TECHNICAL NOTE

GSuite HyperBrowser: integrative analysis of dataset collections across the genome and epigenome

Boris Simovski[1†], Daniel Vodák[2†], Sveinung Gundersen[1,3], Diana Domanska[1], Abdulrahman Azab[1,4], Lars Holden[5], Marit Holden[5], Ivar Grytten[1], Knut Rand[3], Finn Drabløs[6], Morten Johansen[7], Antonio Mora[1,8], Christin Lund-Andersen[2], Bastian Fromm[2], Ragnhild Eskeland[8,9], Odd Stokke Gabrielsen[8], Egil Ferkingstad[10], Sigve Nakken[2], Mads Bengtsen[8], Alexander Johan Nederbragt[1,11], Hildur Sif Thorarensen[1], Johannes Andreas Akse[1], Ingrid Glad[3], Eivind Hovig[1,2,5,7] and Geir Kjetil Sandve[1*]

[*]Correspondence: geirksa@ifi.uio.no
[1]Department of Informatics, University of Oslo, Oslo, Norway
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Background: Recent large-scale undertakings such as ENCODE and Roadmap Epigenomics have generated experimental data mapped to the human reference genome (as genomic tracks) representing a variety of functional elements across a large number of cell types. Despite the high potential value of these publicly available data for a broad variety of investigations, little attention has been given to the analytical methodology necessary for their widespread utilisation.

Findings: We here present a first principled treatment of the analysis of collections of genomic tracks. We have developed novel computational and statistical methodology to permit comparative and confirmatory analyses across multiple and disparate data sources. We delineate a set of generic questions that are useful across a broad range of investigations and discuss the implications of choosing different statistical measures and null models. Examples include contrasting analyses across different tissues or diseases. The methodology has been implemented in a comprehensive open-source software system, the GSuite HyperBrowser. To make the functionality accessible to biologists, and to facilitate reproducible analysis, we have also developed a web-based interface providing an expertly guided and customizable way of utilizing the methodology. With this system, many novel biological questions can flexibly be posed and rapidly answered.

Conclusions: Through a combination of streamlined data acquisition, interoperable representation of dataset collections and customizable statistical analysis with guided setup and interpretation, the GSuite HyperBrowser represents a first comprehensive solution for integrative analysis of track collections across the genome and epigenome. The software is available at: https://hyperbrowser.uio.no

Keywords

genomics, epigenomics, statistical genomics, genome analysis, genomic track, Galaxy, data integration

Background

Improvements in sequencing technologies in recent decades have enabled the determination of the DNA sequences of many large genomes as well as their functional interrogation. Genome-wide profiles for a variety of biological features are being systematically generated for a wide range of cell types, often via concentrated efforts by dedicated consortia. The Encyclopedia of DNA Elements (ENCODE) [1]

project marked a substantial leap in this respect, by making available to the human genomics community a broad collection of cell line-specific data on DNA accessibility and transcription factor binding. The NIH Roadmap Epigenomics Mapping Consortium further contributed a significant amount of additional tissue- and cell-type-specific data to the public domain, including DNA methylation and histone modification profiles for a large number of primary cells. Kundaje et al. [2] refer to the combined collection of ENCODE and Roadmap data as 127 human reference epigenomes. Most of these datasets are in the form of genomic tracks, i.e. sets of elements anchored to locations in a reference genome, which provide a good foundation for the integration of data representing disparate genomic features.

The widespread utilization of these immense amounts of available data is hampered by a lack of tools providing automatic data integration and sound statistical analysis of large collections of diverse datasets. Frameworks and toolkits such as Bioconductor [3] (R), bedtools [4] (command line), Galaxy [5] and HyperBrowser [6] (web interface) have enabled the robust processing and analysis of genomic tracks with reduced development effort using a variety of interfaces. However, these tools are essentially limited to analyses involving either a single track or a pair of tracks, with no support for the analysis of track collections beyond the trivial concatenation of results per track. For investigations aiming to exploit larger data collections through comparative analyses across epigenomes or across genomic features, no general solutions are available (on any platform). Dedicated solutions do exist for specific applications (e.g., assessing a cell type-specific accessibility of a set of single nucleotide polymorphisms (SNPs) [7, 8] or annotating genomic variants [9, 10, 11, 12]), for specific analytical scenarios (e.g., enrichment analysis of one track against a collection [13]), and for specific basic operations (e.g., calculating the number of base pairs covered by all tracks in a collection [14] or computing the intersection of a collection of tracks with the elements of a single query track [10]). Figure 1 presents these different frameworks and dedicated solutions in context. The lack of comprehensive methodologies leads to ad hoc development of analytical solutions in attempts to answer novel questions that draw on the power of large public or in-house data collections. This may severely limit exploitation of the full potential of current experimental technologies and public data repositories, particularly by research groups with limited bioinformatics resources. Furthermore, the prevalence of ad hoc solutions has a negative impact on reproducibility. A new layer of computational methodology is thus needed to directly approach generic questions formulated in the domain of track collections.

Here, we present GSuite HyperBrowser, the first comprehensive solution for the analysis of track collections across the genome and epigenome. GSuite HyperBrowser is an open-source, web-based system that enables analysis of a broad array of both hypothesis-driven and data-driven questions that may be posed using large collections of genomic tracks. We focus on questions of a comparative nature, where a track is contrasted to (or analyzed in the context of) other tracks. The intended input is one or more carefully assembled collections of tracks, with the tracks of a collection typically varying along a single dimension of interest. The input could be a collection of tracks for the same histone modification across cell types or a collection of tracks representing different histone modifications in the same cell type.

The system uses a formalized representation of track collections and includes tools for compiling new collections from local files or public repositories. Analytical questions may relate to which tracks stand out from such a collection, which tracks of a collection are the most similar to a separate (query) track, or how the occurrence or co-occurrence of elements from individual tracks in the collection varies along the genome. Included within the system is guidance on how these generic questions can be meaningfully interpreted with respect to a specific genomic feature.

Results

Overview

The present work is concerned with sets of information elements anchored to specific coordinates in a reference genome, which we refer to as genomic tracks (short form: tracks). A genomic track may consist of e.g. the genome-wide set of experimentally determined locations of DNA methylation or DNA binding by a transcription factor. Often, an investigation may involve a carefully selected collection of tracks representing either different genomic features for a single cell type or a single feature for multiple cell types. We refer to a collection of tracks selected for a particular analytical purpose as a suite of tracks (short: suite).

We define a simple and intuitive tabular format, GSuite, to represent suites of tracks. The GSuite format can represent data at a local or remote server, can include metadata, and can be seamlessly exchanged between individual tools in an analysis workflow. To allow efficient compilation of track suites from a variety of public repositories (like ENCODE and Roadmap Epigenomics) and thus enable integration of disparate data sources, we propose that rather than downloading and reorganizing tracks according to a unified structure, a concept akin to database views is preferable; tracks can be browsed and selected in a unified manner but are retrieved from their respective sources only when a user assembles a track suite.

Even for a pair of tracks, many different questions can be asked regarding their relations [15]. In principle, the number of possible relations that can be queried for multiple tracks grows exponentially with the number of tracks involved. Also, the complexity of defining and interpreting analyses involving multiple heterogeneous tracks is very high. A particularly useful type of question is the comparative assessment of tracks in a suite, where the tracks may be contrasted based on their relation to one another, to a particular separate track or to tracks of another suite. We delineate a set of generic questions that are useful across a broad range of investigations, explore their characteristics, and present a statistical methodology for their resolution. Table 1 lists five of the main questions, along with associated descriptive statistics and hypothesis tests (details provided in Additional file 1). The descriptive statistics can be based on different measures of similarity, and the hypothesis tests can be based on different null models [16]. A schematic view of the statistical analysis related to one of these questions is provided in Figure 2.

The representation, acquisition and analysis of track suites are implemented in a comprehensive, open-source software system, GSuite HyperBrowser. The system builds on the Genomic HyperBrowser [6, 15] and offers a web-based interface powered by Galaxy [5], with several separate tools for the compilation, preprocessing and analysis of track suites (Figure 3). The web interface includes an interactive

tutorial to help new users quickly get up to speed with meaningful analyses, guidance for every tool, results in the form of sortable tables and customizable plots, and a set of thoroughly annotated examples of biological investigations.

Illustrative example

As an illustrative example, consider the exploration of how binding sites for a given transcription factor (TF) co-occur with binding sites of other TFs and with various epigenomic marks. Because TF binding varies between cell types, such an exploration should be conducted in a cell type-specific context. Here, we describe a process for determining the co-occurrence of ChIP-seq peaks for the GATA1 TF versus other TFs and functional epigenomic elements in K562 cells, an established cell line for which abundant experimental data are available. All analysis steps are performed using tools within the GSuite HyperBrowser system. Further details of the analysis and biological interpretations are discussed in Additional file 2.

The first step is to browse available experimental datasets for K562 cells in the ENCODE repository, compile a GSuite file referring all K562 ENCODE tracks and download these to the server (318 tracks). Using tools for GSuite customization, we isolated a single GATA1 track and compiled a suite of the 317 remaining tracks.

We then determined which tracks (in the suite) exhibit the strongest similarity (in terms of peak co-occurrence) with the GATA1 track. The most critical aspect of such an analysis is the precise specification of the measure of similarity (co-occurrence). By selecting the tetrachoric correlation [17, 18, 19] as similarity measure, we obtain a ranking of tracks that is not too dominated by the strongly varying number of elements per track. The tetrachoric correlation, ρ, is defined by assuming that the two tracks are generated by thresholding an underlying continuous, bivariate normally distributed variable, where ρ is then defined as the correlation in the underlying bivariate normal. The tetrachoric correlation ρ can be easily estimated from given tracks, for example using maximum likelihood techniques; we have used the R-package polycor [20] to estimate ρ. Using this measure, the transcription factors SMARCA4, TAL1, EP300 and STAT5A were identified as high-ranking. These TFs have all been previously reported as relevant for GATA1 (see discussion in Additional file 2).

