International Journal of Pharmacy and Pharmaceutical Sciences

ISSN- 0975-1491 Vol 8, Issue 1, 2016

Original Article

# MINING SINGLE NUCLEOTIDE POLYMORPHISM FROM PUBLICLY AVAILABLE ESTS OF BREAD WHEAT (*TRITICUM AESTIVUM* L.)

SAKET CHANDRA, KUNAL MUKHOPADHYAY, MANISH KUMAR*

Department of Bio-Engineering, Birla Institute of Technology, Mesra, Ranchi 835215 Jharkhand, India
Email: manish@bitmesra.ac.in

*Received: 09 Oct 2015 Revised and Accepted: 18 Nov 2015*

## ABSTRACT

**Objective:** The present study was undertaken to discover Single Nucleotide Polymorphisms (SNPs) in bread wheat with reference to leaf rust disease.

**Methods:** Next Generation Sequencing platform sequencing by Oligonucleotide Ligation and Detection (SOLiD) was performed on four Serial Analysis of Gene Expression (SAGE) libraries of mock and leaf rust pathogen infected near-isogenic lines HD2329±Lr28. CLC Genomics Workbench was used for computational prediction of the SNPs. The predicted SNPs were filtered by Blast using wheat Expressed Sequence Tags (ESTs). The SNP-containing ESTs were annotated, and their expression was checked in response to inoculation of *Puccinia triticina.*

**Results:** We have identified 191 SNPs from data obtained through the These EST-SNPs participated in various physiological and biochemical processes that influence important traits, such as cell rescue, defense and disease resistance.

**Conclusion:** Very little knowledge exists on SNPs in hexaploid bread wheat (*Triticum aestivum* L.) because of the difficulty to discern the true polymorphic loci. This study has revealed fast and costs effective approach for SNP discovery which will be helpful in molecular breeding with important agronomic traits.

**Keywords:** Wheat (*Triticum aestivum*), SNPs, SOLiD-SAGE, ESTs, Leaf rust.

© 2016 The Authors. Published by Innovare Academic Sciences Pvt Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

## INTRODUCTION

There has been a recent inclination for single nucleotide polymorphism (SNP)-based markers to substitute other marker types in many crop species, because, in general, SNPs are widespread in the genome, both within and between genes. Major resources have been devoted for the development of SNPs as high-throughput markers and also to SNP discovery. SNP discovery projects have been undertaken in many plant species, such as *Arabidopsis thaliana*, barley, maize, rice, soybean and wheat [1-8]. In species for which no reference sequence is available, large-scale SNP discovery in genes is generally dependent on sequence information in libraries of expressed sequence tags (ESTs) for either direct discovery or as the source for primer design for re-sequencing [9-12]. ESTs have been mined as a source of SNPs in sugarcane [13-15]. The cost of cloning and conventional sequencing of more than a modest number of products is excessive for most budgets. In addition, haplotype assignment can be confusing with this system as a result of bacterial host mismatch repair of cloned PCR heteroduplexes–which can produce apparent 'recombinant' haplotypes [16]. Although SNPs can be typed rapidly when identified, the process of genome-wide SNP discovery has been performed for several crop species.

Bread wheat (*Triticum aestivum* L.) is a key cereal crop in both human and animal nutrition. Its huge genome consists of three highly related sub-genomes (homoeologous A, B and D genomes), originated from two independent polyploidization events [17] (Dubcovsky and Dvorak, 2007). The first event involved the hybridization of two diploid progenitors, an ancestor of *Triticum urartu* (AA genome) and a species related to *Aegilops speltoides* (BB genome), which resulted in wild and cultivated allotetraploid wheats (*T. turgidum* ssp.). The second hybridization occurred between ancestors of the diploid *Aegilops tauschii* var. *strangulate* (DD genome) and an allotetraploid wheat resulting in allohexaploid. Some studies have been carried out on nucleotide diversity in wheat because of the presence of two or three homoeologous genome copies. Cultivated wheat species are reported to have a low level of nucleotide diversity due to their evolutionary history and several demographic bottlenecks and selective events [8, 18]. Therefore, to date, SNP discovery in these species has been a tough task.

Fungal pathogens are a major cause of yield losses in wheat and resistance to fungal pathogens is fundamental to global food security. To reduce crop losses, wheat production is dependent on new and improved cultivars with resistance to the rapidly evolving biotrophic wheat rust diseases, such as leaf rust (*Puccinia triticina*), stripe rust (*Puccinia striiformis*) and stem rust (*Puccinia graminis*). Introducing genes from related species could enhance resistance to pests and diseases, and increase crop yields. Development in next-generation sequencing (NGS) and the unraveling of wheat's complex genome will help the process to identify molecular markers for useful wheat characteristics, to improve this development of novel wheat cultivars.

To identify new gene-associated SNPs, we have taken advantage of the rapidly developing databases of partial cDNA sequences, ESTs that have been generated from many different tissues of the wheat plant. Because the majority of these libraries have been obtained from different individuals, assembly of overlapping sequences for the same region can lead to the identification of new SNPs. In this report, we describe a strategy for rapidly identifying candidate SNPs within ESTs. We attempted to utilize SNPs discovered from ESTs in the public domain for the development of markers.

## MATERIALS AND METHODS

### Plant materials, sequencing and library construction

Near-isogenic lines (NILs) of *Triticum aestivum* cultivar, HD2329 was used in this study. One of the NILs has *Lr28* gene and absent in the other which makes it resistant and susceptible respectively. The seeds were grown to single leaf stage in the growth chamber available at National Phytotron Facility, IARI, New Delhi. Leaf rust pathogen *Puccinia triticina* pathotype 77-5 was used in the study. The pathogen inoculum was prepared by addition urediospores of *P. triticina* pathotype 77-5 and talcum powder (ratio 1:1) and applied smoothly on leaves of HD2329+*Lr28* and HD2329 with the help of a paint brush. Another set of plants belonging to HD2329+*Lr28* and HD2329 were inoculated with only talcum powder and used as a control. After inoculation, misting of the growth chamber was performed and plants were placed under a high humidity of>90% for 24 h post inoculation (hpi) in the dark to facilitate infection [19].

