Bioinformatics, 32(10), 2016, 1502–1508
doi: 10.1093/bioinformatics/btv770
Advance Access Publication Date: 18 January 2016
Original Paper

OXFORD

Systems biology

Multi-omics enrichment analysis using the GeneTrail2 web service

Daniel Stöckel[1,*], Tim Kehl[1], Patrick Trampert[1], Lara Schneider[1], Christina Backes[1], Nicole Ludwig[3], Andreas Gerasch[2], Michael Kaufmann[2], Manfred Gessler[4], Norbert Graf[5], Eckart Meese[3], Andreas Keller[1] and Hans-Peter Lenhof[1]

[1]Center for Bioinformatics, Saarland University, Saarbrücken D-66041, [2]Center for Bioinformatics, Eberhard-Karls-University, Tübingen, D-72076, [3]Department of Human Genetics, Medical School, Saarland University, Homburg D-66421, [4]Theodor-Boveri-Institute/Biocenter, Developmental Biochemistry, and Comprehensive Cancer Center Mainfranken, Würzburg University, Würzburg D-97074 and [5]Department of Pediatric Oncology and Hematology, Medical School, Saarland University, Homburg, D-66421, Germany

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on 19 October 2015; revised on 17 December 2015; accepted on 28 December 2015

Abstract

Motivation: Gene set analysis has revolutionized the interpretation of high-throughput transcriptomic data. Nowadays, with comprehensive studies that measure multiple -omics from the same sample, powerful tools for the integrative analysis of multi-omics datasets are required.

Results: Here, we present GeneTrail2, a web service allowing the integrated analysis of transcriptomic, miRNomic, genomic and proteomic datasets. It offers multiple statistical tests, a large number of predefined reference sets, as well as a comprehensive collection of biological categories and enables direct comparisons between the computed results. We used GeneTrail2 to explore pathogenic mechanisms of Wilms tumors. We not only succeeded in revealing signaling cascades that may contribute to the malignancy of blastemal subtype tumors but also identified potential biomarkers for nephroblastoma with adverse prognosis. The presented use-case demonstrates that GeneTrail2 is well equipped for the integrative analysis of comprehensive -omics data and may help to shed light on complex pathogenic mechanisms in cancer and other diseases.

Availability and implementation: GeneTrail2 can be freely accessed under https://genetrail2.bioinf.uni-sb.de.

Contact: dstoeckel@bioinf.uni-sb.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The advent of high-throughput analysis technologies during the last decades has revolutionized biology and medicine and has lead to a tremendous increase in the amount of available, high-dimensional biological data. However, technologies like microarrays, next-generation sequencing or mass spectrometry are highly sensitive to changes in the experimental conditions or protocols, giving rise to technological noise that must be carefully accounted for (Leek *et al.*, 2010). Due to the above reasons, the development of automated and robust statistical analysis methods for high-throughput data has become a necessity. Enrichment methods are one fundamental class of analysis procedures for the study of pathogenic mechanisms and the identification of deregulated pathways and categories. Widely-applied enrichment methods are *Over Representation Analysis* (*ORA*) (Drăghici *et al.*, 2003) and *Gene Set Enrichment Analysis* (*GSEA*) (Subramanian *et al.*, 2005). Other approaches, e.g. Kim and Volsky (2005), Newton *et al.* (2007), Tian

© The Author 2016. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

et al. (2005), employ straightforward averaging methods that, despite their simplicity, offer competitive performance. Building upon such averaging approaches, Efron and Tibshirani (2007) presented the *maxmean* statistics. More specialized approaches, which complement the generally applicable methods presented above, are e.g. topGO (Alexa *et al.*, 2006) and GO-Bayes (Zhang *et al.*, 2010), which operate on the *Gene Ontology* (*GO*) (Ashburner *et al.*, 2000) and account for its hierarchical structure. Another class of algorithms uses the topology of known biological networks to improve the computed enrichments. Examples are GGEA (Geistlinger *et al.*, 2011) and EnrichNet (Glaab *et al.*, 2012).

Due to the popularity of enrichment methods, many implementations are available, both as stand-alone applications and as web services. Some focus on a certain database while others are limited to one or two algorithms (Fig. 1). Examples for available tools are the Broad Institute's GSEA implementation (Subramanian *et al.*, 2005) or the GSEA-SNP (Holden *et al.*, 2008) R package. Also a wide variety of web services for enrichment analysis exists, which we discussed briefly in Supplementary Note S1. Moreover, a comprehensive collection of enrichment tools can be found in the OMICtools database (Henry *et al.*, 2014). Extensive reviews on enrichment methods (Ackermann and Strimmer, 2009; Efron and Tibshirani, 2007; Huang *et al.*, 2009; Hung *et al.*, 2011; Khatri *et al.*, 2012; Naeem *et al.*, 2012) have been published and reveal that no real gold standard exists. This is due to the fact that each of the proposed methods is based on differing definitions of enriched categories (differing null hypotheses), making their results incomparable in general. Instead of using a single 'magic bullet', an appropriate algorithm needs to be chosen carefully for each individual research task.