Because we did not filter out any K562 tracks included in the suite, the ranking includes experimental replicates for GATA1 as well as non-TF datasets such as histone modifications and DNase I accessibility. This provides a broad view of co-occurrence, including indications for TF cooperation, consistency across experimental replicates for the same TF, and the association of GATA1 with different chromatin states. As a confirmatory extension of the analysis, one can examine whether the high-ranked tracks are significantly more similar to GATA1 than the average for all tracks in the suite. This question can be answered by a hypothesis test available in the same tool used to produce the ranking; it uses a test statistic comparing the similarity of each track to the average of the suite. Different null models may be reasonable; for instance, a null model may assume that the data in the whole suite are fixed, whereas the peak locations in the separate track (GATA1) are assumed to be stochastic according to a distribution that preserves the empirical distribution of lengths and distances between the peaks [15]. Because

an average across the suite forms part of this test statistic, data for the whole suite are required to compute each single measure, meaning that the analysis is at the integrative multiplicity level (as defined in the section on Classes of multiplicity).

Representing suites of genomic tracks: the GSuite format

Fundamentally, a collection of datasets is fully defined by a set of references to its constituents. For convenience, a plain text file of Uniform Resource Locators (URLs) for the contained datasets should be valid as a representation of a dataset collection. To further support relevant analyses, the format should permit inclusion of metadata defining important attributes of each individual dataset.

We have defined a simple format that meets these requirements: GSuite. A plain text file of one URL per line is a valid GSuite instance. The format further allows the definition of headers that, among other functions, declare whether the included datasets are available locally or remotely. A tool that downloads datasets referred to by a collection can then iterate through the source GSuite, download each referred file, and replace the URLs with paths to the locally stored files. In addition to the URLs of the tracks, a GSuite file may include tab-separated columns representing metadata values for each dataset. A full definition of the GSuite format is provided in Additional file 3.

Compiling suites from public repositories

Although repositories such as ENCODE and Roadmap Epigenomics provide free access to large amounts of data, they are not designed for the extraction of large numbers of datasets according to shared characteristics, e.g., extracting large suites of tracks tailored toward a particular analysis. Furthermore, the different repositories do not use a common nomenclature, hindering the integration of related data from several repositories.

A common solution to the integration of data from multiple repositories is to download all data from their respective sources, and construct a meta-repository structured according to a common terminology (e.g., [15, 21, 22]). However, such manual curation and organization is laborious, susceptible to errors or misunderstanding, and can easily become outdated. We therefore adopted a different approach to integrate tracks from multiple sources. Rather than downloading and re-organizing genomic tracks, we use a concept akin to database views; users can browse and select remotely located tracks based on metadata, resulting in a list of URLs of the chosen tracks (GSuite). The GSuite can be further modified and shared as a simple text file. The underlying genomic tracks are only downloaded when a user explicitly asks to create a local copy of the data.

As a low-level access point, we provide a single interface for accessing different repositories according to their original (repository-specific) metadata terminology. This interface avoids the loss or misrepresentation of the exact metadata provided by the individual repositories.

We also provide a high-level access point that sacrifices some degree of metadata precision to permit selection of related tracks across sources according to a unified vocabulary (e.g., all tracks for a particular histone modification across repositories). The high-level access point builds on the low-level access point and is based on a

curated transformation of individual repository-level vocabularies into the unified vocabulary.

The low-level and high-level access points currently support ENCODE [1], Roadmap Epigenomics [2], the International Cancer Genome Consortium data portal [23] and the NHGRI-EBI GWAS Catalog [24].

Classes of multiplicity for analyses of track suites

The analysis of multiple tracks ranges from simple repetition of the same computation on each track to analyses in which the tracks are highly intertwined in the computations and interpretations.

To better delineate the different levels of integration associated with various analyses, we define the following classes of multiplicity for track suite analyses:

Trivial multiplicity (Figure 4(A)): A statistic is computed for each track in a suite, but the computed values are neither compared nor integrated across tracks in the suite of interest. This resulting list of values per track can be convenient for obtaining an overview of a suite. Because it is merely a repetition of computations, it does not introduce any challenges related to multiplicity. An example of trivial multiplicity is to count the number of peaks for each track for transcription factor binding sites in a given cell type.

Contrasting multiplicity (Figure 4(B)): A statistic is computed separately for each track of a suite, possibly in relation to reference tracks (outside the suite), with an aim of contrasting (typically ranking) the values computed for each track from the suite. Co-occurrence is typically of main interest. Although the computations are performed separately (as for trivial multiplicity), the aim of comparing the computed values puts additional requirements on the statistics used. As discussed in Additional file 2, measures designed to capture the similarity/co-occurrence of tracks may be inappropriately affected by the number of elements in each track. An example of contrasting multiplicity is evaluating the co-occurrence of binding sites of a selected transcription factor (TF) against each track from a suite of transcription factor ChIP-seq peak tracks (note: this is the same example as was presented in the ,,Illustrative example" section above)[1]. In this example, using the Jaccard index [25] as the similarity measure produced a ranking that appeared severely affected by the overall number of peaks in each track from the suite. The severity of this effect is also shown on simulated data (Additional file 2). Use of the Forbes coefficient [26] or tetrachoric correlation [17, 18, 19] did not show such an effect, and resulted in a markedly different ranking. Especially the use of the tetrachoric correlation resulted in a biologically very reasonable ranking of potentially cooperating TFs. Since track size has such a strong influence on the Jaccard index, we generally don't recommend its use in situations where tracks are to be ranked.

Integrative multiplicity (Figure 4(C)): A statistic is computed based on pair-wise measures across all tracks in a suite. The statistic may be a single value representing the suite as a whole or it may be in the form of one value per track from the suite.

[1]As in ,,Exploring transcription factor co-occurrence using two alternative measures of similarity", one of the complex example analyses on the GSuite HyperBrower website, which is also briefly presented in the ,,illustrative example" section above

For descriptive statistics computed per track, integrative multiplicity implies that the value of a given track will depend on the context of other tracks included in the suite. An example of integrative multiplicity is the computation of how typical each track in a suite is with respect to the suite, i.e., its average co-occurrence with other tracks in the suite. A computational challenge associated with the integrative multiplicity class is that the data for each track are typically used in several parts of the computations. A simple algorithm would thus either need to read the same data repeatedly from physical storage or simultaneously store the data for all tracks in memory. More advanced algorithms based on map-reduce and memoization of intermediate computations would therefore generally be preferable (and are applied in GSuite HyperBrowser).

Higher-order multiplicity (Figure 4(D)): A statistic is defined based on higher-order relations (beyond pairwise) between the tracks in a suite, implying that a computation must work on elements from many or all tracks from a particular genomic region simultaneously. Then, the statistic cannot be subdivided into multiple pairwise across-track computations. An example is the computation of how many base pairs across the genome are associated with open chromatin in more than half of a set of considered cell types (covered by more than half of the genomic tracks of a suite).

Hypothesis testing

A hypothesis test for multiple tracks investigates whether the aspect of interest for the track or tracks in question is present in the data more/less than what is expected by chance. For all questions in Table 1, we have defined an associated statistical test that can facilitate the assessment of the robustness of the effects observed from the descriptive statistics (Additional file 1).

Statistical tests can be based on parametric distributions or Monte Carlo simulations. Due to the complex structure of a genome, genomic data sets are often not well described by simple parametric distributions. For this reason, simulation has been the preferred choice even for relations involving only a pair of tracks [15, 27]. We have further demonstrated that the simplifying assumptions that are typically required to allow parametric testing on genomic track data will often increase the risk of false-positive findings [16]. Based on such considerations, we find that for the questions of Table 1, the limitations and simplifying assumptions required for parametric testing make Monte Carlo-based simulation a more promising direction.

The following are the main elements of a Monte Carlo-based statistical test:

1. *a test statistic:* a measure that describes the aspect of interest;
2. *a null model:* a model that tracks would follow if generated by chance;
3. *a null distribution:* the distribution of the test statistic when data follow the null model; and
4. *a p-value:* the proportion of the null distribution that is more extreme than the value of the test statistic on the observed (real) data.

For statistical testing to be meaningful, a test statistic must be specified that precisely matches a particular aspect (question) of interest and assumes a realistic (relevant) null model.

Our approach follows [15]: we argue that good, robust results can be obtained by preserving some structure from the tracks while performing the randomization

algorithm. After specifying what we consider relevant null model assumptions, we derive algorithms for sampling tracks from a particular null model and computing the test statistic for each simulated track. We observe that the relevant null models (and thus the associated simulation algorithms) are mostly shared between questions and can be divided into the following three categories (described in terms of simulation algorithms):

- Sampling algorithms that treat each track separately. Any sampling algorithm for single tracks can be extended in this manner to suites, e.g., those presented in [15].
- Sampling algorithms that sample elements across tracks from a suite. Track segments (pairs of reference genome coordinates) can be placed in a single pool shared across tracks and sample segments for each track with or without replacements from this pool and with or without preserving the variation of frequency and length of segments across the tracks. A particular challenge with this sampling approach is how to handle intra-track overlap of segments without introducing sampling biases. Further details on alternative sampling algorithms are provided in Additional file 1.
- Sampling algorithms sampling across suites. These fall into the following two types: one type that pools track elements across both tracks and suites and thus represents a (slight) further complication of the previous category and a second type that permutes entire tracks between suites. Further details are provided in Additional file 1.

