Visual Analytics for Supporting Evidence-Based Interpretation of Molecular Cytogenomic Findings

Paul Parsons
Computer Graphics Technology
Purdue University
West Lafayette, IN, USA
parsonsp@purdue.edu

Kamran Sedig
Computer Science
Western University
London, ON, Canada
sedig@uwo.ca

Robert. E. Mercer
Computer Science
Western University
London, ON, Canada
mercer@csd.uwo.ca

Maryam Khordad
Computer Science
Western University
London, ON, Canada
mkhordad@uwo.ca

Joan Knoll
Pathology and Laboratory Medicine
Western University
London, ON, Canada
jknoll3@uwo.ca

Peter Rogan
Biochemistry
Computer Science
Western University
London, ON, Canada
progran@uwo.ca

ABSTRACT

Interpreting molecular cytogenomic findings that cover the human genome (e.g., microarray results) is challenging, as it requires accessing and working with multiple, diverse sources of data that are often large and heterogeneous. These data need to be accessed, queried, and simultaneously integrated to achieve open-ended goals, such as interpreting findings to make diagnoses and engage in genetic counselling. Currently, typical workflows of users are laborious, as data sources are often not integrated and must be accessed separately. Furthermore, large document sets often have to be combed through to assist in interpretation. Analytics tools are needed to help users process and distill large bodies of information into manageable sizes so the most relevant portions can be focused on. Current tools typically do not offer support for interactively exploring and engaging with visual representations of important entities and relationships (e.g., chromosomes, gene-phenotype relationships, and scientific articles). We present VErdICT, a visual analytics tool that can support users in their interpretation of molecular cytogenomic findings. A participatory design approach was taken to make VErdICT human-centered. We describe its development, usability and usefulness, and outline some future research challenges.

CCS Concepts

•**Human-centered computing** → **Interaction design; Visual analytics;** •**Computing methodologies** → *Information extraction;* •**Applied computing** → *Genomics; Document management and text processing;*

VAHC '15 Chicago, Illinois USA

DOI: 10.xxx/xxx

Keywords

Visual analytics; complex tasks and activities; genomics; interaction design; scientific document collections; evidence-based decision-making

1. INTRODUCTION

While computational techniques and methods for analyzing health and medical data are improving at a rapid pace, many activities still require human experts to be involved in the analysis. In the domain of molecular cytogenomics, relationships among important entities (e.g., genes and phenotypes) are continually being discovered, and views on the significance of genetic mutations and abnormalities are continually being revised. Databases and catalogs that contain the most current evidence are also updated and revised on a regular basis. As a result, it is impossible to remain fully abreast of relevant developments in the field. This poses a challenge for clinicians and researchers, as findings (e.g., structural variations in a patient's genome) from molecular cytogenomic techniques and analyses must be interpreted within the context of the latest available evidence. Such interpretation is essential for a number of important tasks that geneticists perform, such as patient diagnosis and subsequent genetic counseling.

Cytogeneticists are relied on to interpret and determine the clinical relevance of molecular cytogenomic findings. Due to the high degree of expertise required, it is impossible for such analysis to be done solely by computers; thus, human experts must be part of the analysis process. Additionally, because of the diversity of data sources, their large size, and the rapidity with which they change, it is impossible for humans to analyze the relevant data without computational support. Visual analytics tools (VATs) can fulfill such requirements, as they combine the strengths of humans and computers by integrating computational processing and interactive visual representations [35]. In doing so, joint cognitive systems are formed in which cognitive activities, such as interpreting molecular cytogenomic findings, emerge from an interactive discourse between a user and a VAT [29].

In this paper we present VErdICT (Visual Evidence-based

Interpretation for CyTogenomics), a VAT designed to support interpretation of molecular cytogenomic findings. VErdICT's visual interface provides a single point of access to multiple data sources that are otherwise separate, including a searchable, custom-built index of all documents in the MEDLINE database, and custom gene and phenotype dictionaries that are linked to the index. VErdICT employs interactive visualizations for working with document search results, and for exploring gene-phenotype relations within chromosomal regions.

VErdICT was developed in response to the needs of users within the cytogenomics domain. To make VErdICT human-centered, we took a participatory design approach, directly involving domain experts in the design and implementation process. The design team was highly interdisciplinary, composed of experts in visualization, human-computer interaction, natural language processing, molecular biology, and cytogenomics. Formative evaluation with target users helped to assess the usability and usefulness of VErdICT throughout the development process.

This paper is structured as follows: Section 2 provides background information regarding the users, their tasks, and some considerations for design. Section 3 briefly discusses related work and identifies existing gaps. Section 4 describes the development, components, usability and usefulness, and current limitations of VErdICT. Finally, Section 5 outlines some research challenges and future work.

2. BACKGROUND

Molecular cytogenomics is the field in which aspects of molecular genetics and cytogenetics are combined. Cytogenetics focuses on studying the structure and function of the chromosome, especially with respect to the number, structure, function, and abnormality. The development of fluorescent probes complementary to chromosomal DNA has led to the development of fluorescent in situ hybridization (FISH), a technique that links cytogenetics to molecular genetics [4, 14, 18, 17]. The field of molecular cytogenomics is particularly important for diagnosis of prenatal, postnatal, and acquired chromosomal abnormalities, as well as for genetic counseling. By adopting FISH and other techniques, such as comparative genomic hybridization, cytogeneticists can examine the association between visible chromosome rearrangements and defects at the DNA level.