Finally, as heterogenous datasets, e.g. datasets comprised of genomic variations, miRNA and mRNA expression measurements, are becoming increasingly common, integrative platforms for enrichment analyses are needed. Unfortunately, only few such tools are readily available. Examples are miRTrail (Laczny *et al.*, 2012), which links mRNA, miRNA and disease phentotypes, Genevar (Yang *et al.*, 2010), focusing on the association between SNPs and eQTLs, or RAMONA using a Bayesian Model for linking arbitrary -omics to ontology terms (Sass *et al.*, 2014).

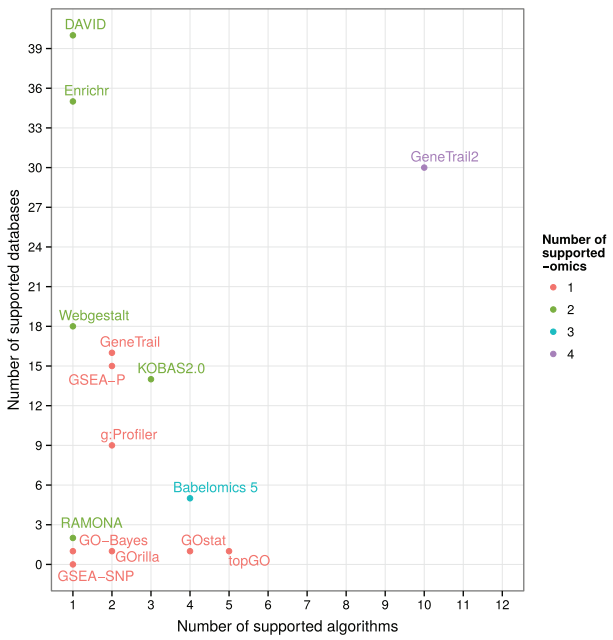

Fig. 1. A comparison of selected enrichment tools. The *number of supported databases* refers to the number of unique data sources from which categories have been obtained. Databases are counted across all supported species and -omics. The *number of supported algorithms* refers to the number of algorithms offered for analysis. Related methods (e.g. network algorithms) have been included. A tool was defined as *supporting an -omics,* if it provides dedicated biological categories for this -omics type

In this study, we present GeneTrail2, a new web server for the analysis of multi-omics datasets, with which we provide one of the most comprehensive tools for enrichment analysis. For human alone, it features over 46 000 categories collected from over 30 databases including KEGG, Reactome, GO, WikiPathways, DrugBank, Pfam, miRWalk and miRDB (cf. Supplementary Table S12). It natively supports transcriptomics, miRNomics, proteomics, and genomics data and can convert between 32 common identifier types. In total, we implemented 13 identifier-level statistics, 10 set-level statistics (see Section 2.3), two *P*-value computation strategies and eight *P*-value adjustment methods. Data from all major -omics are supported, making it possible to analyze and explore heterogeneous datasets in an interactive fashion using GeneTrail2's web interface. The web interface is built on top of modern web technologies with special attention to usability. Non-expert users can quickly perform comprehensive analyses using the predefined workflow, which is complemented with thorough documentation. Moreover, the interface enables users to integrate enrichments obtained from multiple -omics using the integrated mapping procedures and our side-by-side view. For further analysis tasks, we offer a deep integration into existing applications like the network visualization tool BiNA (Gerasch *et al.*, 2014) or the NetworkTrail (Stöckel *et al.*, 2013) web service. Another key feature of GeneTrail2 is its RESTful API, through which power users can script the web service. This scripting interface allows the seamless integration of GeneTrail2 into workflow systems such as Galaxy (Goecks *et al.*, 2010) or Taverna (Wolstencroft *et al.*, 2013). The backend of GeneTrail2 relies on highly optimized C++ code leading to excellent computation times. GeneTrail2 can be accessed under https://genetrail2.bioinf.uni-sb.de.

We demonstrate the capabilities of GeneTrail2 by applying it to a Wilms tumor expression dataset with the goal of identifying molecular determinants for the increased malignancy of certain Wilms tumor subtypes. Further use cases are discussed on the website.

2 Methods

Since GeneTrail2 is a comprehensive software platform comprising more than 31 000 lines of code, we start with a short introduction to the architecture of our web service. Then, we present an example workflow serving as guide through the available functionality and implemented methods.

2.1 Architecture

GeneTrail2 is a complete rewrite of its predecessor, the GeneTrail (Backes *et al.*, 2007) web service, with the goal to provide more, readily available identifier- and set-level statistics and to greatly increase its flexibility. Furthermore, it supports user accounts and, for increased reproducibility, fully automatic documentation of all used parameters. To accommodate this, the new server is based on a modular architecture (cf. Fig. 2) allowing to easily add new features, exchange implementation details or perform general maintenance. We implemented a user interface based on HTML5 and JavaEE technology using the Thymeleaf template engine and JQuery. This interface interacts with the web service using a RESTful API. This API allows access to the complete functionality of GeneTrail2. It offers interfaces for starting and managing computationally

