Genome Annotation using Nanopublications: An Approach to Interoperability of Genetic Data

Rajaram Kaliyaperumal[1], Peter A.C. 't Hoen[1], Zuotian Tatum[1], Mark Thompson[1], Eelke van der Horst[1], Erik Schultes[1], Ivo F.A.C. Fokkema[1], Johan T. den Dunnen[1], Jeroen F.J. Laros[1], José Luis Oliveira[2], Pedro Lopes[2], Pedro Sernadela[2], Marco Roos[1]

[1] Department of Human and Clinical Genetics, Leiden University Medical Centre, Leiden, The Netherlands {R.Kaliyaperumal,P.A.C._t_Hoen,Z.Tatum,M.Thompson,E.van_der_Horst,E.Schultes,I.F.A.C.Fokkema,J.F.J.Laros,M.Roos}@lumc.nl ddunnen@humgen.nl

[2] DETI/IEETA, University of Aveiro, Aveiro, Portugal {jlo,pedrolopes,sernadela}@ua.pt

Abstract

With the widespread use of Next Generation Sequencing (NGS) technologies, the primary bottleneck of genetic research has shifted from data production to data analysis. However, annotated datasets produced by different research groups are often in different formats, making genomic comparisons and integration with other datasets challenging and time consuming tasks. Here, we propose a new data interoperability approach that provides unambiguous (machine readable) description of genomic annotations based on a novel method of data publishing called nanopublication. A nanopublication is a schema built on top of existing semantic web technologies that consists of three components: an individual assertion (i.e., the genomic annotation); provenance (containing links to the experimental information and data processing steps); and publication info (information about data ownership and rights, allowing each genomic annotation to be citable and its scientific impact tracked) [1]. We use nanopublications to demonstrate automatic interoperability between individual genomic annotations from the functional annotation of the mammalian genome 5 (FANTOM5) consortium (transcription start sites) and the Leiden Open Variation Database (genomic variants). The nanopublications can also be integrated with the data of the other semantic web frameworks like COEUS. Exposing legacy information and new NGS data as nanopublications promises tremendous scaling advantages when integrating very large and heterogeneous genomic datasets.

References

[1] Paul Groth, Andrew Gibson, and Jan Velterop. 2010. The anatomy of a nanopublication. Inf. Serv. Use 30, 1-2 (January 2010), 51-56

Genome Annotation using Nanopublications: An Approach to Interoperability of Genetic Data

Rajaram Kaliyaperumal[1], Peter A.C. 't Hoen[1], Zuotian Tatum[1], Mark Thompson[1], Eelke van der Horst[1], Erik Schultes[1], Ivo F.A.C. Fokkema[1], Johan T. den Dunnen[1], Jeroen F.J. Laros[1], José Luis Oliveira[2], Pedro Lopes[2], Pedro Sernadela[2], Marco Roos[1]

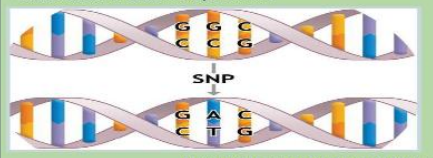
[1] Department of Human and Clinical Genetics, Leiden University Medical Centre, Leiden, The Netherlands, [2] DETI/IEETA, University of Aveiro, Aveiro, Portugal

Abstract

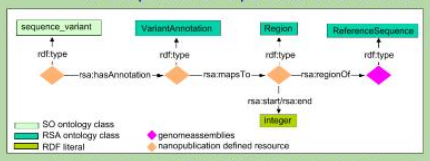
With the widespread use of Next Generation Sequencing (NGS) technologies, the primary bottleneck of genetic research has shifted from data production to data analysis. However, annotated datasets produced by different research groups are often in different formats, making genomic comparisons and integration with other datasets challenging and time consuming tasks. Here, we propose a new data interoperability approach that provides unambiguous (machine readable) description of genomic annotations based on a novel method of data publishing called nanopublication. A nanopublication is a schema built on top of existing semantic web technologies that consists of three components: an individual assertion (i.e., the genomic annotation); provenance (containing links to the experimental information and data processing steps); and publication info (information about data ownership and rights, allowing each genomic annotation to be citable and its scientific impact tracked) [1]. We use nanopublications to demonstrate automatic interoperability between individual genomic annotations from the functional annotation of the mammalian genome 5(FANTOM5) consortium (transcription start sites) and the Leiden Open Variation Database (genomic variants). The nanopublications can also be integrated with the data of the other semantic web frameworks like COEUS. Exposing legacy information and new NGS data as nanopublications promises tremendous scaling advantages when integrating very large and heterogeneous genomic datasets.