There is a crucial difference in the interpretation between hypothesis tests at the contrasting and integrative multiplicity levels. A statistical test that uses a pairwise track similarity measure as a test statistic and a sampling algorithm that treats each track separately will result in p-values at the contrasting multiplicity level (p-values relate to the null hypothesis for each track from a suite in isolation). Such p-values do not provide information about how a particular track is differentiated from other tracks in a collection, but the p-values of different tracks can be compared to assess the relative confidence. By contrast, if either the test statistic is defined across tracks from the suite or if the sampling algorithm draws elements across tracks, the resulting p-values will be at the integrative multiplicity level. Such p-values may represent null hypotheses related to whole suites or how a given track is differentiated from the remaining tracks in the suite.

The basic mode as an interactive tutorial of the system

To accommodate a broad range of usage scenarios, the main tools in the GSuite HyperBrowser are defined in a generic and highly customizable manner. Generality of tools and a rich palette of parameter options are often indispensable for appropriate handling of data during the course of an actual project (and often have important consequences for the interpretation of results), but might mean unnecessary complexity for new users who wish to first familiarize themselves with the system. The system therefore includes a dedicated tutorial version of the tool interface, which simplifies the definitions of basic analyses and streamlines the learning experience. This ,,basic mode" of the system offers a simplified view of a tool's parameter list, hiding options that are typically sufficiently represented by the default values during initial exploratory test runs by users. Perhaps most importantly,

the entry point of the basic mode is a set of interactive analysis examples that illustrate the typical usage of the GSuite tools within particular domains (e.g., the study of genome variation or the study of transcription factor binding). Each example includes detailed instructions for performing a simple integrative analysis and provides relevant datasets necessary for its execution. The examples also offer information regarding generalization of the presented analyses and guidance for utilizing one's own datasets. Entering and leaving the tutorial mode is possible at any time, which will respectively hide or reveal the full set of parameters defined for each tool. Figure 5 shows a screenshot of the tool interface as it appears on the web server.

Examples of biological investigations using the system

While the interactive tutorial illustrates core analytical approaches for a breadth of biological questions, a full investigation will usually involve its own specific steps for data preparation and supporting analysis. To provide an impression of the variety of aspects that may be involved, we include a set of transparent and reproducible examples of biological investigations using the system. The investigation examples are available under the "Examples" tab on the system front page and include an example that reproduces individual findings from the literature (relationship between mutations in a given cancer and cell-specific open chromatin), an example of novel investigations (whether SNPs associated with various diseases are located in miRNA genes), an example of studying experimental biases/artifacts (clustering of tracks associated with different cell types and experimental setups) and an example of studying computational biases (how the exact formula used to measure track similarity has a decisive impact on the results and interpretations).

Discussion and Conclusions

Reference genomes have allowed a broad range of genomic features to be represented in a uniform manner, which facilitates data integration and the discovery of relations and interplay between various features. With recent initiatives to unravel data from multiple epigenomes (cell-type-specific data for a variety of epigenetic marks), a new layer of computational methodology is needed. Similar to the previous generation of computational tools that allowed a question regarding a genome-scale data set to be resolved through a single operation, the next generation of tools (or an updated version of existing tools) should directly approach questions formulated in the domain of collections of genomic tracks.

The most trivial level of functionality for analyzing data collections, based on iterative, single or pair-wise analysis of genomic tracks, is already available on various platforms for genomic track analysis. More complex solutions regarding track collections have been provided only for specific questions by means of dedicated tools (e.g., LOLA [13]). The analysis of track collections (e.g., analysis across a set of functional elements or cell types) has received little attention in the literature. We present here a first step in this direction.

The present work includes three distinct contributions: 1) a computational and statistical methodology for compiling and analyzing collections of genomic tracks; 2) an implementation of the proposed methodology in the form of a large open-source,

integrated software system; and 3) a web-based interface to the developed functionality. The user interface enables meaningful analysis customization by providing expert guidance.

The main approach for the integration of data in the bioinformatics field has been to download data from multiple sources and restructure it according to a uniform hierarchy ([21, 28]). Here, we adopted a different approach by developing solutions to allow users to retrieve data from databases when a specific collection of tracks is needed (instead of downloading and re-organizing data in a general manner in advance). This approach has advantages and disadvantages. Downloading and integrating track collections as needed introduces a delay for users at the time of compilation compared to relying on pre-collected data. This delay is to some degree rectified by a scheme for locally caching data previously downloaded (by any user). The advantage of the chosen approach is that as long as the repositories continue to release their data according to the same protocol, the tool will continuously provide access to all available data in their latest versions. Another strong advantage is the transparency of the approach—users can directly view the URLs at which data were retrieved and the exact time the data were retrieved from a given repository. The currently supported repositories all contain data for the human genome, but the methodology can be readily applied to data connected to any reference genome.

Due to the size and heterogeneity of the genomes of higher organisms, even analyses of single genomic tracks can be complex. Integrative analyses across multiple tracks (typically across cell types or features) add a further layer of complexity. To cope with this complexity, highly customizable tools and extensive user guidance are essential. By developing an integrated software system with a set of robust components for data handling and statistical analysis at the core, we have enabled a range of sophisticated analyses to be performed with limited effort. The developed methodology is accessible to a broad user base via the system's web interface, which provides inbuilt tool guidance and offers an interactive tutorial with a rich list of domain-specific analysis suggestions. Transparency and reproducibility of analyses are ensured by integration with the Galaxy framework, where data, tool and parameter choices are automatically tracked in the background and any step in the analysis can be repeated with the option of changing the original data or parameters.

The Galaxy system also includes a native way of representing multiple datasets, termed dataset lists/collections, which we consider mostly complementary to GSuite. A strong aspect of dataset lists is their tight integration with Galaxy tool execution, which allows any standard Galaxy tool to be executed iteratively on each dataset of a collection. Through its representation as a tabular text file, GSuite is interoperable across systems and can be easily manipulated using any tool or software that operates on tabular datasets, inside or outside the Galaxy system. Furthermore, GSuite supports the specification of custom metadata for each dataset in a collection, which is exploited extensively in our tools and example analyses. We believe a general integration of the GSuite format within the Galaxy system, including functionality for converting between GSuite and dataset lists, could improve the usability of both the GSuite HyperBrowser and the standard Galaxy platform.

The methodology presented here does not cover the full spectrum of analyses that can be envisioned for collections of genomic tracks. First, the current statistics

and null models only relate to pure location data (Point and Segment tracks [29]). Extending the work to handle *Valued Points* and *Segments* (e.g., genes with expression values and tracks from case vs. control elements) as well as *Function* tracks (e.g., signal tracks with ChIP-seq intensities) would clearly broaden the range of supported biological investigations. Second, the present methodology is primarily focused on questions that can be reduced to pairwise track relations. Analysis of higher-order relations between functional elements is a very interesting challenge but requires methodological development beyond what is described here. Third, even for the class of analyses considered here, there are many further questions for which statistical methodology would be useful. Fourth, although data from any source can be uploaded to the system, a consistent terminology for track metadata would enable better unified access to track data sources and their content. We believe that the development of a widely accepted ontology for describing biological and experimental characteristics of tracks should be given high priority to ease data integration and avoid misinterpretation of results achieved when employing public data for research. Ideally, this should be organized as a community effort to ensure international uptake. Fifth, experimental data at the single-cell level is rapidly becoming a powerful tool in biomedical research [30, 31]. Although the methodology presented here can be used directly on single-cell data, these data may give rise to a range of additional questions beyond what is considered in the present work. Through a principled methodological approach and implementation based on generic core components, the open-source GSuite HyperBrowser system is prepared for future extensions in a variety of dimensions.

In conclusion, we believe the GSuite HyperBrowser would permit robust and reproducible solutions to a breadth of cases for which ad hoc development is the only current possibility.

Methods

System implementation

The GSuite HyperBrowser is an integrated software system written mainly in Python, with extensive use of the NumPy library for efficient data handling, as well as some supporting code in R and Javascript (in total, 170,000 lines of code). The GSuite HyperBrowser makes use of code components from the Genomic HyperBrowser [15] to represent individual tracks and to analyze single tracks and pairwise relations between tracks. The user interface is based on the Galaxy system [5], which ensures robust user and dataset management, and includes features supporting reproducible research. To provide users with a more dynamic user interface, the tools in GSuite HyperBrowser is based upon Galaxy ProTo (https://github.com/elixir-no-nels/proto), an alternative tool definition API for the Galaxy framework. To ensure computational efficiency, track data are preprocessed into an indexed, binary format based upon arrays written consecutively to disk [29], while analysis computations are based on a map-reduce scheme that limits memory requirements and a scheme for memoizing intermediate computations [15].

GSuite representation

Collections of tracks are represented as lists of references (URLs) with corresponding metadata in the GSuite tabular text format. The system includes robust functionality for composing, modifying and validating collections in this format. The

system also includes functionality for crawling and for searching and retrieving data from public repositories. The crawling functionality works similarly to a web crawler, accessing metadata from supported repositories to generate a database of the available datasets in the form of URLs along with metadata accompanying each dataset. This database can then be queried on metadata contents, resulting in a novel GSuite file containing Uniform Resource Identifiers (URIs) to original, remotely stored datasets. Before analysis, remote datasets of a GSuite file can be retrieved and stored locally on the web server in hidden Galaxy history elements, resulting in a transformed GSuite file with custom Galaxy URIs that point to such storage. A caching scheme is also implemented, making sure that the datasets for each unique URI that refers to stable content is only retrieved once. The caching simply stores the Galaxy URI for the first retrieval in a register and makes sure that consecutive retrievals result in the same URI.

