Rendiconto OLEA – I Anno

U.O. ENEA

Responsabile: Gaetano Perrotta

WorkPackage 3 – Genomica Funzionale

Obiettivi della ricerca

La principale finalità per il primo anno di progetto era la ma messa a punto e l'esecuzione di azioni per determinare l'ambito funzionale di geni e proteine coinvolti nell'espressione di caratteri di interesse agronomico.
Gli obiettivi specifici comprendevano:
- caratterizzazione ed assemblaggio de novo del trascrittoma attraverso il sequenziamento ultra-massivo di ESTs e la descrizione di repertori di geni espressi su larga scala, in almeno due cultivar con caratteristiche contrastanti a partire dai seguenti organi: fiore, frutto in diversi stadi di sviluppo, foglia, gemma, internodo, apice radicale;
- Identificazione di marcatori SNP tra varietà a livello di singoli geni per caratterizzare la variabilità genetica totale entro la specie e per sviluppare nuovi marcatori da impiegare nel sistema di tracciabilità dei prodotti olivicoli;
- messa a punto di una piattaforma microarray per la determinazione dei livelli di espressione genica ad alta risoluzione e l'analisi dell'espressione genica su larga scala per lo studio comparato di target biologici di interesse.
- analisi dell'espressione proteica su larga scala per la determinazione dell'assetto proteico durante la maturazione del frutto.

Attività svolte

3.1 Caratterizzazione ed assemblaggio de novo del trascrittoma attraverso il sequenziamento ultra-massivo di ESTs

La determinazione del repertorio di geni espressi in olivo è stata eseguita su campioni di fiori, drupa, figlie e radici.
Sono stati utilizzati fiori di tre cultivar (cv), Leccino, Frantoio e Dolce Agogia per identificare geni espressi durante diverse fasi di sviluppo e per evidenziare i profili di espressione genica legati al processo dell'aborto dell'ovario ed al fenomeno dell'auto-incompatibilità.
La caratterizzazione su larga scala di EST da drupa è stata programmata in primo luogo per integrare dati già disponibili, ottenuti dal nostro laboratorio recentemente, sulle cv Coratina e Tendellone. Nell'ambito del progetto Olea si è scelto di proseguire il sequenziamento utilizzando drupe delle cv Ortice e Ruveia per le loro particolari caratteristiche di resistenza/suscettibilità all'attacco di Bactrocera oleae. La determinazione dei geni espressi da questi campioni sarà, dunque, utilizzabile anche per studiare i meccanismi molecolari legati alla suscettibilità e risposta all'attacco del patogeno.
Infine sono stati caratterizzati i trascritti da campioni di foglia e radice della cv Leccino e di una sua variante "dwarf".

Sequenziamento 454 di campioni di cDNA

I campioni di cDNA utilizzati per la costruzione delle librerie da sequenziare sono stati preparati dai partner CNR-IGV e da UNITUS e ottenuti da RNA estratti da fiori, drupe, foglie e radici:
1) Leccino fiore (stadi 1-6 in pool)
2) Leccino fiore (stadi 7-8 in pool)
3) Dolce Agogia fiore (stadi 1-6 in pool)
4) Frantoio fiore (stadi 7-8 in pool)
5) Ortice drupa (undamaged)
6) Ortice drupa (damaged vari stadi in pool)
7) Ruveia drupa (undamaged)
8) Ruveia drupa (damaged vari stadi in pool)
9) Leccino foglia wt
10) Leccino foglia dwarf
11) Leccino radice wt
12) Leccino radice dwarf

