

Quick Start and Tutorial for Web-Based MatchMiner

Overview

Using MatchMiner

 Interactive LookUp

 Batch LookUp

 Batch Merge

Additional Help and Information

 Hints

 Selecting algorithms

 Identifiers and Data Sources

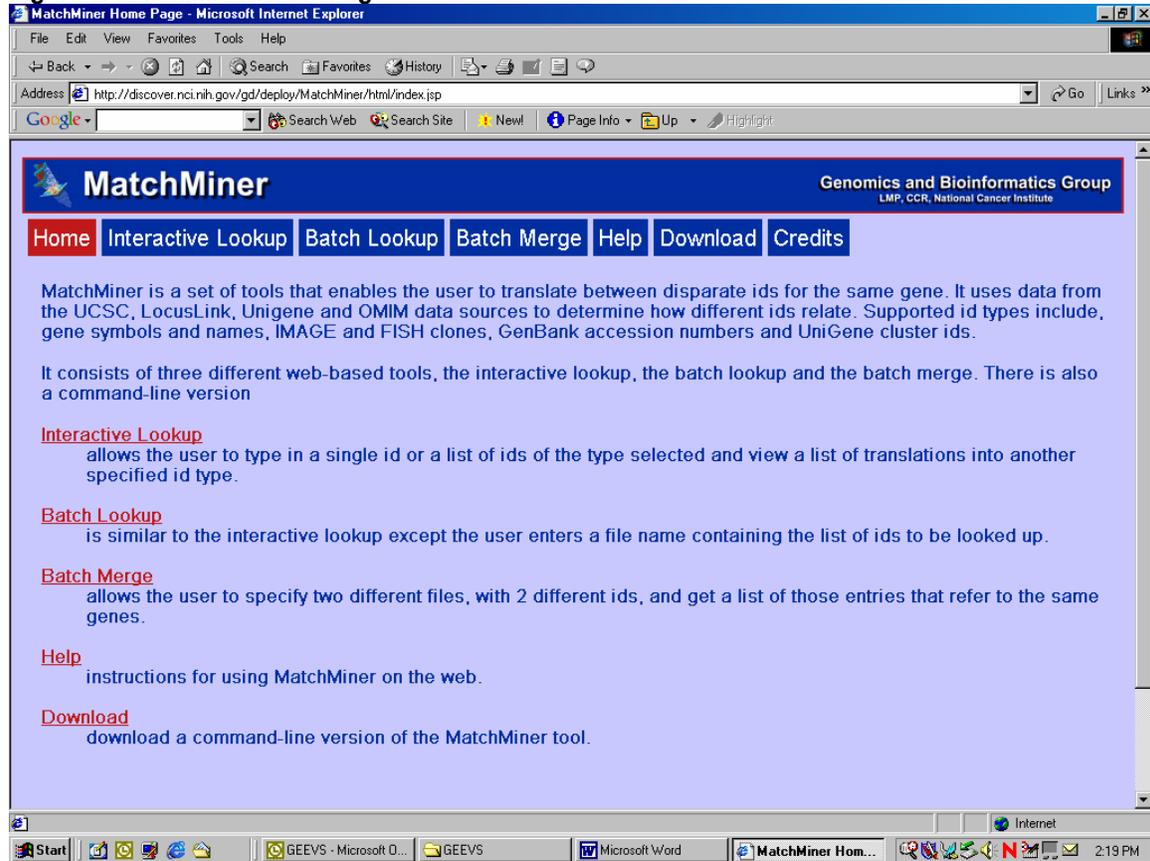
Overview:

MatchMiner is a freely available tool for the navigation among gene and gene product Identifiers. It has two functions: 1) LookUp: translates from one type of Identifier to another, either interactively or in batch; and 2) Merge: Identifies and outputs the one-to-one, one-to-many, and many-to-many relationships between two lists of Identifiers of either the same or different types. This is done by translating the input into an internal gene index and then retrieving the requested information linked to that index.

Using MatchMiner:

The homepage for MatchMiner looks as follows:

