# 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

*Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.*

***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

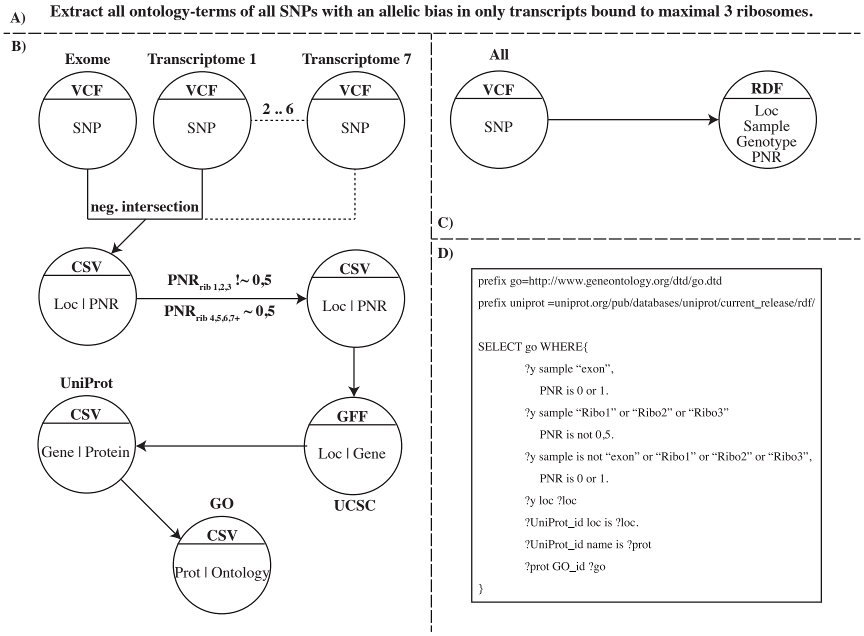
The amount of (public) biological data has exploded in the last years -even outpacing Moore's law- which is the result of the advances in omics-technologies, like Next-Generation Sequencing (NGS) and Mass-Spectrometry (MS), in both performance and costs. Aside from the sheer size, a second factor for the highly complex nature of current biomedical research is the addition of other dimensions, like time-series or treatments to the aforementioned omics-levels. 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[2;4;13;3]. 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. 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[1;9;12]. 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.[8]) 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.[10] 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[5].

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[15] describe this field in 2005 as "*Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces.*". It uses analytical methods from fields as computer science and statistics to perform analyses 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.

Usage case: structural variants of non-coding regions in cancer


A) Extract all ontology-terms of all SNPs with an allelic bias in only transcripts bound to maximal 3 ribosomes.

```
prefix go=http://www.geneontology.org/dtd/go.dtd
prefix uniprot =uniprot.org/pub/databases/uniprot/current_release/rdf/

SELECT go WHERE{
        ?y sample “exon”,
            PNR is 0 or 1.
        ?y sample “Ribo1” or “Ribo2” or “Ribo3”
            PNR is not 0,5.
        ?y sample is not “exon” or “Ribo1” or “Ribo2” or “Ribo3”,
            PNR is 0 or 1.
        ?y loc ?loc
        ?UniProt_id loc is ?loc.
        ?UniProt_id name is ?prot
        ?prot GO_id ?go
}
```

**Figure 1:** ***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 easliy changed without having to juggle the data again.*

## Originality and innovative character

EDWINS PRESENTATIE: ONE-DIMENSIONAL DATA STILL REQUIERS OTHER DATAT TYPES - SMALL RNA-SEQ, CHIP-SEQ ETC INTERGRATION IS VERY HARD: MET ALBERT

HECK KOSTTE HET TWEE MAANDEN OM DATA TE GENEREREN EN TWEE JAAR OM TE INTEGREREN EN ANALYSEREN.

