The Genomic HyperBrowser

– extensive framework for integrative analysis across the genome and epigenome

Elixir All Hands, 4–7 June 2018, Berlin, Germany

Geir K Sandve, Sveinung Gundersen, Boris Simovski, Morten Johansen, Arnoldo Frigessi, and Eivind Hovig

Published contributions by: O Abul, JA Akse, A Azab, M Bengtsen, Ø Borgan, T Clancy, D Domanska, F Drabløs, R Eskeland, E Ferkingstad, B Fromm, OS Gabrielsen, IK Glad, I Grytten, K Gunathasan, L Holden, M Holden, M Kalaš, TG Lien, K Liestøl, C Lund-Andersen, S Nakken, AJ Nederbragt, V Nygaard, S Nygård, A Mora, J Paulsen, KD Rand, H Rydbeck, HS Thorarensen, K Trengereid, E Tøstesen, and D Vodák

Complete analysis framework for genomic tracks - The Genomic HyperBrowser[1,2] (https://hyperbrowser.uio.no) is a web-based framework designed to support any kind of analysis for any kind of data that can be positioned relative to a reference genome assembly (i.e., for any genomic track).

Background

Through international projects like ENCODE, Roadmap Epigenomics, and International Cancer Genomics Consortium (ICGC), in addition to manually curated repositories like the GWAS Catalog, UCSC Genome Browser and Ensembl, a wealth of track data related to, e.g., gene regulation, epigenomic modifications, 3D chromatin conformation, or individual variation, is all available at the fingertips of life science researchers. However, combining such track data into meaningful analyses is difficult, both at the practical level (i.e., handling file formats and tool parameters) and at the analytical level (i.e., selecting an appropriate statistics and making sure that the assumptions of the methodology fits with the data).

The Genomic HyperBrowser

The Genomic HyperBrowser project has, since its initiation in 2008, aimed to provide a complete solution for all parts of the typical analysis workflow, starting after data processing steps like ChIP-seq peak calling or variant calling has taken place. The first iteration of the system focused on the analysis of a pair of tracks. Even with this limitation, the methodological complexities have proven to be immense. Several extensions[3-8] have, over the years, been published based upon the initial two-track design, all developed in conjunction with experts in statistics.

GSuite HyperBrowser

The latest expansion, GSuite HyperBrowser[9], constitutes a complete redesign of the analysis workflow with a focus on handling large collections of tracks, not just pairs, all the way from search and retrieval from data repositories, through diverse manipulation steps if needed, and throughout the statistical analyses themselves. A range of new collection-aware analyses have been developed, along with a sizeable list of pre-defined biological and biomedical analysis scenarios, with easy-to-follow tutorials.

coloc-stats

A recent development is the coloc-stats web page[10] (https://hyperbrowser.uio.no/coloc-stats), which collects the HyperBrowser and 6 other methods for colocalization analysis in the same web interface. The methods are based on different statistical assumptions and might thus give different results. It is thus an advantage for researchers to be able to easily compare the different results.
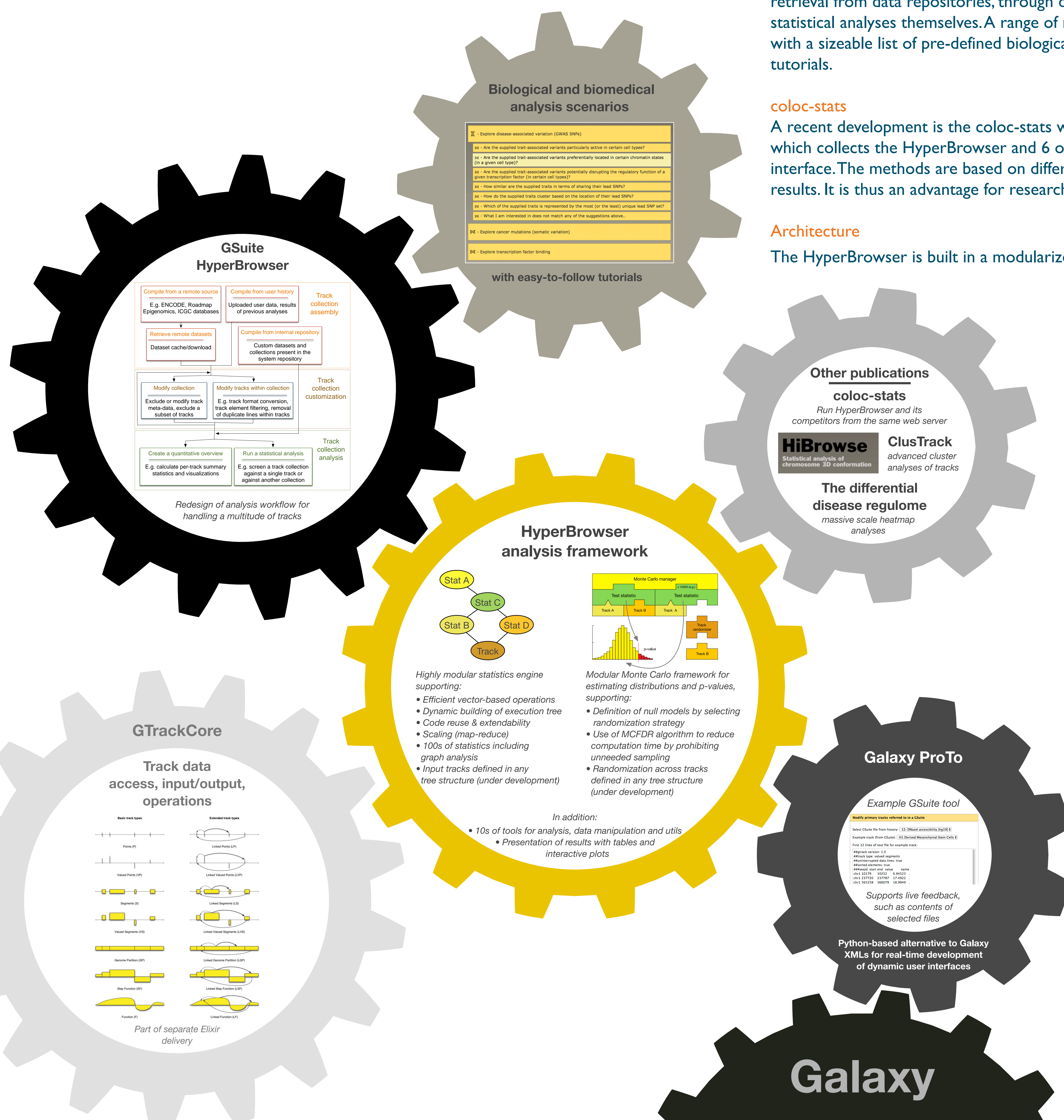
Architecture