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

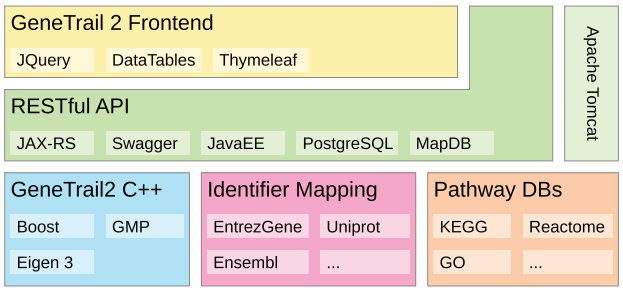

Fig. 2. Architecture of GeneTrail2. Core algorithms are implemented as an optimized C++ library based on Boost, Eigen 3 and GMP. On top of this library we implemented a JAX-RS-based RESTful API. The frontend is based on the Thymeleaf template engine and JQuery. As application server we use Apache Tomcat

intensive analysis tasks as well as querying computed results. Providing a RESTful API has the additional advantage that users can write custom scripts in any programming language, e.g. for batch processing.

All compute-intensive tasks were implemented using highly optimized C++ code in order to achieve maximum performance. This code relies on the Boost library for the implementation of probability distributions and auxiliary algorithms. Eigen 3 is used for matrix–vector algebra and the GMP library provides multi precision integer and floating point operations.

2.2 Workflow

GeneTrail2 offers a considerable number of analysis workflows (Fig. 3). Due to space constraints, we only discuss one common scenario. First, the user uploads the data to be analyzed, e.g. a matrix containing normalized expression measurements. Then, the data can be divided into sample and reference sets in order to enable the computation of identifier-level scores. After score computation, a set-level statistic must be selected. The user can then choose biological categories that should be analyzed.

For each step the user can adjust the parameters of the employed method. As this usually requires considerable expert knowledge, we provide defaults that should be applicable for most use-cases and that have been chosen conservatively in order to prevent false discoveries.

2.3 Method overview

GeneTrail2 accepts the following simple, tab-delimited text files as inputs: identifier lists, score lists and data matrices. In addition, it is possible to analyze microarray expression datasets from the *Gene Expression Omnibus* repository (Barrett *et al.*, 2013) directly.

2.3.1 Identifier and organism estimation

A key challenge when dealing with uploaded data is the automatic detection of the identifiers used in the dataset. We solved this problem by creating a MapDB-based database of supported identifiers, against which user data are validated. In addition, the database supports mapping between different identifier types such as *UniProt Accessions* and *Entrez Gene Ids*. Coupled with automated file type detection, no user intervention should be required when uploading data.

2.3.2 Identifier-level statistics

Whereas identifier lists and score lists can be directly used as input for computing enrichments, expression matrices need to be

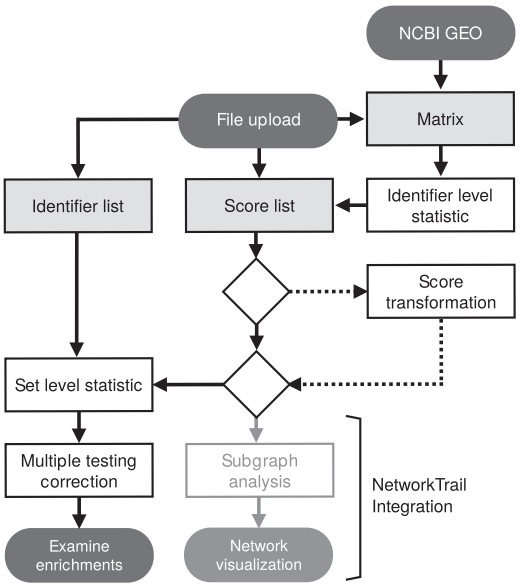

Fig. 3. Simplified flowchart of GeneTrail2. Round boxes depict start/end states. Boxes with gray background represent input file types, whereas a white background represents processing steps. Diamonds are decision nodes

processed to (identifier-level) scores first. To this end, we provide a comprehensive set of statistics (see Supplementary Note S2.1). Among these are the fold-change, Wilcoxon test and a (regularized) version of the *t*-test. If possible and applicable, paired versions of the statistics are provided. For supporting count data we integrated the DESeq2 (Love *et al.*, 2014), edgeR (Robinson *et al.*, 2010) and RUVSeq (Risso *et al.*, 2014) approaches.

Simple transformations such as absolute value, square, square root or logarithm can be applied to computed scores.

2.3.3 Set-level statistics and categories

Once the input data have been prepared, one of the following set-level statistics can be applied: the weighted and unweighted Kolmogorov-Smirnov (KS) test, the Wilcoxon test, ORA, sum-, mean-, median- and maxmean statistics, as well as the one and two sample *t*-test. After choosing an algorithm, the user can select categories for which an enrichment should be computed. The categories integrated into GeneTrail2 were collected from many, commonly used biological databases (cf. Supplementary Table S12). Among others, these databases cover signaling pathways, genomic features, transcription factor (TF) targets, miRNA targets, drug targets and (disease) phenotypes. In addition, user-defined categories can be uploaded in gene matrix transposed (Subramanian *et al.*, 2005) format. For ORA, it is also required to select a reference (background) set. For human alone, we offer over 40 predefined reference sets (see Supplementary Tables S13–S17) covering many experimental platforms. If no applicable reference set is available, a custom set can be uploaded.