One important activity in microarray testing is identifying Copy Number Variations (CNVs), which are alterations of DNA in which there is variation in the number of copies of one or more DNA sections. Some detected CNVs have no role in causing disease; some, however, have been associated with diseases such as autism and schizophrenia. While techniques can help identify CNVs, determining their significance is not a simple task—doing so requires expert judgment and interpretation of the findings by a highly skilled cytogeneticist.

The interpretation of CNVs is not a simple task. It requires a review of genes that exist within a chromosomal region, phenotypes that may be related to those genes, various annotations of the human genome, and the relevant scientific literature. This requires accessing and integrating disparate and diverse sources of data. Furthermore, the data sources (e.g., gene databases, phenotype catalogs, scientific articles) can be quite large and are continually changing—often on a daily basis. Due to the dynamic nature of the data, they have to be reviewed regularly to be well informed and to make interpretations consistent with the most current evidence and accepted scientific consensus. This is especially true of the scientific literature, which is generally considered to be the gold standard of data sources. The challenge is compounded by the fact that the published research can contain inconsistencies. For example, CNVs of one particular chromosomal region (16p13.11) have been linked with autism in one paper [33], whereas another proposed that the same CNVs might be benign [11]. A number of such inconsistencies exist, and expert human judgment is required to assess and interpret them.

2.1 Users and Tasks

The target users in which we are interested are geneticists and cytogeneticists, in both research and clinical settings. For the sake of consistency, we will focus primarily on cytogeneticists in clinical settings. The overarching activity in which such users engage is interpreting the findings from patient analysis to determine their clinical significance. In this context, "interpretation" is a high-level label of a complex activity that involves sub-activities such as analytical reasoning, sense making, and decision-making. To interpret the significance of a given set of findings (e.g., CNVs in a patient sample), cytogeneticists perform a number of interrelated tasks. These include, among others, identifying chromosomal regions of interest, assessing the importance of specific structural variations, determining relevant links between genes and phenotypes, and comparing findings with those reported in the literature. Complex activities of this nature do not take place in a linear fashion; thus, VATs should be designed such that users are provided with different interaction mechanisms for working with visual representations in order to carry out such aforementioned tasks [28].

Interpretation can be a time-consuming activity, depending on the complexity and types of CNVs, often requiring several hours to interpret findings from a single patient. Thus, cytogeneticists can often handle results from only a small number of patients in a day. A number of factors contribute to the time it takes to interpret findings: inconsistencies in the literature; rarity of certain findings; constantly changing evidence; uncertain significance of some genomic variants; number and size of data sources that must be accessed and queried; and non-integrated nature of the data sources. This is compounded by the fact that such interpretation cannot be done computationally, as expert human judgment is required at each step of the process. Moreover, this judgment cannot come from any individual, but must rather come from highly trained specialists.

While such characteristics make it impossible to interpret findings in a fully automatic, computational manner, there still is much potential for computational tools to increase the speed and effectiveness of the process. This is especially true for the large, heterogeneous, non-integrated sources of data that must be accessed and used. VATs can provide such computational support, with the added benefit of interactive visual representations of data that can facilitate perceptual and cognitive tasks.

3. RELATED WORK

Existing relevant work can be roughly categorized into three approaches: 1) developing tools that visualize genomic

or phenomic data; 2) developing search literature search interfaces; and 3) developing VATs to work with document collections for specific users and domains.

The first approach has resulted in a number of visualization tools for specific purposes—e.g., genome assemblies [22], sequence variants [8], and phenomic relations [34], and SNP genotype assignments for personalized medicine [36]. Also related are genome browsers (e.g., [13]), which visualize genomes and annotations such as gene predictions, variations, and expression. While tools in this category tend to be valuable for specific users, data, and tasks, they do not integrate the necessary data sources, including the scientific literature, to comprehensively support interpretation of molecular cytogenomic findings.

The second approach has primarily been aimed at providing alternatives to PubMed, a search engine and interface to the MEDLINE (Medical Literature Analysis and Retrieval System Online) database, which contains much of the literature on biology, biochemistry, and other health- and medicine-related fields. Tools in this category are generally not VATs; rather, they use traditional search interfaces and focus on providing different ways of organizing, analyzing, or offering search capabilities on the MEDLINE data, such as using the Gene Ontology in search [3] or summarizing subjects [25]. Lu [20] provides a survey of alternative interfaces to PubMed. Such tools tend to return results of queries as textual lists, often distributed across multiple pages. For simple search tasks, such as finding a specific article, they are sufficient. For more complex and open-ended search tasks, however, this type of strategy can result in information overload, as users have to scan through long lists of results and examine each one. Using interactive visualizations can make search systems more human-centered [12].

The third approach, which is seen mainly in the visual analytics literature, has been aimed at developing VATs to support users in their work with document collections. Such tools are often aimed at specific users in specific domains, such as journalism [2] and intelligence analysis [9]. However, because such approaches are not directed towards the life sciences, they do not integrate the necessary data, extract the necessary entities from text, or properly support the types of tasks outlined in Section 2.1 for activities in molecular cytogenomic contexts.

