GeneSpider User Manual

Release Date: 17 April 2000



Copyright 2000 Silicon Genetics. All rights reserved. GeneSpring, GeneSpider, GenEx, and MicroSift are trademarks of Silicon Genetics. All other products, including but not limited to GeneBank, Microsoft Excel, Microsoft Notepad and Adobe FrameMaker, are the trademarks of their respective holders.

Table of Contents

Chapter 1 : The GeneSpider	
1.1 What does the GeneSpider do?	3
1.2 What data do you need to use the GeneSpider?	3
1.2.1 The GeneSpider file format	3
1.3 Running the GeneSpider	6
1.3.1 To add or update the information associated with each GenBank	
accession number	6

Chapter 1: The GeneSpider

1.1 What does the GeneSpider do?

Given a list of GenBank accession numbers the GeneSpider searches either GenBank or Locus Link for information associated with those accession numbers and caches it for your future use. Searching Locus Link is currently only useful for human genes.

1.2 What data do you need to use the GeneSpider?

To use the GeneSpider you need a list of GenBank accession numbers (these are also known as GenBank identifiers or the GenBank locus). In addition, you may also have alternative names, functional information, map positions, EC numbers, and so on associated with each gene. This information should be saved in the same file, using GeneSpider file format described below. Hereafter the file containing this information will be referred to as the gene list file. The gene list file must be a tab-delineated text file. (When saving their gene list files, Windows users should look for the document type "Text(Tab delimited)(*.txt)" in the "Save as type:" dialog box.) The gene list file may be created in a spreadsheet program, such as Microsoft Excel, if it is saved as a tab-delineated text file.

1.2.1 <u>The GeneSpider file format</u>

The GeneSpider file format is a tab delimited text file consisting of one line per gene, with several fields separated by tabs. The first field (systematic name) must be included for every gene; the other nine fields are optional, but must be entered in the order presented here.

- 1. Systematic Name: The normal way of referring to this gene. This name must be unique and every gene must have one. Frequently the name used in this column is the gene's GenBank accession number, otherwise it may be the name which labels the gene's raw signal strength values in your array (or other experimental) data files, it may be the gene's location on the array, or it may be any other unique identifier you wish associated with each gene. This column must be included in your gene list file, and it must be filled for every gene.
- 2. Common Name: An alternative way of referring to this gene. Genes are not required to have a common name, and common names do not have to be unique. This column may contain the GenBank accession numbers.
- **3. Map:** Mapping information for this gene. (For example, 16q12.1 or 123...358 if the full nucleotide sequence is known for your organism.)
- 4. EC number: The EC number for this gene.
- 5. **Description:** A description of this gene.

The GeneSpider

- 6. **Product:** What this gene produces.
- 7. **Phenotype:** A description of the phenotype for this gene.
- 8. Function: A description of the function of this gene.
- 9. Keywords: Keywords associated with this gene.
- 10. GenBank locus: The GenBank accession number for this gene. If the GenBank identifiers for your genes were not used as either their systematic or common names, then they must be included in this field.

When creating your gene list file, these ten fields should be entered in the order they are listed here. Remember to include any blank fields in their appropriate columns. The gene's systematic name should always be in the first column, its common name is in the second, and its mapping information in the third column, even if the second column is completely blank because there are no common names for any of your genes. Frequently only the first field (systematic name) is used in the gene list file, in this case it must contain the genes' GenBank accession numbers. If you do this your file will be similar to the one illustrated in Figure 3 (page 7). Otherwise your gene list file should resemble the file illustrated in Figure 1 and Figure 2. Entries with spaces in them, such as "Gene 1" are perfectly acceptable. Each field must be separated from the next one by a tab character. You do not need to have information about every gene. In the example below nothing is known about Gene 14, so the line after its name is left blank. If you have a list of genes and text information about them in a spreadsheet formatted as ten columns with one row per gene, simply save this file as a tab-delineated text file. A note about the figures: do not include the titles of the fields in your gene list file; titles are included here only for clarity.

