**How to Deal With 3 Partial Genome/EST Assemblies**

By: William Green

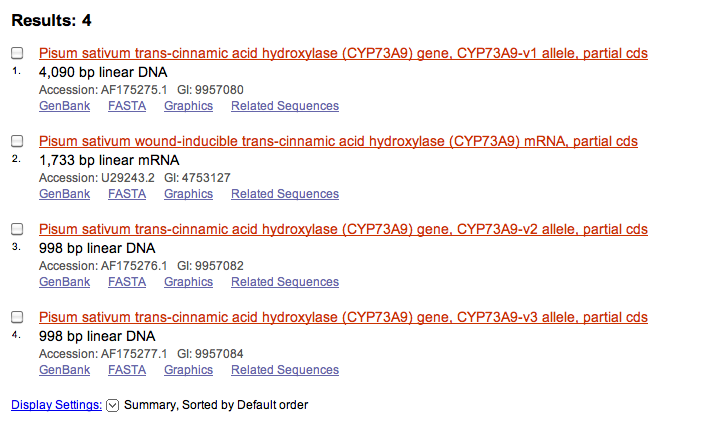
This tutorial attempts to uncover troubleshooting problems when working with multiple genome databases. While the incorporation of multiple genome annotation databases has been shown to improve the overall interpretation of microarray gene expression data (Yin 2010), this tutorial will examine deficiencies in BLASTn results between three partial assemblies: 454, Illumina and EST.

In order to examine key differences, this tutorial will specifically address steps involved in the annotation of the Cytochrome P450 gene (CYP73A9) from *Pisum Sativum* (pea) against the blueberry genome. It will address problems encountered when blasting between the different databases for: CYP73A9-v1 allele cds vs. CYP73A9 mRNA cds.

Note: For background information on the different types of sequencing methods please refer [here](http://gcat.davidson.edu/mediawiki-1.15.0/index.php/Jared). This tutorial assumes the user is familiar with BLASTn and has BLAST version 2.2.24+ installed locally. Procedure for creation of databases and using BLAST locally was followed as described in class. For more information on P450s and specifically CYP73A9 please refer to the online wiki ([link](http://gcat.davidson.edu/mediawiki-1.15.0/index.php/Blueberry_Genome_Project_for_Bio343)).

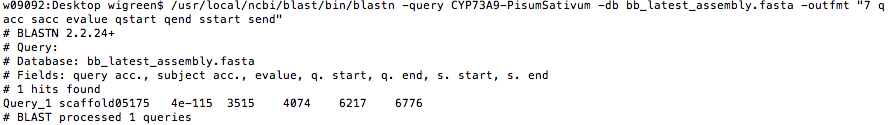
**Procedure:**

**Step 1**: To begin, a nucleotide search was preformed in NCBI for the sequence encoding CYP73A9 (\*the first identified P450 gene in plants). Four sequences were identified in pea (*Pisum Sativum*), two of which encoded for variants of the same exon (CYP73A9-v2 & CYP73A9-v3).

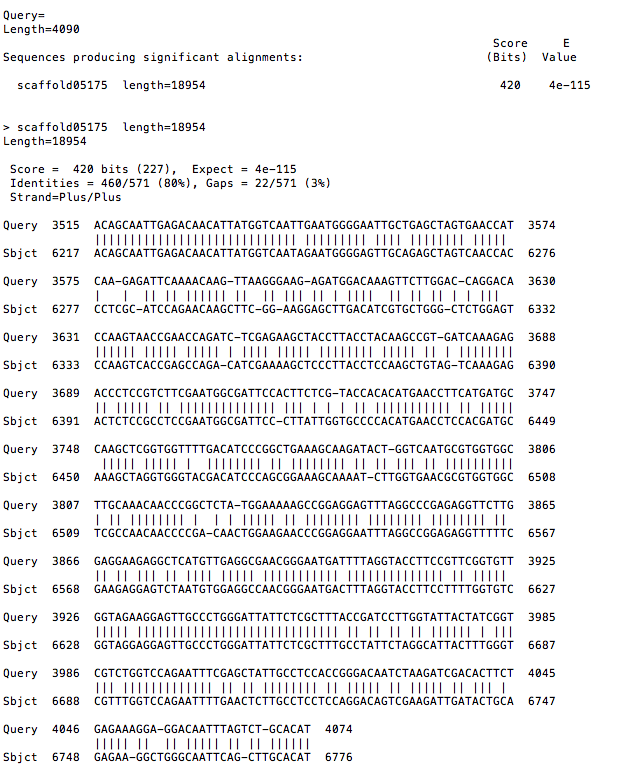


CYP73A9-v1 allele & CYP73A9 mRNA were converted to FASTA files in text edit and were used to BLAST against the three blueberry genome databases.

