In silico analysis and expression profiling of *Expansin A4, BURP domain protein RD22-like* and *E6-like* genes associated with fiber quality in cotton

Farzana Ashraf1 · Asif Ali Khan1 · Nadia Iqbal2 · Zahid Mahmood3 · Abdul Ghaffar4 · Zulqurnain Khan1

Received: 4 February 2022 / Accepted: 25 March 2022 / Published online: 12 May 2022 © The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract

**Background** To supply high-quality cotton fibre for the textile industry, the development of long, strong and fine fibre cotton varieties is imperative. An interlinked approach was used to comprehend the role of fibre genes by analyzing interspecific progenies of cotton species. Wild *Gossypium* species and races are rich source of genetic polymorphism due to environmental dispersal and continuous natural selection. These genetic resources hold mass of outclass genes that can be used in cotton improvement breeding programs to exploit possible traits such as fibre quality, abiotic stress tolerance, and disease and insect resistance. Therefore, use of new molecular techniques such as genomics, transcriptomics and bioinformatics is very important to utilize the genetic potential of wild species in cotton improvement programs.

**Methods** Interspecific lines and *Gossypium* species used in the study were grown at Central Cotton Research Institute (CCRI), Multan. After retrieving DNA sequence of the genes from NCBI, the primers for gene expression and full-length gene sequence were designed. Expression profiling of *Expansin A4, BURP Domain protein RD22-like* and *E6-like* fibre genes was performed through Real Time PCR. BLAST and DNA sequence alignment was conducted for sequence comparison of interspecific lines and *Gossypium* species. Different in silico analysis were used for characterization of fibre genes and identification of cis acting promoter elements in promoter region.

**Results** Variable expression of genes related to fibre development was observed at different stages. BLAST and DNA sequence alignment demonstrated resemblance of interspecific lines with *G. hirsutum*. In silico analysis on the sequence data also confirmed the role of *Expansin A4, BURP Domain protein RD22-like* and *E6-like* fibre genes in fibre development. Genetic engineering is also recommended by transferring *E6-like, Expansin A4* and *BURP Domain RD22-like* genes in local cotton cultivars. Similarly, several stress tolerant and light responsive cis acting elements were identified through promotor analysis, which may contribute for fibre development in the breeding programs.

**Conclusion** *Expansin A4, BURP Domain RD22-like* and *E6-like* have positive role in fibre development with variable expression at fiber length and strength associated stages.

**Keywords** DNA sequencing · Expression analysis · Fibre genes · In silico analysis · Interspecific lines · Cotton

**Abbreviations**

- BLAST: Basic local alignment search tool
- cDNA: Complementary deoxyribonucleic acid
- DNA: Deoxyribonucleic acid
- dNTP: Deoxyribonucleotide triphosphate
- DPA: Day’s post anthesis
- EDTA: Ethylene diamine tetraacetic acid
- KDa: Kilodalton
- LTP: Lipid transfer protein
- mg: Milligram
- mm: Millimeter
- mRNA: Messenger RNA
- MgCl\(_2\): Magnesium chloride
Introduction

Globally, synthetic fibre consumption is continuously increasing and projected to reach at 130 million tons by 2030. The consumption of synthetic fibre is 62.7% compared to 24.3% cotton fibre consumption [1]. Competition of cotton fibre with polyester is creating negative influence on the demand of cotton. Genetic improvement of cotton for fibre traits is very crucial to meet the challenges of the textile industry. So, there is need to devise clear-cut policies for cotton breeding program to enhance the quality cotton production. In a breeding program, germplasm collection, its conservation and utilization, trait specific screening programs and modern genomics have key role in variety development [2]. Cotton genetic resources have been extensively studied over the last many decades to introduce valuable traits in cotton [3–5]. These genetic resources include wild Gossypium germplasm, innovative cygnetic stocks with specific chromosomes additions or deletions in different species, large mapping families, recombinant inbred lines, near isogenic lines and interspecific lines. While there are some queries about narrow genetic base of these cultivars and most breeders would admit that in breeding programs maximum utilization of genetic diversity within their material should be ensured. Breeders will have to utilize wild cotton relatives, as well as advance lines or cultivars to develop cotton varieties with superior traits.

Cotton fibre is made of a highly stretched and condensed epidermal trichome single cell. Generally, fibre cell progresses into four overlapping stages including initiation, elongation, cell wall biosynthesis, and finally maturation [6–10]. Cotton fibre start to yield from 3 days before anthesis to 3 days post anthesis by means of epidermal cells enlargement [11]. Elongation starts from 2 DPA and continue unto 20 DPA after the initiation. These elongated fibres get twisted and produce bundles of fibre [12, 13]. At the cellular level, cotton fibre development is supported by several genes which facilitates the elongation process, for example, Expansins are involved in fibre elongation at various development stages [14]. High transcript abundance of GhEXP1 was observed in cotton fibre during the elongation phase of fibre development, which steadily decreased from 16 to 20 DPA [15, 16]. In cotton, GhEXP1 along with GhRDL1 showed an increase in fiber length and an enlargement of endopleura cells of ovules [17]. The BURP Domain is a plant-specific protein characterized by repetitive units of amino acid [18]. This protein is mainly involved in promoting the fibre cells elongation when over-expressed. Because GhRDL1 directly interacts with cotton α-Expansin fibre gene therefore, Expansins mediate GhRDL1’s effect on overall fibre cell enlargement [17, 19]. It was suggested that E6 protein is involved in fibre development, but no support was present to justify this hypothesis as no conclusive evidence was presented [20]. When E6 antisense suppression construct was used, there was knockdown to uncover a phenotype E6-like. E6 proteins play a comprehensive role in cell wall, and are deposited during fibre elongation, which give high transcripts in fibre cell during transcriptomic analysis [21].