SAGE libraries were prepared for four selected wheat lines; susceptible HD2329 mock, susceptible HD2329 infected, resistant HD2329+*Lr28* mock and resistant HD2329+*Lr28* infected using SOLiD SAGE kit (Applied Biosystems, CA, USA) following recommended protocol. S-M library corresponds to reads generated from susceptible wheat variety HD2329 after mock inoculation. S-PI library stands for reads generated after challenging HD2329 with leaf rust pathogen*.* R-M library is formed after the resistant variety of wheat HD2329+*Lr28* is mock inoculated. R-PI is created after the resistant variety HD2329+Lr28 is challenged with *Puccinia triticina.*

### *In silico* discovery of SNPs

Computational methods nowadays dominate in SNP discovery. We used CLC Genomics Workbench 6.5.1 (CLC bio, Aarhus, Denmark; http://www. clcbio. com) for predicting the SNPs. The SOLiD SAGE reads were first trimmed for quality and adapter. The reads of each library were first screened for a minimum length of 20 bases and a minimum Phred quality score of 20. The sequences were then trimmed of poly A/Ts. This step eliminates low-quality portions of reads, thereafter, *Puccinia* sequences were discarded by allowing the reads to map against the transcripts of *Puccinia* available at The Broad Institute (www.broadinstitute.org/annotation/genome/ puccinia_group/Multi Home.html). The reads that did not match to *Puccinia* transcripts were considered for the discovery of SNPs. The trimmed and *Puccinia* removed reads from each of the four libraries were mapped separately to the reference available at Gene Indices (ftp://occams.dfci.harvard.edu/pub/bio/tgi/data/Triticum_aestivu m/). Two mapping strategies were used to generate and compare the output of SNP numbers and frequencies (table 1). The first strategy involved mapping the sequence reads to the reference at default parameter *i.e.* using a length fraction of 0.5, the similarity of 0.8 and random handling of non-specific reads. In other words, 50% of the reads must have 80% identity to the reference. Insertion, deletion and mismatch costs were 3, 3, and 2 respectively. The second strategy involved mapping the sequence reads to the reference at stringent parameters with length fraction 0.9, the similarity of 0.9 and non-specific mapping of reads were ignored i.e. gene paralogues were minimized by setting the match mode to 'ignore' which meant that those reads aligning to more than one position would be ignored or discarded. Putative SNPs from both the relaxed and stringent mapping were called using the Quality based variant detection tool in CLC Genomics Workbench, which is based on the Neighbourhood Quality Standard (NQS) algorithm [20]. This algorithm uses a combination of quality filters and user-specified thresholds for coverage and frequency to find SNPs. We required an 11-base NQS 20/20, i.e. Phred score of 20 or higher at the central base, and a window of five bases on each side, with a quality score of 20. The minimum variant frequency was set to 0.5% in order to capture a large dataset including rare alleles while the minimum coverage was set to 20x for sensitivity [21]. The resulting SNPs and allele frequencies were tabulated automatically and exported to Excel.

**Table 1: Parameters used for default and stringent mapping**

| Parameter | Default | Stringent |
|---|---|---|
| Masking | No | No |
| Mismatch cost | 2 | 2 |
| Deletion cost | 3 | 3 |
| Length fraction | 0.5 | 0.9 |
| Similarity | 0.8 | 0.9 |
| Colorspace alignment | Yes | Yes |
| Non-specific matches | Random | Ignored |

To reduce the rates for misidentification of SNPs or removal of uninformative SNPs, post-processing of the predicted SNP data was done. BLAST was performed considering the consensus sequence which comprises of predicted SNPs as a query against NCBI wheat EST database. The best hit which do not contain any gaps or Ns and has mismatch only at the position of predicted SNPs was selected for further processing (fig. 1).

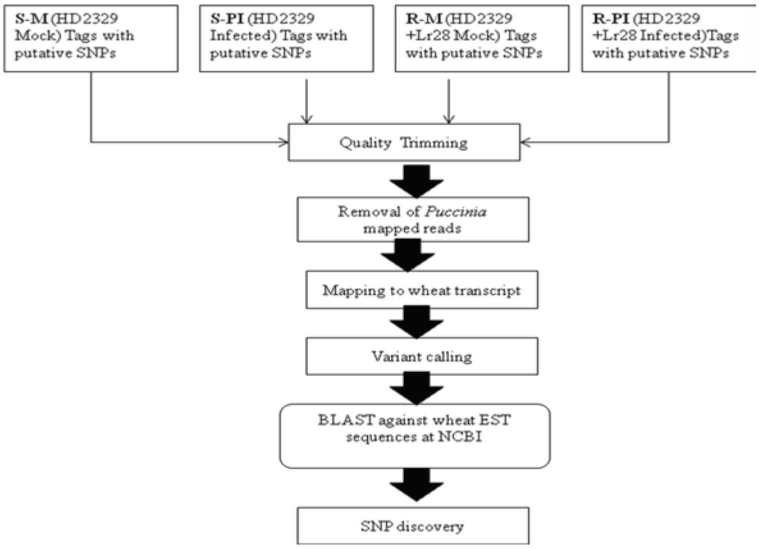

S-M (HD2329 Mock) Tags with putative SNPs
S-PI (HD2329 Infected) Tags with putative SNPs
R-M (HD2329 +Lr28 Mock) Tags with putative SNPs
R-PI (HD2329 +Lr28 Infected)Tags with putative SNPs
Quality Trimming
Removal of *Puccinia* mapped reads
Mapping to wheat transcript
Variant calling
BLAST against wheat EST sequences at NCBI
SNP discovery

**Fig. 1: Pipeline for SNP discovery**

### Annotation of the sequence showing predicted polymorphism

To know the function and pathway in which the particular EST is involved, the sequences were annotated using the software Blast2GO [22]. Functional annotations of polymorphic SNPs containing sequences were analysed for gene content by blastx to non-redundant protein database at NCBI using an e-value cutoff of 1e-5. The blastx search results were filtered to remove non-specific homologies using the following filtration: (1) for each EST sequence with a gene hit results were filtered to keep only the hits with the minimal e-value score; and (2) EST sequence with several hits having the same minimal e-value were further filtered to keep the hits with the highest HSP (high-scoring segment pairs; calculated as the product of % identity multiplied by alignment length). Only SNP-containing EST sequences having a gene hit were used for further analysis.