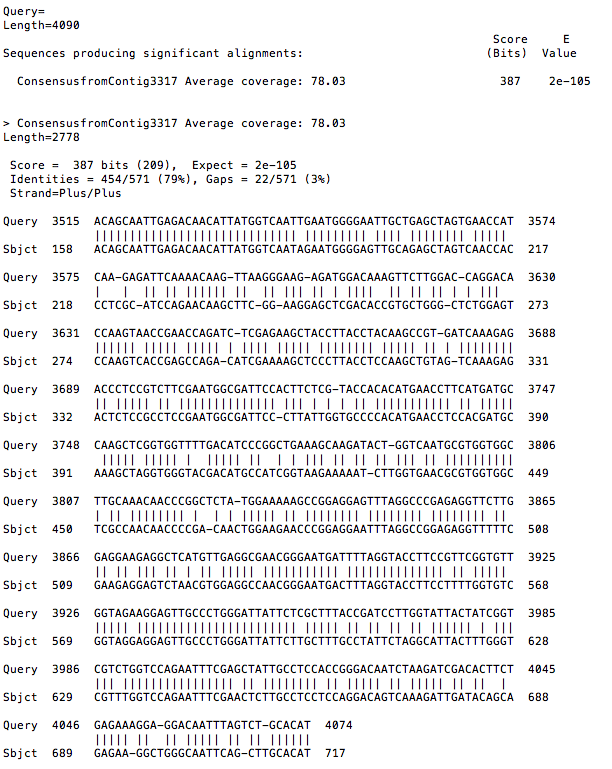
**Step 2**: CYP73A9-v1 sequence (or v1 as shorthand) was blasted locally against the 454 database. One hit was found.



The blueberry compliment was then determined by re-submitting the BLAST, but this time without the output format (/usr/local/ncbi/blast/bin/blastn -query CYP73A9-PisumSativum -db bb\_latest\_assembly.fasta).

