro), mentre nei geni della pelle questo non si verifica. Pertanto, some annotato sopra, sono stati osservati alcuni valori CU gene-specifici, dipendentemente dal tessuto studiato. I valori di utilizzo dei codoni di stop sono stati calcolati nei tre gruppi di geni. Nei geni della pelle e del muscolo, tutti i codoni di stop sono usati uniformemente,

con UAApresente in maniera più frequente.Nei geni del rene, UAG è usato molto raramente, mentre UAA e UGA sono ugualmente rappresentati.

Impronte CUB tra i mammiferi

Abbiamo analizzato valori CU in tutti i 20 geni tessuto-specifici tra 15 specie di mammiferi dell'albero filogenetico dei metazoi (Figura 2A, B, e C) o i valori CU in tutti i mammiferi e tra i tre tessuti genici (Figura 2D, E, e F). I valori CU tra i mammiferi mostrano evidenti impronte CUB tessuto-specifiche dovute a differenti usi del tipo di codone (Figura 2A, B, C).

CAG, AAG, CAC, GAC, GAG, AUC, AAC, UAC, UGC e UUC sono i codoni usati più frequentemente (punti rossi) in tutti i tessuti genici e tra i mammiferi, mentre UUA, CUA, UCG, CGU, CUU, GUA, CGA, AUA, UCA, UUG e GCG sono i codoni più rari (punti blu) nei geni del muscolo e della pellema non dei reni. I CUB del muscolo e della pelle hanno un comportamento simile del raggruppamento dei valori CU, sebbene con valori CU più bassi nella pelle (più punti gialli), simile ai geni del rene che mostrano un differenteraggruppamento gerarchico.

Guardando al raggruppamento dei geni in tutti i mammiferi (Figura 2D, E, and F), i raggruppamenti di valori CU non possono essere visti. Tutti i tessuti hanno impronte CUB differenti, con anche un differente raggruppamento gerarchico dei codoni.

I dendrogrammi legati ai geni sono anche diversi dal momento che i geni del muscolo (13/20), del rene (8/20) e della pelle (4/20) si raggruppano insieme con diversi valori CU legati al diverso tipo di codone. Infatti considerando tutti i geni in tutte le specie i tipi di codone variano. pertanto, sebbene alcune impronte CUB legate ai tessuti sono ancora riconoscibili nei mammiferi, non può essere osservato nessun chiaro comportamento tissutale o raggruppamento legato al gene specifico. Questa è una dimostrazione che i valori CU mostrano una correlazione con i mammiferi ma con importanti differenze gene-specifiche, che cointribuiscono a generare le impronte CUB.

Valori CU e impronte CUB nai geni DC e NDC nell'HSA

Abbiamo raggruppato i geni basandoci sulla loro propensione ad essre sito di variazioni patogeniche (mutazioni) causando malattie rare (geni DC o NDC). Abbiamo profilato i valori di CU nei geni DC e NDC, preservando la distinzione del tessuto (geni del muscolo, della pelle e del rene) tra i mammiferi. La figura 3,pannelli dalla A alla F, mostra i valori CU assoluti nei geni DC e NDC del muscolo, della pelle e del rene, rispettivamente, con Nessun raggruppamento gerarchico e identico ordine di tipo di codone. I valori CU in questi 6 pannelli mostrano che il tipo di codoni più frequenti o più

rari sono molto simili in tutti i geni e tra i mammiferi. Questo è supportato nell'HSA dall'analisi della correlazione di Spearman, che ha dimostrato che i valori CU dei geni DC e NDC sono significativamente correlati nei gruppii del muscolo, della pelle e del rene ($p<0.05$) (Supplementary Figura 1).

Più variabilità di valori CU può essere vista nei codoni con frequenza intermedia, dove una tendenza di gene o di tessuto può essre vista. In particolare, i geni NDC del muscolo hanno valori CU più alti (è visible qualche punto giallo, vedi Figura 3 A, B), mentre i geni DC e NDC del muscolo edel rene (Figura 3, pannelli A, B, E, F) hanno valori CU molto simili. Da notare, CAG è il codone usato più frequentemente e UUA quello usato meno di frequente in tutti i mammiferi.

La figura 3, pannelli da G ad L, mostrano un raggruppamento gerarchico di valori CU nelle stesse categorie di geni di cui sopra. Guardando ai geni DC e NDC, possono essere osservate impronte CUB riconoscibili. I geni del muscolo DC e NDC (Figura 3, pannnelli G, H) mostrano differenti impronte e raggruppamenti. I geni DC del muscolo hanno un raggruppamento compatto di codoni frequentemente utilizzati (AAG, CAG, GAG) o di codoni estremamente rari (UGG, UUA, CUA). I valori CU sono omogenei tra i mammiferi con gruppi di codoni charamente definiti in termini di distanza dall'albero. Nei geni NDC del muscolo, l'impronta CUB cambia. Dominano punti rossi e blu scuro, con pochi codoni con valori CU intermedi (punti gialli). Quest'impronta indica che un CUB più forte si è verificato nei geni NDC del muscolo. Di conseguenza, il dendrogramma basato sul codone nei geni NDC, ma non nei geni DC, mostra che valori CU più alti e più bassi sono raggruppati insieme, sottolineando un possible comportamento diverso tra i mammiferi.

I geni della pelle DC e NDC (Figura 3I-J) mostrano simili impronte CUB con differenze minime. I geni della pelle NDC mostrano più codoni con bassi valori CU (figura superiore 3J) e viceversa meno codoni con valori CU alti o intermedi (figura inferiore 3J), paragonati ai geni DC. Questo implica che Valori CU intermedi si verificano più frequentemente, una scoperta opposta a quella vista nei geni dei muscoli.

Infatti, l'impronta dei geni DC della pelle è simile a quella vista nei geni dei muscoli DC. Possiamo concludere che i geni DC sia del muscolo che della pelle mostrano un tipico modello "no CUB estremo".

Le impronte dei geni DC e NDC del rene (Figura 3K-L) e i dendrogrammi dei codoni differiscono grandemente dagli altri due tessuti, dal momento che le distanze gerarchiche tra raggruppamenti di valore sono opposti. Sebbene la conservazione di valori CU tra i mammiferi si verifichi, la gerarchia del dendrogramma dei geni del rene

mostra un antenato commune per valori CU intermedi e bassi, al contrario dei geni del muscolo e della pelle, dove due line distinte (valori CU alti e bassi) sono visibili, come già osservato prima (Figura 1C).

Le impronte CUB dei geni DC e NDC sono simili, sebbene debba essere notato che i geni NDC mostrano un più alto numero di valori CU bassi (punti gialli).

Abbiamo anche comparato i valori CU tra geni dell'HAS alto-, medio-, e basso-espressi. I geni DC e NDC sono stati quindi divisi in tre categorie dipendenti dal livello di espressione del gene (per la piega dell'RNA e i valori di cut-off delle proteine. Le impronte Cub hanno grande similarità, significand oche infatti, raggruppare i geni secondo il loro livello di espressione porta a una tendenza simile di valori CU. Abbiamo anche analizzato se potessero esserci codoni usati di più secondo il livello di espressione del gene. I codoni AAC, GAC, UGC, UAC, CAC, UUC, AUC, AAG, GAG e CAG sono più frequentemente usati sia nei geni alto- che medio-espressi, mentre GUG è presente solo in geni altamente espressi. I codoni GAC, CAC, UUC, CAG, UAC, UGC, AAG, AUC, GAG, AAC, ACC, GGC, GUC, UCC e GCG sono più frequentemente usati in geni con un basso livello di espressione. In particolare, mentre la maggioranza dei codoni frequentemente usati sono similmente rappresentati in geni ad alta, media e bassa espressione, pertanto per tanto con un CU non influenzato dal livello di espressione, solo pochi codoni con valori CU più bassi come UCC (Ser), ACC (Thr), GGC (Gly), GUC (Val) e GCG (Ala) sono presenti in geni basso-espressi. Similarmente, codoni con valori CU bassi sono sono gli stessi nei geni altamente espresso contro i geni mediamente espressi. Alcuni geni DV altamente espressi hanno più codoni con valori CU alti, come DYS, LMNA e DES (nel muscolo scheletrico), UMOD e PKD1 (nel rene) e FGFR3 (nella pelle). La tendenza è opposta nei geni mediamente espressi, dove alcuni geni NDC, come MLPF, TNNC2, TMEM3BA (nel muscolo scheletrico), e NCLZ2, MCX (nel rene), mostrano valori CU più alti.

È interessante notare che, UAA, che è stato riconosciuto come induttore della terminazione della traduzione con più velocità e accuratezza a livello ribosomiale e può essere letto sia da fattori di rilascio eRF1 che eRF2, è il codone di stop più utilizzato nei geni altamente espressi. UAG e UGA hanno una frequenza simile in tutti i geni.

<u>Prioritizzazione dei 343 codoni usati più differentemente tra i geni tessuto-specifici</u>

Abbiamo applicato la stessa strategia di utilizzare il calcolo dei valori CU per paragonare diversi geni umani per identificare evntuali codoni che possono essere più

differentemente/preferibilmente utilizzati in alcuni geni causanti malattia, comparati a geni non causanti malattia.

Basandoci sul valore di uso del codone p ($P < 0,05$), abbiamo prioritizzato codoni con valori CU differenti tra i tessuti genici. I risultati sono mostrati nella Figura 4. Cinque codoni sono stati prioritizzati, essendo usati in maniera significativamente diversa nei tre tessuti umani: CGU (Arg), CCA (Pro), GAC (Asp), GAU (Asp) and GUA (Val) (Figura 4 A-E). CCA e CGU sono i codoni meno frequentement eusati nel muscolo, GUA nel rene, e GAU e GAC sono I codoni più frequentemente usati nella pelle.

Considerando i geni DC e NDC, ulteriori differenze nei valori CU possono essere osservate per CGU e CCA, che sono usati più frequentemente nei geni DC, e per GUA e GAU, che, sebbene meno significativamente, sono usati più frequentemente nei geni NDC (Figura 4 F-L). Questi dati suggeriscono che i valori CU possono essere influenzati dalla propensione del fenotipo e del gene a causare malattie genetiche. La tendenza del valore CU di questi 5 codoni sembra essersi conservata tra i mammiferi (Figura 4F- L).

Infine, abbiamo contato il numero di codoni con bias estremo nei geni e tra i mammiferi. I geni DC mantengono un uso più elevato di codoni multipli, con solo pochi codoni con bias estremo, paragonati ai geni NDC (Tavola 5, pannelli A-B).

Tavola 5

Panel A

NDC															
	Rhinolophus_ferrumequinum	Mus_musculus	Felis_catus	Canis_lupus	Equus_caballus	Bos_taurus	Microcebus_murinus	Galeopterus_variegatus	Callithrix_jacchus	Macaca_mulatta	Nomascus_leucogenys	Pongo_abelii	Gorilla_gorilla	Pan_troglodytes	homo_sapiens
kidney	85	93	97	98	111	108	100	83	105	109	116	120	118	117	114
muscle	85	88	91	97	105	99	100	88	89	96	97	107	97	98	104
skin	62	62	54	62	61	75	64	50	57	75	65	64	59	62	63

DC															
	Rhinolophus_ferrumequinum	Mus_musculus	Felis_catus	Canis_lupus	Equus_caballus	Bos_taurus	Microcebus_murinus	Galeopterus_variegatus	Callithrix_jacchus	Macaca_mulatta	Nomascus_leucogenys	Pongo_abelii	Gorilla_gorilla	Pan_troglodytes	homo_sapiens
kidney	39	35	33	28	42	38	41	45	38	38	38	32	31	31	31
muscle	46	35	46	53	49	47	52	46	46	45	44	51	44	45	45
skin	37	66	73	64	67	37	74	52	65	74	56	74	64	57	78

Inglese	Italiano
Kidney	Rene
Muscle	Muscolo
Skin	Pelle

Tavola 5. Numero di codoni con bias estremo (“codoni-zero”) trovati in geni DC e NDC tessuto specifici tra i mammiferi

Tavola 5 (pannello A) mostra il numero di codoni con bias estremo (“codoni-zero”) nei tessuti genici e tra i mammiferi. I geni sono raggruppati in DC e NDC. Il numero di “codoni-zero” è basato sul valore di uso dle codone (CU) che abbiamo calcolato in ogni gruppo di geni e tra i mammiferi. I geni DC mantengono un uso multiplo del codone (CUB basso), pertanto con pochi “codoni-zero” paragonati ai geni NDC. Da notare che, il numero di “codoni-zero” è molto più alto nei geni NDC del muscolo e del rene e nei geni DC della pelle, quindi con un comportamento guidato alla malattia. Il CUB è cresciuto progressivamente durante l’evoluzione, sebbene non allo stesso livello in tutti i gruppi di geni. La differenza più marcata nel numero del “codone-zero” è tra HSA rene DC (11) e NDC (94) geni e muscolo DC (49) e geni NDC (109). I pannelli B e C mostrano gli stessi dati riportati sui grafici per apprezzare meglio la tendenza del CUB e il numero di “codoni-zero”.

Comportamento unico del CUB DMD

Abbiamo usato i valori CU calcolati e abbiamo mappato 2828 mutazioni patogeniche missenso e nonsense conosciute nel gene distrofina (DMD), prese da database pubblici (LOVD) o database interni, sui codoni DMD. Abbiamo chiamato il nostro approccio alle mutazioni “map-on-codon”.

Abbiamo verificato se codoni DMD raramente/frequentemente usati sono conseguentemente di rado o di frequente sito di variazioni patogeniche missenso e nonsense provate. Questo supporta che alcuni codoni sono più o meno pronti ad essere il sito di variazioni patogeniche in un contest di gene specifico (DMD, in questo caso), e i “codoni più-meno mutati” sono rilevanti per la capacità di traduzione del gene (DMD), e/o per alter funzioni legate alla traduzione, significando che esse devono essere considerate quando si svolge un’ottimizzazione del codone del gene artificiale.

Il gene DMD non ha CUB estremo e mantiene tutti i tipi di codone usati nella sua sequenza codificata, anche tra i mammiferi studiati. Abbiamo contato il numero dei tipi di codoni nella sequenza codificata DMD e identificato solo 4 codoni con bias estremo, UCG, CCG, ACG e GCG, conseguentemente il parametro noto di bias cut-off, basati sulla ridondanza dei codoni di 2, 3, 4 e 6 triplette (Weissbach H. Syntax of referencing in Molecular Mechanisms of Protein Biosynthesis (cap. Lipmann, F. Twenty Years of Molecular Biology)(ed. Nutley, New Jersey : Elsevier). (2012); 736: 3-5. LOVD. https://databases.lovd.nl/shared/genes/DMD/ (2019)), (Figura 5A).

La figura 5B mostra il numero di mutazioni verificatesi a tutti i tipi di codone. Ogni numero, in cima alle barre, rappresenta quante volte quello specific codone è stato il sito

di una variazione DMD missenso o nonsenso. E' interessante notare che, questi numeri non sono legati a valori CU. Sebbene CCG (Pro), UCG (Ser), GCG (Ala) e ACG (Thr) in rosso nelle figure 5A e B, sono, come aspettato, meno di frequente sito di mutazioni, essendo i codoni più rari usati nel gene DMD, i valori CU degli altri codoni non sono correlati con il verificarsi delle mutazioni. Questo è il caso di CAG (Glu), che mostra valori CU intermedi ma è il codone più frequentemente sito di mutazioni DMD, e UAU (Tyr) e UUU (Phe), che sono i codoni DMD più frequentemente utilizzati (Figura 5A) ma sono raramente sito di mutazioni (Figura 5B).

DISCUSSIONE

Abbiamo calcolato i valori CU in tre piccolo gruppi di geni che sono tessutospecifici e maggiormente espresso nella pelle, nei reni e nel muscolo scheletrico. Abbiamo poi innovativamente comparato i valori CUtra geni DC e NDC, e tra i mammiferi, pertanto utilizzando un approccio guidato alla malattia, per esplorare i valori di CU e il comportamento CUB.

Abbiamo confermato che la tessuto specificità influenza il CUB, e abbiamo osservato una tendenza tessuto-specifica, guardando il CUB in geni alto-, medio-, e basso-espressi. Eì interessante notare che, alcuni codoni sono più rappresentati in geni alto- o basso-espressi, un fatto possibilmente legato auna selezione positiva di codoni attività traslazionale più alta o più bassa durante l'evoluzione, dipendentemente dal ruolo del gene in tessuti specifici e/o organi specifici. Di conseguenza, il codone di stop UAA è preferenzialmente usato nei geni del muscolo e, generalmente, nei geni altamente espressi, probabilmente riflettendo il bisogno di un'ottimale ricognizione ribosomiale dle codone per fermare efficientemente la traduzione. Abbiamo quindi confermato che, tramite anche il calcolo dei valori CU e comparandoli in piccolo gruppi di geni umani selezionati, altamente tessutospecifici, le differenze di valori CU possono essere osservate, dipendentemente dal tessuto e dal livello di espressione del gene.

<u>Impronte tessuto-speifiche nei geni di HSA</u>

Comparando i valori CU in raggruppamenti gerarchici nei geni del muscolo, della pelle e del rene nell'HSA, possono essere osservati diversi pattern. Tipi di codone raramente o frequentemente usati variano tra i tre tessuti, essendo i geni del muscolo e della pelle più simili in termini di raggruppamenti e gerarchia dei valori CU. Queste differenti impronte CUB possono essere dovute alla tessuto specificità dei geni analizzati, e noi supponiamo che i geni aventi qualche funzione tissutale che può richiedere un'efficienza di traduzione più alta o più bassa, usano diversi codoni sinonimi. E' interessante notare

che, molti geni del muscolo e del rene condividono gli stessi codoni raggruppati, frequentemente usati, supportando che questi possano essere i codoni chiave per regolare la traduzione tessuto-specifica. Da notare che, muscolo e rene, insieme con fegato e polmone, sono tessuti parenchimali, che subiscono una simile scarsa capacità di rigenerazione degli organi, come conseguenza dei compromessi evolutivi, che è specialmente legata agli effetti di bilanciamento tra il Sistema immune e la forma di processi fisiologici e patologici. Inoltre, un simile raggruppamento di alti valori CU è ancora più evidente per qualche gene del muscolo (COL6A1, RPL3L, MYLPF, TMEM38A, TNNC2, LMNA, DES, LBX1, SGCA, ANKRD23, CAPN3, DYSF, ACTN3) e del rene(UMOD, BSND, SLC22A8, MIOX, AQP6, PKD1, SLC12A3, GGACT), sottolieando qualche funzione gene-specifica e/o legata all'organo.

E' interessante notare che questi geni sono tutti geni DC, fatto consistente con i nostri risultati di paragone tra geni DC e NDC (vedi sotto). Di conseguenza, molti geni della pelle mostrano valori CU bassi o intermedi, con pochi valori CU alti mai raggruppati. la pelle è riconosciuta come un "microambiente immunologico" che regola la rigenerazione della cellula, da ciò la sua tendenza CUB opposta, paragonata ai geni del muscolo e del rene, può riflettere la sua diversa funzione organica e di sviluppo.

Impronte CUB tra i mammiferi

Le differenze di impronte CUB che abbiamo osservato nei hgeni del muscolo, dle rene e della pelle sono osservabili anche tra i mammiferi. La conservazione evolutive del CUB è stata ampiamente studiata nell'HSA, ma non in una tale granularità genetica. I geni del muscolo e della pelle, ma non del rene, mostrano simili pattern CUB con due maggiori linee gerarchiche, una per i codoni più frequenti e una per quelli più rari. Queste tendenze suggeriscono che il CUB nei geni del muscolo e della pelle possono aver possibilmente seguito qualche percorso evoluzionistico comune. I geni del muscolo sono estremamente important nei mammiferi, dove contribuiscono per circa l'80% della massa corporea. Nell'HSA, L'acquisizione di bipedalismo ha certamente richiesto una robusta forza selettiva per guidare il rimodellamento muscolare, specialmente per i muscoli legati alle giunture degli arti. Può essere trovata un origine commune tra la pelle e il muscolo striato nel panniculus carnosus, un sottile strato muscolare striato attaccato alla pelle e alla fascia della maggior parte dei mammiferi, che fornisce support per le funzioni di pulsazione e contrazione del muscolo. Il panniculus carnosus è ancora conservato negli umani, sebbene non sia considerato avere alcun significato funzionale, ed è un residuo dell'evoluzione, riflettendo la commune origine del muscolo e della pelle. A supportare questo legame, una coorte di rare sindromi muscolo-cutanee dovute a mutazioni nel

percorso dei geni RAS/MAPK sono state descritte nell'uomo e viste le loro impronte CUB silimi, meriterebbero di essere studiate con la nostra strategia.

Impronte CUB nei geni causanti malattia

Comparando i valori CU nei geni causanti malattie rare, abbiamo mostrato che, sebbene i percorsi tessuto-specifici sono ancora riconoscibili, alcuni geni DC (specialmente nelmuscolo) hanno impronte CUB differenti paragonate ai geni NDC.

In generale, i geni DC mostrano valori CU meno estremi paragonati ai geni NDC, fatto molto evidente se abbiamo calcolato i valori CU senza raggruppamento gerarchico. I geni DC del muscolo e, meno evidentemente, della pelle e del rene mostrano valori più intermedi comparati ai geni NDC, suggerendo un diverso comportamento CUB nei geni causanti malattia.

Soprattutto, i geni DC del muscolo mostrano un'impronta legata alla malattia più riconoscibile, suggerendo che un CUB "orientato alla malattia" può parzialmente prevalere nelle scelte sulla tessuto-specificità CUB nel muscolo.

I geni DC e NDC della pelle e del rene mostrano differenze meno evidenti. Nondimeno, alcuni codoni, come AAG, CAG e GAG, hanno valori CU più alti nei geni DC, ipotizzando la possibilità che la loro frequenza sia orientate alla malattia.

E' noto che il CUB aumenta durante l'evoluzione e diventa estremo con la completa mancanza di rappresentazione di qualche codone, codoni che abbiamo definite "codoni zero", sebbene le ragioni di questo non siano pienamente comprese. Lw impronte CU dei geni DC del muscolo orientate alla malattia possono suggerire una diversa pressione di selezione naturale, come già identificato in alcune categorie del gene umano. Infatti, un CUB gene-specifico è stato identificato in alcuni geni di malattie umane, suggerendo un impatto sull'interpretazione di variazioni sinonime. nei geni *CFTR* e *GATA4* (le cui mutazioni causano fibrosi cistica e una malattia cardiaca congenita, rispettivamente), mutazioni sinonime possono alterare le cinetiche di traduzione e il ripiegamento delle proteine introducendo codoni rari o non-ottimali; alti valori di ACT, AGG, ATT e AGC, o AGA CUB sono stati visti in *HPRT1* (le cui mutazioni si verificano nella syndrome Lesch-Nyhn) e *GALC* (le cui mutazioni causano la malattia di Krabbe); infine i geni *BRCA1* e *BRAC2* (geni maggiormente coinvolti nel cancro mendeliano della mammella) mostrano un CUB estremamente basso paragonati ad alti oncogeni. Mettendo insieme questi e i nostri dati, possiamo ipotizzare che alcuni geni DC possono aver attraversato una differente pressione CUB durante l'evoluzione. A support di questa evidenza, è risaputo che alcuni geni di malattia hanno ancora un alto tasso di variazione, come il gene *DMD*,

che possono avere impatto sulla frequenza del tipo di codone. Da notare che , è stato mostrato molto recentemente che l'evoluzione del gene/della proteina può verificarsi non solo in termini di forze orientate all'evoluzione ma anche in relazione con le malattia che le loro mutazioni possono causare.

Pertanto, nei geni causanti malattia da noi studiati forniamo evidenza preliminare che il CUB può essere orientate alla malattia, suggerendo che comparare valori CU in questi geni, che sono più di 6000 già identificati finora negli umani, può essere una buona strategia per capire le regole che governano l'utilizzo dei codoni.

Prioritizzazione dei codoni e comportamento CUB del gene DMD

Cinque codoni, GUA (Val), GAU (Asp), GAC (Asp), CCA (Pro) e CGU (Arg) hanno mostrato i valori CU più differenti tra i geni e i tessuti dell'*HSA*.

GAC e GAU sono i codoni più utilizzati nei geni della pelle, contrariamente CAA e CGU e GUA sono i codoni meno usati nei geni del muscolo e del rene, rispettivamente. E' interessante notare che, questi ultimi tre codoni, insieme con il codone ATC (Ile), hanno una significative correlazione positive con l'espressione dei geni dovuto al loro alto contenuto di GC. Questa scoperta supporta l'uso eccessivo di alcuni codoni orientato ai tessuti, simile a quell oche abbiamo trovato per questi 5 codoni nel nostro studio, che può riflettere uno specific processo di selezione nell'HSA.

Supportando la nostra ipotesi sulle imponte CUB dei geni DC, l'utilizzo tessuto-specifico di questi 5 codoni è anch'esso orientate alla malattia. CCA e CGU sono i codoni più utilizzati nel muscolo e hanno valori CU più alti nei geni DC. Similarmente, GAC è un codone molto utilizzato nella pelle ed è più frequentemente usato nei geni DC, mentre il codono più utilizzato nella pelle, GAU, è più frequentemente usato nei geni NDC. Questa tendenza orientate alla malattia sembra essersi conservata trai mammiferi.

Alcuni report disponibili suggeriscono che una sovrarappresentazione di codoni specifici può caratterizzare i geni del disordine mendeliano e, secondo la teoria della pressione di mutazione direzionale, una pressione selettiva negative ai codoni conferendo un più altro rischio del verificarsi di mutazioni. Infine, è risaputo che alcuni geni dell'HSA mostrano un CUB più forte o più basso in base alla loro funzione. Basandoci su questa conoscenza, possiamo ipotizzare che alcuni geni DC con funzioni uniche possono aver avuto una transizione evoluzionaria differente in termini di CUB, per alcune/tutte le ragioni sopra menzionate.

Il gene DMD è l'unico gene dell'HSA che non presenta alcun "codone zero" eppure usa tutti i codoni della sua sequenza codificata, che è una tendenza conservata tra i

mammiferi. Le mutazioni DMD, incluse variazioni missenso e nonsenso, causano la distrofia muscolare di Duchenne (DMD, OMIM* 300377) una rara, grave e fatale distrofia muscolare, o la forma più mite, la distrofia muscolare di Becker (BMD; OMIM *300376), entrambe ereditate come malattie recessive legate al cromosoma X, con un incidenza di 1:5000 maschi neonati.

Tramite il nostro approccio “mapping-on-codons” delle 2828 variazioni DMD patogeniche missenso e nonsense causanti DMD o BMD, abbiamo trovato una mancanza di correlazione tra frequenza di mutazione e valori CU. I codoni DMD meno frequentemente usati (UCG, CCG, ACG e GCG) sono ricchi si GC e infatti raramente ospitano mutazioni. A ogni modo, altri codoni ricchi di GC come CGC, sono raramente usati nel *DMD* ma sono frequentemente sito di mutazioni (67) e, viceversa, il codone usato veramente più di frequente, UGC, ospita solo 9 mutazioni. CAG è il codone mutato più di frequente (56,5% con 249 mutazioni) ma non il più usato dal gene *DMD*.

Pertanto, i siti di mutazioni mappati sui codoni *DMD* non sono correlati con la frequenza di utilizzo del codone.

E’ interessante notare che, il locus di architettura del *DMD* suggerisce possibili spiegazioni del suo comportamento CUB unico. E’ risaputo che lunghezza degli introni e della proteina, pattern di espressione e valori CUB sono variant genomiche che influenzano il tasso di evoluzione dei geni. I geni con bassi livelli di espressione di mRNA/proteine tendon ad evolversi rapidamente, hanno introni grandi, codice per protein più grandi e hanno un CUB molto basso. Il gene *DMD* incontra tuute queste regole: ha introni enormi, codifica una proteina di grosso peso molecolare, che è altamente tessuto-specifica ma scarsamente abbondante e coerentemente mostra assenza di CUB estremo e nessun “codone zero”. Basandoci su queste metriche, il gene *DMD* potrebbe essersi rapidamente evoluto durante l’evoluzione, e noi possiamo affermare che un simile comportamento si è verificato in altri geni DC che mostrano simili caratteristiche di CUB basso come quelle osservate in alcuni geni DC della pelle e del rene. A ulteriore sostegno della nostra osservazione, gli eucarioti mostrano una correlazione negative tra la lunghezza del gene e il CUB, mentre nelle mosche, una relazione antagonistica tra il CUB e il numero e la lunghezza degli introni è stata descritta e, più intrigantemente, i geni senza introni janno un CUB molto alto.

Le impronte CUB nrei geni DC dovrebbero avere implicazioni sull’ottimizzazione dei codoni. Infatti, i gen DC, specialmente quelli del muscolo che abbiamo studiato, mostrano un’impronta mostrante un CUB basso. Questo significan che i geni DC del muscolo applicano ancora la piena ridindanza del codone e, viceversa, ch ei geni sintetici

corrispondenti progettati per terapie geniche avranno bisogno di una dirompente ottimizzazione del codone tramite applicazione del Codon Adaptation Index (CAI).

Al contrario, i geni NDC mostrano un CUB alto spontaneo, che porta a molti “codoni zero” nella sequenza del codice, quindi mancando molti tipi di codoni sinonimi. Dal moment oche abbiamo mostrato che i geni DC (specialmente il muscolo) hanno un CUB orientate alla malattia, la piena pplicabilità a “tutti” i geni della malattia degli algoritmi di computazione correntemente usati sviluppati per l’ottimizzazione del codone e gli approcci della terapia genica possono essere messi in dubbio (Gould N, Hendy O, & Papamichail D. Computational tools and algorithms for designing customized synthetic genes. Frontiers in bioengineering and biotechnology. (2014); 2, 41). Infatti, alcuni codoni rari non ottimali potrebbero dover essere preservati per l’efficienza di traduzione di proteine o, più interessantemente, per la regolazione dell’espressione gene- e tessuto-specifica come precedentemente riportato, sebbene non negli umani.

RIVENDICAZIONI

1. Un metodo implementato al computer per determinare il valore del bias di utilizzo del codone di un selezionato gene causante malattia, comprendendo i seguenti passaggi:

(i) raccogliere le sequenze di uno o più geni causanti malattia espressi in uno o più tessuti di un organismo;

(ii) raccogliere le sequenze di una pluralità di geni non causanti malattia espressi nello stesso tessuto e organismo dei geni nel passaggio (i);

iii) determinare il calcolo indipendente della frequenza di utilizzo del codone per i 19 aminoacidi essenziali (metionina e triptofano esclusi) in ogni gene raccolto nel passaggio (i) e nel passaggo (ii);

iv) comparare la frequenza di utilizzo del codone determinata nel passaggio (iii) in modo da ottenere il valore del bias di utilizzo del codone.