4. VERDICT

VErdICT is a web-based VAT that has been developed to address the lack of tools specifically supporting interpretation of molecular cytogenomic findings. VErdICT is based on a traditional client-server architecture, in which the server is responsible for most data processing, and the client is responsible for generating visualizations and handling user input, both through textual queries and interactions with the visualizations. An Apache handles different kinds of requests from the client, and performs relevant analytic operations. A separate Apache Solr server functions as a search server, which maintains and handles queries on the document index.

4.1 Development

The main source of scientific articles for the life sciences community is PubMed, which is a search engine and interface to the MEDLINE database. The National Library of Medicine (NLM) hosts the database for free so that it can be downloaded and used for research purposes. MEDLINE consists of article "citations", which are comprised of article metadata, including authors, journal title, Medical Subject Heading (MeSH) keywords, publication date, and so on. Also included in each citation is the abstract text. Currently, the MEDLINE database consists of approximately 24 million citations. In this article, we consider each citation as a document.

We have downloaded the entire MEDLINE database, and have developed a custom index using the open-source Apache Solr/Lucene project. Lucene is essentially an information retrieval library that supports full-text indexing and search functionality. Solr is a search platform that runs on the Lucene index. To rank documents, we are using the well-known term frequency-inverse document frequency (tf-idf) scheme [26]. Lucene also ranks based on an internal similarity measure that generates a vector space model (VSM) score [27], using index terms as dimensions and tf-idf values as weights. In addition to common operations that Solr/Lucene supports, such as stemming and tokenizing, we have developed a UIMA (Unstructured Information Management Architecture) pipeline. The UIMA pipeline can be integrated into the Solr/Lucene indexing process, to extract entities of interest (genes and phenotypes) from the text.

To find our desired entities, Concept Mapper (CM), a high performance dictionary lookup tool, has been used as a component in the UIMA pipeline. We have developed two dictionaries: 1) a phenotype dictionary, based on the Human Phenotype Ontology (HPO [19]), that contains all human phenotype names along with their synonyms; 2) a gene dictionary, based on the standard gene catalog from the Human Genome Nomenclature Committee (HGNC [10]), that contains all gene names in their canonical forms, along with their synonyms and variant names. During indexing, in which all fields of the MEDLINE database are examined, CM looks for matches on dictionary terms and synonyms, and adds a field to the index to indicate a match. For example, the canonical form for a gene is BRCA1. A number of synonyms for BRCA1 are listed in the dictionary (e.g., "BRCA1/BRCA2-containing complex, subunit 1", BRCC1, "Fanconi anemia, complementation group S"). If CM detects the canonical form or any of the synonyms in a document's text, a field will be added to the indexed document to indicate the presence of the BRCA1 gene in the text.

In a second stage of the UIMA pipeline, we perform a form of concept normalization for genes, which can be referred to as gene normalization [7]. Authors often write gene names in short forms. For example, IL3/5 can be used to mean IL3 and IL5; or freac1-freac7 can be used to mean freac1, freac2, and so on, to freac7. Our gene normalization component takes the output from CM, and applies a series of regular expressions to each token that is found in the text, in an attempt to capture the correct genes when such shortcuts are used. The genes that are found are also added to the index.

4.2 Components

From a user's perspective, VErdICT has 3 major components: 1) interactive visual cytogenomic queries, in which users can interact with visual representations of chromosomes, genes, and phenotypes; 2) visual search and document analytics, in which users can search and explore the full set of citations (metadata) from MEDLINE; and 3)

a genome browser, which is a custom implementation of JBrowse [31], that allows users to view and interact with annotations of the human reference genome. Because genome browsers are common and are described in many other places, we will focus only on the first two components.

4.2.1 Interactive Visual Cytogenomic Queries

Without the support of VErdICT, users have to construct queries for the literature in a traditional textual-input fashion. Users may need to initially consult a number of different data sources—e.g., to determine which genes are within a chromosomal region of interest, and then to determine which phenotypes, if any, may be associated with genes in the region. Additionally, when users are not presented with a visual summary of a chromosomal region, it can be easy to miss potentially relevant genes and/or phenotypes. In such cases, where exact queries are not known in advance, visualizations have been shown to help users develop more accurate and relevant queries, especially in the context of searching large document collections [12].

Cytogeneticists often begin an activity with a set of chromosomal regions of interest, which arise from the findings of molecular cytogenomic techniques. Thus, VErdICT allows users to start by selecting the chromosome(s) in which they are interested. Once selected, visual representations of the selected chromosomes that use a familiar style of coloring different chromosomal bands are displayed, as in Figure 1. From this point, users can select a band or other larger or smaller regions of interest on the chromosome.

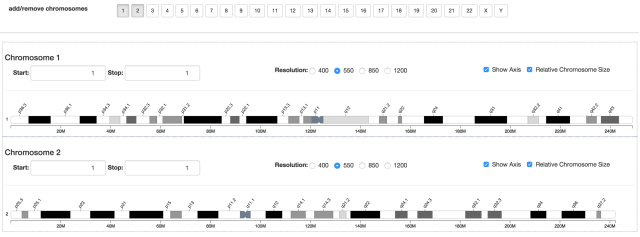

Figure 1: Visual representations of chromosomes. The user has chosen chromosomes 1 and 2.