Figure 1: MatchMiner Home Page

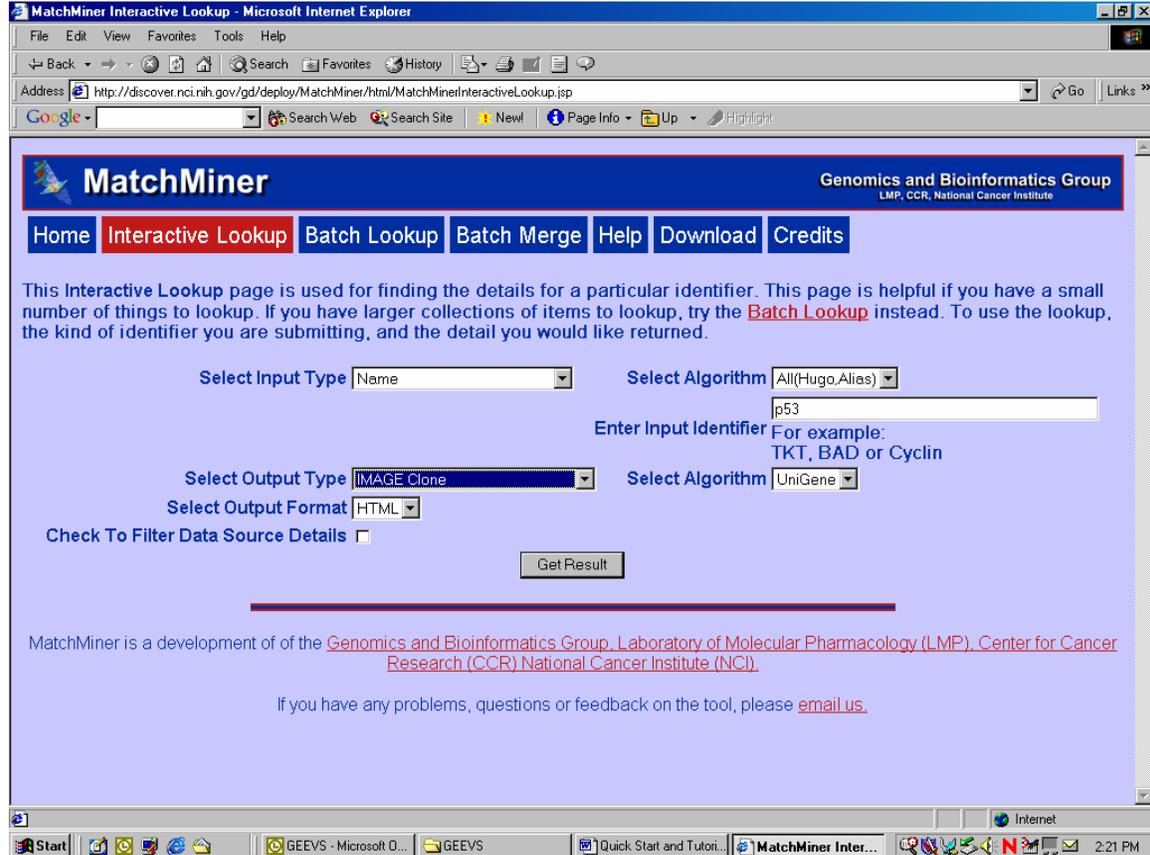


By clicking on the buttons, the user can choose the appropriate searches described in detail below. Examples are provided as well. Before proceeding, please download the example files located under Help to your home machine. Keep track of where you save them.

Interactive Lookup:

The Interactive LookUp permits the input of a single Identifier for quick querying.

Figure 2: Interactive LookUp Screen



1. Specify the type of Identifier you will input. The current choices are Chromosome Location (1p36.3, 1p36, 1p, etc.), Name (gene symbol, alias, or descriptive name), GenBank Accession number, Id (currently restricted to UniGene cluster, but will include LocusLink, OMIM, and others), FISH clone, and IMAGE clone Id. *Example: Name*
2. Choose what search strategy you want to use to match your input to a gene index (see Table 1 in Helpful Hints). *Example: All(HUGO, Alias)*
3. Enter your Identifier. *Example: p53*
4. Choose the type of output you want. Currently, you can request Cytogenetic Location (1p36.3, 1p36, 1p, etc.), Name (symbol, alias, or descriptive name), GenBank Accession number, UniGene cluster Id, FISH clone, IMAGE clone Id, and Sequence Location Number (in bp). *Example: IMAGE clone*
5. Choose the algorithm for returning the output type you want for the matching gene index (Table 1 in Helpful Hints). *Example: UniGene (this will be the only choice)*
6. Choose the output format, either HTML, text, or Excel (use text or Excel if you wish to save your results locally). The Excel format will put a space in front of text entries. The text format has no such extra characters. In Excel, you can save the file by choosing "Save." *Example: HTML*

7. Optional: Check the “Filter Data Source Details” button to suppress the meta-details for the match.
8. Click the “Get Result” button.

Figure 3a: LookUp Result (HTML) Summary

The screenshot shows a web browser window titled "MatchMiner Lookup Results - Microsoft Internet Explorer". The address bar shows the URL: <http://discover.nci.nih.gov/gd/deploy/MatchMiner/html/InteractiveLookupResults.jsp>. The MatchMiner logo is visible at the top left, and the Genomics and Bioinformatics Group logo is at the top right. A navigation menu includes: Home, Interactive Lookup, Batch Lookup, Batch Merge, Help, Download, Credits.

