

## **Text-mining to identify genetically-related cancers**

Chris D Bajdik <sup>1\*</sup>

Byron Kuo <sup>1</sup>

Shawn Rusaw <sup>2</sup>

Steven Jones <sup>2</sup>

Angela Brooks-Wilson <sup>1,2</sup>

<sup>1</sup> Cancer Control Research Program, BC Cancer Agency

<sup>2</sup> Genome Sciences Centre, BC Cancer Agency

\* to whom correspondence should be sent, at 201-601 West Broadway, Vancouver BC, Canada  
V5Z 4C2

Running Title: Text-mining for genetically-related cancers

## Abstract

Objective: Online Mendelian Inheritance in Man (OMIM) is a computerized database of information about genes and heritable traits in human populations, based on information reported in the scientific literature. Our objective was to establish an automated text-mining system for OMIM that will identify genetically-related cancers. Methods: We developed the computer program CGMIM to search for entries in OMIM that are related to one or more cancer types. We performed manual searches of OMIM to verify the program results. Results: In the OMIM database on October 1, 2003, CGMIM identified 1826 genes related to one or more cancer types. BRAF (OMIM \*164757) and CDKN2A (OMIM \*600160) were each related to 14 types of cancer. There were 38 genes related to cancer of the esophagus, and nine genes related to both brain and prostate cancer. Analysis of the results indicate that roughly two genes in OMIM should mention both cancer of the esophagus and cancer of the stomach by chance, but the number of genes in OMIM that mention both cancers is 18. This nine-fold discrepancy suggests cancer of the esophagus and cancer of the stomach are more genetically related than current literature suggests. Conclusions: CGMIM can identify groups of genetically-related cancers. In several ways, groups based on shared genetic factors are anticipated to lead to further etiologic hypotheses and advances regarding environmental agents. CGMIM can be obtained free of charge from the Genome Sciences Centre website at the BC Cancer Agency (<http://www.bcgsc.bc.ca/>).

## Introduction

Cancer is the result of both genetic and environmental factors. Historically, epidemiologists have studied cancer of a specific anatomic site. This is partly motivated by the assumption that cancers arising at different sites are distinct forms of disease. Genetic studies have demonstrated that some genes are related to several types of cancer. Hereditary non-polyposis colorectal cancer (involving colorectal, urinary tract, ovarian and endometrial cancer) and Li-Fraumeni syndrome (involving breast cancer, brain cancer, leukemia, soft tissue sarcomas, bone sarcomas, and adrenal-cortical carcinoma) are examples where groups of cancers are genetically related. Groups of cancers have been previously identified using family histories as reported at a cancer clinic.<sup>1</sup>

Online Mendelian Inheritance in Man (OMIM; <http://www.ncbi.nlm.nih.gov/omim/>) is a computerized database of information about genes and heritable traits in human populations. The database is created and edited by Victor McKusick at Johns Hopkins University and colleagues around the world.<sup>2</sup> We consider it to be a particularly high-quality data source because it is curated by a genetically-knowledgeable team, based on information reported in the scientific literature, and continuously updated. OMIM is maintained on the Internet by the National Center for Biotechnology Information at the US National Institutes of Health.

Data mining aims to discover unexpected trends and patterns from large sets of data.<sup>3</sup> The rapid growth of biomedical literature databases enhances the value of mining text in particular. Text-mining has been described as a modular process involving document categorization, named

entity tagging, fact and information extraction, and collection-wide analysis.<sup>4</sup> In document categorization, a subset of potentially relevant documents is retrieved to increase the efficiency of subsequent steps. Named entity tagging identifies the important entities or objects mentioned in the article, often using a list of synonyms. In fact and information extraction, the relationships between entities are identified. Finally, in collection-wide analysis, information extracted from different documents is integrated.

Many studies aim to explore genes and gene-environment interactions in cancer etiology. The design of these studies requires (1) definition and ascertainment of appropriate case and control groups, (2) identification of important clinical and environmental factors, and (3) identification of candidate genes and variants that are hypothesized to be associated with disease. Each of these steps can benefit from effective text-mining of public data sources. Our objective was to establish an automated text-mining system for OMIM that will identify genetically-related cancers. This will help us to design studies of genes and gene-environment interactions in cancer.

