EpiGRAPH: User-friendly software for statistical analysis and prediction of (epi-) genomic data

Christoph Bock[§], Konstantin Halachev, Joachim Büch and Thomas Lengauer

Max-Planck-Institut für Informatik, Saarbrücken, Germany
[§]Corresponding author, e-mail address: cbock@mpi-inf.mpg.de

Keywords:
Bioinformatics, epigenome, statistics, machine learning, computational epigenetics

Abstract

The EpiGRAPH web service (http://epigraph.mpi-inf.mpg.de/) enables biologists to uncover hidden associations in vertebrate genome and epigenome datasets. Users can upload sets of genomic regions and EpiGRAPH will test a wide range of attributes (including DNA sequence and structure, gene density, chromatin modifications and evolutionary conservation) for enrichment or depletion among these regions. Furthermore, EpiGRAPH learns to predictively identify genomic regions that exhibit similar properties. This paper demonstrates EpiGRAPH's practical utility in a case study on monoallelic gene expression and describes its novel approach to reproducibility and data sharing.

Rationale

EpiGRAPH addresses two tasks that are common in genome biology: (i) discovering novel associations between a set of genomic regions with a specific biological role (e.g. experimentally mapped enhancers, hotspots of epigenetic regulation or sites exhibiting disease-specific alterations) and the bulk of genome annotation data that are available from public databases; and (ii) assessing whether it is possible to predictively identify additional genomic regions with a similar role without the need for further wet-lab experiments.

The increasing relevance of analyzing sets of genomic regions arises from technical innovations such as tiling microarrays and next-generation sequencing (Bock and Lengauer 2008; Mardis 2008; Schones and Zhao 2008; van Steensel 2005), which can be used to scan the genome for specific types of regions (e.g. transcription factor binding sites or cancer-specific genomic alterations). The resulting datasets are difficult to analyze with existing toolkits for genomic data mining – such as GSEA (Subramanian et al. 2007) and DAVID (Huang et al. 2007) – because most existing tools are gene-centric and cannot easily account for genomic regions that are located outside of genes. In the absence of a suitable tool for statistical analysis and prediction of genomic region data, researchers have performed the necessary steps by hand, downloading relevant datasets from existing repositories and writing one-time use scripts for data integration, statistical analysis and prediction (e.g. Allen et al. 2003; Berry et al. 2006; Bock et al. 2006; Cohen et al. 2006; Das et al. 2006; Derti et al. 2006; Fang et al. 2006; Guelen et al. 2008; Luedi et al. 2007; Luedi et al. 2005; Montgomery et al. 2007; Wang et al. 2006). Such manual analyses are time-consuming to perform, difficult to reproduce and require bioinformatic skills that are beyond the reach of most biologists. Hence, these studies support demand for a software toolkit that facilitates statistical analysis and prediction of region-based genome and epigenome data.

With the development of EpiGRAPH, we have pulled together our experiences and established workflows from several studies (Bock et al. 2006; Bock et al. 2007; Bock et al. 2008; Liu et al. 2007; Moser et al. 2008) and incorporated them into a powerful and easy-to-use web service. In the remainder of this paper, we outline the basic concepts of EpiGRAPH, demonstrate its practical use and utility in a case study on monoallelic gene expression and outline how the UCSC Genome Browser (Karolchik et al. 2008), Galaxy (Blankenberg et al. 2007;

Giardine et al. 2005) and EpiGRAPH integrate into a comprehensive pipeline for (epi-) genome analysis and prediction. Finally, the Methods section provides extensive bioinformatic background on EpiGRAPH's software architecture and describes how the software can be extended and customized. This paper is supplemented by a step-by-step, tutorial-style description of two example analyses (http://epigraph.mpi-inf.mpg.de/documentation/EpiGRAPH_tutorial.pdf) and by three tutorial videos that demonstrate EpiGRAPH "in action" (http://epigraph.mpi-inf.mpg.de/videos/).

Concept

EpiGRAPH (http://epigraph.mpi-inf.mpg.de/) is designed to facilitate complex bioinformatic analyses of genome and epigenome datasets. Such datasets frequently consist of sets of genomic regions that share certain properties, e.g. being bound by a specific transcription factor or exhibiting characteristic patterns of evolutionary conservation. Typically, these genomic regions fall into opposing classes, e.g. transcription factor bound vs. unbound promoter regions or significantly conserved vs. nonconserved regulatory elements. Even when this convenient situation does not emerge by default, it is straightforward and common practice to establish it artificially, by generating a randomized set of control regions to complement a given set of genomic regions. EpiGRAPH thus focuses on the analysis of sets of genomic regions that fall into two classes, which we denote as "positives" (cases) and "negatives" (controls).

EpiGRAPH provides four analytical modules (see Figures 1, 2 and 3 for screenshots of illustrative results and Figure 4 for an overview of EpiGRAPH's software architecture). (i) The statistical analysis module identifies attributes that differ significantly between the sets of positives and negatives, based on a large attribute database comprising a broad range of genome and epigenome datasets. (ii) The diagram generation module draws boxplots that visualize the distribution of a selected attribute among the sets of positives vs. negatives. (iii) The machine learning analysis module evaluates how well prediction algorithms – such as support vector machines – can discriminate between positives and negatives in the input dataset, based on different combinations of prediction attributes. (iv) The prediction analysis module predicts whether a genomic region that is not contained in the input dataset belongs to the set of positives or negatives, thus exploiting any correlations detected by the machine learning analysis module for the prediction of new data.

Typical EpiGRAPH analyses follow a pre-defined workflow. The starting point is a dataset of genomic regions, which the user may have obtained through wet-lab analysis (e.g. ChIP-seq analysis of transcription factor binding) or bioinformatic calculations (e.g. computational screening for regions that are under evolutionary constraint). This dataset is uploaded to the EpiGRAPH web service, as a table of genomic regions with separate columns for chromosome name, start position, end position, and a binary class value specifying for each region whether it belongs to the positives or negatives. (When no class value is provided, EpiGRAPH regards all genomic regions of the input dataset as positives and assists the user with calculating a set of random control regions to be used as negatives.) Next, EpiGRAPH calculates a large number of potentially relevant attributes for each genomic region in the input dataset. Most of these attributes are overlap frequencies or score values, quantifying the co-localization of the genomic regions in the input dataset with publicly available annotation data for the respective genome. Upon completion of the attribute calculation (which can take several hours or even days when the input dataset is large), EpiGRAPH's statistical and machine learning modules calculate an initial assessment of significant differences between the positives and negatives in the input dataset and an assessment of whether or not these differences are sufficient for bioinformatic prediction. After inspection of the initial results, the user can specify follow-up analyses based on the pre-calculated attributes. In particular, the diagram generation module can be used to visualize the most interesting differences between positives and negatives as detected by the statistical analysis, and the prediction analysis module lets the user predict the class value of new genomic regions, for example in order to extrapolate experimental data to regions that were not covered by wet-lab experiments.

The key to EpiGRAPH's practical utility is its database, for which we collected a large number of attributes that are likely to play a role in genome function and epigenetic regulation. For the most thoroughly annotated

human genome, EpiGRAPH currently includes almost a thousand attributes (see Table 1 for an overview and http://epigraph.mpi-inf.mpg.de/attributes/ for details), which fall into ten attribute groups: (i) DNA sequence, (ii) DNA structure, (iii) repetitive DNA, (iv) chromosome organization, (v) evolutionary history, (vi) population variation, (vii) genes, (viii) regulatory regions, (ix) transcriptome and (x) epigenome and chromatin structure. EpiGRAPH also incorporates the genomes of chimp, mouse and chicken (with slightly lower numbers of attributes) and can easily be extended to support genomes of other species. In addition to using EpiGRAPH's default attributes, researchers can upload their own datasets and incorporate them as custom attributes in subsequent analyses. Adding relevant experimental data as custom attributes frequently improves the prediction accuracy that EpiGRAPH achieves for a given biological phenomenon.

Application

The best starting point for getting acquainted with the practical use of EpiGRAPH are the tutorial videos at http://epigraph.mpi-inf.mpg.de/videos/ and the step-by-step guide available from http://epigraph.mpi-inf.mpg.de/documentation/EpiGRAPH_tutorial.pdf. In the following case study, we take a slightly more high-level view, focusing on how to plan and interpret an EpiGRAPH analysis and highlighting potential sources of misinterpretation. Nevertheless, all raw data, settings and results of this case study are available online (http://epigraph.mpi-inf.mpg.de/casestudy/), and readers are encouraged to download the analysis description file, upload it into their own EpiGRAPH accounts, reproduce the results and perform follow-up analyses.