The main content area is titled "Lookup Results" and "Results Summary". It contains a table with the following data:

Result Category	Total
Items from the input list that has output	1
Items from the input list with no output	0
Items from the input list that were not found in the database	0

Below the summary is a "Results" table with the following columns: Function, Position, Input Name, Output Clone, Index, and Input Data Source (UniGene, LocusLink, UCSC, OMIM, HugoName). The Output Data Source columns are UniGene, LocusLink, UCSC, OMIM, HugoN.

Function	Position	Input Name	Output Clone	Index	Input Data Source					Output Data Source				
					UniGene	LocusLink	UCSC	OMIM	HugoName	UniGene	LocusLink	UCSC	OMIM	HugoN
Lookup Output	1	p53	5447739	6853	-	LocusLink	-	OMIM	-	UniGene	-	-	-	-
Lookup Output	1	p53	4554240	6853	-	LocusLink	-	OMIM	-	UniGene	-	-	-	-
Lookup Output	1	p53	5457909	6853	-	LocusLink	-	OMIM	-	UniGene	-	-	-	-
Lookup Output	1	p53	4341365	6853	-	LocusLink	-	OMIM	-	UniGene	-	-	-	-

The browser's taskbar at the bottom shows the Start button, several icons, and the system tray with the time 2:21 PM.

Figure 3b: LookUp Results Table (HTML)

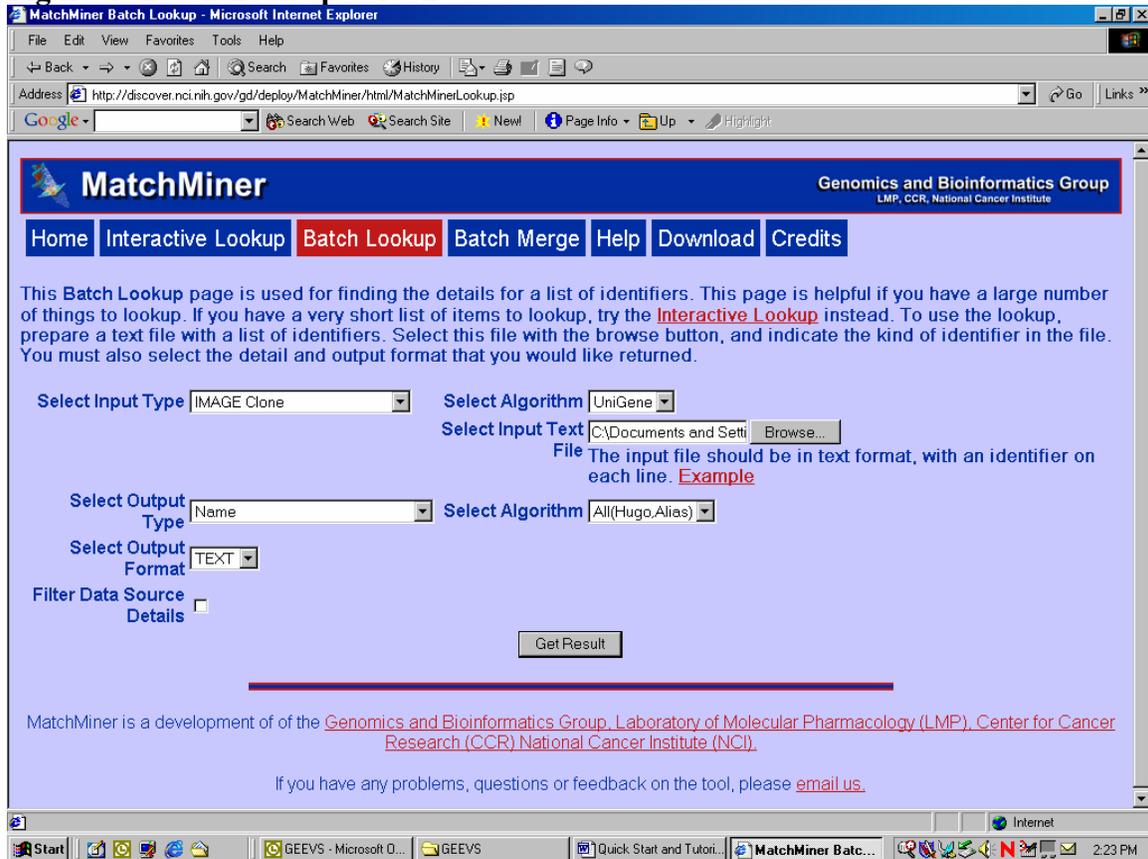
Function	Position	Input Name	Output Clone	Index	Input Data Source					Output Data Source				
					UniGene	LocusLink	UCSC	OMIM	HugoName	UniGene	LocusLink	UCSC	OMIM	HugoName
Lookup Output	1	p53	5447739	6853	-	LocusLink	-	OMIM	-	UniGene	-	-	-	-
Lookup Output	1	p53	4554240	6853	-	LocusLink	-	OMIM	-	UniGene	-	-	-	-
Lookup Output	1	p53	5457909	6853	-	LocusLink	-	OMIM	-	UniGene	-	-	-	-
Lookup Output	1	p53	4341365	6853	-	LocusLink	-	OMIM	-	UniGene	-	-	-	-
Lookup Output	1	p53	4508539	6853	-	LocusLink	-	OMIM	-	UniGene	-	-	-	-
Lookup Output	1	p53	4524419	6853	-	LocusLink	-	OMIM	-	UniGene	-	-	-	-
Lookup Output	1	p53	6056857	6853	-	LocusLink	-	OMIM	-	UniGene	-	-	-	-
Lookup Output	1	p53	5452758	6853	-	LocusLink	-	OMIM	-	UniGene	-	-	-	-
Lookup Output	1	p53	5804474	6853	-	LocusLink	-	OMIM	-	UniGene	-	-	-	-
Lookup Output	1	p53	4339142	6853	-	LocusLink	-	OMIM	-	UniGene	-	-	-	-
Lookup Output	1	p53	5587461	6853	-	LocusLink	-	OMIM	-	UniGene	-	-	-	-
Lookup	1	p53	5245722	6853	-	LocusLink	-	OMIM	-	UniGene	-	-	-	-

