

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

Summary of the research

Effects of structural variation in the non-coding regions of the human genome are rarely studied which is strong contrast to their known role in disease³⁹. Systematic approaches for elucidating their functional effects are rarely successful, mainly due to the complexity of non-coding regions (e.g. cis- and trans-acting elements, co-activation, non-coding RNA interaction). Predictions on functionality are further complicated by the diversity in types and consequences of structural variants. Integration and visualisation of complex, multi-layered datasets are needed to better understand the functional effects of these events²².

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^{7;8;10;33}.

The vast majority of large-scale integrative studies have been conducted on the coding-regions of the genome⁶. 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 genome¹⁹. Of the remaining 98%, approximately 6.2% is theorized to be biologically functional (i.e. is under negative evolutionary pressure)²⁹. 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 traceable²¹.

This is in contrast to the non-coding regions, which often do not show a linear effect on other levels⁴. A good illustration of the complexity of the non-coding regions is the ENCyclopedia Of DNA Elements (ENCODE)-project⁶, 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^{3;18}.

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. This -of course- to be expected, with the discoveries of various functions of the non-coding regions: from (in)directly regulating the transcription-machinery, to playing mayor roles in mRNA-degradation and from post-transcriptional modifications, to directly affecting the localisation of the transcript^{2;25}. Pseudogenes are also categorised as non-coding, but can be resurrected by gene conversion due to structural variants¹⁶. Pseudogenes can perform as decoy for their coding homologs²⁶ (as was found to be the case in both oncogenes and tumour suppressors) and lead to RNA-interference by pseudogene-transcribed endo-siRNAs³⁶.

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

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^{15;23}. Studies on cancer-specific causative non-coding variation are beginning to emerge in the last two years, including colorectal- and skin-cancer^{9;23}, and computational methods for non-coding regions are just starting to come up in the literature of 2014^{15;17}.

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.²²: 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.¹³ and Lerman et al.²⁰ 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 abstract and 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 (remote) 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.

There have been various studies on the integration of biological signals with the aid of semantic web technologies: the power of ontology-based entailment[†] reasoning is widely acknowledged³⁰. However, the momentum was lacking: until 2014, big databases were not available in RDF-format. This meant that bioinformatical research, using graph-based methodology through RDF, had little to no outside support, as they could only integrate proprietary data. An example of this are the methods used in microarray analyses by Szpakowski et al.³⁵ in 2009. Recently, EMBL-EBI has opened their RDF-platform, boasting six big data-sources (Gene Expression Atlas, ChEMBL, BioModels, Reactome, BioSamples and UniProt)¹¹. This was the boost needed to further incorporate graph-based integration via RDF in biological analyses.

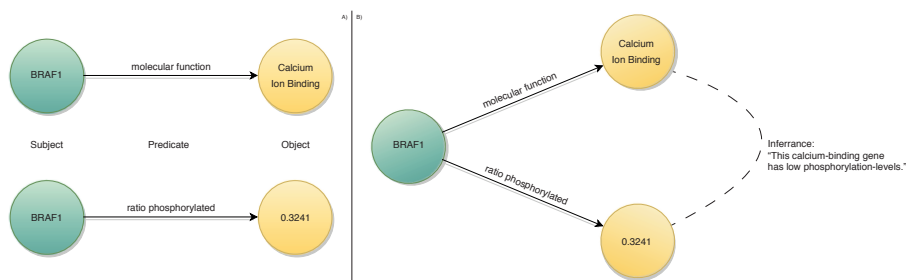


Figure 1: General outline of RDF. A) By linking two triples by their common subjects, **B)** one can infer the relationship between the two objects via the predicates and find patterns.

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.²⁴ 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-EBI¹¹. 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.

Combining these improvements in searching and linking networks with web-based visual analytics 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. Not only can data be more effectively analysed, but it can also be better understood and presented, due to the ability to provide an overview of the complete dataset^{14;37}.

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

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 at the Sanger Institute[‡]. Furthermore, we show through 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 decided to use graph-based data integration methods by linking the SNPs to regions in the TF-ChIP data of ENCODE⁶. Furthermore, two triple stores were added: TcoF³² -containing TF-interacting proteins and co-factors- and gene ontology terms form the GO consortium¹. 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.