Per ciascuna delle dodici librerie *shotgun* sono stati utilizzati 500 ng di cDNA. Ogni cDNA è stato frammentato mediante nebulizzazione con azoto ad alta pressione e utilizzando reagenti ROCHE (GS FLX Titanium Rapid Library Preparation Kit). L'acido nucleico nebulizzato è stato poi recuperato e purificato per selezionare i frammenti di lunghezza compresa tra 500 e 800 nucleotidi.
Dopo avere valutato la qualità dei frammenti, mediante elettroforesi su chip (2100 Bioanalyzer - Agilent), ciascuna libreria è stata quantificata tramite Nano-Fluorimetro e diluita alla concentrazione di 1 x 107 molecole/µl. Il protocollo di nebulizzazione dei campioni 8-12 è stato modificato in modo da ottenere frammenti più lunghi (1400-1800 nucleotidi) compatibili con la versione aggiornata del sistema di sequenziamento Roche GS FLX+ che genera tipicamente sequenze con una lunghezza media incrementata a circa 750 nucleotidi.
La quantità ottimale di ogni libreria da utilizzare per la fase successiva di amplificazione, determinata mediante una reazione di PCR in emulsione, è risultata compresa fra 1 e 4 molecole/biglia.
In seguito, i prodotti della reazione di amplificazione in emulsione sono stati sottoposti ad un processo di arricchimento in modo tale da selezionare solo le biglie contenenti un prodotto di amplificazione e di conseguenza sequenziabile. Dopo la fase di arricchimento, è stata stimata, tramite un contatore di microparticelle (bead-counter), la percentuale di biglie recuperate che legavano il cDNA amplificato. I valori ottenuti, compresi fra il 15% il 20% sono indice di una buona qualità delle librerie amplificate.
Le dodici librerie (circa 750.000 biglie per ognuna delle librerie) sono state utilizzate per tre corse di sequenziamento 454 (Roche) su GS FLX (o GS FLX nel caso delle librerie 9-12) Titanium PicoTiterPlate suddivise in quattro settori, ciascuna.

Assemblaggio de-novo e annotazione funzionale del trascrittoma della drupa e del fiore
Al fine di generare un repertorio completo di geni espressi nel frutto di olivo, le sequenze ottenute dalle quattro librerie sono state opportunamente filtrate ed assemblate mediante l'uso del software GS-Assembler v2.5.3 (Roche). I geni espressi sono stati poi annotati e classificati funzionalmente confrontando le sequenze dei geni espressi con sequenze di proteine note nella banca dati non-ridondante (nr) di proteine attraverso analisi BLAST.
Si è proceduto analogamente con le sequenze ottenute dalle quattro librerie di cDNA di fiore provenienti da tre differenti cultivar di olivo Leccino, Dolce Agogia e Frantoio.

3.2. Scoperta di marcatori SNP

Al fine di identificare potenziali polimorfirsmi di lunghezza SSRs tra varietà da usare per la caratterizzazione della variabilità genetica totale entro la specie e per sviluppare nuovi marcatori da impiegare nel sistema di tracciabilità dei prodotti olivicoli, è stata messa appunto una pipeline computazionale per l'identificazione automatizzata di SSRs all'interno di sequenze generate mediante il sequenziamento 454 di campioni di cDNA.

La procedura ha previsto:

1- l'assemblaggio indipendente delle ESTs ottenute da ciascuna libreria in esame e la generazione di geni espressi unici attraverso l'ausilio del software GS-Assembler v2.5.3 (Roche);
2- l'analisi BLAST per identificare i geni unici omologhi tra le due librerie da confrontare;
3- l'identificazione parallela e automatizzata di SSRs all'interno delle librerie in esame mediante l'uso del software open-source MISA-MIcro SAtellite identification tool;
4- il "parsing" dei dati prodotti dall'analisi BLAST e dal software MISA;
5- l'identificazione degli SSRs polimorfici mediante confronto del numero di "repeats" dei singoli SSRs all'interno dei geni omologhi tra le due librerie da paragonare.

L'utilizzo della suddetta "pipeline" computazionale ha permesso di identificare:

1- 24 potenziali SSRs polimorfici tra le cultivar di olivo Tendellone e Coratina, espressi nel frutto;
2- 72 potenziali SSRs polimorfici tra le cultivar di olivo Ortice e Ruveia, espressi nel frutto;
3- 52 potenziali SSRs polimorfici tra le cultivar di olivo Leccino e Dolce Agogia, espressi nel fiore;
4- 27 potenziali SSRs polimorfici tra le cultivar di olivo Leccino e Frantoio, espressi nel fiore.

3.3 Messa a punto di una piattaforma microarray per la determinazione dei livelli di espressione genica ad alta risoluzione

Per una caratterizzazione estesa dei profili di espressione genica è stato progettato e prodotto, un microarray ad alta densità. Lo sviluppo del microarray ha previsto il disegno di sonde molecolari con opportune caratteristiche di sequenza e struttura, impiegando le informazioni genetiche custodite nel repertorio di geni espressi ottenute attraverso il sequenziamento 454 di campioni di cDNA, come descritto in precedenza.