2. Il metodo secondo la rivendicazione 1, in cui detto organismo è un mammifero.

3. Il metodo secondo qualsiasi delle rivendicazioni 1 o 2, in cui detto tessuto è selezionato tra muscolo, pelle, rene o qualsiasi altro tessuto coinvolto nella malattia gestibile con terapia genica.

4. Il metodo secondo qualsiasi delle rivendicazioni da 1 a 3, in cui detta malattia è una malattia rara.

5. Il metodo secondo qualsiasi delle rivendicazioni da 1 a 4, in cui detta malattia è selezionata tra distrofie muscolari, miopatie congenite, malattia renale tubulointerstiziale, rene policistico di tipo 1, ipercheratosi epidermolitica, displasia ectodermica.

6. Il metodo secondo qualsiasi delle rivendicazioni da 1 a 5, in cui il numero di detti geni raccolti nel passaggio (i) e/o nel passaggio (ii) è il più alto numero possible nella banca dati per la malattia selezionata, o almeno dall’80 al 99%.

7. il metodo secondo qualsiasi delle rivendicazioni da 1 a 6, in cui detto passaggio iii) di determinare la frequenza di uso del codone è svolta tramite il calcolo del Codon Adaptation Index (CAI).

8. Il metodo secondo qualsiasi delle rivendicazioni da 1 a 7, comprendente un ulteriore passaggio di calcolo della conservazione del codone del selezionato gene causante malattia tra i mammiferi, in modo da utilizzare detto calcolo per ottenere il valore del bias di uso del codone.

9. Il metodo secondo qualsiasi delle rivendicazioni da 1 a 7, in cui detto mammifero è selezionato da uno o più tra *R. ferrumequinum* (ferro di cavallo maggiore), *M. musculus* (topo)*, F. catus* (gatto), *C. lupus familiaris* (cane)*, E. caballus* (cavallo), *B. Taurus* (toro)*, M. murinus* (lemure topo grigio)*, G. variegatus* (colugo della sonda)*, C. jacchus* (uistitì)*, M. mulatta* (macaco)*, N. leucogenys* (gibbone)*, P. abelii* (orangotango)*, G. gorilla* (gorilla)*, P. troglodytes* (scimpanzè) *e H. sapiens* (umano).

10. Un metodo implementato al computer per progettare una molecola sintetica di acido nucleico di un selezionato gene causante malattia espresso in uno o più tessuti di un organismo, comprendente i passaggi dei metodi secondo qualsiasi delle rivendicazioni da 1 a 9 e un ulteriore passaggio di ottimizzazione del codone di detta molecola sintetica di acido nucleico basata sul valore del bias di utilizzo del codone ottenuto nel passaggio iv).

11. Un metodo secondo la rivendicazione 10 comprendente i seguenti passaggi:

(i) raccogliere le sequenze di uo o più geni causanti malattia espresso in uno più tessuti di un organismo, preferibilmente calcolando la conservazione del codone del gene causante malattia selezionato tra i mammiferi;

(ii) raccogliere le sequenze di una pluralità di geni non causanti malattia espressi nello stesso tessuto e organismo dei geni nel passaggio (i);

iii) determinare il calcolo indipendente della frequenza di utilizzo dei codoni per i 19 aminoacidi essenziali (escluse metionina e triptofano) in ogni gene raccolto nei passaggi (i) e (ii);

iv) paragonare la frequenza di utilizzo del codone determinata nel passaggio (iii) in modo da ottenere il valore del bias di utilizzo del codone, per identificare codoni a comparsa (codoni tessuto-specifici, codoni tessuto-specifici e codoni gene-specifici) in modo da prioritizzare i codoni più diversamente usati nel gene e nel/i tessuto/i di interesse;

v) progettare la molecola sintetica di acido nucleico di detto gene causante malattia, modificando la struttura secondaria o terziaria di detto gene utilizzando i codoni prioritizzati ottenuti nel passaggio iv).

12. Il metodo secondo la rivendicazione 11, in cui detto passaggio vi) di prioritizzare i codoni usati più differentemenete è svolto attraverso il raggruppamento dei valori di frequenza di utilizzo del codone ottenuti nel passaggio iii) utilizzando un algoritmo di raggruppamento gerarchico con un valore p minore di 0.05 e selezionando i codoni meno usati e/o più usati per le specie, il tessuto e il gene di interesse.

13. Il metodo secondo qualsiasi delle rivendicazioni da 11 a 12, in cui in detto passaggio v) di modificare la struttura secondaria o terziaria di detto gene causante malattia è eseguito tramite la sostituzione e/o rimozione dei codoni meno utilizzati e i codoni più utilizzati, ottenuti nel passaggio iv), nella sequenza di detto gene causante malattia.

14. Un metodo per preparare una molecola sintetica di acido nucleico di un selezionato gene causante malattia, comprendente i passaggi di qualsiasi delle rivendicazioni da 10 a 13 e un ulteriore passaggio vi) di sintetizzare una molecola di acido nucleico comprendente la sequenza di acido nucleico ottimizzata per il codone del passaggio (v).

15. Una proteina, un reagente ricombinante o uno strumento molecolare comprendente la sequenza di acido nucleico ottimizzata per il codone ottenibile tramite il metodo secondo le rivendicazioni da 10 e 13.

16. La proteina, il reagente ricombinante o lo strumento molecolare secondo la rivendicazione 15 per l'uso nel trattamento di malattie gestibili con terapia genica, per esempio in cui detta malattia è selezionata tra distrofie muscolari, miopatie congenite, malattia renale tubulointerstiziale, rene policistico tipo 1, ipercheratosi epidermolitica, displasia ectodermica.

17. Un programma per computer comprendente istruzioni che, quando il programma è eseguito da un computer, comporta che il computer svolga i passaggi del metodo secondo le rivendicazioni 1-13.

18. Un supporto di memorizzazione leggibile dal computer comprendente istruzioni che, quando eseguito da un computer, comporta che il computer svolga passaggioi passaggi del metodo secondo le rivendicazioni 1-13.

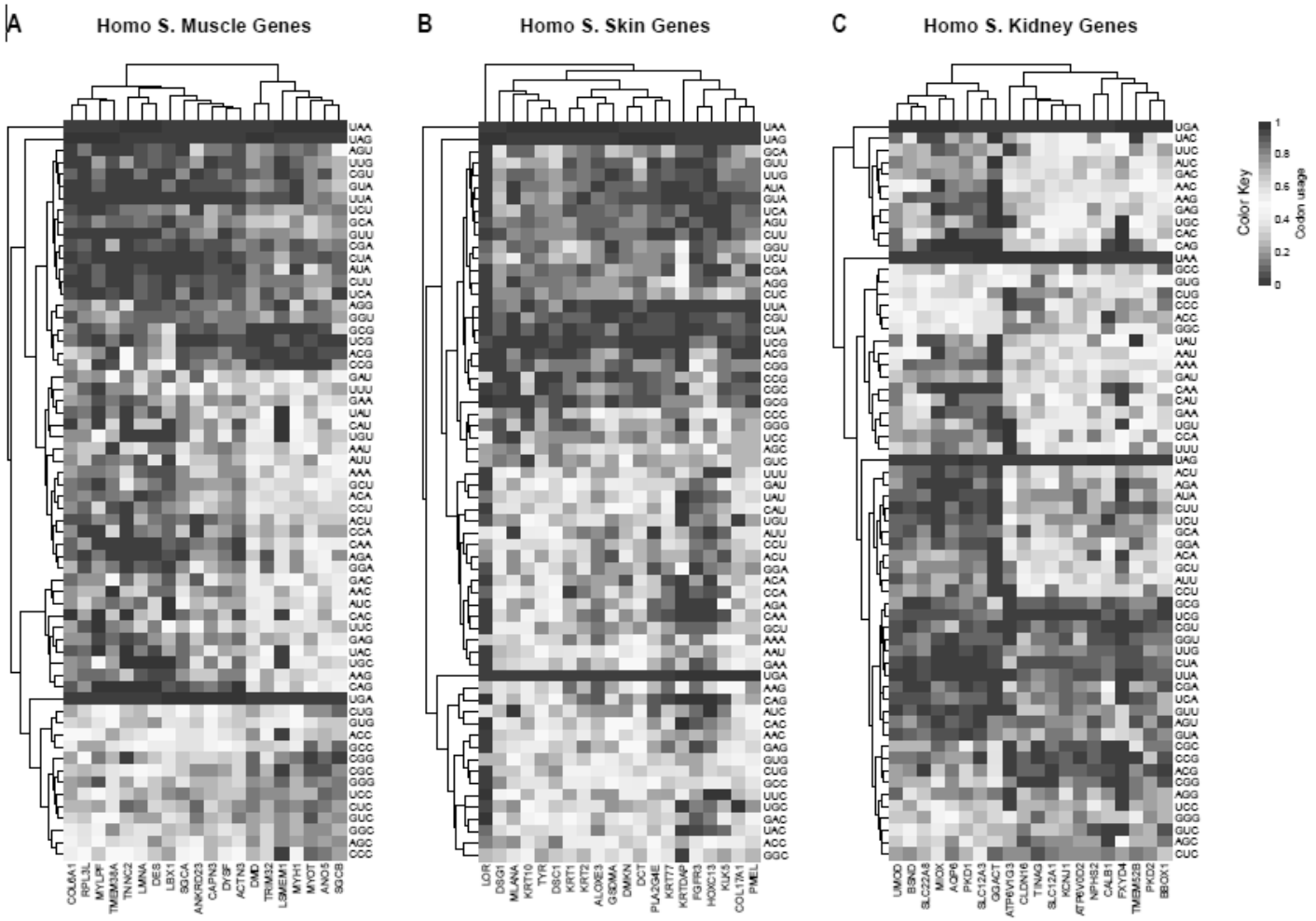

Figure 1

Inglese	**Italiano**
Muscle Genes	Geni del muscolo
Skin Genes	Geni della pelle
Kidney Genes	Geni del rene
Color key	Chiave colore
Codon Usage	Utilizzo del codone

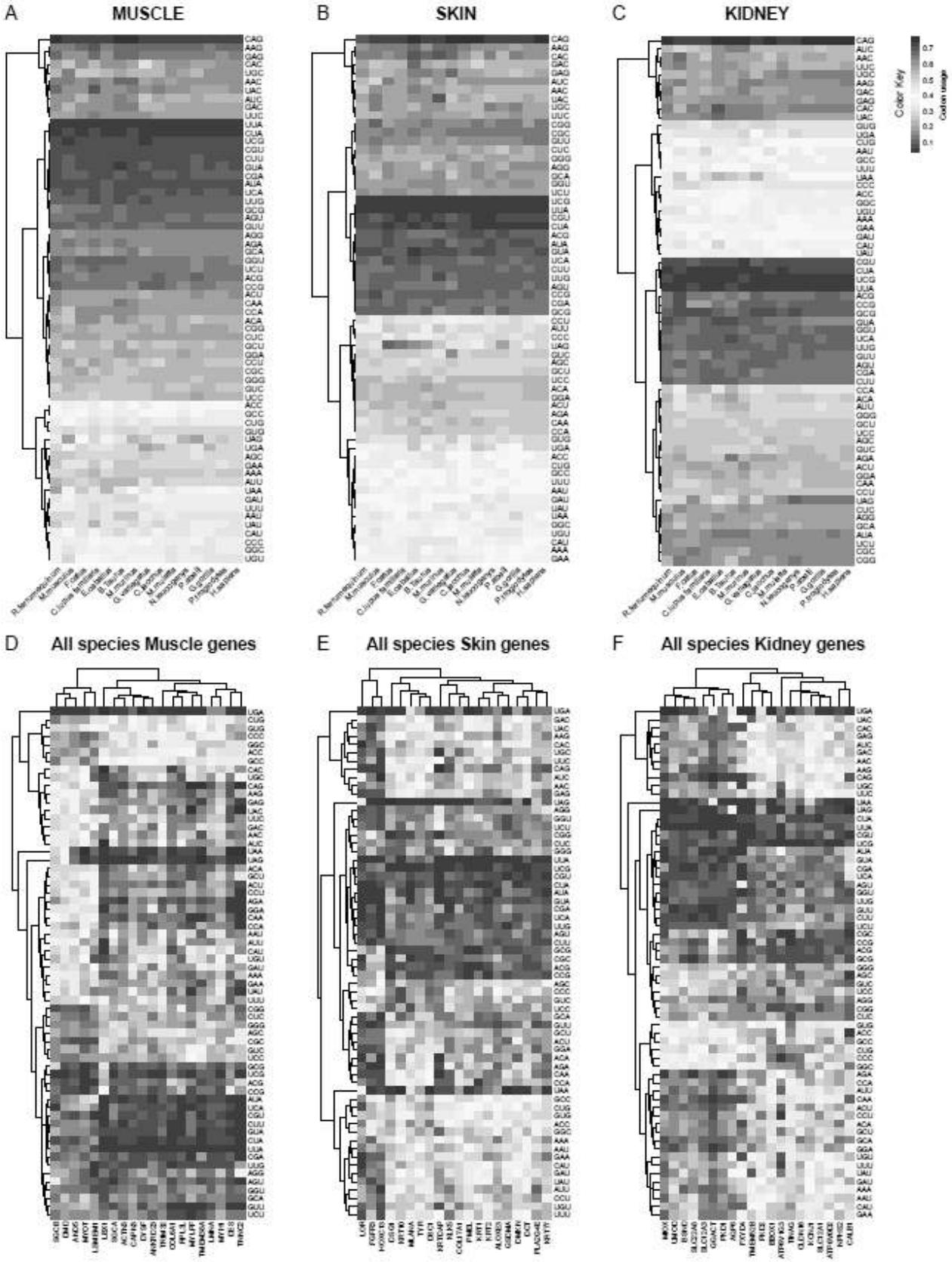
A
MUSCLE
B
SKIN
C
KIDNEY
Color Key
Codon Usage
D
All species Muscle genes
E
All species Skin genes
F
All species Kidney genes

Figure 2

Inglese	**Italiano**
Muscle Genes	Geni del muscolo
Skin Genes	Geni della pelle
Kidney Genes	Geni del rene
Color key	Chiave colore
Codon Usage	Utilizzo del codone
Muscle	Muscolo
Skin	Pelle
Kidney	Rene
All species	Tutte le specie

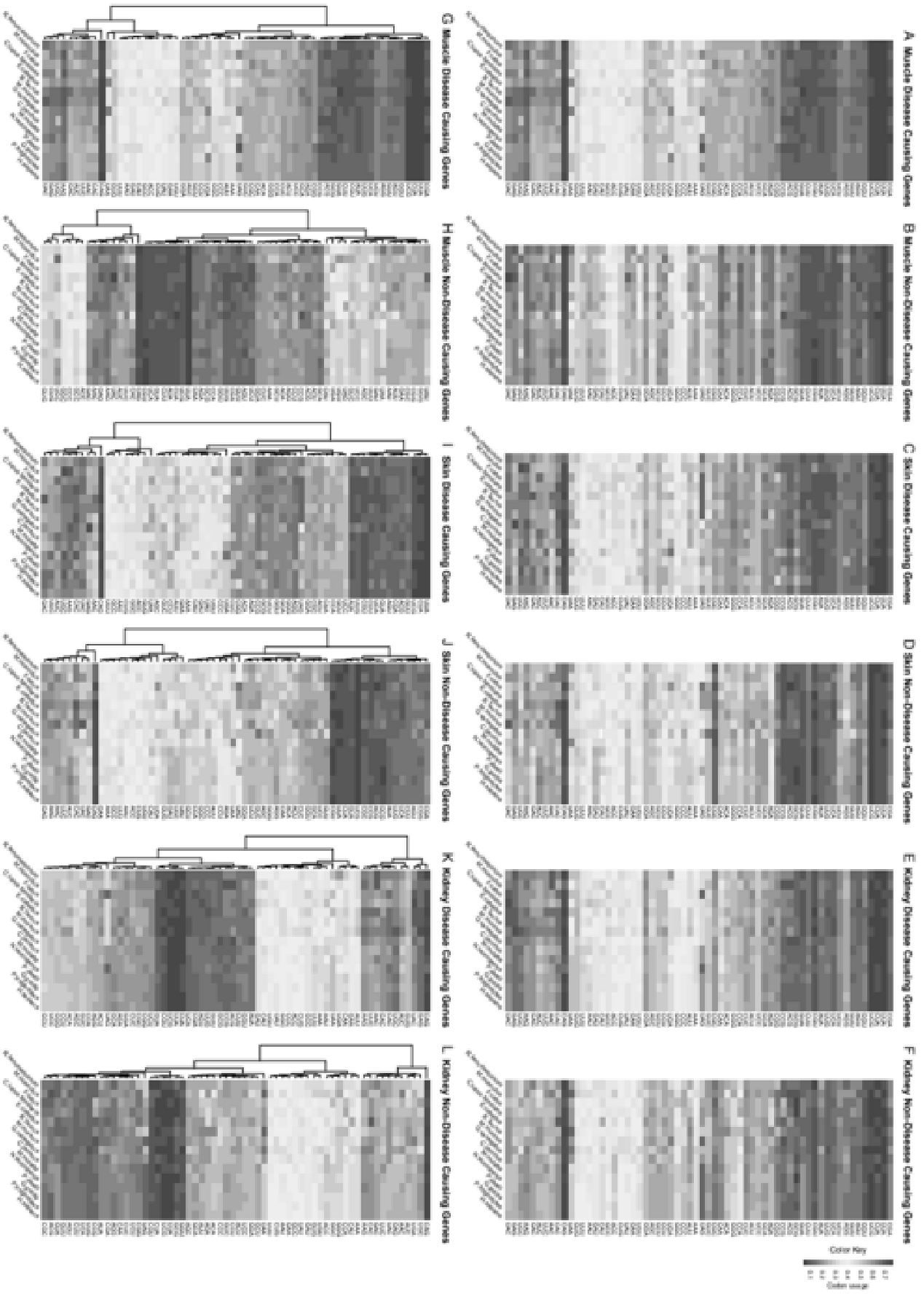
A Muscle Disease Causing Genes
B Muscle Non-Disease Causing Genes
C Skin Disease Causing Genes
D Skin Non-Disease Causing Genes
E Kidney Disease Causing Genes
F Kidney Non-Disease Causing Genes
G Muscle Disease Causing Genes
H Muscle Non-Disease Causing Genes
I Skin Disease Causing Genes
J Skin Non-Disease Causing Genes
K Kidney Disease Causing Genes
L Kidney Non-Disease Causing Genes
Color Key

Figure 3

Inglese	**Italiano**
Muscle disease causing Genes	Geni del muscolo causanti malattia
Skin disease causing Genes	Geni della pelle causanti malattia
Kidney disease causing Genes	Geni del rene causanti malattia
Color key	Chiave colore
Codon Usage	Utilizzo del codone
Muscle non-disease causing Genes	Geni del muscolo non causanti malattia
Skin non-disease causing Genes	Geni della pelle non causanti malattia
Kidney non-disease causing Genes	Geni del rene non causanti malattia

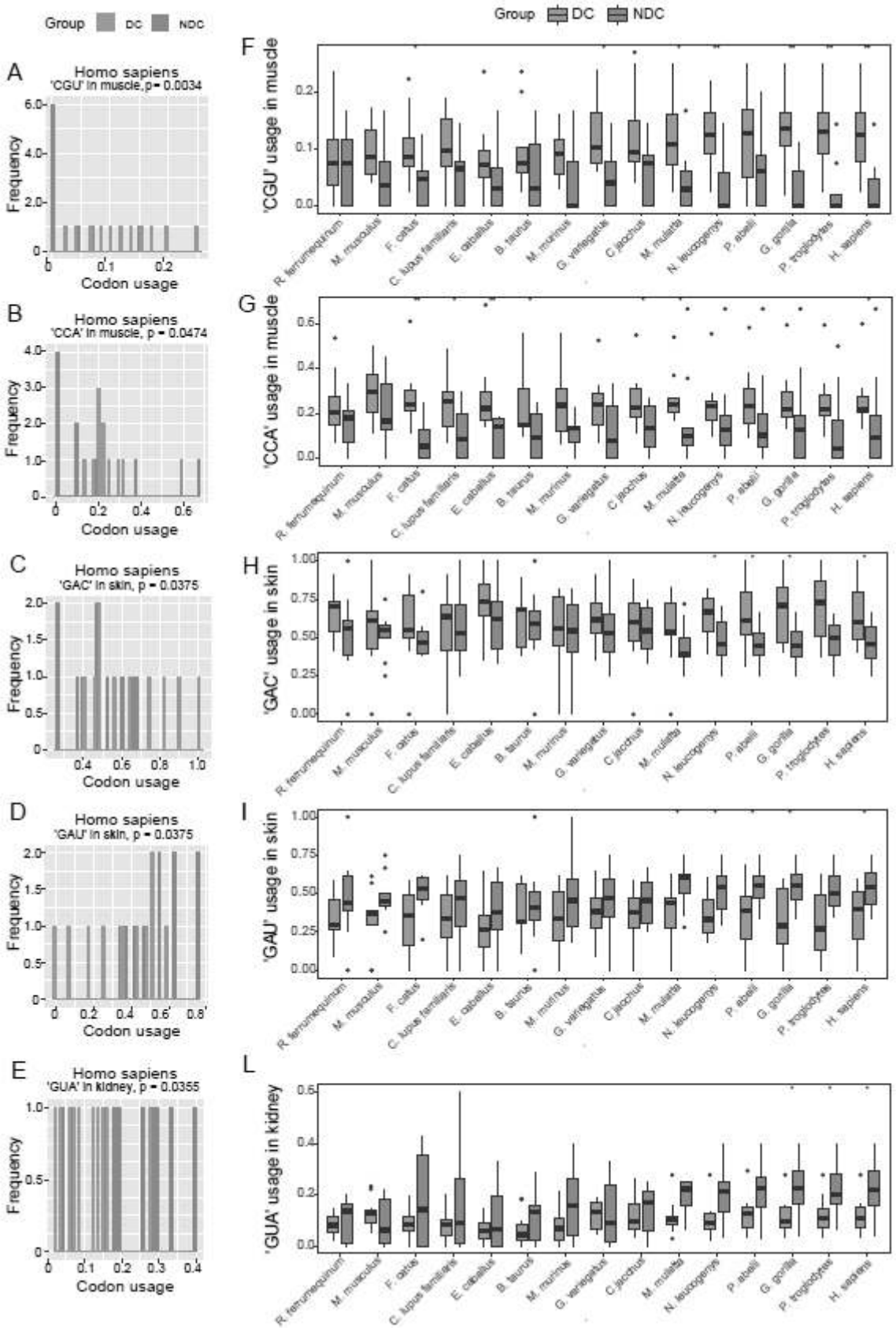

Figure 4

Inglese	Italiano
Group	Gruppo
DC (Disease causing)	Causante la malattia

NDC (Non-Disease causing)	Non causante la malattia
Usage in muscle	Uso nel muscolo
Usage in skin	Uso nella pelle
Usage in kidney	Uso nel rene
Frequency	Frequenza
Codon usage	Utilizzo del codone

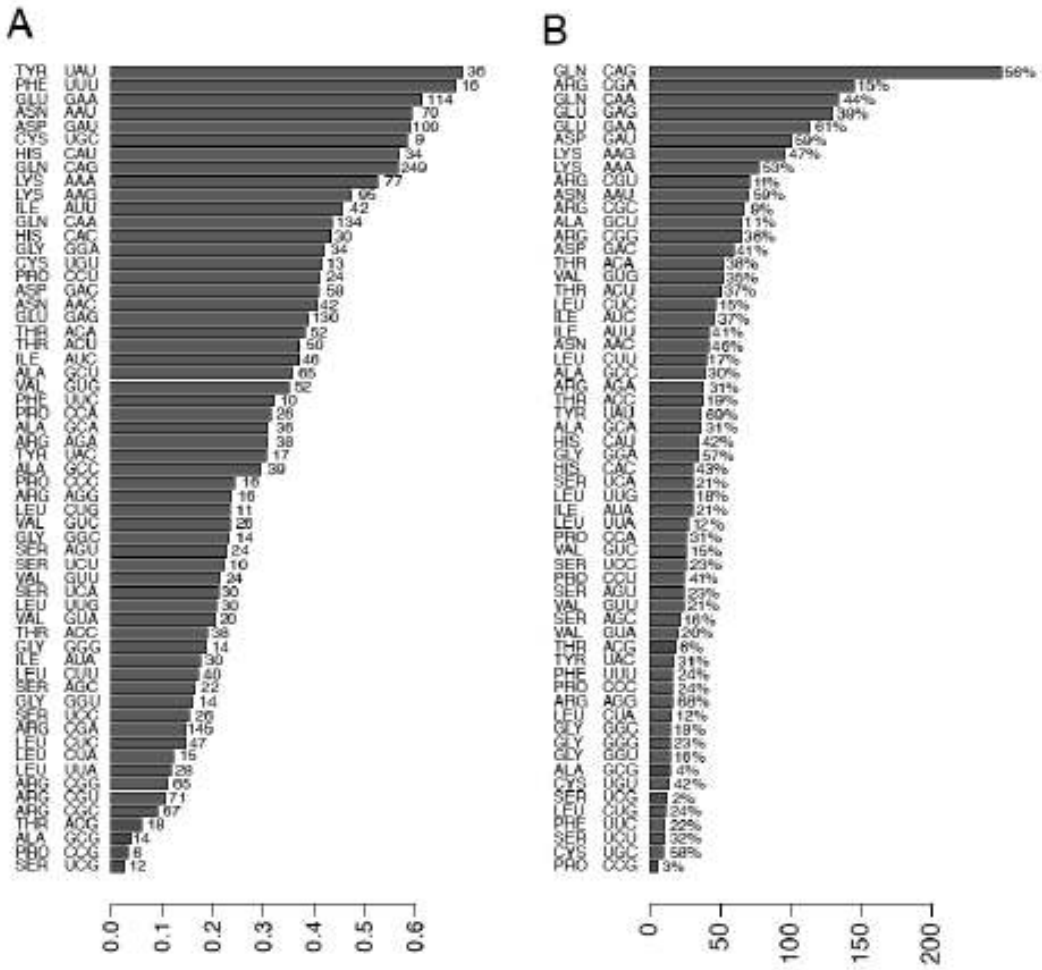

Figure 6

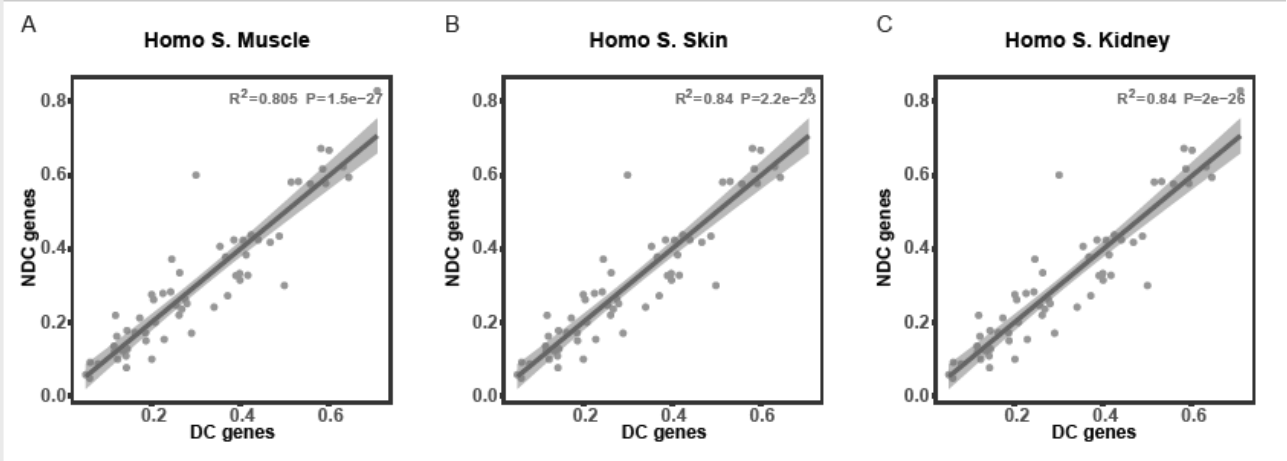

Figure 6

Inglese	**Italiano**
Homo S. Muscle	Muscolo Homo S.
Homo S. Skin	Pelle Homo S.
Homo S. Kidney	Rene Homo S.
DC (disease causing) genes	Geni causanti la malattia
NDC (Non-disease causing) genes	Geni non causanti la malattia

