

## **Allegato PROGETTO di sequenziamento del genoma di Coffea arabica**

Il progetto si è posto dunque l'obiettivo di decodificare il DNA di Coffea arabica che rappresenta il 70% della produzione mondiale di caffè. Il genoma di Arabica è costituito da circa 2,6 miliardi di basi per una lunghezza fisica di circa 140 cm. Viste queste dimensioni, è stato necessario suddividere le attività in varie fasi, ciascuna delle quali è stata accompagnata dallo sviluppo di strategie diverse. Si tratta di un progetto di ricerca molto complesso che ha previsto per la prima volta il sequenziamento e la ricostruzione del genoma di un organismo tetraploide, cioè 4n (quartetti di cromosomi omologhi), derivante dall'unione dei due progenitori, verosimilmente **Coffea canephora e Coffea eugenioides**.

Le cinque fasi principali in cui si è articolata la ricerca erano: 1) Preparazione di una genoteca BAC; 2) Sequenziamento fisico del DNA dei BAC; 3) Sequenziamento fisico del DNA genomico intero; 4) Ricostruzione informatica delle sequenze; 5) Annotazione ed identificazione dei geni codificanti.

### **1) Preparazione di una genoteca BAC**

La grandezza e la complessità del DNA di Coffea arabica rendono complesso il sequenziamento diretto dell'intero genoma soprattutto a causa della sua natura poliploide che porta alla presenza di coppie di sequenze molto simili ma non identiche, ciascuna derivante da uno dei due genomi ancestrali. Per tale ragione è stata scelta una innovativa strategia ibrida che combina, assieme al tradizionale sequenziamento diretto dell'intero genoma, il sequenziamento di cloni BAC. E' stato quindi necessario frammentare il genoma in tanti segmenti più piccoli, ciascuno dei quali è stato successivamente inserito in un vettore, costituito da cromosomi artificiali batterici (BAC: Bacterial Artificial Chromosome). Questa strategia ha consentito di ottenere una collezione di colonie batteriche (cloni), contenente almeno una copia di tutte le sequenze di DNA presenti nel genoma. Tale collezione è definita libreria genomica o genoteca. Questa prima fase si è conclusa con la costituzione di una genoteca BAC pari a circa 175.000 cloni, ciascuno dei quali contiene un frammento di DNA di Arabica di circa 100.000 nucleotidi.

### **2) Sequenziamento fisico del DNA dei BAC**

Ciascun clone è stato posto in coltura in modo che le cellule batteriche si moltiplicassero e automaticamente moltiplicassero le copie del frammento di DNA di Arabica producendo una quantità di DNA sufficiente alle esigenze del processo di sequenziamento. Si è quindi proceduto alla fase di sequenziamento vero e proprio, che per motivi di costo, nonché di ridondanza della genoteca, ha riguardato soltanto 36.864 dei 175.000 cloni ottenuti, riuniti a loro volta in 96 gruppi (o pools) contenenti ciascuno 384 cloni.

### **3) Sequenziamento fisico del DNA genomico intero**

E' stata adottata una strategia che ha previsto la produzione di collezioni di frammenti di DNA genomico di diverse dimensioni, ciascuno dei quali è stato sequenziato ad entrambe le estremità fino ad ottenere circa 100 gigabasi di sequenza complessiva.

### **4) Ricostruzione informatica delle sequenze**

Al termine del sequenziamento si ottiene un'ampia collezione di piccole sequenze casuali, che devono essere assemblate ad ottenere una sequenza consenso che copra l'intero genoma. Perché questo sia possibile, i dati di sequenza bruti vengono innanzitutto ripuliti dalle sequenze del DNA del vettore batterico in modo da avere solo le sequenze dei frammenti di DNA di Coffea arabica. La sequenza di ogni singolo frammento di Arabica è stata quindi confrontata con la sequenza di tutti gli altri frammenti per stabilire se la coda di un frammento corrispondesse alla testa di qualche altro frammento e potessero essere riuniti in una ricostruzione almeno parziale del genoma. Questo processo ha consentito la ricostruzione di oltre il 50% del DNA di Arabica ed è in corso il lavoro che, combinando le due strategie descritte ai punti 2 e 3, dovrebbe consentire di avvicinarsi ad una ricostruzione completa.

### **5) Annotazione e identificazione dei geni codificanti**

Questa fase del progetto consiste principalmente nell'applicazione di diversi sistemi bioinformatici per l'analisi dei dati di sequenziamento genomico, in particolare per quanto riguarda la predizione e l'annotazione genica. La predizione consiste nell'identificazione dei geni, ovvero di quelle parti del DNA codificanti per le proteine che sono effettivamente responsabili delle caratteristiche della pianta e del chicco. Per annotazione si intende invece la descrizione funzionale delle proteine codificate da ogni singolo gene identificato nel processo di predizione.

Il percorso che porta al riconoscimento dei geni a livello del DNA prende in considerazione metodi diversi.