2.3.4 P-value computation

For determining the significance levels of the computed set-level scores, GeneTrail2 offers the *gene set* and the *phenotype* strategy. While the *gene set* strategy is based on permuting the identifier-level scores, the *phenotype* strategy randomly redistributes the measurements between the sample and reference group. An advantage of the *gene set* strategy is that it allows the direct computation of *P*-values for some methods and thus avoids costly permutation tests (see

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

Supplementary Note S2.2). This leads to a higher resolution of the computed *P*-values and very low computation times (see Section 2.4). For an in-depth discussion of the advantages and disadvantages of the respective methods we refer the reader to Tian *et al.* (2005). The *phenotype* strategy always requires that a permutation test is performed. As new identifier-level scores must be derived for every permutation, the method can only be used if a data matrix was supplied.

Finally, a method for multiple testing correction must be chosen. To this end, we provide a large set of correction methods (see Supplementary Table S6) including the popular Benjamini–Hochberg FDR adjustment (Benjamini and Hochberg, 1995).

2.3.5 Examining computed results

Enrichments can be visualized online or downloaded as compressed archives and MS Excel tables. When viewing an enrichment online, we provide detailed statistics about significant categories and their members, which are fully sort- and searchable. Besides the traditional list of significant categories, we also offer an *inverse enrichment* view, which for every gene displays the significant categories it is a member of. In addition, the user can select a set of enrichments that can be viewed side-by-side, providing a compact and clearly arranged overview. This allows to quickly detect similarities and differences with respect to the detected categories. Here, two view modes are available. While the *union* mode displays all categories that are significant in *at least one* enrichment, the *intersection* mode only displays categories that are significant in *all* enrichments. Whereas the union is useful for detecting variability between related enrichments, the intersection can be used to reduce the number of false positives by computing and comparing two or more enrichments using different algorithms. Using these modes, the user is able to effectively balance the sensitivity and specificity of an analysis in a straightforward manner. Combining the side-by-side view with GeneTrail2's mapping features allows for an integrated analysis of gene lists obtained from different -omics. For example, given transcriptomics and proteomics data, protein abundance scores can be mapped onto the corresponding genes allowing to compute enrichments using gene categories for both score lists. Via the union mode it is possible to spot similarities and differences between the enrichments obtained from both -omics.

2.4 Performance

The performance of basic tools such as enrichment methods is critical, as good performance significantly shortens the development cycles of data analysis workflows. To this end, GeneTrail2 uses a C++ implementation to guarantee optimal performance. In general, the *gene set* strategy is one order of magnitude quicker to compute than the *phenotype* strategy (see Table 1). The used set-level statistics has little influence on the computation time. GeneTrail2 significantly outperforms the Broad Institute GSEA application (Subramanian *et al.*, 2005) for both, the *gene set* and the *phenotype* strategy.

3 Results

3.1 Case study: Wilms tumors

Wilms tumors (WTs), or *nephroblastomas*, are childhood renal tumors. While in general WTs are associated with survival rates >90%, some subtypes with much higher relapse rates are known (Sredni *et al.*, 2009). For proper risk-assessment and therapy stratification it is thus crucial to identify and understand the differences in the pathogenic processes between the tumor subtypes.

Table 1. Performance data for enrichments computed on the KEGG categories using the (unweighted) KS and mean set-level statistics

	Broad GSEA	GeneTrail 2
KS		
Gene set	400 s (± 7.3 s)	9.3 s* (± 0.15 s)
Phenotype	428.8 s (± 4.32 s)	84.5 s (± 0.6 s)
Mean		
Gene set	N/A	3 s (± 0.02 s)
Phenotype	N/A	74.8 s (± 1.7 s)

For comparison the KS implementation of the Broad GSEA package was used. Mean run times over five runs are given in seconds; standard deviations are provided in parenthesis. In the comparison, the *t*-test was used as scoring method, no *P*-value correction has been performed, and 10 000 iterations were used for permutation tests. Results marked with a * used an exact *P*-value computation method. Timings were obtained on an Intel Core i7-3770 processor

We used GeneTrail2 to analyze a WT expression dataset in order to determine key players influencing the malignancy of WTs. The dataset consists of 40 mRNA and 47 miRNA expression profiles from 47 tumor biopsies from 39 patients, containing four healthy tissue samples as well as 17 blastemal, nine mixed type and 17 miscellaneous tumor samples (cf. Supplementary Section S3.1). As our dataset stems from a study following the SIOP protocol, patients underwent chemotherapy before surgery and sample collection. Approximately 25% of initial blastemal predominant tumors do not respond to preoperative chemotherapy and have a poor prognosis. Hence, special attention was put on identifying factors leading to an increased resistance to chemotherapy in this blastemal subtype, which accounts for 10% of all WTs. In the following, we refer to those tumors as *blastemal tumors*.

We describe the results of our WT analysis as a learning process that was guided by the computed enrichments, simple statistical analyses and the study of literature. For a review of the implemented enrichment algorithms, we refer the reader to the available, extensive literature (Ackermann and Strimmer, 2009; Efron and Tibshirani, 2007; Huang *et al.*, 2009; Hung *et al.*, 2011; Khatri *et al.*, 2012; Naeem *et al.*, 2012).

