Understanding the functional 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, structural variation in the non-coding regions of the genome is one of the least studied genomic events. Systematic linear approaches for elucidating the functional effects are rarely successful, which is mainly due to the functional complexity of non-coding regions (e.g. cis- and trans-acting elements, co-activation, non-coding RNA interaction). These analyses are further complicated by the diversity in types and consequences of these genomic alterations. Integration and visualisation of complex, multi-layered datasets are needed to better understand the functional effects of these events[1].

Within the realm of systems biology, graph-based approaches 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. 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).

For larger and more complex datasets like these, the bioinformatics community -following other big data sciences- is starting to gravitate towards the Resource Description Framework (RDF). This is a simple and flexible graph-integration approach, which also allows for easy connection to (web-based) public repositories. The formation and testing of hypotheses on these created and/or combined networks is made straightforward by using simple SPARQL-queries and subsequent visual analytics-methodologies.

Here, we propose the use of graph-based methods to decrease the complexity of integrating and visualizing multi-level and -dimensional biological data. By integrating patient-derived data of the UMC Utrecht, we are in the premier position to uncover new biological insights into the complex biology of non-coding structural variants. Our resulting methodologies and discoveries could aid a large community of both scientists and patients, by enabling further elucidation in congenital disease and cancer.

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

Background

The amount of (public) biological data has exploded in the last years (even outpacing Moore's law*). 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 [2–5].

The vast majority of large-scale integrative studies have been conducted on the coding-regions of the genome [6]. Although finding functional genetic variation in the protein-coding regions of the genome has thus been the focus, these regions amount only to approximately two percent of the genome [7]. One of the primary reasons behind this focus is the relative uncomplicated nature of studying coding regions, as consequences on lower levels (e.g. transcription, proteins) are linearly traceable [8]. This is in contrast to the non-coding regions, which often do not show a linear effect on other levels [9]. A good illustration of the complexity of the non-coding regions is the ENCyclopedia Of DNA Elements (ENCODE)-project [6], which contains over fifty different signals (e.g. histone methylation, DNase1 hypersensitivity). The fact that non-coding regions have roles in the regulation of both close and distant genes (i.e. *cis*- and *trans*-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 [10;11].

Genome-Wide Association Studies (GWASs) on a broad range of congenital and acquired diseases (e.g. cancer) have shown that non-coding locations are associated with these diseases. Until 2013, however, tools and sources to systemically explore and analyse the functional consequences of 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 [12;13]. Studies on cancer-specific causative non-coding variation are beginning to emerge in the last two years, including colorectal- and skin-cancer [12;14], and computational methods for non-coding regions are just starting to come up in the literature of 2014 [13;15]. However, no large-scale integrative studies have been performed, which is partly due to the current state of integrative methods.

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

hoe valt dit te halen? (graphs: integratie en visualisatie)

big data graph-methods

Waar stond RDF vijf jaar geleden? En waar staat RDF nu? VA is nodig en heul handig, ook voor leken

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. [1]: 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 a valid strategy, scaling the analysis from the bacteria used by Karr et al. [16] and Lerman et al. [17] 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 graph-based integration methods.** The most apparent method for this is from 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 big data science, enabling users to integrate and search data based on semantics. 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: there is no need for additional (file)formats. By linking triples to each other by either a common object or subject (essentially constructing a graph-based network), new relationships can be inferred (fig.1).

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 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 an important issue and challenge since the beginning of big data, as Pavlopoulos et al. [18] 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 [19]. 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.