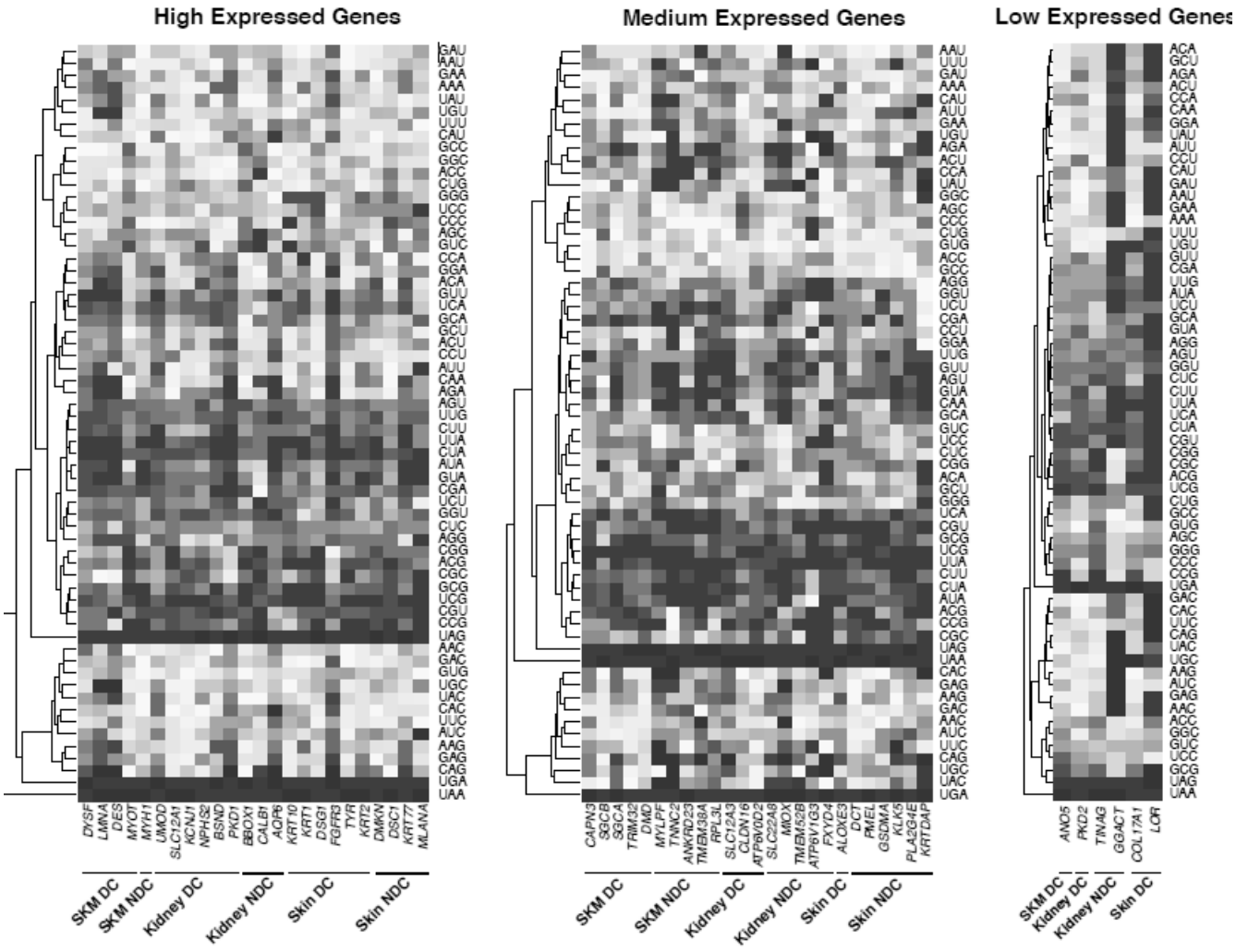

Figure 7

Inglese	**Italiano**
High Expressed Genes	Geni altamente espressi
Medium Expressed Genes	Geni mediamente espressi
Low Expressed Genes	Geni poco espressi
Kidney DC	Rene causante malattia
Kidney NDC	Rene non causante malattia
Skin DC	Pelle causante malattia
Skin NDC	Pelle non causante malattia

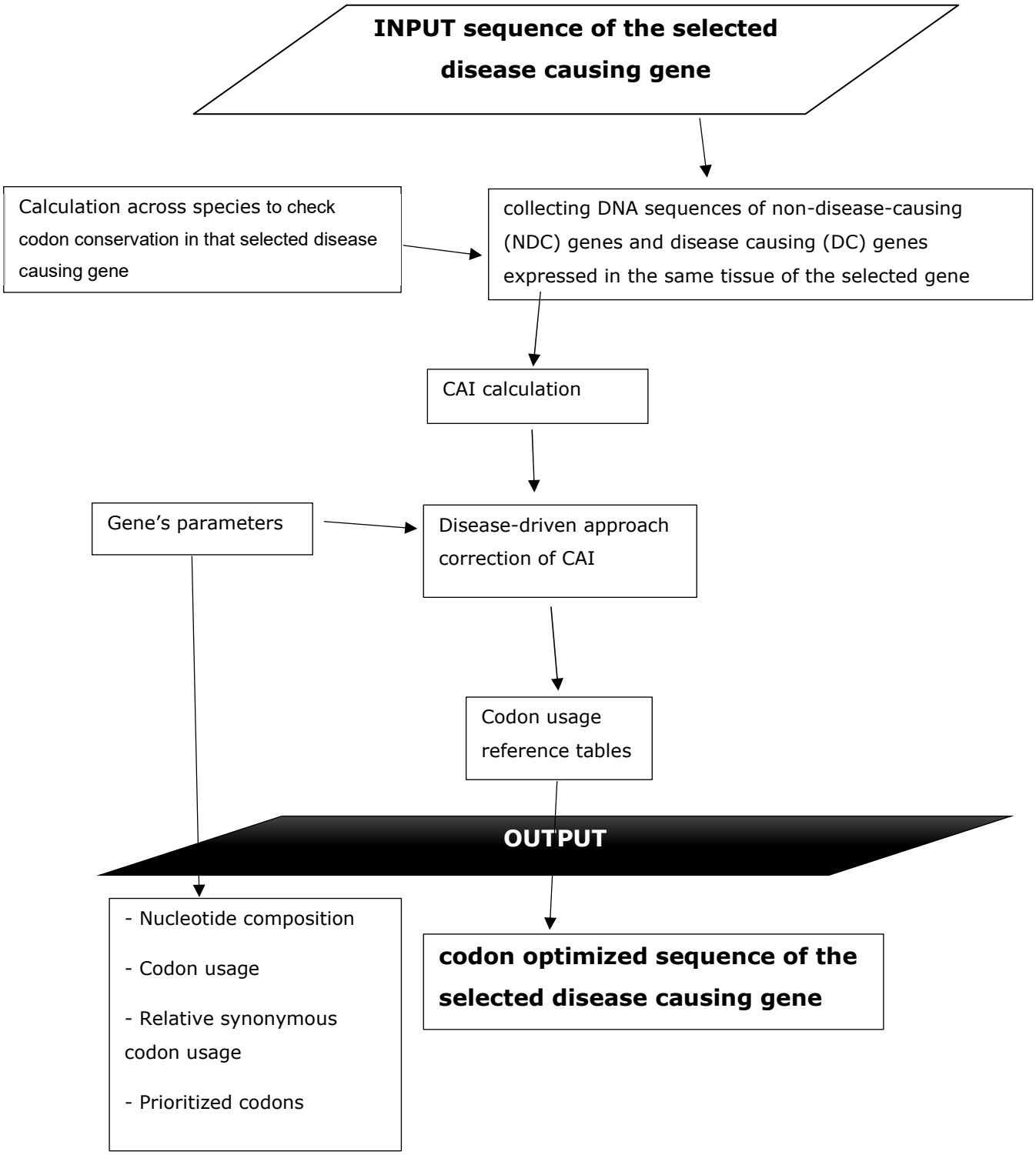
INPUT sequence of the selected disease causing gene
Calculation across species to check codon conservation in that selected disease causing gene
collecting DNA sequences of non-disease-causing (NDC) genes and disease causing (DC) genes expressed in the same tissue of the selected gene
CAI calculation
Gene's parameters
Disease-driven approach correction of CAI
Codon usage reference tables
OUTPUT
- Nucleotide composition
- Codon usage
- Relative synonymous codon usage
- Prioritized codons
codon optimized sequence of the selected disease causing gene

FIGURE 8

Inglese	**Italiano**
sequence of the selected disease-causing gene	sequenza del gene selezionato che causa la malattia
Calculation across species to check codon conservation in that selected disease causing gene	Calcolo tra specie per verificare la conservazione del codone in quel gene selezionato che causa la malattia
collecting DNA sequences of non-disease-causing (NDC) genes and disease causing (DC) genes expressed in the same tissue of the selected gene	raccogliere sequenze di DNA di geni non causanti la malattia (NDC) e geni causanti la malattia (DC) espressi nello stesso tessuto del gene selezionato
CAI calculation	Calcolo CAI
Gene's parameters	I parametri del gene
Disease-driven approach correction of CAI	Approccio guidato dalla malattia correzione del CAI
Codon usage reference tables	Tabelle di riferimento sull'utilizzo dei codoni
codon optimized sequence of the selected disease causing gene	sequenza ottimizzata del codone del gene selezionato che causa la malattia
Nucleotide composition	Composizione nucleotidica
Codon usage	Utilizzo del codone
Relative synonymous codon usage	Uso del relativo sinonimo del codone
Prioritized codons	Codoni prioritari

Ministero dello Sviluppo Economico
DIREZIONE GENERALE SVILUPPO PRODUTTIVO E COMPETITIVITA'-
UFFICIO ITALIANO BREVETTI E MARCHI

RAPPORTO DI RICERCA

Numero della domanda
IO 117572
IT 202200006119

DOCUMENTI CONSIDERATI DI RILIEVO

Categoria	Citazione del documento con indicazione, se appropriata, delle parti rilevanti	Rivendicazioni rilevanti	CLASSIFICAZIONE DELLA DOMANDA (IPC)
	INCOMPLETE SEARCH see sheet C -----		INV. G16B30/00 G16B40/30 G16B20/20
X	FANG M ET AL: "O.11Dystrophin gene codon usage lacks extreme codon bias and shows non-random codon distribution of disease-causing mutations", NEUROMUSCULAR DISORDERS, ELSEVIER LTD, GB, vol. 29, 29 September 2019 (2019-09-29), XP085844575, ISSN: 0960-8966, DOI: 10.1016/J.NMD.2019.06.029 [retrieved on 2019-09-29] * the whole document * -----	1-9,17, 18	
X	US 2017/368198 A1 (XIAO XIAO [US] ET AL) 28 December 2017 (2017-12-28) * the whole document * -----	10-14	
X	WO 2021/226461 A1 (TRANSLATE BIO INC [US]) 11 November 2021 (2021-11-11) * the whole document * -----	10-14	CAMPI TECNICI RICERCATI (IPC) G16B
X	GOULD N ET AL: "Computational tools and algorithms for designing customized synthetic genes", FRONTIERS IN BIOENGINEERING AND BIOTECHNOLOGY, FRONTIERS RESEARCH FOUNDATION, CH, vol. 2, 6 October 2014 (2014-10-06), pages 41-1, XP002755946, ISSN: 2296-4185, DOI: 10.3389/FBIOE.2014.00041 [retrieved on 2014-01-01] * the whole document * -----	10-14	
	-/--		

Questo rapporto di ricerca è stato redatto sulla base di tutte le rivendicazioni

Data di completamento della ricerca	Esaminatore
Munich	21 November 2022	Lüdemann, Susanna

CATEGORIA DEI DOCUMENTI CITATI

X : di particolare rilevanza se considerato singolarmente
Y : di particolare rilevanza se combinato con un altro documento della stessa categoria
A : informazione generica
O : divulgazione orale
P : documento intermedio

T : teoria o principio alla base dell'invenzione
E : documento brevettuale antecedente, ma pubblicato dopo o alla data di deposito
D : documento citato nella domanda
L : documento citato per altre ragioni
.....
& : membro della stessa famiglia di brevetti, documento corrispondente

3
EPO FORM 1503 07.08 (P04C74)

Ministero dello Sviluppo Economico
DIREZIONE GENERALE SVILUPPO PRODUTTIVO E COMPETITIVITA'-
UFFICIO ITALIANO BREVETTI E MARCHI

RAPPORTO DI RICERCA

Numero della domanda
IO 117572
IT 202200006119

DOCUMENTI CONSIDERATI DI RILIEVO

Categoria	Citazione del documento con indicazione, se appropriata, delle parti rilevanti	Rivendicazioni rilevanti	CLASSIFICAZIONE DELLA DOMANDA (IPC)
X,P	ROSSI RACHELE ET AL: "Calculating and comparing codon usage values in rare disease genes highlights codon clustering with disease-and tissue- specific hierarchy", PLOS ONE, vol. 17, no. 3, 31 March 2022 (2022-03-31), page e0265469, XP055982992, DOI: 10.1371/journal.pone.0265469 * the whole document *	1-13,17, 18	

CAMPI TECNICI RICERCATI (IPC)

Questo rapporto di ricerca è stato redatto sulla base di tutte le rivendicazioni

Luogo della ricerca	Data di completamento della ricerca	Esaminatore
Munich	21 November 2022	Lüdemann, Susanna

CATEGORIA DEI DOCUMENTI CITATI

X : di particolare rilevanza se considerato singolarmente
Y : di particolare rilevanza se combinato con un altro documento della stessa categoria
A : informazione generica
O : divulgazione orale
P : documento intermedio

T : teoria o principio alla base dell'invenzione
E : documento brevettuale antecedente, ma pubblicato dopo o alla data di deposito
D : documento citato nella domanda
L : documento citato per altre ragioni
.....
& : membro della stessa famiglia di brevetti, documento corrispondente

3

EPO FORM 1503 07.08 (P04C74)

RICERCA INCOMPLETA
ALLEGATO C

Numero della domanda

IO 117572
IT 202200006119

This search report has not been established in respect of certain claims, because they relate to parts of the application that do not comply with the prescribed requirements to such an extent that no meaningful search can be carried out, specifically:

Claim(s) completely searchable:
1-14, 17, 18

Claim(s) not searched:
15, 16

Reason for the limitation of the search:

Present claim 15,16 encompass compounds defined only by their desired function, contrary to the requirement of clarity, because the "result-to-be-achieved" type of definition does not allow the scope of the claim to be ascertained. The fact that any compound could be screened does not overcome this objection, as the skilled person would not know beforehand whether it fell within the scope claimed. Undue experimentation would be required to screen compounds randomly. Non-compliance with the substantive provisions is such that no meaningful search of claim 15,16 could be carried out at all.

ALLEGATO AL RAPPORTO DI RICERCA SULLA DOMANDA DI BREVETTO ITALIANO N. IO 117572 IT 202200006119

Questo allegato enumera i membri della famiglia di brevetti relativi a documenti brevettuali citati nel summenzionato rapporto di ricerca.
I membri sono indicati come da database dell'Ufficio Europeo dei Brevetti al 21-11-2022
L'Ufficio Europeo dei Brevetti non si assume alcuna responsabilità per queste indicazioni, che vengono fornite a solo scopo informativo.

Documenti brevettuali citati nel rapporto di ricerca	Data di pubblicazione	Membri della famiglia di brevetti	Data di pubblicazione
US 2017368198 A1	28-12-2017	AU 2017281983 A1	13-12-2018
		BR 112018076394 A2	26-03-2019
		CA 2971303 A1	21-12-2017
		CN 109641944 A	16-04-2019
		CO 2019000395 A2	30-04-2019
		EP 3472194 A1	24-04-2019
		IL 263199 A	31-12-2018
		JP 6793758 B2	02-12-2020
		JP 2019525740 A	12-09-2019
		JP 2021045136 A	25-03-2021
		KR 20190027358 A	14-03-2019
		PE 20190563 A1	22-04-2019
		PH 12018502645 A1	14-10-2019
		RU 2019101208 A	21-07-2020
		SG 11201811115T A	30-01-2019
		TW 201812010 A	01-04-2018
		US 2017368198 A1	28-12-2017
		US 2020376141 A1	03-12-2020
		WO 2017221145 A1	28-12-2017
WO 2021226461 A1	11-11-2021	NONE	

Ministero dello Sviluppo Economico
DIREZIONE GENERALE SVILUPPO PRODUTTIVO E COMPETITIVITA' -
UFFICIO ITALIANO BREVETTI E MARCHI

OPINIONE SCRITTA

N. dossier	Data di deposito *(gg/mm/aa)*	Data di priorità *(gg/mm/aa)*	N. domanda
IO117572	29.03.2022		IT202200006119

Classificazione Internazionale dei Brevetti (IPC)
INV. G16B30/00 G16B40/30 G16B20/20

Richiedente
UNIVERSITÀ DEGLI STUDI DI FERRARA

Questa opinione fornisce indicazioni riguardanti i seguenti elementi:

- ☒ Riquadro N. I — Base dell'opinione
- ☐ Riquadro N. II — Priorità
- ☒ Riquadro N. III — Non-redazione di un'opinione a riguardo di novità, attività inventiva e applicazione industriale
- ☐ Riquadro N. IV — Violazione del requisito d'unità dell'invenzione
- ☒ Riquadro N. V — Dichiarazione motivata a riguardo di novità, attività inventiva o applicazione industriale; citazioni e spiegazioni giustificative della dichiarazione
- ☐ Riquadro N. VI — Particolari documenti citati
- ☐ Riquadro N. VII — Difetti particolari nella domanda
- ☒ Riquadro N. VIII — Osservazioni particolari a riguardo della domanda

	Esaminatore
	Lüdemann, Susanna

N. domanda

OPINIONE SCRITTA

IT202200006119

Riquadro N. I Base dell'opinione

1. Questa opinione è stata redatta sulla base delle ultime rivendicazioni depositate prima dell'inizio della ricerca nella tecnica anteriore.

2. Per quanto concerne eventuali sequenze di nucleotidi e/o amminoacidi descritte nella domanda e necessarie per l'invenzione di cui oggetto nelle rivendicazioni, questa opinione è stata redatta sulla base di:

   a. tipo di materiale:

   ☐ una sequenza di DNA

   ☐ una o più tabelle relative alla sequenza di DNA

   b. formato del materiale:

   ☐ cartaceo

   ☐ elettronico

   c. momento di deposito o presentazione:

   ☐ depositato insieme alla domanda al momento del deposito della medesima

   ☐ depositato insieme alla domanda in formato elettronico

   ☐ presentato successivamente al fine della ricerca d'anteriorità

3. ☐ Inoltre, ove sia stata depositata o presentata più di una versione o copia di una sequenza di DNA e/o tabella ad essa relativa, è stata presentata anche la dichiarazione obbligatoria che le informazioni contenute nelle copie successive o addizionali sono identiche a quelle nella domanda come depositata o che, in ogni caso, non vanno oltre il contenuto della domanda depositata originariamente.

4. Note aggiuntive:

N. domanda

OPINIONE SCRITTA

IT202200006119

Riquadro N. III Non-redazione di un'opinione a riguardo di novità, attività inventiva e applicazione industriale

L'analisi concernente la novità, l'attività inventiva e l'applicazione industriale dell'invenzione rivendicata non è stata condotta per quanto riguarda

☐ la domanda in toto

☒ rivendicazioni N. 15, 16

poiché:

☐ la suddetta domanda, o le suddette rivendicazioni N. hanno il seguente oggetto, per il quale non si necessita di fare una ricerca d'anteriorità *(specificare)*:

☐ la descrizione, le rivendicazioni o i disegni *(specificarne i dettagli di seguito)* o le suddette rivendicazioni N. , mancano a tal punto di chiarezza che non è possibile redigere un'opinione significativa *(specificare)*:

☐ le rivendicazioni, o suddette rivendicazioni N. sono insufficientemente fondate sulla descrizione a tal punto che risulta impossibile redigere un'opinione significativa *(specificare)*:

☒ un rapporto di ricerca non è stato redatto sulla domanda in toto o sulle suddette rivendicazioni N. 15, 16

☐ non è risultato possibile redigere un'opinione significativa in quanto la sequenza di DNA non era presente, oppure non è stata presentata nel formato internazionale (WIPO ST.25).

☐ non è risultato possibile redigere un'opinione significativa senza le tabelle relative alle sequenze di DNA; oppure tali tabelle non era presenti nel formato elettronico.

☒ Si veda il Riquadro Supplementare per ulteriori dettagli.

Riquadro N. VDichiarazione motivata a riguardo di novità, attività inventiva o applicazione industriale; citazioni e spiegazioni giustificative della dichiarazione

1. 1. Dichiarazione

Novità (N)	Sì:	Rivendicazioni	1-14, 17, 18
	No:	Rivendicazioni	
Attività inventiva (IS)	Sì:	Rivendicazioni	
	No:	Rivendicazioni	1-14, 17, 18
Applicazione industriale (IA)	Sì:	Rivendicazioni	1-14, 17, 18
	No:	Rivendicazioni	

2. 2. Citazioni e spiegazioni

si veda l'allegato

Riquadro N. VIII Osservazioni particolari a riguardo della domanda

si veda l'allegato

1 **Re Item III**

Non-establishment of opinion with regard to novelty, inventive step and industrial applicability

The subject-matter of claims 15 and 16 are considered as reach-through claims.

As a matter of fact, the subject matter of said claims covers limitless and untried downstream developments in relation to yet to be demonstrated functions/ activities. The claims amount to no more than an invitation to set up further research programmes for which no guidance is forthcoming and, therefore, it is an undue burden to put the claimed subject matter into practice, i.e. to identify all the relevant compounds having said desired property without indication of any structural limitation. A further lack of clarity arises because the public can not ascertain whether or not a particular compound falls within the scope of such a claim. Consequently, the examination can never with any certainty, ascertain whether or not such claims are distinguished over the state-of-the-art.

2 **Re Item V**

Reasoned statement with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement

2.1 Reference is made to the following documents:

D1	FANG M ET AL: "O.11Dystrophin gene codon usage lacks extreme codon bias and shows non-random codon distribution of disease-causing mutations", NEUROMUSCULAR DISORDERS, ELSEVIER LTD, GB, vol. 29, 29 September 2019 (2019-09-29), XP085844575, ISSN: 0960-8966, DOI: 10.1016/J.NMD.2019.06.029 [retrieved on 2019-09-29]
D2	US 2017/368198 A1 (XIAO XIAO [US] ET AL) 28 December 2017 (2017-12-28)
D3	WO 2021/226461 A1 (TRANSLATE BIO INC [US]) 11 November 2021 (2021-11-11)
D4	GOULD N ET AL: "Computational tools and algorithms for designing customized synthetic genes", FRONTIERS IN BIOENGINEERING AND BIOTECHNOLOGY, FRONTIERS RESEARCH FOUNDATION, CH,

vol. 2, 6 October 2014 (2014-10-06), pages 41-1, XP002755946, ISSN: 2296-4185, DOI: 10.3389/FBIOE.2014.00041 [retrieved on 2014-01-01]

D5 Rossi Rachele ET AL: "Calculating and comparing codon usage values in rare disease genes highlights codon clustering with disease-and tissue- specific hierarchy", PLOS ONE, vol. 17, no. 3, 31 March 2022 (2022-03-31), page e0265469, XP055982992, DOI: 10.1371/journal.pone.0265469

3 D1 discloses the study of codon usage in muscle disease causing genes. Codon usage was analysed for diseases and non disease coding genes. Codon usage bias was identified and "studying the codon bias phenomenon in genes causing rare diseases may help in identifying critical codons which might play a role in mutational events or gene variation interpretation. Since codon optimization is routinely used to implement translational efficiency of gene therapy constructs, elucidating the codon usage choices might impact therapy designing."

The determination of codon usage bias is explicitly disclosed, the comparison of the codon-usage frequency is implied from the teaching of D1. 19 essential amino acids are not mentioned. Since, however, it is well known in the art that methionine and tryptophan are single codon amino acids, claim 1 is not considered to be inventive over D1 and common general knowledge.

4 Determining the codon usage bias value according to claims 1-9 is considered a method to gain scientific knowledge about codon bias fingerprinting of tissue specific disease and non-disease coding genes.

The case law of the underlying Search authority (G 1/19) underlines that

- "calculated numerical data reflecting the physical behaviour of a system modelled in a computer usually cannot establish the technical character of an invention in accordance with the COMVIK approach, even if the calculated behaviour adequately reflects the behaviour of a real system underlying the simulation (points 127-135)."

and that

"potential technical effects can be relied on only if the claim at least implies the intended technical use of the data (94). This is also not the case for the presently claimed methods. The claims do not specify the use of the results directly affect a technical system or process, e.g. a manufacturing or controlling step based on the simulation results."

5 Claim 10 relates to a computer implemented method for designing a synthetic nucleic acid molecule of a selected disease-causing gene expressed in one or more tissues of an organism, comprising the steps of the methods according to any one of the claims from 1 to 9 and a further step of a codon-optimization of said synthetic nucleic acid molecule based on the codon usage bias value obtained in step iv).

This subject-matter seems to relate to a technical use of the method of the underlying invention and is therefore considered as relating to a intended technical purpose.

However, claim 10 does not clarify how the codon-optimization is performed so that the claim fails to establish a sufficient link between the technical purpose and the mathematical/algorithmic processing steps " a codon-optimization of said synthetic nucleic acid molecule based on the codon usage bias value obtained in step iv)". In particular, there is no *causal link* between the inputs of the method, the mathematical/algorithmic processing steps and the actual provision of a technical effect. The claim is thus not functionally limited to the technical purpose identified above and a technical effect is not *credibly* achieved.

Said identified "gap" could be closed by introducing the features of claims 12 and 13.

6 However, it is not clear whether such a designing process would lead to nucleic acid molecules having a changed function such as codon optimization. Therefore even if introducing the features of claims 12 or 13, it is not clear whether the alleged effect would be achieved over the whole scope. No experimental data has been provided in the underlying application.

7 Independent claims 17 and 18 are not inventive over D1 for the same reasons as claim 1, mutatis mutandis.

8 **Re Item VIII**

Certain observations on the application

8.1 Claims 1, 11 and 14 are not clear.

8.2 The features "disease-causing genes " and "non- disease causing genes" in claim 1 are not clear in that said features are not supported by the description, as its scope is broader than justified by the description. The description under gene selection on p. 29 clarifies that not all disease genes should be analysed and that polygenic or cancer genes are not included. It appears that introducing the features of claims 4 and 5 would overcome said objection.

8.3 The feature "comparing the codon-usage frequency" is not clear in that it is not clear what is compared with what. Step (iv) relates back to step (iii), which als odes not clarify which genes are compared with. A comparison between disease-causing,only or only non-disease causing genes could also be encompassed by said claim formulation.

8.4 Claim 11 has been drafted as dependent on claim 10. Claim 10 is dependent on claim 1. Therefore the features (i) -(iv) in claim 11 are repetitions of the steps (i)-(iv) of claim 1. The aforementioned claim therefore lacks conciseness.

9 Claim 14 includes an in silico screening/deigning step of identifying synthetic nucleic acid molecules (reference to method according to claims 10-13) and a further step of synthesizing a nucleic acid molecule.

This amounts to a combination of two different and irreconcilable types of process claims: the first step of said claim relates to the use of an entity to achieve a technical effect and the further step relates to a process for the production of a product. The second step builds on the effect achieved by the first step, rather than feeding into the second production step a specific starting material and **resulting in a specific product of defined structure**. This results in an unclear claim.

Società Italiana Brevetti
Intellectual Property

Roma, 21 dicembre 2023

Ministero delle Imprese e del Made in Italy
Direzione Generale per la Tutela della
Proprietà Industriale
Ufficio Italiano Brevetti e Marchi
Via Molise 19
00187 Roma RM

Università degli Studi di Ferrara
Domanda di Brevetto per Invenzione n. 102022000006119 depositata il 29 marzo 2022
Ns. Rif.: BI5666R/RPDG/rrm

Spettabile Ufficio,
con la presente si fa riferimento alla domanda n. 102022000006119 depositata il 29 marzo 2022 e alla rispettiva Comunicazione Ministeriale emessa in data 30 novembre 2022 nel corso del procedimento di esame.

Il Rapporto di Ricerca emesso cita cinque documenti di anteriorità, alla luce dei quali L'Esaminatore riconosce il requisito di novità delle rivendicazioni 1-14, 17 e 18, ma solleva obiezioni di mancanza di attività inventiva per le stesse.

Con la presente si fornisce un set di rivendicazioni emendate e argomenti relativi alle obiezioni sollevate nell'Opinione di Brevettabilità (WO) allegata al Rapporto di Ricerca (RdR).

1) Emendamenti alle rivendicazioni e alla descrizione

La rivendicazione 1 è stata emendata introducendo le caratteristiche delle rivendicazioni 4, 11, 12 e 13.
L'introduzione della rivendicazione 4 è atta a specificare che la malattia relativa al gene causante la malattia è una malattia rara, come suggerito dall'Esaminatore al punto 8.2 dell'Opinione di brevettabilità per risolvere l'obiezione di mancanza di

chiarezza riferita al termine "gene causante la malattia" e "gene non causante la malattia".
La rivendicazione 11 come depositata comprende un passaggio v) di progettare la molecola sintetica di acido nucleico di detto gene causante la malattia. Pertanto, il metodo della presente invenzione non solo atto a determinare il valore del bias di utilizzo del codone di un gene causante la malattia selezionato, ma è per progettare una molecola sintetica di acido nucleico di un gene causante malattia selezionato, espresso in uno o più tessuti di un organismo.
Il passaggio iv) della rivendicazione 1 è stato emendato in modo da adeguarlo alla terminologia della rivendicazione 11 come depositata, specificando che detto passaggio iv) è atto ad identificare codoni a comparsa (codoni tessuto-specifici e codoni gene-specifici) in modo da prioritizzare i codoni più diversamente usati nel gene nel/i tessuto/i di interesse.
Le rivendicazioni 12 e 13 sono state introdotte, come suggerito dall'Esaminatore al punto 5 dell'Opinione di Brevettabilità, in modo da rendere chiaro come è eseguita l'ottimizzazione dei codoni e rendendo credibile l'ottenimento dell'effetto tecnico rivendicato, ovvero quello di progettare una molecola di acido nucleico ottimizzata.
Le rivendicazioni 15 e 16, non ricercate dall'Esaminatore, sono state cancellate.
Le rivendicazioni 4 e 10-13 sono state cancellate in modo da evitare ridondanza, e le altre rivendicazioni e dipendenze sono state rinumerate di conseguenza.

2. Osservazioni sulla brevettabilità

2.1 Chiarezza (Art. 51 CPI)

Nell'Opinione di Brevettabilità allegata al Rapporto di Ricerca, l'Esaminatore obietta la mancanza di chiarezza delle rivendicazioni 1, 11 e 14.
In riferimento alla rivendicazione 1, l'Esaminatore afferma che i termini "gene causante la malattia" e "gene non causante la malattia" non sono chiari e non sono supportati dalla descrizione. Come suggerito dall'Esaminatore, il Richiedente ha emendato al rivendicazione 1 introducendo le caratteristiche della rivendicazione 4 come depositata, specificando che la malattia è una malattia rara. L'Esaminatore afferma inoltre che per la caratteristica di paragonare la frequenza di utilizzo dei codoni non è chiaro quale comparazione di esegue. Il Richiedente ha emendato la rivendicazione 1 in modo da comprendere le caratteristiche della rivendicazione 12, esplicitando il modo in cui vengono prioritizzati i codoni.
Per quanto riguarda l'obiezione sollevata per la rivendicazione 11 per mancanza di concisione dovuta alle dipendenze, il Richiedente ha cancellato la rivendicazione e ritiene pertanto che l'obiezione debba essere ritirata.

2.2. Novità (Art. 46 CPI)

Nell'Opinione di Brevettabilità allegata al Rapporto di Ricerca, l'Esaminatore riconosce il requisito di novità per tutte le rivendicazioni ricercate 1-14, 17, 18. Gli emendamenti introdotti nella rivendicazione 1, che trovano tutti basi nelle rivendicazioni e nel testo come depositati, non introducono nuova materia.
Pertanto, le rivendicazioni 1-11 come emendate devono essere anch'esse considerate nuove secondo l'Art. 46 CPI.