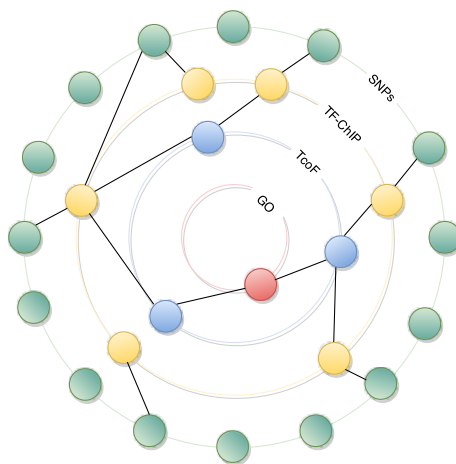


Figure 2: Graph of GO:0042809. From a large dataset of mutations in melanoma-patients (outside ring) to a new -testable- hypothesis, alteration in vitamin D receptor binding (inside ring), via a TFBS-regions and a TF-protein interaction database.

[‡]Unpublished data, discretion appreciated.

Ribosomal profiling and gene ontology: logistic efficiency

A small-scale pilot-study was performed on the data of van Heesch et al.³⁸. 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 conventional methods, twelve set-operations (e.g. intersections, unions) have to be performed on $\pm 5\text{gb}$ of data. Furthermore, three datasets ($\pm 15\text{gb}$) 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.

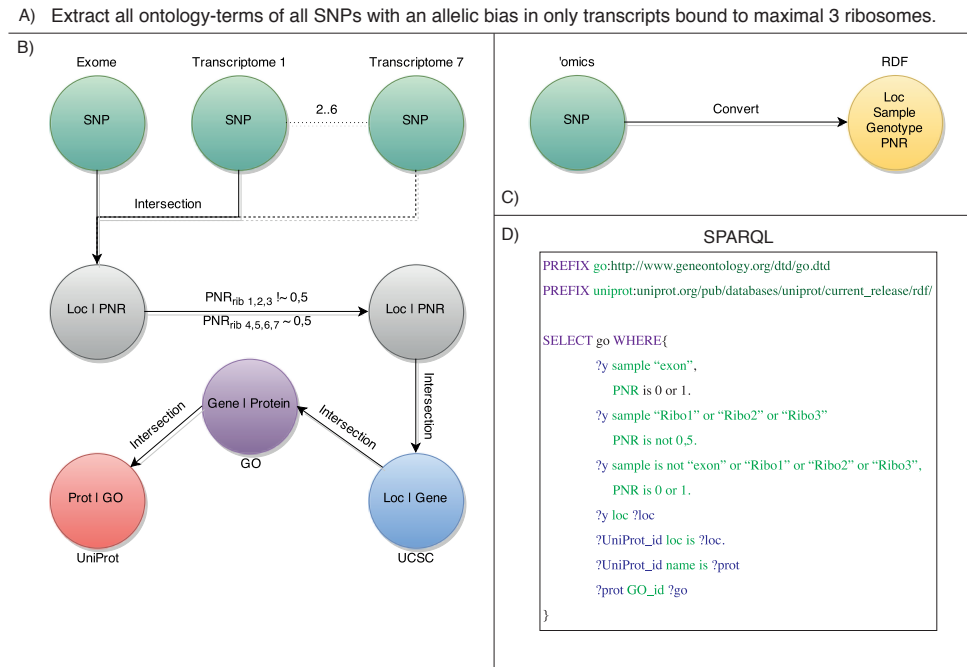


Figure 3: Differences between current integration techniques and RDF. For a question like **A**, series of parsing and interception steps are needed (**B**). External sources have to be fully downloaded and converted. When using RDF, all data is converted into triples once (**C**) and subsequent queries can be made in SPARQL (**D**).

SCOPE & AIMS

The aim of this project is to create new biological insights in effects 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 and adaptation of novel methodologies for integration and analysis of large scale, multi-dimensional biological data. And, more specifically, for structural variants in non-coding regions.
2. **Multi-level analysis** Multi-level and -dimensional integrative analysis to elucidate the consequences of genomic structural variations in non-coding regions, which are found in patients with congenital disease and cancer.