### Localization of SNP-containing ESTs in *T. aestivum* genome

While no complete physical map has yet been developed for *T. aestivum*, chromosome and chromosome arm-specific scaffolds are available at the IWGSC Survey Sequence repository (http://wheat-urgi. versailles. inra. fr/Seq-Repository/) with access to blasting and download. Thus, it was interesting to determine the genomic distribution of *T. aestivum,* SNP-containing sequences, at chromosomes, sub-genomes and arms levels. To this aim, DNA sequences of gene models identified in this study were individually used to perform Blastn against the full set of scaffolds from the IWGSC's wheat chromosome survey sequence (CSS), including repeats.

### Expression studies of the SNP predicted sequences

In order to identify the role of SNP-containing sequences, expression analysis was performed based on the abundance of reads within a particular library. For expression study, reads from four SOLiD SAGE libraries, as mentioned earlier was used to decipher the expression profiles of predicted SNP-containing sequences when challenged with the leaf rust pathogen *P. triticina* to the mock-inoculated controls. Comparison of S-M vs. S-PI, R-M vs. R-PI and S-PI vs. R-PI, were performed by taking sequences containing SNPs as a reference. High-quality reads from the individual library were mapped to the reference to obtain total mapped reads. Analysis of gene expression between the above-mentioned pair of libraries were assessed using Reads Per Kilobase of transcript per Million mapped reads (RPKM) were read counts of a particular contig explain its expression. RPKM, allows measuring even sparsely expressed transcripts considering read count as fundamental. The contigs were considered to be differentially expressed when the average fold change was abs≤ 2; the other criteria was false discovery rate (FDR) p-value correction<0.05 and the difference in absolute value>10 [23]. All post-trimmed reads were mapped to *de novo* assembled contigs using the minimum read length fraction set at 0.9, minimum similarity set at 0.95, and up to 10 non-specific matches were allowed. RPKM was selected as expression value. Uniquely mapped reads were assigned to each contig, allowing a maximum of two mismatches. Statistical difference in expression level was calculated using Kal's test at CLC Genomics Workbench 6.5.1 [24].

## RESULTS

### Creation of sequencing libraries and mapping of sequencing reads

Using SOLiD sequencing, we generated four high-quality libraries of SOLiD-SAGE reads namely, S-M, S-PI, R-M and R-PI. In total, 1, 65, 767, 777 reads with an average length of 34.85 bases were generated (table 2). After trimming low-quality reads, poly A/T tails, adaptor sequences, about 38, 180, 500 reads with an average length of 28.9 were retained (table 2). The library S-PI and R-PI contain *Puccinia* reads so, it was necessary to remove these reads. The *Puccinia* reads were removed by mapping it to the reference available at The Broad Institute. The libraries of S-PI and R-PI, as expected, mapped more i.e. 20.9% and 19.4% to *Puccinia* specific reads (table 3). After removing *Puccinia* specific reads about 30,894,161 reads were retained for subsequent analysis and SNP discovery.

**Table 2: Summary of trimming report of SOLiD SAGE libraries**

| Library name | No. of reads | Average length (nucleotide) | No. of reads after trim | Percentage trimmed (%) | Average length of read after trim |
|---|---|---|---|---|---|
| S-M | 48,782,889 | 34.9 | 12,247,862 | 25.11 | 29.5 |
| S-PI | 37,756,220 | 34.9 | 12,924,486 | 34.23 | 29.2 |
| R-M | 40,118,870 | 34.8 | 6,780,611 | 16.90 | 28.6 |
| R-PI | 39,109,798 | 34.8 | 6,227,541 | 15.92 | 28.3 |

**Table 3: Summary of mapping with *Puccinia* transcripts**

| Library name | Total no. of reads after trim | No. of reads mapped *Puccinia* transcripts | Percentage of reads mapped to *Puccinia* transcripts |
|---|---|---|---|
| S-M | 12,247,862 | 2,154,694 | 17.6 |
| S-PI | 12,924,486 | 2,70,1621 | 20.9 |
| R-M | 6,780,611 | 1,232,943 | 18.2 |
| R-PI | 6,227,541 | 1,208,473 | 19.4 |

The main aim of this study was to discover SNPs in a large number of wheat genes. For this purpose two mapping strategies were employed. The first mapping was performed at relaxed parameters and the second at stringent parameters (table 4 and 5). About 23,981,205 reads are mapped with the reference. The S-PI library has the maximum percentage of mapped reads (table 4). In stringent parameters as expected only 7,124,560 numbers of reads mapped to the reference (table 5). As the majority of the reads were based on the expressed part of the genome, the Transcript Assembly available at Gene Indices was selected as the main reference for aligning the SOLiD SAGE reads from the four libraries for SNP detection.

