Identification, Nomenclature, and Evolutionary Relationships of Mitogen-Activated Protein Kinase (MAPK) Genes in Soybean

Achal Neupane, Madhav P. Nepal, Sarbottam Piya, Senthil Subramanian, Jai S. Rohila, R. Neil Reese and Benjamin V. Benson

1Department of Biology and Microbiology, South Dakota State University, Brookings SD, USA.
Corresponding author email: madhav.nepal@sdstate.edu

Abstract: Mitogen-activated protein kinase (MAPK) genes in eukaryotes regulate various developmental and physiological processes including those associated with biotic and abiotic stresses. Although MAPKs in some plant species including Arabidopsis have been identified, they are yet to be identified in soybean. Major objectives of this study were to identify GmMAPKs, assess their evolutionary relationships, and analyze their functional divergence. We identified a total of 38 MAPKs, eleven MAPKKs, and 150 MAPKKKs in soybean. Within the GmMAPK family, we also identified a new clade of six genes: four genes with TEY and two genes with TQY motifs requiring further investigation into possible legume-specific functions. The results indicated the expansion of the GmMAPK families attributable to the ancestral polyploidy events followed by chromosomal rearrangements. The GmMAPK and GmMAPKKK families were substantially larger than those in other plant species. The duplicated GmMAPK members presented complex evolutionary relationships and functional divergence when compared to their counterparts in Arabidopsis. We also highlighted existing nomenclatural issues, stressing the need for nomenclatural consistency. GmMAPK identification is vital to soybean crop improvement, and novel insights into the evolutionary relationships will enhance our understanding about plant genome evolution.

Keywords: soybean genomics, MAPK family, gene evolution, homology, nomenclature, signal transduction

Evolutionary Bioinformatics 2013:9 363–386
doi: 10.4137/EBO.S12526

This article is available from http://www.la-press.com.

© the author(s), publisher and licensee Libertas Academica Ltd.

This is an open access article published under the Creative Commons CC-BY-NC 3.0 license.
Background

Plants are perpetually exposed to both biotic and abiotic stresses, and they have evolved sophisticated cellular and physiological mechanisms to cope with some severe forms of stresses (i.e., drought, heat, cold, wounding, disease, and so on).1–3 Perception of these stresses and the initiation of appropriate molecular responses occur through an intricate network of proteins involved in signal transductions.4,5 Mitogen-activated protein kinase (MAPK) genes play crucial roles in stress signaling pathways and regulate myriad of cellular and physiological processes.6–9 MAPK was originally discovered in animal cells as a microtubule-associated-protein-kinase in 1986 by Sturgill and Ray.10 It was later found to be a group of proteins phosphorylated at tyrosine residues in response to mitogens; hence, the name “mitogen-activated protein kinase”.11,12 Studies of genome sequences from diverse species at various taxonomic levels have shown that these genes always occur as gene families.13 The MAPK gene members belong to three functionally linked families called MAPK (=MPK), MAPKK (=MEK or MKK), and MAPKKK (=MEKK), forming a cascade of signaling networks associated with various cellular functions, including plant response to biotic and abiotic stresses.2,8,9 The number of MAPK genes within each family varies widely across species. For example, the number of MAPKs and MAPKKs are five and one in Chlamydomonas, six and six in Sacharomyces, 21 and 11 in Populus,13 16 and 12 in Brachypodium,14 20 and 10 in Arabidopsis,2,8 and 15 and eight in Oryza,13 respectively. The diversity of MAPKKs is less understood, and the estimated numbers in Chlamydomonas, Arabidopsis, Oryza, and Sorghum are 8–10,2,8,15,16 80,2,75,15 and 40–60,2 respectively. The MAPK genes in complex organisms have diversified into various clades over time.8,13 In Arabidopsis, MAPK genes are classified into four different subgroups (A, B, C, and D) based on their evolutionary relationships and the presence of TGY and TGY phosphorylation motifs.2,8 Similarly, MAPK genes in Arabidopsis form four distinct clades,8 and the three subgroups of MAPKK genes—MEKK-like, ZIK-like, and Raf-like members—generally occur as monophyletic groups.2 Although the biological functions of all the MAPKs are not fully understood, the MAPKs of same subgroups are likely to involve in similar physiological responses or are functionally redundant, and their gene structures are highly conserved across species.8

Initiation of a MAPK signaling module involves catalytic activation of MAPKKK by upstream cellular receptors, G-proteins, and sometimes by phosphorylated MAPKKK and other protein kinases.2,6 Once phosphorylated, the MAPKKK then activates MAPKK through phosphorylation of two serine/threonine residues in the activation segment, followed by dual phosphorylation of MAPK by MAPKK through phosphorylation of both threonine and tyrosine residues at the TXY motif17–19 in the activation or T-loop located in between the kinase subdomains VII and VIII. This phosphorylation results in an activation of certain transcription factors along with the transduction of MAPK signaling to numerous substrates and downstream protein kinases.20 With the onset of the signal transduction module, the MAPK proteins are promptly translocated from the cytoplasm to the nucleus,21 while the substrates of MAPKs are abundantly present in cytoplasm (for example, phospholipase A2,22 and transmembrane proteins, such as the endothelial growth factor receptor).23 These genes have a conserved domain for docking to their cognate activators, suppressors, and protein substrates to increase the efficiency during protein–protein interactions.24 Inactivation of MAPK genes is carried out by dephosphorylation of the activation motif by protein phosphatase, including tyrosine phosphatases, serine/threonine-specific phosphatase, and dual specificity MAPK phosphatases,25 a process equally important in establishing the physiological equilibriums in living cells. Some of the functional significances of the dephosphorylation of MAPKs can be described by the role of AP2C3 MAPK phosphatase-regulating ectopic cell development and epidermal cell conversion leading to stomatal development in plants.26 Another MAPK phosphatase known as PP2C5 has been found to enhance the ABA sensitivity, and therefore ABA induced seed dormancy and stomatal closure.27

MAPK regulatory pathways are involved in several developmental and defense mechanisms underlining the functional importance of the gene families—MAPK, MAPKK, and MAPKKK (Table 1). In Arabidopsis, some of the MAPKs, like MAPKKK6 (AtMAPKK6) and MAPKKKe2 (AtMAPKK7), are found to be vital for the normal development and functioning of pollen28 and YODA (AtMAPKK4) in cell fate during embryonic development.29,30 Similarly, MAPK3 and MAPK6
Table 1. A synopsis of some of the known MAPK functions in plant species.

| Gene names | Functions | References |
|------------|-----------|------------|
| MAPKKK     |           |            |
| AtMEKK1    | Bacterial and fungal pathogens | 31 |
| ANP1 (=AtMAPKKK1) | Oxidative stress | 92 |
| MAPKKKα1 (=AtMAPKKK6) | Pollen viability | 28 |
| MAPKKKα2 (=AtMAPKKK7) | Pollen viability | 28 |
| YODA, YDA (=MAPKK4) | Extra-embryonic cell fate | 29,30 |
| CTR1 (Raf1) | Oxidative stress, ethylene signaling | 93 |
| ANP3 (=AtMAPKKK2) | Cytokinesis, phragmoplast assembly | 92 |
| EDR1 (Raf2) | Defensive responses | 94 |
| MAPKK      |           |            |
| AtMKK1     | Cold stress | 96 |
| GmMKK1     | Downey mildew, soybean mosaic virus | 38 |
| OsMEK1     | Root apex proliferation | 97 |
| ZmMEK1     | Cytokinesis, cell death, bacterial elicitor signaling | 37,98–100 |
| AtMAPKK2   | Cold and salt stress | 101 |
| NtMEK1     | Ethylene signaling, fruit ripening | 102 |
| AtMKK3     | Participate in JA signaling | 103,104 |
| GmMKK2     | Downey mildew, soybean mosaic virus | 38 |
| AtMAPKK4   | Response to microbial pathogens, stomatal development | 31,105 |
| OsMAPKK4   | Fungal pathogens | 35 |
| AtMAPKK5   | Bacterial and fungal pathogens, stomatal development | 31,105 |
| AtMKK6     | Cytokinesis | 37,106 |
| AtMKK7     | Polar auxin transport | 107 |
| AtMAPKK9   | Leaf senescence, ethylene biosynthesis | 108 |
| MAPK       |           |            |
| AtMAPK1    | Activated by JA, ABA and H$_2$O$_2$ | 109 |
| AtMAPK2    | Activated by JA, ABA and H$_2$O$_2$ | 109 |
| PsMAPK2    | Seed germination | 110 |
| AtMAPK3    | Bacterial and fungal pathogens | 31,111 |
| MsMAPK3    | Cell cycle regulation (after metaphase), cytokinesis | 112 |
| OsMAPK3    | Fungal pathogen | 35 |
| OsMAPK3    | Stomatal development | 105 |
| AtMAPK4    | Cold and salt stress | 101 |
| AtMAPK4    | Immune responses | 113 |
| GmMAPK4    | Downey mildew, soybean mosaic virus | 38 |
| AtMAPK6    | Bacterial and fungal pathogens, cold and salt stress, Leaf senescence, stomatal development | 31,101,105,108,111 |
| OsMAPK6    | Fungal pathogen | 35 |
| GhMAPK7    | Colletotrichum nicotianae (fungus), PVY virus | 114 |
| AtMPK12    | Auxin signaling | 115 |
| AtMPK18    | Stabilization of microtubule | 116 |