Monoallelic gene expression – the focus of our case study – is a common phenomenon in vertebrate genomes. While the majority of human genes are expressed from both alleles, a sizable proportion is expressed exclusively from a single allele, with important biological consequences. Genomic imprinting, i.e. parent-specific monoallelic gene expression, plays a critical role in normal development and gives rise to non-Mendelian patterns of inheritance (Reik 2007). X-chromosome inactivation leads to mitotically heritable silencing of the surplus X chromosome in females (Heard 2004). And random monoallelic gene expression, which is common among odorant receptor genes and immune-system related genes, increases the phenotypic diversity among clonal cells (Gimelbrant et al. 2007).

In an attempt to identify potential determinants of monoallelic gene expression, several bioinformatic studies analyzed the DNA sequence in the vicinity of known monoallelically expressed genes, comparing its properties with control regions from biallelically expressed genes. These studies reproducibly found enrichment of LINE repeats and depletion of SINE repeats to be associated with monoallelic gene expression (Allen et al. 2003; Bailey et al. 2000; Greally 2002; Ke et al. 2002). Encouraged by these results, attempts have been made to predict – based on the genomic DNA sequence – which genes are subject to imprinting and X-chromosome inactivation (Luedi et al. 2007; Luedi et al. 2005; Wang et al. 2006). However, the conclusiveness of these prior studies is somewhat diminished by the fact that most of them relied on small gene lists curated from the literature and that none took epigenetic data into account.

Here, we revisit the relationship between DNA characteristics and monoallelic gene expression based on genome-scale datasets, including a recent assessment of monoallelic vs. biallelic gene expression for 4,000 genes in human lymphoblastic cells (Gimelbrant et al. 2007) and extensive epigenome maps of human T-cell lymphocytes (Barski et al. 2007). To start with, we obtain lists of monoallelically and biallelically expressed genes from the supplementary material of the corresponding paper (Gimelbrant et al. 2007), and we use Galaxy (http://main.g2.bx.psu.edu/) to retrieve the chromosomal location of these genes. For simplicity, we merge the two lists into a single EpiGRAPH input dataset, which contains all 464 positives (monoallelically expressed genes) as well as an equally sized sample of negatives (biallelically expressed genes). Random down-sampling of the set of negatives is performed in order to limit bias toward predicting the majority class, which is a common issue in machine learning. In general, we recommend that the number of positives should never exceed twice the number of negatives, and vice versa. EpiGRAPH automatically enforces this upper limit for the class imbalance, unless the user deselects the corresponding option.

Before we can submit our dataset to EpiGRAPH, we have to decide exactly which regions we want to analyze, i.e. whether we expect DNA signals relating to monoallelic gene expression distributed across the entire gene or preferentially located in specific regions, such as promoters, exons or introns. Since monoallelic gene expression is a consequence of specific transcriptional processes, we believe that promoter regions have the highest probability of containing relevant regulatory elements. For each gene, we thus select a main window from 1,250 basepairs upstream to 250 basepairs downstream of the annotated transcription start site as the region of interest (the necessary calculation is easily performed within any spreadsheet software). However, as we cannot exclude that important regulatory elements might be located further upstream or downstream, we activate EpiGRAPH's option to cover four additional sequence windows ranging from -50 kilobases to +50 kilobases around the region of interest.

Next, we have to decide which groups of attributes from EpiGRAPH's database should be included in our analysis. While it is always possible to perform hypothesis-free screening by selecting all default attributes, focusing the analysis only on promising attribute groups can significantly increase statistical power and also decreases computation time. Based on prior knowledge, we choose four attribute groups that are likely to be related to monoallelic gene expression, namely "repetitive DNA", "regulatory regions", "transcriptome", and "epigenome and chromatin structure".

Having made all relevant decisions, we can now start the analysis, log out of the web service and wait for EpiGRAPH to perform the necessary calculations. Assuming that e-mail notification has been enabled, EpiGRAPH will inform us as soon as it has completed an initial analysis based on default parameters. At that point, we can log into the web service again, review the results and define follow-up analyses.

Our inspection of the results starts with the statistical analysis table (Figure 1). This table summarizes pairwise statistical comparisons between positives and negatives, which were performed for each attribute using the Wilcoxon test (for numerical attributes) and Fisher's exact test (for categorical attributes). Focusing on the 1.5 kilobase core promoter region (the main window of our analysis), a total of 72 out of 563 attributes differ significantly between monoallelically vs. biallelically expressed genes, at a false discovery rate of 5%. Furthermore, similar but weaker differences are observed for four additional sequence windows upstream and downstream of the promoter region (data not shown), indicating that the contrasting genomic properties of monoallelically vs. biallelically expressed genes are strong for the core promoter, but also present in a wider genomic region surrounding the genes.

In their core promoter regions, biallelically expressed genes on average exhibit twice the amount of histone H3 lysine 4 trimethylation (which is indicative of open chromatin) as the promoters of monoallelically expressed genes. Conversely, the latter are almost threefold enriched in terms of repressive histone H3 lysine 27 trimethylation. Consistent with the conclusion that promoters of monoallelically expressed genes generally exhibit a more repressed chromatin state than their biallelic counterparts, we also observe significant under-representation of their associated transcripts in EST libraries and decreased expression according to microarray data (Figure 1). Interestingly, out of the 28 tissues covered by EpiGRAPH, the difference in gene expression is most significant for thymus, consistent with the fact that monoallelic gene expression is prominent among genes related to the immune system.

To illustrate the distinct chromatin structure at the core promoters of monoallelically vs. biallelically expressed genes, we select H3 lysine 4 trimethylation and H3 lysine 27 trimethylation for visualization using EpiGRAPH's diagram generation module (Figure 2). Boxplots confirm that the differences are not only significant, but also substantial in quantitative terms. This confirmation is an important first step toward establishing the biological relevance of our finding, given that even minor and biologically irrelevant differences can become highly significant when sample sizes are large. In general, to demonstrate both significance and strength of an observed difference, we recommend that EpiGRAPH users should report not only *P*-values, but also the corresponding boxplot diagrams or at least the attributes' mean values for the sets of positives and negatives, respectively.

Further support for a strong association between (repressive) chromatin structure and monoallelic gene expression comes from EpiGRAPH's machine learning analysis. Based on the values of 83 chromatin-related

attributes measured across the core promoter regions and four adjacent windows (415 variables in total), EpiGRAPH could predict with an accuracy of 73.8% (sensitivity: 73.4%; specificity 74.2%; correlation: 0.47) whether a gene is monoallelically or biallelically expressed (Figure 3A). Substantially lower prediction performance was observed for the other attribute groups, namely repetitive DNA (accuracy: 58.3%; correlation: 0.17), regulatory regions (accuracy: 51.2%; correlation: 0.03) and the transcriptome (accuracy: 66.5%; correlation: 0.33). We thus conclude that attributes relating to epigenome and chromatin structure are the strongest predictors of monoallelic gene expression. Importantly, all measures of prediction performance reported by EpiGRAPH are calculated exclusively based on test set results in a cross-validation design, thereby minimizing the risk of overtraining and irreproducibly optimistic performance evaluations that is inherent in the use of machine learning methods (Tarca et al. 2007).