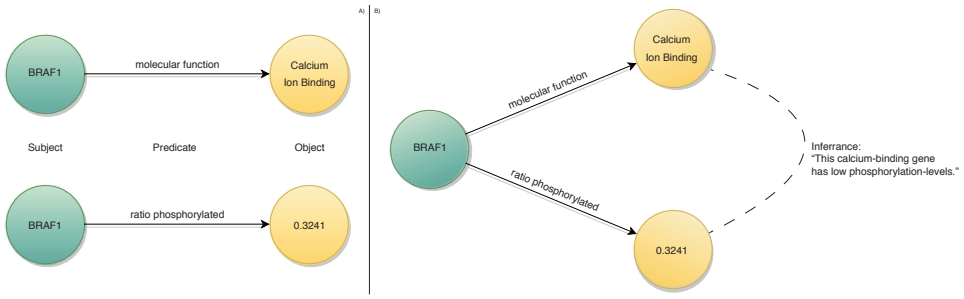


Figure 1: ***General outline of RDF.*** *Two triples (as shown in* ***A****) can be linked by their common subjects (as shown in* ***B****), one can infer the relationship between the two objects via the predicates and find patterns: a gene, responsible for calcium ion binding, has a low phosphorylation level in the investigated sample.*

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† reasoning is widely acknowledged [20]. 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. [21] in 2009. Recently, EMBL-EBI has opened their RDF-platform, boasting six big data-sources (Gene Expression Atlas, ChEMBL, BioModels, Reactome, BioSamples and UniProt) [19]. This was the boost needed to further incorporate RDF in biological analyses.

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 [22;23].

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

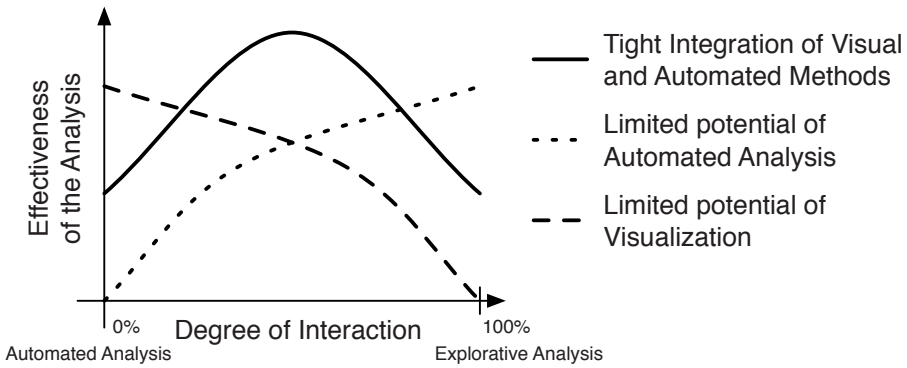


Figure 2: ***Trade-offs between automated and explorative analysis.*** *By combining automated analyses, where appropriate, with the background and intuition of the user, an optimal amount of effectivity can be attained. Picture taken from Keim et al.* [23].

Preliminary studies

During my Masters studies, I have already implemented graph-based methods for different purposes. As an illustration of the strengths of multilayer network-analysis, I describe a small example from the work I have done as an intern an the Sanger Institute[‡]. Furthermore, we show trough a pilot-study the added efficiency of our proposed methods.

Cis-regulatory regions: potential

When analysing predicted altered and/or de novo transcription factor binding sites (TFBS) by single-nucleotide polymorphisms (SNPs) in non-coding regions of over 1300 melanoma-patients, no overrepresented TFBS were found. We then discided to use graph-based data integration methods by linking the SNPs to regions in the TF-ChIP data of ENCODE[6]. Furthermore, two triple stores were added: TcoF -containing TF-interacting proteins and co-factors- and gene ontology terms form the GO consortium. We found that there was a clear overrepresentation of TFs binding to transcription co-activators, like NCOA6, which have a functional role in *vitamin D receptor binding*. Further analyses showed more evidence of the importance of these findings in familial melanoma (e.g. co-segregation in melanoma-prone families). Without the multilayer network-analysis, in this case of genomics and epigenomics data, such new testable hypotheses are often not found. And without RDF, different databases and -sources would be much harder to integrate for (exploratory) analyses.

[‡]Unpublished data, discretion appreciated.