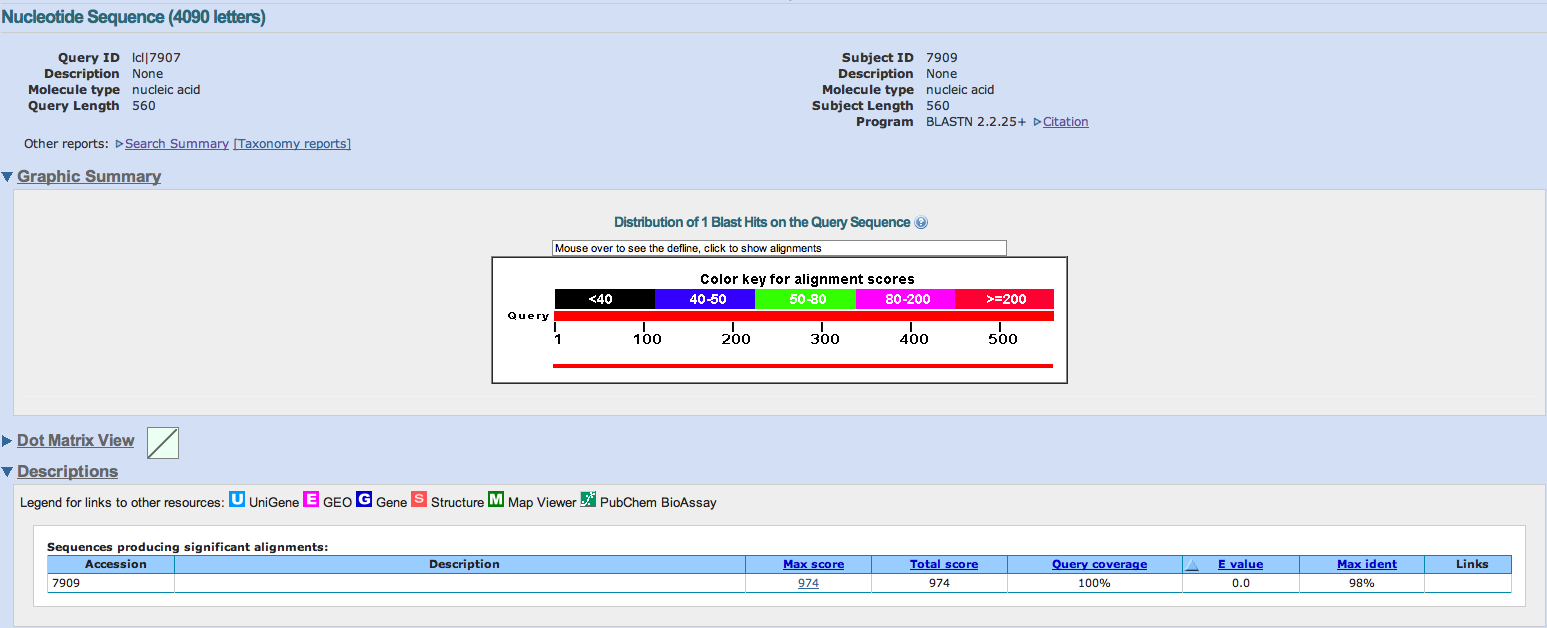


This returned an 80% identity match with blueberry 454 scaffold05175, yielding an e value of 4e-115. The same sequence (v1) was then blasted against the Illumina database. This too found 1 hit that matched 79% with blueberry Contig3317.



**Step 3:** These two hits were then blasted against each other (via NCBI online) as a comparison to see if the different databases returned similar results. The blast returned an e value of 0.0 indicating a perfect match between the two sequences.

Note: this comparison could also be achieved by blasting the determined blueberry FASTA sequence against the other database (ex. BB sequence from 454 against Illumina database). This also returned an e value of 0.0, again indicating a perfect match.

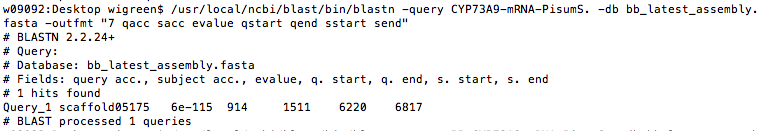


**Step 4:** The v1 sequence was then blasted against the EST database. No hits were found. Since the v1 sequence encodes the entire gene, it contains both introns and exons. Whereas the EST database was created solely from isolated mRNA therefore it contains only exons.

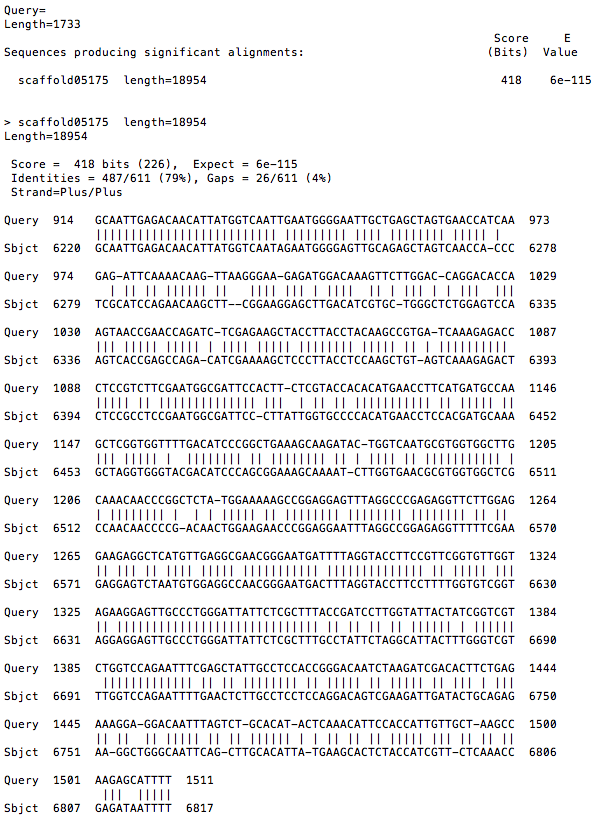
The two identified blueberry consensus hits (step 2) were then blasted against the EST database, again resulting in no hits found. One explanation for the lack of hits could be that the mRNA encoded by the CYP73A9 gene was not expressed at the time of collection for the EST database.

However, since our original file contained the whole gene and not just exons, some undetermined error, whether it was from the assembly or blast, could account for the lack of hits seen with the EST database. Therefore, the CYP73A9 mRNA (or mRNA for shorthand) sequence was used to both verify the aforementioned results and to examine the lack of return seen in the EST database.