**Table 4: Mapping report using default parameter**

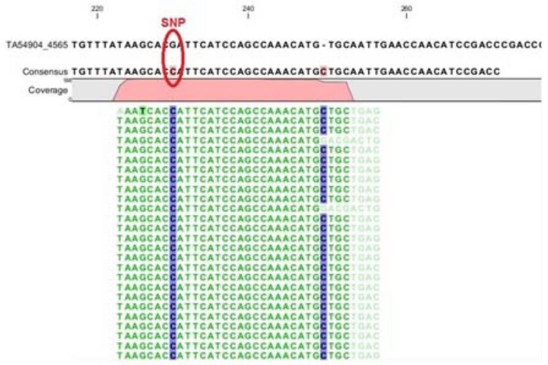
| Library name | No. of reads after removing *Puccinia* matched reads | No. of reads mapped to wheat transcript assembly | Percentage of reads mapped to wheat transcript assembly |
|---|---|---|---|
| S-M | 10,093,168 | 7,781,231 | 77.1 |
| S-PI | 10,225,473 | 8,014,500 | 78.4 |
| R-M | 5,551,126 | 4,092,737 | 73.7 |
| R-PI | 5,024,394 | 4,092,737 | 74.15 |

**Table 5: Mapping report using stringent parameter**

| Library name | No. of reads after removing *Puccinia* matched reads | No. of reads mapped to wheat transcript assembly | Percentage of reads mapped to wheat transcript assembly |
|---|---|---|---|
| S-M | 10,093,168 | 2,159,634 | 21.40 |
| S-PI | 10,225,473 | 2,302,277 | 22.52 |
| R-M | 5,551,126 | 1,376,298 | 24.79 |
| R-PI | 5,024,394 | 1,286,351 | 25.60 |

### Discovery of single nucleotide polymorphisms

SNP discovery was carried out on the reads mapped to the transcript assembly of wheat sequence after depleting reads that matched to the chloroplast, mitochondrial or known repeat sequences. A pipeline developed is mentioned in fig. 1. The main focus was to find SNPs between the homologous loci (fig. 2). About 10,012 numbers of candidate SNPs were initially identified from the sequence alignments.



```
TA54904_4565 TGTTTATAAGCACGATTCATCCAGCCAAACATG-TGCAATTGAACCAACATCCGACCCGACC
Consensus    TGTTTATAAGCACCATTCATCCAGCCAAACATGCTGCAATTGAACCAACATCCGACC
                    AAATCACCATTCATCCAGCCAAACATGCTGCTGAG
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAC
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAC
                    TAAGCACCATTCATCCAGCCAAACATGGACGACTG
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAC
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAG
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAG
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAG
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAC
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAG
                    TAAGCACCATTCATCCAGCCAAACATGGACGACTG
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAC
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAG
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAG
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAG
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAG
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAC
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAG
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAC
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAC
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAC
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAC
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAC
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAC
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAC
                    TAAGCACCATTCATCCAGCCAAACATGCTGCTGAG
```

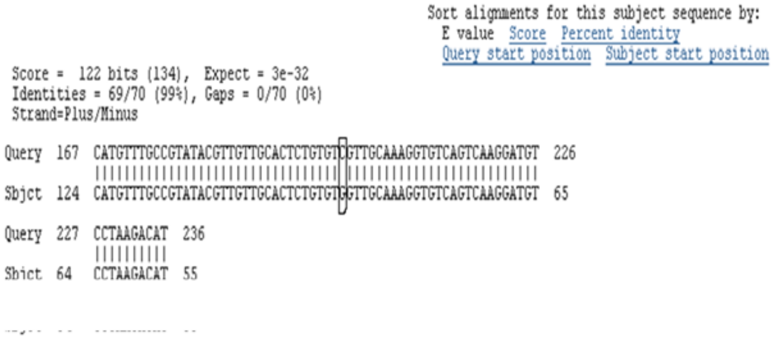
**Fig. 2: CLC Genomics workbench snapshot showing putative SNPs**

The default parameter predicted about 9428 SNPs and even with stringent parameters for SNP detection, 584 putative SNPs were detected (table 6). In S-PI library a maximum number of putative SNPs (3348) were identified.

The sequences containing the putative SNPs were extracted. Uninformative SNPs or false SNPs were removed by BLAST filtering performed against wheat ESTs at NCBI. Each SNP-containing sequence was checked for no gaps, mismatch or N's at either side of the SNPs and only those SNPs fulfilling these criteria were selected (fig. 3). After BLAST filtering 191 EST containing SNPs were selected. The number of SNPs remained in each library after blast filtering is shown in table 7.

**Table 6: Summary of SNPs detected in respective library**

| Library name | Default parameter | Stringent parameter | Total |
|---|---|---|---|
| S-M | 3065 | 180 | 3245 |
| S-PI | 3162 | 186 | 3348 |
| R-M | 1684 | 112 | 1796 |
| R-PI | 1517 | 106 | 1623 |
| Total | 9428 | 584 | 10,012 |



```
                                              Sort alignments for this subject sequence by:
                                                E value  Score  Percent identity
                                                Query start position  Subject start position

 Score =  122 bits (134),  Expect = 3e-32
 Identities = 69/70 (99%), Gaps = 0/70 (0%)
 Strand=Plus/Minus

Query  167  CATGTTTGCCGTATACGTTGTTGCACTCTGTGTCGTTGCAAAGGTGTCAGTCAAGGATGT  226
            |||||||||||||||||||||||||||||||| |||||||||||||||||||||||||||
Sbjct  124  CATGTTTGCCGTATACGTTGTTGCACTCTGTGTGGTTGCAAAGGTGTCAGTCAAGGATGT  65

Query  227  CCTAAGACAT  236
            ||||||||||
Sbjct  64   CCTAAGACAT  55
```

Result of BLAST :

consensus sequence as query<br>
EST sequence as subject,<br>
position of putative SNP encircled

**Fig. 3: Blastn filtering for selecting putative SNPs**

**Table 7: Summary of SNPs detected in respective library after BLAST filtering**

| Library name | Default parameter | Stringent parameter | Total |
|---|---|---|---|
| S-M | 29 | 26 | 55 |
| S-PI | 38 | 19 | 57 |
| R-M | 25 | 19 | 44 |
| R-PI | 19 | 16 | 35 |
| Total | 111 | 80 | 191 |

### Functional annotation of SNPs containing sequences

The SNP-containing genes were identified by blastx search against the non-redundant protein database at NCBI and putative functional annotation were assigned based on homology. In total, 136 SNP-containing sequences were putatively annotated. These genes encode proteins mainly participating in the biological processes of the biosynthetic process, response to stress and DNA metabolic process (fig. 4). In molecular function the most represented process was nucleotide binding, DNA binding and catalytic activity (fig. 5). In cellular component most of the sequences are localized on cytoplasm, plastid and mitochondrion (fig. 6).