ALBERT HECK: GROOT ONDERZOEK OVER STAMCELLEN: DATA INTEGRATIE AND MINING IN AUSTRALIE... PCA VAN PROT, RNA, GEN AND METH KOMEN OVEREEN

There have already been various studies on integration of biological signals with the aid of semantic web technologies. 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.[14] 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[5]. 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[11]. 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. 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.

added value of linked data visual analytics TOV standaard methodes

innovatie van cancer NC-SV

## Methods and techniques

# Research plan

## Timetable

| | Semesters | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 |
| **Data acquisition** | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| **Aim 1: Data-integration** | ■ | ■ | ■ | | | | | |
| Developing omics-specific triples | ▨ | | | | | | | |
| Coding conversion-tools | | ▨ | ▨ | | | | | |
| Writing Best-Practices | | | ▨ | | | | | |
| **Aim 2: Visual analytics** | | | ■ | ■ | | | | |
| Coding SPARQL+D3 endpoint | | | ▨ | | | | | |
| Adding visualisation-methods | | | ▨ | ▨ | | | | |
| Adding filtering and output | | | | ▨ | | | | |
| **Aim 3: Multi-level analysis** | | | | | ■ | ■ | ■ | ■ |
| Consequences of Cancer-SV's | | | | | ▨ | ▨ | ▨ | ▨ |
| **Writing thesis** | | | | | | | | ■ |

### Collaboration

possibles:
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] Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., and Morissette, J. (2008). Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.*, 41(5):706–16.

[2] Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., and Tegnér, J. (2014). Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.*, 8(Suppl 2):I1.

[3] Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M. T., and Beyene, J. (2009). Data integration in genetics and genomics: methods and challenges. *Hum. Genomics Proteomics*, 2009(1):869093–.

[4] Huttenhower, C. and Hofmann, O. (2010). A quick guide to large-scale genomic data mining. *PLoS Comput. Biol.*, 6(5):e1000779.

[5] Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., Wimalaratne, S. M., Martin, M., Le Novère, N., Parkinson, H., Birney, E., and Jenkinson, A. M. (2014). The EBI RDF platform: linked open data for the life sciences. *Bioinformatics*, 30(9):1338–9.

[6] Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., Assad-Garcia, N., Glass, J. I., and Covert, M. W. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401.

[7] Lerman, J. A., Hyduke, D. R., Latif, H., Portnoy, V. A., Lewis, N. E., Orth, J. D., Schrimpe-Rutledge, A. C., Smith, R. D., Adkins, J. N., Zengler, K., and Palsson, B. O. (2012). In silico method for modelling metabolism and gene product expression at genome scale. *Nat. Commun.*, 3:929.

[8] Low, T. Y., van Heesch, S., van den Toorn, H., Giansanti, P., Cristobal, A., Toonen, P., Schafer, S., Hübner, N., van Breukelen, B., Mohammed, S., Cuppen, E., Heck, A. J. R., and Guryev, V. (2013). Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. *Cell Rep.*, 5(5):1469–78.

[9] Neumann, E. K. and Quan, D. (2006). BioDash: a Semantic Web dashboard for drug development. *Pac. Symp. Biocomput.*, pages 176–87.

[10] Pavlopoulos, G. A., Wegener, A.-L., and Schneider, R. (2008). A survey of visualization tools for biological network analysis. *BioData Min.*, 1:12.

[11] Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., Doherty, D., Forsberg, K., Gao, Y., Kashyap, V., Kinoshita, J., Luciano, J., Marshall, M. S., Ogbuji, C., Rees, J., Stephens, S., Wong, G. T., Wu, E., Zaccagnini, D., Hongsermeier, T., Neumann, E., Herman, I., and Cheung, K.-H. (2007). Advancing translational research with the Semantic Web. *BMC Bioinformatics*, 8 Suppl 3(Suppl 3):S2.

[12] Sahoo, S. S., Bodenreider, O., Rutter, J. L., Skinner, K. J., and Sheth, A. P. (2008). 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.

[13] Searls, D. B. (2005). Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.*, 4(1):45–58.

[14] Szpakowski, S., McCusker, J., and Krauthammer, M. (2009). Using semantic web technologies to annotate and align microarray designs. *Cancer Inform.*, 8:65–73.

[15] Thomas, J. and Cook, K. (2005). Illuminating the path: The research and development agenda for visual analytics. *IEEE Comput. Soc.*, pages 19–32.