Transcriptional profiling is a unique tool to gain knowledge about gene mechanisms, regulatory pathways, and gene expression [22, 23]. Number of techniques are used for specific gene expression studies but Real Time PCR is the most reliable technology for absolute and comparative quantification of the gene transcription [24]. This comprehensive wide-ranging gene expression study is supportive to sightsee the role of genes, which are up regulated, entirely expressed, or down regulated during different cotton fibre development stages. Through transcriptomic data, one can explain the fibre expansion process and can discover highly expressed genes for the development of transgenic cotton varieties with superior fibre traits. Profiling of fibre genes in interspecific lines will enable us to unravel variable expression pattern of selected fibre genes.

Application of in silico methods along with expression profiling is important for characterization of fibre genes. DNA sequence of interspecific lines and Gossypium species were aligned to have information about differences and similarities. Diploid and tetraploid genomes of various Gossypium species have repeatedly sequences making their entire genome sequences. These valuable repeatedly sequenced data revealed the evolutionary history of the cotton with polyploidization and decaploidization leading to the of the formation of genus Gossypium [25]. Multiple sequence alignment approaches envisage algorithmic explanation about evolutionarily sequences alignments. Fibre genes were subjected to BLAST analysis for expression validation and multiple DNA sequence alignment for similarities and differences of interspecific lines and parent species. Genomics combines recombinant DNA technology, DNA sequencing and bioinformatics sequence to analyze the structure and function of genes [26]. Bioinformatics is a systematic field that utilizes advance approaches for computational analysis of biological data [26]. Bioinformatics also aids to recognize different promoters involve in fibre yield and quality, abiotic stress tolerance and disease resistance. Strength and specificity related character of promoter sequence can be demonstrated through expression profiling. Strong promoters predict high expression and vice versa.
Fibre genes protein E6, Expansin A4 and BURP Domain RD22-like also have strong promoters, which can be used in future breeding program.

Cotton breeders have extensively carried out interspecific hybridization for utilization of desirable genes from wild species to cultivated cotton and developed interspecific cotton varieties. Among them, a lot of upland cotton lines with improved traits including fibre quality and insect pest resistance have been developed [27–30]. All these upland cotton lines are designated as introgression lines of interspecific hybridization. These interspecific lines with their practical value in cotton breeding program have changed genetic basis from narrow line to a wide broad base in the present upland cotton germplasm and have broken the bottlenecks of breeding. However, the full potential of interspecific lines have not yet been obtained for beneficial traits exploitation in traditional and advanced breeding programs [31]. Therefore, this study was designed to evaluate the expression of fibre genes in diverse interspecific lines and Gossypium species and their role in different fibre development stages. Results of this study will be directive for development of high-quality cotton varieties.

Materials and methods

DNA sequence retrieval and primer designing

DNA sequences of selected fibre genes (Expansin A4, BURP Domain protein RD22-like and E6-like) were retrieved from NCBI website https://www.ncbi.nlm.nih.gov/. RT-PCR Primers were designed using PRIMER 3.0 software (Table 1).

Collection of fibre tissues

Three interspecific lines (SL-19, SL-79 and SL-369) of varying fibre length categorized as long fibre (34.7 mm), medium fibre (28.5 mm) and short fibre (24.0 mm) along with three parent species (G. arboreum, G. anomalum and G. hirsutum) were used for fibre tissue collection. Cotton bolls were collected at different stages (0, 0.5, 10, 15 and 20 days after anthesis). Collected bolls were rinsed with diethyl pyrocarbonate (DEPC) treated water and were stored in liquid nitrogen. These frozen bolls were further used for RNA extraction.

Plant RNA extraction and cDNA synthesis

RNA was extracted following Gynidium isothiocynate method [32, 33]. RNA quality was observed by electrophoresis and monitored under UV light. RNA samples were quantified through nanodrop (Thermo Scientific ND 2000) and concentrations was optimized prior to cDNA synthesis. Extracted RNA from fibre tissues was used for cDNA synthesis. Working solution of synthesized cDNA was prepared according to thermo scientific cDNA kit (K1622) by diluting it to 25 ng. This mixture was partitioned into two parts Total RNA was reverse transcribed to cDNA using 2 µl of the dNTPs mix, 0.5 µl Ribo Lock RNase inhibitor and 0.5 µl Revert Aid Reverse Transcriptase. Reverse transcriptase was used to reverse transcribe the RNA into cDNA.

