Metadata harmonization at scale:

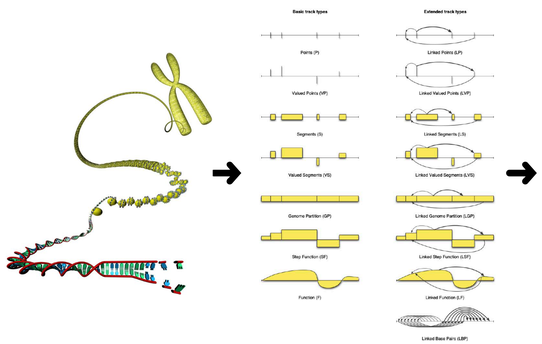
FAIRification of Genomic Annotations WG

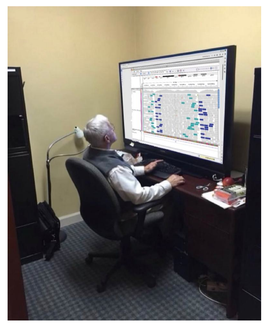
Sveinung Gundersen [1*], Anna Bernasconi [2*], Nathan Sheffield [3*], Adam Wright [4*]

[1] *ELIXIR Norway / University of Oslo (UiO), Norway*
[2] *Politecnico di Milano, Italy*
[3] *University of Virginia, Virginia, USA*
[4] *Ontario Institute for Cancer Research, Canada*
* *WG Co-chairs*

Immense resources have gone into the generation of data that relates genomic positions to functional and structural aspects, in large consortia such as ENCODE, ICGC, FANTOM and FAANG, as well as in smaller research projects. This is complemented by a current push to sequence millions of species in biodiversity projects – all of which needs characterisation of genes and other genomic features.

International Cancer Genome Consortium — 2008 →
PCAWG PanCancer Analysis of WHOLE GENOMES — 2014 → 2020
EARTH BIOGENOME PROJECT — 2018 →
ENCODE — 2003 → 2021
ROADMAP epigenomics — 2008 – 2018
FAANG Functional Annotation of Animal Genomes — 2015 →
Darwin TREE OF LIFE — 2019 →
BLUEPRINT epigenome — 2011 – 2017
FANTOM — 1995 →
THE CANCER GENOME ATLAS — 2005 – 2014
HUMAN CELL ATLAS — 2016 →

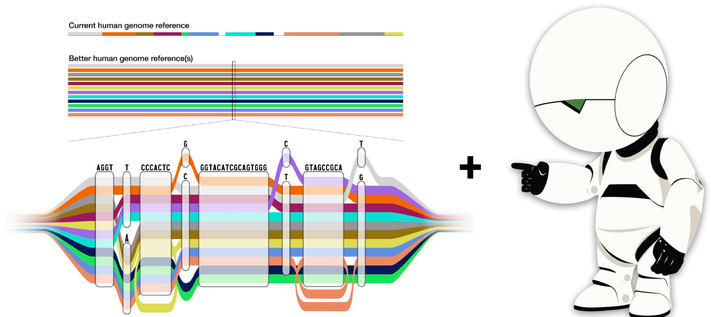




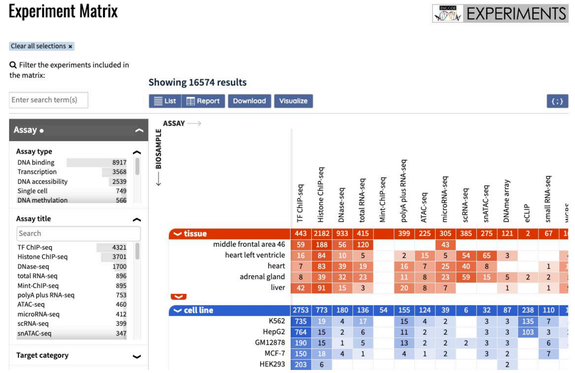
Genome browser meme by Jedidiah Carlson

Functional genomics data are typically provided as tracks for visualisation in genome browsers, or made down-loadable for non-visual analysis. The use of reference genomes as one-dimensional coordinate systems constitutes a powerful unified model for data analysis. Interestingly, the condensed nature of coordinate-based data is ideally suited for data-driven approaches!

