GenomeSnip: Fragmenting the Genomic Wheel to augment discovery in cancer research

Maulik R. Kamdar[1], Aftab Iqbal[1], Muhammad Saleem[2], Helena F. Deus[3], and Stefan Decker[1]

[1] Insight Center for Data Analytics, National University of Ireland, Galway
{maulik.kamdar,aftab.iqbal,stefan.decker}@deri.org
[2] Universität Leipzig, IFI/AKSW, PO 100920, D-04009 Leipzig
saleem@informatik.uni-leipzig.de
[3] Foundation Medicine Inc., 150 Second Street, Cambridge, MA 02141
hdeus@foundationmedicine.com

Abstract. Cancer genomics research has greatly benefited from high-throughput technologies for the characterization of genomic alterations in patients. These voluminous genomics datasets when supplemented with the appropriate computational tools have led towards the identification of *'oncogenes'* and cancer pathways. However, if a researcher wishes to exploit the datasets in conjunction with this extracted knowledge his cognitive abilities need to be augmented through advanced visualizations. In this paper, we present GenomeSnip, a visual analytics platform, which facilitates the intuitive exploration of the human genome and displays the relationships between different genomic features. Knowledge, pertaining to the hierarchical categorization of the human genome, *oncogenes* and abstract, co-occurring relations, has been retrieved from multiple data sources and transformed *a priori*. We demonstrate how cancer experts could use this platform to interactively isolate genes or relations of interest and perform a comparative analysis on the 20.4 billion triples Linked Cancer Genome Atlas (TCGA) datasets.

Keywords: Visualization, Cancer Research, Integrative Genomics, Linked Data, Visual Knowledge Exploration

1 Introduction

1.1 Integrative Genomics

Since the completion of the Human Genome Project in 2003 and the rising influence of high-throughput gene sequencing technologies, the biomedical domain has witnessed a huge increase in genomics data. Advanced computation analysis of these datasets, popularly grouped under Genome-wide Association studies (GWAS) [20], allows the identification of several susceptible loci associated with various types of cancer [8,32]. The knowledge on point alterations and *oncogenes*, so extracted and catalogued, are available through multiple data sources [10,11,29]. While GWAS have been successful to characterize isolated genomic

loci, network-based approaches [31] have also been devised to examine whether a group of genes are implicated together in a disease of interest. Most of these approaches rely on the coincidental, yet synchronized, involvement of a set of genes to reveal interesting functional relationships [18,13]. 'Co-occurrence' of a set of genes does not mandatorily imply physical 'interactions' of their products, but rather macro-molecular associations on a higher, abstract level i.e. pathways and diseases. As the volume and the heterogeneity of genomics data rises, providing scalable, integrated solutions for data management becomes difficult.

To address these integrative challenges, experts have resorted to Linked Data and Semantic Web Technologies [6] for publishing biomedical datasets using Resource Description Framework (RDF) [2,24] to form the Life Sciences Linked Open Data (LSLOD) Cloud. The newest member on the LSLOD Cloud is the Linked Cancer Genome Atlas (TCGA) [26], the RDFized version of the Cancer Genome Atlas[4], a 20.4 billion triples data source characterizing the genomic and clinical profiles of cancer patients. Linked TCGA contains information related to the genomic alterations like methylations, gene expression changes, single nucleotide polymorphisms (SNPs) and copy-number changes (CNVs), mapped to the genomic loci. Along with the ease of representation and querying, usage of these technologies have presented advantages like integration with other datasets, data mining and knowledge extraction [25,3]. Even though it has become commonplace for computer scientists to use semantic web technologies, the cognitive processes of a biomedical researcher need to be augmented through the use of improved, interactive, intuitive, visual approaches [14].

1.2 Genomic Visualization

Genome browsers [16] have provided the simplest, yet effective mode of navigation through genomic datasets in an intuitive fashion. One of the most common methods of implementation is to map the data points against the genomic coordinates and visualize the datasets as charts or heatmaps [27]. Whereas linear genome browsers have been very common until now [23,9,21], a new category of circular visualizations is emerging to meet the needs of comparative genomics [19]. Even as circular plots overcome the cognitive barriers for grasping relationships between disjoint genomic features on an abstract level, the applications miss the 'interactive' essence and compel the biomedical researcher to focus on the entire genome.