2.3 Attività inventiva (Art. 48 CPI)
Nell'Opinione di Brevettabilità allegata al Rapporto di Ricerca, l'Esaminatore ritiene che tutte le rivendicazioni ricercate 1-14, 17, 18 non siano inventive alla luce del documento D1.
In particolare, relativamente alla rivendicazione 1 afferma che l'unica differenza rispetto a D1 risiede nel fatto che non è eseguita in D1 una comparazione della frequenza di utilizzo dei codoni tra 19 amminoacidi essenziali, ma che comunque tale comparazione, seppure non per i 19 amminoacidi menzionati, è implicita dagli insegnamenti di D1.
Inoltre, l'Esaminatore afferma che le rivendicazioni 1-9 come depositate si riferiscono ad un metodo per ottenere conoscenza scientifica, e non rifletterebbe un comportamento tecnico come riportato nella decisione G1/19.
Per quanto riguarda la rivendicazione 10, l'Esaminatore non identifica documenti che descrivono la progettazione di una molecola di acido nucleico come rivendicato, ma afferma che l'effetto tecnico rivendicato non sia credibilmente ottenuto dal momento che la rivendicazione non è funzionalmente limitata allo scopo tecnico di ottimizzazione dei codoni. L'Esaminatore afferma anche che questo "gap" può essere superato introducendo le caratteristiche delle rivendicazioni 12 e 13, ma ritiene che non sarebbe comunque chiaro se tale processo di progettazione porterebbe a molecole di acido nucleico con una funzione modificata come l'ottimizzazione dei codoni.
Infine, l'Esaminatore applica tali obiezioni *mutatis mutandis* alle rivendicazioni 17 e 18.

Sebbene rispettosamente in disaccordo con quanto affermato dall'Esaminatore, il Richiedente ha emendato la rivendicazione 1 introducendo le caratteristiche delle rivendicazioni 4, 11, 12 e 13.
Il documento D1, documento di arte anteriore più vicino nonché l'unico documento su cui l'Esaminatore basa i suoi argomenti, è un abstract in cui compaiono gli stessi inventori della domanda di brevetto in oggetto, e descrive uno studio della frequenza di utilizzo dei codoni nel gene della distrofina (DMD), esaminando l'utilizzo dei codoni in geni causanti malattia muscolare (DC) e geni non causanti malattia muscolare (NDC) attraverso specie diverse.

La rivendicazione 1 come emendata recita:

"1. *Un metodo implementato al computer per progettare una molecola sintetica di acido nucleico di un gene causante malattia selezionato, espresso in uno o più tessuti di un organismo, in cui detta malattia è una malattia rara, comprendente i passaggi:*

(i) raccogliere le sequenze di uno o più geni causanti malattia espressi in uno o più tessuti di un organismo, preferibilmente calcolando la conservazione del codone del gene causante la malattia selezionato tra i mammiferi;

(ii) raccogliere le sequenze di una pluralità di geni non causanti malattia espressi nello stesso tessuto e organismo dei geni nel passaggio (i);

iii) determinare il calcolo indipendente della frequenza di utilizzo del codone per i 19 aminoacidi essenziali (metionina e triptofano esclusi) in ogni gene raccolto nel passaggio (i) e nel passaggio (ii);

iv) paragonare la frequenza di utilizzo del codone determinata nel passaggio (iii) in modo da ottenere il valore del bias di utilizzo del codone, per identificare codoni a comparsa (codoni tessuto-specifici e codoni gene-specifici) in modo da prioritizzare i codoni più diversamente usati nel gene e nel/i tessuto/i di interesse;

v) progettare la molecola sintetica di acido nucleico di detto gene causante malattia, modificando la struttura secondaria o terziaria di detto gene utilizzando i codoni prioritizzati ottenuti nel passaggio iv),

in cui detto passaggio iv) di prioritizzare i codoni usati più differentemente è svolto attraverso il raggruppamento dei valori di frequenza di utilizzo del codone ottenuti nel passaggio iii) utilizzando un algoritmo di raggruppamento gerarchico con un valore p minore di 0.05 e selezionando i codoni meno usati e/o più usati per le specie, il tessuto e il gene di interesse, e
in cui detto passaggio v) di modificare la struttura secondaria o terziaria di detto gene causante malattia è eseguito tramite la sostituzione e/o rimozione dei codoni meno utilizzati e i codoni più utilizzati, ottenuti nel passaggio iv), nella sequenza di detto gene causante malattia."

Pertanto, non solo la rivendicazione 1 differisce da D1 nel fatto che non è esplicitamente descritta la comparazione tra le frequenze di utilizzo dei codoni per 19 amminoacidi, ma anche nel fatto che il passaggio iv) è atto alla prioritizzazione dei codoni più diversamente usati nel gene e nel/i tessuto/i di interesse. Inoltre, non è descritto in D1 il tipo di prioritizzazione (raggruppamento gerarchico con un valore p minore di 0.05 e selezione di codoni meno usati e/o più usati per le specie, tessuto e

gene di interesse), né tantomeno l'ottimizzazione della sequenza del gene causante la malattia mediante sostituzione e/o rimozione dei codoni meno utilizzati e più utilizzati identificati nel passaggio iv).

Le differenze sopra elencate permettono di progettare una molecola sintetica di acido nucleico ottimizzata di un gene causante la malattia, risultato che non è stato in alcun modo descritto o ottenuto nei documenti di arte anteriore citati dall'Esaminatore.

Il problema tecnico oggettivo affrontato dalla presente invenzione è quindi quello di fornire un metodo per la progettazione di una molecola sintetica di acido nucleico per un gene causante malattia rara, espressa in uno o più tessuti di un organismo.

La soluzione proposta dalla rivendicazione 1 emendata introduce differenze sostanziali rispetto a D1. Mentre D1 si limita a studiare la frequenza di utilizzo dei codoni nei geni causanti malattia e non causanti malattia, la rivendicazione 1 emendata propone un metodo specifico per la progettazione di una molecola sintetica di acido nucleico basato sulla frequenza di utilizzo dei codoni per 19 amminoacidi essenziali e sulla loro prioritizzazione e ottimizzazione.

La specificità del passaggio iv), che coinvolge un algoritmo di raggruppamento gerarchico con un valore p minore di 0.05 e la selezione dei codoni meno usati e/o più usati per specie, tessuto e gene, distingue chiaramente l'invenzione da D1. Inoltre, l'ottimizzazione della sequenza del gene causante la malattia mediante sostituzione e/o rimozione dei codoni meno utilizzati e più utilizzati è un aspetto unico introdotto nella presente invenzione.

La combinazione di queste caratteristiche non è suggerita o implicita in D1, e la loro introduzione risolve, con l'impiego di attività inventiva, il problema tecnico oggettivo di progettare una molecola sintetica di acido nucleico per un gene causante malattia rara.

Inoltre, al contrario di quanto affermato dall'Esaminatore al punto 6 dell'Opinione di brevettabilità, il Richiedente ritiene rispettosamente che sia chiaro che la molecola di acido nucleico progettata mediante il metodo della presente invenzione sarà una molecola con codoni ottimizzati, e questo è realizzato in tutto lo scopo rivendicato, dal momento che il metodo comprende la prioritizzazione dei codoni usati più differentemente attraverso un raggruppamento gerarchico, e comprende la modifica della struttura terziaria dell'acido nucleico mediante sostituzione e/o rimozione di codoni meno utilizzati e più utilizzati. Come riportato a pagina 42 del testo come depositato, le impronte CUB (Codon Usage Bias) nei geni causanti malattie possono influenzare l'ottimizzazione dei codoni. In particolare, i geni causanti malattie,

soprattutto quelli nel muscolo che sono stati studiati, mostrano una fingerprint con basso CUB. Ciò significa che i geni muscolari causanti malattia continuano a utilizzare pienamente la ridondanza dei codoni. Al contrario, i geni non causanti malattie mostrano un CUB spontaneamente alto, il che porta a molti "zero-codoni" nelle sequenze codificanti, perdendo così molti tipi di codoni sinonimi.
In sintesi, i geni causanti malattie mostrano uno specifico modello di impronta CUB, il che suggerisce che richiedono un'ottimizzazione dei codoni specifica, sulla base della prioritizzazione descritta nel passaggio iv) della rivendicazione 1.

Pertanto, il Richiedente ritiene rispettosamente che la molecola di acido nucleico progettata mediante il metodo della presente invenzione sarà una molecola con codoni ottimizzati, e che questo è realizzato in tutto lo scopo rivendicato.

Alla luce di quanto sopra, il Richiedente ritiene che il set di rivendicazioni come emendato soddisfi il requisito di attività inventiva secondo l'Art. 48 CPI.

3. Conclusioni

Sulla base di quanto sopra esposto, si richiede a questo Spettabile Ufficio di voler riconoscere che l'intero set di rivendicazioni come emendate soddisfi i requisiti di brevettabilità degli articoli 46, 48, 49, CPI e di voler pertanto procedere al rilascio del brevetto.
Nel caso in cui codesto Ufficio non intenda procedere al rilascio e dovesse emettere un rilievo ai sensi dell'Art. 173(1) CPI, si richiede l'emissione di una comunicazione Ministeriale specifica per il caso in oggetto che illustri adeguatamente, in Italiano, le motivazioni alla base della convinzione avversa dell'Ufficio medesimo. Si richiede, inoltre, di avere un'altra possibilità di depositare argomentazioni e/o emendamenti ai sensi dell'Art. 172(2), 172(3) o 173(7) CPI.

Con osservanza.

Società Italiana Brevetti
Roma, 21 dicembre 2023
Ministero delle Imprese e del Made in Italy

SOCIETÀ ITALIANA BREVETTI

Paolo di Giovine

All. Rivendicazioni emendate revisioni visibili, Rivendicazioni emendate copia pulita

RIASSUNTO

La presente invenzione è legata a metodi che possono essere usati per identificare codoni critici e progettare molecule sintetiche di acido nucleico di geni causanti malattia. Le sequenze sintetiche di acido nucleico possono essere progettate da una sequenza di acido nucleico di riferimento, per esempio, per ottimizzare l'espressione eterologa della sequenza di acido nucleico in un particolare tessuto di un organism ospitante. Alternativamente, sequenze sintetiche di acido nucleico possono essere progettate de novo per codificare un polypeptide desiderato. dette sequenze sintetiche di acido nucleico possono esser eusate per esempio in terapie geniche o alter applicazioni terapeutiche.

Metodo per implementare la progettazione di molecole di acido nucleico sintetico per terapie geniche in malattie rare.

DESCRIZIONE

Ambito tecnico

La presente invenzione si riferisce a metodi usati per identificare codoni critici che possono essere necessari nella progettazione di effettive e appropriate molecole di acido nucleico sintetico con capacità di traslazione applicando il calcolo CUB in specie, tessuti e geni specifici della malattia. Le molecole ottimizzate di acido nucleico sintetico sono pensate per essere usate per la terapia genica o altre applicazioni terapeutiche, oppure per una produzione in larga scala di proteine, reagenti per la ricerca ricombinante e strumenti molecolari in un organismo ospitante. La presente invenzione include ma non è limitata a nuovi parametri, definiti dagli Autori, trattandosi di specie, tessuti e geni di interesse CUB trend, che possono essere usati per l'implementazione e l'ottimizzazione degli algoritmi già esistenti e dei software per la progettazione di molecole sintetiche.

Stato dell'arte

Il codice genetico consiste di 64 triplette di codoni che codificano 20 aminoacidi e tre codoni d'arresto, questi ultimi sono stati riconosciuti dal macchinario transazionale come interruttori della sintesi di proteine. Con l'eccezione di due aminoacidi, triptofano e metionina, che sono codificati da un unico codone, tutti gli altri aminoacidi riconoscono codoni sinonimi multipli basati su due, tre, quattro o sei ridondanze di tripletta, un fenomeno noto come degenerazione del codone. C'è l'intrigante evidenza che la ridondanza del codice genetico abbia giocato un ruolo evoluzionario cruciale nel permettere alla sintesi di proteine di trasformare il mondo RNA nel mondo delle proteine. Per ragioni non completamente comprese, alcuni codoni diventano poco utilizzati, un fenomeno noto come bias di utilizzo del codone (CUB), oppure tende a sparire (extreme CUB) durante l'evoluzione. Nonostante il CUB sia stato ampiamente studiato in varie categorie di geni (ontologia genetica o mappe interactome) e attraverso le specie, il suo

valore evolutivo è ancora incerto. L'originale teoria neutra sull'evoluzione molecolare potrebbe non applicarsi alla selezione di codoni e la pressione mutazionale e la selezione naturale potrebbero aver giocato un ruolo maggiore nel contribuire all'uso dei codoni. Infatti, nonostante solo in pochi casi come per la cheratina e qualche gene ribosomiale o mitocondriale, è stato identificato un extreme CUB negli umani e negli scimpanzé paragonati ad altri mammiferi, suggerire il suo valore evolutivo nei processi gioca un ruolo esclusivo in una linea eucariota specifica.

Esiste un largo consenso sul concetto che la scelta di un codone sinonimo influisce sull'efficienza traslazionale della proteina, sul livello di espressione sulla struttura, e sulla funzione, una nozione che ha suggerito la designazione di codoni ottimali e codone ottimizzazione, che è un processo di routine utilizzato in biologia sintetica per aumentare l'espressione delle proteine. Tuttavia, c'è poco consenso tra i vari algoritmi di ottimizzazione dei codoni, e le metriche correntemente usate potrebbero non essere appropriate per tutti i geni.

Il valore delle variazioni sinonime nel genoma umano e il loro effetto sulle malattie ereditarie e largamente sconosciuto. Interpretare il loro impatto funzionale sui geni è difficile, se non impossibile, senza saggi dedicati funzionali.

Gli strumenti in silicio sono al momento usati per decifrare variazioni sinonime, ma sono inaccurati. Per di più, cambiamenti sinonimi sono completamente ignorati e filtrati da output di dati genomici, un fatto che causa l'omissione degli stessi nella scoperta e validazione della variazione patogena e la mancanza di potenziali nuovi geni di malattia o l'identificazione di genotipi patogeni.

In termini di energia, un CUB estremo viene predetto con il mantenimento basso della richiesta di energia per la traslazione delle proteine, stando al principio massimo dell'entropia, che può portare al progressivo aumento di CUB durante l'evoluzione e attraverso le specie. Questo trend evidenzia certi percorsi funzionali che possono essere stati di priorità energetica (supponendo bias verso codoni preferiti) via selezione naturale e possono essersi verificati in famiglie di geni con particolare rilevanza in una data linea. Per esempio, in genomi con alto contenuto GC (come l'Homo sapiens, HSA), che possono innescare

cambiamenti SNP, l'extreme CUB si verifica frequentemente e si pensa che riduca il rischio del verificarsi di variazioni nonsenso.

Il ruolo dei codoni "rari" (extremely biased) attraverso l'evoluzione è ancora controverso. Nei batteri, i codoni rari vicino all'estremità 5' facilitano la rimozione della repressione della traslazione e sono considerati "rampe autostradali" per innescare e accelerare la traslazione di proteine di più di 60 volte, con un ruolo chiave nella regolazione del traffico ribosomiale. Sinergicamente, le proprietà di ripiegamento del mRNA ricco di codoni rari vicino all'estremità 5 incrementa la velocità di traslazione, come nelle cellule che si dividono rapidamente. Al contrario, alcuni codoni usati frequentemente hanno l'effetto opposto, rallentando l'efficienza della traslazione. Alla luce di quanto detto, c'è ancora necessità di provvedere metodo e modi di migliorare il bias di utilizzo dei codoni e le sue applicazioni.

Sommario dell'invenzione

La presente invenzione è basata sull'inattesa scoperta che l'approccio basato sulla malattia genica aiuta a identificare codoni critici, che possono giocare un ruolo in eventi di mutazione genica, interpretazione di variazioni sinonime e progettazione dell'algoritmo per l'ottimizzazione dei codoni.

Gli autori della presente invenzione, come divulgato in dettaglio nella sezione sperimentale della presente applicazione, hanno comparato innovativamente i valori di utilizzo dei codoni tra geni causanti-malattia (DC) e non-causanti-malattia (NDC) come espresso in specifici tessuti, e attraverso mammiferi, perciò utilizzare un approccio basato sulla malattia, per esplorare i valori CU e il comportamento bias di utilizzo dei codoni (CUB), per poter esplorare l'influenza di questi fattori su valori CU.

Pertanto, gli Autori della presente invenzione per la prima volta hanno identificato 3 nuove metriche utili per calcolare il CUB in geni specifici causanti rare malattie per poter ridefinire e integrare i parametri usati per la progettazione di geni sintetici:

1. Per poter ottenere un gene CUB, è necessario un calcolo tra le specie (mammiferi) per controllare la conservazione del codone in quel gene

specifico. Questo trend di conservazione varia a seconda del tipo di gene. Questa analisi deve essere fatta confrontando la conservazione del codone di quel gene specifico tra i mammiferi.

2. Per conferire un robusto significato statistico, gruppi di geni che causano le stesse malattie legate ai tessuti sono confrontati tra i mammiferi, raccogliendo le sequenze di una pluralità di geni non causanti malattia esclusivamente o preferenzialmente espresse nello stesso tessuto e organismo come un controllo di gruppo;
3. Comparare i valori CU tra i geni di gruppo di controllo (non causanti malattie) e gruppi di geni causanti malattie rare.

La presente invenzione offre anche un metodo che ridefinisce i criteri attualmente usati nella preparazione di prodotti di geni sintetici che possono essere proteine, reagenti ricombinanti, e strumenti molecolari che in fase di deposito della presente invenzione erano in ordine: riposizionare i codoni con l'incremento di quelli ricchi in GC (guanina e citosina) e ridurre quelli ricchi in AT (adenina, timina) contenuto delle 3 metriche di cui sopra; correggere la struttura dell'mRNA secondario e terziario riposizionando alcuni codoni considerati perturbanti la disponibilità delle molecole di mRNA per la traslazione.

Un primo oggetto della presente invenzione è un metodo implementato per computer per la determinazione del valore del bias di uso del codone di un selezionato gene causante malattia in uno o più tessuti di un organismo.

Un ulteriore oggetto della presente invenzione è un metodo implementato per computer per progettare una molecola sintetica di acido nucleico di un selezionato gene causante malattia espressa in uno o più tessuti di un organismo.

Un ulteriore oggetto della presente invenzione è un metodo per preparare una molecola sintetica di acido nucleico di un selezionato gene causante malattia.

Un ulteriore oggetto della presente invenzione è una proteina, un reagente ricombinante o uno strumento molecolare comprendente la sequenza di acido nucleico ottimizzata per il codone ottenibile dall'eseguimento dei metodi della presente invenzione.

Un ulteriore oggetto della presente invenzione è l'uso di codoni prioritarizzati come determinato nei metodi della presente invenzione, in un metodo per la valutazione dello sviluppo di una malattia correlata alla mutazione di un selezionato gene.

Un ulteriore oggetto della presente invenzione è un programma per computer comprendente le istruzioni che, quando il programma è eseguito da un computer, comporta che il computer svolga passaggioi passaggi dei metodi quivi divulgati.

Vantaggi aggiuntivi e/o forme di realizzazione della presente invenzione saranno evidenti dalla successiva descrizione dettagliata.

Breve descrizione dei disegni

La presente invenzione e la seguente descrizione dettagliata delle preferibili realizzazioni della stessa possono essere meglio comprese con riferimento alle seguenti figure:

Figura 1. Valori CU nei geni del muscolo, della pelle e dei reni nell'*Homo sapiens.* Diagrammi di calore sono stati generati usando il pacchetto R gplots. I file sono stati raggruppati basandosi sulla distanza euclidea. La codifica dei colori varia dal blu scuro al rosso con valori CU da bassi ad alti rispettivamente. I raggruppamenti gerarchici di uso del codone sinonimo in tutti i geni studiati in differenti tessuti (muscolo, pelle, rene) sono stati generati in HSA. Il grafico del calore mostra che il raggruppamento di codoni usato frequentemente (in rosso) e dei codoni usati raramente (blu scuro) varia enormemente tra geni e tessuti. Nei geni muscolari, i codoni estremamente degeneri (valori di CU bassi, colore chiave blu scuro, o valori CU alti, colore chiave rosso scuro) sono raggruppati strettamente in termini di genere e di tipo di codone, mentre valori intermedi di CU (colore chiave azzurro o giallo) sono più dispersi negli alberi (pannello A). Tra i geni muscolari, solo il DMD non mostra CUB estremo, dal momento che non compaiono punti rossi (corrispondenti a valori più alti di CU) (figura 1A). Le impronte CUB dei geni della pelle (pannello B) mostra una prevalenza di valori CU bassi di 726 codoni (blu scuro), con pochi dispersi, non raggruppati, punti giallo-rossi distribuiti non omogeneamente (valori di CU alti e intermedi). I raggruppamenti CU sono meno definiti se comparati al muscolo. Nei geni dei reni (pannello C), le impronte CUB differiscono dagli altri due gruppi di geni. La

stragrande maggioranza dei geni hanno valori CU bassi o intermedi (raggruppamenti di punti giallo-blu ampi e diffusi) con un raggruppamento di alti valori di CU strettamente raggruppati e legati a UMOD, BSND, SLC22A8, MIOX, AQP6, PKD1, SLC12A3 e geni GGACT. Interessante in fatto che tutti questi geni sono DC.

Figura 2. Valori di CU tra i mammiferi. I diagrammi di calore sono stati generati usando il pacchetto R gplots. I file sono stati raggruppati basandosi sulla distanza euclidea. Il colore chiave dei valori CU varia dal blu al rosso con valori del CUB rispettivamente bassi e alti. I valori CU in tutti i 60 geni tessuto-specifici tra 15 specie mammifere nell'albero filogenetico dei metazoi sono mostrati nei pannelli A, B, e C, e i valori CU dei singoli geni in tutte le specie sono mostrati nei pannelli D, E, ed F. Impronte CUB tesuto-specifiche sono molto evidenti con un trend conservato tra i mammiferi (pannelli A, B, C). I codoni CAG, AAG, CAC, GAC, GAG, AUC, AAC, UAC, UGC, e UUC sono quelli usati più frequentemente (punti rossi) in tuti i tessuti e tra i mammiferi, mentre i codoni UUA, CUA, UCG, CGU, CUU, GUA, CGA, AUA, UCA, UUG e GCG sono i più rari (punti blu) nei geni muscolari e della pelle, ma non dei reni. I CUB del muscolo (pannello A) e della pelle (pannello B) hanno un’impronta simile, sebbene i geni muscolari hanno molti più codoni con valori CU più bassi paragonati ai geni della pelle; i geni dei reni mostrano un raggruppamento molto diverso.

I pannelli D, E, ed F mostrano valori CU del gene che raggruppa tutti i mammiferi. Benché un chiaro raggruppamento è scarsamente visibile, i tessuti hanno differenti impronte CUB caratterizzate da diversi dendrogrammi legati al gene. Infatti, 13/30 geni nel muscolo (pannello D), 4/40 geni nella pelle (pannello E) e 8/20 geni nei reni (pannello F) mostrano raggruppamenti di alti valori CU. I tipi dei codoni variano anche di conseguenza a valori CUdal momento che i valori alti di CU sono raggruppati nei gruppi di geni elencati di cui sopra. Questa scoperta sostiene che i valori CU hanno un’impronta legata al tessuto che è ancora mantenuta e che raggruppa tutti i mammiferi, e che alcuni valori CU gene-specifici e tipi di codoni sono osservabili.

Figura 3. Comparazione di valori relativi (CUB) di uso di codoni sinonimi in geni causanti malattia (DC) e non causanti malattia (NDC) tra le specie.

I diagrammi di calore sono stati generati usando il pacchetto R gplots. I file sono stati raggruppati in base alla distanza euclidea. Il colore chiave varia dal blu al rosso con valori del CUB rispettivamente da bassi ad alti. Abbiamo raggruppato i geni basandoci sulla loro propensione ad essere il sito di variazioni patogene (mutazioni) causanti rare malattie (geni causanti malattia o DC). I pannelli da A ad F mostrano tutti i tipi di codoni senza raggruppamento gerarchico, elencati nello stesso ordine in tutti i geni e tessuti, e sotto-raggruppati in DC e NDC. I valori CU in questi pannelli mostrano che il più frequente o il più raro tipo di codone sono sovrapponibili nei gruppi di geni.

Due gruppi di codoni di colore chiave blu scuro e rosso scuro si verificano in tutti i 6 pannelli, indicando che l'assoluta frequenza del tipo di codone è simile anche tra geni e mammiferi, possibilmente supportando una tendenza evoluzionaria del CUB. Nondimeno, la variabilità valori più alti di CU può essere visibile in codoni con frequenza intermedia (colore chiave da azzurro a giallo) che può differenziarsi tra gruppi di geni. I geni muscolari NDC (pannello B) hanno un numero più basso di valori intermedi di CU (colore chiave giallo), seguiti dai geni della pelle DC (pannello C) e i geni dei reni (pannello F). CAG è il codone usato più frequentemente in tutti i geni e tra i mammiferi. I pannelli G-L mostrano il raggruppamento gerarchico di valori di CU nelle stesse categorie di gruppi. Impronte CUB chiaramente riconoscibili possono essere osservate nei geni NDC e DC. Questo è maggiormente evidente nei geni muscolari (pannelli G e H) e parzialmente nei geni dei reni (pannelli K e L). I geni muscolari DC hanno raggruppamenti compatti di codoni estremamente frequenti (AAG, CAG, GAG) e codoni estremamente rari (UGG, UUA, CUA), una tendenza conservata tra i mammiferi e con 774 gruppi di codoni (in termini di distanza dall'albero). I colori blu scuro e rosso (codoni più rari e frequenti) predominano in geni muscolari NDC, con pochi codoni con valori intermedi (colore chiave giallo). Questo suggerisce che un forte CUB si è verificato nei geni NDC. Di conseguenza, i geni muscolari NDC mostrano valori di CU più alti e più bassi raggruppati insieme, suggerendo una possibile tendenza evoluzionistica diversa. I geni della pelle NDC e DC (pannelli I-J) mostrano simili impronte CUB con poche differenze. I geni della pelle NDC mostrano nel complesso valori di CU più bassi (pannello J,

parte superiore) e viceversa più codoni con valori di CU intermedi (pannello J, parte inferiore) paragonati ai geni DC. Le impronte CUB dei geni del rene DC e NDC (pannelli K-L) differiscono grandemente dagli altri due gruppi dal momento che i raggruppamenti gerarchici sono opposti. Anche se la conservazione di valori di CU si verifica anche tra i mammiferi, la gerarchia dei dendrogrammi del gene del rene mostra un antenato comune per i valori di CU intermedi e bassi e non due distinte linee (valori di CU alti e bassi) come osservabile nei geni muscolari e della pelle. Le due impronte CUB del rene sono in qualche modo simili; comunque, i geni DC mostrano un più alto numero di valori CU molto bassi.

Figura 4. I 5 codoni estremamente degeneri, usati differentemente dall'*Homo sapiens* in geni DC vs NDC, e tra i tessuti. I grafici sulla sinistra (A, B, C, D, E) mostrano la frequenza di codoni sull'asse delle x e il numero dei geni che usano quello specifico codone (con la frequenza annotata nell'asse delle x) sull'asse delle y. I grafici sulla destra (F, G, H, I, L) mostrano i valori di CU in tutti i mammiferi studiati sull'asse delle x e i valori di uso dei codoni nei tessuti genetici sull'asse delle y. Cinque codoni erano usati più differentemente nell'*HSA*, CGU (Arg), CCA (Pro), GAC (Asp), GAU (Asp), e GUA (Val). Le barre rosa indicano i geni DC, le barre blu i geni NDC. CGU (A, F) è il codone usato meno di frequente nei geni muscolari NDC; CCA (B, G) è il codone usato meno di frequente nei geni muscolari DC; GAC (C, H) è il codone usato più di frequente nei geni della pelle DC; GAU (D, I) è il 798° codone usato più frequentemente nei geni della pelle NDC; GUA (E, L) è il codone usato meno di frequente nei geni del rene DC. Comparando i valori di CU di questi 5 codoni estremamente degeneri tra gruppi di geni DC e NDC e tra i mammiferi, può essere apprezzata la tendenza verso una pesante estremizzazione dei codoni durante l'evoluzione. Durante l'evoluzione, CGU e CCA hanno iniziato ad essere usati di più nei geni muscolari DC, GAC ha iniziato ad essere più utilizzato nei geni della pelle DC, GAU ha iniziato ad essere più utilizzato nei geni della pelle NDC, e GUA ha iniziato ad essere più utilizzato nei geni del rene NDC. Questo suggerisce uno specifico codone, orientato alla malattia CUB, apparentemente conservato tra i mammiferi.

Figura 5. Uso del codone del gene DMD negli umani e approccio "mapping-on-codon". Pannello A valori di uso del codone DMD umano e

percentuali di mutazione. Le barre rappresentano i valori CU nei geni *DMD* nell'*HSA*. Sull'asse delle x sono elencati i tipi di codone e i relativi aminoacidi, sull'asse delle y sono riportati i valori di CU. Le barre rosse rappresentano i 4 codoni *DMD* usati raramente, UCG, CCG, ACG E GCG, in base al nostro taglio, che è basato sulla ridondanza del codone di 2, 3, 4 e 6 triplette (vedi Metodi).

In cima alle barre (lato destro), sono riportati i numeri di mutazioni missenso e nonsenso verificatesi ai relativi codoni *DMD*. Tutti i codoni sono ancora usati dal gene umano *DMD*, e la frequenza del verificarsi della mutazione non è collegata ai valori di uso del codone. Esempi sono UAU, che è il codone più utilizzato, ma con solo 36 mutazioni "mappate", oppure CGA (Arg) che è usata raramente ma ha 114 mutazioni "mappate".

Pannello B. Mappatura delle mutazioni missenso e nonsenso della *DMD* umana sui tipi di codoni. Sull'asse delle x, ci sono i codoni e i relativi aminoacidi, sull'asse delle y, c'è il numero di mutazioni che si sono verificate e "mappate". Secondo il nostro taglio le barre rosse rappresentano i 4 codoni usati raramente nel gene umano *DMD*. In cima alle barre c'è la percentuale di valori CU 822. Il numero delle mutazioni verificatesi e i valori CU non sono strettamente collegati. Esempi sono CAG, che è il codone mutato "più frequentemente" (56%), con alti valori di CU, e UGC (Cys), raramente sito di mutazioni, ma con valori CU molto alti.

