Molecular evidence for segmental duplication across chromosomes of soybean using transcription factor gene family

Manoj Kumar Srivastava and Gyanesh Kumar Satpute
ICAR- Indian Institute of Soybean Research
Khandwa Road, Indore (Madhya Pradesh) 452001, India.

Abstract:

Duplication of genome is an important genetic innovation. Large genome size (1.1 Gb) along with ancient and recent duplication events make the soybean genome more complex. Analyzing the distribution and duplication event in soybean transcription family genes, the segmental duplication within chromosomes was revealed. Our study provides a strong evidence that the large segmental duplication event in genome architecture and evolution of soybean genome using simple method of sequence and order analysis of TF genes. Finally, a scheme for interrelationship of different chromosomes has been proposed.

Keywords: *Glycine max*, Soybean, Chromosome, segmental duplication, transcription factor

Introduction:

Duplication of genome is an important event leading to evolution of organism. Gene and genome duplication may contribute to evolution and domestication of crops. The role of genome duplication in present architecture/ topology of soybean genome have vital influence on agronomic traits, yield potential and adaptation of crop plants. The redundant copy of gene arising by the duplication accumulates the beneficial mutations resulting in new function of duplicated gene product. Thus gene duplication provides a source of plasticity to genome for adapting to changing environment. Large genome size (1.1 Gb) along with ancient and recent duplication events make the soybean genome more complex. Having complex genome structure makes it rather difficult to design and develop effective breeding strategies in soybean for desired traits. Several QTLs for various traits and linkage maps have been developed for 20 chromosomes of soybean genome. Translocation, inversion, deletion and duplication play important roles in creating small duplication events.

In flowering plants, two ancestral whole genome duplication (WGD) is reported. In soybean, two additional sequential WGDs are established: one had occurred 59 MYA in the common ancestor of legumes and other about 8-13 MYA in *Glycine* lineage

(Schmutz et al., 2010). Due to multiple genome duplication events, the number of predicted coding genes in soybean is much higher than in *Arabidopsis* and grapes (Sterck et al., 2007; Cannon and Shoemaker, 2012). Several small blocks of homeologus retention and chromosomal arrangements are shown to exist in 20 chromosomes of soybean (Schmutz et al., 2010; Lestari et al., 2013). The segmental duplication in soybean has been reported to result in the evolution of several phenotypic traits such as disease resistance (Shin et al., 2008; Kim et al., 2009). Several QTLs associated with seed related traits, disease resistance and high content of carbohydrates, proteins and oil were reported to be conserved in the duplicated segment of the soybean genome. The major seed protein QTL is mapped on chromosome 20 (Brummer et al., 1997). This QTL have been studied using several different approaches (Wang et al., 2008; Qi et al., 2011).

Transcription factors are DNA binding regulatory proteins, which interact with other proteins and regulate the process of transcription. Most of the TF genes have several family members. In soybean, there are 57 families of TF genes reported with are distributed at 3747 locus in 20 chromosomes. In the present study the duplication and order of these TF genes were analyzed.

Materials and methods:

All the transcription factor genes of soybean were downloaded from Plant transcription factor database (http://planttfdb.gao-lab.org/index.php?sp=Gma). Duplication of TF genes across all chromosomes was studied by BLAST analysis of each gene against total transcript database (Wm82.a2.v1 Transcript Sequences). The top BLAST hit having E value of 0.0 was considered as transcript arising of duplicated gene. The E value having more than 0.0 was considered as non-duplicated gene product. The chromosome wise duplicated gene pair was recorded. The putative chromosome pair for duplicated segment was observed for similarity in order of TF genes.

Results and discussion:

Most of the high copy number TF genes are distributed among all the chromosomes, however, the relative number of TF varied for each chromosome (Table 1). Five out of 20 chromosomes (Chr 2, Chr 6, Chr 8, Chr 10 and Chr 13) contain more than 200 loci of TF genes. The chromosome 16 has the lowest number of TF genes while chr 13 has the maximum number of TF genes. No specific pattern of chromosomal location for different TF gene family was observed i.e. different TF gene family have different distribution pattern.