Homology distribution showed a maximum hit to *Aegilops tauschii* followed by *Triticum urartu* and *Hordeum vulgare* (fig. 7).

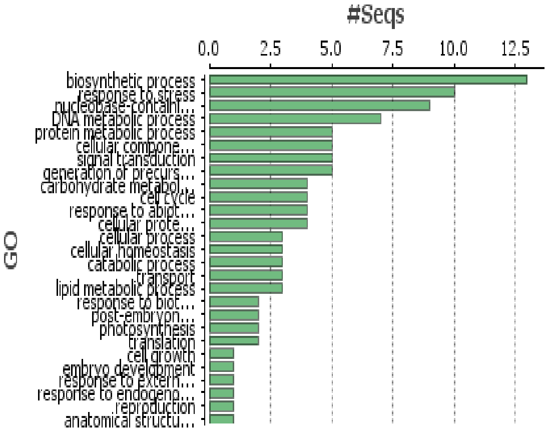


**Fig. 4: Distribution of GO terms in the biological process category**

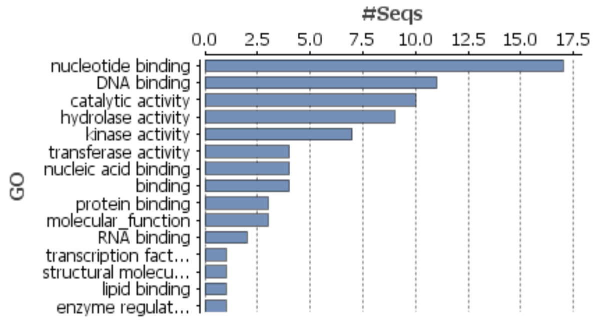


**Fig. 5: Distribution of GO terms in the Molecular function category**

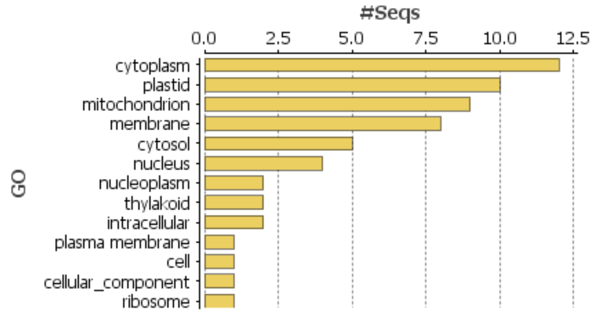


**Fig. 6: Distribution of GO terms in the Cellular Component category**

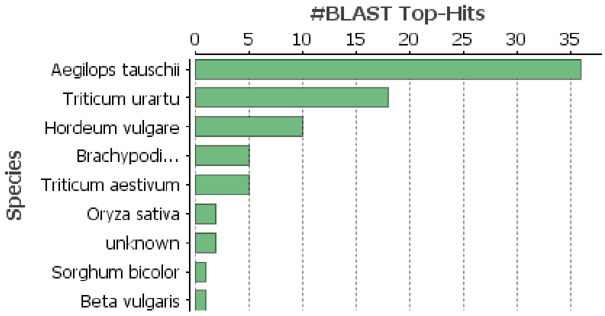


**Fig. 7: Ranking based on the number of hits matching SNPs containing sequences using Non-redundant protein database**

### EST distribution in sub-chromosome arms of *Triticum aestivum*

The chromosome arm specific distribution (fig. 8) showed chromosome 3B has the highest number of SNP-containing EST (15) followed by chromosome arm 4AL (14). On comparing the homologous groups of wheat chromosomes, group 7 had the greatest number of SNP-containing sequences (30). At the sub-genome level, the distribution of SNP-containing genes was almost balanced, with sub-genomes A, B, and D containing 45, 45 and 36 SNP-containing sequences respectively.

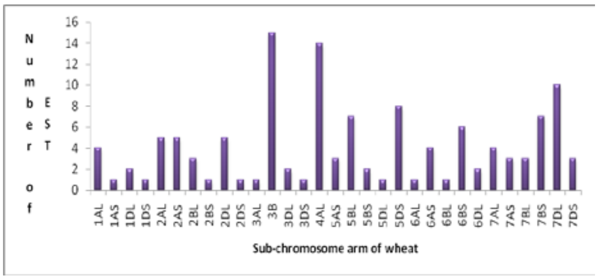


**Fig. 8: Distribution of SNPs containing ESTs across *T. aestivum* chromosomes and chromosome arms**

### Expression study of the SNP-containing sequences

To know the expression pattern of SNP-containing sequences in response to leaf rust infection, the SOLiD SAGE reads, mentioned earlier were used. The comparison was made between the reads of S-M *vs.* S-PI, R-M *vs.* R-PI and S-PI *vs.* R-PI. The SNP-containing sequences were taken as reference. On the comparison between S-M *vs.* S-PI, 71 sequences showed differential expression of which 60 were unregulated in S-M and 11 sequences have more expression in S-PI. The unregulated sequence of S-PI has shown homology with Fructose-bisphosphate aldolase, E3 ubiquitin-protein ligase RLIM, Cyclin-D1-binding protein 1 FAMILY, heat shock factor A6, CBL-interacting protein kinase 10 (table 8). Disease resistance protein RPM1 and a hypothetical protein (Armadillo-type fold) were expressed exclusively in S-PI..

Comparison between R-M *vs.* R-PI revealed 42 sequences to be differentially regulated of which 31 sequences are up regulated in R-M and 11 are up regulated in R-PI. The up regulated sequence of R-PI has shown homology with Fructose-bisphosphate aldolase, Peroxisomal membrane protein 2 and DEAD-box ATP-dependent RNA helicase 20 (table 9). Hypothetical protein F775_18732 and S-norcoclaurine synthase were uniquely expressed in R-PI.

