Everything should be linked: linking and visualising data for dynamic multilevel and multidimensional biological data interpretation.

Exploring multi-level effects of structural variations in non-coding genomic regions in cancer

RHWE (Robin) van der Weide*

Cancer Stem cells & Developmental biology
Utrecht Graduate School of Life Sciences

*Supervisor: Joep de Ligt, PhD

Summary of the research

With the increase in popularity and cost-effectiveness of various omics-approaches, more and more data is becoming available to researchers of different fields. The complexity of integrating and analysing information of these approaches increases with every added omics-layer and/or other dimension (e.g. time-series, treatments). The current tools and frameworks for these approaches have two major limitations in their design: scalability and generality (i.e. the possibility to add of more levels and/or dimensions). Moreover, there isn't an option to overview a dataset without filtering, dividing or structuring the data. These limitations restrict the integration of complex dataset, needed to truly understand biology.

Enter the Semantic Web and its Resource Description Framework (RDF). A simple and flexible framework for describing anything about anything. An example of such a RDF-instance (a triple) is "BRAF1 has the molecular function of binding calcium ion", which has these three parts: a subject (BRAF1), a predicate (molecular function) and an object (binding calcium ion). Another Triple can then say something about the phosphorylated protein levels of this gene in a sample. Connecting these two Triples would enable a researcher to find a possible pattern in the data (i.e. a gene, responsible for calcium ion binding, has a low phosphorylation level in the investigated sample). Since every type of data can be translated to RDF, integration of large datasets of different levels and dimensions becomes possible and a lot more feasible. Both local and remote triples can be easily combined (EMBL-EBI has already launched six databases, including UniProt and Reactome), which makes analyses even more powerful. By using the SPARQL Protocol and RDF Query Language (SPARQL), retrieving and manipulating data in RDF is easily readable by both humans and computers. The SPARQL-results can subsequently be visualized as a whole, or filtered by the user.

Here, we propose the use of semantic web technologies and visual analytics to decrease the complexity of integrating and visualizing multi-level and -dimensional biological data. Firstly, we will create the framework needed to design the missing tools for converting the most-used NGS-formats to RDF. Next, methods and tools for visual analytics of the biological RDF-data will be created. Previously unmanageable integration-focussed analyses on the consequences of structural variation in the non-coding regions of cancer-genomes are used to showcase the proposed methods.

Layman's summary

The biomedical research community wants to be able to study more biological signals than one at the same time, because the biology of, for example, cancer is so complex. However, integrating diverse sets of biological signals is currently one of the main challenges in multiple fields of study. To overcome this, we propose the use of Semantic Web-methods: these are specifically designed for integrating vast amounts of different data (from the internet). Furthermore, it allows users to describe, analyse and test their data interactively and dynamically via visual representations displayed in the browser. One of the fields that would benefit greatly from these methods are those, that find the cancer-causing consequences of changes in parts of chromosomes that do not hold a gene: the non-coding genomic regions. A preliminary study shows the added performance of these methods in biology: researchers are able to describe, analyse and test 20 times more biological questions in the same time, compared to conventional methods. We thus propose to develop these methods further for the research-community (and biology in particular) and show their added benefits by performing research on variations in the non-coding genomic regions in cancer.

Keywords: *structural variation, multi-level data integration, next-generation sequencing, cancer, visual analytics*

Background, aims and approach

Overall aim

The aim of this project is to integrate and visualise multiple levels and dimensions of (NGS-based) omics-data with methods of the semantic web and, using these methods, further understand the consequences of structural variations in the non-coding regions of the genome on other biological levels, like the transcriptome and proteome. Thus, this proposal has three sub-projects, which rely heavily on each other:

1. **Data-integration** Integration of NGS-based data by using Semantic Web-methodologies to improve integrative bioinformatics in general and NGS-based multi-level and -dimensional research in particular.

2. **Visual analytics** Linking the Semantic-Web data to D3.js, enabling researchers to dynamically and interactively visualise RDF.

3. **Multi-level analysis** Multi-level and -dimensional integrative bioinformatical analysis to elucidate the consequences of genomic structural variations in non-coding regions in cancer.