## Materials and Methods

We developed the computer program CGMIM to text-mine entries in OMIM. The program was written in the Perl computer language and implemented on a Linux workstation. CGMIM was used to search OMIM entries for each type of cancer, and those results were then cross-classified to identify genes common to groups of cancer types. OMIM is updated daily and we created static copies of the database to provide a stable reference for search evaluation. Copies of OMIM

were downloaded between March and October of 2003, and each copy contained more than 14,000 entries.

### *The text-mining algorithm*

The text-mining algorithm begins by separating paragraphs into constituent sentences, assuming sentences end with a period followed by a space. Although there are degenerate cases (e.g., “Dr. Smith”), our experience was that these are uncommon in OMIM. “Stemming” was then used to remove capitalization and common suffixes from words. The following example demonstrates the stemming process:

UNSTEMMED: Large-cell lymphomas comprise approximately 25% of all non-Hodgkin lymphomas in children and young adults, and approximately one-third of these tumors have a t(2;5)(p23;q35) translocation.

STEMMED: larg-cell lymphoma compris approxim 25% of all non-hodgkin lymphoma in children and young adult, and approxim on-third of these tumor have a t(2;5)(p23;q35) transloc

Stemming simplifies the text-mining process by changing similar words to identical word fragments. There are many words and phrases that refer to cancer. Natural Language Processing attempts to recognize concepts that might be obscured by language style, grammar, nomenclature and syntax. In medicine, much of the variation occurs in words that refer to the disease. A breast

cancer might be described as a breast tumor, breast carcinoma or mammary gland neoplasm. Other text variations might be the result of English grammar. For example, breast cancer might be referred to as cancer of the breast, and several cancers might be referred to in a list. E.g., “cancer of the ovary, breast, and skin”. A list of synonyms for each cancer type was developed based on the International Classification of Disease for Oncology.<sup>5</sup>

Our list of synonyms was stemmed, and then compared to the stemmed sentences in OMIM. For each phrase in the synonym list, CGMIM searched for sentences that contain all of the individual words in a phrase.

An OMIM entry may contain alternative entry names, mapping information, a text summary, references to key publications, examples of known allelic variants, and a clinical synopsis of the corresponding phenotype. Some of these fields are subjective, such as the examples of allelic variants, and we restricted our search to the text summary. Finally, not all OMIM entries refer to specific genes. Some entries refer to heritable traits for which no gene has been identified. In addition, there can be more than one OMIM entry that refers to a specific gene. This typically occurs when the entry for a trait is linked to a gene that was previously identified and described in a separate OMIM entry. Because OMIM is dynamically organized and updated, this type of multiple referencing is unavoidable. To restrict searches to only the OMIM entries for genes, CGMIM compares each OMIM entry name and alternative name with a list of gene names assigned by the Human Genome Organization (HUGO; <http://www.gene.ucl.ac.uk/hugo/>).

## *Cancer groups*

CGMIM considers 21 anatomic sites based on the major sites of cancer identified by the National Cancer Institute of Canada [ref 6, page 18, Table 1]. One way CGMIM recognizes genetically-related cancers is by identifying all of the cancer types associated with a specific gene. A second way is by identifying the number of genes that are mentioned by OMIM for any combination of cancer types. This was done by generating a table with rows and columns labeled according to cancer type, in which the table cells contain the number of genes associated with the cancers that define the row and column. We refer to this table as the siteXsite matrix. These numbers provide a measure of the genetic “relatedness” of two different cancer types, and CGMIM provides gene names in its output. If several genes are related to a type of cancer, and several genes are related to another type of cancer, some genes will be related to mention both types of cancer by chance alone. The “expected” number of genes (E) related to both types of cancer can be estimated as the total number of genes related to cancer, multiplied by the probabilities for individual cancer types. The latter probabilities are estimated as the proportion of genes in the matrix that are related to each site. The observed number of genes (O) is the number in the matrix cell, and O/E indicates whether more genes are associated with a pair of the cancer sites than chance alone would predict. An O/E value of 1.0 would indicate the number of genes observed is the number expected by chance. An approximate 95% confidence interval (95% CI) is  $O/E \pm (1.96/\sqrt{E})$ .