The GenBank accession numbers are the only information absolutely necessary to include in the gene list file. They may be included as the systematic name, the common name, or as the GenBank locus. All of the GenBank accession numbers for your genes must be included in the same column. If that column is either the systematic name column or the common name column then not every entry in that column must be a GenBank accession number. The GeneSpider will search GenBank for every name given in the column containing the GenBank accession numbers; if it does not find anything, it will copy any information you already had about that gene into the new file it is creating and go on to the next gene.

Systematic Name	Common Name	Мар	EC Number	Description	Product	Phenotype	Function	Keywords	GenBank locus
gene 1	luck1		1.1.1.1	gene somehow causes rats to be very lucky	protein A				g763402
gene 2		16q21.2	1.1.1.2		co-produc	es protein E	3		g764509
gene 3			1.1.1.3						g587439
gene 4			1.1.1.4			deletion ca	uses immort	ality	g093285
gene 5			1.1.1.5						g460389
gene 6	charm5	15q42.3	1.1.1.6	rats with this gene are very cute	protein C			cell cycle	g932509
gene 7			1.1.1.7						g328506
gene 8		9q11.0	1.1.1.8		protein D	possibly			g234876
gene 9	beauty3	16q14.1	1.1.1.9				involved in metabolism	metabolism	1
gene 9.5		19q76.7	1.1.2.0		protein D	possibly			g239857
gene 10			1.1.2.1						g238456
gene 10.2						possibly a mutation			
gene 11			1.1.2.2						g239845
gene 12	weird2	16q44.2	1.1.2.3	rats with this gene have two tails	protein E		involved in DNA synthesis	DNA synthesis	g290030
gene 13		16q87.9	1.1.2.4		protein F				g321197
gene 14									

Figure 1 Example of what the GeneSpider format looks like in Excel.

Systemati	ic Name	Common Name	Map	EC Number	Descrip	otion Proo	duct Phenoty	pe
	Function	nKeywords	GenBa	nk locus				
gene 1 l	luck1	1.1.1.1	gene so	omehow causes ra	ts to be ve	ery lucky prot	tein A	
		g763402						
gene 2		16q21.2 1.1.1.2		co-produces pro	otein B			g764509
gene 3		1.1.1.3				g58	7439	
gene 4		1.1.1.4		deletic	on causes i	mmortality		
Į	g093285	5						
gene 5		1.1.1.5				g46	0389	
gene 6	charm5	15q42.3 1.1.1.6	rats wi	th this gene are v	ery cute	protein C		
	cell cycle	e g932509)					
gene 7		1.1.1.7				g32	8506	
gene 8		9q11.0 1.1.1.8		protein D possi	bly			g234876
gene 9 l	beauty3	16q14.1 1.1.1.9			involve	d in metabolis	m metabol	ism
gene 9.5		19q76.7 1.1.2.0		protein D possi	bly			g239857
gene 10		1.1.2.1				g23	8456	
gene 10.2	2				possibly	y a mutation		
gene 11		1.1.2.2				g23	9845	
gene 12	weird2	16q44.2 1.1.2.3	rats wi	th this gene have	two tails	protein E		involved
in DNA s	synthesis	sDNA synthesis	g29003	30				
gene 13		16q87.9 1.1.2.4		protein F			g32119′	7
gene 14								

Figure 2 Example of the same GeneSpider file shown in Figure 1, saved as a tab-delimited text file.

1.3 Running the GeneSpider

The accession numbers must be included in the gene list file in one of three columns: as the genes' systematic name, common name, or as the entry in the GenBank locus field. Do not worry if your gene list file includes genes without GenBank accession numbers. When you update a gene list file using the GeneSpider, the Spider copies your current gene list file so you do not loose any information regarding non-GenBank genes. After the GeneSpider copies this file, it updates the fields of the copied file associated with a GenBank accession number. While it is copying and updating you will see changes in the GeneSpider window reflecting the information it has processed. After the gene list file has been updated you are given the option to permanently save the updated information. When you save this information, it is saved in the GeneSpider format (tab-delimited text file) in the same directory as your original gene list file. You may then open this text file using a spreadsheet program, such as Microsoft Excel®. The updated file is saved in the GeneSpider format so the next time you wish to update the information associated with your set of genes you can use this new file as the initial gene list file.

