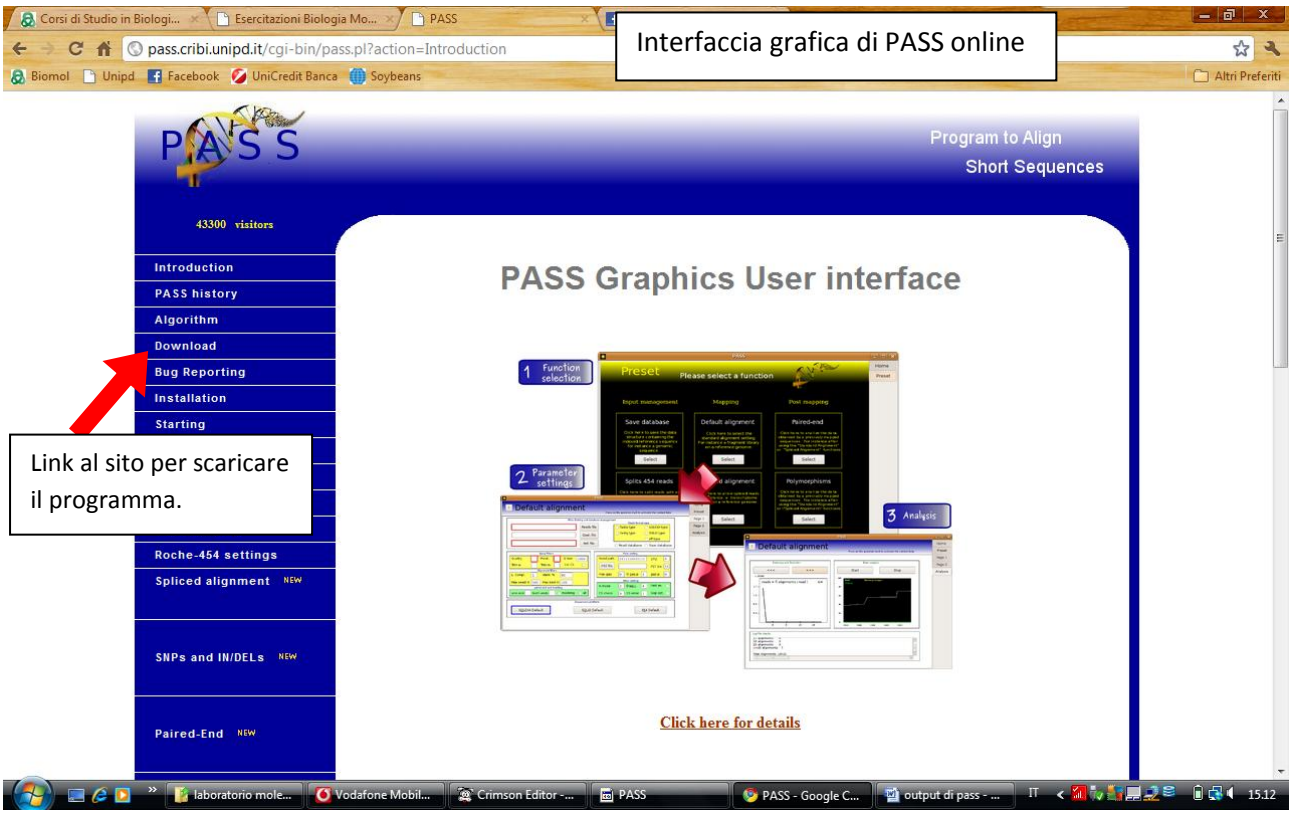
Stevanoni Martina matricola 591705

# BREVE TUTORIAL SULL'USO DI PASS PER ALLINEAMENTI

## Informazioni generali sul programma:

PASS è un programma che permette di allineare corte sequenze (reads) contro una sequenza reference iniziale in modo veloce e anche con introduzione di gap. Può essere usato con tutti i maggiori sistemi operativi e liberamente scaricato dalla rete. È in grado di maneggiare un elevatissimo numero di reads allineate poi alla reference e consente all'utente di modificare la sensibilità dell'allineamento stesso semplicemente settando i diversi parametri. Inoltre permette di utilizzare diversi formati per i dati in input e diverse modalità di visualizzazione per quelli in output.