Users can then drill into the specified region of the chromosome to see the genes within the region. VErdICT presents a visualization of all genes within the region (encoded as orange rectangles in Figure 2). The location and size of genes are encoded within the space according to their location on the chromosome. Genes that inhabit the white region are within the selected area—q12.1, which can be seen in the label above the genes. The gray regions at both ends encode sensitivity information—i.e., genetic information that is slightly outside of the user's query specifications, but can be very useful for exploratory types of tasks [32].

One task that is often required in interpreting molecular cytogenomic findings is the identification of gene-phenotype relations. Typically, users have to consult other catalogs and/or websites (e.g., OMIM, UCSC genome browser) to perform such a task. In VErdICT, however, relations between genes and phenotypes are explicitly encoded and visually represented with links, as shown in Figure 2.

Phenotypes are encoded and colored according to the status of their relationship with a given gene, which is classified in the Online Mendelian Inheritance in Man catalog (OMIM, www.omim.org). The relationship status is encoded using one of 4 colors: pink for relations in which genetic mutations contribute to susceptibility to multifactorial disorders (e.g., diabetes, asthma) or susceptibility to infection (e.g., malaria); green for unconfirmed or possibly spurious mappings from genes to phenotypes; yellow for relations in which genetic variations result in "nondiseases"—genetic variations resulting in apparently abnormal laboratory test values (e.g., dysalbuminemic euthyroidal hyperthyroxinemia); and blue for known causal relations between genetic variations and phenotypes.

At this stage, users can refine their queries by adjusting or entirely changing their selected region on the chromosome. Additionally, they can choose to be taken to external websites that provide more information about genes (genenames.org) or phenotypes (omim.org). Users can also view the selected region in the genome browser component. Finally, users can select any gene or phenotype, and view all of the articles in which it appears in the MEDLINE database. This will take the user to the visual search and document analytics component of VErdICT.

4.2.2 Visual Search and Document Analytics

In this component users can interactively search and explore the set of documents (i.e., scientific articles) from MEDLINE. Users typically come to this component via the cytogenomic visualization component, in which they select a gene or phenotype to view its relevant literature. In such a case, a query is sent from the client to our web server, which then sends a query to the search server to find the relevant documents. For instance, if a user has selected the BRCA1 gene, all articles in which BRCA1 or one of its synonyms appears in the text will be returned from the query. A similar process will take place if a user has selected a phenotype instead of a gene.

Although many genes and phenotypes are mentioned in only a small number of articles, a number of them have been well studied and appear in thousands of articles. As a result, a query can return thousands of documents. As discussed in Section 3, the typical way of displaying a result set of this magnitude is not useful. To address this issue, when results are large, VErdICT employs a clustering algorithm and an interactive visualization that encodes the results of the clustering. Currently, 2 different clustering algorithms are provided. The first is the Lingo algorithm, which looks for meaningful phrases to use as cluster labels, and then assigns documents to the labels to form groups [23]. The second algorithm uses suffix tree clustering (STC), which finds groups of documents sharing a high ratio of frequent phrases; cluster descriptions are a subset of the same phrases used to form the cluster [37]. Generally speaking, the Lingo algorithm provides more meaningful labels, but takes more time than STC, especially when the number of documents is in the thousands. Because the clustering is happening in "real-time"—i.e., it is happening as the user's queries are being submitted—the time it takes has a significant effect on the user experience. Currently, VErdICT defaults to Lingo if the results are fewer than 1,000 and STC if greater than 1,000. Users can change these default settings if they prefer.

Figure 3 shows the result from searching for "BRCA2 and endometrial cancer". Within the time range, which defaults to the past 10 years, there are 41 articles found. A bar graph encodes the number of article matches per year to show the