Scientific relevance and challenges

The amount of (public) biological data has exploded in the last years -even outpacing Moore's law- which is the result of the advances in omics-technologies, like Next-Generation Sequencing (NGS) and Mass-Spectrometry (MS), in both performance and costs. Aside from the sheer size, a second factor for the highly complex nature of current biomedical research is the addition of other dimensions, like time-series or treatments to the aforementioned omics-levels and other types of data from the same level (e.g. miRNA-seq combined with RNA-seq). While there are plenty of studies on single-level data analysis, both academia and industry agree that data-integration is key to understanding the complex nature of biology more thoroughly [1;2;3;4]. However, only a few layers and/or dimensions have been integrated per study and results are -for the most part- cherry picked, instead of data-wide. This is mainly due to the methods used in integration-studies, which are limited due to the large amounts of parsing-time (i.e. the time to convert various file/region-formats): most of them are set up in the same manner as individual level-experiments, whereafter they are combined. These methods lead high amounts of analytical time, as was the case in the study of Munoz et al. [5]: every two months of data-accumulation costed two years of analysis. The limited number of truly integrative studies use computational approaches to reconstruct biological networks. While this is a valid strategy, scaling the analysis from the bacteria used by Karr et al. [6] and Lerman et al. [7] to multi-cellular organisms proves to be difficult. The most obvious reasons for this are the complexity of the used mathematical methods, the integration of multiple data-sources (with varying file-formats) and/or due to the use of a set-in-stone database-structure.

To overcome these scaling issues, **we propose the use of the Semantic Web: the *Resource Description Framework* (RDF) and its query-language *SPARQL Protocol and RDF Query Language* (SPARQL)**. RDF is a general and simple framework for making statements about subjects, which is already heavily used in fields outside of biology, enabling users to integrate and search data based on semantics. Within biology, RDF is only used sparsely and mainly focussed on external data-source integration and not on own data [8;9;10]. Every RDF-statement (i.e. a Triple) has three parts: a subject, a predicate and an object (e.g. BRAF1 :: molecular function :: calcium ion binding). This makes it possible to link every object to another and denote the relationship between them: no additional (file)formats are needed.

Compared to other relational database management systems, RDF is completely flexible: no database-schemas (pre-specified structures for the data, like the mySQL-method of Low et al. [11]) are needed. Aside from the low complex, flexible and self-describing nature of the RDF-data, triples can be seen as a modular directed graph: users can combine multiple relevant RDF-sources (e.g. UniProt and Proteomics-data). Every additional RDF-source results in a more relevant and heterogeneous population of triples, making the network more complex and informative. Extracting relevant information from this "hairball" of linked objects and subjects has been a major issue and challenge since the beginning of big data, as Pavlopoulos et al. [12] stated in 2008. SPARQL provides this ability to filter on an arbitrary number of (human-readable) expressions and can combine multiple databases to query, like the RDF-databases of EMBL-EBI [13].

When data is integrated in Semantic Web RDF-database (TripleStore) and a relevant set of subjects, predicates and/or object is extracted using SPARQL, the remaining dataset is still very large. The abstract and complex nature of this "hairball" makes it hard to formalise an analytical problem to solve. **To create interactive and dynamic visual representations of a dataset, we propose to use of the multidisciplinary theories and methods of *visual analytics***. Thomas and Cook [14] describe this field in 2005 as "*Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces.*". It uses analytical and statistical methods from fields as computer science and statistics and visualisation-techniques from cognitive and design sciences, such that the data can be effectively analysed (i.e. hypotheses formed and analysed) by the user. The JavaScript library Data Driven Documents (*D3.js*) is perfect implementing linked data -which itself is already a graph- within visual analytics, as it is focussed on structuring data for dynamic, web-based (using current standards like HTML & SVG) visualisations [15]. And since it is embedded in HTML, additional operators (e.g. buttons, SPARQL-forms) can be added. Due to these benefits, the use of D3.js in visual analytics is increasing, with a notable biology-specific example in Epiviz2 [16]. However, this tool only takes a specific set of data-formats and -levels and -more importantly- only shows a specific genomic region, instead of the complete scope. This leads to "*cherry picking*", instead of data-focussed formulation and analysis of hypotheses.

To illustrate to possibilities of the fully integrative multi-level and -dimensional visual analytics, **we propose to use our methods to study to consequences of structural variations at non-coding loci in cancer on other levels.** The heterogeneous samples and datasets of cancer make it one of the most computationally demanding types of integrative biology. This technical challenging topic is therefore a perfect illustration of the possibilities of our methods.