A tal fine ci si è avvalsi della tecnologia CustomArrayTM basata sull'uso di semiconduttori modificati (CMOS) adattati per applicazioni biologiche (CustomArray Inc., http://customarrays.com). Il circuito integrato creato nei CustomArraysTM contiene microelettrodi controllati individualmente tramite un circuito logico inserito nel microarray. Ogni micro-elettrodo è controllato digitalmente per indirizzare la generazione selettiva di acidi che controllano le reazione di de-protezione del gruppo DMT-OH. Gli oligonucleotidi sono sintetizzati tramite un programma digitale ed assemblati dentro una Porous Reaction Layer (PRL) che riveste il microarray. Durante questo processo possono essere sintetizzate contemporaneamente più di 90,000 sequenze oligonucleotidiche differenti.

Nello specifico, il DNA microarray di olivo contiene circa 90,000 oligonucleotidi disegnati con caratteristiche termodinamiche simili (35 ± 5 paia di basi (bp), con il 60 ± 5 % del contenuto in basi GC ed una temperatura di Melting (Tm) di 70 ± 5 °C) mediante il software OligoArray v2.1. Successivamente gli stessi oligonucleotidi sono stati sintetizzati in situ attraverso l'uso del CustomArrayTM DNA Synthesizer (CustomArray Inc.).

3.4. Analisi dell'espressione genica su larga scala

Il sequenziamento casuale su larga scala delle quattro librerie di cDNA provenienti dai frutti delle cultivar di olivo Ortice e Ruveia campionati antecedentemente e successivamente all'attacco di mosca (Bactrocera oleae), ha prodotto molte informazioni relativamente alla variazione estesa dei

profili di espressione genica. In linea di principio, più è alto il numero di ESTs assemblate in uno specifico TC, più è alto il numero di molecole di mRNA codificanti quel particolare gene in un dato tessuto. Per tale ragione è stato valutato l'apporto delle 4 librerie a ciascun TC in termini di numero di ESTs. Per misurare l'abbondanza relativa dei trascritti genici nelle quattro librerie è stato applicato il test statistico R (Stekel et al., 2000; Alagna et al., 2009). Tutti i TCs con R>8 (true positive ratio ~99%) sono stati considerati differenzialmente espressi tra le quattro librerie. In totale sono stati identificati 425 trascritti differenziali.
Successivamente, al fine di eseguire un confronto diretto tra le cultivar Ortice e Ruveia, per ciascuna di esse è stato generato un "dataset" unico di ESTs unendo le sequenze relative alle librerie di cDNA provenienti dai frutti delle due cultivar campionati prima e dopo l'attacco di mosca. I due dataset sono stati confrontati e ad essi è stato applicato il test statistico R. In totale sono stati identificati 325 geni unici con R>8 e quindi considerati differenzialmente espressi tra le due cultivar.
Inoltre per ciascuna cultivar sono stati eseguiti due assemblaggi *de-novo* indipendenti al fine di generare un repertorio completo di geni espressi nel frutto di olivo in seguito a danneggiamento prodotto dall'attacco di mosca. I TCs ottenuti da ciascun assemblaggio sono stati annotati e classificati funzionalmente seguendo la "pipeline" illustrata precedentemente. In seguito all'assemblaggio delle ESTs ottenute tramite il sequenziamento delle due librerie di cDNA relativi alla cultivar Ortice, sono stati generati 8981 TCs. Di questi, 16 geni unici hanno mostrato un R>9 e pertanto sono stati considerati differenzialmente espressi in seguito al danneggiamento da mosca. Relativamente all'assemblaggio delle ESTs della cultivar Ruveia, degli 8050 TCs generati, ne sono stati identificati 25 differenziali in seguito al danneggiamento da mosca.