We also performed manual searches of the OMIM database to verify the CGMIM output. The manual searches allowed us to identify the strengths and weakness of the computerized search,

and iteratively modify the software. We considered a sample of OMIM entries and read through the entire text to determine whether the entries referred to a cancer, or if entries were identified by the program where, in reality, there was no true cancer reference. This type of review is based on the expertise of the readers and is not strictly objective. We selected a random sample of OMIM entries and manually reviewed the text to identify sentences that referred to evidence of no association with a particular cancer. (E.g. “An early study showed the gene was not related to breast cancer.”) While an OMIM entry might include a sentence of that sort, another sentence in the entry might cite evidence supporting the same association. (E.g. “A subsequent study showed the gene was related to breast cancer.”) Despite the negative evidence, the OMIM entry cites evidence supporting the association and hence is proper to include when tallying entries associated with the cancer.

## Results

In the OMIM database on October 1, 2003, CGMIM identified 1826 genes related to cancer. Figure 1 shows the number of genes related to one or more of the 21 cancer sites. The genes BRCA2 (OMIM \*600185), BRAF (OMIM \*164757) and CDKN2A (OMIM \*600160) were related to the greatest number of cancer types. The OMIM entries for all three genes mention leukemia, melanoma, breast cancer, colorectal cancer, pancreatic cancer, stomach cancer, ovarian cancer and prostate cancer. The entry for BRCA2 also mentions cancer of the brain, larynx, cervix, uterus, thyroid and kidney. The entry for BRAF also mentions lymphoma and cancer of the lung, bladder, testes, cervix and uterus. The entry for CDKN2A also mentions

lymphoma and cancer of the lung, bladder, brain, esophagus and kidney. Each gene defines a large group of related cancers.

The siteXsite matrix of cancers is given in Table 1. Diagonal cells in the matrix contain the total numbers of genes identified for each cancer type; off-diagonal cells are the numbers of genes identified by both the row and the column titles. For example, there were 38 genes related to cancer of the esophagus and nine genes related to both brain and prostate cancer. The greatest number of genes was related to leukemia. The greatest number of genes related to a combination of two cancers was for lymphoma and leukemia. For some combinations of cancer sites, no genes were identified.

The numbers in the off-diagonal cells measure the absolute relatedness of two cancers, but depend on the number of genes related to the corresponding individual cancers. Based on the number of genes related to leukemia and lymphoma individually, the number expected to involve both is 93.8 and the ratio of the observed and expected values is 1.4 (95% CI 1.2–1.6). This indicates there are roughly 40% more genes related to both cancers than would be expected by chance. Table 2 provides a list of 26 combinations of cancer where the number of genes in the siteXsite matrix exceeds the expected number by three. The table indicates that roughly two genes in OMIM should mention both cancer of the esophagus and cancer of the stomach by chance, but the number of genes in OMIM that mention both cancers is 18. This nine-fold discrepancy suggests that cancer of the esophagus and cancer of the stomach are more related than current literature suggests. Similar conclusions might be made for the other cancer combinations in Table 2. Only three of the confidence intervals for the O/E values in Table 2

include 1, indicating that the observed excess of genes for those cancer combinations could easily have occurred by chance.

For groups of three or more cancers, a subset of cells in Table 1 can be used. For example, it might be of interest to consider the genes related to brain, kidney and ovarian cancer. Table 1 indicates 11 genes are related to both brain and ovarian cancer, 19 genes are related to brain and kidney cancer, and 15 genes are related to kidney and ovarian cancer. Several of the genes in those table cells are the same. There are 36 distinct genes related to one or more of brain, kidney and ovarian cancer, and five genes are related to all of them: CDKN2A (OMIM \*600160), BRCA2 (OMIM \*600185), MADD (OMIM \*603584), PTTG1 (OMIM \*604147) and CHEK2 (OMIM \*604373). CGMIM allows users to specify an arbitrary set of cancer types and returns the names of genes related to all cancers in the set.

We randomly selected 25 genes related to cancer and manually reviewed text of the corresponding OMIM entries. All of the entries correctly mention one or more types of cancer, but for 20% of those entries, one of the cancers was only mentioned in the context of evidence suggesting no association.