in Arabidopsis are found to mediate signals in response to microbe-associate molecular patterns (MAMPs), such as flagellin (flg22) and chitin elicitors. The same MAPKs in Arabidopsis are known to control cell death along with the pathogenic responses by producing reactive oxygen species and indole-derived phytoalexin (camalexin). Orthologs of these two genes in rice are known to play similar roles in fungal resistance, but their upstream MAPKK pathways and complete roles are not yet fully understood. Similarly, MAPKKs such as NtMEK1 in tobacco and AtMKK6 in Arabidopsis, are found to regulate cytokinesis, indicating their crucial role in early-life activities. In soybean, these genes are reported to be involved in developmental and various stress responses, yet largely remain uncharacterized. Homologs of MAPK4 in soybean are reported to negatively regulate defense responses to diseases such as downy mildew and soybean mosaic virus, and positively regulate plant development and growth. Some of the MAPKs in
soybean, such as MAPK1 homolog and another MAPK homolog of 49 kDa (wound-activated protein kinase), are activated in response to salt stress and to elevated phosphatidic acid in wounded soybean plants, respectively.

With the recent completion of the soybean genome sequencing project, it has now become possible to identify and characterize GmMAPKs for the advancement of soybean research. In addition, comparative genomics of legume species along with their nonlegume relatives would allow us to identify any legume-specific MAPK genes that might regulate legume-specific processes such as symbiotic nodule development and isoflavonoid biosynthesis. Arabidopsis MAPKs and MAPKKs were the first published plant MAPKs to be systematically named. There are a few studies, however, that have applied the Arabidopsis model in their MAPK nomenclature, and nomenclatural inconsistencies are very common even in some recent literature, making the communication about these genes very difficult. Although the MAPK genes have been identified in a few plant species including Arabidopsis, only a few of them are characterized and our overall knowledge about these genes in many other plant species including soybean is very limited. The primary objectives of this study were to identify the members of all three subfamilies of GmMAPKs and assess their functional and evolutionary relationships. Identification of GmMAPK genes is the vital first step towards understanding their roles in stress response, growth, development, and defense mechanism in soybean. Understanding evolutionary relationships and functions of these genes is crucial to soybean crop improvement with potential implications in other plant species.

Results
We identified 38 GmMAPKs, 11 GmMAPKKs, and 150 GmMAPKKKs. The results for each MAPK family are described below. We also propose a nomenclature schema consistent with founder MAPKs in Arabidopsis to enable efficient comparative genomics (Table 2). The phylogenetic analysis using both maximum parsimony (MP) and maximum likelihood (ML) resulted into trees with similar topologies, but with slight variation in branch support for MAPK, MAPKK, and MAPKKK datasets. The model test performed showed that the Jones—Taylor—Thornton model with discrete gamma distribution and invariant sites (G+I) was the best-fit evolutionary model for each dataset.

MAPKs
The GmMAPK amino acid sequence length ranged from 326 to 615. The identified 38 GmMAPKs were nested into five distinct clades as shown in Figure 1 and Additional file 1. Phylogenetic placement of the MAPK genes in the ML tree (Fig. 1) is similar to that in the MP tree, but there was a slight variation in the bootstrap support (Additional file 1). Among these five clades, four clades (A, B, C, and D) are well supported and corresponded to those of their homologs in Arabidopsis, rice, and poplar. The members with phosphorylation motif TEY were found to be nested in clade A, B (except GmMAPK5-2 with the TVY motif), and C, while those genes with the TDY motif were nested in clade D. The numbers of GmMAPKs in clade A, B, C, D, and E were four, ten, four, 14, and six, respectively. The fifth clade, E, contained genes that were not orthologs of any of the Arabidopsis MAPKs, four of which had TEY motif and two genes (GmMAPK22-1 and GmMAPK22-2, denoted by * in Fig. 1 and Additional file 1) had a TQY motif. The occurrence of MAPK with the TQY motif has not been reported in other plant species before. Four paralogs of five MAPK genes (GmMAPK5, GmMAPK9, GmMAPK16, GmMAPK20, and GmMAPK23), along with two paralogs of each of six different MAPK genes (GmMAPK3, GmMAPK4, GmMAPK6, GmMAPK11, GmMAPK13, and GmMAPK22) were identified. We did not find the orthologs of AtMAPK15, AtMAPK17, and OsMAPK21 in soybean.

Functional divergence among the paralogs of MAPK gene members, including their functional conservation across taxa, were assessed through transcriptomic data analyses. Expression profiles based on transcriptomic data showed that GmMAPK5-2, GmMAPK5-3, GmMAPK5-4, GmMAPK9-3, GmMAPK9-4, and GmMAPK11-1 had the lowest or no expression in the tissues examined under the given experimental conditions, while the GmMAPK1, GmMAPK2, GmMAPK20-4, GmMAPK23-1, GmMAPK23-2, GmMAPK23-3, and GmMAPK23-4 had relatively higher expression values (Fig. 2).

MAPKKs
The GmMAPKK amino acid sequence length ranged from 227 to 526. As illustrated in Figure 3 and
Table 2. Nomenclature of Arabidopsis MAPK orthologs across five plant species (Arabidopsis, poplar, rice, grapes, and soybean). The names in the bold letters within parentheses are the suggested names.

| Arabidopsis | Poplar | Rice | Grapes | Soybean | Phytozome ID |
|-------------|--------|------|--------|---------|--------------|
| AtMAPK1     | PtMPK1 | OsMPK3 (OsMPK1) | VvMPK2 (VvMPK1) | GmMAPK1 | Glyma04g03210 |
| AtMAPK2     | PtMPK2 | OsMPK4 (OsMPK2) | GmMAPK2 | Glyma06g03270 |
| AtMAPK3     | PtMPK3-1 | OsMPK5 (OsMPK3) | GmMAPK3-1 | Glyma11g15700 |
|             | PtMPK3-2 |         | GmMAPK3-2 | Glyma12g07770 |
| AtMAPK4     | PtMPK4 | OsMPK6 (OsMPK4) | VvMPK5 (VvMPK4) | GmMAPK4-1 | Glyma07g07270 |
| AtMAPK5     | PtMPK5-1 |         | GmMAPK4-2 | Glyma16g03670 |
|             | PtMPK5-2 |         | GmMAPK5-2 | Glyma11g02420 |
|             |         |         | GmMAPK5-3 | Glyma08g02060 |
|             |         |         | GmMAPK5-4 | Glyma05g37480 |
| AtMAPK6     | PtMPK6-1 | OsMPK1 (OsMPK6) | VvMPK6 (VvMPK6) | GmMAPK6-1 | Glyma07g32750 |
|             | PtMPK6-2 | OsMPK4 (OsMPK7) | GmMAPK6-2 | Glyma02g15690 |
| AtMAPK7     | PtMPK7 | OsMPK7 | GmMAPK7 | Glyma05g28980 |
| AtMAPK8     | PtMPK8 | OsMPK14 (OsMPK9) | VvMPK7 (VvMPK9) | GmMAPK9-1 | Glyma05g33980 |
| AtMAPK9     | PtMPK9-1 | OsMPK2 (OsMPK11) | GmMAPK11-1 | Glyma18g47140 |
|             | PtMPK9-2 | GmMAPK11-2 | Glyma09g39190 |
| AtMAPK10    | PtMPK10 | VvMPK9 (VvMPK13) | GmMAPK13-1 | Glyma12g07850 |
| AtMAPK11    | PtMPK11 | OsMPK3 (OsMPK14) | VvMPK1 (VvMPK14) | GmMAPK14 | Glyma08g12150 |
| AtMAPK12    | PtMPK12 | VvMPK10 (VvMPK16) | GmMAPK16-1 | Glyma13g28120 |
| AtMAPK13    | PtMPK13 | OsMPK17 (OsMPK19) | GmMAPK18 | Glyma15g38490 |
| AtMAPK14    | PtMPK14 | OsMPK10 (OsMPK20-1) | GmMAPK19 | Glyma13g33860 |
|             | PtMPK20-2 | OsMPK9 (OsMPK20-2) | GmMAPK20-1 | Glyma18g12720 |
|             | PtMPK20-2 | OsMPK11 (OsMPK20-3) | GmMAPK20-2 | Glyma14g30310 |
|             | PtMPK20-2 | OsMPK8 (OsMPK20-4) | GmMAPK20-3 | Glyma02g45630 |
|             | PtMPK20-2 | OsMPK7 (OsMPK20-5) | GmMAPK20-4 | Glyma08g22240 |
| AtMAPK17    | PtMPK17 | VvMPK11 (VvMPK19) | GmMAPK22-1 | Glyma03g21610 |
| AtMAPK18    | PtMPK18 | VvMPK12 (VvMPK20) | GmMAPK22-2 | Glyma16g10820 |
| AtMAPK19    | PtMPK19 | GmMAPK23-1 | Glyma01g35190 |
|             | PtMPK20-1 | GmMAPK23-2 | Glyma09g34610 |
|             | PtMPK20-2 | GmMAPK23-3 | Glyma16g08080 |
|             | PtMPK20-2 | GmMAPK23-4 | Glyma16g17580 |
| MAPKK       | AtMAPKK1 | GmMAPKK1 | Glyma15g18860 |
| AtMAPKK2    | PtMKK2-1 | GmMAPKK2-1 | Glyma13g16650 |
| AtMAPKK3    | PtMKK2-2 | GmMAPKK2-2 | Glyma17g06020 |
| AtMAPKK4    | PtMKK3 | GmMAPKK3-1 | Glyma05g08720 |
| AtMAPKK5    | PtMKK4 | GmMAPKK3-2 | Glyma19g00220 |
| AtMAPKK6    | PtMKK5 | GmMAPKK4 | Glyma07g0520 |
|             | PtMKK6 | GmMAPKK5 | Glyma08g23900 |
|             |         | GmMAPKK6-1 | Glyma10g15850 |
|             |         | GmMAPKK6-2 | Glyma02g32980 |

