Everything should be linked: linking and visualising data for dynamic 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

To date, studies on non-coding regions of the genome, specifically in cancer, have been limited. This is mainly due to the complex nature of putative functional elements in these regions. In parallel with the ENCODE-project, the interest in these regions has increased: researchers are beginning to study causal non-coding variations in cancer[4;5]. Due to the increase in popularity and cost-effectiveness of various omics-approaches, more and more data is becoming available. The complexity of integrating and analysing information of these approaches increases with every added omics-layer or dimension (e.g. time-series, treatments). When studying the effects of structural variants in non-coding regions in cancer, this complexity is further increased due to cancer-specific (e.g. heterogeneous samples, rapid evolution) and structural variant-specific (e.g. multiple types and consequences) factors.

The current methods for integrating and analysing these layers and dimensions have two significant limitations in their design: scalability and generality (i.e. the possibility to add more levels or dimensions). Moreover, there isn't an option to overview a dataset without filtering, dividing or restructuring the data. The integration of complex datasets is needed to understand the complex biology of cancer betterMunoz et al.[14], but is restricted by these limitations.Ã§

Enter the Semantic Web and its Resource Description Framework (RDF). A simple and flexible framework for describing anything about anything. Since every type of data can be translated to this universal language, integration of large datasets of different levels and dimensions becomes possible and a lot more feasible. When researchers have converted their local data to RDF, they can easily connect and combine it with public repositories, 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 user can subsequently visualise the SPARQL-results as a whole or filter them further.

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. These methods will enable further elucidation of the complex biology of, for example, cancer. Firstly, we will create the framework needed to design the missing tools for converting the most-used NGS-formats to RDF. Next, visualisations (based on visual analytics) of the biological RDF-data will be created, which will be used to perform previously impossible integration-focussed analyses on the consequences of structural variation in the non-coding regions of cancer-genomes.

Layman's summary

The biomedical research community wants to be able to combine and analyse a multitude of biological signals in one experiment because the biology of, for example, cancer is so complex. However, integrating diverse sets of biological signals is currently a serious challenge. To overcome this, we propose the use of Semantic Web-methods: these are specially designed for integrating vast amounts of different data. Furthermore, it allows users to describe, analyse and test their data interactively and dynamically via visual representations displayed in the browser.

Research on non-coding genomic regions, for example, would benefit greatly from these methods. It would enable studies on the complex cancer-causing consequences of changes in parts of chromosomes that do not contain a gene.

A preliminary study shows the added value of such methods in biology: enabling researchers to describe, analyse and test 20 times more biological questions in the same time, compared to conventional methods. We propose to develop these methods further for the research-community (and biology in particular), enabling us to perform 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. Moreover, these methods will be used to study the inter-level (e.g. transcriptome, proteome) consequences of structural variation in non-coding regions of the genome. 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 dynamic and interactive visualisation of RDF-data.

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

Finding causal genetic variation in the protein-coding regions of the genome has been the focus in the majority of genomics studies. However, these regions amount only to approximately two percent[1]. One of the primary reasons behind this is the relative uncomplicated nature of studying coding regions, as consequences on lower levels (e.g. transcription, proteins) are traceable[2]. This is in contrast to the non-coding regions, which often do not show a linear effect on other levels[?] . A good illustration of the complexity of the non-coding regions is the ENCyclopedia Of DNA Elements (ENCODE)-project[3], which contains over fifty different signals (e.g. histone methylation, DNase1 hypersensitivity).

The fact that non-coding regions often have roles in the regulation of distant genes (i.e. cis-acting) provides even more complexity to the analysis of structural variants (SVs) in these regions. For example, the Pierre Robin Syndrome (PRS): SVs (deletions or duplications) in the 3Mb surrounding the SOX9-gene in particular tissues are causative of the striking phenotype of undeveloped mandibles and tongue in children[4;5]. Studies on cancer-specific causative non-coding variation are beginning to emerge in the last two years, including colorectal- and skin-cancer[6;7], and computational methods for non-coding regions are just starting to come up in the literature of 2014[8;9].