Locus frequency of less than 40% represents the TF genes which produce more copies of products from the same gene (Table 2 and S1). These include BBR-BCP (31.03%; 10/29), E2F/DP (36.84%; 14/38), VOZ (23.08%; 6/28), Whirly (38.89%; 7/18) and YABBY (34.04%; 17/47). Five out of 57 families have same number of gene products and gene locus (HRT- like, LFY, RAV, S1Fa-like and SAP). These are also low copy

Table 1: Distribution of TF gene products in soybean chromosomes

S. No.	Gene	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	U	Total
1.	AP2	8	6	5	2	4	2	4	7	4	3	7	6	5	1	4	2	8	4	6	0	11	**99**
2.	ARF	4	8	4	3	10	2	10	6	1	2	3	14	12	5	7	4	2	4	2	1	3	**107**
3.	ARR B	2	2	0	1	7	1	5	4	4	1	7	0	2	2	6	1	6	1	3	0	0	**55**
4.	B3	4	8	9	6	0	5	11	5	6	4	22	8	2	14	1	7	6	12	8	10	0	**148**
5.	BBR BPC	0	0	0	2	7	8	3	7	2	0	0	0	0	0	0	0	0	0	0	0	0	**29**
6.	BES1	2	0	0	1	0	1	1	0	2	0	1	3	4	2	0	0	1	1	0	0	0	**19**
7.	BHLH	33	**46**	42	20	21	29	36	40	16	40	14	23	45	15	33	14	26	13	17	21	4	**548**
8.	BZIP	17	18	23	21	23	15	6	**37**	5	14	33	20	23	10	9	12	7	18	20	16	5	**352**
9.	C2H2	12	29	17	11	13	11	18	14	3	**31**	14	20	30	12	12	8	13	11	16	24	2	**321**
10.	C3H	2	11	11	10	8	7	4	11	7	13	6	8	7	7	10	5	9	3	6	5	0	**150**
11.	CATMA	0	0	0	0	7	0	2	6	2	0	1	0	0	0	2	0	2	2	0	0	0	**24**
12.	CO LIKE	0	4	1	1	1	1	3	2	3	2	0	0	6	5	0	1	1	2	4	3	0	**40**
13.	CPP	3	0	0	1	1	3	6	1	3	1	1	0	0	0	0	0	4	1	0	1	0	**26**
14.	DBB	6	0	1	3	0	3	0	0	1	0	8	16	4	1	1	2	1	0	0	0	0	**47**
15.	DoF	3	4	4	8	4	7	10	7	2	2	2	2	11	0	10	2	5	4	6	2	2	**97**
16.	E2F/DP	1	1	0	6	2	4	0	0	0	2	11	7	0	0	0	0	4	0	0	0	0	**38**
17.	eil	0	1	0	0	1	1	0	1	0	0	1	0	3	1	1	0	0	1	0	2	0	**13**
18.	ERF	15	16	22	17	15	19	17	25	10	24	11	11	**27**	16	14	17	19	11	16	16	0	**338**
19.	FAR1	3	2	6	15	1	12	5	14	7	7	5	4	6	2	26	1	1	8	5	8	0	**138**
20.	G2 LIKE	9	15	21	3	7	3	12	4	19	13	6	10	14	5	**23**	2	9	17	21	9	0	**222**
21.	GATA	4	10	2	6	3	4	5	15	2	6	7	5	2	4	2	6	6	0	2	1	0	**92**
22.	gebp	0	0	1	0	2	0	0	0	0	2	0	0	1	0	1	0	0	0	1	2	0	**10**
23.	GRAS	5	8	6	5	7	6	7	4	6	5	19	17	13	3	10	5	7	12	2	3	1	**151**
24.	GRF	2	0	3	7	0	2	1	0	4	1	4	2	3	0	2	2	4	0	2	0	3	**42**
25.	HB OTHERS	0	2	1	3	0	5	4	8	0	0	0	1	2	1	0	0	0	2	1	0	0	**30**
26.	HBPHD	0	0	0	0	0	0	0	0	1	2	0	0	4	0	3	0	5	0	0	1	0	**16**
27.	HDZIP	15	6	10	3	8	5	12	15	18	12	10	14	10	1	7	6	4	10	6	7	1	**180**
28.	HRT LIKE	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**
29.	HSF	8	1	4	7	8	1	2	7	4	9	5	0	4	5	2	2	5	0	4	3	0	**81**