(Continued)
Table 2. (Continued)

| Arabidopsis | Poplar | Rice | Grapes | Soybean | Phytozome ID |
|-------------|--------|------|--------|---------|--------------|
| AtMAPKK7    | PtMKK7 |      |        |         |              |
| AtMAPKK8    | PtMKK11-1 | PtMKK11-2 |        |         |              |
| AtMAPKK9    | PtMKK9 |      |        |         | Glyma09g30310 |
| AtMAPKK10   | PtMKK10 |      |        |         | Glyma01g01980 |

Figure 1. Maximum likelihood analysis of GmMAPKs and their orthologs in Arabidopsis, poplar, and rice.

Notes: The values above the branches are bootstrap support of 100 replicates. The JTT+G+I evolutionary model was employed in MEQ5.2.2 to perform maximum likelihood analysis. The members with phosphorylation motif TEY are included in clades A, B, and C; TDY in clade D, and members with the TQY (denoted by *) and the TVY (denoted by **) motif in clades E and B, respectively. The MAPK gene models were accepted for phylogenetic analysis using protein sequences of the serine/threonine kinase subfamily having conserved aspartate and lysine residues in their catalytic domain with the (D[K/L/V]) motif and the TXY phosphorylation motif in their activation loop.
Mitogen-activated protein kinase (MAPK) genes in soybean

Evolutionary Bioinformatics 2013:9

0.00 3.11 6.23

Young leaf
Flower
One cm pod
Pod shell 10DAF
Pod shell 14DAF
Seed 10DAF
Seed 14DAF
Seed 21DAF
Seed 25DAF
Seed 28DAF
Seed 35DAF
Seed 42DAF
Root
Nodule

GmMAPK9-4
GmMAPK9-3
GmMAPK9-1
GmMAPK8
GmMAPK6-2
GmMAPK6-1
GmMAPK5-4
GmMAPK5-3
GmMAPK5-2
GmMAPK5-1
GmMAPK4-2
GmMAPK4-1
GmMAPK3-2
GmMAPK3-1
GmMAPK23-4
GmMAPK23-3
GmMAPK23-2
GmMAPK23-1
GmMAPK22-2
GmMAPK22-1
GmMAPK20-4
GmMAPK20-3
GmMAPK20-2
GmMAPK20-1
GmMAPK21
GmMAPK19
GmMAPK18
GmMAPK16-4
GmMAPK16-3
GmMAPK16-2
GmMAPK16-1
GmMAPK14
GmMAPK13-2
GmMAPK13-1
GmMAPK11-2
GmMAPK11-1
GmMAPK1

Figure 2. Heatmap visualization of GmMAPKs.
Note: Log 2-based value was employed to construct the heatmap for MAPK gene expression in different tissues and treatment conditions.

Additional file 2, eleven GmMAPKKs were identified including one new MAPKK (ie, GmMAPKK11) that lacked a potential ortholog in Arabidopsis. The GmMAPKK members formed four distinct clades that corresponded to the four clades of the Arabidopsis MAPKKs (A–D). The topologies in the ML tree and MP tree (Additional file 2) were similar, with slight variations in bootstrap support. We did not find the orthologs of AtMAPKK7 and AtMAPKK9 in soybean. Clade A with five GmMAPKKs included GmMAPKK1 and two paralogs of each GmMAPKK2 and GmMAPKK6; clade B included two paralogs of GmMAPKK3; clade C had two genes—GmMAPKK4 and GmMAPKK5; whereas, clade D included two genes, GmMAPKK8 and GmMAPKK10. Expression data showed that GmMAPKK8 and GmMAPKK10 had the lowest expression values, while GmMAPKK4 and GmMAPKK5 had the highest expression values among GmMAPKK genes (Fig. 4).

MAPKKs
The GmMAPKK amino acid sequence length ranged from 228 to 1411. Altogether, 150 GmMAPKKs were identified (Fig. 5 and Additional file 3). The GmMAPKK members formed three distinct clades: MEKK-like (34 genes), Raf-like (92 genes), and ZIK-like (24 genes), consistent to those in Arabidopsis. The GmMEKK, GmRaf, and GmZIK-like members are color-coded: MEKK, Raf, and ZIK-like genes are represented by blue, black, and red branches, respectively (Fig. 5). The ML tree topologies were similar to MP tree (Additional file 3), with slight variations in branch supports. We found multiple paralogs of MAPKKs in soybean, orthologs of which were not recognized as paralogs in Arabidopsis. Orthologs of some of the MEKK-like gene members (AtMAPKKK8, AtMAPKKK9, AtMAPKKK12, AtMAPKKK15, AtMAPKKK16, AtMAPKKK19, AtMAPKKK20, and AtMAPKKK21), Raf-like members (AtRaf5, AtRaf7–AtRaf9, AtRaf14, AtRaf15, AtRaf24, AtRaf25, AtRaf44–AtRaf46, and AtRaf48), and a ZIK-like member (AtZIK7) of Arabidopsis could not be recovered in soybean. However, new MAPKKK members of each MEKK-like, Raf-like, and ZIK-like subgroup were recovered in soybean. The new MEKK-like gene members in soybean included the paralogs of GmMAPKK22, GmMAPKK22, and GmMAPKK24; the new Raf-like gene members included GmRaf49, GmRaf50, GmRaf51, GmRaf52, and GmRaf53, and the new ZIK-like members that included two paralogs of GmZIK12. Expression profiles showed: (1) GmMAPKKK5-1, GmMAPKKK17, GmMAPKKK18-1, GmMAPKKK18-2, GmMAPKKK18-3, GmMAPKKK22-1, and GmMAPKKK22-2 had the lowest to no expression values, while GmMAPKKK3-1, GmMAPKKK3-2, GmMAPKKK4-4, and GmMAPKKK23-1 had the highest expression values.
Figure 3. Maximum likelihood analysis of GmMAPKKs and their orthologs in Arabidopsis, poplar, and rice. Notes: in the ML phylogram, the values above the branches are bootstrap support of 100 replicates. The JTT+G+I evolutionary model was employed in MEGA5.2.2 to perform maximum likelihood analysis. The MAPKK gene models were accepted for phylogenetic analysis using dual-specificity protein kinases having conserved aspartate and lysine residues in their catalytic domain with the (D[L][I]V/K) motif and the S-X$_5$T phosphorylation motif along with their activation loop.

values among MEKK-like genes (Fig. 6); (2) GmRaf4-1, GmRaf4-2, GmRaf6-2, GmRaf6-3, GmRaf6-4, GmRaf12, GmRaf16-1, GmRaf16-2, GmRaf19-1, GmRaf19-2, GmRaf19-3, GmRaf19-4, GmRaf26, GmRaf30-2, GmRaf34-2, GmRaf41-2, GmRaf42-3, GmRaf49-3, GmRaf51-1, GmRaf51-2, and GmRaf52 had the lowest expression values, while GmRaf2-2, GmRaf17-2, GmRaf20-1, GmRaf20-2, GmRaf21, GmRaf28, GmRaf29, GmRaf33-2, GmRaf49-2, and GmRaf54-1 had the highest expression values among Raf-like genes (Fig. 7A and B); and (3) GmZIK2-2, GmZIK2-3, GmZIK5, GmZIK8-1, GmZIK8-2, GmZIK8-3, GmZIK12-1, and all the paralogs of GmZIK1 except in root tissue) had the lowest expression values, while GmZIK6-1, GmZIK6-2, GmZIK9-2, and GmZIK11 had the highest expression values among ZIK-like genes in all examined tissues (Fig. 8).
Disc strands of non-coding RNA, the RNAs that could not be used to construct the mRNA and the protein. In this regard, a large number of long non-coding RNA and microRNAs are abundant in the soybean genome. The long non-coding RNA such as antisense RNA, sense-antisense RNA, and sense-like RNA were found in many soybean genotypes. The microRNAs in soybean are abundant and have a wide range of functions. The microRNAs have been found to regulate a variety of processes such as development, stress response, and metabolism.