3.1.1 Consensus of enrichment approaches

We used GeneTrail2 to examine the differences between blastemal and non-blastemal tumors using the independent shrinkage *t*-test (Opgen-Rhein and Strimmer, 2007) to compute scores for all genes and miRNAs (see Supplementary Note S3.5). For *P*-value computation the *gene set* strategy with Benjamini–Hochberg adjustment was chosen. In order to compare the implemented set-level statistics based on our data, we applied all available set-level statistics excluding ORA to the score lists. The resulting enrichments are given in Supplementary Tables S18 and S19 and under https://genetrail2.bioinf.uni-sb.de/results.html?session=a9e84e92-ea41-42ab-9ee7-c0f8515f9234. We observed that, despite a considerable overlap, the differences between the enrichments are substantial (cf. Supplementary Fig. S1). For example, while the union of all enrichments contains 1436 *GO-Biological Process* categories, only 343 categories are contained in their intersection. As expected, this effect is especially pronounced for databases containing categories close to the significance level, which may indicate that these categories were only reported due to idiosyncrasies of the corresponding method.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

Table 2. Number of significantly enriched ($P < 0.05$) mRNA categories found by each enrichment method for the blastemal versus non-blastemal scores

Two-sample *t*-test	3866	Sum	3406
One-sample *t*-test	3852	Weighted KS	2497
Two-sample Wilcoxon	3685	Maxmean	2057
KS	3518	Median	1989
Mean	3424		

The number of significant categories per method can be found in Table 2.

In order to focus on highly relevant processes, we only consider categories consistently reported by at least seven of nine set-level statistics, which can be easily achieved using our comparative enrichment view. Further discussion only considers this intersection.

3.1.2 General observations

For mRNA the upregulation of categories like *mRNA Processing*, *Cell Cycle* and *DNA Replication* suggest a clear increase in mitotic activity in blastemal tumors. For miRNA, categories associated with various cancer types, including *HMDD-renal cell carcinoma*, are significantly enriched. The same is true for miRNA categories involved in hormone regulation, immune response, apoptosis and tumor suppression. None of the miRNA families is significant in all enrichments; however, the families *miR-302*, *miR-515*, *miR-30*, *miR-17* and *let-7* are significant for at least seven out of nine tests. The miR-302 and miR-515 families are associated with the activation of the canonical WNT pathway (Anton *et al.*, 2011). In addition, the miR-17 family is known for its oncogenic role in cancer and stem cell development (Mogilyansky and Rigoutsos, 2013).

3.1.3 Deregulation of let-7 via LIN28B and TRIM71

The let-7 miRNA family has previously been reported to play a vital role in WT suppression (Urbach *et al.*, 2014). However, many high abundant family members are upregulated (see Supplementary Table S3), which is unexpected due to let-7 miRNAs acting as tumor suppressors. A possible explanation for this behavior may be provided by TRIM71, which is among the top-scoring genes. TRIM71 degrades LIN28B via ubiquitin-mediated proteosomal degradation (Lee *et al.*, 2014) (cf. Fig. 4). LIN28B in turn suppresses the maturation of pri-let-7 miRNA. Hence, the upregulation of TRIM71 induces an upregulation of the let-7 family, which, in theory, promotes cell differentiation. However, Chang *et al.* (2012) found that TRIM71 can promote rapid embryonic stem cell (ESC) proliferation in mice and report that it represses the expression of CDKN1A, a cyclin-dependent kinase inhibitor, which acts as a cell-cycle regulator. As high expression levels of TRIM71 are commonly observed in undifferentiated cells, the authors conclude that TRIM71 is an important factor for maintaining proliferation in stem cells. Urbach *et al.* (2014) report that LIN28B is able to induce WTs under certain conditions and note that in these tumors the cap mesenchymal (CM) specific stem cell markers CITED1, EYA1 and SIX2 are upregulated. In general, this trend is present in our data, however, we find that the marker's expression is more consistent with TRIM71's expression pattern (Fig. 5 and Supplementary Fig. S2). In summary, our initial results indicate that miRNA and genes associated with stem cell fate play an essential role in blastemal tumors.

3.1.4 Activation of cancer-related WNT signaling

Deregulation of the WNT signaling pathway is often prevalent in cancer samples (Polakis, 2012). Indeed, our enrichment analysis

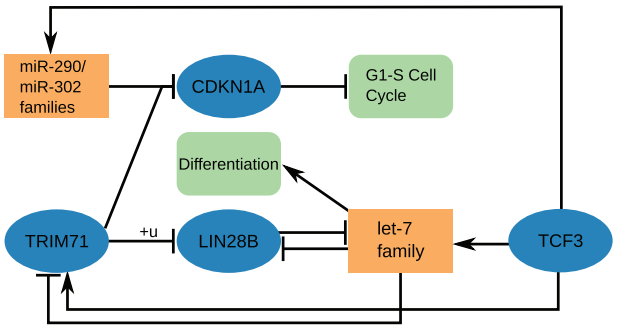

Fig. 4. TRIM71 degrades LIN28B using ubiquitin-mediated proteosomal degradation. LIN28B regulates the maturation of let-7 miRNAs (Piskounova et al., 2011), which promote cell differentiation and act on TRIM71 and LIN28B via a negative feedback loop (Marson *et al.*, 2008). TRIM71 as well as AGO2 in complex with miR-290/302 miRNAs repress CDKN1A expression leading to increased proliferation (Lee *et al.*, 2014). TCF3 acts on most of the above players (Marson *et al.*, 2008), effectively amplifying the currently predominant signal in the feedback loop