Analyzing genomic datasets using the knowledge extracted through GWAS has become of vital importance for the discovery of newer tumour risk hypotheses and diagnosing cancer on a personalized basis [17,5]. On the other hand, studying co-occurrence networks of genes present in common operons could lead to the prediction of hidden protein interactions and functions [12]. Advanced genomic visualization approaches, which infuse these insights into cancer research, still need to be perfected to augment discovery.

[4] `http://cancergenome.nih.gov/`

In this paper, we present our approach towards the development of a semantic visual analytics prototype, GenomeSnip, for the intuitive exploration of the human genome and the interactive analysis of cancer datasets in conjunction with insights from GWAS and co-occurrence. Knowledge was retrieved from multiple data sources, pertaining to the classification of the human genome, relationships between different genomic features, established *oncogenes* and somatic alterations. We discuss how we generate the co-occurrence data cube, based on which genes are implicated together in the same pathway/disease, or are mentioned in the same publication. We then demonstrate how these datasets are assembled into a single, interactive visualization, the 'Genomic Wheel', which displays the chromosomes as arcs in a circular layout and the chromosome co-occurrence data cube as an overlay of chords. We finally showcase a comparative evaluation against other state-of-the-art genomic visualization tools and discuss the applicability of GenomeSnip using the Linked TCGA datasets.

2 Methodology

The GenomeSnip platform was conceptualized with the idea of 'snipping' or clipping the human genome informatively in fragments through interaction with an aggregative, circular visualization, the 'Genomic Wheel', and introspectively analyzing those fragments in a 'Genomic Tracks' display.

2.1 Integration of Data Sources

The 'Genomic Wheel' comprises of two sets of geometrical structures, **arcs** which form the perimeter of the 'Genomic Wheel' and **chords** which connect two arcs. The information used to render these geometrical structures is derived from a multitude of data sources and combined to form the 'Genomic Wheel'.

Arcs Arcs illustrate the rich, hierarchical classification as subsequent layers.

1. **UCSC Genome Browser:** Ideograms are a schematic representation to depict staining patterns on a tightly-coiled chromosome. These Chromosome Bands (Ideograms) were downloaded from the Mapping and Sequencing Tracks Table in the Human Genome Assembly (GRCh37/hg19, Feb 2009), available at the UCSC Genome Browser[5] [16].
2. **CellBase:** Apart from the coordinates and descriptions of the protein-coding genes, the REST API[6] exposed by CellBase [4] also provides external information on genomic variants like cancer-related mutations [10] and SNPs [29] along the human genome. The genes are annotated using the HGNC Nomenclature [22] and the positions are indicated by start/stop attributes.

GenomeSnip also incorporates the Cancer Gene Census[7] [11], a catalogue of *oncogenes*, which bear somatic and germline mutations in their sequences.

[5] `http://genome.ucsc.edu/`
[6] `http://docs.bioinfo.cipf.es/projects/cellbase/wiki/`
[7] `http://cancer.sanger.ac.uk/cancergenome/projects/census/`

Chords Genes may be involved in various pathways in the form of protein inputs or catalysts, implicated in a certain disease, or mentioned in publications.

1. **UniProt:** The UniProt dataset exposed by the EBI RDF Platform[8] provides disease-gene mappings on the execution of the question *'What are the preferred gene name and disease annotations of all human UniProt entries that are known to be involved in a disease?'* as a SPARQL Query.
2. **Kyoto Encyclopedia of Genes and Genomes (KEGG):** KEGG [15] is exposed as a SPARQL Endpoint [3] for providing pathway-gene linkages.
3. **Pubmed2Ensembl:** Pubmed2Ensembl[9] is a customized service extended to provide gene-related publication information. [1].