Common Alligation
The common alligation, a type of multiplication, is a method of calculating a proportion of a given quantity. For example, if a farmer has 100 bushels of corn and he wants to give 25% of it to his children, he can use the common alligation to find the amount of corn each child will receive. The common alligation is also used to calculate the price of a product when the price of the product is known and the quantity of the product is unknown. The common alligation is also used to calculate the price of a product when the price of the product is unknown and the quantity of the product is known.

Review of Previous Studies
Previous studies have shown that the MAPKs in soybean have a variety of functions, including growth, development, stress response, and metabolism. The MAPKs in soybean have been found to be involved in a variety of processes such as development, stress response, and metabolism. The MAPKs in soybean have been found to be involved in the regulation of the expression of genes involved in the processes such as development, stress response, and metabolism.

Discussion
Genomic structure and comparative genomics
Soybean is an allopolyploid species, which has undergone two major polyploidization events approximately 59 million and 13 million years ago, followed by chromosomal rearrangements that perhaps resulted into extinction and diversification of its genes. Approximately 25% of the duplicated genes of the soybean genome are estimated to have been lost, averaging 3.1 retained copies. Ancient events conforming duplication of individual genes, whole genomes, or segmental duplication of chromosomes are thought to have contributed to the evolutionary novelties resulting in functional complexities. Gene duplication could occur by three processes: unequal crossing over (results in tandem duplication), retroposition (results in random gene insertions), and chromosomal rearrangement (results in tandem gene duplication), retroposition could occur by three processes: unequal crossing over (results in functional complexities).

In soybean, we found MAPK1 and two paralogs of MAPK2 and MAPK6 in clade A, two paralogs of MAPK3 in clade B, MAPK4 and MAPK5 in clade C, and MAPK8 and MAPK10 in clade D. Similarly, newly identified GmMAPKK members formed three clades (34 genes in the MEKK-like, 92 in the Raf-like, and 24 in the ZIK-like subgroup) that are consistent with those reported in Arabidopsis and Oryza. Multiple Expectation-maximization for Motif Elicitation (MEME) analysis for the approximate sequence pattern, including logos of protein motifs related to MAPK-specific “signals” also showed the conservation of domains throughout the gene members of each subfamily and of the subgroups (Figs. 9–11). Some of these conserved domains as predicted by Pfam and PROSITE for MAPK and MAPKKs included protein kinase catalytic domain, MAPK-conserved site (specific to MAPKs), serine/threonine-active site, and the Adenosine Triphosphate (ATP) binding site. Some of the domains specific to MAPKs included
Notes: Phylogenetic representation of GmMAPKKKs in circular tree format shows the three subgroups: GmMEKK-like, GmRaf-like, and GmZIK-like are indicated by blue, black, and red branches, respectively. The JTT+G+I evolutionary model was employed in MEGA5.2.2 to perform maximum likelihood analysis with 100 bootstrap replicates. The MAPKKK gene models were accepted for phylogenetic analysis using protein sequences of serine/threonine kinase subfamily having conserved aspartate and lysine residues in their catalytic domain with the (D/L/I/V)K motif, and the members in each subgroup were categorized based on their signature motifs.

The genome size of soybean (1115 MB; 46,430 protein coding genes)\textsuperscript{41,47} is larger than that of Arabidopsis (125 MB; 27,416 genes),\textsuperscript{48} grapevine (487 MB; 30,434 genes),\textsuperscript{49} poplar (~485 MB; 45,555 genes),\textsuperscript{50} and rice (389 MB; ~37,544 genes).\textsuperscript{51} The number of
paralogs in the MAPK family is greater in soybean than in Arabidopsis, grapes, poplar, and in rice as a result of evolutionary processes involving gene duplication and protein diversification. For example, MAPK16 has four paralogs in soybean, two in poplar, and one each in Arabidopsis and Brachypodium, and none in rice, suggesting higher diversity of these genes in dicots. Similarly, MAPKK10 paralogs are more diversified in monocots than in dicots, as evidenced by three paralogs in rice, five in Brachypodium, and one in each of soybean, poplar, and Arabidopsis. Several of these duplication events are deemed to have taken place long before the eudicot-monocot branching. For example, the number of paralogs of MAPK20 in soybean, rice, and poplar are four, five, and two, respectively. This inference in soybean is consistent to that in rice and poplar. Previous studies in rice and Brachypodium showed the absence of MAPK1, MAPK2, MAPK5, MAPK9, and MAPK11,
whereas the orthologs of these genes are found to be conserved in most eudicot species, suggesting the evolution of these orthologs in eudicot species. Conversely, the MAPK21 paralogs were found to have evolved exclusively in monocot species. Within the eudicot species, our analyses revealed the absence of some Arabidopsis orthologs in soybean and poplar. For example, the Arabidopsis ortholog of MAPK12 is absent in both poplar and soybean. Additionally, MAPK13 and MAPK10 are absent in poplar and soybean, respectively. In our separate analysis of MAPK gene members from four legumes and three nonlegume species, we consistently recovered the putative legume-specific clade (clade E) consisting of both TEY and TQY members, which was distinctly separated from the four traditional clades of the MAPK subfamily. The presence of the new clade (clade E; Fig. 1) – including its members with the TQY motif, most likely descended from the TEY type – is an evolutionary innovation within the legume species.

As shown in Figure 1, the five paralogs of OsMAPK20 are present in three different chromosomes: OsMAPK20-1 and OsMAPK20-4 in chromosome 1; OsMAPK20-2 and OsMAPK20-5 in chromosome 5; and MAPK20-3 in chromosome 6. A restriction fragment length polymorphism
Mitogen-activated protein kinase (MAPK) genes in soybean

Evolutionary Bioinformatics 2013:9

Figure 9. Predicted domain structure of GmMAPKs.
Notes: Conserved domain structures as predicted by MEME analysis of the GmMAPK subfamily. Ten different sites were analyzed for the prediction of conserved domain structures in the MAPK subfamily. Each stack height in the logos for ten different predicted motifs represents the sequence conservation in that region which is measured in bits, whereas the height of each residue within the stack indicates the frequency of corresponding amino acid competing for that position.

test performed in rice showed that chromosome 1 and chromosome 5 are ancient duplicates. Likewise, chromosome 2 and chromosome 6 of rice are believed to be ancient duplicates, each housing a paralog of OsMAPK17 (Phytozone: LOC_Os02g04230 and Phytozone: LOC_Os06g49430). Orthologs of AtMAPK20 in Glycine max are present in four paralogs, but all in different chromosomes that possibly resulted from two rounds of whole genome duplications. This duplication process seems distant possibility for segmental or chromosomal duplication, as there is a slight chance for the same locus to undergo two events of duplications. Similarly, the ortholog of AtMAPK16 in soybean has evolved four paralogs, compared to only one in rice and two paralogs in poplar, possibly as a result of successive whole genome duplication events. These differences in the number of duplicates make sense considering the number of chromosomes (2n = 40) in soybean compared to that in rice (2n = 24), along with the differences in their genome sizes. From our combined phylogenetic analysis of Arabidopsis, rice, poplar, and soybean (Fig. 1), the paralogous relationship between MAPK7 and MAPK14 was also recognized in soybean. In parallel to the findings from Arabidopsis, rice, grapevine, and poplar, genes of clades A, B, and C of GmMAPKs (Fig. 1) have TEY, and those of clade D have TDY amino acid motifs. Interestingly, two new types of MAPKs were identified in soybean that were not previously reported in plants—MAPKs containing TVY (GmMAPK5-2) and TQY motifs (GmMAPK22-1 and GmMAPK22-2) in clade B and clade E, respectively. Surprisingly, when soybean MAPK amino acid sequences with the TQY motif were searched for using the Basic Local Alignment Search Tool in the National Center for Biotechnology Information, they were also found in nematodes such as Caenorhabditis elegans (GenBank: NP_494947.2), Caenorhabditis brenneri (GenBank: EGTS1072.1), Brugia malayi (GenBank: XP_001896626.1), Loa loa...
Neupane et al

Figure 10. Predicted domain structure of GmMAPKKs.
Notes: Conserved domain structures as predicted by MEME analysis of the GmMAPKK subfamily. Ten different sites were analyzed for the prediction of conserved domain structures in the MAPKK subfamily. Each stack height in the logos for ten different predicted motifs represents the sequence conservation in that region which is measured in bits, whereas the height of each residue within the stack indicates the frequency of the corresponding amino acid competing for that position.

