

Finding novel associations across domains using linked data: a case study on genetic variants disrupting transcription start sites

Eelke van der Horst^{*1}, Rajaram Kaliyaperumal^{*1}, Zuotian Tatum¹, Mark Thompson¹, Erik Schultes^{1,2}, Eleni Mina¹, Ivo F.A.C. Fokkema¹, Johan T. den Dunnen¹, Marco Roos¹, Jeroen F.J. Laros¹, Barend Mons¹, Kristina M. Hettne^{##1}, Peter A.C. 't Hoen^{##1}

¹Leiden University Medical Center, Leiden, the Netherlands

²Leiden Institute of Advanced Computer Science, Leiden, the Netherlands
eelkevanderhorst@gmail.com, r.kaliyaperumal@lumc.nl,
ztatum@live.com, {m.thompson, E.A.Schultes, E.Mina,
I.F.A.C.Fokkema}@lumc.nl, ddunnen@humgen.nl, {m.roos,
J.F.J.Laros, B.Mons, k.m.hettne, P.A.C._t_Hoen}@lumc.nl

##shared last author

Abstract. With the widespread use of Next Generation Sequencing technologies, the primary bottleneck of genetic research has shifted from data production to data analysis. However, heterogeneous data sets makes comparisons and integration challenging and time consuming. Here, we apply a data interoperability approach that provides unambiguous (machine readable) description of genomic annotations based on nanopublications. We show that novel associations can be discovered by performing queries that span genomic annotations from the FANTOM5 consortium (catalogue of transcription start site(s) (TSS)) and an instance of the Leiden Open Variation Database containing population-based variant frequency data. A correlation was established between the tissue specificity of a TSS and the frequency and size of genetic variants in the genomic region covering the TSS: TSS that are used in many tissues are less tolerant to disruption by insertions and deletions than those specific to just a few tissues.

Keywords: Linked Data, Nanopublication, Transcription start sites, Genetic variants

1 Introduction

Next Generation Sequencing (NGS) technologies produce huge amounts of data on genetic variation between organisms and between individuals. NGS data are reported and annotated in different formats that reflect the provenance and use case of the data.

For example, RefSeq, Ensembl, and dbSNP each have their own annotation format. Feature annotations (such as genes, transcripts, transcription start sites) can be downloaded as BED files, while variants are usually exported to Variant Call Format (VCF) format and subsequently stored in dedicated databases such as Locus Specific Databases (LSDB). One way to approach the interoperability of genomic data is to homogenize the data before analysis. This often requires downloading data from multiple sources into a central location and then standardizing the data structure (*e.g.* by importing into a fixed schema relational database). However, there are several limitations to this approach. First, data in the centralized warehouse may not be kept up to date. The process of downloading and normalizing data is usually time consuming and error prone, especially if the data structure from the source changes over time. Secondly, once meaningful results are derived from such a central data warehouse, it may not be possible to trace back the original data used to generate the results. If the large and ever growing amounts of (NGS) data are to be transformed into actionable information for the biologist, they must be represented in a machine readable format that permits automatic interoperation and *in silico* reasoning even when the data reflect diverse and heterogeneous annotations.

To enhance interoperability of diverse and distributed datasets, we introduced a framework to expose genomic annotation data as linked data with nanopublication as the schema backbone. A nanopublication is the smallest unit of publishable information: an assertion with provenance information and metadata about the nanopublication. Nanopublications build upon Linked Data, which provides a suitable platform for the exchange and integration of data [1, 2]. Nanopublications can be cited, offering an incentive for authors to publish their data in interoperable format [3, 4]. Patrinos *et al.* proposed nanopublications to encourage attribution and dissemination of genetic variant data [3]. Subsequently, we demonstrated how to model and publish genomic annotations, *e.g.* transcription start site(s) (TSS) from the FANTOM5 project as interoperable nanopublications [5].

In this paper, we demonstrate the power of the improved linking of heterogeneous and distributed variant and variant annotation data. We show how nanopublications can lead to novel discoveries through querying of different linked data resources and how shared ontological models allows can be exploited. In addition, the RDF [8, 9] query language SPARQL [10] allows queries over hierarchical graphs (subsumption queries) such as the FANTOM5 sample ontology. We demonstrate how federated queries across these three linked data sets (of which one is hierarchical) can lead to novel discoveries.