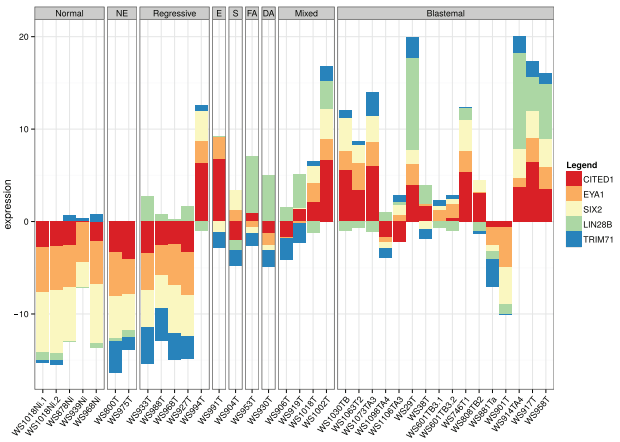

Fig. 5. Expression of TRIM71 and LIN28B in comparison with the cap mesenchyme stem cell markers CITED1, EYA1 and SIX2. Samples are classified as *normal, necrotic (NE), regressive, epithelial (E), stromal (S), focal anaplasia (FA), diffuse anaplasia (DA)* and *blastemal*. The size of a colored bar represents the absolute expression strength of the associated gene

contains categories associated with the WNT pathway (see Supplementary Table S5). This is consistent with its previously reported central role in most WTs and especially in blastema-rich WTs (Fukuzawa *et al.*, 2009). The activation of the canonical WNT pathway usually leads to degradation of the destruction complex that, as long as it is functional, degrades the transcriptional coactivator β-catenin. Thus, degradation of the destruction complex leads to higher amounts of β-catenin in the cytoplasm that is transported to the nucleus where it builds complexes with TCF/LEF proteins. Degradation of the destruction complex lies at the core of developmental processes, ESC self-renewal and differentiation and changes the transcriptional landscape of the cell dramatically.

3.1.5 TCF3 as potential WT master regulator

We argued that factors associated with stem cell fate and the canonical WNT pathway play an essential role in blastemal tumors. A well-known link between the WNT pathway and the core regulatory

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

circuitry of ESCs is TCF3 that together with the pluripotency factors OCT4 (POU5F1), NANOG and SOX2 builds the set of 'ESC master regulators' (Cole *et al.*, 2008). If the WNT pathway is inactive, TCF3 is mainly repressing pluripotency factors and promoting differentiation, however, if the WNT pathway is activated, the repressive complex converts to an activating complex, promoting pluripotency (Cole *et al.*, 2008). To study the influence of ESC master regulators, we constructed a new set of gene categories that we subjected to the KS test using the blastemal versus non-blastemal scores as input. For each of the four transcription factors (TFs), we defined two categories containing genes for which 'strong evidence' exists that they are regulated by the respective TF. In particular, we add a gene to a category for a TF if the TF occupies a site in the gene's promoter region and the correlation between the TF's and the gene's expression is >0.5 (positive category) or smaller than -0.5 (negative category). For the identification of the promoters occupied by the master regulators, we used the mouse ESC ChIP-Chip dataset of Cole *et al.* (2008) and the ChIP-Seq data set of Marson *et al.* (2008). Using this procedure, we obtained more than 1500 genes, including many other TFs and genes involved in ESC fate, influenced by mainly TCF3 and OCT4. For a selection, see Supplementary Table S4. Our KS enrichment (Supplementary Table S20, https://genetrail2.bioinf.uni-sb.de/results.html?session=cbc86903-4248-47a2-b916-bc682924c242) revealed that genes positively regulated by TCF3 ($p \approx 10^{-40}$) and NANOG ($p \approx 10^{-13}$) are strongly enriched, whereas genes negatively regulated by TCF3 ($p \approx 10^{-40}$) are strongly depleted. Conversely, genes positively regulated by OCT4 are strongly depleted, and genes negatively regulated by OCT4 are strongly enriched. This is consistent with a correlation of TCF3 with OCT4 of -0.7. SOX2 and NANOG both seem to be of lesser importance in our data.

However, the four master regulators do not only regulate protein coding genes. Marson *et al.* (2008) revealed that they are also 'associated with promoters for miRNAs that are preferentially expressed in ESCs'. Examples are the miR-302, miR-515 and let-7 families which we previously discussed (cf. Fig. 4). In addition, our data indicate that TCF3 regulates the expression of the miR-17 cluster (all correlations >0.5).