Descriptive statistics and null models

The test statistic needs to be custom-tailored to a particular question. It will thus vary between different questions involving suites of tracks, and will also vary according to slight variations of each question. Still, we find that test statistics for the whole range of questions we have studied can be defined based on a shared hierarchy:

- Pairwise track statistic (T): computes a measure of co-occurrence between a pair of tracks, e.g. the Forbes measure ($\frac{N*|A_i \cap A_j|}{|A_i|*|A_j|}$, where A_i and A_j are the set of genome locations (bps) covered by two tracks i and j, while N is the size of the genome) [26]. This can be a final per-track result in itself (at the contrasting multiplicity level) or part of a higher order computation.
- Integrative statistic (Q): combines values of T for multiple track pairs. This operates on a structure of track pairs (and corresponding T values), e.g. a single track paired with each other track of a suite. The combination of T values can e.g. be the average, max or min of values of T (e.g., $\frac{1}{n-1} * \sum_{j \neq i} T(A_i, A_j)$, where n is the number of tracks in the suite). Analyses based on a Q-statistic are by definition at the integrative or higher-order multiplicity levels.
- Suite statistic (R): Statistic that describes an entire suite. It may combine multiple values of Q. Each Q-value will typically represent a one-to-many computation between tracks in a suite, with the R-value typically representing a many-to-many combination of tracks in a suite. The combination of Q values can e.g. be the average, max or min of values of Q (e.g., $\frac{1}{n} * \sum_i Q(A_i, A_{-i})$). Analyses based on a R-statistic are by definition at the integrative multiplicity level.
- Pairwise suite statistic (S): Statistic that describe the relationship between two suites. Also this statistic may combine multiple values of Q in the same manner as the R-statistic. Analyses based on an S-statistic are by definition at the integrative or higher-order multiplicity level.

Most hypothesis tests in the system are based on Monte Carlo evaluation of p-values, where a particular simulation algorithm produces explicit tracks for the null model and a particular test statistic is used to generate values for the null distribution. Several alternative simulation algorithms are proposed, preserving distinct properties within the scope of individual tracks or across the collection.

Detailed formulas for descriptive and test statistics, as well as detailed sampling algorithms for Monte Carlo evaluation of statistical significance, are provided in Additional file 1.

List of abbreviations

ENCODE, The Encyclopedia of DNA Elements; SNP, Single Nucleotide Polymorphism; TF, Transcription Factor; URI, Uniform Resource Identifier; URL, Uniform Resource Locator

Declarations

Availability and requirements
Project name: GSuite HyperBrowser
Project home page: https://hyperbrowser.uio.no/
Github repository: https://github.com/hyperbrowser/genomic-hyperbrowser
Operating system(s): Platform independent
Programming language: Python 2.7
Other requirements: None
License: GPLv3
Any restrictions to use by non-academics: None
All data and analyses referred to in the manuscript are available from the ,,Examples" tab on the front page of the GSuite HyperBrowser web page. The analyses are available as Galaxy histories, which can be viewed or "imported" for further inspection. Full analysis specifications are available through the "run this job again" button present on history elements (this functionality also allows the analyses to be re-run in original or modified form). Data and results can be directly viewed or downloaded.

Availability of supporting data
A snapshot of the version of the Gsuite HyperBrowser source code used in this paper is archived in the GigaScience database GigaDB [32]

Competing interests
The authors declare that they have no competing interests.

Funding
This work was supported by the Research Council of Norway (under grant agreements 221580, 218241 and 231217/F20), by the Norwegian Cancer Society (under grant agreements 71220—PR-2006-0433 and 3485238-2013) and by the South-Eastern Norway Regional Health Authority (under grant agreement 2014041).

Author's contributions
LH, MH, IG, KR, EF, IKG and GKS developed the statistical methodology. SG, AA, BS and GKS defined the GSuite format and the suite compilation approach. BS, SG, DD, AA, IG, KR, MJ, AM, HST, JAA and GKS implemented the software system. BS, DV, SG, DD, FD, AM, CLA, SN, MB, AJN, EH and GKS tested and validated the system. DV, DD, FD, CLA, BF, RE, OSG, SN, MB and GKS developed the biological examples. BS, DV, FD, EH and GKS drafted the initial manuscript. GKS conceived the approach. All authors participated in the manuscript development, and read and approved the final manuscript.

Acknowledgments
We thank ELIXIR Norway for important contributions to the development of the GSuite HyperBrowser.

Additional files
Additional file 1: A text document describing statistical measures and hypothesis tests for suites of genomic tracks. The document contains detailed formulas and algorithms for statistical methodology used by the GSuite HyperBrowser system. (PDF format, 13 806 KB)
Additional file 2: A text document providing a critical evaluation (on simulated and real data) of how particular choices of similarity measures may influence genome-level analysis results. (PDF format, 1233 KB)
Additional file 3: A text document with a detailed specification of the GSuite file format. (PDF format, 65 KB)

Author details
[1]Department of Informatics, University of Oslo, Oslo, Norway. [2]Dept. of Tumor Biology, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway. [3]Department of Mathematics, University of Oslo, Oslo, Norway. [4]Research Support Services Group, University Center for Information Technology, Oslo, Norway. [5]Statistics For Innovation, Norwegian Computing Center, Oslo, Norway. [6]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. [7]Institute for Medical Informatics, The Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway. [8]Department of Biosciences, University of Oslo, Oslo, Norway. [9]Norwegian Center for Stem Cell Research, Department of Immunology, Oslo University Hospital, Oslo, Norway. [10]Science Institute, University of Iceland, Reykjavik, Iceland. [11]Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Oslo, Norway.