30.	LBD	8	8	4	8	9	6	3	8	2	6	6	2	8	13	4	5	4	5	10	4	2	**125**
31.	LFY	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**2**
32.	LSD	3	0	0	0	2	0	3	1	2	0	0	0	0	0	4	0	3	0	0	0	0	**18**
33.	M TYPE MADS	1	7	3	2	5	0	6	7	0	16	8	0	2	2	0	1	3	10	2	8	4	**87**
34.	MICK MADS	12	19	5	12	23	14	7	**30**	13	4	5	2	22	4	9	5	5	10	3	4	1	**209**
35.	MYB	22	24	18	18	14	23	29	24	19	25	19	19	26	18	14	10	17	**40**	24	24	3	**430**
36.	MYB RELATED	14	10	15	19	15	17	16	10	18	32	25	11	**34**	9	12	25	12	19	12	16	1	**342**
37.	NAC	13	11	3	18	20	**21**	15	17	8	10	8	18	17	12	12	13	9	11	20	13	0	**269**
38.	NFX1	0	0	0	0	0	0	2	1	2	0	1	0	0	0	0	0	0	4	0	0	0	**10**
39.	nfya	0	18	3	0	2	0	12	5	16	5	0	5	3	3	23	5	4	4	9	0	0	**117**
40.	nfyb	0	3	5	2	5	5	3	5	3	8	1	1	1	0	2	0	4	1	1	7	1	**58**
41.	nfyc	0	3	2	2	0	3	0	3	0	2	2	2	8	3	3	0	0	1	3	1	0	**38**
42.	NIN LIKE	1	1	0	15	1	10	0	0	2	1	4	4	9	1	6	2	1	0	0	3	0	**61**
43.	RAV	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**5**
44.	s1fa like	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	**4**
45.	sap	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	**2**
46.	SBP	12	6	6	3	9	6	10	4	3	1	21	1	5	0	4	4	3	5	3	2	3	**111**
47.	SRS	2	7	0	6	0	4	3	0	0	0	4	1	2	2	2	4	3	0	0	2	0	**42**
48.	STAT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	**2**
49.	TALE	9	5	6	15	4	17	1	2	3	5	6	17	6	7	4	2	13	2	3	2	4	**133**
50.	TCP	1	2	1	6	12	8	2	5	2	4	1	6	9	0	1	2	7	4	3	5	0	**81**
51.	Trihelix	2	3	11	5	1	6	4	3	5	13	4	1	8	0	5	6	8	4	6	9	0	**104**
52.	VOZ	0	0	0	0	0	3	3	0	0	12	3	4	1	0	0	0	0	0	0	0	0	**26**
53.	Whirly	2	6	1	0	0	0	0	4	0	0	0	0	0	0	0	0	0	3	2	0	0	**18**
54.	WOX	1	3	1	2	2	2	4	1	1	3	6	1	2	2	1	0	1	4	1	3	0	**41**
55.	WRKY	16	24	19	16	17	20	11	21	**29**	18	4	4	10	16	12	8	19	19	10	3	0	**296**
56.	YABBY	4	1	1	1	2	8	0	1	0	0	0	16	8	0	0	0	4	1	0	0	0	**47**
57.	ZFHD	3	8	0	2	2	2	3	6	5	0	2	3	2	1	0	4	2	5	1	7	0	**58**
		289	**368**	**298**	**315**	**305**	**333**	**322**	**398**	**265**	**362**	**328**	**309**	**424**	**206**	**306**	**195**	**277**	**289**	**262**	**248**	**51**	**6150**