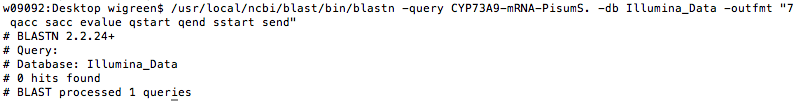
**Step 5:** To verify, the mRNA sequence was blasted against the 454 database. Again one hit was found at relatively the same position (3bp shift) on 454 scaffold05175 with an e value of 6e-115.



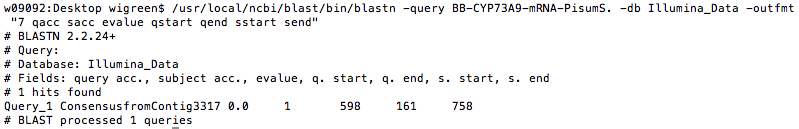
Like before the blueberry compliment was determined, yielding a 79% identity match.



The mRNA sequence was then blasted against the Illumina database, again for verification of the v1 results. Shockingly no hits were found!!



This was not expected. How could one assembly return a hit but not the other? Even more shocking was when the blueberry compliment of mRNA was blasted against the Illumina database, a 100% match was found against the Illumina Contig 3317--as indicated the e value (0.0).

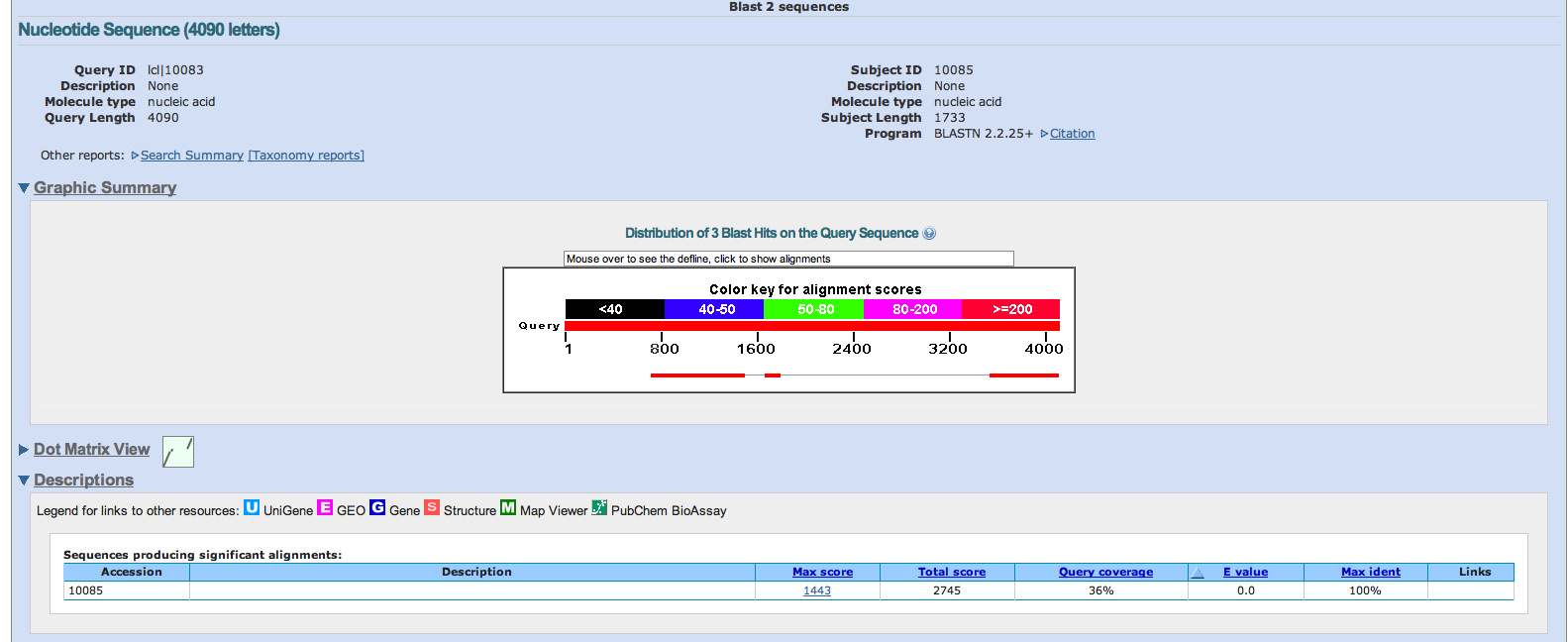


How could this happen? How could one database return a highly conserved hit while the other found nothing at all? Given that both database sequences are of the same species, how could an orthologous gene sequence (v1) differ from the mRNA? Is assembly method at fault?