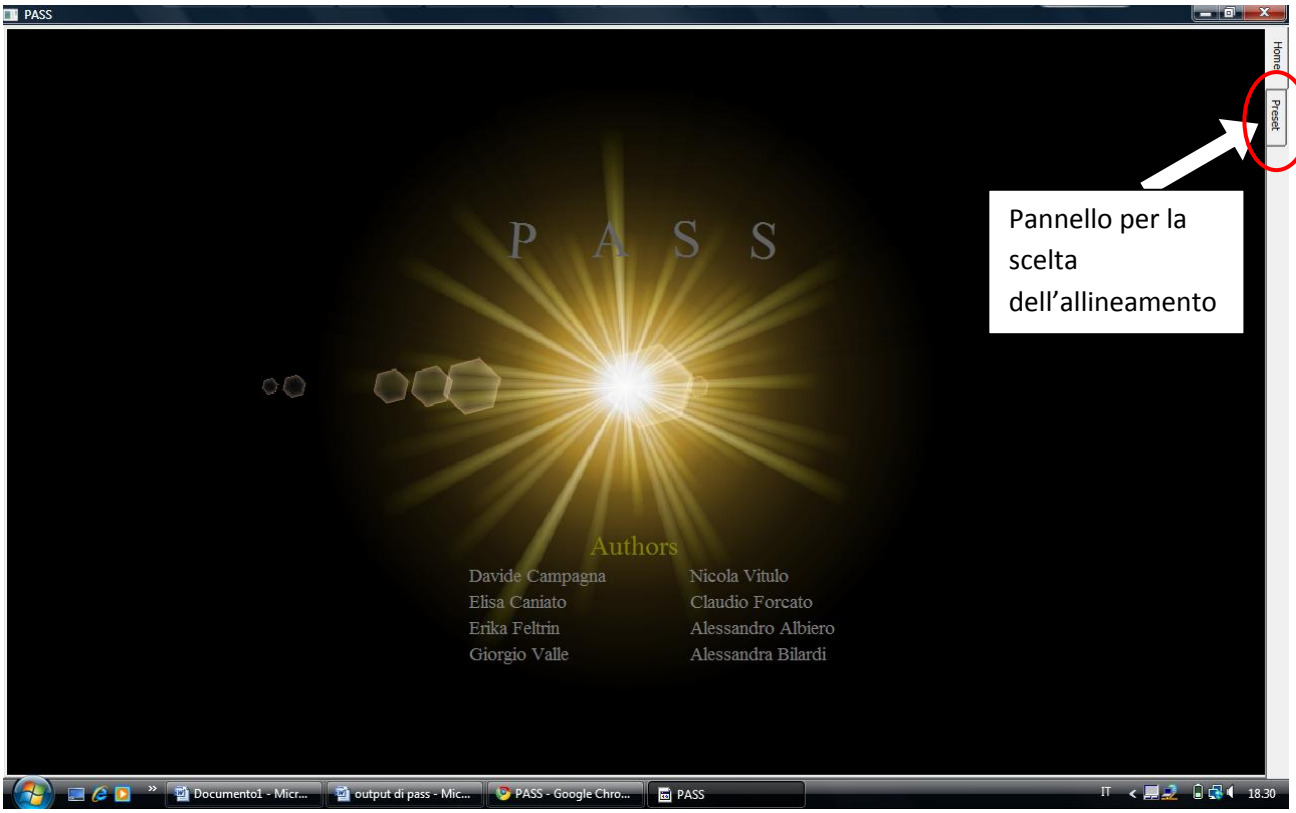



I principali vantaggi di PASS sono:

