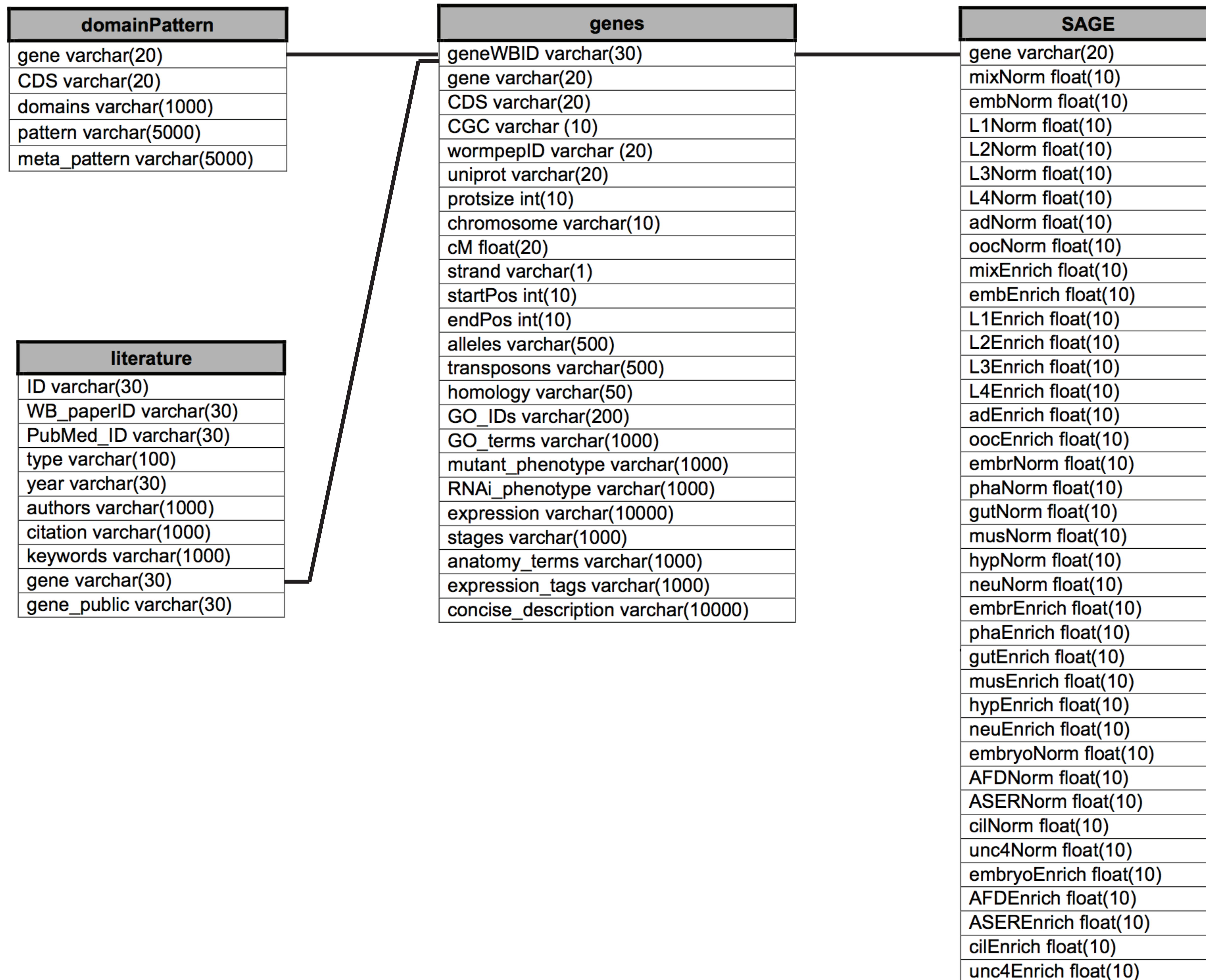


GExplore database schema



Supplementary data: Description of database fields

Table: genes (main table)

Field	Content
geneWBID	Wormbase gene ID (e.g. WBGene00016354)
gene	gene name (e.g. C33F10.5)
CDS	protein identifier (e.g. C33F10.5b)
CGC	CGC gene name (e.g. rig-6)
wormpepID	wormpep protein ID (e.g. CE25803)
uniprot	Uniprot identifier (e.g. Q9BIA2)
protsize	protein size
chromosome	chromosome
cM	genetic position
strand	strand orientation
startPos	physical position of gene start
endPos	physical position of gene end
alleles	all alleles
transposons	transposon insertion in the gene
homology	condensed homology information
GO_IDs	GO term identifiers
GO_terms	GO terms
mutant_phenotype	all mutant phenotypes
RNAi_phenotype	all RNAi phenotypes
expression	combined expression information
stages	stages of expression
anatomy_terms	anatomy terms associated with expression
expression_tags	expression tags generated from microarray data sets
concise_description	concise gene description (see wormbase data field)

Table: literature

Field	Content
ID	unique number
WB_paperID	Wormbase identifier (e.g. WBPaper00000031)
PubMed_ID	PubMed identifier (e.g. 4366476)
type	type of publication (e.g. ARTICLE)
year	year of publication
authors	list of authors
citation	citation (as in wormbase)
keywords	keywords associated with the publication
gene*	gene associated with the publication (e.g. F41C6.1)
gene_public	CGC gene name (e.g. unc-6)

*note: this field contains only one gene name. If a publication is associated with several genes, separate entries for each gene will be generated.

Table: SAGE

Field*	Content
gene	gene name (e.g. C33F10.5)
mix, emb, L1, L2, L3, L4, ad, ooc	SAGE data for the different stages
embr, pha, gut, mus, hyp, neu	SAGE data for the different tissues
embryo, AFD, ASER, cil, unc4	SAGE data for different neuronal subpopulation

*note: fields ending in 'Norm' contains normalized SAGE tags, fields ending in 'Enrich' contain the enrichment factor with respect to the reference library

Table: domainPattern

Field	Content
gene	gene name (e.g. C33F10.5)
CDS	protein identifier (e.g. C33F10.5b)
domains	all domains found in the protein
pattern	sequence of domains in the protein (excluding metapattern*)
meta_pattern	sequence of domains in the protein

*note: currently not used. Could be used to remove metadomains from the display to show the underlying domains (like transmembrane domains)