The LookUp Result Page begins with a description of the parameters for the search, followed by a summary of the number of hits. This is then followed by a table indicating the type of output, the input Identifiers, the matching output Identifier, the internal gene index, and a summary of the data source in which the input Identifier is found, followed by a summary of the output Identifier data sources. These details are important because they allow the user to determine how the matches are made. As seen in figure 3, the input Identifier “p53” is found in LocusLink and OMIM, but not UCSC or UniGene. The IMAGE clone Ids corresponding to the p53 gene index are found in UniGene. This represents a case where a match is made strictly by association through the gene index and is not present explicitly in the database.

Batch LookUp:

Batch LookUp is operationally similar to the Interactive LookUp except that instead of entering a single Identifier, you input a text file with a list of Identifiers. The input file should be formatted as plain text with one Identifier per line (see the example file). To use the Batch LookUp:

Figure 4: Batch LookUp Screen



1. Specify the type of Identifier your input file contains. The current choices are Chromosome Location (1p36.3, 1p36, 1p, etc.), Gene Name (symbol, alias, or descriptive name), GenBank Accession number, UniGene cluster Id, FISH clone, and IMAGE clone Id. *Example: IMAGE Clone*
2. Choose what search strategy you want to use to match your input to a gene index (see Table 1 in Helpful Hints). *Example: UniGene (note: this will be the only choice available)*
3. Enter the filename of your input file. Clicking on the “Browse” button will allow you to browse to your file. *Example: Use the file called “top100cdnaclids.txt”*
4. Choose the type of output you want. Currently, you can request Cytogenetic Location (1p36.3, 1p36, 1p, etc.), Gene Name (symbol, alias, or descriptive name), GenBank Accession number, UniGene cluster Id, FISH clone, IMAGE clone Id, and Sequence Location Number (in bp). *Example: Name*
5. Choose the algorithm for returning the output type you want for the matching gene index (Table 1 in Helpful Hints). *Example: ALL(HUGO, Alias)*
6. Choose the output format, either HTML, Excel or text (use Excel or text if you wish to save your results locally).
7. Optional: Check the “Filter Data Source Details” button to suppress the meta-details for the match.
8. Click the “Get Result” button.

The results will be returned as with the Interactive LookUp. There will be a summary of the choices made to generate the output, a summary of number hits and misses, and the table with the results. There is one additional column, entitled "Position" which indicates the position of the input identifier in the input file. Sorting using this column will reorder the results to match in the input order. (Note: This function only works if the files are saved locally as text and manipulated in another program such as Excel.) Figure 5 shows what the results would look like after importing to Excel. Excel as the file format will do two things: [1] all text will be preceded with a space and [2] if Excel is installed on your machine, the results will be displayed in Excel automatically. To save the file, simply click save and give the file a name.