Figura 6. L'analisi della correlazione di Spearman tra geni DC e NDC nei tessuti genetici del muscolo, della pelle e del rene di *HSA.* Il test ha dimostrato che i valori CU dei geni DC e NDC si correlano significativamente nel muscolo nella pelle e nel rene ($p<0.05$).

Figura 7. I diagrammi di calore sono stati generati utilizzando il pacchetto R gplots. I file sono stati raggruppati in base al sistema della distanza euclidea. Il codice dei colori varia dal blu scuro al rosso con valori di CU rispettivamente da alti a bassi. L'impronta CUB e i valori di CU tra le espressione di geni *HSA* alto-, medio-, e basso. I geni DC NDC erano considerati dipendenti dal loro livello di espressione. Le impronte CUB hanno una forte similitudine, il che significa che infatti, raggruppare i geni secondo il loro livello di espressione produce una tendenza di valori CU. I codoni AAC, GAC, UGC, UAC, CAC, UUC, AUC, AAG,

GAG e CAG sono usati più frequentemente sia in geni alto-espressi che in geni medio-espressi mentre GUG è presente solo in geni alto-espressi. I codoni GAC, CAC, UUC, CAG, UAC, UGC, AAG, AUC, GAG, AAC, ACC, GGC, GUC, UCC e GCG sono usati più frequentemente in geni espressi di basso livello. Pochi codoni hanno valori di CU più bassi come UCC (Ser), ACC (Thr), GGC (Gly), GUC (Val) e GCG (Ala) in geni basso-espressi. Alcuni geni DC alto-espressi hanno più codoni con valori di CU più alti. Come *DYS, LMNA* e *DES* (muscolo), *UMOD* e *PKD1* (rene) e *FGFR3* (pelle). Nei geni medio-espressi la tendenza è opposta, con alcuni geni NDC che mostrano valori di CU più alti come *MLPF, TNNC2, TMEM3BA* (muscolo) e *NCLZ2, MCX* (rene). È interessante notare che UAA è il codone-stop più utilizzato nei geni alto-espressi, dal momento che induce il termine della traduzione con maggiore velocità e accuratezza a livello ribosomiale e può essere letta sia da fattori di rilascio eRF1 che eRF2. UAG e UGA hanno frequenza simile in tutti i tessuti genetici e livelli di espressione.

Figura 8. Il diagramma di flusso schematico di un metodo secondo il corpo della presente invenzione.

Glossario

L'uso delle forme singolari "un", "uno", "una", "il", "lo", "la" includono riferimenti plurali a meno che il contesto non indichi diversamente. Per esempio, riferimenti a "polinucleotide" include una pluralità di polinucleotidi, riferimenti a "substrato" include una pluralità di detti substrati, riferimenti a "una variante" includono una pluralità di varianti, ecc.

Dove è riportata una gamma di valori, deve essere compreso che ogni valore intero intervenuto, e ogni frazione dello stesso, tra i limiti riportati in alto e in basso di quella serie è inoltre specificamente divulgato, insieme a ogni sottoserie tra detti valori. I limiti superiori e inferiori di ogni serie possono essere indipendentemente inclusi nella serie, o esclusi da essa, e ogni serie in cui uno, nessuno o entrambi i limiti sono inclusi sono anche comprese nell'invenzione. Dove un valore discusso ha limiti inerenti (per esempio, dove un componente può essere presente a una concentrazione da 0 a 100%, o dove il pH di una soluzione

acquosa può variare da 1 a 14), quei limiti inerenti sono specificatamente discussi.

Dove un valore è esplicitamente riportato, deve essere compreso che valori che sono della stessa quantità o ammontare del valore riportato rientrano anche nello scopo dell'invenzione. Dove è riportata una combinazione, ogni sotto-combinazione degli elementi di quella combinazione è specificamente discusso e rientra nello scopo dell'invenzione. Al contrario, dove elementi diversi o gruppi di elementi sono discussi individualmente, combinazioni dello stesso sono anche discusse. Dove qualsiasi elemento di un'invenzione è descritto come avente una pluralità di alternative, esempi di quell'invenzione in cui ogni alternativa è esclusa singolarmente, o in qualsiasi combinazione con le altre alternative, sono discusse con la presente (più di un elemento di un'invenzione può avere queste esclusioni, e tutte le combinazioni di elementi aventi queste esclusioni sono discussi nella presente).

A meno che diversamente previsto, tutti i termini tecnici e scientifici usati qui hanno lo stesso significato come comunemente intese da una delle abilità ordinarie nell'arte della genetica, della bioinformatica e di progettazione genica. Ogni metodo e materiale simile o equivalente a quelli qui descritti può essere usato nella pratica o in fase di test delle realizzazioni dell'invenzione, anche se alcuni metodi e materiali sono esemplificati da quelli qui discussi.,

Il bias di uso del codone: come qui usato, il termine "bias di uso del codone", o semplicemente "uso del codone", si riferisce alle differenze nella frequenza del verificarsi di un particolare codone come opposto ad altri codoni sinonimi, nella codificazione del DNA, per la codificazione di un aminoacido all'interno di un organismo. Un bias di uso del codone può essere espresso come una misurazione quantitativa del tasso al quale un particolare codone è usato nel genoma di un particolare organismo, per esempio, se comparato ad altri codoni che codificano lo stesso aminoacido. Negli oggetti elencati in questo documento sono considerati o prioritizzati codoni la cui frequenza di uso è statisticamente differente paragonata agli altri codoni sinonimi, sia alti che bassi.

Vari metodi sono conosciuti da quelli di competenza nell'arte di determinare l'uso del metodo indice di adattamento del codone (CAI), che è essenzialmente

una misurazione della distanza di uso del codone di un gene all'uso del codone di un set predefinito di geni altamente espressi.

Sharp e Li (1987) Nucleic Acids Res. 15:1281-95. Quindi, il bias di uso del codone include le frequenze relative dell'uso dei codoni che codificano lo stesso aminoacido ("codoni sinonimi"). Un bias può verificarsi naturalmente; per esempio, il bias del codone nel genoma di un organismo riflette l'uso complessivo di codoni sinonimi all'interno di tutti i geni in quell'organismo. Un bias può anche essere usato in un algoritmo computazionale, dove, per esempio, può essere usato per determinare la frequenza relativa con cui codoni sinonimi differenti sono selezionati per essere utilizzati nella progettazione di una sequenza polinucleotidica. Similarmente, la frequenza "relativa" di ogni elemento della sequenza usato per codificare un polipeptide all'interno di una sequenza nucleotidica con la quale quell'elemento della sequenza è utilizzato per codificare una caratteristica del polipeptide, diviso dal numero degli accadimenti all'interno del polipeptide in una data cornice di lettura di caratteristiche cheche possono essere codificate da quell'elemento della sequenza.

Il bias di uso del codone può essere dedotto da una tavola di uso del codone per una particolare espressione di un organismo ospitante. Le tavole di uso dei codoni sono prontamente reperibili per molte delle espressioni di organismi ospiti. Vedi. E.g., Nakamura et al. (2000) Nucleic Acids Res. 28:292 (Database di Uso del Codone- versioni aggiornate disponibili a kazusa.or.jp/codon).

I termini "tavola di uso del codone", oppure "tavola del bias del codone), oppure "tavola della frequenza del codone" sono usate indifferentemente e descrivono una tavola che correla ogni codone che può essere usato per codificare un particolare aminoacido con le frequenze con cui ogni codone è usato per codificare quell'aminoacido in un organismo specifico, all'interno di una specifica classe di geni all'interno di quell'organismo, o all'interno di uno o più polinucleotidi sintetici.

Frequenza assoluta del codone: come qui utilizzato, il termine "frequenza assoluta del codone" si riferisce alla frequenza con cui appare un codone relativa al numero totale di codoni sinonimi all'interno di un polinucleotide o set di polinucleotidi in una data cornice di lettura (e.g., una cornice di lettura che è usata

per codificare un polipeptide di interesse). Similarmente, la frequenza “assoluta” di ogni elemento della sequenza usato per codificare un polipeptide all’interno di un polinucleotide è la frequenza con cui quell’elemento della sequenza è usato per codificare una caratteristica (e.g., aminoacido, coppia di aminoacidi, ecc.) del polipeptide, diviso dal numero di occorrenze all’interno del polipeptide di caratteristiche della stessa misura di quelle che potrebbero essere codificate da quell’elemento della sequenza.

Cornice di lettura aperta: come qui utilizzato, il termine “cornice di lettura aperta” si riferisce a tutte le possibili sequenze di polinucleotidi che possono essere usate per codificare uno specifico polipeptide, attraverso la variazione dei codoni usati per codificare aminoacidi all’interno del polipeptide.

Sostituzione del codone: come qui utilizzato, il termine “sostituzione del codone” si riferisce all’alterazione di una sequenza codificante nucleotidica attraverso il cambiamento di uno o più codoni che codificano uno o più aminoacidi di un polipeptide codificato, senza alterare la sequenza di aminoacidi del polipeptide codificato.

Ottimizzazione del codone: Come qui utilizzato, il temine “ottimizzazione del codone” si riferisce ai processi messi in atto per modificare una sequenza codificatrice esistente, oppure per la progettazione di una sequenza codificatrice in prima istanza, per esempio, per migliorare la traduzione in una cellula o organismo ospite di una molecola di RNA trascritto, trascritta dalla sequenza codificante, o per migliorare la trascrizione di una sequenza codificante. L’ottimizzazione del codone include, ma non è limitata, a processi che includono la selezione di codoni per la sequenza codificante per adattarsi alla preferenza di codone di espressione dell’organismo ospite. L’ottimizzazione del codone include anche, per esempio, il processo a cui a volte ci si riferisce come “armonizzazione del codone”, in cui i codoni di una sequenza di codoni che vengono riconosciuti come codoni a basso-utilizzo nell’organismo di origine sono alterati a codoni che vengono riconosciuti come a basso utilizzo nella nuova espressione ospite. Questo processo può permettere ai polipeptidi espressi di piegarsi normalmente tramite l’introduzione di pause naturali e appropriate durante la traduzione/estensione. Birkholtz et al. (2008) Malaria J. 7:197-217.

Modificare: Come qui utilizzati, i termini “modificare” o “alterare”, o qualsiasi altra forma dello stesso, significano modificare, alterare, riposizionare, cancellare, sostituire, rimuovere, variare, o trasformare.

Molecola di acido nucleico: come qui utilizzato, il termine “molecola di acido nucleico” si riferisce a una forma polimerica di nucleotidi, che può includere filamenti di RNA senso e anti-senso, cDNA, DNA genomico, e forme sintetiche e polimeri misti di cui sopra. Un nucleotide può riferirsi a un ribonucleotide, deossiribonucleotide, o una forma modificata dei due tipi di nucleotide. Una “molecola di acido nucleico” come qui usato è sinonimo con “acido nucleico” e “polinucleotide”. Una molecola di acido nucleico ha generalmente la lunghezza di 10 basi, a meno che diversamente specificato. Il termine include forme a filamento singolo o doppio di DNA. Una molecola di acido nucleico può includere uno o entrambi tra nucleotidi verificatisi naturalmente o modificati, legati insieme da legami nucleotidici verificatisi naturalmente o non-naturalmente.

Proteina/polipeptide: i termini “proteina” e “polipeptide” sono qui usati indifferentemente. I termini si riferiscono alla catena molecolare contigua di aminoacidi legati attraverso legami peptidici. I termini non si riferiscono a una specifica lunghezza del prodotto. Perciò, “peptidi”, “oligopeptidi”, e “proteine” sono inclusi nella definizione di polipeptide. I termini includono polipeptidi contenenti modifiche co- e/o post-traslazionali del polipeptide fatti in vitro o in vivo; per esempio e senza limitazione: glicosilazione, acetilazione, fosforilazione, PEGilazione e solfatazione. Inoltre, frammenti di proteine, analoghi (compresi gli aminoacidi non codificati nel codice genetico: e.g., omocisteina, ornitina, p-acetilfenilalanina, D-amminoacidi e creatina), mutanti naturali o artificiali, varianti, proteine di fusione, residui derivatizzati (per esempio, alchilazione di gruppi amminici, acetilazione o esterificazione di gruppi carbossilici), e combinazioni di qualsiasi di cui sopra sono inclusi nel significato di polipeptide.

Come qui utilizzato, il termine “percentuale di identità di sequenza” può riferirsi al valore determinato dall’allineamento di due o più sequenze (e.g., sequenze di acido nucleico e sequenze di aminoacidi) su una finestra di confronto, in cui la porzione della sequenza nella finestra di confronto può comprendere addizioni o rimozioni) per l’allineamento ottimale delle due sequenze. La percentuale è

calcolata dalla determinazione del numero di posizioni alle quali il nucleotide identico o il residuo aminoacido si verifica in entrambe le sequenze per produrre il numero di posizioni abbinate, dividendo il numero delle posizioni abbinate per il numero totale di posizioni nella finestra di confronto, e moltiplicando il risultato per 100 per produrre la percentuale di identità di sequenza. Metodi per l'allineamento delle sequenze per il confronto sono ben conosciute nel campo.

Sintetico: come qui utilizzato in riferimento a una sequenza nucleotidica (o molecola di acido nucleico comprendente una sequenza nucleotidica sintetica), il termine "sintetico" si riferisce a una sequenza che è progettata (e.g., in silicio), per esempio, con lo scopo si esprimere un polipeptide codificato di interesse. Il termine "nucleotide sintetico" include anche il prodotto della manifattura di una molecola di acido nucleico per mezzo di oligonucleotidi sintetizzati chimicamente attraverso metodologie in vitro o in vivo conosciute da coloro i quali sono esperti nella sintesi di geni, o attraverso combinazioni di metodi in vitro o in vivo.

Descrizione dettagliata

Nella seguente, saranno descritte diverse realizzazioni dell'invenzione. È inteso che le caratteristiche delle varie realizzazioni possono essere combinate, laddove è compatibile. In generale, realizzazioni susseguenti saranno discusse solo in relazione alle differenze con quelle precedentemente descritte.

Come precedentemente menzionato, un primo oggetto della presente invenzione è rappresentato da un metodo implementato al computer per determinare il valore del bias di uso del codone di un selezionato gene causante malattia.

In una forma di realizzazione questo metodo comprende i seguenti passaggi:

(i) raccogliere le sequenze di uno o più geni causanti-malattia espressi in uno o più tessuti di un organismo;

(ii) raccogliere le sequenze di una pluralità di geni non-causanti-malattia espressi nello stesso tessuto e organismo dei geni nel passaggiopassaggio (i);

(iii) determinare il calcolo indipendente della frequenza di uso del codone per i 19 aminoacidi essenziali (metionina e triptofano esclusi) in ogni gene raccolto negli passaggio (i) e (ii);

(iv) comparare la frequenza di uso del codone determinata nello passaggio (iii) in modo da ottenere il valore del bias di utilizzo del codone.

Preferibilmente, il metodo della presente invenzione comprende un ulteriore passaggio in cui, per poter ottenere un gene CUB, è necessario un calcolo tra le specie (mammiferi) per controllare la conservazione del codone in quel gene specifico, laddove il gene specifico è il gene selezionato causante-malattia. Questa tendenza di conservazione varia a seconda del tipo di gene. Questa analisi deve essere svolta comparando la conservazione del codone di quel gene specifico tra i mammiferi. Detto calcolo tra le specie è utilizzato nel metodo per ottenere il valore del bias di uso del codone.

In una forma di realizzazione dell'invenzione detto organismo è un mammifero. In un'altra realizzazione dell'invenzione, detti geni causanti-malattia sono raccolti da diversi mammiferi differenti. Preferibilmente, detti mammiferi sono selezionati da *R. ferrumequinum* (pipistrello ferro di cavallo maggiore), *M. musculus* (topo), *F. catus* (gatto), *C. lupus familiaris* (cane), *E. caballus* (cavallo), *B. taurus* (bovino), *M. murinus* (lemure topo grigio), *G. variegatus* (colugo della Sonda), *C. jacchus* (uistitì comune), *M. mulatta* (macaco), *N. leucogenys* (gibbone), *P. abelii* (orangotango), *G. gorilla* (gorilla), *P. troglodytes* (scimpanzé) e *H. sapiens* (umano).

In una ulteriore forma di realizzazione questo metodo è un metodo implementato per computer per la progettazione di una molecola sintetica di acido nucleico di un selezionato gene causante-malattia, che comprende i seguenti passaggi:

(i) raccogliere le sequenze di uno o più geni causanti-malattia esclusivamente o preferibilmente espressi in uno o più tessuti di un organismo;

(ii) raccogliere le sequenze di una pluralità di geni non-causanti-malattia esclusivamente o preferibilmente espressi nello stesso tessuto ed organismo dei geni nello passaggio (i);

(iii) determinare il calcolo indipendente della frequenza di utilizzo del codone per i 19 aminoacidi essenziali (metionina e triptofano esclusi) in ogni gene raccolto negli passaggio (i) e (ii);

(iv) comparare la frequenza di utilizzo del codone determinata nello passaggio (iii) quindi ottenendo il valore del bias di utilizzo del codone, per identificare codoni a comparsa (codoni tessuto-specifici e codoni gene-specifici) per poter dare priorità ai codoni più usati in modo diverso nel gene e nel/nei tessuto/i di interesse;

(v) progettare una molecola di acido nucleico di detto gene causante-malattia, modificando la struttura secondaria o terziaria di detto gene utilizzando i codoni prioritizzati nello passaggio (iv).

Secondo il punteggio delle banche dati pubbliche nel descrivere o valutare l'espressione di un gene in un tessuto, “esclusivamente” è usato quando un gene è espresso solo in quel particolare tessuto, mentre “altamente” è usato quando un gene è espresso anche in altri tessuti ma (risulta) sostanzialmente inferiore rispetto al tessuto considerato, “preferenzialmente” se il gene è espresso in altri tessuti ma il tessuto considerato mostra un’espressione più alta. Detti tessuti sono per esempio il muscolo, la pelle, il rene o qualsiasi altro tessuto coinvolto in malattie suscettibili con terapia genica.

In una forma di realizzazione, detta malattia è selezionata da distrofie muscolari, miopatie congenite, malattia tubulointestinale del rene, rene policistico di tipo 1, ipercheratosi epidermolitica, displasia ectodermica.

In alcune forme di realizzazione dette malattie sono tipi di malattia rara (RD), per esempio legate al muscolo, alla pelle, al rene o altre mutazioni del gene dei tessuti. Malattie rare del muscolo includono per esempio distrofie muscolari e miopatie congenite, i cui geni causanti sono espressi in maniera predominante e alta nei muscoli scheletrici. Da cui, lo scopo del metodo è l’ottimizzazione della sequenza nucleotidica di un gene selezionato espresso in uno o più tessuti, che è coinvolto in una particolare malattia, in particolare tutte le malattie suscettibili con terapia genica.

La pluralità di geni nei passaggi (i) e/o (ii) significa un gruppo di geni dei quali si conosce il coinvolgimento o meno nella malattia del gene selezionato e che sono espressi, preferibilmente in maniera esclusiva, altamente o preferenzialmente espressi, nello stesso tessuto dello stesso organismo del gene selezionato. Preferibilmente, il numero di geni raccolti nei passaggi (i) e/o (ii) è il più alto numero disponibile nella banca dati per la malattia selezionata, oppure per esempio dal 70 al 99%, preferibilmente dall'80 al 99%.

Il bias di utilizzo del codone può essere determinato secondo la procedura di calcolo conosciuta nel campo, per esempio può essere determinata dal metodo di indice di adattamento del codone (CAI), che è essenzialmente una misurazione della distanza dell'utilizzo del codone di un gene all'utilizzo del codone di un set predefinito di geni altamente espressi, usando in questo caso come set di dati di riferimento i geni causanti-malattia e/o geni non-causanti-malattia, il CAI sarà in questo caso modificato introducendo una correzione basata sul set di dati di riferimento usati e piò essere determinato un bias di uso del codone migliorato. Eppure, secondo l'invenzione il CAI sarà modificato utilizzando come set di dati di riferimento predefinito i geni causanti-malattia e/o non-causanti-malattia, e successivamente introducendo una correzione basata sul diverso set di dati di riferimento e può essere determinato un bias di utilizzo del codone migliorato.

In una forma di realizzazione, detto passaggio (iv), di dare la priorità ai codoni più differentemente usati è effettuato raggruppando i valori di frequenza di utilizzo del codone ottenuti nello passaggio (iii) usando un algoritmo di raggruppamento gerarchico con un valore-p minore dello 0.05 oppure applicando un'analisi prioritaria in termini di percentuali CUB e selezionando i codoni meno/più utilizzati per le specie, il tessuto e il gene di interesse.

In un'altra forma di realizzazione dell'invenzione, detto passaggio (v), di modificare la struttura secondaria o terziaria di detto gene causante-malattia, è svolto tramite la sostituzione e/o rimozione dei codoni meno utilizzati e più utilizzati, ottenuti nel passaggio (iv), nella sequenza di detto gene causante-malattia.

Un ulteriore oggetto della presente invenzione è un metodo per preparare una molecola di acido nucleico di un selezionato gene causante-malattia,

comprendendo i passaggi discussi secondo ogni realizzazione qui discussa e un ulteriore passaggio (vi) di sintetizzare di una molecola di acido nucleico comprendente la sequenza di acido nucleico ottimizzata per il codone del passaggio (v).

Il passaggio ulteriore (vi) rispetto al precedentemente descritto metodo di sintetizzazione di una molecola di acido nucleico comprendente o consistente nella sequenza di acido nucleico ottimizzata per il codone progettata secondo il precedente passaggio (v), è svolto con procedure e protocolli per la sintetizzazione di molecole di acido nucleico di una data sequenza nota nell'arte.

In una forma di realizzazione della presente invenzione il metodo comprende i seguenti passaggi:

(i) raccogliere le sequenze di uno o più geni causanti-malattia esclusivamente o preferenzialmente espressi in uno o più tessuti di un organismo;

(ii) raccogliere le sequenze di una pluralità di geni non-causanti-malattia esclusivamente o preferenzialmente espressi nello stesso tessuto e organismo dei geni nello passaggio (i);

(iii) determinare il calcolo indipendente della frequenza di utilizzo del codone per i 19 aminoacidi essenziali (escluse metionina e triptofano) in ogni gene raccolto nei passaggio (i) e (ii);

(iv) comparare la frequenza di utilizzo del codone determinata nel passaggio (iii) in modo da ottenere il valore del bias di utilizzo del codone, per identificare codoni a comparsa (codoni tessuto-specifici e codoni gene-specifici) per poter prioritizzare i codoni usati più diversamente usati nel gene e nel/nei tessuto/i di interesse;

(iv-a) ridurre i codoni rari (eccetto i codoni "a comparsa" identificati nei passaggi precedenti);

(iv-b) aumentare il contenuto di guanina (G) e citosina (C) senza intaccare i codoni "a comparsa", ciò significa che se la sostituzione di una A con una G eliminerà un codone "pop up", questo non dovrebbe essere fatto;

(v) progettare una molecola di acido nucleico di detto gene causante-malattia, modificando la struttura secondaria o terziaria di detta molecola di acido nucleico senza eliminare nessun codone a comparsa.

In una forma di realizzazione il metodo della presente invenzione è un metodo per preparare un prodotto genico sintetico che potrebbero essere le proteine, reagenti ricombinanti, e strumenti molecolari. La sintesi di questi prodotti comprendenti o consistenti nella sequenza di acido nucleico ottimizzata per il codone progettata secondo qualunque realizzazione dei metodi qui discussi. Procedure e protocolli per la sintetizzazione di qualsiasi delle sopra-citate molecole di una data sequenza sono conosciute nel campo.

La molecola di acido nucleico sintetica ottenuta con questi metodi può essere usata per esempio nei trattamenti di malattie suscettibili con terapia genica o in esprimenti in vitro per valutare l'espressione di detta molecola di acido nucleico.

Un altro oggetto della presente invenzione è l'uso dei codoni prioritizzati come ottenuti nel passaggio (iv) di entrambi i metodi descritti nella presente invenzione, in un metodo per la valutazione dello sviluppo di una malattia legata alla mutazione di un selezionato gene. In alcune realizzazioni, le sequenze di acido nucleico estratte da geni causanti-malattia che codificano una trascrizione tradotta o non tradotta e/o da geni non-causanti-malattia che codificano una trascrizione tradotta o non tradotta possono essere importate (e.g., individualmente importate da una banca dati) in un programma software implementato per computer che è capace di ottimizzare la sequenza di codifica secondo i metodi qui discussi.

In un ulteriore aspetto l'invenzione è programma per computer comprendente istruzioni che, quando il programma è eseguito da un computer, comportano che il computer svolga i passaggi del metodo secondo ogni realizzazione qui discussa, per esempio un programma che gira su un web server.

Un ulteriore oggetto della presente invenzione è un supporto di memorizzazione leggibile dal computer che comprende istruzioni che, quando eseguito da un computer, comporta che il computer svolga i passaggi del metodo secondo ogni forma di realizzazione qui discussa, per esempio una chiavetta di

memoria, un CD-ROM o un dispositivo comprendente detto supporto di memorizzazione leggibile dal computer.

Un ulteriore oggetto della presente invenzione è un metodo per identificare specie, tessuti o codoni gene-critici per metodi implementati al computer, al fine di valutare il potenziale di sviluppo di una malattia causata delle varianti della sequenza del DNA. I codoni sono definiti critici basandosi sulla loro frequenza di utilizzo, e questi sono codoni estremamente rari o di uso estremamente comune per la specie, il tessuto e il gene di interesse. L'estremizzazione dell'uso potrebbe influenzare l'efficienza di traduzione e/o il ripiegamento del peptide.

Sono sotto riportati esempi che hanno lo scopo di illustrare meglio le metodologie discusse nella presente descrizione, detti esempi non devono in alcun modo essere considerati come limitazioni della precedente descrizione e delle susseguenti rivendicazioni.

Esempi e dati sperimentali

METODI

La nostra strategia era basata sulla comparazione dei valori CU e i loro raggruppamenti gerarchici nei tessuti del muscolo, del rene e della pelle nell'*Homo Sapiens,* tra i mammiferi selezionati e nei geni DC versus i geni NDC.

Selezione delle specie, dati di sequenza e risorse per il calcolo dei valori CU

Abbiamo selezionato i seguenti 15 mammiferi dell'albero filogenetico dei metazoi: *R. ferrumequinum* (Ferro di cavallo maggiore), *M. musculus* (topo), *F. catus* (gatto), *C. lupus familiaris* (cane), *E. caballus* (cavallo), *B. Taurus* (toro), *M. murinus* (lemure topo grigio), *G. variegatus* (colugo della sonda), *C. jacchus* (uistitì comune), *M. mulatta* (macaco), *N. leucogenys* (gibbone), *P. abelii* (orangotango), *G. gorilla* (gorilla), *P. troglodytes* (scimpanzé) e *H. sapiens* (umano).

Tavola 1

SEQUENCES REFERENCES	NCBI LINK
Rhinolophus ferrumequinum (Greater horseshoe bat)	https://www.ncbi.nlm.nih.gov/assembly/GCA_004115265.3
Mus musculus (House mouse)	https://www.ncbi.nlm.nih.gov/gene/13405
Felis catus (Cat)	https://www.ncbi.nlm.nih.gov/assembly/GCF_000181335.3
Canis lupus familiaris (Dog)	https://www.ncbi.nlm.nih.gov/assembly/GCF_000002285.3/
Equus caballus (Horse)	https://www.ncbi.nlm.nih.gov/assembly/GCF_002863925.1
Bos taurus (Cattle)	https://www.ncbi.nlm.nih.gov/assembly/GCF_002263795.1
Microcebus murinus (Gray mouse lemur)	https://www.ncbi.nlm.nih.gov/nuccore/1135509992
Galeopterus variegatus (Sunda flying lemur)	https://www.ncbi.nlm.nih.gov/nuccore/640470054
Callithrix jacchus (Common marmoset)	https://www.ncbi.nlm.nih.gov/assembly/GCF_000004665.1/
Macaca mulatta (Rhesus macaque)	https://www.ncbi.nlm.nih.gov/assembly/GCF_000772875.2/
Nomascus leucogenys (Northern white-cheeked gibbon)	https://www.ncbi.nlm.nih.gov/nuccore/350542784

Pongo abelii (Sumatran orangutan)	https://www.ncbi.nlm.nih.gov/nuccore/180204958
Gorilla gorilla (Western gorilla)	https://www.ncbi.nlm.nih.gov/assembly/GCF_000151905.2/
Pan troglodytes (Chimpanzee)	https://www.ncbi.nlm.nih.gov/assembly/GCF_000001515.7/
Homo sapiens (Human)	Various genes see supp table 2, 3 and 4

La tavola 1 mostra la lista di specie e sequende di assemblaggio di genoma. Le sequenze di riferimento dell'mRNA di tutti i gruppi di geni dell' *H. sapiens* sono stati recuperate da RefSeq e GeneBank al National Center for Biotechnology Information come illustrato nelle tavole 2, 3 e 4.