The HyperBrowser is built in a modularized fashion, making use of the Galaxy architecture for the basic user interface. Instead of defining static tools via the XML-based solution in Galaxy, the system uses Galaxy ProTo, an extension to Galaxy that provides an alternative way to create tools. ProTo allows for dynamic feedback to the user, with the interface being defined by real-time execution of Python code. The analysis framework consists of a highly extendable statistics engine, and a modular Monte Carlo simulation solution, with carefully defined null model specification by various randomization schemes. Track data access and manipulation is provided by the GTrackCore library. A range of statistics and tools are included, and the system is easily extended if needed in specific research scenarios. The source code is available on GitHub (https://github.com/hyperbrowser/genomic-hyperbrowser).

The Genomic HyperBrowser is one of the official services developed by ELIXIR Norway for the European research community.

[1] Sandve, GK, et al. "The Genomic HyperBrowser: inferential genomics at the sequence level." Genome biology 11.12 (2010): 1-12.
[2] Sandve, GK, et al. "The Genomic HyperBrowser: an analysis web server for genome-scale data." Nucleic acids research 41.W1 (2013): W133-W141.
[3] Sandve, GK., et al. "The differential disease regulome." BMC genomics 12.1 (2011): 353.
[4] Sandve, GK, Ferkingstad E, and Nygård S. "Sequential Monte Carlo multiple testing." Bioinformatics 27.23 (2011): 3235-3241.
[5] Gundersen, S, et al. "Identifying elemental genomic track types and representing them uniformly." BMC Bioinformatics 12.1 (2011): 1.
[6] Paulsen, J, et al. "Handling realistic assumptions in hypothesis testing of 3D co- localization of genomic elements." Nucleic acids research (2013): gkt227.
[7] Paulsen, J, et al. "HiBrowse: multi-purpose statistical analysis of genome-wide chromatin 3D organization." Bioinformatics (2014): btu082.
[8] Rydbeck, H, et al. "ClusTrack: feature extraction and similarity measures for clustering of genome-wide data sets." PloS one 10.4 (2015): e0123261.
[9] Simovski, Boris, et al. "GSuite HyperBrowser: integrative analysis of dataset collections across the genome and epigenome." GigaScience 6.7 (2017): 1-12.
[10] Simovski, Boris, et al. "Coloc-stats: a unified web interface to perform colocalization analysis of genomic features." Nucleic Acids Research 46.W1 (2018) [in press]

GSuite HyperBrowser

Redesign of analysis workflow for handling a multitude of tracks

Biological and biomedical analysis scenarios

with easy-to-follow tutorials

HyperBrowser analysis framework

Highly modular statistics engine supporting:
- Efficient vector-based operations
- Dynamic building of execution tree
- Code reuse & extendability
- Scaling (map-reduce)
- 100s of statistics including graph analysis
- Input tracks defined in any tree structure (under development)

Modular Monte Carlo framework for estimating distributions and p-values, supporting:
- Definition of null models by selecting randomization strategy
- Use of MCFDR algorithm to reduce computation time by prohibiting unneeded sampling
- Randomization across tracks defined in any tree structure (under development)

In addition:
- 10s of tools for analysis, data manipulation and utils
- Presentation of results with tables and interactive plots

GTrackCore

Track data access, input/output, operations

Part of separate Elixir delivery

Other publications

coloc-stats
Run HyperBrowser and its competitors from the same web server

HiBrowse — Statistical analysis of chromosome 3D conformation

ClusTrack — advanced cluster analyses of tracks

The differential disease regulome
massive scale heatmap analyses

Galaxy ProTo

Example GSuite tool

Supports live feedback, such as contents of selected files

Python-based alternative to Galaxy XMLs for real-time development of dynamic user interfaces

Galaxy

Contact:

Sveinung Gundersen
sveinung.gundersen@medisin.uio.no
Elixir Norway, Oslo Node, Univ. of Oslo,
Pb 1080 Blindern, 0316 Oslo, Norway
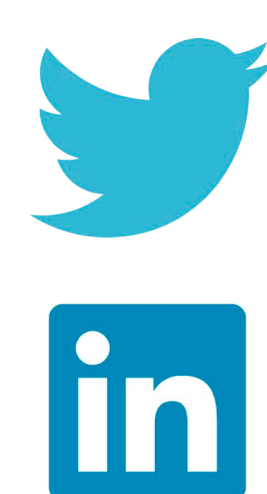
@ELIXIREurope

/company/elixir-europe

UiO : University of Oslo

elixir NORWAY