Table 2: Distribution of duplicated genes (%) in soybean chromosome

Chromosome Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0.0	23.1	11.1	0.0	0.0	0.6	2.7	2.4	11.4	0.0	35.1	0.0	1.0	1.0	0.0	0.0	0.0	0.9	0.0	0.0
2	30.9	0.0	0.0	0.0	0.0	0.0	3.4	0.0	2.4	19.6	0.0	0.0	0.0	48.0	0.0	23.3	1.5	2.6	0.0	1.6
3	8.1	0.0	0.0	0.0	0.8	0.0	8.9	0.0	0.0	0.0	0.0	0.7	0.0	1.0	0.7	3.5	0.0	0.9	65.0	0.0
4	0.0	0.0	0.0	0.0	0.0	76.6	0.0	0.0	0.8	0.0	1.5	0.0	0.0	1.0	0.0	0.0	0.0	0.9	0.0	0.0
5	0.0	0.0	0.9	0.0	0.0	0.0	0.0	37.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	46.6	0.0	5.1	0.0
6	0.0	0.0	0.0	93.7	0.0	0.0	0.0	0.0	0.8	0.0	0.0	18.5	0.0	1.0	1.5	0.0	0.0	0.0	0.0	0.8
7	4.1	3.8	10.3	0.0	0.0	0.0	0.0	15.4	9.8	0.0	0.0	0.7	3.9	0.0	0.0	31.4	10.5	11.1	0.0	18.0
8	2.4	0.0	0.9	0.0	48.4	0.0	17.1	0.0	0.0	0.7	0.7	0.7	0.5	0.0	17.8	2.3	0.0	33.3	0.0	0.0
9	11.4	1.9	0.0	0.8	0.0	0.0	9.6	0.0	0.0	0.0	0.0	2.7	1.0	1.0	27.4	18.6	0.8	22.2	0.7	0.8
10	0.0	18.8	0.9	0.0	0.0	0.6	0.0	0.0	0.0	0.0	2.2	0.0	10.2	0.0	0.0	0.0	0.8	0.0	0.7	73.0
11	37.4	0.0	0.0	1.6	0.8	0.0	0.0	0.6	0.0	2.7	0.0	37.0	0.0	0.0	0.0	0.0	0.0	27.4	0.0	0.0
12	0.0	0.6	0.0	0.0	0.0	17.7	0.0	1.8	3.3	0.0	37.3	0.0	22.9	0.0	0.0	2.3	0.0	0.0	1.5	0.0
13	0.8	0.0	0.0	0.8	0.0	0.0	6.2	0.6	0.8	14.9	0.0	33.6	0.0	8.8	51.9	0.0	14.3	0.9	13.1	4.1
14	0.8	31.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.9	0.0	0.0	0.0	23.3	0.0	0.7	0.0
15	0.0	0.0	0.0	0.0	0.0	1.3	0.7	14.8	30.1	0.0	0.0	0.0	35.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
16	0.0	11.3	2.6	0.0	0.0	0.0	17.8	1.2	16.3	0.0	0.0	2.1	0.0	0.0	0.0	0.0	0.0	0.0	10.9	0.0
17	0.0	1.3	0.0	1.6	46.1	0.6	10.3	0.6	0.8	0.7	0.0	0.0	10.2	31.4	0.0	0.0	0.0	0.0	1.5	0.8
18	0.8	1.9	0.9	0.0	0.0	0.0	8.9	25.4	22.0	0.0	21.6	0.0	0.5	0.0	0.7	0.0	0.8	0.0	0.0	0.0
19	0.0	0.6	72.6	0.0	3.9	0.0	0.0	0.0	0.8	0.0	0.0	0.0	8.3	2.0	0.0	18.6	0.0	0.0	0.0	0.8
20	0.0	4.4	0.0	0.0	0.0	1.9	14.4	0.0	0.8	60.8	0.0	0.0	2.4	0.0	0.0	0.0	0.0	0.0	0.7	0.0
Unassigned	3.3	0.6	0.0	1.6	0.0	0.6	0.0	0.0	0.0	0.7	1.5	4.1	0.0	4.9	0.0	0.0	1.5	0.0	0.0	0.0
	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