## Discussion

Our text-mining tool, CGMIM, will assist in designing effective studies for groups of genetically-related cancers. The groups can be identified by genes that are related to several types of cancer, or by combinations of cancer for which there are more related genes than

expected by chance. A group of cancers might be related by physical proximity in the body (e.g., prostate and bladder cancer), a shared physiologic function (e.g., cancers involving the digestive tract), a common exposure (e.g., cancers caused by air pollution) or a common genetic characteristic (e.g., cancers in tissues that express BRCA1). In several ways, groups based on shared genetic factors are anticipated to lead to further etiologic hypotheses and advances regarding environmental agents. First, grouping cancers will be especially useful if a group combines several cancers that are rare and difficult to study individually. Second, genetic pathways might suggest an environmental factor associated with all of the cancers. For example, a grouping defined by a vitamin receptor gene would suggest vitamin intake as a possible environmental agent in the etiology of all of the cancers. Third, a review of public databases will allow us to design studies that will confirm reported associations and possibly extend them to include other cancer sites. The groups can be used to identify cancers that should be considered together in a definition of family history, and in selection of genetic tests that might be adopted in high risk families. During development of the tool, we observed changes in OMIM and the cancer groups that it produced from one week to another. This illustrates the need for a text-mining tool to perform the analysis, as opposed to reporting a static set of results based on the OMIM contents from a particular day. CGMIM is available from the Genome Sciences Centre website at the BC Cancer Agency (<http://www.bcgsc.bc.ca/>).

For each combination of two cancer types, the program produces the number of genes for which the OMIM entry mentions both of the cancers, as well as a ratio of the observed and expected number of genes for the combination. The observed number indicates the absolute relatedness of the cancers. The ratio of observed and expected numbers O/E indicates the extent to which the

cancers might be related in excess of what the current literature suggests. OMIM is based on published material in the scientific literature. The number of genes identified by our program does not necessarily indicate the relatedness of two or more cancer types, but rather what is known about those cancers. This reflects what research has been funded, performed and published. There is more funding for some certain types of cancer, there are more journals that address certain types of cancer, and there are more scientists studying certain types of cancer. More published information about certain types of cancer reflects our knowledge base, and the scientific literature is hence a valid basis for identifying cancer groups and genes for further study. We noticed that not all mentions of cancer in OMIM are based on equal evidence. In some cases, evidence about an association was based on studies of cell lines or non-human organisms. In other cases, evidence is based on anecdotal observations in a small number of people. Some associations are based on several independent studies that each involves hundreds of patients. It is difficult to rank the evidence that supports an association objectively.

There were sentences in OMIM that referred to negative evidence such as " ... is not related to breast cancer." We could not create an algorithm that recognized all negative references without overlooking valid ones. It was not a major problem for our application because an OMIM entry that reports negative evidence often reports positive evidence as well. Because we are interested in any positive evidence of an association, these "mixed" references are correctly included in the tallies. Another problem was sentences in OMIM that described both evidence of gene expression in cancerous and normal tissue. E.g. " ... has been shown to be expressed in breast cancer cells and prostate cells". The sentences are incorrectly interpreted as mentions of prostate cancer. Our manual review of OMIM indicates about 20% of associations with cancer are tallied

based on incorrect interpretations of sentences in OMIM, indicating groups identified by CGMIM should always be verified by a manual review of the evidence. We assume the excess 20% is included in every cell of the siteXsite matrix. Thus expected values based on the matrix also include the 20% excess, and the O/E ratios are not affected.

Other databases might be used as the basis for assessing scientific knowledge regarding genetic cancer groupings, but OMIM offers several advantages. OMIM is based on all publications that appear in the PubMed database and restricted to information about *human* genetics. Results similar to those reports but based on mining PubMed would be of interest, but our purpose was to provide information human genes and cancers only. In addition, the information in OMIM is curated by people with genetic expertise, and results based on unreliable or inaccurate methods are not included. While the database is large, OMIM condenses information from much bigger sources. More specialized cancer groupings might be created using computerized conference proceedings or journal contents. Likewise, a list of synonyms might be determined from other sources such as the UMLS (Unified Medical Language System) Specialist Lexicon of the National Cancer Institute. We used ICD-O terminology for this application because it is the basis for most scientific writing on cancer.