2.2 Generating Co-occurrence Data Cubes and Similarity Measures

The mappings of genes with external resources were extracted from the **Chord** data sources *a priori* to instantiate naive matching between a pair of genes. In certain cases, conversion of the native identifiers (Ensembl, EntrezGene, KEGG GeneIds) of the genes to the HGNC nomenclature was carried out using the HGNC website [28]. For instance, if a pathway resource indicates that a set of 10 genes are involved, then we establish a co-occurrence pair between all possible pairs of genes in the set. Hence, we would have 45, i.e. n(n - 1)/2 co-occurrence pairs. It can then be inferred that a pair of genes co-occurs together in a certain number of pathways as well as diseases and publications.

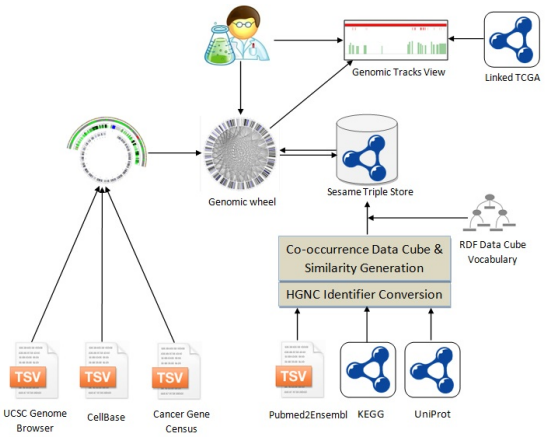

Fig. 1. GenomeSnip Architecture

A co-occurrence data cube of 3 dimensions (segment 1, segment 2 and data source type) and 2 measures (co-occurrence count and names of mapped resources) is created between the 20000 genes extracted from CellBase. We also store the total number of co-occurrence pairs generated for each chromosome and the entire human genome (HG). This approach was extended to the upper levels of the genomic heirarchy, based on the location of the genes, and we obtained similar matrices indicating the co-occurrence of a pair of ideograms and subsequently chromosomes. A slice of the chromosome data cube (shown in Table 1) describes the total number of co-occurrence pairs between two chromosomes, obtained from diseases (*Dis*), pathways (*Path*) and publications (*Pub*).

[8] `http://www.ebi.ac.uk/rdf/`
[9] `http://pubmed2ensembl56.smith.man.ac.uk/`

Table 1. Slice of the Chromosome co-occurrence Data Cube.

	Chr 1			Chr 2			Chr 3			...	Total		
	Dis	Path	Pub	Dis	Path	Pub	Dis	Path	Pub	...	Dis	Path	Pub
Chr 1	28	4049	4522	**15**	**4251**	**2915**	19	4811	2667	...	226	75044	49827
Chr 2	15	4251	2915	14	1662	3589	14	2532	1748	...	155	42513	31424
Chr 3	19	4811	2667	14	2532	1748	16	1338	1390	...	224	44310	26817

The data cubes are transformed to RDF (N-triples syntax) using the RDF Data Cube Vocabulary [7] and are stored in a Triple Store using the Sesame API. The 'Genomic Wheel' retrieves slices of these data cubes for its assembly (Fig. 1).

We also calculate the similarity between any two chromosomes based on their co-occurrence data. This calculation is inspired from Tversky's feature-based similarity measure [30], where the similarity between two entities is a weight-based summation of their common features. For example, the similarity index between Chromosome 1 and 2 is calculated as shown in Equation 1. The total number of co-occurrence pairs between Chromosome 1 and 2, retrieved from either diseases, pathways and publications (marked cells in Table 1), is divided by the total number of co-occurrence pairs registered across the entire human genome from that particular data source (for instance, $HG_{TotalDis}$). α, β, γ are the weights assigned to these fractions respectively.

$$Sim_{12} = \alpha * \frac{Chr_1 \bigcap Chr_2|_{Dis}}{HG_{TotalDis}} + \beta * \frac{Chr_1 \bigcap Chr_2|_{Path}}{HG_{TotalPath}} + \gamma * \frac{Chr_1 \bigcap Chr_2|_{Pub}}{HG_{TotalPub}} \quad (1)$$

To calculate the relative similarity impact on the side of Chromosome 1, we calculate the maximum similarity measure possible for Chromosome 1 using the values of the last three columns in Table 1 (shown in Equation 2). We then divide the similarity index previously calculated with this measure.