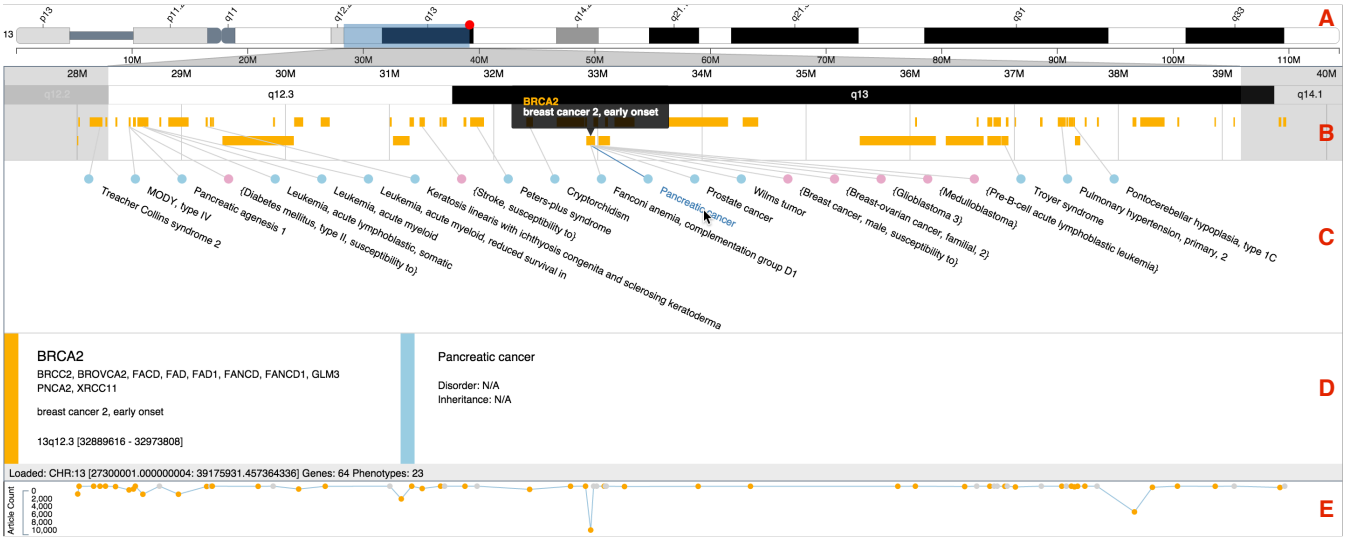

Figure 2: Gene-phenotype relations within a region that the user has selected on chromosome 13. A: visual representation of chromosome 13; B: genes that exist within the specified region of the chromosome; C: phenotypes that have known or possible relations with the displayed genes; D: detail view for a specific gene-phenotype relation. E: article-count view showing number of articles in PubMed/MEDLINE per displayed gene.

user the temporal distribution of results. The blue bars fall within the selected (or default) time range and the gray bars fall outside of the range. This encodes sensitivity information [32] and may help the user to further explore relevant documents. The slider underneath can be used to adjust the range, allowing for dynamic querying of the document index [30]. Below the bar graph are three other bar graphs that encode the values (number of articles) for the top ten phenotypes, genes, and MeSH keywords that appear in the document result set. The number of articles for phenotypes and genes is calculated based on our custom index. MeSH keywords are supplied by authors. Users can interact with each bar, to remove it from the list, go to an external source (e.g., OMIM, HGNC), or view only document subset for a particular phenotype, gene, or keyword. Below this, the result of the document clustering is shown. In this case, the user selected the Lingo clustering algorithm. The 41 documents are categorized into 15 different clusters, each of which is labeled with the cluster name. The size of each cluster encodes the number of documents within it. In Figure 3C, the user has selected the "germline mutations" cluster. When a cluster is selected, the documents within it will populate the list on the right. Each document in the list is tagged with the gene(s), phenotype(s), and keyword(s) that appear within it. For instance, within the selected cluster we can see that the one of the documents has within it the MSH6, MLH1, and other genes. Each document also has 3 buttons that allow the user to add the document to a primary category, add it to a secondary category, or completely remove it from the list. This can help the user to triage the document result set. By clustering the documents, users can usually disregard a number of the clusters, making it easier to find the information they are looking for.

Users can adjust their query at any time in a number of different ways. First, they can adjust the time range that is being considered in the search. There may be a number of reasons why a user would want to do this: perhaps the user is specifically interested in early primary references documenting a gene or phenotype; or s/he may be interested in the most current information that has been published in the past few years. Second, they can adjust other parameters, such as which fields are being considered in the search (e.g., journal, article title). Third, they can change or add to the search box and resubmit the query—e.g., change it to "BRCA2 gene mutation". Any number of Boolean operators can also be used and combined. Fourth, users can go back to the chromosome/gene/phenotype visualization and select another gene or a phenotype. VErdICT has been designed in such a way that states of the tool (i.e., underlying data models, visual representations, and the visual state of the interface) are maintained while users move from one component to another, which can help create a seamless, fluid experience [1, 5]. Thus, users can go back to viewing genes and phenotypes and continue their activity exactly where they previously left off.

4.3 Usability and Usefulness

Usability is typically assessed along three lines: efficiency, effectiveness, and satisfaction. Although a formal summative evaluation has not been completed, we have been conducting ongoing formative evaluations, and there are a number of positive indications of VErdICT's usability. A participatory design and evaluation approach was taken in the development of VErdICT. Two intended users—one a cytogeneticist and one a molecular geneticist—were involved in the design process from the beginning. Through brainstorming, discussion meetings, and feedback from interactive prototype demos, VErdICT was continually refined to meet the needs and expectations of the target audience. A student working on a research project also used VErdICT extensively and provided feedback. While not guaranteeing usability, a participatory approach is known to help identify

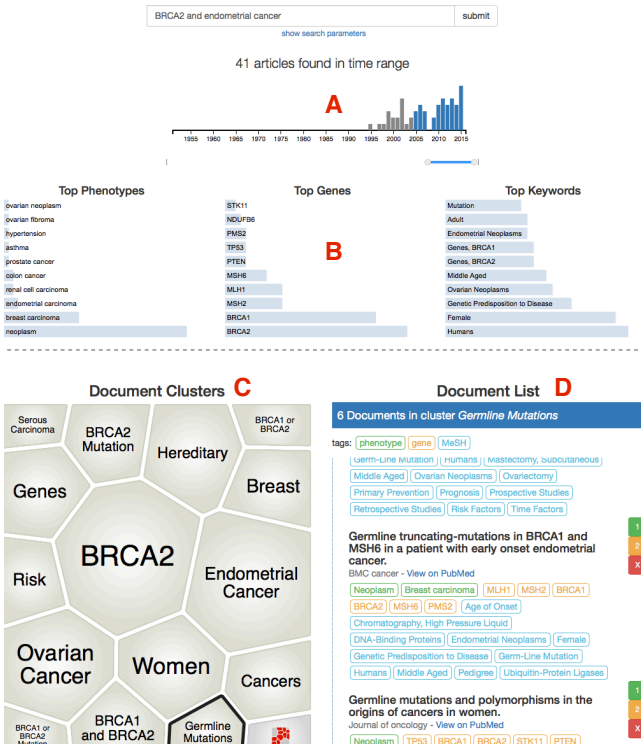

Figure 3: VErdICT interface showing result of searching for "BRCA2 AND endometrial cancer". A: Users see the temporal distribution of results, and can adjust the time range. B: Top phenotypes, genes, and keywords found within the results. C: Document cluster visualization; size encodes number of documents in cluster. D: List of documents in a selected cluser; custom dictionary is used to extract genes and phenotypes in the text.