Figure 5: Batch LookUp results in text file format

The screenshot shows a web browser window with the address <http://discover.nci.nih.gov/gd/deploy/MatchMiner/html/LookupResultsText.jsp>. The page content is rendered as an Excel spreadsheet. The first section is an 'Input Summary' with the following data:

Date	Wednesday, October 9, 2002
Operation	Lookup
Input Source Name	C:\Documents and Settings\busseyk.000\My Documents\GEEVS\MatchMiner tests\top 100 cdna clids.txt
Input Type	IMAGE Clone
Input Algorithm	UniGene
Output Type	Name
Output Algorithm	All(Hugo, Alias)
Data Source Filter	No
Output Format	Text

The second section is an 'Output Summary' with the following data:

Items from the input	102
Items from the input	3
Items from the input	4

The third section is a table of results with the following columns: Function, Position, Input Clon, Output Na, Index, UniGene, LocusLink, UCSC, OMIM, HugoName, UniGene, LocusLink, UCSC, OMIM, HugoName. The data rows are:

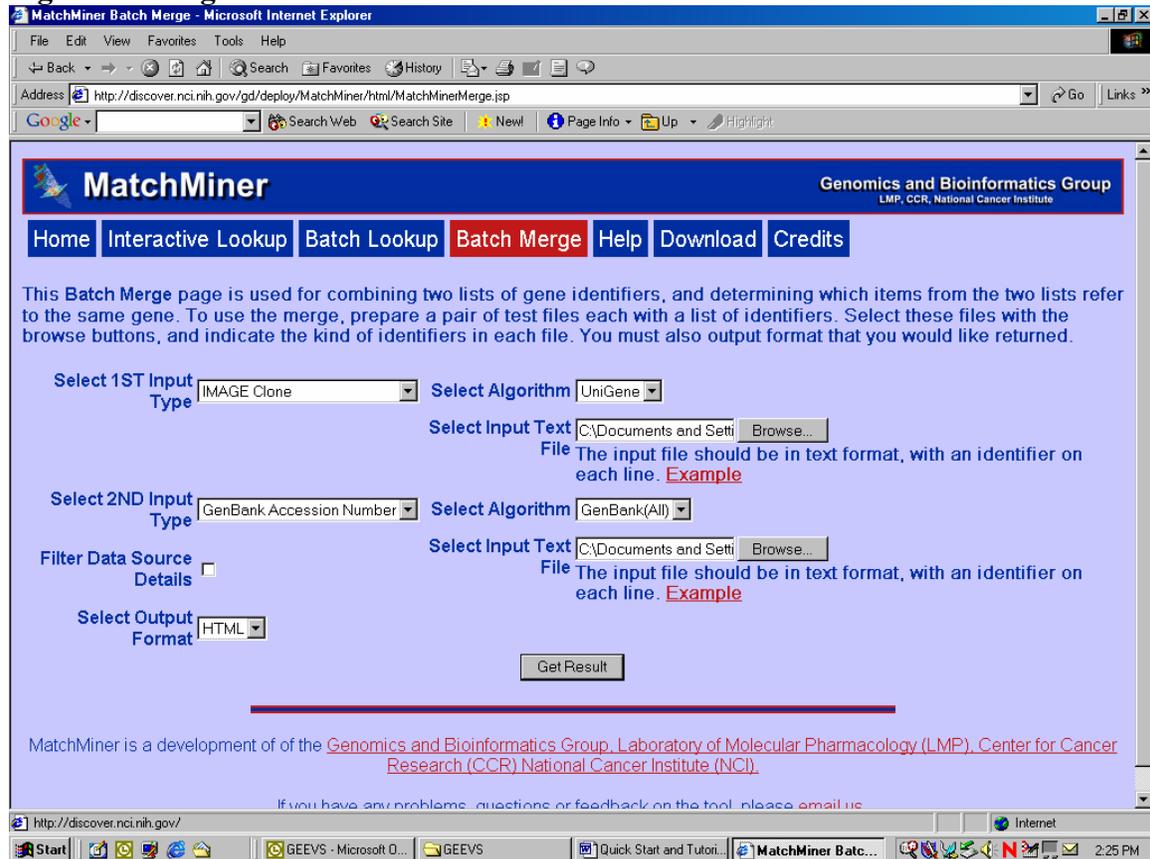
Function	Position	Input Clon	Output Na	Index	UniGene	LocusLink	UCSC	OMIM	HugoName	UniGene	LocusLink	UCSC	OMIM	HugoName
Lookup Output	89	110022	CCND1	572	UniGene	-	-	-	-	UniGene	LocusLink	-	OMIM	Hug
Lookup Output	32	429934	ID2	3239	UniGene	-	-	-	-	UniGene	LocusLink	-	OMIM	Hug
Lookup Output	99	271478	MXI1	4385	UniGene	-	-	-	-	UniGene	LocusLink	-	OMIM	Hug
Lookup Output	63	364752	ARHC	374	UniGene	-	-	-	-	UniGene	LocusLink	-	OMIM	Hug
Lookup Output	49	488479	TPM1	6864	UniGene	-	-	-	-	UniGene	LocusLink	-	OMIM	Hug

Batch Merge:

