Information about the upgraded GA Tool:

The Gene Annotator (GA) tool was built to allow for functional analysis of arbitrary gene sets. Gene sets can be entered into the tool in multiple ways, the first being a list of gene symbols or gene identifiers. Valid identifier types include gene symbols and any of the following types of accession IDs: RGD, NCBI (Entrez) Gene, Ensembl Gene, Ensembl Protein, Affymetrix, GenBank Nucleotide, GenBank Protein, UniProtKB and UniGene. For convenience, identifier types can be supplied in a mixed manner. For instance, a symbol for one gene can be submitted with a UniProt identifier for another in the same input list. The only caveat is that only one type of numeric (that is all number) ID can be entered at a time since the tool will ask you what type of ID the numbers are and will assume that all numbers are that type of ID.

In addition to gene-specific identifiers, Ontology Accession IDs or KEGG Pathway IDs can be supplied. For each Ontology ID, all genes annotated to that term or one of its more specific "child" terms in the Rat Genome Database will be added to the gene list. If a KEGG Pathway ID is supplied, all genes associated with that pathway will be added to the gene list.

In some cases it may be advantageous to supply a genomic region. Gene lists can be generated by entering a chromosome, start position, stop position, and Assembly. All genes located in the input region are added to the gene list. Multiple assemblies are available for rat (*Rattus norvegicus*), mouse (*Mus musculus*), and human (*Homo sapiens*). Genome assemblies include rat versions 3.1, 3.4, 5.0, and Celera, human versions 36, 37, and Celera, and mouse versions 34, 36, 37, 38, and Celera.

Once a "bucket" of genes is defined, the user is brought to the annotation selection screen. This screen allows the user to specify the types of annotations to be included in the report. The first section displays a list of ontologies. Ontologies available include Pathway, Disease, GO, Mammalian Phenotype and ChEBI Chemical Interactions. When selected, annotations to that ontology for each gene in the gene list are retrieved from RGD and displayed on the report page.

The second section includes selection of possible external database identifiers and links to be included in the report. External links available include:

Gene-associated	Nucleotide sequence-associated	Protein-associated	References
Ensembl Gene	Ensembl Transcript	BIND	PubMed
Entrez Gene	GenBank Nucleotide	Ensembl Protein	
Germonline	IMAGE_CLONE	GenBank Protein	
HGNC ID	MGC_CLONE	Gene3D-CATH	
KEGG Report		HPRD ID	
MGD		InterPro	
UniGene		IPI	
		PANTHER	
		Pfam	
		PIRSF	
		PRINTS	
		ProDom	
		PROSITE	
		SMART	
		Superfamily-SCOP	
		TIGRFAMs	
		UniProtKB	

In addition there are other database IDs available. Here is a list of other external database links with the type of data they hold:

Database:	Type of data:
COSMIC	somatic mutations in cancer
NCBI's HomoloGene	orthology
KEGG Pathway	pathways
OMIM	disease
Pathway Interaction Database (PID)	pathways
PharmGKB	chemical interactions
TIGR	gene and gene variant predictions based on ESTs
Transposagen	knockout and transgenic rats

The last section allows for the selection of ortholog information that should be included. Rat, mouse, and human are available as orthologs.

Following selection of annotations, the report page is generated incorporating the user's preferences. Across the top is a scrollable listing of all the genes in the gene list. (If symbol conversions are made, or an identifier cannot be resolved, the user is notified via the status report at the bottom of the page.) Near the top of the report is a gene description. Descriptions are generated from annotation data and include information about the type of molecule the gene encodes, plus some or all of the functions it exhibits, processes it is involved in, pathways it participates in, diseases or phenotypes it is associated with, subcellular locations it is found in, and/or chemicals it interacts with. Below the description, ortholog information is provided. Ortholog information includes the orthologous symbol, RGD ID, and a link to the RGD gene report page.

Under the ortholog information is a listing of all annotations in RGD for the selected gene and any orthologs selected. Each annotation is listed with its term accession number, ontology term, evidence code and a link to the reference. Clicking the accession number will take the user to the RGD ontology report for the selected term, allowing the user to see all of the RGD genes annotated to that term, as well as information about the term, including its more general parent terms and, where applicable, more specific child terms. Returning to the results of the GA Tool, clicking the evidence code in the "Annotations" list takes the user to more information on how the annotation was made at RGD or acquired from other sources, including links to the original papers where applicable. The "External Links" section of the report includes identifiers for and links to the external sites that were selected in the previous screen and that have information on the gene selected. Clicking a link takes the user outside the GA tool to that third party site.

In addition to individual functional analysis of genes, the GA tool has the ability to display a distribution of genes across each of the ontologies selected. This feature can be accessed via the “Ontology Distribution” item in the menu bar across the top of the page. The top 50 terms from each ontology that have gene annotations associated with the genes in the list are displayed in descending order from most annotated to least annotated. The percentage of genes from the input gene set that are annotated to each term is listed along with the term and accession ID. Annotation counts include genes annotated to the term itself as well as to more specific child terms. Clicking the plus sign (+) next to each term displays the list of genes from the input gene set that have related annotations. If a gene is included in the list due to a child term annotation, that child term is listed in parentheses next to the gene symbol.

A checkbox is located to the right of each term. Clicking a box selects that term for "Cross Term Analysis". Terms may be selected from a single ontology or across multiple ontologies. For each term that is selected, the list of genes annotated to that term is compared to the list(s) of genes annotated to any other selected terms and the overlap is displayed in the box at the top left side of the page. The cross term analysis shows the percentage of genes in the original input list annotated to all of the terms selected. Clicking the plus icon (+) displays the list of genes annotated to all selected terms and gives the user the option of "re-entering" that subset of the original list into the GA Tool for further analysis ("Explore this Gene Set").

The menu bar at the top of the page includes options to download the data on the report page as a CSV download. The user has the option of downloading the annotation information for an individual gene or for all of the genes in the bucket. The data format has been designed to facilitate import into spreadsheet programs such as Microsoft Excel. In addition, the "Genome" link will display all genes in the bucket in the RGD Genome Viewer along with genomic position information for each gene.

About the GA Tool's "back-end":

The GA tool is a web-based application built on standard web technologies. Browsers officially supported include Internet Explorer 8/9, Mozilla Firefox 11/12, Google Chrome 18/19 and Safari 5.0/5.1. The user interface is built on HTML, javascript, css, and AJAX. The server side takes advantage of J2EE technologies and an Oracle database.

The tool is dependent on the data and pipelines that drive the Rat Genome Database. Ontology annotations are both manually curated by RGD staff and imported from 3rd parties. External database identifiers and links are imported via a variety of automated pipelines.