For this, we converted variant data from the Leiden Open Variant Database (LOVD), which is a platform for storing high quality, curated variant data [6] adopted by more than 95% of the gene variant databases, into a linked data resource. To better understand the effect of a variant, annotation about the variant region is required. LOVD currently uses reference transcript annotations to determine if a variant is potentially affecting the sequence of transcripts. However, the FANTOM5 greatly extended the catalogue of transcripts and (tissue-specific) TSS. Where previous studies mainly focused at variants affecting the coding exons, we were interested in variants affecting TSS. In particular larger insertions and deletions overlapping TSS

may affect transcription levels or the sequence of the 5'-end of transcripts. To understand the relationship between the tissue specificity (FANTOM5, sample ontology) of a TSS and the observed disruption of its genomic region by variants (LOVD), we converted population-based genetic variant data stored in LOVD to nanopublications. We test the hypothesis that the genomic regions of tissue specific TSS are less conserved than those of ubiquitous TSS and therefore contain more and larger genetic variants. The number of tissues in which a TSS controls gene expression indicates the number of sites that may be affected by TSS disruption. Disruption of TSS that exert regulatory control at many sites have a higher probability of causing a detrimental effect than TSS that control expression in only a few tissues.

2 Materials and Methods

2.1 Modelling

LOVD. To publish LOVD variant data as interoperable nanopublications, attributes of the variant entity were mapped to a semantic representation. Attributes describing the variant type (insertion, deletion, and substitution), Human Genome Variation Society (HGVS) variant description, and genomic position were selected for conversion, as well as allele frequencies in European American and African American population.

A variant was modeled as a genomic annotation that maps to a region on a reference sequence, and was expressed using the reference sequence annotation (RSA)[7] ontology. Apart from covering most of the terms and the predicates to represent the genomic annotation, the ontology also gives us flexibility to represent an annotation across different reference sequence builds. The Sequence Ontology (SO)[8] class was used for representing the variant and the Genome Assemblies [7] to reference an instance of a reference sequence, which maps to a version of the genome assembly. A schematic overview of the variant annotation model is shown in Figure 1. Example LOVD nanopublications can be viewed using the nanopublication viewer that is available at <http://rdf.biosemantics.org>.

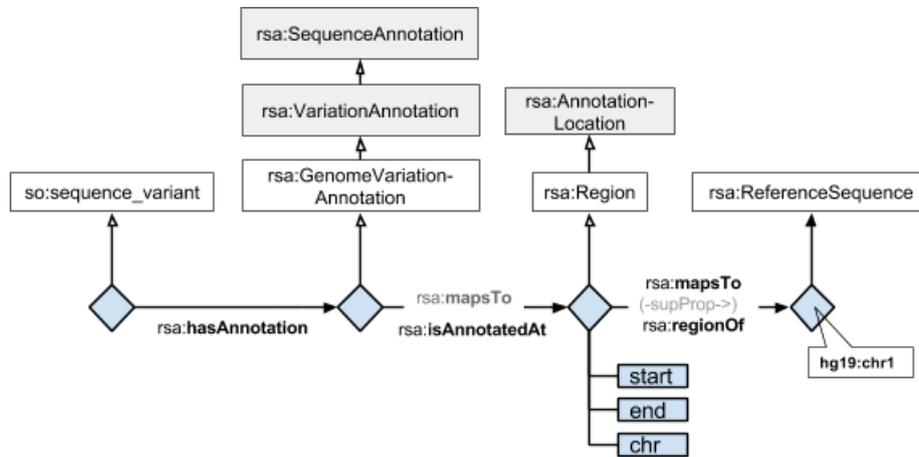


Fig. 1. The variant annotation model. The model states that a variant is observed at a genomic region, which has start and end position, and maps to a genomic component (chromosome) of a reference sequence. Instances of these classes (represented by a diamond shape) are the RDF resources from which a nanopublication assertion is constructed. Sequence ontology (SO), reference sequence annotation (RSA) ontology and human genome 19 reference sequence (HG19) RDF resource were used for this assertion model.

FANTOM5. The modelling of FANTOM5 data as nanopublications is described in supplemental material of [5]. In short, the core FANTOM5 data was exposed as three sets of nanopublications. The first set reported the genomic region of Cap Analysis of Gene Expression (CAGE) clusters in terms of start/end position and orientation on a reference sequence. The second set expressed that a CAGE cluster was also a TSS region, and connected them, when possible, to a specific gene. A CAGE cluster is observed in tissue with an abundance measured in TPM (tags per million). A schematic overview of all three assertion models and their connection is provided in Figure 2. Note that the genomic region is the same model as the variant annotation assertions (see Figure 1). Property `rsa:is_observation_of` was added as an update to the RSA ontology.