Tavola 2

GENI CAUSANTI MALATTIA

	GENE	**NCBI LINK**	**RNA TS TPM***	**PROTEIN EXPRESSION (score)****	**OMIM NUMBER**	**TISSUE SPECIFICITY**
1	**DYSF**: Homo sapiens dysferlin (DYSF), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001130987.1	32,8	High	603009	predominantly
2	**CAPN3**: Homo sapiens calpain 3 (CAPN3), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000070.2	336,8	Medium	114240	Predominantly
3	**SGCB**: Homo sapiens sarcoglycan beta (SGCB), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000232.4	35,4	Medium	600900	Predominantly (the highest of two)
4	**SGCA**: Homo sapiens sarcoglycan alpha (SGCA), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000023.3	181,4	Medium	600119	Predominantly

5	**LMNA**: Homo sapiens lamin A/C (LMNA), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_170707.3	55	High	150330	All
6	**DES**: Homo sapiens desmin (DES), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001927.3	5462	High	125660	Only
7	**MYOT**: Homo sapiens myotilin (MYOT), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_006790.2	766,8	High	604103	Only
8	**ANO5**: Homo sapiens anoctamin 5 (ANO5), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_213599.2	23,6	Low	608662	One of the five
9	**COL6A1**: Homo sapiens collagen type VI alpha 1 chain (COL6A1), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001848.2	29,1	Only smooth muscle	120220	Predominantly
10	**TRIM32**: Homo sapiens tripartite motif containing 32 (TRIM32), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_012210.	5,4	Medium	602290	All
11	**DMD**: Homo sapiens dystrophin (DMD), transcript variant Dp427m, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_004006.2	34,8	Medium	300377	Only

GENI NON CAUSANTI MALATTIA

5

	GENE	**NCBI LINK**	**RNA TS TPM***	**PROTEIN EXPRESSION (score)****	**TISSUE SPECIFICITY *****
1	**ACTN3**: Homo sapiens actinin alpha 3 (gene/pseudogene) (ACTN3), transcript variant 1, coding, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001104.3	556,5	Not performed	Only

2	**MYLPF**: Homo sapiens myosin light chain, phosphorylatable, fast skeletal muscle (MYLPF), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_013292.4	6541,1	Medium	Predominantly (the highest of three)
3	**TNNC2**: Homo sapiens troponin C2, fast skeletal type (TNNC2), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_003279.2	9898,9	Medium	Only
4	**ANKRD23**: Homo sapiens ankyrin repeat domain 23 (ANKRD23), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_144994.7	495,5	Medium	Predominantly (the highest of two)
5	**LBX1**: Homo sapiens ladybird homeobox 1 (LBX1), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_006562.4	9,1	Not performed	Only
6	**LSMEM1**: Homo sapiens leucine rich single-pass membrane protein 1 (LSMEM1), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_182597	31,8	Not performed	Predominantly
7	**TMEM38A**: Homo sapiens transmembrane protein 38A (TMEM38A), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_024074.2	199,7	Medium	One of two
8	**RPL3L**: Homo sapiens ribosomal protein L3 like (RPL3L), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_005061.2	323,9	Medium	Only
9	**MYH1**: Homo sapiens myosin heavy chain 1 (MYH1), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_005963.3	2753,9	High	Only

La tavola mostra la lista dei geni Homo sapiens, prioritizzati nel tessuto muscolare scheletrico, utilizzando la Human Protein Atlas database (https://www.proteinatlas.org/) Per poter prioritizzare i geni del muscolo, abbiamo selezionato quelli con una più alta

espressione dalla lista di geni arricchiti del muscolo scheletrico della Human Protein Atlas database (https://www.proteinatlas.org/search/tissue_specificity_rna:skeletal%20muscle;Tissue%20enriched+AND+sort_by:tissue+specific+score+AND+show_columns:groupenriched). Tutti i dati (RNA, TS, TPM, and Protein expression scores) sono stati ottenuti tramite Human Protein Atlas database. *RNA TS TPM indica un livello di RNA riportato come media TPM (transcripts per million), nel tessuto di riferimento, muscolo scheletrico in questo caso. **I punteggi di espressione proteica sono basati su una stima ottimale dalla "vera" espressione proteica da un'annotazione basata sulla conoscenza nel tessuto selezionato,in questo caso il muscolo scheletrico. ***La specificità del tessuto è basata su dati trovati nel grafico nominato "HPA tissue dataset", un sub-categoria della sezione "RNA sample summary" nel sito HPA, per ogni gene.

La sezione RNA summary mostra una normale distribuzione di campioni individuali nei set di dati di analisi multiple RNA-seq visualizzati con i box plot. "Solo" è usato trascrizione di un gene presente solo nel tessuto specifico (muscolo scheletrico). "Principalmente" è usato quando la maggior parte della trascrizione di un gene è presente nel tessuto specifico (muscolo scheletrico). "Tutto/i"è usato per la trascrizione di un gene presente in tutti i tessuti.

Tavola 3

GENI CAUSANTI MALATTIA

	GENE	NCBI LINK	RNA TS TPM*	PROTEIN EXPRESSION (score) **	OMIM NUMBER	TISSUE SPECIFICITY
1	**KRT10**: Homo sapiens keratin 10 (KRT10), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000421.3	18886	High	148080	Only
2	**KRT1**: Homo sapiens keratin 1 (KRT1), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_006121.3	15454,7	High	139350	Predominantly (one of three)
3	**DSG1**: Homo sapiens desmoglein 1 (DSG1), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001942.3	725,7	High	125670	Predominantly

4	**ALOXE3**: Homo sapiens arachidonate lipoxygenase 3 (ALOXE3), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001165960.1	36,5	Medium	607206	Predominantly (the highest of two)
5	**COL17A1**: Homo sapiens collagen type XVII alpha 1 chain (COL17A1), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000494.3	769,3	Medium / Low	113811	Predominantly
6	**FGFR3**: Homo sapiens fibroblast growth factor receptor 3 (FGFR3), transcript variant 3, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001163213.1	333,1	High	134934	Predominantly
7	**TYR**: Homo sapiens tyrosinase (TYR), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000372.4	29,8	High	604103	Only
8	**LOR**: Homo sapiens loricrin (LOR), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000427.2	532	Medium / Low	606933	Only
9	**HOXC13**: Homo sapiens homeobox C13 (HOXC13), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_017410.2	4,6	Only in hair	142976	Only
10	**KRT2**: Homo sapiens keratin 2 (KRT2), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000423.2	2600	High	600194	Predominantly (one of three)

GENI NON CAUSANTI MALATTIA

5

	GENE	**NCBI LINK**	**RNA TS TPM***	**PROTEIN EXPRESSION (score)****	**TISSUE SPECIFICITY**
1	**DCT**: Homo sapiens dopachrome tautomerase (DCT), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001922.4	126,8	Medium/ High	Only

2	**PMEL**: Homo sapiens premelanosome protein (PMEL), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001200054.1	133,4	Medium/ High	Only
3	**GSDMA**: Homo sapiens gasdermin A (GSDMA), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_178171.4	58,4	Medium/ High	Only
4	**KLK5**: Homo sapiens kallikrein related peptidase 5 (KLK5), transcript variant 2, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001077491.1	213,5	Medium	Only
5	**DMKN**: Homo sapiens dermokine (DMKN), transcript variant 2, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_033317.4	3991,5	High	Predominantly
6	**DSC1**: Homo sapiens desmocollin 1 (DSC1), transcript variant Dsc1a, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_024421.2	273	High	Predominantly (one of three)
7	**KRT77**: Homo sapiens keratin 77 (KRT77), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_175078.2	308,7	High	Predominantly (the highest of four)
8	**PLA2G4E**: Homo sapiens phospholipase A2 group IVE (PLA2G4E), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001206670.1	20,4	Medium/ High	Predominantly
9	**KRTDAP**: Homo sapiens keratinocyte differentiation associated protein (KRTDAP), transcript variant 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_207392.2	4647,8	Medium/ High	Predominantly
10	**MLANA**: Homo sapiens melan-A (MLANA), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_005511.1	50,4	High	Only

La tavola mostra la lista dei geni Homo sapiens, prioritizzati nei tessuti della pelle, usando il database Human Protein Atlas (https://www.proteinatlas.org/).

Per prioritizzare i geni della pelle, abbiamo selezionato quelli con espressioni più alte dalla lista dei geni arricchiti della pelle del database Human Protein Atlas

(https://www.proteinatlas.org/search/tissue_specificity_rna:skin;Tissue%20enriched+AND+sort_by:tissue+specific+score+AND+show_columns:groupenriched).

Tutti i dati (RNA, TS, TPM, punteggi di espressione proteicae specificità del tessuto) sono anch'esse state ottenute dal database Human Protein Atlas. *RNA TS TPM indica un livello di RNA riportato come media TPM (transcripts per million), nel tessuto di riferimento,in questo caso la pelle. **I punteggi di espressione proteica sono basati su una stima ottimale della "vera" espressione proteica da un annotazione basata sulla conoscenza nel tessuto selezionato,in questo caso la pelle. ***La specificità dei tessuti è basata su dati trovati nel grafico nominato "HPA tissue dataset", una sub-categoria della sezione "RNA sample summary" nel sito HPA, per ogni gene. La sezione RNA summary mostra la normale distribuzione di campioni individuali nei set di dati di analisi multiple RNA-seq visualizzati con i box plot. "Solo" è usato per la trascrizione di un gene presente solo nel tessuto specific (pelle). "Principalmente" è usato quando la maggior parte della trascrizione di un gene è presente nel tessuto specifico (pelle). "Tutto/i"è usato per la trascrizione di un gene presente in tutti i tessuti.

Tavola 4

GENI CAUSANTI MALATTIA

	GENE	LINK NCBI	RNA TS TPM*	ESPRESSIONE PROTEINA (PUNTEGGIO)**	NUMERO OMIM	SPECIFICITA' TESSUTALE
1	**UMOD**: Homo sapiens uromodulina (UMOD), variante trascrtto 2, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001008389.2	2392	Alta	191845	Unica
2	**SLC12A1**: Homo sapiens solute carrier famiglia 12 membro 1 (SLC12A1), variante trascritto 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000338.2	780	Alta	600839	Unica
3	**KCNJ1**: Homo sapiens sottofamiglia canale voltaggio-dipendente J membro 1 (KCNJ1), variante trascritto 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000220.4	210,8	Alta	600359	Unica

4	**SLC12A3**: Homo sapiens solute carrier famiglia 12 membro 3 (SLC12A3), variante trascritto 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000339.2	109,8	Media	600968	Unica
5	**NPHS2**: Homo sapiens, membro della famiglia della stomatina, podocina (NPHS2), variante trascritto 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_014625.3	93,6	Alta	604766	Unica
6	**BSND**: Homo sapiens barttin subunità beta accessoria tipo CLCNK (BSND), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_057176.2	10,2	Alta	606412	Prevalentemente (il più alto dei due)
7	**CLDN16**: Homo sapiens claudin 16 (CLDN16), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_006580.3	51,7	Media	603959	Unica
8	**PKD1**: Homo sapiens policistina 1, recettore transitorio che interagisce con il canale potenziale (PKD1), variante del trascritto 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001009944.2	2,3	Alta	601313	Tutte
9	**PKD2**: Homo sapiens polycystin 2, recettore transitorio del canale cationico potenziale (PKD2), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_000297.3	54,2	Media/Bassa	173910	Tutte
10	**ATP6V0D2**: Homo sapiens ATPasi H+ trasportante la subunità V0 d2 (ATP6V0D2), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_152565.1	59,4	Media	618072	Unica

GENI NON CAUSANTI MALATTIA

5

	GENE	LINK NCBI	RNA TS TPM*	ESPRESSIONE PROTEINA (PUNTEGGIO)**	NUMERO OMIM
1	**BBOX1**: Homo sapiens gamma-butirrobetaina idrossilasi 1 (BBOX1), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_003986.2	473	Alta	Prevalentemente (il più alto dei due)
2	**SLC22A8**: Homo sapiens solute carrier famiglia 22 membro 8 (SLC22A8), variante trascritto 2, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001184732.1	339,8	Media	Unica
3	**MIOX**: Homo sapiens mio-inositolo ossigenasi (MIOX), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_017584.5	821,5	Media	Unica
4	**TMEM52B**: Homo sapiens proteina transmembrana 52B (TMEM52B), variante trascritto 2, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001079815.1	183,5	Media	Prevalentemente (il più alto dei due)
5	**TINAG**: Homo sapiens antigene della nefrite tubulointerstiziale (TINAG), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_014464.3	113,9	Bassa	Prevalentemente (uno dei due)
6	**CALB1**: Homo sapiens calbindina 1 (CALB1), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_004929.3	336,9	Alta	Prevalentemente (uno dei due)
7	**ATP6V1G3**: Homo sapiens ATPasi H+ trasportante la subunità V1 G3 (ATP6V1G3), variante trascritto 3, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001320218.1	21,3	Media	Unica
8	**AQP6**: Homo sapiens aquaporina 6 (AQP6), mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_001652.3	20,8	Alta	Unica

9	**FXYD4**: Homo sapiens dominio FXYD contenente ioni regolatori di trasporto 4 (FXYD4), variante trascritto 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_173160.2	127,4	Media	Unica
10	**GGACT**: Homo sapiens gamma-glutamilammina ciclotransferasi (GGACT), variante trascritto 1, mRNA	https://www.ncbi.nlm.nih.gov/nuccore/NM_033110.2	75,9	Bassa	Unica

La tavola mostra la lista di geni Homo sapiens, prioritizzati nel tessuto del rene, utilizzando il database Human Protein Atlas (https://www.proteinatlas.org/). Per prioritizzare I geni del rene abbiamo selezionato quelli con espressione più alta dalla lista di geni arricchiti del rene del database Human Protein Atlas (https://www.proteinatlas.org/search/tissue_specificity_rna:kidney;Tissue%893 20enriched+AND+sort_by:tissue+specific+score+AND+show_columns:groupenriched).

895 Tutti i dati (RNA, TS, TPM, punteggi di espressione proteica e specificità del tessuto) sono anch'essi stati ottenuti dal database Human Protein Atlas. *RNA TS TPM indica un livello di RNA riportato come media TPM (transcripts per million), nel tessuto di riferimento,in questo caso il rene. **I punteggi di espressione proteica sono basati su una stima ottimale della "vera" espressione proteica da un'annotazione basata sulla conoscenza nel tessuto selezionato,in questo caso il rene. ***La specificità del tessuto è basata su dati trovati nel grafico nominato "HPA tissue dataset", una sub-categoria della sezione "RNA sample summary" nei sito HPA,per ogni gene.

La sezione RNA summary mostra la normale distribuzione di campioni individuali nei set di dati di analisi multiple RNA-seq visualizzati con i box plot. "Solo" è usato per la trascrizione di un gene presente solo nel tessuto specifico (rene). "Principalmente" è usato quando la maggior parte della trascrizione di un gene è presente nel tessuto specifico (rene). "Tutto/i"è usato per la trascrizione di un gene presente in tutti i tessuti.

I mammiferi sono stati selezionati basandosi sul numero più alto di geni annotati.

Abbiamo selezionato geni con differenti lunghezze come l'mRNAs e, per massimizzare il numero di codoni analizzati ed evitare pregiudizio contro le sequenze brevi o parziali, sono state selezionate solo sequenze codificate di piena lunghezza.se è stata annotate più di una isoforma di splicing, sono state selezionate le isoforme più

lunghe, dal moment oche recenti studi sulla singola cellula di RNAseq hanno dimostrato che la lunghezza dell'mRNA non influenza il livenno di espressione nel tessuto isoforme.

Selezione dei geni

I geni DC sono stati prioritizzati basandosi sul loro coinvolgimento nelle malattie Mendeliane rare, con incidenza minore di 1:5000 (secondo il catalogo OMIM, www.omim.org), e con un fenotipo omogeneo e legato a tessuti/organi specifici (renale, muscolo scheletrico e pelle). Criteri di esclusione per i geni sono stati il coinvolgimento im malattie poligeniche o cancro (entrambi Mendelianie somatici) e l'assenza di specificità di tessuto (geni di manutenzione). Abbiamo poi selezionato i geni DC basandoci sulla malattia causata e la tessuto-specificità e I genoi NDC basandoci sulla loro maggiore espressione negli stessi tessuti nei quali erano espresso i geni DC (muscolo, pelle e rene).

I criteri di selezione dei geni (Tavole 2, 3 e 4)erano basati su: i) geni pienamente annotati riconosciuti come causanti malattie Mendeliane rare (geni DC), ii) geni pienamente annotati non causanti malattie Mendeliane rare (geni NDC).

Il database dle catalogo OMIM è stato usato per categorizzare i geni come DC o NDC. I geni DC dovevano essere associate con una malattia Mendeliana in uno dei tre tessuti selezionati (rene, muscolo scheletrico e pelle)in almeno 5 famiglie/pazienti riportati, quindi essere confermate nel database OMIM, e con pattern ereditati definiti (autosomico recessivo, autosomico dominante o X-linked recessivo). Evidenza del coinvolgimento del gene in cancri Mendeliani e somatici, come la suscettibilità dei geni, sono stati considerati criteri di esclusione per la selezione dei geni, dal momento che i geni del cancro Mendeliano sono spesso anche geni predisponendi al cancro, e pertento possono rappresentare fattori di confusione nel nostro studio, che si concentra solo sulle malattie Mendeliane.

I tipi di malattie rare sono legate alle mutazioni dei geni dei tessuti del muscolo, della pelle e del rene. Malattie rare del muscolo includono distrofie muscolari e miopatie congenite, i cui geni causative sono principalmente e maggiormente espresso nel muscolo scheletrico come distrofina, disferlina, onella matrice extracellulare del muscolo, come il gene del collagene 6A1. Malattie rare del rene includono gene dell'uromodulina e della policistina 1, le cui mutazioni causano Malattia renale tubulointerstiziale o rene policistico di tipo 1 rispettivamente, gene altamente espresso in questi due differenti compartimenti del rene. Infine, malattie rare della pelle includono geni della cheratina 10 e HOXC13, le cui mutazioni sono associate con l'ipercheratosi

epidermolitica e la displasia ectodermica 9, due malattie differenti in termini di fenotipo e coinvolgimento dello strato cutaneo. Le tavole 2, 3, e 4, riportano l'intera lista di geni RD con tutti i numeri OMIM corrispondenti.

Due geni , TMEM52B and PLA2G4E, non elencati del database OMIM dal moment oche non sono mai stati associate con alcuna delle malattie umane, sono stati controllati usando i database PubMed, ClinVar e DMDM e quindi esclusi per essere causativi di malattie Mendeliane.

Abbiamo selezionato i tessuti basandoci sul loro alto arricchimento genico nel database Human Protein Atlas (HPA). la nostra prioritizzazione era basata su metriche HPA usate pe ril livello di RNA (Transcripts Per Million, TPM), il punteggio di espressione proteica (alto, medio, basso, nullo)e valori di tessuto-specificità. Questi punteggi ci hanno permesso di classificarei profili di espressione di geni NDC e DC secondo la loro specificità di tessuto. Il più alto valore di espressione implica almeno livelli di mRNA quattro volte più alti nei tessuti selezionati comparata a ogni altro tessuto, mentre i punteggi proteici erano alti o medi, bassi livelli di espressione sono stati esclusi.

Database delle mutazioni

Abbiamo esaminato i database pubblici OMIM, ExAC and ClinVar insieme con i nostri database interni UNIFE per variazioni genetiche a singolo nucleotide DMD. Sono state considerate unicamente le variazioni patogeniche missenso e nonsenso, dal momento che il loro significato e la loro identificazione nei database non sono equivoci. Cambiamenti sinonimi nel gene DMD non sono considerati in questo studio dal moment oche sono quasi invariabilemnte definiti come variant dal significato incerto (VUS) ovarianti benigne, decondo le line guida dell'ACMGG.

Calcolo dei valori CU e analisi statistica

La frequenza di utilizzo del codone (CU) è stata calcolata indipendentemente per ognuno dei 19 anminoacidi considerati in ogni gruppo di geni. E' stato valutato anche l'uso dei tre codoni di stop. Metionina (AUG) and triptofano (UGG) non sono stati inclusi dal momento che sono codificati fa un'unica tripletta. Tutte le analisi statistiche sono state svolte con R-3.4.4 (R version 3.4.4. R. https://www.r-project.org/ (2018)).

La significanza statistica è stata definite come valore $P < 0.05$. Test statistici basilari e generazione di diagrammi a barre e diagrammi a scatola sono state svolte utilizzando funzioni built-in incluse con la distribuzione base di R o funzioni nel pacchetto ggplot2 (Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. (2016); 978-

3-319-24277-4).

Per un paragone di utilizzo del codone tra geni DC e NDC, tipi di tessuti e specie, un sum test Wilcox on rank è stato applicato per calcolare valori P a due code usando la funzione 'Wilcox test' in R (Bauer DF. Constructing confidence sets using rank statistics. Journal of the American Statistical Association. (1972); s*67*, 687-690), e questi sono stati visualizzati utilizzando il pacchetto R ggplot2.

La frequenza di CU in geni DC eNDC è stata anche comparata per identificare i codoni usati più differentemente nei tessuti genici, tra le specie e nei geni DC versus NDC, e i dati sono stati visualizzati indiagrammi a scatola.

Il coefficiente di correlazione Spearman di frequenza di utilizzo dei codoni nei geni DC e NDC del muscolo, della pelle e del rene nell'HSA è stato usato e visualizzato utilizzando il pacchetto R ggplot2.

La funzione heatmap.2 nel pacchetto R gplots è stata usata per il raggruppamento dei codoni e la loro visualizzazione.

Nel raggruppamento, poiché gli usi dei codoni sono dati di intervallo e non sono influenzati da outlier con valori estremamente larghi, la distanza metrica euclidea è stata selezionata per una facile implementazione e semplice interpretazione. Il raggruppamento agglomerativo gerarchico è stato svolto utilizzando il metodo standard "completo" in funzione hclust.

I geni sono sempre elencati in grafici in ordine decrescente secondo il numero 224 di esoni, e il numero di esoni è stato calcolato basandosi su annotazioni di dati scaricate dal database Ensembl Genome. I dati sono stati analizzati usando l'algoritmo di raggruppamento gerarchico o applicando un'analisi prioritaria in termini di percentuali CUB.

I codoni sinonimi che non sono affatto utilizzati dai geni, pertanto avendo un bias di codone estremo, sono stati chiamati "codoni-zero".

RISULTATI

Impronta CUB dei geni tessuto-specifici dell'Homo sapiens (HSA)

Abbiamo inizialmente verificato il raggruppamento gerarchico di uso dei codoni sinonimi in tutti i geni studiati con diversa specificità di tessuto (muscolo, pelle e rene) nell'HSA. Abbiamo osservato un raggruppamento di codoni tessuto-specifici, che abbiamo definite "imprinta CUB". I grafici dei diagrammi di calore (Figura 1) mostrano

che raggruppamenti di codoni frequentemente utilizzati (in rosso) e raramente utilizzati (blu scuro) variano grandemente tra i tessuti umani.

Nei geni del muscolo, codoni con un bias estremo (valori CU bassi, colore blu scuro, o alti valori CU, colore chiave rosso scuro) sono strettamente raggruppati in termini sia di gene che di tipo di codone, mentre valori CU intermedi (colore chiave azzurro o giallo) sono disperse negli alberi (Figura 1A).

L'impronta CUB dei geni della pelle (Figura 1B) è caratterizzata da una predominanza di valori CU bassi (punti gialli) con pochi punti rossi disperse e un chiaro raggruppamento di codoni più rari (punti blu scuro), mentre i pochi codoni con valori CU intermedi o alti sono distribuiti in maniera disomogenea tra i geni con grandi distanze nell'albero. I raggruppamenti di codoni generate sono anche meno definiti se paragonati al muscolo e al rene.

Nei geni del rene, l'impronta CUB differisce dagli altri due gruppi di geni e la gerarchia dei codoni è più definite (solo due grandi linee, Figura 1C). I codoni con valori intermedi sono raggrippati con tendenze legate al gene, come visible nei geni PKD2, KCNJ1 e MIOX. La stragrande maggioranza dei geni hanno valori CU bassi o intermedi con ampi e diffusi raggruppamenti di punti blu. Un piccolo gruppo di geni (UMOD, BSND, SLC22A8, MIOX, AQP6, PKD1, SLC12A3, and GGACT) mostrano alti valori CU dei codoni UGA, UAC, UUC, AUC, GAC, AAC, AAG, GAG, UGC, CAC, e CAG, appartenenti a linee ben definite nel raggruppamento (impronta CUB gene-codone specifica).

Guardando al raggruppamento dei codoni, i codoni più utilizzati, CAG, AAG, UGC, GAG, UAC, UUC, CAC, AAC, AUC e GAC, hanno raggruppamento identico e si sovrappongono nedi geni del muscolo e del rene ma non nei geni della pelle (Figura 1A e 1C, lato sinistro). Tra i geni dei muscoli, solo DMD non mostra un CUB estremo, dal moment oche non si verificano punti rossi (Figura 1A), con l'unica ovvia eccezione dell'unico codone di stop UAG.

Guardando al raggruppamento dei geni, i dendrogrammi dei geni del muscolo e del rene mostrano impronte riconoscibili (gerarchia vertical nelle mappe di calore nella Figura 1A e 1C), dal momento che i geni con valori CU alti o bassi (arricchiti con punti rossi o blu) sono chiaramente raggruppati (13 geni nella Figura 1A, e 8 geni nella Figura C, lato sinistro), mentre nei geni della pelle questo non si verifica. Pertanto, some annotato sopra, sono stati osservati alcuni valori CU gene-specifici, dipendentemente dal tessuto studiato. I valori di utilizzo dei codoni di stop sono stati calcolati nei tre gruppi di geni. Nei geni della pelle e del muscolo, tutti i codoni di stop sono usati uniformemente,

con UAApresente in maniera più frequente.Nei geni del rene, UAG è usato molto raramente, mentre UAA e UGA sono ugualmente rappresentati.

Impronte CUB tra i mammiferi

Abbiamo analizzato valori CU in tutti i 20 geni tessuto-specifici tra 15 specie di mammiferi dell'albero filogenetico dei metazoi (Figura 2A, B, e C) o i valori CU in tutti i mammiferi e tra i tre tessuti genici (Figura 2D, E, e F). I valori CU tra i mammiferi mostrano evidenti impronte CUB tessuto-specifiche dovute a differenti usi del tipo di codone (Figura 2A, B, C).

CAG, AAG, CAC, GAC, GAG, AUC, AAC, UAC, UGC e UUC sono i codoni usati più frequentemente (punti rossi) in tutti i tessuti genici e tra i mammiferi, mentre UUA, CUA, UCG, CGU, CUU, GUA, CGA, AUA, UCA, UUG e GCG sono i codoni più rari (punti blu) nei geni del muscolo e della pellema non dei reni. I CUB del muscolo e della pelle hanno un comportamento simile del raggruppamento dei valori CU, sebbene con valori CU più bassi nella pelle (più punti gialli), simile ai geni del rene che mostrano un differenteraggruppamento gerarchico.

Guardando al raggruppamento dei geni in tutti i mammiferi (Figura 2D, E, and F), i raggruppamenti di valori CU non possono essere visti. Tutti i tessuti hanno impronte CUB differenti, con anche un differente raggruppamento gerarchico dei codoni.

I dendrogrammi legati ai geni sono anche diversi dal momento che i geni del muscolo (13/20), del rene (8/20) e della pelle (4/20) si raggruppano insieme con diversi valori CU legati al diverso tipo di codone. Infatti considerando tutti i geni in tutte le specie i tipi di codone variano. pertanto, sebbene alcune impronte CUB legate ai tessuti sono ancora riconoscibili nei mammiferi, non può essere osservato nessun chiaro comportamento tissutale o raggruppamento legato al gene specifico. Questa è una dimostrazione che i valori CU mostrano una correlazione con i mammiferi ma con importanti differenze gene-specifiche, che cointribuiscono a generare le impronte CUB.

Valori CU e impronte CUB nai geni DC e NDC nell'HSA

Abbiamo raggruppato i geni basandoci sulla loro propensione ad essre sito di variazioni patogeniche (mutazioni) causando malattie rare (geni DC o NDC). Abbiamo profilato i valori di CU nei geni DC e NDC, preservando la distinzione del tessuto (geni del muscolo, della pelle e del rene) tra i mammiferi. La figura 3,pannelli dalla A alla F, mostra i valori CU assoluti nei geni DC e NDC del muscolo, della pelle e del rene, rispettivamente, con Nessun raggruppamento gerarchico e identico ordine di tipo di codone. I valori CU in questi 6 pannelli mostrano che il tipo di codoni più frequenti o più

rari sono molto simili in tutti i geni e tra i mammiferi. Questo è supportato nell'HSA dall'analisi della correlazione di Spearman, che ha dimostrato che i valori CU dei geni DC e NDC sono significativamente correlati nei gruppii del muscolo, della pelle e del rene ($p<0.05$) (Supplementary Figura 1).

Più variabilità di valori CU può essere vista nei codoni con frequenza intermedia, dove una tendenza di gene o di tessuto può essre vista. In particolare, i geni NDC del muscolo hanno valori CU più alti (è visible qualche punto giallo, vedi Figura 3 A, B), mentre i geni DC e NDC del muscolo edel rene (Figura 3, pannelli A, B, E, F) hanno valori CU molto simili. Da notare, CAG è il codone usato più frequentemente e UUA quello usato meno di frequente in tutti i mammiferi.

La figura 3, pannelli da G ad L, mostrano un raggruppamento gerarchico di valori CU nelle stesse categorie di geni di cui sopra. Guardando ai geni DC e NDC, possono essere osservate impronte CUB riconoscibili. I geni del muscolo DC e NDC (Figura 3, pannnelli G, H) mostrano differenti impronte e raggruppamenti. I geni DC del muscolo hanno un raggruppamento compatto di codoni frequentemente utilizzati (AAG, CAG, GAG) o di codoni estremamente rari (UGG, UUA, CUA). I valori CU sono omogenei tra i mammiferi con gruppi di codoni charamente definiti in termini di distanza dall'albero. Nei geni NDC del muscolo, l'impronta CUB cambia. Dominano punti rossi e blu scuro, con pochi codoni con valori CU intermedi (punti gialli). Quest'impronta indica che un CUB più forte si è verificato nei geni NDC del muscolo. Di conseguenza, il dendrogramma basato sul codone nei geni NDC, ma non nei geni DC, mostra che valori CU più alti e più bassi sono raggruppati insieme, sottolineando un possible comportamento diverso tra i mammiferi.

I geni della pelle DC e NDC (Figura 3I-J) mostrano simili impronte CUB con differenze minime. I geni della pelle NDC mostrano più codoni con bassi valori CU (figura superiore 3J) e viceversa meno codoni con valori CU alti o intermedi (figura inferiore 3J), paragonati ai geni DC. Questo implica che Valori CU intermedi si verificano più frequentemente, una scoperta opposta a quella vista nei geni dei muscoli.

Infatti, l'impronta dei geni DC della pelle è simile a quella vista nei geni dei muscoli DC. Possiamo concludere che i geni DC sia del muscolo che della pelle mostrano un tipico modello "no CUB estremo".

Le impronte dei geni DC e NDC del rene (Figura 3K-L) e i dendrogrammi dei codoni differiscono grandemente dagli altri due tessuti, dal momento che le distanze gerarchiche tra raggruppamenti di valore sono opposti. Sebbene la conservazione di valori CU tra i mammiferi si verifichi, la gerarchia del dendrogramma dei geni del rene

mostra un antenato commune per valori CU intermedi e bassi, al contrario dei geni del muscolo e della pelle, dove due line distinte (valori CU alti e bassi) sono visibili, come già osservato prima (Figura 1C).

Le impronte CUB dei geni DC e NDC sono simili, sebbene debba essere notato che i geni NDC mostrano un più alto numero di valori CU bassi (punti gialli).