$$Sim_{1Max} = \alpha * \frac{Chr_{1TotalDis}}{HG_{TotalDis}} + \beta * \frac{Chr_{1TotalPath}}{HG_{TotalPath}} + \gamma * \frac{Chr_{1TotalPub}}{HG_{TotalPub}} \quad (2)$$

2.3 Technologies and Availability

The GenomeSnip platform is a web-based client application developed using native web technologies like HTML5 Canvas, JavaScript and JSON. The advent of HTML5 in the recent years allows the application to remove the dependence on proprietary frameworks like Adobe Flash and Silverlight for interactivity, whereas the support across traditional browsers improves the interoperability. Whereas visualization libraries like D3JS[10], which primarily rely on SVG, are suitable for developing interactive visualizations for smaller datasets, the functionality is deeply impacted when rendering larger datasets as SVG stores the rendered objects directly in the browser DOM. Hence, it was prudent to use HTML5 Canvas, which creates a raster graphic of the entire visualization prior

[10] `http://d3js.org`

to rendering in the browser window. GenomeSnip uses the KineticJS[11] library, an HTML5 Canvas JavaScript framework, that enables node nesting, layering, caching and event handling.

In essence, GenomeSnip could be deployed from any Apache Server with PHP5 and PHP-CURL support enabled. The platform communicates with the Sesame Triple Store and the TCGA Endpoints using the SPARQL 1.1 protocol and retrieves the results in JSON format. A public instance of GenomeSnip is deployed at `http://srvgal78.deri.ie/genomeSnip` and could be accessed using any modern browser with HTML5 support like Google Chrome 8+, Mozilla Firefox 4.0+ and Internet Explorer 10. The distribution of the Linked TCGA datasets and its respective endpoints are available at `http://tcga.deri.ie`.

3 GenomeSnip Platform

GenomeSnip is a semantic, visual analytics prototype devised to expedite knowledge exploration and discovery in cancer research. Our approach relies on the knowledge retrieval from various data sources to determine the co-occurrence matrices between different genomic features and fuse these generated insights with genomic datasets into an aggregated, interactive visualization. We have combined the salient features of the linear genome browsers and circular plots, and present an interactive alternative displaying only those genomic regions and its inherent relations which are of actual interest to the cancer expert.

3.1 Genomic Wheel

The human genome is laid in a circular layout with different chromosomes forming the arcs on the perimeter of the 'Genomic Wheel'. The size of each chromosome is directly proportional to the arc length. The hierarchical categorization of each chromosome forms the subsequent layers in the representative arc. Currently, the visualization takes into account four levels of this rich hierarchy - chromosome, ideogram, gene and cancer point mutations along the sequence of each gene. Due to the rich genetic diversity, the intial stage of the 'Genomic Wheel' hides the 'gene' and the 'point mutations' layers and displays only the 'chromosome' and 'ideogram' layer, as shown in Fig. 2(A). By clicking down on each arc, the represented chromosome is highlighted and flares out to display the subsequent layers. The final stage of the visualization depicts all the layers, as shown in Fig. 2(B). Those genes catalogued in the Cancer Gene Census, whose gene sequences bear somatic and germline mutations responsible for cancer, are represented using different shades of red to allow the cancer researcher to intuitively differentiate between them. Hovering the mouse pointer over any gene displays an information box, showing extra information on the gene.

Chords connect different components of the human genome based on the co-occurrence data cubes and the similarity measures generated. At the chromosome level, the thickness of the chords is proportional to the relative similarity

