Coloc-stats

– a unified web interface to perform colocalization analysis of genomic features

Boris Simovski, Chakravarthi Kanduri*, Sveinung Gundersen*, Dmytro Titov, Diana Domanska, Christoph Bock, Lara Bossini-Castillo, Maria Chikina, Alexander Favorov, Ryan M Layer, Andrey A Mironov, Aaron R Quinlan, Nathan C Sheffield, Gosia Trynka, and Geir K Sandve*

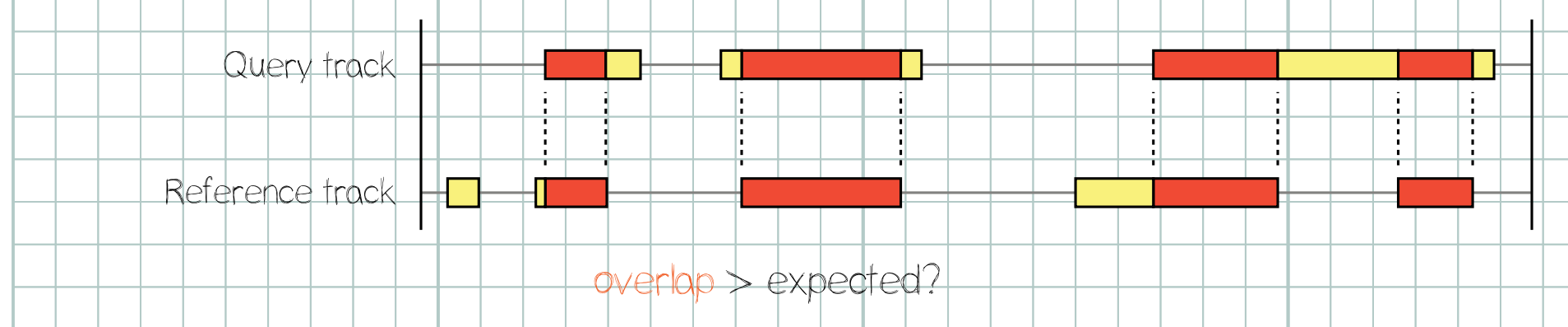