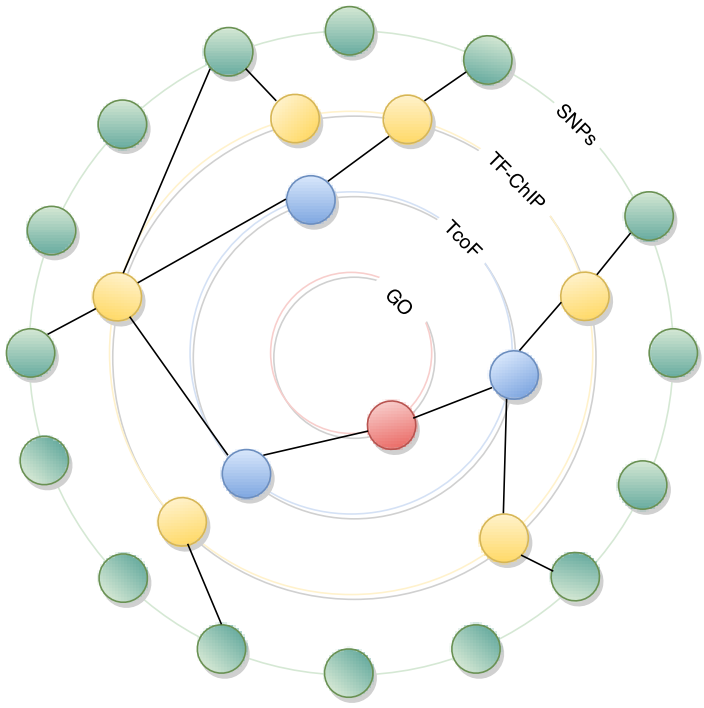


Figure 3: ***Graph of GO:0042809.*** *From a large dataset of mutations in melanoma-patients (outside ring) to a new -testable- hypothesis, alteration in vitamin D receptor binding (inside ring), via a TFBS-regions and a TF-protein interaction database.*

Ribosomal profiling and gene ontology: logistic efficiency

To asses to feasibility of this proposal, a small-scale pilot-study was performed on the data of van Heesch et al. [24]. 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. 4). With the conventional 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 data to a triple-store: 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.

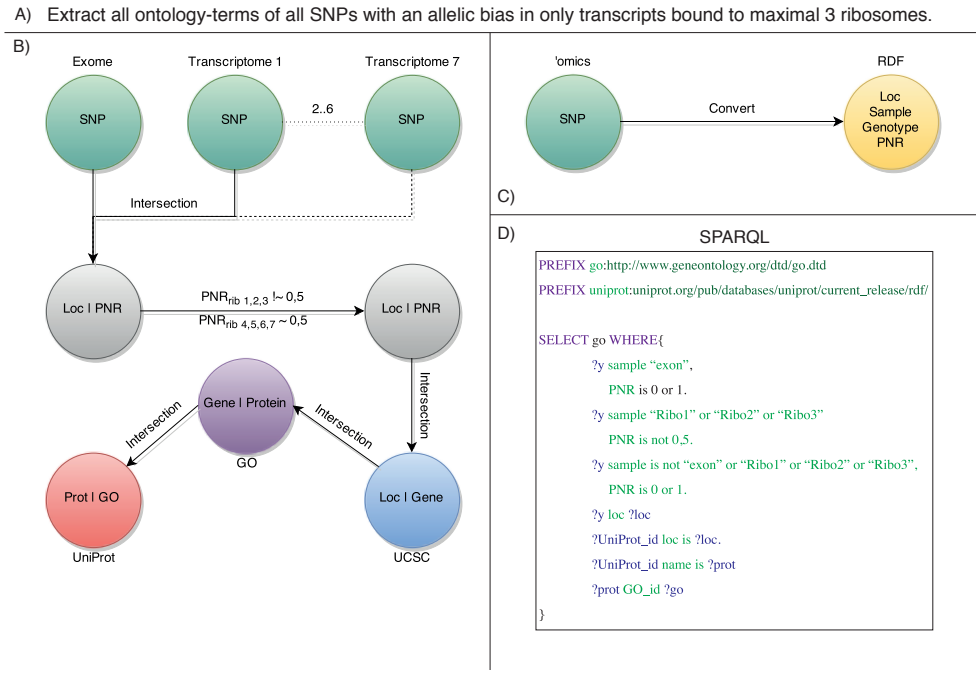


Figure 4: ***Differences between current integration techniques and RDF.*** *When a researcher has a question like* ***A****, they have to go through a series of parsing and interception steps, like in* ***B****. External sources have to be fully downloaded and converted, before use. Our proposed pipeline (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.*

Scope & aims

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.