Per identificare i trascritti diffirenzialemente espressi in tessuti di fiore è stato valutato l'apporto delle 4 librerie di fiore a ciascun TC in termini di numero di ESTs. Tutti i TCs con R>8 (true positive ratio ~98%) sono stati considerati differenzialmente espressi tra le quattro librerie. In totale sono stati identificati 391 trascritti differenziali.
Inoltre, per identificare i soli geni alterati durante lo sviluppo del fiore è stato eseguito un assemblaggio *de-novo* delle sole due librerie di cDNA ottenute dai fiori della cultivar Leccino campionati in diverse fasi del loro sviluppo, pre- e post- antesi. Sono stati prodotti 5292 TCs che sono stati successivamente annotati e classificati funzionalmente. Di questi, 197 trascritti unici hanno mostrato un R>8 e pertanto sono stati considerati differenzialmente regolati durante lo sviluppo del fiore.
Per generare un database di geni coinvolti nel processo dell'aborto dell'ovario sono state confrontate due cultivar di olivo, Leccino (bassa % aborto dell'ovario) e Dolce Agogia (alta % aborto dell'ovario). E' stato eseguito l'assemblaggio delle sole due librerie di cDNA ottenute dai fiori delle due cultivar in esame campionati nelle medesime fasi dello sviluppo. Degli 8235 TCs prodotti, ne sono stati identificati 123 differenzialmente espressi tra le due cultivar.
Infine, al fine di identificare i geni coinvolti nel processo di auto-incompatibilità, sono state confrontate due cultivar di olivo, Leccino (auto-incompatibile) e Frantoio (auto-compatibile). Anche in questo caso è stato eseguito l'assemblaggio delle sequenze ottenute da campioni di fiore nelle medesime fasi dello sviluppo. Dei 6155 TCs generati, 84 trascritti unici hanno mostrato un R>8 e di conseguenza sono stati considerati differenzialmente espressi tra le due cultivar.

3.5. Analisi dell'espressione proteica su larga scala
Lo sviluppo e la maturazione della drupa in olivo rappresentano eventi chiave capaci di influenzare l'accumulo di olio nel mesocarpo e la composizione finale dell'olio. Di fatto, la composizione dell'olio extra vergine d'oliva dipende principalmente dalla composizione finale della drupa. Il crescente interesse dalla comunità scientifica verso l'uso e la composizione degli oli d'oliva è strettamente correlato agli affetti benefici sulla salute umana, in particolare alla capacità di ridurre il

colesterolo LDL. Una dieta ricca di olio d'oliva è generalmente associata ad una minore incidenza di patologie coronariche e cancerose.
Lo scopo principale dell'indagine proteomica messa in atto è monitorare le variazioni del proteoma della drupa d'olivo durante la maturazione, al fine di rivelare modulazioni nella biosintesi di composti correlabili ai tratti qualitativi delle drupe e dell'olio.

Il genotipo investigato nel corso del presente lavoro è rappresentato dalla cv Coratina.
Il contenuto proteico totale è stato estratto da drupe campionate a tre diversi stadi di maturazione: 45, 110 e 150 DAF (DAF-day after flowering). Gli estratti proteici sono stati ottenuti utilizzando un protocollo multi-step, basato sulla rimozione di contaminanti prima dell'estrazione proteica vera e propria.
Pool di almeno 4 frutti per stadi di maturazione (circa 5 g di tessuto totale) sono stati pestellati in azoto liquido. La polvere risultante è stata lavata più volte con acqua ed acetone, al fine di rimuovere i principali contaminanti, in particolare sostanze fenoliche ed olio. Successivamente, il contenuto proteico totale è stato estratto utilizzando la classica procedura d'estrazione proteica basata sull'uso di fenolo.
Gli estratti proteici così ottenuti sono stati sottoposti ad un processo di clean up, mediante l'uso di apposito kit commerciale -2D- Clean up kit (GE Healthcare) e successivamente quantificati, usando il metodo dell'amido nero (metodo Popov).
Per l'analisi 2DE, 200 µg di proteine sono state mescolate con opportune quantità di soluzione di reidratazione DeStreak Rehydration solution (GE Healthcare) e caricate mediante reidratazione passiva su strip 4-7, 18 cm (GE Healthcare). Per ciascuno stadio di maturazione, sono state condotte 4 repliche tecniche.
La focalizzazione isoelettrica è stata condotta utilizzando la seguente apparecchiatura : ETTAN IPGphor II (GE Healthcare).
Di seguito viene riportato il protocollo di corsa applicato:
1) 300V per 5 ore
2) 1000V per 7 ore (step in gradiente)
3) 8000V per 3 ore (step in gradiente)
4) 8000 V per 1 ora e 30 minuti.
20°C ; max 50 µA/strip; circa 29 KV/hr raggiunti.

