Biomedical Data Mining

Extended Abstract

# De novo RNA seq assembly and annotation of important legume-Vicia sativa L. (SRR403901)

**Hetalkumar J Panchal**

## Abstract

*Vicia sativa L.* which is otherwise called normal vetch; is a nitrogen fixing leguminous plant in the family Fabaceae. As of late, cutting edge sequencing innovation, named RNA-seq, has given an incredible way to deal with breaking down the Transcriptome. This investigation is center around RNA-seq of *Vicia sativa L.* of SRR403901 from NCBI database for all over again Transcriptome examination. An aggregate of 12.4 million single peruses were created with N50 of 588 bp. Succession get together contained absolute 22748 contigs which is further inquiry with known proteins, an aggregate of 7652 qualities were distinguished. Among these, solitary 500 unigenes were explained with 18761 quality cosmology (GO) practical classes and groupings planned to 122 pathways via looking against the Kyoto Encyclopedia of Genes and Genomes pathway database (KEGG). These information will be helpful for quality disclosure and useful investigations and the huge number of records revealed in the current examination will fill in as an important hereditary asset of the *Vicia sativa L.*

## Introduction

Cutting edge sequencing techniques for high throughput RNA sequencing (transcriptome) is getting progressively used as the innovation of decision to identify and evaluate known and novel records in plants. This Transcriptome investigation technique is quick and straightforward on the grounds that it doesn't require cloning of the cDNAs. Direct sequencing of these cDNAs can create short peruses at a remarkable profundity. In the wake of sequencing, the subsequent peruses can be gathered into a genome-scale record profile. It is an increasingly far reaching and effective approach to gauge Transcriptome sythesis, get RNA articulation examples, and finds new exons and qualities; sequencing information of Transcriptome was gathered utilizing different get together apparatuses, utilitarian explanation of qualities and pathway investigation conveyed with different Bioinformatics devices. The huge number of records announced in the current examination will fill in as an important hereditary asset for *Vicia sativa L.*

High-throughput short-read sequencing is one of the most recent sequencing advances to be discharged to the genomics network. For instance, on normal a solitary sudden spike in demand for the Illumina Genome Analyser can result in more than 30 to 40 million single-end (~35 nt) groupings. Be that as it may, the subsequent yield can without much of a stretch overpower genomic investigation frameworks intended for the length of conventional Sanger sequencing, or even the littler volumes of information coming about because of 454 (Roche) sequencing innovation. Regularly, the underlying utilization of short-read sequencing was kept to coordinating information from genomes that were almost indistinguishable from the reference genome. Transcriptome examination on a worldwide quality articulation level is a perfect utilization of short-read sequencing. Customarily such examination included integral DNA (cDNA) library development, Sanger sequencing of ESTs, and microarray investigation. Cutting edge sequencing has become a possible technique for expanding sequencing profundity and inclusion while diminishing time and cost contrasted with the customary Sanger strategy.

## Methods

### 1. Sequence Retrieval:

This investigation is center around the again get together and succession explanation of Vicia sativa L.

Hetalkumar J Panchal
Navsari Agricultural University, India, E-mail: swamihetal@gmail.com

of SRR403901 from NCBI database. Crude information downloaded from NCBI SRA (which is from Illumina HiSeq 2000 stage and the example is single finished with 12.4 M spots and 42.4% GC content. Crude grouping was changed over in to fastq record design for additional a documentation with the utilization of SRA TOOL KIT from NCBI.

**2. NGS QC Toolkit**

NGS QC Toolkit, it is an application for quality check and sifting of top notch information. The toolbox is included easy to understand apparatuses for QC of sequencing information produced utilizing Roche 454 and Illumina stages, and extra instruments to help QC (succession group converter and cutting devices) and examination (measurements devices). An assortment of choices have been given to encourage the QC at client characterized boundaries. The toolbox is required to be valuable for the QC of NGS information to encourage better downstream analysis.

**3. De novo sequence assembly by CLC GENOMICS WORKBENCH 7**

A far reaching and easy to use investigation bundle for examining, contrasting, and picturing cutting edge sequencing information. This bundle was utilized for all over again succession get together of grouping with naturally boundaries of all over again get together device.

**4. BLASTX**

The gathered record was additionally considered for explanation in which initial step was to recognize deciphered protein groupings from contigs. BLASTX at NCBI performed with changing barely any boundaries like non excess protein database (nr) chose as Database; Eudicots chose in creature alternative and in Algorithm boundaries Max target Sequences set to 10 and Expect limit set to 6.

**5. Blast2GO**

Blast2GO is an ALL in ONE device for useful explanation of (novel) groupings and the investigation of comment information. In light of the consequences of the protein database comment, Blast2GO was utilized to acquire the practical characterization of the unigenes dependent on GO terms. The record contigs were arranged under three GO terms, for example, atomic capacity, cell process and natural procedure WEGO instrument was utilized to play out the GO utilitarian characterization for the entirety of the unigenes and to comprehend the dispersion of the quality elements of this species at the large scale level. The KEGG database was utilized to explain the pathway of these unigenes.