Another method to identify genetically-related cancer is to examine the extent of people diagnosed with multiple primary cancers in a medical registry. This method has strengths and weaknesses that are distinct from a literature-based strategy. A major advantage is that a medical registry is typically restricted to a specific geographical area, and the population is exposed to similar environmental co-factors. On the other hand, someone's survival will determine the

probability that he or she will be diagnosed with a subsequent malignancy, cancers occurring before or after the registry time period will be missed, and some forms of cancer are more likely to be diagnosed following others cancer because of diagnostic, treatment and disease recording practices. For example, men with bladder cancer might be more likely to be diagnosed with prostate cancer because both are treated by a urologist trained to detect cancer at these sites.

This project used resources that have been developed by the US National Institutes of Health and Human Genome Project.<sup>7</sup> Other algorithms that use these resources to determine disease/gene associations have been described.<sup>8,9</sup> Our approach is expected to be exhaustive of the information reported in OMIM, will produce a computer algorithm for near-automatic updating of the review, and has the potential to be extended to other computerized databases. We will use the tool along with other criteria to guide the design of larger studies of genes and environment in cancer etiology.

#### Acknowledgments

Chris Bajdik was supported by a Scholar Award from the Michael Smith Health Research Foundation. This work was also supported by a research grant from the Canadian Cancer Etiology Research Network. We thank Chris Young for helping with manual searches of the OMIM database.

## References

1. Lindor NM, Greene and the Mayo Family Cancer Program. 1998. The concise handbook of family cancer syndromes. *J Natl Cancer Inst* 90:1039-71
2. Hamosh, A., Scott, A.F., Amberger, J., Bocchini, C., Valle, D., and McKusick V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 30:52-5, 2002.
3. Han, J., and Kamber, M. *Data Mining: Concepts and Techniques. First Edition*. Morgan Kaufmann Publishers, 2001.
4. de Bruin, B., and Martin, J. Getting to the (c) core of knowledge: mining biomedical literature. *Int J Medical Informatics*, 67:7-18, 2002.
5. Fritz, A., Percy, C., Jack, A., Shanmugaratnam, K., Sobin, L., Parkin, D.M., and Whelan, S. *International Classification of Diseases for Oncology. Third Edition*. World Health Organization, 2000.
6. NCIC. *National Cancer Institute of Canada: Canadian Cancer Statistics 2003*, Toronto, Canada 2003

7. Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., Schuler, G.D., Schrimi, L.M., Tatusova, T.A., Wagner, L., and Rapp, B.A. Database resources of the Boyer, T.G., Chen, P. and Lee, W. Genome mining for human cancer genes: wherefore art thou? *Trends Mol. Medicine*, 7:187-9, 2001.
8. National Center for Biotechnology Information: 2002 update. *Nucleic Acids Research*, 30:13-6, 2002
9. Perez-Iratxeta, C., Bork, P. and Andrade, M.A. Association of genes to genetically inherited diseases using data mining. *Nature Genetics*, 31:316-9, 2002

## Figure Legend

Figure 1. The number of genes associated with one or more cancers, based on cancers mentioned in Online Mendelian Inheritance in Man (OMIM; <http://www.ncbi.nlm.nih.gov/omim/>) searched October 1, 2003.

Table 1. A siteXsite matrix for 21 major cancer types. Matrix cells indicate the number of genes related to cancers named in the row and column labels. Cell entries are based on cancers mentioned in Online Mendelian Inheritance in Man (OMIM; <http://www.ncbi.nlm.nih.gov/omim/>) searched October 1, 2003.