and fix usability issues during design, and to increase the probability of a usable final product [21]. With respect to satisfaction, we can briefly say that our participatory design approach has led to a number of improvements that help to align VErdICT with the desires and expectations of users.

4.3.1 *Efficiency*

Multiple factors influence the efficiency of user activities. Users typically have to consult multiple sources of data (e.g., HPO, PubMed, genome browsers) to assist in interpretation, which can be laborious. VErdICT integrates a number of data sources and provides a single point of access through its visual interface. For example, users can discover gene-phenotype relations within a chromosomal region (Figure 2), then submit a query to find articles that contain a particular phenotype, and then work with the document result set (Figure 3). Users can go right back to where they were in the chromosome view, as the underlying state of the system is kept persistent throughout an activity. This allows for seamless interaction among various components, and a unified point of access to data that would otherwise be located in different databases or websites. User feedback thus far strongly suggests that VErdICT can increase the efficiency with which activities are performed.

4.3.2 *Effectiveness*

Effectiveness can be examined in terms of how well common tasks are supported. Some tasks that were outlined in Section 2.1 are listed below, with an example of how VErdICT supports each one's performance:

- Identifying chromosomal regions of interest: after searching the literature for a disease or condition, VErdICT identifies the top 10 genes that are found in the results (Figure 3B). Users can select a gene, and be taken to the chromosome visualization showing the region in which the gene exists (Figure 2). Users can explore the region to find potentially relevant genes, phenotypes, and relations between them.

- Assessing importance of specific structural variations: users can select a region in which patient CNVs exist (Figure 2A), find phenotypes that are associated with the region (Figure 2C), and then have VErdICT display articles in which a given phenotype appears (Figure 3). Scientific evidence within the articles helps users assess the degree of importance of a particular CNV with respect to the selected phenotype.

- Determining relevant links between genes and phenotypes: VErdICT explicitly encodes relationships between genes and phenotypes (see Figure 2), helping users to identify potentially important relationships within a region of interest. This information is available in the HPO, but it is more tedious to uncover without visual, spatial encodings of the relationships. Furthermore, VErdICT connects genes and phenotypes to the literature in a seamless manner, allowing users to access detailed information about gene-phenotype relationships to help determine which are relevant.

- Comparing findings with those reported in the literature: users typically have to cross-reference findings (e.g., CNVs) with the most recent published evidence to determine their significance. VErdICT allows users to easily select a chromosomal region in which a CNV exists (Figure 2A), view genes and phenotypes within the region (Figure 2B,C), and see the number of published articles mentioning those genes (Figure 2E). Typically, users would have to consult multiple data sources to perform such a task, but VErdICT supports its performance in a seamless, integrated manner.

Besides the data-integration aspect of VErdICT, another indicator of its effectiveness is with respect to how document search results are displayed. Users typically use PubMed to perform the literature search aspect of their activities. Results from PubMed are displayed in a simple textual list, through which users must scroll to see each article (see Figure 4). VErdICT, on the other hand, provides visual representations of document metadata and document clusters. Using custom dictionaries and NLP techniques during document indexing, phenotypes and genes within documents are identified in the text and displayed to the users. The document cluster visualization can help users incrementally narrow down relevant information and focus on a single concept with a fewer number of documents than the original search—e.g., 9 documents for germline mutations, as in Figure 3D. Preliminary feedback from users suggests that this ability to narrow down increases the effectiveness of tasks performed

using VErdICT compared to using PubMed. Figure 4 shows the result of searching for "BRCA2 AND endometrial cancer" on PubMed, which is the same query that was submitted to VErdICT in Figure 3.

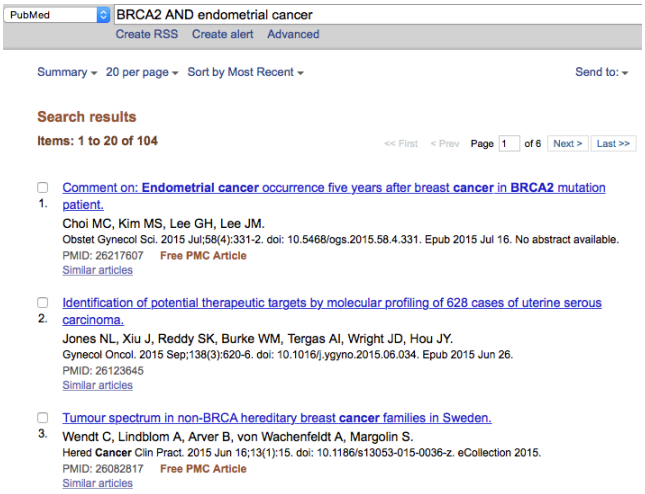

Figure 4: PubMed interface showing result of searching for "BRCA2 AND endometrial cancer".