Real time PCR analysis

Real Time PCR was performed with SYBR Green Super Mix (Bio-Rad, USA) and 10 ng/µl of listed primers (Table 1). 18S rRNA constitutive gene primers were used as data normalizer in this assay. A master mix of iQ SYBR Green Supermix (BioRad) was prepared containing the primers and all other reagents. A 25 µl reaction contains iQ SYBR Green Supermix 2× (12.5 µl), 1 µl Forward primer (25 ng/µl), 1 µl Reverse primer (25 ng/µl), 3 µl cDNA (1:10) dilution and 6.5 µl Sterile water. The PCR was performed using the following cycling conditions: initial denaturation at 95 °C for 4 min, followed by 40 cycles of 95 °C for 1 min, 55 °C for 1 min, and 72 °C for 50 s, with a final extension

| Gene annotation     | Primer pair | Primer sequence (5’–3’) | Primer length | Product length (bp) | Accession no. |
|---------------------|-------------|-------------------------|---------------|---------------------|---------------|
| 18S rRNA            | RT18S-F     | AAACGGCTACCACTCCAAG     | 20            | 153                 | U42827.1      |
|                     | RT18S-R     | CCTTCAATGGATCTCTGGTA    | 20            |                     |               |
| E6-like             | RTE6-F      | ATGCGTCTCCTACAAATCTCTCTTCTTCT | 25           | 211                 | DQ023519      |
|                     | RTE6-R      | TTTCAGAGGATACCTGTGCTTCTT | 24            |                     |               |
| Expansin A4         | RT EXPF     | ATGGCAACAAAAAGATGTG     | 22            | 220                 | DQ204495      |
|                     | RT EXPFR    | AAGCTGCTGCTGTGCTTCCCAT  | 21            |                     |               |
| BURP Domain RD22-like | RTRD22-F  | ATGAAGGTTCTCTCCCAATTCTCTCTCT | 23           | 198                 | XM_016894801  |
|                     | RTRD22-R    | GACGTTTACACACCACCTCCTCTCCTC | 22           |                     |               |
at 72 °C for 10 min. Melt curve analysis (55 °C to 95 °C) followed by 81 cycles for 0.5 min.

**Data analysis and relative quantification**

Comparative expression of each fibre gene was known by reference gene 18S rRNA as internal control. Replicated reaction was done for individually fibre gene to minimize error. For template normalization, similar reactions with 18S rRNA primers were also accomplished. *G. hirsutum* 10 DPA was used as calibrator. Data were normalized by average Ct values of 18S rRNA reference gene with average Ct values of samples. Expression was calculated by using standard Ct method [34].

**Full length gene specific primer designing**

Full length primers (Table 2) were retrieved from phytozome [https://phytozome.jgi.doe.gov/pz/portal.html].

**Sequencing of PCR product**

PCR products of full-length primers were sent to Macrogen Korea for Sanger sequencing. Sequencing PCR was performed using gene specific forward primers.

**Sequencing comparison of interspecific lines and *Gossypium* species**

Multiple alignment of predicted DNA sequences and phylogenetic tree analysis was performed at [https://www.ebi.ac.uk/Tools/msa/clustalo] [35].

**In silico analysis of fibre genes**

Sequences of fibre genes were taken from NCBI database [https://www.ncbi.nlm.nih.gov/] by searching accession number in all data bases. Coding sequences were identified with amino acid residues. Translation of gene sequence into amino acid sequences was done through EXPASY [https://web.expasy.org/translate/] into six reading frames.

**Theoretical computation of physicochemical properties**

Basic physiochemical properties and hydropathy index of protein sequences were computed through Expasy’s ProtParam Proteomic server [http://web.expasy.org/protparam/].

**Functional annotation of protein**

For Subcellular Location DeepLoc-1.0 [http://www.cbs.dtu.dk/services/DeepLoc] databases was used. Moreover, SignalP 4.0 [http://www.cbs.dtu.dk/services/SignalP/] was used to check existence of signal peptide.

**Promoter sequence analysis**

Promoter analysis was carried out at [http://bioinformatics.psb.ugent.be/webtools/plantcare/html/].

**Results**

**Expression profiling of *Expansin A4*, *BURP Domain protein RD22-like* and *E6-like***

Overall expression of *Expansin A4* gene was remarkably high in rapid elongating fibre during 10 DPA in all interspecific lines and *Gossypium* species. Maximum transcripts were found in SL-19 (Fig. 1). Expression of *BURP Domain protein RD22-like* remained constant from 10 to 20 DPA fibre in all genotypes except in *Gossypium anomalum*. Transcripts of *BURP Domain protein RD22-like* gene were maximum in 10 DAP fibre as compared to 5 DPA. In all three interspecific lines, highest expression was detected at 15 and 20 DPA fibre stages in SL-19, SL-79 and SL-369 respectively (Fig. 2). Expression pattern of *E6-like* showed that high expression was detected at 10 and 15 DPA fibre stages predicting its main role in fibre elongation. In interspecific lines, transcripts of *E6-like* gene were variable from 0 DPA till 20 DPA. In SL-19, expression of fibre gene starts to increase from 0 DPA and reached at maximum level at 15 DPA and then slightly decreases at 20 DPA (Fig. 3).