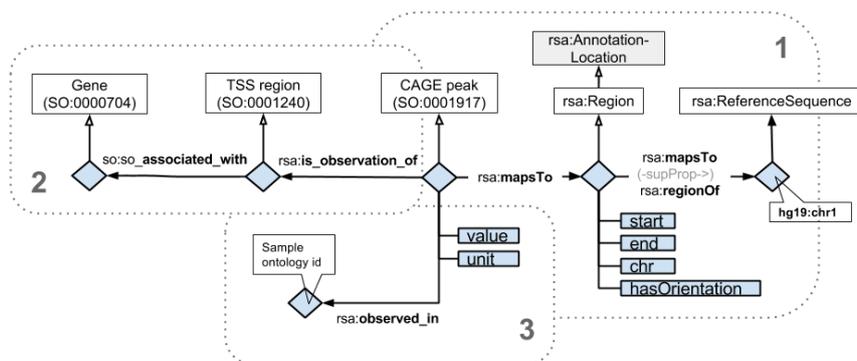


Fig. 2. Combined assertion models of the FANTOM5 nanopublications. Type 1 – cage cluster; type 2 – transcriptional start sites and gene associations; type 3 – cage peak level for muscle cell samples. Instances of these classes (represented by a diamond shape) are the RDF resources from which a nanopublication assertion is constructed. Sequence ontology (SO), reference sequence annotation (RSA) ontology and human genome 19 reference sequence (HG19) RDF resource were used for these assertions models.

Conversion. The LOVD 3.0 whole genome dataset was converted into RDF. This set consists of variant data copied from the Exome Variant Server (EVS) and is available at: http://databases.lovd.nl/whole_genome/genes. The EVS is a publicly available collection of genetic variant data from the NHLBI GO Exome Sequencing Project [9], which aims to identify new genes and mechanisms to help diagnosis, management and treatment of heart, lung and blood disorders by screening (near) the protein coding regions of genomes of phenotypically diverse European American and African American individuals. The EVS data consists of a large number (~2M in LOVD) of mostly non-disease associated or recessive disease variants, and can be considered a representative sample of the variations that occur in or near protein coding regions in genomes of the two populations.

Conversion performed using a custom script that fetched data directly from the underlying MySQL data dump, producing approximately 46 million triples. The database identifier of a variant consists of the HGNC gene symbol and number, and groups multiple observations of the variant together. The variant description at the DNA level was based on the genomic DNA reference sequence following HGVS recommendations. FANTOM5 nanopublications were generated as described previously [5 - Supplementary Note 1], using the phase 1.3 release and phase 1 prerelease version of the sample ontology. The scripts used for converting these datasets into nanopublications are publicly available and can be downloaded at <https://git.lumc.nl/biosemanitics/fantom5-lovd-data-integration>.

Sample classification. Samples are systematically classified using the FANTOM Five (FF) Sample Ontology, which aggregates other ontologies such as the cell

ontology [10] for cell types and Uberon [11] for anatomical systems. The sample ontology is a class hierarchy that was used to annotate experimental artifacts of FANTOM5, such as origin and anatomical location of tissue and cell samples. In this study, we focused on experimental results from adult human tissues, excluding experimental data from (carcinoma) cell-lines, developmental.

Tissue-specific sample retrieval. To retrieve the samples we used SPARQL to query the FANTOM5 sample ontology, making use of the hierarchical structure of the ontology to select samples that fall under classes that are both a subclass of "human adult sample" and "adult tissue sample" class. We defined expression as follows: a TSS is expressed in a certain tissue if it has TPM>0 in at least one sample belonging to that tissue type. The SPARQL query can be accessed via the link https://git.lumc.nl/biosemanitics/fantom5-lovd-data-integration/blob/master/rscript/sparql_queries/get_tissue_specificity_of_tss.sparql

Overlap calculation. A SPARQL query was used to find overlaps between variants and TSS regions. For insertion, deletion and duplication variants, the fraction of a variant or the whole variant should overlap with the TSS region. For the substitution variants the variant should occur within the TSS region. This also includes start and end positions of the TSS. The SPARQL query can be accessed via the link https://git.lumc.nl/biosemanitics/fantom5-lovd-data-integration/blob/master/rscript/sparql_queries/get_variants_overlapping_with_tss.sparql