Due to the complex structure and organization of mammalian genomes, the attribute groups included in our analysis are not statistically independent. On the contrary, strong biological interdependencies exist between attribute groups, for example between chromatin structure and the transcriptome (open chromatin structure facilitates transcription), between regulatory regions and repetitive DNA (evolutionary conserved regulatory regions are largely absent from repetitive regions), and between repetitive DNA and chromatin structure (repetitive regions most commonly exhibit repressive chromatin structure). Therefore, the predictiveness of some attribute groups included in our analysis could be indirect and mediated by their correlation with other, more predictive, attributes. EpiGRAPH helps us better understand such relationships, by measuring whether any combination of two or more attribute groups gives rise to higher prediction performance than each attribute group on its own right (which indicates that all attribute groups contribute to the overall prediction performance) or whether a single attribute group dominates the other attribute groups (in which case the other attribute groups are likely to "borrow" predictiveness from the former, rather than being independently predictive). To perform such an analysis, we restart the machine learning analysis with custom settings, requesting EpiGRAPH to include all possible combinations of attribute groups while focusing on the core promoter regions. The results table reports prediction performances separately for linear support vectors trained on each of the 15 possible combinations of attribute groups (Figure 3B). These data clearly indicate that a single attribute group – epigenome and chromatin structure – is more predictive than all others. In fact, there is no evidence of complementarity for any combination of attribute groups (i.e. no set of attribute groups outperforms the single highest-scoring attribute group contained in the set). In the light of these results, it seems unlikely that repetitive elements are directly causal for monoallelic gene expression at a genomic scale. Rather, the predictiveness of specific repetitive elements observed in prior studies as well as in this analysis appears to be largely due to the fact that certain types of repeats (such as LINE elements) are enriched in regions that exhibit repressive chromatin structure, while other types of repeats (such as SINE elements) are depleted in these regions.

In a final step, we want to use EpiGRAPH to predict for all genes in the human genome whether their tendency is toward monoallelic or biallelic gene expression. To that end, we first verify that a linear support vector machine (EpiGRAPH's default prediction algorithm) indeed provides competitive prediction performance when compared to other machine learning algorithms. Such benchmarking is achieved by restarting the machine learning analysis with custom settings, selecting all available machine learning algorithms for inclusion (Figure 3C). EpiGRAPH's cross-validation results indicate that the linear support vector machine performs on par with the best method, an ensemble learning algorithm (AdaBoost on tree stumps). We thus conclude that a linear support vector machine trained on epigenome and chromatin structure data provides a suitable setup for genome-wide prediction of monoallelic gene expression. Next, we obtain a list of RefSeq-annotated genes from the UCSC Genome Browser, determine 1.5 kilobase core promoter regions for all genes and submit this dataset to EpiGRAPH's prediction analysis, specifying the prediction setup as described. Upon submission of the analysis, EpiGRAPH starts to calculate the relevant attributes and predicts the expression status of all 25,419 RefSeq-annotated genes in the human genome. The results, which are available from http://epigraph.mpi-inf.mpg.de/casestudy/, provide a first genome-wide prediction of monoallelic gene expression in the human genome. Although the accuracy of our predictions is far from perfect (cf. Figure 3C) and further experimental

analysis is clearly warranted, these predictions could be useful for identifying new candidate genes that contribute to the many biological roles of monoallelic gene expression.

In summary, this case study illustrates how EpiGRAPH can be applied to analyzing a genomic feature of interest (here: monoallelic gene expression) in the context of publicly available genome annotations and epigenome data. Two main conclusions emerge from our analysis. First, monoallelically expressed genes exhibit a substantially more repressed chromatin structure in their promoter regions than biallelically expressed genes. This observation is consistent with a model in which monoallelic gene expression is the direct consequence of opposing chromatin states at the two alleles of a gene within a diploid cell. Indeed, Wen et al. recently showed that an experimental search for genomic regions that exhibit activating as well as repressive chromatin marks can identify monoallelically expressed genes (Wen et al. 2008). Second, chromatin structure clearly emerges as the strongest predictor of monoallelic gene expression, as compared to the overall level of gene expression or the enrichment/depletion of specific types of repeats and regulatory regions. In fact, none of the other attribute groups included in our analysis could increase prediction performance after chromatin structure had been accounted for. This observation is not necessarily in contradiction with an (indirectly) causal model in which local enrichment of LINE elements fosters repressive chromatin structure, which in turn facilitates random silencing of a single allele. However, the weak predictiveness of attributes relating to repetitive DNA suggests that such a model omits important additional drivers of monoallelic gene expression.

Integration

EpiGRAPH integrates well with existing bioinformatics resources and infrastructure. It can be regarded as part of a three-step data analysis pipeline involving genome browsers, genome calculators and genome data analysis tools (Figure 5): (i) Researchers typically start the analysis of new genome-scale datasets by uploading preprocessed and quality-controlled data into a genome browser, which facilitates data visualization and manual inspection. The UCSC Genome Browser (Karolchik et al. 2008) is popular for this task, due to the ease with which custom data tracks can be displayed alongside public genome annotations, while Ensembl is an alternative option (Flicek et al. 2008). (ii) Based on initial observations, it is usually necessary to pick a subset of genomic regions for further analysis, e.g. all promoter regions that are bound by a specific transcription factor. The Galaxy web service (Blankenberg et al. 2007; Giardine et al. 2005) is specifically designed to perform the necessary calculations and filtering, in order to select biologically interesting regions for further analysis. (iii) Finally, it is often desirable to perform statistical analysis and data mining on the potentially large set of interesting regions, in order to discover, test and interpret correlations with other genomic data. For this step, a comprehensive and easy-to-use toolkit has been lacking. We developed EpiGRAPH to fill this gap, thereby enabling biologists to perform advanced bioinformatic analysis and prediction with little need for bioinformatic support. We demonstrate the interplay of UCSC Genome Browser, Galaxy and EpiGRAPH in a case study focusing on the (epi-) genomic characteristics of highly polymorphic promoter regions in the human genome (see http://epigraph.mpi-inf.mpg.de/documentation/EpiGRAPH_tutorial.pdf and video tutorial 4 available from http://epigraph.mpi-inf.mpg.de/videos/).

In the future, we anticipate that the three layers of genome browsing, calculation and analysis tools will increasingly merge into a single application, for which “statistical genome browser” might be an appropriate term. To that end, it will be neither necessary nor beneficial to integrate all functionality and underlying databases into a single monolithic tool. Instead, a distributed network of interoperable web services for genome analysis is likely to emerge. Genome browsers could act as single points of entry, from which the user initiates a complex analysis. The analysis is then split into separate subtask, encoded in an XML-based analysis description language (such as the X-GRAF format prototyped in EpiGRAPH) and distributed over the Internet to calculation servers at which all relevant datasets and software components for a specific type of analysis are available. Finally, the decentrally calculated results are merged and displayed to the user at the central genome browser frontend. EpiGRAPH was developed with this scenario in mind and prototypes software paradigms required for distributed genome analysis by concerted action of specialized tools.

Conclusion

The EpiGRAPH web service enables biologists to perform complex bioinformatic analyses online – without having to learn a programming language or to download and manually process large datasets. Compared to related tools such as Galaxy (Blankenberg et al. 2007; Giardine et al. 2005) and Taverna (Hull et al. 2006; Oinn et al. 2004), its main emphasis lies on exploratory statistical analysis, hypothesis generation and bioinformatic prediction, based on large datasets of genomic regions. EpiGRAPH facilitates reproducibility and data sharing by encoding all analyses in standardized analysis description files that can be re-run by other users. We highlighted EpiGRAPH's utility by a case study on monoallelic gene expression, and we provide extensive additional material online (including tutorial videos at http://epigraph.mpi-inf.mpg.de/videos/ and a step-by-step guide available from http://epigraph.mpi-inf.mpg.de/documentation/EpiGRAPH_tutorial.pdf).

Methods

EpiGRAPH software architecture and analysis workflow

The key design decision underlying EpiGRAPH's software architecture is to store each EpiGRAPH analysis in a single XML file. This XML file contains not only a detailed specification of the analysis and its supplementary attributes, but also its current processing status and, upon completion, its results. All XML files processed by EpiGRAPH conform to the standardized X-GRAF format (discussed in more detail below) and are stored in an XML database.

EpiGRAPH's XML-based, analysis-centric design offers a number of advantages over alternative architectures: (i) *Reproducibility*: All information relevant to an analysis, including its specifications and results, are bundled in a single file, which provides a complete documentation of the analysis. The same analysis can be re-run at any time simply by uploading its XML file back into the EpiGRAPH web service. (ii) *Parallel processing*: Because the different analysis modules operate on different parts of the XML tree, they can work in parallel without generating write-write conflicts. (iii) *Interoperability and error checking*: The use of XML files facilitates data exchange with other software systems, and the X-GRAF format provides error checking when XML files are constructed manually or exchanged between different software systems.