[11] `http://kineticjs.com/`

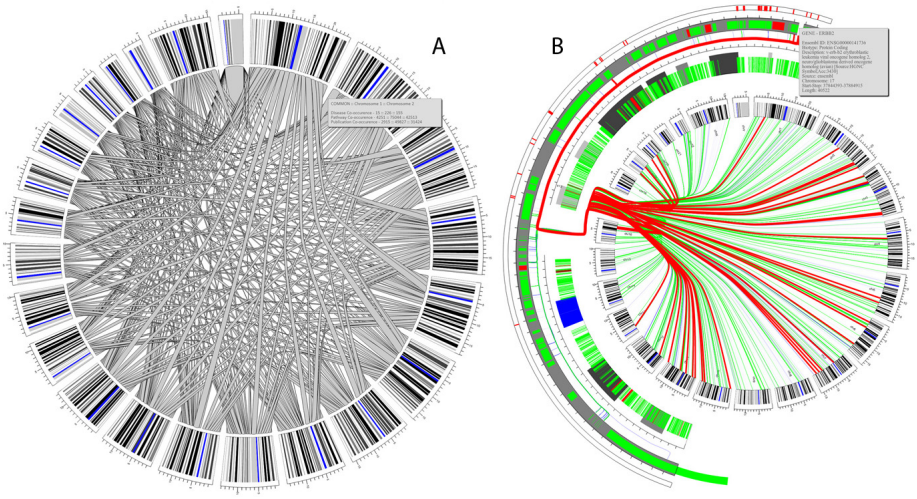

Fig. 2. Genomic Wheel - Initial Stage (A) and Final Stage (B)

impact between two chromosomes i.e. the chord tapers between the connected chromosomes based on the similarity impact on each side. The values of α, β and γ used for generating the similarity measures in the initial layout is 1, 0.4 and 0.1 respectively. Hovering the mouse over each chord displays the slice of co-occurrence matrix. At the genetic level the relation chords are represented using distinct colors (Red, green and blue for diseases, pathways and publications respectively) to enable visual discernibility. Hovering over these chords display the diseases, pathways or publications in which the connecting genes co-occur.

3.2 Genomic Tracks

On clicking any gene, the 'Genomic Tracks' display is launched with the instance of the clicked gene visualized. As the public prototype of GenomeSnip is configured for interactive analysis of the Linked TCGA datasets, the cancer researcher has the option to select any tumor category and load the DNA methylation and the exon expression datasets of the patients diagnosed with that tumor. Selection of the patient executes the SPARQL Query (shown below) against the corresponding TCGA Endpoints and retrieves his sequencing results in real-time. These datasets are represented using bar

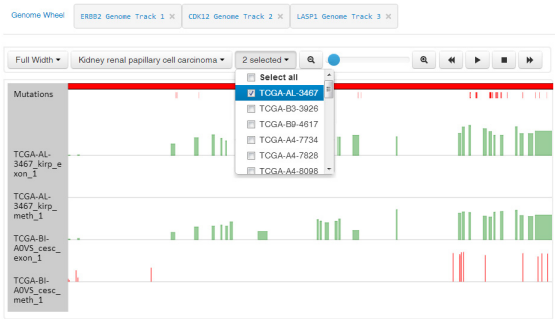

Fig. 3. Genomic Tracks View

charts (red and green respectively), whose X-coordinates are mapped to the genomic coordinates of the gene and the Y-coordinates indicate the beta value or the RKPM value at that chromosomal position. The cancer researcher has the option to launch multiple genes simultaneously in separate tabular panels as shown in Fig. 3, and perform a comparative analysis. Clicking on a relation chord linking two genes in the 'Genomic Wheel' launches the connected genes in separate panels simultaneously. The 'Genomic Tracks' interface provides the basic features, for zooming and panning across the length of the clicked gene.