**These authors contributed equally*

2018 Galaxy Community Conference (GCCBOSC2018), June 27–28, Portland, Oregon, USA

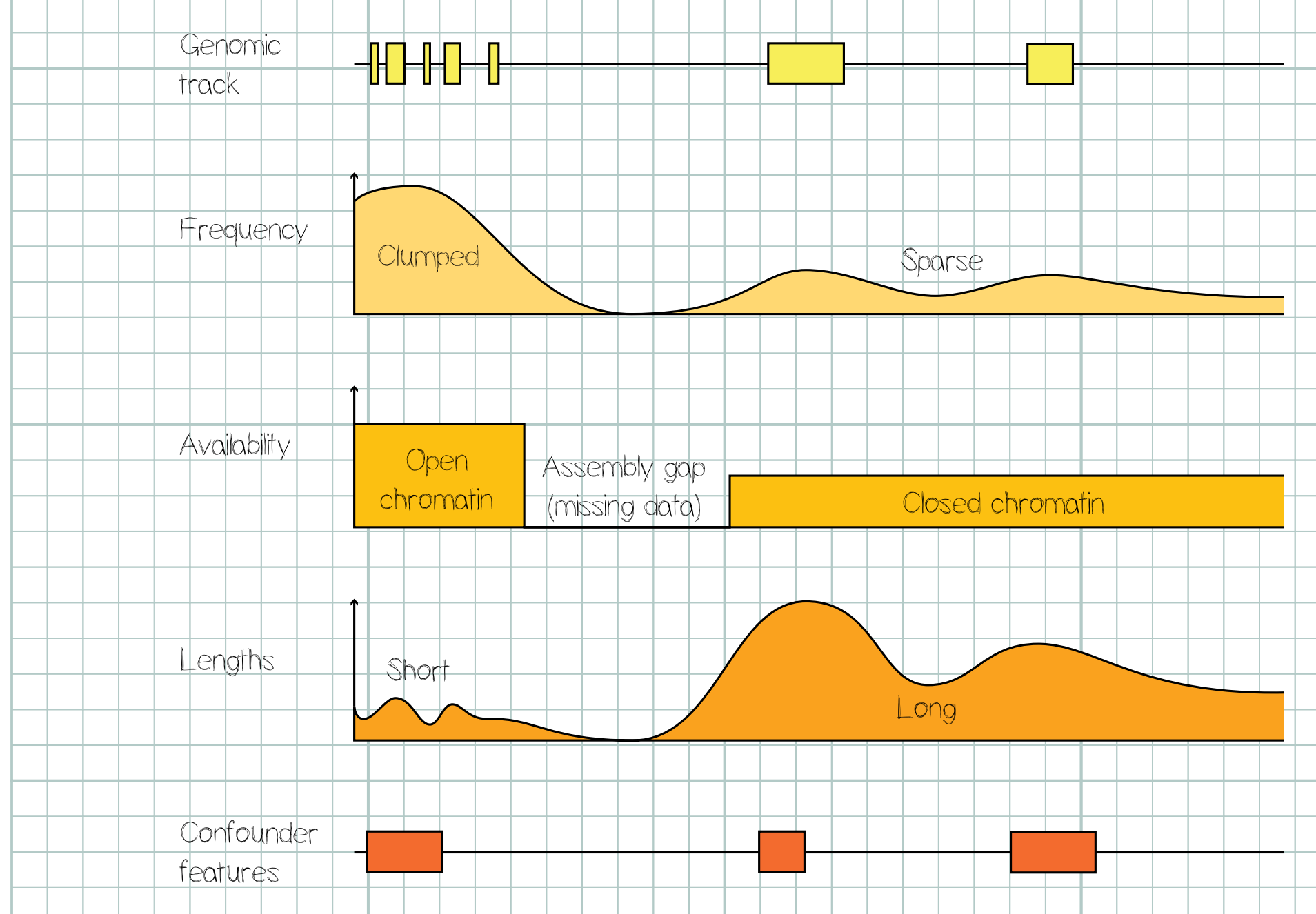
The problem

TFs, SNPs, transcripts, histone modifications, enhancers, repeats, DNA methylation, DNAseI HS, ...

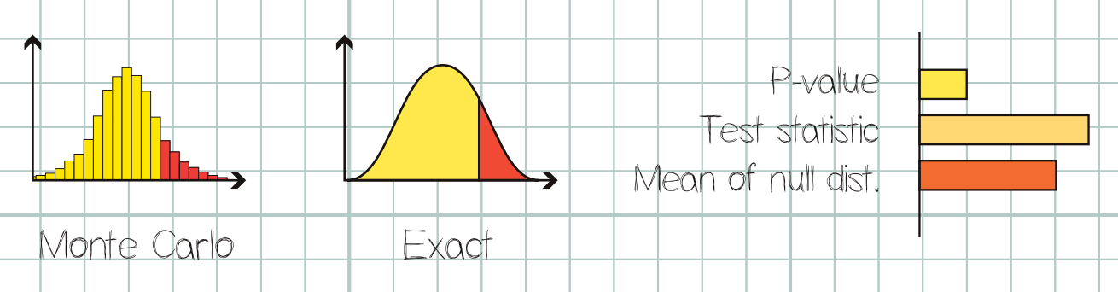
1. Functionally related GENOMIC FEATURES often tend to co-localize:



2. The genomic features do not occur independently, but are known to follow various properties:



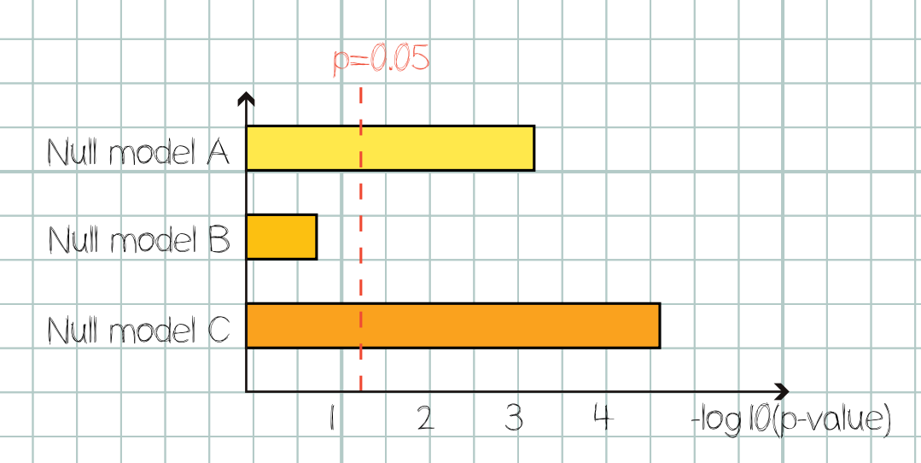
3. Testing for significant co-localization assumes a model of the biological randomness (the NULL MODEL):



4. Existing co-localization analysis tools assume quite different null models

and

5. The choice of null model greatly affects the conclusions[1]:



Thus, there is a high risk of FALSE POSITIVES!

SO, WHAT SHOULD ONE DO?

Run multiple tests with varying methods, parameters, and null models → Compare the results → If they differ, reason about the assumptions, and be conservative

'BUT HOW TO EASILY RUN ALL THE TOOLS?'