Internally, the EpiGRAPH web service consists of three software components and two logical databases (Figure 4). (i) The *web-based frontend* provides convenient access to EpiGRAPH's functionality over the Internet. The frontend is implemented in Java (http://www.java.com/), utilizing the JavaServer Faces framework for its user interface and Java servlets as well as JavaServer Pages for operating as a web application. (ii) The *process control middleware* provides a single point of access to the analyses and custom attributes stored in the XML database and it enforces compliance with the X-GRAF XML format. The middleware is implemented as a Java servlet and makes its services available via XML-RPC (http://www.xmlrpc.com/). (iii) The *analysis calculation backend* performs all attribute calculations and bioinformatic analyses required to execute an EpiGRAPH analysis request. It submits its results to the middleware, which stores them in the XML database. The backend is implemented in Python (http://www.python.org/), using the R package (http://www.r-project.org/) for statistical analysis and diagram generation, and the Weka package (http://www.cs.waikato.ac.nz/~ml/weka/) for machine learning and prediction analysis. (iv) The *relational database* stores EpiGRAPH's default attributes. Oracle Database 11g (http://www.oracle.com/database/) is used with pre-calculated indices in order to achieve high-performance database retrieval. (v) The *XML database* provides central storage of all XML files and enables parallelized access to the XML files as a whole as well as to specific subnodes. EpiGRAPH makes use of Oracle XML DB (http://www.oracle.com/technology/tech/xml/xmldb/index.html), which is an XML database extension of the Oracle database. Technically, Oracle XML DB decomposes all XML files into relational database tables, based on the X-GRAF schema definition and object-relational mapping. Hence, while the relational database and the XML database behind EpiGRAPH are logically distinct and used for different types of data (de-

fault attributes vs. analysis requests and custom attributes), both types of data are ultimately stored in the same database management system.

Importantly, the choice of technologies for each component reflects the specific requirements of the tasks it performs. The frontend has to provide a user-friendly interface in a variety of web browsers, which is best achieved using a web application framework such as JavaServer Faces. The middleware makes connections with the XML database and performs extensive XML processing, hence the use of Java with its well-established libraries for Oracle XML DB access (http://www.oracle.com/technology/tech/xml/xmldb/index.html), StAX (http://jcp.org/en/jsr/detail?id=173) and JAXB processing (https://jaxb.dev.java.net/) is an appropriate choice. The backend implements most of EpiGRAPH's application logic and is likely to be extended by other researchers, therefore Python (http://www.python.org/) was selected due to its proven track record for fast and robust software engineering in scientific applications, its platform independence and its wide acceptance within the bioinformatics community.

The internal workflow of an EpiGRAPH analysis is depicted in Figure 4, illustrating how the different components interact when fulfilling an EpiGRAPH analysis request.

Genomes, annotations and attributes included in EpiGRAPH

EpiGRAPH currently supports five genome assemblies from four species: (i) *hg18*: the latest assembly of the human genome (NCBI36.1); (ii) *hg17*: the genome assembly used for the ENCODE project pilot phase (NCBI35); (iii) *mm9*: the latest assembly of the mouse genome (NCBI37); (iv) *panTro2*: the latest assembly of the chimp genome; (v) *galGal3*: the latest assembly of the chicken genome. For each of these genomes, we manually selected a large number of genomic attributes that are likely to be predictive of interesting genomic phenomena (see Table 1 and http://epigraph.mpi-inf.mpg.de/attributes/ for details). When calculated for a specific genomic region, most of these attributes take the form of overlap frequencies (e.g. how many exons overlap with the genomic region?), overlap lengths (e.g. how many basepairs of exonic DNA overlap with the genomic region?) or DNA sequence pattern frequencies (e.g. how many times does the pattern "TATA" occur in the genomic region?). All of these attributes are standardized to a default region size of one kilobase in order to be comparable between genomic regions of different size. In addition, EpiGRAPH uses score attributes, which are averaged across all overlapping regions of a specific type (e.g. what is the average exon number of all genes overlapping with the genomic region?), and category attributes, which split up an attribute into subattributes (e.g. how many coding vs. non-coding SNPs overlap with the genomic region?).

The data for most of these attributes were collected from annotation tracks of the UCSC Genome Browser (Karolchik et al. 2008), using an automated data retrieval pipeline. In addition, published genomic datasets that appear to be of particular interest are imported into the database on a regular basis. Currently, this includes data on histone modifications (Barski et al. 2007), DNA methylation (Meissner et al. 2008; Rollins et al. 2006), regulatory CpG islands (Bock et al. 2007), DNA helix structure (Gardiner et al. 2003), DNA solvent accessibility (Greenbaum et al. 2007), tissue-specific gene expression (Su et al. 2004), isochores (Costantini et al. 2006) and transcription initiation events (Carninci et al. 2006). Finally, users can upload custom datasets into the database, which makes them available for inclusion in further analyses by the same user.

Attribute calculation

The basic functionality of EpiGRAPH's attribute calculation module is to calculate a large number of genomic attributes (such as frequency and length of overlap with EpiGRAPH's default attributes) for any set of genomic regions submitted to the web service. This step is a prerequisite for all further analyses, and it is typically the most computationally intensive and time-consuming part of an EpiGRAPH analysis. The attribute calculation makes extensive use of multithreading in order to increase performance.

Beyond its core task of deriving hundreds or even thousands of different attribute values for each genomic region in the input dataset, the attribute calculation module provides three additional features that add to its utility as a general genome calculator. First, the user can define derived attributes, thus augmenting genomic attributes that are already contained in the database (e.g. deriving a set of putative promoter regions from a gene

attribute). Second, random control regions can be calculated such that they match a given set of genomic regions in terms of chromosome and length distribution, GC content, repeat content and/or exon overlap. Technically, this is achieved by repeatedly sampling random genomic regions of given length from a specific chromosome and retaining a region only if its GC content, repeat content and/or exon overlap are within a user-specified interval around the corresponding value of the source region. Third, attributes can be calculated not only for the genomic regions provided in the input dataset, but also for fixed sequence windows left and right of these regions, in order to capture significant differences in the upstream or downstream neighborhood of a specific set of genomic regions. All results calculated by the attribute calculation module can be used as basis for further EpiGRAPH analyses or downloaded in tab-separated value format for analysis outside EpiGRAPH.

Statistical analysis and diagram generation

Two of EpiGRAPH's four analytical modules – statistical analysis and diagram generation – help the user identify individual attributes that differ between two sets of genomic regions, which we denote as "positives" and "negatives". The statistical analysis module calculates pairwise statistical tests between the positives and negatives, separately for each genomic attribute. The nonparametric Wilcoxon rank-sum test is used for numeric attributes and Fisher's exact test is used for discrete attributes. *P*-values are adjusted for multiple testing by the highly conservative Bonferroni method, which controls the family-wise error rate, and by a more recent and usually preferred method that controls the false discovery rate (Benjamini and Hochberg 1995). While EpiGRAPH applies an overall significance threshold of 5% by default, the user is free to select different values. If multiple windows around the genomic regions of interest are taken into account and tested simultaneously, the user can specify weights to control how the *P*-value threshold is distributed when testing for significant attributes in each of these windows. A typical choice is to use a relatively high *P*-value of, say, 3% for the central window (i.e. the regions provided by the input dataset), while distributing the remaining 2% equally among the upstream and downstream windows. This way, the additional testing for strong effects in the upstream and downstream neighborhoods comes at the cost of only a limited decrease in statistical power for the genomic regions of interest.

While the statistical analysis module focuses on the question whether or not a specific attribute differs significantly between the sets of positives and negatives, the diagram generation module can help assess the effect size, i.e. the quantitative difference between positives and negatives. For any selected attribute, this module derives boxplots contrasting the attribute's distribution among the positives with that among the negatives.

Machine learning analysis and prediction analysis

In contrast to the statistical analysis module, which focuses on individual attributes, the machine learning analysis module assesses how well attribute groups collectively differentiate between the sets of positives and negatives. We treat this question as a machine learning task, predicting for each genomic region whether it is likely to belong to the set of positives or to the set of negatives and interpreting the prediction performance achieved for a specific attribute group as a measure of how well this group discriminates between positives and negatives.

Technically, a machine learning algorithm (e.g. a support vector machine) is repeatedly trained and tested on partitions of the training dataset following a four-step procedure (all parameters mentioned below are default values and can be changed by the user): (i) If the set of positives contains more than twice as many genomic regions as the set of negatives (or vice versa), the larger set is randomly downsampled such that the class imbalance never exceeds 67% vs. 33%, thus limiting potential prediction bias toward the majority class. (ii) Using 10-fold cross-validation, the machine learning algorithm is repeatedly trained on 90% of the genomic regions and tested on the remaining 10%. (iii) Cross-validation is repeated ten times with random partition assignments. (iv) The overall prediction performance is measured by the correlation coefficient between the predictions and the correct values on the cross-validation test sets, as well as by the corresponding values for percent accuracy, sensitivity and specificity, averaged over all cross-validation runs.