References

1. Consortium, E.P., Aldred, S.F., Collins, P.J., Doyle, F., Epstein, C.B., Frietze, S., Kaul, R., Lee, B.-K., Rosenbloom, K.R., Sabo, P., Safi, A., Shoresh, N., Simon, J.M., Altshuler, R.C., Djebali, S., Dong, X., Furey, T.S., Gerstein, M., Greven, M., Hardison, R.C., Harris, R.S., Hoffman, M.M., Kelllis, M., Kundaje, A., Lassman, T., Li, Q., Merkel, A., Parker, S.C.J., Reddy, T.E., Rozowsky, J., Schlesinger, F., Ward, L.D., Whitfield, T.W., Wu, W., Bernstein, B.E., Dunham, I., Gunter, C., Pazin, M.J., Lowdon, R.F., Dillon, L.A.L., Adams, L.B., Kelly, C.J., Zhang, J., Wexler, J.R., Green, E.D., Good, P.J., Feingold, E.A., Birney, E., Elinitski, L., Gingeras, T.R., Hubbard, T.J., Myers, R.M., Snyder, M., Starnatoyannopoulos, J.A., Tennebaum, S.A., White, K.P., Wrobel, J., Risk, B.A., Kuiper, H.C., Maier, C.W., Xie, L., Ernst, J., Kheradpour, P., Mikkelsen, T.S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M.J., Durham, T., Ku, M., Truong, T., Eaton, M.L., Davis, C.A., Dobin, A., Lassmann, T., Lagarde, J., Lin, W., Xue, C., Marinov, G.K., Khatun, J., Williams, B.A., Zaleski, C., Röder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chen, X., Curado, J., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H.P., Howald, C., Jha, S., Johnson, R., Kapranov, P., Kingswood, C., Luo, O.J., Park, E., Preall, J.B., Presaud, K., Ribeca, P., Robyr, D., Ruan, X., Sammeth, M., Sandu, K.S., Schaeffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Wang, H., Yu, Y., Hayashizaki, Y., Harrow, J., Hubbard, T.J., Reymond, A., Antonarakis, S.E., Hannon, G.J., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Sloan, C.A., Learned, K., Malladi, V.S., Wong, M.C., Barber, G.P., Cline, M.S., Dreszer, T.R., Heitner, S.G., Karolchik, D., Kent, W.J., Kirkup, V.M., Meyer, L.R., Long, J.C., Maddren, M., Raney, B.J., Song, L., Grasfeder, L.L., Giresi, P.G., Lee, B.-K., Battenhouse, A., Sheffield, N.C., Showers, K.A., London, D., Bhinge, A.A., Shestak, C., Schaner, M.R., Kim, S.K., Zhang, Z.Z., Mieczkowski, P.A., Mieczkowska, J.O., Liu, Z., McDaniell, R.M., Ni, Y., Rashid, N.U., Kim, M.J., Adar, S., Zhang, Z., Wang, T., Winter, D., Iyer, V.R., Lieb, J.D., Crawford, G.E., Li, G., Sandhu, K.S., Zheng, M., Wang, P., Fullwood, M.J., Pauli, F., Gertz, J., Vielmetter, J., Partridge, E.C., Varley, K.E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K.M., Anaya, M., Cross, M.K., King, B., Muratet, M.A., Newberry, K.M., McCue, K., Nesmith, A.S., Fisher-Aylor, K.I., Pusey, B., DeSalvo, G., Parker, S.L., Balasubramanian, S., Davis, N.S., Meadows, S.K., Eggleston, T., Newberry, J.S., Levy, S.E., Absher, D.M., Mortazavi, A., Wong, W.H., Blow, M.J., Visel, A., Pennachio, L.A., Elnitski, L., Margulies, E.H., Petrykowska, H.M., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Chrast, J...: An integrated encyclopedia of DNA elements in the human genome. Nature **489**(7414), 57–74 (2012)
2. Kundaje, A., Meuleman, W., Kheradpour, P., Wang, J., Ziller, M.J., Quon, G., Sandstrom, R.S., Eaton, M.L., Pfenning, A., ClaussnitzerYaping Liu, M., Coarfa, C., Alan Harris, R., Gjoneska, E., Xie, W., David Hawkins, R., Lister, R., Gascard, P., Moore, R., Chuah, E., Canfield, T.K., Scott Hansen, R., Kaul, R., Bansal, M.S., Carles, A., Feizi, S., Lowdon, R., Elliott, G., Mercer, T.R., Neph, S.J., Rajagopal, N., Ray, P., Siebenthall, K.T., Sinnott-Armstrong, N.A., Stevens, M., Wu, J., Zhou, X., Abdennur, N., Adli, M., Akerman, M., Barrera, L., Antosiewicz-Bourget, J., Ballinger, T., Barnes, M.J., Bates, D., Bell, R.J.A., Bennett, D.A., Bianco, K., Bock, C., Boyle, P., Brinchmann, J., Caballero-Campo, P., Camahort, R., Carrasco-Alfonso, M.J., Charnecki, T., Chen, H., Chen, Z., Cheng, J.B., Cho, S., Chu, A., Chung, W.-Y., Cowan, C., Athena Deng, Q., Deshpande, V., Diegel, M., Ding, B., Durham, T., Echipare, L., Edsall, L., Flowers, D., Genbacev-Krtolica, O., Gifford, C., Gillespie, S., Giste, E., Glass, I.A., Gnirke, A., Gormley, M., Gu, H., Gu, J., Hafler, D.A., Hangauer, M.J., Hariharan, M., Hatan, M., Haugen, E., He, Y., Heimfeld, S., Herlofsen, S., Hou, Z., Humbert, R., Issner, R., Jackson, A.R., Jia, H., Jiang, P., Johnson, A.K., Kadlecek, T., Kamoh, B., Kapidzic, M., Kent, J., Kim, A., Kleinewietfeld, M., Klugman, S., Krishnan, J., Kuan, S., Kutyavin, T., Lee, A.Y., Lee, K., Li, J., Li, N., Li, Y., Ligon, K.L., Lin, S., Lin, Y., Liu, J., Liu, Y., Luckey, C.J., Ma, Y.P., Maire, C., Marson, A., Mattick, J.S., Mayo, M., McMaster, M., Metsky, H., Mikkelsen, T., Miller, D., Miri, M., Mukame, E., Nagarajan, R.P., Neri, F., Nery, J., Nguyen, T., O'Geen, H., Paithankar, S., Papayannopoulou, T., Pelizzola, M., Plettner, P., Propson, N.E., Raghuraman, S., Raney, B.J., Raubitschek, A., Reynolds, A.P., Richards, H., Riehle, K., Rinaudo, P., Robinson, J.F., Rockweiler, N.B., Rosen, E., Rynes, E., Schein, J., Sears, R., Sejnowski, T., Shafer, A., Shen, L., Shoemaker, R., Sigaroudinia, M., Slukvin, I., Stehling-Sun, S., Stewart, R., Subramanian, S.L., Suknuntha, K., Swanson, S., Tian, S., Tilden, H., Tsai, L., Urich, M., Vaughn, I., Vierstra, J., Vong, S., Wagner, U., Wang, H., Wang, T., Wang, Y., Weiss, A., Whitton, H., Wildberg, A., Witt, H., Won, K.-J., Xie, M., Xing, X., Xu, I., Xuan, Z., Ye, Z., Yen, C.-A., Yu, P., Zhang, X., Zhang, X., Zhao, J., Zhou, Y., Zhu, J., Zhu, Y., Ziegler, S., Beaudet, A.E., Boyer, L.A., Haussler, D., Jones, S.J.M., Li, W., McManus, M.T., Tlsty, T.D., Tsai, L.-H., Wang, W., Costello, J.F., Hirst, M., Meissner, A., Milosavljevic, A., Stamatoyannopoulos, J.A., Wang, T., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Zhang, Z., Wang, J., Amin, V., Whitaker, J.W., Schultz, M.D., Ward, L.D., Sarkar, A., Wu, Y.-C., Pfenning, A.R., Wang, X., Claussnitzer, M., Liu, Y., Harris, R.A., Shoresh, N., Epstein, C.B., Leung, D., Hawkins, R.D., Hong, C., Mungall, A.J., Tam, A., Hansen, R.S., Sabo, P.J., Dixon, J.R., Farh, K.-H., Karlić, R., Kim, A.-R., Kulkarni, A., Li, D., Onuchic, V., Polak, P., Sallari, R.C., Thurman, R.E., Wu, J., Zhang, B., De Jager, P.L., Farnham, ...: Integrative analysis of 111 reference human epigenomes. Nature **518**(7539), 317–330 (2015)
3. Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K.D., Irizarry, R.A., Lawrence, M., Love, M.I., MacDonald, J., Obenchain, V., Oleś, A.K., Pagès, H., Reyes, A., Shannon, P., Smyth, G.K., Tenenbaum, D., Waldron, L., Morgan, M.: Orchestrating high-throughput genomic analysis with Bioconductor. Nature Publishing Group **12**(2), 115–121 (2015)
4. Quinlan, A.R.: Bedtools: the swiss-army tool for genome feature analysis. Current protocols in bioinformatics, 11–12 (2014)
5. Goecks, J., Nekrutenko, A., Taylor, J.: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology **11**(8), 1–13 (2010). doi:10.1186/gb-2010-11-8-r86

6. Sandve, G.K., Gundersen, S., Johansen, M., Glad, I.K., Gunathasan, K., Holden, L., Holden, M., Liestøl, K., Nygård, S., Nygaard, V., Paulsen, J., Rydbeck, H., Trengereid, K., Clancy, T., Drabløs, F., Ferkingstad, E., Kalas, M., Lien, T., Rye, M.B., Frigessi, A., Hovig, E.: The Genomic HyperBrowser: an analysis web server for genome-scale data. Nucleic Acids Research **41**(W1), 133–141 (2013)
7. Schmidt, E.M., Zhang, J., Zhou, W., Chen, J., Mohlke, K.L., Chen, Y.E., Willer, C.J.: Gregor: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. Bioinformatics, 201 (2015)
8. Dunham, I., Kulesha, E., Iotchkova, V., Morganella, S., Birney, E.: FORGE: A tool to discover cell specific enrichments of GWAS associated SNPs in regulatory regions. F1000Research, 1–44 (2015)
9. McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., Cunningham, F.: Deriving the consequences of genomic variants with the ensembl api and snp effect predictor. Bioinformatics **26**(16), 2069–2070 (2010)
10. Hinrichs, A.S., Raney, B.J., Speir, M.L., Rhead, B., Casper, J., Karolchik, D., Kuhn, R.M., Rosenbloom, K.R., Zweig, A.S., Haussler, D., et al.: Ucsc data integrator and variant annotation integrator. Bioinformatics, 766 (2016)
11. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M.: A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff. Fly **6**(2), 80–92 (2012). doi:10.4161/fly.19695. PMID: 22728672
12. Wang, K., Li, M., Hakonarson, H.: Annovar: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research **38**(16), 164–164 (2010)
13. Sheffield, N.C., Bock, C.: LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics **32**(4), 587–589 (2016)
14. Aszódi, A.: Multovl: fast multiple overlaps of genomic regions. Bioinformatics **28**(24), 3318–3319 (2012)
15. Sandve, G.K., Gundersen, S., Rydbeck, H., Glad, I.K., Holden, L., Holden, M., Liestøl, K., Clancy, T., Ferkingstad, E., Johansen, M., Nygaard, V., Tøstesen, E., Frigessi, A., Hovig, E.: The Genomic HyperBrowser: inferential genomics at the sequence level. Genome Biology **11**(12), 121 (2010)
16. Ferkingstad, E., Holden, L., Sandve, G.K.: Monte Carlo Null Models for Genomic Data. Statistical Science **30**(1), 59–71 (2015)
17. Drasgow, F.: Polychoric and polyserial correlations. Encyclopedia of statistical sciences (1988)
18. Olsson, U.: Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika **44**(4), 443–460 (1979)
19. Pearson, K.: Mathematical contributions to the theory of evolution. vii. on the correlation of characters not quantitatively measurable. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **195**, 1–405 (1900)
20. Fox, J.: Polycor: Polychoric and Polyserial Correlations. (2016). R package version 0.7-9. https://CRAN.R-project.org/package=polycor
21. Gundem, G., Perez-Llamas, C., Jene-Sanz, A., Kedzierska, A., Islam, A., Deu-Pons, J., Furney, S.J., Lopez-Bigas, N.: IntOGen: integration and data mining of multidimensional oncogenomic data. **7**(2), 92–93 (2010)
22. Liu, T., Ortiz, J.A., Taing, L., Meyer, C.A., Lee, B., Zhang, Y., Shin, H., Wong, S.S., Ma, J., Lei, Y., Pape, U.J., Poidinger, M., Chen, Y., Yeung, K., Brown, M., Turpaz, Y., Liu, X.S.: Cistrome: an integrative platform for transcriptional regulation studies. Genome Biology **12**(8), 83 (2011)
23. Zhang, J., Baran, J., Cros, A., Guberman, J.M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., *et al.*: International cancer genome consortium data portal—a one-stop shop for cancer genomics data. Database **2011**, 026 (2011)
24. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., *et al.*: The nhgri gwas catalog, a curated resource of snp-trait associations. Nucleic acids research **42**(D1), 1001–1006 (2014)
25. Jaccard, P.: Etude Comparative de la Distribution Florale dans Une Portion des Alpes et du Jura. Impr. Corbaz, n.p. (1901)
26. Forbes, S.A.: On the Local Distribution of Certain Illinois Fishes: an Essay in Statistical Ecology vol. 7. Illinois State Laboratory of Natural History, n.p. (1907)
27. Bickel, P.J., Boley, N., Brown, J.B., Huang, H., Zhang, N.R.: Subsampling methods for genomic inference. The Annals of Applied Statistics, 1660–1697 (2010)
28. Li, M.J., Wang, L.Y., Xia, Z., Sham, P.C., Wang, J.: GWAS3D: detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. Nucleic Acids Research (2013)
29. Gundersen, S., Kalas, M., Abul, O., Frigessi, A., Hovig, E., Sandve, G.K.: Identifying elemental genomic track types and representing them uniformly. BMC Bioinformatics **12**(1), 494 (2011)
30. Tang, F., Barbacioru, C., Nordman, E., Li, B., Xu, N., Bashkirov, V.I., Lao, K., Surani, M.A.: Rna-seq analysis to capture the transcriptome landscape of a single cell. Nature protocols **5**(3), 516–535 (2010)
31. Rotem, A., Ram, O., Shoresh, N., Sperling, R.A., Goren, A., Weitz, D.A., Bernstein, B.E.: Single-cell chip-seq reveals cell subpopulations defined by chromatin state. Nature biotechnology **33**(11), 1165–1172 (2015)
32. Simovski, B., Vodák, D., Gundersen, S., Domanska, D., Azab, A., Holden, L., Holden, M., Grytten, I., Rand, K., Drabløs, F., Johansen, M., Mora, A., Lund-Andersen, C., Fromm, B., Eskeland, R., Gabrielsen, O.S., Ferkingstad, E., Nakken, S., Bengtsen, M., Nederbragt, A.J., Sif Thorarensen, H., Andreas Akse, J., Glad, I., Hovig, E., Sandve, G.K.: Gsuite hyperbrowser version2.0b3. GigaScience Database (2017). http://dx.doi.org/10.5524/100292
33. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M.J.L.: Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics **25**(11), 1422–1423 (2009)
34. Favorov, A., Mularoni, L., Cope, L.M., Medvedeva, Y., Mironov, A.A., Makeev, V.J., Wheelan, S.J.: Exploring Massive, Genome Scale Datasets with the GenometriCorr Package. PLoS Computational Biology **8**(5), 1002529