	B L A D D E R	B R A I N	B R E A S T ( <sup>1</sup> )	C E R V I X	C O L O R E C T A L	E S O P H A G U S	L Y M P H O M A ( <sup>2</sup> )	K I D N E Y	L A R Y N X	L E U K E M I A	L U N G	M O U T H ( <sup>3</sup> )	M Y E L O M A	O V A R Y	P A N C R E A S	P R O S T A T E	M E L A N O M A	S T O M A C H	T E S T I S	T H Y R O I D	U T E R U S ( <sup>4</sup> )	
<b>BLADDER</b>	59	6	26	7	22	7	10	5	0	9	22	1	3	11	10	16	16	11	16	4	4	
<b>BRAIN</b>		107	24	3	31	4	9	19	1	25	21	0	2	11	13	9	19	11	13	6	9	
<b>BREAST<sup>1</sup></b>			383	21	99	13	42	25	1	62	75	8	8	83	30	55	37	29	28	16	30	
<b>CERVIX</b>				43	20	6	6	6	2	10	18	3	2	11	5	7	8	9	4	5	8	
<b>COLORECTAL</b>					362	20	52	29	1	64	72	9	7	46	40	40	36	50	69	22	25	
<b>ESOPHAGUS</b>						38	7	4	0	6	13	3	2	4	6	5	4	18	8	5	2	
<b>LYMPHOMA<sup>2</sup></b>							279	13	1	134	30	3	17	19	9	16	24	17	21	14	13	
<b>KIDNEY</b>								88	1	17	30	3	0	15	12	11	6	13	12	8	7	
<b>LARYNX</b>									6	2	1	2	0	1	1	1	2	2	0	1	2	
<b>LEUKEMIA</b>										614	34	8	17	23	9	22	38	22	27	13	12	
<b>LUNG</b>											228	11	5	34	23	22	32	29	23	24	15	
<b>MOUTH<sup>3</sup></b>												44	0	5	5	6	5	4	7	5	4	
<b>MYELOMA</b>													37	3	1	1	0	3	4	3	0	
<b>OVARY</b>														153	18	27	17	13	23	8	19	
<b>PANCREAS</b>															94	13	14	21	21	13	6	
<b>PROSTATE</b>																140	15	7	16	4	11	
<b>MELANOMA</b>																	198	11	35	9	10	
<b>STOMACH</b>																		108	15	10	11	
<b>TESTIS</b>																			141	6	10	
<b>THYROID</b>																				85	5	
<b>UTERUS<sup>4</sup></b>																						64

Footnotes:

<sup>1</sup> includes male and female breast cancer

<sup>2</sup> includes Hodgkin and non-Hodgkin disease

<sup>3</sup> includes cancers of the lip, tongue, salivary gland, mouth, and pharynx

<sup>4</sup> does not include cervix

Table 2. Combinations of two cancers for which the observed number of genes (O) in the siteXsite matrix exceeds the expected number (E) by three or more. The O/E ratio and a 95% confidence interval (95%CI) are provided. All values of the 95%CI exclude 1.0, except those for larynx – thyroid, larynx – kidney and larynx – pancreas.

Combination of Cancers	<u>Number of Genes Related to Both</u>		O/E Ratio and 95%CI
	Observed (O)	Expected (E)	
cervix – larynx	2	0.1	14.2 ± 5.2
larynx – mouth	2	0.1	13.8 ± 5.1
larynx – uterus	2	0.2	9.5 ± 4.3
esophagus – stomach	18	2.2	8.0 ± 1.3
cervix – esophagus	6	0.9	6.7 ± 2.1
bladder – esophagus	7	1.2	5.7 ± 1.8
larynx – stomach	2	0.4	5.6 ± 3.3
cervix – uterus	8	1.5	5.3 ± 1.6
bladder – cervix	7	1.4	5.0 ± 1.7
pancreas – stomach	21	5.6	3.8 ± 1.8
brain – kidney	19	5.2	3.7 ± 0.9
larynx – thyroid	1	0.3	3.6 ± 3.7
ovary – uterus	19	5.4	3.5 ± 0.8
cervix – stomach	9	2.5	3.5 ± 1.2
bladder – prostate	16	4.5	3.5 ± 0.9
bladder – testis	16	4.6	3.5 ± 0.9
larynx – kidney	1	0.3	3.5 ± 3.6
bladder – pancreas	10	3.0	3.3 ± 1.1
esophagus – mouth	3	0.9	3.3 ± 2.0
larynx – pancreas	1	0.3	3.2 ± 3.5
cervix – lung	18	5.6	3.2 ± 0.8
bladder – stomach	11	3.5	3.2 ± 1.0
larynx – melanoma	2	0.7	3.1 ± 2.4
esophagus – pancreas	6	2.0	3.1 ± 1.4
cervix – ovary	11	3.6	3.1 ± 1.0
lymphoma – myeloma	17	5.7	3.0 ± 0.8