To examine this discrepancy, the following steps were preformed.

(Note: no hits were determined from the EST database by the mRNA sequence—suggesting that no mRNA from CYP73A9 was produced during sample collection).

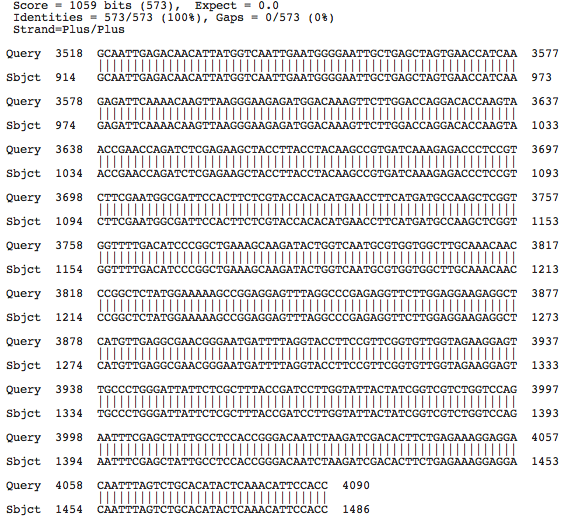
**Step 6:** The v1 sequence was blasted against the mRNA in order to determine the position of exons and introns. This was done online via BLASTn comparing two sequences. ([BLAST link](http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=MegaBlast&PROGRAM=blastn&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&BLAST_SPEC=blast2seq))



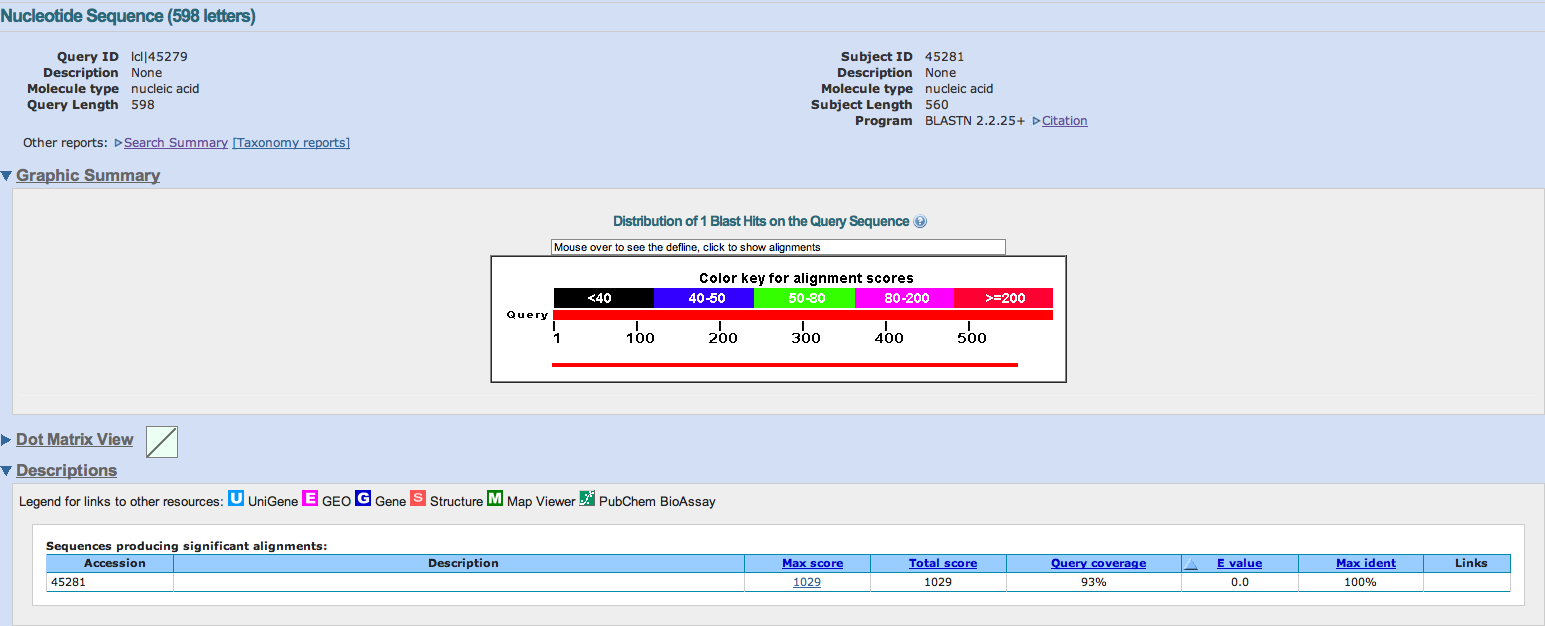
The v1 gene sequence was used as the query and the mRNA for the subject. The above image illustrates this comparison: the 3-exon regions encoded in the mRNA (red rectangles indicated by arrows) with the gaps between them representing the introns.