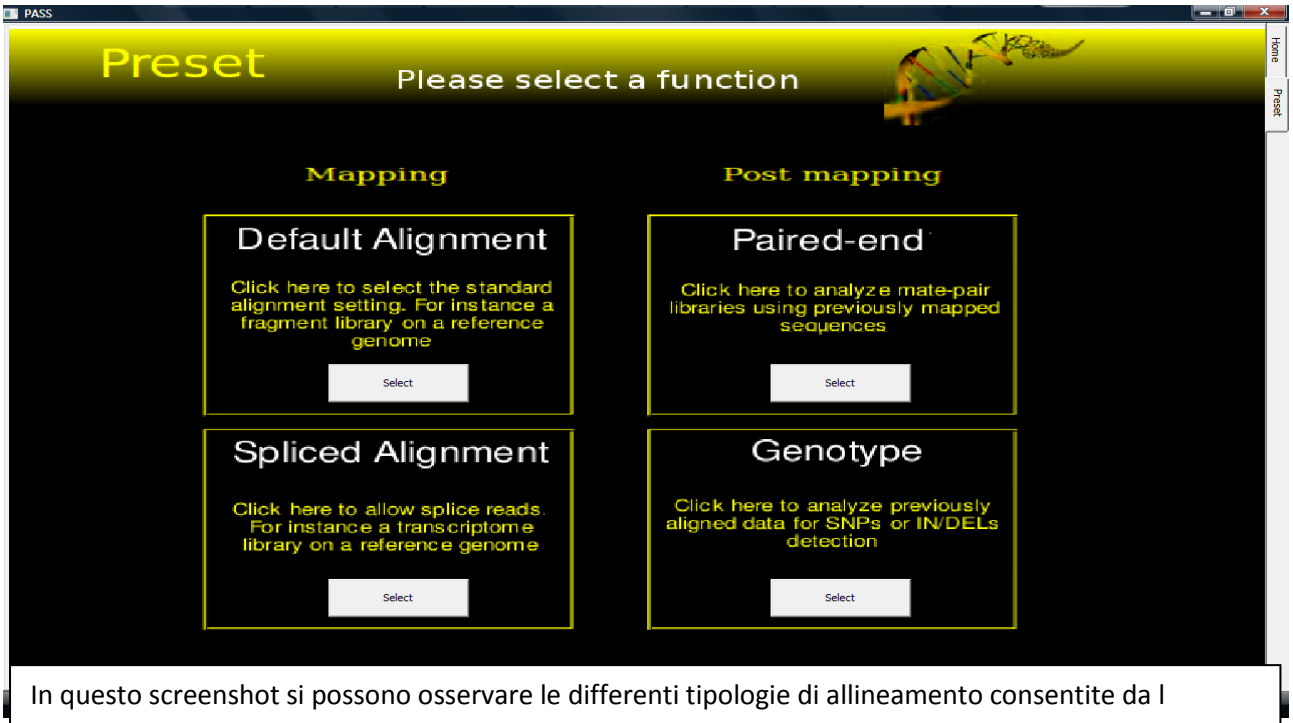
- la velocità nell'esecuzione dell'allineamento;
- i diversi parametri che permettono di modulare la sensibilità;
- la totale compatibilità con le sequenze ottenute con Solexa, Solid e 454;
- altre funzioni per allineamenti locali e supporti per allineamenti con long e spliced reads.

## Descrizione di PASS:

L'interfaccia grafica di PASS è di semplice utilizzo e consente di scegliere la tipologia di allineamento preferita con il pannello **Preset**.




Una volta stabilito quale allineamento utilizzare vengono inseriti dei parametri di default, mentre altri devono essere correttamente compilati in modo da definire i dati in input, le condizioni dell'allineamento e il formato dell'output. Questo viene fatto procedendo nei tre pannelli successivi: **Page1**, **Page2** e **Analysis**.