Relevance

The 2014 survey of Gomez-Cabrero et al. [2] 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.*

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.

Experimental strategy

Placement and institute

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.

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.

Technische themes:

Biologische themes:

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. **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 [25], 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.

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.

Timetable

References

[1] 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.

[2] 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.

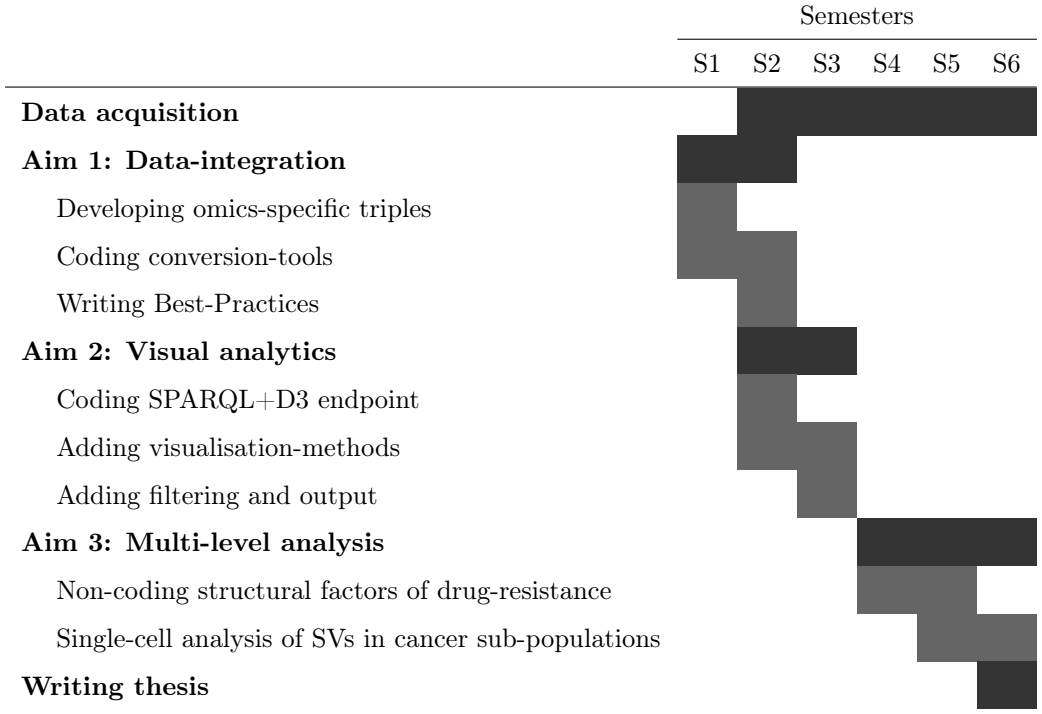
[3] Curtis Huttenhower and Oliver Hofmann. A quick guide to large-scale genomic data mining. *PLoS Comput. Biol.*, 6(5):e1000779, May 2010. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000779. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2877728&tool=pmcentrez&rendertype=abstract.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Christine P Bird, Barbara E Stranger, and Emmanouil T Dermitzakis. Functional variation and evolution of non-coding DNA. *Curr. Opin. Genet. Dev.*, 16(6):559–64, December 2006. ISSN 0959-437X. doi: 10.1016/j.gde.2006.10.003. URL http://www.sciencedirect.com/science/article/pii/S0959437X06002024.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.