While the consequences of structural variations in coding regions are well studied due to the relative ease of finding consequencial changes in the transcript and proteins [17], computational methods for the non-coding regions are just starting to come up in the literature of 2014 [18;19]. There already have been successfully studies, linking non-coding regions to colorectal- and skin-cancer [20;21]. The young nature of the used methods is visible in their output: only a prioritisation-score, based on the probability to be a causal variant, is given. Combining these scores with our linked-data methods -using both self-generated and public data- would lead to further understanding of the mechanisms of cancer.

Originality and innovative character

The 2014 survey of Gomez-Cabrero et al. [1] showed that the questioned biomedical academics had the highest interest (78.2 percent) in the integration of multiple omics-datasets and that there was a high need for standardized tools and data-types. Especially data-storage, -exploration and -exploitation were found the be key: their conclusions were best summarized by *the need for having exploration tools, which combine summery statistics and interactive visualisations, to analyse heterogeneous data-sets.*

There have already been various studies on integration of biological signals with the aid of semantic web technologies, as the power of ontology-based entailment reasoning is widely acknowledged [10]. However, the momentum is lacking: until 2014, no big databases were available in RDF-format. This meant that bioinformamtical research involving RDF had no momentum, as they could only integrate their own data, like the integration of RDF-methods in microarray analyses by Szpakowski et al. [22] in 2009. Recently, EMBL-EBI has opened their own RDF-platfom, boasting six big data-sources (Gene Expression Atlas, ChEMBL, BioModels, Reactome, BioSamples and UniProt [13]. This was the boost of momentum needed to further incorporate RDF in biological analyses.

However, there are two main limitations of this relatively young incorporation: a standard language for denoting triples (e.g. chromosome locations) is missing and the focus lies at linking database-accessions [23]. While the first limitation is also a strength (everybody can use their own dialect), a standardisation-step will enable researchers in all fields of biology to fully benefit from the integrative benefits of the Semantic Web. The second limitation is severely restricting the use of RDF in NGS- and MS-based methods: there are no tools to convert the common formats, like the *Variant Call Format* (VCF) and *Sequence Alignment Format* (SAM), to triples. An example of this is *bio2rdf* [8]: a "*RDFizer*", which converts common databases, like the ones from NCBI, to triplestores. One of the main innovative points of our proposal is the development of methods to handle these NGS- and MS-based formats for use in the Semantic Web. This will result in a broader use of semantic web-technologies for the research community, by enabling the coupling of (own) NGS- and MS-data to existing RDF-databases.

The implementation of web-based visual analytics for RDF-databases is another major innovative point in this proposal. Combining Semantic Web-technology with this will create a paradigm shift in the way integrative analysis of (biological) data is done. Visual analytics have been shown to result in the most optimal analysis-effectivity, as it allows the user to combine the data with their own background and intuition (fig. 1). Not only can data be more effectively analysed, but it can also be better understood and presented, due to the ability to provide an overview of the complete dataset [14;24].

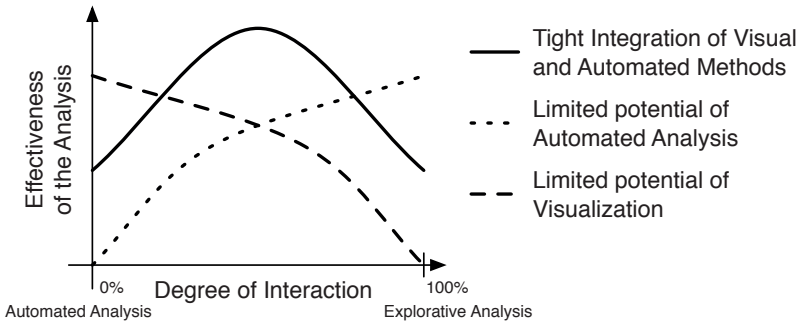


Figure 1: ***Trade-offs between automated and explorative analysis.*** *By combining automated analyses, where appropriate, with the background and intuition of the user, an optimal amount of effectivity can be attained. Picture taken from Keim et al.* [24].

While there has already been a large amount of work in the field of computational cancer research, the vast majority of large-scale integrative studies have been performed on the coding-regions (which account for 1,5 percent of the total size) of the genome [25]. Genome-Wide Association Studies (GWASs) on a broad range of (hereditary) cancers have shown, that non-coding locations are associated with these diseases. Until 2013, however, tools and sources to find the precise causative variations in non-coding genomic regions were limited. In the last two years, several advances have made it possible to asses the consequences of individual variations in non-coding regions [20;18], but

no large-scale integrative studies have been performed (partly due to the current state of integrative methods). With our new methods, we will be in the position to do fully integrative studies on the underlying mechanisms and consequences of (structural) variation in cancer.