3.1.6 IGF2 as putative WNT activator

In the above section, we have outlined how the ESC regulatory circuitry is driven by TCF3 through the WNT pathway. However, the mechanisms that activate WNT signaling still remain unclear. Whereas certain genetic mutations occur with relatively low frequency ($\leq 30\%$), among them genes that may induce WNT signaling, epigenetic lesions and especially loss of imprinting at the IGF2/H19 locus have been found for 81% of all blastemal subtype WTs (Wegert *et al.*, 2015) leading to an overexpression of IGF2. In 2001, Morali *et al.* (2001) showed that IGF2 can induce the expression and import of β-catenin and TCF3 into the nucleus even in the absence of WNT proteins. This triggers a switchover from the epithelial to mesenchymal cell state, which is in accordance with the expression patterns of the CM stem cell markers shown in Figure 5. In addition, TCF3 binding sites have been found in the IGF2 gene (Cole *et al.*, 2008). Remarkably, we observe an extreme correlation of 0.9 between the TCF3 and IGF2 expression (see Supplementary Fig. S3). This suggests that TCF3 in turn regulates IGF2 leading to a self-sustaining feedback loop which is likely to be causal for the stem cell character of, e.g. blastemal tumors.

4 Discussion

We presented GeneTrail2, a platform for the enrichment analysis of multi-omics datasets. GeneTrail2 was designed to offer users a maximal amount of flexibility while keeping the common workflow accessible to non-expert users. This is achieved by offering a user friendly, well-documented web interface. In turn, scripting capabilities allow expert users to conduct fully automated large-scale analyses and the integration into third-party applications. To further this flexibility, we provide, to the best of our knowledge, the largest number of enrichment algorithms (Fig. 1) available in a web service. This was motivated by the lack of gold standards for enrichment analysis and enables researchers to choose the appropriate tool for the task at hand. Through our comparative enrichment view, we additionally offer a straightforward way for balancing the sensitivity and specificity of analyses. Furthermore, we integrated an extensive collection of biological databases allowing to choose appropriate prior knowledge for a research task. Owing to today's capabilities and ubiquity of high-throughput measurement techniques, we implemented support for the analysis of multi-omics datasets. Moreover, the modular structure of GeneTrail2 ensures that new algorithms, databases, organisms and identifiers can easily be added. This allows to track the current state-of-the-art and continuously improve GeneTrail2.

We demonstrated GeneTrail2's capabilities by applying it to a Wilms tumor mRNA/miRNA expression dataset. Here, the major goal was to determine key molecular features that result in lower susceptibility to preoperative chemotherapy of blastemal subtype tumors. We were able to identify a substantial amount of highly significant, cancer-associated categories that were deregulated in this subtype. The unusual expression profiles of the let-7 miRNA family lead to LIN28B, which is well studied in the context of WTs, and in turn to its upstream regulator TRIM71. We discussed the role of TRIM71 and in particular its function as a stem cell regulator and we observed that its expression is consistent with the expression of the CM stem cell markers EYA1, SIX2 and CITED1 (Fig. 5). These first results indicated that stem cell regulators play a central role in blastemal tumors. This is in agreement with the assumption that the ability of WT cells to maintain stem cell character is a main determinant for tumor malignancy. Furthermore, our results revealed an upregulation of the canonical WNT pathway. Searching for a link between WNT signaling and the regulatory circuitry of stem cells, we arrived at the TF TCF3. To examine possible effects of a TCF3 deregulation, we searched for factors influenced by TCF3. We found a large number of genes, including well-known oncogenes, with TCF3 binding sites in their promoters and exceptionally high correlations with TCF3 expression. Interestingly, IGF2 belongs to this group of genes suggesting that TCF3 regulates IGF2. Loss of imprinting at the H19/IGF2 locus is with an extreme rate of occurrence of 80% the most abundant epigenetic lesion in WTs and leads to an IGF2 overexpression that can activate the canonical WNT pathway and, hence, trigger the epithelial to mesenchymal transition (EMT). Our findings indicate that the resulting TCF3 activation closes a self-sustaining feedback loop further boosting IGF2 production. This offers a potential explanation on how malignant tumor cells maintain their stem cell character. The genes discussed above clearly separate WTs according to tumor malignancy and may hence be promising candidates for prognostic biomarkers that may also be valuable for therapy stratification.

The presented use case shows that GeneTrail2 is able to uncover biologically informative signals in -omics data making it an important tool for the elucidation of pathogenic processes. The ability to

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016

compare multiple enrichment results from the same or different -omics allows to identify differences and similarities between experiments. These capabilities are central for modern *in silico* analyses and set GeneTrail2 apart from other approaches.

Funding

This work was supported by the [SPP 1335] (Scalable Visual Analytics) of the DFG and by the DFG projects [LE 952/5-1] and [LE 952/3-2]; the p-medicine project with funding from the European Union's 7th Framework Program for research, technological development and demonstration [270089]; the SIOP-2001/GPOH clinical trial received financial support by 'Deutsche Krebshilfe' [50-2709-Gr2].

Conflict of Interest: none declared.

References

Ackermann,M. and Strimmer,K. (2009) A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, **10**, 47.

Alexa,A. *et al.* (2006) Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, **22**, 1600–1607.

Anton,R. *et al.* (2011) A systematic screen for micro-RNAs regulating the canonical WNT pathway. *PLoS One*, **6**, e26257.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Backes,C. *et al.* (2007) GeneTrail—advanced gene set enrichment analysis. *Nucleic Acids Res.*, **35**(suppl 2), W186–W192.