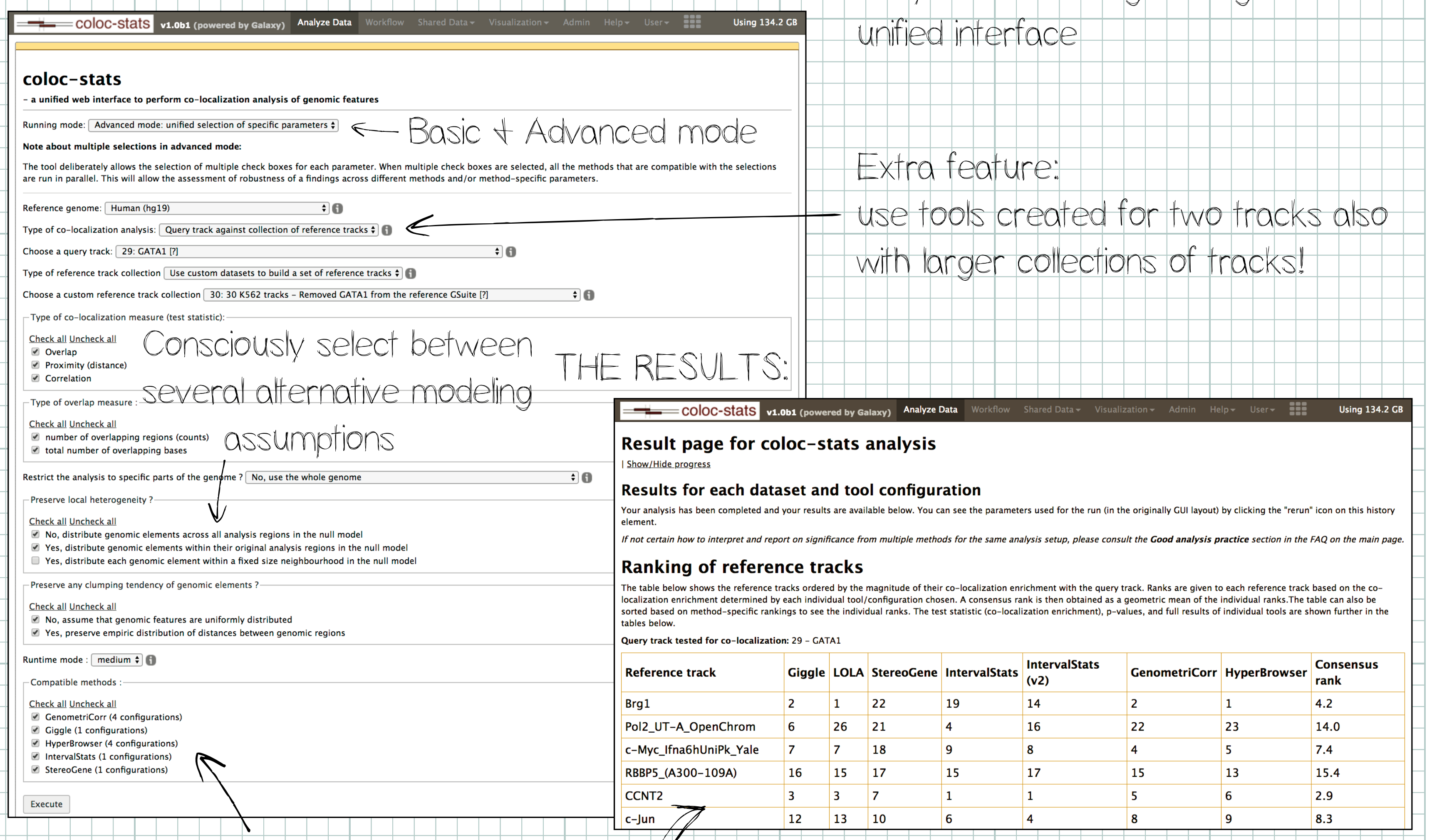
The goal

One tool to ~~rule~~ contain them all:

Coloc-stats

Explore and use multiple co-localization analysis tools through a single unified interface

THE GRAPHICAL USER INTERFACE (GUI):



coloc-stats

– a unified web interface to perform co-localization analysis of genomic features

Running mode: Advanced mode: unified selection of specific parameters ← Basic + Advanced mode

Note about multiple selections in advanced mode:

The tool deliberately allows the selection of multiple check boxes for each parameter. When multiple check boxes are selected, all the methods that are compatible with the selections are run in parallel. This will allow the assessment of robustness of a findings across different methods and/or method-specific parameters.

Reference genome: Human (hg19)

Type of co-localization analysis: Query track against collection of reference tracks

Choose a query track: 29: GATA1 [?]

Type of reference track collection: Use custom datasets to build a set of reference tracks

Choose a custom reference track collection: 30: 30 K562 tracks – Removed GATA1 from the reference GSuite [?]

Type of co-localization measure (test statistic):
Check all Uncheck all
☑ Overlap
☑ Proximity (distance)
☑ Correlation

Type of overlap measure :
Check all Uncheck all
☑ number of overlapping regions (counts)
☑ total number of overlapping bases

Restrict the analysis to specific parts of the genome ? No, use the whole genome

Preserve local heterogeneity ?
Check all Uncheck all
☑ No, distribute genomic elements across all analysis regions in the null model
☑ Yes, distribute genomic elements within their original analysis regions in the null model
☐ Yes, distribute each genomic element within a fixed size neighbourhood in the null model

Preserve any clumping tendency of genomic elements ?
Check all Uncheck all
☑ No, assume that genomic features are uniformly distributed
☑ Yes, preserve empiric distribution of distances between genomic regions

Runtime mode : medium

Compatible methods :
Check all Uncheck all
☑ GenometriCorr (4 configurations)
☑ Giggle (1 configurations)
☑ HyperBrowser (4 configurations)
☑ IntervalStats (1 configurations)
☑ StereoGene (1 configurations)

Execute

Consciously select between several alternative modeling assumptions

Arrive at a list of methods that are compatible with chosen modeling assumptions

Extra feature: use tools created for two tracks also with larger collections of tracks!

THE RESULTS:

Result page for coloc-stats analysis

| Show/Hide progress

Results for each dataset and tool configuration

Your analysis has been completed and your results are available below. You can see the parameters used for the run (in the originally GUI layout) by clicking the "rerun" icon on this history element.