Pilot-study

To asses to feasibility of our proposal, a small-scale pilot-study was done on de data of van Heesch et al.[26]. This dataset includes transcriptome data of mRNA's, bound to a number (1-7+) of ribosomal units and matching exome-data. If one would be interested to look at the molecular functions of a gene, which transcripts have have a bias for one SNP, a disproportional amount of time is lost on parsing, intersecting and downloading various types of data (fig. 2). With the current methods, twelve set-operations (e.g. intersections, unions) have to performed on approximately 5gb of data and three datasets (±15gb) have to be completely downloaded, before a simple curiosity-driven question can be answered. Approximately three and a half hours was needed to perform this, in contrast to one hour with our methods. Of this hour, more than fifty minutes were used to convert VCF to RDF and load the 4store database: every query hereafter takes up approximately 10 minutes. This pilot illustrates two important factors: the low-complex nature of the proposed methods and the valuable property of having a separate query-stage, which results in being able to make more than ($\frac{(8*60)-50}{10}$ =) forty queries in eight hours, instead of approximately ($\frac{8-3,5}{3,5}$ =) two queries in the currently used methods.

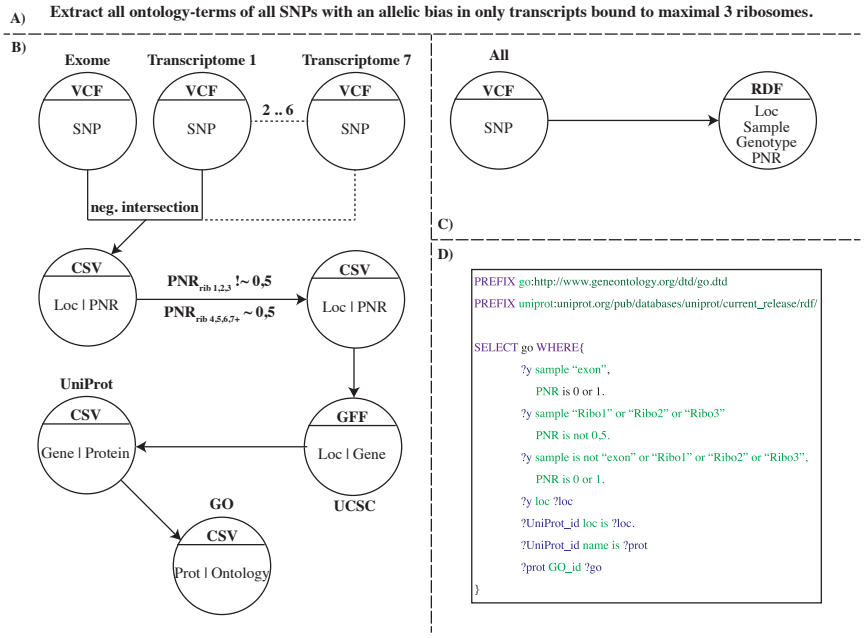
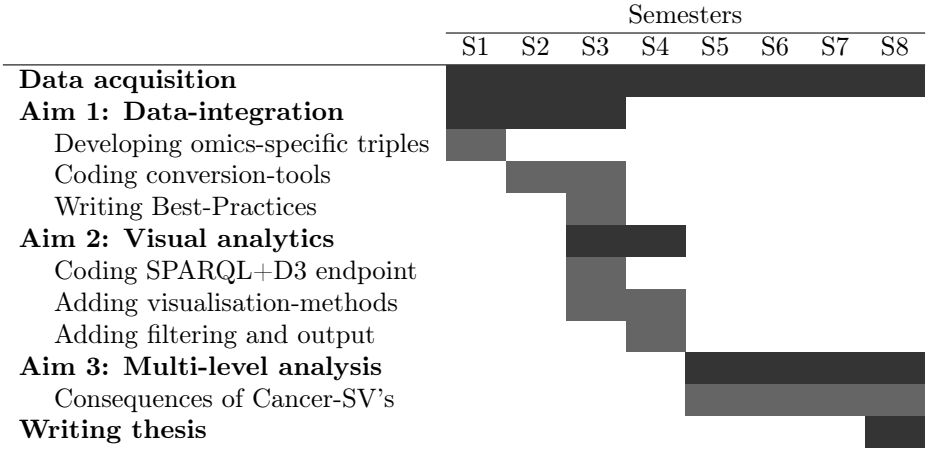