Barrett,T. *et al.* (2013) NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Res.*, **41**, D991–D995.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological)*, **57**, 289–300.

Chang,H.M. *et al.* (2012) Trim71 cooperates with microRNAs to repress CDKN1a expression and promote embryonic stem cell proliferation. *Nat. Commun.*, **3**, 923.

Cole,M.F. *et al.* (2008) Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells. *Genes Dev.*, **22**, 746–755.

Drăghici,S. *et al.* (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**, 107–129.

Fukuzawa,R. *et al.* (2009) Canonical WNT signalling determines lineage specificity in Wilms tumour. *Oncogene*, **28**, 1063–1075.

Geistlinger,L. *et al.* (2011) From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics*, **27**, i366–i373.

Gerasch,A. *et al.* (2014) BiNA: a visual analytics tool for biological network data. *PLoS One*, **9**, e87397.

Glaab,E. *et al.* (2012) EnrichNet: network-based gene set enrichment analysis. *Bioinformatics*, **28**, i451–i457.

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Henry,V.J. *et al.* (2014) Omictools: an informative directory for multi-omic data analysis. *Database*, **2014**, bau069.

Holden,M. *et al.* (2008) GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics*, **24**, 2784–2785.

Huang,D.W. *et al.* (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Hung,J.H. *et al.* (2011) Gene set enrichment analysis: performance evaluation and usage guidelines. *Brief. Bioinform.*, **13**, 281–291.

Khatri,P. *et al.* (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol*, **8**, e1002375.

Kim,S.Y. and Volsky,D.J. (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, **6**, 144.

Laczny,C. *et al.* (2012) miRTrail—a comprehensive webserver for analyzing gene and miRNA patterns to enhance the understanding of regulatory mechanisms in diseases. *BMC Bioinformatics*, **13**, 36.

Lee,S.H. *et al.* (2014) The ubiquitin ligase human Trim71 regulates let-7 microRNA biogenesis via modulation of Lin28b protein. *Biochimica Et Biophysica Acta (BBA) Gene Regulatory Mechanisms*, **1839**, 374–386.

Leek,J.T. *et al.* (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, **11**, 733–739.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Marson,A. *et al.* (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, **134**, 521–533.

Mogilyansky,E. and Rigoutsos,I. (2013) The mir-17/92 cluster: a comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease. *Cell Death Differ.*, **20**, 1603–1614.

Morali,O.G. *et al.* (2001) Igf-ii induces rapid beta-catenin relocation to the nucleus during epithelium to mesenchyme transition. *Oncogene*, **20**, 4942–4950.

Naeem,H. *et al.* (2012) Rigorous assessment of gene set enrichment tests. *Bioinformatics*, **28**, 1480–1486.

Newton,M.A. *et al.* (2007) Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann. Appl. Stat.*, **1**, 85–106.

Opgen-Rhein,R. and Strimmer,K. (2007) Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat. Appl. Genet. Mol. Biol.*, **6**, Article9

Piskounova,E. *et al.* (2011) Lin28b and Lin28b inhibit let-7 microrna biogenesis by distinct mechanisms. *Cell*, **147**, 1066–1079.

Polakis,P. (2012) Wnt signaling in cancer. *Cold Spring Harb. Perspect. Biol.*, **4**, a008052.

Risso,D. *et al.* (2014) Normalization of rna-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.*, **32**, 896–902.

Robinson,M.D. *et al.* (2010) edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Sass,S. *et al.* (2014) Ramona: a web application for gene set analysis on multi-level omics data. *Bioinformatics*,**31**, 128–130.

Sredni,S.T. *et al.* (2009) Subsets of very low risk Wilms tumor show distinctive gene expression, histologic, and clinical features. *Clin. Cancer Res.*, **15**, 6800–6809.

Stöckel,D. *et al.* (2013) NetworkTrail—a web service for identifying and visualizing deregulated subnetworks. *Bioinformatics*, **29**, 1702–1703.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. U. S. A.*, **102**, 15545–15550.

Tian,L. *et al.* (2005) Discovering statistically significant pathways in expression profiling studies. *Proc. Natl Acad. Sci. U. S. A.*, **102**, 13544–13549.

Urbach,A. *et al.* (2014) Lin28 sustains early renal progenitors and induces Wilms tumor. *Genes Dev.*, **28**, 971–982.

Wegert,J. *et al.* (2015) Mutations in the six1/2 pathway and the drosha/dgcr8 miRNA microprocessor complex underlie high-risk blastemal type Wilms tumors. *Cancer Cell*, **27**, 298–311.

Wolstencroft,K. *et al.* (2013) The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. *Nucleic Acids Res.*, **41**, W557–W561.

Yang,T.P. *et al.* (2010) Genevar: a database and java application for the analysis and visualization of snp-gene associations in eqtl studies. *Bioinformatics*, **26**, 2474–2476.

Zhang,S. *et al.* (2010) GO-bayes: gene ontology-based overrepresentation analysis using a Bayesian approach. *Bioinformatics*, **26**, 905–911.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of California, Los Angeles on August 30, 2016