Variant disruption calculation. Any type of variant may disrupt a TSS depending on the extent of overlap between them. Disruptions due to deletions, insertions and substitutions are calculated using the formulas given in equations 1-3. For deletions, the number of base pairs of the variant overlapping with the TSS, was divided by the TSS length (number of base pairs) to get a disruption value (equation 1). For insertions, the length of the variant is divided by TSS length (equation 2). Single nucleotide substitutions are less likely to disrupt a TSS. Still, we were interested to use this most common class of genetic variation to estimate the level of sequence conservation in the TSS, as a measure for evolutionary constraint. For this, the number of unique substitutions in the EVS database per TSS was divided by the TSS length (equation 3).

$$\text{Deletion disruption} = (\text{variant overlapping length}) / (\text{TSS length}) \quad (1)$$

$$\text{Insertion disruption} = (\text{variant length}) / (\text{TSS length}) \quad (2)$$

$$\text{Substitution rate} = (\text{number of unique substitutions in the TSS}) / (\text{TSS length}) \quad (3)$$

3 Results and discussion

In this work, we focused on the genomic annotations of variants and exposed this information as nanopublications. Population-wide genomic variant data from the EVS project were preprocessed and stored in an installation of LOVD 3.0 was converted to nanopublications consisting of the following type of assertions: a variant of a certain type (insertion, deletion, substitution) exists at a defined genomic position, *i.e.* has a specific start and end position on a specific chromosome. We used EVS data from LOVD and not the original source so that we could make use of preprocessed EVS data from LOVD. CAGE clusters and inferred TSS from FANTOM5 (184.827 CAGE clusters and 82.094 TSS) and EVS variants nanopublications were made available as a static RDF files dump (<http://beehub.nl/biosemanitics/fantom5-evs-integration-resources/>) through a public repository that was supported by the Dutch national e-infrastructure. To demonstrate the automatic interoperation between these two datasets, we made federated queries to find variants that overlap with TSS. Samples were limited to human adult tissues, resulting in 129 tissue samples, spanning 57 tissue types. Among these, the CAGE cluster expression (TPM value > 0) is observed in 111 tissue samples and these tissue samples span 44 tissue types (according to the FANTOM5 ontology). There are only 4,073 genes with a single TSS, compared to 14,866 with more than one TSS. The tissue specificity of TSS is displayed in Figure 3. The majority of TSS are used in multiple tissues, and it is difficult to draw a line between tissue-specific and ubiquitous TSS. Even more so, because there are many tissue samples with TPM value between 0 and 1 and any cut-off on expression levels is arbitrary. For this paper, we chose to divide TSS in two equal bins, where the 50% TSS observed in the lowest number of tissues are the tissue-specific TSS and the 50% TSS observed in the highest number of tissues are the ubiquitous TSS.

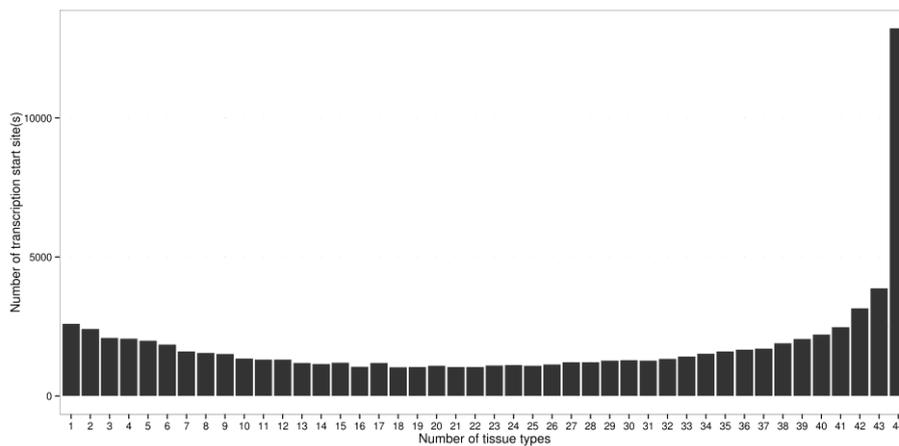


Fig. 3. Tissue specificity of transcription start sites. The x-axis shows the number of tissue types in adult human tissue in which a TSS is expressed. The y-axis represents the number of TSS with TPM > 0 in at least one sample of a tissue type. The maximum number of tissue types is 44. The peak at 44 is a bit artefactual because it is more likely that a TSS is present in all tissues studied than in exactly 43 tissues.