Figure 2: ***Differences between current integration techniques and RDF.*** *When a researcher has a question like* ***A****, he/she has to go through a series of parsing and interception steps (data juggling), like in* ***B****. External sources have to be fully downloaded and converted, before use. Using our proposed pipeline (shown in* ***C****), results in the standard conversion to RDF. Then, a question can be formulated in SPARQL (****D****), incorporating relevant outside sources, which can be easily changed without having to juggle the data again.*

Methods and techniques

Research plan

Timetable



Collaboration

possibles:

hubrecht: organoids! veel cancer-dingetjes te krijgen dus (clonal)

Horizon-groep (cancer gneomics centre)

rdf-mensen (Marco Roos, Joachim baran, japanner&rus)

Knowledge utilisation

stukje over implementatie van RDF: bestaande grote bronnen en mijn uitbreiding tools en vocabulair geeft RDF onderzoekers de mogelijkehid om daadwerkelijke data-integratie studies op te zetten met easy-of-install and -use. Het feit dat er al statistische paketten zijn in de statistical software enivorment of choice -R- betekend dat gebruikers alleen RDF+SPARQL hoeven te leren, maar dat onderliggende statistical analyses op the gefilterede SPARQL-queries gewoon in R kunnen worden gedaan.

In a broader perspective: het uitrbreiden van het semantic web (door NGS-based triple stores) leidt tot een

linked data visual analystic zijn in alle velden te grebruiken, die RDf gebruiken. Ook voor bedrijven (pharma!). Super handig!

Vrij snel te incorpporeren: de meeste dingen zijn er al

cancer NC-SV is nog weinig over bekend: mogelijke nieuwe targets voor cancer screening and or treatment: sociaal en pharma.

References

[1] David Gomez-Cabrero, Imad Abugessaisa, Dieter Maier, Andrew Teschendorff, Matthias Merkenschlager, Andreas Gisel, Esteban Ballestar, Erik Bongcam-Rudloff, Ana Conesa, and Jesper Tegnér. Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.*, 8(Suppl 2):I1, 2014. ISSN 1752-0509. doi: 10.1186/1752-0509-8-S2-I1. URL http://www.biomedcentral.com/1752-0509/8/S2/I1.

[2] Curtis Huttenhower and Oliver Hofmann. A quick guide to large-scale genomic data mining. *PLoS Comput. Biol.*, 6(5):e1000779, May 2010. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000779. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2877728&tool=pmcentrez&rendertype=abstract.

[3] David B Searls. Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.*, 4(1):45–58, January 2005. ISSN 1474-1776. doi: 10.1038/nrd1608. URL http://www.ncbi.nlm.nih.gov/pubmed/15688072.

[4] Jemila S Hamid, Pingzhao Hu, Nicole M Roslin, Vicki Ling, Celia M T Greenwood, and Joseph Beyene. Data integration in genetics and genomics: methods and challenges. *Hum. Genomics Proteomics*, 2009(1):869093–, January 2009. ISSN 1757-4242. doi: 10.4061/2009/869093. URL http://hgp.sagepub.com.proxy.library.uu.nl/content/1/1/869093.full.

[5] Javier Munoz, Teck Y Low, Yee J Kok, Angela Chin, Christian K Frese, Vanessa Ding, Andre Choo, and Albert J R Heck. The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. *Mol. Syst. Biol.*, 7:550, January 2011. ISSN 1744-4292. doi: 10.1038/msb.2011.84. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3261715&tool=pmcentrez&rendertype=abstract.

[6] Jonathan R Karr, Jayodita C Sanghvi, Derek N Macklin, Miriam V Gutschow, Jared M Jacobs, Benjamin Bolival, Nacyra Assad-Garcia, John I Glass, and Markus W Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401, July 2012. ISSN 1097-4172. doi: 10.1016/j.cell.2012.05.044. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3413483&tool=pmcentrez&rendertype=abstract.