**6. SSR mining**

We utilized MIcroSAtellite (MISA) for microsatellite mining which gives different factual yields of records with helpful data.

**7. Plant transcription factor**

PlantTFcat: An Online Plant Transcription Factor and Transcriptional Regulator Categorization and Analysis Tool utilized for distinguishing plant record factor in groupings.

**Results and Discussions**

**1. NGS QC Toolkit**

Arrangement was separated with this apparatus by evacuating connectors and other defiled materials then nature of grouping additionally checked with this device lastly great channel succession document considered for anew grouping get together.

**2. De novo Sequence Assembly**

CLC GENOMICS WORKBENCH 7 considered for all over again grouping get together with naturally boundaries like Mismatch Cost = 2, Insertion Cost = 3, Deletion Cost = 3, Length Fraction = 0.5, Similarity Fraction = 0.8, Word size = 21 lastly 22748 contigs produced with normal estimation of 503 by this product and different subtleties are appeared.

**3. Functional annotation with BLASTX and blast2GO**

3.1 BLASTX

BLASTX was performed to adjust the contigs against non-excess successions database utilizing an E esteem edge of 10-6. Out of 22748 record contigs, 13482 were having BLAST hits to known proteins with high huge closeness and 1114 had no BLAST hits. Out of all out records contigs, and shows that species conveyance where 9819 groupings demonstrated noteworthy likeness with *Medicago truncatula* and least closeness was found *with Prunus mume*.

3.2 Enzyme Code (EC) Classification

Chemical arranged with complete of 2336 groupings which is additionally characterized into six classes which are of Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases and Ligases.

3.3 Gene Ontology (GO) Classification

To practically classify Vicia sativa L. record contigs, Gene Ontology (GO) terms were alloted to each amassed record contigs. Out of 22748 record contigs, 18761 unigenes were assembled into GO utilitarian classifications, which are conveyed under the three primary classifications of Molecular Function, Biological Process and Cellular Components which is yield of WEGO apparatus; It shows that, inside the Molecular Function classification, qualities encoding restricting proteins and proteins identified with reactant action were the most advanced. Proteins identified with metabolic procedures and cell forms were improved in the Biological Process class. With respect to the Cellular Components classification, the cell and cell part were the most exceptionally spoken to classifications.

A sum of 500 unigenes were explained with 122 pathways in the KEGG database. Numerous records incorporate different pathways like metabolic pathways, plant-microbe collaboration pathways, unsaturated fat digestion pathway and unsaturated fat biosynthesis.

**4. SSR mining**

Microsatellite markers (SSR markers) are the absolute best sub-atomic markers in the development of a *Vicia sativa L.* hereditary guide and in assorted variety investigation. For recognizable proof of SSRs, all records were looked with perl content MISA. We distinguished an aggregate of 1150 SSRs in 1055 records. The mono-nucleotide SSRs spoke to the biggest division of SSRs distinguished followed by tri-nucleotide and di-nucleotide SSRs. Albeit just a little division of tetra-, penta-and hexa-nucleotide SSRs were distinguished in records, the number is very noteworthy.

**5. Plant Transcription Factor**

Further, record factor encoding records were recognized by succession correlation with realized record factor quality families. Result shows that record factor qualities circulated with in any event 82 families were distinguished. The general circulation of record factor encoding records among the different known protein families is fundamentally the same as with that of different vegetables as anticipated before.

**Conclusions**

This investigation is center around Vicia sativa L. species (SRR403901) from NCBI database for once more Transcriptome examination by RNA-seq utilizing cutting edge Illumina sequencing. The transcriptome sequencing empowers different practical genomics reads for a living being. Albeit a few high throughput advances have been produced for quick sequencing and portrayal of transcriptomes, communicated grouping information are as yet not accessible for some, life forms, including many yield plants. In this investigation, we performed all over again practical comment of the *Vicia sativa L.* transcriptome without considering any reference species with huge non-excess

Hetalkumar J Panchal
Navsari Agricultural University, India, E-mail: swamihetal@gmail.com

arrangement of 34678 records. The nitty gritty examinations of the informational index has given a few significant highlights of *Vicia sativa L.* transcriptome, for example, GC content, conserved genes across legumes and other plant species, assignment of functional categories by GO terms and identification of SSRs by MISA tool. It is noted that this study of *Vicia sativa* L. will be useful for further functional genomics studies as it includes useful information of each transcript.

**Acknowledgment**

We are heartily thankful to Prof. (Dr.) P.V. Virparia, Director, GDCST, Sardar Patel University, Vallabh Vidyanagar, for providing us facilities for the research work.

This work is partly presented at International Conference on Transcriptomics on July 27-29, 2015 Orlando, USA