number transcription factor genes. The genes having locus to gene product ratio of more than 0.8 produce less gene product variants. These are BES1 (0.84), Dof (0.82), EIL (0.92), ERF (0.88), GeBP (0.9), M-type_MADS (0.95), WOX (0.80) and ZF-HD (0.89).

This analysis for cDNA of all the 3747 locus genes revealed that at least 72.63% of the TF genes were duplicated paralog pair (Table 3 and S2). Also, there was a specific preference of duplication among various chromosomes, e.g. chromosome 1 have more duplication segments from chr 2, chr 9 and chr 11; chromosome 2 have more duplication fragments from chr 1, chr 10, chr 14 and chr 16; Chromosome 3 have major segment from chr 19 and minor contribution from chr 1 and chr 7; Chromosome 4 has almost entire (95%) TF gene duplication from chr 6. The detail distribution is given in Fig 1, Table 3 and supplementary Table S2.

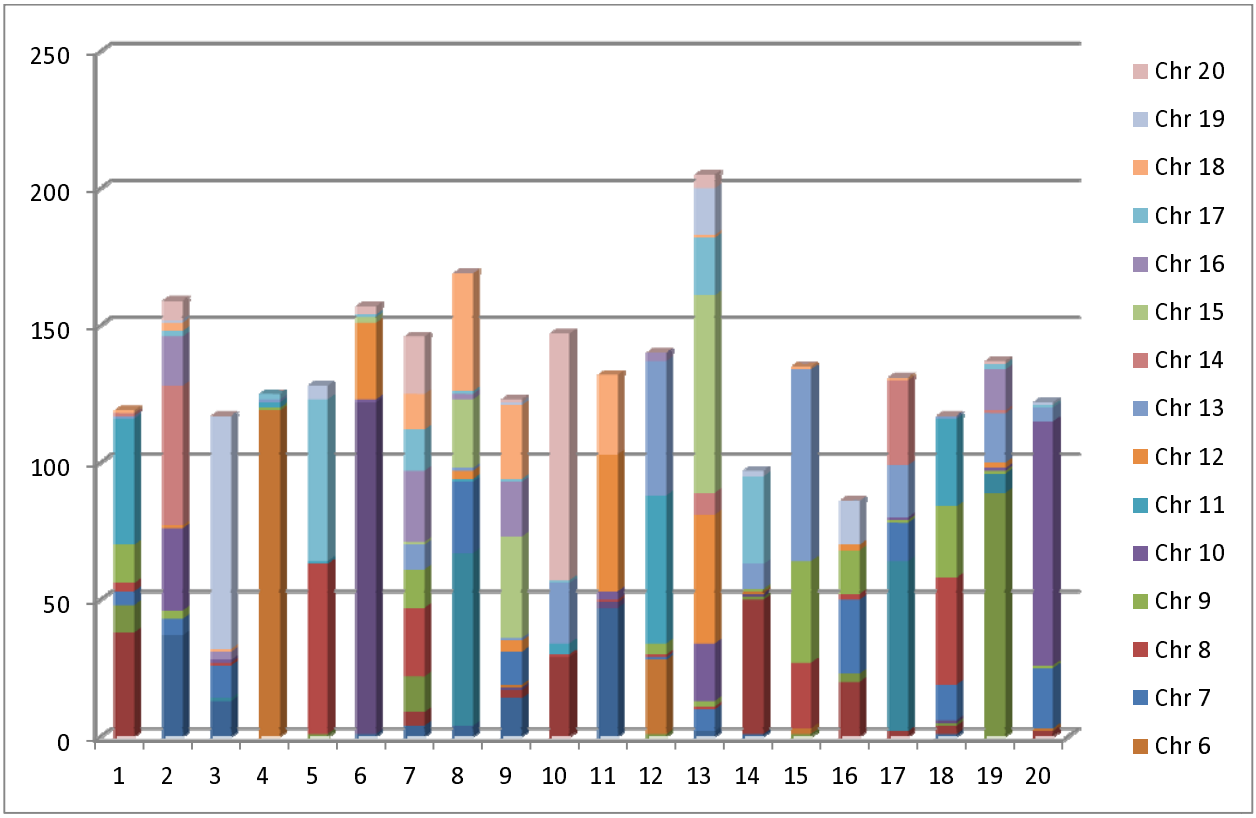

Fig 1: Distribution of duplicated transcription factor genes in different chromosomes of soybean

The gene order of different TFs was then compared and similarity was investigated among putative chromosome pairs (Table S3). The data supported the finding of preferential chromosome duplication. Furthermore, in the segment of similar TF genes the order of various TF genes was found to be collinear either in direct or reverse direction. This indicated that long segmental duplication event in soybean is major event in the evolution of soybean chromosome. The preferential segmental duplication is

presented in Fig S1. This is in accordance with the distribution data of duplicated TF genes. However this may be noted that careful sequence analysis is required at gene to gene basis for the accumulated mutations in duplicated TF genes. This will lead to assigning new function to the duplicated TF gene.

It has been established that the recent genome duplication occurred on many soybean chromosomes (Cannon and Shoemaker, 2012). QTLs across duplicated regions of chr 4/ chr 6, chr 3/ chr 19 and chr 10/chr 20 were shown to be correlated (Lestani et al., 2013). Small rearrangements were found in duplicated homeologus regions of QTLs due to recent duplication event. Soybean gene duplication may also lead to gene regulation (Shoemaker et al., 1996). A large inversion with synteny in the corresponding regions of chr 10 and chr 20 has also been reported (Cannon et al., 2004).

Yang et al (2013) studied similarities between soybean and Arabidopsis genomes using dot- plot analysis and reported that whole genome duplication event occurred more than once during the evolution of soybean genome. About 70% of total genes in soybean genome have duplicated paralog pairs. The block of 2140 genes were found to be the largest pair of duplicated paralog gene in chromosome 3 and 19 of soybean (Yang et al., 2013). The data presented also indicated towards the ongoing duplication event in soybean genome. DNA rearrangement and codon mutations resulted in emergence of new gene sequences which may have new functions leading to better adaptation in new challenging environment.