During prediction analysis, a machine learning algorithm is trained as described above, but now on a bootstrapped sample drawn from the entire training dataset (downsampling is used if necessary to enforce a maxi-

mum class imbalance of 67% vs. 33%). The trained prediction model is then applied to predict the likelihood of belonging to the set of positives for all genomic regions in a user-supplied set of target regions. The resulting quantitative prediction for each region can assume values between zero and one, with a value of zero corresponding to a high-confidence negative prediction, a value of 0.5 to a borderline case, and a value of one to a high-confidence positive prediction. This process is repeated ten times with different bootstrapped samples in order to obtain an additional criterion for the reliability of the predictions. Finally, the consensus prediction, the mean confidence value and the standard deviation of the confidence values are calculated for each genomic region and each prediction setup.

For both machine learning analysis and prediction analysis, EpiGRAPH currently supports the use of seven different machine learning methods/configurations: (i) Support vector machine with linear kernel; (ii) support vector machine with RBF kernel, (iii) AdaBoost on tree stumps, (iv) logistic regression, (v) random forest, (vi) C4.5 tree generator, and (vii) naïve Bayes, all of which are implemented using functions from the Weka package (http://www.cs.waikato.ac.nz/~ml/weka/) with default parameters. For comparison and to give a baseline for the expected accuracy, we also include a trivial algorithm that always predicts the majority class.

X-GRAF format

Throughout EpiGRAPH's workflow (Figure 4), analyses and custom attributes are stored in XML files. In order to standardize the format of these XML files and to facilitate interoperability between the frontend, middleware and backend components, we defined the XML Genomic Relationship Analysis Format (X-GRAF). X-GRAF consists of an XML schema, against which any X-GRAF-compatible XML file has to validate in order to be regarded as syntactically correct, and a set of rules that describe the semantic interpretation of X-GRAF-compliant XML files (see http://epigraph.mpi-inf.mpg.de/xml/ for details). X-GRAF-compatible XML files can incorporate two major subtrees, "attribute definition" and "analysis" (see Supplementary Figure 1 for illustration). The attribute definition section keeps track of genomic attributes, which are organized in attribute groups and can be defined by embedded tab-separated tables or by referring to external data sources (such as a database or a URL). The analysis section documents all analysis steps, including attribute calculation, statistical analysis, diagram generation, machine learning analysis and prediction analysis. Each of these subsections comprises an analysis configuration (a description of what is to be calculated), analysis tracking information (e.g. submission data, current state and error messages) and the results of the analysis (in the form of tables and diagrams that are directly embedded in the XML file).

Although X-GRAF was created for EpiGRAPH, it is designed with additional applications in mind. Being both formalized and sufficiently easy-to-understand, X-GRAF may provide a suitable basis for analysis specification, results documentation and data exchange of future genome analysis tools and statistical genome browsers.

Adapting and extending EpiGRAPH

EpiGRAPH provides multiple options for customization, adaptation and extension, which are outlined below in increasing order of complexity and power.

First, it is possible to use EpiGRAPH for attribute calculation only, thus profiting from EpiGRAPH's large and carefully selected set of default attributes, while performing follow-up analyses offline (e.g. with the R statistics package). To that end, the user performs a normal EpiGRAPH analysis and presses the "Download Data Table" button on the results page to obtain a tab-separated data file that contains all attribute values for all genomic regions in the input dataset.

Second, the user can add custom genomic attributes to EpiGRAPH, using the "Upload Custom Attribute Dataset" button on the overview page. A new custom attribute can be defined in three ways: (i) by uploading a set of genomic regions; (ii) by specifying how the attribute can be calculated from other attributes that are already present in the database (e.g. filtering rows that match a specific condition or defining additional columns); and (iii) by deriving a randomized control attribute that matches an existing attribute in terms of its GC content,

repeat content and/or exon overlap. Custom attributes can be included in EpiGRAPH analyses in the same way as the default attributes, but they are exclusively accessible to the user who has defined them.

Third, the user can specify advanced analysis requests and attribute calculations directly in EpiGRAPH's internal X-GRAF format. Any XML file that adheres to the X-GRAF format can be uploaded through the "Execute Analysis Based on Existing XML File" button, bypassing the interactive "Define New Analysis" pages. This can be useful for several reasons: (i) When running the same analysis on different datasets, it is often convenient to design the analysis once using the web frontend, then download its specifications in X-GRAF format and use a text editor or a custom script to produce separate versions for each dataset. (ii) Sharing X-GRAF files with other researchers (e.g. by inclusion in the supplementary material of a paper) will enable them to reproduce the analysis by simply submitting the X-GRAF files back to the EpiGRAPH web service, thus contributing to reproducible research (Gentleman 2005). (iii) Some of the more advanced features (e.g. calculated attributes with multiple new columns) are supported by the calculation engine but cannot be specified easily using the web frontend.

Fourth, the user can download a "light" version of the EpiGRAPH calculation engine for local installation, which runs on any computer with recent versions of Python (http://www.python.org/), R (for statistical analysis, http://www.r-project.org/) and Weka (for machine learning analysis, http://www.cs.waikato.ac.nz/~ml/weka/), after a few additional libraries have been installed. The "light" version (source code available from http://epigraph.mpi-inf.mpg.de/sourcecode/) is particularly useful for researchers developing new bioinformatic methods for genome analysis, such as new flavors of the statistical analysis, diagram generation, machine learning analysis and prediction analysis, but who do not want to spend their time writing code for attribute calculation. The main disadvantage of the "light" version is that in the absence of a relational database all genomic attributes have to be stored in flat files. However, the "light" version is code-compatible with the full version of EpiGRAPH. Hence it is possible to develop and test new modules using the "light" version and to incorporate the completed modules into the EpiGRAPH web service.

Fifth, the user can obtain and install the full version of EpiGRAPH (release package and source code available on request), which includes the process control middleware and the web frontend components as well as a version of the calculation engine that provides full database support. While running a full-blown EpiGRAPH server locally is a non-trivial task and requires both a Java application server (e.g. Tomcat, http://tomcat.apache.org/) and an Oracle 11g database server (http://www.oracle.com/database/), this setting gives the user full flexibility for customizing EpiGRAPH and a powerful infrastructure for genome analysis.

Authors' contributions

CB initiated the project, conceptualized the software, implemented the frontend, middleware and database components as well as an early backend prototype, performed the case study and drafted the paper. KH designed and implemented a substantially enhanced version of the backend, performed extensive testing and contributed important ideas to all aspects of the project. JB set up and maintained the technical infrastructure. All authors provided relevant input at different stages of the project and contributed to the writing of the paper.

Acknowledgements

We would like to thank Jörn Walter, Martina Paulsen, Eivind Hovig and the Galaxy team for helpful discussions, Yassen Assenov, Barbara Hutter and Fang Liu for testing earlier version of EpiGRAPH and Holger Jung for contributing source code to the attribute calculation backend. This work was partially funded by the European Union through the CANCERDIP project (HEALTH-F2-2007-200620; http://www.cancerdip.eu/).

Competing interests

The authors declare that no competing financial interests exist.