SPARQL Query . Exon Expression results for a patient for the ERBB2 gene

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX tcga: <http://tcga.deri.ie/schema/>
SELECT DISTINCT * WHERE {
<PatientID> tcga:result ?exonResult.
?exonResult tcga:chromosome ''17''; tcga:RPKM ?value;
            tcga:start ?start; tcga:stop ?stop
FILTER(xsd:double(?start) > 37844393 &&
       xsd:double(?stop) < 37884915) }
```

4 Comparative Evaluation of Genomic Visualizations

We carried out a preliminary evaluation to compare the features of GenomeSnip against a few of the popular, free genomic visualization tools available. Insights were gathered through a questionnaire (`http://goo.gl/vnLtX4`) embedded within the GenomeSnip platform from 4 PhD students and 5 researchers. Evaluators from the Biotechnology domain were accustomed with the UCSC Genome Browser [16] and the Integrative Genomics Viewer (IGV) [23] due to usage during their research in protein engineering. Evaluators researching in Bioinformatics were familiar with modern genomic visualization tools.

A summarized list of features compared across the different solutions is shown in Table 2. Whereas browsers like Savant [9] and IGV are established desktop applications to visualize the human genome on a linear scale, they are unable to display genomic relationships on an abstract level (location independent). They provide extensive application programming interfaces (APIs) for the integration of third-party knowledge or visualization modules, but it necessitates manual intervention. On the other hand, Circos [19] renders high-resolution, static images of circular plots for the use in comparative genomics, but these cannot be dynamically updated to focus on any genomic region of interest. In terms of feature comparison, TCGA's Regulome Explorer[12] is closely related to GenomeSnip, with additional visualization of datasets in the form of networks and scatter plots and an example on the integration of PubMed literature for analysis. However, no information was currently available on the integration of GWAS insights into the platform and simultaneous analysis of the datasets.

[12] `http://explorer.cancerregulome.org/`

Table 2. Comparative Evaluation against popular genomic visualization applications.

	UCSC	IGV	Savant	GenomeMaps	Circos	Regulome	GenomeSnip
Linear Coordinates	✓	✓	✓	✓	✗	✓	✓
Circular Plot	✗	✗	✗	✗	✓	✓	✓
Web Interface	✓	✗	✗	✓	✗	✓	✓
Data Upload	✗	✓	✓	✓	✗	✗	✗
Third-party modules	✗	✓	✓	✗	✗	✗	✗
TCGA Demonstration	✓	✓	✗	✗	✓	✓	✓
GWAS Insights	✗	✗	✗	✗	✓	✗	✓
Chromosomal Relations	✗	✗	✗	✗	✓	✓	✓
Simultaneous Analysis	✓	✓	✗	✗	✓	✗	✓
Knowledge Integration	✗	✓	✓	✗	✓	✓	✓
Other Visualizations	Heatmap	Heatmap	✗	✗	✗	Network	✗
Dynamicity	✓	✓	✓	✓	✗	✓	✓

5 Discussion

With the rise of high-throughput gene sequencing technologies, data analysis has replaced data generation as the rate-limiting step for the interpretation of genomic patterns and discovery of newer insights. Most of the popular genome browsers provide navigation across the human genome in a linear fashion. Whereas automated mining tools are more adept towards the linear genomic analysis, the perceptive faculties of humans are more developed towards interpreting patterns in depth compared to length. Moreover, linear visualizations fail to account for inter- and intra-chromosomal relations, which could be easily interpreted and used by humans but difficult for machines. Our approach leading towards the development of GenomeSnip was inspired from circular plots but it overcomes the shortcomings of popular circular visualization tools available. As the assembly of the 'Genomic Wheel' is comprised of the hierarchical classification of the human genome and the co-occurrence relationships, represented as RDF Data Cubes, the ability to integrate other data sources is leveraged and information like disease-associated mutations can be easily visualized.

5.1 Applicability

1. **Formulating Improved Hypotheses:** By integrating the extracted knowledge pertaining to established cancer genes and pathways to the largest dataset of the LSLOD Cloud, and making it accessible through an intuitive, interactive visualization, the GenomeSnip Browser would allow cancer experts to formulate newer risk hypotheses. Interpreting these insights in the context of available knowledge in literature and pathways and isolating genomic segments of interests, cancer experts could perform a comparative analysis using the Linked TCGA datasets as a training set and visually validate their hypotheses.