Table 3: Distribution of duplicated genes in soybean chromosome

Chromosome Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Unassigned	Not duplicated	TOTAL
1	0	37	13	0	0	1	4	4	14	0	47	0	2	1	0	0	0	1	0	0	4	44	172
2	38	0	0	0	0	0	5	0	3	29	0	0	0	49	0	20	2	3	0	2	1	69	221
3	10	0	0	0	1	0	13	0	0	0	0	1	0	1	1	3	0	1	89	0	0	57	177
4	0	0	0	0	0	121	0	0	1	0	2	0	0	1	0	0	0	1	0	0	2	60	188
5	0	0	1	0	0	0	0	63	0	0	0	0	0	0	0	0	62	0	7	0	0	43	176
6	0	0	0	119	0	0	0	0	1	0	0	27	0	1	2	0	0	0	0	1	1	58	210
7	5	6	12	0	0	0	0	26	12	0	0	1	8	0	0	27	14	13	0	22	0	52	198
8	3	0	1	0	62	0	25	0	0	1	1	1	1	0	24	2	0	39	0	0	0	70	230
9	14	3	0	1	0	0	14	0	0	0	0	4	2	1	37	16	1	26	1	1	0	43	164
10	0	30	1	0	0	1	0	0	0	0	3	0	21	0	0	0	1	0	1	89	0	74	221
11	46	0	0	2	1	0	0	1	0	4	0	54	0	0	0	0	0	32	0	0	2	42	184
12	0	1	0	0	0	28	0	3	4	0	50	0	47	0	0	2	0	0	2	0	6	38	181
13	1	0	0	1	0	0	9	1	1	22	0	49	0	9	70	0	19	1	18	5	0	68	274
14	1	51	0	0	0	0	0	0	0	0	0	0	8	0	0	0	31	0	1	0	5	39	136
15	0	0	0	0	0	2	1	25	37	0	0	0	72	0	0	0	0	0	0	0	0	29	166
16	0	18	3	0	0	0	26	2	20	0	0	3	0	0	0	0	0	0	15	0	0	31	118
17	0	2	0	2	59	1	15	1	1	1	0	0	21	32	0	0	0	0	2	1	2	41	181
18	1	3	1	0	0	0	13	43	27	0	29	0	1	0	1	0	1	0	0	0	0	50	170
19	0	1	85	0	5	0	0	0	1	0	0	0	17	2	0	16	0	0	0	1	0	47	175
20	0	7	0	0	0	3	21	0	1	90	0	0	5	0	0	0	0	0	1	0	0	46	174
Unassigned	4	1	0	2	0	1	0	0	0	1	2	6	0	5	0	0	2	0	0	0	0	7	31
																							3747
	123	160	117	127	128	158	146	169	123	148	134	146	205	102	135	86	133	117	137	122	23	1008	

Table 4: Distribution and orientation of genes in duplicated segment pairs of soybean chromosomes

Chromosome Number	Segment number	Number of TF genes in the segment	Total number of genes in the segment	Chromosome number of Duplicated segment	Number of TF genes in the segment	Total number of genes in the segment	Orientation
Chr 1	1	11	130	Chr 9	11	117	REVERSE
	2	16	214	Chr 2	16	227	REVERSE
	3	22	239	Chr 2	22	203	DIRECT
	4	4	29	Chr 3	4	40	DIRECT
	5	8	110	Chr 3	8	123	REVERSE
	6	59	630	Chr 11	54	619	REVERSE
Chr 2	7	24	291	Chr 16	24	345	DIRECT
	8	9	114	Chr 10	9	75	REVERSE
	9	17	224	Chr 14	20	253	DIRECT
	10	5	33	Chr 14	5	20	DIRECT
	11	12	143	Chr 14	13	168	REVERSE
	12	10	170	Chr 14	10	186	REVERSE
Chr 3	13	19	166	Chr 7	14	88	REVERSE
	14	105	1356	Chr 19	105	1326	DIRECT
Chr 4	15	37	407	Chr 6	35	418	DIRECT
	16	41	586	Chr 6	36	574	DIRECT
	17	25	255	Chr 6	25	247	REVERSE
	18	28	350	Chr 6	27	354	REVERSE
	19	18	293	Chr 6	19	319	REVERSE
Chr 5	20	7	50	Chr 17	7	53	REVERSE
	21	15	149	Chr 17	15	151	REVERSE
	22	20	260	Chr 17	22	261	DIRECT
	23	14	99	Chr 17	13	81	REVERSE
	24	6	107	Chr 8	6	104	DIRECT
	25	35	476	Chr 8	33	502	DIRECT
	26	37	485	Chr 8	38	497	DIRECT

Chr 6	27	25	257	Chr 12	25	269	REVERSE
Chr 7	28	11	103	Chr 8	11	93	REVERSE
	29	28	243	Chr 16	27	236	DIRECT
	30	19	223	Chr 9	17	215	REVERSE
	31	22	262	Chr 20	26	323	DIRECT
	32	17	361	Chr 17	19	354	REVERSE
Chr 8	33	7	58	Chr 15	6	44	DIRECT
	34	11	168	Chr 18	12	185	DIRECT
Chr 9	35	46	942	Chr 15	49	972	DIRECT
	36	20	423	Chr 16	20	423	DIRECT
	37	38	465	Chr 18	40	541	REVERSE
Chr 10	38	35	342	Chr 13	34	314	DIRECT
	39	107	1431	Chr 20	110	1361	REVERSE
Chr 11	40	43	486	Chr 12	37	460	DIRECT
	41	46	676	Chr 18	40	503	REVERSE
Chr 12	42	40	441	Chr 13	40	460	REVERSE
Chr 13	43	19	186	Chr 19	22	265	REVERSE
	44	17	239	Chr 17	18	244	REVERSE
	45	40	488	Chr 15	39	468	REVERSE
	46	38	417	Chr 15	42	420	REVERSE
Chr 14	47	37	425	Chr 17	36	392	REVERSE
Chr 16	48	15	146	Chr 19	15	192	REVERSE