(GenBank: XP_003140630.1), Caenorhabditis remanei (GenBank: XP_003109019.1), and Caenorhabditis briggsae (GenBank: XP_002630655.1). Further investigation into the role of these genes in legume—nematode interactions would provide valuable insights as to whether these genes with the TQY motif have evolved independently in legume and nematode species. The presence of MAPKs with the TQY motif in distantly related groups (legumes and nematodes) is perhaps due to evolutionary convergence.

As in MAPKK10 of Arabidopsis, poplar, and rice, GmMAPKK10 also lacks a complete activation loop (S/TxxxxxS/T) that is required for the phosphorylation of MAPKKs in plants. Nonetheless, GmMAPKK10, as in Arabidopsis, poplar, and rice orthologs, retains the lysine and aspartate residues. These residues are believed to be required within the designated motif—D(L/I/V)K—of the catalytic loop for its kinase activity. It is also worth noting that GmMAPKKs in clades C and D (GmMAPKK4, GmMAPKK5, GmMAPKK8, and GmMAPKK10) have only one exon, consistent with the Arabidopsis model. MAPKKK members in soybean are found in relatively large numbers (150 MAPKKKs) compared to 80 members in Arabidopsis and 75 in rice. The comparative genomics of MAPKK genes is beyond the scope of this study and will be discussed elsewhere.

Functional genomics

Presence of large numbers of duplicate MAPK genes in the soybean genome led us to survey the functional genomics of these genes. Gene duplication may result in functional redundancy, divergence, and diversification, including neofunctionalization (where one of the copies is assigned a new function), nonfunctionalization (where one of the copies is destined to lose function), subfunctionalization (where duplicated gene copies evolve to complement one another to retain the ancestral gene function), or hypofunctionalization (where expression of one of the copies is diminished). With the recent surge in whole genome sequencing and the availability of sequencing data, it is not feasible to characterize each and every single gene in different species through laboratory experiments. Therefore, integrated approaches including bioinformatics and functional genomics are vital to the study of gene functions and their evolution. We used transcriptomic data (www.soybase.org, http://plantgrn.noble.org/LegumeIP and http://mpss.udel.edu/) and mapped them onto phylogenies to assess evolution and functional divergence of GmMAPKs. Genes involved in stress-specific physiological responses are evolutionarily more conserved and the genes involved in deoxyribonucleic acid (DNA) repair are likely to be lost; this is as expected of allopolyploidization. Some genes are deemed to become nonfunctional and some will evolve to gain new functions. Using phylogenetic placements and expression profiles, we inferred functional divergence in paralogs of GmMAPK3, GmMAPK4, GmMAPK5, and GmMAPK11 (Fig. 2). In addition, GmMAPK5-2 with the presence of a TVY motif in the activation loop differs from its paralog.
GmMAPK5-1, though the functional significance for this change is yet to be investigated. Our inference about the subfunctionalization MAPK16 paralogs in soybean (Fig. 2) is consistent with the functional divergence reported for their orthologs in poplar. On the other hand, those genes with very low levels to null expression in all tissues are perhaps undergoing constraints of nonfunctionalization. Interestingly, out of six GmMAPK genes nested into clade E, four paralogs of GmMAPK23 have the TEY motif and two paralogs of GmMAPK22 have the TQY motif (Fig. 1). The majority of these genes have higher expression values. The presence of the novel clade of these genes (two of them with the TQY motif) led us to predict that this clade of genes is perhaps involved in legume-specific physiology and organogenesis. Previous work has shown that nematodes such as C. elegans play an important mediating role that is helpful in establishing legume—rhizobia symbiotic interactions by transferring the Sinorhizobium meliloti to the root tissues of Medicago truncatula in response to nematode-attracting signaling molecules released by the plant.

Interestingly, C. elegans, which has some MAPKs with the TQY motif, is able to identify the bacteria for their food needs under varying environmental conditions. Further investigation of these genes and potential nematode—rhizobia—legume interaction(s) would provide valuable insights into their roles in root nodulation or any legume-specific physiology.

In terms of transcriptomic data survey for MAPKK genes, Figure 4 shows relatively higher levels of expression values for GmMAPKK4 and GmMAPKK5 than other GmMAPKs. The expression pattern of these two genes is consistent with that of their orthologs in Arabidopsis (source: MPSS database). Protein—protein interactions among AtMAPKK4 and AtMAPKK5 with AtMAPK6 have also been well established in a previous study, as compared to other AtMAPKKs. Interestingly, the ortholog of AtMAPKK4 in rice (Phytozome: LOC_Os02g54600) also shows higher protein expression values relative to other OsMAPKs. The ortholog of this gene in rice is associated with OsMAPK3 (Phytozome: LOC_Os03g17700) acting as an upstream MAPKK. Phosphorylation of OsMAPKK4 is undertaken by six upstream MAPKKKs: Phytozome: LOC_Os01g50370; LOC_Os02g54600; LOC_Os03g17700; LOC_Os04g56200; LOC_Os05g56400; LOC_Os06g56600.

![Figure 11. (Continued, figure legend on page 17)](image-url)
Figure 11. (Continued, figure legend on page 17)
In soybean, GmMAPKK8 and GmMAPKK10 show no expression in any of the tissues examined (Fig. 4). This result in soybean was also consistent to the MPSS expression data for MAPKK8 and MAPKK10 of Arabidopsis, including all the paralogs of MAPKK10 in poplar and in rice, except for OsMAPKK10-2, as previously shown. Protein—protein interaction assays on MAPK and MAPKK of Arabidopsis also suggest no evidence of the interaction of AtMAPKK8 with any of the AtMAPKs, and a very weak interaction of AtMAPKK10 with only AtMAPK17. In the same study, there was no evidence of the interaction of AtMAPK8, AtMAPK9, AtMAPK12, AtMAPK16, AtMAPK18, and AtMAPK19 with any of the AtMAPKK members. This perhaps could serve as evidence that the majority of the plant signal transductions involve a few regulatory MAPKs, and the same signaling pathway might be used for multiple responses.

In the case of GmMAPKKs, a large number of MEKK-like, Raf-like and ZIK-like genes present a more complex evolutionary pattern. Relative to evolutionarily, less dynamic MAPKKs of Arabidopsis, the GmMAPKK subfamily seems to be extensively amplified and functionally diversified. Differentially expressed paralogs of different genes of the MAPKK subfamily also showed an interesting pattern of functional divergence. One or more paralogs of GmMAPKKs such as GmMAPKK3 (GmMAPKK3-1 and GmMAPKK3-3), GmMAPKK4 (GmMAPKK4-1, GmMAPKK4-2, and GmMAPKK4-3), GmMAPKK11 (GmMAPKK11-1 and GmMAPKK11-2), GmMAPKK13 (GmMAPKK13-1 and GmMAPKK13-2), GmMAPKK24 (GmMAPKK24-1 and GmMAPKK24-2), GmRaf3 (GmRaf3-1 and GmRaf3-2), GmRaf20 (GmRaf20-1 and GmRaf20-2), GmRaf23 (GmRaf23-1 and GmRaf23-2), GmRaf27 (GmRaf27-1 and GmRaf27-2), GmRaf31 (GmRaf31-1, GmRaf31-2, and GmRaf31-3), GmRaf32 (GmRaf32-1 and GmRaf32-2), GmRaf33 (GmRaf33-1 and GmRaf33-2), GmRaf35 (GmRaf35-1 and GmRaf35-2), GmRaf36 (GmRaf36-1, GmRaf36-2, and GmRaf36-3), GmZIK6 (GmZIK6-1 and GmZIK6-2), and GmZIK9 (GmZIK9-1 and GmZIK9-2) are inferred to have undergone functional divergence either through subfunctionalization or neo-functionalization of the duplicated copies (Figs. 6–8). Similarly, some of the duplicated genes are retaining or gaining significant levels of expression in one or few tissues, as...
found in the case of *GmMAPKKK22* (*GmMAPKKK22-1* and *GmMAPKKK22-2*), *GmRaf16-2*, *GmRaf19-3*, *GmRaf51-1*, and *GmZIK1* (*GmZIK1-1* and *GmZIK1-2*). We speculate this gain in function by these genes might have occurred after the duplication of genes that had undergone severe mutation, destroying their ability to process information for biological processes and functions. From the expression data, one or more duplicated copies of the MAPKK genes in soybean, such as *GmMAPKKK5-1*, *GmMAPK18* (*GmMAPKKK18-1*, *GmMAPKKK18-2*, and *GmMAPKKK18-3*), *GmMAPKKK17*, *GmRaf4* (*GmRaf4-1* and *GmRaf4-2*), *GmRaf6* (*GmRaf6-2*, *GmRaf6-3*, and *GmRaf6-4*), *GmRaf16* (*GmRaf16-1* and *GmRaf16-2*), *GmRaf19* (*GmRaf19-1*, *GmRaf19-2*, *GmRaf19-3*, and *GmRaf19-4*), *GmRaf26*, *GmRaf30* (*GmRaf30-2*, *GmRaf34* (*GmRaf34-2*), *GmRaf41* (*GmRaf41-2*), *GmRaf42* (*GmRaf42-3*), *GmRaf49* (*GmRaf49-3*), *GmRaf51* (*GmRaf51-1* and *GmRaf51-2*), *GmZIK1* (*GmZIK1-3*), *GmZIK8* (*GmZIK8-1* and *GmZIK8-3*), and *GmZIK2* (*GmZIK2-1*) are inferred to be undergoing nonfunctionalization (Figs. 6–8). A thorough study of the expression data for these genes is required to understand the evolutionary processes involved in gene/genome duplications in polyploid species such as soybean. We also realized the need for functional genomic and rigorous analyses to test our inferences, but they were beyond the scope of this project.