Passaggio successivo è, infine, la caratterizzazione dei geni predetti, in modo da assegnare loro una possibile funzione biologica. Questa fase, nota come annotazione, necessita di considerevoli risorse computazionali che consentono di eseguire confronti con diversi database di sequenze proteiche, domini proteici e motivi funzionali già noti. Al momento è stato stilato un elenco preliminare dei geni predetti, basato sulla ricostruzione del genoma ad oggi disponibile. Tale caratterizzazione dei geni è incrementale e procede man mano che si aggiungono nuovi elementi alla ricostruzione del genoma.

### **Conclusione**

Il sequenziamento del genoma di *Coffea arabica* fa parte di un ambizioso programma di ricerca che si pone come obiettivo la ricostruzione, finora mai affrontata, dell'intero genoma di un organismo tetraploide, i cui due genomi ancestrali sono estremamente simili.

L'obiettivo, per quanto difficile, è quello di poter distinguere le sequenze derivate dal progenitore *C. eugenioides* da quelle derivanti dal progenitore *C. canephora*. Una buona conoscenza del materiale genetico è di fondamentale importanza per la qualità della pianta, e quindi del seme, in quanto tutte le sue strutture e funzioni sono dettate dalla sequenza di nucleotidi del DNA che determina, a sua volta, la sequenza degli amminoacidi nelle molecole proteiche.

Queste considerazioni hanno rappresentato il punto di partenza del progetto, che ha portato ad oggi ad una buona decifrazione della base genetica della pianta del caffè Arabica. Sono stati infatti sequenziati cloni BAC ed è stata effettuata una prima ricostruzione del genoma. Sono stati, inoltre, sequenziati dei trascritti di foglie e radici di alcune varietà di Arabica per poter identificare i geni codificanti, ovvero quei tratti di DNA effettivamente attivi dai quali dipendono le caratteristiche della pianta e quindi del chicco di caffè. Potenti mezzi di bioinformatica stanno, infine, lavorando per identificare le possibili funzioni dei geni codificanti. Dal sequenziamento del genoma di Arabica sono attese ricadute in vari settori, non ultimo quello economico, con possibili applicazioni agronomiche ed industriali. In più, le analisi genomiche strutturali e funzionali potranno permettere di comprendere come i due genomi di *C. eugenioides* e *C. canephora* si combinano ed interagiscono in quello di *C. arabica* per conferirgli le sue caratteristiche distintive che non rappresentano la semplice somma delle caratteristiche delle due specie ancestrali e che lo rendono così ricercato per la produzione della bevanda caffè.

### **Altre ricerche**

Il sequenziamento del genoma del caffè (*Coffea canephora*) ha messo in luce la singolare evoluzione della capacità di produrre caffeina della pianta e ha chiarito alcuni complessi rapporti fra numerosi diversi geni. Entrambi i risultati permetteranno un miglioramento sia delle qualità organolettiche del prodotto sia la resistenza della pianta del caffè alle malattie.

La ricerca – [pubblicata su "Science"](#) e condotta da un gruppo internazionale di ricercatori afferenti a diversi istituti, fra cui il centro dell'Enea della Casaccia e l'Università di Trieste – si è concentrata sulla specie *C. canephora*, da cui deriva la varietà robusta del caffè, e non su *Coffea arabica*, quella domesticata in tempi più antichi e considerata più pregiata, perché la prima è diploide (ossia ha due copie per ciascun cromosoma), mentre *C. arabica* è tetraploide (quattro copie del corredo cromosomico), una circostanza che rende molto più complessa la lettura e l'interpretazione dei dati ottenuti.

La prima sorpresa che ha riservato il sequenziamento è stata la scoperta del fatto che rispetto ad altre specie di piante, come per esempio il pomodoro o l'uva, il caffè ha un numero molto più elevato di grandi classi di geni responsabili della produzione di alcaloidi e flavonoidi. Per esempio, il principale acido grasso insaturo dei semi di caffè, l'acido linoleico, che contribuisce significativamente al suo aroma, è controllato da ben sei geni, contro l'unico gene destinato a questo scopo presente nella maggior parte delle piante che lo producono.

Ma la sorpresa più grande è venuta dall'analisi dei percorsi di biosintesi della caffeina da parte di *C. canephora*. I ricercatori hanno scoperto che la caffeina è prodotta in modo differente rispetto ad

altre piante che contengono questa sostanza, come cacao e tè. In altre parole, la capacità di sintetizzare la caffeina non è stata ereditata da un antenato comune di caffè, cacao e tè, ma si è evoluta in momenti diversi e in maniera indipendente.

Si ritiene che la caffeina aiuti le piante a tenere lontano i parassiti e a frenare la crescita di altri vegetali che competono per le risorse. [Un altro recente studio](#) ha dimostrato che, al pari degli umani, gli insetti impollinatori possono sviluppare l'abitudine alla caffeina, inducendoli a prediligere le piante che sintetizzano questo composto.