Finally, on comparing S-PI *vs.* R-PI 82 sequences were found to be differentially expressed, of these 25 sequences have more expression in S-PI. In R-PI 57 sequences have more expression as compared to S-PI. The up regulated sequence of R-PI has shown homology with ELAV-like protein 1, Beta-1,3-galactosyltransferase 15, WW domain-containing oxidoreductase, Thioredoxin H-type, Disease resistance protein RPM1, Ubiquitin carboxyl-terminal hydrolase 12, Ring finger and transmembrane domain-containing protein 2, Defensin-like protein, Putative inactive receptor kinase, RNA polymerase Rpb7, Cryptochrome-1, Putative salt tolerance-like protein, Glutaredoxin-C1, DEAD-box ATP-dependent RNA helicase 20, Serine carboxypeptidase-like 19, ATP-dependent RNA helicase dhx8, CBS domain-containing protein etc. (table 10). Ubiquitin carrier protein E2 was uniquely expressed in R-PI. List of up regulated annotated ESTs with fold change has been provided in table 11. It was observed that many of the highly up regulated genes were not annotated.

**Table 8: SNP-containing annotated ESTs and fold changes with higher expression in S-PI as compared to S-M**

| SNP-containing EST | Fold change | Annotation |
|---|---|---|
| BQ237017 | 36.24 | Fructose-bisphosphate aldolase |
| CD490585 | 3.04 | E3 ubiquitin-protein ligase RLIM |
| CD491095 | 3.41 | Cyclin-D1-binding protein 1 FAMILY |
| CJ555209 | 5.29 | heat shock factor A6 |
| CJ564155 | 5.04 | Not available |
| CJ585290 | ∞ | Not available |
| CJ600598 | ∞ | Disease resistance protein RPM1 |
| CJ714721 | ∞ | hypothetical protein, Armadillo-type fold |
| CK163754 | 2.86 | CBL-interacting protein kinase 10 |
| CO346053 | 2.01 | Not available |
| CO349287 | 2.03 | Putative mediator of RNA polymerase II transcription subunit 6 |

∞ stands for infinity

**Table 9: SNP containing annotated ESTs and fold changes with higher expression in R-PI as compared to R-M**

| SNP-containing EST | Fold change | Annotation |
|---|---|---|
| BQ237017 | 2.42 | Fructose-bisphosphate aldolase |
| CJ632153 | 2.32 | Not available |
| CJ677583 | ∞ | hypothetical protein F775_18732 |
| CJ684250 | ∞ | Not available |
| CJ717347 | 3.87 | Peroxisomal membrane protein 2 |
| CJ725154 | ∞ | S-norcoclaurine synthase |
| CJ731128 | 2.03 | Not available |
| CJ849990 | ∞ | Not available |
| DR731556 | 2.5 | DEAD-box ATP-dependent RNA helicase 20 |
| HX181880 | 2.39 | Not available |
| HX194755 | 2.9 | Not available |

∞ stands for infinity

## DISCUSSION

Identification of SNPs in crop plants has been a challenging endeavour, irrespective of whether the whole genome or transcriptome is surveyed for SNPs [25]. Currently, no whole genome reference sequence is publicly available for wheat due to the large genome size and complexity of the genome. We utilized next generation sequencing data to identify SNPs. The strategy involved read mapping to a Transcript Assembly available at Gene Indices and crosschecked against EST reference database. The SNP outputs were annotated, and expression analysis was performed. Our strategy was to reduce the likelihood of false positive SNP discovery by setting stringent SNP discovery parameters and post-SNP discovery processing and minimize the possibility of false SNP identification from gene paralogues.

Defining robust SNP calling software parameters and minimum acceptable coverage is vital [26]. SNPs had to be represented on at least two independent reads, with stringent quality scores both for the SNP itself and the surrounding window of bases. The high-quality neighbourhood SNP scoring algorithm used in this study is very consistent for polymorphism calling and, where high coverage is present, very high specificity can be reached (<10 false positives per Mb) [21]. We chose a minimum base coverage of 20x for SNP calling as increasing minimum coverage to 25x and 30x was found by others to result in only modest gains in sensitivity, that is, the ability to detect a SNP [21]. When the stringency of the assembly parameters length fraction and similarity were increased from 0.5 and 0.8 to 0.9 and 0.9 respectively, the SNP output was significantly changed. The possibility of these SNPs being false due to the alignment of gene paralogues cannot be discounted, however, and could be stringently screened for by discarding a sequence that contained more than one SNP [27].

To assign putative functions of SNPs, we performed blastx searches of corresponding EST sequences against the non-redundant protein database available at NCBI. Blastx search results made it possible to assign putative functions of EST sequences. Of these, some EST sequences showed higher expression in response to infection with *Puccinia triticina.*

The greater part of these annotated contigs showed homology with plants and many of the top hits were from *Aegilops tauschii* whose genome sequence information is available [28]. SNPs in some important gene like Ubiquitin-related will be helpful in countering disease resistance as Ubiquitin-mediated protein modification contributes towards a defensive role in wheat against *P. triticina* [29]. This is particularly important since they could be considered as a valuable candidate gene for polymorphisms underlying important traits leading to the identification of resistance genes. However, these predictions were conducted using computational tools and functional data analyses are therefore needed to validate.