LOVD Dataset

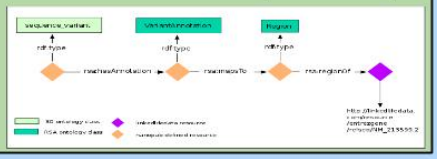
A SNP variant on a DNA which is result of substitution of Guanine by Adenine.



A variant is annotated to a genome with a start and an end position on a particular chromosome.



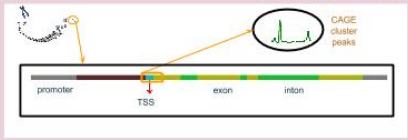
A variant is annotated with a particular transcript.



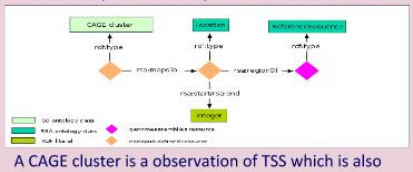
LOVD Nano-publication Assertions

FANTOM5 Dataset

CAGE cluster peaks on a transcription start site(TSS) of a gene.



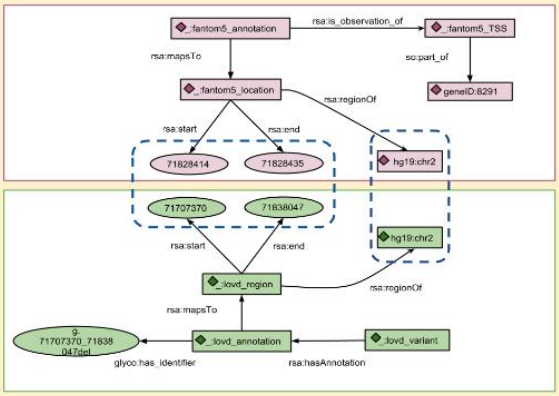
A CAGE cluster is annotated to a genome with a start and an end position on a particular chromosome.



A CAGE cluster is a observation of TSS which is also part of a gene.

FANTOM5 Nano-publication Assertions

Integration of FANTOM5 with LOVD



Integrating the LOVD dataset with the FANTOM5 dataset to find all the LOVD variants that are in the transcription start site(TSS) of a gene.

Result

Nucleotide change	Substitutions	Deletions	Duplications	Indels
Single	22	2	-	-
Multiple	-	6	4	1

Table 1 : Result of the integration of FANTOM5 dataset with the LOVD variant dataset of ten genes associated with Limb- Girdle Muscular Dystrophy (LGMD) disease.

Variant DB-ID	gChange	No.of TSS deleted
SGCB_00043	g. 52886872_52904485del	1
DYSF_00176	g.71707370_71838047del	9
LMNA_00209	g.156052975_156100408del	12

Table 2 : Some variants have deleted entire transcription start sites.

Conclusions

Genomic annotations are crucial for interpreting NGS data. Integrating genomic annotations from different sources is not trivial and time consuming. We proposed the Framework for Genomic Annotations as an example of a data integration strategy. We demonstrated that by adopting such a framework, genomic annotations from different sources with different schema can be easily integrated and kept up to date. Moreover, the provenance of each annotation is preserved, making the tracking of the ownership of the data easy.

[1] Paul Groth, Andrew Gibson, and Jan Velterop. 2010. The anatomy of a nanopublication. Inf. Serv. Use 30, 1-2 (January 2010), 51-56