1.3.1 <u>To add or update the information associated with each GenBank accession</u> <u>number</u>

1. Create your gene list file.

Z97181.1
AL022401
AC004386
AC004388
Z98950
AC004478
AC003666
AC002549
AC003669
Z82204
AC004383
AC004072
AC003683
AC003658
Y15994
AL009175
AC003037
AL008713
AC002422
AC002523
M22332
AQ409366
AQ356884
AQ309743
X61295
AQ572229
AI683867
AQ390430
AQ355719

The GeneSpider

AQ559819			
AQ557343			
AQ573089			
M54985			
AQ549999			
AI421777			
AQ420901			
AQ382430			
AQ440210			
AQ536099			
AI475350			
S67068			
AI821169			
AL045241			
AQ357079			
S80119			
X61294			
AQ377979			
AQ342069			
K02590			
AQ545809			
Z96215			
AQ572877			
AQ554929			
AQ545915			
AQ344044			
AQ378406			
U70924			
Z78996			
J00338			
AQ547459			

Figure 3 Example GeneSpider gene list file. The only column required in the gene list file is systematic name column if it contains the GenBank accession number for each gene. This is the simplest gene list file to create.

2. Click the GeneSpider icon to bring up the first GeneSpider window.



Figure 4 The GeneSpider icon

SeneSpider (c)) 1999 Silicon Genetics	_ 🗆 🗡
	GeneSpider allows you to extract some information from GenBank for a series of genes. You must give GeneSpider a file containing a list of the genes. Please enter the filename below.	
File Name		Browse
Database to search	GenBank	•
	OK Cancel Help	

Figure 5 The initial GeneSpider window

- 3. In the first box, labeled "File Name", enter the complete file name and the pathway of your gene list file. If you are a windows user be sure to include the .txt suffix. To enter this information either write the complete directory pathway in the "File Name" box or:
 - a. Click the "Browse" button. A browse window appears.

Browne		2 X	Browne	2 X
Laok ja: 📴 da	8		Lask in: 🔁 Human	· • • •
arabidopsis Bacteria Cache Com E.colie	Hutan Human Human Human Human House House House House House House House	Visues peast ElensSpring Prelever ElenesPFINGDE.G genomes.tut ElecalNumber.tut	Cache B Genefikanes.txt B Genefikanes2.txt Phanan penomedel	
-				
File pane:		<u>Open</u>	File parter Distribution	<u>Open</u>
Free or gipter Juli He	s(".")	Carcal	Last or Obs [1711-461 ()	Carcal

Figure 6 The "Browse" directory

- **b.** Find your gene list file.
- c. Select your gene list file by clicking it. This will enter the gene list file name in the "File name" box of the "Browse" window.
- d. Click the "Open" button. This writes the complete file name and pathway in the "File Name" box in the GeneSpider window.

CeneSpider (c) 1999 Silicon Genetics	- 🗆 ×
GeneSpider allows you to extract some information from GenBank for a series of genes. You must give GeneSpider a file containing a list of the genes. Please enter the filename below.	
File Name D:\GeneSpider\Genes.txt	Browse
Database to search GenBank	-
OK Cancel Help	

Figure 7 The initial GeneSpider window, with a file indicated in the "File Name" box.

4. Click the arrow at the right hand corner of the box labeled "Database to search". This opens a pulldown menu.

GenBank		
Locus Link		

Figure 8 The "Databases to search" menu

- 5. Choose which database you wish to search by clicking it. The name of the database you will be searching should now be in the "Database to search" box.
- 6. Click the "OK" button. The GeneSpider window will change:

👷 Update Genome Genes.txt by GenBank 📃 🗖 🗙
Options
Use for GenBank identifier Column 1 (Systematic Name) 💌
Cverwrite existing information
Status
Processed: 0
Found: 0
Enhanced: 0
To Go: 0
Waiting to start
Save gene list to D:\GeneSpider\Genes.txt
Start Stop Save and close

Figure 9 The second GeneSpider window, the "Update Genome from GenBank" window.