**Table 10: SNP-containing annotated ESTs and fold changes with higher expression in R-PI as compared to S-PI**

| SNP-containing EST | Fold change | Annotation |
|---|---|---|
| BJ220374 | 2.3 | Not available |
| BJ314338 | 5.48 | hypothetical protein TRIUR3_09559 |
| CA744898 | 3.28 | ELAV-like protein 1 |
| CJ509267 | 6.33 | Beta-1,3-galactosyltransferase 15 |
| CJ511172 | 4.64 | predicted protein |
| CJ526779 | 2 | Not available |
| CJ536987 | 5.16 | Not available |
| CJ538458 | ∞ | Ubiquitin carrier protein E2 |
| CJ547191 | 2.88 | WW domain-containing oxidoreductase |
| CJ552019 | 2.95 | hypothetical protein F775_31570 |
| CJ557944 | 5.23 | Not available |
| CJ576459 | 8.29 | Not available |
| CJ583250 | 3.68 | Not available |
| CJ583301 | 2.63 | Not available |
| CJ584805 | 6.67 | Not available |
| CJ600022 | 6.3 | Thioredoxin H-type |
| CJ600598 | 16 | Disease resistance protein RPM1 |
| CJ608054 | 3.37 | Not available |
| CJ609740 | 4.7 | hypothetical protein F775_10103 |
| CJ611385 | 4.67 | Not available |
| CJ615506 | 7.73 | predicted protein |
| CJ622247 | 101.34 | Ubiquitin carboxyl-terminal hydrolase 12 |
| CJ632153 | 3.88 | Not available |
| CJ632301 | 4 | Not available |
| CJ653541 | 28.8 | predicted protein |
| CJ655821 | 5.33 | Not available |

| | | |
|---|---|---|
| CJ661752 | 2.89 | Not available |
| CJ665107 | 2.09 | Ring finger and transmembrane domain-containing protein 2 |
| CJ670233 | 6.3 | Defensin-like protein |
| CJ676039 | 4.05 | Not available |
| CJ677583 | 5.33 | hypothetical protein F775_18732 |
| CJ680257 | 2.75 | Not available |
| CJ681303 | 2.46 | Putative inactive receptor kinase |
| CJ685340 | 3.56 | Not available |
| CJ688850 | 3.56 | Not available |
| CJ706431 | 2.67 | Not available |
| CJ710387 | 6.47 | RNA polymerase Rpb7 |
| CJ714199 | 4.62 | Cryptochrome-1 |
| CJ725461 | 4 | 50S ribosomal protein L27 |
| CJ731128 | 18.67 | Not available |
| CJ848862 | 7.65 | Putative salt tolerance-like protein |
| CJ849990 | ∞ | Not available |
| CJ884208 | 4.06 | hypothetical protein TRIUR3_20989 |
| CJ907530 | 6.88 | Glutaredoxin-C1 |
| CJ910160 | 4.8 | Not available |
| CK161193 | 2.01 | Not available |
| DR731556 | 7.87 | DEAD-box ATP-dependent RNA helicase 20 |
| GH726620 | 13.33 | Serine carboxypeptidase-like 19 |
| GH726775 | ∞ | predicted protein |
| GH731418 | 2.4 | Not available |
| HX085954 | 11.26 | ATP-dependent RNA helicase dhx8 |
| HX103789 | 3.96 | uncharacterized protein |
| HX103790 | 4.69 | uncharacterized protein |
| HX107602 | 9.86 | CBS domain-containing protein |
| HX167374 | 5.33 | zinc finger CCCH domain-containing protein |
| HX181880 | 2.22 | Not available |
| HX194755 | 64 | Not available |

∞ stands for infinity

**Table 11: SNP-containing annotated ESTs and fold changes with higher expression in S-PI as compared to R-PI**

| SNP-containing EST | Fold change | Annotation |
|---|---|---|
| BQ170192 | 2.26 | Not available |
| BQ237017 | 21.56 | Fructose-bisphosphate aldolase class-I |
| CD490585 | 5.17 | E3 ubiquitin-protein ligase RLIM |
| CD491095 | ∞ | Cyclin-D1-binding protein 1 FAMILY |
| CJ531178 | ∞ | Not available |
| CJ532803 | ∞ | Not available |
| CJ555209 | 3.94 | heat shock factor A6 |
| CJ555694 | ∞ | Not available |
| CJ563575 | ∞ | Not available |
| CJ585290 | ∞ | Not available |
| CJ622441 | ∞ | putative poly(A) polymerase |
| CJ672790 | ∞ | Homeobox-leucine zipper protein ROC8 |
| CJ699321 | ∞ | hypothetical protein F775_01328 |
| CJ725154 | 7.69 | S-norcoclaurine synthase |
| CJ734830 | 3.89 | Not available |
| CJ807624 | 7.14 | synbindin-like |
| CJ917632 | 5.7 | Ribulose bisphosphate carboxylase/oxygenase activase |
| CK163754 | 13.31 | CBL-interacting protein kinase 10 |
| CK192956 | 544.84 | hypothetical protein TRIUR3_30972 |
| CK205634 | 23.59 | Not available |
| CO346053 | ∞ | Not available |
| CO347122 | ∞ | Not available |
| CO349287 | 8.27 | Putative mediator of RNA polymerase II transcription subunit 6 |
| EF473215 | 25.5 | Not available |
| GR304906 | 3.06 | Not available |

We have demonstrated an approach for the rapid identification and verification of SNP-based genetic markers using EST data sources. The use of EST sequence data for the identification of SNPs has many advantages that can be exploited to facilitate the development of highly complex genetic maps of wheat. One of the main advantages of using EST sources is that markers closely associated with, or directly in the coding region of genes, can be identified, thus maximizing the density of a map toward gene-associated markers. In addition to finding variants in new genes, it is also possible that this approach could identify a large number of sequence variants. Discovering SNP with reference to leaf rust which is one of the major threats to wheat production will be very beneficial. Since, computational approaches dominate SNP discovery methods due to the ever-increasing sequence information in public databases, CLC genomics Workbench was employed for predicting the SNPs. In order to ensure that the discovered SNP is a Mendelian locus, it has to be validated. The validation of a SNP marker is the process of designing an assay based on the discovered polymorphism and then genotyping a panel of diverse germ plasm. Working with wheat is challenge where useful SNPs are only a small percentage of the total available polymorphisms. The present study will pitch light on the little-understood interaction of leaf rust with the wheat.