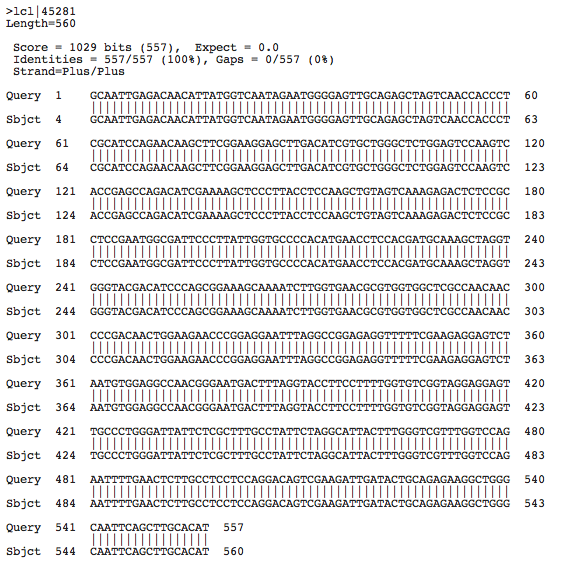
Interestingly, exon 3 appears to match base pair positions for all three called blueberry consensus sequences: v1-454-scaffold05175, v1-Illumina-Contig3317 & mRNA-454-scaffold05175

Shown below is Exon 3 from v1 (query) vs. mRNA (subject) BLAST. (\*refer to BLAST results above for nucleotide positions of identified scaffolds & contigs)



**Step 7:** Next, the consensus blueberry sequences from the two CYP73A9 (v1 vs. mRNA) sources were compared to see if these were identifying the same region &/or base pairs. Of the 598bp identified with 454-mRNA blasts, 557bp matched 100% with the 454 & Illumina blueberry consensus sequences.

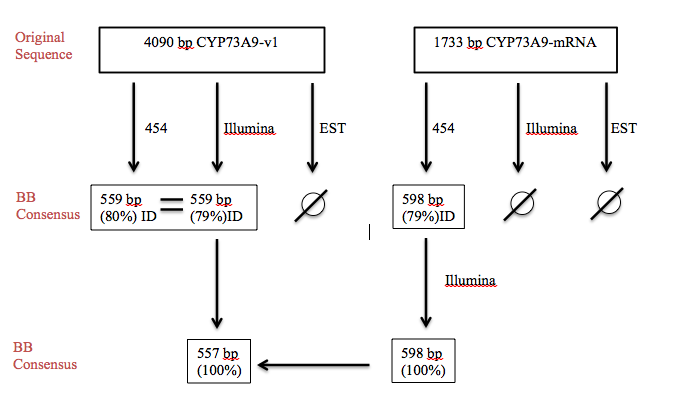




Again, the consensus sequences match perfectly. This indicates that each hit was in fact the same region of DNA located on either the 454-scaffold05175 or Illumina-Contig3317, specifically in exon 3 of the gene.

For simplification purposes let us consider the diagram presented below which illustrates the steps preformed up until this point.