2. **Discovering Protein Interactions:** Use of co-occurrence networks of genes has led to the prediction of hidden protein-protein interactions and discovery of protein functions [12]. The visual discernibility of the co-occurrence pairs has been increased by proportionating the strength of co-occurrence

(more common pathways and publications) to the thickness of the chords connecting them. Analysing highly co-occurrent gene pairs in the context of gene/protein expression datasets of Linked TCGA would lead towards the *in silico* discovery of hidden interactions between their translated proteins.

3. **Predicting Tumour Risk:** One of the most pressing challenges in GWAS is the application of the study findings for the development of personalized genomic medicine and diagnostics, by improved integration of the genetic studies and the generated insights and considering the genetic variation between different individuals. As such, we hope to allow clinical practitioners to upload their patient's genomic datasets and evaluate the alterations and expression levels against those patients registered under the TCGA project. This would facilitate the clinician to make informed, medical decisions, augmented through input from other knowledge sources.

5.2 Future Work

Prostate adenocarcinoma, which is one of the most common malignancy to affect men, could be diagnosed through a combined evaluation of an individual's genomic and clinical data on a personalized scale [5]. We would like to integrate existing models into the GenomeSnip platform, to further assist the cancer researcher in the task of predicting prostate cancer risk in new patients. We hope to provide an 'interaction' overlay as an extra dimension to our visualization, along with the current 'co-occurrence' overlay, by integrating knowledge on protein-protein interactions, gene co-expression and functions. In the current version, the 'point mutations' layer in the 'Genomic Wheel' is not interlinked with other segments. We could extract further information like disease variant mutations from UniProt to address this. Finally, we would like to improve the granularity of the 'Genomic Wheel' visualization and extensively test the user experience and the usability of this platform by conducting a user-driven evaluation.

6 Conclusion

In this paper, we present our approach leading towards the conceptualization of a semantic, visual analytics prototype GenomeSnip, developed for the intuitive exploration of the human genome with embedded insights from Genome-wide Association Studies (GWAS) and Co-occurrence data of genomic features. We present the selection of various data sources which catalogue the extracted GWAS insights and gene-related mappings, and the transformation of this knowledge to generate co-occurrence data cubes and similarity measures between different genomic features. We assemble all this information to develop an aggregative, interactive visualization, the 'Genomic Wheel', for the cancer researchers to intuitively navigate across the human genome and select fragments of interest. The cancer researchers can analyze the 20.4 billion triples Linked TCGA datasets in the context of different selections simultaneously. We finally discuss the different usage scenarios and future directions for this project.

Acknowledgments

This research has been supported in part by the Science Foundation Ireland under Grant Number SFI/12/RC/2289. The authors would like to acknowledge Shanmukha Sampath and Oya Deniz Beyan for the discussions during the conceptualization phase of the GenomeSnip platform.

References

1. Baran, J., Gerner, M., Haeussler, M., Nenadic, G., Bergman, C.M.: Pubmed2Ensembl: A Resource for Mining the Biological Literature on Genes. PLoS ONE 6(9), e24716 (09 2011)
2. Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P., Morissette, J.: Bio2RDF: towards a mashup to build bioinformatics knowledge systems. Journal of biomedical informatics 41(5), 706–716 (2008)
3. Beyan, O.D., Iqbal, A., Khan, Y., Antoniades, A., Keane, J., Hasapis, P., Georgousopoulos, C., Ioannidi, M., Decker, S., Sahay, R.: Querying Phenotype-Genotype Associations across Multiple Knowledge Bases using Semantic Web Technologies. In: IEEE International Conference on BioInformatics and BioEngineering (2013)
4. Bleda, M., Tarraga, J., de Maria, A., Salavert, F., Garcia-Alonso, L., Celma, M., Martin, A., Dopazo, J., Medina, I.: CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources. Nucleic acids research 40(W1), W609–W614 (2012)