4.4 Current Limitations

Two limitations of VErdICT should be noted. First, we are working only with article citations—including abstract texts—and not full texts. Thus, important information that does not appear in the abstracts or metadata is missed. Second, gene normalization is currently returning many false positives. This is due to the ambiguous nature of acronyms used by authors. No solution to this problem currently exists, but two of the authors are working on an NLP solution.

5. CHALLENGES AND FUTURE WORK

While developing VErdICT and engaging in this line of research, a number of research challenges and opportunities for future work have come to light:

- Natural Language Processing: Currently, relations in the text are identified based only on co-occurrence. Some previous research has focused on identifying implicit gene-phenotype relations within text [15, 16]. Future work should integrate this type of sophisticated relationship extraction into the UIMA pipeline.
- Clustering Algorithms and Analytics Models: While VErdICT provides two different algorithms for clustering document query results, users cannot interact with the underlying models or algorithms to adjust their parameters. Providing such features may help to increase the quality of human-VAT coupling [6, 24].
- Ontologies: Although we have made use of the HPO to support document search, we have discovered that complex ontologies can be difficult to understand, and even expert users often do not have accurate mental models of them. Future research is needed to explore this issue and determine how complex ontologies can be best linked to document collections.
- Interactive Triaging: Helping users focus on relevant information within large collections of documents requires further work. There is an need to develop interactive visual analytics techniques to support iterative, nested triaging to help users quickly navigate through and make sense of document collections at various levels. Such work could certainly be generalized to other domains in which large document collections are used.

6. ACKNOWLEDGMENTS

The authors would like to thank the IBM Canada Research and Development Centre, the SOSCIP consortium, Cytognomix, and the Natural Sciences and Engineering Research Council of Canada for their support.

7. REFERENCES

[1] B. B. Bederson. Interfaces for staying in the flow. *Ubiquity*, 2004(September):1–1, 2004.

[2] M. Brehmer, S. Ingram, J. Stray, and T. Munzner. Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool for Investigative Journalists. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2271–2280, 2014.

[3] A. Doms and M. Schroeder. GoPubMed: Exploring PubMed with the gene ontology. *Nucleic Acids Research*, 33(SUPPL. 2), 2005.

[4] S. N. Dorman, B. C. Shirley, J. H. M. Knoll, and P. K. Rogan. Expanding probe repertoire and improving reproducibility in human genomic hybridization. *Nucleic Acids Research*, 41(7), 2013.

[5] N. Elmqvist, a. V. Moere, H.-C. Jetter, D. Cernea, H. Reiterer, and T. Jankun-Kelly. Fluid interaction for information visualization. *Information Visualization*, 10(4):327–340, 2011.

[6] A. Endert, C. North, R. Chang, and M. Zhou. Toward Usable Interactive Analytics: Coupling Cognition and Computation. In *KDD 2014 Workshop on Interactive Data Exploration and Analytics (IDEA)*, 2014.

[7] H. Fang, K. Murphy, Y. Jin, and J. Kim. Human gene name normalization using text matching with automatically extracted synonym dictionaries. *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*, pages 41–48, 2006.

[8] J. A. Ferstay, C. B. Nielsen, and T. Munzner. Variant view: Visualizing sequence variants in their gene context. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2546–2555, 2013.

[9] C. Görg, Z. Liu, J. Kihm, J. Choo, H. Park, and J. Stasko. Combining computational analyses and interactive visualization for document exploration and sensemaking in jigsaw. *IEEE Transactions on Visualization and Computer Graphics*, 19(10):1646–1663, 2013.

[10] K. A. Gray, L. C. Daugherty, S. M. Gordon, R. L. Seal, M. W. Wright, and E. A. Bruford. Genenames.org: The HGNC resources in 2013. *Nucleic Acids Research*, 41(D1), 2013.

[11] F. D. Hannes, A. J. Sharp, H. C. Mefford, T. de Ravel, C. A. Ruivenkamp, M. H. Breuning, J.-P. Fryns, K. Devriendt, G. Van Buggenhout, A. Vogels,

H. Stewart, R. C. Hennekam, G. M. Cooper, R. Regan, S. J. L. Knight, E. E. Eichler, and J. R. Vermeesch. Recurrent reciprocal deletions and duplications of 16p13.11: the deletion is a risk factor for MR/MCA while the duplication may be a rare benign variant. *Journal of Medical Genetics*, 46(4):223–32, apr 2009.

[12] O. Hoeber. Visual Search Analytics: Combining Machine Learning and Interactive Visualization to Support Human-Centred Search. In *Beyond Single-Shot Text Queries: Bridging the Gap(s) Between Research Communities*, pages 37–43, 2014.

[13] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and a. D. Haussler. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, 2002.

[14] W. A. Khan, J. H. Knoll, and P. K. Rogan. Context-based FISH localization of genomic rearrangements within chromosome 15q11.2q13 duplicons. *Molecular Cytogenetics*, 4(1):15, jan 2011.

[15] M. Khordad. Investigating Genotype-Phenotype relationship extraction from biomedical text. Unpublished Doctoral Dissertation, Western University, Canada, 2014.