Dopo la focalizzazione isoelettrica, i campioni sono stati equilibrati in due fasi:
1) Riduzione 50mM TrisHCl, pH 8.8, 6M urea, 2% SDS, 30% Glicerolo, con DTT (2%) -15 min
2) Alchilazione 50mM TrisHCl, pH 8.8, 6M urea, 2% SDS, 30% Glicerolo, con IAA (2.5%) -15 min.
L'analisi SDS PAGE è stata condotta utilizzando il sistema ETTAN DALTtwelve (GE Healthcare) e gel di poliacrilamide (conc 12%). Sono stati utilizzati i seguenti parametri di corsa:
1) 5W/gel per 45 min
2) 15W/gel per 4 ore

Gli spot proteici separati mediante 2DE sono stati visualizzati grazie all'utilizzo del reagente fluorescente SYPRO Ruby; le relative immagini sono state acquisite utilizzando opportuna apparecchiatura scanner (Typhoon- GE Healthcare). Successivamente, le mappe proteiche sono state colorate al Coomassie colloidale ed utilizzate per l'escissione manuale degli spot d'interesse.
Le immagine acquisite al Thyphoon sono state analizzate utilizzando il software SameSpots Progenesis (Nonlinear Dynamics). Questo software include la possibilità di analisi statistiche, in particolare ANOVA e possibilità di determinare FDR (false discovery rate). Tutti gli spot che mostrano p value ≤ 0.02, q value ≤ 0.01 e contemporaneamente fold change ≥ 2 sono stati indicati come spot che mostrano accumulo differenziale durante la maturazione.

Gli spot d'interesse sono stati manualmente escissi dal gel e sottoposti a reazione di riduzione con DTT ed alchilazione con IAA. Successivamente, sono stati digeriti ON (37°C) con Tripsina (Promega).
Le miscele peptidiche così ottenute sono state applicate sulla piastra Anchorchip targetTM (Bruker Daltonics) e processate attraverso spettrometria di massa MALDI-TOF. I relativi spettri MS e MS/MS sono stati acquisiti.
L'identificazione delle proteine d'interesse mediante ricerca in banca dati degli spettri collezionati è attualmente in corso.

Nel corso del presente lavoro, un protocollo d'estrazione delle proteine totali dalla drupa d'olivo è stato messo a punto ed ottimizzato. Esso è stato utilizzato per isolare le proteine presenti nel frutto a tre diversi stadi di maturazione: 45, 110 e 150 DAF
L'analisi differenziale del proteoma è stata raggiunta attraverso l'applicazione dell'elettroforesi 2D accoppiata alla spettrometria di massa MALDI TOF. Mediante elettroforesi 2D, approssimativamente 1600 spot proteici sono stati isolati e visualizzati su gel in tutti in campioni oggetto di studio.
Complessivamente, 247 spot proteici sono risultati differenzialmente accumulati nella drupa durante la maturazione. Un'immagine rappresentativa del proteoma del frutto 45 giorni dopo la fioritura è riportata nella figura 1. Questa immagine è stata utilizzata come riferimento per le procedure di gel matching durante l'analisi dell'espressione differenziale, realizzata utilizzando il software SameSpots Progenesis (Nonlinear Dynamics).
Fig.1 Immagine rappresentativa del proteoma della drupa a 45 DAF.
Gli spot proteici evidenziati sono risultati differenzialmente accumulati
nel corso della maturazione.

I dati elaborati sono stati visualizzati mediante analisi delle componenti principali (PCA). Come atteso, i risultati della PCA mostrano completa separazione degli stadi di maturazione (Fig. 2). La principale componente della PCA spiega, infatti, oltre il 72% della varianza: lo stadio di maturazione rappresenta, perciò, la più larga fonte di variazione quando si comparano stadi successivi di sviluppo.
Tra i 247 spot proteici differenzialmente accumulati durante la maturazione, circa 220 sono stati manualmente escissi dal gel e sottoposti a digestione con tripsina. Le miscele peptidiche sono state analizzate mediante spettrometria MALDI TOF ed i relativi spettri MS e MS/MS sono stati collezionati.