The Batch Merge function is designed to Identify all of the one-to-one, one-to-many, and many-to-many relationships between two lists of Identifiers. These lists can have the same type of Identifier or different types of Identifiers. The resulting output will consist of a summary of the search parameters, as well as a summary of the items matched, unmatched but in the database, and items missing from the database. A table follows, giving all of the possible matches between the two lists, with Identifiers with ambiguous

assignments to gene indexes flagged. Currently, the results are sorted by those items that match between the lists, followed by items unmatched but present in the database sorted by file, and items not found in the database, again sorted by file. Two columns indicate the original order of the input files entries and can be used to resort the data appropriately. (Note: This function only works if the files are saved locally as text and manipulated in another program such as Excel)

Figure 6: Merge Function Screen



To perform a Batch Merge:

1. Specify the type of Identifier your first input file contains. The current choices are Chromosome Location (1p36.3, 1p36, 1p, etc.), Gene Name (symbol, alias, or descriptive name), GenBank Accession number, UniGene cluster Id, FISH clone, and IMAGE clone Id. *Example: IMAGE Clone*
2. Choose what search strategy you want to use to match your first input file to a gene index (see Table 1 in Helpful Hints). *Example: UniGene (note: this will be the only choice available)*
3. Enter the filename of your first input file. Clicking on the “Browse” button will allow you to browse to your file. *Example: Use the file “top100cdnaclids.txt”*
4. Specify the Identifier type of the second file. Currently, you can request Cytogenetic Location (1p36.3, 1p36, 1p, etc.), Gene Name (symbol, alias, or

- descriptive name), GenBank Accession number, UniGene cluster Id, FISH clone, and IMAGE clone Id. *Example: GenBank Accession Number*
5. Choose what search strategy you want to use to match your second input file to a gene index (see Table 1 in Helpful Hints). *Example: GenBank(All)*
 6. Enter the filename of your second input file. Clicking on the “Browse” button will allow you to browse to your file. *Example: Use the file “top100oligoacid.txt”*
 7. Choose the output format, either HTML, Excel, or text (use Excel or text if you wish to save your results locally).
 8. Optional: Check the “Filter Data Source Details” button to suppress the meta-details for the match.
 9. Click the “Get Result” button.