The amount of (public) biological data has exploded in the last years (even outpacing Moore's law[1]). This is the result of the advances in omics-technologies like Next-Generation Sequencing (NGS) and Mass-Spectrometry (MS), in both performance and costs. The addition of other dimensions, like time-series or treatments, is a second factor for the highly complex nature of current biomedical research. While there are plenty of studies on single-level data analysis, both academia and industry agree that data-integration is essential to understanding the complex nature of biology more thoroughly[10–13].

However, only a few layers and dimensions have been integrated per study and results are -for the most part- cherry picked, instead of systematic. This is mainly due to the methods used in integration-studies: most of them are set up in the same manner as individual-level experiments, whereafter they are combined. These methods are limited due to the large amounts of parsing-time (i.e. the time to convert various file/region-formats). An example of the large amount of analytical time needed, when using these methods is the study of Munoz et al.[14]: every two months of data-accumulation costed two years of analysis. The limited number of truly integrative studies use

[1] A two-fold in- or decrease of a variable (here: dollar/nt) per two years.

computational approaches to reconstruct biological networks. While a valid strategy, scaling the analysis from the bacteria used by Karr et al. [15] and Lerman et al. [16] 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 the use of an inflexible 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, 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 [17–19]. 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 (essentially constructing a graph-based network): there is no need for additional (file)formats (fig.1).

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. [20]) are needed. Aside from the non-complex, flexible and self-describing nature of the RDF-data, triples can be seen as a modular directed graph: users can connect multiple relevant RDF-sources (e.g. UniProt andProteomics-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 an important issue and challenge since the beginning of big data, as Pavlopoulos et al. [21] stated in 2008. SPARQL provides the 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 [22]. Another advantage of using SPARQL is the increase in scalability by including multiple triplestores in the same query. By enabling the use of small and specific triplestores, such a federated query results in faster retrieval of the data.

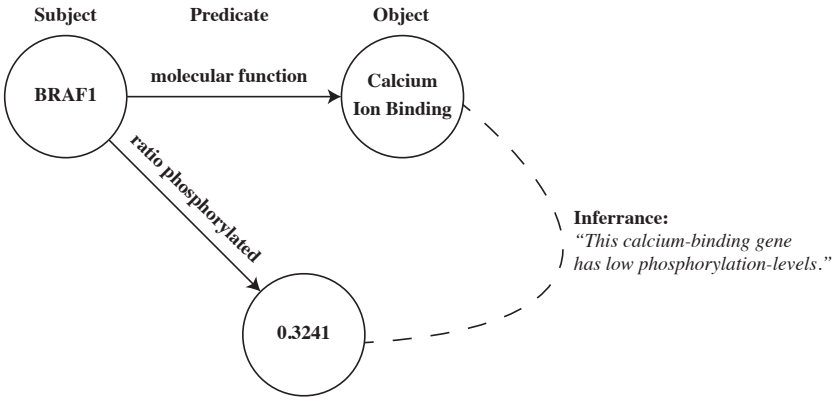


Figure 1: ***General outline of RDF.*** *By linking two triples by their common subjects, one can infer the relationship between the two objects via the predicates and find patterns: a gene, responsible for calcium ion binding, has a low phosphorylation level in the investigated sample.*

When data is incorporated in a Semantic Web RDF-database (TripleStore) and a relevant set of subjects, predicates and objects are extracted using SPARQL, the remaining dataset is still enormous. 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 [23] describe this field in 2005 as "*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. Visual analytics enables efficient exploratory analysis of the data by the user. Drug discovery is one of the leading areas in biological visual analytics, as it provides a more cost-effective method for analysing data of clinical trails [24].

The JavaScript library "Data-Driven Documents" (*D3.js*) is focussed on structuring data for dynamic web-based visualisation, which makes it well-suited for implementing linked data within visual analytics [25]. Moreover, 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: a notable biology-specific example of this is Epiviz2 [26]. However, this tool only takes an explicit set of data-formats and -levels and -more importantly- only shows a particular genomic region, instead of the complete scope. This can easily lead to cherry-picking, instead of data-focussed formulation and analysis of hypotheses.

The heterogeneous samples and datasets of cancer make it one of the most computationally demanding types of integrative biology. **We propose to use our methods to study to consequences of structural variations at non-coding loci in cancer on other levels.** These methods will enable research in this technical challenging topic, by decreasing the computational burden of data-handling, and increase the cognitive abilities of the user, by providing integrative visual interfaces.