**MAPK nomenclature in plants**

*Arabidopsis* is the first plant species to have its complete genome sequenced, and also to have its two MAPK gene families systematically named. We used the *Arabidopsis* model for the nomenclature of *GmMAPKs* and *GmMAPKKs*. There is no published literature on the MAPKK nomenclature, except for the identification of putative MAPKKks of *Arabidopsis* and rice. For *GmMAPKKK* nomenclature (see Fig. 5 and Additional file 3), we followed “The Arabidopsis Information Resources” website (TAIR, http://www.arabidopsis.org). Although the nomenclature model presented in the study of *Arabidopsis* MAPK is described as being robust enough to be adopted and expanded to the MAPKs of newly sequenced genomes, the model alone does not seem to capture the paralogous/orthologous status of all MAPK genes. Later in the study of rice and poplar MAPKs, the nomenclature model was redesigned to manifest the evolutionary relationships among the duplicated MAPK genes. MAPK nomenclature adopted by Hamel et al seems to be an appropriate step towards naming MAPK genes that resulted from duplication events including ancient polyploidization. One of these ancient polyploid genomes is *Arabidopsis thaliana* itself. Homology inferred on the basis of phylogenetic placements and sequence identity of MAPKs in *Arabidopsis* revealed several gene members such as *AtMAPK1* and *AtMAPK4*, *AtMAPK11*, *AtMAPK7*, *AtMAPK8* and *AtMAPK15*, *AtMAPK18* and *AtMAPK19* in the MAPK subfamily, as well as *AtMAPKK1* and *AtMAPKK2*, *AtMAPKK4*, *AtMAPKK3*, and *AtMAPK7*; and *AtMAPKK8* and *AtMAPKK9* in the MAPK subfamily to be paralogous. Based on the phylogenetic analyses of the *Arabidopsis* MAPKK protein sequences from soybean, we have assigned paralogous status to numerous genes in soybean, which could be attributable to both recent and ancient duplication events. A comparative genomic analysis of the dataset, including both legume and nonlegume species, was also performed to further confirm the paralogous status of some of these gene members.

We encountered several plant MAPK nomenclatural inconsistencies while searching for nomenclatural codes for naming the identified *GmMAPKs*. In poplar, most of the genes were systematically named using the *Arabidopsis* model. Nonetheless, most of the MAPKs previously studied in rice and grapes were found to be incoherently named, without following the *Arabidopsis* model (Table 2). In the worst scenario, genes of a species with the same names, described by different authors, do not correspond to each other. Oftentimes, the same gene name/number has been used to name different genes as orthologs in different species of prokaryotes, neglecting the fact that the genes thus named could not be automatically orthologs. A common problem is that without knowing all of the several alternative names of MAPK genes, it is difficult to figure out exactly what gene is being presented in the literature. This might be partly due to evolutionary processes including gene and genome duplication or chromosomal rearrangements, complicating the identification and nomenclatural processes. The problem of nomenclature lies not only within the MAPK family, but the lack of allegiance in nomenclature can also be realized to a wider extent from desultory labeling of “p21” from p21ras to p21waf (also known as WAF1,
CIP1, SDII, and CAP20), each with strict functional differences, to the lack of uniformity in nomenclature of regulatory proteins named variously as FLIP, Casper, FLAME, CASH, and I-FLICE. Inconsistent practices of gene nomenclature and a lack of a universal code could potentially reduce the reliability of cross-species functional analysis.77

One of the most challenging tasks is the nomenclature of the MAPKKK subfamily, particularly in the polyploid species (such as soybean). For the purpose of nomenclature of MAPK in plants, including those with duplicated genomes, we suggest the comparison among multiple species to identify the homologs of these genes. Comparative studies including species with a larger number of identified MAPK orthologs, and perhaps the most basal diploid plant species, can facilitate the naming process. If the genome of the species does not have orthologs of a particular gene in Arabidopsis or in previously studied species, a new name should be proposed. It is therefore important to confirm whether the intended names have already been established to any MAPKs in previously studied organisms, and to determine whether the homology model could be applied to the nomenclature of MAPKs of interest. Such homology-based nomenclature, however, is not always the ultimate and error-free method used to name the genes. In such cases, a suit of practices including phylogenetics, protein function analysis, and structural comparisons may be required for the purpose of correct gene nomenclature.

Conclusion
Systematic identification of MAPK genes in soybean is vital to our understanding of their roles in stress response, growth, development, and defense mechanisms. In this study, we have presented genome-wide identification and nomenclature of GmMAPK families. A total of 38 GmMAPKs, 11 GmMAPKKs, and 150 GmMAPKKKs are identified, and our results suggest the expansion of GmMAPK families in soybean due to ancient genome duplications and recent chromosomal rearrangements. Expression profiles based on transcriptomic data showed expression patterns on a continuum between null expressions to a high level of expression in almost all examined tissues under the given experimental conditions. The expression profiles, when mapped onto the phylogeny, provided evidence of strong functional divergence in GmMAPKs. Comparative genomics of legume and nonlegume species and characterization of legume-specific genes using ribonucleic acid interference (RNAi) and protein—protein interaction is underway in the authors’ labs. In addition, our ongoing study on structural conservation, selection pressure, and expression experiment is expected to add another dimension to the current findings. Evolutionary processes driving the expansion of the MAPK families can be better understood through comparative genomics of MAPKs from multiple representative species at various taxonomic levels. Advancement in DNA sequencing technology is yielding a wealth of genome sequence data from an increasing number of taxa, thus making it challenging for the systematic identification and nomenclature of their genes. The results from this study may facilitate functional dissection of the important genes and their communication among scientific communities.

Materials and Methods
Profile hidden Markov model (HMMs)
In order to perform a thorough search for MAPK gene members in soybean, we performed HMM searches based on the multiple sequence alignments of 20 MAPKs, ten MAPKKs, and 80 MAPKKKs of Arabidopsis. The HMM profile was built using HMM version 3.0 (HMMER Project; Howard Hughes Medical Institute, Chevy Chace, MD, USA) in our Linux system to execute homology searches against whole protein dataset from soybean genome available at phytozome.net. The resulted gene models from hmmbuild/hmmsearch options were accepted only if they were within the inclusion threshold of e-value 0.01. The resulted sequences were aligned in ClustalW version 2.0,79 as described previously.13 The MAPK gene models were accepted only if they displayed the consensus for MAPK-specific motifs and domains.2,8

Identification of conserved domains
MEME was used to predict similarities among protein sequences and visualize conserved motifs in specific subdomains of MAPK, MAPKK, and MAPKKK employing two conditions: (1) the ideal motif widths were set to be between six and 50; and (2) the motif...
search was set to identify ten motif regions. We also chose to use a bulk of the protein sequences from each subfamily of MAPK and MAPKK, and from each subgroup of MAPKKK (MEKK-like, Raf-like, and ZIK-like) for each input dataset. This allowed us to predict the higher number of conserved domains by avoiding the maximum amount of noise. The identified genes were confirmed using their signature motifs and published domains. The conserved protein domains of the MAPK gene members were analyzed using the PROSITE protein database, and the domains within each query sequence were analyzed for the presence of serine and threonine kinase residues using Pfam (pfam.sanger.ac.uk) and SMART (www.smart.embl-heidelberg.de). We also confirmed the presence of residues required for kinase activity in the catalytic loop.