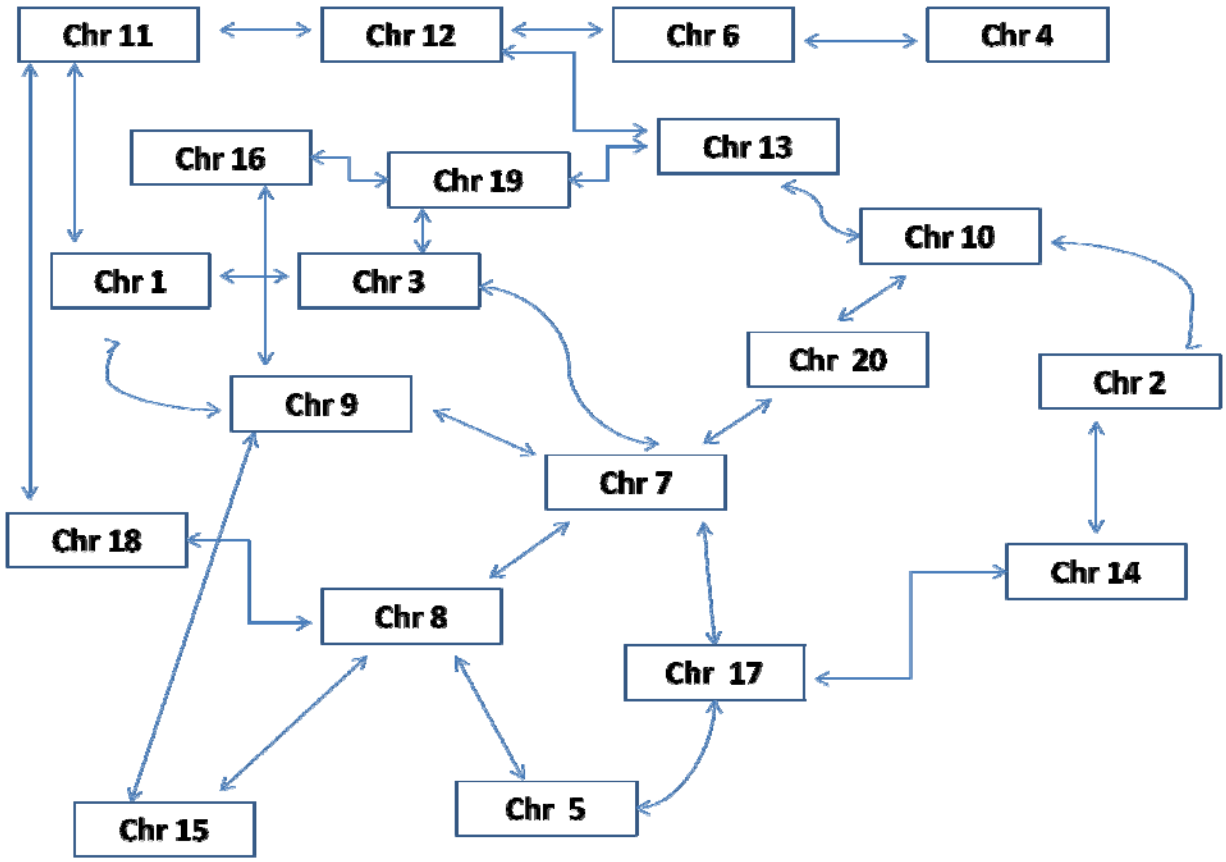

Figure 2: Inter relationship of different soybean chromosomes

Chromosome 1- contains 6 duplicated TF gene segments from chr 9, chr 2, chr 3 and chr 11. The largest segment was from chr 11 having 519 genes.

Chromosome 2- contains 8 distinct segments from chr 1, chr10, chr 14 and chr 16. There were four different fragments duplicated from chr 14, two in forward direction and two in reverse direction.

Chromosome 3- Block of 1355 genes is directly duplicated in chromosome 19 (1325 genes). All the TF genes are collinear. 160, 29 and 88 genes are duplicated in smaller segments and the order of TFs are collinear but in reverse and forward direction.

Chromosome 4- has 5 different segments from chr 6. The first two segments are in forward direction while last three are in reverse direction. Chromosome 6 has an additional segment from chromosome 12 in reverse direction.

Chromosome 5- has seven smaller segments, four from chromosome 17 and three from chromosome 8. Three segments of chr 17 were duplicated in reverse direction. All the three segments duplicated from chromosome 8 were in forward direction. Most of the TF gene order was conserved in duplicated segments.