[16] M. Khordad and R. E. Mercer. Identifying genotype-phenotype relationships in biomedical text. *BMC Bioinformatics*, submitted.

[17] J. H. M. Knoll, P. Lichter, K. Bakdounes, and I.-E. A. Eltoum. In situ hybridization and detection using nonisotopic probes. *Current Protocols in Molecular Biology*, Chapter 14:Unit 14.7, 2007.

[18] J. H. M. Knoll and P. K. Rogan. Sequence-based, in situ detection of chromosomal abnormalities at high resolution. *American Journal of Medical Genetics. Part A*, 121A(3):245–57, 2003.

[19] S. Köhler, S. C. Doelken, C. J. Mungall, S. Bauer, H. V. Firth, I. Bailleul-Forestier, G. C. M. Black, D. L. Brown, M. Brudno, J. Campbell, D. R. Fitzpatrick, J. T. Eppig, A. P. Jackson, K. Freson, M. Girdea, I. Helbig, J. A. Hurst, J. Jähn, L. G. Jackson, A. M. Kelly, D. H. Ledbetter, S. Mansour, C. L. Martin, C. Moss, A. Mumford, W. H. Ouwehand, S. M. Park, E. R. Riggs, R. H. Scott, S. Sisodiya, S. V. Vooren, R. J. Wapner, A. O. M. Wilkie, C. F. Wright, A. T. Vulto-Van Silfhout, N. D. Leeuw, B. B. A. De Vries, N. L. Washingthon, C. L. Smith, M. Westerfield, P. Schofield, B. J. Ruef, G. V. Gkoutos, M. Haendel, D. Smedley, S. E. Lewis, and P. N. Robinson. The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42(D1), 2014.

[20] Z. Lu. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database: The Journal of Biological Databases and Curation*, 2011(0):baq036, 2011.

[21] M. Maguire. Methods to support human-centred design. *International Journal of Human-Computer Studies*, 55(4):587–634, 2001.

[22] C. B. Nielsen, S. D. Jackman, I. Birol, and S. J. M. Jones. ABySS-explorer: Visualizing genome sequence assemblies. In *IEEE Transactions on Visualization and Computer Graphics*, volume 15, pages 881–888, 2009.

[23] S. Osinski and D. Weiss. A concept-driven algorithm for clustering search results. *IEEE Intelligent Systems*, 20(3):48–54, 2005.

[24] P. Parsons, K. Sedig, A. Didandeh, and A. Khosravi. Interactivity in Visual Analytics: Use of Conceptual Frameworks to Support Human-Centered Design of a Decision-Support Tool. In *Hawaii International Conference on System Sciences*, pages 1138–1147. IEEE, 2015.

[25] C. Perez-Iratxeta, A. J. Pérez, P. Bork, and M. A. Andrade. Update on XplorMed: A web server for exploring scientific literature. *Nucleic Acids Research*, 31(13):3866–3868, 2003.

[26] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[27] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[28] K. Sedig and P. Parsons. Interaction Design for Complex Cognitive Activities with Visual Representations: A Pattern-Based Approach. *AIS Transactions on Human-Computer Interaction*, 5(2):84–133, 2013.

[29] K. Sedig, P. Parsons, and A. Babanski. Towards a characterization of interactivity in visual analytics. *Journal of Multimedia Processing Technologies*, 3(1):12–28, 2012.

[30] B. Shneiderman. Dynamic queries for visual information seeking. *IEEE Software*, 11(6):70–77, 1994.

[31] M. E. Skinner, A. V. Uzilov, L. D. Stein, C. J. Mungall, and I. H. Holmes. JBrowse: A next-generation genome browser. *Genome Research*, 19(9):1630–1638, 2009.

[32] R. Spence. Sensitivity encoding to support information space navigation: a design guideline. *Information Visualization*, 1(2):120–129, 2002.

[33] R. Ullmann, G. Turner, M. Kirchhoff, W. Chen, B. Tonge, C. Rosenberg, M. Field, A. M. Vianna-Morgante, L. Christie, A. C. Krepischi-Santos, L. Banna, A. V. Brereton, A. Hill, A.-M. Bisgaard, I. Müller, C. Hultschig, F. Erdogan, G. Wieczorek, and H. H. Ropers. Array CGH identifies reciprocal 16p13.1 duplications and deletions that predispose to autism and/or mental retardation. *Human Mutation*, 28(7):674–82, jul 2007.

[34] J. L. Warner, J. C. Denny, D. A. Kreda, and G. Alterovitz. Seeing the forest through the trees: uncovering phenomic complexity through interactive network visualization. *Journal of the American Medical Informatics Association*, 22(2):324–9, 2015.

[35] P. C. Wong and J. Thomas. Visual analytics. *IEEE Computer Graphics and Applications*, 24(5):20–21, 2004.

[36] E. Worbis, R. Machiraju, C. Bartlett, and W. Ray. Visual Interactive Quality Assurance of Personalized Medicine Data and Treatment Subtype Assignment. In *Workshop on Visual Analytics in Healthcare*, pages 33–36, 2011.

[37] O. Zamir and O. Etzioni. Grouper: A dynamic clustering interface to Web search results. *Computer Networks*, 31(11):1361–1374, 1999.