7. Click the arrow to the right of the box labeled "Use for GenBank identifier". This opens a pulldown menu.

Column 10 (GenBank Locus)
Column 1 (Systematic Name)
Column 2 (Common Name)

Figure 10 The "Use for GenBank identifier" menu

8. Click the column in your gene list file containing the GenBank accession numbers.

The GeneSpider

- 9. If you have information beyond the GenBank accession number in your gene list file the "Overwrite existing information" checkbox pertains to you. If you select the checkbox you tell the Spider to overwrite any information already in your gene list file if it finds similar information for that gene on the web. If you do not select in the "Overwrite existing information" checkbox any information about the genes already in your gene list file will not be modified or updated by information from the web.
- 10. Click the "Start" button. The Spider will process the data from the web, displaying how far it has gotten in the box labeled "Status". This search may take awhile. When you are updating long lists of genes it is better to leave this process running overnight, as it may take a few hours. Messages saying the Spider is caching information may appear during this process.

🏽 Update Genome Genes.txt by GenBank 📃 🗖 🗙
Options
Use for GenBank identifier 🛛 Column 1 (Systematic Name) 💌
Overwrite existing information
Status
Processed: 5
Found: 4
Enhanced: 4
To Go: 56
Processing gene Z98950
Save gene list to D:\GeneSpider\Genes.txt
Start Stop Save and close

Figure 11 The "Update Genome from GenBank" window during the search of GenBank.

Update Genome Genes.txt by GenBank								
Use for GenBank identifier Column 1 (Systematic Name)								
Status								
Processed : 61								
Found: 60								
Enhanced : 59								
To Go: 0								
Done but not Saved								
Save gene list to D:\GeneSpider\Genes.txt								
Start Stop Save and close								

Figure 12 The "Update Genome from GenBank" window after the GenBank search is complete

- 11. Write the name you would like the updated gene list saved as in the box labeled "Save gene list to". If you are a Windows user, remember to include the .txt suffix. The updated gene list file will be saved in the same directory as the original gene list file, unless you type a different pathway in this box.
- 12. Click the "Save and close" button to save the updated gene list file. This also closes the GeneSpider.
- The file resulting from the search illustrated in this document is illustrated in Figure 13 and Figure 14.

Z97181	HTG; DXS7; GT repeat							
polymorphism; GTG repeat polymorphism								
AL022401 CHM	5' part of gene beyond this clone; match: proteins P24386							
P37727 P26374 dJ93L7.1 (RAB Escort p.	rotein 1 (REP-1, RAB proteins geranylgeranyltransferase							
component A 1, Choroideraemia protein, 7	Гареtochoroidal Dystrophy (TCD) protein)							
HTG; CHM; Choroideraemia; gen	ranylgeranyltransferase component A 1; RAB Escort; REP-1;							
REP1; Tapetochoroidal Dystrophy; TCD								
AC004386	HTG							
AC004388	HTG							
Z98950 dJ507I15.1	match: multiple proteins; match: CE02123 P90702 Q96499							
P10661 P65027; match: P09896 P31866 P	P10661 P65027; match: P09896 P31866 P02405 P31028 P52809; match: Q00477 Q00494 P49213 P17843							
P27076; match: cDNAs M19635 M15661 AB000910; match: multiple ESTs; match: T87328 T87321								
AA181201 T41136; match: AA244162 R0	5264 N93353 AA191627; match: AA411822 AA328207							
AA342359 T89286; 60S ribosomal protein	n L44 (L41, L36) like 60S							
ribosomal; L36; L41; L44; Xq26.3-27.3								
AC004478	HTG							
AC003666	HTG							
AC002549	HTG							
AC003669	HTG							
Z82204	repeat polymorphism; X							
AC004383	HTG							