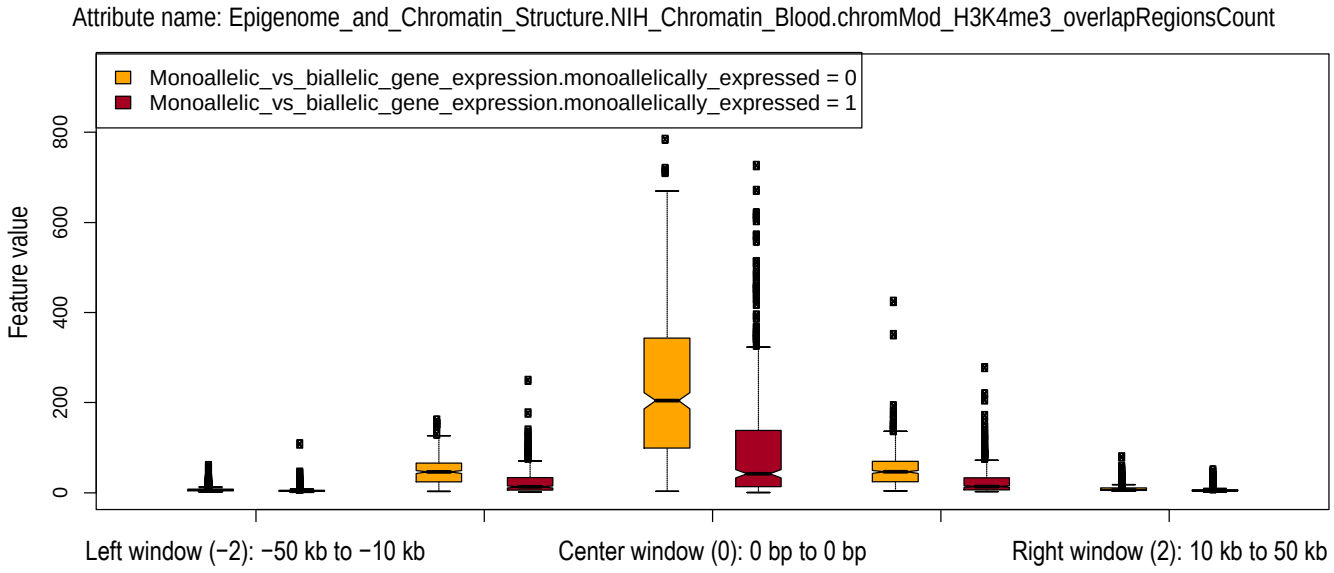
Figures

id	var name	att name	group name	P-val raw	sig bonf	sig fdr	mean class=0	mean class=1	method
1	chromMod_CTCF_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	8.960	4.282	wilcox
2	chromMod_CTCF_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	106.1	59.93	wilcox
3	chromMod_H2A_Z_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	47.85	21.63	wilcox
4	chromMod_H2A_Z_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	376.6	236.3	wilcox
5	chromMod_H2BK5me1_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	7.819	4.772	wilcox
6	chromMod_H2BK5me1_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	132.6	87.38	wilcox
7	chromMod_H3K27me1_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	6.461	3.774	wilcox
8	chromMod_H3K27me1_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	128.7	79.57	wilcox
9	chromMod_H3K4me2_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	18.60	11.34	wilcox
10	chromMod_H3K4me2_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	258.5	177.5	wilcox
11	chromMod_H3K4me3_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	228.4	100.7	wilcox
12	chromMod_H3K9me1_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	19.46	13.86	wilcox
13	chromMod_PolII_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	27.24	8.970	wilcox
14	chromMod_PolII_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	294.5	124.2	wilcox
15	overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	0	Yes	Yes	431.3	232.4	wilcox
16	thymus	GNF_Atlas_2	Transcriptome	0	Yes	Yes	0.176	0.010	wilcox
17	overlapRegionsCount	Human_ESTs	Transcriptome	0	Yes	Yes	101.3	37.77	wilcox
18	overlapRegionsCount	Spliced_ESTs	Transcriptome	0	Yes	Yes	93.35	32.82	wilcox
19	chromMod_H3K27me3_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	4.91e-41	Yes	Yes	2.241	5.902	wilcox
20	chromMod_H3K27me3_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	1.98e-40	Yes	Yes	46.95	113.1	wilcox

Figure 1. Results screenshot of EpiGRAPH's statistical analysis identifying significant differences between the promoter regions of monoallelically vs. biallelically expressed genes

Comparing the promoter regions of monoallelically expressed genes (class = 1) with those of biallelically expressed genes (class = 0), EpiGRAPH's statistical analysis detects highly significant differences in terms of chromatin structure and transcriptional activity. *P*-values in this table are based on the non-parametric Wilcoxon rank-sum test (cf. "method" column). Multiple hypothesis testing was accounted for with both the highly conservative Bonferroni method ("sig bonf" column) and the false-discovery-rate method ("sig fdr" column). A global significance threshold of 0.05 was used in both cases. Attributes highlighted in red are discussed in the main text. An explanation of attribute names is available from the EpiGRAPH website (http://epigraph.mpi-inf.mpg.de/attributes/).

A. Boxplot diagram for (open-chromatin associated) histone H3 lysine 4 trimethylation



B. Boxplot diagram for (repressive) histone H3 lysine 27 trimethylation

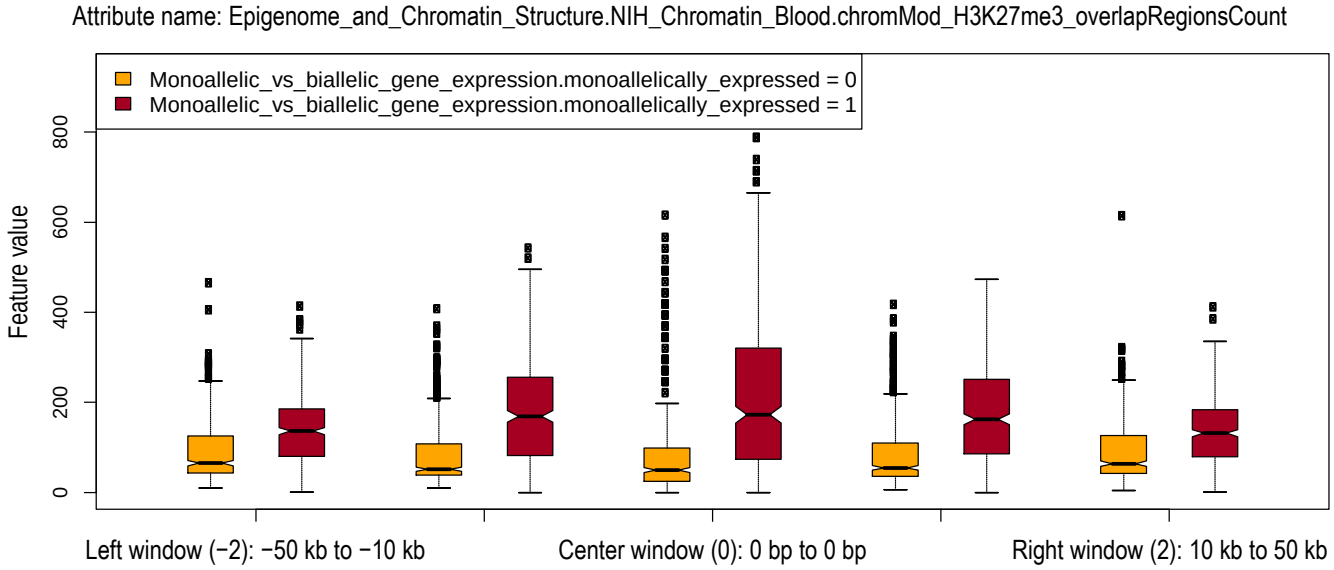


Figure 2. EpiGRAPH-generated diagrams highlighting differential histone modification patterns for the promoters of monoallelically vs. biallelically expressed genes

This figure displays EpiGRAPH-generated boxplots comparing the promoter regions of genes exhibiting monoallelic (red boxplots) vs. biallelic gene expression (yellow boxplots) with respect to their enrichment for two histone modifications. The y-axis plots the frequency of overlap with ChIP-seq tags (Barski et al. 2007), which is indicative of the strength of enrichment of the corresponding histone modification. Boxplots are in standard format (boxes show center quartiles, whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box) and outliers are shown as dots.