The duplicated segments have almost similar number of transcription factor genes (Table 4). However, the number of embedded genes is quite variable in respective duplicated segments (Fig S1). There are 48 segmental pairs detected having 4-107 conserved TF genes and 29-1431 embedded genes. Only segment number 36 has identical 20 TF genes and 423 embedded genes on chr 9 and chr 16. Other segment pairs have variable number of genes. This may be due to ongoing translocation, mutations, and deletions of gene within the segments. Based on these segment pairs an inter-relation of various chromosomes of soybean has been proposed and shown as Fig 2.

Duplication creates genetic redundancy leading to evolutionary innovation. Over the passage of time the duplicated copy acquire a beneficial mutation resulting in retention of both copies. Alternatively the mutation in duplicated segment may make it non functional. The recognition of fact that a single protein can have a multiple catalytic or structural functions supports the contribution of gene duplication. In recent studies, the genome- wide analysis of duplication of individual transcription factors have been reported (Liu et al. 2020; Chen et al., 2019; Li et al., 2019; Ullah et al., 2019). Our results also validate these studies; however, there are some minor differences about the position of duplicated segments. Our study has provided a strong evidence that the large segmental duplication event in genome architecture and evolution of soybean genome using simple method of sequence and order analysis of TF genes. A detailed analysis of these genes using Bioinformatics tools may help in establishing the process of gene duplication in other species and genera.

Conclusions

By analyzing the distribution and order of transcription factor genes the early mode of genome duplication was established. This method provides an easy and effective tool to study genome duplication in different species and genera. The functional analysis of duplicated genes is required for complete elucidation of the process of genome duplication.

Data Availability

The data underlying this article are available in the article and in its online supplementary material. The raw data of transcription factor sequences can be found at http://planttfdb.gao-lab.org/index.php?sp=Gma.

Acknowledgments

We thank Dr Sanjay Gupta, Dr Mehar H Asif and Dr. A. Radhakrishna for discussion during the preparation of the manuscript. We are thankful to Director, ICAR-Indian Institute of Soybean Research, Indore for encouragement and facility provided by for this work.

References

Alseekh, S. et al. Mobile transposable elements shape plant genome diversity. *Trends in Plant Science* **25 (11),** 1062-1064 (2020).

Brummer, E.C. et al. Mapping QTL for seed protein and oil content in eight soybean populations. *Crop Sci.* **37**, 370–378 (1997).

Cannon, S.B & Shoemaker, R.C. Evolutionary and comparative analyses of the soybean genome. *Breeding Science* **61**, 437–444 (2012).

Cannon, S.B. et al. The roles of segmental and tandem gene duplication in the evolution of large families in *Arabidopsis thaliana*. *BMC Plant Pathol.* **4**, 10 (2004).

Chen, F. et al. Genome- wide identification of GRF transcription factors in soybean and expression analysis of GmGRF family under shade stress. *BMC Plant Biology* **19**, 269 (2019).

Kim, K.D. et al. Dynamic rearrangements determine genome organization and useful traits in soybean. *Plant Physiol.* **151**, 1066–1076 (2009).

Lestari, P. et al. Gene divergence of homeologous regions associated with a major seed protein content QTL in soybean. *Front. Plant Sci.* **4**, 176 (2013).

Li, Q. et al. *In silico* genome- wide identification and characterization of BES1 gene family in soybean. *Heliyon* **5**, e01868 (2019).

Liu, W. et al. Genome- wide characterization and expression analysis of soybean trihelix gene family. *Peer J* **8**, e8753 (2020).

Qi, Z.M. et al. Soybean oil content QTL mapping and integrating with meta-analysis method for mining genes. *Euphytica* **179**, 499–514 (2011).

Qiao, X. et al. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biology* **20**, 38 (2019).

Schmutz, J. et al. Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).

Shin, J.H. et al. The lipoxygenase gene family: a genomic fossil of shared polyploidy between *Glycine max* and *Medicago truncatula*. *BMC Plant Biol.* **8**, 133 (2008).

Shoemaker, R.C. et al. Genome duplication in soybean (*Glycine* subgenus *soja*). *Genetics* **144**, 329-338 (1996).

Shu, Y. et al. Genome- wide survey and expression analysis of the MADS- box gene family in soybean. *Mol. Biol. Rep.* **40**, 3901-3911 (2013).

Sterck, L. et al. How many genes are there in plants (and way are they there)? *Curr. Opin. Plant Biol.* **10**, 199–203 (2007).

Ullah, I. et al. Genome- wide identification and evolutionary analysis of TGA transcription factors in soybean. *Scientific Reports* **9**, 11186 (2019).

Yang, Y. et al. Comparative inference of duplicated genes produced by polyploidization in soybean genome. *International Journal of Genomics,* Article ID 275616 (2013).

Yin, G. et al. The large soybean (*Glycine max*) WRKY TF family expanded by segmental duplication events and subsequent divergent selection among subgroups. *BMC Plant Biology* **13**, 148 (2013).