Understanding the effects of structural variation in non-coding regions.

Analysing multi-level 'omics using graph-based integration methods.

RHWE (Robin) van der Weide*

BSc. Biology

MSc. Cancer Stem cells & Developmental biology (honours program)

Utrecht Graduate School of Life Sciences

*Promotor: Edwin Cuppen, Prof. PhD | Copromotor: Joep de Ligt, PhD

Summary of the research

To date, studies on non-coding regions of the genome have been limited. This is mainly due to the complex nature of putative functional elements in these regions. Increased knowledge of these regions, through efforts such as the ENCODE-project, enabled researchers to initiate studies on the causality of non-coding variations[1;2]*.*

The ongoing cost reduction of various omics-approaches coupled to high throughput research, has led an explosion of available data. However, the complexity of integrating and analysing these large datasets increases with every added omics-layer or dimension (e.g. time-series, treatments). Furthermore, the elucidation of structural variants is further complicated by the diversity in types and consequences of these genetic alterations. This is especially true for structural variants in non-coding regions, where finding the affected gene is a challenge all by itself.

The current methods for integrating and analysing multiple layers or dimensions have two major limitations in their design: scalability and generality (i.e. the possibility to easily add more levels or dimensions). Moreover, the sheer amount of data points hampers data exploration without the need for filtering, dividing or restructuring the data. Integration and visualisation of complex datasets are needed to better understand the complex biology of disease[3]*, but is greatly restricted by aforementioned limitations.*

Within the realm of systems biology, graph-based methods are well-known and -used for analysis of interaction-networks. The main benefits of abstracting complex data-sources to graphs are enhanced exploratory analysis via visual analytics and integration of datasets of different levels and dimensions.

For larger and more complex datasets, big data scientists gravitate to the Resource Description Framework (RDF): a simple and flexible graph-framework, which also allows for easy connection to (web-based) public repositories. Researchers can start to form hypotheses of both local and remote parts of a graph using simple SPARQL-queries and subsequently visualise the results using advanced interactive visualisation interfaces.

Here, we propose the use of graph-based methods, like RDF, to decrease the complexity of integrating and visualizing multi-level and -dimensional biological data. These methods will enable us to create new biological insights in the complex biology of non-coding structural variants. Subsequentially, it will enable further elucidation in, for example, congenital disease and cancer.

Keywords: graph-based methodology, structural variation, multi-level data integration, non-coding genomics, visual analytics

Background, aims and approach

Overall aim

The aim of this project is to create new biological insights in the inter-level (e.g. transcriptome, proteome) consequences of structural variations in the non-coding regions of the genome. For this, a large number of datasets and -sources have to be integrated and analysed. Big data graph-based methods, like RDF, allow us to do this. Thus, this proposal has two sub-projects, which rely heavily on each other:

1. **Graph-based data integration & visualisation** Generation of novel methodologies for integration and analysis of large scale, multi-dimensional biological data
2. **Multi-level analysis** Multi-level and -dimensional integrative analysis to elucidate the consequences of genomic structural variations in non-coding regions, found in patients with congenital disease and cancer.

Scientific relevance and challenges

- Non-coding: lastig en weinig onderzocht
- Grootste winst is te halen in het combineren van lagen (distant relationships)
- Distant relationships vindt je het makkelijkst met een graph
- Er is veel data om te gebuiken! Maar het is wel erg veel... en heterogeen

Originality and innovative character

- Weinig onderzoek naar NC-SV
- awesome locatie en onderzoek
- Veel vraag naar integratie
- Waar stond RDF vijf jaar geleden? En waar staat RDF nu?
- VA is nodig en heul handig, ook voor leken

Methods and techniques

Data acquisition will be performed throughout the study. Ongoing sequencing efforts from the research group of Prof. Dr. Cuppen and collaborators will ensure more than adequate amounts of data will be at our disposal. Due to the accompanied scientific questions of this data, the method-development stages of this study will continue to be focussed and inspired by the end-goal: answering biological questions. Furthermore, continuing the data-acquisition in the second half of the study to will enable us to collaborate with the research community and showcase our innovative technology with new and exciting integrative biology studies.

To ensure the greatest compatibility and effectiveness, tight collaborations will be established between leading RDF-users and -developers in- and outside of biology. Triples for NGS- and MS-based data will be developed, taking into account the most commonly used format first. Since different databases can require a specific triple-structure, RDF-databases will be investigated on their ability to handle the large datasets efficiently, including their in- and output options. Selected research groups in Utrecht will be attracted to provide early feedback-rounds, focussed on usability and compatibility.