To validate expression results, the target gene transcriptomic profiles (*E6-like*, *Expansin A4* and *BURP Domain protein RD22-like*) were validated by using existing RNA-seq
Fig. 1 Expression profiling of Expansin A4 in Gossypium species and interspecific lines: a expression of Expansin A4 in G. arboreum, b expression of Expansin A4 in G. hirsutum, c expression of Expansin A4 in G. anomalum, d expression of Expansin A4 in SL-19, e expression of Expansin A4 in SL-79, f expression of Expansin A4 in SL-369
Fig. 2 Expression profiling of BURP Domain RD-22 in Gossypium species and interspecific lines: a (Expression of RD-22 in G. arboreum), b expression of RD-22 in G. hirsutum, c expression of RD-22 in G. anomalum, d expression of RD-22 in SL-19, e expression of RD-22 in SL-79, f expression of RD-22 in SL-369)
Fig. 3 Expression profiling of E6-like in Gossypium species and interspecific lines: a expression of E6-like in G. arboreum, b expression of E6-like in G. hirsutum, c expression of E6-like in G. anomalum, d expression of E6-like in SL-19, e expression of E6-like in SL-79, f expression of E6-like in SL-369
data on Cotton FGD. The results of available fibre specific genes were generally similar with our expression analysis results. Heat map was created on the basis of RNA-seq data of related expressed in transcript per million (TPM) during different fibre development stages. E6-like, Expansin A4 and BURP Domain RD22-like showed similarity with gene Gh-D05G160200, Gh_A10G149600 and Gh_D05G052400 respectively. (Fig. 4). An expression trend of gradual increasing from 5 to 10 DPA were identified, while similar tendencies were also observed in our experiment.

### In silico analysis of E6-like, Expansin A4 and BURP Domain RD-22

**Physicochemical properties**

Expasy’s Protpam analysis of predicted protein showed that Protein E6-like and BURP Domain RD22-like was characterized as unstable as value of instability index was 47.75 and 44.72 respectively (Table 3). Expansin A4 was characterized as a stable protein with value of instability index of 29.01.

**Subcellular localization**

DeepLoc analysis designated that protein. Proteins E6-like, Expansin A4 and BURP Domain RD22-like were a membrane soluble protein family. Location in different organelles with the approximate values (Table 4) predicted the probability of protein location in different organelles. Highest Extracellular values of Proteins E6-like, Expansin A4 and BURP Domain RD22-like (0.819, 0.729 and 0.843 respectively) showed that these proteins are extracellular.

---

![Fig. 4 Heat map of expression levels (log-transformed transcript per kilobase million (TPM) values). Figure was generated based on available RNA-seq data of BURP Domain RD2-like-2, Expansin and E6-like submitted bio projects from cotton FGD data base. Red indicates high expression, yellow indicates intermediate expression and green indicates no expression. It is straightforward to identify highly expressed genes in specific tissues from this figure. Tissues are labeled with days after post anthesis (DPA). Rows indicates the fibre gens and column show the fibre stages (250DPA ovule -25 DPA fibre). The data denotes the logarithm-transformed values of log2, day post anthesis and fragments per kilobase of transcript per million mapped reads](image-url)
Fig. 5 DNA sequence alignments of fibre genes. a E6-like, b Expansin A4, c BURP Domain RD22-like. White shadings indicate the polymorphic nucleotides. Interspecific lines and *Gossypium* species names are indicated in the left and number of bases depicted in each line is marked by the number shown at the top right of each section.

Table 3 Physicochemical properties of E6 like, Expansin A4 and BURP Domain RD22-like

| Physicochemical properties                                             | E6-like | Expansin A4 | BURP Domain RD22-like |
|------------------------------------------------------------------------|---------|-------------|-----------------------|
| Number of amino acids                                                  | 241     | 258         | 335                   |
| Total negatively amino acid charged residues (Asp + Glu)              | 37      | 13          | 35                    |
| Total positively amino acid charged residues (Arg + Lys)               | 25      | 16          | 34                    |
| Molecular weight                                                      | 28.223.37 | 27.936.46   | 36.595.05             |
| Theoretical pI                                                          | 5.00    | 8.36        | 6.89                  |
| Aliphatic index                                                        | 32.37   | 62.83       | 75.64                 |
| Grand average of hydropathicity (GRAVY)                               | - 1.356 | - 0.090     | - 0.266               |
| Instability index (II)                                                 | 47.75   | 29.01       | 44.72                 |
**Signal peptide analysis**

In *E6-like*, *Expansin A4* and BURP Domain RD22 were characterized as extracellular membrane that's why signal peptide was present in protein coding sequence. Score values of C, S, 3Y is more than 0.45 (Table 5) that shows that peptide signal is present.