AC004072			HTG						
AC003683			HTG						
AC003658			HTG						
Y15994		MTM1	gene						
AI 009175 RFP1	match	SW P24386 FMBL X781	$21 \cdot (RAR)$	FSCORT					
DECOMPLY AND ADDITION OF A CONTRACT ADDITION ADDITIONA AD									
PROTEINS GEDANVI GEDANVI	PROTEINS (CEPANYL CEPANYL TRANSFERACE COMPONENT A 1)								
PROTEINS GERAINTLOERAINT	lananyi teonofor	$\sum_{n=1}^{\infty} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \sum_{i$							
A CO02027	rgeranyi transfer	ase; Aq21.1-Aq21.5	UTC						
AC005057		5 . 1. 1 1	ПIU M20212	1102002 1					
AL008/13	steroid	5-aipna-reductase; match:	M32313	dJ95C25.1					
3-oxo-5-alpha-ste	eroid delta(4)-del	nydrogenase; dinydrotestos	sterone; p	seudogene; X					
AC002422			HIG						
AC002523			HIG						
M22332	ORF; putative	unknown protein		L1 insertion					
element									
AQ409366			GSS						
AQ356884			GSS						
AQ309743			GSS						
X61295	L1 retroposon, a	a portion of its ORF2 sequ	ence						
L1 retroposon; reverse tra	nscriptase-like p	rotein							
AQ572229			GSS						
AI683867			EST						
AQ390430			GSS						
AQ355719			GSS						
AQ559819			GSS						
AO557343			GSS						
AO573089			GSS						
M54985		nsi-eta	beta-like	globin pseudogene					
A0549999		por eta	GSS	groom poeudogene					
AI421777			EST						
A0420901			GSS						
A0382430			GSS						
AQ302430			G85						
AQ440210 AQ536000			GSS						
AU75250			USS EST						
AI4/3330			ESI						
50/008			DOT						
A1821109			ESI						
AL045241			ESI						
AQ357079			GSS						
S80119 reverse transcriptase home	olog	This sequence of	comes fro	m Fig.2. Protein					
sequence is in conflict with the con	ceptual translati	on.							
X61294	LI retroposon, a	a portion of its ORF2 sequ	ence						
L1 retroposon; reverse tra	nscriptase-like p	rotein							
AQ377979			GSS						
AQ342069			GSS						
K02590 "pseudo-h3" /pseudo				beta-globin; beta-					
globin h3; globin; pseudogene									
AQ545809			GSS						
Z96215		genom	ic fragme	ent; subtelomeric					
DNA									
AQ572877			GSS						
AQ554929			GSS						
AQ545915			GSS						
AO344044			GSS						
AO378406			GSS						
U70924 reverse transcriptase			000						
c.ore receive numberipuloe									

Z78996	Anonymous marker; single read
J00338	repeat region
AQ547459	GSS

Figure 13 The updated version of the file illustrated in Figure 3.