## CONCLUSION

The SOLiD reads were processed, and the putative SNPs were discovered by CLC Genomics Workbench. The predicted SNPs were filtered individually by performing BLAST of the sequence containing the SNPs with wheat ESTs. After screening, 191 SNPs were finally selected out of 10,012 SNPs. All the 191 SNP-containing sequences were annotated using the Blast2GO. In the lack of a reference genome, EST resources represent an attractive approach for *in silico* SNP identification. The SNP discovery method and application system established in this study was fast and cost effective.

## ACKNOWLEDGEMENT

The authors are thankful to BTISNet SubDIC (BT/BI/04/065/04) for providing facilities for bioinformatics analyses. This work was supported by Department of Biotechnology, Government of India (Grant No. BT/PR6037/AGR/02/308/05 and BT/PR3925/BID/7/384/2011) and Centre of Excellence, Technical Education Quality Improvement Program-II, (Grant No. NPIU/TEQIP II/FIN/31/158).

## CONFLICT OF INTERESTS

Declared none

## REFERENCES

1. Jander G, Norris SR, Rounsley SD, Bush DF, Levin IM, Last RL. *Arabidopsis* map-based cloning in the post-genome era. Plant Physiol 2002;129:440–50.
2. Rostoks N, Borevitz JO, Hedley PE, Russell J, Mudie S, Morris J, *et al.* Single-feature polymorphism discovery in the barley transcriptome. Genome Biol 2005;6:R54.
3. Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, *et al.* SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 2002;3:19.
4. Shen YJ, Jiang H, Jin JP, Zhang ZB, Xi B, He YY, *et al.* Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. Plant Physiol 2004;135:1198–205.
5. McNally KL, Bruskiewich R, Mackill D, Buell CR, Leach JE, Leung H. Sequencing multiple and diverse rice varieties. Connecting whole-genome variation with phenotypes. Plant Physiol 2006;141:26–31.
6. Zhu YL, Song QJ, Hyten DL, Van, Tassell CP, Matukumalli LK, Grimm DR, *et al.* Single-nucleotide polymorphisms in soybean. Genetics 2003;163:1123–34.
7. Ablett G, Hill H, Henry RJ. Sequence polymorphism discovery in wheat microsatellite flanking regions using pyrophosphate sequencing. Mol Breed 2006;17:281–9.
8. Ravel C, Praud S, Murigneux A, Canaguier A, Sapet F, Samson D, *et al.* Single-nucleotide polymorphism frequency in a set of selected lines of bread wheat (*Triticum aestivum* L.). Genome 2006;49:1131–9.
9. Bundock PC, Cross MJ, Shapter FM, Henry RJ. Robust allele-specific polymerase chain reaction markers developed for single nucleotide polymorphisms in expressed barley sequences. Theor Appl Genet 2006;112:358–65.
10. Somers DJ, Kirkpatrick R, Moniwa M, Walsh A. Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. Genome 2003;46:431–7.
11. Choi IY, Hyten DL, Matukumalli LK, Song Q, Chaky JM, Quigley CV, *et al.* A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. Genetics 2007;176:685–96.
12. Rostoks N, Mudie S, Cardle L, Russell J, Ramsay L, Booth A, *et al.* Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. Mol Genet Genomics 2005;274:515–27.
13. Grivet L, Glaszmann JC, Arruda P. Sequence polymorphism from EST data in sugarcane: a fine analysis of 6-phosphogluconate dehydrogenase genes. Genet Mol Biol 2001;24:161–7.
14. Grivet L, Glaszmann JC, Vincentz M, da Silva F, Arruda P. ESTs as a source for sequence polymorphism discovery in sugarcane: an example of the *Adh* genes. Theor Appl Genet 2003;106:190–7.
15. Cordeiro GM, Eliott F, McIntyre CL, Casu RE, Henry RJ. Characterisation of single nucleotide polymorphisms in sugarcane ESTs. Theor Appl Genet 2006;113:331–43.
16. Longeri M, Zanotti M, Damiani G. Recombinant DRB sequences produced by mismatch repair of heteroduplexes during cloning in *Escherichia coli.* Eur J Immunogenet 2002;29:517-23.
17. Dubcovsky J, Dvorak J. Genome plasticity a key factor in the success of polyploid wheat under domestication. Science 2007;316:1862–6.
18. Haudry A, Cenci A, Ravel C, Bataillon T, Brunel D, Poncet C, *et al.* Grinding up wheat: a massive loss of nucleotide diversity since domestication. Mol Biol Evol 2007;24:1506–17.
19. Singh D, Bhaganagare G, Bandopadhyay R, Prabhu KV, Gupta PK, Mukhopadhyay K. Targeted spatio-temporal expression based characterization of state of infection and time-point of maximum defense in wheat NILs during leaf rust infection. Mol Biol Rep 2012;39:9373-82.
20. Altshuler D, Pollara VJ, Cowles CR, Etten WJV, Baldwin J, Linton L, *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 2000;407:513-6.
21. Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, Lee WL, *et al.* Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res 2008;18:763-70.
22. Conesa A, Götz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005;21:3674-6.
23. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Statistical Society: Series B 1995;57:289–300.
24. Kal AJ, van Zonneveld AJ, Benes V, den Berg MV, Koerkamp MG, Albermann K, *et al.* Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. Mol Biol Cell 1999;10:1859-72.
25. Ganal MW, Altmann T, Roder MS. SNP identification in crop plants. Curr Opin Plant Biol 2009;12:211-7.
26. Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, *et al.* Identification of genetic variants using barcoded multiplexed sequencing. Nat Methods 2008;5:887-93.
27. Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, *et al.* SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods 2008;5:247-52.
28. Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, *et al. Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. Nature 2013;496:91–5.
29. Fofana B, Banks TW, McCallum B, Strelkov SE, Cloutier S. Temporal gene expression profiling of the wheat leaf rust pathosystem using cDNA microarray reveals differences in compatible and incompatible defence pathways. Int. J. Plant Genomics 2007. doi: 10.1155/2007/17542. [Article in Press].