**Promoter sequence analysis**

Sequence analysis of cotton *E6-like*, *Expansin A4* and BURP Domain protein RD22-like promoter using PlantCARE predicted many vital motifs in this region related to gene expression. There are few transcriptions activation related motifs along with core promoter elements like TATA and CAAT boxes. These motifs are light responsive, hormone and stress regulated cis elements. These motifs are involved in the light, stress and hormones responsiveness. There were other vital core promoter elements required for promoter activity including TATA box and CAAT box (Tale-6). Cis-acting essential element for the abscisic acid reaction (*Hordeum vulgare*), light response elements (*Arabidopsis thaliana*), gibberellin-enhancer element (*Brassica oleracea*) and element for variation of the palisade mesophyll cells (*Arabidopsis thaliana*) were present in *E6-like* promoter region. Similarly, in *Expansin A4* various cis acting pre-motor elements were identified. Abscisic acid responsiveness elements were identified in *Arabidopsis thaliana*, light responsiveness in *Zea mays*, element responsive for transcription start in *Brassica oleracea* and MeJA-responsiveness in *Hordeum vulgare*. In BURP Domain RD22-like, elements essential for light responsiveness were present in *Petroselinum crispum* while promoter and enhancer regions were identified in *Arabidopsis thaliana*. MYBHv1 binding site, MeJA and anaerobic induction responsive elements were present in *Hordeum vulgare* and *Zea mays* respectively.

**Discussion**

Realistic genetic resources are accessible for innovative cotton breeders to make more perfection in crop improvement. Transcriptomic analysis of interspecific lines and *Gossypium* species for fibre traits identified in this study will improve our understanding of fibre genes that have key role in fibre development. Transcriptomic analysis simplifies the breeding through expression profiling of highly expressed genes. Transcriptomic analysis was performed for the identification of differentially expressed genes at different fibre growth stages in interspecific lines and three *Gossypium* species. Our study predicts expression analysis of selected fibre genes during 0, 5, 10, 15 and 20 DPA fibre stages. High level variable regulation of genes encoding for fibre development was observed at different stages. Transcriptomic profiling has been effectively used for gene identification in cotton crop [36–40]. Here, we describe transcriptome profiling of genes in cotton fibre through quantitative Real Time PCR.

This is the initial comprehensive expression profiling that identified the differentially expressed genes with different stages contributing to fibre development in contrasting interspecific lines of cotton. Real Time PCR results predicted high expression levels specifically in the interspecific lines SL-19 (long staple line) as compared to parent species (Figs. 1–3) envisaging that when genome of two different species merge with each other, its progenitors possess more DNA content, which can be associated with fibre elongation.

### Table 4 Predicted subcellular localization of *E6-like*, *Expansin A4* and BURP Domain RD22-like

| Fibre gene          | Extracellular | Lysosome | Endoplasmic reticulum | Cell membrane | Golgi apparatus | Cytoplasm |
|---------------------|---------------|----------|-----------------------|---------------|----------------|-----------|
| *E6-like*           | 0.8195        | 0.1706   | 0.0083                | 0.0013        | 0.0002         | 0.0002    |
| *Expansin A4*       | 0.7293        | 0.2373   | 0.0329                | 0.0005        |               | 0         |
| BURP Domain RD-22   | 0.8435        | 0.1316   | 0.0237                | 0.0008        | 0              | 0.0003    |

### Table 5 Signal peptide analysis of *E6-like*, *Expansin A4* and BURP Domain RD22-like

| Fibre gene          | Measure | Position | Value | Cut Off | Signal Peptide |
|---------------------|---------|----------|-------|---------|----------------|
| *E6-like*           | Max.C   | 26       | 0.792 |         |                |
|                     | Max.Y   | 26       | 0.840 |         |                |
|                     | Max.S   | 15       | 0.941 |         |                |
|                     | Mean S  | 1–25     | 0.891 |         |                |
|                     | D       | 1–25     | 0.868 | 0.45    | Yes            |
| *Expansin A4*       | Max.C   | 30       | 0.427 |         |                |
|                     | Max.Y   | 30       | 0.586 |         |                |
|                     | Max.S   | 9        | 0.950 |         |                |
|                     | Mean S  | 1–29     | 0.821 |         |                |
|                     | D       | 1–29     | 0.713 | 0.45    | Yes            |
| BURP Domain RD22    | Max.C   | 30       | 0.427 |         |                |
| Domain RD22-like    | Max.Y   | 30       | 0.586 |         |                |
|                     | Max.S   | 9        | 0.950 |         |                |
|                     | Mean S  | 1–29     | 0.821 |         |                |
|                     | D       | 1–29     | 0.713 | 0.45    | Yes            |
and amplified size of single-celled fibres. It was also concluded that transgressive segregates are possible with hybrid vigor because of different genome groups of *Gossypium*, which make it possible to get interspecific lines with good fibre length, fibre strength and fibre fineness [41–44].