Systematic	Common	Мар	EC	Description	Product	Phenotype	Function	Keywords	GenBank
Name 707191 1	Name		Number					HTC: DYS7: CT	ropost
29/101.1								polymorphism: 6	TG repeat
								polymorphism	ri e repour
AL022401	CHM			5' part of gene	dJ93L7.1 (RAB	Escort prote	in 1	HTG; CHM;	
				beyond this clone;	(REP-1, RAB pr	roteins		Choroideraemia;	,
				match: proteins	geranylgeranylt	ransferase c	omponent	geranylgeranyltra	ansferase
				P24300 P3/727 P26374	Tapetochoroida	lenna proteir I Dystrophy i	і, (ТСП)	Escort: REP-1: F	
				1 2007 4	protein)	i Bystrophy ((100)	Tapetochoroidal	Dvstrophy:
								TCD	
AC004386								HTG	
AC004388								HTG	
298950	dJ507115.	I		match: multiple proteins; match: CE02123 P90702 Q96499 P10661 P65027; match: P09896 P31866 P02405 P31028 P52809; match: Q00477 Q00494 P49213 P17843 P27076; match: cDNAs M19635 M15661 AB000910; match: cDNAs M19635 M15661 AB000910; match: multiple ESTs; match: T87328 T87321 AA181201 T41136; match: AA244162 R05264 N93353 AA191627; match: AA411822 AA328207 AA342359 T89286; 60S ribosomal protein L44 (L41, L36) like	Match: Multiple CE02123 P907(P65027; match: P02405 P31028 Q00477 Q00494 P27076; match: M15661 AB000 ESTs; match: T AA181201 T411 R05264 N93353 AA411822 AA3 T89286; 60S ritt (L41, L36) like	proteins; ma 2 Q96499 F 2 P09896 P3 3 P52809; ma 4 P49213 P1 2 cDNAs M19 910; match: 87328 T873 136; match: 3 AA191627; 28207 AA34 posomal prot	Alcri: P10661 1866 atch: 7843 9635 multiple 21 AA244162 match: 2359 ein L44	L44; Xq26.3-27.3	.30; L41; }
AC004478								HIG	
AC003666								HIG	
AC002549								HIG	
AC003669								TIG	hiam. V
L022U4									nisin, A
AUUU4383					1			пю	1

AC004072							HTG		
AC003683							HTG		
AC003658							HTG		
Y15994							MTM1 gene		
AL009175	REP1		match: SW P24386 EMBL X78121; (RAB ESCORT PROTEIN 1) (REP-1) (CHOROIDERAE MIA PROTEIN) (TCD PROTEIN)	dA43C13.1 (RAB PROTEINS choroideremia; rab GERANYLGERANYLTRANSFERASE geranylgeranyl transfe COMPONENT A 1) Xq21.1-Xq21.3				ab ansferase;	
AC003037							HTG		
AL008713			steroid 5-alpha- reductase; match: M32313	dJ93C23.1			3-oxo-5-alpha-steroid delta(4)-dehydrogenase; dihydrotestosterone; pseudogene: X		
AC002422							HTG		
AC002523							HTG		
M22332			ORF; putative	unknown protein			L1 insertion element		
AQ409366							GSS		
AQ356884							GSS		
AQ309743							GSS		
X61295 AQ572229			L1 retroposon, a p	ortion of its ORF	2 sequence		L1 retroposon; reverse transcriptase-like protein GSS		
AI683867							EST		
AQ390430							GSS		
AQ355719							GSS		
AQ559819				-			GSS		
AQ557343							GSS		
AQ573089							GSS		
M54985							psi-eta beta-like globin pseudogene		
AQ549999							GSS		
AI421777							EST		
AQ420901							GSS		
AQ382430							GSS		
AQ440210							GSS		
AQ536099							GSS		
AI475350							EST		
S67068									
AI821169							EST		
AL045241							EST		
AQ357079							GSS		
S80119	reverse trar homolog	nscriptase	This sequence con translation.	nes from Fig.2. F	Protein seque	ence is in o	conflict with the c	onceptual	

X61294				L1 retroposon, a portion of its ORF2 sequence			ence	L1 retroposon; reverse transcriptase-like protein		
AQ377979								GSS		
AQ342069								GSS		
K02590	pseudo-h3	3 /pse	eudo					beta-globin; beta-globin h globin; pseudogene		
AQ545809								GSS		
Z96215								genomic fragment; subtelomeric DNA		
AQ572877								GSS		
AQ554929								GSS		
AQ545915								GSS		
AQ344044								GSS		
AQ378406								GSS		
U70924	reverse tra	anscr	iptase							
Z78996								Anonymous marker; singl read		
J00338								repeat region		
AQ547459								GSS		

Figure 14 The same file as illustrated in Figure 13, shown in Excel format. The first row of column headings is not automatically included in new gene list files; they have been added here for clarity.