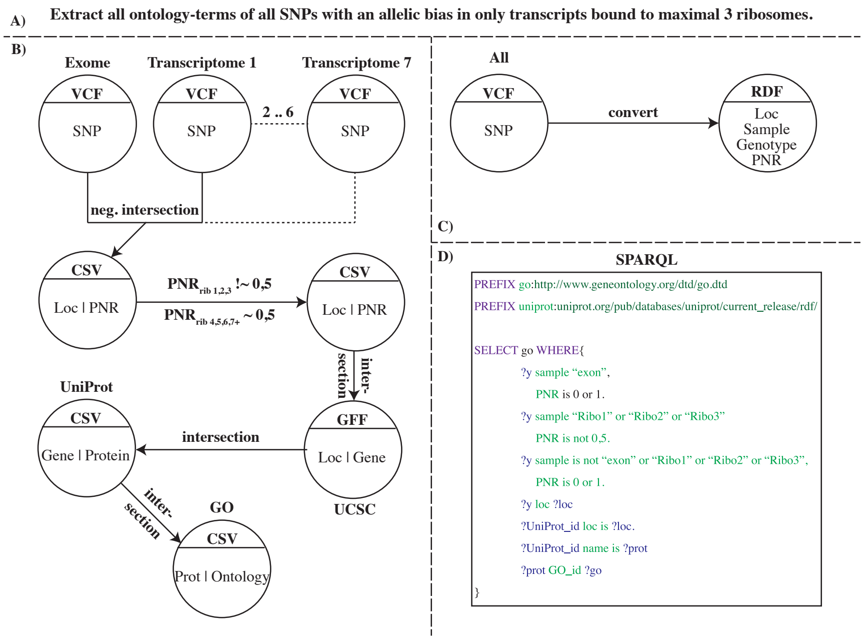
After completion of the triple-development of a data-format (i.e. end-users and the RDF-community have provided positive feedback), development of the conversion-tools is next. In this stage, we seek to expand the capabilities of current leading bioinformatical tools like Sambamba[4] and BIO-VCF[5], to capitalise on their multi-core capabilities. Furthermore, we will seek to collaborate with the current (public data-focussed) initiatives, like Bio2RDF[6] and BioInterchange[7] to ensure software-compatibility and limit redundancy.

The visualisation-subproject will have two phases. In the first phase, we will use a minimalistic model to develop the link between the SPARQL in- and output and *d3.js*-visualisation. The minimalistic model comprises of SNP- and RNA-based visual analytics-based solutions. Resulting methods can be directly used in other projects, focussing on the role of SNPs and transcription(-levels).

After a successful first phase, the second phase will broaden the available visualisations, by creating a modular dashboard. Every module will provide a particular visual (e.g. heat-map, scatter-plot) and will interact with both the SPARQL-input, -output and the other active modules. If a user would, for example, select a specific gene in the scatter-plot, the same data-point will be highlighted in the other modules. The cross-talk between modules has already been implemented in Epiviz2[8], which is highly appraised for this by its users. The order of development of specific modules will be primarily based on the wants and needs of the community, which will be gathered with the above-mentioned feedback-rounds.