RELEVANCE

The 2014 survey of Gomez-Cabrero et al.⁷ 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 to 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*.

HIER NOG IETS OVER NON_CODING_NC

The implementation of our proposed graph-based methods 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 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 our approaches, this research will, in a broader perspective, have a direct effect on the semantic web and biological databases. 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 trials and business-statistics with more efficiency. Moreover, the research-community will be one significant step further in dissecting the complex biology of (structural variation in) non-coding regions of the genome.

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 <http://hub4organoids.eu/> in the Hubrecht Institute enables us to perform analyses on organoids, providing us a stable and homogeneous *in vitro* platform for validations in (non)cancer-samples. With the proposed methods, we will be in a position to perform integrative studies with appropriate biological validation on the mechanisms and consequences of (structural) variation.

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.

A considerable amount of Dutch research groups make use of the Utrecht DNA-sequencing Facility(UDsF) and the Netherlands Proteomics Centre, which ensures adequate opportunities for collaborations and data-integration-based research questions. Furthermore, the newest generation of DNA-sequencing methods from Oxford NanoPore and Pacific Biosciences are being deployed in the UDsF. These technologies result in larger sequenced stretches of DNA, which will make a considerable impact in the identification and analysis of larger structural variants.

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.

Technical themes

Before any analysis on the effects of non-coding structural variants can occur, we firstly lay a strong technical foundation. The application of graph-based methods will be largely based on the RDF-platform. While biology-related triplestores (RDF-databases), conversion-tools and basic ontologies are already made, we only have to focus on the specific missing elements.

Data usage

Local data will be acquired per experiment (see *biological themes*). Conversion of these datasets into triples is done after variant calling, as this is the earliest time in the experiment where we can filter and annotate. Post-calling integration keeps the size of the network to a minimum and allows us to create a general pipeline for conversion into RDF[§].

As with the local data, public data sources will be linked on a per-experiment base. Relevant non-RDF databases (e.g. GO¹, KEGG¹², TcoF³², RegulomeDB⁵) have to be converted, most preferably with the aid of the database-curators (ensuring a compatible and a stable resource for the community).

The Cancer Genome Atlas (TCGA) has been translated into a RDF-resource in 2014³¹. The atlas includes whole-genome (*Affy SNP 6.0*-) copy-number variation data. The fact that this resource also includes clinical, transcriptomic, epigenomic and proteomic data, makes it a very valuable resource for our experiments. Other existing triple-store, like from EBI¹¹, will also be very valuable and used as-is.

Integration and visualisation

iets over wat we gaan doen voor visual analytics van de graphs en de mogelijkheid om een centraal -uitbreidbaar-platform te maken voor sparql-queries icm VA. (voortborduren op dingen als Epiviz en rdf:SynopsViz. http://swui.semanticweb.org/swui06/papers/Karger/Pathetic_Fallacy.html laat mooi een groot probleem zien: een rdf-hairball zit niemand op te wachten (i.e. de echte graph is dus niet goed voor analyse: VA is nodig: VA is het beste als een browser waarin je lekker op los kan klikken.)

In vitro validation

crispr/cas9: mogelijkheid om een select aantal candidates in deatil uit te werken

[§]Tools for VCF-file conversion to triples are already in development by both F. Strozzi and P. Prins^{28;34}

Biological themes

This set of questions (ordered descending on $\frac{Pay}{Risc}$) shows the main innovative point of our proposed methods: they enable us to analyse and integrate several 'omics-levels and multiple dimensions (e.g. drug-resistance, cancer types), while allowing easy connection to public data.

Study recurrent non-coding SVs in public databases

adaasafssf

Identify effects of Chromotripsis induced SVs ??

Wigard heeft data van chromotripsis data patienten die volgens mij deels publiek is. [ref: Deciphering the pathogenic consequences of chromosomal aberrations in human genetic disease.] Wij willen dit natuurlijk verder trekken maar is een waardevolle dataset.

Study non-coding SVs in healthy Adult Stemcells

Volgens mij zou Ruben dat helemaal cool vinden

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.