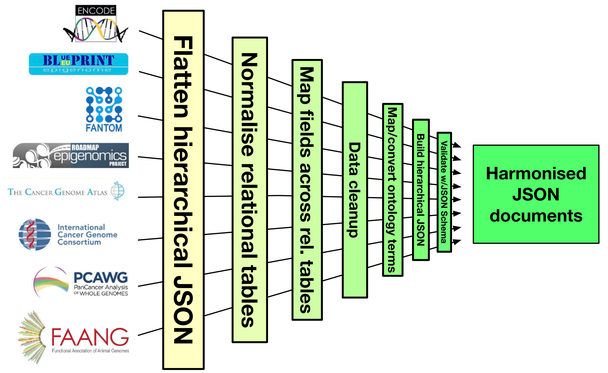
Pangenomes challenge the concept of one-dimensional reference genomes. While a consensus has not yet formed on how to annotate pangenomes, the value of reusing the abundance of existing functional genomics data in this new paradigm is clearly evident. Moreover, data-driven discovery by applying AI to functional annotations still holds consider-able untapped potential. To these ends, FAIR (Findable, Accessible, Interoperable, Reusable) metadata remains key.



Credit: Darryl Leja, NHGRI

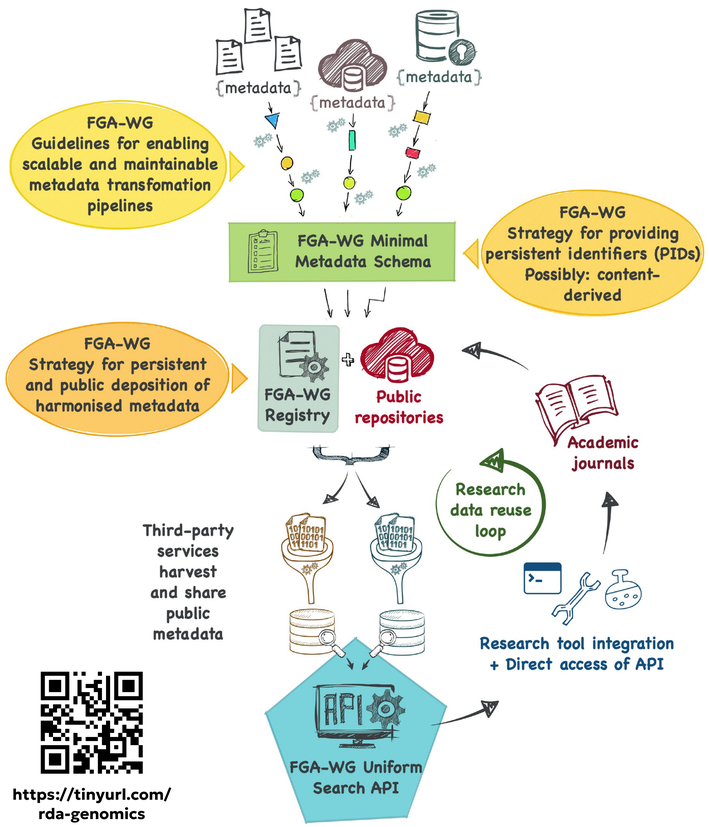


Functional genomics data are often provided through dedicated data portals, such as the ENCODE data portal (left). However, the metadata are provided according to distinct models and interfaces. Additionally, data from smaller research projects are often available only with limited metadata.



Harmonisation and FAIRification of metadata from different sources is a time-consuming and error-prone process, often implemented in *ad hoc* scripts that are difficult to maintain and scale. While each source exhibits particularities in schema, API and content, many operations are common. It should therefore be possible to streamline the process of developing maintainable metadata transformation pipelines.

Planned FAIRification infrastructure & WG deliverables



https://tinyurl.com/rda-genomics

Join our working group in RDA:

FAIRification of Genomic Annotations WG (FGA-WG)

We aim to produce a set of recommendations that together will define a community-oriented infrastructure to make it easier to discover and reuse genomic annotations in a range of contexts.

WG meetings currently every 1st and 3rd Tuesday per month, at 2pm UTC.

New Global Support Services for RDA Working Groups

Relevant roles:

- data producer
- tool developer
- biocurator
- domain expert
- analytical end user

We do not want to reinvent the wheel.

If you have know-ledge to share, please come see us!

elixir
RDA RESEARCH DATA ALLIANCE
Global Alliance for Genomics & Health
RDA TIGER
eosc