A. Initial results using EpiGRAPH's default settings

run	group name	#vars	prediction method	mean corr	corr sd	mean acc	acc sd	sens	spec	#cases
1	Epigenome_and_Chromatin_Structure	415	svm_linear	0.477	0.011	0.738	0.006	0.734	0.742	928
2	Regulatory_Regions	1395	svm_linear	0.025	0.017	0.512	0.009	0.516	0.508	928
3	Repetitive_DNA	982	svm_linear	0.166	0.027	0.583	0.013	0.574	0.592	928
4	Transcriptome	245	svm_linear	0.332	0.009	0.665	0.005	0.625	0.706	928
5	Epigenome_and_Chromatin_Structure+Regulatory_Regions+Repetitive_DNA+Transcriptome	3037	svm_linear	0.304	0.021	0.652	0.011	0.652	0.652	928

B. Follow-up analysis for all possible combinations of attribute groups

run	group name	#vars	prediction method	mean corr	corr sd	mean acc	acc sd	sens	spec	#cases
8	Epigenome_and_Chromatin_Structure	83	svm_linear	0.493	0.009	0.746	0.004	0.748	0.744	928
12	Transcriptome+Epigenome_and_Chromatin_Structure	132	svm_linear	0.485	0.010	0.742	0.005	0.740	0.745	928
10	Regulatory_Regions+Epigenome_and_Chromatin_Structure	362	svm_linear	0.480	0.010	0.740	0.005	0.743	0.736	928
9	Repetitive_DNA+Epigenome_and_Chromatin_Structure	235	svm_linear	0.473	0.016	0.736	0.008	0.738	0.735	928
13	Repetitive_DNA+Transcriptome+Epigenome_and_Chromatin_Structure	284	svm_linear	0.472	0.014	0.736	0.007	0.741	0.731	928
14	Regulatory_Regions+Transcriptome+Epigenome_and_Chromatin_Structure	411	svm_linear	0.469	0.019	0.735	0.009	0.743	0.726	928
11	Repetitive_DNA+Regulatory_Regions+Epigenome_and_Chromatin_Structure	514	svm_linear	0.444	0.020	0.722	0.010	0.728	0.716	928
15	Repetitive_DNA+Regulatory_Regions+Transcriptome+Epigenome_and_Chromatin_Structure	563	svm_linear	0.438	0.015	0.719	0.007	0.721	0.717	928
4	Transcriptome	49	svm_linear	0.346	0.011	0.673	0.006	0.688	0.658	928
5	Repetitive_DNA+Transcriptome	201	svm_linear	0.318	0.010	0.659	0.005	0.637	0.681	928
6	Regulatory_Regions+Transcriptome	328	svm_linear	0.300	0.018	0.650	0.009	0.655	0.645	928
7	Repetitive_DNA+Regulatory_Regions+Transcriptome	480	svm_linear	0.277	0.013	0.638	0.006	0.643	0.634	928
1	Repetitive_DNA	152	svm_linear	0.207	0.014	0.601	0.007	0.489	0.713	928
3	Repetitive_DNA+Regulatory_Regions	431	svm_linear	0.178	0.016	0.589	0.008	0.598	0.580	928
2	Regulatory_Regions	279	svm_linear	0.113	0.020	0.556	0.010	0.624	0.488	928

C. Follow-up analysis with all available machine learning algorithms

run	group name	#vars	prediction method	mean corr	corr sd	mean acc	acc sd	sens	spec	#cases
1	Epigenome_and_Chromatin_Structure	83	ada_stump	0.498	0.006	0.749	0.003	0.738	0.760	928
1	Epigenome_and_Chromatin_Structure	83	svm_linear	0.493	0.009	0.746	0.004	0.748	0.744	928
1	Epigenome_and_Chromatin_Structure	83	svm_rbf	0.468	0.005	0.733	0.002	0.780	0.686	928
1	Epigenome_and_Chromatin_Structure	83	logistic	0.456	0.010	0.728	0.005	0.719	0.737	928
1	Epigenome_and_Chromatin_Structure	83	bayes_naive	0.445	0.009	0.722	0.004	0.764	0.679	928
1	Epigenome_and_Chromatin_Structure	83	tree_forest	0.434	0.017	0.717	0.008	0.697	0.737	928
1	Epigenome_and_Chromatin_Structure	83	tree_c45	0.367	0.015	0.683	0.007	0.639	0.726	928
1	Epigenome_and_Chromatin_Structure	83	rule_trivial	-0.009	1.83e-18	0.496	5.85e-17	0.397	0.595	928

Figure 3. Results screenshots of EpiGRAPH's machine learning module predicting monoallelic gene expression

These screenshots display the results of machine learning analyses comparing the promoter regions of monoallelically expressed genes (class = 1) with those of biallelically expressed genes (class = 0), in each panel based on different EpiGRAPH settings. The values in the tables summarize the average performance of a linear support vector machine or alternative machine learning algorithms (panel C) that were trained and evaluated in ten repetitions of a tenfold cross-validation. Performance measures include mean correlation ("mean corr" column), prediction accuracy ("mean acc" column), sensitivity ("sens" column) and specificity ("spec" column). Additional columns display standard deviations observed among the repeated cross-validations with random partition assignment ("corr sd" and "acc sd"), the number of variables in each attribute group ("#vars") and the total number of genomic regions included in the analysis ("#cases").

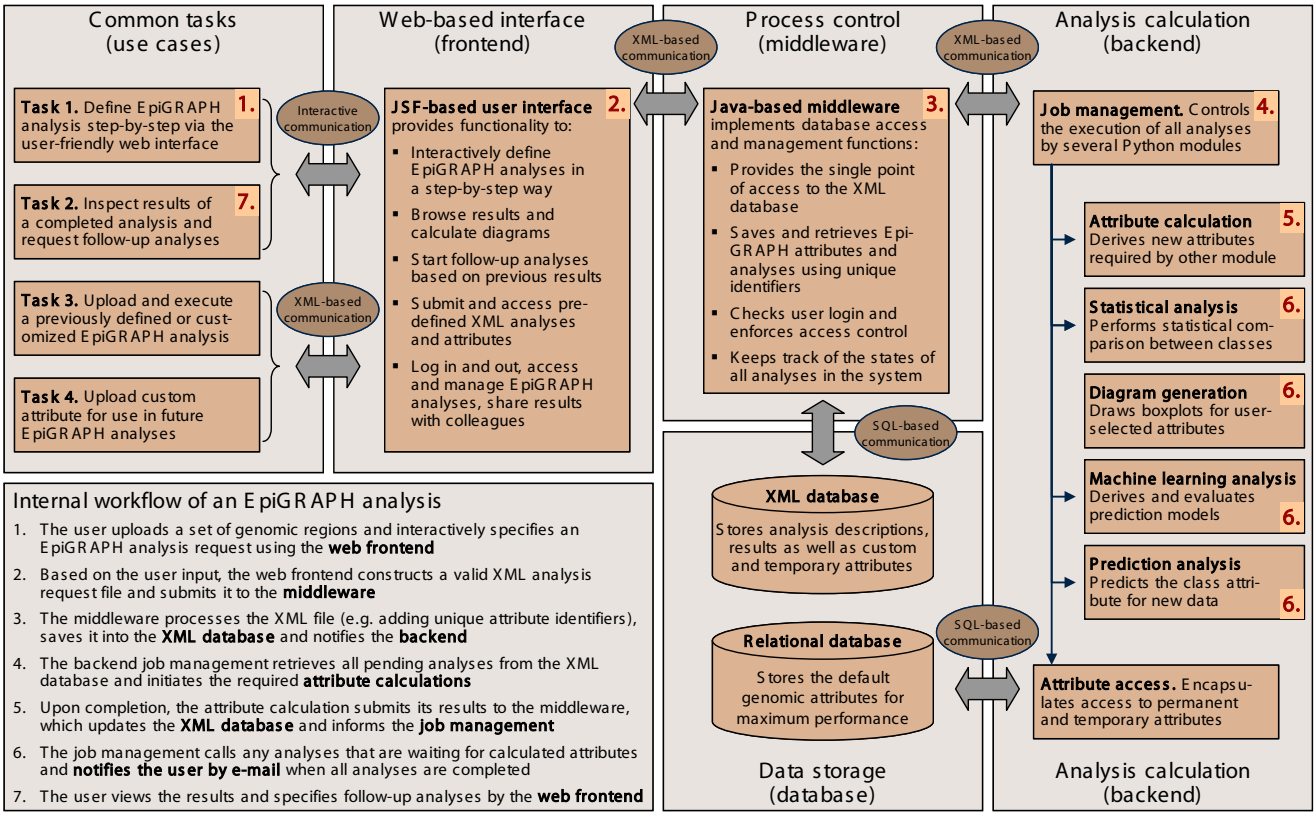


Figure 4. Outline of EpiGRAPH's software architecture

This figure displays a schematic overview of EpiGRAPH's software components, and it describes their interaction in a typical analysis workflow. The red numbers indicate the key component(s) for each step of the workflow description outlined in the bottom left of the figure. Abbreviation: JSF – Java Server Faces (which is a Java-based web application framework).