Risultati ottenuti

Sequnaziamento 454 del cDNA
La prima corsa, effettuata con i quattro campioni di drupa, ha prodotto 695.511 reads (251.090.724 nucleotidi) di buona qualità (Quality Average: 30,59) della lunghezza media di 361 nucleotidi.
I dati suddivisi per settore sono i seguenti:
Settore 1 - Ortice drupa (undamaged) – 173.118 reads (63.611.956 nucleotidi);
Settore 2 - Ortice drupa (damaged vari stadi in pool) – 197.782 reads (73.464.937 nucleotidi);
Settore 3 - Ruveia drupa (undamaged) – 146.765 reads (52.792.933 nucleotidi);
Settore 4 - Ruveia drupa (damaged vari stadi in pool) – 177.846 reads (61.220.898 nucleotidi).

La seconda corsa, effettuata con i quattro campioni di fiore, ha prodotto 465.331 reads (167.774.040 nucleotidi) di buona qualità (Quality Average: 32,51) della lunghezza media di 361 nucleotidi.
I dati suddivisi per settore sono i seguenti:
Settore 1 - Leccino fiore (stadi 1-6 in pool) – 113.134 reads (42.568.083 nucleotidi);

Settore 2 - Leccino fiore (stadi 7-8 in pool) - 1465.76 reads (54.807.276 nucleotidi);
Settore 3 - Dolce Agogia fiore (stadi 1-6 in pool) – 67.797 reads (21.939.320 nucleotidi);
Settore 4 - Frantoio fiore (stadi 7-8 in pool) – 137.824 reads (48.459.361 nucleotidi).

I risultati della terza corsa di sequnziamento sono ancora preliminari ed in corso di analisi.

Assemblaggio de-novo e annotazione funzionale del trascrittoma della drupa
Tale analisi ha permesso di generare un database di 87.720 geni espressi unici rappresentati da 15.058 tentantive consensus (TCs) e 72.662 singletons. Sono stati annotati 48.063 geni espressi con un hit significativo con il database nr (E-value < 1e-5), pari al 54.8 % del totale. Al fine di accelerare i tempi di esecuzione dell'analisi è stato adoperato il programma MPI-BLAST, un'implementazione gratuita e open source del BLAST, in versione parallelizzata.
I geni espressi annotati sono stati classificati funzionalmente integrando le informazioni ottenute dall'analisi BLAST con quelle dei database UniprotKB (AC), KEGG (KO) e Gene Ontology (GO), mediante l'ausilio di appositi script scritti nel linguaggio PERL. L'analisi ha evidenziato che il 38.9 % dei geni espressi, pari a 34.158 trascritti, codifica per proteine con funzione nota (database UniprotKB), di cui 21.393 annotate secondo la classificazione GENE ONTOLOGY e 21.609 secondo quella KEGG .

Assemblaggio de-novo e annotazione funzionale del trascrittoma del fiore
Tale analisi ha permesso di generare un database di 106.598 geni espressi rappresentati da 14.599 tentantive consensus (TCs) e 91.999 singletons. Sono stati annotati 56.700 geni espressi con un hit significativo con il database nr (E value < 1e-5), pari al 53.2 % del totale. Inoltre si è osservato che il 38.2 % dei geni espressi, pari a 40.770 trascritti, codifica per proteine con funzione nota (database UniprotKB), di cui 26.114 annotate secondo la classificazione GENE ONTOLOGY e 25.599 secondo quella KEGG.

Analisi differenziale dell'espressione genica
Le analisi su campioni di drupa di Ortice e Ruveia campionati antecedentemente e successivamente all'attacco di mosca (Bactrocera oleae), hanno identificato in totale 425 trascritti differenziali.
Per quanto riguarda il trascrittoma di fiore, le analisi di espressione comparativa fra campioni di leccino frantoio e dolce agogia, hanno identificato 391 trascritti alterati.

Identificazione di proteine della drupa
L'analisi del proteoma attraverso l'applicazione dell'elettroforesi 2D accoppiata alla spettrometria di massa MALDI TOF, ha permesso di identificare complessivamente, 247 spot proteici risultati differenzialmente accumulati nella drupa durante la maturazione.

Timbro Istituzione	**Firma Responsabile Scientifico U.O.**