(2012)
35. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., Manke, T.: deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research (2016)
36. McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., Bejerano, G.: GREAT improves functional interpretation of cis-regulatory regions. Nature Biotechnology **28**(5), 495–501 (2010)
37. Bock, C., Halachev, K., Büch, J., Lengauer, T.: EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)genomic data. Genome Biology **10**(2), 14 (2009)
38. Halachev, K., Bast, H., Albrecht, F., Lengauer, T., Bock, C.: Epiexplorer: live exploration and global analysis of large epigenomic datasets. Genome biology **13**(10), 1 (2012)

Figure 1

The GSuite HyperBrowser in the context of existing tools and frameworks for genomic track analysis. The codebases of frameworks such as bedtools [4], BioPython [33], Bioconductor [3], Galaxy [5] and the Genomic HyperBrowser [15] add a domain-specific layer on top of general programming languages, providing generic constructs for representing genomic track data and core operations on tracks (including some minimal support for analyzing multiple tracks). The GSuite HyperBrowser codebase is the first general platform to add a new layer of constructs for directly representing collections of tracks and providing core operations (analyses) to be applied to such track collections. Although the functionality of this codebase is provided through a web interface, the codebase is open source, and the same constructs may be used by any other relevant Python-based platform. Also, the underlying approach is general and could be correspondingly implemented in other programming languages. In addition to such general purpose framework, there are a variety of purpose-specific tools for track data. GenometriCorr [34], deepTools2 [35] and GREAT [36] are examples of tools that operate on single/pairs of tracks and support specific analyses or domains. Furthermore, several tools implicitly make use of collections of genomic tracks for analyses in specific domains (e.g., FORGE [8], GREGOR [7] and CISTROME [22]) or for specific types of analyses (e.g., EpiGraph [37], MULTOVL [14], EpiExplorer [38] and LOLA [13].

Figure 2

Illustration of the analysis question ,,Which tracks (in a suite) coincide most strongly with a separate single track?" (see Additional file 1). The input to the tool is a single query track (Q) and a set of reference tracks (R1, R2 and R3). The contingency tables show the pairwise overlap between the query track and each of the reference tracks. The Forbes coefficient [26] is calculated from each contingency table, and used to rank the reference tracks according to similarity to the query.

Figure 3

Overview of typical analysis phases and the tools included in the GSuite HyperBrowser system. A set of tools for assembly and customization of track collections (GSuites) lead up to a diverse range of tools for statistical and visual analysis of relations between a multiplicity of tracks.

Figure 4

Schematic illustration of the four defined classes of multiplicity:

A) Trivial multiplicity - a statistical analysis is executed independently per dataset in the collection.

B) Contrasting multiplicity - a statistic is computed per dataset in the collection and the results are interpreted relative to each other, e.g. ranked from highest to lowest.
C) Integrative multiplicity - a statistic is computed on all pairs of datasets in the the collection. Results are aggregated either per dataset or for the collection as a whole.
D) Higher-order multiplicity - a statistic is defined on higher-order relations between datasets (beyond the pairwise level).

Figure 5

Interactive basic mode of operation. Guided analysis starting with a pre-defined catalog of biological questions, leading to a finalized answer in few simple steps.

Table 1: Analytical questions on track collections

Question	Input data	Descriptive results	Hypothesis testing focused on individual tracks	Hypothesis testing focused on full suite	Example of usage
Which tracks (in a suite) are most representative and most atypical for the suite?	A single suite of tracks	Ranking of tracks based on aggregated (*C) co-occurrence against all other tracks of the suite	Is the most representative track of the suite more similar to the rest than one would expect any of the tracks to be representative of by chance? (*A)	Are the tracks in the suite (as a whole) more similar than expected by chance?	Check for outliers in a collection of replicate experimental tracks of DNaseI hypersensitivity
Which tracks (in a suite) coincide most strongly with a separate track?	A single suite of tracks and a single track	Ranking of tracks based on co-occurrence against the separate track	Does a given track from the suite co-occur with the separate track more than one would expect by chance? (*B)	Do the tracks in the suite (as a whole) coincide with the separate track more than expected by chance?	Compare the enrichment of a set of trait-associated SNPs in open chromatin regions of different tissues
Do certain tracks of one suite coincide particularly strongly with certain tracks of another suite?	Two suites of tracks	A heatmap of co-occurrence for all pairwise combinations of tracks from the two suites	Is a track from one suite co-occurring with a track from the second suite more than expected by chance (given the general propensity of each of the two tracks to co-occur with tracks of the other suite)?	Does the distribution of co-occurrence values for pairwise track combinations have more extreme values than it would be expected by chance?	Assess the enrichment of somatic variants of different cancer types in heterochromatin of different cell types
In which regions of the genome do tracks of a suite have the most occurrences?	A single suite of tracks and a set of genome regions to be used as bins	Ranking of bins based on aggregated (*C) coverage by tracks in the bin	Is the aggregated (*C) coverage by tracks in the given bin higher than one would assume from the coverages of different tracks across the genome as a whole?	Is the occurrence of segments for tracks of a suite varying between bins more than expected by chance?	Find genes with particularly high frequency of somatic variants across a set of cancer patients
In which regions of the genome do tracks of a suite exhibit the strongest tendency to co-occur?	A single suite of tracks and a set of genome regions to be used as bins	Ranking of bins based on aggregated (*C) pairwise co-occurrence of all tracks of the suite against each other	Do the segments co-occur more than expected in a given bin (given their general propensity to co-occur across the genome)?	Does the degree of co-occurrence between segments for tracks of a suite vary more between bins than expected by chance?	Find regions of the genome where ChIP-seq peaks representing binding of a set of transcription factors co-occur frequently