Pilot-study

To asses to feasibility of this proposal, a small-scale pilot-study was performed on de data of van Heesch et al.[9]. This dataset includes transcriptome data of mRNA's, bound to a number of ribosomal units (1 to 7+) and matching exome-data. If one would be interested in the molecular functions of a gene with an allelic bias, a disproportional amount of time is lost on parsing, intersecting and downloading various types of data (fig. 1). With the current methods, twelve set-operations (e.g. intersections, unions) have to be performed on ±5gb of data. Furthermore, three datasets (±15gb) have to be completely downloaded once -until a new version is launched- before a simple exploratory question can be answered. Approximately three and a half hours was needed to perform this, in contrast to one hour with the proposed 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. First and foremost, this pilot illustrates the low-complex nature of the proposed methods. Secondly, it shows 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 with the currently used methods.



```
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**, they have to go through a series of parsing and interception steps, like in **B**. External sources have to be fully downloaded and converted, before use. Our proposed pipeline (shown in **C**), firstly converts the data to RDF. Then, a question can be formulated in SPARQL (**D**), incorporating relevant outside sources, which can be easily changed without having to juggle/download the data again.*

Research plan

Timetable

Proposed biological questions

Semester four to six will be used to perform multi-level analysis with the resulting methods and tools from the previous three semesters. The overall aim is to combine both proprietary and public data to execute previously impossible analyses. Of the public databases available, the most valuable for our purposes are those, that link omics-data to pathways and ontologies: *Pathway Commons*[10] and *Reactome*[11] on perturbed pathways in cancer, *RegulomeDB*[12] on linking non-coding regions to genes and *Gene Ontology*[13] and *KEGG-pathway*[14] for general ontologies and pathways. A subset of the addressed questions and proposed analyses is depicted below.

	Semesters					
	S1	S2	S3	S4	S5	S6
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				■	■	■
Non-coding structural factors of drug-resistance				▒	▒	
Single-cell analysis of SVs in cancer sub-populations					▒	▒
Writing thesis						■

i. **Non-coding structural factors of drug-resistance**

The Cancer Genomics Centre Netherlands (CGC.nl) is in the process of studying the effects of variants in coding regions in cancer, including factors of drug resistance. However, the data has not been used to study the non-coding regions, primarily because of the aforementioned limitations in both non-coding analysis and data-integration. Firstly, we will analyse the (epi)genomic data to identify cancer-specific SVs that are cis-acting on specific genes. Secondly, the data of the products of these genes (e.g. transcripts, proteins and metabolites) is integrated to infer possible consequences on these levels. Linking specific drug-resistance information will also enable us find patterns between the SVs, the identified (consequences of) genes and specific drugs. Since treatment of a single drug often leads to resistance by a bypass in the drug-inhibited pathway[15], we also integrate public and CGC.nl-data on perturbed pathways in cancer. This will elucidate the mechanisms of non-coding SV-induced drug-resistance in cancer samples and potentially identify new targets for treatment.

ii. **Single-cell analysis of SVs in cancer sub-populations**

Due to the advances in single-cell sequencing (CELL-seq) of both DNA and RNA, we are able to look at consequences of SVs on transcription in single cells. The innovation here is the fact that signals are not averaged out by multiple (asynchronous) cells and we can thus analyse the

> cell as part of a sub-population. By integrating CELL-seq DNA- and RNA-data of different sub-populations of (heterogeneous) cancer-samples, we can find the previously obscured direct (i.e. cis-acting) and indirect (i.e. trans-acting) consequences of SVs in specific sub-populations. Furthermore, integrating data of lineage-tracing between and within different sub-populations could identify causal non-coding SV-events in the progression of cancer. Linking the public data of ontology- and pathway-databases will enable us to infer specific sub-population changes in pathways as the consequence of SVs or de-regulated genes due to SVs.

This subset of questions shows the main innovative point of our proposed methods: they make it possible to efficiently and systematic analyse and integrate several omics-levels and multiple dimensions (e.g. drug-resistance, cancer type/sub-population), while allowing easy connection to public data.

Collaboration

By performing this research in the Hubrecht Institute, we surround ourself with various research-fields within the scope of Developmental Biology. One of the newest findings of the Hubrecht are Organoids, which provide a method to study heterogeneous tissues (e.g. cancer) in more detail, by providing clonal (i.e. homogeneous) cultured tissues. Groups in the Hubrecht are heavily involved in (inter)national consortia, like the *Cancer Genomics Centre*. This national consortium of research-groups, predominantly of the Hubrecht Institute and the Netherlands Cancer Institute (NKI), focusses on cancer's (epi)genetic alterations and responses to drugs. Data from this project will include various levels (e.g. (epi)genomics, phospho-proteomics) and dimensions (e.g. drug-responses, time-series). For the cancer sub-population study, a collaboration between the *van Oudenaarden*-group (lineage-tracing and CELL-seq) and the *Clevers*-group (cancer-biobank) will be formed.