Originality and innovative character

The 2014 survey of Gomez-Cabrero et al. [10] showed that 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. Data-storage, -exploration and -exploitation were found the be key. Their conclusions were best summarized by *the need for having exploration tools, which combine summary statistics and interactive visualisations, to analyse heterogeneous data-sets.*

There have been various studies on the integration of biological signals with the aid of semantic web technologies as the power of ontology-based entailment [2] reasoning is widely acknowledged [19]. However, the momentum was lacking: until 2014, big databases were not available in RDF-format. This meant that bioinformatical research involving RDF had little to no outside support, as they could only integrate proprietary data, like the RDF-methods used in microarray analyses by Szpakowski et al. [27] in 2009. Recently, EMBL-EBI has opened their RDF-platform, boasting six big data-sources (Gene Expression Atlas, ChEMBL, BioModels, Reactome, BioSamples and UniProt) [22]. This was the boost needed to further incorporate RDF in biological analyses.

However, there are two main limitations of this relatively young incorporation: a standard language for denoting biological triples (e.g. chromosome locations) is missing and the focus lies at linking database-accessions [28]. While the first limitation could also be a strength, as everybody can use their own dialect. However, a standardisation-step will lower the learning-curve, which will enable researchers in all fields of biology to benefit fully 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 to triples, like the *Variant Call Format* (VCF) and *Sequence Alignment Format* (SAM). An example of this is *Bio2RDF* [17]: an "*RDFizer*", which converts conventional databases, like the ones from NCBI, to triplestores. One of the leading innovative points of this proposal is the development of methods to handle the NGS- and MS-based

[2] The logical consequence of having two linked ontologies, thereby inferring an additional, encompassing relationship on the shared object/subject

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 proprietary NGS- and MS-data to existing RDF-databases.

The implementation of web-based visual analytics for RDF-databases is another leading 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 has 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. 2). 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 [23;29].

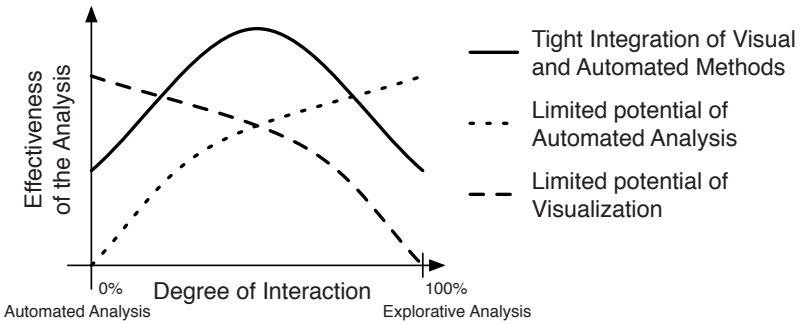


Figure 2: ***Trade-offs between automated and explorative analysis.*** *By combining automated analyses, where appropriate, with the background and intuition of the user, an optimal amount of effectivity can be attained. Picture taken from Keim et al.* [29].

While there has already been a considerable amount of work in the field of computational cancer research, the vast majority of large-scale integrative studies have been conducted on the coding-regions of the genome [3]. Genome-Wide Association Studies (GWASs) on a broad range of (hereditary) cancer-types 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 assess the consequences of individual variations in non-coding regions [6;8]. However, no large-scale integrative studies have been performed, which is partly due to the current state of integrative methods.

Due to the affiliation with the University Medical Center Utrecht (UMCU), we are in a unique position to test our hypotheses and methods in both research and clinical settings. Furthermore, the HUB-biobank in the Hubrecht Institute also enables us to perform analyses on organoids, providing us a stable and homogeneous *in vitro* platform for (validation-)studies on cancer-samples. With the new methods, we will be in a position to perform fully integrative studies on the underlying mechanisms and consequences of (structural) variation in cancer.

Pilot-study

To asses to feasibility of this proposal, a small-scale pilot-study was performed on de data of van Heesch et al. [30]. 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. 3). 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.