Figure 7a: Merge Results Summary

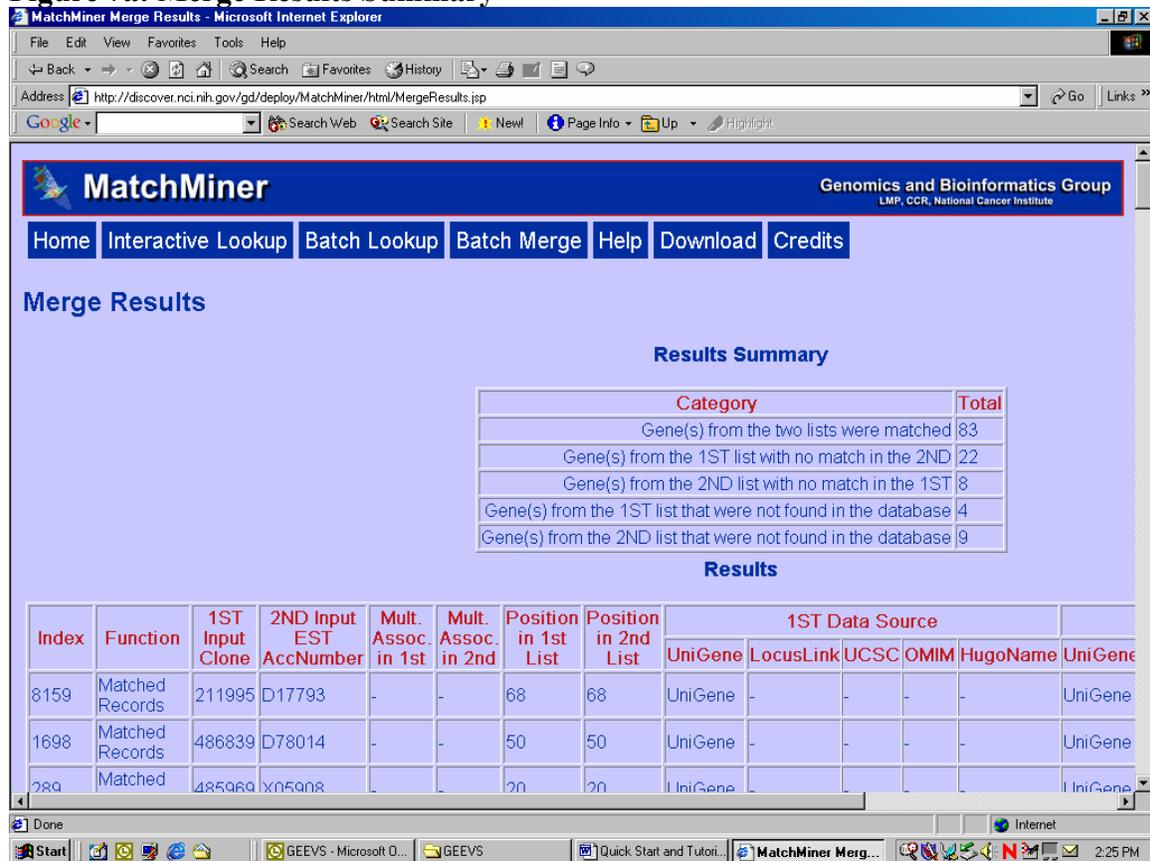


Figure 7b: Merge Results Table

1ST Input Clone	2ND Input EST AccNumber	Mult. Assoc. in 1st	Mult. Assoc. in 2nd	Position in 1st List	Position in 2nd List	1ST Data Source					2ND Data Source		
						UniGene	LocusLink	UCSC	OMIM	HugoName	UniGene	LocusLink	UCSC
211995	D17793	-	-	68	68	UniGene	-	-	-	-	UniGene	-	-
486839	D78014	-	-	50	50	UniGene	-	-	-	-	UniGene	-	-
485969	X05908	-	-	20	20	UniGene	-	-	-	-	UniGene	-	-
415495	Y00433	-	-	31	31	UniGene	-	-	-	-	UniGene	-	-
486676	J02923	-	-	28	28	UniGene	-	-	-	-	UniGene	-	-
428443	M61916	-	-	60	60	UniGene	-	-	-	-	UniGene	-	-
271985	M27160	-	-	1	1	UniGene	-	-	-	-	UniGene	-	-
345538	X12451	-	-	25	25	UniGene	-	-	-	-	UniGene	-	-
510377	X91247	-	-	75	75	UniGene	-	-	-	-	UniGene	-	-
110022	X59798	-	-	89	89	UniGene	-	-	-	-	UniGene	-	-
363919	U70063	-	-	61	61	UniGene	-	-	-	-	UniGene	-	-

Additional Help and Information

Hints

1. A record marked “No Gene Index Found” indicates that the entry could not be mapped to a gene index. Either the entry isn’t in the database or there has been a mismatch between entry identifier type and search algorithm.
2. Currently, files can contain only one type of identifier. Files with mixed identifier types will be processed but will return “No Gene Index Found” for those entries that are of a different identifier class than the specified search.

Selecting Search Algorithms

The variety of both identifier classes and potential database sources requires specifying how matches should be made. This decision is influenced by several factors including type of identifiers, primary and secondary data sources that contain the identifier, and the context in which the identifiers will be used. In MatchMiner, selecting algorithms for identifier translation has been divided into a two step process. The first step is to tell the program how to match the input to an internal gene index. For example, if the input is a

file containing gene symbols, MatchMiner can either match the entries to gene indexes assuming all entries are HUGO (official) names or it can perform a broader search that matches all entries possible to HUGO names first and then goes back to the entries that were not matches and re-scans the list for matches to aliases. The second step selects the way the program should retrieve the requested information. Several of the data sources have similar but not the same information (e.g. cytogenetic location). The information retrieved will be dependent on how that information is going to be used. For instance, if the task is to look up the cytogenetic location of UniGene clusters, one might want to make sure that the locations returned are those from UniGene. However, suppose the list of UniGene clusters was restricted initially to “known” genes. Then, instead of wanting what matched UniGene, a user might want to have a more precise location based on sequence to cytogenetic map translation from UCSC.

Table 1 gives a description of the various algorithms associated with different identifier classes. Additional algorithms are in the works and suggestions are always welcome.

Table 1: MatchMiner Lookup Search Options

Identifier Type	Input Algorithm	Output Algorithm
Name (Gene Symbol, alias, or descriptive name)	ALL (HUGO, Alias) Starts with HUGO. If not found, proceeds through all other sources until a matching alias is found.	ALL (HUGO, Alias) Returns the HUGO name. If no name is flagged as HUGO, returns all aliases.
	Official Searches all data sources for match as the HUGO name.	Official Returns the HUGO name. If not found, nothing is returned.
	Long Searches all data sources for descriptive name	Long Returns all descriptive names
GenBank Accession Number	GenBank (ALL) Searches all data sources starting with UCSC known genes, then LocusLink, UniGene, and UCSC ESTs until match found.	GenBank (ALL) Returns accession number from UCSC known genes. If not found, proceeds through UniGene then UCSC EST
	Data Source Specific Lookup input in a specific data source	Data Source Specific Returns accession numbers found in a particular data source

Identifier Type	Input Algorithm	Output Algorithm
IMAGE clone Id	UniGene Only data source with IMAGE clone Ids	UniGene Returns all IMAGE clone Ids associated with the UniGene Cluster corresponding to the matching gene
Cytogenetic Location	ALL Searches all gene indexes for matching chromosome band	ALL Returns chromosome band from UCSC sequence to band translation. If not found, proceeds through all other sources with multiple bands listed separately.
		UCSC Returns chromosome band from UCSC sequence to band translation.
Database Id	UniGene Searches gene index for matching UniGene Id	UniGene Returns UniGene Id.
Sequence Location Number (BP)	Not Implemented	Transcription Start Returns transcription start from UCSC Known Genes. If not found, proceeds to UCSC EST
FISH Clone	UCSC Searches UCSC FISH clones for match to gene index based on sequence position overlap with UCSC known genes.	UCSC Returns FISH clone Id.