The set of 1,998,175 unique genetic variants in the LOVD whole genome instance of EVS consists of 33,646 insertions, 84,761 deletions, 6,875 duplications, and 1,872,893 substitutions. The length of deletion and insertion variants range from 1 to 102 and 1 to 61. From the set of 1,998,175 variants (EVS), 18,988 showed overlap with 9,949 TSS (Among them 9896 TSS had expression in adult human tissues). Of those variants there were 479 insertions, 1018 deletions, 72 duplications, and 17,419 substitutions. The maximum number of variants per TSS was 30, this TSS was associated with the RNU11 gene (EntrezGene ID 26824).

Association of the fraction of TSS disrupted by insertions and deletions with tissue specificity. Larger insertions and deletion are the classes of variants that are likely most disruptive for TSS. We assumed that the size of the insertion or deletion was an important determinant for its disruptiveness and therefore evaluated the fraction of the total TSS region length overlapping with the insertion or deletion. When a TSS overlapped with multiple insertions and deletions, we only included the variant with the largest overlap. The TSS were ranked according to their tissue specificity and to the fraction of its length overlapping with insertions or deletions. The ordered TSS are divided at the 50th percentile, both for tissue specificity and for disruption. The resulting distributions and p-values according to the chi-square test are displayed in Figure 4. The value in each quarter of the figure represents the percentage of TSS distribution in each quarter. For both deletions (Figure 4A) and insertions (Figure 4B), we observed that tissue-specific TSS tolerate a significantly higher level of disruption (larger fraction deleted or larger insertions) than ubiquitous TSS.

For single nucleotide substitutions, we concentrated on the number of unique positions with variants in EVS, normalized by the length of the TSS regions (Figure 4C). Here, we observed more substitutions in tissue-specific TSS than in ubiquitous TSS, indicative for a higher selective pressure on ubiquitous TSS indels.

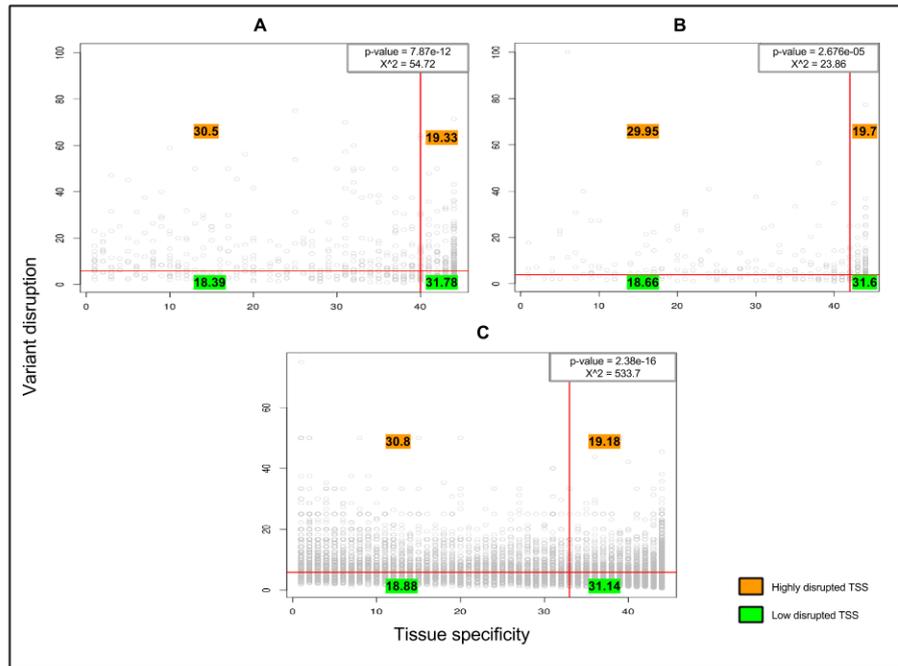


Figure 4. Tissue-specific TSS are more tolerant towards disruption by genetic variants than ubiquitous TSS. TSS are ranked according to the number of tissues in which the TSS is expressed (x-axis). The fraction of the length of the TSS region affected by genetic variants is plotted on the y-axis. The sub plots A, B and C represent the TSS disruptions caused by deletions, insertions and substitutions, respectively. The red horizontal and vertical lines in the graphs represent 50% of the ranked samples along each axis. The values in each quarter of the plot is the percentage of TSS distribution in that quarter. Chi-square statistics and their associated p-values are also indicated in each panel.