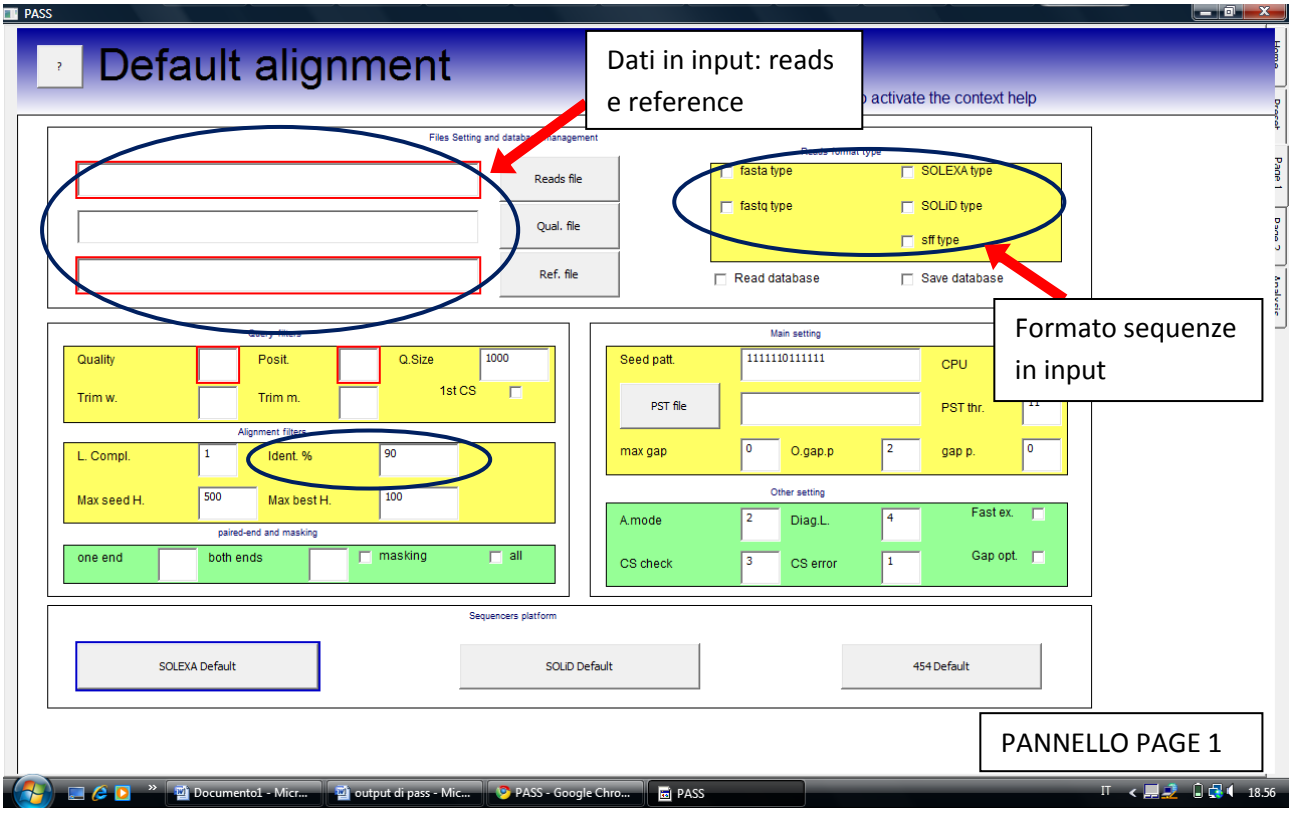

In questo screenshot si possono osservare le differenti tipologie di allineamento consentite da l programma. Per procedere con l'esecuzione poi si devono utilizzare i pannelli successivi.

Per poter dunque meglio descrivere tutti i passaggi chiave di PASS si può considerare un esempio di allineamento:

*allineamento del genoma di Drosophila contro un dataset di cromatogrammi con identità del 90%.*