Expression profiling was compared with RNA sequence data submitted in different bio projects on FGD (Fig. 5). In *Expansin A 4*, our results were according to PRJNA490626 project in which transcripts were detected in 5 experiments including fibre development at various stages (0–25 DPA). Maximum expression was at 10 DPA which was similar to our results. GhEXPA4a and GhEXPA4b are specific fibre related genes that had high expression during the fibre initiation and elongation stages (0 to 15 DPA). Over-expression of *GhEXPA8* predicted that these genes have ability to improve the fibre length and fineness in cotton crop [14]. Expansin proteins indorse the spillage between different microfibrils by Hemicellulose and cellulose cleavage [45]. Moreover, our data also suggested that Expansin protein has essential role in cotton fibre development by enlargement of fibre cells through sliding apart cellulose micro fibrils. Expression levels for *E6-like* genes was also compared. *E6-like* gene has similarity with genes Gh-D05G160200 for fibre related gene. It also plays its role fibre development. *E6* gene was firstly recognized as fibre gene with high expression during cotton fibre development and similar *E6-like* was predicted in Angiosperms [21].

**BURP Domain** proteins are known as important proteins that has significant roles in plant growth and stress responses [46, 47]. Number of BURP proteins have been recognized and characterized on the basis of sequences features. However, different members from different subfamilies predicted variable expression patterns. In our findings, **BURP Domain RD22-like** genes actually execute main function in fibre elongation and maturation. Although low copy number of TPM of **BURP Domain RD22-like** gene were observed but this has a role in fibre development. The cotton fibre related gene (AtRD-22-Like) with over expression in elongating fibre cells, translates a BURP Domain-containing protein [17]. Cotton plants with high expression of GhRD1L1 and *GhEXPA1* give more number of bolls, resulting up to 40% more lint yield plant−1 without disturbing fibre quality and non-reproductive growth [17].

It is further concluded from the study that there is a direct association between *Expansin A4*, *E6-like*, **BURP Domain protein RD22-like** and fibre quality traits. Thus, these are key target for improving the fibre characteristics. Transformation of these highly expressed genes in local cotton varieties can fulfill the mechanized textile industry requirements. Moreover, genetically modified cotton produced by over expression of these genes will be the best source for use as a long staple variety or use as a parent in breeding program.

Biological sequences comparison in molecular biology and bioinformatics has been an imperative approach to supports analysis, such as prediction of protein sub-cellular localization [48], Physio chemical properties [49] and the field of taxonomy [50]. *E6-like* was characterized as unstable as value of instability index was 47.75. A protein whose instability index is less than 40 is expected as stable while a value greater than 40 indicates that the protein may be unstable. Similarly, *Expansin A4* was characterized as a stable protein with value of instability index of 29.01. An imperative step on this mode is prediction of subcellular localization of each protein. *E6-like*, *Expansin A4* and **BURP Domain RD22-like** were characterized as a membrane soluble protein family. In silico analysis also confirm the role of genes in fibre elongation, *Expansin-A4*, **BURP Domain protein RD22-like** and *E6-like* play its main role in rapid elongation and also with predominantly effect in transition stage of elongation supporting to secondary cell wall synthesis.

DNA sequence alignment is a criterion for almost all comparative genomic analyses, including documentation of well-preserved sequence motifs and investigation of genes and species historical relationships [51]. *E6-like*, *Expansin A4* and **BURP Domain RD22-like** PCR amplified full length gene was sequenced and subjected to BLAST analysis followed by multiple sequence alignment of DNA sequence and protein sequence for similarities and differences of interspecific lines and parent species (Fig. 5). It was concluded from the sequence comparison of interspecific lines and species of cotton that tri-species introgression lines are more closely related to *Gossypium hirsutum* as compared to *Gossypium arboreum* and *Gossypium anomalum* depicted. This confirms its back crossing with *G. hirsutum* for yield improvement. These interspecific lines were also originate from BC$_2$S$_2$ population (*G. hirsutum* × 2(*G. arboreum* × *G. anomalum*)) developed at Cytogenetics Section, CCRL, Multan [30]. In interspecific hybrids of *Gossypium*, a greater proportion of female gametes than male gametes is generally useful with few exceptions [52], hence backcross breeding should be subjugated. Review of backcrossing with distinct reference to cotton traits improvement showed that during repeated backcrossing one set of chromosomes retained with genes balanced. This technique has been used successfully in crosses of different *Gossypium* species [53–55].