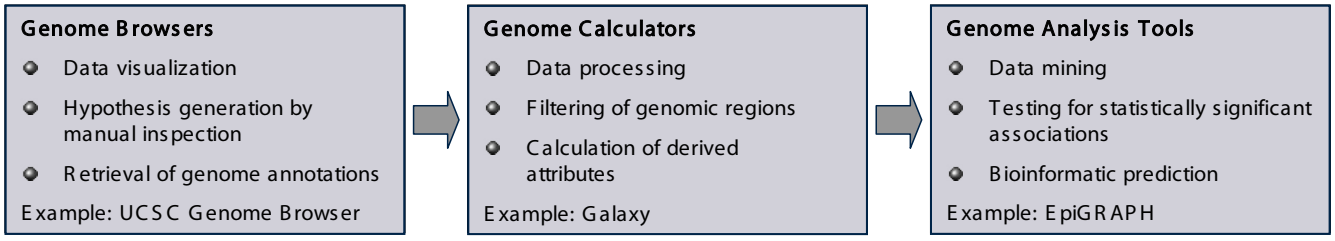


Figure 5. Workflow for web-based analysis of large genome and epigenome datasets

This figure outlines a workflow for the analysis of genome and epigenome data using publicly available web services. Initially, the user uploads a newly generated dataset into a genome browser, which visualizes the data and facilitates hypothesis generation by manual inspection (left box). Next, data can be processed with a genome calculator such as Galaxy, in order to extract interesting regions for in-depth analysis (center box). Finally, genome analysis tools such as EpiGRAPH facilitate the search for significant associations with genome annotation data and enable bioinformatic prediction of genomic regions with similar characteristics as the input dataset (right box).

Tables

Attribute Groups	Total Number of Attributes					Attributes (Examples)
	hg18	*hg17*	*mm9*	*panTro2*	*galGal3*	
DNA Sequence	178	178	178	178	178	Frequency of "TATA" pattern, cytosine content, CpG frequency
DNA Structure	21	21	21	21	21	Predicted DNA helix twist, predicted solvent accessibility
Repetitive DNA	95	95	91	94	94	Overlap with Alu elements, LINE elements and tandem repeats
Chromosome Organization	18	29	15	-	-	Overlap with chromosomal bands and isochores
Evolutionary History	94	101	-	-	86	Overlap with evolutionary conserved regions
Population Variation	75	75	-	-	-	SNP density and overlap with specific SNP types (e.g. non-synonymous exonic or splice site)
Genes	37	60	20	10	10	Overlap with annotated genes, pseudogenes and predicted mi-croRNA genes
Regulatory Regions	249	259	5	5	5	Overlap with CpG islands and predicted transcription factor binding sites
Transcriptome	49	65	9	9	9	Overlap with ESTs and mRNA sequences
Epigenome and Chromatin Structure	80	17	114	-	-	Overlap with ChIP-seq tags indicating enrichment for specific histone modifications
Sum	896	900	453	317	403	

Table 1. List of default attributes included in EpiGRAPH

This table summarizes the collection of default attribute that are currently included in EpiGRAPH. Due to different degrees of annotation, the numbers differ between the genomes of human (*hg18* and *hg17*), mouse (*mm9*), chimp (*panTro2*) and chicken (*galGal3*).

References

Allen, E. et al. 2003. High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes. *Proc Natl Acad Sci U S A* 100: 9940-9945.

Bailey, J.A. et al. 2000. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proc Natl Acad Sci U S A* 97: 6634-6639.

Barski, A. et al. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* 129: 823-837.

Benjamini, Y. et al. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* 57: 289-300.

Berry, C. et al. 2006. Selection of target sites for mobile DNA integration in the human genome. *PLoS Comput Biol* 2: e157.

Blankenberg, D. et al. 2007. A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. *Genome Res* 17: 960-964.

Bock, C. et al. 2008. Computational epigenetics. *Bioinformatics* 24: 1-10.

Bock, C. et al. 2006. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet* 2: e26.

Bock, C. et al. 2007. CpG island mapping by epigenome prediction. *PLoS Comput Biol* 3: e110.

Bock, C. et al. 2008. Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. *Nucleic Acids Res* 36: e55.

Carninci, P. et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* 38: 626-635.

Cohen, S.M. et al. 2006. Genome-wide sequence and functional analysis of early replicating DNA in normal human fibroblasts. *BMC Genomics* 7: 301.

Costantini, M. et al. 2006. An isochore map of human chromosomes. *Genome Res* 16: 536-541.

Das, R. et al. 2006. Computational prediction of methylation status in human genomic sequences. *Proc Natl Acad Sci U S A* 103: 10713-10716.

Derti, A. et al. 2006. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat Genet* 38: 1216-1220.

Fang, F. et al. 2006. Predicting methylation status of CpG islands in the human brain. *Bioinformatics* 22: 2204-2209.

Flicek, P. et al. 2008. Ensembl 2008. *Nucleic Acids Res* 36: D707-714.

Gardiner, E.J. et al. 2003. Sequence-dependent DNA structure: a database of octamer structural parameters. *J Mol Biol* 332: 1025-1035.
Gentleman, R. 2005. Reproducible Research: A Bioinformatics Case Study. *Statistical Applications in Genetics and Molecular Biology* 4.
Giardine, B. et al. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* 15: 1451-1455.
Gimelbrant, A. et al. 2007. Widespread monoallelic expression on human autosomes. *Science* 318: 1136-1140.
Greally, J.M. 2002. Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. *Proc Natl Acad Sci U S A* 99: 327-332.
Greenbaum, J.A. et al. 2007. Construction of a genome-scale structural map at single-nucleotide resolution. *Genome Res* 17: 947-953.
Guelen, L. et al. 2008. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* 453: 948-951.
Heard, E. 2004. Recent advances in X-chromosome inactivation. *Curr Opin Cell Biol* 16: 247-255.
Huang, D.W. et al. 2007. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res* 35: W169-175.
Hull, D. et al. 2006. Taverna: a tool for building and running workflows of services. *Nucleic Acids Res* 34: W729-732.
Karolchik, D. et al. 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* 36: D773-779.
Ke, X. et al. 2002. A novel approach for identifying candidate imprinted genes through sequence analysis of imprinted and control genes. *Hum Genet* 111: 511-520.
Liu, F. et al. 2007. The human genomic melting map. *PLoS Comput Biol* 3: e93.
Luedi, P.P. et al. 2007. Computational and experimental identification of novel human imprinted genes. *Genome Res* 17: 1723-1730.
Luedi, P.P. et al. 2005. Genome-wide prediction of imprinted murine genes. *Genome Res* 15: 875-884.
Mardis, E.R. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet* 24: 133-141.
Meissner, A. et al. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454: 766-770.
Montgomery, S.B. et al. 2007. A Survey of Genomic Properties for the Detection of Regulatory Polymorphisms. *PLoS Comput Biol* 3: e106.
Moser, D. et al. 2008. Functional Analysis of a Potassium-Chloride Co-Transporter 3 (SLC12A6) Promoter Polymorphism Leading to an Additional DNA Methylation Site. *Neuropsychopharmacology*: advance online publication, 4 June 2008; doi: 2010.1038/npp.2008.2077.
Oinn, T. et al. 2004. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 20: 3045-3054.
Reik, W. 2007. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* 447: 425-432.
Rollins, R.A. et al. 2006. Large-scale structure of genomic methylation patterns. *Genome Res* 16: 157-163.
Schones, D.E. et al. 2008. Genome-wide approaches to studying chromatin modifications. *Nat Rev Genet* 9: 179-191.
Su, A.I. et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* 101: 6062-6067.
Subramanian, A. et al. 2007. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* 23: 3251-3253.
Tarca, A.L. et al. 2007. Machine learning and its applications to biology. *PLoS Comput Biol* 3: e116.
van Steensel, B. 2005. Mapping of genetic and epigenetic regulatory networks using microarrays. *Nat Genet* 37 Suppl: S18-24.
Wang, Z. et al. 2006. Evidence of influence of genomic DNA sequence on human X chromosome inactivation. *PLoS Comput Biol* 2: e113.
Wen, B. et al. 2008. Overlapping euchromatin/heterochromatin-associated marks are enriched in imprinted gene regions and predict allele-specific modification. *Genome Res*: advance online publication, 10 October 2008; doi: 2010.1101/gr.067587.067108.