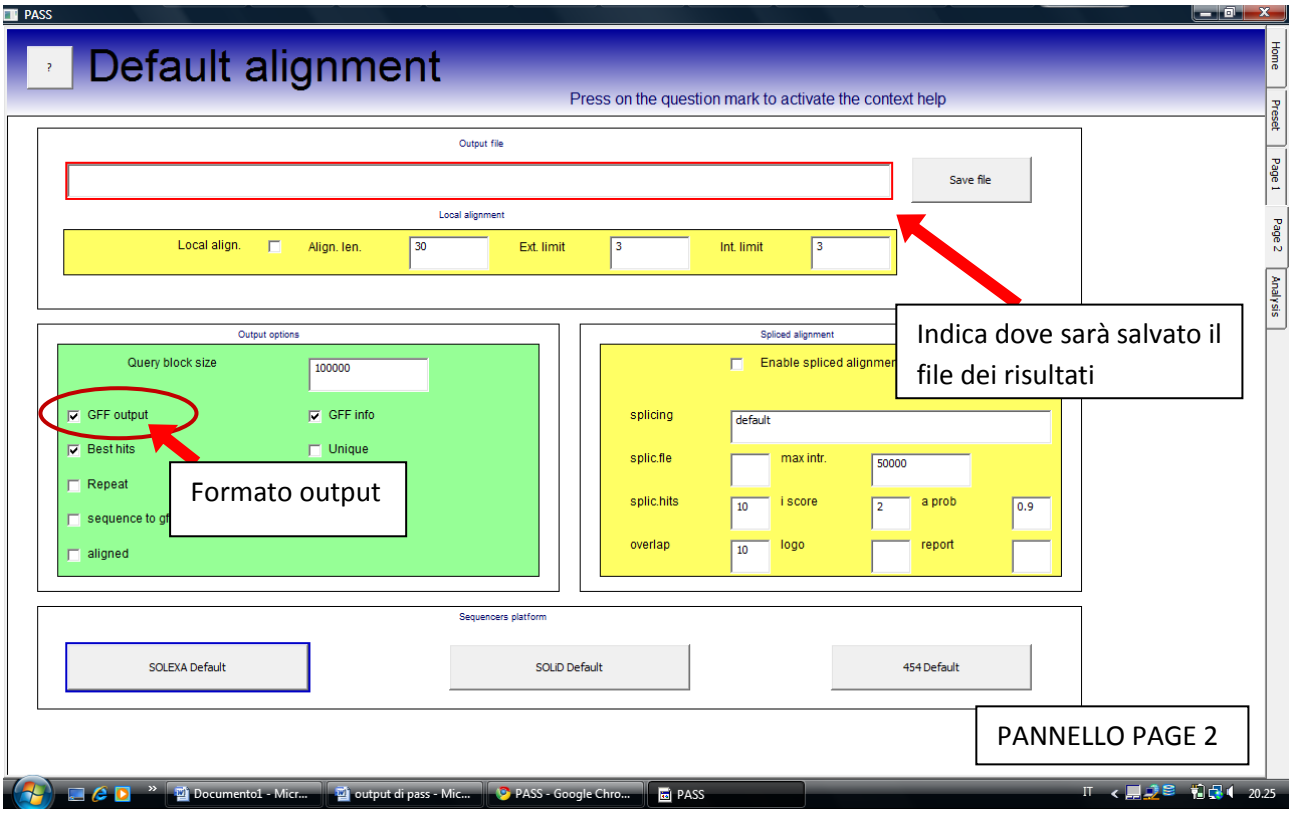
Si tratta di utilizzare un file in formato multifasta costruito con tutte le brevi sequenze dei cromatogrammi usati ad esercitazione, che costituiranno in questo caso le reads. Queste dunque dovranno essere allineate contro l'intero genoma, ma data la sua estensione nello specifico si è scelto di utilizzare la sola sequenza del cromosoma 2L, poiché era noto che le reads erano tutte state ottenute con sequenziamento del gene tim di Drosophila melanogaster situato in esso.

Perciò dal pannello **Preset** si seleziona l'allineamento "Default Alignment" e compare una finestra dove si notano dei parametri già settati di default e dei campi in rosso da compilare.