Furthermore, the affiliation between the institute and Utrecht University will lead the more possibilities. A considerable amount research groups make use of the Utrecht DNA-sequencing Facility and the Netherlands Proteomics Centre, which ensures adequate amounts and variation in data and data-integration-based research questions.

Ties with international leaders in biology-related semantic web and visualisation technologies have been made and will continue to be expanded. Joachim Baran and Pjotr Prins have been heavily involved in the planning stages, being key players in handling various data-formats (into RDF) with *BIO-Ruby*. Communications with Artem Tarasov of *Sambamba* and Jerven Bolleman -key engineer of the *UniProt-RDF* project- have also been established.

Knowledge utilisation

The implementation of RDF will be swift since RDF already is a web-standard and a significant number of public biology-related sources are already in RDF-format. This will enable users of our methodologies to efficiently connect and integrate their data with public resources. Current statistical software, like the R environment, have packages (made by researchers from computer sciences) to extract and further analyse SPARQL-output. This means that users only have to learn SPARQL-queries, in order to use the proposed methods.

By enabling more users to use methods and sources of RDF, this research will, in a broader perspective, have a direct effect on the Semantic Web. By lowering the (bioinformatical) threshold for analysis, more data can be faster analysed by more people, further accelerating research. Users will also be able to tell their story (i.e. results) better. Psychologist will, for example, be able to get a better visualisation and thus understanding of a neuroscientist's work. Big pharmaceutical companies will be able to further include and analyse data of basic science, clinical trails and business-statistics with more efficiency. Moreover, the research-community will be one significant step further in dissecting the complex biology of cancer.

References

[1] Sabina Benko, Judy A Fantes, Jeanne Amiel, Dirk-Jan Kleinjan, Sophie Thomas, Jacqueline Ramsay, Negar Jamshidi, Abdelkader Essafi, Simon Heaney, Christopher T Gordon, David McBride, Christelle Golzio, Malcolm Fisher, Paul Perry, Véronique Abadie, Carmen Ayuso, Muriel Holder-Espinasse, Nicky Kilpatrick, Melissa M Lees, Arnaud Picard, I Karen Temple, Paul Thomas, Marie-Paule Vazquez, Michel Vekemans, Hugues Roest Crollius, Nicholas D Hastie, Arnold Munnich, Heather C Etchevers, Anna Pelet, Peter G Farlie, David R Fitzpatrick, and Stanislas Lyonnet. Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nat. Genet.*, 41(3):359–64, March 2009. ISSN 1546-1718. doi: 10.1038/ng.329. URL http://www.nature.com.proxy.library.uu.nl/ng/journal/v41/n3/abs/ng.329.html.

[2] Ingo Kurth, Eva Klopocki, Sigmar Stricker, Jolieke van Oosterwijk, Sebastian Vanek, Jens Altmann, Heliosa G Santos, Jeske J T van Harssel, Thomy de Ravel, Andrew O M Wilkie, Andreas Gal, and Stefan Mundlos. Duplications of noncoding elements 5' of SOX9 are associated with brachydactyly-anonychia. *Nat. Genet.*, 41(8):862–3, August 2009. ISSN 1546-1718. doi: 10.1038/ng0809-862. URL http://www.nature.com.proxy.library.uu.nl/ng/journal/v41/n8/full/ng0809-862.html.

[3] 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.

[4] Artem Tarasov. Sambamba, 2014.

[5] Naohisa Goto, Pjotr Prins, Mitsuteru Nakao, and Raoul Bonnal. BioRuby: bioinformatics software for the Ruby programming language. . . . , 26(20):2617–9, October 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq475. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2951089&tool=pmcentrez&rendertype=abstracthttp://bioinformatics.oxfordjournals.org/content/26/20/2617.short.

[6] 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.

[7] Joachim Baran. BioInterchange: An Open Source Framework for Transforming Heterogeneous Data Formats Into RDF. *In preperation.*

[8] 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.

[9] 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.

[10] Ethan G Cerami, Benjamin E Gross, Emek Demir, Igor Rodchenkov, Ozgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D Bader, and Chris Sander. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.*, 39(Database issue):D685–90, January 2011. ISSN 1362-4962. doi: 10.1093/nar/gkq1039. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013659&tool=pmcentrez&rendertype=abstract.

[11] 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.

[12] Alan P Boyle, Eurie L Hong, Manoj Hariharan, Yong Cheng, Marc A Schaub, Maya Kasowski, Konrad J Karczewski, Julie Park, Benjamin C Hitz, Shuai Weng, J Michael Cherry, and Michael Snyder. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.*, 22(9):1790–7, September 2012. ISSN 1549-5469. doi: 10.1101/gr.137323.112. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3431494&tool=pmcentrez&rendertype=abstract.

[13] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1):25–9, May 2000. ISSN 1061-4036. doi: 10.1038/75556. URL http://dx.doi.org/10.1038/75556.

[14] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30, January 2000. ISSN 0305-1048. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102409&tool=pmcentrez&rendertype=abstract.

[15] Anirudh Prahallad, Chong Sun, Sidong Huang, Federica Di Nicolantonio, Ramon Salazar, Davide Zecchin, Roderick L Beijersbergen, Alberto Bardelli, and René Bernards. Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature*, 483(7387):100–3, March 2012. ISSN 1476-4687. doi: 10.1038/nature10868. URL http://www.ncbi.nlm.nih.gov/pubmed/22281684.