[7] Joshua A Lerman, Daniel R Hyduke, Haythem Latif, Vasiliy A Portnoy, Nathan E Lewis, Jeffrey D Orth, Alexandra C Schrimpe-Rutledge, Richard D Smith, Joshua N Adkins, Karsten Zengler, and Bernhard O Palsson. In silico method for modelling metabolism and gene product expression at genome scale. *Nat. Commun.*, 3:929, January 2012. ISSN 2041-1723. doi: 10.1038/ncomms1928. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3827721&tool=pmcentrez&rendertype=abstract.

[8] François Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and Jean Morissette. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.*, 41(5):706–16, October 2008. ISSN 1532-0480. doi: 10.1016/j.jbi.2008.03.004. URL http://www.sciencedirect.com/science/article/pii/S1532046408000415.

[9] Eric K Neumann and Dennis Quan. BioDash: a Semantic Web dashboard for drug development. *Pac. Symp. Biocomput.*, pages 176–87, January 2006. ISSN 2335-6936. URL http://europepmc.org/abstract/MED/17094238.

[10] Satya S Sahoo, Olivier Bodenreider, Joni L Rutter, Karen J Skinner, and Amit P Sheth. An ontology-driven semantic mashup of gene and biological pathway information: application to the domain of nicotine dependence. *J. Biomed. Inform.*, 41(5):752–65, October 2008. ISSN 1532-0480. doi: 10.1016/j.jbi.2008.02.006. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2766186&tool=pmcentrez&rendertype=abstract.

[11] Teck Yew Low, Sebastiaan van Heesch, Henk van den Toorn, Piero Giansanti, Alba Cristobal, Pim Toonen, Sebastian Schafer, Norbert Hübner, Bas van Breukelen, Shabaz Mohammed, Edwin Cuppen, Albert J R Heck, and Victor Guryev. Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell Rep.*, 5(5):1469–78, December 2013. ISSN 2211-1247. doi: 10.1016/j.celrep.2013.10.041. URL http://www.cell.com/article/S2211124713006402/fulltext.

[12] Georgios A Pavlopoulos, Anna-Lynn Wegener, and Reinhard Schneider. A survey of visualization tools for biological network analysis. *BioData Min.*, 1:12, January 2008. ISSN 1756-0381. doi: 10.1186/1756-0381-1-12. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2636684&tool=pmcentrez&rendertype=abstract.

[13] Simon Jupp, James Malone, Jerven Bolleman, Marco Brandizi, Mark Davies, Leyla Garcia, Anna Gaulton, Sebastien Gehant, Camille Laibe, Nicole Redaschi, Sarala M Wimalaratne, Maria Martin, Nicolas Le Novère, Helen Parkinson, Ewan Birney, and Andrew M Jenkinson. The EBI RDF platform: linked open data for the life sciences. *Bioinformatics*, 30(9):1338–9, May 2014. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt765. URL http://bioinformatics.oxfordjournals.org.proxy.library.uu.nl/content/30/9/1338.

[14] JJ Thomas and KA Cook. Illuminating the path: The research and development agenda for visual analytics. *IEEE Comput. Soc.*, pages 19–32, 2005. URL http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Illuminating+the+Path:+The+Research+and+Development+Agenda+for+Visual+Analytics#0http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Illuminating+the+path:+The+research+and+development+agenda+for+visual+analytics#0.

[15] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. DÂş: Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2301–9, December 2011. ISSN 1941-0506. doi: 10.1109/TVCG.2011.185. URL http://www.computer.org/csdl/trans/tg/2011/12/ttg2011122301.html.

[16] Florin Chelaru, Llewellyn Smith, Naomi Goldstein, and Héctor Corrada Bravo. Epiviz: interactive visual analytics for functional genomics data. *Nat. Methods*, 11(9):938–940, August 2014. ISSN 1548-7091. doi: 10.1038/nmeth.3038. URL http://www.nature.com.proxy.library.uu.nl/nmeth/journal/v11/n9/abs/nmeth.3038.html.

[17] William McLaren, Bethan Pritchard, Daniel Rios, Yuan Chen, Paul Flicek, and Fiona Cunningham. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, 26(16):2069–70, August 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq330. URL http://bioinformatics.oxfordjournals.org.proxy.library.uu.nl/content/26/16/2069.short.