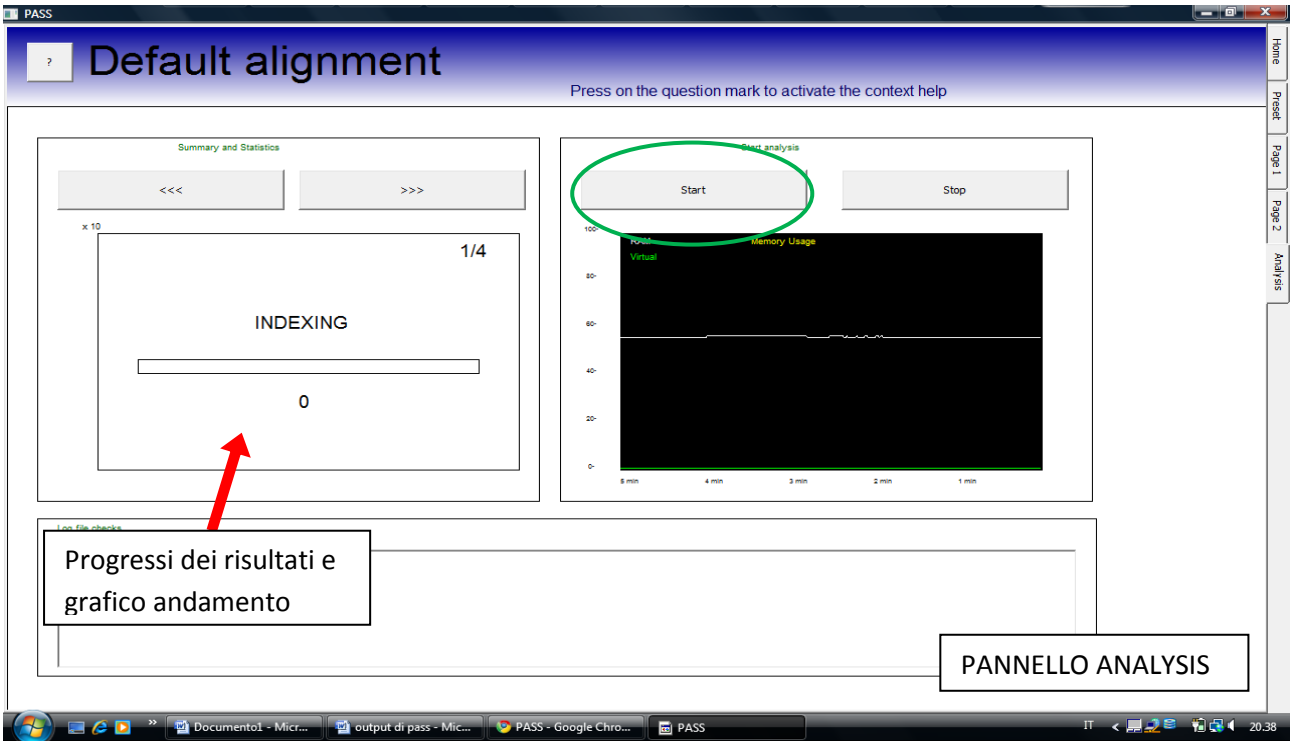

Ci si trova quindi nel pannello **Page1** dove per prima cosa si deve definire il formato dei dati in input. In questo caso si tratta del formato fasta perciò si seleziona la casella corrispondente "fasta type" sulla destra. Ora si devono compilare i campi "reads file" e "ref. file" inserendo rispettivamente il multifasta e la sequenza genomica detta reference. Sotto ad essi poi si va ad impostare la percentuale di identità voluta, in questo caso del 90%, inserendo il valore nel campo "Ident. %".

Compilati correttamente tutti i campi, si può passare al secondo pannello **Page2** dove si indica il nome e la posizione nel computer del file da salvare contenente i risultati calcolati. Questo va inserito nel campo "Save file".