Phylogenetic analysis of GmMAPK families

The MAPK and MAPKK protein sequences for rice, poplar, and Arabidopsis (TAIR—http://www.arabidopsis.org) were directly adopted from the published data sources. The MAPK protein dataset was aligned using ClustalW. The aligned amino acid sequences were manually edited using the program Se-AL (v2.0a11 Carbon; http://tree.bio.ed.ac.uk/software/seal) in order to avoid errors related to residue misplacements due to highly variable indel sizes. The data matrices were analyzed using the MP method in the program PAUP* 4.0b10. Heuristic searches were performed treating all characters equally weighted and unordered. Portions of the data matrices with ambiguous alignments were excluded from the analysis. Data matrices were also analyzed using the ML method using the best evolutionary model in the program, MEGA5.2.2. Branch supports were computed using bootstrapping of 2,000 and 100 replicates for MP and ML analyses, respectively. Human MAPK sequences, HsMAPK1 (GenBank: NP_002736.3), HsMAPKK1 (GenBank: AA137460.1), HsMAPKKK10 (GenBank: NP_002437.2) were used as out-groups for the analyses of MAPK, MAPKK, and MAPKKK families, respectively. Gene—gene relationships were examined based on phylogenetic tree, genetic distance for each member of all three MAPK subfamilies, and their sequence identities (Additional file 5).

Homology assessment, nomenclature, and transcriptomic data

A sequence identity greater than 40% was used to detect homology among the protein sequences. Homology confidence among the protein sequences was established using sequence identity (40% or higher identity), ML estimates of pairwise genetic distances (smallest genetic distance among the potential homologs), and phylogenetic placements to define paralogous status of the duplicated genes and also to assign a number to the identified soybean MAPK genes. Orthologous and paralogous relationships were also inferred through both parametric (ML) and nonparametric (MP) methods. Available RNAseq expression data were downloaded using RNAseq atlas of soybean (www.soybase.org). Arabidopsis MAPK expression data were also downloaded from the MPSS database (http://mpss.udel.edu/; accessed May 12th, 2012), which was the database used for comparative analysis. Mayday workbench was used to create a heatmap visualization for the expression data of soybean and Arabidopsis MAPK genes. For heatmap visualization, normalized gene expression at different growth conditions, tissues, and at different stress levels were log 2-based to explore the functional divergence across the soybean MAPK genes and their orthologs in Arabidopsis.

Acknowledgements

Bibek Koirala, Spencer Schreier, Kenton MacArthur, and Kevin Murray provided constructive criticism on the manuscript.

Author’s Contributions

AN carried out data mining, performed in silico and phylogenetic analyses, investigated the gene family structure and their functional and evolutionary relationships through comparative genomics, and drafted the manuscript. MPN and JSR together conceived the original project, MPN designed and coordinated the project, supervised data analyses, and contributed on drafting the manuscript. SP helped in data mining, phylogenetic analysis, and contributed helpful discussion on the genomic analyses. SS contributed in the data analysis and contributed helpful discussions. JSR helped design the original projects and contributed helpful discussions. RNR contributed helpful discussion and technical consultation to reshape the
manuscript. BVB helped in phylogenetic analysis and provided constructive criticism on the manuscripts. All authors reviewed and approved the final manuscript.

**Funding**

This research was supported by South Dakota Soybean Research and Promotion Council (SDSRPC), Center for Excellence in Drought Tolerance Research (CEDTR), South Dakota Agricultural Experiment Station (SDAES), and Faculty Start up fund to MPN from the Department of Biology and Microbiology at South Dakota State University.

**Competing Interests**

Author(s) disclose no potential conflicts of interest.

**Disclosures and Ethics**

As a requirement of publication the authors have provided signed confirmation of their compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

**References**

1. Rohila JS, Yang Y. Rice mitogen-activated protein kinase gene family and its role in biotic and abiotic stress response. *J Integr Plant Biol*. 2007;49(6):751–9.
2. Jonak C, Okrész L, Bögre L, Hirt H. Complexity, cross talk and integration in biotic and abiotic stress response. *Curr Opin Plant Biol*. 2002;5(5):415–24.
3. Rodriguez MC, Petersen M, Mundy J. Mitogen-activated protein kinase signaling in plants. *Annu Rev Plant Biol*. 2010;61:621–49.
4. Barabási AL, Oltvai ZN. Network biology: understanding the cell’s functional organization. *Nat Rev Genet*. 2004;5(2):101–13.
5. Huang GT, Ma SL, Bai LP, et al. Signal transduction during cold, salt, and drought stresses in plants. *Mol Biol Rep*. 2012;39(2):969–87.
6. Madhani HD, Fink GR. The riddle of MAP kinase signaling specificity. *Trends Genet*. 1998;14(4):151–5.
7. Widmann C, Gibson S, Jarpe MB, Johnson GL. Mitogen-activated protein kinase: conservation of a three-kinase module from yeast to human. *Physiol Rev*. 1999;79(1):143–80.
8. MAPK Group. Mitogen-activated protein kinase cascades in plants: a new nomenclature. *Trends Plant Sci*. 2002;7(7):301–8.
9. Tena G, Asai T, Chua WL, Sheen J. Plant mitogen-activated protein kinase signaling cascades. *Curr Opin Plant Biol*. 2001;4(5):392–400.
10. Sturgill TW, Ray LB. Muscle proteins related to microtubule-associated protein-2 are substrates for an insulin-stimulatable kinase. *Biochem Biophys Res Commun*. 1986;134(2):565–71.
Berlin, NY: Springer-Verlag; 2006;18(5):57. Madhusudan, Trafny EA, Xuong NH, et al. cAMP-dependent protein kinase:...

Liu X, Wang H, Tang Y, et al. Preparation of single rice chromosome for...

Kishimoto N, Higo H, Abe K, Arai S, Saito A, Higo K. Identification of the...

Lazcano A, Miller SL. How long did it take for life to begin and evolve to...

International Rice Genome Sequencing Project. The map-based sequence of...

Tuskan GA, Difazio S, Jansson S, et al. The genome of black cottonwood, Populus trichocarpa...

Zhang J. Evolution by gene duplication: an update.

Lynch M, Conery JS. The evolutionary demography of duplicate genes.

Gill N, Findley S, Walling JG, et al. Molecular and chromosomal evidence for allopolyploidy in soybean. Plant Physiol. 2009;151(3):1167–74.

Shoemaker RC, Polzin K, Labate J, et al. Genome duplication in soybean (Glycine subgenus soja). Genetics. 1996;144(1):329–38.

Lynch M, Conery JS. The evolutionary demography of duplicate genes. J Struct Funct Genomics. 2003;3(1–4):35–44.

Zhang J. Evolution by gene duplication: an update. Trends Ecol Evol. 2003;18(6):292–298.

Arumuganathan K, Earle ED. Nuclear DNA content of some important plant species. Plant Mol Biol Rep. 1999;17(3):208–18.

Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000;408(6814):796–815.

Jailon O, Aury JM, Noel B, et al. French-Italian Public Consortium for Grapevine Genome Characterization. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007;449(7161):463–7.

Tuskan GA, Difazio S, Jansson S, et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science. 2006;313(5793):1594–604.

International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature. 2005;436(7052):793–800.

Lazcano A, Miller SL. How long did it take for life to begin and evolve to cyanobacteria? J Mol Evol. 1994;39(6):546–54.

Kishimoto N, Higo H, Abe K, Arai S, Saito A, Higo K. Identification of the duplicated segments in rice chromosomes 1 and 5 by linkage analysis of cDNA markers of known functions. Theor Appl Genet. 1994;88:722–6.

Throude M, Bolot S, Bosio M, et al. Structure and expression analysis of rice paleo duplications. Nucleic Acids Res. 2009;37(4):1248–59.

Liu X, Wang H, Tang Y, et al. Preparation of single rice chromosome for construction of a DNA library using a laser microbeam trap. J Biotechnol. 2004;109(3):217–26.

Hyun TK, Kim JS, Kwon SY, Kim SH. Comparative genomic analysis of mitogen activated protein kinase gene family in grapevine. Genes & Genomics. 2010;32:275–81.

Madhusudan, Traffy EA, Xuong NH, et al. CAMP-dependent protein kinase: crystalllographic insights into substrate recognition and phosphorylation.

Protein Sci. 1994;3(2):176–87.

Kannan N, Neuwald AF. Evolutionary constraints associated with functional specificity of the CMGC protein kinases MAPK, CDK, GSK, SRPK, DYRK, and CK2alpha. Protein Sci. 2004;13(8):2059–77.

Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000;290(5494):1151–5.

Prince VE, Pickett FB. Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet. 2002;3(11):827–37.

Duarte JM, Cui L, Wall PK, et al. Expression pattern shifts following duplication indicative of subfunctionalization and nonfunctionalization in regulatory genes of Arabidopsis. Mol Biol Evol. 2006;23(2):469–78.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 1999;151(4):151–45.

Ohno S. Evolution by Gene Duplication. Berlin, NY: Springer-Verlag; 1970.

Adams KL, Wendel JF. Exploring the genomic mysteries of polyploidy in cotton. Biological Journal of the Linnean Society. 2004;82(4):573–81.

Blanc G, Wolfe KH. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell. 2004;16(7):1679–91.

Horiuchi J, Prithiviraj B, Bais HP, Kimball BA, Vivanco JM. Soil nema-todes mediate positive interactions between legume plants and rhizobium bacteria. Planta. 2005;222(5):848–57.