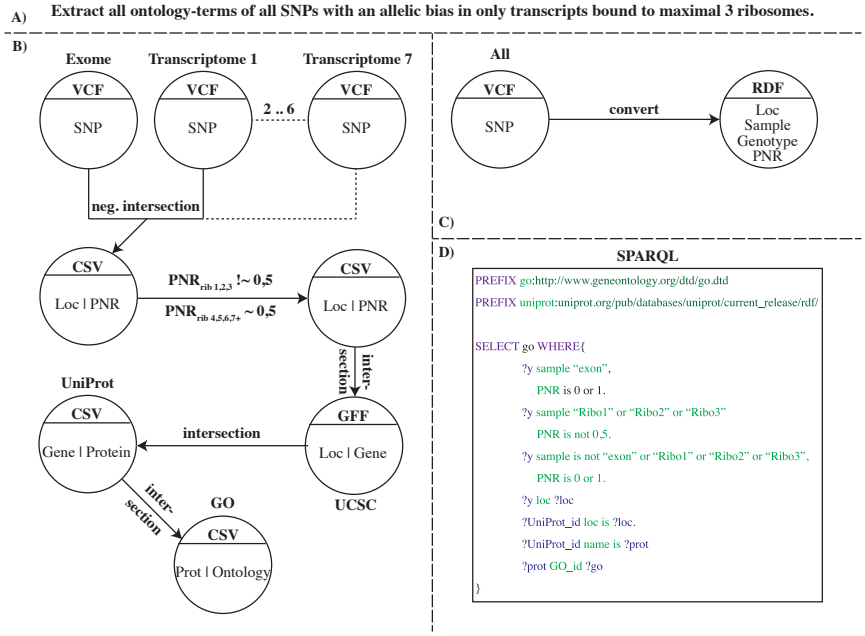


Figure 3: ***Differences between current integration techniques and RDF.*** *When a researcher has a question like* ***A****, they have to go through a series of parsing and interception steps, like in* ***B****. External sources have to be fully downloaded and converted, before use. Our proposed pipeline (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.*

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.

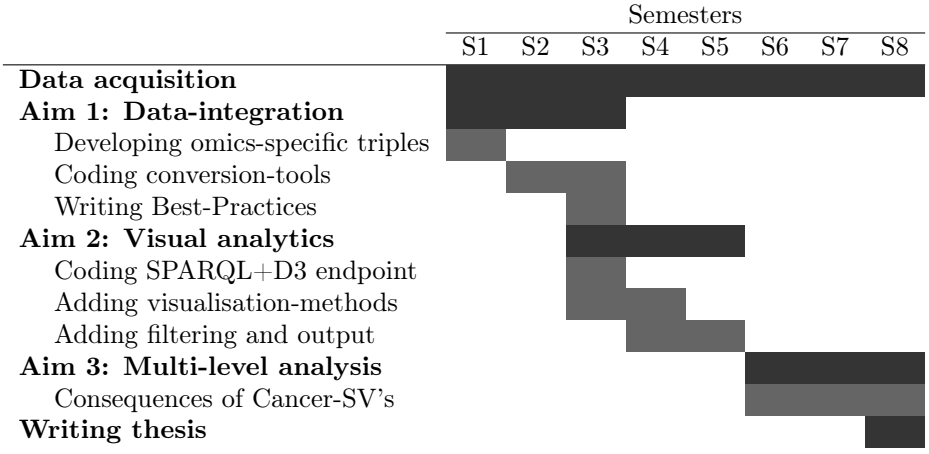
After completion of the triple-development of a data-format (i.e. end-users and the RDF-community have provided positive feedback), development of the conversion-tools is next. In this stage, we seek to expand the capabilities of current leading bioinformatical tools like Sambamba [31] and BIO-VCF [32], to capitalise on their multi-core capabilities. Furthermore, we will seek to collaborate with the current (public data-focussed) initiatives, like Bio2RDF [17] and BioInterchange [33] 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 [26], 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.

Research plan

Timetable



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, phosphomics) and dimensions (e.g. drug-responses, time-series).

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] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, N Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowski, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, and J Szustakowki. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001. ISSN 0028-0836. doi: 10.1038/35057062. URL http://www.ncbi.nlm.nih.gov/pubmed/11237011.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] Xiaohong Cao, Karen B Maloney, and Vladimir Brusic. Data mining of cancer vaccine trials: a bird's-eye view. *Immunome Res.*, 4:7, January 2008. ISSN 1745-7580. doi: 10.1186/1745-7580-4-7. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2639543&tool=pmcentrez&rendertype=abstract.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] Artem Tarasov. Sambamba, 2014.

[32] 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.

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