In silico analysis tries to find proteins with consistent annotations about their interaction and functions in the cellular machinery. An imperative step on this mode is prediction of subcellular localization of each protein. *E6-like*, *Expansin A4* and **BURP Domain RD22-like** were characterized as a membrane soluble protein family. In *E6-like*, *Expansin-A4* and **BURP Domain RD22-like** were characterizes as extracellular membrane that’s why signal peptide was present in protein coding. As validation of specific genes for crop
### Table 6 Cis acting promoter elements in promoter region

| Site name        | Organism                | Position | Strand | Score | Sequence       | Function                                                                 |
|------------------|-------------------------|----------|--------|-------|----------------|--------------------------------------------------------------------------|
| **E6-like**      |                         |          |        |       |                |                                                                          |
| ABRE             | Hordeum vulgare         | 425      | −      | 9     | GCAACGTGTC     | Cis-acting element involved in the abscisic acid responsiveness            |
| AE-box           | Arabidopsis thaliana    | 748      | +      | 8     | AGAAACAA       | Part of a module for light response                                      |
| CAAT-box         | Arabidopsis thaliana    | 638      | +      | 5     | CCAAT          | Common cis-acting element in promoter and enhancer regions              |
| CAAT-box         | Pisum sativum           | 852      | −      | 5     | CAAAT          | Common cis-acting element in promoter and enhancer regions              |
| GARE motif       | Brassica oleracea       | 615      | −      | 7     | TCTGTG        | Gibberellin-responsive element                                          |
| HD-Zip 1         | Arabidopsis thaliana    | 564      | −      | 8     | CAAT(A/T) ATTG | Element involved in differentiation of the palisade mesophyll cells     |
| TATA-box         | Arabidopsis thaliana    | 575      | −      | 4     | TATA           | Core promoter element around -30 of transcription start                 |
| TC-rich repeats  | Nicotiana tabacum       | 380      | +      | 9     | GTTTTCCTAC     | Cis-acting element involved in defense and stress responsiveness        |
| TCT-motif        | Arabidopsis thaliana    | 384      | +      | 6     | TCTTAC         | Part of a light responsive element                                      |
| Expansin A-4     |                         |          |        |       |                |                                                                          |
| G-Box            | Pisum sativum           | 507      | −      | 6     | CACGTT         | Cis-acting regulatory element involved in light responsiveness          |
| ABRE             | Arabidopsis thaliana    | 508      | +      | 5     | ACGTG          | Cis-acting element involved in the abscisic acid responsiveness          |
| ABRE             | Arabidopsis thaliana    | 508      | +      | 5     | ACGTG          | Cis-acting element involved in the abscisic acid responsiveness          |
| ATC-motif        | Zea mays                | 384      | −      | 9     | TGCTATCCG      | Part of a conserved DNA module involved in light responsiveness        |
| CAAT-box         | Pisum sativum           | 361      | −      | 5     | CAAAT          | Common cis-acting element in promoter and enhancer regions             |
| CAAT-box         | Arabidopsis thaliana    | 581      | −      | 8     | CCAATTT        | Common cis-acting element in promoter and enhancer regions             |
| CAAT-box         | Petunia hybrida         | 694      | −      | 7     | TGCCAAC        | Common cis-acting element in promoter and enhancer regions             |
| TATA-box         | Arabidopsis thaliana    | 527      | −      | 4     | TATA           | Core promoter element around -30 of transcription start                 |
| TGACG-motif      | Hordeum vulgare         | 532      | −      | 5     | TGACG          | Cis-acting regulatory element involved in the MeJA-responsiveness        |
| **BURP Domain RD22-like** |                  |          |        |       |                |                                                                          |
| ABRE             | Triticum aestivum       | 181      | −      | 9     | GACACGTGGC     | Cis-acting element involved in the abscisic acid responsiveness          |
| ARE              | Zea mays                | 542      | +      | 6     | AAACCA         | Cis-acting regulatory element essential for the anaerobic induction     |
| Box 4            | Petroselinum crispum    | 450      | −      | 6     | ATTAAT         | Part of a conserved DNA module involved in light responsiveness        |
| CAAT-box         | Arabidopsis thaliana    | 55       | +      | 5     | CCAAT          | Common cis-acting element in promoter and enhancer regions             |
| CCAAT-box        | Hordeum vulgare         | 440      | +      | 6     | CAACGG         | MYBHv1 binding site                                                     |
| CGTCA-motif      | Hordeum vulgare         | 515      | +      | 5     | CGTCA          | Cis-acting regulatory element involved in the MeJA-responsiveness        |
| TATA-box         | Arabidopsis thaliana    | 291      | +      | 4     | TATA           | Core promoter element around -30 of transcription start                 |
| TGACG-motif      | Hordeum vulgare         | 512      | +      | 5     | TGACG          | Cis-acting regulatory element involved in the MeJA-responsiveness        |
| TGACG-motif      | Hordeum vulgare         | 515      | −      | 5     | TGACG          | Cis-acting regulatory element involved in the MeJA-responsiveness        |
improvement programs is also becoming popular engendering novel properties [56–58]. Promoter regions In silico analysis of fibre related gene could be used to predict gene expression profiles in cotton plant. Many stresses resistant, light responsive which can contribute for fibre development were present in E6-like, Expansin A4 and BURP Domain RD22-like (Table 6). To explore the molecular mechanisms regulating cotton fibre development, promoters of several cotton fibre genes have been identified. E6 was the first of such genes to be reported, and the E6 promoter has been used for engineering cotton fibre quality [59]. GhRDL1, a gene highly expressed in cotton fibre cells at the elongation stage, encodes a BURP domain-containing protein [60], and the GaRDL1 promoter showed a trichrome-specific activity in transgenic Arabidopsis plants [61]. The aim of our analysis was to predict promoter and regulatory elements of genes encoding useful stress responsive leading to fibre production. In cotton, basic information related to different cis acting elements was generated to support the effort of improving cotton plant for a stress resistant with more fibre production.