*If not certain how to interpret and report on significance from multiple methods for the same analysis setup, please consult the **Good analysis practice** section in the FAQ on the main page.*

Ranking of reference tracks

The table below shows the reference tracks ordered by the magnitude of their co-localization enrichment with the query track. Ranks are given to each reference track based on the co-localization enrichment determined by each individual tool/configuration chosen. A consensus rank is then obtained as a geometric mean of the individual ranks.The table can also be sorted based on method-specific rankings to see the individual ranks. The test statistic (co-localization enrichment), p-values, and full results of individual tools are shown further in the tables below.

Query track tested for co-localization: 29 – GATA1

Reference track	Giggle	LOLA	StereoGene	IntervalStats	IntervalStats (v2)	GenometriCorr	HyperBrowser	Consensus rank
Brg1	2	1	22	19	14	2	1	4.2
Pol2_UT-A_OpenChrom	6	26	21	4	16	22	23	14.0
c-Myc_Ifna6hUniPk_Yale	7	7	18	9	8	4	5	7.4
RBBP5_(A300-109A)	16	15	17	15	17	15	13	15.4
CCNT2	3	3	7	1	1	5	6	2.9
c-Jun	12	13	10	6	4	8	9	8.3

Examine the robustness of conclusions by comparing results across several different methods/null models

+ Access full results of each method through detailed results pages

The challenges

1. THE TOOLS ARE AVAILABLE IN DIFFERENT FRAMEWORKS:

Linux command line	R libraries	Python / Galaxy
GIGGLE GENOMIC INQUIRY[2] GoShifter[3] IntervalStats[4] StereoGene[5]	GenometriCorr[6] LOLA[7]	GSuite HyperBrowser[8, 9]

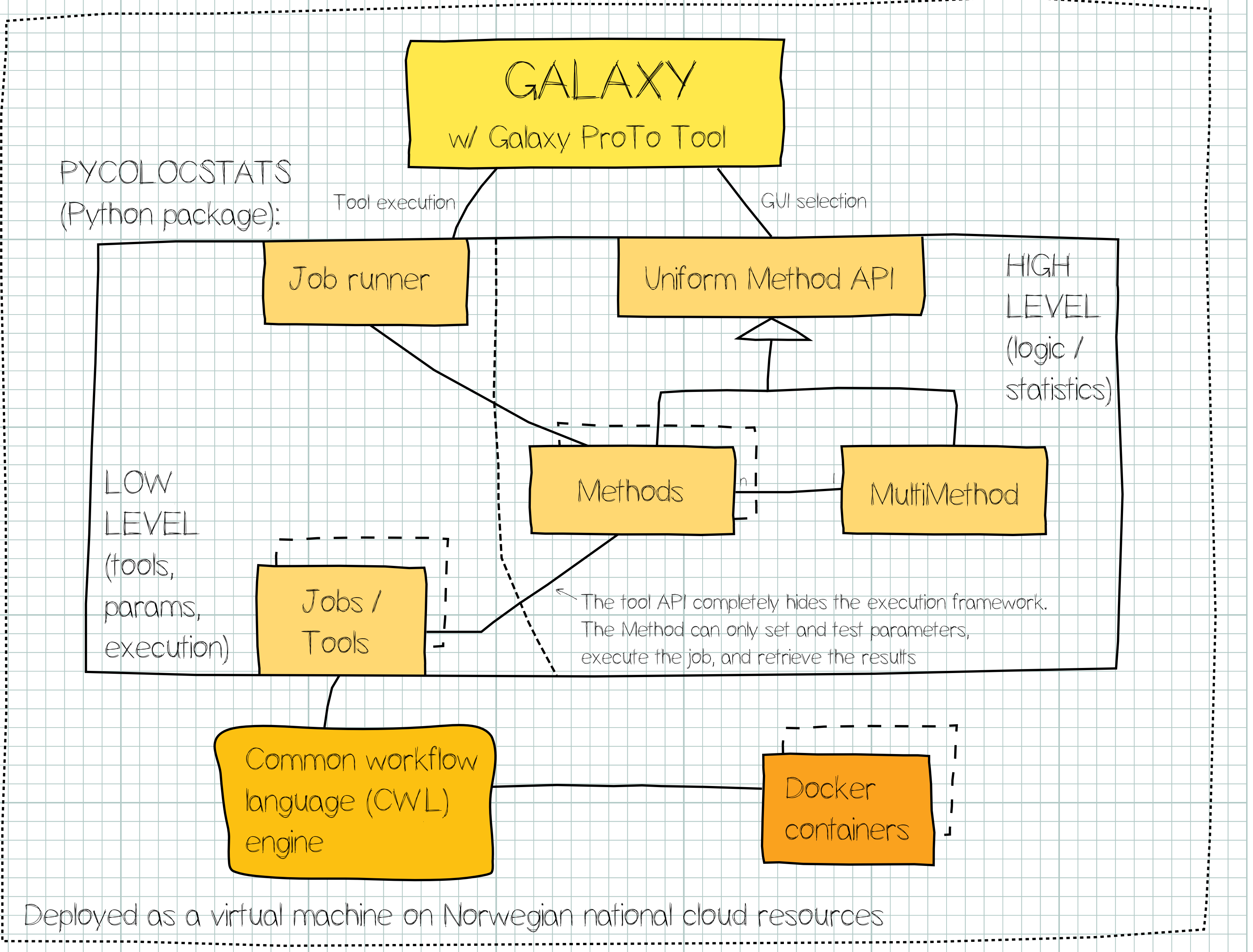
2. REQUIRES SINGLE USER INTERFACE WITH REPRODUCIBILITY AND ADVANCED LOGIC:
→ Galaxy framework: good
Galaxy tool XML: hmm...
Galaxy workflow engine: no good