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 pathway²⁷, 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.

TIMETABLE

	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						

REFERENCES

- [1] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1):25–9.
- [2] Barrett, L. W., Fletcher, S., and Wilton, S. D. (2012). Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cell. Mol. Life Sci.*, 69(21):3613–34.
- [3] Benko, S., Fantes, J. A., Amiel, J., Kleinjan, D.-J., Thomas, S., Ramsay, J., Jamshidi, N., Essafi, A., Heaney, S., Gordon, C. T., McBride, D., Golzio, C., Fisher, M., Perry, P., Abadie, V., Ayuso, C., Holder-Espinasse, M., Kilpatrick, N., Lees, M. M., Picard, A., Temple, I. K., Thomas, P., Vazquez, M.-P., Vekemans, M., Roest Crolius, H., Hastie, N. D., Munnich, A., Etchevers, H. C., Pelet, A., Farlie, P. G., Fitzpatrick, D. R., and Lyonnet, S. (2009). Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nat. Genet.*, 41(3):359–64.
- [4] Bird, C. P., Stranger, B. E., and Dermitzakis, E. T. (2006). Functional variation and evolution of non-coding DNA. *Curr. Opin. Genet. Dev.*, 16(6):559–64.
- [5] Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park,

- J., Hitz, B. C., Weng, S., Cherry, J. M., and Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.*, 22(9):1790–7.
- [6] ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696):636–40.
- [7] Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merckenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., and Tegnér, J. (2014). Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.*, 8(Suppl 2):I1.
- [8] Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M. T., and Beyene, J. (2009). Data integration in genetics and genomics: methods and challenges. *Hum. Genomics Proteomics*, 2009(1):869093–.
- [9] Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V., Chin, L., and Garraway, L. A. (2013). Highly recurrent TERT promoter mutations in human melanoma. *Science*, 339(6122):957–9.
- [10] Huttenhower, C. and Hofmann, O. (2010). A quick guide to large-scale genomic data mining. *PLoS Comput. Biol.*, 6(5):e1000779.
- [11] Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., Wimalaratne, S. M., Martin, M., Le Novère, N., Parkinson, H., Birney, E., and Jenkinson, A. M. (2014). The EBI RDF platform: linked open data for the life sciences. *Bioinformatics*, 30(9):1338–9.
- [12] Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30.
- [13] Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., Assad-Garcia, N., Glass, J. I., and Covert, M. W. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401.
- [14] Keim, D., Mansmann, F., and Thomas, J. (2010). Visual analytics: how much visualization and how much analytics? *ACM SIGKDD Explor. ...*, 11(2):5–8.
- [15] Khurana, E., Fu, Y., Colonna, V., Mu, X. J., Kang, H. M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., Das, J., Abyzov, A., Balasubramanian, S., Beal, K., Chakravarty, D., Challis, D., Chen, Y., Clarke, D., Clarke, L., Cunningham, F., Evani, U. S., Flicek, P., Fragoza, R., Garrison, E., Gibbs, R., Gümüs, Z. H., Herrero, J., Kitabayashi, N., Kong, Y., Lage, K., Liliushvili, V., Lipkin, S. M., MacArthur, D. G., Marth, G., Muzny, D., Pers, T. H., Ritchie, G. R. S., Rosenfeld, J. A., Sisú, C., Wei, X., Wilson, M., Xue, Y., Yu, F., Dermitzakis, E. T., Yu, H., Rubin, M. A., Tyler-Smith, C., and Gerstein, M. (2013). Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science*, 342(6154):1235587.
- [16] Kidd, J. M., Graves, T., Newman, T. L., Fulton, R., Hayden, H. S., Malig, M., Kallicki, J., Kaul, R., Wilson, R. K., and Eichler, E. E. (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell*, 143(5):837–47.
- [17] Kircher, M., Witten, D. M., Jain, P., O’Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, 46(3):310–5.
- [18] Kurth, I., Klopocki, E., Stricker, S., van Oosterwijk, J., Vanek, S., Altmann, J., Santos, H. G., van Harssel, J. J. T., de Ravel, T., Wilkie, A. O. M., Gal, A., and Mundlos, S. (2009). Duplications of noncoding elements 5’ of SOX9 are associated with brachydactyly-anonychia. *Nat. Genet.*, 41(8):862–3.
- [19] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A.,

- Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissole, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., and Szustakowski, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.
- [20] Lerman, J. A., Hyduke, D. R., Latif, H., Portnoy, V. A., Lewis, N. E., Orth, J. D., Schrimpe-Rutledge, A. C., Smith, R. D., Adkins, J. N., Zengler, K., and Palsson, B. O. (2012). In silico method for modelling metabolism and gene product expression at genome scale. *Nat. Commun.*, 3:929.
- [21] McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, 26(16):2069–70.
- [22] Munoz, J., Low, T. Y., Kok, Y. J., Chin, A., Frese, C. K., Ding, V., Choo, A., and Heck, A. J. R. (2011). The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. *Mol. Syst. Biol.*, 7:550.
- [23] Ongen, H., Andersen, C. L., Bramsen, J. B., Oster, B., Rasmussen, M. H., Ferreira, P. G., Sandoval, J., Vidal, E., Whiffin, N., Planchon, A., Padioleau, I., Bielser, D., Romano, L., Tomlinson, I., Houlston, R. S., Esteller, M., Orntoft, T. F., and Dermizakis, E. T. (2014). Putative cis-regulatory drivers in colorectal cancer. *Nature*.
- [24] Pavlopoulos, G. A., Wegener, A.-L., and Schneider, R. (2008). A survey of visualization tools for biological network analysis. *BioData Min.*, 1:12.
- [25] Pichon, X., A. Wilson, L., Stoneley, M., Bastide, A., A King, H., Somers, J., and E Willis, A. (2012). RNA Binding Protein/RNA Element Interactions and the Control of Translation. *Curr. Protein Pept. Sci.*, 13:294–304.
- [26] Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W. J., and Pandolfi, P. P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, 465(7301):1033–8.

- [27] Prahallad, A., Sun, C., Huang, S., Di Nicolantonio, F., Salazar, R., Zecchin, D., Beijersbergen, R. L., Bardelli, A., and Bernards, R. (2012). Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature*, 483(7387):100–3.
- [28] Prins, P. (2014). Bioruby-rdf.
- [29] Rands, C. M., Meader, S., Ponting, C. P., and Lunter, G. (2014). 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLoS Genet.*, 10(7):e1004525.
- [30] Sahoo, S. S., Bodenreider, O., Rutter, J. L., Skinner, K. J., and Sheth, A. P. (2008). An ontology-driven semantic mashup of gene and biological pathway information: application to the domain of nicotine dependence. *J. Biomed. Inform.*, 41(5):752–65.
- [31] Saleem, M., Padmanabhuni, S. S., Ngonga Ngomo, A.-C., Iqbal, A., Almeida, J. S., Decker, S., and Deus, H. F. (2014). TopFed: TCGA Tailored Federated Query Processing and Linking to LOD. *J. Biomed. Semantics*, 5(1):47.
- [32] Schaefer, U., Schmeier, S., and Bajic, V. B. (2011). TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins. *Nucleic Acids Res.*, 39(Database issue):D106–10.
- [33] Searls, D. B. (2005). Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.*, 4(1):45–58.
- [34] Strozzi, F. (2013). Bioruby-vcf2rdf.
- [35] Szpakowski, S., McCusker, J., and Krauthammer, M. (2009). Using semantic web technologies to annotate and align microarray designs. *Cancer Inform.*, 8:65–73.
- [36] Tam, O. H., Aravin, A. A., Stein, P., Girard, A., Murchison, E. P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R. M., and Hannon, G. J. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, 453(7194):534–8.
- [37] Thomas, J. and Cook, K. (2005). Illuminating the path: The research and development agenda for visual analytics. *IEEE Comput. Soc.*, pages 19–32.
- [38] van Heesch, S., van Iterson, M., Jacobi, J., Boymans, S., Essers, P. B., de Bruijn, E., Hao, W., Macinnes, A. W., Cuppen, E., and Simonis, M. (2014). Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biol.*, 15(1):R6.
- [39] Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.*, 14(2):125–38.