Abada EA, Sung H, Dwivedi M, Park BJ, Lee SK, Ahn J. C. elegans behavior of preference choice on bacterial food. Mol Cells. 2009;28(3):209–13.

Coolon JD, Jones KL, Todd TC, Carr BC, Herman MA. Caenorhabditis elegans genomic response to soil bacteria predicts environment-specific genetic effects on life history traits. PLoS Genet. 2009;5(6):e1000503.

Lee JS, Huh KW, Bhargava A, Ellis BE. Comprehensive analysis of protein-protein interactions between Arabidopsis MAPKs and MAP kinase helps define potential MAPK signalling modules. Plant Signal Behav. 2008;3(12):1037–41.

Ding X, Richter T, Chen M, et al. A rice kinase-protein interaction map. Plant Physiol. 2009;149(3):1478–92.

Jung KH, Cao P, Seo YS, Dardick C, Ronald PC. The Rice Kinase Phylogenomes Database: a guide for systematic analysis of the rice kinase superfamily. Trends Plant Sci. 2010;15(11):595–9.

Kishi-Kaboshi M, Takahashi A, Hirochika H. MAMP-responsive MAPK cascade regulates phytoalexin biosynthesis. Plant Signal Behav. 2010;5(12):1653–5.

Suarez-Rodriguez MC, Adams-Phillips L, Liu Y, et al. MEKK1 is required for flg22-induced MPK4 activation in Arabidopsis plants. Plant Physiol. 2007;143(2):661–9.

Colcombet J, Hirt H. Arabidopsis MAPKs: a complex signalling network involved in multiple biological processes. Biochem J. 2008;413(2):217–26.

Schraeke ME, Mitchell-Olds T. Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. Plant Cell. 2008;18(5):1152–65.

Agraval GK, Rakwal R, Iwashashi H. Isolation of novel rice (Oryza sativa L.) multiple stress responsive MAP kinase gene, OsMSRMK2, whose mRNA accumulates rapidly in response to environmental cues. Biochem Biophys Res Commun. 2002;294(5):1099–16.

Lesk AM, Parkinson H, Whisstock JC. Classification of protein function. In: Lesk AM, editor. Database Annotation in Molecular Biology. Hoboken, NJ: John Wiley & Sons; 2005.167–80.

Obstacles of nomenclature. Nature. 1997;389(6664):1.

Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23(21):2947–8.

Bailey TL, Elkan C. The value of prior knowledge in discovering motifs with MEME. Proc Int Conf Intell Syst Mol Biol. 1995;3:21–9.

Hamel LP, Nicole MC, Duplessis S, Ellis BE. Mitogen-activated protein kinase signaling in plant-interacting fungi: distinct messages from conserved messengers. Plant Cell. 2012;24(4):1327–51.

Sigrist CJ, Cerutti L, de Castro E, et al. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. 2010;38(Database issue):D161–6.
Mitogen-activated protein kinase (MAPK) genes in soybean

83. Swofford DL. PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods). Version 4. Sunderland, MA: Sinauer Associates; 2000.
84. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011;28(10):2731–9.
85. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. 1985;39(4):783–91.
86. Petsko GA, Ringe D. Protein Structure and Function. Oxford, UK: New Science Press; 2004.
87. Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999;12(2):85–94.
88. Lesk AM. Introduction to Protein Architecture: The Structural Biology of Proteins. Oxford, NY: Oxford University Press; 2001.
89. Severin AJ, Woody JL, Bolon YT, et al. RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome. BMC Plant Biol. 2010;10:160.
90. Nakano M, Nobuta K, Vernaraju K, Tej SS, Skogen JW, Meyers BC. Plant MPSS databases: signature-based resources for analyses of mRNA and small RNA. Nucleic Acids Res. 2006;34(Database issue): D731–5.
91. Battke F, Symons S, Nieselt K. Mayday—integrative analytics for expression data. BMC Bioinformatics. 2010;11:121.
92. Krysan PJ, Jester PJ, Gottwald JR, Sussman MR. An integrative pipeline to identify and characterize novel plant signal transduction pathways. MAPK phosphatase 3 regulates Arabidopsis seedling development. J Biol Chem. 2001;276(21):18139–45.
93. Calderoni O, Glib A, Bergouinioux C, Heberle-Bors E, Wilson C. A novel tobacco mitogen-activated protein (MAP) kinase, NIM1, is required to activate the cell cycle-regulated p43 NtHt MAP kinase. J Biol Chem. 2001;276(21):18139–45.
94. de los Poo O, Pedley KF, Martin GB. MAPKKKalpha is a positive regulator of cell death associated with both plant immunity and disease. EMBO J. 2004;23(15):3072–82.
95. Ekengren SK, Lin Y, Schiff M, Dinesh-Kumar SP, Martin GB. Two MAPK cascades, NPR1, and TGA transcription factors play a role in Pto-mediated disease resistance in tomato. Plant J. 2003;36(6):905–17.
96. Nakagami H, Pitzschke A, Hirt H. Emerging MAP kinase pathways in plant stress signalling. FEBS Lett. 2007;581(9):1834–40.
97. Ortiz-Masía D, Perez-Amador MA, Carbonell J, Marcote MJ. Diverse stress signals activate the C1 subgroup MAP kinases of Arabidopsis. FEBS Lett. 2008;581(9):1834–40.
98. Barba-Espin G, Diaz-Vivancos P, Job D, Belghazi M, Job C, Hernández JA. Understanding the role of H(2)O(2) during pea seed germination: a combined proteomic and hormone profiling approach. Plant Cell Environ. 2011;34(11):1907–19.
99. Eckardt NA. Induction of phytoalexin biosynthesis: WRKY33 is a target of MAPK signaling. Plant Cell. 2011;23(4):1190.
100. Bögre L, Calderini O, Binarova P, et al. A MAP kinase is activated late in plant mitosis and becomes localized to the plane of cell division. Plant Cell. 1999;11(1):101–13.
101. Brodersen P, Petersen M, Bjørn Nielsen H, et al. Arabidopsis MAP kinase 4 regulates salicylic acid- and jasmonic acid/ethylene-dependent responses via EDS1 and PAD4. Plant J. 2006;47(4):532–46.
102. Shi J, HL, Zhang L, Gao Z, Guo QX. GhMPK7, a novel multiple stress-responsive cotton group C MAPK gene, has a role in broad spectrum disease resistance and plant development. Plant Mol Biol. 2010;74(1–2):1–17.
103. Lee JS, Wang S, Sritubtim S, Chen JG, Ellis BE. Arabidopsis mitogen-activated protein kinase MPK12 interacts with the MAPK phosphatase IDR5 and regulates auxin signaling. Plant J. 2009;57(6):975–85.
104. Walia A, Lee JS, Wasteneys G, Ellis B. Arabidopsis mitogen-activated protein kinase MPK18 mediates cortical microtubule functions in plant cells. Plant J. 2009;59(4):565–75.
Supplementary Data

Additional file 1: maximum parsimony analysis of GmMAPKs and their orthologs in Arabidopsis, poplar, and rice

The values above the branches are bootstrap support of 2,000 replicates. The members with the phosphorylation motif TEY are included in clades A, B, and C, TDY in clade D, and the members with the TQY (denoted by *) and TVY (denoted by **) motif in clades E and B, respectively. The MAPK gene models were accepted for phylogenetic analysis using protein sequences of the serine/threonine kinase subfamily having conserved aspartate and lysine residues in their catalytic domain with the (D[L/I/V]K) motif and TXY phosphorylation motif in their activation loop.

Additional file 2: maximum parsimony analysis of GmMAPKKs and their orthologs in Arabidopsis, poplar, and rice

The values above the branches are bootstrap support of 2,000 replicates. The MAPKK gene models were accepted for phylogenetic analysis using dual-specificity protein kinases having conserved aspartate and lysine residues in their catalytic domain with the (D[L/I/V]K) motif and the S-X_5^-T phosphorylation motif along their activation loop.

Additional file 3: maximum parsimony analysis of GmMAPKKKs and their orthologs in Arabidopsis

Phylogenetic representation of GmMAPKKKs in circular tree format shows the three subgroups: GmMEKK-like, GmRaf-like, and GmZIK-like, as labeled. The values above the branches are bootstrap support of 2,000 replicates. The MAPKKK gene models were accepted for phylogenetic analysis using protein sequences of the serine/threonine kinase subfamily having conserved aspartate and lysine residues in their catalytic domain with the (D[L/I/V]K) motif, and the members in each subgroup were confirmed based on their signature motifs.

Additional file 4: chromosomal distribution of GmMAPKs, GmMAPKKs, and GmMAPKKKs

Distribution of gene members of the GmMAPK family in 20 sets of soybean chromosomes.

Additional file 5: plant MAPK nomenclature

MAPK gene nomenclature and homology assessment based on sequence identity, genetic distance, and phylogenetic placement.