Figure 1

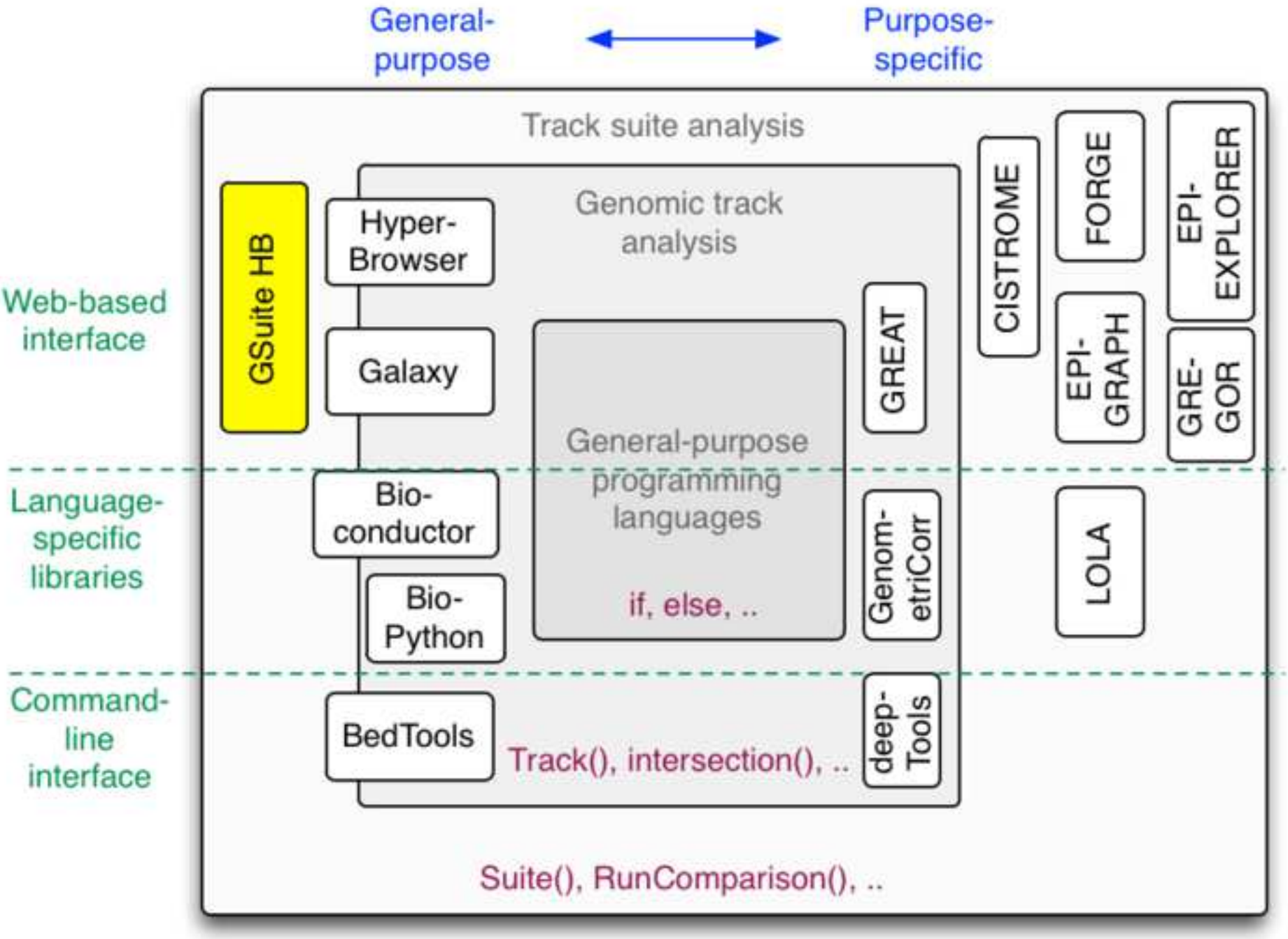

General-purpose
Purpose-specific
Track suite analysis
GSuite HB
Hyper-Browser
Galaxy
Bio-conductor
Bio-Python
BedTools
Genomic track analysis
General-purpose programming languages
if, else, ..
Track(), intersection(), ..
Suite(), RunComparison(), ..
GREAT
Genom-etriCorr
deep-Tools
CISTROME
EPI-GRAPH
LOLA
FORGE
GRE-GOR
EPI-EXPLORER
Web-based interface
Language-specific libraries
Command-line interface

Figure 2

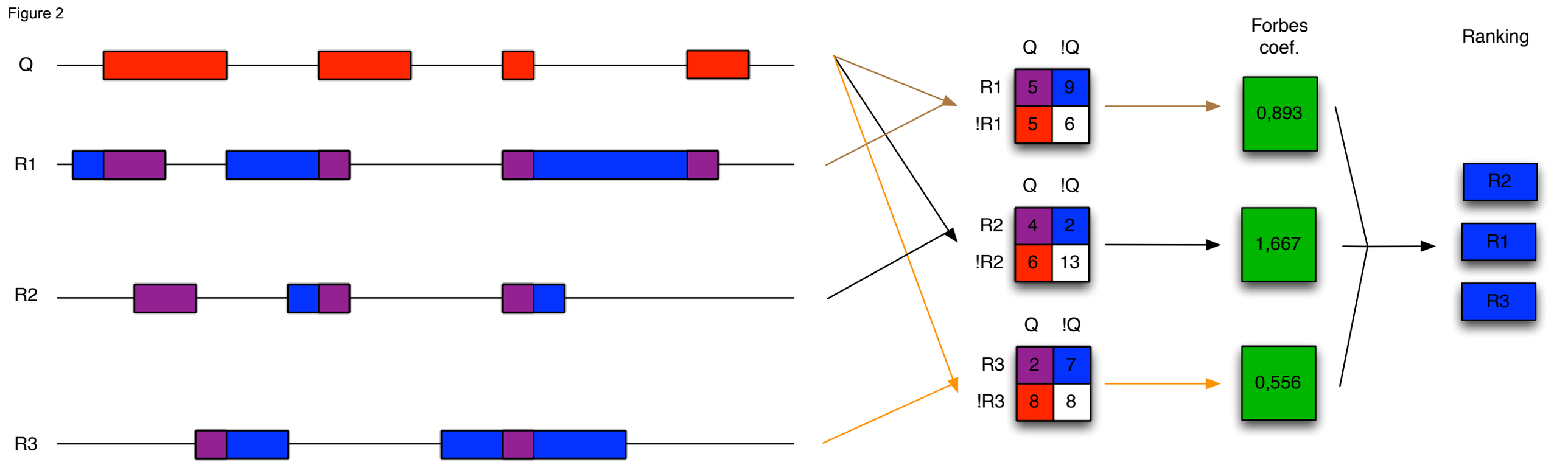

Q
R1
R2
R3
Q !Q
R1 5 9
!R1 5 6
Q !Q
R2 4 2
!R2 6 13
Q !Q
R3 2 7
!R3 8 8
Forbes coef.
0,893
1,667
0,556
Ranking
R2
R1
R3

Figure 3

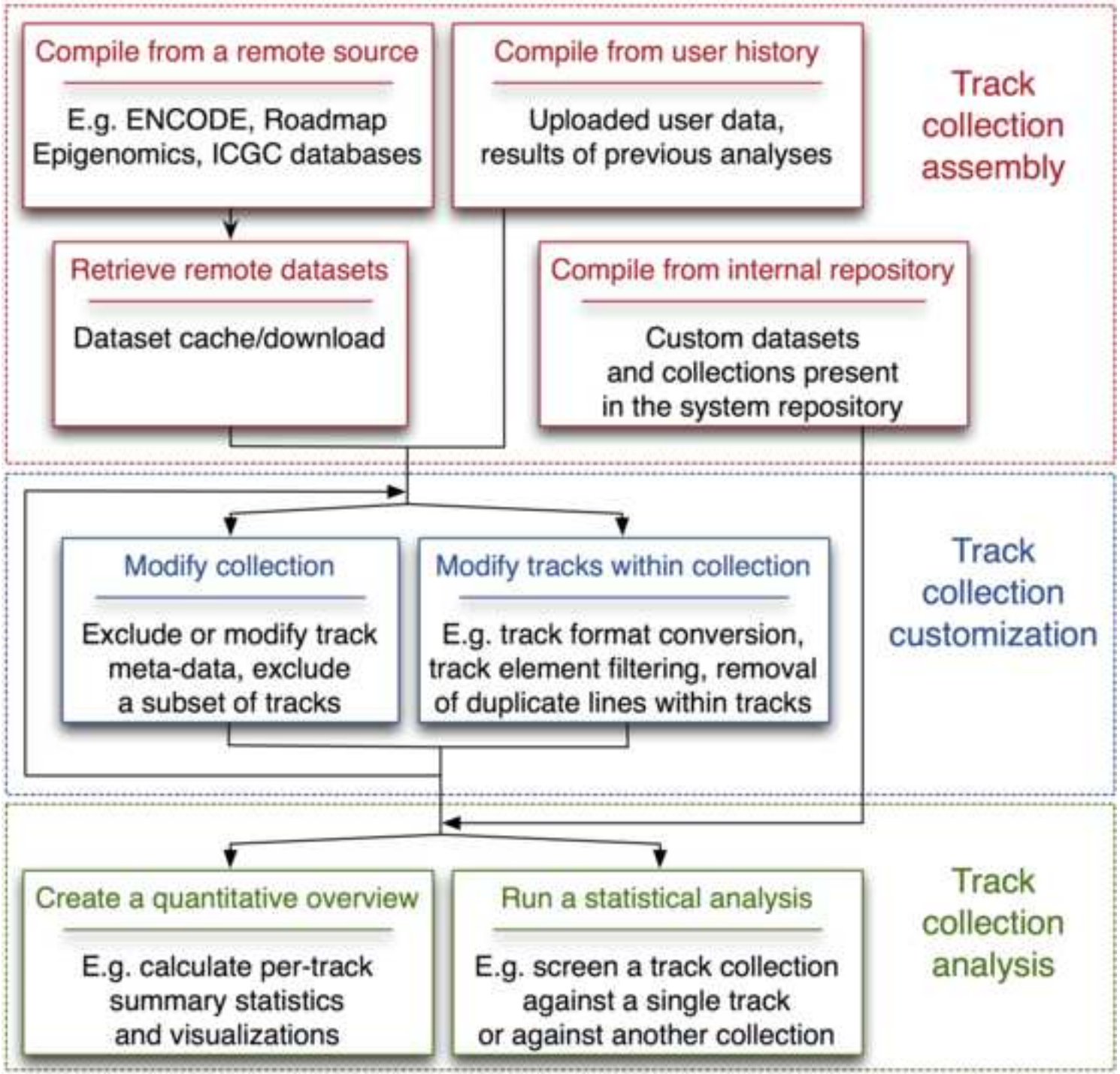

Compile from a remote source
E.g. ENCODE, Roadmap Epigenomics, ICGC databases
Compile from user history
Uploaded user data, results of previous analyses
Track collection assembly
Retrieve remote datasets
Dataset cache/download
Compile from internal repository
Custom datasets and collections present in the system repository
Modify collection
Exclude or modify track meta-data, exclude a subset of tracks
Modify tracks within collection
E.g. track format conversion, track element filtering, removal of duplicate lines within tracks
Track collection customization
Create a quantitative overview
E.g. calculate per-track summary statistics and visualizations
Run a statistical analysis
E.g. screen a track collection against a single track or against another collection
Track collection analysis