3. TRYING TO COORDINATE SEVEN RESEARCH GROUPS...

4. TOO MANY CHEFS (DEVELOPERS) IN THE SAME CODE
→ Good modularization and interfaces needed to divide responsibilities between low-level (tools, parameters) and high-level (method logic) implementation

5. NUCLEIC ACIDS WEB SERVER ISSUE[10] DEADLINE:
= Too little time !!

The solution



The tool API completely hides the execution framework. The Method can only set and test parameters, execute the job, and retrieve the results

Deployed as a virtual machine on Norwegian national cloud resources

[1] Ferkingstad, E, Holden L, and Sandve, GK. "Monte carlo null models for genomic data." Statistical Science 30.1 (2015): 59-71.
[2] Layer, RM, et al. "GIGGLE: a search engine for large-scale integrated genome alanysis." Nature methods (2018).
[3] Trynka, G, et al. "Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci." The American Journal of Human Genetics 97.1 (2015): 139-152.
[4] Chikina, MD, and Troyanskaya, OG. "An effective statistical evaluation of ChIPseq dataset similarity." Bioinformatics 28.5 (2012): 607-613.
[5] Stavrovskaya, ED, et al. "StereoGene: rapid estimation of genome-wide correlation of continuous or interval feature data." Bioinformatics 33.20 (2017): 3158-3165.
[6] Favorov, A, et al. "Exploring massive, genome scale datasets with the GenometriCorr package." PLoS computational biology 8.5 (2012): e1002529.
[7] Sheffield, NC, and Bock, C. "LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor." Bioinformatics 32.4 (2015): 587-589.
[8] Sandve, GK, et al. "The Genomic HyperBrowser: inferential genomics at the sequence level." Genome biology 11.12 (2010): 1-12.
[9] Sandve, GK, et al. "The Genomic HyperBrowser: an analysis web server for genome-scale data." Nucleic acids research 41.W1 (2013):W133-W141.
[10] Simovski, Boris, et al. "Coloc-stats: a unified web interface to perform colocalization analysis of genomic features." Nucleic acids research 46.WI (2018).

Background created by Freepik

K.G. JEBSEN Center for Medical Research
ÖAW AUSTRIAN ACADEMY OF SCIENCES
NIH National Institutes of Health
wellcome trust
NIH National Human Genome Research Institute
The Research Council of Norway
elixir NORWAY
HORIZON 2020
RFBR RUSSIAN FOUNDATION FOR BASIC RESEARCH
MRC Medical Research Council
NATIONAL CANCER INSTITUTE
UiO University of Oslo