Identifiers and Data Sources

The ability to obtain meaningful results with MatchMiner is greatly enhanced by understanding what types of identifiers can be encountered, where they are used in public databases, and how they relate to one another. The identifiers classes encountered in omic research refer to genes at several levels of resolution from sequence to function to all information about the gene.

GenBank is the repository for sequence information maintained and curated by the NCBI (Benson et al. 2002). It is one of three databases involved in the International Nucleotide Sequence Database Collaboration; the European Molecular Biology Laboratory (EMBL) data library and the DNA Data Bank of Japan (DDBJ) are the other partners. Each of

these databases share a standardized record format that permits daily alignment of data between them so that any sequence is available through any of the databases. A sequence record in GenBank is identified by two numbers: a GenBank accession number and a gi number (Benson et al. 2002). The accession number is unique and fixed. It always references a particular sequence record. The GI number, in contrast, is a unique number that references a particular edited state of the sequence. As the sequence in the record is changed, the gi number will also change. An indication of how many times this has taken place in a given record is indicated by the “accession number.version” that appears on the same line as the gi number. The distinction between gi number and accession number is important to keep in mind when working with these records as it will influence the ease by which an analysis or mapping can be retraced in the future.

Certain classes of sequences carry a stereotypical GenBank accession number. RefSeq is the most prominent of these classes and represents an attempt to define a “canonical” sequence for reference to describe various genes (Pruitt and Maglott 2001). These records are denoted by NG_, NT_, and NC_ for genomic sequences, NM_ and XM_ for mRNA sequence, and NP_ and XP_ for protein. RefSeq records are derived from both computational and curated methods. As such, they carry a tag indicating the quality of the sequence ranging from genome annotation corresponding to modeled contigs, mRNA, and proteins, to reviewed records which have been curated by NCBI staff (see (Pruitt and Maglott 2001) or <http://www.ncbi.nlm.nih.gov/LocusLink/RSfaq.html> for an in depth description).

The association of one sequence from a gene with all other sequences from that gene, especially at the mRNA level is the purpose of UniGene (Wheeler et al. 2002). A UniGene cluster is a non-redundant grouping mRNA and EST sequences that have been clustered together by virtue of sequence similarity. This permits grouping of non-overlapping sequences that span the length of a gene’s mRNA. These records, denoted by a unique UniGene cluster id, include all mRNA and EST GenBank sequence records mapping to the cluster as well as the RefSeq GenBank accession number for the protein of product of the cluster. These records also carry the official name for the gene if it has been assigned and a chromosomal location if known.

LocusLink is a non-redundant repository of annotated gene information including descriptions of function, mappings in the GeneOntology™ as provided by Proteome Inc., chromosomal location, and STS and other sequence variations (Wheeler et al. 2002). In addition, there are entries for the official gene name, or in its absence a provisional name, aliases including common names and alternative gene symbols, and cross-references to UniGene cluster ids, RefSeq GenBank accession numbers, and OMIM records.

The cataloging of syndromes and their genetic etiology is the focus of OMIM. These highly curated records contain descriptions of identified genetic mutations and the resulting phenotypes as well as syndromes with as of yet unidentified genetic components. All entries are highly referenced to primary literature. As such, OMIM is an excellent source of gene aliases and common names.

Not all sequences in an omic-level project are gene based. STS sequences are useful because they precisely locate a particular region of DNA to a location in the genomic contig. Such non-coding sequences can be particularly useful in gaining information at the DNA level about regulatory regions. The human genome sequence builds available from University of California Santa Cruz (Cheung et al. 2001; Lander et al. 2001; McPherson et al. 2001) give access to STS and their aliases along with their genomic location. In addition, this resource also encompasses other types of records from GenBank and contains curated links back to UniGene, LocusLink, and OMIM. NCBI also provides a build of the human genome with similar data.

References

1. Benson DA et al (2002) GenBank. *Nucleic Acids Res* 30 (1):17-20.
2. Cheung VG et al (2001) Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* 409 (6822):953-958.
3. Lander ES et al (2001) Initial sequencing and analysis of the human genome. *Nature* 409 (6822):860-921.
4. McPherson JD et al (2001) A physical map of the human genome. *Nature* 409 (6822):934-941.
5. Pruitt KD and Maglott DR (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* 29 (1):137-140.
6. Wheeler DL et al (2002) Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res* 30 (1):13-16.