Figure 4

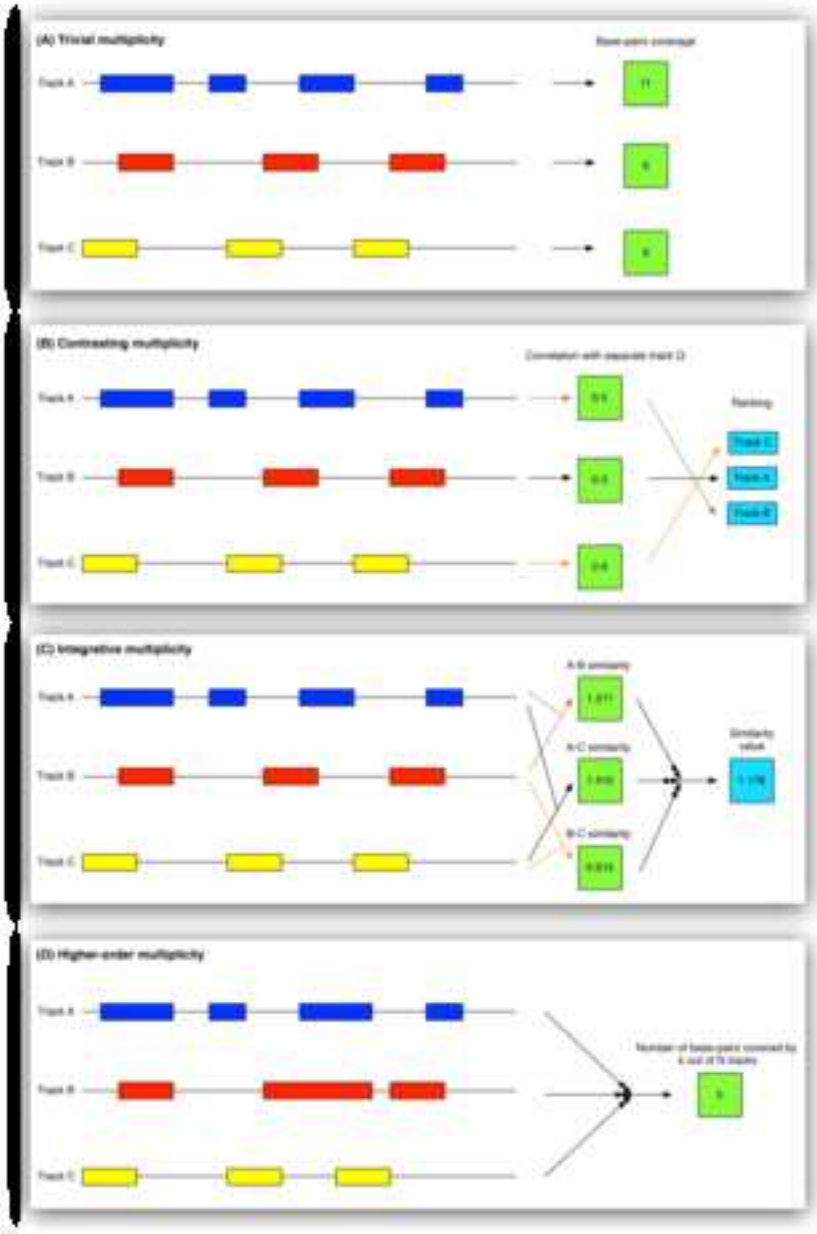
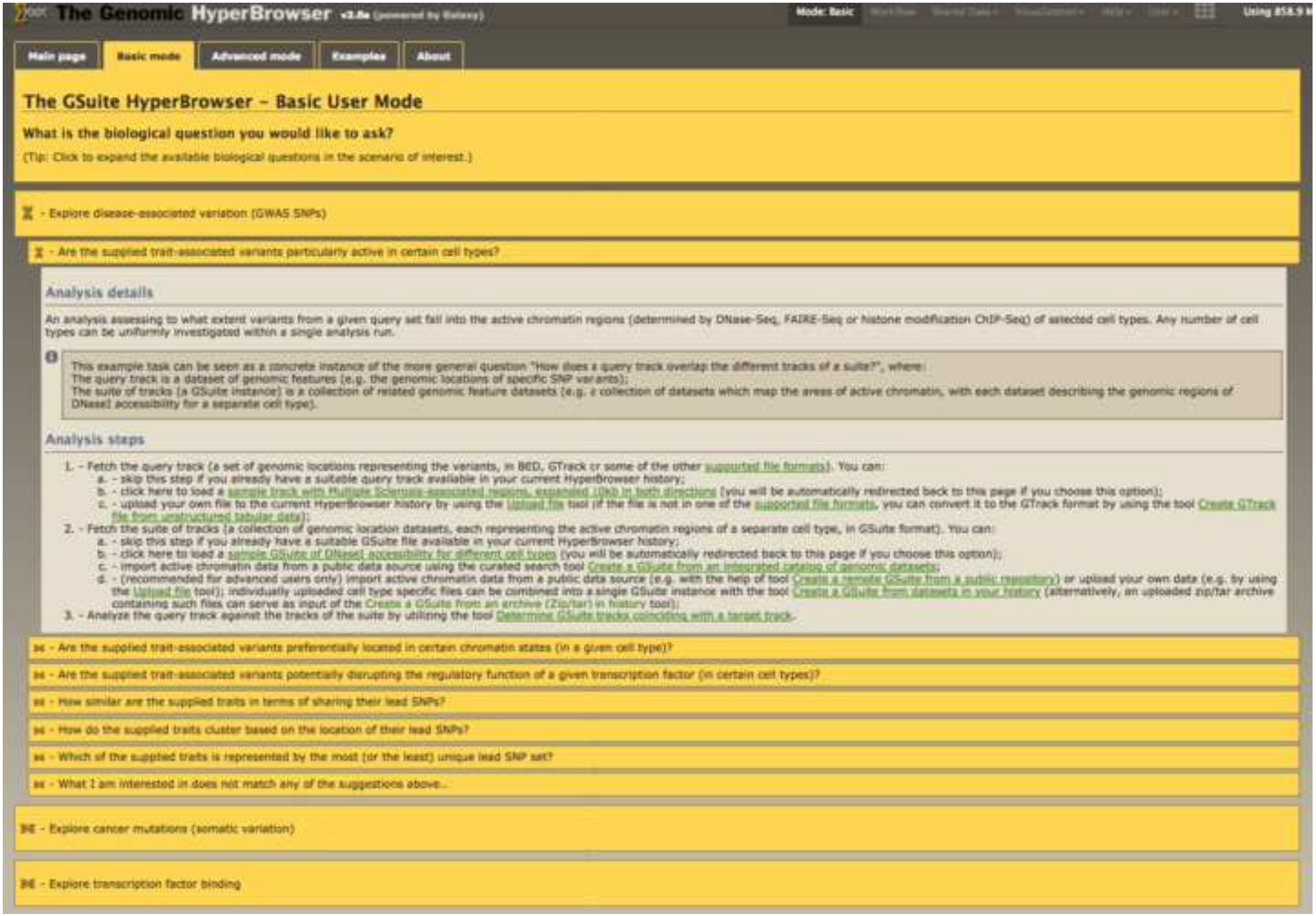

(A) Trivial multiplicity
Track A
Track B
Track C
(B) Contrasting multiplicity
Track A
Track B
Track C
Ranking
(C) Integrative multiplicity
Track A
Track B
Track C
(D) Higher-order multiplicity
Track A
Track B
Track C

Figure 5


The Genomic HyperBrowser
Mode: Basic
Main page
Basic mode
Advanced mode
Examples
About
The GSuite HyperBrowser – Basic User Mode
What is the biological question you would like to ask?
(Tip: Click to expand the available biological questions in the scenario of interest.)
- Explore disease-associated variation (GWAS SNPs)
- Are the supplied trait-associated variants particularly active in certain cell types?
Analysis details
An analysis assessing to what extent variants from a given query set fall into the active chromatin regions (determined by DNase-Seq, FAIRE-Seq or histone modification ChIP-Seq) of selected cell types. Any number of cell types can be uniformly investigated within a single analysis run.
This example task can be seen as a concrete instance of the more general question "How does a query track overlap the different tracks of a suite?", where:
The query track is a dataset of genomic features (e.g. the genomic locations of specific SNP variants);
The suite of tracks (a GSuite instance) is a collection of related genomic feature datasets (e.g. a collection of datasets which map the areas of active chromatin, with each dataset describing the genomic regions of DNase1 accessibility for a separate cell type).
Analysis steps
1. - Fetch the query track (a set of genomic locations representing the variants, in BED, GTrack or some of the other supported file formats). You can:
a. - skip this step if you already have a suitable query track available in your current HyperBrowser history;
b. - click here to load a sample track with Multiple Sclerosis-associated regions, expanded 10kb in both directions (you will be automatically redirected back to this page if you choose this option);
c. - upload your own file to the current HyperBrowser history by using the Upload file tool (if the file is not in one of the supported file formats, you can convert it to the GTrack format by using the tool Create GTrack file from unstructured tabular data);
2. - Fetch the suite of tracks (a collection of genomic location datasets, each representing the active chromatin regions of a separate cell type, in GSuite format). You can:
a. - skip this step if you already have a suitable GSuite file available in your current HyperBrowser history;
b. - click here to load a sample GSuite of DNaseI accessibility for different cell types (you will be automatically redirected back to this page if you choose this option);
c. - import active chromatin data from a public data source using the curated search tool Create a GSuite from an integrated catalog of genomic datasets;
d. - (recommended for advanced users only) import active chromatin data from a public data source (e.g. with the help of tool Create a remote GSuite from a public repository) or upload your own data (e.g. by using the Upload file tool); individually uploaded cell type specific files can be combined into a single GSuite instance with the tool Create a GSuite from datasets in your history (alternatively, an uploaded zip/tar archive containing such files can serve as input of the Create a GSuite from an archive (Zip/tar) in history tool);
3. - Analyze the query track against the tracks of the suite by utilizing the tool Determine GSuite tracks coinciding with a target track.
- Are the supplied trait-associated variants preferentially located in certain chromatin states (in a given cell type)?
- Are the supplied trait-associated variants potentially disrupting the regulatory function of a given transcription factor (in certain cell types)?
- How similar are the supplied traits in terms of sharing their lead SNPs?
- How do the supplied traits cluster based on the location of their lead SNPs?
- Which of the supplied traits is represented by the most (or the least) unique lead SNP set?
- What I am interested in does not match any of the suggestions above...
- Explore cancer mutations (somatic variation)
- Explore transcription factor binding