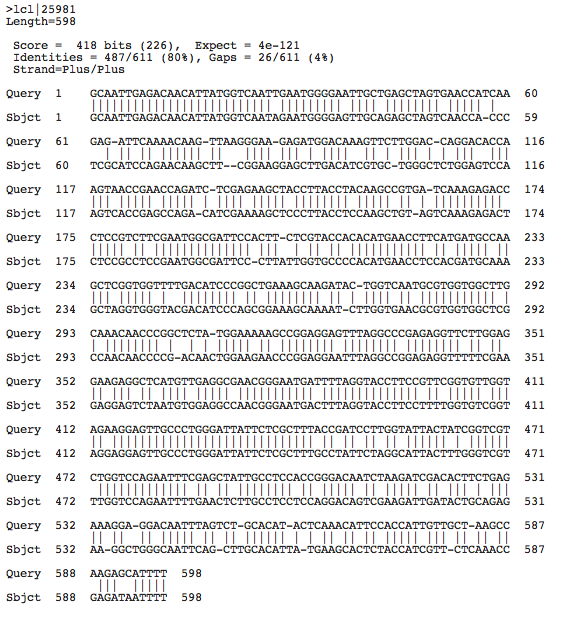
(% ID indicates amount conserved nucleotides between query sequence and blueberry sequence; Key feature: no hit was found with initial BLAST of mRNA against Illumina database but 100% match with either BB consensus sequence)



Each time either the original or BB consensus sequences were BLAST against the databases the same regions were identified in both 454 and Illumina. How then could no hits be found for the mRNA sequence against the Illumina database if both location and % ID were identical in all other BLAST results? (Original & BB consensus)

**Step 8:** To examine the character of the mRNA sequence, the corresponding 598bp from CYP73A9-mRNA that complimented with blueberry (for 454) was isolated from the original 1733bp mRNA sequence. This was then blasted against the Illumina database.

Again, there were no hits found. However when comparing the 598bp sequence from the original 1733bp CYP73A9-mRNA with the 598bp blueberry consensus sequence, we find there is 79% convergence with a very good e value of 4e-121.



**Step 9:**  Finally, in a last ditch effort to discover why this BLAST discrepancy occurred; the 454-scaffold05175 was compared against the Illumina-Contig3317 in an effort to discover differences between the assembly sequences. To preform this comparison exon 3 was located and extracted from both assembly segments separately. They were then blasted against each other using a nucleotide BLAST for 2 or more sequences.

For results of BLAST click [here](454%20vs.%20Illumina%20(all).docx). (file is too large to display accurately in this document, 2778bp).

**Conclusion:** While no conclusive explanation can be made at this point as to why the BLAST result for CYP73A9-mRNA against the Illumina database did not recognize the identified orthologous sequence in blueberry, several possible explanations can be hypothesized.

1) The difference in sequencing methods for 454 vs. Illumina (as described [here](http://gcat.davidson.edu/mediawiki-1.15.0/index.php/Jared)) could have resulted in higher reads for the Illumina database. This higher coverage could raise the threshold value for error rate when using BLAST, resulting in lower identification rates for the Illumina genome database. Additionally, since the Illumina is composed of significantly shorter reads than the 454, less overlap could occur in the CYP73A9 gene region of the unfinished blueberry annotation.

2) Referring to the BLAST results from the last step described in the procedure, there are 4 gaps in the 454 sequence that are not present in the Illumina. While these four nucleotide differences, at positions -790, -2727, -2738 & -2739 on the leading strand, causing the mRNA sequence to be undetectable is unlikely, the nucleotide difference at position 790 falls closely after the position identified to be exon 3 in the blueberry genome (~160-750bp on Illumina-Contig3317). This close proximity to exon 3 could lead to an offset shift in the nucleotide sequence alignment for the Illumina file that could account for the variance seen in the BLAST with CYP73A9-mRNA results. This shift would not occur when using BLAST with the CYP73A9-v1 allele sequence because the intron region would fall at this nucleotide position.

3) Also since both sequences of pea (*Pisum Sativum*) are only partial coding sequences, some un-identified error could have caused the discrepancy in annotations described above. A more finished genome sequence could also increase the accuracy and validity of the blueberry genome.

Even though some problems can/may arise when comparing and using multiple genome assemblies, this integration allows for increased coverage of the genome. Additionally, by using similar methods described in this tutorial, researchers can parse shared characteristics between the assemblies and by doing so can in a sense troubleshoot complications as they arise.

**Reference:**

Yin, Jun, Sarah Mcloughlin, Ian B. Jeffery, Antonino Glaviano, Breandan Kennedy, and Desmond G. Higgins. "Integrating Multiple Genome Annotation Databases Improves the Interpretation of Microarray Gene Expresssion Data." *BMC Genomics* 11.50 (2010). *BioMed Central | The Open Access Publisher*. BioMed Central Ltd. Web. 21 Feb. 2011. <http://www.biomedcentral.com/1471 2164/11/50/abstract/>.