Abbiamo anche comparato i valori CU tra geni dell'HAS alto-, medio-, e basso-espressi. I geni DC e NDC sono stati quindi divisi in tre categorie dipendenti dal livello di espressione del gene (per la piega dell'RNA e i valori di cut-off delle proteine. Le impronte Cub hanno grande similarità, significand oche infatti, raggruppare i geni secondo il loro livello di espressione porta a una tendenza simile di valori CU. Abbiamo anche analizzato se potessero esserci codoni usati di più secondo il livello di espressione del gene. I codoni AAC, GAC, UGC, UAC, CAC, UUC, AUC, AAG, GAG e CAG sono più frequentemente usati sia nei geni alto- che medio-espressi, mentre GUG è presente solo in geni altamente espressi. I codoni GAC, CAC, UUC, CAG, UAC, UGC, AAG, AUC, GAG, AAC, ACC, GGC, GUC, UCC e GCG sono più frequentemente usati in geni con un basso livello di espressione. In particolare, mentre la maggioranza dei codoni frequentemente usati sono similmente rappresentati in geni ad alta, media e bassa espressione, pertanto per tanto con un CU non influenzato dal livello di espressione, solo pochi codoni con valori CU più bassi come UCC (Ser), ACC (Thr), GGC (Gly), GUC (Val) e GCG (Ala) sono presenti in geni basso-espressi. Similarmente, codoni con valori CU bassi sono sono gli stessi nei geni altamente espresso contro i geni mediamente espressi. Alcuni geni DV altamente espressi hanno più codoni con valori CU alti, come DYS, LMNA e DES (nel muscolo scheletrico), UMOD e PKD1 (nel rene) e FGFR3 (nella pelle). La tendenza è opposta nei geni mediamente espressi, dove alcuni geni NDC, come MLPF, TNNC2, TMEM3BA (nel muscolo scheletrico), e NCLZ2, MCX (nel rene), mostrano valori CU più alti.

È interessante notare che, UAA, che è stato riconosciuto come induttore della terminazione della traduzione con più velocità e accuratezza a livello ribosomiale e può essere letto sia da fattori di rilascio eRF1 che eRF2, è il codone di stop più utilizzato nei geni altamente espressi. UAG e UGA hanno una frequenza simile in tutti i geni.

<u>Prioritizzazione dei 343 codoni usati più differentemente tra i geni tessuto-specifici</u>

Abbiamo applicato la stessa strategia di utilizzare il calcolo dei valori CU per paragonare diversi geni umani per identificare evntuali codoni che possono essere più

differentemente/preferibilmente utilizzati in alcuni geni causanti malattia, comparati a geni non causanti malattia.

Basandoci sul valore di uso del codone p ($P < 0,05$), abbiamo prioritizzato codoni con valori CU differenti tra i tessuti genici. I risultati sono mostrati nella Figura 4. Cinque codoni sono stati prioritizzati, essendo usati in maniera significativamente diversa nei tre tessuti umani: CGU (Arg), CCA (Pro), GAC (Asp), GAU (Asp) and GUA (Val) (Figura 4 A-E). CCA e CGU sono i codoni meno frequentement eusati nel muscolo, GUA nel rene, e GAU e GAC sono I codoni più frequentemente usati nella pelle.

Considerando i geni DC e NDC, ulteriori differenze nei valori CU possono essere osservate per CGU e CCA, che sono usati più frequentemente nei geni DC, e per GUA e GAU, che, sebbene meno significativamente, sono usati più frequentemente nei geni NDC (Figura 4 F-L). Questi dati suggeriscono che i valori CU possono essere influenzati dalla propensione del fenotipo e del gene a causare malattie genetiche. La tendenza del valore CU di questi 5 codoni sembra essersi conservata tra i mammiferi (Figura 4F- L).

Infine, abbiamo contato il numero di codoni con bias estremo nei geni e tra i mammiferi. I geni DC mantengono un uso più elevato di codoni multipli, con solo pochi codoni con bias estremo, paragonati ai geni NDC (Tavola 5, pannelli A-B).

Tavola 5

Panel A

NDC															
	Rhinolophus_ferrumequinum	Mus_musculus	Felis_catus	Canis_lupus	Equus_caballus	Bos_taurus	Microcebus_murinus	Galeopterus_variegatus	Callithrix_jacchus	Macaca_mulatta	Nomascus_leucogenys	Pongo_abelii	Gorilla_gorilla	Pan_troglodytes	homo_sapiens
kidney	85	93	97	98	111	108	100	83	105	109	116	120	118	117	114
muscle	85	88	91	97	105	99	100	88	89	96	97	107	97	98	104
skin	62	62	54	62	61	75	64	50	57	75	65	64	59	62	63

DC															
	Rhinolophus_ferrumequinum	Mus_musculus	Felis_catus	Canis_lupus	Equus_caballus	Bos_taurus	Microcebus_murinus	Galeopterus_variegatus	Callithrix_jacchus	Macaca_mulatta	Nomascus_leucogenys	Pongo_abelii	Gorilla_gorilla	Pan_troglodytes	homo_sapiens
kidney	39	35	33	28	42	38	41	45	38	38	38	32	31	31	31
muscle	46	35	46	53	49	47	52	46	46	45	44	51	44	45	45
skin	37	66	73	64	67	37	74	52	65	74	56	74	64	57	78

Inglese	Italiano
Kidney	Rene
Muscle	Muscolo
Skin	Pelle

Tavola 5. Numero di codoni con bias estremo (“codoni-zero”) trovati in geni DC e NDC tessuto specifici tra i mammiferi

Tavola 5 (pannello A) mostra il numero di codoni con bias estremo (“codoni-zero”) nei tessuti genici e tra i mammiferi. I geni sono raggruppati in DC e NDC. Il numero di “codoni-zero” è basato sul valore di uso dle codone (CU) che abbiamo calcolato in ogni gruppo di geni e tra i mammiferi. I geni DC mantengono un uso multiplo del codone (CUB basso), pertanto con pochi “codoni-zero” paragonati ai geni NDC. Da notare che, il numero di “codoni-zero” è molto più alto nei geni NDC del muscolo e del rene e nei geni DC della pelle, quindi con un comportamento guidato alla malattia. Il CUB è cresciuto progressivamente durante l’evoluzione, sebbene non allo stesso livello in tutti i gruppi di geni. La differenza più marcata nel numero del “codone-zero” è tra HSA rene DC (11) e NDC (94) geni e muscolo DC (49) e geni NDC (109). I pannelli B e C mostrano gli stessi dati riportati sui grafici per apprezzare meglio la tendenza del CUB e il numero di “codoni-zero”.

Comportamento unico del CUB DMD

Abbiamo usato i valori CU calcolati e abbiamo mappato 2828 mutazioni patogeniche missenso e nonsense conosciute nel gene distrofina (DMD), prese da database pubblici (LOVD) o database interni, sui codoni DMD. Abbiamo chiamato il nostro approccio alle mutazioni “map-on-codon”.

Abbiamo verificato se codoni DMD raramente/frequentemente usati sono conseguentemente di rado o di frequente sito di variazioni patogeniche missenso e nonsense provate. Questo supporta che alcuni codoni sono più o meno pronti ad essere il sito di variazioni patogeniche in un contest di gene specifico (DMD, in questo caso), e i “codoni più-meno mutati” sono rilevanti per la capacità di traduzione del gene (DMD), e/o per alter funzioni legate alla traduzione, significando che esse devono essere considerate quando si svolge un’ottimizzazione del codone del gene artificiale.

Il gene DMD non ha CUB estremo e mantiene tutti i tipi di codone usati nella sua sequenza codificata, anche tra i mammiferi studiati. Abbiamo contato il numero dei tipi di codoni nella sequenza codificata DMD e identificato solo 4 codoni con bias estremo, UCG, CCG, ACG e GCG, conseguentemente il parametro noto di bias cut-off, basati sulla ridondanza dei codoni di 2, 3, 4 e 6 triplette (Weissbach H. Syntax of referencing in Molecular Mechanisms of Protein Biosynthesis (cap. Lipmann, F. Twenty Years of Molecular Biology)(ed. Nutley, New Jersey : Elsevier). (2012); 736: 3-5. LOVD. https://databases.lovd.nl/shared/genes/DMD/ (2019)), (Figura 5A).

La figura 5B mostra il numero di mutazioni verificatesi a tutti i tipi di codone. Ogni numero, in cima alle barre, rappresenta quante volte quello specific codone è stato il sito

di una variazione DMD missenso o nonsenso. E' interessante notare che, questi numeri non sono legati a valori CU. Sebbene CCG (Pro), UCG (Ser), GCG (Ala) e ACG (Thr) in rosso nelle figure 5A e B, sono, come aspettato, meno di frequente sito di mutazioni, essendo i codoni più rari usati nel gene DMD, i valori CU degli altri codoni non sono correlati con il verificarsi delle mutazioni. Questo è il caso di CAG (Glu), che mostra valori CU intermedi ma è il codone più frequentemente sito di mutazioni DMD, e UAU (Tyr) e UUU (Phe), che sono i codoni DMD più frequentemente utilizzati (Figura 5A) ma sono raramente sito di mutazioni (Figura 5B).

DISCUSSIONE

Abbiamo calcolato i valori CU in tre piccolo gruppi di geni che sono tessutospecifici e maggiormente espresso nella pelle, nei reni e nel muscolo scheletrico. Abbiamo poi innovativamente comparato i valori CUtra geni DC e NDC, e tra i mammiferi, pertanto utilizzando un approccio guidato alla malattia, per esplorare i valori di CU e il comportamento CUB.

Abbiamo confermato che la tessuto specificità influenza il CUB, e abbiamo osservato una tendenza tessuto-specifica, guardando il CUB in geni alto-, medio-, e basso-espressi. Eì interessante notare che, alcuni codoni sono più rappresentati in geni alto- o basso-espressi, un fatto possibilmente legato auna selezione positiva di codoni attività traslazionale più alta o più bassa durante l'evoluzione, dipendentemente dal ruolo del gene in tessuti specifici e/o organi specifici. Di conseguenza, il codone di stop UAA è preferenzialmente usato nei geni del muscolo e, generalmente, nei geni altamente espressi, probabilmente riflettendo il bisogno di un'ottimale ricognizione ribosomiale dle codone per fermare efficientemente la traduzione. Abbiamo quindi confermato che, tramite anche il calcolo dei valori CU e comparandoli in piccolo gruppi di geni umani selezionati, altamente tessutospecifici, le differenze di valori CU possono essere osservate, dipendentemente dal tessuto e dal livello di espressione del gene.

<u>Impronte tessuto-speifiche nei geni di HSA</u>

Comparando i valori CU in raggruppamenti gerarchici nei geni del muscolo, della pelle e del rene nell'HSA, possono essere osservati diversi pattern. Tipi di codone raramente o frequentemente usati variano tra i tre tessuti, essendo i geni del muscolo e della pelle più simili in termini di raggruppamenti e gerarchia dei valori CU. Queste differenti impronte CUB possono essere dovute alla tessuto specificità dei geni analizzati, e noi supponiamo che i geni aventi qualche funzione tissutale che può richiedere un'efficienza di traduzione più alta o più bassa, usano diversi codoni sinonimi. E' interessante notare

che, molti geni del muscolo e del rene condividono gli stessi codoni raggruppati, frequentemente usati, supportando che questi possano essere i codoni chiave per regolare la traduzione tessuto-specifica. Da notare che, muscolo e rene, insieme con fegato e polmone, sono tessuti parenchimali, che subiscono una simile scarsa capacità di rigenerazione degli organi, come conseguenza dei compromessi evolutivi, che è specialmente legata agli effetti di bilanciamento tra il Sistema immune e la forma di processi fisiologici e patologici. Inoltre, un simile raggruppamento di alti valori CU è ancora più evidente per qualche gene del muscolo (COL6A1, RPL3L, MYLPF, TMEM38A, TNNC2, LMNA, DES, LBX1, SGCA, ANKRD23, CAPN3, DYSF, ACTN3) e del rene(UMOD, BSND, SLC22A8, MIOX, AQP6, PKD1, SLC12A3, GGACT), sottolieando qualche funzione gene-specifica e/o legata all'organo.

E' interessante notare che questi geni sono tutti geni DC, fatto consistente con i nostri risultati di paragone tra geni DC e NDC (vedi sotto). Di conseguenza, molti geni della pelle mostrano valori CU bassi o intermedi, con pochi valori CU alti mai raggruppati. la pelle è riconosciuta come un "microambiente immunologico" che regola la rigenerazione della cellula, da ciò la sua tendenza CUB opposta, paragonata ai geni del muscolo e del rene, può riflettere la sua diversa funzione organica e di sviluppo.

Impronte CUB tra i mammiferi

Le differenze di impronte CUB che abbiamo osservato nei hgeni del muscolo, dle rene e della pelle sono osservabili anche tra i mammiferi. La conservazione evolutive del CUB è stata ampiamente studiata nell'HSA, ma non in una tale granularità genetica. I geni del muscolo e della pelle, ma non del rene, mostrano simili pattern CUB con due maggiori linee gerarchiche, una per i codoni più frequenti e una per quelli più rari. Queste tendenze suggeriscono che il CUB nei geni del muscolo e della pelle possono aver possibilmente seguito qualche percorso evoluzionistico comune. I geni del muscolo sono estremamente important nei mammiferi, dove contribuiscono per circa l'80% della massa corporea. Nell'HSA, L'acquisizione di bipedalismo ha certamente richiesto una robusta forza selettiva per guidare il rimodellamento muscolare, specialmente per i muscoli legati alle giunture degli arti. Può essere trovata un origine commune tra la pelle e il muscolo striato nel panniculus carnosus, un sottile strato muscolare striato attaccato alla pelle e alla fascia della maggior parte dei mammiferi, che fornisce support per le funzioni di pulsazione e contrazione del muscolo. Il panniculus carnosus è ancora conservato negli umani, sebbene non sia considerato avere alcun significato funzionale, ed è un residuo dell'evoluzione, riflettendo la commune origine del muscolo e della pelle. A supportare questo legame, una coorte di rare sindromi muscolo-cutanee dovute a mutazioni nel

percorso dei geni RAS/MAPK sono state descritte nell'uomo e viste le loro impronte CUB silimi, meriterebbero di essere studiate con la nostra strategia.

Impronte CUB nei geni causanti malattia

Comparando i valori CU nei geni causanti malattie rare, abbiamo mostrato che, sebbene i percorsi tessuto-specifici sono ancora riconoscibili, alcuni geni DC (specialmente nelmuscolo) hanno impronte CUB differenti paragonate ai geni NDC.

In generale, i geni DC mostrano valori CU meno estremi paragonati ai geni NDC, fatto molto evidente se abbiamo calcolato i valori CU senza raggruppamento gerarchico. I geni DC del muscolo e, meno evidentemente, della pelle e del rene mostrano valori più intermedi comparati ai geni NDC, suggerendo un diverso comportamento CUB nei geni causanti malattia.

Soprattutto, i geni DC del muscolo mostrano un'impronta legata alla malattia più riconoscibile, suggerendo che un CUB "orientato alla malattia" può parzialmente prevalere nelle scelte sulla tessuto-specificità CUB nel muscolo.

I geni DC e NDC della pelle e del rene mostrano differenze meno evidenti. Nondimeno, alcuni codoni, come AAG, CAG e GAG, hanno valori CU più alti nei geni DC, ipotizzando la possibilità che la loro frequenza sia orientate alla malattia.

E' noto che il CUB aumenta durante l'evoluzione e diventa estremo con la completa mancanza di rappresentazione di qualche codone, codoni che abbiamo definite "codoni zero", sebbene le ragioni di questo non siano pienamente comprese. Lw impronte CU dei geni DC del muscolo orientate alla malattia possono suggerire una diversa pressione di selezione naturale, come già identificato in alcune categorie del gene umano. Infatti, un CUB gene-specifico è stato identificato in alcuni geni di malattie umane, suggerendo un impatto sull'interpretazione di variazioni sinonime. nei geni *CFTR* e *GATA4* (le cui mutazioni causano fibrosi cistica e una malattia cardiaca congenita, rispettivamente), mutazioni sinonime possono alterare le cinetiche di traduzione e il ripiegamento delle proteine introducendo codoni rari o non-ottimali; alti valori di ACT, AGG, ATT e AGC, o AGA CUB sono stati visti in *HPRT1* (le cui mutazioni si verificano nella syndrome Lesch-Nyhn) e *GALC* (le cui mutazioni causano la malattia di Krabbe); infine i geni *BRCA1* e *BRAC2* (geni maggiormente coinvolti nel cancro mendeliano della mammella) mostrano un CUB estremamente basso paragonati ad alti oncogeni. Mettendo insieme questi e i nostri dati, possiamo ipotizzare che alcuni geni DC possono aver attraversato una differente pressione CUB durante l'evoluzione. A support di questa evidenza, è risaputo che alcuni geni di malattia hanno ancora un alto tasso di variazione, come il gene *DMD*,

che possono avere impatto sulla frequenza del tipo di codone. Da notare che , è stato mostrato molto recentemente che l'evoluzione del gene/della proteina può verificarsi non solo in termini di forze orientate all'evoluzione ma anche in relazione con le malattia che le loro mutazioni possono causare.

Pertanto, nei geni causanti malattia da noi studiati forniamo evidenza preliminare che il CUB può essere orientate alla malattia, suggerendo che comparare valori CU in questi geni, che sono più di 6000 già identificati finora negli umani, può essere una buona strategia per capire le regole che governano l'utilizzo dei codoni.

Prioritizzazione dei codoni e comportamento CUB del gene DMD

Cinque codoni, GUA (Val), GAU (Asp), GAC (Asp), CCA (Pro) e CGU (Arg) hanno mostrato i valori CU più differenti tra i geni e i tessuti dell'*HSA*.

GAC e GAU sono i codoni più utilizzati nei geni della pelle, contrariamente CAA e CGU e GUA sono i codoni meno usati nei geni del muscolo e del rene, rispettivamente. E' interessante notare che, questi ultimi tre codoni, insieme con il codone ATC (Ile), hanno una significative correlazione positive con l'espressione dei geni dovuto al loro alto contenuto di GC. Questa scoperta supporta l'uso eccessivo di alcuni codoni orientato ai tessuti, simile a quell oche abbiamo trovato per questi 5 codoni nel nostro studio, che può riflettere uno specific processo di selezione nell'HSA.

Supportando la nostra ipotesi sulle imponte CUB dei geni DC, l'utilizzo tessuto-specifico di questi 5 codoni è anch'esso orientate alla malattia. CCA e CGU sono i codoni più utilizzati nel muscolo e hanno valori CU più alti nei geni DC. Similarmente, GAC è un codone molto utilizzato nella pelle ed è più frequentemente usato nei geni DC, mentre il codono più utilizzato nella pelle, GAU, è più frequentemente usato nei geni NDC. Questa tendenza orientate alla malattia sembra essersi conservata trai mammiferi.

Alcuni report disponibili suggeriscono che una sovrarappresentazione di codoni specifici può caratterizzare i geni del disordine mendeliano e, secondo la teoria della pressione di mutazione direzionale, una pressione selettiva negative ai codoni conferendo un più altro rischio del verificarsi di mutazioni. Infine, è risaputo che alcuni geni dell'HSA mostrano un CUB più forte o più basso in base alla loro funzione. Basandoci su questa conoscenza, possiamo ipotizzare che alcuni geni DC con funzioni uniche possono aver avuto una transizione evoluzionaria differente in termini di CUB, per alcune/tutte le ragioni sopra menzionate.

Il gene DMD è l'unico gene dell'HSA che non presenta alcun "codone zero" eppure usa tutti i codoni della sua sequenza codificata, che è una tendenza conservata tra i

mammiferi. Le mutazioni DMD, incluse variazioni missenso e nonsenso, causano la distrofia muscolare di Duchenne (DMD, OMIM* 300377) una rara, grave e fatale distrofia muscolare, o la forma più mite, la distrofia muscolare di Becker (BMD; OMIM *300376), entrambe ereditate come malattie recessive legate al cromosoma X, con un incidenza di 1:5000 maschi neonati.

Tramite il nostro approccio “mapping-on-codons” delle 2828 variazioni DMD patogeniche missenso e nonsense causanti DMD o BMD, abbiamo trovato una mancanza di correlazione tra frequenza di mutazione e valori CU. I codoni DMD meno frequentemente usati (UCG, CCG, ACG e GCG) sono ricchi si GC e infatti raramente ospitano mutazioni. A ogni modo, altri codoni ricchi di GC come CGC, sono raramente usati nel *DMD* ma sono frequentemente sito di mutazioni (67) e, viceversa, il codone usato veramente più di frequente, UGC, ospita solo 9 mutazioni. CAG è il codone mutato più di frequente (56,5% con 249 mutazioni) ma non il più usato dal gene *DMD*.

Pertanto, i siti di mutazioni mappati sui codoni *DMD* non sono correlati con la frequenza di utilizzo del codone.

E’ interessante notare che, il locus di architettura del *DMD* suggerisce possibili spiegazioni del suo comportamento CUB unico. E’ risaputo che lunghezza degli introni e della proteina, pattern di espressione e valori CUB sono variant genomiche che influenzano il tasso di evoluzione dei geni. I geni con bassi livelli di espressione di mRNA/proteine tendon ad evolversi rapidamente, hanno introni grandi, codice per protein più grandi e hanno un CUB molto basso. Il gene *DMD* incontra tuute queste regole: ha introni enormi, codifica una proteina di grosso peso molecolare, che è altamente tessuto-specifica ma scarsamente abbondante e coerentemente mostra assenza di CUB estremo e nessun “codone zero”. Basandoci su queste metriche, il gene *DMD* potrebbe essersi rapidamente evoluto durante l’evoluzione, e noi possiamo affermare che un simile comportamento si è verificato in altri geni DC che mostrano simili caratteristiche di CUB basso come quelle osservate in alcuni geni DC della pelle e del rene. A ulteriore sostegno della nostra osservazione, gli eucarioti mostrano una correlazione negative tra la lunghezza del gene e il CUB, mentre nelle mosche, una relazione antagonistica tra il CUB e il numero e la lunghezza degli introni è stata descritta e, più intrigantemente, i geni senza introni janno un CUB molto alto.

Le impronte CUB nrei geni DC dovrebbero avere implicazioni sull’ottimizzazione dei codoni. Infatti, i gen DC, specialmente quelli del muscolo che abbiamo studiato, mostrano un’impronta mostrante un CUB basso. Questo significan che i geni DC del muscolo applicano ancora la piena ridindanza del codone e, viceversa, ch ei geni sintetici

corrispondenti progettati per terapie geniche avranno bisogno di una dirompente ottimizzazione del codone tramite applicazione del Codon Adaptation Index (CAI).

Al contrario, i geni NDC mostrano un CUB alto spontaneo, che porta a molti “codoni zero” nella sequenza del codice, quindi mancando molti tipi di codoni sinonimi. Dal moment oche abbiamo mostrato che i geni DC (specialmente il muscolo) hanno un CUB orientate alla malattia, la piena pplicabilità a “tutti” i geni della malattia degli algoritmi di computazione correntemente usati sviluppati per l’ottimizzazione del codone e gli approcci della terapia genica possono essere messi in dubbio (Gould N, Hendy O, & Papamichail D. Computational tools and algorithms for designing customized synthetic genes. Frontiers in bioengineering and biotechnology. (2014); 2, 41). Infatti, alcuni codoni rari non ottimali potrebbero dover essere preservati per l’efficienza di traduzione di proteine o, più interessantemente, per la regolazione dell’espressione gene- e tessuto-specifica come precedentemente riportato, sebbene non negli umani.

RIVENDICAZIONI

1. Un metodo implementato al computer per progettare una molecola sintetica di acido nucleico di un selezionato gene causante malattia espresso in uno o più tessuti di un organismo, in cui detta malattia è una malattia rara, comprendente i passaggi:

(i) raccogliere le sequenze di uno o più geni causanti malattia espressi in uno o più tessuti di un organismo, preferibilmente calcolando la conservazione del codone del gene causante malattia selezionato tra i mammiferi;

(ii) raccogliere le sequenze di una pluralità di geni non causanti malattia espressi nello stesso tessuto e organismo dei geni nel passaggio (i);

iii) determinare il calcolo indipendente della frequenza di utilizzo del codone per i 19 aminoacidi essenziali (metionina e triptofano esclusi) in ogni gene raccolto nel passaggio (i) e nel passaggio (ii);

iv) paragonare la frequenza di utilizzo del codone determinata nel passaggio (iii) in modo da ottenere il valore del bias di utilizzo del codone, per identificare codoni a comparsa (codoni tessuto-specifici e codoni gene-specifici) in modo da prioritizzare i codoni più diversamente usati nel gene e nel/i tessuto/i di interesse;

v) progettare la molecola sintetica di acido nucleico di detto gene causante malattia, modificando la struttura secondaria o terziaria di detto gene utilizzando i codoni prioritizzati ottenuti nel passaggio iv),

in cui detto passaggio vi) di prioritizzare i codoni usati più differentemente è svolto attraverso il raggruppamento dei valori di frequenza di utilizzo del codone ottenuti nel passaggio iii) utilizzando un algoritmo di raggruppamento gerarchico con un valore p minore di 0.05 e selezionando i codoni meno usati e/o più usati per le specie, il tessuto e il gene di interesse, e
in cui detto passaggio v) di modificare la struttura secondaria o terziaria di detto gene causante malattia è eseguito tramite la sostituzione e/o rimozione dei codoni meno utilizzati e i codoni più utilizzati, ottenuti nel passaggio iv), nella sequenza di detto gene causante malattia.

2. Il metodo secondo la rivendicazione 1, in cui detto organismo è un mammifero.

3. Il metodo secondo qualsiasi delle rivendicazioni 1 o 2, in cui detto tessuto è selezionato tra muscolo, pelle, rene o qualsiasi altro tessuto coinvolto nella malattia gestibile con terapia genica.

4. Il metodo secondo qualsiasi delle rivendicazioni da 1 a 3, in cui detta malattia è selezionata tra distrofie muscolari, miopatie congenite, malattia renale tubulointerstiziale, rene policistico di tipo 1, ipercheratosi epidermolitica, displasia ectodermica.

5. Il metodo secondo qualsiasi delle rivendicazioni da 1 a 4, in cui il numero di detti geni raccolti nel passaggio (i) e/o nel passaggio (ii) è il più alto numero possibile nella banca dati per la malattia selezionata, o almeno dall'80 al 99%.

6. il metodo secondo qualsiasi delle rivendicazioni da 1 a 5, in cui detto passaggio iii) di determinare la frequenza di uso del codone è svolta tramite il calcolo del Codon Adaptation Index (CAI).

7. Il metodo secondo qualsiasi delle rivendicazioni da 1 a 6, comprendente un ulteriore passaggio di calcolo della conservazione del codone del selezionato gene causante malattia tra i mammiferi, in modo da utilizzare detto calcolo per ottenere il valore del bias di uso del codone.

8. Il metodo secondo qualsiasi delle rivendicazioni da 1 a 7, in cui detto mammifero è selezionato da uno o più tra R. ferrumequinum (ferro di cavallo maggiore), M. musculus (topo), F. catus (gatto), C. lupus familiaris (cane), E. caballus (cavallo), B. Taurus (toro), M. murinus (lemure topo grigio), G. variegatus (colugo della sonda), C. jacchus (uistitì), M. mulatta (macaco), N. leucogenys (gibbone), P. abelii (orangotango), G. gorilla (gorilla), P. troglodytes (scimpanzè) e H. sapiens (umano).

9. Un metodo per preparare una molecola sintetica di acido nucleico di un selezionato gene causante malattia, comprendente i passaggi di qualsiasi delle rivendicazioni da 1 a 8 e un ulteriore passaggio vi) di sintetizzare una molecola di acido nucleico comprendente la sequenza di acido nucleico ottimizzata per il codone del passaggio (v).

10. Un programma per computer comprendente istruzioni che, quando il programma è eseguito da un computer, comporta che il computer svolga i passaggi del metodo secondo le rivendicazioni 1-8.

11. Un supporto di memorizzazione leggibile dal computer comprendente istruzioni che, quando eseguito da un computer, comporta che il computer svolga i passaggi del metodo secondo le rivendicazioni 1-8.