Conclusion

The SL-19 appeared to be a promising source for cotton quality improvement with maximum expression for all fibre genes. To address the negative correlation between yield and fibre quality, use of genetic engineering is recommended to break this linkage by transferring E6-like, Expansin A4 and BURP Domain RD22-like genes in local cotton cultivars.

Acknowledgements Authors are the whole cotton group working at MNS University of Agriculture, Multan and Central Cotton Research Institute, Multan, for providing technical support and germplasm for this study.

Author contributions AAK, NI and ZK designed the research plan. FA carried out the experiments and drafted the manuscript. ZM and AG supported in experimentation and manuscript review and improvement. ZK and NI helped in data analysis. AAK, NI and ZK reviewed the final manuscript.

Funding Funding for this study was provided by Pakistan Science Foundation (PSF), Islamabad under PSF-NSFC-IV/Agr/P-MNSUAM (30). The authors are thankful to PSF for supporting this research work.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Sandin G, Peters GM (2018) Environmental impact of textile reuse and recycling—a review. J Clean Prod 184:353–365
2. Stelly DM, Lacape J-M, Dessauw D, Kohel RJ, Mergeai G et al (2008) International genetic, cyrogenic and germplasm resources for cotton genomics and genetic improvement. Ominpress, Madison
3. Campbell B, Saha S, Percy R, Frelichowski J, Jenkins JN et al (2010) Status of the global cotton germplasm resources. Embrapa Recursos Genéticos e Biotecnologia-Artigo em periódico indexado (ALICE)
4. Lubbers EL, Chee PW (2009) The worldwide gene pool of G. hirsutum and its improvement. In: Genetics and genomics of cotton. Springer, Berlin, pp 23–52
5. Percy RG (2009) The worldwide gene pool of Gossypium barbadense L. and its improvement. Genetics and genomics of cotton. Springer, Berlin, pp 53–68
6. Kim HJ, Triplett BA (2001) Cotton fiber growth in planta and in vitro. Models for plant cell elongation and cell wall biogenesis. Plant Physiol 127:1361–1366
7. Ruan Y-L (2005) Recent advances in understanding cotton fibre and seed development. Seed Sci Res 15:269–280
8. Pang C-Y, Wang H, Pang Y, Xu C, Jiao Y et al (2010) Comparative proteomics indicates that biosynthesis of pectic precursors is important for cotton fiber and Arabidopsis root hair elongation. Mol Cell Proteomics 9:2019–2033
9. Zhang F, Zuo K, Zhang J, Liu X, Zhang L et al (2010) An L1 box binding protein, GbML1, interacts with GbMYB25 to control cotton fibre development. J Exp Bot 61:3599–3613
10. Zhang B, Yang YW, Zhang Y, Liu JY (2013) A high-confidence reference dataset of differentially expressed proteins in elongating cotton fiber cells. Proteomics 13:1159–1163
11. Bauer E, Schmutzer T, Barilar I, Mascher M, Gundlach H et al (2017) Towards a whole-genome sequence for rye (Secale cereale L.). Plant J 89:853–869
12. Jiang H, Li H, Bu Q, Li C (2009) The RHA2a-interacting proteins ANAC019 and ANAC055 may play a dual role in regulating ABA response and jasmonate response. Plant Signal Behav 4:464–466
13. Haigler CH, Betancur L, Stiff MR, Tuttle JR (2012) Cotton fiber: a powerful single-cell model for cell wall and cellulose research. Front Plant Sci 3:104
14. Lv L-M, Zuo D-Y, Wang X-F, Cheng H-L, Zhang Y-P et al (2020) Genome-wide identification of the expansin gene family reveals that expansin genes are involved in fibre cell growth in cotton. BMC Plant Biol 20:1–13
15. Orford SJ, Timmins JN (2000) Expression of a lipid transfer protein gene family during cotton fibre development. Biochim Biophys Acta (BBA) Mol Cell Biol Lipids 1483:275–284
16. Ruan Y-L, Llewellyn DJ, Furbank RT (2001) The control of single-celled cotton fiber elongation by developmentally reversible gating of plasmodesmata and coordinated expression of sucrose and K+ transporters and expansin. Plant Cell 13:47–60
17. Xu B, Gou J-Y, Li F-G, Shangguan X-X, Zhao B et al (2013) A cotton BURP domain protein interacts with α-expansin and their co-expression promotes plant growth and fruit production. Mol Plant 6:945–958
18. Xu H, Li Y, Yan Y, Wang K, Gao Y et al (2010) Genome-scale identification of soybean BURP domain-containing genes and their expression under stress treatments. BMC Plant Biol 10:1–16
19. Park J, Cui Y, Kang B-H (2015) AtPGL3 is an Arabidopsis BURP domain protein that is localized to the cell wall and promotes cell enlargement. Front Plant Sci 6:412
Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.