[18] Ekta Khurana, Yao Fu, Vincenza Colonna, Xinmeng Jasmine Mu, Hyun Min Kang, Tuuli Lappalainen, Andrea Sboner, Lucas Lochovsky, Jieming Chen, Arif Harmanci, Jishnu Das, Alexej Abyzov, Suganthi Balasubramanian, Kathryn Beal, Dimple Chakravarty, Daniel Challis, Yuan Chen, Declan Clarke, Laura Clarke, Fiona Cunningham, Uday S Evani, Paul Flicek, Robert Fragoza, Erik Garrison, Richard Gibbs, Zeynep H Gümüs, Javier Herrero, Naoki Kitabayashi, Yong Kong, Kasper Lage, Vaja Liluashvili, Steven M Lipkin, Daniel G MacArthur, Gabor Marth, Donna Muzny, Tune H Pers, Graham R S Ritchie, Jeffrey A Rosenfeld, Cristina Sisu, Xiaomu Wei, Michael Wilson, Yali Xue, Fuli Yu, Emmanouil T Dermitzakis, Haiyuan Yu, Mark A Rubin, Chris Tyler-Smith, and Mark Gerstein. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science*, 342(6154):1235587, October 2013. ISSN 1095-9203. doi: 10.1126/science.1235587. URL http://europepmc.org/articles/PMC3947637/?report=abstract.

[19] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, 46(3):310–5, March 2014. ISSN 1546-1718. doi: 10.1038/ng.2892. URL http://www.ncbi.nlm.nih.gov/pubmed/24487276.

[20] Halit Ongen, Claus L. Andersen, Jesper B. Bramsen, Bodil Oster, Mads H. Rasmussen, Pedro G. Ferreira, Juan Sandoval, Enrique Vidal, Nicola Whiffin, Alexandra Planchon, Ismael Padioleau, Deborah Bielser, Luciana Romano, Ian Tomlinson, Richard S. Houlston, Manel Esteller, Torben F. Orntoft, and Emmanouil T. Dermitzakis. Putative cis-regulatory drivers in colorectal cancer. *Nature*, July 2014. ISSN 0028-0836. doi: 10.1038/nature13602. URL http://www.nature.com.proxy.library.uu.nl/nature/journal/vaop/ncurrent/full/nature13602.html.

[21] Franklin W Huang, Eran Hodis, Mary Jue Xu, Gregory V Kryukov, Lynda Chin, and Levi A Garraway. Highly recurrent TERT promoter mutations in human melanoma. *Science*, 339(6122):957–9, February 2013. ISSN 1095-9203. doi: 10.1126/science.1229259. URL http://www.sciencemag.org.proxy.library.uu.nl/content/339/6122/957.short.

[22] Sebastian Szpakowski, James McCusker, and Michael Krauthammer. Using semantic web technologies to annotate and align microarray designs. *Cancer Inform.*, 8:65–73, January 2009. ISSN 1176-9351. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4042255&tool=pmcentrez&rendertype=abstract.

[23] Alan Ruttenberg, Tim Clark, William Bug, Matthias Samwald, Olivier Bodenreider, Helen Chen, Donald Doherty, Kerstin Forsberg, Yong Gao, Vipul Kashyap, June Kinoshita, Joanne Luciano, M Scott Marshall, Chimezie Ogbuji, Jonathan Rees, Susie Stephens, Gwendolyn T Wong, Elizabeth Wu, Davide Zaccagnini, Tonya Hongsermeier, Eric Neumann, Ivan Herman, and Kei-Hoi Cheung. Advancing translational research with the Semantic Web. *BMC Bioinformatics*, 8 Suppl 3(Suppl 3):S2, January 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-S3-S2. URL http://www.biomedcentral.com/1471-2105/8/S3/S2.

[24] Daniel Keim, Florian Mansmann, and Jim Thomas. Visual analytics: how much visualization and how much analytics? *ACM SIGKDD Explor. . . .*, 11(2):5–8, 2010. URL http://dl.acm.org/citation.cfm?id=1809403.

[25] The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696):636–40, October 2004. ISSN 1095-9203. doi: 10.1126/science.1105136. URL http://www.ncbi.nlm.nih.gov/pubmed/15499007.

[26] Sebastiaan van Heesch, Maarten van Iterson, Jetse Jacobi, Sander Boymans, Paul B Essers, Ewart de Bruijn, Wensi Hao, Alyson W Macinnes, Edwin Cuppen, and Marieke Simonis. Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biol.*, 15(1):R6, January 2014. ISSN 1465-6914. doi: 10.1186/gb-2014-15-1-r6. URL http://genomebiology.com/2014/15/1/R6.