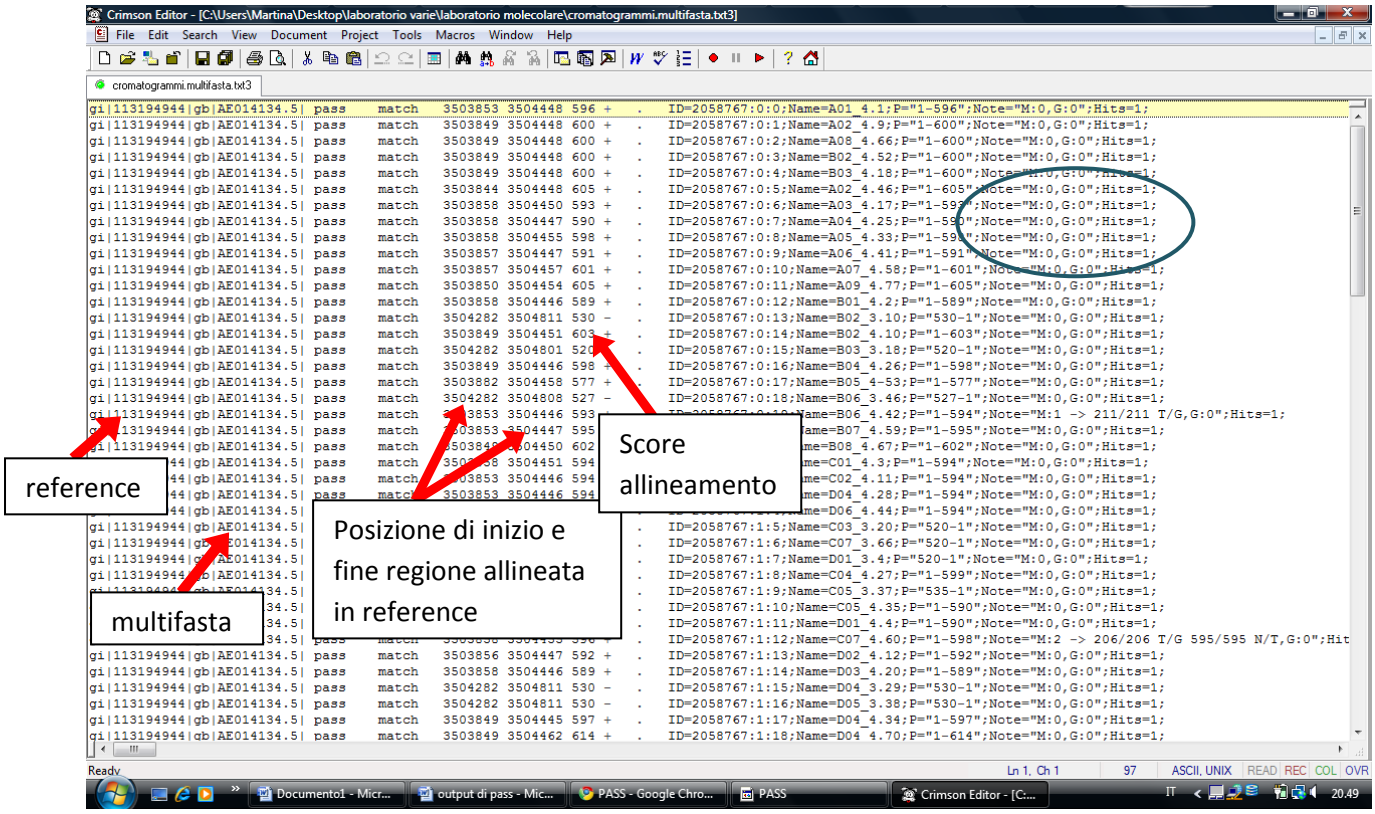



In questo pannello viene anche indicato il formato con cui si visualizzano i risultati, in questo caso si mantiene la tipologia GFF (General Feature Format); in alternativa è possibile visualizzare i risultati con formato "full alignment" come BLAST.

Terminata la compilazione si può passare all'ultimo pannello **Analysis** dove si dà l'avvio all'allineamento impostato tramite il tasto "start" per poi poter visualizzare i risultati.



Il calcolo degli allineamenti fatto da PASS prevede che vengano allineate tutte le reads, ciascuna contro la reference. In questo modo si arriva ad ottenere i risultati in un documento leggibile tramite il programma Crimson Editor SVN286, dove compare la schermata contenente circa 15 colonne di numeri coma previsto dal formato GFF.




I risultati visualizzati devono però essere interpretati. Dall'allineamento tra reads e genoma (cromosoma 2L) di Drosophila si osserva che: la prima colonna contiene il nome con cui è indicata la reference (sequenza genomica) e la seconda il nome con cui appaiono le sequenze brevi o reads (multifasta); la quinta e la sesta colonna indicano le posizioni di inizio e di fine dell'allineamento sulla sequenza reference perciò la regione sul cromosoma dove si trova la read corrispondente; la settima colonna riporta il punteggio dell'allineamento (abbastanza elevato per tutti i casi); la colonna otto definisce su quale filamento della reference si sono allineate le singole reads (5'→3' (+) o 3'→5'(-)); la colonna 11 contenente il campo "Name" indica il nome di ciascuna reads a cui di fatto vanno a corrispondere tutte le voci lungo la riga; la colonna successiva invece riporta le coordinate dei nucleotidi di inizio e fine dell'allineamento delle reads e si vede che tutte sono totalmente allineate; i due campi successivi indicano rispettivamente se sono presenti mismatch "M" specificando poi le posizioni dove si trovano e i gap "G" non trovati con queste reads; l'ultimo campo infine definisce gli "Hits" stabilendo dunque quante volte la reads considerata è presente e nel caso specifico si nota che tutti i valori danno 1 cioè tutte le sequenze non vengono ripetute.

Concludendo dunque si può affermare che tramite PASS è possibile fare degli allineamenti con brevi sequenze contro ad esempio interi genomi ottenendo risultati in tempi ristretti e allo stesso tempo accurati. Ciò significa che si è in grado di stabilire quanti e quali mismatch sono presenti, se sono stati inseriti gap e gli hits che si sono verificati.