In summary, variants in coding parts genes are widely studied. However, variants in regulatory domains may also affect gene function. By integrating FANTOM5 data on TSS and LOVD data with population-based variant frequencies, we were able to study variants in TSS. From an evolutionary point of view, important regulatory elements are more conserved. This translates to finding less disruption by variants when a TSS is more important. A TSS is considered more important when it is active in many tissues. Variants in these TSS disrupt a gene's function in many tissues, thus probably having a larger impact on the organism than variants in tissue-specific TSS. Similar studies with large whole genome sequencing studies should corroborate these findings as our current study is limited by the use of exome data that are not covering the full repertoire of TSS annotated in the FANTOM5 project.

Future work. In the future we would like to integrate the FANTOM5 and EVS nanopublications with other genomic RDF resources. The clear challenge here would be integrating our nanopublications with other RDF resources where different ontologies were used to describe the genomic annotations. For example, we are aware that some major data providers such as the Database Center for Life Science in Japan are using the FALDO ontology instead of the RSA ontology to expose genomic annotation. The major challenge with mapping these ontologies would be to overcome the ontological groundings.

Applying our approach to VCF files might be another topic for future research. One of the challenges here lies in the enormous number of triples that would be generated based on the VCF format.

4 Conclusions

Genomic annotations are crucial for interpreting NGS data. Integrating genomic annotations from different sources is a time consuming and non-trivial task. In this study, we apply a framework for genomic annotations based on nanopublications. We demonstrated automatic interoperability between TSS from the FANTOM5 consortium and population-based variant frequency data from LOVD. We showed that novel associations can be discovered by performing queries that span the integrated domains. Specifically, a correlation was established between the tissue specificity of a TSS and the frequency and size of genetic variants in the genomic region covering the TSS: TSS that are expressed in many tissues are less tolerant to disruption by insertion and deletions than those expressed in only a few tissues. We demonstrated that use of an ontology enabled subsumption queries on FANTOM5 dataset, and we also demonstrated reusability of an existing data source by using the Human genome 19 reference sequence RDF resource in the EVS nanopublication model.

5 Acknowledgements

We gratefully acknowledge the statistical advice from Erik van Zwet. Funding from ODEX4all (NWO 560 650.002.002) and the European Commission (FP-7 project RD-Connect, grant agreement No. 305444) are gratefully acknowledged. This work was carried out on the Dutch national e-infrastructure with the support of SURF Foundation.

References

1. Berners-Lee, T., Bizer, C., Heath, T.: Linked data-the story so far. *Int. J. Semantic Web Inf. Syst.* 5, 1–22 (2009).
2. Heath, T., Bizer, C.: *Linked Data: Evolving the Web into a Global Data Space*. Synth. Lect. *Semantic Web Theory Technol.* 1, 1–136 (2011).

3. Patrinos, G.P., Cooper, D.N., van Mulligen, E., Gkantouna, V., Tzimas, G., Tatum, Z., Schultes, E., Roos, M., Mons, B.: Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain. *Hum. Mutat.* 33, 1503–1512 (2012).
4. Gibson, A., Van Dam, J.C., Schultes, E.A., Roos, M., Mons, B.: Towards Computational Evaluation of Evidence for Scientific Assertions with Nanopublications and Cardinal Assertions. In: *Proceedings of the 5th International Workshop on Semantic Web Applications and Tools for Life Sciences (SWAT4LS) Paris, France*. pp. 28–30 (2012).
5. The FANTOM Consortium and the RIKEN PMI and CLST (dgt): A promoter-level mammalian expression atlas. *Nature*. 507, 462–470 (2014).
6. Fokkema, I.F.A.C., Taschner, P.E.M., Schaafsma, G.C.P., Celli, J., Laros, J.F.J., Dunnen, J.T. den: LOVD v.2.0: the next generation in gene variant databases. *Hum. Mutat.* 32, 557–563 (2011).
7. Tatum, Z., Roos, M., Gibson, A.P., Taschner, P.E., Thompson, M., Schultes, E.A., Laros, J.F.: Preserving sequence annotations across reference sequences. *J. Biomed. Semant.* 5, S6 (2014).
8. Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R., Ashburner, M.: The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* 6, R44 (2005).
9. NHLBI GO Exome Sequencing Project: NHLBI GO Exome Sequencing Project (ESP) Exome Variant Server, <http://evs.gs.washington.edu/EVS/>.
10. Bard, J., Rhee, S.Y., Ashburner, M.: An ontology for cell types. *Genome Biol.* 6, R21 (2005).
11. Mungall, C.J., Torniai, C., Gkoutos, G.V., Lewis, S.E., Haendel, M.A.: Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* 13, R5 (2012).