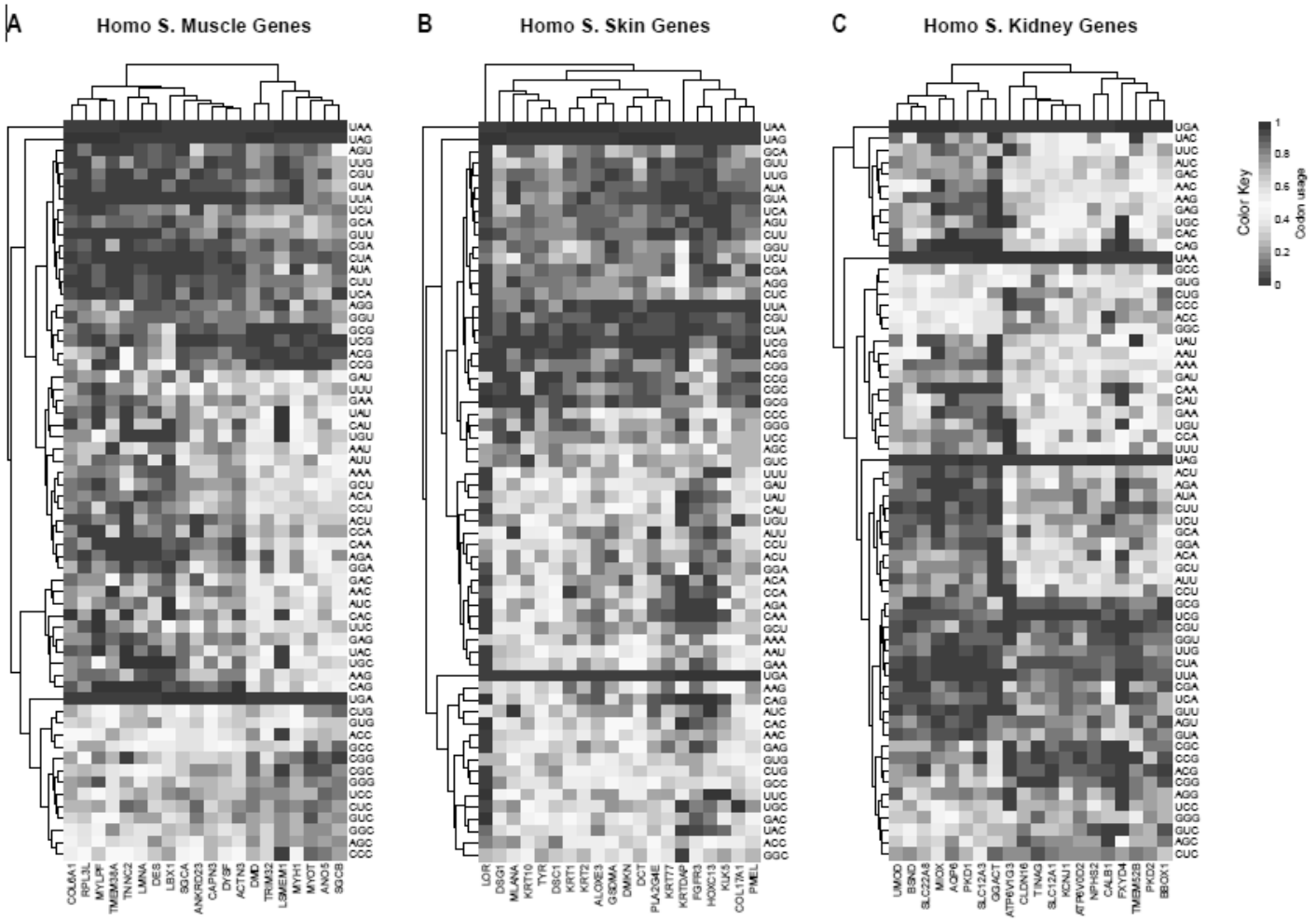

Figure 1

Inglese	**Italiano**
Muscle Genes	Geni del muscolo
Skin Genes	Geni della pelle
Kidney Genes	Geni del rene
Color key	Chiave colore
Codon Usage	Utilizzo del codone

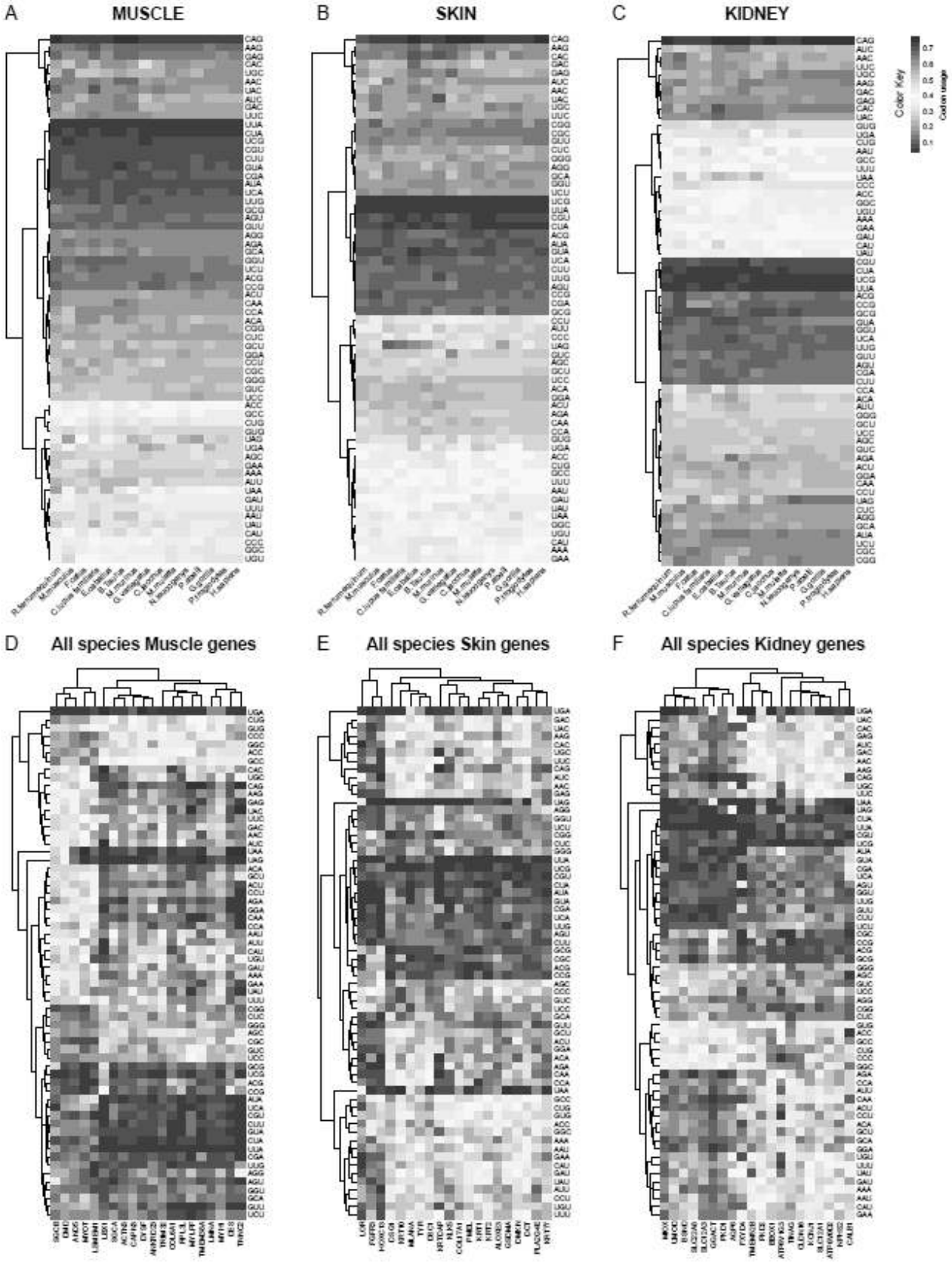
A
MUSCLE
B
SKIN
C
KIDNEY
Color Key
Codon Usage
D
All species Muscle genes
E
All species Skin genes
F
All species Kidney genes

Figure 2

Inglese	**Italiano**
Muscle Genes	Geni del muscolo
Skin Genes	Geni della pelle
Kidney Genes	Geni del rene
Color key	Chiave colore
Codon Usage	Utilizzo del codone
Muscle	Muscolo
Skin	Pelle
Kidney	Rene
All species	Tutte le specie

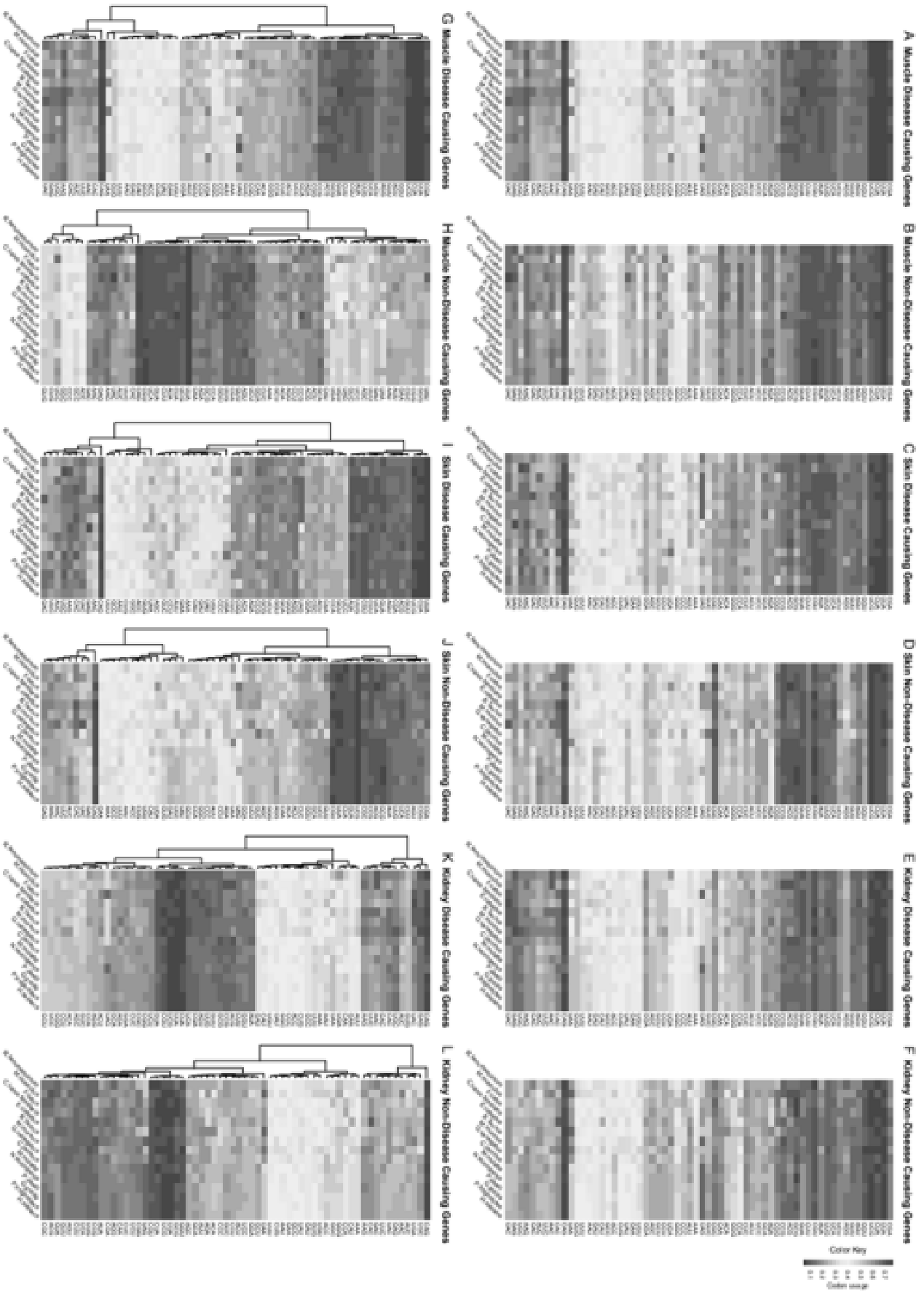
A Muscle Disease Causing Genes
B Muscle Non-Disease Causing Genes
C Skin Disease Causing Genes
D Skin Non-Disease Causing Genes
E Kidney Disease Causing Genes
F Kidney Non-Disease Causing Genes
G Muscle Disease Causing Genes
H Muscle Non-Disease Causing Genes
I Skin Disease Causing Genes
J Skin Non-Disease Causing Genes
K Kidney Disease Causing Genes
L Kidney Non-Disease Causing Genes
Color Key

Figure 3

Inglese	**Italiano**
Muscle disease causing Genes	Geni del muscolo causanti malattia
Skin disease causing Genes	Geni della pelle causanti malattia
Kidney disease causing Genes	Geni del rene causanti malattia
Color key	Chiave colore
Codon Usage	Utilizzo del codone
Muscle non-disease causing Genes	Geni del muscolo non causanti malattia
Skin non-disease causing Genes	Geni della pelle non causanti malattia
Kidney non-disease causing Genes	Geni del rene non causanti malattia

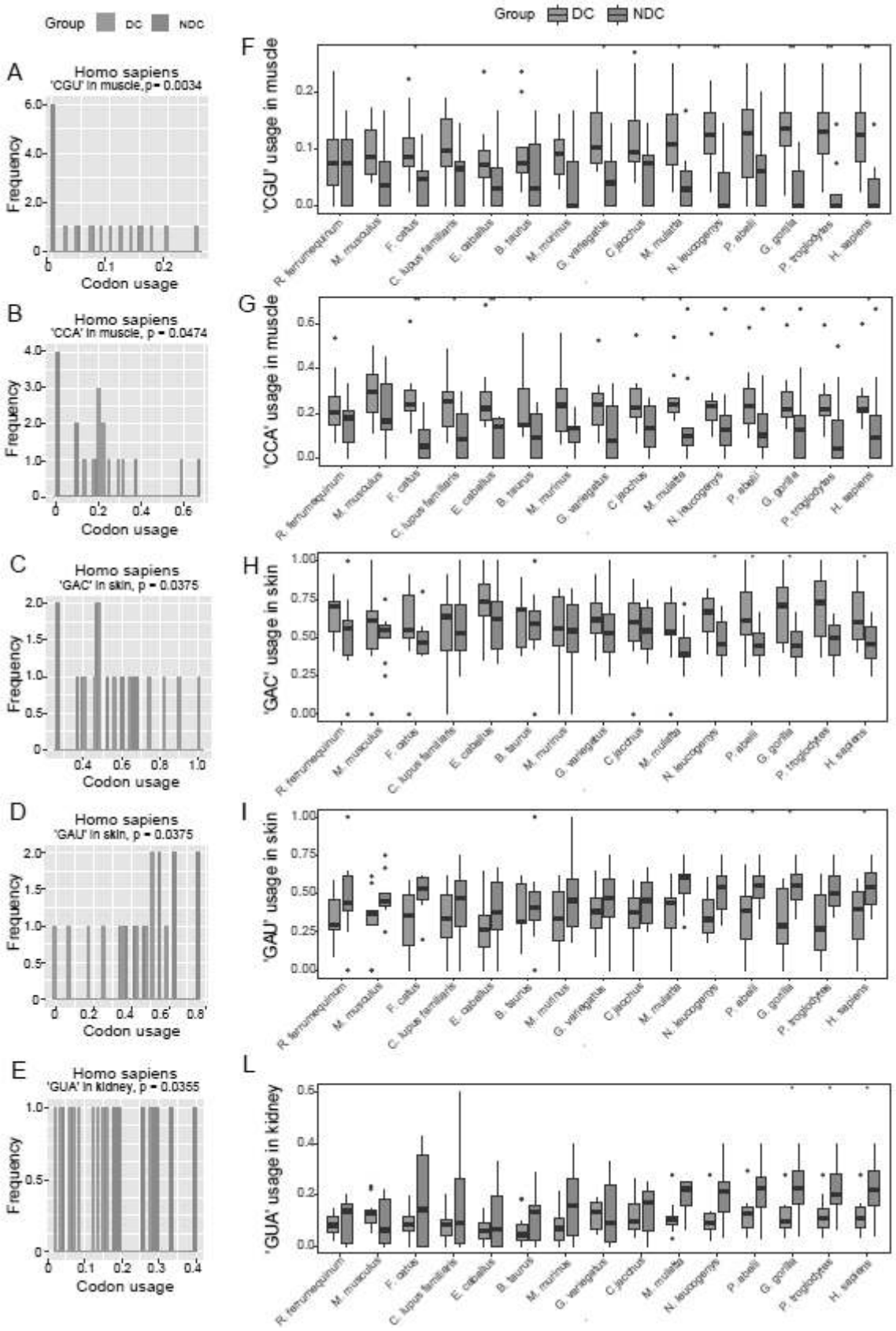

Figure 4

Inglese	Italiano
Group	Gruppo
DC (Disease causing)	Causante la malattia

NDC (Non-Disease causing)	Non causante la malattia
Usage in muscle	Uso nel muscolo
Usage in skin	Uso nella pelle
Usage in kidney	Uso nel rene
Frequency	Frequenza
Codon usage	Utilizzo del codone

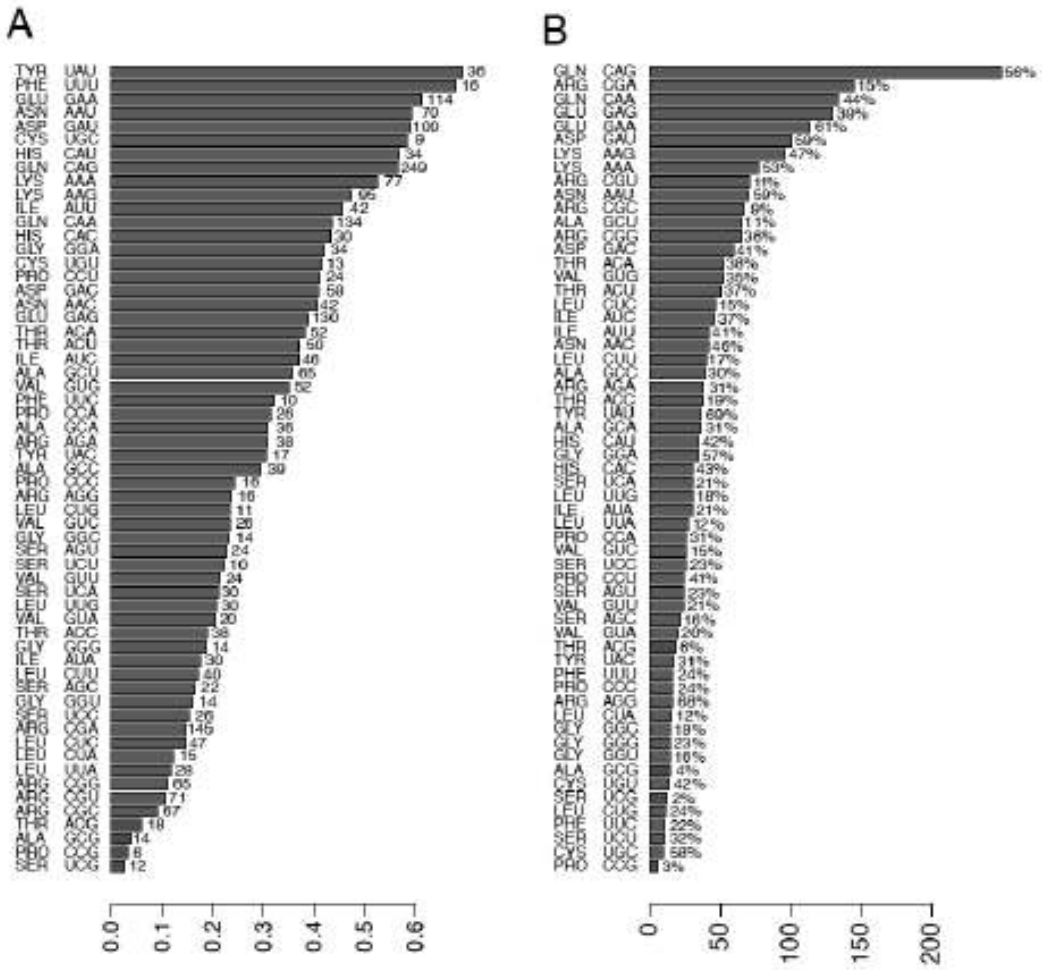

Figure 6

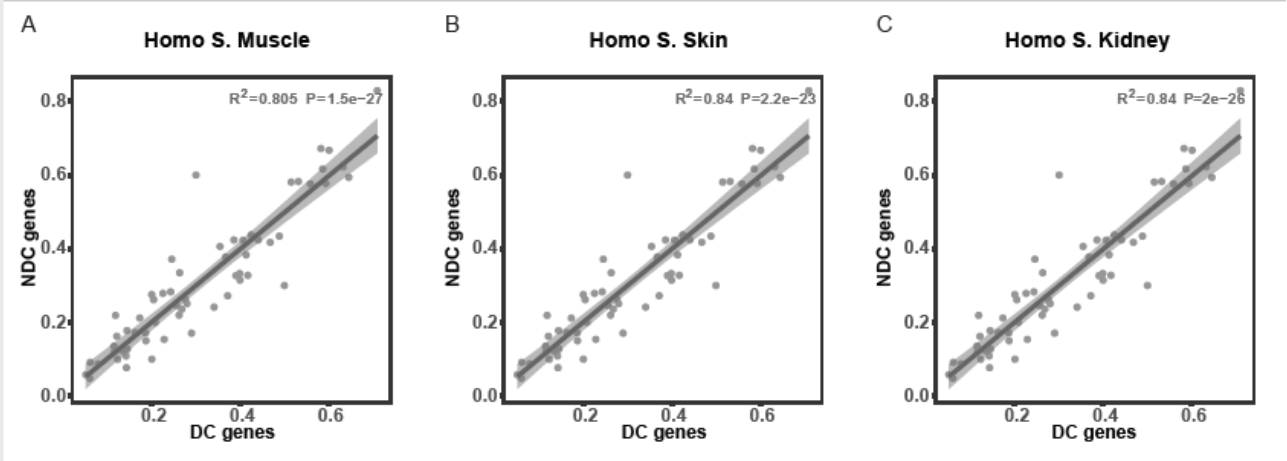

Figure 6

Inglese	**Italiano**
Homo S. Muscle	Muscolo Homo S.
Homo S. Skin	Pelle Homo S.
Homo S. Kidney	Rene Homo S.
DC (disease causing) genes	Geni causanti la malattia
NDC (Non-disease causing) genes	Geni non causanti la malattia

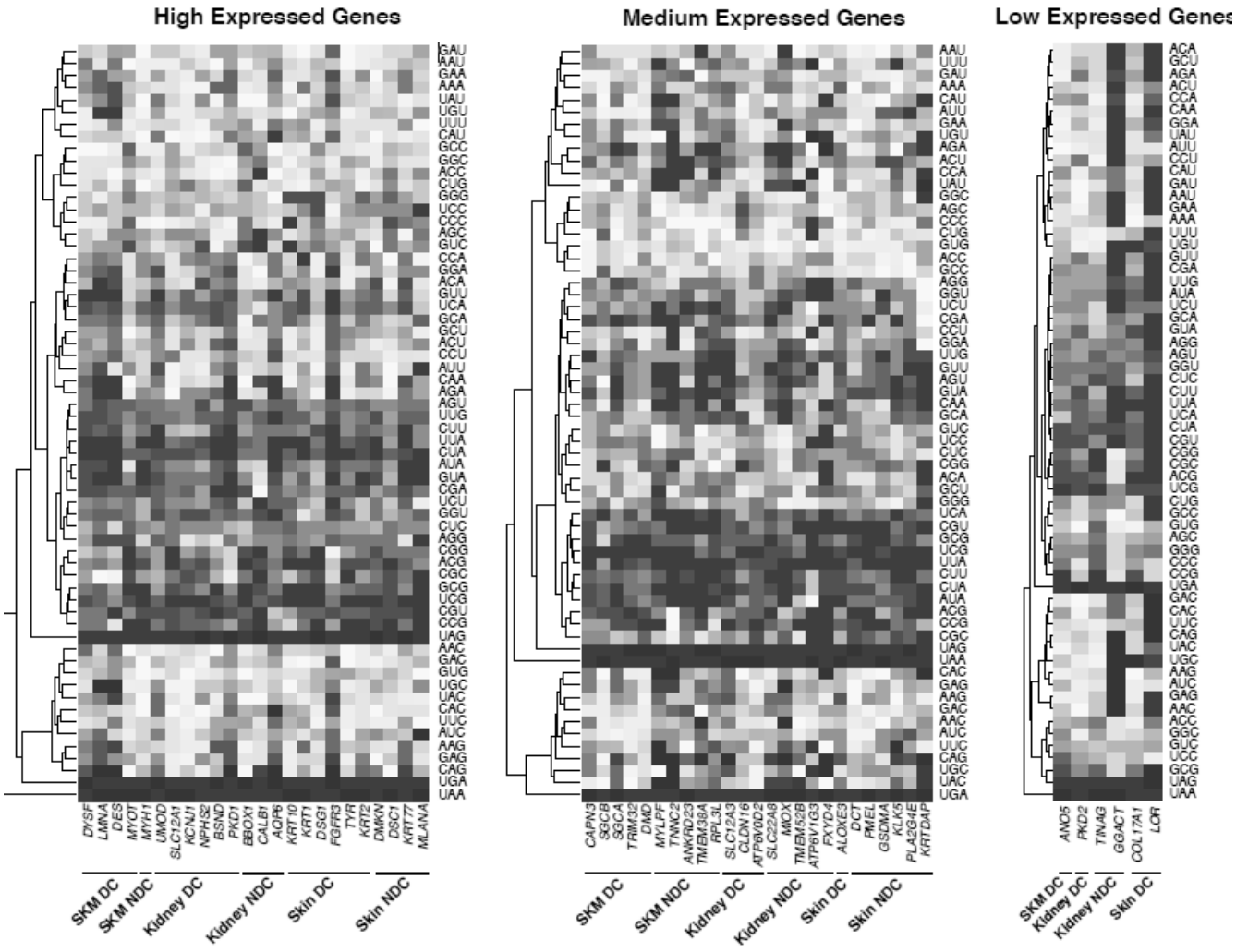

Figure 7

Inglese	**Italiano**
High Expressed Genes	Geni altamente espressi
Medium Expressed Genes	Geni mediamente espressi
Low Expressed Genes	Geni poco espressi
Kidney DC	Rene causante malattia
Kidney NDC	Rene non causante malattia
Skin DC	Pelle causante malattia
Skin NDC	Pelle non causante malattia

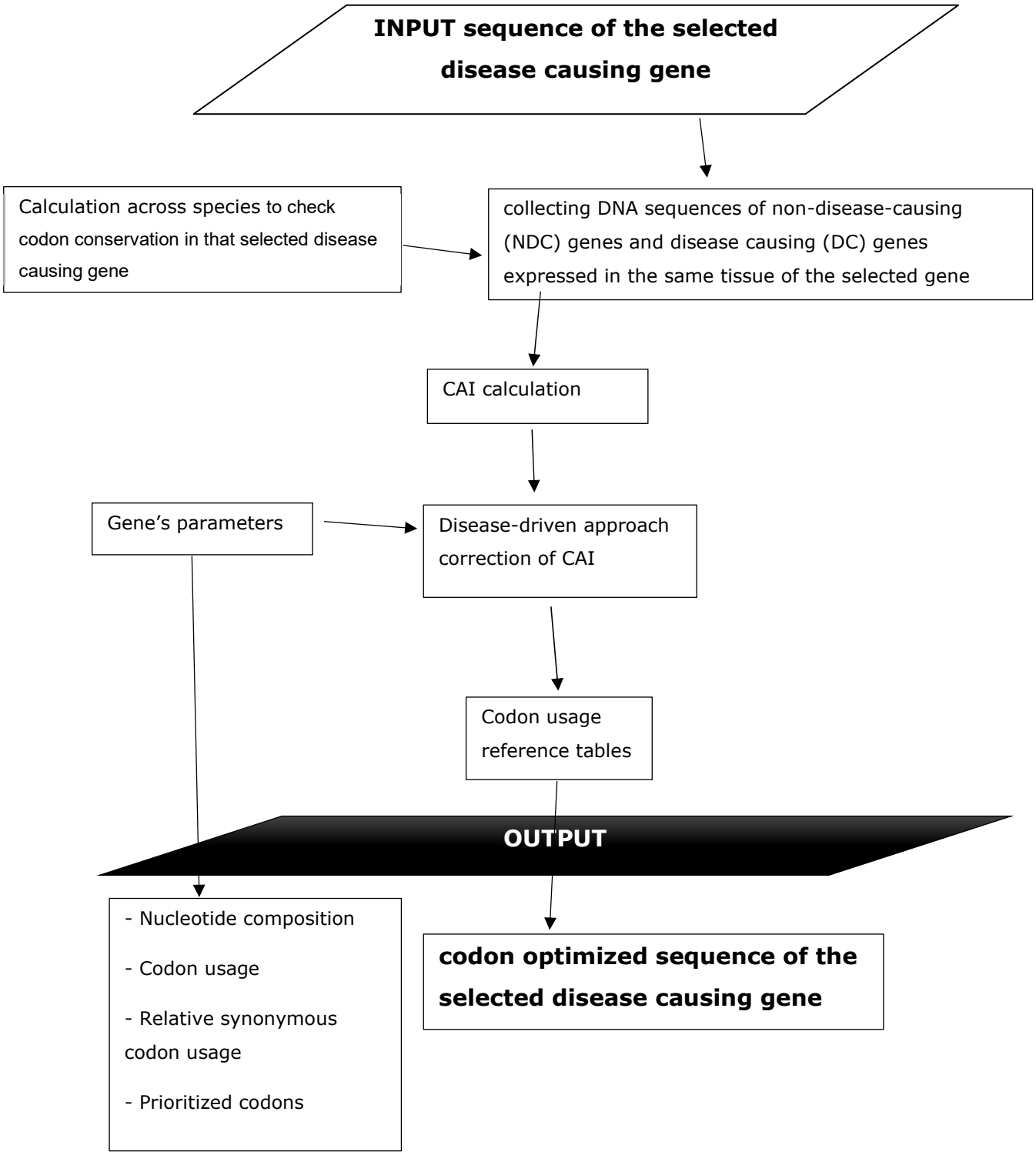
INPUT sequence of the selected disease causing gene
Calculation across species to check codon conservation in that selected disease causing gene
collecting DNA sequences of non-disease-causing (NDC) genes and disease causing (DC) genes expressed in the same tissue of the selected gene
CAI calculation
Gene's parameters
Disease-driven approach correction of CAI
Codon usage reference tables
OUTPUT
- Nucleotide composition
- Codon usage
- Relative synonymous codon usage
- Prioritized codons
codon optimized sequence of the selected disease causing gene

FIGURE 8

Inglese	**Italiano**
sequence of the selected disease-causing gene	sequenza del gene selezionato che causa la malattia
Calculation across species to check codon conservation in that selected disease causing gene	Calcolo tra specie per verificare la conservazione del codone in quel gene selezionato che causa la malattia
collecting DNA sequences of non-disease-causing (NDC) genes and disease causing (DC) genes expressed in the same tissue of the selected gene	raccogliere sequenze di DNA di geni non causanti la malattia (NDC) e geni causanti la malattia (DC) espressi nello stesso tessuto del gene selezionato
CAI calculation	Calcolo CAI
Gene's parameters	I parametri del gene
Disease-driven approach correction of CAI	Approccio guidato dalla malattia correzione del CAI
Codon usage reference tables	Tabelle di riferimento sull'utilizzo dei codoni
codon optimized sequence of the selected disease causing gene	sequenza ottimizzata del codone del gene selezionato che causa la malattia
Nucleotide composition	Composizione nucleotidica
Codon usage	Utilizzo del codone
Relative synonymous codon usage	Uso del relativo sinonimo del codone
Prioritized codons	Codoni prioritari

Ministero delle Imprese e del Made in Italy

DIPARTIMENTO MERCATO E TUTELA
DIREZIONE GENERALE PER LA PROPRIETÀ INDUSTRIALE - UIBM

ATTESTATO DI BREVETTO PER INVENZIONE INDUSTRIALE

Il presente brevetto viene concesso per l'invenzione oggetto della domanda:

N. 102022000006119

TITOLARE/I:	• UNIVERSITÀ DEGLI STUDI DI FERRARA 100.0% Di Giovine Paolo
DOMICILIO:	Società Italiana Brevetti S.p.A. piazza di Pietra 39 00186 Roma
INVENTORE/I:	• Ferlini Alessandra • Rossi Rachele
TITOLO:	Metodo per implementare la progettazione di molecole di acido nucleico sintetico per terapie geniche in malattie rare
CLASSIFICA:	G16B
DATA DEPOSITO:	29/03/2022

Roma, 13/03/2024

Il Dirigente
Loredana Guglielmetti