5. Boyd, L.K., Mao, X., Lu, Y.J.: The complexity of prostate cancer: genomic alterations and heterogeneity. Nature Reviews Urology 9(11), 652–664 (2012)
6. Chen, H., Yu, T., Chen, J.Y.: Semantic web meets integrative biology: a survey. Briefings in bioinformatics 14(1), 109–125 (2013)
7. Cyganiak, R., Reynolds, D., Tennison, J.: The RDF Data Cube Vocabulary, W3C Recommendation 16 January 2014. World Wide Web Consortium (2012)
8. Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D., Thompson, D., Ballinger, D.G., Struewing, J.P., Morrison, J., Field, H., Luben, R., et al.: Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447(7148), 1087–1093 (2007)
9. Fiume, M., Williams, V., Brook, A., Brudno, M.: Savant: genome browser for high-throughput sequencing data. Bioinformatics 26(16), 1938–1944 (2010)
10. Forbes, S.A., Tang, G., Bindal, N., Bamford, S., Dawson, E., Cole, C., Kok, C.Y., Jia, M., Ewing, R., Menzies, A., et al.: COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic acids research 38(suppl 1), D652–D657 (2010)
11. Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., Stratton, M.R.: A census of human cancer genes. Nature Reviews Cancer 4(3), 177–183 (2004)
12. Huynen, M., Snel, B., Lathe, W., Bork, P.: Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome research 10(8), 1204–1210 (2000)
13. Jenssen, T.K., Lægreid, A., Komorowski, J., Hovig, E.: A literature network of human genes for high-throughput analysis of gene expression. Nature genetics 28(1), 21–28 (2001)

14. Kamdar, M.R., Zeginis, D., Hasnain, A., Decker, S., Deus, H.F.: ReVeaLD: A user-driven domain-specific interactive search platform for biomedical research. Journal of biomedical informatics (2013)
15. Kanehisa, M., Goto, S.: KEGG: Kyoto encyclopedia of genes and genomes. Nucleic acids research 28(1), 27–30 (2000)
16. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D.: The human genome browser at UCSC. Genome research 12(6), 996–1006 (2002)
17. Khoury, M.J., Gwinn, M., Yoon, P.W., Dowling, N., Moore, C.A., Bradley, L.: The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? Genetics in Medicine 9(10), 665–674 (2007)
18. Kim, P.J., Price, N.D.: Genetic Co-Occurrence Network across Sequenced Microbes. PLoS Comput Biol. 7(12), e1002340 (12 2011)
19. Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A.: Circos: an information aesthetic for comparative genomics. Genome research 19(9), 1639–1645 (2009)
20. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P., Hirschhorn, J.N.: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetics 9(5), 356–369 (2008)
21. Medina, I., Salavert, F., Sanchez, R., de Maria, A., Alonso, R., Escobar, P., Bleda, M., Dopazo, J.: Genome Maps, a new generation genome browser. Nucleic Acids Research (2013)
22. Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., Wain, H.: The HUGO gene nomenclature committee (HGNC). Human genetics 109(6), 678–680 (2001)
23. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P.: Integrative genomics viewer. Nature biotechnology 29(1), 24–26 (2011)
24. Ruttenberg, A., Rees, J.A., Samwald, M., Marshall, M.S.: Life sciences on the Semantic Web: the Neurocommons and beyond. Briefings in bioinformatics 10(2), 193–204 (2009)
25. Saleem, M., Kamdar, M.R., Iqbal, A., Sampath, S., Deus, H.F., Ngonga, A.C.: Fostering Serendipity through Big Linked Data. In: Semantic Web Challenge at ISWC13 (2013)
26. Saleem, M., Shanmukha, S., Ngonga Ngomo, A.C., Almeida, J.S., Decker, S., Deus, H.F.: Linked cancer genome atlas database. In: I-Semantics 2013 (2013)
27. Schroeder, M.P., Gonzalez-Perez, A., Lopez-Bigas, N., et al.: Visualizing multidimensional cancer genomics data. Genome medicine 5(1), 1–13 (2013)
28. Seal, R.L., Gordon, S.M., Lush, M.J., Wright, M.W., Bruford, E.A.: genenames. org: the HGNC resources in 2011. Nucl. Acids Res. 39(Suppl 1), D514–D519 (2011)
29. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., Sirotkin, K.: dbSNP: the NCBI database of genetic variation. Nucleic acids research 29(1), 308–311 (2001)
30. Tversky, A.: Features of similarity. Psychological review 84(4), 327 (1977)
31. Wang, K., Li, M., Hakonarson, H.: Analysing biological pathways in genome-wide association studies. Nature Reviews Genetics 11(12), 843–854 (2010)
32. Yeager, M., Orr, N., Hayes, R.B., Jacobs, K.B., Kraft, P., Wacholder, S., Minichiello, M.J., Fearnhead, P., Yu, K., Chatterjee, N., et al.: Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nature genetics 39(5), 645–649 (2007)