In [un articolo di commento](#) al sequenziamento, Dani Zamir della Hebrew University of Jerusalem a Rehovot, sottolinea l'importanza di tradurre il sequenziamento del genoma del caffè in un'opportunità per la coltivazione della pianta, in un momento in cui la diversità delle sue varietà sta diminuendo a livello globale sotto la pressione di cambiamenti climatici, deforestazione e diffusione di malattie. In particolare, Zamir osserva che sarebbe importante sviluppare e condividere una mappa che colleghi i dati genetici ottenuti e quelli fenotipici, in particolare quelli relativi a tratti come aroma e sapore.

In questo modo sarebbe possibile sfruttare la ricca riserva di diversità genetica ancora disponibile nelle varietà di caffè meno sfruttate presenti in Africa, terra d'origine della pianta, per assicurare un futuro a un settore agricolo che per molti paesi tropicali rappresenta la principale fonte di guadagno. L'attuale produzione mondiale di caffè arabica, ricorda Zamir, si basa infatti su un piccolo numero di cultivar con una diversità genomica e fenotipica estremamente ridotta.

Nel 2017 anche la bevanda più amata dagli italiani ha un DNA e in questi giorni è stato in parte svelato: alcuni ricercatori della California hanno sequenziato il **genoma del caffè Arabica**. Un traguardo importante a meno di tre anni dalle scoperte sulla *Coffea Robusta*. I risultati sono stati resi noti durante la *Conferenza sul Genoma Animale e Vegetale 2017*, tenutasi a San Diego in Gennaio 2017. Lo studio sul genoma del caffè Arabica è stato sviluppato presso la [UC Davis](#), università della California da sempre incentrata sulla genetica. Per l'analisi è stato scelto un team diversificato nelle competenze: **Juan Medrano**, genetista animale e promotore dello studio; **Allen Van Deynze**, direttore di ricerca per il dipartimento di Biotecnologie della UC Davis; **Dario Cantù**, genetista specializzato in viticoltura; **Amanda Hulse-Kemp**, ricercatrice. Agli scienziati si è aggiunto **Jay Ruskey**, il primo imprenditore a creare una piantagione industriale di caffè negli Stati Uniti continentali. È stata proprio la **Good Land Organics** di Santa Barbara, la piantagione di Ruskey, il terreno dal quale sono stati prelevati i campioni di materiale genetico. Tra tutte le piante la scelta è ricaduta sulla **varietà Geisha**, una pregiata tipologia di caffè Arabica proveniente dall'Etiopia. Il caffè Arabica è uno dei più pregiati e, sotto forma di monorigine o in miscela, copre il **70% della produzione mondiale**. La *Coffea Arabica* è un ibrido derivato dall'incrocio tra la *C. Canephora* (comunemente definita robusta) e la *C. Eugenioides*: questo fa sì che il suo genoma sia formato da quattro serie di cromosomi, a differenza della maggior parte delle piante. Il risultato dello studio ha scoperto 70.830 geni e **1,19 miliardi di paia di basi**, pari a circa 1/3 di quelli dell'uomo. I risultati dello studio di Medrano hanno risvolti su più fronti. È indubbio come una ricerca di questo tipo possa aiutare i coltivatori a conoscere meglio e a migliorare le proprie piante, ma sono le prospettive future a destare il maggiore interesse. Il sequenziamento della *Coffea Arabica* contiene informazioni cruciali per sviluppare varietà di caffè di alta qualità, **resistenti alle malattie** e in grado di **adattarsi ai cambiamenti climatici** che minacciano la produzione dei prossimi 30 anni. Si pongono quindi le basi per **produzioni meno a rischio**, in grado quindi di stabilizzare anche i prezzi delle materie prime in futuro. È possibile visionare il genoma su [Phytozome.net](#) (il database del Dipartimento dell'Energia degli Stati Uniti) e il report completo sul sito di [UC Davis](#).

Anche l'azienda giapponese Suntory, specializzata nella produzione di bevande tra le quali il caffè, ha finanziato una ricerca tesa a decodificare il DNA del caffè, **specie *Coffee Arabica* condotta da un team di ricercatori della University of California, Davis (UC Davis)**.

I risultati sono stati annunciati a febbraio del 2017 alla 25<sup>a</sup> **Plant and Animal Genome Conference** di San Diego e resi pubblici sul sito Phytozome, **il portale di genomica del Joint Genome Institute del Dipartimento dell'Energia americano**. I ricercatori hanno prelevato campioni da 23 piante di caffè Arabica denominata Geisha, originaria dell'Etiopia, e da essi hanno estratto il genoma da sequenziare, scoprendo che esso è costituito da 1,19 miliardi di basi e 70.830 geni stimati, circa un terzo del DNA umano.

I risultati della ricerca sono molto importanti per il futuro della varietà Arabica, le cui coltivazioni sempre più a rischio a causa dei cambiamenti climatici e dell'azione di parassiti che ogni anno determinano la distruzione di ettari di piantagioni.